ΑΝΑΠΤΥΞΗ ΑΛΓΟΡΙΘΜΟΥ
ΜΕΙΞΗΣ ΚΥΤΤΑΡΙΚΩΝ ΣΕΙΡΩΝ
ΓΙΑ ΤΗΝ ΠΑΡΑΓΩΓΗ ΟΜΟΙΟΜΟΡΦΟΥ ΔΕΙΓΜΑΤΟΣ

Μπαλούγια Μαγδαληνή

Τομέας Τηλεπικοινωνιών

Ανάπτυξη Αλγορίθμου Μείξης Κυτταρικών Σειρών
για την Παραγωγή Ομοιόμορφου Δείγματος

Μπαλούγια Μαγδαληνή

Επιβλέπων Καθηγητής : Μιχάλης Ζερβάκης

Εξεταστική Επιτροπή
Ζερβάκης Μιχάλης (Καθηγητής)
Καρυστινός Γεώργιος (Επίκουρος Καθηγητής)
Μπάλας Κωνσταντίνος (Αναπληρωτής Καθηγητής)

Πολυτεχνείο Κρήτης, Χανιά
Σεπτέμβριος 2008

ΠΕΡΙΛΗΨΗ (ABSTRACT)

Τα πρότυπα μείγματα αναφοράς έχουν ευρείες εφαρμογές στον τομέα της Μοριακής Βιολογίας, με κυριότερη τη συμμετοχή τους σε υβριδοποιήσεις με μικροσυστοιχίες. Τα μείγματα αναφοράς δημιουργούνται συνήθως από διάφορες κυτταρικές σειρές, οι οποίες περιλαμβάνουν γονιδιακή πληροφορία. Ο πιο κοινός λόγος για τον οποίο πραγματοποιούνται υβριδοποιήσεις μεταξύ πρότυπων μειγμάτων αναφοράς και μικροσυστοιχιών που περιλαμβάνουν γονιδιακή πληροφορία είναι οι συγκρίσεις των επιπέδων έκφρασης για πολλαπλές υβριδοποιήσεις, αλλά και ο προσδιορισμός των τμημάτων των γονιδίων που περιλαμβάνουν χρήσιμη πληροφορία. Όλη η παραπάνω ανάλυση πάνω σε μία μικροσυστοιχία επηρεάζεται από τον αριθμό των κυτταρικών σειρών και τον τρόπο συμμετοχής τους στο μείγμα αναφοράς. Γι' αυτό είναι αναγκαία η ανάπτυξη ενός πεδίου έρευνας, το οποίο στοχεύει στη δημιουργία μειγμάτων αναφοράς που κατά την υβριδοποίησή τους θα δημιουργούν γονίδια με έντονο σήμα, ώστε να είναι εύκολη η οποιαδήποτε περαιτέρω ανάλυση.

Η δημιουργία ενός μείγματος αναφοράς που θα ικανοποιεί όλες αυτές τις απαιτήσεις ανάγεται λοιπόν στην εύρεση των βέλτιστων αναλογιών με τις οποίες πρέπει να αναμείξουμε τις κυτταρικές σειρές που συμμετέχουν στο μείγμα. Η έννοια "βέλτιστες" ανάγει το πρόβλημα της δημιουργίας του μείγματος αναφοράς, ενός δηλαδή προβλήματος που υπάγεται στα πλαίσια της Μοριακής Βιολογίας σε ένα μαθηματικό πρόβλημα βελτιστοποίησης. Βελτιστοποίηση καλείται η διαδικασία εύρεσης των ακρότατων μίας συνάρτησης, η οποία μπορεί να έχει ελάχιστα, μέγιστα ή και τα δύο. Οι τιμές του σημείου για το οποίο η συνάρτηση παρουσιάζει ακρότατο, αντιστοιχούν στις βέλτιστες τιμές της ανεξάρτητης μεταβλητής ως προς την οποία βελτιστοποιούμε.

Σκοπός της εργασίας αυτής είναι η εύρεση των αναλογιών πρόσμειξης δεδομένου αριθμού κυτταρικών σειρών, με τη χρήση μεθόδων βελτιστοποίησης, ώστε να καταλήξουμε σε ένα ομοιόμορφο μείγμα αναφοράς. Τα αποτελέσματα στα οποία καταλήξαμε για τις αναλογίες προήλθαν από την επίλυση ενός γραμμικού και ενός μη γραμμικού προβλήματος βελτιστοποίησης. Πρόκειται για τα προβλήματα των Ελαχίστων Τετραγώνων και της Ελαχιστοποίησης της Εντροπίας αντίστοιχα, ενώ για την επίλυσή τους χρησιμοποιήθηκε η αναλυτική μέθοδος Gradient Descent, αλλά και κάποιες άμεσες μέθοδοι βελτιστοποίησης. Οι αναλογίες πρόσμειξης όλων των περιπτώσεων είναι αρκετά συγκρίσιμες και ικανοποιούν και τους απαραίτητους περιορισμούς. Προέκυψαν διάφορα μείγματα αναφοράς, εξαιτίας των διαφορετικών προβλημάτων που εξετάσαμε, τα οποία εμφανίζουν την απαιτούμενη ομοιομορφία και περιλαμβάνουν γονίδια με τιμές έντασης τέτοιες ώστε να παρέχουν έντονο σήμα (μέγιστη δυνατή κάλυψη) και να διευκολύνουν την ανάλυση στις μικροσυστοιχίες με τις οποίες θα υβριδοποιηθούν. Όλα τα μείγματα μπορούν να είναι αντιπροσωπευτικά και να χρησιμοποιηθούν ως μείγματα αναφοράς, αλλά προτιμότερα είναι αυτά που προέκυψαν από την υλοποίηση του προβλήματος Ελαχιστοποίησης της Εντροπίας.

ΕΥΧΑΡΙΣΤΙΕΣ

Πρώτα απ' όλα θα ήθελα να ευχαριστήσω των επιβλέποντά Καθηγητή μου, κύριο Μιχάλη Ζερβάκη, για την υπέροχη συνεργασία, την πολύτιμη βοήθεια και στήριξή του, καθώς και για την εμπιστοσύνη και την ελευθερία που μου έδωσε όσον αφορά την έρευνά μου. Είμαι επίσης ευγνώμων στον Καθηγητή κύριο Καρυστινό Γεώργιο και στον Καθηγητή κύριο Μπάλα Κωνσταντίνο, οι οποίοι δέχτηκαν να αξιολογήσουν αυτήν την εργασία.

Θα ήθελα επίσης να αναφέρω συγκεκριμένα κάποιους συναδέλφους μου, οι οποίοι με βοήθησαν πάρα πολύ ώστε να ολοκληρώσω επιτυχώς την εργασία αυτή. Ευχαριστώ θερμά τον Μεταπτυχιακό φοιτητή Κωνσταντίνο Παπαδόπουλο για την ανεκτίμητη βοήθειά του όσον αφορά την κατανόηση των μαθηματικών σημείων της εργασίας αυτής και τον Διδακτορικό φοιτητή Γρηγόρη Χρυσό για τις συμβουλές του, όταν τις χρειαζόμουν. Επιπλέον θα ήθελα να ευχαριστήσω και τη Γεωργία Τσιλίκη, μέλος του Εργαστηρίου Μοριακής Βιολογίας και Βιοτεχνολογίας του Πανεπιστημίου Ηρακλείου, για την άμεση ανταπόκρισή της στην αποστολή των αποτελεσμάτων που προέκυψαν από τη δική της έρευνα και τα οποία ήταν απαραίτητα για την επιβεβαίωση της δικής μου εργασίας.

Θα ήθελα να αφιερώσω αυτήν την εργασία στους γονείς μου για την απεριόριστη βοήθειά τους, ώστε να γίνω αυτό που είμαι σήμερα, αλλά και για την υποστήριξή τους στις ιδέες μου και στη δουλειά μου. Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στον Κωνσταντίνο, για την στήριξη και την εμπιστοσύνη που έδειξε σε μένα και στις δυνατότητές μου, αλλά και για την ανεκτικότητά του απέναντι στα άγχη και τους φόβους μου όλο αυτόν τον καιρό.

ΠΕΡΙΕΧΟΜΕΝΑ

Περίληψη
Ευχαριστίες
Περιεχόμενα
Λίστα Σχημάτων
Λίστα Πινάκων

1 Εισαγωγή
1.1 Σκοπός Εργασίας
1.2 Σχετική Έρευνα
1.3 Δομή Εργασίας

2 Θεωρητικό Υπόβαθρο
2.1 Βελτιστοποίηση
2.2 Βελτιστοποίηση Με ή Χωρίς Περιορισμούς
2.2.1 Βελτιστοποίηση Χωρίς Περιορισμούς
2.2.2 Βελτιστοποίηση Με Περιορισμούς
2.2.3 Βασικά Σημεία Ανάλυσης για Βελτιστοποίηση με ή χωρίς περιορισμούς
2.3 Βελτιστοποίηση για Διάφορες Αντικειμενικές Συναρτήσεις
2.3.1 Βελτιστοποίηση Κυρτών Συναρτήσεων (Convex Optimization)
2.3.1.1 Προβλήματα Χωρίς Περιορισμούς
2.3.1.2 Προβλήματα με Περιορισμούς Ισότητας
2.3.1.3 Ελαχιστοποίηση ως προς μη αρνητική μεταβλητή
2.3.2 Τετραγωνική Βελτιστοποίηση (Quadratic Optimization)
2.3.3 Βελτιστοποίηση Μη Κυρτών Συναρτήσεων
2.4 Μέθοδοι Επίλυσης Κυρτών Προβλημάτων Βελτιστοποίησης (Convex)
2.4.1 Εισαγωγικά Στοιχεία
2.4.2 Επίλυση Προβλημάτων Χωρίς Περιορισμούς
2.4.2.1 Μέθοδος Gradient Descent
2.4.3 Επίλυση Προβλημάτων με Περιορισμούς
2.4.3.1 Lagrange Multipliers
2.4.3.1.1 Lagrangian & Περιορισμοί Ισότητας
2.4.3.1.2 Lagrangian & Περιορισμοί Ανισότητας
2.5 Μέθοδος Ελαχίστων Τετραγώνων
2.6 Πρόβλημα Ελαχιστοποίησης Εντροπίας
2.6.1 Εντροπία – Γενικός Ορισμός
2.6.2 Βελτιστοποίηση Εντροπίας

3 Υλοποίηση
3.1 Καθορισμός Δεδομένων
3.2 Επεξεργασία Αρχικών Δεδομένων
3.2.1 Μείωση Διάστασης Δεδομένων (Συσχέτιση)
3.2.2 Κανονικοποίηση Δεδομένων
3.3 Υλοποίηση Αλγορίθμου Gradient Descent
3.3.1 Υλοποίηση Μεθόδου Ελαχίστων Τετραγώνων
3.3.2 Υλοποίηση Μεθόδου Ελαχιστοποίησης Εντροπίας
3.4 Linear & Non Linear Least Squares in Matlab
3.4.1 Lsqlin
3.4.2 Lsqnonlin

4 Αποτελέσματα
4.1 Αποτελέσματα
4.2 Έλεγχος Αποτελεσμάτων για το Σύνολο των Δεδομένων
4.3 Υπολογισμός Κάλυψης Μείγματος Αναφοράς
4.4 Σύγκριση Αποτελεσμάτων Υπολογιστικής Μεθόδου – Μεθόδων Βελτιστοποίησης

5 Συμπεράσματα
5.1 Συνολικά Συμπεράσματα
5.2 Μελλοντική Εργασία

Παράρτημα Α
Βιβλιογραφία

ΛΙΣΤΑ ΣΧΗΜΑΤΩΝ

2.1 Γεωμετρική Ερμηνεία Συνθήκης Βελτιστότητας
2.2 Γεωμετρική Αναπαράσταση Τετραγωνικού Προβλήματος Βελτιστοποίησης
2.3 Σχηματική Αναπαράσταση Αλγορίθμου Gradient Descent (μία μεταβλητή)
2.4 Σχηματική Αναπαράσταση Αλγορίθμου Gradient Descent (πολλές μεταβλητές)
2.5 Σχηματική Αναπαράσταση Επιλογής Βήματος Αλγορίθμου
2.6 Γεωμετρική Ερμηνεία Πολλαπλασιαστών Lagrange
3.1 Διάγραμμα Περιγραφής Διαδικασίας Αρχικοποίησης του Target Distribution
4.1 Γραφικές Παραστάσεις Μειγμάτων για Αρχικοποίηση Μείγματος $1/27$
4.2 Γραφικές Παραστάσεις Μειγμάτων για Αρχικοποίηση Μείγματος $9{,}1194$
4.3 Γραφικές Παραστάσεις Μειγμάτων για Αρχικοποίηση Μείγματος ≥ 350
4.4 Σύγκριση Μειγμάτων Gradient Descent – Lsqlin, για λ = 1,967 και Αρχικοποίηση $1/27$
4.5 Σύγκριση Μειγμάτων Gradient Descent – Lsqlin, για λ = 10 και Αρχικοποίηση $1/27$
4.6 Σύγκριση Μειγμάτων Gradient Descent – Lsqlin, για λ = 1,967 και Αρχικοποίηση $9{,}1194$
4.7 Σύγκριση Μειγμάτων Gradient Descent – Lsqlin, για λ = 10 και Αρχικοποίηση $9{,}1194$
4.8 Σύγκριση Μειγμάτων Gradient Descent – Lsqlin, για λ = 1,967 και Αρχικοποίηση ≥ 350
4.9 Σύγκριση Μειγμάτων Gradient Descent – Lsqlin, για λ = 10 και Αρχικοποίηση ≥ 350
4.10 Γραφικές Παραστάσεις Μειγμάτων για την Περίπτωση της Εντροπίας
4.11 Σύγκριση Μειγμάτων Gradient Descent – Lsqnonlin, για λ = 1,967
4.12 Σύγκριση Μειγμάτων Gradient Descent – Lsqnonlin, για λ = 10
4.13 Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, λ = 1,967 ($A_{init} = 1/27$)
4.14 Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, λ = 10 ($A_{init} = 1/27$)
4.15 Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, Εντολές ($A_{init} = 1/27$)
4.16 Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, λ = 1,967 ($A_{init} = 9{,}1194$)
4.17 Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, λ = 10 ($A_{init} = 9{,}1194$)
4.18 Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, Εντολές ($A_{init} = 9{,}1194$)
4.19 Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, λ = 1,967 ($A_{init} \geq 350$)
4.20 Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, λ = 10 ($A_{init} \geq 350$)
4.21 Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, Εντολές ($A_{init} \geq 350$)
4.22 Ιστόγραμμα Μείγματος για 34772 γονίδια (λ = 1,967 - $A_{init} = 1/27$)
4.23 Ιστόγραμμα Μείγματος για 34772 γονίδια (λ = 10 - $A_{init} = 1/27$)
4.24 Ιστόγραμμα Μείγματος για 34772 γονίδια (Lsqlin - $A_{init} = 1/27$)
4.25 Ιστόγραμμα Μείγματος για 34772 γονίδια (λ = 1,967 - $A_{init} = 9{,}1194$)
4.26 Ιστόγραμμα Μείγματος για 34772 γονίδια (λ = 10 - $A_{init} = 9{,}1194$)
4.27 Ιστόγραμμα Μείγματος για 34772 γονίδια (Lsqlin - $A_{init} = 9{,}1194$)
4.28 Ιστόγραμμα Μείγματος για. 34772 γονίδια (λ = 1,967 - $A_{init} \geq 350$)
4.29 Ιστόγραμμα Μείγματος για 34772 γονίδια (λ = 10 - $A_{init} \geq 350$)

4.30 Ιστόγραμμα Μείγματος για 34772 γονίδια (Lsqlin - $A_{init} \geq 350$)
4.31 Ιστόγραμμα Μείγματος για 34772 γονίδια (λ = 1,967)
4.32 Ιστόγραμμα Μείγματος για 34772 γονίδια (λ = 10)
4.33 Ιστόγραμμα Μείγματος για 34772 γονίδια (Lsqnonlin)
4.34 Ιστόγραμμα Μείγματος για Αναλογίες 1/22

ΛΙΣΤΑ ΠΙΝΑΚΩΝ

3.1 Μέγεθος Συσχέτισης Μεταβλητών
3.2 Αριθμός Επαναλήψεων Εκτέλεσης Αλγορίθμου (Least Squares)
3.3 Αριθμός Επαναλήψεων Εκτέλεσης Αλγορίθμου (Entropy)
4.1 Κανονικοποιημένες Αναλογίες & Μείγματα για την Αρχικοποίηση $1/27$
4.2 Κανονικοποιημένες Αναλογίες & Μείγματα για την Αρχικοποίηση $9{,}1194$
4.3 Κανονικοποιημένες Αναλογίες & Μείγματα για την Αρχικοποίηση ≥ 350
4.4 Μέσες Τιμές & Διασπορές για την 1η Αρχικοποίηση
4.5 Μέσες Τιμές & Διασπορές για την 2η Αρχικοποίηση
4.6 Μέσες Τιμές & Διασπορές για την 3η Αρχικοποίηση
4.7 Κανονικοποιημένες Αναλογίες & Πρότυπα Μείγματα
4.8 Μέσες Τιμές & Διασπορές των 3 Μειγμάτων
4.9 Σύγκριση Μέσων Τιμών & Διασπορών Εντροπίας – Ελαχίστων Τετραγώνων($A_{init} = 1/27$)
4.10 Σύγκριση Μέσων Τιμών & Διασπορών Εντροπίας – Ελαχίστων Τετραγώνων ($A_{init} = 9{,}1194$)
4.11 Σύγκριση Μέσων Τιμών & Διασπορών Εντροπίας – Ελαχίστων Τετραγώνων($A_{init} \geq 350$)
4.12 Σύγκριση Μέσων Τιμών & Διασπορών για τα 2 Set Δεδομένων $(1/27)$
4.13 Σύγκριση Μέσων Τιμών & Διασπορών για τα 2 Set Δεδομένων $(9{,}1194)$
4.14 Σύγκριση Μέσων Τιμών & Διασπορών για τα 2 Set Δεδομένων (≥ 350)
4.15 Σύγκριση Μέσων Τιμών & Διασπορών για τα 2 Set Δεδομένων
4.16 Ποσοστά Κάλυψης για το Πρόβλημα Ελαχίστων Τετραγώνων ($A_{initial} = 1/27$)
4.17 Ποσοστά Κάλυψης για το Πρόβλημα Ελαχίστων Τετραγώνων ($A_{initial} = 9.1194$)
4.18 Ποσοστά Κάλυψης για το Πρόβλημα Ελαχίστων Τετραγώνων ($A_{initial} \geq 350$)
4.19 Ποσοστά Κάλυψης για το Πρόβλημα Ελαχιστοποίησης της Εντροπίας
4.20 Σύγκριση Τιμών Κάλυψης Αναλογίες 1-1 v.s. $a_i^{optimal}$
4.21 Βαρύτητα Αναλογιών Πρόσμειξης Κυτταρικών Σειρών
4.22 Σύγκριση Κυτταρικών Σειρών – Υπολογιστική Μέθοδος v.s. Μέθοδος Βελτιστοποίησης
A.1 Συντομογραφίες & Πηγές Καρκινωμάτων Κυτταρικών Σειρών
A.2 Προσδιορισμός 27 γονιδίων που προέκυψαν με τη χρήση συσχέτισης
A.3 Βαρύτητα Κυτταρικών Σειρών στο Βέλτιστο Μείγμα

ΚΕΦΑΛΑΙΟ 1 - ΕΙΣΑΓΩΓΗ

Το αρχικό αυτό κεφάλαιο παρουσιάζει το κίνητρο – σκοπό για τον οποίο εκπονήσαμε την εργασία αυτή και συνοψίζει όλη τη σχετική επιστημονική έρευνα που έχει γίνει, τόσο σε ζητήματα βελτιστοποίησης, όσο και για το πρόβλημα δημιουργίας πρότυπου μείγματος αναφοράς, προερχόμενου από τη μείξη κυτταρικών σειρών γονιδίων. Τέλος παρουσιάζει τη δομή της εργασίας αυτής.

1.1 Σκοπός Εργασίας

Οι συγκρίσεις των επιπέδων έκφρασης σε πολλαπλά πειράματα υβριδοποιήσεων μικροσυστοιχιών DNA (microarray) και η αναγνώριση πληροφοριακών προτύπων των εκφράσεων των γονιδίων τους είναι πιο εύκολες όταν ένα *κοινό πρότυπο μείγμα αναφοράς* υβριδοποιείται σε κάθε μικροσυστοιχία. Μια μικροσυστοιχία (microarray) ή κοινώς γνωστή ως γονιδιακό chip είναι μία συλλογή από μικροσκοπικές κηλίδες (spots) DNA, οι οποίες αντιπροσωπεύουν συνήθως μοναδικά γονίδια διατεταγμένα σε μία στερεή επιφάνεια, μέσω ομοιοπολικών δεσμών με ένα χημικό πίνακα. Οι μικροσυστοιχίες DNA χρησιμοποιούνται συνήθως για καταγραφή εκφράσεων γονιδίων (expression profiling) ή για συγκριτική υβριδοποίηση γονιδιώματος. Έτσι, η ισχύς της ανάλυσης μίας μικροσυστοιχίας μπορεί να αυξηθεί εξαρτώμενη από την κάλυψη των γονιδίων που είναι διάστικτα στην συστοιχία. Ως κάλυψη συστοιχίας ορίζουμε το ποσοστό των κηλίδων, οι οποίες έχουν υβριδοποιηθεί με το μείγμα αναφοράς και παράγουν σήμα, πάνω από ένα συγκεκριμένο κατώφλι που έχει ορίσει ο χρήστης.

Για να παρέχουμε λοιπόν τη μέγιστη κάλυψη των γονιδίων που είναι διάστικτα στη συστοιχία και έτσι να έχουμε την ευκολότερη δυνατή ανάλυση των δεδομένων που προκύπτουν από τη μικροσυστοιχία, χρησιμοποιούμε πρότυπα μείγματα αναφοράς, τα οποία κατασκευάζονται από διάφορες κυτταρικές σειρές. Και αυτό γιατί χρησιμοποιώντας ως μείγμα αναφοράς ένα μείγμα που προήλθε από μία μόνο κυτταρική σειρά, το δείγμα ελέγχου δεν παρέχει σήμα σε όλες τις κηλίδες. Για να ξεπεράσουμε λοιπόν αυτό το πρόβλημα, χρησιμοποιούμε τελικά μείγματα αναφοράς που προέρχονται από διάφορες κυτταρικές σειρές και τα οποία μας οδηγούν στη μέγιστη κάλυψη, όπως προαναφέρθηκε. Η κάλυψη όμως δεν εξαρτάται μόνο από τον αριθμό των κυτταρικών σειρών που συμμετέχουν στο μείγμα αναφοράς, αλλά και από τη δομή του μείγματος, δηλαδή τον τρόπο κατά τον οποίο οι σειρές συμμετέχουν σ' αυτό.

Στόχος της εργασίας αυτής είναι *να δημιουργήσει ένα πρότυπο μείγμα αναφοράς, το οποίο θα μπορεί να χρησιμοποιηθεί για σύγκριση των επιπέδων έκφρασης γονιδίων σε πολλαπλές υβριδοποιήσεις μικροσυστοιχιών.* Το πρότυπο μείγμα αυτό προέρχεται από διάφορες κυτταρικές σειρές, οι οποίες αναμειγνύονται με τέτοιο τρόπο, ώστε κατά την υβριδοποίησή του στη μικροσυστοιχία να *επιτυγχάνεται η μέγιστη κάλυψη*. Η δική μας προσέγγιση στο πρόβλημα *υπολογίζει το ποσοστό της κάλυψης* το οποίο επιτυγχάνεται, ως κριτήριο για την καταλληλότητα του μείγματος, *μέσω γραφικών μεθόδων* και επιπλέον έχει ως απαίτηση *τη ύπαρξη ομοιομορφίας του μείγματος. μελέτη της στις επιθυμητές μαθηματικές ιδιότητες που θέλουμε να έχει το μείγμα*. Συγκεκριμένα, αρχικά ελέγχουμε την καταλληλότητα των μειγμάτων αναφοράς στηριζόμενοι στο αν ικανοποιεί τις βασικές μαθηματικές ιδιότητες της ομοιομορφίας. Σύμφωνα με τις ιδιότητες αυτές, ένα μείγμα είναι ομοιόμορφο όταν όλα τα στοιχεία του έχουν τις ίδιες τιμές, πράγμα το οποίο συνεπάγεται μηδενική διασπορά μείγματος (θα υπάρξει περαιτέρω ανάλυση

στο επόμενο κεφάλαιο 3). Έπειτα, ελέγχουμε ποιος είναι ο αριθμός των γονιδίων που βρίσκονται πάνω από ένα προκαθορισμένο κατώφλι, μέσω γραφικών μεθόδων. Το σύνολο των γονιδίων που βρίσκονται πάνω απ' το κατώφλι αυτό αντιστοιχεί στο ποσοστό κάλυψης που παρέχει το μείγμα.

Οι κυτταρικές σειρές που χρησιμοποιήθηκαν για τη δημιουργία του μείγματος είναι 22 σε αριθμό, αποτελούνται η καθεμία από 34772 γονίδια και περιγράφουν ανθρώπινα καρκινώματα[(1)]. Το ομοιόμορφο πρότυπο μείγμα αναφοράς που δημιουργήθηκε, στηρίχθηκε στην ανάμειξη των κυτταρικών σειρών σε συγκεκριμένες αναλογίες. Οι αναλογίες πρόσμειξης δεν είναι τίποτε άλλο από το ποσοστό συμμετοχής που έχει κάθε κυτταρική σειρά στο μείγμα, δηλαδή το ποσοστό εμφάνισης κάθε περίπτωσης ανθρώπινου καρκινώματος στο μείγμα. Για να να προσδιοριστεί επομένως το ζητούμενο μείγμα αναφοράς, *αρκούσε να βρούμε τις κατάλληλες αναλογίες πρόσμειξης*. Αναπαριστώντας με μαθηματικό τρόπο τη δημιουργία του επιθυμητού πρότυπου μείγματος αναφοράς καταλήγουμε στην παρακάτω έκφραση

$$data \times vital\ \ statistics\ \ = reference\ \ _ mixture\ \ .$$

Τα δεδομένα μας αποτελούνται από τις 22 κυτταρικές σειρές, των 34772 γονιδίων η καθεμία (34772×22) και οι αναλογίες πρόσμειξης είναι 22 σε αριθμό (22×1). Το αποτέλεσμα του αλγεβρικού πολλαπλασιασμού τους αντιπροσωπεύει το μείγμα αναφοράς, το οποίο αποτελείται από 34772 γονίδια (34772×1). Βασικοί περιορισμοί στους οποίους πρέπει να υπόκεινται οι αναλογίες πρόσμειξης των κυτταρικών σειρών και οι οποίοι προκύπτουν από την καθεαυτή έννοια της αναλογίας ή ποσοστού, είναι το να περιγράφονται από πραγματικούς, θετικούς αριθμούς, οι οποίοι είναι μικρότεροι ή το πολύ ίσοι με τη μονάδα και το άθροισμά τους (των αναλογιών) να είναι ίσο με 1.
Το μείγμα το οποίο δημιουργήθηκε ικανοποιεί όλες τις απαιτήσεις της ομοιομορφίας και αποτελείται από γονίδια με αρκετά ικανοποιητικές τιμές έντασης, έτσι ώστε κατά την υβριδοποίηση με τη μικροσυστοιχία να έχουμε τη μέγιστη κάλυψη, που είναι και ο βασικός στόχος της εργασίας αυτής.

1.2 Σχετική Έρευνα

Τα πρότυπα μείγματα αναφοράς έχουν ευρείες εφαρμογές στον τομέα της Μοριακής Βιολογίας, κυρίως λόγω της συμμετοχής τους σε υβριδοποιήσεις μικροσυστοιχιών. Γι' αυτό και η δημιουργία τους αποτελεί ένα σημαντικό σημείο έρευνας τόσο για το πεδίο της Μοριακής Βιολογίας, όσο και για άλλα επιστημονικά και ερευνητικά πεδία. Πέρα από την έρευνα που εκπονήθηκε στην εργασία αυτή και η οποία προσεγγίζει το πρόβλημα με καθαρά μαθηματική σκοπιά υπάρχουν και αντίστοιχες έρευνες, οι οποίες χρησιμοποιούν διαφορετικές προσεγγίσεις για τη δημιουργία των μειγμάτων αναφοράς. Οι έρευνες αυτές δημιουργούν μείγματα αναφοράς, των οποίων η καταλληλότητα πιστοποιείται υπολογίζοντας το ποσοστό κάλυψης της μικροσυστοιχίας.
Ένα διαθέσιμο (εμπορικό) πρότυπο μείγμα αναφοράς, το οποίο χρησιμοποιείται συχνά σε πειράματα μικροσυστοιχιών δύο χρωμάτων είναι το διεθνές μείγμα ανθρώπινου RNA της *Stratagene* (www.stratagene.com). Η δημιουργία του μείγματος αυτού στηρίχθηκε σε δύο βασικές ιδέες: στην ανάγκη ύπαρξης πολλών κυτταρικών σειρών και στην ανάγκη χρήσης του μείγματος αυτού σε πολλαπλές υβριδοποιήσεις μικροσυστοιχιών. Πρόκειται για ένα μείγμα, το οποίο περιλαμβάνει υψηλής ποιότητας RNA, αποτελείται από 10 ως 14 κυτταρικές σειρές και

κατασκευάστηκε σε βιομηχανικό επίπεδο. Εκτός, όμως από αυτό το διεθνές και ευρέως χρησιμοποιούμενο μείγμα αναφοράς, έγιναν προσπάθειες για να δημιουργηθούν και άλλα πρότυπα μείγματα.

Το 2002 ο Yang πρότεινε να χρησιμοποιηθεί ένας περιορισμένος αριθμός κυτταρικών σειρών, κάθε μία από τις οποίες εκφράζει ένα μεγάλο αριθμό διαφόρων γονιδίων. Κατασκευάστηκαν λοιπόν δύο διαφορετικά πρότυπα μείγματα από αυτές τις κυτταρικές σειρές, με τη μέγιστη παρουσία των μοναδικών γονιδίων, τα οποία είναι επίσης εύκολο να μεγαλώσουν και αποφέρουν μεγάλες ποσότητες RNA. Στο πρώτο πρότυπο μείγμα αναφοράς αναμείχτηκαν ίσες ποσότητες κυτταρικών σειρών CaCO2 που περιγράφουν καρκινώματα του παχέως εντέρου, KM12L4A και της κυτταρικής σειράς OVCAR3 που περιγράφει καρκίνωμα των ωοθηκών. Στο δεύτερο μείγμα αναφοράς αντικαταστάθηκε η κυτταρική σειρά OVCAR3 με τη σειρά U118MG που περιγράφει καρκίνωμα του εγκεφάλου. Το πρώτο μείγμα αναφοράς έδωσε κάλυψη της τάξεως του 75%, ενώ το δεύτερο μείγμα είχε κάλυψη ίση με 80%.

Επίσης υπάρχουν μείγματα αναφοράς RNA ανθρώπων, ποντικιών και αρουραίων, που αναφέρονται ως Universal Reference RNA και τα οποία προέρχονται από ξεχωριστές κυτταρικές σειρές, που αντιπροσωπεύουν διαφορετικούς ιστούς. Συγκεκριμένα, κάθε ένα μείγμα από τα 3 αποτελείται από 10 ανθρώπινες κυτταρικές σειρές, 11 σειρές ποντικιού και 14 σειρές αρουραίου αντίστοιχα. Η κάλυψη της μικροσυστοιχίας αξιολογείται με βάση ένα προκαθορισμένο κατώφλι (κατώτατο όριο), το οποίο ισούται με το με την ένταση του φόντου ή το διπλάσιο της έντασης κάθε καναλιού. Για κατώφλι ίσο με την ένταση του φόντου επιτυγχάνεται κάλυψη μεγαλύτερη από 80%, ενώ για κατώφλι ίσο με το διπλάσιο της έντασης του φόντου επιτυγχάνεται κάλυψη μεγαλύτερη από 60%. Τα μείγματα αναφοράς αυτά, όπως και όλα τα προηγούμενα που αναφέρθηκαν επιβεβαιώνουν τη βασική ιδέα ότι, αναμειγνύοντας RNA που προέρχεται από περιορισμένες, αλλά διαφορετικές κυτταρικές σειρές μπορούμε να οδηγηθούμε σε βέλτιστα πρότυπα μείγματα αναφοράς.

Μία ακόμα προσπάθεια που έγινε προς αυτήν την κατεύθυνση είναι η εφαρμογή μίας υπολογιστικής μεθόδου για τη δημιουργία πρότυπου μείγματος αναφοράς. Στη μέθοδο αυτή χρησιμοποιούνται 22 κυτταρικές σειρές[(1)], οι οποίες περιγράφουν ανθρώπινα καρκινώματα. Ο αριθμός των γονιδίων που συμμετέχουν στις σειρές αυτές είναι 34772, ενώ κάθε γονίδιο περιγράφεται από μία τιμή έντασης για κάθε κυτταρική σειρά. Οι τιμές έντασης κάποιων γονιδίων δεν υπάρχουν. Για το λόγο αυτό χρησιμοποιείται η μέθοδος των k – Κοντινότερων Γειτόνων, με κριτήριο αντικατάστασης τη συσχέτιση μεταξύ του δείγματος και των 10 κοντινότερων γειτόνων του και με στόχο την αντικατάσταση των τιμών που λείπουν.

Έχοντας λοιπόν ως βάση τα δεδομένα αυτά υπολογίστηκε η κάλυψη κάθε μίας από τις 22 κυτταρικές σειρές και τελικά του μείγματος που δημιουργείται από την ανάμειξη των σειρών αυτών, σε αναλογίες 1 προς 1. Χρησιμοποιώντας ως κατώφλι τη μέση τιμή των εντάσεων για όλες τις κυτταρικές σειρές, η οποία είναι ίση με $C_0 = 350$ υπολογίστηκε ο αριθμός των γονιδίων που έχουν τιμές πάνω από το κατώφλι. Για την εύρεση της μέγιστης κάλυψης αναζητήθηκε ο αριθμός κυτταρικών σειρών που θα είναι μικρότερος από 22, η ανάμειξη των οποίων θα οδηγήσει στο βέλτιστο μείγμα. Χρησιμοποιήθηκαν δύο τρόποι για να βρεθεί ο βέλτιστος αριθμός κυτταρικών σειρών που θα αποτελέσουν το μείγμα αναφοράς και οι αναλογίες με τις οποίες θα συμμετέχουν στο μείγμα αυτό. Σύμφωνα με τον πρώτο τρόπο, δοκιμάστηκαν όλοι οι πιθανοί συνδυασμοί κυτταρικών σειρών χρησιμοποιώντας τη μέθοδο της *εξαντλητικής έρευνας (exhaustive search)*. Για την ανάμειξη των κυτταρικών σειρών σε κάθε πιθανό συνδυασμό χρησιμοποιήθηκαν αναλογίες πρόσμειξης, όπου οι περισσότερες είναι μικροί αριθμοί και μόνο

κάποιες αντιστοιχούν σε σχετικά μεγάλα νούμερα, πράγμα το οποίο σημαίνει ότι κάποιες συνεισφέρουν λιγότερο και κάποιες περισσότερο.
Για κάθε πιθανό συνδυασμό υπολογίζεται το ποσοστό κάλυψης. Η κάλυψη αντιστοιχεί και πάλι στο ποσοστό των εντάσεων των γονιδίων που βρίσκονται πάνω από το προηγούμενο κατώφλι. Το μείγμα λοιπόν που προέρχεται από έναν συγκεκριμένο συνδυασμό κυτταρικών σειρών και που έχει τη μέγιστη κάλυψη (η τιμή της πρέπει να βρίσκεται πάνω από το κατώφλι) είναι το επιθυμητό πρότυπο μείγμα αναφοράς. Για τον πρώτο λοιπόν τρόπο έχουμε το βέλτιστο μείγμα αναφοράς, το οποίο αποτελείται από 22 κυτταρικές σειρές, αλλά μόνο οι 11 από αυτές έχουν βαρύτητα και για το οποίο η μέγιστη κάλυψη που επιτυγχάνεται είναι 57,12%.
Ο δεύτερος τρόπος εύρεσης των κυτταρικών σειρών και των αναλογιών τους περιλαμβάνει καθαρά *υπολογιστικές μεθόδους (simulated annealing)*. Ο δεύτερος τρόπος παρέχει την επιθυμητή λύση πιο εύκολα και πιο γρήγορα και χρησιμοποιείται για να πιστοποιήσει απλά τα αποτελέσματα του πρώτου. Και με τον τρόπο αυτό καταλήγουμε στο ότι το βέλτιστο μείγμα περιλαμβάνει 11 σημαντικές κυτταρικές σειρές και αντιστοιχεί σε ποσοστό κάλυψης 57,12%.
Η τελευταία αυτή έρευνα θα χρησιμοποιηθεί και στη δική μας εργασία, με σκοπό τη σύγκριση των αντίστοιχων αποτελεσμάτων και την επιβεβαίωση της ορθότητας της δικής μας υλοποίησης.

(1) : Βλέπε Παράρτημα Α.1

1.3 Δομή Εργασίας

Το υπόλοιπο κομμάτι της εργασίας αυτής δομείται ως εξής:

- Κεφάλαιο 2 – παρουσιάζει όλο το θεωρητικό υπόβαθρο που χρειάστηκε για να επιλυθεί το πρόβλημα εύρεσης των βέλτιστων αναλογιών πρόσμειξης των κυτταρικών σειρών ως ένα πρόβλημα βελτιστοποίησης. Αναλύθηκαν δηλαδή οι έννοιες της βελτιστοποίησης με ή χωρίς περιορισμούς, οι μέθοδοι επίλυσης προβλημάτων βελτιστοποίησης, καθώς και δύο από τα κυριότερα προβλήματα βελτιστοποίησης.
- Κεφάλαιο 3 – παρουσιάζει αναλυτικά τα βήματα υλοποίησης, για τις δύο μεθόδους που χρησιμοποιήθηκαν (Μέθοδος Ελαχίστων Τετραγώνων & Μέθοδος Ελαχιστοποίησης Εντροπίας).
- Κεφάλαιο 4 – παρουσιάζει τα αποτελέσματα των δύο μεθόδων για το πρόβλημα εύρεσης των βέλτιστων αναλογιών πρόσμειξης.
- Κεφάλαιο 5 – παρουσιάζει τα συνολικά συμπεράσματα που προέκυψαν από την εργασία αυτή, καθώς και τις μελλοντικές επεκτάσεις ή αλλαγές που προτείνονται.

ΚΕΦΑΛΑΙΟ 2 - ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 Βελτιστοποίηση (Optimization)

Στη γλώσσα των μαθηματικών, ο όρος βελτιστοποίηση η μαθηματικός προγραμματισμός αναφέρεται στη μελέτη προβλημάτων στα οποία κάποιος στοχεύει στο να ελαχιστοποιήσει ή να μεγιστοποιήσει μία πραγματική συνάρτηση, επιλέγοντας συστηματικά τις τιμές πραγματικών ή ακεραίων μεταβλητών, μέσα από ένα επιτρεπόμενο set τιμών.

Ένα πρόβλημα βελτιστοποίησης μπορεί να αναπαρασταθεί με τον ακόλουθο τρόπο:

Δοθέντος: μίας συνάρτησης $f : A \to \Re$, από κάποιο set A στους πραγματικούς αριθμούς

Να βρεθεί: ένα στοιχείο x_0 στο A , τέτοιο ώστε $f(x_0) \leq f(x)$, για όλα τα x στο A (ελαχιστοποίηση), ή τέτοιο ώστε $f(x_0) \geq f(x)$, για όλα τα x στο A (μεγιστοποίηση).

Μία τέτοια διατύπωση ονομάζεται πρόβλημα βελτιστοποίησης ή πρόβλημα μαθηματικού προγραμματισμού. Συνήθως, το A είναι ένα υποσύνολο του Ευκλείδειου διαστήματος $\Re^n$, το οποίο συχνά καθορίζεται από ένα σύνολο από περιορισμούς, εξισώσεις ή ανισώσεις, τις οποίες τα στοιχεία του A πρέπει να ικανοποιούν. Ο τομέας A της f ονομάζεται διάστημα αναζήτησης, ενώ τα στοιχεία του A ονομάζονται υποψήφιες λύσεις ή εφικτές λύσεις. Η συνάρτηση f ονομάζεται αντικειμενική συνάρτηση ή συνάρτηση κόστους. Η μεταβλητή x ονομάζεται *μεταβλητή βελτιστοποίησης*. Μία εφικτή λύση (feasible), η οποία ελάχιστοποιεί ή μεγιστοποιεί την αντικειμενική συνάρτηση ονομάζεται *βέλτιστη λύση (optimal)*.

2.2 Βελτιστοποίηση με ή χωρίς Περιορισμούς

Μία χαρακτηριστική οικογένεια κλάσεων για προβλήματα βελτιστοποίησης είναι αυτή που χαρακτηρίζεται από την ύπαρξη ή όχι συναρτήσεων περιορισμών. Οι συναρτήσεις περιορισμών (εάν υπάρχουν) μπορεί να είναι συναρτήσεις ισότητας ή ανισότητας ή και τα δύο μαζί.

2.2.1 Unconstraint Optimization (Βελτιστοποίηση χωρίς περιορισμούς)

Το πιο απλό πρόβλημα βελτιστοποίησης είναι να βρεθεί το ελάχιστο (ή το μέγιστο) μίας scalar – valued συνάρτησης ως προς μία scalar – valued μεταβλητή – της λεγόμενης *αντικειμενικής συνάρτησης* – και το που τοποθετείται αυτό το ελάχιστο (μέγιστο). Υποθέτοντας ότι η συνάρτηση είναι διαφορίσιμη, οι γνωστές συνθήκες που απαιτούνται για να να βρεθεί αυτό το ελάχιστο (μέγιστο) – τοπικό και ολικό – είναι

$$\frac{\partial}{\partial_x} f(x) = 0$$

$$\frac{\partial^2}{\partial_x{}^2} f(x) \succ 0 \text{ , για εύρεση ελαχίστου}$$

$$\frac{\partial^2}{\partial_x{}^2} f(x) \prec 0 \text{ , για εύρεση μεγίστου}$$

Όλες οι τιμές της ανεξάρτητης μεταβλητής x που ικανοποιούν αυτές τις συνθήκες, είναι περιοχές τοπικού ελαχίστου (ή μεγίστου). Χωρίς τη δεύτερη συνθήκη, λύσεις τις πρώτης συνθήκης μπορεί να αποτελούν είτε ελάχιστα, είτε μέγιστα σημεία, είτε σημεία καμπής. Οι λύσεις της πρώτης εξίσωσης ονομάζονται *στατικά σημεία* της αντικειμενικής συνάρτησης. Για να βρεθεί ένα *ολικό ελάχιστο (ή μέγιστο)*, κάθε υποψήφια τιμή πρέπει να εξεταστεί: η αντικειμενική συνάρτηση πρέπει να υπολογιστεί για κάθε στατικό σημείο, ώστε να επιλεχθεί αυτό που οδηγεί στη μικρότερη (ή στη μεγαλύτερη) τιμή της. Εάν, παρ' όλα αυτά, η αντικειμενική συνάρτηση μπορεί να αποδειχτεί ότι είναι αυστηρώς κυρτή (convex), η μοναδική λύση της σχέσης $\frac{\partial}{\partial x} f(x) = 0$ υπάρχει και η λύση αυτή ανταποκρίνεται σε ολικό ελάχιστο (μέγιστο). Η συνάρτηση $f(x)$ είναι *αυστηρώς κυρτή*, εάν για οποιαδήποτε επιλογή των x_1, x_2 και της βαθμωτής μεταβλητής a, $f(ax_1 + (1-a)x_2) \prec af(x_1) + (1-a)f(x_2)$. Κυρτές αντικειμενικές συναρτήσεις εμφανίζονται συχνά στην πράξη και ελαχιστοποιούνται (ή μεγιστοποιούνται) ευκολότερα, εξαιτίας της παραπάνω ιδιότητας.

Όταν η αντικειμενική συνάρτηση εξαρτάται από μία vector – valued ποσότητα x (δηλαδή η είσοδος του προβλήματος είναι ένα διάνυσμα (vector) τιμών), ο υπολογισμός των στατικών στοιχείων της συνάρτησης είναι μία απλή επέκταση της παραπάνω απλής περίπτωσης. Παρ' όλα αυτά όμως, το να εξεταστούν όλα τα στατικά σημεία ως πιθανές τοποθεσίες ελαχίστου (ή μεγίστου) είναι λίγο πιο περίπλοκο σ' αυτήν την περίπτωση. Η παράγωγος – κλίση της συνάρτησης $f(x)$, ως προς το διάνυσμα x (διαστάσεων N) ισοδυναμεί με ένα N - διάστατο διάνυσμα, του οποίου τα στοιχεία είναι οι μερικές παράγωγοι της $f()$ ως προς κάθε στοιχείο του x. Δηλαδή,

$$\nabla_x(f(x)) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(x) \\ \text{.....} \quad \text{....} \\ \frac{\partial}{\partial x_N} f(x) \end{bmatrix}$$

Η δεύτερη παράγωγος της συνάρτησης $f(x)$, η οποία περιγράφεται ως $\nabla_x^2(f(x))$ είναι ένας πίνακας όπου η j^{th} στήλη είναι η παράγωγος του j^{th} στοιχείου της παραγώγου της f.

$$\nabla_x^2(f(x))_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(x)$$

Αυτή η ποσότητα είναι γνωστή και ως *Hessian*, η οποία ορίζεται ως ο πίνακας όλων των δεύτερων μερικών παραγώγων της $f()$. Η Hessian είναι πάντα ένας συμμετρικός πίνακας.

Οι συνθήκες λοιπόν που απαιτούνται για να βρεθεί το ελάχιστο (ή μέγιστο) της αντικειμενικής μας συνάρτησης είναι

$$\nabla_x(f(x)) = 0$$

$$\nabla_x^2(f(x)) \succ 0 \text{, για εύρεση ελαχίστου}$$

$$\nabla_x^2(f(x)) \prec 0 \text{, για εύρεση μεγίστου}$$

Για να είναι ένα στατικό σημείο ελάχιστο, θα πρέπει η Hessian η οποία υπολογίζεται για το σημείο αυτό να είναι ένας θετικά ορισμένος πίνακας. Όταν η αντικειμενική συνάρτηση είναι

αυστηρώς κυρτή, ο παραπάνω έλεγχος δε χρειάζεται να γίνει. Για παράδειγμα, η αντικειμενική συνάρτηση $f(x) = x^T Ax$ είναι κυρτή (convex), όποτε το A είναι ένας συμμετρικός και θετικά ορισμένος πίνακας.
Για την εύρεση του βέλτιστου, η παράγωγος πρέπει να "δείχνει" στην κατεύθυνση του μέγιστου ρυθμού αύξησης της συνάρτησης $f()$. Το στοιχείο αυτό χρησιμοποιείται συχνά σε αριθμητικούς αλγορίθμους βελτιστοποίησης.

2.2.2 Constraint Optimization (Βελτιστοποίηση με περιορισμούς)

Προκείται για την ελαχιστοποίηση (ή μεγιστοποίηση) της αντικειμενικής συνάρτησης, η οποία υπόκειται σε περιορισμούς για τις πιθανές τιμές της ανεξάρτητης μεταβλητής, ως προς την οποία ελαχιστοποιούμε (ή μεγιστοποιούμε). Οι περιορισμοί μπορεί να είναι είτε *περιορισμοί ισότητας (equality constraints), είτε περιορισμοί ανισότητας (inequality constraints)*, είτε και τα δύο μαζί. Επειδή η περίπτωση της scalar μεταβλητής είναι πιο απλή απ' αυτή της μεταβλητής με μορφή vector, η παρακάτω ανάλυση θα αφορά την πρώτη περίπτωση. Φυσικά, η δομή και η ανάλυση της δεύτερης περίπτωσης είναι παρόμοια, άρα η scalar – variable περίπτωση είναι αντιπροσωπευτική.

1. **Περιορισμοί Ισότητας**
   Ένα τυπικό πρόβλημα βελτιστοποίησης με περιορισμούς ισότητας έχει τη μορφή
   $$\min_x \{ f(x) \}$$
   $$subject \quad to \; g_i(x) = 0, \quad i = 1,..., m$$
   όπου $f(.)$ είναι μία scalar – valued αντικειμενική συνάρτηση και $g_i(.)$ είναι vector – valued συναρτήσεις περιορισμών ισότητας. Αυστηρή κυρτότητα της αντικειμενικής συνάρτησης δε συνιστά επαρκές στοιχείο για μοναδικότητα του ελαχίστου (ή μεγίστου). Επιπρόσθετα, κάθε στοιχείο του περιορισμού πρέπει να είναι αυστηρώς κυρτό, ώστε να εγγυάται ότι το πρόβλημα έχει μία μοναδική λύση. Εξαιτίας του περιορισμού, τα στατικά σημεία της $f(.)$ μόνο, μπορεί να μην είναι λύσεις για το πρόβλημα βελτιστοποίησης με περιορισμούς, μπορεί να μην ικανοποιούν τους περιορισμούς. Μάλιστα, οι λύσεις του προβλήματος με περιορισμούς είναι συχνά μη στατικά σημεία της αντικειμενικής συνάρτησης. Επιπλέον, η ad hoc τεχνική για εύρεση όλων των στατικών σημείων της αντικειμενικής συνάρτησης που ικανοποιούν τους περιορισμούς, δε λειτουργεί.

