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

Μάθηση Κατανομών από Ελλιπή Δείγματα

Διπλωματικη Εργασια

της

ΜΑΜΑΛΗ ΑΙΚΑΤΕΡΙΝΗΣ

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

Εργαστηριο Λογικης και Επιστημης Υπολογιστων
Αθήνα, Φεβρουάριος 2022

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Λογικής και Επιστήμης Υπολογιστών

Distribution Learning from Truncated Samples

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΜΑΜΑΛΗ ΑΙΚΑΤΕΡΙΝΗΣ

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

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 4η Φεβρουαρίου 2022.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
....	....	....
Δημήτρης Φωτάκης	Αριστείδης Παγουρτζής	Χρήστος Τζάμος
Αν. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επ. Καθηγητής UW

Αθήνα, Φεβρουάριος 2022

(Υπογραφή)

....

Μαμαλη Αικατερινη

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

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Λογικής και Επιστήμης Υπολογιστών

Περίληψη

Η παρούσα διπλωματική εργασία ασχολείται με το θεμελιώδες πρόβλημα της εκμάθησης κατανομών από ελλιπή δείγματα. Στο πλαίσιο αυτό στόχος είναι η εκτίμηση μιας κατανομής πιθανότητας με βάση ένα σύνολο δειγμάτων που μπορεί να είναι ελλιπές. Αυτό σημαίνει ότι όποια δείγματα της κατανομής δεν ανήκουν σε ένα συγκεκριμένο, άγνωστο σύνολο, που ονομάζουμε σύνολο αποκοπής, αφαιρούνται από το σύνολο των δειγμάτων. Η δυσκολία κλιμακώνεται όταν απαιτούμε η εκτίμηση αυτή να δίνεται από έναν αλγόριθμο αποδοτικά, δηλαδή με τη χρήση πολυωνυμικού πλήθους δειγμάτων και μετά από πολυωνυμικό αριθμό επαναλήψεων. Μελετούμε την περίπτωση δύο συγκεκριμένων κατανομών: της Διωνυμικής κατανομής του Poisson και του μοντέλου Mallows για κατανομές κατάταξης. Σκοπός μας είναι η εύρεση εκείνων των συνθηκών για το σύνολο αποκοπής που είναι τόσο αναγκαίες όσο και ικανές για την επιτυχή εκμάθηση των κατανομών. Στην περίπτωση της Διωνυμικής κατανομής Poisson, αποδεικνύουμε ότι το πρόβλημα είναι, γενικά, αδύνατο, αλλά γίνεται ευκολότερο καθώς η κατανομή πλησιάζει την Κανονική κατανομή. Το γεγονός αυτό δηλώνει μία ενδιαφέρουσα μετάβαση στη δυσκολία του προβλήματος. Στην περίπτωση του μοντέλου Mallows, διατυπώνουμε μία ικανή συνθήκη για την επιλυσιμότητα του προβλήματος.

Λέξεις Κλειδιά

Θεωρία Μάθησης, Μάθηση Κατανομών, Μάθηση από Ελλιπή Δείγματα, Αποκομμένες Κατανομές, Θεωρία Πιθανοτήτων, Διωνυμική Κατανομή Poisson, Μοντέλο Mallows

Abstract

This thesis is concerned with the fundamental problem of learning distributions from truncated samples. In this setting the purpose is to estimate a probability distribution based only on truncated samples. That means that samples falling outside a specific, unknown set are not available. The challenge becomes greater when we demand that these estimations are given by an efficient -in terms of sample and traditional complexity-algorithm. We study the learnability of two specific distributions in this setting: the Poisson Binomial Distribution and the Mallows Distribution. We are interested in those conditions on the truncation set that care both sufficient and necessary to learn these distributions. In the first case, we are faced with an impossible problem that becomes easier as the distribution gains structure, thus indicating an interesting transition on the difficulty of the problem. For the Mallows Model we give a sufficient condition and recognise the sub-optimality of a well-established method in the field of rank aggregation.

Keywords

Learning Theory, Distribution Learning, Learning from Truncated Samples, Truncated Distribution, Probability Theory, Poisson Binomial Distribution, Mallows Model

Ευχαριστίες

Κατ΄ αρχάς, θα ήθελα να ευχαριστήσω τον κ. Φωτάκη για όλη την υποστήριξη κατά την εκπόνηση αυτής της εργασίας, την ευκαιρία που μου έδωσε να μελετήσω ένα τόσο ενδιαφέρον αντικείμενο καθώς και επειδή αποτέλεσε έμπνευση ώστε να ασχοληθώ με αυτό.

Ομοίως, θα ήθελα να ευχαριστήσω τον Άλκη Καλαβάση που κατέστησε την παρούσα εργασία εφικτή, από όλες τις οπτικές γωνίες, για τα θαυμάσια πράγματα που έμαθα με τη βοήθειά του, για όλο το χρόνο που αφιέρωσε και για την απεριόριστη υπομονή.

Contents

List of Figures

List of Tables

Κεφάλαιο 1

Εκτεταμένη Ελληνική Περίληψη

1.1 Εισαγωγή

Ένας από τους πιο πολυσυζητημένους τομείς της σύγχρονης επιστήμης είναι, αδιαμφισβήτητα, η Μηχανική Μάθηση. Η τεχνική του συμπερασμού με βάση δεδομένα επιτυγχάνει να αντιμετωπίσει πολύπλοκα προβλήματα που δυσκολεύουν ακόμα και τους ανθρώπους όπως η αναγνώριση συναισθημάτων [VdMH08]. Η Θεωρία Μάθησης παρέχει τη μαθηματική θεμελίωση και επιστημονική τεκμηρίωση των εν λόγω τεχνικών [Vap99], [Val84].

Στο αυτά τα πλαίσια, η Μάθηση Κατανομών, η οποία πρώτη φορά μελετάτε στο [KMR$^+$94], κατέχει πρωταγωνιστικό ρόλο. Ο Στατιστικός συμπερασμός μελετούσε τα φαινόμενα μέσω πιθανοτικών μοντέλων τους εδώ και δεκαετίες. Συγκεκριμένα, βασίζεται στα δεδομένα για τη συμπεριφορά των συστημάτων ώστε να προσδιορίσει το σωστό μοντέλο που περιγράφη το φαινόμενο. Η αυτοματοποίηση αυτής της μεθόδου ώστε να επιτελείται από μηχανές-υπολογιστές εγκαθιδρύει ένα νέο πεδίο έρευνας, εισάγει πρωτοφανείς προκλήσεις και παρέχει πληροφορίες σχετικά με τα όρια των ικανοτήτων μας.

Η Μάθηση Κατανομών από Ελλιπή Δείγματα συγκαταλέγεται μεταξύ των πιο επιτακτικών από αυτές τις προκλήσεις. Σε αυτό το πρόβλημα υποθέτουμε ότι κάποιες συμπεριφορές-δεδομένα του συστήματος δεν είναι ανιχνεύσιμες από τον παρατηρητή (ή απλώς χάνονται στην πορεία). Έτσι, έχουμε πρόσβαση αποκλειστικά σε δεδομένα που ανήκουν σε ένα συγκεκριμένο σύνολο. Είναι φανερό ότι το εν λόγω πρόβλημα παρέχει μία μετάβαση μεταξύ του εφικτού και του ανέφικτου. Στην τετρτιμμένη περίπτωση που μόλις ένα δείγμα παρέχεται, η ανάκτηση οποιασδήποτε πληροφορίας σχετικά με τη συμπεριφορά του φαινομένου είναι αδύνατη. Αντίθετα, αν τα δεδομένα δεν είναι ελλιπή το πρόβλημα ανάγεται στην παραδοσιακή μορφή του, που θα είναι επιλύσιμη (εδώ δεχόμαστε την υπόθεση ότι κανείς ακολουθεί πορεία από το απλούστερο στο συνθετότερο κατλα τη διαδικασία επίλυσης ενός προβλήματος). Συνεπώς, η Μάθηση Κατανομών από Ελλιπή Δείγματα ασχολείται με ένα σύγχρονο πρόβλημα που βρίσκεται στο σύνορο του πραγματοποιήσιμου. Η ακριβής θέση του συνόρου εξαρτάται από το σύνολο αποκοπής S. Σε αυτή τη διπλωματική εργασία μας απασχολεί ο χαρακτηρισμός του συνόλου S ώστε να εγγυάται τη μάθηση της κατανομής. Ακόμα περισσότερο, επιθυμούμε να βρούμε εκείνα τις συνθήκες για το S που είναι απαραίτητες για την εκμάθησης μιας κατανομής.

