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

Παπάζογλου Άννα

Εφαρμογές των Reproducing Kernel Hilbert Spaces στη Μηχανική Μάθηση και Υλοποίηση Αλγορίθμων

Επιβλέπουσα καθηγήτρια
Λ. Ευαγγελάτου - Δάλλα

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

ΑΘΗΝΑ 2014

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 06/10/2014 από Εξεταστική Επιτροπή

αποτελούμενη από τους :

Ονοματεπώνυμο	Βαθμίδα	Υπογραφή
1. Ευαγγελάτου-Δάλλα (επιβλέπων Καθηγητής)	Αν. Καθ.	
2. Θεοδωρίδης	Καθηγητής	
Δ. Τσαρπαλιάς	Καθηγητής	

Περιεχόμενα

Εισαγωγή **1**

1 Reproducing Kernel Hilbert Spaces (RKHS) **3**
1.1 Ιστορικά Στοιχεία των Kernels 3
1.2 Βασική Θεωρία Χώρων Hilbert 5
1.3 Η Γενική Θεωρία των RKHS 8
1.3.1 Βασικοί Ορισμοί και Θεωρήματα 8
1.3.2 Μιγαδοποίηση ενός RKHS πραγματικών συναρτήσεων 10
1.3.3 Ιδιότητες ενός RKHS 11
1.3.4 Χαρακτηρισμός των Reproducing Kernels 12
1.4 Παραδείγματα Kernel Συναρτήσεων 15
1.5 Χρήσιμες Προτάσεις που αφορούν Reproducing Kernels 22

2 Εφαρμογές Μηχανικής Μάθησης σε Γραμμικά Προβλήματα **27**
2.1 Εισαγωγή 27
2.2 Παραμετρική Μοντελοποίηση 28
2.3 Γραμμική Παλινδρόμηση και Κατηγοριοποίηση 30
2.4 Εκτίμηση Παραμέτρων 31
2.5 Ο Αλγόριθμος Least Mean Squares (LMS) 33
2.6 Ο Αλγόριθμος Recursive Least Squares (RLS) 36
2.6.1 Ο Αλγόριθμος Exponentially Weighted Recursive Least Squares (E-WRLS) 38

3 Εφαρμογές Μηχανικής Μάθησης σε Μη Γραμμικά Προβλήματα **41**
3.1 Εισαγωγή 41
3.2 Μη Γραμμική Παλινδρόμηση και Κατηγοριοποίηση 43
3.3 Το Kernel Τέχνασμα 44
3.4 Απεικονίζοντας ένα Μη Γραμμικό σε ένα Γραμμικό Πρόβλημα 46
3.5 Ο Αλγόριθμος Kernel Least Mean Squares (KLMS) 47
3.5.1 Αραίωση της Λύσης 49
3.5.2 Quantized Kernel Least Mean Squares (QKLMS) 51
3.6 Ο Αλγόριθμος Kernel Recursive Least Squares (KRLS) 52

4 Πειράματα και Εφαρμογές **57**
4.1 Ταυτοποίηση Καναλιού (Channel Identification) 57
4.1.1 Ταυτοποίηση Γραμμικού Καναλιού (Linear Channel Identification) . 58
4.1.2 Ταυτοποίηση Μη Γραμμικού Καναλιού (Nonlinear Channel Identification) 60
4.2 Αντιστάθμιση Καναλιού (Channel Equalization) 64
4.2.1 Αντιστάθμιση Γραμμικού Καναλιού (Linear Channel Equalization) . 64
4.2.2 Αντιστάθμιση Μη Γραμμικού Καναλιού (Nonlinear Channel Equalization) 65

Εισαγωγή

Ένας Reproducing Kernel Hilbert Space (RKHS) στη Συναρτησιακή Ανάλυση είναι μία πλούσια δομή. Συγκεκριμένα, είναι ένας χώρος Hilbert με στοιχεία μιγαδικές συναρτήσεις στον οποίο κάθε γραμμικό evaluation συναρτησοειδές είναι συνεχές. Το αντικείμενο αναπτύχθηκε αρχικά το 1950, ταυτόχρονα από τους Nachman Aronszajn (1907-1980) και Stefan Bergman (1895-1977). Η συσχέτιση των συγκεκριμένων χώρων με τις θετικά ορισμένες συναρτήσεις οδηγεί σε ένα ευρύ πεδίο εφαρμογών τους, όπως για παράδειγμα στη μιγαδική και την αρμονική ανάλυση, την κβαντική μηχανική και τη στατιστική. Στην παρούσα εργασία θα εξετάσουμε εφαρμογές στον κλάδο της τεχνητής νοημοσύνης που αποκαλείται Μηχανική Μάθηση.

Η Μηχανική Μάθηση (Machine Learning) είναι μια περιοχή της τεχνητής νοημοσύνης η οποία αφορά αλγορίθμους και μεθόδους που επιτρέπουν στους υπολογιστές να «μαθαίνουν», να αναπτύσσουν δηλαδή συμπεριφορές βασιζόμενοι σε εμπειρικά δεδομένα. Τα δεδομένα που τροφοδοτούνται στον υπολογιστή κατά το στάδιο της εκπαίδευσης (δεδομένα εκπαίδευσης), τα οποία θα μπορούσαν να έχουν καταγραφεί από κάποιους αισθητήρες ή να προέρχονται από βάσεις δεδομένων, αντικατοπτρίζουν τις σχέσεις μεταξύ των μεταβλητών που τα περιγράφουν. Στόχος είναι η διατύπωση μεθόδων αυτόματης αναγνώρισης πολύπλοκων προτύπων από την «εκπαιδευόμενη» μηχανή, που την καθιστούν ικανή να εξάγει αποφάσεις βάσει των δεδομένων που της παρέχονται. Η δυσκολία έγκειται στο γεγονός πως βασιζόμενη σε έναν περιορισμένο όγκο δεδομένων εκπαίδευσης πρέπει η μηχανή να μπορεί να διεξάγει χρήσιμα συμπεράσματα τα οποία να ανταποκρίνονται στο πλήθος δυνατών συμπεριφορών που ενδέχεται να εμφανίζουν όλα τα πιθανά δεδομένα.

Στις μεθόδους που βασίζονται σε kernels η έννοια των RKHS παίζει αποφασιστικό ρόλο. Με τη χρήση των kernels μία νέα τεχνική ήρθε στο προσκήνιο και έτσι οι μέθοδοι αυτές χρησιμοποιούνται σε ολοένα περισσότερες επιστημονικές περιοχές, ιδίως σε αυτές που απαιτούνται μη γραμμικά μοντέλα. Μη γραμμικά προβλήματα μετασχηματίζονται σε ισοδύναμα γραμμικά, καθώς επαναδιατυπώνονται σε ένα χώρο μεγαλύτερης (ενδεχομένως άπειρης) διάστασης, ο οποίος είναι ένας Reproducing Kernel Hilbert Space. Έτσι, αποφεύγονται υπολογιστικά προβλήματα και θέματα γενίκευσης που προέκυπταν με τις παραδοσιακές τεχνικές όταν η διάσταση του προβλήματος αυξανόταν. Παρόμοιες προσεγγίσεις έχουν χρησιμοποιηθεί στην Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis - PCA), στη Γραμμική Διακριτική Ανάλυση κατά Fisher (Fisher Linear Discriminant Analysis), στην Ομαδοποίηση (Clustering), στην Παλινδρόμηση (Regression), στην Επεξεργασία Εικόνας (Image Processing) και σε πολλά άλλα. Τελευταία, οι RKHS κερδίζουν έδαφος στο χώρο της Επεξεργασίας

Σήματος (Signal Processing).
Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του Προγράμματος Μεταπτυχιακών Σπουδών στα Εφαρμοσμένα Μαθηματικά του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών. Στο πρώτο Κεφάλαιο παρουσιάζονται ορισμένα ιστορικά στοιχεία που αφορούν τους Reproducing Kernel Hilbert Spaces, καθώς και η βασική θεωρία που αναπτύχθηκε γύρω από αυτούς. Τα επόμενα δύο Κεφάλαια αφορούν εφαρμογές Μηχανικής Μάθησης σε γραμμικά και μη γραμμικά προβλήματα. Το Κεφάλαιο 2 αποτελεί ένα εισαγωγικό Κεφάλαιο για το θέμα αυτό, καθώς μας ενδιαφέρει κυρίως να μελετήσουμε την αντιμετώπιση μη γραμμικών προβλημάτων, στα οποία γίνεται εκτενής αναφορά στο Κεφάλαιο 3. Στο τέταρτο και τελευταίο Κεφάλαιο παρουσιάζονται τα πειράματα στα οποία εφαρμόσαμε την παραπάνω θεωρία, καθώς και τα συμπεράσματά μας. Έχει γίνει προσπάθεια να μπορεί κανείς να περιηγηθεί στον κορμό της εργασίας κλιμακωτά, μεταβαίνοντας από τη θεωρία στην εφαρμογή της θεωρίας κατά το δυνατόν ομαλότερα.
Θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα Καθηγήτρια κυρία Λεώνη Ευαγγελάτου-Δάλλα για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον και διαφορετικό θέμα αλλά και για την πολύτιμη βοήθεια και καθοδήγησή της κατά τη διάρκεια της επεξεργασίας της εργασίας. Επίσης, ευχαριστώ τα υπόλοιπα μέλη της τριμελούς επιτροπής, τους κυρίους Θεοδωρίδη Σέργιο, Καθηγητή του Τμήματος Πληροφορικής, και Αθανάσιο Τσαρπαλιά, Καθηγητή του Τμήματος Μαθηματικών, καθώς η ολοκλήρωση της εργασίας αυτής θα ήταν αδύνατη χωρίς τη συμβολή τους. Επίσης, ευχαριστώ τον υποψήφιο διδάκτορα Λευτέρη Καστή (Lancaster University), που συνέβαλε με τα σχόλια και τις υποδείξεις του στη διαμόρφωση του πρώτου Κεφαλαίου. Τέλος, θα ήταν παράλειψη από πλευράς μου να μην ευχαριστήσω το διδάκτορα Παντελή Μπουμπούλη (Τμήμα Πληροφορικής) και τον υποψήφιο διδάκτορα Γιώργο Παπαγεωργίου (Τμήμα Πληροφορικής) για την υπομονή και το χρόνο που μου διέθεσαν, για τις επισημάνσεις τους, καθώς και για την πολύτιμη βοήθειά τους στον προγραμματισμό με Matlab.

Κεφάλαιο 1

Reproducing Kernel Hilbert Spaces (RKHS)

Στο κεφάλαιο αυτό περιγράφουμε ειδικούς χώρους Hilbert που ονομάζονται Reproducing Kernel Hilbert Spaces (RKHS), αφού κάνουμε μία ιστορική αναδρομή στην έννοια του πυρήνα (kernel). Παραθέτουμε, επίσης, αποδείξεις χρήσιμων συμπερασμάτων που αφορούν τους χώρους αυτούς καταλήγοντας στο Θεώρημα Moore. Το αποτέλεσμα αυτό μας εξασφαλίζει το μαθηματικό υπόβαθρο για τη διατύπωση του kernel τεχνάσματος, που αποτελεί το βασικό εργαλείο στις εφαρμογές του Κεφαλαίου 3.
Όσον αφορά το συμβολισμό, τα έντονα στοιχεία στις μαθηματικές εκφράσεις δηλώνουν διανύσματα. Επίσης, επειδή θα αναφερθούμε σε χώρους Hilbert άλλοτε επί του σώματος των πραγματικών αριθμών, $\mathbb{R}$, και άλλοτε επί του σώματος των μιγαδικών αριθμών, $\mathbb{C}$, θα χρησιμοποιούμε το σύμβολο $\mathbb{F}$ χωρίς περαιτέρω διευκρινίσεις.

1.1 Ιστορικά Στοιχεία των Kernels

Παραδείγματα πυρήνων (kernels) όπως αυτοί με τους οποίους θα ασχοληθούμε ήταν γνωστά από το 19ο αιώνα, καθώς σ΄ αυτή την κατηγορία ανήκουν όλες οι συναρτήσεις Green των αυτοσυζυγών συνήθων διαφορικών εξισώσεων (όπως επίσης και κάποιες συναρτήσεις Green μερικών διαφορικών εξισώσεων, συγκεκριμένα οι φραγμένες). Οι χαρακτηριστικές ιδιότητες αυτών των πυρήνων όμως, όπως τις καταλαβαίνουμε τώρα, έχουν τονιστεί και χρησιμοποιηθεί σε εφαρμογές από τις αρχές του 20ού αιώνα.
Υπήρξαν και υπάρχουν ακόμα δύο προσεγγίσεις σ΄ αυτό το θέμα. Προτού τις εξηγήσουμε, πρέπει να αναφέρουμε ότι ένας τέτοιος πυρήνας $K(x, y)$ είναι μία συνάρτηση δύο μεταβλητών, $K : X \times X \to \mathbb{C}$, που χαρακτηρίζεται από τη σχέση (1.1), την οποία εισήγαγε το 1909 ο J. Mercer [25]. Στον πυρήνα K αντιστοιχεί μία καλώς ορισμένη κλάση F συναρτήσεων ως προς την οποία η K ικανοποιεί την «reproducing» ιδιότητα (E. H. Moore [28]). Αντιστρόφως, σε μία κλάση συναρτήσεων F μπορεί να αντιστοιχεί ένας πυρήνας K με την «reproducing» ιδιότητα (N. Aronszajn [2]).
Όσοι ακολουθούν την πρώτη προσέγγιση θεωρούν δεδομένο έναν πυρήνα K και τον μελετούν

μόνο του ή τελικά τον εφαρμόζουν σε διάφορους τομείς (όπως τις ολοκληρωτικές εξισώσεις, τη θεωρία ομάδων, τη γεωμετρία Riemman). Η κλάση F που αντιστοιχεί στον K μπορεί να χρησιμοποιηθεί ως εργαλείο, αλλά παρουσιάζεται εκ των υστέρων, όπως στην εργασία του E. H. Moore [28] αλλά και σε πιο πρόσφατες [38], [20], [21], [22]. Αυτοί που ακολουθούν τη δεύτερη προσέγγιση ενδιαφέρονται πρωτίστως για την κλάση συναρτήσεων F και ο αντίστοιχος πυρήνας K χρησιμοποιείται ως εργαλείο για τη μελέτη των συναρτήσεων της κλάσης αυτής. Ένα από τα βασικά προβλήματα σ΄ αυτή την αναζήτηση είναι ο υπολογισμός του πυρήνα μιας δεδομένης κλάσης F.

Η πρώτη από αυτές τις προσεγγίσεις έχει τις ρίζες της στη θεωρία των ολοκληρωτικών εξισώσεων όπως αναπτύχθηκε από τον D. Hilbert. Οι πυρήνες που μελετήθηκαν τότε ήταν συνεχείς πυρήνες θετικά ορισμένων ολοκληρωτικών τελεστών. Τη θεωρία αυτή ανέπτυξε ο J. Mercer [25], [26] με το όνομα «θετικά ορισμένοι πυρήνες», η οποία αξιοποιήθηκε στη μελέτη ολοκληρωτικών εξισώσεων, κυρίως κατά τη δεύτερη δεκαετία του 20ού αιώνα. Ο Mercer χαρακτήρισε τους θετικά ορισμένους πυρήνες ως τους συνεχείς πυρήνες ολοκληρωτικών εξισώσεων, οι οποίοι πληρούν επιπλέον την ιδιότητα:

$$\sum_{i,j=1}^{n} K(y_i, y_j)\overline{\xi_i}\xi_j \geq 0, \quad \forall y_i \in X, \quad \forall \xi_i \in \mathbb{C}. \tag{1.1}$$

Να σημειώσουμε ότι ο Mercer χρησιμοποίησε μόνο πραγματικούς αριθμούς ξ_i, καθώς μελέτησε μόνο πραγματικούς πυρήνες K. Την προσέγγιση αυτή ακολούθησε ο E. H. Moore [27], [28], ο οποίος την περίοδο 1910-1930 χρησιμοποίησε αυτούς τους πυρήνες, ονομάζοντάς τους «θετικούς ερμιτιανούς πίνακες», ώστε να παρουσιάσει γενικευμένες μορφές ολοκληρωτικών εξισώσεων. Ο Moore όρισε πυρήνες $K(x, y)$ σε ένα τυχόν σύνολο X οι οποίοι απαίτησε να ικανοποιούν την ιδιότητα (1.1). Το θεώρημά του αποτελεί έναν από τους συνδετικούς κρίκους μεταξύ των δύο προσεγγίσεων. Συγκεκριμένα, απέδειξε ότι σε κάθε θετικό ερμιτιανό πίνακα αντιστοιχεί ένα σύνολο συναρτήσεων, το οποίο δέχεται δομή χώρου Hilbert και στον οποίο ο πυρήνας ικανοποιεί την «reproducing» ιδιότητα:

$$f(y) = (f(x), K(x, y)). \tag{1.2}$$

Επίσης στην ίδια προσέγγιση (παρότι μοιάζει να μην υπάρχει κάποια συσχέτιση) εντάσσεται η έννοια των «θετικά ορισμένων συναρτήσεων» που πρότεινε στις αρχές της δεκαετίας του 1940 ο S. Bochner [10]. Ο Bochner θεώρησε συνεχείς συναρτήσεις $\phi(x)$ μίας πραγματικής μεταβλητής x, τέτοιες ώστε οι πυρήνες $K(x, y) = \phi(x - y)$ ικανοποιούσαν τη σχέση (1.1). Εισήγαγε αυτές τις συναρτήσεις με σκοπό την εφαρμογή τους στη θεωρία των μετασχηματισμών Fourier. Η έννοια αυτή γενικεύτηκε αργότερα από τον A. Weil [38] και εφαρμόστηκε ([20], [21], [22]) στη μελέτη τοπολογικών αβελιανών ομάδων με το όνομα «θετικά ορισμένες συναρτήσεις» ή «συναρτήσεις θετικού τύπου». Αυτές οι συναρτήσεις βρήκαν επίσης εφαρμογή στη γεωμετρική και συναρτησιακή ανάλυση [33], [34], [29], [11].

Τη δεύτερη προσέγγιση εισήγαγε ο S. Zaremba [39], [40] την πρώτη δεκαετία του 20ού αιώνα δουλεύοντας σε προβλήματα συνοριακών τιμών για αρμονικές και διαρμονικές (biharmonic)

συναρτήσεις. Ο Zaremba ήταν ο πρώτος που παρουσίασε, σε μία συγκεκριμένη περίπτωση, τον πυρήνα που αντιστοιχούσε σε μία κλάση συναρτήσεων και που διατύπωσε την «reproducing» ιδιότητά του (1.2). Ωστόσο, δεν ανέπτυξε τη γενική θεωρία ούτε έδωσε κάποιο όνομα στους πυρήνες που εισήγαγε. Δεν έγινε πρόοδος προς αυτή την κατεύθυνση έως το 1922, όταν ο S. Bergman [4] παρουσίασε πυρήνες που αντιστοιχούσαν σε κλάσεις αρμονικών συναρτήσεων και αναλυτικών συναρτήσεων μίας ή περισσότερων μεταβλητών. Τις ονόμασε «συναρτήσεις kernel». Παρουσιάστηκαν ως πυρήνες ορθογώνιων συστημάτων σ΄ αυτές τις κλάσεις για μια επαρκή μετρική. Ο Bergman παρατήρησε την «reproducing» ιδιότητά τους, αλλά αυτή δε χρησιμοποιήθηκε ως η βασική χαρακτηριστική τους ιδιότητα όπως γίνεται σήμερα.
Τις δεκαετίες 1920-1940 περισσότερη έρευνα έγινε με πυρήνες που ονομάζονται πυρήνες Bergman, δηλαδή πυρήνες κλάσεων αναλυτικών συναρτήσεων f μίας ή πολλών μιγαδικών μεταβλητών, οι οποίες είναι αναλυτικές και τετραγωνικά ολοκληρώσιμες σε ένα χωρίο D:

$$\int_D |f|^2 \, \partial\tau < +\infty. \tag{1.3}$$

Πολλά σημαντικά αποτελέσματα προέκυψαν από τη χρήση αυτών των πυρήνων στη θεωρία συναρτήσεων μίας και πολλών μιγαδικών μεταβλητών, στις σύμμορφες απεικονίσεις σε απλά και πολλαπλά συνεκτικά χωρία, στις ψευδοσύμμορφες απεικονίσεις, καθώς και στη μελέτη αναλλοίωτων μετρικών Riemann.
Η αρχική ιδέα του Zaremba να εφαρμόσει τους πυρήνες στη λύση προβλημάτων συνοριακών τιμών εκπροσωπείται αυτά τα 20 χρόνια (1920-1940) αποκλειστικά από εργασίες του Bergman, καθώς η μαθηματική κοινότητα δεν της έδωσε την απαιτούμενη προσοχή. Όμως, μετά το Β΄ Παγκόσμιο Πόλεμο, η ιδέα αυτή ήρθε στο προσκήνιο με μια σειρά εργασιών των Bergman και Schiffer [5], [6], [7], [8]. Σε αυτές τις μελέτες, ο πυρήνας αποδείχθηκε ένα πανίσχυρο εργαλείο για την επίλυση προβλημάτων συνοριακών τιμών μερικών διαφορικών εξισώσεων ελλειπτικού τύπου. Χρησιμοποιώντας τον κλασικό τύπο μεταβολών του Hadamard, εντοπίστηκαν σχέσεις ανάμεσα στους πυρήνες που αντιστοιχούν σε κλάσεις λύσεων διαφόρων εξισώσεων και για διάφορα χωρία. Για μια μερική διαφορική εξίσωση, ο πυρήνας της κλάσης λύσεων σε ένα χωρίο αποδείχθηκε ότι είναι η διαφορά των αντίστοιχων συναρτήσεων Neumann και Green - μια τέτοια σχέση είχε παρατηρήσει ήδη ο Zaremba στην ειδική περίπτωση της διαρμονικής εξίσωσης. Παράλληλα με την αναζωπύρωση του μαθηματικού ενδιαφέροντος για την εφαρμογή των πυρήνων σε μερικές διαφορικές εξισώσεις, μελετάται η συσχέτιση μεταξύ αυτών των πυρήνων και των πυρήνων αναλυτικών συναρτήσεων του Bergman. Επίσης, η εφαρμογή των πυρήνων στις σύμμορφες απεικονίσεις σε πολλαπλά συνεκτικά χωρία παρουσίασε μεγάλη πρόοδο, καθώς αποδείχθηκε ότι όλες οι απαραίτητες απεικονίσεις εκφράζονται απλά μέσω του πυρήνα Bergman. Πιο πρόσφατα, εντοπίστηκε η σύνδεση μεταξύ του πυρήνα Bergman και του πυρήνα που παρουσίασε ο G. Szegö.

1.2 Βασική Θεωρία Χώρων Hilbert

Είναι απαραίτητο να υπενθυμίσουμε ορισμένες βασικές έννοιες, ώστε στη συνέχεια να ορίσουμε τους Reproducing Kernel Hilbert χώρους. Παραθέτουμε επίσης το Θεώρημα Αναπαράστασης του Riesz, το οποίο διαδραματίζει πολύ σημαντικό ρόλο στη θεωρία των RKHS.

Ορισμός 1.2.1. *Έστω V ένας χώρος με νόρμα επί ενός σώματος $\mathbb{F}$ και Z ένας κλειστός γραμμικός υπόχωρος του V. Ο* **χώρος πηλίκο** *του V με τον Z λέγεται ο χώρος $(V/Z, \|\cdot\|)$ όπου*

$$V/Z := \{v + Z, v \in V\}, \text{ όπου } v + Z := \{v + z, z \in Z\}.$$

και

$$\|v + Z\| := inf\{\|v + z\|_V, z \in Z\}.$$

Ορισμός 1.2.2. *Έστω V, W δύο χώροι με νόρμα επί ενός σώματος $\mathbb{F}$ και μια γραμμική απεικόνιση $T : V \to W$. Η απεικόνιση T λέγεται* **φραγμένος τελεστής** *αν υπάρχει πραγματικός αριθμός $M \geq 0$, τέτοιος ώστε*

$$\|Tv\|_W \leq M \|v\|_V \quad \forall v \in V.$$

Θεώρημα 1.2.3. *Έστω V, W δύο χώροι με νόρμα επί ενός σώματος $\mathbb{F}$ και μια γραμμική απεικόνιση $T : V \to W$. Τα ακόλουθα είναι ισοδύναμα:*

(i) *Ο T είναι φραγμένος τελεστής.*

(ii) *Ο T είναι συνεχής τελεστής.*

(iii) *Ο T είναι συνεχής τελεστής στο 0.*

Ορισμός 1.2.4. *Η* **νόρμα** *ενός* **φραγμένου τελεστή** *$T : V \to W$ ορίζεται ως:*

$$\|T\| = inf\{M \geq 0 : \|Tv\|_W \leq M \|v\|_V \quad \forall v \in V\}.$$

Ορισμός 1.2.5. *Έστω $\mathcal{H}$ ένας χώρος Hilbert και ένα υποσύνολο $S \subset \mathcal{H}$. Το* **ορθογώνιο συμπλήρωμα του S**, *το οποίο συμβολίζουμε $S^\perp$, είναι το σύνολο των κάθετων (ή ορθογώνιων) προς το S στοιχείων:*

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

Παρατήρηση 1.2.6. *Λόγω της συνέχειας του εσωτερικού γινομένου στον $\mathcal{H}$, το ορθογώνιο συμπλήρωμα είναι πάντα κλειστός υπόχωρος. Πράγματι, εάν $\{s_n\}_{n\in\mathbb{N}} \in S^\perp$ και $s_n \to s \in \mathcal{H}$, τότε για $x \in S$ ισχύει $\langle s, x\rangle = \left\langle \lim_{n\to\infty} s_n, x \right\rangle = \lim_{n\to\infty} \langle s_n, x\rangle = \lim_{n\to\infty} 0 = 0$. Άρα $s \in S^\perp$.*

Ορισμός 1.2.7. *Έστω χώρος Hilbert $\mathcal{H}$ και M κλειστός υπόχωρος του $\mathcal{H}$. Ο M λέγεται* **συμπληρωματικός υπόχωρος** *στον $\mathcal{H}$ αν υπάρχει κλειστός υπόχωρος N του $\mathcal{H}$, τέτοιος ώστε:*

(i) $\mathcal{H} = M + N$, *όπου* $M + N = \{m + n : m \in M, n \in N\}$

(ii) $M \cap N = \{0\}$

Τότε ο $\mathcal{H}$ λέμε ότι αποτελεί το **ευθύ άθροισμα** *των M και N.*
Συμβολίζουμε $\mathcal{H} = M \oplus N$ και, αν $h \in \mathcal{H}$, γράφουμε $h = m \oplus n$, $m \in M, n \in N$.

Πόρισμα 1.2.8. *Έστω M κλειστός υπόχωρος ενός χώρου Hilbert $\mathcal{H}$. Τότε $\mathcal{H} = M \oplus M^{\perp}$.*

Ορισμός 1.2.9. *Έστω $\mathcal{H}$ ένας χώρος Hilbert και M κλειστός υπόχωρός του. Έστω $x \in \mathcal{H}$, με $x = x_1 \oplus x_2,\ x_1 \in M, x_2 \in M^{\perp}$. Η απεικόνιση $x \mapsto x_1$ ορίζει ένα γραμμικό τελεστή P τέτοιο ώστε $P : \mathcal{H} \to M,\ Px = x_1$. Ο τελεστής P ονομάζεται **ορθογώνια προβολή** του $\mathcal{H}$ πάνω στον M.*

Παρατήρηση 1.2.10. *Ισχύει $Py = y$ για κάθε $y \in M$. Επίσης, ο τελεστής P είναι φραγμένος και $\|P\| = 1$, με εξαίρεση την περίπτωση $M = \{0\}$, όπου $P = 0$.*

Ορισμός 1.2.11. *Έστω $A = (a_{ij})$ ένας $n \times n$ πίνακας, όπου $a_{ij} \in \mathbb{C}, i, j \in \{1, 2, \ldots, n\}$. Ο A ονομάζεται **ερμιτιανός**, εάν ισχύει $A = A^*$, όπου $A^* = (\overline{A})^T = \overline{(A^T)}$ είναι ο ανάστροφος μιγαδικός συζυγής πίνακας του A.*

Ορίζουμε παρακάτω το *θετικό* και τον *αυστηρά θετικό* πίνακα. Κάποιοι προτιμούν να ονομάζουν τους πρώτους *θετικά ημιορισμένους* ή *μη αρνητικούς* και τους δεύτερους *θετικούς*. Εδώ προτιμήσαμε να κρατήσουμε την ορολογία που χρησιμοποιείται πιο συχνά στη Θεωρία Τελεστών.

Ορισμός 1.2.12. *Έστω $A = (a_{ij})$ ένας $n \times n$ ερμιτιανός πίνακας. Ο A ονομάζεται **θετικός πίνακας** εάν για κάθε $b_1, b_2, \ldots, b_n \in \mathbb{C}$ ισχύει $\sum_{i,j=1}^{n} \overline{b_i} b_j a_{ij} \geq 0$, όπου $\overline{b_i}$ είναι ο μιγαδικός συζυγής του b_i.*
Συμβολίζουμε $A \geq 0$.

Παρατήρηση 1.2.13. *Παρατηρούμε ότι για το θετικό πίνακα $A = (a_{ij})$ επιτρέπεται η ικανοποίηση της συνθήκης $\sum_{i,j=1}^{n} \overline{b_i} b_j a_{ij} = 0$ από διανύσματα $\boldsymbol{b} = (b_1, b_2, \ldots, b_n) \neq \mathbf{0} \in \mathbb{C}^n$.*

Αντίστοιχα, ορίζεται ο αυστηρά θετικός πίνακας.

Ορισμός 1.2.14. *Έστω $A = (a_{ij})$ ένας $n \times n$ ερμιτιανός πίνακας. Ο A ονομάζεται **αυστηρά θετικός πίνακας** εάν για κάθε $b_1, b_2, \ldots, b_n \in \mathbb{C}$ με $(b_1, b_2, \ldots, b_n) \neq \mathbf{0} \in \mathbb{C}^n$ ισχύει $\sum_{i,j=1}^{n} \overline{b_i} b_j a_{ij} > 0$.*
Συμβολίζουμε $A > 0$.

Παρατηρήσεις 1.2.15. (i) *Παρατηρήστε ότι όταν ο A είναι ερμιτιανός πίνακας, ισχύει $\sum_{i,j=1}^{n} \overline{b_i} b_j a_{ij} \in \mathbb{R}$.*

(ii) *Ένας ισοδύναμος ορισμός του θετικού πίνακα δίνεται παρακάτω.*
Ισχύει $A \geq 0$ αν και μόνο αν $\langle A\boldsymbol{b}, \boldsymbol{b} \rangle \geq 0$ για κάθε $\boldsymbol{b} = (b_1, b_2, \ldots, b_n) \in \mathbb{C}^n$, όπου με $\langle \cdot, \cdot \rangle$ συμβολίζεται το σύνηθες εσωτερικό γινόμενο.

(iii) *Αν A είναι ένας ερμιτιανός πίνακας, αποδεικνύεται ότι $A \geq 0$ αν και μόνο αν για κάθε ιδιοτιμή του, $\lambda \in \mathbb{R}$, ισχύει $\lambda \geq 0$.*
Αντίστοιχα, $A > 0$ αν και μόνο αν για κάθε ιδιοτιμή του A, $\lambda \in \mathbb{R}$, ισχύει $\lambda > 0$. Ισοδύναμα, $A > 0$ αν και μόνο αν ισχύει $A \geq 0$ και ο A είναι αντιστρέψιμος, καθώς αν ο A είναι μη αντιστρέψιμος έχει ιδιοτιμή το $\lambda = 0$.

Ορισμός 1.2.16. *Έστω ένα σύνολο X, μια συνάρτηση δύο μεταβλητών $K : X \times X \to \mathbb{C}$ και ένα υποσύνολο $\{x_1, x_2, \ldots, x_n\} \subseteq X$. Ο τετραγωνικός $n \times n$ πίνακας με στοιχεία $K(x_i, x_j)$ για $i, j = 1, 2, \ldots, n$, ονομάζεται **Gram πίνακας** (ή **kernel πίνακας**) της συνάρτησης K ως προς τα $\{x_1, x_2, \ldots, x_n\}$.*

Ορισμός 1.2.17. *Έστω σύνολο X και μια συνάρτηση δύο μεταβλητών $K : X \times X \to \mathbb{C}$. Η K ονομάζεται **θετικά ορισμένη συνάρτηση** (ή **kernel συνάρτηση**) αν για κάθε $n \in \mathbb{N}$ και για κάθε επιλογή n στοιχείων $x_1, x_2, \ldots, x_n \in X$ με $x_i \neq x_j$, $i \neq j$, ο Gram πίνακας της K ως προς τα $\{x_1, x_2, \ldots, x_n\}$ είναι θετικός.*
Συμβολίζουμε $K \geq 0$.

Θεώρημα 1.2.18 (Αναπαράστασης του Riesz). *Έστω $\mathcal{H}$ ένας χώρος Hilbert επί σώματος $\mathbb{F}$ και $T : \mathcal{H} \to \mathbb{F}$ ένας φραγμένος γραμμικός τελεστής. Τότε υπάρχει μοναδικό στοιχείο $h_0 \in \mathcal{H}$ τέτοιο ώστε $Th = \langle h, h_0 \rangle_{\mathcal{H}}$ για κάθε $h \in \mathcal{H}$. Επίσης, ισχύει $\|T\| = \|h_0\|_{\mathcal{H}}$.*

1.3 Η Γενική Θεωρία των RKHS

1.3.1 Βασικοί Ορισμοί και Θεωρήματα

Στην ενότητα αυτή ορίζουμε τους reproducing kernel Hilbert χώρους, καθώς και τις reproducing kernel συναρτήσεις.

