

## Κεφάλαιο 8

# Ιδιοτιμές και ιδιαζουσες τιμές

Σε αυτό το κεφάλαιο θα μελετήσουμε αλγόριθμους για την επίλυση του προβλήματος ιδιοτιμών

$$A\mathbf{x} = \lambda \mathbf{x}$$

και για τον υπολογισμό της διάσπασης ιδιαζουσών τιμών (singular value decomposition, SVD)

$$A = U\Sigma V^T$$

Θα αναφερθούμε επίσης σε σχετικές εφαρμογές.

Θυμηθείτε από την Ενότητα 4.1 (σελίδα 137) ότι οι ιδιοτιμές, άρα και οι ιδιαζουσες τιμές, δεν μπορούν γενικά να υπολογιστούν επακριβώς σε πεπερασμένο πλήθος βημάτων, ακόμα και αν δεν υπάρχει σφάλμα κινητής υποδιαστολής. Επομένως, όλοι οι αλγόριθμοι για τον υπολογισμό ιδιοτιμών και ιδιαζουσών τιμών είναι υποχρεωτικά επαναληπτικοί, σε αντίθεση με τους αλγόριθμους των Κεφαλαίων 5 και 6.

Οι μέθοδοι για την εύρεση ιδιοτιμών μπορούν να χωριστούν σε δύο κατηγορίες. Η πρώτη βασίζεται σε παραγοντοποιήσεις που περιλαμβάνουν μετασχηματισμούς ομοιότητας (similarity transformations) για την εύρεση αρκετών (ή όλων των) ιδιοτιμών. Η άλλη κατηγορία ασχολείται κυρίως με πολύ μεγάλες και συνήθως αραιές μήτρες για τις οποίες αναζητούνται μόνο λίγες ιδιοτιμές ή/και ιδιοδιανύσματα: γι' αυτές τις περιπτώσεις υπάρχουν αλγόριθμοι που βασίζονται σε μεγάλο βαθμό σε γινόμενα μήτρας-διανύσματος.

Θα ξεκινήσουμε στην Ενότητα 8.1 περιγράφοντας το αποτέλεσμα του επανειλημμένου πολλαπλασιασμού ενός διανύσματος με τη δοθείσα μήτρα, το οποίο μπορεί να σας θυμίσει τις Ενότητες 7.4 και 7.5. Αυτό θα μας οδηγήσει στη θεμελιώδη μέθοδο δυνάμεων, η οποία βρίσκει εφαρμογή σε αλγόριθμους αναζήτησης για μεγά-

λα δίκτυα. Στη συνέχεια θα παρουσιάσουμε τις σημαντικές έννοιες της μετατόπισης (shift) και της αντιστροφής (invert), που μας οδηγούν στην αντίστροφη επανάληψη (inverse iteration). Θα δούμε πώς αυτή μπορεί να επιταχύνει τη σύγκλιση, με το ανάλογο αντίτιμο.

Στην Ενότητα 8.2 θα στρέψουμε την προσοχή μας στη διάσπαση ιδιαζουσών τιμών, την οποία παρουσιάσαμε στην Ενότητα 4.4, και θα εξηγήσουμε πώς μπορεί να χρησιμοποιηθεί στην πράξη. Μια τέτοια εφαρμογή θα μας φέρει ξανά στην επί-λυση προβλημάτων ελάχιστων τετραγώνων όπως στο Κεφάλαιο 6, με τη σημαντική διαφορά όμως ότι οι στήλες της δοθείσας μήτρας μπορεί να είναι γραμμικά εξαρτημένες (linearly dependent).

Οι πλήρεις περιγραφές ανθεκτικών αλγόριθμων για την εύρεση όλων των ιδιοτιμών ή ιδιαζουσών τιμών μιας γενικής μήτρας περιλαμβάνουν πολλές μη τετριμμένες πτυχές και ανήκουν σε πιο εξειδικευμένα βιβλία. Θα αναφέρουμε κάποιες λεπτομέρειες τέτοιων αλγόριθμων στην Ενότητα 8.3.

## 8.1 Η μέθοδος δυνάμεων και οι παραλλαγές της

Η μέθοδος δυνάμεων είναι μια απλή τεχνική για τον υπολογισμό της κυρίαρχης ιδιοτιμής και του κυρίαρχου ιδιοδιανύσματος μιας μήτρας. Βασίζεται στην ιδέα ότι ο επανειλημμένος πολλαπλασιασμός ενός (σχεδόν οποιουδήποτε) τυχαίου διανύσματος με τη δοθείσα μήτρα δίνει διανύσματα που τείνουν στην κατεύθυνση του κυρίαρχου ιδιοδιανύσματος.

**Σημείωση:** Σε σύγκριση με την περιγραφή των Κεφαλαίων 5 έως 7, ακολουθούμε την αντίστροφη σειρά εδώ: Πρώτα περιγράφουμε μεθόδους που βασίζονται σε γινόμενα μήτρας-διανύσματος και αποσκοπούν δυνητικά στην εύρεση λίγων μόνο ιδιοτιμών, αναβάλλοντας για την Ενότητα 8.3 την περιγραφή των γενικών μεθόδων εύρεσης όλων των ιδιοτιμών για μήτρες χωρίς ειδική δομή. Ο λόγος γι' αυτό είναι ότι οι μέθοδοι της δεύτερης κατηγορίας χρησιμοποιούν ως δομικά στοιχεία τις μεθόδους που παρουσιάζονται στην Ενότητα 8.1.

Όμως γιατί να θέλει κανείς να υπολογίσει το κυρίαρχο ιδιοδιάνυσμα μιας μήτρας; Ακολουθεί μια περίπτωση που θα αποκαλύψει τα σχετικά κίνητρα.

**Παράδειγμα 8.1.** Η μηχανή αναζήτησης (search engine) της Google είναι η κυρίαρχη τεχνολογία αναζήτησης στο Διαδίκτυο σήμερα: Πρόκειται για έναν μηχανισμό κατάταξης σελίδων και εμφάνισης των κορυφαίων αποτελεσμάτων –δηλαδή ιστοσελίδων οι οποίες, σύμφωνα με την κρίση της μηχανής αναζήτησης, είναι οι τοποθεσίες που εμφανίζονται με τη μεγαλύτερη συνάφεια με το ερώτημα του χρήστη.

Η εταιρεία δεν θέλει να αποκαλύψει τις ακριβείς λεπτομέρειες των αλγόριθμών της –μάλιστα, οι φήμες λένε ότι αυτοί τροποποιούνται διαρκώς. Ο βασικός αλγόριθμος που έχει οδηγήσει στη δημιουργία αυτού του διαδικτυακού κολοσσού αναζήτησης ονομάζεται **PageRank** και δημοσιεύθηκε το 1998 από τους ιδρυτές της Google, τον Sergey Brin και τον Larry Page. Βασίζεται στον υπολογισμό του κυρίαρχου ιδιοδιανύσματος μιας μεγάλης και πολύ αραιής μήτρας.

Πριν εμβαθύνουμε στις δύσκολες λεπτομέρειες, όμως, πρέπει να έχουμε μια ευρύτερη εικόνα. Ο Ιστός είναι μια οντότητα τεράστιου μεγέθους η οποία έχει δισεκατομμύρια ιστοσελίδες και αλλάζει δυναμικά: Σελίδες ενημερώνονται, προστίθενται και διαγράφονται διαρκώς. Οι μηχανές αναζήτησης είναι συνεχώς απασχολημένες, χρησιμοποιώντας την ασύλληπτη υπολογιστική ισχύ τους και σαρώνοντας τον Ιστό για να ενημερώνουν τις εγγραφές τους. Όταν ένας χρήστης πληκτρολογεί ένα ερώτημα αναζήτησης (search query), η μηχανή αναζήτησης είναι έτοιμη με δισεκατομμύρια εγγραφές και, ανάλογα με τις λεπτομέρειες του ερωτήματος, μπορεί να ανακτήσει αμέσως τις σχετικές ιστοσελίδες. Το δύσκολο κομμάτι είναι η κατάταξη αυτών των σχετικών ιστοσελίδων και ο καθορισμός της σειράς με την οποία θα εμφανιστούν στον χρήστη· όντως, ο τυπικός περιηγητής του Ιστού συνήθως περιορίζεται στο να ελέγχει λίγες μόνο από τις σελίδες που εμφανίζονται πρώτες. Η ιδέα που αποτέλεσε τη βάση για τη μηχανή αναζήτησης ήταν ότι, πέρα από τις ερωτήσεις περιεχομένου (και να είστε βέβαιοι ότι αυτές δεν παραβλέπονται) είναι επίσης σημαντικό να λαμβάνεται υπόψη η δομή των συνδέσμων (links) του Ιστού.

Έστω ότι δίνεται ένα γράφημα (ή γράφος) συνδέσμων δικτύου (network linkage graph) με  $n$  κόμβους (ιστοσελίδες)· η σπουδαιότητα μιας ιστοσελίδας καθορίζεται από το πλήθος και τη σπουδαιότητα των σελίδων που συνδέονται με αυτή. Με μαθηματικούς όρους, έστω ότι η σπουδαιότητα, ή κατάταξη (rank),<sup>42</sup> της σελίδας  $i$  δίνεται από το  $x_i$ . Για να βρούμε την τιμή του  $x_i$ , πρώτα καταγράφουμε όλες τις σελίδες που συνδέονται με τη σελίδα που μας ενδιαφέρει. Ας υποθέσουμε ότι οι θέσεις, ή δείκτες (indices), αυτών των σελίδων δίνονται από το σύνολο  $\{B_i\}$ . Αν μια ιστοσελίδα της οποίας ο δείκτης είναι  $j \in B_i$  δείχνει σε  $N_j$  σελίδες, συμπεριλαμβανομένης της σελίδας  $i$ , και η δική της κατάταξη είναι  $x_j$ , τότε λέμε ότι συνεισφέρει ένα μερίδιο  $\frac{x_j}{N_j}$  στην κατάταξη  $x_i$  της σελίδας  $i$ . Επομένως, με έναν μόνο μαθηματικό τύπο μπορούμε να ορίσουμε ότι

$$x_i = \sum_{j \in B_i} \frac{1}{N_j} x_j, \quad i = 1, \dots, n$$

<sup>42</sup> Σε αυτό το πλαίσιο, ο όρος παραπέμπει περισσότερο στους βαθμούς των αξιωματικών του στρατού παρά στην τάξη (rank) μιας μήτρας.

Αν εξετάσουμε προσεκτικά την παραπάνω παράσταση, θα δούμε ότι στην πραγματικότητα δεν είναι τίποτε άλλο παρά ένα πρόβλημα ιδιοτιμών! Ψάχνουμε ένα διάνυσμα  $\mathbf{x}$  τέτοιο ώστε να ισχύει  $\mathbf{x} = A\mathbf{x}$ , όπου οι μη μηδενικές τιμές  $a_{ij}$  της μήτρας  $A$  είναι τα στοιχεία  $1/N_j$  που συσχετίζονται με τη σελίδα  $i$ . Με άλλα λόγια, θέλουμε να υπολογίσουμε ένα ιδιοδιάνυσμα της  $A$  το οποίο να αντιστοιχεί σε μια ιδιοτιμή ίση με 1. Επειδή το πλήθος των εισερχόμενων και εξερχόμενων συνδέσμων μιας ορισμένης ιστοσελίδας είναι μικρότερο κατά πολλές τάξεις μεγέθους από το συνολικό πλήθος των ιστοσελίδων, η μήτρα  $A$  είναι εξαιρετικά αραιή.

Υπάρχουν μερικά αναπάντητα ερωτήματα σε αυτό το σημείο. Για παράδειγμα, πώς ξέρουμε καν ότι η μήτρα  $A$  έχει μια ιδιοτιμή ίση με 1; Αν όντως έχει, είναι αυτή η ιδιοτιμή μεγάλη ή μικρή συγκριτικά με τις άλλες ιδιοτιμές της μήτρας; Αν υπάρχει μια λύση  $\mathbf{x}$  του προβλήματος, είναι μοναδική εκτός από την επιλογή κάποιου μη μηδενικού πολλαπλασιαστικού παράγοντα; Ακόμα και αν είναι μοναδική, είναι εγγυημένα πραγματική (δηλαδή περιέχει το διάνυσμα πραγματικούς αριθμούς); Και ακόμα και αν είναι πραγματική, θα είναι όλα τα στοιχεία του διανύσματος υποχρεωτικά θετικά, όπως θα έπρεπε για να έχουμε μια «κατάταξη»; Κάποιες από τις απαντήσεις σε αυτά τα ερωτήματα είναι απλές, ενώ άλλες είναι πιο περίπλοκες. Θα αναφερθούμε σε αυτές στο Παράδειγμα 8.3. Προς το παρόν θα σας ζητήσουμε να μας εμπιστευτείτε και να δεχθείτε ότι, με λίγες προσαρμογές αυτού του βασικού μοντέλου (δείτε το Παράδειγμα 8.3), η μήτρα  $A$  έχει όντως μια ιδιοτιμή 1 η οποία τυγχάνει να είναι η κυρίαρχη ιδιοτιμή, με αλγεβρική πολλαπλότητα 1, και ότι το πρόβλημα έχει μοναδική πραγματική θετική λύση  $\mathbf{x}$ . Αυτό ακριβώς είναι και το διάνυσμα κατάταξης που ψάχνουμε· για την εύρεσή του μπορεί να χρησιμοποιηθεί η μέθοδος δυνάμεων, την οποία θα περιγράψουμε στη συνέχεια. ■

## Η μέθοδος δυνάμεων

Έστω ότι οι ιδιοτιμές μιας μήτρας  $A$  είναι  $\{\lambda_j, \mathbf{x}_j\}$  για  $j = 1, \dots, n$ . Σημειώστε ότι εδώ τα  $\mathbf{x}_j$  είναι ιδιοδιανύσματα όπως στην Ενότητα 4.1, και όχι τιμές προσέγγισης όπως στο Κεφάλαιο 7. Έστω ότι  $\mathbf{v}_0$ ,  $\|\mathbf{v}_0\| = 1$ , είναι μια αυθαίρετη αρχική εικασία, και θεωρήστε τον παρακάτω αλγόριθμο:

for  $k = 1, 2, \dots$  μέχρι τον τερματισμό do

$$\tilde{\mathbf{v}} = A\mathbf{v}_{k-1}$$

$$\mathbf{v}_k = \tilde{\mathbf{v}} / \|\tilde{\mathbf{v}}\|$$

end

Το αποτέλεσμα αυτού του αλγόριθμου στο  $k$ -οστό βήμα είναι ένα διάνυσμα  $\mathbf{v}_k = \gamma_k A^k \mathbf{v}_0$ , όπου  $\gamma_k$  είναι ένας αριθμός ο οποίος εγγυάται ότι  $\|\mathbf{v}_k\| = 1$ . Ο λόγος για την επανειλημμένη κανονικοποίηση είναι ότι δεν υπάρχει τίποτα στον ορισμό

ενός ιδιοδιανύσματος το οποίο να αποτρέπει την αύξηση των τιμών προσέγγισης, κάτι που θα μπορούσε όντως να συμβεί για μεγάλες ιδιοτιμές, επιταχύνοντας έτσι τη συσσώρευση των σφαλμάτων στρογγυλοποίησης. Επομένως, πρέπει να κρατήσουμε το μέγεθος των τιμών προσέγγισης υπό έλεγχο.

Θα υποθέτουμε συνεχώς ότι η μήτρα  $A$  έχει  $n$  γραμμικά ανεξάρτητα ιδιοδιανύσματα. Άρα, είναι εφικτό να εκφράσουμε το  $\mathbf{v}_0$  ως γραμμικό συνδυασμό των ιδιοδιανύσματων  $\{\mathbf{x}_j\}$ : Υπάρχουν συντελεστές  $\beta_j$  τέτοιοι ώστε να ισχύει

$$\mathbf{v}_0 = \sum_{j=1}^n \beta_j \mathbf{x}_j$$

Πολλαπλασιάζοντας το  $\mathbf{v}_0$  με την  $A$  παίρνουμε

$$A\mathbf{v}_0 = A \left( \sum_{j=1}^n \beta_j \mathbf{x}_j \right) = \sum_{j=1}^n \beta_j A\mathbf{x}_j = \sum_{j=1}^n \beta_j \lambda_j \mathbf{x}_j$$

Τα ιδιοδιανύσματα που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές είναι πιο ευδιάκριτα στον νέο γραμμικό συνδυασμό. Συνεχίζοντας, για οποιονδήποτε θετικό ακέραιο  $k$  έχουμε

$$A^k \mathbf{v}_0 = \sum_{j=1}^n \beta_j \lambda_j^k \mathbf{x}_j$$

Κατόπιν υποθέτουμε ότι οι ιδιοτιμές  $\lambda_1, \lambda_2, \dots, \lambda_n$  ταξινομούνται κατά φθίνουσα σειρά ως προς το μέγεθός τους, και το μέγεθος της δεύτερης ιδιοτιμής είναι μικρότερο από το μέγεθος της πρώτης, άρα

$$|\lambda_1| > |\lambda_j|, \quad j = 2, \dots, n$$

(Δηλαδή, αποκλείεται το ενδεχόμενο της ισότητας των μεγεθών.) Έστω επίσης ότι το  $\mathbf{v}_0$  έχει μια συνιστώσα στην κατεύθυνση του  $\mathbf{x}_1$ , δηλαδή  $\beta_1 \neq 0$ . Τότε,

$$\mathbf{v}_k = \gamma_k \lambda_1^k \sum_{j=1}^n \beta_j \left( \frac{\lambda_j}{\lambda_1} \right)^k \mathbf{x}_j = \gamma_k \lambda_1^k \beta_1 \mathbf{x}_1 + \gamma_k \lambda_1^k \sum_{j=2}^n \beta_j \left( \frac{\lambda_j}{\lambda_1} \right)^k \mathbf{x}_j$$

όπου  $\gamma_k$  είναι ένας παράγοντας κανονικοποίησης για τον οποίο ισχύει  $\|\mathbf{v}_k\| = 1$ . Εφόσον έχουμε υποθέσει ότι η πρώτη ιδιοτιμή είναι κυρίαρχη, έπειτα ότι για  $j > 1$  έχουμε  $\left| \frac{\lambda_j}{\lambda_1} \right|^k \rightarrow 0$  καθώς το  $k \rightarrow \infty$ . Άρα, όσο μεγαλύτερη είναι η τιμή του  $k$ , τόσο πιο κυρίαρχο είναι το  $\mathbf{x}_1$  στο  $\mathbf{v}_k$ , και στο όριο παίρνουμε ένα μοναδιαίο διάνυσμα στην κατεύθυνση του  $\mathbf{x}_1$ .

Συνεπώς, έχουμε βρει έναν απλό τρόπο για να υπολογίζουμε κατά προσέγγιση το κυρίαρχο ιδιοδιανύσμα. Τι γίνεται όμως με την κυρίαρχη ιδιοτιμή; Στην περίπτω-

ση των Παραδειγμάτων 8.1 και 8.3 η ιδιοτιμή αυτή είναι γνωστή, αλλά στις περισσότερες περιπτώσεις δεν ισχύει κάτι τέτοιο. Εδώ μπορούμε να χρησιμοποιήσουμε το **πηλίκο Rayleigh** (Rayleigh quotient), το οποίο ορίζεται για οποιοδήποτε διάνυσμα ως

$$\mu(\mathbf{v}) = \frac{\mathbf{v}^T A \mathbf{v}}{\mathbf{v}^T \mathbf{v}}$$

Αν το  $\mathbf{v}$  ήταν ιδιοδιάνυσμα,<sup>43</sup> το  $\mu(\mathbf{v})$  θα μας έδινε απλώς τη συσχετισμένη ιδιοτιμή. Αν το  $\mathbf{v}$  δεν είναι ιδιοδιάνυσμα, το πηλίκο Rayleigh του  $\mathbf{v}$  δίνει την πλησιέστερη δυνατή προσέγγιση της ιδιοτιμής με την έννοια των ελάχιστων τετραγώνων. (Θα σας ζητηθεί να το αποδείξετε αυτό στην Άσκηση 2. Δεν είναι δύσκολο!)

Σημειώστε ότι λόγω της κανονικοποίησης του  $\mathbf{v}_k$  έχουμε

$$\mu(\mathbf{v}_k) = \mathbf{v}_k^T A \mathbf{v}_k$$

Αυτή λοιπόν είναι η εκτίμησή μας για την ιδιοτιμή  $\lambda_1$  στην  $k$ -οστή επανάληψη. Ακολουθεί η μέθοδος δυνάμεων για τον υπολογισμό του κυρίαρχου ιδιοζεύγους μιας μήτρας, με βάση τις συνθήκες που έχουμε καθορίσει μέχρι στιγμής.

### Αλγόριθμος: Μέθοδος δυνάμεων.

Είσοδος: η μήτρα  $A$  και η αρχική εικασία  $\mathbf{v}_0$ .

for  $k = 1, 2, \dots$  μέχρι τον τερματισμό

$$\tilde{\mathbf{v}} = A\mathbf{v}_{k-1}$$

$$\mathbf{v}_k = \tilde{\mathbf{v}} / \|\tilde{\mathbf{v}}\|$$

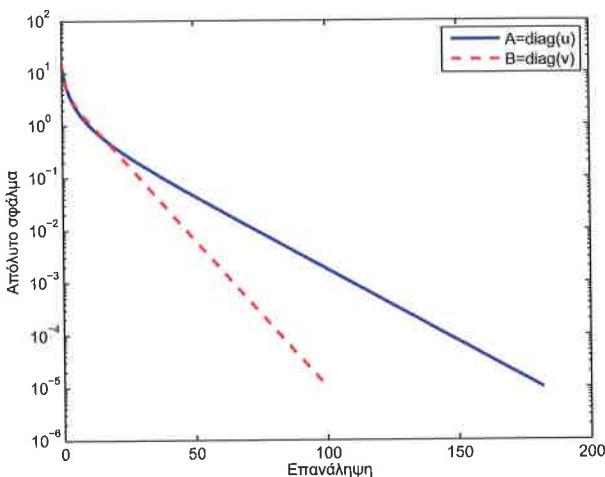
$$\lambda_1^{(k)} = \mathbf{v}_k^T A \mathbf{v}_k$$

end

Η επιλογή ενός κριτηρίου τερματισμού γι' αυτή τη μέθοδο εξαρτάται έως έναν βαθμό από τον σκοπό του υπολογισμού.

**Παράδειγμα 8.2.** Είναι αξιοσημείωτο το πόσες πληροφορίες μπορεί να μας δώσει μια διαγώνια μήτρα σχετικά με την ποιότητα μιας μεθόδου υπολογισμού ιδιοτιμών για μια πιο γενική συμμετρική μήτρα  $A$ . Αυτό οφείλεται κυρίως στο γεγονός ότι τέτοιες μήτρες έχουν φασματική διάσπαση  $A = QDQ^T$ , με την  $Q$  να είναι ορθογώνια και την  $D$  να είναι διαγώνια, οπότε η διαγώνια μήτρα εδώ είναι κατά κάποιον τρόπο αντιπροσωπευτική του «μη ορθογώνιου τμήματος». Στο Σχήμα 8.1 θα παρουσιάσουμε δύο πειράματα που εφαρμόζουν τη μέθοδο δυνάμεων σε διαγώνιες μήτρες.

<sup>43</sup> Για λόγους απλοποίησης του συμβολισμού, ας υποθέσουμε ότι το  $\mathbf{v}$  έχει μόνο πραγματικά στοιχεία. Διαφορετικά, αντικαθιστούμε το  $\mathbf{v}^T$  με  $\mathbf{v}^H$  σε όλη την παραπάνω παράσταση, όπου με  $\mathbf{v}^H$  συμβολίζουμε το «ανεστραμμένο και συζυγές»  $\mathbf{v}$ .



**Σχήμα 8.1:** Σύγκλιση της μεθόδου δυνάμεων για δύο διαγώνιες μήτρες (Παράδειγμα 8.2).

Στο MATLAB οι μήτρες ορίζονται ως εξής:

```
u = [1:32]; v = [1:30,30,32];
A = diag(u); B = diag(v);
```

Άρα, τόσο στην  $A$  όσο και στη  $B$  η μεγαλύτερη ιδιοτιμή είναι  $\lambda_1 = 32$ . Όμως, ενώ στην  $A$  η δεύτερη ιδιοτιμή είναι  $\lambda_2 = 31$ , στη  $B$  είναι  $\lambda_2 = 30$ .

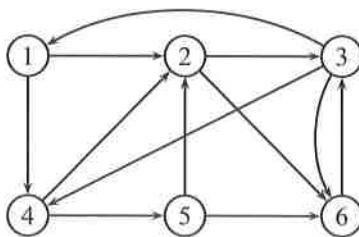
Στο σχήμα απεικονίζεται το απόλυτο σφάλμα, που είναι

$$|\lambda_1^{(k)} - \lambda_1| \equiv |\lambda_1^{(k)} - 32|$$

Τα αποτελέσματα δείχνουν τη σημαντική βελτίωση της σύγκλισης για τη  $B$ . Μπορούμε να κατανοήσουμε καλύτερη την επιταχυνθείσα σύγκλιση συγκρίνοντας τον λογάριθμο  $\log(\frac{31}{32})$  με τον  $\log(\frac{30}{32})$ . Ο λόγος μεταξύ αυτών των δύο τιμών είναι κατά προσέγγιση ίσος με 2.03, το οποίο υποδεικνύει έναν περίπου παρόμοιο παράγοντα ως προς την ταχύτητα σύγκλισης.

Στην Ασκηση 3 θα μελετήσετε ένα παρόμοιο πείραμα για μη διαγώνιες μήτρες.

**Σημείωση:** Στο Παράδειγμα 8.3 θα συνεχίσουμε τη συναρπαστική ιστορία του αλγόριθμου PageRank. Ο υπολογισμός των ιδιοδιανυσμάτων καθαυτός δεν είναι το στοιχείο που κάνει τον αλγόριθμο να ξεχωρίζει. Αν θέλετε να μάθετε λεπτομέρειες, μπορείτε να διαβάσετε το παράδειγμα. Διαφορετικά, θα μπορούσατε να το παραλείψετε χωρίς κανένα πρόβλημα.



**Σχήμα 8.2:** Ένα απλοϊκό δίκτυο για τον αλγόριθμο PageRank (Παράδειγμα 8.3).

**Παράδειγμα 8.3.** Όπως σας υποσχεθήκαμε, θα επιστρέψουμε στο πρόβλημα που περιγράφεται στο Παράδειγμα 8.1 και θα δείξουμε πώς ακριβώς η μέθοδος δυνάμεων παράγει το διάνυσμα PageRank που ορίζεται εκεί. Έστω ότι  $\mathbf{e}$  είναι ένα διάνυσμα μήκους  $n$  με στοιχεία που είναι όλα ίσα με 1, και ορίζουμε ότι  $\mathbf{v}_0 = \frac{1}{n}\mathbf{e}$ . Τότε, για  $k = 0, 1, \dots$ , η επανάληψη ορίζεται ως

$$\mathbf{v}_i^{(k+1)} = \sum_{j \in B_i} \frac{1}{N_j} \mathbf{v}_j^{(k)}, \quad i = 1, \dots, n$$

Αυτό είναι ισοδύναμο με την εφαρμογή της μεθόδου δυνάμεων χωρίς κανονικοποίηση διανυσμάτων και εκτίμηση ιδιοτιμών –πράξεις που δεν απαιτούνται για το συγκεκριμένο απλό παράδειγμα.

Για να δείτε πώς λειτουργεί αυτό, θεωρήστε τη δομή συνδέσμων του απλοϊκού δικτύου που απεικονίζεται στο Σχήμα 8.2. Εδώ, οι ιστοσελίδες είναι κόμβοι του γραφήματος, με αρίθμηση από 1 έως 6. Παρατηρήστε ότι το γράφημα είναι κατευθυνόμενο (directed) ή, ισοδύναμα, η μήτρα που συσχετίζεται με αυτό δεν είναι συμμετρική. Πράγματι, το γεγονός ότι μια σελίδα περιέχει σύνδεσμο προς μια άλλη σελίδα δεν συνεπάγεται υποχρεωτικά ότι η δεύτερη σελίδα θα περιέχει σύνδεσμο προς την πρώτη.

Το έργο της κατασκευής της μήτρας συνδέσμων έχει απλοποιηθεί πλέον. Η  $j$ -οστή στήλη αναπαριστά τους εξερχόμενους συνδέσμους από τη σελίδα  $j$ . Για παράδειγμα, η τρίτη στήλη της μήτρας υποδεικνύει ότι ο κόμβος 3 έχει εξερχόμενους συνδέσμους προς τρεις κόμβους, τους 1, 4 και 6. Σε καθέναν από αυτούς ανατίθεται η τιμή  $1/3$  στην αντίστοιχη θέση της στήλης, και τα υπόλοιπα στοιχεία της στήλης τίθενται ίσα με το 0. Έτσι προκύπτει η μήτρα

$$A = \begin{pmatrix} 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{3} & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

Στη μήτρα  $A$ , οι γραμμές αντιστοιχούν στους εισερχόμενους συνδέσμους. Για παράδειγμα, η γραμμή 2 υποδεικνύει ότι οι κόμβοι που συνδέονται με τον κόμβο 2 έχουν τους αριθμούς 1, 4 και 5. Προσέξτε ότι μόνο οι στήλες έχουν άθροισμα ίσο με 1· δεν μπορούμε να ελέγξουμε με κάποιον τρόπο το άθροισμα των στοιχείων σε οποιαδήποτε γραμμή. Μια τέτοια μήτρα ονομάζεται *στοχαστική* ως προς τις στήλες (column stochastic).<sup>44</sup>

Όλα τα στοιχεία της  $A$  είναι μη αρνητικά. Υπάρχει μια πολύ κομψή θεωρία, η-λικίας ενός αιώνα, η οποία εξασφαλίζει ότι θα υπάρχει μια μοναδική, απλή ιδιοτιμή 1 γι' αυτή τη μήτρα.<sup>45</sup> Έπειτα ότι όλες οι υπόλοιπες ιδιοτιμές της  $A$  είναι μικρότερες σε μέγεθος, και το συσχετισμένο ιδιοδιάνυσμα της κυρίαρχης ιδιοτιμής έχει πραγματικούς αριθμούς ως στοιχεία.

Για να υπολογίσουμε το PageRank, ξεκινάμε με το διάνυσμα

$$\mathbf{v}_0 = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)^T$$

και το πολλαπλασιάζουμε επανειλημμένα με τη μήτρα  $A$ . (Αυτή η επιλογή για την αρχική εικασία βασίζεται περισσότερο σε έναν «δημοκρατικό» τρόπο σκέψης παρά σε κάποια μαθηματική ή υπολογιστική διαίσθηση. Εφόσον δεν γνωρίζουμε τίποτα για τη λόση, γιατί να μην ξεκινήσουμε από ένα διάνυσμα το οποίο υποδεικνύει ότι η κατάταξη είναι ίδια για όλες τις σελίδες;) Τελικά, η μέθοδος συγκλίνει στο

<sup>44</sup> Στη σχετική βιβλιογραφία η μήτρα  $A$  συχνά συμβολίζεται με  $P^T$ , το οποίο δείχνει ότι η αναστροφή της,  $P$ , είναι στοχαστική ως προς τις γραμμές, στις οποίες περιέχει τις πληροφορίες για τους εισερχόμενους συνδέσμους ανά κόμβο. Για να αποφύγουμε τη σύγχυση στον συμβολισμό, δεν χρησιμοποιούμε την αναστροφή στο συγκεκριμένο παράδειγμα.

<sup>45</sup> Η θεωρία στην οποία αναφερόμαστε εδώ ονομάζεται θεωρία Perron-Frobenius, και μπορείτε να τη βρείτε στην πλούσια διαθέσιμη βιβλιογραφία για τις μη αρνητικές μήτρες.

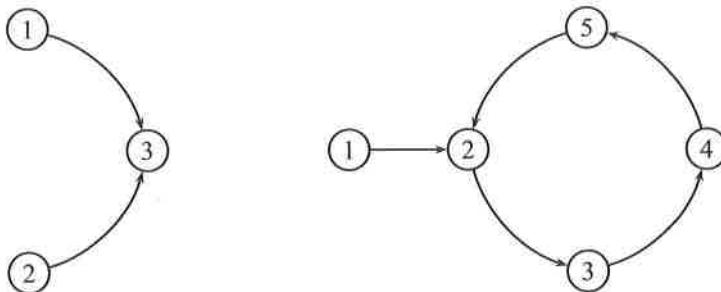
$$\mathbf{x} = \begin{pmatrix} 0.0994 \\ 0.1615 \\ 0.2981 \\ 0.1491 \\ 0.0745 \\ 0.2174 \end{pmatrix}$$

το οποίο είναι και το επιθυμητό διάνυσμα PageRank. Αυτό δείχνει ότι ο κόμβος 3 είναι το στοιχείο με την υψηλότερη κατάταξη. Στην πραγματικότητα, το αποτέλεσμα ενός ερωτήματος δεν συμπεριλαμβάνει συνήθως τις τιμές του  $\mathbf{x}$ : αντ' αυτού δίνεται η κατάσταση, η οποία σύμφωνα με τις τιμές του  $\mathbf{x}$  είναι (3, 6, 2, 4, 1, 5).

Σημειώστε ότι σε αυτή την ειδική περίπτωση, και επειδή ισχύει  $\|\mathbf{v}_0\|_1 = 1$ , όλες οι επόμενες τιμές προσέγγισης ικανοποιούν την ισότητα  $\|\mathbf{v}_k\|_1 = 1$  (γιατί;), και άρα η επιλογή μας να παραλείψουμε το βήμα της κανονικοποίησης στην επανάληψη της μεθόδου δυνάμεων είναι καθ' όλα δικαιολογημένη.

Υπάρχει επίσης μια πιθανοτική ερμηνεία αυτού του μοντέλου. Ένας τυχαίος χρήστης του Ιστού ακολουθεί συνδέσμους με τρόπο που παραπέμπει σε τυχαίο περίπατο (random walk), όπου η πιθανότητα να ακολουθηθεί ένας συγκεκριμένος σύνδεσμος από την ιστοσελίδα  $j$  δίνεται από την τιμή  $1/N_j$  της αντίστοιχης ακμής. Ουσιαστικά το ερώτημα που μας ενδιαφέρει είναι το εξής: Ποια είναι η πιθανότητα να παραμείνει ένας χρήστης σε μια συγκεκριμένη ιστοσελίδα μετά από πολύ («άπειρο») χρόνο, ανεξάρτητα από το πού ξεκίνησε την περιήγησή του;

Μπορούμε να αναμένουμε από αυτό το βασικό μοντέλο να λειτουργεί πάντα για ένα γενικό δίκτυο όπως το Διαδίκτυο; Όχι χωρίς να κάνουμε κάποιες προσαρμογές. Στο Σχήμα 8.3 μπορείτε να δείτε δύο πιθανές δυσκολίες. Στο αριστερό διάγραμμα απεικονίζεται ένας αιωρούμενος κόμβος (dangling node). Μια τέτοια περίπτωση προκύπτει όταν δεν υπάρχουν εξερχόμενοι σύνδεσμοι από μια συγκεκριμένη ιστοσελίδα. Σημειώστε ότι αυτό δεν συνεπάγεται οτιδήποτε για την κατάταξη της ιστοσελίδα! Μπορεί όντως να πρόκειται για μια πολύ σημαντική ιστοσελίδα η οποία απλώς δεν έχει συνδέσμους: σκεφτείτε, για παράδειγμα, μια ιστοσελίδα που περιέχει το σύνταγμα μιας χώρας. Η δυσκολία την οποία προκαλεί ένας αιωρούμενος κόμβος είναι ότι η αντίστοιχη στήλη της  $A$  είναι μηδενική. Στο απλοϊκό παράδειγμά μας, φανταστείτε να μην είχε ο κόμβος 6 έναν σύνδεσμο προς τον κόμβο 3. Σε αυτή την περίπτωση, η 6η στήλη της μήτρας  $\theta$ α είχε μηδενικά στοιχεία. Αυτό σημαίνει πρακτικά ότι μόλις φτάσουμε στον κόμβο 6, θα «κολλήσουμε» και δεν θα μπορούσαμε να συνεχίσουμε να ακολουθούμε συνδέσμους.



**Σχήμα 8.3:** Πράγματα που μπορεί να εξελιχθούν άσχημα στο βασικό μοντέλο: Στα αριστερά απεικονίζεται ένας αιωρούμενος κόμβος και στα δεξιά μια τερματική ισχυρή συνιστώσα που περιλαμβάνει μια κυκλική διαδρομή.

Ένας απλός τρόπος για να διορθωθεί αυτό στο μοντέλο PageRank είναι να αλλαχθούν οι τιμές όλων των στοιχείων της στήλης από 0 σε  $1/n$  (στο παράδειγμά μας,  $1/6$ ). Όσον αφορά την πιθανοτική ερμηνεία, αυτό ισοδυναμεί με το να υποθέσουμε ότι, αν οι χρήστες ακολουθήσουν συνδέσμους και φτάσουν σε μια ιστοσελίδα χωρίς εξερχόμενους συνδέσμους, μπορούν να μεταβούν σε οποιαδήποτε ιστοσελίδα με την ίδια πιθανότητα. Αυτό είναι γνωστό ως «διόρθωση αιωρούμενου κόμβου» (dangling node correction).

Μια άλλη πιθανή δυσκολία στο βασικό μοντέλο είναι το ενδεχόμενο να εισέλθουμε σε «αδέξιδο», όπως φαίνεται στο δεξιό τμήμα του Σχήματος 8.3. Στην ορολογία των γραφημάτων, αυτή είναι μια τερματική ισχυρή συνιστώσα (terminal strong component). Στο σχήμα, ο κόμβος 1 οδηγεί σε κυκλική διαδρομή· μόλις ο χρήστης φτάσει σε αυτόν, δεν υπάρχει τρόπος να φύγει. Η κατάσταση θα ήταν χειρότερη αν ο κόμβος 1 δεν υπήρχε καν. Είναι φυσικό ο Ιστός να περιέχει πολυνάριθμους «κλειστούς βρόχους» που είναι απομονωμένοι από τον εξωτερικό κόσμο. Υπάρχουν πολλά παραδείγματα στα οποία μπορεί να προκύψει αυτή η γενική περίπτωση (χωρίς να περιέχει υποχρεωτικά κάποια κυκλική διαδρομή): σκεφτείτε, για παράδειγμα, την ογκώδη τεκμηρίωση μιας γλώσσας προγραμματισμού όπως η Java. Θα υπάρχουν ιστοσελίδες που δείχνουν η μία προς την άλλη. Είναι επίσης πιθανό να υπάρχουν εξωτερικοί σύνδεσμοι προς αυτή τη συνιστώσα του γραφήματος. Μπορεί, όμως, να υπάρχουν μόνο εσωτερικοί σύνδεσμοι, και η τεκμηρίωση να μην έχει εξερχόμενους συνδέσμους προς τον έξω κόσμο.

Στη μήτρα  $A$ , αυτό μεταφράζεται σε ένα διαγώνιο μπλοκ που «συνδέεται» σπάνια, ή και καθόλου, με οποιεσδήποτε γραμμές ή στήλες της μήτρας που δεν ανήκουν στο συγκεκριμένο μπλοκ. Πρώτον, κάτι τέτοιο μπορεί να καταστρέψει τη μοναδικότητα του διανύσματος PageRank. Για να διορθωθεί αυτό στο μοντέλο PageRank,

κατασκευάζεται ένας κυρτός συνδυασμός (convex combination) της  $A$  (μετά από τη διόρθωση του αιωρούμενου κόμβου) με μια μήτρα τάξης 1. Μια δημοφιλής στρατηγική είναι η αντικατάσταση της  $A$  με

$$\alpha A + (1 - \alpha)\mathbf{u}\mathbf{e}^T$$

όπου  $\alpha$  είναι ένας παράγοντας άμβλυνσης (damping factor),  $\mathbf{e}$  είναι ένα διάνυσμα με στοιχεία που έχουν όλα την τιμή 1, και  $\mathbf{u}$  είναι ένα διάνυσμα εξατομίκευσης (personalization vector). Στη συνέχεια υπολογίζεται το κυρίαρχο ιδιοδιάνυσμα αυτής της νέας μήτρας. Για παράδειγμα, αν ισχύει  $\alpha = 0.85$  και  $\mathbf{u} = \frac{1}{n}\mathbf{e}$ , η διόρθωση μπορεί να ερμηνευθεί ως ένα μοντέλο στο οποίο ένας χρήστης ακολουθεί συνδέσμους με πιθανότητα 0.85, αλλά υπάρχει και μια πιθανότητα 0.15 να μεταβεί με τυχαίο τρόπο σε οποιοδήποτε σημείο του Ιστού. Αν το διάνυσμα  $\mathbf{u}$  περιέχει μηδενικά, τότε το τυχαίο άλμα γίνεται επιλεκτικά, μόνο σε εκείνα τα τμήματα του γραφήματος που αντιστοιχούν σε μη μηδενικά στοιχεία του  $\mathbf{u}$ . έτσι προκύπτει και ο όρος «εξατομίκευση».

Η μήτρα  $\alpha A + (1 - \alpha)\mathbf{u}\mathbf{e}^T$  έχει μερικές πολύ ενδιαφέρουσες φασματικές ιδιότητες. Συγκεκριμένα, ενώ η κυρίαρχη ιδιοτιμή της είναι το 1, οι υπόλοιπες ιδιοτιμές της (που είναι γενικά μιγαδικές) είναι φραγμένες ως προς το μέγεθος από το  $\alpha$ . Αυτό οδηγεί αμέσως στο συμπέρασμα ότι όσο μικρότερο είναι το  $\alpha$ , τόσο πιο γρήγορα συγκλίνει η μέθοδος δυνάμεων. Το πρόβλημα είναι ότι μια μικρή τιμή του  $\alpha$  σημαίνει πως δίνουμε λιγότερο βάρος στο γράφημα συνδέσμων και την ιδέα του να ακολουθούνται σύνδεσμοι, και περισσότερο βάρος στην επιλογή τυχαίων αλμάτων προς ιστοσελίδες χωρίς να ακολουθούνται σύνδεσμοι. Το αν αυτό είναι «καλό» ή όχι είναι καθαρά ζήτημα της μοντελοποίησης. Πράγματι, η επιλογή της «βέλτιστης» παραμέτρου άμβλυνσης, αν βέβαια υπάρχει, έχει αποτελέσει αντικείμενο πολλών συζητήσεων τα τελευταία χρόνια. ■

## Αξιολόγηση των περιορισμών της μεθόδου δυνάμεων

Ας αφιερώσουμε λίγο χρόνο να σκεφτούμε περαιτέρω κάποια ζητήματα που σχετίζονται με τη μέθοδο δυνάμεων. Έχουμε κάνει αρκετές υποθέσεις στην πορεία, κάποιες εκ των οποίων είναι περισσότερο περιοριστικές από τις άλλες.

Η απαίτηση να ισχύει  $\beta_1 \neq 0$  στην αρχική μας εικασία μοιάζει αυθαίρετη εκ πρώτης όψεως, αλλά στην πράξη δεν είναι πολύ περιοριστική. Παραδόξως, ακόμα και αν ισχύει  $\beta_1 = 0$ , τα σφάλματα στρογγυλοποίησης συνήθως σώζουν την κατάσταση! Ο λόγος είναι ότι ένα σφάλμα στρογγυλοποίησης στον υπολογισμό εισάγει συνήθως μια μικρή συνιστώσα στις κατευθύνσεις όλων των ιδιοδιανυσμάτων.

Δεν έχουμε εξετάσει ακόμα την περίπτωση στην οποία η μήτρα έχει αρκετές κυρίαρχες ιδιοτιμές ίδιου μεγέθους. Αυτό αποκλείει από την ανάλυση μερικά πολύ σημαντικά στιγμιότυπα, συμπεριλαμβανομένων και των περιπτώσεων να υπάρχει μια κυρίαρχη μιγαδική ιδιοτιμή (για την οποία η συζυγής της είναι επίσης ιδιοτιμή και έχει το ίδιο μέγεθος) ή κυρίαρχες ιδιοτιμές με αντίθετα πρόσημα. Αποκλείει ακόμα και την ταυτοτική μήτρα! (Ευτυχώς, γνωρίζουμε τις ιδιοτιμές και τα ιδιοδιανύσματα της ταυτοτικής μήτρας.)

Επιπλέον, έχουμε υποθέσει ότι όλα τα ιδιοδιανύσματα της μήτρας είναι γραμμικά ανεξάρτητα. Αυτό δεν ισχύει πάντα. Θυμηθείτε από την Ενότητα 4.1 ότι οι μήτρες των οποίων τα ιδιοδιανύσματα δεν καλύπτουν το  $\mathcal{R}^n$  χαρακτηρίζονται ελλιπείς. Εδώ παίζει ρόλο η έννοια της γεωμετρικής πολλαπλότητας των ιδιοτιμών. Η μέθοδος δυνάμεων μπορεί να εφαρμοστεί και για τέτοιες μήτρες, αλλά η σύγκλιση είναι επίπονα αργή, και σίγουρα πιο αργή από ότι αναμένεται με βάση τους λόγους των κυρίαρχων ιδιοτιμών. Υπάρχει μια σχετικά πλήρης θεωρία για τέτοιες περιπτώσεις αλλά, και πάλι, ανήκει σε ένα πιο εξειδικευμένο βιβλίο.

## Η αντίστροφη επανάληψη και η επανάληψη του πηλίκου Rayleigh

Η μέθοδος δυνάμεων συγκλίνει γραμμικά και η σταθερά του ασυμπτωτικού σφάλματός της είναι  $\left| \frac{\lambda_2}{\lambda_1} \right|$ . Αν η ιδιοτιμή  $\lambda_2$  είναι παραπλήσια στη  $\lambda_1$  (μια περίπτωση που προκύπτει συχνά σε εφαρμογές), η σύγκλιση ίσως είναι υπερβολικά αργή. Ας δούμε ένα παράδειγμα: Αν ισχύει  $\lambda_1 = 1$  και  $\lambda_2 = 0.99$ , τότε για  $k = 100$  έχουμε  $\left( \frac{\lambda_2}{\lambda_1} \right)^k = 0.99^{100} \approx 0.36$ . Αυτό σημαίνει ότι μετά από 100 επαναλήψεις δεν είμαστε καν κοντά στο να κερδίσουμε έστω και ένα δεκαδικό ψηφίο!

Η αντίστροφη επανάληψη (inverse iteration) παρακάμπτει αυτή τη δυσκολία με τη λεγόμενη τεχνική μετατόπισης και αντιστροφής, που οδηγεί σε σημαντικά πιο γρήγορη σύγκλιση, με το ανάλογο αντίτιμο βέβαια – την αναγκαστική επίλυση ενός γραμμικού συστήματος σε κάθε επανάληψη.

## Τεχνική μετατόπισης και αντιστροφής

Η ιδέα είναι η εξής. Αν οι ιδιοτιμές της μήτρας  $A$  είναι  $\lambda_j$ , τότε οι ιδιοτιμές της μήτρας  $A - \alpha I$  είναι  $\lambda_j - \alpha$  και οι ιδιοτιμές της μήτρας  $B = (A - \alpha I)^{-1}$  είναι

$$\mu_j = \frac{1}{\lambda_j - \alpha}$$

Επιπλέον, όσο πιο κοντά βρίσκεται το  $\alpha$  στη  $\lambda_1$ , τόσο πιο κυρίαρχη είναι η μεγαλύτερη ιδιοτιμή της  $B$ . Πράγματι, στο όριο, αν  $\alpha \rightarrow \lambda_1$ , η πρώτη ιδιοτιμή της  $B$  τείνει στο  $\infty$  ενώ οι άλλες ιδιοτιμές τείνουν σε πεπερασμένες τιμές, και συγκεκριμένα στο

$\frac{1}{\lambda_j - \lambda_1}$ . Ας υποθέσουμε ότι θέλουμε να εφαρμόσουμε τη μέθοδο δυνάμεων στη μήτρα  $B = (A - \alpha I)^{-1}$  και όχι στην  $A$ , και έστω ότι η  $\lambda_2$  είναι η ιδιοτιμή της  $A$  που βρίσκεται πιο κοντά στη  $\lambda_1$ . Τότε, μια τέτοια επανάληψη εξακολουθεί να συγκλίνει γραμμικά αλλά με βελτιωμένο ρυθμό:

$$\left| \frac{\mu_2}{\mu_1} \right| = \left| \frac{\frac{1}{\lambda_2 - \alpha}}{\frac{1}{\lambda_1 - \alpha}} \right| = \left| \frac{\lambda_1 - \alpha}{\lambda_2 - \alpha} \right|$$

Όσο πιο κοντά είναι αυτός ο αριθμός στο 0, τόσο πιο γρήγορη είναι η σύγκλιση. Πράγματι, αν το  $\alpha$  βρίσκεται πολύ κοντά στη  $\lambda_1$ , η σύγκλιση αναμένεται να είναι πολύ γρήγορη, και ενδεχομένως πολύ γρηγορότερη από τη σύγκλιση που εμφανίζει η εφαρμογή της μεθόδου δυνάμεων στη μήτρα  $A$ .

Ίσως έχετε παρατηρήσει ήδη ότι η ίδια τεχνική είναι εφικτή με τη χρήση μιας τιμής του  $\alpha$  που είναι παραπλήσια σε οποιαδήποτε ιδιοτιμή  $\lambda_j$ , και όχι υποχρεωτικά στην κυρίαρχη ιδιοτιμή  $\lambda_1$ . Αυτό ισχύει όντως, και η μέθοδος που περιγράφεται εδώ λειτουργεί γενικά για τον υπολογισμό οποιασδήποτε ιδιοτιμής, εφόσον γνωρίζουμε περίπου ποια είναι αυτή. Ειδικότερα, όταν υπάρχουν περισσότερες από μία κυρίαρχες απλές ιδιοτιμές, μπορούμε να υπερκεράσουμε τη δυσκολία της μεθόδου δυνάμεων εφαρμόζοντας μετατοπίσεις.

Υπάρχουν δύο ζητήματα που χρήζουν διευθέτησης εδώ. Το πρώτο αφορά την επιλογή της παραμέτρου  $\alpha$ . Πώς επιλέγουμε μια τιμή του  $\alpha$  η οποία (i) να είναι εύκολο να τη σκεφτούμε και (ii) να βρίσκεται επαρκώς κοντά στη  $\lambda_1$  ώστε να εγγυάται τη γρήγορη σύγκλιση; Ευτυχώς, υπάρχουν αποτελεσματικές και υπολογιστικά φθηνές εκτιμήσεις, ειδικά για την κυρίαρχη ιδιοτιμή, που μπορούν να χρησιμοποιηθούν. Για παράδειγμα, είναι γνωστό ότι για μια μήτρα  $A$  ισχύει  $\rho(A) \leq \|A\|$ , όπου  $\rho$  είναι η φασματική ακτίνα, και  $\rho(A) = \max_i |\lambda_i(A)|$ . Ορισμένες νόρμες μητρών, όπως η 1-νόρμα ή η  $\infty$ -νόρμα, είναι εύκολο να υπολογιστούν, και άρα η επιλογή  $\alpha = \|A\|_1$ , για παράδειγμα, μπορεί σε πολλές περιπτώσεις να είναι μια λογική επιλογή μετατόπισης για τον υπολογισμό της κυρίαρχης ιδιοτιμής της  $A$ .

Ένα άλλο ζήτημα, και μάλιστα σημαντικό, είναι το υπολογιστικό κόστος. Στη μέθοδο δυνάμεων, κάθε επανάληψη περιλαμβάνει ουσιαστικά ένα γινόμενο μήτρας-διανύσματος, ενώ η αντίστροφη επανάληψη απαιτεί την επίλυση ενός γραμμικού συστήματος. Αυτό φέρνει στο προσκήνιο τα ζητήματα των Κεφαλαίων 5 και 7, ειδικά τις επαναληπτικές μεθόδους για γραμμικά συστήματα. Τελικά η μέθοδός μας μπορεί να απαιτεί εκατοντάδες ή και χιλιάδες πολλαπλασιασμούς μήτρας-διανύσματος για μία επανάληψη της αντίστροφης επαναληπτικής διαδικασίας. Συνεπώς, η σύγκλιση της αντίστροφης επανάληψης πρέπει να είναι πολύ γρήγορη για να είναι αποτελεσματική. Για τεράστια προβλήματα όπως η αναζήτηση στο Διαδίκτυο, η χρήση αυ-

τής της μεθόδου δεν τίθεται καν ως επιλογή. Για μικρότερα προβλήματα, ή προβλήματα με ειδική δομή που επιτρέπουν τη χρήση γρήγορων άμεσων μεθόδων, η αντίστροφη επανάληψη είναι πιο ελκυστική. Σημειώστε ότι εφόσον το  $\alpha$  είναι σταθερό, θα είχε νόημα να παραγοντοποιηθεί η μετατόπισμένη και αντεστραμμένη μήτρα  $B$  μία φορά μόνο πριν ξεκινήσει η επανάληψη, οπότε το κόστος κατά τη διάρκεια της επανάληψης θα οφείλεται μόνο στις προς τα εμπρός/προς τα πίσω αντικαταστάσεις για τις επιλύσεις των τριγωνικών συστημάτων.

Ακολουθεί ο αλγόριθμος αντίστροφης επανάληψης.

#### Αλγόριθμος: Αντίστροφη επανάληψη.

Είσοδος: η μήτρα  $A$ , η αρχική εικασία  $v_0$  και η μετατόπιση  $\alpha$ .

for  $k = 1, 2, \dots$  μέχρι τον τερματισμό

$$\text{λύσε το } (A - \alpha I)\tilde{v} = v_{k-1}$$

$$v_k = \tilde{v}/\|\tilde{v}\|$$

$$\lambda^{(k)} = v_k^T A v_k$$

end

Ας επανέλθουμε για λίγο στην επιλογή του  $\alpha$ . Έχουμε ήδη τεκμηριώσει ότι το πηλίκο Rayleigh αποτελεί καλή προσέγγιση μιας ιδιοτιμής για ένα δοθέν διάνυσμα. Επομένως, μπορούμε να επιλέξουμε τη μετατόπιση  $\alpha$  δυναμικά, δηλαδή  $\alpha = \alpha_k$ , θέτοντάς την ίση με το πηλίκο Rayleigh. Με αυτή την επιλογή, η ταχύτητα σύγκλισης αυξάνεται καθώς πλησιάζουμε στη ζητούμενη ιδιοτιμή· άρα, η τάξη σύγκλισης είναι καλύτερη από γραμμική. Στην πραγματικότητα, στις περισσότερες περιπτώσεις είναι κυνηγική. Στη συγκεκριμένη περίπτωση μπορεί να αξίζει να πληρώσουμε το αντίτιμο της αναγκαστικής εκ νέου παραγοντοποίησης της μήτρας σε κάθε επανάληψη.

Ο αλγόριθμος επανάληψης με πηλίκο Rayleigh παρουσιάζεται παρακάτω.

#### Αλγόριθμος: Επανάληψη με πηλίκο Rayleigh.

Είσοδος: η μήτρα  $A$  και η κανονικοποιημένη αρχική εικασία  $v_0$ : επίσης,  $\lambda^{(0)} = v_0^T A v_0$ .

for  $k = 1, 2, \dots$  μέχρι τον τερματισμό

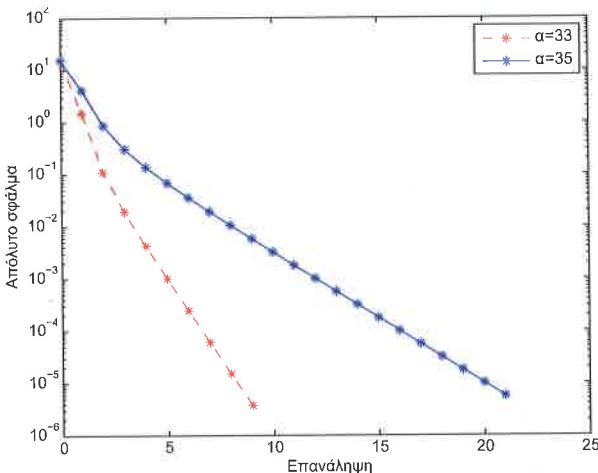
$$\text{λύσε το } (A - \lambda^{(k-1)} I)\tilde{v} = v_{k-1}$$

$$v_k = \tilde{v}/\|\tilde{v}\|$$

$$\lambda^{(k)} = v_k^T A v_k$$

end

**Παράδειγμα 8.4.** Για τη μήτρα  $A$  του Παραδείγματος 8.2 θα εκτελέσουμε την αντίστροφη επανάληψη με δύο σταθερές παραμέτρους:  $\alpha = 33$  και  $\alpha = 35$ . Και πάλι, υπολογίζουμε τα απόλυτα σφάλματα όπως κάναμε στο Παράδειγμα 8.2. Τα αποτελέσματα καταγράφονται στο Σχήμα 8.4. Παρατηρούμε ότι μια μετατόπιση πιο κοντά στην κυρίαρχη ιδιοτιμή (32) οδηγεί σε πολύ πιο γρήγορη σύγκλιση. Επίσης, οι τιμές του πλήθους επαναλήψεων είναι σημαντικά μικρότερες στο Σχήμα 8.4 από τις αντίστοιχες τιμές που απεικονίζονται στο Σχήμα 8.1 για τη μέθοδο δυνάμεων που είδαμε στο προηγούμενο παράδειγμα. s



**Σχήμα 8.4:** Σύγκλιση της αντίστροφης επανάληψης (Παράδειγμα 8.4).

Αν εκτελέσουμε την επανάληψη με πηλίκο Rayleigh, τα πράγματα επιταχύνονται ακόμα περισσότερο. Για το συγκεκριμένο παράδειγμα και μια τυχαία αρχική εικασία, μια τυπική ακολουθία απόλυτων σφαλμάτων που προέκυψε σε μία από τις τέσσερις εκτελέσεις μας ήταν  $3.71\text{e-}1, 9.46\text{e-}2, 2.34\text{e-}4, 2.16\text{e-}11$ . Παρατηρούμε ότι η σύγκλιση είναι εξαιρετικά γρήγορη και το πλήθος των ψηφίων περίπου τριπλασιάζεται σε κάθε επανάληψη. ■

Όταν χρησιμοποιούνται μετατοπίσεις, συνήθως απαιτούνται λιγότερες από 10 επαναλήψεις για να επιτευχθεί το ίδιο επίπεδο ακρίβειας που επιτυγχάνεται με εκατοντάδες (αν όχι περισσότερες) επαναλήψεις της μεθόδου δυνάμεων. Όπως είναι φυσικό, κάθε επανάληψη της αντίστροφης επαναληπτικής διαδικασίας συνήθως έχει σημαντικά μεγαλύτερο κόστος από ότι μια επανάληψη της μεθόδου δυνάμεων, ενώ οι επαναλήψεις με πηλίκο Rayleigh έχουν ακόμα μεγαλύτερο κόστος επειδή απαιτούν την εκ νέου παραγοντοποίηση της μήτρας σε κάθε επανάληψη. Δεν είναι δύσκολο να φανταστούμε αρκετά σενάρια στα οποία καθεμία από αυτές τις μεθόδους

είναι ανώτερη από τις υπόλοιπες, αλλά η πρόσθετη επιβάρυνση που προκύπτει στις περισσότερες περιπτώσεις ενδέχεται να αξίζει τον κόπο.

Ασκήσεις γι' αυτή την ενότητα: 1–6.

## 8.2 Διάσπαση ιδιαζουσών τιμών

Όταν μια τετραγωνική μήτρα  $A$  έχει πολύ μεγάλο δείκτη κατάστασης, δηλαδή όταν πλησιάζει στο να είναι ιδιαζουσα, διάφοροι υπολογισμοί μπορεί να μην έχουν αίσιο τέλος λόγω της μεγάλης μεγέθυνσης την οποία υφίστανται τα μικρά σφάλματα (όπως τα σφάλματα στρογγυλοποίησης). Πολλοί αλγόριθμοι που λειτουργούν σε άλλες περιπτώσεις γίνονται λιγότερο αξιόπιστοι, συμπεριλαμβανομένων και των αλγόριθμων για την εκτίμηση του ίδιου του δείκτη κατάστασης.

Αυτό ισχύει και για τα υπερκαθορισμένα συστήματα. Θυμηθείτε από την Εξισωση (6.1) στη σελίδα 256 ότι αν η  $A$  έχει διαστάσεις  $m \times n$  με  $\text{rank}(A) = n$ , ο δείκτης κατάστασης είναι  $\kappa(A) = \kappa_2(A) = \frac{\sigma_1}{\sigma_n}$ , δηλαδή ο λόγος της μεγαλύτερης προς τη μικρότερη ιδιαζουσα τιμή. Αν είναι πολύ μεγάλος, δηλαδή όταν οι στήλες της  $A$  είναι σχεδόν γραμμικά εξαρτημένες, η επίλυση του αντίστοιχου προβλήματος ελάχιστων τετραγώνων μπορεί να εμφανίσει μεγάλα σφάλματα.

Η μέθοδος SVD, την οποία περιγράψαμε στην Ενότητα 4.4, μπορεί να φανεί χρήσιμη εδώ. Παρακάτω θα δείξουμε τη χρήση της σε αρκετά δύσκολες ή ενδιαφέρουσες καταστάσεις. Θυμηθείτε επίσης το Παράδειγμα 4.18.

### Επίλυση σχεδόν ιδιαζόντων γραμμικών συστημάτων

Έστω ότι η  $A$  έχει διαστάσεις  $n \times n$  και πραγματικούς αριθμούς ως στοιχεία. Αν ο δείκτης κατάστασης  $\kappa(A)$  είναι πολύ μεγάλος, η επίλυση ενός γραμμικού συστήματος  $Ax = b$  μπορεί να είναι ένα πρόβλημα κακής κατάστασης.<sup>46</sup>

Φυσικά, με δεδομένη τη διάσπαση ιδιαζουσών τιμών της  $A$ , είναι εύκολο να βρούμε ότι  $x = V\Sigma^{-1}U^T b$ . Αυτό, όμως, δεν προσδίδει περισσότερο νόημα στην αριθμητική λύση που προκύπτει, όπως έχουμε επισημάνει στην Ενότητα 5.8. Ουσιαστικά, για ένα πρόβλημα κακής κατάστασης, παρότι οι μικρότερες ιδιαζουσες τιμές είναι θετικές, είναι και πολύ μικρές («σχεδόν μηδενικές» κατά κάποια έννοια). Αν το πρόβλημα είναι υπερβολικά κακής κατάστασης για να είναι χρήσιμο, συχνά εφαρμόζουμε μια διαδικασία που αποκαλείται **ομαλοποίηση** (regularization). Αυτό

<sup>46</sup> Ένα πρόβλημα  $Ax = b$  για μεγάλες τιμές του  $\kappa(A)$  δεν είναι υποχρεωτικά κακής κατάστασης με την έννοια που περιγράφεται εδώ. Πιο συγκεκριμένα, το πρόβλημα του Παραδείγματος 7.1 μπορεί να επιλυθεί με ασφάλεια για οποιαδήποτε μεγάλη τιμή του  $n$ .

σημαίνει ότι αντικαθιστούμε το δοθέν πρόβλημα, με έξυπνο τρόπο, με ένα παραπλήσιο πρόβλημα καλύτερης κατάστασης.

**Σημείωση:** Παρατηρήσετε πόσο έχουμε απομακρυνθεί εδώ από οτιδήποτε άλλο έχουμε κάνει μέχρι στιγμής. Αντί να ψάχνουμε αριθμητικούς αλγόριθμους για την επίλυση του εκάστοτε προβλήματος, αλλάζουμε πρώτα το πρόβλημα και λύνουμε ένα παραπλήσιο πρόβλημα με έναν τρόπο για τον οποίο ευελπιστούμε ότι θα έχει περισσότερο νόημα.

Με τη μέθοδο SVD μπορούμε να το πετύχουμε αυτό θέτοντας τις ιδιάζουσες τιμές που είναι μικρότερες από κάποια ανοχή απόρριψης ίσες με το 0 και ελαχιστοποιώντας την  $\ell_2$ -νόρμα της λύσης του υποκαθορισμένου προβλήματος που προκύπτει. Προχωράμε ως εξής:

1. Ξεκινώντας από το  $n$ , πηγαίνουμε προς τα πίσω μέχρι να βρεθεί ένα τέτοιο  $r$  ώστε το  $\frac{\sigma_1}{\sigma_r}$  να είναι ανεκτό σε μέγεθος. Αυτός είναι ο δείκτης κατάστασης του προβλήματος το οποίο όντως λύνουμε.
2. Υπολογίζουμε το  $\mathbf{z} = U^T \mathbf{b}$ : στην πραγματικότητα απαιτούνται μόνο τα πρώτα  $r$  στοιχεία του  $\mathbf{z}$ . Με άλλα λόγια, αν  $\mathbf{u}_i$  είναι το  $i$ -οστό διάνυσμα-στήλη της  $U$ , τότε  $z_i = \mathbf{u}_i^T \mathbf{b}, i = 1, \dots, r$ .
3. Υπολογίζουμε τα  $y_i = \sigma_i^{-1} z_i, i = 1, 2, \dots, r$ , και θέτουμε  $y_i = 0, i = r+1, \dots, n$ .
4. Υπολογίζουμε το  $\mathbf{x} = Vy$ . Αυτό πραγματικά περιλαμβάνει μόνο τις πρώτες  $r$  στήλες της  $V$  και τα πρώτα  $r$  στοιχεία του  $\mathbf{y}$ . Με άλλα λόγια, αν  $\mathbf{v}_i$  είναι το  $i$ -οστό διάνυσμα-στήλη της  $V$ , τότε  $\mathbf{x} = \sum_{i=1}^r y_i \mathbf{v}_i$ .

Βέβαια, στη γενική περίπτωση, το διάνυσμα  $\mathbf{x}$  που προκύπτει μπορεί να μην ικανοποιεί την ισότητα  $A\mathbf{x} = \mathbf{b}$  (αν και την ικανοποιεί στο Παράδειγμα 8.5 που ακολουθεί). Όμως αυτό είναι ό,τι καλύτερο μπορεί να πετύχει κάποιος υπό ορισμένες συνθήκες, και παράγει μια λύση  $\mathbf{x}$  της μικρότερης νόρμας για ένα προσεγγιστικό πρόβλημα επαρκώς καλής κατάστασης.

**Παράδειγμα 8.5.** Θυμηθείτε το Παράδειγμα 4.3 και έστω ότι

$$A = \begin{pmatrix} 1 & 1 \\ 3 & 3 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 2 \\ 6 \end{pmatrix}$$

Εδώ, η μήτρα  $A$  είναι ιδιάζουσα αλλά το  $\mathbf{b}$  εμπίπτει στον χώρο στηλών της. Επομένως, υπάρχουν πολλές λύσεις για τις εξισώσεις  $A\mathbf{x} = \mathbf{b}$ . Με απλή απαλοιφή Gauss προκύπτει διαίρεση με το 0.

