Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Πειραματική αξιολόγηση αλγορίθμων εκμάθησης αποκομμένων πολυδιάστατων Γκαουσσιανών κατανομών

Διπλωματικη Εργασια

του

χριστοδουλου γ. σαντοριναιου

Επιβλέπων: Δημήτριος Φωτάκης
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εργαστηριο Λογικης και Επιστημης Υπολογιστων
Αθήνα, Νοέμβριος 2020

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Πειραματική αξιολόγηση αλγορίθμων εκμάθησης αποκομμένων πολυδιάστατων Γκαουσσιανών κατανομών

Διπλωματικη Εργασια

του

χριστοδουλου γ. σαντοριναιου

Επιβλέπων: Δημήτριος Φωτάκης
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 12η Νοεμβρίου 2020

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Δημήτριος Φωτάκης	Αριστείδης Παγουρτζής	Νικόλαος Παπασπύρου
Αν. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Εργαστηριο Λογικης και Επιστημης Υπολογιστων
Αθήνα, Νοέμβριος 2020

(Υπογραφή)

.....

Χριστοδουλος Γ. Σαντοριναιος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Περίληψη

Το πρόβλημα της εκμάθησης αποκομμένων κανονικών κατανομών έχει ιστορία τριών αιώνων. Το φαινόμενο της αποκοπής στην περίπτωση μίας d-διάστατης κανονικής κατανομής παρατηρείται όταν τα δείγματα αποκαλύπτονται μόνο εάν ανήκουν σε κάποιο σύνολο $S \subseteq \mathbb{R}^{\mathrm{d}}$. Διαφορετικά παραμένουν κρυφά και ο λόγος τους προς τον συνολικό παραμένει άγνωστος. Συναντάται στην πράξη σε διάφορους τομείς: από την προστασία ευαίσθητων προσωπικών δεδομένων σε οικονομικές εφαρμογές μέχρι την αδυναμία καταγραφής ακραίων τιμών που εμφανίζουν τα όργανα μέτρησης πειραματικών διατάξεων.

Το 2018, οι Daskalakis et al. παρουσίασαν τον πρώτο αλγόριθμο που εκτιμάει τις παραμέτρους της αποκομμένης κατανομής με αυθαίρετη ακρίβεια σε πολυωνυμικό χρόνο. Μοναδικές προϋποθέσεις η πρόσβαση μέσω μαντείου στο σύνολο S, το οποίο να έχει μη τετριμμένο μέτρο υπό την άγνωστη κατανομή. Η καινοτομία τους αφορούσε στην εισαγωγή ενός συνόλου προβολής με κατάλληλες ιδιότητες ώστε ο αλγόριθμος Στοχαστικής Κατάβασης Κλίσης(Stochastic Gradient Descent) να επιτυγχάνει (πρακτικά βέλτιστη) πολυωνυμική πολυπλοκότητα, τόσο δειγματική όσο και υπολογιστική.

Στην παρούσα διπλωματική παρουσιάζουμε μία πειραματική υλοποίηση και αξιολόγηση του αλγορίθμου και ελέγχουμε εάν η πράξη συμφωνεί με την θεωρία, και έαν ναι υπό ποιες συνθήκες αυτό επιτυγχάνεται. Συγκεκριμένα, προσαρμόζοντας τον ρυθμό μάθησης του αλγορίθμου επιβεβαιώνουμε τα θεωρητικά αποτελέσματα αναφορικά με την διάσταση του προβλήματος, το μέτρο αποκοπής και την ακτίνα προβολής. Αφήνουμε ένα ανοικτό ερώτημα σχετικά με τη φύση του συνόλου αποκοπής και προτείνουμε μια ευρεστική μέθοδο που επιτυγχάνει ταχύτατη σύγκλιση στην πράξη.

Λέξεις Κλειδιά

Μηχανική Μάθηση, Πολυδιάστατη Κανονική κατανομή, Στοχαστική κατάβαση κλίσης, Αποκομμένη Στατιστική

Abstract

The problem of learning truncated normal distributions spans three centuries. In the case of d-variate normal distribution, truncation occurs when the samples are revealed ony if they fall in some subset $S \subseteq \mathbb{R}^{\mathrm{d}}$. Otherwise, they remain hidden and their count in proportion to the total number of samples remains unknown. In practice, it manifests in several areas: from protecting sensitive information in financial applications to the incapacity of observing extreme values in experiments with current measuring instruments.

In 2018, Daskalakis et al. presented the first algorithm that estimates the parameters of the truncated distribution with arbitrary accuracy and in polynomial time. Sole requirement is oracle access to the set S, which has non trivial measure under unknown distribution. Their innovation included the introduction of a projection set with properties that allow Stochastic Gradient Descent algorithm to achieve (practically optimal) polynomial complexity, both sample and computational.

In this thesis, we present an experimental implementation and evaluation of the algorithm and check whether or not practice and theory agree, and if so under which conditions. More specifically, by adjusting the learning rate of the algorithm we confirm the theoretical results about the problem's dimension, the truncation measure and the radius of the projection domain. We leave an open question regarding the nature of the truncation set and suggest a heuristic method that achieves fast convergence.

Keywords

Machine Learning, Multivariate Normal Distribution, Stochastic gradient descent, Truncated Statistics

Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαίτερα τον επιβλέποντα καθηγητή μου κ.Δημήτριο Φωτάκη για όλη την βοήθεια και την καθοδήγηση που μου προσέφερε κατά τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας.

Έπειτα, θα ήθελα να εντείνω τις ευχαριστίες μου τόσο στον κ.Φωτάκη όσο και στα άλλα δύο μέλη της τριμελούς επιτροπής, τους καθηγήτες κ.Αριστείδη Παγουρτζή και κ.Νικόλαο Παπασπύρου, για την διδασκαλία τους στα προπτυχιακά μου χρόνια. Κατέστησαν την επιλογή να ακολουθήσω την κατεύθυνση Πληροφορικής πραγματικά πολύ εύκολη.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Αντικείμενο της διπλωματικής

Στην επιστήμη των Πιθανοτήτων και της Στατιστικής κατέχει εξέχουσα θέση η κανονική κατανομή. Από την μία, οι αναλυτικές της ιδιότητες την καθιστούν βασικό εργαλείο μελετής και, από την άλλη, η εμφάνισή της σε πολλές και διάφορες πτυχές της καθημερινότητας την φέρνουν συνεχώς στο προσκήνιο σε πολλές στατιστικές έρευνες.

Είναι επόμενο να έχει διατηρήσει, αν όχι ενισχύσει, αυτήν της την θέση και στην μελέτη της Στατιστικής Μάθησης. Αποτέλει μία από τις πρώτες κατανομές για τις οποίες εξάχθηκαν μέθοδοι υπολογισμού και εκτίμησης από σύνολα δειγμάτων. Συνεπώς, είναι αναμενόμενο να την συναντάμε ξανά και ξανά στις διάφορες κατηγορίες μάθησης.

Μία από αυτές, που σφύζει έρευνας σήμερα, είναι η μάθηση από αποκομμένα δείγματα. Με τον όρο αποκοπή περιγράφουμε το φαινόμενο κατά το οποίο δείγματα από μία κατανομή μας αποκρύπτονται, εκτός κι αν ανήκουν σε ένα συγκεκριμένο σύνολο αποκάλυψης. Ίσως το απλούστερο παράδειγμα που μπορούμε να δώσουμε αφορά σε δείγματα που παράγονται από πειραματικές μετρήσεις. Η περιορισμένη ακρίβεια των οργάνων μέτρησης περιορίζει το χώρο από τον οποίο μπορούμε να λάβουμε αξιόπιστα δεδομένα, επί της ουσία αποκόπτωντας την κατανομή. Ειδική περίπτωση αποκοπής αποτελεί και η περίπτωση λογοκρισίας των δειγμάτων κατά την οποία ο συλλέκτης των δεδομένων έχει πλήρη γνώση αλλά για να προστατέψει ευαίσθητες πληροφορίες αποκρύπτει ηθελημένα πληροφορίες από τον στατιστικό αναλυτή.

Οι πρώτες προσπάθειες μάθησης των παραμέτρων μίας αποκομμένης κανονικής κατανομής τοποθέντουνται στα τέλη του 19ου αιώνα. Την άρχη έκανε ο Galton στο [25] για να ακολουθήσουν λίγο αργότερα οι Pearson και Lee στα [46], [47] και [42]. Το πρόβλημα στην μονοδιάστατη περίπτωση λύθηκε τελικά από τον Fisher στο [22]. Η λύση του, ωστόσο, δεν γενικευόταν αποδοτικά στις υψηλότερες διαστάσεις.

Στην σημερινή εποχή που χαρακτηρίζεται από τεράστιο όγκο δεδομένων, με όρους όπως *Big Data* και *Internet of Things* να ακούγοντα όλο και συχνότερα, ήταν μάλλον επιτακτικότερη από πότε η ανάπτυξη ενός αλγορίθμου που είναι υπολογιστικά αποδοτικός ακόμη και σε πολύ μεγάλες διαστάσεις. Αυτό το κενό στην βιβλιογράφια κάλυψαν οι Daskalakis et al στο Efficient Statistics, in High Dimensions, from Truncated Samples,[11], αποδεικνύοντας τον

πρώτο αλγόριθμο πολυωνυμικής πολυπλοκότητας τόσο ως προς την διάσταση των δειγμάτων όσο και ως προς το πλήθος αυτών.

Υποθέτοντας πρόσβαση σε μαντείο με δυνατότητες παραγωγής τυχαίων δειγμάτων από την πραγματική κατανομή στο σύνολο αποκοπής και απαντήσεων σε ερωτήματα ιδιότητας μέλους σε συνδυασμό με μη αμελητέα μάζα πιθανότητας πάνω στο σύνολο, όρισαν έναν συνολο προβολής τέτοιο ώστε να περιέχει και την πραγματική κατανομή και την εκτίμησης που μπορούν να παράξουν αποκομμένα δείγματα με την μέθοδο της πιθανοφάνειας. Ακόμη, κάθε κατανομή που βρίσκεται εντός του συνόλου αποδεικνύεται να έχει τουλάχιστον κάποιο σταθερό μέτρο πάνω στο σύνολο αποκοπής. Εφοδιάσαν τον αλγόριθμος Στοχαστικής Κατάβασης Κλίσης με το σύνολο προβολής για να επιτύχουν σύγκλιση στις πραγματικές παραμέτρους.

1.2 Συνεισφορά

Η συνεισφορά της διπλωματικής αποδίδεται στην πειραμάτικη υλοποίηση του αλγορίθμου. Πιο αναλυτικά, εργαστήκαμε ως εξής: ξεχωρίσαμε τις παραμέτρους που εισήχθησαν στην θεωρητική ανάλυση:

- ⋆ Τη διάσταση του προβλήματος
- ⋆ Το είδος του συνόλου αποκοπής
- ⋆ Το μέτρο της κατανομής πάνω στο σύνολο
- ⋆ Την ακτίνα του συνόλου προβολής

και προσθέσαμε σε αυτές κάποιες μεταβλητές που επηρεάζουν την εξέλιξη του αλγορίθμου στην πράξη:

- ⋆ Τον ρυθμό μάθησης
- ⋆ Το μέγεθος της συστάδας δειγμάτων

Σταθεροποιώντας κάθε φόρα όλες τις παράμετρους πλην μίας, αποτυπώσαμε την εξέλιξη του αλγορίθμου. Στο θεωρητικό σκέλος, επιβεβαιώσαμε την σύγκλιση που προέβλεπε η θεωρία αναφορικά με την διάσταση του προβλήματος, το μέτρο της κατανομής επί του συνόλου αποκοπής και την ακτίνα του συνόλου προβολής. Αναφορικά με τα σύνολα αποκοπής, παράξαμε κάποια αποτελέσματα που επιβεβαίωνουν την θεωρία αλλά αντιμετωπίσαμε μία μη αναμενόμενη αδυναμία σύγκλισης για μια συγκεκριμένη οικογένεια συνόλων. Στο πρακτικό κομμάτι, μελετήσαμε το tradeoff μεταξύ του μεγέθους της συστάδας δειγμάτων και του ρυθμού μάθησης. Ο τελευταίος αποτέλεσε και την ουσιαστική αλλαγή που απαιτεί η προσαρμογή της θεωρίας στην πράξη αφού τελικά υιοθέτησαμε διαφορετικό σχήμα από το θεωρητικά ταχύτερο. Επιδείξαμε, όμως, καλύτερα πρακτικά αποτελέσματα. Ταυτόχρονα, μας δόθηκε η ευκαιρία να σχολιάσουμε τον θόρυβο που εισάγει η διακύμανση της εκτίμησης κλίσης, που όπως θα δούμε έχει εξαιρετική σημασία στην πράξη. Φυσικά, είναι κάτι με το οποίο δεν ασχολείται η θεωρία. Τέλος, μέσα από μία απλή ευρεστική τεχνική, που πατάει κατάλληλα στα μαθηματικά θεμέλια των αλγορίθμων, διαπιστώσαμε μια άκρως ποιοτική σύγκλιση σε εξαιρετικά μικρό αριθμό βημάτων.

1.3 Οργάνωση του τόμου

Το κύριο σώμα της εργασίας χωρίζεται σε 4 κεφάλαια.

- Στο Κεφάλαιο 2 γίνεται μια σύντομη εισαγωγή στο χώρο των Πιθανοτήτων. Θυμόμαστε τους ορισμούς βασικών ποσοτήτων, όπως της έννοιας της κατανομής και της μέσης τιμής, σε μία ή περισσότερες διαστάσεις. Προχωράμε στον ορισμό της Γκαουσσιανής κατανομής και των αποκομμένων κατανομών. Ακόμα, βλέπουμε δύο χρήσιμα εργαλεία για την μετέπειτα μελέτη μας: την απόσταση ολικής μεταβολής και την συνάρτηση πιθανοφάνειας.

- Στο Κεφάλαιο 3 ενισχύουμε το θεωρητικό μας υπόβαθρο με αλγορίθμους από το χώρο της Μηχανικής Μάθησης. Μετά από κάποιες εισαγωγικές σημειώσεις για αυτή, βλέπουμε την μέθοδο της Εκτίμησης Μέγιστης Πιθανοφάνειας, μίας από τις πρώτες που αναπτύχθηκαν για μάθηση από δείγματα. Κατόπιν, γνωρίζουμε τον αλγόριθμο Κατάβασης Κλίσης, το γνωστό gradien descent, και διατυπώνουμε κάποια βασικά αποτελέσματα από την μελέτη του, κυρίως για κυρτές συναρτήσεις. Έπειτα, αφήνουμε πίσω μας τις μεθόδους ντετερμινιστικής βελτιστοποίησης και μελετάμε την Στοχαστική Κατάβαση Κλίσης, ή stochastic gradient descent - SGD. Διατυπώνουμε τα αντίστοιχα θεωρήματα και συγκρίνουμε τα οφέλη που παρέχει η στοχαστική εκδοχή αλλά και τις επιπρόσθετες δυσκολίες που φέρνει.

- Στο Κεφάλαιο 4 πραγματοποιούμε μία ιστορική αναδρομή στην μάθηση από αποκομμένα δείγματα, δίνοντας φυσικά εξέχουσα θέση στις προσπάθειες που αφορούσαν την κανονική κατανομή. Με αύτον τον τρόπο, όταν έρχεται η ώρα να γνωρίσουμε τις μεθόδους που εισήγαγαν οι Daskalakis et al. έχουμε εξοικειωθεί με την αντίστοιχη βιβλιογραφία και έχουμε κατανοήσει την αξία της εργασίας τους, που συνεπάγεται την αξία της παρούσας διπλωματικής, καθώς και τα εμπόδια που υπερκέρασαν. Ολοκληρώνουμε την αναδρομή με σύγχρονες ερευνητικές εργασίες, του τρέχοντος έτους, που εφαρμόζουν σε διάφορα προβλήματα τις τεχνικές που ανέπτυξαν οι συγγραφείς.

- Στο Κεφάλαιο 5 περνάμε από την θεωρία στην πράξη. Βλέπουμε πως τα θεωρήματα μεταφράζονται σε αλγορίθμους αλλά και ποιες προσαρμογές είναι απαραίτητες κατά τη διαδικασία. Οδηγούμαστε, έτσι, στις υλοποιήσεις μας και παρουσιάζουμε τα πειραματικά δεδομένα. Επιβεβαιώνουμε σχεδόν κάθε πτυχής της θεωρίας ενώ προσθέτουμε και την μελέτη της διακύμανσης της εκτίμησης κλίσης. Κλείνουμε το κεφάλαιο με μία ευρεστική που μας χαρίζει ταχύτατη σύγκλιση.

Κλείνουμε την εργασία μας με το Κεφάλαιο 6 το οποίο συνοψίζει την δουλειά μας και προτείνει ενδιαφέρουσες επεκτάσεις αυτής.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο I: Πιθανότητες & Στατιστική

Στην παρούσα εργασία καταπιανόμαστε αποκλειστικά με αποκομμένες Γκαουσσιανές κατανομές. Συνεπώς, θα αποδειχθεί χρήσιμο για την συνέχεια να ορίσουμε κάθε έναν από τους τρεις όρους ξεχωριστά. Αυτό θα αποτελέσει και το αντικείμενο της πρώτης ενότητας του παρόντος κεφαλαίου. Επιπρόσθετα, στην ενότητα 2.2 θα εισάγουμε δύο ακόμη έννοιες, αυτές της απόστασης κατανομών και των συναρτήσεων πιθανοφάνειας, που θα αποτελέσουν για μας βασικά εργαλεία. Ο λόγος θα φανεί στο αμέσως επόμενο κεφάλαιο.

2.1 Στοιχεία Θεωρίας Πιθανοτήτων & Στατιστικής

2.1.1 Η έννοια της κατανομής πιθανότητας

Για έναν έμπειρο αναγνώστη, ο όρος «κατανομή πιθανότητας» ακούγεται μάλλον αμφίσημος, ή, τουλάχιστον, ελλειπής. Οι δύο όροι που θα συναντήσει κάποιος μελετώντας την αντίστοιχη βιβλιογραφία είναι οι «κατανομή μάζας πιθανότητας» και «κατανομή πυκνότητας πιθανότητας». Ο πρώτος αναφέρεται σε διακριτές τυχαίς μεταβλητές ενώ ο δεύτερος σε συνεχείς. Εμείς ασχολούμαστε μονάχα με συνεχείς μεταβλητές οπότε ας θυμηθούμε τον αντίστοιχο ορισμό:

Ορισμός 2.1.1. *Για μια συνεχή τυχαία μεταβλητή X, ορίζουμε ως συνάρτηση πυκνότητας πιθανότητας μία μη αρνητική συναρτήση f τέτοια ώστε:*

$$P(X \in A) = \int_A f_X(x)dx$$

Για περισσότερες πληροφορίες σχετικά με την θεμελίωση της Πιθανότητας όπως μελετάται σήμερα, παραπέμπουμε τον αναγνώστη στο [39].

Για το υπόλοιπο της εργασίας, όποτε αναφέρομαστε σε κάποια κατανομή πιθανότητας θα εννοούμε την αντίστοιχη συνάρτηση πυκνότητας πιθανότητας, εκτός αν ρήτα αναφέρεται διαφορετικά.

Θα δώσουμε ακόμη τους ορισμούς δύο κεντρικών ποσοτήτων στην μελέτη πιθανοτικών κατανομών -ποσότητες στις οποίες θα επιστρέφουμε ξανά και ξανά- και θα παραπέμψουμε τον αναγνώστη στο [3] για περαιτέρω μελέτη.

Ορισμός 2.1.2. *Για μία συνεχή τυχαία μεταβλητή X με κατανομή πιθανότητας f_X ορίζουμε ως μέση ή αναμενόμενη τιμή, και συμβολίζουμε με* $\mathrm{E}[X]$*, την ποσότητα:*

$$\mathrm{E}[X] = \int_{\mathbb{R}} x f_X(x) dx$$

Αν και μάλλον η μέση τιμή αποτελεί την απλούστερη αλλά και συνάμα κεντρικότερη έννοια της Θεωρίας Πιθανοτήτων, αυστηρός μαθηματικός ορισμός αυτής δεν διατυπώθηκε μέχρι και το 1812, ενάμιση αιώνα δηλάδη από την γέννηση του κλάδου, από τον Laplace στο [13].

Εναλλακτικά, την μέση τιμή συμβολίζουμε και με μ_x αντί για $\mathrm{E}[X]$. Ειδικά δε, αν είναι ξεκάθαρο σε ποια μεταβλητή αναφερόμαστε, απλοποιούμε τον συμβολισμό σε μ. Χάριν απλότητας, στα πλαίσια των επόμενων κεφαλαίων, αυτός θα είναι και ο συμβολισμός της επιλογής μας. Προχωράμε, τώρα στον επόμενο ορισμό μας, από το [20].

Ορισμός 2.1.3. *Για μία συνεχή τυχαία μεταβλητή X με κατανομή πιθανότητας f_X ορίζουμε ως διακύμανση και συμβολίζουμε με* $\mathrm{var}[X]$*, την ποσότητα:*

$$\mathrm{var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \int_{\mathbb{R}} (x - \mathrm{E}[X])^2 f_X(x) dx$$

Όπως και στην περίπτωση της μέσης τιμής, έτσι και εδώ υπάρχει ο εναλλακτικός συμβολισμός $\mathrm{var}[X] = \sigma_x^2$, η απλά σ^2. Αντίθετα, όμως, με την προηγούμενη περίπτωση, ο απλούστερος συμβολισμός αναφέρεται μόνο σε μονοδιάστατα προβλήματα. Παρακάτω, θα δούμε τον συμβολισμό για τα πολυδιάστατα, τον οποίο και θα χρησιμοποιούμε.

Και οι τρεις ορισμοί που προηγήθηκαν αφορούν μονοδιάστατες τυχαίες μεταβλητές. Δεδομένου, όμως, ότι ο στόχος μας είναι να εργαστούμε σε υψηλότερες διαστάσεις χρειαζόμαστε και τους αντίστοιχους ορισμούς των πολυδιάστατων ποσοτήτων. Αφήνουμε τον αναγνώστη να δει πώς οι ορισμοί 2.1.1 και 2.1.2 γενικεύονται κατά φυσικό τρόπο, αντικαθιστώντας την μονοδιάστατη τυχαία μεταβλητή $x \in \mathbb{R}$ με το d-διάστατο τυχαίο διάνυσμα $\boldsymbol{x} \in \mathbb{R}^d$ και τον μονοδιάστατο χώρο ολοκλήρωσης $\mathbb{R}$ με τον d-διάστατο $\mathbb{R}^d$. (Από αυτό το σημείο κι έπειτα, με την χρήση του λατινικού γράμματος d θα δηλώνουμε την διάσταση ενώ με την χρήση έντονων χαρακτήρων, π.χ. $\boldsymbol{\mu}$ αντί μ, θα εννοούμε πολυδιάστατα μεγέθη) Αντίθετα, ο ορισμός 2.1.3 δεν γενικεύεται με τόσο άμεσο τρόπο.

Είναι πλέον το κατάλληλο σημείο να εξηγήσουμε συνοπτικά την φυσική σημασία της μέσης τιμής και της διακύμανσης για να κατανοησούμε την προσαρμογή που απαιτείται για την δεύτερη. Η μέση τιμή μπορεί να χαρακτηρισθεί αντιπρόσωπος της κατανομής, υπό την έννοια ότι βρίσκεται στο κέντρο της. Με άλλα λόγια, 50% της κατανομής αντιστοιχεί σε τιμές μικρότερες της μέσης και 50% σε μεγαλύτερες. Η διακύμανση, τώρα, είναι ένα μέτρο

συγκέντρωσης γύρω από την μέση τιμή. Όσο μικρότερη η τιμή της διακύμανσης τόσο μικρότερη η πιθανότητα να απομακρυνθούμε από την μέση τιμή και το αντίστροφο. Η λέξη που μας υποδηλώνει ότι η γενίκευσή μας δεν επιτυγχάνεται με άμεσο τρόπο είναι το «γύρω». Ας σκεφτεί ο αναγνώστης για λίγο πώς ορίζουμε αυτό το «γύρω» σε υψηλότερες διαστάσεις, π.χ. για $d = 2$. Πιθανώς η πρώτη εικόνα που του έρχεται στο μυαλό είναι ένας κύκλος γύρω από την μέση τιμή. Ο κύκλος, όμως, εγγενώς περιέχει μια εξάρτηση μεταξύ των αξόνων x και y. Ας το δούμε και μέσα από ένα παράδειγμα: έστω ότι μελετάμε το ύψος h και το βάρος w μίας ομάδας ατόμων. Υπολογίζουμε τις μέσες τιμές μ_h και μ_w αντίστοιχα καθώς και την διακύμανση του ύψους σ_h^2, την οποία και βρίσκουμε να είναι αρκετά μεγάλη. Η εμπειρία μας μας λέει να περιμένουμε και αντίστοιχα μεγάλη διακύμανση του βάρους αφού ψηλότερα άτομα τείνουν να είναι βαρύτερα. Ωστόσο, αυτήν μας την εμπειρική παρατήρηση δεν είμαστε ακόμη σε θέση να την διατυπώσουμε μαθηματικά. Αυτό το πρόβλημα επιλύει ο επόμενος ορισμός:

Ορισμός 2.1.4. *Για δύο τυχαίες μεταβλητές X και Y ορίζουμε ως συνδιακύμανση, και συμβολίζουμε με* $\mathrm{Cov}(X, Y)$*, την ποσότητα:*

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$

Σημείωση 2.1.1. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$

Σημείωση 2.1.2. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$

Επιστρέφοντας στο παράδειγμά μας, βλέπουμε ότι χρειαζόμαστε τελικά τέσσερις ποσότητες για να περιγράψουμε την συνολική διακύμανση γύρω από την μέση τιμή: τις διακυμάνσεις σ_h^2 και σ_w^2 και τις συνδιακυμάνσεις $\mathrm{Cov}(h, w)$ και $\mathrm{Cov}(w, h)$. Η σημείωση 2.1.1 μας λέει ότι στην πραγματικότητα αρκούν τρεις αλλά ας το αγνοήσουμε αυτό προς ώρας και ας ξεκινήσουμε με την εφαρμογή της σημείωσης 2.1.2, ξαναγράφοντας τις απαιτούμενες ποσότητες ως: $\mathrm{Cov}(h, h), \mathrm{Cov}(h, w), \mathrm{Cov}(w$ και $\mathrm{Cov}(w, w)$. Σε μεγαλύτερες διαστάσεις θα ήταν κουραστικό να παραθέτουμε κάθε φορά όλες τις συνδιακυμάνσεις. Ευτυχώς, παρατηρούμε ένα μοτίβο που μας επιτρέπει να τις γράψουμε με πιο συμπαγή τρόπο:

Ορισμός 2.1.5. *Για ένα τυχαίο* d*-διάστατο διάνυσμα $\boldsymbol{X} = [X_1, \ldots, X_d]$ ορίζουμε ως πίνακα συνδιακύμανσης, και συμβολίζουμε με* $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{X})$*, την ποσότητα:*

$$\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{X}) = \mathrm{E}[(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])^T]$$

Ή σε μορφή πίνακα:

$$\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{X}) = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \ldots & \mathrm{Cov}(X_1, X_d) \\ \vdots & \ddots & \\ \mathrm{Cov}(X_d, X_1) & \ldots & \mathrm{Cov}(X_d, X_d) \end{bmatrix}$$

Όπως και παραπάνω, έχει καθιερωθεί ο απλούστερος συμβολισμός $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}}$, ή $\boldsymbol{\Sigma}$ αν δεν υπάρχει κίνδυνος σύγχυσης, τον οποίο και θα ακολουθήσουμε.