Ορισμός 1.3.1. *Ο $\mathcal{H}$ είναι ένας **reproducing kernel Hilbert space (RKHS)** του X επί του $\mathbb{F}$, όταν ισχύουν τα εξής:*

(i) *ο $\mathcal{H}$ είναι υπόχωρος του $\mathcal{F}(X, \mathbb{F})$, όπου $\mathcal{F}(X, \mathbb{F})$ είναι ο διανυσματικός χώρος όλων των συναρτήσεων $f : X \to \mathbb{F}$, όπου οι γραμμικές πράξεις ορίζονται κατά σημείο.*

(ii) *ο $\mathcal{H}$ είναι εφοδιασμένος με ένα εσωτερικό γινόμενο $\langle \cdot, \cdot \rangle$ που τον μετατρέπει σε χώρο Hilbert, δηλαδή σε πλήρη μετρικό χώρο ως προς τη μετρική που ορίζει η επαγόμενη νόρμα $\|\cdot\|$*

και

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

Έστω $\mathcal{H}$ ένας RKHS ενός συνόλου X. Το Θεώρημα Αναπαράστασης του Riesz (1.2.18) μας εξασφαλίζει ότι για κάθε $y \in X$ υπάρχει ***μοναδικό*** στοιχείο $k_y \in \mathcal{H}$ τέτοιο ώστε να ισχύει $E_y(f) = f(y) = \langle f, k_y \rangle_{\mathcal{H}}$ για κάθε $f \in \mathcal{H}$.

Ορισμός 1.3.2. *Έστω $\mathcal{H}$ ένας RKHS ενός συνόλου X. Έστω $y \in X$. Ορίζουμε τη συνάρτηση*

$$k_y \in \mathcal{H},\ k_y : X \to \mathbb{F}, \quad \text{ώστε} \quad E_y(f) = f(y) = \langle f, k_y \rangle_{\mathcal{H}},\ f \in \mathcal{H}.$$

Η k_y ονομάζεται η ***reproducing kernel συνάρτηση για το σημείο*** *y. Η συνάρτηση $K : X \times X \to \mathbb{F}$ που ορίζεται ως*

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

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

Συμβολισμός
Γράφουμε $k_y(\cdot) = K(\cdot, y), y \in X$.

Επομένως, η K είναι η **μοναδική** συνάρτηση η οποία ικανοποιεί την ιδιότητα

$$\text{για } y \in X \quad f(y) = \langle f, K(\cdot, y) \rangle_{\mathcal{H}} \quad \forall f \in \mathcal{H}.$$

Παρατηρήσεις 1.3.3. (i) *Έστω $y \in X$. Η ιδιότητα*

$$f(y) = \langle f, k_y \rangle_{\mathcal{H}} \quad \forall f \in \mathcal{H} \tag{1.4}$$

την οποία ικανοποιεί η συνάρτηση k_y λέγεται ***reproducing ιδιότητα (reproducing property)****.*

(ii) *Παρατηρούμε ότι ισχύουν τα εξής:*

$$K(x, y) = k_y(x) = \langle k_y, k_x \rangle_{\mathcal{H}} \tag{1.5}$$

$$\|E_y\|^2 = \|k_y\|^2_{\mathcal{H}} = \langle k_y, k_y \rangle_{\mathcal{H}} = K(y, y) \tag{1.6}$$

Η σχέση (1.5) διαδραματίζει σημαντικό ρόλο στις εφαρμογές της θεωρίας των reproducing kernel Hilbert χώρων σε προβλήματα Μηχανικής Μάθησης. Επιτρέπει την αντικατάσταση του δύσκολου υπολογισμού ενός εσωτερικού γινομένου με τον εύκολο υπολογισμό της τιμής μιας συνάρτησης δύο μεταβλητών. Μένει, ωστόσο, να εξασφαλιστεί η ύπαρξη κατάλληλων τέτοιων χώρων και συναρτήσεων για ένα οποιοδήποτε σύνολο X.

Πρόταση 1.3.4. *Έστω $\mathcal{H}$ ένας RKHS συνόλου X με reproducing kernel συνάρτηση K. Τότε ισχύει $K(x, y) = \overline{K(y, x)}, \quad \forall x, y \in X$.*
Εάν η reproducing kernel συνάρτηση του $\mathcal{H}$ είναι πραγματική συνάρτηση, τότε είναι συμμετρική, δηλαδή $K(x, y) = K(y, x)$.

Απόδειξη. Έστω $x, y \in X$. Παρατηρούμε ότι

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

και

$$\overline{K(y,x)} = \overline{k_x(y)} = \overline{\langle k_x, k_y \rangle}_{\mathcal{H}}.$$

Όμως, ισχύει $\langle k_y, k_x \rangle_{\mathcal{H}} = \overline{\langle k_x, k_y \rangle}_{\mathcal{H}}$ και έτσι έχουμε $K(x,y) = \overline{K(y,x)}$.
Εάν η K είναι πραγματική συνάρτηση, τότε ισχύει $\langle k_y, k_x \rangle_{\mathcal{H}} = \langle k_x, k_y \rangle_{\mathcal{H}}$ και το ζητούμενο έπεται όμοια. □

Επίσης, ισχύει η ανισότητα Cauchy-Schwarz σε χώρους RKHS για την reproducing kernel συνάρτηση του χώρου.

Πρόταση 1.3.5 (**Ανισότητα Cauchy-Schwarz**). *Έστω $\mathcal{H}$ ένας RKHS συνόλου X με reproducing kernel συνάρτηση K. Τότε ισχύει*

$$\|K(x,y)\|_{\mathcal{H}}^2 \leq K(x,x) \cdot K(y,y).$$

Απόδειξη. Η απόδειξη είναι άμεση, καθώς $K(x,y)$ είναι το εσωτερικό γινόμενο $\langle k_y, k_x \rangle_{\mathcal{H}}$ στο χώρο $\mathcal{H}$. □

1.3.2 Μιγαδοποίηση ενός RKHS πραγματικών συναρτήσεων

Στην ενότητα αυτή φαίνεται ότι κάθε RKHS πραγματικών συναρτήσεων μπορεί να μιγαδοποιηθεί.

Ας θεωρήσουμε τον $\mathcal{H}$ ως έναν RKHS πραγματικών συναρτήσεων ενός συνόλου X με reproducing kernel τη συνάρτηση $K(x,y), x, y \in X$. Ας θεωρήσουμε επίσης το διανυσματικό χώρο $\mathcal{W}$ μιγαδικών συναρτήσεων του X, $\mathcal{W} = \{f_1 + if_2 : f_1, f_2 \in \mathcal{H}\}$. Ο $\mathcal{W}$ μετατρέπεται σε χώρο Hilbert μέσω της πολικής ταυτότητας:

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

η οποία, όπως εύκολα διαπιστώνει κανείς, ορίζει μιγαδικό εσωτερικό γινόμενο στον $\mathcal{W}$ και η αντίστοιχη νόρμα του είναι

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

Είναι προφανές ότι ο $(\mathcal{W}, \langle \cdot, \cdot \rangle_{\mathcal{W}})$ είναι πλήρης, καθώς εάν $\{h_n\}_{n \in \mathbb{N}} \in \mathcal{W}$ με $h_n = f_n + ig_n$, όπου $f_n, g_n, n \in \mathbb{N}$, είναι πραγματικές συναρτήσεις, και $h_n \to h \in \mathcal{F}(X, \mathbb{C})$, τότε ισχύουν $f_n \to \Re(h) \in \mathcal{H}$ και $g_n \to \Im(h) \in \mathcal{H}$. Συνεπώς $h \in \mathcal{W}$. Επιπλέον, δοθέντος $y \in X$ έχουμε

$$\tilde{E}_y(f_1 + f_2) := E_y(f_1) + iE_y(f_2) = f_1(y) + if_2(y) = \langle f_1 + if_2, k_y \rangle_{\mathcal{W}}.$$

Έτσι, καταλήγουμε ότι ο $\mathcal{W}$ εφοδιασμένος με το παραπάνω εσωτερικό γινόμενο αποτελεί RKHS μιγαδικών συναρτήσεων του X με reproducing kernel τη συνάρτηση $K(x,y), x, y \in X$. Ο χώρος αυτός καλείται **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 του Κεφαλαίου θα θεωρούμε τους RKHS στους οποίους αναφερόμαστε ως μιγαδικούς.

1.3.3 Ιδιότητες ενός RKHS

Έστω σύνολο X και $\mathcal{H}$ ένας RKHS του X με reproducing kernel συνάρτηση την K. Θα αποδείξουμε ότι η K καθορίζει πλήρως το χώρο $\mathcal{H}$.

Σύμφωνα με την Πρόταση που ακολουθεί, οποιοσδήποτε RKHS, $\mathcal{H}$, μπορεί να παραχθεί από την αντίστοιχη reproducing kernel συνάρτηση, K.

Πρόταση 1.3.6. *Έστω $\mathcal{H}$ ένας RKHS συνόλου X με reproducing kernel συνάρτηση K. Τότε η γραμμική θήκη των συναρτήσεων $k_y(\cdot) = K(\cdot, y)$, $y \in X$, είναι πυκνός υπόχωρος του $\mathcal{H}$ ως προς τη νόρμα που επάγει το εσωτερικό γινόμενο στον $\mathcal{H}$, δηλαδή*

$$\mathcal{H} = \overline{span\{k_y(\cdot), y \in X\}}.$$

Απόδειξη. Έστω $\mathcal{M} = span\{k_y(\cdot), y \in X\}$. Θα αποδειχθεί ότι $\mathcal{H} = \overline{\mathcal{M}}$. Σύμφωνα με το Πόρισμα (1.2.8), γράφουμε $\mathcal{H} = \overline{\mathcal{M}} \oplus \mathcal{M}^{\perp}$. Αρκεί να αποδειχθεί ότι $\mathcal{M}^{\perp} = \{0\}$.
Έστω $f \in \mathcal{M}^{\perp}$. Τότε ισχύει $\langle f, k_y \rangle = 0$ για κάθε $y \in X$ και ισοδύναμα $f(y) = 0$ για κάθε $y \in X$. Η σχέση αυτή ισχύει αν και μόνο αν $f \equiv 0$. Επομένως, $\mathcal{M}^{\perp} = \{0\}$ και $\mathcal{H} = \overline{\mathcal{M}}$. □

Στην επόμενη Πρόταση αποδεικνύεται ότι σε έναν RKHS η σύγκλιση ως προς νόρμα συνεπάγεται σύγκλιση κατά σημείο.

Πρόταση 1.3.7. *Έστω $\mathcal{H} \subseteq \mathcal{F}(X, \mathbb{F})$ ένας χώρος Hilbert. Τα ακόλουθα είναι ισοδύναμα.*

(i) *Ο $\mathcal{H}$ είναι RKHS του συνόλου X.*

(ii) *Εάν $\{f_n\}_{n\in\mathbb{N}}$ ακολουθία του $\mathcal{H}$ με $\lim_{n\to\infty} \|f_n - f\|_{\mathcal{H}} = 0$, τότε $\lim_{n\to\infty} f_n(x) = f(x)$ για κάθε $x \in X$.*

Απόδειξη. Έστω ακολουθία $\{f_n\}_{n\in\mathbb{N}}$ σε ένα χώρο RKHS, $\mathcal{H}$, συνόλου X, τέτοια ώστε $\|f_n - f\|_{\mathcal{H}} \overset{n\to\infty}{\longrightarrow} 0$. Έστω $x \in X$. Ισχύει

$$|f_n(x) - f(x)| \overset{(1.4)}{=} |\langle f_n, k_x \rangle_{\mathcal{H}} - \langle f, k_x \rangle_{\mathcal{H}}| = |\langle f_n - f, k_x \rangle_{\mathcal{H}}| \leq \|f_n - f\|_{\mathcal{H}} \|k_x\|_{\mathcal{H}} \overset{n\to\infty}{\longrightarrow} 0.$$

Επομένως, $f_n(x) \overset{n\to\infty}{\longrightarrow} f(x)$ για κάθε $x \in X$.
Αντιστρόφως:
Έστω μια ακολουθία $\{f_n\}_{n\in\mathbb{N}}$ σε ένα χώρο Hilbert, $\mathcal{H}$, τέτοια ώστε $\lim_{n\to\infty} \|f_n - f\| = 0$. Θεωρούμε το συναρτησοειδές

$$E_y : \mathcal{H} \to \mathbb{R}, \quad E_y(f) = f(y), \quad \text{για κάποιο} \quad y \in X.$$

Θα αποδειχθεί ότι το E_y είναι συνεχές για κάθε $y \in X$. Ισχύει

$$|E_y(f_n) - E_y(f)| = |f_n(y) - f(y)| \overset{n\to\infty}{\longrightarrow} 0,$$

αφού $\lim_{n\to\infty} f_n(x) = f(x)$ για κάθε $x \in X$ από την υπόθεση (ii). Επομένως, έχουμε ότι $\lim_{n\to\infty} E_y(f_n) = E_y(f)$ για κάθε $y \in X$ και οποιαδήποτε συγκλίνουσα ακολουθία $\{f_n\}_{n\in\mathbb{N}}$ του $\mathcal{H}$, δηλαδή το E_y είναι συνεχές για κάθε $y \in X$. Επομένως, ο $\mathcal{H}$ είναι ένας RKHS. □

Είναι φανερό, από την Πρόταση που ακολουθεί, ότι δύο διαφορετικοί RKHS δεν μπορεί να έχουν την ίδια reproducing kernel συνάρτηση.

Πρόταση 1.3.8. *Έστω $\mathcal{H}_i,\ i=1,2,$ δύο RKHS συνόλου X με αντίστοιχες συναρτήσεις kernel $K_i,\ i=1,2.$ 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 $f\in\mathcal{H}_1,$ όπου $\|\cdot\|_i$ η νόρμα του χώρου $\mathcal{H}_i,\ i=1,2,$ αντίστοιχα.*

Απόδειξη. Αρχικά παρατηρούμε ότι

$$K_1(x,y)=K_2(x,y),\quad \forall x,y\in X\Leftrightarrow$$
$$k_{y,1}(x)=k_{y,2}(x)\quad \forall x,y\in X\Leftrightarrow$$
$$k_{y,1}=k_{y,2}\quad \forall y\in X.$$

Επομένως, από την Πρόταση (1.3.6) έπεται $\mathcal{H}_1=\mathcal{H}_2$. Για να αποδείξουμε ότι $\|f\|_1=\|f\|_2$ χρησιμοποιούμε την ίδια Πρόταση και το γεγονός ότι η νόρμα είναι συνεχής. Υποθέτουμε ότι $f=\sum_i a_i k_{x_i},\ a_i\in\mathbb{F}, x_i\in X.$ Τότε ισχύει

$$\|f\|_1^2=\langle f,f\rangle_1=\sum_{i,j}a_i\overline{a_j}\left\langle k_{x_i},k_{x_j}\right\rangle_1=\sum_{i,j}a_i\overline{a_j}K_1(x_j,x_i)=\sum_{i,j}a_i\overline{a_j}K_2(x_j,x_i)=$$
$$\sum_{i,j}a_i\overline{a_j}\left\langle k_{x_i},k_{x_j}\right\rangle_2=\langle f,f\rangle_2=\|f\|_2^2$$

και άρα $\|f\|_1=\|f\|_2$. □

1.3.4 Χαρακτηρισμός των Reproducing Kernels

Θα εξετάσουμε ικανές και αναγκαίες συνθήκες ώστε μια συνάρτηση να αποτελεί reproducing kernel συνάρτηση για κάποιον RKHS.

Η ακόλουθη Πρόταση αποτελεί τον πρώτο συνδετικό κρίκο μεταξύ των θετικά ορισμένων συναρτήσεων και των reproducing kernel συναρτήσεων.

Πρόταση 1.3.9. *Έστω σύνολο X και $\mathcal{H}$ ένας RKHS του X με reproducing kernel τη συνάρτηση K. Τότε η K είναι θετικά ορισμένη συνάρτηση.*

Απόδειξη. Έστω $n\in\mathbb{N}$ και στοιχεία $x_1,x_2,\ldots,x_n\in X$ με $x_i\neq x_j,\ i\neq j$. Θα αποδειχθεί ότι ο Gram πίνακας της K ως προς τα $\{x_1,x_2,\ldots,x_n\}$ είναι θετικός.
Έστω $b_1,b_2,\ldots,b_n\in\mathbb{C}$. Τότε:

$$\sum_{i=1}^n\sum_{j=1}^n\overline{b_i}b_jK(x_i,x_j)\overset{(1.5)}{=}\sum_{i=1}^n\sum_{j=1}^n\overline{b_i}b_j\left\langle k_{x_j},k_{x_i}\right\rangle=\sum_{i=1}^n\sum_{j=1}^n\overline{b_i}b_j\left\langle K(\cdot,x_j),K(\cdot,x_i)\right\rangle=$$
$$\sum_{i=1}^n\overline{b_i}\left\langle\sum_{j=1}^n b_jK(\cdot,x_j),K(\cdot,x_i)\right\rangle=\left\langle\sum_{j=1}^n b_jK(\cdot,x_j),\ \sum_{i=1}^n b_iK(\cdot,x_i)\right\rangle=$$
$$\left\langle\sum_{j=1}^n b_jk_{x_j}(\cdot),\ \sum_{i=1}^n b_ik_{x_i}(\cdot)\right\rangle=\left\|\sum_{i=1}^n b_ik_{x_i}(\cdot)\right\|^2\geq 0.$$

Επομένως, πράγματι η K είναι θετικά ορισμένη συνάρτηση. □

Παρατήρηση 1.3.10. *Η έρευνα στρέφεται κυρίως σε χώρους όπου ο Gram πίνακας μιας reproducing kernel συνάρτησης K είναι αυστηρά θετικός, δηλαδή ισχύει $(K(x_i, x_j)) > 0$. Αν κάτι τέτοιο δεν ισχύει, τότε υπάρχει τουλάχιστον ένα διάνυσμα $\boldsymbol{a} = (a_1, a_2, \ldots, a_N)$ μη μηδενικό, τέτοιο ώστε $\left\| \sum_{j=1}^{N} a_j k_{x_j} \right\|^2 = 0 \Leftrightarrow \sum_{j}^{n} a_j k_{x_j} = 0$. Συνεπώς, για κάθε $f \in \mathcal{H}$ έχουμε $\sum_{j=1}^{N} \overline{a_j} f(x_j) = \left\langle f, \sum_{j=1}^{N} a_j k_{x_j} \right\rangle_{\mathcal{H}} = 0$. Δηλαδή, στην περίπτωση αυτή, υπάρχει σχέση γραμμικής εξάρτησης ανάμεσα στις τιμές όλων των συναρτήσεων του $\mathcal{H}$ για κάποιο πεπερασμένο σύνολο σημείων. Υπάρχουν τέτοια παραδείγματα (π.χ. χώροι Sobolev), αλλά τουλάχιστον όσον αφορά εφαρμογές στη Μηχανική Μάθηση οι reproducing kernel συναρτήσεις ορίζουν Gram πίνακες αυστηρά θετικούς.*

Ενώ η τελευταία Πρόταση (1.3.9) είναι αρκετά στοιχειώδης, το αντίστροφο συμπέρασμα είναι πολύ σημαντικό και χαρακτηρίζει τις reproducing kernel συναρτήσεις. Ταυτόχρονα, αποτελεί το δεύτερο συνδετικό κρίκο μεταξύ των θετικά ορισμένων συναρτήσεων και των reproducing kernel συναρτήσεων. Πρόκειται για το Θεώρημα Moore, το οποίο πρωτοπαρουσιάστηκε στο [3] από τον Nachman Aronszajn και αποδόθηκε στον E. H. Moore.

Θεώρημα 1.3.11 (Moore). *Έστω ένα σύνολο X και μια θετικά ορισμένη συνάρτηση $K : X \times X \to \mathbb{C}$. Τότε υπάρχει μοναδικός χώρος $\mathcal{H}$ συναρτήσεων ορισμένων στο X, ο οποίος είναι RKHS, έτσι ώστε η K να αποτελεί τη reproducing kernel συνάρτηση του $\mathcal{H}$.*

Απόδειξη. Έστω $y \in X$. Θέτουμε $k_y(x) = K(x, y), x \in X$, και θεωρούμε $W \subseteq \mathcal{F}(X, \mathbb{C})$, το χώρο που παράγεται από το σύνολο των συναρτήσεων $\{k_y, y \in X\} = \{K(\cdot, y), y \in X\}$, δηλαδή $W = span\{k_y, y \in X\}$.
Θεωρούμε επίσης την απεικόνιση $\langle \cdot, \cdot \rangle_W : W \times W \to \mathbb{C}$, με

$$\left\langle \sum_j a_j k_{y_j},\ \sum_i b_i k_{y_i} \right\rangle_W = \sum_{i,j} a_j \overline{b_i} K(y_i, y_j),$$

όπου $a_j, b_i \in \mathbb{C}$.

Ισχυρισμός: Η απεικόνιση $\langle \cdot, \cdot \rangle_W$ ορίζει εσωτερικό γινόμενο στον W.

Απόδειξη Ισχυρισμού. Έστω $f_1, f_2, f_3 \in W$ και $a_1, a_2 \in \mathbb{C}$:

(i) $\langle f_1, f_1 \rangle_W \geq 0$:
Έστω $f_1 = \sum_j a_j k_{y_j}$. Αφού η K είναι θετικά ορισμένη, έχουμε

$$\langle f_1, f_1 \rangle_W = \sum_{i,j} a_j \overline{a_i} K(y_i, y_j) \geq 0.$$

(ii) $\overline{\langle f_1, f_2\rangle_W} = \langle f_2, f_1\rangle_W$:
Για $f_1 = \sum_j a_j k_{y_j},\ f_2 = \sum_i b_i k_{y_i}$ έχουμε

$$\overline{\langle f_1, f_2\rangle_W} = \overline{\sum_{i,j} a_j \overline{b_i} K(y_i, y_j)} = \sum_{i,j} \overline{a_j} b_i \overline{K(y_i, y_j)} = \sum_{i,j} b_i \overline{a_j} K(y_j, y_i) = \langle f_2, f_1\rangle_W .$$

(iii) $\langle a_1 f_1 + a_2 f_2, f_3\rangle_W = a_1 \langle f_1, f_3\rangle_W + a_2 \langle f_2, f_3\rangle_W$:
Η ισότητα είναι προφανής.

Από τα παραπάνω έπεται ότι η απεικόνιση $\langle\cdot,\cdot\rangle_W$ είναι ημιεσωτερικό γινόμενο. Για να αποδειχθεί ότι είναι εσωτερικό γινόμενο αρκεί να αποδειχθεί ότι

$$\text{αν } f \in W \quad \text{με} \quad \langle f, f\rangle_W = 0,\ \text{τότε } f \equiv 0.$$

Έστω $x \in X$ και $f \in W$ με $f = \sum_j a_j k_{y_j}$. Ισχύει

$$\langle f, k_x\rangle_W = \left\langle \sum_j a_j k_{y_j}, k_x \right\rangle_W = \sum_j a_j \langle k_{y_j}, k_x\rangle_W = \sum_j a_j K(x, y_j) = \sum_j a_j k_{y_j}(x) = f(x).$$

Εφόσον η απεικόνιση $\langle\cdot,\cdot\rangle_W$ είναι ημιεσωτερικό γινόμενο, ισχύει η ανισότητα Cauchy-Schwarz, η οποία με τη βοήθεια της παραπάνω σχέσης γίνεται

$$|\langle f, k_x\rangle_W| \le \|f\|_W \|k_x\|_W \ \Rightarrow\ |f(x)| \le \|f\|_W \|k_x\|_W ,$$

όπου $\|\cdot\|_W$ η ημινόρμα που επάγεται από το ημιεσωτερικό γινόμενο $\langle\cdot,\cdot\rangle_W$. Έστω $\langle f, f\rangle_W = 0$. Ισχύει $\langle f, f\rangle_W = 0 \Leftrightarrow \|f\|_W = 0$, επομένως $|f(x)| \le \|f\|_W \|k_x\|_W = 0$. Άρα $f(x) = 0$ για κάθε $x \in X$, δηλαδή $f \equiv 0$.

■

Συνεπώς, ορίζεται στον W εσωτερικό γινόμενο, το οποίο του δίνει αναλυτική δομή, και άρα υπάρχει μια αφηρημένη πλήρωση $\mathcal{H}$ του χώρου W, η οποία αποτελεί χώρο Hilbert. Θα αποδείξουμε ότι η πλήρωση $\mathcal{H}$ αναπαρίσταται ως υπόχωρος του $\mathcal{F}(X, \mathbb{C})$.
Θεωρούμε μια ακολουθία Cauchy στον W, έστω $\{f_n\}_{n\in\mathbb{N}}, f_n = \sum_j a_j^{(n)} k_{y_j^{(n)}}$, όπου $a_j^{(n)}$ συμβολίζουμε τον $j-$οστό συντελεστή του αθροίσματος του $n-$οστού όρου της ακολουθίας και $k_{y_j^{(n)}}$ συμβολίζουμε την $j-$οστή συνάρτηση του αθροίσματος του $n-$οστού όρου της ακολουθίας. Θεωρούμε, επίσης, τυχόν $x \in X$. Θα αποδείξουμε ότι η ακολουθία $f_n(x) = \sum_j a_j^{(n)} k_{y_j^{(n)}}(x)$ είναι ακολουθία Cauchy μιγαδικών αριθμών. Έχουμε:

$$|f_n(x) - f_m(x)| = |\langle f_n, k_x\rangle_W - \langle f_m, k_x\rangle_W| = |\langle f_n - f_m, k_x\rangle_W| \le \|f_n - f_m\|_W \|k_x\|_W \to 0.$$

Πράγματι, λοιπόν, η ακολουθία $\{f_n(x)\}_n$ είναι Cauchy και, αφού ο χώρος των μιγαδικών αριθμών $\mathbb{C}$ είναι πλήρης, η ακολουθία συγκλίνει σε κάποιο μιγαδικό αριθμό. Έπεται ότι ορίζεται κατά σημείο το όριο της Cauchy ακολουθίας $\{f_n\}_n = \left\{\sum_j a_j^{(n)} k_{y_j^{(n)}}\right\}_n$ και ας είναι $\lim_{n\to\infty} f_n(x) = f(x), x \in X$. Ορίζουμε $\|f\|_{\mathcal{H}} = \lim_{n\to\infty} \|f_n\|_W$. Τότε

$$\|f\|_{\mathcal{H}}^2 = \langle f, f\rangle_{\mathcal{H}} = \lim_{n\to\infty} \langle f_n, f_n\rangle_W = \lim_{n\to\infty} \sum_{i,j} a_j^{(n)} \overline{a_i^{(n)}} K(y_i^{(n)}, y_j^{(n)}),$$

το οποίο είναι ανεξάρτητο της επιλογής της ακολουθίας $\{f_n\}_n$.
Επομένως, ο W έχει μοναδική επέκταση $\mathcal{H}$ στο χώρο $\mathcal{F}(X, \mathbb{C})$, ώστε ο W να είναι πυκνός στον $\mathcal{H}$.
Όσον αφορά την reproducing ιδιότητα, έστω $w \in \mathcal{H}, w = \sum_{j=1}^{\infty} a_j k_{x_j}$ και $y \in X$. Λόγω συνέχειας του εσωτερικού γινομένου έχουμε:

$$\langle w, k_y\rangle_{\mathcal{H}} = \left\langle \sum_{j=1}^{\infty} a_j k_{x_j}, k_y \right\rangle_{\mathcal{H}} = \sum_{j=1}^{\infty} a_j \langle k_{x_j}, k_y\rangle_{\mathcal{H}} = \sum_{j=1}^{\infty} a_j K(y, x_j) = \sum_{j=1}^{\infty} a_j k_{x_j}(y) = w(y).$$

Η μοναδικότητα του χώρου $\mathcal{H}$ προκύπτει από την Πρόταση (1.3.8). □

Το Θεώρημα Moore (1.3.11), σε συνδυασμό με την Πρόταση (1.3.9), την Πρόταση (1.3.8) και τη μοναδικότητα της reproducing kernel συνάρτησης ενός RKHS (Θεώρημα Riesz (1.2.18)), εξασφαλίζει μία ένα προς ένα αντιστοιχία ανάμεσα στους RKHS ενός συνόλου και τις θετικά ορισμένες συναρτήσεις που ορίζονται στο σύνολο. Αυτή αποτελεί μια πολύ ισχυρή ιδιότητα της θεωρίας των RKHS, στην οποία στηρίζεται και το τέχνασμα που περιγράφεται στο Κεφάλαιο 3.

Συμβολισμός
Έστω $K : X \times X \to \mathbb{C}$ μια θετικά ορισμένη συνάρτηση. Συμβολίζουμε με $\mathcal{H}(K)$ τον μοναδικό RKHS που έχει ως reproducing kernel συνάρτηση την K.

Η διαδικασία κατασκευής του χώρου $\mathcal{H}(K)$ με δεδομένη μια συνάρτηση K θετικά ορισμένη ονομάζεται Πρόβλημα Ανοικοδόμησης (Reconstruction Problem) και αποτελεί μια από τις δυσκολότερες προκλήσεις της θεωρίας των RKHS.

1.4 Παραδείγματα Kernel Συναρτήσεων

Προτού προχωρήσουμε, είναι σημαντικό να αναφέρουμε ορισμένα παραδείγματα kernels που απαντώνται συχνά στη βιβλιογραφία και χρησιμοποιούνται σε ποικίλες εφαρμογές.
Οι συναρτήσεις που παρατίθενται παρακάτω ορίζονται στο $X \times X$, όπου $X \subseteq \mathbb{R}^m$.

Gaussian kernel (Gaussian Radial Basis Function (RBF)):

$$K_\sigma(x,y) = exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right),\ \text{όπου } \sigma > 0.$$

Λαπλασιανή kernel (Laplacian kernel):

$$K_\sigma(x,y) = exp\left(-\frac{\|x-y\|}{\sigma}\right),\ \text{όπου } \sigma > 0.$$

Μη ομογενής πολυωνυμική kernel (inhomogeneous polynomial kernel):

$$K_d(x,y) = (\alpha\,\langle x,y\rangle + c)^d,\ \text{όπου } c \geq 0 \text{ σταθερά.}$$

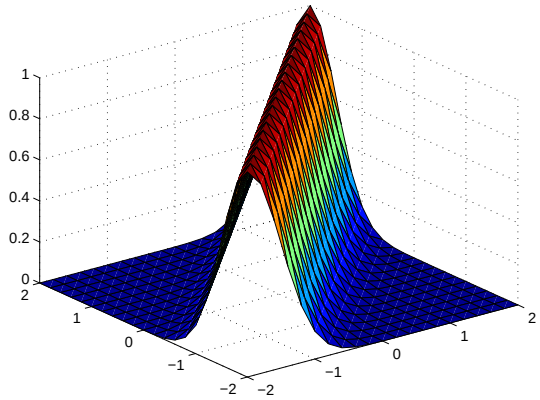
Ομογενής πολυωνυμική kernel (homogeneous polynomial kernel):

$$K_d(x,y) = \langle x,y\rangle^d.$$

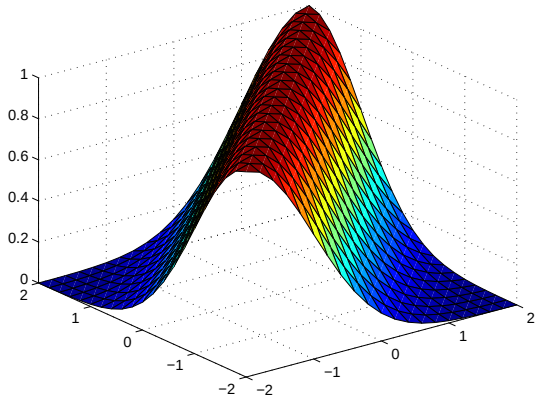
Spline kernel: $K_p(x,y) = B_{2p+1}(\|x-y\|^2)$, όπου $B_n = \bigoplus_{i=1}^{n} I_{[\frac{-i}{2},\frac{i}{2}]}$.

Cosine kernel: $K(x,y) = cos(\angle(x,y))$.

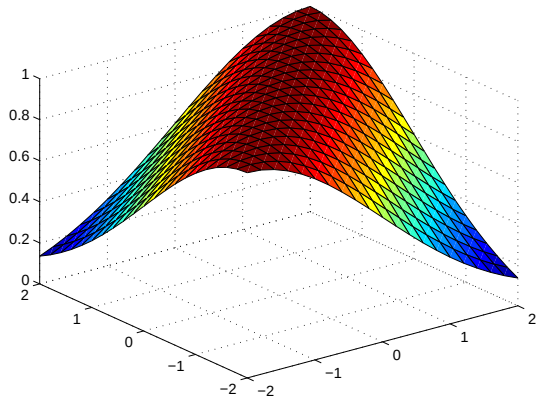
Ίσως η πιο ευρέως διαδεδομένη reproducing kernel συνάρτηση είναι η Gaussian kernel. Αξίζει να σημειωθεί ότι οι RKHS που σχετίζονται με την Gaussian RBF kernel και την Laplacian kernel είναι χώροι άπειρης διάστασης, ενώ οι RKHS που σχετίζονται με τις πολυωνυμικές kernels έχουν πεπερασμένη διάσταση. Περισσότερες πληροφορίες μπορεί κανείς να βρει στο [35].
Τα Σχήματα (1.1) έως (1.5) απεικονίζουν τις γραφικές παραστάσεις κάποιων από τις παραπάνω συναρτήσεις για διάφορες τιμές των αντίστοιχων παραμέτρων, για $X = \mathbb{R}$.