Το MATLAB παράγει τη διάσπαση ιδιαίτερων τιμών

$$U = \begin{pmatrix} -0.316227766016838 & -0.948683298050514 \\ -0.948683298050514 & 0.316227766016838 \end{pmatrix}$$

$$V = \begin{pmatrix} -0.707106781186547 & 0.707106781186548 \\ -0.707106781186548 & -0.707106781186547 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 4.47213595499958 & 0 \\ 0 & 4.01876204512712e-16 \end{pmatrix}$$

Αποφασίζουμε (σοφά) ότι η μικρότερη ιδιάτερη τιμή είναι υπερβολικά μικρή και, άρα, η πραγματική τάξη αυτής της μήτρας είναι  $r = 1$ . Στη συνέχεια υπολογίζουμε τα εξής:

$$z_1 = -6.32455532033676 \rightarrow y_1 = -1.4142135623731 \rightarrow x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Θυμηθείτε ότι όλες οι λύσεις για το πρόβλημα αυτό έχουν τη μορφή  $\tilde{x} = (1 + \alpha, 1 - \alpha)^T$ . Επειδή ισχύει  $\|\tilde{x}\|_2^2 = (1 + \alpha)^2 + (1 - \alpha)^2 = 2(1 + \alpha^2)$ , η διαδικασία SVD που χρησιμοποιούμε έχει βρει με ευσταθή τρόπο τη λύση ελάχιστης  $\ell_2$ -νόρμας γι' αυτό το ιδιάζον πρόβλημα. ■

Σε αυτό το σημείο θέλουμε να επισημάνουμε ξανά μια σημαντική λεπτομέρεια: Η διαδικασία SVD που χρησιμοποιούμε λύνει ένα πρόβλημα διαφορετικό από το αρχικό  $Ax = b$  και, πιο συγκεκριμένα, ένα πρόβλημα που έχει ακριβώς μία λύση η οποία στη γενική περίπτωση ικανοποιεί τις δοθείσες εξισώσεις μόνο κατά προσέγγιση. Η διαδικασία που περιγράφηκε παραπάνω για την επίλυση συστημάτων εξαιρετικά κακής κατάστασης χρησιμοποιεί το θεώρημα της βέλτιστης προσέγγισης κατώτερης (ή μειωμένης) τάξης (best lower rank approximation) που παρουσιάζεται παρακάτω. Η προσέγγιση είναι γνωστή ως **αποκομμένη SVD** (truncated SVD).

## Συμπίεση μιας εικόνας

Το γεγονός ότι η βέλτιστη προσέγγιση κατώτερης τάξης μπορεί να υπολογιστεί τόσο άμεσα με τη μέθοδο SVD καθιστά εφικτή την επινόηση μιας μεθόδου συμπίεσης

**Θεώρημα:** Βέλτιστη προσέγγιση κατώτερης τάξης.

Η βέλτιστη, τάξης  $r$ , προσέγγιση  $A_r$ , μιας μήτρας  $A = U\Sigma V^T$ , με την έννοια ότι η νόρμα  $\|A - A_r\|_2 = \sigma_{r+1}$  είναι ελάχιστη, είναι η μήτρα

$$A_r = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

όπου  $\mathbf{u}_i$  και  $\mathbf{v}_i$  είναι τα  $i$ -οστά διανύσματα-στήλες των  $U$  και  $V$  αντίστοιχα.

(compression scheme): Αποθηκεύοντας τις πρώτες  $r$  στήλες των  $U$  και  $V$ , καθώς και τις πρώτες  $r$  ιδιάζουσες τιμές, παίρνουμε μια προσέγγιση της μήτρας  $A$  χρησιμοποιώντας μόνο  $r(m + n + 1)$  θέσεις αντί των αρχικών  $mn$ .

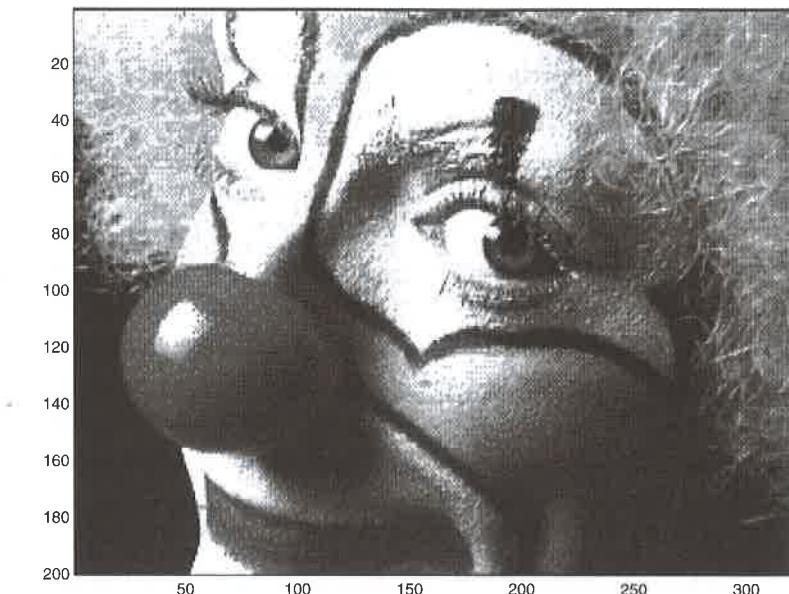
**Παράδειγμα 8.6.** Θεωρήστε τον ακόλουθο κώδικα MATLAB.

```
colormap('gray')
load clown.mat;
figure(1)
image(X);

[U,S,V] = svd(X);
figure(2)
r = 20;
colormap('gray')
image(U(:,1:r)*S(1:r,1:r)*V(:,1:r)');
```

Ο κώδικας φορτώνει την εικόνα ενός κλόουν από τη συλλογή εικόνων του MATLAB σε έναν πίνακα  $X$  διαστάσεων  $200 \times 320$ , εμφανίζει την εικόνα σε ένα σχήμα, βρίσκει τη διάσπαση ιδιαζουσών τιμών (SVD) της μήτρας  $A$ , και εμφανίζει σε ένα άλλο σχήμα την εικόνα που προκύπτει από μια προσέγγιση SVD τάξης 20 της  $A$ . Η αρχική εικόνα φαίνεται στο Σχήμα 8.5 και το συμπιεσμένο αποτέλεσμα στο Σχήμα 8.6.

Οι αρχικές απαιτήσεις σε αποθηκευτικό χώρο για την  $A$  είναι  $200 \cdot 320 = 64,000$ , ενώ η συμπιεσμένη αναπαράσταση απαιτεί  $20 \cdot (200 + 320 + 1) \approx 10,000$  θέσεις αποθήκευσης. Παρόλο που μπορεί να μην φαίνεται ιδιαίτερα εντυπωσιακό, το αποτέλεσμα είναι σίγουρα καλύτερο από κάποια τυχαία προσέγγιση (δείτε τα πρόσθετα σχόλια στο Παράδειγμα 4.18 για το ζήτημα αυτό). ■



**Σχήμα 8.5:** Η αρχική εικόνα ενός κλόουν με  $200 \times 320$  πίξελ.



**Σχήμα 8.6:** Μια προσέγγιση SVD τάξης 20 της εικόνας του κλόουν.

## Λανθάνουσα σημασιολογική ανάλυση

Ένα κεντρικό ζήτημα της **ανάκτησης πληροφοριών** (information retrieval) είναι ο πολύ γρήγορος εντοπισμός εγγράφων που είναι σχετικά με ένα ερώτημα του χρήστη. Η συνάφεια εδώ δεν ορίζεται υποχρεωτικά με την έννοια των μηχανών αναζήτησης, τις οποίες περιγράψαμε εκτενώς στην Ενότητα 8.1. Το ζητούμενο είναι αν ένα ερώτημα και ένα έγγραφο μοιράζονται κάποιο κοινό **θέμα** (theme). Η κατάσταση περιπλέκεται εν μέρει λόγω της μεγάλης κλίμακας των βάσεων δεδομένων και λόγω διάφορων γλωσσολογικών ζητημάτων όπως η ύπαρξη πολλών συνωνύμων. Οι μεγάλες βάσεις δεδομένων δεν είναι απλώς πιο δύσκολες στον χειρισμό, αλλά περιέχουν αναπόφευκτα και «θόρυβο» (noise) με τη μορφή άσχετων εγγράφων, σπάνια χρησιμοποιούμενων λέξεων κ.ο.κ. Στην πράξη είναι ανέφικτο να στηριχθεί κανείς σε ανθρώπους-εμπειρογνόμονες, όχι μόνο επειδή οι άνθρωποι δεν έχουν τον χρόνο ή την αντοχή να διαβάσουν εκατομμύρια έγγραφα, αλλά και επειδή θα μπορούσαν να υπάρχουν τεράστιες διαφορές στην ερμηνεία των ίδιων δεδομένων από διαφορετικά άτομα. Επομένως, απαιτούνται μαθηματικά μοντέλα. Αρκετά τέτοια μοντέλα βασίζονται στην προβολή (projection) των συνόλων δεδομένων και των ερωτημάτων, σε έναν μικρότερο χώρο στον οποίο τα παραπάνω ζητήματα μπορούν να αντιμετωπιστούν με καλύτερο τρόπο. Αυτή είναι, με λίγα λόγια, η ιδέα στην οποία στηρίζεται η **λανθάνουσα σημασιολογική ανάλυση** (latent semantic analysis), που είναι γνωστή και ως **λανθάνουσα σημασιολογική δεικτοδότηση** (latent semantic indexing).

Σημαντική για την αποθήκευση δεδομένων είναι η **μήτρα όρων-εγγράφων** (term-document matrix), μια μήτρα διαστάσεων  $n \times m$ , όπου  $n$  είναι το πλήθος των όρων (λέξεων) και  $m$  είναι το πλήθος των εγγράφων. Το στοιχείο  $i$  στη στήλη  $j$  της μήτρας είναι μια συνάρτηση της συχνότητας με την οποία η λέξη  $i$  εμφανίζεται στο έγγραφο  $j$ . Η απλούστερη συνάρτηση είναι μια απλή καταμέτρηση. Όμως υπάρχουν και πιο σύνθετα μέτρα, τα οποία αμβλύνουν τη σπουδαιότητα των πολύ συχνά χρησιμοποιούμενων όρων και δίνουν μεγαλύτερο βάρος (ή συντελεστή στάθμισης) στους σπάνια χρησιμοποιούμενους όρους. Αυτοί οι όροι συχνά είναι πιο αποτελεσματικοί για τη διάκριση ενός εγγράφου από ένα άλλο. Θα δείξουμε τι ακριβώς αναπαριστά αυτή η μήτρα παρουσιάζοντας ένα πολύ μικρό παράδειγμα.

**Παράδειγμα 8.7.** Θεωρήστε ένα απλό σύνολο δεδομένων το οποίο περιέχει τα ακόλουθα δύο έγγραφα μίας πρότασης (στα Αγγλικά):

«Numerical computations are fun».

«Numerical algorithms and numerical methods are interesting».

Έχουμε δύο έγγραφα και οκτώ διαφορετικές λέξεις. Σε μια ρεαλιστική εφαρμογή ανάκτησης πληροφοριών, λέξεις όπως το «are» και το «and» είναι υπερβολικά συχνές και γενικές για να παρέχουν οποιαδήποτε χρήσιμη πληροφορία, και γι' αυτό συνήθως εξαιρούνται. Έτσι, ορίζουμε και για τα δύο έγγραφα μια μήτρα όρων-εγγράφων με διαστάσεις  $6 \times 2$ , για τις έξι (αλφαριθμητικά διατεταγμένες) λέξεις «algorithms», «computations», «fun», «interesting», «methods» και «numerical».

Αν στη μήτρα αποθηκεύονται οι συχνότητες των λέξεων που χρησιμοποιούνται, η μήτρα όρων-εγγράφων γι' αυτή την περίπτωση είναι

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 2 \end{pmatrix}$$

Ορίζουμε επίσης διανύσματα μήκους 6 για την αναπαράσταση ερωτημάτων. Για παράδειγμα, το ερώτημα «numerical methods» αναπαρίσταται ως  $\mathbf{q} = (0, 0, 0, 0, 1, 1)^T$ .

Είναι πιθανό ένας λογικός άνθρωπος που θα δει τα δύο έγγραφα του παραδείγματός μας να συμφωνήσει ότι αφορούν ένα παρόμοιο ζήτημα. Και όμως, η επικάλυψη των δύο στηλών στη μήτρα  $A$  είναι σχετικά μικρή. Κάτι ανάλογο μπορεί να ισχύει και για ερωτήματα τα οποία χρησιμοποιούν παρόμοιες αλλά όχι τις ίδιες λέξεις. ■

Θα χρησιμοποιήσουμε ένα γενικό μοντέλο διανυσματικού χώρου (vector space model) για να εξακριβώσουμε αν ένα ερώτημα και ένα έγγραφο έχουν πολλά κοινά στοιχεία. Ένα δημοφιλές μέτρο σχετίζεται με εσωτερικά γινόμενα ή το συνημίτονο της γωνίας μεταξύ μιας διανυσματικής αναπαράστασης ενός ερωτήματος και των διανυσμάτων που αναπαριστούν τα έγγραφα. Έστω ότι η μήτρα όρων-εγγράφων  $A$  έχει διαστάσεις  $n \times m$ . Για ένα συγκεκριμένο ερώτημα  $\mathbf{q}$  θέτουμε

$$\cos(\theta_j) = \frac{(A\mathbf{e}_j)^T \mathbf{q}}{\|A\mathbf{e}_j\| \|\mathbf{q}\|}, \quad j = 1, \dots, m$$

όπου  $\mathbf{e}_j$  είναι η  $j$ -οστή στήλη της ταυτοτικής μήτρας διαστάσεων  $m \times m$ .

Όπως έχουμε ήδη αναφέρει, η λανθάνουσα σημασιολογική ανάλυση προβάλλει τα δεδομένα σε έναν μικρότερο χώρο ο οποίος είναι πιο εύκολος στον χειρισμό από υπολογιστική άποψη, και όπου τα παρόμοια έγγραφα και ερωτήματα μπορούν να

προσδιοριστούν με σχετικά εύκολο τρόπο. Ένας προφανής υποψήφιος για μια τέτοια συμπίεση πληροφοριών είναι η βέλτιστη προσέγγιση κατώτερης τάξης, που ορίζεται στη σελίδα 376 και προκύπτει από την αποκομμένη SVD. Αν ισχύει  $A_r = U_r \Sigma_r V_r^T$  και το  $r$  έχει μικρή τιμή, οι γωνίες που ορίστηκαν παραπάνω προσεγγίζονται πλέον σε αυτόν τον μειωμένο χώρο με τη βοήθεια της σχέσης

$$\cos(\theta_j) = \frac{\mathbf{e}_j^T V_r \Sigma_r (U_r^T \mathbf{q})}{\|\Sigma_r V_r^T \mathbf{e}_j\| \|\mathbf{q}\|}$$

Άρα, στον μειωμένο χώρο, το ερώτημα  $\mathbf{q}$  μήκους  $n$  μετασχηματίζεται στο διάνυσμα

$$\tilde{\mathbf{q}} = U_r^T \mathbf{q}$$

μήκους  $r$ . Μπορούμε πλέον να δουλέψουμε στον μειωμένο χώρο λόγω της αποκομμένης SVD.

Για να εκτιμήσουμε πλήρως το αποτέλεσμα αυτής της τεχνικής, θα έπρεπε να εξετάσουμε μια μεγάλη μήτρα όρων-εγγράφων, που θα περιλάμβανε πολύ περισσότερα έγγραφα από τα δύο του Παραδείγματος 8.7.

Σημειώστε ότι ενώ οι τεχνικές που χρησιμοποιούνται στα Παραδείγματα 8.6 και 8.7 παρουσιάζουν κάποια ομοιότητα, η φύση των προβλημάτων οδηγεί σε διαφορετικές υπολογιστικές προκλήσεις: Οι μήτρες όρων-εγγράφων είναι σχεδόν πάντα μεγάλες και αραιές, ενώ οι εικόνες, που θα μπορούσαν επίσης να είναι μεγάλες, συχνά είναι μικρότερες και συνήθως πυκνές. Κατά συνέπεια, και οι αλγόριθμοι για την εύρεση της αποκομμένης SVD θα ήταν επίσης διαφορετικοί, χρησιμοποιώντας τεχνικές παρόμοιας λογικής με αυτές της Ενότητας 7.5. Μια σχετική εντολή του MATLAB για την παρούσα εφαρμογή είναι η svds.

Υπάρχουν πολλά ζητήματα τα οποία δεν έχουμε επιλύσει σε αυτή τη σύντομη ανάλυση. Ένα από αυτά είναι η τιμή του  $r$ , συγκεκριμένα αν υπάρχει κάποιο  $r$  το οποίο να είναι και επαρκώς μικρό και αποτελεσματικό. Η απάντηση σε αυτή την ερώτηση συνήθως δεν μπορεί να δοθεί αναλυτικά και συχνά εξαρτάται από τα δεδομένα. Δείτε επίσης το Παράδειγμα 4.18 (σελίδα 159).

## Γραμμικά ελάχιστα τετράγωνα ελλιπούς τάξης (rank deficient linear least squares)

Η διάσπαση ιδιαζουσών τιμών μπορεί να εκληφθεί ως μια γενίκευση της φασματικής διάσπασης για μη τετραγωνικές μήτρες. Μια σημαντική κατηγορία προβλημάτων που οδηγούν σε μη τετραγωνικές μήτρες αφορά υπερκαθορισμένα συστήματα εξισώσεων.

Το πρόβλημα και οι συνήθεις μέθοδοι ελάχιστων τετραγώνων για την επίλυσή του ορίζονται και αναλύονται στο Κεφάλαιο 6. Όμως, όταν η μήτρα  $A$  διαστάσεων  $m \times n$  ( $n \leq m$ ) είναι ελλιπούς ή σχεδόν ελλιπούς τάξης στηλών, η παρακάτω εναλλακτική επιλογή αξίζει τον κόπο. Μάλιστα, περιλαμβάνει την περίπτωση της ομαλοποίησης τετραγωνικών συστημάτων την οποία θεωρήσαμε παραπάνω ως ειδική περίπτωση.

Εστω ότι θέλουμε να ελαχιστοποιήσουμε το

$$\|\mathbf{b} - A\mathbf{x}\| = \|\mathbf{b} - U\Sigma V^T \mathbf{x}\|$$

ως προς την  $\ell_2$ -νόρμα, όπου η μήτρα  $\Sigma$ , διαστάσεων  $m \times n$ , έχει μόνο  $r$  μη μηδενικές ιδιαζουσες τιμές  $\sigma_1, \dots, \sigma_r$  στην κύρια διαγώνιο της,  $r \leq n \leq m$ . Με παρόμοιο συλλογισμό όπως στην Ενότητα 6.2, μπορούμε να γράψουμε

$$\|\mathbf{b} - A\mathbf{x}\| = \|\mathbf{z} - \Sigma\mathbf{y}\|, \quad \mathbf{z} = U^T \mathbf{b}, \quad \mathbf{y} = V^T \mathbf{x}$$

Αν ισχύει  $r = n$ , δηλαδή αν η  $A$  είναι πλήρους τάξης στηλών, η μοναδική λύση είναι  $\mathbf{x} = Vy$ , όπου

$$y_i = \frac{z_i}{\sigma_i}, \quad i = 1, \dots, r$$

## Λύση ελάχιστης νόρμας

Αν ισχύει  $r < n$ , ακόμα και τότε το καλύτερο που μπορούμε να κάνουμε για να ελαχιστοποιήσουμε τη νόρμα  $\|\mathbf{z} - \Sigma\mathbf{y}\|$  είναι να ορίσουμε τα  $y_1, \dots, y_r$  όπως παραπάνω. Ωστόσο αυτό δεν είναι αρκετό για να οριστεί ένα μοναδικό  $\mathbf{x}$ . Στην πραγματικότητα, σε οποιαδήποτε επιλογή  $y_{r+1}, \dots, y_n$  που σχηματίζει το διάνυσμα  $\mathbf{y}$  μήκους  $n$  μαζί με τα πρώτα σταθερά  $r$  στοιχεία, αντιστοιχεί ένα  $\mathbf{x}$  το οποίο ελαχιστοποιεί τη νόρμα  $\|\mathbf{b} - A\mathbf{x}\|$ .

Στη συνέχεια επιλέγουμε, όπως και προηγουμένως, τη λύση  $\mathbf{x}$  με την ελάχιστη νόρμα (minimum norm). Επομένως, λύνουμε το πρόβλημα της ελαχιστοποίησης της  $\|\mathbf{x}\|$  για όλες τις λύσεις του δοθέντος γραμμικού προβλήματος ελάχιστων τετραγώνων.

Η (μοναδική πλέον!) λύση αυτού του διπλού προβλήματος βελτιστοποίησης είναι ευτυχώς εύκολη. Προφανώς, η  $\|\mathbf{y}\|$  ελαχιστοποιείται με την επιλογή

$$y_i = 0, \quad i = r + 1, r + 2, \dots, n$$

Τότε, όμως, το  $\mathbf{x} = Vy$  έχει επίσης ελάχιστη νόρμα επειδή η  $V$  είναι ορθογώνια.

Όσον αφορά τις στήλες  $\mathbf{u}_i$  της  $U$  και τις στήλες  $\mathbf{v}_i$  της  $V$ , μπορούμε σε συνεπυνγμένη μορφή να γράψουμε

$$\mathbf{x} = \sum_{i=1}^r \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i = A^\dagger \mathbf{b}$$

με την ψευδοαντίστροφη να ορίζεται ως  $A^\dagger = V \Sigma^\dagger U^T$ , όπου

$$\Sigma^\dagger = \begin{cases} 0, & \sigma_i = 0, \\ \frac{1}{\sigma_i}, & \sigma_i \neq 0 \end{cases}$$

Όπως ισχύει για όλα σχεδόν τα ιδιάζοντα γραμμικά συστήματα, εδώ η μήτρα  $A$  μπορεί να έχει σχεδόν γραμμικά εξαρτημένες στήλες, με την έννοια ότι ο δείκτης κατάστασης  $\kappa(A) = \sigma_1/\sigma_n$  είναι πεπερασμένος αλλά απαράδεκτα μεγάλος. Στη συνέχεια εφαρμόζεται μια διαδικασία απόρριψης παρόμοια με την περίπτωση  $m = n$  που περιγράφαμε νωρίτερα. Συνεπώς, η επίλυση του γραμμικού προβλήματος ελάχιστων τετραγώνων μέσω της διάσπασης ιδιαζουσών τιμών (SVD) περιλαμβάνει τα πέντε βήματα που καθορίζονται στον αλγόριθμο που παρουσιάζεται σε αυτή τη σελίδα.

#### Αλγόριθμος: Ελάχιστα τετράγωνα μέσω της SVD.

1. Σχημάτισε την  $A = U \Sigma V^T$ .

2. Αποφάσισε για την απόρριψη  $r$ .

3. Υπολόγισε το  $\mathbf{z} = U^T \mathbf{b}$ .

4. Θέσε  $y_i = \begin{cases} z_i/\sigma_i, & 1 \leq i \leq r, \\ 0, & i > r \end{cases}$

5. Υπολόγισε το  $\mathbf{x} = V \mathbf{y}$ .

**Παράδειγμα 8.8.** Θα πάρουμε τα δεδομένα του Παραδείγματος 6.8 για τη μήτρα  $A$  και το διάνυσμα  $\mathbf{b}$ , και θα προσθέσουμε εσκεμμένα μια ακόμα στήλη στην  $A$  η οποία είναι το άθροισμα των τριών υφιστάμενων στηλών της. Στο MATLAB αυτό επιτυγχάνεται με την εντολή  $B = [A, \text{sum}(A, 2)]$ . Έτσι σχηματίζεται μια μήτρα  $B$ , διαστάσεων  $5 \times 4$ , η οποία –εφόσον δεν υπάρχουν σφάλματα στρογγυλοποίησης– θα είναι τάξης 3. Ειδικότερα, με χρήση ακριβούς αριθμητικής,  $r < n < m$ . Κατόπιν θα ελέγξουμε πώς αποδίδουν οι μέθοδοι που βασίζονται σε ορθογώνιους μετασχηματισμούς.

Η εντολή  $x = A \setminus b$  παράγει για την καλής κατάστασης μήτρα  $A$  τη λύση που δίνεται στο Παράδειγμα 6.8, η οποία ικανοποιεί τις σχέσεις  $\|x\| \approx 0.9473$ ,  $\|b - Ax\| \approx 5.025$ .

Η εντολή  $x = B \setminus b$ , η οποία χρησιμοποιεί μια μέθοδο βασισμένη στην παραγοντοποίηση QR, παράγει ένα προειδοποιητικό μήνυμα σχετικά με την ελλιπή τάξη και μια λύση που ικανοποιεί τις σχέσεις  $\|x\| \approx 1.818$ ,  $\|b - Bx\| \approx 5.025$ . Προσέξτε την αύξηση της  $\|x\|$  λόγω της μεγέθυνσης που υφίστανται τα σφάλματα στρογγυλοποίησης κοντά σε ιδιάζουσα κατάσταση, αν και η νόρμα του βέλτιστου υπολοίπου παραμένει η ίδια σε ακριβή αριθμητική, και περίπου η ίδια και στον αντίστοιχο υπολογισμό.

Κατόπιν θα λύσουμε το πρόβλημα  $\min_x \{ \|x\| \text{ } \text{έτσι ώστε } x \text{ να ελαχιστοποιεί } \|Bx - b\| \}$  στο MATLAB, χρησιμοποιώντας την εντολή `svd` όπως περιγράφεται στον αλγόριθμο που μόλις παρουσιάσαμε. Αυτό μας δίνει

$$x \approx (0.3571, 0.4089, -0.7760, -0.9922 \times 10^{-2})^T, \quad \|x\| \approx 0.9471, \quad \|b - Bx\| \approx 5.025$$

Εδώ, η λύση  $x$  με ελάχιστη νόρμα παράγει επίσης την ίδια νόρμα υπολοίπου, και δεν φαίνεται να επηρεάζεται ιδιαίτερα από την κακή κατάσταση. Στην πραγματικότητα, η λύση αυτή μπορεί να θεωρηθεί ως διαταραχή μιας επαυξημένης εκδοχής της λύσης για το πρόβλημα του Παραδείγματος 6.8, που είχε καλή κατάσταση. ■

## Αποδοτικότητα του βασισμένου στην SVD αλγόριθμου ελάχιστων τετραγώνων

Μπορούμε να αξιοποιήσουμε το γεγονός ότι σε αυτή την εφαρμογή ισχύει πάντα η ανισότητα  $n \leq m$  για να επινοήσουμε μια *SVD* οικονομικού μεγέθους, όπως ακριβώς κάναμε για την παραγοντοποίηση QR στις Ενότητες 6.2 και 6.3, όπου η  $U$  έχει μόνο  $n$  στήλες και η  $\Sigma$  είναι τετραγωνική με διαστάσεις  $n \times n$ . Επιπλέον, σημειώστε ότι στον αλγόριθμο ελάχιστων τετραγώνων χρησιμοποιούνται μόνο οι πρώτες  $r$  στήλες των μητρών  $U$  και  $V$ . Όμως, γενικά δεν γνωρίζουμε την ακριβή τιμή του  $r$  εκ των προτέρων, παρά μόνο ότι ισχύει  $r \leq n$ , οπότε πρέπει να υπολογιστούν όλες οι  $n$  ιδιάζουσες τιμές.

Στο κόστος του αλγόριθμου κυριαρχεί η διάσπαση ιδιαζουσών τιμών, που αποδεικνύεται ότι απαιτεί περίπου  $2mn^2 + 11n^3$  flop. Για  $m \gg n$ , αυτό είναι κατά προσέγγιση ίδιο με το κόστος της μεθόδου που βασίζεται στην παραγοντοποίηση QR, όμως για  $m \approx n$  η μέθοδος SVD έχει σημαντικά μεγαλύτερο κόστος.

*Ασκήσεις γι' αυτή την ενότητα: 7–10.*

### 8.3 Γενικές μέθοδοι για τον υπολογισμό ιδιοτιμών και ιδιαζουσών τιμών

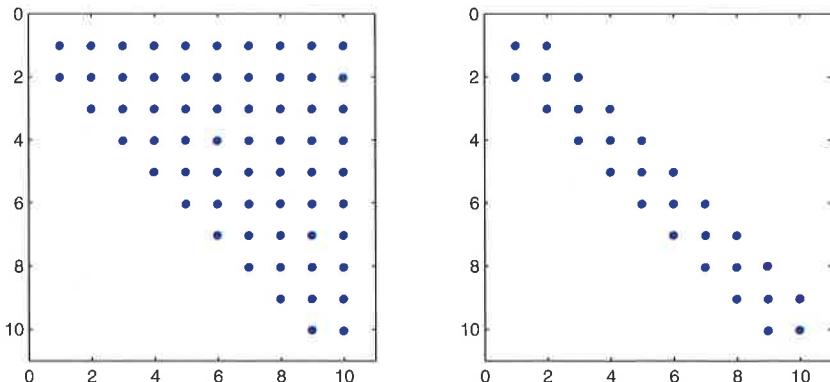
Στην Ενότητα 8.1 περιγράψαμε λεπτομερώς τον τρόπο υπολογισμού του κυρίαρχου ιδιοζεύγους μιας μήτρας, με τη χρήση της μεθόδου δυνάμεων. Η μέθοδος μετατόπισης και αντιστροφής, που οδηγεί στην αντίστροφη επανάληψη και στην επανάληψη με πηλίκο Rayleigh, επιτρέπει τον υπολογισμό –και μάλιστα πάρα πολύ γρήγορα– ενός ιδιοζεύγους το οποίο δεν είναι υποχρεωτικά κυρίαρχο, αλλά ενός μόνο ζεύγους μιλαταύτα. Μια λογική ερώτηση που προκύπτει είναι η εξής: Πώς μπορούμε να υπολογίσουμε πολλές ιδιοτιμές, ή όλες τις ιδιοτιμές, μιας μήτρας; Η απάντηση συνδέεται με μερικούς πολύ κομψούς αριθμητικούς αλγόριθμους. Θα παρέχουμε μια σύντομη περιγραφή σε αυτή την ενότητα και θα εξηγήσουμε πώς υπολογίζεται η διάσπαση ιδιαζουσών τιμών. Όμως θα περιγράψουμε μόνο αλγόριθμους για μήτρες που είναι δυνητικά πυκνές αλλά επαρκώς μικρές ώστε να χωρούν στη μνήμη.

#### Ορθογώνιος μετασχηματισμός ομοιότητας

Θυμηθείτε από την περιγραφή στη σελίδα 137 ότι αν  $S$  είναι μια μη ιδιάζουσα μήτρα που έχει το ίδιο μέγεθος με μια μήτρα  $A$ , τότε  $B = S^{-1}AS$  έχει τις ίδιες ιδιοτιμές με την  $A$ . Επιπλέον, αν το  $\mathbf{x}$  είναι ιδιοδιάνυσμα της  $A$ , τότε το  $S^{-1}\mathbf{x}$  είναι ιδιοδιάνυσμα της  $B$ . Αν  $\mathbf{Q} = S$  είναι ορθογώνια, δηλαδή αν ισχύει  $\mathbf{Q}^T\mathbf{Q} = I$ , τότε ισχύει  $B = S^{-1}AS = \mathbf{Q}^T\mathbf{A}\mathbf{Q}$ . Η χρήση ενός ορθογώνιου μετασχηματισμού ομοιότητας είναι υπολογιστικά ελκυστική επειδή η αντιστροφή σε αυτή την περίπτωση είναι τετριμμένη. Επιπλέον, οι ορθογώνιοι μετασχηματισμοί διατηρούν τη 2-νόρμα και, άρα, είναι λιγότερο επιρρεπείς στη συσσώρευση σφαλμάτων στρογγυλοποίησης συγκριτικά με όλες εναλλακτικές επιλογές.

Η βασική ιδέα των εξελιγμένων αλγόριθμων για τον υπολογισμό όλων των ιδιοτιμών ή ιδιαζουσών τιμών μιας μήτρας είναι η εκτέλεση του υπολογισμού σε δύο στάδια. Το πρώτο στάδιο, που περιλαμβάνει έναν σταθερό αριθμό βημάτων, στοχεύει στον ορθογώνιο μετασχηματισμό της μήτρας σε «απλή» μήτρα. Γενικά, αυτή η απλούστερη μήτρα θα είναι σε άνω Hessenberg μορφή (δείτε τη σελίδα 237). Για το συμμετρικό πρόβλημα των ιδιοτιμών, η μορφή αυτή ανάγεται σε τριδιαγώνια μήτρα (δείτε το Σχήμα 8.7). Στην περίπτωση της SVD, η διαδικασία είναι ελαφρώς διαφορετική: Θέλουμε να μετασχηματίσουμε τη μήτρα σε διδιαγώνια (bidiagonal), όπως στο Σχήμα 8.8.

Το δεύτερο στάδιο δεν μπορεί θεωρητικά να δώσει τις ακριβείς ιδιοτιμές σε πεπερασμένες πράξεις, με εξαίρεση κάποιες πολύ ειδικές περιπτώσεις. Συνήθως πε-



**Σχήμα 8.7:** Το αποτέλεσμα του πρώτου σταδίου ενός τυπικού ιδιοεπιλυτή (eigensolver): Γενική άνω Hessenberg μορφή για μη συμμετρικές μήτρες (αριστερά) και τριδιαγώνια μορφή για συμμετρικές μήτρες (δεξιά).

ριλαμβάνει μια ακολουθία ορθογώνιων μετασχηματισμών ομοιότητας με στόχο να έρθει η μετασχηματισμένη μήτρα όσο το δυνατόν πιο κοντά στην άνω τριγωνική μορφή. Τότε, οι κατά προσέγγιση ιδιοτιμές μπορούν να εξαχθούν από τη διαγώνιο. Αυτή η άνω τριγωνική μορφή είναι όντως διαγώνια αν η  $A$  είναι συμμετρική. Η διαδικασία για τις ιδιάζουσες τιμές είναι παρόμοια.

Ίσως αναρωτηθείτε για ποιον λόγο χωρίζεται ο υπολογισμός σε δύο στάδια, όταν το πρώτο είναι ακριβές (εφόσον δεν υπάρχουν σφάλματα στρογγυλοποίησης) και το δεύτερο είναι επαναληπτικό. Η απάντηση είναι ότι αν δεν μετασχηματιστεί η μήτρα στο πρώτο στάδιο σε άνω Hessenberg μορφή, οι επαναλήψεις στο δεύτερο στάδιο θα μπορούσαν να έχουν μη αποδεκτό κόστος. Αυτό θα αποσαφηνιστεί περαιτέρω στην περιγραφή που ακολουθεί.

## Ο αλγόριθμος QR για ιδιοτιμές

Έστω ότι η μήτρα  $A$  είναι πραγματική, τετράγωνη, αλλά όχι υποχρεωτικά συμμετρική. Θα υποθέσουμε για λόγους απλότητας ότι η  $A$  έχει μόνο πραγματικές ιδιοτιμές και ιδιοδιανύσματα. Αυτή η απλουστευτική υπόθεση δεν είναι αναγκαία για το πρώτο στάδιο του αλγόριθμου που περιγράφεται παρακάτω, αλλά απλοποιεί σημαντικά την περιγραφή του δεύτερου σταδίου. Θυμηθείτε από την Ενότητα 4.1 ότι, γενικά, οι μη συμμετρικές μήτρες έχουν μιγαδικές ιδιοτιμές και ιδιοδιανύσματα, οπότε η συνάρτηση  $\text{qreig}$  της οποίας τον κάθικα θα αναπτύξουμε παρακάτω δεν είναι πραγματικά τόσο γενική όσο η ενσωματωμένη συνάρτηση  $\text{eig}$  του MATLAB· ευελπιστούμε, ωστόσο, ότι η μορφή και η λειτουργία της θα είναι πιο ξεκάθαρη.

## Το πρώτο στάδιο

Θυμηθείτε από την Ενότητα 6.2 τις ανακλάσεις Householder. Τις χρησιμοποιήσαμε για τον σχηματισμό της παραγοντοποίησης QR, και μπορούμε να τις χρησιμοποιήσουμε το ίδιο επιτυχώς και εδώ για να αναπτύξουμε έναν ορθογώνιο μετασχηματισμό ομοιότητας της μορφής  $Q^T A Q$  (που θα έχει μορφή άνω Hessenberg). Η μαθηματική εξήγηση για το πώς μπορεί να γίνει αυτό είναι λίγο ανιαρή, οπότε θα δώσουμε ένα μικρό παράδειγμα, ακολουθούμενο από ένα γενικό πρόγραμμα.

**Παράδειγμα 8.9.** Θεωρήστε τη μήτρα διαστάσεων  $4 \times 4$

$$A = \begin{pmatrix} 0.5 & -0.1 & -0.5 & 0.4 \\ -0.1 & 0.3 & -0.2 & -0.3 \\ -0.3 & -0.2 & 0.6 & 0.3 \\ 0.1 & -0.3 & 0.3 & 1 \end{pmatrix}$$

Η αναγωγή σε άνω Hessenberg μορφή γίνεται σε δύο βήματα, για την πρώτη και τη δεύτερη γραμμή και στήλη. Θα συμβολίζουμε τις μήτρες των στοιχειώδων μετασχηματισμών γι' αυτά τα δύο βήματα με  $Q_1$  και  $Q_2$ . Πρώτον, για  $k = 1$ , εφαρμόζουμε μια ανάκλαση Householder η οποία μηδενίζει τα τελευταία δύο στοιχεία της πρώτης στήλης. Με άλλα λόγια, ψάχνουμε για έναν ανακλαστή  $\mathbf{u}_1$  τέτοιον ώστε το διάνυσμα  $(-0.1, -0.3, 0.1)^T$  να μετατραπεί σε ένα διάνυσμα της μορφής  $(\alpha, 0, 0)^T$  με την ίδια  $\ell_2$ -νόρμα. Ένα τέτοιο διάνυσμα, έως τέσσερα δεκαδικά ψηφία, είναι το  $\mathbf{u}_1 = (-0.8067, -0.5606, 0.1869)^T$ , και προκύπτει μια ορθογώνια μήτρα διαστάσεων  $3 \times 3$  της μορφής  $P^{(1)} = I_3 - 2\mathbf{u}_1\mathbf{u}_1^T$ . Στη συνέχεια ορίζουμε την

$$Q_1^T = \begin{pmatrix} 1 & & \\ & P^{(1)} & \end{pmatrix}$$

όπου η πρώτη γραμμή και η πρώτη στήλη έχουν μηδενικά στοιχεία παντού εκτός από τη θέση  $(1, 1)$ . Πλέον, η μήτρα  $Q_1^T A$  έχει δύο μηδενικά στις θέσεις  $(3, 1)$  και  $(4, 1)$ .

Η σημαντική λεπτομέρεια εδώ είναι ότι, όπως ο πολλαπλασιασμός με την  $Q_1^T$  από τα αριστερά δεν επηρεάζει την πρώτη γραμμή της  $A$ , έτσι και ο πολλαπλασιασμός με την  $Q_1$  από τα δεξιά δεν επηρεάζει την πρώτη στήλη της  $Q_1^T A$ . Σημειώστε ότι αν θέλαμε να βρούμε μια άνω τριγωνική μορφή, θα μπορούσαμε εύκολα να βρούμε μια μήτρα  $Q_1^T$  που να μηδενίζει τα τρία στοιχεία κάτω από το  $a_{1,1} = 0.5$  σε μια τέτοια περίπτωση, όμως, μόλις πολλαπλασιάζαμε με την  $Q_1$  από τα δεξιά, όλος ο κόπος μας θα πήγαινε στράφι επειδή θα επηρεάζονταν και οι τέσσερις στήλες.

Επομένως, ο πρώτος μετασχηματισμός ομοιότητας δίνει

$$Q_1^T A Q_1 = \begin{pmatrix} 0.5 & 0.6030 & -0.0114 & 0.2371 \\ 0.3317 & 0.3909 & -0.1203 & 0.0300 \\ 0 & -0.1203 & 0.6669 & 0.5255 \\ 0 & 0.03 & 0.5255 & 0.8422 \end{pmatrix}$$

και, εξ ορισμού, οι ιδιοτυπές της  $A$  διατηρούνται. Η διαδικασία ολοκληρώνεται με ένα παρόμοιο βήμα για  $k = 2$ . Αυτή τη φορά ο ανακλαστής είναι  $\mathbf{u}_2 = (-0.9925, 0.1221)^T$ , και η μήτρα  $Q_2^T$  είναι συνένωση της  $I_2$  με την ορθογώνια μήτρα, διαστάσεων  $2 \times 2$ ,  $P^{(2)} = I_2 - 2\mathbf{u}_2\mathbf{u}_2^T$ . Προκύπτει τελικά μια μήτρα σε άνω Hessenberg μορφή, όπως ακριβώς θέλαμε:

$$Q_2^T Q_1^T A Q_1 Q_2 = \begin{pmatrix} 0.5 & 0.6030 & 0.0685 & 0.2273 \\ 0.3317 & 0.3909 & 0.1240 & 0 \\ 0 & 0.1240 & 0.4301 & -0.4226 \\ 0 & 0 & -0.4226 & 1.0790 \end{pmatrix}$$

Οι πραγματικές επιπτώσεις αυτού του σταδίου γίνονται πλήρως αντιληπτές μόνο για μεγαλύτερες μήτρες, όπου σε ολόκληρο το αυστηρά κάτω αριστερό τρίγωνο μένουν μόνο  $n - 1$  μη μηδενικά στοιχεία από τα  $\frac{(n-1)n}{2}$  συνολικά. ■

Τα νούμερα στο Παράδειγμα 8.9 υπολογίστηκαν με τη χρήση της παρακάτω συνάρτησης. Προσέξτε ιδιαίτερα τις τελευταίες δύο γραμμές, όπου πολλαπλασιάζουμε την τρέχουσα μήτρα  $A$  με τον κατασκευασμένο (συμμετρικό) ορθογώνιο μετασχηματισμό από τα αριστερά και μετά από τα δεξιά.

```

function A = houseeig (A)
%
% function A = houseeig (A)
%
% Ανάγει την A σε άνω Hessenberg μορφή με τη χρήση ανακλάσεων
% Householder
n = size(A,1);
for k = 1:n-2

z=A(k+1:n,k);
e1=[1; zeros(n-k-1,1)];
u=z+sign(z(1))*norm(z)*e1;

```

```

u = u/norm(u);
% πολλαπλασιασμός από αριστερά και από δεξιά με Q = eye(n-k)
% -2*u*u';
A(k+1:n,k:n) = A(k+1:n,k:n) - 2*u*(u'*A(k+1:n,k:n));
A(1:n,k+1:n) = A(1:n,k+1:n) - 2*(A(1:n,k+1:n)*u)*u';
end

```

## Το δεύτερο στάδιο: Επανάληψη υπόχωρου

Σε αυτό το σημείο μπορούμε να υποθέσουμε ότι η μήτρα  $A$  είναι σε άνω Hessenberg μορφή. Το ζητούμενο είναι να προχωρήσουμε με επαναληπτικό τρόπο. Για τον σκοπό αυτόν θα επεκτείνουμε τη μέθοδο δυνάμεων. Έστω ότι έχουμε μια αρχική ορθογώνια μήτρα  $V_0$  μεγέθους  $n \times m$ ,  $1 \leq m \leq n$ , καθώς και την επανάληψη

for  $k = 1, 2, \dots$

$$\text{Θέσε } \widehat{V} = AV_{k-1}$$

$$\text{Υπολόγισε την παραγοντοποίηση QR της } \widehat{V}: \widehat{V} = V_k R_k$$

Επειδή οι στήλες της  $V_k$  ορίζουν τον ίδιο χώρο με τις στήλες της  $\widehat{V}$  για δεδομένο  $k$ , και επειδή ισχύει  $\widehat{V} = AV_{k-1}$ , μπορούμε να προχωρήσουμε με αναδρομικό τρόπο και να συμπεράνουμε ότι ο χώρος αυτός ορίζεται από την  $A^k V_0$ . Το βήμα ορθογωνιοποίησης (δηλαδή η παραγοντοποίηση QR) είναι καθοριστικής σημασίας: Χωρίς αυτό, θα βρίσκαμε απλώς πολλές προσεγγίσεις του κυρίαρχου ιδιοδιανύσματος (δείτε την Ενότητα 8.1) οι οποίες δεν μας είναι χρήσιμες.

Η παραπάνω επανάληψη είναι γνωστή με διάφορες ονομασίες: επανάληψη υπόχωρων (subspace iteration), ορθογώνια επανάληψη (orthogonal iteration) ή ταντόχρονη επανάληψη (simultaneous iteration). Προσέξτε ότι το πλήθος των στηλών της αρχικής εικασίας  $V_0$ , και άρα των επόμενων τιμών προσέγγισης  $V_k$ , εξακολουθεί να είναι ανοικτό ζήτημα. Για να βρούμε όλες τις ιδιοτιμές της  $A$ , θέτουμε  $m = n$ . Όπως μπορεί να αποδειχθεί, αν όλες οι ιδιοτιμές της  $A$  είναι διαφορετικές μεταξύ τους και αν όλες οι κύριες υπομήτρες της  $A$  έχουν πλήρη τάξη, τότε ξεκινώντας από το  $V_0 = I$  η επαναληπτική διαδικασία δίνει, καθώς το  $k \rightarrow \infty$ , μια άνω τριγωνική μήτρα της οποίας η διαγώνιος έχει ως στοιχεία τις ιδιοτιμές της  $A$ .

## Το δεύτερο στάδιο: Μη αποδοτικός αλγόριθμος ιδιοτιμών QR

Θα προχωρήσουμε υποθέτοντας ότι ισχύει  $m = n$ . Από την ισότητα  $AV_{k-1} = V_k R_k$  έπειται ότι  $V_k^T A V_{k-1} = R_k$ . Ξέρουμε επίσης ότι

$$V_{k-1}^T A V_{k-1} = V_k^T V_k R_k$$

Η τελευταία εξίσωση δεν είναι τίποτε άλλο παρά μια παραγοντοποίηση QR: Πράγματι,  $V_{k-1}^T A V_{k-1} = QR$ , με  $Q = V_{k-1}^T V_k$  και  $R = R_k$ . Άλλα αυτό συνεπάγεται ότι

$$V_k^T A V_k = (V_k^T A V_{k-1})(V_{k-1}^T V_k) = RQ$$

Μόλις ορίσαμε τον πασίγνωστο αλγόριθμο ιδιοτιμών QR στην απλούστερη μορφή του:

Θέσε  $A_0 = A$ .

for  $k = 0, 1, 2, \dots$  μέχρι τον τερματισμό

Παραγοντοποίησε την  $A_k = Q_k R_k$

Κατασκεύασε την  $A_{k+1} = R_k Q_k$

Η διαδικασία είναι εντυπωσιακά παλινδρομική. Παρατηρήστε ότι ο αλγόριθμος ιδιοτιμών QR δεν είναι το ίδιο πράγμα με τον αλγόριθμο παραγοντοποίησης QR· απλώς ο πρώτος χρησιμοποιεί τον δεύτερο. Είναι προφανές ότι οι  $A_{k+1}$  και  $A_k$  είναι ορθογώνια όμοιες (orthogonally similar), και επειδή μπορούμε να ερμηνεύσουμε αυτή τη διαδικασία σε σχέση με την προαναφερθείσα επανάληψη υπόχωρου, η  $A_k$  θα συγκλίνει τελικά σε μια άνω τριγωνική μήτρα με στοιχεία στη διαγώνιο που προσεγγίζουν τις ιδιοτιμές της  $A_0$ . Επειδή έχουμε να κάνουμε με προσεγγιστικές τιμές, για καθεμία από τις οποίες υπολογίζεται η παραγοντοποίηση QR, είναι σημαντικό οι υπολογισμοί αυτοί να έχουν όσο το δυνατόν χαμηλότερο κόστος. Ο υπολογισμός της παραγοντοποίησης QR μιας άνω Hessenberg μήτρας απαιτεί  $O(n^2)$  πράξεις συγκριτικά με τις  $O(n^3)$  πράξεις που απαιτούνται για την ίδια παραγοντοποίηση μιας πλήρους μήτρας. Επιπλέον, η κατασκευή του γινομένου μητρώου  $RQ$  διατηρεί τη διάταξη των μη μηδενικών στοιχείων! Στις Ασκήσεις 12 και 13 θα σας ζητηθεί να επαληθεύσετε τα παραπάνω.

**Παράδειγμα 8.10.** Σε συνέχεια του Παραδείγματος 8.9, θα εφαρμόσουμε τον αλγόριθμο ιδιοτιμών QR με τη μορφή που μόλις ορίσαμε. Οι ιδιοτιμές της αρχικής μήτρας, στρογγυλοποιημένες στο πλήθος των ψηφίων που φαίνεται εδώ, είναι  $-0.0107$ ,  $0.2061$ ,  $0.9189$  και  $1.2857$ , και η άνω Hessenberg μήτρα που υπολογίζεται στο τέλος του Παραδείγματος 8.9 τις διατηρεί.

Χρησιμοποιώντας την άνω Hessenberg μήτρα ως αρχική εικασία  $A_0$  και εκτελώντας τρεις επαναλήψεις του αλγόριθμου αυτού, βρίσκουμε ότι τα στοιχεία στην κύρια διαγώνιο της μήτρας παίρνουν τις τιμές  $(-0.0106, 0.2111, 0.9100, 1.2896)$ . Τα στοιχεία της υποδιαγωνίου μειώνονται σταδιακά σε  $(0.0613, 0.0545, -0.0001)^T$ .

Έτσι, στην τρίτη επανάληψη έχουμε μια χονδρική προσέγγιση των ιδιοτιμών, ενώ τα στοιχεία της υποδιαγωνίου είναι πολύ μικρότερα από τα αρχικά, χωρίς όμως να είναι πολύ μικρά. ■

## Το δεύτερο στάδιο: Αποδοτικός αλγόριθμος ιδιοτιμών QR

Τα αποτελέσματα του Παραδείγματος 8.10 δείχνουν ότι είναι εφικτό να βρούμε εκτιμήσεις των ιδιοτιμών από την κύρια διαγώνιο μετά από λίγες επαναλήψεις. Η σύγκλιση, όμως, η οποία σχετίζεται σε μεγάλο βαθμό με τον ρυθμό μείωσης των στοιχείων της υποδιαγωνίου στην άνω Hessenberg μήτρα, είναι αργή ακόμα και σε αυτή την περίπτωση, όπου η μήτρα είναι εξαιρετικά μικρή και οι ιδιοτιμές είναι σχετικά καλά διαχωρισμένες.

Πράγματι, ο αλγόριθμος QR στην μορφή που περιγράφηκε παραπάνω δεν είναι ιδιαίτερα αποδοτικός. Βρισκόμαστε και εδώ αντιμέτωποι με τα ίδια προβλήματα που παρατηρήθηκαν στη μέθοδο δυνάμεων, τα οποία αποτέλεσαν και το κίνητρο για να βρούμε μια καλύτερη εναλλακτική επιλογή, τη μέθοδο μετατόπισης και αντιστροφής. Ευτυχώς, οι μετατόπισεις μπορούν εύκολα να ενσωματωθούν στην επανάληψη QR και συχνά δίνουν εντυπωσιακά αποτελέσματα.

Αν  $a_k$  είναι μια μετατόπιση κοντά σε μια ιδιοτιμή, τότε μπορούμε να ορίσουμε τον αλγόριθμο QR με μετατόπισεις (QR algorithm with shifts). Και εδώ είναι σχετικά απλό να δείξουμε ότι οι  $A_k$  και  $A_{k+1}$  είναι ορθογώνια όμοιες. Ο αλγόριθμος παρουσιάζεται παρακάτω.

Ναι, αλλά πώς πρέπει να επιλέξουμε τις μετατόπισεις αυτές; Σε πρακτικές εφαρμογές δεν υπάρχει τρόπος να αποφύγουμε τη χρήση μετατόπισεων για την επιτάχυνση της σύγκλισης, και η επιλογή αυτών είναι στην πραγματικότητα μια τέχνη από μόνη της. Παρατηρήστε ότι μπορούμε να μεταβάλλουμε το  $a_k$  δυναμικά καθ' όλη τη διάρκεια της επανάληψης. Ευτυχώς, συχνά αρκεί να το θεωρήσουμε ως μια τιμή κατά μήκος της διαγωνίου. Αυτό ακριβώς θα κάνουμε στη συνέχεια. Μπορεί να αποδειχθεί ότι μια τέτοια επιλογή είναι παρόμοια με τον υπολογισμό του πηλίκου Rayleigh. Όντως, αυτό γίνεται στον επονομαζόμενο αλγόριθμο QR με ρητές μονές μετατόπισεις (QR algorithm with explicit single shifts).

Μια πιθανή υλοποίηση λειτουργεί ως εξής. Χρησιμοποιούμε το τελευταίο διαγώνιο στοιχείο της μήτρας ως μετατόπιση και εκτελούμε τον παραπάνω αλγόριθμο. Σε κάθε επανάληψη, ελέγχουμε για να διαπιστώσουμε αν όλα τα στοιχεία της τελευταίας γραμμής, εκτός από το διαγώνιο στοιχείο, είναι επαρκώς μικρά: Σε μια άνω Hessenberg μήτρα υπάρχει ένα μόνο τέτοιο στοιχείο για έλεγχο. Αν ισχύει κάτι τέ-

### Αλγόριθμος: Επανάληψη QR.

Μετασχημάτισε τη δοθείσα μήτρα  $A_0$  σε άνω Hessenberg μορφή.

```
for k = 1, 2, ... μέχρι τον τερματισμό
     $A_k - \alpha_k I = Q_k R_k$ 
     $A_{k+1} = R_k Q_k + \alpha_k I$ 
end
```

τοιο, τότε μπορούμε να δηλώσουμε ότι έχουμε συγκλίνει σε μία μόνο ιδιοτυπή, η οποία είναι ουσιαστικά το αντίστοιχο στοιχείο της διαγωνίου. Έπειτα μπορούμε να διαγράψουμε τη γραμμή και τη στήλη που αντιστοιχούν σε αυτό το στοιχείο της διαγωνίου και να ξεκινήσουμε από την αρχή, εκτελώντας την ίδια επανάληψη σε μια μήτρα της οποίας η διάσταση έχει μειωθεί κατά ένα. Ακολουθεί ο σχετικός κώδικας.

```
function [lambda,itn] = qreig (A,tol)
%
% function [lambda,itn] = qreig (A,Tol)
%
% Βρίσκει όλες τις πραγματικές ιδιοτιμές της A
% Επιστρέφει επίσης το πλήθος των επαναλήψεων
% στη μεταβλητή itn

% πρώτο στάδιο: μετασχηματισμός σε άνω Hessenberg μορφή
A = houseeig(A);

% δεύτερο στάδιο: βρόχος απομείωσης (deflation)
n = size(A,1); lambda = []; itn = [];
for j = n:-1:1
    % εύρεση της j - οστής ιδιοτιμής
    [lambda(j),itn(j),A] = qrshift (A(1:j,1:j),tol);
end

function [lam,iter,A] = qrshift (A,tol)
%
% function [lam,iter,A] = qrshift (A,tol)
%
% Βρίσκει μία ιδιοτιμή lam της A σε άνω Hessenberg μορφή
% Επιστρέφει επίσης το πλήθος των επαναλήψεων
```

% Επίσης, βελτιώνει την A για το μέλλον

```
m = size(A,1); lam = A(m,m); iter=0; I = eye(m);
if m == 1, return, end
while (iter < 100)
    % μέγιστο πλήθος επαναλήψεων
    if (abs(A(m,m-1)) < tol), return, end      % έλεγχος για σύγκλιση
    iter=iter+1;
    [Q,R]=qr(A-lam*I);                         % υπολογισμός της
                                                % παραγοντοποίησης QR
    A=R*Q+lam*I;                                % εύρεση του A για την
                                                % επόμενη επανάληψη
    lam = A(m,m);                                % επόμενη μετατόπιση
end
```

**Παράδειγμα 8.11.** Θα εκτελέσουμε τη συνάρτηση `qreig` για την απλοϊκή μήτρα διαστάσεων  $4 \times 4$  του Παραδείγματος 8.9 με μια «σφιχτή» τιμή ανοχής, `1e-12`. Έτσι προκύπτει το αποτέλεσμα

$$\begin{aligned} \text{lambda} &= [-0.010679, 0.20608, 0.91887, 1.2857] \\ \text{itn} &= [0, 3, 3, 5] \end{aligned}$$

Τώρα μάλιστα! Προσέξτε ότι οι ιδιοτιμές ανακτώνται με την αντίστροφη σειρά, οπότε η τιμή του πλήθους των επαναλήψεων μειώνεται καθώς προχωράει ο βρόχος απομείωσης (deflation). Επομένως, ο αλγόριθμος ανακτά την  $j$ -οστή ιδιοτιμή, αλλά ταυτόχρονα βελτιώνει και όλες τις υπόλοιπες  $j - 1$  ιδιοτιμές. ■

Μπορεί το πρόγραμμά μας να λύνει και πιο σοβαρά προβλήματα ιδιοτιμών; Στη συνέχεια θα εξετάσουμε μια πιο δύσκολη περίπτωση.

**Παράδειγμα 8.12.** Προβλήματα ιδιοτιμών προκύπτουν με φυσικό τρόπο και στις διαφορικές εξισώσεις. Όμως δεν χρειάζεται να κατανοήσετε πραγματικά τα ζητήματα των διαφορικών εξισώσεων που αναλύονται παρακάτω για να εκτιμήσετε δεόντως τα αποτελέσματα.

Θεωρήστε το πρόβλημα της εύρεσης των ιδιοτιμών  $\lambda$  και των αντίστοιχων ιδιοσυναρτήσεων (eigenfunctions)  $u(t)$  που ικανοποιούν τη διαφορική εξίσωση

$$u''(t) - u'(t) = \lambda u(t), \quad 0 < t < L$$

και τις συνοριακές συνθήκες  $u(0) = u(L) = 0$ . Σε αυτό το σημείο θα ήταν χρήσιμο να θυμηθούμε τον συμβολισμό που χρησιμοποιήσαμε στο Παράδειγμα 4.17 (σελίδα

159). Θεωρούμε το μήκος  $L$  του διαστήματος ως παράμετρο και αναζητάμε μη τετριμμένες λύσεις, το οποίο σημαίνει ότι για κάποια τιμή του  $t$  ισχύει  $u(t) \neq 0$ .

Όπως αποδεικνύεται, οι (πραγματικές και πολλές) ιδιοτιμές γι' αυτό το πρόβλημα δίνονται από τη σχέση

$$\lambda_j^{de} = -\frac{1}{4} - \left(\frac{j\pi}{L}\right)^2, \quad j = 1, 2, \dots$$

Για καθεμία από αυτές τις τιμές υπάρχει αντίστοιχη ιδιοσυνάρτηση η οποία, ακριβώς όπως το ιδιοδιάνυσμα στην αλγεβρική περίπτωση, είναι μη τετριμμένη λύση για ένα ιδιάζον γραμμικό σύστημα.

Στη συνέχεια, για να βρούμε μια αριθμητική προσέγγιση των πρώτων  $n$  τιμών  $\lambda_j^{de}$ , διακριτοποιούμε αυτό το πρόβλημα με έναν τρόπο που επεκτείνει απευθείας την απόδειξη του Παραδείγματος 4.17. Για μια επιλεγμένη μικρή τιμή  $h$ , θέλουμε να βρούμε μια ιδιοτιμή  $\lambda$  και ένα ιδιοδιάνυσμα  $\mathbf{u} = (u_1, u_2, \dots, u_{N-1})^T$  που να ικανοποιούν την εξίσωση

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} - \frac{u_{i+1} - u_{i-1}}{2h} = \lambda u_i, \quad i = 1, 2, \dots, N-1$$

θέτοντας  $u_0 = u_N = 0$ , για  $N = L/h$ . Γράφοντας τα παραπάνω ως το αλγεβρικό πρόβλημα ιδιοτιμών  $A\mathbf{u} = \lambda\mathbf{u}$ , παίρνουμε μια μη συμμετρική, δυνητικά μεγάλη τριδιαγώνια μήτρα  $A$  μεγέθους  $n = N - 1$ , για την οποία ευελπιστούμε ότι θα έχει πραγματικές ιδιοτιμές.

Έστω ότι εκτελούμε τη συνάρτηση `qreig` για το πρόβλημα αυτό με  $L = 10$  και  $h = 0.1$ , δηλαδή το μέγεθος της μήτρας είναι  $n = 99$ , με ανοχή 1.e-4. Τα αποτελέσματα είναι καθησυχαστικά, με έναν μέσο όρο 2.7 επαναλήψεων ανά ιδιοτιμή. Αν ταξινομήσουμε τις ιδιοτιμές κατά φθίνουσα σειρά, η μέγιστη απόλυτη διαφορά μεταξύ των πρώτων έξι ιδιοτιμών  $\lambda_j$  και των αντίστοιχων συνεχών ομολόγων τους,  $\lambda_j^{de}$ , είναι ίση με 0.015. Η απόκλιση αυτή οφείλεται στο σφάλμα διακριτοποίησης, το οποίο εξαρτάται με τη σειρά του από την τιμή του  $h$ , αλλά αυτό δεν μας απασχολεί εδώ: Δεν μεταβάλλεται ιδιαίτερα αν μειώσουμε την ανοχή στη συνάρτηση που έχουμε γράψει για τον υπολογισμό των ιδιοτιμών.

Κατόπιν θα προσπαθήσουμε να λύσουμε για  $L = 80$  με  $h = 0.1$ , δηλαδή το μέγεθος της μήτρας είναι  $n = 799$ . Δυστυχώς, ο αλγόριθμος δεν καταφέρνει να συγκλίνει.

Για να καταλάβουμε τι ακριβώς συμβαίνει, θα εκτελέσουμε τη συνάρτηση `eig` του MATLAB για τα ίδια προβλήματα. Για  $L = 10$ , τα αποτελέσματα είναι συγκρίσιμα με εκείνα της δικής μας συνάρτησης `qreig`. Για  $L = 80$ , όμως, προκύπτουν

μιγαδικές ιδιοτιμές, δηλαδή ιδιοτιμές με μη μηδενικό φανταστικό μέρος, παρότι οι  $\lambda_j^{de}$  παραμένουν πραγματικές.

Ο λόγος γι' αυτή την ξαφνική εμφάνιση μιγαδικών ιδιοτιμών έχει να κάνει με την κακή κατάσταση του προβλήματος, αλλά αυτό δεν μας ενδιαφέρει εδώ. Το δίδαγμα είναι ότι η χρήση της  $qreig$  ως γενικού εργαλείου διερεύνησης για μη συμμετρικές μήτρες μπορεί να αποδειχθεί επικίνδυνη, επειδή ίσως παραβιαστεί –πιθανώς χωρίς έγκαιρη προειδοποίηση– η υπόθεση που έχουμε κάνει ότι οι ιδιοτιμές και τα ιδιοδιανύσματα είναι πραγματικά. ■

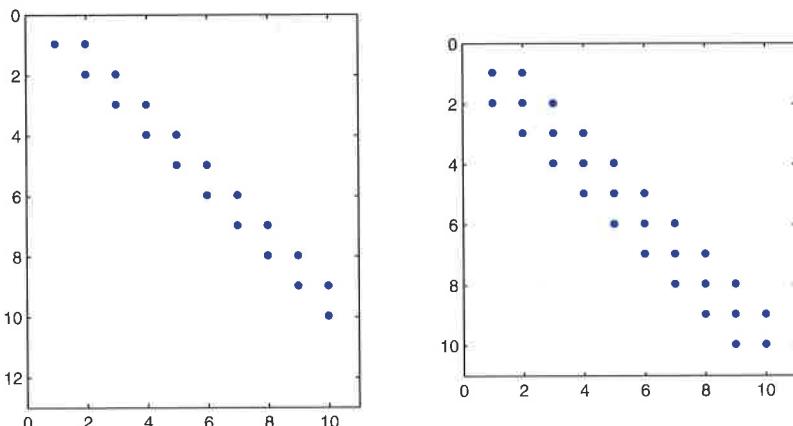
Όντως, αν μελετήσει κανείς τον αλγόριθμο QR που παρουσιάζεται στη σελίδα 391 και τον τρόπο που επιλέξαμε τις μετατοπίσεις, θα διαπιστώσει ότι δεν υπάρχει κάποιος μηχανισμός ο οποίος μπορεί να μας «μεταφέρει» από την ευθεία των πραγματικών αριθμών στο μιγαδικό επίπεδο (με άλλα λόγια, το σφάλμα βρίσκεται στην προδιαγραφή και όχι στην υλοποίηση του αντίστοιχου προγράμματος).

Οι σημαντικοί αλγόριθμοι της σύγχρονης βιβλιογραφίας στηρίζονται στις λεγόμενες έμμεσες μετατοπίσεις (implicit shifts). Γενικά, οι ιδιοτιμές μπορούν να είναι μιγαδικές για μη συμμετρικές μήτρες, και η επιθυμία μας να μην μπλέξουμε με μιδαγική αριθμητική μάς οδηγεί στην έννοια των διπλών μετατοπίσεων. Μια άλλη σημαντική λεπτομέρεια στους αλγόριθμους, που εγγυάται ότι οι παραγοντοποιήσεις QR καθ' όλη τη διάρκεια της επανάληψης QR υπολογίζονται σε  $O(n^2)$  πράξεις, είναι γνωστή ως το *θεώρημα της έμμεσης Q* (implicit Q theorem). Δυστυχώς, δεν έχουμε αρκετό χώρο εδώ για να εμβαθύνουμε περισσότερο.

Θα ολοκληρώσουμε την περιγραφή μας με μια σημείωση για τις συμμετρικές μήτρες. Όχι μόνο οι ιδιοτιμές και τα ιδιοδιανύσματά τους είναι πραγματικοί αριθμοί και πραγματικά διανύσματα αντίστοιχα (με την προϋπόθεση ότι και η μήτρα  $A$  είναι πραγματική), αλλά και η αντίστοιχη άνω Hessenberg μορφή των συμμετρικών μητρών είναι τριδιαγώνια συμμετρική μήτρα. Επιπλέον, οι υπολογισμοί έχουν σημαντικά χαμηλότερο κόστος και είναι πιο ευσταθείς από ό,τι στη γενική περίπτωση. Επομένως, αν ξέρετε πώς να λύνετε μη συμμετρικά προβλήματα ιδιοτιμών, η εφαρμογή των αλγόριθμων αυτής της ενότητας σε συμμετρικά προβλήματα θα είναι ανώδυνη. Σίγουρα, τα συμμετρικά προβλήματα ιδιοτιμών προσφέρονται για ακόμα πιο εξειδικευμένους αλγόριθμους και θεωρία, αλλά θα σας παραπέμψουμε για μια ακόμα φορά σε κάποιο πιο εξειδικευμένο βιβλίο.