2.1.2 Η Γκαουσσιανή κατανομή

Συνεχίζουμε με τη δεύτερη έννοια, αυτή της Γκαουσσιανής κατανομής. Όπως και στην προηγούμενη υποενότητα, αν και χρειαζόμαστε την πολυδιάστατη κατανομή, θα ξεκινήσουμε πρώτα με τον ορισμό της μονοδιάστατης περίπτωσης:

Ορισμός 2.1.6. *Μία συνεχής τυχαία μεταβλητή X λέμε ότι ακολουθεί Γκαουσσιανή ή κανονική κατανομή αν η συνάρτηση κατανομής της είναι:*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu)^2/2\sigma^2}$$

Όπου μ η μέση τιμή της κατανομής και σ^2 η διακύμανσή της, με $\sigma > 0$. Συμβολίζουμε ως $X \sim \mathcal{N}(\mu, \sigma^2)$ όπου το σύμβολο $\sim$ σημαίνει «ακολουθεί».

Η κανονική κατανομή έχει καθιερωθεί ως η κατανομή του Gauss, ο οποίος την όρισε σε ένα από τα σημαντικότερα έργα του([26]), αλλά ενίοτε αναφέρεται και ως κατανομή Gauss - Laplace λόγω της εκτενούς μελέτης του δεύτερου. Ωστόσο, η πρώτη αναφορά σε αυτήν γίνεται κιόλας από τον de Moivre στο [14]. Να προσθέσουμε σε αυτό το σημείο ότι το *Doctrine of chances*, όπως αναφέρεται για συντομία, αποτελεί το πρώτο σύγγραμα Θεωρίας Πιθανότητων. Μερικά παραδείγματα γκαουσσιανών κατανομών απεικονίζονται στο σχήμα 2.1. Να αναφέρουμε ότι συχνά η $\mathcal{N}(0,1)$(μπλε καμπύλη) συμβολίζεται με ϕ και αναφέρεται ως *τυπική κανονική κατανομή*.

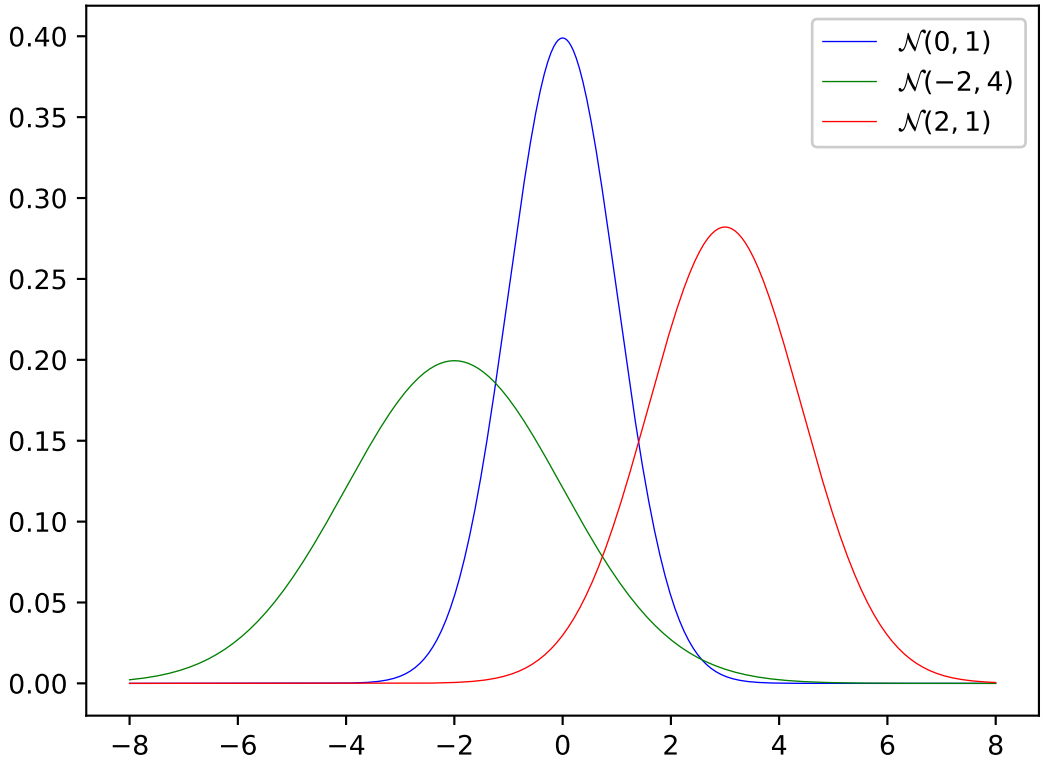
Αν και ξεφεύγει λίγο από την σκοπιά της εργασίας μας, δεν μπορούμε να αναφερθούμε στην κανονική κατανομή χωρίς να κάνουμε κάποια αναφορά και στο κεντρικό ρόλο που διαδραματίζει στην μελέτη της Θεωρίας Πιθανοτήτων. Ένα από τα διασημότερα αποτελέσματα αυτής είναι το *Κεντρικό Οριακό Θεώρημα*. Άτυπα, το θεώρημα αυτό διατυπώνει πώς ο μέσος όρος από ανεξάρτητες επαναλήψεις του ίδιου τυχαίου πειράματος ακολουθεί κανονική κατανομή. Για τον λόγο αυτό βρίσκουμε ότι πολλές ποσότητες που μετράμε στην καθημερινότητά μας ακολουθούν κάποια γκαουσσιανή κατανομή. Τέτοιες ποσότητες είναι το ύψος, το βάρος κατά την γέννηση ή το μέγεθος του παπουτσιού. Για εκτενέστερη μελέτη σχετικά με το Κεντρικό Οριακό Θεώρημα και τις εφαρμογές του θα παραπέμψουμε στο [19] και θα κλείσουμε την παρένθεσή μας λέγοντας πως μια πρώιμη μορφή του θεωρήματος αποδεικνύεται ήδη στη δεύτερη έκδοση του Doctrine of chances! Περισσότερες πληροφορίες σχετικά με το έργο στο [55].

Ας περάσουμε τώρα και στον αντίστοιχο ορισμό υψηλότερων διαστάσεων.

Ορισμός 2.1.7. *Ένα τυχαίο d-διάστατο διάνυσμα $\boldsymbol{X}$ λέμε ότι ακολουθεί d-διάστατη Γκαουσσιανή ή κανονική κατανομή αν η συνάρτηση κατανομής του είναι:*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Όπου $\boldsymbol{\mu}$ η μέση τιμή της κατανομής και $\boldsymbol{\Sigma}$ ο πίνακας συνδιακύμανσή της, με $\boldsymbol{\Sigma}$ θετικά ημιορισμένο.



Σχήμα 2.1: Μονοδιάστατες Γκαουσσιανές κατανομές για διάφορες τιμές των παραμέτρων

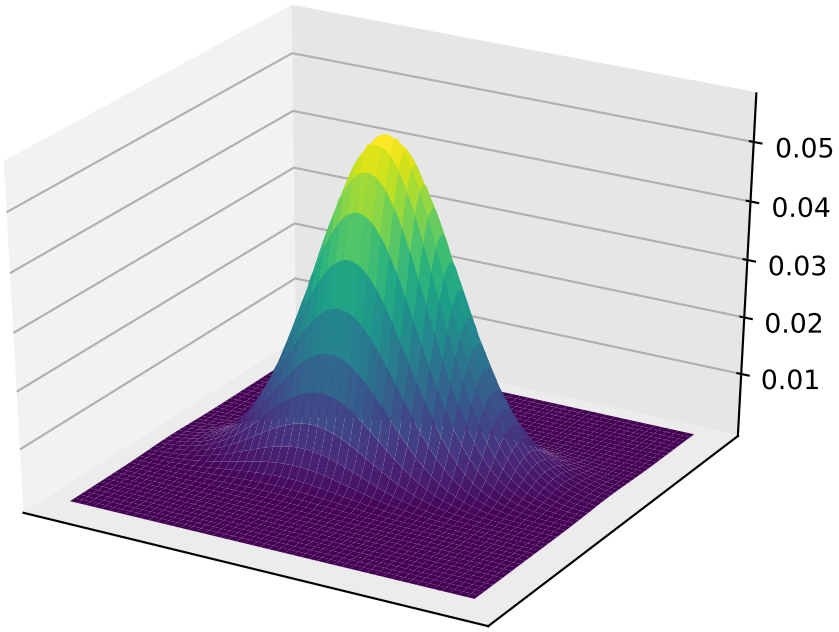
Να εξηγήσουμε συνοπτικά ότι ένας πίνακας είναι θετικά ημιορισμένος αν όλες οι ιδιοτιμές του είναι μη αρνητικές και να στρέψουμε τον αναγώστη στο [61] για μια εισαγωγή στην σχετιζόμενη Γραμμική Άλγεβρα. Ακόμη, κατά αντιστοιχία με την μονοδιάστατη περίπτωση, γράφουμε $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Φυσικά, όσο μεγαλώνουν οι διαστάσεις χάνουμε την δυνατότητα να τις οπτικοποιήσουμε αλλά ας δούμε ένα παράδειγμα για $d = 2$ στο σχήμα 2.2.

Για αναλυτικότερη μελέτη σχετικά με τς πολυδιάστατες κανονικές κατανομές παραπέμπουμε στο ομότιτλο βιβλίο του Tong, [63].

2.1.3 Αποκομμένη Στατιστική

Με πολύ απλά λόγια, η έννοια της αποκοπής στη Στατιστική περιγράφει το εξής φαινόμενο: έστω ότι επιστρέφουμε στο σχήμα 2.1 και, ενώ μελετάμε τις τρεις καμπύλες, αποφασίζουμε να καλύψουμε, να «αποκόψουμε» δηλαδή, ένα τμήμα του σχήματος με το χέρι μας. Προφανώς, δεν έχουμε αλλάξει τις ίδιες τις κατανομές, αλλά αυτό που εμείς βλέπουμε και συνεπώς μπορούμε να μελετήσουμε έχει μεταβληθεί. Θα ξεκινήσουμε με τον μαθηματικό ορισμό της αποκοπής και εν συνέχεια θα δούμε τι το σημαντικό έχει και μας οδήγησε στην μελέτη της.

Ορισμός 2.1.8. *Για μια συνεχή τυχαία μεταβλητή X με συνάρτηση πυκνότητας πιθανότητας*



Σχήμα 2.2: Διδιάστατη Γκαουσσιανή κατανομή

f ορίζουμε ως αποκομμένη συνάρτηση πυκνότητας πιθανότητας g σε ένα ***άγνωστο*** *σύνολο S:*

$$g(x) = P(x \in A | x \in S) = \begin{cases} \frac{\int_A f_X(x)dx}{\int_S f_X(x)dx} & A \in S \\ 0 & A \notin S \end{cases}$$

Το σύνολο S καλείται σύνολο αποκοπής.

Εκ πρώτης όψεως ίσως φαίνεται λίγο δύσκολο να σκεφτούμε κάποιες πρακτικές εφαρμογές της αποκομμένης Στατιστικής οπότε, πρώτα, ας χαλαρώσουμε τον περιορισμό επιτρέποντάς μας γνώση του συνόλου S. Σε αυτήν την περίπτωση μιλάμε για *Στατιστική με λογοκρισία*. Κάθε φορά που συναντάμε ένα ερωτηματολόγιο με επιλογές του τύπου «Περισσότερες από 8 ώρες» ή «Λιγότερες από 2 φορές» βλέπουμε ένα παράδειγμα Στατιστικής με λογοκρισία. Γνωρίζουμε την ύπαρξη ενός δείγματος αλλά το μόνο που γνωρίζουμε για την τιμή του είναι ότι ανήκει σε διαστήματα της μορφής $(8, \infty]$ ή $[0, 2)$. Συνεπώς μιλάμε για *δεξιά* ή *αριστερή λογοκρισία* αντίστοιχα. Επιστρέφοντας στο πιο σθεναρό πλαίσιο της αποκοπής, η μόνη διαφορά έγκειται στο γεγονός ότι δεν γνωρίζουμε την ύπαρξη ή μη του δείγματος. Για παράδειγμα, έστω ότι προσπαθούμε να εκτιμήσουμε τον αριθμό των άστρων σε κάποιο γαλαξία πολύ πολύ μακρυά. Ξέρουμε ότι υπάρχουν ορισμένα άστρα των οποίων το φως δεν μπορούμε να ανιχνεύσουμε ακόμη αλλά δεν ξέρουμε πού και πόσα. Αυτά αποτελούν αποκομμένα δείγματα.

Ελπίζουμε ο αναγνώστης να συμμερίζεται πλέον την τάση να μελετούνται κοινά τα δύο πλαίσια. Για την ώρα, θα παραπέμπουμε στα [1] και [9] για περισσότερη εξοικείωση με αυτόν τον κλάδο της Στατιστικής και θα επιστρέψουμε στο Κεφάλαιο 4 με περισσότερες αναφορές στην μελέτη και την εκμάθηση από αποκομμένα ή/και λογοκριθέντα δείγματα.

2.2 Εργαλεία από τον χωρό των Πιθανοτήτων

Στην παρούσα ενότητα θα δούμε δύο θέματα που θα μας εξοπλίσουν για την συνέχεια. Το πρώτο αφορά στην απόσταση μεταξύ κατανομών και θα μας προσφέρει μετρικές αξιολόγησης των διάφορων αλγορίθμων που θα μελετήσουμε αργότερα. Το δεύτερο θα μας εισάγει την συνάρτηση πιθανοφάνειας ώστε να είμαστε σε θέση να κατανοήσουμε την πρώτη μέθοδο στατιστικής μάθησης που θα συναντήσουμε στην ενότητα 3.1.

2.2.1 Αποστάση μεταξύ κατανομών

Είμαστε αρκετά εξοικειωμένοι με την έννοια της απόστασης όταν πρόκειται για απόσταση μεταξύ δύο σημείων. Πώς, όμως, επεκτείνουμε αυτήν την έννοια στην περίπτωση κατανομών πιθανότητας; Το ερώτημα δεν είναι απλό και απόδειξη για αυτό αποτελεί το γεγονός ότι διάφορες μετρικές έχουν προταθεί στο πέρασμα των ετών, προερχόμενες από κάποια φυσική διαίσθηση ή από κάποια μαθηματική ανάγκη, και όλες έχουν βρει τις χρήσεις τους. Ίσως η πιο διαδεδομένη είναι η *απόσταση ολικής μεταβολής*.

Ορισμός 2.2.1. *Για δύο μέτρα πιθανότητας $\mathcal{P}$ και $\mathcal{Q}$ ορισμένα σε μία σ-άλγεβρα $\mathcal{F}$ των υποσυνόλων του δειγματικού χώρου Ω ορίζουμε την απόσταση ολικής μεταβολής d_{TV} ως:*

$$\mathrm{d}_{TV} = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$$

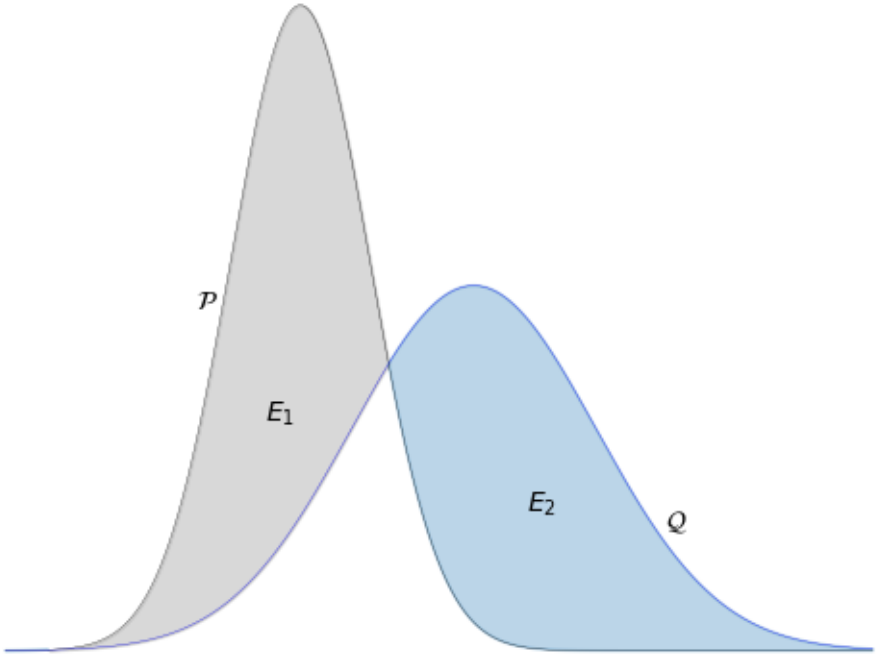
Δεν θα επεκταθούμε στους μαθηματικούς ορισμούς των μέτρων πιθανότητας και της σ-άλγεβρας από την σκοπιά της Θεωρίας Μέτρου. Θα στρέψουμε τον ενδιαφερόμενο αναγνώστη στο [36] και θα απλοποιήσουμε την δουλειά μας λογιζόμενοι την περίπτωση συνεχών μέτρων, όπως ισχύει και με τις πολυδιάστατες Γκαουσσιανές. Τότε, η απόσταση ολικής μεταβολής δίνεται από τον τύπο:

$$\mathrm{d}_{TV} = \frac{1}{2} \int_{\mathbb{R}^d} |p(x) - q(x)| dx \tag{2.1}$$

όπου p, q οι αντίστοιχες συναρτήσεις πυκνότητας πιθανότητας.

Διαισθητικά, η απόσταση ολικής μεταβολής δηλώνει την μεγαλύτερη δυνατή διαφορά πιθανότητας που μπορούν οι δύο κατανομές να αναθέσουν στο ίδιο γεγονός. Στο σχήμα 2.3 βλέπουμε ένα μονοδιάστατο παράδειγμα της απόστασης μεταξύ δύο κανονικών κατανομών. Αφήνουμε στον αναγνώστη να συσχετίσει τα E_1 και E_2 μέσως της σχέσης 2.1.

Συνεχίζοντας, πρέπει να τονίσουμε ότι δυστυχώς δεν υπάρχει κλειστος τύπος για την 2.1. Αυτό σημαίνει ότι αργότερα θα κληθούμε να λάβουμε μία απόφαση: είτε θα υπολογίζουμε



Σχήμα 2.3: Απόσταση ολικής μεταβολής μεταξύ γκαουσσιανών

προσεγγιστικά τα πολυδιάστατα ολοκληρώματα που υπεισέρχονται στην σχέση είτε θα χρησιμοποιούμε κάποια φράγματα, προερχόμενα είτε άμεσα από την απόσταση ολικής μεταβολής είτε έμμεσα από την σχέση της με άλλες αποστάσεις. Ενδεχομένως η πρώτη επιλογή να φαίνεται αναμενόμενη στον αναγνώστη. Ωστόσο, η μελέτη των φραγμάτων παρουσιάζει μεγάλο ενδιαφέρον, τόσο απο θεωρητικής άποψης όσο και από πρακτικής σκοπιάς. Στο [16] μπορεί ο αναγνώστης να μελέτησει τα καλύτερα, τουλάχιστον εξ όσων γνωρίζουμε κατά τη στιγμή της συγγραφής, φράγματα για την περίπτωση των Γκαουσσιανών κατανομών. Τέλος, θα κλείσουμε την παρούσα ενότητα με μία παρατήρηση: η επιλογή της μετρικής(ή του φράγματος αυτής) δεν είναι απλή υπόθεση και μπορεί να επηρεάσει την σύγκλιση ενός αλγορίθμου, ποσοτικά ή ποιοτικά. Στο [27] μπορούμε να δούμε μία επισκόπηση των γνωστότερων μετρικών απόστασης καθώς και πώς αυτές επιδρούν στα αποτελέσματα.

2.2.2 Η συνάρτηση πιθανοφάνειας

Πριν κλείσουμε το πρώτο κεφάλαιο μαθηματικής θεμελιώσης της διπλωματικής, θα μελετήσουμε έναν ακόμη ορισμό. Στην υποενότητα 2.1.1 είδαμε την έννοια της κατανομής πιθανότητας, όπου με δεδομένες τις παραμέτρους κάποιας κατανομής υπολογίζαμε την πιθανότητα ενός δείγματος. Όταν θα θελήσουμε να μάθουμε την κατανομή, δηλαδή τις παραμέτρους αυτής, ο στόχος μας θα είναι το αντίστροφο, με δεδομένα τα δείγματα να υπολογίζουμε την πιθανότητα των παραμέτρων. Αυτή ακριβώς την έννοια όρισε ο Fisher στο [23].

Ορισμός 2.2.2. *Για μία συνεχή τυχαία μεταβλητή X με κατανομή πιθανότητας f_θ ορίζουμε*

την συνάρτηση πιθανοφάνειας $\mathcal{L}$ δεδομένου δείγματος x ως:

$$\mathcal{L}(\theta|x) = f_\theta(x)$$

Να τονίσουμε ότι η συνάρτηση πιθανοφάνειας είναι συνάρτηση της άγνωστης παραμέτρου θ και όχι του δείγματος x το οποίο θεωρείται γνωστό. Εναλλακτικά συμβολίζεται και ως $\mathcal{L}(\theta;x)$. Θα κρατήσουμε αυτόν για λόγους συνέπειας με την εργασία των Daskalakis et al.

Η αξία της έγκειται στο γεγονός ότι κλιμακώνεται πολύ εύκολα για πολλά δείγματα x. Έστω ότι έχουμε δύο *ανεξάρτητα* δείγματα της άγνωστης, ως προς τις παραμέτρους, κατανομής x_1 και x_2. Εφαρμόζωντας τον πολλαπλασιαστικό κανόνα των πιθανοτήτων λαμβάνουμε:

$$\mathcal{L}(\theta;x_1, x_2) = f_\theta(x_1) \cdot f_\theta(x_2)$$

Και γενικεύοντας σε n δείγματα $x_1, \ldots, x_n$:

$$\mathcal{L}(\theta;x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

Για να την κατανοήσουμε καλύτερα, ας υπολογίσουμε την συνάρτηση πιθαναφάνειας στην περίπτωση μιας μονοδιάστατης κανονικής κατανομής, που θα μας χρησιμεύσει και στην συνέχεια. Σε αυτήν την περίπτωση η παράμετρος θ είναι το διάνυσμα των παραμέτρων της κανονικής, $\theta = (\mu, \sigma)$.

Παράδειγμα 2.1.

$$\begin{aligned}
\mathcal{L}(\mu, \sigma;x_1, \ldots, x_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x_i-\mu)^2/2\sigma^2} \\
&= \frac{1}{(\sqrt{2\pi}\sigma)^n} \cdot \prod_{i=1}^{n} \mathrm{e}^{-(x_i-\mu)^2/2\sigma^2} \\
&= \frac{1}{(\sqrt{2\pi}\sigma)^n} \cdot \mathrm{e}^{\left(\sum_{i=1}^{n} -(x_i-\mu)^2/2\sigma^2\right)}
\end{aligned}$$

Εν προκειμένω, θα θέλαμε πολύ να απαλλαχθούμε από το εκθετικό. Και πράγματι, αυτή είναι μια τόσο συχνή επιθυμία κατά την μελέτη των συναρτήσεων πιθανοφάνειας ώστε να μας οδηγήσει στην εισαγωγή του συμβόλου ℓ για τον λογάριθμο της $\mathcal{L}$. Έτσι, καταλήγουμε πως

$$\begin{aligned}
\ell(\mu, \sigma;x_1, \ldots, x_n) &= \ln\left(\mathcal{L}(\mu, \sigma;x_1, \ldots, x_n)\right) \\
&= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}$$

Κεφάλαιο 3

Θεωρητικό Υπόβαθρο ΙΙ: Μηχανική Μάθηση

Σε αντίθεση με το προηγούμενο κεφάλαιο, εδώ δεν θα ξεκινήσουμε με ορισμούς. Για να καταφέρουμε να δώσουμε έναν αυστηρό ορισμό της Μηχανικής Μάθησης χρειαζόμαστε πρώτα έναν αυστηρό ορισμό για την έννοια της Μάθησης. Καλούμε τον αναγνώστη, που μάλλον έχει αφιερώσει αρκετά χρόνια από την ζωή του στην διαδικασία της μάθησης, να σκεφτεί έναν τέτοιο ορισμό. Μάλλον θα δυσκολευτεί. Και δικαίως, καθώς γενικά αποδεκτός ορισμός δεν υπάρχει. Βέβαια, όλοι οι ορισμοί περιγράφουν με τον ένα ή με τον άλλο τρόπο την Μάθηση σαν μια διαδικασία απόκτησης γνώσης και κατανόησης. Σε αντιστοιχία με την περίπτωση της Μάθησης στους ανθρώπους, ή και στα ζώα, μπορούμε να πούμε ότι η Μηχανική Μάθηση είναι μία διαδικασία μάθησης κατά την οποία η οντότητα που μαθαίνει είναι μία μηχανή, ένας υπολογιστής. Ίσως ο πιο διάσημος ορισμός της έννοιας είναι αυτός στο [44]:

> "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."
>
> *Tom Mitchell*

Πριν απαριθμήσουμε μερικές κατηγορίες Μηχανικής Μάθησης, ας δούμε την σχέση μεταξύ Μηχανικής Μάθησης και Τεχνητής Νοημοσύνης. Για να την καταλάβουμε, ας γυρίσουμε για λίγο στο γενικό πλαίσιο της Μάθησης και την σχέση της με την Νοημοσύνη. Η ικανότητα της μάθησης είναι ένα χαρακτηριστικό των νοήμονων όντων. Ωστόσο, δεν αρκεί από μόνη της για να χαρακτηρίσει την νόηση. Η ίδια σχέση υποσυνόλου διατηρείται όταν μεταβαίνουμε και στον ψηφιακό κόσμο. Η Μηχανική Μάθηση είναι ένας κλάδος της Τεχνητής Νοημοσύνης. Ίσως ο πιο ραγδαία αναπτυσσόμενος κλάδος αυτής, αλλά και πάλι, μόνο ένας κλάδος. Οπότε, σαφώς, οι δύο έννοιες δεν πρέπει να συγχέονται ή να θεωρούνται ταυτόσημες, όπως συχνά παρατηρείται. Μία απλή διάκριση είναι η εξής: η Τεχνητή Νοημοσύνη έχει συμβολική φύση και προσπαθεί να συμπεράνει κανόνες ενώ η Μηχανική Μάθηση έχει στατιστική φύση και προσπάθει να αποκτήσει νέα γνώση μέσα από τα δεδομένα. Με άλλα λόγια, η πρώτη επιδιώκει να λύσει ένα σύνθετο πρόβλημα μιμούμενη την ανθώπρινη νοημοσύνη ενώ η δεύτερη να βελτιώσει την απόδοσή της στο ίδιο πρόβλημα με βάση την εμπειρία της σε αυτό.