Η Μάθησης από Ελλιπή Δεδομένα μπορεί να θεωρηθεί ένα υπο-πρόβλημα της Σθεναρής Στατιστικής, βλ. [Hub65], [Hub92]. Σε αυτό το πλαίσιο, επιθυμούμε να αναπτύξουμε μεθόδους που δεν επηρεάζονται από μικρές ανακρίβειες στα δεδομένα τους. Η ανθεκτικότητα σε σφάλματα είναι ζωτικής σημασίας για τα πραγματικά συστήματα που είναι ευάλωττα απέναντι στην κακοβουλία ή, απλώς, την άγνοια των χρηστών. Η Σθεναρή Μάθηση Κατανομών ([DKK$^+$19]) αναπτύσσει αλγορίθμους που είναι ανθεκτικοί σε μία τάξης ε φθορά των δειγμάτων. Οποιοδήποτε είδος φθοράς είναι επιτρεπτό σε αυτα τό το πρόβλημα, συμπεριλαμβανομένης της διαγραφής, πρόσθήκης ή αντικατάστασης των δεδομένων, εφόσον αυτή περιορίζεται σε ένα ε ποσοστό αυτών. Προκύπτει ότι το αποτέλεσμα αυτών των αλγορίθμων δεν μπορεί να είναι περισσότερο από ε-ακριβές. Δηλαδή, η κατανομή που θα βρει ο αλγόριθμος θα έχει ε-σφάλμα στη χειρότερη περίπτωση [JO19]. Πίσω στο πρόβλημα των Ελλιπών Δεδομένων, σημειώνουμε ότι επιτρέπουμε αποκλειστικά τη διαγραφή δειγμάτων. Έτσι είμαστε σε θέση να αποσυνδέσουμε το ποσοστό της φθοράς των δεδομένων (έστω α) από την ανακρίβεια του αποτελέσματος (έστω ε).

Παρά την προφανή χρησιμότητα λύσεων για το παραπάνω πρόβλημα μόλις πρόσφατα δόθηκε μία πλήρης λύση για την περίπτωση της Κανονικής κατανομής σε d διαστάσεις στο [DGTZ18]. Ακολούθησε μία σειρά αποτελεσμάτων που μελετά το πρόβλημα μάθησης κατανομών από ελλιπή δείγματα στα πλαίσια των συνεχών κατανομών [KTZ19], [DKTZ21], των διακριτών κατανομών [FKT20], της μίξης κατανομών [NP20] και της παλινδρόμησης [DGTZ19], [IZD20], [DSYZ21]. Στο [FKT20] πραγματοποιείται μια προσαρμογή των τεχνικών για την μάθηση της Κανονικής κατανομής (που δόθηκε στο [DGTZ18]) για την περίπτωση μιας διακριτής κατανομής, της κατανομής της μίξης n κατανομών Βερνουλλι. Η διαδικασία αυτή φανερώνει μία εγγενή ευαισθησία των διακριτών κατανομών στην αποκοπή τως δεδομένων τους. Αποδεικνύεται ότι ορισμένες επιπρόσθετες (σε σχέση με την περίπτωση της Κανονικής κατανομής) υποθέσει για το σύνολο αποκοπής S είναι *αναγκαίες* για την αποδοτική εκμάθησης της εν λόγω διακριτής κατανομής.

Στην εργασία αυτή μελετάμε σε βάθος το φαινόμενο αυτό. Εξετάζουμε δύο συγκεκριμένες διακριτές κατανομές και αναλύουμε τις συνθήκες που πρέπει να ισχύουν για το S και που ελέγχουν την ικανότητα εκμάθησης.

1.2 Θεωρία Πιθανοτήτων

Στην παρούσα ενότητα παρουσιάζουμε κάποιες βασικές έννοιες της θεώριας πιθανοτήτων. Αυτές θα αποτελέσουν τα εργαλεία που θα οδηγήσουν στα συμπεράσματα αυτής της εργασιας.

Αρχικά, υπενθυμίζουμε κάποιες βασικές κατανομές που εμφανίζονται κατ΄ εξακολούθηση στα παρακάτω.

- *Κατανομή Bernoulli*
  Μια τυχαία μεταβλητή ακολουθεί την κατανομή Bernoulli, $X \sim \mathrm{Be}(p)$ όταν παίρνει την τιμή 1 με πιθανότητα p και την τιμή 0 διαφορετικά. Στην ουσία μια τυχαία κατανομή Bernoulli προσδιορίζει αν ένα γεγονός που έχει πιθανότητα p θα συμβεί. Γι΄ αυτό η X ονομάζεται και *δείκτρια* τυχαία μεταβλητή.

- *Διωνυμική Κατανομή*
  Μια τυχαία μεταβλητή ακολουθεί τη Διωνυμική κατανομή, $X \sim \mathrm{Bin}(n, p)$ αν

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}, k \in [n] \,.$$

- *Κανονική Κατανομή*

Μια τυχαία μεταβλητή ακολουθεί τη Κανονική κατανομή, $X \sim N(\mu, \sigma^2)$ αν

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}.$$

Στη συνέχεια υπενθυμίζουμε το Κεντρικό Οριακό Θεώρημα. Το ΚΟΘ αποτελεί το πιο θεμελώδες αποτέλεσμα της θεωρίας πιθανοτήτων. Σύμφωνα με αυτό, η κατανομή του αθροίσματος μιας ακολουθίας ανεξάρτητων, ισόνομων τυχαίων μεταβλητών προσεγγίζει την κανονική κατανομή καθώς το πλήθος των όρων του αθροίσματος τείνει στο άπειρο.

Θεώρημα 1.2.1 (Κεντρικό Οριακό Θεώρημα)**.** *Έστω $X_1, \ldots, X_n$ μια ακολουθία ανεξάρτητων, ισόνομων τυχαίων μεταβλητών, με μέση τιμή $\mathbb{E}[X_i] = \mu$ και διασπορά $\mathrm{Var}[X_i] = \sigma^2$. Συμβολίζουμε:*

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

Τότε η αθροιστική συνάρτηση κατανομής της Z_n συγκλίνει στην αθροιστική συνάρτηση κατανομής της Κανονικής κατανομής, δηλαδή

$$\lim_{n\to\infty} \Pr\left[Z_n \leq z\right] = \Phi\left(z\right),$$

για κάθε $z \in \mathbb{R}$.

Σημειώνουμε ότι η απαίτηση οι τυχαίες μεταβλητές να είναι ισόνομες μπορεί να αρθεί υπό κάποιους άλλους, πολύ ελαστικούς περιορισμούς.

Σημαντικό ρόλο στα αποτελέσματα αυτής της διπλωματικής διαδραματίζει η σύγκριση μεταξύ κατανομών. Για το σκοπό αυτό, ορίζουμε την TV απόσταση μεταξύ κατανομών.

Ορισμός 1.2.1 (TV απόσταση)**.** *Έστω οι κατανομές πιθανότητας P, Q σε ένα χώρο πιθανότητας $(\mathbb{R}, \mathcal{B})$. Ορίζουμε την TV απόσταση μεταξύ των P, Q ως:*

$$\mathrm{TV}\left(P, Q\right) = \sup_{B \in \mathcal{B}} \left|P\left(B\right) - Q\left(B\right)\right|.$$

Σημειώνουμε ότι στην περίπτωση που οι κατανομές είναι διακριτές με πεδίο ορισμού F ισχύει η ισοδύναμη σχέση:

$$\mathrm{TV}\left(P, Q\right) = \frac{1}{2} \sum_{k \in F} \left|p\left(k\right) - q\left(k\right)\right|.$$

1.3 Θεωρία Μάθησης

Το γενικό πρόβλημα με το οποίο ασχολείται η παρούσα διπλωματική εργασία αφόρα τη μάθηση κατανομών. Θεωρούμε ότι έχουμε πρόσβαση σε ένα σύνολο δεδομένων που προέρχονται από μια συγκεκριμένη, άγνωστη κατανομή. Σκοπός μας είναι η ανάπτυξη αλγορίθμων που να βρίσκουν μια καλή προσέγγιση της άγνωστης κατανομής αποδοτικά.

Ορίζουμε την έννοια της αποδοτικότητας με δύο κριτήρια: πρώτον, ο αλγόριθμος χρειάζεται πολυωνυμικό πλήθος δειγμάτων από την άγνωστη κατανομή και, δεύτερον, τερματίζει μετά από πολυωνυμικό αριθμό δειγμάτων. Φυσικά, ο όρος 'πολυωνυμικό' υπονοεί ότι υπάρχουν κάποιες παράμετροι ως προς τις οποίες θα έχουμε πολυωνυμική εξάρτηση. Οι παράμετροι αυτοί καθορίζονται από το εκάστοτε πρόβλημα μάθησης.

Επιπλέον, πρέπει να προσδιορίσουμε τον όρο 'καλή προσέγγιση'. Σε αυτήν την εργασία στόχος μας είναι να προσεγγίσουμε την άγνωστη κατανομή σε TV απόσταση το πολύ ε, δηλαδή όσο μικρή θέλουμε.