2. **Περιορισμοί Ανισότητας**
   Το τυπικό πρόβλημα βελτιστοποίησης με περιορισμούς ανισότητας έχει τη μορφή
   $$\min_x \{ f(x) \}$$
   $$subject \quad to \; h_i(x) \leq 0, \quad i = 1,..., p$$
   όπου $f(.)$ είναι μία scalar – valued αντικειμενική συνάρτηση και $h_i(.)$ είναι συναρτήσεις περιορισμών ανισότητας.

3. **Περιορισμοί Ισότητας & Ανισότητας**
   Ένα πρόβλημα βελτιστοποίησης, το οποίο περιλαμβάνει συναρτήσεις περιορισμού ισότητας και ανισότητας ταυτόχρονα έχει τη μορφή

$$\min imize_{x} \{ f(x) \}$$
$$subject \quad to \quad g_i(x) = 0, \quad i = 1,..., m$$
$$h_i(x) \le 0, \quad i = 1,..., p$$

Όπως και παραπάνω $f(.)$ είναι μία scalar – valued αντικειμενική συνάρτηση, $g_i(.)$ είναι vector – valued συναρτήσεις περιορισμών ισότητας και $h_i(.)$ είναι συναρτήσεις περιορισμών ανισότητας. Το σύνολο των σημείων για τα οποία ορίζονται η αντικειμενική συνάρτηση και όλες οι συναρτήσεις περιορισμών, ονομάζεται πεδίο ορισμού του προβλήματος βελτιστοποίησης και ορίζεται ως

$$D = \bigcap_{i=1}^{m} dom \ g_i \cap \bigcap_{i=1}^{p} dom \ h_i$$

Ένα σημείο $x \in D$ είναι υποψήφια – εφικτή λύση, εάν ικανοποιεί τους παραπάνω περιορισμούς. Η βέλτιστη τιμή $p*$ του προβλήματός μας ορίζεται ως

$$p* = \inf\{ f(x) \mid g_i(x) = 0, \quad i = 1,..., m, \quad h_i(x) \le 0, \quad i = 1,..., p\}$$

Το x επιτρέπεται να πάρει τις ακραίες τιμές $\pm\infty$. Εάν το πρόβλημα είναι μη εφικτό, τότε ισχύει $p* = \infty$ (ακολουθώντας την standard σύμβαση ότι το infnum ενός κενού συνόλου είναι ∞). Εάν υπάρχουν εφικτές λύσεις x_k με $f(x_k) \to -\infty$ όσο $k \to \infty$, τότε $p* = -\infty$ και λέμε ότι το πρόβλημα βελτιστοποίησης είναι απεριόριστο προς τα κάτω (unbounded below).

2.2.3 Βασικά σημεία ανάλυσης για Βελτιστοποίηση με ή χωρίς Περιορισμούς

Εκτός από τις κατηγοριοποίησεις που μπορεί να υπόκειται ένα πρόβλημα βελτιστοποίησης υπάρχουν και κάποια ζητήματα, σχετικά με την καλύτερη κατανόηση της διαδικασίας βελτιστοποίησης. Τα βασικότερα σημεία ανάλυσης είναι η διαφορά που υπάρχει μεταξύ τοπικών και ολικών βέλτιστων σημείων, οι συμβάσεις που πρέπει να ακολουθούνται για τη μαθηματική περιγραφή ενός προβλήματος βελτιστοποίησης και τέλος, η αντιμετώπιση προβλημάτων, στα οποία απαιτείται η μεγιστοποίηση κάποιας αντικειμενικής συνάρτησης. Όλα τα σημεία αυτά αναλύονται περαιτέρω παρακάτω:

a. Όταν λέμε ότι το x είναι βέλτιστο σημείο, τότε σημαίνει ότι το $x*$ είναι πιθανή λυση και ότι $f(x*) = p*$. Το σύνολο όλων των βέλτιστων σημείων περιγράφεται από τη σχέση

$$X_{opt} = \{x \mid g_i(x) = 0, \quad i = 1,..., m, \quad h_i \le 0, \quad i = 1,..., p, \quad f(x) = p*\}.$$

Εάν υπάρχει ένα βέλτιστο σημείο για το πρόβλημα που μελετάμε, τότε λέμε ότι η βέλτιστη τιμή επετεύχθει και το πρόβλημα είναι επιλύσιμο. Εάν το X_{opt} είναι κενό, λέμε ότι η βέλτιστη τιμή δεν επετεύχθει (αυτό ισχύει όταν το πρόβλημα είναι απεριόριστο προς τα κάτω). Μία πιθανή λύση x, για την οποία ισχύει $f(x) \le p* + \varepsilon$ (όπου $\varepsilon \succ 0$) ονομάζεται $\varepsilon - υποβέλτιστη$, και το σύνολο όλων των σημείων αυτών καλείται ε – υποβέλτιστο σύνολο του προβλήματος.

Λέμε ότι το x είναι τοπικό βέλτιστο, εάν υπάρχει ένα $R \succ 0$, τέτοιο ώστε

$$f(x) = \inf\{ \ f(z) \mid g_i(z) = 0, \quad i = 1,..., m \quad h_i(z) \le 0, \quad i = 1,..., p, \quad \|z - x\|_2 \le R\}$$

ή εκφραζόμενο στη standard του μορφή, το x επιλύει το πρόβλημα βελτιστοποίησης

$$\begin{aligned} &\min \ imize_x \ \{ f(z)\} \\ &subject \ \ to \ \ g_i(z) = 0, \quad i = 1,..., m \\ &\qquad\qquad h_i(z) \le 0, \quad i = 1,..., p \\ &\qquad\qquad \|z - x\|_2 \le R \end{aligned}$$

με μεταβλητή το z. Τέλος, αναφέρουμε ότι ο όρος "Ολικό Βέλτιστο" χρησιμοποιείται πολλές φορές αντί του όρου "Βέλτιστο", για να μπορέσουμε να το διακρίνουμε από το τοπικό βέλτιστο.

b. Η μαθηματική περιγραφή που πρέπει να ακολουθείται για ένα πρόβλημα βελτιστοποίησης ορίζει ότι το δεξί μέλος των περιορισμών ισότητας και ανισότητας θα είναι πάντοτε 0. Αυτό μπορεί να κανονιστεί με το να αφαιρεθεί κάθε μη μηδενικός παράγοντας του δεξιού μέλους του περιορισμού. Για έναν περιορισμό ισότητας της μορφής $z_i(x) = \tilde{z}_i(x)$, η standard μορφή $g_i(x) = 0$ επιτυγχάνεται θέτοντας $g_i(x) = z_i(x) - \tilde{z}_i(x) = 0$. Η διαδικασία είναι ίδια και για έναν περιορισμό ανισότητας. Για παράδειγμα όταν έχουμε έναν περιορισμό της μορφής $h_i(x) \ge 0$, τότε η standard μορφή επιτυγχάνεται γράφοντας $-h_i \le 0$.

c. Όπως φαίνεται από όλη την παραπάνω ανάλυση, ένα τυπικό πρόβλημα βελτιστοποίησης περιλαμβάνει την ελαχιστοποίηση κάποιας συνάρτησης, η οποία υπόκειται σε κάποιους περιοιρισμους ή όχι. Υπάρχουν περιπτώσεις, όμως, στις οποίες επιθυμούμε να μεγιστοποιήσουμε μία συνάρτηση. Σ' αυτήν την περίπτωση μετατρέπουμε το πρόβλημα μεγιστοποίησης, σε ένα ισοδύναμο πρόβλημα ελαχιστοποίησης, το οποίο τελικά ικανοποιεί τη standard μορφή και επιλύεται όπως αναφέραμε παραπάνω.
Συγκεκριμένα θέλουμε να επιλύσουμε το παρακάτω πρόβλημα μεγιστοποίησης

$$\begin{aligned} &\max \ imize \quad f(x) \\ &subject \ \ to \quad g_i(x) = 0, \quad i = 1,..., m \\ &\qquad\qquad h_i(x) \le 0, \quad i = 1,..., p \end{aligned}$$

Σύμφωνα με τα παραπάνω, μπορούμε αντί να μεγιστοποιήσουμε τη συνάρτηση $f(x)$, να ελαχιστοποιήσουμε τη συνάρτηση $-f(x)$, ικανοποιώντας πάντοτε τους περιορισμούς. Έτσι η βέλτιστη τιμή του προβλήματός μας ορίζεται ως

$$p^* = \sup\{ \ f(x) \mid g_i(x) = 0, \quad i = 1,..., m, \quad h_i(x) \le 0, \quad i = 1,..., p\}$$

Τέλος, όταν έχουμε να επιλύσουμε ένα *πρόβλημα μεγιστοποίησης*, η αντικειμενική του συνάρτηση καλείται *επίπεδο ικανοποίησης ή ωφέλειας*, αντί για συνάρτηση κόστους.

2.3 Βελτιστοποίηση για διάφορες Αντικειμενικές Συναρτήσεις

Μία νέα οικογένεια προβλημάτων βελτιστοποίησης είναι αυτή που χαρακτηρίζεται από τη διαφορετική κάθε φορά αντικειμενική συνάρτηση προς ελαχιστοποίηση. Το είδος της

συνάρτησης προς ελαχιστοποίηση δίνει και το αντίστοιχο όνομα για το πρόβλημα βελτιστοποίησης.

2.3.1 Βελτιστοποίηση Κυρτών Συναρτήσεων (Convex Optimization)

Ένα πρόβλημα βελτιστοποίησης κυρτών συναρτήσεων (convex) είναι αυτό που "υπακούει" στη μορφή

$$\begin{aligned} &\min imize && f(x) \\ &subject\ \ to && h_i(x) \le 0, \quad i = 1,\dots, p \\ & && a_i^T x = b_i, \quad i = 1,\dots, m \end{aligned}$$

όπου $f, h_1,\dots, h_p : \Re^n \to \Re$ είναι κυρτές συναρτήσεις. Συγκρίνοντας τη standard μορφή του προβλήματος βελτιστοποίησης κυρτών συναρτήσεων, με αυτήν του γενικού προβλήματος βελτιστοποίησης παρατηρούμε ότι το convex πρόβλημα έχει τρεις επιπλέον απαιτήσεις:

- Η αντικειμενική συνάρτηση πρέπει να είναι κυρτή (convex).
- Οι συναρτήσεις περιορισμών ανισότητας πρέπει να είναι κυρτές (convex), δηλαδή να ισχύει $h_i(ax + by) \le ah_i(x) + bh_i(y)$, για όλα τα $x, y \in \Re^n$ και όλα τα $a, b \in \Re$, για τα οποία ισχύει ότι $a + b = 1,\ a \ge 0,\ b \ge 0$.
- Οι συναρτήσεις περιορισμών ισότητας $g_i(x) = a_i^T x - b_i$ να είναι συγγενικές (affine).

Το σύνολο που περιλαμβάνει όλες τις εφικτές λύσεις ενός κυρτού προβλήματος βελτιστοποίησης είναι και αυτό κυρτό (convex), αφού προκύπτει από την τομή του πεδίου ορισμού του προβλήματος,

$$D = \bigcap_{i=0}^{p} dom\ \ h_i$$

η οποία είναι ένα κυρτό σύνολο, με p (κυρτά) υποεπίπεδα σύνολα $\{x \mid h_i(x) \le 0\}$ και m υπερεπίπεδα $\{x \mid a_i^T x = b_i\}$. Δηλαδή, σε ένα convex πρόβλημα βελτιστοποίησης ελαχιστοποιούμε μία κυρτή αντικειμενική συνάρτηση, πάνω σε ένα κυρτό σύνολο. Εφόσον τα υποεπίπεδα σύνολα μίας κυρτής συνάρτησης είναι κυρτά, τότε τα ε - υποβέλτιστα σύνολα είναι κυρτά. Το βέλτιστο σύνολο είναι κυρτό. Επίσης, εάν η αντικειμενική συνάρτηση είναι αυστηρώς κυρτή, τότε το βέλτιστο σύνολο περιέχει το πολύ ένα σημείο. Τέλος, μία σημαντική ιδιότητα των convex προβλημάτων βελτιστοποίησης είναι το ότι <u>ένα τοπικό βέλτιστο σημείο είναι επίσης και ολικό βέλτιστο</u>.

Κριτήριο Βελτιστότητας για Διαφορίσιμη Αντικειμενική Συνάρτηση

Υποθέτοντας ότι η αντικειμενική συνάρτηση f ενός κυρτού (convex) προβλήματος βελτιστοποίησης είναι διαφορίσιμη, έτσι ώστε για όλα τα $x, y \in dom\ \ f$ να ισχύει

$$f(y) \ge f(x) + \nabla f(x)^T (y - x).$$

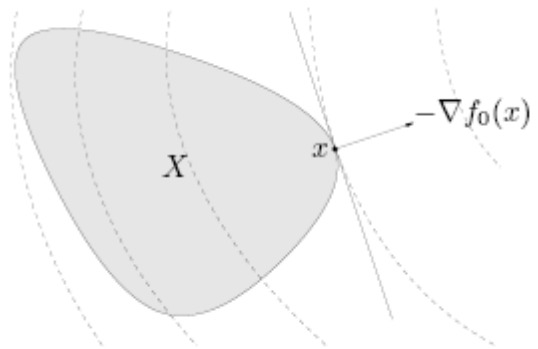
Εάν το X είναι το σύνολο των εφικτών λύσεων του προβλήματος βελτιστοποίησης, δηλαδή,

$$X = \{x \mid g_i(x) = 0, \quad i = 1,\dots, m, \quad h_i(x) \le 0, \quad i = 1,\dots, p\}$$

τότε, το x είναι βέλτιστο, εάν και μόνο εάν το $x \in X$ και ταυτόχρονα ισχύει

$$\boxed{\nabla f(x)^T (y - x) \ge 0 \quad for\ \ all\ \ y \in X} \qquad \textbf{(1)}$$

Το κριτήριο βελτιστότητας αυτό μπορεί να κατανοηθεί και γεωμετρικά: Εάν $\nabla f(x) \neq 0$ σημαίνει ότι το $-\nabla f(x)$ ορίζει ένα υποστηρικτικό υπερεπίπεδο, στο σύνολο των εφικτών λύσεων στο x. Μία σχηματική αναπαράσταση του παραπάνω φαίνεται στο παρακάτω σχήμα:



Σχήμα 2.1. Γεωμετρική ερμηνεία της συνθήκης βελτιστότητας. Η σκιασμένη περιοχή αντιπροσωπεύει το σύνολο των εφικτών λύσεων. Κάποια επίπεδα κυρτότητας της συνάρτησης f περιγράφονται με τις διακεκομμένες γραμμές. Το σημείο x είναι βέλτιστο : $-\nabla f(x)$ ορίζει ένα υποστηρικτικό υπερεπίπεδο (περιγράφεται με την κανονική ευθεία γραμμή) του X στο x.

Για κάθε κατηγορία κυρτού (convex) προβλήματος βελτιστοποίησης, η συνθήκη ή οι συνθήκες βελτιστότητας διαφοροποιούνται. Τα πιο συχνά εμφανιζόμενα προβλήματα βελτιστοποίησης και οι αντίστοιχες συνθήκες βελτιστότητας παρουσιάζονται παρακάτω.

2.3.1.1 Προβλήματα Χωρίς Περιορισμούς

Για ένα πρόβλημα χωρίς περιορισμούς, η συνθήκη βελτιστότητας ανάγεται στην πολύ γνωστή και επαρκή συνθήκη

$$\boxed{\nabla f(x) = 0}$$

ώστε το x να είναι βέλτιστο. Η συνθήκη βελτιστότητας αυτή προκύπτει από τη γενική συνθήκη (1) ως εξής: Υποθέτουμε ότι το x είναι εφικτή λύση, πράγμα το οποίο σημαίνει ότι $x \in dom\ f$ και ότι για όλα τα πιθανά y έχουμε $\nabla f(x)^T (y - x) \geq 0$. Εφόσον η f είναι διαφορίσιμη, το πεδίο ορισμού της (εξ ορισμού) θα είναι ανοιχτό, άρα όλα τα y αρκετά κοντά στο x είναι εφικτές λύσεις. Ας ορίσουμε το $y = x - t * \nabla f(x)$, όπου $t \in \Re$ είναι μία παράμετρος. Για μικρό και θετικό t το y είναι εφικτή λύση, άρα

$$\nabla f(x)^T (y - x) = -t\|\nabla f(x)\|_2^2 \geq 0$$

από το όποιο συμπεραίνουμε ότι $\nabla f(x) = 0$.

Υπάρχουν πολλές πιθανές καταστάσεις, εξαρτώμενες από τον αριθμό των λύσεων της συνθήκης βελτιστότητας. Εάν δεν υπάρχουν λύσεις που να ικανοποιούν το κριτήριο βελτιστότητας, τότε δεν υπάρχουν και βέλτιστα σημεία. Απ' την άλλη μπορούμε να έχουμε πολλαπλές λύσεις για το κριτήριο, περίπτωση στην οποία κάθε λύση είναι και ένα ελάχιστο σημείο για τη συνάρτηση f.

2.3.1.2 Προβλήματα με Περιορισμούς Ισότητας

Ας υποθέσουμε ότι έχουμε το παρακάτω convex πρόβλημα βελτιστοποίησης

$$\begin{aligned} &\min\ imize \quad f(x) \\ &subject\ \ to\ \ Ax = b \end{aligned}$$

Υποθέτουμε ότι το σύνολο των εφικτών λύσεων δεν είναι κενό. Η συνθήκη βελτιστότητας για μία εφικτή λύση x είναι η

$$\boxed{\nabla f(x)^T (y - x) \geq 0}$$

η οποία θα πρέπει να περιλαμβάνει τα y τα οποία ικανοποιούν τη σχέση $Ay = b$. Εφόσον το x είναι εφικτή λύση, κάθε εφικτή λύση y έχει τη μορφή $y = x + v$, για κάποιο $v \in N(A)$. Η συνθήκη βελτιστότητας μπορεί επομένως να εκφραστεί ως:

$$\nabla f(x)^T v \geq 0 \quad for\ \ all\ \ v \in N(A)$$

2.3.1.3 Ελαχιστοποίηση ως προς μη αρνητική μεταβλητή

Ένα άλλο, συχνά συναντώμενο πρόβλημα convex βελτιστοποίησης είναι αυτό της μορφής

$$\begin{aligned} &\min\ imize \quad f(x) \\ &subject\ \ to \quad x \geq 0 \end{aligned}$$

όπου οι μόνοι περιορισμοί ανισότητας που υπάρχουν είναι οι μεταβλητές μας να είναι θετικές. Τότε η συνθήκη βελτιστότητας παίρνει τη μορφή

$$\boxed{x \geq 0, \quad \nabla f(x)^T (y - x) \geq 0 \quad for\ \ all\ \ y \geq 0}$$

Ο όρος $\nabla f(x)^T y$, ο οποίος είναι μία γραμμική συνάρτηση του y δεν περιλαμβάνει κάτω όριο στο $y \geq 0$, εκτός και αν ισχύει $\nabla f(x) \geq 0$. Η συνθήκη τότε απλοποιείται στην $-\nabla f(x)^T x \geq 0$. Όμως ισχύει ότι $x \geq 0$ και $\nabla f(x) \geq 0$, άρα πρέπει να έχουμε $\nabla f(x)^T x = 0$,

$$\sum_{i=1}^{n} (\nabla f(x))_i x_i = 0 .$$

Τώρα, κάθε όρος του παραπάνω αθροίσματος είναι ένα γινόμενο δύο μη αρνητικών αριθμών, επομένως συμπεραίνουμε ότι κάθε όρος πρέπει να είναι μηδέν. Έτσι, η συνθήκη βελτιστότητας μπορεί πλέον να εκφραστεί ως

$$\boxed{x \geq 0, \quad \nabla f(x) \geq 0, \quad x_i (f(x))_i = 0, \quad i = 1,\ldots, n}$$

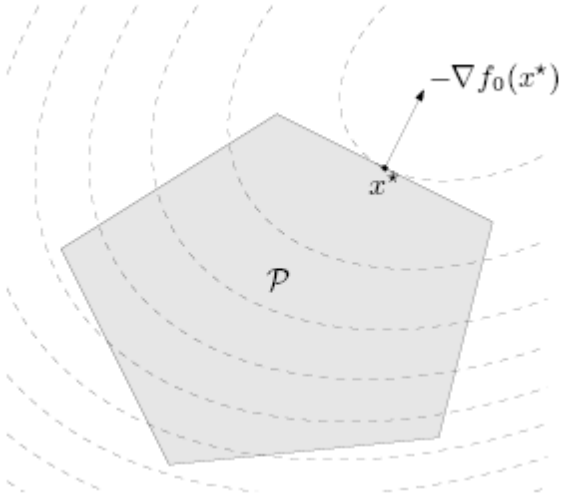
Η τελευταία συνθήκη ονομάζεται *συμπληρωματική (complementary)*, αφού σημαίνει ότι τα δείγματα σποραδικότητας (το σύνολο των σημείων που αντιστοιχεί σε μη μηδενικά στοιχεία) του διανύσματος x και $\nabla f(x)$ είναι συμπληρωματικά.

2.3.2 Τετραγωνική Βελτιστοποίηση (Quadratic Optimization)

Ένα κυρτό (convex) πρόβλημα βελτιστοποίησης καλείται *τετραγωνικό πρόβλημα (Quadratic Problem - QP)* εάν η αντικειμενική συνάρτηση είναι (κυρτή) τετραγωνική και οι συναρτήσεις περιορισμού είναι συγγενικές. Ένα τετραγωνικό πρόβλημα μπορεί να εκφραστεί με την παρακάτω μορφή

$$\begin{aligned} &\min\ imize \quad (1/2)x^T Px + q^T x + r \\ &subject\ \ to \quad Gx \leq h \\ &\qquad\qquad\quad Ax = b \end{aligned}$$

όπου $P \in S_+^n$, $G \in R^{mxn}$, και $A \in R^{pxn}$. Σε ένα τετραγωνικό πρόβλημα ελαχιστοποιούμε την κυρτή τετραγωνική συνάρτηση πάνω σε ένα πολύεδρο, όπως φαίνεται και στο παρακάτω σχήμα.



Σχήμα 2.2. Γεωμετρική αναπαράσταση Τετραγωνικού Προβλήματος. Το σύνολο των εφικτών λύσεων P, το οποίο είναι ένα πολύεδρο, αντιπροσωπεύεται από τη σκιασμένη επιφάνεια. Οι γραμμές περιγράμματος της αντικειμενικής συνάρτησης, η οποία είναι κυρτή τετραγωνική παρουσιάζονται με τη μορφή διακεκομμένων γραμμών. Το σημείο $x*$ είναι το βέλτιστο.

Όταν και η αντικειμενική συνάρτηση, αλλά και οι συναρτήσεις περιορισμού ανισότητας είναι τετραγωνικές, δηλαδή όταν ισχύει

$$\begin{aligned} &\min\ imize \quad (1/2)x^T Px + q^T x + r \\ &subject\ \ to \quad (1/2)x^T P_i x + q_i^T x + r_i \le 0, \quad i = 1,\dots, m \\ &\qquad\qquad Ax = b \end{aligned}$$

(όπου $P_i \in S_+^n$ και $i = 1,\dots, m$), τότε το πρόβλημα ονομάζεται *Τετραγωνικό Πρόβλημα Τετραγωνικών Περιορισμών (Quadratically Constrained Quadratic Program - QCQP)*. Σε ένα τέτοιου είδους πρόβλημα, ελαχιστοποιούμε μία κυρτή τετραγωνική συνάρτηση, πάνω σε μία πιθανή περιοχή, η οποία είναι η τομή ελλειψοειδών (όταν $P_i \ge 0$). Τα τετραγωνικά προβλήματα βελτιστοποίησης συμπεριλαμβάνουν γραμμικά προβλήματα βελτιστοποίησης ως ειδική περίπτωση, θέτοντας στην εξίσωση περιγραφής του QP, $P = 0$. Τα τετραγωνικά προβλήματα με τετραγωνικούς περιορισμούς συμπεριλαμβάνουν τα απλά τετραγωνικά προβλήματα (άρα και τα γραμμικά προβλήματα) ως ειδική περίπτωση θέτοντας στην εξίσωση περιορισμού ανισότητας του QCQP $P_i = 0$, για $i = 1,\dots, m$.

2.3.3 Βελτιστοποίηση Μη Κυρτών Συναρτήσεων

Εκτός από τα προβλήματα κυρτής βελτιστοποίησης έχουμε και τα προβλήματα βελτιστοποίησης αντικειμενικών συναρτήσεων οι οποίες είναι *κοίλες (concave optimization)*. Στα προβλήματα αυτού του είδους η αντικειμενική συνάρτηση είναι κοίλη και οι περιορισμοί είναι κυρτοί, ενώ

για να βρούμε τη βέλτιστη λύση, αρκεί να μεγιστοποιήσουμε τη δοθείσα κάθε φορά συνάρτηση. Η μορφή τέτοιων προβλημάτων περιγράφεται παρακάτω

$$\begin{aligned} &\max\ imize \quad f(x) \\ &subject\ \ to \quad h_i(x) \le 0, \quad i = 1,\dots, p\ . \\ &\qquad\qquad a_i^T x = b_i, \quad i = 1,\dots, m \end{aligned}$$

Για την επίλυση ενός κοίλου προβλήματος, στο οποίο απαιτείται η μεγιστοποίηση της αντικειμενικής του συνάρτησης, χρειάζεται απλά η μετατροπή του στο ισοδύναμό του κυρτό πρόβλημα, στο οποίο θα ελαχιστοποιείται το αντίθετο της συνάρτησης, δηλαδή το $-f(x)$.

Τα προβλήματα κυρτής βελτιστοποίησης (και τα τετραγωνικά προβλήματα βελτιστοποίησης) και κοίλης βελτιστοποίησης ανήκουν στην ευρύτερη κατηγορία των *Μη Γραμμικών Προβλημάτων Βελτιστοποίησης (Non Linear Programming)*. Υπάρχει όμως και μία άλλη μεγάλη κατηγορία προβλημάτων, τα *Γραμμικά Προβλήματα Βελτιστοποίησης (Linear Programming)*, τα οποία έχουν αντίστοιχα ευρείες εφαρμογές με τα μη γραμμικά. Τα προβλήματα αυτά περιλαμβάνουν γραμμικές συναρτήσεις προς βελτιστοποίηση και περιορισμούς ισότητας και ανισότητας, οι οποίοι είναι γραμμικοί. Για να κατανοήσουμε τη λειτουργία των γραμμικών προβλημάτων βελτιστοποίησης χρησιμοποιούμε έναν πολυχώρο, για παράδειγμα ένα πολύγωνο ή ένα πολύεδρο και μία πραγματική συγγενική (affine) συνάρτηση

$$f(x_1, x_2, \dots, x_n) = c_1 x_1 + c_2 x_2 + \dots + c_n x_n + d$$

η οποία ορίζεται για τον πολυχώρο αυτό. Μία μέθοδος γραμμικού προγραμματισμού (linear programming) ψάχνει στον πολυχώρο αυτό και βρίσκει το σημείο που ανήκει σ' αυτόν και για το οποίο η συνάρτηση εμφανίζει τη μεγαλύτερη ή τη μικρότερη τιμή της. Τα γραμμικά προβλήματα βελτιστοποίησης μπορούν να εκφραστούν σύμφωνα με την παρακάτω κανονική μορφή

$$\begin{aligned} &optimize \quad c^T x \\ &subject\ \ to\ Ax \le b \end{aligned}$$

όπου το x αντιπροσωπεύει το διάνυσμα των μεταβλητών ως προς τις οποίες βελτιστοποιούμε, τα c, b είναι διανύσματα γνωστών συντελεστών και το A είναι ένας πίνακας με γνωστούς συντελεστές. Οι εξισώσεις περιορισμών ορίζουν ένα κυρτό πολύεδρο, πάνω στο οποίο η αντικειμενική συνάρτηση πρέπει να βελτιστοποιηθεί.

2.4 Μέθοδοι Επίλυσης Κυρτών Προβλημάτων Βελτιστοποίησης (Convex)

2.4.1 Εισαγωγικά Στοιχεία

Η μέθοδος επίλυσης για μία κλάση προβλημάτων βελτιστοποίησης (και συνεπώς και για ένα convex πρόβλημα βελτιστοποίησης) είναι ουσιαστικά ένας αλγόριθμος, ο οποίος υπολογίζει μία λύση για ένα πρόβλημα που ανήκει σε κάποια κλάση (με μία συγκεκριμένη ακρίβεια). Από το τέλος της δεκαετίας του 1940 έχει καταβληθεί μεγάλη προσπάθεια για την ανάπτυξη αλγορίθμων επίλυσης διαφόρων κλάσεων προβλημάτων βελτιστοποίησης, αναλύοντας τις ιδιότητές τους και αναπτύσσοντας καλές υλοποιήσεις λογισμικού. Η αποτελεσματικότητα των αλγορίθμων αυτών, δηλαδή η ικανότητά μας να λύνουμε προβλήματα βελτιστοποίησης ποικίλει και εξαρτάται από παράγοντες όπως η συγκεκριμένες μορφές της αντικειμενικής συνάρτησης, αλλά και των συναρτήσεων περιορισμών, τον αριθμό των μεταβλητών και των περιορισμών και την ύπαρξη ή όχι ειδικής δομής, όπως για παράδειγμα της σποραδικότητας (ένα πρόβλημα

εμφανίζει σποραδικότητα, εάν κάθε συνάρτηση περιορισμού εξαρτάται μόνο από ένα μικρό αριθμό μεταβλητών).
Οι προσεγγίσεις του γενικού προβλήματος βελτιστοποίησης περιλαμβάνουν και κάποιου είδους συμβιβασμούς, όπως ένα αρκετά μεγάλο χρόνο για τους υπολογισμούς, ή την πιθανότητα να μην μπορέσουμε να βρούμε κάποια εφικτή λύση. Παρ' όλα αυτά, για πάρα πολλές κλάσεις προβλημάτων υπάρχουν αποτελεσματικοί αλγόριθμοι, οι οποίοι μπορούν να επιλύσουν μεγάλα προβλήματα, με πάρα πολλούς περιορισμούς και μεταβλητές. Εκτός από το πιο κοινό σε όλους πρόβλημα γραμμικού προγραμματισμού (το οποίο ουσιαστικά είναι μία ειδική περίπτωση convex προβλήματος βελτιστοποίησης), τα πιο γνωστά και σημαντικά προβλήματα convex βελτιστοποίησης είναι τα *προβλήματα Ελαχίστων Τετραγώνων (Least Squares)*, η δομή των οποίων θα αναλυθεί σε επόμενο κεφάλαιο.
Σύμφωνα πάντα και με όλη την παραπάνω ανάλυση, οι μέθοδοι επίλυσης convex προβλημάτων βελτιστοποίησης θα διαχωρίζονται σε αυτές για προβλήματα χωρίς περιορισμούς και σ' αυτές για προβλήματα με περιορισμούς ισότητας ή ανισότητας ή και τα δύο.

2.4.2 Επίλυση Προβλημάτων Χωρίς Περιορισμούς

Στο κεφάλαιο αυτό θα παρουσιάζουμε κάποια γενικά στοιχεία και κοινές παραδοχές που ακολουθούν οι μέθοδοι επίλυσης προβλημάτων βελτιστοποίησης χωρίς περιορισμούς και θα αναλύσουμε διεξοδικά μία απ' τις πιο ευρέως χρησιμοποιούμενες μεθόδους, η οποία ονομάζεται *Gradient Descent*. Όπως αναφέραμε και σε προηγούμενο κεφάλαιο θέλουμε να επιλύσουμε το πρόβλημα

$$\min \; imize \quad f(x) \tag{1}$$

όπου η συνάρτηση f είναι convex και δύο φορές διαφορίσιμη. Υποθέτουμε ότι για το πρόβλημα αυτό υπάρχει εφικτή λύση και είναι μοναδική, ενώ θεωρούμε ότι η βέλτιστη λύση $\inf_x f(x) = f(x^*)$ ορίζεται ως p^*. Εφόσον η f έχει όλες τις παραπάνω ιδιότητες η ικανή και αναγκαία συνθήκη ώστε ένα σημείο x^* να είναι βέλτιστο, είναι

$$\nabla f(x^*) = 0 \tag{2}$$

Έστι, λύνοντας το πρόβλημα χωρίς περιορισμούς (1) είναι το ίδιο με το να βρούμε λύση στο πρόβλημα (2), το οποίο είναι ένα σύνολο από n εξίσωσεις στις n μεταβλητές $x_1,..., x_n$. Σε ορισμένες ειδικές περιπτώσεις, μπορούμε να βρούμε λύση για το πρόβλημα (1), λύνοντας αναλυτικά τη συνθήκη βελτιστότητας (2), όμως, συνήθως, το πρόβλημα πρέπει να επιλυθεί χρησιμοποιώντας έναν *επαναληπτικό αλγόριθμο*. Αυτό σημαίνει ότι χρησιμοποιούμε έναν αλγόριθμο ο οποίος υπολογίζει μία ακολουθία από σημεία $x^{(0)}, x^{(1)},..... \in dom \; f$, με $f(x^{(k)} \to p^* \; as \; k \to \infty)$. Μία τέτοια ακολουθία σημείων ονομάζεται *ακολουθία ελαχιστοποίησης* του προβλήματος (1). Ο αλγόριθμος τερματίζει όταν $f(x^{(k)}) - p^* \le \varepsilon$, όπου $\varepsilon \ge 0$ είναι μία προσδιορισμένη *ανεκτικότητα (tolerance)*.
Ο κάθε επαναληπτικός αλγόριθμος επίλυσης προβλημάτων ελαχιστοποίησης απαιτεί ένα κατάλληλο αρχικό σημείο. Το αρχικό σημείο πρέπει να ανήκει στο πεδίο ορισμού της f και επιπλέον το σύνολο υπο – επιπέδου

$$S = \{x \in dom \; f \mid f(x) \le f(x^{(0)})\}$$

πρέπει να είναι κλειστό. Η συνθήκη αυτή ικανοποιείται για όλα τα $x^{(0)} \in dom \; f$, εάν η συνάρτηση f είναι κλειστή και αν όλα τα σύνολα των υπο – επιπέδων είναι κλειστά. Συνεχείς

συναρτήσεις με $dom\ f = \Re^n$ είναι κλειστές, έτσι αν $dom\ f = \Re^n$, τότε η αρχική συνθήκη του συνόλου υπο – επιπέδων ικανοποιείται από κάθε $x^{(0)}$. Άλλη μία σημαντική κλάση κλειστών συναρτήσεων είναι οι συνεχείς συναρτήσεις με ανοιχτά πεδία ορισμού, για τα οποία η $f(x)$ τείνει στο άπειρο, όσο το x πλησιάζει το $bd\ dom\ f$.

2.4.2.1 Gradient Descent Method

Η <u>μέθοδος του Gradient Descent</u> (όπως και κάθε descent μέθοδος) παράγει μία ακολουθία ελαχιστοποίησης $x^{(k)}$, $k = 1,\ldots$, όπου

$$x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)}$$

και $t^{(k)} \succ 0$ (εκτός όταν το $x^{(k)}$ είναι βέλτιστο). Ο όρος Δx αποτελεί ένα διάνυσμα στο $\Re^n$, το οποίο καλείται *βήμα ή κατεύθυνση εύρεσης (step or search direction)* και η μεταβλητή $k = 0, 1,\ldots\ldots$ αντιπροσωπεύει τον αριθμό της επανάληψης. Η βαθμωτή μεταβλητή $t^{(k)} \succ 0$ ονομάζεται *μέγεθος βήματος ή μήκος βήματος (step size or step length)* στην k επανάληψη (παρ' ότι δεν είναι ίσο με το $\left\|x^{(k+1)} - x^{(k)}\right\|$, παρά μόνο όταν $\left\|\Delta x^{(k)}\right\| = 1$). Οι όροι *βήμα εύρεσης (search step) και βαθμωτός παράγοντας (scale factor)* είναι πιο ακριβής όσον αφορά την περιγραφή του $t^{(k)} \succ 0$, αλλά οι δύο πρώτοι όροι που αναφέρθηκαν είναι πιο ευρέως χρησιμοποιούμενοι.

Η μέθοδος που μελετάμε είναι descent μέθοδος (μέθοδος πτωτική), πράγμα το οποίο σημαίνει ότι

$$f(x^{(k+1)}) \prec f(x^{(k)}),$$

εκτός όταν το $x^{(k)}$ είναι βέλτιστο. Αυτό συνεπάγεται ότι για όλα τα k έχουμε $x^{(k)} \in S$, το αρχικό σύνολο υπο – επιπέδων, και πιο συγκεκριμένα έχουμε $x^{(k)} \in dom\ f$. Από την κυρτότητα της συνάρτησης f γνωρίζουμε ότι εάν $\nabla f(x^{(k)})^T (y - x^{(k)}) \geq 0$, τότε $f(y) \geq f(x^{(k)})$, έτσι η κατεύθυνση της αναζήτησης για κάθε <u>descent μέθοδο πρέπει</u> να ικανοποιεί την

$$\boxed{\nabla f(x^{(k)})\Delta x^{(k)} \prec 0}$$

Και γι' αυτό καλείται και *κατεύθυνση πτώσης (descent direction)*. Συγκεκριμένα, για τη <u>μέθοδο Gradient Descent</u> που μελετάμε, η κατεύθυνση πτώσης ισούται με την αρνητική παράγωγο της αντικειμενικής συνάρτησης, δηλαδή $\boxed{\Delta x = -\nabla f(x)}$.

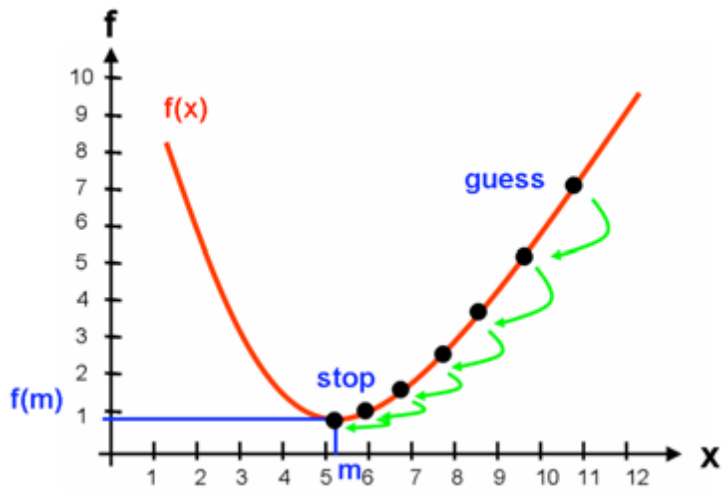
Η <u>μέθοδος Gradient Descent</u> μπορεί να εφαρμοστεί για ελαχιστοποίσης αντικειμενικής συνάρτησης ως προς μία μεταβλητή , αλλά και ως προς περισσότερες από μία μεταβλητές. Έτσι, ανάλογα με τον αριθμό των μεταβλητών ως προς τις οποίες γίνεται η βελτιστοποίηση, η <u>μέθοδος Gradient Descent</u> μπορεί να πάρει τις δύο ακόλουθες μορφές:

Αλγόριθμος Gradient Descent – Μία Μεταβλητή Ελαχιστοποίησης

---

given a starting point $x \in dom\ f$ (**guess**)
repeat
1. $\Delta x = -\nabla f(x)$
2. Choose step size t via a specific search method*
3. Update: $x = x + t\Delta x$

until stopping criterion is satisfied**



Σχήμα 2.3. Σχηματική αναπαράσταση αλγορίθμου Gradient Descent (μία μεταβλητή)

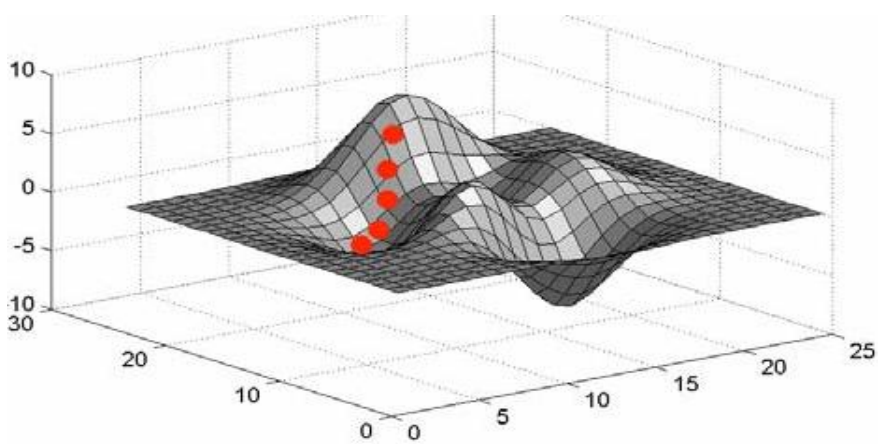
Αλγόριθμος Gradient Descent – Πολλές Μεταβλητές Ελαχιστοποίησης

---

given a starting point $x \in dom \quad f$ (**guess**)
repeat

1. $\Delta x = -\nabla f(x_1, x_2, ..., x_n) = -\{\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_n}\}$
2. Choose step size t via a specific search method*
3. Update: $x = x + t\Delta x$

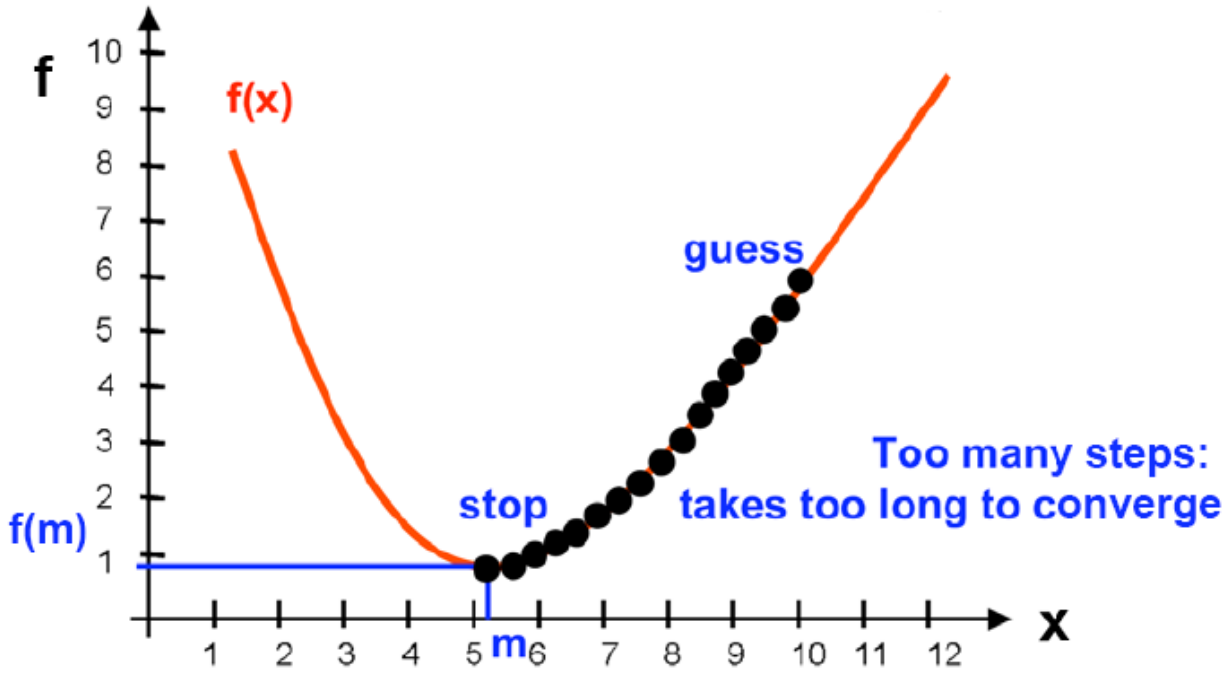
until stopping criterion is satisfied**



Σχήμα 2.4. Σχηματική αναπαράσταση αλγορίθμου Gradient Descent (πολλές μεταβλητές)

***Επιλογή Μεγέθους Βήματος (step size)**

Επιλέγοντας το βήμα του αλγορίθμου πρέπει να ληφθούν υπόψιν τρία βασικά ζητήματα : μετά από κάθε επανάληψη, η αντικειμενική συνάρτηση να ελαχιστοποιείται και ο αλγόριθμος να μπορεί να οδηγηθεί σε σύγκλιση, η σύγκλιση να είναι όσο το δυνατόν γρηγορότερη και τέλος, τα αποτελέσματα της ελαχιστοποίησης να είναι ικανοποιητικά και σωστά. Αυτό σημαίνει ότι για μικρό βήμα αλγορίθμου, έχουμε πολλή αργή σύγκλιση, όπως φαίνεται και στην παρακάτω σχηματική αναπαράσταση, ενώ για πολύ μεγάλο βήμα οδηγούμαστε συνήθως σε κακά αποτελέσματα. Τέλος πρέπει να σημειωθεί ότι κατά το πέρας των επαναλήψεων το βήμα μπορεί να μεταβάλλεται, εάν αυτό οδηγεί σε καλύτερη σύγκλιση και σε πιο ακριβή αποτελέσματα.



Σχήμα 2.5. Σχηματική αναπαράσταση επιλογής βήματος του αλγορίθμου

****Επιλογή Κριτηρίου Σύγκλισης (stopping criterion)**

Η μορφή του κριτηρίου σύγκλισης του αλγορίθμου εξαρτάται επίσης από το αν η ελαχιστοποίηση γίνεται σε μία διάσταση ή σε περισσότερες διαστάσεις. Έτσι μπορεί να πάρει τις δύο παρακάτω μορφές:

Μία διάσταση :

$\left|\nabla f(x)\right| \le \varepsilon$, όπου το ε είναι ένας μικρός, θετικός αριθμός.

Πολλές διαστάσεις :

$\left\|\nabla f(x_1,\ldots, x_n)\right\|_2 \le \varepsilon \Rightarrow \sqrt{\sum_{i=1}^{n}(\frac{\partial f}{\partial x_i})^2} \le \varepsilon \Rightarrow \sqrt{(\frac{\partial f}{\partial x_1})^2 + (\frac{\partial f}{\partial x_2})^2 + \ldots + (\frac{\partial f}{\partial x_n})^2} \le \varepsilon$, όπου το ε είναι ένας μικρός, θετικός αριθμός.

Σε κάποιες υλοποιήσεις του αλγορίθμου το κριτήριο σύγκλισης ελέγχεται αμέσως μετά από το βήμα 1° του αλγορίθμου, ενώ σε κάποιες άλλες, μετά απ' το update της μεταβλητής ως προς την οποία ελαχιστοποιούμε (βήμα 3°), όπως δηλαδή περιγράφεται και παραπάνω στη δομή του αλγορίθμου.

2.4.3 Επίλυση Προβλημάτων με Περιορισμούς

Κάθε πρόβλημα βελτιστοποίησησης με περιορισμούς μπορεί να αναχθεί σε ένα ισοδύναμο πρόβλημα χωρίς περιορισμούς και να επιλυθεί με τις μεθόδους που αναφέραμε παραπάνω. Αυτό γίνεται με το να λάβουμε υπόψιν μας τους περιορισμούς, προσθέτοντας στην αντικειμενική συνάρτηση ένα άθροισμα διαφόρων βαρών πολλαπλασιασμένων με τις συναρτήσεις περιορισμών. Έτσι έχουμε πλέον μόνο μία νέα αντικειμενική συνάρτηση προς ελαχιστοποίηση, η οποία είναι και αυτή διαφορίσιμη και η ελαχιστοποίησή της γίνεται σύμφωνα με τη μέθοδο του Gradient Descent που αναλύσαμε παραπάνω.

2.4.3.1 Lagrange Multipliers

Η πιο κλασική προσέγγιση για την επίλυση προβλημάτων ελαχιστοποίησης με περιορισμούς είναι η *μέθοδος των Πολλαπλασιαστών Lagrange*. Όπως αναφέρθηκε και στα εισαγωγικά στοιχεία για την επίλυση μεθόδων με περιορισμούς, η μέθοδος αυτή μετατρέπει το πρόβλημα ελαχιστοποίησης με περιορισμούς σε ένα πρόβλημα χωρίς περιορισμούς. Η *Lagrangian* ενός προβλήματος με περιορισμούς είναι μία scalar – valued συνάρτηση, η οποία είναι και διαφορίσιμη. Ανάλογα με το αν υπάρχουν μόνο περιορισμοί ισότητας ή περιορισμοί ισότητας και ανισότητας διακρίνουμε δύο μορφές για τη συνάρτηση αυτή, η ανάλυση των οποίων γίνεται παρακάτω.