Συνήθως, χωρίζουμε την Μηχανική Μάθηση σε 3 μεγάλες κατηγορίες:

- *Επιβλεπόμενη μάθηση*
  Σε αυτό το είδος μάθησης, τα δεδομένα είναι εφοδιασμένα με μία *ετικέτα* που παρέχει επιπρόσθετη πληροφορία. Ας το δούμε, πρώτα, μέσα από ένα παράδειγμα: έστω ότι έχουμε κάποιες συνόψεις ταινιών με ετικετά αν είναι επιστημονικής φαντασίας ή όχι. Προφανώς, θέλουμε ο αλγόριθμος να μαθαίνει όταν διαβάζει μία καινούρια σύνοψη αν η ταινία ανήκει στο είδος. Πρακτικά αυτό σημαίνει ότι απαιτούμε την κατασκευή ενός κανόνα που αντιστοιχεί συνόψεις σε ένα ναι/όχι για την επιστημονική φαντασία, βασισμένο στις υπάρχουσες ετικέτες και εφαρμόσιμο σε άγνωστες ταινίες. Ο όρος επιβλεπόμενη σημαίνει ότι κατά τη διαδικασία της μάθησης «επιβλέπουμε» τον αλγόριθμο, βοηθώντας να αναπτύξει τον κανόνα μέσω της παροχής ετικετών.

- *Μη Επιβλεπόμενη μάθηση*
  Όπως προδίδει το όνομα, πλέον δεν έχουμε την δυνατότητα να τοποθετούμε ετικέτες στα δείγματά μας. Οπότε, στο κινηματογραφικό μας παράδειγμα δεν μπορούμε να ζητήσουμε από τον αλγόριθμο να ξεχωρίσει τις ταινίες επιστημονικής φαντασίες από τις υπόλοιπες. Πολύ απλά γιατί δεν γνωρίζει πια την έννοια της επιστημονικής φαντασίας. Ωστόσο, μπορούμε ακόμη να μάθουμε χρήσιμες πληροφορίες, μέσα από τις κατάλληλες ερωτήσεις. Για παράδειγμα, μπορούμε να μάθουμε αν δύο ταινίες ανήκουν στο ίδιο είδος, αρκεί ο αλγόριθμος να ομαδοποιεί κατά κάποιον τρόπο τις ταινίες. Η *ομαδοποίηση*(Clustering) αποτελεί το τυπικότερο παράδειγμα μη επιβλεπόμενης μάθησης.

- *Ενισχυτική μάθηση*
  Η τρίτη κατηγορία δεν αφορά ετικέτες, τουλάχιστον όχι άμεσα. Η ενισχυτική μάθηση μελετάει καταστάσεις που απαιτούν ενέργειες από έναν ή περισσότερους πράκτορες με στόχο την μεγιστοποίηση κάποιας ανταμοιβής. Πολύ συχνά την συναντάμε στην προσπάθεια βελτίωσης του υπολογιστή σε κάποιο παιχνίδι. Για παράδειγμα, στο σκάκι ένα πρόγραμμα ενισχυτικής μάθησης βελτιώνεται βλέποντας όλες τις κινήσεις μίας παρτίδας αλλά γνωρίζοντας μόνο το τελικό αποτέλεσμα. Οπότε η ενισχυτική μάθηση βρίσκεται, κατά κάποιον τρόπο, μεταξύ των άλλων δύο αφού υπάρχει μία ετικέτα για το αποτέλεσμα μεν αλλά ο αλγόριθμος παράγει πολύ περισσότερα από μία ετικέτα για κάθε νέα παρτίδα δε. Για περισσότερα σχετικά με την ενισχυτική μάθηση σε παιχνίδια στρατηγικής ο αναγνωστής μπορεί να ξεκινήσει από το [59].

Τόσο για την εισαγωγή όσο και για το υπόλοιπο μέρος του κεφαλαίου ακολουθούμε το [56], για να συμφωνούμε και με τον φορμαλισμό του [11]. Πριν προχωρήσουμε, όμως, είναι μάλλον απαραίτητο να δούμε πως αντιμετωπίζουμε την μηχανική μάθηση στα πλαίσια της μάθησης κατανομών. Ακριβώς αυτό όρισαν οι Kearns et al. στην σεμιναριακή εργασία τους στο [38].

Ορισμός. *Έστω C μία κλάση κατανομών. Λέμε ότι η C μαθαίνεται αποδοτικά αν για κάθε $\varepsilon > 0$, $0 < \delta \leq 1$ και δεδομένης πρόσβασης σε έναν γεννήτορα μιας άγνωστης κατανομής*

$D \in C$ υπάρχει ένας αλγόριθμος πολυωνυμικού χρόνου A, που καλείται ο αλγόριθμος μάθησης της C, με έξοδο έναν γεννήτορα ή εκτιμητή της κατανομής D' τέτοιον ώστε

$$\mathbb{P}[d(D, D') \leq \varepsilon] \geq 1 - \delta$$

όπου $d(\cdot, \cdot)$ κάποια συνάρτηση απόστασης.

Μάλιστα, αν οι κατανομές που ανήκουν στην C είναι της ίδιας οικογένειας, όπως και στην περίπτωσή μας με τις κανονικές, τότε αρκεί ο προσδορισμός των παραμέτρων τους. Οπότε ο A καλείται *αλγόριθμος μάθησης παραμέτρων*.

3.1 Στοιχεία Μηχανικής Μάθησης

Σε αυτήν την ενότητα, στόχος μας είναι να μελετήσουμε τις τεχνικές της *Εκτιμήτριας Μέγιστης Πιθανοφάνειας*(Ενότητα 3.1.1) και της *Στοχαστικής Κατάβασης Κλίσης*(Ενότητα 3.1.3) που θα μας χρειαστούν στην συνέχεια. Βέβαια, δεν μπορούμε να μιλήσουμε για στοχαστική κατάβαση κλίσης αν δεν αναφερθούμε πρώτα στην απλή Κατάβαση Κλίσης(Ενότητα 3.1.2).

3.1.1 Εκτιμήτρια Μέγιστης Πιθανοφάνειας

Η μέθοδος αυτή, όπως προδίδει το όνομά της, προσπαθεί να μεγιστοποιήσει την συνάρτηση πιθανοφάνειας που ορίσαμε στο προηγούμενο κεφάλαιο. Εισήχθηκε και αυτή από τον Fisher, [21], λίγο αργότερα από τον ορισμό της συνάρτησης. Για να την κατανοήσουμε, ας αφήσουμε για λίγο στην άκρη τις πιθανότητες και ας την αντιμετώπισουμε ως μία απλή συνάρτηση των παραμέτρων. Θέλουμε να βρούμε ένα βρούμε ένα $\hat{\boldsymbol{\theta}}$ τέτοιο ώστε:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\theta)$$

Όπως ξέρουμε από την Ανάλυση, μπορούμε να βρούμε το $\hat{\boldsymbol{\theta}}$ μηδενίζοντας τις μερικές παραγώγους, $\dfrac{\partial \mathcal{L}}{\partial \theta_i} = 0$. Ας το δούμε στην πράξη, συνεχίζοντας από εκεί που είχαμε σταματήσει στο παράδειγμα 2.1.

Παράδειγμα 3.1. Για την ειδική περίπτωση μίας κανονικής κατανομής, είχαμε υπολογίσει ότι:

$$\mathcal{L}(\mu, \sigma; x_1, \ldots, x_n) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \cdot \mathrm{e}^{\left(\sum_{i=1}^{n} -(x_i - \mu)^2 / 2\sigma^2\right)}$$

Η παραγώγιση, είτε ως προς μ είτε ως προς σ, αυτής της έκφρασης δεν φαντάζει ιδανική. Ενθυμούμενοι ότι ο λογάριθμος είναι φθίνουσα συνάρτηση, ξέρουμε ότι η συμπεριφορά ως προς τα ακρότατα της $\mathcal{L}(\mu, \sigma; x_1, \ldots, x_n)$ είναι ίδια με της $-\ell(\mu, \sigma; x_1, \ldots, x_n)$ οπότε χρησιμοποιώντας

το αποτέλεσμα του προηγούμενου παραδείγματος λαμβάνουμε:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$
$$\frac{\partial \ell}{\partial \sigma} = \frac{n}{\sigma} - \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2$$

Εξισώνοντας τις δύο παραγώγους με το μηδέν λαμβάνουμε:

$$\left.\begin{array}{l}\frac{\partial \ell}{\partial \mu} = 0\\ \frac{\partial \ell}{\partial \sigma} = 0\end{array}\right\} \implies \begin{array}{l}\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i\\ \hat{\sigma} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}\end{array}$$

Απλή στην εφαρμογή αλλά τα πρώτα χρόνια χρησιμοποιούνταν ως ευρεστική τεχνική. Την πρώτη απόδειξη κατανομής του σφάλματος παρουσίασε ο Wilks στο [68]. Αν και εξαιρετικά χρήσιμο εργαλείο, δεν μπορεί να αντιμετώπισει με επιτυχία περιπτώσεις που δεν επιδέχονται κλειστό τύπο. Θα δούμε σε λίγο πώς θα υπερκέρασουμε αυτό το εμπόδιο στην δική μας περίπτωση.

3.1.2 Κατάβαση Κλίσης

Στην παρούσα ενότητα θα μελετήσουμε τον αλγόριθμο Κατάβασης Κλίσης(Gradient Descent ή για συντομία GD). Η βασική ιδέα πηγαίνει πίσω στον Cauchy, [8], και είναι η εξής: γνωρίζουμε από την Ανάλυση ότι η παράγωγος μίας συνάρτησης f δηλώνει την κατεύθυνση προς την οποία μεγιστοποιείται και άρα προς την αντίθετη κατεύθυνση ελαχιστοποιείται. Οπότε, αν από ένα σημείο w μεταβούμε σε ένα νεό σημείο w' σύμφωνα με τον κανόνα:

$$\mathbf{w}' = \mathbf{w} - \eta \nabla f(\mathbf{w})$$

για κατάλληλα μικρό η θα έχουμε $f(\mathbf{w}') \leq f(\mathbf{w})$. Αν, τώρα, το $\mathbf{w}$ είναι ελάχιστο της συνάρτησης ξέρουμε από το Θεώρημα $Fermat$ ότι $\nabla f(\mathbf{w}) = 0$ οπότε θα παραμείνουμε στο ίδιο σημείο. Συνεπώς, μπορούμε να επαναλάβουμε την ίδια διαδικασία εώς ότου βρεθούμε στο ελάχιστο ή αρκετά κοντά σε αυτό. Ο γενικός κανόνας τότε γίνεται:

$$\mathbf{w}^{(i+1)} = \mathbf{w}^{(i)} - \eta \nabla f(\mathbf{w}^{(i)})$$

όπου ο εκθέτης i δηλώνει το βήμα και $\mathbf{w}^{(0)}$ την (κατάλληλη) αρχική εκτίμηση.

Μάλλον ο αναγνώστης περίμενε η τελευταία εντολή να επιστρέφει το $\mathbf{w}^{(M)}$ οπότε έχουμε κάποιες εξηγήσεις να δώσουμε. Ιδανικά, θα επιστρέφαμε πράγματι το αποτελέσμα της τελευταίας επανάληψης. Ωστόσο, μία τέτοια επιλογή δεν γενικεύει αρκετά καλά. Για παράδειγμα, αν η f δεν είναι παραγωγίσιμη μπορούμε να εφαρμόσουμε τον αλγόριθμο χρησιμοποιώντας κάποια προσέγγιση του gradient. αρκεί να χρησιμοποιήσουμε τον μέσο όρο ως έξοδο για να διορθώσουμε τυχόν σφάλματα προσέγγισης. Βέβαια, το πιο κλασικό παράδειγμα που απαιτεί χρήση του μέσου όρου είναι η Στοχαστική Κατάβαση Κλίσης. Πριν πάμε, όμως, στην επόμενη ενότητα ας διατυπώσουμε κάποιους ορισμούς και ένα φράγμα για το σφάλμα του αλγορίθμου:

Αλγόριθμος 1 Κατάβαση Κλίσης

```
1: procedure GD(M,η)                ▷ M: αριθμός επαναλήψεων, η: παράμετρος
2:     initialize w^(0) ← 0
3:     for i = 1, ..., M
4:         v_i ← ∇f(w^(i−1))
5:         w^(i) ← w^(i−1) − ηv_i
6:     w̄ ← (1/M) Σ_{i=1}^{M} w^(i)
7:     return w̄
```

Ορισμός 3.1.1. *Ένα σύνολο C ενός διανυσματικού χώρου καλείται κυρτό αν για οποιαδήποτε δύο διανύσματα $\mathbf{u}, \mathbf{v} \in C$ το ευθύγραμμο τμήμα που τα ενώνει περιέχεται στο C. Δηλαδή:*

$$\forall \alpha \in [0,1] : \alpha\mathbf{u} + (1-\alpha)\mathbf{v} \in C$$

Ορισμός 3.1.2. *Έστω κυρτό σύνολο C. Μία συνάρτηση $f : \mathbb{R} \to C$ καλείται κυρτή αν για κάθε $\mathbf{u}, \mathbf{v} \in C$ και $\alpha \in [0,1]$ ισχύει:*

$$f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \leq \alpha f(\mathbf{u}) + (1-\alpha) f(\mathbf{v})$$

Θεώρημα 3.1. *Έστω $B, \rho > 0$ και f κυρτή συνάρτηση με $\mathbf{w}^* \in \arg\min_{\|\mathbf{w}\| \leq B} f(\mathbf{w})$. Αν ισχύει ότι $\|\mathbf{v}_i\| \leq \rho$ τότε:*

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{B\rho}{\sqrt{M}}$$

όπου $\bar{\mathbf{w}}$ η έξοδος του αλγορίθμου 1 εκτελεσμένου με παράμετρο $\eta = \frac{B}{\rho\sqrt{M}}$.

3.1.3 Στοχαστική Κατάβαση Κλίσης

Οι Robbins και Monro πρότειναν το 1951, [51], μια μικρή τροποποίηση της μεθόδου: αντί να βρίσκουμε το gradient της f και να υπολογίζουμε την τιμή της σε κάθε $\mathbf{w}^{(i)}$ θα επιλέγουμε το $\mathbf{v}_i$ τυχαία, με μοναδική προϋπόθεση ότι η αναμενόμενη τιμή του για κάθε επανάληψη θα δείχνει την κατεύθυνση του gradient. Με αυτήν την αλλαγή περνάμε πλεόν από την ντετερμινιστική διαδικασία της Κατάβασης Κλίσης σε μία στοχαστική, εξ ου και το όνομα Στοχαστική Κατάβαση Κλίσης(Stochastic Gradient Descent ή απλά SGD).

Αλγόριθμος 2 Στοχαστική Κατάβαση Κλίσης

```
1: procedure SGD(M,η)               ▷ M: αριθμός επαναλήψεων, η: παράμετρος
2:     initialize w^(0) ← 0
3:     for i = 1, ..., M
4:         v_i : E[v_i|w^(i−1)] = ∇f(w^(i−1))
5:         w^(i) ← w^(i−1) − ηv_i
6:     w̄ ← (1/M) Σ_{i=1}^{M} w^(i)
7:     return w̄
```

Ας δούμε τι κερδίζουμε εισάγοντας την τυχαιότητα, αλλά και τι χάνουμε. Το δεύτερο είναι μάλλον πιο απλό: χάνουμε την σιγουριά που μας παρέχει ο ντετερμινισμός. Θα δούμε σε λίγο το «τίμημα» που πρέπει να πληρώσουμε για αυτό. Ωστόσο, απολαμβάνουμε τόσο θεωρητικά οφέλη όσο και πρακτικά. Τα θεωρητικά, όπως θα διαπιστώσουμε και στα πλαίσια της διπλωματικής, πηγάζουν από το γεγονός ότι μπορούμε να δουλέψουμε και με συναρτήσεις για τις οποίες η παράγωγος(ή το gradient) δεν μπορεί να εκφραστεί με κλειστό μαθηματικό τύπο. Τα πρακτικά έχουν να κάνουν με την παρατήρηση ότι πολλές φορές είναι πιο εύκολο να υπολογιστεί η εκτίμηση της παραγώγου παρά η ίδια η παράγωγος, προσφέροντας σημαντικές βελτιώσεις στην απόδοση των εφαρμογών, ειδικά αν αυτές διαχειρίζονται μεγάλο όγκο δεδομένων. Για περισσότερα σχετικά με το ζήτημα, ο αναγνώστης μπορεί να ξεκινήσει από τα [5] και [6]. Ας συζητήσουμε τώρα το «τίμημα» στο οποίο αναφερθήκαμε προηγουμένως. Ξεκινάμε με το αντίστοιχο του θεωρήματος 3.1 στο στοχαστικό πλαίσιο:

Θεώρημα 3.2. *Έστω $B, \rho > 0$ και f κυρτή συνάρτηση με $\mathbf{w}^* \in \arg\min_{\|\mathbf{w}\| \leq B} f(\mathbf{w})$. Αν ισχύει με πιθανότητα 1 ότι $\|\mathbf{v}_i\| \leq \rho$ τότε:*

$$E[f(\bar{\mathbf{w}}) - f(\mathbf{w}^*)] \leq \frac{B\rho}{\sqrt{M}}$$

όπου $\bar{\mathbf{w}}$ η έξοδος του αλγορίθμου 2 εκτελεσμένου με παράμετρο $\eta = \frac{B}{\rho\sqrt{M}}$.

Υπάρχουν δύο μικρές αλλαγές. Η πρώτη αφορά το φράγμα του $\|\mathbf{v}_i\|$. Δεδομένου ότι πλέον είναι τυχαία μεταβλητή, προσαρμοζόμαστε και απαιτούμε το φράγμα να ισχύει στην πιθανότητα. Η δεύτερη αλλαγή είναι πιο ουσιαστική. Αφού το $\mathbf{v}_i$ είναι η αναμενόμενη τιμή της κλίσης, εργαζόμαστε με αναμενόμενες τιμές και για τα υπόλοιπα μεγέθη με αποτέλεσμα και το φράγμα να αφορά αναμενόμενες τιμές. Ωστόσο, εμείς θα θέλαμε το φράγμα να ισχύει στην πιθανότητα οπότε απαιτούνται κάποια ακόμη βήματα από μεριάς μας. Κατασκεύαζουμε ένα φράγμα αντισυγκέντρωσης της μορφής:

$$P\left(f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \geq c_1\right) \leq c_2$$

όπου το c είναι κάποια σταθερά ενώ το c_1 εξαρτάται από τις παραμέτρους. Επιλέγουμε τις παραμέτρους ώστε το c_1 να λαμβάνει κάποια πολύ μικρή τιμή και εκτέλουμε K ανεξάρτητες επαναλήψεις του αλγορίθμου ώστε να βελτιώσουμε τις πιθανότητές μας. Για παράδειγμα, για $c_2 = \frac{1}{2}$ και $K = 10$ έχουμε πιθανότητα επιτυχίας μεγάλυτερη από $99,9\%$. Από δω και πέρα, θα στοχεύουμε σε μία εξίσωση της μορφής:

$$P\left(f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \geq \varepsilon\right) \leq \delta$$

και θα εκφράζουμε τα K, M σαν συνάρτησει των δ, ε. Ας το δούμε μέσα από ένα παράδειγμα:

Παράδειγμα 3.2. *Ξεκινώντας από το αποτέλεσμα του θεωρήματος 3.2 μας αρκεί*

$$\frac{B\rho}{\sqrt{M}} \leq \varepsilon \implies M \geq \frac{B^2\rho^2}{\varepsilon^2}$$

Έπειτα, εφαρμόζουμε την ανισότητα Markov, [43],*:*

$$P(X \geq c) \leq \frac{E[X]}{c}$$

για $X = f(\bar{\mathbf{w}}) - f(\mathbf{w}^*)$ *και* $c = 2E[X]$ *και λαμβάνουμε*

$$P\left(f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \geq 2\varepsilon\right) \leq \frac{1}{2}$$

Και μετά από K *επαναλήψεις*

$$P\left(f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \geq 2\varepsilon\right) \leq \frac{1}{2^K}$$

Οπότε αρκεί

$$\delta \geq \frac{1}{2^K} \implies K \geq \log\left(\frac{1}{\delta}\right)$$

Θα συνεχίσουμε την μελέτη μας με κάποιες παραλλαγές ή προσαρμογές του αλγορίθμου 2. Η πρώτη επικεντρώνεται γύρω από την συνθήκη $\mathbf{w}^* \in \arg\min_{\|\mathbf{w}\| \leq B} f(\mathbf{w})$, ή πιο συγκεκριμένα την απλούστευσή της ως $\mathbf{w}^* \in \mathcal{H} = \{\mathbf{w} : \|\mathbf{w}\| \leq B\}$. Αφού το $\bar{\mathbf{w}}$ αποτελεί την εκτίμηση μας για το $\mathbf{w}^*$ είναι μάλλον επιθυμητό να ανήκει και αυτό στο ίδιο σύνολο, αλλά δεν διαθέτουμε κάποια εγγύηση για αυτό. Ένας εύκολος τρόπος για να το καταφέρουμε είναι να απαιτήσουμε $\mathbf{w}^{(i)} \in \mathcal{H}\ \forall i$. Με αυτόν τον τρόπο, και από την κυρτότητα του $\mathcal{H}$, συμπεραίνουμε ότι και η τελική μας εκτίμηση ανήκει στο σύνολο.

Αλγόριθμος 3 Στοχαστική Κατάβαση Κλίσης με προβολή

procedure SGD(M,η) ▷ M: αριθμός επαναλήψεων, η: παράμετρος
  initialize $\mathbf{w}^{(0)} \leftarrow 0$
  for $i = 1, \ldots, M$
    $\mathbf{v}_i : \mathrm{E}[\mathbf{v}_i | \mathbf{w}^{(i-1)}] = \nabla f(\mathbf{w}^{(i-1)})$
    $\mathbf{r}_i \leftarrow \mathbf{w}^{(i-1)} - \eta \mathbf{v}_i$
    $\mathbf{w}^{(i)} \leftarrow \arg\min_{\mathbf{w} \in \mathcal{H}} \|\mathbf{w} - \mathbf{r}_i\|$ ▷ Προβολή στο $\mathcal{H}$
  $\bar{\mathbf{w}} \leftarrow \frac{1}{M} \sum_{i=1}^{M} \mathbf{w}^{(i)}$
  return $\bar{\mathbf{w}}$

Βλέπουμε, λοιπόν, μία παραλλαγή του αλγορίθμου με την προσθήκη ενός βήματος προβολής. Η προβολή μας εξασφάλιζει το απαιτούμενο φράγμα για την νόρμα ενώ ταυτόχρονα διατηρεί και τα αποτελέσματα του θεωρήματος 3.2.

Χρήσιμη για την συνέχεια θα μας φανεί και η περίπτωση που η συνάρτηση f δεν είναι απλώς κυρτή.

Ορισμός 3.1.3. *Μία συνάρτηση* f *καλείται* λ *- ισχυρά κυρτή αν για κάθε* $\mathbf{u}, \mathbf{v}$ *και* $\alpha \in [0, 1]$ *ισχύει:*

$$f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \leq \alpha f(\mathbf{u}) + (1-\alpha) f(\mathbf{v}) - \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{u} - \mathbf{v}\|^2$$

Προφανώς, μια ισχυρά κυρτή συνάρτηση είναι και κυρτή($\lambda = 0$) άρα εξακολουθεί να ισχύει η παραπάνω ανάλυση. Βέβαια, αφού έχουμε στην διάθεσή μας έναν πιο ισχυρό περιορισμό είναι αναμενόμενο να μπορούμε να βελτιώσουμε τα αποτελέσματά μας. Πράγματι, οι Hazan et al., [31], παρουσίασαν μια τροποποίηση του αλγορίθμου για ισχυρά κυρτές συναρτήσεις και απέδειξαν ισχυρότερα φράγματα.

Αλγόριθμος 4 Στοχαστική Κατάβαση Κλίσης για μία λ-ισχυρά κυρτή f

1: **procedure** SGD(M) ▷ M: αριθμός επαναλήψεων
2: **initialize** $\mathbf{w}^{(0)} \leftarrow 0$
3: **for** $i = 1, \ldots, M$
4: $\mathbf{v}_i : \mathrm{E}[\mathbf{v}_i|\mathbf{w}^{(i-1)}] = \nabla f(\mathbf{w}^{(i-1)})$
5: $\eta_i = 1/(\lambda i)$
6: $\mathbf{r}_i \leftarrow \mathbf{w}^{(i-1)} - \eta_i \mathbf{v}_i$
7: $\mathbf{w}^{(i)} \leftarrow \arg\min_{\mathbf{w}\in\mathcal{H}} \|\mathbf{w} - \mathbf{r}_i\|$ ▷ Προβολή στο $\mathcal{H}$
8: $\bar{\mathbf{w}} \leftarrow \frac{1}{M}\sum_{i=1}^{M} \mathbf{w}^{(i)}$
9: **return** $\bar{\mathbf{w}}$

Πριν διατυπώσουμε το θεώρημά τους, μας δίνεται εδώ η ευκαιρία για ένα ακόμη σχόλιο. Παρατηρήστε ότι στο βήμα 5 ορίζουμε διαφορετικά η για κάθε επανάληψη. Είναι η πρώτη φορά που βλέπουμε κάτι αντίστοιχο στα πλαίσια της εργασίας μας, ωστόσο στην πράξη είναι αρκετά συνηθισμένο. Η αλλαγή αυτή ποσοτικοποιεί την διαίσθηση πως όσο πλησιάζουμε το ελάχιστο τόσο πιο προσεκτικά, δηλαδή με μικρά βήματα, πρέπει να κινούμαστε. Αντίθετα, στην αρχή της αναζήτησής μας έχουμε την ευχέρεια μεγαλύτερων κινήσεων.

Περνάμε τώρα στο αντίστοιχο του θεωρήματος 3.2 για ισχυρά κυρτές συναρτήσεις.