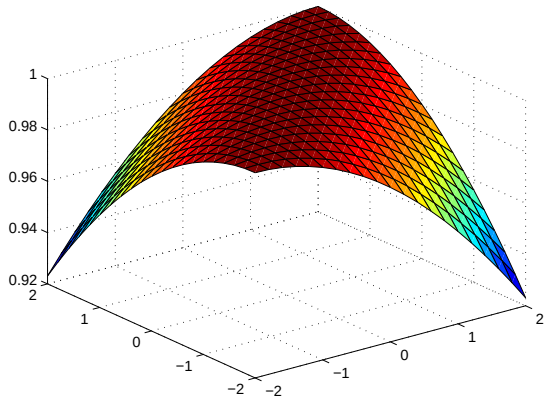
(α′) σ=0,5



(β′) σ=1

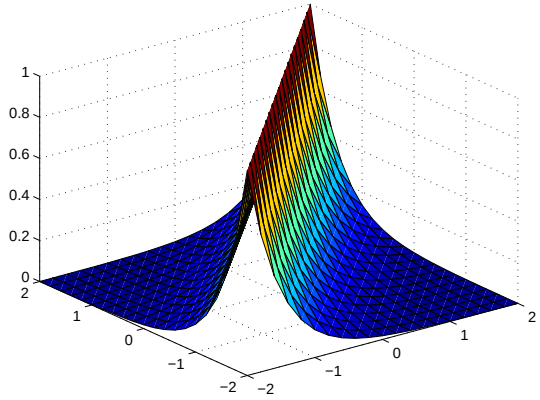


(γ′) σ=2

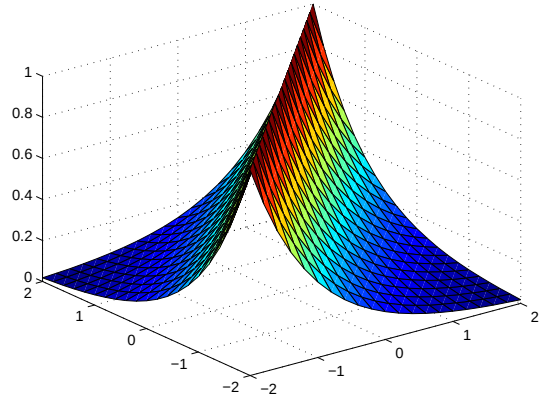


(δ′) σ=10

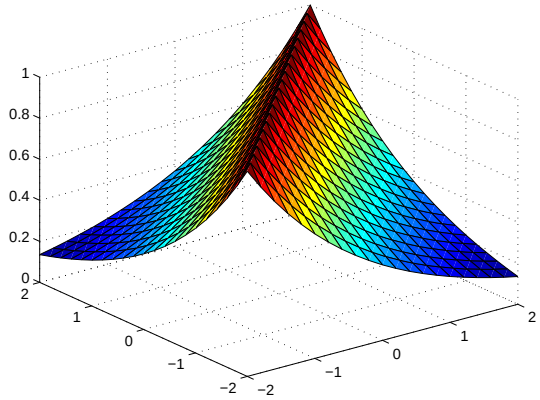
Σχήμα 1.1: Η Gaussian kernel $K_\sigma(x,y) = exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$ για διάφορες τιμές του σ.



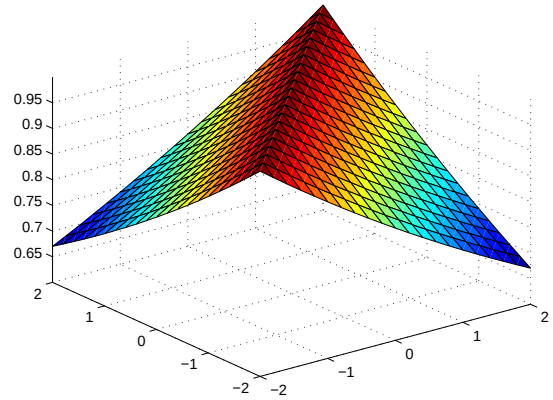
(α′) σ=0,5



(β′) σ=1

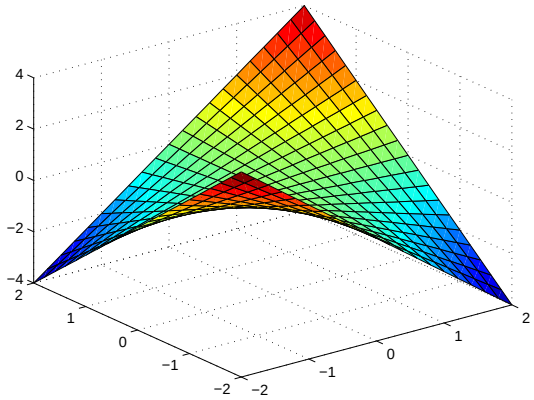


(γ′) σ=2

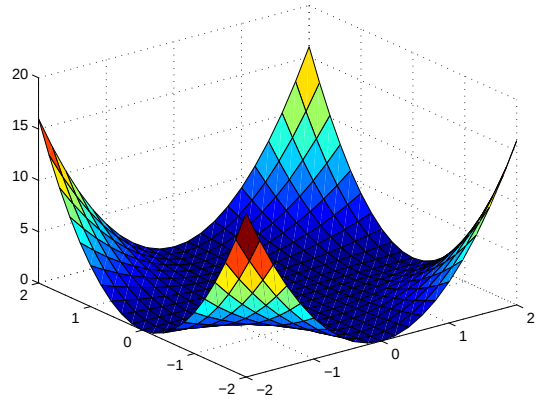


(δ′) σ=10

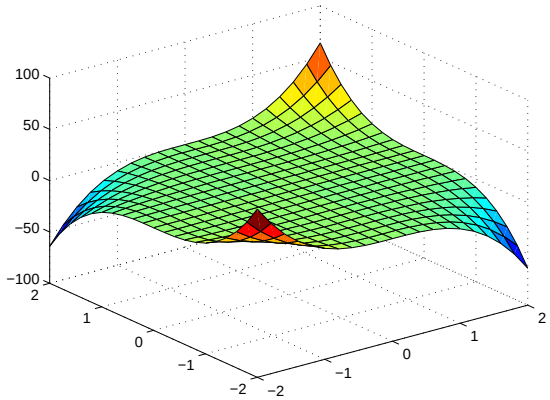
Σχήμα 1.2: Η Laplacian kernel $K_\sigma(x,y) = exp\left(-\frac{\|x-y\|}{\sigma}\right)$ για διάφορες τιμές του σ.



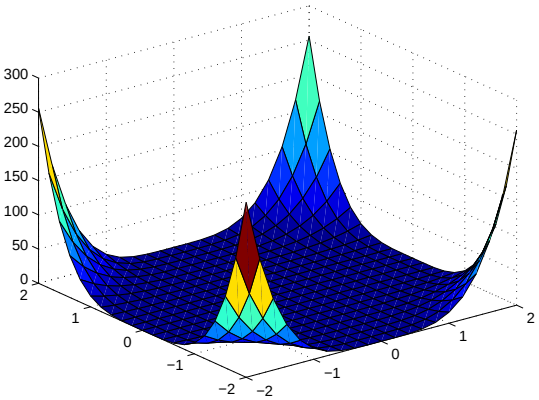
(α′) $d = 1$



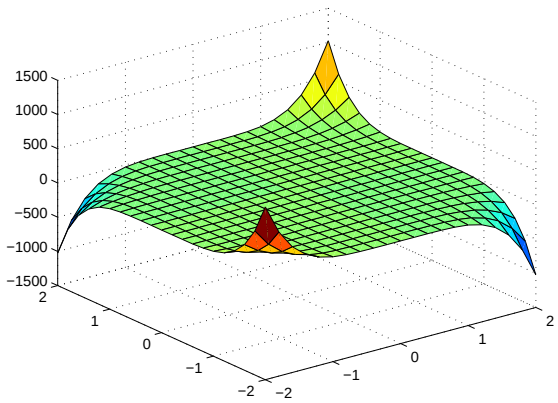
(β′) $d = 2$



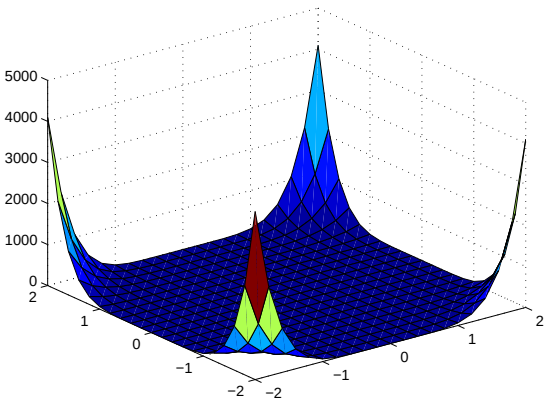
(γ′) $d = 3$



(δ′) $d = 4$

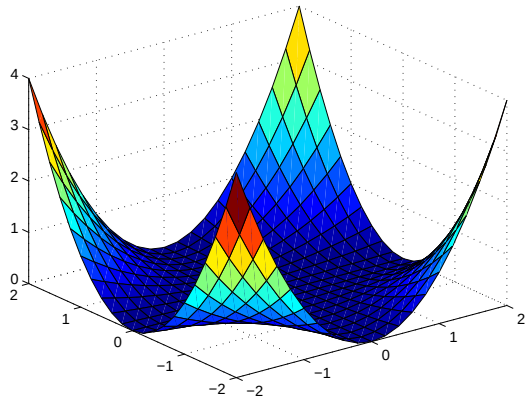


(ε′) $d = 5$

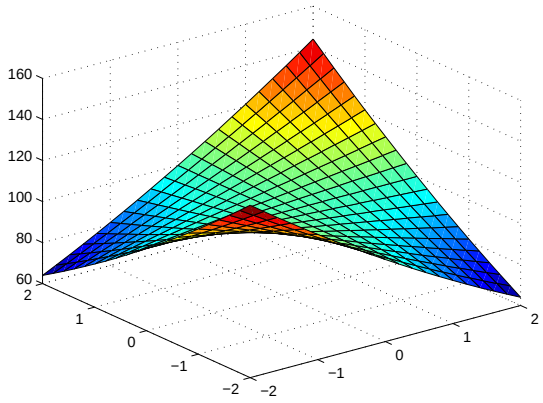


(ϛ′) $d = 6$

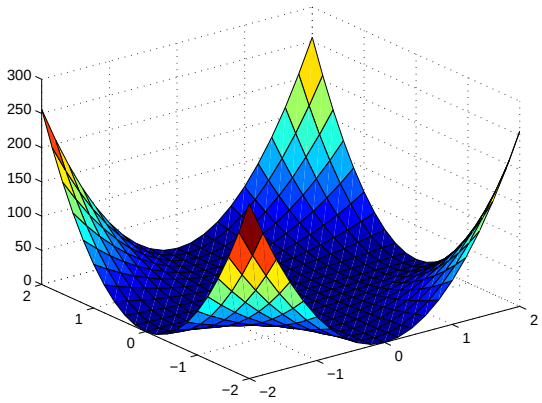
Σχήμα 1.3: Η ομογενής πολυωνυμική kernel $K_d(x, y) = \langle x, y \rangle^d$ για διάφορες τιμές του d.



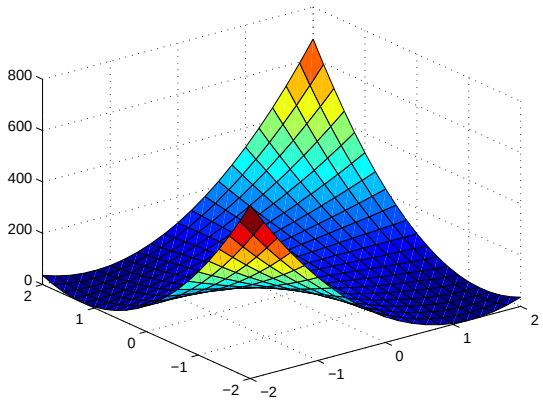
(α′) $\alpha = \frac{1}{2}, \quad c = 0$



(β′) $\alpha = \frac{1}{2}, \quad c = 10$

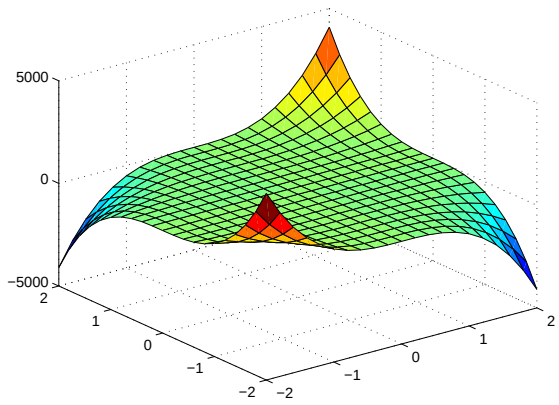


(γ′) $\alpha = 4, \quad c = 0$

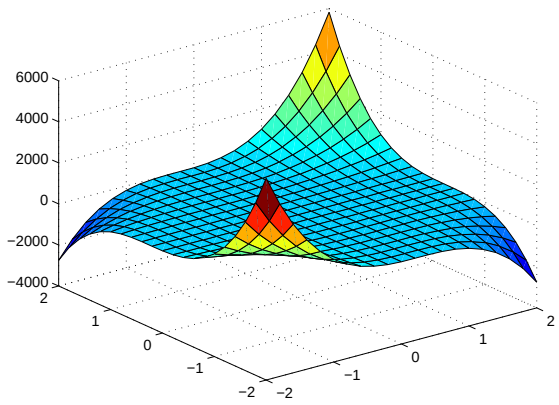


(δ′) $\alpha = 4, \quad c = 10$

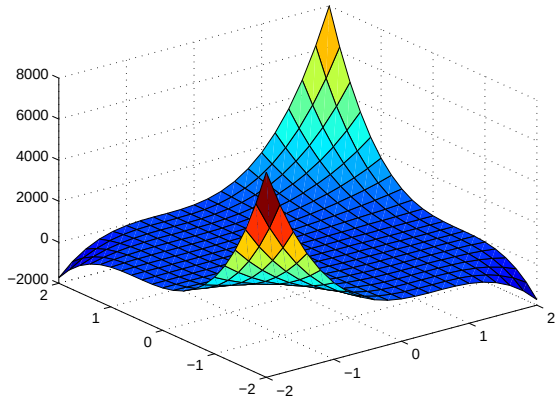
Σχήμα 1.4: Η μη ομογενής πολυωνυμική kernel $K_d(x, y) = (\alpha \langle x, y \rangle + c)^d$ για $d = 2$ και διάφορες τιμές των α και c.



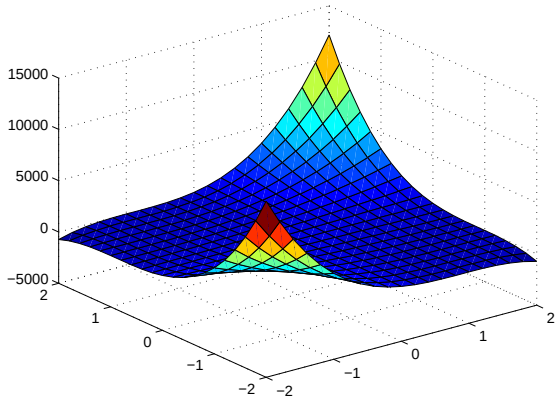
(α′) $c = 0$



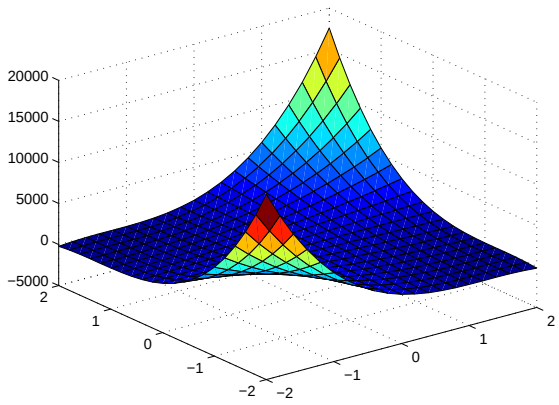
(β′) $c = 2$



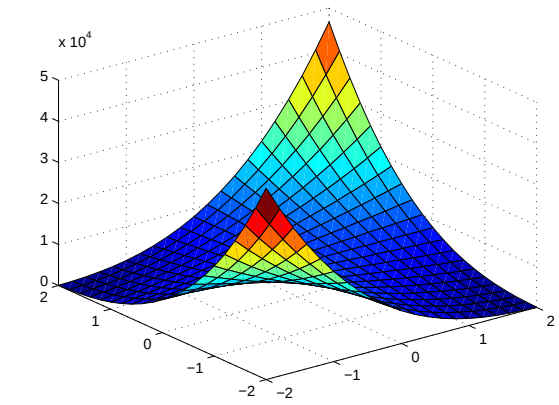
(γ′) $c = 4$



(δ′) $c = 7$



(ε′) $c = 10$



(ϛ′) $c = 20$

Σχήμα 1.5: Η μη ομογενής πολυωνυμική kernel $K_d(x, y) = (\alpha \langle x, y \rangle + c)^d$ για $d = 3,\ \alpha = 4$ και διάφορες τιμές του c.

1.5 Χρήσιμες Προτάσεις που αφορούν Reproducing Kernels

Στην ενότητα αυτή παρατίθενται μερικές χρήσιμες προτάσεις που αφορούν reproducing kernel συναρτήσεις. Επίσης, παρουσιάζονται κάποιες από τις αποδείξεις των προτάσεων.

Πρόταση 1.5.1. *Έστω $\mathcal{H}$ ένας RKHS του X με reproducing kernel συνάρτηση την K. Τότε η $\hat{K}(x,y) = \dfrac{K(x,y)}{\sqrt{K(x,x)K(y,y)}}$ είναι επίσης μία θετικά ορισμένη συνάρτηση στο X. Επίσης, $\left\|\hat{K}(x,y)\right\| \leq 1 \quad \forall x,y \in X$.*

Απόδειξη. Έστω $n \in \mathbb{N}$ και $x_1, x_2, \ldots, x_n \in X$ με $x_i \neq x_j,\ i \neq j$. Θα αποδειχθεί ότι ο Gram πίνακας της K ως προς τα $\{x_1, x_2, \ldots, x_n\}$ είναι θετικός.
Έστω $b_1, b_2, \ldots, b_n \in \mathbb{C}$. Τότε:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \overline{b_i} b_j \hat{K}(x_i,x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \overline{b_i} b_j \frac{K(x_i,x_j)}{\sqrt{K(x_i,x_i)(K(x_j,x_j)}} =$$
$$\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\overline{b_i}}{\sqrt{K(x_i,x_i)}} \frac{b_j}{\sqrt{K(x_j,x_j)}} K(x_i,x_j) \geq 0,$$

αφού η K είναι θετικά ορισμένη συνάρτηση.

Επίσης, για $x, y \in X$ έχουμε $\left\|\hat{K}(x,y)\right\| = \left\|\dfrac{K(x,y)}{\sqrt{K(x,x)K(y,y)}}\right\| \leq 1$ από την ανισότητα Cauchy-Schwarz (1.3.5). □

Θεώρημα 1.5.2. *Έστω $\mathcal{H}$ ένας RKHS του X με reproducing kernel συνάρτηση την K. Τότε η K περιορισμένη στο σύνολο $X_1 \subset X$, $K|_{X_1}$, είναι η reproducing kernel συνάρτηση της κλάσης $\mathcal{H}_1$ των περιορισμών των συναρτήσεων του $\mathcal{H}$ στο υποσύνολο X_1. Η αντίστοιχη νόρμα οποιουδήποτε περιορισμού $f_1 \in \mathcal{H}_1$, που προέρχεται από την $f \in \mathcal{H}$, είναι*

$$\|f_1\|_{\mathcal{H}_1} = min\{\|f\|_{\mathcal{H}} : f \in \mathcal{H}, f|_{X_1} = f_1\}.$$

Πρόταση 1.5.3. *Έστω $\mathcal{H}$ ένας RKHS του X με reproducing kernel συνάρτηση την K. Αν ο $H_0 \subseteq \mathcal{H}$ είναι κλειστός υπόχωρος του $\mathcal{H}$, τότε ο H_0 είναι ένας RKHS του X.*

Παρατήρηση 1.5.4. *Αν k_y είναι η reproducing kernel συνάρτηση στον $\mathcal{H}$ ενός σημείου $y \in X$, τότε η reproducing kernel συνάρτηση στον H_0 για το ίδιο σημείο είναι η συνάρτηση $P(k_y)$, όπου με $P : \mathcal{H} \to H_0$ συμβολίζουμε την ορθογώνια προβολή του $\mathcal{H}$ πάνω στον H_0. Έτσι, θα υπάρχει η reproducing kernel συνάρτηση $K_0(x,y), x, y \in X$, για τον υπόχωρο H_0.*

Θεώρημα 1.5.5. *Έστω $\mathcal{H}_1, \mathcal{H}_2$ δύο RKHS του X με reproducing kernel συναρτήσεις τις K_1, K_2, αντίστοιχα. Τότε η $K = K_1 + K_2$ είναι επίσης μία reproducing kernel συνάρτηση. Ο αντίστοιχος RKHS, $\mathcal{H}$, περιέχει τις συναρτήσεις $f = f_1 + f_2$, όπου $f_i \in \mathcal{H}_i,\ i = 1, 2$. Η αντίστοιχη νόρμα ορίζεται ως*

$$\|f\|^2_{\mathcal{H}} = min\{\|f_1\|^2_{\mathcal{H}_1} + \|f_2\|^2_{\mathcal{H}_2} : f = f_1 + f_2,\ f_i \in \mathcal{H}_i,\ i = 1, 2\}.$$

Απόδειξη. Η απόδειξη ότι η $K = K_1 + K_2$ είναι θετικά ορισμένη συνάρτηση είναι τετριμμένη. Η δυσκολία στην απόδειξη του θεωρήματος έγκειται στο συσχετισμό αυτής της συνάρτησης με το συγκεκριμένο χώρο RKHS $\mathcal{H}$.
Ας θεωρήσουμε το χώρο Hilbert $F = \mathcal{H}_1 \times \mathcal{H}_2$.
Το αντίστοιχο εσωτερικό γινόμενο και η αντίστοιχη νόρμα ορίζονται ως:

$$\langle (f_1, f_2), (g_1, g_2) \rangle_F = \langle f_1, g_1 \rangle_{\mathcal{H}_1} + \langle f_2, g_2 \rangle_{\mathcal{H}_2} ,$$

$$\|(f_1, f_2)\|_F^2 = \|f_1\|_{\mathcal{H}_1}^2 + \|f_2\|_{\mathcal{H}_2}^2 ,$$

όπου $f_1, g_1 \in \mathcal{H}_1$ και $f_2, g_2 \in \mathcal{H}_2$.
Εάν η τομή $\mathcal{H}_0 = \mathcal{H}_1 \cap \mathcal{H}_2$ είναι $\mathcal{H}_0 = \{0\}$, αποδεικνύεται εύκολα ότι υπάρχει μία αντιστοιχία ένα προς ένα μεταξύ των F και $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$, καθώς κάθε $f \in \mathcal{H}$ μπορεί να γραφεί κατά μοναδικό τρόπο ως $f = f_1 \oplus f_2$, όπου $f_1 \in \mathcal{H}_1$ και $f_2 \in \mathcal{H}_2$.
Εάν, όμως, η τομή $\mathcal{H}_0$ είναι μεγαλύτερη από το $\{0\}$, $\mathcal{H}_0 \supset \{0\}$, τότε η απεικόνιση μεταξύ του $f = f_1 + f_2 \in \mathcal{H}$ και του $(f_1, f_2) \in F$ δεν είναι ένα προς ένα. Μπορούμε, ωστόσο, σε αυτές τις περιπτώσεις να βρούμε ένα μικρότερο υπόχωρο, F_0, του F, ο οποίος μπορεί να ταυτιστεί με τον $\mathcal{H}$.
Έστω $F_0 = \{(f, -f) : f \in \mathcal{H}_0\}$. Είναι προφανές ότι ο F_0 είναι κλειστός γραμμικός υπόχωρος του F, άρα θεωρούμε το συμπληρωματικό του υπόχωρο $G = F_0^{\perp}$. Ισχύει $F = F_0 \oplus G$.
Στη συνέχεια, θεωρούμε το γραμμικό μετασχηματισμό $T : F \to \mathcal{H}$, με $T(f_1, f_2) = f_1 + f_2$. Ο πυρήνας αυτού του μετασχηματισμού είναι ο υπόχωρος F_0, από όπου έπεται ότι υπάρχει μία ένα προς ένα αντιστοιχία μεταξύ του G και του $\mathcal{H}$. Θεωρούμε τον αντίστροφο μετασχηματισμό $(T|_G)^{-1} : \mathcal{H} \to G$ με $(T|_G)^{-1}(f) = (\tilde{f}_1, \tilde{f}_2)$ για $f \in \mathcal{H}$, όπου $f = \tilde{f}_1 + \tilde{f}_2$, μέσω του οποίου η $f \in \mathcal{H}$ αναλύεται κατά μοναδικό τρόπο στο G σε δύο συνιστώσες, μία στον $\mathcal{H}_1$ και μία στον $\mathcal{H}_2$. Η ανάλυση αυτή μας επιτρέπει να ορίσουμε ένα εσωτερικό γινόμενο στον $\mathcal{H}$. Για $f, g \in \mathcal{H}$ έχουμε

$$\langle f, g \rangle_{\mathcal{H}} = \left\langle \tilde{f}_1 + \tilde{f}_2, \tilde{g}_1 + \tilde{g}_2 \right\rangle_{\mathcal{H}} = \left\langle \tilde{f}_1, \tilde{g}_1 \right\rangle_{\mathcal{H}_1} + \left\langle \tilde{f}_2, \tilde{g}_2 \right\rangle_{\mathcal{H}_2} = \left\langle (\tilde{f}_1, \tilde{f}_2), (\tilde{g}_1, \tilde{g}_2) \right\rangle_F ,$$
$$\text{όπου} \quad \tilde{f}_1, \tilde{g}_1 \in \mathcal{H}_1, \ \tilde{f}_2, \tilde{g}_2 \in \mathcal{H}_2.$$

Θα αποδείξουμε ότι σε αυτόν το χώρο $\mathcal{H}$ αντιστοιχεί η συνάρτηση $K = K_1 + K_2$.

Ισχυρισμός: Η συνάρτηση $K = K_1 + K_2$ ικανοποιεί τη reproducing ιδιότητα στο χώρο $\mathcal{H}$, δηλαδή

$$\text{αν } y \in X, \quad \text{ισχύει } \langle f, k_y \rangle_{\mathcal{H}} = f(y) \quad \forall f \in \mathcal{H}, \quad \text{όπου } k_y(\cdot) = K(\cdot, y).$$

Απόδειξη Ισχυρισμού. Έστω $y \in X$. Συμβολίζουμε $k_y = K(\cdot, y)$, $k_{y,1} = K_1(\cdot, y)$ και ακόμη $k_{y,2} = K_2(\cdot, y)$. Ισχύει

$$k_y = k_{y,1} + k_{y,2} \in \mathcal{H}, \tag{1.7}$$

δηλαδή $K(\cdot, y) = K_1(\cdot, y) + K_2(\cdot, y) \in \mathcal{H}$.
Επίσης, $T^{-1}(k_y) = (k_y^{(1)}, k_y^{(2)})$, όπου $k_y^{(1)} \in \mathcal{H}_1, k_y^{(2)} \in \mathcal{H}_2$. Έτσι,

$$k_y = k_y^{(1)} + k_y^{(2)}. \tag{1.8}$$

Από τις (1.7) και (1.8) προκύπτει η σχέση $k_{y,1} - k_y^{(1)} = -(k_{y,2} - k_y^{(2)})$, που σημαίνει ότι $(k_{y,1} - k_y^{(1)}, k_{y,2} - k_y^{(2)}) \in F_0$.
Έστω $f \in \mathcal{H}$. Ισχύει $T^{-1}(f) = (\tilde{f}_1, \tilde{f}_2)$ και η f γράφεται $f = \tilde{f}_1 + \tilde{f}_2$, όπου $\tilde{f}_1 \in \mathcal{H}_1, \tilde{f}_2 \in \mathcal{H}_2$. Έχουμε

$$f(y) = \tilde{f}_1(y) + \tilde{f}_2(y) = \left\langle \tilde{f}_1, k_{y,1} \right\rangle_{\mathcal{H}_1} + \left\langle \tilde{f}_2, k_{y,2} \right\rangle_{\mathcal{H}_2} = \left\langle (\tilde{f}_1, \tilde{f}_2), (k_{y,1}, k_{y,2}) \right\rangle_F =$$

$$\left\langle (\tilde{f}_1, \tilde{f}_2), (k_{y,1} - k_y^{(1)}, k_{y,2} - k_y^{(2)}) \right\rangle_F + \left\langle (\tilde{f}_1, \tilde{f}_2), (k_y^{(1)}, k_y^{(2)}) \right\rangle_F =$$

$$\left\langle (\tilde{f}_1, \tilde{f}_2), (k_y^{(1)}, k_y^{(2)}) \right\rangle_F,$$

καθώς από τις σχέσεις $(k_{y,1} - k_y^{(1)}, k_{y,2} - k_y^{(2)}) \in F_0$ και $(\tilde{f}_1, \tilde{f}_2) \in G = F_0^{\perp}$ έπεται ότι $\left\langle (\tilde{f}_1, \tilde{f}_2), (k_{y,1} - k_y^{(1)}, k_{y,2} - k_y^{(2)}) \right\rangle_F = 0$.
Επομένως, $f(y) = \left\langle (\tilde{f}_1, \tilde{f}_2), (k_y^{(1)}, k_y^{(2)}) \right\rangle_F = \left\langle \tilde{f}_1 + \tilde{f}_2, k_y^{(1)} + k_y^{(2)} \right\rangle_{\mathcal{H}} = \langle f, k_y \rangle_{\mathcal{H}}$. Άρα ισχύει $f(y) = \langle f, k_y \rangle_{\mathcal{H}}$ για κάθε $f \in \mathcal{H}$. Επίσης, αφού το y ήταν τυχόν, ισχύει η reproducing ιδιότητα για την K στον $\mathcal{H}$, δηλαδή η $K = K_1 + K_2$ είναι η reproducing kernel συνάρτηση του $\mathcal{H}$.

■

Μένει να αποδείξουμε το τελευταίο μέρος του θεωρήματος.
Θεωρούμε $f \in H$ και υποθέτουμε ότι $\tilde{f}_1$ και $\tilde{f}_2$ είναι οι μοναδικές συνιστώσες στις οποίες αναλύεται η f μέσω του T^{-1}, δηλαδή $f = \tilde{f}_1 + \tilde{f}_2,\ \tilde{f}_1 \in \mathcal{H}_1,\ \tilde{f}_2 \in \mathcal{H}_2$.
Επίσης, $f_i \in \mathcal{H}_i,\ i = 1, 2$, τέτοιες ώστε $f = f_1 + f_2$. Έχουμε $f_1 + f_2 = \tilde{f}_1 + \tilde{f}_2 \Leftrightarrow$
$f_1 - \tilde{f}_1 = -(f_2 - \tilde{f}_2)$, που σημαίνει ότι $(f_1 - \tilde{f}_1, f_2 - \tilde{f}_2) \in F_0$, όπως πριν. Επομένως, ισχύει

$$\|f_1\|_{\mathcal{H}_1}^2 + \|f_2\|_{\mathcal{H}_2}^2 = \|(f_1, f_2)\|_F^2 = \left\|(\tilde{f}_1, \tilde{f}_2)\right\|_F^2 + \left\|(f_1 - \tilde{f}_1, f_2 - \tilde{f}_2)\right\|_F^2 =$$
$$\left\|\tilde{f}_1\right\|_{\mathcal{H}_1}^2 + \left\|\tilde{f}_2\right\|_{\mathcal{H}_2}^2 + \left\|(f_1 - \tilde{f}_1, f_2 - \tilde{f}_2)\right\|_F^2 = \|f\|_{\mathcal{H}}^2 + \left\|(f_1 - \tilde{f}_1, f_2 - \tilde{f}_2)\right\|_F^2.$$

Από την τελευταία σχέση συμπεραίνουμε ότι $\|f\|_{\mathcal{H}}^2 = \|f_1\|_{\mathcal{H}_1}^2 + \|f_2\|_{\mathcal{H}_2}^2$ αν και μόνο αν ισχύουν $f_1 = \tilde{f}_1$ και $f_2 = \tilde{f}_2$.
Από τον ορισμό του εσωτερικού γινομένου στον $\mathcal{H}$ προκύπτει

$$\|f\|_{\mathcal{H}}^2 = \left\|\tilde{f}_1\right\|_{\mathcal{H}_1}^2 + \left\|\tilde{f}_2\right\|_{\mathcal{H}_2}^2 = \left\|(\tilde{f}_1, \tilde{f}_2)\right\|_F^2 = \left\|(\tilde{f}_1, \tilde{f}_2)\right\|_G^2 = \|(f_1, f_2)\|_{F/F_0}^2$$

Από τον ορισμό της νόρμας του χώρου πηλίκο έχουμε

$$\|(f_1, f_2)\|_{F/F_0} = min\{\|(f_1, f_2)\|_F : f = f_1 + f_2,\ f_1 \in \mathcal{H}_1,\ f_2 \in \mathcal{H}_2\}$$

άρα

$$\|(f_1, f_2)\|_{F/F_0}^2 = min\{\|f_1\|_{\mathcal{H}_1}^2 + \|f_2\|_{\mathcal{H}_2}^2 : f = f_1 + f_2,\ f_i \in \mathcal{H}_i,\ i = 1, 2\}.$$

Έτσι, ολοκληρώνεται η απόδειξη. □

Παρατηρήσεις 1.5.6. (i) *Επαγωγικά, το Θεώρημα (1.5.5) μπορεί να επεκταθεί στην περίπτωση που* $K(x,y) = \sum_{j=1}^{n} K_j(x,y), x, y \in X$.

(ii) *Παρατηρούμε ότι αν για τους* $\mathcal{H}_1, \mathcal{H}_2$ *ισχύει* $\mathcal{H}_1 \cap \mathcal{H}_2 = \{0\}$, *τότε η νόρμα δίνεται από τη σχέση* $\|f\|_{\mathcal{H}}^2 = \|f_1\|_{\mathcal{H}_1}^2 + \|f_2\|_{\mathcal{H}_2}^2$. *Αυτή είναι η μοναδική περίπτωση που οι* $\mathcal{H}_1, \mathcal{H}_2$ *είναι συμπληρωματικοί υπόχωροι στον* $\mathcal{H}$.

