ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΗΣ ΑΝΑΛΥΣΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟΥΠΟΛΗ, ΑΘΗΝΑ 15784
ΤΗΛ 210 - 7276397, FAX 210 - 7276398

Μητσάκος Νικόλαος
(MathMits@yahoo.gr)

Εφαρμογές της Συναρτησιακής Ανάλυσης στη Μηχανική Μάθηση

(Μηχανές Διανυσμάτων Υποστήριξης - Ο Adaptive kernel LMS Αλγόριθμος)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
ΓΙΑ ΤΟ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΟΥ ΤΜΗΜΑΤΟΣ ΜΑΘΗΜΑΤΙΚΩΝ
ΤΟΥ
ΕΘΝΙΚΟΥ ΚΑΠΟΔΙΣΤΡΙΑΚΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΑΘΗΝΩΝ
ΣΤΗΝ ΚΑΤΕΥΘΥΝΣΗ
ΕΦΑΡΜΟΣΜΕΝΑ ΜΑΘΗΜΑΤΙΚΑ

ΑΘΗΝΑ 2012

Η παρούσα Διπλωματική Εργασία

εκπονήθηκε στα πλαίσια των σπουδών

για την απόκτηση του

Μεταπτυχιακού Διπλώματος Ειδίκευσης

στα...

...Εφαρμοσμένα Μαθηματικά...

που απονέμει το

Τμήμα Μαθηματικών

του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών

Εγκρίθηκε την ...21/6/2012... από Εξεταστική Επιτροπή αποτελούμενη από τους :

Ονοματεπώνυμο	Βαθμίδα	Υπογραφή
Γ. Ευαγγελάτου-Δάλλα (επιβλέπων Καθηγητής)	Αν. Καθ.	....
Αθανάσιος Τσαρπαλιάς	Καθηγητής	....
Σέργιος Θεοδωρίδης	Καθηγητής	....

Στην Αγγελικούλα,
που τελικά κανείς μας δεν προστάτεψε ...

Περιεχόμενα

Πρόλογος

Στη Συναρτησιακή Ανάλυση, ένας Reproducin Kernel Hilbert Space είναι ένας Hilbert χώρος συναρτήσεων στον οποίο το γραμμικό evaluation συναρτησιακό είναι συνεχές, οπότε οι τιμές του εκφράζονται και ως εσωτερικά γινόμενα μεταξύ κάποιον στοιχείων του χώρου. Το αντικείμενο αναπτύχθηκε αρχικά, και ταυτόχρονα, από τους Nachman Aronszajn (1907–1980) και Stefan Bergman (1895–1977) το 1950. Η συσχέτιση των συγκεκριμένων χώρων με τις θετικά ορισμένες συναρτήσεις οδηγεί σε ένα ευρύ πεδίο εφαρμογών τους, όπως για παράδειγμα στη μιγαδική ανάλυση, στην αρμονική ανάλυση, στην κβαντική μηχανική και τη στατιστική. Στην παρούσα εργασία θα εξετάσουμε εφαρμογές στον κλάδο της τεχνητής νοημοσύνης που αποκαλείται Μηχανική Μάθηση.

Με τον όρο Μηχανική Μάθηση περιγράφεται η επιστημονική αρχή που ασχολείται με το σχεδιασμό και την ανάπτυξη αλγορίθμων που επιτρέπουν στους υπολογιστές να αναπτύσσουν συμπεριφορές βασιζόμενοι σε εμπειρικά δεδομένα. Τα δεδομένα που τροφοδοτούνται στη μηχανή κατά το στάδιο της «εκπαίδευσης», καλούνται και «δεδομένα εκπαίδευσης», θα μπορούσαν για παράδειγμα να έχουν καταγραφεί από κάποιους αισθητήρες ή να προέρχονται από άλλες βάσεις δεδομένων και αποτελούν παρατηρήσεις που αντικατοπτρίζουν τις σχέσεις μεταξύ των μεταβλητών που τα περιγράφουν, παρέχοντας έτσι πληροφορίες για την, άγνωστη σε εμάς, κατανομή πιθανότητας που τα διέπει. Σημαντικό στόχο της έρευνας στο πεδίο της Μηχανικής Μάθησης αποτελεί η διατύπωση μεθόδων αυτόματης αναγνώρισης πολύπλοκων προτύπων από μια «εκπαιδευόμενη» μηχανή, καθιστώντας την έτσι ικανή να εξάγει «έξυπνες», βάση των δεδομένων που της παρέχονται, αποφάσεις. Η μεγάλη δυσκολία έγκειται, βεβαίως, στο γεγονός πως βασιζόμενη σε έναν, περιορισμένο συνήθως, όγκο δεδομένων εκπαίδευσης η μηχανή πρέπει να μπορεί να γενικεύει τη διεξαγωγή χρήσιμων συμπερασμάτων ώστε να ανταποκρίνονται στο, συνήθως τεράστιο, πλήθος δυνατών συμπεριφορών που ενδέχεται να εμφανίζουν όλα τα πιθανά δεδομένα.

Έχει καταβληθεί μεγάλη προσπάθεια ώστε το πλήρες κείμενο να γίνεται κατανοητό από αναγνώστες που έχουν ακαδημαϊκού επιπέδου γνώση της μαθηματικής επιστήμης, ταυτόχρονα όμως αναγνώστες που ενδιαφέρονται μόνο για τις εφαρμογές των συμπερασμάτων στην πληροφορική να μπορούν να περιορίσουν τη μελέτη τους στα Κεφάλαια 2 και 3, ανατρέχοντας κατά το ελάχιστο δυνατόν στο περισσότερο

θεωρητικό 1ο Κεφάλαιο.

Η συγκεκριμένη εργασία εκπονήθηκε στα πλαίσια της ολοκλήρωσης του Μεταπτυχιακού Προγράμματος σπουδών στα Εφαρμοσμένα Μαθηματικά του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών. Θα ήταν παράληψη από πλευράς μου να μην ευχαριστήσω την κ. Δάλλα, τον κ. Τσαρπαλιά και τον κ. Θεοδωρίδη τόσο για την εμπιστοσύνη που έδειξαν σε έναν μαθηματικό να διαχειριστεί έννοιες και μεθόδους πέραν των αυστηρών μαθηματικών «συνηθειών» του, όσο και για τις πολύ χρήσιμες συμβουλές και επισημάνσεις τους. Ειδικά οφείλω να ευχαριστήσω τον Παντελή Μπουμπούλη για την υπομονή και το χρόνο που μου διέθεσε αλλά και για το υλικό που ευγενικά μου παραχώρησε. Τέλος, ευχαριστώ τους καλούς φίλους Βασίλη, Γιώργο και Ευαγγελία, χάρης εις τους οποίους η, ούτως ή άλλως, ενδιαφέρουσα περίοδος των μεταπτυχιακών μου σπουδών μετατράπηκε σε μια απολαυστική διαδρομή.

Κεφάλαιο 1

Reproducing Kernel Hilbert Spaces (RKHS)

Στο Κεφάλαιο αυτό αρχικά θα περιγράψουμε την ειδική κατηγορία χώρων Hilbert που αποκαλούνται RKHS, αναφέροντας, πέρα από τους αυστηρούς μαθηματικούς ορισμούς, και αρκετά παραδείγματα. Στη συνέχεια θα αποδείξουμε χρήσιμα συμπεράσματα της θεωρίας που έχει αναπτυχθεί, όσον αφορά στους συγκεκριμένους χώρους, καταλήγοντας στο Θεώρημα Moore. Το αποτέλεσμα αυτό θα μας εξασφαλίσει το μαθηματικό υπόβαθρο για τη διατύπωση του kernel Τεχνάσματος, που θα αποτελέσει το βασικό εργαλείο στις εφαρμογές των Κεφαλαίων 2 και 3. Μιας και θα αναφερθούμε σε χώρους Hilbert είτε επί του σώματος των πραγματικών αριθμών, $\mathbb{R}$, είτε επί εκείνου των μιγαδικών αριθμών, $\mathbb{C}$, θα χρησιμοποιούμε το σύμβολο $\mathbb{F}$ κάθε φορά που θα διατυπώνουμε έναν ορισμό ή κάποιο συμπέρασμα που ισχύει και στα δύο αυτά σύνολα.

1.1 Ορισμοί και Βασικά Παραδείγματα

Ορισμός 1.1 Δεδομένων δύο διανυσματικών χώρων με νόρμα V και W, επί ενός σώματος $\mathbb{F}$, μια γραμμική απεικόνιση $T : V \longrightarrow W$ καλείται **φραγμένος τελεστής** ανν υπάρχει πραγματικός αριθμός $M \geq 0$ τ.ω.

$$\|Tv\|_W \leq M\|v\|_V \qquad \text{για κάθε } v \in V$$

Αποδεικνύεται ότι ένας τελεστής είναι φραγμένος αν και μόνο αν είναι συνεχής, με τη συνήθη έννοια. Μάλιστα, ακόμα ισχυρότερα, ένας τελεστής είναι φραγμένος αν και μόνο αν είναι συνεχής στο 0. Επίσης οι εικόνες φραγμένων συνόλων, μέσω συνεχών (ή ισοδύναμα φραγμένων) τελεστών, είναι επίσης φραγμένα σύνολα.

Ως **νόρμα** ενός **φραγμένου τελεστή** ορίζουμε :

$$\|T\| = inf\{M : \|Tv\| \leq M\|v\| \ \text{ για κάθε } v \in V\}$$

ή ισοδύναμα:

$$\begin{aligned}\|T\| &= sup\{\|Tv\| : v \in V \ \text{ με } \|v\| \leq 1\} \\ &= sup\{\|Tv\| : v \in V \ \text{ με } \|v\| = 1\} \\ &= sup\{\tfrac{\|Tv\|}{\|v\|} : v \in V \ \text{ με } v \neq 0\}\end{aligned}$$

Υπενθυμίζουμε ότι, το σύνολο όλων των συναρτήσεων που ορίζονται σε ένα δεδομένο σύνολο X και δίνουν τιμές στο $\mathbb{F}$, συμβ. $\mathcal{F}(X, \mathbb{F})$, όταν εφοδιαστεί με τις συνήθεις πράξεις της πρόσθεσης, $(f+g)(x) = f(x) + g(x)$, και του βαθμωτού γινομένου, $(\lambda \cdot f)(x) = \lambda \cdot (f(x))$ για κάθε $\lambda \in \mathbb{F}$, μετατρέπεται σε **διανυσματικό χώρο** επί του σώματος $\mathbb{F}$.

Ορισμός 1.2 Δεδομένου ενός συνόλου X, θα λέμε ότι ο $\mathcal{H}$ είναι ένας **reproducing kernel Hilbert space (RKHS)** του X επί του $\mathbb{F}$, όταν ισχύουν τα εξής:

(α) ο $\mathcal{H}$ είναι διανυσματικός υπόχωρος του $\mathcal{F}(X, \mathbb{F})$,

(β) ο $\mathcal{H}$ είναι εφοδιασμένος με ένα εσωτερικό γινόμενο, $\langle \cdot, \cdot \rangle$, που τον μετατρέπει σε χώρο Hilbert, δηλαδή σε μετρικό χώρο πλήρη ως προς την απόσταση που ορίζει η επαγόμενη από το εσωτερικό γινόμενο νόρμα, $\| \cdot \|$,

(γ) για κάθε $y \in X$, το γραμμικό **evaluation συναρτησιακό**, $E_y : \mathcal{H} \longrightarrow \mathbb{F}$, που ορίζεται από τη σχέση $E_y(f) = f(y)$, είναι φραγμένο.

Οι γραμμικοί, φραγμένοι τελεστές σε χώρους Hilbert χαρακτηρίζονται από την ακόλουθη, σπουδαία, ιδιότητα:

Θεώρημα 1.1 (Αναπαράστασης του Riesz) Έστω H χώρος Hilbert επί σώματος $\mathbb{F}$ και $T : H \longrightarrow \mathbb{F}$ ένας φραγμένος, γραμμικός, τελεστής. Τότε υπάρχει μοναδικό στοιχείο $h_0 \in H$ τέτοιο ώστε $Th = \langle h_0, h \rangle_H$ για κάθε $h \in H$.

Έστω λοιπόν $\mathcal{H}$ ένας RKHS συνόλου X. Το παραπάνω Θεώρημα μας εξασφαλίζει ότι, για κάθε $y \in X$ θα υπάρχει ένα μοναδικό διάνυσμα, $k_y \in \mathcal{H}$, τέτοιο ώστε για κάθε $f \in \mathcal{H}$ να ισχύει $E_y(f) = f(y) = \langle f, k_y \rangle$. Έτσι,

Ορισμός 1.3 Έστω $\mathcal{H}$ ένας RKHS συνόλου X. Για κάθε $y \in X$, η συνάρτηση $k_y \in \mathcal{H}$ με την ιδιότητα

$$E_y(f) = f(y) = \langle f, k_y \rangle, \quad \text{για κάθε } f \in \mathcal{H}$$

καλείται η **reproducing kernel (συνάρτηση) για το σημείο** y, ενώ η συνάρτηση δύο μεταβλητών που ορίζεται ως

$$K(x, y) = k_y(x) \quad \text{για κάθε } x, y \in X$$

καλείται η **reproducing kernel συνάρτηση του** $\mathcal{H}$.

Παρατηρείστε ότι ισχύει:

$$K(x,y) = k_y(x) = \langle k_y, k_x \rangle$$

καθώς και

$$\|E_y\|^2 = \|k_y\|^2 = \langle k_y, k_y \rangle = K(y,y).$$

Η προτελευταία σχέση διαδραματίζει σημαντικό ρόλο στις εφαρμογές της συγκεκριμένης θεωρίας σε πραγματικά προβλήματα, καθώς επιτρέπει την αντικατάσταση του δύσκολου υπολογισμού ενός εσωτερικού γινομένου με τον εύκολο υπολογισμό της τιμής μιας συνάρτησης δύο μεταβλητών. Έχουμε όμως ακόμα αρκετό δρόμο έως ότου εξασφαλίσουμε, στο τέλος του Κεφαλαίου, την ύπαρξη κατάλληλων τέτοιων χώρων και συναρτήσεων για ένα οποιοδήποτε σύνολο X.

Πόρισμα 1.1 Αν ο $\mathcal{H}$ είναι RKHS του X με reproducing kernel την $K(x,y)$, τότε ισχύει

$$K(y,x) = \overline{K(x,y)}.$$

όπου με $\overline{K(x,y)}$ συμβολίζουμε το συζυγή του $K(x,y)$. Προφανώς, όσων αφορά στις πραγματικές kernel συναρτήσεις, ισχύει $K(y,x) = K(x,y)$.

Ορισμός 1.4 Έστω $\mathcal{H}$ ένας RKHS συνόλου X. Θα λέμε ότι ο $\mathcal{H}$ **διαχωρίζει σημεία** ανν για κάθε $x, y \in X$ με $x \neq y$ υπάρχει $f \in \mathcal{H}$ τέτοια ώστε $f(x) \neq f(y)$.

Αποδεικνύεται ότι, αν στον RKHS $\mathcal{H}$ θεωρήσουμε την απεικόνιση $d(x,y) = sup\{|f(x) - f(y)| : f \in \mathcal{H}, \|f\| \leq 1\}$, αυτή ορίζει μια μετρική του X ανν ο $\mathcal{H}$ διαχωρίζει σημεία.

Επίσης αποδεικνύεται ότι, αν $\mathcal{H}_0 \subseteq \mathcal{H}$ είναι κλειστός υπόχωρος τότε ο $\mathcal{H}_0$ είναι επίσης ένας RKHS του X, ενώ η reproducing kernel συνάρτηση, στον $\mathcal{H}_0$, ενός σημείου y είναι η συνάρτηση $P_0(k_y)$, όπου k_y είναι η reproducing kernel συνάρτηση για το ίδιο σημείο στον $\mathcal{H}$ και με $P_0 : \mathcal{H} \longrightarrow \mathcal{H}_0$ συμβολίζουμε την ορθογώνια προβολή του $\mathcal{H}$ πάνω στον $\mathcal{H}_0$. Έτσι, θα υπάρχει μια reproducing kernel συνάρτηση, $K_0(x,y)$, για τον υπόχωρο $\mathcal{H}_0$.

Ένα από τα προβλήματα που εμφανίζουν ιδιαίτερο θεωρητικό ενδιαφέρον είναι ο καθορισμός σχέσεων ανάμεσα στην $K_0(x,y)$ του κλειστού υποχώρου και στη reproducing kernel συνάρτηση, $K(x,y)$, ολόκληρου του χώρου. Παρά το γεγονός ότι είναι σχετικά απλή η διατύπωση ορισμένων γενικών θεωρημάτων που να αφορούν στη σχέση αυτή, ο υπολογισμός συγκεκριμένων παραδειγμάτων είναι αρκετά περίπλοκος.

Στο σημείο αυτό θα ήταν χρήσιμο να παρουσιάσουμε ορισμένα βασικά παραδείγματα, ώστε να γίνει περισσότερο κατανοητή η φύση των συγκεκριμένων χώρων καθώς και η λογική προσδιορισμού των reproducing kernel συναρτήσεών τους:

Ο χώρος Hardy, $H^2(\mathbb{D})$, του Μοναδιαίου Δίσκου.

Έστω $\mathbb{D}$ ο μοναδιαίος δίσκος του μιγαδικού επιπέδου. Για την κατασκευή του χώρου $H^2(\mathbb{D})$, ο οποίος διαδραματίζει σημαντικό ρόλο στη Θεωρία Τελεστών, θεωρούμε τις τυπικές *(formal)* μιγαδικές δυναμοσειρές, $f \backsim \sum_{n=0}^{\infty} a_n z^n$, $g \backsim \sum_{n=0}^{\infty} b_n z^n$, όπου $z \in \mathbb{D}$, τ.ω. $\sum_{j=0}^{\infty} |a_n|^2 < \infty$ και $\sum_{j=0}^{\infty} |b_n|^2 < \infty$, τις οποίες εφοδιάζουμε με το εσωτερικό γινόμενο $\langle f, g \rangle = \sum_{n=0}^{\infty} a_n \overline{b_n}$, οπότε θα ισχύει και $\|f\|^2 = \sum_{n=0}^{\infty} |a_n|^2$. Μπορούμε να δείξουμε ότι κάθε δυναμοσειρά του $H^2(\mathbb{D})$ συγκλίνει σε μια συνάρτηση στο δίσκο. Πράγματι, αν $z \in \mathbb{D}$, τότε:

$$|E_z(f)| = |\sum_{n=0}^{\infty} a_n z^n| \leq \sum_{n=0}^{\infty} |a_n||z|^n \leq \left(\sum_{n=0}^{\infty} |a_n|^2\right)^{\frac{1}{2}} \left(\sum_{n=0}^{\infty} |z|^{2n}\right)^{\frac{1}{2}} = \|f\| \frac{1}{\sqrt{1-|z|^2}}$$

Έτσι, κάθε δυναμοσειρά ορίζει μια συνάρτηση στο $\mathbb{D}$, ενώ οι πράξεις του διανυσματικού χώρου των τυπικών δυναμοσειρών αντιστοιχίζονται ευθέως στις πράξεις μεταξύ των συναρτήσεων του διανυσματικού χώρου $\mathbb{D}$, δηλαδή ικανοποιείται η συνθήκη (α) του ορισμού (1.2).

Η απεικόνιση $L : H^2(\mathbb{D}) \longrightarrow \ell^2(\mathbb{N})$ τώρα, με $L(f) = (a_0, a_1, \ldots)$, είναι γραμμικός ισομορφισμός που διατηρεί το εσωτερικό γινόμενο. Συνεπώς ο $H^2(\mathbb{D})$ μπορεί να ταυτιστεί με τον Hilbert χώρο $\ell^2(\mathbb{Z}^+)$, όπου $\mathbb{Z}^+ = \{0, 1, 2, \ldots\}$ το σύνολο των φυσικών αριθμών $\mathbb{N}$ συμπεριλαμβανομένου και του 0, άρα είναι και ο ίδιος ένας Hilbert χώρος. Δηλαδή ικανοποιείται η συνθήκη (β) του ορισμού (1.2).

Η συνθήκη (γ) του ορισμού ικανοποιείται, επίσης, χάρης εις την παραπάνω ανισότητα, η οποία εξασφαλίζει ότι η απεικόνιση E_z είναι φραγμένη με $\|E_z\| \leq \frac{1}{\sqrt{1-|z|^2}}$. Συνεπώς ο $H^2(\mathbb{D})$ είναι ένας RKHS του $\mathbb{D}$.

Για τον προσδιορισμό της kernel συνάρτησης ενός σημείου $w \in \mathbb{D}$, παρατηρούμε ότι $g(z) = \sum_{n=0}^{\infty} \overline{w}^n z^n \in H^2(\mathbb{D})$ ενώ παράλληλα για κάθε $f(z) = \sum_{n=0}^{\infty} a_n z^n \in H^2(\mathbb{D})$ έπεται $\langle f, g \rangle = \sum_{n=0}^{\infty} a_n w^n = f(w)$. Οπότε η g είναι η reproducing kernel του w. Συνεπώς

$$K(z,w) = k_w(z) = g(z) = \sum_{n=0}^{\infty} \overline{w}^n z^n = \frac{1}{1-\overline{w}z}$$

Η τελευταία καλείται η **Szego Kernel συνάρτηση** του δίσκου.

Τέλος, υπολογίζοντας $\|E_z\| = (K(z,z))^{1/2} = \dfrac{1}{\sqrt{1-|z|^2}}$ διαπιστώνουμε ότι η ανισότητα που είχαμε υπολογίσει νωρίτερα είναι η βέλτιστη δυνατή.

Χώροι Sobolev στο $[0,1]$

Για την κατασκευή των συγκεκριμένων χώρων, οι οποίοι αποτελούν παράδειγμα χώρων που προκύπτουν από διαφορικές εξισώσεις, θεωρούμε τον χώρο $\mathcal{H} = \{f : [0,1] \longrightarrow \mathbb{R} : f$ απόλυτα συνεχής, $f(0) = f(1) = 0, f' \in L^2[0,1]\}$. (Υπενθυμίζουμε ότι όταν μια συνάρτηση είναι απόλυτα συνεχής είναι διαφορίσιμη σχεδόν παντού και ισούται με το ολοκλήρωμα της παραγώγου της). Προφανώς, ο $\mathcal{H}$ είναι διανυσματικός χώρος συναρτήσεων στο $[0,1]$ τον οποίο εφοδιάζουμε με τη μη-αρνητική διγραμμική μορφή $\langle f, g \rangle = \int_0^1 f'(t)g'(t)dt$. Αφού η f ειναι απόλυτα συνεχής και $f(0) = 0$, για κάθε $0 \leq x \leq 1$ έχουμε

$$f(x) = \int_0^x f'(t)dt = \int_0^1 f'(t)\chi_{[0,x]}(t)dt.$$

Έτσι, από την ανισότητα των Cauchy-Schwartz, έχουμε :

$$|f(x)| \leq \left(\int_0^1 f'(t)^2 dt\right)^{\frac{1}{2}} \left(\int_0^1 \chi_{[0,x]}(t)dt\right)^{\frac{1}{2}} = (\langle f, f \rangle)^{1/2}\sqrt{x}$$

από την οποία συμπεραίνουμε ότι $\langle f, f \rangle = 0$ ανν $f = 0$. Οπότε η $\langle \cdot, \cdot \rangle$ αποτελεί εσωτερικό γινόμενο στον $\mathcal{H}$, ενώ $\forall x \in [0,1]$ η E_x είναι φραγμένη και μάλιστα $\|E_x\| \leq \sqrt{x}$.

Για να δείξουμε ότι ο $\mathcal{H}$ είναι RKHS αρκεί να δείξουμε ότι είναι πλήρης. Έστω (f_n) μια Cauchy, ως προς την επαγόμενη από το εσωτερικό γινόμενο νόρμα, ακολουθία.

Τότε και η (f_n') είναι Cauchy στον $L^2[0,1]$, συνεπώς θα συγκλίνει σε κάποια $g \in L^2[0,1]$. Λόγω της παραπάνω ανισότητας η (f_n) πρέπει να είναι κατά σημείο Cauchy, οπότε μπορούμε να ορίσουμε συνάρτηση $f(x) = \lim_n f_n(x)$. Έτσι έχουμε

$$f(x) = \lim_n f_n(x) = \lim_n \int_0^x f_n'(t)dt = \int_0^x g(t)dt$$

που σημαίνει ότι η f είναι απόλυτα συνεχής και $f' = g$ σχεδόν παντού, άρα $f' \in L^2[0,1]$. Επιπλέον,

$$f(0) = \lim_n f_n(0) = 0 = \lim_n f_n(1) = f(1)$$

οπότε, τελικά, $f \in \mathcal{H}$. Συνεπώς, ο $\mathcal{H}$ είναι RKHS του $[0,1]$.

Απομένει να ανακαλύψουμε την kernel συνάρτηση. Ξεκινάμε με την τυπική επίλυση μιας διαφορικής εξίσωσης και κατόπιν αποδεικνύουμε ότι η συνάρτηση που λαμβάνουμε ως λύση ανήκει στον $\mathcal{H}$. Προκειμένου να βρούμε την $k_y(t)$ εφαρμόζουμε ολοκλήρωση κατά παράγοντες και έχουμε

$$f(y) = \langle f, k_y \rangle = \int_o^1 f'(t)k_y'(t)dt = f(t)k_y'(t) - \int_0^1 f(t)k_y''(t)dt = -\int_0^1 f(t)k_y''(t)dt$$

Συμβολίζοντας με δ_y την τυπική Dirac-delta συνάρτηση, έχουμε

$$f(y) = \int_0^1 f(t)\delta_y(t)dt$$

Οπότε το προς επίλυση πρόβλημα συνοριακών τιμών είναι:

$$\begin{gathered} -k_y''(t) = \delta_y(t) \\ k_y(0) = k_y(1) = 0 \end{gathered}$$

Για να βρούμε τη λύση του παραπάνω συστήματος εξισώσεων, η οποία αποκαλείται συνάρτηση Green της διαφορικής εξίσωσης, ολοκληρώνουμε δύο φορές και ελέγχουμε τις συνοριακές συνθήκες, οπότε καταλήγουμε στη συνάρτηση

$$K(x,y) = k_y(x) = \begin{cases} (1-y)x, & x \leq y \\ (1-x)y, & x \geq y \end{cases}$$

Η συνάρτηση αυτή ικανοποιεί πράγματι τις απαραίτητες εξισώσεις ώστε να αποτελεί τη reproducing kernel συνάρτηση του $\mathcal{H}$.