Θεώρημα 3.3. *Έστω f λ-ισχυρά κυρτή συνάρτηση και $E[\|\mathbf{v}_i\|^2] \leq \rho^2$. Αν $\mathbf{w}^* \in \arg\min_{\mathbf{w}\in\mathcal{H}} f(\mathbf{w})$ ελάχιστο αυτής τότε:*

$$E[f(\bar{\mathbf{w}}) - f(\mathbf{w}^*)] \leq \frac{\rho^2}{2\lambda M}(1 + \log(M))$$

όπου $\bar{\mathbf{w}}$ η έξοδος του αλγορίθμου 4.

Ένα σχόλιο για τον ρυθμό σύκλισης των δύο αλγορίθμων: ο αλγόριθμος 3 είναι της τάξης του $\mathcal{O}\left(\frac{1}{\sqrt{M}}\right)$ ενώ ο 4 είναι $\mathcal{O}\left(\frac{\log M}{M}\right)$. Ξαναγράφοντας την πρώτη ως $\mathcal{O}\left(\frac{\sqrt{M}}{M}\right)$ παρατηρούμε αμέσως ότι μοιράζονται έναν όρο $\frac{1}{M}$ που προκύπτει από την διαίρεση κατά τον υπολογισμό του μέσου όρου αλλά διαφέρουν ως προς τον όρο που προκύπτει από την άθροιση των $\mathbf{w}^{(i)}$: $\sqrt{M}$ εναντίον $\log M$. Οι Rahklin et al στο [50] βελτίωσαν το φράγμα στο βέλτιστο $\mathcal{O}\left(\frac{1}{M}\right)$. Ένα ακόμη ενδιαφέρον αποτέλεσμα έδειξαν και οι Shamir & Zhang στο [58] όπου για την περίπτωση ισχυρά κυρτών συναρτήσεων η τελευταία επανάληψη, $\mathbf{w}^{(M)}$ επιτυγχάνει από μόνη της φράγμα της τάξης του $\mathcal{O}\left(\frac{\log M}{M}\right)$, δηλαδή γλυτώνουμε την ανάγκη υπολογισμού του μέσου όρου.

Θα κλείσουμε το κεφάλαιο αιτιολογώντας την επιλογή να εστιάσουμε στις κυρτές συναρτήσεις. Από την μία πλευρά, θα χρειαστούμε αργότερα τα παραπάνω αποτελέσματα αφού το πρόβλημα

των αποκομμένων πολύδιαστατων κανονικών κατανομών ανάγεται πολύ εύκολα σε πρόβλημα κυρτής ελαχιστοποίησης, και το ίδιο ισχύει και για πληθώρα προβλημάτων. Από την άλλη, οι ιδιότητες των κυρτών συναρτήσεων τις καθιστούν ιδανικές στην μελέτη της Βελτιστοποίησης. Το γεγονός αυτό έχει οδηγήσει σε έναν χωρισμό του κλάδου σε *Κυρτή Βελτιστοποίηση* και *Μη κυρτή Βελτιστοποίηση*, με τον πρώτο να απολαμβάνει περισσότερης προσοχής και, συνεπώς, αποτελεσμάτων. Για περισσότερα, παραπέμπουμε τον αναγνώστη στο [7].

Κεφάλαιο 4

Ιστορική αναδρομή στην μάθηση αποκομμένων κατανομών

Στην παρούσα ενότητα θα πραγματοποιήσουμε μία σύντομη αναδρομή στην ιστορία της εκμάθησης αποκομμένων κατανομών. Φυσικά, θα επικεντρωθούμε σε θέματα που αφορούν τις κανονικές κατανομές, αλλά θα δούμε και γνωστές εργασίες, κυρίως από το χώρο της Στατιστικής με λογοκριθέντα δείγματα, που δεν μπορούν να απουσιάσουν. Μέσα από αυτήν την αναδρομή, ευελπιστούμε ο αναγνώστης να κατανοήσει καλύτερα τόσο τις προόδους που σημειώθηκαν στον κλάδο όσο και τα εμπόδια που επέμειναν. Ώστε, όταν θα φτάσουμε στο [11] να μπορούμε να μελετήσουμε τις καινοτόμες τεχνικές που παρουσίασε προς υπερπήδηση αυτών και να καταλήξουμε στο σήμερα για να δούμε το νέο δρόμο έρευνας που άνοιξε στην περιοχή.

4.1 18ος αιώνας

Θα ξεκινήσουμε το ταξίδι μας από τα μέσα του 18ου αιώνα. Σε μία Ευρώπη χτυπημένη από την ευλογιά, υπεύθυνη για έναν στους δέκα θανάτους στα μεγάλα αστικά κέντρα, επικρατεί διχασμός για την χρήση ή μη των εμβολίων. Ο Ελβετός μαθηματικός και φυσικός Daniel Bernoulli παρουσιάζει το 1760,[2], μια εργασία του αναφορικά με τον εμβολιασμό. Μοντελοποιεί το πρόβλημα με τις εξής παραμέτρους:

- n : σύνολο ατόμων εκ των οποίων ένα μολύνεται με ευλογιά
- m : σύνολο ασθενών εκ των οποίων ένα πεθαίνει από ευλογιά
- x : ηλικία
- $\xi(x)$: πλήθος ατόμων ηλικιάς x που επέζησαν του ιού
- $s(x)$: πλήθος ατόμων ηλικιάς x που δεν μολύνθηκαν

Έπειτα, επιλύει το μοντέλο σε δύο διαφορετικές περιπτώσεις, με ή χωρίς εμβόλιο κατά τη γέννηση. Για να παράξει αριθμητικά αποτελέσματα, έπρεπε να χρησιμοποιήσει κάποιες εκτιμήσεις

των σταθερών του. Αν και υπήρχε καταγραφή των γεννήσεων και των θανάτων, ήταν πολλές φορές ελλειπής, καθιστώντας μέρος του δείγματος μη αξιοποιήσιμο. Για τον λόγο αυτό, αναγκάστηκε να αφαιρέσει δεδομένα εώς ότου εξισορροπήσει τους αριθμούς γεννήσεων-θανάτων. Πράττοντας με αυτόν τον τρόπο, έγινε ο πρώτος που χρησιμοποίησε αποκομμένα δείγματα. Τελικά, κατέληξε πως εμβολιάζοντας τον πληθυσμό κατά τη γέννηση αυξάνεται το προσδόκιμο ζωής περίπου κατά τρία χρόνια, από τα 26 στα 29 έτη! Συνεπώς, ακόμη και ένα εμβόλιο με θανατηφόρες παρενέργειες στο 10% της εφαρμογής του, θα οδηγούσε σε αύξηση του προσδόκιμου. Για περισσότερες πληροφορίες σχετικά με την μελέτη του Bernoulli αλλά και την διαμάχη που αυτή προκάλεσε ο αναγνωστής μπορεί να διαβάσει το [10].

4.2 19ος αιώνας

Για περισσότερο από έναν αιώνα δεν υπήρξε κάποια εξέλιξη, τουλάχιστον όχι κάποια αξιωσημείωτη. Ίσως αναμενόμενο, αν αναλογιστούμε ότι διανύουμε τις πρώτες περιόδους του κλάδου και αν αντιμετωπίσουμε την δουλειά του Bernoulli ως ένα outlier ωθούμενο από την επιδημία. Η επόμενη προσπάθεια που αξίζει να σημειωθεί είναι αυτή του Francis Galton στο [25]. Ο Galton προσπάθησε να εκτιμήσει την μέση τιμή και την διακύμανση της ταχύτητας των αλόγων κούρσας, υποθέτοντας ότι αύτη ακολουθεί γκαουσσιανή κατανομή. Τα δεδομένα του ήταν αποκομμένα καθώς αναλυτική καταγραφή των χρόνων γινόταν μόνο για όσα ολοκλήρωναν τον προκριματικό γύρο του ενός μιλίου σε λιγότερο από 2 λεπτά και 30 δευτερόλεπτα. Η μέθοδος που ακολούθησε ήταν αρκετά απλή: υπέθεσε ότι η συχνότερη τιμή των δειγμάτων αντιστοιχεί στην μέση τιμή και έπειτα απεικόνισε τις μετρήσεις του σε άξονες χωρισμένους σε τεταρτημόρια και υπολόγισε τις προσεγγίσεις του με επισκοπήση των σχημάτων.

4.3 20ος αιώνας

Για ευκολότερη μελέτη, θα χωρίσουμε τον αιώνα σε 4 μέρη.

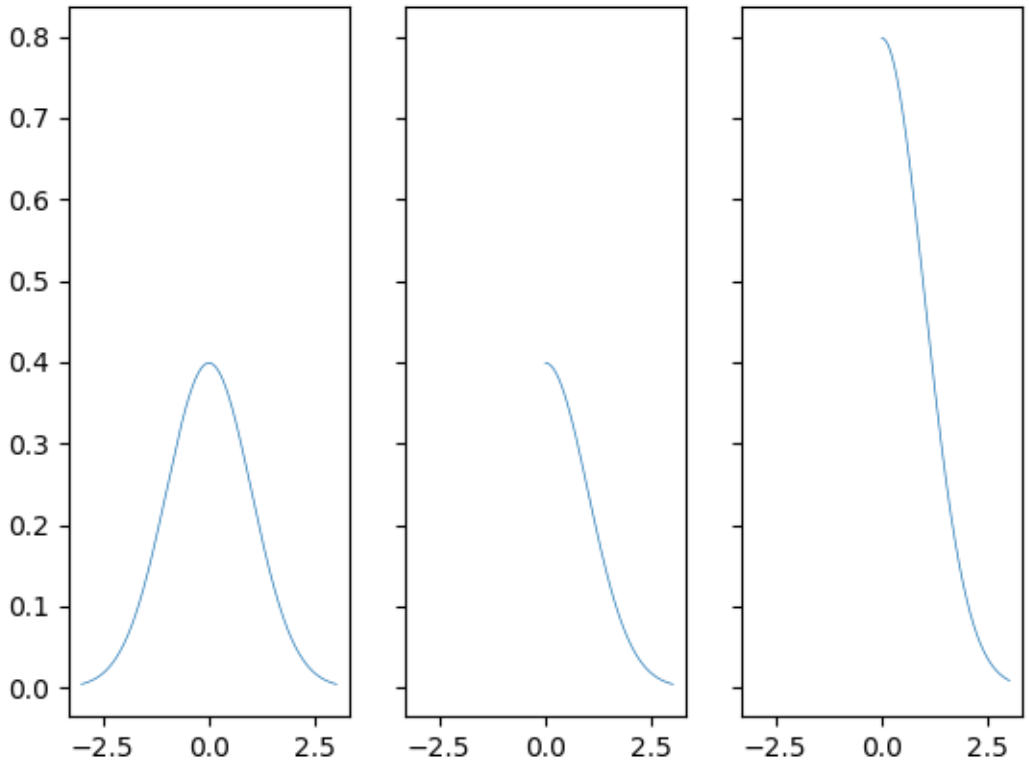
4.3.1 1ο μέρος:Pearson και Lee

Στο πρώτο μέρος, θα δούμε την δουλειά των Karl Pearson και Alice Lee στα [42], [46] και [47]. Ο πρώτος, αφού δεν έμεινε απόλυτα ικανοποιημένος με την δουλειά του Galton, αποφάσισε να εφαρμόσει την *Μεθόδου των Ροπών*.

Ορισμός 4.3.1. *Για μία τυχαία μεταβλητή X ορίζουμε την ροπή τάξης k γύρω από σταθερό σημείο c ως:*

$$\mu_c^{(k)} = \mathrm{E}[(X-c)^k]$$

Η μέθοδος είναι αρκετά απλή: αν χρειάζεται να εκτιμήσουμε μια κατανομή με n παραμέτρους, χρησιμοποιούμε τις n πρώτες ροπές εξισώνοντάς τες με τις εκτιμήσεις που λαμβάνουμε για αυτές από τα δείγματα και επιλύουμε το $n \times n$ σύστημα που προκύπτει. Στην περίπτωση των γκαουσσιανών κατανομών μας αρκούν οι δύο πρώτες ροπές. Ενδιαφέρον παρουσιάζει το



Σχήμα 4.1: $\mathcal{N}(0,1)$ αποκομμένη στα θετικά του άξονα x

γεγονός ότι παρά την πιο αυστηρή μέθοδο που ακολούθησε ο Pearson, τα αποτέλεσματά του δεν διέφεραν και πολύ από αυτά του Galton.

4.3.2 2ο μέρος:Fisher

Στις μέρες μας, η μέθοδος των ροπών δεν χρησιμοποιείται ιδιαίτερα. Έχει αποκατασταθεί από την εκτίμηση μέγιστης πιθανοφάνειας, που όπως είδαμε εισήγαγε ο Fisher. Όπως αναμενόταν, δεν άργησε να χρησιμοποιήσει την τεχνική του και στην περίπτωση των κανονικών κατανομών, π.χ. στο [22]. Γενικά, ο Fisher παρήγαγε μεγάλο όγκο δημοσιεύσεων σε σχετικά προβλήματα και δεν θα μας αρκούσε μία υποενότητα για να τις αναλύσουμε. Την δουλειά του Fisher με απεκομμένα δείγματα γενίκευσε ο Halperin στο [29].

4.3.3 3ο μέρος:Hotelling και Tukey

Εδώ θα δούμε ένα αποτέλεσμα που ο αναγνώστης ίσως ανέμενε. Βλέπουμε στο σχήμα 4.1 πώς ξεκινώντας από την $\mathcal{N}(0,1)$ καταλήγουμε στην αποκομμένη εκδοχή της για $x > 0$ πρώτα αφαιρώντας τα αρνητικά και έπειτα κάνοντας μία κανονικοποίηση(εδώ έχουμε επιλέξει βολικά το σύνολο ώστε ο συντελεστής να είναι απλά ένα $\times 2$.) Το ερώτημα είναι: μπορούμε να

ξεκινήσουμε από μία άλλη κανονική κατανομή και με μία διαφορετική αποκοπή να φτάσουμε στο ίδιο τελικά σχήμα; Ο Hotelling το 1948,[33], και ανεξάρτητα ο Tukey,[64], ένα χρόνο αργότερα απάνησαν αρνητικά. Δηλαδή, υπάρχει μία $1-1$ αντιστοιχία μεταξύ του τοπικού σχήματος της αποκομμένης καμπύλης και της αρχικής. Δυστυχώς, για να παράξουμε το σχήμα, δεδομένου ότι μιλάμε για συνεχή καμπύλη, χρειαζόμαστε άπειρο πληθος σημείων οπότε το αποτέλεσμα, αν και έχει θεωρητική αξία, δεν μπορεί ποτέ να εφαρμοστεί στην πράξη.

4.3.4 4ο μέρος:Tobin και Rubin

Στο τελευταίο μέρος, θα δούμε δύο δουλειές που σχετίζονται περισσότερο με απεκομμένα ή και λογοκριθέντα στατιστικά και λιγότερο με τις κανονικές κατανομές αλλά αποτελούν κομβικά σημεία στην εξέλιξη του κλάδου.
Το πρώτο είναι το *Μοντέλο Tobit*, που πήρε το όνομά του από τον James Tobin[1],[62]. Για να το δούμε πρώτα θα χρειαστούμε έναν ακόμη ορισμό:

Ορισμός 4.3.2. *Για μία τυχαία μεταβλητή X με συνάρτηση πυκνότητας πιθανότητας f_X ορίζουμε την αθροιστική συνάρτηση κατανομής F_X ως:*

$$F_X(x) = P[X \leq x] = \int_{-\infty}^{x} f_X(x)dx$$

Όταν $X \sim \mathcal{N}(0,1)$, σε αντιστοιχία με την περίπτωση της συνάρτησης πυκνότητας πιθανότητας, είθισται να χρησιμοποιείται ο συμβολισμός Φ για την αθροιστική κατανομή. Η αρχική δουλειά του Tobin αφορούσε την περίπτωση που τα δείγματα αποκαλύπτονται αν υπερβαίνουν μία τιμή κατωφλίου y_L. Αν $I(y) = \begin{cases} 1 & y > y_L \\ 0 & y \leq y_L \end{cases}$ η δείκτρια συνάρτηση τότε το μοντέλο υπολογίζει την συνάρτηση πιθανοφάνειας ως:

$$\mathcal{L}(\beta,\sigma) = \prod_{j=1}^{N} \left(\frac{1}{\sigma}\phi\left(\frac{y_j - \beta X_j}{\sigma}\right)\right)^{I(y_j)} \left(1 - \Phi\left(\frac{\beta X_j - y_j}{\sigma}\right)\right)^{1-I(y_j)}$$

Βλέπουμε, λοιπόν, ότι η συνάρτηση πιθανοφάνειας είναι συνδυασμός μίας αθροιστικής συνάρτησης και μίας πυκνότητας πιθανότητας, με την πρώτη να αντιστοιχεί στα λογοκριθέντα ή απεκομμένα δείγματα και τον συντελεστή β να εκφράζει τον λόγο της πιθανότητας να βλέπουμε ένα δείγμα προς την αναμενόμενη τιμή του. Ένα σχεδόν πανομοιότυπο μοντέλο είχε προταθεί και από τον Hald στο [28] αλλά δεν καθιερώθηκε.
Θα ολοκληρώσουμε με μια μικρη αναφορά στο [15]. Οι Rubin et al. δούλεψαν με ελλειπή δείγματα που δεν τους επέτρεπαν κλειστό τύπο για την συνάρτηση πιθανοφάνειας. Για να υπερκέρασουν αυτό το εμπόδιο, εισήγαγαν μία επαναληπτική τεχνική που εν τέλει απέκτησε ευρεία χρήση: τον αλγόριθμο *Προσδοκίας - Μεγιστοποίησης* ή EM όπως έχει καθιερωθεί στην βιβλιογραφία από το αγγλικό *Expectation - Maximization*. Ο αλγόριθμος είναι αρκετά απλός, απολείται από το βήμα Expectation κι από το βήμα Maximization. Αν συμβολίσουμε με X

[1] Δεν πρόκειται για τυπογραφικό λάθος. Το μοντέλος «βαπτίσθηκε» έτσι από τον Arthur Goldberger αλλά ο λόγος αλλαγής της κατάληξης δεν είναι ξεκάθαρος

τα γνωστά δείγματα, Z τα άγνωστα και $\theta(t)$ την εκτίμηση των παραμέτρων κατά το βήμα t τότε έχουμε:

Expectation:

$$Q(\theta|\theta(t)) = E_{Z|X,\theta(t)}[\mathcal{L}(X, Z;\theta)]$$

Maximization:

$$\theta(t+1) = arg\max_{\theta} Q(\theta|\theta(t))$$

Και επαναλαμβάνουμε εώς ότου επιτύχουμε κάποια συνθήκη σύγκλισης.

Θα αφήσουμε πίσω μας τον 20ο αιώνα παραπέμποντας τον αναγνώστη στο [54] για περισσότερες πληροφορίες αναφορικά με την εκμάθηση αποκομμένων στατιστικών, με έμφαση πάντα στις κανονικές κατανομές.

4.4 21ος αιώνα

Χωρίζουμε αυτήν την ενότητα σε 3 μέρη. Στο δεύτερο, θα δούμε λίγο αναλυτικότερα το [11] αλλά πριν πάμε εκεί θα κάνουμε μια μικρή αναφορά στα σθεναρά στατιστικά ώστε να πάρει ο αναγνώστης μια γεύση για το τι μελετάται τα τελευταία χρόνια και για να μην του δημιουργήσουμε την ψευδαίσθηση ότι ο κλάδος ήταν ανενεργός ενώ απλά είχε πάρει μια ελαφρώς διαφορετική κατεύθυνση. Θα κλείσουμε την ενότητα βλέποντας μερικές από τις εφαρμογές των τεχνικών που ανέπτυξαν οι Daskalalis et al..

4.4.1 Σθεναρά στατιστικά

Θα ξεκινήσουμε υπενθυμίζοντας στον αναγνώστη την συζήτησή μας στο τέλος της ενότητας 2.1.3. Ξεχωρίσαμε τα λογοκριθέντα από τα απεκομμένα δείγματα με βάση το αν έχουμε κάποια γνώση σχετικά με την αφαίρεση των δειγμάτων ή όχι. Κλείσαμε την ενότητα με ένα παράδειγμα από το χώρο της αστρονομίας. Ενδεχομένως, όμως, να αφήσαμε τον αναγνώστη με μια απορία: τι γίνεται εάν από ένα σημείο κι έπειτα τα όργανα μέτρησης που διαθέτουμε αντί να μην μπορούν να ανιχνεύσουν το φως καθόλου το ανιχεύουν μεν αλλά ίσως με κάποιο μεγάλο σφάλμα, π.χ. αν ανιχνεύουν ταυτόχρονα φως από περισσότερα άστρα; Φυσικά, μπορούμε να ακολουθήσουμε τον εύκολο δρόμο και απλά να πετάξουμε αυτές τις μετρήσεις. Ή, μπορούμε να το αντιμετωπίσουμε ως μία νέα πρόκληση, στοχεύοντας στην εξαγωγή συμπερασμάτων που αντιστέκονται σε τέτοια «κακά» δείγματα. Αυτή ακριβώς είναι και η έννοια της σθεναρής στατιστικής. Προφανώς, δεν χρειάζεται να στραφούμε σε διαστημικά μεγέθη για να δούμε τις πρακτικές εφαρμογές της. Κάθε φορά που υπολογίζουμε την μέση τιμή και την διακύμανση του ύψους σε μία σχολική αίθουσα εισάγουμε ένα συστηματικό σφάλμα αναμειγνύοντας τις κατανομές αγοριών και κοριτσιών.

Ζούμε σε μία εποχή με πρόσβαση σε συνεχώς αυξανόμενο όγκο πληροφοριών. Φυσική συνέχεια είναι να βρισκόμαστε όλο και πιο συχνά αντιμέτωποι με δεδομένα που επιδρούν αρνητικά στην στατιστική μελέτη και μάθηση οπότε η αξία της σθεναρής στατιστικής όλο και μεγαλώνει. Προφανώς, δεν μπορούμε στα πλαίσια μιας υποενότητας να καλύψουμε όλο αυτόν τον όγκο μελέτης, ούτε είναι κι αυτός ο στόχος μας βέβαια. Πιστεύουμε ότι εξηγήσαμε ικανοποιητικά το πρίσμα υπό το οποίο μελετάται σήμερα η αποκομμένη στατιστική. Για περισσότερη μελέτη παραπέμπουμε στα [30] και [34].

4.4.2 Αποδοτική Στατιστική, σε υψηλές διαστάσεις, από αποκομμένα δείγματα

Έχοντας εξηγήσει το κενό στην αναδρομή μας, μπορούμε πλέον να δούμε πιο αναλυτικά το [11]. Ακολουθώντας τους συγγραφείς, θα μοιράσουμε την συνεισφορά σε δύο κομμάτια. Το δεύτερο, αλλά απλούστερο, μας επιβεβαιώνει ένα αρκετά διαισθητικό αποτελέσμα: αν το σύνολο αποκοπής μάς είναι παντελώς άγνωστο και δεν έχουμε πρόσβαση σε αυτό μέσω του μαντείου, αλλά ούτε και κάποια άλλη υπόθεση που να μας προσφέρει πληροφορίες για τη δομή του, είναι αδύνατο να μάθουμε τις πραγματικές παραμέτρους της κατανομής. Ας στρέψουμε τώρα την προσοχή μας στο κύριο αποτέλεσμα της εργασίας: τον αλγόριθμο στοχαστικής κατάβασης κλίσης με προβολή. Στο κεφάλαιο 3 είδαμε ότι το διπλό βήμα ανανέωσης δεν αποτελεί κάποια καινοτομία της δουλειάς των Daskalakis et al.. Ωστόσο, σε εκείνη την περίπτωση το βλέπαμε απλώς ως ένα μέσο επίτευξης γρηγορότερης σύγκλισης. Αντιθέτως, στο [11] αποτελεί θεμέλιο λίθο του αλγορίθμου. Η εισαγωγή της αποκοπής καθίστα δυσκολότερη τη δουλειά μας καθώς περιόριζει τα «λάθη» που μπορεί να διορθώσει ο αλγόριθμος. Εν προκειμένω, τα διορθώνει αποκλειστικά και μόνο όταν η τρέχουσα εκτίμηση αφήνει μη αμελητέο όγκο εντός του συνόλου αποκοπής. Ίσως ο αναγνώστης βοηθηθεί αν το σκεφτεί ως εξής: βρισκόμαστε στην διαδρομή μας προς τις πραγματικές παραμέτρους της κατανομής αλλά μπορούμε να αντλούμε πληροφορίες μόνο από το αρχικό περιορισμένο σύνολο S. Την στιγμή που θα αφήσουμε αμελητέα μάζα στο S δεν θα μπορούμε να πάρουμε οδηγίες για το επόμενο βήμα μας, οπότε μάλλον θα χαθούμε. Αυτός ο «μαγικός» χώρος που μας επιτρέπει να πηγαίνουμε προς τον στόχο αλλά και να βλέπουμε αρκετό από το S είναι ο D_r, για να μείνουμε πιστοί στην σημειογραφία του [11]. Οπότε, πρακτικά, χρειαζόμαστε 3 συστατικά:

1. Ύπαρξη ενός σημείου-αφετηρίας στο D_r
2. Δυνατότητα ενός βήματος εντός χωρού, και
3. Ύπαρξη του στόχου στο D_r

Με αυτά τα 3 στη διάθεσή μας, το βήμα της προβολής αποτελεί μια δικλείδα ασφαλείας που μας απαγορεύει να εξέλθουμε του χώρου. Χρησιμοποιώντας την παραμετροποίηση του χώρου: $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \to (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1})$ οι συγγραφείς αποδεικνύουν τα παραπάνω 3 σημεία καθώς και την ιδιότητα της κυρτότητας που θα δώσει τα αποτελέσματα σύγκλισης. Έχοντας εξηγήσει τον σκελετό του [11], ας προχωρήσουμε σε πιο λεπτομερή ανάλυση.

Μάλλον μαντεύει ο αναγνώστης ότι θα χρησιμοποιήσουμε έναν(ή περισσότερους) από τους αλγορίθμους της ενότητας 3.1.3. Όλοι οι αλγόριθμοι που είδαμε εργάζονται με ένα διάνυσμα $\mathbf{w}$ ενώ εμείς αναζητούμε ένα ζεύγος παραμέτρων $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, όπου η μία παράμετρος είναι διάνυσμα του $\mathbb{R}^d$ και η άλλη πίνακας του $\mathbb{R}^{d\times d}$. Ξεπερνάμε αυτή τη δυσκολία σε δύο βήματα. Πρώτα ορίζουμε την *διανυσματοποίηση*(vectorization) του πίνακα συνδιακύμανσης, δηλαδή την παράθεση της μίας στήλης μετά την άλλη. Έτσι, παράγουμε ένα διάνυσμα στον $\mathbb{R}^{d^2}$. Το συμβολίζουμε ως $\boldsymbol{\Sigma}^b$. Προφανώς, στην περίπτωσή μας και λόγω συμμετρίας το αποτελέσμα είναι το ίδιο ακόμα και αν ακολουθήσουμε την διαδικασία κατά γραμμές αλλά ας δουλεύουμε με τον γενικό ορισμό. Έπειτα, συνενώνουμε τα δύο διανύσματα για να παράξουμε το διάνυσμα $\begin{bmatrix}\boldsymbol{\Sigma}^b & \boldsymbol{\mu}\end{bmatrix}^T \in \mathbb{R}^{d^2+d}$. Βέβαια, αναφέραμε ήδη ότι θα χρειαστούμε την παραμετροποίηση: $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \to (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1})$ οπότε αν $\mathbf{T} = \boldsymbol{\Sigma}^{-1}$ και $\mathbf{u} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ τότε τελικά $\mathbf{w} = \begin{bmatrix}\mathbf{T}^b \\ \mathbf{u}\end{bmatrix}$.

