ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

Γρηγόριος Ταχυρίδης

Μέθοδοι Υποχώρων Krylov για την Επίλυση Γραμμικών Συστημάτων Toeplitz

---

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Ιωάννινα, 2022

Στη μνήμη του
πατέρα μου

Η παρούσα Διδακτορική Διατριβή εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Διδακτορικού Διπλώματος στα Μαθηματικά που απονέμει το Τμήμα Μαθηματικών του Πανεπιστημίου Ιωαννίνων.

Εγκρίθηκε την 19/01/2022 από την εξεταστική επιτροπή:

- Δημήτριος Νούτσος (Επιβλέπων, Ομότιμος Καθηγητής, Πανεπιστήμιο Ιωαννίνων, Ελλάδα)
- Ευστράτιος Γαλλόπουλος (Μέλος συμβουλευτικής επιτροπής, Καθηγητής, Πανεπιστήμιο Πατρών, Ελλάδα)
- Παρασκευάς Βασσάλος (Μέλος συμβουλευτικής επιτροπής, Αναπληρωτής Καθηγητής, Οικονομικό Πανεπιστήμιο Αθηνών, Ελλάδα)
- Μιχαήλ Βραχάτης (Καθηγητής, Πανεπιστήμιο Πατρών, Ελλάδα)
- Φωτεινή Καρακατσάνη (Επίκουρη Καθηγήτρια, Πανεπιστήμιο Ιωαννίνων, Ελλάδα)
- Μιχαήλ Τσατσόμοιρος (Καθηγητής, Washington State University, USA)
- Παναγιώτης Ψαρράκος (Καθηγητής, Εθνικό Μετσόβιο Πολυτεχνείο, Ελλάδα)

Η έγκριση της διδακτορικής διατριβής από το Τμήμα Μαθηματικών της Σχολής Θετικών Επιστημών του Πανεπιστημίου Ιωαννίνων δεν υποδηλώνει την αποδοχή γνωμών του συγγραφέα (Νόμος 5343/1932, άρθρο 202, παρ. 2 και Νόμος 1268/1982, άρθρο 50, παρ. 8).

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

“Δηλώνω υπεύθυνα ότι η παρούσα διατριβή εκπονήθηκε κάτω από τους διεθνείς ηθικούς και ακαδημαϊκούς κανόνες δεοντολογίας και προστασίας της πνευματικής ιδιοκτησίας. Σύμφωνα με τους κανόνες αυτούς, δεν έχω προβεί σε ιδιοποίηση ξένου επιστημονικού έργου και έχω πλήρως αναφέρει τις πηγές που χρησιμοποίησα στην εργασία αυτή.”

Γρηγόριος Ταχυρίδης

Ευχαριστίες

Θα ήθελα να ευχαριστήσω εκ βαθέων τον επιβλέποντα καθηγητή της διδακτορικής μου διατριβής, Δημήτριο Νούτσο, για την αφοσίωση, τη στήριξη, τη συνεισφορά και την αστείρευτη καθοδήγηση κατά τη διάρκεια της έρευνας και υλοποίησης της παρούσας διδακτορικής διατριβής. Θα ήθελα επίσης να ευχαριστήσω τα δύο μέλη της Συμβουλευτικής Επιτροπής, καθηγητές Ευστράτιο Γαλλόπουλο και Παρασκευά Βασσάλο, για τις χρήσιμες υποδείξεις τους και τη βοήθεια που μου παρείχαν. Θα ήθελα να αποδώσω ιδιαίτερες ευχαριστίες στον καθηγητή Απόστολο Χατζηδήμο για τη συνεργασία και τις επισημάνσεις του, καθώς επίσης και στη δόκτορα Chaysri Thaniporn για την αμφίδρομη αλληλεπίδραση. Τέλος, θα ήθελα να ευχαριστήσω το Τμήμα Μαθηματικών του Πανεπιστημίου Ιωαννίνων και ιδιαιτέρως τον Τομέα Εφαρμοσμένων και Υπολογιστικών Μαθηματικών για την παροχή χώρου εργασίας και ηλεκτρονικού υπολογιστή, καθώς και για τις ποικίλες διευκολύνσεις σε επιστημονικό και τεχνικό επίπεδο, που οδήγησαν στην ομαλότερη συνεργασία και διευκόλυνση της έρευνας.

Μέρος της έρευνας συγχρηματοδοτήθηκε από την Ελλάδα και την Ευρωπαϊκή 'Ενωση (Ευρωπαϊκό Κοινωνικό Ταμείο) μέσω του Επιχειρησιακού Προγράμματος Ανάπτυξη Ανθρώπινου Δυναμικού, Εκπαίδευση και Διά Βίου Μάθηση, στα πλαίσια του έργου με τίτλο ""Krylov subspace methods and Perron-Frobenius theory"" (MIS 5047643).

Περίληψη

Στην παρούσα διδακτορική διατριβή γίνεται μελέτη της προρρύθμισης τετραγωνικών, μη-συμμετρικών και πραγματικών συστημάτων Toeplitz. Αποδεικνύονται θεωρητικά αποτελέσματα, τα οποία αποτελούν ικανές συνθήκες για την αποτελεσματικότητα των προτεινόμενων προρρυθμιστών και την ταχεία σύγκλιση στη λύση του συστήματος, με μεθόδους όπως η Προρρυθμισμένη Γενικευμένη μέθοδος Ελαχίστων Υπολοίπων (PGMRES) και η Προρρυθμισμένη μέθοδος Συζυγών Κλίσεων για το σύστημα των Κανονικών Εξισώσεων (PCGN).

Στο πρώτο κεφάλαιο παραθέτουμε βασικές εισαγωγικές έννοιες, ορισμούς και θεωρητικά αποτελέσματα, τα οποία χρησιμοποιήσαμε για να αποδείξουμε τα θεωρητικά αποτελέσματα της διατριβής. Αυτά έχουν να κάνουν κυρίως με τη συσσώρευση του φάσματος, αλλά και των ιδιαζουσών τιμών, αφού αυτή αποτελεί κριτήριο για το πόσο αποτελεσματικός είναι κάποιος προρρυθμιστής.

Στο δεύτερο κεφάλαιο κατασκευάζουμε έναν ταινιωτό Toeplitz προρρυθμιστή, για συστήματα με καλή, αλλά και κακή κατάσταση. Η τεχνική προρρύθμισης βασίζεται στην εύρεση ενός κατάλληλου τριγωνομετρικού πολυωνύμου για την άρση των ριζών της γεννήτριας συνάρτησης (αν υπάρχουν), σε συνδυασμό με προσέγγιση από κάποιο άλλο τριγωνομετρικό πολυώνυμο. Αποδείχθηκε η συσσώρευση των ιδιοτιμών, καθώς και των ιδιαζουσών τιμών του προρρυθμισμένου συστήματος.

Στο τρίτο κεφάλαιο κατασκευάζουμε έναν κυκλοειδή (circulant) προρρυθμιστή για συστήματα Toeplitz με καλή κατάσταση, καθώς κι έναν ταινιωτό-επί-κυκλοειδή (band-times-circulant) προρρυθμιστή για συστήματα με κακή κατάσταση. Αποδεικνύονται αντίστοιχα θεωρητικά αποτελέσματα, ενώ γίνεται και σύγκριση με τον προρρυθμιστή του προηγούμενου κεφαλαίου στα αριθμητικά αποτελέσματα, της τελευταίας ενότητας.

Στο τέταρτο και τελευταίο κεφάλαιο της διατριβής μελετάμε συστήματα Toeplitz, των οποίων η γεννήτρια συνάρτηση υπάρχει, αλλά δεν είναι γνωστή εκ των προτέρων. Γίνεται κατάλληλη προσαρμογή των προρρυθμιστών που κατασκευάστηκαν

στα προηγούμενα κεφάλαια. Με τεχνικές εκτίμησης της γεννήτριας συνάρτησης, των ριζών αυτής και της πολλαπλότητας των εν λόγω ριζών, κατασκευάζουμε αντίστοιχους προρρυθμιστές.

Abstract

In this thesis we study the preconditioning of square, non-symmetric and real Toeplitz systems. We prove theoretical results, which constitute sufficient conditions for the efficiency of the proposed preconditioners and the fast convergence to the solution of the system, by the Preconditioned Generalized Minimal Residual method (PGMRES) as well as by the Preconditioned Conjugate Gradient method applied to the system of Normal Equations (PCGN).

As introduction, in the first chapter, we give the basic definitions and theorems/lemmas that we use to prove the theoretical results of the thesis. These are dealing with the clustering of the eigenvalues, as well as of the singular values, which is a criterion for the efficiency of the preconditioner.

In the second chapter we construct a band Toeplitz preconditioner for well-conditioned, as well as for ill-conditioned systems. The preconditioning technique is based on the elimination of the roots of the generating function (if there exist), by a trigonometric polynomial, and on a further approximation. The clustering of the eigenvalues and the singular values of the preconditioned system has been proven.

In the next chapter we construct a circulant preconditioner dealing with well-conditioned Toeplitz systems and a band-times-circulant preconditioner for ill-conditioned ones. We prove analogous theoretical results and we give a comparison with the preconditioner proposed previously at the numerical results of the last section.

In the fourth and last chapter of the thesis we study Toeplitz systems, having an unknown generating function. We adapt the preconditioners constructed at the previous chapters. After estimating the generating function, its roots and the multiplicities of them, we construct the corresponding preconditioners.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Στην παρούσα διδακτορική διατριβή ασχολούμαστε με την προρρύθμιση τετραγωνικών, μη-συμμετρικών και πραγματικών συστημάτων Toeplitz διάστασης n. Πίνακες Toeplitz εμφανίζονται σε πληθώρα εφαρμογών όπως στην επεξεργασία σήματος, επεξεργασία εικόνας [11, 47], σε εφαρμογές που προκύπτουν από τη διακριτοποίηση διαφορικών και ολοκληρωτικών εξισώσεων [2, 3, 70], καθώς κι εφαρμογές που σχετίζονται με οικονομικά, μηχανική και πιθανότητες [23, 36]. Σε πολλές εξ αυτών μπορεί να συναντήσουμε ακόμα πιο συγκεκριμένες μορφές πινάκων Toeplitz, όπως τριδιαγώνιους μη-συμμετρικούς πίνακες, που εμφανίζονται στην ανάλυση χρονοσειρών και στην επίλυση μερικών διαφορικών εξισώσεων ([48] και αναφορές εντός αυτής). Θα μπορούσαμε να πούμε ότι η ταχεία επίλυση συστημάτων Toeplitz είναι επιτακτική ανάγκη και αποτελεί πρόκληση, διότι πολλές από τις παραπάνω εφαρμογές αναζητούν επίλυση σε πραγματικό χρόνο (real time).

Τα στοιχεία ενός πίνακα Toeplitz είναι οι συντελεστές του αναπτύγματος Fourier μιας συνάρτησης f, οι οποία καλείται γεννήτρια συνάρτηση του πίνακα. Έτσι, αν t_{jk} συμβολίζει το στοιχείο που βρίσκεται στη j-οστή γραμμή και k-οστή στήλη, ισχύει:

$$t_{jk} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \mathrm{e}^{-\mathrm{i}(j-k)x} \mathrm{d}x, \ 1 \leq j, k \leq n, \ \mathrm{i} = \sqrt{-1}. \tag{1.1}$$

Η γεννήτρια συνάρτηση του πίνακα Toeplitz παίζει σημαντικό ρόλο στην επιλογή του προρρυθμιστή. Όταν αυτή δεν έχει ρίζες λαμβάνουμε συστήματα με καλή κατάσταση (well-conditioned), δηλαδή συστήματα των οποίων ο δείκτης κατάστασης (condition number) είναι φραγμένος από σταθερά ανεξάρτητη της διάστασης n, ενώ όταν μηδενίζεται σε κάποια σημεία (ή και διαστήματα) του πεδίου ορισμού της λαμβάνουμε συστήματα με κακή κατάσταση (ill-conditioned). Περισσότερες λεπτομέρειες θα δοθούν στα επόμενα κεφάλαια.

Από τη σχέση (1.1), γίνεται κατανοητό ότι τα στοιχεία ενός πίνακα Toeplitz T_n, εξαρτώνται μόνο από τη διαφορά $j-k$ κι επομένως, ένας πίνακας Toeplitz έχει τα ίδια στοιχεία κατά μήκος των διαγωνίων του, δηλαδή είναι της μορφής:

$$T_n = \begin{pmatrix} t_0 & t_{-1} & t_{-2} & \cdots & t_{-(n-2)} & t_{-(n-1)} \\ t_1 & t_0 & t_{-1} & \cdots & t_{-(n-3)} & t_{-(n-2)} \\ t_2 & t_1 & t_0 & \cdots & t_{-(n-4)} & t_{-(n-3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ t_{n-2} & t_{n-3} & t_{n-4} & \cdots & t_0 & t_{-1} \\ t_{n-1} & t_{n-2} & t_{n-3} & \cdots & t_1 & t_0 \end{pmatrix}.$$

Σκοπός μας είναι η ταχεία και αποτελεσματική λύση ενός πραγματικού και μη συμμετρικού συστήματος Toeplitz της μορφής:

$$T_n(f)x = b, \tag{1.2}$$

όπου $f = f_1 + \mathrm{i} f_2$, με f_1 μια άρτια, 2π-περιοδική συνάρτηση και f_2 περιττή κι επίσης 2π-περιοδική. Τόσο η f_1, όσο και η f_2 ορίζονται στο διάστημα $(-\pi, \pi]$.

Για την αρτιότητα της διατριβής, δίνουμε τον ορισμό του δείκτη κατάστασης ενός αντιστρέψιμου πίνακα:

Ορισμός 1.0.1. Δείκτης κατάστασης ενός αντιστρέψιμου πίνακα $A \in \mathbb{C}^{n,n}$, ως προς μία φυσική νόρμα $\|\cdot\|$, καλείται ο αριθμός $\kappa(A) = \|A\|\|A^{-1}\|$.

'Οσο πιο κοντά στη μονάδα είναι ο δείκτης κατάστασης του πίνακα συντελεστών, τόσο καλύτερη κατάσταση μπορούμε να πούμε ότι έχει το αντίστοιχο σύστημα.

Στη βιβλιογραφία υπάρχουν γνωστοί αλγόριθμοι για την επίλυση συστημάτων Toeplitz, όπως ο αλγόριθμος του N. Levinson [44], του J. Durbin [24] και του W. Trench [83]. Αυτοί αν και αποτελούν βελτίωση των κλασικών άμεσων μεθόδων, όπως της Απαλοιφής Gauss (η οποία αν εφαρμοστεί αυτούσια χρειάζεται $\mathcal{O}(n^3)$ πράξεις), μπορούν να αντιμετωπίσουν ειδικές μορφές συστημάτων Toeplitz [32], απαιτούν $\mathcal{O}(n^2)$ πράξεις και είναι αποτελεσματικοί για συστήματα που έχουν καλή κατάσταση, χωρίς όμως να γνωρίζουμε κάτι για συστήματα με κακή κατάσταση [8, 9]. Αναφέρουμε ότι αναπτύχθηκαν και πιο γρήγορες άμεσες μέθοδοι, οι οποίες μας δίνουν τη λύση του συστήματος σε $\mathcal{O}(n \log^2 n)$ πράξεις ([12] και αναφορές εντός αυτής). Ωστόσο, οι επαναληπτικές μέθοδοι όπως η μέθοδος Συζυγών Κλίσεων (CG) [34, 65], η Γενικευμένη μέθοδος Ελαχίστων Υπολοίπων (GMRES) [68] και άλλες μέθοδοι υποχώρων Krylov, υπερτερούν των άμεσων αφού ο πολλαπλασιασμός πίνακα Toeplitz επί διάνυσμα μπορεί να γίνει σε $\mathcal{O}(n \log n)$ πράξεις με χρήση του Ταχέως Μετασχηματισμού Fourier (Fast Fourier Transform), γνωστό ως FFT [80]. Η διαφορά (στο σύνολο των πράξεων) ανάμεσα στις άμεσες κι επαναληπτικές μεθόδους είναι πολύ μεγαλύτερη σε block Toeplitz συστήματα, όπου το κάθε block είναι πίνακας Toeplitz. Πιο συγκεκριμένα, με αλγορίθμους τύπου Levinson χρειάζονται $\mathcal{O}\left(N^2 M^3\right)$ πράξεις, αντί $\mathcal{O}\left(NM \log NM\right)$ που χρειάζονται οι επαναληπτικές μέθοδοι [57], όπου M είναι η διάσταση του κάθε block και N είναι το πλήθος των blocks σε κάθε γραμμή του πίνακα.

Αν και οι επαναληπτικές μέθοδοι χρειάζονται $\mathcal{O}(n \log n)$ πράξεις σε κάθε επανάληψη, δεν είναι ιδιαίτερα αποτελεσματικές για συστήματα με κακή κατάσταση, με την έννοια ότι χρειάζονται πολλές επαναλήψεις (ίσως όσες και η διάσταση του συστήματος). Αυτό το φαινόμενο μπορεί να εξαλειφθεί με την τεχνική της προρρύθμισης. Ουσιαστικά προσπαθούμε να βρούμε έναν αντιστρέψιμο πίνακα ο οποίος όχι μόνο κατασκευάζεται γρήγορα, αλλά δίνει και τη λύση του αντίστοιχου συστήματος επίσης γρήγορα. ‘Οσον αφορά στα συστήματα Toeplitz, θέλουμε/απαιτούμε η κατασκευή και επίλυση του προρρυθμισμένου συστήματος να γίνεται το πολύ σε $\mathcal{O}(n \log n)$ πράξεις, διότι όπως προαναφέραμε για τον πολλαπλασιασμό πίνακα Toeplitz επί διάνυσμα απαιτούνται $\mathcal{O}(n \log n)$ πράξεις. Θα θέλαμε να σχολιάσουμε ότι για να γίνει χρήση του FFT, θα πρέπει αρχικά να εμβαπτίσουμε τον πίνακα T_n, σε μια συγκεκριμένη κυκλοειδή (circulant) μορφή διπλάσιας διάστασης. Περισσότερες λεπτομέρειες μπορούν να βρεθούν στο κλασικό βιβλίο του M. Ng [47].

1.1 Βασική θεωρία

Η προρρύθμιση συμμετρικών και θετικά ορισμένων συστημάτων Toeplitz έχει μελετηθεί εκτενώς από πολλούς ερευνητές και είναι πλέον ευρέως γνωστό ότι αυτά μπορούν να επιλυθούν αποτελεσματικά με την Προρρυθμισμένη μέθοδο Συζυγών Κλίσεων (PCG). Οι προρρυθμιστές που εισήχθησαν τα τελευταία χρόνια είναι ιδιαιτέρως αποτελεσματικοί και κάποιοι από αυτούς, όπως για παράδειγμα αυτός στην [52], οδηγούν στην υπεργραμμική (superlinear) σύγκλιση της PCG, παρέχοντας τη λύση του συστήματος σε λίγες επαναλήψεις, ανεξάρτητες της διάστασης n.

Ειδικότερα, το 1991 ο R. Chan [10] εισήγαγε έναν ταινιωτό Toeplitz προρρυθμιστή για συμμετρικά συστήματα Toeplitz με κακή κατάσταση. Η γεννήτρια συνάρτηση του προτεινόμενου προρρυθμιστή ήταν ένα τριγωνομετρικό πολυώνυμο g, το οποίο είχε τις ίδιες ρίζες, καθώς επίσης και την ίδια πολλαπλότητα ριζών, με τη γεννήτρια συνάρτηση (του αρχικού συστήματος) f. Αυτή η τεχνική οδηγεί στην άρση της κακής κατάστασης του συστήματος, αφού η $\frac{f}{g}$ λαμβάνει μόνο θετικές τιμές. Το 1994 οι R. Chan και P. Tang [13] εισήγαγαν έναν ταινιωτό Toeplitz προρρυθμιστή, ο οποίος προέκυψε από το τριγωνομετρικό πολυώνυμο g, το οποίο ελαχιστοποιούσε το μέγιστο σχετικό σφάλμα $\|\frac{f-g}{f}\|_{\infty}$. Το 1997 ο S. Serra-Capizzano [73] πρότεινε εναλλακτικούς τρόπους για την ελαχιστοποίηση του $\|\frac{f-g}{f}\|_{\infty}$ και εισήγαγε με τη σειρά του έναν ταινιωτό Toeplitz προρρυθμιστή, η κατασκευή του οποίου συνδυάζει την άρση της κακής κατάστασης (με κάποιο τριγωνομετρικό πολυώνυμο z_k) και την προσέγγιση της συνάρτησης $\frac{f}{z_k}$, με παρεμβολή στα σημεία Chebyshev πρώτου είδους [66] ή με βέλτιστη ομοιόμορφη προσέγγιση λαμβάνοντας ως κόμβους τα ίδια σημεία. Το 2002 οι D. Noutsos και P. Vassalos [51] πρότειναν έναν προρρυθμιστή, ο οποίος είναι το αποτέλεσμα της άρσης των ριζών της f, με το z_k, και της ρητής προσέγγισης της συνάρτησης $\sqrt{\frac{f}{z_k}}$. Οι ίδιοι συγγραφείς το 2008 [52] πρότειναν έναν προρρυθμιστή, ο οποίος κατασκευάζεται ως το γινόμενο ενός τριγωνομετρικού πολυωνύμου g, που αίρει τις ρίζες της μη-αρνητικής γεννήτριας συνάρτησης f, και πίνακες που ανήκουν σε τριγωνομετρική άλγεβρα [47] και αντιστοιχούν στη θετική συνάρτηση $\frac{f}{g}$. Το 2016 οι D. Noutsos, S. Serra-Capizzano και P. Vassalos [59] πρότειναν έναν προρρυθμιστή ο οποίος ανήκει στην τ-άλγεβρα [5, 22], για συστήματα Toeplitz των οποίων η γεννήτρια συνάρτηση έχει ρίζες μη-ακέραιας τάξης.

Σε πολλές εφαρμογές εμφανίζονται Block Toeplitz πίνακες με Toeplitz Blocks (BTTB) [11, 47, 40]. Αυτοί αποτελούν μια γενίκευση της κλάσης των πινάκων Toeplitz. Οι BTTB πίνακες καλούνται δι-διάστατοι (two-level) πίνακες Toeplitz και συμβολίζονται ως $T_{nm}(f)$, όπου m είναι η διάσταση του κάθε block και n είναι το πλήθος των blocks σε κάθε γραμμή του πίνακα, που σημαίνει ότι $T_{nm}(f) \in \mathbb{R}^{nm \times nm}$. Η γεννήτρια συνάρτηση αυτών είναι η 2π-περιοδική συνάρτηση δύο μεταβλητών $f = f(x, y) : [-\pi, \pi]^2 \to \mathbb{C}$. Κάθε στοιχείο t_{rs} χαρακτηρίζεται από τις παραμέτρους r και s, όπου r συμβολίζει τη block διαγώνιο και s τη διαγώνιο εντός των blocks. ʻΟπως και στη μονοδιάστατη περίπτωση οι τιμές του πίνακα BTTB είναι οι συντελεστές του αναπτύγματος Fourier της f:

$$t_{rs} = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} f(x, y) \mathrm{e}^{-\mathrm{i}(rx+sy)} \mathrm{d}x \mathrm{d}y.$$

Συμμετρικά και θετικά ορισμένα BTTB συστήματα, έχουν μελετηθεί επίσης από πολλούς ερευνητές. Αποτελεσματικοί BTTB προρρυθμιστές προτάθηκαν από τον M. Ng [46], καθώς επίσης και από τους D. Noutsos, S. Serra-Capizzano και P. Vassalos [56, 57, 58]. Σημειώνουμε ότι οι τελευταίοι προρρυθμιστές κατασκευάστηκαν μετά από την προσέγγιση της γεννήτριας συνάρτησης, προσαρμόζοντας την ιδέα της [71] που αφορούσε στη μονοδιάστατη περίπτωση. Προρρυθμιστές που ανήκουν σε άλγεβρα πινάκων προτάθηκαν στην [53] για συστήματα με καλή κατάσταση. Σχολιάζουμε ότι για συστήματα με κακή κατάσταση έχουν αποδειχθεί ορισμένα αρνητικά αποτελέσματα [54, 55, 74, 76, 77]. Εν ολίγοις, προρρυθμιστές από οποιαδήποτε τριγωνομετρική άλγεβρα πινάκων δεν είναι ικανοί να οδηγήσουν σε υπεργραμμική σύγκλιση, συστήματα με κακή κατάσταση από μόνοι τους, που σημαίνει ότι δεν είναι αποτελεσματικοί χωρίς κάποια τεχνική άρσης της κακής κατάστασης. Από την άλλη, ο συνδυασμός ταινιωτών BTTB πινάκων με πίνακες οι οποίοι ανήκουν σε κάποια άλγεβρα (πινάκων), είναι ιδιαιτέρως αποτελεσματικός όπως φαίνεται στην [53], αφού επιτεύχθηκε η συσσώρευση των ιδιοτιμών γύρω από το 1.

Παραπάνω παρουσιάσαμε συνοπτικά ποικίλες τεχνικές προρρύθμισης από τη βιβλιογραφία, συμπεραίνοντας ότι η συμμετρική περίπτωση συστημάτων Toeplitz μελετήθηκε εκτενώς. Ωστόσο, η περίπτωση μη-συμμετρικών συστημάτων Toeplitz χρήζει μελέτης, αφού υπάρχουν πολλά σημεία περαιτέρω ανάλυσης. Ο αναγνώστης μπορεί να βρει διάφορα θεωρητικά εποτελέσματα για τη συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών στις [61, 81, 82, 86, 87]. Σε αυτές δίνονται γενικεύσεις ενός βασικού θεωρήματος για την ισοκατανομή των ιδιοτιμών. Αυτό είναι το θεώρημα ισοκατανομής (equally distribution) του G. Szegö [33]. Θα το δώσουμε, αφού αρχικά ορίσουμε το ουσιώδες άνω και κάτω φράγμα μιας συνάρτησης.

Ορισμός 1.1.1. Ο μέγιστος αριθμός m για τον οποίο ισχύει η ανισότητα $f(x) \geq m$, με εξαίρεση ένα σύνολο μέτρου Lebesgue μηδέν, λέγεται ουσιώδες κάτω φράγμα (essential lower bound) της συνάρτησης f.

Ορισμός 1.1.2. Ο ελάχιστος αριθμός M για τον οποίο ισχύει η ανισότητα $f(x) \leq M$, με εξαίρεση ένα σύνολο μέτρου Lebesgue μηδέν, λέγεται ουσιώδες άνω φράγμα (essential upper bound) της συνάρτησης f.

Θεώρημα 1.1.3 (G. Szegö). Ας είναι f μία πραγματική συνάρτηση, ολοκληρώσιμη κατά Lebesgue και $\lambda_1, \lambda_2, \ldots, \lambda_n$ οι ιδιοτιμές του πίνακα $T_n(f)$. Συμβολίζουμε με m και M το ουσιώδες κάτω και ουσιώδες άνω φράγμα της f, αντίστοιχα. Αν $F(\lambda)$ είναι μία οποιαδήποτε συνεχής συνάρτηση ορισμένη στο διάστημα $[m, M]$ ισχύει:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} F(\lambda_k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F\left(f(x)\right) \mathrm{d}x.$$

Αν στο παραπάνω θεώρημα επιλέξουμε ως $F(x) = x$, προφανώς:

$$\lim_{n\to\infty} \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_n}{n} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\mathrm{d}x.$$

Αυτό σημαίνει ότι ο μέσος όρος των ιδιοτιμών του $T_n(f)$, συγκλίνει στο στοιχείο $t_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\mathrm{d}x$. Γενικότερα, η κατάλληλη επιλογή της F είναι αυτή που μπορεί να μας οδηγήσει στο επιθυμητό αποτέλεσμα (βλ. απόδειξη Θεωρήματος 2.1.2).

Καταλαβαίνουμε ότι ο τρόπος κατανομής/συγκέντρωσης των ιδιοτιμών παίζει σημαντικό ρόλο για την αποτελεσματικότητα μεθόδων Krylov. Είναι μάλιστα γνωστό ότι η αποτελεσματικότητα της μεθόδου GMRES εξαρτάται από τον τρόπο συσσώρευσης των ιδιοτιμών του πίνακα συντελεστών, ενώ αυτή της CGN από τη συσσώρευση των ιδιαζουσών τιμών [45]. Αναφερόμαστε σε αυτές τις δύο μεθόδους, διότι μπορούν να επιλύσουν ένα μη-συμμετρικό σύστημα, σε αντίθεση με τη PCG, η οποία ναι μεν επιλύει με αποτελεσματικότητα συμμετρικά και θετικά ορισμένα συστήματα, αλλά δε μπορεί να εφαρμοστεί σε μη-συμμετρικούς πίνακες. Παρακάτω ορίζουμε δύο είδη συσσώρευσης των ιδιοτιμών, όπως δόθηκαν από τον E. Tyrtyshnikov στην [86]:

Ορισμός 1.1.4. Ένα σύνολο $\Phi \subset \mathbb{C}$ καλείται σύνολο γενικής συσσώρευσης (general cluster) των ιδιοτιμών μιας ακολουθίας πινάκων $\{A_n\}$, $A_n \in \mathbb{C}^{n\times n}$, αν και μόνο αν για κάθε $\varepsilon > 0$:

$$\lim_{n\to\infty} \frac{\gamma_n(\varepsilon)}{n} = 0,$$

ενώ αν:

$$\gamma_n(\varepsilon) \leq c(\varepsilon),$$

όπου $c(\varepsilon)$ είναι σταθερά ανεξάρτητη της διάστασης n, το Φ καλείται σύνολο κύριας συσσώρευσης (proper cluster) των ιδιοτιμών.

Παραπάνω, με $\gamma_n(\varepsilon)$ συμβολίζουμε τον αριθμό των ιδιοτιμών που κυμαίνονται εκτός του πεδίου Φ_ε (outliers), δηλαδή εκτός της ε-επέκτασης του Φ (ένωση του Φ με όλες τις ε-μπάλες, που έχουν ως κέντρο τα σημεία του Φ).

Παρατήρηση. Προσαρμόσαμε τον ορισμό από την [86] για μη-Ερμιτιανούς πίνακες, διότι οι ιδιοτιμές του μη-συμμετρικού πίνακα Toeplitz είναι μιγαδικές. Υπάρχει ανάλογος ορισμός για τη συσσώρευση των ιδιαζουσών τιμών [78].

Δίνουμε ένα θεώρημα, το οποίο θα φανεί χρήσιμο στην απόδειξη της συσσώρευσης των ιδιαζουσών τιμών του προρρυθμισμένου συστήματος, στο επόμενο κεφάλαιο. Αυτό είναι το Θεώρημα 4.5 της [81], το οποίο δόθηκε από τον P. Tilli:

Θεώρημα 1.1.5 (P. Tilli). Έστω $f \in \mathrm{L}^2(Q, \mathbb{C}^{h\times k})$ και ας είναι $\{T_n\}$ ένα σύνολο block πινάκων Toeplitz με γεννήτρια συνάρτηση f. Τότε, το $[\sigma_{\min}(f), \sigma_{\max}(f)]$ αποτελεί διάστημα συσσώρευσης για τις ιδιάζουσες τιμές του $\{T_n\}$.

Σημειώνουμε ότι $Q = (-\pi, \pi)$ και $f \in \mathrm{L}^2(Q, \mathbb{C}^{h\times k})$, σημαίνει ότι η f είναι μια συνάρτηση πινάκων στο $\mathbb{C}^{h\times k}$ με οποιοδήποτε στοιχείο $f_{ij} \in \mathrm{L}^2(-\pi, \pi)$.

Παρατήρηση. Η συσσώρευση στο παραπάνω θεώρημα χαρακτηρίζεται ως γενική, αφού στην [81] περιγράφεται ότι o(n) ιδιάζουσες τιμές του T_n είναι μικρότερες από $\sigma_{\min}(f)$ και όλες τους μικρότερες από $\sigma_{\max}(f)$.

Χρήσιμο επίσης είναι και το παρακάτω λήμμα (Λήμμα 2.1 της [87]), το οποίο δόθηκε από τους E. Tyrtyshnikov και N. Zamarashkin:

Λήμμα 1.1.6 (E. Tyrtyshnikov, N. Zamarashkin). Δοσμένων δύο ακολουθιών πινάκων $\{\mathcal{A}_n\}$ και $\{\mathcal{B}_n\}$, υποθέτουμε ότι $\|\mathcal{A}_n - \mathcal{B}_n\|_F^2 = o(n)$ και $\|\mathcal{B}_n\|_2 \leq M$ ομοιόμορφα ως προς το n. Τότε, το $\{z : |z| \leq M\}$ αποτελεί σύνολο γενικής συσσώρευσης των ιδιοτιμών του $\mathcal{A}_n$.

Παρατήρηση. Αν $\|\mathcal{A}_n - \mathcal{B}_n\|_F^2 = \mathcal{O}(1)$ (και $\|\mathcal{B}_n\|_2 \leq M$ ομοιόμορφα ως προς το n), εύκολα βλέπουμε, από την απόδειξη του αντίστοιχου λήμματος της [87], ότι το $\{z : |z| \leq M\}$ αποτελεί σύνολο κύριας συσσώρευσης των ιδιοτιμών του $\{\mathcal{A}_n\}$.

Είναι προφανές ότι η συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών αναφέρεται σε ακολουθία πινάκων. Ωστόσο, για λόγους απλούστευσης, στην πορεία της διατριβής όταν αναφερόμαστε στη συσσώρευση πίνακα, θα εννοούμε συσσώρευση της αντίστοιχης ακολουθίας πινάκων.

Κεφάλαιο 2

Ταινιωτοί Προρρυθμιστές

Σε αυτό το κεφάλαιο θα μελετήσουμε την προρρύθμιση $n \times n$ πραγματικών και μη-συμμετρικών συστημάτων Toeplitz, χρησιμοποιώντας ως προρρυθμιστές ταινιωτούς πίνακες Toeplitz. Η γεννήτρια συνάρτηση που αντιστοιχεί σε αυτά είναι μιγαδική, της μορφής $f = f_1 + \mathrm{i} f_2$, όπου $\mathrm{i} = \sqrt{-1}$. Σημειώνουμε επίσης ότι η f_1 είναι 2π-περιοδική και άρτια συνάρτηση ενώ η f_2 είναι μεν 2π-περιοδική, αλλά περιττή. Η λύση των προρρυθμισμένων συστημάτων λαμβάνεται με την Προρρυθμισμένη μέθοδο Γενικευμένων Ελαχίστων Υπολοίπων (PGMRES) [67, 69] και την Προρρυθμισμένη μέθοδο Συζυγών Κλίσεων για το σύστημα των Κανονικών Εξισώσεων (PCGN). Θα παρουσιάσουμε μια τεχνική προρρύθμισης η οποία συνδυάζει την άρση των ριζών της γεννήτριας συνάρτησης (αν υπάρχουν) και βέλτιστη ομοιόμορφη προσέγγιση ή παρεμβολή με τριγωνομετρικά πολυώνυμα. Ο προρρυθμιστής που προκύπτει είναι ο πίνακας $T_n(p)$, όπου $p = gq$, με g να είναι το τριγωνομετρικό πολυώνυμο το οποίο έχει τις ίδιες ρίζες με τη γεννήτρια συνάρτηση f και q το τριγωνομετρικό πολυώνυμο που προκύπτει κατόπιν προσέγγισης της $\frac{f}{g}$. Με χρήση του προαναφερθέντος προρρυθμιστή επιτυγχάνεται συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών του προρρυθμισμένου συστήματος σε μια μικρή περιοχή μακριά από το 0 και σε ένα μικρό διάστημα κοντά στο 1, αντίστοιχα. Αυτό σημαίνει ότι μπορούμε να λάβουμε τη λύση του συστήματος με λίγες επαναλήψεις των μεθόδων που αναφέραμε [45], γεγονός το οποίο επιβεβαιώνεται και στα διάφορα αριθμητικά παραδείγματα που δίνουμε στο τέλος του κεφαλαίου.

2.1 Θεωρητικά αποτελέσματα

Αρχικά θα δώσουμε τα θεωρητικά αποτελέσματα, που αφορούν στη σύγκλιση των μεθόδων PGMRES και PCGN, δηλαδή στη συσσώρευση του φάσματος των ιδιοτιμών και ιδιαζουσών τιμών του προρρυθμισμένου συστήματος. Όπως αναφέραμε, στόχος μας είναι να βρούμε έναν κατάλληλο ταινιωτό πίνακα Toeplitz $T_n(p)$, τον οποίο θα χρησιμοποιήσουμε ως προρρυθμιστή, για την αποτελεσματική επίλυση του συστήματος $T_n(f)x = b$. Για να επιτύχουμε τον στόχο μας, θα πρέπει να επιλέξουμε ένα τριγωνομετρικό πολυώνυμο $p = p_1 + \mathrm{i}p_2$ έτσι ώστε $\mathrm{Re}\left(\frac{f}{p}\right) > 0$ και το εύρος (range) της $\left|\frac{f}{p}\right|$, ορισμένο ως:

$$\mathrm{range}\left(\left|\frac{f}{p}\right|\right) = \left[\min_{-\pi\le x\le\pi}\left|\frac{f(x)}{p(x)}\right|, \max_{-\pi\le x\le\pi}\left|\frac{f(x)}{p(x)}\right|\right],$$

να είναι ένα θετικό διάστημα μακριά από το 0.

Ο τρόπος επιλογής του τριγωνομετρικού πολυωνύμου p, θα περιγραφεί λεπτομερώς στην επόμενη ενότητα, όπου δίνεται επίσης και ο τρόπος κατασκευής του προρρυθμιστή. Στη συνέχεια αυτής της ενότητας, θεωρούμε ότι το p έχει ήδη βρεθεί και θα μελετήσουμε τις ιδιότητες που αφορούν στη συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών του προρρυθμισμένου πίνακα. Έτσι, δίνουμε τα ακόλουθα θεωρήματα με τις αποδείξεις τους:

Θεώρημα 2.1.1. Έστω $f \in \mathrm{L}^2(-\pi,\pi)$ και p ένα τριγωνομετρικό πολυώνυμο τέτοιο ώστε $\mathrm{Re}\left(\frac{f}{p}\right) > 0$ και $0 < \alpha = \operatorname*{ess\,inf}_{-\pi\le x\le\pi}\left|\frac{f(x)}{p(x)}\right| \le \operatorname*{ess\,sup}_{-\pi\le x\le\pi}\left|\frac{f(x)}{p(x)}\right| = \beta < \infty$. Τότε, το διάστημα $[\alpha,\beta]$ αποτελεί ένα σύνολο γενικής συσσώρευσης για τις ιδιάζουσες τιμές του προρρυθμισμένου πίνακα $T_n^{-1}(p)T_n(f)$.

Απο404δειξη. Είναι ευρέως γνωστό ότι οι ιδιάζουσες τιμές ενός μη-συμμετρικού και πραγματικού πίνακα A είναι οι τετραγωνικές ρίζες των ιδιοτιμών του A^TA. Επομένως, θα μελετήσουμε τη συμπεριφορά των ιδιοτιμών του πίνακα που αντιστοιχεί στις κανονικές εξισώσεις του προρρυθμισμένου συστήματος.

$$\begin{aligned}\left(T_n^{-1}(p)T_n(f)\right)^T T_n^{-1}(p)T_n(f) &= T_n^T(f)T_n^{-T}(p)T_n^{-1}(p)T_n(f)\\ &= T_n(\bar{f})T_n^{-1}(\bar{p})T_n^{-1}(p)T_n(f).\end{aligned}$$

Ο τελευταίος πίνακας είναι όμοιος με τον:

$$A_n(f,p) = T_n^{-1}(p)T_n(f)T_n(\bar{f})T_n^{-1}(\bar{p}).$$

Για απλούστευση θα μελετήσουμε τη συμπεριφορά των ιδιοτιμών του $A_n(f,p)$. Επίσης, θα κάνουμε χρήση της παρακάτω σχέσης:

$$T_n(f) = T_n(p)T_n\left(\frac{f}{p}\right) + E_n, \tag{2.1}$$

όπου E_n είναι ένας πίνακας που έχει βαθμίδα ίση με $d-1$, d είναι το πλάτος ταινίας (bandwidth) του $T_n(p)$. Εφόσον ο $T_n(p)$ είναι ένας ταινιωτός πίνακας, είναι εύκολο να παρατηρήσουμε ότι οι πίνακες $T_n(f)$ και $T_n(p)T_n\left(\frac{f}{p}\right)$ διαφέρουν μόνο στις $\frac{d-1}{2}$ πρώτες και τελευταίες γραμμές, γεγονός το οποίο αποδεικνύει τη σχέση (2.1). Ομοίως, ισχύει ότι:

$$T_n(\bar{f}) = T_n\left(\frac{\bar{f}}{\bar{p}}\right)T_n(\bar{p}) + E_n^T.$$

Τώρα θα μελετήσουμε το φάσμα του $A_n(f,p)$.

$$\begin{aligned}
A_n(f,p) &= T_n^{-1}(p)T_n(f)T_n(\bar{f})T_n^{-1}(\bar{p}) \\
&= T_n^{-1}(p)\left(T_n(p)T_n\left(\frac{f}{p}\right) + E_n\right) \\
&\qquad \cdot \left(T_n\left(\frac{\bar{f}}{\bar{p}}\right)T_n(\bar{p}) + E_n^T\right)T_n^{-1}(\bar{p}) \\
&= T_n\left(\frac{f}{p}\right)T_n\left(\frac{\bar{f}}{\bar{p}}\right) + T_n^{-1}(p)E_nT_n\left(\frac{\bar{f}}{\bar{p}}\right) \\
&\qquad + T_n\left(\frac{f}{p}\right)E_n^TT_n^{-1}(\bar{p}) + T_n^{-1}(p)E_nE_n^TT_n^{-1}(\bar{p}) \\
&= T_n\left(\frac{f}{p}\right)T_n\left(\frac{\bar{f}}{\bar{p}}\right) + R_n
\end{aligned} \tag{2.2}$$

όπου R_n είναι πίνακας βαθμίδας ίσης το πολύ με $2d-2$.

Από το Θεώρημα 1.1.5 (Θεώρημα 4.5 της [81]), λαμβάνουμε ότι οι ιδιάζουσες τιμές του $T_n\left(\frac{f}{p}\right)$ ή ισοδύναμα οι τετραγωνικές ρίζες των ιδιοτιμών του $T_n\left(\frac{f}{p}\right)T_n\left(\frac{\bar{f}}{\bar{p}}\right)$, συσσωρεύονται στο διάστημα $[\alpha,\beta]$. Η φύση της συσσώρευσης (στο $[\alpha,\beta]$) μέσω του Θεωρήματος 1.1.5 είναι ότι $o(n)$ ιδιάζουσες τιμές του $T_n\left(\frac{f}{p}\right)$ είναι μικρότερες από α και όλες τους μικρότερες από β. Από την (2.2) έχουμε ότι:

$$A_n(f,p) = T_n\left(\frac{f}{p}\right)T_n\left(\frac{\bar{f}}{\bar{p}}\right) + R_n.$$

Επομένως, $o(n)$ ιδιάζουσες τιμές του $T_n^{-1}(p)T_n(f)$ είναι μικρότερες από α και το πολύ $2d-2$ επιπλέον (ιδιάζουσες τιμές) κυμαίνονται εκτός του $[\alpha,\beta]$, το οποίο σημαίνει ότι ένας σταθερός αριθμός ιδιαζουσών τιμών μπορεί να έχουν τιμή μεγαλύτερη του β. Η συσσώρευση αυτού του είδους, έχει οριστεί από τον E. Tyrtyshnikov ως γενική (συσσώρευση) [86]. Επομένως λαμβάνουμε τη γενική συσσώρευση των ιδιαζουσών τιμών του προρρυθμισμένου πίνακα $T_n^{-1}(p)T_n(f)$ και η απόδειξη ολοκληρώθηκε. □

Πρέπει να αναφέρουμε ότι το Θεώρημα 4.5 στην [81], έχει δοθεί σε ένα γενικότερο πλαίσιο, για πίνακες block Toeplitz, όπου η f είναι συνάρτηση πίνακας. Η περίπτωση που μας ενδιαφέρει, σχετίζεται με ένα απλούστερο πλαίσιο, όπου όλες οι υποθέσεις του εν λόγω θεωρήματος επίσης ισχύουν.

Θεώρημα 2.1.2. Έστω $f \in \mathrm{L}^1([-\pi,\pi])$ και p ένα τριγωνομετρικό πολυώνυμο τέτοιο ώστε $\operatorname*{ess\,inf}_{-\pi\le x\le\pi} \mathrm{Re}\left(\frac{f(x)}{p(x)}\right) > 0$. Τότε, οι ιδιοτιμές του προρρυθμισμένου πίνακα $T_n^{-1}(p)T_n(f)$ βρίσκονται εντός του ορθογωνίου $\mathcal{R} = [\alpha,\beta]\times[-\gamma,\gamma]$, όπου $\alpha = \operatorname*{ess\,inf}_{-\pi\le x\le\pi} \mathrm{Re}\left(\frac{f(x)}{p(x)}\right) > 0$, $\beta = \operatorname*{ess\,sup}_{-\pi\le x\le\pi} \mathrm{Re}\left(\frac{f(x)}{p(x)}\right)$ και $\gamma = \operatorname*{ess\,sup}_{-\pi\le x\le\pi} \mathrm{Im}\left(\frac{f(x)}{p(x)}\right)$, εκτός ίσως από το πολύ $2d-2$ ιδιοτιμές, οι οποίες ανήκουν σε μια ε-επέκταση του $\mathcal{R}$. Άρα, το $\mathcal{R}$ αποτελεί σύνολο κύριας συσσώρευσης των ιδιοτιμών, όπου d συμβολίζει το πλάτος ταινίας του $T_n(p)$.

Απο404δειξη. Θα μελετήσουμε το εύρος φάσματος (range) του προρρυθμισμένου πίνακα $T_n(p)^{-1}T_n(f)$:

$$\begin{aligned}
A = \frac{x^H T_n(p)^{-1}T_n(f)x}{x^H x} &= \frac{1}{2}\frac{x^H\left(T_n(p)^{-1}T_n(f) + T_n(\bar{f})T_n(\bar{p})^{-1}\right)x}{x^H x}\\
&+ \frac{1}{2}\mathrm{i}\frac{x^H\left(T_n(p)^{-1}T_n(f) - T_n(\bar{f})T_n(\bar{p})^{-1}\right)x}{\mathrm{i}x^H x}\\
&= \frac{1}{2}\frac{y^H\left(T_n(f)T_n(\bar{p}) + T_n(p)T_n(\bar{f})\right)y}{y^H T_n(p)T_n(\bar{p})y}\\
&+ \frac{1}{2}\mathrm{i}\frac{y^H\left(T_n(f)T_n(\bar{p}) - T_n(p)T_n(\bar{f})\right)y}{\mathrm{i}y^H T_n(p)T_n(\bar{p})y}\\
&= \frac{1}{2}\frac{y^H T_n(f\bar{p}+p\bar{f})y + y^H R_1 y}{y^H T_n(|p|^2)y - y^H R_2 y} + \frac{1}{2}\mathrm{i}\frac{y^H T_n(f\bar{p}-p\bar{f})y + y^H R_3 y}{\mathrm{i}y^H T_n(|p|^2)y - y^H R_2 y},
\end{aligned} \tag{2.3}$$

όπου $y = T_n(\bar{p})^{-1}x$ και R_1, R_2, R_3 είναι πίνακες χαμηλής βαθμίδας (low-rank), ίσης με $d-1$. Αυτοί οι πίνακες έχουν μη-μηδενικά στοιχεία στις $\frac{d-1}{2}$ πρώτες και τελευταίες γραμμές και στήλες, λόγω της ταινιωτής δομής των $T_n(p)$ και $T_n(\bar{p})$.

Είναι προφανές ότι ο πίνακας $T_n(f)T_n(\bar{p}) + T_n(p)T_n(\bar{f})$ είναι συμμετρικός, ενώ ο $\frac{1}{\mathrm{i}}\left(T_n(f)T_n(\bar{p}) - T_n(p)T_n(\bar{f})\right)$ Ερμιτιανός. Διαγράφοντας τις $\frac{d-1}{2}$ πρώτες και τελευταίες γραμμές και στήλες του $T_n(f)T_n(\bar{p}) + T_n(p)T_n(\bar{f})$, καθώς επίσης και του $T_n(f\bar{p} + p\bar{f})$, λαμβάνουμε ότι οι εναπομείναντες πίνακες ταυτίζονται. Το ίδιο ισχύει και για τους $\frac{1}{\mathrm{i}}\left(T_n(f)T_n(\bar{p}) - T_n(p)T_n(\bar{f})\right)$ και $\frac{1}{\mathrm{i}}T_n(f\bar{p} - p\bar{f})$, όπως επίσης και για τους $T_n(p)T_n(\bar{p})$ και $T_n(|p|^2)$.

Χρησιμοποιούμε τη σχέση (2.3), όπου το x επιλέγεται από τον υπόχωρο $\mathcal{V}$ του $\mathbb{R}^n$, με $\dim \mathcal{V} = n-(d-1)$ και τέτοιο ώστε το $y = T_n(\bar{p})^{-1}x$ να έχει μηδενικά στις πρώτες και τελευταίες $\frac{d-1}{2}$ συνιστώσες. Έχουμε:

$$\begin{aligned}
&\operatorname*{range}_{x\in\mathcal{V}}\left(\mathrm{Re}\left(T_n(p)^{-1}T_n(f)\right)\right) = \operatorname*{range}_{x\in\mathcal{V}}\left(\mathrm{Re}\left(T_n\left(\frac{f}{p}\right)\right)\right)\\
&\subset \operatorname*{range}_{x\in\mathbb{R}^n}\left(\mathrm{Re}\left(T_n\left(\frac{f}{p}\right)\right)\right) = \mathcal{ER}\left(\mathrm{Re}\left(\frac{f}{p}\right)\right),
\end{aligned}$$

όπου με $\mathcal{ER}(h)$ συμβολίζουμε το ουσιώδες εύρος (essential range) της συνάρτησης h.

Αυτό σημαίνει ότι το πολύ $d-1$ ιδιοτιμές του συμμετρικού μέρους του πίνακα $T_n(p)^{-1}T_n(f)$ κυμαίνονται εκτός του $\mathcal{ER}\left(\mathrm{Re}\left(\frac{f}{p}\right)\right)$, το οποίο είναι το διάστημα:

$$[\alpha,\beta] = \left[\operatorname*{ess\,inf}_{-\pi\le x\le\pi}\mathrm{Re}\left(\frac{f(x)}{p(x)}\right), \operatorname*{ess\,sup}_{-\pi\le x\le\pi}\mathrm{Re}\left(\frac{f(x)}{p(x)}\right)\right], \text{ όπου } \alpha > 0.$$

Ομοίως, για το αντισυμμετρικό μέρος προκύπτει ότι το πολύ $d-1$ ιδιοτιμές κυμαίνονται εκτός του $\mathcal{ER}\left(\mathrm{Im}\left(\frac{f}{p}\right)\right)$, δηλαδή του διαστήματος:

$$[-\gamma,\gamma] = \left[-\operatorname*{ess\,sup}_{-\pi\le x\le\pi}\mathrm{Im}\left(\frac{f(x)}{p(x)}\right), \operatorname*{ess\,sup}_{-\pi\le x\le\pi}\mathrm{Im}\left(\frac{f(x)}{p(x)}\right)\right].$$

Συμπερασματικά, το πολύ $2(d-1)$ ιδιοτιμές κυμαίνονται εκτός του ορθογωνίου $\mathcal{R} = [\alpha,\beta]\times[-\gamma,\gamma]$, το οποίο αποδεικνύει την κύρια συσσώρευση.

Το ερώτημα που γεννάται είναι: ""Πόσο μακριά από το σύνορο του $\mathcal{R}$, κυμαίνονται οι ιδιοτιμές?"" Προκειμένου να δώσουμε απάντηση σε αυτό, θα πρέπει να εκτιμήσουμε το πηλίκο Rayleigh $\frac{x^H T_n(p)^{-1}T_n(f)x}{x^H x}$, θεωρώντας ότι το x είναι ένα ιδιοδιάνυσμα του $T_n(p)^{-1}T_n(f)$, το οποίο αντιστοιχεί στην ιδιοτιμή λ. Οπότε,

$$\lambda = \frac{x^H T_n(p)^{-1}T_n(f)x}{x^H x} \overset{y=T_n(\bar{p})^{-1}x}{=} \frac{y^H T_n(f)T_n(\bar{p})y}{y^H T_n(p)T_n(\bar{p})y}.$$

Είναι γνωστό ότι παρόλο που οι $T_n(f)T_n(\bar{p})$ και $T_n(p)T_n(\bar{p})$ δεν είναι πίνακες Toeplitz, όσο $n \to \infty$, συμπεριφέρονται ως τελεστές Toeplitz που γεννιούνται από τις $f\bar{p}$ και $p\bar{p}$, αντίστοιχα. Σε κάθε $x \in [-\pi, \pi]$ αντιστοιχεί μια ιδιοτιμή $f(x)\bar{p}(x)$ ή $p(x)\bar{p}(x)$ με αντίστοιχο άπειρο ιδιοδιάνυσμα $y = \frac{1}{\sqrt{n}} \begin{pmatrix} 1 & \mathrm{e}^{\mathrm{i}x} & \mathrm{e}^{\mathrm{i}2x} & \cdots \end{pmatrix}^T$ [71, 84, 19]. Επομένως, για κάποιο αρκετά μεγάλο n, το y είναι κοντά στο διάνυσμα $\frac{1}{\sqrt{n}} \begin{pmatrix} 1 & \mathrm{e}^{\mathrm{i}x} & \mathrm{e}^{\mathrm{i}2x} & \cdots & \mathrm{e}^{\mathrm{i}(n-1)x} \end{pmatrix}^T$.

Χωρίζουμε το y ως $y = \left[y_1^T | y_2^T | y_3^T\right]^T$, όπου $y_1, y_3 \in \mathbb{R}^{\frac{d-1}{2}}$ και $y_2 \in \mathbb{R}^{n-(d-1)}$. Είναι προφανές ότι $\|y_1\| = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ και $\|y_3\| = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$. Εν συνεχεία, εφαρμόζουμε τη σχέση (2.3) για το συγκεκριμένο y. Τότε, προφανώς $y^H R_1 y = \mathcal{O}\left(\frac{1}{n}\right)$, $y^H R_2 y = \mathcal{O}\left(\frac{1}{n}\right)$ και $y^H R_3 y = \mathcal{O}\left(\frac{1}{n}\right)$. Αυτό σημαίνει ότι κάθε ιδιοτιμή του $T_n(p)^{-1}T_n(f)$ ανήκει σε μια $\varepsilon-$επέκταση του ορθογωνίου που προαναφέραμε, με $\varepsilon = \mathcal{O}\left(\frac{1}{n}\right)$. □

Το Θεώρημα 2.1.1 εγγυάται την ταχεία σύγκλιση της μεθόδου PCGN, αφού μας δίνει τη συσσώρευση των ιδιαζουσών τιμών, ενώ το Θεώρημα 2.1.2 εγγυάται την ταχεία σύγκλιση της PGMRES, αφού μέσω του εν λόγω προρρυθμιστή επιτυγχάνεται η συσσώρευση των ιδιοτιμών.

2.2 Κατασκευή του προρρυθμιστή

Θα παρουσιάσουμε την κατασκευή δύο διαφορετικών μεταξύ τους προρρυθμιστών. Ο πρώτος θα αντιστοιχεί σε συστήματα με καλή κατάσταση (well-conditioned), όπου η γεννήτρια συνάρτηση του πίνακα Toeplitz δεν έχει ρίζες, ενώ ο δεύτερος θα αφορά σε συστήματα, των οποίων η γεννήτρια συνάρτηση έχει ρίζες και χαρακτηρίζονται ως συστήματα με κακή κατάσταση (ill-conditioned). Προφανώς, και στις δύο περιπτώσεις η γεννήτρια συνάρτηση του προρρυθμιστή θα είναι ένα τριγωνομετρικό πολυώνυμο $p = p_1 + \mathrm{i}p_2$, με το p_1 να είναι άρτιο πολυώνυμο βαθμού d_1 και το p_2 περιττό, βαθμού d_2. Η κύρια διαφορά εντοπίζεται στο γεγονός ότι σε συστήματα με καλή κατάσταση, τα p_1 και p_2 προκύπτουν αποκλειστικά από την προσέγγιση της f, ενώ σε συστήματα με κακή κατάσταση είναι γινόμενα κατάλληλων τριγωνομετρικών πολυωνύμων, για τα οποία θα δώσουμε περισσότερες λεπτομέρειες στην υποενότητα 2.2.2.

2.2.1 Συστήματα με καλή κατάσταση

Γενικά, η καλή κατάσταση ενός μη-συμμετρικού και πραγματικού συστήματος Toeplitz χαρακτηρίζεται από το γεγονός ότι η γεννήτρια συνάρτηση f δεν έχει ρίζες. Ωστόσο, για τον προρρυθμιστή που κατασκευάζουμε σε αυτό το κεφάλαιο, θα θέλαμε να ικανοποιείται ένας ισχυρότερος περιορισμός. Πιο συγκεκριμένα, θα θέλαμε η f_1 να είναι θετική συνάρτηση και η $|f|$ φραγμένη, για να εξασφαλίσουμε ότι το ουσιώδες εύρος των ιδιοτιμών του προρρυθμισμένου συστήματος, θα είναι ένα συμπαγές υποσύνολο στο δεξιό ημιεπίπεδο, του μιγαδικού επιπέδου.

Ο ταινιωτός προρρυθμιστής, για συστήματα με καλή κατάσταση, θα κατασκευαστεί θεωρώντας κάποιου είδους προσέγγιση των συναρτήσεων που απαρτίζουν την f, δηλαδή των f_1 και f_2. Προτείνουμε δύο τύπους προσέγγισης:

1. Τη βέλτιστη ομοιόμορφη προσέγγιση της f_1 από ένα άρτιο τριγωνομετρικό πολυώνυμο και της f_2 από ένα περιττό, στο διάστημα $[-\pi, \pi]$.
2. Την τριγωνομετρική παρεμβολή των f_1 και f_2 στο ίδιο διάστημα.

Για τη βέλτιστη ομοιόμορφη προσέγγιση εφαρμόζουμε τον αλγόριθμο εναλλαγής σημείων του Remez, παίρνοντας τους κόμβους σε ένα πλέγμα (grid) με k σημεία Chebyshev πρώτου είδους, απεικονισμένα στο διάστημα $[0, \pi]$:

$$x_j = \frac{\pi}{2}\left(\cos\left(\frac{2(k-j)-1}{2k}\pi\right)+1\right),\ j = 1, 2, \dots, k,$$

όπου $k >> d_1, d_2$. Ο λόγος που επιλέγουμε τους κόμβους προσέγγισης στο υποδιάστημα $[0, \pi]$, είναι ότι η f_1 είναι άρτια και η f_2 περιττή, στο $[-\pi, \pi]$.

Αναφέρουμε ότι ο αλγόριθμος εναλλαγής σημείων του Remez δεν επιβαρύνει το κόστος των μεθόδων PGMRES και PCGN, κατά τάξη μεγέθους, αφού εξαρτάται από τη μεταβλητή k, η οποία είναι σταθερή και ανεξάρτητη της διάστασης n.

Το παρακάτω θεώρημα εγγυάται τη συσσώρευση των ιδιαζουσών τιμών του προρρυθμισμένου πίνακα, μετά από ομοιόμορφη προσέγγιση των f_1 και f_2 με χρήση του αλγορίθμου Remez.

Θεώρημα 2.2.1. ‘Εστω $f = f_1 + \mathrm{i} f_2$, η γεννήτρια συνάρτηση του πίνακα $T_n(f)$, όπου f_1 και f_2 συνεχείς, $2\pi-$περιοδικές και φραγμένες συναρτήσεις. ‘Εστω επιπλέον ότι η f_1 είναι θετική και άρτια συνάρτηση, ενώ η f_2 περιττή και $p = p_1 + \mathrm{i} p_2$, όπου p_1 είναι το άρτιο τριγωνομετρικό πολυώνυμο βέλτιστης ομοιόμορφης προσέγγισης της f_1, με μέγιστο σφάλμα ϵ_1 και p_2 είναι το περιττό τριγωνομετρικό πολυώνυμο βέλτιστης ομοιόμορφης προσέγγισης της f_2, με μέγιστο σφάλμα ϵ_2. Τότε, για κάποιο δεδομένο $\epsilon > 0$, υπάρχουν p_1 και p_2, καταλλήλων βαθμών, τέτοια ώστε οι ιδιάζουσες τιμές του προρρυθμισμένου πίνακα $T_n^{-1}(p)T_n(f)$ να έχουν γενική συσσώρευση στο διάστημα $[1 - M\epsilon, 1 + M\epsilon]$, όπου $M = \max\limits_{-\pi \le x \le \pi} \left(\frac{1}{p_1^2(x) + p_2^2(x)}\right)^{\frac{1}{2}}$.

