
Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

τομεας σηματων, ελεγχου και ρομποτικης


# Τροπική Γεωμετρία και Εφαρμογές σε Μηχανική Μάθηση και Βελτιστοποίηση

## ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του


**Δημητριάδη Νικόλαου**

**Επιβλέπων:** Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

ΕΡΓΑΣΤΗΡΙΟ ΟΡΑΣΗΣ ΥΠΟΛΟΓΙΣΤΩΝ, ΕΠΙΚΟΙΝΩΝΙΑΣ ΛΟΓΟΥ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑΣ ΣΗΜΑΤΩΝ



Αθήνα, Οκτώβριος 2020

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

# Τροπική Γεωμετρία και Εφαρμογές σε Μηχανική Μάθηση και Βελτιστοποίηση

## ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

**Δημητριάδη Νικόλαου**

**Επιβλέπων:** Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 7η Οκτωβρίου, 2020.

| | | |
|---|---|---|
| ..... | ..... | ..... |
| Πέτρος Μαραγκός | Αριστείδης Παγουρτζής | Αλέξανδρος Ποταμιάνος |
| Καθηγητής Ε.Μ.Π. | Καθηγητής Ε.Μ.Π. | Αναπληρωτής Καθηγητής Ε.Μ.Π. |

Αθήνα, Οκτώβριος 2020

.....
**ΔΗΜΗΤΡΙΑΔΗΣ ΝΙΚΟΛΑΟΣ**
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

*στους γονείς μου,*
*Ματούλα και Στάθη*

# Ευχαριστίες

Στο σημείο αυτό ολοκληρώνεται ένας πολύ σημαντικός κύκλος της ζωής μου και κλείνει το ιδιαίτερα όμορφο κεφάλαιο των προπτυχιακών σπουδών στο Εθνικό Μετσόβιο Πολυτεχνείο. Η πορεία αυτή εμπλουτίστηκε από πολλούς ανθρώπους, τους οποίους νιώθω την ανάγκη να ευχαριστήσω από τα βάθη της καρδιάς μου.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της παρούσας διπλωματικής εργασίας, κ. Πέτρο Μαραγκό. Ο χρόνος που αφιέρωσε, η καθοδήγηση που παρείχε ανελλιπώς, η υπομονή του και η παρότρυνσή του να ασχοληθώ ερευνητικά με ζητήματα που με ενδιαφέρουν αποτελούν καθοριστικούς παράγοντες για τη διεκπεραίωση της διπλωματικής και, φυσικά, θα με συνοδεύουν στα επόμενα βήματά μου. Όσον αφορά αυτά τα βήματα, τον ευχαριστώ πραγματικά για την πολύτιμη και ανεκτίμητη αρωγή τους στις αιτήσεις για διδακτορικές σπουδές και την προθυμία του να συμβάλει σε αυτή την προσπάθεια.

Θα ήθελα να ευχαριστήσω τους φίλους μου που βρίσκονται στο πλευρό μου χρόνια και ομορφαίνουν τη ζωή μου. Τα παιδιά από το σχολείο, τον Αριστοτέλη, το Γιώργο και το Στέφανο, το Χρήστο και το Στέφανο που είναι συνοδοιπόροι από την τρυφερή ηλικία του νηπιαγωγείου, που μεγαλώσαμε μαζί και έχουμε μοιραστεί τόσες ωραίες στιγμές. Ακόμη, τους ανθρώπους που γνώρισα στα φοιτητικά μου χρόνια, τον Τόλη, το Γιώργο και το Στέλιο, το Βασίλη και το Βασίλη, το Θανάση που αποτελεί την πρώτη επιλογή συνεργάτη σε όλες τις εργασίες και τον Αστέριο που βαδίσαμε μαζί το μονοπάτι για τις αιτήσεις και πολλά άλλα. Τους ευχαριστώ πολύ ιδιαίτερα για την τελευταία περίοδο που υπέμειναν τις ιδιοτροπίες μου και τις ανησυχίες μου σε καθημερινή βάση και με βοήθησαν με τις πολύτιμες συμβουλές τους.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου που από μικρή ηλικία έθεσαν τα θεμέλια για την πορεία μου, με παρότρυναν να αδράξω ευκαιρίες και προσέφεραν καθοδήγηση και ενθάρρυνση σε κάθε βήμα. Στέκονται δίπλα μου στα καλά και τα κακά, πάντα με υπομονή και αγάπη. Μέσα από προσωπικές και καθημερινές θυσίες καλλιέργησαν το περιβάλλον που μου επιτρέπει να κυνηγήσω τα όνειρά μου.

Σας ευχαριστώ πολύ.

Δημητριάδης Νικόλαος
Οκτώβριος 2020

# Περίληψη

Η μηχανική μάθηση έχει γνωρίσει ιδιαίτερη άνθηση την τελευταία δεκαετία με το ερευνητικό ενδιαφέρον να επικεντρώνεται στις βαθιές αρχιτεκτονικές. Στο πλαίσιο αυτό, το υποκείμενο μαθηματικό υπόβαθρο δεν είναι σαφώς θεμελιωμένο, με αποτέλεσμα οι διάφοροι μέθοδοι να χρησιμοποιούνται ως μαύρα κουτιά με ελλιπή κατανόηση των εσωτερικών τους διεργασιών. Τα τελευταία χρόνια, λοιπόν, ένθερμες προσπάθειες και σημαντικός ερευνητικός ζήλος κατευθύνονται στην εισαγωγή και μελέτη μαθηματικών εργαλείων που προσφέρουν εξηγήσεις και διαίθηση στα εργαλεία της μηχανικής μάθησης. Μία από αυτές τις προσεγγίσεις εντοπίζεται στον κλάδο της τροπικής γεωμετρίας καθώς και των μορφολογικών μαθηματικών.

Με αφετηρία την τομή μεταξύ μαθηματικής βελτιστοποίησης και τροπικής γεωμετρίας, εξερευνούμε προβλήματα με τροπικούς περιορισμούς. Αναδεικνύεται η μη-κυρτή φύση τους και σκιαγραφείται μία μέθοδος διάβασης από την τροπική και μη-κυρτή βελτιστοποίηση σε μία συλλογή προβλημάτων με κυρτούς περιορισμούς.

Στη συνέχεια, στρέφουμε την προσοχή μας στα μορφολογικά νευρωνικά δίκτυα, τα οποία θεμελιώνονται σε μη-γραμμικούς νευρώνες διαστολής και συστολής. Εμπλουτίζουμε την εκφραστικότητά τους συνδυάζοντας πολυάριθμους νευρώνες τόσο διαστολής και συστολής στο κρυφό επίπεδο και συνδέουμε το σύνορο απόφασης τους με την τροπική γεωμετρία. Μελετώνται δύο παραλλαγές των Πυκνών Μορφολογικών Δικτύων. Η πρώτη αφορά την επέκταση σε βαθιές αρχιτεκτονικές, όπου σχηματίζονται ενεργοποιήσεις εφάμιλλες του μορφολογικού ανοίγματος και κλεισίματος. Η δεύτερη εντοπίζεται στη χρήση της αποκβαντοποίησης κατά Maslov για την ομαλοποίηση των επιφανειών και την αντιμετώπιση του προβλήματος διάδοσης κλίσης που χαρακτηρίζει η μη-παραγωγισμότητα των μορφολογικών τελεστών.

Επιπρόσθετα, διατυπώνεται τρόπος έκφρασης ενός προβλήματος ταξινόμησης δυαδικών προτύπων από ένα perceptron διαστολής-συστολής στη γλώσσα της μη-κυρτής βελτιστοποίησης. Βασιμένη στη θεωρία πλεγμάτων, η προσέγγιση αυτή υποθέτει μερική διάταξη των δεδομένων, ζήτημα που ανταπεξέρχεται με χρήση μειωμένης διάταξης. Στη συνέχεια, επεκτείνεται η δυαδική φύση του ταξινομητή σε διεργασίες πολλών κλάσεων.

Παράλληλα, εξερευνούνται οι δυνατότητες των Μορφολογικών Δικτύων με μεθόδους εκπαίδευσης βασισμένες στην Κατάβαση Κλίσεων. Παρουσιάζονται αριθμητικά πειράματα σε γνωστά benchmarks για διάφορες επιλογές αρχιτεκτονικών, ακόμα και για τις ομαλοποιημένες εκδόσεις των μορφολογικών τελεστών ως νευρώνες. Οι προσπάθειες επικεντρώνονται στην αραιότητα των μοντέλων αξιολογώντας την απόδοσή τους κατόπιν pruning. Εξετάζονται δύο μέθοδοι αφαίρεσης παραμέτρων και συγκρίνεται η διατήρηση πληροφορίας των μορφολογικών δικτύων με τα αντίστοιχα πλήρως συνδεδεμένα νευρωνικά δίκτυα, καταλήγοντας στο συμπέρασμα ότι τα μορφολογικά δίκτυα κωδικοποιούν πολύ οικονομικά τις υποκείμενες αναπαραστάσεις των δεδομένων.

Τέλος, τα τελευταία χρόνια αναπτύσσεται η τάση της ερμηνευσιμότητας, όπου οι μελετητές επιθυμούν εγγυήσεις για τη μορφή της εξόδου των μοντέλων. Υπό το πρίσμα αυτό, μελετάται η ιδιότητα της μονοτονίας. Προτείνεται μία αμιγώς τροπικά αλγεβρική επίλυση του προβλήματος μέσω εναλλάξ min-plus και max-plus παλινδρόμησης. Ακόμη, εξερευνούμε την ικανοποίηση του περιορισμού μέσω συγκεκριμένων αρχιτεκτονικών επιλογών σε μορφολογικά δίκτυα, εξετάζοντας και ομαλοποιημένους τελεστές.

**Λέξεις Κλειδιά** — Τροπική Γεωμετρία, Μορφολογικά Μαθηματικά, Τροπική Βελτιστοποίηση, Νευρωνικά Δίκτυα, Μηχανική Μάθηση, Ερμηνευσιμότητα, Μονοτονία

# Abstract

Machine learning has flourished over the last decade with research focusing on deep architectures. In this context, the underlying mathematical background is not well-founded, resulting in the use of various methods as black boxes without a complete understanding of their internal processes. In recent years, therefore, intense efforts and significant research zeal have been directed towards the introduction and study of mathematical tools that offer explanations and intuition to machine learning tools. One of these approaches lies on the field of tropical geometry as well as morphological mathematics.

Starting with the intersection between mathematical optimization and tropical geometry, we explore problems with tropical constraints. Their non-convex nature is highlighted and a method of transition from tropical and non-convex optimization to a collection of problems with convex constraints is outlined.

Next, we turn our attention to morphological neural networks, which are based on morphological dilation and erosion neurons. We enrich their expressiveness by combining numerous neurons, both dilation and erosion, at the hidden level and express their decision boundary with tropical geometry. Two variants of Dense Morphological Networks are studied. The first concerns the extension to deep architectures, where activations resembling morphological opening and closing are formed. The second is found in the use of Maslov Dequantization to smooth surfaces and to address the problem of gradient propagation characterized by the non-differentiability of morphological operators.

Additionally, a way of formulating a binary pattern classification problem by a dilation-erosion perceptron is expressed in the language of non-convex optimization. Based on lattice theory, this approach presupposes a partial ordering of the data, an issue that is alleviated using a reduced ordering. Next, the binary nature of the classifier is extended to multiclass problems.

At the same time, the capabilities of Morphological Networks are explored with training methods based on Gradient Descent. Numerical experiments are presented in known benchmarks for various architectural choices, even for softened versions of morphological operators as neurons. Efforts focus on the sparsity of the models by evaluating their performance after pruning. Two methods of parameter pruning are examined and the information retention of the morphological networks is compared with the corresponding fully connected neural networks, concluding that the morphological networks encode the underlying data representations more economically.

Finally, in recent years the trend of interpretability has been developing, where researchers want guarantees for the form of output of the models. In this light, the property of monotonicity is studied. A purely tropical algebraic solution to the problem is proposed through alternating min-plus and max-plus regression. We also explore satisfying the constraints via specific architectural choices in morphological networks, and consider softened operators.

**Keywords** — Tropical Geometry, Morphological Mathematics, Tropical Optimization, Neural Networks, Machine Learning, Interpretability, Monotonicity

# Περιεχόμενα

# Λίστα Σχημάτων

# Κατάλογος Πινάκων

# Χρωματικός Κώδικας

Παρακάτω παρατίθενται ο "χρωματικός κώδικας" που θα χρησιμοποιηθεί.

**Theorem 0.0.1: τίτλος**

> απλό θεώρημα

**Definition 0.0.2: τίτλος**

> Απλός ορισμός

**Lemma 0.0.3: τίτλος**

> Απλό λήμμα

**Example 0.0.4: τίτλος**

> Απλό παράδειγμα

# Κεφάλαιο 1

# Εισαγωγή

## 1.1 Μηχανική Μάθηση

Η μάθηση ως απόρροια του συνδυασμού προσωπικής εμπειρίας και γνώσης και η διάδοσή της από γενιά σε γενιά κατέχει κεντρικό ρόλο στην ανθρώπινη εξέλιξη και ευφυία. Η επιστημονική μοντελοποίηση, παρομοίως, αποσκοπεί στην επεξήγηση του περιβάλλοντος και των πειραματικών δεδομένων. Για παράδειγμα, ο επιστημονικός κλάδος της Φυσικής θεμελιώνεται στη λογική της παρατήρησης και στο σχεδιασμό μαθηματικών θεωριών που την εξηγούν. Ανέκαθεν, ωστόσο, επίκεντρο της επιστημονικές σκέψης ήταν η κατασκευή μηχανών με τη δυνατότητα σκέψης.

Με την αυγή των υπολογιστών κατά τον 20ο αιώνα, αυτές οι τάσεις άρχισαν να αγγίζουν τη σφαίρα της πραγματικότητας. Με τα χρόνια, εδραιώθηκε και εξελίχθηκε ο κλάδος της *τεχνητής νοημοσύνης* (artifical intelligence - AI). Πλέον, η καθημερινή ζωή είναι άρρηκτα συνδεδεμένη με εργαλεία της τεχνητής νοημοσύνης, όπως η αυτοματοποίηση διεργασιών, η διάγνωση ασθενειών και η κατανόηση εικόνων και φυσικής γλώσσας.

Τα πρώτα βήματα είδαν την αντιμετώπιση προβλημάτων που κρίνονται διανοητικά δύσκολα για τον ανθρώπινο νου αλλά με ευθύ τρόπο επίλυσης, καθώς χαρακτηρίζονται από σύνολο αυστηρών κανόνων. Μία από τις πρώτες και πιο γνωστές επιτυχίες του κλάδου εντοπίζεται στο σχεδιασμό του συστήματος *Deep Blue* που επιβλήθηκε του παγκόσμιου πρωταθλητή σκάκι Garry Kasparov το 1997. Ωστόσο, αυτός ο υπολογιστής δεν είχε δυνατότητα γενίκευσης, καθώς η διεργασία προς επίλυση είναι ιδιαίτερα απλή και αποτελείται από 64 τετράγωνα και 32 κομμάτια. Συνεπώς, είναι εφικτή η περιγραφή του συστήματος με ένα απλό σύνολο κανόνων. Αν και οι υπολογιστές μπορούν να επιλύσουν προβλήματα καλώς ορισμένα, ακόμα και αφηρημένα όπως το σκάκι, πολύ καλύτερα από τον άνθρωπο, η απόδοσή τους σε προβλήματα καθημερινά, όπως η αναγνώριση αντικειμένων και κειμένων, ωχριούσε μέχρι και την προηγούμενη δεκαετία.

Η πραγματική πρόκληση για την Τεχνητή Νοημοσύνη, λοιπόν, είναι η επίλυση γενικών προβλημάτων, τα οποία ίσως να είναι εύκολα για τον άνθρωπο αλλά απαιτούν κατανόηση του περιβάλλοντος. Ένα απλό παράδειγμα που επιλύθηκε[1] την τελευταία δεκαετία είναι η κατηγοριοποίηση ψηφίων από εικόνες. Σε αυτό το πλαίσιο είναι αδύνατη η κατασκευή αυστηρού συνόλου κανόνων με το χέρι και απαιτείται η στροφή σε άλλου είδους αλγορίθμους που χρήζουν τους υπολογιστές ικανούς να αποκτήσουν γνώση εξάγοντας πρότυπα μέσα από τα δεδομένα. Αυτή η οικογένεια αλγορίθμων αντιστοιχεί στη *μηχανική μάθηση* (machine learning). Απλοί αλγόριθμοι αντιμετωπίζουν προβλήματα της καθημερινότητας που απαιτούν κατανόηση του πραγματικού κόσμου, όπως ο διαχωρισμός της ηλεκτρονικής αλληλογραφίας σε χρήσιμη και μη (spam mail filtering). Ο Arthur Samuel, ένας από τους πρωτοπόρους της επιστήμης, όρισε τον όρο ως:

> *Η Μηχανική Μάθηση είναι το πεδίο σπουδών*
> *που προσδίδει στον υπολογιστή τη δυνατότητα να μαθαίνει*
> *δίχως να προγραμματίζεται ρητά.*

Στη μηχανική μάθηση, συνεπώς, το βάρος μετατοπίζεται από τη δημιουργία κανόνων στην κατασκευή χαρακτηριστικών (features). Το σύνολο των χαρακτηριστικών διαμορφώνει την αναπαράσταση των δεδομένων (representation). Η επιλογή της αναπαράστασης αποτελεί το κλειδί της απόδοσης στη μηχανική μάθηση. Σε ορισμένα προβλήματα, η επιλογή των χαρακτηριστικών δεν δημιουργεί προκλήσεις. Για παράδειγμα, στην πρόβλεψη της τιμής της μετοχής της εταιρίας $X$ είναι λογικό να συμπεριληφθούν τα έσοδα και τα έξοδά της. Παρομοίως, για την αναγνώριση ομιλητή μέσω ήχου κρίνεται χρήσιμο χαρακτηριστικό η εκτίμηση του μεγέθους της φωνητικής οδού ή των συχνοτήτων ομιλίας ώστε να συμπεράνουμε αν ο ομιλητής είναι άντρας, γυναίκα ή παιδί.

Ωστόσο, η εξαγωγή μίας ορθής αναπαράστασης δεν είναι τόσο άκοπη σε άλλα πεδία εφαρμογών. Χαρακτηριστικό παράδειγμα είναι η όραση υπολογιστών. Έστω ότι σκοπός μας είναι η ανίχνευση προσώπων σε μία φωτογραφία. Ένα πολύ σημαντικό και προεξέχον χαρακτηριστικό είναι η παρουσία μύτης ή ματιών. Ωστόσο, η περιγραφή τέτοιων χαρακτηριστικών σε μορφή pixels είναι δύσκολη. Ο ανθρώπινος νους λαμβάνει μεν ένα σήμα παρόμοιο με αυτό μίας συλλογής pixels αλλά η διεργασία που ακολουθεί για την εύρεση των περιοχών που εντοπίζονται πρόσωπα διέρχεται πολλά στάδια, ή *επίπεδα*, προτού ληφθεί η απόφαση. Επομένως, μία ιδέα είναι η μοντελοποίηση συστημάτων που μιμούνται αυτή τη διαδικασία.

Μία λύση έγκειται στη χρήση μηχανικής μάθησης για την ανακάλυψη όχι μόνο του τρόπου με τον οποίο η αναπαράσταση παράγει την έξοδο αλλά και στην εξαγωγή της ίδιας της αναπαράστασης. Χρησιμοποιούμε, λοιπόν, low-level χαρακτηριστικά, όπως οι τιμές των pixels, και επιτρέπουμε στον υπολογιστή να κατασκευάσει high-level αναπαραστάσεις, όπως μάτια και φρύδια, εκφρασμένες ως συνδυασμός απλότερων αναπαραστάσεων. Αυτή

[1] δηλαδή υπάρχουν αλγόριθμοι που επιλύουν με ακρίβεια ανάλογη του μέσου ανθρώπου

η πειθαρχία ονομάζεται *Βαθιά Μάθηση* (Deep Learning), καθώς τα ακατέργαστα δεδομένα περνούν από πολλά επίπεδα πριν την πρόβλεψη, παράγοντας "βαθιές" αρχιτεκτονικές. Η γέννηση του κλάδου πυροδοτήθηκε από αξιοσημείωτη πρόοδο στις δυνατότητες των υπολογιστών και στη χρήση καρτών γραφικών (Graphics Processing Units - GPUs) για την εκπαίδευση των μοντέλων.

Χαρακτηριστικό παράδειγμα αποτελεί το πρόσθιο νευρωνικό δίκτυο ή το πολυεπίπεδο perceptron ή, στη διεθνή ορολογία, Multi-Layer Perceptron - MLP (βλ. κεφάλαιο 4). Συνοπτικά, πρόκειται για μία απεικόνιση (mapping) από ένα διάνυσμα εισόδου σε τιμές εξόδου που κατασκευάζεται ως σύνθεση πολλών απλότερων συναρτήσεων. Στα πιο βαθιά επίπεδα, δηλαδή, η συνάρτηση χρησιμοποιεί ολοένα και πιο πολύπλοκα και εκλεπτυσμένα χαρακτηριστικά. Επιστρέφουμε στο παράδειγμα αναγνώρισης προσώπου. Στο επίπεδο εισόδου, το MLP λαμβάνει τις τιμές των pixel. Στο πρώτο κρυφό επίπεδο εντοπίζει ακμές, στο δεύτερο γωνίες και περιγράμματα, στο τρίτο μέρη αντικειμένων. Συνδυάζοντας τις εξόδους του τρίτου επιπέδου, είναι η δυνατή η πρόβλεψη.

Αξίζει, λοιπόν, να εντρυφήσουμε στα προβλήματα που αντιμετωπίζει ο κλάδος της μηχανικής μάθησης. Η κατηγοριοποίηση των αλγορίθμων έγκειται στη δοκιμασία που χρησιμοποιούνται. Η ανάλυση περιορίζεται στις βασικές κατηγορίες και γίνεται μία σύντομη αναφορά σε αξιοσημείωτες περιπτώσεις.

### 1.1.1 Επιβλεπόμενη Μάθηση

Στα προβλήματα επιβλεπόμενης μάθησης, κάθε στοιχείο του συνόλου δεδομένων συνδεύεται από μία επισημείωση (label). Οι αλγόριθμοι στοχεύουν στην πρόβλεψη αυτής της τιμής. Συνεπώς, σε αυτή την κλάση προβλημάτων, η αξιολόγηση της απόδοσης των αλγορίθμων είναι διαισθητικά απλή και έγκειται στην απόκλιση από τις πραγματικές τιμές (ground truth). Το είδος της επισημείωσης καθορίζει τη διεργασία.

#### Ταξινόμηση

Στα προβλήματα ταξινόμησης (classification), οι επισημειώσεις λαμβάνουν διακριτές τιμές. Κάθε τιμή αντιστοιχεί σε μία κατηγορία αντικειμένων, ή κλάση, όπως σκύλος ή γάτα. Σκοπός, λοιπόν, είναι η κατασκευή ενός συστήματος που προβλέπει την κλάση ενός διανύσματος δεδομένων εισόδου. Για παράδειγμα, σε συστήματα οπτικής αναγνώρισης χαρακτήρων (Optical Character Recognition - OCR) κάθε εικόνα αντιστοιχεί σε ένα γράμμα της αλφαβήτου, ενώ σε διαγνωστικά εργαλεία ο στόχος μπορεί να είναι η κατηγοριοποίηση των όγκων σε καλοηθείς και κακοηθείς. Σε προβλήματα ταξινόμησης επικρατεί το πιθανοκρατικό μοντέλο, όπου σε κάθε κλάση ανατίθεται μία πιθανότητα και επιλέγεται η κατηγορία με τη μεγαλύτερη τιμή. Ο χώρος εισόδου, λοιπόν, διαιρείται σε περιοχές όπου κάθε κατηγορία υπερισχύει. Το μεταίχμιο δύο περιοχών ονομάζεται *σύνορο απόφασης*. Ένα παράδειγμα απεικονίζεται στο σχήμα 1.1.1a. Μία γραμμή διαχωρίζει τα πρότυπα και η θέση ενός σημείου στο χώρο καθορίζει την πρόβλεψη.

#### Παλινδρόμηση

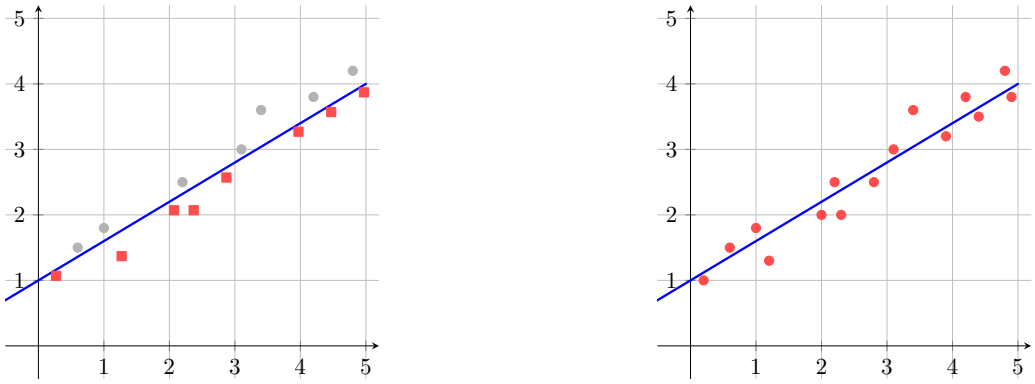
Στα προβλήματα παλινδρόμησης (regression), η έξοδος (επισημείωση) δεν είναι διακριτή αλλά λαμβάνει συνεχείς τιμές. Στόχος είναι η εκτίμηση μίας συνάρτησης $f$ που ταιριάζει βέλτιστα στα δεδομένα. Υπάρχουν διάφορα κριτήρια αξιολόγησης (βλ. §4.3). Η διεργασία της παλινδρόμησης έχει πλούσια ιστορία και πρόδρομός της μπορεί να θεωρηθεί η ειδική περίπτωση της μεθόδου ελαχίστων τετραγώνων. Το φάσμα των εφαρμογών είναι αξιοσημείωτα ευρύ. Για παράδειγμα, έστω το πρόβλημα πρόβλεψης της τιμής μίας οικείας με δεδομένα όπως εμβαδόν σε $m^2$, ταχυδρομικός κώδικας περιοχής και χρονιά κατασκευής. Στο σχήμα 1.1.1b παρουσιάζεται ένα απλό παράδειγμα όπου μία ευθεία προσαρμόζεται στα δεδομένα ώστε να τα "εξηγεί" βέλτιστα.

### 1.1.2 Μη-Επιβλεπόμενη Μάθηση

Σε αντίθεση με την επιβλεπόμενη μάθηση, αυτή η κατηγορία δε φέρει επισημειώσεις στα δεδομένα. Αντί, λοιπόν, ο στόχος να είναι η πρόβλεψη κάποιας τιμής, οι διεργασίες της μη-επιβλεπόμενης μάθησης επικεντρώνονται στην ανακάλυψη της δομής των δεδομένων αναλύοντας τις συσχετίσεις μεταξύ των ελεύθερων μεταβλητών εισόδου. Ακολουθεί η περιγραφή ορισμένων ιδιαίτερα σημαντικών εφαρμογών.

#### Συσταδοποίηση

Η συσταδοποίηση (clustering) των δεδομένων αφορά την ομαδοποίησή τους σε διακριτές ομάδες. Σε αντίθεση με την ταξινόμηση, οι κλάσεις δεν είναι γνωστές εκ των προτέρων. Αντίθετα, οι αλγόριθμοι εξερευνούν τα δεδομένα




(a) Ταξινόμηση

(b) Παλινδρόμηση

Σχήμα 1.1.1: Προβλήματα Επιβλεπόμενης Μηχανικής Μάθησης

για ομοιότητες και διαφορές ανάμεσα στα πρότυπα προσπαθώντας να τα κατατάξουν ανάλογα. Η συσταδοποίηση κατέχει κεντρικό ρόλο σε πεδία εφαρμογών όπως έρευνα αγοράς, όπου οι μελετητές προσπαθούν να διασπάσουν τον πληθυσμό σε ομάδες με κοινά χαρακτηριστικά με σκοπό να κατευθύνουν τη διαφήμιση και την ανάπτυξη προϊόντων. Παρομοίως, η συσταδοποίηση χρησιμοποιείται στον εντοπισμό κοινοτήτων στα κοινωνικά δίκτυα.

#### Μείωση Διαστατικότητας

Στην εποχή των Μεγάλων Δεδομένων (Big Data), είναι πολλές φορές ανώφελο να συμπεριλάβουμε όλες τις διαθέσιμες πηγές πληροφορίας στην ανάπτυξη συστημάτων. Κρίνεται επιθυμητή, λοιπόν, η απομάκρυνση των μεταβλητών που φέρουν μη χρήσιμη πληροφορία και η διατήρηση των χρήσιμων πηγών. Αυτή η διεργασία αντιστοιχεί στη μείωση διαστατικότητας (dimensionality reduction). Ένας από τους πιο γνωστούς αλγορίθμους είναι η Ανάλυση σε Κύριες Συνιστώσες (Principal Component Analysis - PCA), όπου επιλέγονται οι μεταβλητές (που πιθανόν να διαφέρουν από τις αρχικές ελεύθερες μεταβλητές και να αντιστοιχούν σε γραμμικό συνδυασμό τους) που χαρακτηρίζονται από τη μέγιστη διακύμανση.

### 1.1.3 Ενισχυτική Μάθηση

Η ενισχυτική μάθηση (reinforcement learning) είναι τομέας της μηχανικής μάθησης με αλγορίθμους που μιμούνται τη διαδικασία μάθησης που ακολουθεί ο ανθρώπινος οργανισμός [SB98]. Ένα σύστημα αλληλεπιδρά με ένα δυναμικό περιβάλλον ώστε να επιτευχθεί κάποιος στόχος. Ωστόσο, δεν υπάρχει κάποια επισημείωση που να ορίζει σωστή συμπεριφορά όπως στην επιβλεπόμενη μάθηση. Έστω, για παράδειγμα, ένα σύστημα που μαθαίνει να παίζει ένα παιχνίδι όπως τα κλασσικά παιχνίδια της `ATARI` [Mni+13]. Σε αυτή την περίπτωση, η αξιολόγηση του μοντέλου συμπίπτει με το σκορ που επιτυγχάνει.

## 1.2 Στόχοι και Οργάνωση Εργασίας

Στο πλαίσιο της Βαθιάς Μηχανικής Μάθησης, παρατηρήθηκε εμφανής άνθιση των εφαρμογών, καθώς οι μηχανικοί δε χρειάζεται να τροφοδοτήσουν έξυπνα χαρακτηριστικά στο μοντέλο αλλά ακατέργαστα δεδομένα και να αφήσουν την υπολογιστική ισχύ να ανιχνεύσει την αναπαράσταση. Σε αυτό το σημείο έχουμε φτάσει σήμερα όπου ο ζήλος δημιουργίας εφαρμογών αποτελεί τον κινητήριο μοχλό της έρευνας. Ωστόσο, τα θεωρητικά θεμέλια των μοντέλων είναι ελλιπή και η κατανόηση των εσωτερικών διεργασιών των σύγχρονων σύνθετων αρχιτεκτονικών της μηχανικής μάθησης παρουσιάζει μεγάλα κενά. Κατά συνέπεια, απουσιάζει η διαίσθηση στην εξαγωγή συμπερασμάτων για τις λειτουργίες των μοντέλων.

Αναλυτικότερα, τα μη-γραμμικά μαθηματικά διέπουν τον πυρήνα των μοντέλων της μηχανικής μάθησης, το νευρώνα. Η μη-γραμμικότητα εντοπίζεται στη συνάρτηση ενεργοποίησης, η οποία εμπεριέχει την πράξη maximum στην πλειονότητα των σύγχρονων αρχιτεκτονικών. Άρα, τα μαθηματικά οικοδομήματα που αναδύονται είναι συμβατά με την τροπική γεωμετρία, η οποία αντιστοιχεί σε min-plus και max-plus μαθηματική θεώρηση.

Με αφετηρία αυτή τη σημαντική παρατήρηση, κρίνουμε σκόπιμο στο πλαίσιο της παρούσας διπλωματικής να εξερευνήσουμε αυτή τη σύνδεση που προσφέρει μαθηματική θεμελίωση και κατανόηση των συστημάτων της μηχανικής μάθησης από μία διαφορετική οπτική γωνία με στιβαρά μαθηματικά θεμέλια.

Με υπόβαθρο τα τροπικά μαθηματικά μελετάται η μοντελοποίηση προβλημάτων τροπικής βελτιστοποίησης και αναζητείται μία διάβαση από την τροπική στην κλασική θεώρηση. Στη συνέχεια, αξιοποιώντας εργαλεία από το γενικότερο πλαίσιο της Θεωρίας Πλεγμάτων, στρεφόμαστε σε νευρωνικά δίκτυα και εφαρμογές μηχανικής μάθησης. Μελετάται η σύνδεση των μορφολογικών τελεστών min και max με τις περιοχές απόφασης που δημιουργούν και πως αρχιτεκτονικές νευρωνικών δικτύων μπορούν να επιβάλλουν επιθυμητά χαρακτηριστικά στη συνάρτηση εξόδου, όπως μονοτονία. Τέλος, ιδιαίτερη προσοχή επικεντρώνεται στην αραιότητα που επιφέρει η χρήση μορφολογικών νευρώνων και η εκμετάλλευση αυτής της ιδιότητας για αποτελεσματική συμπίεση του δικτύου με τεχνικές pruning.

Η διπλωματική εργασία χωρίζεται σε 7 κεφάλαια. Ακολουθεί μία επιγραμματική παρουσίαση των περιεχομένων κάθε κεφαλαίου:

- Στο κεφάλαιο 1 πραγματοποείται μία εισαγωγή στην εργασία, αναλύοντας το αντικείμενο, τη δομή και τους στόχους. Συμπεριλαμβάνεται μία επισκόπηση του κλάδου της μηχανικής μάθησης, τα προβλήματα του οποίου κινητοποιούν τις εφαρμογές που αναλύονται στη συνέχεια.

- Στο κεφάλαιο 2 παρουσιάζεται η θεωρία των τροπικών μαθηματικών που αποτελεί το θεμέλιο λίθο για τη μετέπειτα ανάλυση. Η ανάλυση των επόμενων κεφαλαίων πραγματοποείται μέσα από το τροπικό πρίσμα. Συμπεριλαμβάνονται παραδείγματα για την καλύτερη κατανόηση των εννοιών. Επιπρόσθετα, εισάγονται γεωμετρικές έννοιες όπως ευθείες και παραβολές στην τροπική άλγεβρα και συγκρίνονται με τα γραμμικά τους ανάλογα.

- Στο κεφάλαιο 3 παρουσιάζεται μία περιήγηση του κλάδου της βελτιστοποίησης, εστιάζοντας σε υποκατηγορίες κρίσιμες στην τροπική ανάλυση αλγορίθμων, όπως ο προγραμματισμός διαφοράς κυρτών συναρτήσεων (Difference of Convex Programming). Ακόμη, προτείνεται ένας τρόπος μετατροπής τροπικών προβλημάτων βελτιστοποίησης σε κλασικά προβλήματα, ξεκλειδώνοντας τα μαθηματικά εργαλεία της κυρτής βελτιστοποίησης.

- Το κεφάλαιο 4 καταπιάνεται με τα νευρωνικά δίκτυα. Αναλύεται η δομή τους και η ιδέα που κινητοποίησε την επινόησή τους. Παρουσιάζονται συναρτήσεις ενεργοποίησης και κόστους, θεμελιώνεται η εκπαίδευση των νευρωνικών δικτύων, δηλαδή πως επιτυγχάνεται η εκμάθηση της "γνώσης" που κρύβεται στα δεδομένα. Τέλος, συζητούνται διάφοροι αλγόριθμοι βελτιστοποίησης της εκπαίδευσης.

- Στο κεφάλαιο 5 στρεφόμαστε σε μία ειδική κατηγορία νευρωνικών δικτύων, τα μορφολογικά νευρωνικά δίκτυα. Η ιδιαιτερότητά τους έγκειται στην επιλογή των συναρτήσεων ενεργοποίησης. Πραγματοποείται μία σύντομη εισαγωγή στα μορφολογικά μαθηματικά και στη Θεωρία Πλεγμάτων, γίνεται σύνδεση με την τροπική άλγεβρα. Παρουσιάζονται διάφορα μοντέλα μορφολογικών δικτύων με αυξανόμενη πολυπλοκότητα και αναλύεται η δυνατότητα προσέγγισης τους και τα σύνορα απόφασης που δημιουργούν. Παρουσιάζονται αρχιτεκτονικές που επιβάλλουν μονοτονικές συνθήκες στην έξοδο. Τέλος, εξετάζεται η εκπαίδευση μορφολογικών νευρωνικών δικτύων μέσω διεργασιών βελτιστοποίησης, η οποία συμπληρώνεται με σχετικά πειράματα.

- Στο κεφάλαιο 6 παρουσιάζονται τα πειραματικά αποτελέσματα των μοντέλων του προηγούμενου κεφαλαίου. Διακρίνουμε τις περιπτώσεις μονοτονικών και μη δικτύων. Στην πρώτη περίπτωση παρουσιάζεται ένας αλγόριθμος παραγωγής τεχνητού συνόλου δεδομένων και η πειραματική αξιολόγηση μονοτονικών δικτύων. Επιπρόσθετα, παρουσιάζεται μία αμιγώς τροπική επίλυση του προβλήματος παλινδρόμησης μονοτονικών δεδομένων. Στη μη μονοτονική περίπτωση, η πειραματική αξιολόγηση επικεντρώνεται σε δημοφιλή σύνολα δεδομένων, όπως `MNIST` [LeC98] και `FashionMNIST` [XRV17], και συγκρίνεται η απόδοση διάφορων αρχιτεκτονικών. Ιδιαίτερο βάρος δίνεται στη συμπίεση των δικτύων με τεχνικές pruning.

- Το κεφάλαιο 7 δρα ως επίλογος της παρούσας διπλωματικής εργασίας. Συνοψίζονται τα αποτελέσματα και παρατίθενται τα κύρια συμπεράσματα. Τέλος, προτείνονται πιθανές μελλοντικές επεκτάσεις των ιδεών της εργασίας.

## 1.3 Συμβολισμός

Προτού εντρυφήσουμε στην ουσία της εργασίας, αξίζει να διαλευκάνουμε το συμβολισμό που θα ακολουθήσει, ώστε ο αναγνώστης να είναι σε θέση να παρακολουθήσει το κείμενο και τη ροή της μαθηματικής σκέψης.

Αναφερόμαστε με τα σύμβολα $\mathbb{N}, \mathbb{Z}, \mathbb{R}$ στα συνολα των φυσικών, των ακεραίων και των πραγματικών αριθμών αντίστοιχα. Αναλυτικότερα, $\mathbb{N} = \{1, 2, 3, \dots\}$, $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$ και με το σύμβολο $\mathbb{R}$ αναφερόμαστε στην ευθεία των πραγματικών αριθμών $(-\infty, \infty)$. Στο πλαίσιο της τροπικής γεωμετρίας, συναντάμε τα επεκτεταμένα σύνολα των πραγματικών αριθμών $\mathbb{R}_{\max} = \mathbb{R} \cup \{-\infty\}$ και $\mathbb{R}_{\min} = \mathbb{R} \cup \{+\infty\}$.

Συμβολίζουμε:

- τα βαθμωτά μεγέθη με μικρά γράμματα, π.χ. $a \in \mathbb{R}$,
- τα διανύσματα ή τους μονοδιάστατους πίνακες (στήλη) με μικρά γράμματα σε bold γραμματοσειρά, π.χ. $\mathbf{a} \in \mathbb{R}^n$,
- τους πίνακες με κεφαλαία γράμματα σε σε bold γραμματοσειρά, π.χ. $\mathbf{A} \in \mathbb{R}^{m \times n}$

Για παράδειγμα,

$$a = 3.14 \qquad \mathbf{a} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 4 & 6 \end{bmatrix}$$

Τέλος, για τις νόρμες $\ell_p$ για $\mathbf{x} \in \mathbb{R}^n$, ισχύει:

$$\|\mathbf{x}\|_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

με $\|\mathbf{x}\|_\infty = \max\{|x_1|, |x_2|, \dots, |x_n|\}$.

# Κεφάλαιο 2

# Τροπική Γεωμετρία

Ο κλάδος της Τροπικής Γεωμετρίας μελετά συστήματα που μοντελοποιούνται με τη χρήση των πράξεων min και $+$. Πρόκειται, λοιπόν, για $(\min, +)$-γεωμετρία και αρχικά είχε λάβει αυτή την ονομασία. Ωστόσο, επικράτησε η "εξωτική" ονομασία προς τιμήν του Βραζιλιάνου μαθηματικού Imre Simon, ο οποίος ήταν από τους πρώτους μελετητές. Η $(\min, +)$-άλγεβρα συνδέεται άμεσα με τη $(\max, +)$ μέσω του ισομορφισμού $\phi(x) = -x$ [Mar15]. Στο πλαίσιο της εργασίας, λοιπόν, υιοθετούμε την ονομασία *τροπική γεωμετρία* για την περιγραφή και των δύο αλγεβρών. Η αναφορά σε μία από αυτές σημειώνεται ρητά ή προκύπτει από το περιεχόμενο.

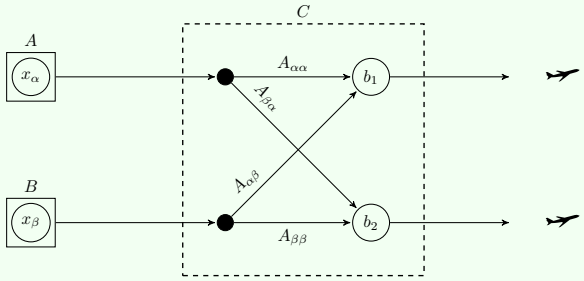
Αφετηρία για τη συζήτηση περί τροπικής γεωμετρίας είναι ένα πρόβλημα της καθημερινότητας:

**Example 2.0.1: Κίνητρο [But10]**

Εξετάζουμε το πρόβλημα χρονοδρομολόγησης που προκύπτει στην περίπτωση συνδεδεμένων πτήσεων (βλ. σχήμα 2.0.1). Έστω τρία αεροδρόμια $A, B, C$ και δύο πτήσεις $\alpha : A \to C$ και $\beta : B \to C$. Οι πτήσεις αναχωρούν τους χρόνους $x_\alpha, x_\beta$ αντίστοιχα, ενώ η διάρκειες τους είναι $d_\alpha, d_\beta$. Από το αεροδρόμιο $C$ υπάρχουν δύο πτήσεις, έστω $\gamma_\alpha, \gamma_\beta$. Έστω $A_{ij}$ ο χρόνος μεταφοράς που απαιτείται από την προσγείωση της πτήσης $j$ μέχρι την άφιξη στην πύλη για την αναχώρηση της πτήσης $\gamma_i$. Θεωρούμε ότι σε μεγάλα αεροδρόμια, ο χρόνος $A_{ij}$ δεν είναι αμελητέος. Αναζητούμε, λοιπόν, τους νωρίτερους δυνατούς χρόνους αναχώρησης για τις πτήσεις $\gamma_\alpha, \gamma_\beta$, τους οποίους συμβολίζουμε με $b_1, b_2$ αντίστοιχα. Τότε, το σύστημα έχει την ακόλουθη μορφή.

$$b_1 = \max\{x_\alpha + d_\alpha + A_{\alpha\alpha}, x_\beta + d_\beta + A_{\alpha\beta}\}$$
$$b_2 = \max\{x_\alpha + d_\alpha + A_{\beta\alpha}, x_\beta + d_\beta + A_{\beta\beta}\}$$

Θεωρώντας τον πίνακα $C_{ij} = A_{ij} + d_j$, το παραπάνω σύστημα γίνεται $\mathbf{b} = \mathbf{C} \boxplus \mathbf{x}$ και αποτελεί γραμμικό σύστημα στο πλαίσιο της max-plus άλγεβρας, όπου με $\boxplus$ συμβολίζουμε τον τροπικό πολλαπλασιασμό πινάκων υπό τη max-plus έννοια.




Σχήμα 2.0.1: Χρονοδρομολόγηση συνδεδεμένων πτήσεων

## 2.1 Βασικές έννοιες

Προτού εντρυφήσουμε στην τροπική άλγεβρα, αξίζει να δοθεί σημασία στο γραμμικό ανάλογό της, τη *Γραμμική Άλγεβρα* και τη σύνδεση μεταξύ των δύο. Η κύρια διαφορά, τουλάχιστον σε άλγεβρικό επίπεδο, είναι η αντικατάσταση των πράξεων $(+, \times)$ της γραμμικής άλγεβρας με το ζεύγος $(\min, +)$ στην τροπική γεωμετρία υπό τη min-plus έννοια και με $(\max, +)$ για τη max-plus έννοια. Επιπρόσθετα, οι δύο περιπτώσεις χαρακτηρίζονται από διαφορετικό ουδέτερο στοιχείο $\epsilon$. Για την περίπτωση min-plus έχουμε $\epsilon_{\min} = +\infty$ καθώς $\forall a \in \mathbb{R}$ ισχύει ότι $\min\{a, \epsilon_{\min}\} = a$. Ομοίως, $\epsilon_{\max} = -\infty$. Στο πλαίσιο της τροπικής γεωμετρίας, λοιπόν, επεκτείνουμε το σύνολο των πραγματικών αριθμών $\mathbb{R}$ ώστε να συμπεριλαμβάνει το ουδέτερο στοιχείο της εκάστοτε θεώρησης. Συνεπώς, $\mathbb{R}_{\min} = \mathbb{R} \cup \{\epsilon_{\min}\} = \mathbb{R} \cup \{+\infty\}$ και $\mathbb{R}_{\max} = \mathbb{R} \cup \{\epsilon_{\max}\} = \mathbb{R} \cup \{-\infty\}$. Τότε, οι πράξεις για δύο στοιχεία $x, y$ ορίζονται ως:

- $x \wedge y = \min\{x, y\}$ για $x, y \in \mathbb{R}_{\min}$ ως την πράξη της τροπικής πρόσθεσης (min-plus έννοια)
- $x \vee y = \max\{x, y\}$ για $x, y \in \mathbb{R}_{\max}$ ως την πράξη της τροπικής πρόσθεσης (max-plus έννοια)
- $x + y$ ως την πράξη του τροπικού πολλαπλασιασμού για $x, y \in \mathbb{R}_{\min}$ *ή* $x, y \in \mathbb{R}_{\max}$

Επεκτείνουμε τις έννοιες αυτές σε πολυδιάστατους χώρους. Τότε, οι πράξεις της τροπικής άλγεβρας για $n$-διάστατα διανύσματα ορίζονται ως:

- $\mathbf{a} \wedge \mathbf{b} = \begin{bmatrix} a_1 \wedge b_1, & \dots & , a_n \wedge b_n \end{bmatrix}^\top$ για $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n_{\min}$
- $\mathbf{a} \vee \mathbf{b} = \begin{bmatrix} a_1 \vee b_1, & \dots & , a_n \vee b_n \end{bmatrix}^\top$ για $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n_{\max}$
- $\mathbf{a} + \mathbf{b} = \begin{bmatrix} a_1 + b_1, & \dots & , a_n + b_n \end{bmatrix}^\top$ για $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n_{\min}$ *ή* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n_{\max}$

Με άλλα λόγια, οι πράξεις γίνονται ανά στοιχείο (element-wise). Οι πράξεις πινάκων στην τροπική άλγεβρα ορίζονται ως:

- $(\mathbf{A} \wedge \mathbf{B})_{ij} = \min(a_{ij}, b_{ij})$ για $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}_{\min}$
- $(\mathbf{A} \vee \mathbf{B})_{ij} = \max(a_{ij}, b_{ij})$ για $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}_{\max}$
- $(\mathbf{A} \boxplus' \mathbf{B})_{ij} = \bigwedge_{q=1}^{k} a_{iq} + b_{qj}$ για $\mathbf{A} \in \mathbb{R}^{m \times k}_{\min}, \mathbf{B} \in \mathbb{R}^{k \times n}_{\min}$
- $(\mathbf{A} \boxplus \mathbf{B})_{ij} = \bigvee_{q=1}^{k} a_{iq} + b_{qj}$ για $\mathbf{A} \in \mathbb{R}^{m \times k}_{\max}, \mathbf{B} \in \mathbb{R}^{k \times n}_{\max}$

**Example 2.1.1: παράδειγμα σε πίνακα**

Παρουσιάζουμε ορισμένα παραδείγματα για την καλύτερη κατανόηση των παραπάνω εννοιών: Έστω $A = \begin{bmatrix} 3 & 3 \\ 0 & 7 \end{bmatrix}, B = \begin{bmatrix} 4 & 1 \\ 5 & 2 \end{bmatrix}$. Τότε, εκτελώντας τις τροπικές εκδόσεις της πρόσθεσης και του πολλαπλασιασμού βρίσκουμε ότι:

$$
\begin{aligned}
C = A \wedge B &= \begin{bmatrix} 3 & 3 \\ 0 & 7 \end{bmatrix} \wedge \begin{bmatrix} 4 & 1 \\ 5 & 2 \end{bmatrix} = \begin{bmatrix} 3 \wedge 4 & 3 \wedge 1 \\ 0 \wedge 5 & 7 \wedge 2 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} \\
C = A \vee B &= \begin{bmatrix} 3 & 3 \\ 0 & 7 \end{bmatrix} \vee \begin{bmatrix} 4 & 1 \\ 5 & 2 \end{bmatrix} = \begin{bmatrix} 3 \vee 4 & 3 \vee 1 \\ 0 \vee 5 & 7 \vee 2 \end{bmatrix} = \begin{bmatrix} 4 & 3 \\ 5 & 7 \end{bmatrix} \\
C = A \boxplus' B &= \begin{bmatrix} 3 & 3 \\ 0 & 7 \end{bmatrix} \boxplus' \begin{bmatrix} 4 & 1 \\ 5 & 2 \end{bmatrix} = \begin{bmatrix} (3+4) \wedge (3+5) & (3+1) \wedge (3+2) \\ (0+4) \wedge (7+5) & (0+1) \wedge (7+2) \end{bmatrix} = \begin{bmatrix} 7 & 4 \\ 4 & 1 \end{bmatrix} \\
C = A \boxplus B &= \begin{bmatrix} 3 & 3 \\ 0 & 7 \end{bmatrix} \boxplus \begin{bmatrix} 4 & 1 \\ 5 & 2 \end{bmatrix} = \begin{bmatrix} (3+4) \vee (3+5) & (3+1) \vee (3+2) \\ (0+4) \vee (7+5) & (0+1) \vee (7+2) \end{bmatrix} = \begin{bmatrix} 8 & 5 \\ 12 & 9 \end{bmatrix}
\end{aligned}
$$

Αξίζει να εξετάσουμε την τροπική άλγεβρα από τη σκοπιά της Θεωρίας Ομάδων. Παρατηρούμε ότι η πράξη $\wedge$ είναι ταυτοδύναμη (idempotent) και ότι αμφότερες είναι αντιμεταθετικές (commutative). Επιπλέον, για την πράξη της τροπικής πρόσθεσης min δεν ορίζεται αντίστροφο στοιχείο[Pin98]. Επομένως, το $(\mathbb{R}_{\min}, \wedge, +)$ είναι ένας ταυτοδύναμος αντιμεταθετικός ημιδακτύλιος (idempotent commutative semiring) [GN04; GM08]. Η τροπική διαίρεση ορίζεται ως η κλασσική αφαίρεση. Παρατηρούμε ότι ο τροπικός ημιδακτύλιος $(\mathbb{R}_{\min}, \wedge, +)$ είναι ισομορφικός με τον $(\max, +)$-ημιδακτύλιο $(\mathbb{R}_{\max}, \vee, +)$ μέσω του ισομορφισμού $\phi(x) = -x$.

Από την παραπάνω συζήτηση περί ουδέτερου στοιχείου καταλήγουμε στην έννοια του μοναδιαίου πίνακα, ο οποίος χαρακτηρίζεται από την ιδιότητα $\mathbf{I} \boxplus' \mathbf{A} = \mathbf{A} \boxplus' \mathbf{I} = \mathbf{A}$ στη min-plus περίπτωση και $\mathbf{I} \boxplus \mathbf{A} = \mathbf{A} \boxplus \mathbf{I} = \mathbf{A}$ στη max-plus. Κατά συνέπεια, ο μοναδιαίος πίνακας έχει ιδιαίτερη μορφή με 0 στα στοιχεία της διαγωνίου και $\epsilon$ στα υπόλοιπα:

$$
\mathbf{I} = \begin{bmatrix} 0 & \epsilon & \dots & \epsilon \\ \epsilon & 0 & \dots & \vdots \\ \vdots & & \ddots & \epsilon \\ \epsilon & \epsilon & \dots & 0 \end{bmatrix} \tag{2.1.1}
$$

Παρατηρούμε ότι έχει την ίδια δομή με το μοναδιαίο πίνακα της γραμμικής άλγεβρας και η διαφορά έγκειται στο γεγονός ότι τα ουδέτερα στοιχεία του πολλαπλασιασμού και της πρόσθεσης είναι διαφορετικά.

## 2.2 Επίλυση Τροπικά Γραμμικών Συστημάτων

Στρέφουμε την προσοχή μας σε ενδιαφέροντα και σημαντικά προβλήματα που έχουν μελετηθεί και επιλύονται στο πλαίσιο των τροπικών μαθηματικών. Λόγω της στενής σχέσης μεταξύ $(\min, +)$ και $(\max, +)$ μαθηματικών, μεταπηδούμε από τη μία άλγεβρα στην άλλη. Υπενθυμίζουμε ότι συμβολίζουμε ως $\boxplus$ και $\boxplus'$ τον πολλαπλασισμό πινάκων σε $(\max, +)$ και $(\min, +)$ άλγεβρα, αντίστοιχα.

Ο κλάδος της γραμμικής άλγεβρας αναπτύχθηκε (εν μέρει) αποσκοπώντας την επίλυση συστημάτων γραμμικών εξισώσεων. Σε συμπαγή μορφή πίνακα:

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b} \tag{2.2.1}$$

με $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{x} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^m$. Το αντίστοιχο πρόβλημα στη $(\max, +)$ άλγεβρα λαμβάνει τη μορφή:

$$\mathbf{A} \boxplus \mathbf{x} = \mathbf{b} \text{ με } \mathbf{A} \in \mathbb{R}_{\max}^{m \times n}, \mathbf{x} \in \mathbb{R}_{\max}^n, \mathbf{b} \in \mathbb{R}_{\max}^m \tag{2.2.2}$$

Το πρόβλημα αυτό μελετήθηκε από τον Cuninghame-Green στο έργο του *Minimax Algebra*[Cun79].

**Theorem 2.2.1: αντίστροφος Cuninghame-Green**

Το πρόβλημα (2.2.2) έχει λύση $\mathbf{x}^*$ μόνο αν το διάνυσμα $\mathbf{x}^*$:

$$\mathbf{x}^* = \mathbf{A}^* \boxplus' \mathbf{b}, \quad \mathbf{A}^* \triangleq -\mathbf{A}^\top \tag{2.2.3}$$

αποτελεί λύση του. Ο πίνακας $\mathbf{A}^*$ λέγεται **αντίστροφος Cuninghame-Green** και ισχύει:

$$\mathbf{A} \boxplus \mathbf{x} \leq \mathbf{A} \boxplus \mathbf{x}^* \leq \mathbf{b}$$

Πολλά προβλήματα μπορούν να μοντελοποιηθούν με αυτό τον τρόπο. Ιστορικά, ένας από τους πρώτους κλάδους όπου εφαρμόστηκαν ιδέες τροπικής άλγεβρας είναι η επιχειρησιακή έρευνα (operations research). Η μαθηματική μοντελοποίηση που προσφέρει η τροπική άλγεβρα χρησιμοποιήθηκε και στην επίλυση προβλημάτων machine-scheduling, shortest-path problems κ.α.

**Example 2.2.2: scheduling**

Ο Cuninghame-Green, ένας από τους πρώτους μελετητές της τροπικής άλγεβρας και συγγραφέας του Minimax Algebra [Cun79], μελέτησε το πρόβλημα *multi-machine interactive production process* (MMIPP). Έστω ότι έχουμε τα προϊόντα $P_1, P_2, \cdots, P_m$ και $n$ μηχανές, καθεμία από τις οποίες συνδράμει μερικά στη δημιουργία ενός τελικού προϊόντος. Έστω, λοιπόν, $a_{ij}$ ο χρόνος που απαιτείται στη μηχανή $j$ για την παραγωγή του προϊόντος $P_i$. Σε περίπτωση που η μηχανή $j$ δε χρησιμοποιείται για την παραγωγή του το προϊόντος $P_i$, θέτουμε $a_{ij} = -\infty$. Επιπλέον, έστω $x_j$ ο χρόνος εκκίνησης της μηχανής $j$. Τότε, όλα τα μερικά προϊόντα που απαιτούνται για το $P_i$ θα είναι έτοιμα τη χρονική στιγμή $b_i$ όπου

$$b_i = \bigvee_k A_{ik} + x_k$$

Σε συμπαγή μορφή

$$A \boxplus x = b \tag{2.2.4}$$

Επιπρόσθετα, τα προβλήματα ιδιοτιμών αντιστοιχούν σε ακρογωνιαίους λίθους στη θεωρία της τροπικής γεωμετρίας. Αναλυτικότερα,

$$\mathbf{A} \boxplus' \mathbf{x} = \lambda \boxplus' \mathbf{x} \tag{2.2.5}$$

$$\mathbf{A} \boxplus' \mathbf{x} \geq \lambda \boxplus' \mathbf{x} \tag{2.2.6}$$

Οι λύσεις τους παρέχονται από δύο θεμελιώδεις πίνακες για την τροπική γεωμετρία.

**Definition 2.2.3: $\Gamma(A)$ και $\Delta(A)$**

Έστω $\mathbf{A} \in \mathbb{R}_{\min}^{n\times n}$. Ορίζουμε τις άπειρες σειρές

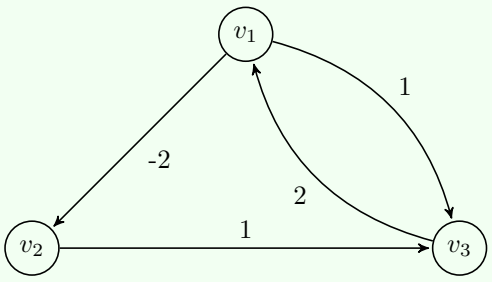
$$\Gamma(\mathbf{A}) = \mathbf{A} \wedge \mathbf{A}^2 \wedge \cdots \wedge \mathbf{A}^n \wedge \cdots = \bigwedge_{k=1}^{\infty} \mathbf{A}^k \tag{2.2.7}$$

$$\Delta(\mathbf{A}) = \mathbf{I} \wedge \mathbf{A} \wedge \mathbf{A}^2 \wedge \cdots \wedge \mathbf{A}^n \wedge \cdots = \bigwedge_{k=0}^{\infty} \mathbf{A}^k \tag{2.2.8}$$

ως ασθενή και ισχυρή μεταβατική κλειστότητα, αντίστοιχα. Οι πίνακες αυτοί παρέχουν λύσεις στα προβλήματα (2.2.5), (2.2.6) αντίστοιχα. Οι λύσεις των προβλημάτων ονομάζονται ιδιοδιανύσματα (eigenvectors) και υπό-ιδιοδιανύσματα (subeigenvectors) όταν δεν ισούνται με $\epsilon$.

Ένα χαρακτηριστικό παράδειγμα προέρχεται από το πρόβλημα ελαχίστων μονοπατιών. Εξετάζουμε την περίπτωση που δεν υπάρχουν αρνητικοί κύκλοι.

**Example 2.2.4: Shortest-Path Problem**




Σχήμα 2.2.1: Γράφος για το min-distance πρόβλημα

Ο παραπάνω γράφος αντιστοιχεί στον πίνακα βαρών:

$$\mathbf{A} = \begin{bmatrix} 0 & -2 & 1 \\ \epsilon & 0 & 1 \\ 2 & \epsilon & 0 \end{bmatrix}$$

Παρατηρούμε ότι δεν έχουμε ανακυκλώσεις (self-loops) και τα στοιχεία του πίνακα που δεν αντιστοιχούν σε κάποια ακμή σημειώνονται με το ουδέτερο στοιχείο $\epsilon = \infty$. Στο απλό αυτό παράδειγμα, είναι εμφανές ότι ο γράφος *δεν* έχει αρνητικούς κύκλους και ότι το μέγιστο μήκος μονοπατιού είναι 2. Επομένως, τα ελάχιστα μονοπάτια χαρακτηρίζονται από την εξίσωση:

$$\Gamma(\mathbf{A}) = \mathbf{A} \boxplus' \mathbf{A}^2$$

Η παραπάνω εξίσωση σκιαγραφεί τη διαπίστωση ότι τα βέλτιστα μονοπάτια έχουν είτε μήκος 1 είτε μήκος 2. Υπολογίζουμε τα επιμέρους βήματα:

$$\mathbf{A}^2 = \begin{bmatrix} 0 & -2 & 1 \\ \epsilon & 0 & 1 \\ 2 & \epsilon & 0 \end{bmatrix} \boxplus' \begin{bmatrix} 0 & -2 & 1 \\ \epsilon & 0 & 1 \\ 2 & \epsilon & 0 \end{bmatrix} = \begin{bmatrix} 0 & -2 & -1 \\ 3 & 0 & 1 \\ 2 & 0 & 0 \end{bmatrix}$$

Παρατηρώντας το γράφο στο σχήμα 2.2.1, συμπεραίνουμε ότι δεν έχει κύκλους αρνητικού κόστους. Άρα, τα περισσότερα βήματα μπορούν μόνο να προσθέσουν κόστος και δεν χρειάζεται να υπολογίσουμε

μεγαλύτερη δύναμη του $\mathbf{A}$. Συνεπώς, ο πίνακας αποστάσεων είναι ίσος με:

$$\begin{aligned}\Gamma(\mathbf{A}) &= \mathbf{A} \wedge \mathbf{A}^2 \\ &= \begin{bmatrix} 0 & -2 & 1 \\ \epsilon & 0 & 1 \\ 2 & \epsilon & 0 \end{bmatrix} \wedge \begin{bmatrix} 0 & -2 & -1 \\ 3 & 0 & 1 \\ 2 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & -2 & -1 \\ 3 & 0 & 1 \\ 2 & 0 & 0 \end{bmatrix}\end{aligned}$$

Ο τρόπος επίλυσης του προβλήματος ελαχίστων μονοπατιών μέσω της εύρεσης της ασθενούς μεταβατικής κλειστότητας $\Gamma(\mathbf{A})$ ταυτίζεται με το γνωστό αλγόριθμο Floyd-Warshall. Ο πίνακας $\mathbf{A}^k$ αντιστοιχεί σε ελάχιστα μονοπάτια μήκους $L = k$ και ο πίνακας $\Gamma(\mathbf{A})$ υπολογίζει τα ελάχιστα μονοπάτια εξετάζοντας τα διάφορα μήκη.