Παρατηρείστε επιπλέον ότι

$$\|E_y\|^2 = \|k_y\|^2 = K(y,y) = y(1-y)$$

η οποία αποτελεί μια καλύτερη εκτίμηση από την $\|E_y\| \leq \sqrt{y}$ που είχαμε επιτύχει νωρίτερα.

Bergman χώροι σε Περιοχές του Μιγαδικού Επιπέδου

Έστω $G \subset \mathbb{C}$ ανοικτό και συνεκτικό σύνολο. Θεωρούμε

$$B^2(G) = \{f : G \longleftrightarrow \mathbb{C} \mid \eta\ f \text{ αναλυτική στο } G \text{ και } \int\int_G |f(x+iy)|^2 dxdy < +\infty\}$$

Επιπλέον, ορίζουμε μια sesquilinear μορφή στο $B^2(G)$ ως εξής

$$\langle f, g\rangle = \int\int_G f(x+iy)\overline{g(x+iy)}dxdy$$

η οποία, όπως αποδεικνύεται εύκολα, αποτελεί εσωτερικό γινόμενο στον $B^2(G)$, δηλαδή ο $B^2(G)$ μετατρέπεται σε χώρο εσωτερικού γινομένου.

Θεώρημα 1.2 Έστω $G \subseteq \mathbb{C}$ ανοικτό και συνεκτικό σύνολο. Τότε ο $B^2(G)$ είναι RKHS στο G.

Απόδειξη. Σταθεροποιώντας ένα $w \in G$ επιλέγουμε $R > 0$ τέτοιο ώστε η κλειστή μπάλα $\overline{B(w,R)}$, κέντρου w και ακτίνας R, να περιέχεται στο G. Από τον τύπο ολοκλήρωσης του Cauchy , για κάθε $0 \leq r \leq R$ έχουμε

$$f(w) = \frac{1}{2\pi}\int_0^{2\pi} f(w + re^{i\theta})d\theta$$

Συνεπώς,

$$f(w) = \frac{1}{\pi R^2}\int_0^R r(2\pi f(w))dr = \frac{1}{\pi R^2}\int_0^R r\int_0^{2\pi} f(w+re^{i\theta})d\theta = \frac{1}{\pi R^2}\int\int_{B(w,R)} f(x+iy)dxdy$$

Επιπλέον, από την ανισότητα Cauchy-Schwartz , έχουμε

$$|f(w)| \leq \frac{1}{\pi R^2}\|f\|_2 \left(\int\int_{B(w,R)} dxdy\right)^{\frac{1}{2}} = \frac{1}{R\sqrt{\pi}}\|f\|_2$$

που μας εξασφαλίζει ότι, για το $w \in G$, το evaluation συναρτησιακό είναι φραγμένο.

Απομένει να αποδείξουμε ότι ο $B^2(G)$ είναι πλήρης ως προς τη συγκεκριμένη νόρμα. Έστω $(f_n) \subseteq B^2(G)$ μια Cauchy ακολουθία. Για κάθε $w \in G$ επιλέγουμε ένα $R > 0$, όπως παραπάνω, και ένα $\delta > 0$ τέτοιο ώστε $0 < \delta < d(B(w, R), G^C)$, όπου με $d(\cdot, \cdot)$ συμβολίζουμε την απόσταση μεταξύ δύο συνόλων. Έτσι, για οποιοδήποτε z εντός της κλειστής μπάλας ακτίνας R και κέντρου w, η κλειστή μπάλα κέντρου z και ακτίνας δ περιέχεται στο G. Συνεπώς, σύμφωνα με την παραπάνω εκτίμηση, έχουμε

$$|f_n(z) - f_m(z)| \leq \frac{1}{\delta\sqrt{\pi}} \|f_n - f_m\|_2$$

Με άλλα λόγια, η ακολουθία συναρτήσεων είναι ομοιόμορφα συγκλίνουσα σε κάθε κλειστή μπάλα που περιέχεται στο G. Συμβολίζοντας με $f(z) = \lim_n f_n(z)$ το όριο της κατά σημείο σύγκλισης της ακολουθίας, έχουμε την (f_n) να συγκλίνει ομοιόμορφα στην f σε κάθε κλειστή μπάλα που περιέχεται στο G, οπότε, από το Θεώρημα του Montel , καταλήγουμε στο συμπέρασμα ότι η f είναι αναλυτική. Τώρα, αφού $B^2(G) \subseteq L^2(G)$ και ο $L^2(G)$ είναι πλήρης, θα υπάρχει $h \in L^2(G)$ τέτοιο ώστε $\|h - f_n\|_2 \longrightarrow 0$. Επιπλέον, η πληρότητα του $L^2(G)$ μας εξασφαλίζει και την ύπαρξη υπακολουθίας (f_{n_k}) της (f_n) τέτοιας ώστε $h(z) = \lim_k f_{n_k}(z)$ σχεδόν παντού, κάτι όμως που σημαίνει $h(z) = f(z)$ σχεδόν παντού, οπότε $\|f - f_n\|_2 \longrightarrow 0$. Καταλήξαμε λοιπόν στο συμπέρασμα ότι η $f \in B^2(G)$, δηλαδή ότι ο $B^2(G)$ είναι πλήρης. ∎

Ορισμός 1.5 Δεδομένου ενός ανοικτού και συνεκτικού υποσυνόλου $G \subseteq \mathbb{C}$ η reproducing kernel συνάρτηση του $B^2(G)$ καλείται **Bergman kernel συνάρτηση** του G.

Το αποτέλεσμα που αποδείξαμε επεκτείνεται και σε ανοικτά συνεκτικά υποσύνολα του $\mathbb{C}^n$ καθώς και σε πολλές μιγαδικές πολλαπλότητες, ενώ η γνώση της Bergman kernel σε τέτοιες περιοχές βρίσκει πολλές εφαρμογές στη μιγαδική ανάλυση.

Επίσης, η ανισότητα $|f| \leq \frac{1}{R\sqrt{\pi}} \|f\|$ υποδεικνύει ότι $B^2(\mathbb{C}) = \{0\}$, καθώς στην περίπτωση αυτή το R μπορεί να γίνει απείρως μεγάλο οπότε $|f| = 0$ για κάθε $f \in B^2(\mathbb{C})$. Συνεπώς, η μόνη αναλυτική συνάρτηση, με $\int\int_{\mathbb{C}} |f(x+iy)|^2 dxdy < +\infty$, που ορίζεται σε ολόκληρο το μιγαδικό επίπεδο είναι η μηδενική.

Έστω, τώρα, ότι το εμβαδόν του G είναι ίσο με $A < +\infty$, οπότε η σταθερή συνάρτηση 1 περιέχεται στον $B^2(G)$ και $\|1\| = \sqrt{A}$. Σε μια τέτοια περίπτωση, είναι λογικό να κανονικοποιήσουμε ούτως ώστε $\|1\| = 1$, κάτι που επιτυγχάνεται εύκολα ορίζοντας εκ νέου το εσωτερικό γινόμενο ως

$$\langle f, g \rangle = \frac{1}{A} \int\int_G f(x+iy)\overline{g(x+iy)} dxdy$$

Στη βιβλιογραφία συχνά η χρήση της ορολογίας «Bergman χώρος» αναφέρεται σε αυτόν, τον κανονικοποιημένο χώρο, μια σύμβαση που θα την υιοθετήσουμε κι εμείς στο παρόν κείμενο. Συγκεκριμένα λοιπόν, ως $B^2(\mathbb{D})$ θα συμβολίζουμε το χώρο των τετραγωνικά ολοκληρώσιμων, αναλυτικών συναρτήσεων στο μοναδιαίο δίσκο $\mathbb{D}$, με εσωτερικό γινόμενο

$$\langle f, g\rangle = \frac{1}{\pi}\int\int_{\mathbb{D}} f(x+iy)\overline{g(x+iy)}dxdy$$

Αποδεικνύεται ότι η **Bergman kernel συνάρτηση** δίνεται από τη σχέση

$$K(z,w) = \frac{1}{(1-z\overline{w})^2}$$

Σταθμισμένοι Χώροι Hardy

Δεδομένης μιας ακολουθίας $\beta = (\beta_n)$, με $\beta_n > 0$, θεωρούμε το χώρο όλων των τυπικών δυναμοσειρών $f(z) = \sum_{n=0}^{\infty} a_n z^n$, με $z \in \mathbb{D}$, για τις οποίες η νόρμα

$$\|f\|_\beta^2 = \sum_{n=0}^{\infty} \beta_n^2 |a_n|^2$$

είναι πεπερασμένη. Ο χώρος αυτός, εφοδιασμένος με το εσωτερικό γινόμενο

$$\langle f, g\rangle = \sum_{n=0}^{\infty} \beta_n^2 a_n b_n$$

όπου $g(z) = \sum_{n=0}^{\infty} b_n z^n$, αποτελεί Hilbert χώρο, ο οποίος συμβολίζεται με H_β^2 και καλείται **σταθμισμένος χώρος Hardy**.

Ένας συνήθης χώρος Hardy , όπως αυτός που αναφέραμε σε προηγούμενο παράδειγμα, θα μπορούσε, προφανώς, να χαρακτηρισθεί ως σταθμισμένος χώρος Hardy όπου για όλα τα βάρη ισχύει $\beta_n = 1$.

Κάθε δυναμοσειρά $f(z) = \sum_{n=0}^{\infty} a_n z^n$ του H^2_β ικανοποιεί τη σχέση

$$\lim_n \beta_n |a_n| = 0$$

Οπότε, για αρκετά μεγάλο n, έχουμε

$$|a_n| \leq \beta_n^{-1}$$

Συνεπώς, η ακτίνα σύγκλισης, R_f, της $\sum_0^{\infty} a_n z^n = f(z)$ ικανοποιεί τη σχέση

$$R_f^{-1} = \limsup_n |a_n|^{1/n} \leq \limsup_n \beta_n^{-1/n} = \left(\liminf_n \beta_n^{1/n}\right)^{-1}$$

Οπότε, η $\sum_0^{\infty} a_n z^n$ θα έχει ακτίνα σύγκλισης μεγαλύτερη, ή ίση, από

$$\liminf_{n\to\infty} (\beta_n)^{-1/n} \equiv R$$

Έτσι, υποθέτοντας ότι $R > 0$, κάθε συνάρτηση του H^2_β θα συγκλίνει, ορίζοντας μια αναλυτική συνάρτηση στο δίσκο ακτίνας R, οπότε ο H^2_β μπορεί να θεωρηθεί ως χώρος αναλυτικών συναρτήσεων στο συγκεκριμένο δίσκο.

Εύκολα διαπιστώνει κανείς ότι, για κάθε $|w| < R$, ισχύει $f(w) = \langle f, k_w \rangle$, όπου η συνάρτηση

$$k_w(z) = \sum_{n=0}^{\infty} \frac{\overline{w}^n z^n}{\beta_n^2}$$

ανήκει στον H^2_β. Έτσι, δεδομένου του περιορισμού για την ακολουθία β ώστε $R > 0$, ο H^2_β είναι ένας RKHS του δίσκου ακτίνας R, με reproducing kernel συνάρτηση την

$$K_\beta(z, w) = k_w(z)$$

Ένας ακόμα σταθμισμένος χώρος Hardy που έχει μελετηθεί εκτενώς είναι ο **Segal-Bargmann χώρος**, ο οποίος προκύπτει θεωρώντας τα βάρη $\beta_n = \sqrt{n!}$. Επειδή $\liminf_{n\to\infty} (n!)^{-1/(2n)} = +\infty$, ο συγκεκριμένος χώρος αποτελεί χώρο ακεραίων συναρτήσεων (entire functions), δηλαδή μιγαδικών συναρτήσεων που είναι ολομορφικές επάνω σε ολόκληρο το μιγαδικό επίπεδο. Η reproducing kernel συνάρτηση

προκύπτει άμεσα πως είναι η $K(z,w) = k_w(z) = \sum_{n=0}^{\infty} \frac{\overline{w}^n z^n}{n!} = e^{z\overline{w}}$.

Παραδείγματα Πολλαπλών Μεταβλητών

Δεδομένου $n \in \mathbb{N}$, με τον όρο πολυ-δείκτης (multi-index) εννοούμε ένα σημείο $I = (i_1, ..., i_n) \in (\mathbb{Z}^+)^n$, οπότε για κάθε $z = (z_1, ..., z_n) \in \mathbb{C}^n$ θεωρούμε $z^I = z_1^{i_1}, ..., z_n^{i_n}$.

Ως **δυναμοσειρά στις n μεταβλητές** εννοούμε μια τυπική έκφραση της μορφής $f(z) = \sum_{I \in (\mathbb{Z}^+)^n} a_I z^I$, όπου οι $a_I \in \mathbb{C}$ καλούνται οι **συντελεστές της f**.

Ορίζουμε το **χώρο Hardy n-μεταβλητών**, $H^2(\mathbb{D}^n)$, ως το σύνολο όλων των δυναμοσειρών $f \sim \sum_{I \in (\mathbb{Z}^+)^n} a_I z^I$ τέτοιων ώστε $\|f\|^2 = \sum_{I \in (\mathbb{Z}^+)^n} |a_i|^2 < +\infty$.
Η λογική του ορισμού είναι η δυναμοσειρά να συγκλίνει για κάθε $z \in \mathbb{D}^n$, ορίζοντας μια αναλυτική συνάρτηση, $f(z)$, όπως και στο Hardy χώρο μιας μεταβλητής, και ο $H^2(\mathbb{D}^n)$ να είναι RKHS του $\mathbb{D}^n$ με reproducing kernel τη συνάρτηση

$$K(z,w) = \sum_{I \in (\mathbb{Z}^+)^n} \overline{w}^I z^I = \prod_{i=1}^{n} \frac{1}{1 - \overline{w}_i z_i}$$

Με παρόμοιο τρόπο ορίζεται ο **χώρος Bergman πολλαπλών μεταβλητών**, $B^2(G)$, για $G \subset \mathbb{C}^n$ ανοικτό και συνεκτικό υποσύνολο, χρησιμοποιώντας $2n$-διάστασης μέτρο Lebesgue.

Επίσης, όπως και στην περίπτωση της μιας μεταβλητής, όταν το μέτρο Lebesgue του G είναι πεπερασμένο μπορούμε να χρησιμοποιήσουμε κανονικοποιημένο μέτρο Lebesgue, ορίζοντας τη νόρμα στον $B^2(G)$ κατά τρόπο ώστε η σταθερή συνάρτηση 1 να έχει νόρμα 1.

1.2 Μιγαδοποίηση ενός RKHS πραγματικών συναρτήσεων

Ας θεωρήσουμε τον $\mathcal{H}$ ως έναν RKHS πραγματικών συναρτήσεων ενός συνόλου X με reproducing kernel τη συνάρτηση $K(x,y)$. Ας θεωρήσουμε επίσης το διανυσματικό

χώρο μιγαδικών συναρτήσεων του X, $W = \{f_1 + if_2 : f_1, f_2 \in \mathcal{H}\}$. Εφοδιασμένος με τη σχέση

$$\langle f_1 + if_2, g_1 + ig_2\rangle_W = \langle f_1, g_1\rangle_{\mathcal{H}} + i\langle f_2, g_1\rangle_{\mathcal{H}} - i\langle f_1, g_2\rangle_{\mathcal{H}} + \langle f_2, g_2\rangle_{\mathcal{H}}$$

η οποία, όπως εύκολα διαπιστώνει κανείς, ορίζει εσωτερικό γινόμενο στο W με αντίστοιχη νόρμα

$$\|f_1 + if_2\|_W^2 = \|f_1\|_{\mathcal{H}}^2 + \|f_2\|_{\mathcal{H}}^2$$

ο W μετατρέπεται σε χώρο Hilbert. Δεδομένου επιπλέον ότι

$$f_1(y) + if_2(y) = \langle f_1 + if_2, k_y\rangle_W$$

καταλήγουμε ότι ο W, εφοδιασμένος με το παραπάνω εσωτερικό γινόμενο, αποτελεί RKHS μιγαδικών συναρτήσεων του X, με reproducing kernel τη συνάρτηση $K(x, y)$. Τον χώρο αυτό θα τον αποκαλούμε **μιγαδοποίηση του $\mathcal{H}$**.

Αφού λοιπόν κάθε πραγματικός RKHS μπορεί να μιγαδοποιηθεί κατά τρόπο τέτοιο ώστε να διατηρεί τη reproducing kernel συνάρτησή του, στη συνέχεια θα θεωρούμε τους RKHS στους οποίους αναφερόμαστε ως μιγαδικούς.

1.3 Η Γενική Θεωρία των RKHS

Έστω σύνολο X και $\mathcal{H}$ ένας RKHS του X, με reproducing kernel τη συνάρτηση K. Θα διαπιστώσουμε ότι η K καθορίζει πλήρως το χώρο $\mathcal{H}$. Για να καταλήξουμε στο πρώτο μας συμπέρασμα θα χρειαστούμε ένα σημαντικό αποτέλεσμα από τη θεωρία τελεστών. Υπενθυμίζουμε λοιπόν τα εξής:

Ορισμός 1.6 Δεδομένου ενός χώρου Hilbert H και ενός υποσυνόλου $S \subset H$, θεωρούμε το σύνολο των **ορθογώνιων (ή κάθετων) προς το S στοιχείων**:

$$S^{\perp} = \{x \in H : \langle x, s\rangle = 0, \forall s \in S\}$$

και το αποκαλούμε **ορθογώνιο σύνολο** του S.

Ορισμός 1.7 Έστω χώρος Hilbert H και M κλειστός υπόχωρος του H. Ο υπόχωρος M καλείται **συμπληρωματικός** στον H αν υπάρχει κλειστός υπόχωρος N του H τέτοιος ώστε:

(α) $H = M + N = \{\mu + \nu : \mu \in M, \nu \in N\}$ και

(β) $M \cap N = (0)$

Τότε ο H λέμε ότι αποτελεί το **ευθύ άθροισμα** των M και N, και γράφουμε $H = M \oplus N$.

Θεώρημα 1.3 (Προβολής) Έστω M κλειστός υπόχωρος ενός χώρου Hilbert H. Τότε

$$H = M \oplus M^{\perp} \tag{1.1}$$

δηλαδή $H = M + M^{\perp}$ και $M \cap M^{\perp} = (0)$.

Στηριζόμενοι λοιπόν σε αυτό το, ούτως ή άλλως πολύ χρήσιμο, Θεώρημα διατυπώνουμε και αποδεικνύουμε την ακόλουθη:

Πρόταση 1.1 Έστω $\mathcal{H}$ ένας RKHS συνόλου X με kernel K. Τότε, η γραμμική θήκη $\mathcal{H}_1$ των συναρτήσεων $k_y(\cdot) = K(\cdot, y)$, δηλαδή ο διανυσματικός υπόχωρος του $\mathcal{H}$ που απαρτίζεται από το σύνολο όλων των γραμμικών συνδυασμών μεταξύ των συναρτήσεων $k_y(\cdot)$, είναι πυκνός υπόχωρος του $\mathcal{H}$.

Απόδειξη. Μια συνάρτηση $f \in \mathcal{H}$ είναι ορθογώνια προς τη γραμμική θήκη $\mathcal{H}_1$ των συναρτήσεων $\{k_y : y \in X\}$ ανν $\langle f, k_y \rangle = f(y) = 0$ για κάθε $y \in X$, κάτι που ισχύει ανν $f = 0$. Δηλαδή η μηδενική συνάρτηση είναι το μοναδικό στοιχείο του $\mathcal{H}$ κάθετο προς τον υπόχωρο $\mathcal{H}_1$. Όμως το εσωτερικό γινόμενο είναι συνεχής απεικόνιση, έτσι η μηδενική συνάρτηση θα είναι το μοναδικό στοιχείο του $\mathcal{H}$ κάθετο και προς την κλειστή θήκη $\overline{\mathcal{H}_1}$, η οποία βεβαίως αποτελεί κλειστό διανυσματικό υπόχωρο του $\mathcal{H}$. Σύμφωνα λοιπόν με το Θεώρημα Προβολής θα ισχύει $\mathcal{H} = \overline{\mathcal{H}_1} \oplus \overline{\mathcal{H}_1}^{\perp}$ δηλαδή $\mathcal{H} = \overline{\mathcal{H}_1}$.
■

Λήμμα 1.1 Έστω $\mathcal{H}$ ένας RKHS συνόλου X και έστω $f_n \subseteq \mathcal{H}$. Αν $\lim_n \|f_n - f\| = 0$, τότε ισχύει $f(x) = \lim_n f_n(x)$ για κάθε $x \in X$.

Απόδειξη. Έχουμε

$$|f_n(x) - f(x)| = |\langle f_n - f, k_x \rangle| \leq \|f_n - f\| \|k_x\| \longrightarrow 0.$$

■

Πρόταση 1.2 Έστω $\mathcal{H}_i,\ i = 1, 2$ δύο RKHS συνόλου X με kernels τις συναρτήσεις $K_i(x, y),\ i = 1, 2$ αντίστοιχα. Αν $K_1(x, y) = K_2(x, y)$ για κάθε $x, y \in X$, τότε $\mathcal{H}_1 = \mathcal{H}_2$ και $\|f\|_1 = \|f\|_2$ για κάθε f.

Απόδειξη. Έστω $K(x, y) = K_1(x, y) = K_2(x, y)$ και $W_i = span\{k_x \in \mathcal{H}_i : x \in X\},\ i = 1, 2$. Λόγω της παραπάνω Πρότασης, κάθε W_i είναι πυκνός στον $\mathcal{H}_i,\ i = 1, 2$, ενώ για κάθε $f \in W_i$ ισχύει $f(x) = \sum_j a_j k_{x_j}(x)$, οπότε οι τιμές των

συναρτήσεων είναι ανεξάρτητες από το αν τις θεωρούμε στο W_1 ή στο W_2. Επιπλέον, ισχύει

$$\|f\|_1^2 = \sum_{i,j} a_i \overline{a_j} \langle k_{x_i}, k_{x_j} \rangle = \sum_{i,j} a_i \overline{a_j} K(x_j, x_i) = \|f\|_2^2 \iff$$
$$\|f\|_1 = \|f\|_2 \quad \text{για κάθε } f \in W_1 = W_2.$$

Τώρα, αν $f \in \mathcal{H}_1$, υπάρχει ακολουθία συναρτήσεων $\{f_n\} \subseteq W_1$ με $\|f - f_n\|_1 \longrightarrow 0$. Αφού η $\{f_n\}$ είναι Cauchy στον W_1 θα είναι Cauchy και στον W_2, δηλαδή θα υπάρχει $g \in \mathcal{H}_2$ τέτοια ώστε $\|g - f_n\| \longrightarrow 0$. Από το Λήμμα όμως, έχουμε $f(x) = \lim_n f_n(x) = g(x)$, οπότε κάθε $f \in \mathcal{H}_1$ βρίσκεται και στον $\mathcal{H}_2$, ενώ όμοια κάθε $g \in \mathcal{H}_2$ βρίσκεται και στον $\mathcal{H}_1$, με άλλα λόγια $\mathcal{H}_1 = \mathcal{H}_2$.

Τέλος, αφού $\|f\|_1 = \|f\|_2$ για κάθε f που ανήκει σε πυκνό υποσύνολο, έπεται ότι οι νόρμες είναι ισοδύναμες για κάθε f. ∎

Πριν προχωρήσουμε σε μια άλλη συνέπεια του παραπάνω Λήμματος, η οποία θα μας παρέχει έναν εναλλακτικό, πολύ χρήσιμο, τρόπο υπολογισμού της kernel συνάρτησης ενός RKHS, ας θυμηθούμε μια, διαφορετική από τη συνήθη, μορφή σύγκλισης.

Δεδομένων διανυσμάτων $\{h_s : s \in S\}$ σε έναν χώρο με νόρμα $\mathcal{H}$, όπου S είναι ένα αυθαίρετο σύνολο, λέμε ότι $h = \sum_{s \in S} h_s$ όταν για οποιοδήποτε $\epsilon > 0$ υπάρχει πεπερασμένο υποσύνολο $F_0 \subseteq S$ τέτοιο ώστε για κάθε πεπερασμένο σύνολο F, με $F_0 \subseteq F \subseteq S$, να ισχύει $\|h - \sum_{s \in F} h_s\| < \epsilon$.

Παραδείγματα τέτοιου είδους σύγκλισης προκύπτουν από τις δύο ταυτότητες Parseval. Αν $\{e_s : s \in S\}$ είναι μια ορθοκανονική βάση ενός χώρου Hilbert $\mathcal{H}$, τότε για κάθε $h \in \mathcal{H}$ έχουμε

$$\|h\|^2 = \sum_{s \in S} |\langle h, e_s \rangle|^2$$

και

$$h = \sum_{s \in S} \langle h, e_s \rangle e_s$$

Παρατηρείστε ότι τα παραπάνω αθροίσματα δεν απαιτούν το σύνολο S να είναι διατεταγμένο. Χαρακτηριστικά, αν θεωρήσουμε $a_n = \frac{(-1)^n}{n}$, $n \in \mathbb{N}$, η σειρά $\sum_{n=1}^{\infty} a_n$ συγκλίνει, ενώ η $\sum_{n \in \mathbb{N}} a_n$ δε συγκλίνει με την έννοια που ορίσαμε.

Στην πραγματικότητα, για μιγαδικούς αριθμούς αποδεικνύεται ότι η $\sum_{n \in \mathbb{N}} z_n$ συγκλίνει ανν η $\sum_{n=1}^{\infty} |z_n|$ συγκλίνει. Οπότε στην περίπτωση των μιγαδικών, αυτή η σύγ-

κλιση είναι ισοδύναμη με την απόλυτη σύγκλιση.

Θεώρημα 1.4 Έστω $\mathcal{H}$ ένας RKHS συνόλου X με reproducing kernel τη συνάρτηση $K(x,y)$. Αν $\{e_s : s \in S\}$ είναι μια ορθοκανονική βάση του $\mathcal{H}$, τότε $K(x,y) = \sum_{s\in S} \overline{e_s(y)} e_s(x)$ όπου η σειρά συγκλίνει κατά σημείο.