## Υπολογισμός της διάσπασης ιδιαζουσών τιμών

Θα ολοκληρώσουμε τη μελέτη μας με μια σύντομη περιγραφή του υπολογισμού της διάσπασης ιδιαζουσών τιμών (SVD), ο οποίος εμφανίζει μια ευελιξία που δεν χαρακτηρίζει τους υπολογισμούς ιδιοτιμών. Στην περίπτωση των ιδιοτιμών, αναγκαστήκαμε να εφαρμόσουμε τον ίδιο ορθογώνιο μετασχηματισμό στα αριστερά και στα δεξιά (ανεστραμμένο), επειδή ήταν απαραίτητο να είναι μετασχηματισμός ομοιότητας που, όπως γνωρίζουμε, διατηρεί αμετάβλητες τις ιδιοτιμές σε ακριβή αριθμητική. Στην περίπτωση της SVD, ωστόσο, δεν υφίσταται η ανάγκη να εκτελεστούν οι ίδιες πράξεις, επειδή η μήτρα  $U$  στα αριστερά δεν είναι απαραίτητα ίδια με τη μήτρα  $V$ . Σημειώστε επίσης ότι οι ιδιοτιμές της συμμετρικής θετικά ημιορισμένης μήτρας  $A^T A$  είναι τα τετράγωνα των ιδιαζουσών τιμών της  $A$ , και για τις πρώτες έχουμε στη διάθεσή μας μια τεχνική για την αναγωγή της μήτρας σε τριδιαγώνια μορφή (δείτε το Σχήμα 8.8).



**Σχήμα 8.8:** Το αποτέλεσμα του πρώτου σταδίου της διάσπασης ιδιαζουσών τιμών (SVD) είναι μια διδιαγώνια μήτρα  $C$  (αριστερά). Η αντίστοιχη τριδιαγώνια μήτρα  $C^T C$  φαίνεται στα δεξιά.

Όλα τα παραπάνω μάς υποχρεώνουν να αναζητήσουμε μια διαδικασία η οποία θα μπορούσε να αναγάγει την  $A$  σε διδιαγώνια μορφή, χρησιμοποιώντας διαφορετικούς ορθογώνιους μετασχηματισμούς στα αριστερά και στα δεξιά. Λειτουργούμε με παρόμοιο τρόπο όπως στην αναγωγή σε μορφή Hessenberg για τον υπολογισμό ιδιοτιμών. Για να δείξουμε αυτή την ιδέα στην πράξη, έστω ότι η διάταξη μη μηδενικών στοιχείων μιας μήτρας  $A$  διαστάσεων  $5 \times 4$  είναι

$$A = \begin{pmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \end{pmatrix}$$

Πολλαπλασιάζοντας με τη  $U_1^T$  από τα αριστερά και χρησιμοποιώντας μετασχηματισμούς Householder, παίρνουμε

$$U_1^T A = \begin{pmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \end{pmatrix}$$

Μπορούμε πλέον να εφαρμόσουμε έναν διαφορετικό ορθογώνιο μετασχηματισμό στα δεξιά. Επειδή, όμως, δεν θέλουμε να πειράξουμε την πρώτη στήλη, συμβιβαζόμαστε με τον μηδενισμό των στοιχείων που βρίσκονται δεξιά από το στοιχείο (1, 2):

$$U_1^T A V_1 = \begin{pmatrix} \times & \times & 0 & 0 \\ 0 & \times & \times & \times \end{pmatrix}$$

Ένα οκόμα βήμα μάς δίνει

$$U_2^T U_1^T A V_1 V_2 = \begin{pmatrix} \times & \times & 0 & 0 \\ 0 & \times & \times & 0 \\ 0 & 0 & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & \times & \times \end{pmatrix}$$

και αυτό συνεχίζεται μέχρι να φτάσουμε σε μια διδιαγώνια μορφή. Όπως είναι φυσικό, το πλήθος των ορθογώνιων μετασχηματισμών στα αριστερά δεν είναι υποχρεωτικά ίσο με το αντίστοιχο πλήθος στα δεξιά.

Μόλις προκύψει μια διδιαιγώνια μορφή, μπορούμε να προχωρήσουμε με διάφορους τρόπους. Οι παραδοσιακές μέθοδοι στηρίζονται σε μια προσαρμογή του αλγόριθμου ιδιοτιμών QR. Οι πιο πρόσφατες και γρήγορες μέθοδοι χρησιμοποιούν μια τεχνική «διαιρεί και βασίλευε» (divide-and-conquer). Και πάλι, δεν θα εμβαθύνουμε περισσότερο.

*Ασκήσεις γι' αυτή την ενότητα: 11–16.*

## 8.4 Ασκήσεις

### 0. Ερωτήσεις επανάληψης

- (α) Τι είναι ένας ορθογώνιος μετασχηματισμός ομοιότητας;
- (β) Δείξτε ότι η τάξη σύγκλισης της μεθόδου δυνάμεων είναι γραμμική, και αναφέρετε ποια είναι η σταθερά του ασυμπτωτικού σφάλματος.
- (γ) Ποιες υποθέσεις πρέπει να γίνουν ώστε η σύγκλιση της μεθόδου δυνάμεων να είναι εγγυημένη;
- (δ) Τι είναι η μέθοδος μετατόπισης και αντιστροφής;
- (ε) Δείξτε ότι η τάξη σύγκλισης της αντίστροφης επανάληψης με σταθερή μετατόπιση είναι γραμμική, και αναφέρετε ποια είναι η σταθερά του ασυμπτωτικού σφάλματος.
- (στ) Ποια είναι η διαφορά κόστους μίας μόνο επανάληψης της μεθόδου δυνάμεων συγκριτικά με την αντίστροφη επανάληψη;
- (ζ) Τι είναι το πηλίκο Rayleigh και πώς μπορεί να χρησιμοποιηθεί για τον υπολογισμό ιδιοτιμών;
- (η) Τι είναι η διάσπαση ιδιαζουσών τιμών (SVD);
- (θ) Ορίστε τη διάσπαση ιδιαζουσών τιμών για μια ορθογώνια μήτρα που έχει λιγότερες γραμμές από ό,τι στήλες.
- (ι) Πότε πρέπει να χρησιμοποιείται η διάσπαση ιδιαζουσών τιμών για την επίλυση ενός προβλήματος ελάχιστων τετραγώνων; Πότε δεν πρέπει να χρησιμοποιείται;
- (ια) Πώς συνδέονται οι ιδιάζουσες τιμές με τη 2-νόρμα μιας μήτρας;
- (ιβ) Πώς συνδέονται οι ιδιάζουσες τιμές με τον φασματικό δείκτη κατάστασης,  $\kappa_2(A)$ ;

- (ιγ) Ποια είναι τα δύο κύρια στάδια των «ιδιοεπιλυτών» (eigensolvers), και ποιος είναι ο κύριος σκοπός του πρώτου σταδίου;
- (ιδ) Ποιο είναι το μοτίβο αραιότητας της μήτρας που προκύπτει από την αναγωγή σε άνω Hessenberg μορφή μιας συμμετρικής μήτρας; Γιατί;
- (ιε) Τι είναι η επανάληψη QR, και πώς συνδέεται με την παραγοντοποίηση QR;
- (ιστ) Γιατί είναι χρήσιμο να εισάγονται μετατοπίσεις στην επανάληψη QR;
- (ιζ) Τι είναι η διδιαγωνιοποίηση και πώς σχετίζεται με τη διάσπαση ιδιαζουσών τιμών; με τη διάσπαση ιδιαζουσών τιμών;
- Η μήτρα προβολής (ή προβολέας) είναι μια μήτρα  $P$  για την οποία ισχύει  $P^2 = P$ .
    - Βρείτε τις ιδιοτιμές ενός προβολέα.
    - Δείξτε ότι αν η μήτρα  $P$  είναι προβολέας, το ίδιο ισχύει και για τη μήτρα  $I - P$ .
  - Δείξτε ότι το πηλίκο Rayleigh μιας πραγματικής μήτρας  $A$  και ενός πραγματικού διανύσματος  $v$ ,  $\mu(v) = \frac{v^T A v}{v^T v}$ , είναι η λύση ελάχιστων τετραγώνων του προβλήματος

$$\min_{\mu} \|A v - \mu v\|$$

όπου το  $v$  δίνεται.

- Ο ακόλουθος κώδικας MATLAB

```
u = [1:32]; v = [1:30,30,32];
M = randn(32,32);
[Q,R] = qr(M);
A = Q*diag(u)*Q';
B = Q*diag(v)*Q';
```

παράγει δύο πλήρεις μήτρες  $A$  και  $B$  με μινιστηριώδη εμφάνιση.

Επαναλάβετε τους υπολογισμούς και τη συζήτηση του Παραδείγματος 8.2 γι' αυτές τις δύο μήτρες. Εξηγήστε τις παρατηρήσεις που μπορείτε να κάνετε.

- Χρησιμοποιήστε ως υπόδειγμα την Άσκηση 3 για να επαναλάβετε τους υπολογισμούς και τη συζήτηση του Παραδείγματος 8.4 για τη μήτρα  $A$  με μετατοπίσεις  $\alpha = 33$  και  $\alpha = 35$ . Εξηγήστε τις παρατηρήσεις που μπορείτε να κάνετε.

5. Έστω ότι

$$A = \begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_2 \end{pmatrix}$$

με  $\lambda_2 = \lambda_1$ .

Πόσο γρήγορα θα συγκλίνει η μέθοδος δυνάμεων στη μοναδική ιδιοτιμή και στο ιδιοδιάνυσμά της; Πώς διαφέρει αυτό από τις παρατηρήσεις που έγιναν στην ανάλυση που παρουσιάστηκε στην Ενότητα 8.1, και γιατί;

6. Σε μια στοχαστική ως προς τις στήλες μήτρα  $P$  τα στοιχεία είναι μη αρνητικά και όλα τα αθροίσματα των στηλών είναι ίσα με 1. Στην πράξη, αυτές οι μήτρες συχνά είναι μεγάλες και αραιές.

Έστω ότι  $E$  είναι μια μήτρα του ίδιου μεγέθους με την  $P$ , π.χ.  $n \times n$ , της οποίας όλα τα στοιχεία είναι ίσα με  $1/n$ , και έστω ότι  $\alpha$  είναι ένας αριθμός,  $0 < \alpha < 1$ .

- (α) Δείξτε ότι η  $A(\alpha) = \alpha P + (1 - \alpha)E$  είναι επίσης μια μήτρα στοχαστική ως προς τις στήλες.
- (β) Ποια είναι η μεγαλύτερη ιδιοτιμή της  $A(\alpha)$ ;
- (γ) Δείτε ότι η δεύτερη μεγαλύτερη ιδιοτιμή της  $A(\alpha)$  είναι φραγμένη (σε απόλυτη τιμή) από το  $\alpha$ .
- (δ) Έστω ότι το κυρίαρχο ιδιοδιάνυσμα της  $A(\alpha)$  πρέπει να υπολογιστεί με τη μέθοδο δυνάμεων. Αυτό το διάνυσμα, εφόσον κανονικοποιηθεί ώστε η  $\ell_1$ -νόρμα του να είναι ίση με 1, ονομάζεται «διάνυσμα στάσιμης κατανομής» (stationary distribution vector).
  - i. Δείξτε πώς μπορούν να υπολογιστούν με αποδοτικό τρόπο, ως προς τον αποθηκευτικό χώρο, γινόμενα μήτρας-διανύσματος με την  $P(\alpha)$ . (Μπορείτε να υποθέσετε ότι το  $n$  έχει πολύ μεγάλη τιμή, και θυμηθείτε ότι η  $E$  είναι πυκνή.)
  - ii. Δείξτε ότι αν εφαρμοστεί η μέθοδος δυνάμεων και αν η αρχική εικασία  $v_0$  ικανοποιεί την ισότητα  $\|v_0\|_1 = 1$ , τότε –και εφόσον δεν υπάρχουν σφάλματα στρογγυλοποίησης– όλα τα επόμενα διανύσματα προσέγγισης  $v_k$  έχουν επίσης μοναδιαία  $\ell_1$ -νόρμα.

[Προσοχή: Το (δ) και, ακόμα περισσότερο, το (γ) είναι αρκετά δυσκολότερα ερωτήματα από τα (α) και (β).]

7. Χρησιμοποιήστε τον ορισμό της ψευδοαντίστροφης μιας μήτρας  $A$  σε σχέση με τις ιδιάζουσες τιμές και τα ιδιοαδιανύσματά της, όπως δίνεται στην περιγραφή της επίλυσης γραμμικών προβλημάτων ελάχιστων τετραγώνων μέσω της διάσπασης ιδιαζουσών τιμών, για να δείξετε ότι ισχύουν οι ακόλουθες σχέσεις:
- $AA^\dagger A = A$
  - $A^\dagger AA^\dagger = A^\dagger$
  - $(AA^\dagger)^T = AA^\dagger$
  - $(A^\dagger A)^T = A^\dagger A$
8. Θεωρήστε το γραμμικό πρόβλημα ελάχιστων τετραγώνων της ελαχιστοποίησης της νόρμας  $\|\mathbf{b} - A\mathbf{x}\|_2$ , όπου  $A$  είναι μια μήτρα διαστάσεων  $m \times n$  ( $m > n$ ) και τάξης  $n$ .
- Χρησιμοποιήστε τη διάσπαση ιδιαζουσών τιμών για να δείξετε ότι η  $A^T A$  δεν είναι ιδιάζουσα.
  - Δίνεται μια μήτρα  $A$  διαστάσεων  $m \times n$  και πλήρους τάξης στηλών. Δείξτε ότι η μήτρα  $A(A^T A)^{-1} A^T$  είναι προβολέας και συμμετρική. Τέτοιοι τελεστές είναι γνωστοί ως *ορθογώνιοι προβολείς* (orthogonal projectors).
  - Δείξτε ότι η λύση του γραμμικού προβλήματος ελάχιστων τετραγώνων ικανοποιεί την εξίσωση
- $$\mathbf{r} = \mathbf{b} - A\mathbf{x} = P\mathbf{b}$$
- όπου  $P$  είναι ένας ορθογώνιος προβολέας. Εκφράστε τον προβολέα  $P$  σε σχέση με την  $A$ .
- Έστω ότι  $Q$  και  $R$  είναι οι μήτρες στην παραγοντοποίηση QR της μήτρας  $A$ . Εκφράστε τη μήτρα  $P$  σε σχέση με τις  $Q$  και  $R$ . Απλοποιήστε το αποτέλεσμά σας όσο το δυνατόν περισσότερο.
  - Με το  $\mathbf{r}$  να έχει οριστεί, όπως συνήθως, ως το υπόλοιπο, αντικαταστήστε το  $\mathbf{b}$  με το  $\hat{\mathbf{b}} = \mathbf{b} + \alpha \mathbf{r}$  για κάποιον αριθμό  $\alpha$ . Δείξτε ότι θα προκύψει η ίδια λύση ελάχιστων τετραγώνων για το πρόβλημα  $\min_{\mathbf{x}} \|A\mathbf{x} - \hat{\mathbf{b}}\|_2$  ανεξάρτητα από την τιμή του  $\alpha$ .
9. Θεωρήστε το πρόβλημα ελάχιστων τετραγώνων
- $$\min_{\mathbf{x}} \|\mathbf{b} - A\mathbf{x}\|_2$$

όπου ξέρουμε ότι η  $A$  είναι κακής κατάστασης. Θεωρήστε επίσης τη μέθοδο ομαλοποίησης, η οποία αντικαθιστά τις κανονικές εξισώσεις με το τροποποιημένο σύστημα καλύτερης κατάστασης

$$(A^T A + \gamma I) \mathbf{x}_\gamma = A^T \mathbf{b}$$

όπου  $\gamma > 0$  είναι μια παράμετρος.

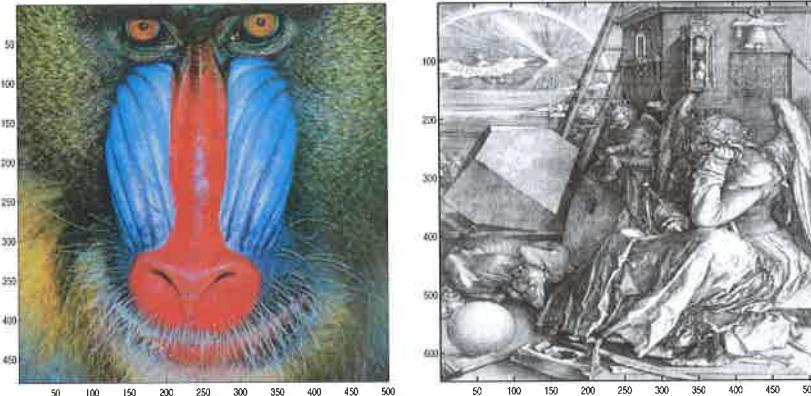
- (α) Δείξτε ότι  $\kappa_2^2(A) \geq \kappa_2(A^T A + \gamma I)$ .
- (β) Ξαναγράψτε τις εξισώσεις για το  $\mathbf{x}_\gamma$  ως γραμμικό πρόβλημα ελάχιστων τετραγώνων.
- (γ) Δείξτε ότι  $\|\mathbf{x}_\gamma\|_2 \leq \|\mathbf{x}\|_2$ .
- (δ) Βρείτε ένα φράγμα για το σχετικό σφάλμα  $\frac{\|\mathbf{x} - \mathbf{x}_\gamma\|_2}{\|\mathbf{x}\|_2}$  σε σχέση είτε με τη μεγαλύτερη είτε με τη μικρότερη ιδιάζουσα τιμή της μήτρας  $A$ .

Αναφέρετε μια ικανή συνθήκη για την τιμή του  $\gamma$  για την οποία το άνω φράγμα του σχετικού σφάλματος είναι μια δοθείσα τιμή  $\varepsilon$ .

- (ε) Γράψτε ένα σύντομο πρόγραμμα το οποίο να λύνει το πρόβλημα  $5 \times 4$  του Παραδείγματος 8.8, ομαλοποιημένου όπως παραπάνω, χρησιμοποιώντας την εντολή της ανάποδης καθέτου (backslash) του MATLAB. Δοκιμάστε τις τιμές  $\gamma = 10^{-j}$  για  $j = 0, 3, 6$  και  $12$ . Για κάθε  $\gamma$ , υπολογίστε την  $\ell_2$ -νόρμα του υπολοίπου,  $\|B\mathbf{x}_\gamma - \mathbf{b}\|$ , και τη λύση,  $\|\mathbf{x}_\gamma\|$ . Συγκρίνετε με τα αποτελέσματα για  $\gamma = 0$ , καθώς και με εκείνα που προκύπτουν από τη διάσπαση ιδιαζουσών τιμών στο Παράδειγμα 8.8. Ποια είναι τα συμπεράσματά σας;
  - (στ) Για μεγάλα προβλήματα ελάχιστων τετραγώνων κακής κατάστασης, ποιο είναι ένα δυνητικό πλεονέκτημα της ομαλοποίησης με το  $\gamma$ , η οποία παρουσιάζεται εδώ, συγκριτικά με τη λύση ελάχιστης νόρμας μέσω αποκομμένης SVD;
10. Σε αυτή την άσκηση θα πειραματιστείτε με δύο εικόνες (δείτε το Σχήμα 8.9) που μπορείτε να βρείτε στη συλλογή εικόνων του MATLAB και τις οποίες μπορείτε να ανακτήσετε πληκτρολογώντας τις εντολές `load mandrill` και `load durer`. Αφού φορτώσετε τα αρχεία, πληκτρολογήστε για καθένα από αυτά την εντολή `colormap(gray)` και μετά την εντολή `image(X)`. Όπως μπορείτε να δείτε, απεικονίζουν τον όμορφο μανδρίλο και έναν πίνακα του καλλιτέχνη Albrecht Dürer, ο οποίος έζησε τον 16ο αιώνα.

- (α) Γράψτε κώδικα στο MATLAB για τον υπολογισμό της αποκομμένης SVD αυτών των εικόνων. Και για τις δύο εικόνες, ξεκινήστε με τάξη (rank)  $r = 2$  και προχωρήστε ακολουθώντας τις δυνάμεις του 2, έως  $r = 64$ . Για μια συνεπτυγμένη παρουσίαση καθεμίας από τις εικόνες σας χρησιμοποιήστε την εντολή `subplot`, με το 3 και το 2 ως τα πρώτα δύο ορίσματα. (Για περισσότερες πληροφορίες πληκτρολογήστε `help subplot`.)
- (β) Σχολιάστε την απόδοση της αποκομμένης SVD για καθεμία από τις εικόνες. Αναφέρετε πόσος αποθηκευτικός χώρος απαιτείται συναρτήσει του  $r$  καθώς και πόσος αποθηκευτικός χώρος απαιτείται για τις αρχικές εικόνες. Εξηγήστε τη διαφορά στην αποτελεσματικότητα της τεχνικής για τις δύο εικόνες, για μικρές τιμές του  $r$ .

[Στην πραγματικότητα, η εικόνα του μανδρίου είναι έγχρωμη· όποτε θελήσετε, μπορείτε να τη δείτε σε όλη της τη μεγαλοπρέπεια αν αντί για `colormap(gray)` πληκτρολογήσετε `colormap(map)`. Όμως για τους σκοπούς των υπολογισμών σας πρέπει να χρησιμοποιήστε την κλίμακα του γκρίζου.]



**Σχήμα 8.9:** Ο μανδρίλος και ένας πίνακας του Albrecht Dürer (δείτε την Άσκηση 11).

11. Δείξτε ότι οι δύο μήτρες,  $A_k$  και  $A_{k+1}$ , σε διαδοχικές επαναλήψεις του αλγόριθμου ιδιοτιμών QR με μία ρητή μετατόπιση είναι ορθογώνια όμοιες.
12. Έστω ότι  $A$  είναι μια συμμετρική τριδιαγώνια τετραγωνική μήτρα διαστάσεων  $n \times n$ .
  - (α) Περιγράψτε τη διάταξη μη μηδενικών παραγόντων της παραγοντοποίησης QR της  $A$ .

- (β) Εξηγήστε πώς μπορούν να χρησιμοποιηθούν περιστροφές Givens στον υπολογισμό της παραγοντοποίησης QR της  $A$ , και δείξτε με συντομία ότι το πλήθος των πράξεων είναι πολύ μικρότερο από αυτό που θα απαιτούσε μια πλήρης μήτρα.
- (γ) Ποια είναι η διάταξη μη μηδενικών παραγόντων του γινομένου  $RQ$ , και πώς μπορεί να φανεί χρήσιμη στην εφαρμογή της επανάληψης QR για τον υπολογισμό ιδιοτιμών;

13. Επαναλάβετε την Άσκηση 12 για μια γενική άνω Hessenberg μήτρα  $A$ .
14. Θυμηθείτε από την Άσκηση 4.3 ότι μια πραγματική μήτρα  $A$  είναι λοξά συμμετρική αν ισχύει  $A^T = -A$ .

Γράψτε ένα πρόγραμμα για τον υπολογισμό των ιδιοτιμών και των ιδιοδιανυσμάτων μιας λοξά συμμετρικής μήτρας. Εκτελέστε τα εξής:

- (α) Αναγάγετε την  $A$  στην τριδιαγώνια μορφή  $A = QJQ^T$  χρησιμοποιώντας μετασχηματισμούς Householder. Δείξτε ότι όλα τα στοιχεία της διαγωνίου της ανηγμένης μήτρας  $J$  είναι μηδενικά.
  - (β) Γράψτε ένα πρόγραμμα που να υπολογίζει την επανάληψη QR για την τριδιαγώνια μήτρα  $J$ .
  - (γ) Εκτελέστε το πρόγραμμά σας με το λοξά συμμετρικό τμήμα του διακριτού τελεστή συναγωγής-διάχυσης που περιγράφεται στο Παράδειγμα 7.13.
15. Εφαρμόστε την επανάληψη QR με μετατοπίσεις στη μήτρα της Άσκησης 3. Εκτελέστε το πρόγραμμά σας με διάφορες παραμέτρους ανοχής και σχολιάστε τόσο την ταχύτητα σύγκλισης όσο και το συνολικό υπολογιστικό έργο που απαιτείται.
  16. Προτείνετε έναν αποδοτικό τρόπο υπολογισμού των ιδιοτιμών της μήτρας

$$M = \begin{pmatrix} A & C \\ B & D \end{pmatrix}$$

όπου  $A \in \mathbb{R}^{k \times k}$ ,  $B \in \mathbb{R}^{j \times k}$ ,  $C \in \mathbb{R}^{k \times j}$  και  $D \in \mathbb{R}^{j \times j}$  είναι πραγματικές, διαγώνιες μήτρες οι οποίες δίνονται.

[Παρατηρήστε ότι τα μεγέθη των μητρών που περιέχονται στην  $M$  είναι γενικά διαφορετικά, καθώς και ότι δεν είναι όλες οι μήτρες τετραγωνικές.]

## 8.5 Πρόσθετες σημειώσεις

Τα θέματα που καλύπτονται σε αυτό το κεφάλαιο περιγράφονται πολύ πιο αναλυτικά σε αρκετά εξειδικευμένα βιβλία της αριθμητικής γραμμικής άλγεβρας. Εδώ θα αναφέρουμε τα βιβλία του Demmel [21], των Golub και van Loan [30], και των Trefethen και Bau [70]. Ένα εγκυκλοπαιδικό βιβλίο το οποίο παρέχει μια διεξοδική θεώρηση έχει γράψει ο Stewart [63]. Ιδιαίτερα εύληπτο είναι και το βιβλίο του Watkins [74]. Ένα κλασικό σύγγραμμα το οποίο έχει παραμείνει στο προσκήνιο για περισσότερα από 40 χρόνια από τότε που εκδόθηκε είναι το βιβλίο του Wilkinson [75].

Υπάρχουν πολυάριθμες εφαρμογές όπου απαιτούνται μερικές ή όλες οι ιδιοτιμές ή ιδιάζουσες τιμές μιας μήτρας, και σε αυτό το κεφάλαιο παρουσιάσαμε δύο παραδείγματα εξόρυξης δεδομένων (data mining). Για περισσότερες πληροφορίες σχετικά με τον αλγόριθμο PageRank, δείτε το βιβλίο των Langville και Meyer [47],<sup>47</sup> καθώς και το βιβλίο των Berry και Browne [7] για μια περιγραφή διάφορων μεθοδολογιών ανάκτησης πληροφοριών, συμπεριλαμβανομένης της λανθάνουσας σημασιολογικής ανάλυσης.

Όπως έχουμε ήδη αναφέρει, παρότι όλες οι μέθοδοι υπολογισμού ιδιοτιμών είναι επαναληπτικές, χωρίζονται σε δύο κατηγορίες, από τις οποίες η μία θυμίζει περισσότερο τις άμεσες μεθόδους του Κεφαλαίου 5 και η άλλη τις επαναληπτικές μεθόδους του Κεφαλαίου 7. Οι μέθοδοι της πρώτης κατηγορίας βασίζονται σε παραγοντοποιήσεις και δεν λαμβάνουν ουσιαστικά υπόψη το μοτίβο αραιότητας της μήτρας. Ένα σχετικό παράδειγμα είναι η παραγοντοποίηση QR. Ο συγκεκριμένος αλγόριθμος, όπως και οι άμεσες μέθοδοι, βασίζεται σε (επανειλημμένες) παραγοντοποιήσεις και είναι αρκετά ανθεκτικός, αν και όχι πάντοτε αλάνθαστος. Παραδόξως, το άτομο που τον επινόησε, ο John Francis, εξαφανίστηκε από τον κόσμο της αριθμητικής ανάλυσης λίγο μετά τη δημοσίευση του πρωτοποριακού άρθρου του το 1961· μάλιστα, μόλις πριν από λίγα χρόνια έμαθε για τον τεράστιο αντίκτυπο που έχει προκαλέσει ο αλγόριθμός του.

Η δεύτερη κατηγορία μεθόδων βασίζονται κυρίως σε γινόμενα μήτρας-διανύσματος και, για τον λόγο αυτόν, λαμβάνουν υπόψη την αραιότητα. Συνήθως εφαρμόζονται σε μεγάλες και αραιές μήτρες. Αναζητούνται μόνο λίγες ιδιοτιμές και ιδιοδιανύσματα. Η μέθοδος δυνάμεων είναι μια βασική τέτοια μέθοδος. Η μέθοδος Lanczos και η μέθοδος Arnoldi, οι οποίες περιγράφηκαν στην Ενότητα 7.5, αποτελούν τα θεμέλια για τέτοιους ιδιοεπιλυτές.

<sup>47</sup> Σ.τ.Μ.: Έχει κυκλοφορήσει στην Ελλάδα με τον τίτλο *H μέθοδος PageRank της Google και άλλα συστήματα κατάταξης* (Πανεπιστημιακές Εκδόσεις Κρήτης, Ηράκλειο Κρήτης, 2010).

Στο MATLAB, ένας τρόπος να διακρίνουμε αυτές τις δύο κατηγορίες μεθόδων είναι κατανοώντας τη διαφορά μεταξύ των «άμεσων» συναρτήσεων `eig` και `svd` και των «επαναληπτικών» συναρτήσεων `eigs` και `svds`.

Υπάρχουν πολλά αξιόπιστα πακέτα λογισμικού για τον υπολογισμό ιδιοτιμών. Το αποθετήριο μαθηματικού λογισμικού `Netlib` περιέχει πολλές τέτοιες συναρτήσεις για τον υπολογισμό όλων των ιδιοτιμών μη τεράστιων μητρώων. Από αυτά τα πακέτα για τον υπολογισμό λίγων ιδιοτιμών για μεγάλες και αραιές μήτρες θα ξεχωρίσουμε ειδικά το κορυφαίο `ARPACK` [48].

## Κεφάλαιο 9

# Μη γραμμικά συστήματα και βελτιστοποίηση

Η βελτιστοποίηση (optimization) φαίνεται να αποτελεί μια σχεδόν αρχέγονη παρόρμηση των επιστημών και μηχανικών. Υπάρχει ένα τεράστιο πλήθος εφαρμογών στις οποίες εμφανίζονται τα μαθηματικά προβλήματα και οι αριθμητικές μέθοδοι που περιγράφονται σε αυτό το κεφάλαιο.

Οι τύποι των προβλημάτων βελτιστοποίησης είναι πολλοί. Το τυπικό πρόβλημα που θα χρησιμοποιήσουμε είναι η ελαχιστοποίηση μιας βαθμωτής συνάρτησης  $\phi$  ως προς  $n$  μεταβλητές  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ . Συμβολίζουμε την ελαχιστοποίηση με

$$\min_{\mathbf{x}} \phi(\mathbf{x})$$

και απαιτούμε το  $\mathbf{x}$  να ανήκει στο  $\mathcal{R}^n$  ή να είναι ένα υποσύνολο το οποίο διέπεται από έναν ή περισσότερους περιορισμούς.

Μια αναγκαία συνθήκη ώστε η συνάρτηση  $\phi(\mathbf{x})$  να έχει ένα ελάχιστο χωρίς περιορισμούς σε κάποιο συγκεκριμένο σημείο είναι ότι όλες οι πρώτες παραγωγοί της, δηλαδή το διάνυσμα κλίσης, πρέπει να μηδενίζονται στο σημείο αυτό. Αυτό αποτελεί γενίκευση της θεωρίας και των αποτελεσμάτων που περιγράψαμε στην Ενότητα 3.5 και αιτιολογείται περαιτέρω στην Ενότητα 9.2. Επομένως, πρέπει να ισχύει  $\frac{\partial \phi}{\partial x_1} = 0, \frac{\partial \phi}{\partial x_2} = 0, \dots, \frac{\partial \phi}{\partial x_n} = 0$ . Γενικά, υπάρχουν  $n$  μη γραμμικές εξισώσεις με  $n$  αγνώστους. Στην Ενότητα 9.1 θα μελετήσουμε το πρόβλημα της επίλυσης συστημάτων μη γραμμικών εξισώσεων. Μάλιστα, θα περιγράψουμε την επίλυσή τους σε ένα πιο γενικό πλαίσιο, χωρίς να πρέπει να λαμβάνουμε υπόψη κάποιο πρόβλημα ελαχιστοποίησης.

Στην Ενότητα 9.2 θα εξετάσουμε το πρόβλημα της άνευ περιορισμών ελαχιστοποίησης μιας βαθμωτής, επαρκώς λείας συνάρτησης ως προς πολλές μεταβλητές.

Θα αναπτύξουμε πολλές χρήσιμες μεθόδους, συμπεριλαμβανομένης και μιας ειδικής μεθόδου για προβλήματα μη γραμμικών ελάχιστων τετραγώνων, με την προϋπόθεση ότι μπορεί να υπολογιστεί η κλίση.

Τέλος, η Ενότητα 9.3 πραγματεύεται με συντομία ένα θέμα που απευθύνεται σε πιο προχωρημένους αναγνώστες, τη βελτιστοποίηση με περιορισμούς (constrained optimization), όπου η ελαχιστοποίηση μιας συνάρτησης υπόκειται σε περιορισμούς, άρα το  $\mathbf{x}$  περιορίζεται να ανήκει σε κάποιο σύνολο  $\Omega$  που περιέχεται αυστηρά στο  $\mathbb{R}^n$ . Αυτό συνήθως δυσχεραίνει την επίλυση του προβλήματος. Η σημαντική ειδική περίπτωση του γραμμικού προγραμματισμού (linear programming), όπου η αντικειμενική συνάρτηση και όλοι οι περιορισμοί είναι γραμμικής φύσης, θα εξεταστεί πιο λεπτομερώς από ότι ο γενικός μη γραμμικός προγραμματισμός με περιορισμούς (constrained nonlinear programming).

## 9.1 Η μέθοδος Newton για μη γραμμικά συστήματα

Θεωρήστε ένα σύστημα  $n$  μη γραμμικών εξισώσεων με  $n$  αγνώστους, γραμμένο στη μορφή

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}$$

όπου

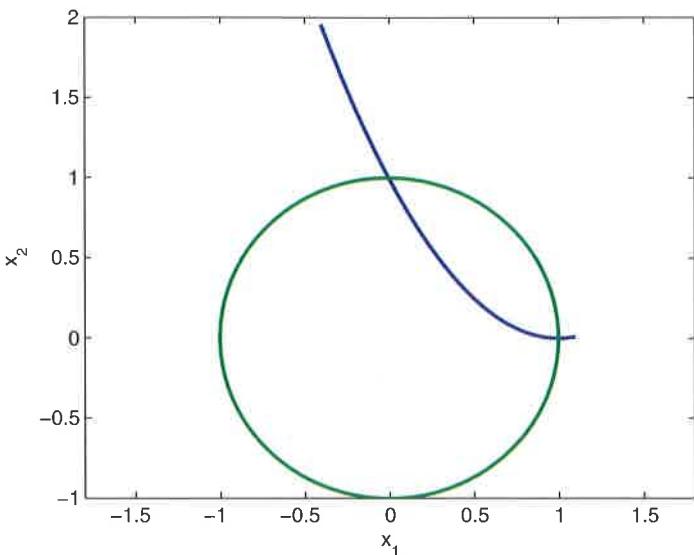
$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{pmatrix}$$

Αναλυτικά, το σύστημα μπορεί να γραφεί ως

$$f_i(x_1, x_2, \dots, x_n) = 0, \quad i = 1, 2, \dots, n$$

Αυτή είναι μια γενίκευση της περίπτωσης  $n = 1$ , την οποία εξετάσαμε στο Κεφάλαιο 3. Στην παρούσα ενότητα δεν θα υποθέσουμε ότι το πρόβλημα προκύπτει υποχρεωτικά από την ελαχιστοποίηση οποιασδήποτε βαθμωτής συνάρτησης. Ο σκοπός μας εδώ είναι να παρουσιάσουμε τη βασική μέθοδο Newton σε ένα πιο γενικό πλαίσιο.

**Παράδειγμα 9.1.** Έστω ότι  $f_1(x_1, x_2) = x_1^2 - 2x_1 - x_2 + 1$  και  $f_2(x_1, x_2) = x_1^2 + x_2^2 - 1$ . Η εξίσωση  $f_1(x_1, x_2) = 0$  περιγράφει μια παραβολή, ενώ η  $f_2(x_1, x_2) = 0$  περιγράφει έναν κύκλο (δείτε το Σχήμα 9.1).



**Σχήμα 9.1:** Μια παραβολή τέμνει έναν κύκλο.

Το σύστημα των εξισώσεων

$$\begin{aligned}x_1^2 - 2x_1 - x_2 + 1 &= 0 \\x_1^2 + x_2^2 - 1 &= 0\end{aligned}$$

έχει δύο ρίζες. Αυτές είναι τα δύο σημεία τομής,  $\mathbf{x}^{*(1)} = (1, 0)^T$  και  $\mathbf{x}^{*(2)} = (0, 1)^T$ . ■

## Η ευρύτερη εικόνα

Θα υποθέτουμε παντού ότι η  $\mathbf{f}$  έχει φραγμένες παραγώγους τουλάχιστον έως δεύτερης τάξης.

Κατ' αναλογία με την περίπτωση της μίας μη γραμμικής εξίσωσης που περιγράφαμε στο Κεφάλαιο 3, ένα σύστημα μη γραμμικών εξισώσεων μπορεί στη γενική περίπτωση να έχει οποιοδήποτε πλήθος λύσεων: πράγματι, το πλήθος των λύσεων δεν σχετίζεται άμεσα με το  $n$ . Επίσης, όπως και προηγουμένως, θα επιδιώξουμε να βρούμε μια επαναληπτική μέθοδο για την επίλυση μη γραμμικών εξισώσεων: Ξεκινώντας από μια αρχική εικασία  $\mathbf{x}_0$ , μια τέτοια μέθοδος παράγει μια ακολουθία τιμών  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k, \dots$ , για την οποία ευελπιστούμε ότι θα συγκλίνει σε μια λύση  $\mathbf{x}^*$  που ικανοποιεί την εξίσωση  $\mathbf{f}(\mathbf{x}^*) = \mathbf{0}$ .

Η επίλυση συστημάτων μη γραμμικών εξισώσεων μπορεί να αποδειχθεί πολύ πιο περίπλοκη από την επίλυση μίας μη γραμμικής εξίσωσης με έναν άγνωστο, αν και κάποιες από τις τεχνικές που περιγράφαμε στο Κεφάλαιο 3 μπορούν να ε-

πεκταθούν και στην οικογένεια των προβλημάτων με τα οποία ασχολούμαστε εδώ. Δυστυχώς, σε αυτή την περίπτωση δεν έχουμε στη διάθεσή μας κάποια μέθοδο όπως η διχοτόμηση (bisection). Σύντομα θα μελετήσουμε μια επέκταση της μεθόδου Newton και των παραλλαγών της, η οποία παρέχει ένα πολύ ισχυρό εργαλείο για τοπική σύγκλιση. Δηλαδή, αν ξεκινήσουμε με το  $x_0$  να βρίσκεται ήδη «κοντά» στο  $x^*$  κατά κάποια άποψη, τότε –υπό λογικές συνθήκες– η μέθοδος θα συγκλίνει πολύ γρήγορα. Ωστόσο η επίτευξη της σύγκλισης για περιπτώσεις όπου το σημείο εκκίνησης  $x_0$  είναι απομακρυσμένο μπορεί να αποδειχθεί ιδιαίτερα προβληματική, λόγω της έλλειψης ενός εργαλείου όπως η μέθοδος διχοτόμησης το οποίο θα μας επέτρεπε να πλησιάσουμε περισσότερο σε μια λύση με ασφαλή τρόπο.

### Θεώρημα: Σειρά Taylor για διανυσματικές συναρτήσεις.

Έστω ότι  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ ,  $\mathbf{f} = (f_1, f_2, \dots, f_m)^T$ , και έστω ότι η  $\mathbf{f}(\mathbf{x})$  έχει φραγμένες παραγώγους τουλάχιστον έως δεύτερης τάξης. Τότε, για ένα διάνυσμα κατεύθυνσης  $\mathbf{p} = (p_1, p_2, \dots, p_n)^T$ , το ανάπτυγμα Taylor για κάθε συνάρτηση  $f_i$  ως προς κάθε συντεταγμένη  $x_j$  δίνει

$$\mathbf{f}(\mathbf{x} + \mathbf{p}) = \mathbf{f}(\mathbf{x}) + J(\mathbf{x})\mathbf{p} + O(\|\mathbf{p}\|^2)$$

όπου  $J(\mathbf{x})$  είναι η **ιακωβιανή** (Jacobian) μήτρα των πρώτων παραγώγων της  $\mathbf{f}$  στο  $\mathbf{x}$ , που ορίζεται ως

$$J(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

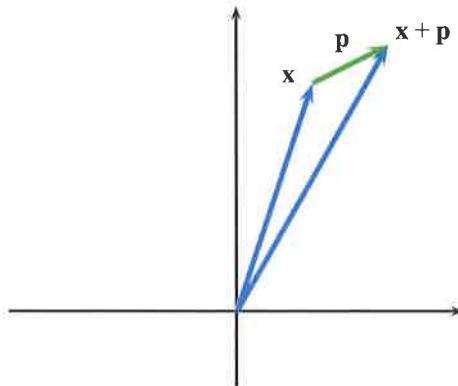
Άρα, έχουμε

$$f_i(\mathbf{x} + \mathbf{p}) = f_i(\mathbf{x}) + \sum_{j=1}^n \frac{\partial f_i}{\partial x_j} p_j + O(\|\mathbf{p}\|^2), \quad i = 1, \dots, m$$

Για να μπορέσουμε να ξεκινήσουμε, χρειαζόμαστε μια επέκταση του θεωρήματος της σειράς Taylor για αριθμούς (σελίδα 39) η οποία να μπορεί να εφαρμοστεί σε συστήματα. Το αντίστοιχο θεώρημα παρουσιάζεται σε αυτή τη σελίδα. Σημειώστε ότι για τους σκοπούς της παρούσας περιγραφής μας ισχύει  $n = m$ . Παρατηρήστε επίσης ότι, στην περίπτωση ενός συστήματος, δεν γίνεται υπολογισμός του επόμενου όρου (υπολοίπου) σε κάποιο ενδιάμεσο  $\xi$ . Επιπλέον, η ακριβής μορφή του όρου που

περιλαμβάνει τις δεύτερες παραγώγους της  $\mathbf{f}$  είναι ένα «τέρας» το οποίο, εντυχώς, δεν θα χρειαστεί να αντιμετωπίσουμε πολύ συχνά στις εφαρμογές μας.

Είναι χρήσιμο να φανταστούμε το  $\mathbf{x}$  ως ένα σημείο στο  $\mathcal{R}^n$  και το  $\mathbf{p} = (p_1, p_2, \dots, p_n)^T$  ως ένα διάνυσμα κατεύθυνσης. Υπό αυτή την έννοια, από ένα σημείο  $\mathbf{x}$  μετακινούμαστε στο σημείο  $\mathbf{x} + \mathbf{p}$  ακολουθώντας την κατεύθυνση  $\mathbf{p}$  (δείτε το Σχήμα 9.2).



**Σχήμα 9.2:** Το σημείο  $x$ , η κατεύθυνση  $p$  και το σημείο  $x + p$ .

## Η μέθοδος Newton

Ας επιστρέψουμε στο σύστημα μη γραμμικών εξισώσεων

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}$$

Η μέθοδος Newton για την επίλυση ενός τέτοιου συστήματος προκύπτει κατά τρόπο παρόμοιο με τη βαθμοτή περίπτωση. Ξεκινώντας από μια αρχική εικασία  $\mathbf{x}_0$ , παράγει μια ακολουθία τιμών  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k, \dots$  ως εξής: Υπολογίζει την επανάληψη  $\mathbf{x}_{k+1}$  προσεγγίζοντας το ανάπτυγμα Taylor της  $\mathbf{f}$  γύρω από το  $\mathbf{x}_k$  με τους γραμμικούς όρους του.

Επομένως, στο  $\mathbf{x} = \mathbf{x}_k$ , αν γνωρίζαμε ότι  $\mathbf{p} = \mathbf{p}^* = \mathbf{x}^* - \mathbf{x}_k$ , τότε για μικρές τιμές του  $\mathbf{p}^*$  θα ίσχυε

$$\mathbf{0} = \mathbf{f}(\mathbf{x}_k + \mathbf{p}^*) \approx \mathbf{f}(\mathbf{x}_k) + J(\mathbf{x}_k)\mathbf{p}^*$$

Οπως είναι φυσικό, δεν γνωρίζουμε το  $\mathbf{p}^*$  επειδή δεν γνωρίζουμε το  $\mathbf{x}^*$ . Μπορούμε όμως να χρησιμοποιήσουμε τα παραπάνω για να ορίσουμε ότι  $\mathbf{p} = \mathbf{p}_k$ , απαιτώντας να ισχύει

$$\mathbf{f}(\mathbf{x}_k) + J(\mathbf{x}_k)\mathbf{p}_k = \mathbf{0}$$

Έτσι προκύπτει ο αλγόριθμος που παρουσιάζεται παρακάτω,

Η παρακάτω συνάρτηση του MATLAB υλοποιεί τον αλγόριθμο.

### Αλγόριθμος: Μέθοδος Newton για συστήματα.

for  $k = 0, 1, \dots$  μέχρι τη σύγκλιση

Λύσε το σύστημα  $J(\mathbf{x}_k)\mathbf{p}_k = -\mathbf{f}(\mathbf{x}_k)$  ως προς  $\mathbf{p}_k$

Θέσε  $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$

end

```

function [x,k] = newtons(f,x,tol,nmax)
%
% function [x,k] = newtons(f,x,tol,nmax)
%
% Η συνάρτηση επιστρέφει στη μεταβλητή x ένα διάνυσμα - στήλη
% x_k τέτοιο ώστε να τσχύει || x_k - x_{k-1} || < tol (1 +
% ||x_k||) και στη μεταβλητή k το πλήθος των απαιτούμενων
% επαναλήψεων ( υπολογισμών της τακωβιανής ). .
%
% Στο ξεκίνημα, το x περιέχει μια αρχική εικασία.
% Αν το k ισούται με nmax, δεν έχει επιτευχθεί σύγκλιση.

% Οι προσεγγιστικές τιμές ||f(x_k)|| αποθηκεύονται.
% Αυτή η επιλογή μπορεί να απενεργοποιηθεί εύκολα.

% Αρχικοποίηση
x = x(:); % το x πρέπει να είναι διάνυσμα - στήλη
fprintf ('k %e \n',k-1,norm(fx) )
format long g

% Μέθοδος Newton
for k=1:nmax
    [fx,Jx] = feval(f,x);
    fprintf ('%d %e \n',k-1,norm(fx) )
    p = -Jx \ fx;
    x = x + p;
    if norm(p) < tol*(1+norm(x))
        fx = feval(f,x);
    end
end

```

```

fprintf ('%d      %e      \n', k, norm(fx) )
return
end
end
k = nmax;

```

Το κριτήριο τερματισμού στη συνάρτηση newtons μπορεί να τροποποιηθεί εύκολα ώστε να περιλαμβάνει και έναν έλεγχο της  $\|\mathbf{f}(\mathbf{x}_k)\|$ .

**Παράδειγμα 9.2.** Για το πρόβλημα που ορίστηκε στο Παράδειγμα 9.1 υπάρχουν δύο λύσεις (ρίζες),  $\mathbf{x}^{*(1)} = (1, 0)^T$  και  $\mathbf{x}^{*(2)} = (0, 1)^T$ . Η ιακωβιανή μήτρα γι' αυτό το πρόβλημα είναι

$$J(\mathbf{x}) = \begin{pmatrix} 2x_1 - 2 & -1 \\ 2x_1 & 2x_2 \end{pmatrix}$$

Σημειώστε ότι η ιακωβιανή δεν είναι υποχρεωτικά μη ιδιάζουσα για οποιαδήποτε επιλογή των  $x_1$  και  $x_2$ . Για παράδειγμα, η εκκίνηση της διαδικασίας των επαναλήψεων από το  $\mathbf{x}_0 = \mathbf{0}$  είναι προβληματική επειδή η  $J(\mathbf{0})$  είναι ιδιάζουσα. Όμως η ιακωβιανή είναι μη ιδιάζουσα στις δύο ρίζες και σε μια γειτονιά αυτών των ριζών.

Η εκκίνηση της μεθόδου Newton από τις δύο αρχικές εικασίες  $\mathbf{x}_0^{(1)} = (1, 1)^T$  και  $\mathbf{x}_0^{(2)} = (-1, 1)^T$  συγκλίνει γρήγορα στις δύο ρίζες αντίστοιχα. Η πρόοδος της  $\|\mathbf{f}(\mathbf{x}_k)\|$  καταγράφεται στον Πίνακα 9.1. Παρατηρήστε πόσο γρήγορα συγκλίνει η επανάληψη μόλις η νόρμα  $\|\mathbf{f}(\mathbf{x}_k)\|$  του υπολοίπου γίνει επαρκώς μικρή. ■

**Πίνακας 9.1.** Σύγκλιση της μεθόδου Newton στις δύο ρίζες του Παραδείγματος 9.1.

$k$	Πρώτη εικασία της $\ \mathbf{f}(\mathbf{x}_k)\ $	Δεύτερη εικασία της $\ \mathbf{f}(\mathbf{x}_k)\ $
0	1.414214e+00	3.162278e+00
1	1.274755e+00	7.218033e-01
2	2.658915e-01	1.072159e-01
3	3.129973e-02	4.561589e-03
4	3.402956e-04	9.556657e-06
5	7.094460e-08	4.157885e-11
6	1.884111e-15	

Μπορεί να αποδειχθεί ότι, αν σε μια γειτονιά κάποιας απομονωμένης ρίζας  $\mathbf{x}^*$  η ιακωβιανή  $J(\mathbf{x})$  διαθέτει φραγμένη αντίστροφη μήτρα και συνεχείς παραγώγους, τότε τοπικά η μέθοδος Newton συγκλίνει **τετραγωνικά**, δηλαδή υπάρχει σταθερά  $M$  τέτοια ώστε να ισχύει

$$\|\mathbf{x}^* - \mathbf{x}_{k+1}\| \leq M\|\mathbf{x}^* - \mathbf{x}_k\|^2$$

με την προϋπόθεση ότι η  $\|\mathbf{x}^* - \mathbf{x}_k\|$  είναι ήδη επαρκώς μικρή. Αυτή η τετραγωνική τάξη σύγκλισης (quadratic convergence order) φαίνεται καλύτερα στον Πίνακα 9.1.

**Παράδειγμα 9.3.** Τα μη γραμμικά συστήματα εξισώσεων ασφαλώς και δεν περιορίζονται σε δύο συνιστώσες. Στην πραγματικότητα, το  $n$  μπορεί εύκολα να πάρει μεγάλες τιμές σε διάφορες εφαρμογές. Ας δούμε ένα απλό παράδειγμα που οδηγεί σε ένα μεγαλύτερο σύνολο μη γραμμικών εξισώσεων.

Επεκτείνοντας το Παράδειγμα 4.17 (σελίδα 159), ας υποθέσουμε ότι πρέπει τώρα να βρούμε μια συνάρτηση  $v(t)$  η οποία να ικανοποιεί τη μη γραμμική συνήθη διαφορική εξίσωση (ΣΔΕ) συνοριακών τιμών

$$\begin{aligned} v''(t) + e^{v(t)} &= 0, \quad 0 < t < 1 \\ v(0) = v(1) &= 0 \end{aligned}$$

και ότι θα το κάνουμε προσεγγιστικά εφαρμόζοντας διακριτοποίηση πεπερασμένων διαφορών. Στην Ενότητα 16.7 θα αναλύσουμε αυτές τις μεθόδους και τα σχετικά προβλήματα, που απευθύνονται γενικά σε πιο προχωρημένους αναγνώστες: εδώ, το μόνο που θα κάνουμε είναι να επεκτείνουμε απευθείας τη μέθοδο του Παραδείγματος 4.17 και να επικεντρωθούμε στο μη γραμμικό σύστημα αλγεβρικών εξισώσεων που προκύπτει.

Διαιρούμε το διάστημα  $[0, 1]$  σε  $n + 1$  ίσα υποδιαστήματα και θέτουμε  $t_i = ih$ ,  $i = 0, 1, \dots, n + 1$ , όπου  $(n + 1)h = 1$ . Θέλουμε να βρούμε μια προσεγγιστική λύση  $v_i \approx v(t_i)$ ,  $i = 1, \dots, n$ , χρησιμοποιώντας τις συνοριακές συνθήκες για να θέσουμε  $v_0 = v_{n+1} = 0$ .

Η διακριτοποίηση της διαφορικής εξίσωσης δίνεται από τη σχέση

$$\frac{v_{i+1} - 2v_i + v_{i-1}}{h^2} + e^{v_i} = 0, \quad i = 1, 2, \dots, n$$

Αυτό είναι ένα σύστημα μη γραμμικών εξισώσεων,  $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ . Εδώ ισχύει  $\mathbf{x} \leftarrow \mathbf{v}$  και  $f_i(\mathbf{v}) = \frac{v_{i+1} - 2v_i + v_{i-1}}{h^2} + e^{v_i}$ .

Το στοιχείο  $(i, j)$  της ιακωβιανής μήτρας είναι το  $\frac{\partial f_i}{\partial v_j}$ . Τελικά προκύπτει ότι

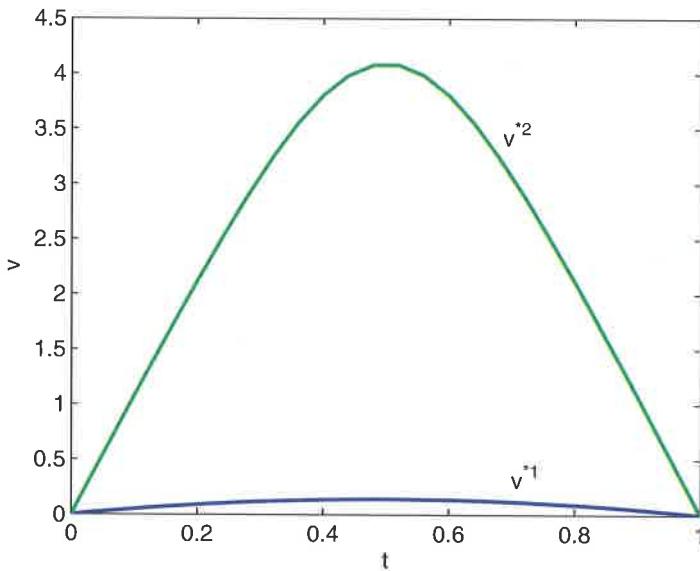
$$J = \frac{1}{h^2} \begin{pmatrix} -2 + h^2 e^{v_1} & 1 & & & \\ 1 & -2 + h^2 e^{v_2} & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 + h^2 e^{v_{n-1}} & 1 \\ & & & 1 & -2 + h^2 e^{v_n} \end{pmatrix}$$

Το μόνο που χρειαζόμαστε για να εκκινήσουμε τη μέθοδο Newton είναι μια αρχική εικασία  $v_0$ . Θα επιλέξουμε τις αντίστοιχες τιμές πλέγματος  $\alpha t(1-t)$  έτσι ώστε να ισχύει

$$v_0 = \alpha(t_1(1-t_1), \dots, t_n(1-t_n))^T$$

και θα δοκιμάσουμε διαφορετικές τιμές για την παράμετρο  $\alpha$ .

Θέτοντας  $tol = 1.e-8$  και  $h = 0.04$  (δηλαδή,  $n = 24$ ) παίρνουμε τα εξής αποτελέσματα. Για μηδενική αρχική εικασία  $\alpha = 0$ , η σύγκλιση επιτυγχάνεται σε 4 επαναλήψεις στη λύση που συμβολίζεται με  $v^{*1}$  στο Σχήμα 9.3. Για  $\alpha = 10$ , βρίσκουμε την ίδια λύση σε 6 επαναλήψεις. Για  $\alpha = 20$ , βρίσκουμε μια άλλη λύση, τη  $v^{*2}$ , σε 6 επαναλήψεις. Στο Σχήμα 9.3 απεικονίζεται και αυτή η λύση.



**Σχήμα 9.3:** Δύο λύσεις για μια συνήθη διαφορική εξίσωση συνοριακών τιμών.

Για  $\alpha = 50$ , ο αλγόριθμος αποκλίνει. Δυστυχώς, η απόκλιση δεν είναι σπάνιο φαινόμενο όταν η μέθοδος Newton χρησιμοποιείται χωρίς τη δέουσα προσοχή. ■

## Τροποποίηση της μεθόδου Newton

Στο Κεφάλαιο 3 περιγράψαμε τρόπους με τους οποίους μπορούν να αντιμετωπιστούν οι διάφορες ανεπάρκειες της μεθόδου Newton. Για παράδειγμα, η μέθοδος ενδέχεται να αποκλίνει όταν η αρχική εικασία δεν βρίσκεται επαρκώς κοντά στη λύση, και ο χρήστης πρέπει να καθορίζει τι αντιστοιχεί στην ιακωβιανή μήτρα. Εδώ προκύπτει επιπλέον η ανάγκη επίλυσης ενός δυνητικά μεγάλου συστήματος εξισώσεων σε κάθε επανάληψη  $k$ , οπότε οι τεχνικές των Κεφαλαίων 5 και 7 μπορούν να εφαρμοστούν άμεσα.

Σε αυτό το κεφάλαιο θα αναλύσουμε αυτά και άλλα ζητήματα. Προς το παρόν, όμως, θα εστιάσουμε στη βελτιστοποίηση χωρίς περιορισμούς, την οποία θα περιγράψουμε στην επόμενη ενότητα.

Πριν προχωρήσουμε, θέλουμε να επισημάνουμε ότι το γενικό πρόβλημα που εξετάζεται σε αυτή την ενότητα μπορεί να διατυπωθεί τεχνητά ως η ελαχιστοποίηση της  $\|\mathbf{f}(\mathbf{x})\|$  ή της  $\|\hat{\mathbf{J}}^{-1}\mathbf{f}(\mathbf{x})\|$  για κάποια σταθερή μήτρα  $\hat{\mathbf{J}}$  που αναπαριστά με κάποιον τρόπο την ιακωβιανή. Δεν συνιστούμε την επίλυση μη γραμμικών εξισώσεων σε αυτό το πλαίσιο, αλλά είναι εφικτό να χρησιμοποιηθούν ως κριτήρια για τη βελτίωση μιας ακολουθίας προσεγγιστικών τιμών. Για παράδειγμα, η επόμενη προσεγγιστική τιμή  $\mathbf{x}_{k+1}$  της μεθόδου Newton που περιγράφεται στη σελίδα 412 μπορεί να θεωρηθεί «καλύτερη» από την τρέχουσα επανάληψη  $\mathbf{x}_k$  αν ισχύει η ανισότητα  $\|\mathbf{f}(\mathbf{x}_{k+1})\| < \|\mathbf{f}(\mathbf{x}_k)\|$ .

*Ασκήσεις γι' αυτή την ενότητα: 1–9.*

## 9.2 Βελτιστοποίηση χωρίς περιορισμούς

Σε αυτή την ενότητα θα περιγράψουμε αριθμητικές μεθόδους για τη βελτιστοποίηση χωρίς περιορισμούς μιας συνάρτησης ως προς  $n$  μεταβλητές. Συνεπώς, το τυπικό πρόβλημα αυτής της κατηγορίας ορίζεται ως

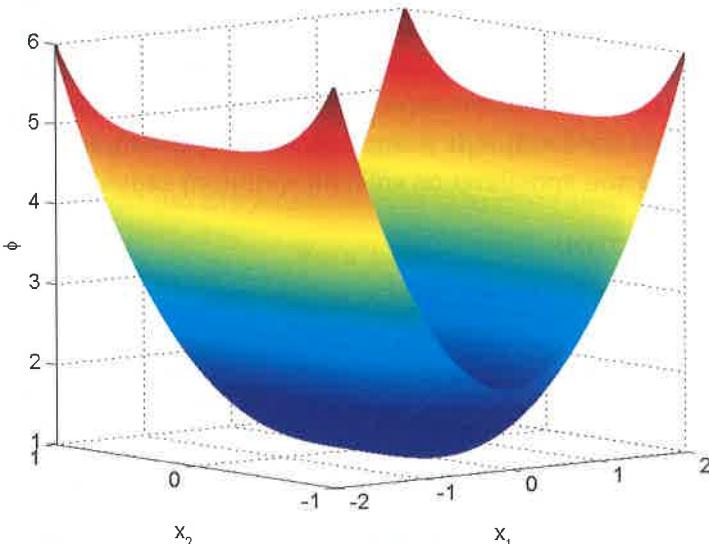
$$\min_{\mathbf{x} \in \mathcal{R}^n} \phi(\mathbf{x})$$

Εδώ, έχουμε  $\phi : \mathcal{R}^n \rightarrow \mathcal{R}$ . Δηλαδή, το όρισμα  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$  είναι διάνυσμα, όπως στην Ενότητα 9.1, αλλά η  $\phi$  παίρνει αριθμητικές τιμές.

**Παράδειγμα 9.4.** Ακολουθεί ένα πολύ απλό παράδειγμα. Για  $n = 2$ ,  $\mathbf{x} = (x_1, x_2)^T$ , καθορίζουμε τη συνάρτηση

$$\phi(\mathbf{x}) = x_1^2 + x_2^4 + 1$$

η οποία προφανώς έχει την ελάχιστη τιμή 1 στο  $\mathbf{x}^* = (0, 0)^T = \mathbf{0}$  (δείτε το Σχήμα 9.4).



**Σχήμα 9.4:** Η συνάρτηση  $x_1^2 + x_2^4 + 1$  έχει μοναδικό ελάχιστο στην αρχή  $(0, 0)$  των αξόνων και δεν έχει μέγιστο. Αν την αναποδογυρίζαμε, η συνάρτηση  $-(x_1^2 + x_2^4 + 1)$  που θα προέκυπτε θα είχε μοναδικό μέγιστο στην αρχή  $(0, 0)$  των αξόνων και δεν θα είχε ελάχιστο.

Γενικά, η εύρεση ενός μεγίστου για τη  $\phi(\mathbf{x})$  είναι η ίδια με την εύρεση ενός ελαχίστου για τη  $-\phi(\mathbf{x})$ . Σημειώστε ότι δεν υπάρχει πεπερασμένο όρισμα για το οποίο η συνάρτηση αυτού του παραδείγματος παίρνει τη μέγιστη τιμή. Επομένως, η συνάρτηση

$$\phi(\mathbf{x}) = -x_1^2 - x_2^4 - 1$$

δεν έχει ελάχιστο πουθενά, παρότι έχει μέγιστο στο  $\mathbf{x} = \mathbf{0}$ . ■

Για ρεαλιστικά, μη τετριμμένα προβλήματα συνήθως δεν μπορούμε να βρούμε σημεία ελαχίστων με απλή εξέταση. Συχνά δεν είναι ξεκάθαρο στην πράξη πόσα τοπικά ελάχιστα έχει μια συνάρτηση  $\phi(\mathbf{x})$  και, αν έχει περισσότερα από ένα, πώς μπορεί να βρεθεί με αποδοτικό τρόπο το ολικό ελάχιστο, που αντιστοιχεί στη συνολική μικρότερη τιμή για τη  $\phi$ .

Τα προβλήματα βελτιστοποίησης χωρίς περιορισμούς αποτελούν μια πολύ πλούσια πηγή συστημάτων μη γραμμικών εξισώσεων. Όμως διακυβεύονται περισσότερα

πράγματα εδώ. Όπως έχουμε δει στην Ενότητα 7.4, η ειδική περίπτωση της ελαχιστοποίησης συναρτήσεων με τετραγωνική αντικειμενική συνάρτηση δίνει ένα γραμμικό σύστημα εξισώσεων, οι οποίες πρέπει να επιλυθούν σε ένα επόμενο υποχρεωτικό στάδιο. Τότε, οι αλγόριθμοι ελαχιστοποίησης «μεταφράζονται» σε μεθόδους επίλυσης γραμμικών συστημάτων εξισώσεων. Μπορούμε να φτάσουμε στις μεθόδους της Ενότητας 7.4 με αυτόν τον τρόπο –δείτε το Παράδειγμα 9.7 παρακάτω. Αυτό είναι σημαντικό όταν η μήτρα είναι μεν πολύ μεγάλη αλλά ο πολλαπλασιασμός μήτρας διανύσματος δεν είναι δαπανηρός. Γενικότερα, για μεγάλα προβλήματα είναι εφικτό και χρήσιμο να σκεφτούμε κάποια παραλλαγή της μεθόδου Newton για τις μη γραμμικές εξισώσεις, σε συνδυασμό με μια μέθοδο από τις Ενότητες 7.4 έως 7.6 για το γραμμικό σύστημα που προκύπτει σε κάθε μη γραμμική επανάληψη.

## Συνθήκες ελαχίστου

**Θεώρημα: Σειρά Taylor για πολλές μεταβλητές.**

Έστω ότι  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ , και έστω ότι η  $\phi(\mathbf{x})$  έχει φραγμένες παραγώγους τουλάχιστον έως τρίτης τάξης. Τότε, για ένα διάνυσμα κατεύθυνσης  $\mathbf{p} = (p_1, p_2, \dots, p_n)^T$ , το ανάπτυγμα Taylor σε κάθε συντεταγμένη δίνει

$$\phi(\mathbf{x} + \mathbf{p}) = \phi(\mathbf{x}) + \nabla\phi(\mathbf{x})^T \mathbf{p} + \frac{1}{2} \mathbf{p}^T \nabla^2\phi(\mathbf{x}) \mathbf{p} + O(\|\mathbf{p}\|^3)$$

Εδώ,  $\nabla\phi(\mathbf{x})$  είναι το διάνυσμα **κλίσης** των πρώτων παραγώγων της  $\phi$  στο  $\mathbf{x}$  και  $\nabla^2\phi(\mathbf{x})$  είναι η **εσσιανή** (Hessian) μήτρα των δεύτερων παραγώγων της  $\phi$  στο  $\mathbf{x}$ , που δίνονται αντίστοιχα από τις σχέσεις

$$\nabla\phi(\mathbf{x}) = \begin{pmatrix} \frac{\partial\phi}{\partial x_1} \\ \frac{\partial\phi}{\partial x_2} \\ \vdots \\ \frac{\partial\phi}{\partial x_n} \end{pmatrix}, \quad \nabla^2\phi(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2\phi}{\partial x_1^2} & \frac{\partial^2\phi}{\partial x_1\partial x_2} & \cdots & \frac{\partial^2\phi}{\partial x_1\partial x_n} \\ \frac{\partial^2\phi}{\partial x_2\partial x_1} & \frac{\partial^2\phi}{\partial x_2^2} & \cdots & \frac{\partial^2\phi}{\partial x_2\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2\phi}{\partial x_n\partial x_1} & \frac{\partial^2\phi}{\partial x_n\partial x_2} & \cdots & \frac{\partial^2\phi}{\partial x_n^2} \end{pmatrix}$$

Ο όρος του υπολοίπου  $O(\|\mathbf{p}\|^3)$  εξαρτάται από παραγώγους της  $\phi$  τάξης 3 ή ανώτερης.

Υποθέτουμε ότι η  $\phi(\mathbf{x})$  έχει συνεχείς παραγώγους τουλάχιστον έως δεύτερης τάξης, τις οποίες συμβολίζουμε με  $\phi \in C^2$ . Επομένως, μπορεί να αναπτυχθεί σε μια σειρά Taylor όπως ορίζεται παρακάτω.