Ας δούμε τώρα που χρειαζόμαστε την αλλαγή των παραμέτρων. Εφαρμόζωντας τον ορισμό 2.1.8 σε μία πολυδιάσταστη κανονική κατανομή λαμβάνουμε:

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S;\mathbf{x}) = \begin{cases} \frac{1}{\int_S \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})} \cdot \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma};\mathbf{x}) & \mathbf{x} \in S \\ 0 & \mathbf{x} \notin S \end{cases}$$

Έπειτα, θα υπολογίσουμε τον αρνητικό λογάριθμο της συνάρτησης πιθανοφάνειας κατ΄ αντιστοιχία με το παράδειγμα 2.1, αλλά για ένα δείγμα:

$$\begin{aligned}\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma};\mathbf{x}) &= -\ln\left(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma};\mathbf{x})\right) + \ln\left(\int_S \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\right) \\ &= -\ln\left(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma};\mathbf{x})\right) + \ln\left(\int_S \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma};\mathbf{y})\, d\mathbf{y}\right)\end{aligned}$$

Αν ανατρέξουμε στον ορισμό 2.1.7 βλέπουμε ότι οι συντελεστές εκτός του εκθετικού είναι σταθεροί ως προς το $\mathbf{y}$ και άρα μπορούμε να τους βγάλουμε έξω από το ολοκλήρωμα του δεύτερου όρου και να τους απλοποιήσουμε με τους αντίστοιχους του πρώτου, λαμβάνοντας την απλοποιημένη μορφή

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma};\mathbf{x}) = -\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) + \ln\left(\int_S \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right) d\mathbf{y}\right)$$

Και εκτελώντας τις επιμεριστικές μπορούμε να κάνουμε το ίδιο κόλπο άλλη μία φορά

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma};\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \ln\left(\int_S \exp\left(-\frac{1}{2}\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y} + \mathbf{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) d\mathbf{y}\right)$$

Αυτό είναι το κατάλληλο σημείο για να αλλάξουμε την παραμετροποίηση του προβλήματος. Αμέσως λαμβάνουμε ότι

$$\ell(\mathbf{u}, \mathbf{T};\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{T}\mathbf{x} - \mathbf{x}^T\mathbf{u} + \ln\left(\int_S \exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{T}\mathbf{y} + \mathbf{y}^T\mathbf{u}\right) d\mathbf{y}\right)$$

Ας εξηγήσουμε αυτήν την αλλαγή. Εκ πρώτης όψεως, φαίνεται να έγινε κυρίως για πρακτικούς λόγους αφού απαλλαχθήκαμε από τον εκθέτη του πίνακα. Ωστόσο, έγινε και μία πιο ουσιώδης αλλαγή: πλέον δεν υπάρχει όρος που να περιέχει και τις δύο μεταβλητές. Αυτό θα παίξει σημαντικό ρόλο στην παραγώγιση που θα ακολουθήσει. Να ανοίξουμε εδώ μία παρένθεση για να σημειώσουμε πως ο στόχος απόδειξης κυρτότητας απλοποείται στην περίπτωση διαφορίσιμων συναρτήσεων στον έλεγχο του προσήμου της δεύτερης παραγώγου ή αντίστοιχα του εσσιανού πίνακα. Προτρέπουμε τον αναγνώστη να δει πόσο πιο απλός θα είναι ο υπολογισμός του προσήμου πλέον. Εκεί που οι όροι γινόμενα θα «επιβίωναν» δύο παραγωγίσεις, τώρα θα έχουμε μόνο κάποιους όρους εξαιτίας του λογαρίθμου. Δεν θα αναλωθούμε σε αναλυτικούς υπολογισμούς των παραγώγων πινάκων(βλ. [48] για περισσότερα) αλλά εύκολα υπολογίζει κανείς πως:

$$\nabla \ell(\mathbf{u}, \mathbf{T}; \mathbf{x}) = -\begin{bmatrix} \left(-\frac{1}{2}\mathbf{x}\mathbf{x}^T\right)^b \\ \mathbf{x} \end{bmatrix} + \mathop{\mathrm{E}}_{\mathbf{y}\sim\mathcal{N}(\mathbf{T}^{-1}\mathbf{u},\mathbf{T}^{-1},S)} \left[\begin{bmatrix} \left(-\frac{1}{2}\mathbf{y}\mathbf{y}^T\right)^b \\ \mathbf{y} \end{bmatrix}\right] \tag{4.1}$$

και

$$\mathbf{H}_{\ell(\mathbf{u},\mathbf{T};\mathbf{x})} = \mathop{\mathrm{Cov}}_{\mathbf{y}\sim\mathcal{N}(\mathbf{T}^{-1}\mathbf{u},\mathbf{T}^{-1},S)} \left[\begin{bmatrix} \left(-\frac{1}{2}\mathbf{y}\mathbf{y}^T\right)^b \\ \mathbf{y} \end{bmatrix}, \begin{bmatrix} \left(-\frac{1}{2}\mathbf{y}\mathbf{y}^T\right)^b \\ \mathbf{y} \end{bmatrix}\right]$$

Αμέσως συμπεραίνουμε την πρώτη απαραίτητη ιδιότητα, αυτή της κυρτότητας. Έπειτα, θέλουμε να μεταβούμε από την εξίσωση 4.1 που αφορά ένα μόλις δειγμα στην αντίστοιχη που αφορά ολόκληρο τον πληθυσμό. Η αλλαγή είναι απλή: αντικαθιστούμε το ένα δείγμα με την αναμενομή τιμή του συνολικού πληθυσμού και έχουμε:

$$\nabla \ell(\mathbf{u}, \mathbf{T}) = -\mathop{\mathrm{E}}_{\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)} \left[\begin{bmatrix} \left(-\frac{1}{2}\mathbf{x}\mathbf{x}^T\right)^b \\ \mathbf{x} \end{bmatrix}\right] + \mathop{\mathrm{E}}_{\mathbf{y}\sim\mathcal{N}(\mathbf{T}^{-1}\mathbf{u},\mathbf{T}^{-1},S)} \left[\begin{bmatrix} \left(-\frac{1}{2}\mathbf{y}\mathbf{y}^T\right)^b \\ \mathbf{y} \end{bmatrix}\right] \tag{4.2}$$

Αν εξισώσουμε το $\nabla \ell(\mathbf{u}, \mathbf{T})$ με το 0, παρόμοια με το παράδειγμα 3.1, λαμβάνουμε αμέσως ότι οι παράμετροι $\mathbf{u}^*, \mathbf{T}^*$, που αντιστοιχούν στις παραμέτρους $\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*$ που αναζητούμε δίνονται από τις σχέσεις:

$$\mathbf{T}^* = \boldsymbol{\Sigma}^{*-1}$$
$$\mathbf{u}^* = \boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*$$

Ως τώρα έχουμε ότι η συνάρτησή μας είναι κυρτή και ότι οι παράμετροι μηδενίζουν το gradient της. Συνεπώς, από Ανάλυση ξέρουμε ότι αυτές οι παράμετροι την ελαχιστοποιούν. Πριν προχωρήσουμε παρακάτω, ας καλύψουμε μία εκκρεμότητα από το προηγούμενο κεφάλαιο. Είχαμε πει ότι η Στοχαστική Κατάβαση Κλίσης αξιοποιείται σε περιπτώσεις που δεν διαθέτουμε κλειστό τύπο για τις συναρτήσεις μας. Δείτε λίγο την εξίσωση 4.2. Μοιάζει με κλειστό τύπο αλλά στην πραγματικότητα δεν είναι καθώς δεν μπορούμε να υπολογίσουμε τις αναμενόμενες τιμές πάνω στο άγνωστο σύνολο αποκοπής. Θα αφήσουμε στον αναγνώστη λίγο χρόνο για να βρει το $\mathbf{v}_i$ των αλγορίθμων 2, 3 και 4, ως υπόδειξη ας δει και πάλι πώς μεταβήκαμε από την σχέση 4.1 στην 4.2, και θα επανέλθουμε σε αυτό αργότερα. Πριν ασχοληθούμε με το σύνολο προβολής ας δώσουμε για λόγους πληρότητας το θεώρημα ισχυρής κυρτότητας του [11]:

Θεώρημα 4.1. *Έστω $\mathbf{H}_\ell$ ο Εσσιανός πίνακας του αρνητικού λογαρίθμου της συνάρτησης πιθανοφάνειας $\ell(\mathbf{u}, \mathbf{T})$ υπό την παρουσία ενός αυθαίρετου συνόλου αποκοπής S τέτοιο ώστε $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S) \geq \beta$ για κάποιο $\beta \in (0, 1]$, όπου $\boldsymbol{\mu} = \mathbf{T}^{-1}\mathbf{u}$ και $\boldsymbol{\Sigma} = \mathbf{T}^{-1}$. Τότε ισχύει ότι*

$$\mathbf{H}_\ell(\mathbf{u}, \mathbf{T}) \succeq \frac{1}{2^{13}} \left(\frac{\beta}{C}\right)^4 \lambda_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot \mathbf{I}$$

όπου C καθολική σταθερά και λ_m πόσοτητα που εξαρτάται από τον τανυστή τέταρτης ροπής της $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

Προχωράμε στην αρχικοποίηση και το σύνολο προβολής D_r. Μέχρι τώρα δεν κάναμε ιδιαίτερη αναφορά στην αρχικοποίση. Θέταμε $\mathbf{w}^{(0)} = 0$ και δεν ασχολούμασταν περαιτέρω. Αν ακολουθήσουμε την ίδια τακτική και εδώ, θα έχουμε μηδενίσει έναν πίνακα συνδιακύμανσης, πράγμα άτοπο. Μία πρώτη σκέψη είναι ότι αφού πριν χρησιμοποιούσαμε το $\mathbf{0}$, δηλαδή το ουδέτερο στοιχείο της πρόσθεσης, και αφού οι συνδιακυμάνσεις λειτουργούν πολλαπλιαστικά να θέσουμε ως αρχική εκτίμηση το $\mathbf{I}$, δηλαδή το ουδέτερο στοιχείο του πολλαπλασιαμού. Κάτι τέτοιο δεν θα πετύχει. Ας το σκεφτούμε πρακτικά μέσα από ένα παράδειγμα: έστω ένα μονοδιάστατο πρόβλημα με πραγματική κατανομή $N(1000, 1)$ και συνολο αποκοπής $S = \{x : x \geq 1000\}$. Αν η αρχική μας εκτίμηση είναι η $N(0, 1)$ η πιθανότητα να παράξουμε ένα δείγμα που να πέφτει εντός του συνόλου αποκοπής είναι πρακτικά μηδέν. Άρα δεν μπορούμε να κάνουμε πρόοδο. Βλέπουμε, δηλαδή, ότι χρειαζόμαστε η αρχική εκτίμηση να βρίσκεται «κοντά» στον στόχο, για κάποια έννοια απόστασης. Εδώ εισέρχεται η έννοια που είπαμε στην περίληψη, της μη αμελητέας μάζας στο D_r. Οι αρχικές παράμετροι που θα μας το δώσουν αυτό προκύπτουν από την εκτιμήτρια μέγιστης πιθανοφάνειας, υπό τον περιορισμό του συνόλου. Προσαρμόζοντας το παράδειγμα 3.1 στις υψηλότερες διαστάσεις έχουμε:

$$\hat{\boldsymbol{\mu}}_S = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$
$$\hat{\boldsymbol{\Sigma}}_S = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_S)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_S)^T$$

Όταν η αρχική εκτίμησή μας ήταν μηδενική, είχαμε κατευθείαν ότι άνηκε στο σύνολο $\mathcal{H}$(το σύνολο προβολής με τις φραγμένες νόρμες). Αν ήταν κάποια σταθερά $\mathbf{w}_0$, θα αλλάζαμε ελαφρώς τον ορισμό σε:

$$\mathcal{H} = \{\mathbf{w} : \|\mathbf{w} - \mathbf{w}_0\| \leq B\}$$

Κάτι αντίστοιχο θα κάνουμε και τώρα. Πρώτα, θα μετονομάσουμε το $\mathcal{H}$ σε D και το B σε r. Έπειτα, παρατηρούμε ότι αφού στην περίπτωσή μας τα $\mathbf{w}, \mathbf{w}_0$ αποτελούν την συνένωση δύο μεταβλητών μπορούμε να το σπάσουμε σε δύο, μία για κάθε μεταβλητή και να πάρουμε $\|\mathbf{u} - \mathbf{u}'\|_2^2 + \|\mathbf{T} - \mathbf{T}'\|_{\mathrm{F}}^2 \leq r^2$, όπου ο τετραγωνισμός απλώς αφαίρεσε την ρίζα που «έμπλεκε» τις δύο μεταβλητές. Έτσι, μπορούμε τώρα να το χωρίσουμε σε δύο προβλήματα ελαχιστοποίησης:

$$\arg\min_{\mathbf{u}} \|\mathbf{u} - \mathbf{u}'\|_2^2$$
$$\text{s.t.} \|\mathbf{u} - \mathbf{u}_S\|_2^2 \leq r^2$$

και

$$\arg\min_{\mathbf{T}} \|\mathbf{T}-\mathbf{T}'\|_{\mathrm{F}}^2$$
$$\text{s.t.} \|\mathbf{T}-\mathbf{T}_S\|_{\mathrm{F}}^2 \leq r^2$$

όπου $\mathbf{T}_S = \mathbf{\Sigma}_S^{-1}$, $\mathbf{u}_S = \mathbf{\Sigma}_S^{-1}\boldsymbol{\mu}_S$ και $\|\cdot\|_{\mathrm{F}}$ η νόρμα Frobenius, [67]. Θα είχαμε τελειώσει αν το πρόβλημα της προβολής σε αυτόν τον χώρο μπορούσε να λυθεί αποδοτικά δεδομένου του ότι είμαστε ήδη αποδοτικοί ως προς τη δειγματική πολυπλοκότητα. Ας δούμε πως οι συγγραφείς ξεπέρασαν και αυτό το εμπόδιο. Είναι εύκολο να δει κανείς πως αν χρησιμοποιούσαμε την αφελή αρχική εκτιμή $\begin{bmatrix}\mathbf{I}^b\\ \mathbf{0}\end{bmatrix}$ θα είχαμε κλειστό τύπο για την προβολή. Οπότε, συνδυάζουμε τα καλά των δύο κόσμων μεταμορφώνοντας τον χώρο μας ώστε $\mathbf{\Sigma}_S \to \mathbf{I}$ και $\boldsymbol{\mu}_S \to \mathbf{0}$. Εφαρμόζουμε δηλαδή έναν αφφινικό μετασχηματισμό, π.χ. [66]. Δυστυχώς, ο μετασχηματισμός μας επιβάλλει μια ακόμη συνθήκη, οπότε καταλήγουμε στο ορισμό του D_r

$$D_r = \{(\mathbf{u},\mathbf{T}) : \|\mathbf{v}\|_2 \leq r, \|\mathbf{T}-\mathbf{I}\|_{\mathrm{F}} \leq r, \|\mathbf{T}^{-1}\|_{\mathrm{F}} \leq r\}$$

που συνεπάγεται το ακόλουθο σύστημα:

$$\arg\min_{\mathbf{u}} \|\mathbf{u}-\mathbf{u}'\|_2^2$$
$$\text{s.t.} \|\mathbf{u}\|_2 \leq r$$

και

$$\arg\min_{\mathbf{T}} \|\mathbf{T}-\mathbf{T}'\|_{\mathrm{F}}^2$$
$$\text{s.t.} \|\mathbf{T}-\mathbf{I}\|_{\mathrm{F}}^2 \leq r^2$$
$$\|\mathbf{T}^{-1}\|_2 \leq r$$

Το πρώτο μισό του προβλήματος είναι αρκετά εύκολο. Θα αφήσουμε στον αναγνώστη μία αναλυτική απόδειξη και θα παρουσιάσουμε μία γεωμετρική ερμηνεία. Θέλουμε να βρούμε το κοντινότερο σημείο στο $\mathbf{u}'$ που ανήκει στην μπάλα κέντρου $\mathbf{0}$ και ακτίνας r. Αν το $\mathbf{u}'$ ανήκει στην μπάλα τότε η απάντηση είναι ο εαυτός του. Διαφορετικά, είναι το σημείο που βρίσκεται στην επιφάνεια της μπάλας και στην ακτινική κατεύθυνση του $\mathbf{u}'$. Δηλαδή,

$$\mathbf{u} = \frac{\min(r, \|\mathbf{u}'\|_2)}{\|\mathbf{u}'\|_2}\mathbf{u}'$$

Για το δεύτερο σκέλος του προβλήματος επιτυγχάνουμε αποδοτικό αλγόριθμο χρησιμοποιώντας την μέθοδο των πολλαπλασιαστών Karush-Kuhn-Tucker,[37] και [41], σε συνδυασμό με μία δυαδική αναζήτηση πάνω στο χώρο των πολλαπλασιαστών. Οπότε αρκεί για κάθε τιμή του λ να λύνουμε αποδοτικά το πρόβλημα:

$$\arg\min_{\mathbf{T}} \|\mathbf{T}-\mathbf{T}'\|_{\mathrm{F}}^2 + \lambda\|\mathbf{T}-\mathbf{I}\|_{\mathrm{F}}^2$$
$$\text{s.t.} \mathbf{T} \succeq \frac{1}{r}$$

Αυτό επιτυγχάνεται με μερικούς αλγεβρικούς μετασχηματισμούς που μας οδηγούν σε κλειστό τύπο. Πλέον, έχουμε όλα τα υπολογιστικά συστατικά και χρειαζόμαστε ένα τελευταία στατιστικό φράγμα: αυτό της διακύμανσης του εκτιμητή του gradient, που καλούσαμε ρ στα θεωρήματα 3.1, 3.2 και 3.3. Ας το διατυπώσουμε:

Θεώρημα 4.2. *Έστω $\mathbf{v}_i$ το gradient όπως υπολογίστηκε στο βήμα i. Έστω $\mathbf{u}, \mathbf{T}$ οι εκτιμήσεις των παραμέτρων μετά το βήμα $i-1$ σύμφωνα με τις οποίες το gradient υπολογίστηκε με $\boldsymbol{\mu} = \mathbf{T}^{-1}\mathbf{u}$ και $\boldsymbol{\Sigma} = \mathbf{T}^{-1}$. Έστω $\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*$ οι παράμετροι που θέλουμε να ανακτήσουμε, με $\mathbf{u}^* = \boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*$ και $\mathbf{T}^* = \boldsymbol{\Sigma}^{*-1}$. Αν υποθέσουμε ότι $(\mathbf{u}, \mathbf{T}) \in D_r, (\mathbf{u}^*, \mathbf{T}^*) \in D_r$ και ότι $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S) \geq \beta$, $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*; S) \geq \beta$ τότε έχουμε ότι*

$$E\left[\|\mathbf{v}_i\|_2^2\right] \leq \frac{100}{\beta} d^2 r^2$$

Θα κλείσουμε την ενότητα αφήνοντας τον αναγνώστη να μελετήσει τις πρωτότυπες αποδείξεις και θα διατυπώσουμε τους αλγορίθμους στο κεφάλαιο 5 ακριβώς πριν τα πειράματα.

4.4.3 Η σύγχρονη γραμμή έρευνας

Ολοκληρώνουμε το κεφάλαιο με μερικές σύντομες αναφορές σε πρόσφατες ερευνητικές εργασίες που αξιοποιούν την τεχνική του pSGD.

- Στο [40], οι Kontonis et al. παρουσίασαν ένα άκρως εντυπωσιακό αποτέλεσμα, όπου πέτυχαν αποδοτική εκμάθηση αποκομμένων πολυδιάστατων κανονικών κατανομών χωρίς την χρήση μαντείου ή οποιασδήποτε άλλης γνώσης του συνόλου αποκοπής. Η μόνη τους υπόθεση ήταν ότι το σύνολο ανήκει σε μία συγκεκριμένη, αλλά ταυτόχρονα αρκετά μεγάλη, οικογένεια συνόλων με κάποιες απαραίτητες «καλές» ιδιότητες. Προφανώς, εξαιτίας αυτής της υπόθεσης, το αποτέλεσμά τους δεν αντιτίθεται στο αποτέλεσμα αδυναμίας εκμάθησης του [11].

- Σε μία πολύ τεχνική δουλειά, οι συγγραφείς του [11] έδειξαν ότι οι αλγόριθμοι που είδαμε προσαρμόζονται με επιτυχία στο πλαίσιο της αποκομμένης παλλινδρόμησης, [12].

- Στο χώρο των Νευρωνικών Δικτύων, οι Wu et al., [69], αντιμετώπισαν μια ανορθωμένη γκαουσσιανή, δηλαδή μια γκαουσσιανή που έχει περάσει μέσα από φίλτρο που μηδενίζει τα αρνητικά, ως αποκομμένη για να αντλήσουν από τις παρούσες μεθοδολογίες και να αποδείξουν εκμάθηση κατανομών που παράγονται από μονοστρωματικά νευρωνικά δίκτυα.

- Οι Fotakis et al., [24], εισήγαγαν μια έννοια «πάχους» του συνόλου αποκοπής στην περίπτωση κατανομών γινομένου και παρουσίασαν την πρώτη εκμάθηση αποκομμένων στατιστικών σε διακριτό πλαίσιο.

- Στο [4], οι Bhattacharyya et al. εφάρμοσαν pSGD σε αραιά γραφικά μοντέλα, δηλαδή σε περιπτώσεις που ο αντίστροφος του πίνακα συνδιακύμανσης μπορεί να απεικονισθεί ως γράφημα με μικρό αριθμό ακμών, που αντιστοιχούν σε μικρό αριθμό συνδιακυμάνσεων.

- Τέλος, οι Nagarajan και Panageas στο [45] συνδύασαν την τεχνική EM που είδαμε στην ιστορική αναδρομή με αυτήν της προσβολής για να εκτιμήσουν παραμέτρους στην περίπτωση της μίξης δύο γκαουσσιανών, όπου η πραγματική κατανομή ισούται με το ημιάθροισμα δύο κανονικών με αντίθετες μέσες τιμές και κοινό πίνακα συνδιακύμανσης.

Κεφάλαιο 5

Πειραματική Αξιολόγηση

Τον παρόν κεφάλαιο συνδέει την θεωρία που συνοπτικά μελετήθηκε στην ενότητα 4.4.2 με τους αλγορίθμους που ανέπτυξαν οι Daskalakis et al. και καταπιάνεται τόσο με τις παραμέτρους που υπεισέρχονται λόγω θεωρητικής αναγκαιότητας όσο και με τις μεταβλητές που πρέπει να ληφθούν υπόψιν σε μία πρακτική υλοποίηση. Παρουσιάζονται τα αποτέλεσμα των πειραμάτων συνοδευόμενα από σχολιασμό αναφορικά με το κατά πόσο επιβεβαιώνουν τη θεωρία και αν, τελικά, συνίστανται να χρησιμοποιηθούν σε πρακτικές εφαρμογές.

5.1 Από την θεωρία στην πράξη

Σκοπός μας είναι να προσαρμόσουμε κάποιον από τους αλγορίθμους 3 και 4 που είδαμε προηγουμένως στο πρόβλημά μας. Για να το πετύχουμε αυτό το πρώτο που χρειαζόμαστε είναι ένα τυχαίο διάνυσμα $\mathbf{v}_i$ με αναμενόμενη τιμή την κατεύθυνση του gradient. Είχαμε υποσχεθεί πως θα το πάρουμε από την εξίσωση 4.2 και είχαμε αφήσει τον αναγνώστη να δοκιμάσει το πώς. Ήρθε η στιγμή να το δούμε, αλλά πρώτα ας θυμήθουμε την εξίσωση:

$$\nabla \ell(\mathbf{u}, \mathbf{T}) = - \mathop{\mathrm{E}}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)} \left[\begin{bmatrix} \left(-\frac{1}{2}\mathbf{x}\mathbf{x}^T\right)^b \\ \mathbf{x} \end{bmatrix} \right] + \mathop{\mathrm{E}}_{\mathbf{y} \sim \mathcal{N}(\mathbf{T}^{-1}\mathbf{u}, \mathbf{T}^{-1}, S)} \left[\begin{bmatrix} \left(-\frac{1}{2}\mathbf{y}\mathbf{y}^T\right)^b \\ \mathbf{y} \end{bmatrix} \right]$$

Μάλλον φαίνεται περίπλοκο αλλά η απάντηση είναι πολύ απλή. Το ερώτημα είναι επί της ουσίας το ίδιο με το ποια συνάρτηση ισούται με την ποσότητα $\mathop{\mathrm{E}}_{z \sim D}[f(z)]$ αν έχουμε ένα $z \sim D$. Προφανώς η $f(z)$! Άρα, στην περίπτωσή μας:

$$\mathbf{v}_i = \begin{bmatrix} \mathbf{T}^b \\ \mathbf{u} \end{bmatrix} = - \left[\begin{bmatrix} \left(-\frac{1}{2}\mathbf{x}\mathbf{x}^T\right)^b \\ \mathbf{x} \end{bmatrix} \right] + \left[\begin{bmatrix} \left(-\frac{1}{2}\mathbf{y}\mathbf{y}^T\right)^b \\ \mathbf{y} \end{bmatrix} \right]$$

Το γεγονός ότι εδώ έχουμε δύο κατανομές δεν πρέπει να μας μπερδεύει. Τα διανύσματα $\mathbf{x}, \mathbf{y}$ ανήκουν το καθένα σε διαφορετική, συνεπώς είναι σταθερά ως προς την άλλη. Βλέπουμε, λοιπόν, στον αλγόριθμο 5 ότι ο υπολογισμός είναι αρκετά ευθύς αλλά και η χρήση του μαντείου, ή Πυθείας(oracle), απαραίτητη. Σε προβλήματα με ύπαρξη μαντείου συνηθίζεται αυτά να αποφαίνονται καταφατικά ή αρνητικά ('yes'/'no' oracle), όπως στην διατριβή του Alan Turing, [65]! Εδώ έχουμε ένα πιο ισχυρό μαντείο, υπό την έννοια ότι μας παρέχει μεν απαντήσεις σε ιδιότητας μέλους αλλά και δείγματα από την κατανομή.