## 2.3 Στοιχεία Γεωμετρίας

### 2.3.1 Τροπικές Καμπύλες

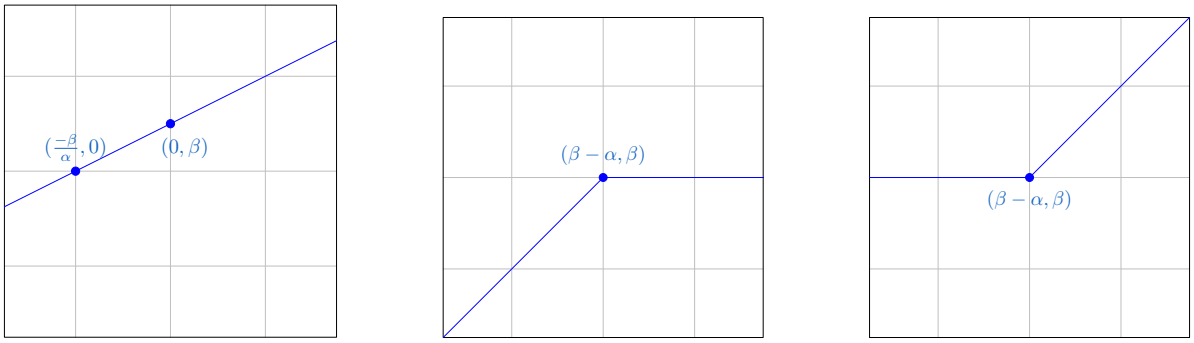
Αρχίζουμε την ανάλυση μας από απλές συναρτήσεις ή απεικονίσεις μίας μεταβλητής, όπως ευθείες και πολυώνυμα στο τροπικό πλαίσιο. Η ευθεία χαρακτηρίζεται από δύο παραμέτρους, $\alpha$ και $\beta$ που συνδέουν την τετμημένη $x$ και την τεταγμένη $y$ με την ακόλουθη σχέση:

$$y = \alpha \cdot x + \beta \tag{2.3.1}$$

Η γραφική παράσταση της ευθείας είναι καθολικά γνωστή. Η τροπική γεωμετρία μεταβάλλει τους τελεστές της πρόσθεσης και του πολλαπλασιασμού. Ενώ ο ορισμός της ευθείας παραμένει αναλλοίωτος, το αποτέλεσμα είναι διαφορετικό:

$$y = (\alpha + x) \wedge \beta = \min(\alpha + x, \beta) \tag{2.3.2}$$

Στο σημείο $x = x_0 = \beta - \alpha$ οι δύο όροι της τροπικής ευθείας είναι ίσοι. Αριστερά του σημείου, η ελεύθερη μεταβλητή $x$ επικρατεί και καθορίζει την καμπύλη, ωστόσο δεξιά του σημείου $x_0$ επικρατεί ο σταθερός όρος. Κατά συνέπεια, εντοπίζεται ένα "σπάσιμο" στο $x_0$ όπου αλλάζει η κλίση της τροπικής ευθείας και η συνάρτηση παύει να είναι παραγωγίσιμη. Γραφικά, έχουμε:




(a) $y = \alpha \cdot x + \beta$ (b) $y = \min(\alpha + x, \beta)$ (c) $y = \max(\alpha + x, \beta)$

Σχήμα 2.3.1: Ευθείες στις διάφορες άλγεβρες

Τα σημεία αυτά είναι χαρακτηριστικά της τροπικής γεωμετρίας. Το πλαίσιο των τροπικών μαθηματικών μελετάει εκ φύσεως **τμηματικά γραμμικές** συναρτήσεις (piecewise linear ή PWL). Συνεχίζουμε την ανάλυση με συναρτήσεις δευτέρου βαθμού, τη γνωστή παραβολή. Στα "κλασικά" μαθηματικά, η παραβολή χαρακτηρίζεται

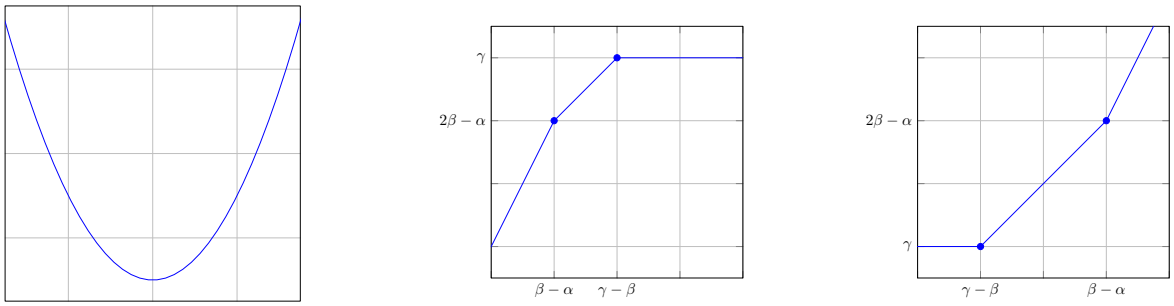
από τη σχέση:

$$y = \alpha \cdot x^2 + \beta \cdot x + \gamma \tag{2.3.3}$$

Τροπικοποιώντας την παραβολή, δηλαδή εναλλάσσοντας τους τελεστές, έχουμε:

$$y = (\alpha + 2 \cdot x) \wedge (\beta + x) \wedge \gamma = \min(\alpha + 2 \cdot x, \beta + x, \gamma) \tag{2.3.4}$$

Γραφικά έχουμε:




(a) $y = \alpha \cdot x^2 + \beta \cdot x + \gamma$ (b) $y = \min(\alpha + 2 \cdot x, \beta + x, \gamma)$ (c) $y = \max(\alpha + 2 \cdot x, \beta + x, \gamma)$

Σχήμα 2.3.2: Παραβολές στις διάφορες άλγεβρες

### 2.3.2 Τροπικά Πολυώνυμα

Από τις γραφικές παραστάσεις 2.3.1 και 2.3.2, παρατηρούμε ότι ο βαθμός της συνάρτησης καθορίζει και το πλήθος των γραμμικών τμημάτων. Επιπλέον, είναι φανερό ότι στη $(\min, +)$ άλγεβρα οι συναρτήσεις είναι κοίλες, ενώ στη $(\max, +)$ οι συναρτήσεις υπό μελέτη είναι κυρτές. Για λόγους πληρότητας, παραθέτουμε τον ορισμό της κυρτότητας για συναρτήσεις:

**Definition 2.3.1: Κυρτή Συνάρτηση**

Μία συνάρτηση $f : \mathbb{R}^n \to \mathbb{R}$ είναι κυρτή αν $\mathbf{dom}\, f$ είναι κυρτό σύνολο και $\forall x, y \in \mathbf{dom}\, f$ και $\theta \in [0, 1]$ ισχύει:

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta) f(y) \tag{2.3.5}$$

Αυτή η ιδιότητα των κυρτών συναρτήσεων παρουσιάζεται γραφικά στο σχήμα 2.3.3.

Στα προηγούμενα παραδείγματα μελετήθηκαν συναρτήσεις μίας μεταβλητής. Η επέκταση των εννοιών σε περισσότερες διαστάσεις είναι άμεση και διαισθητικά απλή. Έστω $x_1, x_2, \cdots, x_n \in \mathbb{R}_{\min}$. Τότε, το τροπικό μονώνυμο, ή απλώς μονώνυμο, είναι οποιοδήποτε γινόμενο, υπό την τροπική έννοια, αυτών των μεταβλητών. Σημειώνεται ότι επιτρέπεται η επανάληψη όρων. Με συμπαγή τρόπο, μπορούμε να γράψουμε ένα τροπικό μονώνυμο ως $\mathbf{k}^\top \mathbf{x}$ όπου $\mathbf{x}, \mathbf{k} \in \mathbb{R}^n_{\min}$. Η κατασκευή ενός πολυωνύμου έγκειται στο συνδυασμό πολλών μονωνύμων. Αυτός ο συνδυασμός αντιστοιχεί στην πράξη της πρόσθεσης στην κλασική ανάλυση. Ωστόσο, στα τροπικά μαθηματικά, το πολυώνυμο είναι το minimum πολλών μονωνύμων.

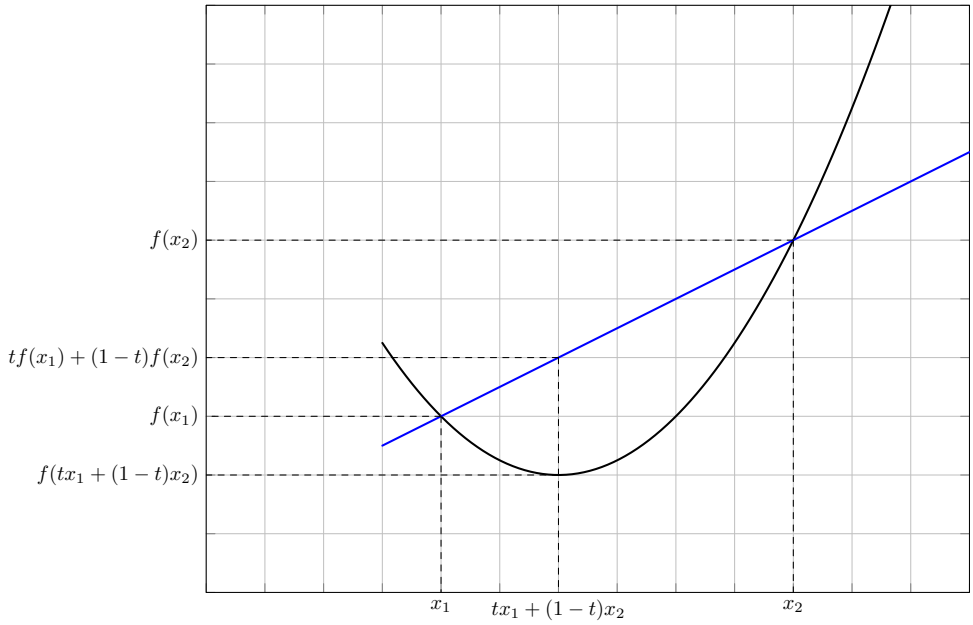
**Definition 2.3.2: Τροπικό πολυώνυμο**

Τροπικό πολυώνυμο (υπό τη min-plus έννοια) $p(x_1, x_2, \cdots, x_n) : \mathbb{R}^n_{\min} \to \mathbb{R}_{\min}$ είναι ο πεπερασμένος γραμμικός συνδυασμός τροπικών μονωνύμων:

$$p(\mathbf{x}) = p(x_1, x_2, \cdots, x_n) = \min_i \{c_{i1}x_1 + \cdots + c_{in}x_n\} = \min_i \{\mathbf{c}_i^\top \mathbf{x}\} \tag{2.3.6}$$

όπου $\mathbf{c}_i \in \mathbb{R}^n$.

Ο ορισμός του τροπικού πολυωνύμου υπό την max-plus έννοια είναι ευθής και αντιστοιχεί στην εναλλαγή του τελεστή min με τον τελεστή max. Στο σχήμα 2.3.4 βλέπουμε παραδείγματα πολυωνύμων μίας μεταβλητής στο πλαίσιο των αλγεβρών $(\max, +)$ και $(\min, +)$ αντίστοιχα.



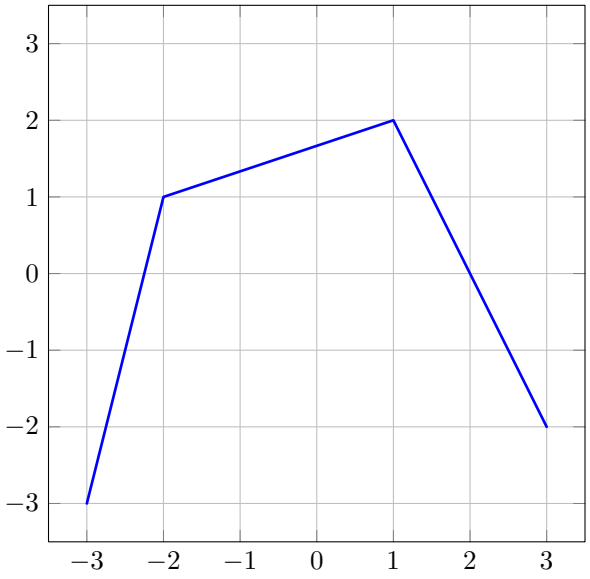

Σχήμα 2.3.3: Ιδιότητα κυρτών συναρτήσεων

Ένας ιδιαίτερα σημαντικός πόλος της κλασικής αλγεβρικής γεωμετρίας είναι η εύρεση των ριζών ενός πολυωνύμου ή ενός συστήματος πολυωνυμικών εξισώσεων. Τα σημεία αυτά ορίζουν μία υπερεπιφάνεια. Η έννοια αυτή έχει αντίστοιχο στην τροπική γεωμετρία:

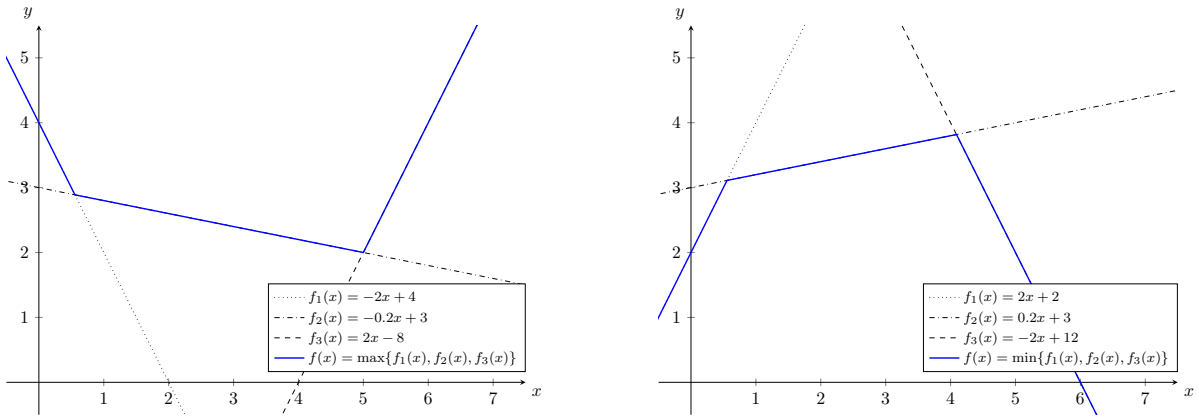
**Definition 2.3.3: Tropical Hypersurface**

Έστω $p(x) : \mathbb{R}^n \to \mathbb{R}$ ένα τροπικό πολυώνυμο $n$ μεταβλητών. Ορίζουμε ως τροπική υπερεπιφάνεια (tropical Hypersurface) $\mathcal{T}(p)$ ως το σύνολο σημείων:

$$\mathcal{T}(p) := \{\mathbf{x} \in \mathbb{R}^n : \text{Το minimum του p επιτυγχάνεται τουλάχιστον 2 φορές}\} \tag{2.3.7}$$



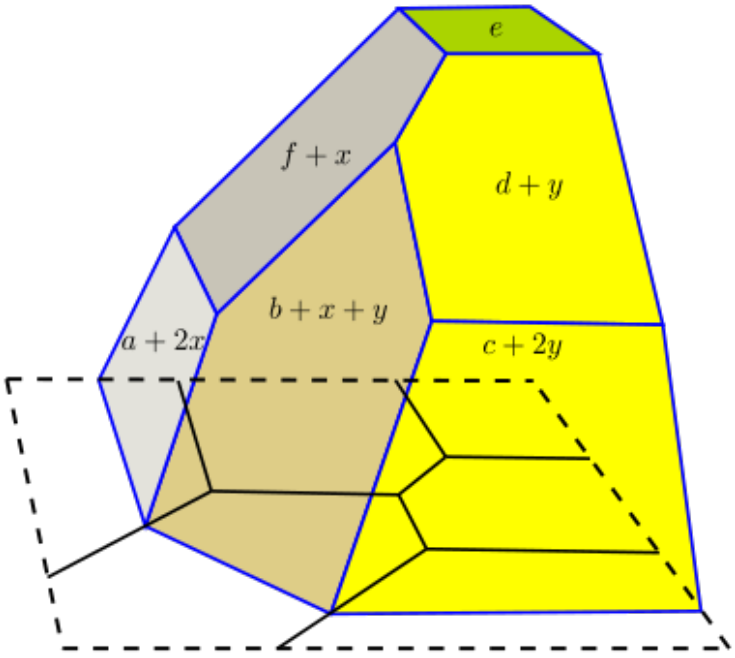

Σχήμα 2.3.5: $\mathcal{T}(p) = \{-2, 1\}$




(a) κατασκευή max-plus πολυωνύμου (b) κατασκευή min-plus πολυωνύμου

Σχήμα 2.3.4: Κατασκευή πολυωνύμων σε max- και min-plus άλγεβρες

Τα σημεία αυτά δεν είναι παραγωγίσιμα και αποτελούν τις ρίζες του τροπικού πολυωνύνου. Στο σχήμα 2.3.2b παρατηρούμε ότι: $\mathcal{T}(p) = \{\beta - \alpha, \gamma - \beta\}$. Η τροπική υπερεπιφάνεια κατέχει έναν ιδιαίτερα σημαντικό ρόλο, καθώς διαχωρίζει το πεδίο ορισμού σε υποχωρία όπου κυριαρχούν διαφορετικά μονώνυμα. Ας εξετάσουμε το παράδειγμα του σχήματος 3. Η τροπική υπερεπιφάνεια είναι $\mathcal{T}(p) = \{-2, 1\}$ και, συνεπώς, το πεδίο ορισμού $\text{dom}(p) = \{-\infty, +\infty\}$ διαχωρίζεται στα υποχωρία $\{(-\infty, -2), (-2, 1), (1, +\infty)\}$. Στο καθένα από αυτά, κυριαρχεί διαφορετικό μονώνυμο. Οι έννοιες αυτές επεκτείνονται και σε περισσότερες διαστάσεις.




Σχήμα 2.3.6: Ο τροπικός γράφος και η τροπική καμπύλη που ορίζονται από ένα τροπικό πολυώνυμο δεύτερης τάξης στο $\mathbb{R}^2_{\text{min}}$. Με διαφορρετικό χρώμα φαίνονατι οι τοπικές συναρτήσεις που αποτελούν τον κυρίραχο όρο στην περιοχή. Η αντίστοιχη περιοχή φαίνεται στο επίπεδο εντός των διακεκομμένων γραμμών. (πηγή: [MS15])

Στο παραπάνω σχήμα (2.3.6) με χρώματα συμβολίζεται το τροπικό πολυώνυμο δύο μεταβλητών $x, y$ και στο $xy$-επίπεδο φαίνεται η προβολή του, η οποία πρόκειται για την τροπική υπερεπιφάνεια. Στην περίπτωση του $\mathbb{R}^2$ είναι πιο προφανής ο διαμερισμός του πεδίου ορισμού σε περιοχές όπου κυριαρχούν τα διάφορα μονώνυμα που

απαρτίζουν το πολυώνυμο. Οι Tarela and Martinez μελετούν piecewise linear συναρτήσεις από τη σκοπιά της Θεωρίας Πλεγμάτων (Lattice) και χαρακτηρίζουν τις διάφορες περιοχές όπου κυριαρχεί διαφορετικός όρος της συνάρτησης ως περιοχές προβολής $\Omega_i$ [TM99]. Αν και η οικογένεια τμηματικά γραμμικών συναρτήσεων είναι σαφώς πιο ευρεία από τον αυστηρό ορισμό κυρτών (κοίλων) συναρτήσεων στο πλαίσιο της max-plus (min-plus) άλγεβρας, η έννοια των περιοχών προβολής βρίσκει αντίκρισμα στα τροπικά μαθηματικά.

Απο το σχήμα του τροπικού γράφου 2.3.6 και της αντίστοιχης τροπικής υπερεπιφάνειας απορρέει άμεσα η έννοια του τροπικού προβολικού χώρου:

$$\mathbb{TP}^{n-1} = \mathbb{R}^n/(1,\dots,1)\mathbb{R} \tag{2.3.8}$$

Αξίζει να σημειωθεί ότι κάθε τροπικά κυρτό υποσύνολο $S \subset \mathbb{R}^n$ είναι κλειστό υπό τροπικό πολλαπλασιασμό $\mathbb{R} + S \subseteq S$. Αυτό σημαίνει ότι αν $x \in S$ τότε $x + \lambda(1,\dots,1) \in S$ για κάθε $\lambda \in S$.

### 2.3.3 Τροπικά Πολύτοπα

Τέλος, μία ανασκόπηση της τροπικής γεωμετρίας είναι ελλιπής χωρίς την αναφορά στο πολύτοπο Newton, το οποίο αντιστοιχίζει τους συντελεστές ενός τροπικού πολυωνύμου σε σημεία στο χώρο, επιτρέποντας την απαλοιφή όρων.

**Definition 2.3.4: Newton Πολύτοπο**

Έστω $p : \mathbb{R}^n \to \mathbb{R}$ ένα τροπικό πολυώνυμο $p(\mathbf{x}) = \max_i\{c_{i1}x_1 + \cdots + c_{in}x_n\} = \max_i\{\mathbf{c}_i^\top \mathbf{x}\}$. Τότε το πολύτοπο Newton για το πολυώνυμο $p$ είναι

$$\text{Newt}(p) = \text{conv}\{\mathbf{c}_i : i \in I\} = \text{conv}\{(c_{i1},\dots,c_{in}) : i \in I\}$$

Πρόκειται, δηλαδή, για το convex hull των τροπικών διανυσματικών εκθετών.

**Example 2.3.5: Newton πολύτοπο**

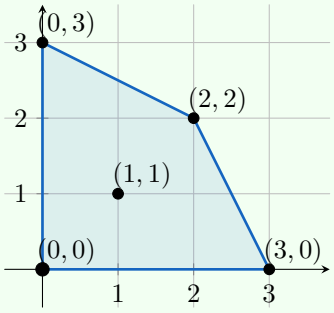
Έστω το ακόλουθο τροπικό πολυώνυμο

$$p(x) = \max\{0, x_1 + x_2, 2x_1 + 2x_2, 3x_1, 3x_2\} \tag{2.3.9}$$

Το κάθε μονώνυμο ορίζει ένα σημείο στο $\mathbb{R}^d$. Εδώ $d = 2$. Αντιστοιχίζουμε τα μονώνυμα σε σημεία του $\mathbb{R}^d$:

$$\begin{aligned} 0 &\to (0,0) \\ x_1 + x_2 &\to (1,1) \\ 2x_1 + 2x_2 &\to (2,2) \\ 3x_1 &\to (3,0) \\ 3x_2 &\to (0,3) \end{aligned}$$

Επομένως, η κυρτή θήκη όλω αυτών των σημείων γραφικά είναι




Σχήμα 2.3.7: Newton πολύτοπο για το τροπικό πολυώνυμο (2.3.9)

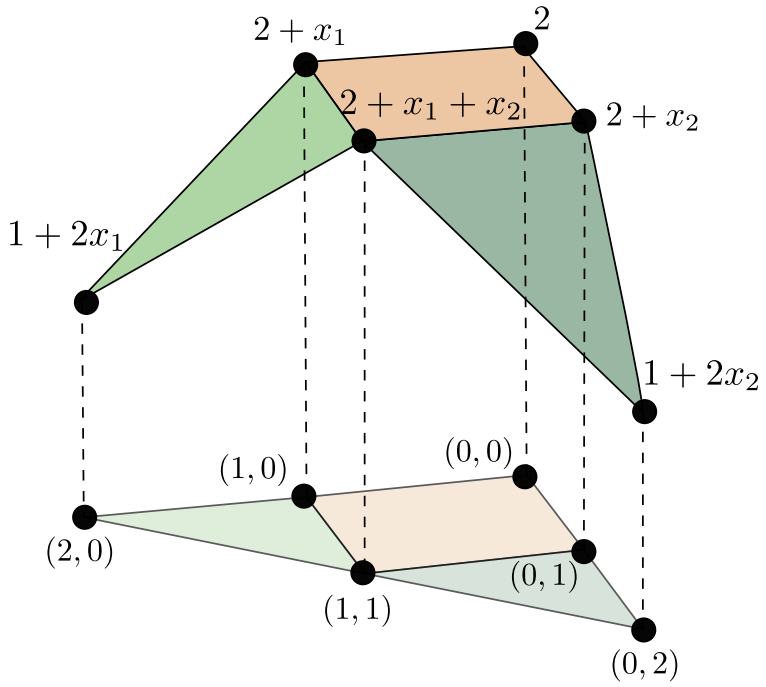
Ερμηνεύοντας γεωμετρικά το παραπάνω σχήμα, βρίσκουμε ότι ο όρος $x_1x_2$ που αντιστοιχεί στο σημείο $(1,1)$ μπορεί να παραλειφθεί, καθώς σε καμία περίπτωση δεν είναι ο όρος που επικρατεί στο πολυώνυμο. Αυτό το συμπέρασμα θυμίζει την ιδιότητα του γραμμικού προγραμματισμού ότι η αντικειμενική συνάρτηση λαμβάνει την ελάχιστη (ή μέγιστη) τιμή της στα άκρα του πολύτοπου (βλ. Simplex). Επιπλέον, μέσα από αυτό το παράδειγμα διαφαίνεται η δύναμη και η διαίσθηση που προσδίδει στην αλγεβρική μοντελοποίηση η τροπική **γεωμετρία**.

Από το παραπάνω παράδειγμα παρατηρούμε τη σημασία που προσφέρει ένα μαθηματικό εργαλείο σαν το πολύτοπο Newton, καθώς και ένα σημαντικό του μειονέκτημα. Αναλυτικότερα, τα μονώνυμα δεν περιέχουν σταθερό όρο, γεγονός που περιορίζει την εκφραστικότητά τους. Ωστόσο, μία πολύ απλή λύση είναι η επέκταση του διανύσματος ελεύθερων μεταβλητών $\mathbf{x} = [x_1, x_2, \ldots, x_d]^\top \in \mathbb{R}^d$ ώστε να περιέχει το σταθερό όρο: $\mathbf{x} \mapsto [\mathbf{x}\ 1]$.

Με άλλα λόγια, έστω ένα τροπικό πολυώνυμο $p$. Σε κάθε κόμβο $\mathbf{c}_i\mathbb{R}^n$ του πολυωνύμου θεωρούμε μία ύψωση (lift) ίση με το σταθερό όρο $b_i \in \mathbb{R}$ των μονωνύμων που απαρτίζουν το πολυώνυμο. Αυτή η κατασκευή δημιουργεί το λεγόμενο επεκτεταμένο πολύτοπο Newton (extended Newton Polytope) eNewt$(p)$:

$$\text{eNewt}(p) = \text{conv}\{(\mathbf{c}_i, b_i) \in \mathbb{R}^n \times \mathbb{R} : i \in I\} \tag{2.3.10}$$

Τότε, καταλήγουμε στην έννοια του άνω καλύμματος $\mathcal{U}(p)$ που αντιστοιχεί στη συλλογή των άνω faces του επεκτεταμένου πολύτοπου Newton eNewt$(p)$. Στο παρακάτω σχήμα 2.3.8 παρουσιάζονται τα Upper και Newton Hulls του τροπικού πολυωνύμου $p(x_1, x_2) = \min\{1 + 2x_1, 2 + x_1, 2 + x_1 + x_2, 2 + x_2, 1 + 2x_2\}$.




Σχήμα 2.3.8: Upper και Newton Hull του πολυωνύμου
$p(x_1, x_2) = \min\{1 + 2x_1, 2 + x_1, 2 + x_1 + x_2, 2 + x_2, 1 + 2x_2\}$

Ως το σημείο αυτό, η συζήτηση έχει περιοριστεί σε συναρτήσεις που αντιστοιχούν στο maximum αφφινικών όρων, δηλαδή σε τροπικά πολυώνυμα. Η ανάλυση είναι ελλιπής, δίχως τη μελέτη μοντέλων που συνδυάζουν τροπικά πολυώνυμα, είτε την πρόσθεσή τους είτε το supremum τους. Η επακόλουθη ανάλυση αφορά τη max-plus θεώρηση.

Στο πλαίσιο της τροπικής γεωμετρίας και με θεμέλιους λίθους τα max-plus πολυώνυμα, οι μαθηματικές κατασκευές που δύναται να ανοικοδομηθούν βασίζονται στους δύο τελεστές, είτε supremum είτε πρόσθεση. Έστω, λοιπόν, max-plus πολυώνυμα $p_i : \mathbb{R}^n \to \mathbb{R}, i \in [m]$.

Σε κάθε περίπτωση, αξίζει να παρατηρήσουμε ότι κάθε τροπικό πολυώνυμο $p_i$ αντιστοιχεί σε ένα πολύτοπο Newton. Κατά συνέπεια, μεταφράζουμε τον αλγεβρικό συνδυασμό των πολυωνύμων στη γεωμετρική σύνθεση πολυτόπων. Η απλή περίπτωση έγκειται στο μοντέλο που αποτελείται από το supremum των $p_i$:

$$f^{\vee}(\mathbf{x}) = \bigvee_{i=1}^{m} p_i(\mathbf{x}) \tag{2.3.11}$$

Γνωρίζουμε ότι ο τελεστής $\vee$ είναι ταυτοδύναμος. Για παράδειγμα:

$$\max\{\max\{a,b,c\},\max\{a,d,e\},\max\{a,b,f,g\}\} = \max\{a,b,c,a,d,e,a,b,f,g\} = \max\{a,b,c,d,e,f,g\}$$

Από τον ορισμό του πολύτοπου Newton (4) και με θεμέλιους λίθους τα πολύτοπα των εκάστοτε πολυωνύμων (αντί για τους κόμβους των εκθετών των μονωνύμων), έχουμε ότι το (συνολικό) πολύτοπο Newton για το μοντέλο $f^{\vee}$ είναι:

$$\mathrm{Newt}(f^{\vee}) = \mathrm{conv}\{\mathrm{Newt}(p_1),\dots,\mathrm{Newt}(p_m)\} \tag{2.3.12}$$

Στρέφουμε την προσοχή μας στο προσθετικό μοντέλο:

$$f^{+}(\mathbf{x}) = \sum_{i=1}^{m} p_i(\mathbf{x}) \tag{2.3.13}$$

**Theorem 2.3.6: Minkowski sum [Zie95, κεφ. 1.1]**

Έστω δύο σύνολα $P, Q \subseteq \mathbb{R}^d$. Τότε το Minkowski άθροισμά τους ορίζεται ως:

$$P \oplus Q = \{\mathbf{x}+\mathbf{y} : \mathbf{x} \in P, \mathbf{y} \in Q\} \tag{2.3.14}$$

Επεκτείνοντας το παραπάνω θεώρημα για τα σύνολα που δημιουργούν τα πολύτοπα Newton καταλήγουμε στο ακόλουθο:

$$\mathrm{Newt}(f^{+}) = \mathrm{Newt}(p_1) \oplus \cdots \oplus \mathrm{Newt}(p_m) \tag{2.3.15}$$

Για την καλύτερη κατανόηση των παραπάνω, εξετάζουμε ένα απλό παράδειγμα. Έστω δύο πολυώνυμα $p_i : \mathbb{R}^2 \to \mathbb{R}, i = 1, 2$ με

$$\begin{aligned} p_1(x,y) &= \max\{0, 2x, 2y, 2x+2y\} \\ p_2(x,y) &= \max\{x+y, x+2y, 2x-y\} \end{aligned}$$

των οποίων τα πολύτοπα Newton φαίνονται στα σχήματα 2.3.9a, 2.3.9b αντίστοιχα. Στη δεύτερη σειρά του σχήματος 2.3.9, παρουσιάζονται οι συνθέσεις των πολυωνύμων από γεωμετρική σκοπιά. Για τη διευκόλυνση του αναγνώστη, χρησιμοποείται πράσινο για το τροπικό πολυώνυμο $p_1$ και μπλε για το $p_2$. Με μαύρη διακεκομμένη γραμμή και μοτίβο (αντί για χρώμα) επισημαίνεται το αποτέλεσμα.

Για το μοντέλο με πράξη supremum βλ. σχήμα 2.3.9c. Παρατηρούμε ότι γεωμετρικά τοποθετούμε τα δύο πολύτοπα Newton σε κοινό διάγραμμα και λαμβάνουμε το κυρτό κάλυμα. Από το σχήμα έχουμε:

$$\begin{aligned} p^{\vee}(x,y) &= \max\{p_1(x,y), p_2(x,y)\} \\ &= \max\{\max\{0,2x,2y,2x+2y\}, \max\{x+y,x+2y,2x-y\}\} \\ &= \max\{0,2x,2y,2x+2y,x+y,x+2y,2x-y\} \\ &= \max\{0,2x-y,2x+2y,2y\} \end{aligned}$$

όπου η απλοποίηση του τελευταίου βήματος γίνεται με τη βοήθεια του σχήματος. Η έξοδος του προσθετικού μοντέλου είναι ελαφρώς περίπλοκη. Παρατηρούμε ότι το πολυώνυμο $p_2$ αποτελείται από 3 όρους. Τοποθετούμε το πολύτοπο $\mathrm{Newt}(p_1)$ σε κάθε μία κορυφή του πολύτοπου Newton για το πολυώνυμο $p_2$. Εδώ χρειάζεται προσοχή, καθώς η τοποθέτηση γίνεται με τρόπο τέτοιο ώστε η αρχή των αξόνων να συμπίπτει με το εκάστοτε σημείο. Χάριν ευκολίας, επιλέγεται το $\mathrm{Newt}(p_1)$ ως στοιχείο μετατόπισης, καθώς περιλαμβάνει την αρχή των αξόνων. Ωστόσο, λόγω της αντιμεταθετικότητας της πράξης πρόσθεσης Minkowski $\oplus$, αυτή η επιλογή δεν μεταβάλλει το αποτέλεσμα. Αλγεβρικά, έχουμε:

$$\begin{aligned} p^+(x,y) &= p_1(x,y) + p_2(x,y) \\ &= \max\{0, 2x, 2y, 2x+2y\} + \max\{x+y, x+2y, 2x-y\} \\ &= \max\{x+y, 2x-y, 4x-y, 4x+y, 3x+4y, x+4y\} \end{aligned}$$

όπου η απλοποίηση του τελευταίου βήματος γίνεται με τη βοήθεια του σχήματος 2.3.15.

Τέλος, εξετάζουμε το μετασχηματισμό που επιφέρει σε ένα πολύτοπο η ύψωση του σχετικού τροπικού πολυωνύμου σε δύναμη $\alpha$. Έστω, $\mathcal{P}$ το πολύτοπο που ορίζει το τροπικό πολυώνυμο $p$. Έστω $p(x,y) = (\max\{x,y\})^{\odot 3}$, όπου επισημειώνουμε τη δύναμη με το σύμβολο $\odot$ για την τροπική δύναμη (δηλαδή κλασικό πολλαπλασιασμό) για αποφυγή σύγχυσης. Τότε:

$$p(x,y) = (\max\{x,y\})^{\odot 3} = \max\{x,y\} + \max\{x,y\} + \max\{x,y\} = 3\max\{x,y\}$$

Εύκολα διαπιστώνουμε ότι η παραπάνω ταυτότητα, η οποία αναφέρεται ως *Freshman's Dream* στη βιβλιογραφία [MS15], ισχύει για κάθε δύναμη $\alpha$. Γεωμετρικά, λοιπόν, η ύψωση σε δύναμη αντιστοιχεί σε μία κλιμάκωση, ώστε $\mathcal{P}(p^{\odot\alpha}) = \alpha\mathcal{P}(p)$ με $\alpha\mathcal{P}(p) = \{ax : x \in \mathcal{P}(p)\}$ [ZNL18]. Διατηρείται, λοιπόν, το σχήμα του πολυτόπου αλλά μεταβάλλεται ο όγκος του.

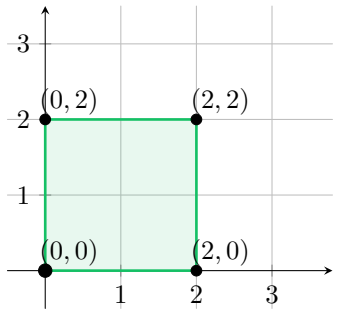
## 2.4 Εφαρμογές

Κατόπιν της εισαγωγής στα τροπικά μαθηματικά, αξίζει να γίνει μία σύντομη αναφορά σε εφαρμογές και συνδέσεις με άλλα επιστημονικά πεδία που χρήζουν μνείας. Η τροπική γεωμετρία χρησιμοποιείται σε πολλούς κλάδους. Εστιάζουμε την προσοχή μας στην Επιστήμη των Υπολογιστών. Στο σχήμα 2.4.1, παρουσιάζονται εποπτικά οι διάφορες περιοχές ενδιαφέροντος.

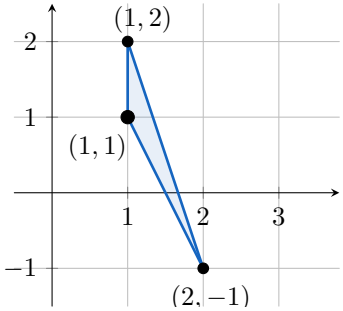
Ιστορικά, η τροπική γεωμετρία θεμελιώθηκε στο πλαίσιο της επιχειρησιακής έρευνας από τον Cuninghame-Green στο έργο του Minimax Algebra [Cun79]. Μελετήθηκε το τροπικό διοειδές (αν και όχι με αυτή την ορολογία) και βρέθηκε η λύση του συστήματος max-plus εξισώσεων στον ομώνυμο πίνακα (βλ. σχέση (2.2.3)).

Οι εφαρμογές της Τροπικής Γεωμετρίας δεν περιορίζονται στην επιχειρησιακή έρευνα. Στενή είναι η σύνδεσή της με την Οικονομική Θεωρία καθώς και τη Θεωρία Παιγνίων. Οι Akian, Gaubert, and Guterman συνδέουν τα τροπικά πολύεδρα με την κλάση zero-sum στοχαστικών παιχνιδιών *mean-payoff games* [AGG12]. Συγκεκριμένα, αποδεικνύουν την ισοδυναμία των τροπικών πολυέδρων με ντετερμινιστικά παιχνίδια με πεπερασμένο αριθμό κινήσεων και συνδέουν την ύπαρξη αρχικών νικητήριων καταστάσεων με το σχετικό τροπικό πολύεδρο. Άλλα μέλη της ίδιας ερευνητικής ομάδας συνδέουν προβλήματα κλασματικού τροπικού προγραμματισμού με παραμετρικά mean-payoff games [GKS12]. Παράλληλα, η τροπική μοντελοποίηση επεκτείνεται και σε άλλους κλάδους της Θεωρίας Παιγνίων, όπως ο σχεδιασμός μηχανισμών (mechanism design) και οι δημοπρασίες (auction theory). Οι Crowell and Tran αναλύουν πολυδιάστατους μηχανισμούς πεπερασμένης αξίας στο τροπικό πλαίσιο και συνδέουν τη συμβατότητα με το κίνητρο (incentive compatibility) με τα κελιά του τροπικού κυρτού καλύμματος [CT16], ενώ οι Tran and Yu εξετάζουν μία άλλη κλάση μηχανισμών, τις λεγόμενες δημοπρασίες μείξης αγαθών (Product-Mix Auctions), υπό το τροπικό πρίσμα [TY15].

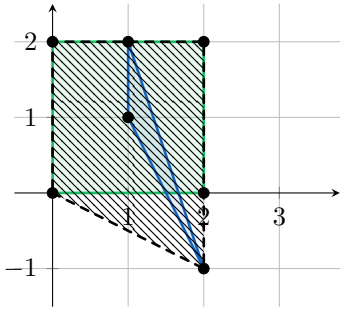
Μνεία χρήζει η σύνδεση του κλάδου της Βελτιστοποίησης με τα τροπικά μαθηματικά. Οι Allamigeon et al. ανέπτυξαν το ανάλογο του γνωστού αλγορίθμου Simplex στον τροπικό ημιδακτύλιο [All+15], ενώ σε προηγούμενη δουλειά τους συνέδεσαν τον τροπικό simplex με mean-payoff games [All+14].



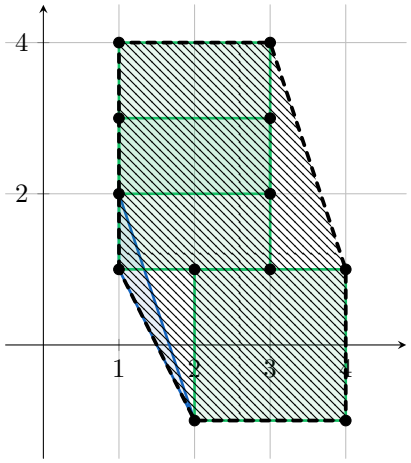

(a) Newt($p_1$)




(b) Newt($p_1$)




(c) Newt($p_1 \vee p_2$)

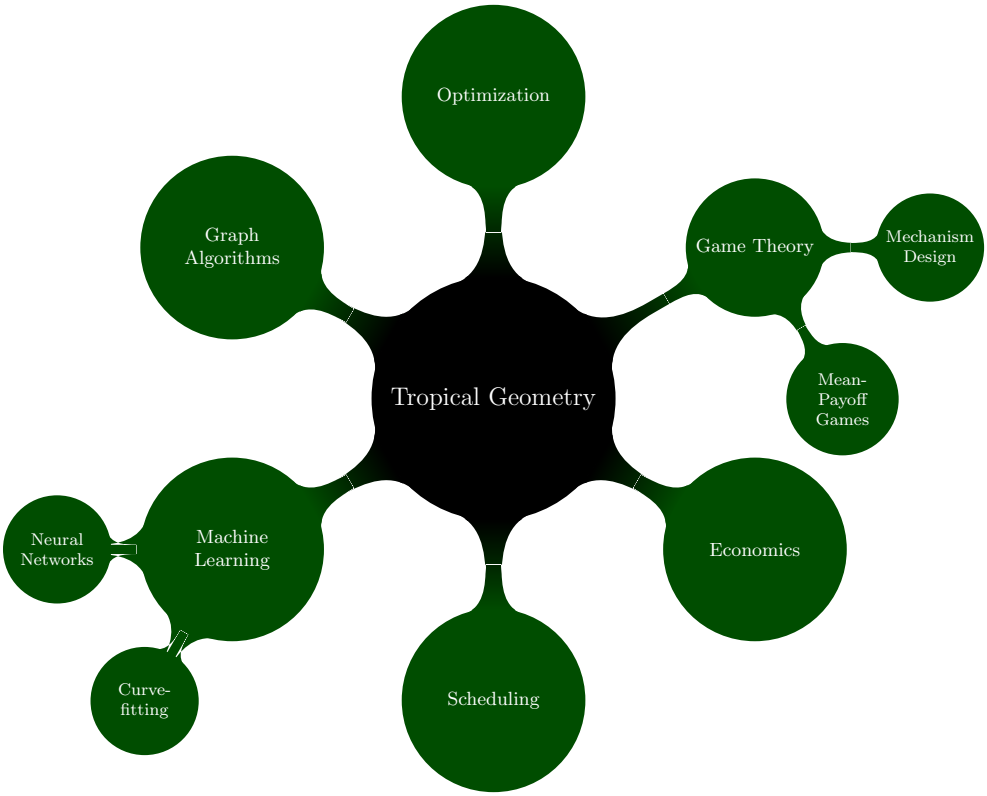



(d) Newt($p_1$) $\oplus$ Newt($p_2$)

Σχήμα 2.3.9: Πολύτοπα Newton από συνδυασμό τροπικών πολυωνύμων

Επιπρόσθετα, η τροπική γεωμετρία έχει αξιοποιηθεί στην ανάλυση γραφικών μοντέλων και σχετικών αλγορίθμων. Οι Pachter and Sturmfels εξερεύνησαν τη σύνδεση με max-product αλγορίθμους όπως Viterbi [PS04], ενώ οι Theodosis and Maragos επέκτειναν σημαντικά τις ιδέες σε πιο σύνθετα μοντέλα, όπως Weighted Finite State Transducers (WFSTs), αναπτύσσοντας ευριστικές μεθόδους pruning [TM18b; TM18c; TM18a].

Τέλος, οι εφαρμογές εντοπίζονται και στη σφαίρα της μηχανικής μάθησης. Οι Maragos and Theodosis ανέπτυξαν μία μέθοδο προσέγγισης τροπικών καμπυλών με ρίζες στην τροπική γεωμετρία και τη θεωρία πλεγμάτων [MT19]. Παράλληλα, οι Charisopoulos and Maragos εντόπισαν τη σύνδεση των μορφολογικών δικτύων με τα τροπικά μαθηματικά, επινοώντας αλγορίθμους [CM17] και χαρακτηρίζοντας γεωμετρικά τα δίκτυα [CM18].




Σχήμα 2.4.1: Πεδία εφαρμογών της Τροπικής Άλγεβρας

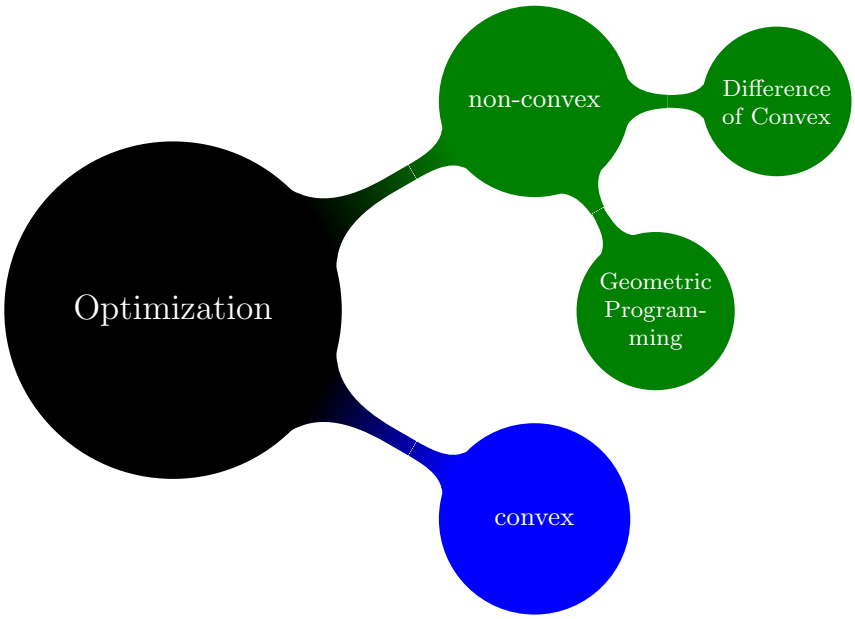
# Κεφάλαιο 3

# Βελτιστοποίηση

Στο κεφάλαιο αυτό θα μελετήσουμε ορισμένες έννοιες της βελτιστοποίησης, τόσο κυρτής όσο και μη κυρτής. Στο πλαίσιο της μη-κυρτής βελτιστοποίησης, παρουσιάζονται κατηγορίες προβλημάτων που δύναται να μετατραπούν σε κυρτά, είτε μέσω μίας χαλάρωσης των μη κυρτών στοιχείων (βλ. παράγραφο 3.2.2) είτε μέσω αλλαγής μεταβλητής (βλ. παράγραφο 3.2.3). Στη συνέχεια, παρατίθενται ορισμένα στοιχεία τροπικής γεωμετρίας, μέσω των οποίων θα συνδέσουμε την τροπική βελτιστοποίηση με την κυρτή και θα εξετάσουμε πως πολυμελετημένοι αλγόριθμοι βελτιστοποίησης εφαρμόζονται στον τροπικό ημιδακτύλιο. Τα προβλήματα βελτιστοποίησης λαμβάνουν τη μορφή:

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in P \end{aligned} \tag{3.0.1}$$

όπου η συνάρτηση $f(\cdot)$ λέγεται αντικειμενική συνάρτηση και το σύνολο $P$ ορίζει το χώρο που μπορεί να κινείται η λύση μας, δηλαδή το $P$ κωδικοποιεί τους περιορισμούς. Η μορφή της αντικειμενικής συνάρτησης και του πεδίου περιορισμών καθορίζουν την τάξη του προβλήματος και, κατά συνέπεια, τους αλγορίθμους επίλυσης. Στο παρακάτω σχήμα, παρουσιάζεται ένας εποπτικός χάρτης των κλάσεων βελτιστοποίησης που θα αναλυθούν. Σε κάθε περίπτωση αναφερόμαστε σε βελτιστοποίηση $\mu\epsilon$ περιορισμούς.




Σχήμα 3.0.1: Optimization mindmap

## 3.1 Κυρτή Βελτιστοποίηση

Ξεκινούμε με την πιο απλή κλάση προβλημάτων. Ο κλάδος της κυρτής βελτιστοποίησης έχει μελετηθεί εκτεταμένα, καθώς βρίσκει αντίκρισμα σε πολλές εφαρμογές και οι λύσεις των προβλημάτων μπορούν να υπολογιστούν με αποδοτικούς αλγορίθμους. Αυτός συνδυασμός απόδοσης και ευρείας εφαρμογής των αλγορίθμων καθιστούν την κυρτή βελτιστοποίηση ένα εξαιρετικά σημαντικό εργαλείο. Οι περισσότερες έννοιες αντλούνται από το βιβλίο των Boyd and Vandenberghe [BV04]. Η κυρτότητα σε επίπεδο συνάρτησης δίνεται στον ορισμό 1. Παρατίθεται και η κυρτότητα σε επίπεδο συνόλων:

**Definition 3.1.1: Κυρτό Σύνολο**

Έστω σύνολο $S \subset \mathcal{V}$, όπου $\mathcal{V}$ ένας διανυσματικός χώρος (ή διατεταγμένο πεδίο γενικότερα). Το σύνολο $S$ αποκαλείται κυρτό αν:

$$x, y \in S \implies tx + (1-t)y \in S, \quad \forall t \in [0,1] \tag{3.1.1}$$

Γιατί, όμως, η ιδιότητα της κυρτότητας είναι τόσο σημαντική; Η ελαχιστοποίηση μίας κυρτής συνάρτησης είναι διαισθητικά απλή. Ας εξετάσουμε την περίπτωση χωρίς περιορισμούς, όπου η κυρτότητα σε συνδυασμό με την τιμή της πρώτης παραγώγου μας πληροφορούν για την εύρεση *ολικού* ελαχίστου. Αναλυτικότερα, και ιδιαίτερα στην περίπτωση αυστηρής κυρτότητας, ο μηδενισμός της πρώτης παραγώγου συνεπάγεται ολικό ελάχιστο και ο οποιοσδήποτε αλγόριθμος τερματίζει. Στην περίπτωση που η παράγωγος δεν είναι μηδενική, η κίνηση στην αντίθετή της κατεύθυνση εγγυάται μείωση του συναρτησιακού. Συνεπώς, είναι γνωστό το κριτήριο τερματισμού και δεν τίθεται λόγος για τοπικά ελάχιστα. Στρέφουμε την προσοχή μας στα κυρτά σύνολα: αν δύο σημεία $x, y$ ανήκουν σε ένα κυρτό σύνολο $S$, τα σημεία που βρίσκονται επί του ευθύγραμμου τμήματος που ενώνει τα $x, y$ ανήκουν στο $S$. Διάφοροι γνωστοί αλγόριθμοι, κυρίως της οικογένειας της κατάβασης κλίσεων, όπως Newton-Raphson, Conjugate Gradient Descent ή Franke-Wolfe εφαρμόζονται σε αυτή την κλάση προβλημάτων. Κοινό στοιχείο αυτών των επαναληπτικών αλγορίθμων είναι το βήμα γραμμικής αναζήτησης (line search techniques) όπου επιλέγεται ένα στοιχείο στο ευθύγραμμο τμήμα ανάμεσα σε $x, y \in S$. Η κυρτότητα, λοιπόν, εγγυάται ότι το σημείο που επιλέγεται ανήκει στο σύνολο περιορισμών $S$.

Τα προβλήματα κυρτής βελτιστοποίησης έχουν αντικειμενική συνάρτηση *και* περιορισμούς κυρτής φύσης.

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0 \quad i \in [m] \end{aligned} \tag{3.1.2}$$

με $f_i(x)$ κυρτές συναρτήσεις.

## 3.2 Μη-Κυρτή Βελτιστοποίηση

Τα προβλήματα κυρτής βελτιστοποίησης έχουν κομψές λύσεις και εγγυήσεις ότι η εύρεση ενός τοπικού ελαχίστου αντιστοιχεί σε global optimum. Ωστόσο, στην πράξη εμφανίζονται πολύ συχνά συναρτήσεις που παρουσιάζουν μη-κυρτή δομή. Χαρακτηριστικό παράδειγμα αποτελούν οι περιοχές απόφασης (decision boundaries) που ορίζουν τα νευρωνικά δίκτυα.

### 3.2.1 Difference of Convex Programming

Τα προβλήματα Difference of Convex (DC) βελτιστοποίησης αποτελούν μία ειδική περίπτωση της μη κυρτής βελτιστοποίησης και λαμβάνουν τη μορφή:

$$\begin{aligned} \min \quad & f_0(x) - g_0(x) \\ \text{s.t.} \quad & f_i(x) - g_i(x) \leq 0 \quad i \in [m] \end{aligned} \tag{3.2.1}$$

όπου $x \in \mathbb{R}^n$ η μεταβλητή βελτιστοποίησης και οι συναρτήσεις $f_i, g_i : \mathbb{R}^n \to \mathbb{R}, i \in [m]$ είναι κυρτές. Σε πρώτη ματιά, φαίνεται πως αυτή η μορφή εμπεριέχει πολύ λίγες συναρτήσεις, ωστόσο ο Hartman απέδειξε ότι το DC πρόβλημα είναι πολύ γενικό. Κατά συνέπεια, πληθώρα μη κυρτών συναρτήσεων μπορούν να εκφραστούν ως διαφορά δύο κυρτών συναρτήσεων.

**Theorem 3.2.1: Hartman [Har59]**

> Έστω $E(\vec{x})$ μία συνάρτηση ενέργειας με φραγμένη Hessian $\frac{\partial^2 E(\vec{x})}{\partial \vec{x} \partial \vec{x}}$. Τότε, μπορούμε να την αναλύσουμε ως το άθροισμα μίας κυρτής και μίας κοίλης συνάρτησης (δηλαδή ως διαφορά δύο κυρτών).

Συνεπώς, όλες οι $C^2$ συναρτήσεις περιλαμβάνονται στην κλάση των προβλημάτων που περιγράφει η 3.2.1. Σε γενικά προβλήματα μη κυρτής βελτιστοποίησης χρησιμοποιούνται μέθοδοι επίλυσης που βασίζονται σε branch-and-bound τεχνικές, με αποτέλεσμα να μην είναι ιδιαίτερα αποδοτικές. Αντιθέτως, τα DC προβλήματα χαρακτηρίζονται από αποδοτικούς αλγορίθμους και εγγυήσεις ως προς τη σύγκλιση σε (τοπικά) ελάχιστα. Μία μεγάλη οικογένεια αλγορίθμων ονομάζεται Difference-of-Convex Algorithms, ή DCA εν συντομία, και βασίζεται στη δυικότητα των προβλημάτων και στη συζυγή συνάρτηση (conjugate function):

$$g^*(\mathbf{y}) = \sup_{\mathbf{x} \in \mathbf{R^n}} \{\mathbf{x}^\top \mathbf{y} - \mathbf{g}(\mathbf{x})\} \tag{3.2.2}$$

Η συζηγής συνάρτηση $g^*$ είναι κυρτή, ακόμα και η $g$ δεν είναι. Η ιδιότητα αυτή επιτρέπει την κατασκευή αποδοτικών αλγορίθμων. Από τον ορισμό 3.2.2 προκύπτει άμεσα η ανισότητα Fenchel:

$$g(\mathbf{x}) + g^*(\mathbf{y}) \geq \mathbf{x}^\top y \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \tag{3.2.3}$$

Η ισότητα ισχύει στην περίπτωση όπου η συνάρτηση $g$ είναι κυρτή. Τότε, η δεύτερη συζηγής συνάρτηση $g^{**}$, δηλαδή η συζηγής της συζηγούς, είναι ίση με την αρχική. Γενικότερα, ισχύει $g^{**}(\mathbf{x}) \leq g(\mathbf{x})\ \forall \mathbf{x} \in \mathbb{R}^n$.

### 3.2.2 Convex Concave Procedure

Θα παρουσιάσουμε τη διαδικασία Convex-Concave [YR03]. Πρόκειται για μία ευριστική μέθοδο για την επίλυση DC προβλημάτων που δεν σχετίζεται στη συζηγή συνάρτηση. Βασίζεται σε μία απλή παρατήρηση: η διαφορά μίας αφφινικής συνάρτησης από μία μία κυρτή συνάρτηση έχει ως αποτέλεσμα μία κυρτή συνάρτηση. Με άλλα λόγια, η (προσθ)αφαίρεση αφφινικών συναρτήσεων από μία κυρτή συνάρτηση δεν επηρεάζει την κυρτότητά της. Έστω, λοιπόν, $f, g : \mathbb{R}^n \to \mathbb{R}$ δύο κυρτές συναρτήσεις και $h(\mathbf{x}) = f(\mathbf{x}) - g(\mathbf{x})$ η διαφορά τους, με $\mathbf{x} \in \mathbf{dom}\, f \cap \mathbf{dom}\, g$, όπως εμφανίζεται τόσο στο σύνολο περιορισμών όσο και στην αντικειμενική συνάρτηση σε DC προβλήματα. Αντικαθιστώντας την κυρτή συνάρτηση $g$ με μία αφφινική προσέγγισή της $\tilde{g}$, η προσέγγιση $\tilde{h}$ μετατρέπεται σε κυρτή συνάρτηση. Η εν λόγω προσέγγιση πραγματοποείται μέσω αναπτύγματος Taylor πρώτου βαθμού. Χρησιμοποιούμε το συμβολισμό ενός πολυ-εκθέτη. Ένας $n$-διάστατος πολυ-εκθέτης είναι μία $n$-τούπλα $\alpha = (\alpha_1, \cdots, \alpha_n)$, με $\alpha_i$ μη αρνητικούς ακεραίους, και τις ιδιότητες $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n$ και $\alpha! = \alpha_1!\alpha_2! \cdots \alpha_n!$. Επιπλέον, $\mathbf{x}^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$. Τότε, το Taylor ανάπτυγμα της διαφορίσιμης συνάρτησης πολλών μεταβλητών $g$ είναι:

$$T_{\mathbf{x}_0}(\mathbf{x}) = \sum_{|\alpha| \geq 0} \frac{(\mathbf{x} - \mathbf{x}_0)^\alpha}{\alpha!} (\partial^\alpha g)(\mathbf{x}_0) \tag{3.2.4}$$

Παρατηρούμε ότι το άθροισμα αφορά όλους τους συνδυασμούς εκθετών. Στο πλαίσιο της διαδικασίας Convex-Concave χρησιμοποείται το ανάπτυγμα Taylor πρώτου βαθμού, το οποίο προκύπτει περιορίζοντας κατάλληλα τη σχέση (3.2.4):

$$T_{\mathbf{x}_0}(\mathbf{x}) = g(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x_0})^\top \nabla g(\mathbf{x}_0) \tag{3.2.5}$$

Μάλιστα, λόγω της κυρτότητας της $g$ στο DC πρόβλημα (3.2.1) έχουμε:

$$g(\mathbf{y}) \geq \mathbf{g}(\mathbf{x}) + \nabla \mathbf{g}(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

και, συνεπώς,

$$\begin{aligned} \tilde{h}(\mathbf{x}; \mathbf{x_k}) &= f(\mathbf{x}) - \hat{\mathbf{g}}(\mathbf{x}; \mathbf{x_k}) \\ &= f(\mathbf{x}) - [\mathbf{g}(\mathbf{x_k}) + \nabla \mathbf{g}(\mathbf{x_k})(\mathbf{x} - \mathbf{x_k})] \\ &\geq f(\mathbf{x}) - \mathbf{g}(\mathbf{x}) \\ &= h(\mathbf{x}) \end{aligned}$$

Η παραπάνω σχέση εξασφαλίζει την πραγματοποιησιμότητα των συνθηκών, καθώς η ικανοποίηση των περιορισμών του γραμμικοποιημένου προγράμματος εγγυάται την ικανοποίηση των περιορισμών του αρχικού. Κυρτοποιώντας την αντικειμενική συνάρτηση και τους περιορισμούς, καταλήγουμε στον αλγόριθμο. Παρουσιάζεται

παρακάτω:

**Algorithm 1:** Convex-Concave Procedure [LB16, §1.2]

**Data:** *feasible* initial point $\mathbf{x}_0$

$k \leftarrow 0$

**repeat**

  `Convexify.` Form $\hat{g}_i(\mathbf{x}; \mathbf{x}_k) \triangleq g_i(\mathbf{x}_k) + \nabla g_i(\mathbf{x}_k)^\top (\mathbf{x} - \mathbf{x}_k)$ for $i = 0, 1, \ldots, m$.

  `Solve.` Set the value of $\mathbf{x}_{k+1}$ to a solution of

$$\begin{aligned} &\text{minimize} && f_0(\mathbf{x}) - \hat{g}_0(\mathbf{x}; \mathbf{x}_k) \\ &\text{subject to} && f_i(\mathbf{x}) - \hat{g}_i(\mathbf{x}; \mathbf{x}_k) \leq 0, \quad i = 1, ...., m \end{aligned}$$

  `Update iteration.` $k \leftarrow k + 1$

**until** *stopping criterion is satisfied*

Ένα δυνατό κριτήριο τερματισμού είναι η βελτίωση της αντικειμενικής συνάρτησης να μην υπερβαίνει ένα κατώφλι $\delta > 0$:

$$\tilde{h}_0(\mathbf{x}; \mathbf{x}_k) - \tilde{h}_0(\mathbf{x}; \mathbf{x}_{k+1}) = (f_0(\mathbf{x}) - \hat{g}_0(\mathbf{x}; \mathbf{x}_k)) - (f_0(\mathbf{x}) - \hat{g}_0(\mathbf{x}; \mathbf{x}_{k+1})) \leq \delta \tag{3.2.6}$$

Αξίζει να σημειωθεί ότι εφόσον η χαλάρωση του αρχικού προβλήματος είναι κυρτή, η αριστερή πλευρά της παραπάνω ανίσωσης δε λαμβάνει αρνητικές τιμές. Όσον αφορά την αρχικοποίηση του αλγορίθμου, οι Yuille and Rangarajan προτείνουν τη χρήση πολλών αρχικών σημείων $\mathbf{x}_0$ και την επιλογή της βέλτιστης λύσης. Αυτό οφείλεται στο γεγονός ότι η διαδικασία Convex-Concave είναι μία τοπική ευριστική μέθοδος και, συνεπώς, η απόδοσή της είναι άρρηκτα συνδεδεμένη με το αρχικό σημείο. Ωστόσο, κάθε σημείο που χρησιμοποείται για αρχικοποίηση πρέπει να ικανοποιεί τους περιορισμούς.

#### Penalty Convex-Concave Procedure

Οι Lipp and Boyd προτείνουν μία επέκταση της διαδικασίας Convex-Concave των Yuille and Rangarajan, η οποία αφαιρεί την απαίτηση το αρχικό σημείο $\mathbf{x}_0$ να ικανοποιεί τους περιορισμούς. Πρόκειται για μία χαλάρωση του αρχικού προβλήματος μέσω της εισαγωγής slack μεταβλητών στις συνθήκες. Αναλυτικότερα, επιτρέπεται η παραβίαση του περιορισμού $f_i(\mathbf{x}) - \hat{g}_i(\mathbf{x}; \mathbf{x}_k) \leq 0$ έως και $s_i$ και η συνθήκη γίνεται $f_i(\mathbf{x}) - \hat{g}_i(\mathbf{x}; \mathbf{x}_k) \leq s_i$. Η ακριβής λογική παρουσιάζεται στον αλγόριθμο 2.

Η χαλάρωση του προβλήματος ενδέχεται να οδηγήσει σε μη εφικτά σημεία και, για αυτό το λόγο, επιβάλλεται ποινή στο άθροισμα των slack μεταβλητών $s_i$. Αυτή μοντελοποίηση είναι παρόμοια με τη χρήση hinge loss. Ακόμη, η ποινή στο άθροισμα $\sum_{i=1}^{m} s_i$ είναι αντίστοιχη με τη χρήση της $\ell_1$ νόρμας και, συνεπώς, παράγει αραιές λύσεις. Αυτό σημαίνει ότι ακόμα και αν δεν ικανοποιούνται όλες οι αρχικές συνθήκες, το πλήθος των παραβάσεων θα είναι μικρό.

Το κριτήριο τερματισμού (3.2.6) αναβαθμίζεται για να περιλαμβάνει τον όρο penalty:

$$\left( f_0(\mathbf{x}) - \hat{g}_0(\mathbf{x}; \mathbf{x}_k) + \tau_k \sum_{i=1}^{m} s_i^{(k)} \right) - \left( f_0(\mathbf{x}) - \hat{g}_0(\mathbf{x}; \mathbf{x}_{k+1}) \ + \tau_k \sum_{i=1}^{m} s_i^{(k+1)} \right) \leq \delta \tag{3.2.7}$$

Το άνω όριο $\tau_{\max}$ έχει διττό λόγο ύπαρξης: αφενός συμβάλλει στην αποφυγή αριθμητικών προβλημάτων σε περίπτωση που το $\tau_i$ λάβει ιδιαίτερα μεγάλη τιμή και, αφετέρου, συνδράμει στη σύγκλιση σε περίπτωση που δε βρεθεί εφικτό σημείο.

### 3.2.3 Geometric Programming

Ο κλάδος του γεωμετρικού προγραμματισμού (ή geometric programming ή GP εν συντομία) ασχολείται με μη-γραμμικά προβλήματα που έχουν ιδιαίτερη μορφή. Αυτή η ιδιαίτερη μορφή επιτρέπει τη μετατροπή του προβλήματος σε κυρτό μέσω κατάλληλης αλλαγής μεταβλητών [Boy+07].

**Algorithm 2:** Penalty Convex-Concave Procedure [LB16, §3.1]

**Data:** initial point $\mathbf{x}_0$, $\tau_0 > 0$, $\tau_{\max}$, $\mu > 1$

$k \leftarrow 0$

**repeat**

`Convexify`. Form $\hat{g}_i(\mathbf{x}; \mathbf{x}_k) \triangleq g_i(\mathbf{x}_k) + \nabla g_i(\mathbf{x}_k)^\top (\mathbf{x} - \mathbf{x}_k)$ for $i = 0, 1, \ldots, m$.

`Solve`. Set the value of $\mathbf{x}_{k+1}$ to a solution of

$$\begin{array}{lll} \text{minimize} & f_0(\mathbf{x}) - \hat{g}_0(\mathbf{x}; \mathbf{x}_k) + \tau_k \sum_{i=1}^{m} s_i & \\ \text{subject to} & f_i(\mathbf{x}) - \hat{g}_i(\mathbf{x}; \mathbf{x}_k) \leq s_i, & i = 1, ...., m \\ & s_i \geq 0, & i = 1, ...., m \end{array}$$

`Update` $\tau$. $\tau_{k+1} \leftarrow \min(\mu \cdot \tau_k, \tau_{\max})$

`Update iteration`. $k \leftarrow k + 1$

**until** *stopping criterion is satisfied*

Η ενασχόληση με τη συγκεκριμένη κλάση μη-κυρτών προβλημάτων βασίζεται σε ορισμένες ισχυρές ιδιότητες που τους προσδίδει η στενή σχέση με την κυρτή βελτιστοποίηση. Αναλυτικότερα, τα προβλήματα γεωμετρικού προγραμματισμού χαίρουν ορισμένων ισχυρών ιδιοτήτων:

- Σε αντίθεση με τη γενική κλάση των μη-κυρτών προβλημάτων βελτιστοποίησης, τα GP προβλήματα επιλύονται αποτελεσματικά.
- μοναδικότητα λύσης: αν υπάρχει ένα τοπικά βέλτιστο σημείο, τότε είναι και το global optimum. Επιπλέον, είναι εγγυημένη η σύγκλιση σε αυτό.
- Οι μέθοδοι επίλυσης είναι εύρωστοι(robust), καθώς δεν χρειάζονται αρχικοποίηση ή σύνθετη εύρεση/fine-tuning παραμέτρων.
- πολλά προβλήματα έχουν "γεωμετρική" φύση και μπορούν να μοντελοποιηθούν με γεωμετρικό προγραμματισμό. Εφαρμογές περιλαμβάνουν σχεδίαση VLSI ή βελτιστοποίηση σχεδιασμού αεροσκαφών [HA14].