Η βαθμωτή συνάρτηση

$$\nabla\phi(\mathbf{x})^T \mathbf{p} = \sum_{i=1}^n p_i \frac{\partial\phi}{\partial x_i}$$

είναι η παράγωγος κατά κατεύθυνση (directional derivative) της  $\phi$  στο  $\mathbf{x}$  στην κατεύθυνση  $\mathbf{p}$ . Ο επόμενος όρος στο ανάπτυγμα είναι η τετραγωνική μορφή

$$\frac{1}{2} \mathbf{p}^T \nabla^2 \phi(\mathbf{x}) \mathbf{p} = \frac{1}{2} \sum_{i,j=1}^n \frac{\partial^2 \phi}{\partial x_i \partial x_j} p_i p_j$$

Αν το  $\mathbf{x}^*$  είναι τοπικό ελάχιστο, δηλαδή ένα σημείο όπου η τιμή της  $\phi$  είναι μικρότερη από ή ίση με την τιμή της σε όλα τα γειτονικά σημεία, τότε για οποιαδήποτε κατεύθυνση  $\mathbf{p}$  έχουμε

$$\phi(\mathbf{x}^* + \mathbf{p}) = \phi(\mathbf{x}^*) + \nabla \phi(\mathbf{x}^*)^T \mathbf{p} + \frac{1}{2} \mathbf{p}^T \nabla^2 \phi(\mathbf{x}^*) \mathbf{p} + O(\|\mathbf{p}\|^3) \geq \phi(\mathbf{x}^*)$$

Από αυτό προκύπτει ότι μια αναγκαία συνθήκη ελαχίστου είναι η

$$\nabla \phi(\mathbf{x}^*) = \mathbf{0}$$

Ένα τέτοιο σημείο στο οποίο μηδενίζεται η κλίση ονομάζεται **κρίσιμο σημείο** (critical point), ή ακρότατο. Για το Παράδειγμα 9.4 έχουμε  $2x_1 = 0$  και  $4x_2^3 = 0$ , από το οποίο βρίσκουμε εύκολα ότι  $x_1 = x_2 = 0$ .

Επιπλέον, μια **ικανή συνθήκη** ώστε ένα κρίσιμο σημείο να είναι στην πραγματικότητα ελάχιστο είναι ότι η εσσιανή μήτρα  $\nabla^2 \phi(\mathbf{x}^*)$  πρέπει να είναι συμμετρική και θετικά ορισμένη.<sup>48</sup>

Πώς μπορούμε να τα αντιληφθούμε όλα αυτά; Φανταστείτε ότι το  $\|\mathbf{p}\|$  είναι πολύ μικρό, ώστε να ισχύει  $\|\mathbf{p}\|^2 \ll \|\mathbf{p}\|$ . Τώρα, αν ισχύει επίσης ότι  $\nabla \phi(\mathbf{x}^*) \neq \mathbf{0}$ , είναι πάντα εφικτό να βρεθεί μια κατεύθυνση  $\mathbf{p}$  τέτοια ώστε να ισχύει η ανισότητα  $\nabla \phi(\mathbf{x}^*)^T \mathbf{p} < 0$ . Άρα,  $\phi(\mathbf{x}^* + \mathbf{p}) < \phi(\mathbf{x}^*)$  και δεν υπάρχει ελάχιστο στο  $\mathbf{x}^*$ . Συνεπώς, πρέπει να ισχύει  $\nabla \phi(\mathbf{x}^*) = \mathbf{0}$ . Με έναν παρόμοιο συλλογισμό μπορούμε να δείξουμε ότι η κλίση πρέπει να μηδενίζεται στα σημεία όπου η  $\phi(\mathbf{x})$  έχει τοπικά ελάχιστα. Η κλίση πρέπει επίσης να μηδενίζεται στα **σαγματικά σημεία** (saddle points), δηλαδή σε σημεία όπου επιτυγχάνεται ένα μέγιστο ως προς κάποιες μεταβλητές και ένα ελάχιστο ως προς άλλες παραμέτρους. Ένα σαγματικό σημείο δεν είναι ούτε ελαχιστοποιητής ούτε μεγιστοποιητής. Ας δούμε ένα απλό παράδειγμα. Η συνάρτηση  $\phi(\mathbf{x}) = x_1^2 - x_2^4 + 1$  έχει ένα σαγματικό σημείο στην αρχή των αξόνων (δείτε την Άσκηση 10).

<sup>48</sup> Θυμηθείτε ότι μια συμμετρική μήτρα  $A$  είναι θετικά ορισμένη αν ισχύει  $\mathbf{x}^T A \mathbf{x} > 0$  για κάθε  $\mathbf{x} \neq 0$ . Οι ιδιοτιμές μιας τέτοιας μήτρας είναι όλες θετικές, και ένα πόρισμα που προκύπτει άμεσα είναι ότι η  $A$  πρέπει να είναι μη ιδιάζουσα. Μια συμμετρική μήτρα  $A$  είναι θετικά ημιορισμένη αν ισχύει  $\mathbf{x}^T A \mathbf{x} \geq 0$  για κάθε  $\mathbf{x}$ . Σε αυτή την περίπτωση, όλες οι ιδιοτιμές είναι μη αρνητικές, οπότε η πιθανότητα να είναι η μήτρα ιδιάζουσα δεν μπορεί να εξαλειφθεί.

Στη συνέχεια, σε ένα κρίσιμο σημείο το οποίο είναι αυστηρό ελάχιστο πρέπει επίσης, για όλες τις κατευθύνσεις  $\mathbf{p}$  που ικανοποιούν την ανισότητα  $0 < \|\mathbf{p}\| \ll 1$ , να ισχύει

$$\phi(\mathbf{x}^* + \mathbf{p}) = \phi(\mathbf{x}^*) + \frac{1}{2} \mathbf{p}^T \nabla^2 \phi(\mathbf{x}^*) \mathbf{p} + O(\|\mathbf{p}\|^3) > \phi(\mathbf{x}^*)$$

Κάτι τέτοιο θα ισχύει αν η εσσιανή μήτρα  $\nabla^2 \phi(\mathbf{x}^*)$  είναι θετικά ορισμένη, επειδή αυτό μας εγγυάται ότι θα ισχύει η ανισότητα  $\mathbf{p}^T \nabla^2 \phi(\mathbf{x}^*) \mathbf{p} > 0$ . ♦

### Θεώρημα: Συνθήκες ελαχιστοποίησης χωρίς περιορισμούς.

Υποθέτουμε ότι η  $\phi(\mathbf{x})$  είναι επαρκώς λεία, π.χ. έστω ότι έχει φραγμένες όλες τις παραγώγους έως τρίτης τάξης. Τότε,

- Μια αναγκαία συνθήκη για την ύπαρξη τοπικού ελαχίστου σε ένα σημείο  $\mathbf{x}^*$  είναι ότι το  $\mathbf{x}^*$  πρέπει να είναι κρίσιμο σημείο, δηλαδή να ισχύει

$$\nabla \phi(\mathbf{x}^*) = \mathbf{0}$$

και η συμμετρική εσσιανή μήτρα  $\nabla^2 \phi(\mathbf{x}^*)$  πρέπει να είναι θετικά ημιορισμένη.

- Μια ικανή συνθήκη για την ύπαρξη τοπικού ελαχίστου σε ένα σημείο  $\mathbf{x}^*$  είναι ότι το  $\mathbf{x}^*$  πρέπει να είναι κρίσιμο σημείο και ότι η μήτρα  $\nabla^2 \phi(\mathbf{x}^*)$  πρέπει να είναι θετικά ορισμένη.

### Βασικές μέθοδοι για βελτιστοποίηση χωρίς περιορισμούς

Σε γενικές γραμμές, η συνθήκη για ένα κρίσιμο σημείο δίνει ένα σύστημα μη γραμμικών εξισώσεων

$$\mathbf{f}(\mathbf{x}) \equiv \nabla \phi(\mathbf{x}) = \mathbf{0}$$

Άρα, η επίλυση ως προς ένα κρίσιμο σημείο αποτελεί ειδική περίπτωση της επίλυσης ενός συστήματος μη γραμμικών εξισώσεων, και επομένως μπορούν να εφαρμοστούν απευθείας μέθοδοι όπως η μέθοδος Newton. Σημειώστε ότι στην ιακωβιανή μήτρα  $J(\mathbf{x})$  αντιστοιχεί η εσσιανή μήτρα της  $\phi$ ,  $J(\mathbf{x}) = \nabla^2 \phi(\mathbf{x})$ . Επειδή η εσσιανή μήτρα συνήθως είναι συμμετρική και θετικά ορισμένη κοντά σε ένα ελάχιστο, αυτή είναι όντως μια σημαντική ειδική περίπτωση συστημάτων μη γραμμικών εξισώσεων.

## Η μέθοδος Newton για ελαχιστοποίηση χωρίς περιορισμούς

Ο αλγόριθμος (της μεθόδου) Newton για το πρόβλημα της ελαχιστοποίησης χωρίς περιορισμούς παρουσιάζεται παρακάτω.

**Αλγόριθμος: Μέθοδος Newton για ελαχιστοποίηση χωρίς περιορισμούς.**

Για το πρόβλημα της ελαχιστοποίησης της  $\phi(\mathbf{x})$  στο  $\mathcal{R}^n$ , έστω ότι  $\mathbf{x}_0$  είναι μια αρχική εικασία που δίνεται.

for  $k = 0, 1, \dots$  μέχρι τη σύγκλιση

Λύσε το σύστημα  $\nabla^2\phi(\mathbf{x}_k) \mathbf{p}_k = -\nabla\phi(\mathbf{x}_k)$  ως προς  $\mathbf{p}_k$

Θέσε  $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$

end

**Παράδειγμα 9.5.** Θέλουμε να ελαχιστοποιήσουμε τη συνάρτηση

$$\phi(\mathbf{x}) = \frac{1}{2} \left( [1.5 - x_1(1 - x_2)]^2 + [2.25 - x_1(1 - x_2^2)]^2 + [2.625 - x_1(1 - x_2^3)]^2 \right)$$

Ένα κρίσιμο σημείο ορίζεται από τις εξισώσεις

$$\nabla\phi(\mathbf{x}) = \mathbf{f}(\mathbf{x}) = \mathbf{0}$$

όπου

$$\begin{aligned} f_1(x_1, x_2) &= -(1.5 - x_1(1 - x_2))(1 - x_2) - (2.25 - x_1(1 - x_2^2))(1 - x_2^2) \\ &\quad - (2.625 - x_1(1 - x_2^3))(1 - x_2^3) \end{aligned}$$

$$\begin{aligned} f_2(x_1, x_2) &= x_1(1.5 - x_1(1 - x_2)) + 2x_1x_2(2.25 - x_1(1 - x_2^2)) \\ &\quad + 3x_1x_2^2(2.625 - x_1(1 - x_2^3)) \end{aligned}$$

Όπως αποδεικνύεται, υπάρχει μοναδικό ελάχιστο γι' αυτό το πρόβλημα στο  $\mathbf{x}^* = (3, 0.5)^T$ ,  $\phi(\mathbf{x}^*) = 0$ . Όμως υπάρχει και ένα σαγματικό σημείο: Στο  $\hat{\mathbf{x}} = (0, 1)^T$  η κλίση μηδενίζεται, άρα  $\mathbf{f}(\hat{\mathbf{x}}) = \mathbf{0}$ , αλλά η εσσιανή μήτρα της  $\phi$  (που είναι η ιακωβιανή της  $\mathbf{f}$ ) είναι  $J = \begin{pmatrix} 0 & 13.875 \\ 13.875 & 0 \end{pmatrix}$ . Οι ιδιοτιμές είναι  $\pm 13.875$ , άρα η εσσιανή μήτρα δεν είναι θετικά ημιορισμένη στο  $\hat{\mathbf{x}}$  και οι αναγκαίες συνθήκες του ελαχίστου δεν ικανοποιούνται στο σημείο αυτό. Για μια γραφική αναπαράσταση δείτε το Σχήμα 9.5.

Ξεκινώντας τη μέθοδο Newton από το  $\mathbf{x}_0 = (8, 0.2)^T$ , παίρνουμε τις επαναλήψεις που παρουσιάζονται στο αριστερό τμήμα του Πίνακα 9.2. Παρατηρούμε ότι επιτυγχάνεται πολύ γρήγορη σύγκλιση στο ελάχιστο. Πρέπει να επισημάνουμε ότι οι τιμές της  $\phi$  μειώνονται μονοτονικά εδώ, παρότι η εσσιανή μήτρα στην αρχική εικασία,  $\nabla^2\phi(\mathbf{x}_0)$ , δεν είναι θετικά ορισμένη: Απλώς η τύχη είναι με το μέρος μας σε

**Πίνακας 9.2.** Παράδειγμα 9.5. Οι πρώτες τρεις στήλες στα δεξιά του μετρητή των επαναλήψεων παρακολουθούν τη σύγκλιση ζεκινώντας από το  $\mathbf{x}_0 = (8, 0.2)^T$ : Η μέθοδος Newton βρίσκει γρήγορα το ελάχιστο. Οι επόμενες τρεις (δεξιές) στήλες παρακολουθούν τη σύγκλιση ζεκινώντας από το  $\mathbf{x}_0 = (8, 0.8)^T$ : Η μέθοδος Newton βρίσκει ένα κρίσιμο σημείο, το οποίο όμως δεν είναι ο ελαχιστοποιητής.

$k$	$\ \mathbf{x}_k - \mathbf{x}^*\ $	$\phi_k - \phi(\mathbf{x}^*)$	$\mathbf{f}_k^T \mathbf{p}_k$	$\ \mathbf{x}_k - \mathbf{x}^*\ $	$\phi_k - \phi(\mathbf{x}^*)$	$\mathbf{f}_k^T \mathbf{p}_k$
0	5.01e+00	4.09e+01	-7.27e+01	5.01e+00	1.02e+00	-1.65e+00
1	8.66e-01	1.21e+00	-2.26e+00	3.96e+00	1.28e-01	-2.94e-02
2	6.49e-02	1.20e-02	-2.32e-02	4.22e+00	1.16e-01	-3.21e-01
3	1.39e-01	1.72e-03	-3.16e-03	1.66e+01	2.87e+02	-4.65e+02
4	2.10e-02	6.91e-05	-1.35e-04	6.14e+00	2.61e+01	-3.16e+01
5	1.38e-03	1.43e-07	-2.86e-07	3.43e+00	7.98e+00	-1.85e+00
6	3.03e-06	1.09e-12	-2.19e-12	2.97e+00	7.10e+00	-4.22e-03
7	2.84e-11	6.16e-23	-1.23e-22	3.04e+00	7.10e+00	-2.84e-08
8				3.04e+00	7.10e+00	-5.42e-19

αυτή την περίπτωση.

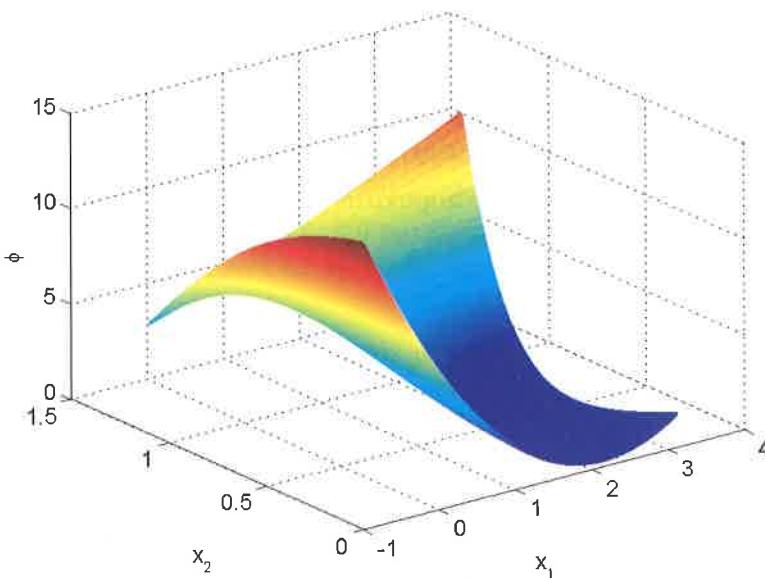
Ζεκινώντας από το  $\mathbf{x}_0 = (8, 0.8)^T$  παίρνουμε τις επαναλήψεις που παρουσιάζονται στο δεξιό τμήμα του Πίνακα 9.2. Και πάλι επιτυγχάνεται πολύ γρήγορη σύγκλιση, δυστυχώς όμως στο σαγματικό σημείο  $\hat{\mathbf{x}} = (1, 0)^T$  και όχι στο ελάχιστο. Σημειώστε ότι όλες οι κατευθύνσεις που παρατηρούνται είναι κατευθύνσεις καθόδου: Για επαρκώς μικρό  $\alpha > 0$  ισχύει  $\phi(\mathbf{x}_k + \alpha \mathbf{p}_k) < \phi(\mathbf{x}_k)$ . Αυτό μπορεί να συμβεί αν το κρίσιμο σημείο είναι σαγματικό (και όχι μέγιστο), παρότι οι εσσιανές μήτρες που προκύπτουν σαφώς και δεν είναι όλες θετικά ορισμένες. ■

### Μια κατηγορία μεθόδων

Το δίδαγμα που προκύπτει από το Παράδειγμα 9.5 είναι ότι το πρόβλημα της ελαχιστοποίησης χωρίς περιορισμούς δεν είναι απλώς μια ειδική περίπτωση επίλυσης μη γραμμικών εξισώσεων αλλά κάτι περισσότερο, αν και είναι χρήσιμο να το δούμε από αυτή την οπτική γωνία. Το πρόβλημα εμπεριέχει περισσότερες πληροφορίες τις οποίες μπορούμε να αξιοποιήσουμε. Ειδικότερα, για οποιαδήποτε κατεύθυνση  $\mathbf{p}$  σε ένα σημείο  $\mathbf{x}$  όπου  $\nabla \phi(\mathbf{x}) \neq \mathbf{0}$ , αν η τιμή της  $\|\mathbf{p}\|$  είναι πολύ μικρή, τότε για να ισχύει η ανισότητα  $\phi(\mathbf{x} + \mathbf{p}) < \phi(\mathbf{x})$  η παράγωγος κατά κατεύθυνση πρέπει να είναι αρνητική, δηλαδή

$$\nabla \phi(\mathbf{x})^T \mathbf{p} < 0$$

Αυτό έπεται άμεσα από το ανάπτυγμα Taylor που παρουσιάζεται στη σελίδα 418. Ένα διάνυσμα  $\mathbf{p}$  για το οποίο ισχύει η συγκεκριμένη ανισότητα ονομάζεται **κατεύθυνση καθόδου** (descent direction) στο σημείο  $\mathbf{x}$ .



**Σχήμα 9.5:** Η συνάρτηση του Παραδείγματος 9.5 έχει μοναδικό ελάχιστο στο  $\mathbf{x}^* = (3, 0.5)^T$  και σαγματικό σημείο στο  $\hat{\mathbf{x}} = (0, 1)^T$ .

Η πλειοψηφία των μεθόδων ελαχιστοποίησης χωρίς περιορισμούς χρησιμοποιούν επαναλήψεις της μορφής

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{x}_k + \alpha_k \mathbf{p}_k, && \text{όπου} \\ \mathbf{p}_k &= -B_k^{-1} \nabla \phi(\mathbf{x}_k)\end{aligned}$$

Συνεπώς, αν η μήτρα  $B_k$  είναι συμμετρική θετικά ορισμένη, το  $\mathbf{p}_k$  είναι κατεύθυνση καθόδου επειδή ισχύει  $\nabla \phi(\mathbf{x}_k)^T \mathbf{p}_k = -\nabla \phi(\mathbf{x}_k)^T B_k^{-1} \nabla \phi(\mathbf{x}_k) < 0$ . Στη συνέχεια θα εξετάσουμε πολλές διαφορετικές περιπτώσεις για την επιλογή της μήτρας  $B_k$ .

Για να κατανοήσουμε τα κίνητρα και να εστιάσουμε μια τέτοια μέθοδο αναζήτησης, αξίζει να αναλύσουμε τα πλεονεκτήματα και τα μειονεκτήματα της βασικής μεθόδου Newton. Στα πλεονεκτήματα συμπεριλαμβάνονται (εκτός από το γεγονός ότι είμαστε ήδη εξοικειωμένοι με τη μέθοδο) η τετραγωνική τοπική σύγκλιση –δείτε την Άσκηση 3– και το γεγονός ότι στην περίπτωση μεγάλων προβλημάτων, αν η εσιανή μήτρα είναι αραιή, τότε και η  $B_k$  είναι στοιχειωδώς αραιή. Στα μειονεκτήματα συμπεριλαμβάνονται τα εξής:

- Απαιτεί να υπάρχει η εσιανή μήτρα.
- Επιπλέον (και αυτό είναι ακόμα χειρότερο), απαιτεί τον υπολογισμό της εσιανής μήτρας.
- Απαιτεί την επίλυση ενός γραμμικού συστήματος σε κάθε επανάληψη.

- Η μήτρα  $B_k = \nabla^2\phi(\mathbf{x}_k)$  ενδέχεται να μην είναι συμμετρική θετικά ορισμένη μακριά από το ελάχιστο.
- Δεν υπάρχει κανένας έλεγχος για τη σύγκλιση. (Για παράδειγμα, συγκλίνει όντως; Αν ναι, συγκλίνει σε σημείο ελαχίστου;)

Όλες οι μέθοδοι που περιγράφονται παρακάτω μπορούν να θεωρηθούν ως απόπειρες να εξαλειφθούν κάποιες από τις δυσκολίες που μόλις παραθέσαμε.

## Καθοδική κλίση

Η απλούστερη επιλογή για τη μήτρα  $B_k$  είναι η ταυτοτική μήτρα  $I$ . Έτσι προκύπτει η κατεύθυνση καθοδικής κλίσης (gradient descent)

$$\mathbf{p}_k = -\nabla\phi(\mathbf{x}_k)$$

Η κατεύθυνση καθοδικής κλίσης αποτελεί εγγυημένα κατεύθυνση καθόδου. Επιπλέον, αποφεύγεται και η επίλυση ενός γραμμικού συστήματος εξισώσεων σε κάθε επανάληψη, όπως απαιτείται από τη μέθοδο Newton.

Όμως πρέπει πάντα να καθορίζουμε το μέγεθος βήματος  $\alpha_k$ . Ακόμα χειρότερα, όπως φαίνεται και στο Παράδειγμα 9.7 (σελίδα 428), το οποίο θα δούμε σύντομα, η συγκεκριμένη μέθοδος ενδέχεται να συγκλίνει βασανιστικά αργά.

Με την κατεύθυνση καθοδικής κλίσης μπορεί να εξαλειφθεί ένα μειονέκτημα της μεθόδου Newton, αν οι δύο μέθοδοι συνδυαστούν:

$$B_k = \nabla^2\phi(\mathbf{x}_k) + \mu_k I$$

Η παράμετρος ανάμειξης  $\mu_k \geq 0$  επιλέγεται έτσι ώστε να διασφαλίζεται ότι η  $B_k$  είναι συμμετρική θετικά ορισμένη μήτρα, και άρα ότι το  $\mathbf{p}_k$  είναι κατεύθυνση καθόδου. Υπάρχουν πολύπλοκες μέθοδοι περιοχής εμπιστοσύνης (trust region) για τον προσδιορισμό αυτής της παραμέτρου σε κάθε επανάληψη  $k$ , οι οποίες όμως ξεφεύγουν από το αντικείμενο αυτού του βιβλίου.

## Ευθύγραμμη αναζήτηση

Οι τεχνικές ευθύγραμμης αναζήτησης (line search) ασχολούνται με την επιλογή του αριθμητικού μήκους βήματος  $\alpha_k$  για την ανανέωση

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$$

της επανάληψης, όπου  $\mathbf{p}_k$  είναι η ήδη καθορισμένη κατεύθυνση αναζήτησης. Για τη μέθοδο («αμιγούς») καθοδικής κλίσης αυτό είναι προαπαιτούμενο.

Για τη μη τροποποιημένη μέθοδο Newton, και σύμφωνα με τη λογική στην οποία στηρίζεται η μέθοδος, η προεπιλογή είναι  $\alpha_k = 1$ . Αυτό δίνει καλά αποτελέ-

σματα όταν η  $\|\mathbf{p}_k\|$  είναι μικρή, αλλά όχι πάντα όταν η  $\|\mathbf{p}_k\|$  είναι μεγάλη. Η έλλειψη εγγυήσεων για τη σύγκλιση από απομακρυσμένα σημεία εκκίνησης  $\mathbf{x}_0$ , όπως στο Παράδειγμα 9.3 για  $\alpha = 50$ , συχνά αποτελεί σημαντική πηγή ανησυχίας στην πράξη. Γι' αυτόν τον λόγο, το μήκος βήματος  $\alpha_k$ ,  $0 < \alpha_k \leq 1$ , επιλέγεται έτσι ώστε να είναι εγγυημένη η μείωση της αντικειμενικής συνάρτησης  $\phi$ .

Επομένως, αν μας δίνεται το  $\mathbf{x}_k$  και μια κατεύθυνση κατάβασης  $\mathbf{p}_k$ , εκτελούμε αναζήτηση κατά μήκος της ευθείας  $\mathbf{x}_k + \alpha \mathbf{p}_k$  για μια τιμή  $\alpha = \alpha_k$  τέτοια ώστε να ισχύει

$$\phi(\mathbf{x}_{k+1}) \equiv \phi(\mathbf{x}_k + \alpha_k \mathbf{p}_k) \leq \phi(\mathbf{x}_k) + \sigma \alpha_k \nabla \phi(\mathbf{x}_k)^T \mathbf{p}_k$$

όπου  $\sigma$  είναι μια σταθερά προστασίας (guard constant), π.χ.  $\sigma = 10^{-4}$ . Συνήθως ξεκινάμε με μια τιμή  $\alpha = \alpha_{\max}$  την οποία μειώνουμε αν χρειαστεί.

Ένας απλός αλγόριθμος **υπαναχώρησης** (backtracking) είναι ο έλεγχος της παραπάνω ανισότητας βελτίωσης για γεωμετρικά φθίνουσες τιμές του  $\alpha$ , που μπορεί να επιλέγονται σύμφωνα με τη σχέση

$$\alpha / \alpha_{\max} = 1, 1/2, 1/4, \dots, (1/2)^j, \dots$$

ο οποίος σταματάει μόλις βρεθεί μια κατάλληλη τιμή για το  $\alpha_k$  που να ικανοποιεί το παραπάνω κριτήριο. Αυτή είναι μια συγκεκριμένη στρατηγική **ασθενούς ευθύγραμμης αναζήτησης** (weak line search).

Μια άλλη τέτοια στρατηγική είναι η ακόλουθη. Παρατηρούμε πως για τη συνάρτηση

$$\psi(\alpha) = \phi(\mathbf{x}_k + \alpha \mathbf{p}_k)$$

γνωρίζουμε ότι  $\psi(0) = \phi(\mathbf{x}_k)$ ,  $\psi'(0) = \mathbf{p}^T \nabla \phi(\mathbf{x}_k)$  και  $\psi(\tilde{\alpha}_k) = \phi(\mathbf{x}_k + \tilde{\alpha}_k \mathbf{p}_k)$ , όπου  $\tilde{\alpha}_k$  είναι η τρέχουσα, μη ικανοποιητική τιμή για το  $\alpha_k$ . Άρα, «περνάμε» ένα τετραγωνικό πολυώνυμο από αυτά τα τρία σημεία δεδομένων (δείτε την Ενότητα 10.7 και ειδικότερα την Άσκηση 10.25) και ελαχιστοποιούμε τη συγκεκριμένη τετραγωνική συνάρτηση παρεμβολής, παίρνοντας

$$\alpha_k = \frac{-\psi'(0)\tilde{\alpha}_k^2}{2(\psi(\tilde{\alpha}_k) - \psi(0) - \tilde{\alpha}_k \psi'(0))}$$

Προσέξτε ότι αυτό μπορεί να λειτουργήσει σωστά μόνο αν ισχύει  $\psi'(0) < 0$ . Ακολουθεί ένα απόσπασμα κώδικα που χρησιμοποιεί ένα μείγμα των δύο παραπάνω στρατηγικών, με την κατεύθυνση αναζήτησης  $\mathbf{p}$ , την τιμή  $\text{phix}$  της συνάρτησης, και την κλίση  $\text{gphix}$  να δίνονται στο τρέχον  $\mathbf{x} = \mathbf{x}_k$ .

```
pgphi = p' * gphix;
alpha = alphamax;
```

```

xn = x + alpha * p; phixn = feval(phi,xn);
while (phixn > phix + sigma * alpha * pgphi) * (alpha > alphamin)
    mu = -0.5 * pgphi * alpha / (phixn - phix - alpha * pgphi );
    if mu < 0.1 || pgphi >= 0
        mu = 0.5;           % δεν εμπιστευόμαστε την τετραγωνική
                           % παρεμβολή από απομακρυσμένα σημεία
    end
    alpha = mu * alpha;
    xn = x + alpha * p;
    phixn = feval(phi,xn);
end

```

Για τη μέθοδο Newton, μόλις το  $\mathbf{x}_k$  πλησιάσει στο  $\mathbf{x}^*$ , η καλύτερη τιμή για το  $\alpha_k$  αναμένεται να είναι  $\alpha_k = 1$ , με αποτέλεσμα να ανακτηθεί η γρήγορη σύγκλιση της αμιγούς μεθόδου. Αν χρησιμοποιούμε τη μέθοδο καθοδικής κλίσης, όμως, δεν υπάρχει κάποια φυσική τιμή για το  $\alpha_k$  στη γενική περίπτωση.

### Παράδειγμα 9.6. Θεωρήστε τη συνάρτηση

$$\phi(\mathbf{x}) = x_1^4 + x_1 x_2 + (1 + x_2)^2$$

Υπάρχει ένα και μοναδικό ελάχιστο στο  $\mathbf{x}^* \approx (0.695884386, -1.34794219)^T$ , όπου  $\phi(\mathbf{x}^*) \approx -0.582445174$ .

Ξεκινώντας από το  $\mathbf{x}_0 = (0.75, -1.25)^T$ , με την αμιγή μέθοδο Newton επιτυγχάνουμε πάρα πολύ γρήγορη σύγκλιση.

Ωστόσο, αν ξεκινήσουμε τις επαναλήψεις της αμιγούς μεθόδου Newton από το  $\mathbf{x}_0 = (0, 0.3)^T$ , δεν επιτυγχάνεται σύγκλιση. Για να κατανοήσουμε γιατί συμβαίνει αυτό, παρατηρούμε ότι η μήτρα  $\nabla^2\phi(\mathbf{x}) = \begin{pmatrix} 12x_1^2 & 1 \\ 1 & 2 \end{pmatrix}$  είναι θετικά ορισμένη μόνο όταν ισχύει η ανισότητα  $24x_1^2 > 1$ . Ειδικότερα, η εσσιανή μήτρα είναι ιδιάζουσα για  $x_1 = \pm\frac{1}{\sqrt{24}}$ . Η αρχική εικασία  $\mathbf{x}_0 = (0, 0.3)^T$  και οι επόμενες προσεγγιστικές τιμές της μεθόδου Newton παγιδεύονται στην «κακή περιοχή» της εσσιανής μήτρας.

Αν ενεργοποιήσουμε την ασθενή ευθύγραμμη αναζήτηση, το «καλό» σημείο εκκίνησης δίνει την ίδια ακολουθία προσεγγιστικών τιμών όπως η αμιγής μέθοδος Newton, δηλαδή  $\alpha_k = 1$  για όλα τα  $k$  που απαντώνται. Αντιθέτως, αν ξεκινήσουμε από το  $\mathbf{x}_0 = (0, 0.3)^T$  και χρησιμοποιήσουμε τη στρατηγική ευθύγραμμης αναζήτησης, επιτυγχάνουμε τη σύγκλιση σε 6 επαναλήψεις. Η σύνεση αποδίδει όντως καρπούς. Τα πρώτα τρία μήκη βήματος είναι μικρότερα από 0.5 το καθένα, ενώ για τα τελευταία τρία οι επαναλήψεις της αμιγούς μεθόδου Newton επιτυγχάνουν πολύ γρήγορη σύγκλιση. ■

Υπάρχει το θεωρητικό υπόβαθρο για να αποδειχθεί ότι, υπό κατάλληλες συνθήκες, είναι εγγυημένη η σύγκλιση σε τοπικό ελάχιστο από οποιαδήποτε αρχική εικασία  $\mathbf{x}_0$  σε μια περιοχή τιμών που δεν είναι πλέον μικρή. Συντρέχουν όμως και λόγοι ανησυχίας. Ένας από αυτούς είναι ότι αν η μέθοδος Newton απαιτεί όντως ευθύγραμμη αναζήτηση, αυτό μπορεί κάλλιστα να αποτελεί ένδειξη ότι η ίδια η κατεύθυνση αναζήτησης  $\mathbf{p}_k$  δεν είναι καλή. Η τροποποίηση αυτής της κατεύθυνσης, όπως συμβαίνει στη μέθοδο περιοχής εμπιστοσύνης, μπορεί τότε να αποδειχθεί πιο προσοδοφόρα. Ένας άλλος λόγος ανησυχίας για όλες αυτές τις μεθόδους εντοπίζεται στο γεγονός ότι είναι **άπληστες** (greedy), προσπαθώντας να πετύχουν αυτό που είναι καλύτερο τοπικά παρά καθολικά.

**Παράδειγμα 9.7 (Κυρτή τετραγωνική ελαχιστοποίηση).** Θα μελετήσουμε το απλούστερο πρόβλημα μη γραμμικής βελτιστοποίησης, δηλαδή το πρόβλημα της ελαχιστοποίησης χωρίς περιορισμούς τετραγωνικών συναρτήσεων της μορφής

$$\phi(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T A \mathbf{x} - \mathbf{b}^T \mathbf{x}$$

όπου  $A$  είναι μια συμμετρική και θετικά ορισμένη μήτρα διαστάσεων  $n \times n$  και  $\mathbf{b}$  είναι ένα  $n$ -διάνυσμα, που δίνονται.

Σημειώστε ότι ισχύει  $\nabla \phi(\mathbf{x}) = A\mathbf{x} - \mathbf{b}$  και  $\nabla^2 \phi(\mathbf{x}) = A$  για κάθε  $\mathbf{x}$ . Συνεπώς, η αναγκαία συνθήκη ελαχίστου,  $\nabla \phi(\mathbf{x}) = \mathbf{0}$ , γίνεται

$$A\mathbf{x} = \mathbf{b}$$

και αυτό συνιστά επίσης ικανή συνθήκη μοναδικού ελαχίστου.

Η μέθοδος Newton γίνεται τετριμμένη εδώ:

$$\mathbf{x}^* = \mathbf{x}_1 = \mathbf{x}_0 + A^{-1}(\mathbf{b} - A\mathbf{x}_0) = A^{-1}\mathbf{b}$$

Συνεπώς, η λύση προκύπτει μετά από μία επανάληψη για οποιαδήποτε αρχική εικασία. Η ζωή αποκτά και πάλι ενδιαφέρον μόνο όταν σκεφτούμε κάποια περίπτωση όπου δεν ενδείκνυται η άμεση επίλυση του συστήματος  $A\mathbf{x} = \mathbf{b}$  και απαιτούνται επαναληπτικές μέθοδοι όπως αυτές που περιγράφονται στο Κεφάλαιο 7. Απλούστερες πλέον μέθοδοι βελτιστοποίησης λειτουργούν ως επαναληπτικοί επιλυτές για το γραμμικό σύστημα εξισώσεων.

Η Ενότητα 7.4 αποκτά ιδιαίτερη συνάφεια εδώ, επειδή έχουμε χρησιμοποιήσει την ίδια αντικειμενική συνάρτηση  $\phi(\mathbf{x})$  για να καταλήξουμε στη μέθοδο συζυγούς κλίσης (CG) και τη μέθοδο καθοδικής κλίσης. Τόσο η μέθοδος απότομης καθόδου (steepest descent, SD) όσο και η μέθοδος CG μπορούν να γραφούν ως

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k, \quad \alpha_k = \frac{\langle \mathbf{r}_k, \mathbf{r}_k \rangle}{\langle \mathbf{p}_k, A\mathbf{p}_k \rangle}$$

Αυτό είναι το αποτέλεσμα μιας **ακριβούς ευθύγραμμης αναζήτησης** (exact line search). Η μέθοδος SD αποτελεί συγκεκριμένη περίπτωση της καθοδικής κλίσης, δηλαδή  $\mathbf{p}_k = \mathbf{r}_k$ , όπου  $\mathbf{r}_k = -\nabla\phi(\mathbf{x}_k) = \mathbf{b} - A\mathbf{x}_k$ , ενώ για τη μέθοδο CG υπάρχει πιο πολύπλοκος μαθηματικός τύπος για το  $\mathbf{p}_k$ .

Για μια μέθοδο καθοδικής κλίσης, όμως, μπορούμε να επιλέξουμε το μέγεθος βήματος  $\alpha_k$  με διαφορετικούς τρόπους, και όχι μόνο με τη μέθοδο SD. Για παράδειγμα, θεωρήστε το ενδεχόμενο «καθυστέρησης» του μεγέθους βήματος SD, χρησιμοποιώντας στην  $k$ -οστή επανάληψη το

$$\alpha_k = \frac{\langle \mathbf{r}_{k-1}, \mathbf{r}_{k-1} \rangle}{\langle \mathbf{r}_{k-1}, A\mathbf{r}_{k-1} \rangle}$$

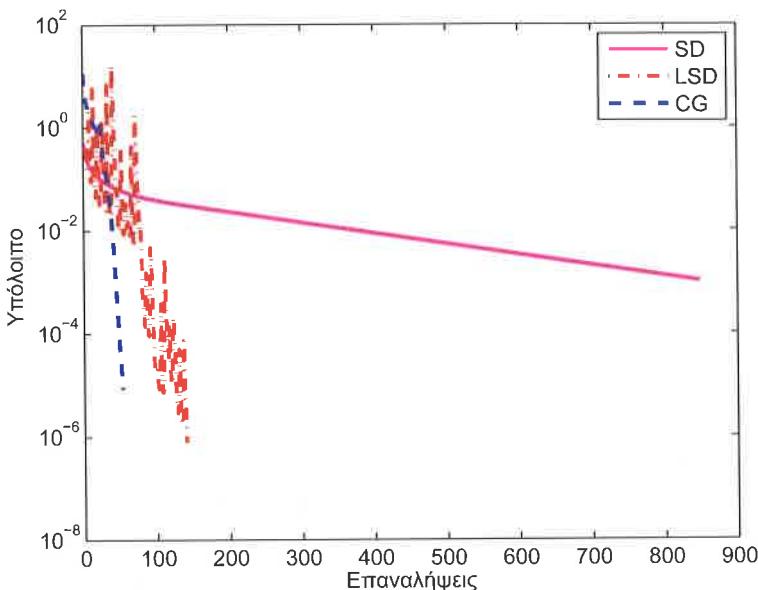
Θα αναφερόμαστε σε αυτή τη μέθοδο ως μέθοδο καθυστερημένης απότομης καθόδου (lagged steepest descent, LSD). δείτε επίσης την Άσκηση 7.14.

Το Σχήμα 9.6 είναι συγκρίσιμο με το Σχήμα 7.5. Δείχνει τη συμπεριφορά σύγκλισης για το ίδιο πρόβλημα Poisson όπως στα Παραδείγματα 7.1 και 7.10 με  $n = 31^2 = 961$  και  $\kappa(A) \sim n$ . Η μέθοδος SD είναι αργή, απαιτώντας  $O(n)$  επαναλήψεις. Η μέθοδος CG συγκλίνει με οργανωμένο τρόπο, απαιτώντας  $O(\sqrt{n})$  επαναλήψεις. Παραδόξως, η παραλλαγή LSD της καθοδικής κλίσης συγκλίνει πολύ πιο γρήγορα από ότι η μέθοδος SD, παρότι η σύγκλιση είναι και λίγο πιο αργή και σημαντικά πιο απρόβλεπτη συγκριτικά με τη μέθοδο CG.

Οι πρώτες επαναλήψεις της μεθόδου SD (έστω 10) μπορούν όντως να είναι αρκετά αποτελεσματικές σε κάποιες εφαρμογές. Μόνο όταν εκτελούνται πολλά τέτοια βήματα μειώνεται σημαντικά η αποδοτικότητα, με τη μέθοδο να ακολουθεί ένα μοτίβο όπου οι επιπλέον επαναλήψεις παύουν να προσθέτουν ιδιαίτερο όφελος. ■

## Μη ακριβής μέθοδος Newton

Για μεγάλα προβλήματα με αραιές εσσιανές μήτρες μπορεί να ενδείκνυται μια επαναληπτική μέθοδος για την εύρεση της κατεύθυνσης  $\mathbf{p}_k$  σε κάθε επανάληψη. Η ενσωμάτωση μιας τέτοιας επιλογής οδηγεί σε μια μέθοδο με έναν εσωτερικό και έναν εξωτερικό βρόχο.



**Σχήμα 9.6:** Σύγκλιση των επαναληπτικών μεθόδων καθοδικής και συζυγούς κλίσης για την εξίσωση Poisson του Παραδείγματος 7.1.

Ο εξωτερικός βρόχος αφορά μια επανάληψη που μοιάζει με τη μέθοδο Newton: Μια δημοφιλής επιλογή είναι η μήτρα  $B_k = \nabla^2 \phi(\mathbf{x}_k) + \mu_k I$  όπως αναφέραμε νωρίτερα, με την ανισότητα  $\mu_k \geq 0$  να εγγυάται ότι είναι τουλάχιστον θετικά ημιορισμένη και ότι διατηρεί την αραιότητά της.

Ο εσωτερικός βρόχος αφορά μια μέθοδο επίλυσης γραμμικών συστημάτων όπως εκείνες που εξετάσαμε στην Ενότητα 7.4 και στο Παράδειγμα 9.7. Η μέθοδος PCG αποτελεί ιδιαίτερα δημοφιλή επιλογή εδώ. Ο συνδυασμός συνήθως παράγει μια **μη ακριβή μέθοδο Newton** (inexact Newton method), όπου η επαναληπτική μέθοδος για το γραμμικό πρόβλημα (η εσωτερική επανάληψη) τερματίζεται νωρίς. Ωστόσο η ακρίβεια του γραμμικού επαναληπτικού επιλυτή αυξάνεται, δηλαδή η ανοχή σύγκλισης μειώνεται, αν η λύση της εξωτερικής επανάληψης βρίσκεται ήδη πολύ κοντά σε μια ρίζα  $\mathbf{x}^*$ , ώστε να επιτευχθεί γρήγορη τοπική σύγκλιση της εξωτερικής επανάληψης. Σε αυτή την κατηγορία εμπίπτουν μερικές πολύ δημοφιλείς μέθοδοι που χρησιμοποιούνται σε εφαρμογές.

## Μέθοδοι οιονεί Newton

Φυσικά, οι δυσκολίες κατά τον υπολογισμό παραγώγων, που ήδη έκαναν την εμφάνισή τους στην Ενότητα 9.1 επειδή εκεί έπρεπε να υπολογίσουμε μια ιακωβιανή μήτρα, γίνονται ακόμα πιο εμφανείς εδώ επειδή πρέπει να υπολογίσουμε και πρώτες και δεύτερες παραγώγους. Επιπλέον, η επέκταση της μεθόδου τέμνουνσας, την οποία περιγράψαμε στην Ενότητα 3.4, γίνεται πιο περίπλοκη και πιο ενδιαφέρουσα επειδή εδώ έχουμε την ευκαιρία να επιμείνουμε στο να διατηρηθεί η μήτρα  $B_k$  θετικά ορισμένη, εξασφαλίζοντας έτσι κατευθύνσεις καθόδου καθ' όλη τη διάρκεια της διαδικασίας των επαναλήψεων.

Επομένως, στην  $k$ -οστή επανάληψη η προσέγγιση  $B_k$  της εσσιανής μήτρας ανανεώνεται προκειμένου να σχηματιστεί η επόμενη προσέγγιση  $B_{k+1}$ . Αυτές οι μέθοδοι όχι μόνο παρακάμπτουν την ανάγκη για τον ρητό υπολογισμό της  $\nabla^2\phi(\mathbf{x})$ , αλλά συχνά (αν και όχι πάντα) είναι και λιγότερο δαπανηρές στη χρήση συγκριτικά με άλλες παραλλαγές της μεθόδου Newton. Επιπλέον, υπό κατάλληλες συνθήκες εγγυώνται τη σύγκλιση σε ένα σημείο  $\mathbf{x}^*$  που ικανοποιεί, με μια ανοχή σφάλματος, την ισότητα  $\nabla\phi(\mathbf{x}^*) = \mathbf{0}$ .

Προσέξτε ότι από το ανάπτυγμα Taylor για την κλίση έχουμε

$$\nabla\phi(\mathbf{x}_k) \approx \nabla\phi(\mathbf{x}_{k+1}) - \nabla^2\phi(\mathbf{x}_{k+1})(\mathbf{x}_{k+1} - \mathbf{x}_k)$$

Η ουσιαστική δράση της εσσιανής μήτρας φαίνεται να είναι στην κατεύθυνση του  $\mathbf{w}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ . Απαιτούμε να ισχύει

$$B_{k+1}\mathbf{w}_k = \mathbf{y}_k, \quad \mathbf{y}_k = \nabla\phi(\mathbf{x}_{k+1}) - \nabla\phi(\mathbf{x}_k)$$

ώστε αυτή η δράση να αναπαράγεται από την προσέγγιση της εσσιανής μήτρας παρότι, κατά τα άλλα, η  $B_{k+1}$  και η  $\nabla^2\phi(\mathbf{x}_{k+1})$  μπορεί να μην βρίσκονται ιδιαίτερα κοντά η μία στην άλλη.

## Η μέθοδος BFGS

Η πιο δημοφιλής παραλλαγή αυτών των μεθόδων, η ανανέωση Broyden-Fletcher-Goldfarb-Shanno, που είναι κοινώς γνωστή με την ονομασία μέθοδος BFGS, παρουσιάζεται στην επόμενη σελίδα. Συνιστά τη ραχοκοκαλιά πολλών πακέτων λογισμικού βελτιστοποίησης γενικής χρήσης. Αντί να ανανεώνει τη μήτρα  $B_k$ , ανανεώνει απευθείας την αντίστροφή της,  $G_k = B_k^{-1}$ .

Η ταχύτητα τοπικής σύγκλισης αυτής της μεθόδου είναι μόνο υπεργραμμική, και όχι τετραγωνική, αλλά αυτός δεν είναι σημαντικός λόγος ανησυχίας. Δυστυχώς, για εσσιανές μήτρες κακής κατάστασης καθώς και για μεγάλες, αραιές εσσιανές μήτρες, η μέθοδος BFGS ενδέχεται να χάνει κάποια από τα πλεονεκτήματά της. Στην

περίπτωση των αραιών εσσιανών μητρών υπάρχουν δημοφιλείς παραλλαγές περιορισμένης μνήμης (limited memory) της μεθόδου BFGS, που είναι γνωστές ως L-BFGS. Η περιγραφή τους ξεφεύγει από το αντικείμενο αυτού του βιβλίου. Σε κάθε περίπτωση, στο προσκήνιο επανεμφανίζονται άλλες παραλλαγές της μεθόδου Newton.

## Μη γραμμικά ελάχιστα τετράγωνα και η μέθοδος Gauss-Newton

Ας επανέλθουμε στο θεμελιώδες πρόβλημα της προσαρμογής δεδομένων που περιγράψαμε στις εισαγωγικές παραγράφους του Κεφαλαίου 6 και δείξαμε στα Παραδείγματα 8.1 και 8.2. Έστω ότι δίνονται τα παρατηρούμενα δεδομένα  $\mathbf{b}$  και μια συνάρτηση-μοντέλο  $\mathbf{g}(\mathbf{x}) = \mathbf{Ax}$  η οποία προβλέπει τα δεδομένα για κάθε  $\mathbf{x}$ : το πρόβλημα είναι να βρούμε το βέλτιστο  $\mathbf{x}$  υπό την έννοια ότι τα προβλεπόμενα δεδομένα συμφωνούν στον μεγαλύτερο δυνατό βαθμό με τα παρατηρούμενα.

Υπάρχουν πολλά πρακτικά παραδείγματα, ωστόσο, όπου η συνάρτηση εξαρτάται μη γραμμικά από το διάνυσμα ορισμάτων  $\mathbf{x}$ . Τότε οδηγούμαστε στο πρόβλημα μη γραμμικής βελτιστοποίησης χωρίς περιορισμούς

$$\min_{\mathbf{x}} \|\mathbf{g}(\mathbf{x}) - \mathbf{b}\|$$

**Αλγόριθμος: Η επανάληψη BFGS.**

Επίλεξε τα  $\mathbf{x}_0$  και  $G_0$  (π.χ.  $G_0 = I$ )

for  $k = 0, 1, \dots$  μέχρι τη σύγκλιση

$$\mathbf{p}_k = -G_k \nabla \phi(\mathbf{x}_k)$$

Βρες κατάλληλο μέγεθος βήματος  $\alpha_k$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$$

$$\mathbf{w}_k = \alpha_k \mathbf{p}_k$$

$$\mathbf{y}_k = \nabla \phi(\mathbf{x}_{k+1}) - \nabla \phi(\mathbf{x}_k)$$

$$G_{k+1} = \left( I - \frac{\mathbf{w}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{w}_k} \right) G_k \left( I - \frac{\mathbf{y}_k \mathbf{w}_k^T}{\mathbf{y}_k^T \mathbf{w}_k} \right) + \frac{\mathbf{w}_k \mathbf{w}_k^T}{\mathbf{y}_k^T \mathbf{w}_k}.$$

end

όπου η  $\mathbf{g} \in C^2$  έχει  $m$  στοιχεία,  $\mathbf{g} : \mathcal{R}^n \rightarrow \mathcal{R}^m$ ,  $m \geq n$ , και η νόρμα είναι εξ ορισμού η 2-νόρμα. Προφανώς, το τυπικό πρόβλημα που μελετήσαμε στο Κεφάλαιο 6 αποτελεί ειδική περίπτωση του παρόντος προβλήματος μη γραμμικών ελάχιστων τετραγώνων.

Όπως και στην Ενότητα 6.1, είναι πιο βολικό να εξετάσουμε το πρόβλημα στη μορφή

$$\min_{\mathbf{x}} \phi(\mathbf{x}) = \frac{1}{2} \|\mathbf{g}(\mathbf{x}) - \mathbf{b}\|^2$$

Με την ιακωβιανή μήτρα  $\mathbf{g}$  να ορίζεται ως

$$A(\mathbf{x}) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_{m-1}}{\partial x_1} & \frac{\partial g_{m-1}}{\partial x_2} & \cdots & \frac{\partial g_{m-1}}{\partial x_n} \\ \frac{\partial g_m}{\partial x_1} & \frac{\partial g_m}{\partial x_2} & \cdots & \frac{\partial g_m}{\partial x_n} \end{pmatrix}$$

οι αναγκαίες συνθήκες ελαχίστου δίνονται από τη σχέση

$$\nabla \phi(\mathbf{x}^*) = A(\mathbf{x}^*)^T (\mathbf{g}(\mathbf{x}^*) - \mathbf{b}) = \mathbf{0}$$

(Στην Άσκηση 14 θα επαληθεύσετε ότι αυτό όντως ισχύει.) Η παράσταση που προκύπτει είναι μια προφανής γενίκευση των κανονικών εξισώσεων (normal equations)· πράγματι, υποθέτουμε επίσης –όπως και στο Κεφάλαιο 6– ότι η ιακωβιανή μήτρα είναι πλήρους τάξης στηλών, τουλάχιστον σε μια γειτονιά του  $\mathbf{x}^*$ , ώστε η μήτρα  $A^T A$  να είναι συμμετρική και θετικά ορισμένη. Εδώ, όμως, εξακολούθουμε να πρέπει να λύσουμε ένα μη γραμμικό σύστημα  $n$  εξισώσεων.

Με ένα άλλο χρήσιμο τέχνασμα του λογισμού προκύπτει η εσσιανή μήτρα

$$\nabla^2 \phi(\mathbf{x}) = A(\mathbf{x})^T A(\mathbf{x}) + L(\mathbf{x})$$

όπου  $L$  είναι μια μήτρα διαστάσεων  $n \times n$  με στοιχεία

$$L_{i,j} = \sum_{l=1}^m \frac{\partial^2 g_k}{\partial x_i \partial x_j} (g_l - b_l)$$

## Η μέθοδος Gauss-Newton

Εστιάζοντας την προσοχή μας στις αριθμητικές μεθόδους, μπορούμε να καταλήξουμε σε μια κλασική μέθοδο ως εξής. Όπως συνήθως, ξεκινάμε με μια αρχική εικασία  $\mathbf{x}_0$  και εξετάζουμε στην  $k$ -οστή επανάληψη το ερώτημα της εύρεσης της επόμενης προσεγγιστικής τιμής,  $\mathbf{x}_{k+1}$ , με δεδομένη την τρέχουσα προσεγγιστική τιμή,  $\mathbf{x}_k$ . Είναι φυσικό να προσεγγίσουμε την  $\mathbf{g}(\mathbf{x}_{k+1})$  με την  $\mathbf{g}(\mathbf{x}_k) + A(\mathbf{x}_k)\mathbf{p}_k$ , όπως και στη μέθοδο

Newton. Συνεπώς, επιλέγουμε για το διάνυσμα διόρθωσης το γραμμικό πρόβλημα ελάχιστων τετραγώνων

$$\min_{\mathbf{p}} \|A(\mathbf{x}_k)\mathbf{p} - (\mathbf{b} - \mathbf{g}(\mathbf{x}_k))\|$$

συμβολίζοντας τον ελαχιστοποιητή με  $\mathbf{p}_k$  και θέτοντας

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$$

Αυτή είναι η επανάληψη *Gauss-Newton*. Παρατηρώντας ότι το  $\mathbf{b} - \mathbf{g}(\mathbf{x}_k)$  είναι το διάνυσμα υπολοίπου για την τρέχουσα προσεγγιστική τιμή, μπορούμε να βρούμε το επόμενο διάνυσμα κατεύθυνσης προσαρμόζοντας αυτό το υπόλοιπο από τον χώρο στηλών της τρέχουσας ιακωβιανής μήτρας  $A(\mathbf{x}_k)$ . Οι κανονικές εξισώσεις που είναι μαθηματικά ισοδύναμες με τη διατύπωση γραμμικών ελάχιστων τετραγώνων της  $k$ -οστίς επανάληψης είναι

$$A(\mathbf{x}_k)^T A(\mathbf{x}_k) \mathbf{p}_k = A(\mathbf{x}_k)^T (\mathbf{b} - \mathbf{g}(\mathbf{x}_k))$$

**Αλγόριθμος: Μέθοδος Gauss-Newton για ελάχιστα τετράγωνα.**

for  $k = 0, 1, \dots$ , μέχρι τη σύγκλιση

1. Λύσε το γραμμικό πρόβλημα ελάχιστων τετραγώνων

$$\min_{\mathbf{p}} \|A(\mathbf{x}_k)\mathbf{p} - (\mathbf{b} - \mathbf{g}(\mathbf{x}_k))\|$$

με τον ελαχιστοποιητή  $\mathbf{p} = \mathbf{p}_k$ .

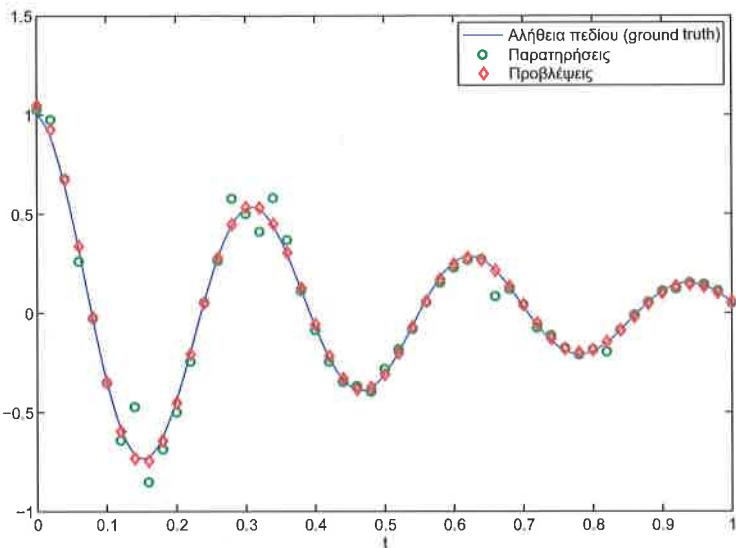
2. Θέσε

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$$

end

**Παράδειγμα 9.8.** Έστω ότι παράγουμε δεδομένα χρησιμοποιώντας τη συνάρτηση  $u(t) = e^{-2t} \cos(20t)$ , που απεικονίζεται στο Σχήμα 9.7 ως συνεχής μπλε καμπύλη. Στα 51 σημεία  $t_i = 0.02(i - 1)$ ,  $i = 1, \dots, 51$ , προσθέτουμε 20% τυχαίο θόρυβο στη  $u(t_i)$  για να παραγάγουμε τα δεδομένα  $b_i$ , τα οποία εμφανίζονται ως πράσινοι κύκλοι στο Σχήμα 9.7. Στη συνέχεια προσποιούμαστε ότι δεν έχουμε δει ποτέ την μπλε καμπύλη και προσπαθούμε να προσαρμόσουμε σε αυτό το διάνυσμα  $\mathbf{b}$  μια συνάρτηση της μορφής

$$v(t) = x_1 e^{x_2 t} \cos(x_3 t)$$



**Σχήμα 9.7:** Μη γραμμική προσαρμογή δεδομένων (δείτε το Παράδειγμα 9.8).

Με βάση τον παραπάνω συμβολισμό έχουμε  $m = 51$ ,  $n = 3$ , και οι  $i$ -οστές γραμμές των  $\mathbf{g}$  και  $A$  ορίζονται ως

$$\begin{aligned} g_i &= x_1 e^{x_2 t_i} \cos(x_3 t_i), \quad a_{i,1} = e^{x_2 t_i} \cos(x_3 t_i) \\ a_{i,2} &= t_i g_i \quad a_{i,3} = -t_i x_1 e^{x_2 t_i} \sin(x_3 t_i), \quad 1 \leq i \leq m \end{aligned}$$

Εκτελώντας τη μέθοδο Gauss-Newton με αφετηρία το  $\mathbf{x}_0 = (1.2, -1.9, 18)^T$ , παίρνουμε μετά από 7 επαναλήψεις τα εξής:  $\|\mathbf{p}_k\| < 10^{-7}$ ,  $\mathbf{x} \approx (1.0471, -2.1068, 19.9588)^T$  και  $\|\mathbf{g} - \mathbf{b}\| \approx 0.42$ . Τα προβλεπόμενα δεδομένα εμφανίζονται ως κόκκινοι ρόμβοι στο Σχήμα 9.7. Αυτό είναι ένα πολύ ικανοποιητικό αποτέλεσμα.

Με το επίπεδο του θορύβου στο 100% παίρνουμε μετά από 15 επαναλήψεις το αποτέλεσμα  $\|\mathbf{g} - \mathbf{b}\| \approx 1.92$  για  $\mathbf{x} \approx (1.0814, -3.0291, 19.8583)^T$ . Προσέξτε μια περίεργη ιδιότητα: Όταν το μοντέλο προσεγγίζει τα δεδομένα λιγότερο καλά, απαιτούνται περισσότερες επαναλήψεις.

Όμως η χρήση της πιο γενικής αρχικής εικασίας  $\mathbf{x}_0 = (1.2, -1.9, 10)^T$  οδηγεί γρήγορα σε μια άλλη λύση γι' αυτό το ευαίσθητο πρόβλημα η οποία δεν προσεγγίζει καλά τα δεδομένα. Ο μη γραμμικός κόσμος είναι γεμάτος εκπλήξεις, ένα γεγονός που δεν πρέπει να υποτιμάει κανείς: Απαιτείται πάντα –και χωρίς καμία εξαίρεση– η προσεκτική αξιολόγηση οποιουδήποτε υπολογισμένου αποτελέσματος. ■

## Η μέθοδος Gauss-Newton και η μέθοδος Newton

Επισημαίνουμε ότι η μέθοδος Gauss-Newton διαφέρει από τη μέθοδο Newton για την ελαχιστοποίηση της  $\phi(\mathbf{x})$ . Πιο συγκεκριμένα, η μέθοδος Newton ορίζεται ως

$$\nabla^2 \phi(\mathbf{x}_k) \mathbf{p}_k = -\nabla \phi(\mathbf{x}_k) = A(\mathbf{x}_k)^T (\mathbf{b} - \mathbf{g}(\mathbf{x}_k))$$

άρα το δεξιό μέλος είναι το ίδιο όπως και για τις κανονικές εξισώσεις της επανάληψης της μεθόδου Gauss-Newton. Όμως η μήτρα  $\nabla^2 \phi(\mathbf{x}_k)$  περιέχει τον επιπλέον όρο  $L$  ο οποίος αγνοείται στην επανάληψη της μεθόδου Gauss-Newton. Αυτό μας οδηγεί σε πολλά ενδιαφέρονται συμπεράσματα:

- Η κατεύθυνση της μεθόδου Gauss-Newton, σε αντίθεση με τη μέθοδο Newton, αποτελεί εγγυημένα κατεύθυνση καθόδου ως προς τη  $\phi$ . Αυτό οφείλεται στο γεγονός ότι η μήτρα  $A^T A$  είναι συμμετρική και θετικά ορισμένη ακόμα και όταν αυτό δεν ισχύει για τη μήτρα  $A^T A + L$ .
- Η επανάληψη της μεθόδου Gauss-Newton είναι λιγότερο δαπανηρή και ενδέχεται να είναι καλύτερης κατάστασης (πιο ευσταθής) από ότι η επανάληψη της μεθόδου Newton.
- Η τάξη σύγκλισης της μεθόδου Gauss-Newton γενικά δεν είναι εγγυημένα τετραγωνική, επειδή η διαφορά ανάμεσα σε αυτήν και στην επανάληψη της μεθόδου Newton δεν μηδενίζεται στο όριο.
- Η μέθοδος Gauss-Newton συγκλίνει πιο γρήγορα για προβλήματα όπου το μοντέλο προσαρμόζεται καλά στα δεδομένα. Αυτό συμβαίνει επειδή, σε μια τέτοια περίπτωση, η τιμή  $\|\mathbf{g}(\mathbf{x}) - \mathbf{b}\|$  είναι «μικρή» κοντά στη λύση, και άρα η  $\|L\|$  είναι μικρή, οπότε η μέθοδος Gauss-Newton πλησιάζει περισσότερο στη μέθοδο Newton.

Ασκήσεις γι' αυτή την ενότητα: 10–18.

## 9.3 \*Βελτιστοποίηση με περιορισμούς

Το γενικό πρόβλημα βελτιστοποίησης με περιορισμούς που εξετάζουμε εδώ είναι, όπως και προηγουμένως, η ελαχιστοποίηση μιας βαθμωτής συνάρτησης  $\phi(\mathbf{x})$  ως προς  $n$  μεταβλητές. Η διαφορά είναι ότι τώρα υπάρχουν *ισοτικοί* και *ανισοτικοί* περιορισμοί (equality and inequality constraints) τους οποίους πρέπει να ικανοποιεί οποιοδήποτε κατάλληλο  $\mathbf{x}$ . Πολλά προβλήματα βελτιστοποίησης σε εφαρμογές εμφανίζονται με αυτή τη μορφή, γι' αυτό και θα διερευνήσουμε τεχνικές και μεθόδους για την επίλυσή τους.

## Περιορισμοί

Το γενικό πρόβλημα γράφεται ως

$$\min_{\mathbf{x} \in \Omega} \phi(\mathbf{x}), \quad \text{όπου}$$

$$\Omega = \{\mathbf{x} \in \mathcal{R}^n \mid c_i(\mathbf{x}) = 0, i \in \mathcal{E}, \quad c_i(\mathbf{x}) \geq 0, i \in \mathcal{I}\}$$

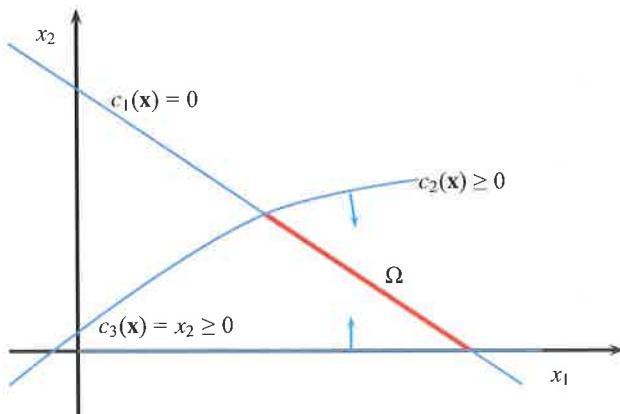
Άρα, το  $\mathcal{E}$  είναι το σύνολο των ισοτικών περιορισμών και το  $\mathcal{I}$  είναι το σύνολο των ανισοτικών περιορισμών. Θα υποθέσουμε ότι το  $c_i(\mathbf{x}) \in C^1$ , για κάθε  $i$ . Οποιοδήποτε σημείο  $\mathbf{x} \in \Omega$  είναι μια *εφικτή λύση* (feasible solution) –σε αντιδιαστολή με τη βέλτιστη λύση. Ας δούμε αυτόν τον συμβολισμό στην πράξη με ένα απλό παράδειγμα.

### Παράδειγμα 9.9.

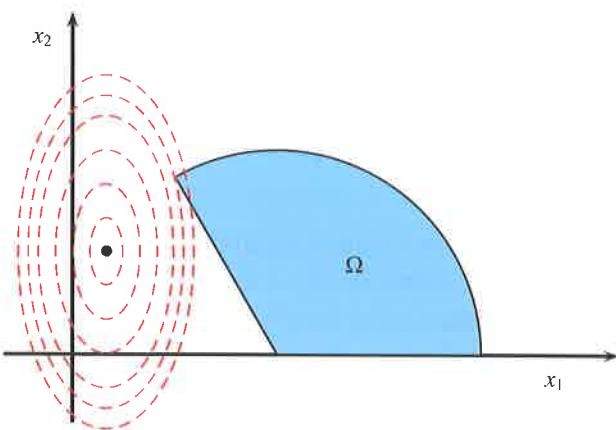
1. Θεωρήστε το σύνολο που απεικονίζεται στο Σχήμα 9.8. Εδώ,  $\mathcal{E} = \{1\}$ ,  $\mathcal{I} = \{2, 3\}$ . Το σύνολο  $\Omega$  αποτελείται από το τμήμα της ευθείας  $c_1 = 0$  το οποίο βρίσκεται ανάμεσα στην καμπύλη  $c_2 = 0$  και στον άξονα  $x_1$ .
2. Συχνά, το  $\mathcal{E}$  είναι κενό. Τότε, το  $\Omega$  μπορεί να έχει μη κενό εσωτερικό όπως στο Σχήμα 9.9.