Πόρισμα 1.5.7. *Έστω* $\mathcal{H}$ *ένας RKHS του* X *με reproducing kernel συνάρτηση την* K *και οι* H_1, H_2 *είναι συμπληρωματικοί υπόχωροι του* $\mathcal{H}$ *με αντίστοιχες reproducing kernel συναρτήσεις τις* K_1, K_2. *Τότε ισχύει* $K = K_1 + K_2$.

Πέραν του αθροίσματος των kernels, υπάρχουν και άλλοι μετασχηματισμοί που διατηρούν τις reproducing kernel συναρτήσεις. Πολλοί από αυτούς παρουσιάζονται παρακάτω. Αναλυτική περιγραφή για τον επαγόμενο RKHS, καθώς και αποδείξεις για τις περιπτώσεις που αναφέρονται στην Πρόταση (1.5.8), μπορεί κανείς να βρει στα [3], [35].

Πρόταση 1.5.8. *Έστω ότι οι* $K_i : X \times X \to \mathbb{C}, i = 0, 1, 2, \ldots,$ *και η* $K : \tilde{X} \times \tilde{X} \to \mathbb{C}$ *είναι θετικά ορισμένες συναρτήσεις και οι* $f_1 : \tilde{X} \to X, \quad f_2 : X \to \mathbb{C}$ *είναι τυχαίες συναρτήσεις. Τότε, ισχύουν τα εξής:*

(i) *Η* λK_0 *είναι θετικά ορισμένη συνάρτηση για κάθε* $\lambda \geq 0$. *Μάλιστα, τότε ισχύει* $\mathcal{H}(\lambda K_0) = \mathcal{H}(K_0)$ *αν* $\lambda > 0$ *και* $\mathcal{H}(0) = \{0\}$ *αν* $\lambda = 0$.

(ii) *Η* $K_1 + K_2$ *είναι θετικά ορισμένη συνάρτηση, όπως αποδείξαμε στο Θεώρημα (1.5.5).*

(iii) *Η* $K_1 \cdot K_2$ *είναι θετικά ορισμένη συνάρτηση.*

(iv) *Αν για τις* K_i, $i = 1, 2, \ldots,$ *ισχύει* $\lim_{n\to\infty} K_n(x,y) = K_0(x,y)$ *για κάθε* $x, y \in X$, *τότε η* K_0 *είναι θετικά ορισμένη συνάρτηση.*

(v) *Αν* $p(z)$ *είναι ένα πολυώνυμο με μη αρνητικούς συντελεστές, τότε η* $p(K_0)$ *είναι θετικά ορισμένη συνάρτηση.*

(vi) *Η* $e^{K_0(x,y)}$ *είναι θετικά ορισμένη συνάρτηση.*

(vii) *Η* $K_0(f_1(x), f_1(y))$ *είναι θετικά ορισμένη συνάρτηση στο* $\tilde{X}$.

(viii) *Η* $f_2(x)K_0(x,y)\overline{f_2(y)}$ *είναι θετικά ορισμένη συνάρτηση στο* X.

(ix) *Το τανυστικό γινόμενο των* K_0 *και* K, $(K_0 \otimes K)(x, y, \tilde{x}, \tilde{y})$, *όπου* $x, y \in X$, $\tilde{x}, \tilde{y} \in \tilde{X}$, *είναι θετικά ορισμένη συνάρτηση στο* $X \times \tilde{X}$.

(x) *Το ευθύ άθροισμα των* K_0 *και* K, $(K_0 \oplus K)(x, y, \tilde{x}, \tilde{y})$, *όπου* $x, y \in X$, $\tilde{x}, \tilde{y} \in \tilde{X}$, *είναι θετικά ορισμένη συνάρτηση στο* $X \times \tilde{X}$.

Παρατήρηση 1.5.9. *Όπως είδαμε στην περίπτωση (vi) της Πρότασης (1.5.8), η $e^{K_0(x,y)}$ είναι θετικά ορισμένη συνάρτηση. Για την απόδειξη, αρκεί να παρατηρήσουμε ότι η συνάρτηση e^z γράφεται ως ανάπτυγμα Taylor, το οποίο μπορεί να θεωρηθεί όριο πολυωνύμων με μη αρνητικούς συντελεστές.*

Υπάρχουν δύο σημαντικές κλάσεις από kernels που ακολουθούν συγκεκριμένους κανόνες και χρησιμοποιούνται ευρέως στην πράξη.
Η πρώτη περιλαμβάνει τις **συναρτήσεις kernel εσωτερικού γινομένου** (**dot product kernels**). Αυτές ορίζονται ως $K(x,y) = f(\langle x,y\rangle)$, για κάποια πραγματική συνάρτηση f.
Τη δεύτερη κλάση αποτελούν οι **συναρτήσεις kernel** που είναι **αναλλοίωτες ως προς μεταφορές** (**translation invariant kernels**) και ορίζονται ως $K(x,y) = f(x-y)$, για κάποια πραγματική συνάρτηση f ορισμένη στο σύνολο X.
Ακολουθούν δύο Θεωρήματα στα οποία διατυπώνονται αναγκαίες και ικανές συνθήκες ώστε τέτοιες συναρτήσεις να είναι reproducing kernels.

Θεώρημα 1.5.10. *Έστω $f : \mathbb{R} \to \mathbb{R}$, όπου $f(t) = \sum_n a_n t^n$. Μία συνάρτηση K ορισμένη στο σύνολο X, όπου $K(x,y) = f(\langle x,y\rangle)$, είναι θετικά ορισμένη συνάρτηση αν και μόνο αν ισχύει $a_n \geq 0$ για κάθε n.*

Θεώρημα 1.5.11. *(**Κριτήριο Bochner-Fourier για kernels αναλλοίωτες ως προς μεταφορές** [11]) Έστω $f : X \to \mathbb{R}$. Μία συνάρτηση $K(x,y) = f(x-y)$ ορισμένη στο σύνολο $X \subseteq \mathbb{R}^m$ είναι θετικά ορισμένη συνάρτηση, εάν ο μετασχηματισμός Fourier*

$$F[K](\boldsymbol{\omega}) = (2\pi)^{-\frac{N}{2}} \int_X e^{-i\langle \boldsymbol{\omega}, x\rangle} f(x)dx$$

είναι μη αρνητικός.

Παρατήρηση 1.5.12. *Αξιοποιώντας τα παραπάνω Θεωρήματα, μπορεί κανείς να αποδείξει ότι κάποιες από τις συναρτήσεις kernel που αναφέρθηκαν προηγουμένως είναι θετικά ορισμένες.*

(i) *Προκύπτει άμεσα από το Θεώρημα (1.5.10) ότι η συνάρτηση $K(x,y) = \langle x,y\rangle$ είναι θετικά ορισμένη.*

(ii) *Η cosine kernel: Αρκεί να παρατηρήσουμε ότι $cos(\angle(x,y)) = \dfrac{\langle x,y\rangle}{\|x\|\,\|y\|}$.*
Μπορεί εναλλακτικά να αποδειχθεί ότι η cosine kernel είναι θετικά ορισμένη συνάρτηση με τη βοήθεια της Πρότασης (1.5.1), αφού πρόκειται για την κανονικοποίηση της απλής συνάρτησης kernel $\langle x,y\rangle$.

(iii) *Η ομογενής πολυωνυμική kernel: Αν το $p(z) = z^d$ είναι πολυώνυμο με μη αρνητικούς συντελεστές, τότε η $p(\langle x,y\rangle) = (\langle x,y\rangle)^d$ είναι θετικά ορισμένη συνάρτηση.*

Σημείωση 1.5.13. *Για να αποδείξει κανείς ότι η Gaussian kernel και η Laplacian kernel είναι θετικά ορισμένες συναρτήσεις χρησιμοποιείται παραγώγιση κατά Fréchet.*

Κεφάλαιο 2

Εφαρμογές Μηχανικής Μάθησης σε Γραμμικά Προβλήματα

Τα προβλήματα Μηχανικής Μάθησης διακρίνονται σε γραμμικά και μη γραμμικά. Τα γραμμικά προβλήματα είναι απλούστερα και ο τρόπος προσέγγισής τους λιγότερο περίπλοκος. Ως εκ τούτου, παρουσιάζονται πρώτα αυτά, ώστε να εξοικειωθεί ο αναγνώστης με τις τεχνικές που χρησιμοποιούνται και στο επόμενο Κεφάλαιο να παρουσιαστούν τα, πιο σύνθετα ως προς την προσέγγιση, μη γραμμικά προβλήματα, όπου χρησιμοποιείται η θεωρία των RKHS.

2.1 Εισαγωγή

Πολλά από τα προβλήματα Μηχανικής Μάθησης είναι ισοδύναμα με την εκτίμηση μιας άγνωστης συναρτησιακής σχέσης που συνδέει τις μεταβλητές εισόδου (αίτιο) με τις μεταβλητές εξόδου (αποτέλεσμα). Συνήθως αυτή η συνάρτηση επιλέγεται να είναι συγκεκριμένου τύπου - παραδείγματος χάριν γραμμική ή τετραγωνική - και περιγράφεται από ένα σύνολο άγνωστων παραμέτρων. Υπάρχουν δύο βασικοί τρόποι προσδιορισμού αυτών των παραμέτρων.
Ο πρώτος τρόπος, γνωστός ως Μπεϋζιανή Συμπερασματολογία ή Συμπερασματολογία κατά Bayes (Bayesian Inference), αντιμετωπίζει τις παραμέτρους ως τυχαίες μεταβλητές [9]. Σύμφωνα με την προσέγγιση αυτή, η σχέση των μεταβλητών εισόδου και εξόδου εκφράζεται μέσω ενός συνόλου συναρτήσεων πυκνότητας πιθανότητας (σ.π.π.). Το πρόβλημα στην περίπτωση αυτή είναι να εκτιμηθεί η σ.π.π. που συσχετίζει τις μεταβλητές εισόδου-εξόδου, καθώς και η σ.π.π. που περιγράφει την τυχαία φύση των παραμέτρων. Ωστόσο, ο πρωταρχικός σκοπός της Συμπερασματολογίας κατά Bayes δεν είναι να ανακαλύψει συγκεκριμένες τιμές για τις παραμέτρους. Αντιθέτως, η δεύτερη προσέγγιση θεωρεί ότι οι άγνωστες παράμετροι είναι ντετερμινιστικές και κύριο μέλημα είναι να προσδιορίσει συγκεκριμένες εκτιμήσεις για τις τιμές τους.
Οι δύο τρόποι προσδιορισμού των παραμέτρων έχουν ένα κοινό σημείο. Ανακαλύπτουν τις απαιτούμενες πληροφορίες, με σκοπό να εκτιμήσουν τις παραμέτρους, από ένα διαθέσιμο σύνολο μετρήσεων εισόδων-εξόδων, που είναι γνωστά ως **δεδομένα εκπαίδευσης**. Το στάδιο της εκτίμησης των σ.π.π. ή των παραμέτρων ονομάζεται **εκπαίδευση** και αυτή η φιλοσοφία

μάθησης λέγεται **Επιβλεπόμενη Μάθηση** (**Supervised Learning**). Υπάρχει επίσης η **Μη Επιβλεπόμενη Μάθηση** (**Unsupervised Learning**), καθώς και η **Ημιεπιβλεπόμενη Μάθηση** (**Semi-Supervised Learning**) [37]. Στην πρώτη δεν υπάρχουν καθόλου μετρήσεις των μεταβλητών εξόδου στα δεδομένα εκπαίδευσης, ενώ στη δεύτερη υπάρχουν πολύ λίγες. Εμείς θα ασχοληθούμε με την Επιβλεπόμενη Μάθηση ενός συνόλου παραμέτρων που θα θεωρούμε ντετερμινιστικές.

2.2 Παραμετρική Μοντελοποίηση

Στη φύση, δεν μπορεί κανείς να ξέρει εάν υπάρχει ένα «πραγματικό» μοντέλο για τα δεδομένα που έχει. Η βεβαιότητά μας για τα «πραγματικά» μοντέλα αφορά μόνο τα δεδομένα προσομοίωσης. Στην πραγματικότητα, πρέπει κανείς να υιοθετήσει ένα μοντέλο και μια μέθοδο εκπαίδευσης για να προσδιορίσει τις παραμέτρους του μοντέλου, ώστε τα αποτελέσματα να είναι χρήσιμα στην πράξη.
Όσον αφορά τη διαδικασία της εκτίμησης, αφότου υιοθετηθεί ένα παραμετρικό μοντέλο, το πρόβλημα της εκτίμησης συνίσταται στην υιοθέτηση ενός κριτηρίου που μετρά πόσο καλά προσεγγίζουν οι τιμές που προβλέφθηκαν μέσω του μοντέλου τις μετρήσεις που έχουμε από τα δεδομένα εκπαίδευσης. Διαφορετικά κριτήρια καταλήγουν σε διαφορετικές εκτιμήσεις και επαφίεται στην κρίση του σχεδιαστή να επιλέξει τελικά αυτό που καλύπτει καλύτερα τις ανάγκες του. Για συγκεκριμένο πλήθος εκπαιδευτικών δεδομένων, χρησιμοποιείται πολύ συχνά η **cross validation** τεχνική για να προσδιοριστεί η «καλύτερη» επιλογή συνάρτησης κόστους. Σύμφωνα με την τεχνική αυτή, το σύνολο S των δεδομένων χωρίζεται σε d μικρότερα σύνολα S_d, κατά το δυνατόν περίπου ίσα. Η διαδικασία επαναλαμβάνεται d φορές και κάθε φορά επιλέγεται διαφορετικό σύνολο S_d για έλεγχο, ενώ τα υπόλοιπα $d-1$ σύνολα χρησιμοποιούνται για την εκπαίδευση. Έτσι, έχουμε το πλεονέκτημα ότι διασταυρώνουμε τα αποτελέσματα με ένα σύνολο S_d από τα δεδομένα το οποίο δεν έχει εμπλακεί στην εκπαίδευση και άρα μπορεί να θεωρηθεί ανεξάρτητο, ενώ ταυτόχρονα χρησιμοποιούνται τελικά όλα τα δεδομένα και για εκπαίδευση και για διασταύρωση. Μετά το πέρας της διαδικασίας,

- είτε συνδυάζονται οι d εκτιμήσεις, για παράδειγμα παίρνοντας τη μέση τιμή τους,
- είτε συνδυάζονται οι παράμετροι βάσει των σφαλμάτων ελέγχου, ώστε να πάρουμε καλύτερη εκτίμηση του γενικού σφάλματος που αναμένουμε να λάβει ο εκτιμητής σε εφαρμογές.

Τελευταίο βήμα στην εκτιμητική διαδικασία αποτελεί η επιλογή ενός αλγορίθμου ο οποίος θα βελτιστοποιεί το κριτήριο που προαναφέρθηκε. Ως προς την επιλογή αυτή, η κλασική προσέγγιση είναι η λεγόμενη **batch processing προσέγγιση**, όπου τα δεδομένα εκπαίδευσης είναι διαθέσιμα όλα ταυτόχρονα. Ο σχεδιαστής έχει πρόσβαση σε όλα τα δεδομένα συγχρόνως, τα οποία στη συνέχεια χρησιμοποιούνται για τη βελτιστοποίηση.
Όλα τα batch σχήματα έχουν ένα μεγάλο μειονέκτημα. Από τη στιγμή που είναι διαθέσιμη μια νέα μέτρηση μετά την ολοκλήρωση της εκπαίδευσης, πρέπει ολόκληρη η διαδικασία εκπαίδευσης να αρχίσει από την αρχή με καινούριο σύνολο μετρήσεων που θα περιλαμβάνει τη νέα μέτρηση. Αναμφίβολα, αυτός δεν είναι αποτελεσματικός τρόπος για να αντιμετωπίσει

κανείς το πρόβλημα. Αντίθετα, αποτελεσματικές είναι μέθοδοι που κάνουν τη βελτιστοποίηση αναδρομικά, υπό την έννοια ότι ενημερώνουν την τρέχουσα εκτίμηση κάθε φορά που παραλαμβάνεται μια νέα μέτρηση, λαμβάνοντας υπόψη μόνο αυτή τη νέα πληροφορία και την εκτίμηση των παραμέτρων που είναι διαθέσιμη εκείνη τη στιγμή. Τα δεδομένα εκπαίδευσης που ήδη χρησιμοποιήθηκαν προηγουμένως δε θα χρησιμοποιηθούν ποτέ ξανά, καθώς όλες οι πληροφορίες που είχαν να συνεισφέρουν στη διαδικασία εκπαίδευσης έχουν πια ενσωματωθεί στην τρέχουσα εκτίμηση. Αναφερόμαστε σ΄ αυτά τα **εκπαιδευτικά σχήματα** ως **online** ή **adaptive**. Ο όρος **Προσαρμοστική Μάθηση (Adaptive Learning)** χρησιμοποιείται κυρίως στην Επεξεργασία Σήματος, αλλά τέτοια σχήματα έχουν χρησιμοποιηθεί επίσης εκτενώς στις Επικοινωνίες, τη Θεωρία Συστημάτων και Αυτόματου Ελέγχου από τη δεκαετία του 1960, στα πλαίσια του LMS, του RLS και του φίλτρου Kalman [23], [31], [18]. Την κινητήρια δύναμη πίσω από την ανάγκη για αυτές τις μεθόδους αποτέλεσε η ανάγκη για έναν εκπαιδευτικό μηχανισμό που θα μπορεί να ενσωματώσει αργές μεταβολές του εκπαιδευτικού περιβάλλοντος στο χρόνο και σταδιακά θα μπορεί να ξεχάσει το παρελθόν. Μάλιστα, αυτή η φιλοσοφία μιμείται καλύτερα τον τρόπο με τον οποίο ο ανθρώπινος εγκέφαλος μαθαίνει και προσπαθεί να προσαρμοστεί σε νέες συνθήκες. Οι μηχανισμοί online μάθησης έχουν γίνει τελευταία πολύ δημοφιλείς σε εφαρμογές όπως την Ανίχνευση Δεδομένων (Data Mining) και τη Βιοπληροφορική (Bioinformatics), αλλά και γενικά σε περιπτώσεις όπου τα δεδομένα βρίσκονται σε πολύ μεγάλες βάσεις δεδομένων.
Στην εργασία αυτή θα ασχοληθούμε με αλγορίθμους προσαρμοστικής μάθησης στην αντιμετώπιση προβλημάτων Μηχανικής Μάθησης.

Παραθέτουμε παρακάτω μερικές βασικές έννοιες που χρησιμοποιούμε στο Κεφάλαιο αυτό, προς διευκόλυνση του μη εξοικειωμένου αναγνώστη.

Δεδομένα εκπαίδευσης (Training data): Ένα σύνολο μετρήσεων εισόδων-εξόδων που χρησιμοποιούνται για την εκπαίδευση μιας συγκεκριμένης μηχανής.

Online learning σχήμα: Ένας τρόπος εκπαίδευσης, στον οποίο τα δεδομένα φθάνουν το ένα μετά το άλλο και χρησιμοποιούνται μόνο μία φορά.

Batch learning σχήμα: Ένας τρόπος εκπαίδευσης, στον οποίο τα δεδομένα είναι όλα γνωστά εκ των προτέρων.

Πρόβλημα παλινδρόμησης (Regression task): Ένα πρόβλημα παραμετρικής μοντελοποίησης, όπου εκτιμάται μία παραμετρική σχέση εισόδων-εξόδων που ταιριάζει καλύτερα στα διαθέσιμα δεδομένα.

Πρόβλημα κατηγοριοποίησης (Classification task): Ένα πρόβλημα παραμετρικής μοντελοποίησης, όπου τα δεδομένα κατηγοριοποιούνται σε μία ή περισσότερες κλάσεις.

2.3 Γραμμική Παλινδρόμηση και Κατηγοριοποίηση

Δύο από τα σημαντικότερα προβλήματα στην παραμετρική μοντελοποίηση στη Μηχανική Μάθηση είναι αυτά της παλινδρόμησης (regression) και της κατηγοριοποίησης (classification). Πρόκειται για δύο στενά συνδεδεμένα προβλήματα, αν και είναι διαφορετικά. Περισσότερες πληροφορίες υπάρχουν στο [36].
Στη **γραμμική παλινδρόμηση (linear regression)**, έχοντας ένα σύνολο N δεδομένων εκπαίδευσης $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^{N},\ \boldsymbol{x_n} \in \mathbb{R}^m,\ y_n \in \mathbb{R}$, επιθυμούμε να εντοπίσουμε τη σχέση εισόδου-εξόδου μέσω ενός μοντέλου της μορφής

$$y_n = f(\boldsymbol{x_n}) + \eta_n,\ n = 1, 2, \ldots, N. \tag{2.1}$$

Γενικά ισχύει $\boldsymbol{y_n} \in \mathbb{R}^l$, αλλά θεωρούμε ότι $y_n \in \mathbb{R}$ χάριν απλότητας.
Η ακολουθία η_n είναι κάποιος θόρυβος, μη παρατηρήσιμος και, στη γραμμική περίπτωση, η συνάρτηση $f(\cdot)$ είναι η

$$f(\boldsymbol{x_n}) = \sum_{k=1}^{K} \theta_k x_n^{(k)}, \tag{2.2}$$

όπου $\boldsymbol{x_n} = [x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(K)}]^T$ και $\theta_k,\ k = 1, 2, \ldots, K$, είναι οι άγνωστες παράμετροι. Συνδυάζοντας τη σχέση (2.1) με τη (2.2) έχουμε $y_n = \sum_{k=1}^{K} \theta_k x_n^{(k)} + \eta_n$ και γράφουμε

$$y_n = \boldsymbol{\theta}^T \boldsymbol{x_n} + \eta_n,$$

όπου $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_K]^T$.
Στόχος του προβλήματος της εκτίμησης παραμέτρων είναι να πάρουμε εκτιμήσεις $\hat{\boldsymbol{\theta}}$ για το $\boldsymbol{\theta}$ χρησιμοποιώντας τα διαθέσιμα δεδομένα εκπαίδευσης. Αφού βρούμε το $\hat{\boldsymbol{\theta}}$ και δώσουμε κάποια τιμή στο $\boldsymbol{x}$, η πρόβλεψη για την αντίστοιχη τιμή εξόδου υπολογίζεται σύμφωνα με το μοντέλο

$$\hat{y} = \hat{\boldsymbol{\theta}}^T \boldsymbol{x}. \tag{2.3}$$

Στην **κατηγοριοποίηση (classification)**, εργαζόμαστε λίγο διαφορετικά. Οι μεταβλητές εξόδου είναι διακριτές, δηλαδή $y_n \in D$, όπου το σύνολο D είναι πεπερασμένο σύνολο με διακριτές τιμές. Για παράδειγμα, για την απλή περίπτωση ενός προβλήματος κατηγοριοποίησης με δύο κλάσεις (δυαδική κατηγοριοποίηση) θα μπορούσε κανείς να επιλέξει $D = \{1, -1\}$ ή $D = \{0, 1\}$. Οι τιμές εξόδου είναι γνωστές ως **ετικέτες κλάσεων (class labels)**. Τα διανύσματα εισόδου λέγονται **χαρακτηριστικά διανύσματα (feature vectors)** και επιλέγονται ούτως ώστε οι αντίστοιχες συνιστώσες να εμπεριέχουν όσο το δυνατόν περισσότερη πληροφορία για τη διάκριση στις κλάσεις.
Στόχος της κατηγοριοποίησης είναι να σχεδιαστεί μία συνάρτηση f (ή στη γενική περίπτωση ένα σύνολο συναρτήσεων), γνωστή ως **ταξινομητής (classifier)**, έτσι ώστε το αντίστοιχο

(υπερ-) επίπεδο $f(\boldsymbol{x}) = 0$ στο χώρο των $\boldsymbol{x}$ να διαχωρίζει τα σημεία που ανήκουν σε διαφορετικές κλάσεις με τον καλύτερο δυνατό τρόπο. Συγκεκριμένα, ο σκοπός ενός ταξινομητή είναι να χωρίσει το χώρο εισόδων σε περιοχές, που η κάθε μία από αυτές συνδέεται με μία και μόνο από τις κλάσεις. Η επιφάνεια (για την περίπτωση του απλού προβλήματος της δυαδικής κατηγοριοποίησης) είναι γνωστή ως **επιφάνεια απόφασης (decision surface)**. Οι συναρτήσεις στη **γραμμική κατηγοριοποίηση (linear classification)** είναι της μορφής (2.2).
Επομένως, ο σχεδιασμός ενός ταξινομητή αντιμετωπίζεται και πάλι ως ένα πρόβλημα εκτίμησης παραμέτρων. Αφού υπολογίζονται εκτιμήσεις $\hat{\boldsymbol{\theta}}$ για τις άγνωστες παραμέτρους $\boldsymbol{\theta}$, δεδομένου ενός χαρακτηριστικού διανύσματος $\boldsymbol{x}$ - που προκύπτει από έναν κανόνα και η ετικέτα κλάσης του, y, είναι άγνωστη - η προβλεπόμενη ετικέτα λαμβάνεται ως

$$\hat{y} = g(f(\boldsymbol{x})), \tag{2.4}$$

όπου η $g(\cdot)$ λέγεται **συνάρτηση ταξινόμησης** και δείχνει σε ποια μεριά του (υπερ-) επιπέδου $f(\boldsymbol{x}) = 0$ βρίσκεται το $\boldsymbol{x}$. Συνήθως χρησιμοποιείται η $g(\cdot) = sgn(\cdot)$, δηλαδή η συνάρτηση ελέγχου προσήμου (sign function) με

$$sgn(\boldsymbol{x}) = \begin{cases} \dfrac{\boldsymbol{x}}{|\boldsymbol{x}|}, \text{ αν } \boldsymbol{x} \neq \mathbf{0} \\ 0, \text{ αν } \boldsymbol{x} = \mathbf{0}. \end{cases}$$

Τα προβλήματα παλινδρόμησης και κατηγοριοποίησης είναι δύο τυπικά προβλήματα Μηχανικής Μάθησης και πλήθος εκπαιδευτικών εφαρμογών μπορεί να διατυπωθεί ως κάποιο από τα δύο. Είδαμε ότι μπορούν και τα δύο να αντιμετωπιστούν ως προβλήματα εκτίμησης παραμέτρων. Φυσικά, μπορούν και τα δύο προβλήματα να προσεγγιστούν διαμέσου άλλων οδών. Για παράδειγμα, μπορεί να χρησιμοποιηθεί η μέθοδος του k-πλησιέστερου γείτονα, η οποία δεν εμπλέκει καθόλου εκτίμηση παραμέτρων [37].

2.4 Εκτίμηση Παραμέτρων

Ας δώσουμε μία πιο τυπική περιγραφή για το πρόβλημα εκτίμησης παραμέτρων.

- Δίνεται ένα σύνολο N δεδομένων εκπαίδευσης $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^{N}$, όπου $\boldsymbol{x_n} \in \mathbb{R}^m$, $y_n \in Y \subseteq \mathbb{R}$.

- Υιοθετούμε μια παραμετρική κλάση συναρτήσεων

  $$\mathcal{F} = \{f_{\boldsymbol{\theta}}(\cdot) : \boldsymbol{\theta} \in \mathcal{A} \subseteq \mathbb{R}^K\},$$

  η οποία, στις περιπτώσεις που έχουμε δει ως τώρα, είναι η κλάση των γραμμικών συναρτήσεων

  $$\mathcal{F}_{lin} = \{f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^T\boldsymbol{x},\ \boldsymbol{\theta} \in \mathbb{R}^K\}.$$

Στόχος μας είναι η εύρεση μίας συνάρτησης από την κλάση $\mathcal{F}$, που θα συμβολίζουμε $f_{\boldsymbol{\theta}*}(\cdot)$, η οποία - δεδομένων μιας τιμής για το $\boldsymbol{x}$ και ενός μοντέλου (π.χ. (2.3) για παλινδρόμηση ή (2.4) για κατηγοριοποίηση) - να υπολογίζει την αντίστοιχη τιμή εξόδου, $\hat{y}$, κατά το βέλτιστο τρόπο. Βέλτιστο, όμως, με ποια έννοια; Για αυτό το λόγο:

- Επιλέγουμε μία συνάρτηση κόστους (loss function - cost function) από ένα σύνολο διαθέσιμων μη αρνητικών συναρτήσεων

$$\mathcal{L} : \mathbb{R} \times \mathbb{R} \to [0, +\infty)$$

και

- Υπολογίζουμε το $\boldsymbol{\theta}_*$ ούτως ώστε να ελαχιστοποιήσουμε τη συνολική απώλεια επάνω σε όλα τα σημεία από τα εκπαιδευτικά δεδομένα, δηλαδή

$$f_{\boldsymbol{\theta}_*}(\cdot) : \boldsymbol{\theta}_* \in \underset{\boldsymbol{\theta}\in\mathcal{A}}{\operatorname{argmin}} J(\boldsymbol{\theta}),$$

όπου

$$J(\boldsymbol{\theta}) := \sum_{n=1}^{N} \mathcal{L}(y_n, f_{\boldsymbol{\theta}}(\boldsymbol{x_n})), \tag{2.5}$$

υποθέτοντας ότι υπάρχει ελάχιστο. Σε μία adaptive διαδικασία, η ελαχιστοποίηση της συνάρτησης κόστους γίνεται αναδρομικά στη μορφή

$$\boldsymbol{\theta_n} = \boldsymbol{\theta_{n-1}} + (\text{όρος διόρθωσης σφάλματος}),$$

καθώς το n αυξάνει.

Διαφορετικές κλάσεις συναρτήσεων, $\mathcal{F}$, και διαφορετικές συναρτήσεις κόστους, $\mathcal{L}$, δίνουν διαφορετικές εκτιμήσεις. Στην πράξη, για συγκεκριμένο πλήθος εκπαιδευτικών δεδομένων, χρησιμοποιείται η cross validation τεχνική για να προσδιοριστεί η «καλύτερη» επιλογή. Φυσικά, στην πράξη, όταν χρησιμοποιούνται adaptive τεχνικές και το n αφήνεται να μεταβάλλεται, οι τεχνικές cross validation δεν έχουν νόημα, ειδικά σε περιβάλλοντα χρονικά μεταβαλλόμενα. Σε τέτοιες περιπτώσεις, η επιλογή της κλάσης συναρτήσεων καθώς και της συνάρτησης κόστους επιλέγονται εμπειρικά. Οι καθοριστικοί παράγοντες για τις επιλογές αυτές είναι η υπολογιστική πολυπλοκότητα, η ταχύτητα σύγκλισης, η ικανότητα του αλγορίθμου να προσαρμόζεται στις αλλαγές, καθώς επίσης και η ευρωστία στο θόρυβο και τη συσσώρευση αριθμητικών σφαλμάτων.
Υπάρχει μια σημαντική δυσκολία στα προβλήματα Μηχανικής Μάθησης που πρέπει να υπερπηδήσουμε και που σχετίζεται με την πολυπλοκότητα του επιλεγμένου μοντέλου. Πρόκειται για το λεγόμενο **overfitting**. Εάν το μοντέλο είναι πολύ περίπλοκο, όσον αφορά το πλήθος των δεδομένων εκπαίδευσης, N, τότε τείνει να «μαθαίνει» πάρα πολλά από το συγκεκριμένο σύνολο δεδομένων. Προσπαθεί δηλαδή να προσαρμοστεί όσο το δυνατόν καλύτερα στα

σημεία του συνόλου εκπαίδευσης, αλλά δε διαχειρίζεται καλά δεδομένα εκτός του συνόλου αυτού, δεν μπορεί να γενικεύσει καλά. Εάν, από την άλλη, το μοντέλο είναι πολύ απλό, δεν έχει την ικανότητα να εξάγει ούτε καν τις απαραίτητες πληροφορίες που παρέχουν τα δεδομένα εισόδου, με αποτέλεσμα να οδηγεί σε κακές εκτιμήσεις. Στην πράξη, λοιπόν, αναζητά κανείς τη χρυσή τομή μετρώντας την επίδοση διάφορων εκτιμητών. Η επίδοση μετράται με τη βοήθεια ενός κριτηρίου απώλειας (π.χ. μέσο τετραγωνικό σφάλμα) ελέγχοντας πόσο κοντά βρίσκονται οι τιμές που προέβλεψε το μοντέλο στις πραγματικές τιμές.
Μπορεί κανείς να αντιμετωπίσει το παραπάνω πρόβλημα χρησιμοποιώντας έναν επιπλέον όρο στη συνάρτηση κόστους, ο οποίος εξισορροπεί/αντισταθμίζει την εκτίμηση μεταξύ των δύο αυτών ακραίων καταστάσεων. Η τεχνική αυτή είναι γνωστή ως **μέθοδος εξομάλυνσης (regularization)**. Με αυτήν τη μέθοδο, ελαχιστοποιείται η συνεισφορά του όρου της συνάρτησης κόστους, κάτι που σχετίζεται με την ακρίβεια του εκτιμητή, ενώ ταυτόχρονα το μοντέλο παραμένει όσο το δυνατόν λιγότερο περίπλοκο. Αυτό γίνεται προσθέτοντας έναν όρο στη συνάρτηση κόστους στη σχέση (2.5), ο οποίος διατηρεί τη νόρμα των παραμέτρων όσο γίνεται μικρότερη. Έτσι, έχουμε

$$J(\boldsymbol{\theta}) := \sum_{n=1}^{N} \mathcal{L}(y_n, f_{\boldsymbol{\theta}}(\boldsymbol{x_n})) + \lambda\Omega(\|\boldsymbol{\theta}\|),$$

όπου $\Omega(\cdot)$ είναι μία γνησίως αύξουσα μη αρνητική συνάρτηση και η $\|\cdot\|$ είναι μια νόρμα, για παράδειγμα η ευκλείδεια νόρμα (L_2) ή η L_1 νόρμα. Η παράμετρος λ λέγεται **παράμετρος εξομάλυνσης** και ελέγχει τη σχετική σημαντικότητα των δύο όρων στη συνάρτηση κόστους. Συχνά επιλέγονται ως συναρτήσεις Ω οι παρακάτω:

$$\Omega(\|\boldsymbol{\theta}\|) = \sum_{k=1}^{K} |\theta_k|^2 \quad \text{ή} \quad \Omega(\|\boldsymbol{\theta}\|) = \sum_{k=1}^{K} |\theta_k| \,.$$

Επίσης, χρησιμοποιούνται πολύ συχνά τεχνικές **αραίωσης** (**sparsification**) για να αποφευχθεί το overfitting. Με τέτοιες τεχνικές η επιλεγμένη μέθοδος εκτίμησης επιτυγχάνει να παρασύρει προς το μηδέν τις λιγότερο σημαντικές, ως προς την ακρίβεια, συνιστώσες του $\boldsymbol{\theta}$.