Οι παραπάνω ιδιότητες αναδεικνύουν τη σημασία αυτού του κλάδου βελτιστοποίησης και αξίζει να εντρυφήσουμε περαιτέρω. Ο θεμέλιος λίθος του γεωμετρικού προγραμματισμού είναι το μονώνυμο (monomial), το οποίο λαμβάνει την ακόλουθη μορφή:

$$f(\mathbf{x}) = c x_1^{a_1} x_2^{a_2} ... x_n^{a_n} = c\mathbf{x}^{\mathbf{a}} \tag{3.2.8}$$

με $a_i \in \mathbb{R}$, ενώ οι μεταβλητές $x_i$ λαμβάνουν θετικές τιμές. Για παράδειγμα, η συνάρτηση $f(x_1, x_2) = 5x_1^3 x_2^{-0.23}$ είναι έγκυρη αλλά όχι κυρτή. Τα posynomials αποτελούν το άθροισμα πολλών monomials:

$$g(x) = \sum_{k=1}^{K} c_k x_1^{a_{1k}} x_2^{a_{2k}} ... x_n^{a_{nk}} = \sum_{k=1}^{K} c_k \mathbf{x}_{\mathbf{k}}^{\mathbf{a_k}} \tag{3.2.9}$$

με $c_k > 0, \forall k \in [K]$. Επομένως, η *τυπική μορφή* των προβλημάτων Γεωμετρικού Προγραμματισμού είναι:

$$\begin{array}{lll} \text{minimize} & g_0(x) & \\ \text{subject to} & f_i(x) = 1, & i = 1, ...., m \\ & g_i(x) \leq 1, & i = 1, ...., n \end{array} \tag{3.2.10}$$

όπου $f_i$ και $g_i$ είναι τα monomials και τα posynomials, αντίστοιχα. Ένα παράδειγμα Γεωμετρικού Προγραμματισμού παρουσιάζεται παρακάτω (όχι σε τυπική μορφή).

$$\begin{aligned} \text{minimize} \quad & x^{-1}y \\ \text{subject to} \quad & 2 \leq x \leq 3, \\ & x^2 + 3y/z \leq \sqrt{y}, \\ & x/y = z^2 \end{aligned}$$

Η τυπική μορφή δεν ορίζει απαραίτητα κυρτό πρόβλημα, καθώς έστω και ένας αρνητικός εκθέτης αρκεί ώστε το πρόβλημα να κατηγοριοποιηθεί ως μη κυρτό. Ωστόσο, η ιδιαίτερη μορφή των posynomials επιτρέπει τη μετατροπή του (3.2.10) σε αντίστοιχο κυρτό. Αυτό επιτυγχάνεται με αλλαγή των μεταβλητών, τόσο στην αντικειμενική συνάρτηση όσο και στις συνθήκες. Θέτουμε $y_i = \log x_i \implies x_i = e^{y_i}$. Εξετάζουμε την απλή περίπτωση του μονωνύμου (3.2.8). Επομένως:

$$\begin{aligned} f(\mathbf{x}) &= f(x_1, x_2, \ldots, x_n) \\ &= f(e^{y_1}, e^{y_2}, \ldots, e^{y_n}) \\ &= c(e^{y_1})^{a_1} \ldots (e^{y_n})^{a_n} \\ &= e^{\mathbf{a}^\top \mathbf{y} + b} \end{aligned}$$

όπου $b = \log c$. Η αλλαγή των μεταβλητών, επομένως, μετατρέπει το μονώνυμο σε εκθετικό αφφινικών συναρτήσεων. Παρομοίως, για το posynomial της εξίσωσης (3.2.9), έχουμε:

$$\begin{aligned} g(x) &= \sum_{k=1}^{K} c_k x_1^{a_{1k}} x_2^{a_{2k}} ... x_n^{a_{nk}} \\ &= \sum_{k=1}^{K} e^{\mathbf{a}_k^\top \mathbf{y} + b_k} \end{aligned}$$

Μετά την αλλαγή των μεταβλητών, το posynomial γίνεται το άθροισμα εκθετικών αφφινικών όρων. Μετατρέποντας την αντικειμενική συνάρτηση και τους περιορισμούς με τον παραπάνω τρόπο, καθώς και λαμβάνοντας το λογάριθμο, το πρόβλημα παίρνει την ακόλουθη μορφή:

$$\begin{aligned} \text{minimize} \quad & \tilde{g}_0(x) = \log\left(\textstyle\sum_{k=1}^{K_0} e^{\mathbf{a}_{0k}^\top \mathbf{y} + b_{0k}}\right) & & \\ \text{subject to} \quad & \tilde{f}_i(x) = e^{\alpha_i^\top \mathbf{y} + \beta_i} = 1, & & i = 1, ...., m \\ & \tilde{g}_i(x) = \log\left(\textstyle\sum_{k=1}^{K_i} e^{\mathbf{a}_{ik}^\top \mathbf{y} + b_{ik}}\right) \leq 1, & & i = 1, ...., n \end{aligned} \tag{3.2.11}$$

Εφόσον οι συναρτήσεις $\tilde{g}_i$ είναι κυρτές και οι $\tilde{f}_i$ είναι αφφινικές, το πρόβλημα (3.2.11) είναι κυρτό. Μάλιστα, αναφερόμαστε στην παραπάνω έκφραση ως *γεωμετρικό πρόβλημα σε κυρτή μορφή*. Αξίζει να επισημανθεί ότι εφόσον τόσο η αντικειμενική συνάρτηση όσο και οι περιορισμοί είναι μονώνυμα ($K_i = 1$ $i = 0, 1, \ldots, n$), το πρόβλημα κατόπιν αλλαγής μεταβλητών κατατάσσεται στο Γραμμικό Προγραμματισμό [BT97].

### 3.2.4 Generalized Geometric Programming

Η τυπική μορφή (3.2.10) είναι περιοριστική και δεν επιτρέπει την έκφραση πολλών προβλημάτων στη γλώσσα του γεωμετρικού προγραμματισμού. Στο πλαίσιο της τροπικής άλγεβρας, η χρήση της κυρτής συναρτησης $max$ έχει σπουδαία και προφανή σημασία. Επομένως, ο Γενικευμένος Γεωμετρικός Προγραμματισμός επεκτείνει τις παραπάνω ιδέες. Η πρώτη επέκταση αφορά τις κλασματικές δυνάμεις των posynomials. Για παράδειγμα, έστω $f_1, f_2$ posynomials. Τότε, ο περιορισμός

$$f_1(\mathbf{x})^a + f_2(\mathbf{x})^b \leq 1 \tag{3.2.12}$$

είναι έγκυρος μόνο για $a, b \in \mathbb{N}$, καθώς μόνο τότε το ανάπτυγμα της έκφρασης διατηρεί τις απαιτήσεις της *τυπικής μορφής* (3.2.10). Το παραπάνω πρόβλημα μπορεί να παρακαμφθεί, ωστόσο, αντικαθιστώντας τα posynomials με μεταβλητές $t_1, t_2$. Αναλυτικότερα, η σχέση (3.2.12) γίνεται:

$$f_1(\mathbf{x})^{3.2} + f_2(\mathbf{x})^{5.3} \leq 1 \implies \begin{cases} t_1^{3.2} + t_2^{5.3} \leq 1 \\ f_1(\mathbf{x}) \leq t_1 \\ f_2(\mathbf{x}) \leq t_2 \end{cases} \tag{3.2.13}$$

Με παρόμοιο τρόπο μετατρέπεται το maximum πολλών posynomials σε *τυπική μορφή*. Ακολουθεί ένα γενικό παράδειγμα μετατροπής ενός προβλήματος Γενικευμένου Γεωμετρικού Προγραμματισμού σε *τυπική μορφή*, ώστε να γίνει σαφής η ισοδυναμία τους. Έστω το πρόβλημα[LB16]:

$$\begin{array}{ll} \min & \max\{x + z, 1 + (y + z)^{1/2}\} \\ \text{s.t.} & \max\{y, z^2\} + \max\{yz, 0.3\} \leq 1 \\ & 3xy/z = 1 \end{array} \tag{3.2.14}$$

το οποίο δεν είναι πρόβλημα ΓΠ. Ωστόσο, με κατάλληλες αλλαγές, όπως (3.2.13), μετατρέπεται στο ισοδύναμο:

$$\begin{array}{lll} \min & t_1 & \\ \text{s.t.} & x + z \leq t_1, & 1 + t_2^{1/2} \leq t_1 \\ & y + z \leq t_2 & \\ & t_3 + t_4 \leq 1 & \\ & y \leq t_3, & z^2 \leq t_3 \\ & yz \leq t_4, & 0.3 \leq t_4 \\ & 3xy/z = 1 & \end{array} \tag{3.2.15}$$

## 3.3 Από τροπική σε κλασική βελτιστοποίηση

Ο κλάδος του optimization είναι ευρύς και αποτελεί αντικείμενο μελέτης πολλές δεκαετίες. Ιδιαίτερη πρόοδος έχει σημειωθεί στον τομέα της κυρτής βελτιστοποίησης [BV04], καθώς αποτελεί ένα εργαλείο με πληθώρα εφαρμογών. Τα τελευταία χρόνια αναπτύσσεται ενδιαφέρον στον κλάδο της τροπικής βελτιστοποίησης. Αρκετές τεχνικές του optimization δεν έχουν αντίστοιχο (ή δεν έχει βρεθεί ακόμη αντίστοιχο) στο τροπικό πλαίσιο. Ωστόσο, υπάρχει μία στενή σχέση ανάμεσα στην "κλασική" και στην τροπική βελτιστοποίηση και αυτή έγκειται ότι το τροπικό πολύτοπο αποτελεί μία συλλογή κυρτών πολυτόπων υπό την κλασική έννοια. Επομένως, δύναται η επίλυση προβλημάτων τροπικής βελτιστοποίησης μέσω της επίλυσης υποπροβλημάτων κλασικής βελτιστοποίησης. Τονίζεται ότι οι μέθοδοι που παρουσιάζονται δεν αποσκοπούν στην εύρεση της βέλτιστης λύσης αλλά στην παραλληλοποίηση της επίλυσης.

Προτού εντρυφήσουμε στο ορισμό και την ανάλυση των προβλημάτων, αξίζει να μελετήσουμε τα μαθηματικά εργαλεία που επιτρέπουν τη διάβαση από την τροπική στην κλασική βελτιστοποίηση.

### 3.3.1 Πολυεδρική Γεωμετρία

Στην ενότητα αυτή εισάγονται έννοιες πολυεδρικής γεωμετρίας, όπως ημίχωρος, πολύεδρο κ.α. Στις προηγούμενες ενότητες, παρουσιάστηκαν αλγεβρικά στοιχεία των τροπικών μαθηματικών. Η σύνδεσή τους με τα γεωμετρικά στοιχεία επιτρέπει την αποτελεσματική ανάλυση και σκιαγραφεί αλγοριθμικές προοπτικές. Η ενότητα βασίζεται σε [Zie95; DS04; CT16].

Η έννοια της κυρτότητας (βλ. ορισμούς 3.1.1 και 2.3.5) είναι κεντρική στην παρακάτω ανάλυση. Ωστόσο, αξίζει να επισημάνουμε ότι η έννοια της ευθείας στον τροπικό ημιδακτύλιο είναι διαφορετική από την καθιερωμένη, όπως είδαμε προηγουμένως. Κατά συνέπεια, τα τροπικά κυρτά σύνολα διαφέρουν από τα κυρτά σύνολα της γραμμικής άλγεβρας.

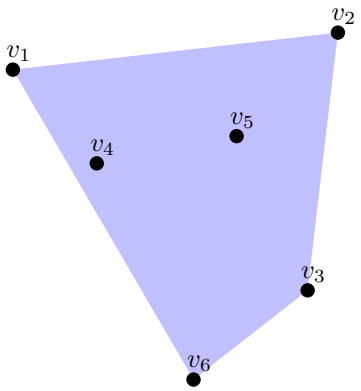
Αυτή η παρατήρηση μας οδηγεί στο κυρτό κάλυμμα ή convex hull, ένα σύνολο που προκύπτει από τον κυρτό συνδυασμό ενός διακριτού συνόλου σημείων.

**Definition 3.3.1: Convex Hull**

Έστω $S = \{\mathbf{x_1}, \mathbf{x_2} \cdots, \mathbf{x_n}\}$ ένα σύνολο σημείων στο $\mathbb{R}^d$. Το κυρτό κάλυμμα του $S$ ορίζεται ως το σύνολο:

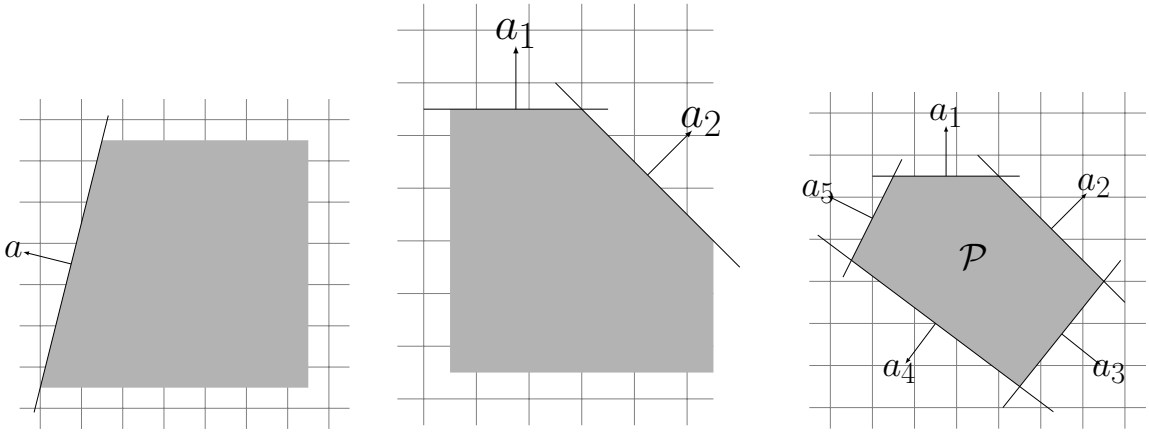
$$\text{conv}(S) = \left\{ \sum_{i=1}^{n} \lambda_i \mathbf{x}_i : \quad \lambda_i \geq 0, \sum_{i=1}^{n} \lambda_i = 1 \right\} \tag{3.3.1}$$

Στο σχήμα 3.3.1 παρατίθεται το κυρτό κάλυμμα ενός συνόλου σημείων στο $\mathbb{R}^2$, χρησιμοποιώντας κλασικά μαθηματικά. Τα σημεία εντός του convex hull μπορούν να παραλειφθούν, καθώς αποτελούν κυρτούς συνδυασμούς των σημείων-κορυφών.




Σχήμα 3.3.1: $\text{conv}(\{v_1, v_2, v_3, v_4, v_5, v_6\}) = \{v_1, v_2, v_3, v_6\}$

Στο σημείο αυτό, είμαστε σε θέση να εξετάσουμε τους θεμέλιους λίθους της αλγεβρικής γεωμετρίας, έννοιες διαμέρισης του χώρου και δημιουργίας συνόλων. Ξεκινάμε από την πιο απλή έννοια: τον ημίχωρο. Ο ημίχωρος ορίζει ένα σύνολο $S = \left\{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top \mathbf{x} \leq b\right\} \subset \mathbb{R}^n$. Το πολύεδρο αποτελεί την τομή ημιχώρων. Τέλος, αναφερόμαστε σε πολύτοπο, όταν το πολύεδρο είναι φραγμένο. Μία γραφική αναπαράσταση παρέχεται στο σχήμα 3.3.2. Με τον τρόπο αυτό εκφράζεται ένα πολύτοπο και ο συγκεκριμένος τρόπος έκφρασης ονομάζεται $\mathcal{H}$-πολύτοπο.




(a) Ημίχωρος (b) Πολύεδρο (c) Πολύτοπο

Σχήμα 3.3.2: Διαφορά μεταξύ ημιχώρου, πολυέδρου και πολυτόπου. Ένα πολύεδρο είναι η τομή πολλών ημιχώρων και ένα πολύτοπο είναι ένα φραγμένο πολύεδρο.

**Definition 3.3.2: $\mathcal{H}$-πολύτοπο**

Ένα $\mathcal{H}$-πολύτοπο είναι η τομή ενός πεπερασμένου πλήθους κλειστών ημιχώρων (halfspaces) : ένα φραγμένο σύνολο $P \subset \mathbb{R}^n$ της μορφής

$$P = P(\mathbf{A}, \mathbf{z}) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{z}\} \tag{3.3.2}$$

με $\mathbf{A} \in \mathbb{R}^{m \times d}$ και $\mathbf{z} \in \mathbb{R}^m$.

Τα περισσότερα προβλήματα βελτιστοποίησης κωδικοποιούν τους περιορισμούς στην παραπάνω μορφή, χρησιμοποιώντας ανισώσεις. Ένα παράδειγμα από Γραμμικό Προγραμματισμό είναι το εξής:

$$\begin{aligned} \text{minimize} \quad & 3x_1 + 2x_2 \\ \text{subject to} \quad & -x_1 + 3x_2 \leq 12 \\ & x_1 + 3x_2 \leq 12 \\ & 2x_1 + x_2 \leq 8 \\ & -x_1 - x_2 \leq 10 \\ & x_1, x_2 \geq 0 \end{aligned}$$

Είναι ξεκάθαρο ότι οι περιορισμοί του προβλήματος έχουν τη μορφή ενός $\mathcal{H}$-πολυτόπου (3.3.2). Ωστόσο, το παράδειγμα του σχήματος 3.3.1 σκιαγραφεί ένα διαφορετικό τρόπο έκφρασης ενός πολυτόπου, ο οποίος βασίζεται στις κορυφές. Αυτή η παρατήρηση εκφράζεται μαθηματικά με το παρακάτω θεώρημα:

**Definition 3.3.3: $\mathcal{V}$-πολύτοπο**

Έστω $S$ ένα πεπερασμένο σύνολο σημείων $\mathbf{x} \in \mathbb{R}^n$. Το κυρτό κάλυμμα (convex hull) του συνόλου $S$ λέγεται $\mathcal{V}$-πολύτοπο.

Ένα πολύτοπο μπορεί να εκφραστεί χρησιμοποιώντας είτε τις κορυφές είτε τους ημιχώρους. Οι δύο παραπάνω εκφράσεις είναι ισοδύναμες.

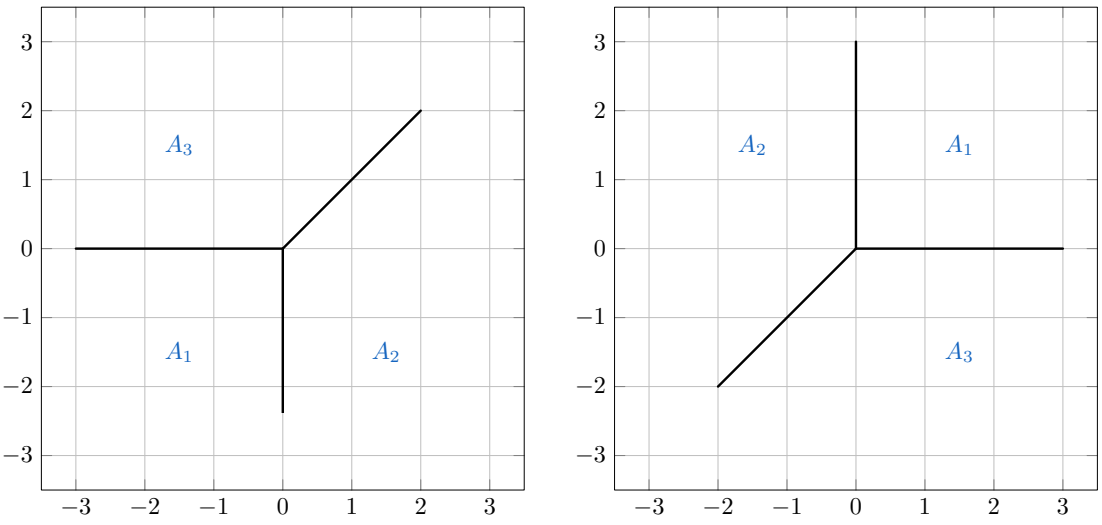
**Theorem 3.3.4: Κεντρικό θεώρημα πολυτόπων**

Ένα σύνολο $P \subseteq \mathbb{R}^d$ είναι το κυρτό κάλυμμα ενός πεπερασμένου συνόλου σημείων ($\mathcal{V}$-πολύτοπο) $P = \text{conv}(\mathbf{V})$ για κάποιο $\mathbf{V} \in \mathbb{R}^{d \times n}$ αν και μόνο αν υπάρχει μία φραγμένη τομή halfspaces ($\mathcal{H}$-πολύτοπο) όπου $P = P(\mathbf{A}, \mathbf{z})$ με $\mathbf{A} \in \mathbb{R}^{m \times d}$ και $\mathbf{z} \in \mathbb{R}^m$.

Ως αυτό το σημείο, οι έννοιες είναι ιδιαίτερα γενικές και βρίσκουν αντίκρισμα τόσο στο μονοειδές της τροπικής γεωμετρίας όσο και στην κλασική θεώρηση της γραμμικής άλγεβρας. Για την ανάλυση τροπικών προβλημάτων, ακολουθεί η εισαγωγή ορισμένων μαθηματικών εργαλείων της τροπικής γεωμετρίας.

Έστω $\mathcal{F}_{min}$ ένα fan στο $\mathbb{TP}^{n-1}$ που ορίζεται από την τροπική καμπύλη με $a_1 = \cdots = a_n = 0$. Με όμοιο τρόπο ορίζουμε το max-plus ανάλογο. Η σχηματική αναπαράσταση στο 3.3.3. Παρατηρούμε ότι το τροπικό fan $\mathcal{F}_{min}$ συμπίπτει με την τροπική υπερεπιφάνεια για το πολυώνυμο $\min\{x_1, x_2, \ldots, x_n, 0\}$. Αντίστοιχα για το max-plus ανάλογο.

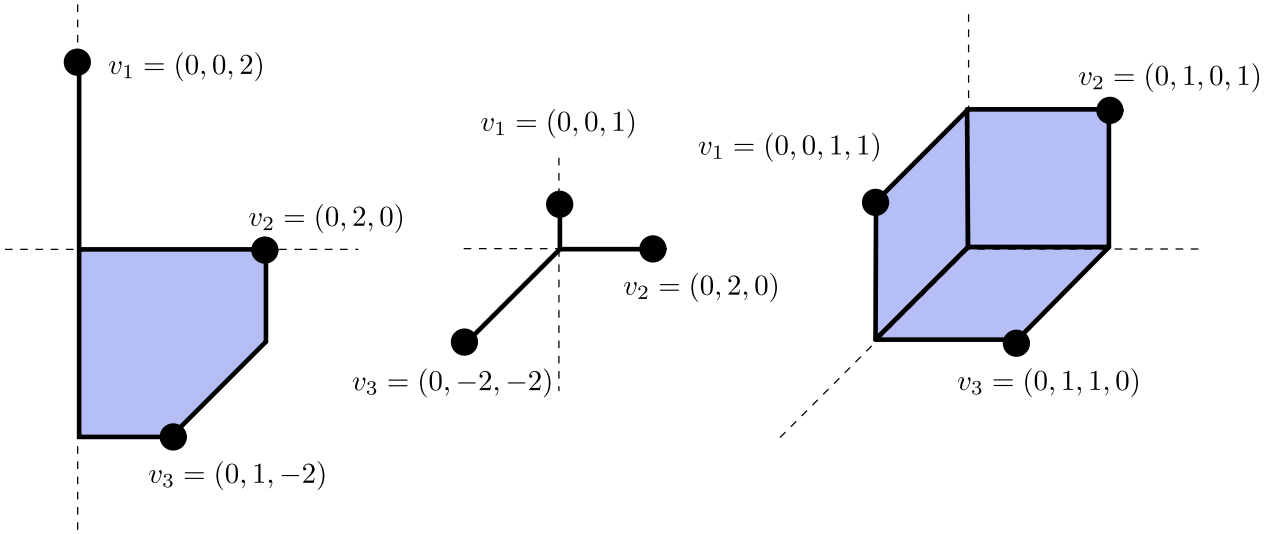
Έστω $V \subset \mathbb{R}^n$ ένα σύνολο σημείων. Τότε η τροπική κυρτή θήκη (tropical convex hull) είναι το ελάχιστο τροπικά κυρτό υποσύνολο του $\mathbb{R}^n$ που περιέχει το $V$. Η μορφή των παραπάνω συνόλων διαφέρει από τα αντίστοιχα γραμμικά. Στο σχήμα 3.3.4 βλέπουμε ορισμένα παραδείγματα που είναι κυρτά στο πλαίσιο των τροπικών μαθηματικών αλλά αποτελούν συλλογή (κλασικά) κυρτών συνόλων. Για τα σημεία έχουμε $v_u \in \mathbb{TP}^2$. Το αριστερό τροπικά κυρτό σύνολο διαιρείται εύκολα στο γαλάζιο πολύτοπο και στο ευθύγραμμο τμήμα που ορίζουν τα σημεία $(0,0,0)$ και $(0,0,2)$. Παρόμοια διαχωρίζονται και τα άλλα τροπικά κυρτά σύνολα. Επιπλέον, τα ευθύγραμμα τμήματα που περικλείουν τις γαλάζιες περιοχές, ή γενικότερα ορίζουν τα πολύτοπο, είναι είτε παράλληλα στους άξονες είτε στην ευθεία $y = x$. Εδώ εντοπίζεται η λειτουργία των fans. Μάλιστα, το μεσαίο




(a) $\mathcal{F}_{max}$ (b) $\mathcal{F}_{min}$

Σχήμα 3.3.3: Max-plus και Min-plus fans σε δύο διαστάσεις

πολύτοπο αντιστοιχεί στο $\mathcal{F}_{min}$. Ο τρόπος διάκρισης των αντικειμένων της συλλογής είναι μέσω του τύπου τους.




Σχήμα 3.3.4: Τροπικά κυρτά σύνολα. Σχήμα από [DS04]

**Definition 3.3.5: τύπος του $x$ σχετικά με $V$**

Έστω $x \in \mathbb{TP}^{n-1}$. Ο τύπος του $x$ σχετικά με $V$ είναι η διατεταγμένη $n$-πλειάδα (ordered $n$-tuple) $S_1, \ldots, S_n$ των υποσυνόλων $S_j \subseteq [r]$ που ορίζεται με τον ακόλουθο τρόπο: ο δείκτης $i \in S_j$ αν:

$$v_{ij} - x_j = \bigwedge_{k=1}^{n} v_{ik} - x_k$$

Το σύνολο όλων των σημείων των οποίων ο τύπος περιέχει το $S$ είναι το $X_S = \{\mathbf{x} \in \mathbb{TP}^{n-1} : S \subseteq \text{type}(\mathbf{x})\}$.

Επισημαίνεται ότι το σύνολο $S$ είναι ένα κλειστό κυρτό πολύεδρο υπό την κλασική έννοια. Αναλυτικότερα:

$$X_S = \{\mathbf{x} \in \mathbb{TP}^{n-1} : x_k - x_j \leq v_{ik} - v_{ij} \quad \forall j,k \in [n] \text{ s.t. } i \in S_j\} \tag{3.3.3}$$

Αυτή η παρατήρηση είναι ιδιαίτερα σημαντική, καθώς αποτελεί τη γέφυρα από την τροπική στην κλασική βελτιστοποίηση. Διακρίνοντας το τροπικό πολύτοπο στους διάφορους τύπους του, γίνεται η κωδικοποίηση των συνθηκών με τρόπο τέτοιο ώστε κλασικοί αλγόριθμοι κυρτής βελτιστοποίησης να μπορούν να εφαρμοστούν. Η διαίρεση αυτή ονομάζεται αποσύνθεση σε κελιά ή cell decomposition[1]:

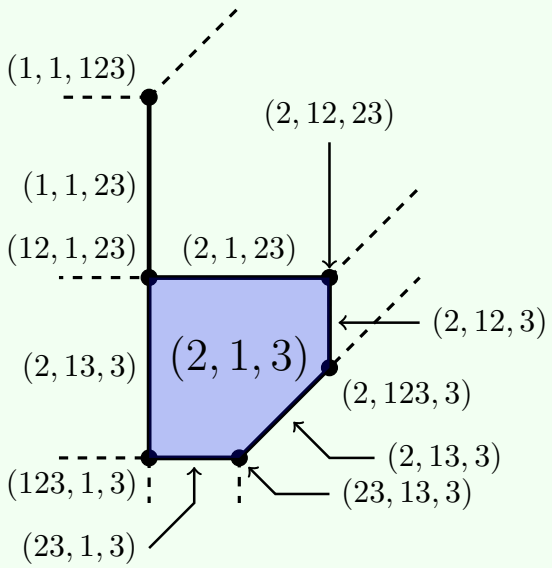
**Theorem 3.3.6: cell decomposition**

Η συλλογή των κυρτών πολυέδρων $X_S$, με $S$ να παίρνει τις τιμές όλων των τύπων, ορίζει ένα cell decomposition $\mathcal{C}_V$ του $\mathbb{TP}^{n-1}$. Το τροπικό πολύτοπο $\mathcal{P} = \text{tconv}(V)$ ισούται με την ένωση όλων των φραγμένων κελιών $X_S$ σε αυτή την αποδόμηση.

Τα fans χωρίζουν το $\mathbb{TP}^{n-1}$ σε υποχώρους. Γεωμετρικά, λοιπόν, το cell decomposition ενός συνόλου $V \subset \mathbb{TP}^{n-1}$ πραγματοποείται μέσω της τοποθέτησης max-plus fans μετατοπισμένων ώστε τα κέντρα τους να συμπίπτουν με τα σημεία $v_i \in V$. Το τροπικό πολύτοπο αντιστοιχεί στην ένωση όλων των φραγμένων τομών που ορίζουν τα εν λόγω fans.

**Example 3.3.7: types και cell decomposition**

Έστω $r = n = 3$ και $V = \{\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3}\}$ με $\mathbf{v_1} = (0,0,2), \mathbf{v_2} = (0,2,0)$ και $\mathbf{v_3} = (0,1,-2)$. Τότε το cell decomposition $\mathcal{C}_V$ φαίνεται στο σχήμα 3.3.5. Παρατηρούμε ότι το γκρι κελί έχει τύπο $(2,1,3)$. Πώς προκύπτει αυτό; Η max-plus καμπύλη 3.3.3a διασπά το $\mathbb{R}^2$ σε $A_1 \cup A_2 \cup A_3$. Το γαλάζιο κελί είναι στο halfspace $A_1$ σχετικά με το $\mathbf{v_2}$, στο $A_2$ σχετικά με το $\mathbf{v_1}$ και στο $A_3$ σχετικά με το $\mathbf{v_3}$.

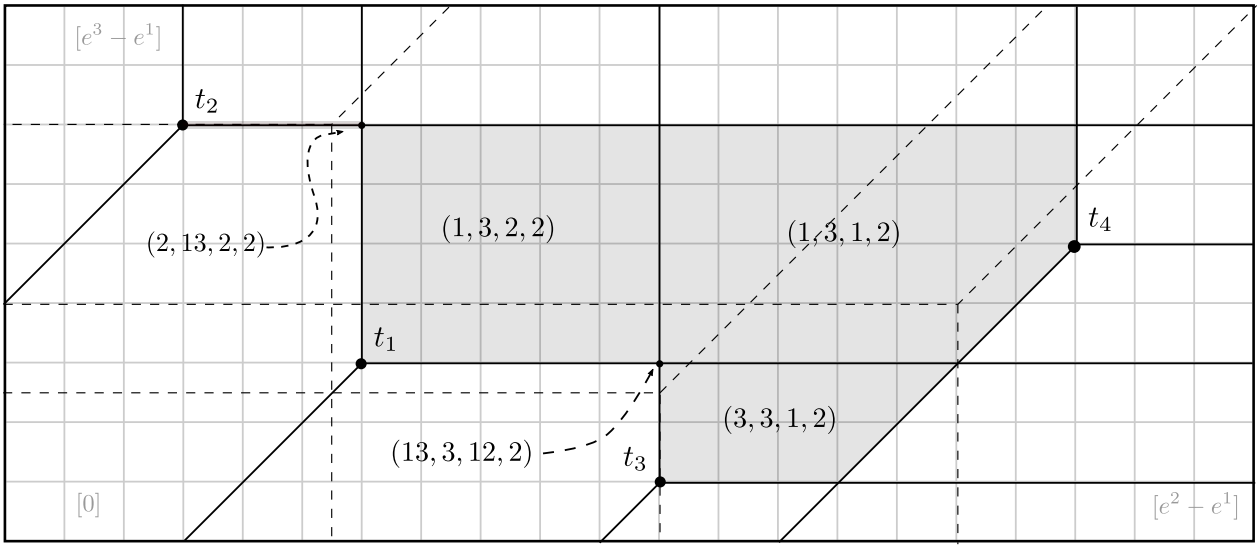



Σχήμα 3.3.5: cell decomposition. Σχήμα από [DS04]

### 3.3.2 tropical linear problems

Στις προηγούμενες ενότητες, παρουσιάστηκαν τα εργαλεία για τη μετατροπή των προβλημάτων κυρτής βελτιστοποίησης σε κλασική μορφή. Στην περίπτωσή μας, έχουμε ένα τροπικό πολύτοπο $\mathcal{P}$. Σύμφωνα με το θεώρημα 6 το πολύτοπο $\mathcal{P}$ μπορεί να εκφραστεί ως η ένωση ενός πεπερασμένου αριθμού, έστω $M$, κυρτών

[1] Χρησιμοποιούμε την αγγλική ορολογία.




Σχήμα 3.3.6: Min-plus arrangement των σημείων $t_1, t_2, t_3, t_4$.

υπό την κλασική έννοια πολύτοπων:

$$\mathcal{P} = \bigcup_{m=1}^{M} P_m \tag{3.3.4}$$

Επομένως το πρόβλημα (3.0.1) γίνεται

$$\min_{m \in [M]} z_m \text{ s.t. } z_m = \min\{f(x) | x \in P\}$$
$$\mathcal{P} = \bigcup_{m=1}^{M} P_m \tag{3.3.5}$$

Η ανάλυση αρχίζει από τα πιο απλά προβλήματα, τα τροπικά γραμμικά προβλήματα. Έστω το πρόβλημα ελαχιστοποίησης:

$$\text{minimize } f(x) + g(x) \tag{3.3.6}$$

όπου $f, g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ είναι κλειστές κυρτές και μπορούν να είναι nonsmooth. Τότε, σύμφωνα με τη μέθοδο Alternating direction method of multipliers (ADMM) [Par14; The15] το πρόβλημα επιλύεται:

$$x_{k+1} = \text{prox}_{\lambda f}(z_k - u_k) \tag{3.3.7}$$
$$z_{k+1} = \text{prox}_{\lambda g}(x_{k+1} + u_k) \tag{3.3.8}$$
$$u_{k+1} = u_k + x_{k+1} - z_{k+1} \tag{3.3.9}$$

Συνήθως χρησιμοποείται η συνάρτηση $f$ ως objective function και η $g$ για κωδικοποίηση των constraints. Έχοντας αποδομήσει το τροπικό πολύτοπο σε μία συλλογή κλασικών πολυτόπων χρησιμοποιούμε τη μέθοδο ADMM για την επίλυση των επιμέρους προβλημάτων. Η "έκδοση" (variant) με τον proximal operator

$$\text{prox}_{\lambda f}(v) = \text{argmin}_x \left\{ f(x) + \frac{1}{2\lambda} \|x - v\|_2^2 \right\} \tag{3.3.10}$$

εγγυάται strong convexity και ταχύτερη σύγκλιση [CP11].

### 3.3.3 tropical fractional problems

Το πρόβλημα έχει μελετηθεί από τους Gaubert, Katz, and Sergeev στο [GKS12]. Η αντικειμενική συνάρτηση λαμβάνει τη μορφή τροπικής διαίρεσης (τροπικών) πολυωνύμων, η οποία έγκειται στην "κλασική" αφαίρεση.

Έστω, λοιπόν, $f, g : \mathbb{R}^n \to \mathbb{R}$ δύο τροπικά πολυώνυμα. Τότε η αντικειμενική συνάρτηση τροπικά κλασματικού προγραμματισμού έχει τη μορφή:

$$h(\mathbf{x}) = f(\mathbf{x}) \oslash g(\mathbf{x}) = f(\mathbf{x}) - g(\mathbf{x}) \tag{3.3.11}$$

Γνωρίζουμε, επιπλέον, ότι τα τροπικά πολυώνυμα έχουν κυρτή φύση. Επομένως, αντιμετωπίζουμε ένα πρόβλημα διαφοράς κυρτών συναρτήσεων [Har59; She+16] (βλ. Difference of Convex (DC) programming) που έχει την παρακάτω μορφή:

$$\begin{aligned} \text{minimize} \quad & \underbrace{\left(\bigvee_{i=1}^{n} x_i + a_i\right) \vee a_{n+1}}_{f(\mathbf{x})} - \underbrace{\left(\bigvee_{i=1}^{n} x_i + b_i\right) \vee b_{n+1}}_{g(\mathbf{x})} \\ \text{subject to} \quad & (\mathbf{A} \boxplus \mathbf{x}) \vee \mathbf{c} \leq (\mathbf{B} \boxplus \mathbf{x}) \vee \mathbf{d} \end{aligned}$$

### 3.3.4 tropical constraint problems

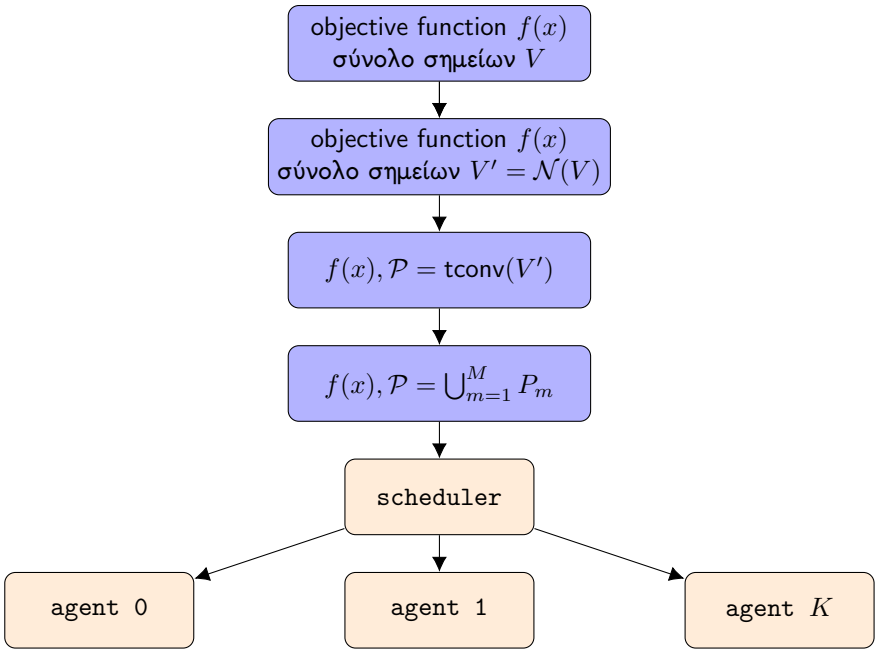
Η αποδόμηση που περιγράφεται στ προηγούμενο section και χρησιμοποείται στις παραπάνω παραγράφους μπορεί να επεκταθεί σε προβλήματα που δεν είναι αμιγώς τροπικά (ή max-plus) αλλά έχουν σύνολο περιορισμών ένα τροπικό πολύεδρο. Τέτοιου είδους προβλήματα περιγράφουν ένα χώρο που ορίζεται από τροπικές ανισότητες και στοχεύουν στη μεγιστοποίηση μίας αυθαίρετης αντικειμενικής συνάρτησης. Εφόσον η αντικειμενική συνάρτηση είναι κυρτή ή μπορεί να εκφραστεί ως διαφορά κυρτών συναρτήσεων, το πρόβλημα μπορεί να επιλυθεί αποτελεσματικά με τεχνικές παρόμοιες με τις παραπάνω παραγράφους.

## 3.4 Δρομολόγηση εργασιών

Το τροπικό πρόβλημα βελτιστοποίησης έχει διαχωριστεί σε $M$ προβλήματα με κυρτό υπό την κλασική έννοια σύνολο περιορισμών. Έστω, λοιπόν, ότι έχουμε $K$ υπολογιστές-agents και θέλουμε να ολοκληρωθεί η διαδικασία όσο πιο σύντομα γίνεται. Υποθέτουμε ότι οι agents έχουν ίση υπολογιστική δυνατότητα και ότι κάθε υπο-πρόβλημα αντιστοιχεί σε μία τιμή δυσκολίας $d_i$ με $i \in [M]$. Για παράδειγμα, η δυσκολία αυτή μπορεί να έγκειται στο πλήθος των ανισοτήτων. Επομένως, η δρομολόγηση των υπο-προβλημάτων στους $K$ υπολογιστές μπορεί να μοντελοποιηθεί ως εξής:

$$\begin{aligned} \min \quad & \max_{k \in K} \sum_{i \in S_k} d_i \\ \text{s.t.} \quad & \bigsqcup_{k=1}^{K} S_k = [M] \\ & S_k \neq \emptyset \quad \forall k \in [K] \end{aligned} \tag{3.4.1}$$

όπου $S_k \subseteq [M]$ είναι το σύνολο των υποπροβλημάτων που ανατίθενται στον agent $k$. Πρόκειται, λοιπόν, για το equal-sum subset πρόβλημα που ανήκει στην κλάση πολυπλοκότητας `NP-hard` αλλά επιδέχεται αλγόριθμο ψευδοπολυωνυμικού χρόνου για $K = O(1)$ [Cie+08]. Ο αλγόριθμος παρουσιάζεται σχηματικά στο 3.4.1.




Σχήμα 3.4.1: Διάγραμμα ροής της διαδικασίας

# Κεφάλαιο 4

# Νευρωνικά Δίκτυα

Ο όρος *νευρωνικά δίκτυα* εμπεριέχει μία συλλογή υπολογιστικών μοντέλων με πύρηνα το νευρώνα. Παρόλο που τα νευρωνικά δίκτυα επινοήθηκαν το 1958 με το perceptron του Rosenblatt, η δημοφιλία τους εξελίχθηκε την περασμένη δεκαετία, όπου η πρόοδος στις δυνατότητες των υπολογιστών επέτρεψε την ευρεία και αποτελεσματική εφαρμογή τους σε πληθώρα προβλημάτων.

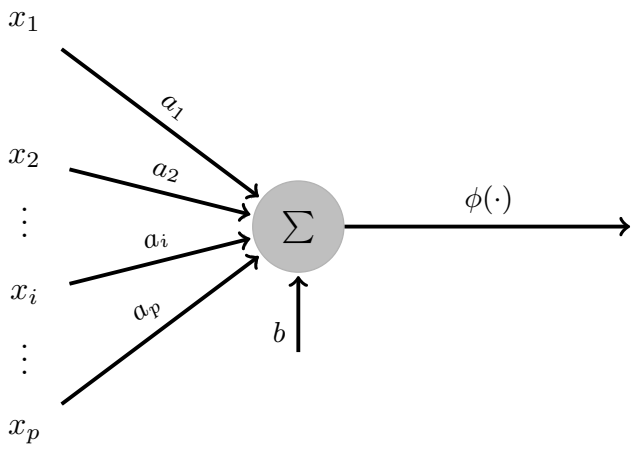
Τα νευρωνικά δίκτυα επιλύουν προβλήματα προσέγγισης συναρτήσεων. Η χρήση τους προσανατολίζεται σε προβλήματα μη-γραμμικά και μη-κυρτής φύσεως. Η ιδέα ξεκινά με τους McCulloch and Pitts [MP43] το 1943 και βασίζεται στον ανθρώπινο εγκέφαλο, ο οποίος συνιστά ένα σύνθετο μη-γραμμικό και παράλληλο υπολογιστή. Προσλαμβάνει πληροφορίες, τις επεξεργάζεται και οδηγείται σε συμπεράσματα/αποφάσεις. Συγκεκριμένα, ο άνθρωπος λαμβάνει ερεθίσματα από το περιβάλλον του, τα οποία συλλαμβάνουν οι υποδοχείς του οργανισμού. Τα σήματα μετατρέπονται σε ηλεκτρικά σήματα, γεγονός που καθιστά εφικτή την επεξεργασία τους. Ακολουθεί η ανάλυση της πληροφορίας.

Μελετητές της βιολογίας έχουν καταλήξει στο συμπέρασμα ότι η εν λόγω επεξεργασία των σημάτων είναι μη-γραμμική. Τα νευρωνικά δίκτυα, λοιπόν, κατασκευάστηκαν με στόχο τη μίμηση αυτής της συμπεριφοράς. Στο κεφάλαιο αυτό πραγματοποείται μία σύντομη εισαγωγή στα δημοφιλή αυτά μοντέλα. Καλύπτεται η δομή τους, οι συναρτήσεις ενεργοποίησης και κόστους. Αναλύεται ο τρόπος που τα νευρωνικά δίκτυα "μαθαίνουν" και συμπεριλαμβάνεται μία σύντομη αναφορά στους αλγορίθμους βελτιστοποίησης της διαδικασίας εκπαίδευσής τους.

## 4.1 Δομή νευρωνικού δικτύου

Ο νευρώνας αποτελεί τον πυρήνα του νευρωνικού δικτύου. Πρόκειται για μία υπολογιστική μονάδα που δέχεται ένα διάνυσμα εισόδου $\mathbf{x} \in \mathbb{R}^n$, το επεξεργάζεται και παράγει μία είσοδο $y \in \mathbb{R}$. Ο νευρώνας παρουσιάζεται στο σχήμα 4.1.1. Αναλυτικότερα:

- Ο νευρώνας χαρακτηρίζεται από ένα διάνυσμα βαρών σύνδεσης $\mathbf{a} \in \mathbb{R}^n$ και το bias $b \in \mathbb{R}$.
- Ο νευρώνας υπολογίζει το σταθμισμένο άθροισμα $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b = \sum_{i=1}^{n} a_i x_i + b$. Τα βάρη σύνδεσης προσδίδουν διαφορετική "σημασία" σε κάθε χαρακτηριστικό εισόδου $x_i, i \in [n]$.
- Τέλος, το σταθμισμένο άθροισμα διέρχεται από ένα φίλτρο, τη λεγόμενη συνάρτηση ενεργοποίησης $\phi(\cdot)$.




Σχήμα 4.1.1: Perceptron

Το μοντέλο αυτό επινοήθηκε από τον Rosenblatt [Ros58] το 1958 και λειτουργεί σωστά, δηλαδή ταξινομεί τις κλάσεις χωρίς λάθη, στην περίπτωση που τα πρότυπα είναι γραμμικά διαχωρίσιμα. Αναλυτικότερα, η εξίσωση $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b = 0$ που ορίζει η έξοδος του perceptron, αντιστοιχεί σε ένα υπερεπίπεδο στο $\mathbb{R}^n$ που δρα ως

διαχωριστική επιφάνεια απόφασης ανάμεσα στις δύο κλάσεις εισόδου $\mathcal{C}_0, \mathcal{C}_1$. Εφόσον οι κλάσεις είναι γραμμικά διαχωρίσιμες, η εκπαίδευση παράγει ένα διάνυσμα βαρών $\mathbf{a}$ και όρο πόλωσης (bias) $b$, τέτοιο ώστε:

$$\begin{cases} \mathbf{a}^\top \mathbf{x} + b \geq 0 & \mathbf{x} \in \mathcal{C}_0 \\ \mathbf{a}^\top \mathbf{x} + b < 0 & \mathbf{x} \in \mathcal{C}_1 \end{cases}$$

Ως συνάρτηση ενεργοποίησης $\phi(\cdot)$ (βλ. σχήμα 4.1.1), χρησιμοποείται η συνάρτηση προσήμου. Σε κάθε εποχή, ο αλγόριθμος εκπαίδευσης εξετάζει όλα τα πρότυπα εισόδου και ενημερώνει τα βάρη προσθαφαιρώντας το πρότυπο εισόδου $\mathbf{x}$, σταθμισμένο με το ρυθμό μάθησης $\eta$, ώστε να τροποποιήσει το σύνορο απόφασης $\mathbf{a}^\top \mathbf{x} + b = 0$ σε περίπτωση λανθασμένης ταξινόμησης. Ο κανόνας ενημέρωσης, καθώς και ο αλγόριθμος παρουσιάζονται παρακάτω σε μορφή ψευδοκώδικα:

**Algorithm 3:** Perceptron Training

**Data:** number of epochs $\mathcal{E}$, dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \{-1, 1\}\}_{i=1}^m$, learning rate $\eta$
**Init:** random weight vector $w = \begin{bmatrix} \mathbf{a} & b \end{bmatrix}$

**do**
  pick random $\mathbf{x} \in \mathcal{D}$ ▷ all patterns are selected once at each epoch
  compute perceptron's output $d \leftarrow \text{sgn}(f(\mathbf{x}))$
  update weights $\mathbf{w} \leftarrow \mathbf{w} + \eta(y - d)\mathbf{x}$
**until** *all patterns are classified correctly or number of epochs is reached*
**if** *number of epochs is reached* **then**
  **return** FALSE
**else**
  **return** $\mathbf{w} = \begin{bmatrix} \mathbf{a} & b \end{bmatrix}$
**end**

Επιλέγοντας κατάλληλο αριθμό εποχών $\mathcal{E}$ και ρυθμό μάθησης $\eta$, το perceptron παράγει ένα σωστό σύνορο απόφασης για γραμμικά διαχωρίσιμα πρότυπα. Ωστόσο, μία τόσο απλή κατανομή, με πρότυπα πάνω και κάτω μίας νοητής γραμμής, δεν παρατηρείται σε πραγματικά προβλήματα. Αντίθετα, χαλαρώνεται η επιταγή του μηδενικού λάθους και στοχεύουμε στην ελαχιστοποίησή του.

Για την εκμάθηση σύνθετων κατανομών, επεκτείνουμε το μοντέλο του απλού perceptron σε πολλά επίπεδα διατάσσοντας στο καθένα πολυάριθμα perceptrons. Οι νευρώνες που λαμβάνουν το ίδιο διάνυσμα εισόδου βρίσκονται στο ίδιο *επίπεδο*. Κατά συνέπεια, κάθε επίπεδο με $n$-διάστατο διάνυσμα εισόδου και $m$ εξόδους ορίζει μία απεικόνιση $\phi : \mathbb{R}^n \to \mathbb{R}^m$. Το διάνυσμα εξόδου $\phi(\mathbf{x})$ αποτελεί είσοδο στο επόμενο επίπεδο και η διαδικασία επαναλαμβάνεται έως το επίπεδο εξόδου που, συνήθως, χαρακτηρίζεται από μία εξοδο $y \in \mathbb{R}$. Τα ενδιάμεσα επίπεδα, δηλαδή όλα εκτός της εισόδου και της εξόδου, λέγονται *κρυφά*. Ένα γενικό νευρωνικό δίκτυο έχει τη μορφή του σχήματος 4.1.2.

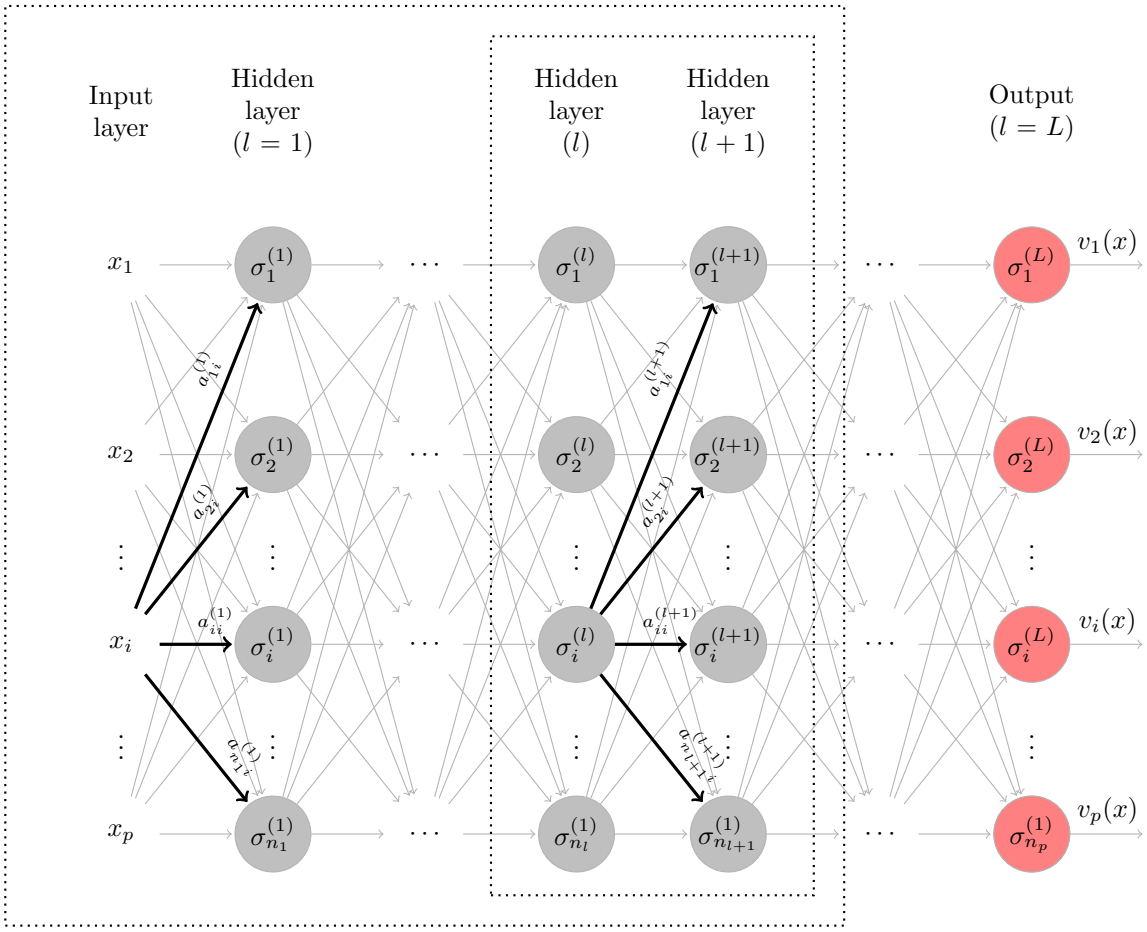
## 4.2 Συναρτήσεις Ενεργοποίησης

Οι συναρτήσεις ενεργοποίησης μιμούνται τη μη-γραμμική απόκριση των νευρώνων του ανθρώπινου οργανισμού. Πολλές διαφορετικές συναρτήσεις έχουν προταθεί στη βιβλιογραφία και η επιλογή τους εξαρτάται σε μεγάλο βαθμό από την εφαρμογή. Ορισμένες γνωστές συναρτήσεις παρουσιάζονται παρακάτω και στο σχήμα 4.2.1.

Η μη γραμμική συμπεριφορά των συναρτήσεων ενεργοποίησης επιτρέπει την εκμάθηση σύνθετων περιοχών απόφασης. Για το λόγο αυτό, τα νευρωνικά δίκτυα χρησιμοποιούνται ευρέως για πληθώρα εφαρμογών. Ωστόσο, η επιθυμητή αυτή ιδιότητα αποτελεί τροχοπέδη στη λεπτομερή ανάλυση και κατανόηση της λειτουργίας και των αποτελεσμάτων των νευρωνικών δικτύων, καθώς εγείρει προβλήματα μη-κυρτής φύσεως.

Στην υποενότητα αυτή μελετούμε τις πιο διαδεδομένες συναρτήσεις ενεργοποίησης και συνδέουμε την τροπική γεωμετρία με ορισμένες. Αυτή η σύνδεση εξηγεί σε μεγάλο βαθμό το αυξάνον ενδιαφέρον της κοινότητας της μηχανικής μάθησης/τεχνητής νοημοσύνης στα μαθηματικά μοντέλα της τροπικής γεωμετρίας.

Παρουσιάζουμε τις διάφορες συναρτήσεις και ακολουθεί σύντομος σχολιασμός.




Σχήμα 4.1.2: Feed-forward Νευρωνικό Δίκτυο

**συνάρτηση κατωφλίου**

$$\phi(x) = \begin{cases} 1 & x \geq \theta \\ 0 & x < \theta \end{cases} \tag{4.2.1}$$

Η συνάρτηση κατωφλίου αποτελεί την εναρκτήρια συνάρτηση ενεργοποίησης στη βιβλιογραφία. Αντιστοιχεί στη διέγερση του νευρώνα που πρότεινε ο Rosenblatt στο ομώνυμο perceptron. Αποσκοπεί στην ταξινόμηση προτύπων από δύο κατηγορίες (binary classification).

**συνάρτηση προσήμου**

$$\phi(x) = \text{ sign}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \tag{4.2.2}$$

Η συνάρτηση προσήμου αποτελεί επέκταση της συνάρτησης κατωφλίου από το πεδίο τιμών $[0, 1]$ στο $[-1, 1]$. Είναι φανερό ότι η συμπεριφορά της είναι παρόμοια.

**σιγμοειδής συνάρτηση**

$$\phi(x) = \sigma_T(x) = \frac{1}{1 + e^{-x/T}} \tag{4.2.3}$$

Αποτελεί μία ομαλή (smooth) προσέγγιση της συνάρτησης κατωφλίου. Ιστορικά, η σιγμοειδής χρησιμοποιούταν κατά κόρον τις προηγούμενες δεκαετίες για προβλήματα δυαδικής ταξινόμησης. Αυτό οφείλεται στο γεγονός ότι προσδίδει πιθανότητες $\phi(x)$ στο πρότυπο 1 και $1 - \phi(x)$ στο πρότυπο 0. Ωστόσο, ο υπολογισμός του εκθετικού και η floating point διαίρεση επιβαρύνουν υπολογιστικά την εκπαίδευση των δικτύων. Συνεπώς, στην εποχή των βαθιών αρχιτεκτονικών (deep learning), η χρήση της έχει περιοριστεί. Αξίζει να σημειωθεί ότι η επέκτασή της (softmax) που αναθέτει πιθανότητες σε προβλήματα κατηγοριοποίησης πολλών προτύπων χρησιμοποείται ως το τελευταίο layer στα νευρωνικά δίκτυα σε προβλήματα ταξινόμησης πολλών κλάσεων (multi-class classification).

**υπερβολική εφαπτομένη**

$$\phi(x) = \alpha \cdot \tanh(\beta x) = \alpha \cdot \frac{e^{\beta x} - e^{-\beta x}}{e^{\beta x} + e^{-\beta x}} \tag{4.2.4}$$

Η υπερβολική εφαπτομένη αποτελεί μία γενίκευση της σιγμοειδούς συνάρτησης με πεδίο τιμών το $[-\alpha, \alpha]$. Συνήθως, $\alpha = 1$.

**Rectified Linear Unit (ReLU)**

$$\phi(x) = \text{ReLU}(x) = \max(0, x) \tag{4.2.5}$$

Η συνάρτηση Rectified Linear Unit ή αλλιώς ReLU [NH] αποτελεί την πλέον διαδεδομένη επιλογή για συνάρτηση ενεργοποίησης για Βαθιά Νευρωνικά Δίκτυα λόγω της απλότητάς της που οδηγεί σε υψηλές ταχύτητες υπολογισμού. Η χρήση του τελεστή max αναδεικνύει άμεσα τη σχέση της ReLU με την τροπική γεωμετρία. Οι Zhang, Naitzat, and Lim εξερεύνησαν αυτή τη σχέση και έδειξαν ότι ένα πρόσθιο νευρωνικό δίκτυο (feedforward neural network) με ReLU ενεργοποιήσεις ορίζει ένα τροπικό κλάσμα (τροπικών) πολυωνύμων, όπως στη σχέση βλ. (3.3.11) [ZNL18].

**Leaky Rectified Linear Unit (LReLU)**

$$\phi(x) = \text{LReLU}(x) = \max(\alpha x, x) \tag{4.2.6}$$

με $a \in (0, 1)$. Η ReLU δεν κωδικοποιεί πληροφορίες για $x < 0$ και η παραλλαγή της, η LReLU, μετριάζει αυτή τη συμπεριφορά, προσδίδοντας υψηλότερη σημασία στις θετικές τιμές αλλά διατηρώντας την πληροφορία των αρνητικών τιμών. Συνήθης τιμή είναι $\alpha = 0.01$[1].

[1] PyTorch documentation για Leaky ReLU

**Softplus**

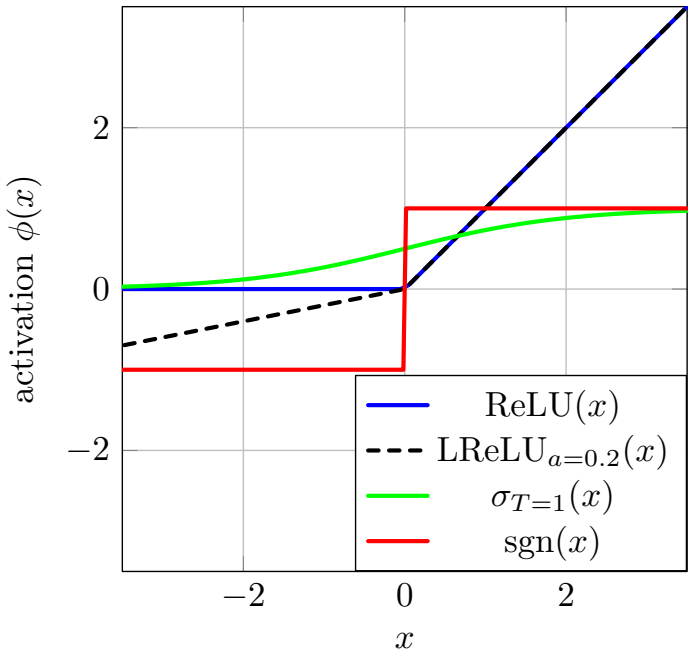
$$\phi(x) = \text{Softplus}(x) = \frac{1}{\beta} \cdot \log(1 + \exp(\beta \cdot x)) \tag{4.2.7}$$

Η συνάρτηση Softplus είναι μία ομαλή προσέγγιση της ReLU και μπορεί να χρησιμοποιηθεί με σκοπό την επιβολή θετικών τιμών στα βάρη. Επιπλέον, η σχέση της με την Τροπική Γεωμετρία είναι στενή [Vir01; Luc10], μέσω της αποκβαντοποίησης κατά Maslov (Maslov Dequantization). Σε αντίθεση με τη ReLU, η Softplus είναι παραγωγίσιμη και δεν εμπλέκονται υποπαράγωγοι κατά την εκπαίδευση νευρωνικών που τη χρησιμοποιούν (περισσότερες λεπτομέρειες στην υποενότητα 4.4).

**Maxout**

$$f(\mathbf{x}) = \bigvee_{i=1}^{k} \{W_{ij}\mathbf{x} + b_{ij}\} = \bigvee_{i=1}^{k} \left[\sum_{m=1}^{n} W_{ijm}x_m\right] + b_{ij} \tag{4.2.8}$$

όπου $\mathbf{W}_{j,:}$ είναι η $j-$οστή σειρά του πίνακα βαρών $\mathbf{W}$. Συνεπώς, η μονάδα Maxout αντιστοιχεί σε ένα τροπικό πολυώνυμο. Τα δίκτυα Maxout προτάθηκαν από Goodfellow [Goo+13a]. Ένας νευρώνας maxout μπορεί να ερμηνευτεί ως μία τμηματική γραμμική προσέγγιση μίας αυθαίρετης κυρτής συνάρτησης.




Σχήμα 4.2.1: Συναρτήσεις ενεργοποίησης

## 4.3 Συναρτήσεις κόστους

Προς το παρόν, η συζήτηση έχει περιοριστεί στο εμπρόσθιο πέρασμα του νευρωνικού δικτύου, δηλαδή πως το διάνυσμα εισόδου $\mathbf{x} \in \mathbb{R}^n$ διέρχεται μέσα από το νευρωνικό δίκτυο και παράγει την είσοδο $y \in \mathbb{R}$. Ωστόσο, η παραπάνω διεργασία δεν εξηγεί τη διαδικασία της *μάθησης*, δηλαδή πως το νευρωνικό δίκτυο εξελίσσεται με κάθε δείγμα.

Οι συναρτήσεις κόστους καθορίζουν αυτή την εξέλιξη μέσω του αλγορίθμου *backpropagation*. Με πολύ απλά λόγια, το κόστος αντιστοιχεί στο σφάλμα και ανάλογα το μέγεθός του, τα βάρη των συνάψεων ενημερώνονται ώστε να συμπεριλάβουν την πληροφορία του τελευταίου δείγματος. Η φύση του προβλήματος καθορίζει και την επιλογή της μετρικής κόστους. Με άλλα λόγια, διαφορετικές συναρτήσεις χρησιμοποιούνται σε προβλήματα ταξινόμησης (classification) και διαφορετικές για προβλήματα παλινδρόμησης (regression).