Η βασική παραπάνω διαφορά στις μορφές του  $\Omega$  μπορεί κάλλιστα να αντικατοπτρίζεται στους αντίστοιχους αλγόριθμους. ■

Ένα πολύ σύνηθες παράδειγμα περιορισμών είναι όταν κάποιες μεταβλητές στο  $\mathbf{x}$  πρέπει να είναι μη αρνητικές επειδή αντιστοιχούν σε φυσικά μεγέθη όπως η αγωγιμότητα, ή σε ένα κοινό αγαθό όπως η ποσότητα των αλεύρων που έχει απομείνει στην αποθήκη ενός αρτοποιείου, τα οποία δεν μπορούν να πάρουν αρνητικές τιμές.



**Σχήμα 9.8:** Ισοτικοί και ανισοτικοί περιορισμοί. Το εφικτό σύνολο  $\Omega$  αποτελείται από τα σημεία πάνω στην έντονη κόκκινη γραμμή.



**Σχήμα 9.9:** Ένα εφικτό σύνολο  $\Omega$  με μη κενό εσωτερικό, και ισοϋψείς της  $\phi$ : οι μεγαλύτερες ελλείψεις αντιστοιχούν σε μεγαλύτερες τιμές της  $\phi$ .

Αν το ελάχιστο χωρίς περιορισμούς της  $\phi$  ανήκει στο  $\Omega$ , το πρόβλημα είναι ουσιαστικά το πρόβλημα ελαχιστοποίησης χωρίς περιορισμούς. Επομένως, για να διαφοροποιήσουμε τα πράγματα και να τα κάνουμε πιο ενδιαφέροντα, υποθέτουμε ότι ο ελαχιστοποιητής χωρίς περιορισμούς της  $\phi$  δεν ανήκει στο  $\Omega$ , όπως στο Σχήμα 9.9. Στο συγκεκριμένο σχήμα και άλλού, είναι βολικό να σχεδιάζονται ισοϋψείς της αντικειμενικής συνάρτησης  $\phi$ . Μια **ισοϋψής** (level set) της  $\phi$  αποτελείται από όλα τα σημεία  $\mathbf{x}$  στα οποία η  $\phi(\mathbf{x})$  έχει την ίδια τιμή. Οι ομόκεντρες ελλείψεις στο Σχήμα 9.9 αντιστοιχούν σε ισοϋψείς για διαφορετικές τιμές της  $\phi$ . Η κλίση της  $\phi$  σε κάποιο σημείο  $\mathbf{x}$  σχηματίζει ορθή γωνία με την ισοϋψή που διέρχεται από το συγκεκριμένο σημείο.

Ορίζουμε σε κάθε σημείο  $\mathbf{x} \in \mathcal{R}^n$  το ενεργό σύνολο (active set)

$$\mathcal{A}(\mathbf{x}) = \mathcal{E} \cup \{i \in \mathcal{I} \mid c_i(\mathbf{x}) = 0\}$$

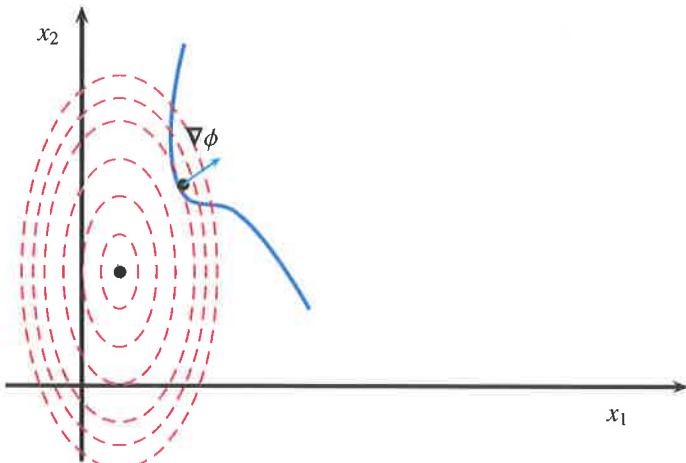
Κατόπιν εξετάζουμε προβλήματα όπου το  $\mathcal{A}(\mathbf{x}^*)$  δεν είναι κενό.

Θα ξεκινήσουμε τη σύντομη περιγραφή μας με τις αναγκαίες συνθήκες για την ύπαρξη βελτίστου. Στη συνέχεια θα μελετήσουμε διάφορες οικογένειες αλγόριθμων που χρησιμοποιούνται στην πράξη για βελτιστοποίηση με περιορισμούς. Τέλος, θα αναφερθούμε με περισσότερες λεπτομέρειες σε έναν πολύ χρήσιμο αλγόριθμο για το πρόβλημα του γραμμικού προγραμματισμού.

## Συνθήκες ελαχίστου με περιορισμούς

Όπως και στην περίπτωση χωρίς περιορισμούς, υπάρχουν αναγκαίες συνθήκες πρώτης τάξης για ένα κρίσιμο σημείο, και υπάρχουν αναγκαίες και ικανές συνθήκες δεύτερης τάξης για ένα τοπικό ελάχιστο. Όμως όλες αυτές οι συνθήκες είναι σημαντικά

πιο περίπλοκες από ό,τι όταν δεν υπάρχουν περιορισμοί. Θα ξεκινήσουμε με τις αναγκαίες συνθήκες πρώτης τάξης, επεκτείνοντας την απαίτηση που ορίσαμε στη βελτιστοποίηση χωρίς περιορισμούς, δηλαδή να μηδενίζεται η  $\nabla\phi(\mathbf{x}^*)$ . Ας δούμε πρώτα ένα παράδειγμα που παρέχει τα απαραίτητα κίνητρα.



**Σχήμα 9.10:** Ένας ισοτικός περιορισμός και οι ισοϋψεις της  $\phi$ . Στο  $\mathbf{x}^*$  η κλίση σχηματίζει ορθή γωνία με την εφαπτομένη του περιορισμού.

**Παράδειγμα 9.10.** Θα ξεκινήσουμε με έναν ισοτικό περιορισμό στο  $\mathcal{R}^2$ :  $c(x_1, x_2) = 0$  (δείτε το Σχήμα 9.10). Σε οποιοδήποτε σημείο  $\mathbf{x}$  η κλίση  $\nabla\phi(\mathbf{x})$  σχηματίζει ορθή γωνία με την εφαπτομένη της ισοϋψούς στο συγκεκριμένο σημείο. Στο σημείο ελαχίστου  $\mathbf{x}^*$  ο περιορισμός και η ισοϋψής για τη  $\phi(\mathbf{x}^*)$  έχουν την ίδια κατεύθυνση εφαπτομένης, άρα η  $\nabla\phi(\mathbf{x}^*)$  είναι παράλληλη με τη  $\nabla c(\mathbf{x}^*)$ . Αυτό σημαίνει ότι υπάρχει σταθερά αναλογίας  $\lambda^*$  τέτοια ώστε να ισχύει

$$\nabla\phi(\mathbf{x}^*) = \lambda^* \nabla c(\mathbf{x}^*)$$

Ας υποθέσουμε τώρα ότι υπάρχει μόνο ένας ανισοτικός περιορισμός,  $c(x_1, x_2) \geq 0$ .

Η εφικτή περιοχή  $\Omega$  βρίσκεται δεξιά από τη συνεχή μπλε καμπύλη στο Σχήμα 9.10. Άρα, η  $\nabla\phi(\mathbf{x}^*)$  όχι μόνο είναι παράλληλη με τη  $\nabla c(\mathbf{x}^*)$ , αλλά πρέπει επίσης να δείχνει στο εσωτερικό του εφικτού συνόλου, κάτι που σημαίνει ότι το  $\lambda^*$  δεν μπορεί να είναι αρνητικό. Κατά συνέπεια, προκύπτει ότι

$$\nabla\phi(\mathbf{x}^*) = \lambda^* \nabla c(\mathbf{x}^*), \quad \lambda^* \geq 0$$

Αυτή η τιμή του  $\lambda^*$  ονομάζεται πολλαπλασιαστής Lagrange (Lagrange multiplier).

Η γενική περίπτωση δεν αποτελεί άμεση επέκταση του Παραδείγματος 9.10, αλλά θα παραλείψουμε την πλήρη απόδειξη εδώ και απλώς θα τη διατυπώσουμε. Πρώτα, όμως, πρέπει να διατυπώσουμε την υπόθεση της εξειδίκευσης περιορισμών (constraint qualification).

### Εξειδίκευση περιορισμών.

Έστω ότι  $\mathbf{x}^*$  είναι ένα τοπικό κρίσιμο σημείο, και έστω ότι με  $\nabla c_i(\mathbf{x})$  συμβολίζονται οι κλίσεις των περιορισμών. Έστω επίσης ότι  $A_*^T$  είναι η μήτρα της οποίας οι στήλες είναι οι κλίσεις  $\nabla c_i(\mathbf{x}^*)$  όλων των ενεργών περιορισμών, δηλαδή εκείνων που ανήκουν στο  $\mathcal{A}(\mathbf{x}^*)$ .

Η υπόθεση της εξειδίκευσης περιορισμών είναι ότι η μήτρα  $A_*^T$  έχει πλήρη τάξη στηλών.

Στη συνέχεια ορίζουμε τη **συνάρτηση Lagrange** (ή Lagrangian)

$$\mathcal{L}(\mathbf{x}, \lambda) = \phi(\mathbf{x}) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(\mathbf{x})$$

Τότε, οι αναγκαίες συνθήκες πρώτης τάξης για την ύπαρξη ενός ελαχίστου είναι οι διάσημες συνθήκες Karush-Kuhn-Tucker (KKT), που παρουσιάζονται σε αυτή τη σελίδα.

Εφόσον ισχύει η συνθήκη εξειδίκευσης περιορισμών που ορίσαμε, υπάρχει ένα τοπικά μοναδικό  $\lambda^*$  το οποίο, μαζί με το  $\mathbf{x}^*$ , ικανοποιεί τις συνθήκες KKT. Επιπλέον, μπορεί να αποδειχθεί ότι χωρίς την εξειδίκευση περιορισμών αυτές οι συνθήκες δεν είναι και τόσο αναγκαίες.

### Θεώρημα: Συνθήκες ελαχιστοποίησης με περιορισμούς.

Υποθέτουμε ότι οι  $\phi(\mathbf{x})$  και  $c_i(\mathbf{x})$  είναι επαρκώς λείες κοντά σε ένα κρίσιμο σημείο  $\mathbf{x}^*$  και ότι ισχύει η συνθήκη εξειδίκευσης περιορισμών. Τότε, υπάρχει διάνυσμα πολλαπλασιαστών Lagrange  $\lambda^*$  τέτοιο ώστε να ισχύουν οι σχέσεις

$$\begin{aligned}\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \lambda^*) &= \mathbf{0} \\ c_i(\mathbf{x}^*) &= 0 \quad \forall i \in \mathcal{E} \\ c_i(\mathbf{x}^*) &\geq 0 \quad \forall i \in \mathcal{I} \\ \lambda_i^* &\geq 0 \quad \forall i \in \mathcal{I} \\ \lambda_i^* c_i(\mathbf{x}^*) &= 0 \quad \forall i \in \mathcal{E} \cup \mathcal{I}\end{aligned}$$

**Παράδειγμα 9.11.** Θέλουμε να ελαχιστοποιήσουμε μια τετραγωνική συνάρτηση με γραμμικούς ισοτικούς περιορισμούς. Το πρόβλημα ορίζεται ως

$$\begin{array}{ll} \min_{\mathbf{x}} & \phi(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T H \mathbf{x} - \mathbf{d}^T \mathbf{x} \\ \text{έτσι ώστε} & A \mathbf{x} = \mathbf{b} \end{array}$$

όπου η μήτρα  $A$  έχει διαστάσεις  $m \times n$ ,  $m \leq n$ . Η συνθήκη εξειδίκευσης περιορισμών «κμεταφράζεται» στην απαίτηση η μήτρα  $A$  να είναι πλήρους τάξης γραμμών.

Αυτό είναι ένα πρόβλημα **τετραγωνικού προγραμματισμού** (quadratic programming) με ισοτικούς περιορισμούς. Η συνάρτηση Lagrange είναι

$$\mathcal{L}(\mathbf{x}, \lambda) = \frac{1}{2} \mathbf{x}^T H \mathbf{x} - \mathbf{d}^T \mathbf{x} - \lambda^T (\mathbf{b} - A \mathbf{x})$$

και οι συνθήκες KKT είναι

$$\begin{aligned} H \mathbf{x} - \mathbf{d} + A^T \lambda &= \mathbf{0} \\ \mathbf{b} - A \mathbf{x} &= \mathbf{0} \end{aligned}$$

Το σύστημα γραμμικών εξισώσεων που προκύπτει, το οποίο είναι γνωστό και ως **σύστημα σαγματικού σημείου** (saddle point system), μπορεί να διαταχθεί ως εξής:

$$\begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbf{d} \\ \mathbf{b} \end{pmatrix}$$

Η μήτρα KKT, ή μήτρα σαγματικού σημείου,  $K = \begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix}$ , είναι μεν συμμετρική αλλά είναι μη ορισμένη ακόμα και αν η μήτρα  $H$  είναι θετικά ορισμένη. Επειδή η μήτρα  $A$  είναι πλήρους τάξης γραμμών, η  $K$  είναι μη ιδιάζουσα αν ισχύει  $\mathbf{y}^T H \mathbf{y} \neq 0$  για κάθε  $\mathbf{y} \in \text{null}(A)$ ,  $\mathbf{y} \neq \mathbf{0}$  (δείτε την Άσκηση 19).

Αν η  $K$  είναι μη ιδιάζουσα, τότε προφανώς υπάρχει ακριβώς ένα κρίσιμο σημείο  $\mathbf{x}^*$  το οποίο ικανοποιεί τις αναγκαίες συνθήκες. ■

Η κατάσταση περιπλέκεται ακόμα πιο πολύ, ακόμα και για προβλήματα τετραγωνικού προγραμματισμού, όταν υπάρχουν ανισοτικοί περιορισμοί.

Λαμβάνοντας υπόψη τι έχει συμβεί στην πολύ απλή συνθήκη πρώτης τάξης χωρίς περιορισμούς η οποία ορίζει ένα κρίσιμο σημείο, μπορείτε να φανταστείτε ότι οι συνθήκες δεύτερης τάξης για την περίπτωση με περιορισμούς (που περιλαμβάνουν την εσσιανή μήτρα της συνάρτησης Lagrange  $\mathcal{L}$  ως προς το  $\mathbf{x}$ ) δεν θα είναι απλές. Πράγματι δεν είναι, και γι' αυτό δεν θα τις συμπεριλάβουμε σε αυτή τη σύντομη παρουσίαση.

## Επισκόπηση αλγόριθμων

Σε γενικές γραμμές, οι ισοτικοί περιορισμοί έχουν μια πιο αλγεβρική αίσθηση [αναζήτηση σε μειωμένο χώρο, σε πολλαπλότητα περιορισμών (constraint manifold) κ.λπ.], ενώ οι ανισοτικοί περιορισμοί ενδέχεται να προσδώσουν στο πρόβλημα μια πρόσθετη συνδυαστική (combinatorial) αίσθηση –συγκεκριμένα, ποιοι περιορισμοί είναι ενεργοί στη λύση και ποιοι όχι. Είναι εφικτή η διάκριση μεταξύ των δύο γενικών μεθοδολογιών για τη βελτιστοποίηση με περιορισμούς:

- **Μέθοδοι ενεργού συνόλου**

Αν υποθέσουμε ότι το ελάχιστο χωρίς περιορισμούς της  $\phi$  δεν βρίσκεται στο εσωτερικό του συνόλου εφικτότητας  $\Omega$ , η λύση μας πρέπει να βρίσκεται πάνω στο σύνορο  $\partial\Omega$ . Οι μέθοδοι ενεργού συνόλου αναζητούν το βέλτιστο κατά μήκος του συνόρου. Στην περίπτωση των ανισοτικών περιορισμών, υπάρχουν μέθοδοι ενεργού συνόλου όπου παρακολουθούμε το  $\mathcal{A}(\mathbf{x}_k)$ , εισάγοντας και αφαιρώντας περιορισμούς από το ενεργό σύνολο καθώς «κατηφορίζουμε» κατά μήκος του συνόρου.

- **Άλλες μέθοδοι**

Εδώ, η βέλτιστη λύση προσεγγίζεται με επαναληπτικό τρόπο, είτε από το εσωτερικό της εφικτής περιοχής  $\Omega$  –αυτές είναι οι μέθοδοι εσωτερικών σημείων (interior point methods)– ή, στη γενική περίπτωση, από μια μέθοδο που ενδέχεται να χρησιμοποιεί και μη εφικτά σημεία αλλά η οποία δεν κινείται κατά μήκος του συνόρου.

Στις μεθόδους της δεύτερης κατηγορίας περιλαμβάνονται και εκείνες όπου η αντικειμενική συνάρτηση τροποποιείται ακολουθιακά: Για κάθε τέτοια τροποποίηση επιλύεται το αντίστοιχο πρόβλημα ελαχιστοποίησης χωρίς περιορισμούς.

- **Μέθοδοι ποινής (penalty methods)**

Αυτές οι μέθοδοι, όπως και η μέθοδος καθοδικής κλίσης, αποτελούν δημοφιλή επιλογή λόγω της απλότητάς τους. Θεωρήστε, για παράδειγμα, την ελαχιστοποίηση χωρίς περιορισμούς της αντικειμενικής συνάρτησης με ποινές

$$\min_{\mathbf{x}} \psi(\mathbf{x}, \mu) = \phi(\mathbf{x}) + \frac{1}{2\mu} \sum_{i \in \mathcal{E}} c_i^2(\mathbf{x})$$

όπου  $\mu > 0$  είναι μια παράμετρος. Αυτό έχει νόημα για προβλήματα στα οποία υπάρχουν μόνο ισοτικοί περιορισμοί: παρατηρήστε ότι ισχύει  $c_i(\mathbf{x}^*) = 0$ . Στη συνέχεια μπορεί κανείς να επιλύσει μια ακολουθία τέτοιων προβλημάτων για

φθίνουσες τιμές μη αρνητικών  $\mu$ ,  $\mu \downarrow 0$ , χρησιμοποιώντας το  $\mathbf{x}(\mu_{k-1})$  ώστε να κατασκευάσει την αρχική εικασία για την επαναληπτική διαδικασία χωρίς περιορισμούς για το  $\mathbf{x}(\mu_k)$ .

- **Μέθοδοι φραγμού** (barrier methods)

Αυτές είναι μέθοδοι εσωτερικών σημείων. Με αφετηρία ένα σημείο εντός του εφικτού συνόλου επιλύεται μια ακολουθία προβλημάτων χωρίς περιορισμούς, έτσι ώστε σε κάθε στάδιο να τροποποιείται η αντικειμενική συνάρτηση για να διασφαλίζεται ότι η λύση στο σύνορο προσεγγίζεται από το εσωτερικό του  $\Omega$ . Για παράδειγμα, θεωρήστε το πρόβλημα

$$\min_{\mathbf{x}} \psi(\mathbf{x}, \mu) = \phi(\mathbf{x}) - \mu \sum_{i \in \mathcal{I}} \log c_i(\mathbf{x})$$

όπου  $\mu \downarrow 0$ .

- **Επαυξημένη συνάρτηση Lagrange**

Για ένα πρόβλημα με αποκλειστικά ισοτικούς περιορισμούς, θα θεωρήσουμε μια παραλλαγή της μεθόδου ποινής που ορίζεται από το ακόλουθο πρόβλημα χωρίς περιορισμούς:

$$\min_{\mathbf{x}} \psi(\mathbf{x}, \lambda, \mu) = \phi(\mathbf{x}) - \sum_{i \in \mathcal{E}} \lambda_i c_i(\mathbf{x}) + \frac{1}{2\mu} \sum_{i \in \mathcal{E}} c_i^2(\mathbf{x})$$

Έστω ότι δίνονται οι εκτιμήσεις  $\lambda_k$ ,  $\mu_k$  επιλύουμε το πρόβλημα ελαχιστοποίησης χωρίς περιορισμούς για  $\mathbf{x} = \mathbf{x}_{k+1}$ , και μετά ανανεώνουμε τους πολλαπλασιαστές στις  $\lambda_{k+1}$ ,  $\mu_{k+1}$ .

Η πιο δημοφιλής μέθοδος που χρησιμοποιείται στην πράξη για κώδικες γενικής χρήσης είναι ο **ακολουθιακός τετραγωνικός προγραμματισμός** (sequential quadratic programming, SQP), αν και την ανταγωνίζονται σχεδόν ισάξια ορισμένες μέθοδοι βασισμένες στην επαυξημένη συνάρτηση Lagrange. Σε κάθε επανάληψη επιλύεται ένα τετραγωνικό πρόγραμμα (quadratic program, QP) που ορίζεται ως

$$\begin{aligned} \min_{\mathbf{p}} \quad & \frac{1}{2} \mathbf{p}^T W_k \mathbf{p} + \nabla \phi(\mathbf{x}_k)^T \mathbf{p} \\ \text{s.t.} \quad & c_i(\mathbf{x}_k) + \nabla c_i(\mathbf{x}_k)^T \mathbf{p} = 0, \quad i \in \mathcal{E}, \quad c_i(\mathbf{x}_k) + \nabla c_i(\mathbf{x}_k)^T \mathbf{p} \geq 0, \quad i \in \mathcal{I} \end{aligned}$$

και το οποίο επιστρέφει μια κατεύθυνση  $\mathbf{p}_k$  στην προσεγγιστική τιμή  $(\mathbf{x}_k, \lambda_k)$ . Εδώ, η αντικειμενική συνάρτηση προσεγγίζει τη συνάρτηση Lagrange  $\mathcal{L}$  κοντά στο  $(\mathbf{x}_k, \lambda_k)$  και οι γραμμικοί περιορισμοί συνιστούν τη γραμμικοποίηση των αρχικών περιορισμών στην τρέχουσα προσεγγιστική τιμή. Για την επίλυση του τετραγωνικού

προγράμματος με ανισοτικούς περιορισμούς χρησιμοποιείται μια μέθοδος ενεργού συνόλου.

## Γραμμικός προγραμματισμός

Το πρόβλημα του γραμμικού προγραμματισμού (linear programming, LP) έχει πολυάριθμες εφαρμογές και βρίσκεται στο επίκεντρο του ενδιαφέροντος στον κλάδο της **επιχειρησιακής έρευνας** (operations research) εδώ και δεκαετίες. Στην πραγματικότητα, αποτελεί ειδική περίπτωση του γενικού προβλήματος βελτιστοποίησης με περιορισμούς, όπου τόσο η αντικειμενική συνάρτηση όσο και οι περιορισμοί είναι γραμμικοί. Όπως αρμόζει σε μια τόσο σημαντική κατηγορία προβλημάτων, διαθέτει το δικό του «οικοσύστημα» ορισμών και συμβολισμών.

## Πρωτεύουσα και δυϊκή μορφή

Στην **πρωτεύουσα μορφή** του (primal form), το πρόβλημα LP ορίζεται ως

$$\begin{array}{ll} \min & \mathbf{c}^T \mathbf{x} \\ \text{έτσι ώστε} & A\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{array}$$

Εδώ πρέπει να επισημάνουμε ότι αυτός ο συμβολισμός δεν ταιριάζει απολύτως με τις προηγούμενες συμβάσεις μας: Το  $\mathbf{c}$  είναι απλώς ένα σταθερό διάνυσμα («κόστους») και οι ανισοτικοί περιορισμοί διαχωρίζονται από τον συμβολισμό μητρώων. Υποθέτουμε ότι η μήτρα  $A \in \mathcal{R}^{m \times n}$  είναι πλήρους τάξης γραμμών  $m$ . Υποθέτουμε επίσης ότι ισχύει  $n > m$ , ώστε να αποφευχθούν οι τετριμμένες περιπτώσεις, καθώς και ότι υπάρχει βέλτιστη λύση.

Μια θεμελιώδης έννοια στη βελτιστοποίηση είναι η έννοια της **δυϊκότητας** (duality). Για το πρόβλημα LP που έχουμε ορίσει, η **δυϊκή μορφή** (dual form) ορίζεται ως

$$\begin{array}{ll} \max & \mathbf{b}^T \mathbf{y} \\ \text{έτσι ώστε} & A^T \mathbf{y} \leq \mathbf{c} \end{array}$$

Η δυϊκή μορφή της δυϊκής είναι η πρωτεύουσα μορφή. Αν το  $\mathbf{x}$  είναι εφικτό για την πρωτεύουσα μορφή και το  $\mathbf{y}$  είναι εφικτό για τη δυϊκή (που σημαίνει ότι κάθε μορφή ικανοποιεί τους αντίστοιχους περιορισμούς), τότε έχουμε  $\mathbf{b}^T \mathbf{y} \leq \mathbf{c}^T \mathbf{x}$ , με την ισότητα να ισχύει στο βέλτιστο σημείο. Αυτό είναι ένα βασικό μοντέλο στο εμπόριο, όπου ο στόχος του πωλητή (το μέγιστο κέρδος) και ο στόχος του αγοραστή (το ελάχιστο κόστος) σχετίζονται με αυτόν τον τρόπο.

**Παράδειγμα 9.12.** Μια επιχείρηση παράγει δύο αγροτικά προϊόντα, το ένα υψηλής και το άλλο χαμηλής ποιότητας, των οποίων οι τιμές πώλησης είναι 150€ και 100€ ανά τόνο αντίστοιχα. Τα προϊόντα απαιτούν τρεις πρώτες ύλες, η διαθεσιμότητα των οποίων είναι αντίστοιχα 75, 60 και 25 τόνοι ανά ώρα. Αν  $x_1$  και  $x_2$  είναι τα αντίστοιχα επίπεδα παραγωγής σε τόνους, οι απαιτήσεις σε πρώτες ύλες δίνονται από τις σχέσεις

$$\begin{aligned} 2x_1 + x_2 &\leq 75 \\ x_1 + x_2 &\leq 60 \\ x_1 &\leq 25 \end{aligned}$$

Η δυσκολία έγκειται στον προσδιορισμό αυτών των επιπέδων παραγωγής ώστε να μεγιστοποιηθεί το κέρδος  $150x_1 + 100x_2$ . Όπως είναι φυσικό, τα επίπεδα παραγωγής δεν μπορούν να έχουν αρνητικές τιμές, άρα  $x_1 \geq 0$  και  $x_2 \geq 0$ .

Για να διατυπώσουμε αυτό το πρόβλημα στην πρωτεύουσα μορφή του, πρώτα θα μετατρέψουμε τους ανισοτικούς περιορισμούς σε ισοτικούς εισαγάγοντας πρόσθετους μη αρνητικούς αγνώστους,  $x_3 = 75 - 2x_1 - x_2$ ,  $x_4 = 60 - x_1 - x_2$  και  $x_5 = 25 - x_1$ . Αυτές οι επιπλέον μεταβλητές είναι επίσης μη αρνητικές, αλλά δεν έχουν κάποιο συσχετισμένο κόστος. Επιπλέον, μπορούμε να γράψουμε την αντικειμενική συνάρτηση ως τη συνάρτηση που ελαχιστοποιεί την ποσότητα  $-150x_1 - 100x_2$ . Επομένως, έχουμε διατυπώσει το πρόβλημα LP στην πρωτεύουσα μορφή (σελίδα 443), η οποία ορίζεται ως

$$\begin{aligned} \mathbf{c}^T &= (-150 \quad -100 \quad 0 \quad 0 \quad 0) \\ A &= \begin{pmatrix} 2 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 75 \\ 60 \\ 25 \end{pmatrix} \end{aligned}$$

όπου το διάνυσμα των αγνώστων  $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)^T$  έχει μόνο μη αρνητικά στοιχεία. Σας ζητούμε να μείνετε συντονισμένοι: Στο Παράδειγμα 9.13 θα αποκαλυφθεί το βέλτιστο νούμερο σε ευρώ.

Στη συνέχεια θα προσπαθήσουμε να ερμηνεύσουμε την εμπορική σημασία που έχει η δυϊκή μορφή. ■

Μια **βασική εφικτή λύση** (basic feasible solution),  $\hat{\mathbf{x}}$ , γι' αυτό το πρόβλημα ικανοποιεί τους ισοτικούς και ανισοτικούς περιορισμούς, και έχει το πολύ μόνο  $m$  μηδενικά στοιχεία. Στην πραγματικότητα, το σύνορο του συνόλου περιορισμών  $\Omega$  είναι πολυγωνικό για το πρόβλημα LP, και μια βασική εφικτή λύση αντιστοιχεί σε

μια κορυφή αυτού του πολυγώνου. Επιπλέον, υπάρχει πάντα μια βέλτιστη λύση για το πρόβλημα LP η οποία είναι βασική εφικτή λύση!

Η συνάρτηση Lagrange μπορεί να γραφεί ως

$$\mathcal{L} = \mathbf{c}^T \mathbf{x} - \sum_{i=1}^m y_i \left( \sum_{j=1}^n a_{ij} x_j - b_i \right) - \sum_{j=1}^n s_j x_j$$

δηλαδή χρησιμοποιούμε και το  $y_i$  και το  $s_j$ , όπου το γράμμα  $s$  προέρχεται από τη λέξη slack (χαλαρός), για να συμβολίζουμε τους πολλαπλασιαστές Lagrange  $\lambda$ . Συνεπώς, οι συνθήκες KKT,

$$\begin{aligned} \mathbf{c} - A^T \mathbf{y} - \mathbf{s} &= \mathbf{0} \\ A\mathbf{x} &= \mathbf{b} \\ s_i x_i &= 0, \quad i = 1, \dots, n \\ \mathbf{x} \geq \mathbf{0}, \quad \mathbf{s} \geq \mathbf{0} \end{aligned}$$

είναι και αναγκαίες και ικανές για την ύπαρξη ελαχίστου. Ονομάζονται συνθήκες **συμπληρωματικής χαλαρότητας** (complementary slackness) και εκφράζουν το πρόβλημα LP χωρίς να υπάρχει κάποια αντικειμενική συνάρτηση που πρέπει να ελαχιστοποιηθεί ή να μεγιστοποιηθεί. Εδώ πρέπει να σημειώσουμε τα εξής: Δεδομένου ότι οι περιορισμοί είναι μη αρνητικοί, οι αισθητές  $x_i s_i = 0$  μπορούν να αντικατασταθούν από την πιο συνεπτυγμένη (αν και αινιγματική) συνθήκη

$$\mathbf{x}^T \mathbf{s} = 0$$

Οι εξισώσεις KKT είναι επίσης γνωστές ως η **πρωτεύουσα-δυϊκή μορφή**.

## Αλγόριθμοι LP

Οι δύο γενικές κατηγορίες μεθόδων που προαναφέρθηκαν οδηγούν σε δημοφιλείς εκδοχές αλγόριθμων γραμμικού προγραμματισμού. Η διάσημη, παλαιότατη πλέον, μέθοδος **simplex** κινείται κατά μήκος του συνόρου του  $\Omega$  από τη μία βασική εφικτή λύση στην επόμενη, αναζητώντας τη βέλτιστη λύση. Αποτέλεσε τον στυλοβάτη ολόκληρου του κλάδου της βελτιστοποίησης για αρκετό καιρό έως και τη δεκαετία του 1980. Ακόμα και σήμερα υπάρχουν πολύπλοκες παραλλαγές της μεθόδου simplex που παραμένουν πολύ ανταγωνιστικές.

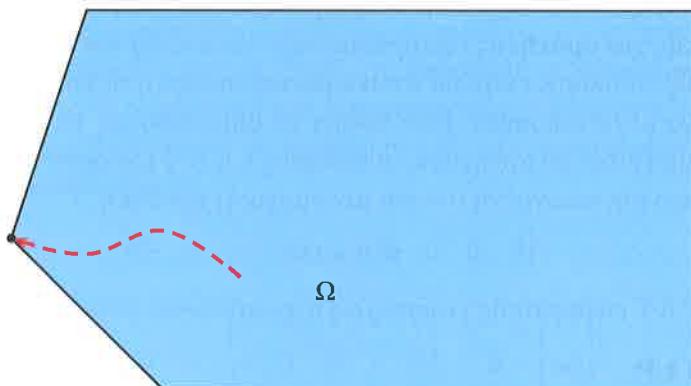
Εδώ, ωστόσο, θα εστιάσουμε σε μια **πρωτεύουσα-δυϊκή** επαναληπτική μέθοδο που, στην πιο καθαρή μορφή της, δεν αγγίζει καν το σύνορο του συνόλου περιορισμών: Όλες οι προσεγγιστικές τιμές ικανοποιούν τις ανισότητες  $\mathbf{x}_k > \mathbf{0}$ ,  $s_k > 0$ . Είναι απλή στην υλοποίηση και έχει πολύ καλή απόδοση στην πλειοψηφία των περιπτώσεων. Μια λεπτομερής υπολογιστική αξιολόγηση θα εξαρτιόταν σε μεγάλο βαθ-

μό από την αποδοτικότητα του χρησιμοποιούμενου γραμμικού επιλυτή, αλλά αυτό ξεφεύγει από το αντικείμενο αυτής της σύντομης εισαγωγής στο θέμα.

Έστω ότι  $X$  είναι η διαγώνια μήτρα με στοιχεία  $x_1, \dots, x_n$  στην κύρια διαγώνιο, και  $S$  είναι η διαγώνια μήτρα με στοιχεία  $s_1, \dots, s_n$  στην κύρια διαγώνιο. Επιπλέον, θέτουμε  $\mathbf{e} = (1, \dots, 1)^T$ . Στη συνέχεια ορίζουμε την **κεντρική διαδρομή** (center path), εξαρτημένη από μια παράμετρο  $\tau \geq 0$ , ως εξής:

$$\begin{aligned} A^T \mathbf{y} + \mathbf{s} &= \mathbf{c} \\ A \mathbf{x} &= \mathbf{b} \\ X S \mathbf{e} &= \tau \mathbf{e} \\ \mathbf{x} > \mathbf{0}, \quad \mathbf{s} > \mathbf{0} \end{aligned}$$

Η λύση  $\mathbf{x}(\tau)$ ,  $\mathbf{y}(\tau)$ ,  $\mathbf{s}(\tau)$  είναι εφικτή για  $\tau \geq 0$ , και μπορούν να κατασκευαστούν μέθοδοι οι οποίες ακολουθούν αυτή τη διαδρομή προς τη βελτιστότητα καθώς το  $\tau \downarrow 0$ . Στο όριο οι συνθήκες KKT είναι ενεργές (δείτε το Σχήμα 9.11).



**Σχήμα 9.11:** Κεντρική διαδρομή στην πρωτεύουσα περιοχή εφικτότητας του γραμμικού προγραμματισμού.

Χρειαζόμαστε έναν ακόμα ορισμό. Για ένα τρέχον σημείο  $\mathbf{z} = (\mathbf{x}, \mathbf{y}, \mathbf{s})$  που ικανοποιεί τις ανισότητες  $\mathbf{x} > \mathbf{0}$ ,  $\mathbf{s} > \mathbf{0}$ , αλλά όχι υποχρεωτικά και κάποια άλλη σχέση, το μέτρο δυϊκότητας (duality measure) ή **χάσμα δυϊκότητας** (duality gap) ορίζεται ως

$$\mu = \frac{1}{n} \mathbf{x}^T \mathbf{s}$$

Στην κεντρική διαδρομή προφανώς ισχύει  $\mu = \tau$ . Πιο γενικά, ισχύει  $\mu > 0$  και θέλουμε το  $\mu \downarrow 0$ .

Στη συνέχεια, μπορούμε να γράψουμε ως εξής το βήμα Newton για το (ήπια μη γραμμικό) υποσύστημα που αποτελείται από τις ισότητες που περιλαμβάνονται τόσο στο σύστημα KKT όσο και στο σύστημα που ορίζει την κεντρική διαδρομή:

$$\begin{pmatrix} 0 & A^T & I \\ A & 0 & 0 \\ S & 0 & X \end{pmatrix} \begin{pmatrix} \delta \mathbf{x} \\ \delta \mathbf{y} \\ \delta \mathbf{s} \end{pmatrix} = \begin{pmatrix} \mathbf{c} - A^T \mathbf{y} - \mathbf{s} \\ \mathbf{b} - A \mathbf{x} \\ \sigma \mu \mathbf{e} - X \mathbf{S} \mathbf{e} \end{pmatrix}$$

όπου  $\sigma \in [0, 1]$  είναι μια παράμετρος κεντραρίσματος. Η τιμή  $\sigma = 1$  δίνει ένα (επιφυλακτικό) βήμα προς την κεντρική διαδρομή με  $\tau = \mu$ , ενώ η τιμή  $\sigma = 0$  δίνει ένα (αισιόδοξο) βήμα Newton για το σύστημα KKT. Μόλις επιλυθεί αυτό το γραμμικό σύστημα, ανανεώνουμε την τρέχουσα λύση προσθέτοντας κατάλληλα πολλαπλάσια  $\delta \mathbf{x}$ ,  $\delta \mathbf{y}$  και  $\delta \mathbf{s}$  στα  $\mathbf{x}$ ,  $\mathbf{y}$  και  $\mathbf{s}$  αντίστοιχα, τέτοια ώστε να εξακολουθούν να ισχύουν οι συνθήκες θετικότητας.

## Ένα βήμα πρόβλεψης-διόρθωσης

Ένας πρακτικός αλγόριθμος του είδους που μόλις αναφέραμε μπορεί να προκύψει αν χρησιμοποιηθεί μια μέθοδος πρόβλεψης-διόρθωσης (predictor-corrector method). Η ιδέα στην οποία στηρίζεται μπορεί να περιγραφεί ως εξής. Αν υποθέσουμε προς το παρόν ότι μπορεί να εκτελεστεί ένα πλήρες βήμα, το ιδανικό σενάριο είναι να πάρουμε στο τέλος του βήματος τα εξής:

$$\begin{aligned} A^T(\mathbf{y} + \delta \mathbf{y}) + (\mathbf{s} + \delta \mathbf{s}) &= \mathbf{c} \\ A(\mathbf{x} + \delta \mathbf{x}) &= \mathbf{b} \\ (X + \delta X)(S + \delta S)\mathbf{e} &= \tau \mathbf{e} \\ \mathbf{x} > \mathbf{0}, \quad \mathbf{s} > \mathbf{0} \end{aligned}$$

Εδώ γράφουμε  $\tau = \sigma\mu$ , όπου  $\mu$  είναι το (υπολογίσιμο) χάσμα δυϊκότητας και  $\sigma$  είναι η παράμετρος κεντραρίσματος. Προκύπτουν δύο ζητήματα:

- Πώς επιλέγουμε τιμή για το  $\sigma$  ώστε να μπορεί να εκτελεστεί ένα μεγάλο βήμα;
- Πώς προσεγγίζουμε τον όρο καμπυλότητας (curvature)  $\Delta = \delta X \delta S$ ; Αυτός είναι και ο μόνος μη γραμμικός όρος εδώ· οι υπόλοιποι όροι συγκροτούν τις γραμμικές εξισώσεις μιας επανάληψης Newton.

Και τα δύο αυτά ζητήματα μπορούν να διευθετηθούν με ένα βήμα πρόβλεψης. Θέτουμε  $\sigma = \sigma_p = 0$ ,  $\Delta_p = 0$ , και λύνουμε τις γραμμικές εξισώσεις που προκύπτουν από το βήμα Newton. Αυτή η κατεύθυνση υποτίθεται ότι αντιστοιχεί στη μέγιστη πρόδο προς το βέλτιστο, και θα είναι επιτυχής αν οι περιορισμοί θετικότητας δεν

αποδειχθούν υπερβολικά περιοριστικοί. Συμβολίζουμε το αποτέλεσμα με  $\delta\mathbf{x}^p, \delta\mathbf{s}^p$ . Αρχικά, αυτό μας δίνει μια προσέγγιση

$$\Delta_c = \delta X^p \delta S^p$$

για τον όρο καμπυλότητας. Κατόπιν πρέπει να βρούμε πόσο μακριά μπορούμε να φτάσουμε στην προβλεπόμενη κατεύθυνση χωρίς να παραβιάσουμε τη θετικότητα, υπολογίζοντας τα

$$\alpha_p = \min \left\{ 1, \min_{\delta x_i < 0} \frac{x_i}{-\delta x_i^p} \right\}, \quad \beta_p = \min \left\{ 1, \min_{\delta s_i < 0} \frac{s_i}{-\delta s_i^p} \right\}$$

Αν εκτελούσαμε το μέγιστο επιτρεπόμενο βήμα, το νέο χάσμα δυϊκότητας θα ήταν

$$\mu_p = \frac{1}{n} (\mathbf{x} + \alpha_p \delta \mathbf{x}^p)^T (\mathbf{s} + \beta_p \delta \mathbf{s}^p)$$

Αυτή η ποσότητα μπορεί να υπολογιστεί. Αν το  $\mu_p$  είναι πολύ μικρότερο από το  $\mu$ , το βήμα πρόβλεψης μπορεί να προκαλέσει μεγάλη πρόοδο. Στην αντίθετη περίπτωση, πρέπει να προστεθεί σημαντικό κεντράρισμα στην κατεύθυνση αναζήτησης. Αυτό αποτυπώνεται στην επιλογή του

$$\sigma = \left( \frac{\mu_p}{\mu} \right)^3$$

Έχοντας σταθεροποιήσει την τιμή της παραμέτρου κεντραρίσματος με αυτόν τον τρόπο και προσεγγίζοντας την καμπυλότητα με το  $\Delta_c$ , μπορούμε στη συνέχεια να λύσουμε το σύστημα

$$\begin{pmatrix} 0 & A^T & I \\ A & 0 & 0 \\ S & 0 & X \end{pmatrix} \begin{pmatrix} \delta \mathbf{x} \\ \delta \mathbf{y} \\ \delta \mathbf{s} \end{pmatrix} = \begin{pmatrix} \mathbf{c} - A^T \mathbf{y} - \mathbf{s} \\ \mathbf{b} - A \mathbf{x} \\ \sigma \mu \mathbf{e} - (XS + \Delta_c) \mathbf{e} \end{pmatrix}$$

ως προς τη διορθωμένη κατεύθυνση. Σημειώστε ότι η ίδια μήτρα και ένα μεγάλο τμήμα του ίδιου δεξιού μέλους χρησιμοποιούνται και για το βήμα της πρόβλεψης και για το βήμα της διόρθωσης.

## Περαιτέρω λεπτομέρειες

Επειδή οι μήτρες  $X$  και  $S$  είναι διαγώνιες, είναι εύκολο (και είθισται) να εκτελούμε απαλοιφή Gauss κατά μπλοκ και να ανάγουμε την παραπάνω μήτρα στη μήτρα σαγματικού σημείου

$$\begin{pmatrix} -X^{-1}S & A^T \\ A & 0 \end{pmatrix}$$

και περαιτέρω στη συμμετρική θετικά ορισμένη μήτρα

$$AS^{-1}XA^T$$

Στον κώδικα που θα παρουσιάσουμε στην επόμενη σελίδα φαίνεται η χρήση της τελευταίας εξίσωσης (πιο συγκεκριμένα, στη συνάρτηση newt1p).

Τέλος, θέλουμε το επόμενο σημείο λύσης να είναι και αυτό θετικό, οπότε η ανανέωση είναι

$$\begin{aligned} \mathbf{x} &= \mathbf{x} + \alpha \delta \mathbf{x}, \quad \alpha = \min \left\{ 1, \min_{\delta x_i < 0} \frac{(1 - \text{tolf})x_i}{-\delta x_i} \right\} \\ \mathbf{s} &= \mathbf{s} + \beta \delta \mathbf{s}, \quad \beta = \min \left\{ 1, \min_{\delta s_i < 0} \frac{(1 - \text{tolf})s_i}{-\delta s_i} \right\} \\ \mathbf{y} &= \mathbf{y} + \beta \delta \mathbf{y} \end{aligned}$$

Θέτουμε, για παράδειγμα,  $\text{tolf} = 0.01$ .

Σε αυτό το σημείο πρέπει να επισημάνουμε ότι ο αλγόριθμός μας δεν είναι αυστηρά μια μέθοδος εσωτερικών σημείων, επειδή οι προσεγγιστικές τιμές δεν ικανοποιούν υποχρεωτικά τους ισοτικούς περιορισμούς των εξισώσεων KKT. Η συνδυασμένη υπό κλίμακα νόρμα αυτών των υπολοίπων, στην οποία αναφερόμαστε ως «μη εφικτότητα», αναμένεται –όπως είναι φυσικό– να συρρικνωθεί έως το 0 καθώς προσεγγίζεται η βέλτιστη λύση.

Εδώ ολοκληρώνεται η περιγραφή του αλγόριθμου για την πρωτεύουσα-δυϊκή μέθοδο, αλλά υπάρχει και κάτι ακόμα που πρέπει να εξετάσουμε πριν δούμε μια υλοποίησή του.

## Άλμα σε κοντινή εφικτή λύση

Θυμηθείτε ότι η ακριβής λύση έχει (δυνητικά πολλά) μηδενικά. Άρα, καθώς η τρέχουσα προσεγγιστική τιμή πλησιάζει στο βέλτιστο, θα υπάρχουν πολύ μεγάλα στοιχεία στις μήτρες  $S^{-1}$  και  $X^{-1}$ , προκαλώντας αναστάτωση στις μεθόδους των Ενοτήτων 7.4 και 7.5. Από την άλλη, η κύρια δυσκολία με τις μεθόδους ενεργού συνόλου –η εκτέλεση αναζήτησης σε εκθετικά μεγάλο πλήθος εφικτών λύσεων– ενδέχεται να περιοριστεί σε σημαντικό βαθμό αν εξεταστούν μόνο κοντινές λύσεις, μια επιλογή που δικαιολογείται επειδή βρισκόμαστε κοντά στο βέλτιστο.

Όταν το χάσμα δυϊκότητας  $\mu$  γίνει μικρότερο από κάποια ανοχή  $\text{tolb}$ , έστω 0.0001, μπορεί να αξίζει τον κόπο να κάνουμε ένα άλμα σε μια κοντινή κορυφή του  $\Omega$  (η οποία αντιστοιχεί σε μια βασική λύση) και να ελέγξουμε τη βελτιστότητα. Αυτό μπορεί να γίνει ως εξής:

1. Έστω ότι  $\mathcal{B} \subset \{1, 2, \dots, n\}$  είναι το σύνολο των  $m$  δεικτών που αντιστοιχούν στα μεγαλύτερα στοιχεία του  $\mathbf{x}$ . Υποψιαζόμαστε ότι αυτή είναι μια «πιθανή βέλτιστη βάση».
2. Θέτουμε

$$\hat{x}_j = 0 \quad \text{αν } j \notin \mathcal{B}, \quad \hat{s}_j = 0 \quad \text{αν } j \in \mathcal{B}$$

Κατασκευάζουμε τη μήτρα  $B$ , διαστάσεων  $m \times m$ , από τις  $m$  βασικές στήλες της  $A$ , και ομοίως το διάνυσμα  $\mathbf{c}_B$  από το  $\mathbf{c}$ .

3. Λύνουμε την εξίσωση

$$B\hat{\mathbf{x}}_B = \mathbf{b}$$

και εισάγουμε τις τιμές του  $\hat{\mathbf{x}}_B$  ως τις βασικές τιμές του  $\hat{\mathbf{x}}$  για  $j \in \mathcal{B}$ .

4. Θέτουμε

$$\hat{\mathbf{y}} = B^{-T} \mathbf{c}_B, \quad \hat{\mathbf{s}} = \mathbf{c} - A^T \hat{\mathbf{y}}$$

5. Αν ισχύουν οι ανισότητες  $\hat{\mathbf{x}} > -\epsilon \mathbf{e}$ ,  $\hat{\mathbf{s}} > -\epsilon \mathbf{e}$  και  $\frac{1}{n} \hat{\mathbf{s}}^T \hat{\mathbf{x}} < \epsilon$  για πολύ μικρή θετική ανοχή  $\epsilon$ , έχει βρεθεί μια βέλτιστη λύση στο  $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{s}})$ . Αν δεν ισχύουν, αυτές οι πληροφορίες αγνοούνται και ο αλγόριθμος συνεχίζει από το  $(\mathbf{x}, \mathbf{y}, \mathbf{s})$ .

Ακολουθεί το πρόγραμμά μας. Δεν πραγματοποιεί εντυπωσιακούς ελέγχους για ειδικές περιστάσεις όπως εκφυλισμένες περιπτώσεις, κενά σύνολα εφικτότητας ή το ενδεχόμενο μη φραγμένης λύσης. Επιπλέον, έχουμε προτιμήσει να κάνουμε πιο ευανάγνωστο τον κώδικα παρά να βελτιστοποιήσουμε την απόδοσή του. Είναι εκπληκτικό το πόσο απλό μπορεί τελικά να γίνει ένα αποδοτικό πρόγραμμα για ένα πρωταρχικής σημασίας μη τετριμμένο πρόβλημα.

```
function [x,gap,nbas] = lpm (A,b,c)
%
% function [x,gap,nbas] = lpm (A,b,c)
%
% Επιλύει το πρόβλημα γραμμικού προγραμματισμού
% min c^T x  s.t. Ax = b, x >= 0
%
% Η μήτρα A έχει διαστάσεις 1 x m, το διάνυσμα b έχει
```

% διαστάσεις  $l \times 1$ , και το διάνυσμα  $s$  έχει διαστάσεις  $m \times 1$   
% Επιστρέφει τη λύση  $x$  και το χάσμα δυϊκότητας που( πρέπει  
% να έχει τιμή κοντά στο 0 αν όλα έχουν πάει καλά)  
% Επίσης, η μεταβλητή  $nbas$  είναι το πλήθος των αλμάτων για  
% τον έλεγχο βασικών λύσεων πριν προσεγγιστεί η βελτιστότητα

```
[l,m] = size(A);
scaleb = norm(b) + norm(A,inf) + 1;
scalec = norm(c) + norm(A,inf) + 1;
tolf = 0.01; otol = 1-tolf;
toln = 1.e-9; tolp = 1.e-10; tolb = 1.e-4;
nbas = 0;

% αρχική εικασία
x = ones(m,1); s = ones(m,1); y = zeros(l,1);

fprintf('itn      gap      infeas      mu\n')
for it = 1:2*m+10
    % επανάληψη, η πρόβλεψη - διόρθωση
    % μετριέται ως ένα βήμα

    % μέτρο δυϊκότητας
    mu = (x'*s)/m;
    % πρόβλεψη διόρθωσης
    [dx,dy,ds] = newtlp(A,b,c,x,y,s,theta);

    % ενσωμάτωση θετικότητας στους περιορισμούς
    alfa = 1; beta = 1;
    for i=1:m
        if dx(i) < 0, alfa = min(alfa, -x(i)/dx(i)); end
        if ds(i) < 0, beta = min(beta, -s(i)/ds(i)); end
    end

    % το υποψήφιο μέτρο δυϊκότητας
    muaff = ( (x+alfa*dx)' * (s+beta*ds) ) / m;
    % παράμετρος κεντραρίσματος
    sigma = (muaff/mu)^3;

    % διόρθωση προς την κεντρική διαδρομή
```

```

smu = sigma * mu;
[dx,dy,ds] = newtlp(A,b,c,x,y,s,smu,dx,ds);

% ενσωμάτωση θετικότητας στους περιορισμούς
alfa = 1; beta = 1;
for i=1:m
    if dx(i) < 0, alfa = min(alfa, -otol*x(i)/dx(i)); end
    if ds(i) < 0, beta = min(beta, -otol*s(i)/ds(i)); end
end

% ανανέωση λύσης
x = x + alfa*dx;
s = s + beta*ds;
y = y + beta*dy;

% έλεγχος προόδου
infeas = norm(b - A*x)/scaleb + norm(c - A'*y - s)/scalec;
gap = (c'*x - b'*y) / m;
if (infeas > 1.e+12)+(gap < -toln)
    fprintf('αδυναμία σύγκλισης: ίσως δεν υπάρχει λύση')
    return
end
fprintf('%d      %e      %e      %e\n',it,gap,infeas,mu)

if (abs(infeas) < tolن)*(abs(gap) < tolن), return, end

% άλμα στην επόμενη βασική λύση
if gap < tolб
    nbas = nbas + 1;
    [xx,sortof] = sort(-x);
    [xx,yy,ss] = basln(A,b,c,sortof);
    gap = (c'*xx - b'*yy) / m;
    if (sum(xx+tolp >= 0) > m-1)*(sum(ss+tolp >= 0) > m-1)...
        *(abs(gap) < tolن)
        x = xx;
        return
    end
end

```

```

end
function [dx,dy,ds] = newtlp(A,b,c,x,y,s,mu,dx,ds)
%
% function [dx,dy,ds] = newtlp(A,b,c,x,y,s,mu,dx,ds)
%
% Βήμα Newton για γραμμικό προγραμματισμό
[1,m] = size(A);
rc = A'*y + s - c;
rb = A*x - b;
rt = x.*s - mu;
if nargin == 9, rt = rt + dx.*ds; end

rhs = [-rb + A * ((rt - x.*rc)./s)];
Mat = A * diag(x./s) * A';
dy = Mat \ rhs;
ds = -rc - A'*dy;
dx = -(x.*ds + rt)./s;

function [x,y,s] = basln(A,b,c,sort)
%
% function [x,y,s] = basln(A,b,c,sort)
%
% Για δοθέν διάνυσμα δεικτών, το πρώτο 1 υποδεικνύει
% μια βάση από την A.
% Βρίσκει την αντίστοιχη βασική λύση.
[1,m] = size(A);
B = zeros(1,1); cb = zeros(1,1);

% κατασκευή βάσης
for j=1:1
    B(:,j) = A(:,sort(j));
    cb(j) = c(sort(j));
end

xb = B \ b;
x = zeros(m,1);
for j=1:1
    x(sort(j)) = xb(j);
end

```

```

end
y = B' \ cb;
s = c - A'*y;

```

**Παράδειγμα 9.13.** Αν η συνάρτηση `lpm` εκτελεστεί για το μικρό πρόβλημα του Παραδείγματος 9.12, η σύγκλιση επιτυγχάνεται μετά από 13 επαναλήψεις. Επιπλέον, ένα άλμα σε μια κοντινή βασική λύση βρίσκει το βέλτιστο  $x_1 = 15$ ,  $x_2 = 45$ . Επομένως, το κέρδος είναι  $150x_1 + 100x_2 = 6750$  €.

Η απόδοση του προγράμματός μας σε αυτό το μικρό πρόβλημα, όπου  $m = 3$ ,  $n = 5$ , δεν είναι και τόσο εξαιρετική, αλλά αυτό δεν είναι πολύ σημαντικό: Περισσότερο ενδιαφέρον έχει να δούμε τι συμβαίνει σε μεγαλύτερα προβλήματα. ■

**Παράδειγμα 9.14.** Θα κατασκευάσουμε ένα δοκιμαστικό πρόβλημα ως εξής. Αρχικά θέτουμε  $m = 260$ ,  $n = 570$ , και παράγουμε μια τυχαία μήτρα  $A$  διαστάσεων  $m \times n$  με τη συνάρτηση `randn` του MATLAB. Στη συνέχεια παράγουμε δύο ακόμα τυχαία (μη αρνητικά)  $n$ -διανύσματα, τα  $\hat{x}$  και  $c$ , με τη συνάρτηση `rand` του MATLAB. Έπειτα θέτουμε  $b = A\hat{x}$  και ξεχνάμε το  $\hat{x}$ . Αυτό εγγυάται ότι η περιοχή εφικτότητας δεν είναι κενή.

Η κλήση `lpm(A, b, c)` οδηγεί στη σύγκλιση σε 9 επαναλήψεις (το ιστορικό της σύγκλισης παρουσιάζεται στον Πίνακα 9.3). Έγιναν δύο άλματα σε κοντινές συνοριακές κορυφές, το πρώτο αποτυχημένο και το δεύτερο επιτυχημένο, τερματίζοντας την επανάληψη με τη βέλτιστη λύση για την οποία και τα δύο μέτρα που έχουν καταγραφεί στον Πίνακα 9.3 έχουν μηδενική τιμή.

**Πίνακας 9.3.** Πρόσθιος του αλγόριθμου *LP* για την πρωτεύονσα-δυϊκή μέθοδο (Παράδειγμα 9.14).

Επανάληψη	Νόρμα μη εφικτότητας	Χάσμα δυϊκότητας
1	6.38e-02	1.78e-01
2	1.36e-02	5.60e-02
3	3.50e-03	1.78e-02
4	7.80e-04	5.89e-03
5	1.65e-04	1.77e-03
6	5.07e-05	6.79e-04
7	1.74e-05	2.15e-04
8	3.53e-06	5.78e-05
9	5.66e-07	2.01e-05

Προσέξτε πόσο μικρότερο από το  $m$  είναι το συνολικό πλήθος των επαναλήψεων. ■

**Παράδειγμα 9.15 (λύσεις ελάχιστης νόρμας για υποκαθορισμένα συστήματα).** Αυτή η μελέτη περίπτωσης είναι μεγαλύτερη από ό,τι συνήθως και επιλύει ένα πρόβλημα το οποίο προς το παρόν παρουσιάζει γενικό ενδιαφέρον.

Θεωρήστε το υποκαθορισμένο γραμμικό σύστημα εξισώσεων

$$J\mathbf{y} = \mathbf{b}$$

όπου  $J$  είναι μια δοθείσα μήτρα διαστάσεων  $m \times \hat{n}$  με πλήρη τάξη γραμμών, και  $\mathbf{b}$  είναι ένα δεξιό μέλος το οποίο επίσης δίνεται. Για να έχετε ένα παράδειγμα στο μυαλό σας, σκεφτείτε ότι η  $J$  συμπίπτει με τη μήτρα  $A$  του προηγούμενου παραδείγματος, με τη διαφορά ότι εδώ δεν υπάρχει αντικειμενική συνάρτηση  $\phi$  προς ελαχιστοποίηση και το  $\mathbf{y}$  δεν έχει καμία σχέση με δυϊκές μεταβλητές. Αν ισχύει  $\hat{n} > m$ , αυτό το σύστημα μπορεί να έχει πολλές λύσεις επειδή ο μηδενόχωρος της  $J$  δεν είναι κενός. Ποια από τις δύο μήτρες πρέπει να επιλέξουμε;

Αν έχετε διαβάσει πρόσφατα την Ενότητα 8.2, μπορείτε να απαντήσετε σχεδόν αυτόματα: Επιλέγουμε εκείνη που έχει τη μικρότερη  $\ell_2$ -νόρμα. Συνεπώς, λύνουμε το εύκολο πρόβλημα βελτιστοποίησης (μη γραμμικού προγραμματισμού) με περιορισμούς

$$\begin{array}{ll} \min_{\mathbf{y}} & \|\mathbf{y}\|_2 \\ \text{έτσι ώστε} & J\mathbf{y} = \mathbf{b} \end{array}$$

Αυτό μπορεί να γίνει με χρήση της διάσπασης ιδιαζουσών τιμών (SVD) της μήτρας  $J$ , όταν η  $J$  είναι επαρκώς μικρή. Θυμηθείτε από την Ενότητα 4.4 ότι μπορούμε να γράψουμε τη διάσπαση ως  $J = U\Sigma V^T$ . Τότε,  $\Sigma(V^T\mathbf{y}) = U^T\mathbf{b}$ , με τη  $\Sigma$  να είναι ουσιαστικά μια διαγώνια μήτρα στην οποία οι ιδιάζουσες τιμές  $\sigma_i$  είναι διατεταγμένες κατά φθίνουσα σειρά στην κύρια διαγώνιο. Επομένως, για  $\mathbf{z} = V^T\mathbf{y}$  θέτουμε

$$z_i = \begin{cases} \frac{(U^T\mathbf{b})_i}{\sigma_i}, & 1 \leq i \leq m \\ 0, & m < i \leq \hat{n} \end{cases}$$

και η λύση της ελάχιστης  $\ell_2$ -νόρμας είναι  $\mathbf{y} = V\mathbf{z}$ . Αυτοί οι μετασχηματισμοί μπροστίσω επιτρέπονται (αν θυμάστε) επειδή οι μήτρες  $U$  και  $V$  είναι ορθογώνιες, οπότε οι μετασχηματισμοί που τις χρησιμοποιούν διατηρούν την  $\ell_2$ -νόρμα. Για μια εναλλακτική μέθοδο επίλυσης δείτε επίσης την Άσκηση 22.

Η διαδικασία επίλυσης που περιγράφηκε παραπάνω εκτελείται συχνά σε πολλές πρακτικές εφαρμογές. Όμως δεν είναι πάντα αυτό που θέλουμε, επειδή όλα τα στοιχεία της λύσης για που προκύπτει συνήθως είναι μη μηδενικά. Υπάρχουν σημαντικές εφαρμογές στις οποίες ξέρουμε εκ των προτέρων ότι το διάνυσμα  $\mathbf{b}$  στο δεξιό μέλος μπορεί να γραφεί ως γραμμικός συνδυασμός ενός σχετικά μικρού πλήθους  $l$ ,  $l \leq m < \hat{n}$ , των στηλών της  $J$ . Μια **αραιή λύση** για το συγκεκριμένο γραμμικό σύστημα θα μπορούσε να κατασκευαστεί στη συνέχεια, μόνο αν γνωρίζαμε όμως ποιες  $l$  στήλες της  $J$  να επιλέξουμε, θέτοντας όλα τα υπόλοιπα στοιχεία του για ίσα με το 0. Αυτό είναι πολύ ενδιαφέρον, επειδή μια τέτοια αραιή λύση ισοδυναμεί με το να επιλέξουμε τι είναι σημαντικό στο μοντέλο που αναπαρίσταται από το γραμμικό σύστημα εξισώσεων.

Επομένως, πώς βρίσκουμε μια τέτοια αραιή λύση; Όπως μπορεί να αποδειχθεί, το πρόβλημα της εύρεσης μιας λύσης με το ελάχιστο πλήθος μη μηδενικών στοιχείων όντως μπορεί να είναι πολύ δύσκολο.

Εντυχώς, όπως επίσης αποδεικνύεται, συχνά (αν και όχι πάντα) μια αραιή λύση προκύπτει από την επίλυση του ακόλουθου προβλήματος βελτιστοποίησης με περιορισμούς:

$$\begin{array}{ll} \min_y & \|y\|_1 \\ \text{έτσι ώστε} & Jy = \mathbf{b} \end{array}$$

Συνεπώς, ψάχνουμε τη λύση με τη μικρότερη  $\ell_1$ -νόρμα.

Το τελευταίο πρόβλημα μπορεί να διατυπωθεί ως πρόβλημα LP. Γράφουμε

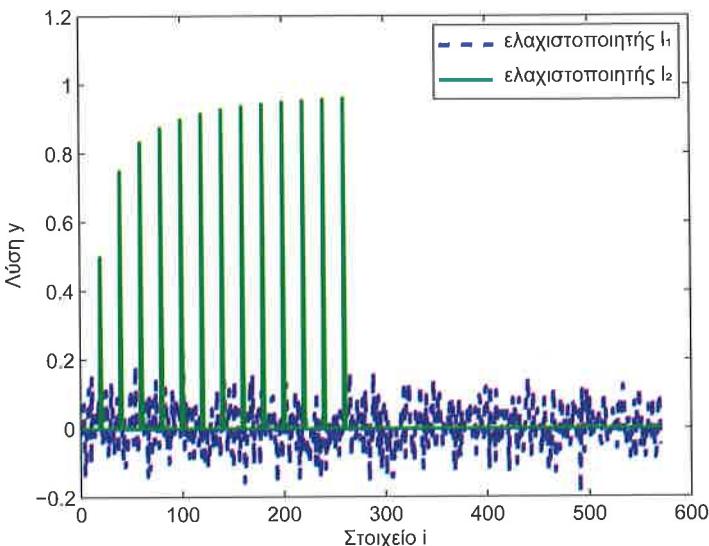
$$y_i = u_i - v_i, \quad 1 \leq i \leq \hat{n}$$

και θέτουμε

$$A = [J, -J], \quad \mathbf{x} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}, \quad \mathbf{c} = \mathbf{e}$$

Θέτοντας επίσης  $n = 2\hat{n}$ , βρισκόμαστε αντιμέτωποι με ένα πρόβλημα LP σε τυπική, πρωτεύουσα μορφή (με βάση τον προηγούμενο συμβολισμό μας), για το οποίο μπορεί να εκτελεστεί το πρόγραμμα 1pm που παρουσιάσαμε παραπάνω.

Μια επιπλέον βελτίωση που μπορεί να γίνει στη συνάρτηση 1pm γι' αυτή την ειδική μεν, αλλά αρκετά σημαντική δε, κατηγορία προβλημάτων αφορά τους πρώτους ελέγχους για μια βέλτιστη λύση. Καθώς η διαδικασία των επαναλήψεων εξελίσσεται και πλησιάζει περισσότερο στη σύγκλιση, τα κυρίαρχα στοιχεία του γιαρχίζουν να ξεχωρίζουν αρκετά σε μέγεθος από τα υπόλοιπα. Επομένως, αντί για τη γενική τακτική των «αλμάτων στην πλησιέστερη βάση», εδώ μπορούμε να συγκε-



**Σχήμα 9.12:** Δύο λύσεις ελάχιστης νόρμας για ένα υποκαθορισμένο γραμμικό σύστημα εξισώσεων. Η λύση  $\ell_1$  είναι βολικά αραιή.

ντρώσουμε τις στήλες της  $J$  που αντιστοιχούν στις πρώτες  $l$  μεγαλύτερες τρέχουσες τιμές  $|y_i|$  σε μια μήτρα διαστάσεων  $m \times l$ . Κατόπιν πρέπει να λύσουμε ένα γραμμικό υπερκαθορισμένο πρόβλημα ελάχιστων τετραγώνων στο οποίο μπορούν να εφαρμοστούν οι τεχνικές του Κεφαλαίου 6, προσπαθώντας να προσαρμόσουμε το  $\mathbf{b}$  σε αυτά τα  $l$  στοιχεία: Ελέγχουμε το υπόλοιπο που προκύπτει και ανακοινώνουμε τη νίκη αν η νόρμα του είναι επαρκώς μικρή, όπου για να βρούμε την πλήρη λύση θέτουμε όλα τα υπόλοιπα στοιχεία του  $y$  ίσα με το 0.

Ας δούμε ένα αριθμητικό παράδειγμα. Κατασκευάζουμε την  $J$  ακριβώς όπως την  $A$  στο Παράδειγμα 9.14, άρα  $m = 260$  και  $n = 2 \cdot 570 = 1140$ . Έπειτα θέτουμε  $\hat{y}_i = 1 - 10/i$  για  $i = 20, 40, 60, \dots, 260$ , ή  $\hat{y}_i = 0$  διαφορετικά, και κατόπιν σχηματίζουμε το  $\mathbf{b} = J\hat{\mathbf{y}}$  και δεν χρησιμοποιούμε άλλο το  $\hat{\mathbf{y}}$ .