2.4.3.1.1 Lagrangian & Περιορισμοί Ισότητας

Θεωρώντας ότι έχουμε να επιλύσουμε το πρόβλημα βελτιστοποίησης

$$\min_x \{ f(x) \}$$

$$subject \quad to \; g_i(x) = 0, \quad i = 1,..., m$$

με πεδίο ορισμού του προβλήματος το μη κενό σύνολο $D = \bigcap_{i=1}^{p} dom \;\; g_i$. Ορίζουμε ως Lagrangian $L : \Re^n x \Re^m \rightarrow \Re$, με πεδίο ορισμού $L = dom \;\; D \times \Re^m$, τη συνάρτηση

$$L(\underline{x}, \lambda) = f(\underline{x}) + \lambda^T g(\underline{x}) \Rightarrow L(\underline{x}, \lambda) = f(\underline{x}) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

όπου το λ_i ονομάζεται *πολλαπλασιαστής Lagrange* και σχετίζεται με τον i^{th} περιορισμό ισότητας $g_i(x) = 0$. Έτσι πλέον έχουμε να ελαχιστοποιήσουμε μία νέα αντικειμενική συνάρτηση, τη Lagrangian η οποία περιλαμβάνει πλέον και τους περιορισμούς ισότητας.

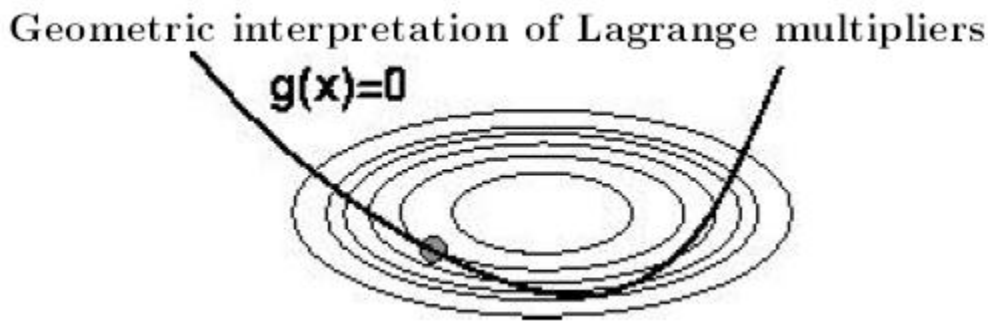
Το ακόλουθο θεώρημα υποδηλώνει ότι τα στατικά σημεία της Lagrangian συνάρτησης είναι δυνατές λύσεις του προβλήματος βελτιστοποίησης με περιορισμούς : κάθε υποψήφια λύση πρέπει να εξεταστεί για να αποφασιστεί εάν ελαχιστοποιεί την αντικειμενική συνάρτηση.

Θεώρημα: Έστω x_o μία τοπική λύση για το πρόβλημα ελαχιστοποίησης με περιορισμούς, όπου οι μερικές παράγωγοι των συναρτήσεων περιορισμών $\nabla_x(g_1(x)), \ldots, \nabla_x(g_M(x))$ ως προς κάθε στοιχείο είναι γραμμικά ανεξάρτητες. Τότε υπάρχει ένα μοναδικό διάνυσμα λ_0 τέτοιο ώστε

$$\nabla_x(L(x_0, \lambda_0)) = 0$$

Επιπλέον, η τετραγωνική μορφή $y^T \nabla_x^2(L(x_0, \lambda_0))\, y$ είναι μη αρνητική για όλα τα y που ικανοποιούν την εξίσωση $(\nabla_x(g(x)))^T y = 0$.

Το τελευταίο αποτέλεσμα του θεωρήματος καταλήγει στο ότι η Hessian της Lagrangian υπολογιζόμενη για κάθε στατικό σημείο είναι σίγουρα μη αρνητική σε σχέση με όλα τα ορθογώνια διανύσματα, των παραγώγων των περιορισμών. Το αποτέλεσμα αυτό γενικεύει την ιδέα της θετικά ορισμένης Hessian σε προβλήματα χωρίς περιορισμούς.
Το παραπάνω θεώρημα έχει και την αντίστοιχη γεωμετρική ερμηνεία όπως περιγράφεται και στο παρακάτω σχήμα. Ο περιορισμός αντιπροσωπεύεται από το περίγραμμα στο x επίπεδο. Ένας χάρτης περιγράμματος της αντικειμενικής συνάρτησης προσδιορίζει εκείνες τις τιμές του x, για τις οποίες ισχύει $f(x) = c$.



Σχήμα 2.6. Γεωμετρική ερμηνεία πολλαπλασιαστών Lagrange. Η έντονη μαύρη γραμμή αντιστοιχεί σε όλες τις τιμές του x που ικανοποιούν τους περιορισμούς. Οι πιο λεπτές γραμμές είναι περιγράμματα σταθερών τιμών της αντικειμενικής συνάρτησης. Το περίγραμμα το οποίο αντιστοιχεί στην ελάχιστη τιμή της αντικειμενικής συνάρτησης και εφάπτεται στη γραμμή που αντιπροσωπεύει τους περιορισμούς είναι και η λύση στο πρόβλημα βελτιστοποίησης με περιορισμούς.

Στο παραπάνω σχήμα, όσο το c γίνεται μικρότερο, το περίγραμμα συρρικνώνεται σε ένα μικρό κύκλο, στο κέντρο του σχήματος. Η λύση για το πρόβλημα βελτιστοποίησης με περιορισμούς εμφανίζεται όταν η μικρότερη τιμή του c επιλέγεται έτσι ώστε το περίγραμμα να "ακουμπάει" στο περίγραμμα του περιορισμού. Στο σημείο αυτό, η παράγωγος της αντικειμενικής συνάρτησης και του περιγράμματος του περιορισμού είναι ανάλογες μεταξύ τους. Το διάνυσμα αναλογίας είναι το λ_0, ο γνωστός σε όλους πολλαπλασιαστής Lagrange. Η ακριβής τιμή του πολλαπλασιαστή Lagrange πρέπει να είναι τέτοια, ώστε ο περιορισμός να ικανοποιείται απολύτως. Σημειώνουμε ότι ο περιορισμός μπορεί να εφάπτεται στο χάρτη περιγράμματος της αντικειμενικής συνάρτησης για μεγαλύτερες τιμές του c. Αυτές οι πιθανές, αλλά λανθασμένες λύσεις μπορεί να αποκλειστούν μόνο με το να υπολογιστεί η αντικειμενική συνάρτηση γι' αυτές.

2.4.3.1.2 Lagrange & Περιορισμοί Ανισότητας (& Ισότητας)

Όταν κάποιοι από τους περιορισμούς είναι και ανισότητες, η τεχνική των πολλαπλασιαστών Lagrange μπορεί να χρησιμοποιηθεί με τον ίδιο τρόπο, αλλά η λύση πρέπει να εξεταστεί προσεκτικά σε όλες τις λεπτομέρειές της. Το πρόβλημα βελτιστοποίησης με περιορισμούς ισότητας και ανισότητας αντιπροσωπεύεται από την Lagrangian συνάρτηση $L : \Re^n \times \Re^m \times \Re^p \to \Re$, η οποία περιγράφεται από την παρακάτω σχέση

$$L(\underline{x}, \lambda, \mu) = f(\underline{x}) + \lambda^T g(\underline{x}) + \mu^T h(\underline{x}) = f(\underline{x}) + \sum_{i=1}^{m} \lambda_i g_i(\underline{x}) + \sum_{i=1}^{p} \mu_i h_i(\underline{x})$$

όπου το λ_i ονομάζεται *πολλαπλασιαστής Lagrange* και σχετίζεται με τον i^{th} περιορισμό ισότητας $g_i(x) = 0$ και το μ_i καλείται επίσης *πολλαπλασιαστής Lagrange* και σχετίζεται με τον i^{th} περιορισμό ανισότητας.

Θεώρημα: Έστω x_o μία τοπική λύση για το πρόβλημα ελαχιστοποίησης με περιορισμούς. Εάν οι μερικές παράγωγοι ως προς κάθε στοιχείο των συναρτήσεων g και οι μερικές παράγωγοι ως προς κάθε στοιχείο των συναρτήσεων h, για τις οποίες ισχύει $h(x_0) = 0$, είναι γραμμικά ανεξάρτητες, τότε

$$\nabla_x (L(x_0, \lambda_0, \mu_0)) = 0$$

όπου $\mu_0 \geq 0$ και $\mu_i^0 h_i(x^0) = 0$.

Το αποτέλεσμα του θεωρήματος αυτού, όσον αφορά την αντιμετώπιση των περιορισμών ανισότητας διαφέρουν σημαντικά απ' αυτό που προέκυψε για την αντιμετώπιση των περιορισμών ισότητας. Είτε ένα στοιχείο από τους περιορισμούς εξισώνεται με τη μέγιστη τιμή του (το μηδέν σ' αυτήν την περίπτωση) και το αντίστοιχο στοιχείο του Lagrange πολλαπλασιαστή του είναι μη – αρνητικό (και είναι συνήθως θετικό), είτε ένα στοιχείο είναι μικρότερο από τον περιορισμό και το αντίστοιχο στοιχείο του Lagrange πολλαπλασιαστή του είναι μηδέν. Απ' το τελευταίο συμπέρασμα προκύπτει ότι κάποια στοιχεία των περιορισμών ανισότητας δεν είναι τόσο αυστηρά όσο άλλα, αλλά η χαλαρότητα αυτών των στοιχείων δεν επηρεάζει τη λύση.

2.5 Μέθοδος Ελαχίστων Τετραγώνων (Least Squares)

Θεωρούμε το παρακάτω γραμμικό σύστημα εξισώσεων $\boxed{A_{m\times n}\vec{x} = \vec{b}}$ **(1)**, όπου $A \in \Re^{m\times n}$ είναι ένας πίνακας με m γραμμές και n στήλες, με $m \succ n$ και το $x \in \Re^n$ είναι το διάνυσμα των μεταβλητών, ως προς τις οποίες θέλουμε να ελαχιστοποιήσουμε την αντικειμενική μας συνάρτηση. Ένα τέτοιο σύστημα εξισώσεων δεν έχει ακριβή λύση, γι' αυτό ψάχνουμε να βρούμε μία λύση $\vec{x}$ με το μικρότερο λάθος για το διάνυσμα $A\vec{x} - \vec{b}$, χρησιμοποιώντας κάποια νόρμα διανύσματος για να προσδιορίσουμε το μέγεθος του διανύσματος του λάθους. Στη μέθοδο των ελαχίστων τετραγώνων, ψάχνουμε για το διάνυσμα του λάθους που έχει τη μικρότερη 2η νόρμα. Πιο συγκεκριμένα αναζητούμε ένα διάνυσμα $\vec{x}$ τέτοιο ώστε

$$\left\| A\overrightarrow{x^*} - \vec{b} \right\|_2 \leq \left\| A\vec{x} - \vec{b} \right\|_2, \text{ για όλα τα διανύσματα } \vec{x}$$

Από τα παραπάνω προκύπτει και το αντίστοιχο πρόβλημα quadratic – convex ελαχστιστοποίησης, least squares, του οποίου η αντικειμενική συνάρτηση είναι το άθροισμα των τετραγώνων των όρων της μορφής $a_i^T x - b_i$, όπου a_i^T είναι οι γραμμές του πίνακα A .

Έτσι το πρόβλημα ελαχιστοποίησης είναι της μορφής

$$\min \ imize \qquad f(x) = \|Ax - b\|_2^2 = \sum_{i=1}^{m} (a_i^T x - b)^2$$

και μπορεί να περιλαμβάνει η όχι περιορισμούς ισότητας, ανισότητας, ή και τα δύο. Το παραπάνω πρόβλημα μπορεί να λάβει τη μορφή

$$\min \ imize \qquad \|Ax - b\|^2 = x^T A^T Ax - 2b^T Ax + b^T b = x^T Px + Q^T x + R$$

$$\acute{o}\pi o\upsilon \quad P = A^T A, \quad Q^T = -2b^T A, \quad R = b^T b$$

Η πιο απλή μέθοδος επίλυσης του προβλήματος των ελαχίστων τετραγώνων, είναι η αναγωγή του σε επίλυση απλού γραμμικού, τετραγωνικού συστήματος εξισώσεων. Συγκεκριμένα πολλαπλασιάζουμε και τα δύο μέλη της σχέσης (1) με τον ανάστροφο του πίνακα A , οποίος ορίζεται ως A^T και έτσι προκύπτει η νέα εξίσωση

$$(A^T A)x = A^T b \Rightarrow$$

η οποία αντιπροσωπεύει πλέον ένα γραμμικό, τετραγωνικό σύστημα εξισώσεων, του οποίου η λύση είναι η $x = (A^T A)^{-1} A^T b$. Για το πρόβλημα των ελαχίστων τετραγώνων υπάρχουν πολλοί αλγόριθμοι επίλυσης (ανάμεσα σε αυτούς και αυτοί που αναλύθηκαν παραπάνω), οι οποίοι επιλύουν το πρόβλημα με μεγάλη ακρίβεια και αξιοπιστία για τα αποτελέσματα. Οι αλγόριθμοι αυτοί μπορούν να επιλύσουν προβλήματα πολλών μεταβλητών και χιλιάδων όρων σε πολύ μικρούς χρόνους. Επίσης, τα προβλήματα ελαχίστων τετραγώνων μπορούν να έχουν ή να μην έχουν περιορισμούς ισότητας και ανισότητας.

2.6 Πρόβλημα Βελτιστοποίησης Εντροπίας

2.6.1 Εντροπία – Γενικός Ορισμός

Υποθέτουμε ότι υπάρχει ένα σύνολο από διακριτά σύμβολα $(x_1, x_2, \ldots, x_n)$ με αντίστοιχες πιθανότητες p_i , όπου το άθροισμα των πιθανοτήτων είναι ίσο με 1. Η εντροπία της διακριτής κατανομής – ένα μέτρο τυχαιότητας ή μη προβλεψιμότητας της ακολουθίας των συμβόλων που την προκαλούν – ορίζεται ως

$$H = \sum_{i=1}^{n} p_i \log(1/p_i) = -\sum_{i=1}^{n} p_i \log p_i, \quad 0 \le p_i \le 1$$

όπου στον τύπο αυτόν χρησιμοποιείται ο δυαδικός λογάριθμος $(\log_2)$. Σε κάθε περίπτωση υπολογισμού των πιθανοτήτων, λαμβάνουμε υπόψιν μας τη σχέση $0 \log 0 = 0$. (Για συνεχείς κατανομές ο λογάριθμος που χρησιμοποιείται συνήθως έχει βάση το e , δηλαδή χρησιμοποιείται το $\ln$). Αξίζει να σημειωθεί ότι η εντροπία δεν εξαρτάται από τα σύμβολα, αλλά μόνο από τις πιθανότητές τους.

2.6.2 Βελτιστοποίηση Εντροπίας

Η βελτιστοποίηση της εντροπίας είναι μία μέθοδος ανάλυσης της διαθέσιμης πληροφορίας, με στόχο τον προσδιορισμό της κατανομής πυκνότητας πιθανότητας που ακολουθούν τα δεδομένα μας. Μέσω της βέλτιστης κατανομής αυτής μπορούμε να εξάγουμε χρήσιμες πληροφορίες για τα δεδομένα μας. Συνήθως, το πρόβλημα που μας ζητείται να επιλύσουμε είναι η *μεγιστοποίηση της εντροπίας*.

Θεωρούμε τη διακριτή κατανομή πιθανότητας ανάμεσα σε n αμοιβαία αποκλειόμενες προτάσεις. Η κατανομή με την περισσότερη πληροφορία εμφανίζεται όταν μία από τις προτάσεις ισχύει. Στην περίπτωση αυτή, η εντροπία της πληροφορίας θα είναι ίση με το μηδέν. Η κατανομή με τη λιγότερη πληροφορία θα εμφανιστεί όταν δεν υπάρχει λόγος να ευννοηθεί καμία πρόταση, έναντι κάποιας άλλης. Σ' αυτήν την περίπτωση, μόνη λογική κατανομή πυκνότητας πιθανότητας είναι η ομοιόμορφη κατανομή και η εντροπία της πληροφορίας θα είναι ίση με τη μέγιστη πιθανή τιμή της ($\log n$). Διαλέγοντας λοιπόν κατανομή με μικρότερη εντροπία συνεπάγεται ότι κάνουμε υποθέσεις ότι έχουμε πληροφορίες για τα δεδομένα, που πραγματικά δεν υπάρχουν. Απ' την άλλη, διαλέγοντας κατανομή με μεγαλύτερη εντροπία συνεπάγεται ότι παραβιάζουμε τους περιορισμούς που ήδη υπάρχουν για τα δεδομένα μας και που είναι επιθυμητοί. Αναζητούμε επομένως τη *μέγιστη κατανομή εντροπίας*, η οποία είναι και η μόνη λογική κατανομή και η οποία εμφανίζεται *για προτάσεις που ακολουθούν την ομοιόμορφη κατανομή,* με πυκνότητα πιθανότητας $p_i = 1/n, \quad i = 1, \ldots, n$.

Η σχέση που περιγράφει τη μεγιστοποίηση της εντροπίας είναι

$$p^* = \arg\max\{ H(p)\} \Rightarrow p^* = \arg\max\{ -\sum_{i=1}^{n} (p_i \log_2 p_i)\} .$$

Η παραπάνω συνάστηση είναι μία κοίλη (concave) συνάρτηση ως προς p. Μπορούμε να επιλύσουμε το πρόβλημα μεγιστοποίησης της αντικειμενικής συνάρτησης αυτής, ανάγοντάς το σε ένα ισοδύναμο πρόβλημα ελαχιστοποίησης (μέθοδος που αναλύθηκε στο κεφάλαιο 2.2.3.3). Συγκεκριμένα, αντί να μεγιστοποιήσουμε την εντροπία μπορούμε να ελαχιστοποιήσουμε την αντίθετη, μη αρνητική συνάρτηση της εντροπίας, όπως περιγράφεται και στην παρακάτω σχέση

$$p^* = \arg\min\{ -H(p)\} \Rightarrow p^* = \arg\min\{ \sum_{i=1}^{n} (p_i \log_2 p_i)\} .$$

Έτσι, η συνάρτηση που έχουμε να ελαχιστοποιήσουμε είναι πλέον κυρτή (convex), το πρόβλημά μας ακολουθεί την standard μορφή ενός convex προβλήματος ελαχιστοποίησης και μπορεί να επιλυθεί με τις μεθόδους που αναλύθηκαν παραπάνω. Σημπληρώνουμε ότι το πρόβλημά μας μπορεί να υπόκειται ή όχι σε περιορισμούς.

ΚΕΦΑΛΑΙΟ 3 - ΥΛΟΠΟΙΗΣΗ

3.1 Καθορισμός Δεδομένων

Τα δεδομένα που θα χρησιμοποιηθούν σε όλη την περαιτέρω ανάλυση της διαδικασίας της υλοποίησης αντιπροσωπεύονται από έναν πίνακα, διαστάσεων 34772×22, όπου οι 34772 γραμμές του αντιστοιχούν σε γονίδια, ενώ οι 22 στήλες του αντιστοιχούν σε κυτταρικές σειρές γονιδίων που περιγράφουν ανθρώπινα καρκινώματα[1]. Η τιμή κάθε γονιδίου στον πίνακα δεδομένων αντιπροσωπεύει την ένταση – φωτεινότητα του γονιδίου, έτσι όπως εμφανίζεται πάνω στην πλακέτα της μικροσυστοιχίας (microarray). Στα δεδομένα αυτά όμως υπάρχουν γονίδια, τα οποία δεν αντιπροσωπεύονταν από κάποια τιμή έντασης. Υπάρχουν δηλαδή τιμές που λείπουν – *missing values* – για κάποια γονίδια, οι οποίες περιγράφονται με την ένδειξη NaN στον πίνακα δεδομένων.

Missing Values

Το φαινόμενο των missing values μπορεί να εμφανιστεί για πολλούς λόγους. Ένας λόγος είναι η μη επαρκής ανάλυση ή ακόμα και η καταστροφή της εικόνας, στην οποία είναι αποθηκευμένη η πληροφορία του slide της μικροσυστοιχίας (πρόκειται για εικόνα μαύρου φόντου, με κουκίδες σε διάφορα χρώματα, που αντιπροσωπεύουν τα γονίδια). Ένας άλλος λόγος είναι η ύπαρξη σκόνης ή γρατζουνιών πάνω στο slide, με αποτέλεσμα την αλλοίωση της εμφάνισης κάποιων γονιδίων. Τέλος, η μη ύπαρξη τιμών για κάποια γονίδια μπορεί να οφείλεται σε "κακό" μέτρο ανάλυσης της ποιότητας της εικόνας, το οποίο διαγράφει γονίδια με μικρή φωτεινότητα (χαμηλή ποιότητα).

Μία πολλή γνωστή μέθοδος για την αντικατάσταση των τιμών που λείπουν είναι αυτή των ***k – Κοντινότερων Γειτόνων (k – Nearest Neighbors (KNN))***. Η μέθοδος αυτή στηρίζεται στην ιδέα ότι *πρότυπα που βρίσκονται κοντά στο χώρο των χαρακτηριστικών, είναι πιθανόν να ανήκουν στην ίδια τάξη*. Η παραπάνω διατύπωση περιέχει τη λέξη "*κοντά*", πράγμα που κάνει απαραίτητη την εισαγωγή μίας μετρικής. Ως μετρική μπορούν να χρησιμοποιηθούν η Ευκλείδεια (euclidean) απόσταση, η Μαχαλανόβιος (mahalanobis) απόσταση, η συσχέτιση, κλπ, με επικρατέστερη την περίπτωση της Ευκλείδειας απόστασης.

Η αναλυτική έκφραση της μεθόδου, η οποία ουσιαστικά αποτελεί μία εκτίμηση της άγνωστης τάξης του προτύπου, ακολουθεί το παρακάτω σχήμα:

- Δίνεται ένα σημείο – πρότυπο x προς ταξινόμηση.
- Επέλεξε τους k κοντινότερους γείτονες του x. Το k είναι συνήθως μονός αριθμός, χωρίς βέβαια αυτό να αποκλείει και τη χρήση ζυγών αριθμών.
- Εάν οι περισσότεροι από τους k κοντινότερους γείτονες ανήκουν σε μία συγκεκκριμένη τάξη, τότε αποδίδουμε το x στην τάξη που πλειοψηφεί. Ο αριθμός και το είδος των τάξεων εξαρτάται από τη μετρική, η οποία καθορίζει και την ακρίβεια της ταξινόμησης.

Πιο αναλυτικά, για μία συγκεκριμένη μετρική θέτουμε ένα *κατώφλι* λ, πέρα από το οποίο η μέθοδος αποφασίζει σε ποια τάξη να τοποθετήσει το πρότυπο. Η τιμή της μετρικής υπολογίζεται για κάθε γείτονα του σημείου x. Εάν είναι πάνω από το κατώφλι τότε το συγκεκριμένο δείγμα – γείτονας ταξινομείται στην αντίστοιχη και προκαθορισμένη κλάση. Εάν είναι κάτω απ' το κατώφλι ταξινομείται σε κάποια άλλη κλάση. Σημειώνουμε ότι μπορούν να υπάρχουν διάφοροι

έλεγχοι για το κατώφλι (πέρα από τους μικρότερο – μεγαλύτερο απ' το λ), πράγμα το οποίο οδηγεί και σε περισσότερες κλάσεις ταξινόμησης. Αφού λοιπόν υπολογιστούν οι μετρικές για όλους τους γείτονες, σε σχέση πάντα με το πρότυπο x, τότε αποφασίζουμε σε ποια κλάση ανήκει το x. Η κλάση στην οποία θα ανήκει το x θα είναι αυτήν στην οποία ανήκουν οι περισσότεροι από τους k κοντινοτερους γείτονες, που ταξινομήθηκαν σύμφωνα με την παραπάνω διαδικασία.
Η μέθοδος των k κοντινότερων γειτόνων είναι μία δυνατή και ταυτόχρονα ευαίσθητη μέθοδος για την εκτίμηση των missing values. Το μεγάλο της πλεονέκτημα είναι η εύκολη υλοποίησή της.

Επιστρέφοντας και πάλι στα δεδομένα του προβλήματός μας χρησιμοποιήσαμε τη μέθοδο που αναλύσαμε παραπάνω για να αντικαταστήσουμε τις ενδείξεις NaN που υπάρχουν στον πίνακα δεδομένων. Η διαδικασία είναι απλή, αφού υπάρχει έτοιμη εντολή που περιέχεται στο Toolbox του Matlab και η οποία υλοποιεί την παραπάνω μέθοδο. Η εντολή είναι η *knnimpute* και η σύνταξή της φαίνεται παρακάτω

$$knnimpute \quad (Data\ ,\ k)$$

όπου $data$ είναι το αρχείο των δεδομένων (34772×22), που περιέχουν και missing values, ενώ k είναι ο αριθμός των κοντινότερων γειτόνων που θα εξεταστούν. Το αποτέλεσμα της εντολής αυτής αποθηκεύτηκε σε έναν πίνακα $C_{all\ data}$ (34772×22). Στην περίπτωσή μας ο αριθμός των κοντινότερων γειτόνων είναι 10, ενώ ως μετρική της μεθόδου χρησιμοποιήθηκε η συσχέτιση (correlation coefficient) και όχι η Ευκλείδεια απόσταση, η οποία είναι η τυπική μετρική που χρησιμοποιείται από την εντολή. Η αλλαγή της μετρικής έγινε μέσα στο ίδιο το σώμα της συνάρτησης knnimpute. Η λέξη 'euclidean' που υπάρχει ως όρισμα στον κώδικα της συνάρτησης, αντικαταστάθηκε από τη λέξη 'correlation'.

3.2 Επεξεργασία Αρχικών Δεδομένων

3.2.1 Μείωση Διάστασης Δεδομένων (Συσχέτιση)

Έπειτα από διεξοδική ανάλυση των απαιτήσεων του προβλήματός μας σε υπολογιστικό χρόνο, σε χωρητικότητα μνήμης για το περιβάλλον του Matlab στο οποίο εργαστήκαμε (οι υπολογισμοί δεν μπορούν να εκτελεστούν, λόγω υπέρβασης μνήμης του Matlab), καταλήξαμε στην ανάγκη εύρεσης μεθόδου, για τη μείωση του μεγέθους του πίνακα των δεδομένων προς επεξεργασία. Στην ανάγκη εύρεσης μίας τέτοιας μεθόδου συνέτεινε και ο αριθμός των μαθηματικών υπολογισμών, ο οποίος πρέπει να είναι μικρός. Αυτό σημαίνει ότι το σύστημα εξισώσεων που θέλουμε να επιλύσουμε πρέπει να έχει όσο το δυνατόν μικρότερο αριθμό εξισώσεων, ώστε οι υπολογισμοί να είναι λιγότεροι και πιο εύκολοι.

Η πιο ευρέως διαδεδομένη μέθοδος για την ελαχιστοποίηση των μεταβλητών ενός προβλήματος είναι η εύρεση της συσχέτισης που υπάρχει μεταξύ τους. Για παράδειγμα, εάν μία μεταβλητή έχει υψηλή συσχέτιση με άλλες τρεις μεταβλητές του ίδιου προβλήματος, τότε μπορούμε να απορρίψουμε τις 3 μεταβλητές και να κρατήσουμε τη μία, η οποία θα τις αντιπροσωπεύει όλες και η επεξεργασία της θα οδηγήσει σε ίδια αποτελέσματα, με το να επεξεργαζούμασταν και τις υπόλοιπες 3.
Συγκεκριμένα, η ***συσχέτιση (correlation coefficient)*** είναι ένας αριθμός, ο οποίος περιγράφει το βαθμό και την κατεύθυνση της γραμμικής σχέσης μεταξύ δύο (ή και περισσότερων) τυχαίων

μεταβλητών. Μπορούμε να πούμε ότι αναφέρεται στην ύπαρξη εξάρτησης ή ανεξαρτησίας των μεταβλητών. Ο στατιστικός τύπος που χρησιμοποιείται ευρέως και περιγράφει τη συσχέτιση δύο μεταβλητών είναι:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y}$$

όπου: $\mu_X = E(X)$, η μέση τιμή του X

$\mu_Y = E(Y)$, η μέση τιμή του Y

σ_X, η τυπική απόκλιση του X $(\sigma_X^2 = E(X^2) - E^2(X))$

σ_Y, η τυπική απόκλιση του Y $(\sigma_Y^2 = E(Y^2) - E^2(Y))$

$\mathrm{cov}(X,Y)$, είναι το γινόμενο των αποκλίσεων 2 τυχαίων μεταβλητών γύρω από το μέσο (covariance)

Η συσχέτιση ορίζεται μόνο όταν και οι δύο τυπικές αποκλίσεις είναι πεπερασμένες και διάφορες του μηδενός. Η μεγαλύτερη κατ' απόλυτο τιμή που μπορεί να πάρει η συσχέτιση είναι το 1. Μπορεί να ισούται με 1, δηλαδή να έχουμε αυξανόμενη γραμμική συσχέτιση, να ισούται με -1, δηλαδή να έχουμε ελαττωμένη γραμμική συσχέτιση και να πάρει οποιαδήποτε τιμή, ανάμεσα σ' αυτό το διάστημα, υποδεικνύοντας τη γραμμική εξάρτηση μεταξύ των μεταβλητών. Όσο πιο κοντά είναι ο συντελεστής στο 1 ή το -1, τοσό ισχυρότερη είναι η συσχέτιση μεταξύ των μεταβλητών. Εάν οι μεταβλητές είναι εξαρτημένες, τότε η συσχέτιση είναι 0, αλλά η αντίστροφη πρόταση δεν ισχύει, αφού η συσχέτιση εμφανίζει μόνο τις γραμμικές ανεξαρτησίες μεταξύ των μεταβλητών.

Το μέγεθος της συσχέτισης δύο μεταβλητών περιγράφεται αναλυτικά στον ***Πίνακα 3.1*** που φαίνεται παρακάτω. Αξίζει να σημειωθεί βέβαια ότι τα κριτήρια που περιγράφονται στον πίνακα αυτόν μπορεί κατά μία έννοια να θεωρηθούν αυθαίρετα και έτσι δεν πρέπει να παρατηρούνται τόσο αυστηρά. Αυτό συμβαίνει γιατί η ερμηνεία της συσχέτισης εξαρτάται από το περιεχόμενο και τους σκοπούς για τους οποίους υπολογίζεται.

Correlation	**Negative**	**Positive**
Small	−0.29 to −0.10	0.10 to 0.29
Medium	−0.49 to −0.30	0.30 to 0.49
Large	−1.00 to −0.50	0.50 to 1.00

Πίνακας 3.1. Μέγεθος Συσχέτισης Μεταβλητών

Έτσι, για το πρόβλημά μας χρησιμοποιήσαμε τη συσχέτιση για να ομαδοποιήσουμε τα γονίδια που έχουν ίδια συσχέτιση. Δημιουργήσαμε δηλαδή ομάδες γονιδίων με ίδια τιμή συσχέτισης για τα γονίδια κάθε ομάδας, τις οποίες τελικά αντικαταστήσαμε με ένα αντιπροσωπευτικό γονίδιο απ' την ομάδα, το οποίο θα συνεισφέρει το ίδιο στους υπολογισμούς μας, όπως και τα υπόλοιπα γονίδια της ομάδας που απορρίφθηκαν. Με τη διαδικασία αυτή κατορθώσαμε να μειώσουμε τον αριθμό των γονιδίων, άρα και των γραμμών του πίνακα δεδομένων.
Για να μειώσουμε το χρόνο και την πολυπλοκότητα των υπολογισμών που θα προκύψουν από την αναλυτική εφαρμογή της σχέσης που υπολογίζει τη συσχέτιση, χρησιμοποιήσαμε την έτοιμη

συνάρτηση *corrcoef* που παρέχει το περιβάλλον του Matlab. Η σύνταξη της εντολής – συνάρτησης φαίνεται παρακάτω:

$$R = corrcoef \quad (x, y)$$

όπου x, y είναι οι 2 τυχαίες μεταβλητές και R είναι η μεταβλητή στην οποία αποθηκεύεται η τιμή της συσχέτισης μεταξύ των δύο μεταβλητών.

Συγκεκριμένα, στην υλοποίησή μας δημιουργήσαμε έναν πίνακα r, μεγέθους $1 \times NumberOfGe \quad nes = 1 \times 34772$, ο οποίος αποκαλείται *πίνακας συσχέτισης (Correlation Matrix)* και στον οποίο αποθηκεύουμε την τιμή της συσχέτισης ενός γονιδίου με όλα τα υπόλοιπα γονίδια, η οποία υπολογίζεται άμεσα από την παραπάνω εντολή. Εδώ πρέπει να σημειώσουμε ότι η συσχέτιση παίρνει και θετικές και αρνητικές τιμές. Εμείς στον πίνακα r αποθηκεύουμε κάθε φορά το μέτρο της συσχέτισης, δηλαδή μόνο θετικές τιμές, χρησιμοποιώντας την έτοιμη εντολή του Matlab, *abs*, χωρίς αυτό να επηρεάζει τα αποτελέσματά μας.
Η εντολή αυτή εκτελείται μέσα σε έναν επαναληπτικό βρόγχο (for loop), και υπολογίζει συσχετίσεις *κάθε γονιδίου ξεχωριστά* με όλα τα υπόλοιπα. Οι υπολογισμοί των συσχετίσεων, που πραγματοποιούνται για κάθε γονίδιο υπολογίζονται από τη σχέση $\frac{N * (N - 1)}{2} = \frac{34772 * (34772 - 1)}{2} = 604528606$, όπου N ο αριθμός των γονιδίων. Πρόκειται για μεγάλο αριθμό υπολογισμών, γι' αυτό και οι απαιτήσεις του κώδικα σε χρόνο εκτέλεσης ήταν μεγάλες ($1^{1/2}$ μέρα). Επομένως, το τελικό αποτέλεσμα που θα έπρεπε να λάβουμε είναι ένας τεράστιος πίνακας με πολλές τιμές συσχετίσεων για κάθε ένα γονίδιο. Όμως, το περιβάλλον του Matlab αδυνατεί να αποθηκεύσει πίνακες με τόσο μεγάλο αριθμό στοιχείων σε ένα αρχείο. Έτσι, θα ήμασταν αναγκασμένοι να εκτελούμε τον κώδικα αυτό για κάθε ένα γονίδιο ξεχωριστά και να αποθηκεύουμε τα αποτελέσματα αυτά σε διαφορετικό αρχείο κάθε φορά. Εξαιτίας, όμως, του μεγάλου χρόνου εκτέλεσης του κώδικα, αυτό θα ήταν χρονοβόρο. Για το λόγο αυτό, δεσμεύσαμε μνήμη και για έναν άλλο πίνακα db, μεγέθους 1×34772, ο οποίος ουσιαστικά αποτελεί και την έξοδο του προγράμματος αυτού. Οι τιμές του πίνακα αυτού προκύπτουν ως εξής: όταν η απόλυτη τιμή της συσχέτισης ενός γονιδίου με τα υπόλοιπα είναι μικρότερη από το *κατώφλι 0.5* (η τιμή για το κατώφλι προέκυψε σύμφωνα με τον Πίνακα 1, ο οποίος υποδεικνύει ότι ισχυρή συσχέτιση παρουσιάζεται για το διάστημα – κατ' απόλυτες τιμές – $[0.5, 1]$), τότε αντικαθιστούμε αυτή την τιμή με 0. Εάν είναι μεγαλύτερη, τότε την αντικαθιστούμε με 1. Η τιμή 0 συνεπάγεται ότι δεν υπάρχει καμία συσχέτιση μεταξύ αυτού του γονιδίου και των υπολοίπων, και επομένως το γονίδιο δε πρέπει να απορριφθεί, ενώ η τιμή 1 υποδηλώνει συσχέτιση και το γονίδιο μπορεί να απορριφθεί.
Για περαιτέρω διευκόλυνσή μας αποθηκεύσαμε το περιεχόμενο του πίνακα db σε ένα αρχείο Excel. Το αρχείο αυτό περιλαμβάνει ουσιαστικά τιμές 0 και 1. Όμως, εμείς με όλη την παραπάνω διαδικασία επιθυμούμε να δημιουργήσουμε έναν νέο, μικρότερο πίνακα, ο οποίος να περιλαμβάνει γονίδια και κυτταρικές σειρές (με τις αντίστοιχες τιμές έντασης που τα αντιπροσωπεύουν) και όχι τιμές 0 και 1. Έτσι δημιουργούμε ένα νέο κομμάτι κώδικα, στο οποίο λαμβάνουμε ως είσοδο το αρχείο Excel με τις τιμές 0 και 1 και βγάζουμε ως έξοδο τον τελικό πίνακα δεδομένων. Η τελική μορφή του πίνακα δεδομένων προκύπτει ως εξής: σε έναν επαναληπτικό βρόγχο και για 34772 επαναλήψεις, όπου συναντάμε 0, αντικαθιστούμε με το αντίστοιχο γονίδιο – σειρά του αρχικού πίνακα δεδομένων, ενώ όπου συναντάμε 1, απορρίπτουμε το αντίστοιχο γονίδιο – σειρά που υπάρχει στα δεδομένα μας.

Μετά από όλους αυτούς τους υπολογισμούς καταλήγουμε να μειώσουμε τον αριθμό των γραμμών – γονιδίων του αρχικού πίνακα, *από 34772 σε 27*[2]. Έτσι τα δεδομένα μας πλέον αντιπροσωπεύονται από έναν πίνακα $C_{smaller}$, μεγέθους 27×22 .

3.2.2 Κανονικοποίηση Δεδομένων

Η ανάγκη κανονικοποίησης των τιμών έντασης των γονιδίων προέκυψε και πάλι από την απαίτηση για μικρότερη πολυπλοκότητα υπολογισμών. Οι τιμές έντασης, έτσι όπως υπάρχουν στον αρχικό πίνακα δεδομένων, αλλά και στο νέο μικρότερο πίνακα $C_{smaller}$ είναι της τάξης των $3^{\omega\nu}$ ή και $4^{\omega\nu}$ ψηφίων. Εμείς επιθυμούμε αριθμούς μικρότερης τάξης, για να μπορούμε να τους μελετήσουμε με μεγαλύτερη ευκολία και οι υπολογισμοί μας να γίνονται γρηγορότερα.
Για το λόγο αυτό αποφασίσαμε να κανονικοποιήσουμε τις τιμές έντασης, χρησιμοποιώντας ως μέσο κανονικοποίησης το δυαδικό λογάριθμο $(\log_2)$. Σε δύο εμφωλευμένους επαναληπτικούς βρόγχους (for loops) και για αριθμό επαναλήψεων από 1 ως 27 και από 1 ως 22 αντίστοιχα αντικαταστήσαμε την τιμή έντασης κάθε γονιδίου, σε κάθε κυτταρική σειρά με το δυαδικό λογάριθμο της τιμής αυτής. Για να είναι απόλυτα ορθή η παραπάνω διαδικασία, μέσα στους επαναληπτικούς βρόγχους εισάγαμε έναν έλεγχο, σύμφωνα με τον οποίο εάν καποια τιμή έντασης ενός γονιδίου είναι 0, τότε αυτή να παραμείνει 0. Αυτό ισχύει γιατί ο λογάριθμος του μηδενός δεν ορίζεται. Σε κάθε άλλη περίπτωση, απλά να αντικαταστήσει την τιμή της έντασης, όπως αναλύσαμε παραπάνω. Έτσι, πλέον έχουμε τον τελικό, κανονικοποιημένο πίνακα δεδομένων C_{final} , ο οποίος μπορεί να χρησιμοποιηθεί ως είσοδος για την υλοποίηση του βασικού αλγορίθμου (Gradient Descent), της εργασίας αυτής. Η οποιαδήποτε περαιτέρω ανάλυση της διαδικασίας υλοποίησης θα αναφέρεται στο νέο πίνακα $C_{final}\,(27 \times 22)$!

3.3 Υλοποίηση Αλγορίθμου Gradient Descent

3.3.1 Υλοποίηση Μεθόδου Ελαχίστων Τετραγώνων

Όπως αναφέρθηκε και στη θεωρία, ο αλγόριθμος Gradient Descent χρησιμοποιείται για την επίλυση προβλημάτων ελαχιστοποίησης χωρίς περιορισμούς. Η μέθοδος που θέλουμε να χρησιμοποιήσουμε για να βρούμε τις ζητούμενες αναλογίες πρόσμειξης των κυτταρικών σειρών, με στόχο τη δημιουργία ενός πρότυπου μείγματος είναι αυτή των Ελαχίστων Τετραγώνων. Το πρόβλημα των Ελαχίστων Τετραγώνων, έτσι όπως ορίζεται για την περίπτωσή μας, υπόκειται σε περιορισμούς για τις τιμές των αναλογιών. Γι' αυτό θα πρέπει να μετατρέψουμε το πρόβλημα ελαχιστοποίησης με περιορισμούς, σε πρόβλημα χωρίς περιορισμούς, το οποίο τελικά θα επιλύεται με τον αλγόριθμο Gradient Descent. Αυτό, όπως θα δούμε και παρακάτω γίνεται με το να δημιουργήσουμε μία νέα αντικειμενική συνάρτηση, η οποία θα περιλαμβάνει και τους περιορισμούς στους οποίους υπόκεινται οι αναλογίες. Οι περιορισμοί θα εισαχθούν στη νέα συνάρτηση, με τη βοήθεια των πολλαπλασιαστών Lagrange.
Όλη η παρακάτω ανάλυση λοιπόν περιλαμβάνει τον ορισμό της αρχικής αντικειμενικής συνάρτησης για το πρόβλημα Ελαχίστων Τετραγώνων, αλλά και της νέας συνάρτησης (με τους περιορισμούς), τη δημιουργία διαφορετικών διανυσμάτων αρχικοποίησης του πρότυπου μείγματος, που χρησιμοποιούνται για τον ορισμό της αντικειμενικής συνάρτησης, τα βασικά βήματα εκτέλεσης του αλγορίθμου και τέλος, την κανονικοποίηση των αποτελεσμάτων, ώστε να υπακούν σε όλους τους περιορισμούς του προβλήματος.

❖ **Ορισμός Αντικειμενικής Συνάρτησης**

Η συνάρτηση που αντιπροσωπεύει το πρόβλημα των Ελαχίστων Τετραγώνων, προσαρμοσμένο πάντα στις απαιτήσεις και τα δεδομένα του προβλήματός μας έχει την παρακάτω μορφή

$$f_{LS} = \frac{1}{27}\left\| C_{final} * \underline{a} - \underline{A_{initial}} \right\|_2^2 .$$

Ο όρος 1/27 χρησιμοποιείται για να κανονικοποιήσει τις τιμές της συνάρτησης. Ο πίνακας C_{final} (27×22) είναι ο πίνακας των δεδομένων του προβλήματός μας, a (22×1) είναι το διάνυσμα των αναλογιών και $A_{initial}$ (27×1) είναι το διάνυσμα τιμών του target distribution. Η ποσότητα $C_{final} * \underline{a} - \underline{A_{initial}}$ αντιστοιχεί ουσιαστικά σε ένα διάνυσμα διαστάσεων 27×1. Εκτελώντας τις αλγεβρικές πράξεις που περιγράφουν το διάνυσμα αυτό, καταλήγουμε στην τελική μορφή του διανύσματος, η οποία φαίνεται παρακάτω:

$$C_{final} * \underline{a} - \underline{A_{initial}} =$$

$$\begin{bmatrix} C_{1,1} & C_{1,2} & \dots & C_{1,22} \\ C_{2,1} & C_{2,2} & \dots & C_{2,22} \\ \vdots & \vdots & \ddots & \vdots \\ C_{27,1} & C_{27,2} & \dots & C_{27,22} \end{bmatrix} * \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{22} \end{bmatrix} - \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_{27} \end{bmatrix} = \underbrace{\begin{bmatrix} C_{1,1} * a_1 + C_{1,2} * a_2 + \dots + C_{1,22} * a_{22} - A_1 \\ C_{2,1} * a_1 + C_{2,2} * a_2 + \dots + C_{2,22} * a_{22} - A_2 \\ \\ C_{27,1} * a_1 + C_{27,2} * a_2 + \dots + C_{27,22} * a_{22} - A_{27} \end{bmatrix}}_{27 \times 1}$$

Σύμφωνα με την αριθμητική άλγεβρα, η νόρμα διανύσματος είναι ίση με την τετραγωνική ρίζα του αθροίσματος των τετραγώνων των στοιχείων του διανύσματος. Συνδυάζοντας τον ορισμό της νόρμας με την αναλυτική μορφή του διανύσματος που παρουσιάστηκε παραπάνω προκύπτει η τελική αναλυτική μορφή της συνάρτησης των Ελαχίστων Τετραγώνων, η οποία είναι

$$f_{LS} = \frac{1}{27} * [(C_{1,1} * a_1 + \dots + C_{1,22} * a_{22} - A_1)^2 + (C_{2,1} * a_1 + \dots + C_{2,22} * a_{22} - A_2)^2 + \dots$$
$$\dots + (C_{27,1} * a_1 + \dots + C_{27,22} * a_{22} - A_{27})^2]$$

Η συνάρτηση αυτή αντιστοιχεί στο πρόβλημα Ελαχίστων Τετραγώνων, χωρίς όμως την ύπαρξη περιορισμών. Το πρόβλημά μας όμως περιλαμβάνει τους εξής περιορισμούς για τις αναλογίες πρόσμειξης κυτταρικών σειρών:

- Οι αναλογίες να είναι θετικοί αριθμοί και όχι πάρα πολύ μικροί.
- Η μέγιστη τιμή που μπορούν να πάρουν να είναι το 1 (ή αλλιώς το 100%).
- Το άθροισμά τους να είναι ίσο με τη μονάδα (δεν μπορεί να είναι μικρότερο ή μεγαλύτερο από 1).

Έτσι είναι αναγκαίο να εισάγουμε στην ήδη υπάρχουσα συνάρτηση που αναλύσαμε παραπάνω, τους ανωτέρω περιορισμούς. Στηριζόμενοι στη μέθοδο των πολλαπλασιαστών Lagrange, προσθέτουμε στη συνάρτηση τον όρο $\|\underline{a}\|_2^2$, ο οποίος αποτελεί τη μαθηματική περιγραφή των ανωτέρω περιορισμών, πολλαπλασιασμένο

πάντα με έναν θετικό αριθμό λ (πολλαπλασιαστής Lagrange). Δημιουργούμε δηλαδή μία νέα αντικειμενική συνάρτηση, η οποία αντιστοιχεί στη Lagrangian του προβλήματος μας. Ο νέος όρος αναλύεται σύμφωνα με τον ορισμό της νόρμας διανύσματος, όπως και παραπάνω. Η τελική μορφή της αντικειμενικής συνάρτησης του προβλήματος Ελαχίστων Τετραγώνων, η οποία θα ελαχιστοποιηθεί με τη βοήθεια του αλγορίθμου Gradient Descent είναι η

$$\boxed{f = \frac{1}{27} * \left\| C_{final} * \underline{a} - \underline{A_{initial}} \right\|_2^2 + \lambda * \left\| \underline{a} \right\|_2^2}$$

Η αναλυτική έκφραση της αντικειμενικής συνάρτησης αυτής είναι αυτή που φαίνεται παρακάτω

$$f = \frac{1}{27} * [(C_{1,1} * a_1 + \ldots + C_{1,22} * a_{22} - A_1)^2 + (C_{2,1} * a_1 + \ldots + C_{2,22} * a_{22} - A_2)^2 + \ldots$$
$$\ldots + (C_{27,1} * a_1 + \ldots + C_{27,22} * a_{22} - A_{27})^2] + \lambda * (a_1^2 + a_2^2 + \ldots + a_{22}^2)$$

- **Διανύσματα Αρχικοποίησης Target Distribution** $(A_{initial})$

Οι τιμές των στοιχείων του διανύσματος A που αντιπροσωπεύει το πρότυπο μείγμα δεν είναι γνωστές, πράγμα το οποίο δυσχεραίνει την επίλυση του προβλήματός μας. Χωρίς τον πίνακα αυτό δεν μπορούμε να ελαχιστοποιήσουμε την παραπάνω αντικειμενική συνάρτηση, μιας και οι άγνωστοι στο πρόβλημά μας θα ήταν πλέον δύο. Έτσι είναι αναγκαίο να αρχικοποιήσουμε τις τιμές των στοιχείων του πρότυπου μείγματος και να δημιουργήσουμε έτσι μία σειρά μειγμάτων $A_{initial}$, από τα οποία θα προκύψουν και οι αντίστοιχες αναλογίες. Χρησιμοποιήσαμε 3 διαφορετικές αρχικοποιήσεις, για να ελέγξουμε το κατά πόσο η επιλογή της αρχικοποίησης επηρεάζει τα αποτελέσματά μας.