Έστω $\mathbf{y} \in \mathbb{R}^n$ το διάνυσμα που περιέχει τις πραγματικές τιμές και $\hat{\mathbf{y}} \in \mathbb{R}^n$ το διάνυσμα προβλέψεων που προκύπτει ως έξοδος του νευρωνικού δικτύου (για $n$ δείγματα). Εξετάζουμε πρώτα τις συναρτήσεις κόστους στο πλαίσιο των προβλημάτων παλινδρόμησης.

**Mean Squared Error (MSE)**

$$\mathrm{MSE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2 \tag{4.3.1}$$

Η μετρική κόστους με τη γηραιότερη ιστορία αντιστοιχεί στο μεσοτετραγωνικό λάθος καθώς έχει μελετηθεί από την εποχή του Gauss στο πλαίσιο της Γραμμικής Άλγεβρας. Η γραμμική παλινδρόμηση (linear regression) είναι ένα από τα πιο απλά αλλά, συγχρόνως, πιο διαδεδομένα μοντέλα και επιλύει το πρόβλημα της ελαχιστοποίησης του μεσοτετραγωνικού λάθους. Στο πλαίσιο της μηχανικής μάθησης, έχει παρατηρηθεί ότι το μεσοτετραγωνικό σφάλμα είναι ευαίσθητο σε outliers (δείγματα που απέχουν αρκετές διασπορές από τη μέση τιμή της κατανομής).

**Mean Absolute Error (MAE)**

$$\mathrm{MAE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n}|y - \hat{y}| \tag{4.3.2}$$

Το μέσο απόλυτο σφάλμα δεν αντιμετωπίζει προβλήματα με outliers. Έχει πιο εύρωστη συμπεριφορά και η ελαχιστοποίηση αυτής της μετρικής οδηγεί στην εύρεση της διαμέσου της κατανομής.

Στρέφουμε την προσοχή μας σε συναρτήσεις κόστους λάθους για ταξινόμηση. Έστω $\mathbf{x} \in \mathbb{R}^n$ το διάνυσμα εισόδου, δηλαδή το πρότυπο υπό εξέταση και $y$ η πραγματική τιμή (label) του προτύπου. Για προβλήματα δυαδικής ταξινόμησης (binary classification), ισχύει $m = 2$. Εξετάζουμε τις ακόλουθες συναρτήσεις κόστους σε σχέση με ένα δείγμα. Για προβλήματα δυαδικής ταξινόμησης σε κλάσεις $\mathcal{C}_0, \mathcal{C}_1$ θεωρούμε:

$$y = \begin{cases} 1 & \mathbf{x} \in \mathcal{C}_1 \\ -1 & \mathbf{x} \in \mathcal{C}_0 \end{cases}$$

**misclassification error (0/1-loss)**

$$\ell_{0/1}(y, \mathbf{x}) = \begin{cases} 1 & y \cdot f(\mathbf{x}) < 0 \\ 0 & \text{αλλιώς} \end{cases} \tag{4.3.3}$$

Το 0/1-σφάλμα, λοιπόν, αντιστοιχεί σε μία δείκτρια συνάρτηση $\mathbb{1}[y \neq \hat{y}]$. Από την απλή της έκφραση και από το σχήμα 4.3.1, συμπεραίνουμε ότι αυτό το σφάλμα είναι μη-συνεχές και μη-κυρτό. Κατά συνέπεια, είναι αδύνατη η εφαρμογή μεθόδων με υποπαραγώγους (subgradients) και η αποτελεσματική βελτιστοποίηση προβλημάτων με αυτό το κριτήριο. Οι επόμενες μετρικές κόστους αποτελούν κυρτές προσεγγίσεις του πραγματικού λάθους και διευκολύνουν την εκπαίδευση του δικτύου στο πλαίσιο του online learning [Sha+11]. Αναλυτικότερα, η χρήση κυρτών συναρτήσεων προσδίδει στους αλγορίθμους ευστάθεια, μειώνοντας τις ταλαντώσεις.

**Hinge Loss**

$$\ell_{\text{hinge}}(y, \mathbf{x}) = \max\{0, 1 - y \cdot f(\mathbf{x})\} \tag{4.3.4}$$

Σκοπεύει στη δημιουργία ενός συνόρου απόφαση (decision boundary) με μέγιστο περιθώριο σφάλματος. Χρησιμοποείται στις Μηχανές Διανυσμάτων Υποστήριξης ή Support Vector Machines (SVMs).

**Cross-Entropy Loss**

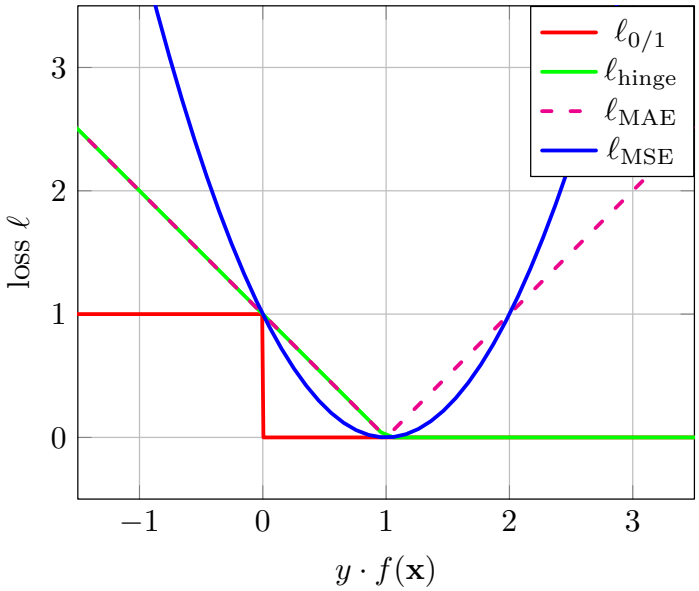
Χρησιμοποείται τόσο για δυαδική ταξινόμηση όσο και για την περίπτωση πολλών κλάσεων. Έστω $p, q : \mathcal{X} \to [0, 1]$ δύο συναρτήσεις μάζας πιθανότητας. Τότε, η διασταυρωμένη εντροπία, ή αλλιώς cross-entropy, ορίζεται ως:

$$H(p, q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x)$$

Στο πλαίσιο της μηχανικής μάθησης, ωστόσο, δύναται η απλοποίηση της παραπάνω έκφρασης. Η έξοδος του δικτύου πρέπει να είναι ένα διάνυσμα $\hat{\mathbf{y}} = f(\mathbf{x})$ έτσι ώστε $f : \mathbb{R}^n \to \Delta_m \subset [0,1]^m$, όπου $\Delta_m$ το probability simplex σε $m$ διαστάσεις. Με άλλα λόγια, ανατίθεται μία πιθανότητα το πρότυπο να ανήκει σε καθεμία από τις $m$ πιθανές κλάσεις. Αυτό επιτυγάνεται με τη χρήση softmax ως συνάρτηση ενεργοποίησης στο επίπεδο εξόδου. Τέλος, το διάνυσμα $\mathbf{y}$ με τα πραγματικά labels είναι ένα one-hot vector, δηλαδή έχει άσσο στη σωστή κλάση και μηδενικά στις υπόλοιπες. Έστω ότι η σωστή κλάση είναι η $i \in [m]$. Τότε:

$$\begin{aligned} H(\mathbf{y}, \hat{\mathbf{y}}) &= -\mathbf{y} \log \hat{\mathbf{y}} \\ &= -\sum_{k=1}^{m} y_k \log \hat{y_k} \\ &= -y_i \log \hat{y_i} = \log \hat{y_i} \end{aligned}$$

Αξίζει να σημειωθεί ότι η διασταυρωμένη εντροπία ταυτίζεται με την απόσταση Kullback-Leibler, καθώς το διάνυσμα πραγματικών labels $\mathbf{y}$ είναι one-hot: $H(\mathbf{y}, \hat{\mathbf{y}}) = D(\mathbf{y}||\hat{\mathbf{y}})$. Το γεγονός αυτό προσδίδει διαίσθηση σε αυτή τη μετρική, καθώς η απόσταση Kullback-Leibler ποσοτικοποιεί την απόσταση δύο κατανομών.




Σχήμα 4.3.1: Συναρτήσεις Κόστους

## 4.4 Εκπαίδευση Νευρωνικού Δικτύου Πολλών Στρωμάτων

Στην ενότητα 4.1 παρουσιάστηκε ο αλγόριθμος εκπαίδευσης του perceptron του Rosenblatt. Ωστόσο, ο αλγόριθμος αυτός δεν είναι εφαρμόσιμος σε πολυεπίπεδα νευρωνικά δίκτυα. Αρκεί να παρατηρήσει κανείς ότι με εξαίρεση το νευρώνα εξόδου, δεν υπάρχει κάποιος ξεκάθαρος "στόχος" για την έξοδο. Οι κρυφοί νευρώνες αποτελούν ενδιάμεσα στάδια στο μαύρο κουτί που ονομάζεται νευρωνικό δίκτυο και, κατά συνέπεια, δεν ορίζεται σωστή έξοδος (target output). Αυτό σημαίνει ότι για τα βάρη των κρυφών νευρώνων, ο κανόνας ενημέρωσής τους εξαρτάται από τους γειτονικούς νευρώνες. Ο αλγόριθμος που επιλύει αυτό το πρόβλημα λέγεται *backpropagation*, δηλαδή backward propagation of errors. Ο ελληνικός όρος είναι *οπισθοδιάδοση*.

Προτού εντρυφήσουμε στη μαθηματική διατύπωση των εξισώσεων που διέπουν την οπισθοδιάδοση, αξίζει να ξεκαθαριστεί ο συμβολισμός. Αναλυτικότερα, συμβολίζουμε με $w_{ij}^{(k)}$ το βάρος του κόμβου $j$ του επιπέδου $k$ για τον κόμβο $i$ (του προηγούμενου επιπέδου). Χάριν ευκολίας, παραλείπουμε τον όρο πόλωσης $b_i^{(k)}$ ενσωματώνοντας τον στο διάνυσμα βαρών ως το μηδενικό όρο $w_{i0}^{(k)} = b_i^{(k)}$. Θεωρούμε ότι το επίπεδο $k$ αποτελείται από $n_k$ νευρώνες και η έξοδος του νευρώνα $i$ είναι $o_i^k$. Συμβολίζουμε με $\phi$ και $\phi'$ τη συνάρτηση ενεργοποίησης και την παράγωγό της, αντίστοιχα. Χρησιμοποείται ο δείκτης $o$ για τη συνάρτηση ενεργοποίησης

του επιπέδου εξόδου, $\phi_o$, καθώς στην πράξη διαφέρει μαθηματικά από τις υπόλοιπες. Τέλος, με $a_i^k$ συμβολίζεται το σταθμισμένο άθροισμα για τον κόμβο $i$ του επιπέδου $k$:

$$a_i^{(k)} = b_i^{(k)} + \sum_{j=1}^{n_k} w_{ij}^{(k)} o_j^{(k-1)} = \sum_{j=0}^{n_k} w_{ij}^{(k)} o_j^{(k-1)} \tag{4.4.1}$$

Σκοπός του αλγορίθμου είναι η ενημέρωση των βαρών $w_{ij}^{(k)}$ ώστε να ελαχιστοποιηθεί μία μετρική κόστους $J$. Η ανάλυση του αλγορίθμου backpropagation συνδέεται ιστορικά με το μεσοτετραγωνικό σφάλμα (4.3.1) λόγω της απλής παραγώγου του. Στην επακόλουθη ανάλυση χρησιμοποείται το ίδιο σφάλμα $J = \frac{1}{2}(y - \hat{y})^2$. Επιθυμούμε, λοιπόν, να υπολογίσουμε τη συμβολή του βάρους $w_{ij}^{(k)}$ στο σφάλμα $\frac{\partial J}{\partial w_{ij}^{(k)}}$, χάριν ευκολίας. Σύμφωνα με τον κανόνα της αλυσίδας:

$$\frac{\partial J}{\partial w_{ij}^{(k)}} = \frac{\partial J}{\partial a_j^{(k)}} \cdot \frac{\partial a_j^{(k)}}{\partial w_{ij}^{(k)}} \tag{4.4.2}$$

$$\triangleq \delta_j^{(k)} \cdot \frac{\partial a_i^{(k)}}{\partial w_{ij}^{(k)}} \tag{4.4.3}$$

όπου με $\delta_j^{(k)}$ συμβολίζουμε τον όρο σφάλματος του κόμβου $j$ στο επίπεδο $k$. Εξετάζουμε το δεύτερο παράγοντα του κανόνα αλυσίδας:

$$\frac{\partial a_i^{(k)}}{\partial w_{ij}^{(k)}} = \frac{\partial}{\partial w_{ij}^{(k)}} \left( \sum_{l=0}^{n_{k-1}} w_{lj}^{(k)} o_l^{(k-1)} \right) = o_i^{(k-1)} \tag{4.4.4}$$

Αντικαθιστώντας στην εξίσωση (4.4.3) έχουμε:

$$\frac{\partial J}{\partial w_{ij}^{(k)}} = \delta_j^{(k)} o_i^{(k-1)} \tag{4.4.5}$$

Συνεπώς, η μερική παράγωγος για το βάρος είναι γινόμενο του όρου διόρθωσης $\delta_j^{(k)}$ για τον κόμβο $j$ του επιπέδου $k$ και της εξόδου του κόμβου $i$ του επιπέδου $k-1$. Διαισθητικά, αυτό είναι λογικό καθώς το βάρος $w_{ij}^{(k)}$ συνδέει τους παραπάνω κόμβους του δικτύου. Η παραπάνω ανάλυση είναι ανεξάρτηση των συναρτήσεων ενεργοποίησης. Για το επίπεδο εξόδου και για συνάρτηση σφάλματος τη μεσοτετραγωνική, υπολογίζουμε επακριβώς την παράγωγο. Για τα κρυφά επίπεδα, δείχνουμε ότι ο όρος σφάλματος εξαρτάται από τους αντίστοιχους όρους του επόμενου επιπέδου. Από την ιδιότητα αυτή προκύπτει και το όνομα του αλγορίθμου backpropagation.

Η έξοδος $\hat{y}$ του δικτύου είναι ίση με $\hat{y} = \phi_o(a_1^{(m)})$, καθώς έχουμε μόνο μία έξοδο στο πρόβλημα παλινδρόμησης. Συνεπώς:

$$\begin{aligned} J = \frac{1}{2}(\hat{y} - y)^2 &= \frac{1}{2}(\phi_o(a_1^{(m)}) - y)^2 \\ \implies \delta_1^{(m)} &= \frac{\partial J}{\partial a_i^{(k)}} \\ &= \frac{\partial}{\partial a_i^{(k)}} \left( \frac{1}{2}(\phi_o(a_1^{(m)}) - y)^2 \right) \\ &= (\phi_o(a_1^{(m)}) - y) \cdot \phi_o'(a_1^{(m)}) \\ &= (\hat{y} - y) \cdot \phi_o'(a_1^{(m)}) \end{aligned} \tag{4.4.6}$$

Συνδυάζοντας τις παραπάνω εξισώσεις, καταλήγουμε στην εξής σχέση για το βάρος $w_{i1}^m$ του επιπέδου εξόδου:

$$\frac{\partial J}{\partial w_{i1}^{(m)}} = \delta_1^{(m)} o_i^{(m-1)} = (\hat{y} - y) \cdot \phi_o'(a_1^{(m)}) \cdot o_i^{(m-1)} \tag{4.4.7}$$

Για τους κόμβους των κρυφών επιπέδων, ο όρος σφάλματος $\delta_j^{(k)}$, $1 \leq k < m$, λαμβάνει την ακόλουθη μορφή:

$$\delta_j^{(k)} = \frac{\partial J}{\partial a_j^{(k)}} = \sum_{l=1}^{r_{k+1}} \frac{\partial J}{\partial a_l^{(k+1)}} \cdot \frac{\partial a_l^{(k+1)}}{\partial a_j^{(k)}} \tag{4.4.8}$$

$$= \sum_{l=1}^{r_{k+1}} \delta_l^{(k+1)} \cdot \frac{\partial a_l^{(k+1)}}{\partial a_j^{(k)}} \tag{4.4.9}$$

Από τον ορισμό του όρου $a_l^{(k+1)}$ (4.4.1) και από το γεγονός ότι $o_j^{(k)} = \phi(a_j^{(k)})$ έχουμε:

$$a_l^{(k+1)} = \sum_{j=1}^{r_k} w_{jl}^{(k+1)} o_j^{(k)} = \sum_{j=1}^{r_k} w_{jl}^{(k+1)} \phi(a_j^{(k)}) \tag{4.4.10}$$

$$\implies \frac{\partial a_l^{(k+1)}}{\partial a_j^{(k)}} = w_{jl}^{(k+1)} \phi'(a_j^{(k)}) \tag{4.4.11}$$

Αντικαθιστώντας στην εξίσωση (4.4.9) λαμβάνουμε τη σχέση οπισθοδιάδοσης (*backpropagation formula*), που συνδέει τους όρους σφάλματος ενός επιπέδου με τους αντίστοιχους του επόμενου:

$$\delta_j^{(k)} = \sum_{l=1}^{r_{k+1}} \delta_l^{(k+1)} \cdot w_{jl}^{(k+1)} \cdot \phi'(a_j^{(k)}) = \phi'(a_j^{(k)}) \sum_{l=1}^{r_{k+1}} \delta_l^{(k+1)} \cdot w_{jl}^{(k+1)} \tag{4.4.12}$$

Καταλήγουμε, λοιπόν, στη μερική παράγωγο των βαρών κρυφών νευρώνων:

$$\frac{\partial J}{\partial w_{ij}^{(k)}} = \delta_j^{(k)} o_i^{(k-1)} = \phi'(a_j^{(k)}) o_i^{(k-1)} \sum_{l=1}^{r_{k+1}} \delta_l^{(k+1)} \cdot w_{jl}^{(k+1)} \tag{4.4.13}$$

Για πολλές συναρτήσεις ενεργοποίησης, η παράγωγος $\phi'(x)$ έχει κλειστό τύπο. Ένας από τους λόγους που η σιγμοειδής συνάρτηση (4.2.3) $\sigma_T(x) = \frac{1}{1+e^{-x/T}}$ επικράτησε πριν την εποχή του deep learning είναι (εν μέρει) και λόγω απλής έκφρασης της παραγώγου της. Για $T = 1$:

$$\begin{aligned} \sigma'(x) &= \frac{\partial \sigma(x)}{\partial x} = \frac{\partial}{\partial x}\left(\frac{1}{1+e^{-x}}\right) \\ &= -\frac{1}{(1+e^{-x})^2} \cdot e^{-x} \cdot (-1) = \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \\ &= \sigma(x) \cdot (1 - \sigma(x)) \end{aligned} \tag{4.4.14}$$

Επομένως, με απλές πράξεις η παράγωγος υπολογίζεται εύκολα εφόσον έχει διατηρηθεί στη μνήμη η έξοδος του perceptron $\sigma(x)$. Για παράδειγμα, για πρόβλημα παλινδρόμησης με μεσοτετραγωνικό σφάλμα, η μερική παράγωγος του σφάλματος για βάρος του νευρώνα εξόδου είναι:

$$(4.4.7) \overset{(4.4.14)}{\implies} \frac{\partial J}{\partial w_{i1}^{(m)}} = (\hat{y} - y) \cdot \sigma(a_1^{(m)}) \cdot (1 - \sigma(a_1^{(m)})) \cdot o_i^{(m-1)}$$

Πλέον, ωστόσο, έχουν επικρατήσει συναρτήσεις που βασίζονται σε μη-γραμμικές συναρτήσεις όπως max. Τέτοιοι τελεστές αφαιρούν την παραγωγισιμότητα από ορισμένα σημεία του πεδίου ορισμού διατηρώντας, ωστόσο,

τη συνέχεια. Χαρακτηριστικά παραδείγματα αποτελούν οι συναρτήσεις της οικογένειας Rectified Linear Units: ReLU (4.2.5) και Leaky ReLU (4.2.6). Στις περιπτώσεις αυτές,χρειάζεται μία πιο ισχυρή έννοια παραγωγισιμότητας για τις ανάγκες του αλγορίθμου backpropagation:

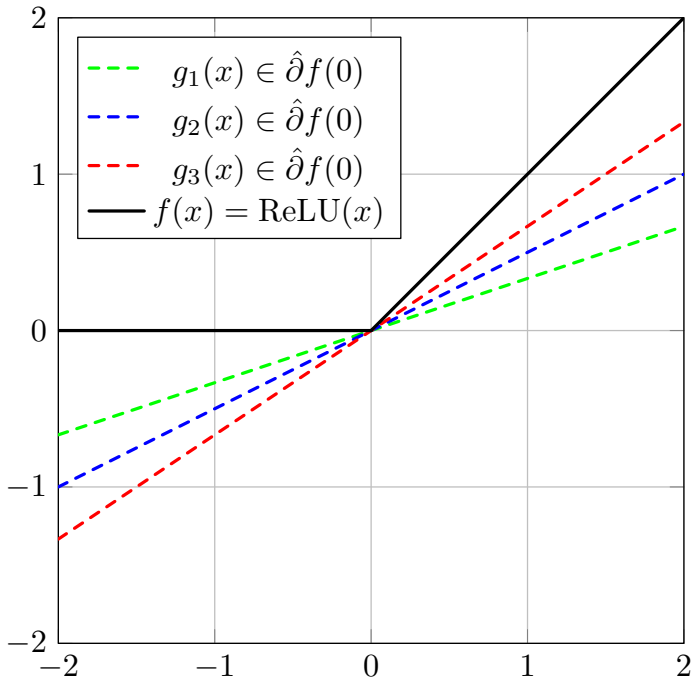
**Definition 4.4.1: Subgradient**

Ένα διάνυσμα $\mathbf{g} \in \mathbb{R}^n$ είναι *υποπαράγωγος* (subgradient) της συνάρτησης $f : \mathbb{R}^n \to \mathbb{R}$ στο σημείο $\mathbf{x} \in \mathbf{dom}\, f$ αν για κάθε $\mathbf{z} \in \mathbf{dom}\, f$ ισχύει:

$$f(\mathbf{z}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{z} - \mathbf{x}) \tag{4.4.15}$$

Αν, επιπλέον, η συνάρτηση $f$ είναι κυρτή, η παράγωγός της στο $\mathbf{x}$ είναι εξ ορισμού υποπαράγωγος. Συμβολίζουμε με $\partial f(\mathbf{x})$ το σύνολο των υποπαραγώγων της $f$ στο σημείο $\mathbf{x}$. Αν $\partial f(\mathbf{x}) \neq \emptyset$, η συνάρτηση λέγεται *υποπαραγωγίσιμη* στο $\mathbf{x}$. Αν η συνάρτηση $f$ είναι υποπαραγωγίσιμη σε όλο το πεδίο ορισμού $\mathbf{dom}\, f$ αποκαλείται υποπαραγωγίσιμη.

Γεωμετρικά, ένα διάνυσμα $\mathbf{g}$ είναι υποπαράγωγος της $f$ στο $\mathbf{x}$, $\mathbf{g} \in \partial f(\mathbf{x})$, αν η αφφινική συνάρτηση $f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{z} - \mathbf{x})$ είναι υποεκτιμητής της $f$ σε όλο το πεδίο ορισμού (global underestimator). Η υποπαράγωγος αποτελεί μία γενίκευση της έννοιας της παραγώγου και χρησιμοποείται σε περιπτώσεις που η συνάρτηση υπό εξέταση δεν είναι παραγωγίσιμη σε υποσύνολο ή σε όλο το πεδίο ορισμού. Στην εικόνα 4.4.1 παρουσιάζεται σχηματικά η έννοια της υποπαραγώγου για τη συνάρτηση ενεργοποίησης ReLU. Στην περίπτωση αυτή, το 0 είναι το μοναδικό σημείο που δεν ορίζεται παράγωγος.




Σχήμα 4.4.1: Sugradient της ReLU συνάρτησης

## 4.5 Αλγόριθμοι Βελτιστοποίησης Κατάβασης Κλίσεων

Στην προηγούμενη ενότητα εξερευνήσαμε τον αλγόριθμο οπισθοδιάδοσης. Αν και έγινε λόγος για τον υπολογισμό των παραγώγων, δεν αναφερθήκαμε με ποιον τρόπο χρησιμοποιούνται αυτές οι ποσότητες στη μάθηση κρυφών αναπαραστάσεων. Σε αυτή την ενότητα, λοιπόν, αναλύεται ο κανόνας ενημέρωσης των βαρών, τόσο σε μαθηματικό όσο και σε διαισθητικό επίπεδο.

Στα πειράματα που ακολουθούν στο κεφάλαιο 6, χρησιμοποιούνται διάφοροι αλγόριθμοι εκπαίδευσης, ή optimizers σύμφωνα με τη βιβλιογραφία. Αξίζει, λοιπόν, να αναφερθούμε στις διαφορές τους. Ο αλγόριθμος της κατάβασης κλίσεων αποτελεί την κατεξοχήν επιλογή για βελτιστοποίηση/εκπαίδευση νευρωνικών δικτύων. Χρησιμοποείται σε μη-κυρτά προβλήματα χωρίς περιορισμούς. Στόχος μας είναι να ελαχιστοποιήσουμε την αναμενόμενη τιμή του σφάλματος:

$$J(\mathbf{w}) = \mathbb{E}_{\mathbf{x},y\sim\mathcal{D}}\ell(\mathbf{w};\mathbf{x},y) \tag{4.5.1}$$

όπου το διάνυσμα $\mathbf{w}$ αντιστοιχεί στις παραμέτρους του μοντέλου.

### 4.5.1 Gradient Descent

Ο αλγόριθμος της κατάβασης κλίσεων, ή η οικογένεια των αλγορίθμων, βασίζονται σε μία απλή ιδέα. Έστω ότι τη χρονική στιγμή $\tau$ το μοντέλο μας έχει τις παραμέτρους $\mathbf{w}^{(\tau)}$. Υπολογίζοντας την παράγωγο της συνάρτησης κόστους $J$ στο σημείο $\mathbf{w}^{(\tau)}$, βρίσκουμε την κατεύθυνση που οδηγεί με βέλτιστο βήμα στη μεγιστοποίηση της εν λόγω μετρικής κόστους. Ωστόσο, σκοπός μας είναι η ελαχιστοποίησή της και, συνεπώς, επιλέγουμε να κινηθούμε στην αντίθετη κατεύθυνση. Άρα , ο κανόνας αναβάθμισης των παραμέτρων είναι:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta\nabla J(\mathbf{w}^{(\tau)}) \tag{4.5.2}$$

όπου με $\eta$ συμβολίζουμε την υπερπαράμετρο του ρυθμού εκμάθησης (learning rate). Από την κατεύθυνση που επιλέγεται, συμπεραίνουμε ότι υπάρχει μία μικρή γειτονιά με κέντρο το $\mathbf{w}^{(\tau)}$ που η κίνηση προς $-\nabla J(\mathbf{w}^{(\tau)})$ θα οδηγήσει σε μείωση της μετρικής κόστους. Έστω $\mathbf{p}_\tau = -\nabla J(\mathbf{w}^{(\tau)})$ η κατεύθυνση καθόδου. Τότε, από το Θεώρημα Μέσης Τιμής (ΘΜΤ) υπάρχει $\mu \in (0,1)$ τέτοιο ώστε $J(\mathbf{w}^{(\tau)} + \alpha_\tau\mathbf{p}_\tau) = J(\mathbf{w}^{(\tau)}) + \alpha_\tau\nabla J(\mathbf{w}^{(\tau)} + \mu\alpha_\tau\mathbf{p}_\tau)^\top\mathbf{p}_\tau$. Για $\alpha_\tau$ αρκετά μικρό και εφόσον $J \in C^1(\mathbb{R}^n)$ έχουμε ότι $J(\mathbf{w}^{(\tau)}+\alpha_\tau\mathbf{p}_\tau) < J(\mathbf{w}^{(\tau)})$. Για απογυγή παρανόησης, το Θεώρημα Μέσης Τιμής εγγυάται ότι υπάρχει βήμα $\alpha_\tau$ που να οδηγεί σε μείωση της αντικειμενικής συνάρτησης $J$. Αντίθετα, ο προκαθορισμένος ρυθμός μάθησης $\eta$ δεν αποτελεί τη βέλτιστη επιλογή και μπορεί να οδηγήσει και σε αύξηση της $J$.

Με μία προσεκτική ματιά του παραπάνω κανόνα συμπεραίνουμε ότι αυτός ο αλγόριθμος δεν είναι αποδοτικός, καθώς απαιτεί τον υπολογισμό της συνάρτησης σφάλματος για όλα τα δείγματα. Ο αριθμός των δειγμάτων ενδέχεται να είναι στις δεκάδες χιλιάδες, καθιστώντας τον απλό αλγόριθμο κατάβασης κλίσεων απαγορευτικό στην πράξη.

### 4.5.2 Stochastic Gradient Descent

Η στοχαστική έκδοση του αλγορίθμου αντιμετωπίζει την παραπάνω συμπεριφορά (μη-υπολογισιμότητα) αναβαθμίζοντας τις παραμέτρους *για κάθε δείγμα του συνόλου εκπαίδευσης*

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta\nabla J(\mathbf{w}^{(\tau)};\mathbf{x}_i,y_i) \tag{4.5.3}$$

όπου το $(\mathbf{x}_i,y_i)$ δειγματολειπτείται τυχαία από την κατανομή $\mathcal{D}$. Καθώς οι παράμετροι αναβαθμίζονται σε κάθε δείγμα, ενδέχεται το εν λόγω δείγμα να μην είναι αντιπροσωπευτικό της συνολικής κατανομής και να μην επιλεχθεί κατάλληλη κατεύθυνση καθόδου. Παρατηρείται, λοιπόν, ιδιαίτερη διακύμανση στις τιμές της αντικειμενικής συνάρτησης. Ωστόσο, αυτή η διακύμανση επιτρέπει τη μεταπήδηση από ένα τοπικό ελάχιστο σε καλύτερη περιοχή παραμέτρων [Rud17].

Στην πράξη, χρησιμοποείται μία ενδιάμεση παραλλαγή που λαμβάνει υπόψιν ένα υποσύνολο (batch) του συνόλου εκπαίδευσης για κάθε βήμα αναβάθμισης. Ο εν λόγω αλγόριθμος ονομάζεται Mini-Batch Stochastic Gradient Descent, αλλά στην πράξη χρησιμοποείται η ορολογία Stochastic Gradient Descent (SGD). Στη συνέχεια, ακολουθούμε αυτή τη σύμβαση.

### 4.5.3 Momentum

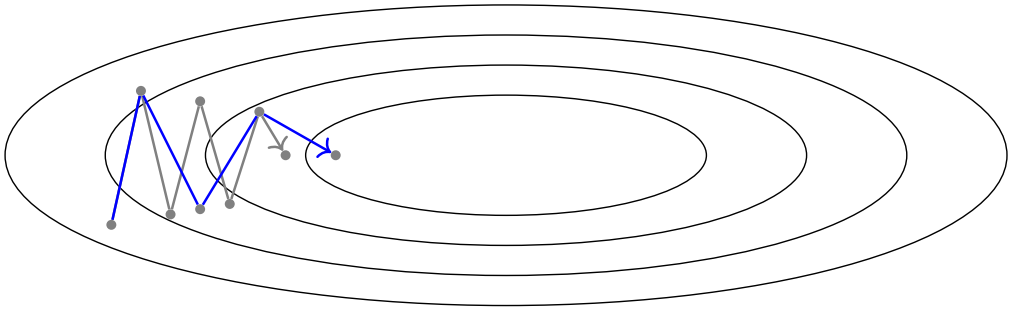
Η στοχαστική κατάβαση κλίσης (SGD) δεν αντιμετωπίζει αποτελεσματικά την πλοήγηση σε περιοχές όπου η επιφάνεια έχει διαφορετικές κλίσεις σε κάθε κατεύθυνση. Τέτοιες περιοχές είναι συχνές γύρω από τοπικά ελάχιστα και η Στοχαστική Κατάβαση Κλίσεων ταλαντώνεται γύρω από σημεία επιτυγχάνοντας αργή πρόοδο προς το τοπικό ελάχιστο.

Ο αλγόριθμος Momentum προσπαθεί να μετριάσει αυτή την ανεπιθύμητη συμπεριφορά προσδίδοντας στον κανόνα ενημέρωσης βαρών το στοιχείο της ορμής. Αναλυτικότερα, ο όρος διόρθωσης εμπεριέχει τον όρο διόρθωσης του προηγούμενου βήματος κλιμακωμένο κατά $\gamma \in (0, 1)$:

$$\mathbf{v}^{(\tau)} = \gamma \mathbf{v}^{(\tau-1)} + \eta \nabla J(\mathbf{w}^{(\tau)}) \tag{4.5.4}$$

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \mathbf{v}^{(\tau)} \tag{4.5.5}$$

Συμπεριλαμβάνοντας τον προηγούμενο όρο διόρθωσης $\mathbf{v}^{(\tau-1)}$, ο τωρινός $\mathbf{v}^{(\tau)}$ αυξάνεται στις κατευθύνσεις που δείχνουν προς την ίδια κατεύθυνση και μειώνεται στις αντίθετες κατευθύνσεις. Κατά συνέπεια, η σύγκλιση είναι ταχύτερη και με λιγότερες ταλαντώσεις. Η ιδέα παρουσιάζεται στο σχήμα 4.5.1. Ωστόσο, αυτή η επιθυμητή συμπεριφορά εξαρτάται σε μεγάλο βαθμό από την επιλογή της υπερπαραμέτρου $\gamma$ και ενδέχεται να έχει δυσμενείς επιπτώσεις (βλ. 4.5.2).



Σχήμα 4.5.1: Στοχαστική Κατάβαση Κλίσεων με και χωρίς momentum

### 4.5.4 Adaptive Momentum Estimation (Adam)

Η απλή εκδοχή του αλγορίθμου κατάβασης κλίσεων δεν εγγυάται καλή σύγκλιση, καθώς χαρακτηρίζεται από ορισμένα σημαντικά προβλήματα. Αρχικά, είναι επιθυμητή η επιλογή μεταβλητού βήματος μάθησης για τις διάφορες παραμέτρους. Αναλυτικότερα, στην περίπτωση που τα δεδομένα υπό εξέταση είναι αραιά και τα χαρακτηριστικά τους εμφανίζονται με διαφορετική συχνότητα, είναι εύλογη η πραγματοποίηση μεγαλύτερης αναβάθμισης σε σπάνια χαρακτηριστικά, ώστε το μοντέλο να τα ενσωματώσει στη γνώση του. Ακόμη, από την ανάλυση της προηγούμενης παραγράφου, προκύπτει το συμπέρασμα ότι η ένταξη όρου ορμής συνδράμει στην ταχύτητα της σύγκλισης και, ενδεχομένως, στην εύρεση καλύτερου τοπικού ελαχίστου.

Σημειώνεται ότι $\nabla \equiv \nabla_{\mathbf{w}_\tau}$ στις ακόλουθες σχέσεις. Ο αλγόριθμος adam χρησιμοποιεί την πρώτη και τη δεύτερη ροπή της παραγώγου της αντικειμενικής συνάρτησης $\nabla J(\mathbf{w}_\tau)$.

$$\mathbf{m}_\tau = \beta_1 \mathbf{m}_{\tau-1} + (1 - \beta_1) \nabla J(\mathbf{w}_\tau) \tag{4.5.6}$$

$$\mathbf{v}_\tau = \beta_2 \mathbf{v}_{\tau-1} + (1 - \beta_2)(\nabla J(\mathbf{w}_\tau))^2 \tag{4.5.7}$$

Παρατηρούμε ότι οι ροπές τη χρονική στιγμή $\tau$ εξαρτώνται από τις ροπές της προηγούμενης χρονικής στιγμής. Με αυτό τον τρόπο υλοποείται ένας όρος running average. Ωστόσο, αντί να χρησιμοποείται μία διπλή ουρά που θα ήταν υπολογιστικά ακριβή καταφεύγουμε στη χρήση συντελεστών $\beta_1, \beta_2 \in (0, 1)$. Συνήθεις τιμές είναι $\beta_1 = 0.99, \beta_2 = 0.9$. Με τον τρόπο αυτό, οι πιο πρόσφατοι όροι καταλαμβάνουν σπουδαιότερο ρόλο στις ροπές.

Καθώς οι ροπές αρχικοποιούνται με μηδενικά διάνυσματα, αυτές τείνουν να κλίνουν (biased) προς το μηδέν. Οι συγγραφείς Kingma and Ba [KB17] αντιμετωπίζουν το παραπάνω πρόβλημα χρησιμοποιώντας τους ακόλουθους διορθωτικούς όρους:

$$\hat{\mathbf{m}}_\tau = \frac{\mathbf{m}_\tau}{1 - \beta_1^\tau} \tag{4.5.8}$$

$$\hat{\mathbf{v}}_\tau = \frac{\mathbf{v}_\tau}{1 - \beta_2^\tau} \tag{4.5.9}$$

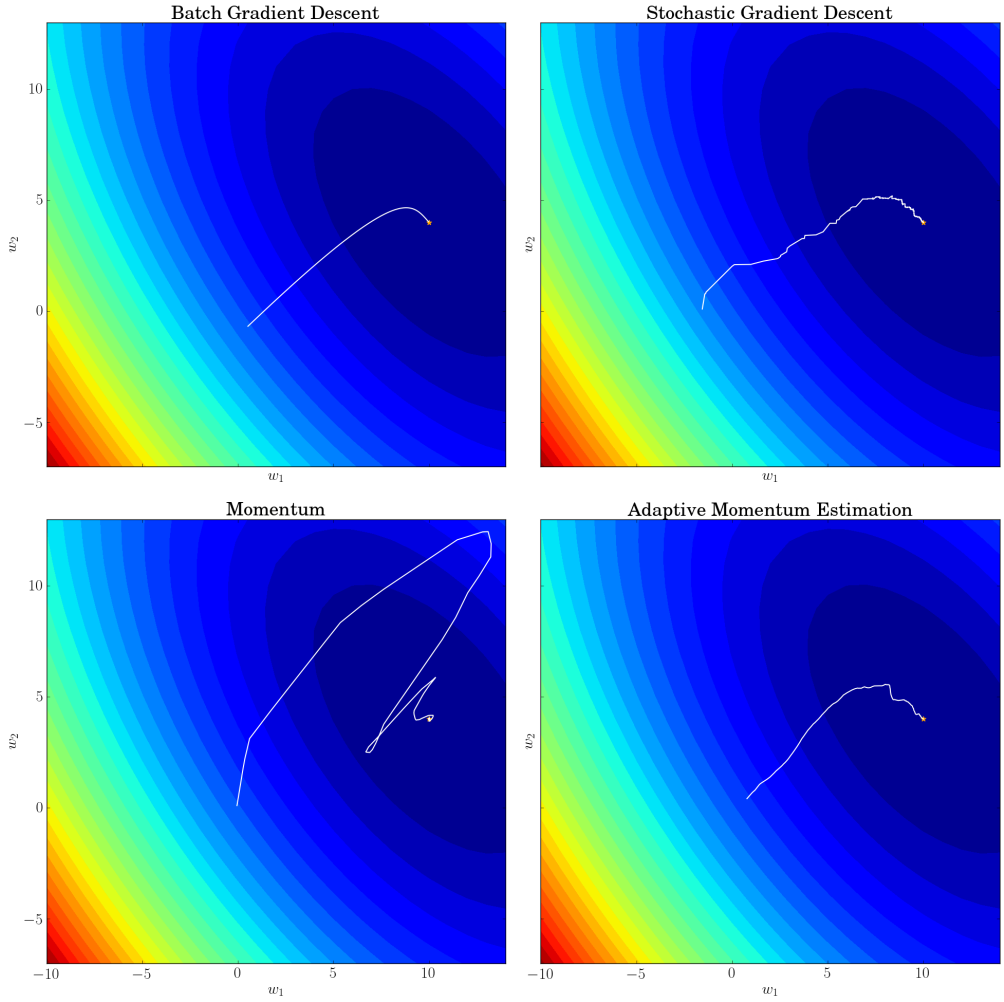
Συνεπώς, ο κανόνας αναβάθμισης των παραμέτρων διαμορφώνεται ως εξής:

$$\mathbf{w}_{\tau+1} = \mathbf{w}_\tau - \frac{\eta}{\sqrt{\hat{\mathbf{v}}_\tau} + \epsilon}\hat{\mathbf{m}}_\tau \tag{4.5.10}$$

Ο αλγόριθμος adam βελτιώνει σε πολλά σημεία την απλή κατάβαση κλίσεων, τόσο θεωρητικά όσο και πειραματικά. Πλέον, αποτελεί μία πολύ σύνηθη επιλογή για την εκπαίδευση των νευρωνικών δικτύων.

Στο σχήμα 4.5.2 απεικονίζονται οι τροχιές των αλγορίθμων βελτιστοποίησης αυτής της ενότητας. Παρατηρούμε ότι η Κατάβαση Κλίσεων που χρησιμοποιεί όλο το σύνολο δεδομένων για τον υπολογισμό της κλίσης σε κάθε βήμα έχει μία πολύ λεία και δίχως ταλαντώσεις πορεία. Ωστόσο, ο χρόνος επίλυσης του προβλήματος είναι απαγορευτικός. Η Στοχαστική Κτάβαση Κλίσεων (εδώ `BATCH SIZE` $= 1$), δεν έχει τόσο λεία πορεία και σχηματικά αφήνει την εντύπωση αργότερης σύγκλισης. Όμως, αυτή η ταλάντωση οφείλεται στο γεγονός ότι η ενημέρωση των βαρών πραγματοποείται *για κάθε* δείγμα. Στη συνέχεια απεικονίζεται η τροχιά του αλγορίθμου Momentum. Παρατηρούμε ότι όρος ορμής "παρασέρνει" την πορεία και οδηγεί σε μεγαλύτερες ταλαντώσεις στην αρχή. Αντιθέτως, καθώς ο αλγόριθμος πλησιάζει στο τοπικό ελάχιστο, ο όρος ορμής εστιάζει την κατεύθυνση με αποτέλεσμα να χρειάζονται λίγοτερα βήματα. Τέλος, ο αλγόριθμος Adaptive Momentum Estimation ελαχιστοποιεί τη συμπεριφορά ταλάντωσης που χαρακτηρίζει τον Momentum και οδηγεί σε σύγκλιση δίχως πολλές αλλαγές κατεύθυνσης.

Αξίζει να σημειωθεί ότι η αντικειμενική συνάρτηση είναι ιδιαίτερα απλή και κυρτή με αποτέλεσμα όλοι οι αλγόριθμοι να βρίσκουν το (ολικό) ελάχιστο. Στην πράξη, οι αντικειμενικές συναρτήσεις παρουσιάζουν μη-κυρτή μορφή και η επιλογή του αλγορίθμου βελτιστοποίησης είναι κρίσιμη.




Σχήμα 4.5.2: Απεικόνιση της τροχιάς των αλγορίθμων βελτιστοποίησης

# Κεφάλαιο 5

# Μορφολογικά νευρωνικά δίκτυα

Στο προηγούμενο κεφάλαιο έγινε μία σύντομη και γενική περιγραφή των νευρωνικών δικτύων. Σε αυτό το κεφάλαιο, επικεντρωνόμαστε σε μία ειδική κατηγορία νευρωνικών δικτύων, τα μορφολογικά νευρωνικά δίκτυα. Τα εν λόγω μοντέλα χρησιμοποιούν συναρτήσεις ενεργοποίησης *μόνο* με $\min, \max$ όρους. Παραδείγματα περιλαμβάνουν τις (4.2.5), (4.2.6), (4.2.8). Ιδιαίτερη προσοχή, ωστόσο, δίνεται σε κόμβους που αντιστοιχούν σε τροπικά πολυώνυμα, min-plus ή max-plus. Οι συναρτήσεις $\min, \max$ είναι εκ φύσεως μη γραμμικές. Αυτό σημαίνει ότι η χρήση τροπικών κόμβων καθιστά ανούσια τη χρήση συνάρτησης ενεργοποίησης.

Τα μορφολογικά νευρωνικά δίκτυα έχουν τις ρίζες τους στη θεωρία Πλεγμάτων (lattice theory) και στα ομώνυμα μορφολογικά μαθηματικά. Η θεωρία πλεγμάτων μπορεί να θεωρηθεί ως μία αλγεβρική γενίκευση της τροπικής άλγεβρας και στο κεφάλαιο αυτό εξερευνούμε τη σύνδεσή τους. Στη συνέχεια, εξετάζουμε ορισμένες αρχιτεκτονικές αναφέροντας τα διάφορα πλεονεκτήματά τους. Τέλος, ακολουθεί μία συζήτηση για μονοτονικά νευρωνικά δίκτυα, δηλαδή για μοντέλα που διατηρούν στην έξοδο τη μονοτονία των μεταβλητών εισόδου.

Ο μορφολογικός νευρώνας εισήχθηκε από τους Davidson and Hummer με στόχο την εκμάθηση μορφολογικών στοιχείων όπως η διαστολή και η συστολή στις εικόνες [DH93]. Η προσπάθεια αυτή εντατικοποιήθηκε και οδήγησε σε πιο γενικά μοντέλα από τους Ritter and Sussner [RS96], οι οποίοι πρότειναν μία απλή αρχιτεκτονική με ένα κρυφό επίπεδο για την ταξινόμηση δυαδικών προτύπων. Τα σύνορα απόφασης αντιστοιχούν σε υπερεπίπεδα παράλληλα στους άξονες. Στο πλαίσιο αυτό προτάθηκαν δύο βελτιώσεις. Αρχικά, ο Sussner επέκτεινε τα δίκτυα συμπεριλαμβάνοντας δεύτερο κρυφό επίπεδο και με τη δυνατότητα ταξινόμησης πολλών κλάσεων [Sus98]. Ακόμη, οι Barmpoutis and Ritter μελέτησαν τη δυνατότητα περιστροφής των υπερεπιπέδων[BR07], αίροντας την περιοριστική ιδιότητα του [RS96]. Οι Ritter and Urcid συνδέουν το μορφολογικό νευρώνα με βιολογικές διεργασίες, θεμελιώνουν τη φύση του στη Θεωρία Πλεγμάτων και αναδεικνύουν την ικανότητα ενός μορφολογικού perceptron με ένα επίπεδο να προσεγγίζει πολυδιάστατες συμπαγείς περιοχές με αυθαίρετο βαθμό ακρίβειας [RU03].

Οι Ritter, Sussner, and Diza-de-Leon εισήγαγαν τον όρο μορφολογικά νευρωνικά δίκτυα [RSD98], αντικαθιστώντας τις πράξεις της πρόσθεσης και του πολλαπλασιασμού με maximum (ή minimum) και πρόσθεση. Πραγματοποίησαν, λοιπόν, μία τροπικοποίηση (tropicalization) των νευρωνικών μαθηματικών. Οι Charisopoulos and Maragos εκμεταλλεύονται αυτή την παρατήρηση και μελετούν την κλάση αυτών των δικτύων με τη βοήθεια της τροπικής θεωρίας [CM17]. Μία παρόμοια κλάση μοντέλων μελέτησαν οι Yang and Maragos στο πλαίσιο της θεωρίας Probably Approximate Correct (PAC) του Valiant, τα min-max δίκτυα, που αντιστοιχούν σε μία γενίκευση Boolean συναρτήσεων με ρίζες στη θεωρία Πλεγμάτων [YM95]. Κοινό στοιχείο σε όλες τις παραπάνω αναφορές είναι η εκμετάλλευση της ιδιαίτερης μορφής που χαρακτηρίζει τους μορφολογικούς τελεστές για το σχεδιασμό αλγορίθμων εκπαίδευσης με ταχύτερη και πιο εύρωστη σύγκλιση, ανταπεξέρχοντας τη μη-παραγωγισιμότητά τους. Ακόμη, οι Pessoa and Maragos παρουσιάζουν την κλάση των morphological/rank/linear (MRL) πολυεπιπέδων πρόσθιων νευρωνικών δικτύων όπου οι νευρώνες αποτελούνται από υβριδικές γραμμικές και μη πράξεις [PM00], ενώ οι Sussner and Esmi μελετούν τα μορφολογικά δίκτυα υπό τη σκοπιά της ανταγωνιστικής μάθησης [SE11].

Πρόσφατα, ερευνητές παρατήρησαν και ανέπτυξαν τη σύνδεση της τροπικής γεωμετρίας με τα νευρωνικά δίκτυα. Οι Montúfar et al. μελετούν την πολυπλοκότητα των βαθιών πρόσθιων νευρωνικών δικτύων με τμηματικά γραμμικές ενεργοποιήσεις και υπολογίζουν ένα άνω όριο για το πλήθος των γραμμικών περιοχών που δημιουργούν[Mon+14]. Οι Charisopoulos and Maragos βελτιώνουν αυτό το άνω όριο, εξετάζοντας το πρόβλημα υπό το τροπικό πρίσμα [CM18]. Τέλος, οι Zhang, Naitzat, and Lim εξετάζουν πρόσθια νευρωνικά δίκτυα με ReLU ενεργοποιήσεις και δείχνουν την αντιστοιχία τους με τροπικές κλασματικές απεικονίσεις [ZNL18].

## 5.1 Μορφολογικά Μαθηματικά

Η θεωρία πλεγμάτων ή lattices (weighted) προσφέρει μία γενίκευση των εννοιών της τροπικής γεωμετρίας. Ένα μερικώς διατεταγμένο σύνολο ή poset (partially ordered set) $(\mathcal{P}, \leq)$ είναι ένα σύνολο $\mathcal{P}$ με τη διμελή πράξη $\leq$ που δημιουργεί μία μερική διάταξη. Αναλυτικότερα, οι εξής ιδιότητες πρέπει να τηρούνται:

- *Ανακλαστική:* Για κάθε $x \in \mathcal{P}$, ισχύει $x \leq x$.
- *Μεταβατική:* Αν $x \leq y$ και $y \leq z$, τότε $x \leq z$.
- *Αντισυμμετρική:* Αν $x \leq y$ και $y \leq x$, τότε $x = y$.

**Definition 5.1.1: Lattice**

Ένα Lattice είναι ένα poset $(\mathcal{L}, \leq)$ με την ιδιότητα ότι κάθε δύο στοιχεία $X, Y \in \mathcal{L}$ έχουν ένα supremum $X \vee Y$ και ένα infimum $X \wedge Y$. Χρησιμοποιείται ο συμβολισμός $(\mathcal{L}, \vee, \wedge)$.

Ένα lattice $\mathcal{L}$ λέγεται πλήρες (complete) αν κάθε υποσύνολο $\mathcal{L}_0 \subset \mathcal{L}$ έχει supremum και infimum in $\mathcal{L}$. Στο πλαίσιο των πλεγμάτων ορίζονται διάφοροι τελεστές, δηλαδή απεικονίσεις (ή mappings) $\psi : \mathcal{L} \to \mathcal{L}$ . Ορισμένες σημαντικές ιδιότητες τελεστών είναι:

| | |
|---|---|
| identity | $\mathbf{id}(X) = X \quad \forall X \in \mathcal{L}$ |
| extensive | $\psi \geq \mathbf{id}$ |
| anti-extensive | $\psi \leq \mathbf{id}$ |
| idempotent | $\psi^2 = \psi$ |
| involution | $\psi^2 = \mathbf{id}$ |

Πίνακας 5.1: Ιδιότητες Μορφολογικών Τελεστών

Ιδιαίτερη σημασία έχουν οι μονότονοι τελεστές, οι οποίοι διατηρούν τη μερική διάταξη. Ένας τελεστής $\psi$ λέγεται increasing αν διατηρεί τη μερική διάταξη: $X \leq Y \implies \psi(X) \leq \psi(Y)$. Αντίστοιχα ορίζουμε και decreasing τελεστές.

Από το μάθημα της Όρασης Υπολογιστών έχουμε δει 4 βασικούς τελεστές:

- dilation $\delta$: $\delta\left(\bigvee_{i\in J} X_i\right) = \bigvee_{i\in J} \delta(X_i)$
- erosion $\epsilon$: $\epsilon\left(\bigwedge_{i\in J} X_i\right) = \bigwedge_{i\in J} \epsilon(X_i)$
- opening $\alpha = \delta\epsilon$: increasing, idempotent, *anti*-extensive
- closing $\beta = \epsilon\delta$: increasing, idempotent, extensive

Οι ελληνικοί όροι των παραπάνω τελεστών είναι (με την ίδια σειρά) διαστολή, συστολή, άνοιγμα και κλείσιμο. Στο κείμενο, χρησιμοποιούμε και τους ελληνικούς και τους διεθνείς όρους. Αξίζει, ωστόσο, να δοθούν και οι ακριβείς ορισμοί των τελεστών διαστολής και συστολής:

**Definition 5.1.2: Διαστολή και Συστολή**

Έστω $\mathcal{L}, \mathcal{M}$ δύο πλήρη lattices. Ενας τελεστής $\epsilon : \mathcal{L} \to \mathcal{M}$ λέγεται συστολή και ένας τελεστής $\delta : \mathcal{L} \to \mathcal{M}$ λέγεται διαστολή αν οι παρακάτω ιδιότητες ισχύουν για κάθε υποσύνολο $X \subseteq \mathcal{L}$:

$$\epsilon\left(\bigwedge X\right) = \bigwedge_{x\in X} \epsilon(x) \tag{5.1.1}$$

$$\delta\left(\bigvee X\right) = \bigvee_{x\in X} \delta(x) \tag{5.1.2}$$

Τέλος, ο δυϊκός τελεστής του $\psi$ χαρακτηρίζεται από τη σχέση $\psi^{\neg}(X) \triangleq [\psi(X^{\neg})]^{\neg}$. Στον τροπικό ημιδακτύλιο, λοιπόν, βλ. θεώρημα 1. Ένας increasing τελεστής $\psi$ σε ένα πλήρες πλέγμα $\mathcal{L}$ λέγεται residuated αν υπάρχει increasing τελεστής $\psi^\sharp$ τέτοιος ώστε

$$\psi\psi^\sharp \leq \mathbf{id} \leq \psi^\sharp\psi \tag{5.1.3}$$

Ο τελεστής $\psi^\sharp$ λέγεται residual (υπόλοιπο) [Mar17]. Οι τελεστές $\psi$ και $\psi^\sharp$ έρχονται πάντα σε ζευγάρι $(\psi, \psi^\sharp)$ που αντιστοιχεί σε dilation και erosion. Το εν λόγω ζευγάρι $(\psi, \psi^\sharp) \equiv (\delta, \epsilon)$ λέγεται adjunction:

$$\delta(X) \leq Y \Leftrightarrow X \leq \epsilon(Y) \quad \forall X, Y \in \mathcal{L} \tag{5.1.4}$$

Το adjunction $(\delta, \epsilon)$ ορίζει δύο προβολές στο πλέγμα:

$$\alpha^2 = \alpha \leq \mathbf{id} \leq \beta = \beta^2 \tag{5.1.5}$$

Πρόκειται για τους τελεστές opening $\alpha$ και closing $\beta$. Ορίζοντας ως $\mathbf{x}\backslash\mathbf{y} = \max\{a \in \mathbb{R} : \mathbf{x} + a \leq \mathbf{y}\}$, η προβολική μετρική Hilbert (Hilbert projective metric) είναι

$$d_H(\mathbf{x}, \mathbf{y}) = -[(\mathbf{x}\backslash\mathbf{y}) + (\mathbf{y}\backslash\mathbf{x})] \tag{5.1.6}$$

Για διανύσματα $x \in \overline{\mathbb{R}}^n$ η προβολική μετρική Hilbert λαμβάνει την απλούστερη μορφή του range semimetric[CGQ04; Aki+11; GK06]:

$$d_H(\mathbf{x}, \mathbf{y}) = \max_i(x_i - y_i) - \min_i(x_i - y_i) \tag{5.1.7}$$

Και το canonical projection του διανύσματος $\mathbf{b}$ στον εαυτό του είναι:

$$P(\mathbf{b}) = \mathbf{A} \boxplus' \mathbf{x}^* = \mathbf{A} \boxplus' \mathbf{A}^* \boxplus \mathbf{b} = \delta(\epsilon(\mathbf{b})) \leq \mathbf{b} \tag{5.1.8}$$

### 5.1.1 Μαθηματική Μορφολογία σε πολλές μεταβλητές

Η άναλυση έχει περιοριστεί σε ιδέες που αφορούν μία μεταβλητή. Αντίστοιχα, ένα πλήρες lattice στο $\mathbb{R}^n$ ορίζει μία μερική διάταξη της μορφής $\mathbf{x} = (x_1, x_2, \ldots, x_n) \preccurlyeq (y_1, y_2, \ldots, y_n) = \mathbf{y} \Leftrightarrow x_i \leq y_i, \forall i \in [n]$. Στην περίπτωση αυτή, τα infimum και supremum ορίζονται με όμοιο τρόπο για το υποσύνολο $X \subseteq \mathbb{R}^n$: $\bigwedge X = (\bigwedge X_1, \bigwedge X_2, \ldots, \bigwedge X_n)$ και $\bigvee X = (\bigvee X_1, \bigvee X_2, \ldots, \bigvee X_n)$ με $X_i = \{\mathbf{x}_i : (x_1, x_2, \ldots, x_n) \in X\}$.

Η μερική διάταξη στα βαθμωτά σύνολα είναι καλά ορισμένη και διαισθητικά απλή. Ωστόσο, η επέκτασή της στο $\mathbb{R}^n$ δημιουργεί προβλήματα, καθώς ακόμα και ένα στοιχείο $i \in [n]$ με αντίθετη φορά ανίσωσης αρκεί για την αδυναμία σύγκρισης δύο σημείων $\mathbf{x}, \mathbf{y} \in X$. Σε διανυσματικό πλαίσιο, λοιπόν, ο παραπάνω ορισμός μερικής διάταξης δεν επιτρέπει τη σύγκριση ανάμεσα στα περισσότερα ζεύγη σημείων. Κατά συνέπεια, χρειάζεται μία εναλλακτική μορφή διάταξης. Αυτή δέχεται το χαρακτηρισμό *μειωμένη* (reduced ordering ή r-ordering). Αναθέτοντας μία τιμή σε κάθε στοιχείο $\mathbf{x} \in \mathbb{R}^n$ του lattice $\mathcal{L}$ είναι πλέον δυνατή η κατασκευή μίας διάταξης. Με μαθηματικό φορμαλισμό, μία $r$-διάταξη που χρησιμοποιεί την απεικόνιση $\rho : \mathcal{L} \to \mathcal{M}$, με $\mathcal{M}$ πλήρες lattice, ορίζεται ως:

$$\mathbf{x} \preccurlyeq_\rho \mathbf{y} \Leftrightarrow \rho(\mathbf{x}) \leq \rho(\mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{L} \tag{5.1.9}$$

Με θεμέλιο λίθο την έννοια της $r$-διάταξης, ακολουθεί η επέκτασή της στους τελεστές:

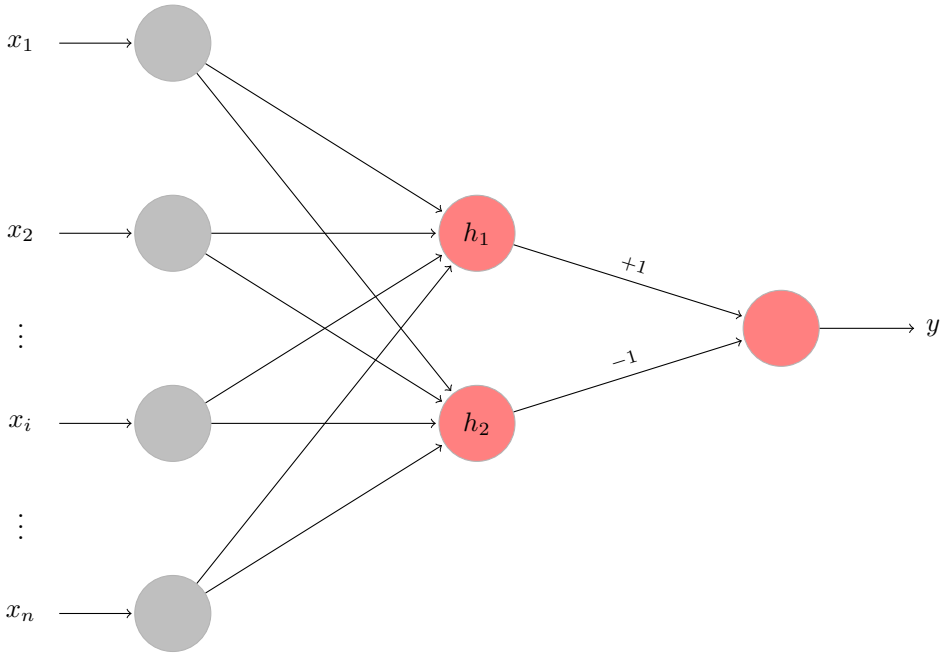
**Definition 5.1.3: $r$-Increasing τελεστής**

> Έστω $\rho : \mathcal{V} \to \mathcal{L}$ και $\sigma : \mathcal{W} \to \mathcal{M}$ δύο επί απεικονίσεις από τα μη κενά σύνολα $\mathcal{V}, \mathcal{W}$ στα πλήρη lattices $\mathcal{L}$ και $\mathcal{M}$. Τότε, ένας τελεστής $\psi : \mathcal{V} \to \mathcal{W}$ είναι $r$-increasing αν $\mathbf{x} \preccurlyeq_\rho \mathbf{y}$ συνεπάγεται ότι $\psi(\mathbf{x}) \leq_\sigma \psi(\mathbf{y})$

## 5.2 Δίκτυα Maxout

Τα δίκτυα Maxout προτάθηκαν από Goodfellow et al. ως μία μέθοδος που αξιοποιεί τα πλεονεκτήματα της τεχνικής *Dropout*[Goo+13b]. Πρόκειται για ένα feed-forward δίκτυο που χρησιμοποιεί την ομώνυμη συνάρτηση ενεργοποίησης. Το δίκτυο που προτείνουν οι συγγραφείς παρουσιάζεται στο σχήμα 5.2.1.

Παρατηρούμε ότι τα δίκτυα maxout έχουν δύο κρυφούς κόμβους, $h_1$ και $h_2$, που χαρακτηρίζονται από την ομώνυμη συνάρτηση ενεργοποίησης (4.2.8). Συνεπώς, οι συναρτήσεις αυτές παράγουν κυρτή έξοδο και η συνολική έξοδος του νευρωνικού $y$ προκύπτει ως η διαφορά αυτών των δύο κυρτών συναρτήσεων. Έστω $y_2 = h(\mathbf{x})$ η έξοδος του νευρώνα $h_2$. Γνωρίζουμε ότι η $y_2$ είναι κυρτή και τμηματικά-γραμμική και, συνεπώς, η $-y_2 = -h(\mathbf{x})$ είναι κοίλη και τμηματικά-γραμμική. Με άλλα λόγια, "περνώντας" το $-1$ στη διέγερση του νευρώνα, λαμβάνουμε μία διαφορετική οπτική γωνία του νευρωνικού, όπου η εξόδος είναι το άθροισμα ενός κυρτού και ενός κοίλου όρου, ενός $(\max, +)$ και ενός $(\min, +)$ όρου. Τα δίκτυα Maxout μπορούν να προσεγγίσουν με αυθαίρετη ακρίβεια οποιαδήποτε συνεχή συνάρτηση $f$.




Σχήμα 5.2.1: Νευρωνικό δίκτυο Maxout[Goo+13b]
με την ομώνυμη συνάρτηση ενεργοποίησης 4.2.8

**Theorem 5.2.1: [Wan04]**

Έστω $m, n \in \mathbb{Z}_+$ υπάρχουν δύο ομάδες $(n+1)$-διάστατων διανυσμάτων $[W_{1j}, b_{1j}]$ και $[W_{2j}, b_{2j}]$ με $j \in [k]$ τέτοια ώστε

$$g(x) = h_1(x) - h_2(x)$$

Δηλαδή, οποιαδήποτε PWL συνάρτηση μπορεί να εκφραστεί ως τη διαφορά δύο κυρτών PWL συναρτήσεων.

Ο Wang μελετά τις τμηματικά γραμμικές συναρτήσεις από τη σκοπιά της θεωρίας πλεγμάτων. Διαισθητικά, επιλέγοντας πολύ μικρά intervals, μία τμηματικά γραμμική συνάρτηση μπορεί να προσεγγίσει πολύ πιστά οποιαδήποτε συνάρτηση, ακόμα και αν αυτή είναι smooth. Πράγματι, το παρακάτω θεώρημα μελετά την ασυμπτωτική συμπεριφορά της κλάσης των τμηματικά γραμμικών συναρήσεων, δηλαδή τη δυνατότητά τους να προσεγγίσουν μία συνάρτηση σε αυθαίρετο βαθμό ακρίβειας.

**Theorem 5.2.2: Stone-Weierstrass**

Από το θεώρημα προσέγγισης Stone-Weierstrass, έστω $C \subset \mathbb{R}^n$ ένα συμπαγές σύνολο, $f :\to \mathbb{R}$ μία συνεχής συνάρτηση και $\epsilon > 0$ ένας θετικός πραγματικός αριθμός. Τότε, υπάρχει συνεχής τμηματικά γραμμική (PWL) συνάρτηση $g = g_\epsilon$ τέτοια ώστε $\forall x \in C : \ |f(x) - g(x)| < \epsilon$.

Από τα θεωρήματα 1,2 συμπεραίνουμε ότι κάθε συνεχής συνάρτηση $f$ μπορεί να προσεγγιστεί με αυθαίρετη ακρίβεια σε ένα συμπαγές πεδίο ορισμού $C \subset \mathbb{R}^n$ από ένα maxout δίκτυο με δύο κρυφές μονάδες maxout.