Απόδειξη. Για κάθε $y \in X$, έχουμε $\langle k_y, e_s\rangle = \overline{\langle e_s, k_y\rangle} = \overline{e_s(y)}$. Έτσι, $k_y = \sum_{s\in S} \overline{e_s(y)} e_s$ όπου τα αθροίσματα αυτά συγκλίνουν ως προς τη νόρμα του $\mathcal{H}$.
Αφού όμως συγκλίνουν ως προς τη νόρμα συγκλίνουν ως προς κάθε σημείο. Οπότε $K(x,y) = k_y(x) = \sum_{s\in S} \overline{e_s(y)} e_s(x)$. ∎

Για παράδειγμα, στο χώρο Hardy, οι συναρτήσεις $e_n(z) = z^n$, $n \in \mathbb{Z}^+$, αποτελούν ορθοκανονική βάση, συνεπώς η reproducing kernel συνάρτηση για το χώρο προκύπτει και ως εξής:

$$\sum_{n\in\mathbb{Z}^+} e_n(z)\overline{e_n(w)} = \sum_{n=0}^{\infty} (z\overline{w})^n = \frac{1}{1-z\overline{w}}$$

Αξίζει να σημειωθεί ότι, η προϋπόθεση το σύνολο που χρησιμοποιούμε στο άθροισμα να είναι βάση δεν είναι αναγκαία. Τέτοια σύνολα όμως έχουν πολύ κομψό και χρήσιμο χαρακτηρισμό.

Ορισμός 1.8 Έστω $\mathcal{H}$ ένας χώρος Hilbert με εσωτερικό γινόμενο $\langle\cdot,\cdot\rangle$. Ένα σύνολο διανυσμάτων $\{f_s : s \in S\} \subseteq \mathcal{H}$ καλείται **Parseval πλαίσιο** αν ισχύει

$$\|h\|^2 = \sum_{s\in S} |\langle h, f_s\rangle|^2$$

για κάθε $h \in \mathcal{H}$.

Για παράδειγμα, αν $\{u_s : s \in S\}$ και $\{v_t : t \in T\}$ είναι δύο ορθοκανονικές βάσεις του $\mathcal{H}$, τότε τα σύνολα $\{u_s : s \in S\} \cup \{0\}$ και $\{u_s/\sqrt{2} : s \in S\} \cup \{v_t/\sqrt{2} : t \in T\}$ είναι και τα δύο Parseval πλαίσια του $\mathcal{H}$.

Τα Parseval πλαίσια δεν είναι απαραίτητο να είναι γραμμικά ανεξάρτητα σύνολα. Ένας από τους συνηθέστερους τρόπους δημιουργίας Parseval πλαισίων εμφανίζεται στην ακόλουθη Πρόταση.

Πρόταση 1.3 Έστω $\mathcal{H}$ ένας χώρος Hilbert, $\mathcal{H}_0 \subseteq \mathcal{H}$ ένας κλειστός υπόχωρος και ας συμβολίσουμε με P_0 την ορθογώνια προβολή του $\mathcal{H}$ πάνω στο $\mathcal{H}_0$. Αν $\{e_s : s \in S\}$

είναι μια ορθοκανονική βάση του $\mathcal{H}$, τότε το $\{P_0(e_s) : s \in S\}$ αποτελεί ένα Parseval πλαίσιο για το $\mathcal{H}_0$.

Απόδειξη. Έστω $h \in \mathcal{H}_0$. Τότε $h = P_0(h)$, οπότε $\langle h, e_s\rangle = \langle P_0(h), e_s\rangle = \langle h, P_0(e_s)\rangle$. Συνεπώς $\|h\|^2 = \sum_{s\in S} |\langle h, P_0(e_s)\rangle|^2$. ∎

Οποιαδήποτε από τις ταυτότητες Parseval θα μπορούσε να χρησιμοποιηθεί για να ορίσουμε πλαίσια Parseval, όπως φαίνεται στην Πρόταση που ακολουθεί

Πρόταση 1.4 Έστω $\mathcal{H}$ ένας Hilbert χώρος και $\{f_s : s \in S\} \subseteq \mathcal{H}$. Το $\{f_s : s \in S\}$ είναι πλαίσιο Parseval ανν $h = \sum_{s\in S} \langle h, f_s\rangle f_s$ για κάθε $h \in \mathcal{H}$. Επιπροσθέτως, αν το $\{f_s : s \in S\}$ είναι πλαίσιο Parseval τότε για κάθε $h_1, h_2 \in \mathcal{H}$ έπεται $\langle h_1, h_2\rangle = \sum_{s\in S} \langle h_1, f_s\rangle\langle f_s, h_2\rangle$.

Απόδειξη. (βλ. [5] σελ. 11-12) ∎

Πρόταση 1.5 (Larson) Έστω $\{f_s : s \in S\}$ ένα πλαίσιο Parseval ενός χώρου Hilbert $\mathcal{H}$. Τότε υπάρχει χώρος Hilbert, $\mathcal{K}$, ο οποίος να περιέχει τον $\mathcal{H}$ ως υπόχωρο, καθώς και μια ορθοκανονική βάση $\{e_s : s \in S\}$ του $\mathcal{K}$, έτσι ώστε $f_s = P_{\mathcal{H}}(e_s)$, $s \in S$, όπου με $P_{\mathcal{H}}$ συμβολίζουμε την ορθογώνια προβολή του $\mathcal{K}$ πάνω στον $\mathcal{H}$.

Απόδειξη. (βλ. [5] σελ. 12) ∎

Θεώρημα 1.5 (Papadakis) Έστω $\mathcal{H}$ ένας RKHS συνόλου X με reproducing kernel τη συνάρτηση $K(x, y)$. Τότε το $\{f_s : s \in S\} \subseteq \mathcal{H}$ είναι ένα πλαίσιο Parseval ανν $K(x, y) = \sum_{s\in S} f_s(x)\overline{f_s(y)}$, όπου η σειρά συγκλίνει κατά σημείο.

Απόδειξη. (βλ. [5] σελ. 12-13) ∎

1.4 Χαρακτηρισμός των Reproducing Kernels

Στην παράγραφο αυτή θα εξετάσουμε ικανές και αναγκαίες συνθήκες ώστε μια συνάρτηση $K(x, y)$ να αποτελεί reproducing kernel συνάρτηση για κάποιον RKHS. Αρχικά, ας υπενθυμίσουμε ορισμένες έννοιες από τη θεωρία πινάκων.

Ορισμός 1.9 Έστω $A = (a_{i,j})$ ένας $n \times n$ ερμητιανός[1] πίνακας. Ο A καλείται **θετικός (ή θετικά ημιορισμένος ή μη αρνητικός) πίνακας**, (συμβ. $A \geq$

[1] $A = A^*$, όπου A^* ο ανάστροφος συζυγής του A.

0), ανν για κάθε $a_1, ..., a_n \in \mathbb{C}$ ισχύει $\sum_{i,j=1}^{n} \overline{a_i} a_j a_{ij} \geq 0$.

(Η ποσότητα $\sum_{i,j=1}^{n} \overline{a_i} a_j a_{ij}$ είναι πάντα πραγματικός αριθμός όταν ο A είναι ερμητιανός).

Ισοδύναμα, συμβολίζοντας με $\langle \cdot, \cdot \rangle$ το σύνηθες εσωτερικό γινόμενο, ο $A \geq 0$ αν και μόνο αν $\langle Ax, x \rangle \geq 0$ για κάθε $x = (a_1, ..., a_n) \in \mathbb{C}^n$.

Για έναν ερμητιανό πίνακα A αποδεικνύεται ότι, $A \geq 0$ αν και μόνο αν για κάθε ιδιοτιμή του λ ισχύει $\lambda \geq 0$.

Παρατηρείστε ότι, όσον αφορά σε θετικούς πίνακες $A = (a_{i,j})$, επιτρέπεται η ικανοποίηση της συνθήκης $\sum_{i,j=1}^{n} \overline{a_i} a_j a_{ij} = 0$ από διανύσματα $(a_1, ..., a_n) \neq 0$. Ερμητιανούς $n \times n$ πίνακες $A = (a_{i,j})$ για τους οποίους απαιτούμε να ισχύει $\sum_{i,j=1}^{n} \overline{a_i} a_j a_{ij} > 0$ για κάθε $(a_1, ..., a_n) \neq 0$ τους αποκαλούμε **αυστηρά θετικούς** (συμβ. $A > 0$).

Για έναν ερμητιανό πίνακα A ισχύει $A > 0$ αν και μόνο αν $\lambda > 0$ για κάθε ιδιοτιμή του λ.

Τέλος, αναφορικά στη σχέση των δύο παραπάνω κλάσεων πινάκων, αποδεικνύεται ότι $A > 0$ αν και μόνο αν $A \geq 0$ και A αντιστρέψιμος.

Ορισμός 1.10 Έστω σύνολο X, μια συνάρτηση $K : X \times X \longrightarrow \mathbb{C}$ δύο μεταβλητών και ένα υποσύνολο $\{x_1, x_2, ..., x_n\} \subseteq X$. Ο τετραγωνικός $n \times n$ πίνακας $(K(x_i, x_j))$ με στοιχεία $(K(x_i, x_j))_{i,j} = K(x_i, x_j)$ για $i, j = 1, ..., n$ καλείται **Gram πίνακας (ή kernel πίνακας) της συνάρτησης** K ως προς τα $\{x_1, ..., x_n\}$.

Ορισμός 1.11 Έστω σύνολο X και $K : X \times X \longrightarrow \mathbb{C}$ μια συνάρτηση δύο μεταβλητών. Η K καλείται **kernel (ή θετικά ορισμένη) συνάρτηση** (συμβ. $K \geq 0$) αν για κάθε $n \in \mathbb{N}$ και για κάθε επιλογή n διακριτών στοιχείων $\{x_1, ..., x_n\} \subseteq X$, ο Gram πίνακας της K ως προς τα $\{x_1, ..., x_n\}$ είναι θετικός.

Αποδεικνύεται ότι τα αθροίσματα kernel συναρτήσεων είναι kernel συναρτήσεις, ενώ αν $K : X \times X \longrightarrow \mathbb{C}$ είναι μια kernel συνάρτηση και $f : X \longrightarrow \mathbb{C}$ είναι μια τυχαία συνάρτηση τότε η συνάρτηση $K_0(x, y) = f(x) \cdot K(x, y) \cdot \overline{f(y)}$ είναι επίσης μια kernel συνάρτηση.

Με όλα τα παραπάνω υπ' όψιν, διατυπώνουμε την ακόλουθη:

Πρόταση 1.6 Έστω σύνολο X και $\mathcal{H}$ ένας RKHS του X με reproducing kernel τη συνάρτηση K. Τότε η K είναι kernel συνάρτηση.

Απόδειξη. Σταθεροποιώντας δυο επιλογές στοιχείων $\{x_1, ..., x_n\} \subseteq X$ και $a_1, ..., a_n \in \mathbb{C}$ έχουμε

$$\sum_{i,j}^{n} \overline{a_i} a_j K(x_i, x_j) = \langle \sum_{j}^{n} a_j k_{x_j}, \sum_{i}^{n} a_i k_{x_i} \rangle = \| \sum_{j}^{n} a_j k_{x_j} \|^2 \geq 0 \qquad (1.2)$$

απ' όπου και έπεται το ζητούμενο. ∎

Σημειώνουμε εδώ ότι, στη γενική περίπτωση, για τον Gram πίνακα μιας reproducing kernel συνάρτησης ισχύει $(K(x_i, x_j)) > 0$. Όταν κάτι τέτοιο δεν ισχύει, οι παραπάνω υπολογισμοί αποκαλύπτουν την ύπαρξη ενός μη-μηδενικού διανύσματος $a = (a_1, ..., a_n)$ τέτοιου ώστε $\| \sum_{j}^{n} a_j k_{x_j} \| = 0 \iff \sum_{j}^{n} a_j k_{x_j} = 0$. Συνεπώς, για κάθε $f \in \mathcal{H}$, έχουμε $\sum_{j}^{n} \overline{a_j} f(x_j) = \langle f, \sum_{j}^{n} a_j k_{x_j} \rangle = 0$. Δηλαδή σε μια τέτοια περίπτωση υπάρχει σχέση γραμμικής εξάρτησης ανάμεσα στις τιμές όλων των συναρτήσεων του $\mathcal{H}$ για κάποιο πεπερασμένο σύνολο σημείων.

Ένα τέτοιο παράδειγμα αποτελούν οι χώροι Sobolev στο $[0, 1]$, στους οποίους, όπως είδαμε νωρίτερα, χρησιμοποιήσαμε χώρους με συνοριακές συνθήκες, της μορφής $f(0) = f(1)$, κάτι που σημαίνει $k_1(t) = k_0(t)$.

Εναλλακτικά, πολλοί χώροι αναλυτικών συναρτήσεων, όπως οι χώροι Hardy και Bergman που έχουμε ήδη αναφέρει, περιέχουν όλα τα πολυώνυμα. Καθώς, όμως, δεν υπάρχει εξίσωση της μορφής $\sum_{j} \beta_j p(x_j) = 0$ η οποία να ικανοποιείται από όλα τα πολυώνυμα, οι reproducing kernel συναρτήσεις τέτοιων χώρων ορίζουν πάντα αυστηρά θετικούς, άρα και αντιστρέψιμους, πίνακες.

Έτσι, αν θεωρήσουμε ως παράδειγμα τη Szego kernel συνάρτηση του χώρου Hardy , για τον πίνακα $(K(x_i, x_j)) = \left(\frac{1}{1-\overline{x_i} x_j}\right)$ προκύπτει άμεσα το συμπέρασμα ότι είναι αντιστρέψιμος για οποιαδήποτε επιλογή $\{x_1, ..., x_n\}$ σημείων του δίσκου. Το συμπέρασμα αυτό δεν είναι καθόλου εύκολο να προκύψει μέσω συνηθισμένων μεθόδων Γραμμικής Άλγεβρας, γεγονός που αποτελεί ένδειξη των δυνατοτήτων που έχει η θεωρία των RKHS.

Και ενώ η τελευταία Πρόταση 1.6 είναι αρκετά στοιχειώδης, το αντίστροφο συμπέρασμα είναι πολύ σημαντικό και χαρακτηρίζει τις reproducing kernel συναρτήσεις.

Θεώρημα 1.6 (Moore) Έστω σύνολο X και μια συνάρτηση $K : X \times X \longrightarrow \mathbb{C}$. Αν η K είναι kernel συνάρτηση τότε υπάρχει ένας RKHS, $\mathcal{H}$, συναρτήσεων ορισμένων στο X, έτσι ώστε η K να αποτελεί τη reproducing kernel συνάρτηση του $\mathcal{H}$.

Απόδειξη. Για κάθε $y \in X$ θέτουμε $k_y(x) = K(x,y)$ και θεωρούμε το χώρο $W \subseteq \mathcal{F}$ που παράγεται από το σύνολο των συναρτήσεων $\{k_y : y \in X\}$.

Ισχυρισμός: Υπάρχει μια καλά ορισμένη, sesquilinear απεικόνιση, $B : W \times W \longrightarrow \mathbb{C}$, με $B\left(\sum_j a_j k_{y_j}, \sum_i b_i k_{y_i}\right) = \sum_{i,j} a_j \overline{b_i} K(y_i, y_j)$, όπου a_j και b_i βαθμωτά μεγέθη.

Απόδειξη Ισχυρισμού: Για να αποδείξουμε ότι η B είναι καλά ορισμένη στο W πρέπει να αποδείξουμε ότι για την ταυτοτικά μηδενική συνάρτηση $f(x) = \sum_j a_j k_{y_j}(x)$ ισχύει $B(f,w) = B(w,f) = 0$ για κάθε $w \in W$. Καθώς όμως ο W παράγεται από τις συναρτήσεις k_y, αρκεί να αποδείξουμε ότι $B(f,k_y) = B(k_y,f) = 0$. Εξ' ορισμού όμως $B(f,k_y) = \sum_j a_j K(y,y_j) = f(y) = 0$. Όμοια επίσης $B(k_y,f) = \sum_j \overline{a_j} K(y_j,y) = \sum_j \overline{a_j}\overline{K(y,y_j)} = \overline{f(y)} = 0$.

Αντίστροφα τώρα, αν $B(f,w) = 0$ για κάθε $w \in W$, τότε θεωρώντας $w = k_y$ έχουμε $f(y) = 0$. Οπότε $B(f,w) = 0$ για κάθε $w \in W$ αν και μόνο αν η f είναι ταυτοτικά μηδενική συνάρτηση στο X.

Έτσι, η B είναι καλά ορισμένη, ενώ επαληθεύεται εύκολα ότι είναι και sesquilinear. Επιπλέον, για κάθε $f \in W$ έχουμε $f(x) = B(f,k_x)$.

Τώρα, αφού η K είναι θετικά ορισμένη, για κάθε $f = \sum_j a_j k_{y_j}$ έχουμε $B(f,f) = \sum_{i,j} a_j \overline{a_i} K(y_i,y_j) \geq 0$, ενώ προκύπτει επιπλέον, με όμοιο τρόπο όπως στης απόδειξη της ανισότητας των Cauchy-Schwartz, ότι $B(f,f) = 0$ αν και μόνο αν $B(w,f) = B(f,w) = 0$ για κάθε $w \in W$. Συνεπώς, $B(f,f) = 0$ αν και μόνο αν η f είναι η ταυτοτικά μηδενική συνάρτηση.

Η B λοιπόν αποτελεί εσωτερικό γινόμενο στον W.

Τώρα, όπως για κάθε δεδομένο εσωτερικό γινόμενο σε διανυσματικό χώρο, έτσι και εδώ μπορούμε να επιτύχουμε πλήρωση του χώρου, λαμβάνοντας ισοδύναμες κλάσεις Cauchy ακολουθιών του W σχηματίζοντας ένα χώρο Hilbert, $\mathcal{H}$.

Πρέπει να αποδείξουμε ότι κάθε στοιχείο του $\mathcal{H}$ είναι πράγματι συνάρτηση ορισμένη στο X (σε αντίθεση με την περίπτωση της πλήρωσης των συνεχών συναρτήσεων του $[0,1]$ ας πούμε, όπου λαμβάνουμε τον $L^2[0,1]$). Για το σκοπό αυτό, ας θεωρήσουμε $h \in H$ και $(f_n) \subseteq W$ μια ακολουθία Cauchy που συγκλίνει στην h. Από την ανισότητα των Cauchy-Schwartz έχουμε $|f_n(x) - f_m(x)| = |B(f_n - f_m, k_x)| \leq \|f_n - f_m\|\sqrt{K(x,x)}$. Η ακολουθία λοιπόν είναι κατά σημείο Cauchy, οπότε μπορούμε να ορίσουμε $h(x) = \lim_n f_n(x)$, όπου φυσικά κάθε τιμή είναι ανεξάρτητη της ακολουθίας Cauchy που επιλέξαμε.

Τέλος, συμβολίζοντας με $\langle \cdot, \cdot \rangle$ το εσωτερικό γινόμενο στον $\mathcal{H}$, για την παραπάνω h έχουμε $\langle h, k_y \rangle = \lim_n \{f_n, k_y\} = \lim_n B(f_n, k_y) = \lim_n f_n(y) = h(y)$. Οπότε ο $\mathcal{H}$ είναι πράγματι RKHS του X, και αφού η k_y είναι η reproducing kernel συνάρτηση για το σημείο y, έχουμε ότι η $K(x, y) = k_y(x)$ είναι η reproducing kernel συνάρτηση του $\mathcal{H}$. ∎

Το Θεώρημα του Moore, σε συνδυασμό με την Πρόταση 1.2, μας εξασφαλίζουν την ένα προς ένα αντιστοιχία ανάμεσα στους RKHS ενός συνόλου και στις kernel (θετικά ορισμένες) συναρτήσεις που ορίζονται στο σύνολο, επιχείρημα που αποτελεί μια πολύ ισχυρή ιδιότητα της θεωρίας των Kernel. Στην ιδιότητα αυτή βασίζεται και το τέχνασμα που περιγράφουμε στην επόμενη παράγραφο, το οποίο θα χρησιμοποιηθεί στις εφαρμογές των Κεφαλαίων 2 και 3.

Ορισμός 1.12 Δεδομένης μιας kernel συνάρτησης $K : X \times X \longrightarrow \mathbb{C}$, συμβολίζουμε με $\mathcal{H}(K)$ τον μοναδικό RKHS που έχει ως reproducing kernel συνάρτηση την K.

Δεδομένης μιας kernel συνάρτησης $K : X \times X \longrightarrow \mathbb{C}$, αποδεικνύεται ότι για δύο σημεία $x_1, x_2 \in X$ με $x_1 \neq x_2$ ο 2×2 πίνακας $(K(x_i, x_j))$ είναι αυστηρά θετικός ανν οι συναρτήσεις k_{x_1} και k_{x_2} είναι γραμμικά ανεξάρτητες, κάτι που συμβαίνει αν και μόνο αν ο $\mathcal{H}(K)$ διαχωρίζει σημεία.

Η διαδικασία κατασκευής του χώρου $\mathcal{H}(K)$ με αφετηρία μια δεδομένη kernel συνάρτηση ονομάζεται **Πρόβλημα Ανοικοδόμησης (Reconstruction Problem)** και αποτελεί μια από τις δυσκολότερες προκλήσεις της θεωρίας των RKHS. Για παράδειγμα, ας υποθέσουμε ότι ξεκινούμε με την Szego kernel συνάρτηση στο δίσκου, $K(z, w) = \frac{1}{(1-\overline{w}z)}$, οπότε ο χώρος W που λαμβάνουμε σύμφωνα με την απόδειξη του Θεωρήματος Moore αποτελείται από τους γραμμικούς συνδυασμούς των συναρτήσεων $k_w(z)$ οι οποίες είναι ακέραιες (rational) συναρτήσεις με έναν μοναδικό πόλο τάξεως ένα εκτός του δίσκου. Έτσι ο χώρος W δεν περιέχει πολυώνυμα, παρόλο που στο χώρο $\mathcal{H}(K) = H^2(\mathbb{D})$ τα πολυώνυμα αποτελούν πυκνό υποσύνολο.

Τουλάχιστον για την αναλυτική περίπτωση, υπάρχουν στη βιβλιογραφία θεωρήματα που επιτρέπουν να καθορίσουμε πότε στο χώρο $\mathcal{H}(K)$ περιέχονται πολυώνυμα.

Κλείνοντας αυτή την παράγραφο, ας δούμε μια απλή εφαρμογή του Θεωρήματος Moore.

Πρόταση 1.7 Έστω σύνολο X, f μια μη μηδενική συνάρτηση ορισμένη στο X και ας θεωρήσουμε και τη συνάρτηση $K(x, y) = f(x)\overline{f(y)}$. Τότε, η K είναι θετικά ορισμένη, ο $\mathcal{H}(K)$ είναι ο χώρος που παράγεται από την f και ισχύει $\|f\| = 1$.

Απόδειξη. Προκειμένου να αποδείξουμε ότι η K είναι θετικά ορισμένη, υπολογίζουμε

$$\sum_{i,j} a_i \overline{a_j} K(x_i, x_j) = |\sum_i a_i f(x_i)|^2 \geq 0$$

Για να βρούμε το χώρο $\mathcal{H}(K)$, παρατηρούμε αρχικά ότι για κάθε συνάρτηση k_y ισχύει $k_y = \overline{f(y)}f$, οπότε ο W δεν είναι άλλος από τον μονοδιάστατο χώρο που παράγεται από την f. Κάθε χώρος, όμως, πεπερασμένης διάστασης είναι πλήρης, άρα ο $\mathcal{H}(K)$ δεν είναι άλλος από το χώρο που παράγεται από την f.

Για το τελικό συμπέρασμα της Πρότασης, υπολογίζουμε τη νόρμα της f. Σταθεροποιούμε ένα σημείο y, τέτοιο ώστε $f(y) \neq 0$, και έχουμε

$$|f|^2 \cdot \|f\|^2 = \|\overline{f(y)}f\|^2 = \|k_y\|^2 = \langle k_y, k_y \rangle = K(y,y) = |f(y)|^2$$

απ΄ όπου έπεται $\|f\| = 1$. ∎

1.5 Το Kernel Τέχνασμα

Καθώς στα επόμενα δύο κεφάλαια θα χρησιμοποιήσουμε το εν λόγω τέχνασμα σε εφαρμογές Μηχανικής Μάθησης και Επεξεργασίας Σήματος, στη συγκεκριμένη παράγραφο θα περιορίσουμε την αναφορά μας σε πραγματικές kernel συναρτήσεις.

Συνοψίζοντας όσα είδαμε προηγουμένως, σε κάθε kernel συνάρτηση $K : X \times X \longrightarrow \mathbb{R}$ αντιστοιχεί ένας, και μόνον ένας, RKHS, δηλαδή ένας, και μόνον ένας, διανυσματικός χώρος συναρτήσεων $\mathcal{H}$, υπόχωρος του $\mathcal{F}(X, \mathbb{R})$, ο οποίος είναι εφοδιασμένος, κατά μοναδικό τρόπο, με ένα εσωτερικό γινόμενο που τον καθιστά χώρο Hilbert με reproducing kernel τη συνάρτηση K. Στην πραγματικότητα, η kernel συνάρτηση K παράγει όλο το χώρο $\mathcal{H}$, δηλαδή $\mathcal{H} = \overline{span\{K(x, \cdot) | x \in X\}}$. Μια από τις ισχυρές ιδιότητες της θεωρίας των Kernel, η οποία στους χώρους της Μηχανικής Μάθησης είναι γνωστή με την ονομασία **Kernel Τέχνασμα**, περιγράφεται ως εξής:

«Δεδομένου ενός αλγορίθμου ο οποίος στους υπολογισμούς χρησιμοποιεί εσωτερικά γινόμενα, μπορούμε να κατασκευάσουμε έναν εναλλακτικό αλγόριθμο, αντικαθιστώντας κάθε ένα από τα εσωτερικά γινόμενα με μία θετικά ορισμένη kernel συνάρτηση.»