Απόδειξη. Από το Θεώρημα 2.1.1 έχουμε ότι οι ιδιάζουσες τιμές του προρρυθμισμένου πίνακα $T_n^{-1}(p)T_n(f)$ έχουν γενική συσσώρευση στο εύρος της $\left|\frac{f}{p}\right|$. Η κατασκευή του p, μέσω βέλτιστης ομοιόμορφης προσέγγισης μας δίνει:

$$f_1(x) = p_1(x) + e_1(x) \text{ και } f_2(x) = p_2(x) + e_2(x),$$

όπου $\|e_1\|_\infty = \epsilon_1$ και $\|e_2\|_\infty = \epsilon_2$, είναι τα αντίστοιχα σφάλματα της βέλτιστης ομοιόμορφης προσέγγισης. Επομένως, $\forall x \in [-\pi, \pi]$ ισχύει ότι:

$$\frac{f}{p} = \frac{f_1 + \mathrm{i} f_2}{p_1 + \mathrm{i} p_2} = \frac{p_1 + e_1 + \mathrm{i}(p_2 + e_2)}{p_1 + \mathrm{i} p_2}.$$

Για κάθε $x \in [-\pi, \pi]$, ορίζουμε το διάνυσμα $e = (e_1 \; e_2)^T$, καθώς επίσης και το διάνυσμα των μεγίστων τιμών $\widehat{\epsilon} = (\epsilon_1 \; \epsilon_2)^T$. Για να μελετήσουμε τη συμπεριφορά της $\left|\frac{f}{p}\right|$ θεωρούμε τη συνάρτηση $\left|\frac{f}{p}\right|^2$:

$$\left|\frac{f}{p}\right|^2 = \frac{(p_1 + e_1)^2 + (p_2 + e_2)^2}{p_1^2 + p_2^2} = 1 + 2\frac{p_1 e_1 + p_2 e_2}{p_1^2 + p_2^2} + \frac{e_1^2 + e_2^2}{p_1^2 + p_2^2}.$$

Για να λάβουμε τα κάτω και άνω φράγματα χρησιμοποιούμε την τριγωνική ανισότητα, καθώς και την ανισότητα Cauchy-Schwarz, με τον τρόπο που φαίνεται παρακάτω:

$$\begin{aligned}
\left|\frac{f}{p}\right|^2 &\geq 1 - 2\left|\frac{p_1 e_1 + p_2 e_2}{p_1^2 + p_2^2}\right| + \frac{e_1^2 + e_2^2}{p_1^2 + p_2^2} \\
&\geq 1 - 2\frac{|p_1||e_1| + |p_2||e_2|}{p_1^2 + p_2^2} + \frac{e_1^2 + e_2^2}{p_1^2 + p_2^2} \\
&\geq 1 - 2\frac{\sqrt{p_1^2 + p_2^2}}{p_1^2 + p_2^2}\sqrt{e_1^2 + e_2^2} + \frac{e_1^2 + e_2^2}{p_1^2 + p_2^2} \\
&= \left(1 - \frac{1}{\sqrt{p_1^2 + p_2^2}}\|e\|_2\right)^2.
\end{aligned} \tag{2.4}$$

Επομένως,

$$\left|\frac{f}{p}\right| \geq 1 - \frac{1}{\sqrt{p_1^2 + p_2^2}}\|e\|_2 \geq 1 - \max_{-\pi \leq x \leq \pi} \frac{1}{\sqrt{p_1^2 + p_2^2}}\|\widehat{\epsilon}\|_2 = 1 - M\|\widehat{\epsilon}\|_2.$$

Σχετικά με το άνω φράγμα, έχουμε ότι ισχύει η αντίστροφη ανισότητα της (2.4), βάζοντας $+$ αντί του $-$ ως πρόσημο στον δεύτερο όρο. Έτσι έχουμε ότι:

$$\left|\frac{f}{p}\right|^2 \leq \left(1 + \frac{1}{\sqrt{p_1^2 + p_2^2}} \|e\|_2\right)^2,$$

και ακολούθως,

$$\left|\frac{f}{p}\right| \leq 1 + \frac{1}{\sqrt{p_1^2 + p_2^2}} \|e\|_2 \leq 1 + \max_{-\pi \leq x \leq \pi} \frac{1}{\sqrt{p_1^2 + p_2^2}} \|\widehat{\epsilon}\|_2 = 1 + M\|\widehat{\epsilon}\|_2.$$

Σημειώνουμε ότι η p_1 μπορεί να είναι θετική για μια κατάλληλη επιλογή του βαθμού d_1, επειδή η f_1 είναι θετική. Αυτό σημαίνει ότι:

$$M = \max_{-\pi \leq x \leq \pi} \left(\frac{1}{p_1^2(x) + p_2^2(x)}\right)^{\frac{1}{2}} < \infty.$$

Τα σφάλματα της βέλτιστης ομοιόμορφης προσέγγισης, ϵ_1 και ϵ_2, μικραίνουν όσο οι βαθμοί των πολυωνύμων αυξάνονται. Μπορούμε να επιλέξουμε τους βαθμούς d_1 και d_2, έτσι ώστε $\|\widehat{\epsilon}\|_2 \leq \epsilon$. Αυτό σημαίνει ότι,

$$1 - M\epsilon \leq \left|\frac{f}{p}\right| \leq 1 + M\epsilon, \tag{2.5}$$

το οποίο μας δίνει τη συσσώρευση γύρω από το 1. □

Σχολιάζουμε ότι μέσω της ανισότητας (2.5), έχουμε ότι η γραφική παράσταση της $\left|\frac{f}{p}\right|$ ανήκει στον δίσκο με κέντρο το σημείο $(1,0)$ και ακτίνα $M\epsilon$. Το Θεώρημα 2.1.2 περιγράφει με ακριβή τρόπο (ορθογώνιο) τη θέση των ιδιοτιμών. Με ανάλογο χειρισμό της παραπάνω απόδειξης λαμβάνουμε ότι αυτό το ορθογώνιο βρίσκεται εντός του δίσκου με κέντρο το $(1,0)$ και ακτίνα $M\epsilon$.

Στο Θεώρημα 2.2.1 υποθέσαμε ότι η f είναι συνεχής. Σε μια πιο γενική περίπτωση όπου η f είναι συνεχής στο $(-\pi, \pi)$, αλλά παρουσιάζει ασυνέχεια στο σημείο π (ειδικότερα στο $\pm\pi$), το οποίο σημαίνει ότι $f_2(\pi) \neq 0$, αλλάζουμε ελαφρώς τη διαδικασία προσέγγισης της f_2: Δε χρησιμοποιούμε, για την προσέγγιση, το διάστημα $[0, \pi]$, αλλά ένα υποδιάστημα $[0, c]$. Αφορμή γι αυτήν την αλλαγή στάθηκε το γεγονός ότι οποιοδήποτε περιττό τριγωνομετρικό πολυώνυμο έχει ρίζα στο π, ενώ η f_2 όχι, το οποίο αυξάνει τη τιμή του μέγιστου σφάλματος προσέγγισης κατά πολύ.

Χρησιμοποιώντας το διάστημα $[0, c]$, $c < \pi$, επιτυγχάνουμε ένα μικρό σφάλμα προσέγγισης στο $[0, c]$, κι ένα μεγαλύτερο στην περιοχή του π. Όπως θα φανεί και από τα αριθμητικά παραδείγματα στην επόμενη ενότητα, αυτό δεν επηρεάζει τη συσσώρευση των ιδιαζουσών τιμών και ιδιοτιμών. Η επιλογή του c γίνεται εμπειρικά. Αναφέρουμε ότι στα αριθμητικά παραδείγματα επιλέξαμε $c = \frac{5\pi}{7}$.

Όσον αφορά στην παρεμβολή της f_1 από κάποιο άρτιο τριγωνομετρικό πολυώνυμο βαθμού d_1, χρησιμοποιούμε τα αντίστοιχα d_1+1 σημεία Chebyshev στο $[0, \pi]$. Για την παρεμβολή της f_2 από κάποιο περιττό τριγωνομετρικό πολυώνυμο βαθμού d_2, χρησιμοποιούμε το σημείο 0 και d_2 σημεία Chebyshev στο $[0, c]$.

Θεώρημα 2.2.2. Έστω $f = f_1 + \mathrm{i} f_2$, όπου $f_1 > 0$ είναι συνεχής, $2\pi-$περιοδική συνάρτηση και $f_2 \in C((-\pi, \pi))$, $2\pi-$περιοδική με $f_2(\pi) \neq 0$. Έστω επίσης $p = p_1 + \mathrm{i} p_2$, όπου p_1 είναι το πολυώνυμο βέλτιστης ομοιόμορφης τριγωνομετρικής προσέγγισης της f_1 στο $[-\pi, \pi]$, με μέγιστο σφάλμα προσέγγισης ϵ_1, και p_2 είναι το τριγωνομετρικό πολυώνυμο βέλτιστης ομοιόμορφης προσέγγισης της f_2 στο $[-c, c] \subset (-\pi, \pi)$, με μέγιστο σφάλμα προσέγγισης ϵ_2, ενώ $\epsilon_2' = \max\limits_{-\pi \leq x \leq \pi} |f_2(x) - p_2(x)|$. Τότε, για κάποιο δεδομένο $\epsilon > 0$, υπάρχουν πολυώνυμα p_1, p_2, καταλλήλων βαθμών, τέτοια ώστε για αρκετά μεγάλη διάσταση n, L ιδιάζουσες τιμές, όπου $L \geq \frac{c}{\pi} n$, να ανήκουν στο διάστημα $I_\epsilon = [1 - M\epsilon, 1 + M\epsilon]$, όπου $M = \max\limits_{-\pi \leq x \leq \pi} \left(\frac{1}{p_1^2(x) + p_2^2(x)} \right)^{\frac{1}{2}}$. Οι υπόλοιπες ιδιάζουσες τιμές κυμαίνονται εκτός του I_ϵ και παρουσιάζουν γενική συσσώρευση στο $I_\epsilon' = [1 - M\epsilon, 1 + M\epsilon']$, όπου $\epsilon' = \sqrt{\epsilon_1^2 + {\epsilon_2'}^2}$.

Απόδειξη. Εύκολα παρατηρούμε ότι όλα τα βήματα της απόδειξης του Θεωρήματος 2.2.1 ισχύουν στο υποδιάστημα $[-c, c]$, αφού ο αλγόριθμος Remez συμπεριφέρεται καλά (παρουσιάζει μικρό σφάλμα) σε αυτό. Έτσι, λαμβάνουμε την ανάλογη συσσώρευση για την $\left|\frac{f}{p}\right|$:

$$\left|\frac{f}{p}\right| \in [1 - M\epsilon, 1 + M\epsilon], \ x \in [-c, c].$$

Έστω $D = \left\{ x \in [-\pi, \pi] : \left|\frac{f(x)}{p(x)}\right| \in [1 - M\epsilon, 1 + M\epsilon] \right\}$. Προφανώς, $[-c, c] \subset D \subset [-\pi, \pi]$. Εφαρμόζουμε το τύπου-Szegö θεώρημα ισοκατανομής των ιδιαζουσών τιμών [86, 81, 33, 61, 1]. Θεωρούμε τη συνεχή συνάρτηση $0 \leq F_h \leq 1$ ως εξής:

$$\begin{aligned} &F_h(z) = 1, \ z \leq 1 - M\epsilon - h, \ z \geq 1 + M\epsilon + h. \\ &F_h(z) = 0, \ z \in [1 - M\epsilon, 1 + M\epsilon], \end{aligned}$$

όπου h είναι κάποιος μικρός θετικός αριθμός. Τότε,

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n}\#\{\sigma_j \leq 1 - M\epsilon - h \ \vee \ \sigma_j \geq 1 + M\epsilon + h\} \\
&\leq \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} F_h(\sigma_j) = \frac{1}{2\pi}\int_{-\pi}^{\pi} F_h\left(\left|\frac{f(x)}{p(x)}\right|\right) \mathrm{d}x \\
&= \frac{1}{2\pi}\int_{-\pi}^{-c} F_h\left(\left|\frac{f(x)}{p(x)}\right|\right) \mathrm{d}x + \frac{1}{2\pi}\int_{c}^{\pi} F_h\left(\left|\frac{f(x)}{p(x)}\right|\right) \mathrm{d}x \\
&\leq \frac{1}{2\pi}\left(\int_{-\pi}^{-c} 1\mathrm{d}x + \int_{c}^{\pi} 1\mathrm{d}x\right) = \frac{2(\pi - c)}{2\pi} = \frac{\pi - c}{\pi},
\end{aligned}$$

όπου # δηλώνει τον πληθικό αριθμό του συνόλου και $\vee$ τη λογική διάζευξη (OR).

Η τευλευταία ανισότητα ισχύει διότι $[-c, c] \subset D$. Παίρνοντας το $h > 0$ να τείνει προς το 0, προκύπτει ότι:

$$\limsup_{n\to\infty} \frac{1}{n}\#\{\sigma_j \leq 1 - M\epsilon \ \vee \ \sigma_j \geq 1 + M\epsilon\} \leq \frac{\pi - c}{\pi}.$$

Επομένως,

$$\limsup_{n\to\infty} \frac{1}{n}\#\{\sigma_j \in [1 - M\epsilon, 1 + M\epsilon]\} \geq \frac{c}{\pi}.$$

Αυτό σημαίνει ότι για αρκετά μεγάλη διάσταση n, L ιδιάζουσες τιμές ανήκουν στο διάστημα $I_\epsilon = [1 - M\epsilon, 1 + M\epsilon]$, όπου $L \geq \frac{c}{\pi}n$. Θα δείξουμε ότι οι υπόλοιπες ιδιάζουσες τιμές παρουσιάζουν γενική συσσώρευση στο $I_{\epsilon'} = [1 - M\epsilon, 1 + M\epsilon']$ και το κύριο σώμα αυτών, στο $[1 + M\epsilon, 1 + M\epsilon']$, όπου $\epsilon' = \sqrt{\epsilon_1^2 + {\epsilon'_2}^2}$.

Παρατηρούμε ότι $p_2(\pi) = 0$, ενώ $f_2(\pi) \neq 0$. Λόγω των υποθέσεων συνέχειας, μπορούμε εύκολα να συμπεράνουμε ότι $|f_2| > |p_2|$ και $\mathrm{sign}(f_2) = \mathrm{sign}(p_2)$ σε μια περιοχή του π. Σχολιάζουμε ότι επιλέγουμε την τιμή του c αρκετά κοντά στο σημείο π, ώστε οι παραπάνω σχέσεις να ισχύουν. Για την ανάλυση, ορίζουμε τη συνάρτηση s ως εξής:

$$s(z) = \begin{cases} 1\,, z \geq 0 \\ -1\,, z < 0 \end{cases}$$

Για κάθε $x \in [-\pi, \pi]$ ισχύει:

$$\begin{gathered} -\epsilon_1 \leq e_1(x) \leq \epsilon_1,\ x \in [-\pi, \pi] \\ -\epsilon_2 \leq e_2(x) \leq \epsilon_2 \Leftrightarrow -\epsilon_2 \leq s(p_2(x))e_2(x) \leq \epsilon_2,\ x \in [-c, c]. \end{gathered} \tag{2.6}$$

Μέσω της παραπάνω υπόθεσης λαμβάνουμε ότι στο διάστημα $[c, \pi]$, η συνάρτηση e_2 έχει το ίδιο πρόσημο με την p_2. Το ίδιο ισχύει επίσης στο $[-\pi, -c]$. Επομένως,

$$0 \leq s(p_2(x))e_2(x) \leq \epsilon_2',\ |x| \in [c, \pi]. \tag{2.7}$$

Συνδυάζοντας τις σχέσεις (2.6) και (2.7) λαμβάνουμε ότι:

$$-\epsilon_2 \leq s(p_2(x))e_2(x) \leq \epsilon_2',\ x \in [-\pi, \pi]. \tag{2.8}$$

Παρακάτω εκτιμούμε τα σφάλματα της $\left|\frac{f}{p}\right|^2$:

$$\begin{aligned} \left|\frac{f}{p}\right|^2 &= \frac{(p_1 + e_1)^2 + (p_2 + e_2)^2}{p_1^2 + p_2^2} = \frac{(p_1 + e_1)^2 + (s(p_2)p_2 + s(p_2)e_2)^2}{p_1^2 + p_2^2} \\ &= \frac{(p_1 + e_1)^2 + (|p_2| + s(p_2)e_2)^2}{p_1^2 + p_2^2}. \end{aligned}$$

Όπως και στην απόδειξη του Θεωρήματος 2.2.1 ορίζουμε τα διανύσματα $e = (e_1\ e_2)^T$, $\widehat{\epsilon} = (\epsilon_1\ \epsilon_2)^T$ και $\widehat{\epsilon}' = (\epsilon_1\ \epsilon_2')^T$. Από τις (2.6),(2.8) και το γεγονός ότι $p_1 > 0$, λαμβάνουμε ότι:

$$\frac{(p_1 - \epsilon_1)^2 + (|p_2| - \epsilon_2)^2}{p_1^2 + p_2^2} \leq \left|\frac{f}{p}\right|^2 \leq \frac{(p_1 + \epsilon_1)^2 + (|p_2| + \epsilon_2')^2}{p_1^2 + p_2^2}.$$

Το κάτω φράγμα μας δίνει:

$$\begin{aligned} \frac{(p_1 - \epsilon_1)^2 + (|p_2| - \epsilon_2)^2}{p_1^2 + p_2^2} &= 1 - 2\frac{p_1\epsilon_1 + |p_2|\epsilon_2}{p_1^2 + p_2^2} + \frac{\epsilon_1^2 + \epsilon_2^2}{p_1^2 + p_2^2} \\ &\geq 1 - 2\frac{\sqrt{p_1^2 + p_2^2}}{p_1^2 + p_2^2}\|\widehat{\epsilon}\|_2 + \frac{\|\widehat{\epsilon}\|_2^2}{p_1^2 + p_2^2} \geq (1 - M\|\widehat{\epsilon}\|_2)^2, \end{aligned}$$

ενώ το άνω φράγμα:

$$\frac{(p_1 + \epsilon_1)^2 + (|p_2| + \epsilon_2')^2}{p_1^2 + p_2^2} \leq (1 + M\|\widehat{\epsilon}'\|_2)^2.$$

Επομένως,

$$1 - M\epsilon \leq 1 - M\|\widehat{\epsilon}\|_2 \leq \left|\frac{f}{p}\right| \leq 1 + M\|\widehat{\epsilon}'\|_2 = 1 + M\epsilon',$$

και η απόδειξη ολοκληρώθηκε. □

Στο σημείο αυτό, θα πρέπει να σχολιάσουμε ότι η ουσιαστική διαφορά των δύο παραπάνω διαστημάτων είναι ότι το ϵ τείνει προς το 0, όσο οι βαθμοί των πολυωνύμων αυξάνονται, ενώ το ϵ' τείνει προς μια σταθερά μεγαλύτερη του 0. Πρακτικά, το ϵ μεγαλώνει όσο το c επιλέγεται πιο κοντά στο σημείο π. Αυτό σημαίνει ότι λαμβάνουμε λιγότερες ιδιάζουσες τιμές εκτός του διαστήματος $[1 - M\epsilon, 1 + M\epsilon]$, αλλά αυτό γίνεται μεγαλύτερο. Αυτός είναι ο λόγος που επιλέγουμε τη σταθερά c εμπειρικά.

2.2.2 Συστήματα με κακή κατάσταση

Προχωρούμε με τον τρόπο κατασκευής του προρρυθμιστή για συστήματα με κακή κατάσταση, όπου σε αντίθεση με την περίπτωση που εξετάσαμε παραπάνω, η συνάρτηση f_1 έχει ρίζες. Εφόσον η f_2, ως περιττή συνάρτηση, έχει πάντα ρίζα στο 0, αρχικά θα εξετάσουμε την περίπτωση όπου η f_1 έχει επίσης ρίζα στο 0.

Η f έχει μοναδική ρίζα στο 0

Παρατηρούμε ότι η f_1 δεν αλλάζει πρόσημο στο $[-\pi, \pi]$, αφού είναι άρτια συνάρτηση. Συμβολίζουμε με m_1 και m_2 την πολλαπλότητα της ρίζας για την f_1 και f_2, αντίστοιχα. Θα μελετήσουμε την περίπτωση όπου m_1 είναι άρτιος ακέραιος, ενώ m_2 περιττός. Αρχικά υποθέτουμε ότι $m_1 < m_2$, το οποίο σημαίνει ότι η πολλαπλότητα της ρίζας, της συνάρτησης f, είναι m_1. Τότε, ακολουθώντας την ευρέως γνωστή τεχνική που προτάθηκε από τον R. Chan στην [10], κάνουμε άρση της κακής κατάστασης, διαιρώντας με το τριγωνομετρικό πολυώνυμο:

$$g(x) = (2 - 2\cos(x))^{\frac{m_1}{2}} .$$

Επομένως, λαμβάνουμε τη συνάρτηση:

$$\widehat{f} = \frac{f}{g} = \frac{f_1 + \mathrm{i} f_2}{g} = \frac{f_1}{g} + \mathrm{i}\frac{f_2}{g} = \widehat{f}_1 + \mathrm{i}\widehat{f}_2.$$

Με αυτόν τον τρόπο, έχουμε ότι $\widehat{f_1} > 0$ κι έτσι ερχόμαστε σε αντιστοιχία με την καλή κατάσταση. Αυτό σημαίνει ότι μπορούμε να προσεγγίσουμε την $\widehat{f}$, αντί της f, με τον τρόπο που προαναφέραμε. Ας είναι $q = q_1 + \mathrm{i}q_2$ το πολυώνυμο με το οποίο προσεγγίζουμε την $\widehat{f}$. Διαιρώντας με αυτό, λαμβάνουμε:

$$\frac{\widehat{f}}{q} = \frac{\frac{f}{g}}{q} = \frac{\frac{f_1+\mathrm{i}f_2}{g}}{q_1+\mathrm{i}q_2} = \frac{f_1+\mathrm{i}f_2}{gq_1+\mathrm{i}gq_2} = \frac{f}{p}.$$

Στη συνέχεια κατασκευάζουμε τον ταινιωτό πίνακα Toeplitz $T_n(p)$, τον οποίο χρησιμοποιούμε και ως προρρυθμιστή, όπου $p = gq_1 + \mathrm{i}gq_2$. Το εύρος της $\frac{f}{p}$ αποτελεί ένα σύνολο συσσώρευσης γύρω από τη μονάδα και από το Θεώρημα 2.1.1, οι ιδιάζουσες τιμές του $T_n^{-1}(p)T_n(f)$ συσσωρεύονται στο εύρος της $\left|\frac{f}{p}\right|$ με την έννοια της γενικής συσσώρευσης.

Θα θέλαμε να σημειώσουμε ότι αν η f είναι συνεχής συνάρτηση, τότε για τον προρρυθμισμένο πίνακα $T_n^{-1}(p)T_n(f)$ ισχύει το Θεώρημα 2.2.1. Από την άλλη, αν η f_2 παρουσιάζει ασυνέχεια στο σημείο π, ισχύει το Θεώρημα 2.2.2. Σχολιάζουμε ότι η $\widehat{f_2}$ διατηρεί τη ρίζα στο 0, αλλά αυτό δεν έχει κανένα αρνητικό αντίκτυπο αφού είναι μια περιττή συνάρτηση.

Στην περίπτωση όπου $m_2 < m_1$, η πολλαπλότητα της ρίζας της f είναι ίση με m_2, δηλαδή εξαρτάται από το φανταστικό μέρος της συνάρτησης. Αν προσπαθήσουμε να προκαλέσουμε άρση της κακής κατάστασης, εφαρμόζοντας την τεχνική που περιγράψαμε παραπάνω, που σημαίνει να διαιρέσουμε με $(2-2\cos(x))^{\frac{m_1}{2}}$, το φανταστικό μέρος του πηλίκου που προκύπτει θα τείνει προς το άπειρο στο σημείο 0. Από την άλλη, αν προσπαθήσουμε να μειώσουμε την επίδραση της ρίζας της f_2, διαιρώντας με $\mathrm{i}\,(\sin(x))^{m_2}$, τότε το φανταστικό μέρος της f μετατρέπεται στο πραγματικό μέρος του πηλίκου και με τη σειρά του, το πραγματικό μέρος της f μετατρέπεται σε φανταστικό (του πηλίκου). Το πρόβλημα σε αυτήν την περίπτωση είναι ότι το νέο πραγματικό μέρος τείνει προς το άπειρο όταν $x \to \pm\pi$.

Για να αποτρέψουμε αυτό το φαινόμενο, όπου το πηλίκο της f προς το τριγωνομετρικό πολυώνυμο απειρίζεται, διαιρούμε με έναν συνδυασμό των δύο συναρτήσεων και πιο συγκεκριμένα με το τριγωνομετρικό πολυώνυμο:

$$g = g_1 + \mathrm{i}g_2 = (2-2\cos(x))^{\frac{m_1}{2}} + \mathrm{i}\,(\sin(x))^{m_2}.$$

Επομένως, λαμβάνουμε τη συνάρτηση:

$$\widehat{f} = \frac{f_1+\mathrm{i}f_2}{g_1+\mathrm{i}g_2} = \frac{f_1g_1+f_2g_2}{g_1^2+g_2^2} + \mathrm{i}\frac{f_2g_1-f_1g_2}{g_1^2+g_2^2} = \widehat{f_1} + \mathrm{i}\widehat{f_2}.$$

Με αυτόν τον τρόπο η $\widehat{f_1}$ είναι θετική και φραγμένη, ενώ η $\widehat{f_2}$ έχει ρίζα στο 0. Γίνεται κατανοητό ότι ερχόμαστε και πάλι σε αντιστοιχία με την καλή κατάσταση. Προσεγγίζοντας την $\widehat{f}$ με τον ίδιο τρόπο, όπως παραπάνω, κατασκευάζουμε το πολυώνυμο $q = q_1 + \mathrm{i}q_2$. Διαιρώντας με αυτό έχουμε:

$$\frac{\widehat{f}}{q} = \frac{\frac{f}{g}}{q} = \frac{f}{gq} = \frac{f}{p},$$

όπου:

$$p = gq = (g_1 + \mathrm{i}g_2)(q_1 + \mathrm{i}q_2) = g_1q_1 - g_2q_2 + \mathrm{i}(g_1q_2 + g_2q_1) = p_1 + \mathrm{i}p_2.$$

'Οπως έχουμε ήδη περιγράψει, η συσσώρευση των ιδιαζουσών τιμών επιβεβαιώνει και την αποτελεσματικότητα της μεθόδου PCGN. Εύκολα μπορούμε να ελέγξουμε ότι τα Θεωρήματα 2.2.1 και 2.2.2 ισχύουν, κάτω από τις αντίστοιχες υποθέσεις. Το Θεώρημα 2.1.2 εγγυάται την αποτελεσματικότητα της μεθόδου PGMRES.

Η υπόθεση ότι η πολλαπλότητα m_1 είναι ίση με κάποιον άρτιο ακέραιο και m_2 με κάποιον περιττό είναι απαραίτητη, διότι υπάρχουν άρτιες συναρτήσεις με ρίζα περιττής πολλαπλότητας στο 0 (π.χ. $f_1 = |x|$), αλλά οι παράγωγοι αυτών να μην ορίζονται στο 0. Ουσιαστικά, θέλουμε να είναι ομαλή συνάρτηση, στην περιοχή της ρίζας. Το ίδιο ισχύει και για την f_2 (π.χ. $f_2 = |x|x$). Αν δεν ισχύει αυτή η υπόθεση, η αποτελεσματικότητα του προρρυθμιστή δεν είναι πάντα σίγουρη.

Η f έχει ρίζα σε σημείο διαφορετικό του 0

'Εστω ότι η f έχει μια ρίζα στο σημείο $x_0 \in [-\pi, \pi)$, $x_0 \neq 0$. 'Αρα, x_0 είναι μια ρίζα των f_1 και f_2 με πολλαπλότητες m_1 και m_2, αντίστοιχα. Σε αυτή την περίπτωση, δεν είναι απαραίτητο ότι η m_1 είναι κάποιος άρτιος αριθμός και η m_2 κάποιος περιττός. Εφόσον η f_1 είναι άρτια συνάρτηση και η f_2 περιττή, το $-x_0$ είναι επίσης ένα σημείο ρίζας για τις συναρτήσεις f_1 και f_2 με τις πολλαπλότητες της ρίζας στο x_0. Η συνάρτηση f_2 έχει μία επιπλέον ρίζα στο 0, με περιττή πολλαπλότητα. Αφού η f_1 έχει ρίζες στο $\pm x_0$, με πολλαπλότητα m_1, σε μικρή περιοχή του $\pm x_0$, $I_\epsilon = [-\epsilon - x_0, -x_0 + \epsilon] \cup [x_0 - \epsilon, x_0 + \epsilon]$, η συνάρτηση f_1 θα έχει τη μορφή:

$$f_1(x) = c_1(x)(x - x_0)^{m_1}(x + x_0)^{m_1} + o\left((x - x_0)^{m_1}(x + x_0)^{m_1}\right), \tag{2.9}$$

όπου $c_1(x)$ είναι φραγμένη συνάρτηση μακριά από το 0, η οποία διατηρεί πρόσημο στο I_ϵ. Στο σημείο αυτό, πρέπει να σημειώσουμε ότι η συνάρτηση f_1 θα πρέπει να είναι αρκετά ομαλή στα σημεία ριζών. Ειδικότερα, θα πρέπει να είναι ομαλή τάξεως

m_1. Για παράδειγμα, η άρτια συνάρτηση $|x-1||x+1|$ έχει ρίζες πολλαπλότητας ίσης με 1 στο σημείο ± 1, αλλά δεν είναι παραγωγίσιμη σε αυτό. Αυτή, δε μπορεί να γραφεί στη μορφή (2.9) και οι ρίζες της δεν αίρονται. Ωστόσο, όταν η f_1 είναι ομαλή τάξεως m_1, οι ρίζες αυτής αίρονται, διαιρώντας με το τριγωνομετρικό πολυώνυμο:

$$\begin{aligned}&\operatorname{sign}(c_1(x))\left(\sin\frac{x-x_0}{2}\right)^{m_1}\left(\sin\frac{x+x_0}{2}\right)^{m_1}\\ &=\operatorname{sign}(c_1(x))\frac{1}{2^{\frac{m_1}{2}}}\left(\cos\left(x_0\right)-\cos\left(x\right)\right)^{m_1}.\end{aligned}$$

Ο συντελεστής $\frac{1}{2^{\frac{m_1}{2}}}$ δεν παίζει κάποιον ρόλο κι επομένως μπορούμε να χρησιμοποιήσουμε την απλούστερη μορφή:

$$\operatorname{sign}(c_1(x))\left(\cos\left(x_0\right)-\cos\left(x\right)\right)^{m_1}. \tag{2.10}$$

Σχολιάζουμε ότι δε θα μπορούσαμε να χρησιμοποιήσουμε το τριγωνομετρικό πολυώνυμο:

$$\operatorname{sign}(c_1(x))\left(\sin\left(x-x_0\right)\right)^{m_1}\left(\sin\left(x+x_0\right)\right)^{m_1},$$

επειδή αυτό έχει δύο επιπλέον ρίζες $|x_0|-\pi$ και $-|x_0|+\pi$ στο $(-\pi,\pi)$.

Παρόμοια ανάλυση ισχύει και για τη συνάρτηση f_2, με τη μικρή διαφορά ότι αυτή έχει και μία επιπλέον ρίζα στο 0, με πολλαπλότητα m_0. Συμπεραίνουμε ότι το κατάλληλο τριγωνομετρικό πολυώνυμο για την άρση των ριζών αυτής, είναι το:

$$q_2(x)=\operatorname{sign}(c_2(x))\left(\cos\left(x_0\right)-\cos\left(x\right)\right)^{m_2}\left(\sin\left(x\right)\right)^{m_0}, \tag{2.11}$$

όπου $c_2(x)$ παίζει τον ρόλο της $c_1(x)$ στη (2.9), αλλά αυτή τη φορά στο σύνολο:

$$I_\epsilon^{'}=[-\epsilon-x_0,-x_0+\epsilon]\cup[-\epsilon,\epsilon]\cup[x_0-\epsilon,x_0+\epsilon].$$

Η ομαλότητα της f_2 απαιτείται όπως περιγράφηκε παραπάνω και για την f_1.

Στην περίπτωση που $m_1\leq m_2$, χρειάζεται να άρουμε τις ρίζες της f_1. 'Αρα το τριγωνομετρικό πολυώνυμο g, δίνεται από τη σχέση (2.10). Αν $m_1<m_2$, οι ρίζες της f_2 στο $\pm x_0$ παραμένουν, αλλά ως ρίζες με μικρότερη πολλαπλότητα. Επομένως, σε αυτή την ειδική περίπτωση, η ομαλότητα της f_2 δεν απαιτείται. Για παράδειγμα, έστω ότι $f(x)=(x-1)(x+1)+\mathrm{i}x|x-1|(x-1)|x+1|(x+1)$. Παρόλο που η f_2 δεν έχει την απαιτούμενη ομαλότητα, μπορούμε να κατασκευάσουμε τον προρρυθμιστή.

Το ίδιο ισχύει όταν η πολλαπλότητα m_2 δεν είναι ίση με κάποιον ακέραιο αριθμό. Στην περίπτωση όπου $m_2 < m_1$, θα πρέπει να χρησιμοποιήσουμε έναν συνδυασμό των (2.10) και (2.11). Επομένως, έχουμε:

$$\begin{aligned} g = g_1 + \mathrm{i}g_2 = &\operatorname{sign}(c_1(x)) \left(\cos(x_0) - \cos(x)\right)^{m_1} \\ &+ \mathrm{i}\operatorname{sign}(c_2(x)) \left(\cos(x_0) - \cos(x)\right)^{m_2} \left(\sin(x)\right)^{m_0}. \end{aligned} \tag{2.12}$$

Στη συνέχεια ακολουθούμε την ίδια τεχνική για την προσέγγιση της $\widehat{f} = \frac{f}{g}$, με τριγωνομετρικά πολυώνυμα, καταλήγοντας στο ότι ο ταινιωτός Toeplitz προρρυθμιστής θα έχει ως γεννήτρια συνάρτηση την $p = gq$ και τα θεωρητικά αποτελέσματα τα οποία παρουσιάσαμε παραπάνω ισχύουν για τον προρρυθμισμένο πίνακα $T_n^{-1}(p)T_n(f)$.

Ρίζες των f_1 και f_2 σε διαφορετικά σημεία

Υποθέτουμε ότι η f_1 έχει ρίζες στο $\pm x_1 \in [-\pi, \pi)$ τάξεως m_1 και η f_2 έχει ρίζες στο $\pm x_2 \in [-\pi, \pi)$ τάξεως m_2 και μία επιπλέον ρίζα στο 0, τάξεως m_0. Προφανώς, η συνάρτηση f δεν έχει ρίζες, αφού οι f_1 και f_2 δε μηδενίζονται στα ίδια σημεία, αλλά λαμβάνει τιμές τόσο στο θετικό, όσο και στο αρνητικό ημιεπίπεδο του μιγαδικού επιπέδου. Προκειμένου να άρουμε τις ρίζες της f_1 και να κατασκευάσουμε έναν αποτελεσματικό προρρυθμιστή, διαιρούμε με μια πιο συγκεκριμένη μορφή του τριγωνομετρικού πολυωνύμου (2.12), η οποία δίνεται παρακάτω:

$$\begin{aligned} g = g_1 + \mathrm{i}g_2 = &\operatorname{sign}(c_1(x)) \left(\cos(x_1) - \cos(x)\right)^{m_1} \\ &+ \mathrm{i}\operatorname{sign}(c_2(x)) \left(\cos(x_2) - \cos(x)\right)^{m_2} \left(\sin(x)\right)^{m_0}. \end{aligned} \tag{2.13}$$

Διαιρώντας με το τριγωνομετρικό πολυώνυμο της (2.13), αίρουμε τις ρίζες της f_1 κι έτσι το εύρος της $\widehat{f_1}$ ανήκει σε ένα φραγμένο σύνολο του θετικού ημιεπιπέδου. Σημειώνουμε ότι αν προσπαθήσουμε να άρουμε τις ρίζες της f_1, διαιρώντας με κάποιο τριγωνομετρικό πολυώνυμο, ανάλογο της (2.10), όπως:

$$\operatorname{sign}(c_1(x)) \left(\cos(x_1) - \cos(x)\right)^{m_1},$$

το φανταστικό μέρος θα τείνει προς το άπειρο στο $\pm x_1$.

Σχολιάζουμε ότι αν η f_1 έχει ρίζες στο $\pm x_1$ και η f_2 έχει ρίζα μόνο στο σημείο 0 ($x_2 = 0$), επιλέγουμε ως g το τριγωνομετρικό πολυώνυμο:

$$g = \operatorname{sign}(c_1(x)) \left(\cos(x_1) - \cos(x)\right)^{m_1} + \mathrm{i}\operatorname{sign}(c_2(x)) \sin(x)^{m_0}.$$

Ρίζες της f σε πολλά σημεία

Υποθέτουμε ότι η f_1 έχει ρίζες στα μη-μηδενικά σημεία $\pm x_1, \pm x_2, \ldots, \pm x_k$, τάξεως $m_1, m_2, \ldots, m_k$, αντίστοιχα και η f_2 έχει επίσης ρίζες στα ίδια σημεία με πολλαπλότητες $\ell_1, \ell_2, \ldots, \ell_k$ και μία επιπλέον ρίζα στο 0, με πολλαπλότητα ℓ_0.

Αν $m_i \leq \ell_i$, $\forall i = 1, 2, \ldots, k$, μπορούμε να επιλέξουμε το τριγωνομετρικό πολυώνυμο το οποίο αίρει όλες τις ρίζες της f_1. Αυτό θα είναι ένα γινόμενο τριγωνομετρικών πολυωνύμων τα οποία δίνονται από τη σχέση (2.10) ως:

$$g = \text{sign}(c_1(x)) \prod_{i=1}^{k} \left(\cos(x_i) - \cos(x)\right)^{m_i}. \tag{2.14}$$

Αν η f_1 έχει κι αυτή ρίζα στο 0, με πολλαπλότητα $m_0 \leq \ell_0$, το τριγωνομετρικό πολυώνυμο g λαμβάνει τη μορφή:

$$g = \text{sign}(c_1(x))(2 - 2\cos(x))^{\frac{m_0}{2}} \prod_{i=1}^{k} \left(\cos(x_i) - \cos(x)\right)^{m_i}.$$

Απαιτούμε την ανισότητα $m_0 \leq \ell_0$, διότι διαφορετικά το φανταστικό μέρος της $\widehat{f}$ θα έτεινε προς το άπειρο στο σημείο 0. Η περίπτωση όπου $m_0 > \ell_0$ μπορεί να καλυφθεί με έναν διαφορετικό τρόπο, ο οποίος περιγράφεται στη συνέχεια.

Ορίζουμε τα σύνολα δεικτών $Q = \{i : m_i \leq \ell_i\}$ και $R = \{i : m_i > \ell_i\}$. Προφανώς, αν $R = \varnothing$, τότε $\#Q = k$ και αυτή η περίπτωση καλύφθηκε παραπάνω. Θα περιγράψουμε την περίπτωση όπου $R \neq \varnothing$. Σε αυτή δε μπορούμε να χρησιμοποιήσουμε το τριγωνομετρικό πολυώνυμο, το οποίο δίνεται από τη σχέση (2.14), επειδή στα σημεία των ριζών $\pm x_i : i \in R$, θα προκαλέσουμε το φανταστικό μέρος να τείνει προς το άπειρο. Επομένως, θα πρέπει να χρησιμοποιήσουμε ένα συνδυασμό τριγωνομετρικών πολυωνύμων, έτσι ώστε να άρουμε τις ρίζες των f_1 και f_2.

Είναι προφανές ότι η f έχει ρίζες στα σημεία $\pm x_i$, με πολλαπλότητα $\min\{m_i, \ell_i\}$. Επομένως είναι απαραίτητο, το g να περιέχει τη συνάρτηση:

$$\widehat{g} = \prod_{i=1}^{k} \left(\cos(x_i) - \cos(x)\right)^{\min\{m_i, \ell_i\}},$$

ως όρο του γινομένου. Ο υπόλοιπος όρος θα πρέπει να άρει τις ρίζες που απομένουν. Οι ρίζες της f_1 που δεν έχουν αρθεί, είναι σε διαφορετικά σημεία, σε σχέση με αυτές της f_2, αφού έχουμε μειώσει την πολλαπλότητα των ριζών (της f) κατά $\min\{m_i, \ell_i\}$. Οι εναπομείνασες ρίζες της f_1 αντιστοιχούν σε δείκτες $i \in R$, ενώ αυτές της f_2 αντιστοιχούν σε δείκτες $i \in Q$ κι έτσι οδηγούμαστε σε μια γενίκευση της περίπτωσης ““Ρίζες των f_1 και f_2 σε διαφορετικά σημεία””, της υποενότητας 2.2.2. Επομένως, λαμβάνοντας υπόψιν τη σχέση (2.13), αυτός ο όρος θα πρέπει να είναι:

$$\begin{aligned}\widetilde{g} &= \operatorname{sign}(c_1(x)) \prod_{i\in R} (\cos(x_i) - \cos(x))^{m_i-\ell_i} \\ &+ \mathrm{i}\operatorname{sign}(c_2(x))\,(\sin(x))^{\ell_0} \prod_{i\in Q} (\cos(x_i) - \cos(x))^{\ell_i-m_i}\,.\end{aligned}$$

Το γινόμενο των $\widehat{g}$ και $\widetilde{g}$ μας δίνει τη συνάρτηση g, η οποία κατόπιν ορισμένων πράξεων γράφεται ως:

$$\begin{aligned}g &= \operatorname{sign}(c_1(x)) \prod_{i=1}^{k} (\cos(x_i) - \cos(x))^{m_i} \\ &+ \mathrm{i}\operatorname{sign}(c_2(x))\,(\sin(x))^{\ell_0} \prod_{i=1}^{k} (\cos(x_i) - \cos(x))^{\ell_i}\,.\end{aligned} \tag{2.15}$$

Παραπάνω δόθηκε το κατάλληλο τριγωνομετρικό πολυώνυμο για την περίπτωση όπου η συνάρτηση f έχει ρίζες σε πολλά σημεία. Αυτή η επιλογή του g (δηλαδή η (2.15)) καλύπτει επίσης την περίπτωση όπου η f έχει ρίζες σε πολλά σημεία, όπως παραπάνω, ενώ επίσης μπορούν να υπάρχουν σημεία όπου η f_1 μηδενίζεται, ενώ η f_2 δεν έχει ρίζες και το αντίστροφο. ‘Εστω $x_j \neq 0$ ένα τέτοιο σημείο, με την f_1 να έχει πολλαπλότητα ρίζας m_j και f_2 να μην έχει ρίζα. Υποθέτουμε ότι $\ell_j = 0$. Αναλόγως, αν x_j είναι ένα σημείο, όπου η f_2 έχει ρίζα πολλαπλότητας ℓ_j και η f_1 δεν έχει ρίζα, υποθέτουμε ότι $m_j = 0$. ‘Οπως προαναφέρθηκε, αυτή η περίπτωση καλύπτεται από τη σχέση (2.15), όπου κάποιες τιμές των m_i και ℓ_i μπορούν να είναι ίσες με 0.

Θα πρέπει να σημειώσουμε ότι αν η f_1 έχει ρίζα στο 0, με πολλαπλότητα m_0, πολλαπλασιάζουμε τον πρώτο όρο της (2.15) με $(2-2\cos(x))^{\frac{m_0}{2}}$, που αντιστοιχεί στην άρση αυτής. Ως εκ τούτου, θα επιτευχθεί η άρση των ριζών και ακολουθούμε την τεχνική προσέγγισης της συνάρτησης $\widehat{f} = \frac{f}{g}$.

2.2.3 Δι-διάστατη περίπτωση

Η προτεινόμενη τεχνική μπορεί να επεκταθεί, κάτω από κατάλληλους μετασχηματισμούς, για την επίλυση μη-συμμετρικών και μη-θετικά ορισμένων δι-διάστατων συστημάτων Toeplitz. Ωστόσο, για την ανάλυση αυτής της περίπτωσης έχουμε να αντιμετωπίσουμε κάποιες επιπλέον δυσκολίες. Η πρώτη εξ αυτών εντοπίζεται στον τρόπο προσέγγισης της συνάρτησης δύο μεταβλητών, αφού δε μπορούμε να χρησιμοποιήσουμε τη βέλτιστη ομοιόμορφη προσέγγιση. Θα μπορούσαμε να ξεπεράσουμε αυτή τη δυσκολία με χρήση παρεμβολής. Μια ακόμα δυσκολία είναι ότι η f (f_1 και/ή f_2) μπορεί να έχει καμπύλες ριζών, αντί για απλά σημεία. Αυτή η περίπτωση μπορεί να καλυφθεί με χρήση ανάλογης ανάλυσης, όπως στην εργασία [56]. Ωστόσο, στην απλούστερη περίπτωση όπου η f είναι χωριζομένων μεταβλητών, η παραπάνω ανάλυση μπορεί να γενικευτεί, αφού ο BTTB πίνακας παράγεται ως το άθροισμα τανυστικών γινομένων των μονοδιάστατων πινάκων Toeplitz. Μετά τον διαχωρισμό των μεταβλητών, μπορούμε να χρησιμοποιήσουμε τον αλγόριθμο Remez για την κάθε συνάρτηση (μίας μεταβλητής) και να προχωρήσουμε στην κατασκευή του προρρυθμιστή με τα αντίστοιχα τανυστικά γινόμενα. Η ισχύς των παραπάνω ισχυρισμών επιβεβαιώνεται στο Παράδειγμα 2.3.6.

2.3 Αριθμητικά αποτελέσματα

Σε αυτή την ενότητα παρουσιάζουμε μια πληθώρα αριθμητικών αποτελεσμάτων, προκειμένου να φανεί η αποτελεσματικότητα της προτεινόμενης τεχνικής προρρύθμισης. Τα αποτελέσματα λήφθηκαν μέσω του Matlab. Σε όλα τα παραδείγματα το διάνυσμα στήλη του δεύτερου μέλους, b, του συστήματος $T_n(f)x = b$, επιλέχθηκε έτσι ώστε η λύση του συστήματος να είναι το διάνυσμα που έχει όλες τις συνιστώσες του ίσες με μονάδα, $(1\ 1 \cdots 1)^T$. Ως αρχική υπόθεση θεωρήσαμε το μηδενικό διάνυσμα κι επιλέξαμε το κριτήριο τερματισμού: $\frac{\|r^{(k)}\|_2}{\|r^{(0)}\|_2} \leq 10^{-6}$, όπου $r^{(k)} = b - Ax^{(k)}$ είναι το διάνυσμα υπόλοιπο της k-οστής επανάληψης και $r^{(0)} = b$.

Στους πίνακες επαναλήψεων χρησιμοποιούμε τον ακόλουθο συμβολισμό: Με I_n δηλώνουμε ότι δε χρησιμοποιήθηκε καμία τεχνική προρρύθμισης, το B δηλώνει ότι ως προρρυθμιστής χρησιμοποιήθηκε ο ταινιωτός πίνακας Toeplitz $T_n(g)$, ενώ το R_{d_1,d_2} δηλώνει τον προρρυθμιστή $T_n(p)$, ο οποίος προέκυψε μετά από βέλτιστη ομοιόμοργη προσέγγιση της $\frac{f}{g}$. Σημειώνουμε ότι $p = gq$, με $q = q_1 + \mathrm{i}q_2$ και d_1 είναι ο βαθμός του q_1, ενώ d_2 ο βαθμός του q_2. Ακολουθώντας τον ίδιο τρόπο συμβολισμού, με In_{d_1,d_2} δηλώνουμε ότι ο προρρυθμιστής προέκυψε από παρεμβολή στις $\widehat{f_1}$ και $\widehat{f_2}$, με τριγωνομετρικά πολυώνυμα βαθμού d_1 και d_2, αντίστοιχα. Δίνουμε επίσης και τον αριθμό των ιδιαζουσών τιμών που κυμαίνονται εκτός του διαστήματος $[1 - M\epsilon, 1 + M\epsilon]$ (βλ. Θεώρημα 2.2.1 και 2.2.2), ως SV − out.

Παράδειγμα 2.3.1. Έστω $\mathfrak{f}_1(x) = x^2 + 1 + \mathrm{i}\mathfrak{h}_1(x)$, όπου:

$$\mathfrak{h}_1(x) = \begin{cases} -\pi - x\,, & -\pi \leq x < -\frac{\pi}{2} \\ x\,, & -\frac{\pi}{2} \leq x < \frac{\pi}{2} \\ \pi - x\,, & \frac{\pi}{2} \leq x \leq \pi \end{cases} .$$

Είναι προφανές ότι η $\mathfrak{h}_1$ είναι μια 2π-περιοδική και συνεχής συνάρτηση. Παρατηρούμε επίσης ότι η $f_1 = x^2 + 1$ είναι μια θετική συνάρτηση. Ο αριθμός των επαναλήψεων που χρειάζονται μέχρι τη σύγκλιση των μεθόδων PGMRES και PCGN, δίνεται στον Πίνακα 2.1.

n	PGMRES				PCGN			
	I_n	$R_{4,4}$	$R_{6,6}$	$R_{8,6}$	I_n	$R_{4,4}$	$R_{6,6}$	$R_{8,6}$
256	31	8	7	6	72	37	34	38
512	30	8	7	6	74	31	30	31
1024	29	8	7	6	73	30	30	29
2048	29	8	6	6	72	30	29	30

Πίνακας 2.1: Επαναλήψεις ($\mathfrak{f}_1$).

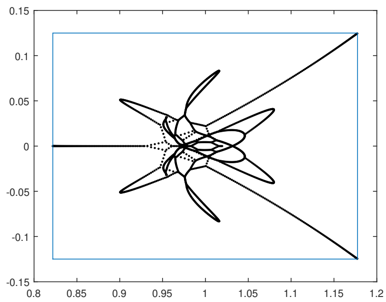
Παρατηρούμε ότι η PGMRES συγκλίνει σε πολύ λιγότερες επαναλήψεις σε σύγκριση με την PCGN. Αυτό εξηγείται από τη διαφορά στον τρόπο συσσώρευσης των ιδιοτιμών και ιδιαζουσών τιμών. 'Εχουμε γενική συσσώρευση για τις ιδιάζουσες τιμές του προρρυθμισμένου συστήματος, μερικές εκ των οποίων μπορεί να είναι κοντά στο 0, ενώ έχουμε κύρια συσσώρευση των ιδιοτιμών, εντός ενός ορθογωνίου μακριά από την αρχή των αξόνων.

Το Σχήμα 2.1 δείχνει τη συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών, όταν $n = 2048$. Οι μπλε γραμμές στο Σχήμα 2.1β' συμβολίζουν τις τιμές $1 - M\epsilon$ και $1+M\epsilon$. Σημειώνουμε ότι η f_1 έχει προσεγγιστεί από το τριγωνομετρικό πολυώνυμο p_1, βαθμού 8 και η $\mathfrak{h}_1$ από το p_2, το οποίο έχει βαθμό ίσο με 6.

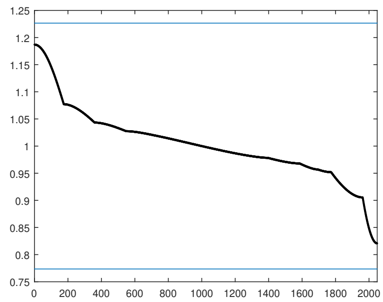
'Οπως μπορούμε να δούμε όλες οι ιδιάζουσες τιμές βρίσκονται ανάμεσα στα $1 - M\epsilon$ και $1 + M\epsilon$ και το Θεώρημα 2.2.1 ισχύει. Στο Σχήμα 2.1α' μπορούμε να παρατηρήσουμε ότι οι ιδιοτιμές του προρρυθμισμένου συστήματος κυμαίνονται εντός του ορθογωνίου $[0.822, 1.178] \times [-0.125, 0.125]$, το οποίο έχει πλευρές που αποτελούν φράγματα του πραγματικού και φανταστικού μέρους της $\frac{\mathfrak{f}_1}{p}$, επομένως το Θεώρημα 2.1.2 ισχύει.

Παράδειγμα 2.3.2. 'Εστω $\mathfrak{f}_2(x) = x^2 + \mathrm{i}x^3$. Προφανώς $f_1 = x^2$, $f_2 = x^3$ και τόσο η f_1, όσο και η f_2 έχουν ρίζα στο 0, με πολλαπλότητα $m_1 = 2$ και $m_2 = 3$, αντίστοιχα. Επομένως, θα άρουμε τη ρίζα της $\mathfrak{f}_2$, διαιρώντας με το τριγωνομετρικό πολυώνυμο $g(x) = 2 - 2\cos(x)$, όπως περιγράφηκε στην υποενότητα 2.2.2. Τότε, προσεγγίζουμε τη συνάρτηση $\frac{\mathfrak{f}_2}{g}$ με τον τρόπο που περιγράψαμε στην ενότητα 2.2 και κατασκευάζουμε τον προρρυθμιστή.

Το Σχήμα 2.2 δείχνει το πραγματικό και φανταστικό μέρος των συναρτήσεων

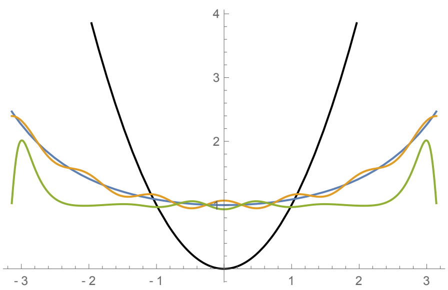


(α') Ιδιοτιμές.

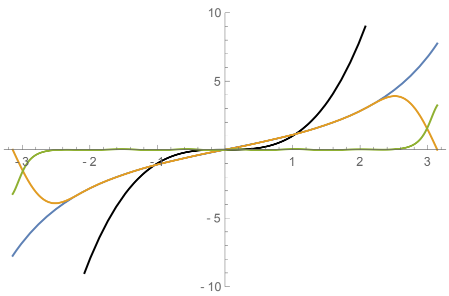


(β') Ιδιάζουσες τιμές.

Σχήμα 2.1: Ιδιοτιμές και ιδιάζουσες τιμές ($\mathfrak{f}_1$).



(α') Πραγματικό μέρος.



(β') Φανταστικό μέρος.

Σχήμα 2.2: $x^2 + \mathrm{i}x^3$.

$\mathfrak{f}_2(x) = x^2 + \mathrm{i}x^3$ (μαύρο), $\widehat{\mathfrak{f}_2} = \frac{\mathfrak{f}_2}{g}$ (μπλε), q: βέλτιστη ομοιόμορφη τριγωνομετρική προσέγγιση με πολυώνυμα 6ου βαθμού για τις $\widehat{\mathfrak{f}_{21}}$ και $\widehat{\mathfrak{f}_{22}}$ (πορτοκαλί), και $\frac{\mathfrak{f}_2}{gq}$ (πράσινο).

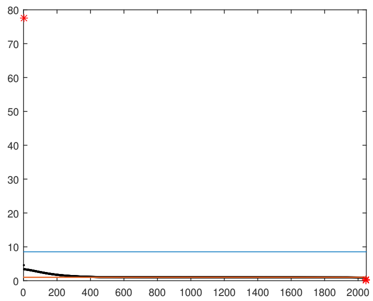
Ο αριθμός επαναλήψεων, καθώς και το πλήθος ιδιαζουσών τιμών που κυμαίνονται εκτός του $[1 - M\epsilon, 1 + M\epsilon]$, όταν χρησιμοποιούμε τον $R_{6,6}$ ως προρρυθμιστή, δίνονται στον Πίνακα 2.2. Προσεγγίζουμε την $\widehat{\mathfrak{f}_{22}}$ στο διάστημα $[-\frac{5\pi}{7}, \frac{5\pi}{7}]$ (που σημαίνει ότι επιλέγουμε $c = \frac{5\pi}{7}$).

n	PGMRES				PCGN				
	I_n	B	$R_{4,4}$	$R_{6,6}$	I_n	B	$R_{4,4}$	$R_{6,6}$	SV − out
256	256	67	24	22	-	80	37	35	58
512	>500	70	27	26	-	93	43	41	114
1024	>500	69	28	27	-	104	47	44	226
2048	>500	68	28	27	-	115	52	48	450

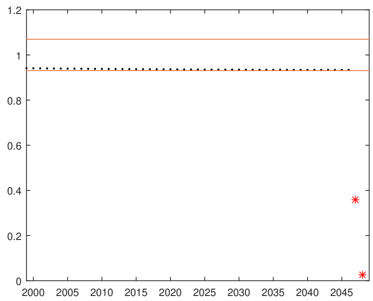
Πίνακας 2.2: Επαναλήψεις ($\mathfrak{f}_2$).

Παρατηρούμε ότι ο αριθμός ιδιαζουσών τιμών που κυμαίνονται εκτός του $I_\epsilon = [1 - M\epsilon, 1 + M\epsilon] = [0.931, 1.069]$, είναι μικρότερος από $n - \frac{c}{\pi}n$, όπως αναμένονταν από το Θεώρημα 2.2.2. Για παράδειγμα, όταν $n = 2048$, έχουμε 450 ιδιάζουσες τιμές εκτός του I_ϵ, ενώ $[n - \frac{c}{\pi}n] = 585$.

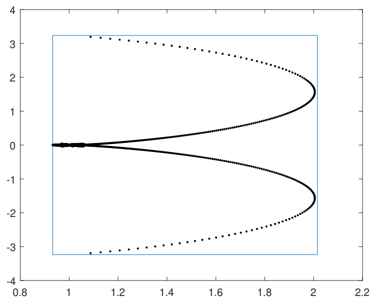
Η συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών, του προρρυθμισμένου συστήματος, όταν $n = 2048$ δίνεται στο Σχήμα 2.3: Το Σχήμα 2.3α' δείχνει τη συσσώρευση των ιδιαζουσών τιμών, το Σχήμα 2.3β' δείχνει τις τελευταίες 50 ιδιάζουσες τιμές και το Σχήμα 2.3γ' τη συσσώρευση των ιδιοτιμών, εντός του ορθογωνίου $[0.933, 2.015] \times [-3.236, 3.236]$. Στα Σχήματα 2.3α', 2.3β' οι πορτοκαλί γραμμές οριοθετούν του διάστημα I_ϵ και η μπλε δείχνει την τιμή $1 + M\epsilon'$. Τα αστέρια κόκκινου χρώματος συμβολίζουν τις ιδιάζουσες τιμές που κυμαίνονται εκτός του $I'_\epsilon = [1 - M\epsilon, 1 + M\epsilon'] = [0.931, 8.535]$ και χαρακτηρίζουν τη γενική συσσώρευση του Θεωρήματος 2.1.1. Το Σχήμα 2.3γ' επιβεβαιώνει την ισχύ του Θεωρήματος 2.1.2. Αξίζει να σημειωθεί ότι η κύρια μάζα τον ιδιοτιμών συσσωρεύεται πολύ κοντά στο σημείο $(1, 0)$.



(α') Ιδιάζουσες τιμές.



(β') Τελευταίες 50 ιδιάζουσες τιμές.



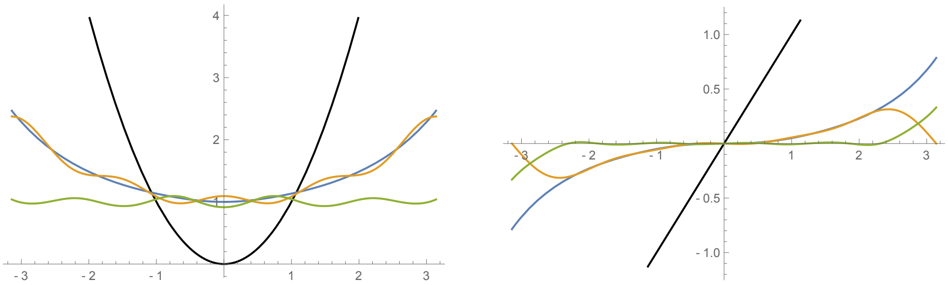
(γ') Ιδιοτιμές.

Σχήμα 2.3: Ιδιοτιμές και ιδιάζουσες τιμές ($\mathfrak{f}_2$).

Στα παραδείγματα που ακολουθούν επικεντρωνόμαστε μόνο στη συμπεριφορά της μεθόδου PGMRES, αφού είδαμε ότι αυτή συγκλίνει (στη λύση) σε πολύ λιγότερες επαναλήψεις, σε σχέση με τη μέθοδο PCGN.