Ταυτόχρονα, δεχόμαστε μία ακόμη πρόκληση. Τα δείγματα της άγνωστης κατανομής στα οποία έχουμε πρόσβαση πρέπει να ανήκουν σε ένα προκαθορισμένο σύνολο S το οποίο είναι άγνωστο στον αλγόριθμο (τουλάχιστον στη γενική περίπτωση). Αυτό είναι το πρόβλημα της μάθησης κατανομών από ελλιπή δείγματα. Ονομάζουμε το σύνολο S σύνολο αποκοπής, αφού όσα δείγματα δεν ανήκουν σε αυτό 'κόβονται'. Τελικά, τα δείγματα στα οποία έχουμε πρόσβαση φαίνεται να ανήκουν σε μια καινούργια κατανομή, την οποία ονομάζουμε αποκομμένη κατανομή.

Ορισμός 1.3.1 (Αποκομμένη Κατανομή)**.** *Έστω μία κατανομή πιθανότητας $\mathcal{D}$ με πεδίο ορισμού $\mathcal{Z}$. Θεωρούμε το σύνολο αποκοπής $S \subseteq \mathcal{Z}$. Η αποκομμένη στο S κατανομή πιθανότητας $\mathcal{D}$, που θα συμβολίζεται με $\mathcal{D}_S$, ορίζεται ως:*

$$\mathcal{D}_S\left(x\right) = \frac{\mathbf{1}_{\{x \in S\}}}{\mathcal{D}(S)} \mathcal{D}\left(x\right)\ , x \in \mathcal{Z}\,.$$

Ο ρόλος της Κανονικής κατανομής στη Θεωρία Πιθανοτήτων και τη Στατιστική είναι αδιαμφισβήτητα θεμελιώδης. Ως αποτέλεσμα βρίσκεται στο επίκεντρο της έρευνας και η συμπεριφορά της έχει αναλυθεί για πληθώρα προβλημάτων. Στο πρόβλημα που μας απασχολεί η απάντηση δόθηκε στο [DGTZ18] για την Κανονική κατανομή σε d διαστάσεις. Στο Θεώρημα 1.3.1 επαναλαμβάνουμε το αποτέλεσμα για την περίπτωση που $d = 1$.

Θεώρημα 1.3.1. *Έστω ένα μετρήσιμο σύνολο S και δίνεται πρόσβαση σε μαντείο που επιβεβαιώνει αν ένα στοιχείο ανήκει στο S. Θεωρούμε ότι μια άγνωστη Κανονική κατανομή $N(\mu, \sigma^2)$ έχει μέτο στο S που ισούται με $a = \Pr_{X \sim N(\mu,\sigma^2)}[X \in S] > 0$. Δίνονται, επίσης, δείγματα $x_1, x_2, \ldots$ από μία αποκομμένη Κανονική κατανομή $N_S(\mu, \sigma^2)$. Τότε, υπάρχει αποδοτικός αλγόριθμος που επιστρέφει εκτιμήσεις $\hat{\mu}$ και $\hat{\sigma}^2$. Για κάθε $\epsilon > 0$, ο αλγόριθμος χρησιμοποιεί $O(1/\epsilon^2)$ δείγματα και κάνει ισάριθμες ερωτήσεις στο μαντείο για δώσει τις παραπάνω εκτιμήσεις που ικανοποιούν τα παρακάτω:*

$$|\mu - \hat{\mu}| < \sigma\epsilon \quad \textit{και} \quad \left|1 - \frac{\hat{\sigma}^2}{\sigma^2}\right| < \epsilon^2$$

με πιθανότητα τουλάχιστον 99%. Επιπλέον, ισχύει ότι:

$$\mathrm{TV}\left(N(\mu, \sigma^2), N(\hat{\mu}, \hat{\sigma}^2)\right) < O(\epsilon)\,.$$

Ο αλγόριθμος που επιτυγχάνει την παραπάνω απόδοση είναι ένας από τους πιο θεμελιώδεις αλγορίθμους της Θεωρίας Μάθησης: ο αλγόριθμος Στοχαστικής Κατάβασης Κλίσης. Παρουσιάζουμε τον αλγόριθμο στο 1.

Ο αλγόριθμος Κατάβασης Κλίσης (η μη-Στοχαστική εκδοχή του) εφαρμόζεται για την εύρεση του ελάχιστου σημείου μιας συνάρτησης-στόχου που επιθυμούμε να ελαχιστοποιήσουμε. Η αποδοτική εύρεση του σημείου επιτυγχάνεται όταν η συνάρτηση-στόχος ικανοποιεί ορισμένες προϋποθέσεις. Η βασικότερη από αυτές είναι η κυρτότητα. Όταν η συνάρτηση είναι κυρτή (βλ. 4.2) η πληροφορία γα την κλίση της σε κάθε σημείο υποδεικνύει την κατεύθυνση προς την οποία βρίσκεται το ελάχιστο. Συνεπώς, περιγράψαμε τη βασική λειτουργία του αλγορίθμου: Ξεκινώντας από ένα τυχαίο σημείο υπολογίζει την κλίση της συνάρτηση-στόχου σε αυτό και κινείται προς την κατεύθυνσή του ελάχιστου σημείου.

Στην περίπτωση του αλγορίθμου Στοχαστικής Κατάβασης Κλίσης δεν έχουμε πρόσβαση στην πραγματική τιμή της Κλίσης σε κάθε σημείο. Αποδεικνύεται ότι αν γνωρίζουμε μία αμερόληπτη εκτίμηση αυτής ο αλγόριθμος θα συγκλίνει και πάλι. Αυτή την ιδιότητα εκμεταλλεύεται το [DGTZ18], όπου προσδιορίζεται ένας αμερόληπτος εκτιμητής της Κλίσης της συνάρτησης-στόχου. Ο εκτιμητής αυτός βασίζεται στην υπόθεση ότι έχουμε πρόσβαση στο σύνολο αποκοπής S. Συγκεκριμένα υποθέτουμε ότι υπάρχει ένα μαντείο όπου μπορούμε να ελέγχουμε αν ένα στοιχείο του πεδίου ορισμού ανήκει στο S. Η πληροφορία αυτή αποδεικνύεται ότι είναι αναγκαία για την σύγκλιση του αλγορίθμου σε πεπερασμένο αριθμό βημάτων.

1.4 Διωνυμική Κατανομή Poisson

Η Διωνυμική Κατανομή Poisson είναι μια από τις πιο βασικές διακριτές κατανομές. Προκύπτει φυσικά στη μοντελοποιήση πληθώρας προβλημάτων και έχει μεγάλη εκφραστική ικανότητα. Η Διωνυμική Κατανομή Poisson ταξης n ορίζεται ως το άθροισμα n ανεξάρτητων τυχαίων μεταβλητών που ακολουθούν την κατανομή Bernoulli. Θα αναφερόμαστε σε αυτή με τη συντομογραφία PBD ή PBD_n όταν η τάξη δεν είναι σαφής από τα συμφραζόμενα. Παρατηρούμε ότι η PBD αποτελεί γενίκευση της Διωνυμικής κατανομής την οποία πρώτος μελέτησε ο Poisson, εξού και η ονομασία της.

Στην Παράγραφο 1.2 παρουσιάσαμε το Κεντρικό Οριακό Θεώρημα. Σημειώσαμε, επίσης, ότι υπάρχουν εκδοχές του ΚΟΘ οι οποίες παρακάμπτουν την υπόθεση για ισόνομες τυχαίες μεταβλητές. Ακολουθεί μια σημαντική παρατήρηση. Η PBD ικανοποιεί τις προϋποθέσεις του θεωρήματος και, συνεπώς, πλησιάζει ασυμπτωτικά την κανονική κατανομή.

Σε ένα περισσότερο σύγχρονο αποτέλεσμα, στο [DP15] δίνεται ένας πλήρης χαρακτηρισμός του συνόλου των PBDs. Συγκεκριμένα αποδεικνύεται ότι μπορούμε να προσεγγίσουμε, με όση ακρίβεια θέλουμε, οποιαδήποτε PBD με μία άλλη PBD έτσι ώστε η τελευταία να έχει μια ορισμένη μορφή. Συγκεκριμένα μπορεί να είναι είτε μια Διωνυμική κατανομή με σχετικά υψηλή διασπορά είτε μια άλλη PBD με αρκετά λιγότερες παραμέτρους $l << n$. Είναι γνωστό ότι μία Διωνυμική κατανομή με υψηλή διασπορά προσεγγίζει την κανονική κατανομή, οπότε επιστρέφουμε στο ΚΟΘ.