Αλγόριθμος 5 Εκτίμηση του gradient

function GRADIENT ESTIMATION($\mathbf{w}$)
  $\mathbf{x} \leftarrow$ Δείγμα από το μαντείο
  $\begin{bmatrix} \mathbf{T}^b \\ \mathbf{u} \end{bmatrix} \leftarrow \mathbf{w}$
  $\boldsymbol{\mu} \leftarrow \mathbf{T}^{-1}\mathbf{u}$
  $\boldsymbol{\Sigma} \leftarrow \mathbf{T}^{-1}$
  repeat
    $\mathbf{y} \leftarrow$ Δείγμα από την $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  until $M_S(\mathbf{y}) = \mathbf{True}$ ▷ M_S: Ερώτημα προς το μαντείο
  return $-\begin{bmatrix} \left(-\frac{1}{2}\mathbf{x}\mathbf{x}^T\right)^b \\ \mathbf{x} \end{bmatrix} + \begin{bmatrix} \left(-\frac{1}{2}\mathbf{y}\mathbf{y}^T\right)^b \\ \mathbf{y} \end{bmatrix}$

Το τελευταίο δομικό συστατικό που χρειαζόμαστε είναι η αποδοτική προβολή. Ο αλγόριθμος 6 είναι αποδοτικός ως προς τη διάσταση καθώς ο εσωτερικός υπολογισμός του κλειστού τύπου(βήμα 5) απαιτεί μόνο πράξεις πινάκων και υπολογισμό μίας αποσύνθεσης ιδιάζουσων τιμών(Singular Value Decomposition - [18]), που εκτελούνται σε πολυωνυμικό ως προς τη διάσταση του πίνακα χρόνο.

Αλγόριθμος 6 Προβολή στο D_r

function PROJECT TO DOMAIN($\mathbf{r}$)
  $\begin{bmatrix} \mathbf{T}^b \\ \mathbf{u} \end{bmatrix} \leftarrow \mathbf{r}$
  $\mathbf{u}' \leftarrow \arg\min_{b:\|b\| \le r_1} \|\mathbf{b} - \mathbf{u}\|_2^2$
  repeat Δυαδική αναζήτηση επί του λ
    Επίλυση του υποπροβλήματος προβολής

$$\mathbf{T}' \leftarrow \arg\min_{\mathbf{T}' \succeq r_3\mathbf{I}} \|\mathbf{T} - \mathbf{T}'\|_{\mathrm{F}}^2 + \lambda\|\mathbf{T} - \mathbf{I}\|_{\mathrm{F}}^2$$

  until $\|\mathbf{T} - \mathbf{I}\|_2^2 \le r_3^2$ και η αντικειμενική τιμή του $\mathbf{T}'$ είναι ελάχιστη
  return $(\mathbf{u}', \mathbf{T}')$

Να εξηγήσουμε τι σημαίνουν τα πρωτοεμφανισθέντα r_1, r_2, r_3, γιατί ως τώρα χρησιμοποιούσαμε μόνο ένα r. Στην πράξη, δύσκολα οι τρεις εμφανίσεις του r στο ορισμό του D_r θα έχουν την ίδια τιμή. Για την εξαγωγή φραγμάτων στην στατιστική ανάλυση όμως ενδιαφερόμαστε για το μέγιστο από τα 3. Οπότε, σιωπηλά, είχαμε θέσει $r = \max(r_1, r_2, r_3)$. Τώρα που προχωράμε στην υλοποίηση, μπορούμε να αξιοποιήσουμε την ελευθερία να έχουμε διαφορετικές ακτίνες που ίσως οδηγήσουν σε ταχύτερη σύγκλιση.

5.1.1 Οι αλλαγές

Όπως, συχνά, συμβαίνει όταν η θεωρία περνά στην πράξη έτσι και δω θα κάνουμε κάποιες αλλαγές. Οι Daskalakis et al. εφάρμοσαν τους αλγορίθμους εκτίμησης του gradient αποδοτικής προβολής στο γενικό πλαίσιο του αλγορίθμου 4 που φέρει τις καλύτερες θεωρητικές εγγυήσεις. Ωστόσο, εμείς παρατηρήσαμε καλύτερα αποτελέσματα εφαρμόζοντας τους στο πλαίσιο του αλγορίθμου 3. Ενδεχομένως επειδή δεν χρησιμοποιήσαμε κάποιο προσεγγιστικά σωστό λ, αφού είναι δύσκολο να υπολογιστεί το πραγματικό, ακόμη και με γνωστές τις υπόλοιπες παραμέτρους. Δηλαδή, αλλάζουμε μόνο το ρυθμό μάθησης η. Η δεύτερη αλλαγή αφορά και πάλι τον ρυθμό μάθησης. Αντί να χρησιμοποιήσουμε τον τύπο $\eta = \frac{B}{\rho\sqrt{M}}$ θα απλοποιήσουμε σε $\eta = \frac{1}{\sqrt{M}}$. Εδώ, ο (προσεγγιστικός) υπολογισμός του ρυθμού μάθησης που μας παρέχει η θεωρία είναι αρκετά πιο εύκολος αλλά στην πράξη δεν αποδίδει. Πιθανώς, γιατί τα θεωρητικά φράγματα δεν είναι οριακά(tight). Θα δικαιολογήσουμε βέβαια την επιλογή μας και πειραματικά λίγο παρακάτω. Συγκεντρώνοντας τις αλλαγές μας και προσθέτοντας τον αφφινικό μετασχηματισμό των συγγραφέων καταλήγουμε στον τελικό αλγόριθμο που θα αξιολογήσουμε:

Πειραματικός Αλγόριθμος

procedure SGD(M) ▷ M: αριθμός επαναλήψεων
  $\begin{bmatrix} \hat{\boldsymbol{\Sigma}}^b \\ \hat{\boldsymbol{\mu}} \end{bmatrix} \leftarrow$ Μέσος όρος δειγμάτων από το μαντείο
  Εφαρμογή αφφινικού μετασχηματισμού: $\begin{bmatrix} \hat{\boldsymbol{\Sigma}}^b \\ \hat{\boldsymbol{\mu}} \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{I}^b \\ \mathbf{0} \end{bmatrix}$
  initialize $\mathbf{w}^{(0)} \leftarrow \begin{bmatrix} \mathbf{I}^b \\ \mathbf{0} \end{bmatrix}$
  $\eta_i \leftarrow \frac{1}{\sqrt{M}}$
  for $i = 1, \ldots, M$
    $\mathbf{v}_i \leftarrow$ Gradient Estimation$(\mathbf{w}^{(i-1)})$
    $\mathbf{r}_i \leftarrow \mathbf{w}^{(i-1)} - \eta \mathbf{v}_i$
    $\mathbf{w}^{(i)} \leftarrow$ Project to domain$(\mathbf{r}_i)$
  $\bar{\mathbf{w}} \leftarrow \frac{1}{M} \sum_{i=1}^{M} \mathbf{w}^{(i)}$
  Εφαρμογή αντίστροφου αφφινικού μετασχηματισμού στο $\bar{\mathbf{w}}$
  return $\bar{\mathbf{w}}$

Να σημειωθεί ότι κατά την εκτίμηση του gradient εννοείται ότι ο αφφινικός μετασχηματισμός εφαρμόζεται και στα δείγματα που μας παρέχει το μαντείο.

5.2 Σύστημα αξιολόγησης

Πριν περάσουμε στην παρουσίαση των αποτελεσμάτων, οφείλουμε να εξηγήσουμε τον τρόπο με τον οποίο θα αξιολογήσουμε τα πειράματα, καθώς και τις παραμέτρους αυτών.

5.2.1 Μετρική

Καλούμε τον αναγνώστη να ανατρέξει στην συζήτησή μας στην ενότητα 2.2.1. Μία οποιαδήποτε μετρική αξιολόγησης των πειραμάτων μας θα είναι εξ ορισμού μετρική απόστασης μεταξύ των δύο κατανομών, της τελικής μας εκτίμησης και των πραγματικών παραμέτρων. Όπως είχαμε τονίσει, θα επιθυμούσαμε να χρησιμοποιήσουμε την απόσταση ολικής μεταβολής. Ωστόσο, η έλλειψη κλειστού τύπου για αυτήν μας οδηγεί στην χρήση της σχέσης (1.1) του [11].

Λήμμα 5.1. *Έστω $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ οι πραγματικές παράμετροι της κατανομής και $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ εκτιμήσεις αυτών. Αν ισχύει ότι*

$$\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})\|_2 \leq \varepsilon \text{ και } \|\mathbf{I} - \boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}\|_{\mathrm{F}} \leq \varepsilon$$

τότε η απόσταση ολικής μεταβολής μεταξύ των $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ και $\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ είναι $\mathcal{O}(\varepsilon)$.

Παρατηρούμε ότι στην πρώτη ανισότητα υπεισέρχονται οι δύο μέσες τιμές ενώ στην δεύτερη καμία. Αντίθετα, ενώ και οι δύο συνδιακυμάνσεις υπάρχουν στην δεύτερη ανισότητα, βλέπουμε ότι η μία εμφανίζεται και στην πρώτη. Παρ΄όλ΄ αύτα, θα χρησιμοποιούμε την πρώτη ανισότητα ως απόσταση των μέσων τιμών και τη δεύτερη ως απόσταση των συνδιακυμάνσεων. Φυσικά, όταν έχουμε επιτύχει ως εκτίμηση την πραγματική τιμή, η αντίστοιχη διαφορά μηδενίζεται. Για να τις διακρίνουμε, θα συμβολίζουμε από δω κι έπειτα με ε_1 την απόσταση των μέσων τιμών και με ε_2 των πινάκων συνδιακύμανσης. Αντίστοιχα, στις εκτιμήσεις με τις οποίες αρχικοποιούμε θα αναφερόμαστε ως *αρχικά* ε_1 και ε_2.

5.2.2 Παράμετροι

Στον πίνακα 5.1 υπάρχουν συγκεντρωμένες όλες οι παράμετροι των πειραμάτων που θα ακολουθήσουν. Όπως φαίνεται και από τις διακεκομένες γραμμές, μπορούμε να τις ξεχωρίσουμε σε 3 κατηγορίες.

Παράμετρος	Σύμβολο	Τιμή
Κατανομή προς εύρεση	$\mathcal{N}$	$\mathcal{N}(\mathbf{0}, \mathbf{I})$
Αριθμός επαναλήψεων	M	10000
Διάσταση του προβλήματος	dim ή d	2
Σύνολο αποκοπής	S	$\{\mathbf{x} = (x_1, \ldots, x_d) : x_1 \geq 0\}$
Μέτρο του συνόλου αποκοπής	α	50%
Ακτίνα του συνόλου προβολής	r	r^*
Ρυθμός μάθησης	η	$\frac{1}{\sqrt{M}}$
Μέγεθος συστάδας δειγμάτων	batch	1

Πίνακας 5.1: Πίνακας παραμέτρων

Στην πρώτη κατηγορία έχουμε την πραγματική κατανομή και τον αριθμό των επαναλήψεων του αλγορίθμου. Για να διατηρήσουμε ένα μετρό σύγκρισης, θα κρατήσουμε αυτές τις μεταβλητές

σταθερές σε όλη την διάρκεια των πειραμάτων μας. Στην δεύτερη κατηγορία έχουμε την διάσταση του προβλήματος, το είδος και το μέτρο του συνόλου αποκοπής και την ακτίνα του συνόλου προβολής. Πρόκειται για τις παραμέτρους που μας εισάγει η θεωρητική ανάλυση. Στην τελευταία έχουμε το ρυθμό μάθησης και το μέγεθος συστάδας, για τα οποία είτε δεν γίνεται λόγος στην θεωρία είτε θεωρούνται γνωστά αλλά στην πράξη θα πρέπει να τα λάβουμε υπόψιν.

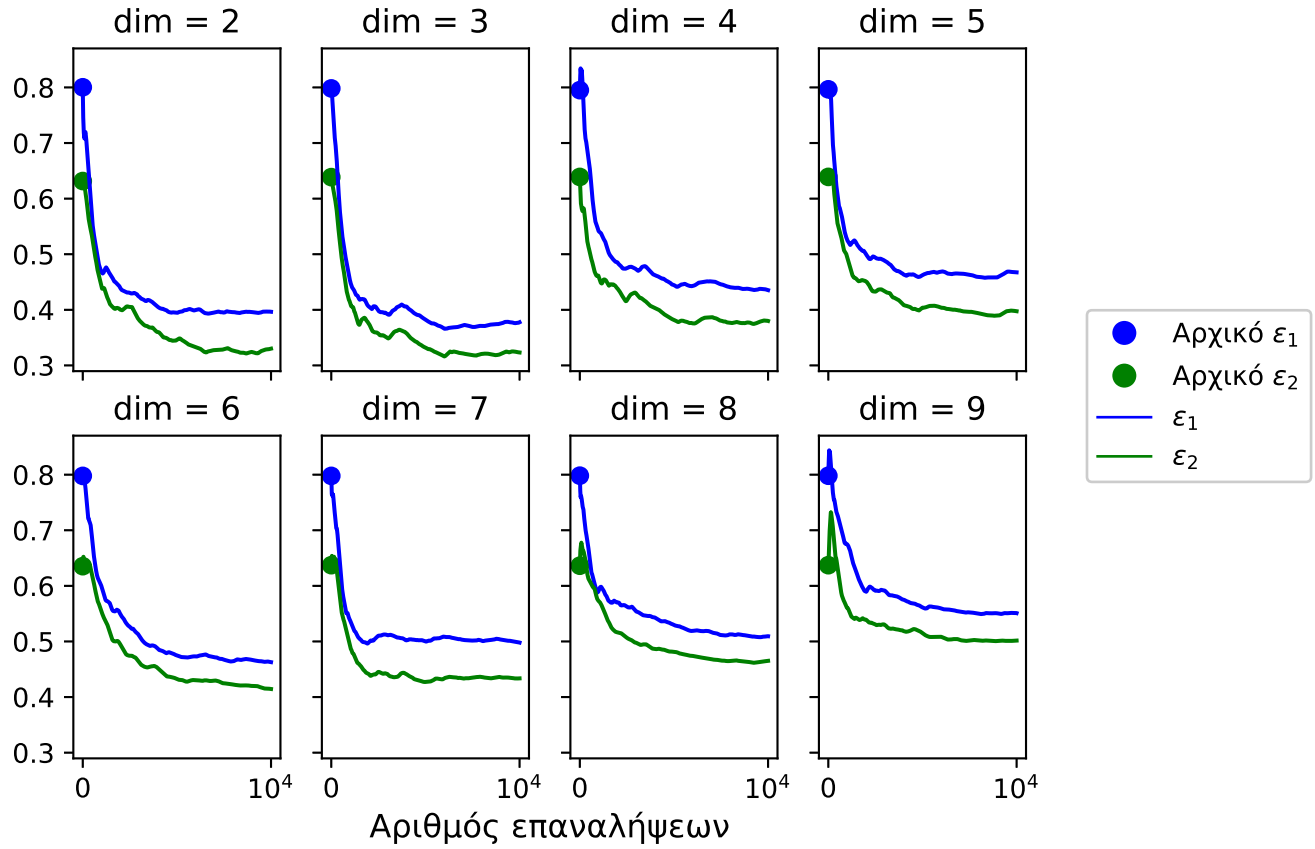
5.3 Μεθοδολογία πειραμάτων

Η μεθοδολογία που ακολουθήσαμε ελέγχει με την σείρα κάθε μία από τις παραμέτρους που μόλις είδαμε. Για να είναι ξεκάθαρη η επίδραση της κάθε παραμέτρου την απομονώνουμε, δηλαδή διατηρούμε σταθερές τις άλλες (και σε «ενδιάμεσες» τιμές για να μην επιδρούν σημαντικά στα αποτελέσματα) και την αντιμετωπίζουμε ως την μοναδική μεταβλητή του κάθε πειράματος. Κάθε μία από τις υποενότητες που ακολουθούν είναι αφιερωμένη και σε μία παράμετρο. Να προσθέσουμε, επίσης, ότι σε όλα τα πειράματα η ακρίβεια ορίζεται ως $\varepsilon = 0.01$. Η υλοποίηση έχει γίνει στην γλώσσα προγραμματισμού Python, κυρίως με χρήση της βιβλιοθήκης Numpy ενώ τα σχήματα παράχθησαν με την βοήθεια της βιβλιοθήκης Matplotlib.

5.3.1 Η επίδραση της διάστασης

Δεδομένου ότι η κύρια υπόσχεση του αλγορίθμου pSGD είναι η πολυωνυμική υπολογιστική πολυπλοκότητα ως προς την διάσταση του προβλήματος, δεν μπορούμε παρά να ξεκινήσουμε την μελέτη μας από το πώς ο αλγόριθμος επηρεάζεται από την αύξηση της διάστασης. Για την αρχική εκτίμηση έχουμε μεταφράσει το $\tilde{\mathcal{O}}(d^2/\varepsilon^2)$ απλά σε d^2/ε^2.

Από την εξέταση του σχήματος 5.1 εξάγουμε δύο συμπεράσματα. Πρώτον, η αρχική εκτίμηση είναι επί της ουσίας ανεπηρέαστη από την διάσταση. Αυτό ήταν και το ζητούμενο, αφού στα δείγματα για τον υπολογισμό της εισαγάγαμε και την διάσταση. Συνεπώς, η απόσταση που βλέπουμε είναι πλέον συνάρτηση μόνο του α άρα ορθώς μένει αναλλοίωτη. Δεύτερον, και κυριότερον, βλέπουμε ότι η αύξηση της διάστασης σταδιακά «σπρώχνει» τις δυο καμπύλες προς τα πάνω. Βέβαια, κάτι τέτοιο είναι και πάλι αναμενόμενο αφού έχουμε χρησιμοποιήσει σταθερό αριθμό επαναλήψεων. Αν λύναμε την ανίσωση που συνδέει το ε με το M και συνεπώς με το d - είναι αντίστοιχη του παραδείγματος 3.2, εναλλακτικά φαίνεται και από την υλοποίηση μας - για σταθερό M θα λαμβάναμε αμέσως ότι η αύξηση του d οδηγεί σε αύξηση των ε.



Σχήμα 5.1: Αξιολόγηση του αλγορίθμου με αύξηση της διάστασης

5.3.2 Η επίδραση του συνόλου S

Σειρά έχει η εξέταση της απόδοσης του αλγορίθμου καθώς εναλλάσονται τα σχήματα. Σε αυτήν την περίπτωση, οι θεωρητικές εγγυήσεις μας απαγορεύουν μόνο τη χρήση συνόλων αμελητέας μάζας. Διαφορετικά, δεν υπάρχουν περιορισμοί. Φυσικά, δεν μπορούμε να ελέγξουμε οποιοδήποτε σύνολο μπορείς να σκεφτεί κάποιος αλλά δοκιμάσαμε τον αλγόριθμο με μερικά διαφορετικά. Αυτά ήταν:

Set 1: $\{(x, y) : x \geq 0\}$

Set 2: $\{(x, y) : x^2 + y^2 \leq 1.4\}$

Set 3: $\{(x, y) : |x| \leq 1.05, |y| \leq 1.05\}$

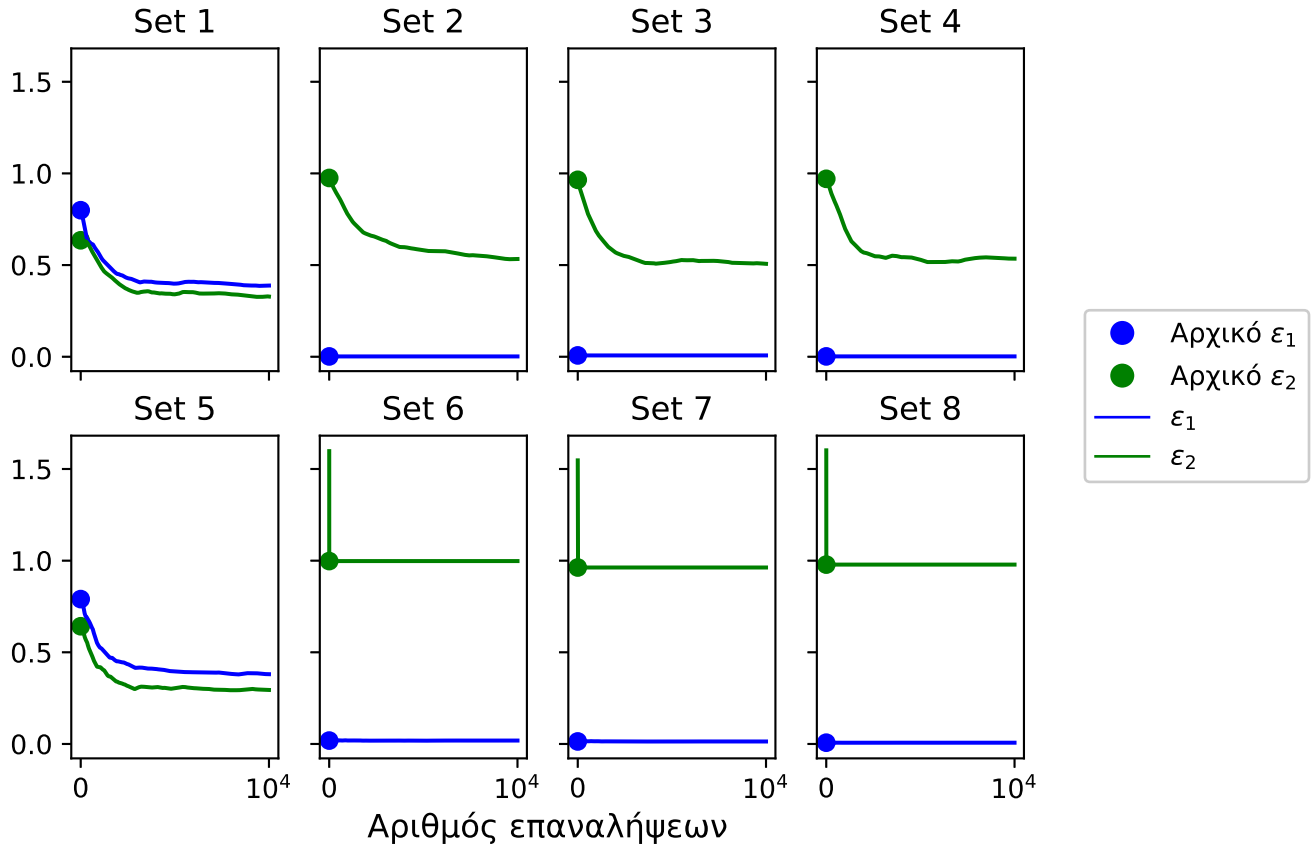
Set 4: $\{(x, y) : (x \geq 0 \cap \text{Set 2}) \cup (x < 0 \cap \text{Set 3})\}$

Set 5: $\overline{\text{Set 1}}$

Set 6: $\overline{\text{Set 2}}$

Set 7: $\overline{\text{Set 3}}$

Set 8: $\overline{\text{Set 4}}$



Σχήμα 5.2: Αξιολόγηση του αλγορίθμου σε διαφορετικά σύνολα αποκοπής

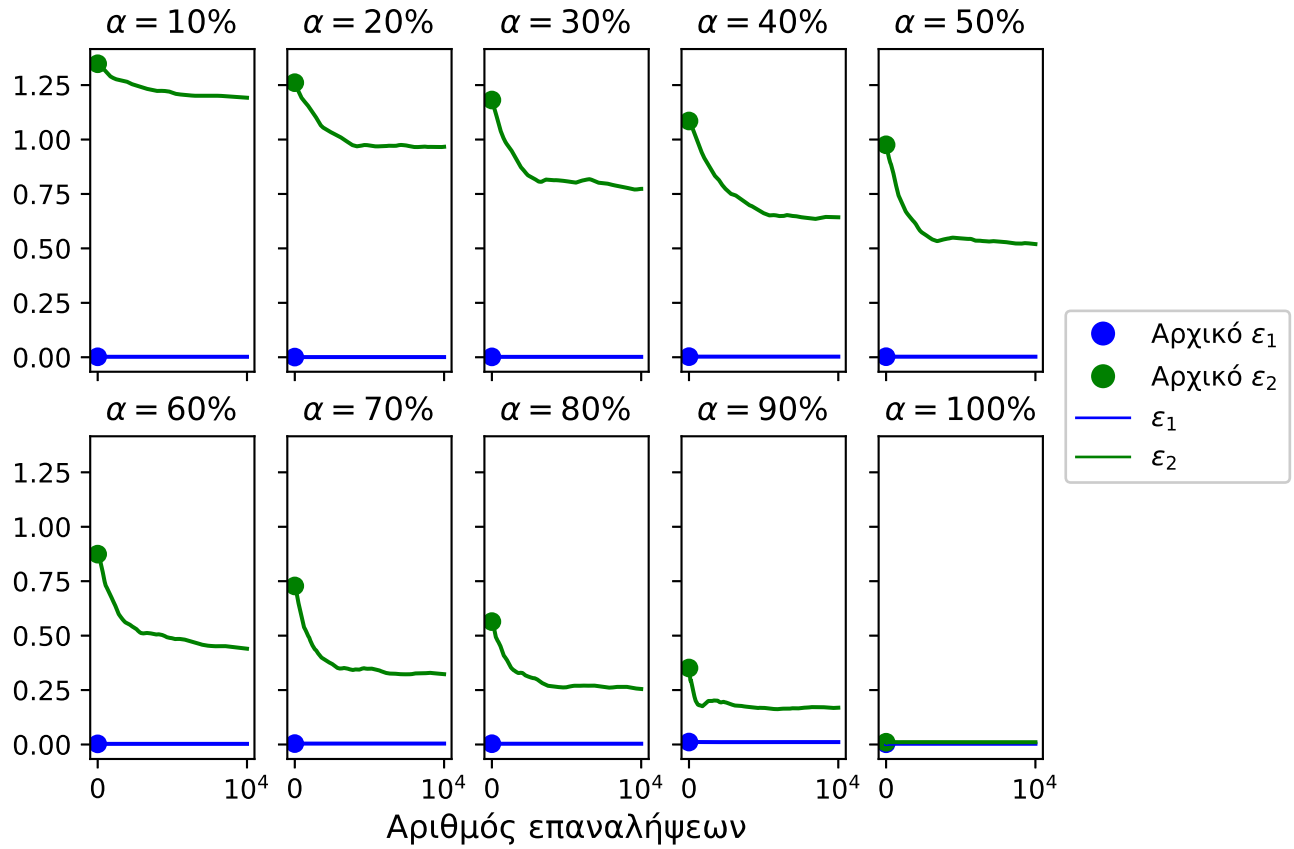
Το πρώτο σύνολο αποκοπής, που αποτελεί και το γενικό μας για όλες τις περιπτώσεις, είναι το θετικό ημιεπίπεδο, αντίστοιχα ημίχωρος στις υψηλότερες διαστάσεις. Το δεύτερο είναι είναι μία μπάλα με κέντρο την αρχή των αξόνων και ακτίνα $\sqrt{1,4}$ και το τρίτο ένα κουτί ίδιου κέντρου και πλευράς $2,1$. Να σημειώσουμε ότι τόσο η ακτίνα της μπάλας όσο και η πλευρά του κουτιού έχουν επιλεχθεί ώστε να ισχύει προσεγγιστικά $\alpha = 0.5$. Με το τέταρτο σύνολο επιδιώξαμε να παράξουμε ένα λίγο πιο σύνθετο και ταυτόχρονα μη κυρτό σχήμα, συνδυάζοντας τα δύο προηγούμενα σχήματα. Για τα υπόλοιπα 4, χρησιμοποιήσαμε την πράξη του συμπληρώματος. Οπότε τα σύνολα 1 και 5 είναι κυρτά και μη φραγμένα, τα σύνολα 2 και 3 κυρτά και φραγμένα, το σύνολο 4 είναι μη κυρτό και φραγμένο και τα 6,7 και 8 μη κυρτά και μη φραγμένα.