Παράδειγμα 2.3.3. Έστω ότι $\mathfrak{f}_3(x) = x^2 + \mathrm{i}x$. Σε αυτό το παράδειγμα $f_1 = x^2$, $f_2 = x$ και τόσο η f_1, όσο και η f_2 έχουν ρίζα στο 0, με πολλαπλότητα $m_1 = 2$ και $m_2 = 1$, αντίστοιχα. Επομένως, θα άρουμε τις ρίζες της γεννήτριας συνάρτησης χρησιμοποιώντας το τριγωνομετρικό πολυώνυμο $g(x) = 2 - 2\cos(x) + \mathrm{i}\sin(x)$. Στη συνέχεια, προσεγγίζουμε τη συνάρτηση $\frac{\mathfrak{f}_3}{g}$ και κατασκευάζουμε τον προρρυθμιστή. Σε αυτό το παράδειγμα δίνουμε και τον αριθμό επαναλήψεων όταν χρησιμοποιούμε βέλτιστη ομοιόμορφη προσέγγιση, καθώς επίσης και παρεμβολή με τριγωνομετρικά πολυώνυμα.

Το Σχήμα 2.4 δείχνει το πραγματικό και φανταστικό μέρος των $\mathfrak{f}_3$, $\widehat{\mathfrak{f}_3} = \frac{\mathfrak{f}_3}{g}$, q που αφορά στον προρρυθμιστή $R_{4,4}$ και $\frac{\mathfrak{f}_3}{gq}$, όπως έγινε και στο προηγούμενο παράδειγ-



(α') Πραγματικό μέρος. (β') Φανταστικό μέρος.

Σχήμα 2.4: $x^2 + \mathrm{i}x$.

μα.

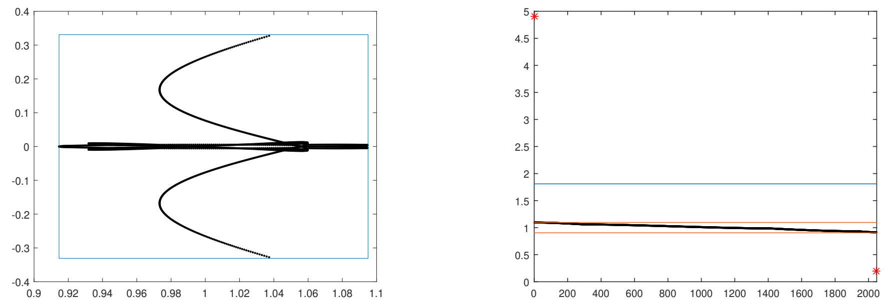
Ο αριθμός επαναλήψεων δίνεται στον Πίνακα 2.3. Σε αυτόν δίνεται επίσης και ο αριθμός των ιδιαζουσών τιμών που κυμαίνονται εκτός του διαστήματος συσσώρευσης και αφορούν στον $R_{4,4}$.

n	I_n	B	$R_{4,4}$	$In_{4,4}$	$R_{10,10}$	$In_{10,10}$	SV − out
256	256	11	6	6	6	12	2
512	>500	11	6	6	6	12	2
1024	>500	10	6	6	5	12	2
2048	>500	10	6	5	5	11	2

Πίνακας 2.3: Επαναλήψεις ($\mathfrak{f}_3$).

Παρατηρούμε ότι οι προρρυθμιστές που κατασκευάστηκαν από βέλτιστη ομοιόμορφη προσέγγιση και παρεμβολή, έχουν σχεδόν τους ίδιους αριθμούς επαναλήψεων, όταν οι βαθμοί των πολυωνύμων q_1 και q_2 είναι μικροί. Ωστόσο, όσο οι βαθμοί μεγαλώνουν, ο προρρυθμιστής που προκύπτει μέσω βέλτιστης ομοιόμορφης προσέγγισης έχει καλύτερη συμπεριφορά, δηλαδή είναι πιο αποτελεσματικός από αυτόν που προέκυψε μέσω παρεμβολής. Αυτό οφείλεται στην ταλάντωση του τριγωνομετρικού πολυωνύμου παρεμβολής, μέσω της οποίας δεν εξασφαλίζεται η μείωση του σφάλματος, όσο ο βαθμός του πολυωνύμου αυξάνεται. ʻΑρα, επιβεβαιώνεται ότι η βέλτιστη ομοιόμορφη προσέγγιση δίνει πιο αποτελεσματικούς προρρυθμιστές.

Η συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών όταν $n = 2048$, χρησιμοποιώντας τον $R_{4,4}$, δίνεται στο Σχήμα 2.5: το Σχήμα 2.5α' τη συσσώρευση των ιδιοτιμών, ενώ το Σχήμα 2.5β' τη συσσώρευση των ιδιαζουσών τιμών.



(α') Ιδιοτιμές.

(β') Ιδιάζουσες τιμές.

Σχήμα 2.5: Ιδιοτιμές και ιδιάζουσες τιμές ($\mathfrak{f}_3$).

Στο Σχήμα 2.5β', οι πορτοκαλί γραμμές είναι τα άκρα του διαστήματος $I_\epsilon = [0.904, 1.096]$ και η μπλε γραμμή λαμβάνει την τιμή $1 + M\epsilon' = 1.809$. ʻΟπως και στο προηγούμενο παράδειγμα, με κόκκινα αστέρια συμβολίζουμε τις ιδιάζουσες τιμές που κυμαίνονται εκτός του διαστήματος γενικής συσσώρευσης. Το σχήμα 2.5α' δείχνει τη συσσώρευση των ιδιοτιμών εντός του ορθογωνίου $[0.915, 1.095] \times [-0.331, 0.331]$.

Παράδειγμα 2.3.4. ʻΕστω $\mathfrak{f}_4(x) = x^2 - 1 + i\mathfrak{h}_2(x)$, όπου:

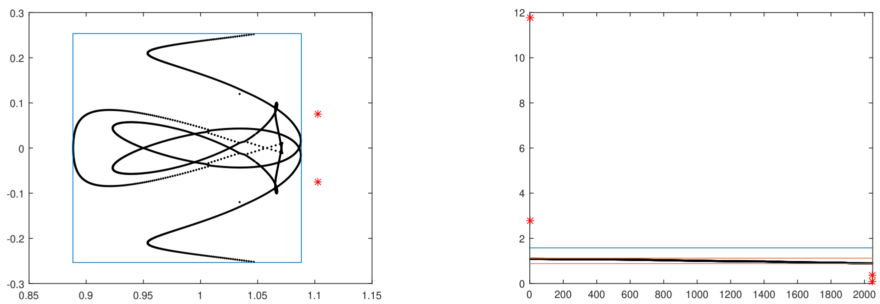
$$\mathfrak{h}_2(x) = \begin{cases} -1 - x \,, -\pi \leq x < -\frac{1}{2} \\ \quad x \quad , -\frac{1}{2} \leq x < \frac{1}{2} \\ \; 1 - x \;, \;\; \frac{1}{2} \leq x \leq \pi \end{cases} .$$

Η συνάρτηση f_1 έχει μια απλή ρίζα στο ± 1 (με πολλαπλότητα $m_1 = 1$) και η $f_2 = \mathfrak{h}_2$ έχει επίσης ρίζα στο ± 1 με πολλαπλότητα $m_2 = 1$ και μία ρίζα στο 0, με πολλαπλότητα $m_0 = 1$. Επομένως, επιλέγουμε ως $g(x) = \cos(1) - \cos(x)$ και προσεγγίζουμε την $\frac{\mathfrak{f}_4}{g}$ με τριγωνομετρικά πολυώνυμα 4ου βαθμού, έτσι ώστε στη συνέχεια να κατασκευάσουμε τον προρρυθμιστή.

Ο αριθμός επαναλήψεων και ιδιαζουσών τιμών που κυμαίνονται εκτός του διαστήματος συσσώρευσης δίνονται στον Πίνακα 2.4. Η συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών, όταν $n = 2048$, δίνεται στα Σχήματα 2.6α' και 2.6β', αντίστοιχα. Παρατηρούμε ότι δύο ιδιοτιμές έχουν πραγματικό μέρους μεγαλύτερο από $\underset{-\pi \leq x \leq \pi}{\text{ess sup}} \ \text{Re}\left(\frac{\mathfrak{f}_4(x)}{p(x)}\right)$, χαρακτηρίζοντας την κύρια συσσώρευση στο ορθογώνιο $[0.888, 1.088] \times [-0.253, 0.253]$.

n	I_n	B	$R_{4,4}$	SV − out
256	256	15	6	4
512	>500	15	6	4
1024	>500	16	6	4
2048	>500	15	6	4

Πίνακας 2.4: Επαναλήψεις ($\mathfrak{f}_4$).



(α') Ιδιοτιμές. (β') Ιδιάζουσες τιμές.

Σχήμα 2.6: Ιδιοτιμές και ιδιάζουσες τιμές ($\mathfrak{f}_4$).

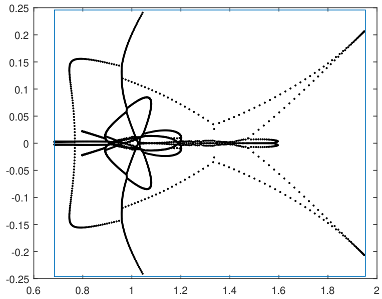
Παράδειγμα 2.3.5. 'Εστω ότι $\mathfrak{f}_5(x) = (x^2-1)^2 + \mathrm{i}x(x^2-4)$. Σε αυτό το παράδειγμα προφανώς το πραγματικό και φανταστικό μέρος της γεννήτριας συνάρτησης έχουν ρίζες σε διαφορετικά σημεία, αφού $f_1 = (x^2-1)^2$ και $f_2 = x(x^2-4)$. Ειδικότερα, η f_1 έχει ρίζα στο ± 1 με πολλαπλότητα $m_1 = 1$, ενώ η f_2 έχει ρίζα στο 0 και στο ± 2, με πολλαπλότητες $m_0 = m_2 = 1$. Επομένως, επιλέγουμε ως $g(x) = (\cos(1) - \cos(x))^2 + \mathrm{i}\sin(x)(\cos(2) - \cos(x))$, σύμφωνα με όσα περιγράψαμε στον τρόπο κατασκευής του προρρυθμιστή. Τελικά, κατασκευάζουμε τον $R_{8,6}$.

Ο αριθμός επαναλήψεων με χρήση της μεθόδου PGMRES δίνεται στον Πίνακα 2.5. Η κύρια συσσώρευση των ιδιοτιμών εντός του ορθογωνίου $[0.684, 1.953] \times [-0.246, 0.246]$ φαίνεται στο Σχήμα 2.7.

Παράδειγμα 2.3.6. 'Εστω η δι-διάστατη συνάρτηση χωριζομένων μεταβλητών $\mathfrak{f}_6(x,y) = x^2 + y^2 + \mathrm{i}(x+y)$. Προφανώς, $f_1(x,y) = x^2 + y^2$ και $f_2(x,y) = x + y$. Η f_1 έχει μια ρίζα στο $(0,0)$ με πολλαπλότητα ίση με 2, ενώ η f_2 έχει μια ρίζα στο ίδιο σημείο, αλλά με πολλαπλότητα 1. Ας είναι $\mathfrak{h}(z) = z^2 + \mathrm{i}z$, $z \in [-\pi, \pi]$, τότε η συνάρτηση $\mathfrak{f}_6$ χωρίζεται ως: $\mathfrak{f}_6(x,y) = \mathfrak{h}(x) + \mathfrak{h}(y)$. Θα άρουμε τη ρίζα της

n	I_n	B	$R_{8,6}$
256	151	25	12
512	197	25	11
1024	223	25	11
2048	229	24	11

Πίνακας 2.5: Επαναλήψεις ($\mathfrak{f}_5$).



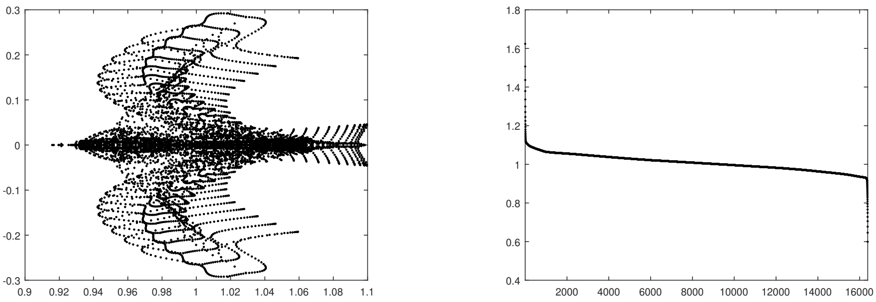
Σχήμα 2.7: Ιδιοτιμές ($\mathfrak{f}_5$).

$\mathfrak{h}(z)$, διαιρώντας με $g(z) = 2 - 2\cos(z) + \mathrm{i}\sin(z)$. Κατόπιν, προσεγγίζουμε την $\frac{h}{g}$ με το τριγωνομετρικό πολυώνυμο $q = q_1 + \mathrm{i}q_2$, όπου q_1 και q_2 είναι 4ου βαθμού. Στη συνέχεια κατασκευάζουμε τον μονοδιάστατο ταινιωτό πίνακα Toeplitz, $T_n(p) = T_n(gq)$, ο οποίος είναι ο αντίστοιχος προρρυθμιστής για τον $T_n(\mathfrak{h})$. Τελικά, ο δι-διάστατος ταινιωτός Toeplitz προρρυθμιστής κατασκευάζεται από το τανυστικό γινόμενο $T_{nm}(\widehat{p}) = I_n \otimes T_m(p) + T_n(p) \otimes I_m$, όπου $\widehat{p}(x,y) = p(x) + p(y)$.

Ο αριθμός επαναλήψεων για διάφορες διαστάσεις των blocks δίνεται στον Πίνακα 2.6. Η συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών δίνεται στο Σχήμα 2.8.

$n \backslash m$	16	32	64	128
16	6	6	6	6
32	6	6	6	6
64	6	6	6	6
128	6	6	6	6

Πίνακας 2.6: Επαναλήψεις ($\mathfrak{f}_6$).



(α') Ιδιοτιμές.

(β') Ιδιάζουσες τιμές.

Σχήμα 2.8: Ιδιοτιμές και ιδιάζουσες τιμές ($\mathfrak{f}_6$).

Το Σχήμα 2.8α' δείχνει τη συσσώρευση των ιδιοτιμών του προρρυθμισμένου συστήματος, όταν $n = m = 128$ και το Σχήμα 2.8β', την αντίστοιχη συσσώρευση των ιδιαζουσών τιμών.

Κεφάλαιο 3

Κυκλοειδείς Προρρυθμιστές

Μια ευρέως γνωστή κατηγορία προρρυθμιστών αποτελούν οι κυκλοειδείς πίνακες [20]. Ένας πίνακας καλείται κυκλοειδής όταν είναι Toeplitz και κάθε επόμενη γραμμή/στήλη αυτού προκύπτει από μια κυκλική μετατόπιση της προηγούμενης. Έτσι οι κυκλοειδείς πίνακες έχουν την παρακάτω μορφή:

$$C_n = \begin{pmatrix} c_0 & c_{-1} & c_{-2} & \cdots & c_2 & c_1 \\ c_1 & c_0 & c_{-1} & \cdots & c_3 & c_2 \\ c_2 & c_1 & c_0 & \cdots & c_4 & c_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{-2} & c_{-3} & c_{-4} & \cdots & c_0 & c_{-1} \\ c_{-1} & c_{-2} & c_{-3} & \cdots & c_1 & c_0 \end{pmatrix}.$$

Στη βιβλιογραφία κάποιος μπορεί να βρει πολλά είδη κυκλοειδών προρρυθμιστών για συστήματα Toeplitz, όπως είναι για παράδειγμα αυτός που εισήγαγε ο G. Strang το 1986, ο βέλτιστος κυκλοειδής προρρυθμιστής που πρότεινε ο T. Chan το 1988 κ.α. Οι προρρυθμιστές του G. Strang και του T. Chan, κατασκευάζονται από τις τιμές του αρχικού πίνακα Toeplitz σύμφωνα με τις παρακάτω σχέσεις, αντίστοιχα:

$$s_k = \begin{cases} t_k, & 0 \leq k \leq \lfloor n/2 \rfloor \\ t_{k-n}, & \lfloor n/2 \rfloor < k \leq n-1 \\ s_{n+k}, & 0 < -k \leq n-1 \end{cases}, \tag{3.1}$$

$$c_k = \begin{cases} \frac{(n-k)t_k + kt_{k-n}}{n}, & 0 \leq k \leq n-1 \\ c_{n+k}, & 0 < -k \leq n-1 \end{cases}.$$

Βλέπουμε ότι ο κυκλοειδής προρρυθμιστής του G. Strang κατασκευάζεται αναδιπλώνοντας τις κεντρικές διαγωνίους του πίνακα των συντελεστών T_n, σύμφωνα με τη σχέση (3.1). Σχολιάζουμε επίσης ότι ο κυκλοειδής προρρυθμιστής του T. Chan ονομάζεται βέλτιστος επειδή ελαχιστοποιεί τη νόρμα Frobenius $\|C_n - T_n\|_F$, ως προς οποιονδήποτε κυκλοειδή πίνακα C_n.

‘Ολοι οι κυκλοειδείς προρρυθμιστές έχουν μια κοινή χαρακτηριστική ιδιότητα, να διαγωνοποιούνται από τον πίνακα διακριτού μετασχηματισμού Fourier, ο οποίος δίνεται ως:

$$\mathcal{F}_n = \frac{1}{\sqrt{n}} \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{n-1} & \omega^{2(n-1)} & \cdots & \omega^{(n-1)^2} \end{pmatrix},$$

όπου $\omega = \mathrm{e}^{2\pi \mathrm{i}/n}$ και προφανώς $\omega^j, j = 0, \ldots, n-1$ είναι οι διακεκριμένες λύσεις της μιγαδικής εξίσωσης $z^n - 1 = 0$, δηλαδή οι n-οστές ρίζες της μονάδας. Αναφέρουμε ότι ο πίνακας $\mathcal{F}_n$ είναι ορθομοναδιαίος (επομένως και αντιστρέψιμος) και συμμετρικός [11, 42, 90]. Λόγω της προαναφερθείσας ιδιότητας μπορούμε να γράψουμε έναν οποιονδήποτε κυκλοειδή προρρυθμιστή C_n, ως:

$$C_n = \mathcal{F}_n^H \Lambda_n \mathcal{F}_n,$$

όπου Λ_n είναι διαγώνιος πίνακας, ο οποίος έχει ως στοιχεία της κυρίας διαγωνίου του τις ιδιοτιμές του C_n. Ως εκ τούτου, είναι προφανές ότι μπορούμε να ορίσουμε έναν κυκλοειδή προρρυθμιστή με ιδιοτιμές της αρεσκείας μας. Λόγω του ότι θέλουμε να προρρυθμίσουμε ένα σύστημα Toeplitz, το οποίο έχει ως γεννήτρια συνάρτηση την $f = f_1 + \mathrm{i} f_2$ (με τις ιδιότητες που αναφέραμε στο εισαγωγικό κεφάλαιο), είναι συνετό να κατασκευάσουμε τον κυκλοειδή προρρυθμιστή:

$$C_n(f) = \mathcal{F}_n^H \Lambda_n(f) \mathcal{F}_n,$$

ο οποίος έχει ως ιδιοτιμές τις τιμές που λαμβάνει η γεννήτρια συνάρτηση f στους κόμβους:

$$\frac{2(j-1)\pi}{n}, \; j = 1, 2, \ldots, n. \tag{3.2}$$

Προφανώς, λόγω περιοδικότητας της γεννήτριας συνάρτησης τα διαστήματα $(-\pi, \pi]$ και $[0, 2\pi)$ περιέχουν τις ίδιες τιμές, για τους αντίστοιχους κόμβους.

3.1 Κατασκευή του προρρυθμιστή

Έστω $f = f_1 + \mathrm{i}f_2$, όπου f_1 είναι άρτια, 2π-περιοδική συνάρτηση και f_2 περιττή κι επίσης 2π-περιοδική, ορισμένες στο $(-\pi, \pi]$. Είναι προφανές, όπως είδαμε και στο προηγούμενο κεφάλαιο, ότι η f_2 έχει ρίζα στο 0, ως περιττή συνάρτηση. Αυτή η ιδιότητα δεν ισχύει πάντα για την f_1, η οποία μπορεί να μην έχει και καμία ρίζα στο διάστημα $(-\pi, \pi]$. Επομένως, θα διακρίνουμε και πάλι δύο περιπτώσεις προρρύθμισης του αρχικού συστήματος, βάσει της φύσης της f, δηλαδή το αν έχει ρίζες ή όχι.

3.1.1 Συστήματα με καλή κατάσταση

Θα ξεκινήσουμε την περιγραφή κατασκευής του προρρυθμιστή για συστήματα με καλή κατάσταση, δηλαδή συστήματα των οποίων η γεννήτρια συνάρτηση f, δεν έχει ρίζες στο πεδίο ορισμού της. Σε αυτή την περίπτωση, ο προτεινόμενος προρρυθμιστής είναι ο κυκλοειδής πίνακας $\mathcal{C}_n(f)$, ο οποίος έχει ως ιδιοτιμές, τις τιμές $f\left(\frac{2\pi(k-1)}{n}\right), k = 1, \ldots, n$.

Σημειώνουμε ότι η f μπορεί να μην έχει ρίζες, ωστόσο δεν είναι απαραίτητο για την f_1 να είναι θετική. Αυτό είναι προφανές αφού οι συναρτήσεις f_1 και f_2, μπορούν να μηδενίζονται σε διαφορετικά σημεία. Σημειώνουμε επίσης ότι η χρήση του $\mathcal{C}_n(f)$ ως προρρυθμιστή είναι επιτρεπτή, διότι είναι αντιστρέψιμος πίνακας. Η κατασκευή αυτού, γίνεται με $\mathcal{O}(n \log n)$ πράξεις, μέσω του ταχύ μετασχηματισμού Fourier (FFT) [80].

3.1.2 Συστήματα με κακή κατάσταση

Συνεχίζουμε με την περίπτωση όπου η f έχει ρίζες. Σε αυτή, προτείνουμε την εύρεση ενός τριγωνομετρικού πολυωνύμου g, έτσι ώστε το πραγματικό μέρος της $\frac{f}{g}$ να είναι θετικό, όπως ακριβώς και στην τεχνική προρρύθμισης του προηγούμενου κεφαλαίου. Η μορφή αυτού δόθηκε στην εργασία [49] (βλ. ενότητα 2.2). Ο προτεινόμενος προρρυθμιστής γι αυτή την περίπτωση είναι ο πίνακας $T_n(g)\mathcal{C}_n\left(\frac{f}{g}\right)$, όπου $\mathcal{C}_n\left(\frac{f}{g}\right)$ είναι ο κυκλοειδής πίνακας, που έχει ιδιοτιμές ίσες με $\frac{f}{g}\left(\frac{2\pi(k-1)}{n}\right)$, $k = 1, \ldots, n$ και $T_n(g)$ είναι ο ταινιωτός πίνακας Toeplitz, ο οποίος έχει ως γεννήτρια συνάρτηση το τριγωνομετρικό πολυώνυμο g. Η κατασκευή του προρρυθμιστή γίνεται επίσης γρήγορα, σε $\mathcal{O}(n \log n)$ πράξεις, χρησιμοποιώντας και πάλι τον ταχύ μετασχηματισμό Fourier. Σημειώνουμε ότι ο $T_n(g)\mathcal{C}_n\left(\frac{f}{g}\right)$ μπορεί να χρησιμοποιηθεί ως προρρυθμιστής και στην περίπτωση όπου η f δεν έχει ρίζες, αλλά αντιθέτως η f_1 μηδενίζεται στο πεδίο ορισμού της.

3.2 Θεωρητικά αποτελέσματα

Σε αυτή την ενότητα δίνουμε τα θεωρητικά αποτελέσματα, που αφορούν στη συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών του προρρυθμισμένου συστήματος. Αυτά αποτελούν ικανές συνθήκες για την ταχεία σύγκλιση των μεθόδων PGMRES και PCGN, με χρήση του προτεινόμενου προρρυθμιστή. Αρχικά θα μελετήσουμε την περίπτωση όπου η γεννήτρια συνάρτηση του συστήματος είναι συνεχής και στη συνέχεια θα επεκταθούμε και στην περίπτωση όπου αυτή έχει σημεία ασυνέχειας.

Επειδή $T_n(f) = T_n(f_1) + \mathrm{i}T_n(f_2)$, είναι ευρέως γνωστό [14] ότι:

$$\|T_n(f)\|_2 \le 2\|f\|_\infty. \tag{3.3}$$

Συμβολίζουμε με $\lambda_k(\mathcal{C}_n(f)),\ k = 1, \ldots, n$, τις ιδιοτιμές του κυκλοειδή πίνακα $\mathcal{C}_n(f)$. Από την κατασκευή αυτού, γνωρίζουμε ότι ισχύει:

$$\lambda_k(\mathcal{C}_n(f)) = f\left(\frac{2\pi(k-1)}{n}\right),\ k = 1, \ldots, n.$$

Λήμμα 3.2.1. Έστω μια 2π-περιοδική και μιγαδική συνάρτηση f. Τότε:

$$\|\mathcal{C}_n(f)\|_2 \le 2\|f\|_\infty. \tag{3.4}$$

Επιπλέον, αν η f δεν έχει ρίζες στο $(-\pi, \pi]$, ισχύει ότι:

$$\|\mathcal{C}_n^{-1}(f)\|_2 \le 2\left\|\frac{1}{f}\right\|_\infty. \tag{3.5}$$

Απο404δειξη. Παρατηρούμε ότι $\mathcal{C}_n(f) = \mathcal{C}_n(f_1) + \mathrm{i}\mathcal{C}_n(f_2)$, όπου $\mathcal{C}_n(f_1)$ και $\mathcal{C}_n(f_2)$ είναι Ερμιτιανοί πίνακες. Εύκολα βλέπουμε ότι:

$$\|\mathcal{C}_n(f_1)\|_2 = \max_k |\lambda_k(\mathcal{C}_n(f_1))| = \max_k \left|f_1\left(\frac{2\pi(k-1)}{n}\right)\right| \le \|f_1\|_\infty,$$

$$\|\mathcal{C}_n(f_2)\|_2 = \max_k |\lambda_k(\mathcal{C}_n(f_2))| = \max_k \left|f_2\left(\frac{2\pi(k-1)}{n}\right)\right| \le \|f_2\|_\infty.$$

Έτσι, λαμβάνουμε την παρακάτω σχέση:

$$\|\mathcal{C}_n(f)\|_2 \le \|\mathcal{C}_n(f_1)\|_2 + \|\mathcal{C}_n(f_2)\|_2 \le \|f_1\|_\infty + \|f_2\|_\infty \le 2\|f\|_\infty.$$

Για το άνω φράγμα της $\|\mathcal{C}_n^{-1}(f)\|_2$, έχουμε:

$$\|\mathcal{C}_n^{-1}(f)\|_2 = \left\|\mathcal{C}_n\left(\frac{1}{f}\right)\right\|_2 \leq 2\left\|\frac{1}{f}\right\|_\infty.$$

Σημειώνουμε ότι αν η f δεν έχει ρίζες στο $(-\pi,\pi]$: $|f(x)| \neq 0, \forall x \in (-\pi,\pi]$. □

3.2.1 Συνεχής περίπτωση

Θεώρημα 3.2.2. Έστω f μια 2π-περιοδική και συνεχής, μιγαδική συνάρτηση. Τότε, για κάθε $\varepsilon > 0$, υπάρχει σταθερά M, τέτοια ώστε για κάθε $n > 2M$, $T_n(f) - \mathcal{C}_n(f) = S_n + L_n$, όπου $\|S_n\|_2 \leq \varepsilon$ και ο πίνακας L_n έχει βαθμίδα το πολύ ίση με $2M$.

Απόδειξη. Έστω f η συνάρτηση με τις ιδιότητες που περιγράψαμε παραπάνω. Από το θεώρημα προσέγγισης Stone-Weierstrass [79], για κάποιο δεδομένο $\epsilon > 0$, υπάρχει ένα τριγωνομετρικό πολυώνυμο

$$p_M(x) = \sum_{k=-M}^{M} \rho_k \mathrm{e}^{\mathrm{i}kx},$$

τέτοιο ώστε:

$$\|f - p_M\|_\infty \leq \epsilon. \tag{3.6}$$

Για κάθε $n > 2M$:

$$T_n(f) - \mathcal{C}_n(f) = T_n(f - p_M) - \mathcal{C}_n(f - p_M) + T_n(p_M) - \mathcal{C}_n(p_M).$$

Εύκολα παρατηρούμε ότι οι πίνακες $T_n(p_M)$ και $\mathcal{C}_n(p_M)$ διαφέρουν μόνο κατά έναν πίνακα χαμηλής βαθμίδας L_n, το πολύ ίσης με $2M$ [64]. Χρησιμοποιώντας τις (3.3), (3.4) και (3.6), καταλήγουμε στο ότι για τους δύο πρώτους όρους του δεξιού μέλους, της παραπάνω εξίσωσης ισχύει:

$$\begin{aligned}\|T_n(f - p_M) - \mathcal{C}_n(f - p_M)\|_2 &\leq \|T_n(f - p_M)\|_2 + \|\mathcal{C}_n(f - p_M)\|_2 \\ &\leq 2\|f - p_M\|_\infty + 2\|f - p_M\|_\infty \leq 4\epsilon.\end{aligned}$$

Επομένως, $S_n = T_n(f - p_M) - \mathcal{C}_n(f - p_M)$ είναι ένας πίνακας με μικρή νόρμα κι επιλέγοντας $\epsilon = \frac{\varepsilon}{4}$ λαμβάνουμε το ζητούμενο αποτέλεσμα. □

Όπως αποδείχθηκε στο Λήμμα 3.2.1, όταν η γεννήτρια συνάρτηση f δεν έχει ρίζες στο $(-\pi, \pi]$, ισχύει η σχέση (3.5). Συνδυάζοντάς την με το Θεώρημα 3.2.2 και το γεγονός ότι

$$\mathcal{C}_n^{-1}(f)T_n(f) - I_n = \mathcal{C}_n^{-1}(f)(T_n(f) - \mathcal{C}_n(f)) = \mathcal{C}_n^{-1}(f)S_n + \mathcal{C}_n^{-1}(f)L_n,$$

μπορούμε να δώσουμε το παρακάτω πόρισμα.

Πόρισμα 3.2.3. Έστω f μια 2π-περιοδική και συνεχής, μιγαδική συνάρτηση, η οποία δεν έχει ρίζες στο $(-\pi, \pi]$. Τότε, για κάθε $\varepsilon > 0$, υπάρχει σταθερά M, τέτοια ώστε για κάθε $n > 2M$, $\mathcal{C}_n^{-1}(f)T_n(f) - I_n = \widehat{S}_n + \widehat{L}_n$, όπου $\|\widehat{S}_n\|_2 \leq \varepsilon$ και ο πίνακας $\widehat{L}_n$ έχει βαθμίδα το πολύ ίση με $2M$.

Θεώρημα 3.2.4. Έστω f μια 2π-περιοδική, συνεχής και μιγαδική συνάρτηση, η οποία δεν έχει ρίζες στο $(-\pi, \pi]$. Τότε, για κάθε $\varepsilon > 0$, το διάστημα $[1-\varepsilon, 1+\varepsilon]$ αποτελεί ένα σύνολο κύριας συσσώρευσης των ιδιαζουσών τιμών του $\mathcal{C}_n^{-1}(f)T_n(f)$.

Απο🕱δειξη. Είναι ευρέως γνωστό ότι οι ιδιάζουσες τιμές του $\mathcal{C}_n^{-1}(f)T_n(f)$ είναι οι τετραγωνικές ρίζες των ιδιοτιμών του πίνακα που σχετίζεται με τις κανονικές εξισώσεις, δηλαδή του $(\mathcal{C}_n^{-1}(f)T_n(f))^H\mathcal{C}_n^{-1}(f)T_n(f)$.

Για να μελετήσουμε τη συσσώρευση των ιδιοτιμών του προαναφερθέντος πίνακα, ακολουθούμε την απόδειξη του Θεωρήματος 2, της [14] και καταλήγουμε στο ότι το πολύ $4M$ ιδιοτιμές του $(\mathcal{C}_n^{-1}(f)T_n(f))^H\mathcal{C}_n^{-1}(f)T_n(f) - I_n$, έχουν απόλυτη τιμή μεγαλύτερη του ε. Το τελευταίο αποτέλεσμα είναι ισοδύναμο με την κύρια συσσώρευση των ιδιαζουσών τιμών του προρρυθμισμένου συστήματος $\mathcal{C}_n^{-1}(f)T_n(f)$ στο $[1-\varepsilon, 1+\varepsilon]$. □

Παρακάτω, με $\mathcal{T}_n$ συμβολίζουμε τον βέλτιστο κυκλοειδή προρρυθμιστή που προτάθηκε από τον T. Chan στην [16].

Θεώρημα 3.2.5. Έστω f μια 2π-περιοδική και συνεχής, μιγαδική συνάρτηση, η οποία έχει ρίζες στο $(-\pi, \pi]$. Έστω επίσης g ένα τριγωνομετρικό πολυώνυμο τέτοιο ώστε η $\frac{f}{g}$ να μην έχει ρίζες στο $(-\pi, \pi]$. Τότε, για κάθε $\varepsilon > 0$, το διάστημα $[1-\varepsilon, 1+\varepsilon]$ αποτελεί σύνολο κύριας συσσώρευσης των ιδιαζουσών τιμών του $\mathcal{P}_n^{-1}\left(\frac{f}{g}\right)T_n^{-1}(g)T_n(f)$, όπου $\mathcal{P}_n\left(\frac{f}{g}\right)$ είναι είτε ο κυκλοειδής πίνακας $\mathcal{C}_n\left(\frac{f}{g}\right)$, είτε ο βέλτιστος κυκλοειδής προρρυθμιστής $\mathcal{T}_n$, που προέκυψε από τον $T_n\left(\frac{f}{g}\right)$.

Απο🕱δειξη. Παρακάτω δίνουμε την απόδειξη για την περίπτωση όπου $\mathcal{P}_n\left(\frac{f}{g}\right) = \mathcal{C}_n\left(\frac{f}{g}\right)$. Η απόδειξη για τη χρήση του $\mathcal{T}_n$ ως προρρυθμιστή θα είναι ακριβώς ίδια, λαμβάνοντας υπόψιν το Πόρισμα 1 της [14], το οποίο είναι το αντίστοιχο του Πορίσματος 3.2.3.

Έστω g το τριγωνομετρικό πολυώνυμο που έχει βαθμό $\deg(g) = d$. Τότε, ο $T_n(g)$ θα είναι ένας ταινιωτός πίνακας Toeplitz με πλάτος ταινίας $2d+1$. Ακολουθώντας την ίδια τεχνική με το Θεώρημα 3.2.4 και το Θεώρημα 2.1.1, θα αναλύσουμε το αντίστοιχο προρρυθμισμένο σύστημα των κανονικών εξισώσεων. Έχουμε:

$$
\begin{aligned}
&\mathcal{C}_n^{-1}\left(\frac{f}{g}\right) T_n^{-1}(g) T_n(f) T_n(\bar{f}) T_n^{-1}(\bar{g}) \mathcal{C}_n^{-1}\left(\frac{\bar{f}}{\bar{g}}\right) = \\
&\mathcal{C}_n^{-1}\left(\frac{f}{g}\right) T_n^{-1}(g) \left[T_n(g) T_n\left(\frac{f}{g}\right) + L_1\right] \\
&\qquad\qquad \left[T_n\left(\frac{\bar{f}}{\bar{g}}\right) T_n(\bar{g}) + L_1^H\right] T_n^{-1}(\bar{g}) \mathcal{C}_n^{-1}\left(\frac{\bar{f}}{\bar{g}}\right) = \\
&\left[\mathcal{C}_n^{-1}\left(\frac{f}{g}\right) T_n\left(\frac{f}{g}\right) + L_2\right] \left[T_n\left(\frac{\bar{f}}{\bar{g}}\right) \mathcal{C}_n^{-1}\left(\frac{\bar{f}}{\bar{g}}\right) + L_2^H\right] = \\
&\mathcal{C}_n^{-1}\left(\frac{f}{g}\right) T_n\left(\frac{f}{g}\right) T_n\left(\frac{\bar{f}}{\bar{g}}\right) \mathcal{C}_n^{-1}\left(\frac{\bar{f}}{\bar{g}}\right) + L_3.
\end{aligned}
$$

Οι πίνακες L_1 και L_1^H παραπάνω είναι χαμηλής βαθμίδας, το πολύ ίσης με $2d$ (d είναι ο βαθμός του g). Το ίδιο ισχύει και για τους L_2 και L_2^H. Έτσι, καταλήγουμε στο ότι ο L_3 είναι Ερμιτιανός πίνακας με βαθμίδα το πολύ ίση με $4d$.

Έστω ότι p_M είναι ένα τριγωνομετρικό πολυώνυμο προσέγγισης της $\frac{f}{g}$ (όπως στο Θεώρημα 3.2.2). Επειδή η $\frac{f}{g}$ δεν έχει ρίζες, από το Θεώρημα 3.2.4 έχουμε ότι $\forall \varepsilon > 0$, το πολύ $4M$ ιδιάζουσες τιμές του $\mathcal{C}_n^{-1}(\frac{f}{g})T_n(\frac{f}{g})$ θα κυμαίνονται εκτός του $[1-\varepsilon, 1+\varepsilon]$. Λαμβάνοντας υπόψιν και αυτές που κυμαίνονται εκτός του διαστήματος, λόγω του πίνακα L_3, έχουμε ότι $\forall \varepsilon > 0$, το πολύ $4M + 4d$ ιδιάζουσες τιμές του προρρυθμισμένου συστήματος $\mathcal{C}_n^{-1}(\frac{f}{g})T_n^{-1}(g)T_n(f)$ θα κυμαίνονται εκτός του διαστήματος $[1-\varepsilon, 1+\varepsilon]$ κι έτσι η απόδειξη ολοκληρώνεται. □

Θεώρημα 3.2.6. Έστω f μια 2π-περιοδική, συνεχής και μιγαδική συνάρτηση, η οποία δεν έχει ρίζες στο $(-\pi, \pi]$. Τότε, οι ιδιοτιμές του $\mathcal{C}_n^{-1}(f)T_n(f)$ συσσωρεύονται, με την έννοια της κύριας συσσώρευσης, γύρω από το σημείο $(1, 0)$.

Αποδειξη. Γνωρίζουμε ότι ένας πίνακας A μπορεί να γραφεί ως το άθροισμα του Ερμιτιανού του και αντι-Ερμιτιανού του μέρους, $A = H(A) + SH(A) = \frac{A+A^H}{2} + \frac{A-A^H}{2}$. Για να μελετήσουμε τη συσσώρευση των ιδιοτιμών του A, μπορούμε να μελετήσουμε τη συσσώρευση των ιδιοτιμών του $H(A)$, καθώς και του $SH(A)$, λαμβάνοντας υπόψιν ότι $\mathrm{Re}(\lambda(A)) \in \mathrm{range}(H(A))$ και $\mathrm{Im}(\lambda(A)) \in \mathrm{range}(SH(A))$ [4, 30, 37].

Από το Θεώρημα 3.2.2, μπορούμε να αντικαταστήσουμε το $T_n(f)$ με $\mathcal{C}_n(f) + S_n + L_n$, όπου $\|S_n\|_2 < \epsilon$ και $\mathrm{rank}(L_n) \leq 2M$. Επομένως,

$$\mathcal{C}_n^{-1}(f)T_n(f) = \mathcal{C}_n^{-1}(f)\left[\mathcal{C}_n(f) + S_n + L_n\right] = I_n + \mathcal{C}_n^{-1}(f)S_n + \mathcal{C}_n^{-1}(f)L_n.$$

Το Ερμιτιανό του μέρος είναι:

$$\begin{aligned} H(\mathcal{C}_n^{-1}(f)T_n(f)) &= H(I_n + \mathcal{C}_n^{-1}(f)S_n + \mathcal{C}_n^{-1}(f)L_n) \\ &= I_n + H(\mathcal{C}_n^{-1}(f)S_n) + H(\mathcal{C}_n^{-1}(f)L_n). \end{aligned}$$

Σύμφωνα με την ανάλυση του Θεωρήματος 3.2.2, ο τελευταίος όρος είναι ένας πίνακας χαμηλής βαθμίδας, το πολύ ίσης με $4M$ και σχετίζεται με το ότι το πολύ $4M$ ιδιοτιμές του προρρυθμισμένου πίνακα έχουν πραγματικό μέρος εκτός του διαστήματος συσσώρευσης. Μένει να μελετήσουμε το φάσμα του $H(\mathcal{C}_n^{-1}(f)S_n)$.

$$\begin{aligned} |\lambda(H(\mathcal{C}_n^{-1}(f)S_n))| &= \left|\lambda\left(\frac{\mathcal{C}_n^{-1}(f)S_n + S_n^H\mathcal{C}^{-1}(\bar{f})}{2}\right)\right| \\ &\leq \left\|\frac{\mathcal{C}_n^{-1}(f)S_n + S_n^H\mathcal{C}^{-1}(\bar{f})}{2}\right\|_2 \\ &\leq \left\|\mathcal{C}_n^{-1}(f)S_n\right\|_2 \leq \left\|\mathcal{C}_n^{-1}(f)\right\|_2 \epsilon \leq 2\left\|\frac{1}{f}\right\|_\infty \epsilon. \end{aligned}$$

Επιλέγοντας ως $\varepsilon = 2\left\|\frac{1}{f}\right\|_\infty \epsilon$, έχουμε ότι τα πραγματικά μέρη των ιδιοτιμών, του προρρυθμισμένου πίνακα, συσσωρεύονται στο $[1-\varepsilon, 1+\varepsilon]$.

Το αντι-Ερμιτιανό μέρος του προρρυθμισμένου πίνακα είναι:

$$SH(\mathcal{C}_n^{-1}(f)T_n(f)) = SH(\mathcal{C}_n^{-1}(f)S_n) + SH(\mathcal{C}_n^{-1}(f)L_n).$$

Μέσω της ίδιας ανάλυσης λαμβάνουμε ότι ο τελευταίος όρος σχετίζεται με το ότι το πολύ $4M$ ιδιοτιμές του προρρυθμισμένου πίνακα έχουν φανταστικό μέρος εκτός του διαστήματος συσσώρευσης. Μελετώντας ανάλογα το $SH(\mathcal{C}_n^{-1}(f)S_n)$ έχουμε:

$$|\lambda(SH(\mathcal{C}_n^{-1}(f)S_n))| \leq \left\|\mathcal{C}_n^{-1}(f)S_n\right\|_2 \leq 2\left\|\frac{1}{f}\right\|_\infty \epsilon.$$

Επομένως, τα φανταστικά μέρη των ιδιοτιμών του προρρυθμισμένου πίνακα συσσωρεύονται στο $[-\varepsilon, \varepsilon]$, και το θεώρημα αποδείχθηκε. □

Θεώρημα 3.2.7. Έστω f μια 2π-περιοδική, συνεχής και μιγαδική συνάρτηση, η οποία έχει ρίζες στο $(-\pi, \pi]$. Έστω επίσης g ένα τριγωνομετρικό πολυώνυμο τέτοιο ώστε η $\frac{f}{g}$ να μην έχει ρίζες στο $(-\pi, \pi]$. Τότε, οι ιδιοτιμές του $\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n^{-1}(g)T_n(f)$ συσσωρεύονται, με την έννοια της κύριας συσσώρευσης, γύρω από το σημείο $(1, 0)$.

Απόδειξη. Έστω ότι h συμβολίζει το πηλίκο Rayleigh του Ερμιτιανού μέρους του προρρυθμισμένου συστήματος $\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n^{-1}(g)T_n(f)$. Έχουμε:

$$
\begin{aligned}
h &= \frac{1}{2}\frac{x^H\left[\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n^{-1}(g)T_n(f) + T_n(\bar{f})T_n^{-1}(\bar{g})\mathcal{C}_n^{-1}\left(\frac{\bar{f}}{\bar{g}}\right)\right]x}{x^Hx} \\
&= \frac{1}{2}\frac{x^H\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n^{-1}(g)\left[T_n(g)T_n\left(\frac{f}{g}\right) + L_1\right]x}{x^Hx} \\
&\qquad + \frac{1}{2}\frac{x^H\left[T_n\left(\frac{\bar{f}}{\bar{g}}\right)T_n(\bar{g}) + L_1^H\right]T_n^{-1}(\bar{g})\mathcal{C}_n^{-1}\left(\frac{\bar{f}}{\bar{g}}\right)x}{x^Hx} \\
&= \frac{1}{2}\frac{x^H\left[\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n\left(\frac{f}{g}\right) + T_n\left(\frac{\bar{f}}{\bar{g}}\right)\mathcal{C}_n^{-1}\left(\frac{\bar{f}}{\bar{g}}\right)\right]x}{x^Hx} + \frac{1}{2}\frac{x^HL_2x}{x^Hx},
\end{aligned}
$$

όπου $L_2 = \mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n^{-1}(g)L_1 + L_1^HT_n^{-1}(\bar{g})\mathcal{C}_n^{-1}\left(\frac{\bar{f}}{\bar{g}}\right)$ είναι ένας Ερμιτιανός πίνακας χαμηλής βαθμίδας, το πολύ ίσης με $4d$ (d είναι ο βαθμός του g).

Το πρώτο πηλίκο Rayleigh μας δίνει το εύρος του $H\left(\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n\left(\frac{f}{g}\right)\right)$, για όλα τα $x \in \mathbb{R}^n$. Επομένως, τα πραγματικά μέρη των ιδιοτιμών του $\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n\left(\frac{f}{g}\right)$ συσσωρεύονται στο range $\left(H\left(\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n\left(\frac{f}{g}\right)\right)\right)$. Λόγω του ότι η συνάρτηση $\frac{f}{g}$ δεν έχει ρίζες, το Θεώρημα 3.2.6 ισχύει για την $\frac{f}{g}$, αντί της f. Αυτό σημαίνει ότι τα πραγματικά μέρη των ιδιοτιμών του $\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n\left(\frac{f}{g}\right)$ συσσωρεύονται στο $[1-\varepsilon, 1+\varepsilon]$, με το πολύ $4M$ ιδιοτιμές εκτός του διαστήματος (συσσώρευσης). Λαμβάνοντας υπόψιν και τον πίνακα χαμηλής βαθμίδας L_2, συμπεραίνουμε ότι τα πραγματικά μέρη των ιδιοτιμών του προρρυθμισμένου συστήματος συσσωρεύονται στο $[1-\varepsilon, 1+\varepsilon]$ κι έχουμε το πολύ $4M + 4d$ ιδιοτιμές εκτός του διαστήματος.

Μελετώντας με τον ίδιο τρόπο το αντι-Ερμιτιανό μέρος του προρρυθμισμένου πίνακα $\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n^{-1}(g)T_n(f)$, έχουμε ότι τα φανταστικά μέρη των ιδιοτιμών του προρρυθμισμένου συστήματος συσσωρεύονται στο $[-\varepsilon, \varepsilon]$ με το πολύ $4M + 4d$ ιδιοτιμές εκτός του διαστήματος. $\square$

Παρατήρηση. Είναι προφανές ότι τα Θεωρήματα 3.2.6 και 3.2.7, ισχύουν και για τον βέλτιστο κυκλοειδή προρρυθμιστή. Οι αποδείξεις είναι ακριβώς οι ίδιες.

3.2.2 Κατά τμήματα συνεχής περίπτωση

Στην προηγούμενη υποενότητα μελετήσαμε την περίπτωση όπου η $f = f_1 + \mathrm{i} f_2$ είναι συνεχής συνάρτηση. Ένα λογικό ερώτημα το οποίο γεννιέται είναι: ""Τι ισχύει όταν η f_2 παρουσιάζει ασυνέχεια στο $-\pi$ και π?"". Γενικότερα, ""Τι ισχύει αν η f είναι κατά τμήματα συνεχής συνάρτηση στο $(-\pi, \pi]$?"". Θα αποδείξουμε ανάλογα θεωρήματα, προκειμένου να απαντήσουμε σε αυτά τα ερωτήματα. Από εδώ και στο εξής, υποθέτουμε ότι η f_1 είναι πραγματική, άρτια και 2π-περιοδική συνάρτηση, η f_2 πραγματική, περιττή κι επίσης 2π-περιοδική, και υποθέτουμε επίσης ότι η f είναι κατά τμήματα συνεχής στο $(-\pi, \pi]$.

Στην υποενότητα αυτή, χρησιμοποιούμε τον $\mathcal{C}_n(f)$ ή τον $\mathcal{C}_n\left(\frac{f}{g}\right)$ (όταν χρειάζεται η άρση των ριζών) για να κατασκευάσουμε τον προρρυθμιστή. Τα θεωρητικά αποτελέσματα, τα οποία θα αποδείξουμε παρακάτω, ισχύουν και για τον βέλτιστο κυκλοειδή προρρυθμιστή $\mathcal{T}_n(f)$ ή $\mathcal{T}_n\left(\frac{f}{g}\right)$. Οι αποδείξεις μπορούν να γίνουν με τον ίδιο τρόπο, αν παρακάτω χρησιμοποιήσουμε το Λήμμα 8 της [89], αντί του Λήμματος 3.2.9.

Σημειώνουμε ότι για κατά τμήματα συνεχείς, καθώς επίσης και για συνεχείς συναρτήσεις, ο βέλτιστος κυκλοειδής προρρυθμιστής $\mathcal{T}_n\left(\frac{f}{g}\right)$, σε συνδυασμό με τον ταινιωτό πίνακα Toeplitz $T_n(g)$, χρησιμοποιείται για πρώτη φορά. Όσον αφορά στην κατασκευή του προρρυθμιστή $\mathcal{T}_n\left(\frac{f}{g}\right)$, ο υπολογισμός του $T_n\left(\frac{f}{g}\right)$ απαιτείται, ενώ ο $\mathcal{C}_n\left(\frac{f}{g}\right)$ κατασκευάζεται εύκολα από τη συνάρτηση $\frac{f}{g}$.

Θεώρημα 3.2.8. Έστω f μια συνάρτηση χωρίς ρίζες, η οποία γράφεται ως $f = f_1 + \mathrm{i} f_2$, f_1 2π-περιοδική και άρτια, ενώ f_2 2π-περιοδική και περιττή συνάρτηση, έχοντας σημεία ασυνέχειας $\xi_1, \xi_2, \ldots, \xi_\nu \in (0, 2\pi]$ με εύρος ασυνέχειας (jump of discontinuity), τη φραγμένη ποσότητα:

$$\alpha_k = \lim_{x \to \xi_k^+} f(x) - \lim_{x \to \xi_k^-} f(x).$$

Τότε, $\mathcal{O}(\log n)$ ιδιοτιμές του $\Delta_n = T_n(f) - \mathcal{C}_n(f)$ κυμαίνονται εκτός του ορθογωνίου $[-\varepsilon, \varepsilon]^2$ του μιγαδικού επιπέδου.

Θα δώσουμε την απόδειξη του Θεωρήματος 3.2.8, αφού πρώτα δώσουμε το παρακάτω λήμμα, με την απόδειξή του.

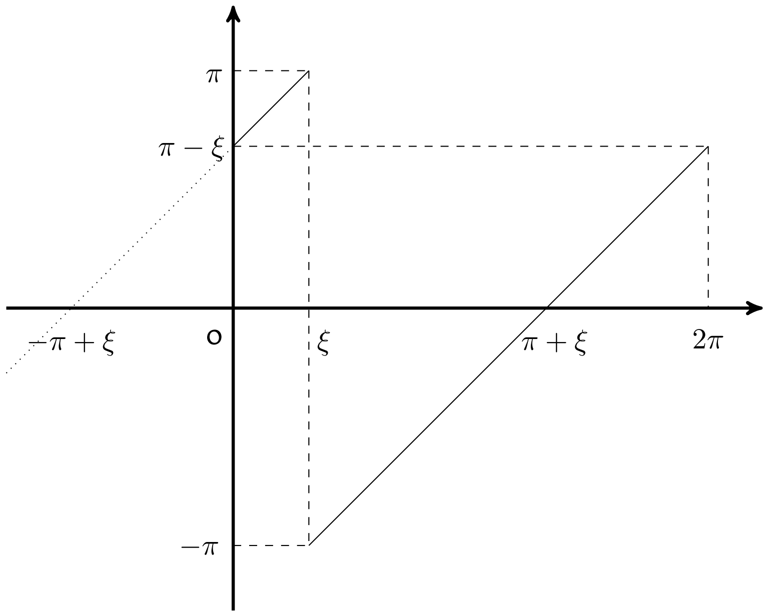
Λήμμα 3.2.9. Έστω ξ ένα τυχαίο σημείο στο διάστημα $(0, 2\pi]$ και g ορισμένη ως:

$$\mathfrak{g}(x) = \begin{cases} x + \pi - \xi, & 0 < x \leq \xi, \\ x - \pi - \xi, & \xi < x \leq 2\pi. \end{cases}$$

Τότε, $T_n(\mathfrak{g}) - \mathcal{C}_n(\mathfrak{g}) = A_n + B_n$, όπου $\|A_n\|_F$ είναι φραγμένη από μια σταθερά ανεξάρτητη της διάστασης n και $\|B_n\|_F$ τείνει στο άπειρο όπως ο $\log n$ ($\|B_n\|_F = \mathcal{O}(\log n)$).

Απόδειξη. Αρχικά, είναι εύκολο να βρει κανείς [89] ότι οι διαγώνιοι του πίνακα $T_n(\mathfrak{g})$ δίνονται ως:

$$t_k = \begin{cases} 0, & k = 0, \\ \frac{\mathrm{i}}{k}\mathrm{e}^{-\mathrm{i}k\xi}, & k \neq 0. \end{cases} \tag{3.7}$$



Σχήμα 3.1: Γραφική παράσταση της $\mathfrak{g}(x)$.

Χωρίς βλάβη της γενικότητας, επιτρέψτε μας να υποθέσουμε ότι $\frac{2\pi m}{n} \leq \xi < \frac{2\pi(m+1)}{n} < \pi$, όπου $m \in \mathbb{N}$. Θα υπολογίσουμε τις τιμές c_k του κυκλοειδούς πίνακα $\mathcal{C}_n(\mathfrak{g}) = \mathcal{F}_n^H \Delta(\mathfrak{g}) \mathcal{F}_n$. Έχουμε:

$$c_k = \frac{1}{n} \sum_{j=0}^{n-1} \mathrm{e}^{-\mathrm{i}kj2\pi/n} \mathfrak{g}\left(\frac{2\pi j}{n}\right) \tag{3.8}$$

Για $k = 0$ είναι προφανές, από τη σχέση (3.8), ότι η κύρια διαγώνιος του $\mathcal{C}_n(\mathfrak{g})$ είναι το άθροισμα των ιδιοτιμών του, διαιρεμένο με n:

$$c_0 = \frac{1}{n}\sum_{j=0}^{n-1}\mathfrak{g}\left(\frac{2\pi j}{n}\right).$$

Αναλυτικότερα, για τις τιμές του $\mathcal{C}_n(\mathfrak{g})$ έχουμε:

$$\begin{aligned}
c_k &= \frac{1}{n}\sum_{j=0}^{m}\left[\frac{2\pi\left(j+\frac{n}{2}\right)}{n}-\xi\right]\mathrm{e}^{-\mathrm{i}kj2\pi/n} + \frac{1}{n}\sum_{j=m+1}^{n-1}\left[\frac{2\pi\left(j-\frac{n}{2}\right)}{n}-\xi\right]\mathrm{e}^{-\mathrm{i}kj2\pi/n}\\
&= \frac{1}{n}\sum_{j=0}^{m}\left[\frac{2\pi\left(j+\frac{n}{2}\right)}{n}-\xi\right]\mathrm{e}^{-\mathrm{i}kj2\pi/n}\\
&\quad+\frac{1}{n}\sum_{j=m+1-n}^{-1}\left[\frac{2\pi\left(j+\frac{n}{2}\right)}{n}-\xi\right]\mathrm{e}^{-\mathrm{i}k(j+n)2\pi/n}\\
&= \frac{1}{n}\sum_{j=m+1-n}^{m}\left[\frac{2\pi\left(j+\frac{n}{2}\right)}{n}-\xi\right]\mathrm{e}^{-\mathrm{i}kj2\pi/n}\\
&= \frac{1}{n}\sum_{j=0}^{n-1}\left[\frac{2\pi\left(j+m+1-\frac{n}{2}\right)}{n}-\xi\right]\mathrm{e}^{-\mathrm{i}k(j+m+1-n)2\pi/n}\\
&= \frac{1}{n}\left(\frac{2\pi\left(m+\frac{1}{2}\right)}{n}-\xi\right)\sum_{j=0}^{n-1}\mathrm{e}^{-\mathrm{i}k(j+m+1)2\pi/n}\\
&\quad+\frac{1}{n}\sum_{j=0}^{n-1}\frac{2\pi\left(j+\frac{1}{2}-\frac{n}{2}\right)}{n}\mathrm{e}^{-\mathrm{i}k(j+m+1)2\pi/n}\\
&= S_k^{(1)}+S_k^{(2)}.
\end{aligned}$$

- Για $k = 0$, $S_0^{(1)} = \frac{1}{n}\left(\frac{2\pi(m+\frac{1}{2})}{n}-\xi\right)\sum\limits_{j=0}^{n-1}1 = \theta$, όπου:

  $$\theta = \frac{2\pi\left(m+\frac{1}{2}\right)}{n}-\xi$$

  και από τον τρόπο που ορίσαμε το m, ισχύει $|\theta| \leq \frac{\pi}{n}$.

- Για $k \neq 0$, $S_k^{(1)} = 0$, επειδή οι όροι του αθροίσματος είναι ισαπέχοντα σημεία στον μοναδιαίο κύκλο (οι n-οστές ρίζες της μονάδας) και το άθροισμα αυτών είναι ίσο με μηδέν.

Τώρα θα υπολογίσουμε το άθροισμα $S_k^{(2)}$:

$$
\begin{aligned}
S_k^{(2)} &= \frac{1}{n}\sum_{j=0}^{n-1} \frac{2\pi\left(j+\frac{1}{2}-\frac{n}{2}\right)}{n} \mathrm{e}^{-\mathrm{i}k(j+m+1)2\pi/n} \\
&= \frac{1}{n}\sum_{j=0}^{\frac{n}{2}-1} \frac{2\pi\left(j+\frac{1}{2}-\frac{n}{2}\right)}{n} \mathrm{e}^{-\mathrm{i}k(j+m+1)2\pi/n} \\
&\quad + \frac{1}{n}\sum_{j=\frac{n}{2}}^{n-1} \frac{2\pi\left(j+\frac{1}{2}-\frac{n}{2}\right)}{n} \mathrm{e}^{-\mathrm{i}k(j+m+1)2\pi/n} \\
&= \frac{1}{n}\sum_{j=1}^{\frac{n}{2}} \frac{2\pi\left(j-\frac{1}{2}-\frac{n}{2}\right)}{n} \mathrm{e}^{-\mathrm{i}k(j+m)2\pi/n} \\
&\quad + \frac{1}{n}\sum_{j=1}^{\frac{n}{2}} \frac{2\pi\left(\frac{n}{2}-j+\frac{1}{2}\right)}{n} \mathrm{e}^{-\mathrm{i}k(n-j+m+1)2\pi/n} \\
&= \frac{1}{n}\sum_{j=1}^{n/2} \frac{2\pi\left(\frac{n}{2}-j+\frac{1}{2}\right)}{n} \left(\mathrm{e}^{-\mathrm{i}k(m+1-j)2\pi/n} - \mathrm{e}^{-\mathrm{i}k(m+j)2\pi/n}\right).
\end{aligned}
$$

Από την παραπάνω ανάλυση, είναι ξεκάθαρο ότι $c_0 = S_0^{(1)} + S_0^{(2)} = \theta$. Επιπλέον, συμπεραίνουμε ότι οι τιμές c_k, για $k \neq 0$ δίνονται ως $c_k = S_k^{(2)}$. Οπότε, έχουμε:

$$
\begin{aligned}
c_k &= \frac{1}{n}\mathrm{e}^{-\mathrm{i}k\left(m+\frac{1}{2}\right)2\pi/n} \sum_{j=1}^{n/2} \frac{2\pi}{n}\left(\frac{n}{2}-j+\frac{1}{2}\right)\left(\mathrm{e}^{\mathrm{i}k\left(j-\frac{1}{2}\right)2\pi/n} - \mathrm{e}^{-\mathrm{i}k\left(j-\frac{1}{2}\right)2\pi/n}\right) \\
&= \frac{2\mathrm{i}}{n}\mathrm{e}^{-\mathrm{i}k\left(m+\frac{1}{2}\right)2\pi/n} \sum_{j=1}^{n/2} \frac{2\pi}{n}\left(\frac{n}{2}-j+\frac{1}{2}\right)\sin\left[k\left(j-\frac{1}{2}\right)\frac{2\pi}{n}\right].
\end{aligned}
$$

Θα υπολογίσουμε την ποσότητα:

$$
\frac{2}{n}\sum_{j=1}^{n/2} \frac{2\pi}{n}\left(\frac{n}{2}-j+\frac{1}{2}\right)\sin\left[k\left(j-\frac{1}{2}\right)\frac{2\pi}{n}\right]. \tag{3.9}
$$

Πολλαπλασιάζοντας με $\sin\left(\frac{k\pi}{n}\right)$, έχουμε:

$$
\begin{aligned}
&\frac{2}{n}\sum_{j=1}^{n/2}\frac{2\pi}{n}\left(\frac{n}{2}-j+\frac{1}{2}\right)\sin\left[k\left(j-\frac{1}{2}\right)\frac{2\pi}{n}\right]\sin\left(\frac{k\pi}{n}\right)\\
&=\frac{1}{n}\sum_{j=1}^{n/2}\frac{2\pi}{n}\left(\frac{n}{2}-j+\frac{1}{2}\right)\left\{\cos\left[k\left(j-1\right)\frac{2\pi}{n}\right]-\cos\left(kj\frac{2\pi}{n}\right)\right\}\\
&=\frac{1}{n}\sum_{j=0}^{n/2-1}\frac{2\pi}{n}\left(\frac{n}{2}-j-\frac{1}{2}\right)\cos\left(kj\frac{2\pi}{n}\right)\\
&\quad-\frac{1}{n}\sum_{j=1}^{n/2}\frac{2\pi}{n}\left(\frac{n}{2}-j+\frac{1}{2}\right)\cos\left(kj\frac{2\pi}{n}\right)\\
&=\frac{1}{n}\frac{2\pi}{n}\frac{n-1}{2}-\frac{1}{n}\frac{2\pi}{n}\sum_{j=1}^{n/2-1}\cos\left(kj\frac{2\pi}{n}\right)-\frac{1}{n}\frac{\pi}{n}\cos\left(k\pi\right).
\end{aligned}
$$

‘Οταν k είναι περιττός αριθμός, η παραπάνω ποσότητα είναι ίση με $\frac{1}{n}\frac{2\pi}{n}\frac{n-1}{2}+\frac{1}{n}\frac{\pi}{n}=\frac{\pi}{n}$. Αντιστοίχως, όταν k είναι άρτιος, $\frac{1}{n}\frac{2\pi}{n}\frac{n-1}{2}-\frac{1}{n}\frac{2\pi}{n}\displaystyle\sum_{j=0}^{n/2-1}\cos\left(kj\frac{2\pi}{n}\right)+\frac{1}{n}\frac{2\pi}{n}-\frac{1}{n}\frac{\pi}{n}=\frac{\pi}{n}$.