1) **Target Distribution για Αρχικοποίηση** $A_{initial} = \frac{1}{27}$

Το μείγμα αυτό δημιουργήθηκε στηριζόμενο στην ανάγκη ύπαρξης ομοιομορφίας. Άλλωστε, στόχος της εργασίας αυτής είναι η εύρεση αναλογιών πρόσμειξης κυτταρικών σειρών, ώστε να δημιουργήσουν ένα ομοιόμορφο πρότυπο μείγμα. Σύμφωνα, λοιπόν, με τη θεωρία πιθανοτήτων, μία κατανομή καλείται ομοιόμορφη εάν όλα τα στοιχεία της είναι ίσα μεταξύ τους και παίρνουν την τιμή $1/n$, όπου n είναι ο αριθμός των στοιχείων που αποτελούν την κατανομή. Χαρακτηριστικό της ομοιόμορφης κατανομής είναι ότι η διασπορά των στοιχείων της κατανομής, γύρω από τη μέση τιμή είναι ίση με το 0. Έτσι, για να είναι το μείγμα μας ομοιόμορφο θα πρέπει και τα 27 στοιχεία του να είναι ίσα με $1/27 = 0.037$, όπως φαίνεται και παρακάτω

$$A_{initial} = [\underbrace{0{,}037 \quad 0{,}037 \quad \ldots\ldots \quad 0{,}037 \quad 0{,}037}_{27\ \text{φορές}}]$$

Η αντίστοιχη αντικειμενική συνάρτηση που προκύπτει για τη συγκεκριμένη αρχικοποίηση είναι

$$f = \frac{1}{27} * [(C_{1,1} * a_1 + \ldots + C_{1,22} * a_{22} - \frac{1}{27})^2 + (C_{2,1} * a_1 + \ldots + C_{2,22} * a_{22} - \frac{1}{27})^2 + \ldots + (C_{27,1} * a_1 + \ldots + C_{27,22} * a_{22} - \frac{1}{27})^2] + \lambda * (a_1^2 + a_2^2 + \ldots + a_{22}^2)$$

2) Target Distribution για Αρχικοποίηση $A_{initial} = 9{,}1194$

Η αρχικοποίηση αυτή προέκυψε στηριζόμενη στην εξής διαδικασία:

- Θεωρούμε ότι όλες οι κυτταρικές σειρές συμμετέχουν στο πρότυπο μείγμα κατά τον ίδιο τρόπο. Δηλαδή αναμειγνύουμε τις κυτταρικές σειρές με αναλογίες ίσες με $1/22$, όπου 22 είναι ο αριθμός των αναλογιών.
- Δημιουργείται έτσι ένα αντίστοιχο μείγμα, με μέση τιμή 9,119447391.

Για να φτιάξουμε λοιπόν την τελική αρχικοποίηση θέσαμε και τα 27 στοιχεία του μείγματος ίσα με την παραπάνω τιμή. Θα μπορούσαμε να χρησιμοποιήσουμε ως αρχικοποίηση τις τιμές μείγματος που προέκυψαν από τη μείξη των κυτταρικών σειρών με ίσες αναλογίες ($1/22$). Όμως, το μείγμα αυτό δεν παρουσιάζει την απαιτούμενη ομοιομορφία, μιας και η διασπορά του είναι αρκετά μεγαλύτερη από το 0. Αντίθετα, το μείγμα με όλα στοιχεία ίσα με 9,1194, έχει διασπορά 0 και έτσι ανταποκρίνεται πλήρως στην ανάγκη ομοιομορφίας του πρότυπου μείγματος. Η νέα αρχικοποίηση είναι φαίνεται παρακάτω

$$A_{initial} = [\underbrace{9{,}1194 \quad 9{,}1194 \quad \ldots\ldots \quad 9{,}1194 \quad 9{,}1194}_{27\ \phi o \rho \acute{\varepsilon} \varsigma}]$$

και η αντίστοιχη αντικειμενική συνάρτηση για την περίπτωση αυτή είναι η

$$f = \frac{1}{27} * [(C_{1,1} * a_1 + \ldots + C_{1,22} * a_{22} - 9{,}1194)^2 + (C_{2,1} * a_1 + \ldots + C_{2,22} * a_{22} - 9{,}1194)^2 + \ldots + (C_{27,1} * a_1 + \ldots + C_{27,22} * a_{22} - 9{,}1194)^2] + \lambda * (a_1^2 + a_2^2 + \ldots + a_{22}^2)$$

3) Target Distribution για Αρχικοποίηση $A_{initial} \geq 350$

Για την τελευταία αρχικοποίηση του target distribution στηριχθήκαμε στην ανάγκη ύπαρξης στοιχείων – γονιδίων (ουσιαστικά τα στοιχεία που αποτελούν το μείγμα δεν είναι τίποτε άλλο από γονίδια), των οποίων η τιμή έντασης στο μείγμα να είναι πάνω από ένα συγκεκριμένο κατώφλι. Επιθυμούμε δηλαδή να δημιουργήσουμε ένα μείγμα, τα γονίδια του οποίου θα είναι διακριτά επάνω στο slide της μικροσυστοιχίας (η έντασή τους θα έχει τέτοια τιμή, ώστε η φωτεινότητα των spots στο slide της μικροσυστοιχίας να είναι μεγάλη και έτσι, τα γονίδια να είναι διακριτά). Για το λόγο αυτό επιλέχθηκε τιμή κατωφλίου έντασης γονιδίων το 350, το οποίο σε κανονικοποιημένη μορφή με χρήση του δυαδικού λογάριθμου αντιστοιχεί στην τιμή $\log_2(350) = 8{,}451211$.

Χρησιμοποιήσαμε και πάλι ως βάση το μείγμα που προκύπτει από την πρόσμειξη των κυτταρικών σειρών σε αναλογίες ίσες με $1/22$. Συγκρίνουμε την τιμή των στοιχείων του με την τιμή του κατωφλίου. Εάν το στοιχείο – γονίδιο έχει τιμή μεγαλύτερη απ' αυτή του κατωφλίου, τότε παραμένει ως έχει στο μείγμα. Εάν η τιμή του είναι μικρότερη, τότε θα αντικατασταθεί με τη μέση τιμή του μείγματος – βάσης (μείγμα που αντιστοιχεί σε αναλογίες ίσες με $1/22$), η οποία ισούται με $9{,}1194$. Έτσι, με την παραπάνω διαδικασία, η οποία περιγράφεται στο ***Σχήμα 3.1*** που ακολουθεί δημιουργήσαμε και την τρίτη κατά σειρά αρχικοποίηση.

Μείγμα - Βάση (α(i) = 1/22)	***Σύγκριση με Κατώφλι***	***Ενέργεια***	***Τελικό Μείγμα***
7,988624281	< 8,451211	Replace	9,119449371
8,139298738	< 8,451211	Replace	9,119449371
7,903450597	< 8,451211	Replace	9,119449371
9,984325188	> 8,451211	Keep	9,984325188
7,905493838	< 8,451211	Replace	9,119449371
7,785679747	< 8,451211	Replace	9,119449371
11,83245661	> 8,451211	Keep	11,83245661
10,77224276	> 8,451211	Keep	10,77224276
8,051658307	< 8,451211	Replace	9,119449371
9,973924616	> 8,451211	Keep	9,973924616
8,904981832	> 8,451211	Keep	8,904981832
8,978116193	> 8,451211	Keep	8,978116193
8,008152273	< 8,451211	Replace	9,119449371
8,122559065	< 8,451211	Replace	9,119449371
8,947933543	> 8,451211	Keep	8,947933543
8,137045091	< 8,451211	Replace	9,119449371
8,436335567	< 8,451211	Replace	9,119449371
8,183792525	< 8,451211	Replace	9,119449371
8,687322606	> 8,451211	Keep	8,687322606
12,88046964	> 8,451211	Keep	12,88046964
12,41099677	> 8,451211	Keep	12,41099677
7,629169044	< 8,451211	Replace	9,119449371
8,653002396	> 8,451211	Keep	8,653002396
7,934621517	< 8,451211	Replace	9,119449371
11,77680091	> 8,451211	Keep	11,77680091
9,478883553	> 8,451211	Keep	9,478883553
8,717795805	> 8,451211	Keep	8,717795805

Σχήμα 3.1. Διάγραμμα περιγραφής της διαδικασίας παραγωγής της 3° κατά σειρά αρχικοποίησης του target distribution. Στην πρώτη στήλη φαίνονται οι τιμές του μείγματος – βάσης, και στην τελευταία οι τελικές τιμές αρχικοποίησης.

Θα πρέπει να σημειώσουμε εδώ, ότι παρ' όλο που τα στοιχεία για την αρχικοποίηση αυτή δεν είναι όλα ίσα μεταξύ τους, το μείγμα εμφανίζει σχετική ομοιομορφία, μιας και η διασπορά του είναι ίση με $1{,}4154$, τιμή σχετικά μικρή και κοντά στο 0. Άρα μπορούμε να συμπεράνουμε ότι και αυτό το μείγμα ικανοποιεί την απαίτηση της ομοιομορφίας.

- **Βήματα Εκτέλεσης Αλγορίθμου Gradient Descent**

Ο αλγόριθμος Gradient Descent υλοποιήθηκε στο προγραμματιστικό περιβάλλον του Matlab και σύμφωνα με την αντίστοιχη θεωρία τα βήματα της υλοποίησής μας είναι τα εξής:

a) Ορίζουμε τα *ζητούμενα* του προβλήματός μας, δηλαδή τους αγνώστους που υπάρχουν στην αντικειμενική συνάρτηση και ως προς τους οποίους επιθυμούμε να ελαχιστοποιήσουμε. Οι άγνωστοι, όπως φαίνεται και από όλη την παραπάνω ανάλυση είναι οι 22 αναλογίες πρόσμειξης κυτταρικών σειρών a_i. Για να μπορέσουμε να δηλώσουμε τις αναλογίες στην υλοποίησή μας ως μεταβλητές των οποίων οι τιμές είναι άγνωστες, χρησιμοποιήσαμε την εντολή του Matlab *syms*. Η εντολή αυτή δημιουργεί συμβολικά αντικείμενα (στην περίπτωσή μας έχουμε συμβολικούς αριθμούς). Το όρισμα εισόδου της είναι συνήθως πολλά string, τα οποία πρέπει να ξεκινούν απαραίτητα με ένα γράμμα και μπορεί να περιλαμβάνουν αλφαριθμητικούς χαρακτήρες. Το αποτέλεσμα της εντολής είναι ένα σύνολο συμβολικών αριθμών ή μεταβλητών (κάθε αριθμός για κάθε string). Οι συμβολικοί αυτοί αριθμοί μπορεί να είναι πραγματικοί, μιγαδικοί, μόνο θετικοί, γι' αυτό και η εντολή παίρνει ως όρισμα τις αντίστοιχες επιλογές (real, unreal, positive). Η σύνταξη της εντολής φαίνεται παρακάτω

$$syms \quad \arg_1 \ \arg_2 \ \ldots\ldots \ real$$

Στην υλοποίησή μας τα ορίσματα της εντολής είναι οι 22 αναλογίες a_i, οι οποίες επιθυμούμε να είναι πραγματικοί αριθμοί και έτσι καταλήγουμε στην εξής σύνταξη

$$syms \quad a_1 \quad a_2 \quad a_3 \ \ldots\ldots \ a_{20} \quad a_{21} \quad a_{22} \quad real$$

Οι αναλογίες στην αντικειμενική συνάρτηση εμφανίζονται με τη μορφή διανύσματος. Έτσι δημιουργούμε ένα αντίστοιχο διάνυσμα 22×1, από συμβολικούς αριθμούς, όπως φαίνεται και παρακάτω

$$a = [a_1 \quad a_2 \quad \ldots \quad a_{22}]$$

b) Εισάγουμε στην υλοποίησή μας τα *δεδομένα* του προβλήματος, δηλαδή τον κανονικοποιημένο πίνακα των 27 γονιδίων και των 22 κυτταρικών σειρών και τα τρία διαφορετικά διανύσματα αρχικοποίησης μείγματος. Το κάθε διάνυσμα αρχικοποίησης του target distribution προκύπτει γράφοντας αναλυτικά τις τιμές του, όπως αυτές προέκυψαν από την ανωτέρω ανάλυση. Τέλος ορίζουμε την τιμή του πολλαπλασιαστή Lagrange λ, μιας και όπως αναφέρθηκε παραπάνω, η αντικειμενική συνάρτηση που θα χρησιμοποιηθεί θα περιλαμβάνει και τους απαραίτητους περιορισμούς (συγκεκριμένα θα περιέχει και τον όρο $\lambda * \|\underline{a}\|_2^2$). Για το λ έχουμε δύο διαφορετικές τιμές που επιθυμούμε να εξετάσουμε. Μία μικρή και μία μεγάλη τιμή, ώστε να καλύψουμε όλο το φάσμα των περιπτώσεων. Εάν

το λ είναι μικρό, τότε η επίδραση του περιορισμού στην αρχική αντικειμενική συνάρτηση είναι μικρή, ενώ αν είναι μεγάλο, η επίδραση του περιορισμού είναι μεγάλη. Ο τρόπος με τον οποίο υπολογίζεται η τιμή του πολλαπλασιαστή λ περιγράφεται παρακάτω:
Πραγματοποιώντας τις αλγεβρικές πράξεις μεταξύ διανυσμάτων – πινάκων και στηριζόμενοι σε αλγεβρικές ιδιότητες υπολογίζουμε την αναλυτική διανυσματική έκφραση της αντικειμενικής μας συνάρτησης

$$\left\| C_{final} * a - A_{initial} \right\|_2^2 + \lambda * \left\| a \right\|_2^2 = (C_{final} * a - A_{initial})^T * (C_{final} * a - A_{initial}) + \lambda * a^T a =$$

$$= C_{final}{}^T a^T C_{final}\ a - C_{final}{}^T a^T A_{initial} - C_{final}\ a\ A_{initial}{}^T + A_{initial} A_{initial}{}^T + \lambda a^T a$$

Υπολογίζουμε την παράγωγο της παραπάνω έκφρασης ως προς τον άγνωστο του προβλήματός μας, δηλαδή ως προς το διάνυσμα των αναλογιών a. Έπειτα μηδενίζουμε την παράγωγο, για να βρούμε τη σχέση από την οποία προκύπτει το διάνυσμα των αναλογιών. Η μαθηματική έκφραση των παραπάνω είναι η εξής

$$\nabla_a \{ \left\| C_{final} * a - A_{initial} \right\|_2^2 + \lambda \left\| a \right\|_2^2 \} = 0 \Rightarrow 2C_{final}{}^T C_{final}\ a - 2C_{final}{}^T a + 2\lambda a = 0 \Rightarrow$$

$$2 * (C_{final}{}^T C_{final}\ a - C_{final}{}^T A_{initial} + \lambda a) = 0 \Rightarrow a * (C_{final}{}^T C_{final} + \lambda I) - C_{final}{}^T A_{initial} = 0$$

$$\Rightarrow a = [C_{final}{}^T C_{final} + \lambda I]^{-1} C_{final}{}^T A_{initial}$$

Δημιουργούμε λοιπόν στο Matlab την παραπάνω σχέση, όπου C_{final} είναι ο πίνακας των δεδομένων μας, I είναι ο μοναδιαίος πίνακας, μεγέθους 22×22 και $A_{initial}$ το διάνυσμα που αντιπροσωπεύει το target distribution. Μέσα σε έναν επαναληπτικό βρόγχο (for loop) και για τιμές του λ από 0 ως 2, με βήμα επανάληψης 0.001 υπολογίζουμε τις τιμές του διανύσματος των αναλογιών, για κάθε διαφορετική τιμή του λ. Ταυτόχρονα, μέσα στον ίδιο βρόγχο υπολογίζουμε την τιμή της αντικειμενικής μας συνάρτησης f, για κάθε διάνυσμα αναλογιών που προκύπτει (και συνεπώς για κάθε διαφορετική τιμή του λ). Η βέλτιστη τιμή του πολλαπλασιαστή λ είναι εκείνη για την οποία η συνάρτηση εμφανίζει το ολικό της ελάχιστο. Από τους υπολογισμούς μας προέκυψε ότι η συνάρτησή μας εμφανίζει ελάχιστο, όταν η τιμή του λ είναι ίση με $1{,}967$. Αυτή είναι και η μία από τις δύο τιμές που θα χρησιμοποιηθούν για τον πολλαπλασιαστή λ, στον αλγόριθμο Gradient Descent. Η άλλη τιμή, η οποία είναι μεγάλη είναι το 10 και επιλέχθηκε τυχαία.
Αξίζει να σημειώσουμε εδώ ότι ο παραπάνω επαναληπτικός βρόγχος εκτελέστηκε και για τις 3 διαφορετικές αρχικοποιήσεις του διανύσματος $A_{initial}$. Όμως και για τις 3 αρχικοποιήσεις, η μικρή του λ που προέκυψε ήταν ίδια και ίση με $1{,}967$. Επομένως, *η αρχικοποίηση του target distribution δεν επηρεάζει την επιλογή της τιμής του πολλαπλασιαστή λ.*

c) Δημιουργούμε την αντικειμενική συνάρτηση (με τους περιορισμούς), σύμφωνα με τον τύπο που παρουσιάστηκε παραπάνω. Η αντικειμενική συνάρτηση έχει και αυτή συμβολική μορφή, αφού περιλαμβάνει το διάνυσμα συμβόλων των αναλογιών πρόσμειξης. Για τη δημιουργία της τελικής μορφής της συνάρτησης f εκτελούνται αριθμητικές πράξεις μεταξύ των πινάκων δεδομένων και αρχικοποίησης μείγματος και του διανύσματος συμβολικών αριθμών, που αντιστοιχεί στις αναλογίες. Οι υπολογισμοί αυτοί όμως, εξαιτίας της ύπαρξης των συμβολικών αριθμών οδηγούν σε μία περίπλοκη μαθηματική έκφραση για τη συνάρτηση f. Οι συντελεστές δηλαδή των a_i είναι της μορφής A/B, όπου A, B είναι πολύ μεγάλοι αριθμοί. Για να καταλήξουμε λοιπόν σε ένα αποτέλεσμα για την αντικειμενική συνάρτηση πιο ευκρινές και πιο απλοποιήμενο χρησιμοποιήσαμε την εντολή του Matlab, *vpa*. Η εντολή αυτή χρησιμοποιεί ακρίβεια αριθμητικής μεταβλητής για να μετατρέψει το όρισμά της σε δεκαδικό αριθμό, με ακρίβεια d δεκαδικών ψηφίων. Ο αριθμός ακριβείας των δεκαδικών ψηφίων είναι ήδη προεπιλεγμένος κατά την εφαρμογή της εντολής, μπορεί όμως και ο χρήστης να θέσει δική του τιμή για την ακρίβεια των ψηφίων. Η σύνταξη της εντολής για τη δική μας υλοποίηση φαίνεται παρακάτω

$$vpa\ (f)$$

όπου f είναι η έκφραση της αντικειμενικής μας συνάρτησης (με περιορισμούς) με τους συντελεστές των a_i να έχουν τη μορφή A/B (περίπλοκη μορφή).

d) Στο βήμα αυτό επιλέγουμε την τιμή του βήματος του αλγορίθμου, καθώς και το starting point του. Το βήμα του αλγορίθμου θα πρέπει να μην είναι πολύ μικρό, ώστε η σύγκλιση να μην είναι αργή, αλλά ούτε και πολύ μεγάλο, γιατί αυτό θα μας οδηγήσει σε πολύ κακά αποτελέσματα. Λαμβάνοντας υπόψην τα παραπάνω και έχοντας ως δεδομένο ότι η αντικειμενική συνάρτηση πρέπει να ελαχιστοποιείται (συνεπώς, η τιμή του κριτηρίου σύγκλισης να μειώνεται), δοκιμάσαμε αρκετές τιμές για το βήμα στο διάστημα $[0.00001,\ 0.01]$. Έπειτα από έλεγχο για το αν η τιμή της συνάρτησης ελαχιστοποιείται και για το αν το κριτήριο σύγκλισης μειώνεται, για κάθε διαφορετικό βήμα αλγορίθμου, η τιμή στην οποία καταλήξαμε είναι $step = 0.0001$.

Τέλος, το σημείο έναρξης του αλγορίθμου στην περίπτωσή μας είναι ένα διάνυσμα a^0, μεγέθους 22×1, το οποίο περιλαμβάνει τις αρχικές τιμές για τις αναλογίες a_i. Οι τιμές του διανύσματος αυτού προκύπτουν από την απαίτηση το άθροισμα των αναλογιών να είναι ίσο με 1. Μία λοιπόν δυνατή αρχικοποίηση που θα μπορούσε να ικανοποιεί αυτήν την απαίτηση είναι αυτή που περιγράφεται από τις σχέσεις $a_{1\ldots21} = 1/22 = 0{,}045$, $a_{22} = 1 - (a_1 + a_2 + \ldots + a_{21}) = 0{,}055$, οι οποίες τελικά δημιουργούν το παρακάτω διάνυσμα

$$a^0 = [\underbrace{0{,}045 \quad 0{,}045 \quad \ldots \quad 0{,}045 \quad 0{,}045}_{1\ldots21} \quad 0{,}055\,]$$

e) Έπειτα υπολογίζουμε τις *μερικές παραγώγους* της αντικειμενικής συνάρτησης ως προς κάθε άγνωστη αναλογία a_i, η μορφή των οποίων είναι και αυτή συμβολική. Εκφράζοντας με μαθηματικούς συμβολισμούς τις μερικές παραγώγους θέλουμε ουσιαστικά να υπολογίσουμε τα $\frac{\partial f}{\partial a_1}$, $\frac{\partial f}{\partial a_2}$, … , $\frac{\partial f}{\partial a_n}$. Ο υπολογισμός των μερικών παραγώγων μπορεί να γίνει με τη χρήση δύο διαφορετικών εντολών του Matlab:

- Εντολή *jacobian*, της οποίας η σύνταξη είναι $grad = jacobian\ (f, [x_1, \ldots, x_n])$, όπου f είναι η συνάρτηση προς παραγώγιση και $x_1, x_2, \ldots, x_n$ οι μεταβλητές ως προς τις οποίες παραγωγίζουμε.
- Εντολή *diff*, της οποίας η σύνταξη είναι $grad = diff\ (f, x_1)$, $grad = diff\ (f, x_2)$, … , $grad = diff\ (f, x_n)$, όπου f είναι η συνάρτηση προς παραγώγιση και $x_1, x_2, \ldots, x_n$ οι μεταβλητές ως προς τις οποίες παραγωγίζουμε.

Χρησιμοποιήσαμε και τις δύο αυτές εντολές για να ελέγξουμε ότι καταλήγουν στην ίδια σωστή αναλυτική έκφραση για κάθε μερική παράγωγο. Και οι δύο καταλήγουν τελικά στην ίδια αναλυτική έκφραση. Γι' αυτό καταλήξαμε στο να χρησιμοποιήσουμε την εντολή diff, η οποία στο πρόβλημά μας συντάσσεται ως εξής

$$fa_1 = diff\ (f, a_1), \quad fa_2 = diff\ (f, a_2), \ldots, \quad fa_{22} = diff\ (f, a_{22})$$

Ο γενικός τύπος που υπολογίζει ***την παράγωγο*** της αντικειμενικής μας συνάρτησης και ο οποίος αναλύθηκε και παραπάνω, για την ανάγκη εύρεσης του πολλαπλασιαστή λ είναι

$$\nabla_a f = \nabla_a \left(\frac{1}{27} * \left\| C_{final}\ \underline{a} - A_{initial} \right\|^2 + \lambda * \left\| \underline{a} \right\|^2\right) =$$

$$= \frac{1}{27} * \nabla_a \left(\left\| C_{final}\ \underline{a} - A_{initial} \right\|^2\right) + \lambda * \nabla_a \left(\left\| \underline{a} \right\|^2\right) =$$

$$= \frac{1}{27} * \nabla_a \{(C_{final} * a - A_{initial})^T * (C_{final} * a - A_{initial})^T\} + \lambda * \nabla_a (a^T a) =$$

$$= \frac{2}{27} * (C_{final} * a - A_{initial}) * C_{final}^T + 2\lambda * a$$

ενώ, ο γενικός τύπος που περιγράφει ***τις μερικές παραγώγους*** και ο οποίος ταυτίζεται με το αποτέλεσμα που προκύπτει από την εντολή diff, φαίνεται παρακάτω

$$\frac{\partial f}{\partial a_{(i)}} = \frac{2}{27} * (C_{final}\ a - A_{initial}) * C_{(j(1,\ldots,27),\ i)} + 2\lambda * a_{(i)}$$

Ο τρόπος που συνδέει την παράγωγο της αντικειμενικής μας συνάρτησης με τις μερικές παραγώγους είναι ουσιαστικά το ότι η παράγωγος της συνάρτησης είναι ένα διάνυσμα του οποίου τα στοιχεία είναι οι μερικές παράγωγοι ως προς κάθε αναλογία. Το παραπάνω περιγράφεται μαθηματικά ως εξής

$$\nabla_a f = \begin{bmatrix} \frac{\partial f}{\partial a_1} \\ \frac{\partial f}{\partial a_2} \\ \vdots \\ \frac{\partial f}{\partial a_{22}} \end{bmatrix}$$

Η έκφραση που προκύπτει για κάθε μερική παράγωγο περιλαμβάνει τα σύμβολα a_i. Για να μπορέσουν να χρησιμοποιηθούν οι μερικές παράγωγοι στον αλγόριθμο όμως, θα πρέπει να γνωρίζουμε την ακριβή αριθμητική τους τιμή. Για να μπορέσει να γίνει αυτό θα πρέπει σε κάθε αναλογία να αναθέτουμε κάθε φορά και μία αριθμητική τιμή, ώστε να μην υπάρχουν άγνωστοι στην έκφραση των παραγώγων. Το διάνυσμα των αριθμητικων τιμων των αναλογιών προκύπτει κάθε φορά από τον ίδιο τον αλγόριθμο. Στην πρώτη επανάληψη του αλγορίθμου, για τον υπολογισμό της τιμής των μερικών παραγώγων χρησιμοποιείται το διάνυσμα αρχικοποίησης a^0, που ορίστηκε παραπάνω. Στις επόμενες επαναλήψεις, για τον υπολογισμό τους χρησιμοποιείται το εκάστοτε διάνυσμα τιμών των αναλογιών που προκύπτει απ' τον αλγόριθμο. Για να μπορέσουμε να πραγματοποιήσουμε τις αναθέσεις τιμών για τις αναλογίες, χρησιμοποιούμε την εντολή του Matlab *subs*. Η εντολή αυτή αντικαθιστά τα σύμβολα των συμβολικών εκφράσεων ή πινάκων με αριθμούς, ενώ η σύνταξή της για το πρόβλημά μας είναι

$$res_1 = subs\ (fa_1, a, a_{from_Gradient_Descent}),\ res_2 = subs\ (fa_2, a, a_{from_Gradient_Descent}),$$
$$\ldots,\ res_{22} = subs\ (fa_{22}, a, a_{from_Gradient_Descent})\ .$$

f) Ορίζουμε το κριτήριο σύγκλισης, το οποίο σύμφωνα με τη θεωρία προκύπτει από τον τύπο

$$stop_criterion = \sqrt{res_1^2 + res_2^{22} + \ldots + res_{22}^2}\ .$$

Οι μεταβλητές res_i αντιστοιχούν στην αριθμητική τιμή κάθε μερικής παραγώγου. Η τιμές των παραγώγων αλλάζουν σε κάθε επανάληψη, επομένως θα αλλάζει και η τιμή του κριτηρίου σύγκλισης. Στόχος του αλγορίθμου είναι το κριτήριο σύγκλισης να μειώνεται σε κάθε επανάληψη, μέχρι να φτάσει στην ελάχιστη δυνατή τιμή του. Η ελάχιστη τιμή πρέπει να είναι ένας θετικός και μικρός αριθμός, η επιλογή του οποίου εξαρτάται το είδος του προβλήματος. Για το πρόβλημα των ελαχίστων τετραγώνων που εξετάζουμε, η ελάχιστη τιμή για το κριτήριο σύγκλισης είναι το $0{,}01$, τιμή που προέκυψε έπειτα από την εκτέλεση του κώδικα και για τις 3 περιπτώσεις αρχικοποίησης μείγματος και για 2 τιμές του πολλαπλασιαστή λ.

g) Τέλος εφαρμόζουμε τον επαναληπτικό τύπο που ουσιαστικά υλοποιεί και τον αλγόριθμο Gradient Descent, για τις 3 διαφορετικές περιπτώσεις αρχικοποίησης του target distribution και για τις 2 διαφορετικές τιμές του πολλαπλασιαστή λ. Από τον τύπο αυτό προκύπτουν κάθε φορά οι νέες τιμές των αναλογιών a_i. Σύμφωνα με τον τύπο αυτό, οι νέες αναλογίες προκύπτουν από τις αναλογίες τις προηγούμενης επανάληψης εάν αφαιρέσουμε την τιμή της παραγώγου της προηγούμενης επανάληψης, πολλαπλασιασμένης με το βήμα του αλγορίθμου. Η μαθηματική αναπαράσταση της παραπάνω έκφρασης είναι η

$$a_i^{k+1} = a_i^k - step * res_i^k .$$

Κατά την πάροδο των επαναλήψεων, οι τιμές των αναλογιών μειώνονται, ώσπου μετά από κάποια επανάληψη προκύπτουν και αρνητικές τιμές. Όμως, ο περιορισμός στον οποίο πρέπει να υπόκεινται οι αναλογίες είναι να είναι θετικοί αριθμοί. Γι' αυτό σε κάθε επανάληψη πραγματοποιήσαμε τον παρακάτω έλεγχο, ο οποίος στην υλοποίησή μας πραγματοποιήθηκε μέσα σε ένα βρόγχο ελέγχου (if – else): εάν κάποια από τις αναλογίες (ή και όλες οι αναλογίες) της προηγούμενης επανάληψης έχει αρνητική τιμή, τότε η τιμή της στην επόμενη επανάληψη αντικαθίσταται με την τιμή $0{,}00001$. Εάν οι αναλογίες έχουν μόνο θετικές τιμές, τότε οι τιμές τους παραμένουν ως έχουν και οι νέες αναλογίες προκύπτουν από την εφαρμογή του παραπάνω επαναληπτικού τύπου. Η επιλογή του συγκεκριμένου αριθμού για την αντικατάσταση των αρνητικών τιμών είναι η καλύτερη δυνατή. Αυτό ισχύει γιατί τυχόν μεγαλύτερες τιμές επηρεάζουν τη σύγκλιση του αλγορίθμου, οδηγώντας σε σκαμπανεβάσματα των τιμών του κριτηρίου σύγκλισης, αλλά ακόμη και σε άνοδο των τιμών του κριτηρίου, αντί για μείωση. Επίσης, τυχόν μικρότερες τιμές θα κατέληγαν σε αναλογίες πολύ μικρές (πολύ κοντά στο μηδέν), το οποίο σημαίνει ότι θα είναι και ασήμαντες.

Τέλος, ο αριθμός των επαναλήψεων εκτέλεσης του αλγορίθμου είναι διαφορετικός για κάθε περίπτωση, γιατί η επιλογή κάθε φορά της αρχικοποίησης του target distribution και του πολλαπλασιαστή λ επηρεάζει το χρόνο σύγκλισης. Δηλαδή, άλλοτε πλησιάζουμε το ελάχιστο κριτήριο σύγκλισης σε μικρό αριθμό επαναλήψεων και άλλοτε σε μεγαλύτερο. Σε όλες πάντως τις περιπτώσεις φροντίζουμε ώστε η σύγκλιση να πραγματοποιείται σε έναν εύλογο αριθμό επαναλήψεων. Στον ***Πίνακα 3.2*** παραθέτουμε τον αριθμό των επαναλήψεων που χρειάστηκε ο αλγόριθμος για να φτάσει την ελάχιστη τιμή του κριτηρίου σύγκλισης, για κάθε μία από τις 6 περιπτώσεις (3 διαφορετικές αρχικοποιήσεις και 2 πολλαπλασιαστές για κάθε αρχικοποίηση).

	$A_{initial} = 1/27$		$A_{initial} = 9{,}1194$		$A_{initial} \geq 350$	
	λ = 1,967	**λ = 10**	**λ = 1,967**	**λ = 10**	**λ = 1,967**	**λ = 10**
Επαναλήψεις	1161	1125	8,676	3382	8105	1748

Πίνακας 3.2. Αριθμός Επαναλήψεων Εκτέλεσης Αλγορίθμου (Least Squares)

- **Κανονικοποίηση Αποτελεσμάτων Αναλογιών Πρόσμειξης**
  Τα αποτελέσματα που προέκυψαν από τον αλγόριθμο Gradient Descent (δηλαδή 22 αναλογίες για κάθε μία από τις 6 περιπτώσεις) ικανοποιούν όλους τους απαιτούμενους περιορισμούς, εκτός απο τον περιορισμό το άθροισμα των αναλογιών πρόσμειξης να είναι ακριβώς ίσο με 1. Συγκεκριμένα, οι αναλογίες που προκύπτουν απ' την εφαρμογή του αλγορίθμου έχουν θετικό άθροισμα, το οποίο όμως δεν είναι ίσο με 1, αλλά λίγο μικρότερο. Για να μπορέσουμε να δημιουργήσουμε άθροισμα ίσο με 1 χρειάστηκε να κανονικοποιήσουμε τα αποτελέσματα των αναλογιών πρόσμειξης. Υπολογίζουμε το άθροισμα των 22 αναλογιών που προκύπτουν από τον αλγόριθμο. Έπειτα διαιρούμε κάθε μία από τις 22 αναλογίες με το άθροισμα αυτό. Οι 22 τιμές που προκύπτουν είναι οι τελικές αναλογίες πρόσμειξης κυτταρικών σειρών, των οποίων το άθροισμα είναι θετικό και ίσο με 1.

3.3.2 Υλοποίηση Μεθόδου Ελαχιστοποίησης Εντροπίας

Για να μπορέσουμε να έχουμε συγκρίσιμα αποτελέσματα με αυτά της μεθόδου των Ελαχίστων Τετραγώνων, χρησιμοποιήσαμε και μία δεύτερη μέθοδο, τη μέθοδο Ελαχιστοποίησης της Εντροπίας. Στην πραγματικότητα, *για να μπορέσουμε να βρούμε τις βέλτιστες αναλογίες πρόσμειξης θα πρέπει να μεγιστοποιήσουμε την εντροπία του πρότυπου μείγματος που θέλουμε να δημιουργήσουμε.* Όμως η standard μορφή ενός convex προβλήματος βελτιστοποίησης (όπως είναι αυτό της εντροπίας) αντιστοιχεί σε ελαχιστοποίηση κάποιας συνάρτησης. Έτσι, σύμφωνα και με την αντίστοιχη θεωρία που αναπτύχθηκε στο κεφάλαιο 2, αντί να μεγιστοποιήσουμε την εντροπία, θα ελαχιστοποιήσουμε το αντίθετο της εντροπίας (αντί για $\max \{-\sum x \log x\}$ θα έχουμε $\min \sum x \log x$). Για δική μας διευκόλυνση και για την αποφυγή παρερμηνειών το πρόβλημα ελαχιστοποίησης του αντίθετου της εντροπίας θα το αποκαλούμε απλά πρόβλημα Ελαχιστοποίησης Εντροπίας.

Το πρόβλημα των Ελαχίστων Τετραγώνων είναι ένα τετραγωνικό πρόβλημα, ενώ η χρήση των μερικών παραγώγων της αντικειμενικής του συνάρτησης, το ανάγει σε ένα γραμμικό πρόβλημα, εύκολο να επιλυθεί. Αντίθετα, το πρόβλημα ελαχιστοποίησης της εντροπίας είναι λογαριθμικό, ενώ και με τη χρήση των παραγώγων του ανάγεται και πάλι σε λογαριθμικό πρόβλημα. Παρ' όλα αυτά όμως και το πρόβλημα ελαχιστοποίησης της εντροπίας μπορεί να επιλυθεί με τον αλγόριθμο Gradient Descent.

Το πρόβλημα ελαχιστοποίησης της εντροπίας υπόκειται στους ίδιους περιορισμούς όσον αφορά τις αναλογίες πρόσμειξης κυτταρικών σειρών, όπως και το πρόβλημα Ελαχίστων Τετραγώνων. Έτσι θα χρειαστεί να μετατραπεί και αυτό σε πρόβλημα χωρίς περιορισμούς, το οποίο τελικά θα επιλύεται με τον αλγόριθμο Gradient Descent. Θα δημιουργήσουμε και πάλι μία νέα αντικειμενική συνάρτηση που θα περιλαμβάνει τους περιορισμούς, πάντα με τη βοήθεια των πολλαπλασιαστών Lagrange. Η διαδικασία που θα ακολουθηθεί είναι παρόμοια με αυτή που ακολουθήθηκε για τη μέθοδο των Ελαχίστων Τετραγώνων. Οι μικρές διαφορές εντοπίζονται στον ορισμό της αντικειμενικής συνάρτησης (διαφορετική συνάρτηση), στον υπολογισμό των μερικών παραγώγων (αναλυτικότερη διαδικασία), στις τιμές του αρχικού διανύσματος των αναλογιών (σημείο έναρξης) και στις τιμές του βήματος και του κριτηρίου σύγκλισης του αλγορίθμου.

❖ **Ορισμός Αντικειμενικής Συνάρτησης**

Η συνάρτηση που αντιστοιχεί στο πρόβλημα Ελαχιστοποίησης της Εντροπίας και που είναι πάντοτε προσαρμοσμένη στα δεδομένα του προβλήματός μας, είναι η

$$f_{ENTROPY} = (C_{final} * \underline{a}) * \log(C_{final} * \underline{a}) = \sum_{j=1}^{27} [(\sum_{i=1}^{22} C_{j,i} * a_i) * \log(\sum_{i=1}^{22} C_{j,i} * a_i)] .$$

Ο πίνακας C_{final} (27×22) είναι ο πίνακας των δεδομένων του προβλήματός μας, a (22×1) είναι το διάνυσμα των αναλογιών. Η ποσότητα $C_{final} * \underline{a}$ αντιστοιχεί ουσιαστικά σε ένα διάνυσμα διαστάσεων 27×1. Εκτελώντας τις αλγεβρικές πράξεις που περιγράφουν το διάνυσμα αυτό, καταλήγουμε στην τελική μορφή του διανύσματος, η οποία φαίνεται παρακάτω:

$$C_{final} * \underline{a} - \underline{A_{initial}} =$$

$$\begin{bmatrix} C_{1,1} & C_{1,2} & \dots & C_{1,22} \\ C_{2,1} & C_{2,2} & \dots & C_{2,22} \\ \vdots & \vdots & \ddots & \vdots \\ C_{27,1} & C_{27,2} & \dots & C_{27,22} \end{bmatrix} * \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{22} \end{bmatrix} = \underbrace{\begin{bmatrix} C_{1,1} * a_1 + C_{1,2} * a_2 + \dots + C_{1,22} * a_{22} \\ C_{2,1} * a_1 + C_{2,2} * a_2 + \dots + C_{2,22} * a_{22} \\ \\ C_{27,1} * a_1 + C_{27,2} * a_2 + \dots + C_{27,22} * a_{22} \end{bmatrix}}_{27 \times 1}$$

Η αναλυτική λοιπόν μορφή της συνάρτησης που υπολογίζει την εντροπία είναι αυτή που φαίνεται παρακάτω

$$\begin{aligned} f = [&(C_{1,1} * a_1 + \dots + C_{1,22} * a_{22}) * \log(C_{1,1} * a_1 + \dots + C_{1,22} * a_{22}) + \\ &(C_{2,1} * a_1 + \dots + C_{2,22} * a_{22}) * \log(C_{2,1} * a_1 + \dots + C_{2,22} * a_{22}) + \\ &\dots + (C_{27,1} * a_1 + \dots + C_{27,22} * a_{22}) * \log(C_{27,1} * a_1 + \dots + C_{27,22} * a_{22})] \end{aligned}$$

Η συνάρτηση αυτή αντιστοιχεί στο πρόβλημα Ελαχιστοποίησης της Εντροπίας, χωρίς όμως την ύπαρξη περιορισμών. Το πρόβλημά μας όμως περιλαμβάνει τους εξής περιορισμούς για τις αναλογίες πρόσμειξης κυτταρικών σειρών:

- Οι αναλογίες να είναι θετικοί αριθμοί και όχι πάρα πολύ μικροί.
- Η μέγιστη τιμή που μπορούν να πάρουν να είναι το 1 (ή αλλιώς το 100%).
- Το άθροισμά τους να είναι ίσο με τη μονάδα (δεν μπορεί να είναι μικρότερο ή μεγαλύτερο από 1).

Έτσι στηριζόμενοι και πάλι στη μέθοδο των πολλαπλασιαστών Lagrange, προσθέτουμε στη συνάρτηση τον όρο $\|\underline{a}\|_2^2$, ο οποίος αποτελεί τη μαθηματική περιγραφή των ανωτέρω περιορισμών, πολλαπλασιασμένο πάντα με έναν θετικό αριθμό λ (πολλαπλασιαστής Lagrange). Ο νέος όρος αναλύεται σύμφωνα με τον ορισμό της νόρμας διανύσματος, όπως και παραπάνω.

Η τελική μορφή της αντικειμενικής συνάρτησης του προβλήματος Ελαχιστοποίησης της Εντροπίας, η οποία θα ελαχιστοποιηθεί με τη βοήθεια του αλγορίθμου Gradient Descent είναι η

$$\boxed{f = (C_{final} * \underline{a}) * \log(C_{final} * \underline{a}) + \lambda * \|\underline{a}\|_2^2}$$

ενώ, η αναλυτική μορφή της αντικειμενικής αυτής συνάρτησης είναι αυτή που φαίνεται παρακάτω

$$\begin{aligned} f = {} & (C_{1,1} * a_1 + \ldots + C_{1,22} * a_{22}) * \log(C_{1,1} * a_1 + \ldots + C_{1,22} * a_{22}) + \\ & (C_{2,1} * a_1 + \ldots + C_{2,22} * a_{22}) * \log(C_{2,1} * a_1 + \ldots + C_{2,22} * a_{22}) + \ldots \\ & (C_{27,1} * a_1 + \ldots + C_{27,22} * a_{22}) * \log(C_{27,1} * a_1 + \ldots + C_{27,22} * a_{22}) + \\ & \lambda * (a_1^2 + a_2^2 + \ldots + a_{22}^2) \end{aligned}$$

- **Βήματα Εκτέλεσης Αλγορίθμου Gradient Descent**

Ο αλγόριθμος Gradient Descent υλοποιήθηκε και πάλι στο προγραμματιστικό περιβάλλον του Matlab. Τα βήματα που είναι ακριβώς ίδια με αυτά του προβλήματος των Ελαχίστων Τετραγώνων θα παρουσιαστούν σε λιγότερο αναλυτική μορφή. Όσα βήματα είναι διαφορετικά θα περιγραφούν εκτενώς στην παρακάτω ανάλυση.

a) Ορίζουμε τα *ζητούμενα* του προβλήματός μας, δηλαδή τους αγνώστους που υπάρχουν στην αντικειμενική συνάρτηση και ως προς τους οποίους επιθυμούμε να ελαχιστοποιήσουμε. Οι άγνωστοι είναι οι 22 αναλογίες πρόσμειξης κυτταρικών σειρών a_i και τους αποθηκεύουμε σε ένα διάνυσμα συμβολικών αριθμών, όπως και στο πρόβλημα των Ελαχίστων Τετραγώνων. Ο ορισμός των αναλογιών, καθώς και το διάνυσμα 22×1, στο οποίο τις αποθηκεύουμε φαίνονται παρακάτω

$$\begin{aligned} & syms \quad a_1 \quad a_2 \quad a_3 \ldots\ldots a_{20} \quad a_{21} \quad a_{22} \quad real \\ & a = [a_1 \quad a_2 \quad \ldots \quad a_{22}] \end{aligned}$$

b) Εισάγουμε στην υλοποίησή μας τα *δεδομένα* του προβλήματος, δηλαδή τον κανονικοποιημένο πίνακα των 27 γονιδίων και των 22 κυτταρικών σειρών και την τιμή του πολλαπλασιαστή Lagrange λ, μιας και όπως αναφέρθηκε παραπάνω, η αντικειμενική συνάρτηση που θα χρησιμοποιηθεί θα περιλαμβάνει και τους απαραίτητους περιορισμούς (συγκεκριμένα θα περιέχει και τον όρο $\lambda * \|\underline{a}\|_2^2$). Για το λ έχουμε και πάλι δύο διαφορετικές τιμές που επιθυμούμε να εξετάσουμε, το $1{,}967$ και το 10. Αυτή τη φορά δε χρησιμοποιήθηκε κάποια μαθηματική ανάλυση για την επιλογή των τιμών του πολλαπλασιαστή λ. Επιλέχθηκαν έτσι ώστε να είναι ίδιες με αυτές που χρησιμοποιήθηκαν στο πρόβλημα των Ελαχίστων Τετραγώνων, για να μπορεί να υπάρχει αντιστοιχία στη σύγκριση των δύο προβλημάτων.

c) Δημιουργούμε την αντικειμενική συνάρτηση (με τους περιορισμούς), σύμφωνα με τον τύπο που παρουσιάστηκε παραπάνω. Η αντικειμενική συνάρτηση έχει και αυτή συμβολική μορφή, αφού περιλαμβάνει το διάνυσμα συμβόλων των αναλογιών πρόσμειξης και αποτελείται και πάλι από συντελεστές για τις

αναλογίες που είναι μεγάλοι αριθμοί, της μορφής A/B. Για να καταλήξουμε επομένως σε ένα πιο ευκρινές και πιο απλοποιήμενο αποτέλεσμα χρησιμοποιήσαμε όπως και στην άλλη μέθοδο, την εντολή του Matlab, *vpa*. Η σύνταξη της εντολής και για αυτή την υλοποίηση είναι

$$vpa\ (f)$$

όπου f είναι η έκφραση της αντικειμενικής μας συνάρτησης με τους συντελεστές των a_i, της μορφής A/B (περίπλοκη μορφή).

d) Στο βήμα αυτό επιλέγουμε την τιμή του βήματος του αλγορίθμου, καθώς και το starting point του. Το βήμα του αλγορίθμου θα πρέπει να μην είναι πολύ μικρό, ώστε η σύγκλιση να μην είναι αργή, αλλά ούτε και πολύ μεγάλο, γιατί αυτό θα μας οδηγήσει σε πολύ κακά αποτελέσματα. Λαμβάνοντας υπόψην τα παραπάνω και έχοντας ως δεδομένο ότι η αντικειμενική συνάρτηση πρέπει να ελαχιστοποιείται, δοκιμάσαμε αρκετές τιμές για το βήμα στο διάστημα $[0.00001\ ,\ \ 0.01]$. Έπειτα από έλεγχο για το αν η τιμή της συνάρτησης ελαχιστοποιείται και για το αν το κριτήριο σύγκλισης μειώνεται, για κάθε διαφορετικό βήμα αλγορίθμου, η τιμή στην οποία καταλήξαμε είναι $step = 0.00001$. Όπως μπορούμε να παρατηρήσουμε το βήμα αυτό είναι μικρότερο από το αντίστοιχο για τη μέθοδο των Ελαχίστων Τετραγώνων, και αυτή είναι και η πρώτη βασική διαφορά των δύο υλοποιήσεων του αλγορίθμου.

Τέλος, το σημείο έναρξης του αλγορίθμου είναι και πάλι ένα διάνυσμα a^0, μεγέθους 22×1, το οποίο περιλαμβάνει τις αρχικές τιμές για τις αναλογίες a_i. Οι τιμές του διανύσματος αυτού προκύπτουν από την απαίτηση το άθροισμα των τετραγώνων των αναλογιών να είναι ίσο με 1. Πρόκειται για μία ακόμα διαφορά μεταξύ των μεθόδων Ελαχίστων Τετραγώνων και Ελαχιστοποίησης της Εντροπίας. Το αντίστοιχο αρχικό διάνυσμα που χρησιμοποιήθηκε στη μέθοδο των Ελαχίστων Τετραγώνων δεν ήταν δυνατό να χρησιμοποιηθεί για την Ελαχιστοποίηση της Εντροπίας. Οι τιμές για τις αναλογίες ήταν πολύ μικρές και οδηγούνταν σε αρνητικές σε πολύ μικρό αριθμό επαναλήψεων, αλλά το κυριότερο είναι το ότι η χρήση τους οδηγούσε σε αύξηση του κριτηρίου σύγκλισης και της τιμής της συνάρτησης. Μία, λοιπόν, δυνατή αρχικοποίηση που θα μπορούσε να ικανοποιεί αυτήν την απαίτηση είναι αυτή που περιγράφεται από το παρακάτω διάνυσμα

$$a^0 = [0{,}2 \quad 0{,}3 \quad 0{,}4 \quad 0{,}1 \quad 0{,}25 \quad 0{,}2 \quad 0{,}17 \quad 0{,}15 \quad 0{,}21 \quad 0{,}3 \quad 0{,}1 \quad 0{,}22$$
$$0{,}13 \quad 0{,}29 \quad 0{,}35 \quad 0{,}23 \quad 0{,}12 \quad 0{,}11 \quad 0{,}09 \quad 0{,}05 \quad 0{,}1 \quad 0{,}16]$$

e) Έπειτα υπολογίζουμε τις μερικές παραγώγους της αντικειμενικής συνάρτησης ως προς κάθε άγνωστη αναλογία a_i, η μορφή των οποίων είναι και αυτή συμβολική. Εκφράζοντας με μαθηματικούς συμβολισμούς τις μερικές παραγώγους θέλουμε ουσιαστικά να υπολογίσουμε τα $\frac{\partial f}{\partial a_1}$, $\frac{\partial f}{\partial a_2}$, ... , $\frac{\partial f}{\partial a_n}$. Οι μερικές παράγωγοι στην περίπτωση της εντροπίας δε θα υπολογιστούν με τη βοήθεια κάποιας

εντολής, αλλά θα γραφτεί ο αναλυτικός μαθηματικός τους τύπος, ο οποίος προέκυψε μετά από υπολογισμό των παραγώγων ‘στο χέρι’.