Θεώρημα 1.4.1. *Έστω $X_1, \dots, X_n$ μια ακολουθία αμοιβαία ανεξάρτητων δεικτριών τυχαίων μεταβλητών, και $k \in \mathbb{N}$. Τότε, υπάρχει μια δεύτερη ακολουθία αμοιβαία ανεξάρτητων δεικτριών τυχαίων μεταβλητών $Y_1, \dots, Y_n$ τέτοια ώστε να ισχύουν τα ακόλουθα:*

- $\text{TV}\left(\sum_i X_i, \sum_i Y_i\right) \leq 41/k$;
- *τουλάχιστον ένα από τα παρακάτω ικανοποιείται:*
  - *(Διωνυμική μορφή) υπάρχουν $l \in \{1, \dots, n\}$ και $q \in \{\frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n}\}$ τέτοια ώστε, για κάθε $i \leq l$, $\mathbb{E}[Y_i] = q_i$ και, για κάθε $i > l$, $\mathbb{E}[Y_i] = 0\cdot$ επιπλέον, τα l και q ικανοποιούν τις σχέσεις $lq \geq k^2$ και $lq(1-q) \geq k^2 - k - 1\cdot$ ή*
  - *(k-αραιή μορφή) υπάρχουν κάποια $l \leq k^3$ τετοια ώστε, για κάθε $i \leq l$, $\mathbb{E}[Y_i] \in \{\frac{1}{k^2}, \frac{2}{k^2}, \dots, \frac{k^2-1}{k^2}\}$ και, για κάθε $i > l$, $\mathbb{E}[Y_i] \in \{0, 1\}$.*

Τότε μπορούμε να ανάγουμε το πρόβλημα της εκμάθησης μιας PBD σε δύο υποπροβλήματα: την εκμάθηση μια Διωνυμικής κατανομής και την εκμάθηση μιας αραιης PBD. Πράγματι, αυτή είναι η τακτική που ακολουθείται στο παραδοσικό πρόβλημα μάθησης, που έχουμε πρόσβαση στο σύνολο του πεδίου ορισμού. Στην επόμενη ενότητα θα προσπαθήσουμε να την εφαρμόσουμε στο πρόβλημα που μας ενδιαφέρει στην παρούσα εργασία: την εκμάθηση από ελλιπή δείγματα.

1.5 Μάθηση Διωνυμικής Κατανομής Poisson από Ελλιπή Δείγματα

Το πρώτο βήμα για τη μελέτη της δυνατότητας εκμάθησης μιας κατανομής από ελλιπή δείγματα είναι η μελέτη της αναγνωρισιμότητας της αποκομμένης κατανομής. Παρατηρείστε ότι όταν αναφερόμαστε σε μία οικογένεια κατανομών που προσδιορίζονται από έναν αριθμό παραμέτρων, για κάθε διαφορετική τιμή των παραμέτρων η κατανομές διαφέρουν, έστω και λίγο. Αυτό είναι απαραίτητο για να είναι καλά ορίσμενο το σύνολο.

Ωστόσο, εάν όταν περιορίζουμε την κατανομή σε ένα συγκεκριμένο υποσύνολο του πεδίου ορισμού της, δημιουργούμε μια νέα οικογένεια κατανομών για την οποία δεν έχουμε τέτοια εγγύηση. Εάν δύο διαφορετικές κατανομές καταλήγουν στην ίδια αποκεμμένη κατανομή είναι αδύνατο να προσδιορίσουμε την αρχική. Αυτό ισχύει ακόμα κι αν γνωρίζουμε επακριβώς την αποκομμένη κατανομή, πόσο μάλλον όταν έχουμε απλώς πρόσβαση σε δείγματά της.

Με βάση την παραπάνω παρατήρηση, εξετάζουμε υπό ποιες προϋποθέσεις για το S το σύνολο των PBDs είναι αναγνωρίσιμο από την αποκομμένη εκδοχή του. Παίρνουμε ένα αρνητικό αποτέλεσμα. Δίνουμε δύο διαφορετικές PBDs, με TV απόσταση τουλάχιστον $1/2$ που αν κρύψουμε μόλις ένα στοιχείο του πεδίου ορισμού τους, ταυτίζονται. Έτσι, στη γενική περίπτωση, είναι αδύνατο να προσδιοριστεί αλγόριθμος που, για κάθε S και για κάθε PBD, να εγγυάται ότι το αποτέλεσμά του προσεγγίζει την πραγματική κατανομή με καλή πιθανότητα. Στο σχήμα 2.1 μπορούμε να δούμε αυτο το ζεύγος κατανομών.

Ωστόσο το Θεωρήμα 1.4.1 αποκαλύπτει ότι ένα μεγάλο υποσύνολο των PBDs μπορεί να προσεγγιστεί από Ќανονικές΄ κατανομές. Όμως γνωρίζουμε, από το θεώρημα 1.3.1 ότι η Κανονική κατανομή μπορεί να βρεθεί αποδοτικά από ελλιπή δείγματα. Μάλιστα, γνωρίζουμε τον ακριβή χαρακτηρισμό του συνόλου αποκοπής S σε αυτή την περίπτωση. Παρατηρούμε ότι όσο πιο κοντά είμαστε σε μια συνεχή κατανομή, η μάθηση από ελλιπή δείγματα γίνεται εφικτή. Η πρόθεσή μας είναι να χρησιμοποιήσουμε τον αλγόριθμο του [DGTZ18] για να βρούμε τη μέση τιμή και την διασπορά της PBD. Έτσι θα μπορέσουμε να προσδιορίσουμε τις παραμέτρους της Διωνυμικής κατανομής που σύμφωνα με το Θεώρημα 1.4.1 την προσεγγίζει.

Παρόλα αυτά, η προσαρμογή του αποτελέσματος 1.3.1 για την περίπτωση των PBDs είναι απαιτητική. Κατ΄ αρχάς, παρατηρούμε ότι η PBD είναι μια διακριτή κατανομή. Αυτό σημαίνει ότι η μάζα της Κανονικής κατανομής σε οποιοδήποτε σύνολο αποκοπής S για την PBD θα είναι μηδενική. Συνεπώς, το θεώρημα δεν προσφέρει κάποια εγγύηση για την αποδοτική υλοποιήση του αλγορίθμου.

Ένα δεύτερο πρόβλημα είναι ότι τα δείγματα δεν προέρχονται πράγματι από μια Κανονική κατανομή. Ωστόσο, η συνάρτηση-στόχος που θέλουμε να ελαχιστοποιήσουμε υποθέτει δείγματα από μια Κανονική κατανομή. Έτσι, η εκτίμηση της κλίσης της συνάρτησης-στόχου δεν είναι αμερόληπτη. Τότε ο αγόριθμος δεν έχει καμία εγγύηση σύγκλισης. Το δεύτερο αυτό πρόβλημα μπορεί να ξεπεραστεί εύκολα δεδομένου ότι το σφάλμα στην εκτίμηση θα είναι μικρό.

Τελικά, περιοριζόμαστε στην πειραματική επιβεβαίωση της εν λόγω τεχνικής. Τα αποτελέσματα παρουσιάζονται στα σχήματα 6.1, 6.2.

1.6 Μάθηση Κατανομής Mallows από Ελλιπή Δείγματα

Η κατανομή Mallows αναφέρεται σε ένα πιθανοτικό μοντέλο που ορίζεται για το σύνολο των διατάξεων-μεταθέσεων $\mathbb{S}_m$. Σύμφωνα με αυτό, έχουμε ένα σύνολο αντικειμένων A και δεχόμαστε ότι υπάρχει μία πραγματική διάταξη, $\pi_0 \in \mathbb{S}_m$. Τότε κάθε π που προκύπτει από το μοντέλο είναι θορυβώδης εκδοχή αυτής της πραγματικής διάταξης. Ο θόρυβος κάθε διάταξης-δείγματος παράγεται σύμφωνα με την εξής διαδικασία: για κάθε ζεύγος αντικειμένων $a, b \in A$ για το οποίο $a \succ_{\pi_0} b$ η πιθανότητα η σειρά να διατηρηθεί στο δείγμα, δηλαδή $a \succ_\pi b$, είναι ανάλογη του ϕ και μεγαλύτερη από $1/2$.

Έτσι η πιθανότητα μιας διάταξης στο μοντέλο Mallows μειώνεται με την απόστασή της από την πραγματική διάταξη. Ως απόσταση μεταξύ διατάξεων θεωρούμε την απόσταση Kendall-Tau. Σύμφωνα με αυτήν, για $\pi, \sigma \in \mathbb{S}_m$,

$$\mathrm{KT}(\pi, \sigma) = |\{\{a, b\} \subseteq A : a \succ_\pi b \quad and \quad b \succ_\sigma a\}| \,.$$

Τελικά προκύπτει ο ακόλουθος ορισμός.

Ορισμός 1.6.1 (Κατανομή Mallows)**.** *Έστω $\phi \in (0,1)$ και $\pi_0 \in \mathbb{S}_m$. Συμβολίζουμε με $\mathcal{M}(\pi_0, \phi)$ την κατανομή Mallows σύμφωνα με την οποία κάθε διάταξη $\pi \in \mathbb{S}_m$ εμφανίζεται με πιθανότητα*

$$p(\pi) = \frac{1}{Z(\phi)} \phi^{\mathrm{KT}(\pi, \pi_0)} \,,$$

όπου $Z(\phi)$ η σταθερά κανονικοποίησης.

Θα μελετήσουμε τις συνθήκες που πρέπει να ισχύουν για το σύνολο αποκοπής S ώστε η πραγματική διάταξη π_0 να μπορεί να βρεθεί με πολυωνυμικό πλήθος δειγμάτων, σε πολυωνυμικό χρόνο. Στην περίπτωση που τα δείγματα είναι πλήρη, δηλαδή $S = \mathbb{S}_m$, η απάντηση δίνεται στο [CPS13]. Το ακόλουθο θεώρημα επαναλαμβάνει αυτό το αποτέλεσμα.