Επομένως, η σχέση (3.9) μπορεί να γραφεί ως $\frac{\pi}{n\sin\left(\frac{k\pi}{n}\right)}$ κι έτσι οι τιμές c_k, όταν $k\neq 0$ δίνονται ως:

$$
c_k=\frac{\mathrm{i}\pi}{n\sin\left(\frac{k\pi}{n}\right)}\mathrm{e}^{-\mathrm{i}k(m+\frac{1}{2})2\pi/n}. \tag{3.10}
$$

Μένει να υπολογίσουμε τον πίνακα (διαφορά πινάκων) $T_n(\mathfrak{g})-\mathcal{C}_n(\mathfrak{g})$. Η κύρια διαγώνιός του δίνεται ως $d_0=-c_0=-\theta$ και λαμβάνοντας υπόψη τις (3.7) και (3.10) παρατηρούμε ότι οι k-οστές διαγώνιοί του d_k, όταν $k\neq 0$ δίνονται ως:

$$
\begin{aligned}
d_k=t_k-c_k&=\frac{\mathrm{i}}{k}\mathrm{e}^{-\mathrm{i}k\xi}-\frac{\mathrm{i}\pi}{n\sin\left(\frac{k\pi}{n}\right)}\mathrm{e}^{-\mathrm{i}k(m+\frac{1}{2})2\pi/n}\\
&=\frac{\mathrm{i}}{k}\left(\mathrm{e}^{-\mathrm{i}k\xi}-\mathrm{e}^{-\mathrm{i}k(m+\frac{1}{2})2\pi/n}\right)+\mathrm{i}\left(\frac{1}{k}-\frac{\pi}{n\sin\left(\frac{k\pi}{n}\right)}\right)\mathrm{e}^{-\mathrm{i}k(m+\frac{1}{2})2\pi/n}.
\end{aligned}
\tag{3.11}
$$

Θα αποδείξουμε ότι ο πρώτος όρος της παραπάνω σχέσης είναι της τάξεως $\frac{1}{n}$. Έχουμε:

$$\begin{aligned}\frac{\mathrm{i}}{k}\left(\mathrm{e}^{-\mathrm{i}k\xi}-\mathrm{e}^{-\mathrm{i}k(m+\frac{1}{2})2\pi/n}\right)&=\frac{\mathrm{i}}{k}\mathrm{e}^{-\mathrm{i}k\xi}\left(1-\mathrm{e}^{-\mathrm{i}k((m+\frac{1}{2})2\pi/n-\xi)}\right)\\&=\frac{\mathrm{i}}{k}\mathrm{e}^{-\mathrm{i}k\xi}\left(1-\mathrm{e}^{-\mathrm{i}k\theta}\right).\end{aligned}$$

Αναπτύσοντας κατά Taylor την ποσότητα $\mathrm{e}^{-\mathrm{i}k\theta}$ γύρω από το 0, λαμβάνουμε:

$$\frac{\mathrm{i}}{k}\mathrm{e}^{-\mathrm{i}k\xi}\left(1-\mathrm{e}^{-\mathrm{i}k\theta}\right)=\frac{\mathrm{i}}{k}\mathrm{e}^{-\mathrm{i}k\xi}\left(-\mathrm{i}k\theta\mathrm{e}^{-\mathrm{i}k\widehat{\theta}}\right)=\theta\mathrm{e}^{-\mathrm{i}k(\xi+\widehat{\theta})},\ |\widehat{\theta}|<|\theta|.$$

Προφανώς, $\left|\theta\mathrm{e}^{-\mathrm{i}k(\xi+\widehat{\theta})}\right|=|\theta|=\mathcal{O}\left(\frac{1}{n}\right)$.

Γράφουμε τη διαφορά πινάκων $T_n(\mathfrak{g})-\mathcal{C}_n(\mathfrak{g})$ ως το άθροισμα δύο πινάκων $T_n(\mathfrak{g})-\mathcal{C}_n(\mathfrak{g})=\widetilde{A}_n+\widetilde{B}_n$, όπου οι τιμές του $\widetilde{A}_n$ είναι της τάξεως $\mathcal{O}\left(\frac{1}{n}\right)$ και αυτές του $\widetilde{B}_n$ είναι $\Omega\left(\frac{1}{n}\right)$. Επομένως, η k-οστή διαγώνιος του $\widetilde{A}_n$ αποτελείται από τις τιμές $\frac{\mathrm{i}}{k}\left(\mathrm{e}^{-\mathrm{i}k\xi}-\mathrm{e}^{-\mathrm{i}k(m+\frac{1}{2})2\pi/n}\right)$.

Στη συνέχεια, θα εκτιμήσουμε την τάξη του δεύτερου όρου της (3.11). Αφού $\mathrm{e}^{-\mathrm{i}k(m+\frac{1}{2})2\pi/n}$ είναι μέτρου 1, προσπαθούμε να εκτιμήσουμε την τάξη των όρων $\frac{1}{k}-\frac{\pi}{n\sin\left(\frac{k\pi}{n}\right)}$ για τις διάφορες τιμές του $k=1,2,\ldots,n-1$.

Διακρίνουμε τρεις περιπτώσεις, λαμβάνοντας υπόψη το μέγεθος του k σε σχέση με το n.
1) $k\sim n$ και $n-k\sim n$: Αυτό σημαίνει ότι $k=\alpha n$, όπου $0<\alpha<1$ είναι μια σταθερά ανεξάρτητη του n. Τότε:

$$\begin{aligned}\frac{1}{k}-\frac{\pi}{n\sin\left(\frac{k\pi}{n}\right)}&=\frac{1}{k}-\frac{1}{\sin(\alpha\pi)}\frac{\pi}{n}=\frac{1}{k}-\beta\frac{\pi}{n}=-\frac{1}{n-k}+\left(\frac{1}{k}-\beta\frac{\pi}{n}+\frac{1}{n-k}\right)\\&=-\frac{1}{n-k}+\left(\frac{1}{\alpha n}-\beta\frac{\pi}{n}+\frac{1}{(1-\alpha)n}\right)=-\frac{1}{n-k}+\mathcal{O}\left(\frac{1}{n}\right).\end{aligned}$$

2) $n-k=o(n)\Leftrightarrow k=n-o(n)$: Παρατηρούμε ότι, $\sin\left(\frac{k\pi}{n}\right)=\sin\left((n-k)\frac{\pi}{n}\right)$. Εφαρμόζοντας ανάπτυγμα Taylor έχουμε:

$$\sin\left((n-k)\frac{\pi}{n}\right)=(n-k)\frac{\pi}{n}-\frac{1}{6}(n-k)^3\frac{\pi^3}{n^3}\cos\widetilde{\theta},\ 0<\widetilde{\theta}<(n-k)\frac{\pi}{n}.$$

Επομένως:

$$\frac{1}{k}-\frac{\pi}{n\sin\left(\frac{k\pi}{n}\right)}=\frac{1}{k}-\frac{1}{(n-k)-\frac{1}{6}(n-k)^3\frac{\pi^2}{n^2}\cos\widetilde{\theta}}$$
$$\simeq\frac{1}{k}-\frac{1}{(n-k)}+\frac{1}{6}(n-k)\frac{\pi^2}{n^2}\cos\widetilde{\theta}=-\frac{1}{n-k}+\mathcal{O}\left(\frac{1}{n}\right).$$

3) $k=o(n)$: Με παρόμοιο τρόπο

$$\frac{1}{k}-\frac{\pi}{n\sin\left(\frac{k\pi}{n}\right)}=\frac{1}{k}-\frac{1}{k-\frac{1}{6}k^3\frac{\pi^2}{n^2}\cos\widehat{\theta}}\simeq\frac{1}{k}-\frac{1}{k}+\frac{1}{6}k\frac{\pi^2}{n^2}\cos\widehat{\theta}$$
$$=-\frac{1}{n-k}+\frac{1}{n-k}+\frac{1}{6}k\frac{\pi^2}{n^2}\cos\widehat{\theta}$$
$$=-\frac{1}{n-k}+\mathcal{O}\left(\frac{1}{n}\right),\ 0<\widehat{\theta}<k\frac{\pi}{n}.$$

Χωρίζουμε τον πίνακα $\widetilde{B}_n$ ως $\widetilde{B}_n=\widehat{A}_n+\widehat{B}_n$, όπου ο $\widehat{A}_n$ έχει τιμές που χαρακτηρίζονται ως $\mathcal{O}\left(\frac{1}{n}\right)$ (στις τρεις περιπτώσεις που προαναφέραμε) και $\left(\widehat{B}_n\right)_k=-\frac{\mathrm{i}}{n-k}\mathrm{e}^{-\mathrm{i}k(m+\frac{1}{2})2\pi/n}$, $k>0$. Επειδή ο πίνακας $\widehat{B}_n$ είναι Ερμιτιανός, οι τιμές των διαγωνίων με αρνητικό δείκτη k, θα είναι οι συζυγείς των αντίστοιχων τιμών για $k>0$. Άρα, $\left(\widehat{B}_n\right)_{-k}=\overline{\left(\widehat{B}_n\right)_k}=\frac{\mathrm{i}}{n-k}\mathrm{e}^{\mathrm{i}k(m+\frac{1}{2})2\pi/n}$, $k>0$.

Τότε, χωρίζουμε επίσης τον $\widehat{B}_n$ ως:

$$\widehat{B}_n=\begin{pmatrix}V_{n/2} & U_{n/2}\\ U_{n/2}^H & V_{n/2}\end{pmatrix}=\begin{pmatrix}V_{n/2} & 0\\ 0 & V_{n/2}\end{pmatrix}+\begin{pmatrix}0 & U_{n/2}\\ U_{n/2}^H & 0\end{pmatrix}=A_n'+B_n.$$

Από την παραπάνω ανάλυση, συμπεραίνουμε ότι $A_n'=\mathcal{O}\left(\frac{1}{n}\right)$. Συνοψίζοντας, ο πίνακας $T_n(\mathfrak{g})-\mathcal{C}_n(\mathfrak{g})$ γράφεται ως A_n+B_n, όπου $A_n=\widetilde{A}_n+\widehat{A}_n+A_n'$ είναι ένας Ερμιτιανός πίνακας Toeplitz, με τιμές τάξεως $\mathcal{O}\left(\frac{1}{n}\right)$. Επομένως, η νόρμα Frobenius αυτού φράσσεται από μια σταθερά ανεξάρτητη του n κι έτσι τόσο οι ιδιοτιμές του, όσο και οι ιδιάζουσες τιμές αυτού, συσσωρεύονται κατά κύριο τρόπο γύρω από το μηδέν [22, 85].

Μένει να μελετήσουμε τον πίνακα B_n. Για να εκπληρώσουμε αυτόν τον σκοπό, ακολουθούμε την ίδια τεχνική που παρουσιάζεται στο Λήμμα 8 της [89]. Θεωρούμε τον πίνακα:

$$J_n = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ \vdots & & 1 & 0 \\ 0 & \iddots & & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}$$

και τους ορθομοναδιαίους πίνακες P_n και Q_n, όπως φαίνεται παρακάτω:

$$P_n = \mathrm{diag}\left(-1, -\mathrm{e}^{\mathrm{i}\zeta}, \ldots, -\mathrm{e}^{\mathrm{i}(\frac{n}{2}-1)\zeta}\right),$$

$$Q_n = \mathrm{diag}\left(-\mathrm{i}\mathrm{e}^{\mathrm{i}\frac{n}{2}\zeta}, -\mathrm{i}\mathrm{e}^{\mathrm{i}(\frac{n}{2}+1)\zeta}, \ldots, -\mathrm{i}\mathrm{e}^{\mathrm{i}(n-1)\zeta}\right),$$

όπου $\zeta = \left(m + \frac{1}{2}\right)\frac{2\pi}{n}$. Τότε, $U_{n/2} = P_{n/2}^H \mathcal{H}_{n/2} J_{n/2} Q_{n/2}$ όπου $\mathcal{H}_{n/2}$ είναι ο $\frac{n}{2} \times \frac{n}{2}$ πίνακας Hilbert. 'Ετσι, ο B_n μπορεί να γραφεί ως:

$$\begin{aligned}
\begin{pmatrix} 0 & U_{n/2} \\ U_{n/2}^H & 0 \end{pmatrix} &= \begin{pmatrix} P_{n/2}^H & 0 \\ 0 & Q_{n/2}^H \end{pmatrix} \begin{pmatrix} 0 & \mathcal{H}_{n/2} J_{n/2} \\ J_{n/2}\mathcal{H}_{n/2} & 0 \end{pmatrix} \begin{pmatrix} P_{n/2} & 0 \\ 0 & Q_{n/2} \end{pmatrix} \\
&= \frac{1}{2} \begin{pmatrix} P_{n/2}^H & 0 \\ 0 & Q_{n/2}^H \end{pmatrix} \begin{pmatrix} I_{n/2} & I_{n/2} \\ J_{n/2} & -J_{n/2} \end{pmatrix} \begin{pmatrix} \mathcal{H}_{n/2} & 0 \\ 0 & -\mathcal{H}_{n/2} \end{pmatrix} \cdot \\
&\quad \cdot \begin{pmatrix} I_{n/2} & J_{n/2} \\ I_{n/2} & -J_{n/2} \end{pmatrix} \begin{pmatrix} P_{n/2} & 0 \\ 0 & Q_{n/2} \end{pmatrix} \\
&= \frac{1}{2} \begin{pmatrix} P_{n/2}^H & P_{n/2}^H \\ Q_{n/2}^H J_{n/2} & -Q_{n/2}^H J_{n/2} \end{pmatrix} \begin{pmatrix} \mathcal{H}_{n/2} & 0 \\ 0 & -\mathcal{H}_{n/2} \end{pmatrix} \begin{pmatrix} P_{n/2} & J_{n/2} Q_{n/2} \\ P_{n/2} & -J_{n/2} Q_{n/2} \end{pmatrix}.
\end{aligned}$$

Παρατηρούμε ότι ο B_n είναι όμοιος με τον μεσαίο πίνακα του παραπάνω γινομένου, δηλαδή του

$$\begin{pmatrix} \mathcal{H}_{n/2} & 0 \\ 0 & -\mathcal{H}_{n/2} \end{pmatrix},$$

αφού ο πίνακας

$$\frac{1}{\sqrt{2}} \begin{pmatrix} P_{n/2}^H & P_{n/2}^H \\ Q_{n/2}^H J_{n/2} & -Q_{n/2}^H J_{n/2} \end{pmatrix}$$

είναι ορθογώνιος. Επειδή ο $\mathcal{H}_{n/2}$ είναι ο $\frac{n}{2} \times \frac{n}{2}$ πίνακας Hilbert, συμπεραίνουμε ότι η νόρμα $\|B_n\|_F$ τείνει στο άπειρο όπως ο $\log\left(\frac{n}{2}\right)$, το οποίο ισοδύναμα δηλώνει ότι το πλήθος των ιδιοτιμών του B_n, που έχουν απόλυτη τιμή μεγαλύτερη του $\varepsilon > 0$, είναι της τάξεως $\mathcal{O}\left(\log\left(\frac{n}{2}\right)\right)$ [88, 89]. $\square$

Απόδειξη του Θεωρήματος 3.2.8. Θεωρούμε ότι οι συναρτήσεις f_1 και f_2 έχουν σημεία ασυνέχειας $\xi_1, \xi_2, \ldots, \xi_\nu \in (0, 2\pi]$ με εύρη ασυνέχειας, τη φραγμένη ποσότητα α_k για την f_1 και β_k για την f_2 στο $\xi_k,\ k = 1, 2, \ldots, \nu$. Σε περίπτωση που η f_1 έχει ασυνέχεια στο σημείο ξ_k, όπου η f_2 είναι συνεχής, θέτουμε $\beta_k = 0$. Φυσικά, ακολουθούμε την ίδια λογική για την αντίστροφη περίπτωση.

Στη συνέχεια ακολουθούμε την τεχνική που εφάρμοσαν οι R. Chan και M-C. Yeung στην [89], αλλά στην περίπτωσή μας, έχουμε να μελετήσουμε μιγαδικές συναρτήσεις, αντί για πραγματικές. Θεωρούμε τις συναρτήσεις $\mathfrak{g}_k$ για κάθε αντίστοιχο σημείο ξ_k, όπως κάναμε στο Λήμμα 3.2.9:

$$\mathfrak{g}_k(x) = \begin{cases} x + \pi - \xi_k, & 0 < x \leq \xi_k, \\ x - \pi - \xi_k, & \xi_k < x \leq 2\pi. \end{cases} \quad k = 1, 2, \ldots, \nu,$$

η οποία είναι ασυνεχής στα σημεία ξ_k, με αντίστοιχο εύρος ασυνέχειας ίσο με -2π.

Ακολούθως, προσθαφαιρούμε στην f τη συνάρτηση,

$$\widehat{\mathfrak{g}}(x) = \sum_{k=1}^{\nu} \left(\frac{\alpha_k}{2\pi} + \mathrm{i}\frac{\beta_k}{2\pi} \right) \mathfrak{g}_k(x),$$

οπότε $f(x) = f_1(x) + \mathrm{i} f_2(x) = f_1(x) + \sum_{k=1}^{\nu} \frac{\alpha_k}{2\pi}\mathfrak{g}_k(x) + \mathrm{i} f_2(x) + \mathrm{i}\sum_{k=1}^{\nu} \frac{\beta_k}{2\pi}\mathfrak{g}_k(x) - \widehat{\mathfrak{g}}(x)$.

'Οπως στην [89], είναι προφανές ότι η $h_1(x) = f_1(x) + \sum_{k=1}^{\nu} \frac{\alpha_k}{2\pi}\mathfrak{g}_k(x)$ και η $h_2(x) = f_2(x) + \sum_{k=1}^{\nu} \frac{\beta_k}{2\pi}\mathfrak{g}_k(x)$ είναι και οι δύο συνεχείς στο $(0, 2\pi]$. Σχηματίζοντας τη διαφορά $\Delta_n = T_n(f) - \mathcal{C}_n(f)$ λαμβάνουμε ότι:

$$\Delta_n = T_n(f) - \mathcal{C}_n(f) = T_n(h) - \mathcal{C}_n(h) - \left(T_n(\widehat{\mathfrak{g}}) - \mathcal{C}_n(\widehat{\mathfrak{g}}) \right), \qquad (3.12)$$

όπου $h(x) = h_1(x) + \mathrm{i} h_2(x)$. Επειδή η h είναι μια συνεχής συνάρτηση, από το Θεώρημα 3.2.2, ο πίνακας $T_n(h) - \mathcal{C}_n(h)$ έχει κύρια συσσώρευση των ιδιοτιμών στο 0. Πρέπει να εκτιμήσουμε τη συμπεριφορά του $T_n(\widehat{\mathfrak{g}}) - \mathcal{C}_n(\widehat{\mathfrak{g}})$. 'Εχουμε:

$$T_n(\widehat{\mathfrak{g}}) - \mathcal{C}_n(\widehat{\mathfrak{g}}) = \sum_{k=1}^{\nu} \frac{\alpha_k}{2\pi} \left(T_n(\mathfrak{g}_k) - \mathcal{C}_n(\mathfrak{g}_k) \right) + \mathrm{i}\sum_{k=1}^{\nu} \frac{\beta_k}{2\pi} \left(T_n(\mathfrak{g}_k) - \mathcal{C}_n(\mathfrak{g}_k) \right).$$

Χρησιμοποιώντας το Λήμμα 3.2.9, λαμβάνουμε ότι κάθε πίνακας $T_n(\mathfrak{g}_k) - \mathcal{C}_n(\mathfrak{g}_k) = A_{n,k} + B_{n,k}$, όπου $\|A_{n,k}\|_F \leq c_k < \infty$ (c_k : σταθερά ανεξάρτητη της n) και

$\|B_{n,k}\|_F = \mathcal{O}(\log n)$. Επειδή $T_n(\widehat{\mathfrak{g}}) - \mathcal{C}_n(\widehat{\mathfrak{g}})$ είναι ένας γραμμικός συνδυασμός των πινάκων $T_n(\mathfrak{g}_k) - \mathcal{C}_n(\mathfrak{g}_k)$, $k = 1, 2, \ldots, \nu$ ισχύει ότι $\|T_n(\widehat{\mathfrak{g}}) - \mathcal{C}_n(\widehat{\mathfrak{g}})\|_F = \mathcal{O}(\log n)$. Ως εκ τούτου, $\mathcal{O}(\log n)$ ιδιοτιμές του $\Delta_n = T_n(f) - \mathcal{C}_n(f)$ κυμαίνονται εκτός του ορθογωνίου $[-\varepsilon, \varepsilon]^2$, του μιγαδικού επιπέδου. □

Στη συνέχεια θα μελετήσουμε τη συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών του προρρυθμισμένου πίνακα $\mathcal{C}_n^{-1}(f)T_n(f)$, όταν η f είναι κατά τμήματα συνεχής, με πεπερασμένο πλήθος σημείων ασυνέχειας.

Θεώρημα 3.2.10. Έστω f μια μιγαδική συνάρτηση, όπως στο Θεώρημα 3.2.8. Τότε, οι ιδιάζουσες τιμές του προρρυθμισμένου πίνακα $\mathcal{C}_n^{-1}(f)T_n(f)$ συσσωρεύονται γύρω από το 1, με την έννοια της γενικής συσσώρευσης, δηλαδή για κάθε $\varepsilon > 0$, $\mathcal{O}(\log n)$ ιδιάζουσες τιμές κυμαίνονται εκτός του $[1-\varepsilon, 1+\varepsilon]$.

Απόδειξη. Είναι προφανές ότι για τον προρρυθμισμένο πίνακα ισχύει:

$$\begin{aligned}\mathcal{C}_n^{-1}(f)T_n(f) &= \mathcal{C}_n^{-1}(f)(T_n(f) - \mathcal{C}_n(f)) + I_n = \mathcal{C}_n^{-1}(f)\Delta_n + I_n \\ &\overset{(3.12)}{=} \mathcal{C}_n^{-1}(f)\Delta_n(h) - \mathcal{C}_n^{-1}(f)\Delta_n(\widehat{\mathfrak{g}}) + I_n,\end{aligned}$$

όπου $\Delta_n(h) = T_n(h) - \mathcal{C}_n(h)$ και $\Delta_n(\widehat{\mathfrak{g}}) = T_n(\widehat{\mathfrak{g}}) - \mathcal{C}_n(\widehat{\mathfrak{g}})$. Από την απόδειξη του Θεωρήματος 3.2.8 έχουμε ότι $\Delta_n(\widehat{\mathfrak{g}}) = A_n + B_n$, όπου $\|A_n\|_F < \infty$, ανεξάρτητη της διάστασης n και $\|B_n\|_F = \mathcal{O}(\log n)$. Επομένως, η παραπάνω σχέση γράφεται ως:

$$\mathcal{C}_n^{-1}(f)T_n(f) - I_n = \mathcal{C}_n^{-1}(f)\left(\Delta_n(h) - A_n\right) - \mathcal{C}_n^{-1}(f)B_n.$$

Υποθέτουμε ότι $A = \mathcal{C}_n^{-1}(f)\left(\Delta_n(h) - A_n\right)$ και $B = -\mathcal{C}_n^{-1}(f)B_n$. Γνωρίζουμε ότι οι ιδιάζουσες τιμές του A συσσωρεύονται κατά κύριο τρόπο γύρω από το 0 και ο B_n, ο οποίος είναι όμοιος με τον $\mathcal{H}_n$, έχει γενική συσσώρευση των ιδιαζουσών τιμών γύρω από το 0. Έστω ότι για κάποιο συγκεκριμένο $\varepsilon > 0$, k ιδιάζουσες τιμές είναι μεγαλύτερες από ε. Από την ανισότητα του Weyl [39] έχουμε ότι:

$$\sigma_{j+k}(A+B) \leq \sigma_{k+1}(A) + \sigma_j(B) \leq \varepsilon + \sigma_j(B),\ 1 \leq j+k \leq n.$$

Συνεπώς, $\mathcal{O}(\log n)$ ιδιάζουσες τιμές του προρρυθμισμένου πίνακα κυμαίνονται εκτός του $[1-\varepsilon, 1+\varepsilon]$ και η απόδειξη ολοκληρώθηκε. □

Λήμμα 3.2.11. Έστω $\{\mathcal{A}_n\}$ και $\{\mathcal{B}_n\}$ δύο ακολουθίες πινάκων, με $\mathcal{A}_n$ να είναι αντιστρέψιμος για κάθε $n \in \mathbb{N}$. Έστω επίσης ότι οι $\{\mathcal{A}_n\}$ και $\{\mathcal{A}_n^{-1}\}$ έχουν φραγμένη νόρμα $\|\cdot\|_2$, το οποίο σημαίνει ότι υπάρχουν θετικές σταθερές c και d τέτοιες ώστε $\|\mathcal{A}_n\|_2 \leq c$ και $\|\mathcal{A}_n^{-1}\|_2 \leq d$, για οποιονδήποτε $n \in \mathbb{N}$. Τότε, για τη νόρμα Frobenius ισχύει ότι, $\|\mathcal{A}_n\mathcal{B}_n\|_F \sim \|\mathcal{B}_n\|_F$, που σημαίνει ότι υπάρχουν $\alpha, \beta > 0$ έτσι ώστε, $\alpha\|\mathcal{B}_n\|_F \leq \|\mathcal{A}_n\mathcal{B}_n\|_F \leq \beta\|\mathcal{B}_n\|_F$.

Απο404δειξη. Γνωρίζουμε ότι $\|\mathcal{A}_n\mathcal{B}_n\|_F = \|\mathcal{B}_n\mathcal{A}_n\|_F = \left(\text{tr}\left(\mathcal{A}_n^H\mathcal{B}_n^H\mathcal{B}_n\mathcal{A}_n\right)\right)^{\frac{1}{2}}$.

Έστω λ_j, $j = 1, 2, \ldots, n$ οι ιδιοτιμές του $\mathcal{A}_n^H\mathcal{B}_n^H\mathcal{B}_n\mathcal{A}_n$ για κάποιο σταθερό n και μ_j οι ιδιοτιμές του $\mathcal{B}_n^H\mathcal{B}_n$, ταξινομημένες σε μη-αύξουσα σειρά. Έστω επίσης ότι με w_j, $j = 1, 2, \ldots, n$ συμβολίζουμε τα ιδιοδιανύσματα του $\mathcal{B}_n^H\mathcal{B}_n$ και $W_j = \text{span}\{w_j, w_{j+1}, \ldots, w_n\}$, $\widetilde{W}_j = \text{span}\{w_1, w_2, \ldots, w_j\}$. Χρησιμοποιούμε το min-max θεώρημα των Courant-Fisher, για να συσχετίσουμε τις ιδιοτιμές λ_j με τις μ_j.

$$
\begin{aligned}
\lambda_j &= \min_{\mathcal{V}:\dim(\mathcal{V})=n-j+1} \max_{x\in\mathcal{V}} \frac{x^H\mathcal{A}_n^H\mathcal{B}_n^H\mathcal{B}_n\mathcal{A}_n x}{x^H x} \leq \max_{x\in\mathcal{A}_n^{-1}W_j} \frac{x^H\mathcal{A}_n^H\mathcal{B}_n^H\mathcal{B}_n\mathcal{A}_n x}{x^H x} \\
&= \max_{y\in W_j} \frac{y^H\mathcal{B}_n^H\mathcal{B}_n y}{y^H\mathcal{A}_n^{-H}\mathcal{A}_n^{-1}y} = \max_{y\in W_j} \frac{y^H\mathcal{B}_n^H\mathcal{B}_n y}{y^H y}\cdot\frac{y^H y}{y^H\mathcal{A}_n^{-H}\mathcal{A}_n^{-1}y} \\
&\leq \max_{y\in W_j} \frac{y^H\mathcal{B}_n^H\mathcal{B}_n y}{y^H y}\cdot\max_{y\in W_j}\frac{y^H y}{y^H\mathcal{A}_n^{-H}\mathcal{A}_n^{-1}y} \\
&= \min_{\mathcal{V}:\dim(\mathcal{V})=n-j+1} \max_{y\in\mathcal{V}} \frac{y^H\mathcal{B}_n^H\mathcal{B}_n y}{y^H y}\cdot\bar{c}_j = \bar{c}_j\mu_j, \text{ όπου } \frac{1}{d^2} \leq \bar{c}_j \leq c^2.
\end{aligned}
$$

Από την άλλη,

$$
\begin{aligned}
\lambda_j &= \max_{\mathcal{V}:\dim(\mathcal{V})=j} \min_{x\in\mathcal{V}} \frac{x^H\mathcal{A}_n^H\mathcal{B}_n^H\mathcal{B}_n\mathcal{A}_n x}{x^H x} \geq \min_{x\in\mathcal{A}_n^{-1}\widetilde{W}_j} \frac{x^H\mathcal{A}_n^H\mathcal{B}_n^H\mathcal{B}_n\mathcal{A}_n x}{x^H x} \\
&= \min_{y\in\widetilde{W}_j} \frac{y^H\mathcal{B}_n^H\mathcal{B}_n y}{y^H\mathcal{A}_n^{-H}\mathcal{A}_n^{-1}y} \geq \min_{y\in\widetilde{W}_j} \frac{y^H\mathcal{B}_n^H\mathcal{B}_n y}{y^H y}\cdot\min_{y\in\widetilde{W}_j}\frac{y^H y}{y^H\mathcal{A}_n^{-H}\mathcal{A}_n^{-1}y} \\
&= \max_{\mathcal{V}:\dim(\mathcal{V})=j} \min_{y\in\mathcal{V}} \frac{y^H\mathcal{B}_n^H\mathcal{B}_n y}{y^H y}\cdot\underline{c}_j = \underline{c}_j\mu_j, \text{ όπου } \frac{1}{d^2} \leq \underline{c}_j \leq c^2.
\end{aligned}
$$

Επομένως, από το θεώρημα ενδιαμέσων τιμών υπάρχουν $\widetilde{c}_j$ ($\underline{c}_j \leq \widetilde{c}_j \leq \bar{c}_j$), έτσι ώστε $\lambda_j = \widetilde{c}_j\mu_j$, $j = 1, 2, \ldots, n$.

Παίρνοντας τη νόρμα Frobenius, λαμβάνουμε:

$$
\begin{aligned}
\|\mathcal{A}_n\mathcal{B}_n\|_F &= \left(\text{tr}\left(\mathcal{A}_n^H\mathcal{B}_n^H\mathcal{B}_n\mathcal{A}_n\right)\right)^{\frac{1}{2}} = \left(\sum_{j=1}^n \lambda_j\right)^{\frac{1}{2}} = \left(\sum_{j=1}^n \widetilde{c}_j\mu_j\right)^{\frac{1}{2}} \\
&= \left(\widetilde{c}\sum_{j=1}^n \mu_j\right)^{\frac{1}{2}} = \sqrt{\widetilde{c}}\left(\sum_{j=1}^n \mu_j\right)^{\frac{1}{2}} = c'\left(\text{tr}\left(\mathcal{B}_n^H\mathcal{B}_n\right)\right)^{\frac{1}{2}} = c'\|\mathcal{B}_n\|_F,
\end{aligned}
$$

όπου, από το θεώρημα ενδιαμέσων τιμών, $\frac{1}{d^2} \leq \widetilde{c} \leq c^2$ και $\frac{1}{d} \leq c' \leq c$. Έτσι, η απόδειξη ολοκληρώθηκε. □

Θεώρημα 3.2.12. Έστω f μια μιγαδική συνάρτηση, όπως περιγράφηκε στο Θεώρημα 3.2.8. Τότε, οι ιδιοτιμές του $\mathcal{C}_n^{-1}(f)T_n(f)$ συσσωρεύονται γύρω από το $(1,0)$, με την έννοια της γενικής συσσώρευσης, που σημαίνει ότι για κάθε $\varepsilon > 0$, $\mathcal{O}(\log n)$ ιδιοτιμές κυμαίνονται εκτός του ορθογωνίου $[1-\varepsilon, 1+\varepsilon] \times [-\varepsilon, \varepsilon]$ του μιγαδικού επιπέδου.

Απόδειξη. Για να μελετήσουμε το φάσμα των ιδιοτιμών διαχωρίζουμε τον προρρυθμισμένο πίνακα στο Ερμιτιανό και αντι-Ερμιτιανό του μέρος. Το Ερμιτιανό μέρος γράφεται ως:

$$\frac{1}{2}\left[\mathcal{C}_n^{-1}(f)T_n(f) + T_n(\bar{f})\mathcal{C}_n^{-1}(\bar{f})\right] = \frac{1}{2}\left[\mathcal{C}_n^{-1}(f)\Delta_n + \Delta_n^H\mathcal{C}_n^{-1}(\bar{f})\right] + I_n,$$

κι έτσι για να αποδείξουμε ότι έχει γενική συσσώρευση γύρω από το 1 με $\mathcal{O}(\log n)$ ιδιοτιμές εκτός του διαστήματος συσσώρευσης, πρέπει να αποδείξουμε ότι ο Ερμιτιανός πίνακας $\mathcal{C}_n^{-1}(f)\Delta_n + \Delta_n^H\mathcal{C}_n^{-1}(\bar{f})$ έχει γενική συσσώρευση γύρω από το 0 με $\mathcal{O}(\log n)$ ιδιοτιμές εκτός του διαστήματος συσσώρευσης. Επειδή η f δεν έχει ρίζες, έχουμε ότι $c \leq \|\mathcal{C}_n^{-1}(f)\|_2 \leq C$, όπου c, C είναι θετικές σταθερές. Στο Θεώρημα 3.2.8 αποδείξαμε ότι $\|\Delta_n\|_F = \mathcal{O}(\log n)$. Χρησιμοποιώντας το Λήμμα 3.2.11 με $\mathcal{A}_n = \mathcal{C}_n^{-1}(f)$ και $\mathcal{B}_n = \Delta_n$, λαμβάνουμε ότι $\|\mathcal{C}_n^{-1}(f)\Delta_n\|_F = \mathcal{O}(\log n)$. Παίρνοντας τη νόρμα Frobenius έχουμε

$$\|\mathcal{C}_n^{-1}(f)\Delta_n + \Delta_n^H\mathcal{C}_n^{-1}(\bar{f})\|_F \leq \|\mathcal{C}_n^{-1}(f)\Delta_n\|_F + \|\Delta_n^H\mathcal{C}_n^{-1}(\bar{f})\|_F = \mathcal{O}(\log n).$$

Επομένως, για κάποιο $\varepsilon > 0$, το πολύ $\mathcal{O}(\log n)$ ιδιοτιμές του $\mathcal{C}_n^{-1}(f)\Delta_n + \Delta_n^H\mathcal{C}_n^{-1}(\bar{f})$ κυμαίνονται εκτός του διαστήματος $[-\varepsilon, \varepsilon]$, το οποίο ισοδύναμα μας δίνει τη γενική συσσώρευση του πραγματικού μέρους των ιδιοτιμών (του προρρυθμισμένου συστήματος) στο $[1-\varepsilon, 1+\varepsilon]$.

Θεωρούμε το αντι-Ερμιτιανό μέρος του προρρυθμισμένου πίνακα:

$$\frac{1}{2}\left[\mathcal{C}_n^{-1}(f)T_n(f) - T_n(\bar{f})\mathcal{C}_n^{-1}(\bar{f})\right] = \frac{1}{2}\left[\mathcal{C}_n^{-1}(f)\Delta_n - \Delta_n^H\mathcal{C}_n^{-1}(\bar{f})\right].$$

Ομοίως έχουμε ότι

$$\|\mathcal{C}_n^{-1}(f)\Delta_n - \Delta_n^H\mathcal{C}_n^{-1}(\bar{f})\|_F \leq \|\mathcal{C}_n^{-1}(f)\Delta_n\|_F + \|\Delta_n^H\mathcal{C}_n^{-1}(\bar{f})\|_F = \mathcal{O}(\log n).$$

Η παραπάνω σχέση μας δίνει τη γενική συσσώρευση του φανταστικού μέρους των ιδιοτιμών στο $[-\varepsilon, \varepsilon]$.

Από την άλλη:

$$\mathcal{O}(\log n) = \|\mathcal{C}_n^{-1}(f)\Delta_n\|_F \le \frac{1}{2}\|\mathcal{C}_n^{-1}(f)\Delta_n + \Delta_n^H \mathcal{C}_n^{-1}(\bar{f})\|_F$$
$$+ \frac{1}{2}\|\mathcal{C}_n^{-1}(f)\Delta_n - \Delta_n^H \mathcal{C}_n^{-1}(\bar{f})\|_F.$$

Αυτό σημαίνει ότι είτε το Ερμιτιανό, είτε το αντι-Ερμιτιανό μέρος (είτε και τα δύο) έχουν νόρμα Frobenius της τάξης $\mathcal{O}(\log n)$. Λαμβάνοντας υπόψη τις ιδιότητες ισοδυναμίας των όρων $\|\mathcal{C}_n^{-1}(f)\Delta_n\|_F$, $\|\Delta_n\|_F$ και $\|\mathcal{H}_{n/2}\|_F$, λαμβάνουμε ότι $\mathcal{O}(\log n)$ ιδιοτιμές κυμαίνονται εκτός του ορθογωνίου $[1-\varepsilon, 1+\varepsilon] \times [-\varepsilon, \varepsilon]$ του μιγαδικού επιπέδου κι έτσι η απόδειξη ολοκληρώθηκε. □

Παρατήρηση. Λόγω της περιοδικότητας της συνάρτησης f, όλα τα παραπάνω θεωρήματα ισχύουν και στο $(-\pi, \pi]$.

Συνεχίζουμε με τη μελέτη της περίπτωσης όπου η γεννήτρια συνάρτηση f έχει πεπερασμένα σημεία ριζών, καθώς επίσης και πεπερασμένα σημεία ασυνέχειας. Αρχικά, μελετάμε τη συμπεριφορά των ιδιαζουσών τιμών.

Θεώρημα 3.2.13. Έστω $f = f_1 + \mathrm{i}f_2$, όπου f_1 είναι άρτια και f_2 περιττή, 2π-περιοδική, έχοντας k ρίζες $x_1, x_2, \ldots, x_k$ στο $(-\pi, \pi]$ και ν σημεία ασυνέχειας $\xi_1, \xi_2, \ldots, \xi_\nu$, επίσης στο $(-\pi, \pi]$, με αντίστοιχα εύρη ασυνέχειας

$$\alpha_j = \lim_{x \to \xi_j^+} f(x) - \lim_{x \to \xi_j^-} f(x),$$

και υποθέτουμε ότι τα σημεία x_j είναι διαφορετικά από τα ξ_j. Έστω επίσης g, το τριγωνομετρικό πολυώνυμο τέτοιο ώστε η $\frac{f}{g}$ να μην έχει ρίζες στο $(-\pi, \pi]$. Τότε, για κάθε $\varepsilon > 0$, το διάστημα $[1-\varepsilon, 1+\varepsilon]$ αποτελεί σύνολο γενικής συσσώρευσης των ιδιαζουσών τιμών του προρρυθμισμένου πίνακα $\mathcal{C}_n^{-1}\left(\frac{f}{g}\right) T_n^{-1}(g) T_n(f)$, με $\mathcal{O}(\log n)$ ιδιάζουσες τιμές εκτός του διαστήματος.

Απο404δειξη. Ακολουθούμε την απόδειξη του Θεωρήματος 3.2.5 προκειμένου να λάβουμε το αντίστοιχο αποτέλεσμα για τον πίνακα των κανονικών εξισώσεων,

$$\mathcal{C}_n^{-1}\left(\frac{f}{g}\right) T_n^{-1}(g) T_n(f) T_n(\bar{f}) T_n^{-1}(\bar{g}) \mathcal{C}_n^{-1}\left(\frac{\bar{f}}{\bar{g}}\right) =$$
$$\mathcal{C}_n^{-1}\left(\frac{f}{g}\right) T_n\left(\frac{f}{g}\right) T_n\left(\frac{\bar{f}}{\bar{g}}\right) \mathcal{C}_n^{-1}\left(\frac{\bar{f}}{\bar{g}}\right) + L,$$

όπου L είναι πίνακας χαμηλής βαθμίδας, το πολύ ίσης με $4d$ (d είναι ο βαθμός του τριγωνομετρικού πολυωνύμου g). Επομένως, οι ιδιάζουσες τιμές του προρρυθμισμένου πίνακα $\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n^{-1}(g)T_n(f)$ συμπεριφέρονται όπως αυτές του $\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n^{-1}\left(\frac{f}{g}\right)$ με τη διαφορά $4d$ επιπλέον ιδιαζουσών τιμών εκτός του διαστήματος συσσώρευσης, που προκύπτουν από τον L.

Η μόνη διαφορά από το Θεώρημα 3.2.5 εντοπίζεται στο ότι η συνάρτηση $\frac{f}{g}$ παρουσιάζει ασυνέχεια στα σημεία ξ_j, $j = 1, 2, \dots, \nu$ με πεπερασμένο εύρος ασυνέχειας $\beta_j = \frac{\alpha_j}{g(\xi_j)}$, $j = 1, 2, \dots, \nu$. Στη συνέχεια, εφαρμόζοντας το Θεώρημα 3.2.10 για την $\frac{f}{g}$ καταλήγουμε στο επιθυμητό αποτέλεσμα. □

Η συμπεριφορά των ιδιοτιμών δίνεται στο επόμενο θεώρημα.

Θεώρημα 3.2.14. Έστω f μια μιγαδική συνάρτηση και g τριγωνομετρικό πολυώνυμο, όπως περιγράφηκε στο Θεώρημα 3.2.13. Τότε, οι ιδιοτιμές του πίνακα $\mathcal{C}_n^{-1}(\frac{f}{g})T_n^{-1}(g)T_n(f)$ συσσωρεύονται γύρω από το $(1, 0)$, με την έννοια της γενικής συσσώρευσης. Ισοδύναμα, για κάθε $\varepsilon > 0$, $\mathcal{O}(\log n)$ ιδιοτιμές κυμαίνονται εκτός του ορθογωνίου $[1-\varepsilon, 1+\varepsilon] \times [-\varepsilon, \varepsilon]$ του μιγαδικού επιπέδου.

Απο404δειξη. Για το Ερμιτιανό μέρος του $\mathcal{C}_n^{-1}(\frac{f}{g})T_n^{-1}(g)T_n(f)$ ισχύει:

$$\begin{aligned}
H &= \frac{1}{2}\left[\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n^{-1}(g)T_n(f) + T_n(\bar{f})T_n^{-1}(\bar{g})\mathcal{C}_n^{-1}\left(\frac{\bar{f}}{\bar{g}}\right)\right] \\
&= \frac{1}{2}\left[\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n^{-1}(g)\left(T_n(g)T_n\left(\frac{f}{g}\right) + L_1\right)\right] \\
&\quad + \frac{1}{2}\left[\left(T_n\left(\frac{\bar{f}}{\bar{g}}\right)T_n(\bar{g}) + L_1^H\right)T_n^{-1}(\bar{g})\mathcal{C}_n^{-1}\left(\frac{\bar{f}}{\bar{g}}\right)\right] \\
&= \frac{1}{2}\left[\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n\left(\frac{f}{g}\right) + L_2 + T_n\left(\frac{\bar{f}}{\bar{g}}\right)\mathcal{C}_n^{-1}\left(\frac{\bar{f}}{\bar{g}}\right) + L_2^H\right] \\
&= \frac{1}{2}\left[\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n\left(\frac{f}{g}\right) + T_n\left(\frac{\bar{f}}{\bar{g}}\right)\mathcal{C}_n^{-1}\left(\frac{\bar{f}}{\bar{g}}\right)\right] + L_3,
\end{aligned}$$

όπου L_3 είναι πίνακας χαμηλής βαθμίδας, το πολύ ίσης με $4d$ (d είναι ο βαθμός του τριγωνομετρικού πολυωνύμου g). Παρατηρούμε ότι ο H, διαφέρει από το Ερμιτιανό μέρος του πίνακα $\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n\left(\frac{f}{g}\right)$, μόνο κατά τον L_3.

Επειδή ο $\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)$ έχει φραγμένη νόρμα $\|\cdot\|_2$ και η $\frac{f}{g}$ έχει σημεία ασυνέχειας, μπορούμε να χρησιμοποιήσουμε το Λήμμα 3.2.11 και το Θεώρημα 3.2.12 για να λάβουμε ότι το Ερμιτιανό μέρος του $\mathcal{C}_n^{-1}\left(\frac{f}{g}\right)T_n\left(\frac{f}{g}\right)$ έχει γενική συσσώρευση των ιδιοτιμών, γύρω από το 1 με $\mathcal{O}(\log n)$ ιδιοτιμές εκτός του διαστήματος συσσώρευσης.

Εργαζόμενοι ανάλογα για το αντι-Ερμιτιανό μέρος λαμβάνουμε τη γενική συσσώρευση, γύρω από το 0 με $\mathcal{O}(\log n)$ ιδιοτιμές εκτός του διαστήματος συσσώρευσης και η απόδειξη ολοκληρώθηκε. □

Παρατήρηση. Αποδείξαμε τη γενική συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών, του προρρυθμισμένου πίνακα, όταν η γεννήτρια συνάρτηση f έχει σημεία ασυνέχειας. Αυτό, εκ πρώτης όψεως είναι ένα αρνητικό αποτέλεσμα σε σχέση με τη συνεχή περίπτωση, όπου αποδείχθηκε κύρια συσσώρευση. Ωστόσο, η συνάρτηση του λογαρίθμου, η οποία χαρακτηρίζει τη γενική συσσώρευση, τείνει πολύ αργά προς το άπειρο και στα αριθμητικά αποτελέσματα η γενική συσσώρευση δε γίνεται αισθητή. Αντιθέτως, η συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών μοιάζει να είναι κύρια.

3.3 Αριθμητικά αποτελέσματα

Σε αυτή την ενότητα δίνουμε διάφορα αριθμητικά παραδείγματα, τα οποία έρχονται σε συμφωνία με τα θεωρητικά αποτελέσματα τα οποία αποδείξαμε, σχετικά με την προτεινόμενη τεχνική προρρύθμισης. Στα αριθμητικά πειράματα το διάνυσμα b, του δεξιού μέλους του αρχικού συστήματος, επιλέχθηκε και πάλι έτσι ώστε η λύση του συστήματος να είναι το διάνυσμα, του οποίου όλες οι συνιστώσες είναι ίσες με μονάδα, δηλαδή το $(1\ 1\ \cdots\ 1)^T$. Ως αρχική προσέγγιση επιλέξαμε το μηδενικό διάνυσμα και ως κριτήριο τερματισμού: $\frac{\|r(k)\|_2}{\|r(0)\|_2} \leq 10^{-6}$, όπου $r(k)$ συμβολίζει, όπως και στο προηγούμενο κεφάλαιο, το διάνυσμα υπόλοιπο της k-οστής επανάληψης και $r(0) = b$.

Στους πίνακες δίνουμε τον αριθμό επαναλήψεων των μεθόδων PGMRES και PCGN, έως ότου να έχουμε την επιθυμητή σύγκλιση στη λύση του συστήματος. Χρησιμοποιούμε τον εξής συμβολισμό: n είναι η διάσταση του συστήματος, με I_n δηλώνουμε ότι δε χρησιμοποιήθηκε κανένας προρρυθμιστής, το $\mathcal{C}_n$ συμβολίζει τον προτεινόμενο προρρυθμιστή, ενώ το $\mathcal{T}_n$ συμβολίζει τον βέλτιστο κυκλοειδή προρρυθμιστή [16, 14, 89]. Με χρήση παρόμοιου συμβολισμού, όταν γίνεται άρση των ριζών της γεννήτριας συνάρτησης, με $B\mathcal{C}_n$ και $B\mathcal{T}_n$ θα δηλώνουμε τον αντίστοιχο ""ταινιωτο-επί-κυκλοειδή"" (Band-times-Circulant) προρρυθμιστή.

Παράδειγμα 3.3.1. Ως πρώτο παράδειγμα αυτού του κεφαλαίου επιλέγουμε τη 2π-περιοδική και συνεχή συνάρτηση που είδαμε και στο Παράδειγμα 2.3.1, $\mathfrak{f}_1(x) =$

n	PGMRES			PCGN		
	I_n	$\mathcal{C}_n$	$\mathcal{T}_n$	I_n	$\mathcal{C}_n$	$\mathcal{T}_n$
256	31	5	5	72	6	6
512	30	5	5	74	6	6
1024	29	5	5	73	6	6
2048	29	4	4	72	6	6

Πίνακας 3.1: Επαναλήψεις ($\mathfrak{f}_1$).

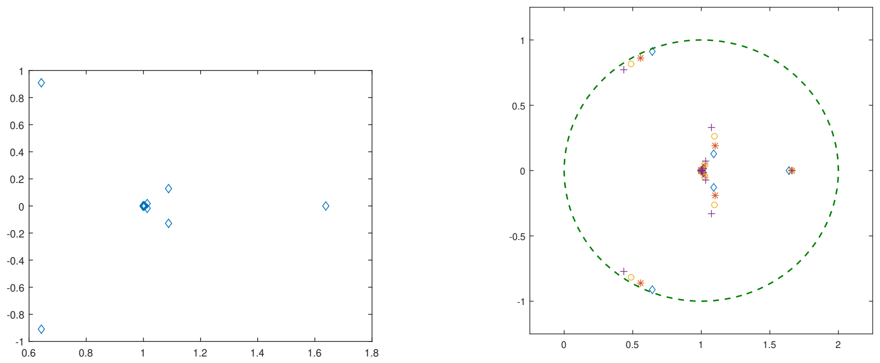
$x^2 + 1 + \mathrm{i}\mathfrak{h}_1(x)$, όπου:

$$\mathfrak{h}_1(x) = \begin{cases} -\pi - x\,, & -\pi \leq x < -\frac{\pi}{2} \\ x\,, & -\frac{\pi}{2} \leq x < \frac{\pi}{2} \\ \pi - x\,, & \frac{\pi}{2} \leq x \leq \pi \end{cases} .$$

Προφανώς, το πραγματικό μέρος της $\mathfrak{f}_1$ είναι μια θετική συνάρτηση στο $(-\pi, \pi]$. Συνεπώς, για την επίλυση του συστήματος χρησιμοποιούμε τους κυκλοειδείς προρρυθμιστές $\mathcal{C}_n$ και $\mathcal{T}_n$. Ο Πίνακας 3.1 δείχνει τον αριθμό επαναλήψεων, μέχρι τη σύγκλιση των μεθόδων PGMRES και PCGN. Παρατηρούμε ότι οι προρρυθμιστές συγκλινουν, στη λύση του συστήματος, με τον ίδιο αριθμό επαναλήψεων. 'Οπως θα δούμε στα παραδείγματα που ακολουθούν, τόσο ο $\mathcal{C}_n$, όσο και ο $\mathcal{T}_n$ είναι αποτελεσματικοί, όμως σημειώνουμε ότι στα περισσότερα εξ αυτών, ο $\mathcal{C}_n$ συγκλίνει στη λύση με λιγότερες επαναλήψεις σε σύγκριση με τον $\mathcal{T}_n$. Επιπλέον, στο Σχήμα 3.4 παρατηρούμε ότι η συσσώρευση των ιδιοτιμών είναι πολύ πιο πυκνή όταν χρησιμοποιούμε τον $B\mathcal{C}_n$, αντί του $B\mathcal{T}_n$.

Παράδειγμα 3.3.2. Σε αυτό το παράδειγμα η γεννήτρια συνάρτηση του πίνακα Toeplitz είναι η $\mathfrak{f}_2(x) = x^2 + \mathrm{i}x^3$, βλ. επίσης Παράδειγμα 2.3.2. Αυτή έχει μια ρίζα στο 0 και το φανταστικό της μέρος έχει σημείο ασυνέχειας στο π. Η πολλαπλότητα της ρίζας εξαρτάται από το πραγματικό μέρος της $\mathfrak{f}_2$, γεγονός το οποίο μας οδηγεί στην επιλογή $g(x) = 2 - 2\cos(x)$, για την άρση των ριζών. Η αναγκαιότητα, καθώς και τα πλεονεκτήματα της προρρύθμισης είναι εξώφθαλμα, όπως φαίνεται στον Πίνακα 3.2.

Η αποτελεσματικότητα της προτεινόμενης τεχνικής προρρύθμισης φαίνεται στον Πίνακα 3.2. Εκεί παρατηρούμε μια ελαφρώς καλύτερη συμπεριφορά του $B\mathcal{C}_n$ σε σχέση με τον ""ταινιωτό-επί-βέλτιστο κυκλοειδή"" (Band-times-optimal Circulant), όταν παίρνουμε τη λύση μέσω της μεθόδου PCGN (βλ. επίσης [85]).



(α') Ιδιοτιμές όταν $n = 256$. (β') Ιδιοτιμές για διάφορες διαστάσεις.

Σχήμα 3.2: Ιδιοτιμές ($\mathfrak{f}_2$).

n	PGMRES				PCGN			
	I_n	$\mathcal{T}_n$	$B\mathcal{C}_n$	$B\mathcal{T}_n$	I_n	$\mathcal{T}_n$	$B\mathcal{C}_n$	$B\mathcal{T}_n$
256	256	22	7	7	-	61	13	14
512	>500	28	7	7	-	90	14	17
1024	>500	36	7	7	-	140	15	17
2048	>500	39	7	8	-	273	16	19

Πίνακας 3.2: Επαναλήψεις ($\mathfrak{f}_2$).

Στο Σχήμα 3.2β' δίνουμε τη συσσώρευση των ιδιοτιμών του προρρυθμισμένου πίνακα $\mathcal{C}_n^{-1}\left(\frac{\mathfrak{f}_2}{g}\right)T_n^{-1}(g)T_n(\mathfrak{f}_2)$ από τη διάσταση $n = 256$ (μπλε διαμάντια), έως $n = 2048$ (μωβ σταυροί). Με πορτοκαλί αστέρια και κίτρινους κύκλους συμβολίζουμε τις αντίστοιχες ιδιοτιμές για $n = 512$ και $n = 1024$, αντίστοιχα. Ο κύκλος με την πράσινη διακεκομμένη γραμμή, έχει κέντρο το $(1, 0)$ και ακτίνα ίση με 1. Παρατηρούμε ότι όλες οι ιδιοτιμές είναι εντός του κύκλου.

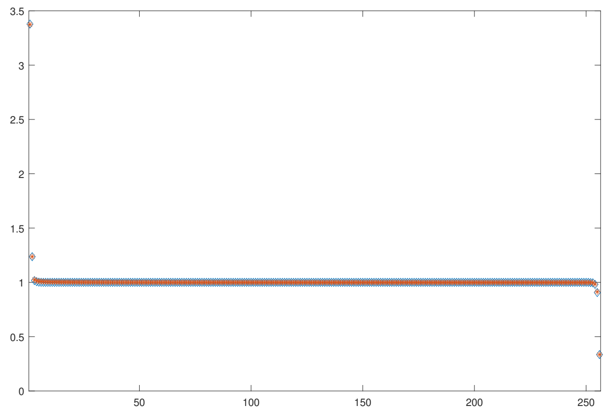
Παράδειγμα 3.3.3. Σε αυτό το παράδειγμα η γεννήτρια συνάρτηση του πίνακα Toeplitz είναι η $\mathfrak{f}_3(x) = x^2 + \mathrm{i}x$, που μελετήθηκε με διαφορετική τεχνική προρρύθμισης στο Παράδειγμα 2.3.3. Αυτή, όπως και στο προηγούμενο παράδειγμα, έχει ρίζα στο 0 και το φανταστικό της μέρος παρουσιάζει ασυνέχεια στο π. Η διαφορά εντοπίζεται στο ότι η πολλαπλότητα της ρίζας εξαρτάται από το φανταστικό μέρος και όχι από το πραγματικό. Η άρση των ριζών είναι απαραίτητη για να επιτύχουμε ταχύτερη σύγκλιση στη λύση του συστήματος. Ο ταινιωτός πίνακας Toeplitz θα έχει ως

n	PGMRES				PCGN			
	I_n	$\mathcal{T}_n$	$B\mathcal{C}_n$	$B\mathcal{T}_n$	I_n	$\mathcal{T}_n$	$B\mathcal{C}_n$	$B\mathcal{T}_n$
256	256	9	5	5	-	10	7	7
512	$>$500	9	5	5	-	11	7	7
1024	$>$500	9	5	5	-	11	8	7
2048	$>$500	9	5	5	-	11	8	8

Πίνακας 3.3: Επαναλήψεις ($\mathfrak{f}_3$).

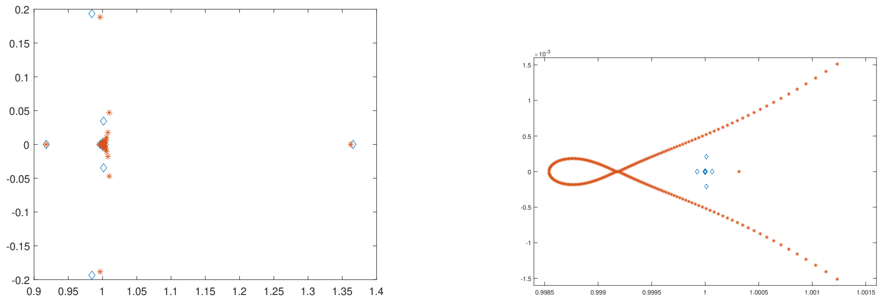
γεννήτρια συνάρτηση την $g(x) = 2 - 2\cos(x) + \mathrm{i}\sin(x)$ (βλ. υποενότητα 2.2.2).

Ο Πίνακας 3.3 δείχνει τον αριθμό επαναλήψεων για τις μεθόδους PGMRES και PCGN. Στα Σχήματα 3.3 και 3.4 παρουσιάζουμε τη συσσώρευση ιδιαζουσών τιμών και των ιδιοτιμών, αντίστοιχα, όταν $n = 256$. Πιο συγκεκριμένα, αυτές που αφορούν στον προρρυθμιστή $B\mathcal{C}_n$ σημειώνονται με μπλε διαμάντια, ενώ αυτές που αφορούν στον $B\mathcal{T}_n$, με πορτοκαλί αστέρια. Σε αυτό το παράδειγμα είναι ολοφάνερη η αναγκαιότητα προρρύθμισης, διότι χωρίς αυτή, η λύση του συστήματος δίνεται σε επαναλήψεις ίσες με τη διάσταση n.