## 5.3 Πυκνά Μορφολογικά Δίκτυα

Μία σαφής επέκταση των δικτύων Maxout [Goo+13b; CM17] είναι η προσθήκη νευρώνων τόσο σε βάθος όσο και σε πλάτος. Εξετάζουμε πρώτα την περίπτωση όπου αυξάνεται το πλάτος, δηλαδή το πλήθος των νευρώνων στο (μοναδικό) κρυφό επίπεδο. Αρχικά, εισάγουμε ορισμένες έννοιες από τη θεωρία πλεγμάτων (lattice theory) που είναι άρρηκτα συνδεδεμένες με την τροπική γεωμετρία και τα μορφολογικά δίκτυα. Με τη γνώση αυτή, θα

μπορούμε να δούμε τα μορφολογικά δίκτυα από μία επιπλέον σκοπιά.

Οι αποκρίσεις ενός νευρώνα που υπολογίζει ένα τροπικό πολυώνυμο αντιστοιχούν σε διαστολή (dilation) στον $(\max, +)$-ημιδακτύλιο και σε συστολή (erosion) στον $(\min, +)-$ημιδακτύλιο. Χρησιμοποιώντας τον κλασικό συμβολισμό της Όρασης Υπολογιστών:

$$\delta(\mathbf{x}) = \bigvee_{i=1}^{n} w_i + a_i x_i = \mathbf{w}^\top \boxplus \mathbf{x}^{\mathbf{a}} \tag{5.3.1}$$

$$\epsilon(\mathbf{x}) = \bigwedge_{i=1}^{n} w_i + a_i x_i = \mathbf{w}^\top \boxplus' \mathbf{x}^{\mathbf{a}} \tag{5.3.2}$$

Δημιουργώντας πολλά αντίγραφα των νευρώνων διαστολής και συστολής προκύπτει ένα πιο εκφραστικό δίκτυο με $n$ νευρώνες διαστολής και $m$ νευρώνες συστολής στο κρυφό επίπεδο [MSC19]. Το εν λόγω νευρωνικό δίκτυο παρουσιάζεται στο σχήμα 5.3.1.

Έστω, λοιπόν, ότι το κρυφό επίπεδο έχει $n$ νευρώνες διαστολής $\delta_i(\cdot), i \in [n]$ και $m$ νευρώνες συστολής $\epsilon_j(\cdot), j \in [m]$. Ακολουθεί ένα πλήρες συνδεδεμένο επίπεδο που αντιστοιχεί σε γραμμικό συνδυασμό. Επομένως, η έξοδος για το πρότυπο $\mathbf{x}$ είναι:

$$f(\mathbf{x}) = \sum_{i=1}^{n} w_i^{\oplus} \delta_i(\mathbf{x}) + \sum_{j=1}^{m} w_j^{\ominus} \epsilon_j(\mathbf{x}) \tag{5.3.3}$$

όπου με $w_i^{\oplus}$ και $w_j^{\ominus}$ συμβολίζουμε τα βάρη των σχετικών νευρώνων στο πλήρως συνδεδεμένο επίπεδο.

### 5.3.1 Σύνορο απόφασης

Εξετάζουμε το σύνορο απόφασης που παράγει η έξοδος του πυκνού μορφολογικού δικτύου (5.3.3). Εισάγουμε τις έννοιες των *αρθρωτικών συναρτήσεων* (hinge functions) και των *αρθρωτικών υπερεπιπέδων* (hinging hyperplanes) [WS05].

**Definition 5.3.1: $k$-order Hinge function [WS05]**

Μία αρθρωτική συνάρτηση βαθμού $k$ αποτελείται από $(k+1)$ υπερεπίπεδα που συνδέονται συνεχώς με τρόπο που να διατηρείται η συνέχεια της συνάρτησης. Ορίζονται ως:

$$h^{(k)}(\mathbf{x}) = \pm \max\{\mathbf{w}_1^\top \mathbf{x} + b_1, \ldots, \mathbf{w}_{k+1}^\top \mathbf{x} + b_{k+1}\} \tag{5.3.4}$$

Επομένως, μία αρθρωτική συνάρτηση αντιστοιχεί σε τροπικά πολυώνυμα με $(k+1)$ όρους. Πιο συγκεκριμένα, το πρόσημο $+$ στην εξίσωση (5.3.4) αντιστοιχεί σε τροπικό πολυώνυμο υπό την max-plus έννοια. Αντίθετα, το πρόσημο $-$ οδηγεί στη min-plus έκδοση, καθώς $-\max\{\mathbf{w}_1^\top \mathbf{x} + b_1, \ldots, \mathbf{w}_{k+1}^\top \mathbf{x} + b_{k+1}\} = \min\{-\mathbf{w}_1^\top \mathbf{x} - b_1, \ldots, -\mathbf{w}_{k+1}^\top \mathbf{x} - b_{k+1}\} = \min\{\mathbf{v}_1^\top \mathbf{x} + \beta_1, \ldots, \mathbf{v}_{k+1}^\top \mathbf{x} + \beta_{k+1}\}$ για $\mathbf{v}_i = -\mathbf{w}_i$ και $\beta_i = -b_i$, $i \in [k+1]$. Συνεχίζουμε με το δεύτερο ορισμό:

**Definition 5.3.2: $d$-order hinging hypeplanes ($d$-HH) [WS05]**

Ένα αρθρωτικό υπερεπίπεδο βαθμού $d$ ορίζεται ως το άθροισμα πολυβάθμιων αρθρωτικών συναρτήσεων:

$$\sum_i a_i h^{(k_i)}(\mathbf{x}) \tag{5.3.5}$$

όπου $\alpha_i \in \{-1, 1\}, k_i \leq d$.

Κατά συνέπεια, το αρθρωτικό υπερεπίπεδο αποτελεί μία γενίκευση των τροπικών πολυωνύμων και του μαθηματικού οικοδομήματος που ορίζει το δίκτυο Maxout (βλ. σχήμα 5.2.1). Αντιστοιχεί σε άθροισμα τροπικών πολυωνύμων, τόσο max-plus όσο και min-plus. Από τον ισομορφισμό $\phi(x) = -x$ που συνδέει

τους ημιδακτυλίους $(\mathbb{R}_{\max}, \max, +)$ και $(\mathbb{R}_{\min}, \min, +)$ προκύπτει και η εξήγηση ότι το αρθρωτικό υπερεπίπεδο αποτελεί τη διαφορά τροπικών πολυωνύμων (είτε min-plus είτε max-plus). Συνεπώς, η συνάρτηση που ορίζει η εξίσωση (2) είναι μη-κυρτή.

Οι Wang and Sun απέδειξαν το ακόλουθο λήμμα για τα αρθρωτικά υπερεπίπεδα:

**Lemma 5.3.3: [WS05, Θεώρημα 1]**

> Για κάθε θετικό ακέραιο $d$ και αυθαίρετη συνεχή τμηματικά γραμμική συνάρτηση $f : \mathbb{R}^d \to \mathbb{R}$ υπάρχει πεπερασμένο πλήθος, έστω $N$, θετικών ακεραίων $\eta(k) \leq d+1, k \in [N]$ και αντίστοιχα $\alpha_i \in \{-1, 1\}$ ώστε:
> $$f(\mathbf{x}) = \sum_{k=1}^{N} a_i h^{(\eta(k))}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^d \tag{5.3.6}$$

Άρα, κάθε συνεχής και τμηματικά γραμμική συνάρτηση $d$ μεταβλητών μπορεί να εκφραστεί ως ένα αρθρωτικό υπερεπίπεδο βαθμού $d$. Επιπλέον, παρατηρούμε ότι η συνάρτηση εξόδου του μορφολογικού δικτύου (5.3.3): $f(\mathbf{x}) = \sum_{i=1}^{n} w_i^{\oplus} \delta_i(\mathbf{x}) + \sum_{j=1}^{m} w_j^{\ominus} \epsilon_j(\mathbf{x})$ αντιστοιχεί σε μία πολυβάθμια αρθρωτική συνάρτηση. Τότε, από το θεώρημα προσέγγισης Stone-Weierstrass 2, προκύπτει ότι αρκεί ένα επίπεδο διαστολής-συστολής σε σειρά με ένα πλήρως συνδεδεμένο επίπεδο για την προσέγγιση οποιασδήποτε συνεχούς και ομαλής συνάρτησης. Ο βαθμός προσέγγισης, δηλαδή το περιθώριο λάθους, καθορίζεται από το πλήθος των νευρώνων.

Το μορφολογικό δίκτυο του σχήματος 5.3.1 μαθαίνει τη συνάρτηση (5.3.3), η οποία αντιστοιχεί σε μία συλλογή πολλών υπερεπιπέδων ενωμένων μαζί. Αυτό σημαίνει ότι το πλήθος των υπερεπιπέδων που το δίκτυο μαθαίνει με $l \triangleq n + m$ νευρώνες στο μορφολογικό επίπεδο διαστολής-συστολής είναι πολύ μεγαλύτερο από $l$. Οι τελεστές $\max, \min$ επιλέγουν μόνο μία είσοδο να περάσει από το νευρώνα.

### 5.3.2 Ομαλοποιημένα Μορφολογικά Δίκτυα

Οι τελεστές max και min παράγουν τμηματικά γραμμικές συναρτήσεις. Όπως επισημάνθηκε στα προηγούμενα κεφάλαια, τα σημεία της τροπικής υπερεπιφάνειας, δηλαδή τα σημεία που το max ή min επιτυγχάνονται πάνω από μία φορά, δεν είναι διαφορίσιμα. Η παραγωγισιμότητα είναι χρήσιμη στην εκπαίδευση των νευρωνικών δικτύων. Η σχετική ανάλυση βρίσκεται στην παράγραφο για την οπισθοδιάδοση 4.4. Η σύνηθης λύση έγκειται στη χρήση υποπαραγώγων (subgradients). Ωστόσο, δύναται και η χρήση ομαλοποιημένων εκδοχών για τους προαναφερθέντες τελεστές. Η ομαλή έκδοση των μορφολογικών τελεστών διαστολής και συστολής είναι [MSC19]:

$$\delta_{(\beta)}(\mathbf{x}) = \frac{1}{\beta} \log \left( \sum_k e^{\beta(x_k + s_k)} \right) \tag{5.3.7}$$

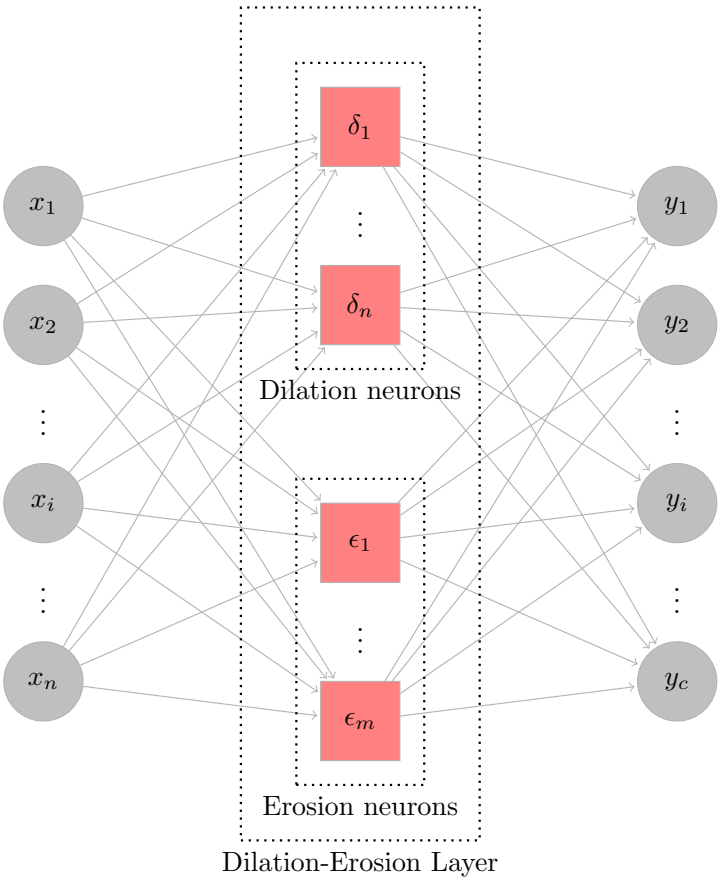
$$\epsilon_{(\beta)}(\mathbf{x}) = -\frac{1}{\beta} \log \left( \sum_k e^{\beta(s_k - x_k)} \right) \tag{5.3.8}$$

Η παράμετρος $\beta$ καθορίζει τη σκληρότητα του ομαλοποιημένου τελεστή. Καθώς $\beta \to \infty$, οι ομαλοποιημένες εκδοχές (5.3.7), (5.3.8) συγκλίνουν στις "κανονικές" (5.3.1), (5.3.2). Στο πλαίσιο της τροπικής γεωμετρίας καθώς και των ταυτοδύναμων μαθηματικών (idempotent mathematics), η εν λόγω προσέγγιση είναι γνωστή και έχει μελετηθεί εκτενώς. Στη βιβλιογραφία αναφέρεται ως Maslov Dequantization [Lit07].

**Lemma 5.3.4**

> Για $\beta \to \infty$, εχουμε $\delta_\beta(\mathbf{x}) \to \delta(\mathbf{x})$.

*Proof.* Θα δείξουμε ότι $\delta_{(\beta)}(\mathbf{x}) = \frac{1}{\beta} \log \left( \sum_k e^{\beta(x_k + s_k)} \right) \to \max\{x_k + s_k\} = \delta(\mathbf{x})$ καθώς $\beta \to \infty$. Χάριν




Σχήμα 5.3.1: Πυκνό μορφολογικά δίκτυο με $n$ νευρώνες διαστολής και $m$ νευρώνες συστολής στο κρυφό επίπεδο.

ευκολίας, έστω $y_k = x_k + s_k$. Τότε:

$$\begin{aligned}
\lim_{\beta\to\infty} \delta_{(\beta)}(\mathbf{x}) &= \lim_{\beta\to\infty} \frac{1}{\beta} \log\left(\sum_k e^{\beta y_k}\right) \\
&= \lim_{\beta\to\infty} \frac{\log\left(\sum_k e^{\beta y_k}\right)}{\beta} && \triangleright \text{ DLH} \\
&= \lim_{\beta\to\infty} \frac{\sum_k y_k e^{\beta y_k}}{\sum_k e^{\beta y_k}} \\
&= \lim_{\beta\to\infty} \sum_j \frac{y_j e^{\beta y_j}}{\sum_k e^{\beta y_k}} \\
&= \lim_{\beta\to\infty} \sum_j \frac{y_j}{1 + \sum_{k\neq j} e^{\beta(y_k - y_j)}} && \triangleright \text{ divide by } e^{\beta y_j} \\
&= \sum_j \underbrace{\lim_{\beta\to\infty} \frac{y_j}{1 + \sum_{k\neq j} e^{\beta(y_k - y_j)}}}_{A_j} && (5.3.9)
\end{aligned}$$

όπου με DLH επισημειώνεται η χρήση του κανόνα *De L'Hôpital*. Έστω ότι υπάρχουν $K$ στοιχεία που είναι ίσα με το maximum. Διακρίνουμε τις εξής περιπτώσεις για την ποσότητα $A_j$:

- $j \neq j^* = \mathbf{argmax}_k\, y_k$. Τότε, υπάρχουν τρία σύνολα $I^>, I^<, I^=$ τέτοια ώστε $I^> = \{i : y_i > y_j\}$, $I^< = \{i : y_i < y_j\}$ και $I^= = \{i : y_i = y_j\}$. Προφανώς $I^> \neq \emptyset$. Άρα,

$$\begin{aligned}
A_j &= \lim_{\beta\to\infty} \frac{y_j}{1 + \sum_{k\neq j} e^{\beta(y_k - y_j)}} \\
&= \lim_{\beta\to\infty} \frac{y_j}{1 + \sum_{k\in I^<} e^{\beta(y_k - y_j)} + \sum_{k\in I^=} e^{\beta(y_k - y_j)} + \sum_{k\in I^>} e^{\beta(y_k - y_j)}} \\
&= \frac{y_j}{1 + \sum_{k\in I^<} e^{-\infty} + \sum_{k\in I^>} e^{0} + \sum_{k\in I^>} e^{\infty}} = 0
\end{aligned}$$

- $j = j^* = \mathbf{argmax}_k\, y_k$. Τότε, $I^> = \emptyset, |I^=| = K-1$ από την υπόθεση και κάθε όρος $y_k - y_j$ του εκθετικού στον παρονομαστή είναι αρνητικός για $k \in I^<$. Τότε:

$$\begin{aligned}
A_j &= \lim_{\beta\to\infty} \frac{y_j}{1 + \sum_{k\neq j} e^{\beta(y_k - y_j)}} \\
&= \frac{y_j}{1 + \sum_{k\in I^=} e^{0} + \sum_{k\in I^<} e^{-\infty}} \\
&= \frac{y_j}{1 + (K-1) + 0} \\
&= \frac{y_j}{K} = \frac{y_{\max}}{K}
\end{aligned}$$

Συνεπώς, η ποσότητα $A_j$ είναι μηδενική για κάθε όρο $y_j$ που δεν αντιστοιχεί στο μέγιστο $y_{\max}$ και ίση με $\frac{y_{\max}}{K}$ για κάθε έναν από τους $K$ όρους που είναι ίσοι με το μέγιστο. Τότε, από τη σχέση 5.3.9 προκύπτει το ζητούμενο.

□

Τα ομαλοποιημένα μορφολογικά δίκτυα συνδέονται στενά με τις αρχιτεκτονικές Log-Sum-Exp [CGP18; Cal+19]. Αναλυτικότερα, οι Calafiore, Gaubert, and Possieri αποδεικνύουν δυνατότητα καθολικής προσέγγισης κυρτών συναρτήσεων από την κλάση συναρτήσεων Log-Sum-Exp (εν συντομία $\mathrm{LSE}_T$) $f : \mathbb{R}^n \to \mathbb{R}$:

$$f_T(\mathbf{x}) = T \log\left(\sum_{k=1}^{K} b_k^{1/T} \exp\left(\langle \alpha^{(k)}, \mathbf{x}/T \rangle\right)\right)$$

με $K \in \mathbb{N}, b_k > 0, \alpha^{(k)} \in \mathbb{R}^n$ και η παράμετρος $T \in \mathbb{R}^{>0}$ αναφέρεται ως παράμετρος θερμοκρασίας. Είναι αντίστροφη της παραμέτρου $\beta$ (5.3.7). Οι συγγραφείς στοχεύουν στην προσέγγιση κυρτών συναρτήσεων μέσω ενός απλού πρόσθιου νευρωνικού δικτύου με ένα κρυφό επίπεδο. Επιλέγουν, λοιπόν, κατάλληλες συναρτήσεις ενεργοποίησης. Οι κρυφοί κόμβοι χαρακτηρίζονται από εκθετική συνάρτηση ενεργοποίησης: $\phi(x) = \exp\left(\frac{x}{T}\right)$, ενώ για την έξοδο ισχύει $\phi(x) = T\log(x)$. Συνεπώς, η έξοδος του νευρωνικού δικτύου είναι $y = T\log(x) = T\log\left(\sum_{k=1}^{K} a_k\right)$ με $a_k$ να σημειώνεται η έξοδος του $k$-οστού κόμβου του κρυφού επιπέδου.

Σαφώς, η κλάση $\mathrm{LSE}_T$ συνδέεται στενά με την κλάση των γενικευμένων posynomials $\mathrm{GPOS}_T$ ορισμένα στην εξίσωση (3.2.9) συμπεριλαμβάνοντας αντίστοιχη παράμετρο θερμοκρασίας $T$. Πρόκειται για μία αντιστοιχία ένα-προς-ένα, σύμφωνα με [CGP18, Proposition 3]. Αν $f(\mathbf{x}) \in \mathrm{LSE}_T$ και $\psi(\mathbf{z}) \in \mathrm{GPOS}_T$:

$$\exp(f(\log(\mathbf{z}))) \in \mathrm{GPOS}_T$$
$$\log(\psi(\exp(\mathbf{x}))) \in \mathrm{LSE}_T$$

## 5.4 Πυκνά και Βαθιά Μορφολογικά Δίκτυα

Επεκτείνουμε τη λογική των Πυκνών Μορφολογικών Δικτύων της προηγούμενης ενότητας σε περισσότερα επιπέδα. Ένα παράδειγμα φαίνεται στο σχήμα 5.4.1. Στοιβάζοντας επίπεδα διαστολής-συστολής, προκύπτουν αποκρίσεις παρόμοιες με μορφολογικούς τελεστές opening και closing. Πιο συγκεκριμένα, έστω $A, B$ δύο σύνολα. Ακολουθώντας τις συμβάσεις της Όρασης Υπολογιστών, συμβολίζουμε με $\oplus, \ominus, \circ, \bullet$ τη διαστολή (dilation), συστολή (erosion), άνοιγμα (opening) και κλείσιμο (closing). Για σύνολα $A, B$ οι τελεστές ορίζονται ως:

$$A \circ B = (A \ominus B) \oplus B \tag{5.4.1}$$
$$A \bullet B = (A \oplus B) \ominus B \tag{5.4.2}$$

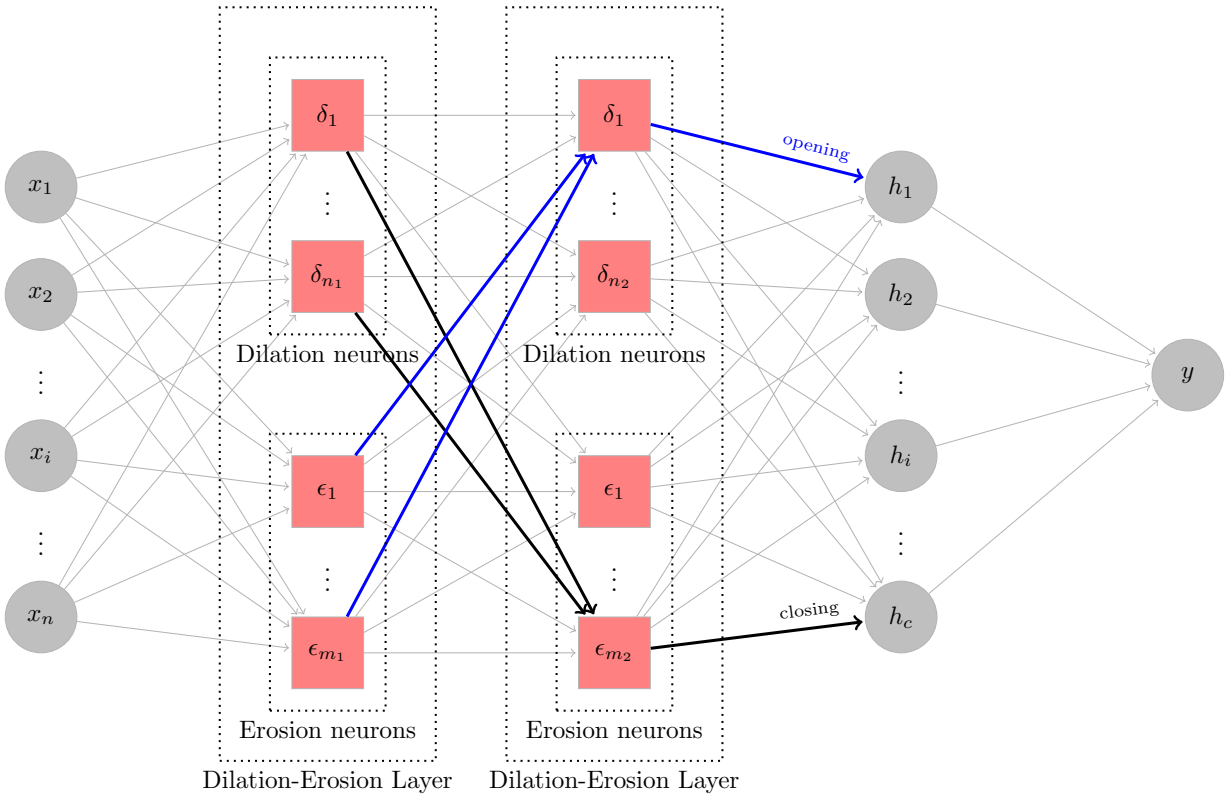
Ο τελεστής opening αντιστοιχεί, λοιπόν, στην εφαρμογή συστολής και, στη συνέχεια, εφαρμογή διαστολής στο αποτέλεσμα. Χρησιμοποείται το ίδιο κατασκευαστικο στοιχείο, εδώ το σύνολο $B$, και για τις δύο μορφολογικές πράξεις. Στο πεδίο της όρασης υπολογιστών, το μορφολογικό άνοιγμα χρησιμοποείται για αποθορυβοποίηση, καθώς απομακρύνει μικρά αντικείμενα διατηρώντας το σχήμα και το μέγεθος μεγαλύτερων. Ενώνει μικρές σχισμές και έχει παρόμοια επίδραση με τη συστολή, αν και σε μικρότερο βαθμό, καθώς αφαιρεί ορισμένα pixels από τις ακμές αντικειμένων στο προσκήνιο.

Το μορφολογικό κλείσιμο εκτελεί τις πράξεις της διαστολής και συστολής με αντίστροφη σειρά από το άνοιγμα. Συνεπώς, είναι χρήσιμο για να γεμίζει τρύπες, διατηρώντας το σχήμα των αντικειμένων. Ενώνει μικρές φωτεινές περιοχές και αφαιρεί μικρά σκοτεινά σημεία. Η επέκταση των πυκνών μορφολογικών δικτύων σε βαθύτερες αρχιτεκτονικές μπορεί να γίνει με δύο τρόπους: με ή χωρίς επανάληψη του πλήρως συνδεδεμένου επιπέδου. Αναλυτικότερα, γίνεται να χρησιμοποιήσουμε πολλά μορφολογικά επίπεδα (διάφορων μεγεθών) σε σειρά και να ολοκληρώσουμε το μοντέλο με ένα πλήρως συνδεδεμένο επίπεδο. Διαφορετικά, μπορούμε να θεωρήσουμε το μορφολογικό επίπεδο μαζί με το πλήρως συνδεδεμένο ως μία μονάδα και το δίκτυο να αποτελείται από πολλές τέτοιες μονάδες εν σειρά.

## 5.5 Μονοτονικά Δίκτυα

### 5.5.1 Μέθοδοι επιβολής ιδιότητας μονοτονίας

Ο τομέας της Μηχανικής Μάθησης άνθισε ιδιαίτερα από τις βαθιές αρχιτεκτονικές (Deep Learning). Πρωτογενής ανάπτυξη της πειραματικής έρευνας συνόδευσε καινοτομίες στη βιομηχανία. Ωστόσο, τα βαθιά μοντέλα




Σχήμα 5.4.1: Πυκνό μορφολογικά δίκτυο με δύο επίπεδα διαστολής-συστολής. Διατάσσοντας σε σειρά μορφολογικούς νευρώνες διαστολής και συστολής προκύπτουν αποκρίσεις παρόμοιες με άνοιγμα (opening) και κλείσιμο (closing).

χαρακτηρίζονται από ορισμένα μείζονα προβλήματα: αφενός απαιτείται πολύς χρόνος για την εκπαίδευσή τους και αφετέρου χρησιμοποιούνται ως μαύρα κουτιά με αποτέλεσμα η συμπεριφορά τους σε άγνωστα δείγματα να είναι απρόβλεπτη.

Τα τελευταία χρόνια, αναδύεται μία νέα τάση που φέρει τον τίτλο *interpretability* (ερμηνευσιμότητα). Η προσπάθεια εντοπίζεται στην κατασκευή μοντέλων που προσφέρουν διαφάνεια στους χρήστες και διαίσθηση στους ερευνητές. Αυτό είναι ιδαίτερα σημαντικό πλέον, καθώς τα συστήματα μηχανικής μάθησης είναι πανταχού παρόντα και η χρήση τους δεν περιορίζεται σε ειδήμονες. Ένα σημαντικό χαρακτηριριστικό ερμηνευσιμότητας είναι η μονοτονία της εξόδου ως προς (ορισμένα) χαρακτηριστικά της εισόδου [Gup+16]. Αυτός ο περιορισμός συνεπάγεται ότι η αύξηση ενός στοιχείου $x_i$ της εισόδου $\mathbf{x}$ δεν μπορεί να οδηγήσει σε μείωση της εξόδου $y$.

Έστω, για παράδειγμα, ότι στόχος είναι η εκτίμηση της τιμής ενός αυτοκινήτου και ένα από τα χαρακτηριστικά είναι ο αριθμός των χιλιομέτρων που έχει διανύσει. Διατηρώντας τα υπόλοιπα χαρακτηριστικά αναλλοίωτα, αναμένουμε ότι η αύξηση των χιλιομέτρων να οδηγήσει σε μείωση της τιμής, λόγω φθοράς. Ωστόσο, ένα μοντέλο εκπαιδευμένο με θορυβώδη δείγματα ενδέχεται να μη συμμορφωθεί στην πρότερη γνώση. Ένα πιθανό σενάριο είναι η έλλειψη πολλών δεδομένων με αποτέλεσμα ένας αλγόριθμος όπως η Στοχαστική Κατάβαση Κλίσεων να παράξει μία επιφάνεια που να μην υπακούει αυτό το διαισθητικό κανόνα σε ένα υποσύνολο του πεδίου ορισμού.

Στην παράγραφο αυτή, λοιπόν, αντιμετωπίζουμε το πρόβλημα της μονότονης παλινδρόμησης από τη σκοπιά της μηχανικής μάθησης. Σε γενικότερο πλαίσιο, το πρόβλημα έχει λάβει την ονομασία *ισότονη παλινδρόμηση* (isotonic regression) [BC90] και η επίλυση βασίζεται στον επαναπροσδιορισμό των ετικετών για τα δείγματα με τιμές κοντά στη μονότονη επιφάνεια. Η διατύπωση του προβλήματος στο $\mathbb{R}$ είναι η εξής:

$$\begin{aligned} \text{minimize} \quad & \sum_i w_i (y_i - \hat{y}_i)^2 \\ \text{subject to} \quad & \hat{y}_i < \hat{y}_j \qquad x_i < x_j \end{aligned}$$

όπου με $\hat{y}_i$ συμβολίζεται η εκτιμώμεμενη τιμή, δηλαδή η έξοδος του προβλήματος και με $x_i, y_i$ σημειώνονται η τετμημένη και η τεταγμένη (πραγματική έξοδος) του δείγματος $i$, αντίστοιχα. Η προκύπτουσα συνάρτηση είναι τμηματικά σταθερή. Το πρόβλημα επιλύεται σε $O(n)$ χρόνο αν τα δεδομένα υπακούν κάποια *ολική* διάταξη. Ωστόσο, στη γενικότερη περίπτωση που συναντάται σε προβλήματα μάθησης, αναζητείται *μερική* διάταξη των δεδομένων και το πρόβλημα τετραγωνικού προγραμματισμού που ορίζει η ισότονη παλινδρόμηση επιλύεται με πολυπλοκότητα $O(n^4)$.

### 5.5.2 Μονοτονία μέσω Μορφολογικών Δικτύων

Η επιβολή αυτής της συνθήκης μπορεί να πραγματοποιηθεί μέσω όρων κανονικοποίησης (regularization) που "τιμωρούν" τιμές που δεν την ικανοποιούν. Αυτές οι μέθοδοι είναι υπολογιστικά ακριβές και απαιτούν τη βελτιστοποίηση επιπρόσθετων (υπερ)παραμέτρων. Μία εναλλακτική προσέγγιση είναι η σχεδίαση της αρχιτεκτονικής του μοντέλου ώστε να επιβάλλει την πρότερη γνώση δίχως το επιπλέον υπολογιστικό βάρος. Αυτό το μονοπάτι ακολούθησαν οι Archer and Wang και όρισαν ένα νευρωνικό δίκτυο για προβλήματα δυαδικής ταξινόμησης που εγγυάται μονοτονία [AW93]. Ο αλγόριθμός τους μειώνει τα βάρη των δειγμάτων με παραγώγους που παραβιάζουν τη συνθήκη της μονοτονίας, ώστε να παράξει ένα δίκτυο με θετικά βάρη. Άλλοι μελετητές πρότειναν την επιβολή θετικών βαρών σε δίκτυο ενός κρυφού επιπέδου που συμπληρώνεται από μία σιγμοειδή (ή αλλο μονότονο και μη γραμμικό μετασχηματισμό) στην έξοδο. Ωστόσο, οι Daniels and Velikova απέδειξαν ότι αυτή η λογική απαιτεί $K$ κρυφά επίπεδα ώστε να επιτύχει προσέγγιση μίας $K$-διάστατης μονότονης συνάρτησης σε αυθαίρετο βαθμό ακριβείας [DV10], γεγονός που καθιστά τη μέθοδο αυτή απαγορευτική.

Μία εναλλακτική προσέγγιση είναι η σχεδίαση αρχιτεκτονικών που εγγυώνται τη συνθήκη της μονοτονίας, Ο Sill προτείνει μία απλή αρχιτεκτονική με min και max συναρτήσεις ενεργοποίησης. Το μοντέλο φαίνεται στο σχήμα 5.5.1 και ένα παράδειγμα καμπύλης εξόδου στο 5.5.2 για το ίδιο μοντέλο στην περίπτωση μονοδιάσταστης εισόδου. Εφόσον τα βάρη ανάμεσα στο επίπεδο εισόδου και στο κρυφό είναι θετικά, η έξοδος είναι μονότονη. Η επιβολή θετικών βαρών μπορεί να επιτευχθεί με διάφορους τρόπους. Ο Sill προτείνει τη χρήση εκθετικού μετασχηματισμού. Έστω πως η μονοτονία είναι επιθυμητή για τη μεταβλητή εισόδου $i$ και $\forall j, k$ τα βάρη που αντιστοιχούν στην είσοδο υπολογίζονται ως εξής $w_i^{j,k} = e^{z_i^{(j,k)}}$. Βελτιστοποιώντας τη μεταβλητή $z_i^{(j,k)}$, τα βάρη

παραμένουν θετικά. Άλλες επιλογές περιλαμβάνουν τις συναρτήσεις ενεργοποίησης της σχετικής παραγράφου 4.2. Αναλυτικότερα, δύναται η χρήση της ReLU (4.2.5) ή της ομαλής της προσέγγισης Softplus (4.2.7). Ωστόσο, η επιλογή της ReLU δε συνίσταται, καθώς δεν αξιοποιεί την πληροφορία για αρνητικές τιμές και απαιτείται η χρήση υποπαραγώγων, δυσχεραίνοντας τη σύγκλιση. Απλές επιλογές θετικών συναρτήσεων, όπως $z \mapsto z^2$, προτιμώνται. Ο Sill ορίζει ως ενεργή (active) την ομάδα $g_l$ που καθορίζει την έξοδο του μοντέλου:

$$l = \arg\min_k g_k(\mathbf{x}) \tag{5.5.1}$$

Οι μονάδες max αντιστοιχούν σε ομάδες (groups). Ο τελεστής max επιτρέπει την προσέγγιση κυρτών επιφανειών εντός κάθε ομάδας, ενώ ο τελεστής min στις ομάδες επιτρέπει την προσέγγιση κοίλων περιοχών της συνάρτησης που προσπαθούμε να προσεγγίσουμε. Ο Sill προτείνει μία ιδιαίτερη παραλλαγή της Στοχαστικής Κατάβασης Κλίσεων, όπου τα δείγματα που σχετίζονται με ένα ενεργό υπερεπίπεδο εξετάζονται σε κάθε επανάληψη (αντί να έρχονται με τυχαία σειρά). Έστω, λοιπόν, ότι δίκτυο αποτελείται από $K$ ομάδες (max όρους δηλαδή), καθεμία από τις οποίες αποτελείται από $J$ αφφινικούς όρους που ορίζονται από το διάνυσμα βαρών $\mathbf{w}_{k,j}$ και τον όρο πόλωσης $b_{k,j}$. Τότε, η έξοδος του δικτύου είναι:

$$y = f(\mathbf{x}) = \min_{k\in[K]} \max_{j\in[J]} \{\mathbf{w}_{k,j}^\top \mathbf{x} + b_{k,j}\} \tag{5.5.2}$$

Αξίζει να εντρυφήσουμε στη διαμόρφωση του αλγορίθμου οπισθοδιάδοσης για το μορφολογικο μοντέλο του Sill. Οι τελεστές $\min, \max$ δεν είναι παραγωγίσιμοι σε όλο το πεδίο ορισμού. Χάριν ευκολίας, χρησιμοποιούμε το γνωστό συμβολισμό $\epsilon$ για τον κόμβο min και $\delta_i$ για τον $i$-οστό όρο max, $i \in [K]$. Για το πρόβλημα παλινδρόμησης χρησιμοποιούμε ως μετρική κόστους $J$ το μεσοτετραγωνικό σφάλμα. Έστω $f : \mathbb{R}^n \to \mathbb{R}$ η έξοδος του δικτύου για δείγματα $\mathbf{x}_i \in \mathbb{R}^n$ και επιθυμητή έξοδο $y_i \in \mathbb{R}$. Με γνώμονα το σχήμα 5.5.1 και από τον κανόνα της αλυσίδας έχουμε:

$$\begin{aligned}
\frac{\partial J}{\partial \epsilon} &= \sum_i \frac{\partial J}{\partial f(\mathbf{x}_i)} \cdot \frac{\partial f(\mathbf{x}_i)}{\partial \epsilon} \\
&= \sum_i (f(\mathbf{x}_i) - y_i) \cdot 1 = \sum_i (f(\mathbf{x}_i) - y_i) \qquad (5.5.3) \\
\frac{\partial J}{\partial \delta_k} &= \sum_i \frac{\partial J}{\partial f(\mathbf{x}_i)} \cdot \frac{\partial f(\mathbf{x}_i)}{\partial \epsilon} \cdot \frac{\partial \epsilon}{\partial \delta_k} \\
&= \sum_i (f(\mathbf{x}_i) - y_i) \cdot \frac{\partial \epsilon}{\partial \delta_k} \qquad (5.5.4) \\
\frac{\partial \epsilon}{\partial \delta_k} &= \begin{cases} 1 & \mathbf{argmax}_h\{\delta_h\} = k \\ 0 & \text{αλλιώς} \end{cases} \qquad (5.5.5) \\
\frac{\partial J}{\partial \mathbf{w}_{k,j}} &= \sum_i \frac{\partial J}{\partial f(\mathbf{x}_i)} \cdot \frac{\partial f(\mathbf{x}_i)}{\partial \epsilon} \cdot \frac{\partial \epsilon}{\partial \delta_k} \cdot \frac{\partial \delta_k}{\partial \mathbf{w}_{k,j}} \\
&= \sum_i (f(\mathbf{x}_i) - y_i) \cdot \frac{\partial \epsilon}{\partial \delta_k} \cdot \frac{\partial \delta_k}{\partial \mathbf{w}_{k,j}} \qquad (5.5.6) \\
\frac{\partial \delta_k}{\partial \mathbf{w}_{k,j}} &= \begin{cases} \mathbf{x} & \mathbf{argmax}_h\{\mathbf{w}_{k,h}^\top \mathbf{x}\} = j \\ 0 & \text{αλλιώς} \end{cases} \qquad (5.5.7)
\end{aligned}$$

Χάριν ευκολίας, χρησιμοποιείται το επαυξημένο διάνυσμα βαρών $\mathbf{w}_{k,j}$. Από τις εξισώσεις (5.5.5) και (5.5.7) συμπεραίνουμε ότι ο αλγόριθμος οπισθοδιάδοσης για το μεμονωμένο δείγμα $\mathbf{x}_i$ ενημερώνει τα βάρη *μόνο* της ενεργούς ομάδας. Για το λόγο αυτό, προτείνεται ο ιδιαίτερος τρόπος εκμάθησης που αναφέρεται παραπάνω.

Ένας πρόγονος του δικτύου του Sill είναι η μελέτη των Yang and Maragos [YM95], οι οποίοι εισήγαγαν τους min-max ταξινομητές και τους μελέτησαν ως μία γενίκευση Boolean συναρτήσεων στο πλαίσιο της Θεωρίας

Πλεγμάτων. Ένας απόγονος του δικτύου του Sill είναι η μελέτη των [VDF06], όπου επεκτείνεται το μοντέλο και σε μερικώς μονότονες συναρτήσεις. Οι εν λόγω συναρτήσεις είναι μονότονες μόνο ως προς ορισμένα χαρακτηριστικά της εισόδου.

Προτού συνεχίσουμε στην τροπική ανάλυση της αρχιτεκτονικής του Sill, αξίζει να σημειωθούν δύο εναλλακτικές προσεγγίσεις. Η πρώτη εντοπίζεται στη μονότονη παλινδρόμηση μέσω lattices [Gup+16]. Στο πλαίσιο αυτό, ο όρος είναι διαφορετικός από την προηγούμενη χρήση του. Χρησιμοποιούμε τον αγγλικό όρο, αντί για τον ελληνικό (πλέγμα), για αποφυγή παρανόησης. Πρόκειται για πίνακες που περιέχουν τις τιμές της συνάρτησης και οι ενδιάμεσες τιμές προκύπτουν με παρεμβολή. Ένα κλασικό παράδειγμα είναι οι πίνακες (look-up tables) με τιμές της κανονική κατανομής $\mathcal{N}(0,1)$ που συναντώνται στο τέλος βιβλίων με θέμα την εισαγωγή στις πιθανότητες και τη στατιστική. Καθώς οι τιμές που περιέχουν είναι διακριτές, ο υπολογισμός της τιμής της συνάρτησης συνάρτησης πυκνότητας πιθανότητας (σ.π.π.) πραγματοποείται με παρεμβολή. Η ιδέα αυτή ξεκίνησε με παλινδρόμηση χωρίς περιορισμούς στα lattices [GG09], όπου χρησιμοποιούνται αλγόριθμοι Empirical Risk Minimization (ERM) για την προσέγγιση συναρτήσεων από δείγματα εκπαίδευσης. Στη συνέχεια, επέκτειναν αυτή την ιδέα σε προβλήματα με περιορισμούς, όπως μονοτονία ή κυρτότητα [Gup+16; Gup+18] και σε τεχνικές όπως νευρωνικά δίκτυα και βαθιά μηχανική μάθηση μέσω Deep Lattice Networks [You+17; Mil+16].

Τέλος, μία πρόσφατη εξέλιξη αφορά την προσέγγιση συνεχών κατανομών πιθανότητας, όπου η αθροιστική συνάρτηση κατανομής (Α.Σ.Κ.) είναι εξ ορισμού μονότονη. Οι Wehenkel and Louppe χρησιμοποιούν μονότονα νευρωνικά δίκτυα με στόχο τον ορισμό *αντιστρέψιμων μετασχηματισμών*. Σημειώνουν ότι οι αρχιτεκτονικές που επιβάλλουν θετικά βάρη ή/και συναρτήσεις ενεργοποίησης επιτυγχάνουν μονοτονία αλλά πληρώνουν το τίμημα σε περιορισμένη εκφραστικότητα. Για το λόγο αυτό προτείνουν ένα νευρωνικό δίκτυο με μοναδικό περιορισμό τη θετική έξοδο [WL19]. Επιλέγουν συνάρτηση ενεργοποίησης το Exponential Linear Unit αυξημένο κατά 1:

$$\mathrm{ELU}(x) = \begin{cases} x & x > 0 \\ \alpha(e^x - 1) & x \leq 0 \end{cases} \tag{5.5.8}$$

Στο πλαίσιο των αντιστρέψιμων μετασχηματισμών, εξετάζουμε τη συμπεριφορά του δικτύου του Sill. Έστω το πρόσθιο πέρασμα για μία μεταβλητή $y = f(x) \triangleq \min_{k\in[K]} \max_{j\in[J]} \{w_{k,j}x + b_{k,j}\}$. Τότε, ο αντίστροφος μετασχηματισμός έχει τον εξής κλειστό τύπο:

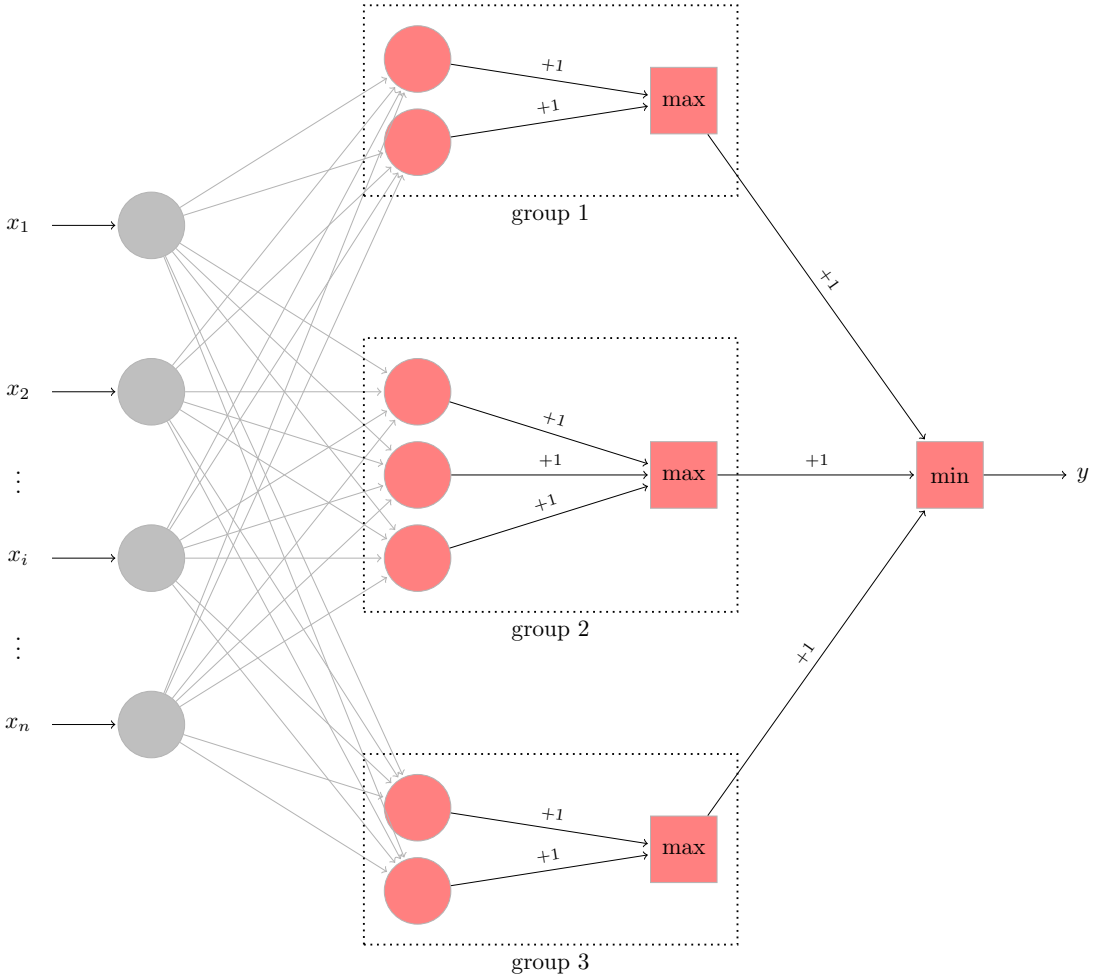
$$x = f^{-1}(y) = \max_{k\in[K]} \min_{j\in[J]} \{w_{k,j}^{-1}(y - b_{k,j})\} \tag{5.5.9}$$

Συνεπώς, ο πρόσθιος μετασχηματισμός αντιστοιχεί σε closing, ενώ ο αντίστροφος σχηματίζει το adjunction, ένα opening. Στις γενικότερες κατανομές που μελετώνται στο [WL19], η εύρεση κλειστού τύπου δεν αποτελεί βιώσιμη λύση. Ωστόσο, η συνάρτηση εξόδου $F$ αντιστοιχεί σε Αθροιστική Συνάρτηση Κατανομής με θετική συνάρτηση πυκνότητας πιθανότητας $f$. Άρα, η $F$ είναι *αυστηρά* μονότονη και δέχεται μοναδικό αντίστροφο σημείο για κάθε $x$ του πεδίου ορισμού. Η αντιστροφή μπορεί να υπολογιστεί με απλούς αλγορίθμους εύρεσης ρίζας, όπως η μέθοδος διχοτόμησης, χάρη στην ιδιαίτερη μορφή της καμπύλης που εγγυάται μοναδική λύση.

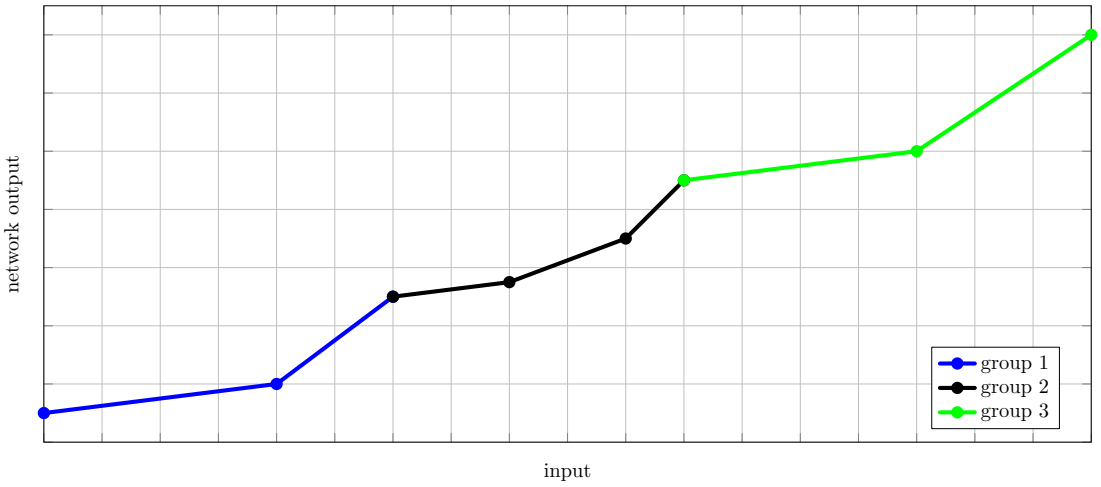
### 5.5.3 Τροπική ανάλυση Μονοτονικών Δικτύων

Η ανάλυση των monotonic networks διευκολύνεται από τη χρήση τροπικών μαθηματικών. Παρατηρούμε ότι οι κόμβοι που χρησιμοποιούν ως συνάρτηση ενεργοποίησης τη Maxout αντιστοιχούν σε τροπικά πολυώνυμα. Χρησιμοποιώντας το πολύτοπο Newton (βλ. 4) μπορούμε να αποφανθούμε για το πλήθος των γραμμικών περιοχών κάθε κόμβου. Παράλληλα, δύναται η απλοποίηση του πολυωνύμου μέσω της αφαίρεσης όρων που δεν συμβάλλουν στην τελική έκφραση της συνάρτησης εξόδου του κόμβου (βλ. παράδειγμα 5).

Εκτός από τη τροπική σκοπιά, παρόμοιες αρχιτεκτονικές έχουν προταθεί στον κόσμο των πλεγμάτων (lattice theory). Αναλυτικότερα, οι Tarela, Alonso, and Martinez πρότειναν ένα πρώιμο νευρωνικό δίκτυο που βασίζεται σε πλεγματικά πολυώνυμα (lattice polynomials) [TAM90] και έχει μορφή παρόμοια με αυτή των min-max δικτύων του Sill [Sil98] (βλ. σχήμα 5.5.1). Ωστόσο, αίρουν την απαίτηση αποκλειστικά θετικών βαρών και, επομένως, μονότονων επιφανειών και μελετούν πιο γενικά μοντέλα. Επιπλέον, στο δίκτυό τους, Boolean μεταβλητές καθορίζουν τη σύνδεση των κόμβων ανάμεσα στα διάφορα επίπεδα. Αντίθετα, ο Sill αναθέτει μέσω αρχιτεκτονικής σε κάθε max κόμβο μεμονωμένα βάρη $\mathbf{w_i}$.




Σχήμα 5.5.1: Monotonic Neural Network [Sil98]




Σχήμα 5.5.2: Η επιφάνεια που δημιουργείται από 3 groups. Παρατηρούμε ότι κάθε group είναι αντιστοιχεί σε ένα κυρτό τμήμα της καμπύλης και το σύνολο τους οδηγεί σε μία μη-κυρτή, μονότονη καμπύλη.

Επιπρόσθετα, σε επόμενη δουλειά τους [TM99] οι συγγραφείς προτείνουν κλάσεις διαμερίσεων του χώρου σύμφωνα με τις τομές των γραμμικών (τοπικών) συναρτήσεων που απαρτίζουν τη συνολική συνάρτηση. Έστω, λοιπόν, η τμηματικά γραμμική συνάρτηση $f : D \subset \mathbb{R}^n \to \mathbb{R}$. Ονομάζουμε *όψη* (face) κάθε $n$-διάστατο συνδεδμένο τμήμα γραμμικού στοιχείου της $f$ και *ακμή* (edge) κάθε $(n-1)$-διάστατο τμήμα ενός ορατού συνόρου δύο γραμμικών στοιχείων. Μελετούν 3 κατηγορίες περιοχών:

1. Έστω ότι η συνάρτηση $f$ αποτελείται από $L$ διακριτές τοπικές συναρτήσεις $f_l$. Τότε η **περιοχή προβολής** $\Omega_i$ απορρέει από την προβολή στο πεδίο ορισμού $D$ κάθε όψης της $f$ που χαρακτηρίζεται από την ίδια τοπική συνάρτηση.
2. Μία **περιοχή συνόρου** $\phi_j$ είναι μία περιοχή της διαμέρισης του πεδίου ορισμού που παράγεται από τη διαμόρφωση του συνόρου ορατών συνόρων $\{\phi_\lambda\}$.
3. Μία **περιοχή μοναδιαίας τάξης** (unique-order) $\Xi_k$ είναι μία περιοχή της διαμέρισης του πεδίου ορισμού που παράγεται από τη διαμόρφωση του συνόρου ορατών συνόρων $\{\phi_{\lambda_v}\}$ και κρυφών συνόρων $\{\phi_{\lambda_h}\}$: $\{\phi_{\lambda_v}\} \cup \{\phi_{\lambda_h}\}$.

Διαγράφεται μία ιεραρχία, καθώς οι περιοχές προβολών σχηματίζονται από την ένωση περιοχών συνόρων, οι οποίες σχηματίζονται από την ένωση περιοχών μοναδιαίας τάξης. Οι περιοχές προβολών αντιστοιχούν σε περιοχές όπου το τροπικό πολυώνυμο έχει μοναδική έκφραση: μία γραμμική συνάρτηση. Οι περιοχές προβολών ορίζουν διακριτές εκφράσεις της συνάρτησης. Εφόσον, η τελική συνάρτηση ενδέχεται να είναι μη-κυρτή, οι περιοχές προβολών μπορεί να είναι μη-κυρτές και μη-συνδεδεμένες. Αντιθέτως, οι περιοχές συνόρων και μοναδιαίας τάξης είναι πάντοτε κυρτές και συνδεδεμένες.

Η συγκεκριμένη διαμέριση του χώρου μελετάται στο πλαίσιο της κλάσης των συνεχών και τμηματικά γραμμικών συναρτήσεων, η οποία είναι υπερσύνολο των τροπικών καμπύλων. Αυτό συμβαίνει διότι οι τροπικές συναρτήσεις χαρακτηρίζονται από την ιδιότητα της κυρτότητας. Στην περίπτωση αυτή, το σύνολο των περιοχών προβολών είναι συνδεδεμένο και τα σύνορα τους σχηματίζουν την τροπική υπερεπιφάνεια.

## 5.6 Εκπαίδευση Μορφολογικών Δικτύων με μεθόδους Βελτιστοποίησης

### 5.6.1 Θεωρητικό Υπόβαθρο

Στο προηγούμενο κεφάλαιο παρουσιάστηκαν ο αλγόριθμος της οπισθοδιάδοσης (βλ. ενότητα 4.4) και οι αλγόριθμοι βελτιστοποίησης του βήματος ενημέρωσης των βαρών (βλ. ενότητα 4.5). Στις περιπτώσεις αυτές, η διαδικασία εκμάθησης των προτύπων είναι στοχαστική και δεν παρέχει εγγυήσεις βελτιστότητας του

αποτελέσματος, καθώς το πρόβλημα είναι μη-γραμμικό και μη-κυρτό. Ωστόσο, υπάρχουν περιπτώσεις ταξινομητών που η ανάθεση βαρών στις μεταβλητές μπορεί να πραγματοποιηθεί εξετάζοντας το πρόβλημα υπό τη σκοπιά της βελτιστοποίησης. Χαρακτηριστικό παράδειγμα αποτελούν οι Μηχανές Διανυσμάτων Υποστήριξης ή, αλλιώς, Support Vector Machines (SVMs). Τα SVMs αποτελούσαν την κατ'εξοχήν επιλογή σε προβλήματα Όρασης Υπολογιστών πριν την άνθηση των Νευρωνικών Δικτύων και, ιδιαίτερα, των βαθιών αρχιτεκτονικών. Η διατύπωσή τους αντιστοιχεί σε πρόβλημα κυρτής βελτιστοποίησης.

Έστω, λοιπόν, δύο κλάσεις $\mathcal{C}_0$ και $\mathcal{C}_1$, που είναι *γραμμικά διαχωρίσιμες*. Ο αλγόριθμος διανυσμάτων υποστήριξης αναζητά ένα διάνυσμα βαρών $\mathbf{w}$ και όρο πόλωσης $b$ που να διαχωρίζει τα πρότυπα των δύο κλάσεων. Εφόσον τα εν λόγω πρότυπα είναι γραμμικά διαχωρίσιμα, υπάρχουν κατάλληλα $\mathbf{w}, b$. Αναθέτοντας τις τιμές $-1, +1$ στις κλάσεις $\mathcal{C}_0, \mathcal{C}_1$, ο ταξινομητής για το πρότυπο $\mathbf{x}_i$ είναι:

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}_i + b \begin{cases} \geq +1 & y_i = +1 \\ < -1 & y_i = -1 \end{cases} \tag{5.6.1}$$

και ο ταξινομητής $\phi$ προκύπτει από την εφαρμογή της συνάρτησης προσήμου $\phi(\mathbf{x}) = \text{sgn}(f(\mathbf{x}))$. Τότε, το πρόβλημα για τη γραμμικά διαχωρίσιμη περίπτωση με $N$ πρότυπα λαμβάνει την ακόλουθη μορφή:

$$\begin{aligned} &\text{minimize} && \|\mathbf{w}\|_2^2 \\ &\text{subject to} && y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \quad i \in [N] \end{aligned} \tag{5.6.2}$$

Η παραπάνω διατύπωση, ωστόσο, δεν συμπεριλαμβάνει την περίπτωση μη-γραμμικά διαχωρίσιμων προτύπων που είναι σαφώς πιο γενική και εμφανίζεται σε πραγματικά δεδομένα. Ωστόσο, με μία μικρή τροποποίηση, δύναται η επίλυση και αυτού του πιο γενικού προβλήματος. Έστω οι βοηθητικές μεταβλητές (slack variables) $\xi_i \geq 0, i \in [N]$. Επιτρέπουμε την παραβίαση της συνθήκης για το πρότυπο $i$ το πολύ κατά $\xi_i$. Συνεπώς, η συνθήκη του προβλήματος (5.6.2) γίνεται $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i$. Όσο μεγαλύτερο το $\xi_i$ τόσο περισσότερο παραβιάζεται η αρχική συνθήκη. Επομένως, για την αποφυγή παθολογικών περιπτώσεων όπου επιλέγονται οι βοηθητικές μεταβλητές με τρόπο τέτοιο ώστε να ικανοποιούνται οι νέες συνθήκες αλλά ο ταξινομητής να μη γενικεύει σε test δεδομένα, πρέπει να ορίσουμε κάποια ποινή που να μην επιτρέπει υψηλές τιμές για τα $\xi_i$. Χρησιμοποιείται το άθροισμα των slack variables και ο ταξινομητής διανυσμάτων υποστήριξης για τη μη-διαχωρίσιμη περίπτωση λαμβάνει την ακόλουθη μορφή:

$$\begin{aligned} &\text{minimize} && \|\mathbf{w}\|_2^2 + C\sum_{i=1}^{N} \xi_i \\ &\text{subject to} && y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \quad i \in [N], \\ & && \xi_i \geq 0 \qquad\quad i \in [N] \end{aligned} \tag{5.6.3}$$

Με τη βοήθεια του hinge loss, το παραπάνω πρόβλημα βελτιστοποίησης μπορεί να εκφραστεί και ως πρόβλημα χωρίς περιορισμούς ως:

$$\text{minimize } \|\mathbf{w}\|_2^2 + \frac{C}{N}\sum_{i=1}^{N} \max\{0, \underbrace{1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)}_{\leq \xi_i}\} \tag{5.6.4}$$

Στην ενότητα αυτή, λοιπόν, αντλείται έμπνευση από τις παραπάνω ιδέες και εξετάζεται μία μέθοδος εκπαίδευσης Μορφολογικών δικτύων ριζωμένη στη θεωρία της Βελτιστοποίησης και, πιο συγκεκριμένα, στη διαδικασία Convex-Concave. Η μέθοδος αυτή προτάθηκε από τους Charisopoulos and Maragos [CM17] και μία επέκταση παρουσιάστηκε πρόσφατα στο [Val20].

Εστιάζουμε στην περίπτωση του απλού μορφολογικού δικτύου που απεικονίζεται στο σχήμα 5.3.1 με ένα νευρώνα διαστολής, ένα νευρώνα συστολής και ένα νευρώνα εξόδου. Συνεπώς, $n = m = c = 1$. Η έξοδος $y$ προκύπτει ως κυρτός συνδυασμός των εξόδων του κρυφού επιπέδου: διαστολή $\delta_{\mathbf{w}}$ και συστολή $\epsilon_{\mathbf{m}}$ όπου με $\mathbf{w}, \mathbf{m}$ συμβολίζουμε τις παραμέτρους των τροπικών perceptrons (βλ. (5.3.1), (5.3.2)). Έστω το σύνολο δεδομένων $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in$

$\mathbb{R}^n \times \{-1, +1\} : i \in [N]\}$, το οποίο διαχωρίζεται στα θετικά πρότυπα $\mathcal{C}_0$ και στα αρνητικά $\mathcal{C}_1$. Τότε, η ανάθεση βαρών στο τροπικό max-min perceptron γίνεται μέσω της επίλυσης του προβλήματος:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{N} v_i \max\{0, \xi_i\} \\ \text{subject to} \quad & \lambda\delta_{\mathbf{w}}(\mathbf{x}_i) + (1-\lambda)\epsilon_{\mathbf{m}}(\mathbf{x}_i) \geq -\xi_i \quad \forall \mathbf{x}_i \in \mathcal{C}_0, \\ & \lambda\delta_{\mathbf{w}}(\mathbf{x}_i) + (1-\lambda)\epsilon_{\mathbf{m}}(\mathbf{x}_i) \leq +\xi_i \quad \forall \mathbf{x}_i \in \mathcal{C}_1, \\ & \lambda \in (0, 1) \end{aligned} \tag{5.6.5}$$

Οι παράμετροι $v_i$ αντιστοιχίζουν βάρη στα πρότυπα [CM17]. Αναλυτικότερα, οι Charisopoulos and Maragos προτείνουν μία ευριστική μέθοδο για την ανάθεση των βαρών $v_i$. Χάριν ευκολίας, χρησιμοποιούμε τους μετρητές $i$ και $k$ για πρότυπα και κλάσεις αντίστοιχα. Τότε, το δειγματικό κέντρο της κάθε κλάσης είναι:

$$\mu_k = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \mathbf{x}_i$$

Στη συνέχεια, ορίζουν τον όρο $\theta_i$ που αντιστοιχεί στον αντίστροφο της απόστασης του προτύπου $i$ από το δειγματικό του κέντρο:

$$\theta_i = \frac{1}{\|\mathbf{x}_i - \mu_k\|}$$

Τότε, τα βάρη $v_i$ προκύπτουν κατόπιν κλιμάκωσης:

$$v_i = \frac{\theta_i}{\max_{j:j \in \mathcal{C}_k} \theta_j}$$

Μία ελαφρώς εναλλακτική προσέγγιση προτείνεται στο [Val20]. Ακολουθείται μία άπληστη (greedy) λογική που βασίζεται στην εύρεση των βαρών του κάθε perceptron ξεχωριστά, αρχικά για τον όρο διαστολής και μετά για τον όρο συστολής. Στη συνέχεια, ακολουθεί ο υπολογισμός της παραμέτρου $\lambda$. Η προσέγγιση αυτή επιτρέπει και τη χρήση όρου κανονικοποίησης. Για συνάρτηση απόφασης $f_{\mathbf{u}}$ το πρόβλημα είναι:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{N} v_i \max\{0, \xi_i\} + C\|\mathbf{u} - \mathbf{r}\|_1 \\ \text{subject to} \quad & f_{\mathbf{u}}(\mathbf{x}) \geq -\xi_i \quad \forall \mathbf{x}_i \in \mathcal{C}_0, \\ & f_{\mathbf{u}}(\mathbf{x}) \leq +\xi_i \quad \forall \mathbf{x}_i \in \mathcal{C}_1 \end{aligned} \tag{5.6.6}$$

οπου $\mathbf{r}$ μία μεταβλητή αναφοράς για το διάνυσμα βαρών $\mathbf{u}$. Συνεπώς, το πρόβλημα (5.6.6) επιλύεται δύο φορές:

- $f_{\mathbf{u}} = \delta_{\mathbf{w}}$ και μεταβλητή αναφοράς $\mathbf{r} = -\bigwedge \mathcal{C}_0$
- $f_{\mathbf{u}} = \epsilon_{\mathbf{m}}$ και μεταβλητή αναφοράς $\mathbf{r} = -\bigvee \mathcal{C}_1$

Αυτές οι επιλογές μεταβλητών αναφοράς επιτρέπουν στο perceptron να ταξινομήσει σωστά το μέγιστο δυνατό αριθμό θετικών και αρνητικών προτύπων από τους νευρώνες συστολής και διαστολής, αντίστοιχα. Εφόσον υπολογιστούν τα συναπτικά βάρη των perceptrons, η παράμετρος $\lambda$ καθορίζεται μέσω της ελαχιστοποίησης του μέσου hinge loss:

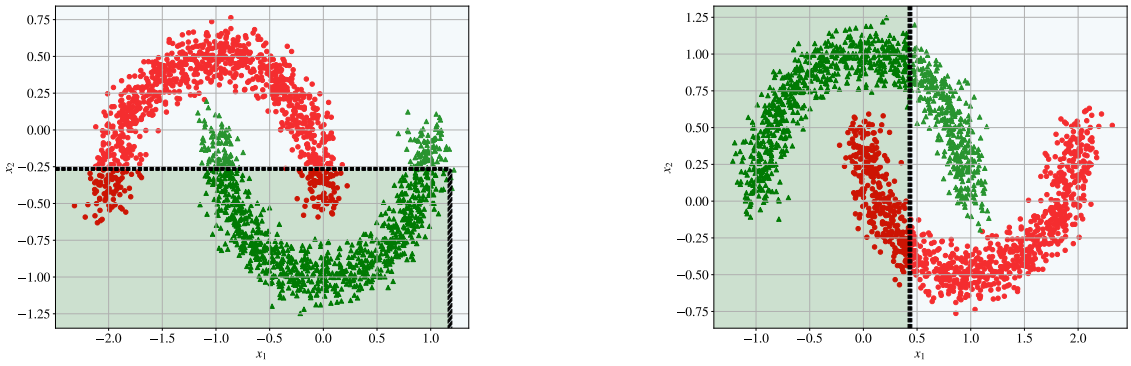
$$\lambda^* = \mathbf{argmin}_{0 \leq \lambda \leq 1} \sum_{i=1}^{N} \max\{0, -y_i \left[\lambda\delta_{\mathbf{w}}(\mathbf{x}_i) + (1-\lambda)\epsilon_{\mathbf{m}}(\mathbf{x}_i)\right]\} \tag{5.6.7}$$

Η παράμετρος $\lambda$ ελέγχει την ισορροπία μεταξύ των μορφολογικών perceptrons διαστολής και συστολής, οι οποίοι επικεντρώνονται στα δεδομένα της αρνητικής και της θετικής κλάσης αντίστοιχα.