Θεώρημα 1.6.1. *Για κάθε $\delta > 0$, υπάρχει πολυωνυμικός αλγόριθμος που βρίσκει την πραγματική διάταξη π_0 του μοντέλου Mallows με πιθανότητα τουλάχιστον $1 - \delta$, χρησιμοποιώντας $O(\log(m/\delta))$ δείγματα από την $\mathcal{M}(\pi_0, \phi)$.*

Η απόδειξη του παραπάνω θεωρήματος υποδεικνύει μία συνθήκη για τη $\mathcal{M}_S(\pi_0, \phi)$ ώστε δείγματα από την αποκομμένη κατανομή να αρκούν για να μάθουμε το π_0. Δηλαδή, έχουμε μια ικανή συνθήκη για μάθηση από ελλιπή δείγματα.

Ορισμός 1.6.2. *Θα λέμε ότι η $\mathcal{M}_S(\pi_0, \phi)$ είναι συνεπής αν για κάθε ζεύγος αντικειμένων $a, b \in A$ τέτοιο ώστε $a \succ_{\pi_0} b$ ισχύει ότι:*

$$p^S_{a \succ b} > p^S_{b \succ a} \,,$$

όπου $p^S_{a \succ b} = \sum_{a \succ_\pi b} \mathcal{M}_S(\pi)$ είναι η πιθανότητα η αποκομμένη κατανομή να διατάξει το a πάνω από το b.

Τότε, εντελώς ανάλογα με το Θεώρημα 1.6.1, αποδεικνύεται το ακόλουθο θεώρημα.

Θεώρημα 1.6.2. *Έστω η αποκομμένη κατανομή Mallows $\mathcal{M}_S(\pi_0, \phi)$. Υποθέτουμε ότι η $\mathcal{M}_S(\pi_0, \phi)$ είναι συνεπής. Επίσης, έστω ότι $\delta_{min} = min_{a \succ_{\pi_0} b}(p^S_{a \succ b} - p^S_{b \succ a})$. Τότε, για κάθε $\delta > 0$, υπάρχει πολυωνυμικός αλγόριθμος που βρίσκει την πραγματική διάταξη π_0 με πιθανότητα τουλάχιστον $1 - \delta$, χρησιμοποιώντας $O(\log(m/\delta)/\delta^2_{min})$ δείγματα από την $\mathcal{M}_S(\pi_0, \phi)$.*

Στη συνέχεια ορίζουμε μία συνθήκη που εγγυάται την αναγνωρισιμότητα της αποκομμένης κατανομής $\mathcal{M}_S(\pi_0, \phi)$.

Θεώρημα 1.6.3. *Η $\mathcal{M}_S(\pi_0, \phi)$ όπου το ϕ είναι γνωστό δεν είναι αναγνωρίσιμη αν και μόνο αν υπάρχει $\pi_1 \in \mathbb{S}_m$ τέτοιο ώστε, για κάθε $\pi_i \in S$, να ισχύει ότι:*

$$\mathrm{KT}(\pi_i, \pi_0) - \mathrm{KT}(\pi_i, \pi_1) = c\,,$$

όπου c σταθερά.

Συνεπώς, το Θεώρημα 1.6.3 υποδεικνύει μία αναγκαία συνθήκη για μάθηση της πραγματικής διάταξης από ελλιπή δείγματα. Ωστόσο, ο στόχος αυτής της εργασίας είναι να βρει μία συνθήκη που να *χαρακτηρίζει* την ικανότητα μάθησης. Δηλαδή, μία ικανή και αναγκαία συνθήκη.

Αρχικά, τονίζουμε ότι οι δύο παραπάνω συνθήκες δεν είναι ισοδύναμες. Δίνουμε ένα παράδειγμα μιας Mallows κατανομής και ένος συνόλου αποκοπής S ώστε η $\mathcal{M}_S(\pi_0, \phi)$ να είναι πάντα αναγνωρίσιμη αλλά να απαιτούνται εκθετικά πολλά δείγματα για να τη αναγνωρίσουμε.

Έστω $\pi_0 = (1, 2, 3)$ και οποιοδήποτε $\phi = 0.3$. Έστω $S = \{(2, 1, 3), (3, 1, 2), (3, 2, 1)\}$. Μπορούμε να επαληθεύσουμε με κώδικα (ή με εξαντλητική αναζήτηση με το χέρι) ότι για τις παραπάνω τιμές η $\mathcal{M}_S(\pi_0, \phi)$ είναι πάντα αναγνωρίσιμη. Θεωρούμε τώρα $S' = \{(2, 1, 3), (3, 1, 2)\}$. Με ανάλογες τεχνικές προκύπτει ότι η $\mathcal{M}_{S'}(\pi_0, \phi)$ δεν είναι αναγνωρίσιμη, άρα το π_0 είναι αδύνατο να βρεθεί από δείγματα της. Ωστόσο, τα δύο σύνολα διαφέρουν μόνο στο στοιχείο $(3, 2, 1)$, το οποίο χρειάζεται εκθετικά πολλά δείγματα (ως προς το m) για να προκύψει ως δείγμα. Αυτό είναι άμεσο αν παρατηρήσουμε ότι $\mathrm{KT}(\pi_0, (3, 2, 1)) = \binom{m}{2}$ και από τον ορίσμό της κατανομής Mallows.

Τελικά δεν έχουμε προσδιορίσει με ασφάλεια την ικανή και αναγκαία συνθήκη για το σύνολο αποκοπής S. Κάνουμε όμως μία σημαντική παρατήρηση. Ο αλγόριθμος που δόθηκε στο [CPS13] βρίσκει το π_0 χρησιμοποιώντας αποκλειστηκά πληροφορία για την ανά ζεύγη διάταξη των αντικειμένων, αγνοώντας την ολική διάταξη που προσφέρουν τα δείγματα του Mallows. Φαίνεται ότι, αν και αυτή η πληροφορία αρκεί στην απλή περίπτωση, όταν τα δείγματα είναι ελλιπή, δεν δίνει η βέλτιστη προσέγγιση για την επίλυση του προβλήματος.

Τέλος, σε μια προσπάθεια να κατανοήσουμε καλύτερα την απαίτηση για *συνέπεια* της $\mathcal{M}_S(\pi_0, \phi)$, μελετάμε κάποια συνήθη σύνολα αποκοπής. Αυτά είναι ένα ομοιόμορφα τυχαίο S και ένα S όπου αφαιρούμε μόνο ένα στοιχείο. Βλέπουμε ότι όσο το ϕ είναι πιο κοντά στο 1 η κατανομή Mallows είναι πολύ ευαίσθητη στο να γίνει *ασυνεπής*.

Θεώρημα 1.6.4. *Έστω $S \sim_u \mathcal{P}(\mathbb{S}_m)$. Τότε, η αποκομμένη κατανομή Mallows $\mathcal{M}_S(\pi_0, \phi)$, όπου $\phi \in (0, 1 - \sqrt{16 \log(\frac{m}{\delta})/m!})$, είναι συνεπής με πιθανότητα τουλάχιστον $1 - \delta$.*

Θεώρημα 1.6.5. *Έστω η κατανομή $\mathcal{M}(\pi_0, \phi)$ και έστω $|S| = |\mathbb{S}_m| - 1$. Τότε η $\mathcal{M}_S(\pi_0, \phi)$ είναι συνεπής αν $\frac{1-\phi}{1+\phi} Z_\phi > 1$, όπου Z_ϕ η σταθερά κανονικοποιήσης της κατανομής Mallows.*

Chapter 2

Introduction

Automated Learning undoubtedly lies among the ’hottest’ topics of modern science. For the first time, instead of making rules and enforcing them to reality, people allow reality to become the rule. The power of inference offers solutions to intricate problems up to recognising emotions [VdMH08]. Learning Theory provides the mathematical foundation that validates these techniques and keeps them in line with the scientific method [Vap99], [Val84].

In this context, Distribution Learning, introduced in [KMR$^+$94], consists a fundamental field. In Statistical Inference it has been a long practice to model phenomena through probability distributions and rely on data to determine them. Automating these methods for machines to perform them efficiently has opened a new world of study, given rise to further challenges and broaden our understanding about the limits of our abilities.

One of the most imposing of these challenges is Learning from Incomplete Data. This is what we will call Truncated Samples. Needless to say, this problem enjoys an innate transition from possible to impossible. In the trivial case that none or only one sample is provided, for example, the impossibility of a solution is immediate. On the opposite side where no truncation is performed we fall back to the traditional setting. Thus, Distribution Learning from Truncated Samples tackles with an important real-world problem, trading the extend of realizable. The trade-off is controlled by the characteristics of the truncation set S. This thesis aspires to specify those conditions on S that *determine* (iff-conditions) learnability from truncated samples.