Το τέχνασμα βασίζεται στην χρήση μιας απεικόνισης $\Phi : X \longrightarrow \mathcal{H} : \Phi(x) = k_x$, η οποία καλείται **χαρακτηριστική απεικόνιση**, και απεικονίζει κάθε στοιχείο του X σε ένα στοιχείο του $\mathcal{H}$ (υπενθυμίζουμε ότι το στοιχείο $k_x \in \mathcal{H}$ είναι η reproducing kernel συνάρτηση για το σημείο x). Επομένως, με βάση την παρατήρησή μας στον

Ορισμό 1.2 και χάρης εις τη συμμετρία των πραγματικών kernel συναρτήσεων, η συγκεκριμένη απεικόνιση έχει την ιδιότητα

$$\langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} = \langle k_x, k_y \rangle_{\mathcal{H}} = K(x, y)$$

Έτσι, μέσω της χαρακτηριστικής απεικόνισης, το Kernel Τέχνασμα επιτυγχάνει να μετατρέψει ένα μη γραμμικό πρόβλημα εντός του συνόλου X σε ένα γραμμικό πρόβλημα εντός του «καλύτερου» χώρου $\mathcal{H}$. Κατόπιν, επιλύουμε το γραμμικό πρόβλημα στο χώρο $\mathcal{H}$, κάτι που συνήθως αποτελεί σχετικά εύκολη εργασία, ενώ η επιστροφή του αποτελέσματος στο χώρο X μας εξασφαλίζει την, τελική, μη γραμμική λύση του αρχικού μας προβλήματος.

Κεφάλαιο 2

Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης - ΜΔΥ (Support Vector Machines - SVM), τις οποίες εισήγαγε αρχικά ο Vapnik και οι συνεργάτες του στην AT& T, καθιερώθηκαν ταχύτατα ως μια αλγοριθμική προσέγγιση στο πρόβλημα της κατηγοριοποίησης (classification) μέσα στο ευρύτερο πλαίσιο που ονομάζεται Επιβλεπόμενη Μηχανική Μάθηση (Supervised Machine Learning). Ένα πλήθος προβλημάτων κατηγοριοποίησης οι λύσεις των οποίων απαιτούσαν νευρωνικά δίκτυα, ή και άλλες, πιο πολύπλοκες, μεθόδους, αποδείχθηκε πως είναι άμεσα επιλύσιμα με ΜΔΥ. Επιπλέον οι ΜΔΥ, εκτός του ότι εφαρμόζονται γενικά ευκολότερα απ΄ ότι τα Νευρωνικά Δίκτυα, έχουν ταυτόχρονα και το σημαντικό πλεονέκτημα να βασίζουν τη λειτουργία τους σε ένα σταθερό μαθηματικό υπόβαθρο, εξασφαλίζοντας έτσι την ύπαρξη βέλτιστης λύσης, σε αντίθεση με ότι συμβαίνει στα νευρωνικά δίκτυα.

2.1 Προβλήματα κατηγοριοποίησης επιβλεπόμενης μάθησης

Γενικά τα προβλήματα κατηγοριοποίησης διαχωρίζονται σε προβλήματα που αφορούν σε κατηγοριοποίηση δεδομένων σε δύο (**δυαδική (binary) κατηγοριοποίηση**) ή περισσότερες κλάσεις (**multiclass κατηγοριοποίηση**). Καθώς πολλές μέθοδοι έχουν αναπτυχθεί ειδικά για τη δυαδική περίπτωση, σε αυτήν θα αναφερθούμε κι εμείς στη συνέχεια. Άλλωστε η multiclass κατηγοριοποίηση συχνά πραγματοποιείται μέσω συνδυασμού πολλών δυαδικών ταξινομητών (binary classifiers).

Στα **προβλήματα κατηγοριοποίησης επιβλεπόμενης μάθησης** λοιπόν, δεδομένης μιας εισροής σημειακών δεδομένων δύο διαφορετικών ειδών, ζητο-

ύμενο αποτελεί η διαμόρφωση μεθόδου αναγνώρισης της κλάσης στην οποία κάθε νέο δεδομένο ανήκει. Για το σκοπό αυτό, διατίθεται αρχικά ένα σύνολο δεδομένων εκπαίδευσης (training data), αποτελούμενο, ας πούμε, από m σημεία της μορφής:

$$(\boldsymbol{x}_i, y_i), \text{ για } i = 1, ..., m \tag{2.1}$$

Τα $\boldsymbol{x}_i$ αποκαλούνται **χαρακτηριστικά διανύσματα**, στις n ας πούμε διαστάσεις ($\boldsymbol{x}_i \in X \subseteq \mathbb{R}^n, \forall i = 1, ..., m$), και περιέχουν τις πληροφορίες που περιγράφουν κάθε σημειακό δεδομένο, ενώ τα αντίστοιχα y_i λαμβάνουν την τιμή ± 1, ανάλογα με το αν το αντίστοιχο σημειακό δεδομένο βρίσκεται εντός της μίας, $(+1)$, ή της άλλης, (-1), εκ των δύο κλάσεων στις οποίες επιθυμούμε να μάθουμε να τα ταξινομούμε.

Στην πράξη, ο στόχος μας είναι να καθορίσουμε, βάση των δεδομένων εκπαίδευσης, έναν **κανόνα απόφασης** (decision rule) με τη μορφή συνάρτησης $f(\boldsymbol{x})$, της οποίας το πρόσημο να προβλέπει την τιμή του y, όχι μόνο για τα δεδομένα εκπαίδευσης αλλά και για νέες τιμές του $\boldsymbol{x}$.

Ανάμεσα στην πληθώρα εφαρμογών των ΜΔΥ σε διαφόρων ειδών χώρους, συναντούμε και εφαρμογές στις οποίες το χαρακτηριστικό διάνυσμα $\boldsymbol{x}$ ανήκει σε συνεκτικό υπόχωρο $X \subseteq \mathbb{R}^n$. Ωστόσο μπορούμε σε τέτοιες περιπτώσεις, σκεπτόμενοι και λίγο δημιουργικά, να αναδιατυπώνουμε το πρόβλημα σύμφωνα με το ακόλουθο πλαίσιο :

το χαρακτηριστικό διάνυσμα ας είναι ένα δυαδικό διάνυσμα που θα κωδικοποιεί την ύπαρξη ή την απουσία διαφόρων «χαρακτηριστικών» (εξ΄ ου και η ονομασία του).

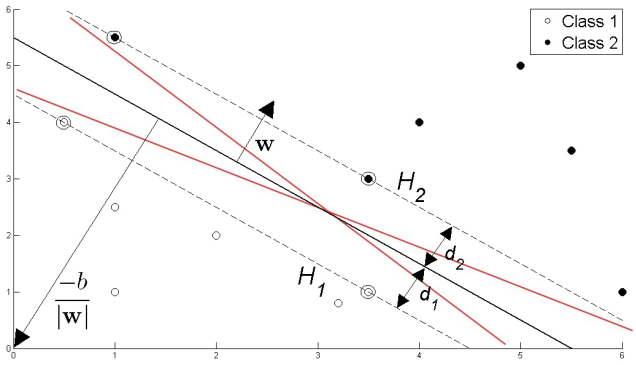
Για παράδειγμα, το χαρακτηριστικό διάνυσμα που περιγράφει μια ακολουθία DNA μήκους p θα μπορούσε να έχει $n = 4p$ διαστάσεις, με κάθε base position να χρησιμοποιεί τέσσερις διαστάσεις, λαμβάνοντας την τιμή ένα σε μία απο τις τέσσερις θέσεις (ανάλογα αν είναι A, C, G ή T), και μηδέν στις υπόλοιπες. Έτσι π.χ. το χαρακτηριστικό διάνυσμα

$$\boldsymbol{x}_0 = (0,0,1,0,1,0,0,0,0,0,0,1,0,0,0,1,1,0,0,0,0,1,0,0,1,0,0,0)$$

διαστάσεως $n = 4 \cdot 7 = 28$ θα αντιστοιχούσε στην ακολουθία GATTACA μήκους 7.

2.2 Η Ειδική Περίπτωση των Γραμμικά Διαχωρίσιμων Δεδομένων

Προκειμένου να γίνει ευκολότερα κατανοητή η έννοια των ΜΔΥ, ας ξεκινήσουμε τη μελέτη τους εστιάζοντας σε μία, μάλλον ουτοπική, εξιδανικευμένη πρώτη περίπτωση, στην οποία ας θεωρήσουμε ότι τα δεδομένα μας είναι **γραμμικά διαχωρίσιμα** *(βλ. Σχήμα 1)*. Υποθέτουμε, δηλαδή, ότι είναι εφικτό να σχεδιάσουμε είτε μια



Σχήμα 2.1: *Οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ) στην εξιδανικευμένη περίπτωση των γραμμικά διαχωρίσιμων δεδομένων. Σκοπός μας είναι να κατηγοριοποιήσουμε τις περιοχές του επιπέδου ως περιέχουσες • ή ο. Η fat plane η οποία ορίζεται από τη σχέση $-1 \leq f(\boldsymbol{x}) \leq 1$ επιλέγεται έτσι ώστε να μεγιστοποιείται το περιθώριο (όπως φαίνεται στο σχήμα). Σε μια τέτοια μεγιστοποίηση, ένα μικρό πλήθος σημείων, τα «διανύσματα υποστήριξης» (σε κύκλο στο σχήμα), θα βρίσκονται επάνω στα επίπεδα οριοθέτησης.*

ευθεία γραμμή στο γράφημα με 2 κατακόρυφους άξονες που αντιστοιχούν στις 2 συντεταγμένες κάθε διανύσματος $\boldsymbol{x}_i$ όταν $n = 2$, είτε ένα υπερεπίπεδο, στην περίπτωσή που $n > 2$, δηλαδή μια επιφάνεια διάστασης $(n-1)$ η οποία να περιγράφεται από μια εξίσωση της μορφής:

$$f(\boldsymbol{x}) \equiv \boldsymbol{w}^T \boldsymbol{x} + b = 0 \tag{2.2}$$

έτσι ώστε να διαχωρίζονται πλήρως τα δεδομένα εκπαίδευσης. Με άλλα λόγια, όλα τα σημεία εκπαίδευσης με $y_i = +1$ θα μπορούν να βρίσκονται από τη μια μεριά του υπερεπιπέδου (συνεπώς θα έχουν $f(\boldsymbol{x}_i) > 0$), ενώ ταυτόχρονα όλα τα σημεία εκπαίδευσης με $y_i = -1$ θα βρίσκονται από την άλλη μεριά (και θα έχουν $f(\boldsymbol{x}) < 0$). Στο σημείο αυτό να διευκρινίσουμε ότι:

- θα χρησιμοποιούμε, από εδώ και στο εξής, το συμβολισμό $\boldsymbol{w}^T \boldsymbol{x}$ για να εκφράσουμε το σύνηθες εσωτερικό γινόμενο των διανυσμάτων $\boldsymbol{w}$ και $\boldsymbol{x}$.
- το $\boldsymbol{w}$ θεωρείται κανονικοποιημένο ως προς το υπερεπίπεδο.
- το κλάσμα $\frac{b}{\|\boldsymbol{w}\|}$ εκφράζει την κάθετη απόσταση του υπερεπιπέδου από την αρχή του συστήματος, η οποία αποκαλείται και **offset**.
- ορισμένες φορές απαιτείται, για λόγους απλότητας, $b = 0$, οπότε το υπερεπίπεδο προφανώς διέρχεται από την αρχή του συστήματος συντεταγμένων και αναφερόμαστε σε αυτό ως **unbiased** υπερεπίπεδο, ενώ στη γενική περίπτωση, που κάτι τέτοιο δε συμβαίνει, καλούμε τα υπερεπίπεδα **biased**.

Στη γενική περίπτωση, περισσότερα από ένα υπερεπίπεδα μπορούν να διαχωρίσουν τα γραμμικά διαχωρίσιμα δεδομένα και θα μπορούσε, ενδεχομένως, κάποιος να αρκεστεί στον εντοπισμό οποιωνδήποτε κατάλληλων $\boldsymbol{w}$ και b που να οδηγούν σε υπερεπίπεδο που να διαχωρίζει πλήρως τα δεδομένα. Τότε, η $f(\boldsymbol{x})$ στη σχέση (2.2) θα αποτελούσε πράγματι έναν κανόνα απόφασης (με κόκκινο χρώμα στο Σχήμα 2.1). Η αλήθεια όμως είναι πως μπορούμε, και θα επιδιώξουμε, να επιτύχουμε κάτι ακόμα καλύτερο.

Οι ΜΔΥ (μεγίστου περιθωρίου) στοχεύουν, λοιπόν, στο να προσανατολίσουν το υπερεπίπεδο με τέτοιο τρόπο ώστε να δημιουργεί το μεγαλύτερο δυνατό περιθώριο, δηλαδή να εμφανίζει τη μεγαλύτερη απόσταση από τα κοντινότερά του, και από τις δυο μεριές, σημεία.

Ειδικότερα, δοθέντος ενός υπερεπιπέδου της μορφής (2.2) που διαχωρίζει τα δεδομένα, μπορούμε πάντα να αλλάζουμε την κλίμακα του $\boldsymbol{w}$ με μια σταθερά και να ρυθμίζουμε το b κατά τέτοιο τρόπο ώστε να επιτυγχάνουμε

$$\begin{aligned} \boldsymbol{w}^T\boldsymbol{x}_i + b \geqslant +1, \text{ όταν } y_i = +1 \\ \boldsymbol{w}^T\boldsymbol{x}_i + b \leqslant -1, \text{ όταν } y_i = -1 \end{aligned} \tag{2.3}$$

Οι εξισώσεις

$$\begin{aligned} \boldsymbol{w}^T\boldsymbol{x}_i + b = +1, \text{ του } H_1 \\ \boldsymbol{w}^T\boldsymbol{x}_i + b = -1, \text{ του } H_2 \end{aligned}$$

αντιπροσωπεύουν παράλληλα υπερεπίπεδα που οριοθετούν και διαχωρίζουν τα δεδομένα (με διακεκομμένες γραμμές στο Σχήμα 2.1), μια δομή που αποκαλείται **fat plane**. Δεδομένου πως το ιδανικό υπερεπίπεδο είναι εκείνο που ισαπέχει από τα H_1 και H_2, δηλαδή θεωρώντας $d_1 = d_2$ (μια απόσταση που είναι γνωστή και ως **SVM's margin**), διαπιστώνουμε πως αρκεί να προσανατολίσουμε το υπερεπίπεδο έτσι ώστε να μεγιστοποιείται η συγκεκριμένη απόσταση. Μέσω αναλυτικής γεωμετρίας καταλήγουμε εύκολα στο συμπέρασμα πως η κατακόρυφη απόσταση ανάμεσα στα υπερεπίπεδα οριοθέτησης (το διπλάσιο του SVM's margin δηλαδή) είναι:

$$2 \times (SVM's\ margin) = \frac{2}{\|\boldsymbol{w}\|} \tag{2.4}$$

Επίσης παρατηρούμε ότι οι δυο εξισώσεις (2.3) μπορούν να γραφούν με τη μορφή μιας εξίσωσης, ως εξής:

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geqslant 1 \tag{2.5}$$

Καταλήγουμε λοιπόν ότι, από όλα τα υπερεπίπεδα που διαχωρίζουν τα δεδομένα, το ιδανικό είναι αυτό που επιτρέπει τη **φαρδύτερη fat plane**, γνωστό επίσης και ως **maximum SVM's margin**, και μπορεί να βρεθεί μέσω της επίλυσης του, ισοδύνα-

μου με όσα αναφέραμε παραπάνω, **προβλήματος τετραγωνικού προγραμματισμού**[1] :

$$\boxed{\begin{array}{rl} \text{ελαχιστοποίησε το:} & \dfrac{1}{2}\|\boldsymbol{w}\|^2 \\ \\ \text{υπο συνθήκες:} & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geqslant 1 \quad i = 1, ..., m \end{array}} \tag{2.6}$$

Παρατηρείστε ότι προτιμούμε να ελαχιστοποιήσουμε την ποσότητα $\dfrac{1}{2}\|\boldsymbol{w}\|^2$ αντί να μεγιστοποιήσουμε, ισοδύναμα, την ποσότητα $\dfrac{2}{\|\boldsymbol{w}\|}$ που βρίσκεται στο δεξί μέλος της σχέσης (2.4). Ο παράγοντας $1/2$ απλώς εισάγεται για να διευκολύνει κάποιες πράξεις αργότερα.

Για την επίλυση τέτοιου είδους προβλημάτων (τετραγωνικού προγραμματισμού) υπάρχουν διαθέσιμοι αλγόριθμοι, οπότε ας θεωρήσουμε την επίλυση του (2.6) ως υπολογιστικά εφικτή.

Παρατηρούμε ότι σε ένα πρόβλημα όπως το (2.6), ορισμένα (συνήθως λίγα το πλήθος) σημεία οφείλουν να βρίσκονται επάνω στα υπερεπίπεδα οριοθέτησης, αλλιώς η fat plane θα μπορούσε να γίνει φαρδύτερη. Αυτά τα σημεία, για τα οποία ισχύει $f(\boldsymbol{x}) = \pm 1$, καλούνται **Διανύσματα Υποστήριξης** της λύσης. Παρά το γεγονός ότι η ονομασία των ΜΔΥ οφείλεται στα συγκεκριμένα διανύσματα, αυτά δεν παίζουν ιδιαίτερο ρόλο στις γενικότερες, και πιο ρεαλιστικές, περιπτώσεις που θα εξετάσουμε παρακάτω.

2.3 Πρωτεύοντα και Δυϊκά Προβλήματα στον Τετραγωνικό Προγραμματισμό

Στην παράγραφο αυτή θα αναφερθούμε σε μια διαδικασία αρκετά δημοφιλή στα προβλήματα βελτιστοποίησης και ειδικότερα θα περιγράψουμε πώς αυτή εφαρμόζεται στα προβλήματα τετραγωνικού προγραμματισμού. Ενδεχομένως η διαδικασία αυτή με μια πρώτη ματιά να μοιάζει ως περιττή, καθώς όπως θα δούμε οδηγεί απλώς στην αντικατάσταση ενός προβλήματος τετραγωνικού προγραμματισμού, της μορφής (2.6), από ένα άλλο, όπως θα διαπιστώσουμε όμως στη συνέχεια, αυτή η αντικατάσταση έχει ισχυρότατες συνέπειες.

[1] quadratic programming problem

Το γενικό πρόβλημα στον τετραγωνικό προγραμματισμό, γνωστό και ως **Πρωτεύον Πρόβλημα**, μπορεί να διατυπωθεί ως εξής:

$$\boxed{\begin{array}{rl} \text{ελαχιστοποίησε την:} & f(\boldsymbol{w}) \\ \\ \text{υπό συνθήκες:} & \begin{array}{l} g_j(\boldsymbol{w}) \leq 0 \\ h_k(\boldsymbol{w}) = 0 \end{array} \end{array}} \tag{2.7}$$

όπου η $f(\boldsymbol{w})$ είναι τετραγωνική στο $\boldsymbol{w}$, γράφεται δηλαδή στη μορφή $f(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T Q \boldsymbol{w} + \boldsymbol{c}^T \boldsymbol{w}$ με τον Q να είναι συμμετρικός πίνακας, οι $g(\boldsymbol{w})$ και $h(\boldsymbol{w})$ είναι affine ως προς $\boldsymbol{w}$, είναι δηλαδή γραμμικές συν μία σταθερά, και οι δείκτες j και k εκφράζουν τα σύνολα των ανισωτικών και εξισωτικών περιορισμών αντίστοιχα.

Η **Αρχή της Δυϊκότητας**, μας επιτρέπει για κάθε πρωτεύον πρόβλημα να αναζητούμε το αντίστοιχο **δυϊκό πρόβλημα**, το οποίο και να θεωρούμε ως έναν εναλλακτικό τρόπο επίλυσης του πρωτεύοντος προβλήματος.

Για να περάσουμε από το πρωτεύον στο δυϊκό, αρχικά γράφουμε τη **Λαγκρανζιανή**

$$L_p(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \equiv \frac{1}{2} f(\boldsymbol{w}) + \sum_j \alpha_j g_j(\boldsymbol{w}) + \sum_k \beta_k h_k(\boldsymbol{w}) \tag{2.8}$$

η οποία ενσωματώνει στην τετραγωνική μορφή όλους τους περιορισμούς πολλαπλασιασμένους με τους αντίστοιχους **πολλαπλασιαστές Lagrange**. Κατόπιν γράφουμε το ακόλουθο υποσύνολο συνθηκών για ένα μέγιστο:

$$\frac{\partial L_p}{\partial w_i} = 0, \qquad \frac{\partial L_p}{\partial \beta_k} = 0 \tag{2.9}$$

και χρησιμοποιούμε άλγεβρα στις εξισώσεις που προκύπτουν ώστε να εξαλείψουμε το $\boldsymbol{w}$ από την L_p, προς όφελος των $\boldsymbol{\alpha}$ και $\boldsymbol{\beta}$ (όπου τα $\boldsymbol{\alpha}$ και $\boldsymbol{\beta}$ περιγράφουν τα διανύσματα των α_j και β_k αντίστοιχα). Αποκαλούμε το αποτέλεσμα **Reduced Lagrangian,** $L_D(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

Το σημαντικό συμπέρασμα, που προκύπτει από την επονομαζόμενη **Ισχυρή Δυϊκότητα** (strong duality) και από τα θεωρήματα **Kuhn-Tucker**, είναι πως η λύση του δυϊκού προβλήματος:

$$\boxed{\begin{array}{rll} \text{μεγιστοποίησε το:} & L_D(\boldsymbol{\alpha}, \boldsymbol{\beta}) & \\ \text{υπό συνθήκες:} & \alpha_j \geqslant 0 & \text{για κάθε} j \end{array}} \tag{2.10}$$

είναι ισοδύναμη με τη λύση του πρωτεύοντος προβλήματος.

Στην πραγματικότητα, αυτό το αποτέλεσμα είναι γενικότερο του τετραγωνικού προγραμματισμού και ισχύει, σε γενικές γραμμές, για κάθε κυρτή $f(\boldsymbol{x})$.

Επιπλέον, αν $\widehat{\boldsymbol{w}}$ είναι η βέλτιστη λύση του πρωτεύοντος προβλήματος και $\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}$ είναι οι βέλτιστες λύσεις του δυϊκού προβλήματος, έχουμε :

$$\begin{array}{rcl} f(\widehat{\boldsymbol{w}}) & = & L_D(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) \\ \widehat{\alpha}_j g_j(\widehat{\boldsymbol{w}}) & = & 0 \quad \text{για κάθε j} \end{array} \tag{2.11}$$

Η τελευταία συνθήκη καλείται η **συμπληρωματική Karush-Kuhn-Tucker συνθήκη**. Μας υποδεικνύει ότι τουλάχιστον ένα από τα $\widehat{\alpha}_j$ και $g_j(\widehat{\boldsymbol{w}})$ πρέπει να είναι μηδέν για κάθε j. Αυτό σημαίνει ότι, από τη λύση του δυϊκού προβλήματος, μπορούμε άμεσα να αναγνωρίσουμε ανισωτικούς περιορισμούς του πρωτεύοντος προβλήματος που είναι «καρφιτσωμένοι» πάνω στο όριό τους, αυτούς δηλαδή με μη μηδενικά $\widehat{\alpha}_j$ στη λύση του δυϊκού.

2.4 Δυϊκή Διατύπωση των ΜΔΥ Μεγίστου Περιθωρίου

Η διαδικασία που περιγράψαμε στην προηγούμενη παράγραφο εφαρμόζεται άμεσα στο πρόβλημα τετραγωνικού προγραμματισμού (2.6) για τις ΜΔΥ μέγιστου περιθωρίου. Από τη στιγμή που δεν υπάρχουν εξισωτικοί περιορισμοί δε θα υπάρχουν βέβαια και β_k.

Θεωρώντας, λοιπόν, το διάνυσμα $\boldsymbol{\alpha} = (a_1, ..., a_m)$ των πολλαπλασιαστών Lagrange, όπου $a_i \geq 0 \ \forall i$, η Λαγκρανζιανή (2.8) γράφεται

$$\begin{array}{rcl} L_p & \equiv & \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} a_i[y_i(\boldsymbol{x}_i^T\boldsymbol{w} + b) - 1] \\ & = & \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} a_i y_i(\boldsymbol{x}_i^T\boldsymbol{w} + b) + \sum_{i=1}^{m} a_i \end{array} \tag{2.12}$$

Επιθυμούμε να βρούμε τα $\widehat{\boldsymbol{w}}$ και $\widehat{b}$ που ελαχιστοποιούν, και το $\widehat{\boldsymbol{\alpha}}$ που μεγιστοποιεί την (2.12). Για να το πετύχουμε, αρχικά παραγωγίζουμε την L_p ως προς $\boldsymbol{w}$ και b και θέτουμε:

$$0 = \frac{\partial L_p}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i \implies \widehat{\boldsymbol{w}} = \sum_{i=1}^{m} \widehat{a}_i y_i \boldsymbol{x}_i \tag{2.13}$$

και

$$0 = \frac{\partial L_p}{\partial b} = \sum_{i=1}^{m} a_i y_i \tag{2.14}$$

Αντικαθιστώντας τις εξισώσεις (2.13) και (2.14) στην (2.12) παίρνουμε την, ανεξάρτητη των $\boldsymbol{w}$ και b, reduced Lagrangian:

$$\begin{aligned} L_D &\equiv \sum_{i=1}^{m} a_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \quad \text{έτσι ώστε } a_i \geq 0\ \forall i \text{ και } \sum_{i=1}^{m} a_i y_i = 0 \\ &= \sum_{i=1}^{m} a_i - \frac{1}{2}\boldsymbol{\alpha}^T diag(\boldsymbol{y}) G diag(\boldsymbol{y}) \boldsymbol{\alpha} \quad \text{έτσι ώστε } a_i \geq 0\ \forall i \text{ και } \sum_{i=1}^{m} a_i y_i = 0 \end{aligned} \tag{2.15}$$

όπου με $diag(\boldsymbol{y})$ έχουμε συμβολίσει το διαγώνιο πίνακα που σχηματίζεται από το διάνυσμα $\boldsymbol{y} = (y_1, ..., y_m)$ κατά τον προφανή τρόπο και με G συμβολίζουμε τον Gram πίνακα των εσωτερικών γινομένων όλων των $\boldsymbol{x}_j$ μεταξύ τους

$$G_{ij} \equiv \boldsymbol{x}_i^T \boldsymbol{x}_j$$

Η τελευταία αυτή μορφή αποκαλείται **Δυϊκή μορφή της Πρωτεύουσας L_p (Dual form of the Primary L_p)** και είναι η μορφή που θα απαιτήσουμε, κατόπιν, να μεγιστοποιηθεί ως προς το $\boldsymbol{\alpha}$. Αυτό όμως που πρώτα έχει μεγάλη αξία να παρατηρήσουμε είναι πως η Δυϊκή αυτή μορφή απαιτεί για τον υπολογισμό της μόνο τα εσωτερικά γινόμενα των διανυσμάτων $\boldsymbol{x}_i$, κάτι που είναι ουσιώδες για την εφαρμογή, αργότερα, του Kernel τεχνάσματος (βλ. §1.5).