Στις ενότητες που ακολουθούν θα δούμε δύο ευρέως διαδεδομένους αλγορίθμους, τους οποίους στο επόμενο Κεφάλαιο θα επεκτείνουμε σε μη γραμμικά προβλήματα με τη βοήθεια της θεωρίας των RKHS.

2.5 Ο Αλγόριθμος Least Mean Squares (LMS)

Ας υποθέσουμε ότι έχουμε στη διάθεσή μας μια ακολουθία παραδειγμάτων εισόδων-εξόδων $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^{N}$, όπου $\boldsymbol{x_n} \in X \subseteq \mathbb{R}^m$ και $y_n \in \mathbb{R},\ n \in \mathbb{N}$, και επιθυμούμε να ανακαλύψουμε το μηχανισμό μιας συνάρτησης $F : X \to \mathbb{R},\ X \subseteq \mathbb{R}^m$ η οποία περιγράφει τη σχέση εισόδων-εξόδων.

Στόχος ενός τυπικού αλγορίθμου προσαρμοστικής μάθησης είναι ο προσδιορισμός σύμφωνα με τα δεδομένα μιας σχέσης εισόδου-εξόδου, $f_{\boldsymbol{\theta}}$, μέσα από μια παραμετρική κλάση συναρτήσεων $\mathcal{F} = \{f_{\boldsymbol{\theta}} : X \to \mathbb{R},\ \boldsymbol{\theta} \in \mathbb{R}^K\}$, έτσι ώστε να ελαχιστοποιείται η τιμή μιας προκαθορισμένης συνάρτησης κόστους $\mathcal{L} : \mathbb{R}^K \to [0, +\infty)$ η οποία σε κάθε βήμα n υπολογίζει το σφάλμα ανάμεσα στο πραγματικό αποτέλεσμα, y_n, και στην εκτίμησή του, $\hat{y}_n = f_{\boldsymbol{\theta}}(\boldsymbol{x_n})$.
Στην πιο συνηθισμένη μορφή του LMS αλγορίθμου υιοθετείται η κλάση των γραμμικών συναρτήσεων

$$\mathcal{F}_{lin} = \{f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x},\ \boldsymbol{\theta} \in \mathbb{R}^K\},$$

ενώ ως συνάρτηση κόστους χρησιμοποιείται το **μέσο τετραγωνικό σφάλμα (mean square error (MSE))**

$$\mathcal{L}(\boldsymbol{\theta}) = E[|y_n - f_{\boldsymbol{\theta}}(\boldsymbol{x_n})|^2] = E[\left|y_n - \boldsymbol{\theta}^T \boldsymbol{x_n}\right|^2].$$

Συμβολίζουμε

$$e_n = y_n - \boldsymbol{\theta}_{\boldsymbol{n-1}}^T \boldsymbol{x_n}, \quad n = 1, 2, \ldots, N$$

το **εκ των προτέρων σφάλμα** (**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

$$-\nabla \mathcal{L}(\boldsymbol{\theta}) = 2E[(y_n - \boldsymbol{\theta}_{\boldsymbol{n-1}}^T \boldsymbol{x_n})\boldsymbol{x_n}] = 2E[e_n \boldsymbol{x_n}]$$

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 **ενημερωμένη εξίσωση βήματος** (**step update** ή **weight update equation**), η οποία προς την κατεύθυνση της ελαχιστοποίησης είναι

$$\boldsymbol{\theta_n} = \boldsymbol{\theta_{n-1}} + \mu e_n \boldsymbol{x_n}, \quad n = 1, 2, \ldots, N,$$

όπου μ είναι η παράμετρος που εκφράζει πόσο μεγάλο είναι το «βήμα» μας προς την κατεύθυνση της καθόδου (αποκαλείται και **βήμα εκμάθησης**). Υποθέτοντας ότι $\boldsymbol{\theta_0} = \mathbf{0}$, η τελευταία εξίσωση οδηγεί στη σχέση

$$\boldsymbol{\theta_n} = \mu \sum_{i=1}^{n} e_i \boldsymbol{x_i}, \quad n = 1, 2, \ldots, N,$$

ενώ η έξοδος που εκτιμάται στο βήμα n είναι

$$\hat{y}_n = f_{\boldsymbol{\theta_{n-1}}}(\boldsymbol{x_n}) = \boldsymbol{\theta}_{\boldsymbol{n-1}}^T \boldsymbol{x_n} = \mu \sum_{i=1}^{n-1} e_i \boldsymbol{x_i}^T \boldsymbol{x_n}, \quad n = 1, 2, \ldots, N.$$

Ο Κώδικας **Least Mean Squares (LMS)**

Είσοδος: Τα δεδομένα $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^{N}$ και η παράμετρος μ του βήματος εκμάθησης.
1: **Αρχικοποίηση**: $\boldsymbol{\theta} = \mathbf{0}$.
2: **for** i=1 to N
3: Υπολόγισε την έξοδο του συστήματος: $\hat{y}_i = \boldsymbol{\theta}^T \boldsymbol{x_i}$
4: Υπολόγισε το σφάλμα: $e_i = y_i - \hat{y}_i$
5: $\boldsymbol{\theta} = \boldsymbol{\theta} + \mu e_i \boldsymbol{x_i}$
6: **end for**
Έξοδος: Το διάνυσμα $\boldsymbol{\theta} \in \mathbb{R}^N$.

Υπάρχουν διάφορα **κριτήρια σύγκλισης** του LMS αλγορίθμου. Ένα από αυτά, ίσως το δημοφιλέστερο, διατυπώνεται ως εξής:
Εφόσον το φίλτρο είναι γραμμικό, δηλαδή $F(\boldsymbol{x_n}) = \boldsymbol{\theta_*}^T \boldsymbol{x_n} + \eta_n$, και η $\boldsymbol{x_n}$ είναι υπό την ασθενή έννοια στάσιμη διαδικασία, τότε $\mu \sum_{n=1}^{\infty} e_n \boldsymbol{x_n} \to \boldsymbol{\theta_*}$ και $\mathcal{L}(\boldsymbol{\theta_{n-1}}) = E[|e_n|^2] \overset{n\to\infty}{\longrightarrow} c$, c σταθερά, αν το μ ικανοποιεί τη συνθήκη $0 < \mu < \dfrac{2}{\lambda_{max}}$, όπου λ_{max} είναι η μεγαλύτερη ιδιοτιμή του πίνακα συσχέτισης $R = E[\boldsymbol{x_n} \boldsymbol{x_n}^T]$. Στην πράξη, αρκεί $0 < \mu < \dfrac{2}{tr(R)}$.
Επίσης, καθώς ο αλγόριθμος LMS εμφανίζει ευαισθησία στην κλίμακα των $\boldsymbol{x_i}$, καθίσταται δύσκολη η επιλογή βήματος εκμάθησης μ που 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 επαναληπτικής διαδικασίας (βήμα 5) αντικατασταθεί από την

$$\boldsymbol{\theta} = \boldsymbol{\theta} + \frac{\mu e_i}{\|\boldsymbol{x_i}\|^2} \boldsymbol{x_i},$$

όπου $\mu \in (0, 2)$. Ο αλγόριθμος καλείται πλέον **Normalized LMS (NLMS)** και έχει αποδειχθεί ότι ο βέλτιστος ρυθμός εκμάθησης επιτυγχάνεται για $\mu = 1$.
Συνεπώς, σε κάθε περίπτωση, μετά από μια εκπαίδευση n βημάτων, το $\boldsymbol{\theta_n}$ εκφράζεται ως γραμμικός συνδυασμός όλων των εισόδων από το $\boldsymbol{x_1}$ έως και το $\boldsymbol{x_n}$, 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:

$$f(\boldsymbol{x_{n+1}}) = \boldsymbol{\theta_n}^T \boldsymbol{x_{n+1}} = \mu \sum_{i=1}^{n} e_i \boldsymbol{x_i}^T \boldsymbol{x_{n+1}},$$

όπου

$$e_i = y_i - \mu \sum_{j=1}^{i-1} e_j \boldsymbol{x_j}^T \boldsymbol{x_i}.$$

Όπως θα δούμε στο επόμενο Κεφάλαιο, αυτό χρησιμοποιείται στο kernel τέχνασμα και ο αλγόριθμος LMS επεκτείνεται στον Kernel LMS.

2.6 Ο Αλγόριθμος Recursive Least Squares (RLS)

Υπενθυμίζουμε ότι θεωρήσαμε δεδομένη μια ακολουθία παραδειγμάτων $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^N$, όπου $\boldsymbol{x_n} \in X \subseteq \mathbb{R}^m,\ y_n \in \mathbb{R}$. Υιοθετούμε και για τον αλγόριθμο RLS την παραμετρική κλάση των γραμμικών συναρτήσεων

$$\mathcal{F}_{lin} = \{f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x},\ \boldsymbol{\theta} \in \mathbb{R}^K\}$$

ως το σύνολο όπου αναζητάμε τη βέλτιστη σχέση εισόδου-εξόδου ανάμεσα στα $\boldsymbol{x_n}$ και y_n, ενώ η συνάρτηση κόστους στην επανάληψη n παίρνει τη μορφή:

$$\mathcal{L}_n(\theta) = \sum_{i=1}^{n} |y_i - f_{\boldsymbol{\theta}}(\boldsymbol{x_i})|^2 + \lambda \|\boldsymbol{\theta}\|^2,\ n = 1, 2, \ldots, N \tag{2.6}$$

για κάποιο επιλεγμένο $\lambda > 0$. Παρατηρήστε ότι η συνάρτηση κόστους αποτελείται από δύο όρους. Ο πρώτος μετρά το σφάλμα μεταξύ της εξόδου που εκτιμάται, $f_{\boldsymbol{\theta}}(\boldsymbol{x_i})$, και της πραγματικής εξόδου, y_i, κάθε χρονική στιγμή i έως τη στιγμή n, ενώ ο δεύτερος είναι ο όρος εξομάλυνσης. Η ύπαρξη του δεύτερου όρου έχει μεγάλη σημασία για τον αλγόριθμο, καθώς συντελεί στο να αποφευχθεί το overfitting. Κάθε χρονική στιγμή n ο RLS βρίσκει τη λύση στο πρόβλημα ελαχιστοποίησης $\min_{\boldsymbol{\theta}} \mathcal{L}_n(\boldsymbol{\theta})$.

Ο αλγόριθμος RLS, συγκρινόμενος με τον LMS, παρουσιάζει συνήθως πολύ γρηγορότερη σύγκλιση με κόστος τη μεγαλύτερη υπολογιστική πολυπλοκότητα.

Η λογική του αλγορίθμου RLS μπορεί να περιγραφεί σε πιο συμπαγή μορφή, αν κανείς χρησιμοποιήσει πίνακες. Ορίζουμε, λοιπόν, τους $m \times n$ πίνακες $X_n = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n})$ και τα διανύσματα $\boldsymbol{y_{(n)}} = (y_1, y_2, \ldots, y_n)^T \in \mathbb{R}^n$, για $n = 1, 2, \ldots, N$. Τότε η συνάρτηση κόστους (2.6) μπορεί ισοδύναμα να γραφεί:

$$\mathcal{L}_n(\boldsymbol{\theta}) = \left\|\boldsymbol{y_{(n)}} - X_n^T \boldsymbol{\theta}\right\|^2 + \lambda \|\boldsymbol{\theta}\|^2 = (\boldsymbol{y_{(n)}} - X_n^T \boldsymbol{\theta})^T (\boldsymbol{y_{(n)}} - X_n^T \boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|^2 =$$
$$\left\|\boldsymbol{y_{(n)}}\right\|^2 - 2\boldsymbol{\theta}^T X_n \boldsymbol{y_{(n)}} + \boldsymbol{\theta}^T X_n X_n^T \boldsymbol{\theta} + \lambda \boldsymbol{\theta}^T \boldsymbol{\theta} = \left\|\boldsymbol{y_{(n)}}\right\|^2 - 2(X_n \boldsymbol{y_{(n)}})^T \boldsymbol{\theta} + \boldsymbol{\theta}^T X_n X_n^T \boldsymbol{\theta} + \lambda \boldsymbol{\theta}^T \boldsymbol{\theta}.$$

Καθώς η $\mathcal{L}_n(\boldsymbol{\theta})$ είναι αυστηρά κυρτή συνάρτηση, έχει μοναδικό ελάχιστο που λαμβάνεται στο σημείο $\boldsymbol{\theta_n} := \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{L}_n(\boldsymbol{\theta})$, όπου μηδενίζεται η κλίση, $\nabla \mathcal{L}_n(\boldsymbol{\theta_n}) = 0$. Παρατηρήστε ότι η κλίση της συνάρτησης κόστους, χρησιμοποιώντας τους γνωστούς κανόνες παραγώγισης, είναι

$$\nabla \mathcal{L}_n(\boldsymbol{\theta_n}) = -2X_n \boldsymbol{y_{(n)}} + 2X_n X_n^T \boldsymbol{\theta} + 2\boldsymbol{\theta}\lambda.$$

Εξισώνοντας την κλίση με 0 παίρνουμε τη λύση $\boldsymbol{\theta_n}$ του προβλήματος ελαχιστοποίησης $\min_{\boldsymbol{\theta}} \mathcal{L}_n(\boldsymbol{\theta})$:

$$\boldsymbol{\theta_n} = (\lambda I_m + X_n X_n^T)^{-1} X_n \boldsymbol{y_{(n)}},$$

όπου I_m είναι ο $m \times m$ ταυτοτικός πίνακας. Παρατηρήστε ότι αυτή η λύση περιέχει σε κάθε βήμα την αντιστροφή ενός πίνακα, κάτι που είναι συνήθως υπολογιστικά απαιτητικό. Για να υπερπηδήσει αυτό το εμπόδιο, ο αλγόριθμος RLS υπολογίζει τη λύση αναδρομικά αξιοποιώντας το γνωστό matrix inversion lemma (τον τύπο των Sherman-Morrison-Woodbury):

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1},$$

όπου $A \in \mathbb{R}^{m \times m},\ B \in \mathbb{R}^{m \times l},\ C \in \mathbb{R}^{l \times k},\ D \in \mathbb{R}^{k \times m}$.
Θεωρώντας $P_n = \lambda I_m + X_n X_n^T$ προκύπτει εύκολα ότι $P_n = P_{n-1} + \boldsymbol{x_n}\boldsymbol{x_n}^T$. Θέτοντας στον παραπάνω τύπο $A = P_{n-1},\ B = \boldsymbol{x_n},\ C = [1]$ και $D = \boldsymbol{x_n}^T$ παίρνουμε:

$$P_n^{-1} = (P_{n-1} + \boldsymbol{x_n}\boldsymbol{x_n}^T)^{-1} = P_{n-1}^{-1} - \frac{P_{n-1}^{-1}\boldsymbol{x_n}\boldsymbol{x_n}^T P_{n-1}^{-1}}{1 + \boldsymbol{x_n}^T P_{n-1}^{-1}\boldsymbol{x_n}}.$$

Παρατηρώντας ότι $X_n \boldsymbol{y_{(n)}} = X_{n-1}\boldsymbol{y_{(n-1)}} + \boldsymbol{x_n} y_n$, η λύση $\boldsymbol{\theta_n}$ γίνεται:

$$\boldsymbol{\theta_n} = P_n^{-1} X_n \boldsymbol{y_{(n)}} = \left(P_{n-1}^{-1} - \frac{P_{n-1}^{-1}\boldsymbol{x_n}\boldsymbol{x_n}^T P_{n-1}^{-1}}{1 + \boldsymbol{x_n}^T P_{n-1}^{-1}\boldsymbol{x_n}}\right)(X_{n-1}\boldsymbol{y_{(n-1)}} + \boldsymbol{x_n} y_n) =$$

$$P_{n-1}^{-1} X_{n-1}\boldsymbol{y_{(n-1)}} + P_{n-1}^{-1}\boldsymbol{x_n} y_n - \frac{P_{n-1}^{-1}\boldsymbol{x_n}\boldsymbol{x_n}^T P_{n-1}^{-1} X_{n-1}\boldsymbol{y_{(n-1)}}}{1 + \boldsymbol{x_n}^T P_{n-1}^{-1}\boldsymbol{x_n}} - \frac{P_{n-1}^{-1}\boldsymbol{x_n}\boldsymbol{x_n}^T P_{n-1}^{-1}\boldsymbol{x_n} y_n}{1 + \boldsymbol{x_n}^T P_{n-1}^{-1}\boldsymbol{x_n}} =$$

$$\boldsymbol{\theta_{n-1}} - \frac{P_{n-1}^{-1}\boldsymbol{x_n}\boldsymbol{x_n}^T\boldsymbol{\theta_{n-1}}}{1 + \boldsymbol{x_n}^T P_{n-1}^{-1}\boldsymbol{x_n}} + \left(P_{n-1}^{-1}\boldsymbol{x_n} y_n - \frac{P_{n-1}^{-1}\boldsymbol{x_n}\boldsymbol{x_n}^T P_{n-1}^{-1}\boldsymbol{x_n} y_n}{1 + \boldsymbol{x_n}^T P_{n-1}^{-1}\boldsymbol{x_n}}\right) =$$

$$\boldsymbol{\theta_{n-1}} - \frac{P_{n-1}^{-1}\boldsymbol{x_n}\boldsymbol{x_n}^T\boldsymbol{\theta_{n-1}}}{1 + \boldsymbol{x_n}^T P_{n-1}^{-1}\boldsymbol{x_n}} + \frac{P_{n-1}^{-1}\boldsymbol{x_n} y_n}{1 + \boldsymbol{x_n}^T P_{n-1}^{-1}\boldsymbol{x_n}} = \boldsymbol{\theta_{n-1}} + \frac{P_{n-1}^{-1}\boldsymbol{x_n}}{1 + \boldsymbol{x_n}^T P_{n-1}^{-1}\boldsymbol{x_n}}(y_n - \boldsymbol{x_n}^T\boldsymbol{\theta_{n-1}}).$$

Επομένως, η λύση $\boldsymbol{\theta_n}$ υπολογίζεται αναδρομικά από τον τύπο

$$\boldsymbol{\theta_n} = \boldsymbol{\theta_{n-1}} + \frac{P_{n-1}^{-1}\boldsymbol{x_n}}{1 + \boldsymbol{x_n}^T P_{n-1}^{-1}\boldsymbol{x_n}} e_n,$$

όπου $e_n = y_n - \boldsymbol{\theta_{n-1}^T}\boldsymbol{x_n}$ είναι το σφάλμα της εκτίμησης στο βήμα n.

Ο Κώδικας **Recursive Least Squares (RLS)**

Είσοδος: Τα δεδομένα $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^{N}$ και η παράμετρος εξομάλυνσης λ.
1: **Αρχικοποίηση**: $P = \lambda I_m,\ \boldsymbol{\theta} = \mathbf{0}$.
2: **for** i=1 to N
3: Υπολόγισε την έξοδο του συστήματος: $\hat{y}_i = \boldsymbol{\theta}^T \boldsymbol{x_i}$
4: Υπολόγισε το σφάλμα: $e_i = y_i - \hat{y}_i$
5: $r = 1 + \boldsymbol{x_i}^T P^{-1} \boldsymbol{x_i}$
6: $\boldsymbol{k} = r^{-1} P^{-1} \boldsymbol{x_i}$
7: Ενημέρωσε τη λύση: $\boldsymbol{\theta} = \boldsymbol{\theta} + e_i \boldsymbol{k}$
8: Ενημέρωσε τον αντίστροφο πίνακα: $P^{-1} = P^{-1} - \boldsymbol{k}\boldsymbol{k}^T r$
9: **end for**
Έξοδος: Το διάνυσμα $\boldsymbol{\theta} \in \mathbb{R}^m$.

2.6.1 Ο Αλγόριθμος Exponentially Weighted Recursive Least Squares (EWRLS)

Ο αλγόριθμος RLS χρησιμοποιεί πληροφορίες από όλα τα προηγούμενα δεδομένα σε κάθε επανάληψη. Παρότι αυτή η τακτική επιτρέπει στον RLS να επιτύχει καλύτερη ταχύτητα σύγκλισης, μπορεί να αποτελέσει τροχοπέδη σε χρονικά μεταβαλλόμενα περιβάλλοντα. Τότε, θα ήταν προτιμότερο να «ξεχαστούν» κάποια από τα προηγούμενα δεδομένα τα οποία αντιστοιχούν σε ένα πρώιμο στάδιο της διαδικασίας εκπαίδευσης. Αυτό γίνεται χρησιμοποιώντας κάποιους συντελεστές βαρύτητας (weighting factors) στη συνάρτηση κόστους:

$$\mathcal{L}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} w(n,i) \left| y_i - f_{\boldsymbol{\theta}}(\boldsymbol{x_i}) \right|^2 + \lambda w(n) \left\| \boldsymbol{\theta} \right\|^2,\ n = 1, 2, \ldots, N,$$

όπου w είναι κάποια κατάλληλα επιλεγμένη συνάρτηση βάρους. Χρησιμοποιείται πολύ συχνά ο **εκθετικός συντελεστής βαρύτητας** (**exponential weighting factor** ή **forgetting factor**), όπου επιλέγονται $w(n,i) = w^{n-i}$ και $w(n) = w^n$, για κάποιο $0 < w \leq 1$. Μικρές τιμές για το w (κοντά στο 0) επιβάλλουν έναν ισχυρό μηχανισμό λήθης, ο οποίος καθιστά τον αλγόριθμο πιο ευαίσθητο σε πρόσφατα δεδομένα και ενδέχεται να οδηγήσει σε κακές επιδόσεις. Έτσι, συνήθως το w επιλέγεται κοντά στο 1. Παρατηρήστε ότι για $w = 1$ προκύπτει ο απλός αλγόριθμος RLS.
Στον αλγόριθμο Exponentially Weighted RLS (EWRLS) ελαχιστοποιούμε τη συνάρτηση κόστους:

$$\mathcal{L}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} w^{n-i} \left| y_i - \boldsymbol{\theta}^T \boldsymbol{x_i} \right|^2 + \lambda w^n \left\| \boldsymbol{\theta} \right\|^2,\ n = 1, 2, \ldots, N.$$

Αν γράψουμε τις σχέσεις χρησιμοποιώντας πίνακες, η συνάρτηση κόστους γίνεται

$$\mathcal{L}(\boldsymbol{\theta}) = \left\| W_n^{1/2} (\boldsymbol{y_{(n)}} - X_n \boldsymbol{\theta}) \right\|^2 + w^n \lambda \boldsymbol{\theta}^T \boldsymbol{\theta},$$

όπου $X_n = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}) \in \mathbb{R}^{m \times n}$, $\boldsymbol{y_{(n)}} = (y_1, y_2, \ldots, y_n)^T \in \mathbb{R}^n$, για $n = 1, 2, \ldots, N$, και ο W είναι ένας διαγώνιος πίνακας με στοιχεία τις φθίνουσες δυνάμεις του w ξεκινώντας από την $n-1$: $W_n = diag\{w^{n-1}, w^{n-2}, \ldots, 1\}$, $n = 1, 2, \ldots, N$. Η λύση στο αντίστοιχο πρόβλημα ελαχιστοποίησης είναι:

$$\boldsymbol{\theta_n} = (X_n W_n X_n^T + w^n \lambda I_m)^{-1} X_n W_n \boldsymbol{y_{(n)}}.$$

Εργαζόμενοι όπως στην ενότητα (2.6) παίρνουμε:

$$P_n^{-1} = w^{-1} P_{n-1}^{-1} - \frac{w^{-2} P_{n-1}^{-1} \boldsymbol{x_n} \boldsymbol{x_n}^T P_{n-1}^{-1}}{1 + w^{-1} \boldsymbol{x_n}^T P_{n-1}^{-1} \boldsymbol{x_n}}$$

και

$$\boldsymbol{\theta_n} = \boldsymbol{\theta_{n-1}} + \frac{w^{-1} P_{n-1}^{-1} \boldsymbol{x_n}}{1 + w^{-1} \boldsymbol{x_n}^T P_{n-1}^{-1} \boldsymbol{x_n}} e_n.$$

Παρατίθεται παρακάτω αναλυτικά ο κώδικας για τον αλγόριθμο EWRLS.

Ο Κώδικας **Exponentially Weighted Recursive Least Squares (EWRLS)**

Είσοδος: Τα δεδομένα $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^N$, το βάρος w και η παράμετρος εξομάλυνσης λ.
1: **Αρχικοποίηση**: $P = \lambda I_m$, $\boldsymbol{\theta} = \boldsymbol{0}$.
2: **for** i=1 to N
3: Υπολόγισε την έξοδο του συστήματος: $\hat{y}_i = \boldsymbol{\theta}^T \boldsymbol{x_i}$
4: Υπολόγισε το σφάλμα: $e_i = y_i - \hat{y}_i$
5: $r = 1 + w^{-1} \boldsymbol{x_i}^T P^{-1} \boldsymbol{x_i}$
6: $\boldsymbol{k} = (rw)^{-1} P^{-1} \boldsymbol{x_i}$
7: Ενημέρωσε τη λύση: $\boldsymbol{\theta} = \boldsymbol{\theta} + e_i \boldsymbol{k}$
8: Ενημέρωσε τον αντίστροφο πίνακα: $P^{-1} = w^{-1} P^{-1} - \boldsymbol{k} \boldsymbol{k}^T r$
9: **end for**
Έξοδος: Το διάνυσμα $\boldsymbol{\theta} \in \mathbb{R}^m$.

Κεφάλαιο 3

Εφαρμογές Μηχανικής Μάθησης σε Μη Γραμμικά Προβλήματα

Στο Κεφάλαιο αυτό θα δούμε μη γραμμικά προβλήματα Μηχανικής Μάθησης και θα διατυπώσουμε και θα σχολιάσουμε το kernel τέχνασμα, προτού το χρησιμοποιήσουμε σε συγκεκριμένους αλγορίθμους.
Όλες τις συναρτήσεις kernel που θα αναφέρουμε θα τις θεωρούμε πραγματικές. Ωστόσο, οι αλγόριθμοι που αξιοποιούν το kernel τέχνασμα έχουν πρόσφατα επεκταθεί και σε μιγαδικές συναρτήσεις.

3.1 Εισαγωγή

Στη μοντελοποίηση, υιοθετείται ένα μοντέλο και μια μέθοδος εκπαίδευσης ώστε να προσδιοριστούν οι παράμετροι του μοντέλου. Όσον αφορά τη μη γραμμική περίπτωση, είναι φανερό ότι διαφορετικές μη γραμμικές συναρτήσεις παράγουν διαφορετικά μοντέλα. Θα μεταχειριστούμε με ενιαίο τρόπο μία μεγάλη κλάση από μη γραμμικά προβλήματα, που απαντώνται συχνά στην πράξη, αξιοποιώντας τη θεωρία των **Reproducing Kernel Hilbert Spaces (RKHS)**. Αυτή η ιδέα αναπτύχθηκε σταδιακά, από τις αρχές του 20ού αιώνα, στο χώρο της Συναρτησιακής Ανάλυσης [3] και έγινε γνωστή στο χώρο της Μηχανικής Μάθησης πιο πρόσφατα, όταν αξιοποιήθηκε στα πλαίσια της μελέτης των Διανυσμάτων Υποστήριξης (Support Vector Machines).
Η μέθοδος που αναφέρθηκε παραπάνω επιτρέπει να χειριστεί κανείς με ενιαίο τρόπο μία μεγάλη κλάση από μη γραμμικά προβλήματα λύνοντας ένα ισοδύναμο γραμμικό πρόβλημα σε ένα διαφορετικό χώρο από αυτόν όπου βρίσκονται τα πραγματικά δεδομένα υιοθετώντας μία ειδική απεικόνιση από το χώρο εισόδων σε έναν άλλο χώρο. Έχει μεγάλη σημασία πως όσον αφορά τους RKHS δε χρειάζεται να μας απασχολεί ούτε η ακριβής φύση του χώρου ούτε καν η διάστασή του. Από τη στιγμή που η παραμετρική εκπαίδευση έχει ολοκληρωθεί, ο σχεδιαστής μπορεί να δοκιμάσει διαφορετικές απεικονίσεις, που αντιστοιχούν σε διαφορετικά μη γραμμικά μοντέλα, και να κρατήσει αυτή που καλύπτει καλύτερα τις ανάγκες του.

Ας υπενθυμίσουμε την τυπική περιγραφή για το πρόβλημα εκτίμησης παραμέτρων.

- Δίνεται ένα σύνολο N δεδομένων εκπαίδευσης $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^{N}$, όπου $\boldsymbol{x_n} \in \mathbb{R}^m$, $y_n \in Y \subseteq \mathbb{R}$.
- Υιοθετούμε μια παραμετρική κλάση συναρτήσεων
$$\mathcal{F} = \{f_{\boldsymbol{\theta}}(\cdot) : \boldsymbol{\theta} \in \mathcal{A} \subseteq \mathbb{R}^K\}.$$
- Επιλέγουμε μία συνάρτηση κόστους (loss function - cost function) από ένα σύνολο διαθέσιμων μη αρνητικών συναρτήσεων
$$\mathcal{L} : \mathbb{R} \times \mathbb{R} \to [0, +\infty)$$
- Υπολογίζουμε το $\boldsymbol{\theta_*}$ ούτως ώστε να ελαχιστοποιήσουμε τη συνολική απώλεια επάνω σε όλα τα σημεία από τα εκπαιδευτικά δεδομένα, δηλαδή
$$f_{\boldsymbol{\theta_*}}(\cdot) : \boldsymbol{\theta_*} \in \underset{\boldsymbol{\theta} \in \mathcal{A}}{\operatorname{argmin}} J(\boldsymbol{\theta}),$$
όπου
$$J(\boldsymbol{\theta}) := \sum_{n=1}^{N} \mathcal{L}(y_n, f_{\boldsymbol{\theta}}(\boldsymbol{x_n})),$$
υποθέτοντας ότι υπάρχει ελάχιστο. Σε μία adaptive διαδικασία, η ελαχιστοποίηση της συνάρτησης κόστους γίνεται αναδρομικά στη μορφή
$$\boldsymbol{\theta_n} = \boldsymbol{\theta_{n-1}} + (\text{όρος διόρθωσης σφάλματος}),$$
καθώς το n αυξάνει.

Μπορεί κανείς να αντιμετωπίσει το overfitting χρησιμοποιώντας τη μέθοδο εξομάλυνσης (regularization). Έτσι, έχουμε
$$J(\boldsymbol{\theta}) := \sum_{n=1}^{N} \mathcal{L}(y_n, f_{\boldsymbol{\theta}}(\boldsymbol{x_n})) + \lambda \Omega(\|\boldsymbol{\theta}\|),$$
όπου $\Omega(\cdot)$ είναι μία γνησίως αύξουσα μη αρνητική συνάρτηση και η $\|\cdot\|$ είναι μια νόρμα, για παράδειγμα η ευκλείδεια νόρμα (L_2) ή η L_1 νόρμα. Συχνά επιλέγονται ως συναρτήσεις Ω οι παρακάτω:
$$\Omega(\|\boldsymbol{\theta}\|) = \sum_{k=1}^{K-1} |\theta_k|^2 \quad \text{ή} \quad \Omega(\|\boldsymbol{\theta}\|) = \sum_{k=1}^{K-1} |\theta_k| \,. \tag{3.1}$$

Επίσης, χρησιμοποιούνται πολύ συχνά τεχνικές αραίωσης (sparsification) για να αποφευχθεί το overfitting.

3.2 Μη Γραμμική Παλινδρόμηση και Κατηγοριοποίηση

Ας δούμε τα προβλήματα της παλινδρόμησης και της κατηγοριοποίησης για τη μη γραμμική περίπτωση.

Στη **μη γραμμική παλινδρόμηση (nonlinear regression)**, έχοντας ένα σύνολο N δεδομένων εκπαίδευσης $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^{N}$, $\boldsymbol{x_n} \in \mathbb{R}^m$, $y_n \in \mathbb{R}$, επιθυμούμε να εντοπίσουμε τη σχέση εισόδου-εξόδου μέσω ενός μοντέλου της μορφής

$$y_n = f(\boldsymbol{x_n}) + \eta_n, \; n = 1, 2, \ldots, N \tag{3.2}$$

Η ακολουθία η_n είναι κάποιος θόρυβος, μη παρατηρήσιμος, και η μη γραμμική συνάρτηση $f(\cdot)$ μοντελοποιείται στη μορφή

$$f(\boldsymbol{x}) = \theta_0 + \sum_{k=1}^{K-1} \theta_k \phi_k(\boldsymbol{x}), \tag{3.3}$$

όπου οι θ_k με $k = 0, 1, 2, \ldots, K-1$, είναι οι άγνωστες παράμετροι και οι $\phi_k(\boldsymbol{x})$ είναι προεπιλεγμένες μη γραμμικές συναρτήσεις $\phi_k(\cdot) : \mathbb{R}^m \to \mathbb{R}$, όπου $k = 1, 2, \ldots, K-1$. Συνδυάζοντας τη σχέση (3.2) με την (3.3) έχουμε $y_n = \theta_0 + \sum_{k=1}^{K-1} \theta_k \phi_k(\boldsymbol{x}) + \eta_n$ και γράφουμε

$$y_n = \boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x_n}) + \eta_n,$$

όπου $\boldsymbol{\phi}(\cdot) = [\phi_1(\cdot), \phi_2(\cdot), \ldots, \phi_{K-1}(\cdot), 1]^T$ και $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_{K-1}, \theta_0]^T$. Το θ_0 είναι γνωστό ως **μεροληψία (bias)** και έχει απορροφηθεί στο διάνυσμα $\boldsymbol{\theta}$ προσθέτοντας το 1 ως το τελευταίο στοιχείο του διανύσματος $\boldsymbol{\phi}$.