Ο γενικός τύπος που αντιπροσωπεύει ***την παράγωγο*** της αντικειμενικής συνάρτησης του προβλήματός μας είναι

$$\nabla_a f = \nabla_a \{(C_{final} * \underline{a}) * \log(C_{final} * \underline{a}) + \lambda \|\underline{a}\|^2\} =$$

$$= \nabla_a \{(C_{final} * \underline{a}) * \log(C_{final} * \underline{a})\} + \lambda * \nabla_a \{\|a\|^2\} =$$

$$= \{\nabla_a ((C_{final} * \underline{a}))\} * \log(C_{final} * \underline{a}) + (C_{final} * \underline{a}) * \{\nabla_a (\log(C_{final} * \underline{a}))\} + \lambda * \nabla_a (a^T a) =$$

$$= C_{final}^T * \log(C_{final} * a) + (C_{final} * a) * \frac{1}{(C_{final} * a) * \ln 2} * \{\nabla_a ((C_{final} * \underline{a}))\} + 2\lambda a =$$

$$= C_{final}^T * \log(C_{final} * a) + \frac{1}{\ln 2} * I * C_{final}^T + 2\lambda a$$

ενώ, η μαθηματική έκφραση για καθεμία από τις 22 ***μερικές παραγώγους*** είναι η

$$\frac{\partial f}{\partial a_{(i)}} = C_{(j(1,\ldots,27),\, i)} * \log(C_{final} * a_{(i)}) + \frac{C_{(j(1,\ldots,27),\, i)}}{\ln 2} + 2\lambda * a_{(i)}$$

Ο τρόπος που συνδέει την παράγωγο της αντικειμενικής μας συνάρτησης με τις μερικές παραγώγους είναι ουσιαστικά το ότι η παράγωγος της συνάρτησης είναι ένα διάνυσμα του οποίου τα στοιχεία είναι οι μερικές παράγωγοι ως προς κάθε αναλογία. Το παραπάνω περιγράφεται μαθηματικά ως εξής

$$\nabla_a f = \begin{bmatrix} \frac{\partial f}{\partial a_1} \\ \frac{\partial f}{\partial a_2} \\ \vdots \\ \frac{\partial f}{\partial a_{22}} \end{bmatrix}$$

Για τον υπολογισμό λοιπόν των μερικών παραγώγων στον κώδικά μας (μιας και θα γίνει αναλυτικά και όχι με τη χρήση εντολής), δημιουργείται η σχέση που περιγράφηκε παραπάνω μέσα σε έναν επαναληπτικό βρόγχο (for loop), για αριθμό επαναλήψεων από 1 ως 22, μιας και οι μερικές παράγωγοι είναι 22, όπως προαναφέραμε. Για να υπολογίσουμε την αριθμητική τιμή της κάθε μερικής παραγώγου χρησιμοποιούμε την εντολή *subs*, η οποία αντικαθιστά τα σύμβολα του διανύσματος των αναλογιών με συγκεκριμένες αριθμητικές τιμές που προκύπτουν από τον ίδιο τον αλγόριθμο. Στην πρώτη επανάληψη του αλγορίθμου, το διάνυσμα συμβόλων των αναλογιών αντικαθίσταται από το διάνυσμα αρχικοποίησης a^0, που ορίστηκε για την περίπτωση της εντροπίας που εξετάζουμε. Στις επόμενες επαναλήψεις, για την αντικατάστασή του χρησιμοποιείται το εκάστοτε διάνυσμα τιμών των αναλογιών.

$$res_1 = subs\ (fa(1),\ a,\ a_{from_Gradient_Descent}),\ res_2 = subs\ (fa(2),\ a,\ a_{from_Gradient_Descent}),$$
$$\ldots,\ res_{22} = subs\ (fa(22),\ a,\ a_{from_Gradient_Descent})$$

f) Ορίζουμε το κριτήριο σύγκλισης, το οποίο σύμφωνα με τη θεωρία προκύπτει από τον τύπο

$$stop_criterion = \sqrt{res_1 + res_2 + \ldots + res_{22}}\ .$$

Οι μεταβλητές res_i αντιστοιχούν στην αριθμητική τιμή κάθε μερικής παραγώγου. Η τιμές των παραγώγων αλλάζουν σε κάθε επανάληψη, επομένως θα αλλάζει και η τιμή του κριτηρίου σύγκλισης. Στόχος του αλγορίθμου είναι το κριτήριο σύγκλισης να μειώνεται σε κάθε επανάληψη, μέχρι να φτάσει στην ελάχιστη δυνατή τιμή του. Για το πρόβλημα της ελαχιστοποίησης της εντροπίας, η ελάχιστη τιμή για το κριτήριο σύγκλισης είναι το $1{,}455583$ για πολλαπλασιαστή Lagrange $\lambda = 1{,}967$ και το $1{,}3904$ για πολλαπλασιαστή Lagrange $\lambda = 10$. Οι τιμές αυτές προέκυψαν έπειτα από την εκτέλεση του κώδικα και για τις 2 τιμές του πολλαπλασιαστή λ. Η ελάχιστη τιμή αυτή είναι λίγο μεγαλύτερη από την αντίστοιχη της μεθόδου των Ελαχίστων Τετραγώνων. Αυτό συμβαίνει γιατί εξαιτίας της ύπαρξης του λογαρίθμου μέσα στον τύπο υπολογισμού του κριτηρίου σύγκλισης, οι τιμές του κριτηρίου κατά την έναρξη των επαναλήψεων είναι αρκετά μεγάλες και με την πάροδο των επαναλήψεων μικραίνουν με αργό ρυθμό (λόγω και του μικρού βήματος), με αποτέλεσμα αυτή να είναι η πιο μικρή τιμή στην οποία μπορεί να φτάσει το κριτήριο. Πάντως, και πάλι, η ελάχιστη τιμή του κριτηρίου είναι μικρή και θετική, σε σύγκριση με το σύνολο των τιμών που λαμβάνει.

g) Τέλος εφαρμόζουμε τον επαναληπτικό τύπο που ουσιαστικά υλοποιεί και τον αλγόριθμο Gradient Descent, για τις 3 διαφορετικές περιπτώσεις αρχικοποίησης μείγματος και για τις 2 διαφορετικές τιμές του πολλαπλασιαστή λ. Από τον τύπο αυτό προκύπτουν κάθε φορά οι νέες τιμές των αναλογιών a_i. Σύμφωνα με τον τύπο αυτό, οι νέες αναλογίες προκύπτουν από τις αναλογίες τις προηγούμενης επανάληψης εάν αφαιρέσουμε την τιμή της παραγώγους της προηγούμενης επανάληψης, πολλαπλασιασμένης με το βήμα του αλγορίθμου. Η μαθηματική αναπαράσταση της παραπάνω έκφρασης είναι η

$$a_i^{k+1} = a_i^k - step * res_i^k\ .$$

Κατά την πάροδο των επαναλήψεων, για την περίπτωση της εντροπίας καταλήγουμε σε πολλές αρνητικές τιμές για τις αναλογίες και μάλιστα πολύ νωρίς. Επιπλέον, η λογαριθμική έκφραση $\log(C_{final} * \underline{a})$ που υπάρχει τόσο στις μερικές παραγώγους, όσο και στον τύπο του κριτηρίου σύγκλισης παίρνει και μιγαδικές τιμές. Αυτό είναι πολύ λογικό να συμβαίνει, αφού όπως γνωρίζουμε από τη θεωρία λογαρίθμων, το πεδίο ορισμού του λογαρίθμου είναι το $(0, 1]$, άρα για οποιαδήποτε τιμή εκτός αυτού, ο λογάριθμος θα οδηγεί σε μιγαδικό αποτέλεσμα. Γι' αυτό σε κάθε επανάληψη πραγματοποιήσαμε έναν πιο σύνθετο έλεγχο, απ' τον αντίστοιχο για το πρόβλημα των Ελαχίστων Τετραγώνων.

Πρόκειται για έναν έλεγχο σε δύο στάδια, ο οποίος στην υλοποίησή μας πραγματοποιήθηκε μέσα σε ένα βρόγχο ελέγχου (if – else) φαίνεται παρακάτω:

Στάδιο 1^{ο} - Έλεγχος Αρνητικών Αναλογιών

Εάν κάποια από τις αναλογίες (ή και όλες οι αναλογίες) της προηγούμενης επανάληψης έχει αρνητική τιμή, τότε η τιμή της στην επόμενη επανάληψη αντικαθίσταται με την τιμή $0{,}00001$, όπως και στον αντίστοιχο έλεγχο για το πρόβλημα των Ελαχίστων Τετραγώνων. Εάν οι αναλογίες έχουν μόνο θετικές τιμές, τότε οι τιμές τους παραμένουν ως έχουν και οι νέες αναλογίες προκύπτουν από την εφαρμογή του παραπάνω επαναληπτικού τύπου. Η επιλογή του συγκεκριμένου αριθμού για την αντικατάσταση των αρνητικών τιμών είναι η καλύτερη δυνατή. Αυτό ισχύει γιατί τυχόν μεγαλύτερες τιμές επηρεάζουν τη σύγκλιση του αλγορίθμου, οδηγώντας σε σκαμπανεβάσματα των τιμών του κριτηρίου σύγκλισης, αλλά ακόμη και σε άνοδο των τιμών του κριτηρίου, αντί για μείωση. Επίσης, τυχόν μικρότερες τιμές θα κατέληγαν σε αναλογίες πολύ μικρές (πολύ κοντά στο μηδέν), το οποίο σημαίνει ότι θα είναι και ασήμαντες.

Στάδιο 2^{ο} – Έλεγχος Τιμών Λογαρίθμου

Στο στάδιο αυτό ελέγχουμε τις τιμές που παίρνει η έκφραση $C_{final} * \underline{a}$, η οποία είναι το όρισμα του λογαρίθμου που υπάρχει στη σχέση των μερικών παραγώγων και του κριτηρίου σύγκλισης. Η ανάγκη ελέγχου του λογαρίθμου είναι και ο λόγος για τον οποίο η μερική παράγωγος της αντικειμενικής συνάρτησης ως προς κάθε a_i χρησιμοποιήθηκε με την αναλυτική, μαθηματική της μορφή και δεν υπολογίστηκε με χρήση κάποιας εντολής. Σύμφωνα με τη θεωρία λογαρίθμων, για τιμές μικρότερες του μηδενός, ο λογάριθμος καταλήγει σε μιγαδικά αποτελέσματα. Για την τιμή 0 δεν ορίζεται, ενώ στο διάστημα $(0,1]$ οδηγεί σε αρνητικές τιμές. Για τιμές μεγαλύτερες του 1 είναι θετικός. Ο έλεγχος που θα πραγματοποιήσουμε επομένως, πρέπει να είναι τέτοιος ώστε να αποφεύγονται τα μιγαδικά αποτελέσματα, αλλά και ταυτόχρονα να μειώνεται η τιμή του κριτηρίου, για να υπάρξει σύγκλιση. Μετά από διάφορες δοκιμές τιμών καταλήξαμε ότι όταν το όρισμα του λογαρίθμου είναι αρνητικό θα πρέπει να αντικαθίσταται με την τιμή $0{,}5$. Έτσι, το αποτέλεσμα του λογαρίθμου δε θα είναι ποτέ μιγαδικό και επειδή για 0.5 ο λογάριθμος θα είναι αρνητικός, το κριτήριο σύγκλισης θα μειώνεται.

Τέλος, ο αριθμός των επαναλήψεων εκτέλεσης του αλγορίθμου είναι διαφορετικός για κάθε περίπτωση, αφού επηρεάζεται από την τιμή του πολλαπλασιαστή λ. Στον ***Πίνακα 3.3*** παραθέτουμε τον αριθμό των επαναλήψεων που χρειάστηκε ο αλγόριθμος για να φτάσει στην ελάχιστη τιμή του κριτηρίου σύγκλισης, για κάθε μία από τις 2 περιπτώσεις πολλαπλασιαστών.

	λ = 1,967	**λ = 10**
Επαναλήψεις	2971	3104

Πίνακας 3.3. Αριθμός Επαναλήψεων Εκτέλεσης Αλγορίθμου (Entropy)

- **Κανονικοποίηση Αποτελεσμάτων Αναλογιών Πρόσμειξης**
  Τα αποτελέσματα που προέκυψαν από τον αλγόριθμο Gradient Descent, όπως και στην περίπτωση των ελαχίστων τετραγώνων (δηλαδή 22 αναλογίες για κάθε μία από τις 6 περιπτώσεις) ικανοποιούν όλους τους απαιτούμενους περιορισμούς, εκτός απο τον περιορισμό το άθροισμα των αναλογιών πρόσμειξης να είναι ακριβώς ίσο με 1. Συγκεκριμένα, οι αναλογίες που προκύπτουν απ' την εφαρμογή του αλγορίθμου έχουν θετικό άθροισμα, το οποίο όμως δεν είναι ίσο με 1, αλλά λίγο μικρότερο. Για να μπορέσουμε να δημιουργήσουμε άθροισμα ίσο με 1 χρειάστηκε να κανονικοποιήσουμε τα αποτελέσματα των αναλογιών πρόσμειξης. Υπολογίζουμε το άθροισμα των 22 αναλογιών που προκύπτουν από τον αλγόριθμο. Έπειτα διαιρούμε κάθε μία από τις 22 αναλογίες με το άθροισμα αυτό. Οι 22 τιμές που προκύπτουν είναι οι τελικές αναλογίες πρόσμειξης κυτταρικών σειρών, των οποίων το άθροισμα είναι θετικό και ίσο με 1.

3.4 Linear & Non Linear Least Squares in Matlab

3.4.1 Linear Least Squares (Lsqlin)

Εκτός από την εφαρμογή του αλγορίθμου Gradient Descent για τη μέθοδο των Ελαχίστων Τετραγώνων υπάρχει και ένας πιο άμεσος τρόπος εύρεσης των βέλτιστων αναλογιών πρόσμειξης. Με τη χρήση μίας μόνο εντολής του Matlab, της *lsqlin*, μπορούμε να βρούμε τις βέλτιστες αναλογίες πρόσμειξης κυτταρικών σειρών, χωρίς να εφαρμόσουμε αναλυτικά τον αλγόριθμο βελτιστοποίησης. Έτσι μπορούμε να έχουμε και ένα μέτρο σύγκρισης για την ορθότητα των αποτελεσμάτων που προέκυψαν από την εφαρμογή του αλγορίθμου Gradient Descent. Επιθυμούμε δηλαδή οι αναλογίες που προκύπτουν από τον αλγόριθμο να είναι ίδιες ή σχεδόν ίδιες με αυτές της lsqlin και έτσι, να δημιουργούν μείγματα με παρόμοια συμπεριφορά.

Η εντολή lsqlin επιλύει ένα γραμμικό πρόβλημα ελαχίστων τετραγώνων, το οποίο υπόκειται σε κάποιους περιορισμούς. Πρόκειται δηλαδή για έναν τρόπο επίλυσης προβλήματος ελαχιστοποίησης με περιορισμούς. Το συγκεκριμένο πρόβλημα ελαχιστοποίησης έχει τη μορφή

$$\begin{aligned} \min \; imize_{x} \quad & \frac{1}{2} * \left\| Cx - d \right\|_2^2 \\ subject \;\; to \quad & Ax \le b \\ & A_{eq} x = b_{eq} \\ & lb \le x \le ub \end{aligned}$$

όπου C είναι ο πίνακας των συντελεστών των αγνώστων του γραμμικού συστήματος εξισώσεων που επιθυμούμε να επιλύσουμε. Το d είναι το διάνυσμα των σταθερών του προς επίλυση συστήματος, A είναι ο πίνακας των συντελεστών του περιορισμού (ή των περιορισμών) ανισότητας που υπόκειται το πρόβλημα ελαχιστοποίησης και b το διάνυσμα των σταθερών του περιορισμού ανισότητας. Αντίστοιχα ορίζονται και τα A_{eq} και b_{eq} για τον περιορισμό ισότητας, ενώ τα lb και ub διανύσματα που αντιστοιχούν στο κάτω και το άνω όριο τιμών, τις οποίες μπορούν να λάβουν τα στοιχεία του διανύσματος x, ως προς το οποίο επιθυμούμε να ελαχιστοποιήσουμε. Η υλοποίηση του παραπάνω προβλήματος μπορεί να περιλαμβάνει ή όχι αρχικό σημείο (starting point) x_0.

Η σύνταξη της εντολής lsqlin με την παρουσία αρχικού σημείου, σύμφωνα με το Toolbox του Matlab είναι η εξής:

$$x = lsqlin\ (C,\ d,\ A,\ b,\ A_{eq},\ b_{eq},\ lb,\ ub,\ x_0)$$

Σε περίπτωση που δεν υπάρχουν περιορισμοί ισότητας ή ανισότητας, ή άνω και κάτω όρια τιμών για το x, τότε τη θέση τους στη σύνταξη της εντολής παίρνουν κενοί πίνακες και διανύσματα, οι οποίοι συντάσσονται ως $[]$. Επειδή, όπως γνωρίζουμε και από τη θεωρία, το πρόβλημα ελαχίστων τετραγώνων είναι πάντοτε ένα convex πρόβλημα ελαχιστοποίησης, η lsqlin θα βρίσκει πάντα ένα ολικό, αλλά όχι απαραίτητα μοναδικό ελάχιστο.

Συγκεκριμένα, το πρόβλημα της εύρεσης των βέλτιστων αναλογιών πρόσμειξης, με στόχο τη δημιουργία του πρότυπου μείγματος παίρνει τη μορφή

$$\begin{aligned} &\min\ imize && \frac{1}{2} * \left\| C_{final} * \underline{a} - \underline{A_{initial}} \right\| \\ &subject\ \ to && A_{eq}\, a = b_{eq} \\ & && lb \le a \le ub \end{aligned}$$

Πρόκειται δηλαδή για ένα πρόβλημα χωρίς περιορισμούς ανισότητας, όπου C_{final} (27×22) είναι ο πίνακας των δεδομένων (γονίδια – κυτταρικές σειρές), a (22×1) το διάνυσμα των αναλογιών ως προς το οποίο επιθυμούμε να ελαχιστοποιήσουμε και $A_{initial}$ (27×1) το διάνυσμα των αρχικών τιμών του target distribution. Ο περιορισμός ισότητας είναι το $\sum_{i=1}^{22} a_i = 1$ (αφού αναφερόμαστε σε αναλογίες, πρέπει το άθροσμά τους εξ ορισμού να είναι 1 ή αλλιώς 100%) και περιλαμβάνει τον πίνακα A_{eq} (27×22) και το διάνυσμα b_{eq} (27×1), οι οποίοι ορίζονται ως

$$\underset{(27 \times 22)}{A_{eq}} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \underset{(27 \times 1)}{b_{eq}} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Οι αναλογίες πρέπει να είναι θετικοί αριθμοί και η μέγιστη τιμή που μπορούν να πάρουν είναι το 1. Για το λόγο αυτό, το άνω όριο τιμών ub θα είναι ίσο με 1 και το κάτω όριο lb θα είναι ίσο με 0,001 (η εντολή δεν μπορεί να χρησιμοποιήσει ως όριο τον αριθμό 0, γι' αυτο χρησιμοποιούμε έναν ούτε πολύ μικρό, αλλά ούτε και πολύ μεγάλο θετικό αριθμό). Επίσης χρησιμοποιήσαμε την ίδια αρχικοποίηση για το διάνυσμα των αναλογιών, με αυτή που χρησιμοποιήθηκε στον αλγόριθμο Gradient Descent. Το διάνυσμα αρχικοποίησης a^0 (22×1) παίρνει τις τιμές

$$(a^0)' = \left[\underbrace{0{,}045 \quad 0{,}045 \quad \dots \quad 0{,}045 \quad 0{,}045}_{21} \quad \underbrace{0{,}055}_{1} \right]$$

Δεσμεύσαμε μνήμη στο Matlab για τον πίνακα των δεδομένων και για τον πίνακα A_{eq} και δημιουργήσαμε αντίστοιχα το διάνυσμα αρχικών τιμών του μείγματος και το διάνυσμα b_{eq}. Για τις αναλογίες πρόσμειξης δεσμεύσαμε μνήμη για ένα διάνυσμα συμβόλων 22×1 (όπως και στη μέθοδο του Gradient Descent), μιας και αποτελούν τους αγνώστους του προβλήματος μας.

Τέλος, για τα άνω και κάτω όρια των τιμών των αναλογιών δεσμεύσαμε μνήμη για δύο διανύσματα (22×1), χρησιμοποιώντας την εντολή *ones*. Η εντολή αυτή δημιουργεί ένα πίνακα ή διάνυσμα με όλα τα στοιχεία του να είναι ίσα με 1. Επειδή όμως το κάτω όριο στο πρόβλημά μας πρέπει να αντιπροσωπεύεται από ένα διάνυσμα που όλα τα στοιχεία του θα είναι ίσα με 0,001 και όχι με 1, πολλαπλασιάσαμε όλα τα στοιχεία του διανύσματος που προέκυψαν από την εντολή ones με τον αριθμό 0,001. Έτσι τα άνω και κάτω όρια περιγράφονται στον κώδικά μας ως

$$lb = 0.001 * ones\ (22, 1)$$

$$ub = ones\ (22, 1)$$

Αφού δημιουργήσαμε όλα τα απαραίτητα ορίσματα, χρησιμοποιήσαμε τελικά την εντολή lsqlin, απ΄την οποία προέκυψαν και τα αποτελέσματα για τις αναλογίες πρόσμειξης, αποθηκευμένα στο διάνυσμα συμβόλων a. Η εντολή εκτελέστηκε και για τις 3 διαφορετικές περιπτώσεις αρχικοποίησης του target distribution ($A_{initial} = 1/27$, $A_{initial} = 9{,}1194$, $A_{initial} \geq 350$).

3.4.2 Non Linear Least Squares (Lsqnonlin)

Η αντίστοιχη εντολή με την lsqlin του Matlab είναι η *lsqnonlin*. Με τη χρήση της εντολής αυτής μπορούμε να βρούμε τις βέλτιστες αναλογίες πρόσμειξης κυτταρικών σειρών, χωρίς να εφαρμόσουμε αναλυτικά τον αλγόριθμο βελτιστοποίησης για την αντικειμενική συνάρτηση της εντροπίας. Έτσι μπορούμε και πάλι να έχουμε ένα μέτρο σύγκρισης για την ορθότητα των αποτελεσμάτων που προέκυψαν από την εφαρμογή του αλγορίθμου Gradient Descent. Επιθυμούμε δηλαδή οι αναλογίες που προκύπτουν από τον αλγόριθμο να είναι ίδιες ή σχεδόν ίδιες με αυτές της lsqnonlin και έτσι, να δημιουργούν μείγματα με παρόμοια συμπεριφορά.

Η εντολή lsqnonlin επιλύει ένα μη γραμμικό πρόβλημα ελαχίστων τετραγώνων, το οποίο υπόκειται σε κάποιους περιορισμούς. Πρόκειται δηλαδή για έναν άμεσο τρόπο επίλυσης προβλήματος ελαχιστοποίησης με περιορισμούς. Το συγκεκριμένο πρόβλημα ελαχιστοποίησης έχει τη μορφή

$$\min_x \quad f(x) = f_1(x)^2 + f_2(x)^2 + \ldots + f_m(x)^2$$

$$subject\ \ to \quad lb \leq x \leq ub$$

όπου $f(x)$ είναι η συνάρτηση προς ελαχιστοποίηση, η οποία αντιπροσωπεύεται από το άθροισμα τετραγωνικών, μη γραμμικών όρων. Τέλος, τα lb και ub είναι διανύσματα που αντιστοιχούν στο κάτω και το άνω όριο τιμών, τις οποίες μπορούν να λάβουν τα στοιχεία του διανύσματος x, ως προς το οποίο επιθυμούμε να ελαχιστοποιήσουμε. Η υλοποίηση του παραπάνω προβλήματος μπορεί να περιλαμβάνει ή όχι αρχικό σημείο (starting point) x_0.

Αντί να χρειάζεται κάθε φορά να υπολογίζουμε την τιμή της $f(x)$ (δηλαδή το άθροισμα των τετραγώνων), στην lsqnonlin αρκεί απλά να ορίζουμε τη διανυσματική συνάρτηση

$$F(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{bmatrix}.$$

Τότε, μιλώντας πάντα σε όρους διανυσματικούς μπορούμε να ξανα – ορίσουμε το παραπάνω μη γραμμικό πρόβλημα ελαχιστοποίησης ως

$$\min_x \quad \frac{1}{2}\|F(x)\|_2^2 = \frac{1}{2}\sum_{i=1}^{m} f_i(x)^2$$

όπου x είναι ένα διάνυσμα και $F(x)$ είναι μία συνάρτηση που επιστρέφει μία διανυσματική τιμή.

Η σύνταξη της εντολής lsqnonlin με την παρουσία αρχικού σημείου, σύμφωνα με το Toolbox του Matlab είναι η εξής:

$$x = lsqnonlin \quad (fun, x_0, lb, ub)$$

Συγκεκριμένα, το πρόβλημα της εύρεσης των βέλτιστων αναλογιών πρόσμειξης, με στόχο τη δημιουργία του πρότυπου μείγματος παίρνει τη μορφή

$$\min \; imize \qquad fun = (C_{final} * \underline{a}) \log(C_{final} * \underline{a})$$

$$subject \quad to \quad lb \le a \le ub$$

Πρόκειται δηλαδή για ένα μη γραμμικό πρόβλημα, όπου C_{final} (27×22) είναι ο πίνακας των δεδομένων (γονίδια – κυτταρικές σειρές) και a (22×1) το διάνυσμα των αναλογιών ως προς το οποίο επιθυμούμε να ελαχιστοποιήσουμε. Οι αναλογίες πρέπει να είναι θετικοί αριθμοί και η μέγιστη τιμή που μπορούν να πάρουν είναι το 1. Για το λόγο αυτό, το άνω όριο τιμών ub θα είναι ίσο με 1 και το κάτω όριο lb θα είναι ίσο με 0,001 (η εντολή δεν μπορεί να χρησιμοποιήσει ως όριο τον αριθμό 0, γι' αυτο χρησιμοποιούμε έναν ούτε πολύ μικρό, αλλά ούτε και πολύ μεγάλο θετικό αριθμό). Επίσης χρησιμοποιήσαμε την ίδια αρχικοποίηση για το διάνυσμα των αναλογιών, με αυτή που χρησιμοποιήθηκε στον αντίστοιχο αλγόριθμο Gradient Descent για την ελαχιστοποίηση της εντροπίας. Το διάνυσμα αρχικοποίησης a^0 (22×1) παίρνει τις τιμές

$$(a^0)' = [0{,}2 \quad 0{,}3 \quad 0{,}4 \quad 0{,}1 \quad 0{,}25 \quad 0{,}2 \quad 0{,}17 \quad 0{,}15 \quad 0{,}21 \quad 0{,}3 \quad 0{,}1 \quad 0{,}22$$
$$0{,}13 \quad 0{,}29 \quad 0{,}35 \quad 0{,}23 \quad 0{,}12 \quad 0{,}11 \quad 0{,}09 \quad 0{,}05 \quad 0{,}1 \quad 0{,}16]$$

Στην περίπτωση της lsqnonlin δεν μπορούμε να αποθηκεύσουμε το διάνυυσμα των αναλογιών, ως ένα διάνυσμα από σύμβολα, γιατί η εντολή δεν δέχεται ως είσοδο συναρτήσεις με συμβολική μορφή. Έτσι ορίζοντας τη συνάρτηση fun εισάγουμε ταυτόχρονα ως όρισμα το διάνυσμα των αναλογιών a, χωρίς να χρειάζεται κάποια περαιτέρω διαδικασία για τον ορισμό των αναλογιών. Τέλος, για τα άνω και κάτω όρια των τιμών των αναλογιών δεσμεύσαμε μνήμη για δύο διανύσματα (22×1), χρησιμοποιώντας και πάλι την εντολή *ones*. Έτσι τα άνω και κάτω όρια περιγράφονται στον κώδικά μας ως

$$lb = 0.001 * ones \quad (22, 1)$$

$$ub = ones \quad (22, 1)$$

Αφού δημιουργήσαμε όλα τα απαραίτητα ορίσματα, χρησιμοποιήσαμε τελικά την εντολή lsqnonlin, απ΄την οποία προέκυψαν και τα αποτελέσματα για τις αναλογίες πρόσμειξης, αποθηκευμένα στο διάνυσμα συμβόλων a. Επειδή το άθροισμα των αναλογιών που προέκυψαν είναι μικρότερο και όχι ίσο με το 1, έπρεπε και πάλι να κανονικοποιήσουμε τα αποτελέσματά μας. Αυτό συμβαίνει γιατί, όπως φαίνεται και στη σύνταξη της εντολής δεν μπορούμε να

εισάγουμε στην lsqnonlin περιορισμούς ισότητας, όπως είναι ο περιορισμός $\sum_{i=1}^{22} a_i = 1$, στον οποίο υπόκεινται οι αναλογίες μας. Η κανονικοποίηση έγινε διαιρώντας τις αναλογίες που προέκυψαν με το άθροισμά τους. Έτσι καταλήγουμε σε αναλογίες που ικανοποιούν όλους τους δυνατούς περιορισμούς. Στην περεταίρω ανάλυση θα χρησιμοποιούνται οι κανονικοποιημένες αναλογίες πρόσμειξης κυτταρικών σειρών.

ΚΕΦΑΛΑΙΟ 4 – ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 Αποτελέσματα

Από την υλοποίηση του αλγορίθμου Gradient Descent, αλλά και από τη χρήση των εντολών του Matlab, lsqlin και lsqnonlin για τη μέθοδο Ελαχίστων Τετραγώνων και της ελαχιστοποίησης της Εντροπίας αντίστοιχα προέκυψαν τα ζητούμενα αποτελέσματα για τις αναλογίες πρόσμειξης των κυτταρικών σειρών (πάντα κανονικοποιημένα ώστε $\sum_{i=1}^{22} a_i = 1$). Έτσι, με τα αποτελέσματα αυτά δημιουργήσαμε τα αντίστοιχα πρότυπα μείγματα A_{final}, το οποίο ήταν και το ζητούμενο της εργασίας. Για τα μείγματα αυτά υπολογίσαμε τη μέση τιμή και τη διασπορά τους και εξετάσαμε την ομοιομορφία τους.

Επομένως, για τη μέθοδο των Ελαχίστων Τετραγώνων, οι αναλογίες και τα αντίστοιχα πρότυπα μείγματα αντιστοιχούν σε 3 διαφορετικές αρχικοποιήσεις (όπως αναλύθηκε και παραπάνω), με 2 υπό – περιπτώσεις εκτέλεσης αλγορίθμου Gradient Descent και 1 υπό – περίπτωση εκτέλεσης της εντολής του lsqlin του Matlab. Συνεπώς έχουμε **9** ***διαφορετικά μείγματα που προκύπτουν απ' τη μέθοδο Ελαχίστων Τετραγώνων***. Αντίθετα για τη μέθοδο ελαχιστοποίησης της Εντροπίας έχουμε 2 μείγματα για την υπό – περίπτωση του Gradient Descent και άλλο 1 για την υπό – περίπτωση της lsqnonlin. Άρα έχουμε ***3 διαφορετικά μείγματα που προκύπτουν από τη μέθοδο ελαχιστοποίησης της Εντροπίας***.

Το κάθε πρότυπο μείγμα και για τις δύο μεθόδους προκύπτει από τη σχέση $\boxed{A_{final} = C_{final} * a_i}$, όπου a_i το διάνυσμα των αναλογιών για κάθε περίπτωση και C_{final} ο κανονικοποιημένος πίνακας δεδομένων, μεγέθους 27×22 .

- **Μέθοδος Ελαχίστων Τετραγώνων**

  Για να δημιουργήσουμε λοιπόν το κάθε μείγμα πραγματοποιήσαμε στο Matlab τον πολλαπλασιασμό του πίνακα δεδομένων (27×22), με το διάνυσμα αναλογιών (22×1), δεσμεύοντας μνήμη για τους πίνακες που συμμετέχουν στον πολλαπλασιασμό, αλλά και για το διάνυσμα στο οποίο αποθηκεύονται οι τιμές του κάθε μείγματος (μέγεθος 27×1). Έπειτα υπολογίσαμε τη μέση τιμή και τη διασπορά του κάθε διανύσματος μείγματος, χρησιμοποιώντας τις συναρτήσεις του Matlab, *mean και var*, με σύνταξη

  $$M = mean\ (A)$$

  $$V = \mathrm{var}(\ A)$$

  Η έτοιμη συνάρτηση mean υπολογίζει το άθροισμα των στοιχείων του διανύσματος και το διαιρεί με το πλήθος των στοιχείων. Η διασπορά αποτελεί ένα μέσο εκτίμησης για το πόσο κοντά η μακριά κατανέμονται οι τιμές του διανύσματος, γύρω από τη μέση τιμή. Η συνάρτηση var του Matlab υπολογίζει τη διασπορά στηριζόμενη στον τύπο $\sigma_x^2 = mean\ (x^2) - \{mean\ (x)\}^2$.

Στους ***Πίνακες 4.1, 4.2, 4.3*** που ακολουθούν παραθέτουμε τα αποτελέσματα για τις κανονικοποιημένες αναλογίες πρόσμειξης των κυτταρικών σειρών και τις τιμές για τα αντίστοιχα μείγματα, για τις τρεις πιθανές περιπτώσεις αρχικοποίησης του διανύσματος που αντιπροσωπεύει το μείγμα.

Gradient Descent				**Lsqlin**	
λ = 1,967		**λ = 10**			
Αναλογίες a_i	**Μείγμα**	**Αναλογίες** a_i	**Μείγμα**	**Αναλογίες** a_i	**Μείγμα**
6,50E-04	7,79E+00	1,86E-02	8,0142416	0,001	8,1176625
7,39E-03	8,26E+00	3,90E-02	8,2503001	0,001	8,4858268
2,09E-02	7,90E+00	4,30E-02	7,9797685	0,001	7,831516
4,21E-03	1,03E+01	3,60E-02	10,075857	0,001	9,3742513
4,19E-03	8,16E+00	3,50E-02	8,0296809	0,001	8,0531513
1,54E-03	7,89E+00	2,75E-02	7,8855727	0,001	7,8408932
3,82E-04	1,20E+01	1,63E-02	11,906545	0,001	12,185033
1,40E-03	1,05E+01	3,28E-02	10,763771	0,001	8,6539116
2,45E-02	8,25E+00	4,30E-02	8,1668024	0,001	7,9658909
2,52E-02	1,01E+01	4,56E-02	10,073377	0,001	10,153414
2,50E-02	9,38E+00	4,61E-02	9,0841616	0,001	8,9324694
2,77E-02	8,98E+00	4,42E-02	9,0637909	0,001	8,8406245
5,12E-03	8,04E+00	3,75E-02	8,087929	0,001	7,9268754
7,15E-03	8,79E+00	3,92E-02	8,3200105	0,001	7,8646063
2,07E-03	9,34E+00	3,03E-02	9,1235101	0,001	8,8779988
5,37E-02	8,67E+00	6,17E-02	8,3085746	0,001	8,9589084
2,77E-02	8,50E+00	5,26E-02	8,5291631	0,979	8,5891816
1,25E-02	8,31E+00	3,92E-02	8,2827085	0,001	8,0989587
4,26E-02	8,55E+00	5,65E-02	8,7645102	0,001	8,4455773
3,74E-04	1,21E+01	2,82E-02	12,758273	0,001	11,772379
5,15E-03	1,27E+01	3,69E-02	12,587458	0,001	13,119131
7,01E-01	7,76E+00	1,92E-01	7,7314227	0,001	7,6399159
	9,1320557		8,8548264		8,9059488
	8,1095773		8,050708		8,122382
	11,710879		11,846592		11,815284
	9,5952925		9,5832743		9,0156164
	8,8691174		8,8530989		8,3764196

Πίνακας 4.1. Κανονικοποιημένες Αναλογίες & Μείγματα για ***την αρχικοποίηση*** 1 / 27

Gradient Descent				Lsqlin	
λ = 1,967		λ = 10			
Αναλογίες a_i	**Μείγμα**	**Αναλογίες** a_i	**Μείγμα**	**Αναλογίες** a_i	**Μείγμα**
1,03E-05	8,0295421	0,0191717	8,0585280	0,001	8,0124966
3,60E-02	8,3034622	0,0459608	8,2402527	0,001	8,2719059
4,39E-02	7,9894500	0,0479483	7,9860656	0,001	7,9470116
1,08E-02	9,8511711	0,0414603	10,003909	0,001	9,4628602
5,37E-03	8,0242430	0,0399016	7,9887082	0,001	8,0553368
1,03E-05	7,9083392	0,0303875	7,8759545	0,001	7,822983
1,03E-05	11,736727	0,0207388	11,866262	0,001	11,708089
5,56E-04	10,556773	0,0386911	10,784706	0,001	10,259389
4,59E-02	8,1630362	0,0469913	8,1351240	0,001	8,1492489
6,91E-02	9,9825886	0,0523338	10,047135	0,001	10,079675
7,08E-02	9,0865331	0,0525605	9,0022893	0,001	9,1324126
4,68E-02	9,0774284	0,047868	9,0705051	0,001	9,0678247
2,77E-02	8,1014937	0,0440611	8,0893100	0,001	8,2331153
4,44E-02	8,1535901	0,0467094	8,1887391	0,001	8,176903
4,11E-06	9,1660345	0,0344696	9,0613622	0,001	9,6425916
1,51E-01	8,2352124	0,0707899	8,2086012	0,491953745	8,2900165
1,16E-01	8,5101163	0,0639223	8,5229582	0,166036092	8,4496423
3,44E-02	8,2376912	0,0445813	8,2619173	0,001	8,1786084
1,26E-01	8,8851342	0,0654796	8,8132269	0,134340957	9,1966281
2,13E-06	12,556804	0,0336154	12,868575	0,001	12,137536
3,11E-02	12,520985	0,0437775	12,539992	0,001	12,404025
1,41E-01	7,7609313	0,0685804	7,7172852	0,189669206	7,7794285
	8,8271019		8,7769547		8,7770619
	8,0500913		8,0254121		8,0390101
	11,808408		11,861581		11,848022
	9,5584006		9,5629594		9,6272417
	8,9125502		8,8402337		8,8921491

Πίνακας 4.2. Κανονικοποιημένες Αναλογίες & Μείγματα για ***την αρχικοποίηση*** 9,1194

Gradient Descent				Lsqlin	
λ = 1,967		λ = 10			
Αναλογίες a_i	**Μείγμα**	**Αναλογίες** a_i	**Μείγμα**	**Αναλογίες** a_i	**Μείγμα**
0,016888	8,064694	0,0394352	8,067868	0,001	7,9186669
0,05483	8,254704	0,0474528	8,228761	0,001	8,3027358
0,033935	7,979027	0,0431634	7,981737	0,001	7,9547106
0,044276	10,01464	0,0455419	10,07213	0,001	10,069841
0,018651	7,997983	0,0401458	7,988309	0,001	8,0747376
0,034793	7,887443	0,0435426	7,868698	0,169919206	7,9153457
0,033733	11,90903	0,0432886	11,94584	0,001	12,005151
0,041257	10,721	0,0450318	10,84289	0,019540662	10,958241
0,043206	8,124687	0,044377	8,131227	0,113787445	8,266955
0,055775	10,02784	0,0473677	10,06542	0,001	9,9894942
0,054102	8,991422	0,0470197	8,993994	0,080048642	9,158395
0,048635	9,052585	0,045637	9,065189	0,018104722	8,9381733
0,053067	8,087349	0,047025	8,088661	0,001	8,161796
0,06374	8,197059	0,04926	8,20452	0,056153693	8,5480977
0,021174	9,01801	0,0398117	9,033547	0,001	9,1461412
0,055421	8,257933	0,0473557	8,230321	0,001	8,5156662
0,080892	8,520873	0,0532499	8,521824	0,031678907	8,4999979
0,067207	8,253745	0,0495052	8,263817	0,135358689	8,283493
0,047464	8,714236	0,0457943	8,761879	0,001	8,6862077
0,021426	12,88711	0,040622	12,98347	0,001	12,888862
0,047452	12,57335	0,0455048	12,54542	0,001	12,612666
0,062075	7,724947	0,0498678	7,709933	0,362408034	7,7128256
	8,775239		8,748661		8,914675
	8,025585		8,0169		8,0499333
	11,83942		11,88441		11,693132
	9,52291		9,563887		9,5762341
	8,788333		8,802413		8,8469975

Πίνακας 4.3. Κανονικοποιημένες Αναλογίες & Μείγματα για ***την αρχικοποίηση*** ≥ 350

Στους ***Πίνακες 4.4, 4.5, 4.6*** περιλαμβάνονται οι μέσες τιμές και οι διασπορές κάθε μείγματος, για κάθε περίπτωση αρχικοποίησης.

	$A_{initial}$ = 1/27		
	Gradient Descent		**Lsqlin**
	λ = 1,967	**λ = 10**	
Μέση Τιμή	9,2464	9,2213	9,035697
Διασπορά	2,0697217	2,2530547	2,184243

Πίνακας 4.4. Μέσες Τιμές & Διασπορές για την 1η αρχικοποίηση

	$A_{initial}$ = 9,1194		
	Gradient Descent		**Lsqlin**
	λ = 1,967	**λ = 10**	
Μέση Τιμή	9,184957	9,1999463	9,171897
Διασπορά	2,103013	2,2905617	1,951897

Πίνακας 4.5. Μέσες Τιμές & Διασπορές για την 2η αρχικοποίηση

	$A_{initial} \geq 350$		
	Gradient Descent		**Lsqlin**
	λ = 1,967	**λ = 10**	
Μέση Τιμή	9,1930058	9,207842	9,247747
Διασπορά	2,2988271	2,3634005	2,271283

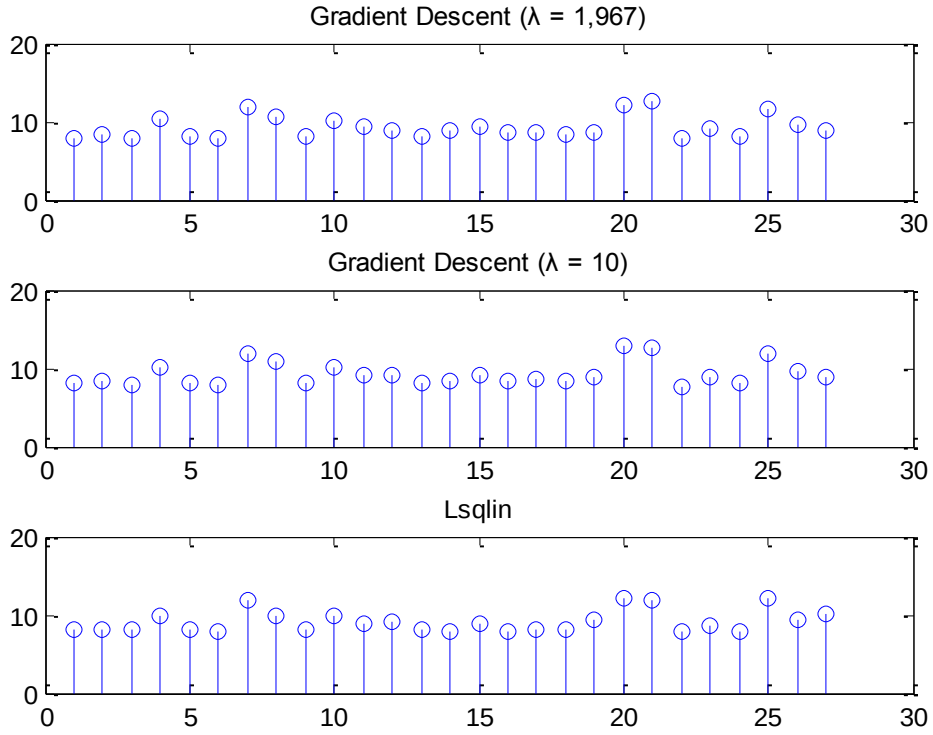
Πίνακας 4.6. Μέσες Τιμές & Διασπορές για την 3η αρχικοποίηση

Εξετάζοντας στο σύνολο τις τιμές των μέσων τιμών συμπεραίνουμε ότι όλα τα μείγματα που προκύπτουν έχουν *σχεδόν ίσες μέσες τιμές*. Επίσης, ελέγχοντας τις διασπορές παρατηρούμε ότι όλες *κυμαίνονται γύρω από την ίδια τιμή* (≈ 2). Η τιμή αυτή μπορεί να μην είναι 0 (πλήρης ομοιομορφία), αλλά είναι αρκετά μικρή, ώστε να εξασφαλίζει *μία καλή ομοιομορφία* για τα πρότυπα μείγματα αναφοράς που δημιουργήθηκαν. Εφόσον, λοιπόν, τα τελικά μείγματα που προέκυψαν από τις 3 διαφορετικές αρχικοποιήσεις έχουν την ίδια συμπεριφορά, τότε μπορούμε να πούμε *ότι η αρχικοποίηση(target distribution) που θα επιλεγεί δεν επηρεάζει τα αποτελέσματα για τις βέλτιστες αναλογίες πρόσμειξης, και συνεπώς ούτε και τα μείγματα αναφοράς που προκύπτουν*. Εάν βέβαια θελήσουμε να λάβουμε υπόψη μας τις διαφορές που παρατηρούνται στις τιμές των διασπορών, παρ' ότι αυτές είναι πολύ μικρές, τότε καταλήγουμε στο ότι η αρχικοποίηση που οδηγεί σε χειρότερα αποτελέσματα όσον αφορά την ομοιομορφία είναι η $A_{initial} \geq 350$. Και αυτό γιατί, οι μεγαλύτερες τιμές διασποράς παρουσιάζονται για την αρχικοποίηση μείγματος με τιμές μεγαλύτερες από το 350. Οι άλλες δύο αρχικοποιήσεις είναι εξίσου καλές όσον αφορά την ομοιομορφία των μειγμάτων αναφοράς.