The above framework can be considered a specific case of Robust Statistics [Hub65], [Hub92]. This area of study is concerned with developing algorithms that are not just *efficient* but also *robust* to *minor* violations of their assumptions. Such properties are vital when constructing practical applications, prominent to adversarial or simply ignorant users. Robust Distribution Learning ([DKK$^+$19]) studies algorithms that are resilient to a small, ε-corruption of their sample data. It allows any kind of corruption, including replacing, adding or removing samples. Since an algorithm has traditionally no way to retrieve the original samples, but may just recognise the adversarial ones, a resulting ε-error on its output is unavoidable [JO19]. In the case of Truncation, however, where only removing data is allowed, one is able to distinguish the level of data corruption (say α) from the quality of result (say ε).

Despite its apparent significance it was only recently that even the elementary problem of learning a multivariate Gaussian distribution from truncated samples was settled

[DGTZ18]. Following this, a series of works considers the problem of truncated samples in the fields of continuous distributions [KTZ19], [DKTZ21], discrete distributions [FKT20], mixtures of distributions [NP20] and regression [DGTZ19], [IZD20], [DSYZ21]. In [FKT20] an adaptation of the techniques for learning a multivariate Gaussian (given in [DGTZ18]) is made for a discrete distribution, the Boolean Product Distribution. This method reveals an inherent sensitivity of discrete distributions to truncation. It is proven that additional assumptions on the truncation set S are necessary to guarantee efficient learnability of the Boolean Product Distribution. We study the learnability of two discrete distributions from truncated samples.

This work further investigates the above result. We examine two discrete, parametric distributions and analyze the conditions on S that secure their learnability.

Setting and Contribution

- *Poisson Binomial Distribution* Consider the distribution of the sum of n independent indicators with potentially different success probabilities p_i i.e.

  $$X = \sum_{i=1}^{n} X_i \, , \, X_i \sim \mathrm{Be}(p_i) \, .$$

  Then X follows the Poisson Binomial Distribution of order n and parameters p_i i.e. $X \sim \mathrm{PBD}([p_i]_{i=1}^{n})$.

  First, we provide an example where the truncation set S hides just one element of the distribution's support. We show there are two different PBDs, with TV distance $1/2$, with the exact same truncation on that set, making the problem non-identifiable. This example can be seen in fig. 2.1[1]. That is, in the general case, learning a PBD from truncated samples is impossible.

  We proceed with an interesting observation about the nature of Poisson Binomial Distributions. The study of sums of independent random variables has interest the scientific world for decades and goes back to the well-known Central Limit Theorem. In [DP15] provide a full characterization for the structure of such a sum, the PBD. In consequence to the CLT-like results, the class of PBDs consists of two kinds of distributions: those close to a Gaussian -which are characterized as heavy Binomials- and those away from it. In other words, those resembling a continuous distribution and those not. We aim to exploit this structure to highlight our claim: learning from truncated samples becomes challenging as we depart from continuity.

  In the sparse case, the previous impossibility result still holds. This follows immediately, since the distributions in our example have constant TV distance from any heavy Binomial distribution.

  In the close-to-Gaussian case, we apply the Stochastic Gradient Descend, as in [DGTZ18], to retrieve the mean value and variance of the PBD. Then, the proximity of a heavy PBD to a heavy Binomial distribution implies that the Binomial with the same mean and variance as the algorithm suggests must be a good solution to our problem. Recalling their main result we have theorem 2.0.1.

[1] The code for all the figures (except cited otherwise) can be found here.

Theorem 2.0.1 (Theorem 1 [DGTZ18]). *Given oracle access to a measurable set S, whose measure under an unknown normal distribution $N(\mu, \sigma^2)$ is $a = \Pr_{X \sim N(\mu,\sigma^2)}[X \in S] > 0$, and samples $x_1, x_2, \ldots$ from the truncated normal distribution $N_S(\mu, \sigma^2)$, there exists a polynomial time algorithm that returns estimates $\hat{\mu}$ and $\hat{\sigma}^2$. For all $\epsilon > 0$, the algorithm uses $O(1/\epsilon^2)$ samples and queries to the oracle and produces estimates that satisfy*

$$|\mu - \hat{\mu}| < \sigma\epsilon \quad \textit{and} \quad |1 - \frac{\hat{\sigma}^2}{\sigma^2}| < \epsilon^2$$

with probability at least 99%. *Moreover, it holds that:*

$$\mathrm{TV}(N(\mu, \sigma^2), N(\hat{\mu}, \hat{\sigma}^2)) < O(\epsilon)\,.$$

Notice there are two basic challenges for adapting the above theorem to the case of heavy PBDs. First, the PBD is a discrete distribution. So the mass of the truncation set α is always zero, implying an infinite number of samples and iterations for the PSGD to converge. Second, the estimators for the sub-gradient of the PSGD are biased. This is because the objective to be minimized assumes an underlying Gaussian distribution. However the samples come from a PBD, close to it. Since the error in the estimation is small, we can prove that PSGD is robust to it. Still the challenge of discretization is hard to overcome and we leave a formal proof of the theorem in the case samples come from a heavy PBD open for future work.

For now, we suffice to validate our intuition through experiments. Using samples from sufficiently *heavy* PBD we run the SGD for the Gaussian and retrieve an ε-close to the PBD result.

- *Mallows Distribution* The Mallows Model is a statistical model defined over the group of permutations $\mathbb{S}_m$. It has two parameters: the central ranking $\pi_0 \in \mathbb{S}_m$ and the dispersion $\phi \in (0, 1)$. Every ranking is given probability inversely proportional to its distance from π_0, where the Kendall-Tau distance is used. In this work we consider the parameter ϕ known. Note that we have now substantially moved away from continuous distributions.

  In the non-truncated case [CPS13] provide an algorithm for retrieving π_0 efficiently (in terms of Mallows samples and runtime). This algorithm is based on the pair-wise comparisons between the alternatives in a ranking. It decisively relies on a property of the Mallows distribution: in a Mallows sample any pair of alternatives is ranked *consistently* (i.e. the same way) to the central ranking with probability at least 1/2. We refer to this property as *consistency.* Based on it, we gain a sufficient condition on S for learning π_0 from truncated Mallows samples.

Informal Theorem 2.0.1. *Let $\mathcal{M}_S(\pi_0, \phi)$ be the truncated on S Mallows distribution. If S does not violate the consistency of the distribution, there is an efficient algorithm that retrieves π_0 given samples from $\mathcal{M}_S(\pi_0, \phi)$.*

In search for a more solid characterization of this property on S we prove a connection to the parameter ϕ. Whenever ϕ is away from 1 (that is, the Mallows distribution is not close to uniform), *consistency* is preserved even if S hides a large number of rankings.

What is more, the specific algorithm of [CPS13] does not work without this property. This is an implication that the above condition on S must be necessary as well. Exploring the necessity of *consistency* we point out a condition that fully characterizes identifiability. That means, this is a necessary condition for retrieving π_0.

Informal Theorem 2.0.2. *$\mathcal{M}_S(\pi_0, \phi)$ is non-identifiable iff there exists π_1 such that $\mathrm{KT}(\pi_0, \pi_i) - \mathrm{KT}(\pi_1, \pi_i) = c$, where $\mathrm{KT}(.)$ is the Kendall-Tau distance and c constant.*

However, these two results are not equivalent which is illustrated through two examples. First, we give an identifiable truncated Mallows $\mathcal{M}_S(\pi_0, \phi)$ that demands exponentially many samples to be distinguished from an non-identifiable one. In other words, this is an identifiable distribution whose parameter π_0 cannot be learnt efficiently. Moreover, we find an in-*consistent* Mallows that is not identifiable.

Putting it all together, we must conclude that the algorithm in [CPS13] is not optimal in the case of truncation. It seems that there exists another algorithm that is more robust against truncation. It could achieve this if it is not solely based on the information about the pair-wise comparisons a Mallows ranking provides. This conclusion is supported by recent results for the Mallows distribution such as [LM21].

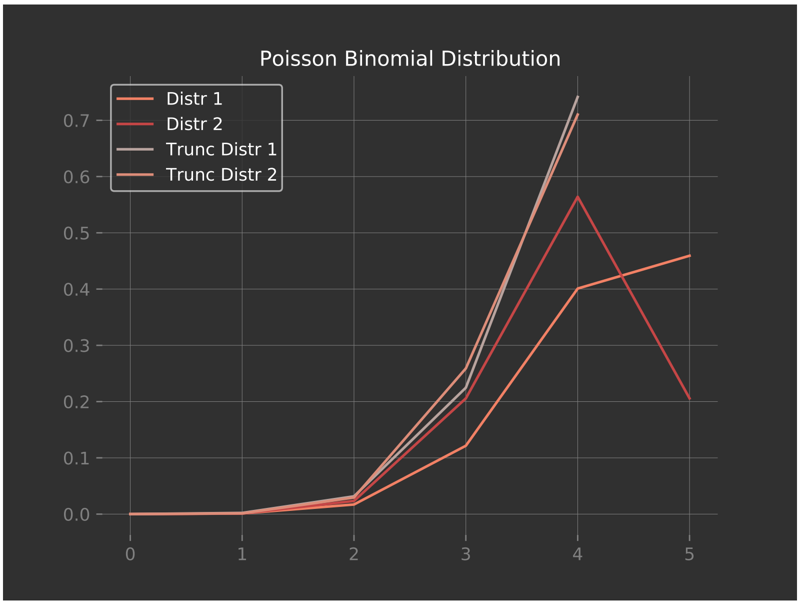

Figure 2.1: Non-identifiable from their truncation PBDs

Organization

This work intends to provide a thorough investigation of the causes behind the success or failure of a distribution learning algorithm. With this in mind Chapter 1 presents the basic probabilistic notions that, assuming they describe phenomena, are responsible for their 'good' properties and enable prediction. In the next Chapter we describe the learning paradigm that provides the goal for our analysis. Moving forward we deal with our actual problem. Chapters 3 and 4 introduce the Poisson Binomial Distribution and study its learnability from truncated samples respectively. Finally, Chapter 5 examines the case Mallows distribution.