Στόχος του προβλήματος της εκτίμησης παραμέτρων είναι να πάρουμε εκτιμήσεις $\hat{\boldsymbol{\theta}}$ για το $\boldsymbol{\theta}$ χρησιμοποιώντας τα διαθέσιμα δεδομένα εκπαίδευσης. Αφού βρούμε το $\hat{\boldsymbol{\theta}}$ και δώσουμε κάποια τιμή στο $\boldsymbol{x}$, η πρόβλεψη για την αντίστοιχη τιμή εξόδου υπολογίζεται σύμφωνα με το μοντέλο

$$\hat{y} = \hat{\boldsymbol{\theta}}^T \phi(\boldsymbol{x}). \tag{3.4}$$

Στην **κατηγοριοποίηση (classification)**, θυμίζουμε ότι οι μεταβλητές εξόδου είναι διακριτές, δηλαδή $y_n \in D$, όπου το σύνολο D είναι πεπερασμένο σύνολο με διακριτές τιμές.

Στόχος της κατηγοριοποίησης είναι να σχεδιαστεί ένας ταξινομητής (classifier), δηλαδή μία συνάρτηση f (ή στη γενική περίπτωση ένα σύνολο συναρτήσεων), έτσι ώστε το αντίστοιχο (υπερ-) επίπεδο $f(\boldsymbol{x}) = 0$ στο χώρο των $\boldsymbol{x}$ να διαχωρίζει τα σημεία που ανήκουν σε διαφορετικές κλάσεις με τον καλύτερο δυνατό τρόπο. Στην εργασία αυτή θα θεωρήσουμε συναρτήσεις της μορφής (3.3).

Ο σχεδιασμός ενός ταξινομητή, λοιπόν, αντιμετωπίζεται και πάλι ως ένα πρόβλημα εκτίμησης παραμέτρων. Αφού υπολογίζονται εκτιμήσεις $\hat{\boldsymbol{\theta}}$ για τις άγνωστες παραμέτρους $\boldsymbol{\theta}$, δεδομένου

ενός χαρακτηριστικού διανύσματος $\boldsymbol{x}$ - που προκύπτει από έναν κανόνα και η ετικέτα κλάσης του, y, είναι άγνωστη - η προβλεπόμενη ετικέτα λαμβάνεται ως

$$\hat{y} = g(f(\boldsymbol{x})), \tag{3.5}$$

όπου η **συνάρτηση ταξινόμησης** $g(\cdot)$ δείχνει σε ποια μεριά του (υπερ-) επιπέδου $f(\boldsymbol{x}) = 0$ βρίσκεται το $\boldsymbol{x}$. Συνήθως χρησιμοποιείται η $g(\cdot) = sgn(\cdot)$, δηλαδή η συνάρτηση ελέγχου προσήμου (sign function) με

$$sgn(\boldsymbol{x}) = \begin{cases} \dfrac{\boldsymbol{x}}{|\boldsymbol{x}|}, \text{ αν } \boldsymbol{x} \neq \mathbf{0} \\ 0, \text{ αν } \boldsymbol{x} = \mathbf{0}. \end{cases}$$

3.3 Το Kernel Τέχνασμα

Αναφέρουμε παρακάτω το kernel τέχνασμα, καθώς και δύο θεωρήματα που θα χρειαστούμε στη συνέχεια του Κεφαλαίου.

Έστω $\mathcal{H} \subseteq \mathcal{F}(X, \mathbb{R})$ ένας RKHS ενός συνόλου X με reproducing kernel τη συνάρτηση $K : X \times X \to \mathbb{R}$. Έστω η απεικόνιση

$$\Phi : X \to \mathcal{H}, \quad \Phi(\boldsymbol{x}) = k_{\boldsymbol{x}},$$

η οποία ονομάζεται **χαρακτηριστική απεικόνιση** και απεικονίζει κάθε στοιχείο του X στη reproducing kernel συνάρτηση για το σημείο αυτό. Έστω τώρα ότι $\boldsymbol{x_1}, \boldsymbol{x_2} \in \mathbb{R}^m$ και ότι $\Phi(\boldsymbol{x_1}), \Phi(\boldsymbol{x_2})$ είναι οι αντίστοιχες εικόνες στον $\mathcal{H}$. Όπως έχουμε ήδη δει στο 1ο Κεφάλαιο, ισχύει $\langle k_{\boldsymbol{x}}, k_{\boldsymbol{y}} \rangle_{\mathcal{H}} = K(\boldsymbol{x}, \boldsymbol{y})$. Επομένως, έχουμε τη σχέση:

$$\langle \Phi(\boldsymbol{x_1}), \Phi(\boldsymbol{x_2}) \rangle_{\mathcal{H}} = K(\boldsymbol{x_1}, \boldsymbol{x_2}).$$

Αυτή είναι μια αξιοσημείωτη ιδιότητα, καθώς μας επιτρέπει να υπολογίζουμε αντί για εσωτερικά γινόμενα σε έναν μεγάλης διάστασης χώρο RKHS τις τιμές μιας συνάρτησης στον μικρότερης διάστασης χώρο εισόδων. Επίσης, η ιδιότητα αυτή επιτρέπει να γίνει η παρακάτω προσέγγιση, η οποία έχει χρησιμοποιηθεί εκτενώς στη Μηχανική Μάθηση.

Το Kernel Τέχνασμα. *Αν σχεδιάσουμε έναν αλγόριθμο στο χώρο εισόδων ο οποίος να υπολογίζει τις εκτιμήσεις των παραμέτρων ενός γραμμικού μοντέλου και στους υπολογισμούς να χρησιμοποιεί μόνο εσωτερικά γινόμενα, τότε μπορούμε να αντικαταστήσουμε κάθε ένα από τα εσωτερικά γινόμενα στον αλγόριθμο με τον υπολογισμό της αριθμητικής τιμής μίας συνάρτησης kernel.*

Έτσι, ένα μη γραμμικό πρόβλημα εντός του συνόλου X μετατρέπεται σε ένα γραμμικό πρόβλημα εντός του χώρου $\mathcal{H}$. Ο νέος αλγόριθμος είναι ισοδύναμος με την επίλυση του γραμμικού προβλήματος εκτίμησης παραμέτρων στον RKHS που αντιστοιχεί στην επιλεγμένη kernel συνάρτηση. Διαφορετικές συναρτήσεις kernel αντιστοιχούν σε διαφορετικούς χώρους RKHS

και άρα σε διαφορετικά μη γραμμικά μοντέλα - κάθε εσωτερικό γινόμενο στον RKHS είναι ένας υπολογισμός στο χώρο εισόδων της τιμής μιας μη γραμμικής συνάρτησης. Το Kernel Τέχνασμα χρησιμοποιήθηκε πρώτη φορά στο [1], [12].

Το παρακάτω Θεώρημα (Representer Theorem) έχει μεγάλη σημασία σε πρακτικές εφαρμογές. Εξασφαλίζει ότι, ακόμα και στην περίπτωση που προσπαθεί κανείς να λύσει ένα πρόβλημα βελτιστοποίησης σε έναν άπειρης διάστασης RKHS $\mathcal{H}$ (όπως αυτός που παράγεται από την Gaussian kernel), η λύση του προβλήματος βρίσκεται στη γραμμική θήκη N συγκεκριμένων kernel συναρτήσεων, αυτών που κεντρικοποιούνται στα N σημεία εκπαίδευσης.

Θεώρημα 3.3.1 (Representer Theorem). *Έστω $\Omega : [0, +\infty) \to \mathbb{R}$ μια γνησίως αύξουσα συνάρτηση, X ένα μη κενό σύνολο και $\mathcal{L} : X \times \mathbb{R}^2 \to \mathbb{R} \cup \{\infty\}$ μια αυθαίρετη συνάρτηση κόστους. Τότε κάθε ελαχιστοποιητής $f \in \mathcal{H}$ του regularized προβλήματος ελαχιστοποίησης*

$$\min_f \ \big(\ \mathcal{L}(\ (x_1, y_1, f(x_1)), \ldots, (x_N, y_N, f(x_N)) \ ) + \Omega(\|f\|_{\mathcal{H}}^2) \ \big)$$

επιδέχεται αναπαράσταση της μορφής

$$f = \sum_{n=1}^{N} \theta_n K(\cdot, x_n),$$

όπου $\theta_n \in \mathbb{R}$ για $n = 1, 2, ..., N$.

Παραδείγματα συναρτήσεων κόστους $\mathcal{L}$ είναι τα εξής:

(i) $\mathcal{L}((x_1, y_1, f(x_1)), \ldots, (x_N, y_N, f(x_N))) = \sum_{n=1}^{N} (f(x_n) - y_n)^2$

(ii) $\mathcal{L}((x_1, y_1, f(x_1)), \ldots, (x_N, y_N, f(x_N))) = \sum_{n=1}^{N} |f(x_n) - y_n|$

Παρατηρήστε ότι στη συνάρτηση Ω συνήθως η μεροληψία, θ_0, δεν περιλαμβάνεται στη νόρμα (σχέση (3.1)), καθώς μπορεί να καταστήσει τη λύση ευαίσθητη σε στροφές και μετατοπίσεις ως προς την αρχή των αξόνων.
Επίσης, στην πράξη συχνά περιλαμβάνουμε έναν **παράγοντα μεροληψίας** στη λύση regularized προβλημάτων ελαχιστοποίησης που βασίζονται σε kernels, δηλαδή υποθέτουμε ότι η f επιδέχεται αναπαράσταση της μορφής

$$f = \sum_{n=1}^{N} \theta_n K(\cdot, x_n) + b,$$

όπου $b \in \mathbb{R}$. Αυτό έχει αποδειχθεί ότι βελτιώνει την επίδοση των αντίστοιχων αλγορίθμων [37], [35], [13] για δύο λόγους. Πρώτον, η μεροληψία b διευρύνει την οικογένεια συναρτήσεων στην οποία αναζητάμε τη λύση οδηγώντας σε ενδεχομένως καλύτερες εκτιμήσεις. Επιπλέον,

καθώς ο παράγοντας $\Omega(\|f\|^2_{\mathcal{H}})$ σταθμίζει τις τιμές της f στα σημεία εκπαίδευσης, η λύση τείνει να πάρει τιμές όσο το δυνατόν πιο κοντά στο 0 για μεγάλες τιμές του λ. Η χρήση του παράγοντα μεροληψίας αιτιολογείται θεωρητικά από το παρακάτω Θεώρημα.

Θεώρημα 3.3.2 (**Semi-parametric Representer Theorem**). *Έστω ότι ισχύουν οι υποθέσεις του Θεωρήματος (3.3.1). Έστω επίσης ένα σύνολο M πραγματικών συναρτήσεων $\{\psi_m\}_{m=1}^M$, όπου $\psi_m : X \to \mathbb{R},\ m = 1, 2, \ldots, M$, για τις οποίες ο $N \times M$ πίνακας $(\psi_m(x_n))_{n,m}$ έχει τάξη M. Τότε κάθε λύση $\tilde{f} := f + h$, όπου $f \in \mathcal{H}, h \in span\{\psi_1, \psi_2, \ldots, \psi_M\}$, του προβλήματος*

$$\min_{\tilde{f}} \left(\mathcal{L}(\ (x_1, y_1, \tilde{f}(x_1)), \ldots, (x_N, y_N, \tilde{f}(x_N))\ ) + \Omega(\|f\|^2_{\mathcal{H}}) \right)$$

επιδέχεται αναπαράσταση της μορφής

$$\tilde{f} = \sum_{n=1}^{N} \theta_n K(\cdot, x_n) + \sum_{m=1}^{M} b_m \psi_m(\cdot),$$

όπου $\theta_n \in \mathbb{R},\ b_m \in \mathbb{R}$ για $n = 1, 2, ..., N,\ m = 1, 2, \ldots, M$.

3.4 Απεικονίζοντας ένα Μη Γραμμικό σε ένα Γραμμικό Πρόβλημα

Ας δούμε το ανάπτυγμα στη σχέση (3.3) σε συνδυασμό με τα μοντέλα στην (3.4) για την παλινδρόμηση και στην (3.5) για την κατηγοριοποίηση. Παρατηρούμε ότι αν απεικονίσουμε τον αρχικό χώρο εισόδων διάστασης m σε ένα νέο M-διάστατο χώρο,

$$\mathbb{R}^m \ni \boldsymbol{x} \mapsto \boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^M,$$

αναμένουμε το πρόβλημά μας να μοντελοποιείται αρκετά καλά υιοθετώντας ένα γραμμικό μοντέλο σε αυτόν το νέο χώρο. Μένει μόνο να επιλέξουμε εμείς τις κατάλληλες συναρτήσεις $\phi_k(\cdot)$, καθώς και την κατάλληλη διάσταση M του νέου χώρου.
Ειδικά για την κατηγοριοποίηση, αυτή η διαδικασία έχει στέρεο θεωρητικό υπόβαθρο το Θεώρημα του Cover.

Θεώρημα 3.4.1 (**Θεώρημα Cover**). *Έστω N το πλήθος σημεία σε τυχαίες θέσεις σε έναν m-διάστατο χώρο και έστω μια απεικόνιση που απεικονίζει αυτά τα σημεία σε ένα χώρο μεγαλύτερης διάστασης. Τότε η πιθανότητα τα σημεία να τοποθετηθούν στο νέο χώρο σε γραμμικά διαχωρίσιμες ομάδες τείνει στο 1 καθώς η διάσταση, M, του νέου χώρου τείνει στο άπειρο.*

Παρότι μια τέτοια μέθοδος μοιάζει ιδανική, έχει πρακτικές δυσκολίες. Για να γίνει ένα πρόβλημα γραμμικό, ίσως χρειάζεται να απεικονίσουμε τον αρχικό χώρο σε ένα χώρο με πολύ μεγάλη διάσταση, κάτι που μπορεί να οδηγήσει σε πολύ μεγάλο αριθμό άγνωστων παραμέτρων προς εκτίμηση. Επιλέγοντας, όμως, να απεικονίσουμε τα σημεία σε έναν RKHS, που μπορεί

να είναι και άπειρης διάστασης, οι υπολογισμοί γίνονται πολύ πιο εύκολα με τη βοήθεια του kernel τεχνάσματος.

Ας υποθέσουμε ότι $\mathcal{H}$ είναι ένας RKHS με αντίστοιχη reproducing kernel συνάρτηση την $K : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$. Ας θεωρήσουμε την απεικόνιση

$$\mathbb{R}^m \ni \boldsymbol{x} \mapsto \phi_k(\boldsymbol{x}) \in \mathcal{H},$$

όπου $\phi_k(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{x_k}),\ k = 1, 2, \ldots, N$. Τότε η (3.3) γίνεται

$$f(\boldsymbol{x}) = \theta_0 + \sum_{k=1}^{N} \theta_k K(\boldsymbol{x}, \boldsymbol{x_k}).$$

Το ανάπτυγμα, λοιπόν, της μη γραμμικής συνάρτησης f περιλαμβάνει ακριβώς τόσους όρους όσα είναι τα σημεία εκπαίδευσης και κάθε όρος συσχετίζεται με ένα σημείο.
Τότε μπορεί να εφαρμοστεί το kernel τέχνασμα και, βεβαίως, ισχύει το Representer Theorem, που διαμορφώνεται ως εξής:

Representer Theorem. Έστω μία συνάρτηση κόστους

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{N} \mathcal{L}(y_n, f_{\boldsymbol{\theta}}(\boldsymbol{x_n})) + \lambda\Omega(\|\boldsymbol{\theta}\|)$$

και υποθέτουμε ότι οι λύσεις βρίσκονται σε έναν RKHS με reproducing kernel συνάρτηση $K(\cdot, \cdot)$. Τότε ο ελαχιστοποιητής της συνάρτησης κόστους J γράφεται στη μορφή

$$f(\cdot) = \sum_{n=1}^{N} \theta_n K(\cdot, \boldsymbol{x_n}) + \theta_0.$$

Αξιοποιώντας τα παραπάνω, θα δούμε στη συνέχεια πώς διαμορφώνονται οι αλγόριθμοι που αναφέρθηκαν στο Κεφάλαιο 2 με τη βοήθεια της θεωρίας των RKHS.

3.5 Ο Αλγόριθμος Kernel Least Mean Squares (KLMS)

Έστω ένας RKHS $\mathcal{H}(K)$. Στους αλγορίθμους που βασίζονται σε kernels η έξοδος εκτιμάται μέσω μιας μη γραμμικής συνάρτησης $f \in \mathcal{H}$. Συγκεκριμένα, η ακολουθία παραδειγμάτων $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^{N}$ μετασχηματίζεται στην $\{(\Phi(\boldsymbol{x_n}), y_n)\}_{n=1}^{N}$ μέσω της χαρακτηριστικής απεικόνισης $\Phi : X \to \mathcal{H},\ \Phi(\boldsymbol{x}) = K(\cdot, \boldsymbol{x})$, ενώ η έξοδος που εκτιμάται στο βήμα n παίρνει τη μορφή $\hat{y}_n = \langle \Phi(\boldsymbol{x_n}), f \rangle_{\mathcal{H}}$, για κάποια συνάρτηση $f \in \mathcal{H}$. Παρότι αυτός είναι ένας γραμμικός αλγόριθμος στον $\mathcal{H}$, αντιστοιχεί σε μια μη γραμμική διαδικασία στο χώρο εισόδων X. Ο μηχανισμός του συνοψίζεται ως εξής:

- Αναζητάμε συνάρτηση $f \in \mathcal{H}$ τέτοια ώστε να ελαχιστοποιείται η συνάρτηση κόστους

$$\mathcal{L}(f) = E[|y_n - \langle \Phi(\boldsymbol{x_n}), f \rangle_{\mathcal{H}}|^2].$$

- Θέτουμε $e_n = y_n - \langle \Phi(\boldsymbol{x_n}), f_{n-1} \rangle_{\mathcal{H}}$, παραγωγίζουμε την $\mathcal{L}$ κατά Fréchet και προσεγγίζουμε με την τιμή της για κάθε χρονική στιγμή n

$$\nabla \mathcal{L}(f) = -2E[e_n \Phi(\boldsymbol{x_n})] \approx -2e_n \Phi(\boldsymbol{x_n}).$$

- Παίρνουμε τελικά προς την κατεύθυνση της ελαχιστοποίησης

$$f_n = f_{n-1} + \mu e_n \Phi(\boldsymbol{x_n}).$$

- Εύκολα προκύπτει ότι

$$f_n = \mu \sum_{i=1}^{n} e_i \Phi(\boldsymbol{x_i}). \tag{3.6}$$

Υποθέτοντας ότι $f_0 = 0$ (η μηδενική συνάρτηση), σε κάθε βήμα n θα έχουμε:

$$\hat{y}_n = \langle \Phi(\boldsymbol{x_n}), f_{n-1} \rangle_{\mathcal{H}} = \left\langle \Phi(\boldsymbol{x_n}), \mu \sum_{i=1}^{n-1} e_i \Phi(\boldsymbol{x_i}) \right\rangle_{\mathcal{H}} = \mu \sum_{i=1}^{n-1} e_i \langle \Phi(\boldsymbol{x_n}), \Phi(\boldsymbol{x_i}) \rangle_{\mathcal{H}} = \mu \sum_{i=1}^{n-1} e_i K(\boldsymbol{x_n}, \boldsymbol{x_i}).$$

Άρα έχουμε

$$\hat{y}_n = \mu \sum_{i=1}^{n-1} e_i K(\boldsymbol{x_i}, \boldsymbol{x_n}). \tag{3.7}$$

Είναι σημαντικό να τονιστεί ότι, σε αντίθεση με τον LMS, όπου η λύση είναι ένα διάνυσμα, στον KLMS η λύση είναι μία συνάρτηση στο χώρο $\mathcal{H}$ και άρα δεν μπορεί να αναπαρασταθεί από μια μηχανή. Ωστόσο, όπως είναι σύνηθες στις μεθόδους που βασίζονται σε kernels, αυτό που χρειάζεται κανείς είναι η εκτιμώμενη έξοδος, $\hat{y}_n$, που γράφεται συναρτήσει των σημείων εισόδου, $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_{n-1}}$ και $\boldsymbol{x_n}$, (σχέση (3.7)) και όχι η πραγματική λύση (3.6). Έτσι, ο αλγόριθμος αποθηκεύει στη μνήμη μόνο τα εξής:

(i) τα **κέντρα** (σημεία εισόδου), $\boldsymbol{x_i}$, του αναπτύγματος (3.7), τα οποία αποθηκεύονται σε ένα **λεξικό** D και

(ii) τους συντελεστές, μe_i, του αναπτύγματος (3.7) που αποθηκεύονται σε ένα διάνυσμα $\boldsymbol{\theta}$.

Παρατηρήστε ότι η σχέση (3.6) συμφωνεί με όσα ξέρουμε από το Representer Theorem (3.3.1).

Ο Κώδικας **Kernel Least Mean Squares (KLMS)**

Είσοδος: Τα δεδομένα $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^{N}$, η παράμετρος μ του βήματος εκμάθησης και οι παράμετροι της kernel συνάρτησης.
1: **Αρχικοποίηση**: $\boldsymbol{\theta} = \mathbf{0}, D = \{\}$.
2: **for** i=1 to N
3: **if** i=1
4: $\hat{y}_i = 0$
5: **else**
6: Υπολόγισε την έξοδο του συστήματος: $\hat{y}_i = \sum_{j=1}^{i-1} \theta_j K(\boldsymbol{u_j}, \boldsymbol{x_i})$
7: **end if**
8: Υπολόγισε το σφάλμα: $e_i = y_i - \hat{y}_i$
9: $\theta_i = \mu e_i$
10: Καταχώρισε το νέο κέντρο $\boldsymbol{u_i} = \boldsymbol{x_i}$ στη λίστα με τα κέντρα: $D = D \cup \{\boldsymbol{u_i}\}$ και $\boldsymbol{\theta} = (\boldsymbol{\theta}^T, \theta_i)^T$
11: **end for**
Έξοδος: Το διάνυσμα $\boldsymbol{\theta} \in \mathbb{R}^N$ και το λεξικό $D = \{\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_N}\}$.

Μπορούμε και εδώ να χρησιμοποιήσουμε μια κανονικοποιημένη μορφή αντί για τη σχέση $f_n = f_{n-1} + \mu e_n \Phi(\boldsymbol{x_n})$, για παράδειγμα την:

$$f_n = f_{n-1} + \frac{\mu e_n}{K(\boldsymbol{x_n}, \boldsymbol{x_n})} \Phi(\boldsymbol{x_n}),$$

όπου $\mu \in (0, 2)$, λαμβάνοντας έτσι τον **Normalized KLMS (NKLMS)**. Στον αλγόριθμο, αυτό υλοποιείται με αντικατάσταση του βήματος 9 με το $\theta_i = \frac{\mu e_i}{\kappa}$, όπου το $\kappa = K(\boldsymbol{x_i}, \boldsymbol{x_i})$ 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 απευθείας 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 (CKLMS)**, ο οποίος δρα απευθείας σε μιγαδικούς RKHS χώρους. Περισσότερες πληροφορίες μπορεί κανείς να βρει στα [14], [15], [16].

3.5.1 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 όσο

εκτελείται ο αλγόριθμος να απαιτείται ολοένα και περισσότερη μνήμη καθώς και υπολογιστική ισχύς. Έτσι, δεν είναι δυνατόν να χρησιμοποιηθεί αυτούσιος ο KLMS σε πραγματικά προβλήματα, καθώς το λεξικό μεγαλώνει απεριόριστα.
Υπάρχουν, ωστόσο, διαθέσιμες ορισμένες στρατηγικές αντιμετώπισης του φαινομένου αυτού, που οδηγούν σε αραιές λύσεις. Αυτό το επιτυγχάνουν διαμορφώνοντας το λεξικό ως ένα βαθμό κατά τα πρώτα στάδια του αλγορίθμου, στη συνέχεια όμως επιτρέπουν σε νέα κέντρα να προστίθενται μόνο αν πληρούν συγκεκριμένα κριτήρια. Η γενική δομή ενός τέτοιου αλγορίθμου είναι η ακόλουθη.

Ο Κώδικας **Kernel Least Mean Squares (KLMS) Με Αραίωση**

Είσοδος: Τα δεδομένα $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^{N}$, η παράμετρος μ του βήματος εκμάθησης και οι παράμετροι της kernel συνάρτησης.
1: **Αρχικοποίηση**: $\boldsymbol{\theta} = \mathbf{0}, D = \{\}, M = 0$.
2: **for** i=1 to N
3: **if** i=1
4: $\hat{y}_i = 0$
5: **else**
6: Υπολόγισε την έξοδο του συστήματος: $\hat{y}_i = \sum_{j=1}^{M} \theta_j K(\boldsymbol{u_j}, \boldsymbol{x_i})$
7: **end if**
8: Υπολόγισε το σφάλμα: $e_i = y_i - \hat{y}_i$
9: $\theta_i = \mu e_i$
Έλεγχος προϋποθέσεων αραίωσης
10: **if** Οι Προϋποθέσεις Αραίωσης ικανοποιούνται
11: $M = M + 1$
12: Καταχώρισε το νέο κέντρο $\boldsymbol{u_M} = \boldsymbol{x_i}$ στη λίστα με τα κέντρα: $D = D \cup \{\boldsymbol{u_M}\}$ και $\boldsymbol{\theta} = (\boldsymbol{\theta}^T, \theta_i)^T$
13: **end if**
14: **end for**
Έξοδος: Το διάνυσμα $\boldsymbol{\theta} \in \mathbb{R}^M$ και το λεξικό $D = \{\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_M}\}$.

Διάφορες δημοφιλείς στρατηγικές αραίωσης αναφέρονται παρακάτω:

- Platt's Novelty Criterion
- Στρατηγική Αραίωσης βάσει Συνεκτικότητας (Coherence Based Sparsification Strategy)
- Surprise Criterion

3.5.2 Quantized Kernel Least Mean Squares (QKLMS)

Ένα σημαντικό μειονέκτημα των παραπάνω μεθόδων αραίωσης είναι ότι διατηρούν, επ΄ αόριστον και απαράλλαχτες, τις παλαιότερες πληροφορίες (με τη μορφή των θ_i που συγκροτούν το $\boldsymbol{\theta}$) αδυνατώντας έτσι να αντεπεξέλθουν σε αλλαγές που ενδέχεται να επηρεάσουν το κανάλι. 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 παρουσιάζεται στον αλγόριθμο που ακολουθεί.

Ο Κώδικας **Quantized Kernel Least Mean Squares (QKLMS)**

Είσοδος: Τα δεδομένα $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^{N}$, η παράμετρος μ του βήματος εκμάθησης, οι παράμετροι της kernel συνάρτησης και το μέγεθος κβαντισμού δ.
1: **Αρχικοποίηση**: $\boldsymbol{\theta} = \mathbf{0}, D = \{\}, M = 0$.
2: **for** i=1 to N
3: **if** i=1
4: $\hat{y}_i = 0$
5: **else**
6: Υπολόγισε την έξοδο του συστήματος: $\hat{y}_i = \sum_{j=1}^{M} \theta_j K(\boldsymbol{u_j}, \boldsymbol{x_i})$
7: **end if**
8: Υπολόγισε το σφάλμα: $e_i = y_i - \hat{y}_i$
9: $\theta_i = \mu e_i$
10: Υπολόγισε την απόσταση $dist$ του $\boldsymbol{x_i}$ από το D: $dist = \min_{\boldsymbol{u_k} \in D} \|\boldsymbol{x_i} - \boldsymbol{u_k}\| = \|\boldsymbol{x_i} - \boldsymbol{u_{k_0}}\|$, για κάποιο $k_0 \in \{1, 2, \ldots, M\}$.
11: **if** $dist \geq \delta$
12: $M = M + 1$
13: Καταχώρισε το νέο κέντρο $\boldsymbol{u_M} = \boldsymbol{x_i}$ στη λίστα με τα κέντρα: $D = D \cup \{\boldsymbol{u_M}\}$ και $\boldsymbol{\theta} = (\boldsymbol{\theta}^T, \theta_i)^T$
14: **else**
15: Κράτησε το λεξικό D ως έχει και ενημέρωσε το συντελεστή θ_{k_0}: $\theta_{k_0} = \theta_{k_0} + \mu e_i$.
16: **end if**
17: **end for**
Έξοδος: Το διάνυσμα $\boldsymbol{\theta} \in \mathbb{R}^M$ και το λεξικό $D = \{\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_M}\}$.

Καθώς, λοιπόν, καταφθάνει κάθε νέα πληροφορία $\boldsymbol{x_n}$, ο αλγόριθμος αποφασίζει αν πρόκειται για νέο κέντρο ή για περιττό σημείο. Συγκεκριμένα, αν η απόσταση του $\boldsymbol{x_n}$ από το λεξικό D_n, όπως αυτό είναι διαμορφωμένο μέχρι εκείνη τη στιγμή, είναι μεγαλύτερη ή ίση από το μέγεθος κβαντισμού δ (κάτι που σημαίνει ότι το $\boldsymbol{x_n}$ δεν μπορεί να «κβαντοποιηθεί» σε κά-

ποιο 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 $D_{n+1} = D_n \cup \{\boldsymbol{x_n}\}$. 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, του $\boldsymbol{u_{k_0}} \in D_n$.

3.6 Ο Αλγόριθμος Kernel Recursive Least Squares (KRLS)

Όπως στην περίπτωση του αλγορίθμου KLMS, στον Kernel Recursive Least Squares (KRLS) εκτιμάται η έξοδος του συστήματος μέσω μιας συνάρτησης f ορισμένης σε ένα χώρο RKHS, $\mathcal{H}(K)$. Η ακολουθία εισόδων μετασχηματίζεται στην $\{(\Phi(\boldsymbol{x_n}), y_n)\}_{n=1}^{N}$ μέσω της χαρακτηριστικής απεικόνισης $\Phi : X \to \mathcal{H},\ \Phi(\boldsymbol{x}) = K(\cdot, \boldsymbol{x})$, και η συνάρτηση κόστους παίρνει τη μορφή

$$\mathcal{L}_n(f) = \sum_{i=1}^{n} |y_i - \langle \Phi(\boldsymbol{x_i}), f \rangle_{\mathcal{H}}|^2 + \lambda \|f\|_{\mathcal{H}}^2,\ n = 1, 2, \ldots, N.$$

Το Representer Theorem εξασφαλίζει ότι η λύση f_n του προβλήματος ελαχιστοποίησης $\underset{f \in \mathcal{H}}{\operatorname{argmin}} \mathcal{L}_n(f)$ βρίσκεται στον πεπερασμένης διάστασης υπόχωρο που παράγεται από τις n συναρτήσεις kernel που είναι κεντρικοποιημένες στα σημεία εκπαίδευσης $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$, δηλαδή

$$f_n = \sum_{i=1}^{n} \boldsymbol{\theta_{n,i}} K(\cdot, \boldsymbol{x_i}),\ n = 1, 2, \ldots, N.$$

Έτσι, με τη βοήθεια του kernel τεχνάσματος το αποτέλεσμα στην επανάληψη n εκτιμάται ως εξής:

$$\hat{y}_n = \langle \Phi(\boldsymbol{x_n}), f_n \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \boldsymbol{\theta_{n,i}} K(\boldsymbol{x_i}, \boldsymbol{x_n}),\ n = 1, 2, \ldots, N.$$

Καθώς το ανάπτυγμα μεγαλώνει απεριόριστα όσο το n αυξάνει, γίνεται φανερή και σε αυτή την περίπτωση η ανάγκη για ένα μηχανισμό αραίωσης που θα διατηρεί τα σημεία εκπαίδευσης που περιέχουν περισσότερη πληροφορία και θα απορρίπτει τα υπόλοιπα. Ο βασικός στόχος του αλγορίθμου KRLS είναι να εκτιμήσει αναδρομικά το (ολοένα μεγαλύτερης διάστασης) διάνυσμα $\boldsymbol{\theta_n} = (\theta_{n,1}, \theta_{n,2}, \ldots, \theta_{n,n})^T$.

Ο Κώδικας **Kernel Recursive Least Squares (KRLS)**

Είσοδος: Τα δεδομένα $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^{N}$, η παράμετρος εξομάλυνσης λ και οι παράμετροι της kernel συνάρτησης.

1: **Αρχικοποίηση**: $\tilde{K}^{-1} = \left[\frac{1}{\lambda + K(\boldsymbol{x_1}, \boldsymbol{x_1})}\right], \tilde{\boldsymbol{\theta}} = \left[\frac{y_1}{\lambda + K(\boldsymbol{x_1}, \boldsymbol{x_1})}\right], D = [\boldsymbol{x_1}]$.
2: **for** i=2 to N
3: $\quad \boldsymbol{\beta} = (K(\boldsymbol{x_1}, \boldsymbol{x_i}), \ldots, K(\boldsymbol{x_{i-1}}, \boldsymbol{x_i}))^T$
4: $\quad$ Υπολόγισε την έξοδο του συστήματος: $\hat{y}_i = \tilde{\boldsymbol{\theta}}^T \boldsymbol{\beta}$
5: $\quad$ Υπολόγισε το σφάλμα: $e_i = y_i - \hat{y}_i$
6: $\quad \boldsymbol{\alpha} = \tilde{K}^{-1}\boldsymbol{\beta}$
7: $\quad \delta = \lambda + K(\boldsymbol{x_i}, \boldsymbol{x_i}) - \boldsymbol{\beta}^T\boldsymbol{\alpha}$
8: $\quad$ Ενημέρωσε τον πίνακα $\tilde{K}^{-1} = \frac{1}{\delta}\begin{bmatrix} \delta\tilde{K}^{-1} + \boldsymbol{\alpha}\boldsymbol{\alpha}^T & -\boldsymbol{\alpha} \\ -\boldsymbol{\alpha}^T & 1 \end{bmatrix}$
9: $\quad$ Ενημέρωσε τη λύση $\tilde{\boldsymbol{\theta}} = \begin{bmatrix} \tilde{\boldsymbol{\theta}} - \frac{\boldsymbol{\alpha}}{\delta} e_i \\ \frac{e_i}{\delta} \end{bmatrix}$
10: $\quad$ Πρόσθεσε το νέο κέντρο στο λεξικό: $D = D \cup \{\boldsymbol{x_i}\}$
11: **end for**

Έξοδος: Το διάνυσμα $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^N$ και το λεξικό $D = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_N}\}$.