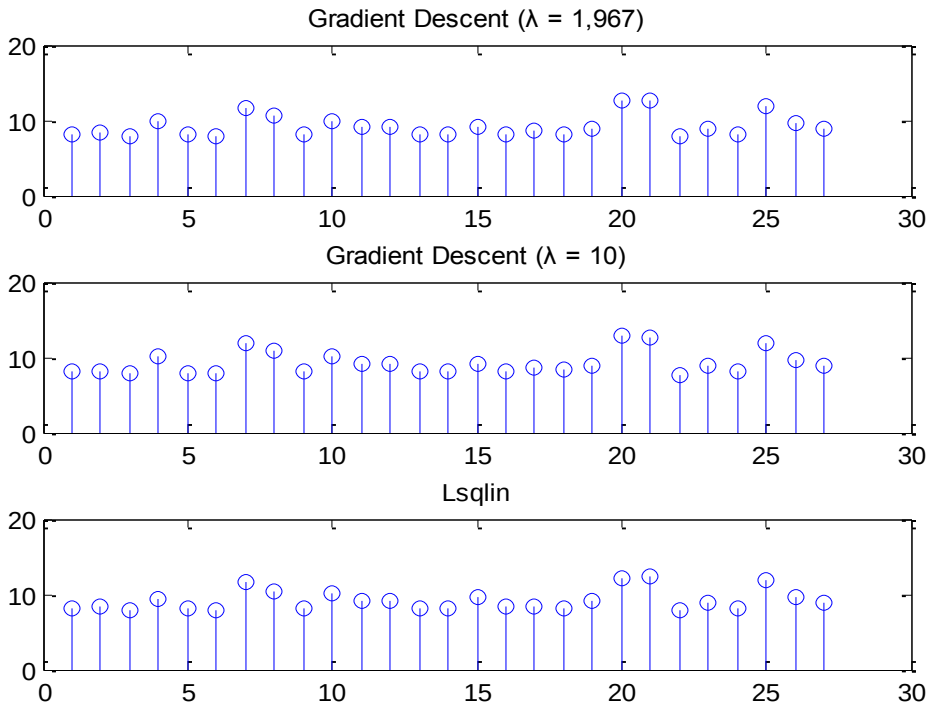
Έπειτα, δημιουργήσαμε τις γραφικές παραστάσεις για τις τιμές κάθε μείγματος αναφοράς, ώστε να μπορέσουμε να ελέγξουμε την ομοιομορφία τους και μέσω σχηματικής αναπαράστασης. Για τη δημιουργία των γραφικών παραστάσεων χρησιμοποιήσαμε την εντολή *stem* του Matlab, η οποία δημιουργεί δισδιάστατες γραφικές παραστάσεις για ακολουθίες διακριτών μεταβλητών και της οποίας η σύνταξη φαίνεται παρακάτω

$$stem\ (x,\ y)$$

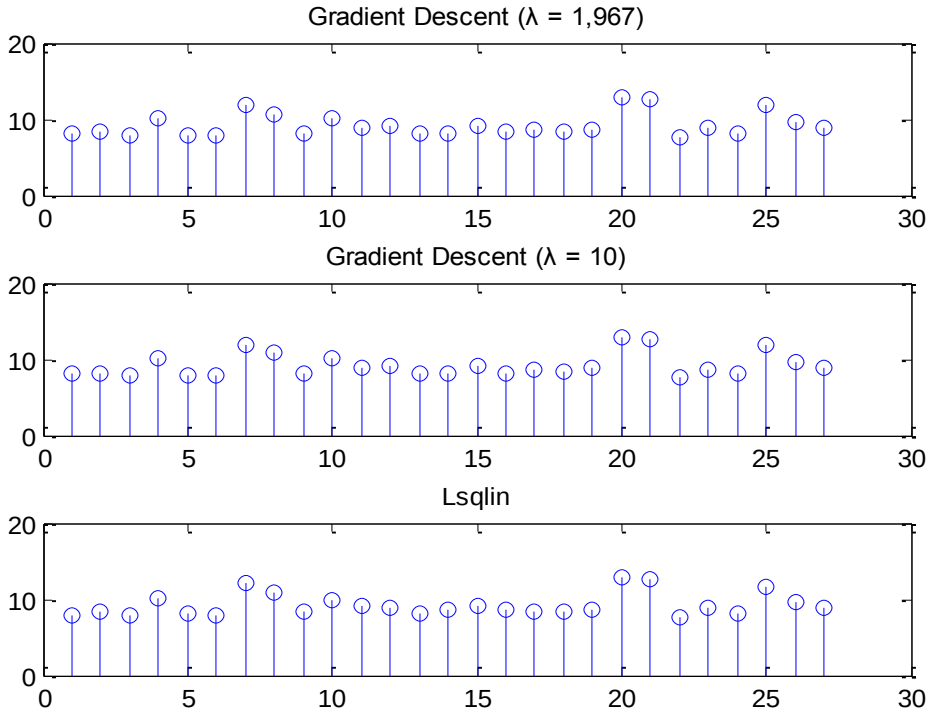
όπου x είναι το διάνυσμα που αποθηκεύει τις τιμές του x – άξονα της γραφικής παράστασης, ενώ y, η συνάρτηση, της οποίας τη γραφική παράσταση επιθυμούμε να δημιουργήσουμε. Τα ***Σχήματα 4.1, 4.2 και 4.3*** αντιστοιχούν στις 3 διαφορετικές αρχικοποιήσεις που χρησιμοποιήθηκαν και περιλαμβάνουν τις γραφικές παραστάσεις των πρότυπων μειγμάτων που προέκυψαν από τον αλγόριθμο Gradient Descent (2 γραφικές για τους 2 διαφορετικούς πολλαπλασιαστές λ) και από την εντολή του Matlab lsqlin. Για να μπορέσουμε να εμφανίσουμε ταυτόχρονα και τις τρεις γραφικές παραστάσεις σε ένα και μόνο σχήμα, χρησιμοποιήσαμε επιπλέον την εντολή του Matlab *subplot*.



Σχήμα 4.1. Γραφικές Παραστάσεις Μειγμάτων ***για Αρχικοποίηση*** 1 / 27



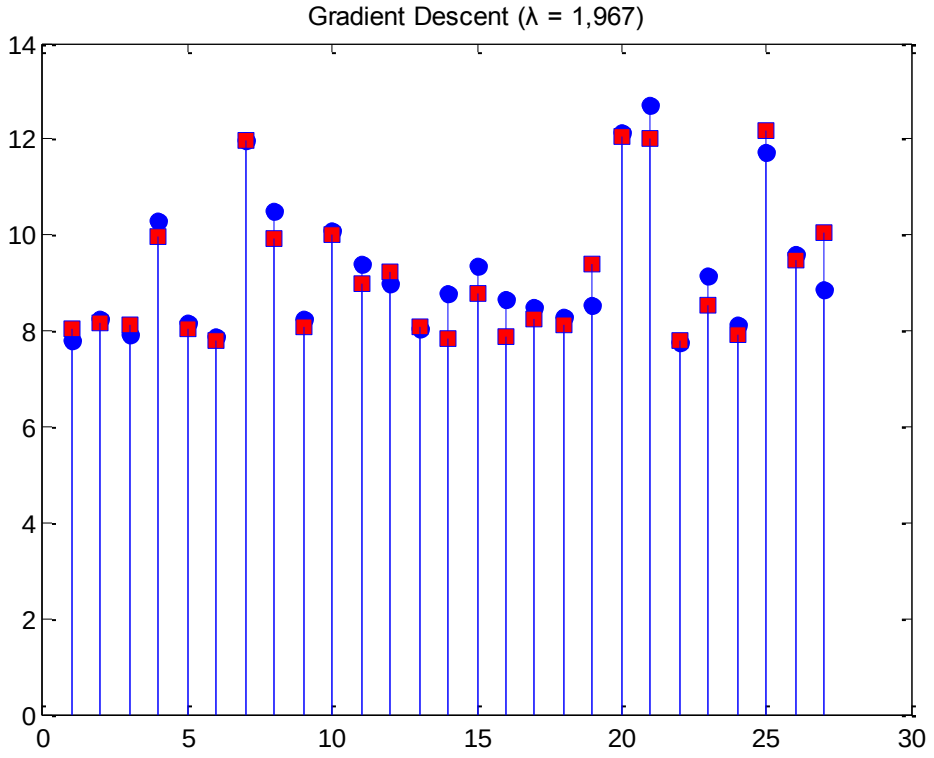
Σχήμα 4.2. Γραφικές Παραστάσεις Μειγμάτων ***για Αρχικοποίηση*** 9,1194



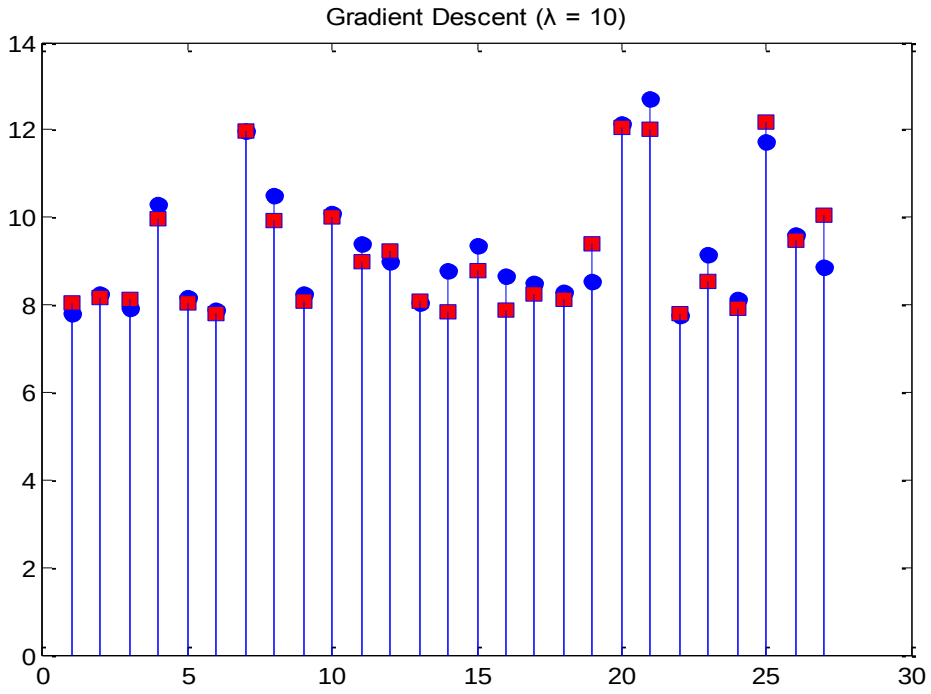
Σχήμα 4.3. Γραφικές Παραστάσεις Μειγμάτων ***για Αρχικοποίηση*** ≥ 350

Η γραφική παράσταση κάθε μείγματος παρουσιάζει μεγάλη ομοιότητα με τη θεωρητική γραφική παράσταση που περιγράφει την ομοιόμορφη κατανομή, της οποίας όλα στα στοιχεία λαμβάνουν την ίδια τιμή στον άξονα των y, σχηματίζοντας μία ευθεία γραμμή, παράλληλη με τον άξονα των x. Δηλαδή μπορούμε να συμπεράνουμε και με γραφικό τρόπο, ότι κάθε ένα από τα μείγματα αναφοράς ικανοποιεί την απαίτηση της ομοιομορφίας, μιας και τα περισσότερα στοιχεία κάθε μείγματος κυμαίνονται γύρω από την ίδια τιμή και μόνο λίγα παρουσιάζουν αποκλίσεις. Οι μικρές αυτές αποκλίσεις των λίγων αυτών στοιχείων είναι ο λόγος που η διασπορά είναι λίγο μεγαλύτερη και όχι ίση ή σχεδόν ίση με το μηδέν (5 στοιχεία αποκλίνουν έντονα). Τέλος, συγκρίνοντας τα μείγματα που προέκυψαν για κάθε αρχικοποίηση παρατηρούμε ότι δεν είναι μόνο ομοιόμορφα, αλλά ότι η συμπεριφορά τους όσον αφορά την ομοιομορφία είναι κοινή (σχεδόν ίδιες μέσες τιμές και διασπορές). Πιστοποιούμε δηλαδή και με γραφικό τρόπο το ότι *η αρχικοποίηση(target distribution) που θα επιλεγεί δεν επηρεάζει τα αποτελέσματα για τα μείγματα αναφοράς που προκύπτουν.*

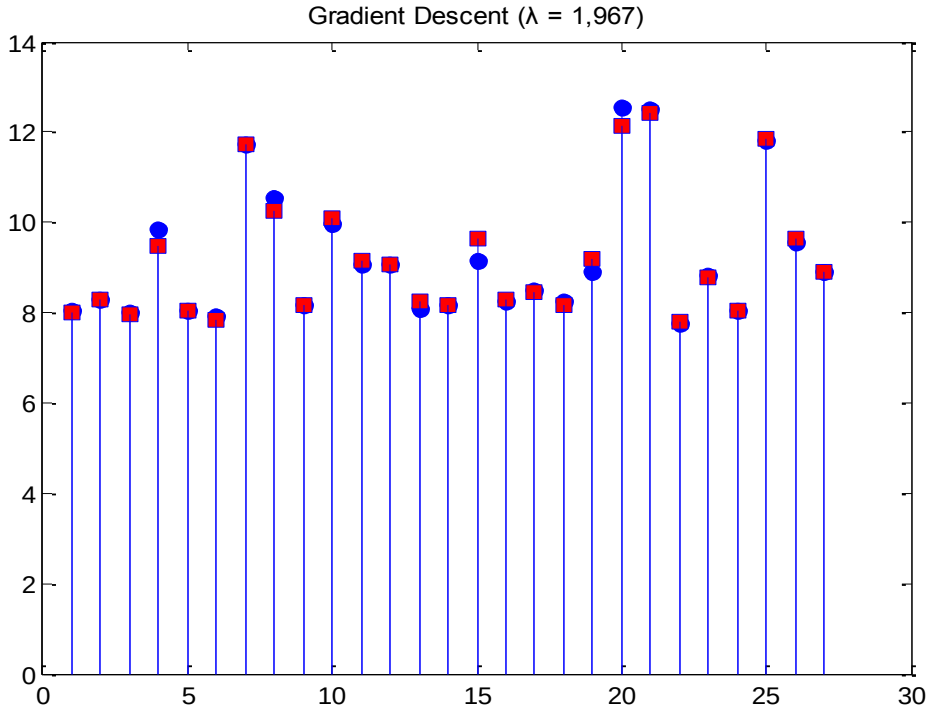
Τέλος, ελέγξαμε εάν τα μείγματα που προκύπτουν από την εφαρμογή του Gradient Descent παρουσιάζουν ίδια ή διαφορετική συμπεριφορά με αυτά που προκύπτουν από τη χρήση της εντολής lsqlin. Δημιουργήσαμε για κάθε αρχικοποίηση τα ***ζεύγη μειγμάτων*** $\boldsymbol{A_{GD_\lambda = 1,967} - A_{lsqlin}}$ και $\boldsymbol{A_{GD_\lambda = 10} - A_{lsqlin}}$. Τα μείγματα δημιουργούνται και πάλι με την εντολή stem και εμφανίζονται στο ίδιο σχήμα, με τη βοήθεια της εντολής *hold* του Matlab (μπλε χρώμα για Gradient Descent και κόκκινο χρώμα για lsqlin).



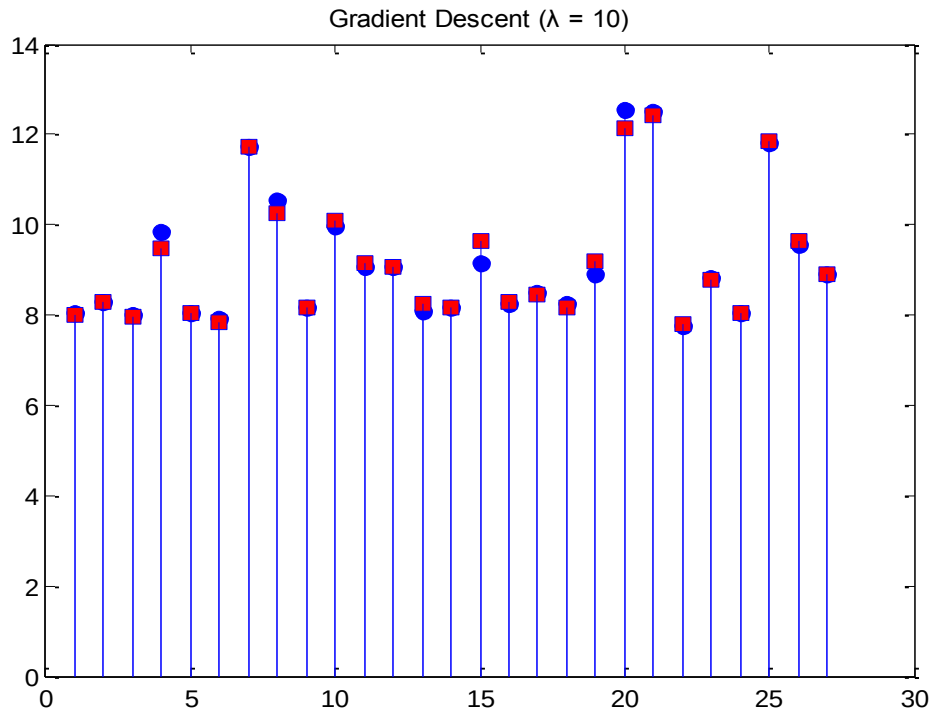
Σχήμα 4.4. Σύγκριση Μειγμάτων που προέκυψαν από τη μέθοδο Gradient Descent και την lsqlin, για λ = 1,967 (χρήση στο Gradient Descent), ***για την αρχικοποίηση*** 1 / 27 .



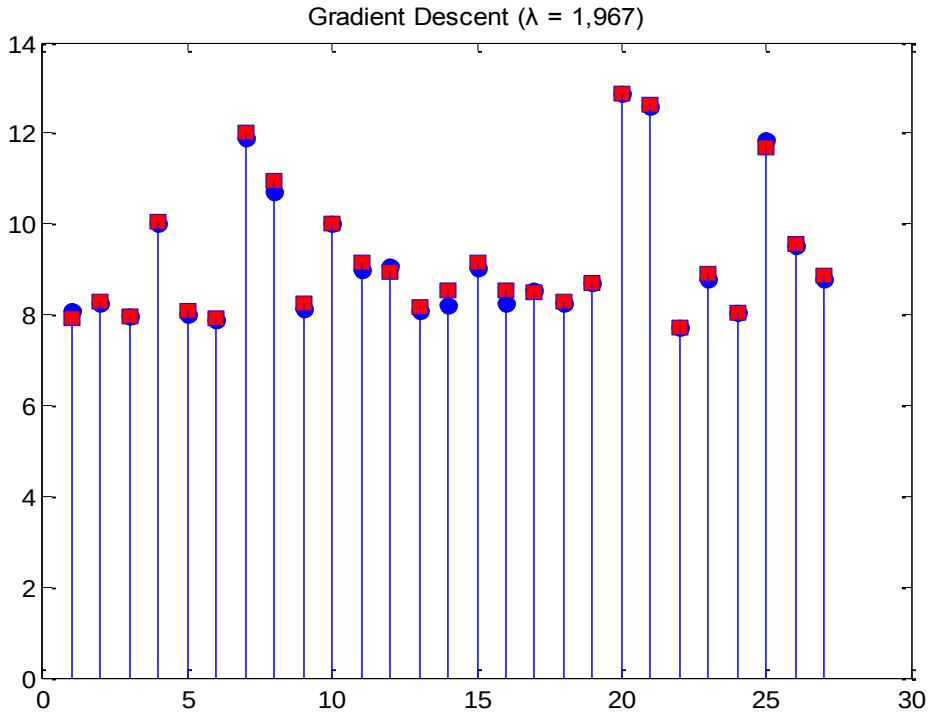
Σχήμα 4.5. Σύγκριση Μειγμάτων που προέκυψαν από τη μέθοδο Gradient Descent και την lsqlin, για λ = 10 (χρήση στο Gradient Descent), ***για την αρχικοποίηση*** 1 / 27



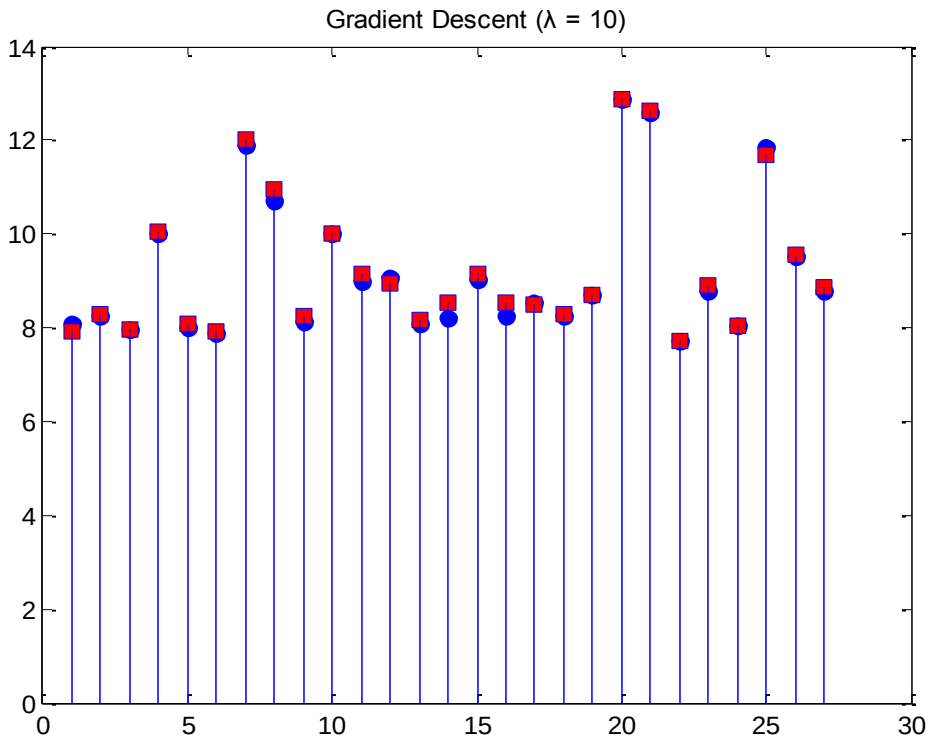
Σχήμα 4.6. Σύγκριση Μειγμάτων που προέκυψαν από τη μέθοδο Gradient Descent και την lsqlin, για λ = 1,967 (χρήση στο Gradient Descent), ***για αρχικοποίηση*** 9,1194 .



Σχήμα 4.7. Σύγκριση Μειγμάτων που προέκυψαν από τη μέθοδο Gradient Descent και την lsqlin, για λ = 10 (χρήση στο Gradient Descent), ***για αρχικοποίηση*** 9,1194 .



Σχήμα 4.8. Σύγκριση Μειγμάτων που προέκυψαν από τη μέθοδο Gradient Descent και την lsqlin, για λ = 1,967 (χρήση στο Gradient Descent), ***για αρχικοποίηση*** ≥ 350 .



Σχήμα 4.9. Σύγκριση Μειγμάτων που προέκυψαν από τη μέθοδο Gradient Descent και την lsqlin, για λ = 10 (χρήση στη μέθοδο Gradient Descent), ***για αρχικοποίηση*** ≥ 350 .

Όπως φαίνεται λοιπόν από τις γραφικές παραστάσεις των ζευγών μειγμάτων, αλλά και από τις τιμές των διασπορών και των μέσων τιμών, *τα μείγματα που προκύπτουν από την αναλυτική εφαρμογή του αλγορίθμου και από την εντολή lsqlin ταυτίζονται αρκετά*. Οι διαφορές είναι μικρές και αποτυπώνονται όπως είδαμε παραπάνω στις διαφορετικές τιμές των διασπορών τους. Μπορούμε λοιπόν να πούμε ότι η συμπεριφορά τους όσον αφορά την ομοιομορφία, θα είναι πολύ παρόμοια, *άρα τα αποτελέσματά μας δεν επηρεάζονται από τον τρόπο με τον οποίο θα υλοποιηθεί το πρόβλημα των Ελαχίστων Τετραγώνων*.
Αν βέβαια θέλαμε να εξετάσουμε τις μικρές διαφορές που υπάρχουν μεταξύ των μειγμάτων που δημιουργούνται από τους δύο τρόπους υλοποίησης θα καταλήγαμε στα εξής: πρώτον, οι τιμές των διασπορών που αντιστοιχούν στην εντολή lsqlin είναι μικρότερες, άρα η υλοποίηση με χρήση της εντολής οδηγεί σε λίγο καλύτερα μείγματα ως προς την ομοιομορφία. Δεύτερον, για τον αλγόριθμο Gradient Descent καλύτερα αποτελέσματα όσον αφορά την ομοιομορφία έχουμε για την τιμή του πολλαπλασιαστή λ = 1,967 (μικρότερες διασπορές).

- **Μέθοδος Ελαχιστοποίησης Εντροπίας**

Για την εξαγωγή των αποτελεσμάτων και των αντίστοιχων συμπερασμάτων για το πρόβλημα Ελαχιστοποίησης της Εντροπίας ακολουθήσαμε την ίδια διαδικασία που ακολουθήθηκε για το πρόβλημα των Ελαχίστων Τετραγώνων. Δημιουργήσαμε τα δύο μείγματα που προκύπτουν από τον αλγόριθμο Gradient Descent, για τις δύο τιμές του πολλαπλασιαστή Lagrange λ και το μείγμα που προέκυψε από την εντολή lsqnonlin. Τα

μείγματα μεγέθους (27×1) παράχθησαν πολλαπλασιάζοντας τον πίνακα δεδομένων C_{final} (27×22) με το διάνυσμα των αναλογιών a (22×1), δηλαδή $C_{final} * \underline{a} = A_{final}$, αφού πρώτα δεσμεύσαμε μνήμη στο Matlab για το καθένα απ' τα δύο. Έπειτα υπολογίσαμε τη μέση τιμή και τη διασπορά κάθε μείγματος, χρησιμοποιώντας τις αντίστοιχες εντολές του Matlab. Στον ***Πίνακα 3.10*** παρουσιάζονται οι τιμές των αναλογιών που προέκυψαν από την ελαχιστοποίηση της συνάρτησης της εντροπίας και των αντίστοιχων πρότυπων μειγμάτων. Στον ***Πίνακα 3.11*** παρουσιάζονται οι τιμές των διασπορών και των μέσων τιμών.

Gradient Descent				**Lsqnonlin**	
λ = 1,967		**λ = 10**			
Αναλογίες a_i	**Μείγμα**	**Αναλογίες** a_i	**Μείγμα**	**Αναλογίες** a_i	**Μείγμα**
2,99E-05	7,96E+00	2,15E-05	7,9653524	0,01177923	8,04394812
9,71E-05	8,35E+00	2,68E-04	8,3289769	0,1287787	8,13604551
3,08E-01	8,0517518	2,32E-01	8,0235167	0,23886919	8,03716361
3,94E-04	9,86E+00	2,15E-04	9,8170664	1,1326E-05	9,91747529
1,88E-04	8,07E+00	1,81E-04	8,0723589	0,06920298	7,93053998
3,94E-04	7,82E+00	1,04E-04	7,8216211	0,01121292	7,80944425
3,19E-05	1,16E+01	2,49E-05	11,639108	1,1326E-05	11,6478865
3,94E-04	1,04E+01	1,87E-04	10,330713	1,1326E-05	11,222975
9,07E-05	8,19E+00	1,72E-05	8,1746094	0,02299215	8,07877238
8,94E-05	1,00E+01	1,93E-04	10,023942	0,13059089	10,0307166
1,01E-04	9,09E+00	2,27E-04	9,1093728	1,1326E-05	8,97932123
2,90E-04	8,92E+00	1,05E-04	8,9512691	0,03714988	9,11176971
2,56E-04	8,11E+00	2,28E-05	8,1196697	1,1326E-05	8,07044112
1,10E-04	8,25E+00	2,83E-04	8,2473417	0,11609337	7,99304834
1,62E-04	9,31E+00	3,94E-04	9,3427704	0,18450352	9,02161452
1,96E-01	8,3874538	2,24E-01	8,3935225	0,04870259	8,36680942
2,10E-01	8,5026736	2,09E-01	8,4884009	1,1326E-05	8,50290735
1,21E-04	8,24E+00	9,09E-05	8,2284149	1,1326E-05	8,24074334
5,76E-02	9,1678967	9,43E-02	9,1289175	1,1326E-05	9,03258001
1,42E-04	1,24E+01	3,94E-04	12,331798	1,1326E-05	13,0776877
6,34E-05	1,28E+01	3,94E-04	12,693683	1,1326E-05	12,5077816
2,26E-01	7,7230958	2,38E-01	7,7338316	1,1326E-05	7,70050124
	8,9724319		8,9404551		8,82269729
	8,0484334		8,0491995		8,00790487
	11,840291		11,842851		11,8788695
	9,4887328		9,5011396		9,53911097
	9,0418796		9,0236111		8,81582097

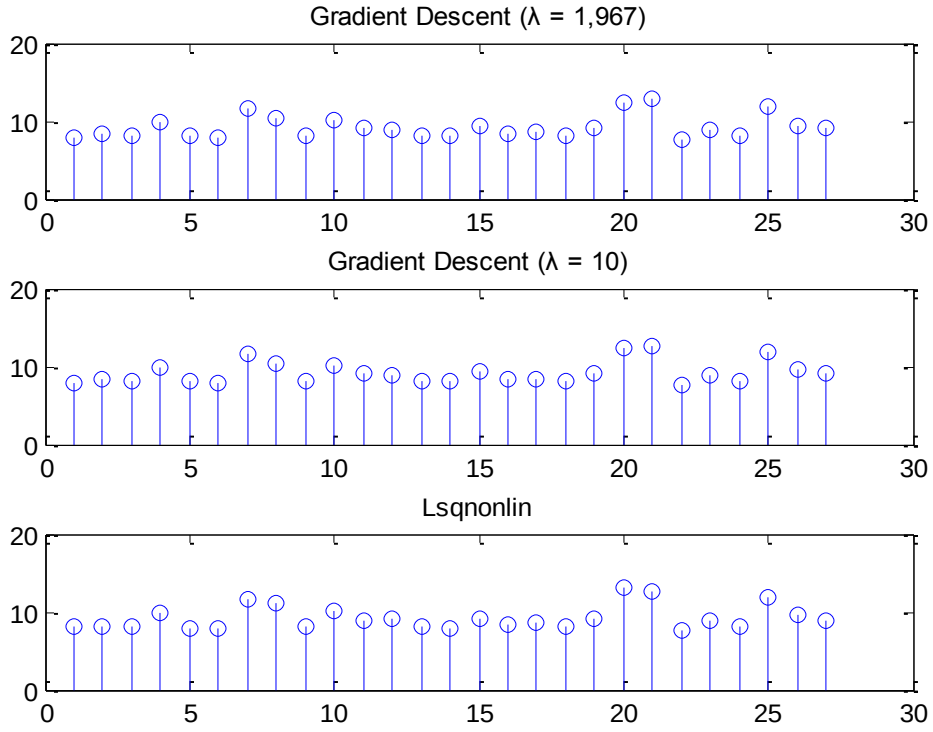
Πίνακας 4.7. Κανονικοποιημένες Αναλογίες και Πρότυπα Μείγματα

	Gradient Descent		Lsqnonlin
	λ = 1,967	**λ = 10**	
Μέση Τιμή	9,2065219	9,1971679	9,2046139
Διασπορά	2,0766857	2,0426964	2,3880722

Πίνακας 4.8. Μέσες Τιμές και Διασπορές των 3 Μειγμάτων

Από τον πίνακα που περιλαμβάνει τα αποτελέσματα για τις μέσες τιμές και τις διασπορές μπορούμε να βγάλουμε και πάλι κάποια συμπεράσματα για την υπάρξη ή όχι ομοιομορφίας των αντίστοιχων μειγμάτων. Εξετάζοντας λοιπόν συνολικά τις μέσες τιμές συμπεραίνουμε ότι και τα τρία μείγματα που προκύπτουν έχουν *σχεδόν ίσες μέσες τιμές*. Επίσης, ελέγχοντας τις διασπορές παρατηρούμε ότι όλες *κυμαίνονται γύρω από την ίδια τιμή* (≈ 2). Η τιμή αυτή μπορεί να μην είναι 0 (πλήρης ομοιομορφία), αλλά είναι αρκετά μικρή, ώστε να εξασφαλίζει *μία καλή ομοιομορφία* για τα πρότυπα μείγματα που δημιουργήθηκαν.

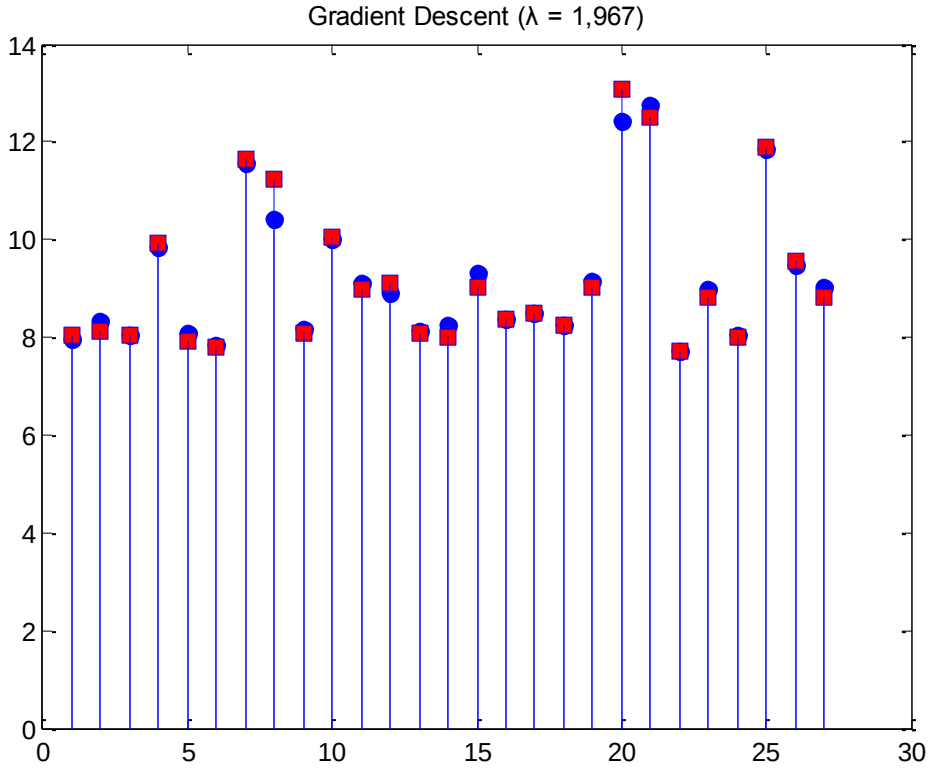
Στη συνέχεια παραθέτουμε σε ένα σχήμα, τις γραφικές παραστάσεις των πρότυπων μειγμάτων που προέκυψαν από τον αλγόριθμο Gradient Descent (2 γραφικές για τους 2 διαφορετικούς πολλαπλασιαστές λ) και από την εντολή του Matlab lsqnonlin (1 γραφική). Οι γραφικές δημιουργήθηκαν με τη χρήση της εντολής *stem* ενώ η ταυτόχρονη παράθεση τους στο ίδιο σχήμα, έγινε χρησιμοποιώντας την εντολή του Matlab *subplot*.



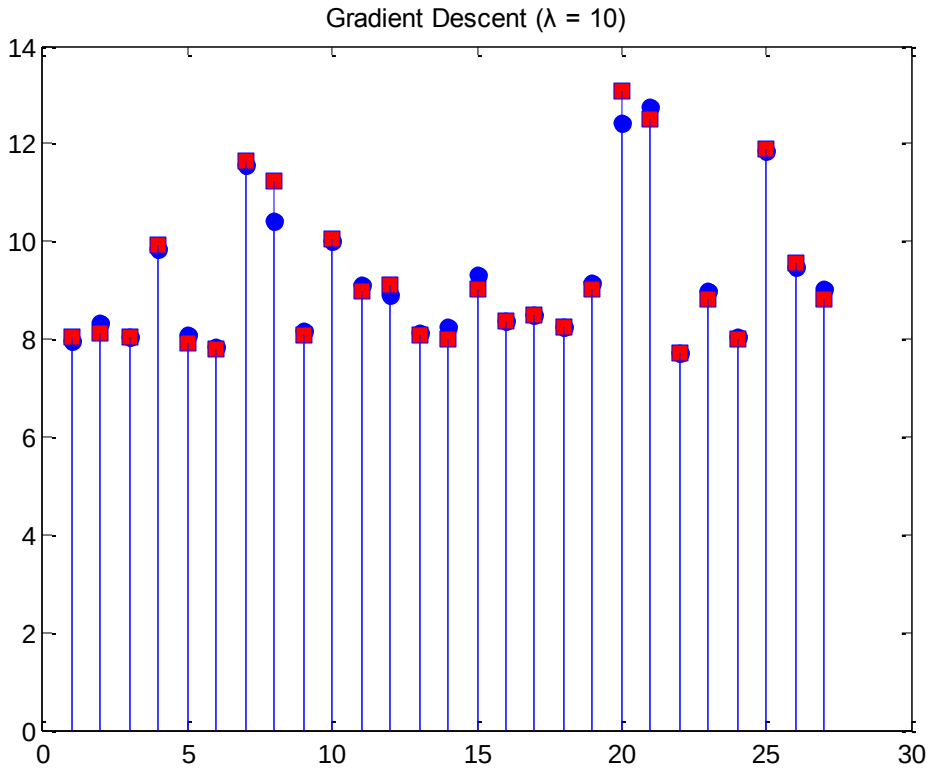
Σχήμα 4.10. Γραφικές Παραστάσεις Μειγμάτων για την περίπτωση της Εντροπίας

Η γραφική παράσταση κάθε μείγματος παρουσιάζει μεγάλη ομοιότητα με τη θεωρητική γραφική παράσταση που περιγράφει την ομοιόμορφη κατανομή, της οποίας όλα στα στοιχεία λαμβάνουν την ίδια τιμή στον άξονα των y, σχηματίζοντας μία ευθεία γραμμή, παράλληλη με τον άξονα των x. *Δηλαδή μπορούμε να συμπεράνουμε και με γραφικό τρόπο, ότι κάθε ένα από τα μείγματα αναφοράς ικανοποιεί την απαίτηση της ομοιομορφίας*, μιας και τα περισσότερα στοιχεία κάθε μείγματος κυμαίνονται γύρω από την ίδια τιμή και μόνο λίγα παρουσιάζουν αποκλίσεις. Οι μικρές αυτές αποκλίσεις των λίγων αυτών στοιχείων είναι ο λόγος που η διασπορά είναι λίγο μεγαλύτερη και όχι ίση ή σχεδόν ίση με το μηδέν.

Τέλος, προσπαθήσαμε να ελέγξουμε εάν τα μείγματα που προκύπτουν από την εφαρμογή του αλγορίθμου Gradient Descent παρουσιάζουν ίδια ή παρόμοια συμπεριφορά με αυτά που προκύπτουν από τη χρήση της εντολής lsqnonlin. Στο πρώτο σχήμα παραθέτουμε τη γραφική παράσταση του ***ζεύγους μειγμάτων*** $\boldsymbol{A_{GD_\lambda = 1,967}}$ **-** $\boldsymbol{A_{lsqnonlin}}$ και στη δεύτερη του ***ζεύγους*** $\boldsymbol{A_{GD_\lambda = 10}}$ **-** $\boldsymbol{A_{lsqnonlin}}$. Τα μείγματα εμφανίζονται στο ίδιο σχήμα, με τη βοήθεια της εντολής *hold* του Matlab. Κάθε μείγμα δημιουργείται και πάλι με την εντολή *stem*, αλλά για να μπορέσουμε να τα ξεχωρίσουμε μιας και βρίσκονται στο ίδιο σχήμα, τους προσδίδουμε διαφορετικό χρώμα (μπλε για Gradient Descent και κόκκινο για lsqnonlin).



Σχήμα 4.11. Σύγκριση Μειγμάτων που προέκυψαν από τη μέθοδο Gradient Descent και την lsqnonlin, για λ = 1,967 (χρήση στο Gradient Descent).



Σχήμα 4.12. Σύγκριση Μειγμάτων που προέκυψαν από τη μέθοδο Gradient Descent και την lsqnonlin, για λ = 10 (χρήση στη μέθοδο Gradient Descent).

Η μόνη σύγκριση που μπορεί να γίνει στην περίπτωση του προβλήματος της εντροπίας (εφόσον δεν απαιτείται να υπάρχει αρχικοποίηση του target distribution, ώστε να έχουμε διαφορετικές περιπτώσεις) είναι μεταξύ του αλγορίθμου Gradient Descent και της εντολής lsqnonlin. Θα μπορούσαμε να πούμε ότι οι δύο τρόποι υλοποίησης του προβλήματος της εντροπίας οδηγούν σε μείγματα με πολύ παρόμοια συμπεριφορά ως προς την ομοιομορφία. *Παρατηρώντας όμως καλύτερα τις διαφορές των διασπορών συμπεραίνουμε ότι ο αλγόριθμος Gradient Descent οδηγεί στην παραγωγή μειγμάτων αναφοράς με καλύτερη ομοιομορφία.* Τέλος, και για τις *δύο διαφορετικές τιμές του πολλαπλασιαστή λ*, τα αποτελέσματα για τις διασπορές είναι ίδια, οπότε μπορούμε να πούμε ότι και τα μείγματα που αντιστοιχούν στις δύο αυτές περιπτώσεις *έχουν ίδια συμπεριφορά ως προς την ομοιομορφία.*

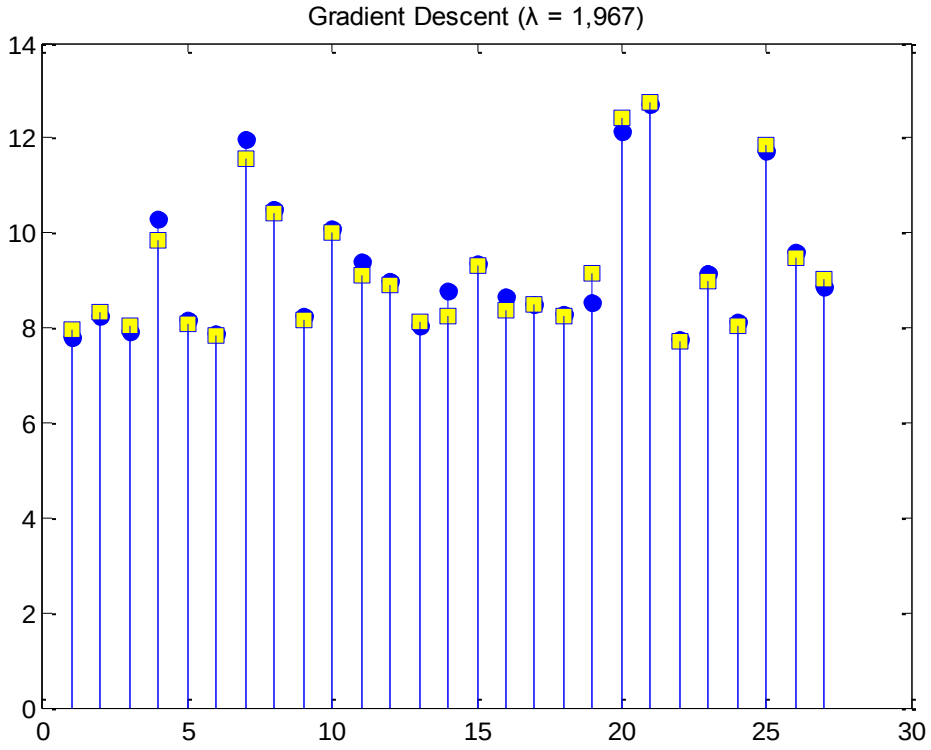
❖ **Σύγκριση Αποτελεσμάτων των δύο Προβλημάτων**

Η χρήση δύο διαφορετικών προβλημάτων ελαχιστοποίησης για την εύρεση των βέλτιστων αναλογιών πρόσμειξης πραγματοποιήθηκε για να διαπιστώσουμε ποια από τις δύο μεθόδους μας παρέχει καλύτερα αποτελέσματα όσον αφορά την ομοιομορφία του πρότυπου μείγματος αναφοράς που θέλουμε να δημιουργήσουμε. Πρόκειται για ένα γραμμικό πρόβλημα, αυτό των ελαχίστων τετραγώνων και ένα μη γραμμικό πρόβλημα, αυτό της ελαχιστοποίησης της εντροπίας. Αυτό που πρέπει να κάνουμε είναι να συγκρίνουμε τα μείγματα που προκύπτουν από κάθε μέθοδο μέσω των γραφικών τους παραστάσεων, συμπεριλαμβανομένων και των μέσων τιμών και διασπορών τους. Επειδή

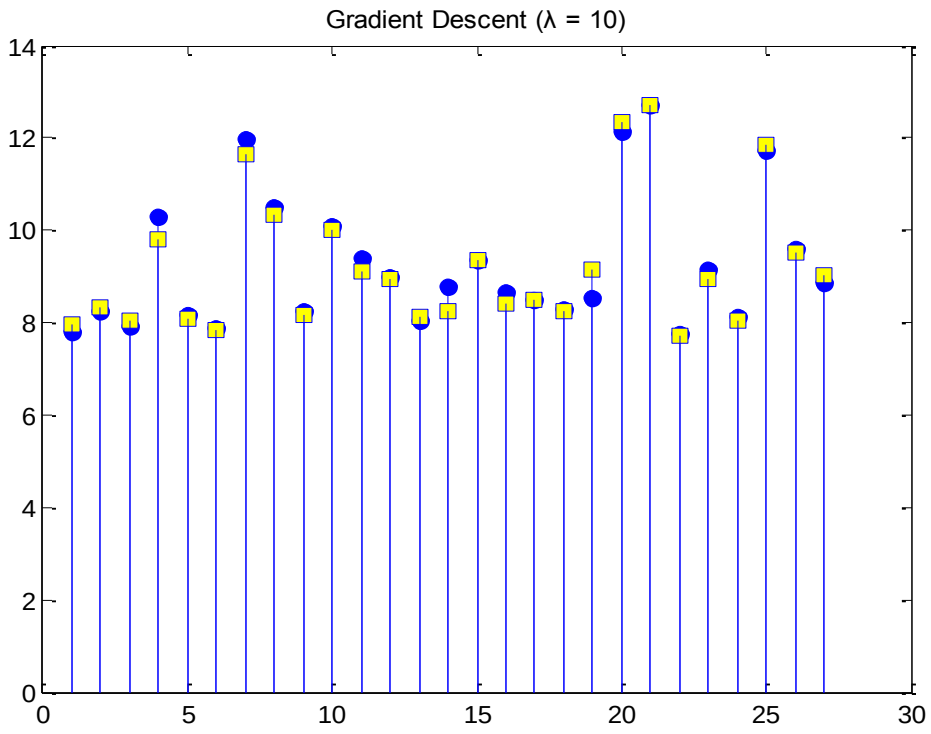
για το πρόβλημα ελαχίστων τετραγώνων εξετάζουμε 3 διαφορετικές περιπτώσεις αρχικοποίησης, ενώ για το πρόβλημα της εντροπίας έχουμε μία περίπτωση, θα συγκρίνουμε τη μία αυτή περίπτωση με καθεμία απ' αυτές των ελαχίστων τετραγώνων ξεχωριστά. Τα οποιαδήποτε συμπεράσματα για το πιο από τα δύο προβλήματα οδηγεί σε καλύτερα αποτελέσματα, θα προκύψουν από την παραπάνω σύγκριση.