### 5.6.2 Πειράματα Εκπαίδευσης με Convex-Concave Procedure

Εξετάζεται η μέθοδος στο σύνολο δεδομένων Double Moons από το βιβλίο του Haykin [Hay09, § 1.5]. Πρόκειται για δύο διεφθαρμένα με θόρυβο ημικύκλια, καθένα από τα οποία αντιστοιχεί σε μία κλάση. Επιλύοντας τα προβλήματα βελτιστοποίησης που περιγράφονται παραπάνω, λαμβάνουμε τα ακόλουθα αποτελέσματα: $\mathbf{m} = \begin{bmatrix} 2.19 & 0.34 \end{bmatrix}, \mathbf{w} = \begin{bmatrix} -1.20, 0.25 \end{bmatrix}$. Κατόπιν, υπολογίζεται το βέλτιστο $\lambda$ από την εξίσωση (5.6.7) και προκύπτει η τιμή $\lambda^* = 1$. Συνεπώς, ο βέλτιστος ταξινομητής συμπίπτει με νευρώνα διαστολής. Το αποτέλεσμα φαίνεται στο σχήμα 5.6.1. Το ποσοστό επιτυχίας είναι 84%.

Επαναλαμβάνουμε το πείραμα με μία διαισθητικά ανούσια αλλά μαθηματικά κρίσιμη αλλαγή: αντιστρέφουμε τις ετικέτες. Η κοινή λογική υπαγορεύει ότι αυτό δεν θα επηρεάσει το μοντέλο. Ωστόσο, η επίδραση είναι καταστροφική. Το προκύπτον σύνορο απόφασης απεικονίζεται στο σχήμα 5.6.1b, όπου το ποσοστό επιτυχίας είναι μόλις 67%.



(a) με κανονικές ετικέτες

(b) με αντιστροφή ετικετών

Σχήμα 5.6.1: Ταξινόμηση στο τεχνητό πρόβλημα Double Moons

Αυτό συμβαίνει διότι ο μεικτός ταξινομητής max-min perceptron βασίζεται στη θεωρία πλεγμάτων και, εξ ορισμού, προϋποθέτει μερική διάταξη των δεδομένων. Επιπλέον, είναι increasing ταξινομητής ως γραμμικός συνδυασμός δύο increasing νευρώνων. Άρα, τα πρότυπα της θετικής κλάσης οφείλουν να χαρακτηρίζονται από υψηλότερες τιμές, σε γενικές γραμμές, από τα αρνητικά πρότυπα.

Μία μέθοδος αντιμετώπισης αυτής της ανεπιθύμητης συμπεριφοράς είναι η χρήση μειωμένης διάταξης (5.1.9). Το αποτέλεσμα έχει λάβει το όνομα reduced Dilation-Erosion Perceptron (r-DEP) [Val20]. Επιλέγεται μία απεικόνιση $\rho : \mathbb{R}^n \to \mathbb{R}^m$ που μετασχηματίζει την είσοδο και ο ταξινομητής εφαρμόζεται στο νέο σύνολο δεδομένων. Αν το αρχικό σύνολο δεδομένων είναι το $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \{-1, +1\} : i \in [N]\}$, η απεικόνιση παράγει το νέο σύνολο δεδομένων $\mathcal{D}_{\text{new}} = \{(\rho(\mathbf{x}_i), y_i) \in \mathbb{R}^m \times \{-1, +1\} : i \in [N]\}$. Η απεικόνιση $\rho(\cdot)$ αποτελείται από $m$ επιμέρους απεικονίσεις $\rho(\mathbf{x}) = [\rho_1(\mathbf{x}), \rho_2(\mathbf{x}), \ldots, \rho_m(\mathbf{x})]^\top$ με $\rho_i(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}, i \in [m]$. Οι εν λόγω συναρτήσεις αποκαλούνται *πυρήνες* και χρησιμοποιούνται συχνά στο πλαίσιο των αλγορίθμων διανυσμάτων υποστήριξης. Ορισμένα παραδείγματα περιλαμβάνουν:

| Πυρήνας | $k(\mathbf{x}, \mathbf{y})$ |
|---|---|
| Linear | $\langle \mathbf{x}, \mathbf{y} \rangle$ |
| Polynomial | $\langle 1 + \mathbf{x}, \mathbf{y} \rangle^d$ |
| Gaussian | $e^{-\|\mathbf{x}-\mathbf{y}\|^2/(2\sigma^2)}$ |
| Sigmoid | $\tanh(\gamma \langle \mathbf{x}, \mathbf{y} \rangle + r)$ |

Πίνακας 5.2: Kernels

Σημειώνεται ότι οι πυρήνες με βάση την κανονική κατανομή αποκαλούνται Radial Basis Functions (RBFs). Η επιλογή των πυρηνών αποτελεί σημαντική πρόκληση και έγκειται σε σχεδιαστική απόφαση. Αυτή η προσέγγιση διαφέρει σε μία κρίσιμη πτυχή από την κλασική θεώρηση των νευρωνικών δικτύων, όπου η εκπαίδευση πραγματοποιείται με τον αλγόριθμο οπισθοδιάδοσης. Στο πλαίσιο της βαθιάς μηχανικής μάθησης, η επιλογή των

features και, συνεπώς, των εισόδων στο μοντέλο είναι απλή, καθώς το δίκτυο επωφελείται από τη συσσώρευση πληροφορίας ανά τα επίπεδα. Εδώ, αντιθέτως, η κατασκευή των χαρακτηριστικών καθορίζει άμεσα την απόδοση του μοντέλου, καθώς δεν υπάρχουν τα παραπάνω επίπεδα για τη αυτόματη εξαγωγή χαρακτηριστικών. Η προσέγγιση σε αυτή την περίπτωση είναι η χρήση πολυάριθμων πυρήνων με στόχο την αποτύπωση της κρυφής γνώσης. Ο συνδυασμός των πυρήνων μπορεί να γίνει με δύο τρόπους:

- Bagging: ίδιος πυρήνας και παράμετροι αλλά εκπαιδευμένοι σε διαφορετικά υποσύνολα του $\mathcal{D}_{\text{new}}$.
- Ensemble: διαφορετικοί πυρήνες εκπαιδευμένοι σε όλο το σύνολο δεδομένων $\mathcal{D}_{\text{new}}$.

Σε αντίθεση με τα νευρωνικά δίκτυα, η χρήση μεθόδων βελτιστοποίησης εγγυάται και γρήγορη σύγκλιση και ευστάθεια της μεθόδου. Αυτό ισχύει ιδιαίτερα για κυρτά προβλήματα, όπως Support Vector Machines, αλλά και σε ένα βαθμό για μη-κυρτά που επιλύονται με την ευριστική μέθοδο Convex-Concave.

Για την αξιολόγηση του r-DEP χρησιμοποιούμε δύο πυρήνες (για λόγους οπτικοποίησης του συνόρου απόφασης), τους οποίους συνδυάζουμε με τις δύο παραπάνω μεθόδους. Αναλυτικότερα, χρησιμοποιούνται πυρήνες Linear και RBF για Ensemble, ενώ στη μέθοδο Bagging αξιοποιούνται μόνο πυρήνες RBF. Στο σχήμα 5.6.2 παρουσιάζονται σχηματικά τα αποτελέσματα. Η πάνω σειρά αφορά τη μέθοδο Ensemble, ενώ η κάτω τη μέθοδο Bagging. Στην αριστερή στήλη απεικονίζεται το σύνορο απόφασης στο μετασχηματισμένο σύνολο δεδομένων $\mathcal{D}_{\text{new}}$, το οποίο αποτελεί μία τμηματικά γραμμική καμπύλη. Στη δεξιά στήλη, επιστρέφουμε στο αρχικό σύνολο δεδομένων. Και στις δύο περιπτώσεις, το ποσοστό επιτυχίας είναι 100%, σαφής βελτίωση από το 84% σε αυτό το ομολογουμένως απλό τεχνητό παράδειγμα. Επιπλέον, η μέθοδος είναι εύρωστη στην αλλαγή ετικετών.

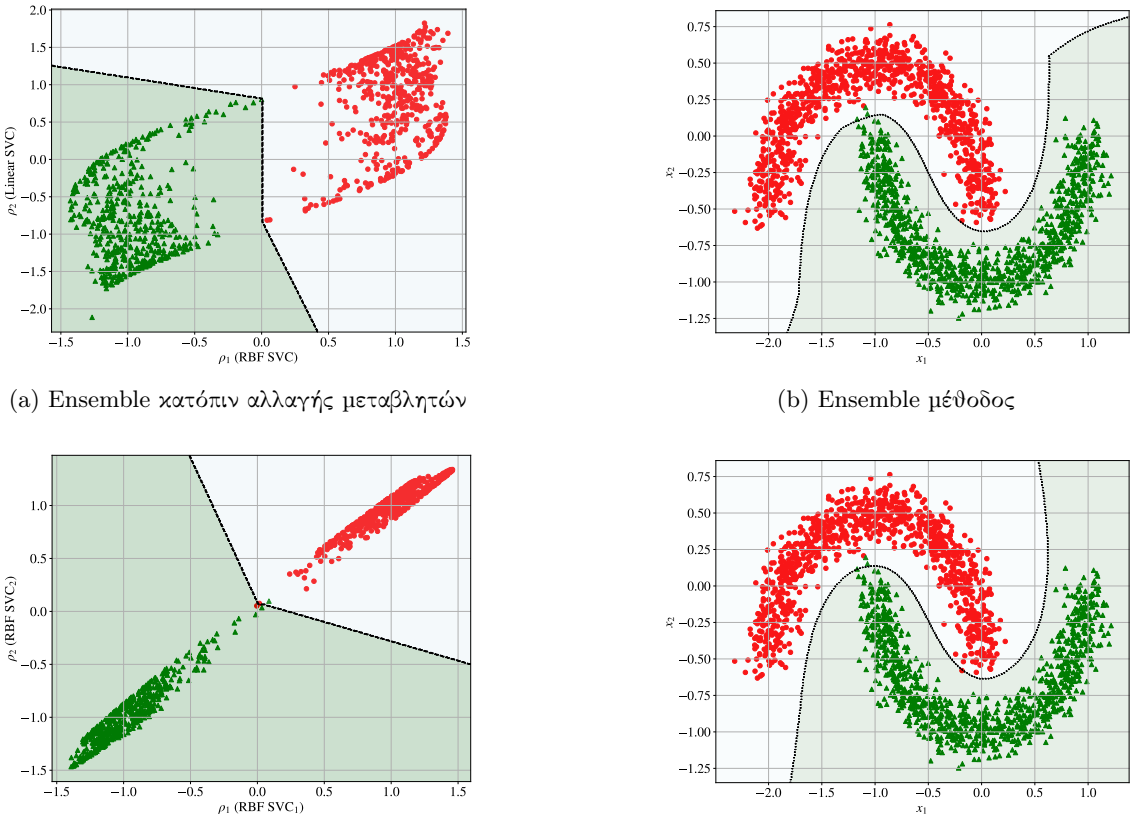
Συνεχίζουμε την πειραματική αξιολόγηση με ακόμα δύο σύνολα δεδομένων. Συγκεκριμένα, επιλέγουμε τα σύνολα δεδομένων Ripley's [Rip07] και Wisconsin Breast Cancer Dataset (WDBC) με στόχο τη σύγκριση με μεθόδους που αναπτύχθηκαν στο [CM17]. Αντλούμε τα αποτελέσματά τους και εξετάζουμε πώς οι ταξινομητές που μελετήθηκαν παραπάνω αποδίδουν.

Αρχικά, λίγα λόγια για τα σύνολα δεδομένων. Το σύνολο Ripley's αποτελείται από 2 κλάσεις, καθεμία από τις οποίες παράγεται με τυχαία δείγματα από ενα μείγμα Γκαουσιανών (Gaussian mixture) με δύο στοιχεία. Χωρίζεται σε 250 training και 1000 testing δεδομένα. Με στόχο το στατιστικά εύρωστο πειραματισμό, ενώνουμε τα δεδομένα training και test και εκτελούμε τα πειράματα με Statified K-Fold των 10 folds. Παρουσιάζουμε τα αποτελέσματα στον πίνακα 5.3. Ο ταξινομητής DEP αντιστοιχεί στο max-min perceptron από [CM17], ενώ με ετικέτα "greedy" επισημειώνουμε τον αντίστοιχο ταξινομητή όπου τα perceptrons διαστολής και συστολής εκπαιδεύονται ανεξάρτητα [Val20] και ενώνονται μετέπειτα σύμφωνα με τη σχέση (5.6.7). Παρατηρούμε ότι οι διάφοροι αλγόριθμοι έχουν εφάμιλλη απόδοση. Αυτό μπορεί να εξηγηθεί σχηματικά, γεγονός που είναι δυνατό αφού το σύνολο δεδομένων Ripley's έχει δύο χαρακτηριστικά. Συγκεκριμένα, στο σχήμα 5.6.3 φαίνονται τα σύνορα απόφασης για τους ταξινομητές r-DEP (Ensemble) και r-DEP (Bagging), τόσο στα αρχικά όσο και στα μετασχηματισμένα χαρακτηριστικά. Από τις εικόνες με τα αρχικά χαρακτηριστικά (βλ. 5.6.3b,5.6.3d) παρατηρούμε ότι μία τροπική καμπύλη μπορεί να διαχωρίσει ικανοποιητικά τα δεδομένα, γεγονός που αντικατοπτρίζεται στο ποσοστό επιτυχίας του DEP ταξινομητή.

Συνεχίζουμε με το σύνολο δεδομένων Wisconsin Breast Cancer Dataset (WDBC). Πρόκειται για πρότυπα για τη διάγνωση του καρκίνου του μαστού που διαχωρίζονται σε δύο κλάσεις, καλοηθείς και κακοηθείς όγκοι. Κάθε πρότυπο χαρακτηρίζεται από 30 features. Τα αποτελέσματα φαίνονται πάλι στον πίνακα 5.3. Είναι φανερό ότι η χρήση πυρήνων είναι ευεργετική, καθώς οι μειωμένοι (reduced) ταξινομητές παρουσιάζουν σημαντική βελτίωση.

| | Ripley's | WDBC |
|---|---|---|
| DEP | $0.902 \pm 0.001$ | $0.908 \pm 0.001$ |
| greedy DEP | $0.894 \pm 0.039$ | $0.912 \pm 0.049$ |
| r-DEP (Ensemble) | $0.889 \pm 0.035$ | $0.972 \pm 0.015$ |
| r-DEP (Bagging) | $0.907 \pm 0.038$ | $0.965 \pm 0.014$ |

Πίνακας 5.3: Πειραματική αξιολόγηση Dilation-Erosion Perceptron και παραλλαγών. Η πρώτη μέθοδος προέρχεται από [CM17], η δεύτερη από [Val20], ενώ οι μέθοδοι με μειωμένη διάταξη βασίζονται στο [Val20] με δικές μας επιλογές πυρήνων.




(a) Ensemble κατόπιν αλλαγής μεταβλητών

(b) Ensemble μέθοδος

(c) Bagging κατόπιν αλλαγής μεταβλητών

(d) Bagging μέθοδος

Σχήμα 5.6.2: Αξιολόγηση του r-DEP στο σύνολο δεδομένων Double Moons

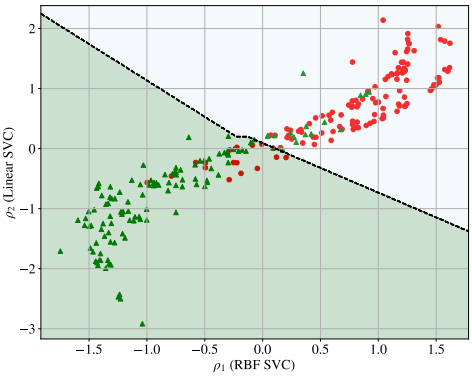
### 5.6.3 Επέκταση Εκπαίδευσης με Convex-Concave Procedure σε multiclass ταξινόμηση

Οι ταξινομητές που θεμελιώνονται στο Dilation-Erosion Perceptron αντιμετωπίζουν εν γένει προβλήματα δυαδικής φύσης. Ωστόσο, σχυνά καλούμαστε να επιλύσουμε διεργασίες που αφορούν $K > 2$ κλάσεις. Διάφορες μέθοδοι έχουν προταθεί στη βιβλιογραφία των Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines - SVMs), τις οποίες προσαρμόζουμε στο πλαίσιο του DEP. Ακολουθεί μία σύντομη ανασκόπηση των πιο σημαντικών μεθόδων.

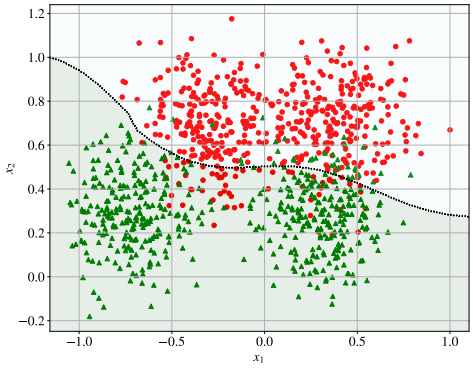
Μία συχνή προσέγγιση είναι η κατασκευή $K$ ξεχωριστών ταξινομητών, στην οποία το $k$-οστό μοντέλο $f_k(\cdot)$ εκπαιδεύεται με δεδομένα από την κλάση $\mathcal{C}_k$ ως θετικά πρότυπα και τις υπόλοιπες $K-1$ κλάσεις ως αρνητικά. Συνεπώς, η μέθοδος αυτή είναι γνωστή ως *one-verus-all* ή *one-versus-the-rest*. Όμως, αναδύονται προβλήματα, τα οποία εντοπίζονται στο γεγονός ότι η πρόβλεψη μπορεί να εμπεριέχει πολλούς ταξινομητές, οδηγώντας σε ασυνέπειες. Μία λύση είναι η λήψη απόφασης σύμφωνα με τον ταξινομητή που μεγιστοποιεί το περιθώριο $y(\mathbf{x}) = \max_k f_k(\mathbf{x})$. Ωστόσο, με αυτή την παραλλαγή οι ταξινομητές εκπαιδεύονται σε διαφορετικες διεργασίες.

Ένα ακόμα πρόβλημα της μεθόδου *one-verus-all* έγκειται στα διαφορετικά μεγέθη των συνόλων εκπαίδευσης. Ας υποθέσουμε ότι ένα σύνολο εκπαίδευσης με $N$ στοιχεία είναι διαχωρισμένο ομοιόμορφα σε $K$ κλάσεις. Τότε, τα θετικά πρόυπα είναι ίσα με $|\mathcal{C}_k| = \frac{N}{K}$, ενώ τα αρνητικά είναι ίσα με $|\mathcal{C}_{-k}| = \frac{(K-1)N}{K}$, επιφέροντας ασυμμετρία. Μία παραλλαγή είναι η χρήση ενός σχήματος ανάθεσης βαρών, όπου η θετική κλάση αντιστοιχίζεται σε βάρος $+1$ και τα στοιχεία της αρνητικής σε $-\frac{1}{K-1}$.

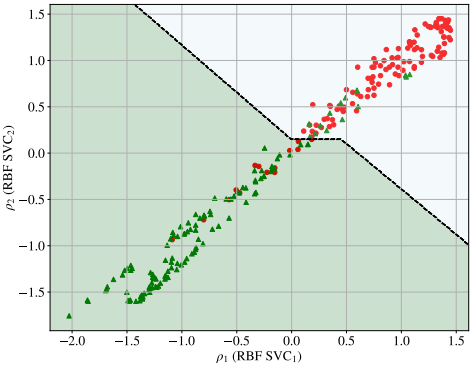
Μία διαφορετική προσέγγιση είναι η χρήση $\frac{K(K-1)}{2}$ δυαδικών ταξινομητών, επιλύοντας ένα πρόβλημα DC για κάθε ζευγάρι κλάσεων. Η μέθοδος αυτή ονομάζεται *one-verus-one*. Είναι σαφές ότι καθώς το πλήθος των κλάσεων αυξάνεται, το υπολογιστικό κόστος αυξάνεται σημαντικά. Για παράδειγμα, για 10 κλάσεις χρειάζονται 45 ταξινομητές. Η πρόβλεψη βασίζεται σε όλους όλους τους δυαδικούς ταξινομητές και αντιστοιχεί στην κλάση



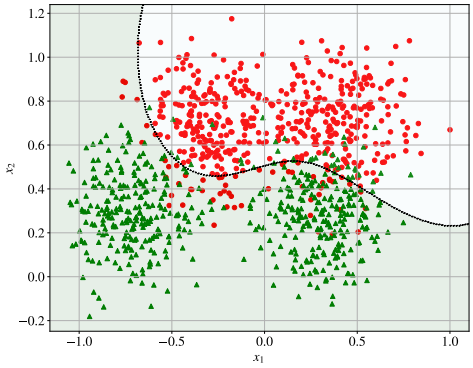

(a) Ensemble κατόπιν αλλαγής μεταβλητών




(b) Ensemble μέθοδος




(c) Bagging κατόπιν αλλαγής μεταβλητών




(d) Bagging μέθοδος

Σχήμα 5.6.3: Αξιολόγηση του r-DEP στο σύνολο δεδομένων Ripley's

που επιλέχθηκε τις περισσότερες φορές.

Ένας τρόπος αντιμετώπισης της υψηλής πολυπλοκότητας της μεθόδου *one-verus-one* είναι η οργάνωση των δυαδικών ταξινομητών σε ένα κατευθυνόμενο άκυκλο γράφο (Directed Acyclic Graph - DAG) ο οποίος αποτελείται μεν από $\frac{K(K-1)}{2}$ ταξινομητές αλλά απαιτείται η αξιολόγηση μόλις $K-1$ για τη λήψη απόφασης.

Εμπνευσμένοι από τις ιδέες αυτές, επεκτείνουμε το μοντέλο του δυαδικού ταξινομητή Dilation-Erosion Perceptron και των παραλλαγών του για διεργασίες ταξινόμησης πολλών κλάσεων. Υλοποιείται η προσέγγιση *one-verus-one* και αξιολογείται στα σύνολα δεδομένων `MNIST` και `FashionMNIST` . Για λεπτομερή περιγραφή των συγκεκριμένων συνόλων δεδομένων, ο αναγνώστης παραπέμπεται στις ενότητες 6.2.1 και 6.2.2, αντίστοιχα. Εν συντομία, τα δεδομένα χωρίζονται σε 60000 training και 10000 testing στοιχεία και αφορούν αναγνώριση ψηφίων. Επομένως, έχουμε $K = 10$ κλάσεις. Εξετάζουμε τη συμπεριφορά του Bagging ταξινομητή με πυρήνες Radial Basis Function για διαφορετικά πλήθη εκτιμητών $n$. Τα αποτελέσματα παρουσιάζονται παρακάτω.

| `MNIST` | | | | `FashionMNIST` | | | |
|---|---|---|---|---|---|---|---|
| $n=5$ | $n=10$ | $n=15$ | $n=20$ | $n=5$ | $n=10$ | $n=15$ | $n=20$ |
| 97.72 | 97.72 | 97.67 | 97.64 | 88.21 | 88.07 | 88.11 | 88.12 |

Πίνακας 5.4: Πειραματική αξιολόγηση r-DEP στα multiclass προβλήματα `MNIST` και `FashionMNIST`. Επιλέγεται μέθοδος Bagging με $n$ πυρήνες RBF.

Η μελέτη των reduced μορφολογικών τελεστών και η επέκτασή τους σε προβλήματα πολλών κλάσεων αναδεικνύει ένα πολύ σημαντικό ζήτημα: τη σχεδιαστική απόφαση που αφορά την επιλογή πυρήνων. Οι επιλογές είναι πολλές και απαιτείται βαθύτερος πειραματισμός με συνδυασμούς πυρήνων.

# Κεφάλαιο 6

# Πειραματική αξιολόγηση

## 6.1 Πειραματισμός με Μονότονες και Μη-Κυρτές Συναρτήσεις

Ο πειραματισμός σε μονότονες κατανομές πραγματοποείται από το ομώνυμα δίκτυα του προηγούμενου κεφαλαίου. Η αξιολόγηση αφορά τόσο τεχνητά δεδομένα όσο και πραγματικά. Ωστόσο, πριν εντρυφήσουμε σε υλοποιήσεις με επίκεντρο τα νευρωνικά δίκτυα, προσφέρουμε μία αλγεβρική επίλυση του προβλήματος παλινδρόμησης (regression).

### 6.1.1 Παρουσίαση των Datasets

Μία μεγάλη οικογένεια προβλημάτων, τόσο ταξινόμησης όσο και παλινδρόμησης, χαρακτηρίζονται από συνθήκες μονοτονίας. Για παράδειγμα, η τιμή ενός αυτοκινήτου μειώνεται με την ηλικία του αυτοκινήτου, ενώ η τιμή ενός ακινήτου αυξάνεται με το μέγεθος σε τετραγωνικά μέτρα $m^2$, ενώ το bond rating μίας εταιρίας αυξάνεται με το debt to capital ratio. Παρόμοιες σχέσεις εγείρονται και στην ιατρική επιστήμη: η αύξηση του σωματικού βάρους οδηγεί σε αύξηση παθήσεων της καρδιάς, καρκίνο κ.α. Χρησιμοποιώντας μοντέλα που συμπεριλαμβάνουν δομικά αυτό τον περιορισμό της μονοτονίας, προκύπτουν εργαλεία με υψηλότερη ακρίβεια και χαμηλότερη διασπορά σε σχέση με μη-μονότονα μοντέλα [DV10].

#### Τεχνητά δεδομένα (Regression)

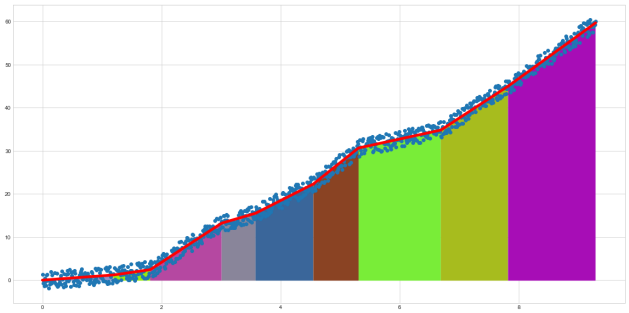
Σκοπός μας είναι η δημιουργία δεδομένων μορφής παρόμοιας με την εικόνας 5.5.2. Ακολουθούμε τη διαδικασία που σκιαγραφούν οι Tarela, Alonso, and Martinez [TM99] για την κατασκευή μονότονης συνάρτησης στο $\mathbb{R}$. Δεδομένου ενός σημείου, επιλέγουμε τυχαία μία (θετική) κλίση και υπολογίζουμε το bias ώστε να συνεχίζει το ευθύγραμμο τμήμα από το προηγούμενό του. Τέλος, επιλέγουμε ένα τυχαία interval οπου το συγκεκριμένο ευθύγραμμο να είναι dominant. Αυτό το interval αντιστοιχεί στην περιοχή προβολής στην εν λόγω διάσταση [TAM90]. Επιλέγοντας ολοένα και μεγαλύτερες κλίσεις για ένα δεδομένο πλήθος τμημάτων $k$, κατασκευάζουμε κυρτές ομάδες με $k$ τμήματα, σύμφωνα με την ορολογία του Sill.

Αφότου έχουμε κατασκευάσει την ευθεία, επιλέγουμε τυχαία σημεία και προσθέτουμε θόρυβο $e$ μηδενικής μέσης τιμής. Έγκυρες επιλογές περιλαμβάνουν $e \sim U[-a, a]$ ή $e \sim \mathcal{N}(0, \sigma^2)$. Επεκτείνοντας αυτή τη λογική σε περισσότερες διαστάσεις, κατασκευάζουμε τεχνητό dataset σε πολλές διαστάσεις. Παρακάτω παρουσιάζεται ο αλγόριθμος 4.

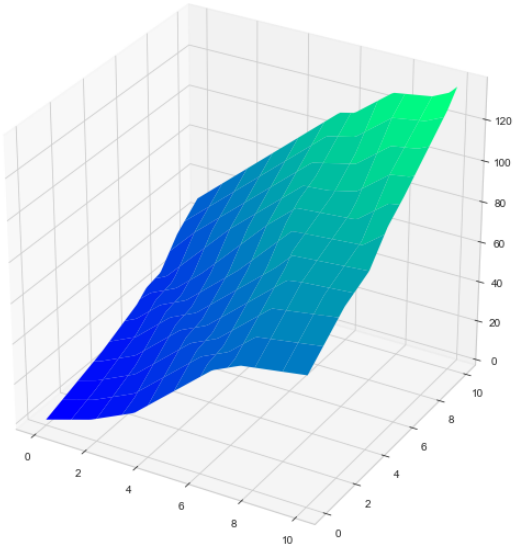
**Algorithm 4:** Artificial Dataset construction

**Data:** number of groups $G$, number of hyperplanes in group $H$, noise $\epsilon \sim \mathcal{X}$, samples $N$, dimensions $d$
**Result:** monotonic dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$

set $y = 0$
**for** $g = 1, \ldots, G$ **do**
  set slope $\mathbf{a}_{g,0} = \mathbf{0} \in \mathbb{R}^d$
  **for** $h = 1, \ldots, H$ **do**
    select $\mathbf{a}_{g,h}$ such that $\mathbf{a}_{g,h} \succcurlyeq \mathbf{a}_{g,h-1}$
    select interval $\Delta\mathbf{x}_{g,h} \in \mathbb{R}_+^d$
    compute bias $b_{g,h} = y - \mathbf{a}_{g,h}^\top \Delta\mathbf{x}_{g,h}$
    update $y \leftarrow y + \mathbf{a}_{g,h}^\top \Delta\mathbf{x}_{g,h}$
  **end**
**end**
construct PWL function $f : \mathbb{R}^d \to \mathbb{R}$ such that $f(\mathbf{x}) = \min_{g \in [G]} \max_{h \in [H]} \mathbf{a}_{g,h}^\top \mathbf{x} + b_{g,h}$
**for** $i = 1, \ldots, N$ **do**
  sample $\mathbf{x}_i \sim \mathcal{U}[\mathbf{0}, \sum_{g,h} \Delta\mathbf{x}_{g,h}]$
  $y_i = f(\mathbf{x}_i) + \epsilon$
**end**
**return** $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$



(a) Παράδειγμα σε 1 διάσταση $G = H = 3$



(b) Παράδειγμα σε 2 διαστάσεις $G = H = 3$

Σχήμα 6.1.1: Παραδείγματα του τεχνητού συνόλου δεδομένων

### 6.1.2 Τροπική Μη-Κυρτή Παλινδρόμηση

Μελετούμε το πρόβλημα της μονότονης παλινδρόμησης από τη σκοπιά της τροπικής άλγεβρας. Αναλυτικότερα, έστω ότι προσπαθούμε να λύσουμε το πρόβλημα

$$\mathbf{A} \boxplus \mathbf{x} = \mathbf{b} \tag{6.1.1}$$

όπου υπενθυμίζεται ότι με $\boxplus$ συμβολίζουμε τον πολλαπλασιασμό πινάκων υπό τη max-plus έννοια. Το πρόβλημα 6.1.1 ενδέχεται να μην έχει ακριβή λύση. Στην περίπτωση αυτή, αποσκοπούμε στη λύση του προβλήματος:

$$\begin{array}{ll} \text{minimize} & \|\mathbf{A} \boxplus \mathbf{x} = \mathbf{b}\|_p \\ \text{subject to} & \mathbf{A} \boxplus \mathbf{x} \leq \mathbf{b} \end{array} \tag{6.1.2}$$

όπου με $\|\cdot\|_p$ συμβολίζεται η $\ell_p$ νόρμα για $p = 1, 2 \ldots, \infty$. Στον τροπικό ημιδακτύλιο, οι λύσεις του παραπάνω προβλήματος υπολογίστηκαν στην πρωταρχική μελέτη του κλάδου από τον Cuninghame-Green [Cun79] για τις ειδικές περιπτώσεις $p - 1$ και $p = \infty$.

**Theorem 6.1.1: Επίλυση max-plus συστήματος [MT19]**

Αν το πρόβλημα (6.1.2) έχει λύση, τότε η μέγιστη λύση του δίνεται από τον πίνακα Cuninghame-Greene:

$$\hat{\mathbf{x}} = \mathbf{A}^* \boxplus' \mathbf{b} = \left[ \bigwedge_i b_i - a_{ij} \right] \tag{6.1.3}$$

με $\mathbf{A}^* = -\mathbf{A}^\top$. Επιπλέον, για την περίπτωση των νορμών $\ell_1$ και $\ell_\infty$ η παραπάνω λύση (6.1.3) είναι και μοναδική.

Το πρόβλημα της παλινδρόμησης είναι κλασικό στη βιβλιογραφία της στατιστικής και της μηχανικής μάθησης. Κεντρική σημασία έχει το πρόβλημα των *ελαχίστων τετραγώνων* (least squares estimate) όπου μία ευθεία $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b$ προσαρμόζεται σε ένα σύνολο δεδομένων με στόχο την ελαχιστοποίηση της $\ell_2$ νόρμας. Η λύση αυτού του προβλήματος πιστώνεται στον Carl Friedrich Gauss και είναι γνωστή από τον 19ο αιώνα.

Όπως έχουμε αναφέρει και στις προηγούμενες ενότητες, η έννοια της ευθείας στα τροπικά μαθηματικά διαφέρει. Ας εξετάσουμε την απλή περίπτωση της μίας μεταβλητής. Η ευθεία στα κλασικά μαθηματικά είναι $f(x) = ax + b$

και η τροπικοποίησή (tropicalization) της είναι $f(x) = \max(a + x, b)$. Συνεπώς, το πρόβλημα ανάγεται στη βελτιστοποίηση δυο μεταβλητών $\mathbf{w} = (a, b)$. Έστω σύνολο δεδομένων $\mathcal{D} = \{(x_i, f_i) \in \mathbb{R}^2\}_{i=1}^m$. Τότε, το πρόβλημα λαμβάνει τη μορφή $\mathbf{X} \boxplus \mathbf{w} = \mathbf{f}$ και:

$$\mathbf{X} \boxplus \mathbf{w} = \mathbf{f} \implies \hat{\mathbf{w}} = \mathbf{X}^* \boxplus' \mathbf{f} \tag{6.1.4}$$

$$\text{ή} \quad \begin{bmatrix} x_1 & 0 \\ \vdots & \vdots \\ x_m & 0 \end{bmatrix} \boxplus \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} f_1 \\ \vdots \\ f_m \end{bmatrix} \implies \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} \bigwedge_i f_i - x_i \\ \bigwedge_i f_i \end{bmatrix} \tag{6.1.5}$$

Η λύση του προβλήματος $\hat{\mathbf{w}}$ αντιστοιχεί σε μία προσαρμογή (fit) στα δεδομένα ώστε να παρέχει το ελάχιστο σφάλμα από τις κάτω προσεγγίσεις. Οι [MT19] χρησιμοποιούν τον όρο *greatest lower estimate (GLE)*. Σύμφωνα με την ενότητα 5.1, ο max-plus πολλαπλασιασμός πινάκων αντιστοιχεί σε διανυσματική διαστολή $\delta_{\mathbf{A}}(\mathbf{x}) = \mathbf{A} \boxplus \mathbf{x}$, ενώ ο min-plus πολλαπλασιασμός σε διανυσματική συστολή $\epsilon_{\mathbf{A}}(\mathbf{x}) = \mathbf{A} \boxplus' \mathbf{x}$. Από τις συνθήκες του προβλήματος (6.1.2), συμπεραίνουμε ότι τα διανύσματα που επιλέγονται αντιστοιχούν σε υπολύσεις υπό την έννοια ότι $\delta_{\mathbf{A}}(\mathbf{x}) = \mathbf{A} \boxplus \mathbf{x} \leq \mathbf{b}$ και επιλέγεται η μέγιστη υπολύση $\hat{\mathbf{x}} = \epsilon_{\mathbf{A}}(\mathbf{b})$. Άρα, η βέλτιστη λύση του max-plus προβλήματος (6.1.2) αντιστοιχεί σε ένα opening που προσεγγίζει το διάνυσμα $\mathbf{b}$ από κάτω $\delta_{\mathbf{A}}(\epsilon_{\mathbf{A}}(\mathbf{b})) \leq \mathbf{b}$.

Η λύση GLE αντιστοιχεί στο πρόβλημα με περιορισμούς (6.1.2). Για την ειδική περίπτωση της $\ell_\infty$, υπάρχει λύση στο πρόβλημα της ελαχιστοποίησης της εν λόγω νόρμας *χωρίς* περιορισμούς. Η λύση έγκειται στην εύρεση ενός βαθμωτού μεγέθους $\mu$ που αντιστοιχεί σε μετατόπιση της λύσης (6.1.3). Για $\mu = \frac{1}{2}\|\mathbf{A} \boxplus \hat{\mathbf{x}} - \mathbf{x}\|_\infty = \frac{1}{2}\|\mathbf{A} \boxplus (\mathbf{A}^* \boxplus' \mathbf{x}) - \mathbf{x}\|_\infty$, η λύση του προβλήματος χωρίς περιορισμούς είναι:

$$\overline{\mathbf{x}} = \mu + \hat{\mathbf{x}} = \mu + \mathbf{A}^* \boxplus' \mathbf{x} \tag{6.1.6}$$

Η λύση $\overline{\mathbf{x}}$ αποκαλείται *Minimum Max Absolute Error* ή *MMAE* εν συντομία [MT19]. Στο παρακάτω σχήμα παρουσιάζεται μία απλή περίπτωση της παλινδρόμησης σε περιπτώσεις θορύβου που δειγματολειπτείται είτε από κανονική κατανομή μηδενικής μέσης τιμής είτε από ομοιόμορφη κατανομή. Στην πρώτη σειρά μελετάται το max-plus πρόβλημα για τη συνάρτηση $\max\{b, a + x\}$ και στη δεύτερη σειρά το αντίστοιχο min-plus. Συνεπώς, αναζητούνται οι τιμές $a$ και $b$.

Η παραπάνω προσέγγιση μπορεί να εφαρμοστεί και σε πιο γενικά προβλήματα: πολυώνυμα υψηλών βαθμών σε πολλές διαστάσεις. Έστω, λοιπόν, το max-plus πολυώνυμο:

$$f(\mathbf{x}) = \bigvee_{i=1}^{K} \mathbf{a}_k^\top \mathbf{x} + b_k$$

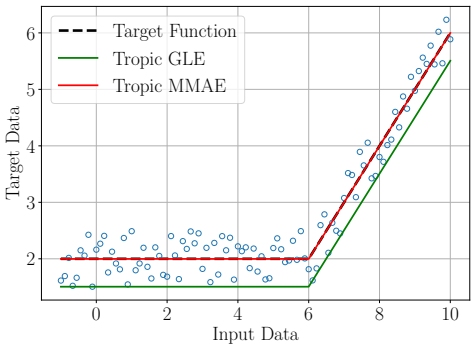
και τα δεδομένα $\mathcal{D} = \{(\mathbf{x}_i, f_i) \in \mathbb{R}^n \times \mathbb{R}\}_{i=1}^m$. Οι εξισώσεις $\mathbf{X} \boxplus' \mathbf{w} = \mathbf{f}$ λαμβάνουν τη μορφή:

$$\begin{bmatrix} \mathbf{a}_1^\top \mathbf{x}_1 & \mathbf{a}_2^\top \mathbf{x}_1 & \dots & \mathbf{a}_K^\top \mathbf{x}_1 \\ \mathbf{a}_1^\top \mathbf{x}_2 & \mathbf{a}_2^\top \mathbf{x}_2 & \dots & \mathbf{a}_K^\top \mathbf{x}_2 \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{a}_1^\top \mathbf{x}_m & \mathbf{a}_2^\top \mathbf{x}_m & \dots & \mathbf{a}_K^\top \mathbf{x}_m \end{bmatrix} \boxplus \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix}$$

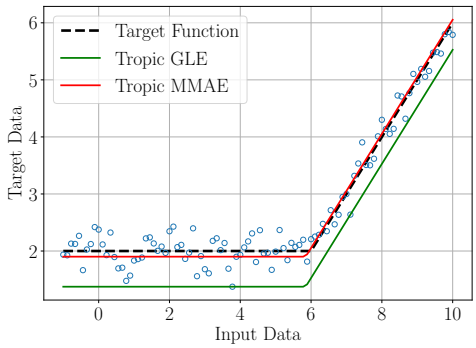
και η λύση είναι:

$$\begin{bmatrix} \hat{b}_1 \\ \vdots \\ \hat{b}_K \end{bmatrix} = \hat{\mathbf{w}} = \mathbf{X}^* \boxplus' \mathbf{f} = \begin{bmatrix} \bigwedge_{i=1}^m f_i - \mathbf{a}_1^\top \mathbf{x}_i \\ \vdots \\ \bigwedge_{i=1}^m f_i - \mathbf{a}_K^\top \mathbf{x}_i \end{bmatrix}$$

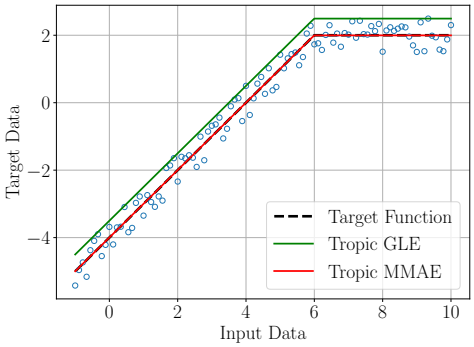
Από τις παραπάνω εξισώσεις φαίνεται ότι οι κλίσεις $\mathbf{a}_i$ θεωρούνται γνωστές και οι μεταβλητές βελτιστοποίησεις έγκεινται στους όρους bias $b_i$. Αυτή η υπόθεση φαίνεται περιοριστική, αλλά μπορεί να αντιμετωπιστεί εύκολα. Αναλυτικότερα, μπορούμε να θεωρήσουμε ότι το διάνυσμα $\mathbf{a}_i$ λαμβάνει όλες τις ακέραιες τιμές έως ένα μέγιστο βαθμό [HKA16; MT19]. Αυτό προϋποθέτει ότι τα δεδομένα μας χαρακτηρίζονται από κάποιο μέτρο ομαλότητας.



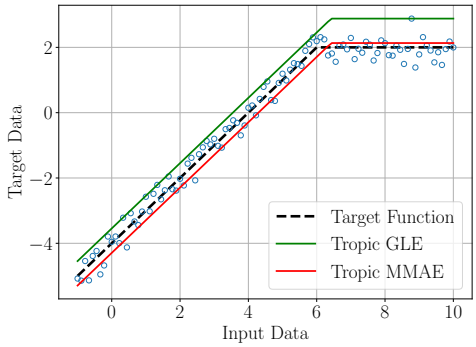

(a) max-plus με Uniform noise




(b) max-plus με Gaussian noise




(c) min-plus με Uniform noise




(d) min-plus με Gaussian noise

Σχήμα 6.1.2: Παραδείγματα Τροπικής Παλινδρόμησης

Για $L$-smooth συνάρτηση, χρησιμοποιούμε κλίσεις έως $\lceil L \rceil$. Η τροπική παλινδρόμηση που σκιαγραφείται παραπάνω μπορεί να εφαρμοστεί και σε min-plus δεδομένα. Τότε, αναφερόμαστε σε κοίλες συναρτήσεις και η βέλτιστη λύση αντιστοιχεί σε διανυσματικό closing.

Πέρα από την τυφλή χρήση κλίσεων, δύναται η αξιοποίηση αλγορίθμων clustering για την εύρεση κλίσεων που ταιριάζουν καλά στα δεδομένα. Μέσα από το σύνολο $\mathcal{D}$, υπολογίζονται οι αριθμητικές παράγωγοι (numerical gradients) και χρησιμοποιείται ο αλγόριθμος συσταδοποίησης $k$-means, ο οποίος παρουσιάζεται παρακάτω.

**Algorithm 5:** $K$-means

**Data:** number of centroids $K$, dataset $\mathcal{D} = \{(\mathbf{x}_i) \in \mathbb{R}^p\}_{i=1}^m$
**Init:** initialize *cluster centroids* $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^p$ randomly

**do**

Assign datapoints to cluster centroids:

$$c^{(i)} \leftarrow \arg\min_{j \in [K]} \|\mathbf{x}^{(i)} - \mu_j\|_2, \quad \forall i \in [m]$$

Update cluster centroids:

$$\mu_j \leftarrow \frac{\sum_{i=1}^{m} \mathbb{1}\{c^{(i)} = j\}\mathbf{x}^{(i)}}{\sum_{i=1}^{m} \mathbb{1}\{c^{(i)} = j\}}, \quad \forall j \in [K]$$

**until** *convergence*

**return** cluster centroids $\mu_1, \mu_2, \ldots, \mu_K$

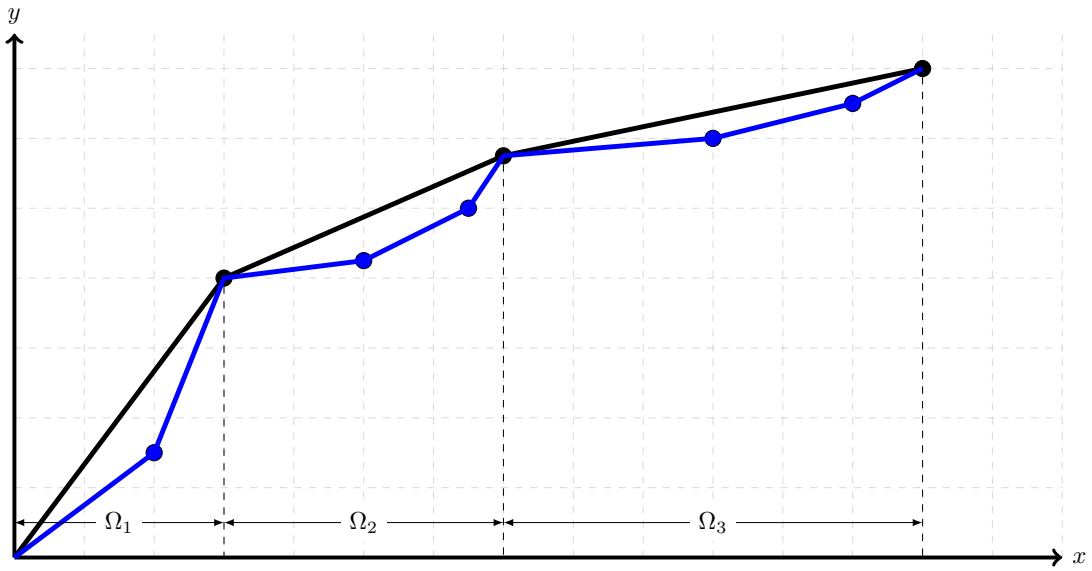
Έστω πως από το σύνολο δεδομένων $\mathcal{D} \subset \mathbb{R}^p$ προκύπτουν τα διανύσματα $\delta_i \in \mathbb{R}^p$. Ο αλγόριθμος προσπα-

θεί να τα διαχωρίσει σε $K$ συστάδες μέσω μίας επαναληπτικής διαδικασίας δύο βημάτων. Η παράμετρος $K$ επιλέγεται από το χρήστη. Στο πρώτο βήμα αναθέτει κάθε σημείο στη συστάδα με τη μικρότερη απόσταση. Στο δεύτερο βήμα, ενημερώνεται το κέντρο της συστάδας λαμβάνοντας υπόψιν *μόνο* τα σημεία που έχουν ανατεθεί σε αυτή. Η διαδικασία επαναλαμβάνεται ως τη σύγκλιση, δηλαδή μέχρι να μην επέλθει μεταβολή των κέντρων $\mu_i, i \in [K]$. Ο αλγόριθμος εξαρτάται σε μεγάλο βαθμό από την αρχικοποίηση των κέντρων. Στην πράξη, λοιπόν, χρησιμοποιείται η παραλλαγή `k-means++`, η οποία εκτελεί πολλές φορές τον αλγόριθμο με διάφορες αρχικοποιήσεις και επιλέγει την καλύτερη λύση. Ο αλγόριθμος $K$-means αποτελεί ειδική περίπτωση του Expectation-Maximization, όπου η ανάθεση σε συστάδα δε γίνεται με δείκτρια συνάρτηση (hard assignment) αλλά με στατιστικούς τρόπους, αναθέτοντας σε κάθε cluster βάρη που αντιστοιχούν σε πιθανότητες το σημείο να ανήκει στην αντίστοιχη συστάδα (soft assignment).

Σε περίπτωση που τα δεδομένα μας είναι μονότονα, μπορούμε να προσαρμόσουμε την παραπάνω τεχνική επιτρέποντας μόνο θετικές κλίσεις. Αυτό γίνεται περιορίζοντας τις υποψήφιες κλίσεις στους φυσικούς αριθμούς, αντί για τους ακεραίους. Στην περίπτωση που οι κλίσεις υπολογίζονται με αλγόρίθμους συσταδοποίησης, προτείνεται ο περιορισμός των κλίσεων που προκύπτουν σε θετικές τιμές. Επιπλέον, ενδέχεται τα δεδομένα μας να *μην* έχουν αμιγώς κυρτή (ή αμιγώς κοίλη) μορφή. Τότε, για την καλύτερη προσαρμογή της ευθείας στα δεδομένα, δύναται η εφαρμογή ενός δεύτερου "περάσματος" στα δεδομένα. Για σχηματική επεξήγηση της ιδέας, βλ. σχήμα 6.1.3:

1. $(\min, +)$ πέρασμα (μαύρο χρώμα). Το αποτέλεσμα είναι μία αύξουσα, κοίλη και τμηματικά γραμμική συνάρτηση η οποία αντιστοιχεί σε ένα projection set $\Omega_i$ [TM99].

2. $(\max, +)$ πέρασμα (μπλε χρώμα). Το ευθύγραμμο τμήμα του αντίστοιχου projection set $\Omega_i$ αντικαθίσταται με ένα $(\max, +)$ πολυώνυμο.




Σχήμα 6.1.3: Σχηματική απεικόνιση tropical min-max regression

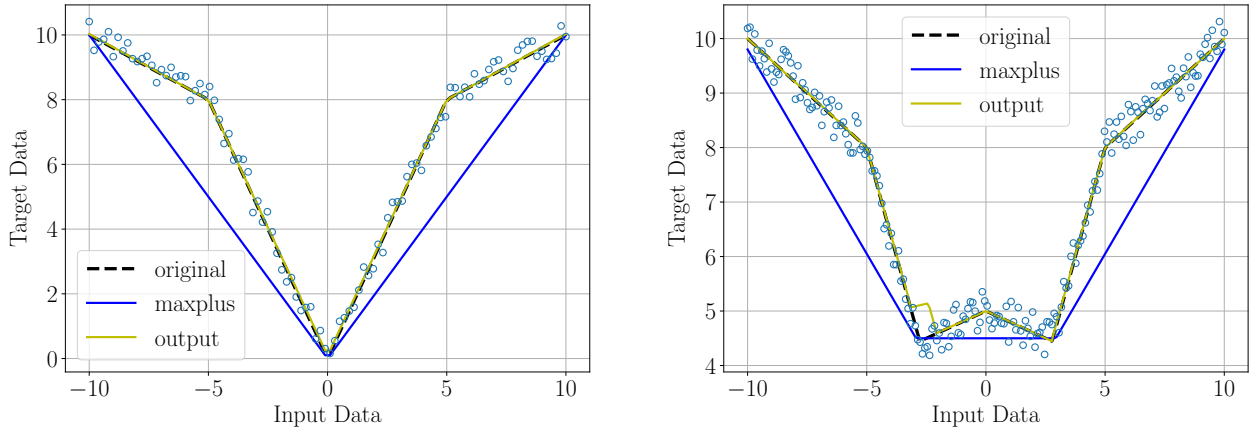
Συνεπώς, αυτός ο αλγόριθμος δύο βημάτων διαχωρίζει το πεδίο ορισμού σε σύνολα προβολής $\Omega_i$ στο αρχικό πέρασμα και κατόπιν προσπαθεί να προσεγγίσει τα δεδομένα του $\Omega_i$ με περισσότερη ακρίβεια. Με παρόμοιο τρόπο, οι Magnani and Boyd διαχωρίζουν το πεδίο ορισμού και προσαρμόζουν αφφινικές συναρτήσεις σε κάθε υποσύνολο, λαμβάνοντας το max τους ως έξοδο, με αποτέλεσμα να προκύπτει μία τμηματικά γραμμική και κυρτή προσέγγιση των δεδομένων.

Συνεχίζουμε με μερικά απλά παραδείγματα για την πειραματική αξιολόγηση αυτής της "διμερούς" παλινδρόμησης. Εξετάζονται δύο συμμετρικές συναρτήσεις, οι απεικονίσεις των οποίων φαίνονται στο σχήμα 6.1.4. Με μαύρη διακεκομμένη γραμμή φαίνεται η αρχική συνάρτηση, γύρω από την οποία υπάρχουν τα σημεία προς εξέταση που προκύπτουν κατόπιν πρόσθεσης θορύβου. Με μπλε γραμμή διαγράφεται ο βέλτιστος διαχωρισμός των περιοχών μετά από ένα max-plus πέρασμα. Σημειώνεται ότι αυτό δεν είναι το αποτέλεσμα που παράγει στην

πραγματικότητα το πρώτο πέρασμα. Με πράσινο απεικονίζεται η (πραγματική) έξοδος του διμερούς αλγορίθμου. Και στις δύο περιπτώσεις χρησιμοποιείται ο αλγόριθμος $K$-means για τον προσδιορισμό των κλίσεων.

Στο πρώτο παράδειγμα εξετάζεται η συνάρτηση $f_1(x) = (-0.4x + 6 \wedge -1.6x) \vee (0.4x + 6 \wedge 1.6x)$, η οποία απεικονίζεται στο σχήμα 6.1.4a. Προστίθεται θόρυβος $\mathcal{U}[-0.5, 0.5]$ και επιλέγεται $K = 2$ για το πρώτο πέρασμα, ώστε να εξακριβωθεί η δυνατότητα διαχωρισμού των κλίσεων σε δύο διακριτά σύνολα, θετικές και αρνητικές κλίσεις. Στη συνέχεια και εφόσον αναγνωρισθμούν οι δύο περιοχές προβολών, οι οποίες συμπίπτουν με θετικούς και αρνητικούς αριθμούς στην προκειμένη περίπτωση, τα δεδομένα επεξεργάζονται με min-plus πέρασμα. Παρατηρούμε ότι ο αλγόριθμος επιστρέφει την πραγματική ευθεία (η πράσινη ευθεία συμπίπτει με τη μαύρη).

Στο δεύτερο παράδειγμα (σχήμα 6.1.4b), η συνάρτηση $f_1(x)$ εμπλουτίζεται με έναν min-plus όρο και προκύπτει η $f_2(x) = f_1(x) \vee (5 - 0.2x \wedge 5 + 0.2x)$. Ο όρος αυτός διακόπτει τη συμμετρία των κλίσεων (θετικές κλίσεις για θετικές τετμημένες και αρνητικές κλίσεις για αρνητικές). Συνεπώς, η αναγνώριση των περιοχών προβολής δεν είναι τόσο απλή. Από το σχήμα φαίνεται ότι υπάρχουν 3 περιοχές, ωστόσο τα πειράματα για $K = 3$ δεν επέστρεψαν σωστά αποτελέσματα. Η έξοδος που απεικονίζεται με το πράσινο χρώμα προκύπτει με $K = 20$ στο πρώτο πέρασμα και $K = 2$ στο δεύτερο. Παρόλο που οι συστάδες είναι πολλές στο πρώτο πέρασμα, η προκύπτουσα max-plus επιφάνεια δεν περιλαμβάνει ισάριθμα τμήματα. Παρατηρούμε, λοιπόν, ότι ο αλγόριθμος επιτυγχάνει μία πολύ καλή προσέγγιση της αρχικής συνάρτησης, καθώς η πράσινη γραμμή συμπίπτει σχεδόν σε όλο το πεδίο ορισμού με τη μαύρη διακεκομμένη.




(a) $f_1(x) = (-0.4x + 6 \wedge -1.6x) \vee (0.4x + 6 \wedge 1.6x)$ (b) $f_2(x) = f_1(x) \vee (5 - 0.2x \wedge 5 + 0.2x)$

Σχήμα 6.1.4: Παραδείγματα min-max Παλινδρόμησης

### 6.1.3 Επίλυση Monotonic Regression μέσω Monotone Neural Networks

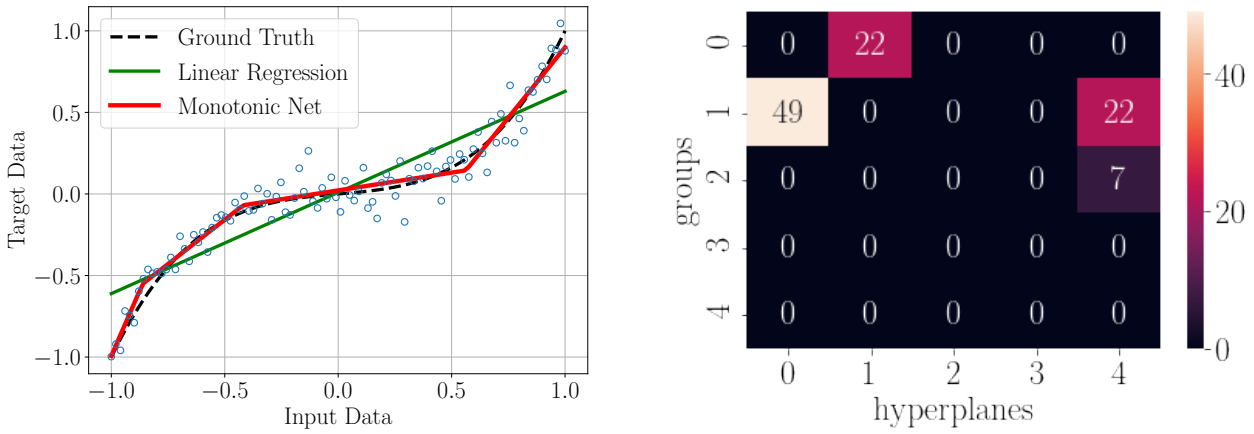
Επιλέγοντας κατάλληλη αρχιτεκτονική για ένα νευρωνικό δίκτυο, είναι δυνατή η παλινδρόμηση σε δεδομένα διατηρώντας την επιφάνεια εξόδου μονότονη. Η σχετική θεωρία βρίσκεται στην ενότητα 5.5.2 και στο σημείο αυτό εξετάζονται τα εν λόγω δίκτυα (βλ. σχήμα 5.5.1) υπό πειραματική σκοπιά. Χωρίς βλάβη της γενικότητας, επικεντρωνόμαστε σε αύξουσες συναρτήσεις.

Έστω η συνάρτηση $f(x) = x^3 + x + \sin x$, η οποία είναι γνησίως αύξουσα καθώς $f'(x) = 3x^2 + 1 + \cos x > 0, \forall x \in \mathbb{R}$. Χρησιμοποιώντας το συμβολισμό της εξίσωσης (5.5.2), επιλέγουμε $K = 5$ ομάδες καθεμία από τις οποίες απαρτίζεται από $J = 5$ υπερεπίπεδα. Κανονικοποιούμε τη συνάρτηση στο $\mathcal{X} \times \mathcal{Y} = [-1, 1]^2$, ώστε η επιλογή θορύβου να συνάδει με τη διαίσθηση. Επιλέγουμε 100 σημεία για το σύνολο δεδομένων, τα οποία διαφθείρουμε με θόρυβο $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Για τη διασφάλιση θετικών κλίσεων, άρα και μονότονης επιφάνειας εξόδου, επιλέγεται η απεικόνιση $z \mapsto z^2$ ώστε να είναι δυνατό να υπάρχει μηδενική κλίση. Με άλλα λόγια, η απεικόνιση αυτή συμπεριλαμβάνει την περίπτωση αύξουσας/φθίνουσας συνάρτησης αντί για αποκλειστικά *γνησίως* αύξουσες/φθίνουσες συναρτήσεις.

Στο σχήμα 6.1.5a παρουσιάζεται η έξοδος του νευρωνικού δικτύου και η ευθεία που προκύπτει από γραμμική παλινδρόμηση για σύγκριση. Η εκπαίδευση διήρκησε 1000 εποχές χρησιμοποιώντας τον αλγόριθμο

βελτιστοποίησης Adaptive Momentum Estimation (adam) με ρυθμό μάθησης $\eta = 0.01$. Εποπτικά, η έξοδος του νευρωνικού προσεγγίζει το σύνολο δεδομένων πολύ καλύτερα από τη γραμμική πανινδρόμηση.

Το σχήμα 6.1.5b δείχνει πόσα σημεία αντιστοιχούν σε κάθε υπερεπίπεδο. Από τα διαθέσιμα 25 μόνο 5 αξιοποιούνται, γεγονός αναμενόμενο αν σκεφτεί κανείς τη μορφή της συνάρτησης $f$. Η είσοδος είναι κοίλη για αρνητικές τιμές $f''(x) < 0$ και ιδιαίτερα για $x \in [-1, 0.5]$. Συνεπώς, σε αυτό το πεδίο τιμών, δύο ομάδες με ένα υπερεπίπεδο προσεγγίζουν αυτό το κοίλο τμήμα. Αντίθετα, για $x \in [0.5, 1]$ η συνάρτηση είναι "ιδιαίτερα" κυρτή και χρησιμοποιούνται δύο υπερεπίπεδα για την προσέγγιση. Για τις ενδιάμεσες τιμές, η είσοδος προσεγγίζεται ικανοποιητικά από ένα ευθύγραμμο τμήμα, το οποίο συμπεριλαμβάνεται στην ίδια ομάδα με τα προηγούμενα υπερεπίπεδα. Η προσέγγιση, λοιπόν, καθρεπτίζει τη συμμετρία της ειδόδου.




(a) Συνάρτηση εξόδου

(b) Ενεργοποιήσεις της συνάρτησης εξόδου

Σχήμα 6.1.5: Παράδειγμα Παλινδρόμησης με Sill Networks

Από τα πειράματα, προέκυψε μία πολύ σημαντική παρατήρηση σχετικά με την αρχικοποίηση των βαρών. Ο αλγόριθμος κατάβασης κλίσεων ενημερώνει τις παραμέτρους του μοντέλου σύμφωνα με τις παραγώγους που προκύπτουν από τον αλγόριθμο οπισθοδιάδοσης. Ωστόσο, οι μορφολογικοί τελεστές $\min, \max$ μονοπωλούν τη διαδικασία αυτή βαρών καθώς ενημερώνονται μόνο τα βάρη του υπερεπιπέδου που αφορά το σημείο εισόδου. Για περισσότερες λεπτομέρειες, βλ. ενότητα 5.5.2. Το γεγονός αυτό αποτελεί τροχοπέδη στην εκπαίδευση, καθώς μία λάθος αρχικοποίηση οδηγεί σε παθολογικές περιπτώσεις όπου ορισμένα υπερεπίπεδα ή ολόκληρες ομάδες δεν ενημερώνονται ποτέ. Μία απλή λύση είναι η χρήση Glorot αρχικοποίησης [GB10] αλλά με κλιμάκωση. Στο παράδειγμα, επιλέγεται παράμετρος κλιμάκωσης (gain) ίση με 25. Αρχικά, λοιπόν, τα βάρη έχουν υψηλές τιμές οδηγώντας σε υψηλό σφάλμα. Κατά την εκπαίδευση ενημερώνονται οι παράμετροι και οι τιμές διορθώνονται σε μικρότερες. Οι ομάδες, όμως, επεξεργάζονται τα υπερεπίπεδα με τελεστή max και, συνεπώς, "επιβιώνουν" εκείνα που οι τιμές τους είναι ακόμη υψηλές, δηλαδή αυτά που δεν έχουν ενημερωθεί ακόμα. Σιγά σιγά, όλα τα υπερεπίπεδα αντιστοιχίζονται σε σημεία ειδόδου και ενημερώνονται τα βάρη τους. Χωρίς την κλιμάκωση, η μέθοδος είναι ασταθής και πολλές φορές η έξοδός της ταυτίζεται με αυτή της γραμμικής παλινδρόμησης.