Σχήμα 3.3: Ιδιάζουσες τιμές ($\mathfrak{f}_3$).

Θα θέλαμε να σημειώσουμε ότι οι αριθμοί επαναλήψεων μεταξύ του $\mathcal{T}_n$ και $B\mathcal{T}_n$ δε διαφέρουν σε μεγάλο βαθμό, γεγονός το οποίο δεν ισχύει στο προηγούμενο παράδειγμα. Εκεί, η διαφορά είναι αισθητή και αυτό οφείλεται στην τάξη της ρίζας, η οποία ήταν ίση με 2, ενώ στο παρόν παράδειγμα ίση με 1. ‘Οσο μεγαλύτερη δηλαδή είναι η τάξη της ρίζας, τόσο μεγαλύτερη είναι και η διαφορά στην αποτελεσματικό-



(α') Ιδιοτιμές. (β') Ιδιοτιμές κοντά στο $(1,0)$.

Σχήμα 3.4: Ιδιοτιμές ($\mathfrak{f}_3$).

τητα των προρρυθμιστών $\mathcal{T}_n$ και $B\mathcal{T}_n$.

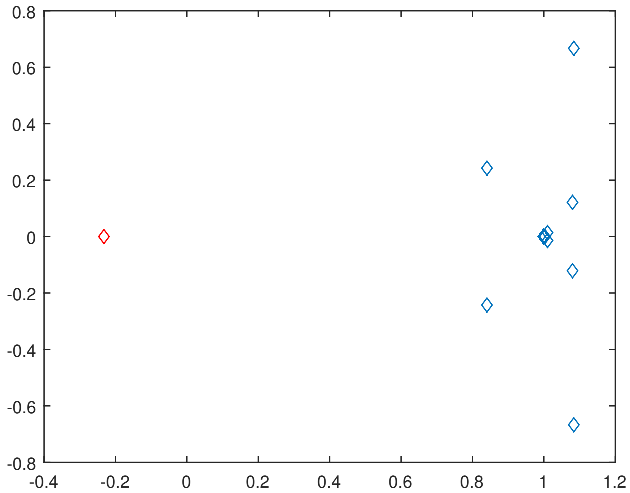
Αν και φαίνεται ότι οι ιδιάζουσες τιμές των προρρυθμισμένων συστημάτων ταυτίζονται, στο Σχήμα 3.4 παρατηρούμε ότι η χρήση του $B\mathcal{C}_n$, αντί του $B\mathcal{T}_n$ οδηγεί σε ένα πιο πυκνό σύνολο συσσώρευσης των ιδιοτιμών, σε μια μικρή περιοχή κοντά στο $(1,0)$.

Παράδειγμα 3.3.4. Δίνουμε ένα παράδειγμα, στο οποίο η γεννήτρια συνάρτηση του πίνακα Toeplitz είναι η $\mathfrak{f}_7(x) = x^2 - 1 + \mathrm{i}x^3$. Αυτή δεν έχει ρίζες στο $(-\pi,\pi]$, αλλά το πραγματικό της μέρος παίρνει τόσο θετικές, όσο και αρνητικές τιμές. Προφανώς, το φανταστικό της μέρος έχει ασυνέχεια στο π. Χρησιμοποιούμε τους κυκλοειδείς προρρυθμιστές $\mathcal{C}_n$ και $\mathcal{T}_n$.

n	PGMRES			PCGN		
	I_n	$\mathcal{C}_n$	$\mathcal{T}_n$	I_n	$\mathcal{C}_n$	$\mathcal{T}_n$
256	160	9	9	-	11	13
512	248	9	9	362	12	13
1024	326	9	9	397	12	13
2048	370	9	10	415	12	13

Πίνακας 3.4: Επαναλήψεις ($\mathfrak{f}_7$).

Οι αριθμοί επαναλήψεων φαίνονται στον Πίνακα 3.4. Στο Σχήμα 3.5 δίνουμε τη συσσώρευση των ιδιοτιμών του $\mathcal{C}_n^{-1}(f)T_n(\mathfrak{f}_7)$, όταν $n = 256$. Βλέπουμε ότι και πάλι αυτές συσσωρεύονται γύρω από το $(1,0)$. Σχολιάζουμε ότι αν και υπάρχει μια αρνητική ιδιοτιμή, η οποία συμβολίζεται ως κόκκινο διαμάντι, η μέθοδος PGMRES είναι αποτελεσματική. Γνωρίζουμε ότι αν μια ιδιοτιμή του προρρυθμισμένου πίνα-



Σχήμα 3.5: Ιδιοτιμές ($\mathfrak{f}_7$).

κα δεν ανήκει στον δίσκο με κέντρο το $(1, 0)$ και ακτίνα ίση με 1, ο αριθμός επαναλήψεων αυξάνεται κατά 1 [31].

Παρατήρηση. Διάφορες τεχνικές προρρύθμισης, με χρήση κυκλοειδών πινάκων, μπορούν να βρεθούν στις [38, 62] και [63]. Στις πρώτες δύο οι συγγραφείς συμμετρικοποιούν τον αρχικό πίνακα συντελεστών και λύνουν το σύστημα που προκύπτει με την Προρρυθμισμένη μέθοδο Ελαχίστων Υπολοίπων (PMINRES) [60]. Στην τελευταία, οι συγγραφείς αναλύουν προρρυθμιστές οι οποίοι ανήκουν σε τριγωνομετρική άλγεβρα, για μη-συμμετρικά συστήματα Toeplitz. Επικεντρώνονται στην επίλυση του συστήματος με τη μέθοδο PCGN, αλλά χρησιμοποιούν και PGMRES. Σημειώνουμε ότι στις παραπάνω εργασίες δεν έγινε χρήση ταινιωτών-επί-άλγεβρα (Band-times-Algebra) προρρυθμιστών. Θεωρητικά αποτελέσματα για τη συσσώρευση των ιδιαζουσών τιμών μπορούν επίσης να βρεθούν στην [28].

Στη συνέχεια, θα δώσουμε τον αριθμό επαναλήψεων χρησιμοποιώντας τη μέθοδο PGMRES, για τα Παραδείγματα 1 και 3 της [38]. Προσαρμόσαμε τις επιλογές στα αριθμητικά πειράματα, έτσι ώστε να είναι ίδιες με αυτές της [38], για να έχουμε αμεσότερη σύγκριση. Πιο συγκεκριμένα, αλλάξαμε το διάνυσμα b, έτσι ώστε αυτό να έχει όλες τις συνιστώσες του ίσες με 1. Επιπλέον αλλάξαμε το κριτήριο τερματισμού σε $\frac{\|r(k)\|_2}{\|r(0)\|_2} < 10^{-7}$ (ακριβώς όπως στην [38]).

Παράδειγμα 3.3.5. Σε αυτό το παράδειγμα, πίνακας συντελεστών είναι ο πίνακας

Grcar:

$$G_n = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & \cdots & 0 \\ -1 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 & -1 & 1 \end{bmatrix}.$$

Ο G_n έχει ως γεννήτρια συνάρτηση το τριγωνομετρικό πολυώνυμο $\mathfrak{f}_8(x) = 1 + \cos(2x) + \cos(3x) + \mathrm{i}\,[-2\sin(x) - \sin(2x) - \sin(3x)]$. Οι αριθμοί επαναλήψεων εφαρμόζοντας PGMRES και τους προρρυθμιστές $\mathcal{C}_n$ και $\mathcal{T}_n$, καθώς επίσης και PMINRES με τον προρρυθμιστή που προτάθηκε στην [38] (αυτός είναι ο $|\mathcal{T}_n|$) δίνεται στον Πίνακα 3.5. Δίνουμε επίσης και τους αριθμούς επαναλήψεων, με χρήση του $|\mathcal{C}_n|$ ως προρρυθμιστή, του οποίου η αποτελεσματικότητα αποδείχθηκε στην [29].

n	PGMRES			PMINRES	
	I_n	$\mathcal{C}_n$	$\mathcal{T}_n$	$\|\mathcal{C}_n\|$	$\|\mathcal{T}_n\|$
128	94	4	6	9	13
256	158	4	6	9	12
512	218	4	6	9	11
1024	213	4	5	9	11

Πι404νακας 3.5: Επαναλήψεις ($\mathfrak{f}_8$).

Αν και το κόστος ανά επανάληψη της μεθόδου PMINRES, είναι λιγότερο από αυτό της PGMRES, διότι το κόστος της τελευταίας αυξάνεται από επανάληψη σε επανάληψη, παρατηρούμε ότι οι αριθμοί επαναλήψεων που δίνονται από τον $\mathcal{C}_n$ είναι επαρκώς μικρότεροι από αυτούς της PMINRES. Οι 4 επαναλήψεις είναι πολύ λίγες κι έτσι το κόστος ανά επανάληψη είναι περίπου το ίδιο με αυτό της PMINRES.

Παρατηρούμε ότι ο $\mathcal{C}_n$ είναι πιο αποτελεσματικός από τον βέλτιστο κυκλοειδή προρρυθμιστή. Αυτό ισχύει κι όταν ο πίνακας συντελεστών του συστήματος δίνεται από τη συνάρτηση υπερβολικού ημιτόνου του G_n. Χρησιμοποιούμε τους προρρυθμιστές $\sinh \mathcal{C}_n$ και $\sinh \mathcal{T}_n$ και δίνουμε τα αντίστοιχα αποτελέσματα στον Πίνακα 3.6.

n	I_n	$\sinh \mathcal{C}_n$	$\sinh \mathcal{T}_n$
64	64	8	14
128	124	9	13
256	240	9	11
512	486	9	10

Πίνακας 3.6: Επαναλήψεις για τον $\sinh G_n$.

Παράδειγμα 3.3.6. Κατόπιν διακριτοποίησης ολοκληρω-διαφορικών (integro-differential) εξισώσεων, υπάρχει περίπτωση ο πίνακας συντελεστών να σχετίζεται με την εκθετική συνάρτηση ενός πίνακα Toeplitz [43]. Έστω $\mathfrak{f}_2(x) = x^2 + \mathrm{i}x^3$ η γεννήτρια συνάρτηση του $T_n(\mathfrak{f}_2)$. Μπορούμε να χρησιμοποιήσουμε τον $\mathrm{e}^{B\mathcal{C}_n}$ ως προρρυθμιστή για τον $\mathrm{e}^{T_n(\mathfrak{f}_2)}$. Στον Πίνακα 3.7 δίνουμε τους αριθμούς επαναλήψεων, εφαρμόζοντας PGMRES. Σημειώνεται ότι αν και ο πίνακας $\mathrm{e}^{T_n(\mathfrak{f}_2)}$ δεν είναι Toeplitz, ο προτεινόμενος προρρυθμιστής επιτυγχάνει την ταχεία σύγκλιση στη λύση του συστήματος.

n	I_n	$\mathrm{e}^{B\mathcal{C}_n}$
256	255	11
512	497	12
1024	$>$500	12
2048	$>$500	13

Πίνακας 3.7: Επαναλήψεις για τον $\mathrm{e}^{T_n(\mathfrak{f}_2)}$.

Κεφάλαιο 4

Συστήματα Toeplitz με ‘Αγνωστη Γεννήτρια Συνάρτηση

Σε αυτό το κεφάλαιο μελετάμε την προρρύθμιση $n \times n$ μη συμμετρικών, πραγματικών συστημάτων Toeplitz, όταν η γεννήτρια συνάρτηση του πίνακα συντελεστών T_n δεν είναι γνωστή εκ των προτέρων, όμως γνωρίζουμε ότι μια γεννήτρια συνάρτηση f, η οποία σχετίζεται με την ακολουθία πινάκων $\{T_n\}$, $T_n = T_n(f)$, όντως υπάρχει. Γίνεται κατάλληλη προσαρμογή, τόσο των ταινιωτών προρρυθμιστών του δευτέρου κεφαλαίου, όσο και των κυκλοειδών/ταινιωτών-επί-κυκλοειδών προρρυθμιστών του προηγούμενου κεφαλαίου. Αναλύεται ο τρόπος κατασκευής των προρρυθμιστών, από τις τιμές του πίνακα συντελεστών και μελετάται η συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών του προρρυθμισμένου συστήματος.

4.1 Ταινιωτοί προρρυθμιστές

Θα ξεκινήσουμε από την παρουσίαση των ταινιωτών Toeplitz προρρυθμιστών, δίνοντας τον τρόπο κατασκευής αυτών και μελετώντας τόσο τη συνεχή, όσο και την ασυνεχή περίπτωση. Λόγω του ότι δεν είναι γνωστό σε ποια από τις δύο περιπτώσεις βρισκόμαστε, θα δώσουμε έναν τρόπο εύρεσης πιθανών σημείων ασυνέχειας. Στο τέλος της ενότητας θα δώσουμε ορισμένα αριθμητικά παραδείγματα, τα οποία υποδεικνύουν την αποτελεσματικότητα του προτεινόμενου προρρυθμιστή.

4.1.1 Κατασκευή του προρρυθμιστή

Αρχικά, θα θέλαμε να σημειώσουμε ότι αφού ο πίνακας T_n είναι πραγματικός και μη-συμμετρικός, προκύπτει από μια συνάρτηση $f = f_1 + \mathrm{i}f_2$, όπου f_1 είναι άρτια, f_2 περιττή και i είναι η φανταστική μονάδα. Για να εκτιμήσουμε τις ρίζες της f, θα πρέπει να προσεγγίσουμε τις συναρτήσεις f_1 και f_2, που την απαρτίζουν, χρησιμοποιώντας τις τιμές του αρχικού πίνακα συντελεστών. Αφού επιλέξουμε ένα ισοκατανεμημένο πλέγμα $G_n = \{\theta_j\}$, όπου

$$\theta_j = -\pi + \frac{2j\pi}{n+1},\ j = 1, \ldots, n,$$

θα προσεγγίσουμε τις συναρτήσεις f_1 και f_2 (στο G_n). Μια προφανής προσέγγιση αποτελεί το ανάπτυγμα Fourier, διότι η γεννήτρια συνάρτηση του πίνακα T_n είναι εξ ορισμού το ίδιο το ανάπτυγμα Fourier, αναλόγως με τη διάσταση n. Το πηλίκο Rayleigh με χρήση κατάλληλων διανυσμάτων αποτελεί ακόμη ένα μαθηματικό εργαλείο για την προσέγγιση της f [71]. Ωστόσο, αν αυτό εφαρμοστεί σε όλο το πλέγμα G_n, το υπολογιστικό κόστος υπερβαίνει το $\mathcal{O}(n \log n)$ κι έτσι η χρήση του γίνεται πρακτικά απαγορευτική. Αυτός είναι ο λόγος που προτιμάμε τον υπολογισμό του αναπτύγματος Fourier στα σημεία του G_n, ο οποίος μπορεί να γίνει σε $\mathcal{O}(n \log n)$, χρησιμοποιώντας τον ταχύ μετασχηματισμό Fourier. $\forall j = 1, \ldots, n$ έχουμε:

$$f(\theta_j) \simeq F_{n-1}(\theta_j) = \sum_{k=-n+1}^{n-1} t_k \mathrm{e}^{\mathrm{i}k\theta_j}. \tag{4.1}$$

Χωρίζοντας το πραγματικό και φανταστικό μέρος, των λαμβανόμενων τιμών, προσεγγίζουμε τις συναρτήσεις f_1 και f_2, αντίστοιχα. Παρακάτω θα αναλύσουμε τη διαδικασία επιλογής μιας πιθανής ρίζας για τη συνάρτηση f_1. Παρόμοια ανάλυση ισχύει και για τη συνάρτηση f_2.

Γνωρίζοντας ότι η f_1 είναι άρτια συνάρτηση, συμπεραίνουμε ότι αυτή θα μπορούσε να έχει ρίζες είτε ανάμεσα σε δύο διαδοχικά σημεία θ_j και θ_{j+1}, όπου λαμβάνει διαφορετικό πρόσημο (σε αυτή την περίπτωση η συνάρτηση f_1 τέμνει τον άξονα), είτε ανάμεσα σε δύο σημεία θ_{j-1} και θ_{j+1}, με $f_1(\theta_j)$ να λαμβάνει μια πολύ μικρή τιμή, σχεδόν ίση με μηδέν κι επιπλέον η ακολουθία $\{f_1(\theta_i),\ i = 1, \ldots, n\}$ να αλλάζει τοπικά μονοτονία στο θ_j (σε αυτή την περίπτωση η f_1 εφάπτεται στον άξονα). Λαμβάνοντας υπόψη τη λεπτότητα του πλέγματος G_n, οι παραπάνω περιπτώσεις μπορούν να ενοποιηθούν ως εξής: Επιλέγουμε το θ_i, $1 \leq i \leq n$ ως σημείο πιθανής ρίζας αν $|\,\mathrm{Re}\,(F_{n-1}(\theta_i))\,|$ λαμβάνει πολύ μικρές τιμές κοντά στο μηδέν, π.χ. $|\,\mathrm{Re}\,(F_{n-1}(\theta_i))\,| < 10^{-6}$. Επισημαίνουμε ότι σε περίπτωση που $|\,\mathrm{Re}\,(F_{n-1}(\theta_i))\,| = |\,\mathrm{Re}\,(F_{n-1}(\theta_{i+1}))\,| \simeq 0$, θέτουμε ως σημείο πιθανής ρίζας τον μέσο όρο των τιμών θ_i και θ_{i+1}.

Σχολιάζουμε ότι η παραπάνω διαδικασία εφαρμόζεται σε συστήματα Toeplitz μεγάλης διάστασης. Για συστήματα μικρής διάστασης θα ήταν καλό να προχωρήσουμε και σε μια τεχνική εκλέπτυνσης κοντά στα σημεία πιθανών ριζών, με χρήση του πηλίκου Rayleigh, όπως έγινε στις [41, 56, 71, 72]).

Για την εύρεση του κατάλληλου τριγωνομετρικού πολυωνύμου, το οποίο αίρει την κακή κατάσταση του αρχικού συστήματος, είναι αναγκαία και η εύρεση της πολλαπλότητας κάθε ρίζας. Έστω m_i^1 και m_i^2 οι πολλαπλότητες των ριζών της f_1 και f_2, αντίστοιχα, οι οποίες αφορούν στο σημείο πιθανής ρίζας x_i, $i = 1, 2, \ldots, \rho$. Από εδώ και στο εξής με m_0^1 και m_0^2 θα συμβολίζουμε τις πολλαπλότητες της ρίζας στο $x_0 = 0$, για την f_1 και f_2, αντίστοιχα.

Θα περιγράψουμε την εκτίμηση της πολλαπλότητας των ριζών για τη συνάρτηση f_1. Αρχικά, υποθέτουμε για απλούστευση ότι $\mathrm{Re}(F_{n-1}(\theta_j)) \geq 0, \forall j = 1, 2, \ldots, n$. Επιλέγουμε τον άνω αριστερά κύριο τετραγωνικό υποπίνακα του T_n, μιας συγκεκριμένης μικρής διάστασης, π.χ. 64×64. Σκοπός μας είναι να μελετήσουμε τη συμπεριφορά της ιδιοτιμής που αντιστοιχεί στη ρίζα που εξετάζουμε πηγαίνοντας από κάποια μικρή διάσταση στη διπλάσιά της. Λαμβάνοντας τον $k \times k$ κύριο υποπίνακα του T_n, π.χ. $k = 16$, υπολογίζουμε το συμμετρικό μέρος του πίνακα, S_k^1, το οποίο αντιστοιχεί στην f_1. Σημειώνουμε ότι για κάθε ρίζα υπάρχει μια αντίστοιχη ιδιοτιμή, η οποία τείνει στο 0, όσο η μεταβλητή k παίρνει μεγαλύτερες τιμές. Έστω x_i το σημείο όπου το $\mathrm{Re}\left(F_{n-1}(\theta_j)\right)$, $j = 1, 2, \ldots, n$ λαμβάνει την ελάχιστη τιμή. Εφαρμόζουμε τη μέθοδο των Αντιστρόφων Δυνάμεων (Inverse Power) [21, 45] στον S_k^1, με αρχικό διάνυσμα $\Theta_{i,k}$, ορισμένο ως:

$$\Theta_{i,k} = \frac{1}{\sqrt{k}} \left(1, \mathrm{e}^{\mathrm{i}x_i}, \ldots, \mathrm{e}^{\mathrm{i}(k-1)x_i}\right)^T .$$

Εκτιμούμε την πολλαπλότητα m_i^1 εφαρμόζοντας την τεχνική που προτάθηκε στην [56]. Με πιο απλά λόγια, εξετάζουμε κατά πόσο η απόσταση μεταξύ δύο διαδοχικών ιδιοτιμών (κατ" αναλογία με το k) γίνεται όλο και πιο μικρή. Διπλασιάζοντας τη μεταβλητή k ξανά και ξανά (ας πούμε από 16 σε 32 και τέλος σε 64), εκτιμούμε το λόγο:

$$s_i^1 = \frac{\widetilde{\lambda^1}_{i,k} - \widetilde{\lambda^1}_{i,2k}}{\widetilde{\lambda^1}_{i,2k} - \widetilde{\lambda^1}_{i,4k}}.$$

Με $\widetilde{\lambda^1}_{i,k}$ συμβολίζουμε την πλησιέστερη στο 0 ιδιοτιμή (που αντιστοιχεί στη ρίζα x_i) του συμμετρικού μέρους του $k \times k$ κύριου υποπίνακα του T_n, υπολογισμένη με λίγες επαναλήψεις της μεθόδου Αντιστρόφων Δυνάμεων.

Εκτιμούμε την πολλαπλότητα m_i^1, ως τον κοντινότερο ακέραιο στο $\log_2\left(s_i^1\right)$. Αυτό αποδεικνύεται ως εξής: Είναι γνωστό ότι η ιδιοτιμή $\lambda_{i,k}^1$ που αντιστοιχεί στη ρίζα x_i με πολλαπλότητα m_i^1, τείνει στο 0 με ταχύτητα $\mathcal{O}\left(\frac{1}{k^{m_i^1}}\right)$. Επομένως, αυτή γράφεται ως:

$$\lambda_{i,k}^1 = c\frac{1}{k^{m_i^1}} + o\left(\frac{1}{k^{m_i^1}}\right).$$

Τότε, η προσέγγιση $\widetilde{\lambda^1}_{i,k}$ της $\lambda_{i,k}^1$ είναι η $\widetilde{\lambda^1}_{i,k} \simeq c\frac{1}{k^{m_i^1}}$ και ο λόγος s_i^1 προσεγγίζεται ως:

$$s_i^1 \simeq \frac{c\frac{1}{k^{m_i^1}} - c\frac{1}{(2k)^{m_i^1}}}{c\frac{1}{(2k)^{m_i^1}} - c\frac{1}{(4k)^{m_i^1}}} = \frac{c\frac{1}{k^{m_i^1}}\left(1-\frac{1}{2^{m_i^1}}\right)}{c\frac{1}{k^{m_i^1}}\frac{1}{2^{m_i^1}}\left(1-\frac{1}{2^{m_i^1}}\right)} = 2^{m_i^1}.$$

Ως επακόλουθο, $\log_2 s_i^1 \simeq m_i^1$.

Αν η γραφική παράσταση της f_1, λαμβάνει θετικές και αρνητικές τιμές, προφανώς υπάρχουν σημεία ριζών όπου υπάρχει τομή με τον άξονα. Για την εκτίμηση της πολλαπλότητας αυτών των ριζών, δεν μπορούμε να χρησιμοποιήσουμε τη διαδικασία που περιγράψαμε παραπάνω. Αυτή μπορεί να εφαρμοστεί μόνο για ρίζες που αντιστοιχούν σε τοπικά ελάχιστα. Έτσι, θα προσπαθήσουμε να εκτιμήσουμε την πολλαπλότητα αυτών των ριζών για τη συνάρτηση $|f_1|$, όπου τα σημεία τομής μετατρέπονται σε τοπικά ελάχιστα. Φυσικά, οι πολλαπλότητες των ριζών της $|f_1|$ παραμένουν ίδιες με αυτές της f_1.

Επομένως, πρέπει να υπολογίσουμε τον πίνακα $\widehat{S}_k^1 = T_k(|f_1|)$, για μικρές τιμές της μεταβλητής k (π.χ., 16, 32 και 64), επειδή δε θέλουμε να αυξήσουμε την πολυπλοκότητα του αλγορίθμου. Θα χρησιμοποιήσουμε τις τιμές $|\operatorname{Re}(F_{n-1})|$ στο $G_n \cup \{-\pi, \pi\}$. Τότε,

$$\begin{aligned}\left(\widehat{S}_k^1\right)_{rq} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |f_1(x)| \mathrm{e}^{-\mathrm{i}(r-q)x}\mathrm{d}x \\ &\simeq \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\operatorname{Re}\left(F_{n-1}(x)\right)\right| \mathrm{e}^{-\mathrm{i}(r-q)x}\mathrm{d}x.\end{aligned} \tag{4.2}$$

Μπορούμε να προσεγγίσουμε το τελευταίο ολοκλήρωμα με χρήση του σύνθετου κανόνα του Simpson, σε σημεία του $G_n \cup \{-\pi, \pi\}$, ή οποιαδήποτε άλλη μέθοδο αριθμητικής ολοκλήρωσης. Ακολουθώντας την τεχνική που προαναφέραμε για τον $\widehat{S}_k^1$, υπολογίζουμε τις πολλαπλότητες των ριζών της f_1.

Είναι προφανές ότι για την εκτίμηση των πολλαπλοτήτων που έχουν οι ρίζες της συνάρτησης f_2, δε μπορούμε να αποφύγουμε τον υπολογισμό του πίνακα $\widehat{S}_k^2 =$

$T_k(|f_2|)$, επειδή η f_2 λαμβάνει πάντα θετικές και αρνητικές τιμές ως περιττή συνάρτηση. ῾Ολα τα υπόλοιπα παραμένουν ίδια, με τη διαφορά ότι το φανταστικό μέρος $\mathrm{Im}(F_{n-1})$ παίρνει τη θέση του $\mathrm{Re}(F_{n-1})$. Θα θέλαμε να σχολιάσουμε ότι σε περίπτωση που υπάρχει μια ρίζα x_i όπου $f_1(x_i) = 0$, ενώ $f_2(x_i) \neq 0$ (ή $f_2(x_i) = 0$ και $f_1(x_i) \neq 0$), τότε θέτουμε $m_i^2 = 0$ (ή $m_i^1 = 0$).

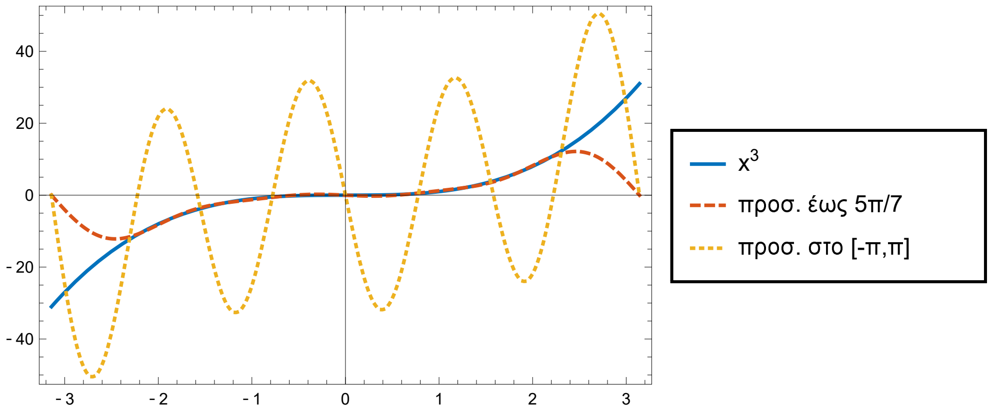
Αναφέρουμε ότι η τεχνική για την οποία γίνεται λόγος κι έχει να κάνει με τον υπολογισμό των πολλαπλοτήτων όταν η f_ℓ, $\ell = 1, 2$ λαμβάνει θετικές και αρνητικές τιμές, αποτελεί μια βελτίωση της αντίστοιχης τεχνικής των [56, 71], όπου οι συγγραφείς μελέτησαν μη-αρνητικές συναρτήσεις.

Λόγω της λεπτότητας του πλέγματος G_n, είναι εύκολο να εκτιμήσουμε τα σημεία πιθανής ασυνέχειας των συναρτήσεων f_1 και f_2. Για απλούστευση, θα περιγράψουμε αυτή την εκτίμηση για την f_1. Το μόνο που έχουμε να κάνουμε είναι να ελέγξουμε το μέγεθος του λόγου $\mathrm{Re}\left(F_{n-1}(\theta_j)\right) - \mathrm{Re}\left(F_{n-1}(\theta_{j+1})\right)$, $j = 1, 2, \ldots, n$, προς $h = \frac{2\pi}{n+1}$, όπου $\theta_{n+1} = \theta_1$. Αυτό σημαίνει ότι ελέγχουμε το πόσο διαφέρουν τα πραγματικά μέρη του αναπτύγματος F_{n-1}, για δύο διαδοχικά σημεία, σε σχέση με το βήμα h. Αν

$$\frac{|\,\mathrm{Re}\left(F_{n-1}(\theta_j)\right) - \mathrm{Re}\left(F_{n-1}(\theta_{j+1})\right)|}{h} = \Omega(n),$$

(αρκούντως μεγάλο), τότε υποθέτουμε ότι υπάρχει ένα σημείο ασυνέχειας ή ταχεία μεταβολή της συνάρτησης (unbounded variation), στο διάστημα $[\theta_j, \theta_{j+1}]$. Προφανώς, ακριβώς η ίδια ανάλυση μπορεί να γίνει και για την f_2. Αυτό σημαίνει ότι αν αντικαταστήσουμε το Re με Im, μπορούμε να εκτιμήσουμε τα σημεία ασυνέχειας της f_2.

Η εκτίμηση των σημείων ασυνέχειας απαιτείται για την πολυωνυμική προσέγγιση της $\frac{f}{g_n}$, όπου g_n είναι το κατάλληλο τριγωνομετρικό πολυώνυμο το οποίο αίρει τις εκτιμώμενες ρίζες της συνάρτησης f. ᾿Οπως περιγράψαμε στο δεύτερο κεφάλαιο, ο προρρυθμιστής γίνεται ιδιαίτερα πιο αποτελεσματικός αν αποκλείσουμε μια περιοχή κοντά στο σημείο ασυνέχειας, πριν εφαρμόσουμε τον αλγόριθμο βέλτιστης ομοιόμορφης προσέγγισης του Remez. Εκεί, η απόσταση από το σημείο ασυνέχειας, επιλέχθηκε εμπειρικά να είναι ίση με $\frac{2\pi}{7}$. Ως παράδειγμα, στο Σχήμα 4.1 παρουσιάζουμε την προσέγγιση της συνάρτησης x^3 στο $[-\pi, \pi]$, όταν αποκλείουμε κάποια σημεία κοντά στο $\pm\pi$ (όπου εντοπίζεται η ασυνέχεια), καθώς κι όταν δεν προχωρούμε σε κάποια εξαίρεση σημείων. Η προσέγγιση έγινε με τριγωνομετρικά πολυώνυμα τετάρτου βαθμού και όπως παρατηρείται, το σφάλμα της προσέγγισης είναι κατά πολύ μεγαλύτερο στη δεύτερη περίπτωση.



Σχήμα 4.1: Προσέγγιση της x^3.

Προκειμένου να αποφύγουμε την κακή κατάσταση θα πρέπει να βρούμε ένα τριγωνομετρικό πολυώνυμο, το οποίο έχει τις ίδιες ρίζες με τη γεννήτρια συνάρτηση f ώστε να άρουμε τις ρίζες αυτής. Έστω x_i, $i = 1, 2, \ldots, \rho$ οι ακριβείς μη-μηδενικές ρίζες της συνάρτησης f στο διάστημα $(0, \pi]$, με πολλαπλότητες m_i^1 για την f_1 και m_i^2 για την f_2, αντίστοιχα, $i = 1, 2, \ldots, \rho$. Λόγω του ότι η f_1 είναι άρτια και η f_2 περιττή, αν x_i είναι μια ρίζα στο $(0, \pi]$, το σημείο $-x_i$ είναι επίσης μια ρίζα στο $[-\pi, 0)$, με την ίδια πολλαπλότητα. Τότε, όπως περιγράψαμε εκτενώς στο δεύτερο κεφάλαιο, η μορφή του τριγωνομετρικού πολυωνύμου g δίνεται ως:

Αν $m_i^1 \leq m_i^2, \forall i = 1, 2, \ldots, \rho$:

$$g = \text{sign}(c_1(x)) \prod_{i=1}^{\rho} \left(\cos(x_i) - \cos(x)\right)^{m_i^1}.$$

Σε περίπτωση που η f_1 έχει επίσης ρίζα στο 0, με πολλαπλότητα $m_0^1 \leq m_0^2$:

$$g = \text{sign}(c_1(x))(2 - 2\cos(x))^{\frac{m_0^1}{2}} \prod_{i=1}^{\rho} \left(\cos(x_i) - \cos(x)\right)^{m_i^1}.$$

Ωστόσο, αν υπάρχει τουλάχιστον μια ρίζα $x_j : m_j^1 > m_j^2$:

$$
\begin{aligned}
g = &\operatorname{sign}(c_1(x)) \prod_{i=1}^{\rho} (\cos(x_i) - \cos(x))^{m_i^1} \\
&+ \mathrm{i}\operatorname{sign}(c_2(x))\,(\sin(x))^{m_0^2} \prod_{i=1}^{\rho} (\cos(x_i) - \cos(x))^{m_i^2}.
\end{aligned} \tag{4.3}
$$

Σημειώνουμε ότι αν η f_1 έχει μια ρίζα στο 0, με πολλαπλότητα m_0^1, πολλαπλασιάζουμε τον πρώτο όρο της (4.3) με $(2-2\cos(x))^{\frac{m_0^1}{2}}$. Οι συναρτήσεις c_1 και c_2 ορίζονται με τον τρόπο που περιγράψαμε στο δεύτερο κεφάλαιο. Πρακτικά, τα πρόσημα αυτών $\operatorname{sign}(c_1)$ και $\operatorname{sign}(c_2)$ επιλέγονται έτσι ώστε $\mathrm{Re}\left(\frac{f}{g}\right) > 0$.

Θα δώσουμε ένα απλό παράδειγμα, ώστε να γίνει περισσότερο κατανοητός στον αναγνώστη ο τρόπος με τον οποίο επιλέγουμε το τριγωνομετρικό πολυώνυμο g. Έστω $f(x) = (x-x_1)(x+x_1) + \mathrm{i}x(x-x_1)^2(x+x_1)^2$, $x_1 \in (0,\pi]$, $x \in [-\pi,\pi]$. Όπως φαίνεται η f έχει ρίζες στο $\pm x_1$ με πολλαπλότητα ίση με 1. Παρατηρούμε ότι η πολλαπλότητα της ρίζας για την f_1 είναι $m_1^1 = 1$, ενώ για την f_2 είναι $m_1^2 = 2$ και $m_0^2 = 1$. Επειδή $m_1^1 < m_1^2$, επιλέγουμε το τριγωνομετρικό πολυώνυμο $g(x) = \cos(x_1) - \cos(x)$, για την άρση της κακής κατάστασης. Τότε, η $\frac{f}{g}$ γράφεται:

$$
\frac{f(x)}{g(x)} = \frac{(x-x_1)(x+x_1)}{\cos(x_1)-\cos(x)} + \mathrm{i}\frac{x(x-x_1)^2(x+x_1)^2}{\cos(x_1)-\cos(x)}.
$$

Είναι εύκολο να ελέγξει κανείς ότι το πραγματικό μέρος της παραπάνω συνάρτησης δεν έχει ρίζες και είναι θετικό μακριά από το 0. Από την άλλη, το φανταστικό μέρος έχει ρίζες στο 0 και $\pm x_1$, με πολλαπλότητα ίση με 1. Ωστόσο, το γεγονός αυτό δεν αποτελεί πρόβλημα, διότι προσπαθούμε να πετύχουμε συσσώρευση του φανταστικού μέρους των ιδιοτιμών κοντά στο 0. Τονίζουμε ότι η συνάρτηση $\frac{f}{g}$ δεν έχει ρίζες, διότι $\mathrm{Re}\left(\frac{f}{g}\right) > 0$.

Σκοπός μας είναι να μελετήσουμε κατά ποιον τρόπο το σφάλμα κατά την εκτίμηση μιας ρίζας της συνάρτησης f επηρεάζει τη σύγκλιση της μεθόδου PGMRES. Σχολιάζουμε ότι στην εργασία [58] οι συγγραφείς έδωσαν αποτελέσματα για τον δείκτη κατάστασης του προρρυθμισμένου πίνακα, για δι-διάστατους (two-level) θετικά ορισμένους πίνακες Toeplitz, όπου η συνάρτηση f είναι δύο μεταβλητών, μη-αρνητική και άρτια, έχοντας ρίζες άρτιας τάξης. Σημειώνουμε ότι τα ίδια αποτελέσματα, λαμβάνοντας ανάλογες υποθέσεις/θεωρήσεις, ισχύουν και στην περίπτωσή μας όταν οι ρίζες έχουν άρτιες πολλαπλότητες. Επειδή αυτό δε συμβαίνει πάντα, θα πρέπει να βρούμε κάποιον εναλλακτικό τρόπο σύγκλισης, μέσω της μελέτης του σφάλματος. Μελετούμε κι εδώ ξεχωριστά το συμμετρικό και αντισυμμετρικό μέρος του προρρυθμισμένου συστήματος.

Πρακτικά εκτιμούμε τις ρίζες στα σημεία $\widetilde{x}_i \simeq x_i$, $i = 1, 2, \ldots, \rho$. Υποθέτουμε ότι οι πολλαπλότητες των ριζών υπολογίστηκαν με ακρίβεια, όπως επίσης και η ρίζα στο 0. Σε περίπτωση που εκτιμούμε δύο ρίζες κοντά στο 0 με απόσταση $o(1)$, θεωρούμε ότι έχουμε μία ρίζα στο 0. Επομένως, δημιουργείται ένα σφάλμα κατά την εκτίμηση των ριζών, που βρίσκονται μακριά από το 0. Το τριγωνομετρικό πολυώνυμο g_n, το οποίο επιτυγχάνει την άρση των εκτιμώμενων ριζών, δίνεται από τη διαδικασία που περιγράψαμε παραπάνω, με τη διαφορά ότι στις αντίστοιχες σχέσεις το $\widetilde{x}_i$ παίρνει τη θέση του x_i.

Για την απλούστευση της ανάλυσης, υποθέτουμε ότι η f έχει δύο ρίζες στα σημεία $\pm x_1$, $x_1 \neq 0$ με πολλαπλότητα κάποιον ακέραιο αριθμό $\alpha > 0$. Σε περίπτωση που έχουμε περισσότερα από ένα ζεύγη ριζών, η ανάλυση γενικεύεται άμεσα. Έστω $f = f_1 + \mathrm{i} f_2$ και υποθέτουμε ότι η f_1 και f_2 έχουν ρίζες στο $\pm x_1$ τάξεως α και β, αντίστοιχα, με $\beta \geq \alpha$, ενώ η f_2 έχει και μια επιπλέον ρίζα στο 0. Σε αυτή την περίπτωση, σύμφωνα με την προαναφερθείσα ανάλυση, το τριγωνομετρικό πολυώνυμο που αίρει την κακή κατάσταση θα έπρεπε να είναι το $g(x) = c\,(\cos(x_1) - \cos(x))^{\alpha}$. Ωστόσο, όπως τονίσαμε, πρακτικά έχουμε ένα σφάλμα στην εκτίμηση της ρίζας, $\widetilde{x}_1 - x_1 = \varepsilon$. Επομένως, αν χρησιμοποιήσουμε ως προρρυθμιστή τον ταινιωτό πίνακα $T_n(g_n)$, όπου $g_n(x) = c\,(\cos(\widetilde{x}_1) - \cos(x))^{\alpha}$ [75], θα πρέπει να μελετήσουμε τη συμπεριφορά του φάσματος του προρρυθμισμένου πίνακα $T_n^{-1}(g_n)T_n(f) = T_n^{-1}(g_n)T_n(f_1) + T_n^{-1}(g_n)T_n(\mathrm{i} f_2)$. Ο πρώτος όρος του αθροίσματος γράφεται ως:

$$\begin{aligned} T_n^{-1}(g_n)T_n(f_1) &= T_n^{-1}(g_n)\left[T_n(g_n)T_n\left(\frac{f_1}{g_n}\right) + L_1\right] \\ &= T_n\left(\frac{f_1}{g_n}\right) + L_2 = T_n\left(h_1\frac{g}{g_n}\right) + L_2, \end{aligned}$$

όπου L_1 και L_2 είναι πίνακες χαμηλής βαθμίδας, ίσης με 2α και h_1 είναι μια θετική και φραγμένη συνάρτηση. Το μη-φραγμένο μέρος της γεννήτριας συνάρτησης του τελευταίου πίνακα είναι ο λόγος $\frac{g}{g_n}$, που έχει ρίζα στο x_1 και πόλο στο $\widetilde{x}_1$. Είναι εύκολο να δει κανείς ότι αυτός ο λόγος είναι φραγμένος στο σύνολο $K = [-\pi, \pi] \setminus [(-\widetilde{x}_1 - \varepsilon, -\widetilde{x}_1 + \varepsilon) \cup (\widetilde{x}_1 - \varepsilon, \widetilde{x}_1 + \varepsilon)]$ και μη-φραγμένος στα διαστήματα $(-\widetilde{x}_1 - \varepsilon, -\widetilde{x}_1 + \varepsilon)$ και $(\widetilde{x}_1 - \varepsilon, \widetilde{x}_1 + \varepsilon)$. Υποθέσαμε, χωρίς βλάβη της γενικότητας ότι $\varepsilon = \widetilde{x}_1 - x_1 > 0$.

Λόγω της ισοκατανομής των ιδιοτιμών των πινάκων Toeplitz [33] το πολύ $4\varepsilon n$ ιδιοτιμές μπορούν να κυμαίνονται εκτός του $[a, b]$, όπου $a = \min\limits_{x \in K} \frac{f_1(x)}{g_n(x)}$ και $b = \max\limits_{x \in K} \frac{f_1(x)}{g_n(x)}$. Το μέγεθος του ε εξαρτάται από το πόσο ομαλή είναι η συνάρτηση f_1.

Σύμφωνα με την [71], αν η f_1 είναι συνεχής το σφάλμα του αναπτύγματος Fourier, για την προσέγγιση της f_1, είναι $\varepsilon = \mathcal{O}\left(\frac{\log n}{n}\right)$, επομένως $\mathcal{O}(\log n)$ ιδιοτιμές μπορούν να κυμαίνονται εκτός του διαστήματος $[a, b]$.

Το αντισυμμετρικό μέρος του προρρυθμισμένου πίνακα μας δίνει:

$$\begin{aligned} T_n^{-1}(g_n)T_n(\mathrm{i}f_2) &= T_n^{-1}(g_n)\left[T_n(g_n)T_n\left(\frac{\mathrm{i}f_2}{g_n}\right) + L_1'\right] \\ &= T_n\left(\mathrm{i}h_2\frac{g}{g_n}\right) + L_2', \end{aligned}$$

όπου h_2 είναι μια φραγμένη και περιττή συνάρτηση, η οποία έχει μια ρίζα στο 0 με την πολλαπλότητα που έχει και η f_2, καθώς επίσης και ρίζες στο $\pm x_1$ με πολλαπλότητα $\beta - \alpha$. Οι πίνακες L_1', L_2' έχουν χαμηλή βαθμίδα ίση με 2α. Ακολουθώντας την ανάλυση που περιγράψαμε για το συμμετρικό μέρος, έχουμε το πολύ $\mathcal{O}(\log n)$ ιδιοτιμές εκτός του $[-c, c]$, όπου $c = \max\limits_{x \in K} \frac{f(x)}{g_n(x)}$.

Αν $\beta < \alpha$, τότε $g(x) = c\left(\cos(x_1) - \cos(x)\right)^{\alpha} + \mathrm{i}c' \sin(x)^{\gamma}\left(\cos(x_1) - \cos(x)\right)^{\beta}$ κι έχουμε ότι:

$$\frac{f}{g_n} = \frac{z^{\beta}}{z_n^{\beta}} \frac{h_2 \sin(x)^{\gamma} - \mathrm{i}h_1 z^{\alpha-\beta}}{c' \sin(x)^{\gamma} - \mathrm{i}c z_n^{\alpha-\beta}} = \frac{z^{\beta}}{z_n^{\beta}} w, \tag{4.4}$$

όπου $z = \cos(x_1) - \cos(x)$, $z_n = \cos(\widetilde{x}_1) - \cos(x)$, h_1, h_2 είναι φραγμένες, θετικές και άρτιες συναρτήσεις. Ο πρώτος λόγος, $\frac{z^{\beta}}{z_n^{\beta}}$, αποτελεί τον μη-φραγμένο παράγοντα του $\frac{f}{g_n}$, ενώ ο δεύτερος, w, είναι φραγμένος. Είναι εύκολο να δούμε ότι το πραγματικό μέρος του w είναι φραγμένη, θετική και άρτια συνάρτηση, ενώ το φανταστικό φραγμένη και περιττή. Επομένως,

$$T_n^{-1}(g_n)T_n(f) = T_n\left(\frac{f}{g_n}\right) + \widehat{L}_1 = T_n\left(w\frac{z^{\beta}}{z_n^{\beta}}\right) + \widehat{L}_2.$$

Ακολουθώντας την ίδια ανάλυση, με χρήση των ιδιοτήτων ισοκατανομής των ιδιοτιμών του συμμετρικού και αντισυμμετρικού μέρους του $T_n\left(w\frac{z^{\beta}}{z_n^{\beta}}\right)$, όπως και στην πρώτη περίπτωση, καταλήγουμε σε ανάλογα αποτελέσματα. Έτσι, το πολύ $\mathcal{O}(\log n)$ ιδιοτιμές κυμαίνονται εκτός του ορθογωνίου $[a, b] \times [-c, c]$. Περισσότερες λεπτομέρειες δίνονται στην απόδειξη του Θεωρήματος 4.1.1.

Όπως προαναφέρθηκε, στο πρόβλημα που μελετάμε, η γεννήτρια συνάρτηση f δεν είναι γνωστή εκ των προτέρων. Ωστόσο, έχουμε ήδη υπολογίσει το ανάπτυγμα Fourier στο G_n. Οπότε, υπολογίζουμε το λόγο $\frac{f}{g}$ προσεγγίζοντας τον από το λόγο

$\widehat{f} = \frac{F_{n-1}}{g_n}$ σε κάποιο υποσύνολο του G_n. Μένει να προσαρμόσουμε κατάλληλα την τεχνική προσέγγισης που προτάθηκε στην [49], για τη βέλτιστη ομοιόμορφη προσέγγιση, με χρήση του αλγορίθμου Remez, της $\widehat{f_1} = \mathrm{Re}(\widehat{f})$ και $\widehat{f_2} = \mathrm{Im}(\widehat{f})$ με άρτια και περιττά τριγωνομετρικά πολυώνυμα κατάλληλων βαθμών, αντίστοιχα. Αρχικά, επιλέγουμε ένα ισοκατανεμημένο σύνολο σημείων από το πλέγμα G_n, έστω G_k, $k \ll n$, στο $(0, \pi)$. Για να βελτιώσουμε την απόδοση του προρρυθμιστή, καλό θα ήταν να αποκλείσουμε κάποια σημεία του G_k, τα οποία ανήκουν σε περιοχές ασυνέχειας της $\widehat{f_1}$ ή της $\widehat{f_2}$ (μειώνοντας το σφάλμα προσέγγισης από τον αλγόριθμο Remez). Εξαιτίας σφαλμάτων κατά την εκτίμηση των πιθανών ριζών και λαμβάνοντας υπόψη τη σχέση (4.4), η συνάρτηση $\widehat{f_1}$, ή η $\widehat{f_2}$ μπορεί να μην είναι φραγμένες σε μικρές περιοχές που περιέχουν τις ρίζες. Για να μειώσουμε το σφάλμα προσέγγισης, καλό είναι να αποκλείσουμε επίσης κάποια σημεία τα οποία ανήκουν σε τέτοιες περιοχές. Τέλος, σχηματίζουμε τα σύνολα G_k^1 και G_k^2 για την $\widehat{f_1}$ και $\widehat{f_2}$, αντίστοιχα, κι εφαρμόζουμε τον αλγόριθμο προσέγγισης Remez.

Προσεγγίζουμε τις $\widehat{f_1}$ και $\widehat{f_2}$ με τις q_1 και q_2, αντίστοιχα. Τότε, ορίζουμε την $q = q_1 + \mathrm{i} q_2$ και σχηματίζουμε τον ταινιωτό πίνακα Toeplitz $T_n(p_n)$, όπου $p_n = g_n q$. Συμβολίζουμε με d_1 και d_2 τους βαθμούς των τριγωνομετρικών πολυωνύμων q_1 και q_2, αντίστοιχα. Από εδώ και στο εξής θα συμβολίζουμε τον προρρυθμιστή $T_n(p_n)$, ως R_{d_1,d_2}, για να φαίνονται οι βαθμοί των πολυωνύμων προσέγγισης. Με χρήση του R_{d_1,d_2} ως προρρυθμιστή, μπορούμε να λύσουμε μη-συμμετρικά και πραγματικά συστήματα Toeplitz, με αποτελεσματικό τρόπο. Θεωρητικά αποτελέσματα για την περίπτωση που η f είναι γνωστή εκ των προτέρων, δόθηκαν στο δεύτερο κεφάλαιο. Παρακάτω παρουσιάζουμε το αντίστοιχο του Θεωρήματος 2.1.2, το οποίο αφορά στη συσσώρευση των ιδιοτιμών, για το πρόβλημα που μας απασχολεί, δηλαδή όταν η f δεν είναι γνωστή εκ των προτέρων.

Θεώρημα 4.1.1. Ας είναι T_n ο $n \times n$, πραγματικός πίνακας Toeplitz, με άγνωστη γεννήτρια συνάρτηση. Έστω p_n το τριγωνομετρικό πολυώνυμο το οποίο προέκυψε από την προτεινόμενη διαδικασία, με σφάλματα στις εκτιμήσεις των ριζών τάξεως το πολύ $\mathcal{O}\left(\frac{\log n}{n}\right)$. Τότε, οι ιδιοτιμές του προρρυθμισμένου συστήματος $T_n(p_n)^{-1}T_n$ συσσωρεύονται στο ορθογώνιο $[a, b] \times [-c, c]$, όπου $a = \min_{x \in K} \mathrm{Re}\left(\frac{F_{n-1}(x)}{p_n(x)}\right)$, $b = \max_{x \in K} \mathrm{Re}\left(\frac{F_{n-1}(x)}{p_n(x)}\right)$, $c = \max_{x \in K} \mathrm{Im}\left(\frac{F_{n-1}(x)}{p_n(x)}\right)$ και $K = [-\pi, \pi] \setminus \bigcup_i ((-\beta_i, -\alpha_i) \cup (\alpha_i, \beta_i))$, με (α_i, β_i) να είναι διαστήματα που περιέχουν τις ρίζες x_i και τους πόλους $\widetilde{x}_i$, $i = 1, 2, \ldots, \rho$, έχοντας μήκος $\beta_i - \alpha_i = \mathcal{O}\left(\frac{\log n}{n}\right)$, καθώς επίσης και $x_i - \alpha_i$, $\widetilde{x}_i - \alpha_i$, $\beta_i - x_i$, $\beta_i - \widetilde{x}_i$ είναι της τάξεως $\mathcal{O}\left(\frac{\log n}{n}\right)$. Τότε, το πολύ $\mathcal{O}(\log n)$ ιδιοτιμές κυμαίνονται εκτός του ορθογωνίου.

Απο404δειξη. Αν μπορούσαμε να προσδιορίσουμε τις ρίζες με ακρίβεια, θα χρησιμοποιούσαμε ως προρρυθμιστή τον πίνακα $T_n(p)$, όπου $p = gq$. Τότε, από το Θεώρημα 2.1.2 θα προέκυπτε μια κύρια συσσώρευση στο ορθογώνιο $[a, b] \times [-c, c]$. Ωστόσο, χρησιμοποιούμε ως προρρυθμιστή τον πίνακα $T_n(p_n)$. Έχουμε επιλέξει το μήκος του $[\alpha_i, \beta_i]$ να είναι $\mathcal{O}\left(\frac{\log n}{n}\right)$, επειδή η απόσταση μεταξύ της ρίζας x_i και του πόλου $\widetilde{x}_i$ είναι τάξεως το πολύ $\mathcal{O}\left(\frac{\log n}{n}\right)$. Έχουμε επιλέξει επίσης τις αποστάσεις $x_i - \alpha_i$, $\widetilde{x}_i - \alpha_i$, $\beta_i - x_i$, $\beta_i - \widetilde{x}_i$ να είναι της τάξεως $\mathcal{O}\left(\frac{\log n}{n}\right)$, έτσι ώστε η συνάρτηση $\bar{f} = \frac{F_{n-1}}{p_n}$ να είναι φραγμένη (άνω και κάτω), ανεξαρτήτως της διάστασης n, στο σύνολο K. Εύκολα βλέπουμε ότι ο λόγος $\frac{\cos(x_i) - \cos(x)}{\cos(\widetilde{x}_i) - \cos(x)}$, ο οποίος χαρακτηρίζει τη συνάρτηση $\bar{f}$ ως μη φραγμένη στο $\widetilde{x}_i$ και δηλώνει ότι μηδενίζεται στο x_i, είναι φραγμένος ανεξαρτήτως της διάστασης n, έξω από το διάστημα (α_i, β_i). Αυτό εξασφαλίζει ότι η $\bar{f}$ είναι φραγμένη σε ένα ορθογώνιο $[a, b] \times [-c, c]$, όταν ορίζεται στο σύνολο K. Οι ιδιοτιμές που κυμαίνονται εκτός του $[a, b] \times [-c, c]$ εξαρτώνται από αυτά τα διαστήματα. Για να εκτιμήσουμε πόσες είναι αυτές, χρησιμοποιούμε το θεώρημα ισοκατανομής των ιδιοτιμών του Szegö. Αναλυτικότερα, θα πρέπει να ελέγξουμε ξεχωριστά το πραγματικό και φανταστικό μέρος της $\bar{f}$. Αρχικά, σταθεροποιούμε έναν ακέραιο n, ο οποίος είναι αρκετά μεγάλος και ουσιαστικά είναι η διάσταση του αρχικού συστήματος προς λύση $T_n x = b$. Στη συνέχεια ορίζουμε ως N, την ακέραια μεταβλητή που χρησιμοποιούμε στο θεώρημα του Szegö. Θα δώσουμε την απόδειξη για το πραγματικό μέρος της $\bar{f}$, καθώς αυτή που αφορά στο φανταστικό μέρος είναι ανάλογη.

Θέτουμε $\bar{f}_1 = \mathrm{Re}(\bar{f})$. Προφανώς η $\bar{f}_1$ είναι μη-φραγμένη στο σύνολο

$$\bigcup_i ((-\beta_i, -\alpha_i) \cup (\alpha_i, \beta_i))$$

και δε μπορούμε να χρησιμοποιήσουμε το θεώρημα του Szegö. Αφού ο n είναι σταθερός και η $\bar{f}_1$ προέρχεται από το πλέγμα G_n, μπορούμε να προσεγγίσουμε την $\bar{f}_1$ με την $\widetilde{f}_1$, η οποία είναι φραγμένη και παράγει τον ίδιο πίνακα $T_n(\bar{f}_1)$. Αυτό γίνεται ως εξής: Αν δεν υπάρχει κανένας πόλος ανάμεσα σε δύο διαδοχικά σημεία του G_n, η $\widetilde{f}_1$ λαμβάνει την ίδια τιμή με την $\bar{f}_1$. Αν υπάρχει κάποιος πόλος ανάμεσα σε δύο διαδοχικά σημεία w_j και w_{j+1}, η $\widetilde{f}_1$ λαμβάνει την τιμή του ευθυγράμμου τμήματος

$$\frac{\bar{f}_1(w_{j+1}) - \bar{f}_1(w_j)}{w_{j+1} - w_j}(x - w_j) + \bar{f}_1(w_j), \ x \in [w_j, w_{j+1}].$$

Αυτό σημαίνει ότι η $\widetilde{f}_1(w_j)$ είναι τοπικό ελάχιστο/μέγιστο, ενώ η $\widetilde{f}_1(w_{j+1})$ είναι τοπικό μέγιστο/ελάχιστο, αντίστοιχα, αλλά φράσονται καθώς η διάσταση είναι σταθερή. Επομένως, η $\widetilde{f}_1$ είναι μια φραγμένη και συνεχής συνάρτηση. Για να εφαρμόσουμε το θεώρημα του Szegö, θεωρούμε τη συνεχή και φραγμένη συνάρτηση F_η:

$$F_\eta(z) = \begin{cases} 1, \; z \leq a-\eta, z \geq b+\eta \\ 0, \; z \in [a,b] \end{cases},$$

η αρκούντως μικρό. Έχουμε:

$$\begin{aligned} &\limsup_{N\to\infty} \frac{1}{N} \#\{\lambda_j(T_n(\widetilde{f}_1)) < a \;\vee\; \lambda_j(T_n(\widetilde{f}_1)) > b\} \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} F_\eta(\widetilde{f}_1(x))\mathrm{d}x \leq \frac{1}{2\pi} \int_{\bigcup_{i=1}^{\rho}((-\beta_i,-\alpha_i)\cup(\alpha_i,\beta_i))} 1\mathrm{d}x \\ &= \sum_{i=1}^{\rho} 2(\beta_i-\alpha_i) = 2\sum_{i=1}^{\rho} c_i \frac{\log n}{n} = c\frac{\log n}{n}, \end{aligned}$$

όπου # δηλώνει τον πληθικό αριθμό του συνόλου και $\vee$ τη λογική διάζευξη (OR). Άρα, $\limsup_{N\to\infty} \#\{\lambda_j(T_n(\widetilde{f}_1)) < a \;\vee\; \lambda_j(T_n(\widetilde{f}_1)) > b\} \leq c\frac{\log n}{n}N$.

Πηγαίνοντας πίσω στο σταθερό n και δεδομένου ότι επιλέγουμε κάποιο (αρκούντως) μικρό η, καταλήγουμε στο ότι ο αριθμός των ιδιοτιμών που κυμαίνονται εκτός του διαστήματος $[a,b]$ (κατεύθυνση του πραγματικού άξονα) είναι το πολύ $c\frac{\log n}{n}n = c\log n$, δηλαδή της τάξεως $\mathcal{O}(\log n)$. Το ίδιο αποτέλεσμα λαμβάνεται για την κατεύθυνση του φανταστικού άξονα. Επομένως, $\mathcal{O}(\log n)$ ιδιοτιμές του προρρυθμισμένου πίνακα κυμαίνονται εκτός του ορθογωνίου $[a,b]\times[-c,c]$. □

Παρατήρηση. Αν δεν έχουν εκτιμηθεί ρίζες σε σημεία διαφορετικά του 0 ή αν όλες οι ρίζες έχουν εκτιμηθεί με ακρίβεια, τότε από το Θεώρημα 2.1.2, η συσσώρευση των ιδιοτιμών στο ορθογώνιο είναι κύρια. Επιπλέον, αν η γεννήτρια συνάρτηση του T_n είναι επαρκώς ομαλή ή κάποιο τριγωνομετρικό πολυώνυμο, περιπτώσεις όπου το σφάλμα προσέγγισης του αναπτύγματος Fourier είναι της τάξεως $\mathcal{O}\left(\frac{1}{n}\right)$ [71], ακολουθώντας ακριβώς την απόδειξη του Θεωρήματος 4.1.1, καταλήγουμε σε κύρια συσσώρευση των ιδιοτιμών του προρρυθμισμένου πίνακα.

Παρατήρηση. Όσον αφορά στη συσσώρευση των ιδιαζουσών τιμών, από το Θεώρημα 2.1.1, αυτή επιτυγχάνεται για τον προρρυθμισμένο πίνακα αν η f είναι γνωστή εκ των προτέρων και ανήκει στην κλάση $\mathrm{L}^2([-\pi,\pi])$. Στην περίπτωσή μας, όπου η συνάρτηση είναι άγνωστη η συσσώρευση των ιδιαζουσών τιμών εξακολουθεί να έχει την ίδια φύση (γενική συσσώρευση). Ωστόσο, αν υπάρχουν σφάλματα κατά την εκτίμηση των ριζών, αυτή γίνεται ακόμα χειρότερη. Επομένως, η σύγκλιση της μεθόδου PCGN, είναι πιο αργή από αυτή της μεθόδου PGMRES. Αυτό φαίνεται και στον Πίνακα 4.2 του Παραδείγματος 4.1.3.