Chapter 3

Probability Theory

This chapter covers the probabilistic background necessary to understand the problem this thesis is concerned with. At the same time, it is hoped the results presented here will extend the knowledge and perspective of the reader about statistics and their power to model real world phenomena. We recall some basic concepts from probability theory. Using these we build on some powerful tools, distribution concentration and distribution distance. These are essential for the formalization of our objective and the proof of our results.

Probability theory can be considered an application of measure theory. That is, a probability space is a measure space whose measure function satisfies some extra properties. Though very elegant, this work does not adopt this approach. In what follows, we wish to simply recall the basic notions in the field so as to introduce two new techniques for handling probabilistic models: concentration and distance. The first one allows us to control the probability of bad events happening. Thus is important in establishing the truth of our theorems. The second defines the properties these theorems guarantee. This thesis' main concern is to construct tools/algorithms that calculate 'close' distributions.

3.1 Basic Concepts

The very first understanding of probability must come from a non-deterministic real life experiment. Indeed, in every day life, one encounters processes with unpredictable outcomes, such as the flipping of a coin, tossing a die, etc. However, some outcomes occur more often than others. This tempts many people to guess on them and, surprisingly, be correct most of the time! Studying such phenomena and trying to give an explanation for this luck is the main concern of probability theory.

Probability Measure

The concept we wish to formalize is as follows: Consider an experiment with a set of possible outcomes. The aim is to quantify the *probability* that the next execution's result belongs to a subset of them. For example, when tossing a die, the possible results are $\{1, 2, 3, 4, 5, 6\}$. The bet, though, might be to get a number ≥ 5. Thus one is interested in the probability of a subset of the outcomes $\{5, 6\}$.

Let Ω denote this set of outcomes which are called elementary events. Then Ω is the space of elementary events. Consider $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ the subsets of Ω whose probability we will quantify. These are called events. The set $\mathcal{A}$ need not include all the possible subsets of Ω but it must have the following properties:

- $\Omega \in \mathcal{A}$.
- If $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$.
- If $A_n \in \mathcal{A}, n = 1, 2, \ldots$, then $\cup_{n=1}^{\infty} A_n \in \mathcal{A}$.

Such a collection of subsets $\mathcal{A}$ is called a σ-algebra. The space $(\Omega, \mathcal{A})$ is a measurable space. The structure of such a space allows to define a *measure* on it. That is, a function to quantify the sets in the space. This abstract description of a measure space should suffice for us here. Then the probability of an event $A \in \mathcal{A}$ is given by a probability measure Pr on $(\Omega, \mathcal{A})$. Denote $\Pr[A]$ the probability of the event. If A is an elementary event, i.e. $A = \{\omega\}, \omega \in \Omega$ we write $\Pr[\{\omega\}] = \Pr[\omega]$. Proceeding to a formal definition we have:

Definition 3.1.1 (Probability Measure). *Consider a set Ω and $\mathcal{A}$ a σ-algebra on this set. The function $\Pr : \mathcal{A} -> [0, 1]$ is a probability measure if:*

- $\Pr[A] \geq 0, \forall A \in \mathcal{A}$.
- $\Pr[\Omega] = 1$.
- $\Pr[\cup_{n=1}^{\infty} A_n] = \sum_{n=1}^{\infty} \Pr[A_n], \forall A_i, A_j$ *such that* $A_i \cap A_j = \emptyset, i \neq j$, $i, j = 1, 2, \ldots$, *i.e. for all disjoint events.*

The triplet $(\Omega, \mathcal{A}, \Pr)$ is called a probability space.

We should define two elementary concepts for a probability measure Pr, independence and conditioning.

Definition 3.1.2 (Independence). *Two events $A, B \in \mathcal{A}$ are independent if*

$$\Pr\left[A \cap B\right] = \Pr\left[A\right] \cdot \Pr\left[B\right] \ .$$

This means that the probability of A happening does not affect B and vice versa. Independence is a common assumption throughout this study. Note that it extends to more than two events in a natural way, i.e.

$$\Pr\left[\bigcap_i A_i\right] = \prod_i \Pr\left[A_i\right] \ ,$$

for any sequence of independent events $(A_i)_i$.

In contrast, the next notion formalizes the non-independence of two events A, B.

Definition 3.1.3 (Conditional Probability). *Let $A, B \in \mathcal{A}$ be two events. The probability of A conditioned on B, denoted as $\Pr[A|B]$, is defined as:*

$$\Pr\left[A|B\right] = \frac{\Pr\left[A \cap B\right]}{\Pr\left[B\right]}, \Pr\left[B\right] > 0 \, .$$

Conditioning gives the probability of an event A given that another event B occurs. This is different than than probability of A, as B might imply A, thus increasing its probability. If A, B are independent $\Pr[A|B] = \Pr[A]$.

Random Variable

We are now ready to define the concept of a random variable. This is simply a function $X : \Omega \to \mathbb{R}$. Apart from the fact that it is possible, let us elaborate on why to work with X. Introducing a random variable changes our definition of events. We no longer care about the actual result of an experiment but rather for the value it induces for X. Such a mapping often allows for flexibility since Ω can be a very experiment-specific set. Moreover, studying different experiments through 'equivalent' mappings offers a more general understanding about phenomena.

Thus we want to formalize the probability that the random variable X takes value in a set $B \subseteq \mathbb{R}$ regardless of which actual event in Ω causes that value. Since

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} = A \in \mathcal{A}\,,$$

the probability in question is actually the probability of A. Real analysis guarantees this property for all open subsets of $\mathbb{R}$, B, and continuous functions X. In effect, we have a new probability measure with respect to X.

Definition 3.1.4 (Probability Distribution)**.** *Consider the probability space* $(\Omega, \mathcal{A}, \Pr)$*. Let* $X : \Omega \to \mathbb{R}_X \subseteq \mathbb{R}$ *be a random variable and* $\mathcal{B}$ *a* σ*-algebra on* $\mathbb{R}_X$*. The probability distribution of* X*,* $\Pr \circ X^{-1}$*, is defined as:*

$$\Pr \circ X^{-1}(B) = \Pr[X \in B] = \Pr\left[X^{-1}(B)\right], \forall B \in \mathcal{B}\,.$$

The triplet $(\mathbb{R}_X, \mathcal{B}, \Pr \circ X^{-1})$ *is called the probability space of the random variable* X*.*

As becomes clear, there is no need for Ω nor even X anymore. A probability distribution can be considered directly as a probability measure on $(I, \mathcal{B}_I)$, where $I \subseteq \mathbb{R}$ and $\mathcal{B}_I$ a σ-algebra on I. Then there must be a mapping X that induces such a distribution for $(\Omega, \mathcal{A})$. The latter will be the space of the specific experiment we wish to formalize through a distribution.

What is more, the notion of a random variable can be extended. Consider any measurable space $(S, \mathcal{S})$ and a function $X : \Omega \to S$. Then X is called a random element. Now the probability distribution $\mathcal{D}$ is a probability measure on $(S, \mathcal{S})$. For instance, the Mallows distribution, presented later on, is a distribution over the set of permutations $\mathbb{S}_m$. In the rest of this work, we refer to probability distributions defined over some domain without mentioning the underlying space Ω or mapping X. We use $\mathcal{D}$ to denote a probability distribution and write $X \sim \mathcal{D}$ for a random variable distributed according to $\mathcal{D}$.

We will now define the common tools for manipulating probability distributions over a subset of the reals $I \subseteq \mathbb{R}$.

Definition 3.1.5 (c.d.f.)**.** *Consider* $(I, \mathcal{B}_I, \mathcal{D})$ *a probability space and* $X \sim \mathcal{D}$*. The function*

$$F(x) = \Pr[X \leq x]\,, x \in I\,,$$

is called the cumulative distribution function (c.d.f.).

Usually and almost always in this thesis, we will be interested in probability distributions that are specified through a parameterized closed form c.d.f. Then the probability distribution will be given in the form $F(x; \vec{\theta})$, where $\vec{\theta}$ is the vector of the parameters.