Στο σχήμα 5.2 βλέπουμε κάποια ενδιαφέροντα αποτελέσματα. Ξεκινώντας από τα αναμενομένα, βλέπουμε ότι τα σχήματα που αντιστοιχούν στα σύνολα 1 και 5 είναι πανομοιότυπα. Αν δεν υπήρχε η τυχαιότητα, θα αναμέναμε να είναι ακριβώς τα ίδια λόγω συμμετρίας. Έπειτα, περνάμε στα σύνολα 2,3 και 4. Εδώ βλέπουμε να έχει ξεκινήσει η σύγκλιση μεν, πιο αργή από το σύνολο 1 δε. Επίσης, παρατηρούμε ότι στη σφαίρα η σύγκλιση είναι ελαφρώς πιο ομαλή σε αντίθεση με τα άλλα δύο που οι καμπύλες πέφτουν πιο απότομα. Το πραγματικά ενδιαφέρον σημείο, όμως, είναι η εικόνα των σχημάτων 6-8. Βλέπουμε ένα καρφί(spike) στα πρώτα βήματα και μετά στασιμότητα σε τιμές περίπου ίσες με την αρχική. Κάτι τέτοιο δεν προβλεπόταν από την θεωρία και πρόκειται για το μοναδκό σημείο που δεν την επιβεβαιώσαμε. Το γιατί χρήζει

περαιτέρω μελέτης.

5.3.3 Η επίδραση του α

Προχωράμε στην διερεύνηση των τιμών του α. Να τονίσουμε ότι εδώ θα κάνουμε μια εξαίρεση στον τρόπο εργασίας μας και θα αλλάξουμε, μόνο για τώρα, και το σύνολο αποκοπής σε μπάλες κέντρου $\mathbf{0}$. Ο λόγος είναι ότι θέλουμε να τροποποιούμε το α οπότε ένα σταθερό σύνολο δεν μας εξυπηρετεί. Επιλέξαμε την μπάλα γιατί θέλουμε να έχουμε μια όσο το δυνατόν ομοιόμορφη μεταβολή στην παράμετρο. Αν επιμέναμε στα ημιεπίπεδα, θα αλλοιώναμε τα αποτέλεσματα λόγω των μεγάλων αλλαγών στην συνδιακύμανση που θα παρατηρούσαμε.

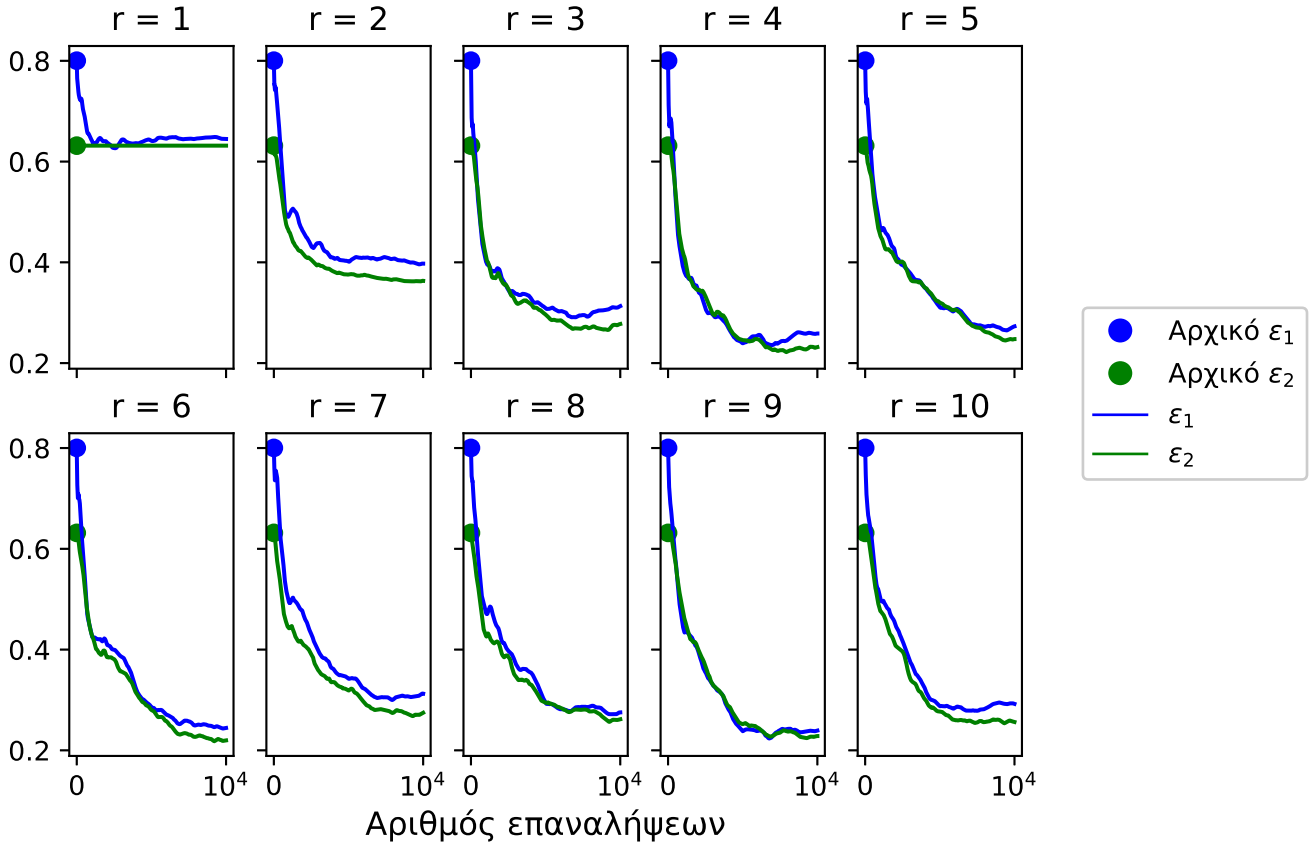


Σχήμα 5.3: Αξιολόγηση του αλγορίθμου για διαφορετικά α

Στο σχήμα 5.3 βλέπουμε τα αποτέλεσματα. Για άλλη μια φορά, υπάρχει συμφωνία με την θεωρία. Όσο μικραίνουμε το α τόσο χειροτερεύουμε την αρχική εκτιμήση και ταυτόχρονα την σύγκλιση. Φυσικά, αφού ο αλγόριθμος είναι βέλτιστος ακόμα κι όταν δεν υπάρχει απόκοπη, $\alpha = 100\%$, ξεκινάμε από μία άριστη αρχική εκτίμηση (θυμηθείτε το παράδειγμα 3.2) και την διατηρούμε.

5.3.4 Η επίδραση του r

Ακολουθεί η ακτίνα r του συνόλου προβολής. Μέχρι τώρα υπολογίζαμε και χρησιμοποιούσαμε την βέλτιστη ακτίνα r^*. Ή, για να είμαστε πιο ακριβείς, τις βέλτιστες ακτίνες r_1^*, r_2^*, r_3^*. Τώρα, θα χρησιμοποιήσουμε ένα κοινό r και για τις 3 ποσότητες.

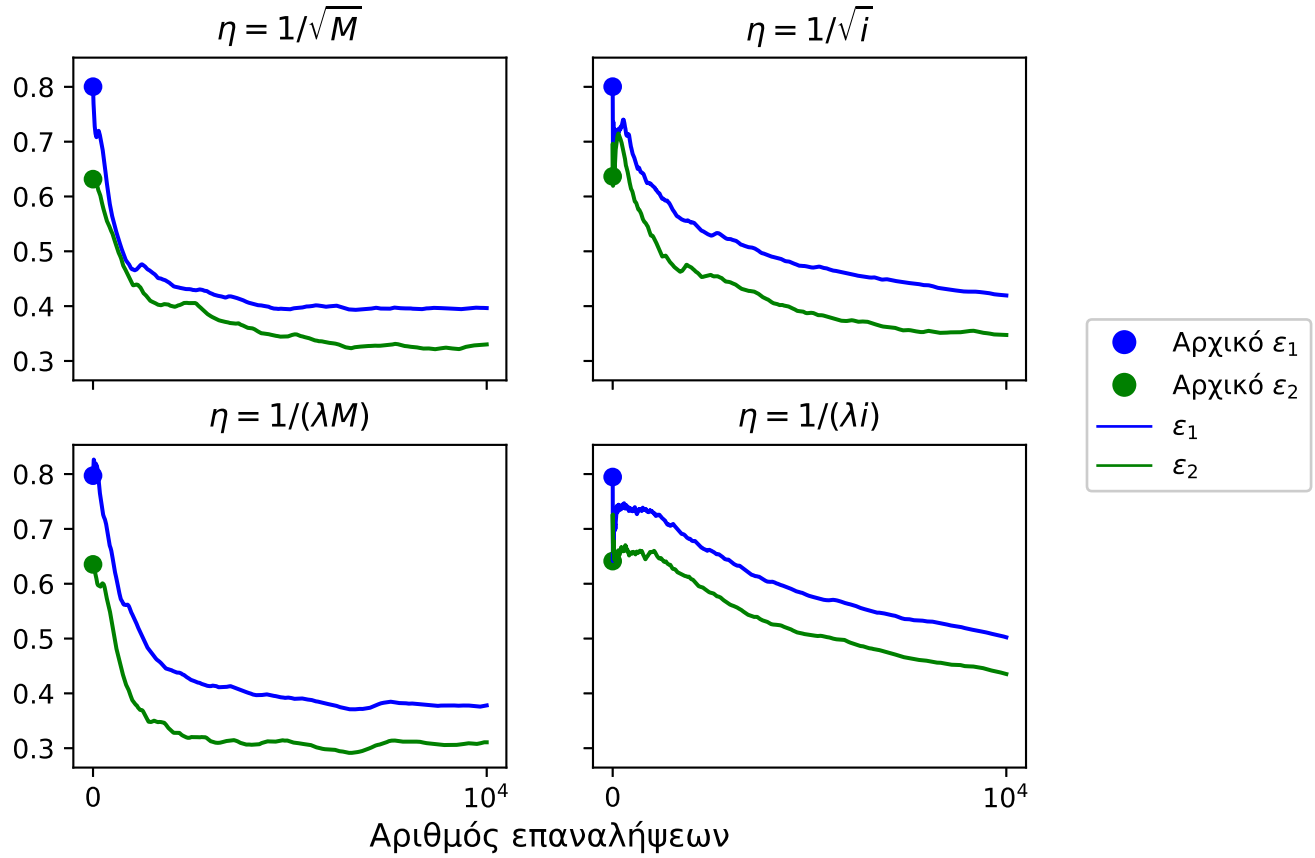


Σχήμα 5.4: Αξιολόγηση του αλγορίθμου για διαφορετικές ακτίνες r του συνόλου προβολής

Κοιτώντας ο αναγνώστης το σχήμα 5.4 έχει μάλλον δύο απορίες. Πώς προέκυψε η ευθεία καμπύλη συνδιακύμανσης για $r = 1$ και πού οφείλεται η βελτιώση που παρατηρείται ανάμεσα στις 3 πρώτες δοκιμές. Για να απαντήσουμε στην πρώτη ερώτηση, πρέπει να θυμηθούμε τον τρόπο με το οποίο προέκυψε η τρίτη συνθήκη του σύνολου προβολής, $\|\mathbf{T}^{-1}\|_{\mathrm{F}} \leq r$. Μας την επέβαλε ο αφφινικός μετασχηματισμός. Πρακτικά, μπορούμε να πούμε ότι αφορά στην κλίμακα(scale) του μετασχηματισμού. Συνεπώς, όταν την θέτουμε ίση με 1 απαιτούμε να μην έχουμε μεταβολή. Για την δεύτερη ερώτηση, πρέπει να πληροφορήσουμε τον αναγνώστη ότι $r^* \approx 2.7$ συνεπώς η βέλτιστη λύση δεν ανήκει στο σύνολο για μικρότερες τιμές. Επιπλέον, έχουμε μία μικρή βελτίωση για μεγαλύτερες τιμές. Ο λόγος είναι ότι μην τοποθετώντας την βέλτιστη λύση στο σύνορο, δίνουμε μία μεγαλύτερη ευχέρεια στον αλγόριθμό μας να την πλησιάσει. Θα προχωρήσουμε αφήνοντας για τον αναγνώστη μια ερώτηση: είναι δυνατόν να μεγαλώσουμε κατάλληλα το r ώστε να καταγράψουμε σχήματα με τις αποστάσεις να αποκλίνουν;

5.3.5 Η επίδραση του η

Στον ρυθμό μάθησης οφείλουμε να σταθούλε λίγο παραπάνω. Όπως είχαμε δηλώσει, ακολουθήσαμε άλλη πορεία από αυτήν που μας υπέδειξε η θεωρία. Για να το δικαιολογήσουμε θα δοκιμασούμε 4 διαφορετικά σχήματα για το η: θα θέσουμε $\lambda = 10^{-2}$ και θα δοκιμάσουμε το θεωρητικό $1/(\lambda i)$ και το αντίστοιχο σταθερό $1/(\lambda M)$ απέναντι στο καθιερωμένο μας πλεόν $1/(\sqrt{M})$ και το μεταβλητό του ανάλογο $1/(\sqrt{i})$.

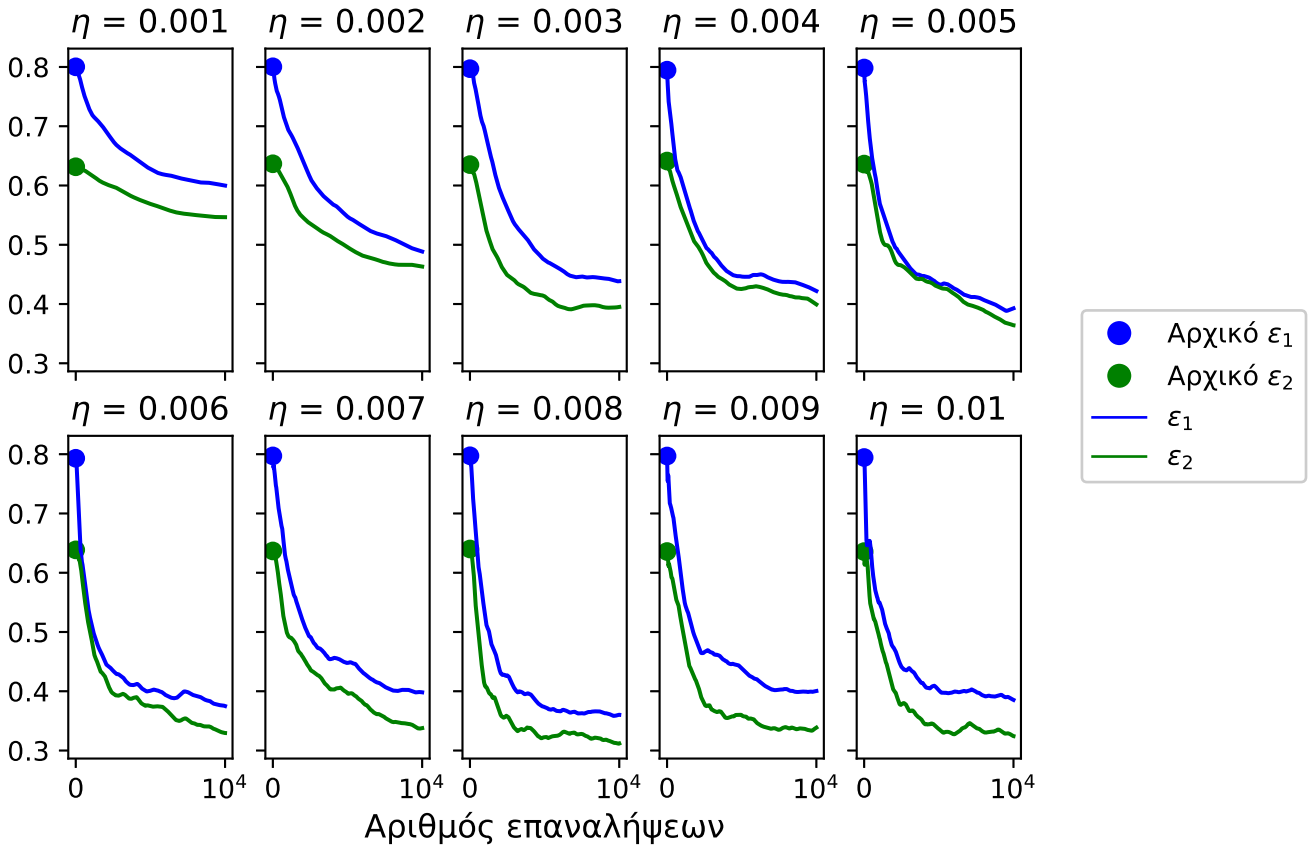


Σχήμα 5.5: Αξιολόγηση του αλγορίθμου για διαφορετικές επιλογές του η

Η σύγκριση φαίνεται στο σχήμα 5.5. Προσέξτε πως η πρόταση της θεωρίας φέρει το χειρότερο αποτέλεσμα. Από κοντά, αλλά λίγοτερο καλύτερα, και το έτερο σχήμα με μεταβλητό βήμα. Ο λόγος που παρατηρείται το φαινόμενο είναι πως η θεωρητική αιτιολόγηση της ιδέας που θέλει το βήμα να μειώνεται όσο πλησιάζουμε στον στόχο αποτυγχάνει στην πράξη να διορθώσει τα εκάστοτε πολύ θορυβώδη βήματα. Θα επανέλθουμε σε αυτό στην αμέσως επόμενη υποενότητα αλλά πρώτα να μιλήσουμε για τους νικητές. Παρατηρούμε ξεκάθαρα καλύτερα αποτελέσματα για σταθερό βήμα. Αυτό είναι γιατί είχαμε μια καλή υποεκτίμηση του λ. Με το σφάλμα που μας δίνει το 3.3, ανάλογο του $1/\lambda$, όσο περισσότερο υποεκτιμάμε την πραγματική κυρτότητα τόσο μεγαλύτερο σφάλμα θα λαμβάνουμε. Αντιθέτως, αν την υπερεκτιμήσουμε μάλλον θα παρατηρήσουμε απόκλιση. Σε ορισμένες πρακτικές εφαρμογές θα συναντήσουμε κάποια μεγάλη υποέκτιμηση, π.χ. $\lambda = 10^{-5}$ που θα αντισταθμίζεται από

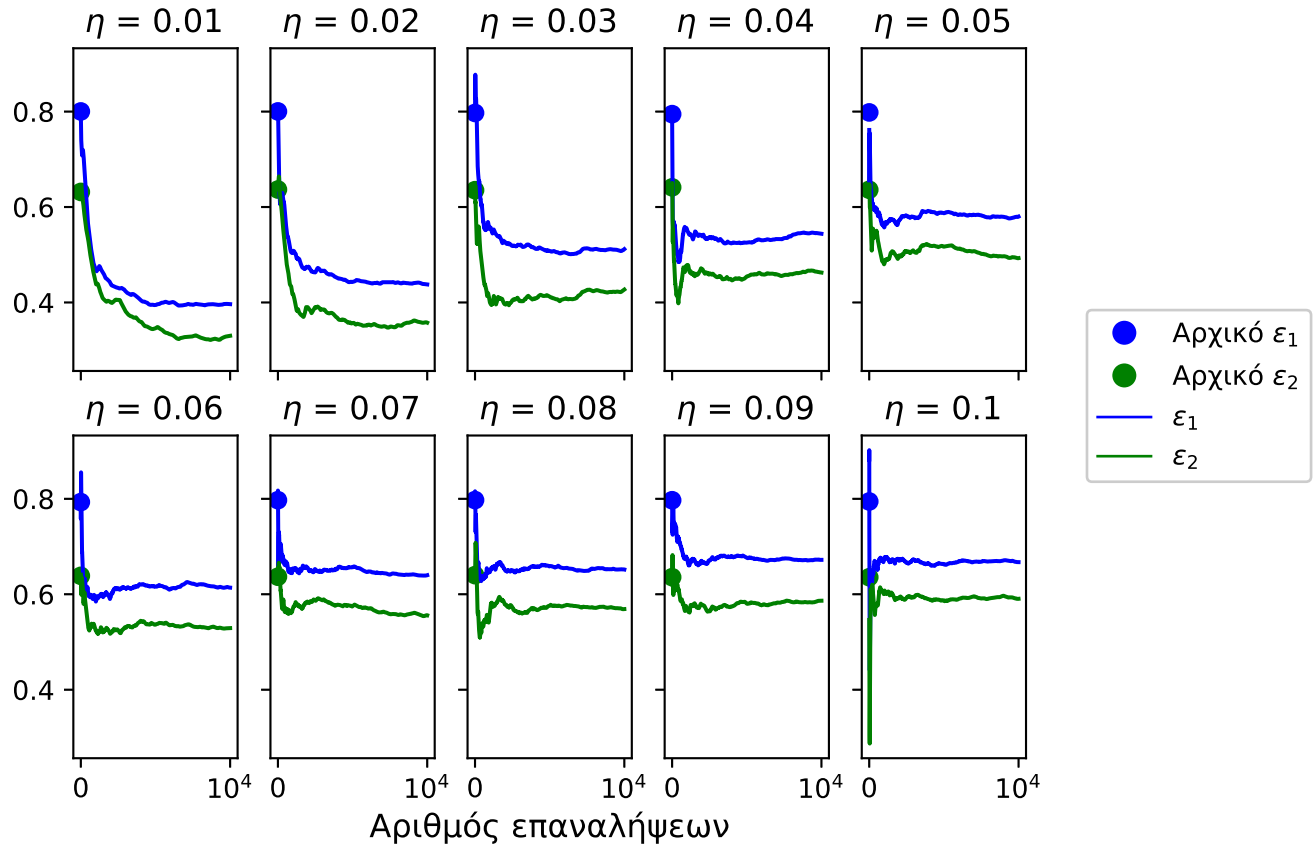
αυξημένο M.

Αφού δείξαμε γιατί επιλέξαμε ως η το $1/(\sqrt{M})$ πρέπει να εξετάσουμε και τον συντελεστή αυτού. Μην ξεχνάμε πως η θεωρία υπαγορεύει $\eta = \frac{B}{\rho\sqrt{M}}$. Για να διαπιστώσουμε την επίδραση του σταθερού συντελεστή πάνω στην υπόρριζη ποσότητα αντικαθιστούμε το $M = 10000$, υπολογίζουμε $\eta = 0,01c$ και μελετάμε την εξέλιξη του αλγορίθμου για διάφορες τιμές του c. Θα φανεί στην πορεία γιατί αλλά θα μοιράσουμε τις τιμές του c σε δύο σχήματα.



Σχήμα 5.6: Αξιολόγηση του αλγορίθμου για μικρότερους πολλαπλασιαστές του η

Στο πρώτο από αυτά 5.6 βλέπουμε πώς διαφοροποιούνται τα πράγματα για $c = \{\frac{1}{10}, \ldots, 1\}$. Στις μικρότερες τιμές η σύγκλιση είναι αρκετά πιο αργή. Ωστόσο, υπάρχει και μία ακόμη διαφορά. Βλέπουμε ότι και οι καμπύλες είναι πιο ομαλές. Χωρίς να δώσουμε κάποιον ορισμό του τι θεωρούμε ομαλό, προτρέπουμε τον αναγνώστει να κοιτάξει την διαφορά αναμέσα στο πρώτο και το τελευταίο σχήμα. Από την μία μεριά, οι καμπύλες για $\eta = 0,01$ πετυχαίνουν πολύ καλύτερο αποτέλεσμα. Από την άλλη, διαφαίνεται ένα «τρεμούλιασμα». Το λογικό επόμενο είναι να αναρωτηθούμε τι θα γίνει αν αυξήσουμε παραπάνω τον ρυθμό μάθησης. Θα αυξηθεί αυτό το «τρεμούλιασμα» με παράλληλη επιτάχυνση της σύγκλισης ή κάτι άλλο. Αξίζει μία σύντομη παύση για να το σκεφτείτε πριν προχωρήσετε στην επόμενη σελίδα.



Σχήμα 5.7: Αξιολόγηση του αλγορίθμου για μεγαλύτερους πολλαπλασιαστές του η

Δεν γνωρίζουμε πόσοι από σας περιμένατε το αποτέλεσμα του σχήματος 5.7. Κατ' αρχάς να διευκρινίσουμε ότι τώρα θέσαμε $c = \{1, 2, \ldots, 10\}$. Το αποτέλεσμα σίγουρα δεν είναι καλύτερο. Άρα η απάντηση στο αν αυξήθηκε ο ρυθμός σύγκλισης είναι αρνητική. Βασικά συνέβη το αντίθετο. Όχι μόνο δεν βελτιώθηκε αλλά όσο αυξάναμε το η χειροτέρευε κιόλας. Τι έγινε όμως με το «τρεμούλιασμα»;. Αν όχι και τα προηγούμενα, σίγουρα μας μαρτυράει την απάντηση το τελευταίο διάγραμμα. Δείτε ότι στα πρώτα βήματα η πράσινη καμπύλη κατεβαίνει μέχρι κάτω κι επιστρέφει, ενώ η μπλε ανεβαίνει απότομα πριν κατέβει. Αυτό το φαινόμενο είναι που μας δίνει έναυσμα για την μελέτη της διακύμανσης του gradient.