Ο Αλγόριθμος 4.1.2 περιγράφει την κατασκευή του προτεινόμενου προρρυθμιστή, σε μορφή ψευδοκώδικα.

Αλγόριθμος 4.1.2 Κατασκευή του Προρρυθμιστή.

Είσοδος: $n \in \mathbb{N}$, T_n: $n \times n$ μη-συμμετρικός, πραγματικός πίνακας Toeplitz.

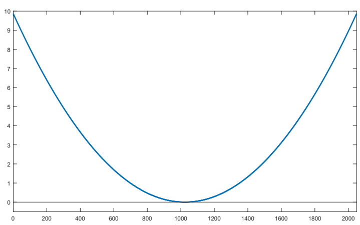
1: Κατασκευάστε το ισοκατανεμημένο πλέγμα G_n, με σημεία: $\theta_j = -\pi + \frac{2\pi j}{n+1}$, $j = 1, 2, \ldots, n$.
2: για $j = 1, 2, \ldots, n$
3: Υπολογίστε το ανάπτυγμα Fourier: $F_{n-1}(\theta_j) = \sum\limits_{k=-n+1}^{n-1} t_k \mathrm{e}^{\mathrm{i}k\theta_j}$, $\theta_j \in G_n$.
4: Εκτιμήστε τις f_1 και f_2 ως τις τιμές των $F_{n-1}^1(\theta_j) = \mathrm{Re}(F_{n-1}(\theta_j))$ και $F_{n-1}^2(\theta_j) = \mathrm{Im}(F_{n-1}(\theta_j))$, αντίστοιχα.
5: τέλος για
6: Επιλέξτε σημεία $\theta_i \in G_n$, κοντά στα τοπικά ελάχιστα της $|F_{n-1}^\ell|$, $\ell = 1, 2$, τέτοια ώστε $|F_{n-1}^\ell(\theta_i)| \simeq 0$ και θεωρήστε τα ως πιθανές ρίζες x_i, $i = 1, \ldots, \rho$.
7: Επιλέξτε διαστήματα $[\theta_j, \theta_{j+1}]$, όπου είναι πιθανόν να υπάρχουν σημεία ασυνέχειας του F_{n-1}.
8: Εκτιμήστε τις πολλαπλότητες των ριζών της f_ℓ, $\ell = 1, 2$:
9: αν F_{n-1}^ℓ λαμβάνει θετικές και αρνητικές τιμές
10: Υπολογίστε το $|F_{n-1}^\ell|$ στο G_n.
11: Υπολογίστε το $\widehat{S}_{4k}^\ell \simeq T_{4k}(|F_{n-1}^\ell|)$ με χρήση του σύνθετου κανόνα του Simpson, $k \ll n$.
12: Για κάθε ρίζα $x_i \in G_n$, εκτιμήστε την αντίστοιχη ιδιοτιμή $\lambda_{i,k}^\ell$ του $\widehat{S}_k^\ell$ χρησιμοποιώντας λίγες επαναλήψεις της μεθόδου Αντιστρόφων Δυνάμεων με αρχικό διάνυσμα $\Theta_{i,k} = \frac{1}{\sqrt{k}}\left(1, \mathrm{e}^{\mathrm{i}x_i}, \mathrm{e}^{2\mathrm{i}x_i}, \ldots, \mathrm{e}^{(k-1)\mathrm{i}x_i}\right)^T$.
13: Επαναλάβετε το 12 για τους $\widehat{S}_{2k}^\ell$ και $\widehat{S}_{4k}^\ell$ για να λάβετε τις $\lambda_{i,2k}^\ell$ και $\lambda_{i,4k}^\ell$, αντίστοιχα.
14: Υπολογίστε το λόγο $s_i^\ell = \frac{\lambda_{i,k}^\ell - \lambda_{i,2k}^\ell}{\lambda_{i,2k}^\ell - \lambda_{i,4k}^\ell}$ και τις πολλαπλότητες m_i^ℓ ως τους πλησιέστερους ακέραιους στον $\log_2 s_i^\ell$.
15: αλλιώς
16: Υπολογίστε το συμμετρικό και αντισυμμετρικό μέρος (του T_n), $S_n^1 = \frac{T_n + T_n^T}{2}$ και $S_n^2 = \frac{T_n - T_n^T}{2}$, αντίστοιχα.

17: Επαναλάβετε τα 12 - 14 για τον S_k^ℓ αντί του $\widehat{S}_k^\ell$.
18: τέλος αν
19: Επιλέξτε το τριγωνομετρικό πολυώνυμο g_n, ώστε $\mathrm{Re}\left(\frac{F_{n-1}(\theta_j)}{g_n(\theta_j)}\right) > 0$.
20: Ορίστε το G_k, υποσύνολο του G_n με k ισοκατανεμημένα σημεία στο $(0, \pi)$.
21: Εξαιρέστε τα σημεία του G_k τα οποία είναι κοντά σε ρίζες ή σημεία ασυνέχειας της f_ℓ, και ορίστε το νέο σύνολο ως G_k^ℓ, $\ell = 1, 2$.
22: Εκτιμήστε τις $\mathrm{Re}\left(\frac{f}{g}\right)$ και $\mathrm{Im}\left(\frac{f}{g}\right)$, ως $\widehat{f}_1$ και $\widehat{f}_2$, αντίστοιχα, υπολογισμένες στο G_k^ℓ, $\ell = 1, 2$.
23: Προσεγγίστε την $\widehat{f}_\ell$, με ένα κατάλληλο τριγωνομετρικό πολυώνυμο q_ℓ, με βέλτιστη ομοιόμορφη προσέγγιση, χρησιμοποιώντας ως κόμβους τα σημεία του G_k^ℓ, $\ell = 1, 2$.
24: Κατασκευάστε τον ταινιωτό προρρυθμιστή Toeplitz $T_n(p_n)$, όπου $p_n = g_n q$ και $q = q_1 + \mathrm{i} q_2$.

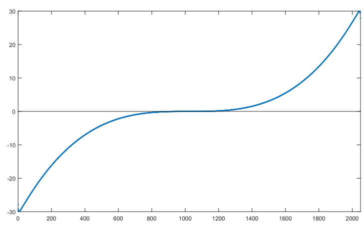
4.1.2 Αριθμητικά αποτελέσματα

Σε αυτή την υποενότητα παρουσιάζουμε μια πληθώρα αριθμητικών παραδειγμάτων, τα οποία δείχνουν την αποτελεσματικότητα της προτεινόμενης τεχνικής προρρύθμισης. Δίνουμε τις επαναλήψεις που χρειάζονται για την επιθυμητή σύγκλιση στη λύση του συστήματος, με σφάλμα το πολύ ίσο με 10^{-6}. Στους Πίνακες 4.2, 4.5 και 4.6 με I_n δηλώνουμε ότι δε χρησιμοποιήθηκε κανένας προρρυθμιστής.

Παράδειγμα 4.1.3. Παίρνοντας τον πίνακα Toeplitz ο οποίος προκύπτει από τη συνάρτηση $\mathfrak{f}_2(x) = x^2 + \mathrm{i}x^3$, θα εκτιμήσουμε τις ρίζες αυτής, καθώς και την πολλαπλότητα αυτών, από τα στοιχεία του πίνακα.



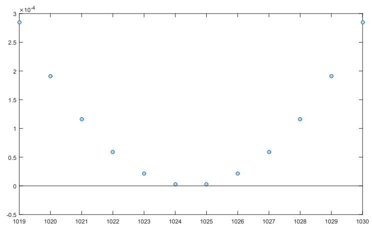
(α') Εκτίμηση της x^2.



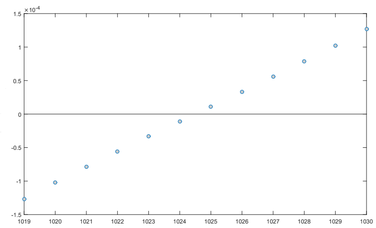
(β') Εκτίμηση της x^3.

Σχήμα 4.2: Ανάπτυγμα Fourier στο G_{2048}.

Τα Σχήματα 4.2 και 4.3 δείχνουν τις τιμές που λαμβάνει το ανάπτυγμα Fourier για το πραγματικό και φανταστικό μέρος της (άγνωστης) γεννήτριας συνάρτησης $\mathfrak{f}_2$.



(α') Εκτίμηση της x^2 (μεγ.).



(β') Εκτίμηση της x^3 (μεγ.).

Σχήμα 4.3: Ανάπτυγμα Fourier κοντά στην αρχή των αξόνων.

Εκτελώντας 4 επαναλήψεις της μεθόδου Αντιστρόφων Δυνάμεων καταλήγουμε στο ότι το πραγματικό μέρος της $\mathfrak{f}_2$ έχει μια ρίζα στο 0 με πολλαπλότητα $m_0^1 = 2$, ενώ το φανταστικό μέρος αυτής έχει ρίζα στο ίδιο σημείο με πολλαπλότητα $m_0^2 = 3$ (για περισσότερες λεπτομέρειες βλ. Πίνακα 4.1). Επομένως, εκτιμήσαμε ότι η $\mathfrak{f}_2$ έχει μια ρίζα στο 0 και $m_0^1 < m_0^2$. Στον Πίνακα 4.1, με $|\lambda_0^2|$ δηλώνουμε την ελάχιστη ιδιοτιμή του $\widehat{S}_k^2$ (υπολογισμένη με τη μέθοδο των Αντιστρόφων Δυνάμεων). Προκειμένου να άρουμε την κακή κατάσταση επιλέγουμε το τριγωνομετρικό πολυώνυμο $g_n = g = 2 - 2\cos(x)$, και προσεγγίζουμε την $\frac{F_{n-1}}{g}$ με τριγωνομετρικά πολυώνυμα τετάρτου βαθμού. Στη συνέχεια, κατασκευάζουμε τον προρρυθμιστή $R_{4,4}$ κι επιλύουμε το σύστημα με χρήση της μεθόδου PGMRES και PCGN. Οι επαναλήψεις των μεθόδων για διάφορες διαστάσεις δίνονται στον Πίνακα 4.2.

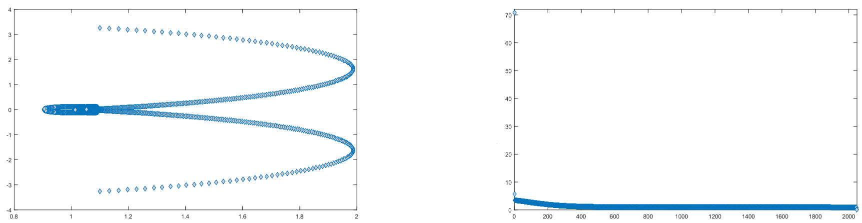
k	λ_0^1	$\log_2(s_0^1)$	$\|\lambda_0^2\|$	$\log_2(s_0^2)$
16	0.0351	1.9224	0.0133	2.6634
32	0.0092		0.0023	
64	0.0024		0.0005	

Πίνακας 4.1: Πολλαπλότητα των ριζών ($\mathfrak{f}_2$).

Σημειώνουμε ότι η λύση του συστήματος χωρίς προρρύθμιση λαμβάνεται σχεδόν μετά από n επαναλήψεις (n είναι η διάσταση του συστήματος). Στο Σχήμα 4.4 φαίνεται η συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών του προρρυθμισμένου συστήματος.

n	PGMRES I_n	PGMRES $R_{4,4}$	PCGN I_n	PCGN $R_{4,4}$
1024	>500	28	-	45
2048	>500	28	-	49
4096	>500	28	-	54
8192	>500	27	-	57

Πινακας 4.2: Επαναλήψεις ($\mathfrak{f}_2$).



(α') Ιδιοτιμές. (β') Ιδιάζουσες τιμές.

Σχημα 4.4: Ιδιοτιμές και ιδιάζουσες τιμές ($\mathfrak{f}_2$).

Παράδειγμα 4.1.4. Θα δώσουμε εν συντομία, ένα ανάλογο παράδειγμα για τον πίνακα Toeplitz, ο οποίος έχει ως γεννήτρια συνάρτηση την $\mathfrak{f}_3(x) = x^2 + \mathrm{i}x$. Εδώ, η ρίζα της $\mathfrak{f}_3$ είναι απλή. Το γεγονός αυτό επιβεβαιώθηκε δουλεύοντας ακριβώς με τον ίδιο τρόπο, όπως στο Παράδειγμα 4.1.3. Πιο συγκεκριμένα, καταλήξαμε στο ότι $m_0^1 = 2$ και $m_0^2 = 1$, αφού $\log_2(s_0^1) = 1.9224$ και $\log_2(s_0^2) = 0.9718$. Επομένως, η $\mathfrak{f}_3$ έχει μια ρίζα στο 0 και $m_0^1 > m_0^2$, που σημαίνει ότι επιλέγουμε ως g το τριγωνομετρικό πολυώνυμο $2 - 2\cos(x) + \mathrm{i}\sin(x)$. Η λύση του συστήματος λαμβάνεται μετά από μόλις 6 επαναλήψεις, όταν $n = 2048$, και 5 επαναλήψεις όταν $n = 4096$ και 8192, με χρήση της μεθόδου PGMRES και τον $R_{4,4}$ ως προρρυθμιστή. Αξίζει να αναφερθεί ότι η λύση του συστήματος λαμβάνεται στον ίδιο αριθμό επαναλήψεων με εκείνον της τεχνικής προρρύθμισης του δευτέρου κεφαλαίου, δηλαδή στην περίπτωση όπου η $\mathfrak{f}_3$ είναι γνωστή εκ των προτέρων. Αυτό πιθανότατα συμβαίνει διότι οι ρίζες έχουν εκτιμηθεί με ακρίβεια.

Παράδειγμα 4.1.5. Σε αυτό το παράδειγμα ασχολούμαστε με την επίλυση του συστήματος Toeplitz που προκύπτει από τη γεννήτρια συνάρτηση $\mathfrak{f}_9(x) = (x^2-1)^2 + \mathrm{i}x(x^2-1)$. Αυτή έχει ρίζες στο ± 1, σημεία τα οποία δεν ανήκουν στο πλέγμα G_n. Αν μπορούσαμε να εκτιμήσουμε τις τιμές της $\mathfrak{f}_9$ με ακρίβεια στο G_n, τότε θα αναμέναμε ένα σφάλμα, στην εκτίμηση της ρίζας, της τάξεως $\mathcal{O}\left(\frac{1}{n}\right)$. Ωστόσο, προσεγγίζουμε την $\mathfrak{f}_9$ μέσω του αναπτύγματος Fourier του T_n κι έτσι το σφάλμα εξαρτάται από τη φύση της (άγνωστης) $\mathfrak{f}_9$, δηλαδή το πόσο ομαλή είναι. Στο παράδειγμα μας αναμένουμε σφάλμα της τάξεως $\mathcal{O}\left(\frac{\log n}{n}\right)$, καθώς εντοπίζεται ασυνέχεια για το φανταστικό μέρος στο $\pm\pi$ [71]. Φυσικά, όσο μεγαλώνει η διάσταση του αρχικού συστήματος, τόσο ακριβέστερη γίνεται η εκτίμηση της ρίζας (βλ. Πίνακα 4.3).

n	x_1^1	x_1^2
2048	1.0012	0.9981
4096	1.0007	0.9991
8192	0.9996	0.9996

Πίνακας 4.3: Εκτίμηση της ρίζας ($\mathfrak{f}_9$).

Στον Πίνακα 4.4 παρουσιάζουμε τον αριθμό επαναλήψεων, χρησιμοποιώντας τη μέθοδο PGMRES και τον $R_{8,4}$, ως προρρυθμιστή, όταν $n = 2048$, για διάφορα υποθετικά σφάλματα στην εκτίμηση των ριζών. ‘Οπως είναι φυσικό, δίνουμε επίσης τον αριθμό επαναλήψεων για τις ρίζες που εκτιμήθηκαν μέσω της διαδικασίας που περιγράψαμε, όπου τα σφάλματα ήταν $\varepsilon_1 = 0.0012$ και $\varepsilon_2 = 0.0019$. Εκτελώντας 4 επαναλήψεις της μεθόδου Αντιστρόφων Δυνάμεων υπολογίζουμε ότι $\log_2(s_1^1) = 1.6780$, $\log_2(s_0^2) = 0.8882$ και $\log_2(s_1^2) = 1.0005$. Αυτό σημαίνει ότι $m_0^1 = 0$, $m_1^1 = 2$, $m_0^2 = 1$ και $m_1^2 = 1$. Επομένως, επιλέγουμε ως τριγωνομετρικό πολυώνυμο το $g_n(x) = (\cos(x_1^1) - \cos(x))^2 + \mathrm{i}\sin(x)(\cos(x_1^2) - \cos(x))$.

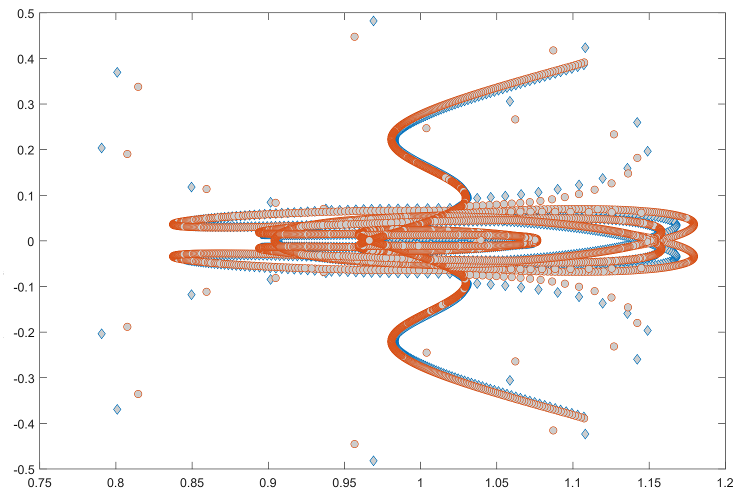
ε_1 \ ε_2	0	0.0001	0.0005	0.0010	0.0019	0.0020
0	7	7	8	9	12	13
0.0001	7	7	8	9	12	13
0.0005	7	7	8	9	12	13
0.0010	7	7	8	9	13	13
0.0012	7	7	8	9	13	13
0.0020	7	7	8	9	13	13

Πίνακας 4.4: Επαναλήψεις με υποθετικά σφάλματα ($\mathfrak{f}_9$).

Στον Πίνακα 4.4 παρατηρούμε ότι ο αριθμός επαναλήψεων είναι σχεδόν ο ίδιος κατά μήκος των στηλών. Αυτό σημαίνει ότι το σφάλμα ε_1 δεν παίζει τοσό καθοριστικό ρόλο για τη σύγκλιση του προρρυθμισμένου συστήματος, όσο το ε_2. Αυτό ισχύει διότι η πολλαπλότητα της ρίζας του φανταστικού μέρους της $\mathfrak{f}_9$ είναι μικρότερη από την πολλαπλότητα της αντίστοιχης ρίζας, του πραγματικού μέρους της $\mathfrak{f}_9$. Μπορούμε να εξηγήσουμε αυτό το φαινόμενο αναλύοντας την $\frac{\mathfrak{f}_9}{g_n}$.

$$\begin{aligned}\frac{\mathfrak{f}_9}{g_n} &= \frac{(x^2-1)^2 + \mathrm{i}x(x^2-1)}{(\cos(x_1^1)-\cos(x))^2 + \mathrm{i}\sin(x)(\cos(x_1^2)-\cos(x))} \\ &= \frac{h_1(\cos(1)-\cos(x))^2 + \mathrm{i}h_2\sin(x)(\cos(1)-\cos(x))}{(\cos(x_1^1)-\cos(x))^2 + \mathrm{i}\sin(x)(\cos(x_1^2)-\cos(x))}.\end{aligned}$$

Λαμβάνοντας υπόψη ότι η μη-μηδενική ρίζα έχει εκτιμηθεί σε διαφορετικά σημεία x_1^1 και x_1^2 για το πραγματικό και φανταστικό μέρος, αντίστοιχα, η συνάρτηση $\frac{\mathfrak{f}_9}{g_n}$ δεν έχει πόλο, αλλά λαμβάνει πολύ μεγάλη τιμή για $x = x_1^1$ και $x = x_1^2$. Μελετώντας τον παρονομαστή της συνάρτησης σε μια περιοχή του 1, που περιέχει τα σημεία x_1^1 και x_1^2, παρατηρούμε ότι ο δεύτερος όρος $\sin(x)(\cos(x_1^2)-\cos(x))$ κυριαρχεί επί του πρώτου $(\cos(x_1^1)-\cos(x))^2$, στην προαναφερθείσα περιοχή, εκτός από μια μικρότερη σε μέγεθος περιοχή του x_1^2, με μήκος της τάξεως $\mathcal{O}\left(\frac{1}{n^2}\right)$, όπου ο πρώτος όρος υπερταιρεί επί του δεύτερου. Επομένως, το σφάλμα ε_1 της εκτίμησης του x_1^1 έχει επιρροή σε μια μικρή περιοχή της τάξεως $\mathcal{O}\left(\frac{1}{n^2}\right)$, αλλά η τεχνική μας είναι κατασκευασμένη στο πλέγμα G_n με βήμα $\frac{2\pi}{n+1}$ κι αυτό σημαίνει ότι μια τόσο μικρή περιοχή δε μπορεί να ανιχνευθεί. Με άλλα λόγια το ε_1 δεν παίζει και τόσο καθοριστικό ρόλο στην αύξηση των επαναλήψεων.



Σχημα 4.5: Ιδιοτιμές ($\mathfrak{f}_9$).

Στο σχήμα 4.5 παρουσιάζουμε τη συσσώρευση των ιδιοτιμών, για το προρρυθμισμένο σύστημα, όταν $n = 2048$ (μπλε διαμάντια) και $n = 4096$ (πορτοκαλί κύκλοι). Παρατηρούμε ότι οι ιδιοτιμές μακριά από το $(1, 0)$, οι οποίες εμφανίζονται ως διακεκριμένα/απομονωμένα σημεία, σχηματίζουν ζεύγη (μπλε διαμαντιών-πορτοκαλί κύκλων), τα οποία κυμαίνονται εκτός της συσσώρευσης. Το γεγονός αυτό εξηγεί τη συσσώρευση με το πολύ $\mathcal{O}(\log n)$ ιδιοτιμές εκτός του φάσματος. Σημειώστε ότι ο ρυθμός αύξησης της τάξεως $\mathcal{O}(\log n)$, είναι ιδιαιτέρως αργός, όταν διπλασιάζεται η διάσταση n. Αυτό φαίνεται επίσης στον Πίνακα 4.5, όπου οι επαναλήψεις που δίνονται μέσω της μεθόδου PGMRES δεν αυξάνονται όταν διπλασιάζουμε το n.

n	I_n	$R_{8,4}$
2048	>500	13
4096	>500	12
8192	>500	12

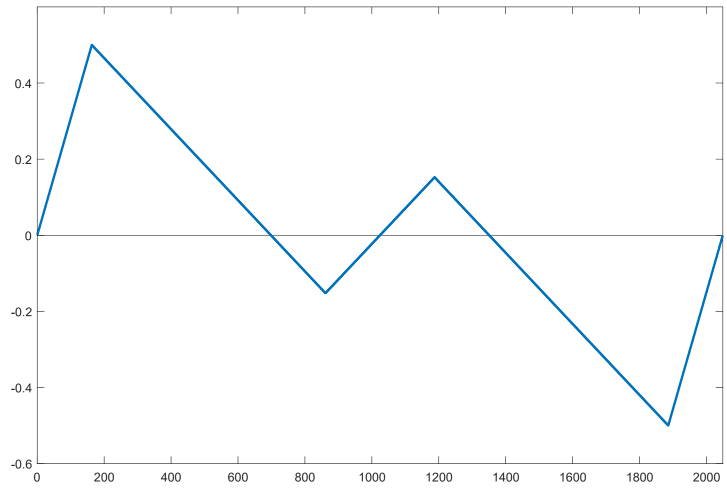
Πι🚧νακας 4.5: Επαναλήψεις ($\mathfrak{f}_9$).

Παράδειγμα 4.1.6. Ως το τελευταίο παράδειγμα αυτής της υποενότητας, παρουσιάζουμε την επίλυση ενός συστήματος Toeplitz, το οποίο γεννάται από μια συνεχή συνάρτηση που έχει ρίζες στο ± 1 με πολλαπλότητες $m_1^1 = m_1^2 = 1$. Πιο συγκεκριμένα, ο πίνακας Toeplitz γεννάται από την $\mathfrak{f}_{10}(x) = (x^2 - 1) + \mathrm{i}\mathfrak{h}_3(x)$, όπου $\mathfrak{h}_3(x)$ είναι η τεθλασμένη γραμμή, ορισμένη ως:

$$\mathfrak{h}_3(x) = \begin{cases} x + \pi, & x \in [-\pi, -\pi + \frac{1}{2}) \\ \frac{x}{-2\pi+3} + \frac{1}{-2\pi+3}, & x \in [-\pi + \frac{1}{2}, -\frac{1}{2}) \\ \frac{x}{2\pi-3}, & x \in [-\frac{1}{2}, \frac{1}{2}) \\ \frac{x}{-2\pi+3} - \frac{1}{-2\pi+3}, & x \in [\frac{1}{2}, \pi - \frac{1}{2}) \\ x - \pi, & x \in [\pi - \frac{1}{2}, \pi] \end{cases}$$

Το φανταστικό μέρος του αναπτύγματος Fourier της $\mathfrak{h}_3$, υπολογισμένο στο πλέγμα G_n όταν $n = 2048$, δίνεται στο Σχήμα 4.6.

Θεωρώντας τη γεννήτρια συνάρτηση ως άγνωστη και χρησιμοποιώντας τον προτεινόμενο αλγόριθμο, εκτιμήσαμε τις μη-μηδενικές ρίζες στα (ίδια) σημεία $x_1^1 = x_1^2$ και μετά από 3 επαναλήψεις της μεθόδου Αντίστροφων Δυνάμεων, συμπεράναμε ότι $m_1^1 = m_1^2 = 1$ και $m_0^2 = 1$. Επομένως, το τριγωνομετρικό πολυώνυμο το οποίο αίρει τις εκτιμώμενες ρίζες της $\mathfrak{f}_{10}$, δίνεται ως $g_n(x) = \cos(x_1^1) - \cos(x)$. Στον Πίνακα 4.6 παρουσιάζουμε τον αριθμό επαναλήψεων, με χρήση της PGMRES όταν δε



Σχήμα 4.6: Ανάπτυγμα Fourier της $\mathfrak{h}_3$.

n	I_n	$R_{4,4}$
2048	>500	11
4096	>500	12
8192	>500	12

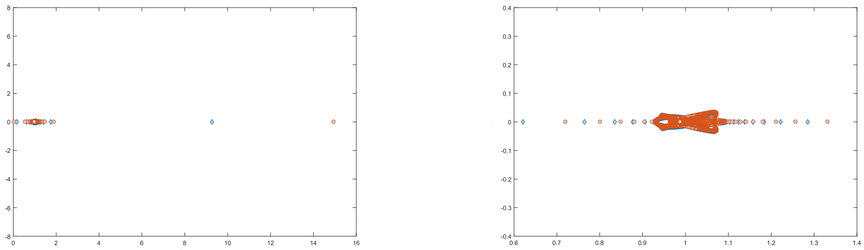
Πίνακας 4.6: Επαναλήψεις ($\mathfrak{f}_{10}$).

χρησιμοποιούμε κάποιον προρρυθμιστή, καθώς κι όταν γίνεται προρρύθμιση με τον $R_{4,4}$.

Στο Σχήμα 4.7, παρουσιάζουμε τη συσσώρευση των ιδιοτιμών του προρρυθμισμένου συστήματος.

4.2 Κυκλοειδείς προρρυθμιστές

Σε αυτή την υποενότητα αρχικά θα περιγράψουμε τον τρόπο με τον οποίο κατασκευάζουμε τον προτεινόμενο προρρυθμιστή και στη συνέχεια θα δώσουμε έναν αλγόριθμο για την κατασκευή αυτού, σε μορφή ψευδοκώδικα. Ο προτεινόμενος προρρυθμιστής λαμβάνει δύο μορφές. Αυτή του κυκλοειδούς πίνακα, για συστήματα με καλή κατάσταση (η γεννήτρια συνάρτηση δεν έχει ρίζες), καθώς και αυτή



(α') Ιδιοτιμές. (β') Ιδιοτιμές κοντά στο $(1, 0)$.

Σχήμα 4.7: Ιδιοτιμές ($\mathfrak{f}_{10}$).

του ταινιωτού-επί-κυκλοειδή πίνακα, για συστήματα με κακή κατάσταση (η γεννήτρια συνάρτηση έχει ρίζες). Αυτομάτως καταλαβαίνει κανείς ότι το πρώτο βήμα είναι να μελετήσουμε αν η γεννήτρια συνάρτηση έχει ρίζες ή όχι. Εφόσον βρεθούν ρίζες, προφανώς θα πρέπει να προσεγγιστούν. Το ίδιο θα πρέπει να γίνει και για τις πολλαπλότητες των ριζών, έτσι ώστε να βρεθεί το κατάλληλο τριγωνομετρικό πολυώνυμο που οδηγεί στην άρση της κακής κατάστασης.

4.2.1 Κατασκευή του προρρυθμιστή

'Οπως και στην προηγούμενη ενότητα, όπου προτάθηκε η χρήση ταινιωτών προρρυθμιστών, έτσι κι εδώ αρχικά θα προσεγγίσουμε τη γεννήτρια συνάρτηση του πίνακα Toeplitz, με τη βοήθεια του αναπτύγματος Fourier, F_{n-1}, στο ισοκατανεμημένο πλέγμα G_n, το οποίο έχει ως σημεία τα $\theta_j = \frac{2(j-1)\pi}{n}$, $j = 1, \ldots, n$. Επιλέξαμε αυτό το πλέγμα, το οποίο διαφέρει από το πλέγμα της προηγούμενης ενότητας (βλ. επίσης [17]), διότι τα σημεία αυτού είναι και σημεία όπου λαμβάνονται οι ιδιοτιμές του κυκλοειδή προρρυθμιστή που προσπαθούμε να κατασκευάσουμε, όπως φαίνεται στη σχέση (3.2).

Οι τιμές του αναπτύγματος F_{n-1}, θα μας υποδείξουν και τα σημεία πιθανών ριζών της γεννήτριας συνάρτησης, σε περίπτωση που το σύστημα είναι κακής κατάστασης. Αυτά επιλέγονται με τον τρόπο που προαναφέραμε στην υποενότητα 4.1.1, δηλαδή επιλέγοντας τα τοπικά ελάχιστα των $|F^1_{n-1}(\theta_j)|$ και $|F^2_{n-1}(\theta_j)|$, τα οποία λαμβάνουν τιμή πολύ κοντά στο 0. Σημειώνουμε για ακόμα μία φορά ότι σε περίπτωση που $|F^\ell_{n-1}(\theta_j)|$ και $|F^\ell_{n-1}(\theta_{j+1})|$, $\ell = 1, 2$ έχουν την ίδια τιμή για κάποιο j, η οποία είναι κοντά στο 0, υποθέτουμε ότι υπάρχει ρίζα στο $\frac{\theta_j + \theta_{j+1}}{2}$.

Παρατήρηση. Αν το πραγματικό μέρος του αναπτύγματος Fourier, F^1_{n-1} δεν έχει ρίζες, δεν προχωρούμε στην εκτίμηση των πιθανών ριζών του F_{n-1}^2 στο G_n, διότι ο προτεινόμενος προρρυθμιστής θα έχει τη μορφή κυκλοειδούς πίνακα, που προκύπτει από τις τιμές της F_{n-1}. Αυτό σημαίνει ότι δε γίνεται άρση των ριζών, όπως και στην περίπτωση όπου η γεννήτρια συνάρτηση είναι γνωστή εκ των προτέρων (βλ. υπενότητα 3.1.1).

Παρατήρηση. Γνωρίζουμε ότι το φανταστικό μέρος του αναπτύγματος Fourier, F_{n-1}^2, έχει ρίζα στο $\pm\pi$, ως περιττό τριγωνομετρικό πολυώνυμο. Προκειμένου να εξετάσουμε αν η f_2 έχει όντως ρίζα στο $\pm\pi$, θα ελέγξουμε τις τιμές της F_{n-1}^2 σε μια περιοχή του π (στο πλέγμα G_n). Αν αυτές είναι κοντά στο 0, θεωρούμε ότι υπάρχει ρίζα στο $\pm\pi$, διαφορετικά το σημείο π είναι σημείο ασυνέχειας (όπως προφανώς και το $-\pi$).

Για την άρση των ριζών της γεννήτριας συνάρτησης χρειαζόμαστε, όπως και στην προηγούμενη ενότητα, τις πολλαπλότητες m_i^1 και m_i^2, των ριζών της f_1 και f_2, αντίστοιχα, για το σημείο x_i, $i = 1, 2, \ldots, \rho$. Υπενθυμίζουμε ότι m_0^1 και m_0^2 δηλώνουν τις αντίστοιχες πολλαπλότητες στο 0. Η εκτίμηση των πολλαπλοτήτων περιγράφηκε στην υποενότητα 4.1.1. Θα θέλαμε να αναφέρουμε ότι στη βιβλιογραφία μπορούν να βρεθούν κι άλλες ενδιαφέρουσες τεχνικές για την προσέγγιση των ιδιοτιμών, όπως για παράδειγμα οι [7, 25, 26, 27].

Για την περίπτωση όπου η f_1 έχει ρίζες οι οποίες τέμνουν τον άξονα, το αντίστοιχο ολοκλήρωμα της (4.2) είναι το:

$$\frac{1}{2\pi}\int_{0}^{2\pi} \left|F_{n-1}^1(x)\right| \mathrm{e}^{-\mathrm{i}(r-q)x}\mathrm{d}x,$$

το οποίο μπορεί να υπολογιστεί από τον σύνθετο κανόνα του Simpson στο $G_n \cup \{2\pi\}$.

Το βήμα που έπεται της εκτίμησης των πολλαπλοτήτων είναι η εύρεση κατάλληλου τριγωνομετρικού πολυωνύμου, το οποίο αίρει τις ρίζες της f. Προφανώς, λόγω του ότι η f_1 είναι άρτια συνάρτηση και η f_2 περιττή, αν x_i είναι ρίζα στο $(0, \pi]$, $-x_i$ θα είναι ρίζα στο $(-\pi, 0)$, με την ίδια πολλαπλότητα. Παρατηρούμε ότι τα σημεία του πλέγματος, της προτεινόμενης τεχνικής προρρύθμισης, ανήκουν στο $[0, 2\pi)$. Ωστόσο χρησιμοποιούμε το διάστημα $(-\pi, \pi]$ για την κατασκευή του τριγωνομετρικού πολυωνύμου, λόγω της προφανούς αντιστοιχίας μεταξύ των σημείων του $(-\pi, 0)$ και $(\pi, 2\pi)$ (μετατόπιση κατά 2π). Το τριγωνομετρικό πολυώνυμο δίνεται όπως περιγράφεται στην υποενότητα 4.1.1 (βλ. επίσης υποενότητα 2.2.2).

Παρατήρηση. Προφανώς υπάρχει περίπτωση οι F_{n-1}^1 και F_{n-1}^2 να έχουν ρίζες σε διαφορετικά σημεία. Τότε, η γεννήτρια συνάρτηση δεν έχει καμία ρίζα και δεν προχωρούμε με την άρση των ριζών. Έτσι, ο προτεινόμενος προρρυθμιστής είναι ο κυκλοειδής πίνακας $C_n(F_{n-1})$. Αυτό αποτελεί μια διαφορά ανάμεσα στην προτεινόμενη τεχνική προρρύθμισης και σε αυτήν της προηγούμενης ενότητας, που είχε να κάνει με ταινιωτούς προρρυθμιστές. Εκεί, υπενθυμίζουμε ότι η άρση των ριζών είναι απαραίτητη σε περίπτωση που η F_{n-1}^1 έχει κάποια ρίζα.

Ο Αλγόριθμος 4.2.1 περιγράφει την κατασκευή του προτεινόμενου προρρυθμιστή, σε μορφή ψευδοκώδικα.

Αλγόριθμος 4.2.1 Κατασκευή του Προρρυθμιστή.

Είσοδος: $n \in \mathbb{N}$, T_n: $n \times n$ μη-συμμετρικός, πραγματικός πίνακας Toeplitz.

1: Κατασκευάστε το ισοκατανεμημένο πλέγμα G_n, με σημεία: $\theta_j = \frac{2(j-1)\pi}{n}$, $j = 1, 2, \ldots, n$.
2: για $j = 1, 2, \ldots, n$
3: Υπολογίστε το ανάπτυγμα Fourier: $F_{n-1}(\theta_j) = \sum_{k=-n+1}^{n-1} t_k \mathrm{e}^{\mathrm{i}k\theta_j}$, $\theta_j \in G_n$.
4: Εκτιμήστε τις f_1 και f_2 ως τις τιμές των $F_{n-1}^1(\theta_j) = \mathrm{Re}(F_{n-1}(\theta_j))$ και $F_{n-1}^2(\theta_j) = \mathrm{Im}(F_{n-1}(\theta_j))$, αντίστοιχα.
5: τέλος για
6: Επιλέξτε σημεία $\theta_i \in G_n$, κοντά στα τοπικά ελάχιστα της $|F_{n-1}^1|$, τέτοια ώστε $|F_{n-1}^1(\theta_i)| \simeq 0$ και θεωρήστε τα ως πιθανές ρίζες της $|F_{n-1}^1|$.
7: αν δεν έχει επιλεχθεί κανένα σημείο ως ρίζα τότε
8: Θέστε $g_n = 1$ και πηγαίνετε στο 33.
9: αλλιώς
10: Επιλέξτε σημεία $\theta_i \in G_n$, κοντά στα τοπικά ελάχιστα της $|F_{n-1}^2|$, τέτοια ώστε $|F_{n-1}^2(\theta_i)| \simeq 0$ και θέστε τα ως πιθανές ρίζες της $|F_{n-1}^2|$.
11: τέλος αν
12: Σχηματίστε το σύνολο πιθανών ριζών $\{x_i,\ i = 1, \ldots, \rho\}$ ως την ένωση των επιλεγμένων ριζών των F_{n-1}^1 και F_{n-1}^2.
13: αν δεν υπάρχει κοινή ρίζα για τις F_{n-1}^1 και F_{n-1}^2 τότε
14: Θέστε $g_n = 1$ και πηγαίνετε στο 33.
15: τέλος αν
16: Υπολογίστε το συμμετρικό και αντι-συμμετρικό μέρος (του T_n), $S_n^1 = \frac{T_n + T_n^T}{2}$ και $S_n^2 = \frac{T_n - T_n^T}{2}$, αντίστοιχα.
17: για $\ell = 1, 2$
18: αν το F_{n-1}^ℓ παίρνει θετικές και αρνητικές τιμές τότε
19: Υπολογίστε την $|F_{n-1}^\ell|$ στο G_n.

20: Υπολογίστε το $\widehat{S}^{\ell}_{4k} \simeq T_{4k}(|F^{\ell}_{n-1}|)$ με χρήση του σύνθετου κανόνα του Simpson, $k \ll n$.
21: αλλιώς
22: Θέστε $\widehat{S}^{\ell}_{4k} = S^{\ell}_{4k}$.
23: τέλος αν
24: για $j = 1, 2, \ldots, \rho$
25: αν x_i είναι μια ρίζα του F^{ℓ}_{n-1} τότε
26: Προσεγγίστε την ιδιοτιμή $\lambda^{\ell}_{i,k}$ του $\widehat{S}^{\ell}_{k}$ ως $\widetilde{\lambda}^{\ell}_{i,k}$ με λίγες επαναλήψεις της μεθόδου Αντιστρόφων Δυνάμεων, με αρχικό διάνυσμα: $\Theta_{i,k} = \frac{1}{\sqrt{k}}\left(1, \mathrm{e}^{\mathrm{i}x_i}, \mathrm{e}^{2\mathrm{i}x_i}, \ldots, \mathrm{e}^{(k-1)\mathrm{i}x_i}\right)^T$.
27: Επαναλάβετε το 26 για $\widehat{S}^{\ell}_{2k}$ και $\widehat{S}^{\ell}_{4k}$ ώστε να λάβετε τις $\widetilde{\lambda}^{\ell}_{i,2k}$ και $\widetilde{\lambda}^{\ell}_{i,4k}$, αντίστοιχα.
28: Υπολογίστε το $\widetilde{s_i}^{\ell} = \frac{\widetilde{\lambda}^{\ell}_{i,k} - \widetilde{\lambda}^{\ell}_{i,2k}}{\widetilde{\lambda}^{\ell}_{i,2k} - \widetilde{\lambda}^{\ell}_{i,4k}}$ για να εκτιμήσετε την πολλαπλότητα m^{ℓ}_i ως τον πλησιέστερο ακέραιο στον $\log_2 \widetilde{s}^{\ell}_i$.
29: τέλος αν
30: τέλος για
31: τέλος για
32: Επιλέξτε το τριγωνομετρικό πολυώνυμο g_n, ώστε $\mathrm{Re}\left(\frac{F_{n-1}(\theta_j)}{g_n(\theta_j)}\right) > 0$.
33: Κατασκευάστε τον κυκλοειδή προρρυθμιστή $C_n\left(\frac{F_{n-1}}{g_n}\right)$, ο οποίος έχει ως ιδιοτιμές, τις τιμές της $\frac{F_{n-1}}{g_n}$ στα σημεία του G_n.
34: Κατασκευάστε τον ταινιωτό-επί-κυκλοειδή προρρυθμιστή $T_n(g_n)C_n\left(\frac{F_{n-1}}{g_n}\right)$.

Στην κατασκευή του ταινιωτού προρρυθμιστή, για μη-συμμετρικά συστήματος Toeplitz με άγνωστη γεννήτρια συνάρτηση, που περιγράψαμε παραπάνω, δώσαμε έναν τρόπο εκτίμησης πιθανών σημείων ασυνέχειας, διότι μας ενδιέφερε να μειώσουμε το σφάλμα προσέγγισης του αλγορίθμου Remez. Αυτή η εκτίμηση δεν είναι απαραίτητη για τον προτεινόμενο προρρυθμιστή, αφού ο κυκλοειδής προρρυθμιστής κατασκευάζεται από τις τιμές της $\frac{F_{n-1}}{g_n}$, στα σημεία του πλέγματος G_n. Σημειώνουμε ότι ο λόγος $\frac{F_{n-1}}{g_n}$ έχει πόλο στα σημεία του G_n, τα οποία είναι πιθανές ρίζες. Για την κατασκευή του $C_n\left(\frac{F_{n-1}}{g_n}\right)$, αντικαθιστούμε την τιμή $\frac{F_{n-1}(\theta_i)}{g_n(\theta_i)}$ με $\frac{F_{n-1}(\theta_{i+1})}{g_n(\theta_{i+1})}$ ή $\frac{F_{n-1}(\theta_{i-1})}{g_n(\theta_{i-1})}$. Μπορούμε επίσης να αντικαταστήσουμε με τον μέσο όρο των δύο τελευταίων όρων. Αυτή η τεχνική μετατόπισης χρησιμοποιήθηκε για συμμετρικά συστήματα Toeplitz στη [15].

4.2.2 Θεωρητικά αποτελέσματα

Αρχικά, θα μελετήσουμε τη συσσώρευση των ιδιοτιμών για το προρρυθμισμένο σύστημα, όταν η f δεν έχει ρίζες και είτε είναι επαρκώς ομαλή, είτε απλώς συνεχής. Η περίπτωση όπου η f έχει σημεία ασυνέχειας θα καλυφθεί στη συνέχεια. Αναφέρουμε ότι στην [71], ο συγγραφέας μελέτησε την προρρύθμιση συμμετρικών συστημάτων Toeplitz με άγνωστη γεννήτρια συνάρτηση και απέδειξε ιδιότητες για διάφορες περιπτώσεις όπου η γεννήτρια συνάρτηση είναι άγνωστη και συνεχώς παραγωγίσιμη ή συνεχής.

Θεώρημα 4.2.2. Έστω T_n ένας πραγματικός πίνακας Toeplitz, η γεννήτρια συνάρτηση του οποίου υπάρχει και είναι άγνωστη. Υποθέτουμε επίσης ότι δεν έχουν εντοπιστεί ρίζες, μέσω της προτεινόμενης τεχνικής. Τότε, οι ιδιοτιμές του προρρυθμισμένου πίνακα $C_n^{-1}(F_{n-1})T_n$ συσσωρεύονται γύρω από το σημείο $(1,0)$ του μιγαδικού επιπέδου και η συσσώρευση χαρακτηρίζεται ως:

1. Κύρια σε μια περιοχή του $(1,0)$, με ακτίνα της τάξεως $\mathcal{O}\left(\frac{1}{n}\right)$, αν η f είναι επαρκώς ομαλή (συνεχώς παραγωγίσιμη).

2. Κύρια σε μια περιοχή του $(1,0)$, με ακτίνα της τάξεως $\mathcal{O}\left(\frac{\log n}{n}\right)$, αν η f είναι συνεχής.

Απο📖δειξη. Ο προρρυθμισμένος πίνακας γράφεται ως:

$$\begin{aligned} C_n^{-1}(F_{n-1})T_n &= C_n^{-1}(F_{n-1})C_n(f)C_n^{-1}(f)T_n = C_n\left(\frac{f}{F_{n-1}}\right)C_n^{-1}(f)T_n \\ &= \left(I_n + C_n\left(\frac{f-F_{n-1}}{F_{n-1}}\right)\right)C_n^{-1}(f)T_n \\ &= C_n^{-1}(f)T_n + C_n\left(\frac{f-F_{n-1}}{fF_{n-1}}\right)T_n. \end{aligned} \tag{4.5}$$

Από το Θεώρημα 3.2.6 έχουμε ότι οι ιδιοτιμές του πρώτου όρου, του παραπάνω αθροίσματος, έχουν κύρια συσσώρευση στο σημείο $(1,0)$ του μιγαδικού επιπέδου αν η f είναι συνεχής.

Στη συνέχεια μελετάμε τον δεύτερο όρο του αθροίσματος της (4.5), παίρνοντας τη νόρμα $\|\cdot\|_2$ αυτού και χρησιμοποιώντας το Λήμμα 1 της [14] και το Λήμμα 3.2.1:

$$\begin{aligned}\left\|C_n\left(\frac{f-F_{n-1}}{fF_{n-1}}\right)T_n\right\|_2 &\leq \left\|C_n\left(\frac{f-F_{n-1}}{fF_{n-1}}\right)\right\|_2 \|T_n\|_2 \\ &\leq 2\left\|\frac{f-F_{n-1}}{fF_{n-1}}\right\|_\infty 2\|f\|_\infty = 4\max\left|\frac{f-F_{n-1}}{fF_{n-1}}\right|\max|f| \\ &\leq 4\frac{\max|f|}{\min|fF_{n-1}|}\max|f-F_{n-1}| \leq c\max|f-F_{n-1}|.\end{aligned}$$

Στην περίπτωση 1 όπου η f θεωρείται συνεχώς παραγωγίσιμη, από την [71] έχουμε ότι $\max|f-F_{n-1}| = \mathcal{O}\left(\frac{1}{n}\right)$. Για τις υποθέσεις της περίπτωσης 2, έχουμε επίσης από την [71] ότι $\max|f-F_{n-1}| = \mathcal{O}\left(\frac{\log n}{n}\right)$.

Για να μελετήσουμε τη συσσώρευση των ιδιοτιμών του προρρυθμισμένου πίνακα $C_n^{-1}(F_{n-1})T_n$, χρησιμοποιούμε το min-max θεώρημα των Courant-Fischer για το συμμετρικό και αντι-συμμετρικό του μέρος. Προχωρούμε με την ανάλυση του συμμετρικού μέρους. 'Οσον αφορά στο αντι-συμμετρικό, αυτή είναι ανάλογη.

'Εστω $A_n = \frac{C_n^{-1}(f)T_n+T_n^TC_n^{-1}(\bar{f})}{2}$ το συμμετρικό μέρος του πρώτου όρου του αθροίσματος στην (4.5) και $B_n = \frac{C_n\left(\frac{f-F_{n-1}}{fF_{n-1}}\right)T_n+T_n^TC_n\left(\frac{\bar{f}-\bar{F}_{n-1}}{\bar{f}\bar{F}_{n-1}}\right)}{2}$ το συμμετρικό μέρος του δευτέρου όρου της ίδιας σχέσης. Τότε, το συμμετρικό μέρος του προρρυθμισμένου πίνακα γράφεται ως $S_n = A_n + B_n$. 'Εστω ότι με λ_k και $\widetilde{\lambda}_k$, $k = 1,2,\ldots,n$ συμβολίζουμε της ιδιοτιμές του A_n και S_n, αντίστοιχα, ταξινομημένες σε μη-αύξουσα σειρά: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ και $\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \cdots \geq \widetilde{\lambda}_n$. Τότε, από το min-max θεώρημα των Courant-Fischer, έχουμε:

$$\begin{aligned}\widetilde{\lambda}_k &= \min_{V\in\mathbb{R}^{n+1-k}}\max_{x\in V}\frac{x^H(A_n+B_n)x}{x^Hx} \leq \max_{x\in W}\frac{x^H(A_n+B_n)x}{x^Hx} \\ &\leq \max_{x\in W}\frac{x^HA_nx}{x^Hx} + \max_{x\in W}\frac{x^HB_nx}{x^Hx} \leq \lambda_k + \|B_n\|_2,\end{aligned}$$

όπου W είναι ο χώρος που επιτυγχάνεται το μέγιστο του $\frac{x^HA_nx}{x^Hx}$. Από την άλλη:

$$\begin{aligned}\widetilde{\lambda}_k &= \max_{V\in\mathbb{R}^k}\min_{x\in V}\frac{x^H(A_n+B_n)x}{x^Hx} \geq \min_{x\in\widetilde{W}}\frac{x^H(A_n+B_n)x}{x^Hx} \\ &\geq \min_{x\in\widetilde{W}}\frac{x^HA_nx}{x^Hx} + \min_{x\in\widetilde{W}}\frac{x^HB_nx}{x^Hx} \geq \lambda_k - \max_{x\in\widetilde{W}}\left|\frac{x^HB_nx}{x^Hx}\right| \\ &\geq \lambda_k - \max_{x\in\mathbb{R}^n}\left|\frac{x^HB_nx}{x^Hx}\right| = \lambda_k - \|B_n\|_2,\end{aligned}$$

όπου $\widetilde{W}$ είναι ο χώρος στον οποίο επιτυγχάνεται το ελάχιστο του $\frac{x^H A_n x}{x^H x}$. Επομένως, για κάθε ιδιοτιμή λ_k του A_n, οι αντίστοιχη ιδιοτιμή $\widetilde{\lambda}_k$ του S_n κυμαίνεται στο διάστημα $[\lambda_k - \|B_n\|_2, \lambda_k + \|B_n\|_2]$. Αυτό σημαίνει ότι η συσσώρευση της ακολουθίας πινάκων $\{S_n\}$ είναι μια $\|B_n\|_2$-επέκταση της συσσώρευσης του $\{A_n\}$.

Η ίδια ανάλυση για τα αντι-συμμετρικά μέρη $A'_n = \frac{C_n^{-1}(f)T_n - T_n^T C_n^{-1}(\bar{f})}{2}$ και $B'_n = \frac{C_n\left(\frac{f-F_{n-1}}{fF_{n-1}}\right)T_n - T_n^T C_n\left(\frac{\bar{f}-\bar{F}_{n-1}}{\bar{f}\bar{F}_{n-1}}\right)}{2}$, μας δίνει ότι για κάθε ιδιοτιμή μ_k του A'_n, η αντίστοιχη ιδιοτιμή $\widetilde{\mu}_k$ του $S'_n = A'_n + B'_n$, κυμαίνεται στο διάστημα $[\mu_k - \|B'_n\|_2, \mu_k + \|B'_n\|_2]$. Η μελέτη του αντι-συμμετρικού μέρους μας οδηγεί στον ίδιο τύπο συσσώρευσης γύρω από το 0.

Λόγω της κύριας συσσώρευσης του πρώτου όρου, οι Ερμιτιανές και αντι-Ερμιτιανές ακολουθίες πινάκων $\{A_n\}$ και $\{A'_n\}$, αντίστοιχα, παρουσιάζουν κύρια συσσώρευση στο 1 και στο 0, αντίστοιχα. Η ιδιότητα που αναφέραμε ισχύει διότι $C_n^{-1}(f)T_n = I_n + S_n + R_n$, όπου S_n είναι πίνακας με μικρή νόρμα και R_n είναι πίνακας χαμηλής βαθμίδας (βλ. Πόρισμα 3.2.3). Επειδή το πραγματικό/φανταστικό μέρος των ιδιοτιμών ενός πίνακα βρίσκεται εντός του εύρους του Ερμιτιανού/αντι-Ερμιτιανού μέρους του πίνακα [4, 37], το αποτέλεσμα προκύπτει από το Λήμμα 1.1.6 (Λήμμα 2.1 της [87]), όπου ως $\mathcal{A}_n$ επιλέξαμε το $C_n^{-1}(f)T_n - I_n$ και ως $\mathcal{B}_n$ το S_n. Επομένως, για κάθε $\varepsilon > 0$, υπάρχουν ακέραιοι $M, M' > 0$, τέτοιοι ώστε M ιδιοτιμές του A_n να κυμαίνονται εκτός του διαστήματος $(1-\varepsilon, 1+\varepsilon)$ και M' ιδιοτιμές του A'_n, εκτός του $(-\varepsilon, \varepsilon)$. Υποθέτουμε ότι οι πρώτες M_1 $(\lambda_1, \lambda_2, \dots \lambda_{M_1})$ και οι τελευταίες M_2 $(\lambda_{n-M_2+1}, \dots, \lambda_n)$ ιδιοτιμές, όπου $M_1 + M_2 = M$, κυμαίνονται εκτός της συσσώρευσης του A_n, ενώ οι πρώτες M'_1 $(\mu_1, \mu_2, \dots \mu_{M'_1})$ και οι τελευταίες M'_2 $(\mu_{n-M'_2+1}, \dots, \mu_n)$, με $M'_1 + M'_2 = M'$, κυμαίνονται εκτός της συσσώρευσης του A'_n. Οι A_n και A'_n γράφονται ως:

$$A_n = \sum_{k=1}^{n} \lambda_k x_k x_k^H, \text{ και } A'_n = \sum_{k=1}^{n} \mu_k y_k y_k^H,$$

όπου x_k, $k = 1, 2, \dots, n$ είναι τα κανονικοποιημένα ιδιοδιανύσματα του A_n, τα οποία σχηματίζουν ορθοκανονική βάση και y_k, $k = 1, 2, \dots, n$ είναι τα αντίστοιχα ιδιοδιανύσματα του A'_n. Χωρίζουμε τους A_n και A'_n ως $A_n = \widetilde{A}_n + \widehat{A}_n$ και $A'_n = \widetilde{A}'_n + \widehat{A}'_n$, αντίστοιχα, όπου:

$$\widetilde{A}_n = \sum_{k=M_1+1}^{n-M_2} \lambda_k x_k x_k^H, \; \widehat{A}_n = \sum_{k=1}^{M_1} \lambda_k x_k x_k^H + \sum_{k=n-M_2+1}^{n} \lambda_k x_k x_k^H \text{ και}$$
$$\widetilde{A}'_n = \sum_{k=M'_1+1}^{n-M'_2} \mu_k y_k y_k^H, \; \widehat{A}'_n = \sum_{k=1}^{M'_1} \mu_k y_k y_k^H + \sum_{k=n-M'_2+1}^{n} \mu_k y_k y_k^H.$$

Είναι προφανές ότι $\|\widetilde{A}_n - I_n\|_2 \le \varepsilon$ και $\|\widetilde{A}'_n\|_2 \le \varepsilon$, ενώ οι $\widehat{A}_n$ και $\widehat{A}'_n$ είναι πίνακες με χαμηλή βαθμίδα.

Χρησιμοποιούμε το Λήμμα 1.1.6, επιλέγοντας $\mathcal{A}_n = S_n + S'_n - I_n = A_n + A'_n + B_n + B'_n - I_n$ και $\mathcal{B}_n = \widetilde{A}_n + \widetilde{A}'_n + B_n + B'_n - I_n$. Έχουμε ότι $\|\mathcal{B}_n\|_2 \le \|\widetilde{A}_n - I_n\|_2 + \|\widetilde{A}'_n\|_2 + \|B_n\|_2 + \|B'_n\|_2$ και $\|\mathcal{A}_n - \mathcal{B}_n\|_F^2 = \|\widehat{A}_n + \widehat{A}'_n\|_F^2 \le c$, αφού $\widehat{A}_n + \widehat{A}'_n$ είναι πίνακας χαμηλής βαθμίδας, σταθερής και ανεξάρτητης της διάστασης n. Επομένως, οι ιδιοτιμές του $\mathcal{A}_n$ συσσωρεύονται με την έννοια της κύριας συσσώρευσης στον δίσκο $\{z : \|z\| \le 2\varepsilon + \|B_n\|_2 + \|B'_n\|_2\}$, για κάθε $\varepsilon > 0$.

Στην περίπτωση 1 (επαρκώς ομαλή συνάρτηση) έχουμε ότι $\|B_n\|_2 + \|B'_n\|_2 = \mathcal{O}\left(\frac{1}{n}\right)$, ενώ στην περίπτωση 2 (συνεχής συνάρτηση), $\|B_n\|_2 + \|B'_n\|_2 = \mathcal{O}\left(\frac{\log n}{n}\right)$. Επομένως, η προρρυθμισμένη ακολουθία πινάκων $\{\mathcal{A}_n + I_n\}$ παρουσιάζει κύρια συσσώρευση των ιδιοτιμών σε μια περιοχή του $(1,0)$, με ακτίνα $\mathcal{O}\left(\frac{1}{n}\right)$ και $\mathcal{O}\left(\frac{\log n}{n}\right)$, στην περίπτωση 1 και 2, αντίστοιχα κι έτσι η απόδειξη ολοκληρώνεται. □

Για να μελετήσουμε την περίπτωση όπου η f έχει σημεία ασυνέχειας, ή την περίπτωση όπου η f έχει ρίζες, αρχικά θα πρέπει να αποδείξουμε κάποιες ιδιότητες που αφορούν στις ιδιοτιμές και ιδιάζουσες τιμές, γινομένων ακολουθιών πινάκων, οι οποίες θα μας φανούν χρήσιμες.

Λήμμα 4.2.3. Έστω $\{\mathcal{A}_n\}$ και $\{\mathcal{B}_n\}$ ακολουθίες πινάκων, των οποίων τα Ερμιτιανά μέρη παρουσιάζουν κύρια συσσώρευση των ιδιοτιμών στο 1 και τα αντι-Ερμιτιανά μέρη αυτών παρουσιάζουν κύρια συσσώρευση των ιδιοτιμών στο 0. Τότε, η ακολουθία $\{\mathcal{C}_n\}$, όπου $\mathcal{C}_n = \mathcal{A}_n\mathcal{B}_n$, έχει κύρια συσσώρευση των ιδιοτιμών στο σημείο $(1,0)$, του μιγαδικού επιπέδου.