Η πρώτη προσέγγιση στον αλγόριθμο KRLS παρουσιάστηκε στο [19] από τους Engel, Mannor και Meir. Η μέθοδός τους αναπτύχθηκε με άξονα μια συγκεκριμένη στρατηγική αραίωσης που λέγεται **Approximate Linear Dependency (ALD)**. Σε αυτή τη στρατηγική αραίωσης, θεωρούμε ένα λεξικό $D_{n-1} = \{\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_{M_{n-1}}}\}$ που έχει μέγεθος M_{n-1} και περιέχει κάποια κέντρα από τα παλαιότερα σημεία εκπαίδευσης. Ελέγχουμε εάν το νέο σημείο $\Phi(\boldsymbol{x_n})$ είναι σχεδόν γραμμικά εξαρτημένο με τα στοιχεία του λεξικού, $\Phi(\boldsymbol{u_1}), \Phi(\boldsymbol{u_2}), \ldots, \Phi(\boldsymbol{u_{M_{n-1}}})$. Στην περίπτωση αυτή, το $\Phi(\boldsymbol{x_n})$ προσεγγίζεται από ένα γραμμικό συνδυασμό των στοιχείων του λεξικού, ειδάλλως προστίθεται στο λεξικό. Συγκεκριμένα, η συνθήκη είναι

$$\delta_n := \min_{\alpha_n} \left\{ \left\| \sum_{m=1}^{M_{n-1}} \alpha_{n,m}\Phi(\boldsymbol{u_m}) - \Phi(\boldsymbol{x_n}) \right\|^2 \right\} \leq \epsilon_0,$$

όπου το ϵ_0 είναι μία παράμετρος που καθορίζει ο χρήστης.
Χρησιμοποιώντας πίνακες, το πρόβλημα ελαχιστοποίησης γράφεται ισοδύναμα:

$$\delta_n = \min_{\boldsymbol{\alpha_n}} \left(\boldsymbol{\alpha_n}^T \tilde{K}_{n-1}\boldsymbol{\alpha_n} - 2\boldsymbol{\alpha_n}^T\boldsymbol{\beta_n} + K(\boldsymbol{x_n}, \boldsymbol{x_n}) \right),$$

όπου $(\tilde{K}_{n-1})_{i,j} = K(\boldsymbol{u_i}, \boldsymbol{u_j})$, για $i, j = 1, 2, \ldots, M_{n-1}$,
και $\boldsymbol{\beta_n} = (K(\boldsymbol{u_1}, \boldsymbol{x_n}), K(\boldsymbol{u_2}, \boldsymbol{x_n}), \ldots, K(\boldsymbol{u_{M_{n-1}}}, \boldsymbol{x_n}))^T$.
Υποθέτοντας ότι ο $\tilde{K}_{n-1}$ είναι αντιστρέψιμος, η λύση στο πρόβλημα ελαχιστοποίησης είναι

$$\boldsymbol{\alpha_n} = \tilde{K}_{n-1}^{-1}\boldsymbol{\beta_n},$$

ενώ το ελάχιστο δίνεται από τη σχέση

$$\delta_n = K(\boldsymbol{x_n}, \boldsymbol{x_n}) - \boldsymbol{\beta_n}^T \boldsymbol{\alpha_n}.$$

Στην περίπτωση που $\delta_n > \epsilon_0$, το $\boldsymbol{x_n}$ προστίθεται στο λεξικό, το οποίο τώρα περιέχει $M_n = M_{n-1} + 1$ κέντρα. Εάν $\delta_n \leq \epsilon_0$, τότε το $\boldsymbol{x_n}$ δεν περιλαμβάνεται στο λεξικό, δηλαδή $D_n = D_{n-1},\ M_n = M_{n-1}$, και προσεγγίζουμε το $\Phi(\boldsymbol{x_n})$ ως $\Phi(\boldsymbol{x_n}) \approx \sum_{m=1}^{M_n} \alpha_{n,m}\Phi(\boldsymbol{u_m})$.
Χρησιμοποιώντας την τελευταία σχέση μπορούμε να προσεγγίσουμε τον πλήρη kernel πίνακα K_n, όπου $(K_n)_{i,j} = K(\boldsymbol{x_i}, \boldsymbol{x_j})$ για $i, j = 1, 2, \ldots, n$, ως

$$K_n \approx A_n \tilde{K_n} A_n^T,$$

όπου $A_n = (\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \ldots \boldsymbol{\alpha_n})^T$ είναι ο $M_n \times M_n$ πίνακας που περιέχει τους συντελεστές του γραμμικού συνδυασμού που σχετίζεται με κάθε $\boldsymbol{u_m},\ m = 1, 2, \ldots, M_n$. Αυτό οφείλεται στο γεγονός ότι κάθε χρονική στιγμή n κάθε στοιχείο $K(\boldsymbol{x_i}, \boldsymbol{x_j})$ υπολογίζεται ως

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \langle \Phi(\boldsymbol{x_i}), \Phi(\boldsymbol{x_j}) \rangle_{\mathcal{H}} = \left\langle \sum_{k=1}^{M_n} \alpha_{i,k}\Phi(\boldsymbol{u_k}), \sum_{l=1}^{M_n} \alpha_{j,l}\Phi(\boldsymbol{u_l}) \right\rangle_{\mathcal{H}} =$$
$$\sum_{k=1}^{M_n}\sum_{l=1}^{M_n} \alpha_{j,l}\alpha_{i,k} \langle \Phi(\boldsymbol{u_k}), \Phi(\boldsymbol{u_l}) \rangle_{\mathcal{H}} = \sum_{k=1}^{M_n}\sum_{l=1}^{M_n} \alpha_{j,l}\alpha_{i,k} K(\boldsymbol{u_k}, \boldsymbol{u_l}),$$

για κάθε $k, l = 1, 2, \ldots, M_n,\ i, j = 1, 2, \ldots, N$. Παρατηρήστε ότι $\alpha_{i,j} = 0$ για κάθε $j > i$. Αυτό ακριβώς είναι το σημείο-κλειδί της λογικής του ALD για τον KRLS.
Σε όρους πινάκων η αντίστοιχη συνάρτηση κόστους του KRLS προβλήματος ελαχιστοποίησης στο βήμα n δίνεται από τη σχέση

$$\mathcal{L}_n(\boldsymbol{\theta}) = \left\|\boldsymbol{y_{(n)}} - K_n\boldsymbol{\theta}\right\|^2 \approx \left\|\boldsymbol{y_{(n)}} - A_n\tilde{K_n}A_n^T\boldsymbol{\theta}\right\|^2 = \left\|\boldsymbol{y_{(n)}} - A_n\tilde{K_n}\tilde{\boldsymbol{\theta}}\right\|^2 = \mathcal{L}_n(\tilde{\boldsymbol{\theta}}),$$

όπου $\tilde{\boldsymbol{\theta}} = A_n^T\boldsymbol{\theta}$ είναι η νέα παράμετρος προς βελτιστοποίηση. Παρατηρήστε ότι ενώ το αρχικό διάνυσμα $\boldsymbol{\theta_n} \in \mathbb{R}^n$ είναι ένα διάνυσμα διάστασης n, το νέο διάνυσμα $\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} \in \mathbb{R}^{M_n}$ έχει σημαντικά μικρότερη διάσταση (η οποία εξαρτάται από τη στρατηγική αραίωσης). Η λύση του τροποποιημένου προβλήματος ελαχιστοποίησης είναι

$$\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} = (\tilde{K_n}^T A_n^T A_n \tilde{K_n})^{-1}\tilde{K_n}^T A_n^T \boldsymbol{y_n} = \tilde{K_n}^{-1}(A_n^T A_n)^{-1}(\tilde{K_n}^T)^{-1}\tilde{K_n}^T A_n^T \boldsymbol{y_n} = \tilde{K_n}^{-1} P_n^{-1} A_n^T \boldsymbol{y_n},$$

όπου $P_n = A_n^T A_n$, υποθέτοντας ότι υπάρχει ο αντίστροφος. Για τις δύο περιπτώσεις της στρατηγικής αραίωσης ALD προκύπτουν, μετά από πράξεις [36], αναδρομικές σχέσεις για το $\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}$ ως εξής:

- Εάν $\delta_n \leq \epsilon_0$, τότε $\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} = \tilde{\boldsymbol{\theta}}_{\boldsymbol{n-1}} + \dfrac{\tilde{K_n}^{-1} P_{n-1}^{-1}\boldsymbol{\alpha_n} e_n}{1 + \boldsymbol{\alpha_n}^T P_{n-1}^{-1}\boldsymbol{\alpha_n}}$.
- Εάν $\delta_n > \epsilon_0$, τότε $\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} = \begin{bmatrix} \tilde{\boldsymbol{\theta}}_{\boldsymbol{n-1}} - \frac{\boldsymbol{\alpha_n}}{\delta_n} e_n \\ \frac{e_n}{\delta_n} \end{bmatrix}$.

Παρατηρήστε ότι αναπτύξαμε τη στρατηγική αυτή υποθέτοντας την αντιστρεψιμότητα των $\tilde{K_n}$ και $A_n^T A_n$ για κάθε n. Για την περίπτωση του RKHS που παράγεται από τον Gaussian πυρήνα, που είναι ο πιο δημοφιλής, αποδεικνύεται ότι ο $\tilde{K_n}$ είναι πράγματι αντιστρέψιμος. Επιπλέον, ο A_n είναι από κατασκευής πίνακας πλήρους τάξης. Επομένως, ο $A_n^T A_n$ είναι αυστηρά θετικά ορισμένος και άρα αντιστρέψιμος.

Ο Κώδικας **Kernel Recursive Least Squares (KRLS) Με Αραίωση ALD**

Είσοδος: Τα δεδομένα $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^N$, οι παράμετροι της kernel συνάρτησης και η ALD παράμετρος ϵ_0.

1: **Αρχικοποίηση**: $\tilde{K}^{-1} = \left[\frac{1}{K(\boldsymbol{x_1}, \boldsymbol{x_1})}\right], P^{-1} = [1], \tilde{\boldsymbol{\theta}} = \left[\frac{y_1}{K(\boldsymbol{x_1}, \boldsymbol{x_1})}\right], D = [\boldsymbol{x_1}], M = 1.$
2: **for** i=2 to N
3: $\quad \boldsymbol{\beta} = (K(\boldsymbol{u_1}, \boldsymbol{x_i}), \ldots, K(\boldsymbol{u_M}, \boldsymbol{x_i}))^T$
4: $\quad$ Υπολόγισε την έξοδο του συστήματος: $\hat{y}_i = \tilde{\boldsymbol{\theta}}^T \boldsymbol{\beta}$
5: $\quad$ Υπολόγισε το σφάλμα: $e_i = y_i - \hat{y}_i$
Για την αραίωση ALD
6: $\quad \boldsymbol{\alpha} = \tilde{K}^{-1}\boldsymbol{\beta}$
7: $\quad \delta = K(\boldsymbol{x_i}, \boldsymbol{x_i}) - \boldsymbol{\beta}^T \boldsymbol{\alpha}$
8: $\quad$ **if** $\delta > \epsilon_0$
9: $\quad\quad$ Πρόσθεσε το νέο κέντρο στο λεξικό: $D = D \cup \{\boldsymbol{x_i}\}, M = M + 1$
10: $\quad\quad$ Ενημέρωσε τον πίνακα $\tilde{K}^{-1} = \frac{1}{\delta}\begin{bmatrix} \delta\tilde{K}^{-1} + \boldsymbol{\alpha}\boldsymbol{\alpha}^T & -\boldsymbol{\alpha} \\ -\boldsymbol{\alpha}^T & 1 \end{bmatrix}$
11: $\quad\quad$ Ενημέρωσε τον πίνακα $P^{-1} = \begin{bmatrix} P^{-1} & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix}$
12: $\quad\quad$ Ενημέρωσε τη λύση $\tilde{\boldsymbol{\theta}} = \begin{bmatrix} \tilde{\boldsymbol{\theta}} - \frac{\boldsymbol{\alpha}}{\delta} e_i \\ \frac{e_i}{\delta} \end{bmatrix}$
13: $\quad$ **else**
14: $\quad\quad \boldsymbol{q} = \dfrac{P^{-1}\boldsymbol{\alpha}}{1 + \boldsymbol{\alpha}^T P^{-1}\boldsymbol{\alpha}}$
15: $\quad\quad$ Ενημέρωσε τον πίνακα $P^{-1} = P^{-1} - \boldsymbol{q}\boldsymbol{\alpha}^T P^{-1}$
16: $\quad\quad$ Ενημέρωσε τη λύση $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}} + \tilde{K}^{-1}\boldsymbol{q}e_i$
17: $\quad$ **end if**
18: **end for**
Έξοδος: Το διάνυσμα $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^M$ και το λεξικό $D = \{\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_M}\}$.

Κεφάλαιο 4

Πειράματα και Εφαρμογές

Προκειμένου να συγκρίνουμε τις επιδόσεις των αλγορίθμων που παρουσιάστηκαν στα προηγούμενα Κεφάλαια, πραγματοποιήσαμε ορισμένα πειράματα χρησιμοποιώντας τη γλώσσα προγραμματισμού Matlab. Άλλα πειράματα ήταν γραμμικά και άλλα μη γραμμικά, ώστε να καταστεί σαφής η χρησιμότητα των αλγορίθμων που χρησιμοποιούν kernels.
Σε όλους τους αλγορίθμους που αναφέρονται στο Κεφάλαιο και αφορούν kernels χρησιμοποιήθηκε η Gaussian kernel συνάρτηση. Επίσης, στις περιπτώσεις των αλγορίθμων τύπου LMS, χρησιμοποιήθηκε η normalized εκδοχή τους, ενώ, όπου ήταν εφικτό, χρησιμοποιήθηκε αλγόριθμος με αραίωση αντί του αντίστοιχου χωρίς αραίωση.
Στα πειράματα χρησιμοποιείται λευκός Gaussian θόρυβος (white Gaussian noise), όπως συνηθίζεται στις περισσότερες εφαρμογές. Πρόκειται για ένα βασικό μοντέλο θορύβου που χρησιμοποιείται στη Θεωρία Πληροφοριών για να μιμηθεί την επίδραση πολλών τυχαίων διαδικασιών που εμφανίζονται στη φύση. Το χαρακτηριστικό λευκός (white) αναφέρεται στην έννοια ότι ο θόρυβος έχει ομοιόμορφη δύναμη σε όλη τη ζώνη συχνοτήτων για το σύστημα πληροφοριών. Πρόκειται για το ανάλογο του λευκού χρώματος το οποίο έχει ομοιόμορφες εκπομπές σε όλες τις συχνότητες του ορατού φάσματος. Ο χαρακτηρισμός Gaussian οφείλεται στο ότι ο θόρυβος ακολουθεί κανονική κατανομή στο χρόνο με μέση τιμή μηδέν.
Οι τιμές των παραμέτρων σε κάθε αλγόριθμο επιλέχθηκαν ώστε να ελαχιστοποιείται το μέσο τετραγωνικό σφάλμα (MSE). Οι παράμετροι βελτιστοποιήθηκαν με cross validation τεχνική.

4.1 Ταυτοποίηση Καναλιού (Channel Identification)

Στο πείραμα **ταυτοποίησης καναλιού**, δεδομένων των παρατηρήσεων x γίνεται εκτίμηση του καναλιού. Αρχικά γίνεται εκτίμηση του καναλιού μέσω εκπαίδευσης σε δείγμα μικρού σχετικά μεγέθους. Αφού ταυτοποιηθεί το κανάλι, προβλέπονται οι τιμές $\hat{y}$, των y.
Στόχος του πειράματος ταυτοποίησης καναλιού είναι ο σχεδιασμός ενός φίλτρου που να δρα επάνω στην είσοδο του καναλιού, x, και να αναπαράγει το γνήσιο σήμα εξόδου, y, όσο καλύτερα γίνεται (Σχήμα 4.1).

Σχήμα 4.1: Channel Identification

4.1.1 Ταυτοποίηση Γραμμικού Καναλιού (Linear Channel Identification)

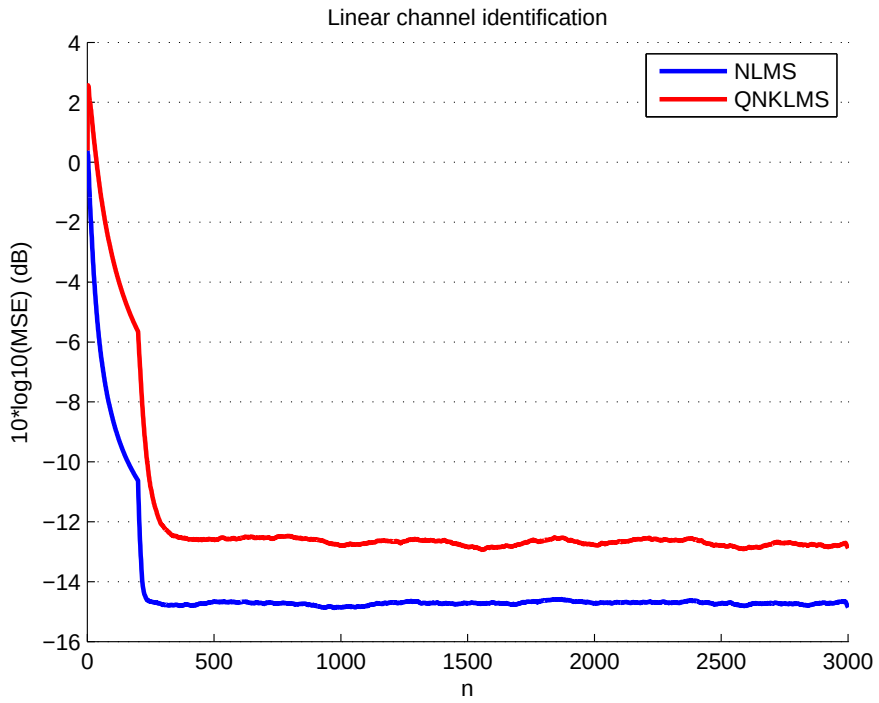
Θεωρούμε ένα τυπικό πείραμα ταυτοποίησης γραμμικού καναλιού.
Το σήμα περνά από το γραμμικό κανάλι

$$t_n = 0.2638 \cdot x_{(n)} - 0.4092 \cdot x_{(n-1)} + 0.2441 \cdot x_{(n-2)} - 0.9049 \cdot x_{(n-3)} + 0.9892 \cdot x_{(n-4)},$$

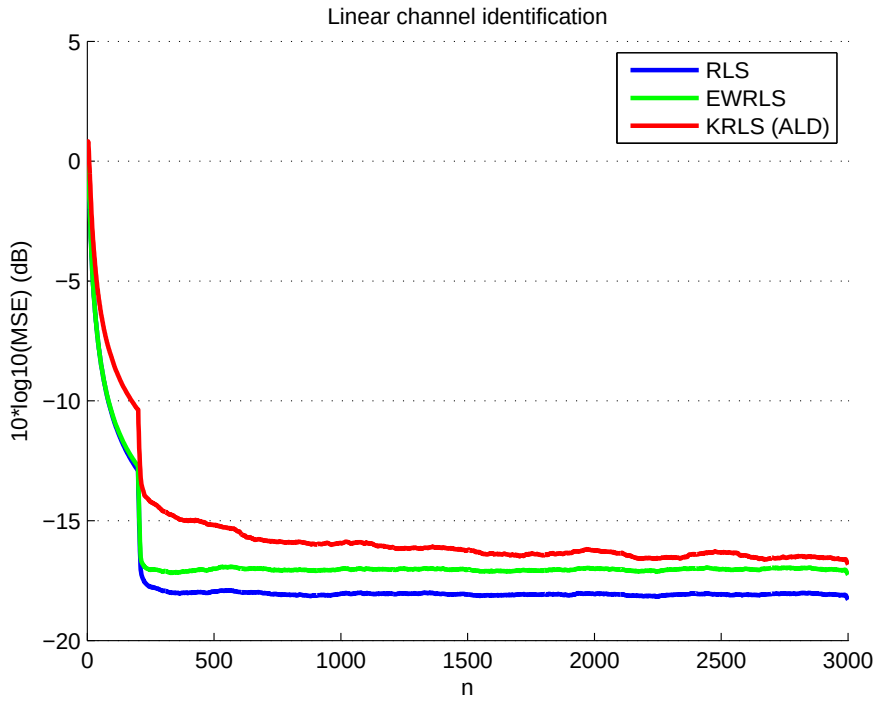
επηρεάζεται από λευκό Gaussian θόρυβο επιπέδου 15 dB και παρατηρείται τελικά ως y_n.
Το πείραμα πραγματοποιήθηκε σε 100 σύνολα από 3000 δείγματα δεδομένων το καθένα και οι συγκρίσεις έγιναν μεταξύ των αλγορίθμων τύπου LMS στη μια περίπτωση και μεταξύ των αλγορίθμων τύπου RLS στην άλλη.

Συγκρίνουμε στο γραμμικό πρόβλημα τον απλό αλγόριθμο LMS με τον Kernel LMS. Επιθυμούμε να ελαχιστοποιηθεί η συνάρτηση κόστους που έχουμε επιλέξει, δηλαδή το μέσο τετραγωνικό σφάλμα (MSE). Οι παράμετροι βελτιστοποιήθηκαν σε: $\mu = 1$ το βήμα εκμάθησης, $\sigma = 3.5$ η παράμετρος της Gaussian kernel συνάρτησης και $\delta = 1.9$ το μέγεθος κβαντισμού. Από τη γραφική παράσταση του μέσου τετραγωνικού σφάλματος (MSE) για το πείραμα (Σχήμα 4.2α΄) παρατηρεί κανείς ότι ο Kernel LMS αλγόριθμος δεν προσφέρει κάτι περισσότερο από τον LMS.

Στη συνέχεια, συγκρίνουμε στο γραμμικό πρόβλημα τους αλγορίθμους RLS, EWRLS με τον Kernel RLS, όπου επιθυμούμε επίσης να ελαχιστοποιηθεί το μέσο τετραγωνικό σφάλμα (MSE). Οι παράμετροι βελτιστοποιήθηκαν σε: $\lambda = 0.15$ η παράμετρος εξομάλυνσης, $w = 0.95$ ο συντελεστής βαρύτητας, $\sigma = 3.5$ η παράμετρος της Gaussian kernel συνάρτησης και $\epsilon_0 = 0.04$ η ALD παράμετρος. Στο Σχήμα 4.2β΄ βλέπει κανείς τη γραφική παράσταση του μέσου τετραγωνικού σφάλματος (MSE) για το πείραμα αυτό. Όπως προηγουμένως, δε φαίνεται να έχει όφελος κανείς χρησιμοποιώντας τον αλγόριθμο KRLS για το γραμμικό πρόβλημα.



(α′) Μέθοδοι LMS. Βήμα εκμάθησης μ=1, kernel παράμετρος σ=3.5, μέγεθος κβαντισμού δ=1.9



(β′) Μέθοδοι RLS, παράμετρος εξομάλυνσης λ=0.15, συντελεστής βαρύτητας w=0.95, kernel παράμετρος σ=3.5, ALD παράμετρος $\varepsilon_0 = 0.04$

Σχήμα 4.2: Μέσο τετραγωνικό σφάλμα (MSE) για το πείραμα ταυτοποίησης γραμμικού καναλιού.

Όπως ήταν αναμενόμενο, όλοι οι αλγόριθμοι ανταποκρίνονται καλά για το γραμμικό πρόβλημα. Οι kernel αλγόριθμοι ωστόσο, είναι πιο πολύπλοκοι υπολογιστικά χωρίς να παρουσιάζουν καλύτερα αποτελέσματα. Αυτό που έχει ενδιαφέρον και αποτελεί την αιτία εφαρμογής της θεωρίας των kernels στη Μηχανική Μάθηση, είναι η συμπεριφορά των αλγορίθμων σε μη γραμμικά προβλήματα.

4.1.2 Ταυτοποίηση Μη Γραμμικού Καναλιού (Nonlinear Channel Identification)

Θεωρούμε αυτή τη φορά ένα πείραμα ταυτοποίησης μη γραμμικού καναλιού.
Το σήμα περνά από το γραμμικό κανάλι

$$t_n = 0.2638 \cdot x_{(n)} - 0.4092 \cdot x_{(n-1)} + 0.2441 \cdot x_{(n-2)} - 0.9049 \cdot x_{(n-3)} + 0.9892 \cdot x_{(n-4)}$$

και στη συνέχεια από το μη γραμμικό κανάλι

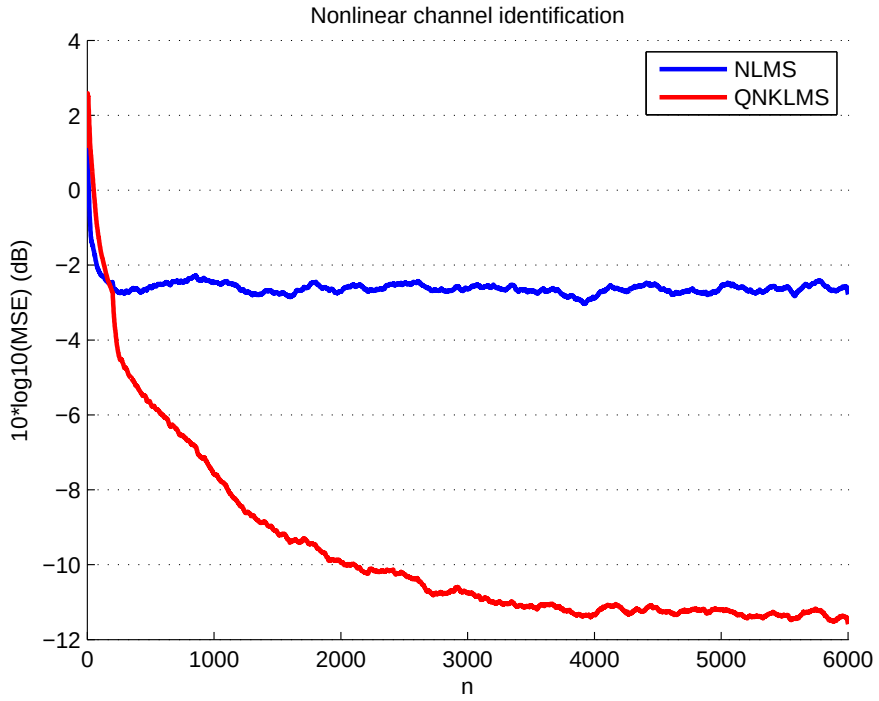
$$q_n = 0.6 \cdot x_{(n)}^2,$$

επηρεάζεται από λευκό Gaussian θόρυβο επιπέδου 15 dB και παρατηρείται τελικά ως y_n.
Το πείραμα πραγματοποιήθηκε σε 100 σύνολα από 6000 δείγματα δεδομένων το καθένα και οι συγκρίσεις έγιναν μεταξύ των αλγορίθμων τύπου LMS στη μια περίπτωση και μεταξύ των αλγορίθμων τύπου RLS στην άλλη. Επιπλέον, συγκρίναμε τις επιδόσεις των αλγορίθμων KLMS και KRLS με αραίωση. Σε όλες τις συγκρίσεις που μόλις αναφέραμε σκοπός μας ήταν να ελαχιστοποιηθεί η επιλεγμένη συνάρτηση κόστους, δηλαδή το μέσο τετραγωνικό σφάλμα (MSE).

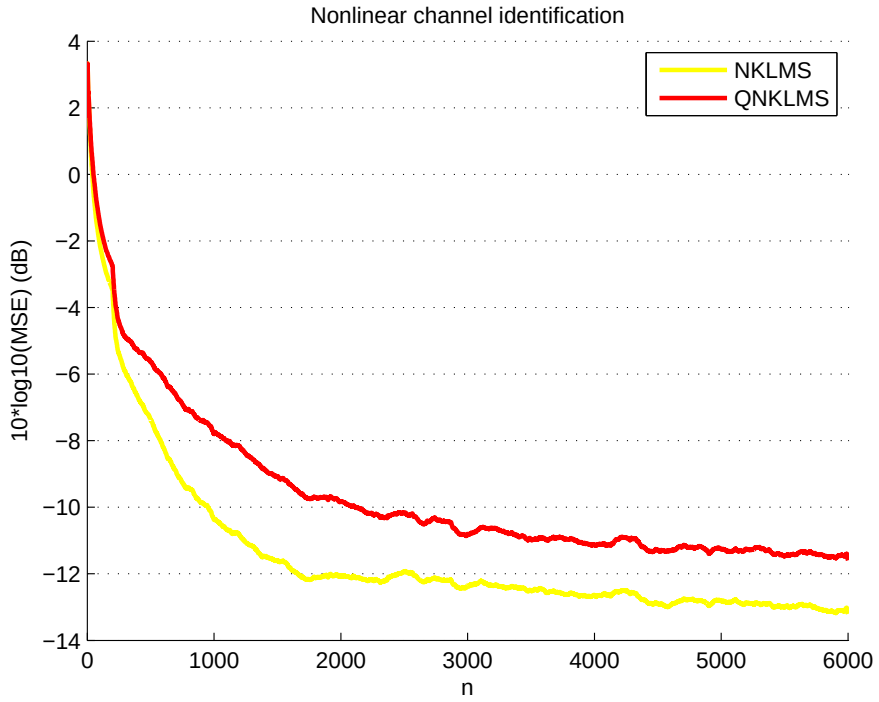
Συγκρίνουμε στο μη γραμμικό πρόβλημα τον απλό αλγόριθμο LMS με τον Kernel LMS. Οι παράμετροι βελτιστοποιήθηκαν σε: $\mu = 1$ το βήμα εκμάθησης, $\sigma = 3$ η παράμετρος της Gaussian kernel συνάρτησης και $\delta = 2$ το μέγεθος κβαντισμού. Είναι φανερό από τη γραφική παράσταση του μέσου τετραγωνικού σφάλματος (MSE) στο Σχήμα 4.3α΄ ότι ο LMS αδυνατεί να αντεπεξέλθει, ενώ, αντίθετα, ο KLMS δίνει πολύ καλό αποτέλεσμα.

Στη συνέχεια, συγκρίναμε τις επιδόσεις του αλγορίθμου KLMS και του KLMS με αραίωση. Στο Σχήμα 4.3β΄ δίνεται η γραφική παράσταση του μέσου τετραγωνικού σφάλματος (MSE). Βλέπουμε εδώ ότι ο αραιός αλγόριθμος δεν ανταποκρίνεται τόσο καλά όσο ο απλός KLMS, συχνά όμως σε εφαρμογές επιλέγεται ο αραιός προς εξοικονόμηση χρόνου. Πρέπει να σχολιάσουμε ότι η ταχύτητα του KLMS με αραίωση είναι μεγαλύτερη περίπου κατά 88% έναντι του KLMS χωρίς αραίωση.

Στη συνέχεια, συγκρίνουμε στο μη γραμμικό πρόβλημα τους αλγορίθμους RLS, EWRLS με τον Kernel RLS. Οι παράμετροι βελτιστοποιήθηκαν σε: $\lambda = 0.05$ η παράμετρος εξομάλυνσης, $w = 0.95$ ο συντελεστής βαρύτητας, $\sigma = 4$ η παράμετρος της Gaussian kernel συνάρτησης και $\epsilon_0 = 0.01$ η ALD παράμετρος. Όπως παρατηρεί κανείς στο Σχήμα 4.4α΄ οι αλγόριθμοι RLS και EWRLS δεν ανταποκρίνονται καλά, σε αντίθεση βεβαίως με τον KRLS, κάτι που ήταν αναμενόμενο, αφού οι δύο πρώτοι αδυνατούν να διαχειριστούν τη μη γραμμικότητα.

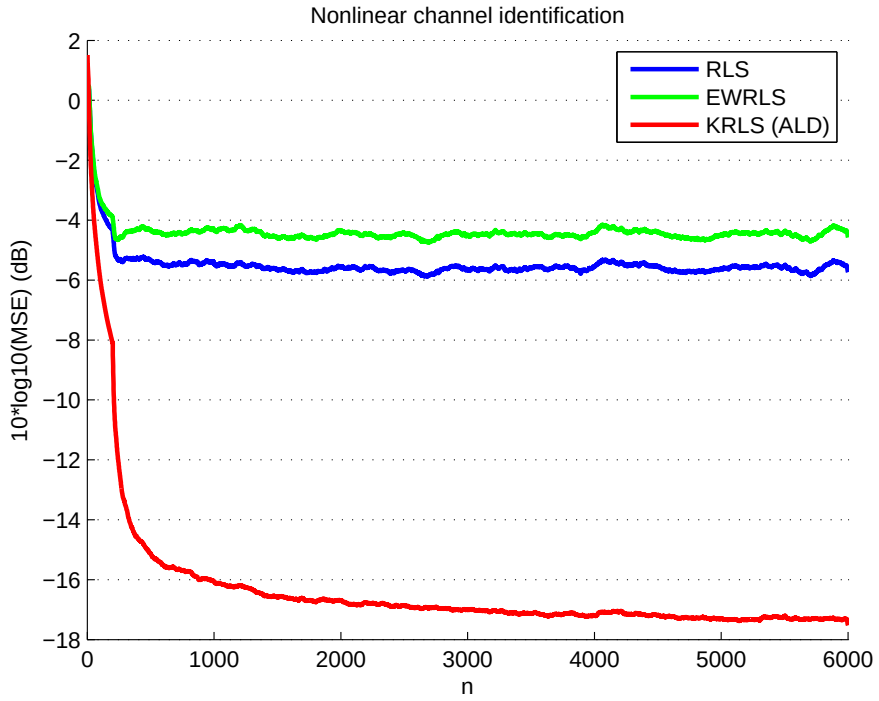


(α′) Σύγκριση των αλγορίθμων NLMS και NKLMS με αραίωση (QNKLMS).