Η λύση ελάχιστης  $\ell_2$ -νόρμας,  $\mathbf{y}^{(2)}$ , δίνει

$$\|\mathbf{y}^{(2)}\|_2 = 2.12, \quad \|\mathbf{y}^{(2)}\|_1 = 34.27$$

Για τη λύση ελάχιστης  $\ell_1$ -νόρμας,  $\mathbf{y}^{(1)}$ , εκτελούμε τη συνάρτηση `l1pr`, με την αναζήτηση για το πρώτο ελάχιστο να είναι αυτή που περιγράφηκε παραπάνω, χρησιμοποιώντας την εκτίμηση  $l = 20$ . Απαιτούνται μόνο 3 επαναλήψεις πριν βρεθεί μια λύση που να ικανοποιεί τις σχέσεις

$$\|\mathbf{y}^{(1)}\|_2 = 3.20, \quad \|\mathbf{y}^{(1)}\|_1 = 11.41$$

Οι δύο λύσεις έχουν σχεδιαστεί στο Σχήμα 9.12. Είναι προφανές ότι η  $\ell_1$  είναι ιδιαίτερα αραιή, αναπαραγάγοντας το  $\hat{y}$  σε αυτή τη συγκεκριμένη περίπτωση, ενώ η  $\ell_2$  δεν είναι. Πολλοί ερευνητές δείχνουν ξεκάθαρα τον ενθουσιασμό τους γι' αυτό το χαρακτηριστικό της  $\ell_1$ -νόρμας. ■

*Ασκήσεις γι' αυτή την ενότητα: 19–22.*

## 9.4 Ασκήσεις

### 0. Ερωτήσεις επανάληψης

- (α) Τι είναι η βελτιστοποίηση με περιορισμούς και η βελτιστοποίηση χωρίς περιορισμούς; Αναφέρετε μια σημαντική διαφορά τους.
- (β) Τι είναι η ιακωβιανή μήτρα;
- (γ) Πόσες λύσεις αναμένεται να έχει ένα μη γραμμικό σύστημα  $n$  αλγεβρικών εξισώσεων;
- (δ) Ποιες είναι οι βασικές πρόσθετες δυσκολίες κατά την αριθμητική επίλυση των προβλημάτων που μελετήσαμε στην Ενότητα 9.1 συγκριτικά με τα προβλήματα που περιγράφαμε στο Κεφάλαιο 3;
- (ε) Αναφέρετε τη συνθήκη τετραγωνικής σύγκλισης ενός μη γραμμικού συστήματος και εξηγήστε τη σπουδαιότητά της.
- (στ) Εξηγήστε πώς η ελαχιστοποίηση μιας μη γραμμικής συνάρτησης ως προς πολλές μεταβλητές (Ενότητα 9.2) οδηγεί στην επίλυση συστημάτων αλγεβρικών εξισώσεων (Ενότητα 9.1).
- (ζ) Ποιες είναι οι αναγκαίες και ποιες οι ικανές συνθήκες για την ύπαρξη ελαχίστου χωρίς περιορισμούς;
- (η) Ποια είναι η διαφορά μεταξύ ενός τοπικού ελαχίστου και ενός ολικού ελαχίστου; Ποιο από τα δύο προσπαθούν συνήθως να βρουν οι μέθοδοι που περιγράφονται σε αυτό το κεφάλαιο;
- (θ) Γιατί είναι σημαντικό να διατηρείται η μήτρα επαναλήψεων  $B_k$  συμμετρική και θετικά ορισμένη κατά την αναζήτηση ενός ελαχίστου μιας λείας συνάρτησης  $\phi(\mathbf{x})$ ; Το εγγυάται αυτόματα αυτό η μέθοδος Newton;
- (ι) Ορίστε την κατεύθυνση καθόδου και την ευθύγραμμη αναζήτηση, και εξηγήστε τη σχέση που τις συνδέει.

- (ια) Τι είναι μια μέθοδος καθοδικής κλίσης; Αναφέρετε δύο πλεονεκτήματα και δύο μειονεκτήματα που έχει μια τέτοια μέθοδος συγκριτικά με τη μέθοδο Newton για βελτιστοποίηση χωρίς περιορισμούς.
- (ιβ) Τι είναι μια μέθοδος οιονεί Newton; Αναφέρετε τρία πλεονεκτήματα που έχει μια τέτοια μέθοδος συγκριτικά με τη μέθοδο Newton.
- (ιγ) Γράψτε τις συνθήκες KKT και εξηγήστε τη σπουδαιότητά τους.
- (ιδ) Τι είναι μια μέθοδος ενεργού συνόλου; Αναφέρετε μια πολύ γνωστή μέθοδο ενεργού συνόλου για το πρόβλημα του γραμμικού προγραμματισμού.
- (ιε) Πώς σχετίζεται η πρωτεύουσα-δυϊκή μορφή για τον γραμμικό προγραμματισμό με την πρωτεύουσα μορφή τέτοιων προβλημάτων;
- (ιστ) Ορίστε την κεντρική διαδρομή και το χάσμα δυϊκότητας για προβλήματα γραμμικού προγραμματισμού.

1. Γράψτε την επανάληψη της μεθόδου Newton για την επίλυση καθενός από τα παρακάτω συστήματα:

(α)

$$\begin{aligned}x_1^2 + x_1 x_2^3 &= 9 \\3x_1^2 x_2 - x_2^3 &= 4\end{aligned}$$

(β)

$$\begin{aligned}x_1 + x_2 - 2x_1 x_2 &= 0 \\x_1^2 + x_2^2 - 2x_1 + 2x_2 &= -1\end{aligned}$$

(γ)

$$\begin{aligned}x_1^3 - x_2^2 &= 0 \\x_1 + x_1^2 x_2 &= 2\end{aligned}$$

2. Θεωρήστε το σύστημα

$$x_1 - 1 = 0$$

$$x_1 x_2 - 1 = 0$$

Σίγουρα θα συμφωνήσετε ότι η άμεση επίλυσή του είναι κάτι το τετριμένο, αλλά έστω ότι θέλουμε να εφαρμόσουμε ούτως ή άλλως τη μέθοδο Newton. Για ποιες αρχικές εικασίες θα αποτύχει η μέθοδος; Αιτιολογήστε την απάντησή σας.

3. Αυτή η άσκηση επικεντρώνεται σε αποδείξεις.

Έστω ότι η  $\mathbf{f}(\mathbf{x})$  είναι συνεχώς παραγωγίσιμη κατά Lipschitz σε ανοικτό κυρτό σύνολο  $\mathcal{D} \subset \mathbb{R}^n$ , δηλαδή υπάρχει σταθερά  $\gamma \geq 0$  για την οποία ισχύει

$$\|J(\mathbf{x}) - J(\mathbf{y})\| \leq \gamma \|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{D}$$

όπου  $J$  είναι η ιακωβιανή μήτρα διαστάσεων  $n \times n$  της  $\mathbf{f}$ . Είναι εφικτό να αποδειχθεί ότι αν τα  $\mathbf{x}$  και  $\mathbf{x} + \mathbf{p}$  ανήκουν στο  $\mathcal{D}$ , τότε

$$\mathbf{f}(\mathbf{x} + \mathbf{p}) = \mathbf{f}(\mathbf{x}) + \int_0^1 J(\mathbf{x} + \tau \mathbf{p}) \mathbf{p} d\tau$$

(α) Θεωρώντας τα παραπάνω ως δεδομένα, δείξτε ότι

$$\|\mathbf{f}(\mathbf{x} + \mathbf{p}) - \mathbf{f}(\mathbf{x}) - J(\mathbf{x})\mathbf{p}\| \leq \frac{\gamma}{2} \|\mathbf{p}\|^2$$

(β) Υποθέστε περαιτέρω ότι υπάρχει μια ρίζα  $\mathbf{x}^* \in \mathcal{D}$  που ικανοποιεί την εξίσωση

$$\mathbf{f}(\mathbf{x}^*) = \mathbf{0}, \quad J(\mathbf{x}^*) \text{ μη ιδιάζουσα}$$

Δείξτε ότι η μέθοδος Newton συγκλίνει τετραγωνικά για  $\mathbf{x}_0$  που είναι επαρκώς κοντά στο  $\mathbf{x}^*$ .

4. Θεωρήστε την ακόλουθη μη γραμμική μερική διαφορική εξίσωση (ΜΔΕ) στις δύο διαστάσεις:

$$-\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + e^u = g(x, y)$$

Εδώ, η  $u = u(x, y)$  είναι μια συνάρτηση ως προς δύο μεταβλητές  $x$  και  $y$ , που ορίζεται στο μοναδιαίο τετράγωνο  $(0, 1) \times (0, 1)$  και υπακούει σε ομογενείς συνοριακές συνθήκες Dirichlet. Τη διακριτοποιούμε σε ομοιόμορφο πλέγμα, ακριβώς όπως στο Παράδειγμα 7.1, χρησιμοποιώντας την τιμή  $h = 1/(N + 1)$ , οπότε παίρνουμε τις εξισώσεις

$$4u_{i,j} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1} + h^2 e^{u_{i,j}} = h^2 g_{i,j}, \quad 1 \leq i, j \leq N \\ u_{i,j} = 0, \quad \text{διαφορετικά}$$

(α) Βρείτε την ιακωβιανή μήτρα  $J$  και αποδείξτε ότι είναι πάντα συμμετρική και θετικά ορισμένη.

(β) Γράψτε ένα πρόγραμμα στο MATLAB το οποίο να λύνει αυτό το σύστημα μη γραμμικών εξισώσεων με τη μέθοδο Newton για  $N = 8, 16, 32$ . Κατασκευάστε μια συνάρτηση  $g(x, y)$  για το δεξιό μέλος τέτοια ώστε η ακριβής λύση του διαφορικού προβλήματος να είναι  $u(x, y) = \sin(\pi x) \sin(\pi y)$ . Ξεκινήστε με μια αρχική εικασία που έχει μόνο μηδενικά στοιχεία, και τερ-

ματίστε την επανάληψη όταν ισχύει  $\|\delta u^{(k)}\|_2 < 10^{-6}$ . Σχεδιάστε τις νόρμες  $\|\delta u^{(k)}\|_2$  και εξηγήστε τη σύγκλιση που παρατηρείτε.

5. Ένα γραμμικό σύστημα εξισώσεων  $Ax = b$ , διαστάσεων  $n \times n$ , τροποποιείται με τον ακόλουθο τρόπο: Για κάθε  $i$ ,  $i = 1, \dots, n$ , η τιμή  $b_i$  στο δεξιό μέλος της  $i$ -οστής εξίσωσης αντικαθίσταται από την τιμή  $b_i - x_i^3$ . Προφανώς, το τροποποιημένο σύστημα εξισώσεων (ως προς τους αγνώστους  $x_i$ ) είναι μη γραμμικό.
  - (α) Βρείτε την αντίστοιχη ιακωβιανή μήτρα.
  - (β) Δεδομένου ότι η μήτρα  $A$  είναι αυστηρά διαγώνια υπερέχουσα με θετικά στοιχεία στη διαγώνιο της, αναφέρετε αν η ιακωβιανή μήτρα σε κάθε προσεγγιστική τιμή θα είναι εγγυημένα μη ιδιάζουσα ή όχι.
  - (γ) Έστω ότι η μήτρα  $A$  είναι συμμετρική θετικά ορισμένη (όχι υποχρεωτικά διαγώνια υπερέχουσα) και ότι εφαρμόζεται η μέθοδος Newton για την επίλυση του μη γραμμικού συστήματος. Είναι εγγυημένη η σύγκλιση;
6. (α) Έστω ότι η μέθοδος Newton εφαρμόζεται σε ένα γραμμικό σύστημα  $Ax = b$ . Ποιος είναι ο επαναληπτικός τύπος και πόσες επαναλήψεις απαιτούνται για να επιτευχθεί σύγκλιση;
  - (β) Έστω ότι η ιακωβιανή μήτρα είναι μη ιδιάζουσα στη λύση ενός μη γραμμικού συστήματος εξισώσεων. Διατυπώστε εικασίες σχετικά με το τι μπορεί να συμβεί όσον αφορά τη σύγκλιση και τον ρυθμό σύγκλισης. Πιο συγκεκριμένα, είναι εφικτό να προκύψει μια κατάσταση στην οποία η επανάληψη της μεθόδου Newton συγκλίνει μεν αλλά όχι τετραγωνικά;
7. Χρησιμοποιήστε τη μέθοδο Newton για να λύσετε μια διακριτοποιημένη εκδοχή της διαφορικής εξίσωσης

$$y'' = -(y')^2 - y + \ln x, \quad 1 \leq x \leq 2, \quad y(1) = 0, \quad y(2) = \ln 2$$

Η διακριτοποίηση σε ομοιόμορφο πλέγμα, με τον συμβολισμό του Παραδείγματος 9.3, μπορεί να είναι

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + \left( \frac{y_{i+1} - y_{i-1}}{2h} \right)^2 + y_i = \ln(1 + ih), \quad i = 1, 2, \dots, n$$

Η λύση για το συγκεκριμένο πρόβλημα είναι  $y(x) = \ln x$ . Συγκρίνετε τα δικά σας αριθμητικά αποτελέσματα με τη λύση  $y(x)$  για  $n = 8, 16, 32$  και  $64$ . Σχολιάστε τη σύγκλιση της μεθόδου Newton ως προς τις επαναλήψεις και το μέγεθος του πλέγματος, καθώς και ως προς το σφάλμα της λύσης.

8. Θεωρήστε το μη γραμμικό πρόβλημα

$$-\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) - e^u = 0$$

που ορίζεται στο μοναδιαίο τετράγωνο  $0 < x, y < 1$  με ομογενείς συνοριακές συνθήκες Dirichlet (δείτε το Παράδειγμα 7.1). Είναι γνωστό ότι το πρόβλημα αυτό έχει δύο λύσεις.

Χρησιμοποιώντας διακριτοποίηση σε ομοιόμορφο πλέγμα με μέγεθος βήματος  $h = 1/(N + 1) = 2^{-7}$  και επεκτείνοντας τη διακριτοποίηση που περιγράφεται στο Παράδειγμα 7.1, βρείτε προσεγγίσεις για τις συναρτήσεις των δύο λύσεων. Χρησιμοποιήστε τη μέθοδο Newton με κατάλληλες αρχικές εικασίες και λύστε απευθείας το γραμμικό σύστημα που προκύπτει [δηλαδή με τον τελεστή ανάποδης καθέτου του MATLAB (backslash)]. Σχεδιάστε τις δύο λύσεις και εμφανίστε τις υπό κλίμακα (scaled) νόρμες τους,  $\|\mathbf{u}\|_2 / \sqrt{n}$  και  $\|\exp(\mathbf{u})\|_\infty$ . Πόσες επαναλήψεις χρειάζεται η μέθοδος Newton για να συγκλίνει;

9. Επαναλάβετε τη διαδικασία που περιγράφεται στην Άσκηση 8, αυτή τη φορά χρησιμοποιώντας για την επίλυση των γραμμικών συστημάτων που προκύπτουν μια μέθοδο υπόχωρου Krylov της επιλογής σας από εκείνες που περιγράφηκαν στην Ενότητα 7.5. Αντί να απαιτήσετε να συγκλίνει ο επαναληπτικός γραμμικός επιλυτής με μια αυστηρή ανοχή, επιβάλλετε ένα «χαλαρό» κριτήριο τερματισμού: Ο επιλυτής τερματίζεται όταν η τιμή του σχετικού υπολοίπου γίνει μικρότερη από 0.01. Αυτό είναι ένα παράδειγμα μιας μη ακριβούς μεθόδου Newton. Σχολιάστε τον ρυθμό σύγκλισης και το συνολικό υπολογιστικό κόστος.

[Προειδοποίηση: Αυτό το ερώτημα είναι πιο δύσκολο από τα υπόλοιπα. Δείτε την Άσκηση 7.25 πριν αποπειραθείτε να απαντήσετε.]

10. Δείξτε ότι η συνάρτηση

$$\phi(\mathbf{x}) = x_1^2 - x_2^4 + 1$$

έχει σαγματικό σημείο στην αρχή των αξόνων, δηλαδή η αρχή των αξόνων είναι κρίσιμο σημείο το οποίο δεν είναι ούτε ελάχιστο ούτε μέγιστο.

Τι θα συμβεί αν για την εύρεση αυτού του σαγματικού σημείου εφαρμοστεί η μέθοδος Newton;

11. Ποιο από τα συστήματα μη γραμμικών εξισώσεων της Άσκησης 1 μπορεί να εκφραστεί ως  $\nabla \phi(\mathbf{x}) = \mathbf{0}$  για κάποια συνάρτηση  $\phi$  που πρέπει να ελαχιστοποιηθεί;

12. Η συνάρτηση Rosenbrock,  $\phi(\mathbf{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$ , έχει μοναδικό ελαχιστοποιητή ο οποίος μπορεί να βρεθεί με απλή εξέταση.

Εφαρμόστε τις μεθόδους Newton και BFGS –και τις δύο με ασθενή ευθύγραμμη αναζήτηση– ξεκινώντας από το  $\mathbf{x}_0 = (0, 0)^T$ . Συγκρίνετε την απόδοσή τους.

13. Θεωρήστε την ελαχιστοποίηση της συνάρτησης  $\phi(\mathbf{x}) = \mathbf{c}^T \mathbf{x} + \frac{1}{2} \mathbf{x}^T H \mathbf{x}$ , όπου  $\mathbf{c} = (5.04, -59.4, 146.4, -96.6)^T$  και

$$H = \begin{pmatrix} 0.16 & -1.2 & 2.4 & -1.4 \\ -1.2 & 12.0 & -27.0 & 16.8 \\ 2.4 & -27.0 & 64.8 & -42.0 \\ -1.4 & 16.8 & -42.0 & 28.0 \end{pmatrix}$$

Δοκιμάστε τις μεθόδους Newton και BFGS, ξεκινώντας από το  $\mathbf{x}_0 = (-1, 3, 3, 0)^T$ . Εξηγήστε γιατί η μέθοδος BFGS απαιτεί πολύ περισσότερες επαναλήψεις.

14. Θεωρήστε το μη γραμμικό πρόβλημα ελάχιστων τετραγώνων της ελαχιστοποίησης της συνάρτησης

$$\phi(\mathbf{x}) = \frac{1}{2} \|\mathbf{g}(\mathbf{x}) - \mathbf{b}\|^2$$

(α) Δείξτε ότι

$$\nabla \phi(\mathbf{x}) = A(\mathbf{x})^T (\mathbf{g}(\mathbf{x}) - \mathbf{b})$$

όπου  $A$  είναι η ιακωβιανή μήτρα, διαστάσεων  $m \times n$ , της συνάρτησης  $\mathbf{g}$ .

(β) Δείξτε ότι

$$\nabla^2 \phi(\mathbf{x}) = A(\mathbf{x})^T A(\mathbf{x}) + L(\mathbf{x})$$

όπου  $L$  είναι μια μήτρα διαστάσεων  $n \times n$  με στοιχεία

$$L_{i,j} = \sum_{k=1}^m \frac{\partial^2 g_k}{\partial x_i \partial x_j} (g_k - b_k)$$

[Θα ήταν καλό να ελέγξετε πρώτα την παράγωγο  $\frac{\partial \phi}{\partial x_i}$  για σταθερό  $i$ : στη συνέχεια μπορείτε να ελέγξετε την παράγωγο  $\frac{\partial^2 \phi}{\partial x_i \partial x_j}$  για σταθερό  $j$ .]

15. Στην Άσκηση 6.4 σας ζητήθηκε να επινοήσετε ένα τέχνασμα μεταβλητού μετασχηματισμού προκειμένου να επιλύσετε ένα απλό μη γραμμικό πρόβλημα προσαρμογής δεδομένων ως γραμμικό.

Επιλύστε το ίδιο πρόβλημα ως προς τις μεταβλητές που δίνονται, δηλαδή  $\mathbf{x} = (\gamma_1, \gamma_2)^T$ , χωρίς τον ειδικό μετασχηματισμό, χρησιμοποιώντας τη μέθοδο Gauss-

Newton. Η επαναληπτική διαδικασία σας πρέπει να σταματάει όταν ισχύει  $\|\mathbf{p}_k\| < \text{tol}(\|\mathbf{x}_k\| + 1)$ . Πειραματιστείτε με μερικές αρχικές εικασίες και ανοχές  $\text{tol}$ . Τι παρατηρείτε;

Αν έχετε λύσει και την Άσκηση 6.4, μπορείτε να χρησιμοποιήσετε τη λύση που προέκυψε εκεί ως αρχική εικασία για τη μέθοδο Gauss-Newton. Αξιολογήστε τη σχετική ποιότητα των λύσεων συγκρίνοντας τα υπόλοιπα που προκύπτουν.

16. Επιλύστε το πρόβλημα της Άσκησης 15 (δηλαδή το πρόβλημα που δίνεται στην Άσκηση 6.4 χωρίς τον ειδικό μετασχηματισμό που εφαρμόζεται εκεί) χρησιμοποιώντας τη μέθοδο Newton. Εφαρμόστε το ίδιο κριτήριο τερματισμού όπως στην Άσκηση 15 και συγκρίνετε το πλήθος των επαναλήψεων των δύο μεθόδων (δηλαδή συγκρίνετε τη μέθοδο Newton με τη μέθοδο Gauss-Newton) για τιμές ανοχής  $\text{tol} = 10^{-6}$  και  $\text{tol} = 10^{-10}$ . Σχολιάστε αυτά που παρατηρείτε.
17. Σε αυτή την άσκηση σας ζητείται να βρείτε μια συνάρτηση  $u(t)$  στο διάστημα  $[0, 1]$  όταν δίνονται «ενθύρυβα» δεδομένα στα σημεία  $t_i = ih$ ,  $i = 0, 1, \dots, N$ , με  $N = 1/h$ . Επειδή οι τιμές των δεδομένων περιέχουν και θύρυβο, δεν μπορούμε απλώς να θέσουμε  $u(t_i) \equiv u_i = b_i$ ; Γνωρίζοντας ότι η  $u(t)$  πρέπει να είναι τμηματικά λεία, προσθέτουμε έναν όρο ομαλοποίησης (regularization) προκειμένου να επιβάλουμε ποινές για την υπερβολική τραχύτητα της  $u$ . Συνεπώς, για το άγνωστο διάνυσμα  $\mathbf{u} = (u_0, u_1, \dots, u_N)^T$  λύνουμε το πρόβλημα

$$\min \phi_2(\mathbf{u}) = \frac{h}{2} \sum_{i=1}^N \frac{1}{2} \left[ (u_i - b_i)^2 + (u_{i-1} - b_{i-1})^2 \right] + \frac{\beta h}{2} \sum_{i=1}^N \left( \frac{u_i - u_{i-1}}{h} \right)^2$$

- (a) Γράψτε την κλίση και την εσσιανή μήτρα αυτής της αντικειμενικής συνάρτησης. Για να περιγράψετε τον όρο ομαλοποίησης χρησιμοποιήστε τη μήτρα

$$W = \frac{1}{\sqrt{h}} \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{N \times (N+1)}$$

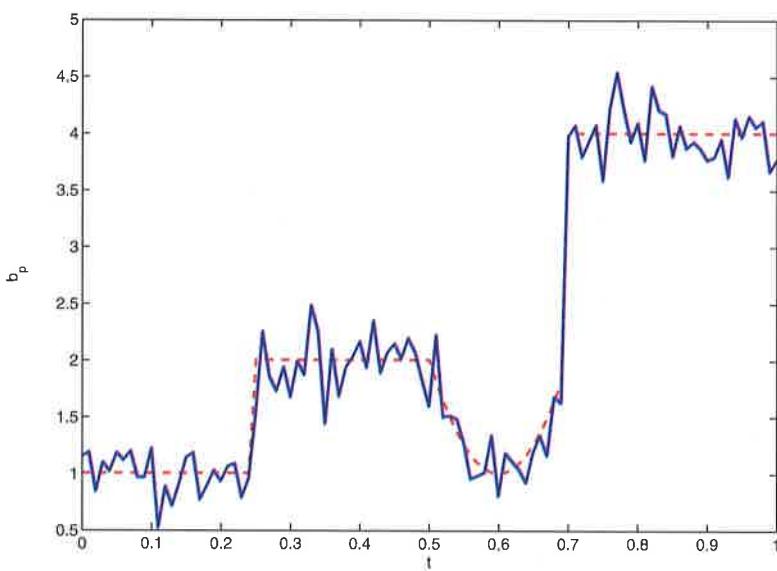
- (β) Επιλύστε αυτό το πρόβλημα αριθμητικά για τα ακόλουθα στιγμιότυπά του. Για να «κατασκευάσετε δεδομένα» για κάποιο συγκεκριμένο  $N$ , ξεκινήστε με

$$b_p(t) = \begin{cases} 1, & 0 \leq t < 0.25 \\ 2, & 0.25 \leq t < 0.5 \\ 2 - 100(t - 0.5)(0.7 - t), & 0.5 \leq t < 0.7 \\ 4, & 0.7 \leq t \leq 1 \end{cases}$$

Υπολογίστε την τιμή της συνάρτησης αυτής στα σημεία του πλέγματος και μετά προσθέστε θόρυβο ως εξής:

```
noisev = randn(size(b_p)) * mean(abs(b_p)) * noise;
data = b_p + noisev;
```

Οι τιμές που προκύπτουν είναι τα δεδομένα που «βλέπει» το πρόγραμμά σας. Ένα παράδειγμα δίνεται στο Σχήμα 9.13.



**Σχήμα 9.13:** Απεικόνιση της «ενθόρυβης» συνάρτησης (με συνεχή μπλε γραμμή) και της συνάρτησης που πρέπει να ανακτηθεί (με κόκκινη διακεκομένη γραμμή) για την Άσκηση 17.

Σχεδιάστε τα δεδομένα αυτά και την ανακτηθείσα καμπύλη  $\mathbf{u}$  για τις τιμές παραμέτρων  $(\beta, noise) = (10^{-3}, 0.01), (10^{-3}, 0.1), (10^{-4}, 0.01)$  και  $(10^{-4}, 0.1)$ . Πραγματοποιήστε δοκιμές για  $N = 64$  ή  $N = 128$ . Τι μπορείτε να παρατηρήσετε;

18. Σε συνέχεια της Άσκησης 17, αν η συνάρτηση δεδομένων περιέχει ασυνέχειες αλμάτων (jump discontinuities), αυτές οι ασυνέχειες θα θολώσουν από την ομαλοποίηση που προτείνεται στην Άσκηση 17. Επομένως, θεωρήστε αυτ' αυ-

τού το πρόβλημα ελαχιστοποίησης

$$\min \phi_1(\mathbf{u}) = \frac{h}{2} \sum_{i=1}^N \frac{1}{2} \left[ (u_i - b_i)^2 + (u_{i-1} - b_{i-1})^2 \right] + \gamma h \sum_{i=1}^N \sqrt{\left( \frac{u_i - u_{i-1}}{h} \right)^2 + \varepsilon}$$

όπου  $\varepsilon = 10^{-6}$ , για παράδειγμα.

- (α) Έστω ότι  $T_{1,h}$  είναι ο όρος που πολλαπλασιάζεται με το  $\gamma$  στην αντικειμενική συνάρτηση  $\phi_1$ . Δείξτε ότι

$$\frac{\partial T_{1,h}}{\partial u_j} = \frac{u_j - u_{j-1}}{\sqrt{(u_j - u_{j-1})^2 + \varepsilon h^2}} + \frac{u_j - u_{j+1}}{\sqrt{(u_j - u_{j+1})^2 + \varepsilon h^2}}$$

Επιπλέον, θέτοντας

$$\hat{D} = \text{diag} \left\{ h / \sqrt{(u_j - u_{j-1})^2 + \varepsilon h^2} \right\}, \quad \hat{B} = \sqrt{\hat{D}}$$

μπορούμε να γράψουμε

$$\nabla T_{1,h} = W^T \hat{B}^T \hat{B} W \mathbf{u} = W^T \hat{D} W \mathbf{u}$$

- (β) Μια μέθοδος για την επίλυση αυτού του προβλήματος είναι προφανής: Στην αρχή κάθε επανάληψης σταθεροποιούμε το  $\hat{D}$  με βάση την τρέχουσα προσεγγιστική τιμή, και μετά εφαρμόζουμε τους συνήθεις αλγόριθμους για ένα γραμμικό συναρτησοειδές σταθμισμένων ελάχιστων τετραγώνων (weighted least squares). Η συγκεκριμένη μέθοδος είναι γνωστή ως επαναλαμβανόμενα ελάχιστα τετράγωνα (iterated least squares).

Χρησιμοποιήστε αυτή τη μέθοδο για να επιλύσετε τα ίδια προβλήματα όπως στο προηγούμενο ερώτημα (δηλαδή τα ίδια «συνθετικά δεδομένα»), για  $\gamma = 10^{-2}$  και  $\gamma = 10^{-3}$ . Σχολιάστε αυτά που παρατηρείτε.

19. Θεωρήστε τη μήτρα σαγματικού σημείου

$$K = \begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix}$$

όπου η μήτρα  $H$  είναι συμμετρική θετικά ημιορισμένη και η μήτρα  $A$  είναι πλήρους τάξης γραμμών.

- (α) Δείξτε ότι η  $K$  είναι μη ιδιάζουσα αν ισχύει  $\mathbf{y}^T H \mathbf{y} \neq 0$  για όλα τα  $\mathbf{y} \in \text{null}(A)$ ,  $\mathbf{y} \neq \mathbf{0}$ .

- (β) Δείξτε με ένα παράδειγμα ότι η  $K$  είναι συμμετρική αλλά μη ορισμένη, δηλαδή ότι έχει και θετικές και αρνητικές ιδιοτιμές.
20. Για τη μήτρα σαγματικού σημείου της Άσκησης 19, αν ήσασταν υποχρεωμένοι να λύσετε το αντίστοιχο σύστημα εξισώσεων επαναληπτικά, ποια από τις μεθόδους που περιγράφηκαν στις Ενότητες 7.4 και 7.5 θα επιλέγατε;
21. Θεωρήστε το πρόβλημα

$$\min_{\mathbf{y}} \|\mathbf{y}\|_p$$

$$\text{έτσι ώστε } J\mathbf{y} = \mathbf{b}$$

όπου η μήτρα  $J$  είναι διαστάσεων  $m \times n$ ,  $m \leq n$ .

Επινοήστε ένα παράδειγμα στο οπόιο η επίλυση με  $p = 2$  έχει όντως περισσότερο νόημα από την επίλυση με  $p = 1$ .

22. Με τον συμβολισμό της Άσκησης 21, για  $p = 2$  το πρόβλημα μπορεί να επιλυθεί όπως στο Παράδειγμα 9.15 με τη χρήση διάσπασης ιδιαζουσών τιμών. Ωστόσο μπορεί επίσης να επιλυθεί με τη χρήση των τεχνικών της βελτιστοποίησης με περιορισμούς:
- (α) Εξηγήστε γιατί μπορεί να δικαιολογηθεί η αντικατάσταση της αντικειμενικής συνάρτησης με το  $\frac{1}{2}\|\mathbf{y}\|_2^2$ .
  - (β) Διατυπώστε τις συνθήκες KKT του προβλήματος βελτιστοποίησης με περιορισμούς που προκύπτει, ώστε να πάρετε ένα γραμμικό σύστημα εξισώσεων.
  - (γ) Επινοήστε ένα παράδειγμα και επιλύστε το πρόβλημα χρησιμοποιώντας τη μέθοδο που μόλις αναπτύξατε και τη μέθοδο του Παραδείγματος 9.15. Συγκρίνετε τα αποτελέσματα και σχολιάστε.

## 9.5 Πρόσθετες σημειώσεις

Η βελτιστοποίηση είναι ένα τεράστιο πεδίο. Η *Mathematical Optimization Society* έχει ως αποστολή της τη διερεύνηση ακαδημαϊκών πτυχών της βελτιστοποίησης, αλλά υπάρχουν πολλά ζητήματα πέραν εκείνων που συνήθως την απασχολούν. Παραδόξως, παλαιότερα το όνομά της ήταν *Mathematical Programming Society*, αλλά το 2010 οι υπεύθυνοι αποφάσισαν να το αλλάξουν. Στο παρόν κεφάλαιο έχουμε περιγράψει με συντομία μόνο ένα μέρος αυτού του μεγαθήριου και, πιο συγκεκριμένα, εκείνο που σχετίζεται πιο άμεσα με άλλα θέματα που αναπτύσσονται σε αυτό το βιβλίο.

Υπάρχουν πολλά διδακτικά βιβλία με θέμα την αριθμητική βελτιστοποίηση. Ένα από αυτά το οποίο καλύπτει τα σχετικά θέματα με τρόπο παραπλήσιο του δικού μας είναι το βιβλίο των Nocedal και Wright [57]. Ένα πιθανώς πιο εύληπτο σύγραμμα, που μάλλον απευθύνεται περισσότερο σε όσους δεν είναι και τόσο εξουκειωμένοι με τους υπολογιστές, είναι το βιβλίο των Griva, Nash και Sofer [34]. Τα βιβλία των Fletcher [25], Dennis και Schnabel [22], και Gill, Murray και Wright [29] συγκαταλέγονται στα αγαπημένα μας κλασικά συγγράμματα.

Υπάρχουν πολλά εξαιρετικά πακέτα λογισμικού που προσφέρουν λύσεις για μια μεγάλη ποικιλία προβλημάτων βελτιστοποίησης με περιορισμούς. Ένα παράδειγμα τέτοιου λογισμικού είναι και το CPLEX: <http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/>.

Η μέθοδος πρόβλεψη-διόρθωσης που περιγράφηκε στην Ενότητα 9.3 αποτελεί παραλλαγή της περιγραφής στο [57] ενός αλγόριθμου που επινόησε ο Mehrotra [53].

Στο πολυμελετημένο πεδίο της συνεχούς βελτιστοποίησης με περιορισμούς εξακολουθεί σήμερα να παρατηρείται μεγάλη κινητικότητα. Ένας από τους αναδυόμενους τομείς του ασχολείται με πολύ μεγάλα αλλά δομημένα προβλήματα όπου οι περιορισμοί είναι μερικές διαφορικές εξισώσεις: δείτε, για παράδειγμα, το άρθρο του Biegler κ.ά. [8].

Ένας άλλος ερευνητικός τομέας που αποτελεί επίκεντρο του ενδιαφέροντος είναι οι τεχνικές αραιής ανακατασκευής (sparse reconstruction) και αραιής a priori κατανομής (sparse prior), οι οποίες στηρίζονται σε μεγάλο βαθμό στις αξιοσημείωτες ιδιότητες που έχει η βελτιστοποίηση με την  $\ell_1$ -νόρμα για την εύρεση αραιών λύσεων καθώς και με άλλες  $\ell_p$ -νόρμες για  $p < 1$ . Μια γεύση αυτών πήρατε στο Παράδειγμα 9.15. Ουσιαστικά, παρέχουν συχνά αξιόπιστες μεθόδους προσέγγισης για το εκθετικά δύσκολο συνδυαστικό πρόβλημα της επιλογής ενός μικρού πλήθους συναρτήσεων βάσης, ή στοιχείων της λύσης, για την περιγραφή ενός παρατηρούμενου φαινομένου με αποδεκτή ακρίβεια. Όπως πολλοί άλλοι, ο Mallat [52] αποσαφηνίζει τη χρήση αυτού του εργαλείου στην επεξεργασία σήματος.

Τα γραμμικά συστήματα που προκύπτουν από προβλήματα βελτιστοποίησης με περιορισμούς συχνά αποκαλούνται και συστήματα σαγματικού σημείου, κυρίως από ερευνητές που δεν επικεντρώνονται αποκλειστικά στη βελτιστοποίηση. Τις δύο τελευταίες δεκαετίες έχει παρατηρηθεί μια κατακόρυφη αύξηση του ενδιαφέροντος για τις μεθόδους επίλυσης τέτοιων προβλημάτων. Μια διεξοδική μελέτη των επαναληπτικών επιλυτών για συστήματα σαγματικού σημείου παρουσιάζεται στο άρθρο των Benzi, Golub και Liesen [6].

Όπως έχουμε ήδη επισημάνει, η βελτιστοποίηση είναι ένα τεράστιο πεδίο, οπότε είναι λογικό να έχουμε παραλείψει αρκετές κατηγορίες προβλημάτων βελτι-

στοποίησης στο βιβλίο που κρατάτε στα χέρια σας. Σε αυτές περιλαμβάνονται τα προβλήματα διακριτής βελτιστοποίησης, όπου κάποιες μεταβλητές περιορίζονται να παίρνουν μόνο μία από πολλές διακριτές τιμές –π.χ. μια μεταβλητή απόφασης (decision variable)  $x_1$  μπορεί να πάρει μόνο την τιμή 0 για το «όχι» ή την τιμή 1 για το «ναι»— και η στοχαστική βελτιστοποίηση. Επίσης, δεν έχουμε αναφερθεί σε βάθος στα προβλήματα ολικής βελτιστοποίησης. Τέλος, όλες οι μέθοδοι που έχουμε εξετάσει υποθέτουν ότι είναι διαθέσιμη η κλίση της αντικειμενικής συνάρτησης. Υπάρχει ένας ξεχωριστός ερευνητικός κλάδος για μεθόδους που λύνουν προβλήματα τα οποία δεν ικανοποιούν αυτή την υπόθεση.



## Κεφάλαιο 10

# Πολυωνυμική παρεμβολή

Τα πολυώνυμα παρεμβολής σπάνια αποτελούν το τελικό προϊόν μιας αριθμητικής διαδικασίας. Η σπουδαιότητά τους έγκειται περισσότερο στο ότι είναι τα δομικά στοιχεία για άλλους, πιο πολύπλοκους αλγόριθμους στην παραγώγιση, στην ολοκλήρωση, στην επίλυση διαφορικών εξισώσεων, στη θεωρία προσέγγισης (approximation theory) γενικά, και σε άλλους τομείς. Συνεπώς, η πολυωνυμική παρεμβολή είναι κάτι που προκύπτει συχνά· πράγματι, είναι μια από τις ευρέως διαδεδομένες εργασίες, τόσο στον σχεδιασμό αριθμητικών αλγόριθμων όσο και στην ανάλυσή τους. Η σπουδαιότητά της και ο κεντρικός ρόλος της δικαιολογούν το μεγάλο μέγεθος αυτού του κεφαλαίου.

Στην Ενότητα 10.1 θα ξεκινήσουμε με μια γενική περιγραφή διαδικασιών προσέγγισης ως προς μία ανεξάρτητη μεταβλητή, καταλήγοντας στην πολυωνυμική παρεμβολή ως μια τέτοια θεμελιώδη οικογένεια τεχνικών. Στις Ενότητες 10.2, 10.3 και 10.4 θα δούμε τουλάχιστον τρεις διαφορετικές μορφές (διαφορετικές βάσεις) πολυωνύμων παρεμβολής. Όλες αυτές οι μορφές έχουν θεμελιώδη σημασία και χρησιμοποιούνται εκτενώς στην πρακτική κατασκευή αριθμητικών αλγόριθμων.

Στην Ενότητα 10.5 θα βρούμε εκτιμήσεις και φράγματα για το σφάλμα στην πολυωνυμική παρεμβολή. Αν η επιλογή των θέσεων για τα δεδομένα παρεμβολής μπορεί να γίνει από τον χρήστη, μια πλεονεκτική επιλογή είναι ένα ειδικό σύνολο τετμημένων (ιόμβων) που ονομάζονται σημεία Chebyshev, το οποίο θα περιγράψουμε στην Ενότητα 10.6. Τέλος, στην Ενότητα 10.7 θα μελετήσουμε την περίπτωση στην οποία είναι διαθέσιμες για παρεμβολή όχι μόνο οι τιμές μιας συνάρτησης αλλά και οι τιμές των παραγώγων.

## 10.1 Γενική προσέγγιση και παρεμβολή

Η παρεμβολή (interpolation) αποτελεί ειδική περίπτωση της προσέγγισης (approximation). Σε αυτή την ενότητα θα εξετάσουμε διαφορετικά περιβάλλοντα στα οποία προκύπτουν προβλήματα προσέγγισης, θα αιτιολογήσουμε την ανάγκη για την εύρεση προσεγγιστικών συναρτήσεων, θα περιγράψουμε μια γενική μορφή για συναρτήσεις παρεμβολής καθώς και σημαντικές ειδικές περιπτώσεις, και θα ολοκληρώσουμε με την πολυωνυμική παρεμβολή.

### Διακριτή και συνεχής προσέγγιση στη μία διάσταση

Οι μέθοδοι προσέγγισης μπορούν να χωριστούν σε δύο κατηγορίες με βάση τον τύπο του προβλήματος:

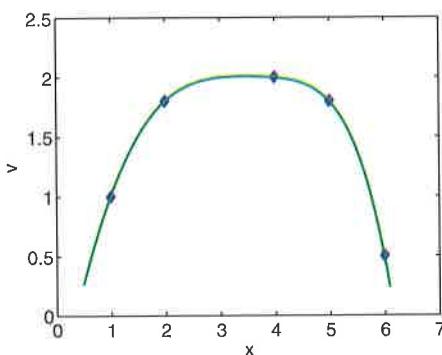
#### 1. Προσαρμογή δεδομένων (πρόβλημα διακριτής προσέγγισης):

Δίνεται ένα σύνολο σημείων δεδομένων  $\{(x_i, y_i)\}_{i=0}^n$  και πρέπει να βρεθεί μια λογική συνάρτηση  $v(x)$  που να προσαρμόζεται στα σημεία δεδομένων.

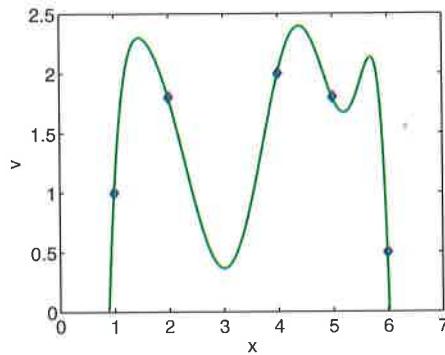
Αν τα δεδομένα είναι ακριβή, τότε ίσως έχει νόημα η απαίτηση η  $v(x)$  να παρεμβάλλεται στα δεδομένα, δηλαδή η καμπύλη να διέρχεται ακριβώς από τα δεδομένα, ικανοποιώντας την εξίσωση

$$v(x_i) = y_i, \quad i = 0, 1, \dots, n$$

Δείτε το Σχήμα 10.1.



(α) Λογική.



(β) Παράλογη.

**Σχήμα 10.1:** Διαφορετικές καμπύλες παρεμβολής που διέρχονται από το ίδιο σύνολο σημείων.

## 2. Προσέγγιση συναρτήσεων:

Δίνεται ρητά ή ορίζεται μόνο έμμεσα μια πολύπλοκη συνάρτηση  $f(x)$  και πρέπει να βρεθεί μια απλούστερη συνάρτηση  $v(x)$  η οποία να προσεγγίζει την  $f(x)$ .

Για παράδειγμα, ας υποθέσουμε ότι πρέπει να βρούμε γρήγορα μια προσεγγιστική τιμή για το  $\sin(1.2)$  έχοντας στη διάθεσή μας μόνο μια απλή αριθμομηχανή (το 1.2 αντιστοιχεί σε ακτίνια, όχι μοίρες). Από τη βασική τριγωνομετρία γνωρίζουμε τις τιμές του  $\sin(x)$  για  $x = 0, \pi/6, \pi/4, \pi/3$  και  $\pi/2$ : Πώς μπορούμε να χρησιμοποιήσουμε αυτές τις τιμές για να εκτιμήσουμε το  $\sin(1.2)$ ;

Ας δούμε ένα άλλο παράδειγμα. Έστω ότι έχουμε ένα πολύπλοκο, ακριβό πρόγραμμα που υπολογίζει το τελικό σημείο (για παράδειγμα, το σημείο προσγείωσης) της τροχιάς ενός διαστημικού λεωφορείου για κάθε δοθείσα τιμή μιας ορισμένης παραμέτρου ελέγχουν. Εκτελούμε αυτόν τον υπολογισμό, δηλαδή καλούμε το πρόγραμμα, για αρκετές τιμές της παραμέτρου. Στη συνέχεια, όμως, μπορεί να θέλουμε να χρησιμοποιήσουμε αυτές τις υπολογισμένες πληροφορίες προκειμένου να αποκτήσουμε μια ιδέα για το σημείο προσγείωσης που προκύπτει για άλλες τιμές της παραμέτρου ελέγχουν –χωρίς να πρέπει να καταφύγουμε στον πλήρη υπολογισμό για κάθε τιμή της παραμέτρου.

Οι τεχνικές παρεμβολής για την προσαρμογή συναρτήσεων κατά αυτόν τον τρόπο είναι πανομοιότυπες με εκείνες για την προσαρμογή δεδομένων, αφού όμως καθορίζουμε πρώτα τα σημεία δεδομένων  $\{(x_i, y_i = f(x_i))\}_{i=0}^n$ . Η διαφορά μεταξύ της παρεμβολής συναρτήσεων και της παρεμβολής προσαρμογής δεδομένων είναι ότι στην πρώτη

- έχουμε κάποια ελευθερία να επιλέξουμε τα  $x_i$  με έξυπνο τρόπο και
- ενδέχεται να είμαστε σε θέση να λάβουμε υπόψη το καθολικό σφάλμα παρεμβολής.

## Η ανάγκη για παρεμβολή

Γιατί όμως θέλουμε να βρούμε μια προσεγγιστική συνάρτηση  $v(x)$  γενικά;

- Για σκοπούς πρόβλεψης: Μπορούμε να χρησιμοποιήσουμε τη  $v(x)$  για να βρίσκουμε προσεγγιστικές τιμές της υποκείμενης συνάρτησης σε σημεία  $x$  διαφορετικά από τις τετμημένες  $x_0, \dots, x_n$  των δεδομένων.

Αν το  $x$  περιέχεται στο μικρότερο διάστημα που περιλαμβάνει όλες τις τετμημένες των δεδομένων, αυτή τη διαδικασία ονομάζεται παρεμβολή: αν το  $x$  δεν ανήκει στο διάστημα, έχουμε παρεκβολή (extrapolation), ή προέκταση. Για παράδειγμα, μπορεί να υπάρχουν δεδομένα σχετικά με την απόδοση μιας μετοχής

στο τέλος κάθε εβδομάδας διαπραγμάτευσης κατά τη διάρκεια του προηγούμενου έτους. Η παρεμβολή ώστε να εκτιμηθεί η τιμή της μετοχής για άλλες ημέρες του προηγούμενου έτους είναι ένα σχετικά ασφαλές εγχείρημα. Η παρεκβολή ώστε να εκτιμηθεί η τιμή της μετοχής σε κάποια χρονική στιγμή του επόμενου έτους είναι ένα πολύ πιο ριψοκίνδυνο εγχείρημα (αν και πιθανώς πιο ενδιαφέρον).

- Για σκοπούς χειρισμού: Ένα παράδειγμα είναι η εύρεση προσεγγίσεων για παραγώγους και ολοκληρώματα της υποκείμενης συνάρτησης.

Η συνάρτηση παρεμβολής πρέπει όχι μόνο να επιτρέπει τον εύκολο υπολογισμό και χειρισμό της, αλλά να είναι και «λογική». Δηλαδή, πρέπει να μοιάζει με μια καμπύλη την οποία θα μπορούσαμε όντως να σχεδιάσουμε έτσι ώστε να διέρχεται από τα σημεία –δείτε το Σχήμα 10.1. Όμως το τι ακριβώς είναι «λογικό» εξαρτάται από το ευρύτερο πλαίσιο και είναι δύσκολο να οριστεί με ακριβείς όρους για γενική χρήση.

## Συναρτήσεις παρεμβολής και η αναπαράστασή τους

Γενικά υποθέτουμε ότι όλες οι συναρτήσεις παρεμβολής (ή, πιο γενικά, προσέγγισης)  $v(x)$  έχουν γραμμική μορφή.<sup>49</sup> Επομένως, γράφουμε

$$v(x) = \sum_{j=0}^n c_j \phi_j(x) = c_0 \phi_0(x) + \cdots + c_n \phi_n(x)$$

όπου  $\{c_j\}_{j=0}^n$  είναι άγνωστοι συντελεστές, ή **παράμετροι**, που καθορίζονται από τα δεδομένα και  $\{\phi_j(x)\}_{j=0}^n$  είναι προκαθορισμένες **συναρτήσεις βάσης** (basis functions). Γι' αυτές τις συναρτήσεις βάσης θεωρούμε περαιτέρω ότι είναι γραμμικά ανεξάρτητες, το οποίο σημαίνει ότι αν βρεθούν συντελεστές  $\{c_j\}_{j=0}^n$  τέτοιοι ώστε να ισχύει  $v(x) = 0$  για κάθε  $x$ , τότε οι συντελεστές πρέπει να μηδενίζονται και οι ίδιοι:  $c_j = 0$ ,  $j = 0, 1, \dots, n$ .

Παρατηρήστε ότι έχουμε κάνει εξ ορισμού την εξής υπόθεση: Το πλήθος των συναρτήσεων βάσης ισούται με το πλήθος  $n+1$  των σημείων δεδομένων. Αν οι συναρτήσεις βάσης είναι λιγότερες από τα σημεία δεδομένων, δεν μπορούμε να ελπίζουμε ότι θα παρεμβάλουμε όλες τις τιμές δεδομένων και συνήθως καταφεύγουμε σε μια μέθοδο ελάχιστων τετραγώνων όπως αυτές που περιγράφαμε στο Κεφάλαιο 6. Στα Κεφάλαια 12 και 13 θα μελετήσουμε άλλες τεχνικές εύρεσης προσεγγιστικών συναρτήσεων που δεν προσαρμόζονται ακριβώς στα δεδομένα. Γενικά, οι συνθήκες παρεμβολής μπορούν να γραφούν ως  $n+1$  γραμμικές σχέσεις για τους  $n+1$

<sup>49</sup> Σημειώστε ότι η μορφή της συνάρτησης παρεμβολής είναι γραμμική υπό την έννοια ότι αποτελεί γραμμικό συνδυασμό συναρτήσεων βάσης σε κάποιον κατάλληλο χώρο, όχι ότι η  $v(x)$  καθαυτή είναι γραμμική συνάρτηση του  $x$ .

άγνωστους συντελεστές. Το γραμμικό σύστημα εξισώσεων που προκύπτει είναι

$$\begin{pmatrix} \phi_0(x_0) & \phi_1(x_0) & \phi_2(x_0) & \cdots & \phi_n(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \phi_2(x_1) & \cdots & \phi_n(x_1) \\ \vdots & \vdots & \vdots & & \vdots \\ \phi_0(x_n) & \phi_1(x_n) & \phi_2(x_n) & \cdots & \phi_n(x_n) \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix} \quad (10.1)$$

Μπορεί να μην χρειαστεί να ορίσουμε και να λύσουμε το σύστημα (10.1) σε μια δεδομένη περίσταση, αλλά αυτή η επιλογή είναι πάντα εφικτή λόγω της υπόθεσης που έχουμε κάνει για τη γραμμική αναπαράσταση της  $v(x)$  σε σχέση με τις συναρτήσεις βάσης.

Ακολουθούν μερικά συνήθη παραδείγματα συναρτήσεων παρεμβολής.

- Στο κεφάλαιο αυτό θα ασχοληθούμε με την πολυωνυμική παρεμβολή

$$v(x) = \sum_{j=0}^n c_j x^j = c_0 + c_1 x^1 + \cdots + c_n x^n$$

Αυτή η απλούστερη και πιο οικεία μορφή αναπαράστασης ενός πολυωνύμου συνεπάγεται την επιλογή μιας μονωνυμικής βάσης

$$\phi_j(x) = x^j, \quad j = 0, 1, \dots, n$$

αλλά, όπως θα δούμε, υπάρχουν και άλλες επιλογές.

- Στο επόμενο κεφάλαιο θα περιγράψουμε την τμηματικά πολυωνυμική παρεμβολή (piecewise polynomial interpolation), η οποία βασίζεται στην εφαρμογή της πολυωνυμικής παρεμβολής κατά «τμήματα» παρά σε ολόκληρο το διάστημα που δίνεται.
- Η τριγωνομετρική παρεμβολή (trigonometric interpolation) είναι επίσης εξαιρετικά χρήσιμη, ειδικά στην επεξεργασία σήματος καθώς και στην περιγραφή κυματικών και άλλων περιοδικών φαινομένων. Για παράδειγμα, θεωρήστε τη συνάρτηση

$$\phi_j(x) = \cos(jx), \quad j = 0, 1, \dots, n$$

Στο Κεφάλαιο 13 θα αναπτύξουμε διεξοδικά πιο γενικές επιλογές ακολουθώντας αυτή τη λογική.

Γενικά, είναι σημαντικό να διακρίνουμε δύο στάδια στη διαδικασία παρεμβολής:

- Κατασκευή** της συνάρτησης παρεμβολής. Πρόκειται για πράξεις που είναι ανεξάρτητες από τα σημεία στα οποία μπορεί κατόπιν να υπολογίσουμε την τιμή της  $v(x)$ . Ένα σχετικό παράδειγμα είναι ο προσδιορισμός των συντελεστών  $c_0, c_1, \dots, c_n$  για μια βάση  $\phi_0(x), \phi_1(x), \dots, \phi_n(x)$ .
- Υπολογισμός** της συνάρτησης παρεμβολής σε ένα σημείο  $x$ .

Η κατασκευή της συνάρτησης παρεμβολής γίνεται μία φορά για ένα συγκεκριμένο σύνολο δεδομένων. Μετά από αυτό, ο υπολογισμός ενδέχεται να εκτελεστεί πολλές φορές.

## Πολυωνυμική παρεμβολή

**Σημείωση:** Στο υόλοιπο αυτού του κεφαλαίου θα ασχοληθούμε αποκλειστικά με την πολυωνυμική παρεμβολή.

Ο κύριος λόγος για τον οποίο η πολυωνυμική προσέγγιση είναι επιθυμητή έγκειται στην απλότητά της. Τα πολυώνυμα

- είναι εύκολο να κατασκευαστούν και να υπολογιστούν (θυμηθείτε επίσης την ένθετη μορφή από το Παράδειγμα 1.4)·
- είναι εύκολο να αθροιστούν και να πολλαπλασιαστούν (και το αποτέλεσμα είναι επίσης πολυώνυμο)·
- είναι εύκολο να τα παραγωγίσουμε και να τα ολοκληρώσουμε (και το αποτέλεσμα είναι επίσης πολυώνυμο)· και
- έχουν πολλά διαφορετικά χαρακτηριστικά παρά την απλότητά τους.

## 10.2 Μονωνυμική παρεμβολή

Έστω ότι συμβολίζουμε ένα πολυώνυμο παρεμβολής βαθμού το πολύ  $n$  με

$$p(x) = p_n(x) = \sum_{j=0}^n c_j x^j = c_0 + c_1 x + \cdots + c_n x^n$$

Για  $n+1$  σημεία δεδομένων

$$(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$$

Θέλουμε να βρούμε  $n+1$  συντελεστές<sup>50</sup>  $c_0, c_1, \dots, c_n$  τέτοιους ώστε να ισχύει

$$p(x_i) = y_i, \quad i = 0, 1, \dots, n$$

<sup>50</sup> Θυμηθείτε ότι ένα πολυώνυμο βαθμού  $n$  δεν έχει  $n$  αλλά  $n+1$  συντελεστές.

Θα υποθέσουμε, μέχρι την Ενότητα 10.7, ότι οι τετμημένες των σημείων δεδομένων είναι διαφορετικές, δηλαδή ισχύει

$$x_i \neq x_j \quad \text{όταν} \quad i \neq j$$

**Παράδειγμα 10.1.** Έστω ότι  $n = 1$  και τα δύο σημεία δεδομένων είναι τα  $(x_0, y_0) = (1, 1)$  και  $(x_1, y_1) = (2, 3)$ . Θέλουμε να προσαρμόσουμε σε αυτά τα δύο σημεία ένα πολυώνυμο βαθμού το πολύ 1 της μορφής

$$p_1(x) = c_0 + c_1 x$$

Οι συνθήκες παρεμβολής είναι

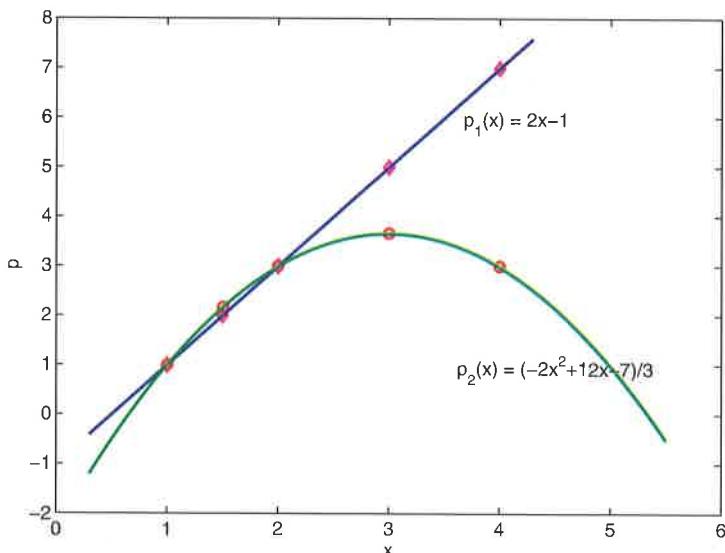
$$p_1(x_0) = c_0 + 1c_1 = 1$$

$$p_1(x_1) = c_0 + 2c_1 = 3$$

Πρόκειται για δύο γραμμικές εξισώσεις με δύο αγνώστους  $c_0$  και  $c_1$ , οι οποίες μπορούν να επιλυθούν με τη χρήση τεχνικών από την άλγεβρα του γυμνασίου. Βρίσκουμε ότι  $c_0 = -1$  και  $c_1 = 2$ , άρα

$$p_1(x) = 2x - 1$$

Αυτή η γραμμική συνάρτηση παρεμβολής απεικονίζεται στο Σχήμα 10.2.



**Σχήμα 10.2:** Τετραγωνική και γραμμική πολυώνυμη παρεμβολή.

Στη συνέχεια, έστω ότι  $n = 2$  και έστω ότι τα τρία σημεία δεδομένων είναι τα  $(1, 1)$ ,  $(2, 3)$  και  $(4, 3)$ . Τα πρώτα δύο είναι τα ίδια όπως παραπάνω, αλλά το τρίτο ζεύγος καθορίζει μια σημαντικά διαφορετική τιμή στο  $x = 4$  από εκείνη που προβλέπει το  $p_1(x)$ : Ενώ ισχύει  $p_1(4) = 7$ , εδώ ψάχνουμε ένα πολυώνυμο του οποίου η τιμή στο  $x = 4$  ισούται με 3. Άρα, θέλουμε να προσαρμόσουμε σε αυτά τα τρία σημεία ένα πολυώνυμο βαθμού το πολύ 2 της μορφής

$$p_2(x) = c_0 + c_1x + c_2x^2$$

Σημειώστε ότι οι συντελεστές  $c_0$  και  $c_1$  στο  $p_2(x)$  δεν αναμένεται γενικά να είναι οι ίδιοι όπως για το  $p_1(x)$ . Οι συνθήκες παρεμβολής είναι

$$p_2(x_0) = c_0 + 1c_1 + 1c_2 = 1$$

$$p_2(x_1) = c_0 + 2c_1 + 4c_2 = 3$$

$$p_2(x_2) = c_0 + 4c_1 + 16c_2 = 3$$

Αυτό είναι ένα γραμμικό σύστημα  $3 \times 3$  ως προς τους άγνωστους συντελεστές  $c_j$ , το οποίο σε μορφή μητρώων μπορεί να γραφεί ως

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 3 \end{pmatrix}$$

Οι ακόλουθες εντολές του MATLAB

`A = [1 1 1; 1 2 4; 1 4 16];`

`y = [1; 3; 3];`

`c = A \ y`

δίνουν το εξής αποτέλεσμα (έως ένα σφάλμα στρογγυλοποίησης):

$$c_0 = -\frac{7}{3}, \quad c_1 = 4, \quad c_2 = -\frac{2}{3}$$

Έτσι κατασκευάζεται η τετραγωνική συνάρτηση παρεμβολής. Το επιθυμητό πολυώνυμο παρεμβολής  $p_2$  είναι

$$p_2(x) = (-2x^2 + 12x - 7)/3$$

και η τιμή του μπορεί να υπολογιστεί για οποιαδήποτε τιμή του  $x$ . Για παράδειγμα, στο  $x = 3$  έχουμε

$$p_2(3) = \frac{11}{3}$$

μια τιμή που είναι αρκετά μικρότερη από το  $p_1(3) = 5$ . Παρατηρήστε επίσης στο Σχήμα 10.2 τις διαφορετικές τιμές των δύο πολυωνύμων στο  $x = 1.5$ , όπως φαίνεται από το ζεύγος των κοντινών αλλά όχι ταυτιζόμενων κόκκινων συμβόλων. ■

## Μοναδικό πολυωνύμιο παρεμβολής

Αν γενικεύσουμε το παραπάνω παράδειγμα για  $n+1$  σημεία δεδομένων, οι συνθήκες παρεμβολής οδηγούν σε ένα σύστημα  $(n+1)$  γραμμικών εξισώσεων ως προς  $(n+1)$  αγνώστους  $c_0, c_1, \dots, c_n$  που ορίζονται ως

$$\begin{pmatrix} 1 & x_0^1 & x_0^2 & \cdots & x_0^n \\ 1 & x_1^1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n^1 & x_n^2 & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Η μήτρα συντελεστών  $X$  ονομάζεται μήτρα *Vandermonde*: πιο συγκεκριμένα, είναι γνωστό από οποιοδήποτε εισαγωγικό βιβλίο της γραμμικής άλγεβρας ότι

$$\det(X) = \prod_{i=0}^{n-1} \left[ \prod_{j=i+1}^n (x_j - x_i) \right]$$

δηλαδή, η ορίζουσα είναι το γινόμενο όλων των πιθανών διαφορών ως προς τα  $x_i$ . Συνεπώς, με την προϋπόθεση ότι οι τετμημένες είναι διαφορετικές,  $\det(X) \neq 0$  και άρα η  $X$  είναι μη ιδιάζουσα.<sup>51</sup> Αυτό το επιχείρημα παρέχει μια απλή απόδειξη για το θεώρημα που παρουσιάζεται στην επόμενη σελίδα και ορίζει ότι υπάρχει μοναδικό πολυωνύμιο παρεμβολής  $p$ . Η μοναδικότητα της πολυωνυμικής παρεμβολής είναι ιδιαίτερα σημαντική εξαιτίας των διαφορετικών μορφών που μπορεί να πάρει αυτό το πολυωνύμιο: Προκύπτει πάντα το ίδιο αποτέλεσμα, ανεξάρτητα από τη μέθοδο ή βάση που χρησιμοποιείται για την εύρεση του πολυωνύμου παρεμβολής.

Αργότερα, στην Ενότητα 10.7, θα γενικεύσουμε την ύπαρξη και τη μοναδικότητα στην περίπτωση που οι τετμημένες των δεδομένων δεν είναι υποχρεωτικά διαφορετικές.

<sup>51</sup> Αυτό επαληθεύει επίσης την υπόθεση που έχουμε κάνει ότι οι συναρτήσεις βάσης, σε αυτή την περίπτωση τα μονώνυμα  $\phi_j(x) = x^j$ ,  $j = 0, 1, \dots, n$ , είναι γραμμικά ανεξάρτητες, επειδή η μήτρα είναι μη ιδιάζουσα για μια αυθαίρετη επιλογή διαφορετικών σημείων  $x_0, x_1, \dots, x_n$ .

### Συμβολισμός γινομένου.

Έστω ότι  $z_1, z_2, \dots, z_n$  είναι  $n$  αριθμητικές τιμές. Πρέπει να είστε εξοικειωμένοι με τον συμβολισμό για το άθροισμά τους, που είναι

$$\sum_{i=1}^n z_i = z_1 + z_2 + \cdots + z_n$$

Κατά τρόπο ανάλογο, ο συμβολισμός για το γινόμενό τους είναι

$$\prod_{i=1}^n z_i = z_1 \cdot z_2 \cdots z_n$$

### Θεώρημα: Ύπαρξη μοναδικού πολυωνύμου παρεμβολής.

Για οποιαδήποτε πραγματικά σημεία δεδομένων  $\{(x_i, y_i)\}_{i=0}^n$  με διαφορετικές τετμημένες  $x_i$ , υπάρχει μοναδικό πολυώνυμο  $p(x)$  βαθμού το πολύ  $n$  το οποίο ικανοποιεί τις συνθήκες παρεμβολής

$$p(x_i) = y_i, \quad i = 0, 1, \dots, n$$

### Χρήση της μονωνυμικής βάσης για την κατασκευή συναρτήσεων παρεμβολής

Μέχρι στιγμής έχουμε αποδείξει τη μοναδικότητα, αλλά έχουμε επίσης προτείνει έναν γενικό τρόπο για την εύρεση του πολυωνύμου παρεμβολής  $p(x)$ : Σχηματίζουμε τη μήτρα Vandermonde και λύνουμε ένα γραμμικό σύστημα εξισώσεων. Το μεγάλο πλεονέκτημα αυτής της τεχνικής είναι η προφανής απλότητα και ευκολία της. Ωστόσο, αν εξετάσουμε τη γενική χρήση της πολυωνυμικής παρεμβολής, θα ανακαλύψουμε και κάποια μειονεκτήματα:

- Οι υπολογιζόμενοι συντελεστές  $c_j$  δεν είναι άμεσα ενδεικτικοί της παρεμβαλλόμενης συνάρτησης, και ενδέχεται να αλλάξουν εντελώς αν θελήσουμε να τροποποιήσουμε ελαφρώς το πρόβλημα παρεμβολής: περισσότερες λεπτομέρειες γι' αυτό το θέμα θα δοθούν στις Ενότητες 10.3 και 10.4.
- Η μήτρα Vandermonde,  $X$ , συχνά είναι κακής κατάστασης (δείτε την Ενότητα 5.8), άρα οι συντελεστές που προσδιορίζονται με αυτόν τον τρόπο μπορεί να μην είναι ακριβείς.
- Αυτή η μέθοδος απαιτεί περίπου  $\frac{2}{3}n^3$  flop για την εκτέλεση της απαλοιφής Gauss (δείτε την Ενότητα 5.1) στο στάδιο της κατασκευής: υπάρχει μια άλλη μέθοδος που απαιτεί μόνο  $n^2$  flop περίπου. Το στάδιο του υπολογισμού, όμως, δεν μπορεί να επιταχυνθεί περαιτέρω: με τη χρήση της ένθετης μορφής, απαιτεί περίπου  $2n$  flop ανά σημείο υπολογισμού.