Απόδειξη. Το Ερμιτιανό μέρος του $\mathcal{A}_n$, για κάθε $\varepsilon_A > 0$ γράφεται ως $\frac{\mathcal{A}_n + \mathcal{A}_n^H}{2} = I_n + S_n + R_n$, όπου S_n είναι ένας πίνακας με μικρή νόρμα $\|S_n\|_2 \le \varepsilon_A$ και R_n είναι ένας πίνακας χαμηλής βαθμίδας $\mathrm{rank}(R_n) = k$. Το αντι-Ερμιτιανό μέρος του $\mathcal{A}_n$, για κάθε $\varepsilon'_A > 0$ γράφεται ως $\frac{\mathcal{A}_n - \mathcal{A}_n^H}{2} = S'_n + R'_n$, όπου $\|S'_n\|_2 \le \varepsilon'_A$ και $\mathrm{rank}(R'_n) = k'$. Ανάλογες ιδιότητες ισχύουν επίσης και για τον $\mathcal{B}_n$. Αυτό σημαίνει ότι, για κάθε $\varepsilon_B > 0$, $\frac{\mathcal{B}_n + \mathcal{B}_n^H}{2} = I_n + Q_n + P_n$, όπου $\|Q_n\|_2 \le \varepsilon_B$ και $\mathrm{rank}(P_n) = l$, καθώς επίσης και για κάθε $\varepsilon'_B > 0$, $\frac{\mathcal{B}_n - \mathcal{B}_n^H}{2} = Q'_n + P'_n$, όπου $\|Q'_n\|_2 \le \varepsilon'_B$ και $\mathrm{rank}(P'_n) = l'$. Καταλήγουμε στο ότι $\mathcal{A}_n = I_n + S_n + S'_n + R_n + R'_n$ και $\mathcal{B}_n = I_n + Q_n + Q'_n + P_n + P'_n$. Επομένως:

$$\mathcal{C}_n = \mathcal{A}_n\mathcal{B}_n = (I_n + S_n + S'_n + R_n + R'_n)(I_n + Q_n + Q'_n + P_n + P'_n) = I_n + \mathcal{S}_n + \mathcal{R}_n,$$

όπου:

$$\mathcal{S}_n = S_n + S'_n + Q_n + Q'_n + S_nQ_n + S_nQ'_n + S'_nQ_n + S'_nQ'_n \text{ και}$$

$$\mathcal{R}_n = (R_n + R'_n)(I_n + Q_n + Q'_n + P_n + P'_n) + (I_n + S_n + S'_n)(P_n + P'_n).$$

Είναι προφανές ότι ο $\mathcal{S}_n$ είναι πίνακας με μικρή νόρμα $\|\mathcal{S}_n\|_2 \leq \varepsilon$, για κάθε $\varepsilon > 0$ (ε είναι επιλεγμένο ως $\varepsilon_A + \varepsilon'_A + \varepsilon_B + \varepsilon'_B + \varepsilon_A\varepsilon_B + \varepsilon_A\varepsilon'_B + \varepsilon'_A\varepsilon_B + \varepsilon'_A\varepsilon'_B$). Επίσης ισχύει ότι $\mathrm{rank}(\mathcal{R}_n) \leq (k + k' + l + l')$, το οποίο σημαίνει ότι ο $\mathcal{R}_n$ είναι πίνακας χαμηλής βαθμίδας. Χρησιμοποιώντας το Λήμμα 1.1.6, θέτοντας ως τον $\mathcal{A}_n$ του Λήμματος 1.1.6 τον $\mathcal{S}_n + \mathcal{R}_n$ και ως τον $\mathcal{B}_n$ του ίδιου λήμματος, τον $\mathcal{S}_n$, προκύπτει το ζητούμενο αποτέλεσμα, επειδή έχουμε ότι $\|\mathcal{R}_n\|_F^2 = \mathcal{O}(1)$ κι επειδή $\|\mathcal{S}_n\|_2 \leq \varepsilon$, για κάθε $\varepsilon > 0$ αρκούντως μικρό, οι ιδιοτιμές του $\mathcal{S}_n + \mathcal{R}_n$ συσσωρεύονται με την έννοια της κύριας συσσώρευσης στον κλειστό δίσκο με κέντρο το $(0, 0)$ και ακτίνα ε. Συνεπώς, οι ιδιοτιμές του $\mathcal{C}_n = I_n + \mathcal{S}_n + \mathcal{R}_n$ έχουν κύρια συσσώρευση στο $(1, 0)$. □

Λήμμα 4.2.4. Έστω $\{\mathcal{A}_n\}$ και $\{\mathcal{B}_n\}$ ακολουθίες πινάκων, των οποίων τα Ερμιτιανά μέρη παρουσιάζουν γενική συσσώρευση των ιδιοτιμών στα διαστήματα $(1-r_A, 1+r_A)$ και $(1-r_B, 1+r_B)$, με $s_A(n)$ και $s_B(n)$ ιδιοτιμές εκτός των διαστημάτων, αντίστοιχα, ενώ τα αντι-Ερμιτιανά μέρη αυτών παρουσιάζουν γενική συσσώρευση των ιδιοτιμών στα διαστήματα $(-r_A, r_A)$ και $(-r_B, r_B)$, με $s_A(n)$ και $s_B(n)$ ιδιοτιμές εκτός των διαστημάτων, αντίστοιχα. Τότε, η ακολουθία $\{\mathcal{C}_n\}$, όπου $\mathcal{C}_n = \mathcal{A}_n\mathcal{B}_n$, έχει γενική συσσώρευση των ιδιοτιμών σε μια περιοχή του $(1, 0)$, του μιγαδικού επιπέδου, με ακτίνα $2r_A + 2r_B + 4r_Ar_B$ και το πολύ $2s_A(n) + 2s_B(n)$ ιδιοτιμές εκτός της συσσώρευσης.

Απόδειξη. Όπως και στην απόδειξη του Λήμματος 4.2.3, το Ερμιτιανό μέρος του $\mathcal{A}_n$ μπορεί να γραφεί ως $\frac{\mathcal{A}_n+\mathcal{A}_n^H}{2} = I_n + S_n + R_n$, όπου $\|S_n\|_2 \leq r_A$ και $\mathrm{rank}(R_n) = s_A(n)$. Το αντι-Ερμιτιανό μέρος μπορεί να γραφεί ως $\frac{\mathcal{A}_n-\mathcal{A}_n^H}{2} = S'_n + R'_n$, όπου $\|S'_n\|_2 \leq r_A$ και $\mathrm{rank}(R'_n) = s_A(n)$. Το ίδιο ισχύει και για τον $\mathcal{B}_n$, δηλαδή $\frac{\mathcal{B}_n+\mathcal{B}_n^H}{2} = I_n + Q_n + P_n$, όπου $\|Q_n\|_2 \leq r_B$ και $\mathrm{rank}(P_n) = s_B(n)$, καθώς επίσης και $\frac{\mathcal{B}_n-\mathcal{B}_n^H}{2} = Q'_n + P'_n$, όπου $\|Q'_n\|_2 \leq r_B$ και $\mathrm{rank}(P'_n) = s_B(n)$. Τότε, $\mathcal{A}_n = I_n + S_n + S'_n + R_n + R'_n$, ενώ $\mathcal{B}_n = I_n + Q_n + Q'_n + P_n + P'_n$. Επομένως:

$$\begin{aligned}
\mathcal{C}_n = \mathcal{A}_n\mathcal{B}_n &= (I_n + S_n + S'_n + R_n + R'_n)(I_n + Q_n + Q'_n + P_n + P'_n) \\
&= I_n + S_n + S'_n + (I_n + S_n + S'_n)(Q_n + Q'_n) \\
&\qquad + (R_n + R'_n)(I_n + Q_n + Q'_n + P_n + P'_n) + (I_n + S_n + S'_n)(P_n + P'_n) \\
&= I_n + \mathcal{S}_n + \mathcal{R}_n.
\end{aligned}$$

Είναι προφανές ότι ο $\mathcal{S}_n$ έχει φραγμένη νόρμα:

$$\begin{aligned}\|\mathcal{S}_n\|_2 &= \|S_n + S_n' + (I_n + S_n + S_n')(Q_n + Q_n')\|_2 \\ &\le \|S_n\|_2 + \|S_n'\|_2 + (1 + \|S_n\|_2 + \|S_n'\|_2)(\|Q_n\|_2 + \|Q_n'\|_2) \\ &\le 2r_A + 2r_B + 4r_A r_B,\end{aligned}$$

ενώ ο $\mathcal{R}_n$ είναι πίνακας χαμηλής βαθμίδας $\operatorname{rank}(\mathcal{R}_n) \le 2s_A(n) + 2s_B(n)$.

Χρησιμοποιούμε το Λήμμα 1.1.6, θέτοντας ως $\mathcal{A}_n$ του Λήμματος 1.1.6, τον $\mathcal{S}_n + \mathcal{R}_n$ και ως $\mathcal{B}_n$ του ίδιου λήμματος, τον $\mathcal{S}_n$. Επειδή $\operatorname{rank}(\mathcal{R}_n) = \mathcal{O}(2s_A(n) + 2s_B(n))$, λαμβάνουμε ότι $\|\mathcal{R}_n\|_F^2 = \mathcal{O}(2s_A(n) + 2s_B(n))$ κι επίσης ισχύει ότι $\|\mathcal{S}_n\|_2 \le 2r_A + 2r_B + 4r_A r_B$. Επομένως, οι ιδιοτιμές του $\mathcal{S}_n + \mathcal{R}_n$ συσσωρεύονται στον κλειστό δίσκο με κέντρο το $(0, 0)$ και ακτίνα $2r_A + 2r_B + 4r_A r_B$, με την έννοια της γενικής συσσώρευσης, με $\mathcal{O}(2s_A(n) + 2s_B(n))$, ιδιοτιμές εκτός της συσσώρευσης. Συμπεραίνουμε ότι οι ιδιοτιμές του $\mathcal{C}_n = I_n + \mathcal{S}_n + \mathcal{R}_n$ έχουν γενική συσσώρευση στον κλειστό δίσκο με κέντρο το $(1, 0)$ και ακτίνα $2r_A + 2r_B + 4r_A r_B$, καθώς επίσης και $\mathcal{O}(2s_A(n) + 2s_B(n))$ ιδιοτιμές εκτός της συσσώρευσης. $\square$

Στο Λήμμα 4.2.4, θεωρήσαμε τις ίδιες ακτίνες r_A και r_B, για τα Ερμιτιανά και αντι-Ερμιτιανά μέρη των $\mathcal{A}_n$ και $\mathcal{B}_n$, αντίστοιχα, καθώς επίσης και τον ίδιο αριθμό ιδιοτιμών εκτός της συσσώρευσης, $s_A(n)$ και $s_B(n)$. Η απόδειξη είναι ανάλογη αν θεωρήσουμε διαφορετικές τιμές, αλλά η τάξη μεγέθους παραμένει η ίδια.

Παρατήρηση. Πρέπει να σχολιάσουμε ότι αν οι ακτίνες r_A και r_B διαφέρουν κατά τάξη μεγέθους, τότε η ακτίνα του κύκλου θα πρέπει να είναι αυτή με τη μεγαλύτερη τάξη. Αν επιπλέον οι $s_A(n)$ και $s_B(n)$ διαφέρουν επίσης κατά τάξη μεγέθους, ο αριθμός των ιδιοτιμών που βρίσκονται εκτός της συσσώρευσης, θα εξαρτάται ομοίως από τη μεγαλύτερη τάξη. Αξίζει να σημειώσουμε επίσης ότι το Λήμμα 4.2.4 περιέχει το αποτέλεσμα του Λήμματος 4.2.3, θέτοντας $r_A = r_B = 0$ και $s_A(n) = c_1$, $s_B(n) = c_2$, όπου c_1 και c_2 είναι σταθερές ανεξάρτητες της διάστασης n.

Προχωρούμε με τη μελέτη των ιδιοτιμών, όταν η f είναι ασυνεχής και δεν έχει ρίζες.

Θεώρημα 4.2.5. Έστω T_n ένας πραγματικός πίνακας Toeplitz, η γεννήτρια συνάρτηση του οποίου υπάρχει και είναι άγνωστη. Υποθέτουμε επίσης ότι δεν έχουν βρεθεί ρίζες μέσω της προτεινόμενης μεθόδου. Αν η f έχει πεπερασμένα σημεία ασυνέχειας, οι ιδιοτιμές του προρρυθμισμένου πίνακα $C_n^{-1}(F_{n-1})T_n$ συσσωρεύονται, με την έννοια της γενικής συσσώρευσης, σε μια περιοχή του $(1, 0)$, με σταθερή ακτίνα, το πολύ ίση με $0.179/(1 - 0.179)$ και με $\mathcal{O}(\log n)$ ιδιοτιμές εκτός της συσσώρευσης.

Απόδειξη. 'Οπως έχει ήδη αποδειχθεί στη σχέση (4.5), ο προρρυθμισμένος πίνακας γράφεται ως:

$$C_n^{-1}(F_{n-1})T_n = C_n\left(\frac{f}{F_{n-1}}\right)C_n^{-1}(f)T_n.$$

Από το Θεώρημα 3.2.8 λαμβάνουμε τη γενική συσσώρευση των ιδιοτιμών του $C_n^{-1}(f)T_n$, στο $(1,0)$ με $\mathcal{O}(\log n)$ ιδιοτιμές εκτός της συσσώρευσης. Το Ερμιτιανό μέρος του πίνακα μπορεί να γραφεί ως $I_n + H(S_n) + H(R_n)$, όπου $H(S_n) = \frac{S_n+S_n^H}{2}$ και $H(R_n) = \frac{R_n+R_n^H}{2}$. Προφανώς, $\|H(S_n)\|_2 \leq \varepsilon$ και $\|H(R_n)\|_F^2 = \mathcal{O}(\log n)$. Ανάλογες ιδιότητες ισχύουν και για το αντι-Ερμιτιανό του μέρος κι έτσι καταλήγουμε στην ίδια κατά τάξη μεγέθους γενική συσσώρευση του Ερμιτιανού και αντι-Ερμιτιανού μέρους του πίνακα $C_n^{-1}(f)T_n$.

Οι ιδιοτιμές του πρώτου όρου $C_n\left(\frac{f}{F_{n-1}}\right)$ είναι οι τιμές της συνάρτησης $\frac{f}{F_{n-1}}$ στα σημεία του πλέγματος G_n. Γνωρίζουμε ότι το ανάπτυγμα Fourier συγκλίνει ομοιόμορφα, όσο το n τείνει στο άπειρο, με ταχύτητα σύγκλισης $\mathcal{O}\left(\frac{\log n}{n}\right)$, σε οποιοδήποτε συμπαγές υποσύνολο του $\mathbb{R}$ που δεν περιέχει σημεία ασυνέχειας. Ωστόσο, σε μικρές περιοχές των σημείων ασυνέχειας, εμφανίζεται το φαινόμενο Gibbs [35], όπου το σχετικό σφάλμα συγκλίνει σε κάποιον σταθερό αριθμό, σχεδόν ίσο με 17.9% της τιμής της f [6]. 'Εχουμε:

$$C_n\left(\frac{f}{F_{n-1}}\right) = C_n\left(1+\frac{f-F_{n-1}}{F_{n-1}}\right) = I_n + C_n\left(\frac{f-F_{n-1}}{F_{n-1}}\right)$$

και ισχύει ότι:

$$\begin{aligned}
\left\|C_n\left(\frac{f-F_{n-1}}{F_{n-1}}\right)\right\|_2 &= \left(\lambda_{\max}\left(C_n\left(\frac{f-F_{n-1}}{F_{n-1}}\right)^H C_n\left(\frac{f-F_{n-1}}{F_{n-1}}\right)\right)\right)^{\frac{1}{2}}\\
&= \left(\lambda_{\max}\left(C_n\left(\left|\frac{f-F_{n-1}}{F_{n-1}}\right|^2\right)\right)\right)^{\frac{1}{2}}\\
&\leq \left(\max_x\left|\frac{f(x)-F_{n-1}(x)}{F_{n-1}(x)}\right|^2\right)^{\frac{1}{2}} \leq \left\|\frac{f-F_{n-1}}{F_{n-1}}\right\|_\infty.
\end{aligned}$$

Γράφουμε το ανάπτυγμα Fourier ως $F_{n-1} = F_{n-1}^1 + \mathrm{i}F_{n-1}^2$. Ισχύει:

$$\begin{aligned}
&F_{n-1}^1 = f_1 + \epsilon_1 f_1 = (1+\epsilon_1)f_1 \text{ και}\\
&F_{n-1}^2 = f_2 + \epsilon_2 f_2 = (1+\epsilon_2)f_2,
\end{aligned}$$

όπου ϵ_1, ϵ_2 είναι οι αντίστοιχες συναρτήσεις σφάλματος. Έτσι:

$$\begin{aligned}\left\| C_n\left(\frac{f-F_{n-1}}{F_{n-1}}\right)\right\|_2 \leq \left\|\frac{f-F_{n-1}}{F_{n-1}}\right\|_\infty &= \max_x \left|\frac{f(x)-F_{n-1}(x)}{F_{n-1}(x)}\right| \\ &= \max_x \frac{\left|f_1(x)-F^1_{n-1}(x)+\mathrm{i}\left(f_2(x)-F^2_{n-1}(x)\right)\right|}{\left|F^1_{n-1}(x)+\mathrm{i}F^2_{n-1}(x)\right|} \\ &= \max_x \frac{|\epsilon_1(x)f_1(x)+\mathrm{i}\epsilon_2(x)f_2(x)|}{|(1+\epsilon_1(x))f_1(x)+\mathrm{i}(1+\epsilon_2(x))f_2(x)|}.\end{aligned}$$

Είναι προφανές ότι το παραπάνω μέγιστο εμφανίζεται σε μικρές περιοχές των σημείων ασυνέχειας όπου το φαινόμενο Gibbs λαμβάνει χώρα. Έστω y το σημείο (σε κάποια μικρή περιοχή σημείου ασυνέχειας) που δίνει τη μέγιστη τιμή. Τότε:

$$\begin{aligned}\left\| C_n\left(\frac{f-F_{n-1}}{F_{n-1}}\right)\right\|_2 &\leq \max_x \frac{|\epsilon_1(x)f_1(x)+\mathrm{i}\epsilon_2(x)f_2(x)|}{|(1+\epsilon_1(x))f_1(x)+\mathrm{i}(1+\epsilon_2(x))f_2(x)|} \\ &= \frac{\left(\epsilon_1^2(y)f_1^2(y)+\epsilon_2^2(y)f_2^2(y)\right)^{\frac{1}{2}}}{\left((1+\epsilon_1(y))^2f_1^2(y)+(1+\epsilon_2(y))^2f_2^2(y)\right)^{\frac{1}{2}}} \\ &\leq \frac{0.179\left(f_1^2(y)+f_2^2(y)\right)^{\frac{1}{2}}}{(1-0.179)\left(f_1^2(y)+f_2^2(y)\right)^{\frac{1}{2}}} = \frac{0.179}{1-0.179}.\end{aligned}$$

Υπολογίζοντας τη νόρμα του γινομένου $\mathcal{A}_n\mathcal{B}_n$, όπως στο Λήμμα 4.2.4 με $\mathcal{A}_n = C_n\left(\frac{f}{F_{n-1}}\right)$ και $\mathcal{B}_n = C_n^{-1}(f)T_n$, και λαμβάνοντας υπόψη ότι $r_A = \frac{0.179}{1-0.179}$ και $r_B = 0$, προκύπτει ότι $\left\|C_n\left(\frac{f}{F_{n-1}}\right)C_n^{-1}(f)T_n\right\|_2 \leq r_A = \frac{0.179}{1-0.179} \simeq 0.218$. Επομένως, ο προρρυθμισμένος πίνακας έχει γενική συσσώρευση των ιδιοτιμών, σε μια περιοχή με κέντρο το $(1,0)$, ακτίνα το πολύ ίση με 0.218 και $\mathcal{O}(\log n)$ ιδιοτιμές εκτός της συσσώρευσης. □

Συνεχίζουμε με τη μελέτη των ιδιοτιμών και ιδιαζουσών τιμών, στην περίπτωση που η f έχει ρίζες, δηλαδή για συστήματα με κακή κατάσταση.

Θεώρημα 4.2.6. Έστω T_n πραγματικός πίνακας Toeplitz, στον οποίο αντιστοιχεί μια άγνωστη και μιγαδική γεννήτρια συνάρτηση f, με ρίζες στο διάστημα $(-\pi,\pi]$. Έστω F_{n-1} το ανάπτυγμα Fourier του T_n και g_n το τριγωνομετρικό πολυώνυμο, το οποίο αίρει τις εκτιμώμενες ρίζες μέσω της τεχνικής που περιγράφηκε. Υποθέτουμε ότι οι εκτιμώμενες μη-μηδενικές ρίζες $\widetilde{x}_i$, $i=1,2,\ldots,\rho$ έχουν ένα σφάλμα $\varepsilon_i = \widetilde{x}_i - x_i$ και ότι οι πολλαπλότητες αυτών έχουν εκτιμηθεί ακριβώς. Τότε, ο προρρυθμισμένος πίνακας $C_n^{-1}\left(\frac{F_{n-1}}{g_n}\right)T_n^{-1}(g_n)T_n$ έχει:

1. Γενική συσσώρευση των ιδιοτιμών σε μια περιοχή του $(1,0)$, με ακτίνα $o(1)$ και $o(n)$ ιδιοτιμές εκτός της συσσώρευσης, αν η f είναι συνεχής.

2. Γενική συσσώρευση των ιδιοτιμών σε μια περιοχή του $(1,0)$, με ακτίνα σχεδόν ίση με $0.179/(1-0.179)$ και $o(n)$ ιδιοτιμές εκτός της συσσώρευσης, αν η f είναι κατά τμήματα συνεχής.

Απόδειξη. Ξεκινάμε με την απόδειξη της περίπτωσης 1. Για απλούστευση, υποθέτουμε ότι οι εκτιμώμενες ρίζες είναι οι $\pm\widetilde{x}_1 \neq 0$ με πολλαπλότητα m. Η ανάλυση για περισσότερα ζεύγη ριζών, αποτελεί μια απλή γενίκευση, όπως θα γίνει εύκολα κατανοητό από την απόδειξη παρακάτω.

Αρχικά υποθέτουμε ότι η πολλαπλότητα της ρίζας αντιστοιχεί στο πραγματικό μέρος της f. Επομένως, χρησιμοποιώντας την προτεινόμενη τεχνική, το τριγωνομετρικό πολυώνυμο είναι το:

$$g_n(x) = (\cos\widetilde{x}_1 - \cos x)^m = \left(2\sin\frac{x+\widetilde{x}_1}{2}\sin\frac{x-\widetilde{x}_1}{2}\right)^m .$$

Έτσι, ο προρρυθμισμένος πίνακας γράφεται ως:

$$\begin{aligned} P_n &= C_n^{-1}\left(\frac{F_{n-1}}{g_n}\right)T_n^{-1}(g_n)T_n(f) \\ &= C_n^{-1}\left(\frac{F_{n-1}}{g_n}\right)T_n^{-1}(g_n)\left(T_n(g)T_n\left(\frac{f}{g}\right)+L\right), \end{aligned} \tag{4.6}$$

όπου L είναι πίνακας χαμηλής βαθμίδας, η οποία εξαρτάται από το εύρος ταινίας του $T_n(g)$.

Θα μελετήσουμε τον πίνακα $P_n' = C_n^{-1}\left(\frac{F_{n-1}}{g_n}\right)T_n^{-1}(g_n)T_n(g)T_n\left(\frac{f}{g}\right)$, ο οποίος διαφέρει από τον P_n, κατά πίνακα χαμηλής βαθμίδας. Αυτός είναι όμοιος με τον:

$$\begin{aligned} \widetilde{P}_n &= C_n\left(\frac{f}{g}\right)P_n'C_n^{-1}\left(\frac{f}{g}\right) \\ &= C_n\left(\frac{fg_n}{F_{n-1}g}\right)T_n^{-1}(g_n)T_n(g)T_n\left(\frac{f}{g}\right)C_n^{-1}\left(\frac{f}{g}\right). \end{aligned} \tag{4.7}$$

Χωρίζουμε αυτό το γινόμενο πινάκων στους εξής παράγοντες:
$C_n\left(\frac{fg_n}{F_{n-1}g}\right)$, $T_n^{-1}(g_n)T_n(g)$ και $T_n\left(\frac{f}{g}\right)C_n^{-1}\left(\frac{f}{g}\right)$. Ο τελευταίος εξ αυτών είναι όμοιως με τον προρρυθμισμένο πίνακα που αντιστοιχεί στην περίπτωση συστημάτων με καλή κατάσταση. Επομένως, αν η f είναι συνεχής, έχει κύρια συσσώρευση των ιδιοτιμών του στο $(1,0)$. Σημειώνουμε επίσης ότι το Ερμιτιανό και αντι-Ερμιτιανό

του μέρος έχει το ίδιο είδος συσσώρευσης, όπως εύκολα μπορεί να δει κανείς στην απόδειξη του Θεωρήματος 4.2.2.

Πρέπει να μελετήσουμε και τη συσσώρευση των δύο παραγόντων που απομένουν. Ο πρώτος γράφεται ως:

$$\begin{aligned} C_n\left(\frac{fg_n}{F_{n-1}g}\right) &= C_n\left(\frac{(F_{n-1}+\varepsilon_f)(g+\varepsilon_g)}{F_{n-1}g}\right) \\ &= I_n + C_n\left(\frac{\varepsilon_f}{F_{n-1}}\right) + C_n\left(\frac{\varepsilon_g}{g}\right) + C_n\left(\frac{\varepsilon_f\varepsilon_g}{F_{n-1}g}\right). \end{aligned} \tag{4.8}$$

Το σφάλμα $\varepsilon_f = f - F_{n-1}$ του αναπτύγματος Fourier εξαρτάται από την ομαλότητα της συνάρτησης f, ας πούμε αν η f είναι συνεχώς παραγωγίσιμη, το σφάλμα είναι τάξεως $\mathcal{O}\left(\frac{1}{n}\right)$, ενώ αν η f είναι απλά συνεχής, είναι τάξεως $\mathcal{O}\left(\frac{\log n}{n}\right)$. Αν η f είναι επαρκώς ομαλή, μπορούμε να λάβουμε, κατά τάξη μεγέθους, μικρότερο σφάλμα ε_f. Παρατηρούμε ότι σε οποιαδήποτε από τις παραπάνω περιπτώσεις το ε_f τείνει προς το 0, όσο το n τείνει στο άπειρο, δηλαδή $\varepsilon_f = o(1)$. Θεωρούμε τα διαστήματα με κέντρο τα σημεία των εκτιμώμενων ριζών $-\widetilde{x}_1$ και $\widetilde{x}_1$, $(-\widetilde{x}_1 - s_n, -\widetilde{x}_1 + s_n)$ και $(\widetilde{x}_1 - s_n, \widetilde{x}_1 + s_n)$, αντίστοιχα, με ακτίνα s_n η οποία εξαρτάται μεν από το n, ισχύει δε ότι $s_n = o(1)$, ή ισοδύναμα $\lim\limits_{n\to\infty} \frac{s_n}{1} = 0$ (ο κλασικός ασυμπτωτικός ορισμός του o). Η συνάρτηση $\frac{\varepsilon_f}{F_{n-1}}$ στο σημείο $\widetilde{x}_1 + s_n$ μας δίνει $\frac{\varepsilon_f(\widetilde{x}_1+s_n)}{F_{n-1}(\widetilde{x}_1+s_n)} = \frac{\varepsilon_f(\widetilde{x}_1+s_n)}{f(\widetilde{x}_1+s_n)-\varepsilon_f(\widetilde{x}_1+s_n)}$. Υποθέτουμε ότι $\varepsilon_f = o(s_n^m)$, ισοδύναμα $\lim\limits_{n\to\infty} \frac{\varepsilon_f}{s_n^m} = 0$. Τότε, ο παραπάνω λόγος προσεγγίζεται ως:

$$\frac{\varepsilon_f(\widetilde{x}_1+s_n)}{F_{n-1}(\widetilde{x}_1+s_n)} \simeq \frac{\varepsilon_f(\widetilde{x}_1+s_n)}{f(\widetilde{x}_1+s_n)} = \frac{\varepsilon_f(\widetilde{x}_1+s_n)}{cs_n^m} = o(1).$$

Θα θέλαμε επίσης να σχολιάσουμε ότι έχουμε τη δυνατότητα επιλογής του s_n. Για παράδειγμα, αν $\varepsilon_f = \mathcal{O}\left(\frac{1}{n}\right)$ και $m = 1$, μπορούμε να επιλέξουμε $s_n = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ ή $s_n = \mathcal{O}\left(\frac{\log n}{n}\right)$, υπάρχουν δηλαδή άπειρες επιλογές.

Το ίδιο ισχύει και για τα σημεία $-\widetilde{x}_1 - s_n$, $-\widetilde{x}_1 + s_n$ και $\widetilde{x}_1 - s_n$. Είναι προφανές ότι ο προαναφερθής λόγος παραμένει τάξεως $o(1)$, για τα σημεία που ανήκουν στο $(-\pi,\pi]\backslash(-\widetilde{x}_1 - s_n, -\widetilde{x}_1 + s_n) \cup (\widetilde{x}_1 - s_n, \widetilde{x}_1 + s_n)$. Επομένως, λαμβάνουμε τη συσσώρευση των ιδιοτιμών του $C_n\left(\frac{\varepsilon_f}{F_{n-1}}\right)$ σε μια περιοχή του $(0,0)$, με ακτίνα $o(1)$. Προκειμένου να βρούμε το είδος συσσώρευσης, μένει να μετρήσουμε πόσες ιδιοτιμές κυμαίνονται εκτός αυτής και αντιστοιχούν σε σημεία των παραπάνω διαστημάτων. Η απόσταση κάθε διαστήματος είναι $2s_n$, επομένως λαμβάνουμε το πολύ $\frac{4s_n}{2\pi}n = o(n)$ ιδιοτιμές εκτός της συσσώρευσης, το οποίο σημαίνει ότι έχουμε γενική συσσώρευση των ιδιοτιμών σε μια μικρή περιοχή του $(0,0)$.

Θα μελετήσουμε τον όρο $C_n\left(\frac{\varepsilon_g}{g}\right)$. Η συνάρτηση σφάλματος ε_g εξαρτάται από το μέγεθος του σφάλματος κατά την εκτίμηση της ρίζας $\varepsilon_1 = \widetilde{x}_1 - x_1$. Ο S. Serra-Capizzano απέδειξε στην [71], ότι αν $f \in C^k$, αυτό το σφάλμα φράσσεται από $|\varepsilon_1| \leq C\left(\frac{\log n}{n^k}\omega(f;n^{-1})\right)^{\frac{1}{m}}$, οπότε $|\varepsilon_1| = \mathcal{O}\left(\frac{\log n}{n^k}\right)^{\frac{1}{m}}$. Στην περίπτωση όπου $f \in C$, λαμβάνουμε το σφάλμα $|\varepsilon_1| = \mathcal{O}\left(\frac{\log n}{n}\right)^{\frac{1}{m}}$, ενώ αν η f είναι συνεχώς παραγωγίσιμη, είναι γνωστό ότι $|\varepsilon_1| = \mathcal{O}\left(\frac{1}{n}\right)^{\frac{1}{m}}$. Στη συνέχεια, μελετάμε το λόγο $\frac{\varepsilon_g}{g}$:

$$\begin{aligned}
\frac{\varepsilon_g(x)}{g(x)} &= \frac{g_n(x)-g(x)}{g(x)} = \frac{(\cos\widetilde{x}_1 - \cos x)^m - (\cos x_1 - \cos x)^m}{(\cos x_1 - \cos x)^m} \\
&= \frac{(\cos\widetilde{x}_1 - \cos x_1)\sum_{i=0}^{m-1}(\cos\widetilde{x}_1 - \cos x)^{m-i-1}(\cos x_1 - \cos x)^i}{(\cos x_1 - \cos x)^m} \\
&= \frac{\cos\widetilde{x}_1 - \cos x_1}{\cos x_1 - \cos x} \times \sum_{i=0}^{m-1}\left(\frac{\cos\widetilde{x}_1 - \cos x}{\cos x_1 - \cos x}\right)^i \\
&= \frac{\sin\frac{x_1+\widetilde{x}_1}{2}\sin\frac{x_1-\widetilde{x}_1}{2}}{\sin\frac{x+x_1}{2}\sin\frac{x-x_1}{2}}\sum_{i=0}^{m-1}\left(\frac{\sin\frac{x+\widetilde{x}_1}{2}\sin\frac{x-\widetilde{x}_1}{2}}{\sin\frac{x+x_1}{2}\sin\frac{x-x_1}{2}}\right)^i.
\end{aligned}$$

'Οπως και στη μελέτη του $\frac{\varepsilon_1}{F_{n-1}}$, έτσι κι εδώ, θεωρούμε τα διαστήματα $(-x_1 - s_n, -x_1 + s_n)$ και $(x_1 - s_n, x_1 + s_n)$, τα οποία τώρα έχουν ως κέντρο τις ακριβείς ρίζες $-x_1$ και x_1, αντίστοιχα και ακτίνα s_n που εξαρτάται από το n, αλλά $s_n = o(1)$. Υπολογίζοντας τη συνάρτηση $\frac{\varepsilon_g}{g}$ στο σημείο $x_1 + s_n$, έχουμε:

$$\frac{\varepsilon_g(x_1+s_n)}{g(x_1+s_n)} = \frac{\sin\frac{x_1+\widetilde{x}_1}{2}\sin\frac{-\varepsilon_1}{2}}{\sin\left(x_1+\frac{s_n}{2}\right)\sin\frac{s_n}{2}}\sum_{i=0}^{m-1}\left(\frac{\sin\frac{x_1+s_n+\widetilde{x}_1}{2}\sin\frac{s_n-\varepsilon_1}{2}}{\sin\left(x_1+\frac{s_n}{2}\right)\sin\frac{s_n}{2}}\right)^i.$$

Η ακτίνα s_n έχει επιλεχθεί έτσι ώστε $\varepsilon_1 = o(s_n)$. Τότε, το τελευταίο άθροισμα είναι φραγμένο και θετικό μακριά από το 0. Επομένως, ο πρώτος λόγος είναι αυτός που χαρακτηρίζει το μέγεθος του $\frac{\varepsilon_g}{g}$, στο $x_1 + s_n$.

$$\frac{\varepsilon_g(x_1+s_n)}{g(x_1+s_n)} = C\frac{\sin\frac{x_1+\widetilde{x}_1}{2}\sin\frac{-\varepsilon_1}{2}}{\sin\left(x_1+\frac{s_n}{2}\right)\sin\frac{s_n}{2}} \sim \frac{\varepsilon_1}{s_n} = o(1).$$

Υπολογίζοντας το λόγο $\frac{\varepsilon_g}{g}$ στα σημεία $-x_1 - s_n$, $-x_1 + s_n$ και $x_1 - s_n$, καταλήγουμε στο ίδιο αποτέλεσμα. 'Οπως και στην προηγούμενη περίπτωση, ο λόγος $\frac{\varepsilon_g}{g}$

παραμένει τάξεως $o(1)$ για τα σημεία που ανήκουν στο $(-\pi,\pi]\backslash(-x_1-s_n,-x_1+s_n)\cup(x_1-s_n,x_1+s_n)$. Επομένως, έχουμε γενική συσσώρευση των ιδιοτιμών του $C_n\left(\frac{\varepsilon_g}{g}\right)$ σε μια περιοχή του $(0,0)$, με ακτίνα τάξεως $o(1)$ και το πολύ $\frac{4s_n}{2\pi}n=o(n)$ ιδιοτιμές εκτός της συσσώρευσης.

Στη συνέχεια θεωρούμε ότι η πολλαπλότητα m της ρίζας, αντιστοιχεί στο φανταστικό μέρος της f. Τότε, ο λόγος $\frac{g_n}{g}$ δίνεται ως:

$$\frac{g_n(x)}{g(x)}=\frac{(\cos\widetilde{x}_1-\cos x)^m}{(\cos x_1-\cos x)^m}\frac{\sin x-\mathrm{i}(\cos\widetilde{x}_1-\cos x)^l}{\sin x-\mathrm{i}(\cos x_1-\cos x)^l}.$$

Αφού $\varepsilon_1=o(1)$, το δεύτερο κλάσμα είναι κοντά στο 1, για κάθε $x\in(-\pi,\pi]$. Πιο συγκεκριμένα, $\frac{\sin x-\mathrm{i}(\cos\widetilde{x}_1-\cos x)^l}{\sin x-\mathrm{i}(\cos x_1-\cos x)^l}=1+o(1)$. Άρα το μέγεθος του $\frac{g_n}{g}$ χαρακτηρίζεται από το πρώτο κλάσμα, το οποίο παραπάνω μελετήθηκε εκτενώς. Προφανώς, η ακολουθία πινάκων του όρου $C_n\left(\frac{\varepsilon_f\varepsilon_g}{F_{n-1}g}\right)$, που είναι το γινόμενο $C_n\left(\frac{\varepsilon_f}{F_{n-1}}\right)C_n\left(\frac{\varepsilon_g}{g}\right)$ έχει γενική συσσώρευση των ιδιοτιμών σε μια περιοχή του $(0,0)$, με ακτίνα $o(1)$ και $o(n)$ ιδιοτιμές εκτός της συσσώρευσης.

Συμπερασματικά, η ακολουθία πινάκων $\left\{C_n\left(\frac{fg_n}{F_{n-1}g}\right)\right\}$ έχει γενική συσσώρευση των ιδιοτιμών, σε μια περιοχή του $(1,0)$ με ακτίνα $o(1)$ και $o(n)$ ιδιοτιμές οι οποίες κυμαίνονται εκτός της συσσώρευσης. Το μέγεθος της ακτίνας χαρακτηρίζεται από το μεγαλύτερο μέγεθος εκ των $C_n\left(\frac{\varepsilon_f}{F_{n-1}}\right)$ και $C_n\left(\frac{\varepsilon_g}{g}\right)$ και ο αριθμός ιδιοτιμών που χαρακτηρίζουν τη συσσώρευση ως γενική, από τον αντίστοιχο μεγαλύτερο αριθμό. Προφανώς, το Ερμιτιανό και αντι-Ερμιτιανό του μέρος έχουν το ίδιο είδος συσσώρευσης.

Μένει να μελετήσουμε τον όρο $T_n^{-1}(g_n)T_n(g)$. Είναι γνωστό ότι το φάσμα αυτού του πίνακα είναι ισοκατανεμημένο με αυτό του $T_n\left(\frac{g}{g_n}\right)$, που σημαίνει ότι μπορούμε να εξετάσουμε τον τελευταίο πίνακα, αντί του $T_n^{-1}(g_n)T_n(g)$. Έχουμε:

$$T_n\left(\frac{g}{g_n}\right)=T_n\left(\frac{g_n-\varepsilon_g}{g_n}\right)=I_n-T_n\left(\frac{\varepsilon_g}{g_n}\right).$$

Θα εξετάσουμε το φάσμα του $T_n\left(\frac{\varepsilon_g}{g_n}\right)$. Προκειμένου να αποδείξουμε τη συσσώρευση των ιδιοτιμών της ακολουθίας πινάκων $\left\{T_n\left(\frac{\varepsilon_g}{g_n}\right)\right\}$, θεωρούμε τα διαστήματα $(-\widetilde{x}_1-s_n,-\widetilde{x}_1+s_n)$ και $(\widetilde{x}_1-s_n,\widetilde{x}_1+s_n)$, τα οποία έχουν το ίδιο μέγεθος $2s_n$, που εξαρτάται από το ε_1 ($\varepsilon_1=o(s_n)$, $s_n=o(1)$), όπως προηγουμένως. Η ίδια ανάλυση μας δίνει ότι $\frac{\varepsilon_g(x)}{g_n(x)}\leq r_n=o(1)$ για όλα τα x, εκτός των παραπάνω διαστημάτων.

Σταθεροποιούμε το n και χρησιμοποιούμε το θεώρημα Szegö με:

$$F(z) = \begin{cases} 1, & |z| \geq r_n + h \\ 0, & |z| \leq r_n \end{cases}, \tag{4.9}$$

για αρκούντως μικρό h. Τότε, ο αριθμός των ιδιοτιμών εκτός της περιοχής με κέντρο το $(1,0)$ και ακτίνα $r_n = o(1)$, δίνεται ως:

$$\lim_{N\to\infty} \frac{1}{N} \#\{\lambda_i : |\lambda_i| > r_n\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} F\left(\frac{\varepsilon_g(x)}{g_n(x)}\right) \mathrm{d}x \leq \frac{1}{2\pi} \int_{x\in I_p} 1 \mathrm{d}x = \frac{4s_n}{2\pi} = \frac{2s_n}{\pi},$$

όπου $I_p = (-\widetilde{x}_1 - s_n, -\widetilde{x}_1 + s_n) \cup (\widetilde{x}_1 - s_n, \widetilde{x}_1 + s_n)$. Επομένως, $\#\{\lambda_i : |\lambda_i| > r_n\} \leq 2\frac{s_n}{\pi} N$ για αρκούντως μεγάλη τιμή του N. Πηγαίνοντας πίσω στο n, λαμβάνουμε $\#\{\lambda_i : |\lambda_i| > r_n\} \leq 2\frac{s_n}{\pi} n = o(n)$.

Με άλλα λόγια αποδείξαμε ότι ο $T_n\left(\frac{g}{g_n}\right)$ γράφεται ως $T_n\left(\frac{g}{g_n}\right) = I_n + S_n + R_n$, όπου $\|S_n\|_2 = r_n = o(1)$ και $\|R_n\|_F^2 \sim 2\frac{s_n}{\pi} n = o(n)$. Εφόσον ο $T_n\left(\frac{g}{g_n}\right)$ προέκυψε από τον $T_n^{-1}(g_n)T_n(g)$ προσθέτοντας έναν πίνακα χαμηλής βαθμίδας, η οποία είναι σταθερή και ανεξάρτητη της διάστασης n, καταλήγουμε στο ότι $T_n^{-1}(g_n)T_n(g) = I_n + S_n + R_n'$, όπου ο R_n' είναι επίσης πίνακας χαμηλής βαθμίδας. Επομένως, $\|R_n'\|_F^2 \sim 2\frac{s_n}{\pi} n = o(n)$. Είναι προφανές ότι οι ιδιοτιμές του Ερμιτιανού και αντι-Ερμιτιανού μέρους του $T_n^{-1}(g_n)T_n(g)$ έχουν το ίδιο είδος γενικής συσσώρευσης.

Στη συνέχεια χρησιμοποιούμε το Λήμμα 4.2.4 για να αποδείξουμε τη γενική συσσώρευση. Επιλέγοντας $\mathcal{A}_n = C_n\left(\frac{fg_n}{F_{n-1}g}\right)$ και $\mathcal{B}_n = T_n^{-1}(g_n)T_n(g)$, έχουμε ότι το γινόμενο $C_n\left(\frac{fg_n}{F_{n-1}g}\right) T_n^{-1}(g_n)T_n(g)$ έχει γενική συσσώρευση των ιδιοτιμών, σε μια περιοχή του $(1,0)$ με ακτίνα $o(1)$, η οποία αντιστοιχεί στη μεγαλύτερη ακτίνα των δύο παραγόντων που το απαρτίζουν, καθώς επίσης και $o(n)$ ιδιοτιμές που κυμαίνονται εκτός της συσσώρευσης, οι οποίες επίσης εξαρτόνται από τους δύο παράγοντες του γινομένου.

Τέλος, χρησιμοποιούμε για άλλη μία φορά το ίδιο λήμμα, επιλέγοντας $\mathcal{A}_n = C_n\left(\frac{fg_n}{F_{n-1}g}\right) T_n^{-1}(g_n)T_n(g)$ και $\mathcal{B}_n = T_n\left(\frac{f}{g}\right) C_n^{-1}\left(\frac{f}{g}\right)$. Έτσι λαμβάνουμε τη γενική συσσώρευση των ιδιοτιμών του προρρυθμισμένου πίνακα σε μια περιοχή του $(1,0)$, με ακτίνα ίση με αυτή του πρώτου όρου και $o(n)$ ιδιοτιμές εκτός της συσσώρευσης.

Αποδείξαμε ότι η ακολουθία πινάκων (4.7) χωρίζεται ως $I_n + S_n + R_n$, όπου $\|S_n\|_2 = o(1)$ είναι η μεγαλύτερη από τις ακτίνες των τριών όρων του γινομένου και η νόρμα $\|R_n\|_F^2 = o(n)$ αντιστοιχεί στον μεγαλύτερο αριθμό των ιδιοτιμών που κυμαίνονται εκτός της συσσώρευσης, βάσει των όρων του γινομένου. Προφανώς, το Ερμιτιανό και αντι-Ερμιτιανό του μέρος έχουν το ίδιο είδος συσσώρευσης των ιδιοτιμών.

Συνοψίζοντας, αποδείξαμε ότι ο πίνακας $\widetilde{P}_n$, ή ο όμοιος αυτού P_n' χωρίζεται ως $P_n' = I_n + S_n + R_n$, όπου $\|S_n\|_2 = o(1)$ είναι η τελική ακτίνα της περιοχής συσσώρευσης και $\|R_n\|_F^2 = o(n)$, ο τελικός αριθμός των ιδιοτιμών που κυμαίνονται εκτός της συσσώρευσης. Επειδή ο P_n διαφέρει από τον P_n' κατά έναν πίνακα χαμηλής και σταθερής βαθμίδας (βλ. 4.6), ο προρρυθμισμένος πίνακας γράφεται ως $P_n = I_n + S_n + R_n'$, όπου $\|R_n'\|_F^2 = o(n)$, το οποίο αποδεικνύει τη γενική συσσώρευση των ιδιοτιμών.

Συνεχίζουμε με την απόδειξη της περίπτωσης 2. Αυτή γίνεται ακολουθώντας τα ίδια βήματα με την απόδειξη της περίπτωσης 1, αλλά εντοπίζονται δύο διαφορές. Η πρώτη έχει να κάνει με τον τρίτο όρο του γινομένου της (4.7), όπου ο αριθμός των ιδιοτιμών που κυμαίνονται εκτός της συσσώρευσης είναι $\mathcal{O}(\log n)$, λόγω της ασυνέχειας της f, όπως αποδείχθηκε στο Θεώρημα 4.2.5. Η δεύτερη και πιο ουσιαστική διαφορά έχει να κάνει με τη μελέτη του πρώτου όρου του γινομένου (4.7), δηλαδή του $C_n\left(\frac{fg_n}{F_{n-1}g}\right)$ και ειδικότερα του πίνακα $C_n\left(\frac{\varepsilon_f}{F_{n-1}}\right)$. Επειδή εμφανίζεται το φαινόμενο Gibbs, αυτός έχει συσσώρευση των ιδιοτιμών σε μια περιοχή του $(0,0)$ με ακτίνα σχεδόν ίση με $0.179/(1-0.179)$, η οποία από τη μία είναι μικρή, αλλά από την άλλη είναι $\mathcal{O}(1)$ (σταθερή). Αυτή η ακτίνα, ακολουθώντας την παραπάνω απόδειξη, είναι η μεγαλύτερη σε μέγεθος κι έτσι καταλήγουμε στη γενική συσσώρευση των ιδιοτιμών του προρρυθμισμένου πίνακα, σε μια περιοχή του $(1,0)$ με ακτίνα το πολύ ίση με $0.179/(1-0.179)$ και $o(n)$ ιδιοτιμές να κυμαίνονται εκτός της συσσώρευσης, βάσει των τριών όρων του γινομένου. □

Παρατήρηση. Στην περίπτωση όπου το πραγματικό και φανταστικό μέρος της γεννήτριας συνάρτησης δεν έχουν τα ίδια σημεία ασυνέχειας, η περιοχή συσσώρευσης των ιδιοτιμών, γύρω από το $(1,0)$, έχει ακτίνα αρκετά μικρότερη του 0.218.

Προχωράμε με τη μελέτη της συσσώρευσης των ιδιαζουσών τιμών του προρρυθμισμένου πίνακα.

Θεώρημα 4.2.7. Έστω T_n ένας πραγματικός πίνακας Toeplitz, στον οποίο αντιστοιχεί μια άγνωστη και μιγαδική γεννήτρια συνάρτηση f, με ρίζες στο διάστημα $(-\pi,\pi]$. Έστω F_{n-1} το ανάπτυγμα Fourier του T_n και g_n το τριγωνομετρικό πολυώνυμο, το οποίο αίρει τις εκτιμώμενες ρίζες μέσω της τεχνικής που περιγράφηκε. Υποθέτουμε ότι οι εκτιμώμενες μη-μηδενικές ρίζες $\widetilde{x}_i$, $i = 1, 2, \ldots, \rho$ έχουν ένα σφάλμα $\varepsilon_i = \widetilde{x}_i - x_i$ και ότι οι πολλαπλότητες αυτών έχουν εκτιμηθεί με ακρίβεια.

Τότε, ο προρρυθμισμένος πίνακας $C_n^{-1}\left(\frac{F_{n-1}}{g_n}\right)T_n^{-1}(g_n)T_n$ έχει:

1. Γενική συσσώρευση των ιδιαζουσών τιμών σε ένα διάστημα γύρω από το 1, με μήκος $o(1)$ και $o(n)$ ιδιάζουσες τιμές εκτός της συσσώρευσης, αν η f είναι συνεχής.

2. Γενική συσσώρευση των ιδιαζουσών τιμών στο διάστημα $[0.6847, 1.2374]$ και $o(n)$ ιδιάζουσες τιμές εκτός της συσσώρευσης, αν η f είναι κατά τμήματα συνεχής.

Απο404δειξη. Όπως εύκολα παρατηρεί κανείς, οι υποθέσεις του θεωρήματος είναι ίδιες με αυτές του Θεωρήματος 4.2.6. Επομένως, αποδεικνύεται ότι στην περίπτωση 1 οι ιδιοτιμές του προρρυθμισμένου πίνακα P_n έχουν γενική συσσώρευση σε μια περιοχή του $(1, 0)$, με ακτίνα $r = o(1)$ και $r = 0.218$ στην περίπτωση 2, καθώς επίσης και $s(n) = o(n)$ ιδιοτιμές εκτός της συσσώρευσης. Αποδείχθηκε ότι ο P_n γράφεται ως $P_n = I_n + \mathcal{S}_n + \mathcal{R}_n$, όπου $\|\mathcal{S}_n\|_2 \le r$ και $\|\mathcal{R}_n\|_F^2 = \mathcal{O}(s(n))$. Προφανώς, το συμμετρικό του μέρος γράφεται ως $\frac{P_n+P_n^T}{2} = I_n + S_n + R_n$, όπου $S_n = \frac{\mathcal{S}_n+\mathcal{S}_n^T}{2}$ και $R_n = \frac{\mathcal{R}_n+\mathcal{R}_n^T}{2}$ και το αντι-συμμετρικό του μέρος ως $\frac{P_n-P_n^T}{2} = S_n' + R_n'$, όπου $S_n' = \frac{\mathcal{S}_n-\mathcal{S}_n^T}{2}$ και $R_n' = \frac{\mathcal{R}_n-\mathcal{R}_n^T}{2}$. Ισχύει ότι $\|S_n\|_2 \le r$, $\|S_n'\|_2 \le r$, $\|R_n\|_F^2 = \mathcal{O}(s(n))$ και $\|R_n'\|_F^2 = \mathcal{O}(s(n))$.

Για να μελετήσουμε τη συσσώρευση των ιδιαζουσών τιμών, μελετάμε τη συσσώρευση των ιδιοτιμών της ακολουθίας πινάκων που αντιστοιχεί στο σύστημα των κανονικών εξισώσεων $\{Q_n\} = \{P_n^T\}\{P_n\}$. Επειδή η ακολουθία πινάκων $\{Q_n\}$ αποτελεί ένα γινόμενο ακολουθιών πινάκων, μπορούμε να χρησιμοποιήσουμε το Λήμμα 4.2.4 με:

$$\begin{aligned}\mathcal{A}_n &= P_n^T = I_n + S_n - S_n' + R_n - R_n' \text{ και}\\ \mathcal{B}_n &= P_n = I_n + S_n + S_n' + R_n + R_n'.\end{aligned}$$

Τότε:

$$\begin{aligned}\mathcal{C}_n \equiv Q_n = P_n^T P_n &= (I_n + S_n - S_n' + R_n - R_n')(I_n + S_n + S_n' + R_n + R_n')\\ &= I_n + 2S_n + S_n^2 - {S_n'}^2 + (R_n - R_n')(I_n + S_n + S_n' + R_n + R_n')\\ &\quad + (I_n + S_n - S_n')(R_n + R_n')\\ &= I_n + \widehat{\mathcal{S}}_n + \widehat{\mathcal{R}}_n,\end{aligned}$$

όπου $\widehat{\mathcal{S}}_n = 2S_n + S_n^2 - {S_n'}^2$ και $\widehat{\mathcal{R}}_n = (R_n - R_n')(I_n + S_n + S_n' + R_n + R_n') + (I_n + S_n - S_n')(R_n + R_n')$. Είναι προφανές ότι $\|\widehat{\mathcal{S}}_n\|_2 \le 2\|S_n\|_2 + \|S_n^2\|_2 + \|{S_n'}^2\|_2 \le 2\|S_n\|_2 + \|S_n\|_2^2 + \|S_n'\|_2^2 \le 2r + 2r^2$, ενώ $\operatorname{rank}(\widehat{R}_n) \le 2\operatorname{rank}(R_n) + 2\operatorname{rank}(R_n') = 4\mathcal{O}(s(n))$.

Επομένως, η ακολουθία πινάκων των κανονικών εξισώσεων $\{Q_n\}$ έχει γενική συσσώρευση των ιδιοτιμών στο $\left(1-2r-2r^2, 1+2r+2r^2\right)$ με $\mathcal{O}(s(n))$ ιδιοτιμές εκτός της περιοχής συσσώρευσης. Επειδή οι ιδιάζουσες τιμές του προρρυθμισμένου πίνακα είναι οι τετραγωνικές ρίζες των ιδιοτιμών του Q_n, λαμβάνουμε ότι, ο προρρυθμισμένος πίνακας έχει γενική συσσώρευση των ιδιαζουσών τιμών στο $\left(\sqrt{1-2r-2r^2}, \sqrt{1+2r+2r^2}\right)$ με $\mathcal{O}(s(n))$ ιδιάζουσες τιμές εκτός του διαστήματος.

Στην περίπτωση 1 από το Θεώρημα 4.2.6, έχουμε ότι $r = o(1)$ και συνεπώς το παραπάνω διάστημα είναι σχεδόν το ίδιο με το $(1-r, 1+r)$. Συνεπώς, το μήκος του είναι σχεδόν ίσο με $2r = o(1)$.

Στην περίπτωση 2 εμφανίζεται το φαινόμενο Gibbs και $r = 0.218$. Επομένως, το παραπάνω διάστημα είναι ίσο με $[0.6847, 1.2374]$.

Αυτή η παρατήρηση μας οδηγεί στο συμπέρασμα ότι η συσσώρευση των ιδιοτιμών συμβαδίζει με αυτή των ιδιαζουσών τιμών, βάσει της ακτίνας r και των αριθμό των $\mathcal{O}(s(n))$ ιδιοτιμών που κυμαίνονται εκτός της συσσώρευσης. $\square$

Παρατήρηση. Πρακτικά, όταν η f είναι κατά τμήματα συνεχής η περιοχή συσσώρευσης των ιδιοτιμών, καθώς επίσης και το διάστημα συσσώρευσης των ιδιαζουσών τιμών, είναι μικρότερη/ο όπως θα δούμε και θα εξηγήσουμε παρακάτω, στα αριθμητικά πειράματα (βλ. Παραδείγματα 4.2.10 και 4.2.11).

4.2.3 Αριθμητικά αποτελέσματα

Εδώ θα δείξουμε την αποτελεσματικότητα της προτεινόμενης τεχνικής προρρύθμισης και την ισχύ των θεωρητικών αποτελεσμάτων που αποδείξαμε. Οι επιλογές υλοποίησης είναι ίδιες με αυτές των προηγούμενων κεφαλαίων, εκτός από τα δύο τελευταία παραδείγματα όπου μεταβάλαμε το κριτήριο τερματισμού των μεθόδων σε $\frac{\|r^{(k)}\|_2}{\|r^{(0)}\|_2} \leq 10^{-7}$, προκειμένου να λάβουμε μεγαλύτερη ακρίβεια. Στους πίνακες επαναλήψεων με n δηλώνουμε προφανώς τη διάσταση του συστήματος, $\mathcal{C}_n(f)$ είναι ο κυκλοειδής προρρυθμιστής του προηγούμενου κεφαλαίου, όπου η γεννήτρια συνάρτηση είναι γνωστή εκ των προτέρων. Αναφέρουμε ότι αυτός προτάθηκε στην [50]. $\mathcal{C}_n$ είναι ο προρρυθμιστής που κατασκευάστηκε με τον τρόπο που περιγράψαμε. Για τα συστήματα με κακή κατάσταση, όπου χρησιμοποιούμε ταινιωτούς-επίκυκλοειδείς προρρυθμιστές, χρησιμοποιούμε αντιστοίχως τον συμβολισμό $\mathcal{BC}_n(f)$ και $\mathcal{BC}_n$. Σε κάποια παραδείγματα γίνεται σύγκριση και με την τεχνική προρρύθμισης της προηγούμενης ενότητας, ενώ σε όλα εξ αυτών δίνονται οι χρόνοι CPU, από την κατασκευή του προρρυθμιστή έως ότου επιτευχθεί η σύγκλιση στη λύση του συστήματος. Ο κώδικας υλοποιήθηκε σε ένα σύστημα με τετραπύρηνο Intel i5-3470 στα 3.2 GHz και 4 GB DDR3 RAM στα 1600 MHz.

Παράδειγμα 4.2.8. Στο πρώτο παράδειγμα αυτής της ενότητας, φαίνεται η ισχύς του Θεωρήματος 4.2.2. Θεωρούμε τις συνεχείς συναρτήσεις, οι οποίες δεν έχουν ρίζες στο $(-\pi,\pi]$, $\mathfrak{f}_{11}(x) = x^2\sin^2(x) + 1 + \mathrm{i}\sin(x)x^2$ και $\mathfrak{f}_{12}(x) = x^2 + 1 + \mathrm{i}\mathfrak{h}_1'(x)$, όπου:

$$\mathfrak{h}_1'(x) = \begin{cases} -\pi - x, & x \in \left(-\pi, -\frac{\pi}{2}\right] \\ x, & x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right] \\ \pi - x, & x \in \left(\frac{\pi}{2}, \pi\right] \end{cases}.$$

Στους Πίνακες 4.7 και 4.8 δίνουμε τους αριθμούς επαναλήψεων της PGMRES και τους χρόνους CPU, έως ότου επιτευχθεί η σύγκλιση στη λύση των συστημάτων που έχουν ως γεννήτρια συνάρτηση την $\mathfrak{f}_{11}$ και $\mathfrak{f}_{12}$, αντίστοιχα. 'Οπως φαίνεται, δίνεται και ο αριθμός επαναλήψεων με χρήση του προρρυθμιστή $\mathcal{C}_n(f)$, όταν δηλαδή η γεννήτρια συνάρτηση είναι γνωστή εκ των προτέρων. Η σύγκλιση επιτυγχάνεται στις ίδιες επαναλήψεις και σε πολύ κοντινό χρόνο. Παρατηρούμε επίσης ότι για τις δύο μεγαλύτερες διαστάσεις, η ταχύτερη σύγκλιση επιτυγχάνεται με προρρύθμιση και όχι χωρίς καμία τεχνική προρρύθμισης. Είναι προφανές ότι οι πίνακες του συγκεκριμένου παραδείγματος έχουν καλή κατάσταση κι επομένως τα πλεονεκτήματα της προρρύθμισης δεν είναι ξεκάθαρα.

Η συσσώρευση των ιδιοτιμών για τα προρρυθμισμένα συστήματα δίνεται στο Σχήμα 4.8. Το Σχήμα 4.9 δείχνει τη συσσώρευση των ιδιοτιμών, κοντά στο $(1,0)$, όταν $n = 2048$, με χρήση του προτεινόμενου προρρυθμιστή (μπλε κύκλοι), καθώς επίσης και με χρήση του βέλτιστου κυκλοειδή προρρυθμιστή [16] (πράσινα διαμάντια). Παρατηρούμε μια πολύ πιο συμπαγή συσσώρευση με χρήση του πρώτου, όπως άλλωστε συνέβαινε και για γνωστές εκ των προτέρων γεννήτριες συναρτήσεις.

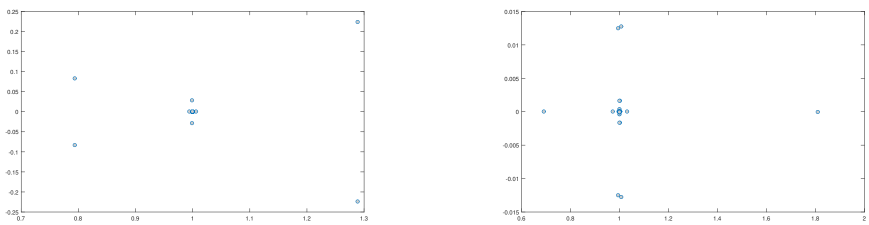
n	I_n	CPU	$\mathcal{C}_n(\mathfrak{f}_{11})$	CPU	$\mathcal{C}_n$	CPU
1024	40	0.0929	6	0.0832	6	0.0959
2048	39	0.1262	6	0.0842	6	0.0990
4096	37	0.1708	6	0.0972	6	0.1080
8192	36	0.2297	6	0.1259	6	0.1493

Πι404νακας 4.7: PGMRES: Επαναλήψεις και χρόνοι CPU ($\mathfrak{f}_{11}$).

n	I_n	CPU	$\mathcal{C}_n(\mathfrak{f}_{12})$	CPU	$\mathcal{C}_n$	CPU
1024	29	0.0968	5	0.1002	5	0.1227
2048	29	0.1116	4	0.1149	4	0.1270
4096	28	0.1455	4	0.1254	4	0.1297
8192	27	0.1865	4	0.1604	4	0.1651

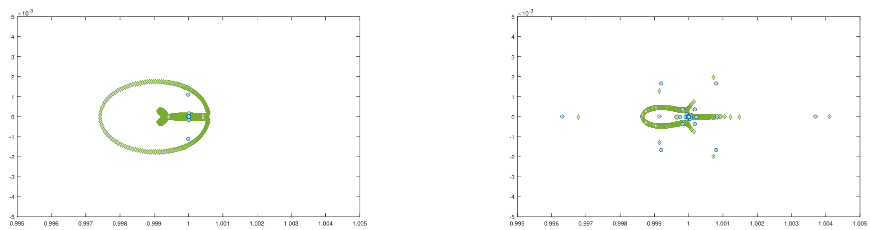
Πίνακας 4.8: PGMRES: Επαναλήψεις και χρόνοι CPU ($\mathfrak{f}_{12}$).