Παρατήρηση 2.1 Η παραπάνω εξίσωσης (2.15) μπορεί, πάντως, να γραφεί και στις ισοδύναμες μορφές:

$$\begin{aligned} L_D &= \sum_{i=1}^{m} a_i - \frac{1}{2}\sum_{i,j} a_i H_{ij} a_j \quad \text{όπου } H_{ij} \equiv y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \\ &= \sum_{i=1}^{m} a_i - \frac{1}{2}\boldsymbol{\alpha}^T H \boldsymbol{\alpha} \quad \text{έτσι ώστε } a_i \geq 0\ \forall i,\ \sum_{i=1}^{m} a_i y_i = 0 \end{aligned}$$

με την τελευταία να αποτελεί τη μορφή που έχουμε επιλέξει να χρησιμοποιήσουμε στην διατύπωση των αλγορίθμων στο τέλος αυτής και των επομένων παραγράφων.

Υπενθυμίζουμε, επίσης, ότι οι δείκτες στα $\boldsymbol{x}$ δεν εκφράζουν συντεταγμένες του διανύσματος, αλλά υποδεικνύουν σε ποιο σημείο (data point) αναφερόμαστε.

Έτσι, το δυϊκό πρόβλημα στην ολοκληρωμένη του μορφή είναι το εξής:

$$\boxed{\begin{aligned} \text{μεγιστοποίησε (ως προς } \boldsymbol{\alpha}\text{) το :} \quad & \sum_{i=1}^{m} a_i - \frac{1}{2}\boldsymbol{\alpha}^T diag(\boldsymbol{y}) G diag(\boldsymbol{y})\boldsymbol{\alpha} \\ \text{υπό συνθήκες :} \quad & a_i \geq 0 \quad \text{για κάθε } i \\ & \sum_{i=1}^{m} a_i y_i = 0 \end{aligned}} \tag{2.16}$$

που είναι ένα **πρόβλημα κυρτής τετραγωνικής βελτιστοποίησης (convex quadratic optimization problem)**. Για την επίλυσή του «τρέχουμε» έναν αλγόριθμο επίλυσης QP ο οποίος θα μας δώσει το $\widehat{\boldsymbol{\alpha}}$, το οποίο, με αντικατάσταση κατόπιν στη (2.13), θα μας δώσει το $\widehat{\boldsymbol{w}}$. Το μόνο που απομένει λοιπόν είναι ο υπολογισμός του $\widehat{b}$.

Κάθε σημείο που ικανοποιεί την (2.14) και είναι Διάνυσμα Υποστήριξης, έστω $\boldsymbol{x}_s$, θα έχει τη μορφή

$$y_s(\boldsymbol{x}_s^T \widehat{\boldsymbol{w}} + \widehat{b}) = 1$$

η οποία λόγω της (2.13) γράφεται

$$y_s \left(\sum_{k \in S} \widehat{a_k} y_k \boldsymbol{x}_k^T \boldsymbol{x}_s + \widehat{b} \right) = 1$$

όπου με S συμβολίζουμε το σύνολο των δεικτών των Διανυσμάτων Υποστήριξης. Ο προσδιορισμός του S έγκειται στον προσδιορισμό των δεικτών i για τους οποίους $a_i > 0$. Πολλαπλασιάζοντας κατά μέλη με το y_s και λαμβάνοντας υπ' όψιν ότι $y_s^2 = 1$ έχουμε

$$y_s^2 \left(\sum_{k \in S} \widehat{a_k} y_k \boldsymbol{x}_k^T \boldsymbol{x}_s + \widehat{b} \right) = y_s \iff$$

$$\widehat{b} = y_s - \sum_{k \in S} \widehat{a_k} y_k \boldsymbol{x}_k^T \boldsymbol{x}_s$$

Εναλλακτικά, και για να αποφύγουμε μέρος από το σφάλμα στρογγυλοποίησης, είναι καλύτερα, αντί να χρησιμοποιήσουμε κάποιο τυχαίο Διάνυσμα Υποστήριξης $\boldsymbol{x}_s$, να χρησιμοποιήσουμε μια μέση τιμή όλων των Διανυσμάτων Υποστήριξης στο S, οπότε

$$\boxed{\widehat{b} = \frac{1}{N_s} \sum_{s \in S} \left(y_s - \sum_{k \in S} \widehat{a_k} y_k \boldsymbol{x}_k^T \boldsymbol{x}_s \right)} \tag{2.17}$$

Έτσι έχουμε προσδιορίσει τις μεταβλητές $\widehat{w}$ και $\widehat{b}$ που καθορίζουν το βέλτιστο προσανατολισμό του διαχωρίζοντος υπερεπιπέδου, με άλλα λόγια τον κανόνα απόφασης $f(\boldsymbol{x}) = \widehat{\boldsymbol{w}}^T \boldsymbol{x} + \widehat{b}$.

Κάποιες χρήσιμες παρατηρήσεις είναι οι εξής:

- Τα σημεία με μη μηδενικό $\widehat{a}_i$ ικανοποιούν τις προϋποθέσεις ως ισότητες, δηλ. είναι διανύσματα υποστήριξης.
- Το μόνο σημείο όπου τα δεδομένα $\boldsymbol{x}_i$ εμφανίζονται στην (2.16) είναι στον υπολογισμό του πίνακα G.
- Το μόνο βήμα των υπολογισμών που είναι τάξεως n (της διάστασης του χαρακτηριστικού διανύσματος) είναι ο υπολογισμός των στοιχείων του πίνακα G.
- Όλα τα άλλα βήματα των υπολογισμών είναι τάξεως m, του πλήθους των δεδομένων σημείων (data points)

Έτσι, πηγαίνοντας από το πρωτεύον στο δυϊκό, αντικαταστήσαμε ένα πρόβλημα τάξεως n^2, της διάστασης του χαρακτηριστικού πίνακα, με ένα πρόβλημα τάξεως (κυρίως) m^2, του πλήθους των δεδομένων σημείων. Κάτι τέτοιο ίσως φαίνεται λίγο περίεργο, αφού προφανώς κάνει τα προβλήματα που περιέχουν τεράστιο πλήθος δεδομένων σημείων δυσκολότερα. Παρ' όλα αυτά διευκολύνει σημαντικά, όπως θα δούμε παρακάτω, τα προβλήματα με μέτριο πλήθος δεδομένων άλλα με τεράστια χαρακτηριστικά διανύσματα. Αυτό είναι στην πραγματικότητα και το πεδίο στο οποίο οι ΜΔΥ διαπρέπουν!

Ο αλγόριθμος SVM για γραμμικά διαχωρίσιμα δεδομένα

Ο αλγόριθμος επίλυσης ενός προβλήματος δυαδικής κατηγοριοποίησης γραμμικά διαχωρίσιμων δεδομένων με τη χρήση των ΜΔΥ ακολουθεί την εξής πορεία:

- Δημιούργησε τον πίνακα H, όπου $H_{ij} = y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$.
- Βρες το διάνυσμα $\boldsymbol{\alpha} = (a_1, ..., a_m)$ έτσι ώστε να μεγιστοποιείται η

$$\sum_{i=1}^{m} a_i - \frac{1}{2} \boldsymbol{\alpha}^T H \boldsymbol{\alpha}$$

υπό συνθήκες

$$a_i \geq 0 \ \forall \ i \text{ και } \sum_{i=1}^{m} a_i y_i = 0. \tag{2.18}$$

Αυτό επιτυγχάνεται με τη χρήση ενός QP solver.

- Υπολόγισε το $\boldsymbol{w} = \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i$.
- Καθόρισε το σύνολο S των Διανυσμάτων Υποστήριξης, βρίσκοντας τους δείκτες i για τους οποίους $a_i > 0$.
- Υπολόγισε το $b = \frac{1}{N_s} \sum_{s \in S} \left(y_s - \sum_{k \in S} a_k y_k \boldsymbol{x}_k^T \boldsymbol{x}_s \right)$.
- Κάθε νέο σημείο $\boldsymbol{x}'$ κατατάσσεται υπολογίζοντας το $y' = sgn(\boldsymbol{w}^T \boldsymbol{x}' + b)$.

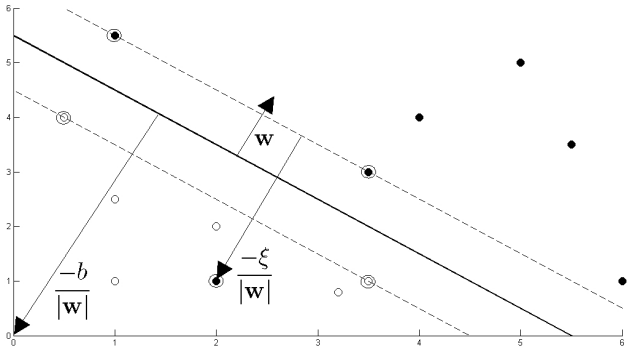
2.5 Η 1-Norm Soft-Margin ΜΔΥ και το Δυϊκό της

Ας εγκαταλείψουμε τώρα την, όχι και τόσο ρεαλιστική, υπόθεση ότι υπάρχει υπερεπίπεδο το οποίο διαχωρίζει πλήρως τα δεδομένα εκπαίδευσης, και ας επεκτείνουμε τη μεθοδολογία των ΜΔΥ προσαρμόζοντάς της τη δυνατότητα να διαχειρίζεται δεδομένα τα οποία δεν είναι πλήρως γραμμικά διαχωρίσιμα (Σχήμα 2.2). Για να το επιτύχουμε, εισάγουμε στις ανισότητες (2.3) μια «χαλαρή» (slack), μη αρνητική, μεταβλητή ξ_i για κάθε δεδομένο $\boldsymbol{x}_i$. Άν το δεδομένο (σημείο) είναι από εκείνα που μπορούν να διαχωριστούν από ένα fat plane, τότε $\xi_i = 0$. Αν δεν μπορεί, τότε η ποσότητα $\xi_i > 0$ εκφράζει το πόσο απέχει από το να μπορούσε (amount of descrepancy), οπότε οι ανισότητες (2.3) τροποποιούνται ως εξής:

$$\begin{gathered} \boldsymbol{w}^T \boldsymbol{x}_i + b \geq +1 - \xi_i, \text{ όταν } y_i = +1 \\ \boldsymbol{w}^T \boldsymbol{x}_i + b \leq -1 + \xi_i, \text{ όταν } y_i = -1 \\ \xi_i \geq 0 \ \forall \ i \end{gathered} \tag{2.19}$$

οι οποίες συνδυαζόμενες με αντίστοιχο τρόπο όπως και στη (2.5), οδηγούν στη σχέση:

$$y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1 + \xi_i \geq 0 \ \text{ όπου } \xi_i \geq 0 \ \forall \ i. \tag{2.20}$$



Σχήμα 2.2: *Οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ) στην περίπτωση των γραμμικά μη διαχωρίσιμων δεδομένων. Σε κάθε σημείο καταλογίζεται μια «ποινή» ή οποία αυξάνει ανάλογα με την απόσταση του σημείου από το επίπεδο οριοθέτησης που όφειλε να το «συγκρατεί», λαμβάνοντας έτσι θετική τιμή όταν το σημείο βρίσκεται από τη «λάθος μεριά» και μηδενική τιμή όταν το σημείο βρίσκεται από τη «σωστή μεριά».*

Με άλλα λόγια λοιπόν, σ΄ αυτήν εδώ την περίπτωση, των **soft margin ΜΔΥ**, σε κάθε σημείο που βρίσκεται από τη λάθος μεριά του αντίστοιχου επιπέδου οριοθέτησης «καταλογίζουμε» μία «ποινή», η οποία αυξάνει ανάλογα με την απόστασή του απ΄ αυτό. Μιας και στόχο μας αποτελεί να ελαχιστοποιήσουμε το πλήθος των εσφαλμένων ταξινομήσεων, ένας λογικός τρόπος να τροποποιήσουμε την αντικειμενική συνάρτηση του προβλήματος (2.6) είναι εισάγοντας έναν όρο που θα «παρακινεί» τα ξ_i να γίνουν όσο το δυνατόν μικρότερα, και μηδενικά όπου αυτό είναι εφικτό, ως εξής:

$$\begin{array}{rl} \text{ελαχιστοποίησε το:} & \dfrac{1}{2}\|\boldsymbol{w}\|^2 + \lambda \displaystyle\sum_{i=1}^{m} \xi_i \\ \\ \text{υπο συνθήκες:} & \xi_i \geq 0 \\ & y_i(w \cdot \boldsymbol{x}_i + b) - 1 + \xi_i \geq 0, \qquad i = 1, ..., m \end{array} \tag{2.21}$$

Η παράμετρος λ ρυθμίζει τη σχέση «ανταλλαγής» (trade-off) που δημιουργείται ανάμεσα στη μείωση των ποινών των slack μεταβλητών ξ_i και στην αύξηση του πλάτους του περιθωρίου. Αντιμετωπίζουμε, δηλαδή, πλέον ένα πρόβλημα το οποίο, πέρα από βελτιστοποίηση, απαιτεί επιπλέον και **(regularization)**, με την 1-norm soft-margin ΜΔΥ να υιοθετεί, όπως και η ονομασία της υπονοεί, ένα γραμμικό άθροισμα των (θετικών) ξ_i ως τον regularization όρο.

Παρατήρηση 2.2 Μια πιθανή εναλλακτική είναι η 2-norm soft-margin ΜΔΥ, όπου

ο regularization όρος θα ήταν $\sum_i^m {\xi_i}^2$. Παρ' όλα αυτά, επειδή αυτή η μέθοδος δίνει κάπως πιο πολύπλοκες εξισώσεις, θα αποφύγουμε να τη χρησιμοποιήσουμε.

Καθώς κινούμαστε κατά μήκος της trade-off καμπύλης $0 < \lambda < \infty$, συναντούμε, εναλλάξ, λύσεις οι οποίες «προτιμούν» ένα πραγματικά φαρδύ fat plane (αδιαφορώντας για το πόσα σημεία βρίσκονται μέσα του, ή από τη λάθος μεριά του) και λύσεις οι οποίες είναι τόσο «φειδωλές» στο να επιτρέπουν ασυμφωνίες ώστε συμβιβάζονται με ένα fat plane με σχεδόν καθόλου περιθώριο. Οι πρώτες είναι λιγότερο ακριβείς με τα δεδομένα εκπαίδευσης αλλά πιθανώς πιο εύρωστες με τα νέα δεδομένα. Οι τελευταίες είναι μεν όσο το δυνατόν περισσότερο ακριβείς με τα δεδομένα εκπαίδευσης αλλά είναι πιθανώς εύθραυστες (και λιγότερο ακριβείς) με τα νέα δεδομένα. (Η επιλογή του λ είναι μια *design trade-off* που πρέπει να πραγματοποιηθεί ανάλογα με τις εκάστοτε ανάγκες. Περισσότερες πληροφορίες σχετικά με τους regularization όρους αλλά και τη μεταβολή της trade-off σχέσης καθώς αυτοί μεταβάλλονται, μπορεί να βρει κάποιος στο [7] (Να προστεθεί επιπλέον βιβλιογραφία)).

Το σημαντικό, πάντως, είναι πως οποιαδήποτε μη αρνητική τιμή του λ επιτρέπει να υπάρξει κάποια λύση, είτε τα δεδομένα είναι γραμμικά διαχωρίσιμα είτε όχι. Αυτό φαίνεται αν παρατηρήσουμε το γεγονός πως η $\boldsymbol{w} = \mathbf{0}$ είναι πάντα μια εφικτή (όχι όμως η βέλτιστη) λύση του (2.21) για αρκετά μεγάλα θετικά ξ_i ανεξάρτητα από την τιμή του λ. Αν όμως υπάρχει εφικτή λύση θα πρέπει, βεβαίως, να υφίσταται και κάποια βέλτιστη λύση.

Υπολογίζοντας, λοιπόν, εκ νέου τη Λαγκρανζιανή, την οποία όπως και πριν θα επιδιώξουμε να ελαχιστοποιήσουμε ως προς τις μεταβλητές $\boldsymbol{w}$, b και ξ_i και να μεγιστοποιήσουμε ως προς την $\boldsymbol{\alpha}$ (όπου $a_i \geq 0$, $\mu_i \geq 0$ για κάθε i), έχουμε:

$$L_p \equiv \frac{1}{2}\|\boldsymbol{w}\|^2 + \lambda \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} a_i[y_i(\boldsymbol{x}_i^T \boldsymbol{w} + b) - 1 + \xi_i] - \sum_{i=1}^{m} \mu_i \xi_i \tag{2.22}$$

Παραγωγίζοντας ως προς $\boldsymbol{w}$, b και ξ_i και θέτοντας τις παραγώγους ίσες με μηδέν έχουμε:

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = \mathbf{0} \Longrightarrow \boldsymbol{w} = \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i \tag{2.23}$$

$$\frac{\partial L_p}{\partial b} = 0 \Longrightarrow \sum_{i=1}^{m} a_i y_i = 0 \tag{2.24}$$

$$\frac{\partial L_p}{\partial \xi_i} = 0 \Longrightarrow \lambda = a_i + \mu_i \tag{2.25}$$

Αντικαθιστώντας τις σχέσεις αυτές στη (2.22) προκύπτει ότι η L_D έχει, παραδόξως, την ίδια μορφή με εκείνη της (2.15), με μόνη διαφορά είναι ότι εδώ η (2.25) σε συνδυασμό με τις $\mu_i \geq 0$ για κάθε i επιβάλουν έναν περιορισμό άνω φράγματος λ στα a_i, επιπροσθέτως του μηδενικού κάτω φράγματος (ένας τέτοιος περιορισμός, $0 \leq a_i \leq \lambda$, καλείται **box contraint**). Το Δυϊκό πρόβλημα λοιπόν, και σε αυτή την περίπτωση των 1-norm soft margin ΜΔΥ, διατυπώνεται ως εξής:

$$\boxed{\begin{array}{rl} \text{μεγιστοποίησε (ως προς } \boldsymbol{\alpha}\text{) το :} & \displaystyle\sum_{i=1}^{m} a_i - \frac{1}{2}\boldsymbol{\alpha}^T diag(\boldsymbol{y}) G diag(\boldsymbol{y})\boldsymbol{\alpha} \\ \text{υπό συνθήκες :} & 0 \leq a_i \leq \lambda \quad \text{για κάθε } i \\ & \displaystyle\sum_{i=1}^{m} a_i y_i = 0 \end{array}} \tag{2.26}$$

με τον τύπο (2.23) υπολογισμού του $\widehat{\boldsymbol{w}}$ να παραμένει ο ίδιος με τον (2.13), ενώ και το $\widehat{b}$ υπολογίζεται επίσης με τον ίδιο τρόπο όπως στη (2.17) νωρίτερα, με τη μόνη διαφορά ότι το σύνολο των Διανυσμάτων Υποστήριξης που χρησιμοποιούνται σε αυτόν τον υπολογισμό καθορίζεται πλέον από την εύρεση των δεικτών i για τους οποίους $0 < \widehat{a}_i \leq \lambda$.

Παρατηρούμε ότι, εκτός από κάποιες εκφυλισμένες περιπτώσεις διπλών μηδενικών (double zeros), ισχύουν τα εξής:

$$\begin{array}{rcl} \widehat{a}_i = 0 & \Longleftrightarrow & \text{το δεδομένο } i \text{ βρίσκεται από τη σωστή μεριά του fat plane} \\ 0 < \widehat{a}_i < \lambda & \Longleftrightarrow & \text{το δεδομένο } i \text{ βρίσκεται ακριβώς επάνω στο όριο του fat plane} \\ & & \text{(είναι δηλαδή διάνυσμα υποστήριξης)} \\ \widehat{a}_i = \lambda & \Longleftrightarrow & \text{το δεδομένο i βρίσκεται είτε εντός είτε από τη λάθος μεριά του fat plane} \end{array} \tag{2.27}$$

Επίσης παρατηρούμε, και πάλι, ότι καθώς μετακινούμε το λ προς την τιμή μηδέν, αποδίδοντας σε όλο και περισσότερα a_i την τιμή λ, παίρνουμε λύσεις με διαρκώς αυξανόμενο πλήθος λανθασμένα ταξινομημένων σημείων, αλλά φαρδύτερα fat planes.

Αν και η υπόθεση της γραμμικότητας (δηλαδή η χρήση υπερεπιπέδου για το διαχωρισμό των δεδομένων) είναι αρκετά περιοριστική, το μοντέλο που περιγράφεται στην (2.26) έχει κάποια πρακτική χρησιμότητα σε προβλήματα όπου υπάρχει κάποιος λόγος να πιστεύουμε ότι η απάντηση είναι (κατά κάποιο τρόπο τουλάχιστον) γραμμική ως προς τις συνιστώσες του χαρακτηριστικού διανύσματος. Ακόμα όμως δεν έχουμε χρησιμοποιήσει το «δυνατό μας χαρτί», τη μαθηματική θεωρία δηλαδή που αναπτύξαμε στο 1ο Κεφάλαιο. Αυτή είναι που, με τον τρόπο που περιγράφουμε στην επόμενη παράγραφο, θα μας απαλλάξει από το «βραχνά» της γραμμικότητας.

Ο αλγόριθμος SVM επίλυσης προβλήματος Δυαδικής Κατηγοριοποίησης όχι πλήρως γραμμικά διαχωρίσιμων δεδομένων

Ο αλγόριθμος SVM για την επίλυση μιας δυαδικής κατηγοριοποίησης δεδομένων, τα οποία δεν είναι πλήρως γραμμικά διαχωρίσιμα, ακολουθεί την εξής πορεία:

- Δημιούργησε τον πίνακα H, όπου $H_{ij} = y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$.
- Αποφάσισε πόσο σημαντικές θα είναι οι μη ορθές ταξινομήσεις, καθορίζοντας μια κατάλληλη τιμή για την παράμετρο λ.
- Βρες το διάνυσμα $\boldsymbol{\alpha} = (a_1, ..., a_m)$ έτσι ώστε να μεγιστοποιείται η

$$\sum_{i=1}^{m} a_i - \frac{1}{2} \boldsymbol{\alpha}^T H \boldsymbol{\alpha}$$

  υπό συνθήκες

$$0 \leq a_i \leq \lambda \ \forall \ i \text{ και } \sum_{i=1}^{m} a_i y_i = 0. \tag{2.28}$$

  Αυτό επιτυγχάνεται με τη χρήση ενός QP solver.

- Υπολόγισε το $\boldsymbol{w} = \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i$.
- Καθόρισε το σύνολο S των Διανυσμάτων Υποστήριξης, βρίσκοντας τους δείκτες i για τους οποίους $0 < a_i \leq \lambda$.
- Υπολόγισε το $b = \frac{1}{N_s} \sum_{s \in S} \left(y_s - \sum_{k \in S} a_k y_k \boldsymbol{x}_k^T \boldsymbol{x}_s \right)$.
- Κάθε νέο σημείο $\boldsymbol{x}'$ κατατάσσεται υπολογίζοντας το $y' = sgn(\boldsymbol{w}^T \boldsymbol{x}' + b)$.

2.6 Το Kernel Τέχνασμα στις ΜΔΥ

Είμαστε σε θέση, πλέον, να παρουσιάσουμε τη μέθοδο που προσδίδει στις ΜΔΥ την πραγματική τους ισχύ. Ας φανταστούμε μία, όχι απαραίτητα γραμμική, απεικόνιση $\boldsymbol{\phi} : X \subseteq \mathbb{R}^n \longrightarrow \mathcal{H}$, η οποία ενσωματώνει, κατά μία έννοια, τα m το πλήθος n-διάστατα χαρακτηριστικά διανύσματα σε έναν, κατά πολύ υψηλότερου βαθμού, N-διάστατο χώρο $\mathcal{H}$,

$$\boldsymbol{x} \;\; n\text{-διάστατο} \longrightarrow \;\; \boldsymbol{\phi}(\boldsymbol{x}) \;\; N\text{-διάστατο} \qquad (n < N) \tag{2.29}$$

Η βασική ιδέα, όπως φαίνεται και στο Σχήμα 2.3, είναι πως μία, σε μεγάλο βαθμό μη-γραμμική, διαχωριστική επιφάνεια στο n-διάστατο χώρο ενδέχεται να απεικονίζεται (ή έστω να προσεγγίζεται καλά) από ένα γραμμικό υπερεπίπεδο στο N-διάστατο χώρο $\mathcal{H}$.