		Gradient Descent		**Lsqlin - Lsqnonlin**
		λ = 1,967	**λ = 10**	
Εντροπία	**Μέση Τιμή**	9,2065219	9,1971679	9,2046139
	Διασπορά	2,0766857	2,0426964	2,3880722
Ελάχιστα Τετράγωνα	**Μέση Τιμή**	9,2464	9,2213	9,035697
	Διασπορά	2,0697217	2,2530547	2,184243

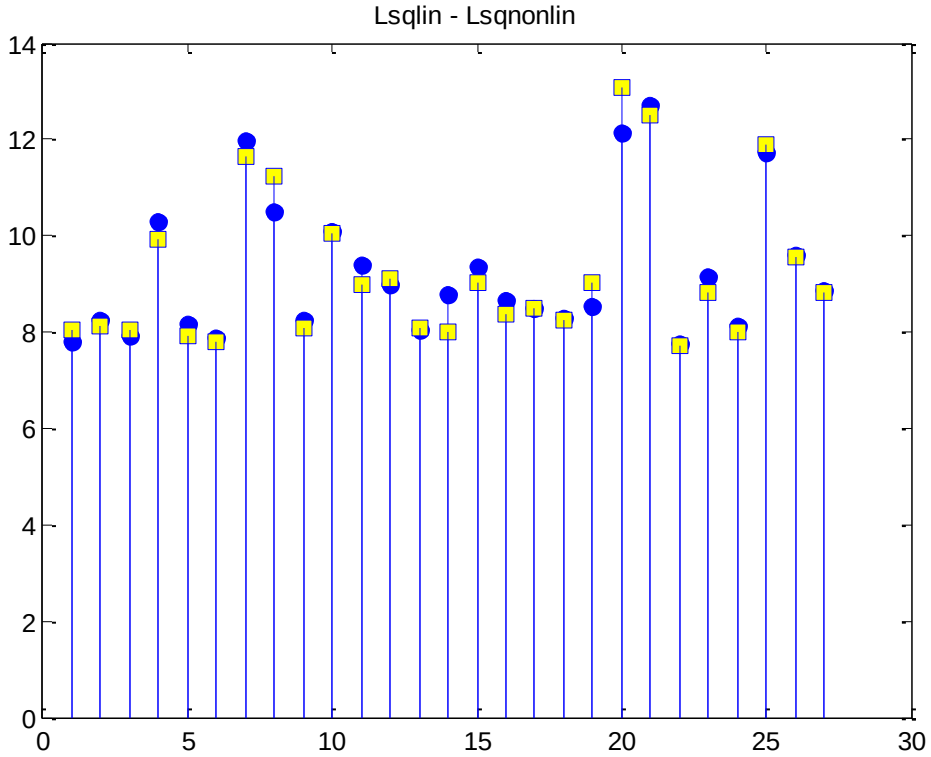
Πίνακας 4.9. Σύγκριση Μέσων Τιμών & Διασπορών Εντροπίας – Ελαχίστων Τετραγώνων ($A_{init} = 1/27$)



Σχήμα 4.13. Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, **λ = 1,967** ($A_{init} = 1/27$)



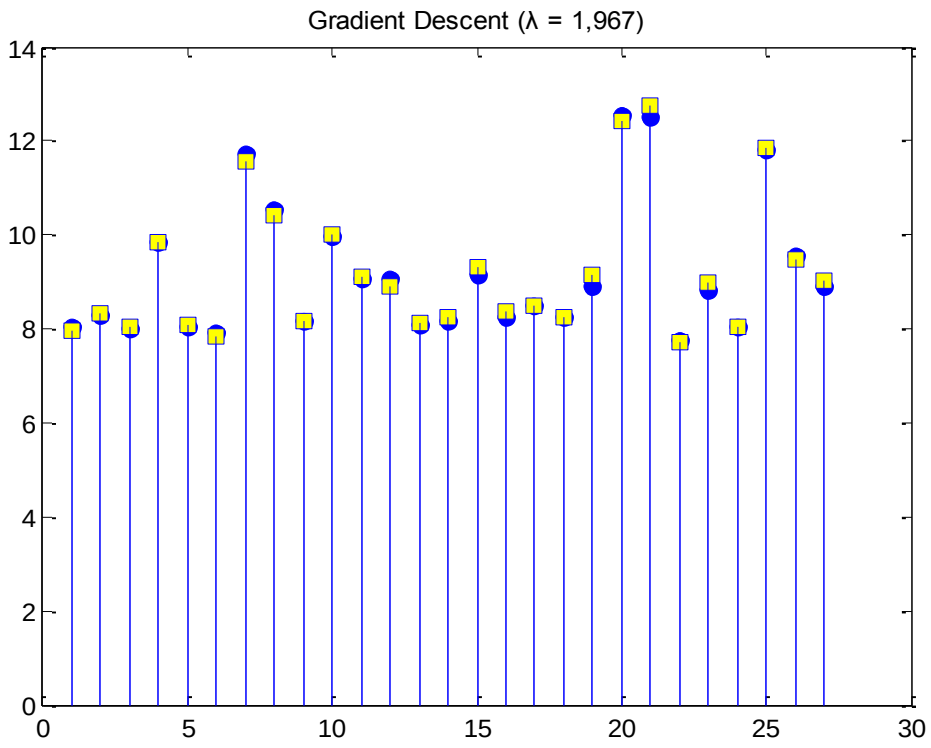
Σχήμα 4.14. Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, **λ = 10** ($A_{init} = 1/27$)



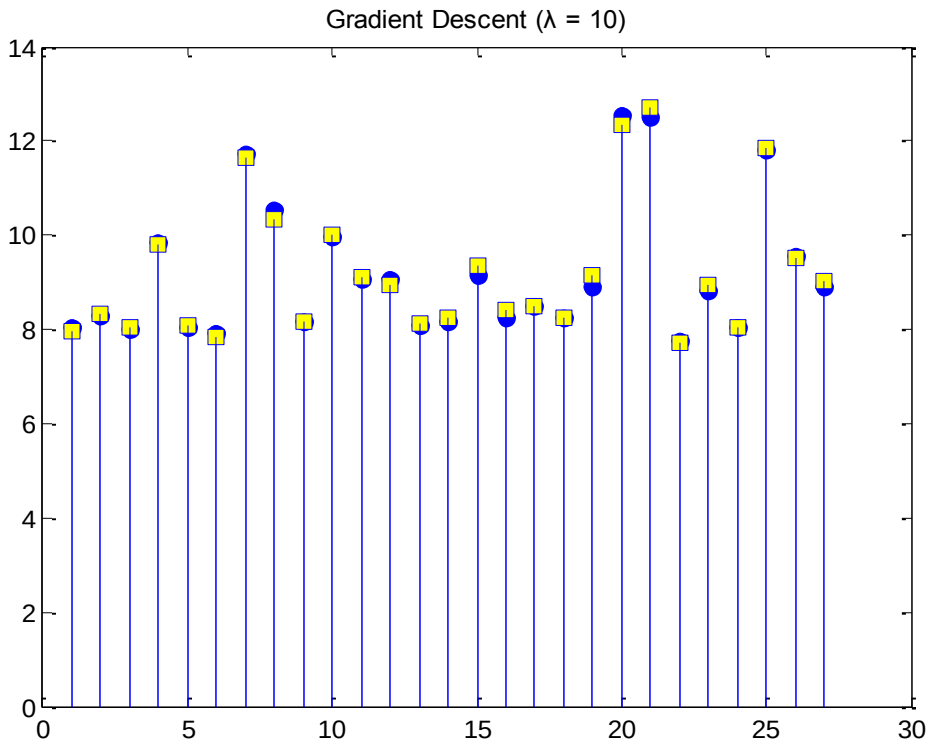
Σχήμα 4.15. Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, **Εντολές** ($A_{init} = 1/27$)

		Gradient Descent		Lsqlin - Lsqnonlin
		λ = 1,967	**λ = 10**	
Εντροπία	**Μέση Τιμή**	9,2065219	9,1971679	9,2046139
	Διασπορά	2,0766857	2,0426964	2,3880722
Ελάχιστα Τετράγωνα	**Μέση Τιμή**	9,184957	9,1999463	9,171897
	Διασπορά	2,103013	2,2905617	1,951897

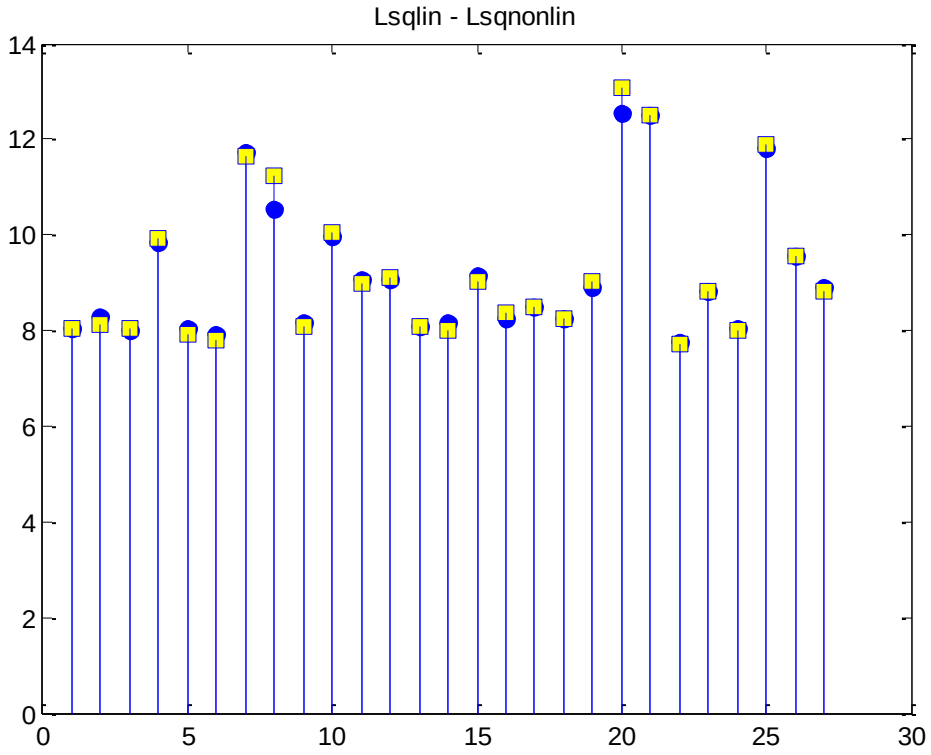
Πίνακας 4.10. Σύγκριση Μέσων Τιμών & Διασπορών Εντροπίας – Ελαχίστων Τετραγώνων (A_{init} = 9,1194)



Σχήμα 4.16. Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, **λ = 1,967** (A_{init} = 9,1194)



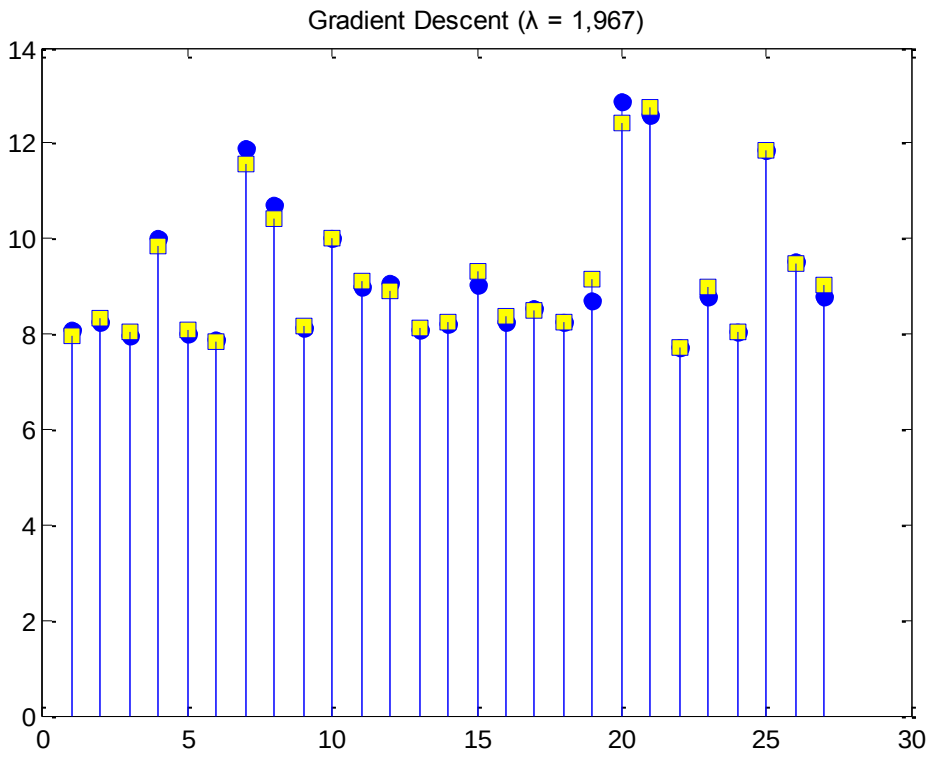
Σχήμα 4.17. Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, **λ = 10** (A_{init} = 9,1194)



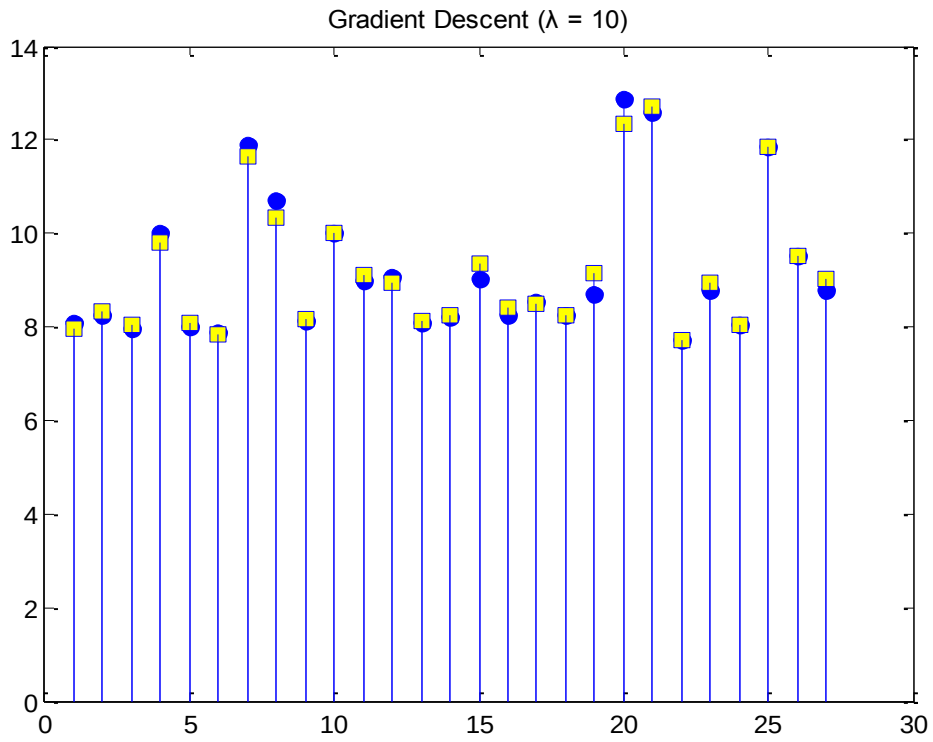
Σχήμα 4.18. Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, **Εντολές** (A_{init} = 9,1194)

		Gradient Descent		Lsqlin - Lsqnonlin
		λ = 1,967	**λ = 10**	
Εντροπία	**Μέση Τιμή**	9,2065219	9,1971679	9,2046139
	Διασπορά	2,0766857	2,0426964	2,3880722
Ελάχιστα Τετράγωνα	**Μέση Τιμή**	9,1930058	9,207842	9,247747
	Διασπορά	2,2988271	2,3634005	2,271283

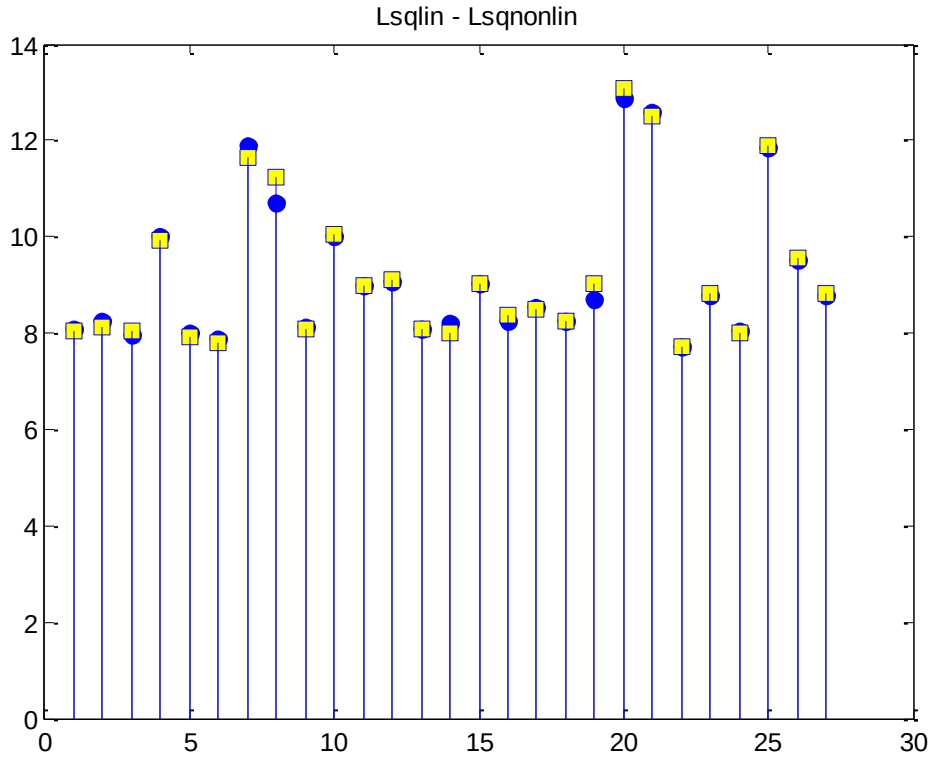
Πίνακας 4.11. Σύγκριση Μέσων Τιμών & Διασπορών Εντροπίας – Ελαχίστων Τετραγώνων ($A_{init} \geq 350$)



Σχήμα 4.19. Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, **λ = 1,967** ($A_{init} \geq 350$)



Σχήμα 4.20. Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, **λ = 10** ($A_{init} \geq 350$)



Σχήμα 4.21. Σύγκριση Ελαχίστων Τετραγώνων – Εντροπίας, **Εντολές** ($A_{init} \geq 350$)

Συγκρίνοντας τα αντίστοιχα μείγματα για καθεμία από τις 3 περιπτώσεις αρχικοποίησης μειγμάτων παρατηρούμε ότι τα στοιχεία κάθε μείγματος έχουν σχεδόν ίδιες τιμές και για τα δύο προβλήματα που εξετάσαμε. Αυτό φαίνεται από τα αντίστοιχα ζεύγη γραφικών παραστάσεων που παρουσιάσαμε παραπάνω, αλλά και από τις μέσες τιμές που είναι σχεδόν ίσες για όλα τα μείγματα. Θα μπορούσαμε λοιπόν να πούμε ότι τα μείγματα που προκύπτουν έχουν πολλά κοινά σημεία και η συμπεριφορά τους όσον αφορά την ομοιομορφία είναι αρκετά παρόμοια. Θα μπορούσαν επομένως όλα να αποτελέσουν πρότυπα μείγματα αναφοράς. Αν όμως θέλαμε να καταλήξουμε στο πρόβλημα (Ελάχιστα Τετράγωνα ή Εντροπία) που δημιουργεί το μείγμα που έχει την καλύτερη συμπεριφορά όσον αφορά την ομοιομορφία και επομένως αυτό που πλησιάζει περισσότερο το ιδανικό, θα έπρεπε να ελέγξουμε τυχόν διαφορές στις διασπορές τους. Παρατηρώντας συνολικά τις τιμές των διασπορών για τα δύο προβλήματα βλέπουμε ότι οι τιμές που προκύπτουν από το πρόβλημα Ελαχιστοποίησης της Εντροπίας είναι μικρότερες. Αυτό συνεπάγεται *ότι το πρόβλημα ελαχιστοποίησης της εντροπίας οδηγεί σε καλύτερα αποτελέσματα και τελικά σε μείγματα αναφοράς με μεγαλύτερη ομοιομορφία, απ' αυτά του αντίστοιχου προβλήματος των Ελαχίστων Τετραγώνων.*

4.2 Έλεγχος Αποτελεσμάτων για το Σύνολο των Δεδομένων

Όπως αναφέρθηκε και στο πρώτο κεφάλαιο της υλοποίησης, όλη η παραπάνω διαδικασία δεν πραγματοποιήθηκε για το σύνολο των γονιδίων, δηλαδή για τα 34772 γονίδια, αλλά για ένα μικρότερο, αντιπροσωπευτικό αριθμό γονιδίων, τα 27, τα οποία προέκυψαν από τη χρήση της συσχέτισης για τη μείωση της διάστασης του πίνακα δεδομένων. Έτσι, τα αποτελέσματα που προέκυψαν για το πρότυπο μείγμα που θέλουμε να δημιουργήσουμε, αφορούν τα 27 γονίδια. Πρέπει όμως να είμαστε σίγουροι ότι η επιλογή των 27 γονιδίων αυτών είναι όντως αντιπροσωπευτική για το σύνολο των γονιδίων. Χρειάζεται δηλαδή να γνωρίζουμε ότι με τη μείωση του αριθμού των γονιδίων δεν έχει χαθεί σημαντική πληροφορία από τα δεδομένα μας, πράγμα το οποίο μπορεί να επηρεάσει τα αποτελέσματά μας, αλλά και ότι τα συγκεκριμένα γονίδια που επιλέχθησαν θα οδηγήσουν στα ίδια αποτελέσματα, με το να είχαμε χρησιμοποιήσει το σύνολο των γονιδίων.

Αντί να εφαρμόσουμε από την αρχή όλους τους αλγορίθμους που προαναφέρθηκαν για το σύνολο των δεδομένων, πράγμα το οποίο θα ήταν χρονοβόρο, μπορούμε να πραγματοποιήσουμε τον έλεγχο αυτό με τον τρόπο που αναλύεται παρακάτω:

Δημιουργούμε το πρότυπο μείγμα $A_{all\ data}(34772 \times 1)$ για το σύνολο πλέον των δεδομένων, το οποίο προκύπτει απ' τη σχέση $\boxed{A_{all\ data} = C_{initial} * a_i}$. Πολλαπλασιάζουμε δηλαδή τον αρχικό πίνακα δεδομένων (πίνακας όπου οι ενδείξεις NaN έχουν αντικατασταθεί) $C_{initial}(34772 \times 22)$, τον οποίο έχουμε επίσης κανονικοποιήσει χρησιμοποιώντας το δυαδικό λογάριθμο, με τις αναλογίες a_i που προέκυψαν από την εφαρμογή του Gradient Descent, αλλά και των εντολών ελαχιστοποίησης του Matlab, για τα προβλήματα ελαχίστων τετραγώνων και ελαχιστοποίησης εντροπίας (πάντα αναφερόμαστε στις κανονικοποιημένες αναλογίες). Συνεπώς έχουμε και πάλι **9 διαφορετικά μείγματα για τη μέθοδο Ελαχίστων Τετραγώνων και 3 για τη μέθοδο ελαχιστοποίησης της Εντροπίας, για το σύνολο πλέον των δεδομένων**. Τα μείγματα αυτά θα συγκριθούν με τα αντίστοιχα των 2 μεθόδων, για δεδομένα 27 γονιδίων (αναλύθηκαν στο προηγούμενο κεφάλαιο). Έπειτα υπολογίζουμε τη μέση τιμή και τη διασπορά για κάθε ένα από

τα 9 μείγματα $A_{all\ data}$, που δημιουργούνται από το σύνολο των δεδομένων, ενώ ταυτόχρονα δημιουργούμε και τη γραφική παράσταση των τιμών του κάθε μείγματος. Για να μπορέσουμε να πιστοποιήσουμε ότι η χρήση μόνο 27 γονιδίων είναι ταυτόσημη με αυτή του συνόλου των γονιδίων, θα πρέπει η διασπορά και η μέση τιμή για το κάθε μείγμα $A_{all\ data}$ να είναι ίδια ή σχεδόν ίδια με την αντίστοιχη του μείγματος A_{final} (μείγμα 27 γονιδίων με κανονικοποιημένες τιμές). Επιπλέον θα πρέπει η γραφικές παραστάσεις των αντίστοιχων μειγμάτων να ταυτίζονται, καθώς επίσης και να προσεγγίζουν τη μορφή της ευθείας, ώστε να ικανοποιείται και η απαιτούμενη ομοιομορφία του πρότυπου μείγματος.

- **Μέθοδος Ελαχίστων Τετραγώνων**

  Πραγματοποιήσαμε τον έλεγχο για το σύνολο των δεδομένων, για τις αναλογίες πρόσμειξης των κυτταρικών σειρών, που προέκυψαν από τη μέθοδο των ελαχίστων τετραγώνων (Least Squares) και από την εντολή lsqlin. Για κάθε διάνυσμα αναλογιών a_i (22×1) εκτελέσαμε στο Matlab τον πολλαπλασιασμό του με τον πίνακα των 34772 γονιδίων $C_{initial}$ και έτσι δημιουργήσαμε το κάθε μείγμα $A_{all\ data}$. Με τις έτοιμες συναρτήσεις του Matlab, *mean και var* υπολογίσαμε τη μέση τιμή και τη διασπορά αντίστοιχα για το κάθε μείγμα, όπως και για τα δεδομένα των 27 γονιδίων. Επιπλέον, με χρήση της εντολής του Matlab *hist* δημιουργήσαμε το ιστόγραμμα κάθε μείγματος. Σύμφωνα με την στατιστική, το ιστόγραμμα είναι μία γραφική αναπαράσταση συχνοτήτων, οι οποίες αναπαρίστανται με τη μορφή ράβδων (bars). Περιγράφει ποια αναλογία περιπτώσεων αντιστοιχεί σε καθεμία από τις διαφορετικές κατηγορίες. Η σύνταξη της εντολής hist είναι αυτή που φαίνεται παρακάτω

  $$hist\ (Y)\ .$$

  όπου Y είναι η συνάρτηση της οποίας το ιστόγραμμα θέλουμε να δημιουργήσουμε. Η εντολή stem που χρησιμοποιήθηκε για τα δεδομένα 27 γονιδίων, δεν ήταν δυνατό να χρησιμοποιηθεί για το σύνολο των δεδομένων, γιατί οδηγούσε σε ένα μη ευδιάκριτο γράφημα μείγματος, από το οποίο δεν μπορούσε να εξαχθεί κανένα συμπέρασμα για την ομοιομορφία. Για τον καλύτερο και ευκολότερο έλεγχο της ομοιομορφίας των μειγμάτων λοιπόν χρησιμοποιήσαμε το ιστόγραμμα. Όσα περισσότερα στοιχεία αντιστοιχούν στην ίδια περίπτωση, τόσο πιο πολύ προσεγγίζουμε το ιδανικό ιστόγραμμα ομοιόμορφης κατανομής., του οποίου τα στοιχεία ανήκουν όλα στην ίδια κατηγορία. Στην περίπτωση των μειγμάτων μας βέβαια έχουμε αρκετά στοιχεία που βρίσκονται γύρω από την ίδια τιμή , αλλά υπάρχουν και κάποια που παρουσιάζουν αποκλίσεις. Για το λόγο αυτό το ιστόγραμμα που αναμένουμε θα περιμβάνει ράβδους μικρού ύψους, σε κάποιες συχνότητες. Αξίζει να συμπληρώσουμε ότι μιας και οι τιμές των στοιχείων των μειγμάτων κυμαίνονται στο διάστημα [4,18] διαβαθμίσαμε τον άξονα των συχνοτήτων με τιμές από [1,20]. Επίσης, το εύρος (μήκος) κάθε ράβδου του ιστογράμματος ορίστηκε ίσο με 0.1 για πιο ακριβή ανάλυση και κατηγοριοποίηση των τιμών των στοιχείων.

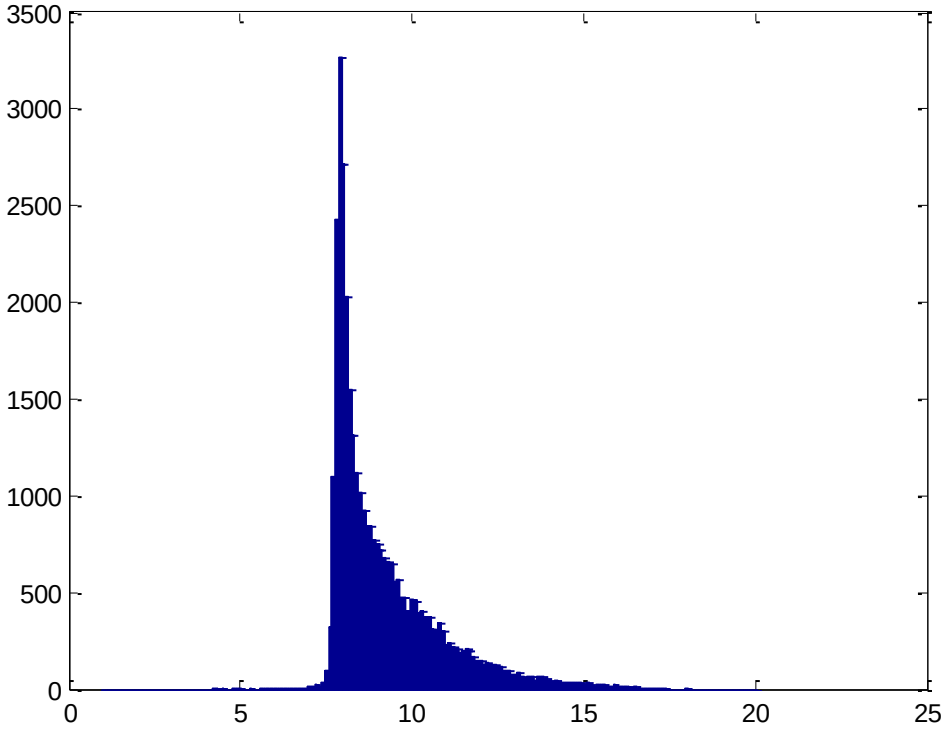
  Τέλος, ο βασικός στόχος όλης της παραπάνω διαδικασίας είναι να διαπιστωθεί το ότι η μείωση της διάστασης των γονιδίων δεν επηρεάζει τα αποτελέσματα για τις αναλογίες πρόσμειξης και συνεπώς για τα μείγματα. Αυτό μπορεί να διαπιστωθεί συγκρίνοντας τις μέσες τιμές και τις διασπορές για όλα τα πιθανά ζεύγη μειγμάτων $A_{all\ data} - A_{final}$. Εάν οι μέσες τιμές και οι διαπορές του συμπίπτουν τότε έχουν την ίδια συμπεριφορά ως προς την ομοιομορφία και άρα οδηγούν σε κοινά αποτελέσματα.

Στους ***Πίνακες 3.15, 3.16, 3.17*** που ακολουθούν, παραθέτουμε τα αποτελέσματα για τις μέσες τιμές και τις διασπορές για τα μείγματα $A_{all\ data}$ και A_{final}, για τις τρεις περιπτώσεις αρχικοποίησης του target distribution. Επίσης, παρακάτω ακολουθούν και τα σχήματα που παρουσιάζουν τα ιστογράμματα των μειγμάτων.

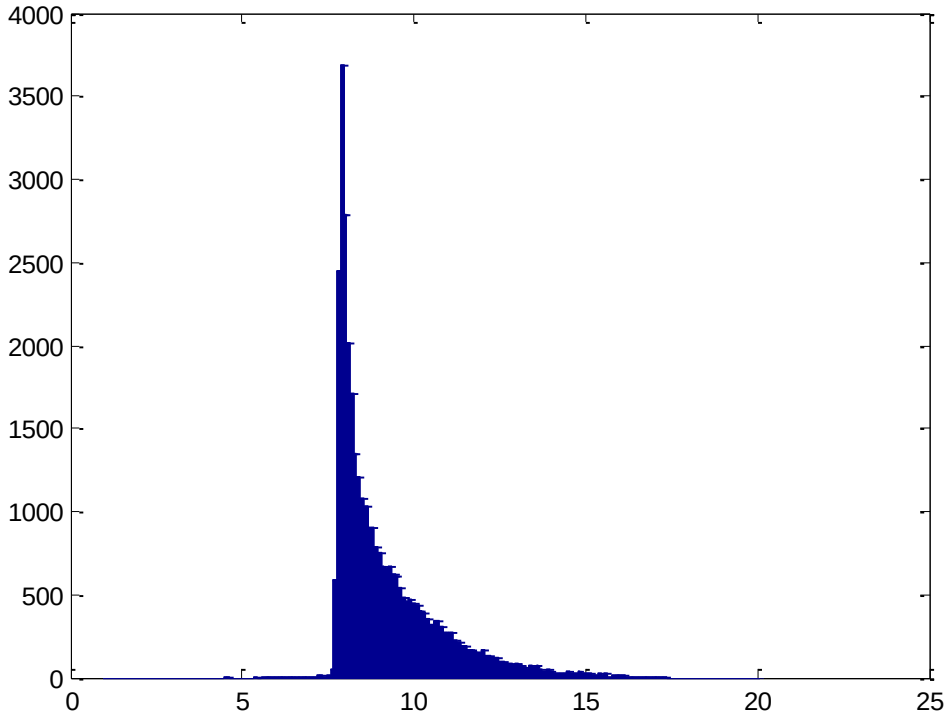
Αρχικοποίηση Διανύσματος $A_{init} = 1/27$

	Grad Desc λ=1,967	**Grad Desc λ=10**	**LSQLIN**
	ALL DATA		
Μέση Τιμή	9,1774	9,1652	9,1308
Διασπορά	2,6399	2,4691	2,5279
	DATA 27x22		
Μέση Τιμή	9,2464	9,2213	9,0357
Διασπορά	2,0697	2,2531	2,1842

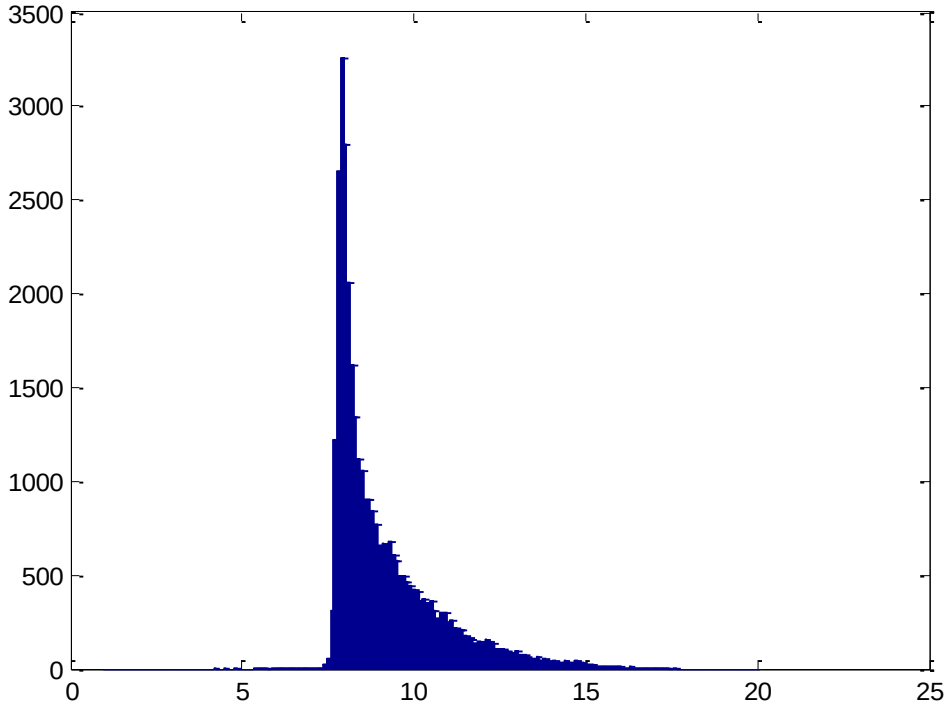
Πίνακας 4.12. Σύγκριση Μέσων Τιμών & Διασπορών για τα 2 set δεδομένων (1 / 27)



Σχήμα 4.22. Ιστόγραμμα Μείγματος για. 34772 γονίδια (λ = 1,967 - A_{init} = 1 / 27)



Σχήμα 4.23. Ιστόγραμμα Μείγματος για 34772 γονίδια (λ = 10 - $A_{init} = 1/27$)

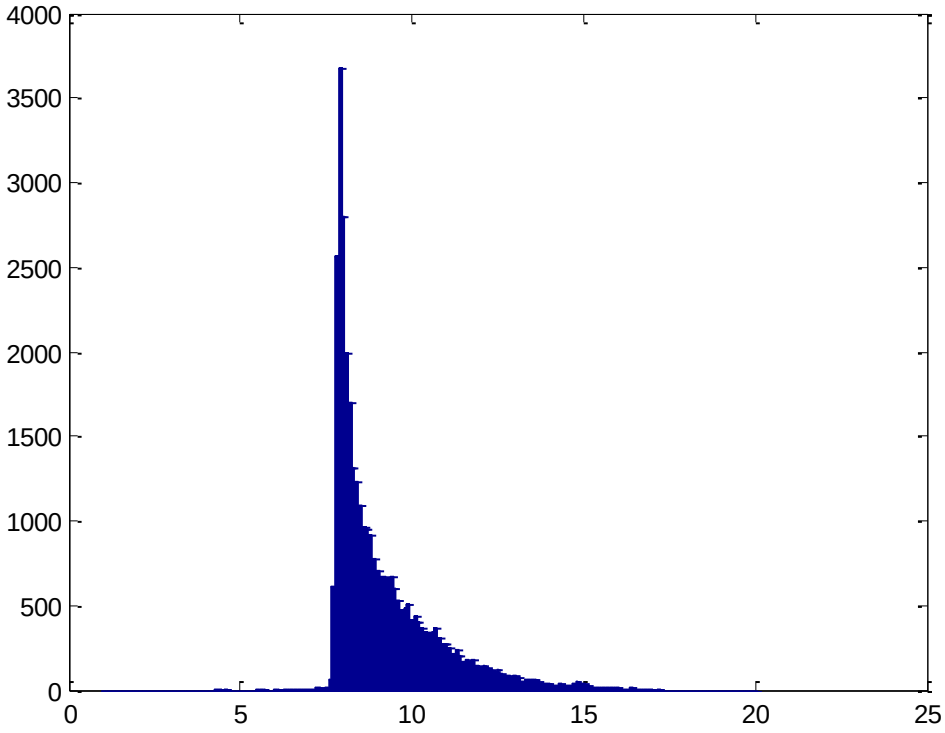


Σχήμα 4.24. Ιστόγραμμα Μείγματος για 34772 γονίδια (Lsqlin - $A_{init} = 1/27$)

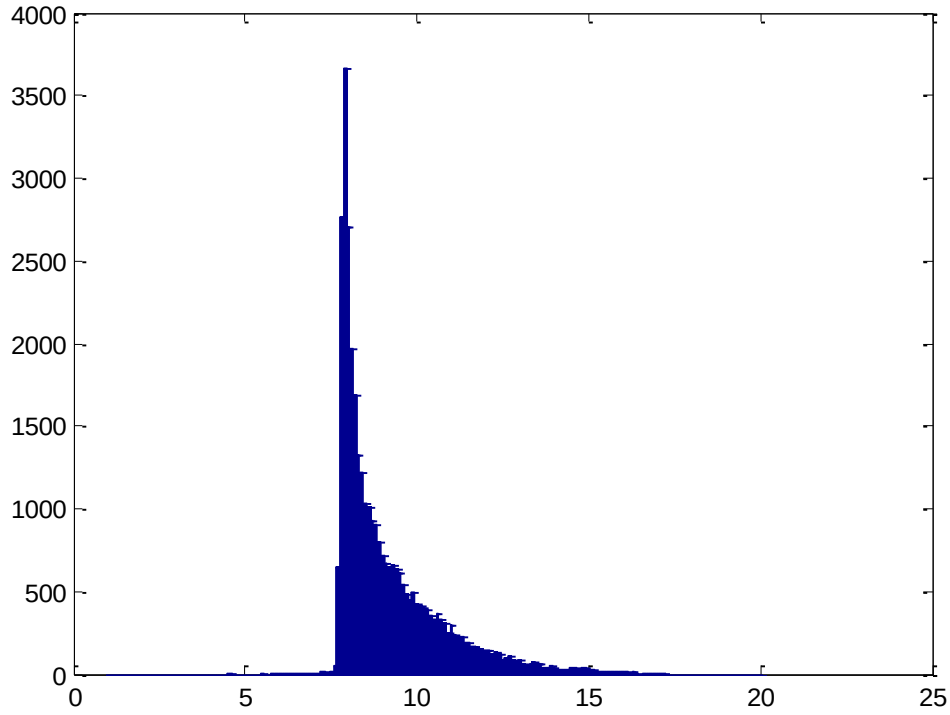
Αρχικοποίηση Διανύσματος A_{init} = 9,1194

	Grad Desc λ=1,967	**Grad Desc λ=10**	**LSQLIN**
	ALL DATA		
Μέση Τιμή	9,1583	9,1493	9,1522
Διασπορά	2,4526	2,4409	2,4816
	DATA 27x22		
Μέση Τιμή	9,1850	9,1999	9,1719
Διασπορά	2,1030	2,2906	1,9519

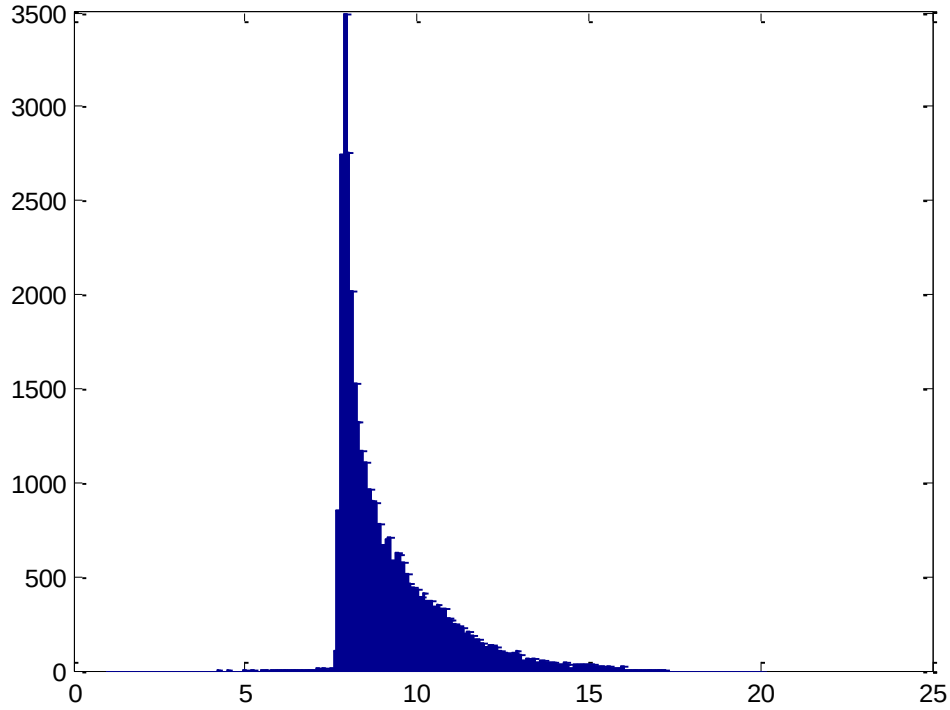
Πίνακας 4.13. Σύγκριση Μέσων Τιμών & Διασπορών για τα 2 set δεδομένων (9,1194)



Σχήμα 4.25. Ιστόγραμμα Μείγματος για 34772 γονίδια (λ = 1,967 - A_{init} = 9,1194)



Σχήμα 4.26. Ιστόγραμμα Μείγματος για 34772 γονίδια (λ = 10 - A_{init} = 9,1194)

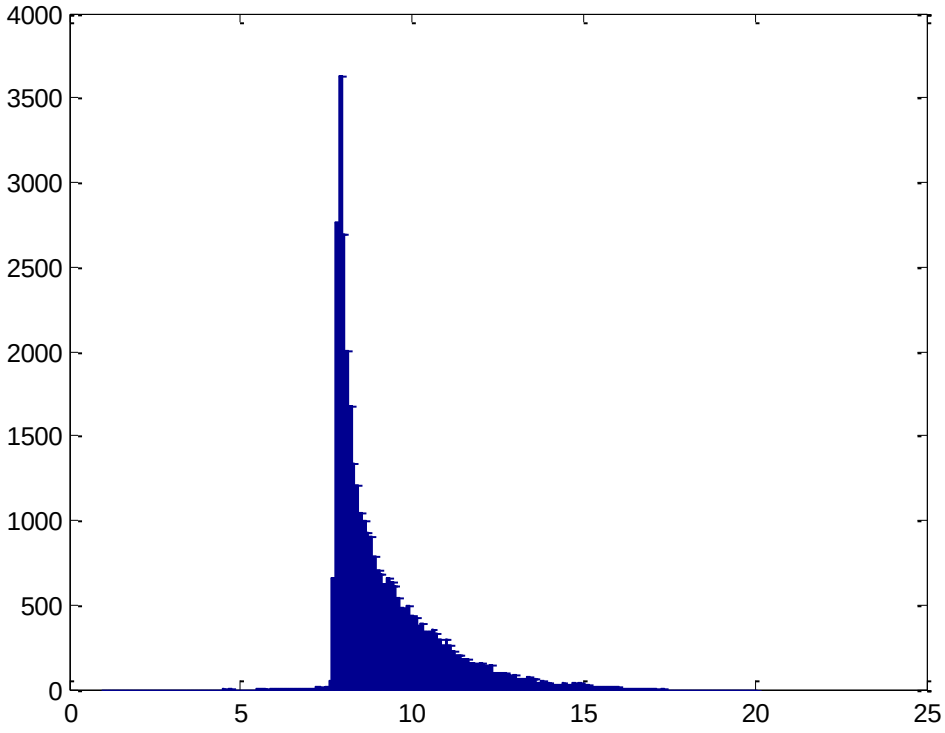


Σχήμα 4.27. Ιστόγραμμα Μείγματος για 34772 γονίδια (Lsqlin - A_{init} = 9,1194)

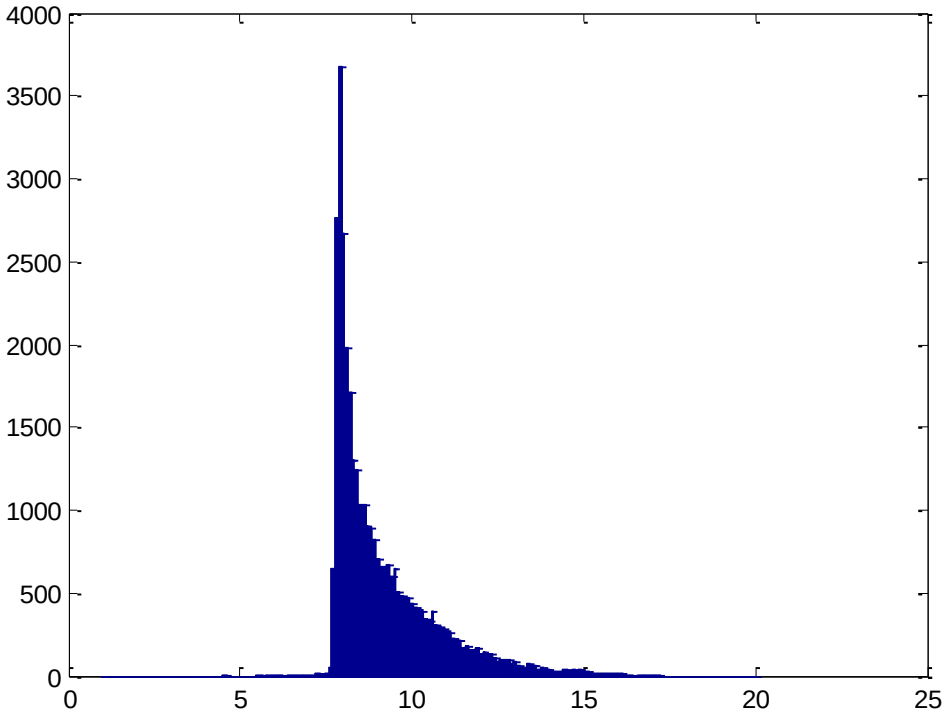
Αρχικοποίηση Διανύσματος $A_{init} \geq 350$ $(\log_2 350 = 8,45)$

	Grad Desc λ=1,967	Grad Desc λ=10	LSQLIN
	ALL DATA		
Μέση Τιμή	9,1503	9,1490	9,1648
Διασπορά	2,4416	2,4407	2,5126
	DATA 27x22		
Μέση Τιμή	9,1930	9,2477	9,2078
Διασπορά	2,2989	2,2713	2,2713

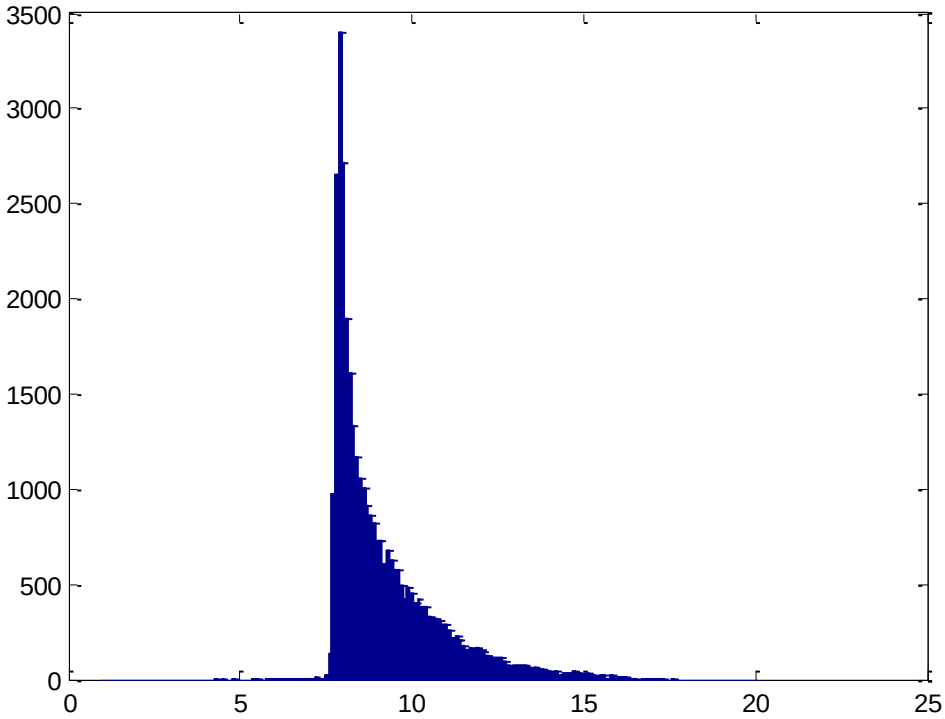
Πίνακας 4.14. Σύγκριση Μέσων Τιμών & Διασπορών για τα 2 set δεδομένων (≥ 350)



Σχήμα 4.28. Ιστόγραμμα Μείγματος για 34772 γονίδια (λ = 1,967 - $A_{init} \geq 350$)



Σχήμα 4.29. Ιστόγραμμα Μείγματος για. 34772 γονίδια (λ = 10 - $A_{init} \geq 350$)



Σχήμα 4.30. Ιστόγραμμα Μείγματος για. 34772 γονίδια (Lsqlin - $A_{init} \geq 350$)

Από τα ιστογράμματα παρατηρούμε ότι το μεγαλύτερο σύνολο των στοιχείων των μειγμάτων και συγκεκριμένα περίπου 13.000 στοιχεία σε ένα εύρος τιμών [7.5, 8.]. Επίσης, ένα σύνολο στοιχείων, αξιόλογου μεγέθους είναι αυτό που περιλαμβάνει περίπου 7000 στοιχεία και αντιστοιχεί σε εύρος [8, 8.5]. Το σύνολο των στοιχείων που αποκλίνουν και έχουν πολύ μικρότερες ή μεγαλύτερες τιμές αποτελείται από 14772 γονίδια. Δεν είναι πολύ μικρό, αλλά είναι μικρότερο από το άθροισμα των στοιχείων που παίρνουν τιμές από 7.5 ως 8.5. Συνεπώς επιτυγχάνεται αρκετά μεγάλη ομοιομορφία, αφού 20000 γονίδια κυμαίνονται γύρω από την ίδια περίπου τιμή.
Επιπλέον παρατηρούμε ότι τόσο η διασπορά, όσο και η μέση τιμή παρουσιάζουν παρόμοιες τιμές για τα δύο διαφορετικά σύνολα δεδομένων. Άρα, τα μείγματα που προκύπτουν για το σύνολο των δεδομένων είναι ισοδύναμα με αυτά που προκύπτουν χρησιμοποιώντας 27 γονίδια. *Μπορούμε δηλαδή να πούμε ότι η χρήση λιγότερων γονιδίων καταλήγει σε ίδια συμπεράσματα, με αυτή του συνόλου των γονιδίων, άρα είναι ασφαλής και δε χάνεται κάποια χρήσιμη πληροφορία*.