Στο Σχήμα 4.9β' παρατηρούμε επίσης ότι η συσσώρευση των ιδιοτιμών του $\mathcal{C}_n^{-1}T_n(\mathfrak{f}_{12})$ γύρω από το $(1,0)$, δεν είναι τόσο συμπαγής, όσο στο Σχήμα 4.9α', που αφορά στον $\mathcal{C}_n^{-1}T_n(\mathfrak{f}_{11})$. Αυτό ήταν αναμενόμενο από το Θεώρημα 4.2.2, διότι η $\mathfrak{f}_{12}$ είναι μια συνεχής συνάρτηση, αλλά όχι συνεχώς παραγωγίσιμη όπως είναι η $\mathfrak{f}_{11}$ και η περιοχή συσσώρευσης είναι της τάξεως $\mathcal{O}\left(\frac{\log n}{n}\right)$, αντί $\mathcal{O}\left(\frac{1}{n}\right)$.



(α') Ιδιοτιμές. (β') Ιδιοτιμές.

Σχήμα 4.8: Ιδιοτιμές (αριστερά $\mathfrak{f}_{11}$) και (δεξιά $\mathfrak{f}_{12}$).



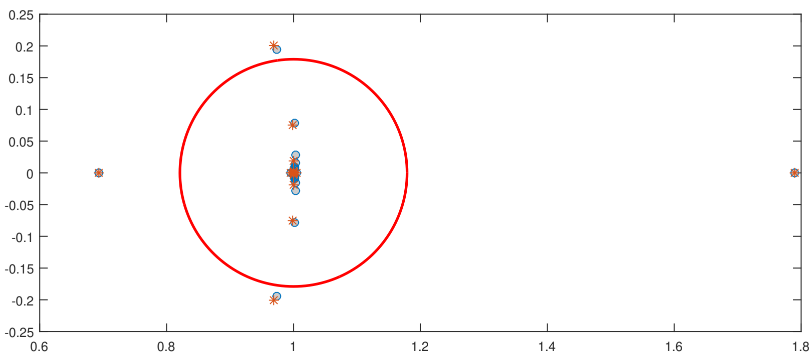
(α') Ιδιοτιμές κοντά στο $(1,0)$. (β') Ιδιοτιμές κοντά στο $(1,0)$.

Σχήμα 4.9: Ιδιοτιμές (αριστερά $\mathfrak{f}_{11}$) και (δεξιά $\mathfrak{f}_{12}$).

Παράδειγμα 4.2.9. Σε αυτό το παράδειγμα ασχολούμαστε με την προρρύθμιση του συστήματος Toeplitz, το οποίο έχει ως γεννήτρια συνάρτηση την $\mathfrak{f}_{13}(x) = x^2 + 1 + \mathrm{i}x$, $x \in (-\pi, \pi]$. Αυτή δεν έχει ρίζες και το φανταστικό της μέρος παρουσιάζει ασυνέχεια στο π, που σημαίνει ότι αντιστοιχεί στο Θεώρημα 4.2.5.

n	I_n	CPU	$\mathcal{C}_n(\mathfrak{f}_{13})$	CPU	$\mathcal{C}_n$	CPU	$R_{4,4}$	CPU
1024	35	0.1023	6	0.0758	6	0.0855	7	0.2011
2048	34	0.1183	6	0.0808	6	0.0917	7	0.2254
4096	33	0.1716	5	0.0773	5	0.0816	7	0.2790
8192	32	0.2142	5	0.0811	5	0.1067	6	0.3790

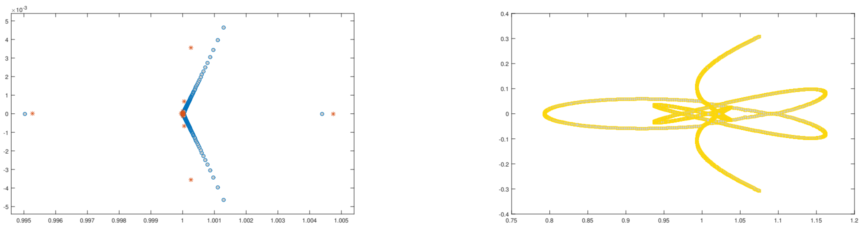
Πίνακας 4.9: PGMRES: Επαναλήψεις και χρόνοι CPU ($\mathfrak{f}_{13}$).



Σχήμα 4.10: Ιδιοτιμές ($\mathfrak{f}_{13}$).

Οι επαναλήψεις και ο χρόνος CPU, με χρήση των κυκλοειδών προρρυθμιστών όπως και στο προηγούμενο παράδειγμα, δίνεται στο Πίνακα 4.9. Σε αυτόν δίνουμε και τον αριθμό επαναλήψεων με χρήση του ταινιωτού Toeplitz προρρυθμιστή $R_{4,4}$, της προηγούμενης ενότητας. Η υπεροχή του $\mathcal{C}_n$ επί του $R_{4,4}$ είναι προφανής. Δίνουμε επίσης τη συσσώρευση των ιδιοτιμών του $\mathcal{C}_n^{-1}T_n(\mathfrak{f}_{13})$ (μπλε κύκλοι) και του $\mathcal{C}_n^{-1}(\mathfrak{f}_{13})T_n(\mathfrak{f}_{13})$ (πορτοκαλί αστέρια) στο Σχήμα 4.10. Ο κόκκινος κύκλος σε αυτό, δηλώνει την περιοχή συσσώρευσης για την ασυνεχή περίπτωση, με ακτίνα 0.179. Παρατηρούμε ότι στην πράξη οι ιδιοτιμές έχουν μια καλύτερη συσσώρευση, απ" ότι αναμένονταν, κοντά στο σημείο $(1, 0)$. Μια εξήγηση γι αυτό το φαινόμενο θα δοθεί στο επόμενο παράδειγμα.

Το Σχήμα 4.11α' δείχνει τη συσσώρευση των ιδιοτιμών κοντά στο $(1, 0)$. Σε αυτό φαίνεται η διαφορά, ανάμεσα στη συσσώρευση σε σημείο και στη συσσώρευση



(α') Ιδιοτιμές κοντά στο $(1, 0)$. (β') Ιδιοτιμές.

Σχήμα 4.11: Ιδιοτιμές ($\mathfrak{f}_{13}$).

σε περιοχή της τάξεως $\mathcal{O}\left(\frac{\log n}{n}\right)$. Το Σχήμα 4.11β' δείχνει την αντίστοιχη συσσώρευση, με χρήση του προρρυθμιστή $R_{4,4}$. Η διαφορά ανάμεσα στις συσσωρεύσεις, αντιστοιχεί και στη διαφορά ανάμεσα στις επαναλήψεις. Ωστόσο, και σε αυτό το παράδειγμα μελετάμε έναν πίνακα με καλή κατάσταση και τα οφέλη της προρρύθμισης δε γίνονται ακόμη αισθητά. Η διαφορά ανάμεσα στις επαναλήψεις θα αυξηθεί στα επόμενα παραδείγματα, όπου μελετάμε την προρρύθμιση συστημάτων με κακή κατάσταση.

Παράδειγμα 4.2.10. Σε αυτό το παράδειγμα η γεννήτρια συνάρτηση του πίνακα Toeplitz είναι η $\mathfrak{f}_2(x) = x^2 + \mathrm{i}x^3$, $x \in (-\pi, \pi]$, η οποία έχει ρίζα στο 0 και προσεγγίζεται με ακρίβεια χρησιμοποιώντας την προτεινόμενη τεχνική προρρύθμισης. Λεπτομέρειες για την προσέγγιση της πολλαπλότητας που έχει η ρίζα δίνονται στον Πίνακα 4.10. Σημειώνουμε ότι αν και στο Παράδειγμα 4.1.3 μελετήσαμε το ίδιο σύστημα, η πολλαπλότητα των ριζών εκτιμήθηκε σωστά, αλλά ελαφρώς διαφορετικά, όπως φαίνεται και στον Πίνακα 4.1. Υπενθυμίζουμε ότι το πλέγμα διαφέρει σε σχέση με την προηγούμενη ενότητα. Για την εκτίμηση των πολλαπλοτήτων, τρέξαμε 4 επαναλήψεις της μεθόδου Αντίστροφων Δυνάμεων, με $\Theta_{i,k} = \frac{1}{\sqrt{k}}\left(1, \mathrm{e}^{\mathrm{i}x_0}, \mathrm{e}^{2\mathrm{i}x_0}, \ldots, \mathrm{e}^{(k-1)\mathrm{i}x_0}\right)^T$ και $x_0 = 0$ ως αρχικό διάνυσμα. Καταλήξαμε στο ότι η πολλαπλότητα της ρίζας του πραγματικού μέρους είναι ίση με 2 και αυτή του φανταστικού με 3. Αυτό συνεπάγεται ότι το τριγωνομετρικό πολυώνυμο που είναι κατάλληλο για την άρση των ριζών δίνεται ως $g(x) = 2 - 2\cos x$, αφού η πολλαπλότητα της ρίζας του πραγματικού μέρους είναι μικρότερη από αυτή του φανταστικού.

Όπως καταλαβαίνει κανείς, στους Πίνακες 4.11 και 4.12, δίνουμε τις επαναλήψεις και τους χρόνους CPU, με χρήση των PGMRES και PCGN. Εκεί, τα πλεονεκτήματα της προρρύθμισης γίνονται πλέον ξεκάθαρα. Σχολιάζουμε ότι χωρίς προρρύθμιση η λύση του 1024×1024 συστήματος λαμβάνεται σε 547.8306 δευτερόλεπτα (δ), με τη

k	$\widetilde{\lambda}_{0,k}^{1}$	$\log_2(\widetilde{s_0}^1)$	$\widetilde{\lambda}_{0,k}^{2}$	$\log_2(\widetilde{s_0}^2)$
16	0.0351	1.9224	0.0118	2.9337
32	0.0092		0.0015	
64	0.0024		0.0002	

Πίνακας 4.10: Πολλαπλότητα των ριζών ($\mathfrak{f}_2$).

μέθοδο GMRES. Σχολιάζουμε επίσης ότι ο χρόνος κατασκευής των προρρυθμιστών αυξάνεται με αργό ρυθμό, όσο η διάσταση του συστήματος μεγαλώνει, ενώ ο χρόνος εκτέλεσης της μεθόδου επίλυσης είναι της τάξεως $\mathcal{O}(n \log n)$, όπως αναμένονταν από τη θεωρία. Για παράδειγμα, όταν $n = 4096$ η κατασκευή του $\mathcal{BC}_n$ γίνεται σε 0.2003 δ και η λύση λαμβάνεται σε 0.1635 δ, ενώ οι αντίστοιχες τιμές όταν $n = 8192$ είναι 0.2211 δ και 0.3533 δ. Η υπεροχή του $\mathcal{BC}_n$ επί του $R_{4,4}$ είναι προφανής, αφού η λύση λαμβάνεται σε λιγότερές από τις μισές επαναλήψεις.

n	I_n	$\mathcal{BC}_n(\mathfrak{f}_2)$	CPU	$\mathcal{BC}_n$	CPU	$R_{4,4}$	CPU
1024	$>$500	7	0.0942	11	0.2890	28	0.4834
2048	$>$500	7	0.1312	11	0.3036	28	0.6360
4096	$>$500	8	0.1660	12	0.3636	28	0.9496
8192	$>$500	8	0.2491	12	0.5744	27	1.4868

Πίνακας 4.11: PGMRES: Επαναλήψεις και χρόνοι CPU ($\mathfrak{f}_2$).

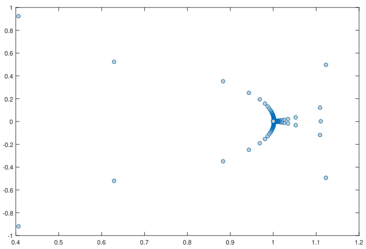
n	I_n	$\mathcal{BC}_n(\mathfrak{f}_2)$	CPU	$\mathcal{BC}_n$	CPU	$R_{4,4}$	CPU
1024	-	15	0.0626	19	0.2658	45	0.5775
2048	-	16	0.1022	22	0.2949	49	0.8101
4096	-	21	0.1750	26	0.3737	54	1.4418
8192	-	23	0.3139	29	0.5115	57	2.6980

Πίνακας 4.12: PCGN: Επαναλήψεις και χρόνοι CPU ($\mathfrak{f}_2$).

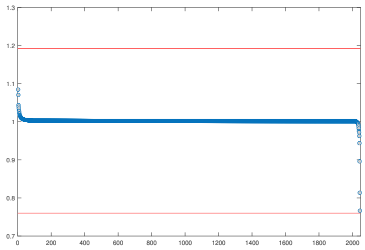
Το Σχήμα 4.12 δείχνει τη συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών, όταν $n = 2048$. Πιο συγκεκριμένα, το Σχήμα 4.12α' δείχνει τη συσσώρευση των ιδιοτιμών του $\mathcal{BC}_n^{-1}T_n(\mathfrak{f}_2)$ και το Σχήμα 4.12β', αυτή των ιδιαζουσών τιμών για τον ίδιο πίνακα. Ομοίως, τα Σχήματα 4.12γ' και 4.12δ' δείχνουν τη συσσώρευση των ιδιοτιμών και ιδιαζουσών τιμών του $R_{4,4}^{-1}T_n(\mathfrak{f}_2)$. Οι κόκκινες γραμμές χαρακτηρίζουν τα άκρα του διαστήματος $[0.7602, 1.1925]$. Η συσσώρευση των ιδιοτιμών στο

Σχήμα 4.12α' μοιάζει να είναι γενική, σε μια περιοχή του $(1, 0)$ με ακτίνα 0.179 και $\mathcal{O}(\log n)$ ιδιοτιμές εκτός του διαστήματος συσσώρευσης, όπως αποδείχθηκε στο Θεώρημα 4.2.6, ενώ στο Σχήμα 4.12γ' η συσσώρευση είναι σε ένα μεγαλύτερο ορθογώνιο, όπως αποδείχθηκε στο Θεώρησα 4.1.1 (βλ. επίσης Θεώρημα 2.1.2).

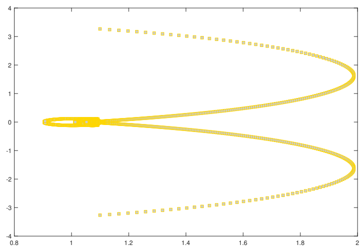
Θα θέλαμε να σχολιάσουμε τη συμπεριφορά των ιδιαζουσών τιμών μεταξύ των προαναφερθέντων πινάκων. Στο Σχήμα 4.12β' παρατηρούμε ότι, αν και η συσσώρευση είναι στο διάστημα $[0.7602, 1.1925]$, μοιάζει να έχουμε συσσώρευση γύρω από το σημείο 1 και μόνο λίγες ιδιάζουσες τιμές να είναι μακριά από αυτό. Αυτό οφείλεται στο φαινόμενο Gibbs, το οποίο εμφανίζεται στις περιοχές ασυνέχειας. Από την άλλη, στο Σχήμα 4.12δ', παρατηρούμε ότι οι ιδιάζουσες τιμές δεν έχουν τόσο αυστηρή δομή κι εξαπλώνονται σε μια περιοχή του 1, αφού μέσω του αλγορίθμου Remez παρουσιάζεται διακύμανση της καμπύλης γύρω από το 1, σε όλο το διάστημα ορισμού της συνάρτησης.



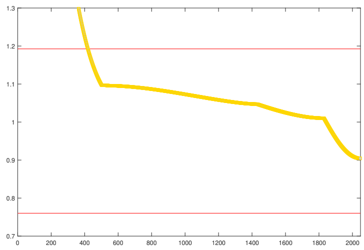
(α') Ιδιοτιμές.



(β') Ιδιάζουσες τιμές.



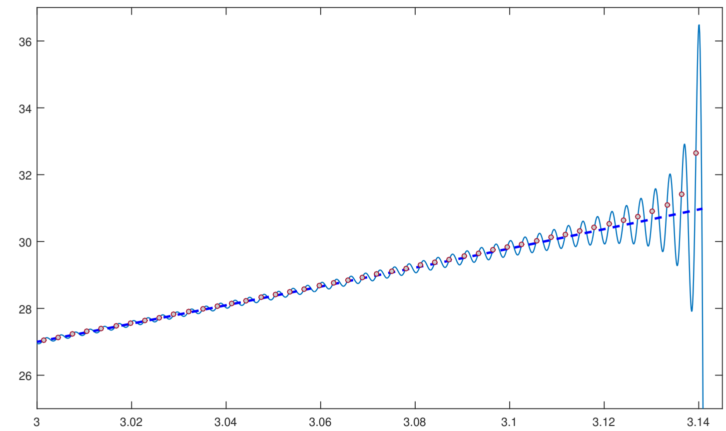
(γ') Ιδιοτιμές.



(δ') Ιδιάζουσες τιμές.

Σχήμα 4.12: Ιδιοτιμές και ιδιάζουσες τιμές ($\mathfrak{f}_2$).

Στο Σχήμα 4.13 δίνουμε το ανάπτυγμα Fourier F^2_{n-1}, του φανταστικού μέρους της $\mathfrak{f}_2$ (μπλε γραμμή) κοντά στο π, όπου εμφανίζεται το φαινόμενο Gibbs. Εκεί δίνουμε και τις τιμές του F^2_{n-1} στα σημεία του πλέγματος G_n (κόκκινοι κύκλοι). Η διακεκομμένη γραμμή είναι η γραφική παράσταση της $\mathrm{Im}(\mathfrak{f}_2)$. Παρατηρούμε ό-



Σχήμα 4.13: Φαινόμενο Gibbs κοντά στο π ($\mathrm{Im}(\mathfrak{f}_2)$).

τι οι τιμές στο G_n δε λαμβάνουν τις μέγιστες δυνατές τιμές της ταλάντωσης που σχετίζεται με το φαινόμενο Gibbs. Αυτός είναι ο λόγος που μπορεί να επιτύχουμε συσσώρευση σε μια μικρότερη περιοχή, όπως είδαμε στο Παράδειγμα 4.2.9.

Παράδειγμα 4.2.11. Σε αυτό το παράδειγμα, ο πίνακας του συστήματος Toeplitz έχει κακή κατάσταση, διότι η γεννήτρια συνάρτηση αυτού είναι η $\mathfrak{f}_9(x) = (x^2-1)^2 + \mathrm{i}x(x^2-1)$, $x \in (-\pi, \pi]$, την οποία μελετήσαμε προγενέστερα στο Παράδειγμα 4.1.5. Βλέπουμε ότι η $\mathfrak{f}_9$ έχει ρίζες στο ± 1 (1 και $2\pi - 1$, στο διάστημα $[0, 2\pi)$), που δεν είναι στοιχεία του G_n. Παρατηρούμε επίσης ότι το φανταστικό μέρος της γεννήτριας συνάρτησης έχει ασυνέχεια στο π, που σημαίνει ότι αναμένουμε μια πιο ελαστική συσσώρευση, λόγω του φαινομένου Gibbs.

Η ρίζα στο 1, εκτιμήθηκε στο πλέγμα G_n, χρησιμοποιώντας το ανάπτυγμα Fourier, ως $\widetilde{x}_1 = 1.0002$ (ακριβέστερα 1.000155), όταν $n = 2048$. Περισσότερες λεπτομέρειες για την εκτίμηση της ρίζας δίνεται, για διάφορες διαστάσεις του συστήματος, στον Πίνακα 4.13. Σημειώνουμε ότι σε αυτό το παράδειγμα οι εκτιμήσεις δόθηκαν στο ίδιο σημείο για διαφορετικές διαστάσεις του G_n. Οι πολλαπλότητες των ριζών, όπως φαίνεται στον Πίνακα 4.14, εκτιμήθηκαν μέσω της διαδικασίας που περιγράψαμε ως $m_1^1 = 2$, $m_0^2 = 1$ και $m_1^2 = 1$. Επομένως, $g_n(x) = (\cos\widetilde{x}_1 - \cos x)^2 + \mathrm{i}\sin x(\cos\widetilde{x}_1 - \cos x)$.

n	$\mathrm{Re}(F_{n-1})$		$\mathrm{Im}(F_{n-1})$	
	$\widetilde{x}_1$	$F^1_{n-1}(\widetilde{x}_1)$	$\widetilde{x}_1$	$F^2_{n-1}(\widetilde{x}_1)$
1024	θ_{164}	$-3.378 * 10^{-5}$	θ_{164}	$-4.422 * 10^{-3}$
2048	θ_{327}	$-8.367 * 10^{-6}$	θ_{327}	$-2.055 * 10^{-3}$
4096	θ_{653}	$-2.019 * 10^{-6}$	θ_{653}	$-8.722 * 10^{-4}$
8192	θ_{1305}	$-4.320 * 10^{-7}$	θ_{1305}	$-2.806 * 10^{-4}$

Πίνακας 4.13: Εκτίμηση της $\widetilde{x}_1$ στο G_n ($\mathfrak{f}_9$).

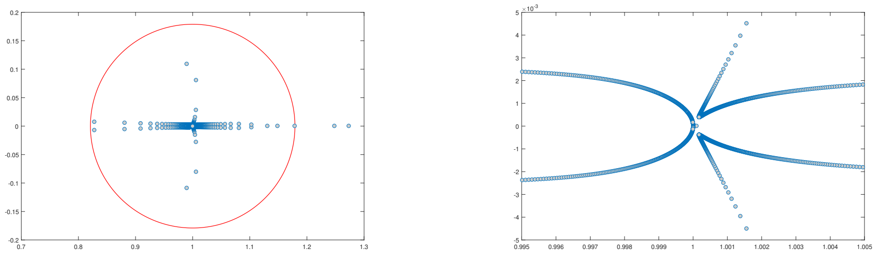
Από την [71] γνωρίζουμε ότι το σφάλμα στην εκτίμηση της ρίζας είναι $\varepsilon_1 = \mathcal{O}\left(\frac{\log n}{n}\right)$. Από τον Πίνακα 4.13, έχουμε ότι $\varepsilon_1 \sim \frac{1}{n}$, το οποίο έρχεται σε συμφωνία με τη θεωρία, αλλά είναι μια καλύτερη προσέγγιση από $\frac{\log n}{n}$. 'Οπως αποδείχθηκε στο Θεώρημα 4.2.6, ο τελευταίος παράγοντας $T_n\left(\frac{\mathfrak{f}_9}{g}\right) C_n^{-1}\left(\frac{\mathfrak{f}_9}{g}\right)$ έχει γενική συσσώρευση των ιδιοτιμών στο σημείο $(1,0)$, με $\mathcal{O}(\log n)$ ιδιοτιμές εκτός του διαστήματος συσσώρευσης, εξαιτίας της ασυνέχειας της $\mathfrak{f}_9$. Για τον δεύτερο παράγοντα, $T_n^{-1}(g_n)T_n(g)$, θεωρούμε ως $s_n \sim \frac{\log n}{n}$, για να λάβουμε γενική συσσώρευση των ιδιοτιμών σε μια περιοχή του $(1,0)$, με ακτίνα $r_n = \mathcal{O}\left(\frac{1}{\log n}\right)$ και $\mathcal{O}(\log n)$ ιδιοτιμές εκτός του διαστήματος συσσώρευσης. Λόγω του φαινομένου Gibbs, οι ιδιοτιμές του πρώτου παράγοντα, $C_n\left(\frac{\mathfrak{f}_9 g_n}{F_{n-1} g}\right)$, έχουν κύρια συσσώρευση σε μια περιοχή του $(1,0)$, με ακτίνα το πολύ 0.179. Επομένως, από το Θεώρημα $4.2.6$ καταλήγουμε σε γενική συσσώρευση των ιδιοτιμών του προρρυθμισμένου πίνακα, σε μια περιοχή του $(1,0)$, με ακτίνα το πολύ ίση με 0.179 και $\mathcal{O}(\log n)$ ιδιοτιμές εκτός του διαστήματος συσσώρευσης.

k	$\widetilde{\lambda}^1_{1,k}$	$\log_2(\widetilde{s_0}^1)$	$\widetilde{\lambda}^2_{0,k}$	$\log_2(\widetilde{s_0}^2)$	$\widetilde{\lambda}^2_{1,k}$	$\log_2(\widetilde{s_1}^2)$
16	0.1128	1.6781	0.1307	0.8825	0.1836	0.8956
32	0.0338		0.0689		0.1083	
64	0.0091		0.0355		0.0678	

Πίνακας 4.14: Πολλαπλότητα των ριζών ($\mathfrak{f}_9$).

Στο Σχήμα 4.14 δίνουμε τη συσσώρευση των ιδιοτιμών. Στο Σχήμα 4.14α', ο κόκκινος κύκλος είναι η περιοχή με ακτίνα 0.179, που αναφέραμε παραπάνω. Θα θέλαμε να σημειώσουμε ότι για $n = 2048$, υπάρχουν 8 επιπλέον ιδιοτιμές, εκτός του διαστήματος συσσώρευσης, που δε φαίνονται στο Σχήμα 4.14α', το οποίο κάνει τις ιδιοτιμές που κυμαίνονται εκτός του διαστήματος συσσώρευσης να είναι 10. Στο Σχήμα 4.14β' δίνεται μια μεγέθυνση πολύ κοντά στο $(1, 0)$.

Παρατήρηση. Στο Σχήμα 4.14α' η συσσώρευση παρατηρείται σε μια μικρότερη περιοχή και φαίνεται ότι το ίδιο φαινόμενο, που περιγράψαμε στο Σχήμα 4.13, του Παραδείγματος 4.2.10, εμφανίζεται και σε αυτό το παράδειγμα.



(α') Ιδιοτιμές. (β') Ιδιοτιμές κοντά στο $(1, 0)$.

Σχήμα 4.14: Ιδιοτιμές ($\mathfrak{f}_9$).

Στους Πίνακες 4.15 και 4.16 παρουσιάζουμε τις επαναλήψεις των μεθόδων PGMRES και PCGN, αντίστοιχα, καθώς επίσης και τους αντίστοιχους χρόνους. Υπενθυμίζουμε ότι το κριτήριο τερματισμού άλλαξε σε $\frac{\|r^{(k)}\|_2}{\|r^{(0)}\|_2} \leq 10^{-7}$.

n	I_n	$\mathcal{BC}_n(\mathfrak{f}_9)$	CPU	$\mathcal{BC}_n$	CPU	$R_{4,4}$	CPU
1024	>500	7	0.1227	11	0.3080	21	1.3286
2048	>500	6	0.1373	11	0.3252	22	2.4072
4096	>500	6	0.1780	12	0.3837	22	4.5923
8192	>500	6	0.2287	13	0.5177	22	9.6205

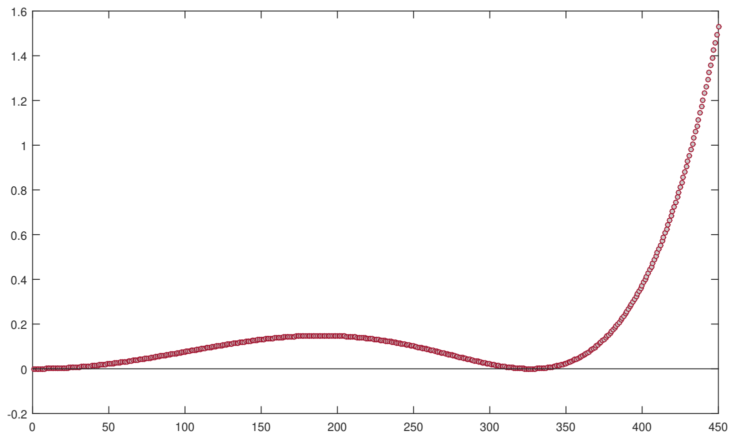
Πίνακας 4.15: PGMRES: Επαναλήψεις και χρόνοι CPU ($\mathfrak{f}_9$).

n	I_n	$\mathcal{BC}_n(\mathfrak{f}_9)$	CPU	$\mathcal{BC}_n$	CPU	$R_{4,4}$	CPU
1024	-	11	0.1398	19	0.3922	45	2.3847
2048	-	11	0.1928	20	0.4966	49	4.8036
4096	-	11	0.3162	19	0.7050	53	10.277
8192	-	11	0.5281	19	1.1097	57	23.194

Πίνακας 4.16: PCGN: Επαναλήψεις και χρόνοι CPU ($\mathfrak{f}_9$).

Παράδειγμα 4.2.12. Ως τελευταίο παράδειγμα, χρησιμοποιούμε την προτεινόμενη τεχνική προρρύθμισης για ένα σύστημα, του οποίου ο πίνακας Toeplitz έχει μια συνεχή γεννήτρια συνάρτηση με ρίζες στο 0 και στο ± 1, δηλαδή στο 1 και στο $2\pi - 1$, (στο διάστημα που μελετάμε). Το πραγματικό μέρος της συνάρτησης που επιλέξαμε, παίρνει τιμές κοντά στο 0 για $x \in [0, 1.2]$, όπως φαίνεται στο Σχήμα 4.15, όπου δίνουμε τις πρώτες 450 τιμές του F^1_{n-1} στο G_n, $n = 2048$. Μετά από αυτή τη μικρή εισαγωγή, σημειώνουμε ότι η γεννήτρια συνάρτηση δίνεται ως $\mathfrak{f}_{14}(x) = x^2(x^2-1)^2 + \mathrm{i}\mathfrak{h}'_3(x)$, όπου:

$$\mathfrak{h}'_3(x) = \begin{cases} x+\pi, & x \in \left(-\pi, -\pi+\frac{1}{2}\right] \\ \frac{x}{-2\pi+3} + \frac{1}{-2\pi+3}, & x \in \left(-\pi+\frac{1}{2}, -\frac{1}{2}\right] \\ \frac{x}{2\pi-3}, & x \in \left(-\frac{1}{2}, \frac{1}{2}\right] \\ \frac{x}{-2\pi+3} - \frac{1}{-2\pi+3}, & x \in \left(\frac{1}{2}, \pi-\frac{1}{2}\right] \\ x-\pi, & x \in \left(\pi-\frac{1}{2}, \pi\right] \end{cases}.$$



Σχήμα 4.15: Πρώτες 450 τιμές του F^1_{n-1} στο G_n ($\mathfrak{f}_{14}$).

Όπως και στο προηγούμενο παράδειγμα, η ρίζα στο 1 εκτιμήθηκε ως $\widetilde{x}_1 = 1.000155$, για διάφορες διαστάσεις του προρρυθμισμένου συστήματος. Σχολιάζουμε μια λεπτομέρεια στην εκτίμηση της $\widetilde{x}_1$. Η απόλυτη τιμή του F^1_{n-1} έχει δύο τοπικά ελάχιστα στην περιοχή της ρίζας, για παράδειγμα, όταν $n = 2048$ λαμβάνουμε τα τοπικά ελάχιστα στα σημεία θ_{325} και θ_{329}. Ωστόσο, θεωρούμε ότι η f_1 έχει μια ρίζα στο θ_{327}, που είναι ο μέσος όρος των παραπάνω ποσοτήτων, διότι τα θ_{325} και θ_{329} διαφέρουν κατά $\mathcal{O}\left(\frac{1}{n}\right)$. Ακριβώς η ίδια συμπεριφορά παρατηρήθηκε σε όλες τις διαστάσεις που εξετάσαμε (n : 1024, 2048, 4096, 8192).

Εκτιμήσαμε τις πολλαπλότητες των ριζών, τρέχοντας μόνο 2 επαναλήψεις της μεθόδου Αντίστροφων Δυνάμεων. 'Οπως φαίνεται στους Πίνακες 4.17 και 4.18, οι πολλαπλότητες εκτιμήθηκαν ορθά ως $m_0^1 = 2$, $m_1^1 = 2$, $m_0^2 = 1$ και $m_1^2 = 1$. Το τριγωνομετρικό πολυώνυμο που χρησιμοποιήσαμε για την άρση των ριζών είναι το $g_n(x) = (2 - 2\cos x)(\cos \widetilde{x}_1 - \cos x)^2 + \mathrm{i} \sin x(\cos x - \cos \widetilde{x}_1)$.

k	$\widetilde{\lambda}_{0,k}^1$	$\log_2(\widetilde{s_0}^1)$	$\widetilde{\lambda}_{1,k}^1$	$\log_2(\widetilde{s_1}^1)$
16	0.0286	1.7869	0.0782	1.8854
32	0.0081		0.2279	
64	0.0021		0.0078	

Πίνακας 4.17: Πολλαπλότητα των ριζών ($\mathrm{Re}(\mathfrak{f}_{14})$).

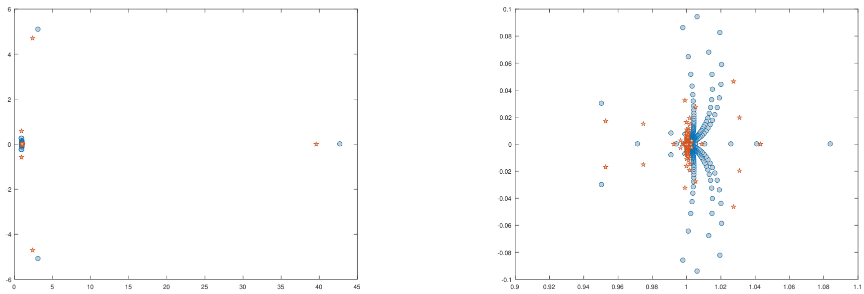
k	$\widetilde{\lambda}_{0,k}^2$	$\log_2(\widetilde{s_0}^2)$	$\widetilde{\lambda}_{1,k}^2$	$\log_2(\widetilde{s_1}^2)$
16	0.0362	0.6224	0.0381	0.7394
32	0.0207		0.0210	
64	0.0107		0.0108	

Πίνακας 4.18: Πολλαπλότητα των ριζών ($\mathrm{Im}(\mathfrak{f}_{14})$).

Οι αριθμοί επαναλήψεων και οι χρόνοι εκτέλεσης, με χρήση της PGMRES και του προτεινόμενου προρρυθμιστή, καθώς επίσης και του ταινιωτού-επί-κυκλοειδή προρρυθμιστή του προηγούμενου κεφαλαίου, όπου η γεννήτρια συνάρτηση θεωρείται γνωστή εκ των προτέρων, δίνεται στον Πίνακα 4.19. Σε αυτόν παρατηρούμε την καλή συμπεριφορά και αποτελεσματικότητα του $\mathcal{BC}_n$, αφού οι επαναλήψεις που λαμβάνουμε είναι κοντά με τις επαναλήψεις κατόπιν χρήσεως του $\mathcal{BC}_n(\mathfrak{f}_{14})$.

n	I_n	$\mathcal{BC}_n(\mathfrak{f}_{14})$	CPU	$\mathcal{BC}_n$	CPU
1024	>500	7	0.2286	9	0.6108
2048	>500	7	0.3982	8	0.7936
4096	>500	6	0.7771	8	1.2602
8192	>500	6	1.4390	8	2.7490

Πίνακας 4.19: PGMRES: Επαναλήψεις και χρόνοι CPU ($\mathfrak{f}_{14}$).



(α') Ιδιοτιμές όταν $n = 1024$ και $n = 8192$. (β') Ιδιοτιμές κοντά στο $(1, 0)$.

Σχήμα 4.16: Ιδιοτιμές ($\mathfrak{f}_{14}$).

‘Οπως και στο προηγούμενο παράδειγμα, αναμένουμε ένα σφάλμα στην εκτίμηση των ριζών $\varepsilon_1 = \mathcal{O}\left(\frac{\log n}{n}\right)$. Ωστόσο, αυτό φαίνεται να είναι $\varepsilon_1 \sim \frac{1}{n}$. ‘Οπως αποδείχθηκε στο Θεώρημα 4.2.6, οι ιδιοτιμές του τελευταίου παράγοντα, $T_n\left(\frac{\mathfrak{f}_{14}}{g}\right) C_n^{-1}\left(\frac{\mathfrak{f}_{14}}{g}\right)$, παρουσιάζουν κύρια συσσώρευση στο σημείο $(1, 0)$, αφού αυτή η συνάρτηση είναι συνεχής. Για τον $T_n^{-1}(g_n)T_n(g)$, θεωρήσαμε ως $s_n \sim \frac{\log n}{n}$ για να λάβουμε τη γενική συσσώρευση των ιδιοτιμών, σε μια περιοχή του $(1, 0)$, με ακτίνα $r_n = \mathcal{O}\left(\frac{1}{\log n}\right)$ και $\mathcal{O}(\log n)$ ιδιοτιμές εκτός της συσσώρευσης. Για τον πρώτο παράγοντα $C_n\left(\frac{\mathfrak{f}_{14} g_n}{F_{n-1} g}\right)$, επειδή η $\mathfrak{f}_{14}$ είναι συνεχής, αλλά όχι επαρκώς ομαλή, μέσω του Θεωρήματος 4.2.6 καταλήγουμε σε κύρια συσσώρευση των ιδιοτιμών σε μια περιοχή του $(1, 0)$, με ακτίνα $r_n = \mathcal{O}\left(\frac{\log n}{n}\right)$. Επομένως, από το ίδιο θεώρημα καταλήγουμε στη γενική συσσώρευση των ιδιοτιμών του προρρυθμισμένου πίνακα σε μια περιοχή του $(1, 0)$, με ακτίνα $\mathcal{O}\left(\frac{1}{\log n}\right)$ και $\mathcal{O}(\log n)$ ιδιοτιμές εκτός της συσσώρευσης. Στο Σχήμα 4.16 παρουσιάζουμε τη συσσώρευση των ιδιοτιμών όταν $n = 1024$ (μπλε κύκλοι) και $n = 8192$ (πορτοκαλί πεντάγραμμα). Στο Σχήμα 4.16α' μπορούμε εύκολα να ξεχωρίσουμε τις περιοχές συσσώρευσης και βλέπουμε ότι η περιοχή που αντιστοιχεί στη διάσταση $n = 8192$ είναι πολύ μικρότερη από αυτή για $n = 1024$. Με άλλα λόγια, όσο μεγαλύτερη είναι η διάσταση n, τόσο μικρότερη γίνεται η περιοχή συσσώρευσης. Το Σχήμα 4.16β' είναι μια μεγέθυνση της συσσώρευσης των ιδιοτιμών κοντά στο $(1, 0)$, με ακτίνα $\frac{1}{\log 1024} = 0.1$.

Σύνοψη

Στην παρούσα διατριβή μελετήσαμε και προτείναμε προρρυθμιστές για την ταχεία επίλυση μη-συμμετρικών και πραγματικών συστημάτων Toeplitz, με μεθόδους υποχώρων Krylov και πιο συγκεκριμένα με την Προρρυθμισμένη Γενικευμένη μέθοδο Ελαχίστων Υπολοίπων (PGMRES) και την Προρρυθμισμένη μέθοδο Συζυγών Κλίσεων για το σύστημα των Κανονικών Εξισώσεων (PCGN). Οι προρρυθμιστές που προτάθηκαν και των οποίων αποδείχθηκε η αποτελεσματικότητα ανήκουν στις κατηγορίες των ταινιωτών πινάκων Toeplitz και των κυκλοειδών πινάκων.

Πιο συγκεκριμένα, μετά από μια σύντομη ιστορική αναδρομή στην προρρύθμιση συστημάτων Toeplitz, στο πρώτο κεφάλαιο δόθηκαν οι βασικοί ορισμοί και κάποια χρήσιμα θεωρητικά αποτελέσματα σχετικά με τη συσσώρευση του φάσματος. Ακολούθως, στο δεύτερο κεφάλαιο προτάθηκε ως προρρυθμιστής ένας ταινιωτός πίνακας Toeplitz, ο οποίος προκύπτει κατόπιν άρσης των ριζών της, γνωστής εκ των προτέρων, γεννήτριας συνάρτησης και βέλτιστης ομοιόμορφης προσέγγισης ή παρεμβολής, με τριγωνομετρικά πολυώνυμα. Τα αποτελέσματα που παρουσιάστηκαν σε αυτό μπορούν να βρεθούν στη [49]. Στο επόμενο κεφάλαιο μελετήθηκε ένα είδος κυκλοειδή προρρυθμιστή για μη-συμμετρικά συστήματα Toeplitz με καλή κατάσταση, δηλαδή συστήματα των οποίων η γεννήτρια συνάρτηση δεν έχει ρίζες. Σε αυτό έγινε επίσης μελέτη ενός ταινιωτού-επί-κυκλοειδή προρρυθμιστή, για συστήματα με κακή κατάσταση, δηλαδή συστήματα που παράγονται από κάποια γεννήτρια συνάρτηση με ρίζες. Τα αποτελέσματα αυτού του κεφαλαίου μπορούν να βρεθούν στη [50]. Αναφέρουμε ότι μελετήθηκε τόσο η συνεχής, όσο και η κατά τμήματα συνεχής περίπτωση και ότι η γεννήτρια συνάρτηση σε αυτό το κεφάλαιο θεωρήθηκε επίσης γνωστή εκ των προτέρων. Αυτή η θεώρηση δεν έγινε στο τελευταίο κεφάλαιο της διατριβής, όπου μελετήθηκαν κατάλληλα προσαρμοσμένοι προρρυθμιστές των προηγούμενων κεφαλαίων για συστήματα με άγνωστη γεννήτρια συνάρτηση. Έγινε εκτίμηση των πιθανών ριζών και των πολλαπλοτήτων αυτών, από τις τιμές του πίνακα συντελεστών με χρήση του αναπτύγματος Fourier. Η προσαρμογή των ταινιωτών Toeplitz προρρυθμιστών δημοσιεύθηκε στα πρακτικά διεθνούς επιστημονικού συνεδρίου [17] και αυτή των ταινιωτών-επί-κυκλοειδών προρρυθμιστών μπορεί να βρεθεί στην εργασία [18]. Στο τέλος του κάθε κεφαλαίου δόθηκαν διάφορα αριθμητικά παραδείγματα, στα οποία φανερώνεται η αποτελεσματικότητα των προτεινόμενων προρρυθμιστών.

Η αναγκαιότητα της παραπάνω έρευνας προέκυψε από το γεγονός ότι σε πολλές εφαρμογές, όπως στην επεξεργασία εικόνας, στην επεξεργασία σήματος και σε εφαρμογές που προκύπτουν από τη διακριτοποίηση διαφορικών εξισώσεων, εμφανίζονται μη-συμμετρικά και πραγματικά συστήματα Toeplitz. Σε ακόμη περισσότερες εφαρμογές εμφανίζονται συστήματα των οποίων ο πίνακας συντελεστών είναι σχεδόν Toeplitz (quasi-Toeplitz). Τέτοιοι πίνακες προκύπτουν κυρίως από τη διακριτοποίηση διαφορικών και ολοκληρωτικών εξισώσεων και είναι Toeplitz πίνακες με μια διαταραχή της (Toeplitz) δομής στο άνω αριστερά και κάτω δεξιά μέρος του πίνακα ή ""Toeplitz συν διαγώνιος"" πίνακας ή ""Toeplitz συν ταινιωτός"" πίνακας με σχετικά μικρό πλάτος ταινίας. Εκτιμούμε ότι οι προτεινόμενοι προρρυθμιστές με κατάλληλη προσαρμογή θα είναι αποτελεσματικοί για τέτοιου είδους συστήματα και αυτό αποτελεί σκέψεις για μελλοντική έρευνα.

Ιδιαίτερο ενδιαφέρον, κυρίως στις εφαρμογές, παρουσιάζουν τα συστήματα που έχουν ως πίνακα αγνώστων δι-διάστατους (ή γενικότερα d-διάστατους) πίνακες Toeplitz που προκύπτουν από διακριτοποίηση μερικών διαφορικών εξισώσεων κυρίως συνοριακών προβλημάτων, καθώς και ρητών διαφορικών εξισώσεων (fractional differential equations). Εδώ προκύπτουν ιδιαίτερες δυσκολίες ως προς την υπολογιστική πολυπλοκότητα των αλγορίθμων, κυρίως των ταινιωτών προρρυθμιστών, καθώς και στην ανάπτυξη θεωρίας σχετιζόμενης με τη συσσώρευση του φάσματος, κυρίως στους δι-διάστατους κυκλοειδείς προρρυθμιστές για τους οποίους έχουν αποδειχθεί αρνητικά αποτελέσματα [54, 55, 76]. Ωστόσο ο συνδυασμός δι-διάστατων ταινιωτών Toeplitz και κυκλοειδών πινάκων φαίνεται να αντιμετωπίζει τα προβλήματα με αποτελεσματικότητα, όπως στη συμμετρική περίπτωση [53]. Ιδιαίτερη δυσκολία παρουσιάζουν δι-διάστατα προβλήματα, όπου η γεννήτρια συνάρτηση είναι άγνωστη, στην εκτίμηση των ριζών και των αντίστοιχων πολλαπλοτήτων. Ωστόσο, θετικά αποτελέσματα που αφορούν στη συμμετρική περίπτωση [56, 57, 58] είναι ενθαρρυντικά. 'Ολα τα παραπάνω αποτελούν σκέψεις για περαιτέρω έρευνα.

Βιβλιογραφία

[1] Avram, F. On bilinear forms in Gaussian random variables and Toeplitz matrices. Probab. Theory Relat. Fields., 79:37 – 45, 1988.

[2] Bai, Z.-Z., Chan, R.H. and Ren, Z.-R. On sinc discretization and banded preconditioning for linear third-order ordinary differential equations. Numer. Linear Algebra Appl., 18(3):471 – 497, 2011.

[3] Bai, Z.-Z., Huang, Y.-M. and Ng, M.K. On preconditioned iterative methods for certain time-dependent partial differential equations. SIAM J. Numer. Anal., 47(2):1019 – 1037, 2009.

[4] Bendixson, I. Sur les racines d'une équation fondamentale. Acta Math., 25:359 – 365, 1902.

[5] Bini, D. and Di Benedetto, F. A New Preconditioner for the Parallel Solution of Positive Definite Toeplitz Systems, In: Proceedings of the Second Annual ACM Symposium on Parallel Algorithms and Architectures, 220 – 223. 1990.

[6] Bôcher, M. Introduction to the Theory of Fourier's Series. Ann. Math., 7(3):81 – 152, 1906.

[7] Bogoya, J.M., Böttcher, A., Grudsky, S.M. and Maximenko, E.A. Eigenvalues of Hermitian Toeplitz matrices with smooth simple-loop symbols. J. Math. Anal. Appl., 422(2):1308 – 1334, 2015.

[8] Bunch, J.R. Stability of methods for solving Toeplitz systems of equations. SIAM J. Sci. Stat. Comput., 6(2):349 – 364, 1985.

[9] Bunch, J.R. The weak and strong stability of algorithms in numerical linear algebra. Linear Algebra Appl., 88:49 – 66, 1987.

[10] Chan, R.H. Toeplitz preconditioners for Toeplitz systems with nonnegative generating functions. IMA J. Numer. Anal., 11(3):333 – 345, 1991.

[11] Chan, R.H. and Jin, X.-Q. An Introduction to Iterative Toeplitz Solvers. SIAM, Philadelphia, 2007.

[12] Chan, R.H. and Ng, M.K. Conjugate gradient methods for Toeplitz systems. SIAM Rev., 38(3):427 – 482, 1996.

[13] Chan, R.H. and Tang, T.P. Fast band–Toeplitz preconditioners for Hermitian Toeplitz systems. SIAM J. Sci. Stat. Comput., 15:164 – 171, 1994.

[14] Chan, R.H. and Yeung, M.-C. Circulant preconditioners for complex Toeplitz matrices. SIAM J. Numer. Anal., 30(4):1193 – 1207, 1993.

[15] Chan, R.H., Potts, D. and Steidl, G. Preconditioners for nondefinite Hermitian Toeplitz systems. SIAM J. Matrix Anal. Appl., 22(3):647 – 665, 2001.

[16] Chan, T.F. An optimal circulant preconditioner for Toeplitz systems. SIAM J. Sci. Statist. Comput., 9(4):766 – 771, 1988.

[17] Chaysri, T., Hadjidimos, A., Noutsos, D. and Tachyridis, G. Band preconditioners for non-symmetric real Toeplitz systems with unknown generating function, In: Proceedings of 25th International Conference on Circuits, Systems, Communications and Computers, 86 – 96, IEEE. 2021.

[18] Chaysri, T., Hadjidimos, A., Noutsos, D. and Tachyridis, G. Band-times-circulant preconditioners for non-symmetric real Toeplitz systems with unknown generating function. 2022. https://doi.org/10.4208/eajam.230721.251121.

[19] Dai, H., Geary, Z. and Kadanoff, L.P. Asymptotics of eigenvalues and eigenvectors of Toeplitz matrices. J. Stat. Mech. Theory Exp., 2009(05):P05012, 2009.

[20] Davis, P.J. Circulant Matrices. John Wiley & Sons, New York, 1979.

[21] Demmel, J.W. Applied Numerical Linear Algebra. SIAM, Philadelphia, 1997.

[22] Di Benedetto, F. and Serra-Capizzano, S. A unifying approach to abstract matrix algebra preconditioning. Numer. Math., 82:57 – 90, 1999.

[23] Duffy, D.J. Finite Difference Methods in Financial Engineering: a Partial Differential Equation approach. John Wiley & Sons, 2013.

[24] Durbin, J. The fitting of time-series models. Rev. Inst. Int. Stat., pages 233 – 244, 1960.

[25] Ekström, S.-E. and Garoni, C. A matrix-less and parallel interpolation-extrapolation algorithm for computing the eigenvalues of preconditioned banded symmetric Toeplitz matrices. Numer. Algor., 80(3):819 – 848, 2019.

[26] Ekström, S.-E. and Serra-Capizzano, S. Eigenvalues and eigenvectors of banded Toeplitz matrices and the related symbols. Numer. Linear Algebra Appl., 25(5):e2137, 2018.

[27] Ekström, S.-E., Garoni, C. and Serra-Capizzano, S. Are the eigenvalues of banded symmetric Toeplitz matrices known in almost closed form? Exp. Math., 27(4):478 – 487, 2018.

[28] Ferrari, P., Barakitis, N. and Serra-Capizzano, S. Asymptotic spectra of large matrices coming from the symmetrization of Toeplitz structure functions and applications to preconditioning. Numer. Linear Algebra Appl., 28(1):e2332, 2021.

[29] Ferrari, P., Furci, I., Hon, S., Mursaleen, M.A. and Serra-Capizzano, S. The eigenvalue distribution of special 2-by-2 block matrix-sequences with applications to the case of symmetrized Toeplitz structures. SIAM J. Matrix Anal. Appl., 40(3):1066 – 1086, 2019.

[30] Garren, K.R. Bounds for the Eigenvalues of a Matrix, In: Dissertations, Theses, and Masters Projects. Paper 1539624585. 1965.

[31] Gmati, N. and Philippe, B. Comments on the GMRES convergence for preconditioned systems, In: Proceedings of the International Conference on Large-Scale Scientific Computing, 40 – 51. 2007.

[32] Golub, G.H. and Van Loan, C.F. Matrix Computations, 4th edition. Johns Hopkins University Press, 2013.

[33] Grenander, U. and Szego, G. Toeplitz Forms and Their Applications. Chelsea, New York, 1984.

[34] Hestenes M.R. and Stiefel, E. Methods of conjugate gradients for solving linear systems. J. Res. Nat. Bur. Stand., 49:409 – 435, 1952.

[35] Hewitt, E. and Hewitt, R.E. The Gibbs-Wilbraham phenomenon: an episode in Fourier analysis. Arch. Hist. Exact Sci., 21(2):129 – 160, 1979.

[36] Higham, N.J. Functions of Matrices: Theory and Computation. SIAM, Philadelphia, PA, 2008.

[37] Hirsch, M. Sur les racines d'une èquation fondamentale. Acta Math., 25:367 – 370, 1902.

[38] Hon, S. and Wathen, A.J. Circulant preconditioners for analytic functions of Toeplitz matrices. Numer. Algor., 79(4):1211 – 1230, 2018.

[39] Horn, R.A. and Johnson, C.R. Topics in Matrix Analysis. Cambridge University Press, Cambridge, 1994.

[40] Jin, X.-Q. Developments and Applications of Block Toeplitz Iterative Solvers. Science Press, Beijing, 2006.

[41] Korovkin, P. P. Linear Operators and Approximation Theory. Hindustan Publishing Corp, Delhi, 1960.

[42] Kra, I. and Simanca, S. On circulant matrices. Notices of the AMS, 59(3):368 – 377, 2012.

[43] Kressner, D. and Luce, R. Fast computation of the matrix exponential for a Toeplitz matrix. SIAM J. Matrix Anal. Appl., 39(1):23 – 47, 2018.

[44] Levinson, N. The Wiener (root mean square) error criterion in filter design and prediction. J. Math. Phys., 25(1-4):261 – 278, 1946.

[45] Nachtigal, N.M., Reddy, S.C. and Trefethen, L.N. How fast are nonsymmetric matrix iterations? SIAM J. Matrix Anal. Appl., 13(3):778 – 795, 1992.

[46] Ng, M.K. Band preconditioners for block-Toeplitz-Toeplitz-block systems. Linear Algebra Appl., 259:307 – 327, 1997.

[47] Ng, M.K. Iterative Methods for Toeplitz Systems. Oxford University Press, New York, 2004.

[48] Noschese, S., Pasquini, L. and Reichel, L. Tridiagonal Toeplitz matrices: properties and novel applications. Numer. Linear Algebra Appl., 20(2):302 – 326, 2013.

[49] Noutsos, D. and Tachyridis, G. Band Toeplitz preconditioners for non-symmetric real Toeplitz systems by preconditioned GMRES method. J. Comput. Appl., 373:112250, 2020.

[50] Noutsos, D. and Tachyridis, G. Band-times-circulant preconditioners for non-symmetric Toeplitz systems. BIT Numer. Math., 2021. https://doi.org/10.1007/s10543-021-00883-y.

[51] Noutsos, D. and Vassalos, P. New band Toeplitz preconditioners for ill-conditioned symmetric positive definite Toeplitz systems. SIAM J. Matrix Anal. Appl., 23:728 – 743, 2002.

[52] Noutsos, D. and Vassalos, P. Superlinear convergence for PCG using band plus algebra preconditioners for Toeplitz systems. Comput. Math. with Appl., 56(5):1255 – 1270, 2008.

[53] Noutsos, D. and Vassalos, P. Band plus algebra preconditioners for two-level Toeplitz systems. BIT Numer. Math., 51:659 – 719, 2011.

[54] Noutsos, D., Serra-Capizzano, S. and Vassalos, P. Spectral equivalence and matrix algebra preconditioners for multilevel Toeplitz systems: a negative result. Contemp. Math., 323:313 – 322, 2003.

[55] Noutsos, D., Serra-Capizzano, S. and Vassalos, P. Matrix algebra preconditioners for multilevel Toeplitz systems do not insure optimal convergence rate. Theor. Comput. Sci., 315(2 - 3):557 – 579, 2004.

[56] Noutsos, D., Serra-Capizzano, S. and Vassalos, P. A preconditioning proposal for ill-conditioned Hermitian two-level Toeplitz systems. Numer. Linear Algebr. with Appl., 12:231 – 239, 2005.

[57] Noutsos, D., Serra-Capizzano, S. and Vassalos, P. Block band Toeplitz preconditioners derived from generating function approximations: Analysis and applications. Numer. Math., 104:339 – 376, 2006.

[58] Noutsos, D., Serra-Capizzano, S. and Vassalos, P. Two-level Toeplitz preconditioning: approximation results for matrices and functions. SIAM J. Sci. Comput., 28:439 – 458, 2006.

[59] Noutsos, D., Serra-Capizzano, S. and Vassalos, P. Essential spectral equivalence via multiple step preconditioning and applications to ill conditioned Toeplitz matrices. Linear Algebra Appl., 491:276 – 291, 2016.

[60] Paige, C.C. and Saunders, M.A. Solution of sparse indefinite systems of linear equations. SIAM J. Numer. Anal., 12(4):617 – 629, 1975.

[61] Parter, S.V. On the distribution of singular values of Toeplitz matrices. Linear Algebr. Appl., 80:115 – 130, 1986.

[62] Pestana, J. and Wathen, A.J. A preconditioned MINRES method for nonsymmetric Toeplitz matrices. SIAM J. Matrix Anal. Appl., 36(1):273 – 288, 2015.

[63] Potts, D. and Steidl, G. Optimal trigonometric preconditioners for nonsymmetric Toeplitz systems. Linear Algebra Appl., 281(1 – 3):265 – 292, 1998.

[64] Potts, D. and Steidl, G. Preconditioners for ill-conditioned Toeplitz matrices. BIT Numer. Math., 39(3):513 – 533, 1999.

[65] Reid, J.K. On the method of conjugate gradients for the solution of large sparse systems of linear equations, In: Proceedings of the Oxford conference of institute of mathmatics and its applications, 231 – 254. 1971.

[66] Rivlin, T.J. An Introduction to the Approximation of Functions. Courier Corporation, 1981.

[67] Saad, Y. Iterative Methods for Sparse Linear Systems. SIAM, Philadelphia, 2003.

[68] Saad, Y. and Schultz, M.H. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Statist. Comp., 7(3):856 – 869, 1986.

[69] Saad, Y. and Schultz M.H. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput., 7:856 – 869, 1986.

[70] Sachs, E.W. and Strauss, A.K. Efficient solution of a partial integro-differential equation in finance. Appl. Numer. Math., 58(11):1687 – 1703, 2008.

[71] Serra-Capizzano, S. How to choose the best iterative strategy for symmetric Toeplitz systems. SIAM J. Numer. Anal., 36:1078 – 1103, 1999.

[72] Serra-Capizzano, S. Some theorems on linear positive operators and functionals and their applications. Comput. Math. Appl., 39(7 - 8):139 – 167, 2000.

[73] Serra-Capizzano, S. Spectral behavior of matrix sequences and discretized boundary value problems. Linear Algebra Appl., 337(1 – 3):37 – 78, 2001.

[74] Serra-Capizzano, S. Matrix algebra preconditioners for multilevel Toeplitz matrices are not superlinear. Linear Algebra Appl., 343:303 – 319, 2002.

[75] Serra-Capizzano, S. Practical band Toeplitz preconditioning and boundary layer effects. Numer. Algorithms, 34(2):427 – 440, 2003.

[76] Serra-Capizzano, S. and Tyrtyshnikov, E.E. Any circulant-like preconditioner for multilevel matrices is not superlinear. SIAM J. Matrix Anal. Appl., 21:431 – 439, 2000.

[77] Serra-Capizzano, S. and Tyrtyshnikov, E.E. How to prove that a preconditioner cannot be superlinear. Math. Comp., 72(243):1305 – 1316, 2003.

[78] Serra, S. Optimal, quasi-optimal and superlinear band-Toeplitz preconditioners for asymptotically ill-conditioned positive definite Toeplitz systems. Math. Comp., 66(218):651 – 665, 1997.

[79] Sohrab, H.H. Basic Real Analysis. Birkhäuser, New York, 2003.

[80] Strang, G. A Proposal for Toeplitz Matrix Calculations. Studies in Applied Mathematics, 74(2):171 – 176, 1986.

[81] Tilli, P. Singular values and eigenvalues of non-Hermitian block Toeplitz matrices. Linear Algebra Appl., 272:59 – 89, 1998.

[82] Tilli, P. Some results on complex Toeplitz eigenvalues. J. Math. Anal. Appl., 239:390 – 401, 1999.

[83] Trench, W.F. An algorithm for the inversion of finite Toeplitz matrices. J. SIAM, 12(3):515 – 522, 1964.

[84] Tyrtyshnikov, E.E. Influence of matrix operations on the distribution of eigenvalues and singular values of Toeplitz matrices. Linear Algebra Appl., 207:225 – 249, 1994.

[85] Tyrtyshnikov, E.E. Circulant preconditioners with unbounded inverses. Linear Algebra Appl., 216:1 – 23, 1995.

[86] Tyrtyshnikov, E.E. A unifying approach to some old and new theorems on distribution and clustering. Linear Algebra Appl., 232:1 – 43, 1996.

[87] Tyrtyshnikov, E.E. and Zamarashkin, N.L. Thin structure of eigenvalue clusters for non-Hermitian Toeplitz matrices. Linear Algebra Appl., 292:297 – 310, 1999.

[88] Widom, H. Hankel matrices. Trans. Am. Math. Soc., 121(1):1 – 35, 1966.

[89] Yeung, M.-C. and Chan, R.H. Circulant preconditioners for Toeplitz matrices with piecewise continuous generating functions. Math Comput, 61(204):701 – 718, 1993.

[90] Zhan, X. Matrix theory. vol. 147, American Mathematical Society, 2013.