Για να αντιληφθούμε καλύτερα πώς μπορεί να γίνεται αυτό, ας θεωρήσουμε την απεικόνιση από τις 2 στις 5 διαστάσεις:

$$(\boldsymbol{x}_0, \boldsymbol{x}_1) \xrightarrow{\phi} ({\boldsymbol{x}_0}^2, \boldsymbol{x}_0\boldsymbol{x}_1, {\boldsymbol{x}_1}^2, \boldsymbol{x}_0, \boldsymbol{x}_1) \tag{2.30}$$

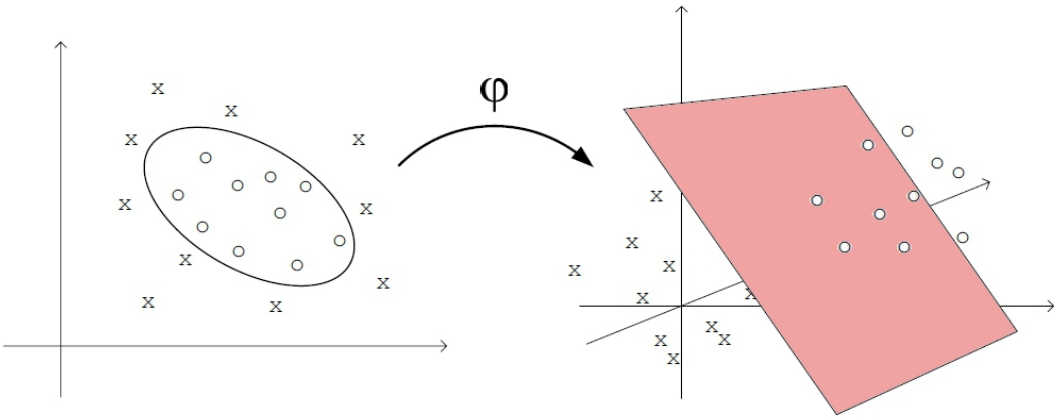
Με αυτήν την απεικόνιση, ένας κανόνας απόφασης $f(\boldsymbol{x})$, ο οποίος κατασκευάζεται ως γραμμικός στον χώρο ενσωμάτωσης, γίνεται αρκετά γενικός ώστε να περιλαμβάνει όλες τις γραμμικές και τις τετραγωνικές μορφές (γραμμές, ελλείψεις, υπερβολές) του αρχικού χαρακτηριστικού χώρου, δηλαδή

$$f(\boldsymbol{x}) = F[\boldsymbol{\phi}(\boldsymbol{x})] \equiv \langle \boldsymbol{W} \cdot \boldsymbol{\phi}(\boldsymbol{x}) \rangle_{\mathcal{H}} + B \tag{2.31}$$

όπου τα κεφαλαία γράμματα αντιστοιχούν σε ποσότητες του χώρου ενσωμάτωσης, ενώ χρησιμοποιούμε το συμβολισμό $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ για να εκφράσουμε το εσωτερικό γινόμενο εντός του χώρου ενσωμάτωσης $\mathcal{H}$.

Αν και σ' αυτό το παράδειγμα $N = 5$, ο χώρος ενσωμάτωσης $\mathcal{H}$ θα μπορούσε να είναι ένας χώρος τεράστιας διάστασης, όπως ένα εκατομμύριο ή ένα δισεκατομμύριο, ή ακόμα και άπειρης διάστασης καθώς, όπως θα δούμε παρακάτω, κάτι τέτοιο δε μας δημιουργεί κανένα πρόβλημα.

Το ερώτημα που ανακύπτει άμεσα είναι: πώς μπορούμε να βρούμε, βασιζόμενοι στα δεδομένα μας, τα $\boldsymbol{W}$ και B στον χώρο ενσωμάτωσης; Ας δοκιμάσουμε να εργαστούμε ακριβώς όπως και πριν, λαμβάνοντας υπ' όψιν ότι τώρα βρισκόμαστε σε έναν, δυνάμει, πολύ υψηλότερης, ενδεχομένως και άπειρης, διάστασης χώρο. Το πρωτεύον πρόβλημα



Σχήμα 2.3: *Όταν τα χαρακτηριστικά διανύσματα απεικονίζονται από έναν χαμηλής διάστασης χώρο (εδώ διάστασης 2) σε έναν υψηλότερης διάστασης χώρο (εδώ διάστασης 3), μη γραμμικές επιφάνειες διαχωρισμού μπορούν να προσεγγιστούν καλώς από γραμμικές. Στην πράξη χρησιμοποιούνται πολύ υψηλής, ακόμα και άπειρης, διάστασης χώροι ενσωμάτωσης, οι οποίοι όμως υπεισέρχονται στον υπολογισμό της ΜΔΥ εμμέσως, διαμέσου του kernel τεχνάσματος.*

(σε αντιστοιχία με το (2.21)) είναι:

$$\begin{aligned} &\text{ελαχιστοποίησε το:} \quad \frac{1}{2}\|\boldsymbol{W}\|^2 + \lambda \sum_i \Xi_i \\ &\text{υπό συνθήκες:} \quad \begin{array}{l} \Xi_i \geq 0, \\ y_i(\langle \boldsymbol{W} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i)\rangle_{\mathcal{H}} + B) - 1 + \Xi_i \geq 0 \quad i = 1, ..., m \end{array} \end{aligned} \tag{2.32}$$

Αυτό είναι ένα πρόβλημα τετραγωνικού προγραμματισμού σε έναν, ενδεχομένως, εκατομμυριο-διάστατο, δισεκατομμυριο-διάστατο ή ακόμα και άπειρης διάστασης (!!) χώρο, κατά πάσα συνεπώς πιθανότητα αδύνατο να ελεγχθεί από ένα συνηθισμένο ηλεκτρονικό υπολογιστή. Ας δούμε όμως και το δυϊκό του, το οποίο προκύπτει πως είναι το εξής:

$$\begin{aligned} &\text{μεγιστοποίησε (ως προς } a) \text{ το :} \quad \sum_{i=1}^{m} a_i - \frac{1}{2}\boldsymbol{\alpha}^T diag(y) \mathcal{K}_{ij} diag(y) \boldsymbol{\alpha} \\ &\text{υπό συνθήκες :} \quad 0 \leq a_i \leq \lambda \quad \text{για κάθε } i \\ &\qquad\qquad \sum_{i=1}^{m} a_i y_i = 0 \end{aligned} \tag{2.33}$$

Παρατηρούμε ότι είναι ακριβώς το ίδιο με το (2.26) μόνο που εδώ ο $Gram$ πίνακας G_{ij} έχει αντικατασταθεί από τον, ας τον αποκαλέσουμε **kernel πίνακα** $\mathcal{K}_{ij}$,

$$\mathcal{K}_{ij} \equiv \langle \boldsymbol{\phi}(\boldsymbol{x}_i) \cdot \boldsymbol{\phi}(\boldsymbol{x}_j) \rangle_{\mathcal{H}} \tag{2.34}$$

Αυτό είναι πολύ σημαντικό επίτευγμα, καθώς το πρόβλημα τετραγωνικού προγραμματισμού (2.33) δεν είναι δυσκολότερο να λυθεί απ' ότι το αυθεντικό πρόβλημα (2.26)! Ζούνε και τα δύο σε έναν χώρο διάστασης m, όσο το πλήθος των δεδομένων σημείων, και τροφοδοτούνται και τα δύο από έναν σταθερό πίνακα, τον G_{ij} στη μια περίπτωση και τον $\mathcal{K}_{ij}$ στην άλλη, ο οποίος προϋπολογίζεται με βάση τα δεδομένα.

Καταφέραμε λοιπόν να «στριμώξουμε» την «κατάρα» της υψηλής διαστασιμότητας του χώρου ενσωμάτωσης σε μια πολύ στενή γωνία, δηλαδή στον υπολογισμό μόνο των m^2 το πλήθος τιμών του πίνακα $\mathcal{K}_{ij}$. Τώρα θα την εξολοθρεύσουμε πλήρως με το **τέχνασμα Kernel**.

Η δυνατότητα εφαρμογής του τεχνάσματος, του οποίου τη φιλοσοφία περιγράψαμε ήδη στην παράγραφο 1.5, βασίζεται στο γεγονός ότι, ουσιαστικά, σε κανένα στάδιο της διαδικασίας δεν χρειάζεται να γνωρίζουμε την απεικόνιση $\boldsymbol{\phi} : X \subseteq \mathbb{R}^n \longrightarrow \mathcal{H}$. Το μόνο που πραγματικά χρειαζόμαστε είναι ένας τρόπος υπολογισμού του kernel πίνακα $\mathcal{K}_{ij} \equiv \langle \boldsymbol{\phi}(\boldsymbol{x}_i) \cdot \boldsymbol{\phi}(\boldsymbol{x}_j) \rangle_{\mathcal{H}}$, δηλαδή ενός πίνακα μεγέθους $m \times m$ με τις μαθηματικές ιδιότητες ενός χώρου με εσωτερικό γινόμενο υψηλότερης διάστασης. Ο υποτιθέμενος αυτός χώρος $\mathcal{H}$ θα μπορούσε, λοιπόν, να είναι ο RKHS χώρος του συνόλου X των $\boldsymbol{x}_i$ (όπως αυτός περιγράφεται αναλυτικά στην Παράγραφο 1.1), αποτελούμενος από τις reproducing kernel συναρτήσεις $k_{x_i} \equiv \boldsymbol{\phi}(x_i)$ των x_i, οι οποίες παράγονται μέσω μιας, κάποιας, χαρακτηριστικής απεικόνισης $\boldsymbol{\phi}(\boldsymbol{x})$. Η ύπαρξη ενός (και μόνον ενός) τέτοιου χώρου για κάθε kernel συνάρτηση $K : X \times X \longrightarrow \mathbb{R}$ εξασφαλίζεται από το Θεώρημα 1.6 (Moore), επιτρέποντάς μας τελικά να υπολογίζουμε:

$$\mathcal{K}_{ij} \equiv \langle \boldsymbol{\phi}(\boldsymbol{x}_i) \cdot \boldsymbol{\phi}(\boldsymbol{x}_j) \rangle_{\mathcal{H}} \equiv K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

Με πιο απλά λόγια, επιλέγοντας μια οποιαδήποτε kernel συνάρτηση $K : X \times X \longrightarrow \mathbb{R}$ και χρησιμοποιώντας τη σχέση $\mathcal{K}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ επιτυγχάνουμε να υπολογίσουμε με έμμεσο τρόπο τα στοιχεία του kernel πίνακα $\mathcal{K}_{ij} \equiv \langle \boldsymbol{\phi}(\boldsymbol{x}_i) \cdot \boldsymbol{\phi}(\boldsymbol{x}_j) \rangle_{\mathcal{H}}$ της αντίστοιχης απεικόνισης $\boldsymbol{\phi} : X \subseteq \mathbb{R}^n \longrightarrow \mathcal{H}$, παρακάμπτοντας στην ουσία τον υπολογισμό των εσωτερικών γινομένων εντός του τεράστιας ή άπειρης διάστασης αντίστοιχου RKHS χώρου $\mathcal{H}$.

Ήδη γνωρίζουμε έναν πιθανό kernel πίνακα, το Gram πίνακα G_{ij}, ο οποίος στην ουσία αντιστοιχεί στον kernel πίνακα που προκύπτει αν επιλέξουμε ως kernel συνάρτηση τη γραμμική $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j$.

Στο σημείο αυτό ας αναφερθούμε σε ορισμένες, αποδείξιμες, γενικές ιδιότητες των *kernel* συναρτήσεων $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$:

- ο $\mathcal{K}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ πρέπει να είναι συμμετρικός (in i and j) και να έχει μη αρνητικές ιδιοτιμές ($Mercer's\ Theorem$)
- Κάθε multinomial combination kernel συναρτήσεων είναι *kernel* συνάρτηση. Δηλαδή μπορούμε ελεύθερα να συνδυάζουμε *kernel* συναρτήσεις μέσω πολλαπλασιασμού, πρόσθεσης και αλλαγής κλίμακας με μια σταθερά.
- η $K(\boldsymbol{h}(\boldsymbol{x}_i), \boldsymbol{h}(\boldsymbol{x}_j))$ είναι *kernel* για κάθε $\boldsymbol{h}$, αρκεί η $K(,)$ να είναι *kernel*. Αυτή η ιδιότητα γενικεύει την αρχική ιδέα του χώρου ενσωμάτωσης.
- η $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = g(\boldsymbol{x}_i)g(\boldsymbol{x}_j)$ είναι πάντα *kernel* για κάθε συνάρτηση g.

Αφού λοιπόν καταλήξουμε σε μια kernel συνάρτηση και λύσουμε το πρόβλημα τετραγωνικού προγραμματισμού (2.33), τότε ο τελικός κανόνας απόφασής μας για κάθε νέο χαρακτηριστικό διάνυσμα $\boldsymbol{x}$ είναι

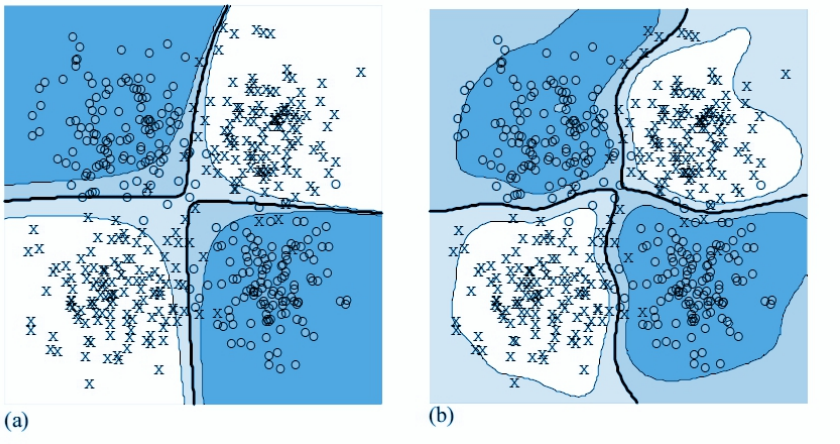
$$f(\boldsymbol{x}) = \sum_{i=1}^{m} \widehat{a}_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + \widehat{b} \tag{2.35}$$

όπου, θεωρώντας και πάλι μέση τιμή, ισχύει:

$$\begin{aligned} b &= \tfrac{1}{N_s} \sum_{s \in S} \left(y_s - \sum_{k \in S} a_k y_k \langle \boldsymbol{\phi}(\boldsymbol{x}_k) \cdot \boldsymbol{\phi}(\boldsymbol{x}_s) \rangle_{\mathcal{H}} \right) \\ &= \tfrac{1}{N_s} \sum_{s \in S} \left(y_s - \sum_{k \in S} a_k y_k K(\boldsymbol{x}_k, \boldsymbol{x}_s) \right) \end{aligned} \tag{2.36}$$

2.7 Η επιλογή της κατάλληλης kernel συνάρτησης

Ενώ η κατασκευή της ιδανικής *kernel* για κάθε συγκεκριμένο πρόβλημα ενδεχομένως να προϋποθέτει και κάποιες «καλλιτεχνικές» ικανότητες, ορισμένες πολύ γενικές *kernel* συναρτήσεις αποδεικνύονται ιδιαιτέρως ισχυρές στην επίλυση προβλημάτων



Σχήμα 2.4: *ΜΔΥ οι οποίες μαθαίνουν να διαχωρίζουν το επίπεδο. Τα δεδομένα εισόδου σχεδιάζονται από τέσσερις 2-διάστατες Gaussians , με μια μικρή επικάλυψη, δίνοντας σε όσα βρίσκονται διαγωνίως αντικριστά την ίδια σήμανση (x ή o). Οι έντονες γραμμές είναι οι επιφάνειες των κανόνων απόφασης $f(\boldsymbol{x}) = 0$ που προκύπτουν από τις ΜΔΥ. Οι απαλότερα σχεδιασμένες γραμμές απεικονίζουν τις $f(\boldsymbol{x}) = \pm 1$.(α) Πολυωνυμική kernel με $d = 8$. (β) Gaussian radial basis function kernel.*

«πραγματικής κατάστασης». Συχνά, μπορούμε απλώς να δοκιμάζουμε ορισμένες από αυτές και να επιλέγουμε αυτήν που δείχνει να δουλεύει καλύτερα. Οι ακόλουθες είναι κάποιες καλές αρχικές επιλογές:

$$
\begin{aligned}
\textbf{γραμμική:} \quad & K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j \\
\textbf{δύναμη:} \quad & K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T \boldsymbol{x}_j)^d, \quad 2 \leq d \leq 20 \quad \text{(για παράδειγμα)} \\
\textbf{πολυωνυμική:} \quad & K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (a\boldsymbol{x}_i^T \boldsymbol{x}_j + b)^d \\
\textbf{σιγμοειδής:} \quad & K(\boldsymbol{x}_i, \boldsymbol{x}_j) = tanh(a\boldsymbol{x}_i^T \boldsymbol{x}_j + b) \\
\textbf{Gaussian radial basis συνάρτηση:} \quad & K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right)
\end{aligned}
\tag{2.37}
$$

Στην παράγρ. 2.3 του [1] θα βρείτε περισσότερα standard kernels ενώ το κεφάλαιο

13 του ίδιου βιβλίου περιγράφει πολλά εξειδικευμένα *kernels*, π.χ. για σύγκριση συμβολοσειρών ή αποσπασμάτων κειμένου, για αναγνώριση εικόνας και για ένα μεγάλο πλήθος άλλων εφαρμογών. Η βιβλιογραφία βεβαίως που ασχολείται με το θέμα δεν περιορίζεται στο παράδειγμα που μόλις αναφέραμε (βλ. επίσης [2] κ.α.)

Στο Σχήμα 2.4 παρουσιάζεται ένα παράδειγμα χρήσης μιας πολυωνυμικού *kernel* συνάρτησης με $d = 8$ και μιας Gaussian radial basis kernel. Είναι χαρακτηριστικό των Gaussian Kernel ότι επηρεάζονται περισσότερο από τοπικές-γειτονικές επιδράσεις (γεγονός που μπορεί να θεωρηθεί είτε ως θετικό είτε ως αρνητικό ανάλογα με την περίπτωση), ενώ τα πολυωνυμικά *kernels* αναζητούν ομαλότερες, πιο σφαιρικές λύσεις.

Αν και ξεφεύγει από το αντικείμενο της συγκεκριμένης εργασίας, να αναφέρουμε πάντως ότι το *kernel* τέχνασμα δεν βρίσκει εφαρμογή μόνο στις ΜΔΥ (δηλαδή σε αλγορίθμους βασιζόμενους σε υπερεπίπεδα διαχωρισμού), αλλά και σε πλήθος άλλων αλγορίθμων της αναγνώρισης προτύπου, για παράδειγμα στην principal component analysis (PCA) και στον Fischer discriminant algorithm. Στα [1] και [3] μπορείτε να βρείτε εκτεταμένες πληροφορίες σχετικά με αυτούς τους kernel-based learning αλγορίθμους.

2.8 Μερικές πρακτικές συμβουλές αναφορικά στις ΜΔΥ

Η Gaussian radial basis kernel συνάρτηση είναι πολύ δημοφιλής, καθώς έχει μόνο μία ρυθμιζόμενη μεταβλητή, σ, και είναι εύκολο να υποθέσουμε μια αρχική τιμή δοκιμής, δηλ. οποιαδήποτε χαρακτηριστική απόσταση ανάμεσα σε κοντινά σημεία στον χώρο των χαρακτηριστικών. Θα μπορούσαμε να πούμε πως η Gaussian kernel κατηγοριοποιεί, σε ένα βαθμό, με την τοπική-γειτονική «συγκατάθεση».

Για τις πολυωνυμικές *kernel* συναρτήσεις, ξεκινήστε επιλέγοντας a και b ώστε να κάνετε την ποσότητα $a\boldsymbol{x}_i \cdot \boldsymbol{x}_j + b$ να βρίσκετε ανάμεσα στα ± 1 για κάθε i και j. Η δύναμη d ερμηνεύεται, πολύ απλοποιημένα, ως το πλήθος των διαφορετικών χαρακτηριστικών που επιθυμούμε η σύγκριση να «αναμιγνύει». Συνεπώς, για $d = 1$ (γραμμική) διαχωρίζει το χώρο κατά ένα χαρακτηριστικό τη φορά, για $d = 2$ κοιτάζει ζεύγη χαρακτηριστικών ταυτόχρονα, κ.ο.κ. Επίσης πολύ απλουστευμένα, η διαφορά ανάμεσα στην kernel δύναμη και στην πολυωνυμική εντοπίζεται στο κατά πόσο επιθυμούμε να λάβουμε υπ όψιν μόνο ακριβώς d χαρακτηριστικά τη φορά (δύναμη), ή όλους τους συνδυασμούς ανά d ή λιγότερων το πλήθος χαρακτηριστικών (πολυωνυμική). Παρ ολ αυτά, αυτές οι ερμηνείες δεν πρέπει να λαμβάνονται υπ όψιν κατά γράμμα. Ειδικότερα, ένα μεγαλύτερο d δεν είναι απαραίτητα καλύτερο.

Σχετικά με την επιλογή της παραμέτρου λ ας αναφέρουμε απλώς ότι στη σχετική βιβλιογραφία θα συναντήσει κανείς διάφορες μεθόδους που αφορούν στο συγκεκριμένο ζήτημα.

Υπάρχουν, τέλος, πολλά κόλπα και παρακάμψεις που μπορούν να επιταχύνουν τη λύση μιας ΜΔΥ, οι οποίες σχετίζονται κυρίως με το γενικό πρόβλημα τετραγωνικού προγραμματισμού, και τα καλά πακέτα ΜΔΥ τα εκμεταλλεύονται. Για παράδειγμα, ένα καλό πακέτο οφείλει να εκμεταλλεύεται την αραιότητα στο χαρακτηριστικό χώρο γλυτώνοντας υπολογισμούς. Ένα πολύ καλό πακέτο είναι το Thorsten Joachim's SVM-light (http://svmlight.joachims.org.) (Implementing Support Vector Machines in C)το οποίο διανέμεται ελεύθερα στο διαδίκτυο. Το Gist [Software Tools for Support Vector Machine Classification and for Kernel Principal Components Analysis at http://microarray.cpmc.columbia.edu/gist.]είναι μία ακόμα δημοφιλής ελεύθερη εφαρμογή. Τέλος, στην ιστοσελίδα http://www.support-vector.net. υπάρχει σελίδα με παραπομπές σε μια ποικιλία ΜΔΥ εφαρμογών.

Ο Αλγόριθμος επίλυσης ΜΔΥ με τη χρήση Kernel

Προκειμένου να χρησιμοποιηθούν οι ΜΔΥ στην επίλυση ενός προβλήματος κατηγοριοποίησης δεδομένων που δεν είναι γραμμικά διαχωρίσιμα, αρχικά πρέπει να επιλέξουμε μια kernel συνάρτηση και να ρυθμίσουμε τις παραμέτρους της κατά τρόπο που οι απεικονίσεις των δεδομένων μας, στον αντίστοιχο χώρο ενσωμάτωσης, να είναι γραμμικά διαχωρίσιμες. Κάτι τέτοιο απαιτεί μάλλον περισσότερο «καλλιτεχνικές» ικανότητες παρά επιστημονική ακρίβεια και είναι κάτι που επιτυγχάνεται εμπειρικά μέσω δοκιμών. Κάποιες καλές επιλογές αποτελούν οι Radial Basis, οι Πολυωνυμικές και οι Sigmoidal kernel συναρτήσεις.

Το πρώτο βήμα είναι, λοιπόν, η επιλογή της kernel συνάρτησης $K : X \times X \longrightarrow \mathbb{R}$, και κατ επέκταση, εμμέσως βέβαια, της χαρακτηριστικής απεικόνισης $\boldsymbol{x} \longmapsto \boldsymbol{\phi}(\boldsymbol{x})$. Κατόπιν:

- Δημιούργησε τον πίνακα H, όπου $H_{ij} = y_i y_j \langle \boldsymbol{\phi}(\boldsymbol{x}_i) \cdot \boldsymbol{\phi}(\boldsymbol{x}_j) \rangle_{\mathcal{H}} = y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$.
- Αποφάσισε πόσο σημαντικές θα είναι οι μη ορθές ταξινομήσεις, καθορίζοντας μια κατάλληλη τιμή για την παράμετρο λ.
- Βρες το διάνυσμα $\boldsymbol{\alpha} = (a_1, ..., a_m)$ έτσι ώστε να μεγιστοποιείται η

$$\sum_{i=1}^{m} a_i - \frac{1}{2} \boldsymbol{\alpha}^T H \boldsymbol{\alpha}$$

υπό συνθήκες

$$0 \leq a_i \leq \lambda \ \forall \ i \text{ και } \sum_{i=1}^{m} a_i y_i = 0. \tag{2.38}$$

Αυτό επιτυγχάνεται με τη χρήση ενός QP solver.

- Υπολόγισε το $\boldsymbol{w} = \sum_{i=1}^{m} a_i y_i \boldsymbol{\phi}(\boldsymbol{x}_i)$.
- Καθόρισε το σύνολο S των Διανυσμάτων Υποστήριξης, βρίσκοντας τους δείκτες i για τους οποίους $0 < a_i \leq \lambda$.
- Υπολόγισε το

$$\begin{aligned} b &= \frac{1}{N_s} \sum_{s \in S} \left(y_s - \sum_{k \in S} a_k y_k \langle \boldsymbol{\phi}(\boldsymbol{x}_k) \cdot \boldsymbol{\phi}(\boldsymbol{x}_s) \rangle_{\mathcal{H}} \right) \\ &= \frac{1}{N_s} \sum_{s \in S} \left(y_s - \sum_{k \in S} a_k y_k K(\boldsymbol{x}_k, \boldsymbol{x}_s) \right). \end{aligned}$$

- Κάθε νέο σημείο $\boldsymbol{x}'$ κατατάσσεται υπολογίζοντας το

$$\begin{aligned} y' &= sgn(\boldsymbol{w}^T \boldsymbol{x}' + b) \\ &= sgn\left(\sum_{i=1}^{m} a_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \right) \end{aligned}$$

Κεφάλαιο 3

Ο Αλγόριθμος Kernel Least-Mean-Square και οι τεχνικές Αραίωσής του.

3.1 Adaptive Learning Αλγόριθμοι στην αντιμετώπιση Προβλημάτων Μάθησης

Ας υποθέσουμε ότι επιθυμούμε να «ανακαλύψουμε» το μηχανισμό μιας συνάρτησης $F : X \subset \mathbb{R}^M \longrightarrow \mathbb{R}$, (θα την αποκαλούμε και **true filter**), έχοντας στη διάθεσή μας μια ακολουθία παραδειγμάτων εισόδων-εξόδων $((\boldsymbol{x}_1, d_1), (\boldsymbol{x}_2, d_2), \ldots, (\boldsymbol{x}_n, d_n), \ldots)$, όπου $\boldsymbol{x}_n \in X \subset \mathbb{R}^M$ και $d_n \in \mathbb{R}$ για κάθε $n \in \mathbb{N}$. Στόχος ενός τυπικού **Adaptive Learning αλγορίθμου** είναι ο προσδιορισμός, με βάση τα δεδομένα, μιας σχέσης εισόδου-εξόδου, $f_{\boldsymbol{w}}$, μέσα από μια παραμετρική κλάση συναρτήσεων $H = \{f_{\boldsymbol{w}} : X \longrightarrow \mathbb{R},\ \boldsymbol{w} \in \mathbb{R}^\nu\}$, έτσι ώστε να ελαχιστοποιείται η τιμή μιας προκαθορισμένης loss function $L(\boldsymbol{w})$ η οποία, σε κάθε βήμα n, υπολογίζει το σφάλμα ανάμεσα στο πραγματικό αποτέλεσμα, d_n, και στην εκτίμησή του $f_{\boldsymbol{w}}(\boldsymbol{x}_n)$.