- **Μέθοδος Ελαχιστοποίησης Εντροπίας**

Πραγματοποιήσαμε τον έλεγχο για το σύνολο των δεδομένων, για τις αναλογίες πρόσμειξης των κυτταρικών σειρών, που προέκυψαν από το πρόβλημα Ελαχιστοποίησης της Εντροπίας και από την εντολή lsqnonlin. Για κάθε διάνυσμα αναλογιών $a_i\ (22 \times 1)$ εκτελέσαμε στο Matlab τον πολλαπλασιασμό του με τον πίνακα των 34772 γονιδίων $C_{initial}$ και έτσι δημιουργήσαμε το κάθε μείγμα $A_{all\ data}$. Με τις έτοιμες συναρτήσεις του Matlab, *mean* και *var* υπολογίσαμε τη μέση τιμή και τη διασπορά αντίστοιχα για το κάθε μείγμα, όπως και για τα δεδομένα των 27 γονιδίων. Επιπλέον, με χρήση της εντολής του Matlab *hist* δημιουργήσαμε τα ιστογράμματα για κάθε μείγμα. Και πάλι ο άξονας των συχνοτήτων του ιστογράμματος είναι διαβαθμισμένος από 1 ως 20, με μήκος ράβδου ίσο με 0.1. Στον ***Πίνακα 3.18*** που ακολουθεί, παραθέτουμε τα αποτελέσματα για τις μέσες τιμές και τις διασπορές για τα μείγματα $A_{all\ data}$ και A_{final}.

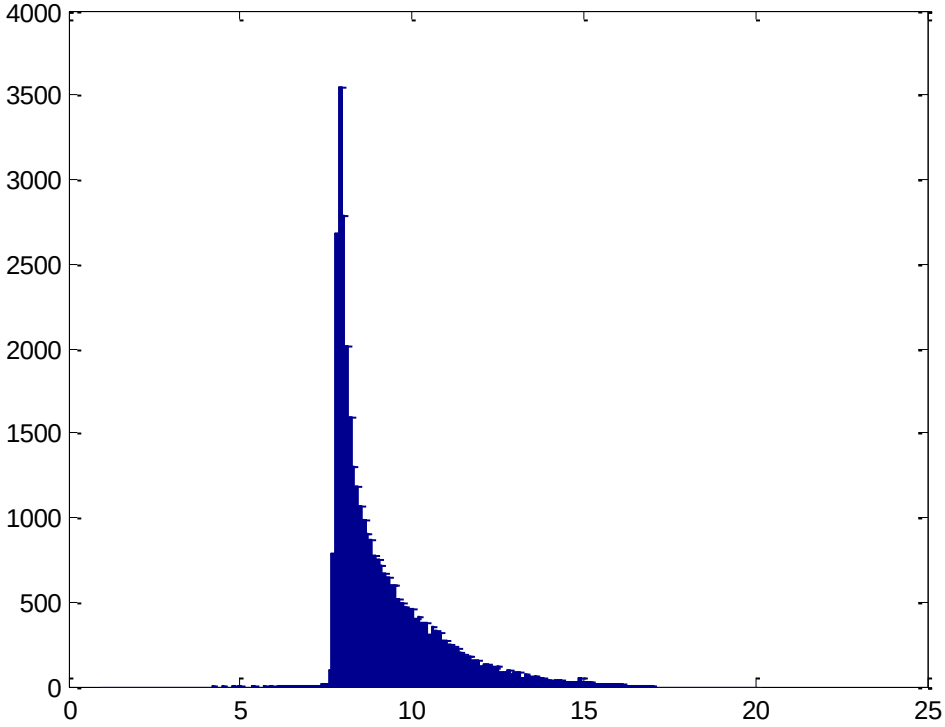
	Grad Desc λ=1,967	**Grad Desc λ=10**	**LSQLIN**
	ALL DATA		
Μέση Τιμή	9,154343479	9,154077731	9,153719616
Διασπορά	2,484305128	2,478782922	2,494639854
	DATA 27x22		
Μέση Τιμή	9,206521904	9,197167887	9,206521904
Διασπορά	2,076685646	2,042696374	2,388072201

Πίνακας 4.15. Σύγκριση Μέσων Τιμών & Διασπορών για τα 2 set δεδομένων

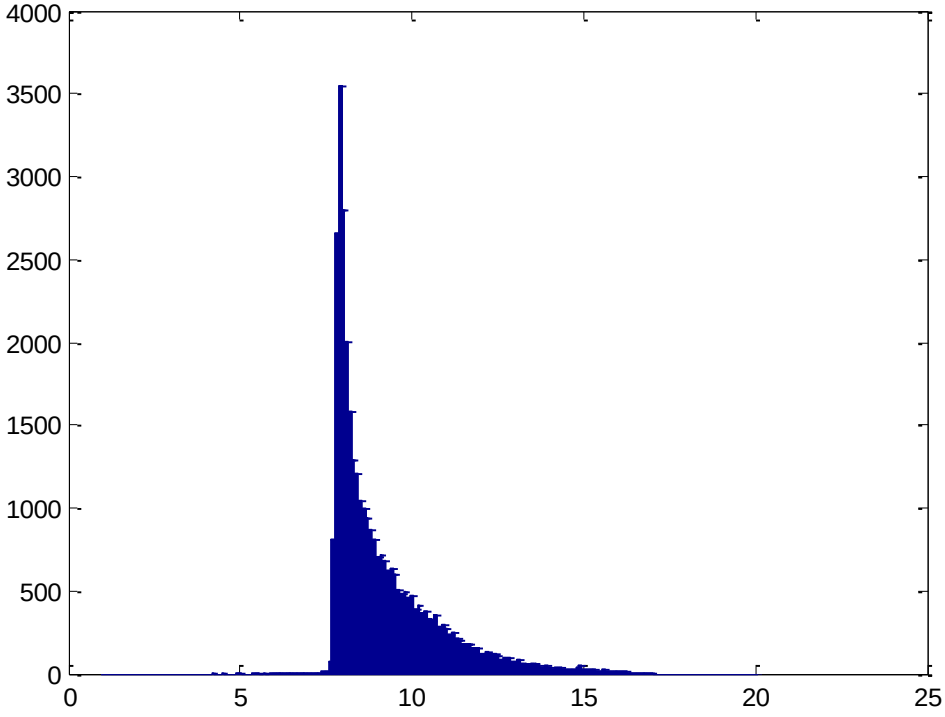
Παρατηρούμε ότι οι μέσες τιμές για τα δύο σύνολα δεδομένων είναι σχεδόν ίδιες. Οι διασπορές και για τα δύο σύνολα δεδομένων κυμαίνονται γύρω απ' την τιμή 2, με τις διασπορές των μειγμάτων για το σύνολο των δεδομένων (34772 γονίδια) να είναι λίγο μεγαλύτερες. Η μικρή αυτή διαφορά στις τιμές των διασπορών μεταξύ των δύο συνόλων δεδομένων οφείλεται στην ύπαρξη περισσότερων στοιχείων στο μείγμα που αντιστοιχεί στο σύνολο των δεδομένων (34772 στοιχεία). Αυτό ισχύει γιατί, περισσότερα στοιχεία στο μείγμα συνεπάγεται πιο μεγάλος αριθμός στοιχείων που παρουσιάζουν αποκλίσεις,

οι οποίες μπορεί να είναι και εντονότερες απ' ότι στο μείγμα με τα 27 στοιχεία (υπήρχαν μόνο 5 στοιχεία με αποκλίσεις). Παρ' όλα αυτά, η διαφορές στις τιμές των διασπορών είναι πολύ μικρές, όπως προαναφέρθηκε και έτσι μπορούμε να πούμε ότι η επιλογή των 27 γονιδίων οδήγησε σε αποτελέσματα ταυτόσημα με αυτά των 34772 γονιδίων. Επομένως, μειώνοντας τα γονίδια καταλήξαμε σε ασφαλή αποτελέσματα, χωρίς να χάσουμε σημαντική πληροφορία.

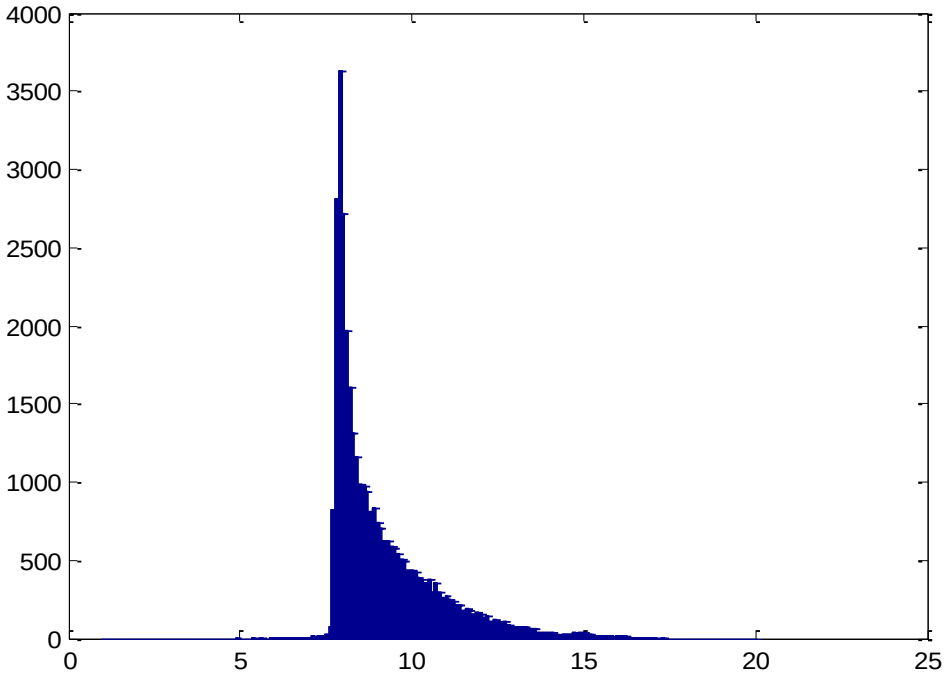
Τέλος, για τις δύο περιπτώσεις τιμών του πολλαπλασιαστή λ για τον αλγόριθμο Gradient Descent και για την περίπτωση της εντολής lsqnonlin παραθέτουμε τα ιστογράμματα που δημιουργήσαμε, με στόχο τον έλεγχο της ομοιομορφίας των μειγμάτων που προέκυψαν από το σύνολο των γονιδίων.



Σχήμα 4.31. Ιστόγραμμα Μείγματος για 34772 γονίδια (λ = 1,967)



Σχήμα 4.32. Ιστόγραμμα Μείγματος για 34772 γονίδια (λ = 10)



Σχήμα 4.33. Ιστόγραμμα Μείγματος για 34772 γονίδια (Lsqnonlin)

Από τα ιστογράμματα παρατηρούμε ότι το μεγαλύτερο σύνολο των στοιχείων των μειγμάτων και συγκεκριμένα περίπου 13.000 στοιχεία σε ένα εύρος τιμών [7.5, 8.]. Επίσης, ένα σύνολο στοιχείων, αξιόλογου μεγέθους είναι αυτό που περιλαμβάνει περίπου 7000 στοιχεία και αντιστοιχεί σε εύρος [8, 8.5]. Το σύνολο των στοιχείων που αποκλίνουν και έχουν πολύ μικρότερες ή μεγαλύτερες τιμές αποτελείται από 14772 γονίδια. Δεν είναι πολύ μικρό, αλλά είναι μικρότερο από το άθροισμα των στοιχείων που παίρνουν τιμές από 7.5 ως 8.5. Συνεπώς επιτυγχάνεται αρκετά μεγάλη ομοιομορφία, αφού 20000 γονίδια κυμαίνονται γύρω από την ίδια περίπου τιμή.

4.3 Υπολογισμός Κάλυψης Μείγματος

Όπως γνωρίζουμε από τη θεωρία, το ποσοστό κάλυψης ορίζεται ως ποσοστό των κηλίδων, οι οποίες έχουν υβριδοποιηθεί με το μείγμα αναφοράς και παράγουν σήμα, πάνω από ένα συγκεκριμένο κατώφλι που έχει ορίσει ο χρήστης. Επομένως, αρκεί τα γονίδια που αποτελούν το μείγμα αναφοράς να περιγράφονται από τιμές έντασης οι οποίες είναι μεγάλες και συγκεκριμένα μεγαλύτερες από ένα προκαθορισμένο κατώφλι. Ως τιμή κατωφλίου επιλέγεται η μέση τιμή έντασης για το σύνολο των κυτταρικών σειρών, η οποία είναι ίση με 350 και σε κανονικοποιημένη μορφή, με τη χρήση δυαδικού λογαρίθμου ίση με $\log_2(350) = 8{,}4512$. Η τιμή αυτή είναι ίδια με την τιμή κατωφλίου που χρησιμοποιείται στην υπολογιστική μέθοδο εύρεσης του πρότυπου μείγματος αναφοράς. Πρόκειται για μία σχετική έρευνα, που αναλύθηκε στο Κεφάλαιο 1, και με την οποία επιθυμούμε να συγκριθούμε για την επιβεβαίωση της ορθότητας των αποτελεσμάτων και συνεπώς και των μεθόδων που χρησιμοποιούμε.
Παρατηρώντας τα ιστογράμματα που παρατέθηκαν στην προηγούμενη ενότητα για τα μείγματα που δημιουργούνται από το σύνολο των γονιδίων (34772) και χρησιμοποιώντας ως κατώφλι την παραπάνω τιμή, μπορούμε να υπολογίσουμε πόσα είναι τα γονίδια, τα οποία βρίσκονται πάνω από το κατώφλι. Ελέγχουμε δηλαδή πόσες ράβδοι βρίσκονται πάνω από τη συγκεκριμένη συχνότητα – κατώφλι και μετά αθροίζουμε τον αριθμό των γονιδίων που ανήκουν σ' αυτές τις ράβδους. Αφού πλέον έχουμε τον αριθμό των γονιδίων που βρίσκονται πάνω από το κατώφλι και αφού γνωρίζουμε ήδη το σύνολο των γονιδίων, μπορούμε να υπολογίσουμε το ποσοστό επί τις 100 (%) της κάλυψης.
Στους 3 πρώτους πίνακες συνοψίζονται τα ποσοστά κάλυψης για κάθε μείγμα που δημιουργείται από το πρόβλημα των Ελαχίστων Τετραγώνων, για κάθε περίπτωση αρχικοποίησης, ενώ στον επόμενο πίνακα συνοψίζονται τα ποσοστά κάλυψης μειγμάτων που προέκυψαν από την ελαχιστοποίηση της Εντροπίας.

	$A_{initial} = 1/27$		
	Gradient Descent		**Lsqlin**
	λ =1,967	**λ = 10**	
# Γονιδίων	19763	19861	19293
Κάλυψη	56.76%	57.12%	55.48%

Πίνακας 4.16. Ποσοστά Κάλυψης για το Πρόβλημα Ελαχίστων Τετραγώνων ($A_{initial} = 1/27$)

	$A_{initial}$ = 9.1194		
	Gradient Descent		**Lsqlin**
	λ =1,967	**λ = 10**	
# Γονιδίων	19951	19822	19801
Κάλυψη	57.38%	57%	56.95%

Πίνακας 4.17. Ποσοστά Κάλυψης για το Πρόβλημα Ελαχίστων Τετραγώνων ($A_{initial}$ = 9.1194)

	$A_{initial}$ ≥ 350		
	Gradient Descent		**Lsqlin**
	λ =1,967	**λ = 10**	
# Γονιδίων	19831	19836	19921
Κάλυψη	57.03%	57.04%	57.29%

Πίνακας 4.18. Ποσοστά Κάλυψης για το Πρόβλημα Ελαχίστων Τετραγώνων ($A_{initial}$ ≥ 350)

	Gradient Descent		**Lsqnonlin**
	λ =1,967	**λ = 10**	
# Γονιδίων	19837	**20004**	19677
Κάλυψη	57.05%	**57.53%**	56.59%

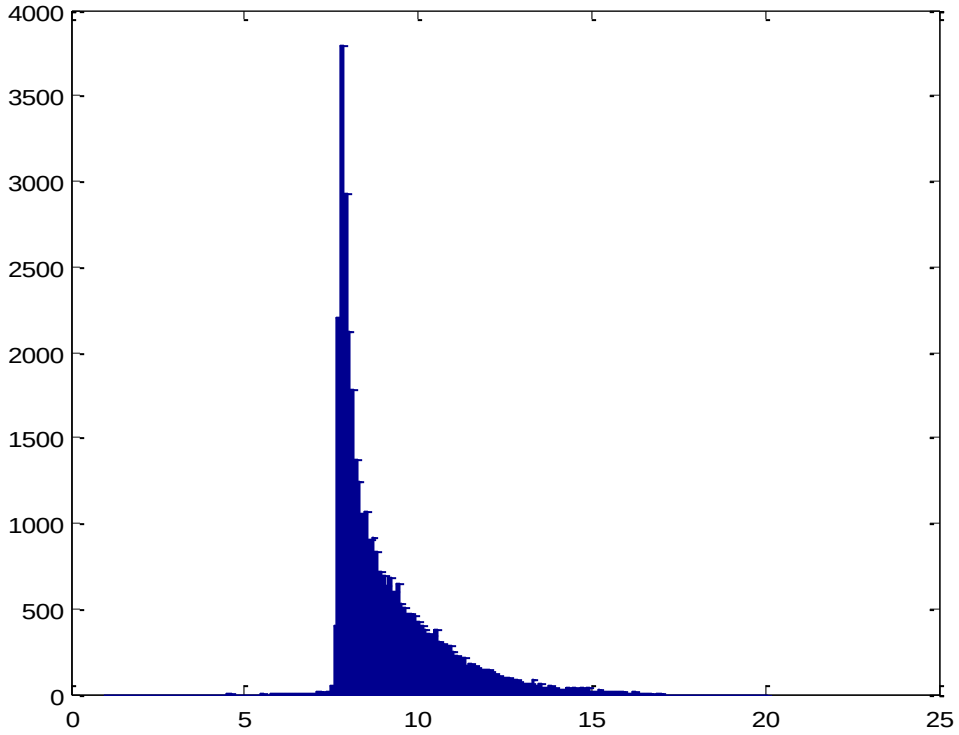
Πίνακας 4.19. Ποσοστά Κάλυψης για το Πρόβλημα Ελαχιστοποίησης Εντροπίας

Όπως αναλύθηκε και παραπάνω, οι βέλτιστες αναλογίες πρόσμειξης και συνεπώς, το μείγμα με την καλύτερη συμπεριφορά ως προς την ομοιομορφία είναι αυτό που προκύπτει από το πρόβλημα Ελαχιστοποίησης της Εντροπίας, για την εφαρμογή του αλγορίθμου Gradient Descent (λ = 10). Όσον αφορά το ποσοστό κάλυψης, το μείγμα που παράγει μεγάλη κάλυψη είναι επίσης αυτό που παράγεται από το πρόβλημα Ελαχιστοποίησης της Εντροπίας, για την εφαρμογή του αλγορίθμου Gradient Descent (λ = 10). Οι διαφορές βέβαια, τόσο μεταξύ των 3 διαφορετικών μειγμάτων που προκύπτουν από το πρόβλημα των Ελαχίστων Τετραγώνων, όσο και μεταξύ των μειγμάτων Ελαχίστων Τετραγώνων - Ελαχιστοποίησης Εντροπίας (για τα δύο δηλαδή διαφορετικά προβλήματα) είναι αρκετά μικρές. *Όμως, το πρόβλημα ελαχιστοποίησης της εντροπίας δεν παύει να καταλήγει σε έστω και λίγο καλύτερη κάλυψη. Η βέλτιστη κάλυψη που επιτυγχάνεται είναι ίση με 57.53%* .

- ✓ Πρώτος μας στόχος είναι να συγκρίνουμε τη μέγιστη κάλυψη, που προκύπτει από τη δική μας υλοποίηση και η *οποία είναι 57.53%,* με την αντίστοιχη που προκύπτει από την υπολογιστική μέθοδο για την εύρεση πρότυπου μείγματος αναφοράς. Η υπολογιστική μέθοδος, όπως αναλύεται και στο Κεφάλαιο 1 καταλήγει σε ένα μείγμα 22 κυτταρικών σειρών, από τις οποίες μόνο οι 11 έχουν μεγάλη συμμετοχή. Το βέλτιστο μείγμα αυτό

καταλήγει *σε ποσοστό κάλυψης ίσο με 57,12%. Παρατηρούμε δηλαδή ότι τα ποσοστά είναι ίδια, με μία μικρή διαφορά (προς το καλύτερο) για τη δική μας κάλυψη*.

- ✓ Τέλος, πραγματοποιήσαμε σύγκριση της μέγιστης κάλυψης που αντιστοιχεί στο βέλτιστο μείγμα (Ελαχιστοποίηση Εντροπίας), με την κάλυψη που προκύπτει από το μείγμα που δημιουργείται εάν αναμείξουμε τις κυτταρικές σειρές με αναλογίες 1/22. Δηλαδή με το μείγμα που προκύπτει από μείξη των σειρών σε αναλογίες 1 προς 1. Αυτό που θέλουμε είναι η μέγιστη κάλυψη που προκύπτει από την 1 προς 1 ανάμειξη να είναι μικρότερη από αυτήν για μείξη με τις βέλτιστες αναλογίες (διαφορετικές μεταξύ τους).
  Παρακάτω παρουσιάζουμε το ιστόγραμμα του μείγματος που προκύπτει από μείξη των κυτταρικών σειρών με 1 προς 1 αναλογίες. Επίσης, στον ***Πίνακα 4.20*** παραθέτουμε τις τιμές κάλυψης που θέλουμε να συγκρίνουμε.



Σχήμα 4.34. Ιστόγραμμα Μείγματος για Αναλογίες 1/22

	Μείγμα 1-1	Βέλτιστο Μείγμα
# Γονιδίων	18776	20004
Κάλυψη	54%	57.53%

Πίνακας 4.20. Σύγκριση Τιμών Κάλυψης Αναλογίες 1-1 v.s. $a_i^{optimal}$

Παρατηρούμε δηλαδή ότι στη δική μας προσέγγιση, με χρήση αλγορίθμων βελτιστοποίησης καταλήγουμε στο ότι *η μείξη κυτταρικών σειρών με αναλογίες 1 προς 1 οδηγεί σε μικρότερη τιμή κάλυψης, από την αντίστοιχη για μείξη των σειρών σε διάφορες αναλογίες* (βέλτιστες αναλογίες από το πρόβλημα Ελαχιστοποίησης της Εντροπίας). Το μείγμα με αναλογίες πρόσμειξης 1 προς 1 άλλωστε δεν έχει καλή συμπεριφορά και ως προς την ομοιομορφία, μιας και η διασπορά του είναι σχετικά μεγάλη.

4.4 Σύγκριση Αποτελεσμάτων με Υπολογιστικής Μεθόδου – Μεθόδων Βελτιστοποίησης

Στην ενότητα αυτή θα προσπαθήσουμε να συγκρίνουμε τα αποτελέσματα που προέκυψαν από τη δική μας προσέγγιση, με την επίλυση των δύο διαφορετικών προβλημάτων ελαχιστοποίησης, με τα αντίστοιχα που προέκυψαν από την εφαρμογή της υπολογιστικής μεθόδου. Δύο είναι τα κύρια σημεία, ως προς τα οποία μπορεί να γίνει η σύγκριση: οι κυτταρικές σειρές που δημιουργούν το μείγμα αναφοράς για κάθε προσέγγιση (πόσες και ποιες) και το ποσοστό κάλυψης που επιτυγχάνεται για το μείγμα κάθε προσέγγισης. Η περίπτωση της κάλυψης εξετάστηκε στην προηγούμενη ενότητα.

Για την περίπτωση, τώρα, των κυτταρικών σειρών, η υπολογιστική προσέγγιση κατέληξε σε βέλτιστο μείγμα με μόνο 11 σημαντικές κυτταρικές σειρές. Η δική μας προσέγγιση κατέληξε σε μείγμα αναφοράς με 22 κυτταρικές σειρές, απ' τις οποίες μόνο οι 5 έχουν μεγάλη συμμετοχή σ' αυτό. Οι 5 αυτές σειρές ξεχώρισαν γιατί, από τις 22 βέλτιστες αναλογίες πρόσμειξης κυτταρικών σειρών που προέκυψαν (από το πρόβλημα Ελαχιστοποίησης της Εντροπίας – Αλγόριθμος Gradient Descent (λ = 10)), μόνο 5 έχουν μεγάλη τιμή. Οι σειρές λοιπόν που αντιστοιχούν σ' αυτές τις αναλογίες έχουν και τη μεγαλύτερη βαρύτητα στο μείγμα. Για να μπορέσουμε όμως να συγκρίνουμε μείγματα με αντίστοιχο αριθμό κυτταρικών σειρών για τις δύο προσεγγίσεις, συμπληρώνουμε στο μείγμα μας ακόμα 6 κυτταρικές σειρές. Το ποιες είναι οι σειρές που θα συμπληρωθούν προκύπτει από την παρακάτω διαδικασία: ταξινομούμε τις αναλογίες από τη μεγαλύτερη στη μικρότερη. Οι 5 πρώτες αντιστοιχούν στις κυτταρικές σειρές με τη μεγαλύτερη βαρύτητα, που ήδη έχουμε συμπεριλάβει στο μείγμα. Οι επόμενες 6, με τη διαδοχική σειρά που εμφανίζονται καθορίζουν και τις αντίστοιχες κυτταρικές σειρές που θα συμπεριληφθούν.

Στον ***πρώτο πίνακα*** παρουσιάζονται οι βέλτιστες αναλογίες πρόσμειξης της δικής μας προσέγγισης, που προέκυψαν από τη μέθοδο της ελαχιστοποίησης της εντροπίας. Στον ίδιο πίνακα, με τη σήμανση √ προσδιορίζονται οι αναλογίες με τις μεγαλύτερες τιμές, άρα και τη μεγαλύτερη βαρύτητα, ώστε να κατανοήσουμε πως προέκυψαν οι 5 σημαντικές για εμάς κυτταρικές σειρές.

Στο ***δεύτερο πίνακα*** περιγράφεται η σύγκριση των αποτελεσμάτων για τις 2 διαφορετικές προσεγγίσεις. Στην πρώτη στήλη παρουσιάζονται οι συντομογραφίες των κυτταρικών σειρών, στη δεύτερη τα αποτελέσματα της υπολογιστικής μεθόδου και στην τρίτη, τα αποτελέσματα της δικής μας προσέγγισης, με τη χρήση μεθόδων βελτιστοποίησης.

Αναλογίες	**Βαρύτητα**
2,15E-05	
2,68E-04	
2,32E-01	√ (3η Κυτταρική Σειρά)
2,15E-04	
1,81E-04	
1,04E-04	
2,49E-05	
1,87E-04	
1,72E-05	
1,93E-04	
2,27E-04	
1,05E-04	
2,28E-05	
2,83E-04	
3,94E-04	
2,24E-01	√ (16η Κυτταρική Σειρά)
2,09E-01	√ (17η Κυτταρική Σειρά)
9,09E-05	
9,43E-02	√ (19η Κυτταρική Σειρά)
3,94E-04	
3,94E-04	
2,38E-01	√ (22η Κυτταρική Σειρά)

Πίνακας 4.21. Βαρύτητα Αναλογιών Πρόσμειξης Κυτταρικών Σειρών

Κυτταρική Σειρά	Υπολογιστική Μέθοδος	Μέθοδοι Βελτιστοποίησης
HL60		
Hs578T	√	√
McF7		√
OVCAR3	√	
Panc1	√	
SKMEL3		
SKMM2		
T47D		
TERA1	√	
U87MG		
Raji	√	√
JAR		
Saos2	√	
SW872		√
THP1	√	√
HCT116		√
HUVEC	√	√
HepG2	√	
HeLa		√
LNCap	√	√
Molt4		√
WERI1	√	√

Πίνακας 4.22. Σύγκριση Κυτταρικών Σειρών – Υπολογιστική Μέθοδος v.s. Μέθοδος Βελτιστοποίησης

Παρατηρώντας τις στήλες του δεύτερου πίνακα, συμπεραίνουμε ότι ο αριθμός των κυτταρικών σειρών που δημιουργούν το κάθε μείγμα αναφοράς και που είναι κοινές και για τα δύο μείγματα είναι 6 στις 11. Έχουμε δηλαδή μία ταύτιση μεγαλύτερη του 50%, το οποίο μας κάνει να συμπεράνουμε ότι τα μείγματα αναφοράς των 2 προσεγγίσεων έχουν αρκετές ομοιότητες.

ΚΕΦΑΛΑΙΟ 5 - ΣΥΜΠΕΡΑΣΜΑΤΑ

5.1 Συμπεράσματα

Στην εργασία αυτή υπολογίσαμε τις βέλτιστες αναλογίες πρόσμειξης κυτταρικών σειρών, με στόχο τη δημιουργία ενός πρότυπου, ομοιόμορφου μείγματος αναφοράς. Για να επιλύσουμε το πρόβλημα αυτό χρησιμοποιήσαμε μεθόδους βελτιστοποίησης και συγκεκριμένα τον αναλυτικό αλγόριθμο Gradient Descent, καθώς και εντολές από το περιβάλλον του Matlab που υλοποιούν προβλήματα βελτιστοποίησης. Χρησιμοποιώντας δύο από τα κυριότερα προβλήματα βελτιστοποίησης, αυτό των Ελαχίστων Τετραγώνων και αυτό της Ελαχιστοποίησης της Εντροπίας και προσαρμόζοντας τα δεδομένα μας στις μαθηματικές απαιτήσεις του καθενός, υπολογίσαμε τις βέλτιστες αναλογίες πρόσμειξης. Από τις αναλογίες αυτές προέκυψαν και τα αντίστοιχα πρότυπα μείγματα αναφοράς, των οποίων η ομοιομορφία ελέγχθηκε χρησιμοποιώντας ως κριτήριο τη διασπορά των μειγμάτων, αλλά και μέσω των γραφικών παραστάσεων που αναπαριστούν τα στοιχεία κάθε μείγματος.

Τα συμπεράσματα που προέκυψαν είναι αρκετά ικανοποιητικά. Στην προσέγγιση του προβλήματός μας με τη μέθοδο των Ελαχίστων Τετραγώνων υπήρξαν 3 διαφορετικές περιπτώσεις αρχικοποιήσεις για το target distribution. Τα πρότυπα μείγματα που προέκυψαν για κάθε περίπτωση είναι ομοιόμορφα μιας και οι τιμές των διασπορών τους είναι περίπου ίσες με 2, τιμή όχι πολύ κοντά στο 0 που είναι το ιδανικό, αλλά σχετικά μικρή, συγκρινόμενη με το εύρος τιμών. Εξετάζοντας στο σύνολό τους τις μέσες τιμές και τις διασπορές για κάθε περίπτωση αρχικοποίησης παρατηρούμε ότι είναι σχεδόν ίσες (οι διαφορές τους είναι αρκετά μικρές), άρα το πρώτο συμπέρασμα το οποίο μπορεί να προκύψει είναι το ότι η αρχικοποίηση, η οποία είναι απαραίτητη για τη μέθοδο των Ελαχίστων Τετραγώνων δεν επηρεάζει το αποτέλεσμα των αναλογιών και συνεπώς και τη σύσταση του μείγματος αναφοράς. Δηλαδή καταλήγουμε σε μείγματα με κοινά χαρακτηριστικά, άρα και με κοινή συμπεριφορά ως προς την ομοιομορφία.

Για τη μέθοδο των Ελαχίστων Τετραγώνων και για κάθε περίπτωση αρχικοποίησης χρησιμοποιήθηκαν δύο τρόποι υλοποίησης, ο αλγόριθμος Gradient Descent και η εντολή lsqlin του Matlab. Συγκρίνοντας τις διασπορές των μειγμάτων αναφοράς για τους δύο αυτούς τρόπους υλοποίησης βλέπουμε ότι είναι σχεδόν ίσες και καταλήγουμε στο ότι τα αποτελέσματά μας δεν επηρεάζονται από τον τρόπο με τον οποίο θα υλοποιηθεί το πρόβλημα των Ελαχίστων Τετραγώνων.

Αντίστοιχα αποτελέσματα προέκυψαν για την προσέγγιση του προβλήματός μας με τη μέθοδο της Ελαχιστοποίησης της Εντροπίας. Όλα τα πρότυπα μείγματα αναφοράς που προέκυψαν είναι ομοιόμορφα. Η συμπεριφορά τους είναι αρκετά παρόμοια, με κάποιες μεγαλύτερες όμως διαφορές στις τιμές των διασπορών, μεταξύ των δύο τρόπων υλοποίησης της μεθόδου (αλγόριθμος Gradient Descent – εντολή lsqnonlin). Για το λόγο αυτό συμπεραίνουμε ότι καλύτερη ομοιομορφία παρουσιάζουν τα μείγματα αναφοράς που προέκυψαν από τη χρήση του αλγορίθμου Gradient Descent. Η σύγκριση με τα αποτελέσματα που προκύπτουν από την εντολή lsqnonlin γίνεται και για τις δύο περιπτώσεις πολλαπλασιαστή Lagrange λ και καταλήγει στο ότι και στις δύο περιπτώσεις, ο αλγόριθμος οδηγεί σε καλύτερα μείγματα αναφοράς.

Συγκρίνοντας τα μείγματα αναφοράς που προέκυψαν από τις δύο διαφορετικές προσεγγίσεις επίλυσης του προβλήματός μας (Ελάχιστα Τετράγωνα & Ελαχιστοποίηση Εντροπίας), καταλήγουμε στο ότι καλύτερα μείγματα ως προς την ομοιομορφία είναι αυτά που προέκυψαν από την Ελαχιστοποίηση της Εντροπίας. Παρ’ ότι οι διαφορές που παρουσιάζονται δεν είναι

μεγάλες και παρ' ότι οι απαιτήσεις του μη γραμμικού προβλήματος της Ελαχιστοποίησης της Εντροπίας σε χρόνο και πολυπλοκότητα υλοποίησης είναι μεγαλύτερες από τις αντίστοιχες του προβλήματος των Ελαχίστων Τετραγώνων, ελαχιστοποιώντας την εντροπία καταλήγουμε σε καλύτερα αποτελέσματα για τα πρότυπα μείγματα αναφοράς.
Τέλος επιθυμούσαμε να συγκρίνουμε τα αποτελέσματα της δικής μας προσέγγισης, με τα αντίστοιχα που προέκυψαν από μία πρόσφατη σχετική προσέγγιση, η οποία χρησιμοποίησε υπολογιστικές μεθόδους για την εύρεση του πρότυπου μείγματος αναφοράς. Αξίζει βέβαια να σημειώσουμε ότι η προσέγγιση αυτή δεν στόχευε στη δημιουργία ομοιόμορφου μείγματος αναφοράς. Γι' αυτό συγκρίναμε τις κυτταρικές σειρές που δημιουργούν το βέλτιστο μείγμα αναφοράς, καθώς και το ποσοστό κάλυψης του βέλτιστου μείγματος για κάθε προσέγγιση. Στην υπολογιστική μέθοδο προκύπτει μείγμα με μόνο 11 είναι σημαντικές κυτταρικές σειρές. Στη δική μας προσέγγιση, σημαντικές είναι οι 5 από τις 22 κυτταρικές σειρές. Για να μπορέσουμε όμως να συγκρίνουμε τα μείγματα, χρειάστηκε να προσθέσουμε στο δικό μας 6 επιπλέον σειρές, των οποίων η βαρύτητα μπορεί να μην ήταν τόσο μεγάλη, όσο των 5, αλλά ήταν αρκετά καλή. Αυτό έγινε για να συγκριθούν μείγματα που προκύπτουν από τη μείξη ίσου αριθμού κυτταρικών σειρών. Καταλήξαμε σε 6 κοινές κυτταρικές σειρές μεταξύ των δύο μειγμάτων, από τις 11 βέλτιστες. Όσο για το ποσοστό κάλυψης, για την υπολογιστική μέθοδο, η μέγιστη κάλυψη είναι ίση με 57,12%, ενώ για τη δική μας υλοποίηση ισούται με 57.53%, δηλαδή οι δύο διαφορετικές προσεγγίσεις καταλήγουν σε κοινά αποτελέσματα.

5.2 Μελλοντική Εργασία

Παρά το γεγονός ότι τα αποτελέσματα τα οποία προέκυψαν από την ανωτέρω ανάλυση είναι αρκετά ικανοποιητικά, δεν παύουν να υπάρχουν ιδέες για μία πιο ολοκληρωμένη αντιμετώπιση του προβλήματος αυτού.

- Μία πρώτη ιδέα είναι η πειραματική χρήση των πρότυπων μειγμάτων αναφοράς που προέκυψαν από την εργασία αυτή, σε εξειδικευμένα εργαστήρια Μοριακής Βιολογίας, που ασχολούνται με ανάλυση μικροσυστοιχιών. Τοποθετώντας τα μείγματα αναφοράς σε μικροσυστοιχίες, αυτά υβριδοποιούνται με τα γονίδια που είναι διάστικτα στις μικροσυστοιχίες. Από την υβριδοποίηση λαμβάνουμε την απαραίτητη πληροφορία για τα γονίδια αυτά. Για να ελέγξουμε πόσο καλό είναι το μείγμα αναφοράς που χρησιμοποιήθηκε, αρκεί να αναλύσουμε την πληροφορία που λαμβάνουμε. Εάν τα γονίδια είναι πλήρως ευδιάκριτα και διαχωρίσιμα μετά την υβριδοποίηση και έχουν κοινές τιμές έντασης, τότε το μείγμα αναφοράς που χρησιμοποιήθηκε είναι ομοιόμορφο και κατάλληλο για χρήση. Έτσι, χρησιμοποιώντας τα μείγματα στην πράξη και έχοντας ως κριτήριο την έννοια της κάλυψης, έναν δηλαδή όρο που χρησιμοποιείται πολύ στην ανάλυση μικροσυστοιχιών, μπορούμε να ελέγξουμε την καταλληλότητα των μειγμάτων αναφοράς που προέκυψαν από την υλοποίησή μας. Επίσης, σε περίπτωση που τα μείγματα αυτά δεν είναι κατάλληλα, μπορούμε μέσω της πειραματικής ανάλυσης να καταλάβουμε που υπάρχει το πρόβλημα και έτσι να βελτιώσουμε την υλοποίησή μας.

- Τέλος, μία δεύτερη ιδέα είναι η χρήση γενετικών αλγορίθμων για την επιλογή των παραμέτρων της βελτιστοποίησης. Πρόκειται για αλγορίθμους επίλυσης προβλημάτων, που είναι εμπνευσμένοι από τις αρχές της Βιολογικής Εξέλιξης. Οι αλγόριθμοι αυτοί εκτελούν αναζήτηση στο χώρο των υποψήφιων λύσεων, με στόχο την εύρεση αποδεκτής λύσης, σύμφωνα με κάποιο κριτήριο. Το πλεονέκτημα των αλγορίθμων αυτών, σε σχέση με αυτούς που χρησιμοποιήθηκαν στη δική μας υλοποίηση είναι το ότι μπορούν να βρουν μία αρκετά καλή λύση, σε ένα εύλογο χρονικό διάστημα.

Συνοψίζοντας, όλα τα παραπάνω αποτελούν σημαντικές επεκτάσεις, ώστε να δημιουργηθούν πρότυπα μείγματα αναφοράς που θα είναι εφαρμόσιμα στο ευρύ πεδίο της ανάλυσης μικροσυστοιχιών.

ΠΑΡΑΡΤΗΜΑ Α

(1)Α.1 Ερμηνεία Κυτταρικών Σειρών

Οι 22 κυτταρικές σειρές που υπάρχουν στα δεδομένα του προβλήματός μας περιγράφουν ανθρώπινα καρκινώματα. Μία κυτταρική σειρά που αντιστοιχεί σε κάποιο καρκίνωμα αποτελείται από συγκεκριμένα γονίδια, με συγκεκριμένες τιμές έντασης των γονιδίων αυτών. Αυτό ισχύει για να μπορούμε να διαχωρίζουμε τη μία κυτταρική σειρά από την άλλη και για να γνωρίζουμε ποια συγκεκριμένη πληροφορία περιλαμβάνουν όταν τις χρησιμοποιούμε. Παρακάτω παρατίθενται οι συντομογραφίες των κυτταρικών σειρών που συμμετέχουν στο πρόβλημά μας, καθώς και ποιο είναι το αντίστοιχο καρκίνωμα που αντιπροσωπεύουν.

Κυτταρική Σειρά	**Πηγή Καρκινώματος**
HL60	Μυελός των Οστών
Hs578T	Θηλαστικός Αδένας
McF7	Θηλαστικός Αδένας
OVCAR3	Ωοθήκες
Panc1	Πάγκρεας
SKMEL3	Δέρμα
SKMM2	Μυελός των Οστών
T47D	Θηλαστικός Αδένας
TERA1	Όρχεις
U87MG	Εγκέφαλος
Raji	Β – Λεμφοκυττάρων
JAR	Γεννητικά Όργανα
Saos2	Οστά
SW872	Λιποσάρκωμα
THP1	Αίμα
HCT116	Έντερο
HUVEC	Αιμοφόρα Αγγεία
HepG2	Συκώτι
HeLa	Αυχένα
LNCap	Προστάτης
Molt4	Τ – Λεμφοκυττάρων
WERI1	Αμφιβληστροειδής Χιτώνας Ματιού

Πίνακας Α.1. Συντομογραφίες & Πηγές Καρκινωμάτων Κυτταρικών Σειρών

(2)*Α.2 Καταγραφή 27 Γονιδίων έπειτα από τη χρήση Συσχέτισης*

Μετά από την εφαρμογή της συσχέτισης στα αρχικά δεδομένα μας και πριν προκύψει ο τελικός πίνακας με τα 27 γονίδια και τις 22 κυτταρικές σειρές υπάρχει ένα αρχείο, στο οποίο αποθηκεύονται τιμές 0 και 1. Η τιμή 0 συνεπάγεται ότι δεν υπάρχει καμία συσχέτιση μεταξύ αυτού του γονιδίου και των υπολοίπων, και επομένως το γονίδιο δε πρέπει να απορριφθεί, ενώ η τιμή 1 υποδηλώνει συσχέτιση και το γονίδιο μπορεί να απορριφθεί. Χρησιμοποιώντας ως βάση το αρχείο αυτό καταγράψαμε ποια είναι τα γονίδια τα οποία τελικά αποτελούν τον πίνακα δεομένων μεγέθους 27x22. Τα γονίδια που κρατήθηκαν είναι αυτά που αντιστοιχούν στην τιμή 0, στο βοηθητικό αρχείο. Στον πίνακα που ακολουθεί παραθέτουμε τα 27 αυτά γονίδια.

H200000001	H200000018	H200000092
H200000002	H200000022	H200000115
H200000005	H200000025	H200000128
H200000006	H200000035	H200000136
H200000008	H200000037	H200000139
H200000010	H200000063	H200000141
H200000011	H200000073	H200000163
H200000012	H200000081	H200000170
H200000014	H200000091	H200000291

Πίνακας Α.2. Προσδιορισμός 27 γονιδίων που προέκυψαν με τη χρήση συσχέτισης

Α.3 Βαρύτητα Κυτταρικών Σειρών για το Βέλτιστο Μείγμα

Το βέλτιστο μείγμα που προέκυψε από την υλοποίησή μας είναι αυτό που προκύπτει από το πρόβλημα Ελαχιστοποίησης της Εντροπίας, έπειτα από την αναλυτική εφαρμογή του αλγορίθμου Gradient Descent και για τιμή του πολλαπλασιαστή λ ίση με 10. Εξετάζοντας τις αναλογίες που δημιούργησαν το βέλτιστο αυτό μείγμα μπορούμε να συμπεράνουμε ποιες κυτταρικές έχουν σημαντική συμμετοχή στη δημιουργία του μείγματος και ποιες όχι. Στον παρακάτω πίνακα με τη σήμανση √ παρουσιάζονται οι σειρές με τη μεγάλη συμμετοχή στο μείγμα.

Κυτταρική Σειρά	**Βαρύτητα**
HL60	
Hs578T	
McF7	√
OVCAR3	
Panc1	
SKMEL3	
SKMM2	
T47D	
TERA1	
U87MG	
Raji	
JAR	
Saos2	
SW872	
THP1	
HCT116	√
HUVEC	√
HepG2	
HeLa	√
LNCap	
Molt4	
WERI1	√

Πίνακας Α.3. Βαρύτητα Κυτταρικών Σειρών στο Βέλτιστο Μείγμα

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Georgia Tsiliki, Sofia Kaforou, Manouela Kapsetaki, Dimitris Kafetzopoulos, "A computational approach to microarray universal reference" , *Institute of Molecular Biology & Biotechnology*, Heraklion, Greece

[2] Anastasios Koutsos, Alexandra Manaia και Julia Wilingale-Theune, "Η αληθινή και εικονική μικροσυστοιχία βήμα προς βήμα.pdf", *European Molecular Biology Laboratory*, Heidelberg

[3] http://en.wikipedia.org/wiki/Optimization

[4] Don Johnson, "Optimization Theory", Version 1.4: Mar 22, 2008 4:24 pm GMT-5

[5] Don Johnson, "Constraint Optimization", Version 1.2: Jul 7, 2003 12:00 am GMT-5

[6] Stephen Boyd, Lieven Vandenberghe, "Convex Optimization.pdf", *Cambridge University Press*, chapter 4.1 – Optimization Problems, pp. 127 – 135, 2004

[7] Stephen Boyd, Lieven Vandenberghe, "Convex Optimization.pdf", *Cambridge University Press*, chapter 4.2 – Convex Optimization, pp. 136 – 143, 2004

[8] Stephen Boyd, Lieven Vandenberghe, "Convex Optimization.pdf", *Cambridge University Press*, chapter 1.3 – Convex Optimization, pp. 7 – 8, 2004

[9] Stephen Boyd, Lieven Vandenberghe, "Convex Optimization.pdf", *Cambridge University Press*, chapter 4.4 – Quadratic Optimization Problems, pp. 152 – 153, 2004

[10] Alexandre M. Bayen, "Descent Methods.pdf", *University of California at Berkeley*, Lecture 10, 2005

[11] Stephen Boyd, Lieven Vandenberghe, "Convex Optimization.pdf", *Cambridge University Press*, chapter 9.2 – Descent Methods, pp. 463 – 469, 2004

[12] http://en.wikipedia.org/wiki/Gradient_descent

[13] S. Sawyer , "The Method of Lagrange Multipliers.pdf" – October 25, 2002

[14] http://en.wikipedia.org/wiki/Lagrange_multipliers

[15] Stephen Boyd, Lieven Vandenberghe, "Convex Optimization.pdf", *Cambridge University Press*, chapters 5.1, 5.2 – The Lagrange Dual Function & The Lagrange Dual Problems, pp. 215 – 231, 2004

[16] Joseph Dennis Kolesar, "Matlab and Max/Min.pdf", *University of Maryland*

[17] Richard O. Duda, Peter E. Hart and David G. Stork, “Pattern Classification.pdf”, 2nd edition, September 3, 1997

[18] Amos Ron & Mark Cowlishaw, “Review of Least Squares Solutions to Overdetermined Systems.pdf”, September 11, 2005

[19] http://en.wikipedia.org/wiki/Principle_of_maximum_entropy

[20] http://en.wikipedia.org/wiki/Entropy

[21] G. Karagiannis & G. Steinhauer, “The Nearest Neighbor Rule.pdf”

[22] Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., and Botstein, D. (1999). "Imputing missing data for gene expression arrays", Technical Report, Division of Biostatistics, *Stanford University*

[23] A. Karpinski, “Correlation.pdf”, 2007

[24] http://en.wikipedia.org/wiki/Correlation

[25] http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

[26] Jonathan Rosenberg, “Gradients, Gradient Plots and Tangent Planes” – *Matlab*, 2005

[27] Coleman, T.F. and Y. Li, "A Reflective Newton Method for Minimizing a Quadratic Function Subject to Bounds on Some of the Variables," *SIAM Journal on Optimization*, Vol. 6, Number 4, pp. 1040-1058, 1996

[28] Coleman, T.F. and Y. Li, "An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds," *SIAM Journal on Optimization*, Vol. 6, pp. 418–445, 1996