Υπάρχουν περιπτώσεις στις οποίες τα δύο τελευταία μειονεκτήματα δεν είναι σημαντικά. Πρώτον, το υψηλότερο υπολογιστικό κόστος, αν όντως προκύπτει, είναι σημαντικό μόνο όταν η τιμή του  $n$  είναι «μεγάλη», και όχι όταν ισούται με το 2 ή το 3. Επίσης, η κακή κατάσταση, δηλαδή ο ισχυρισμός ότι η βάση  $\phi_j(x) = x^j$ ,  $j = 0, 1, \dots, n$ , δεν είναι «καλή» υπό την έννοια ότι τα σφάλματα στρογγυλοποίησης μεγεθύνονται σε παράλογο βαθμό, παρατηρείται κυρίως όταν το διάστημα της παρεμβολής είναι μεγάλο ή όταν η τιμή του  $n$  δεν είναι μικρή. Αν όλα τα σημεία που μας ενδιαφέρουν, συμπεριλαμβανομένων όλων των σημείων δεδομένων  $x_i$  και των σημείων υπολογισμού  $x$ , περιέχονται σε ένα μικρό διάστημα κοντά, έστω, σε κάποιο σημείο  $\hat{x}$ , τότε η βάση

$$\{1, (x - \hat{x}), (x - \hat{x})^2, (x - \hat{x})^3\}$$

είναι απολύτως «λογική» για ένα κυβικό πολυώνυμο. Περισσότερες λεπτομέρειες γι' αυτό το θέμα μπορείτε να βρείτε στην Ενότητα 11.2. Στην πραγματικότητα, το πολυώνυμο

$$p(x) = c_0 + c_1(x - \hat{x}) + \dots + c_n(x - \hat{x})^n$$

θυμίζει ανάπτυγμα σειράς Taylor (δείτε τη σελίδα 39),

$$f(x) = f(\hat{x}) + f'(\hat{x})(x - \hat{x}) + \dots + \frac{f^{(n)}(\hat{x})}{n!}(x - \hat{x})^n + R_n(x)$$

όπου ο όρος του υπολοίπου μπορεί να γραφεί ως

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - \hat{x})^{n+1}$$

για κάποιο  $\xi$  μεταξύ των  $x$  και  $\hat{x}$ .

Στο μεγαλύτερο μέρος αυτού του κεφαλαίου οι τιμές του  $n$  έχουν μόνο ένα δεκαδικό ψηφίο. Μεγαλύτεροι βαθμοί πολυωνύμων εμφανίζονται μόνο στην Ενότητα 10.6, όπου η μονωνυμική βάση είναι όντως ανεπαρκής.

Υπάρχουν αρκετά πρόσθετα επιθυμητά χαρακτηριστικά –που συχνά είναι πολύ πιο σημαντικά από τα ζητήματα αποδοτικότητας που αναφέρθηκαν παραπάνω– τα οποία η απλή μονωνυμική βάση  $\{\phi_j(x) = x^j\}$  δεν διαθέτει. Θα τα περιγράψουμε στις δύο επόμενες ενότητες, μαζί με βάσεις που έχουν όντως τέτοια χαρακτηριστικά και οι οποίες συμβάλλουν στην πιο φυσική κατανόηση του ευρύτερου πλαισίου και των δύο προβλημάτων αλλά και της επίλυσής τους.

Ασκήσεις γι' αυτή την ενότητα: 1–3.

### 10.3 Παρεμβολή Lagrange

Οι συντελεστές  $c_j$  των πολυωνύμων  $p_1(x)$  και  $p_2(x)$  στο Παράδειγμα 10.1 δεν σχετίζονται άμεσα με τις τιμές δεδομένων  $y_j$ . Θα ήταν χρήσιμο να βρεθεί μια πολυωνυμική βάση τέτοια ώστε να ισχύει  $c_j = y_j$ , η οποία θα έδινε

$$p(x) = p_n(x) = \sum_{j=0}^n y_j \phi_j(x)$$

Ο χειρισμός μιας τέτοιας αναπαράστασης θα ήταν ιδιαίτερα εύκολος, για παράδειγμα όταν ψάχνουμε μαθηματικούς τύπους παραγώγων και ολοκληρωμάτων. Αυτό θα αποδειχθεί σε επόμενα κεφάλαια, και ιδιαίτερα στα Κεφάλαια 14 έως 16.

Μια τέτοια πολυωνυμική βάση παρέχεται από τη διαδικασία της παρεμβολής Lagrange. Για τον λόγο αυτόν ορίζουμε τα **πολυώνυμα Lagrange**,  $L_j(x)$ , τα οποία είναι πολυώνυμα βαθμού  $n$  που ικανοποιούν την εξίσωση

$$L_j(x_i) = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$$

Αν δίνονται τα δεδομένα  $y_i$  στις τετμημένες  $x_i$  όπως προηγουμένως, τότε το μοναδικό πολυώνυμο παρεμβολής βαθμού το πολύ  $n$  μπορεί πλέον να γραφεί ως

$$p(x) = \sum_{j=0}^n y_j L_j(x)$$

Πράγματι, το  $p$  έχει βαθμό το πολύ ίσο με  $n$  (αφού είναι ο γραμμικός συνδυασμός πολυωνύμων βαθμού  $n$ ), και ικανοποιεί τις συνθήκες παρεμβολής επειδή ισχύει

$$p(x_i) = \sum_{j=0}^n y_j L_j(x_i) = 0 + \cdots + 0 + y_i L_i(x_i) + 0 + \cdots + 0 = y_i$$

**Παράδειγμα 10.2.** Θα χρησιμοποιήσουμε τα ίδια τρία ζεύγη δεδομένων όπως στο Παράδειγμα 10.1, συγκεκριμένα τα  $(1, 1)$ ,  $(2, 3)$  και  $(4, 3)$ , για να δείξουμε πώς κατασκευάζονται τα πολυώνυμα Lagrange. Για να αναγκάσουμε το  $L_0(x)$  να μηδενιστεί στα  $x = 2$  και  $x = 4$ , γράφουμε

$$L_0(x) = a(x - 2)(x - 4)$$

Τότε, από την απαίτηση να ισχύει  $L_0(1) = 1$  προκύπτει ότι  $a(1 - 2)(1 - 4) = 1$ , δηλαδή  $a = \frac{1}{3}$ , και άρα

$$L_0(x) = \frac{1}{3}(x - 2)(x - 4)$$

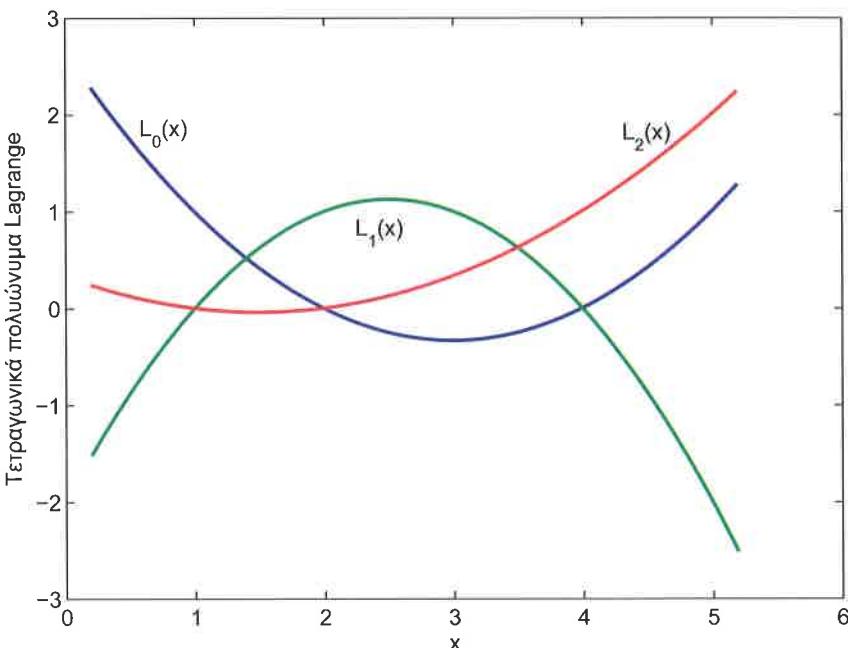
Ομοίως, βρίσκουμε ότι

$$L_1(x) = -\frac{1}{2}(x-1)(x-4), \quad L_2(x) = \frac{1}{6}(x-1)(x-2)$$

Αυτά τα πολυώνυμα Lagrange απεικονίζονται στο Σχήμα 10.3. Επομένως, παίρνουμε τη συνάρτηση παρεμβολής

$$\begin{aligned} p_2(x) &= \frac{y_0}{3}(x-2)(x-4) - \frac{y_1}{2}(x-1)(x-4) + \frac{y_2}{6}(x-1)(x-2) \\ &= \frac{1}{3}(x-2)(x-4) - \frac{3}{2}(x-1)(x-4) + \frac{3}{6}(x-1)(x-2) \end{aligned}$$

Παρά τη διαφορετική μορφή της, αυτή είναι ακριβώς η ίδια τετραγωνική συνάρτηση παρεμβολής όπως στο Παράδειγμα 10.1, άρα στην πραγματικότητα έχουμε  $p_2(x) = (-2x^2 + 12x - 7)/3$ . Επίσης, είναι εύκολο να επαληθεύσουμε ότι και εδώ ισχύει  $p_2(3) = \frac{11}{3}$ . Όλα αυτά δεν θα έπρεπε να μας προκαλούν έκπληξη –δείχνουν απλώς τη μοναδικότητα της πολυωνυμικής παρεμβολής (δείτε το θεώρημα στη σελίδα 480). ■



**Σχήμα 10.3:** Τα τετραγωνικά πολυώνυμα Lagrange  $L_0(x)$ ,  $L_1(x)$  και  $L_2(x)$  που βασίζονται στα σημεία  $x_0 = 1$ ,  $x_1 = 2$ ,  $x_2 = 4$  (Παράδειγμα 10.2).

## Ιδιότητες των πολυωνύμων Lagrange

Ποιες ιδιότητες έχουν τα πολυώνυμα Lagrange; Ποια ακριβώς είναι η μορφή τους;

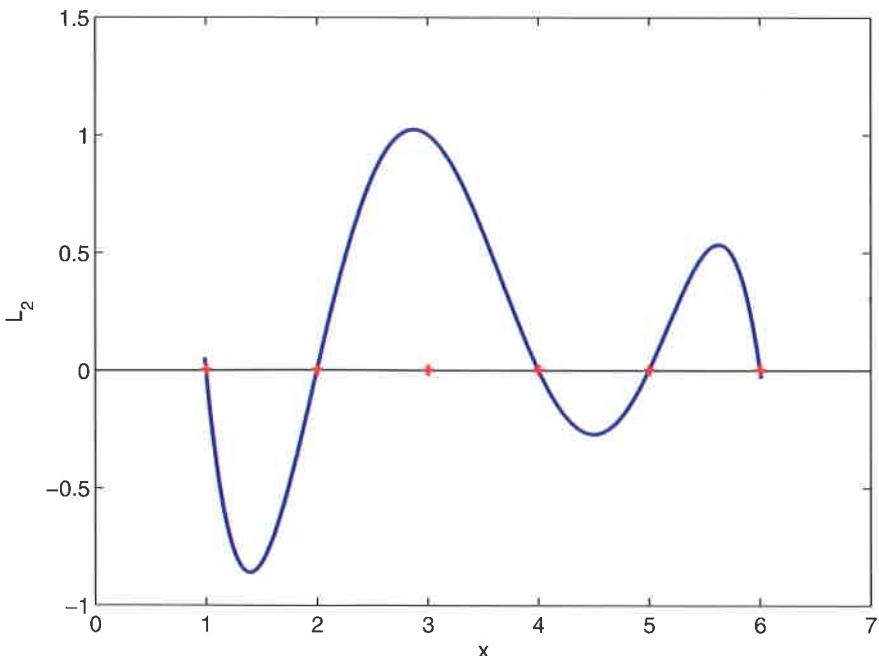
Στη γενική περίπτωση, όπου καθορίζονται  $n + 1$  τετμημένες δεδομένων  $x_i$ , τα πολυώνυμα Lagrange είναι μοναδικά επειδή δεν είναι τίποτε άλλο παρά πολυώνυμα παρεμβολής για ειδικά δεδομένα.<sup>52</sup> Αυτό μπορούμε εύκολα να το επαληθεύσουμε απευθείας, καθορίζοντας ρητά ότι

$$L_j(x) = \frac{(x - x_0) \cdots (x - x_{j-1})(x - x_{j+1}) \cdots (x - x_n)}{(x_j - x_0) \cdots (x_j - x_{j-1})(x_j - x_{j+1}) \cdots (x_j - x_n)} = \prod_{\substack{i=0 \\ i \neq j}}^n \frac{(x - x_i)}{(x_j - x_i)}$$

Πράγματι, το πολυώνυμο βαθμού  $n$ , γραμμένο σε σχέση με τις ρίζες του ως

$$(x - x_0) \cdots (x - x_{j-1})(x - x_{j+1}) \cdots (x - x_n)$$

παρεμβάλλεται ξεκάθαρα στις μηδενικές τιμές όλων των τετμημένων των δεδομένων εκτός από τα  $x_j$ , και η διαίρεση με την τιμή του  $x_j$  κανονικοποιεί την παράσταση και δίνει  $L_j(x_j) = 1$ . Μια άλλη εικόνα ενός πολυωνύμου Lagrange παρέχεται στο Σχήμα 10.4.



**Σχήμα 10.4:** Το πολυώνυμο Lagrange  $L_2(x)$  για  $n = 5$ . Μαντέψτε ποιες είναι οι τετμημένες δεδομένων  $x_i$ .

<sup>52</sup> Για κάθε  $j$ , θέτουμε  $y_i \leftarrow 1$  αν  $i = j$  και  $y_i \leftarrow 0$  αν  $i \neq j$ ,  $i = 0, 1, \dots, n$ .

### Αλγόριθμος: Πολυωνυμική παρεμβολή Lagrange.

- Κατασκευή:** Για τα δεδομένα  $\{(x_i, y_i)\}_{i=0}^n$ , υπολόγισε τους βαρυκεντρικούς συντελεστές  $w_j = 1 / \prod_{i \neq j} (x_j - x_i)$  και τις ποσότητες  $w_j y_j$ , για  $j = 0, 1, \dots, n$ .
- Υπολογισμός:** Για ένα σημείο υπολογισμού  $x$  που δεν είναι ίσο με κάποιο από τα σημεία δεδομένων  $\{x_i\}_{i=0}^n$ , υπολόγισε το πολυώνυμο

$$p(x) = \frac{\sum_{j=0}^n \frac{w_j y_j}{(x-x_j)}}{\sum_{j=0}^n \frac{w_j}{(x-x_j)}}$$

Τα πολυώνυμα Lagrange σχηματίζουν μια βάση ιδανικής κατάστασης,  $\phi_j(x) = L_j(x)$ , για όλα τα πολυώνυμα βαθμού το πολύ  $n$ . Στην περίπτωση της παρεμβολής Lagrange, τα στοιχεία της μήτρας του συστήματος (10.1) στη σελίδα 475 είναι  $\phi_j(x_i) = L_j(x_i)$ , και διαπιστώνουμε ότι η δυνητικά προβληματική μήτρα Vandermonde για τη μονωνυμική βάση στην Ενότητα 10.2 είναι πλέον ταντοτική μήτρα. Συνεπώς, το σύστημα (10.1), που δεν ορίζεται ποτέ με αυτόν τον τρόπο, δίνει τη λύση  $c_j = y_j$ ,  $j = 1, \dots, n$ .

Σημειώστε ότι το  $L_j$  έχει  $n$  ρίζες και άρα  $n - 1$  ακρότατα (εναλλασσόμενα ελάχιστα και μέγιστα). Παρεμπιπτόντως, το  $L_j(x_j) = 1$  δεν χρειάζεται να είναι μέγιστο για την  $L_j(x)$ .

### Κατασκευή και υπολογισμός

Από την περιγραφή μας μέχρι στιγμής πρέπει να έχει γίνει σαφές τι είναι τα πολυώνυμα Lagrange και πώς λειτουργούν για σκοπούς παρεμβολής. Στη συνέχεια θα επινοήσουμε έναν αποδοτικό τρόπο για να εκτελούμε τα στάδια της κατασκευής και του υπολογισμού. Αυτό θα μπορούσε να αποδειχθεί σημαντικό αν η τιμή του  $n$  είναι μεγάλη.

Το στάδιο της κατασκευής περιλαμβάνει οτιδήποτε δεν εξαρτάται από κάποιο σημείο υπολογισμού  $x$ . Εδώ, αυτό ισοδυναμεί με την κατασκευή των παρονομαστών των  $n + 1$  πολυωνύμων Lagrange. Θα ορίσουμε ότι

$$\rho_j = \prod_{i \neq j} (x_j - x_i), \quad w_j = \frac{1}{\rho_j}, \quad j = 0, 1, \dots, n$$

Αυτή η κατασκευή απαιτεί περίπου  $n^2$  flop. Οι ποσότητες  $w_j$  ονομάζονται **βαρυκεντρικοί συντελεστές** (barycentric weights).

Ας προχωρήσουμε στο στάδιο του υπολογισμού. Για ένα συγκεκριμένο σημείο  $x$  διαφορετικό από την τετμημένη στην οποία απαιτείται η τιμή του πολυωνύμου  $p(x)$ , παρατηρούμε ότι όλοι οι αριθμητές του  $L_j(x)$  περιέχουν μικρές παραλλαγές της συνάρτησης

$$\psi(x) = (x - x_0) \cdots (x - x_n) = \prod_{i=0}^n (x - x_i)$$

Τότε, η συνάρτηση παρεμβολής είναι

$$p(x) = \psi(x) \sum_{j=0}^n \frac{w_j y_j}{(x - x_j)}$$

Είναι προφανές ότι, για κάθε τιμή του  $x$ , το κόστος του υπολογισμού είναι  $O(n)$  flop (περίπου  $5n$ , για να είμαστε πιο ακριβείς).

Μπορούμε να κάνουμε μια μικρή περαιτέρω τροποποίηση, ουσιαστικά για λόγους αισθητικής, παρατηρώντας ότι για τη συγκεκριμένη συνάρτηση  $f(x)$  που έχει τιμή 1 παντού ισχύει  $y_j = 1$  για κάθε  $j$ . Επιπλέον, η συνάρτηση παρεμβολής έχει επίσης τιμή 1 στα ίδια ακριβώς σημεία, επομένως ισχύει

$$1 = \psi(x) \sum_{j=0}^n \frac{w_j \cdot 1}{(x - x_j)}$$

για κάθε  $x$ . Άρα, η  $\psi(x)$  ορίζεται σε σχέση με ποσότητες που υπολογίζονται ούτως ή άλλως. Ο μαθηματικός τύπος που προκύπτει ονομάζεται **βαρυκεντρική παρεμβολή** (barycentric interpolation) και οδηγεί στον αλγόριθμο που παρουσιάστηκε στην προηγούμενη σελίδα. Θα άξιζε τον κόπο να εφαρμόσετε αυτόν τον αλγόριθμο στο Παράδειγμα 10.2 και να ελέγξετε τις ενδιάμεσες ποσότητες που προκύπτουν.

*Ασκήσεις γι' αυτή την ενότητα: 4–6.*

## 10.4 Διαιρεμένες διαφορές και η μορφή Newton

Σε αυτή την ενότητα θα συνεχίσουμε να μελετάμε την πολυωνυμική παρεμβολή σε  $n + 1$  σημεία δεδομένων με διαφορετικές τετμημένες. Θα εισαγάγουμε μια ακόμα αναπαράσταση για την πολυωνυμική παρεμβολή (Θα είναι η τελευταία, σας το υποσχόμαστε!), επειδή οι προηγούμενες αναπαραστάσεις μας δεν καλύπτουν δύο σημαντικές πτυχές. Η πρώτη είναι η ελκυστικότητα της εισαγωγής των δεδομένων παρεμβολής  $(x_i, y_i)$  κατά ένα ζεύγος τη φορά, αντί για την ταυτόχρονη εισαγωγή όλων των ζευγών από την αρχή. Η άλλη πτυχή που έχουμε αποφύγει μέχρι στιγμής είναι η εκτίμηση του σφάλματος στην προσέγγιση της παρεμβολής, και τα όσα

Θα περιγράψουμε εδώ θα μας βοηθήσουν να διαμορφώσουμε την ανάλυση που θα παρουσιάσουμε στην Ενότητα 10.5.

## Η πολυωνυμική βάση Newton

Προς το παρόν έχουμε δει την τυπική μονωνυμική βάση,  $\{\phi_j(x) = x^j\}_{j=0}^n$ . Αυτό είχε ως αποτέλεσμα να προκύψει μια όχι και τόσο ιδανική διαδικασία για την κατασκευή του  $p = p_n$ , αλλά και μια εύκολη διαδικασία για τον υπολογισμό του  $p_n(x)$  σε ένα ορισμένο  $x$ . Από την άλλη, με την πολυωνυμική βάση Lagrange,  $\{\phi_j(x) = \prod_{i \neq j, i=0}^n \frac{(x-x_i)}{(x_j-x_i)}\}_{j=0}^n$ , το στάδιο της κατασκευής είναι εύκολο, αλλά ο υπολογισμός του  $p_n(x)$  είναι σχετικά περίπλοκος θεωρητικά. Η πολυωνυμική βάση Newton μπορεί να εκληφθεί ως ένας χρήσιμος συμβιβασμός: θέτουμε

$$\phi_j(x) = \prod_{i=0}^{j-1} (x - x_i), \quad j = 0, 1, \dots, n$$

Στην περιγραφή που ακολουθεί αποκαλύπτονται τα πλεονεκτήματα αυτής της επιλογής.

**Παράδειγμα 10.3.** Για μια τετραγωνική συνάρτηση παρεμβολής έχουμε τις συναρτήσεις βάσης

$$\phi_0(x) = 1, \quad \phi_1(x) = x - x_0, \quad \phi_2(x) = (x - x_0)(x - x_1)$$

Θεωρήστε ξανά το ίδιο σύνολο δεδομένων όπως στα Παραδείγματα 10.1 και 10.2, συγκεκριμένα τα σημεία  $(1, 1)$ ,  $(2, 3)$  και  $(4, 3)$ . Η συνθήκη παρεμβολής στο  $x_0 = 1$  δίνει

$$1 = p(1) = c_0\phi_0(1) + c_1\phi_1(1) + c_2\phi_2(1) = c_0 \cdot 1 + c_1 \cdot 0 + c_2 \cdot 0 = c_0$$

Στη συνέχεια, για  $c_0 = 1$ , η συνθήκη παρεμβολής στο  $x_1 = 2$  είναι

$$3 = p(2) = 1 + c_1(2 - 1) + c_2 \cdot 0$$

το οποίο μας δίνει  $c_1 = 2$ .

Τέλος, από την τρίτη συνθήκη παρεμβολής παίρνουμε

$$3 = p(4) = 1 + 2(4 - 1) + c_2(4 - 1)(4 - 2)$$

το οποίο μας δίνει  $c_2 = -\frac{4}{3}$ . Συνεπώς, το πολυώνυμο παρεμβολής είναι

$$p_2(x) = 1 + 2(x - 1) - \frac{2}{3}(x - 1)(x - 2)$$

Στο  $x = 3$  η τιμή του είναι, φυσικά, ίση με  $p_2(3) = \frac{11}{3}$ , δηλαδή η ίδια τιμή όπως στα Παραδείγματα 10.1 και 10.2, επειδή το πολυώνυμο παρεμβολής είναι μοναδικό: Αλ-

λάζει μόνο η μορφή του ανάλογα με την αναπαράσταση. Όπως και στα προηγούμενα παραδείγματά μας, θέλουμε να παροτρύνουμε τους σκεπτικιστές (και, ως γνωστόν, κάθε καλός επιστήμονας πρέπει να διακατέχεται από μια υγή δόση σκεπτικισμού!) να επαληθεύσουν, για μία ακόμα φορά, ότι προκύπτει το ίδιο πολυώνυμο όπως και στα προηγούμενα παραδείγματα.

Το σημαντικό στοιχείο που πρέπει να θυμόμαστε από αυτό το παράδειγμα είναι ότι για να υπολογίσουμε τον συντελεστή  $c_0$  χρησιμοποιήσαμε μόνο το πρώτο σημείο δεδομένων, και για να υπολογίσουμε τον συντελεστή  $c_1$  χρειαστήκαμε μόνο τα πρώτα δύο σημεία. ■

## Υπαρξη προσαρμοστικής συνάρτησης παρεμβολής

Το πιο σημαντικό χαρακτηριστικό της αναπαράστασης Newton είναι ότι είναι εξελικτική, ή αναδρομική: Έχοντας προσδιορίσει το  $p_{n-1}(x)$  που παρεμβάλλεται στα πρώτα  $n$  σημεία δεδομένων, το χρησιμοποιούμε για να κατασκευάσουμε χωρίς 1-διαίτερο κόστος το  $p_n(x)$  που παρεμβάλλεται σε όλα τα προηγούμενα δεδομένα και το  $(x_n, y_n)$ . Επομένως, δεν χρειάζεται να γνωρίζουμε όλα τα σημεία δεδομένων ή να προσδιορίσουμε τον βαθμό  $n$  εκ των προτέρων. Αντιθέτως, μπορούμε να υιοθετήσουμε έναν προσαρμοστικό (adaptive) τρόπο. Αυτό το χαρακτηριστικό ενδέχεται να αποδειχθεί υπερπολύτιμο όταν κανείς επεξεργάζεται εργαστηριακές μετρήσεις, για παράδειγμα, όπου δεν είναι όλα τα δεδομένα διαθέσιμα ταυτόχρονα. Δείτε επίσης τις Ασκήσεις 9 και 10.

Η παρούσα μορφή του  $p_n(x)$  είναι

$$p_n(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + \cdots + c_n(x - x_0)(x - x_1) \cdots (x - x_{n-1})$$

Είναι πάντα εφικτός ο προσδιορισμός των συντελεστών  $c_j$ ; Και αν ναι, είναι αυτή η συγκεκριμένη αναπαράσταση μοναδική;

Μπορούμε να βρούμε τις (καταφατικές) απαντήσεις εξετάζοντας το γενικό σύστημα (10.1) στη σελίδα 475. Λαμβάνοντας υπόψη τη μορφή των συναρτήσεων βάσης, είναι προφανές ότι στην παρούσα περίπτωση έχουμε

$$\phi_j(x_i) = 0, \quad i = 0, 1, \dots, j-1, \quad \phi_j(x_j) \neq 0$$

Άρα, η μήτρα του συστήματος (10.1) είναι κάτω τριγωνική και τα στοιχεία στην κύρια διαγώνιο της είναι μη μηδενικά. Επειδή η ορίζουσα μιας τριγωνικής μήτρας είναι το γινόμενο των διαγώνιων στοιχείων της, βρίσκουμε ότι, για αυθαίρετες (διαφορετικές) τετμημένες δεδομένων, η μήτρα είναι μη ιδιάζουσα. Συνεπώς, υπάρχει μοναδική λύση για τους άγνωστους συντελεστές  $c_0, \dots, c_n$  ως προς τα δεδομένα  $y_0, \dots, y_n$ .

Το παραπάνω επιχείρημα στηρίζεται σε βασικές έννοιες της γραμμικής άλγεβρας. Μάλιστα, υπάρχει και ένας αποδοτικός αλγόριθμος προς τα εμπρός αντικατάστασης (σελίδα 174). Επομένως, θα μπορούσαμε να κατασκευάσουμε το σύστημα (10.1) και να εφαρμόσουμε τον αλγόριθμο της Ενότητας 5.1 για να ολοκληρώσουμε την τρέχουσα περιγραφή μας σε αυτό ακριβώς το σημείο. Θα επιλέξουμε, ωστόσο, να προχωρήσουμε εκτελώντας τον αλγόριθμο προς τα εμπρός αντικατάστασης συμβολικά, χωρίς να κατασκευάσουμε το σύστημα (10.1) και χωρίς να στηριχθούμε στις γνώσεις που έχουμε αποκομίσει από το Κεφάλαιο 5, επειδή αυτό θα έχει ως αποτέλεσμα να προκύψουν σημαντικές πρόσθετες πληροφορίες σχετικά με τις διαδικασίες προσέγγισης καθώς και ένας πιο άμεσος, αποδοτικός αλγόριθμος.

## Αναπαράσταση με διαιρεμένες διαφορές

Για να απλοποιήσουμε τον συμβολισμό, θα χρησιμοποιούμε το  $f(x_i)$  αντί του  $y_i$ : Θέλουμε να παρακολουθούμε ποια δεδομένα χρησιμοποιούνται στην αναδρομική διαδικασία κατασκευής. Συνεπώς, προχωράμε ως εξής.

- Προσδιορίζουμε τον συντελεστή  $c_0$  χρησιμοποιώντας τη συνθήκη  $p_n(x_0) = f(x_0)$ :

$$\begin{aligned} f(x_0) &= p_n(x_0) = c_0 + 0 + \cdots + 0 = c_0 \\ \Rightarrow c_0 &= f(x_0) \end{aligned}$$

- Έπειτα προσδιορίζουμε τον συντελεστή  $c_1$  χρησιμοποιώντας τη συνθήκη  $p_n(x_1) = f(x_1)$ :

$$\begin{aligned} f(x_1) &= p_n(x_1) = c_0 + c_1(x_1 - x_0) + 0 + \cdots + 0 = c_0 + c_1(x_1 - x_0) \\ \Rightarrow c_1 &= \frac{f(x_1) - f(x_0)}{x_1 - x_0} \end{aligned}$$

- Στη συνέχεια επιβάλλουμε επίσης τη συνθήκη  $p_n(x_2) = f(x_2)$ . Έχοντας ήδη προσδιορίσει τους συντελεστές  $c_0$  και  $c_1$ , παίρνουμε μια συνθήκη που περιλαμβάνει μόνο τον συντελεστή  $c_2$ . Σας παρακαλούμε να επαληθεύσετε (μπορείτε να χρησιμοποιήσετε τη διατύπωση της Άσκησης 13 γι' αυτόν τον σκοπό) ότι το αποτέλεσμα μπορεί να γραφεί ως

$$c_2 = \frac{\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0}$$

- Συνεχίζουμε αυτή τη διαδικασία μέχρι να προσδιοριστούν όλοι οι συντελεστές  $c_j$ .

Ο συντελεστής  $c_j$  του πολυωνύμου παρεμβολής σε μορφή Newton είναι γνωστός ως η  $j$ -οστή διαιρεμένη διαφορά (divided difference) και συμβολίζεται με  $f[x_0, x_1, \dots, x_j]$ . Επομένως, γράφουμε

$$f[x_0] = c_0, f[x_0, x_1] = c_1, \dots, f[x_0, x_1, \dots, x_n] = c_n$$

Αυτό υποδεικνύει ρητά τα σημεία δεδομένων από τα οποία εξαρτάται ο κάθε συντελεστής  $c_j$ .

**Σημείωση:** Ο συμβολισμός των διαιρεμένων διαφορών είναι αρκετά λεπτομερής, ίσως με έναν τρόπο που αρχικά δεν φαίνεται και τόσο ελκυστικός. Σας ζητάμε να κάνετε λίγη υπομονή επειδή πολύ σύντομα θα αναδειχθούν πολλές σημαντικές ιδέες.

Άρα, ο μαθηματικός τύπος της παρεμβολής με διαιρεμένες διαφορές Newton στην πλήρη μορφή του είναι

$$\begin{aligned} p_n(x) &= f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) \\ &\quad + \cdots + f[x_0, x_1, \dots, x_n](x - x_0)(x - x_1) \cdots (x - x_{n-1}) \\ &= \sum_{j=0}^n \left( f[x_0, x_1, \dots, x_j] \prod_{i=0}^{j-1} (x - x_i) \right) \end{aligned}$$

Οι συντελεστές διαιρεμένων διαφορών ικανοποιούν την αναδρομική εξίσωση

$$f[x_0, x_1, \dots, x_j] = \frac{f[x_1, x_2, \dots, x_j] - f[x_0, x_1, \dots, x_{j-1}]}{x_j - x_0}$$

Δείτε την Άσκηση 7. Γενικότερα, το  $x_0$  μπορεί να αντικατασταθεί από το  $x_i$  για οποιαδήποτε  $0 \leq i < j$ . Ο μαθηματικός τύπος που προκύπτει παρουσιάζεται παρακάτω. Η συγκεκριμένη αναδρομή σημαίνει ότι δεν χρειάζεται να κατασκευάζουμε από την αρχή τους συντελεστές κάθε φορά που προσθέτουμε ένα νέο σημείο παρεμβολής.

### Πίνακας διαιρεμένων διαφορών και υπολογισμός συνάρτησης παρεμβολής

Πώς μπορούμε να χρησιμοποιήσουμε στην πράξη όλες αυτές τις αναδρομές και τους μαθηματικούς τύπους; Σημειώστε ότι για να υπολογίσουμε το  $\gamma_{n,n} = f[x_0, x_1, \dots, x_n]$

πρέπει να υπολογίσουμε όλα τα

$$\gamma_{j,l} = f[x_{j-l}, x_{j-l+1}, \dots, x_j], \quad 0 \leq l \leq j \leq n$$

**Διαιρεμένες διαφορές.** Για τα σημεία  $x_0, x_1, \dots, x_n$  και για αυθαίρετους δείκτες  $0 \leq i < j \leq n$ , θέσε

$$f[x_i] = f(x_i)$$

$$f[x_i, \dots, x_j] = \frac{f[x_{i+1}, \dots, x_j] - f[x_i, \dots, x_{j-1}]}{x_j - x_i}$$

Συνεπώς, κατασκευάζουμε έναν πίνακα διαιρεμένων διαφορών, ο οποίος είναι ο παρακάτω άνω τριγωνικός πίνακας.

$i$	$x_i$	$f[x_i]$	$f[x_{i-1}, x_i]$	$f[x_{i-2}, x_{i-1}, x_i]$	$\dots$	$f[x_{i-n}, \dots, x_i]$
0	$x_0$	$f(x_0)$				
1	$x_1$	$f(x_1)$	$\frac{f[x_1] - f[x_0]}{x_1 - x_0}$			
2	$x_2$	$f(x_2)$	$\frac{f[x_2] - f[x_1]}{x_2 - x_1}$	$f[x_0, x_1, x_2]$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	
$n$	$x_n$	$f(x_n)$	$\frac{f[x_n] - f[x_{n-1}]}{x_n - x_{n-1}}$	$f[x_{n-2}, x_{n-1}, x_n]$	$\dots$	$f[x_0, x_1, \dots, x_n]$

Εξάγοντας τα διαγώνια στοιχεία του πίνακα παίρνουμε τους συντελεστές  $c_j = \gamma_{j,j} = f[x_0, \dots, x_j]$  για το πολυώνυμο παρεμβολής Newton.

**Παράδειγμα 10.4.** Για το ίδιο πρόβλημα όπως στο Παράδειγμα 10.3 βρίσκουμε ότι

$$f[x_0, x_1] = \frac{3-1}{2-1} = 2, \quad f[x_1, x_2] = \frac{3-3}{4-2} = 0, \quad f[x_0, x_1, x_2] = \frac{0-2}{4-1} = -\frac{2}{3}$$

Ο αντίστοιχος πίνακας διαιρεμένων διαφορών είναι

$i$	$x_i$	$f[\cdot]$	$f[\cdot, \cdot]$	$f[\cdot, \cdot, \cdot]$
0	1	1		
1	2	3	2	
2	4	3	0	$-\frac{2}{3}$

Επομένως, το πολυώνυμο παρεμβολής  $p_2(x)$  είναι ίδιο με εκείνο που προσδιορίζεται στο Παράδειγμα 10.3.

Προσέξτε ότι οι πρώτοι δύο όροι του  $p_2(x)$  σχηματίζουν το γραμμικό πολυώνυμο παρεμβολής  $p_1(x)$  που βρήκαμε στο Παράδειγμα 10.1. Αυτό αποδεικνύει το

δυνατό σημείο της βάσης Newton. Επίσης, σε ένθετη μορφή έχουμε

$$p_2(x) = 1 + (x - 1) \left( 2 - \frac{2}{3}(x - 2) \right)$$

Αν θελήσουμε να προσθέσουμε ένα ακόμα σημείο δεδομένων, έστω το  $(x_3, f(x_3)) = (5, 4)$ , χρειάζεται να προσθέσουμε μόνο μία γραμμή στον πίνακα διαιρεμένων διαφορών. Υπολογίζουμε τα

$$\begin{aligned} f[x_3] &= 4, \quad f[x_2, x_3] = \frac{4 - 3}{5 - 4} = 1, \quad f[x_1, x_2, x_3] = \frac{1 - 0}{5 - 2} = \frac{1}{3} \\ f[x_0, x_1, x_2, x_3] &= \frac{(1/3) - (-2/3)}{5 - 1} = \frac{1}{4} \end{aligned}$$

Ο αντίστοιχος πίνακας είναι

$i$	$x_i$	$f[\cdot]$	$f[\cdot, \cdot]$	$f[\cdot, \cdot, \cdot]$	$f[\cdot, \cdot, \cdot, \cdot]$
0	1	1			
1	2	3	2		
2	4	3	0	$-\frac{2}{3}$	
3	5	4	1	$\frac{1}{3}$	$\frac{1}{4}$

Άρα, για το  $p_3$  ποιόρνουμε την παράσταση

$$\begin{aligned} p_3(x) &= p_2(x) + f[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2) \\ &= 1 + (x - 1) \left( 2 - \frac{2}{3}(x - 2) \right) + \frac{1}{4}(x - 1)(x - 2)(x - 4) \end{aligned}$$

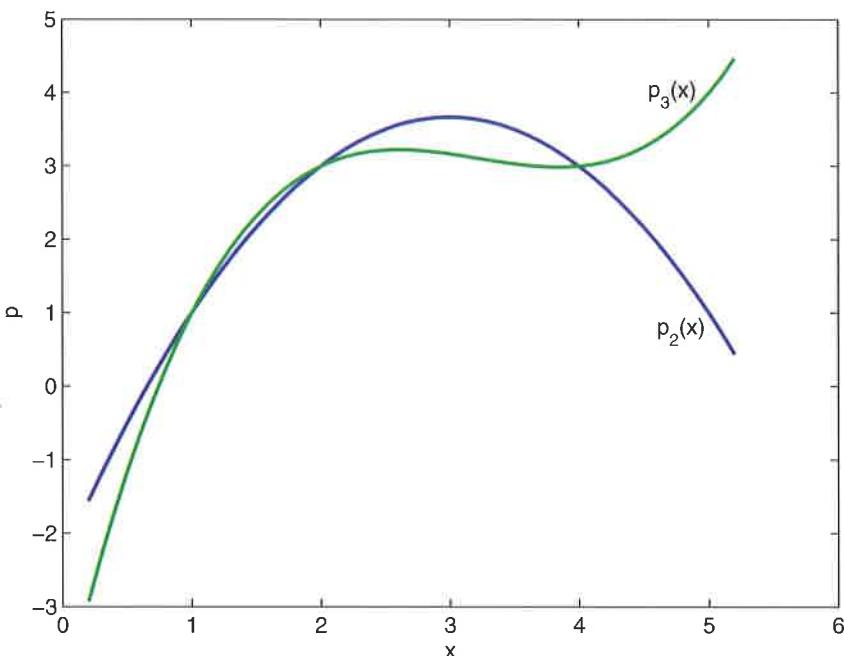
Σε ένθετη μορφή, αυτή η παράσταση γράφεται ως

$$p_3(x) = 1 + (x - 1) \left( 2 + (x - 2) \left( -\frac{2}{3} + \frac{1}{4}(x - 4) \right) \right)$$

Για να βρούμε μια προσέγγιση μεγαλύτερου βαθμού, το μόνο που έχουμε να κάνουμε είναι να προσθέσουμε έναν όρο.

Στο Σχήμα 10.5 φαίνονται τα δύο πολυώνυμα  $p_2$  και  $p_3$ . Προσέξτε ότι το  $p_2$  προβλέπει μια αρκετά διαφορετική τιμή για  $x = 5$  συγκριτικά με ό,τι επιβάλλει αργότερα το πρόσθετο στοιχείο δεδομένων  $f(x_3)$ , και έτσι εξηγείται η σημαντική διαφορά που έχουν οι δύο καμπύλες παρεμβολής. ■

Ακολουθούν τρεις συναρτήσεις στο MATLAB οι οποίες υλοποιούν τη μέθοδο που περιγράφηκε παραπάνω. Η πρώτη κατασκευάζει τον πίνακα διαιρεμένων διαφορών και επιστρέφει τους συντελεστές του πολυωνύμου παρεμβολής.



**Σχήμα 10.5:** Τα πολυωνύμα παρεμβολής  $p_2$  και  $p_3$  για το Παράδειγμα 10.4.

```

function [coef,table] = divdif (xi, yi)
%
% function [coef,table] = divdif (xi, yi)
%
% Κατασκευάζει έναν πίνακα διαιρεμένων διαφορών
% με βάση τα σημεία δεδομένων (xi,yi).
% Επιστρέφει τους συντελεστές παρεμβολής Newton στη
% μεταβλητή coef

np1 = length(xi); n = np1-1;
table = zeros(np1,np1); xi = shiftdim(xi); yi = shiftdim(yi);
% Κατασκευή του πίνακα διαιρεμένων διαφορών μία στήλη τη φορά
table(1:np1,1) = yi;
for k = 2:np1
    table(k:np1,k) = (table(k:np1,k-1) - table(k-1:n,k-1)) ./ ...
                      (xi(k:np1) - xi(1:np1-k+1));
end
coef = diag(table); % τα διαγώνια στοιχεία του πίνακα

```

Σημειώστε ότι η μεταβλητή `table` είναι ένας διδιάστατος πίνακας, οπότε μπορείτε με ασφάλεια να αγνοήσετε τις μηδενικές τιμές στο αυστηρά άνω τριγωνικό κομμάτι του. Στην πραγματικότητα, πρέπει να μπορείτε να διακρίνετε από τον κώδικα ότι δεν είναι αυστηρά απαραίτητο να αποθηκεύεται ολόκληρος ο πίνακας, αλλά εμείς το κάνουμε για λόγους κατανόησης.

Η επόμενη συνάρτηση υπολογίζει με ένθετο τρόπο το πολυώνυμο παρεμβολής σε μορφή Newton:

```
function p = evalnewt (x, xi, coef)
%
% function p = evalnewt (x, xi, coef)
%
% Υπολογίζει στο x το πολυώνυμο παρεμβολής σε μορφή Newton
% με βάση τα σημεία παρεμβολής xi και τους συντελεστές coef

np1 = length(xi);
p = coef(np1)*ones(size(x));
for j=np1-1:-1:1
    p = p.* (x - xi(j)) + coef(j);
end
```

Ένας αλγόριθμος για μια πιο γενική περίπτωση (η οποία περιλαμβάνει και τιμές παραγώγων) παρέχεται στην Ενότητα 10.7 (σελίδα 509).

Το Παράδειγμα 10.4 δείχνει επίσης τη διαδικασία της προσθήκης ενός μόνο επιπλέον ζεύγους δεδομένων  $(x_{n+1}, f(x_{n+1}))$  σε ένα πολυώνυμο παρεμβολής  $p_n$  των  $n+1$  πρώτων ζευγών δεδομένων, από την οποία προκύπτει ότι

$$p_{n+1}(x) = p_n(x) + f[x_0, x_1, \dots, x_n, x_{n+1}] \prod_{i=0}^n (x - x_i)$$

Αυτό υλοποιείται στην ακόλουθη συνάρτηση.

```
function [coef,table] = divdifadd (xi, yi, table)
%
% function [coef,table] = divdifadd (xi, yi, table)
%
% Κατασκευάζει μία ακόμα γραμμή ενός υφιστάμενου πίνακα
% διαιρεμένων διαφορών, και προσθέτει συντελεστή παρεμβολής,
% με βάση ένα πρόσθετο (τελευταίο στα xi και yi)
% σημείο δεδομένων
```

```

np1 = length(xi); n = np1 - 1;
table = [table zeros(n,1); yi(np1) zeros(1,n)];
for k=2:np1
    table(np1,k) = (table(np1,k-1) - table(n,k-1)) / ...
                    (xi(np1) - xi(np1-k+1));
end
coef = diag(table);

```

Παρακάτω μπορείτε να δείτε τον κώδικα (εκτός από κάποιες εντολές σχεδίασης) που χρησιμοποιήθηκε για τη δημιουργία του Σχήματος 10.5:

```

x = 0.2:0.01:5.2; % πλέγμα υπολογισμού
% τετραγωνική συνάρτηση παρεμβολής
xi = [1,2,4]; yi = [1,3,3];
[coef2,table] = divdif(xi,yi);
% υπολογισμός τετραγωνικής στο x
y2 = evalnewt(x,xi,coef2);
% προσθήκη σημείου δεδομένων
xi = [xi,5]; yi = [yi,4];
% κυβική συνάρτηση παρεμβολής
[coef3,table] = divdifadd(xi,yi,table);
% υπολογισμός κυβικής στο x
y3 = evalnewt(x,xi,coef3);
plot (x,y2,'b',x,y3,'g')

```

## Σύγκριση αλγόριθμων

Αν και μια ακριβής καταμέτρηση των πράξεων δεν είναι πάρα πολύ σημαντική πρακτικά (εν μέρει επειδή η τιμή του  $n$  δεν αναμένεται να είναι μεγάλη, με εξαίρεση ίσως την περιγραφή στην Ενότητα 10.6), πρέπει να επισημάνουμε ότι η κατασκευή του πίνακα διαιρεμένων διαφορών απαιτεί  $n^2/2$  διαιρέσεις και  $n^2$  προσθέσεις/αφαιρέσεις. Επιπλέον, ο ένθετος υπολογισμός της μορφής Newton απαιτεί περίπου  $2n$  πολλαπλασιασμούς και προσθέσεις: Αυτές οι καταμετρήσεις των πράξεων σε κάθε στάδιο είναι τουλάχιστον το ίδιο καλές όπως για τις προηγούμενες συναρτήσεις παρεμβολής που έχουμε δει.

Έχοντας περιγράψει τουλάχιστον τρεις βάσεις για πολυωνυμική παρεμβολή, και μάλιστα αρκετά λεπτομερώς, θα κάνουμε μια παύση και θα συνοψίσουμε τις αντίστοιχες ιδιότητές τους σε έναν πίνακα. Το κόστος της κατασκευής και του υπολογισμού σε ένα σημείο δίνεται ως προς τον μεγιστοβάθμιο όρο της καταμέτρησης των flop.

Ονομασία βάσης	$\phi_j(x)$	Κόστος κατασκευής	Κόστος υπολογισμού	Ελκυστικό χαρακτηριστικό
Μονωνυμική	$x^j$	$\frac{2}{3}n^3$	$2n$	απλή
Lagrange	$L_j(x)$	$n^2$	$5n$	$c_j = y_j$ πιο ευσταθής
Newton	$\prod_{i=0}^{j-1} (x - x_i)$	$\frac{3}{2}n^2$	$2n$	προσαρμοστική

## Διαιρεμένες διαφορές και παράγωγοι

**Σημείωση:** Η σημαντική σχέση που συνδέει τις διαιρεμένες διαφορές με τις παραγώγους, την οποία θα αποδείξουμε παρακάτω, δίνεται στο θεώρημα που παρουσιάζεται στην επόμενη σελίδα. Αν είστε διατεθειμένοι να την αποδεχθείτε χωρίς απόδειξη, μπορείτε να παραλείψετε την αρκετά τεχνική περιγραφή που ακολουθεί.

Η συνάρτηση διαιρεμένων διαφορών μπορεί να θεωρηθεί ως επέκταση της έννοιας της παραγώγου μιας συνάρτησης. Πράγματι, αν θεωρήσουμε μια παραγωγίσιμη συνάρτηση  $f(x)$  σε δύο σημεία  $z_0$  και  $z_1$ , και φανταστούμε το  $z_1$  να πλησιάζει στο  $z_0$ , τότε θα ισχύει

$$f[z_0, z_1] = \frac{f(z_1) - f(z_0)}{z_1 - z_0} \rightarrow f'(z_0)$$

Αλλά ακόμα και αν τα  $z_0$  και  $z_1$  παραμείνουν διαφορετικά, το θεώρημα μέσης τιμής (σελίδα 45) ορίζει ότι υπάρχει σημείο  $\zeta$  μεταξύ τους τέτοιο ώστε να ισχύει

$$f[z_0, z_1] = f'(\zeta)$$

Αντό συσχετίζει απευθείας την πρώτη διαιρεμένη διαφορά με την πρώτη παράγωγο της  $f$ , στην περίπτωση βέβαια που η παράγωγος υπάρχει.

Υπάρχει μια παρόμοια σχέση που συσχετίζει την  $k$ -οστή διαιρεμένη διαφορά της  $f$  με την  $k$ -οστή παράγωγό της. Θα αποδείξουμε την ύπαρξή της λεπτομερώς.

Έστω ότι  $z_0, z_1, \dots, z_k$  είναι  $k + 1$  διαφορετικά σημεία που περιέχονται σε ένα διάστημα στο οποίο μια συνάρτηση  $f$  είναι ορισμένη και έχει  $k$  φραγμένες παραγώγους. Πριν ξεκινήσουμε, προσέξτε ότι οι συντελεστές των διαιρεμένων διαφορών είναι συμμετρικοί ως προς τα ορίσματα. Δηλαδή, αν  $(\hat{z}_0, \hat{z}_1, \dots, \hat{z}_k)$  είναι μια μετάθεση (permutation) των τετμημένων  $(z_0, z_1, \dots, z_k)$ , τότε

$$f[\hat{z}_0, \hat{z}_1, \dots, \hat{z}_k] = f[z_0, z_1, \dots, z_k]$$

**Θεώρημα: Διαιρεμένη διαφορά και παράγωγος.**

Έστω ότι η συνάρτηση  $f$  είναι ορισμένη και έχει  $k$  φραγμένες παραγώγους στο διάστημα  $[a, b]$ , και έστω ότι  $z_0, z_1, \dots, z_k$  είναι  $k+1$  διαφορετικά σημεία στο  $[a, b]$ . Τότε, υπάρχει σημείο  $\zeta \in [a, b]$  τέτοιο ώστε να ισχύει

$$f[z_0, z_1, \dots, z_k] = \frac{f^{(k)}(\zeta)}{k!}$$

Δείτε την Άσκηση 13. Συνεπώς, μπορούμε κάλλιστα να υποθέσουμε ότι τα σημεία  $z_i$  είναι ταξινομημένα κατά αύξουσα σειρά και να γράψουμε

$$a = z_0 < z_1 < \dots < z_k = b$$

Χωρίς να υπολογίσουμε τίποτα, έστω ότι  $p_k$  είναι το πολυώνυμο παρεμβολής βαθμού το πολύ  $k$  που ικανοποιεί τη σχέση  $p_k(z_i) = f(z_i)$ , και έστω ότι το σφάλμα είναι  $e_k(x) = f(x) - p_k(x)$ . Τότε,  $e_k(z_i) = 0$ ,  $i = 0, 1, \dots, k$ , δηλαδή το  $e_k$  έχει  $k+1$  ρίζες.

Εφαρμόζοντας το θεώρημα του Rolle (δείτε τη σελίδα 45) παρατηρούμε ότι μεταξύ κάθε ζεύγους  $z_{i-1}$  και  $z_i$  του  $e_k(x)$  υπάρχει μια ρίζα της εξίσωσης του σφάλματος της παραγώγου,  $e'_k(x)$ , για  $k$  ρίζες συνολικά στο διάστημα  $[a, b]$ . Εφαρμόζοντας επανειλημμένα το θεώρημα του Rolle βρίσκουμε ότι το  $e_k^{(k-l)}(x)$  έχει  $l+1$  ρίζες στο  $[a, b]$ ,  $l = k, k-1, \dots, 1, 0$ . Ειδικότερα, η  $e_k^{(k)}$  έχει μία τέτοια ρίζα, που συμβολίζεται με  $\zeta$ , όπου

$$0 = e_k^{(k)}(\zeta) = f^{(k)}(\zeta) - p_k^{(k)}(\zeta)$$

Έπειτα πρέπει να χαρακτηρίσουμε την  $p_k^{(k)}$ , δηλαδή την  $k$ -οστή παράγωγο του πολυωνύμου παρεμβολής βαθμού  $k$ . Αυτή η εργασία είναι απλή αν θυμηθούμε τη μορφή Newton της παρεμβολής, την οποία μπορούμε να χρησιμοποιήσουμε για να γράψουμε το πολυώνυμο παρεμβολής ως

$$p_k(x) = f[z_0, z_1, \dots, z_k] x^k + q_{k-1}(x)$$

με το  $q_{k-1}(x)$  να είναι πολυώνυμο βαθμού  $< k$ .

Όμως η  $k$ -οστή παράγωγος του  $q_{k-1}(x)$  μηδενίζεται παντού:  $q_{k-1}^{(k)} \equiv 0$ . Άρα, το μόνο στοιχείο που παραμένει είναι η  $k$ -οστή παράγωγος της  $f[z_0, z_1, \dots, z_k] x^k$ . Η  $k$ -οστή παράγωγος του  $x^k$  είναι η σταθερά  $k!$  (γιατί;), και συνεπώς  $p_k^{(k)} = f[z_0, z_1, \dots, z_k](k!)$ . Αντικαθιστώντας το  $e_k^{(k)}(\zeta)$  στην παράσταση αυτή παίρνουμε τον σημαντικό μαθηματικό τύπο

$$f[z_0, z_1, \dots, z_k] = \frac{f^{(k)}(\zeta)}{k!}$$

Επομένως, έχουμε αποδείξει το θεώρημα που διατυπώθηκε παραπάνω, το οποίο συνδέει με μεθοδικό τρόπο διαιρεμένες διαφορές και παραγώγους ακόμα και όταν οι τετμημένες των δεδομένων δεν είναι κοντά η μία στην άλλη. ♦

Αυτός ο μαθηματικός τύπος συνηγορεί υπέρ της άποψης ότι η μορφή Newton της παρεμβολής αποτελεί, κατά κάποιον τρόπο, επέκταση του αναπτύγματος της σειράς Taylor. Θα αποδειχθεί χρήσιμος στην Ενότητα 10.7 καθώς και στην εκτίμηση του σφάλματος παρεμβολής, μια διαδικασία που θα περιγράψουμε στη συνέχεια.

*Ασκήσεις γι' αυτή την ενότητα: 7–14.*

## 10.5 Το σφάλμα της πολυωνυμικής παρεμβολής

Στην περιγραφή μας για τις διαιρεμένες διαφορές και τις παραγώγους στο τέλος της προηγούμενης ενότητας, εστιάσαμε στις τεταγμένες δεδομένων  $f(x_i) = y_i$ ,  $i = 0, 1, \dots, n$ , και αγνοήσαμε όλες τις άλλες τιμές της  $f(x)$ . Η συνάρτηση  $f$  που πιθανότατα παρήγαγε τα παρεμβληθέντα δεδομένα θα μπορούσε ακόμα και να μην είναι ορισμένη σε άλλες τιμές του ορίσματος. Εδώ, όμως, υποθέτουμε ότι η  $f(x)$  είναι ορισμένη στο διάστημα  $[a, b]$  το οποίο περιέχει τα σημεία παρεμβολής, και θέλουμε να αξιολογήσουμε πόσο μεγάλη ενδέχεται να είναι η διαφορά μεταξύ της  $f(x)$  και του πολυωνύμου  $p_n(x)$  σε οποιοδήποτε σημείο αυτού του διαστήματος. Θα υποθέσουμε επίσης ότι υπάρχουν κάποιες παράγωγοι της  $f(x)$ , οι οποίες είναι φραγμένες στο διάστημα που μας ενδιαφέρει.

### Μια παράσταση για το σφάλμα

Θα ορίσουμε τη συνάρτηση σφάλματος της κατασκευασμένης συνάρτησης παρεμβολής μας ως

$$e_n(x) = f(x) - p_n(x)$$

Μπορούμε να βρούμε μια παράσταση για το σφάλμα αυτό χρησιμοποιώντας ένα απλό τέχνασμα: Σε οποιοδήποτε σημείο  $x \in [a, b]$  στο οποίο θέλουμε να υπολογίσουμε το σφάλμα του  $p_n(x)$ , θα προσποιούμαστε ότι το  $x$  είναι ένα ακόμα σημείο παρεμβολής. Τότε, από την εξίσωση στην Ενότητα 10.4 που συνδέει το  $p_{n+1}$  με το  $p_n$  προκύπτει ότι

$$f(x) = p_{n+1}(x) = p_n(x) + f[x_0, x_1, \dots, x_n, x] \psi_n(x)$$

όπου ισχύει η σχέση

$$\psi_n(x) = \prod_{i=0}^n (x - x_i)$$

η οποία ορίζεται επίσης στη σελίδα 486. Επομένως, το σφάλμα είναι

$$e_n(x) = f(x) - p_n(x) = f[x_0, x_1, \dots, x_n, x] \psi_n(x)$$

Όσο απλή και αν είναι αυτή η παράσταση για το σφάλμα, εξαρτάται τόσο από τα δεδομένα όσο και από ένα μεμονωμένο σημείο υπολογισμού  $x$ . Θέλουμε έναν πιο γενικό, ποιοτικό χειρισμό του σφάλματος. Γι' αυτό θα βρούμε περαιτέρω μορφές εκτιμήσεων και φραγμάτων για το σφάλμα.

### Εκτιμήσεις και φράγματα του σφάλματος

Το πρώτο πράγμα που θα κάνουμε είναι να αντικαταστήσουμε τη διαιρεμένη διαφορά στην εξίσωση του σφάλματος με την αντίστοιχη παράγωγο, υποθέτοντας ότι η  $f$  είναι επαρκώς λεία. Παρατηρούμε ότι αν η προσεγγιζόμενη συνάρτηση  $f$  έχει  $n+1$  φραγμένες παραγώγους, το θεώρημα που παρουσιάζεται στη σελίδα 497 συνεπάγεται ότι υπάρχει σημείο  $\xi = \xi(x)$ , όπου  $a \leq \xi \leq b$ , τέτοιο ώστε να ισχύει  $f[x_0, x_1, \dots, x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!}$ . Αυτή η σχέση προκύπτει μόλις προσδιοριστούν το  $k$  για  $n+1$  και τα  $(z_0, z_1, \dots, z_k)$  για  $(x_0, \dots, x_n, x)$  στο συγκεκριμένο θεώρημα. Δίνει την εκτίμηση σφάλματος

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \psi_n(x)$$

**Θεώρημα: Σφάλμα πολυωνυμικής παρεμβολής.**

Αν το πολυώνυμο  $p_n$  παρεμβάλλεται στη συνάρτηση  $f$  στα  $n+1$  σημεία  $x_0, \dots, x_n$  και η  $f$  έχει  $n+1$  φραγμένες παραγώγους στο διάστημα  $[a, b]$  το οποίο περιέχει αυτά τα σημεία, τότε για κάθε  $x \in [a, b]$  υπάρχει σημείο  $\xi = \xi(x) \in [a, b]$  τέτοιο ώστε να ισχύει

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^n (x - x_i)$$

Επιπλέον, το φράγμα σφάλματος είναι

$$\max_{a \leq x \leq b} |f(x) - p_n(x)| \leq \frac{1}{(n+1)!} \max_{a \leq t \leq b} |f^{(n+1)}(t)| \max_{a \leq s \leq b} \prod_{i=0}^n |s - x_i|$$

Αν και δεν γνωρίζουμε ότι  $\xi = \xi(x)$ , μπορούμε να βρούμε ένα φράγμα για το σφάλμα σε όλα τα σημεία υπολογισμού  $x$  σε κάποιο διάστημα  $[a, b]$  που περιέχει τις τετμημένες δεδομένων  $x_0, \dots, x_n$ , αν υπάρχει ήδη ένα άνω φράγμα για την

$$\|f^{(n+1)}\| = \max_{a \leq t \leq b} |f^{(n+1)}(t)|$$

Ομοίως, μπορούμε να βρούμε ένα φράγμα για το σφάλμα παρεμβολής αν επιπροσθέτως μεγιστοποιήσουμε την  $|\psi_n(x)|$  για τις τετμημένες που δίνονται στο διάστημα υπολογισμού. Επιπλέον, επειδή για κάθε σημείο υπολογισμού  $x \in [a, b]$  ισχύει μια παρόμοια σχέση για το σφάλμα, το μέγιστο μέγεθος του σφάλματος στο συγκεκριμένο διάστημα είναι επίσης φραγμένο από την ίδια ποσότητα. Τελικά παίρνουμε το θεώρημα σφάλματος πολυωνυμικής παρεμβολής.

Σημειώστε ότι, αν και χρησιμοποιήσαμε τη μορφή Newton για να βρούμε την παραπάνω παράσταση για το σφάλμα, η παράσταση αυτή είναι ανεξάρτητη από τη βάση που χρησιμοποιείται για το πολυώνυμο παρεμβολής. Υπάρχει μόνο ένα τέτοιο πολυώνυμο, ανεξάρτητα από τη μέθοδο αναπαράστασης.

## Πρακτικά ζητήματα

Στην πράξη, η  $f$  συνήθως είναι άγνωστη, και το ίδιο ισχύει για τις παραγώγους της. Μερικές φορές μπορούμε να κάνουμε τα εξής:

- Να μετρήσουμε την ακρίβεια μιας τέτοιας προσέγγισης υπολογίζοντας μια ακολουθία πολυωνύμων  $p_k(x), p_{k+1}(x), \dots$  με τη χρήση διαφορετικών υποσυνόλων των σημείων και συγκρίνοντας τα αποτελέσματα για να δούμε πόσο «καλά» συμφωνούν. (Αν συμφωνούν «καλά», αυτό υποδηλώνει σύγκλιση, ή επαρκή ακρίβεια.) Αυτό θα έπρεπε να συμβαίνει όταν οι τιμές της  $f$  δεν διαφέρουν υπερβολικά μεταξύ τους και το διάστημα  $[a, b]$  που περιέχει τις τετμημένες των δεδομένων και τα σημεία υπολογισμού δεν είναι υπερβολικά μεγάλο.
- Να κρατήσουμε τα σημεία δεδομένων σε εφεδρεία για χρήση στην κατασκευή πολυωνύμων ανώτερης τάξης για την εκτίμηση του σφάλματος. Συνεπώς, αν έχουμε περισσότερα από τα  $n + 1$  ζεύγη δεδομένων που χρησιμοποιήθηκαν για την κατασκευή του  $p_n(x)$ , τότε μπορούμε να υπολογίσουμε το  $p_n$  σε αυτές τις πρόσθετες τετμημένες και να συγκρίνουμε με τις τιμές που δίνονται για να βρούμε μια υποεκτίμηση του μέγιστου σφάλματος στη συνάρτηση παρεμβολής μας. Ομως μπορεί να χρειαστούμε πολλά επιπλέον σημεία για μια ρεαλιστική εκτίμηση.

**Παράδειγμα 10.5.** Οι παρακάτω μέγιστες ημερήσιες θερμοκρασίες (σε βαθμούς Κελσίου) καταγράφηκαν κάθε τρίτη μέρα του Αιγαίου στου σε μια πόλη της Μέσης Ανατολής:

Ημέρα	3	6	9	12	15	18	21	24	27
Θερμοκρασία (°C)	31.2	32.0	35.3	34.1	35.0	35.5	34.1	35.1	36.0

Προσέξτε ότι αυτές οι τιμές δεδομένων δεν είναι μονοτονικά φθίνουσες ή αύξουσες. Έστω ότι θέλουμε να εκτιμήσουμε τη μέγιστη θερμοκρασία την ημέρα  $x = 13$  του συγκεκριμένου μήνα.

Εδώ είναι χρήσιμο να επιδείξουμε κάποια κριτική ικανότητα: Το γεγονός ότι υπάρχει διαθέσιμος ένας πίνακας με πολλά δεδομένα δεν συνεπάγεται υποχρεωτικά ότι πρέπει να τα χρησιμοποιήσουμε όλα για να αποκτήσουμε μια εικόνα σχετικά με το τι συμβαίνει γύρω από ένα συγκεκριμένο σημείο. Αυτό που θα μπορούσαμε να κάνουμε, για παράδειγμα, είναι να κατασκευάσουμε ένα κυβικό ( $n = 3$ ) πολυώνυμο χρησιμοποιώντας τέσσερα συγκεκριμένα σημεία κοντά στο  $x = 13$ . Ακολουθώντας αυτή τη στρατηγική, επιλέγουμε  $x_0 = 9$ ,  $x_1 = 12$ ,  $x_2 = 15$ ,  $x_3 = 18$ . Έτσι παίρνουμε  $p_3(13) = 34.29$ . Αν, αντιθέτως, καταφύγουμε στη γραμμική παρεμβολή των δύο πλησιέστερων γειτόνων, στα  $x = 12$  και  $x = 15$ , παίρνουμε  $p_1(13) = 34.4$ . Αυτό μας επιτρέπει να έχουμε κάποια εμπιστοσύνη ότι οι υπολογισμοί μας είναι πιθανώς σωστοί.

Συνοψίζοντας, μπορούμε να πούμε ότι έκανε πολύ ζέστη στην υπό εξέταση πόλη κατά τη διάρκεια και της συγκεκριμένης ημέρας. ■

Πώς πρέπει να επιλέγονται οι τετμημένες δεδομένων, αν αυτό είναι εφικτό, για προσεγγίσεις μικρότερου βαθμού;

Για να διατηρήσουμε την  $|\psi_n(x)|$  μικρή πρέπει να προσπαθήσουμε να συσταδοποιήσουμε τα δεδομένα κοντά στα σημεία όπου θέλουμε να κάνουμε τις προσεγγίσεις. Επίσης, καλό είναι να αποφεύγεται η παρεκβολή αν υπάρχει τέτοια επιλογή. Τελευταίο αλλά εξίσου σημαντικό είναι και το γεγονός ότι πρέπει να αντιμετωπίζουμε την πολυωνυμική παρεμβολή ως έναν μηχανισμό για τοπική προσέγγιση για την καθολική προσέγγιση συγκεκριμένων δεδομένων πρέπει να χρησιμοποιούμε τμηματικά πολυώνυμα (δείτε το Κεφάλαιο 11).

Ασκήσεις γι' αυτή την ενότητα: 15–16.

## 10.6 Παρεμβολή Chebyshev

Έστω ότι δίνεται μια λεία συνάρτηση  $f(x)$  και θέλουμε να βρούμε μια παρεμβολή καλής ποιότητας σε ολόκληρο το διάστημα  $[a, b]$ . Είμαστε ελεύθεροι να επιλέξουμε τις  $n + 1$  τετμημένων δεδομένων,  $x_0, x_1, \dots, x_n$ : Πώς πρέπει να επιλέξουμε αυτά τα σημεία;

Θεωρήστε την εξίσωση για το σφάλμα παρεκβολής από την προηγούμενη ενότητα (σελίδα 499). Ας υποθέσουμε περαιτέρω ότι δεν υπάρχουν πρόσθετες πληροφορίες για τα  $\xi$  ή ακόμα και για την ίδια την  $f$ , εκτός από την εγγύηση ότι μπορεί να δειγματοληφθεί οπουδήποτε, και ότι υπάρχει η φραγμένη  $(n + 1)$ -οστή παράγωγός

της: Αυτό είθισται να συμβαίνει σε πολλές εφαρμογές. Τότε, το καλύτερο που μπορούμε να κάνουμε για να ελαχιστοποιήσουμε το σφάλμα  $\max_{a \leq x \leq b} |f(x) - p_n(x)|$  είναι να επιλέξουμε τέτοιες τετμημένες δεδομένων ώστε να ελαχιστοποιείται η ποσότητα  $\max_{a \leq x \leq b} |\psi_n(x)|$ .