3.2 Ο αλγόριθμος Least Mean Square (LMS)

Στην πιο συνηθισμένη μορφή του LMS αλγορίθμου, υιοθετείται ως χώρος υποθέσεως η κλάση των γραμμικών συναρτήσεων $H_1 = \{f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^M\}$, ενώ ως loss function χρησιμοποιείται το **μέσο τετραγωνικό σφάλμα (mean square error - MSE)**

$$L(\boldsymbol{w}) \equiv E[|d_n - f_{\boldsymbol{w}}(\boldsymbol{x})|^2] = E[|d_n - \boldsymbol{w}^T\boldsymbol{x}|^2]$$

Συμβολίζουμε, επίσης, την ποσότητα

$$e_n = d_n - \boldsymbol{w}_{n-1}^T\boldsymbol{x}_n$$

και την αποκαλούμε **a priori σφάλμα** σε κάθε βήμα n. Χρησιμοποιώντας, τώρα, τη Stochastic gradient descent μέθοδο, σε κάθε χρονική στιγμή $n = 1, 2, \ldots, N$, η κλίση του μέσου τετραγωνικού σφάλματος

$$-\nabla L(w) = 2E[(d_n - \boldsymbol{w}_{n-1}^T \boldsymbol{x}_n)(\boldsymbol{x}_n)] = 2E[e_n \boldsymbol{x}_n]$$

προσεγγιζόμενη από την τιμή της για κάθε δεδομένη χρονική στιγμή n

$$E[e_n \boldsymbol{x}_n] \approx e_n \boldsymbol{x}_n$$

οδηγεί στη **step update (ή weight-update) εξίσωση**, η οποία προς την κατεύθυνση της μείωσης είναι:

$$\boldsymbol{w}_n = \boldsymbol{w}_{n-1} + \mu e_n \boldsymbol{x}_n \tag{3.1}$$

όπου μ είναι η παράμετρος που εκφράζει πόσο μεγάλο είναι το «βήμα» μας προς την κατεύθυνση της καθόδου (αποκαλείται και **βήμα εκμάθησης**). Η τελευταία εξίσωση μας οδηγεί στις ακόλουθες σχέσεις:

$$\begin{aligned}
&\boldsymbol{w}_0 = \mathbf{0} \quad (\text{υπόθεση}) \\
&\boldsymbol{w}_1 = \boldsymbol{w}_0 + \mu e_1 \boldsymbol{x}_1 = \mu e_1 \boldsymbol{x}_1 \\
&\boldsymbol{w}_2 = \underbrace{\mu e_1 \boldsymbol{x}_1}_{\boldsymbol{w}_1} + \mu e_2 \boldsymbol{x}_2 \\
&\vdots \\
&\boldsymbol{w}_n = \mu \sum_{k=1}^{n} e_k \boldsymbol{x}_k
\end{aligned} \tag{3.2}$$

Αναλυτικότερα, τα βήματα του αλγορίθμου ακολουθούν την εξής πορεία:

$$\textbf{Αρχικοποίηση:}\quad \boldsymbol{w}_0 = \mathbf{0}$$

$$\textbf{Βήμα 1:}\quad \text{έρχεται το } (\boldsymbol{x}_1, d_1)$$

$$\text{Βήμα 2:}\quad f(\boldsymbol{x}_1) \equiv \boldsymbol{w}_0^T \boldsymbol{x}_1 = 0$$

$$\text{Βήμα 3:}\quad e_1 = d_1 - f(\boldsymbol{x}_1) = d_1$$

$$\text{Βήμα 4:}\quad \boldsymbol{w}_1 = \boldsymbol{w}_0 + \mu e_1 \boldsymbol{x}_1 = \mu e_1 \boldsymbol{x}_1$$

$$\textbf{Βήμα 5:}\quad \text{έρχεται το } (\boldsymbol{x}_2, d_2)$$

$$\text{Βήμα 6:}\quad f(\boldsymbol{x}_2) \equiv \boldsymbol{w}_1^T \boldsymbol{x}_2$$

$$\text{Βήμα 7:}\quad e_2 = d_2 - f(\boldsymbol{x}_2)$$

$$\text{Βήμα 8:}\quad \boldsymbol{w}_2 = \boldsymbol{w}_1 + \mu e_2 \boldsymbol{x}_2$$

$$\textbf{Βήμα 9:}\quad \text{έρχεται το } (\boldsymbol{x}_3, d_3)$$

$$\vdots$$

Ο Κώδικας Least-Mean Square

- $\boldsymbol{w}$=**0**
- for i=1 to N (π.χ. N=5000)

  $f \equiv \boldsymbol{w}^T \boldsymbol{x}_i$

  $e = d_i - f$

  $\boldsymbol{w} = \boldsymbol{w} + \mu e \boldsymbol{x}_i$
- end for

Από τα διάφορα **κριτήρια σύγκλισης** του LMS αλγορίθμου, μάλλον το δημοφιλέστερο διατυπώνεται ως εξής: εφόσον το φίλτρο (true filter) είναι γραμμικό, δηλαδή $F(\boldsymbol{x}_n) = \boldsymbol{w}_*^T \boldsymbol{x}_n$ και η $\boldsymbol{x}_n$ είναι WSS (weak-sense stationarity) process, τότε

$\mu \sum_{n=1}^{\infty} e_i \boldsymbol{x}_i \longrightarrow \boldsymbol{w}_*$ και $E[|e_n|^2] \longrightarrow c$(σταθερά), καθώς $n \longrightarrow \infty$, αν το μ ικανοποιεί τη συνθήκη $0 < \mu < \frac{1}{\lambda_{\max}}$, όπου με $\lambda_{\max}$ συμβολίζουμε τη μεγαλύτερη ιδιοτιμή του πίνακα συσχέτισης (correlation matrix) $R = E[\boldsymbol{x}_n \boldsymbol{x}_n^T]$. Στην πράξη, αρκεί $0 < \mu < \frac{1}{tr(R)}$.

Καθώς, επίσης, ο LMS εμφανίζει ευαισθησία στην κλίμακα των $\boldsymbol{x}_i$, καθίσταται δύσκολη η επιλογή βήματος εκμάθησης μ τέτοιου ώστε να εξασφαλίζει τη σταθερότητα του αλγορίθμου για τις διάφορες τιμές των δεδομένων εισόδου. Για να παρακάμψουμε το συγκεκριμένο πρόβλημα μπορούμε να χρησιμοποιήσουμε μια παραλλαγή του LMS αλγορίθμου, η οποία προκύπτει αν η τελευταία εξίσωση της παραπάνω επαναληπτικής διαδικασίας αντικατασταθεί απο την

$$\boldsymbol{w} = \boldsymbol{w} + \frac{\mu e}{\|\boldsymbol{x}_i\|^2} \boldsymbol{x}_i \tag{3.3}$$

Ο αλγόριθμος καλείται πλέον **Normalized LMS** και έχει αποδειχθεί ότι ο βέλτιστος ρυθμός εκμάθησης του επιτυγχάνεται για $\mu = 1$.

Συνεπώς, σε κάθε περίπτωση, μετά από μια εκπαίδευση n βημάτων, το **βάρος** $\boldsymbol{w}_n$ εκφράζεται ως γραμμικός συνδυασμός των προηγούμενων και του τελευταίου δεδομένου εισόδου, σταθμισμένων από τα αντίστοιχα a priori σφάλματα. Ακόμα σημαντικότερο είναι το γεγονός ότι η διαδικασία εισόδου-εξόδου του συγκεκριμένου συστήματος εκπαίδευσης μπορεί να εκφραστεί αποκλειστικά με όρους εσωτερικών γινομένων:

$$f(\boldsymbol{x}_{n+1}) = \boldsymbol{w}_n^T \boldsymbol{x}_{n+1} = \mu \sum_{k=1}^{n} e_k \boldsymbol{x}_k^T \boldsymbol{x}_{n+1} \tag{3.4}$$

όπου

$$e_n = d_n - \mu \sum_{k=1}^{n-1} e_k \boldsymbol{x}_k^T \boldsymbol{x}_n$$

Άρα, χρησιμοποιόντας το τέχνασμα Kernel, ο αλγόριθμος LMS επεκτείνεται εύκολα στον Kernel LMS, που αποτελεί το αντικείμενο της επόμενης παραγράφου.

3.3 Ο αλγόριθμος Kernel LMS (KLMS)

Υπενθυμίζουμε ότι ως kernel συνάρτηση θεωρούμε μια συνεχή, συμμετρική, θετικά ορισμένη συνάρτηση $K : X \times X \longrightarrow \mathbb{R}$ (βλ. Κεφάλαιο 1). Στις περισσότερο διαδεδομένες kernel συναρτήσεις περιλαμβάνονται η **Gaussian kernel**:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-a\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2} \tag{3.5}$$

και η **πολυωνυμική kernel**:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T \boldsymbol{x}_j + 1)^p \tag{3.6}$$

Για κάθε kernel συνάρτηση $K : X \times X \longrightarrow \mathbb{R}$, το Θεώρημα 1.6 (Moore) μας εξασφαλίζει την ύπαρξη, μοναδικού, χώρου RKHS, $\mathcal{H}$, (όπως αυτός ορίστηκε στην Παράγραφο 1.1) ο οποίος παράγεται μέσω μιας αντίστοιχης, όχι απαραίτητα γραμμικής, χαρακτηριστικής απεικόνισης $\boldsymbol{\phi} : X \longrightarrow \mathcal{H}$, που ενσωματώνει τα δεδομένα $\boldsymbol{x} \in X$ στο χώρο συναρτήσεων $\mathcal{H}$ με τη μορφή $\boldsymbol{\phi}(\boldsymbol{x}) = k_{\boldsymbol{x}}$, δηλαδή με τη μορφή των αντίστοιχων, για κάθε σημείο $\boldsymbol{x}$, kernel συναρτήσεων. Η σημαντική ιδιότητα που συνδέει όλα τα παραπάνω, είναι βεβαίως πως, για κάθε $\boldsymbol{x}_1, \boldsymbol{x}_2 \in X$ ισχύει:

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle k_{\boldsymbol{x}_1}, k_{\boldsymbol{x}_2} \rangle_{\mathcal{H}} = \langle \boldsymbol{\phi}(\boldsymbol{x}_1), \boldsymbol{\phi}(\boldsymbol{x}_2) \rangle_{\mathcal{H}}$$

(Υπενθυμίζουμε ότι η συμμετρία συμπεριλαμβάνεται στις ιδιότητες των πραγματικών kernel συναρτήσεων).

Θεωρώντας λοιπόν το χώρο των γραμμικών συναρτησιακών $H_2 = \{T_{\boldsymbol{w}} : \mathcal{H} \longrightarrow \mathbb{R}, T_{\boldsymbol{w}}(\boldsymbol{\phi}(\boldsymbol{x})) = \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle_{\mathcal{H}}, \boldsymbol{w} \in \mathcal{H}\}$ ως νέο χώρο υποθέσεως, ο αλγόριθμος Kernel LMS (KLMS) δεν είναι άλλος από τον LMS εκτελούμενο για την ακολουθία παραδειγμάτων $((\boldsymbol{\phi}(\boldsymbol{x}_1), d_1), \ldots, (\boldsymbol{\phi}(\boldsymbol{x}_n), d_n))$, αντιστοιχίζοντας έτσι μια μη-γραμμική διαδικασία εντός του Ευκλειδείου χώρου εισόδων X σε μια γραμμική εντός του RKHS χώρου $\mathcal{H}$. Ο μηχανισμός του συνοψίζεται ως εξής:

Αναζητούμε συνάρτηση

$$f(\boldsymbol{x}_n) \equiv T_{\boldsymbol{w}}(\boldsymbol{\phi}(\boldsymbol{x}_n)) =< \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}_n) >_{\mathcal{H}} \quad , \boldsymbol{w} \in \mathcal{H}$$

έτσι ώστε να ελαχιστοποιείται η loss function

$$L(w) \equiv E[|d_n - f(\boldsymbol{x}_n)|^2] = E[|d_n - \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}_n) \rangle_{\mathcal{H}}|^2]$$

Θέτοντας και πάλι

$$e_n = d_n - f(\boldsymbol{x}_n)$$

παραγωγίζουμε κατά Frechet

$$\nabla L(\boldsymbol{w}) = -2E[e_n \boldsymbol{\phi}(\boldsymbol{x}_n)]$$

την οποία και πάλι, κατά τη λογική του LMS, προσεγγίζουμε με την τιμή της για κάθε χρονική στιγμή n

$$\nabla L(\boldsymbol{w}) = -2e_n \boldsymbol{\phi}(\boldsymbol{x}_n)$$

λαμβάνοντας, τελικά, προς την κατεύθυνση της ελαχιστοποιήσεως

$$\boldsymbol{w}_n = \boldsymbol{w}_{n-1} + \mu e_n \boldsymbol{\phi}(\boldsymbol{x}_n) \tag{3.7}$$

Ο αλγόριθμος «τρέχει» ως εξής:

$$\boldsymbol{w}_0 = 0$$

$$\boldsymbol{w}_1 = \mu e_1 \boldsymbol{\phi}(\boldsymbol{x}_1)$$

$$\boldsymbol{w}_2 = \mu e_1 \boldsymbol{\phi}(\boldsymbol{x}_1) + \mu e_2 \boldsymbol{\phi}(\boldsymbol{x}_2)$$

$$\vdots$$

$$\boldsymbol{w}_n = \mu \sum_{k=1}^{n} e_k \boldsymbol{\phi}(\boldsymbol{x}_k)$$

Σε κάθε βήμα n θα έχουμε λοιπόν:

$$\begin{aligned} f(\boldsymbol{x}_n) = T_{\boldsymbol{w}_{n-1}}(\boldsymbol{\phi}(\boldsymbol{x}_n)) &=< \boldsymbol{w}_{n-1}, \boldsymbol{\phi}(\boldsymbol{x}_n) >_{\mathcal{H}} \\ &=< \mu \sum_{k=1}^{n-1} e_k \boldsymbol{\phi}(\boldsymbol{x}_k), \boldsymbol{\phi}(\boldsymbol{x}_n) >_{\mathcal{H}} \\ &= \mu \sum_{k=1}^{n-1} e_k < \boldsymbol{\phi}(\boldsymbol{x}_k), \boldsymbol{\phi}(\boldsymbol{x}_n) >_{\mathcal{H}} \\ &= \mu \sum_{k=1}^{n-1} e_k K(\boldsymbol{x}_k, \boldsymbol{x}_n) \end{aligned}$$

Τελικά, η σχέση εισόδου-εξόδου, μετά από N βήματα εκπαίδευσης του αλγορίθμου εκμάθησης, είναι

$$\boldsymbol{w}_n = \mu \sum_{k=1}^{n} e_k \boldsymbol{\phi}(\boldsymbol{x}_k) \tag{3.8}$$

$$f(\boldsymbol{x}_n) = \mu \sum_{k=1}^{n-1} e_k K(\boldsymbol{x}_k, \boldsymbol{x}_n) \tag{3.9}$$

Ο Κώδικας Kernel Least-Mean Square

- **Είσοδος**: τα δεδομένα $(\boldsymbol{x}_n, y_n)$ και το πλήθος τους N
- **Έξοδος**: το ανάπτυγμα $\boldsymbol{w} = \sum_{k=1}^{N} \alpha_k K(\cdot, \boldsymbol{u}_k)$, όπου $\alpha_k = \mu e_k$

- **Αρχικοποίηση**:
  $f^0 = 0$
  n: το βήμα μάθησης
  μ: η παράμετρος μ του βήματος εκμάθησης
  Όρισε το διάνυσμα $\boldsymbol{\alpha} = 0$, τον πίνακα $D = \{\}$ και τις παραμέτρους της *kernel* συνάρτησης.

- **for** $n = 1 \dots N$ **do**

  if $n == 1$ **then**

  $f_n = 0$

  else

  Υπολόγισε το filter output $f_n = \sum_{k=1}^{M} \alpha_k K(\boldsymbol{u}_k, \boldsymbol{x}_n)$

  end if

  Υπολόγισε το σφάλμα : $e_n = d_n - f_n$

  $\alpha_n = \mu e_n$

  Καταχώρησε το νέο κέντρο $\boldsymbol{u}_n = \boldsymbol{x}_n$ στη λίστα με τα κέντρα, δηλαδή, $D = \{D, \boldsymbol{u}_n\}$, $\boldsymbol{\alpha}^T = \{\boldsymbol{\alpha}^T, \alpha_n\}$

- **end for**

Μπορούμε και εδώ, στη θέση της (3.7), να χρησιμοποιήσουμε μια κανονικοποιημένη μορφή, όπως για παράδειγμα την:

$$\boldsymbol{w}_n = \boldsymbol{w}_{n-1} + \frac{\mu e_n}{K(\boldsymbol{x}_n, \boldsymbol{x}_n)} \boldsymbol{\phi}(\boldsymbol{x}_n) \tag{3.10}$$

λαμβάνοντας έτσι τον **normalized KLMS (KNLMS)**. Στον αλγόριθμο, αυτό μεταφράζεται ως αντικατάσταση του βήματος $a_n = \mu e_n$ με το $a_n = \frac{\mu e_n}{\kappa}$, όπου το $\kappa = K(\boldsymbol{x}_n, \boldsymbol{x}_n)$ μπορεί να έχει υπολογιστεί σε προηγούμενο βήμα.

Οι ιδιότητες σύγκλισης και ευστάθειας του KLMS αποτελούν, ακόμα, ανοικτό πεδίο έρευνας. Λαμβάνοντας υπ όψιν ότι ο αλγόριθμος KLMS είναι ο LMS εκτελούμενος σε RKHS χώρο, οι ιδιότητες του LMS μεταφέρονται απ ευθείας στον KLMS. Όμως, ενώ οι ιδιότητες του LMS έχουν αποδειχθεί για Ευκλείδειους χώρους, οι RKHS χώροι που χρησιμοποιούνται στην πράξη είναι άπειρης διάστασης χώροι Hilbert. Αξίζει επίσης να επισημάνουμε ότι, πρόσφατα, αναπτύχθηκε μια γενίκευση του KLMS, ο **complex kernel LMS**, ο οποίος δρα απ ευθείας σε μιγαδικούς RKHS χώρους.

3.4 Αραίωση της Λύσης

Το σημαντικότερο μειονέκτημα του αλγορίθμου KLMS είναι πως το πλήθος των σημείων $\boldsymbol{x}_n$ που εμπλέκονται στην εκτίμηση του αποτελέσματος αυξάνεται διαρκώς, με αποτέλεσμα να απαιτείται ολοένα και περισσότερη μνήμη, καθώς και υπολογιστική ισχύς, όσο ο αλγόριθμος «τρέχει». Το σύνολο των σημείων αυτών μπορεί να θεωρηθεί ως ένα «**λεξικό**» D το οποίο αποθηκεύεται στη μνήμη.

Έτσι, δεν είναι δυνατόν να χρησιμοποιηθεί αυτούσιος ο KLMS σε πραγματικά προβλήματα, μιας και το λεξικό μεγαλώνει απεριόριστα με αποτέλεσμα, μετά από ορισμένο πλήθος επαναλήψεων, το ανάπτυγμα (3.9) να γίνεται τόσο μεγάλο ώστε να γεμίζει τη μνήμη του υπολογιστή ενώ παράλληλα θα απαιτεί και τεράστιο χρόνο υπολογισμού, καθιστώντας έτσι την εφαρμογή μας άχρηστη. Είναι λοιπόν αναγκαία η εύρεση μεθόδων που να περιορίζουν το μέγεθος του αναπτύγματος.

Υπάρχουν, πράγματι, διαθέσιμες ορισμένες στρατηγικές αντιμετώπισης του φαινομένου περιγράψαμε παραπάνω, οδηγώντας σε αραιές λύσεις. Αυτό το επιτυγχάνουν διαμορφώνοντας το λεξικό, έως ένα βαθμό, κατά τα πρώτα στάδια του αλγορίθμου καταχωρώντας αρκετά νέα σημεία (ή αλλιώς νέα **κέντρα**, όπως θα αποκαλούμε στη συνέχεια τα σημεία που καταχωρούνται στο λεξικό) αυξάνοντας έτσι την έκτασή του, στη συνέχεια όμως επιτρέπουν σε νέα σημεία να προστίθενται ως κέντρα μόνο αν πληρούν συγκεκριμένα κριτήρια. Η γενική δομή ενός τέτοιου αλγορίθμου είναι η ακόλουθη:

Ο Κώδικας Kernel Least-Mean Square με αραίωση

- **Είσοδος**: τα δεδομένα $(\boldsymbol{x}_n, y_n)$ και το πλήθος τους N

- **Έξοδος**: το ανάπτυγμα $\boldsymbol{w} = \sum_{k=1}^{M} \alpha_k K(\cdot, \boldsymbol{u}_k)$, όπου $\alpha_k = \mu e_k$

- **Αρχικοποίηση**:
  $f^0 = 0, M = 0$
  n: το βήμα μάθησης
  μ: η παράμετρος μ του βήματος εκμάθησης
  Όρισε διάνυσμα $\boldsymbol{\alpha} = 0$, πίνακα $D = \{\}$ και τις παραμέτρους του *kernel*

- **for** $n = 1 \ldots N$ **do**

  if $n == 1$ **then**

  $f_n = 0$

else

Υπολόγισε το filter output $f_n = \sum_{k=1}^{M} \alpha_k K(\boldsymbol{u}_k, \boldsymbol{x}_n)$

end if

Υπολόγισε το σφάλμα : $e_n = d_n - f_n$

$\alpha_n = \mu e_n$

Έλεγχος Προϋποθέσεων Αραίωσης

if Οι Προϋποθέσεις Αραίωσης Ικανοποιούνται **then**

$M = M + 1$

Καταχώρησε το νέο κέντρο $\boldsymbol{u}_M = \boldsymbol{x}_n$ στη λίστα με τα κέντρα, δηλαδή, $D = \{D, \boldsymbol{u}_M\}$, $\boldsymbol{\alpha}^T = \{\boldsymbol{\alpha}^T, \alpha_n\}$

end if

- **end for**

3.4.1 Platt's novelty criterion

Παράδειγμα μιας διάσημης στρατηγικής αραίωσης αποτελεί το **Platt's novelty criterion**, σύμφωνα με το οποίο για κάθε νέο ζεύγος $(\boldsymbol{x}_n, d_n)$ που εξετάζεται, η απόφαση για το αν θα καταχωρηθεί το $\boldsymbol{x}_n$ στο D λαμβάνεται άμεσα με βάση τους εξής απλούς κανόνες:

- Αρχικά, υπολογίζεται η απόσταση του νέου σημείου $\boldsymbol{x}_n$ από το λεξικό D_{n-1}

$$dist = \min_{\boldsymbol{u}_k \in D_{n-1}} \{\|\boldsymbol{x}_n - \boldsymbol{u}_k\|\}$$

- Άν η απόσταση αυτή είναι μικρότερη από ένα προκαθορισμένο κατώτατο όριο δ_1, γεγονός που σημαίνει ότι το υπό εξέταση διάνυσμα βρίσκεται «πολύ» κοντά σε κάποιο από τα ήδη υπάρχοντα στο λεξικό διανύσματα, τότε το νέο διάνυσμα δεν καταχωρείται στο λεξικό D_{n-1} οπότε παραμένει $D_n = D_{n-1}$.

- Αλλιώς, υπολογίζεται το σφάλμα $e_n = d_n - f_n$. Αν η τιμή $|e_n|$ είναι μικρότερη από ένα προκαθορισμένο όριο δ_2 τότε το νέο σημείο πάλι δεν καταχωρείται και παραμένει $D_n = D_{n-1}$. Μόνο στην περίπτωση που $|e_n| \geq \delta_2$ το $\boldsymbol{x}_n$ καταχωρείται στο D_{n-1} και το λεξικό διαμορφώνεται ως $D_n = D_{n-1} \cup \{\boldsymbol{x}_n\}$.

Βεβαίως, κάθε φορά που καταχωρούμε ένα νέο σημείο στο λεξικό D δεν θα πρέπει να παραλείπουμε να καταχωρούμε και τον αντίστοιχο συντελεστή $a_n = \mu e_n$ στη λίστα των συντελεστών a.

3.4.2 Coherence Based Sparsification Strategy

Ένα άλλο γνωστό σχήμα είναι η αποκαλούμενη **στρατηγική αραίωσης βάση συνεκτικότητας (coherence based sparsification strategy)**, σύμφωνα με την οποία ένα σημείο $\boldsymbol{x}_n$ καταχωρείται στο λεξικό αν η συνεκτικότητά του ξεπερνά ένα δεδομένο όριο ϵ_0, δηλαδή αν

$$\max_{\boldsymbol{u}_k \in D_n} \{|K(\boldsymbol{x}_n, \boldsymbol{u}_k)|\} \leq \epsilon_0 \tag{3.11}$$

Έχει αποδειχθεί ότι η διάσταση ενός λεξικού που διαμορφώνεται σύμφωνα με τον παραπάνω κανόνα παραμένει πεπερασμένη, καθώς το n τείνει στο άπειρο.

3.4.3 Surprise Criterion

Μια περισσότερο εκλεπτυσμένη στρατηγική είναι το αποκαλούμενο **surprise criterion**. Surprise ενός νέου ζεύγους δεδομένων $(\boldsymbol{x}_n, d_n)$ ως προς ένα σύστημα εκμάθησης T ορίζουμε την αρνητική λογαριθμική πιθανότητα του $(\boldsymbol{x}_n, d_n)$ δεδομένης της learning system's hypothesis της κατανομής των δεδομένων, δηλαδή

$$S_T(\boldsymbol{x}_n, d_n) = -\ln p((\boldsymbol{x}_n, d_n)|T)$$

Με βάση το παραπάνω μέτρο, το ζεύγος δεδομένων μπορεί να ταξινομηθεί σε μία από τις ακόλουθες κατηγορίες:

- Abnormal: $S_T(\boldsymbol{x}_n, d_n) > \delta_1$,
- Learnable: $\delta_1 \geq S_T(\boldsymbol{x}_n, d_n) \geq \delta_2$,
- Reduntant: $S_T(\boldsymbol{x}_n, d_n) < \delta_2$,

όπου δ_1, δ_2 είναι παράμετροι που σχετίζονται με το πρόβλημα. Σύμφωνα με αυτή τη μέθοδο, μόνο όταν ένα νέο ζεύγος δεδομένων ταξινομείται ως learnable καταχωρείται στο λεξικό. Στην περίπτωση του $KLMS$ με Gaussian εισόδους και δίχως a priori πληροφορίες, το surprise measure είναι

$$S_n = \frac{1}{2} ln(r_n) + \frac{e_n^2}{2r_n} \tag{3.12}$$

όπου

$$r_n = \lambda + K(\boldsymbol{x}_n, \boldsymbol{x}_n) - \max_{\boldsymbol{u}_k \in D_n} \{\frac{K^2(\boldsymbol{x}_n, \boldsymbol{u}_k)}{K(\boldsymbol{u}_k, \boldsymbol{u}_k)}\} \tag{3.13}$$

με λ μια ορισμένη από τον χρήστη παράμετρο τακτοποίησης.