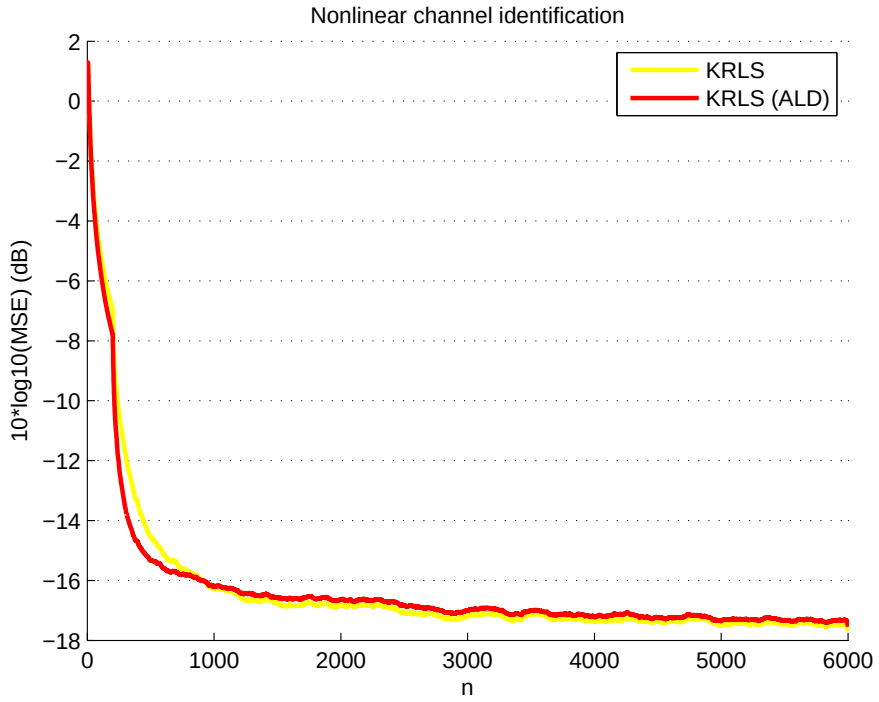


(β′) Σύγκριση των αλγορίθμων NKLMS και NKLMS με αραίωση (QNKLMS).

Σχήμα 4.3: Μέσο τετραγωνικό σφάλμα (MSE) για το πείραμα ταυτοποίησης μη γραμμικού καναλιού.
Σύγκριση των αλγορίθμων τύπου LMS. Βήμα εκμάθησης μ=1, kernel παράμετρος σ=3, μέγεθος κβαντισμού δ=2.

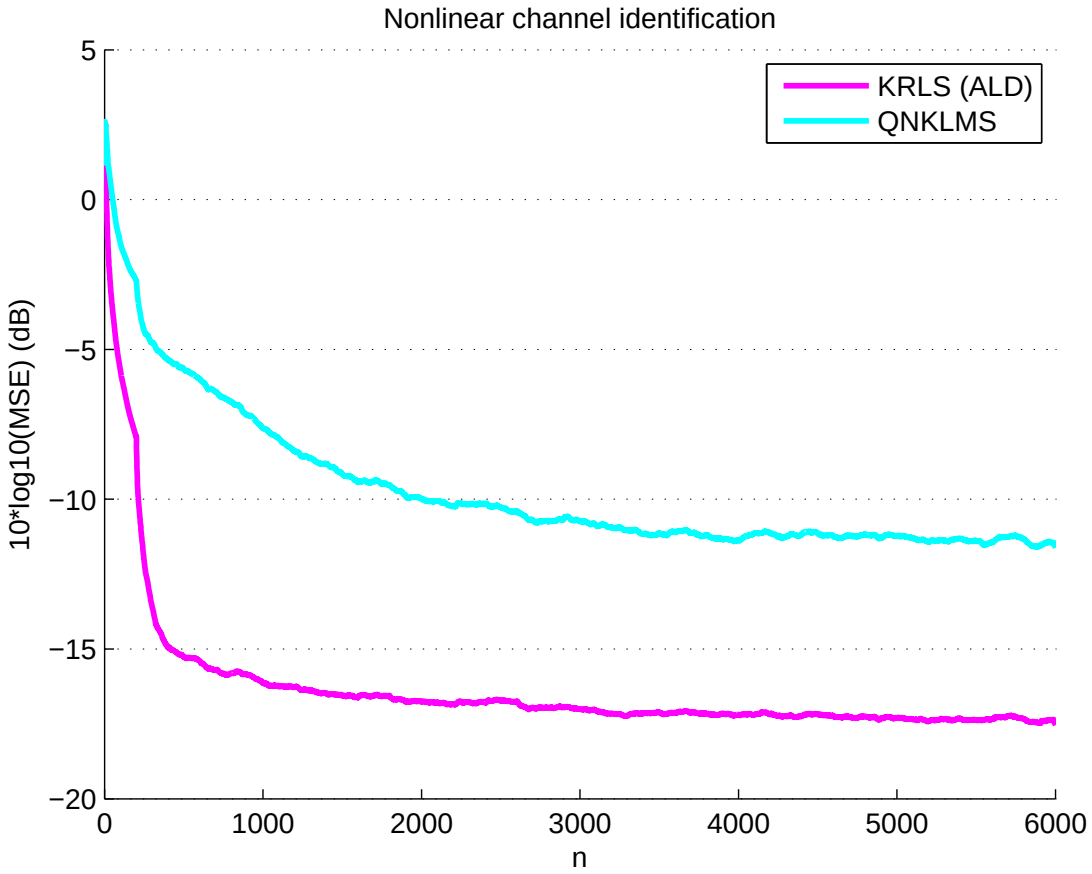


(α′) Σύγκριση των αλγορίθμων RLS, EWRLS και KRLS με αραίωση (ALD).



(β′) Σύγκριση των αλγορίθμων KRLS και KRLS με αραίωση (ALD).

Σχήμα 4.4: Μέσο τετραγωνικό σφάλμα (MSE) για το πείραμα ταυτοποίησης μη γραμμικού καναλιού.
Σύγκριση των αλγορίθμων τύπου RLS. Παράμετρος εξομάλυνσης λ=0.05, συντελεστής βαρύτητας w=0.95, kernel παράμετρος σ=4, ALD παράμετρος $\varepsilon_0 = 0.01$.



Σχήμα 4.5: Μέσο τετραγωνικό σφάλμα (MSE) για τους αλγορίθμους QNKLMS και KRLS με αραίωση (ALD).
Για τον QNKLMS: βήμα εκμάθησης μ=1, kernel παράμετρος σ=3, μέγεθος κβαντισμού δ=2.
Για τον KRLS: kernel παράμετρος σ=4, ALD παράμετρος $\varepsilon_0 = 0.01$.

Επίσης, εξετάσαμε την επίπτωση της αραίωσης στην επίδοση του αλγορίθμου KRLS. Το Σχήμα 4.4β΄ δείχνει τη γραφική παράσταση του μέσου τετραγωνικού σφάλματος (MSE). Παρότι εδώ δε φαίνεται να υπάρχει σημαντική διαφορά στην επίδοση των δύο αλγορίθμων, ωστόσο ο αραιός έχει ταχύτερη σύγκλιση. Επίσης, πρέπει να επισημάνουμε ότι η ταχύτητα αυξάνεται περίπου κατά 99% στον KRLS με αραίωση, σε σχέση με τον KRLS χωρίς αραίωση.

Μια τελευταία σύγκριση που κάναμε είναι αυτή μεταξύ των αλγορίθμων KLMS και KRLS, όπου χρησιμοποιήθηκαν αντίστοιχα οι αλγόριθμοι QNKLMS και KRLS με αραίωση ALD. Οι παράμετροι ήταν αντίστοιχα $\mu = 1$ το βήμα εκμάθησης, $\sigma = 3$ η παράμετρος της Gaussian kernel συνάρτησης, $\delta = 2$ το μέγεθος κβαντισμού και $\lambda = 0.05$ η παράμετρος εξομάλυνσης, $w = 0.95$ ο συντελεστής βαρύτητας, $\sigma = 4$ η παράμετρος για τη Gaussian kernel συνάρτηση, $\epsilon_0 = 0.01$ η ALD παράμετρος. Από τη γραφική παράσταση του μέσου τετραγωνικού σφάλματος (MSE) (Σχήμα 4.5) βλέπουμε πως ο αλγόριθμος KRLS συγκλίνει πολύ πιο γρήγορα από τον KLMS. Παρατηρήσαμε, επίσης, ότι ο KRLS χρειάζεται περίπου 52% περισσότερο χρόνο από τον KLMS για να δώσει αποτέλεσμα.

4.2 Αντιστάθμιση Καναλιού (Channel Equalization)

Στο πείραμα **αντιστάθμισης καναλιού**, η διαδικασία είναι η αντίστροφη από αυτή στο πείραμα ταυτοποίησης. Συγκεκριμένα, μέσω της εκπαίδευσης σε δείγμα μικρού σχετικά μεγέθους γίνεται εκτίμηση του καναλιού με δεδομένες τις εξόδους y, οι οποίες έχουν προκύψει από τις παρατηρήσεις x. Αφού ολοκληρωθεί η εκπαίδευση, αναπαράγεται το γνήσιο σήμα εισόδου, x, όσο καλύτερα γίνεται.
Στόχος του πειράματος αντιστάθμισης καναλιού είναι ο σχεδιασμός ενός «αντίστροφου» φίλτρου το οποίο δρώντας στην έξοδο, y, του καναλιού να αναπαράγει το γνήσιο σήμα εισόδου x με όσο μικρότερο σφάλμα γίνεται (Σχήμα 4.6). Επίσης, θεωρούμε καθυστέρηση (delay) D, η οποία λαμβάνεται για τις διάφορες καθυστερήσεις που σχετίζονται με το σύστημά μας.

4.2.1 Αντιστάθμιση Γραμμικού Καναλιού (Linear Channel Equalization)

Θεωρούμε ένα τυπικό πείραμα αντιστάθμισης γραμμικού καναλιού.
Το σήμα περνά από το γραμμικό κανάλι

$$t_n = -0.21 \cdot x_{(n)} + 0.05 \cdot x_{(n-1)} + 0.45 \cdot x_{(n-2)},$$

επηρεάζεται από λευκό Gaussian θόρυβο επιπέδου 15 dB και παρατηρείται τελικά ως y_n. Στη διάρκεια της εκπαίδευσης το y αποτελεί την είσοδο, ενώ το x την έξοδο και σκοπός είναι η εκτίμηση του καναλιού.
Το πείραμα πραγματοποιήθηκε σε 100 σύνολα από 3000 δείγματα δεδομένων το καθένα και οι συγκρίσεις έγιναν μεταξύ των αλγορίθμων τύπου LMS στη μια περίπτωση και μεταξύ των αλγορίθμων τύπου RLS στην άλλη. Η καθυστέρηση (delay) τέθηκε ίση με 5.

Συγκρίνουμε στο γραμμικό πρόβλημα τον απλό αλγόριθμο LMS με τον Kernel LMS ως

Σχήμα 4.6: Channel Equalization

προς το μέσο τετραγωνικό σφάλμα (MSE). Οι παράμετροι βελτιστοποιήθηκαν σε: $\mu = 0.8$ το βήμα εκμάθησης για τον LMS και $\mu = 1$ το βήμα εκμάθησης για τον KLMS, $\sigma = 1$ η παράμετρος της Gaussian kernel συνάρτησης και $\delta = 0.5$ το μέγεθος κβαντισμού. Από τη γραφική παράσταση του μέσου τετραγωνικού σφάλματος (MSE) για το πείραμα (Σχήμα 4.7α΄) παρατηρεί κανείς ότι ο Kernel LMS αλγόριθμος δεν προσφέρει κάτι περισσότερο από τον LMS.

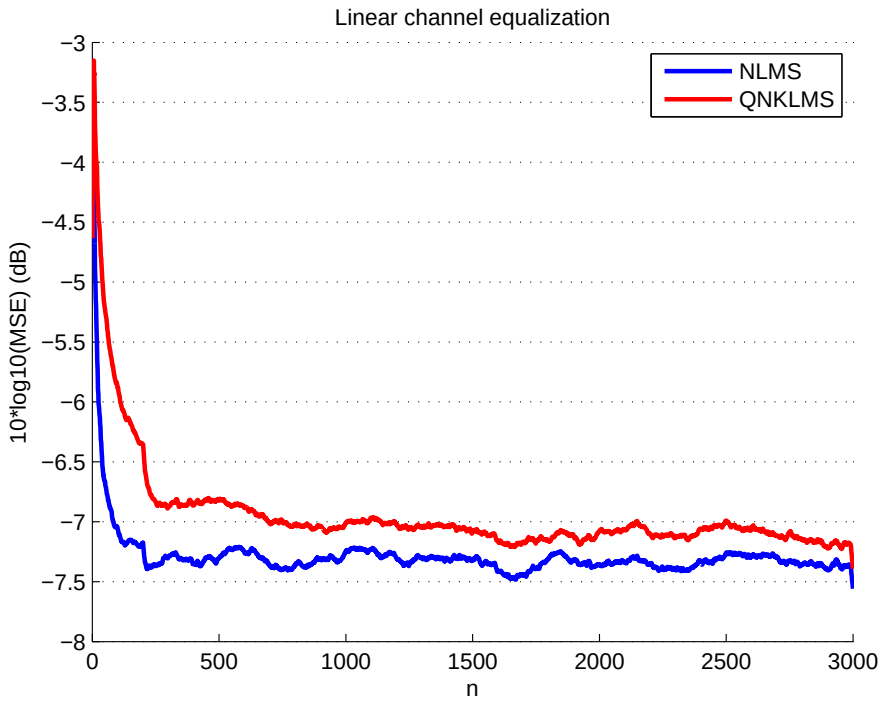
Στη συνέχεια, συγκρίνουμε στο γραμμικό πρόβλημα τους αλγορίθμους RLS, EWRLS με τον Kernel RLS, όπου επιθυμούμε επίσης να ελαχιστοποιηθεί το μέσο τετραγωνικό σφάλμα (MSE). Οι παράμετροι βελτιστοποιήθηκαν σε: $\lambda = 0.1$ η παράμετρος εξομάλυνσης, $w = 0.95$ ο συντελεστής βαρύτητας και $\sigma = 2$ η παράμετρος της Gaussian kernel συνάρτησης. Στο Σχήμα 4.7β΄ βλέπει κανείς τη γραφική παράσταση του μέσου τετραγωνικού σφάλματος (MSE) για το πείραμα αυτό. Όπως προηγουμένως, δε φαίνεται να έχει όφελος κανείς χρησιμοποιώντας τον αλγόριθμο KRLS για το γραμμικό πρόβλημα.

Όπως ήταν αναμενόμενο, όλοι οι αλγόριθμοι ανταποκρίνονται καλά στο πείραμα αντιστάθμισης καναλιού για το γραμμικό πρόβλημα. Στην επόμενη υποενότητα ωστόσο, είναι εμφανές ότι μόνο οι αλγόριθμοι που βασίζονται σε kernels μπορούν να αντεπεξέλθουν ικανοποιητικά στα μη γραμμικά προβλήματα.

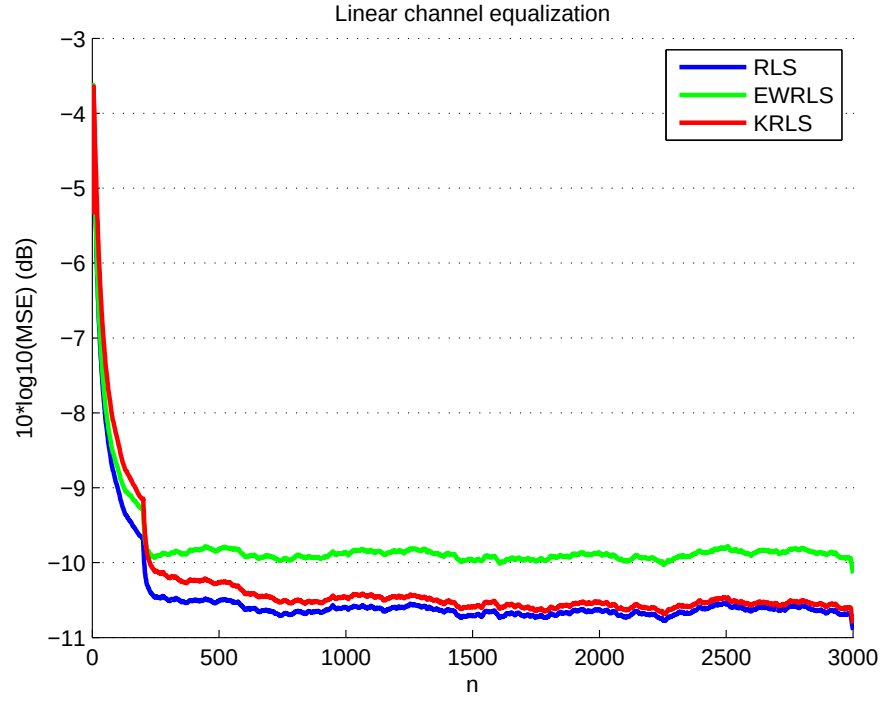
4.2.2 Αντιστάθμιση Μη Γραμμικού Καναλιού (Nonlinear Channel Equalization)

Θεωρούμε αυτή τη φορά ένα πείραμα αντιστάθμισης μη γραμμικού καναλιού.
Το σήμα περνά από το γραμμικό κανάλι

$$t_n = -0.21 \cdot x_{(n)} + 0.05 \cdot x_{(n-1)} + 0.45 \cdot x_{(n-2)}$$



(α′) Βήμα εκμάθησης μ=0.8 για τον LMS, μ=1 για τον KLMS, kernel παράμετρος σ=1, μέγεθος κβαντισμού δ=0.5



(β′) Παράμετρος εξομάλυνσης λ=0.1, συντελεστής βαρύτητας w=0.95, kernel παράμετρος σ=2

Σχήμα 4.7: Μέσο τετραγωνικό σφάλμα (MSE) για το πείραμα αντιστάθμισης γραμμικού καναλιού.

και στη συνέχεια από το μη γραμμικό κανάλι

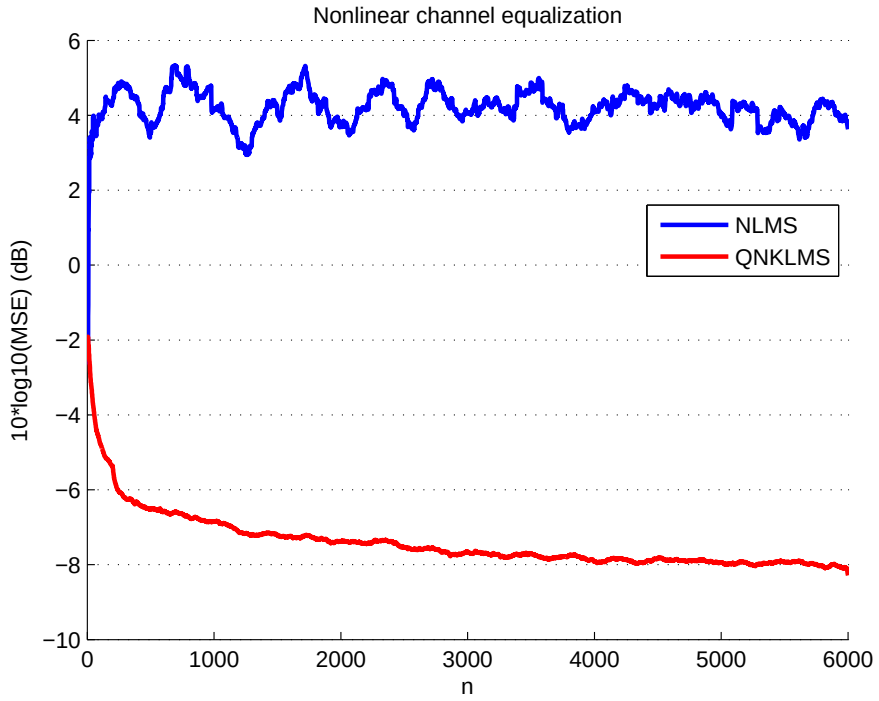
$$q_n = -3.7 \cdot x_{(n-1)}^3 + 0.6 \cdot x_{(n-2)}^2,$$

επηρεάζεται από λευκό Gaussian θόρυβο επιπέδου 15 dB και παρατηρείται τελικά ως y_n. Στη διάρκεια της εκπαίδευσης το y αποτελεί την είσοδο, ενώ το x την έξοδο και σκοπός είναι η εκτίμηση του καναλιού.
Το πείραμα πραγματοποιήθηκε σε 100 σύνολα από 6000 δείγματα δεδομένων το καθένα και οι συγκρίσεις έγιναν μεταξύ των αλγορίθμων τύπου LMS στη μια περίπτωση και μεταξύ των αλγορίθμων τύπου RLS στην άλλη. Η καθυστέρηση (delay) τέθηκε ίση με 5.

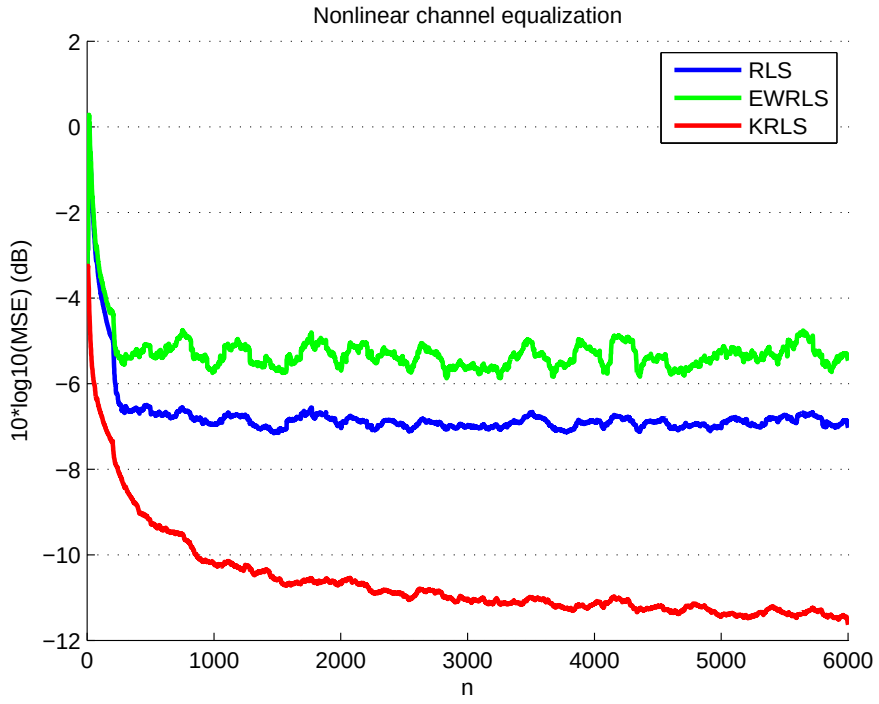
Συγκρίνουμε στο μη γραμμικό πρόβλημα τον απλό αλγόριθμο LMS με τον Kernel LMS. Επιθυμούμε να ελαχιστοποιηθεί η συνάρτηση κόστους που έχουμε επιλέξει, δηλαδή το μέσο τετραγωνικό σφάλμα (MSE). Οι παράμετροι βελτιστοποιήθηκαν σε: $\mu = 1$ το βήμα εκμάθησης, $\sigma = 10$ η παράμετρος της Gaussian kernel συνάρτησης και $\delta = 6$ το μέγεθος κβαντισμού. Είναι φανερό από τη γραφική παράσταση του μέσου τετραγωνικού σφάλματος (MSE) στο Σχήμα 4.8α΄ ότι ο LMS αδυνατεί να αντεπεξέλθει, ενώ, αντίθετα, ο KLMS δίνει πολύ καλό αποτέλεσμα.

Στη συνέχεια, συγκρίνουμε στο μη γραμμικό πρόβλημα τους αλγορίθμους RLS, EWRLS με τον Kernel RLS ως προς το μέσο τετραγωνικό σφάλμα (MSE). Οι παράμετροι βελτιστοποιήθηκαν σε: $\lambda = 1$ η παράμετρος εξομάλυνσης, $w = 0.95$ ο συντελεστής βαρύτητας για τους αλγορίθμους RLS, EWRLS και $\lambda = 0.1$ η παράμετρος εξομάλυνσης, $\sigma = 10$ η παράμετρος της Gaussian kernel συνάρτησης για τον αλγόριθμο KRLS. Όπως παρατηρεί κανείς στο Σχήμα 4.8β΄ οι αλγόριθμοι RLS και EWRLS δεν ανταποκρίνονται καλά, σε αντίθεση βεβαίως με τον KLMS, κάτι που ήταν αναμενόμενο, αφού οι δύο πρώτοι δεν έχουν σχεδιαστεί για να αντιμετωπίζουν μη γραμμικά προβλήματα.

Είναι φανερό, λοιπόν, από τα Σχήματα 4.8α΄ και 4.8β΄ ότι σε μη γραμμικά προβλήματα οι kernel αλγόριθμοι είναι οι μόνοι - μεταξύ αυτών που μελετήσαμε - που μπορούν να δώσουν καλό αποτέλεσμα.



(α′) Βήμα εκμάθησης μ=1, kernel παράμετρος σ=10, μέγεθος κβαντισμού δ=6



(β′) Παράμετρος εξομάλυνσης λ=1 για τους RLS και EWRLS, συντελεστής βαρύτητας w=0.95, παράμετρος εξομάλυνσης λ=0.1 για τον KRLS, kernel παράμετρος σ=10

Σχήμα 4.8: Μέσο τετραγωνικό σφάλμα (MSE) για το πείραμα αντιστάθμισης μη γραμμικού καναλιού.

Βιβλιογραφία

[1] Aizerman M. A. , Braverman E. M. , Rozonoer L. I. , *Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning*, Automation and Remote Control, Vol. 25, pp. 821-837, 1964.

[2] Aronszajn N. , *La Théorie Générale des Noyaux Reproduisants et ses Applications*, Première Partie, Proceedings of the Cambridge Philosophical Society, Vol. 39, p.133, 1944.

[3] Aronszajn N. , *Theory of Reproducing Kernels*, Transactions of the American Mathematical Society, Vol. 68, No. 3, pp. 337-404, 1950.

[4] Bergman S. , *Über die Entwicklung der harmonischen Funktionen der Ebene und des Raumes nach Orthogonalfunktionen*, Mathematische Annalen, Vol. 86, pp. 238-271, 1922. (Thesis, Berlin, 1921)

[5] Bergman S. , *Functions satisfying certain Partial Differential Equations of Elliptic Type and their Representation*, Duke Mathematical Journal, Vol. 14, pp. 349-366, 1947.

[6] Bergman S. , Schiffer M. , *A Representation of Green's and Neumann's functions in the Theory of Partial Differential Equations of Second Order*, Duke Mathematical Journal, Vol. 14, pp. 609-638, 1947.

[7] Bergman S. , Schiffer M. , *On Green's and Neumann's functions in the Theory of Partial Differential Equations*, Bulletin of the American Mathematical Society, Vol. 53, pp. 1141-1151, 1947.

[8] Bergman S. , Schiffer M. , *Kernel functions in the Theory of Partial Differential Equations of Elliptic Type*, Duke Mathematical Journal, Vol. 15, pp. 535-566, 1948.

[9] Bishop C. M. , *Pattern Recognition and Machine Learning*, Springer, corr. 2nd printing Ed. , 2007.

[10] Bochner S. , *Vorlesungen über Fouriersche Integrale*, Leipzig, 1932.

[11] Bochner S. , *Hilbert Distances and Positive Definite Functions*, Annals of Mathematics, Vol. 42, pp. 647-656, 1941.

[12] Boser B. , Guyon I. , Vapnik V. , *A Training Algorithm for Optimal Margin Classifiers*, In: "Fifth Annual Workshop on Computational Learning Theory", pp. 144-152, Pittsburgh, 1992.

[13] Bouboulis P. , Slavakis K. , Theodoridis S. , *Adaptive Kernel-based Image Denoising employing Semi-Parametric Regularization*, IEEE Transactions on Image Processing, Vol. 19, No. 6, pp. 1465-1479, June 2010.

[14] Bouboulis P. , Theodoridis S. , *Extension of Wirtinger's Calculus to Reproducing Kernel Hilbert Spaces and the Complex Kernel LMS*, IEEE Transactions on Signal Processing, Vol. 59, No. 3, pp. 964-978, 2011.

[15] Bouboulis P. , Slavakis K. , Theodoridis S. , *Adaptive Learning in Complex Reproducing Kernel Hilbert Spaces Employing Wirtinger's Subgradients*, IEEE Transactions on Neural Networks and Learning Systems, Vol. 23, No. 3, pp. 425-438, March 2012.

[16] Bouboulis P. , Theodoridis S. , Mavroforakis M. , *The Augmented Complex Kernel LMS*, IEEE Transactions on Signal Processing, 2012.

[17] Conway J. , *A Course in Functional Analysis*, Graduate Texts in Mathematics, Springer-Verlag, New York, 2nd Ed. ,1990.

[18] Diniz P. S. R. , *Adaptive Filtering: Algorithms and Practical Implementation*, Springer, 3rd Ed. , 2008.

[19] Engel Y. , Mannor S. , Meir R. , *The Kernel Recursive Least-Squares Algorithm*, IEEE Transactions on Signal Processing, Vol. 52, No. 8, pp. 2275-2285, 2004.

[20] Gelfand I. , Raikoff D. , *Irreducible Unitary Representation of Arbitrary Locally Bicompact Groups*, Recreational Mathematics (Matematicheskii Sbornik), Vol. 13, p. 316, 1943.

[21] Godement R. , *Sur les Fonctions de Type Positif. Sur les Propriétés Ergodiques des Fonctions de Type Positif. Sur les Partitions Finies des Fonctions de Type Positif. Sur Certains Opérateurs Définis dans l' Espace d' une Fonction de Type Positif. Sur Quelques Propriétés des Fonctions de Type Positif Définies sur un Groupe Quelconque.* , Comptes Rendus de l' Académie des Sciences Paris, Vol. 221 (1945), p. 69, p. 134, Vol. 222 (1946), p. 36, p. 213, p. 529.

[22] Godement R. , *Les Fonctions de Type Positif et la Théorie des Groupes*, Transactions of the American Mathematical Society, Vol. 63, pp. 1-84, Paris, 1948. (Thesis, Paris)

[23] Haykin S. , *Adaptive Filter Theory*, Prentice-Hall, New Jersey, 3rd Ed. , 1996.

[24] Liu W. , Príncipe J. , Haykin S. , *Kernel Adaptive Filtering: A Comprehensive Introduction*, Wiley, Hoboken: New Jersey, 2010.

[25] Mercer J. , *Functions of Positive and Negative Type and their connection with the Theory of Integral Equations*, Philosophical Transactions of the Royal Society of London, Ser. A, Vol. 209, pp. 415-446, 1909.

[26] Mercer J. , *Sturm-Liouville Series of Normal Functions in the Theory of Integral Equations*, Philosophical Transactions of the Royal Society of London, Ser. A, Vol. 211, pp. 111-198, 1911.

[27] Moore E. H. , *On Properly Positive Hermitian Matrices*, Bulletin of the American Mathematical Society, Vol. 23, p. 59, 1916.

[28] Moore E. H. , *General Analysis*, Memoirs of the American Philosophical Society, Part I, 1935, Part II, 1939.

[29] v. Neumann J. , Schönberg I. J. , *Fourier Integrals and Metric Geometry*, Transactions of the American Mathematical Society, Vol. 50, pp. 226-251, 1941.

[30] Paulsen V. I. , *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, September 2009. Notes.

[31] Sayed A. H. , *Fundamentals of Adaptive Filtering*, John Wiley & Sons, New Jersey, 2003.

[32] Shawe-Taylor J. , Cristianini N. , *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[33] Schönberg I. J. , *Metric Spaces and Positive Definite Functions*, Transactions of the American Mathematical Society, Vol. 44, pp. 522-536, 1938.

[34] Schönberg I. J. , *Positive Definite Functions on Spheres*, Duke Mathematical Journal, Vol. 9, pp. 96-108, 1942.

[35] Schölkopf B. , Smola A. , *Learning with Kernels*, The MIT Press, Cambridge, MA, 2001.

[36] Slavakis K. , Bouboulis P. , Theodoridis S. , *Online Learning in Reproducing Kernel Hilbert Spaces*, Academic Press Library in Signal Processing, Academic Press, 2014.

[37] Theodoridis S. , Koutroumbas K. , *Pattern Recognition*, Academic Press, 4th Ed. , 2009.

[38] Weil A. , *L' Intégration dans les Groupes Topologiques et ses Applications*, Actualités Scientifiques et Industrielles, Vol. 869, Paris, 1940.

[39] Zaremba S. , *L' Équation Biharmonique et une Classe Remarquable de Fonctions Fondamentales Harmonique*, Bulletin International de l' Académie des Sciences de Cracovie, pp. 147-196, 1907.

[40] Zaremba S. , *Sur le Calcul Numérique des Fonctions Demandées dans le Problème de Dirichlet et le Problème Hydrodynamique*, Bulletin International de l' Académie des Sciences de Cracovie, pp. 125-195, 1908.