Ένας τρόπος αντιμετώπισης του προβλήματος αυτού είναι η χρήση ομαλών προσεγγίσεων των μορφολογικών τελεστών. Για τη σχετική θεωρία, βλ. υποενότητα 5.3.2. Επομένως, για ένα διάνυσμα εισόδου έχουμε τις εξής επιλογές:

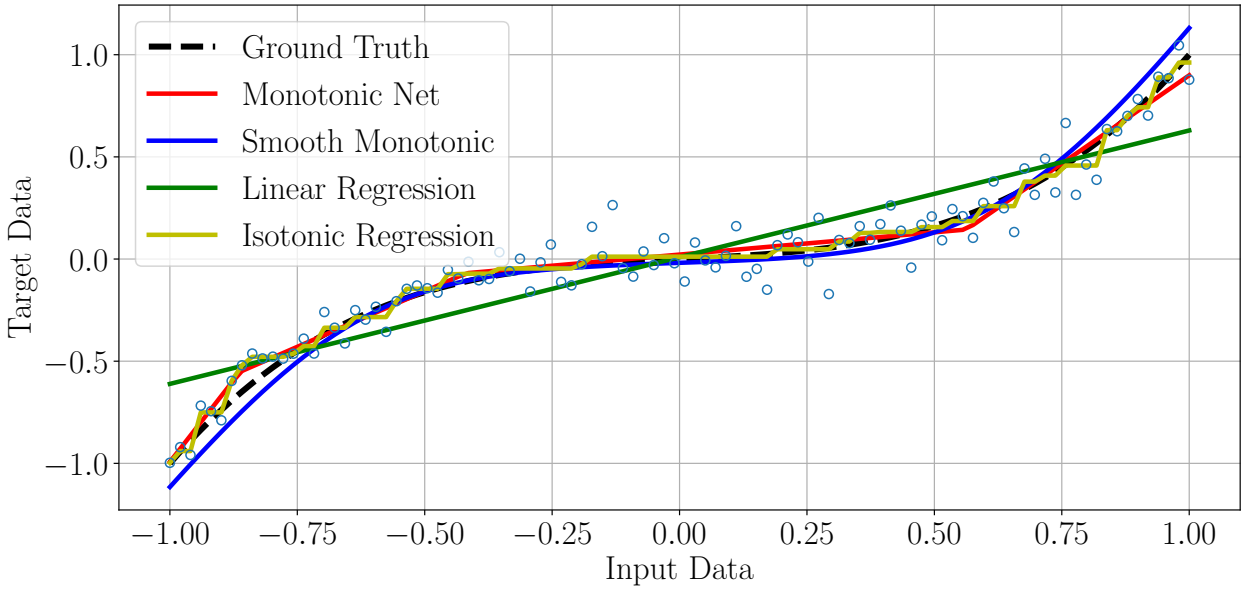
$$\mathbf{x} \in \mathbb{R}^n \mapsto \begin{cases} \max\limits_{i \in [n]} x_i & \text{μορφολογικός τελεστής} \\ \frac{1}{\beta} \log \left( \sum\limits_{i=1}^{n} \exp(\beta \cdot x_i) \right) & \text{ομαλή προσέγγισή του max (βλ. (5.3.7))}, \beta > 0 \end{cases}$$

Με την ομαλή προσέγγιση, δεν υπάρχουν τα παραπάνω ζητήματα στην εκπαίδευση, εφόσον πλέον κάθε ενημέρωση δεν επηρεάζει μόνο τις παραμέτρους του ενεργού υπερεπιπέδου. Αυτό επιβεβαιώνουν και τα πειράματά μας, όπου η παράμετρος κλιμάκωσης δεν είναι απαραίτητη, αν και κρίνεται βοηθητική σε μικρές τιμές. Στο σχήμα 6.1.6 βλέπουμε τις επιφάνειας που παράγουν οι διάφορες μέθοδοι. Όσον αφορά το σφάλμα, οι δύο μέθοδοι είναι συγκρίσιμες και τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα. Τέλος, συγκριτικά με

την ισοτονική παλινδρόμηση, το μονότονο νευρωνικό δίκτυο επιτυγχάνει καλύτερη προσέγγιση σύμφωνα με το μεσοτετραγωνικό λάθος.

| Μέθοδος | $\sigma = 0.05$ | $\sigma = 0.1$ | $\sigma = 0.15$ | $\sigma = 0.2$ |
|---|---|---|---|---|
| Linear Regression | 0.0236 | 0.03077 | 0.04827 | 0.0505 |
| Isotonic Regression | 0.0042 | 0.01112 | 0.02557 | 0.0417 |
| Monotonic Net [Sil98] | 0.00305 | 0.01107 | 0.02401 | 0.0395 |
| Smooth Monotonic Net [ours] | **0.00294** | **0.00938** | **0.02302** | **0.0386** |

Πίνακας 6.1: Σύγκριση RMS error των μονοτονικών μεθόδων για θόρυβο $\mathcal{N}(0, \sigma^2)$

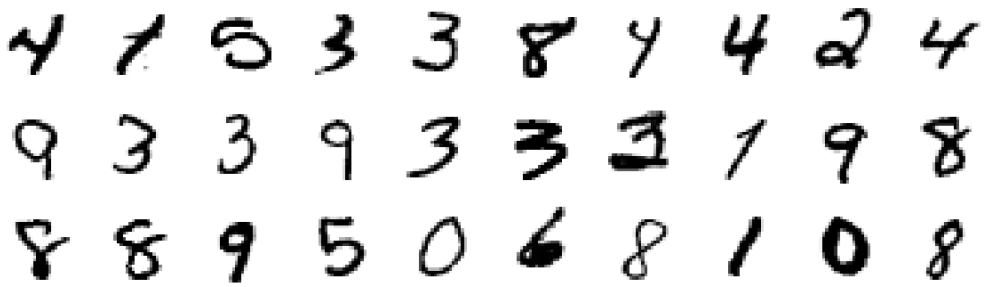



Σχήμα 6.1.6: Απεικόνιση των μεθόδων Παλινδρόμησης για τις διάφορες μεθόδους. Με μπλε απεικονίζεται το προτεινόμενο μοντέλο με χρήση ομαλοποιημένων μορφολογικών τελεστών.

## 6.2 Πειραματισμός με Μορφολογικά Δίκτυα σε σύνολα δεδομένων Όρασης Υπολογιστών

### 6.2.1 Σύνολο Δεδομένων MNIST

Το σύνολο δεδομένων MNIST (Modified National Institute of Standards and Technology)[LeC98] αποτελείται από εικόνες μεγέθους $28 \times 28$ pixels με χειρόγραφα ψηφία στο δεκαδικό σύστημα (0-9). Πρόκειται για ένα από τα πιο διαδεδομένα datasets και χρησιμοποιείται κατά κόρον ως benchmark στο χώρο της Τεχνητής Νοημοσύνης και ιδιαίτερα στο υποπεδίο της Μηχανικής Μάθησης. Ιστορικά, η δημοσίευση του MNIST dataset συμπίπτει με την επινόηση των συνελικτικών δικτύων [LeC+98]. Στην εικόνα 6.2.1 παρατίθενται ορισμένα παραδείγματα των χειρόγραφων ψηφίων.



Σχήμα 6.2.1: Παραδείγματα από το `MNIST` dataset

Το σύνολο δεδομένων αποτελείται από $70,000$ πρότυπα: $60,000$ πρότυπα εκπαίδευσης (training) και $10,000$ πρότυπα αξιολόγησης (test). Για την πειραματική αξιολόγηση διαχωρίζουμε περαιτέρω το σύνολο δεδομένων εκπαίδευσης σε training $50,000$ και validation $10,000$. Χρησιμοποιούμε διάφορους αλγορίθμους βελτιστοποίησης, όπως Stochastic Gradient Descent (SGD) (βλ. § 4.5.2), Adaptive Moment Estimation (Adam) (βλ. § 4.5.4). Εξετάζουμε τη συμπεριφορά των αλγορίθμων για διάφορους ρυθμούς μάθησης $\eta \in \{0.001, 0.003, \cdots, 0.009, 0.01, 0.03, \cdots, 0.09\}$. Η ποικιλία της υπερπαραμέτρου πηγάζει από το γεγονός ότι οι χαμηλές τιμές προτείνονται για Adaptive Moment Estimation, ενώ οι υψηλότερες για Στοχαστική Κατάβαση Κλίσεων. Για λόγους πληρότητας, συνεπώς, καλύπτουμε το παραπάνω φάσμα τιμών.

Πειραματιζόμαστε με μεγάλη οικογένεια αρχιτεκτονικών, μεταβλητή τόσο σε βάθος όσο σε πλάτος επιπέδων. Ξεκινούμε την ανάλυση από μικρές και συνεχίζουμε σε μεγάλες (ως προς το πλήθος των νευρώνων) αρχιτεκτονικές. Καθώς επιλύουμε πρόβλημα ταξινόμησης, το τελευταία επίπεδο παραμένει αμετάβλητο: Softmax με 10 κλάσεις (μία για κάθε ψηφίο) και ως συνάρτηση κόστους επιλέγεται η διασταυρωμένη εντροπία (cross-entropy). Για την αρχικοποίηση των βαρών του νευρωνικού δικτύου, χρησιμοποείται Glorot ομοιόμορφη αρχικοποίηση [GB10]. Επιλέγουμε 50 εποχές εκπαίδευσης στον παρακάτω πίνακα

Εξετάζουμε τη συμπεριφορά των μορφολογικών δικτύων μόνο με όρους dilations, μόνο με όρους erosion καθώς και υβριδικά δίκτυα για διάφορες τιμές νευρώνων στο (μοναδικό) κρυφό επίπεδο. Επιπρόσθετη παραμετροποίηση εντοπίζεται στην επιλογή μεθόδου εκπαίδευσης[1] και στο ρυθμό μάθησης (learning rate). Παρουσιάζουμε τα αποτελέσματα για το σύνολο δεδομένων `MNIST` για το μορφολογικό δίκτυο μόνο με όρους dilation.

| | Adaptive Momentum Estimation | | | | Stochastic Gradient Descent | | | |
|---|---|---|---|---|---|---|---|---|
| $\eta$ | 24 | 32 | 64 | 128 | 24 | 32 | 64 | 128 |
| 0.001 | 92.77 | 94.36 | 95.57 | 96.90 | 55.24 | 53.65 | 56.37 | 63.68 |
| 0.003 | 92.42 | 94.08 | 96.12 | 96.71 | 70.90 | 78.23 | 80.92 | 83.54 |
| 0.005 | 91.83 | 93.92 | 94.99 | 96.63 | 81.52 | 83.96 | 85.93 | 87.48 |
| 0.007 | 91.79 | 93.49 | 94.66 | 96.56 | 83.31 | 85.55 | 88.78 | 89.07 |
| 0.009 | 91.50 | 93.46 | 94.91 | 95.38 | 85.59 | 87.33 | 89.33 | 90.26 |
| 0.01 | 92.39 | 92.93 | 95.25 | 95.85 | 85.82 | 87.60 | 89.26 | 91.03 |
| 0.03 | 91.37 | 92.38 | 93.40 | 92.29 | 89.95 | 90.96 | 93.36 | 94.12 |
| 0.05 | 88.87 | 92.01 | 92.55 | 92.76 | 91.74 | 92.07 | 93.94 | 95.01 |
| 0.07 | 87.15 | 89.51 | 90.78 | 90.93 | 91.32 | 92.78 | 94.65 | 93.73 |
| 0.09 | 84.49 | 86.63 | 86.15 | 88.42 | 92.23 | 92.83 | 93.72 | 95.29 |

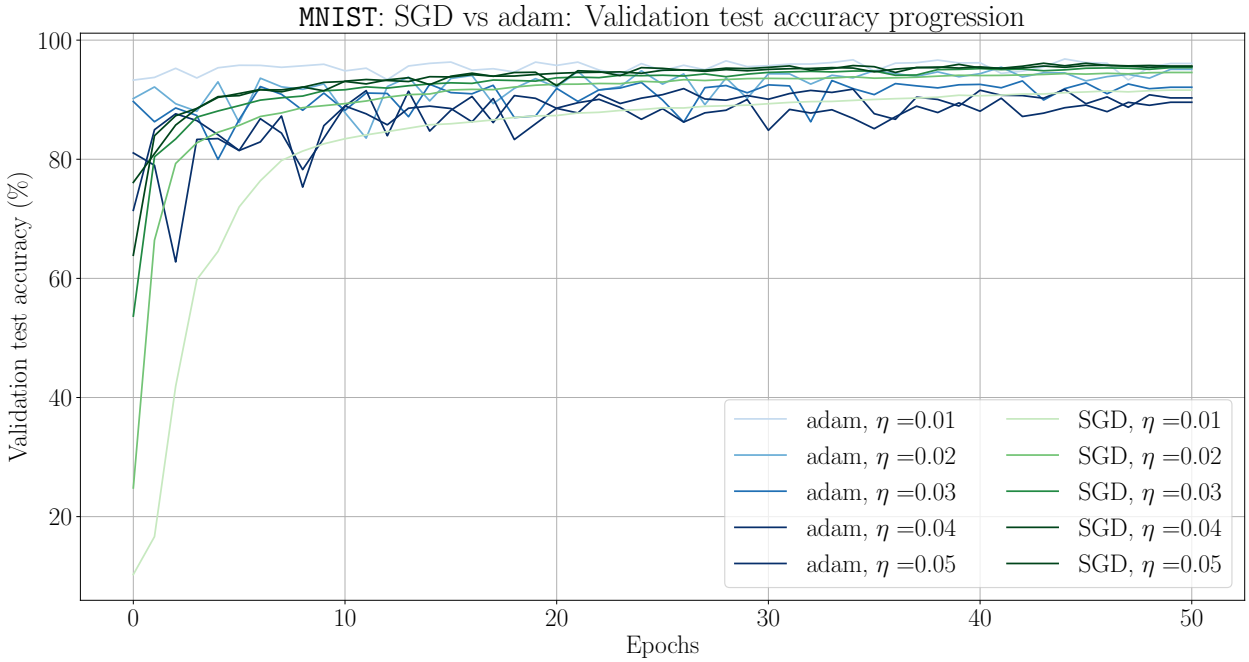
Πίνακας 6.2: `MNIST`: Ακρίβεια Μορφολογικού δικτύου μόνο με όρους dilation στο κρυφό επίπεδο για διάφορους ρυθμούς μάθησης $\eta$. Εξετάζονται διάφορα πλήθη νευρώνων στο κρυφό επίπεδο.

Παρατηρούμε ότι η αύξηση των νευρώνων στο κρυφό επίπεδο οδηγεί σε βελτίωση της ακρίβειας. Αυτό είναι αναμενόμενο, καθώς η αύξηση των νευρώνων συνεπάγεται μεγαλύτερη εκφραστικότητα του μοντέλου. Επιπλέον, παρατηρούμε ότι ο αλγόριθμος βελτιστοποίησης adam παράγει καλύτερα αποτελέσματα. Η συμπεριφορά των

[1] υπερπαράμετροι για τον αλγόριθμο Adam επιλέγονται οι προκαθορισμένες/προτεινόμενες: $\beta_1 = 0.99, \beta_2 = 0.9$.

αλγορίθμων βελτιστοποίησης είναι αξιοσημείωτη. Η Στοχαστική Κατάβαση Κλίσεων παράγει πολύ χαμηλά ποσοστά επιτυχίας για χαμηλούς ρυθμούς μάθησης. Αυτό σημαίνει ότι απαιτούνται πολύ περισσότερες εποχές από τις 50 που χαρακτηρίζουν το παραπάνω πειραματικό πλαίσιο. Μία άλλη εξήγηση είναι ότι ο αλγόριθμος έχει φτάσει σε κάποιο τοπικό ελάχιστο και η χαμηλή τιμή του ρυθμού μάθησης, σε συνδυασμό με την έλλειψη momentum, αποτελεί τροχοπέδη στην απόδραση από αυτό. Αντίθετα, ο αλγόριθμος adam συμπεριφέρεται βέλτιστα στο άλλο άκρο του φάσματος τιμών ρυθμού μάθησης. Παράλληλα, όμως, επιδεικνύει πολύ χαμηλότερη διακύμανση, γεγονός που επισημαίνει την ευρωστία του (robustness).

Εκτός από τις απόλυτες τιμές των μέτρων απόδοσης στο σύνολο των test δεδομένων, αξίζει να εντρυφήσουμε στην πρόοδο της ακρίβειας στο validation set κατά τις εποχές. Στο σχήμα 6.2.2 φαίνεται η εν λόγω πρόοδος, όπου με μπλε και πράσινες αποχρώσεις απεικονίζονται τα δεδομένα για τους αλγορίθμους Adaptive Moment Estimation (adam) και Stochastic Gradient Descent (SGD) αντίστοιχα. Είναι φανερό ότι ο adam επιτυγχάνει υψηλά ποσοστά επιτυχίας από την πρώτη κιόλας εποχή, ενώ ο απλούστερος αλγόριθμος της στοχαστικής κατάβασης κλίσεων (SGD) απαιτεί αρκετές εποχές για να συγκλίνει. Ωστόσο, οι υψηλές τιμές ρυθμού μάθησης αποδεικνύονται ζημιογόρες για τον αλγόριθμο adam και, για το λόγο αυτό, συμπεραίνουμε ότι είναι βέλτιστη η χρήση χαμηλών ρυθμών. Τα πειράματά μας επιβεβαιώνουν ότι μία καλή επιλογή είναι η $\eta = 0.001$.




Σχήμα 6.2.2: Σύγκριση ακρίβειας στο validation set του `MNIST` ανά τις εποχές εκπαίδευσης για τους optimizers Stochastic Gradient Descent & Adaptive Momentum Estimation.

Στη συνέχεια, παρουσιάζουμε τα αποτελέσματα για μορφολογικά δίκτυα ενός κρυφού επιπέδου με αποκλειστικά όρους erosion (βλ. πίνακα 6.3). Κατά τα άλλα, το πειραματικό πλαίσιο παραμένει αναλλοίωτο. Η απόδοση των μοντέλων εμφανίζει σαφή πτώση από τα μοντέλα διαστολής. Ωστόσο, η συμπεριφορά του ταξινομητή είναι ποιοτικά εφάμιλλη με το προηγούμενο δίκτυο, καθώς ο adam εξακολουθεί να είναι ο πιο εύρωστος αλγόριθμος. Αντίθετα με το δίκτυο διαστολής, οι βέλτιστες τιμές ρυθμών μάθησης εντοπίζονται στο κέντρο του φάσματος. Οι χαμηλοι ρυθμοί μάθησης αποτελούν απαγορετική επιλογή για τη Στοχαστική Κατάβαση Κλίσεων, σύμφωνα και με τα πειράματα στα μοντέλα συστολής.

| | Adaptive Momentum Estimation | | | | Stochastic Gradient Descent | | | |
|---|---|---|---|---|---|---|---|---|
| $\eta$ | 24 | 32 | 64 | 128 | 24 | 32 | 64 | 128 |
| 0.001 | 79.19 | 80.39 | 89.27 | 92.62 | 40.13 | 49.65 | 50.70 | 54.44 |
| 0.003 | 72.28 | 82.94 | 89.97 | 93.57 | 58.87 | 59.23 | 77.01 | 81.62 |
| 0.005 | 70.06 | 82.36 | 91.16 | 94.09 | 68.81 | 72.62 | 78.67 | 85.70 |
| 0.007 | 82.12 | 85.19 | 91.45 | 94.91 | 68.25 | 72.10 | 80.04 | 85.65 |
| 0.009 | 80.47 | 86.24 | 91.14 | 94.13 | 71.76 | 73.70 | 81.27 | 85.95 |
| 0.01 | 78.40 | 84.89 | 91.74 | 94.46 | 67.76 | 69.72 | 83.08 | 88.12 |
| 0.03 | 84.30 | 87.12 | 89.77 | 93.07 | 75.01 | 77.57 | 87.98 | 89.87 |
| 0.05 | 82.33 | 82.61 | 86.85 | 89.92 | 74.08 | 82.40 | 85.08 | 90.67 |
| 0.07 | 81.31 | 84.37 | 86.48 | 87.63 | 75.75 | 78.66 | 85.25 | 91.68 |
| 0.09 | 78.08 | 82.19 | 86.76 | 83.70 | 74.01 | 75.39 | 86.21 | 91.63 |

Πίνακας 6.3: `MNIST`: Ακρίβεια Μορφολογικού δικτύου μόνο με όρους erosion στο κρυφό επίπεδο για διάφορους ρυθμούς μάθησης $\eta$. Εξετάζονται διάφορα πλήθη νευρώνων στο κρυφό επίπεδο.

Στο σημείο αυτό, συνδυάζουμε τις παραπάνω αρχιτεκτονικές χρησιμοποιώντας και από τις δύο κατηγορίες όρων (dilation, erosion). Διατηρούμε το συνολικό πλήθος των νευρώνων του κρυφού επιπέδου ίδιο, διασπώντας το σε μισούς όρους dilation και μισούς erosion. Επιπλέον, επεκτείνουμε το πειραματικό περιβάλλον και σε μοντέλα με περισσότερους νευρώνες. Τα αποτελέσματα παρουσιάζονται στον πίνακα 6.4, όπου παρατηρούμε σαφή βελτίωση στους ταξινομητές.

Οι νευρώνες dilation και erosion δημιουργούν διαφορετικά features ως εισόδους για τα επόμενα επίπεδα. Καθώς τα πειράματα αφορούν Όραση Τπολογιστών, αξίζει να σημειωθεί ότι για τις μονοκάναλες (γκρίζες) εικόνες, υψηλές τιμές σε pixels αντιστοιχούν σε φωτεινές περιοχές, ενώ χαμηλές τιμές αντιστοιχούν σε σκοτεινές. Συνεπώς, οι νευρώνες dilation επικεντρώνονται στην πληροφορία που κωδικοποιούν τα φωτεινά σημεία και οι όροι erosion στις σκοτεινές περιοχές. Οι μεικτές αρχιτεκτονικές περιλαμβάνουν και τους δύο τύπους νευρώνουν. Άρα, ο εν λόγω ταξινομητής αξιοποιεί και τις δύο πηγές πληροφοριών. Το αποτέλεσμα είναι η βελτίωση της απόδοσης, γεγονός που είναι εμφανές στον πίνακα 6.4.

| | Adaptive Momentum Estimation | | | | | | Stochastic Gradient Descent | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\eta$ | 24 | 32 | 64 | 128 | 256 | 400 | 24 | 32 | 64 | 128 | 256 | 400 |
| 0.001 | 89.75 | 92.19 | 94.92 | 96.47 | 97.42 | 97.63 | 51.17 | 49.12 | 56.19 | 55.22 | 61.15 | 61.84 |
| 0.003 | 90.08 | 91.57 | 94.76 | 96.59 | 96.83 | 97.26 | 63.58 | 73.65 | 79.86 | 81.9 | 83.8 | 84.38 |
| 0.005 | 90.16 | 90.55 | 94.67 | 96.44 | 97.17 | 97.09 | 76.52 | 81.65 | 84.37 | 86.63 | 88.07 | 87.88 |
| 0.007 | 89.51 | 91.33 | 94.15 | 95.52 | 95.81 | 95.55 | 77.42 | 81.66 | 85.76 | 87.61 | 89.19 | 89.24 |
| 0.009 | 89.79 | 92.16 | 94.67 | 95.44 | 96.13 | 96.49 | 78.16 | 84.49 | 88.38 | 88.69 | 90.01 | 90.28 |
| 0.01 | 90.22 | 91.62 | 93.99 | 95.80 | 95.47 | 95.96 | 79.57 | 84.43 | 88.60 | 89.88 | 90.39 | 90.81 |
| 0.03 | 89.80 | 90.53 | 94.06 | 92.79 | 93.85 | 94.09 | 86.42 | 88.72 | 91.67 | 93.21 | 93.68 | 94.06 |
| 0.05 | 87.44 | 89.73 | 91.74 | 91.91 | 90.52 | 87.68 | 87.08 | 89.01 | 93.00 | 94.14 | 94.64 | 94.32 |
| 0.07 | 85.74 | 86.83 | 90.35 | 88.15 | 91.33 | 91.78 | 88.32 | 89.80 | 92.79 | 94.36 | 95.05 | 95.68 |
| 0.09 | 82.97 | 85.37 | 87.56 | 87.23 | 89.01 | 87.14 | 89.10 | 90.00 | 92.73 | 94.73 | 95.74 | 96.07 |

Πίνακας 6.4: `MNIST`: Ακρίβεια Μορφολογικού δικτύου με ισάριθμους όρους dilation και erosion στο κρυφό επίπεδο για διάφορους ρυθμούς μάθησης $\eta$. Εξετάζονται διάφορα πλήθη νευρώνων στο κρυφό επίπεδο.

Επεκτείνουμε το παραπάνω υβριδικό δίκτυο προσθέτοντας ένα δεύτερο κρυφό επίπεδο. Από τα προηγούμενα αποτελέσματα παρατηρούμε ότι οι βέλτιστες αποδόσεις επιτυγχάνονται με τον αλγόριθμο adam για χαμηλούς ρυθμούς μάθησης. Για το λόγο αυτό, μεταβάλλουμε το πειραματικό πλαίσιο ώστε να περιλαμβάνει αποκλειστικά μοντέλα εκπαιδευμένα με Adaptive Moment Estimation και ρυθμούς μάθησης $\eta \in \{0.001, 0.002, \ldots, 0.005\}$. Όσον αφορά το πλήθος και το είδος των νευρώνων των κρυφών επιπέδων, επιλέγουμε $n$ όρους dilation και $n$ όρους erosion στο πρώτο κρυφό επίπεδο, ενώ το δεύτερο επίπεδο αποτελείται από τους μισούς νευρώνες

διατηρώντας την $50 - 50$ αναλογία.

Μία ιδιαίτερα σημαντική αλλαγή εντοπίζεται στην αύξηση των εποχών μάθησης από 50 σε 200. Η χρήση μεικτών επιπέδων διαστολής και συστολής *σε σειρά* παράγει ενεργοποιήσεις εφάμιλλες με μορφολογικά openings και closings (βλ. § 5.4). Αυτό, ωστόσο, επιδεινώνει το πρόβλημα διάδοσης κλίσης (gradient propagation) και, ως αποτέλεσμα, απαιτείται περισσότερος χρόνος για την εκπαίδευση του δικτύου συγκριτικά με τις προηγούμενες αρχιτεκτονικές ενός κρυφού επιπέδου. Τα αποτελέσματα παρουσιάζονται στον πίνακα 6.5. Παρατηρούμε ότι η αύξηση της πολυπλοκότητας του μοντέλου δεν επιφέρει βελτίωση στην απόδοση αφού τα αποτελέσματα είναι εφάμιλλα του δικτύου με ένα μεικτό κρυφό επίπεδο (βλ. πίνακα 6.4). Ενδέχεται ο πιο ενδελεχής πειραματισμός για τις αρχιτεκτονικές, ως προς το πλήθος κόμβων σε κάθε επίπεδο, και τις υπερπαραμέτρους, όπως πλήθος εποχών μάθησης καθώς και ρυθμός μάθησης, να οδηγήσουν σε καλύτερα αποτελέσματα για το δίκτυο 2 κρυφών επιπέδων.

| $\eta$ \ $n$ | 24 | 32 | 64 | 128 | 256 | 400 |
|---|---|---|---|---|---|---|
| 0.001 | 86.57 | 90.76 | 94.21 | 96.22 | 96.6 | 96.91 |
| 0.002 | 91.20 | 91.66 | 94.76 | 95.90 | 96.61 | 96.85 |
| 0.003 | 90.76 | 90.94 | 95.15 | 95.94 | 96.67 | 96.82 |
| 0.004 | 90.19 | 91.06 | 94.49 | 95.76 | 97.08 | 96.74 |
| 0.005 | 91.58 | 91.52 | 94.55 | 96.33 | 95.92 | 97.09 |
| 0.006 | 90.75 | 92.51 | 94.99 | 95.99 | 95.52 | 95.93 |
| 0.007 | 90.39 | 92.57 | 94.94 | 96.20 | 95.93 | 95.84 |
| 0.008 | 90.29 | 91.81 | 94.76 | 95.79 | 96.44 | 95.49 |
| 0.009 | 91.07 | 92.73 | 94.00 | 96.00 | 95.94 | 96.47 |
| 0.01 | 90.01 | 92.43 | 94.67 | 94.37 | 96.45 | 96.68 |

Πίνακας 6.5: `MNIST`: Ακρίβεια Μορφολογικού δικτύου με δύο κρυφά επίπεδα για διάφορους ρυθμούς μάθησης $\eta$. Εξετάζονται διάφορα πλήθη νευρώνων στο κρυφό επίπεδο. Κάθε κρυφό επίπεδο αποτελείται από ισάριθμους όρους dilation και erosion.

### 6.2.2 Σύνολο Δεδομένων FashionMNIST

Στην εποχή του Deep Learning, το σύνολο δεδομένων MNIST έχει πάψει να αποτελεί πρόκληση για τους ερευνητές, καθώς αρχιτεκτονικές έχουν καταφέρει ποσοστά επιτυχίας άνω του 99.7%. Επιπλέον, σε πολλές περιπτώσεις μπορούμε να αποφανθούμε για ένα πρότυπο εστιάζοντας σε λίγα pixels ή ακόμα και σε ένα. Για το λόγο αυτό, οι Xiao, Rasul, and Vollgraf προτείνουν το σύνολο δεδομένων FashionMNIST [XRV17] ως άμεσο αντικαταστάτη του MNIST, αφού αποτελείται από τον ίδιο αριθμό προτύπων (10) και οι εικόνες έχουν την ίδια διάσταση ($28 \times 28$ pixels). Το σύνολο δεδομένων αποτελείται από $70,000$ εικόνες αντικειμένων ρουχισμού. Στην εικόνα 6.2.3 παρατίθενται ορισμένα παραδείγματα των εικόνων.

Σχήμα 6.2.3: Παραδείγματα από το `FashionMNIST` dataset

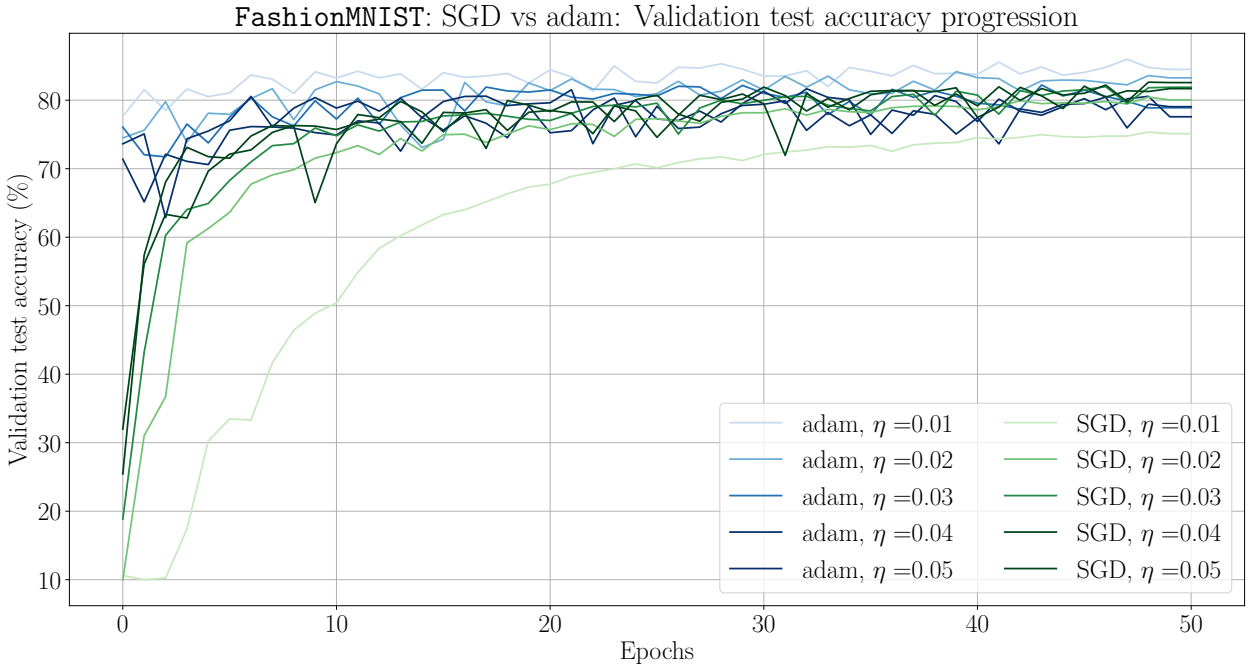
Σε συνέχεια των πειραμάτων στο σύνολο δεδομένων `MNIST`, χρησιμοποιούμε το ίδιο πειραματικό πλαίσιο για το σύνολο δεδομένων `FashionMNIST`, διασπώντας το σύνολο δεδομένων σε $50,000$ πρότυπα training, $10,000$

validation και 10,000 testing. Ξεκινούμε, λοιπόν, με μορφολογικό δίκτυο ενός κρυφού επιπέδου με αποκλειστικά νευρώνες διαστολής. Τα αποτελέσματα παρουσιάζονται στον πίνακα 6.6. Η επιλογή βήματος μάθησης $\eta = 0.007$ για τον αλγόριθμο adam αποφέρει τα βέλτιστα αποτελέσματα. Τα αποτελέσματα επιβεβαιώνουν ότι η αύξηση των νευρώνων αποτελεί ευεργετικό παράγοντα γαι την απόδοση του μοντέλου.

| | Adaptive Momentum Estimation | | | | Stochastic Gradient Descent | | | |
|---|---|---|---|---|---|---|---|---|
| $\eta$ | 24 | 32 | 64 | 128 | 24 | 32 | 64 | 128 |
| 0.001 | 78.89 | 80.98 | 83.22 | 84.70 | 29.67 | 30.21 | 30.74 | 29.84 |
| 0.003 | 79.63 | 80.92 | 84.09 | 84.79 | 54.54 | 57.42 | 55.19 | 58.35 |
| 0.005 | 77.84 | 80.72 | 82.81 | 85.24 | 59.45 | 66.08 | 68.04 | 66.60 |
| 0.007 | 79.20 | 80.74 | 83.01 | 85.92 | 67.89 | 67.73 | 71.84 | 71.77 |
| 0.009 | 79.03 | 80.19 | 83.63 | 85.23 | 69.40 | 69.07 | 73.24 | 73.07 |
| 0.01 | 79.49 | 80.66 | 83.88 | 79.28 | 71.06 | 71.63 | 73.10 | 74.45 |
| 0.03 | 79.58 | 80.30 | 82.31 | 82.33 | 75.48 | 73.33 | 77.75 | 78.88 |
| 0.05 | 78.36 | 78.61 | 80.26 | 82.22 | 76.46 | 77.51 | 79.14 | 80.47 |
| 0.07 | 76.46 | 77.08 | 79.82 | 77.36 | 74.98 | 77.75 | 80.41 | 81.21 |
| 0.09 | 74.11 | 75.64 | 76.90 | 77.75 | 76.46 | 77.56 | 80.12 | 78.72 |

Πίνακας 6.6: `FashionMNIST`: Ακρίβεια Μορφολογικού δικτύου μόνο με όρους dilation στο κρυφό επίπεδο για διάφορους ρυθμούς μάθησης $\eta$. Εξετάζονται διάφορα πλήθη νευρώνων στο κρυφό επίπεδο.

Επιπλέον, παρατηρούμε ότι η πρόδοος στην απόδοση του νευρωνικού δικτύου στο validation set είναι σαφώς ανώτερη με την επιλογή του αλγορίθμου adam για εκπαίδευση. Σύμφωνα με το σχήμα 6.2.4 ο αλγόριθμος Stochastic Gradient Descent απαιτεί πολλές εποχές ώστε να επιτύχει απόδοση συγκρίσιμη με αυτή του adam. Μάλιστα, για τη μικρότερη τιμή του ρυθμού μάθησης, $\eta = 0.001$, η απόδοση του μοντέλου υστερεί για πολλές εποχές σε σχέση με τις άλλες παραμετροποιήσεις. Η εκπαίδευση εγκλωβίζεται σε τοπικό ελάχιστο. Αυτό είναι εμφανές από το γεγονός ότι η ανοιχτή πράσινη γραμμή αντιστοιχεί σε ιδιαίτερα χαμηλότερα αποτελέσματα, τα οποία σταθεροποιούνται στο ήμισυ της εκπαίδευσης.




Σχήμα 6.2.4: Σύγκριση ακρίβειας στο validation set του `FashionMNIST` ανά τις εποχές εκπαίδευσης για τους optimizers Stochastic Gradient Descent & Adaptive Momentum Estimation.

Τα πειράματα συνεχίζουν με μορφολογικά νευρωνικά δίκτυα ενός κρυφού επιπέδου με αποκλειστικά κόμβους συστολής. Τα αποτελέσματα παρουσιάζονται στον πίνακα 6.7 και εμφανίζουν τόσο ομοιότητες όσο και διαφορές με την προηγούμενη αρχιτεκτονική. Πιο συγκεκριμένα, η αύξηση των νευρώνων εξακολουθεί να βελτιώνει την απόδοση του μοντέλου. Αντιθέτως, ο αλγόριθμος Στοχαστικής Κατάβασης Κλίσεων παρουσιάζει καλύτερα αποτελέσματα για μεγαλύτερους ρυθμούς μάθησης.

| | Adaptive Momentum Estimation | | | | Stochastic Gradient Descent | | | |
|---|---|---|---|---|---|---|---|---|
| $\eta$ | 24 | 32 | 64 | 128 | 24 | 32 | 64 | 128 |
| 0.001 | 79.17 | 81.48 | 83.79 | 86.13 | 48.05 | 61.74 | 62.82 | 69.08 |
| 0.003 | 80.40 | 81.10 | 83.48 | 85.47 | 63.38 | 65.09 | 71.47 | 76.54 |
| 0.005 | 80.53 | 80.78 | 84.38 | 85.46 | 67.89 | 74.42 | 75.75 | 78.40 |
| 0.007 | 77.78 | 80.76 | 84.00 | 85.41 | 70.15 | 73.96 | 76.84 | 79.60 |
| 0.009 | 79.20 | 81.87 | 84.04 | 85.63 | 72.94 | 70.91 | 77.81 | 80.35 |
| 0.01 | 79.91 | 80.94 | 83.83 | 85.16 | 73.29 | 75.23 | 78.89 | 80.19 |
| 0.03 | 79.85 | 80.90 | 83.26 | 81.13 | 77.22 | 77.79 | 81.11 | 82.75 |
| 0.05 | 77.18 | 79.24 | 79.55 | 75.69 | 77.11 | 79.43 | 81.74 | 83.95 |
| 0.07 | 76.12 | 75.30 | 78.88 | 79.28 | 77.38 | 79.04 | 81.90 | 84.18 |
| 0.09 | 75.49 | 75.73 | 74.07 | 76.11 | 77.55 | 78.95 | 82.44 | 84.17 |

Πίνακας 6.7: `FashionMNIST`: Ακρίβεια Μορφολογικού δικτύου μόνο με όρους erosion στο κρυφό επίπεδο για διάφορους ρυθμούς μάθησης $\eta$. Εξετάζονται διάφορα πλήθη νευρώνων στο κρυφό επίπεδο.

Προτού εξετάσουμε αρχιτεκτονικές πολλών κρυφών επιπέδεων, στρέφουμε την προσοχή μας σε αρχιτεκτονικές μεικτής φύσεως που απαρτίζονται τόσο από νευρώνες erosion όσο και dilation. Χάριν ευκολίας, επιλέγουμε ισάριθμους νευρώνες από κάθε κατηγορία. Τα αποτελέσματα φαίνονται στον πίνακα 6.8. Ο αλγόριθμος adam με βήμα μάθησης $\eta = 0.001$ επιτυγχάνει τη βέλτιστη απόδοση σε αυτή την παραμετροποίηση, αλλά και συγκριτικά με τις προηγούμενες αρχιτεκτονικές, βλ. πίνακες 6.6, 6.7.

| | Adaptive Momentum Estimation | | | | | | Stochastic Gradient Descent | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\eta$ | 24 | 32 | 64 | 128 | 256 | 400 | 24 | 32 | 64 | 128 | 256 | 400 |
| 0.001 | 82.05 | 82.05 | 85.11 | 86.76 | 87.11 | 88.34 | 41.54 | 50.99 | 60.39 | 65.79 | 68.94 | 66.9 |
| 0.003 | 81.21 | 83.00 | 84.74 | 86.53 | 86.89 | 87.37 | 62.02 | 64.10 | 69.33 | 73.84 | 74.34 | 76.42 |
| 0.005 | 81.76 | 83.16 | 85.20 | 86.74 | 85.21 | 86.93 | 62.65 | 69.25 | 72.82 | 76.03 | 78.09 | 79.46 |
| 0.007 | 80.41 | 82.66 | 83.22 | 84.12 | 83.38 | 86.34 | 71.36 | 72.80 | 75.35 | 77.61 | 79.32 | 80.55 |
| 0.009 | 81.13 | 82.34 | 83.99 | 85.78 | 85.22 | 85.46 | 71.80 | 71.54 | 76.26 | 79.11 | 80.86 | 81.72 |
| 0.01 | 81.51 | 82.40 | 84.48 | 86.10 | 84.34 | 86.79 | 72.58 | 75.08 | 76.02 | 78.44 | 80.66 | 81.74 |
| 0.03 | 79.46 | 80.92 | 82.85 | 84.78 | 83.73 | 83.13 | 76.08 | 78.37 | 80.28 | 82.32 | 83.15 | 84.40 |
| 0.05 | 78.28 | 79.24 | 82.08 | 82.82 | 82.56 | 82.13 | 76.00 | 79.82 | 82.42 | 82.91 | 83.54 | 83.98 |
| 0.07 | 76.85 | 78.34 | 78.24 | 80.85 | 77.49 | 78.04 | 76.96 | 79.70 | 83.10 | 83.98 | 85.25 | 85.47 |
| 0.09 | 74.62 | 74.68 | 75.68 | 77.83 | 75.20 | 75.58 | 80.08 | 78.86 | 83.03 | 82.74 | 84.89 | 86.21 |

Πίνακας 6.8: `FashionMNIST`: Ακρίβεια Μορφολογικού δικτύου με ισάριθμους όρους dilation και erosion στο κρυφό επίπεδο για διάφορους ρυθμούς μάθησης $\eta$. Εξετάζονται διάφορα πλήθη νευρώνων στο κρυφό επίπεδο.

Τέλος, επεκτείνουμε το μοντέλο σε δύο κρυφά επίπεδα με μεικτούς νευρώνες. Τα αποτελέσματα παρουσιάζονται στον πίνακα 6.9. Τα αποτελέσματα κινούνται στις ίδιες γραμμές με το σύνολο δεδομένων `MNIST`, καθώς η αύξηση της πολυπλοκότητας του δικτύου δεν επιφέρει βελτίωση στην απόδοση. Αντίθετα, παρατηρείται μικρή πτώση σχετικά με το μεικτό μορφολογικό επίπεδο (βλ. πίνακα 6.8). Επομένως, απαιτείται βαθύτερη αναζήτηση των αρχιτεκτονικών, των συνδυασμών που αφορούν τους αριθμούς κόμβων ανά επίπεδο και το πλήθος εποχών εκπαίδευσης.

| $\eta$ \ $n$ | 24 | 32 | 64 | 128 | 256 | 400 |
|---|---|---|---|---|---|---|
| 0.001 | 81.21 | 81.95 | 83.78 | 85.83 | 86.53 | 85.8 |
| 0.002 | 80.63 | 82.43 | 84.07 | 86.12 | 85.69 | 87.23 |
| 0.003 | 81.67 | 82.19 | 84.98 | 85.88 | 86.50 | 85.46 |
| 0.004 | 81.70 | 83.60 | 84.67 | 85.97 | 86.84 | 87.14 |
| 0.005 | 81.16 | 82.07 | 85.05 | 84.73 | 85.00 | 87.47 |
| 0.006 | 82.49 | 82.83 | 85.34 | 86.13 | 86.17 | 86.37 |
| 0.007 | 82.14 | 82.78 | 85.28 | 85.80 | 87.12 | 86.65 |
| 0.008 | 81.18 | 82.52 | 85.20 | 86.05 | 86.11 | 85.91 |
| 0.009 | 79.99 | 83.03 | 84.73 | 85.70 | 87.04 | 84.12 |
| 0.01 | 81.00 | 81.62 | 85.20 | 86.51 | 85.69 | 86.76 |

Πίνακας 6.9: `FashionMNIST`: Ακρίβεια Μορφολογικού δικτύου με δύο κρυφά επίπεδα για διάφορους ρυθμούς μάθησης $\eta$. Εξετάζονται διάφορα πλήθη νευρώνων στο κρυφό επίπεδο. Κάθε κρυφό επίπεδο αποτελείται από ισάριθμους όρους dilation και erosion.

### 6.2.3 Ομαλοποιημένα Μορφολογικά Δίκτυα

Στην παράγραφο 5.3.2, μελετήθηκε μία ομαλοποιημένη έκδοση των μορφολογικών δικτύων. Αξίζει, λοιπόν, η αξιολόγηση τέτοιων αρχιτεκτονικών. Τα πειράματα επικεντρώνονται στην επιρροή της υπερπαραμέτρου "σκληρότητας" $\beta$. Για το λόγο αυτό, επιλέγονται οι υπερπαράμετροι που οδηγούν στη βέλτιστη απόδοση: αλγόριθμος βελτιστοποίησης Adaptive Momentum Estimation με χαμηλούς ρυθμούς μάθησης $\eta \in \{0.001, 0.003, 0.005, 0.007, 0.009\}$. Η εκπαίδευση διαρκεί 100 εποχές. Η αρχιτεκτονική έγκειται σε μεικτό μορφολογικό επίπεδο με $n = 200$ ομαλοποιημένες διαστολές και ισάριθμες ομαλοποιημένες συστολές. Τα αποτελέσματα για τα σύνολα δεδομένων `MNIST` και `FashionMNIST` παρουσιάζονται στους πίνακες 6.10 και 6.11, αντίστοιχα.

| $\eta$ \ $\beta$ | β=1 | β=2 | β=4 | β=8 | β=20 |
|---|---|---|---|---|---|
| 0.001 | 88.10 | 93.85 | 96.22 | 97.78 | 97.30 |
| 0.003 | 89.03 | 94.69 | 96.54 | 97.24 | 97.08 |
| 0.005 | 89.11 | 96.67 | 97.84 | 96.50 | 96.55 |
| 0.007 | 94.79 | 96.90 | 97.48 | 97.14 | 96.63 |
| 0.009 | 95.64 | 96.59 | 97.25 | 97.03 | 96.28 |

Πίνακας 6.10: `MNIST`: Ακρίβεια Μορφολογικού δικτύου με 200 όρους ομαλοποιημένων dilation και 200 όρους ομαλοποιημένων erosion στο κρυφό επίπεδο για διάφορους ρυθμούς μάθησης $\eta$. Εξετάζεται η επίδραση της υπερπαραμέτρου σκληρότητας $\beta$.

| $\eta$ \ $\beta$ | β=1 | β=2 | β=4 | β=8 | β=20 |
|---|---|---|---|---|---|
| 0.001 | 76.95 | 82.28 | 85.11 | 86.67 | 87.08 |
| 0.003 | 72.41 | 82.51 | 86.17 | 84.95 | 86.44 |
| 0.005 | 80.80 | 85.40 | 87.13 | 86.89 | 86.34 |
| 0.007 | 83.76 | 85.29 | 84.19 | 85.04 | 85.02 |
| 0.009 | 82.24 | 85.04 | 86.29 | 85.48 | 85.57 |

Πίνακας 6.11: `FashionMNIST`: Ακρίβεια Μορφολογικού δικτύου με 200 όρους ομαλοποιημένων dilation και 200 όρους ομαλοποιημένων erosion στο κρυφό επίπεδο για διάφορους ρυθμούς μάθησης $\eta$. Εξετάζεται η επίδραση της υπερπαραμέτρου σκληρότητας $\beta$.

Μία παρατήρηση που χρήζει μνείας αφορά το χρόνο εκπαίδευσης. Ακόμα και για το ρηχό επιπεδο που εξετάζουμε (μόνο ένα κρυφό επίπεδο), η ομαλοποιημένη εκδοχή των μορφολογικών τελεστών οδηγεί σε σημαντική αύξηση του χρόνου εκπαίδευσης. Αυτό οφείλεται στην πολυπλοκότητα ενός Log-Sum-Exp όρου σε αντίθεση με τον τελεστή max.

### 6.2.4 Pruning Νευρωνικών Δικτύων

Τα state-of-the-art μοντέλα της Βαθιάς Μηχανικής Μάθησης (Deep Learning) αποτελούνται από εκατομμύρια παραμέτρους διασκορπισμένες σε δεκάδες επίπεδα. Στον πίνακα 6.12 παρέχονται ορισμένα παραδείγματα. Οι απαιτήσεις σε μνήμη είναι υψηλές και ο όγκος δεδομένων καθιστά αναποτελεσματική τη χρήση τους. Αντιθέτως, η εξέλιξη των βιολογικών συστημάτων έχει οδηγήσει σε αραιές αναπαραστάσεις. Συνεπώς, ερευνητική προσπάθεια καταβάλλεται για τη συμπίεση των σύνθετων μοντέλων μέσω της μείωσης των παραμέτρων δίχως να θυσιάζεται η απόδοση [HZS17]. Αυτές οι τεχνικές μειώνουν τη μνήμη, την μπαταρία και, γενικότερα, τις απαιτήσεις στο hardware. Επιπλέον, επιτρέπουν τη χρήση μοντέλων στο υπολογιστικό νέφος (cloud) για πολλούς χρήστες ή σε φορητές συσκευές, όπως smartphones, για προσωπική χρήση. Το τελευταίο πεδίο ενδιαφέροντος αποκτά ιδιαίτερη σημασία και για λόγους ιδιωτικότητας δεδομένων.

| Μοντέλο | Έτος Δημοσίευσης | # Επιπέδων | # Παραμέτρων |
|---|---|---|---|
| AlexNet | 2012 | 11 | 62,378,344 |
| Resnet | 2015 | - | 25,636,712 |
| VGG16 | 2015 | 23 | 138,357,544 |

Πίνακας 6.12: Πλήθος παραμέτρων σε γνωστές αρχιτεκτονικές. Πηγή

Στο πλαίσιο των τροπικών μαθηματικών έχει μελετηθεί το πρόβλημα εύρεσης αραιών λύσεων και το ζήτημα του pruning. Οι Smyrnis, Maragos, and Retsinas αναπτύσσουν μία γεωμετρική μέθοδο διαίρεσης πολυωνύμων στον τροπικό ημιδακτύλιο με στόχο την ελαχιστοποίηση νευρωνικών δικτύων [SMR20], ενώ στη συνέχεια οι Smyrnis and Maragos επεκτείνουν τον αλγόριθμο σε προβλήματα ταξινόμησης πολλών κλάσεων [SM20]. Παράλληλα, οι Tsiamis and Maragos μελετούν την αραιότητα max-plus συστημάτων, όπως (6.1.1), συνδέοντας τις λύσεις με το πρόβλημα του pruning.

Τα μορφολογικά νευρωνικά δίκτυα παράγουν εκ φύσεως αραιές ενεργοποιήσεις, καθώς οι τελεστές επιλέγουν μονοσήμαντα εισόδους. Συνεπώς, οι τεχνικές μείωσης παραμέτρων δύναται να εφαρμοστούν επιτυχώς για τις αρχιτεκτονικές που εξερευνήθηκαν στα προηγούμενα πειράματα. Εξετάζονται οι αρχιτεκτονικές με ένα κρυφό επίπεδο. Μπορεί να εφαρμοστεί κλάδεμα στις ακμές μεταξύ επιπέδων εισόδου και κρυφού είτε μεταξύ κρυφού επιπέδου και εξόδου. Απεικονίζονται κατ' αντιστοιχία οι ενεργοποιήσεις στις εικόνες 6.2.7 και 6.2.5.

Οι εικόνες είναι διαφωτιστικές. Επιλέγεται η απεικόνιση δικτύου με 64 νευρώνες διαστολής και 0 διαστολής (ώστε να προκύπτει πλέγμα $8 \times 8$). Κάθε εικόνα του πλέγματος αντιστοιχεί σε ένα νευρώνα του κρυφού επιπέδου και έχει μέγεθος $28 \times 28$ pixels όπως και τα διανύσματα εισόδου (εικόνες του συνόλου `MNIST`). Ο αλγόριθμος εκπαίδευσης είναι η Στοχαστική Κατάβαση Κλίσεων (SGD). Αναλυτικότερα, το επίπεδο γραμμικού συνδυασμού, το οποίο συμπίπτει με την έξοδο του δικτύου με 10 νευρώνες, δεν είναι καθόλου αραιό. Αυτό είναι αναμενόμενο, καθώς υπολογίζει ένα σταθμισμένο άθροισμα των ενεργοποιήσεων του προηγούμενου επιπέδου. Αντίθετα, το μορφολογικό επίπεδο είναι ιδιαίτερα αραιό. Αυτό είναι εμφανές από το σχήμα 6.2.7, όπου η ένταση του pixel συνεπάγεται την τιμή της ενεργοποίησης: τα πιο ανοιχτά pixels αφορούν υψηλές τιμές. Σε ορισμένους νευρώνες, αρκούν πολύ λίγα pixels από τα $28 \times 28 = 784$ διαθέσιμα, γεγονός που υποδηλώνει δυνατότητες pruning.

Εξετάζονται δύο μέθοδοι pruning:

- **Πρώτη Μέθοδος.** Αντιστοιχεί σε pruning σύμφωνα με τη νόρμα $\ell_1$. Αφαιρείται ένα συγκεκριμένο ποσοστό από τις μονάδες με τη χαμηλότερη $\ell_1$ νόρμα. Εξετάζεται η απόδοση του μοντέλου για διάφορα ποσοστά *διατήρησης* μονάδων $p \in \mathcal{P}$. Επιλέγεται έλεγχος ανά διαστήματα της τάξης του 10% έως ότου διατηρηθεί το 10% και στη συνέχεια μελετάται η απόδοση του μοντέλου για μονοψήφια ποσοστά διαχείρησης. Με άλλα λόγια, $\mathcal{P} = \{100, 90, 80, 70, 60, 50, 40, 30, 20, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1\}\%$. Αξίζει να σημειωθεί ότι οι ενεργοποιήσεις του κρυφού επιπέδου λαμβάνουν αρνητικές τιμές αλλά πολύ χαμηλές.

Ουσιαστικά, λοιπόν, το pruning με $\ell_1$ νόρμα αντιστοιχεί σε διατήρηση μόνο των υψηλών θετικών τιμών σε αυτή την περίπτωση. Τα αποτελέσματα αφορούν δίκτυα διαστολής και μεικτά. Για το σύνολο δεδομένων `MNIST`, τα αποτελέσματα για δίκτυο διαστολής και μεικτό παρουσιάζονται στους πίνακες 6.13 και 6.14, αντίστοιχα. Για το σύνολο δεδομένων `FashionMNIST`, τα αποτελέσματα βρίσκονται στους πίνακες 6.15 και 6.16, αντίστοιχα.

- **Δεύτερη Μέθοδος.** Η δεύτερη μέθοδος είναι ευριστική. Αναζητά τις μονάδες με μέγιστη ενεργοποίηση σε κάθε νευρώνα. Στη συνέχεια, διατηρεί αποκλειστικά τις μονάδες του εν λόγω νευρώνα που υπερβαίνουν ένα κατώφλι. Έστω, λοιπόν, $M$ η μέγιστη ενεργοποίηση του νευρώνα και $p \in (0, 1)$ η παράμετρος κατωφλίου. Τότε, διατηρούνται οι ενεργοποιήσεις $m$ που ικανοποιούν τη συνθήκη $m > p \cdot M$. Τα πειράματα περιορίζονται σε μεικτές αρχιτεκτονικές με υψηλό αριθμό νευρώνων. Παρουσιάζονται τα αποτελέσματα για τα σύνολα `MNIST` και `FashionMNIST` στους πίνακες 6.17 και 6.18. Σε παρενθέσεις, επισημειώνεται το ποσοστό (%) του αρχικού δικτύου που διατηρείται.

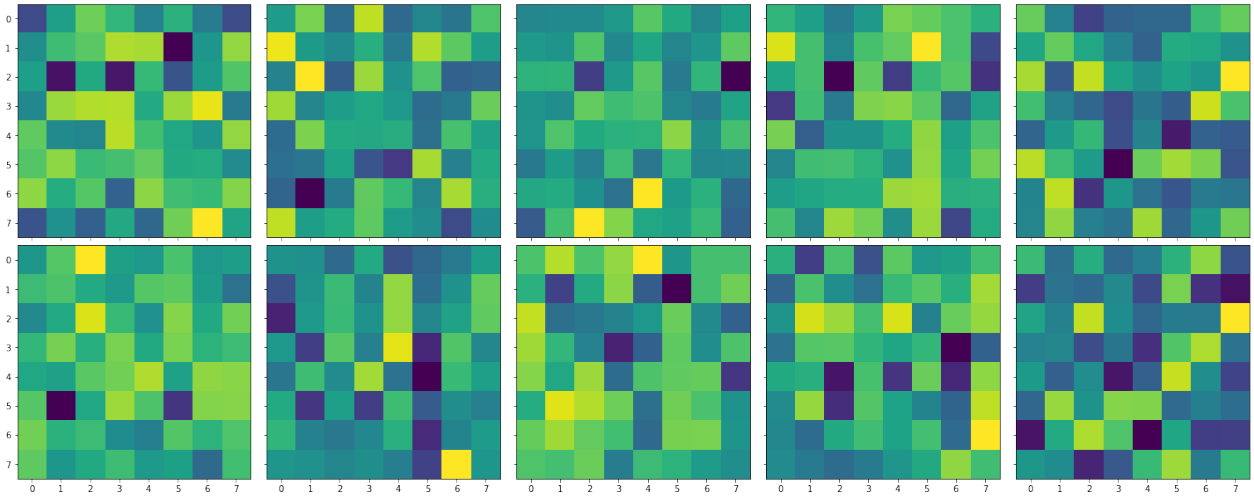
**Συμπεράσματα** Αμφότερες οι μέθοδοι επιτυγχάνουν συμπίεση του δικτύου. Μάλιστα, υψηλές τιμές συμπίεσης $(90 - 95\%)$ επιφέρουν μηδαμινή πτώση του ποσοστού επιτυχίας. Η παρατήρηση αυτή ισχύει ακόμα και για πολύ μικρά δίκτυα (πχ με 24 νευρώνες στο κρυφό επίπεδο), όπου η χρήση αποκλειστικά 3% των μονάδων συνεπάγεται ότι χρησιμοποιούνται λιγότερες από 1000 μονάδες στο μορφολογικό επίπεδο.

Αξιοσημείωτη είναι η ευρωστία που επιδεικνύουν οι αναπαραστάσεις που προκύπτουν από τον αλγόριθμο βελτιστοποίησης της Στοχαστικής Κατάβασης Κλίσης. Χαρακτηριστικό παράδειγμα αποτελεί η περίπτωση του μεικτού δικτύου για το σύνολο δεδομένων `MNIST` (βλ. πίνακα 6.14), όπου οι στήλες για τον αλγόριθμο SGD παρουσιάζουν ελάχιστη διακύμανση. Ιδιαίτερα για μεγαλύτερα δίκτυα ($l \geq 128$), η απόδοση της πρώτης και της τελευταίας γραμμής, δηλαδή ολόκληρου του δικτύου και του δικτύου με το 1% του κρυφού επιπέδου είναι πανομοιότυπη.

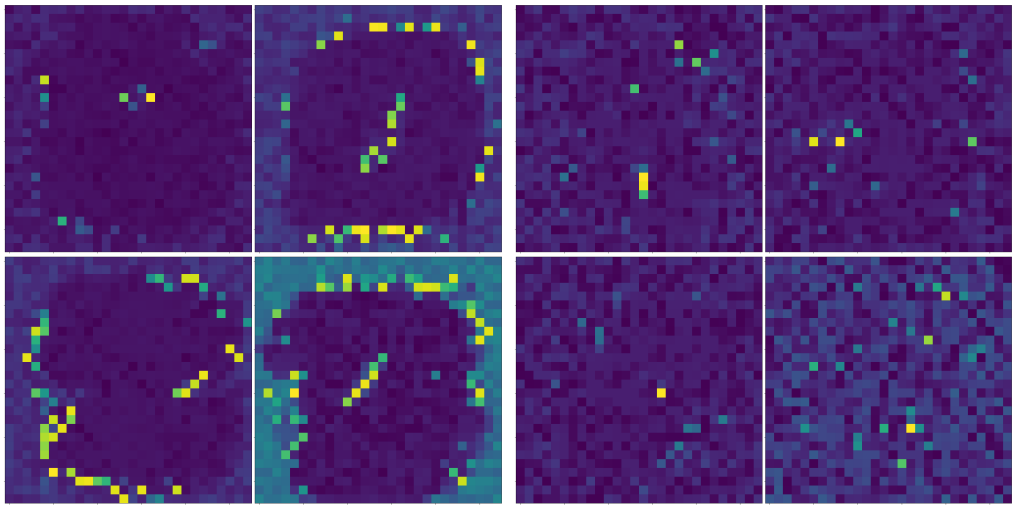
Παρόμοια συμπεριφορά χαρακτηρίζει τον αλγόριθμο Adaptive Momentum Estimation (adam), μόνο που το κατώφλι φαίνεται να είναι στο 3%, καθώς η συμπίεση πέρα από αυτό οδηγεί σε μεγάλες πτώσεις στην απόδοση. Ποιοτικά, αυτό εξηγείται από την εικόνα 6.2.6, όπου συγκρίνονται οι ενεργοποιήσεις που προκύπτουν από τους δύο αλγορίθμους για ίδια αρχιτεκτονική (επιλέγονται οι 4 πρώτοι νευρώνες για δίκτυο με μόνο 64 διαστολές στο κρυφό επίπεδο). Ο αλγόριθμος Adam χρησιμοποιεί περισσότερες μονάδες και σκιαγραφεί, τρόπον τινά, τα διάφορα ψηφεία. Συνεπώς, το κατώφλι 3% επιτρέπει τη διατήρηση αυτών των μονάδων και δεν επιφέρει μεταβολές στην απόδοση.

Τέλος, εστιάζουμε στη σύγκριση των δύο μεθόδων. Η πρώτη μέθοδος παράγει ομοιόμορφο κλάδεμα στους νευρώνες, ενώ η δεύτερη εξαρτάται από τις τιμές ενεργοποίησης στον καθένα. Και στις δύο περιπτώσεις, η συμπίεση του δικτύου είναι σημαντική και οι αποδόσεις των δύο μεθόδων συγκρίσιμες. Από τους πίνακες, μπορούμε να αποφανθούμε ότι η δεύτερη μέθοδος συνδυάζεται βέλτιστα με τον αλγόριθμο Adaptive Momentum Estimation, ενώ τα μοντέλα που έχουν εκπαιδευθεί με Στοχαστική Κατάβαση Κλίσης συμπεριφέρονται βέλτιστα στην πρώτη μέθοδο pruning. Αυτές οι παρατηρήσεις βασίζονται σε σύγκριση των ποσοστών επιτυχίας στα test δεδομένα για ίδια ποσοστά διατήρησης των συνδέσεων του αρχικού δικτύου. Τέλος, αξίζει να σημειωθεί ότι η δεύτερη μέθοδος είναι πιο εύχρηστη, καθώς η επιλογή τιμών για την παράμετρο $p$ είναι πιο διαισθητική από την επιλογή ακριβούς ποσοστού για κλάδεμα.

Για παράδειγμα, για $p = 0.2$ διατηρείται λιγότερο από 5% των συνδέσεων για το σύνολο δεδομένων `MNIST` και 10% των συνδέσεων για το σύνολο δεδομένων `FashionMNIST`. Η επιλογή μίας τέτοιας τιμής συμβαδίζει με τη διαίσθηση, καθώς γνωρίζουμε ότι διατηρούνται οι "σημαντικές" συνδέσεις και γνωρίζουμε εκ των προτέρων ότι η απόδοση του εναπομείνοντος δικτύου θα είναι συγκρίσιμη με το αρχικό. Αντίθετα, η επιλογή των αντίστοιχων ποσοστών μπορεί να αποδειχθεί ζημιογόνα, ιδιαίτερα σε πιο σύνθετα σύνολα δεδομένων.