Έτσι οδηγούμαστε στην επιλογή των **σημείων Chebyshev**. Τα σημεία αυτά ορίζονται στο διάστημα  $[-1, 1]$  ως

$$x_i = \cos\left(\frac{2i+1}{2(n+1)}\pi\right), \quad i = 0, \dots, n$$

Για ένα γενικό διάστημα  $[a, b]$ , εφαρμόζουμε τον αφινικό μετασχηματισμό (affine transformation) που απεικονίζει το  $[-1, 1]$  στο  $[a, b]$  για να μετατοπίσουμε και να αλλάξουμε την κλίμακα των σημείων Chebyshev. Επομένως, με τον μετασχηματισμό

$$x = a + \frac{b-a}{2}(t+1), \quad t \in [-1, 1]$$

ορίζουμε εκ νέου τις τετμημένες της παρεμβολής ως

$$x_i \leftarrow a + \frac{b-a}{2}(x_i + 1), \quad i = 0, \dots, n$$

## Σφάλμα παρεμβολής με τη χρήση σημείων Chebyshev

Θα επιλέξουμε να παραμείνουμε και εδώ στο διάστημα  $[-1, 1]$ . Τα σημεία Chebyshev είναι ρίζες του πολυωνύμου Chebyshev, το οποίο ορίζεται και περιγράφεται λεπτομερώς στην Ενότητα 12.4. Επομένως, το μονικό πολυώνυμο Chebyshev (δηλαδή το πολυώνυμο Chebyshev που έχει υποστεί τέτοια αλλαγή κλίμακας ώστε ο μεγιστοβάθμιος συντελεστής του να είναι ίσος με 1) συμβολίζεται με  $\psi_n(x)$ , όπου τα  $x_i$  είναι τα σημεία Chebyshev. Όπως θα εξηγήσουμε λεπτομερώς στην Ενότητα 12.4, η μέγιστη απόλυτη τιμή αυτού του πολυωνύμου στο διάστημα παρεμβολής είναι ίση με  $2^{-n}$ . Άρα, τα  $n+1$  σημεία Chebyshev που ορίστηκαν παραπάνω επιλύουν το πρόβλημα **min-max**

$$\beta = \min_{x_0, x_1, \dots, x_n} \max_{-1 \leq x \leq 1} |(x - x_0)(x - x_1) \cdots (x - x_n)|$$

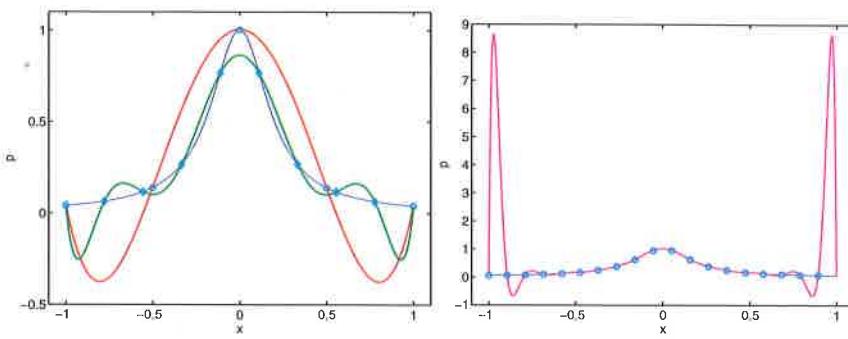
και δίνουν την τιμή  $\beta = 2^{-n}$ . Αυτό οδηγεί στο ακόλουθο φράγμα του σφάλματος παρεμβολής:

$$\max_{-1 \leq x \leq 1} |f(x) - p_n(x)| \leq \frac{1}{2^n(n+1)!} \max_{-1 \leq t \leq 1} |f^{(n+1)}(t)|$$

**Παράδειγμα 10.6.** Πριν από πολύ καιρό ο C. Runge περιέγραψε το φαινομενικά αθώο παράδειγμα

$$f(x) = \frac{1}{1 + 25x^2}, \quad -1 \leq x \leq 1$$

Τα αποτελέσματα της πολυωνυμικής παρεμβολής σε 5, 10 και 20 ισαπέχοντα σημεία έχουν σχεδιαστεί στο Σχήμα 10.6. Προσέξτε τη διαφορετική κλίμακα στον άξονα γ μεταξύ των δύο γραφικών παραστάσεων: Η προσέγγιση γίνεται χειρότερη για μεγαλύτερες τιμές του  $n$ !

(α)  $n = 4, 9$ .(β)  $n = 19$ .

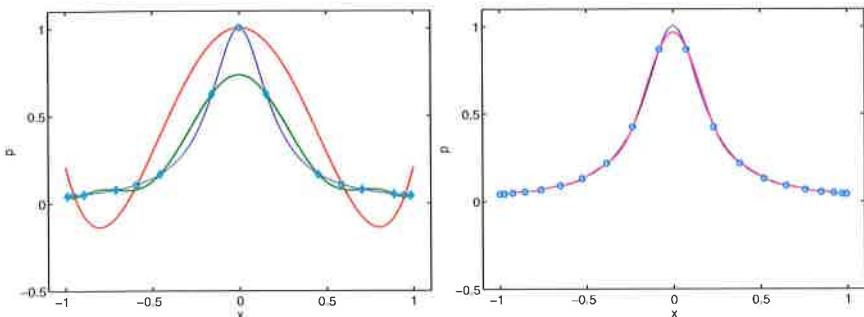
**Σχήμα 10.6:** Η καθολική πολυωνυμική παρεμβολή σε ομοιόμορφα τοποθετημένες τετμημένες μπορεί να μην είναι καλή. Εδώ, η μπλε καμπύλη με ένα μέγιστο είναι η συνάρτηση Runge του Παραδείγματος 10.6, και οι άλλες καμπύλες είναι πολυώνυμα παρεμβολής βαθμού  $n$ .

Υπολογίζοντας το  $\frac{f^{(n+1)}}{(n+1)!}$  μπορούμε να δούμε ότι ο όρος του σφάλματος μεγαλώνει κοντά στα άκρα του διαστήματος, και αυτό εξηγεί το γεγονός ότι τα αποτελέσματα δεν βελτιώνονται καθώς αυξάνεται ο βαθμός του πολυωνύμου.<sup>53</sup>

Στη συνέχεια επαναλαμβάνουμε το πείραμα, αυτή τη φορά χρησιμοποιώντας τα σημεία Chebyshev για τετμημένες των σημείων δεδομένων τίποτε άλλο δεν αλλάζει. Τα πολύ καλύτερα αποτελέσματα που προκύπτουν έχουν σχεδιαστεί στο Σχήμα 10.7. ■

Η βελτίωση στο Παράδειγμα 10.6 που προκύπτει από την παρεμβολή στα σημεία Chebyshev, συγκριτικά με την πολυωνυμική παρεμβολή σε ισαπέχοντα σημεία, είναι αρκετά σημαντική. Ίσως μπείτε στον πειρασμό να σκεφτείτε ότι είναι λίγο «ειδική», με την έννοια ότι τα σημεία Chebyshev είναι περισσότερο συγκεντρωμένα

<sup>53</sup> Ωστόσο υπάρχει ένα κλασικό θεώρημα από τον Weierstrass που ορίζει ότι είναι εφικτό να βρεθούν αυθαίρετα κοντινές πολυωνυμικές προσεγγίσεις οποιασδήποτε συνάρτησης.

(a)  $n = 4, 9.$ (b)  $n = 19.$ 

**Σχήμα 10.7:** Πολυωνυμική παρεμβολή στα σημεία Chebyshev (Παράδειγμα 10.6). Τα αποτελέσματα είναι πολύ πιο βελτιωμένα συγκριτικά με το Σχήμα 10.6, ειδικά για μεγαλύτερες τιμές του  $n$ .

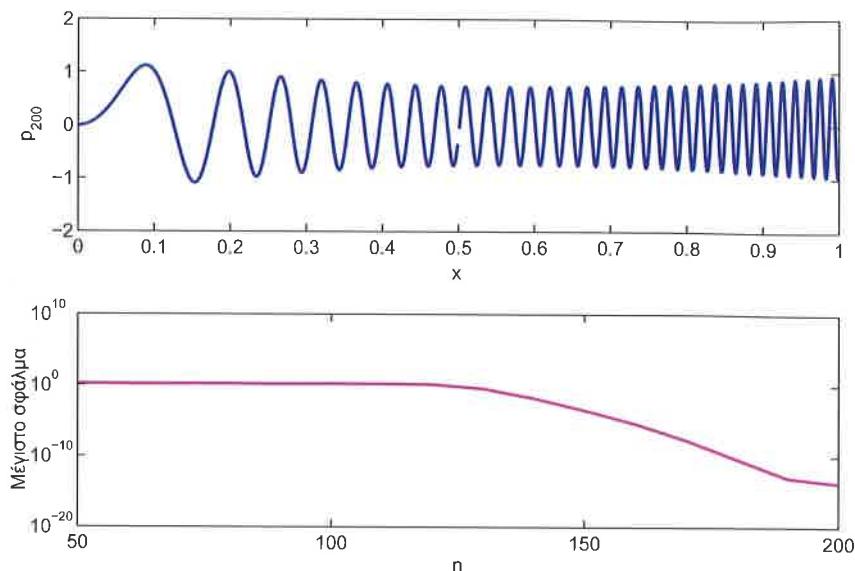
κοντά στα άκρα του διαστήματος, όπου ακριβώς μεγαλώνει η τιμή του  $\frac{f^{(n+1)}}{(n+1)!}$  για το συγκεκριμένο παράδειγμα, ενώ η υπόθεση με βάση την οποία ξεκινήσαμε ήταν ότι δεν γνωρίζουμε πολλά για την  $f$ . Δεν είναι δύσκολο να κατασκευάσει κανείς παραδείγματα όπου η παρεμβολή Chebyshev δεν θα έχει καλά αποτελέσματα. Παρ' όλα αυτά, η παρεμβολή στα σημεία Chebyshev δίνει πολύ καλά αποτελέσματα σε αρκετές περιπτώσεις, και πολλές σημαντικές αριθμητικές μέθοδοι για την επίλυση διαφορικών εξισώσεων βασίζονται σε αυτή.

**Σημείωση:** Μια πιο ολοκληρωμένη αιτιολόγηση της εντυπωσιακής συμπειροφάς της παρεμβολής Chebyshev θα απαιτούσε περισσότερη εμβάθυνση στη θεωρία προσέγγισης από αυτή που σκοπεύουμε να κάνουμε εμείς σε αυτό το βιβλίο. Ένα πρακτικό αποτέλεσμα είναι ότι αυτό είναι το μόνο σημείο στο παρόν κεφάλαιο όπου είναι φυσικό να επιτρέψουμε στο  $n$  να πάρει μεγάλες τιμές. Ένα συμπέρασμα που μπορεί να εξαχθεί από αυτή την τελευταία παρατήρηση είναι το εξής: Για να είναι ακριβής η παρεμβολή Chebyshev, πρέπει γενικά να χρησιμοποιείται (βαρυκεντρική) παρεμβολή Lagrange (δείτε τη σελίδα 485).

**Παράδειγμα 10.7.** Θέλουμε να βρούμε μια όσο το δυνατόν ακριβέστερη πολυωνυμική προσέγγιση για τη συνάρτηση

$$f(x) = e^{3x} \sin(200x^2)/(1 + 20x^2), \quad 0 \leq x \leq 1$$

Η γραφική παράσταση της συνάρτησης αυτής παρουσιάζεται στο πάνω τμήμα του Σχήματος 10.8. Σίγουρα δεν μοιάζει με πολυώνυμο μικρού βαθμού.



**Σχήμα 10.8:** Πάνω: Η συνάρτηση του Παραδείγματος 10.7 είναι πανομοιότυπη με το πολυώνυμο παρεμβολής της στα 201 σημεία Chebyshev. Κάτω: Το μέγιστο σφάλμα πολυωνυμικής παρεμβολής ως συνάρτηση του βαθμού των πολυωνύμων. Όταν ο βαθμός διπλασιάζεται από  $n = 100$  σε  $n = 200$ , το σφάλμα μειώνεται από τη μη αποδεκτή τιμή ( $> 1$ ) σχεδόν στο επίπεδο της μονάδας στρογγυλοποίησης.

Έχουμε παρεμβάλλει τη συγκεκριμένη συνάρτηση στα  $n+1$  σημεία Chebyshev για τιμές του  $n$  από 50 έως 200 με βήμα αύξησης το 10. Χρησιμοποιήθηκε η αναπαράσταση Lagrange. Στο Σχήμα 10.8 έχει σχεδιαστεί επίσης η γραφική παράσταση του συνολικού μέγιστου σφάλματος (μετρημένου στο ομοιόμορφο πλέγμα υπολογισμού  $x = 0:0.001:1$ ). Παρατηρήστε ότι, για βαθμούς πολυωνύμων έως 100 ή μεγαλύτερους, το σφάλμα παρεμβολής είναι αρκετά μεγάλο. Άλλα καθώς το  $n$  αυξάνεται περαιτέρω, το σφάλμα τελικά μειώνεται, και μάλιστα αρκετά γρήγορα: Τότε αντιστοιχεί περίπου σε  $O(q^{-n})$  για κάποιο  $q > 1$ . Αυτή είναι η επονομαζόμενη **φασματική ακρίβεια** (spectral accuracy). Για  $n = 200$ , το σφάλμα βρίσκεται ήδη στο επίπεδο της μονάδας στρογγυλοποίησης.

Για τη συνάρτηση Runge του Παραδείγματος 10.6 τα πράγματα φαίνονται ακόμα καλύτερα: Η φασματική ακρίβεια παρατηρείται ήδη για μέτριες τιμές του  $n$  (δείτε την Άσκηση 19). Από την άλλη, όμως, πρόκειται για μια ειδική περίπτωση συνάρτησης.

Σε πολλές εφαρμογές απαιτείται λελογισμένη αλλά όχι υπέρμετρα μεγάλη ακρίβεια, και η αρχική καθυστέρηση στη βελτίωση που παρατηρείται στο κάτω τμήμα του Σχήματος 10.8 μπορεί να θεωρηθεί ανησυχητική: Μόνο για περίπου  $n = 150$  τα

αποτελέσματα είναι οπτικά αποδεκτά, και το γεγονός ότι για ακόμα μεγαλύτερες τιμές του  $n$  αρχίζουμε να έχουμε εντυπωσιακή ακρίβεια δεν είναι πάντα κρίσιμης σημασίας, αν και δεν πιστεύουμε ότι υπάρχει κανείς που θα παραπονιόταν για την υπερβολική ακρίβεια. ■

Ασκήσεις γι' αυτή την ενότητα: 17–20.

## 10.7 Παρεμβολή των τιμών των παραγώγων

Συχνά θέλουμε να βρούμε ένα πολυώνυμο παρεμβολής  $p_n(x)$  το οποίο να παρεμβάλλεται όχι μόνο στις τιμές μιας συνάρτησης αλλά και στις τιμές των παραγώγων της σε σημεία που δίνονται.

**Παράδειγμα 10.8.** Ένας άνθρωπος πετάει μια πέτρα υπό γωνία  $45^\circ$ , και η πέτρα πέφτει στο έδαφος μετά από πτήση πέντε μέτρων. Η τροχιά της πέτρας,  $y = f(x)$ , υπακούει στις εξισώσεις της κίνησης, λαμβάνοντας υπόψη τη βαρύτητα και ενδεχομένως και την αεροδυναμική οπισθέλκουσα (air drag). Άλλα εμείς θέλουμε να βρούμε μια γρήγορη προσέγγιση και γνωρίζουμε ήδη κάποιες πληροφορίες για την  $f$ , συγκεκριμένα ότι  $f(0) = 1.5$ ,  $f'(0) = 1$  και  $f(5) = 0$ . Επομένως, προσαρμόζουμε μια τετραγωνική συνάρτηση σε αυτά τα σημεία δεδομένων.

Χρησιμοποιώντας μια μονωνυμική βάση γράφουμε

$$p_2(x) = c_0 + c_1x + c_2x^2, \quad p'_2(x) = c_1 + 2c_2x$$

Αντικαθιστώντας τα δεδομένα, παίρνουμε τρεις εξισώσεις για τους συντελεστές:  $1.5 = p_2(0) = c_0$ ,  $1 = p'_2(0) = c_1$  και

$$0 = p_2(5) = c_0 + 5c_1 + 25c_2$$

Το πολυώνυμο παρεμβολής που προκύπτει (παρακαλούμε να το επαληθεύσετε) είναι

$$p_2(x) = 1.5 + x - 0.26x^2$$

και έχει σχεδιαστεί στο Σχήμα 10.9. ■

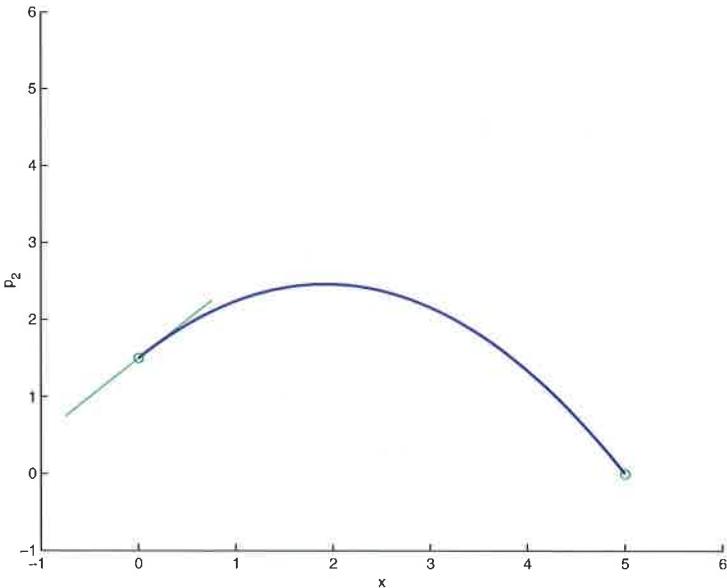
## Το γενικό πρόβλημα

Η γενική περίπτωση απαιτεί πιο περίπλοκο συμβολισμό. Ας υποθέσουμε ότι υπάρχουν  $q + 1$  διαφορετικές τετμημένες

$$t_0, t_1, t_2, \dots, t_q$$

και  $q + 1$  μη αρνητικοί ακέραιοι

$$m_0, m_1, m_2, \dots, m_q$$



**Σχήμα 10.9:** Τετραγωνικό πολυώνυμο παρεμβολής  $p_2(x)$  που ικανοποιεί τις ισότητες  $p_2(0) = 1.5$ ,  $p'_2(0) = 1$  και  $p_2(5) = 0$ .

Θέλουμε να βρούμε το μοναδικό **εφάπτον πολυώνυμο** (osculating polynomial)<sup>54</sup> μικρότερου βαθμού που ικανοποιεί την εξίσωση

$$p_n^{(k)}(t_i) = f^{(k)}(t_i) \quad (k = 0, 1, \dots, m_i), \quad i = 0, 1, \dots, q$$

Αν μετρήσουμε πόσες συνθήκες πρέπει να ικανοποιούνται, είναι προφανές ότι ο βαθμός του πολυωνύμου παρεμβολής είναι το πολύ ίσος με

$$n = \sum_{k=0}^q m_k + q$$

Άρα, οι τιμές της πολυωνυμικής συνάρτησης  $p_n$  και των πρώτων  $m_i$  παραγάγων της συμπίπτουν με τις αντίστοιχες τιμών των πρώτων  $m_i$  παραγάγων της  $f$  σε κάθε σημείο παρεμβολής. Αυτό αποτελεί μια γενίκευση που δεν περιορίζεται μόνο στα όσα έχουμε δει προηγουμένως. Στα σχετικά παραδείγματα περιλαμβάνονται τα εξής:

1. Αν ισχύει  $q = n$ ,  $m_i = 0$  για κάθε  $i$ , προκύπτει η πολυωνυμική παρεμβολή στις τιμές της συνάρτησης όπως έχουμε δει και προηγουμένως.
2. Αν ισχύει  $q = 0$ , προκύπτει το πολυώνυμο *Taylor* βαθμού  $m_0$  στο  $t_0$ .
3. Αν ισχύει  $n = 2q + 1$ ,  $m_i = 1$  για κάθε  $i$ , προκύπτει η **παρεμβολή Hermite** (Hermite interpolation).

<sup>54</sup> Η αγγλική λέξη «osculating» προέρχεται από τη λατινική λέξη για το «φιλί» (kiss).

Αυτή η λίστα περιέχει τις πιο χρήσιμες περιπτώσεις, αν και η παραπάνω διατύπωση δεν περιορίζεται σε αυτές τις επιλογές.

## Κυβική παρεμβολή Hermite

Η πιο δημοφιλής συνάρτηση παρεμβολής με εφάπτοντα πολυώνυμα είναι η *κυβική συνάρτηση Hermite* (Hermite cubic), που προκύπτει για  $q = m_0 = m_1 = 1$ . Χρησιμοποιώντας τη μονωνυμική βάση μπορούμε να γράψουμε τις συνθήκες παρεμβολής ως

$$\begin{aligned} c_0 + c_1 t_0 + c_2 t_0^2 + c_3 t_0^3 &= f(t_0), & c_1 + 2c_2 t_0 + 3c_3 t_0^2 &= f'(t_0) \\ c_0 + c_1 t_1 + c_2 t_1^2 + c_3 t_1^3 &= f(t_1), & c_1 + 2c_2 t_1 + 3c_3 t_1^2 &= f'(t_1) \end{aligned}$$

και να λύσουμε αυτές τις τέσσερις γραμμικές εξισώσεις ως προς τους συντελεστές  $c_j$ .

## Κατασκευή του εφάπτοντος πολυωνύμου

Μια γενική μέθοδος κατασκευής του εφάπτοντος πολυωνύμου μπορεί να επινοηθεί εύκολα με την επέκταση της μορφής Newton και τις διαιρεμένες διαφορές της Ενότητας 10.4. Ορίζουμε το σύνολο των τετμημένων επαναλαμβάνοντας  $m_i + 1$  φορές καθένα από τα σημεία  $t_i$  στην ακολουθία, οπότε προκύπτει η ακολουθία

$$(x_0, x_1, x_2, \dots, x_n) = \left( \underbrace{t_0, t_0, \dots, t_0}_{m_0+1}, \underbrace{t_1, \dots, t_1}_{m_1+1}, \dots, \underbrace{t_q, \dots, t_q}_{m_q+1} \right)$$

Οι αντίστοιχες τιμές δεδομένων είναι

$$\begin{aligned} (y_0, y_1, y_2, \dots, y_n) \\ = (f(t_0), f'(t_0), \dots, f^{(m_0)}(t_0), f(t_1), \dots, f^{(m_1)}(t_1), \dots, f(t_q), \dots, f^{(m_q)}(t_q)) \end{aligned}$$

Στη συνέχεια χρησιμοποιούμε τη μορφή παρεμβολής Newton

$$p_n(x) = \sum_{j=0}^n f[x_0, x_1, \dots, x_j] \prod_{i=0}^{j-1} (x - x_i)$$

όπου

$$f[x_k, \dots, x_j] = \begin{cases} \frac{f[x_{k+1}, \dots, x_j] - f[x_k, \dots, x_{j-1}]}{x_j - x_k}, & x_k \neq x_j, \\ \frac{f^{(j-k)}(x_k)}{(j-k)!}, & x_k = x_j \end{cases}$$

για  $0 < k \leq j \leq n$ . Προσέξτε ότι οι επαναλαμβανόμενες τιμές πρέπει να είναι συνεχόμενες στην ακολουθία  $\{x_i\}_{i=0}^n$ . Στη συνέχεια παρουσιάζεται ο αλγόριθμος που προκύπτει.

### Αλγόριθμος: Πολυωνυμική παρεμβολή σε μορφή Newton.

1. *Κατασκευή:* Για τα δεδομένα  $\{(x_i, y_i)\}_{i=0}^n$ , όπου οι τετμημένες δεν είναι υποχρεωτικά διαφορετικές,

$$\begin{aligned} &\text{for } j = 0, 1, \dots, n \\ &\text{for } l = 0, 1, \dots, j \\ &\gamma_{j,l} = \begin{cases} \frac{\gamma_{j,l-1} - \gamma_{j-1,l-1}}{x_j - x_{j-l}} & \text{αν } x_j \neq x_{j-l}, \\ \frac{f^{(l)}(x_j)}{l!} & \text{διαφορετικά} \end{cases} \end{aligned}$$

2. *Υπολογισμός:* Για το σημείο υπολογισμού  $x$ ,

$$\begin{aligned} p &= \gamma_{n,n} \\ &\text{for } j = n-1, n-2, \dots, 0 \\ p &= p(x - x_j) + \gamma_{j,j} \end{aligned}$$

Με τον ορισμό αυτής της μεθόδου, η εξίσωση για το σφάλμα της πολυωνυμικής παρεμβολής που βρήκαμε στην Ενότητα 10.5 επίσης επεκτείνεται απρόσκοπτα στην περίπτωση της παρεμβολής με εφάπτοντα πολυώνυμα.

**Παράδειγμα 10.9.** Για τη συνάρτηση  $f(x) = \ln(x)$  έχουμε τις τιμές  $f(1) = 0$ ,  $f'(1) = 1$ ,  $f(2) = 0.693147$ ,  $f'(2) = 0.5$ . Θα κατασκευάσουμε την αντίστοιχη κυβική συνάρτηση παρεμβολής Hermite.

Χρησιμοποιώντας την απλή αλλά όχι γενική διαδικασία της μονωνυμικής βάσης που περιγράψαμε παραπάνω, βρίσκουμε άμεσα τη συνάρτηση παρεμβολής

$$p(x) = -1.53426 + 2.18223x - 0.761675x^2 + 0.113706x^3$$

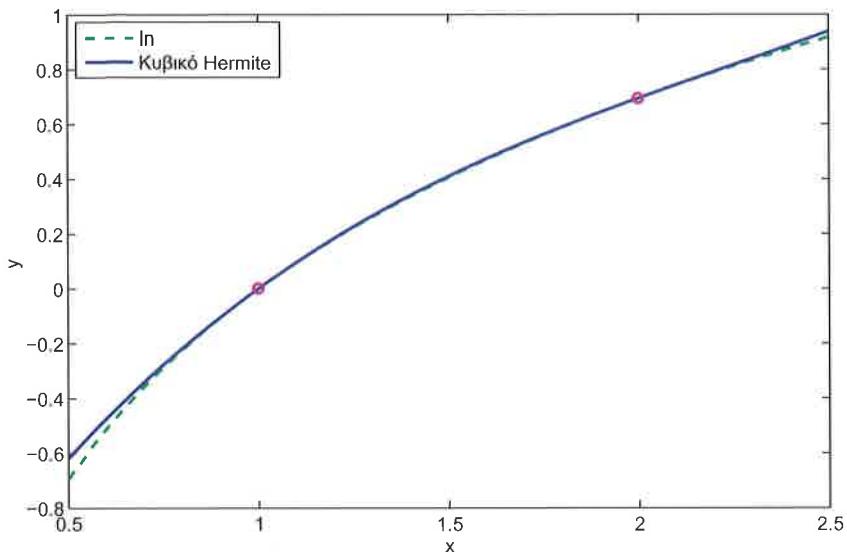
Χρησιμοποιώντας τον γενικό αλγόριθμο που παρουσιάζεται σε αυτή τη σελίδα παίρνουμε την εναλλακτική αναπαράσταση

$$p(x) = (x - 1) - 0.30685(x - 1)^2 + 0.113706(x - 1)^2(x - 2)$$

Η συνάρτηση και η αντίστοιχη συνάρτηση παρεμβολής της απεικονίζονται στο Σχήμα 10.10.

Για παράδειγμα,  $p_3(1.5) = 0.409074$ , ενώ  $\ln(1.5) = 0.405465$ .

Για να εκτιμήσουμε το σφάλμα στο  $x = 1.5$ , προσποιούμενοι ότι δεν μπορούμε να υπολογίσουμε την τιμή της  $f$  εκεί, θα χρησιμοποιήσουμε την ίδια εξίσωση όπως



**Σχήμα 10.10:** Το κυβικό εφάπτον πολυώνυμο Hermite για τη συνάρτηση  $\ln(x)$  στα σημεία 1 και 2.

και προηγουμένως: Επειδή ισχύει  $|f'''(x)| = |6x^{-4}| \leq 6$ , προκύπτει ότι το σφάλμα είναι φραγμένο από το

$$\frac{1}{4}(1.5 - 1)^2(1.5 - 2)^2 = 0.015625$$

Όπως αποδεικνύεται, το νούμερο αυτό είναι περίπου 4 φορές μεγαλύτερο από το πραγματικό σφάλμα. ■

**Παράδειγμα 10.10.** Δίνονται οι πέντε τιμές δεδομένων

$t_i$	$f(t_i)$	$f'(t_i)$	$f''(t_i)$
8.3	17.564921	3.116256	0.120482
8.6	18.505155	3.151762	

Κατασκευάζουμε τον πίνακα διαιρεμένων διαφορών

$$(x_0, x_1, x_2, x_3, x_4) = \left( \underbrace{8.3, 8.3, 8.3}_{m_0=2}, \underbrace{8.6, 8.6}_{m_1=1} \right)$$

$$f[x_0, x_1] = \frac{f'(t_0)}{1!} = f'(8.3), \quad f[x_1, x_2] = \frac{f'(t_0)}{1!}$$

$$f[x_0, x_1, x_2] = \frac{f''(t_0)}{2!} = \frac{f''(8.3)}{2}, \quad f[x_3, x_4] = \frac{f'(t_1)}{1!} = f'(8.6)$$

κ.ο.κ. Στον παρακάτω πίνακα έχουν υπογραμμιστεί οι τιμές των παραγώγων που δίνονται αρχικά.

$x_i$	$f[\cdot]$	$f[\cdot, \cdot]$	$f[\cdot, \cdot, \cdot]$	$f[\cdot, \cdot, \cdot, \cdot]$	$f[\cdot, \cdot, \cdot, \cdot, \cdot]$
8.3	17.564921				
8.3	17.564921	<u>3.116256</u>			
8.3	17.564921	<u>3.116256</u>	<u>0.060241</u>		
8.6	18.505155	3.134113	0.059524	-0.002389	
8.6	18.505155	<u>3.151762</u>	0.058829	-0.002319	0.000233

Η τεταρτοβάθμια συνάρτηση παρεμβολής που προκύπτει είναι

$$\begin{aligned} p_4(x) &= \sum_{k=0}^4 f[x_0, \dots, x_k] \prod_{j=0}^{k-1} (x - x_j) \\ &= 17.564921 + 3.116256(x - 8.3) + 0.060241(x - 8.3)^2 \\ &\quad - 0.002389(x - 8.3)^3 + 0.000233(x - 8.3)^3(x - 8.6) \quad \blacksquare \end{aligned}$$

Ασκήσεις γι' αυτή την ενότητα: 21–25.

## 10.8 Ασκήσεις

### 0. Ερωτήσεις επανάληψης

- (α) Πώς διακρίνονται μεταξύ τους οι όροι προσαρμογή δεδομένων, παρεμβολή, και πολυωνυμική παρεμβολή;
- (β) Πώς διακρίνεται η (διακριτή) προσαρμογή δεδομένων από την προσέγγιση μιας συνάρτησης;
- (γ) Τι είναι οι συναρτήσεις βάσης; Πρέπει μια συνάρτηση παρεμβολής  $v(x)$  η οποία είναι γραμμένη ως γραμμικός συνδυασμός συναρτήσεων βάσης να είναι γραμμική ως προς το  $x$ ;
- (δ) Ένα πολυώνυμο παρεμβολής είναι μοναδικό ανεξάρτητα από την επιλογή της βάσης. Εξηγήστε γιατί.

- (ε) Αναφέρετε ένα πλεονέκτημα και δύο μειονεκτήματα που έχει η χρήση της μονωνυμικής βάσης για πολυωνυμική παρεμβολή.
- (στ) Τι είναι τα πολυνόμια Lagrange; Πώς χρησιμοποιούνται για πολυωνυμική παρεμβολή;
- (ζ) Τι είναι οι βαρυκεντρικοί συντελεστές;
- (η) Αναφέρετε τα κύρια πλεονεκτήματα και το κύριο μειονέκτημα που έχει η χρήση της αναπαράστασης Lagrange.
- (θ) Τι είναι ένας πίνακας διαιρεμένων διαφορών και πώς κατασκευάζεται;
- (ι) Γράψτε την εξίσωση για την πολυωνυμική παρεμβολή σε μορφή Newton.
- (ια) Αναφέρετε δύο πλεονεκτήματα και δύο μειονεκτήματα που έχει η χρήση της αναπαράστασης Newton για πολυωνυμική παρεμβολή.
- (ιβ) Περιγράψτε τα γραμμικά συστήματα που λύνονται ως προς τη μονωνυμική βάση, την αναπαράσταση Lagrange και την αναπαράσταση Newton.
- (ιγ) Περιγράψτε τη σχέση που συνδέει την  $k$ -οστή διαιρεμένη διαφορά μιας συνάρτησης  $f$  με την  $k$ -οστή παράγωγό της.
- (ιδ) Γράψτε μια εξίσωση για το σφάλμα της πολυωνυμικής παρεμβολής και μια εξίσωση για το φράγμα του σφάλματος.
- (ιε) Πώς επηρεάζει γενικά η λειότητα (δηλαδή η παραγωγισμότητα) μιας συνάρτησης και οι παράγωγοί της την ποιότητα των πολυωνύμων παρεμβολής που την προσεγγίζουν;
- (ιστ) Δώστε ένα παράδειγμα στο οποίο προκύπτει το φράγμα του σφάλματος.
- (ιζ) Όταν παρεμβάλλουμε μια συνάρτηση  $f$  έχοντας στη διάθεσή μας μόνο σημεία δεδομένων, δηλαδή όταν δεν γνωρίζουμε την  $f$  ή τις παραγώγους της, πώς μπορούμε να μετρήσουμε την ακρίβεια της προσέγγισής μας;
- (ιη) Τι είναι τα σημεία Chebyshev και γιατί είναι σημαντικά;
- (ιθ) Περιγράψτε την παρεμβολή με εφάπτοντα πολυνόμια. Πώς διαφέρει από τη συνήθη πολυωνυμική παρεμβολή;
- (ικ) Τι είναι μια κυβική συνάρτηση παρεμβολής Hermite;
1. Βρείτε τη γραμμική συνάρτηση παρεμβολής που διέρχεται από τα σημεία δεδομένων  $(1.0, 2.0)$  και  $(1.1, 2.5)$ . Βρείτε επίσης την τετραγωνική συνάρτηση παρεμβολής που διέρχεται από αυτά τα δύο σημεία και από το  $(1.2, 1.5)$ . Δείξτε ότι η κατάσταση μπορεί να απεικονιστεί όπως στο Σχήμα 10.11.

2. Ορισμένα ζητήματα μοντελοποίησης καθιστούν υποχρεωτική την αναζήτηση για μια συνάρτηση

$$u(x) = \gamma_0 e^{\gamma_1 x + \gamma_2 x^2}$$

όπου οι άγνωστοι συντελεστές  $\gamma_1$  και  $\gamma_2$  αναμένεται να είναι μη θετικοί. Δίνονται τα ζεύγη δεδομένων προς παρεμβολή,  $(x_0, z_0)$ ,  $(x_1, z_1)$  και  $(x_2, z_2)$ , όπου  $z_i > 0$ ,  $i = 0, 1, 2$ . Άρα, απαιτούμε να ισχύει  $u(x_i) = z_i$ .

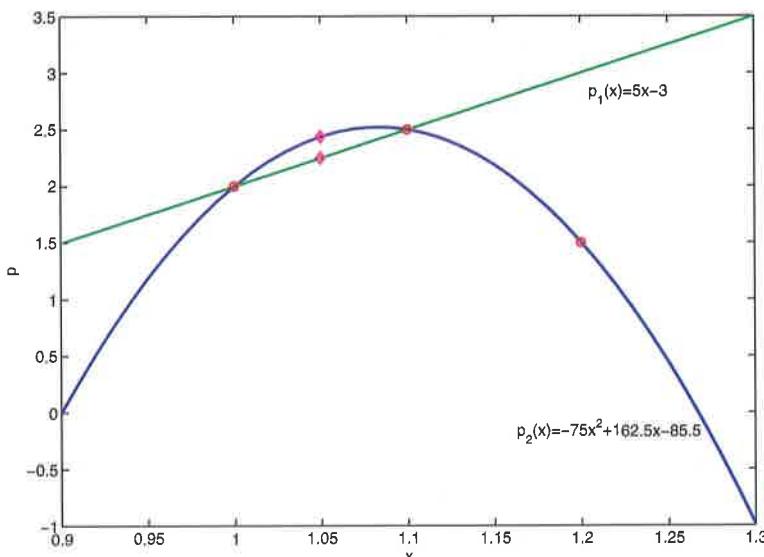
Η συνάρτηση  $u(x)$  δεν είναι γραμμική ως προς τους συντελεστές της, αλλά η  $v(x) = \ln(u(x))$  είναι γραμμική ως προς τους δικούς της.

- (α) Βρείτε ένα τετραγωνικό πολυώνυμο  $v(x)$  το οποίο να παρεμβάλλεται σε τρία κατάλληλα επιλεγμένα ζεύγη δεδομένων, και μετά βρείτε μια εξίσωση για τη  $u(x)$  σε σχέση με τα αρχικά δεδομένα.

[Αυτή είναι μια άσκηση που πρέπει να λυθεί με μολύβι και χαρτί: στην επόμενη άσκηση πρέπει να αφιερώσετε πολύ λιγότερο από τον χρόνο σας.]

- (β) Γράψτε κώδικα ο οποίος θα βρίσκει τη  $u$  για τα ζεύγη δεδομένων  $(0, 1)$ ,  $(1, 0.9)$ ,  $(3, 0.5)$ . Παρέχετε τους συντελεστές  $\gamma_i$  και σχεδιάστε τη συνάρτηση παρεμβολής που προκύπτει στο διάστημα  $[0, 6]$ . Με ποιον τρόπο συμπεριφέρεται η καμπύλη διαφορετικά –από ποιοτική άποψη– συγκριτικά με μια τετραγωνική;

3. Χρησιμοποιήστε τις γνωστές τιμές της συνάρτησης  $\sin(x)$  στα  $x = 0, \pi/6, \pi/4, \pi/3$  και  $\pi/2$  για να βρείτε ένα πολυώνυμο παρεμβολής  $p(x)$ . Ποιος είναι ο βαθμός του πολυωνύμου σας;



**Σχήμα 10.11:** Τετραγωνική και γραμμική πολυωνυμική παρεμβολή.

Ποιο είναι το μέγεθος  $|p(1.2) - \sin(1.2)|$  του σφάλματος παρεμβολής;

4. Δίνονται  $n + 1$  ζεύγη δεδομένων  $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$ . Ορίστε για  $j = 0, 1, \dots, n$  τις συναρτήσεις  $\rho_j = \prod_{i \neq j} (x_j - x_i)$ , και έστω επίσης ότι  $\psi(x) = \prod_{i=0}^n (x - x_i)$ .

(α) Δείξτε ότι

$$\rho_j = \psi'(x_j)$$

(β) Δείξτε ότι το πολυώνυμο παρεμβολής βαθμού το πολύ  $n$  είναι

$$p_n(x) = \psi(x) \sum_{j=0}^n \frac{y_j}{(x - x_j)\psi'(x_j)}$$

5. Κατασκευάστε ένα πολυώνυμο  $p_3(t)$  βαθμού το πολύ τρία σε μορφή Lagrange το οποίο να παρεμβάλλεται στα δεδομένα

t	-1.1	1.1	2.2	0.0
y	0.0	6.75	0.0	0.0

6. Δίνονται τα τέσσερα σημεία δεδομένων  $(-1, 1), (0, 1), (1, 2), (2, 0)$ . Βρείτε το κυβικό πολυώνυμο παρεμβολής χρησιμοποιώντας

- τη μονωνυμική βάση·
- τη βάση Lagrange·
- τη βάση Newton.

Δείξτε ότι οι τρεις αναπαραστάσεις δίνουν το ίδιο πολυώνυμο.

7. Για τη βάση Newton, αποδείξτε ότι ο συντελεστής  $c_j = f[x_0, x_1, \dots, x_j]$  ικανοποιεί την εξίσωση αναδρομής όπως αυτή ορίζεται στην Ενότητα 10.4 (σελίδα 491).

[Η συγκεκριμένη απόδειξη είναι σύντομη αλλά ενδέχεται να έχει κάποιον βαθμό δυσκολίας.]

8. Μια μυστική φόρμουλα για την αιώνια νεότητα,  $f(x)$ , ανακαλύφθηκε από τον Δρα Ταχή, ο οποίος εργάζεται σε μια εταιρεία βιοτεχνολογίας. Ωστόσο ο Δρ. Ταχής έχει εξαφανιστεί και, σύμφωνα με κάποιες φήμες, βρίσκεται σε διαπραγματεύσεις με έναν ανταγωνιστή της εταιρείας.

Από τις σημειώσεις που άφησε πίσω ο Δρ. Ταχής καθώς έφευγε βιαστικά είναι σαφές ότι  $f(0) = 0$ ,  $f(1) = 2$ , καθώς και ότι  $f[x_0, x_1, x_2] = 1$  για οποιαδήποτε τρία σημεία  $x_0, x_1, x_2$ . Βρείτε την  $f(x)$ .

9. Ο Γιάννης αποφασίζει να αγοράσει μετοχές μιας πολλά υποσχόμενης διαδικτυακής εταιρείας. Η τιμή της μετοχής είναι 100. Ο Γιάννης αρχίζει να καταγράφει την τιμή στο τέλος κάθε εβδομάδας. Με τις τετμημένες να μετριούνται σε μέρες, προέκυψαν τα ακόλουθα δεδομένα: (0, 100), (7, 98), (14, 101), (21, 50), (28, 51), (35, 50).

Για να αναλύσει το τι ακριβώς συνέβη, πρέπει να υπολογίσει κατά προσέγγιση την τιμή της μετοχής μερικές ημέρες πριν την κατάρρευσή της.

- (α) Βρείτε μια γραμμική συνάρτηση παρεμβολής που να διέρχεται από τα σημεία με τετμημένες 7 και 14. Έπειτα προσθέστε σε αυτό το σύνολο δεδομένων την τιμή στο 0 και (ξεχωριστά) την τιμή στο 21 για να πάρετε δύο τετραγωνικές συναρτήσεις παρεμβολής. Υπολογίστε την τιμή και για τις τρεις συναρτήσεις παρεμβολής στο  $x = 12$ . Ποια πιστεύετε ότι έχει τη μεγαλύτερη ακρίβεια; Αιτιολογήστε την απάντησή σας.
- (β) Σχεδιάστε τις δύο παραπάνω τετραγωνικές συναρτήσεις παρεμβολής μαζί με τα δεδομένα (χωρίς κάποια τεθλασμένη γραμμή που να διέρχεται από τα δεδομένα) στο διάστημα  $[0, 21]$ . Τι μπορείτε να παρατηρήσετε;
10. Δίνονται 137 ζεύγη δεδομένων ομοιόμορφα τοποθετημένα σε διαφορετικές τετμημένες:  $(x_i, y_i)$ ,  $i = 0, 1, \dots, 136$ . Αυτά τα δεδομένα υποτίθεται ότι αναπαριστούν μια συνάρτηση  $f(x)$  που παρήγαγε αυτές τις τιμές δεδομένων έχει πολλές φραγμένες παραγώγους παντού εκτός από λίγα σημεία όπου εμφανίζει ασυνέχειες αλμάτων. (Φανταστείτε ότι σχεδιάζετε μια καμπύλη ομαλά από τα αριστερά προς τα δεξιά, αλλά κάποιες φορές σηκώνετε το μολύβι από το χαρτί και το μετακινείτε οριζόντια, και μετά συνεχίζετε να σχεδιάζετε.) Για κάθε υποδιάστημα  $[x_{i-1}, x_i]$  θέλουμε να βρούμε το (ας ελπίσουμε όσο το δυνατόν) καλύτερο κυβικό πολυώνυμο  $p_3(x)$  που παρεμβάλλεται με ακρίβεια στην  $f(x)$  και το οποίο διέρχεται από τα σημεία  $x_{i-1} < x < x_i$ . Αυτό εμπεριέχει την επιλογή καλών γειτόνων στους οποίους θα γίνει η παρεμβολή.
- Προτείνετε έναν αλγόριθμο γι' αυτή την εργασία. Αιτιολογήστε την απάντησή σας.
11. Δίνεται μια ακολουθία  $y_0, y_1, y_2, \dots$  Ορίστε τον τελεστή της προς τα εμπρός (ή κατάντη) διαφοράς (forward difference operator)  $\Delta$  ως

$$\Delta y_i = y_{i+1} - y_i$$

Οι δυνάμεις του  $\Delta$  ορίζονται αναδρομικά από τις σχέσεις

$$\begin{aligned}\Delta^0 y_i &= y_i, \\ \Delta^j y_i &= \Delta(\Delta^{j-1} y_i), \quad j = 1, 2, \dots\end{aligned}$$

Επομένως,  $\Delta^2 y_i = \Delta(y_{i+1} - y_i) = y_{i+2} - 2y_{i+1} + y_i$  κ.λπ.

Θεωρήστε την πολυωνυμική παρεμβολή σε ισαπέχοντα σημεία,  $x_i = x_0 + ih$ ,  $i = 0, 1, \dots, n$ .

(α) Δείξτε ότι

$$f[x_0, x_1, \dots, x_j] = \frac{1}{j! h^j} \Delta^j f(x_0)$$

[Υπόδειξη: Χρησιμοποιήστε μαθηματική επαγωγή.]

(β) Δείξτε ότι το πολυώνυμο παρεμβολής βαθμού το πολύ  $n$  δίνεται από την εξίσωση *Newton* της προς τα εμπρός διαφοράς

$$p_n(x) = \sum_{j=0}^n \binom{s}{j} \Delta^j f(x_0)$$

όπου  $s = \frac{x-x_0}{h}$  και  $\binom{s}{j} = \frac{s(s-1)\cdots(s-j+1)}{j!}$  [με  $\binom{s}{0} = 1$ ].

12. Δίνεται μια ακολουθία  $y_0, y_1, y_2, \dots$ . Ορίστε τον τελεστή της προς τα πίσω (ή ανάντη) διαφοράς (backward difference operator)  $\nabla$  ως

$$\nabla y_i = y_i - y_{i-1}$$

Οι δυνάμεις του  $\nabla$  ορίζονται αναδρομικά από τις σχέσεις

$$\begin{aligned}\nabla^0 y_i &= y_i, \\ \nabla^j y_i &= \nabla(\nabla^{j-1} y_i), \quad j = 1, 2, \dots\end{aligned}$$

Συνεπώς,  $\nabla^2 y_i = \nabla(y_i - y_{i-1}) = y_i - 2y_{i-1} + y_{i-2}$  κ.λπ.

Θεωρήστε την πολυωνυμική παρεμβολή σε ισαπέχοντα σημεία,  $x_i = x_0 + ih$ ,  $i = 0, 1, \dots, n$ .

(α) Δείξτε ότι

$$f[x_n, x_{n-1}, \dots, x_{n-j}] = \frac{1}{j! h^j} \nabla^j f(x_n)$$

[Υπόδειξη: Χρησιμοποιήστε μαθηματική επαγωγή.]

(β) Δείξτε ότι το πολυώνυμο παρεμβολής βαθμού  $n$  δίνεται από την εξίσωση Newton της προς τα πίσω διαφοράς

$$p_n(x) = \sum_{j=0}^n (-1)^j \binom{s}{j} \nabla^j f(x_n)$$

όπου  $s = \frac{x_n - x}{h}$  και  $\binom{s}{j} = \frac{s(s-1)\cdots(s-j+1)}{j!}$  [με  $\binom{s}{0} = 1$ ].

13. Έστω ότι  $(\hat{x}_0, \hat{x}_1, \dots, \hat{x}_k)$  είναι μια μετάθεση των τετμημένων  $(x_0, x_1, \dots, x_k)$ . Δείξτε ότι

$$f[\hat{x}_0, \hat{x}_1, \dots, \hat{x}_k] = f[x_0, x_1, \dots, x_k]$$

[Υπόδειξη: Θεωρήστε την  $k$ -οστή παράγωγο του μοναδικού πολυωνύμου βαθμού  $k$  το οποίο παρεμβάλλεται στην  $f$  σε αυτά τα  $k + 1$  σημεία, ανεξάρτητα από το πώς είναι διατεταγμένα.]

14. Έστω ότι τα σημεία  $x_0, x_1, \dots, x_n$  είναι σταθερά. Θεωρήστε τη διαιρεμένη διαφορά  $f[x_0, x_1, \dots, x_n, x]$  ως συνάρτηση του  $x$ . (Η συνάρτηση αυτή εμφανίζεται στην εξίσωση για το σφάλμα της πολυωνυμικής παρεμβολής.)

Υποθέστε ότι η  $f(x)$  είναι πολυώνυμο βαθμού  $m$ . Δείξτε ότι

- αν  $\text{ισχύει } m \leq n$ , τότε  $f[x_0, x_1, \dots, x_n, x] \equiv 0$
- διαφορετικά, η  $f[x_0, x_1, \dots, x_n, x]$  είναι πολυώνυμο βαθμού  $m - n - 1$ .

[Υπόδειξη: Αν  $\text{ισχύει } m > n$ , αποδείξτε το ζητούμενο πρώτα για την περίπτωση  $n = 0$ . Στη συνέχεια προχωρήστε με επαγωγή, εξετάζοντας τη συνάρτηση  $g(x) = f[x_1, \dots, x_n, x]$ .]

15. Θέλουμε να προσεγγίσουμε τη συνάρτηση  $e^x$  στο διάστημα  $[0, 1]$  χρησιμοποιώντας πολυωνυμική παρεμβολή με  $x_0 = 0$ ,  $x_1 = 1/2$  και  $x_2 = 1$ . Έστω ότι το πολυώνυμο παρεμβολής είναι  $p_2(x)$ .

(α) Βρείτε ένα άνω φράγμα για το μέγεθος του σφάλματος,

$$\max_{0 \leq x \leq 1} |e^x - p_2(x)|$$

(β) Βρείτε το πολυώνυμο παρεμβολής χρησιμοποιώντας την αγαπημένη σας τεχνική.

(γ) Σχεδιάστε στο ίδιο σχήμα τη συνάρτηση  $e^x$  και το πολυώνυμο παρεμβολής που βρήκατε, χρησιμοποιώντας εντολές plot.

- (δ) Σχεδιάστε το μέγεθος  $|e^x - p_2(x)|$  του σφάλματος στο παραπάνω διάστημα χρησιμοποιώντας λογαριθμική κλίμακα (δηλαδή την εντολή `semilogy`) και επαληθεύστε με απλή εξέταση ότι βρίσκεται χαμηλότερα από το φράγμα που υπολογίσατε στο (α).
16. Για το πρόβλημα της Άσκησης 3, βρείτε ένα φράγμα για το σφάλμα της πολυωνυμικής παρεμβολής στο  $[0, \pi/2]$ .  
 Συγκρίνετε αυτό το φράγμα με το πραγματικό σφάλμα παρεμβολής στο  $x = 1.2$ .
17. Κατασκευάστε δύο παραδείγματα για οποιονδήποτε θετικό ακέραιο  $n$ : ένα όπου η παρεμβολή σε  $n + 1$  ισαπέχοντα σημεία είναι πιο ακριβής από ό,τι η παρεμβολή στα  $n + 1$  σημεία Chebyshev, και ένα όπου η παρεμβολή Chebyshev είναι πιο ακριβής.  
 Τα παραδείγματά σας πρέπει να είναι πειστικά χωρίς τη βοήθεια υπολογιστή.
18. (α) Παρεμβάλετε τη συνάρτηση  $f(x) = \sin(x)$  σε 5 σημεία Chebyshev στο διάστημα  $[0, \pi/2]$ . Συγκρίνετε τα αποτελέσματά σας με εκείνα των Ασκήσεων 3 και 16.  
 (β) Επαναλάβετε την παρεμβολή, αυτή τη φορά χρησιμοποιώντας 5 σημεία Chebyshev στο διάστημα  $[0, \pi]$ . Σχεδιάστε την  $f(x)$  και τη συνάρτηση παρεμβολής. Ποια είναι τα συμπεράσματά σας;
19. Παρεμβάλετε τη συνάρτηση Runge του Παραδείγματος 10.6 σε σημεία Chebyshev για τιμές του  $n$  από 10 έως 170 με βήμα αύξησης το 10. Υπολογίστε το μέγιστο σφάλμα παρεμβολής στο ομοιόμορφο πλέγμα υπολογισμού  $x = -1 : 0.001 : 1$  και σχεδιάστε το σφάλμα ως προς τον βαθμό των πολυωνύμου, όπως στο Σχήμα 10.8, χρησιμοποιώντας την εντολή `semilogy`. Παρατηρήστε τη φασματική ακρίβεια.
20. Τα **ακρότατα σημεία Chebyshev** (Chebyshev extremum points) είναι στενοί συγγενείς των σημείων Chebyshev. Ορίζονται στο διάστημα  $[-1, 1]$  ως

$$x_i = \xi_i = \cos\left(\frac{i}{n}\pi\right), \quad i = 0, 1, \dots, n$$

Για περισσότερες πληροφορίες σχετικά με αυτά τα σημεία δείτε την Ενότητα 12.4. Όπως και τα σημεία Chebyshev, τα ακρότατα σημεία Chebyshev τείνουν να συγκεντρώνονται κοντά στα άκρα του διαστήματος.

Επαναλάβετε τις Ασκήσεις 18 και 19, καθώς και το Παράδειγμα 10.7, παρεμβάλλοντας στα  $n + 1$  ακρότατα σημεία Chebyshev ένα πολυώνυμο βαθμού το πολύ  $n$ .

Συγκρίνετε με τα αντίστοιχα αποτέλεσματα που προκύπτουν από τη χρήση των σημείων Chebyshev και δείξτε ότι, παρόλο που τα ακρότατα σημεία Chebyshev είναι ελαφρώς χειρότερα, η διαφορά δεν είναι ποτέ μεγάλη. Στην πραγματικότητα, η διάσημη ευστάθεια του Chebyshev και η φασματική ακρίβεια παρατηρούνται και εδώ.

21. Παρεμβάλετε τη συνάρτηση  $f(x) = \ln(x)$  με ένα κυβικό πολυώνυμο που διέρχεται από τα σημεία  $x_i = (0.1, 1, 2, 2.9)$ . Υπολογίστε την τιμή του πολυωνύμου παρεμβολής στο  $x = 1.5$  και συγκρίνετε το αποτέλεσμα με την ακριβή τιμή και με την τιμή του κυβικού εφάπτοντος πολυωνύμου Hermite που διέρχεται από τα σημεία  $x_i = (1, 1, 2, 2)$ , οι οποίες δίνονται στο Παράδειγμα 10.9.

Εξηγήστε τις παρατηρήσεις σας εξετάζοντας τους όρους σφάλματος και για τα δύο κυβικά πολυώνυμα παρεμβολής.

22. Για κάποια συνάρτηση  $f$ , έχετε στη διάθεσή σας έναν πίνακα επεκτεταμένων διαιρεμένων διαφορών της μορφής

$i$	$z_i$	$f[\cdot]$	$f[\cdot, \cdot]$	$f[\cdot, \cdot, \cdot]$	$f[\cdot, \cdot, \cdot, \cdot]$
0	5.0	$f[z_0]$			
1	5.0	$f[z_1]$	$f[z_0, z_1]$		
2	6.0	4.0	5.0	-3.0	
3	4.0	2.0	$f[z_2, z_3]$	$f[z_1, z_2, z_3]$	$f[z_0, z_1, z_2, z_3]$

Συμπληρώστε τον πίνακα.

23. Για τα δεδομένα της Άσκησης 22, ποιο είναι το εφάπτον πολυώνυμο  $p_2(x)$  βαθμού το πολύ 2 το οποίο ικανοποιεί τις ισότητες

$$p_2(5.0) = f(5.0), \quad p'_2(5.0) = f'(5.0), \quad p_2(6.0) = f(6.0)$$

24. (α) Γράψτε κώδικα που να παρεμβάλλει την  $f(x) = \cosh(x) = \frac{e^x + e^{-x}}{2}$  με ένα εφάπτον πολυώνυμο το οποίο συμπίπτει και με την  $f(x)$  και με την  $f'(x)$  στις τετμημένες  $x_0 = 1$  και  $x_1 = 3$ . Σχεδιάστε ένα γράφημα (με λογαριθμικό κατακόρυφο άξονα) στο οποίο να γίνεται σύγκριση της  $f(x)$  με το πολυώνυμο παρεμβολής και ένα άλλο γράφημα στο οποίο να φαίνεται το σφάλμα του πολυωνύμου παρεμβολής σας στο συγκεκριμένο διάστημα.

Χρησιμοποιήστε την εντολή `semilogy` για να δημιουργήσετε τα γραφήματά σας.

- (β) Τροποποιήστε τον κώδικα για να παραγάγετε ένα άλλο πολυώνυμο παρεμβολής που να συμπίπτει με την  $f(x)$  και την  $f'(x)$  στις τετμημένες  $x_0 = 1$ ,  $x_1 = 2$  και  $x_2 = 3$ . Σχεδιάστε δύο γραφήματα, ένα για τη σύγκριση της συνάρτησης με το πολυώνυμο παρεμβολής, και ένα για το σφάλμα. Συγκρίνετε την ποιότητα αυτού του νέου πολυωνύμου με την ποιότητα του πολυωνύμου στο (α).
- (γ) Τροποποιήστε ξανά τον κώδικα για να παραγάγετε ένα πολυώνυμο παρεμβολής που να συμπίπτει με την  $f(x)$ , την  $f'(x)$  και την  $f''(x)$  στις τετμημένες  $x_0 = 1$  και  $x_1 = 3$ . Σχεδιάστε τα δύο γραφήματα και σχολιάστε την ποιότητα αυτής της προσέγγισης συγκριτικά με τα δύο προηγούμενα πολυώνυμα παρεμβολής.
25. Μια δημοφιλής τεχνική που χρησιμοποιείται σε μεθόδους ελαχιστοποίησης συναρτήσεων ως προς πολλές μεταβλητές περιλαμβάνει μια ασθενή ενθύγραμμη αναζήτηση κατά την οποία προσδιορίζεται ένα προσεγγιστικό ελάχιστο  $x^*$  για μια συνάρτηση μίας μεταβλητής,  $f(x)$ , για την οποία δίνονται οι τιμές  $f(0)$ ,  $f'(0)$  και  $f(1)$ . Η  $f(x)$  ορίζεται για κάθε μη αρνητικό  $x$ , έχει συνεχή δεύτερη παράγωγο, και ικανοποιεί τις ανισότητες  $f(0) < f(1)$  και  $f'(0) < 0$ . Στη συνέχεια οι τιμές που δίνονται παρεμβάλλονται από ένα τετραγωνικό πολυώνυμο, και το  $x^*$  ορίζεται ως το ελάχιστο του πολυωνύμου παρεμβολής.
- (α) Βρείτε το  $x^*$  για τις τιμές  $f(0) = 1$ ,  $f'(0) = -1$ ,  $f(1) = 2$ .
- (β) Αποδείξτε ότι το τετραγωνικό πολυώνυμο έχει μοναδικό ελάχιστο το οποίο ικανοποιεί την ανισότητα  $0 < x^* < 1$ . Μπορείτε να αποδείξετε το ίδιο για την  $f$ ;
26. Έχουμε υπολογίσει ένα πολυώνυμο παρεμβολής για τα σημεία δεδομένων  $\{(x_i, y_i)\}_{i=0}^n$  χρησιμοποιώντας καθεμία από τις τρεις πολυωνυμικές βάσεις τις οποίες περιγράψαμε σε αυτό το κεφάλαιο (μονωνυμική, Lagrange και Newton), και έχουμε ήδη κατασκευάσει τυχόν απαραίτητες μήτρες, διανύσματα, συντελεστές παρεμβολής, ή/και συναρτήσεις βάσης. Ξαφνικά συνειδητοποιούμε ότι το τελευταίο σημείο δεδομένων ήταν λάθος. Θεωρήστε τις εξής περιπτώσεις:
- (α) Το τελευταίο σημείο δεδομένων θα έπρεπε να είναι το  $(\tilde{x}_n, y_n)$  (δηλαδή, το  $y_n$  είναι ίδιο όπως και προηγουμένως).

- (β) Το τελευταίο σημείο δεδομένων θα έπρεπε να είναι το  $(x_n, \tilde{y}_n)$  (δηλαδή, το  $x_n$  είναι ίδιο όπως και προηγουμένως).

Για καθένα από αυτά τα δύο σενάρια, προσδιορίστε το υπολογιστικό κόστος της εύρεσης των τροποποιημένων πολυωνύμων παρεμβολής. Δώστε τις απαντήσεις σας χρησιμοποιώντας τον συμβολισμό  $O$  (δείτε τη σελίδα 41).

[Οι απαντήσεις αυτού του ερωτήματος ενδέχεται να απαιτούν κάποιες επιπλέον γνώσεις γραμμικής άλγεβρας.]

## 10.9 Πρόσθετες σημειώσεις

Ο θεμελιώδης ρόλος τον οποίο καλείται να παίξει η πολυωνυμική παρεμβολή σε πολλές διαφορετικές πτυχές των αριθμητικών υπολογισμών είναι και η αιτία για το μεγάλο μέγεθος αυτού του κεφαλαίου, το οποίο κατά τα άλλα πραγματεύεται ουσιαστικά μια σχετικά απλή διαδικασία. Πολλά διδακτικά συγγράμματα για τις αριθμητικές μεθόδους και την αριθμητική ανάλυση αφιερώνουν ένα σημαντικό κομμάτι τους στο θέμα αυτό: Αν θέλετε να πάρετε μια γεύση, αναφέρουμε επιγραμματικά τα βιβλία των Burden και Faires [11], των Cheney και Kincaid [12], και το κλασικό σύγγραμμα των Conte και de Boor [13].

Τα πολυώνυμα Lagrange χρησιμοποιούνται στον σχεδιασμό μεθόδων αριθμητικής ολοκλήρωσης και επίλυσης διαφορικών εξισώσεων. Θα αναφερθούμε εν μέρει σε αυτό το θέμα στα Κεφάλαια 14, 15 και 16. Για περισσότερες λεπτομέρειες μπορείτε να ανατρέξετε σε άλλες πηγές, όπως τα βιβλία των Davis και Rabinowitz [17], των Ascher και Petzold [5], και του Ascher [3].

Για την πολυωνυμική παρεμβολή σε ομοιόμορφα τοποθετημένες (ή ισαπέχουσες) τετμημένες, όπου υπάρχει μια τιμή απόστασης  $h$  τέτοια ώστε να ισχύει  $x_i - x_{i-1} = h$  για κάθε  $i = 1, \dots, n$ , ο πίνακας διαιρεμένων διαφορών έχει πιο απλή μορφή. Διάφορες ονομασίες έχουν χρησιμοποιηθεί στο παρελθόν γι' αυτή την περίπτωση: εξισώσεις Newton της προς τα εμπρός και της προς τα πίσω διαφοράς, εξίσωση κεντρικής διαφοράς του Stirling (δείτε τις Ασκήσεις 11 και 12) κ.ά.

Στον σχεδιασμό μεθόδων για σύνθετα προβλήματα κυριαρχεί επίσης η χρήση διαιρεμένων διαφορών και η παρεμβολή μορφής Newton. Όπως αποσαφηνίζεται και σε κάποιες από τις ασκήσεις, αυτό είναι φυσικό ειδικά αν προσθέτουμε σημεία παρεμβολής ένα τη φορά, π.χ. επειδή θέλουμε να παραμείνουμε στη μία πλευρά μιας ασυνέχειας στην παρεμβαλλόμενη συνάρτηση (αντί να «διασχίσουμε» την ασυνέχεια με ένα απειρως παραγωγίσιμο πολυώνυμο). Κάποιες μέθοδοι αυτού του είδους για την επίλυση προβλημάτων με «κραδασμούς» αποκαλούνται ουσιαστικά μη ταλαντευόμενες μέθοδοι (essentially nonoscillatory, ENO) –δείτε, για παράδειγμα, το [3].

Η πολυωνυμική παρεμβολή Chebyshev αποτελεί την πιο σημαντική εξαίρεση στον πρακτικό κανόνα που ορίζει ότι πρέπει να διατηρούμε μικρό τον βαθμό των πολυωνύμων παρεμβολής και να τα χρησιμοποιούμε τοπικά, και μόνο όταν θέλουμε μεγαλύτερη ακρίβεια να καταφεύγουμε στις μεθόδους του Κεφαλαίου 11. Τουλάχιστον για απείρως λείες συναρτήσεις  $f(x)$ , η μεγάλου βαθμού πολυωνυμική παρεμβολή Chebyshev μπορεί να επιτύχει πολύ μεγάλη ακρίβεια με λογικό κόστος, όπως φαίνεται από το Παράδειγμα 10.7 και την Άσκηση 19. Αυτό μας επιτρέπει να αντιμετωπίζουμε τις λείες συναρτήσεις ως αντικείμενα, αντικαθιστώντας τες εσωτερικά με τα αντίστοιχα πολυώνυμα παρεμβολής Chebyshev κατά τρόπο που δεν είναι ορατός στον χρήστη. Έτσι καθίστανται εφικτές «συμβολικές» πράξεις όπως η παραγώγιση, η ολοκλήρωση, ή ακόμα και η επίλυση διαφορικών εξισώσεων ως προς μία μεταβλητή. Ο Trefethen κ.ά. έχουν ασχοληθεί με αυτό το θέμα στο πλαίσιο του MATLAB· για περισσότερες λεπτομέρειες μπορείτε να μεταβείτε στη διεύθυνση <https://www.mathworks.com/matlabcentral/mlc-downloads/downloads/submissions/23972/versions/22/previews/chebfun/guide/html/guide4.html>.

Πολυώνυμα Chebyshev προκύπτουν συχνά κατά την αριθμητική ανάλυση φαινομενικά άσχετων μεταξύ τους μεθόδων λόγω της ιδιότητας min-max που έχουν. Για παράδειγμα, φιγουράρουν σε περίοπτη θέση στην ανάλυση των ρυθμών σύγκλισης για τη διάσημη μέθοδο συζυγούνς κλίσης που παρουσιάσαμε στην Ενότητα 7.4· δείτε, για παράδειγμα, το βιβλίο της Greenbaum [32] ή του LeVeque [50]. Μπορούν επίσης να εφαρμοστούν στην αποδοτική επίλυση μερικών διαφορικών εξισώσεων με τη χρήση μεθόδων που είναι συλλογικά γνωστές ως φασματικός συνεγγισμός (spectral collocation)· δείτε, μεταξύ άλλων, το βιβλίο του Trefethen [69].