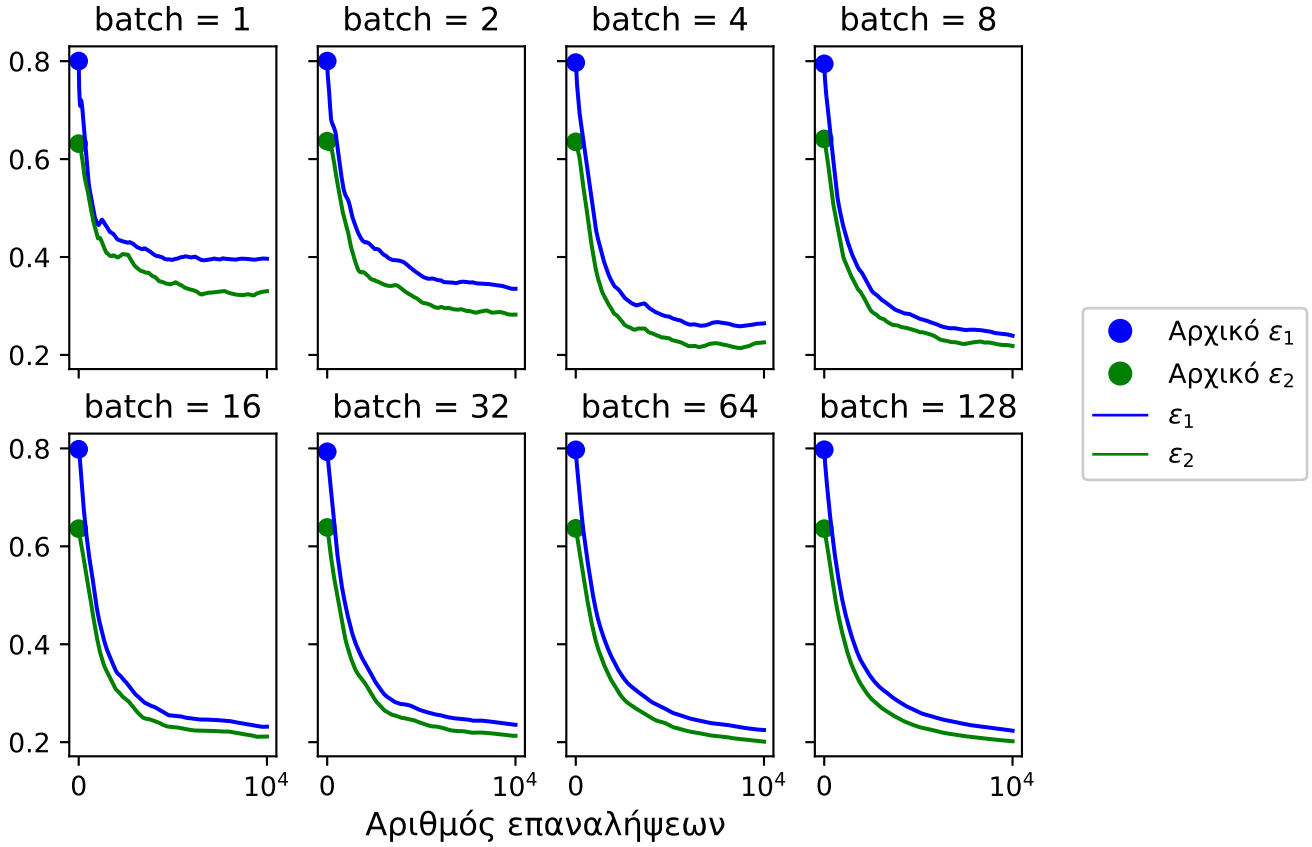
5.3.6 Η επίδραση της διακύμανσης του Gradient estimation

Το μέγεθος της συστάδας δειγμάτων αποτελεί την μοναδική παράμετρο την οποία μελετάμε ενώ δεν αναφερέται πουθενά στο [11]. Αυτό που βαπτίσαμε «τρεμούλιασμα» στην προηγούμενη ενότητα δεν είναι άλλο από την διακύμανση της εκτίμησης του gradient. Ενώ στο κεφάλαιο 3 είχαμε αρκεστεί στο πώς η χαλάρωση της απαίτησης από το πλήρες gradient σε μία εκτίμηση αυτού μας οδήγησε στον SGD επηρεάζοντας την ανάλυση της εξόδου του, στην πράξη μπορεί η τυχαία μεταβλητή $\mathbf{v}_i$ να έχει τόσο μεγάλη διακύμανση που να επηρεάζει σημαντικά ένα μεμονωμένο βήμα. Κοινά αποδεκτός τρόπος αντιμετώπισης δεν υπάρχει. Μπορούμε να χωρίσουμε τις προσεγγίσεις σε τρεις κατηγορίες:

- Θεωρητικές προσέγγισεις μείωσης της διακύμανσης
  Οι βασικότερες μέθοδοι αυτής της κατηγορίας είναι η *Προβλεπτική μείωση της διακύμανσης* - [35], η *Στοχαστική μέση κατάβαση κλίσης(SAG)* - [52] και η *Στοχαστική ανάβαση διπλών συντεταγμένων(SDCA)* - [57].

- Τροποποιήσεις του βασικού αλγορίθμου
  Διασημότερες τροποποιήσεις είναι η *Στοχαστική Κατάβαση Κλίσης με μέθοδο ροπών* - [53] και *Προσαρμοστικός αλγόριθμος κλίσης(AdaGrad)* - [17].

- Χρήση συστάδας δειγμάτων



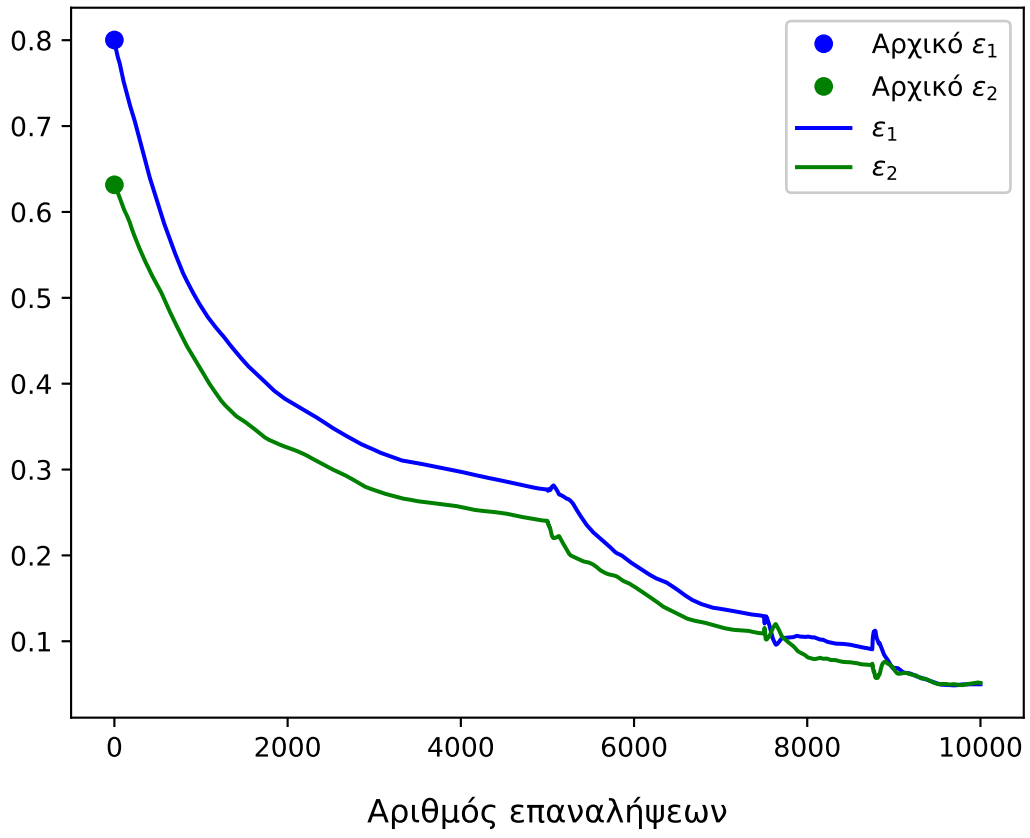
Σχήμα 5.8: Αξιολόγηση του αλγορίθμου για διαφορετικά μεγέθη συστάδας

Εμείς θα αξιοποιήσουμε την τεχνική της συστάδας δειγμάτων. Τα αποτελέσματά της διαπιστώνονται από το σχήμα 5.8. Η γενική ιδέα είναι αντί να χρησιμοποιήσουμε ένα $\mathbf{v}_i$ να χρησιμοποιήσουμε μια συστάδα από αυτά και να τα συνδυάσουμε καταλλήλως για μία βελτιωμένη εκτίμηση του gradient, π.χ. υπολογίζοντας τον μέσο όρο όπως πράξαμε και μεις εδώ. Παρά την απλότητά της, η μέθοδος χαίρει μεγάλη πρακτικής εφαρμογής και δεν υπολείπεται θεωρητικής αναλύσης. Στο [60], οι συγγραφείς προτρέπουν να αυξάνουμε το μέγεθος της συστάδας αντί να μειώνουμε τον ρυθμό μάθησης. Γενικά, όπως φάνηκε και από την σύνδεση των ενοτήτων της εργασίας μας, οι δύο αυτές ποσότητες αλληλοεξαρτώνται.

Απλοϊκά, όσο πιο σίγουροι είμαστε ότι μαθαίνουμε σωστά, τόσο πιο γρήγορα έχουμε τη δυνατότητα να μάθουμε. Περισσότερα μπορεί ο αναγνώστης να διαβάσει στο [49].

5.3.7 Βελτιώση της σύγκλισης

Θα ολοκληρώσουμε τα πειράματα με μία ολοκληρωμένη σύγκλιση. Ενώ μελετήσαμε ποιοτικά την επίδραση των παραμέτρων με όλα τα προηγούμενα σχήματα, σε κανένα δεν πετύχαμε εξαιρετικά καλή εκτίμηση των παραμέτρων. Κάτι που στην τελική είναι και το ζητούμενο της υλοποίησης ενός αλγορίθμου.



Σχήμα 5.9: Βελτίωση της σύγκλισης με χρήση συστάδας δειγμάτων και εποχών

Αυτό αλλάζει με το σχήμα 5.9. Όπως φαίνεται από την ομαλότητα της καμπύλης, χρησιμοποιήθηκε συστάδα δειγμάτων, μεγέθους 16. Επιπλεόν, προστέθηκε μία νέα στρατηγική. Η υλοποίηση είναι παρόμοια με τον αλγόριθμο Epoch-GD-Proj των Hazan and Kale στο [32]. Μία διαφορά είναι ότι σε κάθε εποχή ανανεώναμε την ακτίνα στη βέλτιστη με βάση την τρέχουσα εκτίμηση. Η χτυπητή διαφορά, όμως, έχει να κάνει με τα βήματα που απαρτίζουν την κάθε εποχή. Οι συγγραφείς διπλασίαζαν τα βήματα κάθε επόχης. Εμείς τα *υποδιπλασιάζαμε*, όπως φαίνεται και από το σχήμα. Παρότι, συνεπώς, δεν έχουμε κάποια θεωρητική εγγύηση για την μέθοδο, πετύχαμε αξιοσημείωτα αποτελέσματα με τις δύο καμπύλες να βρίσκονται περίπου στο $0,05$ μετά το περας των βημάτων.

5.4 Αποτελέσματα της μελέτης

Θα ολοκληρώσουμε το κεφάλαιο επισκεπτόμενοι ξανά τα αποτελέσματα των πειραμάτων μας.

- **Η διάσταση του προβλήματος**
  Πρώτο πειραματικό σταθμό μας αποτέλεσε η διάσταση της Γκαουσσιανής κατανομής που αναζητούμε. Ο αλγόριθμος είναι πολυωνυμικός ως προς τα δείγματα που χρειάζεται για την αρχική εκτίμηση, τα δείγματα ανά βήμα και τα ερωτήματα προς το μαντείο. Επαληθεύσαμε την πρώτη ιδιότητα υπολογίζοντας αρχικές εκτιμήσεις ίδιας αξιολόγησης ανεξαρτήτως διάστασης και την δεύτερη μέσα από την πρόοδο του αλγορίθμου για σταθερό αριθμό βημάτων

- **Το άγνωστο σύνολο αποκοπής**
  Έπειτα στραφήκαμε στο σύνολο αποκοπής. Είδαμε διάφορους συνδυασμούς συνόλων, σχετικούς με το αν είναι φραγμένα και κυρτά. Θεωρητικά όλα πληρούσαν τις προϋποθέσεις σύγκλισης όμως δεν είδαμε κάτι τέτοιο για μη κύρτα και μη φραγμένα.

- **Το μέτρο στο συνόλου αποκοπής**
  Μετά είδαμε πώς επηρεάζει τον αλγόριθμο το α. Για το ίδιο πρόβλημα, με σταδιακή αύξηση του μέτρου περάσαμε από μία πολύ κακή αρχική εκτιμήση με μικρή πρόοδο στην επιβεβαίωση ότι ο αλγορίθμος είναι βέλτιστος όταν δεν υπάρχει αποκοπή.

- **Η ακτίνα του συνόλου προβολής**
  Συνεχίσαμε αυξομειώνοντας την ακτίνα του συνόλου προβολής. Παρατηρήσαμε ότι για μικρότερες της πραγματικής ακτίνας υπάρχει ένα φράγμα στην απόσταση που μπορούμε να πλησιάσουμε την πραγματική κατανομή ενώ για όχι πολυ μεγαλύτερες η σύγκλιση δεν επηρεάζεται.

- **Ο ρυθμός μάθησης**
  Ακολούθως αλλάξαμε το η που προβλέπει η θεωρία σε κάποιο πιο συντηρητικό αλλά παραδόξως δείξαμε ότι επιτυγχάνει καλύτερα αποτελέσματα. Ταυτόχρονα, ελέγξαμε διάφορες τιμές που αντιστοιχούν στα φράγματα της μέσης τιμής και διακύμανσης του τυχαίου διανύσματος εκτίμησης κλίσης και ανακαλύψαμε σημαντικό θόρυβο στις πόσοτητες.

- **Η διακύμανση της εκτίμησης κλίσης**
  Οδηγούμενοι από αυτό μας το συμπέρασμα, μελετήσαμε την διακύμανση και αποφασίσαμε να εφαρμόσουμε την τεχνική της συστάδας δειγμάτων. Λάβαμε έτσι κάποια πρώτα βελτιωμένα αποτελέσματα.

- **Μια βελτιωμένη σύγκλιση**
  Τελικά, συνδυάσαμε την χρήση συστάδας δειγμάτων με την χρήση εποχών εκθετικά φθίνουσας διάρκειας κατά τις οποίες επανεκκινούσαμε τον αλγόριθμο με βάση την τελική εκτίμηση της κάθε εποχής και με αντίστοιχη μικρότερη ακτίνα του συνόλου προβολής για να καταφέρουμε να βρεθούμε πολύ κοντά στις αρχικές παραμέτρους.

Κεφάλαιο 6

Επίλογος

6.1 Σύνοψη

Σε αύτη τη διπλωματική που σιγά σιγά ολοκληρώνεται παρουσιάσαμε μία υλοποίηση του αλγορίθμου Projected Stochastic Gradient Descent που εισήγαγαν οι Daskalakis et al. το 2018 για την μάθηση των παραμέτρων μιας πολυδιάστατης Γκαουσσιανής κατανομής υπό καθεστώς αποκοπής σε πολυωνυμικό χρόνο ως προς τα δείγματα αλλά και ως προς τη διάσταση.

Χωρίσαμε την μελέτη μας σε 3 στάδια. Στο πρώτο από αυτά αποκτήσαμε ένα θεωρητικό υπόβαθρο από το χώρο των Πιθανοτήτων που απαιτήθηκε για να κατανοήσουμε τον αλγόριθμο. Θυμηθήκαμε τι είναι Γκαουσσιανή κατανομή και πραγματοποιήσαμε μια εισαγωγή στην αποκομμένη Στατιστική. Επιπλέον, στην απόσταση ολικής κατανομής βρήκαμε ένα εργαλείο αξιολόγησης του τελικού αλγορίθμου.

Επεκτείναμε το υπόβαθρό μας από τον χώρο της Μηχανικής Μάθησης. Πρώτα μέσα από την μύηση στην τεχνική της Εκτίμησης Μέγιστης Πιθανοφάνειας και εν συνέχεια στους Αλγορίθμους Κατάβασης Κλίσης. Ξεκινήσαμε από την ντετερμινιστική εκδοχή, γνωρίσαμε μέσα από αυτήν τις βασικές παραμέτρους αυτής της οικογένειας αλγορίθμων και εν τέλει μεταβήκαμε στην στοχαστική. Εκεί διατυπώσαμε τα θεωρήματα που αποτέλεσαν την ραχοκοκκαλιά της δουλειάς των Daskalakis et al., και άρα και της δικής μας.

Πριν φτάσουμε στη διατύπωση του αλγορίθμου, προβήκαμε σε μια ιστορική αναδρομή που εκτεινόταν σε 4 αιώνες. Είδαμε όλα τα προβλήματα που αντιμετώπισαν οι ερευνητές στην μάθηση από αποκομμένα δείγματα. Από την πρώτη επαφή του Bernoulli με τον κλάδο φτάσαμε στην λύση του ζητήματος μέσω της εισαγωγής νέων τεχνικών, τις οποίες είδαμε να βρίσκουν ευρεία εφαρμογή στο σήμερα.

Έχοντας αντιληφθεί τη δυσκολία του προβλήματος και συνεπώς την θεωρητική αξία της λύσης αντιλαμβανόμαστε και τα οφέλη που μπορούν να προκύψουν και από μία πρακτική υλοποίησή του pSGD. Πράγματι, αφού τον υλοποιήσαμε με επιτυχία, βλέπουμε μέσα από ένα σύνολο πειραμάτων να επιβεβαιώνονται όλα πλην ενός τα θεωρητικά πορίσματα και τελικά καταφέρνουμε να εκτιμήσουμε τις παραμέτρους με μεγάλη ακρίβεια.

6.2 Μελλοντικές προτάσεις

Δεν μπορούμε παρά να αναφέρουμε ως πρώτη πρόταση, δεδομένου ότι πρόκειται για το μόνο σημείο που η θεωρία δεν συμφωνεί με την πράξη 100% την ανάγκη για περαιτέρω μελέτη σχετικά με το σύνολο αποκοπής. Απαιτείται σχεδιασμος περισσότερων συνόλων με στόχο την κατανόηση της απουσίας σύγκλισης. Πιθανώς να πρέπει να αξιποιηθούν τεχνικές που αναπτύχθηκαν ή σχολιάστηκαν στην Ενότητα 5.3.6 για την επίτευξη ικανοποιητικών αποτελεσμάτων. Αφού συζητάμε ακόμη για το σύνολο, μια πιθανή επέκταση της δουλειάς μας θα ήταν η υλοποίηση του συγγενικού αλγορίθμου των Kontonis et al. που ανταλλάσει την απαίτηση χρήσης μαντείου με κάποιες ιδιότητες του συνόλου αποκοπής.

Ενδιαφέρον παρουσιάζει ακόμη η μελέτη και ενσωμάτωση τεχνικών βελτίωσης της σύγκλισης που συχνά χρησιμοποιούνται σε αντίστοιχες υλοποιήσεις αλλά και των θεωρητικών μεθόδων μείωσης της διακύμανσης. Προς την αντίθετη κατεύθυνση, αξίζει να μελετηθεί από θεωρητική σκοπιά η ευρεστική τεχνική της Ενότητας 5.3.7 που έδωσε τα αποτελέσματα.

Εξαιρετικές προοπτικες ενασχολήσης παρουσιάζει και η δοκιμάσια αυτοματοποίησης της επιλογής των παραμέτρων του αλγορίθμου μας. Τόσο του ρυθμού μάθησης και του μέγεθους συστάδας όσο και της ακτίνας του σύνολου προβολής. Ένας θεωρητικός αλγόριθμος εκτίμησης της ακτίνας ή μια ευρεστική τεχνική θα ενίσχυε σημαντικά την υλοποίηση μας καθώς θα της έδινε την δυνατότητα να εφαρμοστεί σε πραγματικά προβλήματα αυτόνομα, χωρίς την ανάγκη εξωτερικής γνώσης.

Βιβλιογραφία

[1] N. Balakrishnan and E. Cramer. *The art of progressive censoring.* Springer, 2014.

[2] D. Bernoulli. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir. *Histoire de l'Acad., Roy.Sci.(Paris) avec Mem*, pages 1–45, 1760.

[3] D. Bertsekas and J. Tsitsiklis. *Introduction to Probability.* Athena Scientific, 2008.

[4] A. Bhattacharyya, R. Desai, S. G. Nagarajan, and I. Panageas. Efficient statistics for sparse graphical models from truncated samples, 2020.

[5] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 161–168. Curran Associates, Inc., 2008.

[6] L. Bottou and Y. L. Cun. Large scale online learning. In *Advances in neural information processing systems*, pages 217–224, 2004.

[7] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[8] A. Cauchy. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

[9] A. C. Cohen. *Truncated and censored samples: theory and applications.* CRC press, 2016.

[10] C. Colombo and M. Diamanti. The smallpox vaccine: the dispute between bernoulli and d'alembert and the calculus of probabilities. *Lettera Matematica*, 2(4):185–192, Mar 2015.

[11] C. Daskalakis, T. Gouleakis, C. Tzamos, and M. Zampetakis. Efficient statistics, in high dimensions, from truncated samples. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 639–649. IEEE, 2018.

[12] C. Daskalakis, T. Gouleakis, C. Tzamos, and M. Zampetakis. Computationally and statistically efficient truncated regression. volume 99 of *Proceedings of Machine Learning Research*, pages 955–960, Phoenix, USA, 25–28 Jun 2019. PMLR.

[13] P. S. de Laplace. *Théorie analytique des probabilités*. Courcier, 1812.

[14] A. de Moivre. *The Doctrine of Chances: or, a method for calculating the probabilities of events in play*. W.Pearson, 1718.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[16] L. Devroye, A. Mehrabian, and T. Reddad. The total variation distance between high-dimensional Gaussians. *arXiv e-prints*, page arXiv:1810.08693, Oct. 2018.

[17] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[18] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[19] H. Fischer. *A history of the central limit theorem: from classical to modern probability theory*. Springer, 2011.

[20] R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

[21] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922.

[22] R. A. Fisher. Properties and applications of hh functions. *Mathematical tables*, 1:815–852, 1931.

[23] R. A. Fisher et al. 014: On the" probable error" of a coefficient of correlation deduced from a small sample. 1921.

[24] D. Fotakis, A. Kalavasis, and C. Tzamos. Efficient parameter estimation of truncated boolean product distributions. In *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 1586–1600. PMLR, 2020.

[25] F. Galton. An examination into the registered speeds of american trotting horses, with remarks on their value as hereditary data. *Proceedings of the Royal Society of London*, 62:310–315, 1897.

[26] C. F. Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Frid. Perthes, Besser, 1809.

[27] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[28] A. Hald. Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. *Scandinavian Actuarial Journal*, 1949(1):119–134, 1949.

[29] M. Halperin. Maximum likelihood estimation in truncated samples. *The Annals of Mathematical Statistics*, 23(2):226–238, 1952.

[30] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

[31] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[32] E. Hazan and S. Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436, 2011.

[33] H. Hotelling. Fitting generalized truncated normal distributions. In *Annals of Mathematical Statistics*, volume 19, pages 596–596. INST MATHEMATICAL STATISTICS IMS BUSINESS OFFICE-SUITE 7, 3401 INVESTMENT . . . , 1948.

[34] P. J. Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

[35] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[36] O. Kallenberg. *Foundations of Modern Probability*. Springer-Verlag New York, 2002.

[37] W. Karush. Minima of functions of several variables with inequalities as side conditions. In *Traces and Emergence of Nonlinear Programming*, pages 217–245. Springer, 2014.

[38] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. E. Schapire, and L. Sellie. On the learnability of discrete distributions. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 273–282, 1994.

[39] A. N. Kolmogorov. On logical foundations of probability theory. In *Probability Theory and Mathematical Statistics*, pages 1–5, Berlin, Heidelberg, 1983. Springer Berlin Heidelberg.

[40] V. Kontonis, C. Tzamos, and M. Zampetakis. Efficient truncated statistics with unknown truncation. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1578–1595, 2019.

[41] H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Traces and emergence of nonlinear programming*, pages 247–258. Springer, 2014.

[42] A. Lee. Table of the gaussian "tail" functions; when the "tail" is larger than the body. *Biometrika*, 10(2/3):208–214, 1914.

[43] A. A. Markov. On a question by di mendeleev. *Zapiski Imperatorskoi Akademii Nauk*, 62(1-24):12, 1890.

[44] T. M. Mitchell. Machine learning, 1997.

[45] S. G. Nagarajan and I. Panageas. On the analysis of em for truncated mixtures of two gaussians. volume 117 of *Proceedings of Machine Learning Research*, pages 634–659, San Diego, California, USA, 08 Feb–11 Feb 2020. PMLR.

[46] K. Pearson. On the systematic fitting of curves to observations and measurements. *Biometrika*, 1(3):265–303, 1902.

[47] K. Pearson and A. Lee. On the generalised probable error in multiple normal correlation. *Biometrika*, 6(1):59–68, 1908.

[48] K. B. Petersen and M. S. Pedersen. The matrix cookbook, 2012.

[49] X. Qian and D. Klabjan. The impact of the mini-batch size on the variance of gradients in stochastic gradient descent. *arXiv preprint arXiv:2004.13146*, 2020.

[50] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

[51] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[52] N. L. Roux, M. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. In *Advances in neural information processing systems*, pages 2663–2671, 2012.

[53] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[54] H. Schneider. *Truncated and Censored Samples from Normal Populations*. Marcel Dekker, Inc., USA, 1986.

[55] I. Schneider. Abraham de moivre, the doctrine of chances (1718, 1738, 1756). *Landmark Writings in Western Mathematics 1640-1940*, pages 105–120, 2005.

[56] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[57] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

[58] O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. volume 28 of *Proceedings of Machine Learning Research*, pages 71–79. PMLR, 2013.

[59] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[60] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2017.

[61] G. Strang. *Introduction to Linear Algebra*. Wellesley Cambridge Press, 2016.

[62] J. Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24–36, 1958.

[63] Y. L. Tong. *The multivariate normal distribution*. Springer Science & Business Media, 2012.

[64] J. W. Tukey. Sufficiency, truncation and selection. *Ann. Math. Statist.*, 20(2):309–311, 06 1949.

[65] A. M. Turing. Systems of logic based on ordinals. *Proceedings of the London Mathematical Society, Series 2*, 45:161–228, 1939.

[66] E. W. Weisstein. Affine transformation. From MathWorld—A Wolfram Web Resource. `https://mathworld.wolfram.com/AffineTransformation.html`.

[67] E. W. Weisstein. Frobenius norm. From MathWorld—A Wolfram Web Resource. `http://mathworld.wolfram.com/FrobeniusNorm.html`.

[68] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9:60–62, 1938.

[69] S. Wu, A. G. Dimakis, and S. Sanghavi. Learning distributions generated by one-layer relu networks. In *Advances in Neural Information Processing Systems 32*, pages 8107–8117. Curran Associates, Inc., 2019.