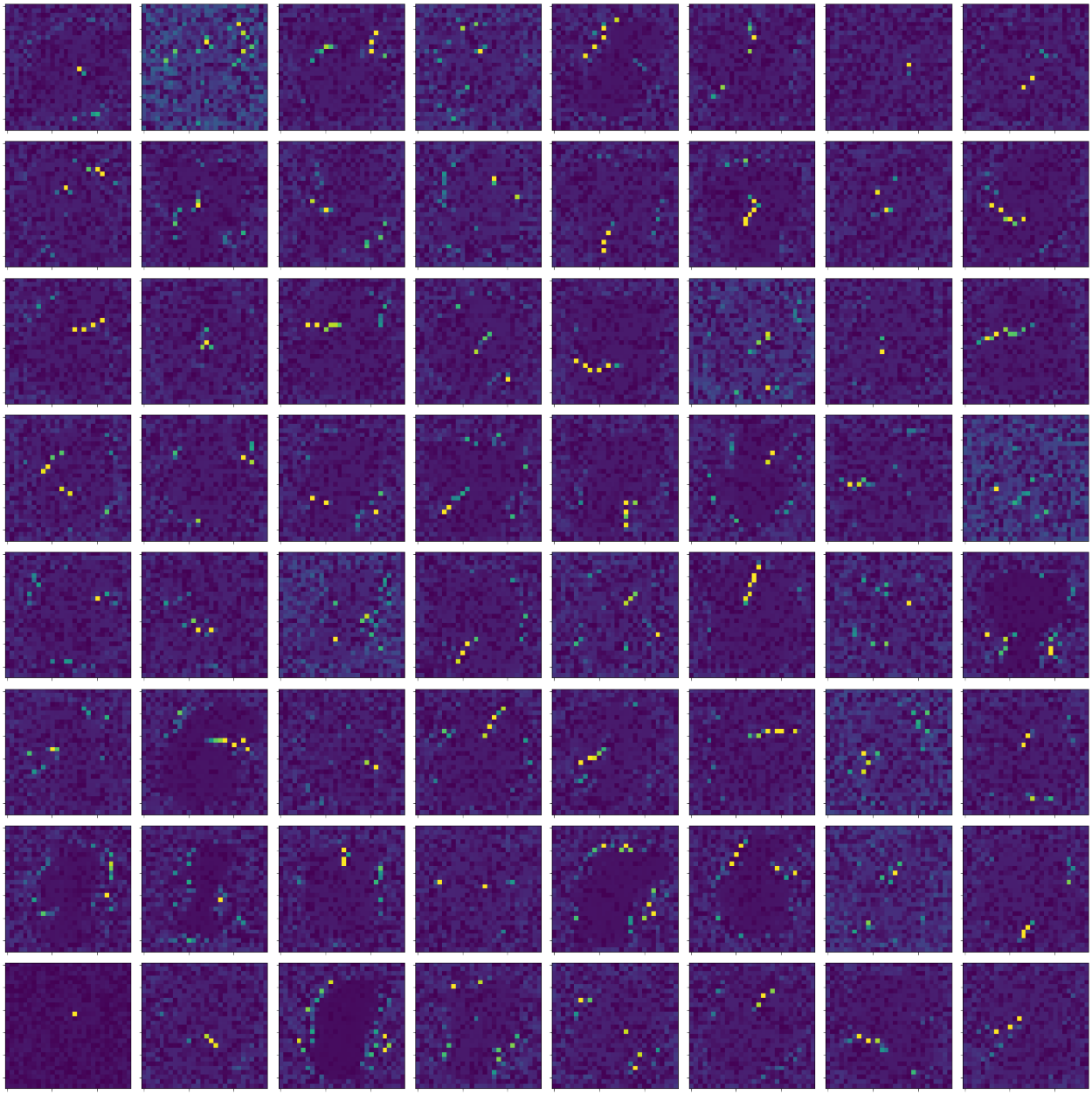


Σχήμα 6.2.5: `MNIST`: ενεργοποιήσεις επιπέδου γραμμικού συνδυασμού



(a) Adaptive Momentum Estimation (b) Stochastic Gradient Descent (SGD)

Σχήμα 6.2.6: Σύγκριση ενεργοποιήσεων σε σχέση με αλγόριθμο βελτιστοποίησης



Σχήμα 6.2.7: MNIST: ενεργοποιήσεις μορφολογικού επιπέδου

| | Adaptive Momentum Estimation | | | | Stochastic Gradient Descent | | | |
|---|---|---|---|---|---|---|---|---|
| | 24 | 32 | 64 | 128 | 24 | 32 | 64 | 128 |
| 100% | 92.47 | 94.36 | 95.57 | 96.90 | 92.23 | 92.83 | 93.72 | 95.29 |
| 90% | 92.46 | 94.35 | 95.57 | 96.89 | 92.23 | 92.83 | 93.72 | 95.29 |
| 80% | 92.46 | 94.34 | 95.52 | 96.89 | 92.23 | 92.83 | 93.72 | 95.30 |
| 70% | 92.46 | 94.34 | 95.51 | 96.89 | 92.24 | 92.84 | 93.71 | 95.30 |
| 60% | 92.45 | 94.29 | 95.54 | 96.91 | 92.22 | 92.82 | 93.72 | 95.31 |
| 50% | 92.45 | 94.33 | 95.47 | 96.92 | 92.23 | 92.80 | 93.71 | 95.33 |
| 40% | 92.52 | 94.17 | 95.37 | 96.86 | 92.20 | 92.80 | 93.73 | 95.31 |
| 30% | 92.13 | 94.03 | 95.26 | 96.79 | 92.17 | 92.78 | 93.72 | 95.33 |
| 20% | 91.94 | 93.75 | 95.10 | 96.70 | 92.17 | 92.73 | 93.80 | 95.30 |
| 10% | 91.53 | 93.28 | 94.68 | 96.47 | 92.18 | 92.73 | 93.77 | 95.32 |
| 9% | 91.25 | 93.16 | 94.46 | 96.38 | 92.18 | 92.74 | 93.77 | 95.32 |
| 8% | 90.66 | 93.05 | 94.42 | 96.22 | 92.18 | 92.74 | 93.75 | 95.32 |
| 7% | 90.39 | 92.98 | 94.54 | 96.19 | 92.18 | 92.74 | 93.76 | 95.31 |
| 6% | 90.36 | 92.92 | 94.37 | 96.10 | 92.18 | 92.74 | 93.77 | 95.31 |
| 5% | 90.31 | 92.39 | 93.80 | 96.07 | 92.17 | 92.74 | 93.77 | 95.32 |
| 4% | 89.96 | 92.31 | 93.61 | 95.89 | 92.17 | 92.74 | 93.78 | 95.28 |
| 3% | 87.60 | 90.35 | 92.95 | 95.44 | 92.13 | 92.75 | 93.76 | 95.27 |
| 2% | 76.16 | 76.90 | 86.73 | 92.27 | 91.98 | 92.67 | 93.68 | 95.23 |
| 1% | 37.68 | 60.74 | 55.74 | 74.15 | 90.02 | 91.95 | 93.21 | 95.12 |

Πίνακας 6.13: `MNIST`: Αξιολόγηση της επίδρασης της πρώτης μεθόδου Pruning στην ακρίβεια του Μορφολογικού δικτύου *μόνο* με όρους Dilation

| | Adaptive Momentum Estimation | | | | | | Stochastic Gradient Descent | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 24 | 32 | 64 | 128 | 256 | 400 | 24 | 32 | 64 | 128 | 256 | 400 |
| 100% | 89.75 | 92.19 | 94.75 | 96.47 | 97.42 | 97.63 | 89.10 | 90.00 | 92.73 | 94.73 | 95.74 | 96.07 |
| 90% | 89.76 | 92.20 | 94.76 | 96.48 | 97.41 | 97.62 | 89.09 | 90.01 | 92.73 | 94.74 | 95.74 | 96.07 |
| 80% | 89.76 | 92.20 | 94.76 | 96.49 | 97.40 | 97.64 | 89.09 | 90.01 | 92.74 | 94.73 | 95.73 | 96.07 |
| 70% | 89.78 | 92.22 | 94.73 | 96.52 | 97.43 | 97.61 | 89.07 | 90.01 | 92.73 | 94.73 | 95.73 | 96.08 |
| 60% | 89.74 | 92.23 | 94.68 | 96.47 | 97.37 | 97.61 | 89.08 | 89.99 | 92.73 | 94.72 | 95.73 | 96.05 |
| 50% | 89.75 | 92.17 | 94.55 | 96.43 | 97.38 | 97.55 | 89.09 | 89.97 | 92.74 | 94.72 | 95.75 | 96.05 |
| 40% | 89.70 | 92.18 | 94.48 | 96.44 | 97.29 | 97.50 | 89.06 | 89.96 | 92.71 | 94.73 | 95.72 | 96.03 |
| 30% | 89.52 | 91.93 | 94.32 | 96.23 | 97.24 | 97.49 | 89.04 | 89.96 | 92.72 | 94.71 | 95.74 | 96.02 |
| 20% | 89.12 | 91.84 | 94.27 | 96.15 | 97.24 | 97.36 | 89.03 | 89.91 | 92.72 | 94.77 | 95.73 | 95.98 |
| 10% | 89.06 | 91.59 | 93.92 | 95.88 | 97.20 | 97.05 | 89.04 | 89.92 | 92.71 | 94.78 | 95.69 | 96.01 |
| 9% | 88.81 | 91.56 | 93.90 | 95.88 | 97.21 | 97.02 | 89.04 | 89.91 | 92.71 | 94.78 | 95.70 | 96.00 |
| 8% | 88.76 | 91.37 | 93.90 | 95.84 | 97.17 | 97.02 | 89.04 | 89.91 | 92.71 | 94.80 | 95.72 | 96.03 |
| 7% | 88.74 | 91.33 | 93.92 | 95.86 | 97.13 | 97.02 | 89.04 | 89.91 | 92.71 | 94.80 | 95.73 | 96.02 |
| 6% | 88.25 | 91.29 | 93.85 | 95.90 | 97.10 | 97.04 | 89.06 | 89.91 | 92.71 | 94.80 | 95.73 | 96.01 |
| 5% | 88.29 | 91.35 | 93.79 | 95.78 | 97.10 | 96.95 | 89.04 | 89.91 | 92.71 | 94.80 | 95.71 | 96.02 |
| 4% | 87.97 | 91.12 | 93.70 | 95.72 | 97.05 | 96.91 | 89.04 | 89.93 | 92.71 | 94.79 | 95.73 | 96.01 |
| 3% | 86.22 | 88.99 | 92.68 | 94.83 | 96.12 | 96.44 | 89.00 | 89.92 | 92.70 | 94.81 | 95.70 | 96.01 |
| 2% | 76.72 | 80.82 | 88.10 | 92.25 | 92.20 | 90.54 | 88.71 | 89.71 | 92.60 | 94.79 | 95.70 | 96.02 |
| 1% | 53.00 | 52.66 | 62.68 | 83.75 | 83.64 | 83.06 | 85.93 | 88.07 | 91.75 | 94.50 | 95.62 | 96.08 |

Πίνακας 6.14: `MNIST`: Αξιολόγηση της επίδρασης της πρώτης μεθόδου Pruning στην ακρίβεια του Μορφολογικού δικτύου με όρους dilation και erosion στο κρυφό επίπεδο

| | Adaptive Momentum Estimation | | | | Stochastic Gradient Descent | | | |
|---|---|---|---|---|---|---|---|---|
| | 24 | 32 | 64 | 128 | 24 | 32 | 64 | 128 |
| 100% | 78.89 | 80.98 | 83.22 | 84.70 | 76.46 | 77.56 | 80.12 | 78.72 |
| 90% | 78.89 | 80.94 | 83.23 | 84.67 | 76.46 | 77.56 | 80.12 | 78.72 |
| 80% | 78.89 | 80.96 | 83.30 | 84.69 | 76.45 | 77.55 | 80.12 | 78.73 |
| 70% | 78.87 | 80.96 | 83.26 | 84.70 | 76.43 | 77.53 | 80.13 | 78.77 |
| 60% | 78.87 | 80.99 | 83.22 | 84.72 | 76.45 | 77.53 | 80.13 | 78.77 |
| 50% | 78.88 | 80.98 | 83.30 | 84.63 | 76.45 | 77.52 | 80.06 | 78.69 |
| 40% | 78.84 | 81.01 | 83.31 | 84.56 | 76.47 | 77.48 | 80.02 | 78.68 |
| 30% | 78.56 | 80.83 | 82.96 | 84.61 | 76.44 | 77.46 | 80.01 | 78.73 |
| 20% | 78.42 | 80.63 | 82.79 | 84.39 | 76.41 | 77.44 | 79.92 | 78.73 |
| 10% | 77.81 | 80.24 | 82.21 | 84.36 | 76.49 | 77.41 | 79.89 | 78.84 |
| 9% | 77.81 | 80.24 | 82.25 | 84.28 | 76.48 | 77.41 | 79.89 | 78.87 |
| 8% | 77.66 | 80.21 | 82.20 | 84.16 | 76.48 | 77.41 | 79.90 | 78.78 |
| 7% | 77.48 | 79.99 | 82.20 | 84.08 | 76.45 | 77.39 | 79.95 | 78.69 |
| 6% | 77.47 | 79.93 | 82.18 | 84.04 | 76.44 | 77.37 | 79.95 | 78.86 |
| 5% | 77.48 | 79.71 | 82.12 | 83.94 | 76.42 | 77.39 | 79.91 | 78.87 |
| 4% | 77.32 | 79.49 | 82.06 | 83.93 | 76.46 | 77.36 | 79.92 | 78.88 |
| 3% | 77.27 | 79.41 | 81.78 | 83.99 | 76.49 | 77.24 | 79.93 | 78.86 |
| 2% | 77.23 | 79.28 | 81.01 | 83.48 | 76.44 | 77.30 | 79.80 | 78.85 |
| 1% | 72.97 | 75.23 | 75.78 | 78.99 | 76.57 | 77.06 | 79.32 | 79.13 |

Πίνακας 6.15: `fashionMNIST`: Αξιολόγηση της επίδρασης της πρώτης μεθόδου Pruning στην ακρίβεια του Μορφολογικού δικτύου *μόνο* με όρους Dilation

| | Adaptive Momentum Estimation | | | | | | Stochastic Gradient Descent | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 24 | 32 | 64 | 128 | 256 | 400 | 24 | 32 | 64 | 128 | 256 | 400 |
| 100% | 82.05 | 82.05 | 85.11 | 86.76 | 87.11 | 88.34 | 80.08 | 78.86 | 83.03 | 82.74 | 84.89 | 86.21 |
| 90% | 82.05 | 82.05 | 85.12 | 86.72 | 87.12 | 88.34 | 80.08 | 78.87 | 83.03 | 82.75 | 84.88 | 86.21 |
| 80% | 82.08 | 82.06 | 85.18 | 86.72 | 87.19 | 88.33 | 80.06 | 78.87 | 83.03 | 82.74 | 84.91 | 86.21 |
| 70% | 82.05 | 82.05 | 85.10 | 86.77 | 87.16 | 88.34 | 80.08 | 78.89 | 82.99 | 82.74 | 84.90 | 86.22 |
| 60% | 82.02 | 82.05 | 85.08 | 86.79 | 87.14 | 88.33 | 80.09 | 78.87 | 83.01 | 82.79 | 84.92 | 86.28 |
| 50% | 82.03 | 82.02 | 85.14 | 86.64 | 87.13 | 88.31 | 80.10 | 78.90 | 83.02 | 82.80 | 84.95 | 86.20 |
| 40% | 82.09 | 82.02 | 85.08 | 86.59 | 87.18 | 88.29 | 80.07 | 78.88 | 83.00 | 82.72 | 84.94 | 86.32 |
| 30% | 82.01 | 81.97 | 85.02 | 86.63 | 87.21 | 88.28 | 80.01 | 78.90 | 82.99 | 82.71 | 84.96 | 86.26 |
| 20% | 81.81 | 81.91 | 85.06 | 86.49 | 86.99 | 88.06 | 79.97 | 78.92 | 83.01 | 82.70 | 84.94 | 86.31 |
| 10% | 81.27 | 81.68 | 84.90 | 86.27 | 86.61 | 87.73 | 79.97 | 78.89 | 82.97 | 82.72 | 84.86 | 86.38 |
| 9% | 81.25 | 81.60 | 84.89 | 86.24 | 86.45 | 87.63 | 79.97 | 78.91 | 82.98 | 82.70 | 84.90 | 86.35 |
| 8% | 81.22 | 81.60 | 84.90 | 86.19 | 86.58 | 87.61 | 79.97 | 78.91 | 82.98 | 82.73 | 84.87 | 86.35 |
| 7% | 81.19 | 81.62 | 84.81 | 86.26 | 86.42 | 87.62 | 79.97 | 78.92 | 82.97 | 82.74 | 84.90 | 86.37 |
| 6% | 81.17 | 81.49 | 84.79 | 86.32 | 86.20 | 87.51 | 79.97 | 78.92 | 82.97 | 82.69 | 84.86 | 86.34 |
| 5% | 81.15 | 81.48 | 84.90 | 86.32 | 86.26 | 87.22 | 79.98 | 78.98 | 82.97 | 82.66 | 84.88 | 86.38 |
| 4% | 81.02 | 81.49 | 84.58 | 86.12 | 86.28 | 87.07 | 79.99 | 78.96 | 82.89 | 82.57 | 84.94 | 86.39 |
| 3% | 81.01 | 81.39 | 84.28 | 86.16 | 86.02 | 86.43 | 79.98 | 78.96 | 82.88 | 82.49 | 84.94 | 86.35 |
| 2% | 80.92 | 81.28 | 84.20 | 86.13 | 85.63 | 84.66 | 79.99 | 78.99 | 82.87 | 82.32 | 85.00 | 86.36 |
| 1% | 78.18 | 78.89 | 83.32 | 83.90 | 83.44 | 83.55 | 79.86 | 79.11 | 82.73 | 82.37 | 84.85 | 86.19 |

Πίνακας 6.16: `fashionMNIST`: Αξιολόγηση της επίδρασης της πρώτης μεθόδου Pruning στην ακρίβεια του Μορφολογικού δικτύου με όρους dilation και erosion στο κρυφό επίπεδο

| | Adaptive Momentum Estimation | | | Stochastic Gradient Descent | | |
|---|---|---|---|---|---|---|
| $p$ | 128 | 256 | 400 | 128 | 256 | 400 |
| - | 96.47 (100.00) | 97.42 (100.00) | 97.63 (100.00) | 94.73 (100.00) | 95.74 (100.00) | 96.07 (100.00) |
| 0.1 | 96.08 (4.60) | 97.26 (4.47) | 97.26 (4.39) | 94.80 (6.94) | 95.72 (9.32) | 96.04 (11.75) |
| 0.2 | 96.06 (3.52) | 97.17 (3.42) | 97.25 (3.35) | 94.70 (4.14) | 95.65 (3.71) | 96.01 (4.55) |
| 0.3 | 95.86 (3.06) | 97.19 (2.91) | 97.14 (2.83) | 94.56 (2.78) | 95.60 (2.22) | 95.91 (2.17) |
| 0.4 | 95.86 (2.67) | 97.11 (2.51) | 96.87 (2.41) | 94.44 (1.94) | 95.34 (1.49) | 95.68 (1.16) |
| 0.5 | 95.58 (2.33) | 96.82 (2.14) | 96.72 (2.04) | 94.05 (1.33) | 94.87 (1.03) | 95.39 (0.71) |
| 0.6 | 94.98 (2.02) | 95.87 (1.82) | 96.26 (1.75) | 93.66 (0.80) | 94.42 (0.71) | 94.87 (0.48) |
| 0.7 | 93.64 (1.74) | 93.93 (1.55) | 94.13 (1.48) | 93.10 (0.63) | 92.25 (0.49) | 94.17 (0.37) |
| 0.8 | 90.50 (1.46) | 89.12 (1.26) | 88.44 (1.21) | 89.90 (0.46) | 90.45 (0.32) | 93.55 (0.28) |
| 0.9 | 81.55 (1.18) | 83.36 (1.02) | 79.93 (0.97) | 86.22 (0.33) | 86.83 (0.23) | 90.28 (0.20) |

Πίνακας 6.17: `MNIST`: Αξιολόγηση της επίδρασης της δεύτερης μεθόδου Pruning στην ακρίβεια του Μορφολογικού δικτύου με όρους dilation και erosion στο κρυφό επίπεδο

| | Adaptive Momentum Estimation | | | Stochastic Gradient Descent | | |
|---|---|---|---|---|---|---|
| $p$ | 128 | 256 | 400 | 128 | 256 | 400 |
| - | 86.76 (100.00) | 87.11 (100.00) | 88.34 (100.00) | 82.74 (100.00) | 84.89 (100.00) | 86.21 (100.00) |
| 0.1 | 86.53 (7.01) | 86.84 (8.87) | 88.15 (10.13) | 82.61 (12.06) | 85.02 (17.57) | 86.28 (19.96) |
| 0.2 | 86.44 (4.81) | 86.83 (5.91) | 88.13 (6.60) | 82.44 (5.01) | 84.95 (8.20) | 86.25 (9.39) |
| 0.3 | 86.36 (3.43) | 86.92 (3.98) | 88.01 (4.33) | 82.31 (2.55) | 84.90 (3.97) | 86.20 (4.32) |
| 0.4 | 86.07 (2.52) | 86.47 (2.69) | 87.63 (2.82) | 82.00 (1.50) | 84.72 (2.16) | 85.79 (2.05) |
| 0.5 | 85.86 (1.91) | 85.80 (1.86) | 87.04 (1.86) | 81.72 (0.93) | 84.24 (1.37) | 85.71 (1.08) |
| 0.6 | 85.02 (1.39) | 85.80 (1.29) | 85.90 (1.23) | 81.20 (0.62) | 82.86 (0.94) | 85.36 (0.63) |
| 0.7 | 84.11 (0.99) | 85.05 (0.90) | 85.18 (0.83) | 80.80 (0.37) | 82.02 (0.63) | 84.82 (0.40) |
| 0.8 | 80.94 (0.70) | 83.12 (0.62) | 81.68 (0.55) | 79.10 (0.28) | 81.63 (0.39) | 82.54 (0.28) |
| 0.9 | 76.90 (0.48) | 78.99 (0.40) | 75.59 (0.34) | 75.09 (0.20) | 78.17 (0.25) | 80.09 (0.20) |

Πίνακας 6.18: `fashionMNIST`: Αξιολόγηση της επίδρασης της δεύτερης μεθόδου Pruning στην ακρίβεια του Μορφολογικού δικτύου με όρους dilation και erosion στο κρυφό επίπεδο

### 6.2.5 Σύγκριση Μορφολογικών Νευρωνικών με Διαφορετικές Αρχιτεκτονικές

Συγκρίνουμε τα μορφολογικά νευρωνικά δίκτυα των προηγούμενων ενοτήτων με παραδοσιακά νευρωνικά δίκτυα. Χρησιμοποιούμε νευρωνικά δίκτυα ίσης διάστασης αντικαθιστώντας το επίπεδο με νευρώνες διαστολής-συστολής με πλήρως συνδεδεμένο. Στο πλαίσιο αυτό, απαιτείται και ο ορισμός μίας συνάρτησης ενεργοποίησης ώστε τα επίπεδα να διαχωρίζονται από ένα μη-γραμμικό μετασχηματισμό. Επιλέγουμε Rectified Linear Unit.

Τα αποτελέσματα για τα σύνολα δεδομένων `MNIST` και `FashionMNIST` φαίνονται στους πίνακες 6.19 και 6.20, αντίστοιχα. Με $\delta$ και $\epsilon$ συμβολίζουμε τα μορφολογικά δίκτυα μόνο με όρους dilation και μόνο με erosion αντίστιχα. Για $\delta, \epsilon$ έχουμε το μεικτό δίκτυο, με $2(\delta, \epsilon)$ το μορφολογικό δίκτυο με δύο μεικτά επίπεδα, ενώ με $\delta_\beta, \epsilon_\beta$ συμβολίζεται το ομαλοποιημένο μορφολογικό επίπεδο με παράμετρο σκληρότητας $\beta$[2]. Παρατηρούμε ότι τα δύο είδη δικτύων, μορφολογικά στις διάφορες παραλλαγές τους και τα "κανονικά", έχουν συγκρίσιμες αποδόσεις, με το Feedforward Νευρωνικό Δίκτυο με ReLU ενεργοποιήσεις (FF-ReLU) να εμφανίζει ελαφρώς καλύτερα αποτελέσματα.

Ωστόσο, ένα από τα βασικά χαρακτηριστικά των μορφολογικών νευρωνικών δικτύων είναι η οικονομία τους, όπως μελετήθηκε στην υποενότητα 6.2.4. Αυτό είναι ιδιαίτερα εμφανές από το σχήμα 6.2.8, όπου απεικονίζονται οι ενεργοποιήσεις του κρυφού επιπέδου για το δίκτυο με 400 νευρώνες σε αυτό. Τα συμπεράσματα είναι αντιδιαμετρικά με αυτά του μορφολογικού δικτύου, βλ. σχήμα 6.2.7. Ποσοτικά, αυτή η σύγκριση παρουσιάζεται στους πίνακες 6.21 και 6.22, όπου εφαρμόζεται η πρώτη μέθοδος pruning. Ας επικεντρώσουμε την προσοχή μας στις στήλες για το κρυφό δίκτυο με τους μέγιστους νευρώνες (400) τόσο για αλγόριθμο βελτιστοποίησης Adaptive Momentum Estimation όσο και για Stochastic Gradient Descent. Ενώ τα μορφολογικά δίκτυα διατηρούν την απόδοσή τους ακόμα και με 1% των αρχικών κόμβων, τα δίκτυα FF-ReLU δεν παρουσιάζουν τον ίδιο βαθμό ευρωστίας. Ιδιαίτερα σε πιο σύνθετα σύνολα δεδομένων, όπως `FashionMNIST`, η απόδοσή τους σταματά να είναι συγκρίσιμη με την αρχική για ποσοστό διατήρησης $\sim 20-30\% \gg 1\%$, γεγονός που συνεπάγεται ότι χρειάζονται 20 με 30 φορές περισσότεροι παράμετροι για το ίδιο αποτέλεσμα.

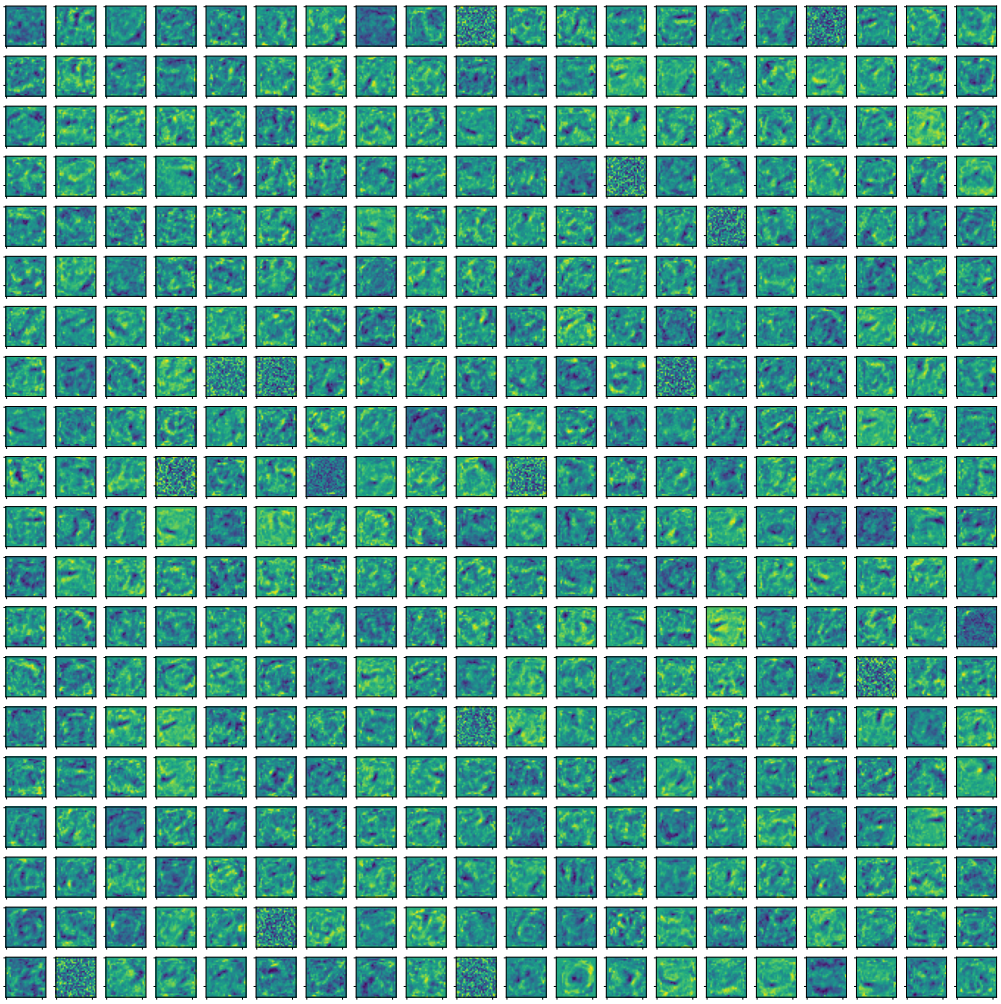
| | Adaptive Momentum Estimation | | | | | | Stochastic Gradient Descent | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ♯ | $\delta$ | $\epsilon$ | $\delta, \epsilon$ | $2(\delta, \epsilon)$ | $\delta_\beta, \epsilon_\beta$ | FF-ReLU | $\delta$ | $\epsilon$ | $\delta, \epsilon$ | FF-ReLU |
| 24 | 92.77 | 84.30 | 90.22 | 91.58 | - | 96.25 | 92.23 | 75.75 | 89.10 | - |
| 32 | 94.36 | 87.12 | 92.19 | 92.73 | - | 96.54 | 92.83 | 82.40 | 90.00 | 97.00 |
| 64 | 96.12 | 91.74 | 94.92 | 95.15 | - | 97.53 | 94.65 | 87.98 | 93.00 | 97.56 |
| 128 | 96.90 | 94.91 | 96.59 | 96.33 | - | 98.07 | 95.29 | 91.68 | 94.73 | 97.92 |
| 256 | - | - | 97.42 | 97.08 | - | 98.15 | - | - | 95.74 | 98.02 |
| 400 | - | - | 97.63 | 97.09 | 97.84 | 98.03 | - | - | 96.07 | 98.08 |

Πίνακας 6.19: Σύγκριση της ακρίβειας των βέλτιστων αρχιτεκτονικών των μεθόδων στο σύνολο δεδομένων `MNIST`. Ο συμβολισμός για τις στήλες περιγράφεται παραπάνω.

| | Adaptive Momentum Estimation | | | | | | Stochastic Gradient Descent | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ♯ | $\delta$ | $\epsilon$ | $\delta, \epsilon$ | $2(\delta, \epsilon)$ | $\delta_\beta, \epsilon_\beta$ | FF-ReLU | $\delta$ | $\epsilon$ | $\delta, \epsilon$ | FF-ReLU |
| 24 | 79.63 | 80.53 | 82.05 | 82.49 | - | 86.78 | 76.46 | 77.55 | 80.08 | - |
| 32 | 80.98 | 81.87 | 83.16 | 83.60 | - | 87.06 | 77.75 | 79.43 | 79.82 | 87.49 |
| 64 | 84.09 | 84.38 | 85.20 | 85.34 | - | 87.83 | 80.41 | 82.44 | 83.10 | 88.29 |
| 128 | 85.92 | 86.13 | 86.76 | 86.51 | - | 88.44 | 81.21 | 84.18 | 83.98 | 88.66 |
| 256 | - | - | 87.11 | 87.12 | - | 89.09 | - | - | 85.25 | 87.69 |
| 400 | - | - | 88.34 | 87.47 | 87.13 | 89.44 | - | - | 86.21 | 88.81 |

Πίνακας 6.20: Σύγκριση της ακρίβειας των βέλτιστων αρχιτεκτονικών των μεθόδων στο σύνολο δεδομένων `FashionMNIST`. Ο συμβολισμός για τις στήλες περιγράφεται παραπάνω.

[2]Ο πίνακας περιλαμβάνει το ομαλοποιημένο δίκτυο με την καλύτερη απόδοση από τις διάφορες παραμέτρους σκληρότητας.



Σχήμα 6.2.8: Ενεργοποιήσεις κρυφού επιπέδου ενός Feedforward Νευρωνικού Δικτύου με ReLU ενεργοποιήσεις

| | Adaptive Momentum Estimation | | | | | | Stochastic Gradient Descent | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 24 | 32 | 64 | 128 | 256 | 400 | 24 | 32 | 64 | 128 | 256 | 400 |
| 100% | 95.80 | 96.34 | 97.47 | 97.93 | 98.15 | 97.84 | 96.26 | 96.82 | 97.41 | 97.80 | 98.00 | 98.16 |
| 90% | 95.83 | 96.36 | 97.49 | 97.93 | 98.15 | 97.85 | 96.21 | 96.81 | 97.43 | 97.80 | 98.00 | 98.17 |
| 80% | 95.76 | 96.37 | 97.44 | 97.93 | 98.14 | 97.85 | 96.21 | 96.82 | 97.42 | 97.82 | 98.00 | 98.13 |
| 70% | 95.85 | 96.31 | 97.46 | 97.94 | 98.14 | 97.84 | 96.18 | 96.75 | 97.44 | 97.81 | 97.97 | 98.14 |
| 60% | 95.43 | 96.25 | 97.31 | 97.95 | 98.18 | 97.83 | 96.24 | 96.79 | 97.44 | 97.82 | 98.01 | 98.12 |
| 50% | 94.66 | 95.42 | 96.99 | 97.82 | 98.13 | 97.79 | 96.05 | 96.70 | 97.51 | 97.76 | 98.02 | 98.10 |
| 40% | 93.44 | 91.84 | 96.22 | 97.65 | 97.98 | 97.79 | 96.05 | 96.54 | 97.48 | 97.71 | 97.93 | 98.08 |
| 30% | 80.01 | 78.53 | 94.01 | 96.78 | 97.59 | 97.51 | 95.72 | 95.90 | 97.38 | 97.75 | 97.91 | 97.97 |
| 20% | 49.73 | 54.14 | 80.00 | 94.17 | 95.77 | 96.70 | 94.71 | 94.72 | 96.91 | 97.45 | 97.79 | 97.89 |
| 10% | 19.68 | 34.91 | 40.67 | 79.93 | 86.12 | 90.12 | 89.09 | 84.37 | 91.61 | 95.57 | 95.60 | 96.84 |
| 9% | 19.86 | 33.98 | 38.12 | 77.54 | 83.23 | 88.46 | 87.00 | 82.83 | 90.47 | 94.57 | 95.18 | 96.56 |
| 8% | 19.22 | 31.20 | 37.09 | 75.89 | 79.17 | 84.66 | 85.22 | 81.57 | 86.66 | 94.03 | 94.35 | 95.93 |
| 7% | 18.22 | 29.15 | 39.21 | 69.21 | 77.50 | 82.87 | 81.47 | 74.58 | 85.72 | 93.05 | 92.53 | 95.48 |
| 6% | 17.05 | 21.86 | 35.12 | 64.14 | 74.42 | 78.76 | 78.39 | 69.18 | 84.20 | 91.50 | 90.27 | 94.93 |
| 5% | 13.87 | 18.75 | 30.69 | 58.40 | 70.92 | 71.07 | 74.09 | 65.21 | 81.88 | 89.06 | 86.02 | 93.82 |
| 4% | 13.19 | 17.16 | 21.99 | 52.24 | 65.27 | 69.24 | 65.58 | 59.61 | 74.62 | 83.57 | 76.20 | 91.19 |
| 3% | 12.11 | 16.28 | 18.64 | 40.01 | 60.21 | 66.09 | 57.45 | 44.78 | 71.68 | 72.84 | 67.50 | 85.80 |
| 2% | 11.07 | 14.09 | 17.93 | 28.56 | 47.46 | 58.85 | 44.14 | 35.63 | 59.04 | 54.79 | 58.10 | 76.93 |
| 1% | 11.28 | 13.69 | 13.69 | 23.56 | 28.38 | 43.86 | 41.41 | 31.55 | 34.47 | 41.38 | 49.83 | 60.06 |

Πίνακας 6.21: `MNIST`: Αξιολόγηση της επίδρασης της πρώτης μεθόδου Pruning στην ακρίβεια του Feedforward Νευρωνικού Δικτύου με ReLU ενεργοποιήσεις

| | Adaptive Momentum Estimation | | | | | | Stochastic Gradient Descent | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 24 | 32 | 64 | 128 | 256 | 400 | 24 | 32 | 64 | 128 | 256 | 400 |
| 100% | 86.05 | 86.71 | 88.06 | 88.49 | 89.06 | 88.82 | 86.04 | 87.20 | 85.58 | 87.87 | 88.01 | 87.79 |
| 90% | 86.07 | 86.72 | 88.14 | 88.43 | 89.04 | 88.81 | 86.08 | 87.21 | 85.60 | 87.83 | 88.05 | 87.74 |
| 80% | 86.01 | 86.80 | 88.19 | 88.53 | 88.97 | 88.76 | 86.11 | 87.27 | 85.72 | 87.84 | 88.01 | 87.82 |
| 70% | 86.01 | 86.67 | 88.19 | 88.51 | 88.86 | 88.70 | 85.75 | 87.20 | 85.54 | 87.95 | 88.11 | 87.92 |
| 60% | 85.98 | 86.17 | 88.24 | 88.61 | 88.78 | 88.74 | 85.45 | 87.23 | 85.43 | 87.92 | 87.91 | 87.68 |
| 50% | 85.75 | 86.30 | 87.67 | 87.93 | 88.50 | 88.20 | 85.15 | 87.18 | 85.52 | 87.97 | 87.77 | 87.03 |
| 40% | 84.71 | 84.38 | 81.63 | 86.91 | 86.72 | 86.74 | 84.01 | 86.39 | 85.19 | 87.43 | 87.45 | 86.64 |
| 30% | 81.39 | 71.50 | 73.27 | 82.92 | 82.88 | 84.24 | 81.28 | 84.00 | 85.21 | 86.40 | 86.89 | 85.73 |
| 20% | 73.04 | 59.27 | 64.21 | 60.74 | 74.20 | 78.48 | 73.65 | 75.01 | 81.11 | 83.56 | 84.33 | 83.62 |
| 10% | 48.61 | 48.94 | 49.44 | 39.24 | 63.77 | 63.85 | 53.85 | 53.48 | 67.06 | 69.28 | 71.63 | 69.59 |
| 9% | 44.34 | 46.00 | 47.15 | 38.14 | 56.50 | 62.20 | 49.35 | 51.79 | 66.16 | 68.12 | 69.37 | 66.75 |
| 8% | 41.80 | 45.20 | 46.59 | 39.07 | 56.74 | 58.11 | 46.59 | 48.21 | 62.06 | 64.04 | 65.31 | 63.55 |
| 7% | 39.94 | 40.11 | 46.17 | 36.05 | 55.68 | 53.74 | 40.15 | 47.12 | 58.81 | 58.85 | 61.57 | 58.22 |
| 6% | 37.32 | 38.86 | 47.69 | 34.69 | 54.07 | 50.97 | 41.54 | 43.62 | 50.15 | 57.65 | 58.65 | 53.39 |
| 5% | 32.03 | 34.67 | 41.31 | 36.48 | 46.71 | 48.19 | 41.48 | 42.81 | 44.91 | 53.46 | 56.56 | 46.12 |
| 4% | 29.64 | 26.76 | 38.98 | 36.68 | 48.43 | 46.44 | 34.61 | 34.35 | 41.94 | 49.13 | 53.67 | 43.30 |
| 3% | 23.29 | 15.79 | 32.19 | 36.70 | 50.56 | 42.96 | 28.98 | 30.74 | 32.26 | 44.08 | 45.77 | 40.51 |
| 2% | 15.99 | 12.95 | 20.86 | 30.87 | 36.62 | 32.12 | 24.70 | 35.42 | 28.50 | 40.95 | 43.06 | 36.14 |
| 1% | 11.55 | 12.16 | 14.20 | 26.54 | 29.43 | 25.79 | 21.65 | 25.16 | 28.15 | 29.40 | 35.43 | 27.39 |

Πίνακας 6.22: `FashionMNIST`: Αξιολόγηση της επίδρασης της πρώτης μεθόδου Pruning στην ακρίβεια του Feedforward Νευρωνικού Δικτύου με ReLU ενεργοποιήσεις

# Κεφάλαιο 7

# Επίλογος

## 7.1 Σύνοψη και Συμπεράσματα

Στο σημείο αυτό ολοκληρώνεται η διπλωματική εργασία. Συνοψίζουμε, λοιπόν, τη συνεισφορά της μελέτης. Αυτή εντοπίζεται στους ακόλουθους άξονες:

- Μελέτη και μοντελοποίηση προβλημάτων τροπικής βελτιστοποίησης ως συλλογή προβλημάτων κλασικής βελτιστοποίησης. Αναλυτικότερα, τα προβλήματα ελαχιστοποίησης τροπικού πολυωνύμου υπό τροπικές συνθήκες αναλύονται σε κυρτά προβλήματα, ενώ τα προβλήματα τροπικού κλασματικού προγραμματισμού μοντελοποιούνται ως διαφορά κυρτών συναρτήσεων (DC programming).
- Ανάλυση υπό την τροπική σκοπιά μονότονων νευρωνικών δικτύων. Συσχέτιση των περιοχών αποφάσεων με ιδέες απο Θεωρία Πλεγμάτων και έκφραση της μονότονης συνάρτησης καθώς και του αντίστροφου μετασχηματισμού της στη γλώσσα των Μορφολογικών Μαθηματκών, σε closing και opening αντίστοιχα.
- Ανάπτυξη αλγορίθμου κατασκευής τμηματικά γραμμικών και μονότονων επιφανειών σε αυθαίρετο πλήθος διαστάσεων.
- Μελέτη και μοντελοποίηση μορφολογικών δικτύων με τροπικούς νευρώνες. Έκφραση της περιοχής απόφασης στα τροπικά μαθηματικά μέσω τους γενικευμένου πλαισίου των αρθρωτικών συναρτήσεων.
- Επέκταση των μορφολογικών δικτύων σε περισσότερα επίπεδα ώστε να διαμορφωθούν ενεργοποιήσεις εφάμιλλες με opening και closing. Περαιτέρω επέκταση σε ομαλοποιημένη έκδοση των μορφολογικών τελεστών και σύνδεση με Maslov Dequantization.
- Εκτενής πειραματισμός των προαναφερθέντων μοντέλων σε γνωστά σύνολα δεδομένων της Όρασης Υπολογιστών που αφορούν προβλήματα κατηγοριοποίησης. Συγκριτική μελέτη της συμπεριφοράς των δικτύων για τους δύο πιο κοινούς αλγορίθμους βελτιστοποίησης της εκπαίδευσης: Stochastic Gradient Descent και Adaptive Momentum Estimation.
- Συμπίεση των μοντέλων μέσω δύο μεθόδων pruning. Αξιοποιείται η αραιότητα που επιφέρουν οι μορφολογικοί τελεστές ώστε να διατηρείται η απόδοσή των μοντέλων με (μικρό) υποσύνολο των αρχικών συνδέσεων.
- Επέκταση μεθόδων βελτιστοποίησης για εκπαίδευση μοντέλων μηχανικής μάθησης σε προβλήματα ταξινόμησης πολλών κλάσεων και πειραματισμός σε γνωστά σύνολα δεδομένων.

## 7.2 Μελλοντικές Επεκτάσεις

Αφετηρία αυτής της εργασίας ήταν η σύνδεση των τροπικών, καθώς και των μορφολογικών, μαθηματικών με μοντέλα της μηχανικής μάθησης. Είναι φανερό ότι το αντικείμενο δεν έχει εξαντληθεί στο στενό πλαίσιο αυτής της προσπάθειας. Μέσω της εργασίας, ωστόσο, εντοπίστηκαν ορισμένες δίοδοι για ανάπτυξη και περαιτέρω

μελέτη των ιδεών από τα προηγούμενα κεφάλαια. Αναλυτικότερα, θεωρούμε πως οι κύριοι άξονες επέκτασης είναι οι εξής:

- Εφαρμογή πιο πολύπλοκων πυρήνων κατά τη σύνθεση των προβλημάτων ταξινόμησης σε μορφή διαφοράς κυρτών συναρτήσεων (DC programming). Ιδιαίτερα σε προβλήματα Όρασης Υπολογιστών, ωφέλιμο κρίνεται να συμπεριλάβουμε πυρήνες ειδικά σχεδιασμένους για τα δεδομένα, επαυξάνοντας τα χαρακτηριστικά με τις εξόδους πυρηνών Sobel κ.α. Όσον αφορά την ταξινόμηση πολλών κλάσεων, αξίζει να διερευνηθούν στο πλαίσιο των Dilation-Erosion Perceptrons και άλλες μέθοδοι από τη εκτενή βιβλιογραφία των Μηχανών Διανυσμάτων Υποστήριξης, όπως *one-versus-the-rest*, soft-assignment αντί για hard και ιδέες όπως Directed Acyclic Graph Support Vector Machines (DAGSVMs) όπου οι δυαδικοί ταξινομητές οργανώνονται σε έναν κατευθυνόμενο άκυκλο γράφο. Σε πρακτικό επίπεδο, χρήσιμος κρίνεται ο πειραματισμός με παράλληλες μεθόδους στην εκπαίδευση των δυαδικών ταξινομητών, καθώς είναι ανεξάρτητοι μεταξύ τους και η όλη διαδικασία δύναται να επιταχυνθεί ιδιαίτερα. Τέλος, θα είχε ενδιαφέρον η αξιολόγηση των μεθόδων και σε άλλα σύνολα δεδομένων, τόσο δυαδικής όσο και μη ταξινόμησης.

- Επέκταση των Πυκνών Μορφολογικών Δικτύων σε βαθιές αρχιτεκτονικές (πέρα από 2 επίπεδα) και αξιολόγηση σε πιο σύνθετα σύνολα δεδομένων όπως CIFAR10. Ενδιαφέρον παρουσιάζει η εξερεύνηση, τόσο πειραματικά όσο και θεωρητικά, της χρήσης εμβολίμων πλήρως συνδεδεμένων επιπέδων μεταξύ των μορφολογικών επιπέδων και πώς αυτά τα υβριδικά δίκτυα, που συνδυάζουν στοιχεία τόσο από μορφολογικά όσο και από παραδοσιακά νευρωνικά δίκτυα, καθορίζουν την εκφραστικότητα του μοντέλου και το σύνορο απόφασης. Επιπρόσθετα, τα επόμενα βήματα είναι απαραίτητο να συμπεριλαμβάνουν τη μελέτη συνελικτικών νευρωνικών δικτύων (Convolutional Neural Networks) και των μορφολογικών αντιστοίχων τους, όπως σε [Mel+19; BCT16; SZS19; FFY20; Nog+19]. Τέλος, οι μέθοδοι pruning έδειξαν πολλά υποσχόμενα αποτελέσματα, γεγονός που ενθαρρύνει την εφαρμογή τους τόσο για συμπίεση ταξινομητών σε πιο σύνθετα σύνολα δεδομένων, όσο και για τη χρήση τους σε μοντέλα με συνελικτικούς όρους.

- Εξερεύνηση των μοντέλων που εγγυώνται μονοτονία σε μεγαλύτερο βάθος. Σε προβλήματα παλινδρόμησης, ενδιαφέρουσα κρίνεται η αξιολόγηση του αμιγώς αλγεβρικού αλγορίθμου σε πιο σύνθετα προβλήματα ακόμα και σε γενικές (μη-κυρτές και όχι αυστηρά μονότονες) συναρτήσεις. Ακόμη, τα μονοτονικά νευρωνικά δίκτυα χρήζουν περαιτέρω μελέτης και ιδιαίτερα η ομαλοποιημένη εκδοχή τους, καθώς αντιμετωπίζει ικανοποιητικά το πρόβλημα διάδοσης κλίσεων (gradient propagation) που χαρακτηρίζει τα μορφολογικά δίκτυα. Στο πλαίσιο αυτό, χρήσιμη κρίνεται η αξιολόγηση σε σύνολα δεδομένων τόσο για διεργασίες παλινδρόμησης όσο και ταξινόμησης.

# Παράρτημα Α

# Βιβλιογραφία


[AGG12] Akian, M., Gaubert, S., and Guterman, A. "Tropical polyhedra are equivalent to mean payoff games". In: *International Journal of Algebra and Computation* 22.01 (Feb. 2012), p. 1250001.

[AGG10] Akian, M., Gaubert, S., and Guterman, A. "The correspondence between tropical convexity and mean payoff games". In: *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2010)*. 2010, pp. 5–9.

[Aki+11] Akian, M. et al. "Best approximation in max-plus semimodules". In: *Linear Algebra and its Applications* 435.12 (2011), pp. 3261–3296.

[All+14] Allamigeon, X. et al. "Combinatorial simplex algorithms can solve mean payoff games". In: *SIAM Journal on Optimization* 24.4 (Jan. 2014), pp. 2096–2117.

[All+15] Allamigeon, X. et al. "Tropicalizing the Simplex Algorithm". In: *SIAM Journal on Discrete Mathematics* 29.2 (Jan. 2015), pp. 751–795.

[AW93] Archer, N. P. and Wang, S. "Application of the Back Propagation Neural Network Algorithm with Monotonicity Constraints for Two-Group Classification Problems*". In: *Decision Sciences* 24.1 (1993), pp. 60–75.

[BR07] Barmpoutis, A. and Ritter, G. X. "Orthonormal basis lattice neural networks". In: *Computational Intelligence Based on Lattice Theory*. Springer, 2007, pp. 45–58.

[BT97] Bertsimas, D. and Tsitsiklis, J. N. *Introduction to linear optimization*. Vol. 6. Athena Scientific Belmont, MA, 1997.

[BC90] Best, M. J. and Chakravarti, N. "Active set algorithms for isotonic regression; a unifying framework". In: *Mathematical Programming* 47.1-3 (1990), pp. 425–439.

[BCT16] Blot, M., Cord, M., and Thome, N. "Max-min convolutional neural networks for image classification". In: *2016 IEEE International Conference on Image Processing (ICIP)*. 2016, pp. 3678–3682.

[BV04] Boyd, S. P. and Vandenberghe, L. *Convex optimization*. Cambridge University Press, 2004. ISBN: 978-0-521-83378-3.

[Boy+07] Boyd, S. et al. "A tutorial on geometric programming". In: *Optimization and Engineering* 8.1 (2007), pp. 67–127.

[But10] Butkovič, P. *Max-linear systems: theory and algorithms*. Springer monographs in mathematics. Springer, 2010. ISBN: 978-1-84996-298-8.

[CGP18] Calafiore, G. C., Gaubert, S., and Possieri, C. "Log-sum-exp neural networks and posynomial models for convex and log-log-convex data". In: *arXiv:1806.07850 [cs]* (2018).

[Cal+19] Calafiore, G. C. et al. "A Universal Approximation Result for Difference of log-sum-exp Neural Networks". In: *arXiv:1905.08503 [cs]* (2019).

[CM17] Charisopoulos, V. and Maragos, P. "Morphological Perceptrons: Geometry and Training Algorithms". In: *Mathematical Morphology and Its Applications to Signal and Image Processing*. Ed. by J. Angulo, S. Velasco-Forero, and F. Meyer. Vol. 10225. Springer International Publishing, 2017, pp. 3–15. ISBN: 978-3-319-57239-0.

[CM18] Charisopoulos, V. and Maragos, P. "A Tropical Approach to Neural Networks with Piecewise Linear Activations". In: *arXiv:1805.08749 [cs, stat]* (May 2018).

[Cie+08] Cieliebak, M. et al. "On the Complexity of Variations of Equal Sum Subsets." In: *Nord. J. Comput.* 14.3 (2008), pp. 151–172.

[CGQ04] Cohen, G., Gaubert, S., and Quadrat, J.-P. "Duality and separation theorems in idempotent semimodules". In: *Linear Algebra and its Applications* 379 (2004), pp. 395–422.

[CP11] Combettes, P. L. and Pesquet, J.-C. "Proximal Splitting Methods in Signal Processing". In: *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Ed. by H. H. Bauschke et al. Vol. 49. Springer New York, 2011, pp. 185–212. ISBN: 978-1-4419-9568-1.

[CT16] Crowell, R. A. and Tran, N. M. "Tropical Geometry and Mechanism Design". In: *arXiv:1606.04880 [cs, math]* (June 2016).

[Cun79] Cuninghame-Green, R. A. *Minimax algebra*. Lecture notes in economics and mathematical systems. Springer, 1979. ISBN: 978-3-540-09113-4.

[DV10] Daniels, H. and Velikova, M. "Monotone and Partially Monotone Neural Networks". In: *IEEE Transactions on Neural Networks* 21.6 (2010), pp. 906–917.

[DH93] Davidson, J. L. and Hummer, F. "Morphology neural networks: An introduction with applications". In: *Circuits, Systems and Signal Processing* 12.2 (1993), pp. 177–210.

[DS04] Develin, M. and Sturmfels, B. "Tropical convexity". In: *Doc. Math* 9.1–27 (2004), pp. 7–8.

[FFY20] Franchi, G., Fehri, A., and Yao, A. "Deep morphological networks". In: *Pattern Recognition* 102 (2020), p. 107246.

[GG09] Garcia, E. and Gupta, M. "Lattice Regression". In: *Advances in Neural Information Processing Systems 22*. Ed. by Y. Bengio et al. Curran Associates, Inc., 2009, pp. 594–602.

[GK06] Gaubert, S. and Katz, R. "Max-Plus Convex Geometry". In: *Relations and Kleene Algebra in Computer Science*. Ed. by R. A. Schmidt. Vol. 4136. Springer Berlin Heidelberg, 2006, pp. 192–206. ISBN: 978-3-540-37873-0.

[GKS12] Gaubert, S., Katz, R. D., and Sergeev, S. "Tropical linear-fractional programming and parametric mean payoff games". In: *Journal of Symbolic Computation* 47.12 (2012), pp. 1447–1478.

[GN04] Gilbert, W. J. and Nicholson, W. K. *Modern algebra with applications*. 2nd ed. Pure and applied mathematics. Wiley-Interscience, 2004. ISBN: 978-0-471-41451-3.

[GB10] Glorot, X. and Bengio, Y. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 2010, pp. 249–256.

[GM08] Gondran, M. and Minoux, M. *Graphs, dioids and semirings: new models and algorithms*. Operations research/computer science interfaces. Springer, 2008. ISBN: 978-0-387-75449-9.

[Goo+13a] Goodfellow, I. J. et al. "Maxout Networks". In: *arXiv:1302.4389 [cs, stat]* (Feb. 2013).

[Goo+13b] Goodfellow, I. J. et al. "Maxout networks". In: *arXiv preprint arXiv:1302.4389* (2013).

[Gup+16] Gupta, M. et al. "Monotonic calibrated interpolated look-up tables". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 3790–3836.

[Gup+18] Gupta, M. et al. "Diminishing Returns Shape Constraints for Interpretability and Regularization". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., 2018, pp. 6834–6844.

[Har59] Hartman, P. "On functions representable as a difference of convex functions". In: *Pacific Journal of Mathematics* 9.3 (Sept. 1959), pp. 707–713.

[Hay09] Haykin, S. S. *Neural networks and learning machines/Simon Haykin*. New York: Prentice Hall, 2009.

[HZS17] He, Y., Zhang, X., and Sun, J. "Channel Pruning for Accelerating Very Deep Neural Networks". In: *arXiv:1707.06168 [cs]* (Aug. 2017).

[HA14] Hoburg, W. and Abbeel, P. "Geometric Programming for Aircraft Design Optimization". In: *AIAA Journal* 52.11 (Nov. 2014), pp. 2414–2426.

[HKA16] Hoburg, W., Kirschen, P., and Abbeel, P. "Data fitting with geometric-programming-compatible softmax functions". In: *Optimization and Engineering* 17.4 (Dec. 2016), pp. 897–918.

[KB17] Kingma, D. P. and Ba, J. "Adam: A Method for Stochastic Optimization". In: *arXiv:1412.6980 [cs]* (2017).

[LeC98] LeCun, Y. "The MNIST database of handwritten digits". In: *http://yann. lecun. com/exdb/mnist/* (1998).

[LeC+98] LeCun, Y. et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[LB16] Lipp, T. and Boyd, S. "Variations and extension of the convex–concave procedure". In: *Optimization and Engineering* 17.2 (June 2016), pp. 263–287.

[Lit07] Litvinov, G. L. "Maslov dequantization, idempotent and tropical mathematics: A brief introduction". In: *Journal of Mathematical Sciences* 140.3 (2007), pp. 426–444.

[Luc10] Lucet, Y. "What shape is your conjugate? A survey of computational convex analysis and its applications". In: *SIAM review* 52.3 (2010), pp. 505–542.

[MS15] Maclagan, D. and Sturmfels, B. *Introduction to tropical geometry*. Vol. 161. American Mathematical Soc., 2015.

[MB09] Magnani, A. and Boyd, S. P. "Convex piecewise-linear fitting". In: *Optimization and Engineering* 10.1 (2009), pp. 1–17.

[Mar15] Maragos, P. "Abstract Algebra, Symmetry Groups and Euclidean Motions". In: *Abstract Algebra* (2015), p. 20.

[Mar17] Maragos, P. "Dynamical systems on weighted lattices: general theory". In: *Mathematics of Control, Signals, and Systems* 29.4 (Dec. 2017).

[MT19] Maragos, P. and Theodosis, E. "Tropical Geometry and Piecewise-Linear Approximation of Curves and Surfaces on Weighted Lattices". In: *arXiv:1912.03891 [cs, math, stat]* (Dec. 2019).

[MP43] McCulloch, W. S. and Pitts, W. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4 (1943), pp. 115–133.

[Mel+19] Mellouli, D. et al. "Morphological Convolutional Neural Network Architecture for Digit Recognition". In: *IEEE Transactions on Neural Networks and Learning Systems* 30.9 (2019), pp. 2876–2885.

[Mil+16] Milani Fard, M. et al. "Fast and Flexible Monotonic Functions with Ensembles of Lattices". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 2919–2927.

[Mni+13] Mnih, V. et al. "Playing atari with deep reinforcement learning". In: *arXiv preprint arXiv:1312.5602* (2013).

[MSC19] Mondal, R., Santra, S., and Chanda, B. "Dense Morphological Network: An Universal Function Approximator". In: *arXiv:1901.00109 [cs, stat]* (2019).

[Mon+14] Montúfar, G. et al. "On the Number of Linear Regions of Deep Neural Networks". In: *arXiv:1402.1869 [cs, stat]* (Feb. 2014).

[NH] Nair, V. and Hinton, G. E. "Rectified Linear Units Improve Restricted Boltzmann Machines". In: (), p. 8.

[Nog+19] Nogueira, K. et al. "An Introduction to Deep Morphological Networks". In: *arXiv:1906.01751 [cs]* (2019).

[PS04] Pachter, L. and Sturmfels, B. "Tropical geometry of statistical models". In: *Proceedings of the National Academy of Sciences* 101.46 (Nov. 2004), pp. 16132–16137.

[Par14] Parikh, N. "Proximal Algorithms". In: *Foundations and Trends® in Optimization* 1.3 (2014), pp. 127–239.

[PM00] Pessoa, L. F. and Maragos, P. "Neural networks with hybrid morphological/rank/linear nodes: a unifying framework with applications to handwritten character recognition". In: *Pattern Recognition* 33.6 (2000), pp. 945–960.

[Pin98] Pin, J.-E. *Tropical semirings*. 1998.

[Rip07] Ripley, B. D. *Pattern recognition and neural networks*. Cambridge university press, 2007.

[RS96] Ritter, G. X. and Sussner, P. "An introduction to morphological neural networks". In: *Proceedings of 13th International Conference on Pattern Recognition*. Vol. 4. IEEE. 1996, pp. 709–717.

[RSD98] Ritter, G. X., Sussner, P., and Diza-de-Leon, J. "Morphological associative memories". In: *IEEE Transactions on neural networks* 9.2 (1998), pp. 281–293.

[RU03] Ritter, G. X. and Urcid, G. "Lattice algebra approach to single-neuron computation". In: *IEEE Transactions on Neural Networks* 14.2 (2003), pp. 282–295.

[Ros58] Rosenblatt, F. "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6 (1958), p. 386.

[Rud17] Ruder, S. "An overview of gradient descent optimization algorithms". In: *arXiv:1609.04747 [cs]* (2017).

[Sha+11] Shalev-Shwartz, S. et al. "Online learning and online convex optimization". In: *Foundations and trends in Machine Learning* 4.2 (2011), pp. 107–194.

[She+16] Shen, X. et al. "Disciplined convex-concave programming". In: *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, Dec. 2016, pp. 1009–1014. ISBN: 978-1-5090-1837-6.

[SZS19] Shen, Y., Zhong, X., and Shih, F. Y. "Deep Morphological Neural Networks". In: *arXiv preprint arXiv:1909.01532* (2019).

[Sil98] Sill, J. "Monotonic networks". In: *Advances in neural information processing systems*. 1998, pp. 661–667.

[SM20] Smyrnis, G. and Maragos, P. "Multiclass Neural Network Minimization via Tropical Newton Polytope Approximation". In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. PMLR, July 2020.

[SMR20] Smyrnis, G., Maragos, P., and Retsinas, G. "Maxpolynomial Division with Application To Neural Network Simplification". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 4192–4196.

[Sus98] Sussner, P. "Morphological perceptron learning". In: *Proceedings of the 1998 IEEE International Symposium on Intelligent Control (ISIC) held jointly with IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA) Intell.* IEEE. 1998, pp. 477–482.

[SE11] Sussner, P. and Esmi, E. L. "Morphological perceptrons with competitive learning: Lattice-theoretical framework and constructive learning algorithm". In: *Information Sciences* 181.10 (2011), pp. 1929–1950.

[SB98] Sutton, R. S. and Barto, A. G. *Reinforcement learning: an introduction*. Adaptive computation and machine learning. MIT Press, 1998. ISBN: 978-0-262-19398-6.

[TAM90] Tarela, J., Alonso, E., and Martinez, M. "A representation method for PWL functions oriented to parallel processing". In: *Mathematical and Computer Modelling* 13.10 (1990), pp. 75–83.

[TM99] Tarela, J. and Martinez, M. "Region configurations for realizability of lattice piecewise-linear models". In: *Mathematical and Computer Modelling* 30.11-12 (1999), pp. 17–27.

[The15] Theodoridis, S. *Machine learning: a Bayesian and optimization perspective*. Elsevier, AP, 2015. ISBN: 978-0-12-801522-3.

[TM18a] Theodosis, E. and Maragos, P. "An Adaptive Pruning Algorithm for Spoofing Localisation Based on Tropical Geometry". In: *arXiv:1811.01017 [cs, stat]* (Nov. 2018).

[TM18b] Theodosis, E. and Maragos, P. "Analysis of the Viterbi Algorithm Using Tropical Algebra and Geometry". In: *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, June 2018, pp. 1–5. ISBN: 978-1-5386-3512-4.

[TM18c] Theodosis, E. and Maragos, P. "Tropical Modeling of Weighted Transducer Algorithms on Graphs". In: *arXiv:1811.00573 [cs, math]* (Nov. 2018).

[TY15] Tran, N. M. and Yu, J. "Product-Mix Auctions and Tropical Geometry". In: *arXiv:1505.05737 [cs, math, q-fin]* (May 2015).

[TM19] Tsiamis, A. and Maragos, P. "Sparsity in max-plus algebra and systems". In: *Discrete Event Dynamic Systems* (May 2019).

[Val20] Valle, M. E. "Reduced Dilation-Erosion Perceptron for Binary Classification". In: *Mathematics* 8.4 (2020), p. 512.

[VDF06] Velikova, M., Daniels, H., and Feelders, A. "Solving Partially Monotone Problems with Neural Networks". In: 12 (2006), p. 6.

[Vir01] Viro, O. *Dequantization of Real Algebraic Geometry on Logarithmic Paper*. Ed. by C. Casacuberta et al. Birkhäuser Basel, 2001, pp. 135–146. ISBN: 978-3-0348-9497-5.

[Wan04] Wang, S. "General constructive representations for continuous piecewise-linear functions". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 51.9 (2004), pp. 1889–1896.

[WS05] Wang, S. and Sun, X. "Generalization of hinging hyperplanes". In: *IEEE Transactions on Information Theory* 51.12 (2005), pp. 4425–4431.

[WL19] Wehenkel, A. and Louppe, G. "Unconstrained Monotonic Neural Networks". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 1545–1555.

[XRV17] Xiao, H., Rasul, K., and Vollgraf, R. "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms". In: *arXiv preprint arXiv:1708.07747* (2017).

[YM95] Yang, P.-F. and Maragos, P. "Min-max classifiers: Learnability, design and application". In: *Pattern Recognition* 28.6 (1995), pp. 879–899.

[You+17] You, S. et al. “Deep Lattice Networks and Partial Monotonic Functions”. In: *arXiv:1709.06680 [cs, stat]* (2017).

[YR03] Yuille, A. L. and Rangarajan, A. “The Concave-Convex Procedure”. In: *Neural Computation* 15.4 (Apr. 2003), pp. 915–936.

[ZNL18] Zhang, L., Naitzat, G., and Lim, L.-H. “Tropical Geometry of Deep Neural Networks”. In: *arXiv:1805.07091 [cs, math, stat]* (May 2018).

[Zie95] Ziegler, G. M. *Lectures on polytopes*. Graduate texts in mathematics. Springer-Verlag, 1995. ISBN: 978-0-387-94329-9.