3.4.4 Quantized Kernel Least-Mean Square (QKLMS)

Ένα σημαντικό μειονέκτημα των μεθόδων αραίωσης που αναφέραμε έως τώρα είναι ότι διατηρούν, επ' αόριστον και απαράλλαχτες, τις παλαιότερες πληροφορίες (με τη μορφή των a_i που συγκροτούν το $\boldsymbol{\alpha}$), αδυνατώντας έτσι να αντεπεξέλθουν σε αλλαγές που ενδέχεται να επηρεάσουν το κανάλι. Υπό την έννοια αυτή θα μπορούσαν να θεωρηθούν περισσότερο ως online παρά ως adaptive filtering algorithms. Μια διαφορετική τεχνική επιβολής αραίωσης στη λύση του $KLMS$, η οποία όμως διαθέτει επιπροσθέτως και την ικανότητα προσαρμογής σε ενδεχόμενες αλλαγές του καναλιού, είναι ο **κβαντισμός (quantization)** των δεδομένων εκπαίδευσης στο χώρο εισόδων. Η φιλοσοφία της μεθόδου εμφανίζεται στον αλγόριθμο που ακολουθεί:

Ο Κώδικας Quantized Kernel Least-Mean Square (QKLMS)

- **Είσοδος**: τα δεδομένα $(\boldsymbol{x}_n, y_n)$ και το πλήθος N
- **Έξοδος**: το ανάπτυγμα $\boldsymbol{w} = \sum_{k=1}^{N} \alpha_k K(\cdot, \boldsymbol{u}_k)$, όπου $\alpha_k = \mu e_k$
- **Αρχικοποίηση**:
  $f^0 = 0$
  n: το βήμα μάθησης
  μ: η παράμετρος μ του βήματος εκμάθησης
  Όρισε διάνυσμα $\boldsymbol{\alpha} = 0$, πίνακα $D = \{\}$, τις παραμέτρους του *kernel* και το μέγεθος δ της κβάντισης.

  $M = 0$
- **for** $n = 1 \ldots N$ **do**

  if $n == 1$ **then**

  $f_n = 0$

  else

Υπολόγισε το filter output $f_n = \sum_{k=1}^{M} \alpha_k K(\boldsymbol{u}_k, \boldsymbol{x}_n)$

end if

Υπολόγισε το σφάλμα : $e_n = d_n - f_n$

$\alpha_n = \mu e_n$

Υπολόγισε την απόσταση μεταξύ $\boldsymbol{x}_n$ και D: $dist = \min_{\boldsymbol{u}_k \in D} \|\boldsymbol{x}_n - \boldsymbol{u}_k\| = \|\boldsymbol{x}_n - \boldsymbol{u}_l\|$, για ένα $l \in \{1, \ldots, M\}$.

if $dist > \delta$ **then**

$M = M + 1$

Καταχώρησε το νέο κέντρο $\boldsymbol{u}_M = \boldsymbol{x}_n$ στη λίστα με τα κέντρα, δηλαδή, $D = \{D, \boldsymbol{u}_M\}$, $\boldsymbol{\alpha}^T = \{\boldsymbol{\alpha}^T, \alpha_n\}$

else

Κράτησε το λεξικό D ως έχει και ενημέρωσε το συντελεστή a_l, δηλαδή $a_l = a_l + \mu e_n$.

end if

- **end for**

Καθώς, λοιπόν, καταφθάνει κάθε νέα πληροφορία $\boldsymbol{x}_n$, διαδεχόμενη την προηγούμενη, ο αλγόριθμος αποφασίζει αν πρόκειται για νέο κέντρο ή για περιττό σημείο. Συγκεκριμένα, αν η απόστασή του $\boldsymbol{x}_n$ από το λεξικό D_n, όπως αυτό είναι διαμορφωμένο μέχρι εκείνη τη στιγμή, είναι μεγαλύτερη ή ίση από το κβαντικό μέγεθος δ (κάτι που σημαίνει ότι το $\boldsymbol{x}_n$ δεν μπορεί να «κβαντοποιηθεί» σε κάποιο από τα σημεία που περιέχονται ήδη στο D_n) τότε το $\boldsymbol{x}_n$ ταξινομείται ως νέο κέντρο και καταχωρείται στο λεξικό, το οποίο γίνεται πλέον $D_{n+1} = \{D_n, \boldsymbol{x}_n\}$. Αλλιώς, το $\boldsymbol{x}_n$ αναγνωρίζεται ως «περιττό» σημείο και ο αλγόριθμος δεν επιβαρύνει άσκοπα το μέγεθος του λεξικού καταχωρώντας το ως επιπλέον κέντρο, εκμεταλλεύεται όμως την πληροφορία που προέκυψε ώστε να ανανεώσει το συντελεστή του πλησιέστερου στο συγκεκριμένο σημείο κέντρου, ας πούμε του $\boldsymbol{u}_l \in D_n$.

3.5 Δοκιμές των αλγορίθμων και συμπεράσματα

3.5.1 Nonlinear Channel Equalization

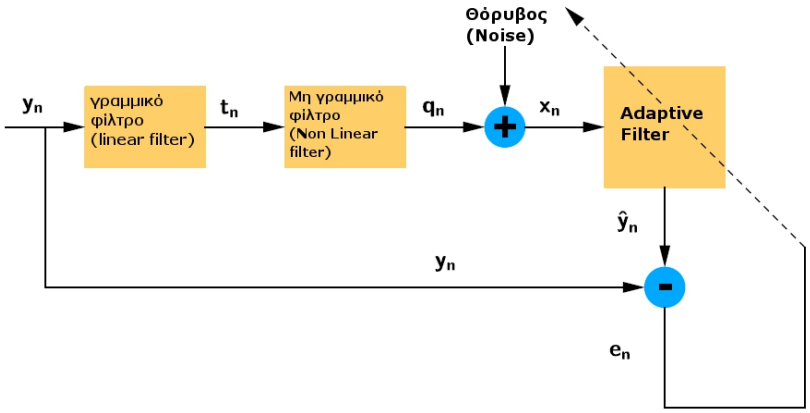
Προκειμένου να ελέγξουμε τις επιδόσεις του KLMS, θεωρούμε ένα τυπικό πείραμα αντιστάθμισης μη γραμμικού καναλιού (nonlinear channel equalization task). (Σχήμα) Το μη γραμμικό κανάλι αποτελείται από ένα γραμμικό φίλτρο (linear filter)

$$t_n = 0.8 \cdot y_n + 0.7 \cdot y_{n-1}$$

και μια memoryless nonlinearity

$$q_n = t_n + 0.8 \cdot t_n^2 + 0.7 \cdot t_n^3$$

Το σήμα, κατόπιν, επηρεάζεται από additive white Gaussian noise και παρατηρείται τελικά ως x_n. Το επίπεδο (level) του θορύβου τέθηκε ίσο με 15 dB.



Σχήμα 3.1: *Equalization Task*

Στόχος του channel equalization task είναι ο σχεδιασμός ενός αντίστροφου φίλτρου που να δρα επάνω στο output του καναλιού, x_n, και να αναπαράγει το γνήσιο input σήμα y_n όσο καλύτερα γίνεται.

Εφαρμόζουμε τον αλγόριθμο KLMS στο σύνολο παραδειγμάτων

$$((x_n, x_{n-1}, \ldots, x_{n-k+1}), y_{n-D})$$

όπου $k > 0$ το μήκος του αντισταθμιστή (equalizer's length) και D η χρονική καθυστέρηση του αντισταθμιστή (equalizer's time delay), η οποία εμφανίζεται σχεδόν σε κάθε equalization set up. Με άλλα λόγια, το αποτέλεσμα του αντισταθμιστή τη χρονική στιγμή x_n αντιστοιχεί στην εκτίμηση του y_{n-D}.

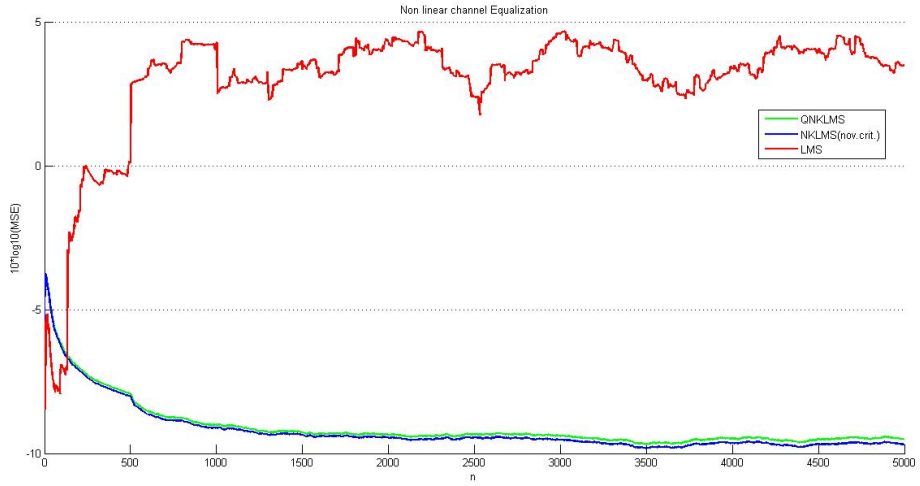
Το πείραμα πραγματοποιήθηκε σε 50 σύνολα από 5000 δείγματα input signals στο καθένα (Gaussian random variable with zero mean and unit variance) συγκρίνοντας τις επιδόσεις του standard LMS με εκείνες του KLMS, εφαρμόζοντας δύο διαφορετικές στρατηγικές αραίωσης, θεωρώντας όλους τους αλγορίθμους στις normalized εκδοχές τους.

Όσον αφορά τις δύο εκδοχές του KLMS, χρησιμοποιήθηκε και στις δύο η Gaussian kernel συνάρτηση (με $\sigma = 5$), και στη μία εκδοχή, NKLMS(nov.crit.), υιοθετήθηκε το Platt's novelty criterion ως τεχνική αραίωσης της λύσης ($\delta_1 = 0.04$, $\delta_2 = 0.04$), ενώ στην άλλη , QNKLMS, υιοθετήθηκε η τεχνική του κβαντισμού των δεδομένων (με μέγεθος κβάντισης $\delta = 0.8$).

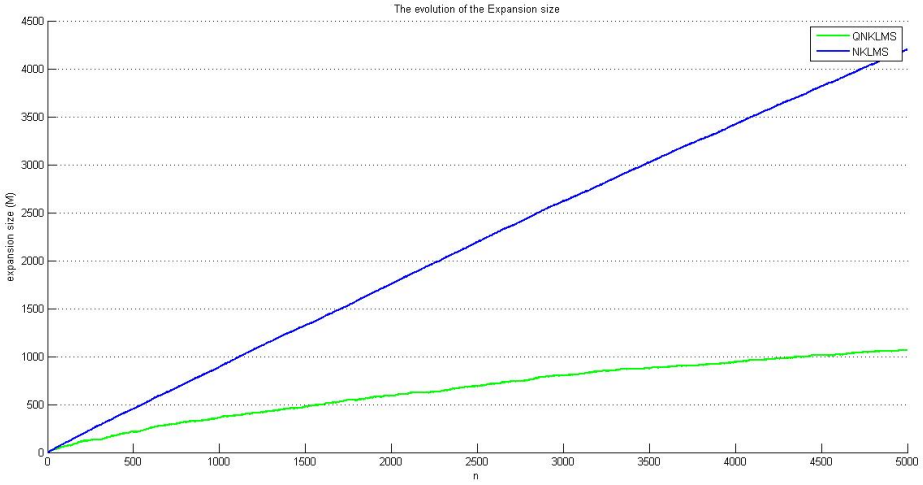
Το βήμα μάθησης (step update parameter) ρυθμίστηκε ώστε να αποδίδει τα καλύτερα δυνατά αποτελέσματα (in terms of the steady-state error rate). Η time delay D ρυθμίστηκε επίσης ως βέλτιστη.

Το Σχήμα 3.2, στην επόμενη σελίδα, παρουσιάζει τις learning curves των KLMS και standard LMS αλγορίθμων. Η υπεροχή του KLMS είναι εμφανής, κάτι που ήταν αναμενόμενο καθώς ο LMS αδυνατεί να διαχειριστεί τη μη γραμμικότητα.

Το Σχήμα 3.3 παρουσιάζει την αύξηση των όρων που εμφανίζονται στο ανάπτυγμα της λύσης, όσον αφορά στις δυο διαφορετικές μεθόδους αραίωσης του KLMS αλγορίθμου. Η οικονομία που επιτυγχάνουμε μέσω της κβάντισης, χωρίς ουσιαστικό κόστος στην αποτελεσματικότητα του αλγορίθμου, είναι αξιοσημείωτη.



Σχήμα 3.2: *Οι καμπύλες μάθησης (Learning curves) των normalized LMS και KLMS αλγορίθμων. Για τον KLMS, χρησιμοποιήθηκε και στις δύο περιπτώσεις η Gaussian kernel συνάρτηση ($\sigma = 5$). Στην NKLMS(nov.crit.), εκδοχή υιοθετήθηκε το Platt's novelty criterion ως τεχνική αραίωσης της λύσης ($\delta_1 = 0.04$, $\delta_2 = 0.04$), ενώ στην QNKLMS, υιοθετήθηκε η τεχνική του κβαντισμού των δεδομένων (με μέγεθος κβάντισης $\delta = 0.8$).*



Σχήμα 3.3: *Η εξέλιξη του πλήθους των όρων του αναπτύγματος της λύσης εφαρμόζοντας τις δυο διαφορετικές μεθόδους αραίωσης: Platt's novelty criterion ($\delta_1 = 0.04$, $\delta_2 = 0.04$) στον NKLMS(nov.crit.) , κβαντισμός των δεδομένων (με μέγεθος κβάντισης $\delta = 0.8$) στον QNKLMS.*

3.5.2 Chaotic Time Series Prediction

Στο συγκεκριμένο πείραμα, επιχειρούμε βραχυπρόθεσμη πρόβλεψη των όρων της χαοτικής χρονοσειράς Mackey-Glass, η οποία περιγράφεται από την ακόλουθη delayed διαφορική εξίσωση:

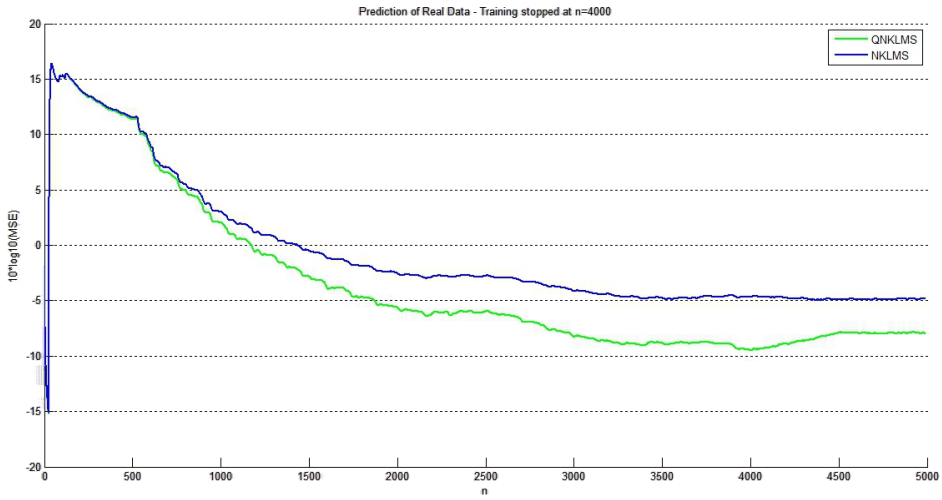
$$\frac{dx(t)}{dt} = \frac{a \cdot x(t-\tau)}{1 + x(t-\tau)^{10}} - b \cdot x(t) \tag{3.14}$$

με τιμή παραμέτρου $\tau = 30$ και $a = 10$, $b = 0.2$. Για την αριθμητική της επίλυση χρησιμοποιούμε την 4ης τάξεως Runge-Kutta μέθοδο.

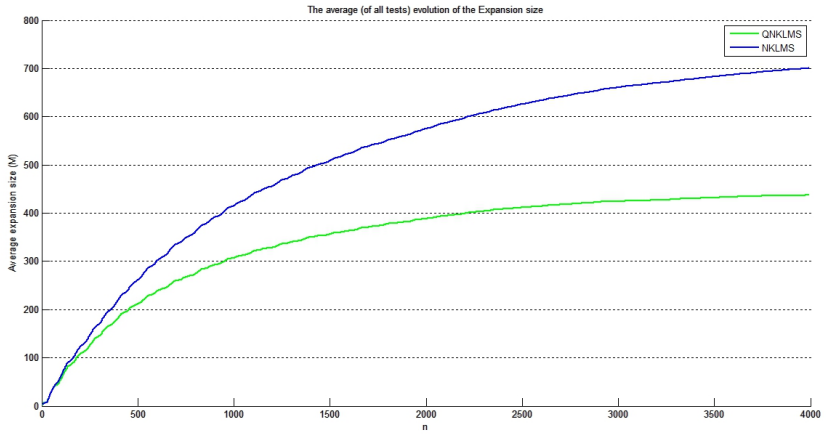
Συγκρίνουμε τις επιδόσεις του αλγορίθμου KLMS, στη normalized εκδοχή του, εφαρμόζοντας δύο τεχνικές αραίωσης : τον κβαντισμό (QNKLMS) (μέγεθος κβάντισης $\delta = 0.1$) και το Platt's Novelty Criterion (NKLMS) ($\delta_1 = 0.04$,$\delta_2 = 0.04$).

Ως kernel συνάρτηση επιλέχθηκε η Gaussian, με τιμή παραμέτρου $\sigma = 3$ και στις δύο περιπτώσεις, ενώ το βήμα μάθησης μ ρυθμίστηκε ώστε να δίνει τα καλύτερα αποτελέσματα.

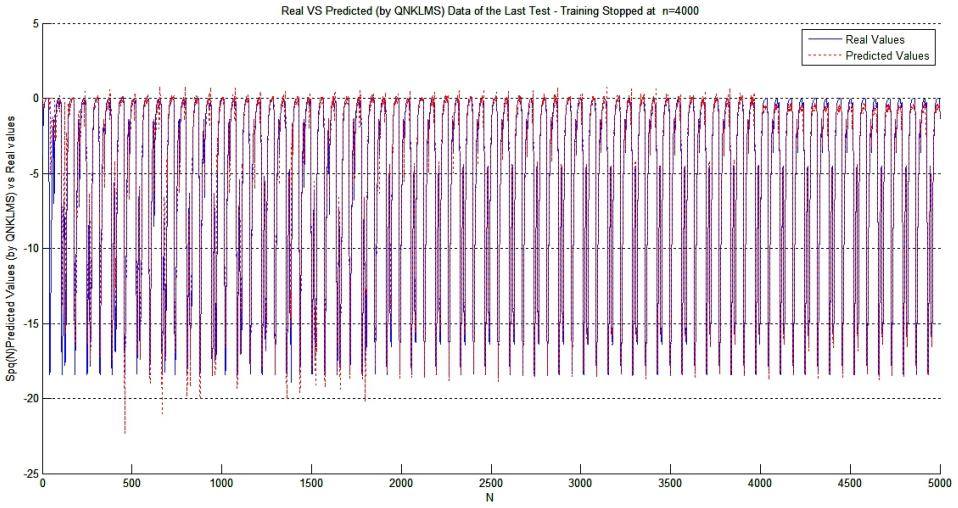
Η εκτίμηση πραγματοποιείται, συνολικά, για τους 5000 πρώτους όρους της χρονοσειράς. Οι πρώτες 4000 τιμές χρησιμοποιούνται ως σημεία εκπαίδευσης ενώ οι επόμενες 1000 ως δεδομένα δοκιμής. Στα δεδομένα έχει προστεθεί και Gaussian θόρυβος. Πραγματοποιούμε 50 Monte Carlo προσομοιώσεις, με διαφορετικές τιμές στην αρχική συνθήκη x_0 της χρονοσειράς Mackey και διαφορετικές τιμές θορύβου.



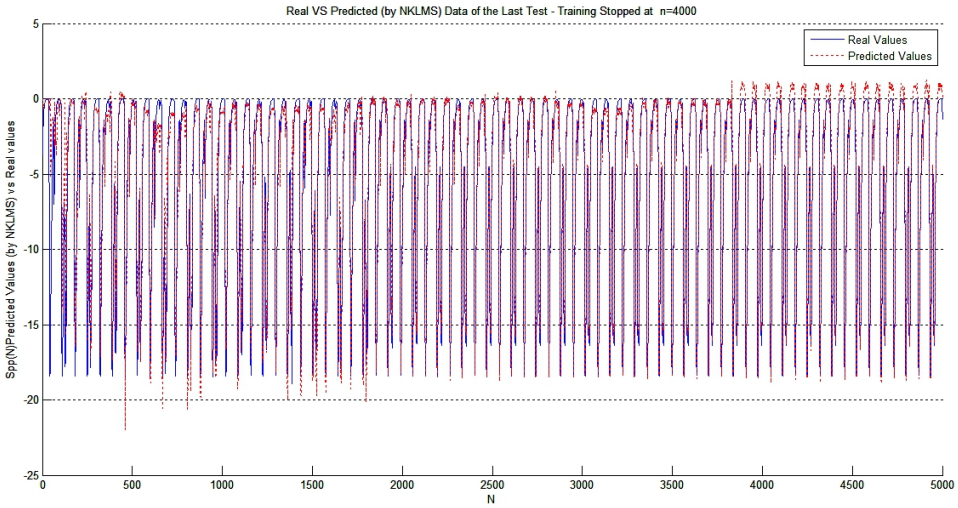
Σχήμα 3.4: *Οι καμπύλες μάθησης (learning curves) του normalized KLMS εφαρμόζοντας κβαντισμό (QNKLMS) (μέγεθος κβάντισης $\delta = 0.1$) και το Platt's Novelty Criterion (NKLMS) ($\delta_1 = 0.04$,$\delta_2 = 0.04$) για την αραίωση της λύσης. Το* ***Μέσο Σχετικό Σφάλμα (Average Relative error)*** *στην εκτίμηση των τιμών από τον QNKLMS για τα 50 τεστ είναι συνολικά: 4.8085% (Training: 4.7266%, Testing: 4.8392%), ενώ για τον NKLMS είναι συνολικά: 7.3752% (Training: 7.5749%, Testing: 6.8555%).*



Σχήμα 3.5: *Η εξέλιξη του πλήθους των όρων στο ανάπτυγμα της λύσης, μέχρι το πέρας της εκπαίδευσης, για τις δύο διαφορετικές μεθόδους αραίωσης: κβαντισμό (QNKLMS) (μέγεθος κβάντισης $\delta = 0.1$) και Platt's Novelty Criterion (NKLMS) ($\delta_1 = 0.04$,$\delta_2 = 0.04$). Με τον κβαντισμό επιτυγχάνεται σημαντική οικονομία, χωρίς κόστος στις επιδόσεις του αλγορίθμου.*



Σχήμα 3.6: *Σύγκριση των εκτιμώμενων (μέσω του QNKLMS) τιμών των όρων της χρονοσειράς Mackey, (κόκκινο χρώμα) με τις πραγματικές τιμές των όρων (μπλε χρώμα). Η αρχική συνθήκη για τη συγκεκριμένη Mackey τέθηκε* $x_0 = -4,56$.



Σχήμα 3.7: *Αντίστοιχη σύγκριση με εκτίμηση των τιμών μέσω του NKLMS αλγορίθμου.*

Βιβλιογραφία

[1] B. Scolkopf and A.Smola, "Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond", MIT Press, 2002.

[2] S. Theodoridis and K. Koutroumbas, "Pattern Recognition", 4th ed. Academic Press, Nov. 2008

[3] J. Shawe-Taylor and N. Cristianini, "Kernel Methods for Pattern Analysis", Cambridge University Press 2004.

[4] P. Bouboulis and M. Mavroforakis, "Reproducing Kernel Hilbert Spaces and Fractal Interpolation", Journal of Computational and Applied Mathematics, 2011.

[5] V.I. Paulsen, "An Introduction to the theory of Reproducing Kernel Hilbert Spaces", September 2009.

[6] I.Steinwart, D. Hush and C.Scovel, "An explicit descripion of the Reproducing Kernel Hilbert spaces of Gaussian RBF kernels", IEEE Trans. Info. Theory, vol. 52, no. 10, pp. 4635-4643, 2006.

[7] Press, Teukolsky, Vetterling, Flannery, "Numerical Recipes - The Art of Scientific Computing".

[8] P.Bouboulis, K.Slavakis, S.Theodoridis, "Adaptive Learning in Complex Reproducing Kernel Hilbert Spaces employing Writinger's subgradients".

[9] N.Aronszajn, "Theory of reproducing Kernels", transactions of the American Mathematical Society, vol. 68, pp. 337-404, 1950

[10] Weifeng Liu, *Student Member,IEEE*, Puskal P. Pokharel, *Student Member, IEEE*, and Jose C.Principe, *Fellow,IEEE*, "The Kernel Least-Mean-Square Algorithm".

[11] Karl R.Stronmberg, "Introduction to classical Real Analyssis", Wadsworth International Group, 1981.

[12] A.Caponnetto and A. Rakhlin, "Stability properties of empirical risk minimization over Donsker classes", J. Mach. Learn. Res., vol. 7, pp. 2565-2583, 2006.