Depending on their domain I, probability distributions are distinguished in discrete and continuous. If I is finite or countable $\mathcal{D}$ is a discrete distribution. For a discrete $\mathcal{D}$ we define the probability mass function (p.m.f.) p as follows:

$$p\left(x\right) = \Pr_{X \sim \mathcal{D}}\left[X = x\right] = F\left(x\right) - F\left(x-1\right)\,, x \in I\,.$$

If the c.d.f. is a continuous function the distribution is continuous. The notion of p.m.f. is meaningless for a continuous distribution since $\Pr_{X \sim D}[X = x] = 0$. However we can define the probability density function (p.d.f.) p as:

$$p\left(x\right) = \frac{dF\left(x\right)}{dx}\,.$$

Again, in case $\mathcal{D}$ is parameterized by $\vec{\theta}$ we write $p(x) = p(x; \vec{\theta})$.

Probability Distributions

In the following we define a number of fundamental probability distributions that will be used extensively in this thesis.

- *Bernoulli* Consider a random variable X that takes values in $\{0, 1\}$ such that $p(0) = 1 - p$ and $p(1) = p$. Then we write $X \sim \mathrm{Be}(p)$ and X is follows the Bernoulli distribution. Consider an event A that happens with probability p. E.g. when flipping a coin, the event of seeing tails happens with some probability p depending on the bias of the coin. A random variable X that equals 1 whenever A happens and 0 otherwise follows a Bernoulli distribution. As a result X is also called indicator random variable. We note that a random variable that indicates an event A is denoted $\mathbf{1}_A$.

- *Binomial* Consider a set of n independent random variables $X_i \sim \mathrm{Be}(p), 1 \leq i \leq n$. Their sum $X \sum_{i=1}^{n} X_i$ takes values in $\{0, \ldots, n\}$ and denotes how many times an event happened in n independent trials. Then X is said to follow the Binomial distribution with parameters n, p, i.e. $X \sim \mathrm{Bin}(n, p)$. The p.m.f. of the distribution is equals:

  $$p\left(k\right) = \binom{n}{k} p^k (1-p)^{n-k}\,,$$

  where $0 \leq k \leq n$.

- *Poisson* A random variable X taking values in $\{0, 1, \ldots, \}$ follows the Poisson distribution with parameter $\lambda > 0$, $X \sim \mathit{Poisson}(\lambda)$, if

  $$p\left(k\right) = e^{-\lambda} \frac{\lambda^k}{k!}\,,$$

  where $k \in \{0, 1, \ldots, \}$. This distribution is used to express how many times an event happens in a certain interval of time (or space). Note that the Binomial distribution also counts the number of occurrences in a discrete interval while the Poisson refers to a continuous amount of 'trials'.

This observation can be formalised. Let $X_n \sim \mathrm{Bin}(n,p)$. If $p \to 0$ and $np \to \lambda$ while $n \to \infty$ it holds that:

$$\lim_{n\to\infty} \Pr\left[X_n = k\right] = e^{-\lambda}\frac{\lambda^k}{k!}\,.$$

This result immediately follows from Stirling's type.

- *Normal* A random variable X over $\mathbb{R}$ follows the Normal distribution with parameters μ, σ, i.e. $X \sim N(\mu, \sigma^2)$, if

  $$p\left(x\right) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\,,$$

  where $x \in \mathbb{R}$. The distribution is also referred to as Gaussian distribution. We will use the two names interchangeably throughout the text. When $\mu = 0$ and $\sigma = 1$ the distribution $N(0,1)$ is called standard Normal. Then the c.d.f. and p.d.f. of the distribution are denoted Φ and ϕ respectively. Thus

  $$\Phi\left(x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \quad \text{and} \quad \phi\left(x\right) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,.$$

  Although the c.d.f. of the Normal distribution does not have a closed form the values of Φ have been calculated and exist on Tables. So the c.d.f. X of any $X \sim N(\mu, \sigma^2)$ on x can be found through the transformation

  $$F\left(x\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)\,.$$

- *Exponential Families* We will now introduce a category of probability distributions called exponential family. The distributions in this category have a common expression and share some useful properties.

Definition 3.1.6 (Exponential Family [Xia19]). *A set of distributions $p_{\vec{\theta}}$ over $\mathbb{R}$ that are parameterized by $\vec{\theta}$ is called an exponential family, if their p.d.f. can be written as:*

$$p(x;\vec{\theta}) = \exp\left(\eta(\vec{\theta})^T T(x) - A(\vec{\theta}) + B(x)\right), \forall x \in \mathbb{R}\,,$$

where $A(\vec{\theta})$ and $B(x)$ are scalars, $\eta(\vec{\theta})$ is a row vector and $T(x)$ is a column vector that represents the sufficient statistics of x.

The above definition allows us to define a probability distribution by specifying three functions: η, T, B. One can derive $A(\vec{\theta})$ by the normalization condition. Then we get that:

$$A\left(\vec{\theta}\right) = \log\int_x \exp\left(\eta(\vec{\theta})^T T(x) + B(x)\right)dx\,.$$

As a result, A is called the *log*-normalizer and the distribution is well defined when $A(\vec{\theta}) < \infty$.

The apparent elegance of the above expression can be attributed to the separation between the parameters and the domain of the distribution. Notice there is a linear relation between the parameters of the distribution η and the function $T(x)$. Thus,

Distribution	Parameters θ	Natural Parameters η	Expression	Exponential Expression
Binomial(n)/ Bernoulli($n=1$)	p	$\log \frac{p}{1-p}$	$\binom{n}{x} p^x (1-p)^{n-x}$	$\binom{n}{x} e^{(\eta x - n \log(1+e^{\eta}))}$
Poisson	λ	$\log \lambda$	$e^{-\lambda} \frac{\lambda^x}{x!}$	$\frac{1}{x!} e^{(\eta x - e^{\eta})}$
Normal	$[\mu, \sigma^2]$	$[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}]$	$\frac{1}{\sigma\sqrt{2\pi}} e^{(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})}$	$\frac{1}{\sqrt{2\pi}} e^{(\eta \cdot [x, x^2] + \frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log(-2\eta_2))}$

Table 3.1

given the mass of the distribution on $d = |\vec{\theta}|$ points, we can derive the parameters solving a linear system. That means we retrieve the distribution that induces this mass on the points. This is a very important property for inference, parameter estimation and learning as will become clear in the chapters that follow.

Note that the aforementioned distributions, except for the Binomial, belong to the exponential family. The Binomial distribution, given the parameter n, also does. Expressing them in the above form often requires a transformation of their parameters. This is given by η function. We call the new parameters natural parameters of the distribution. Table 3.1 presents the transformation of the distributions defined here in their exponential form.

3.2 Probability Concentration

Since a probability space is a measurable space the Lebesgue integral is defined. Some of the quantities that arise by such integrals fully characterize a probability distribution. In many cases, just a couple of these quantities suffices to reveal the important properties of a distribution. For example, it is a common practice to retrieve the parameters of a distribution by their moments. We refer to the integral of any function of a random variable by defining expectation as follows.

Definition 3.2.1 (Expected Value)**.** *Consider* $(\mathbb{R}_X, \mathcal{B}, \Pr \circ X^{-1})$ *the probability space of a random variable* X *. Let* $f : \mathbb{R}_X \to \mathbb{R}$ *be a function of the random variable. The expected value of* $f(X)$ *is defined as:*

$$\mathbb{E}\left[f(X)\right] = \int_{\mathbb{R}_X} f(x) p_D\left(x\right) dx\,,$$

whenever the integral exists.

Note that the expected value, as an integral, has all their good properties such as linearity. Moreover, for a discrete random variable X the above integral turns into a sum, i.e. $\mathbb{E}\left[f(X)\right] = \sum_{\mathbb{R}_X} f(x) p_D\left(x\right)$.

There are certain functions f whose expected value gives important information about the distribution of a random variable X. The most common among them are the mean value and the variance of X. They arise for $f(x) = x$ and $f(x) = (x - \mathbb{E}[x])^2$ respectively.

Definition 3.2.2 (Mean Value - Variance)**.** *Consider the probability space of a random variable* X $(\mathbb{R}_X, \mathcal{B}, \Pr \circ X^{-1})$*. We call mean value of* X *the expectation of* $\mathbb{E}[X]$ *and denote the variance of* X *as* $\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

Concentration

As its name implies, the expected value is the average value a distribution induces. There is a large class of distributions that are concentrated around their expected values. That is the difference between any sample of the distribution and the mean of the distribution is small with high probability. This is a very useful property as it allows us to gain a lot of information about the distribution, knowing just its first moment.

Formally, consider a random variable $Z \sim \mathcal{D}$ following some distribution $\mathcal{D}$. In the following we study the probability that $|Z - \mathbb{E}[Z]|$ is large. What is more we establish some sufficient conditions for a distribution to have good concentration.

Consider a nonnegative random variable Y. By a mere observation, for all $t > 0$,

$$t\mathbf{1}_{Y \geq t} \leq Y\,.$$

Taking the expectations of both sides we can derive *Markov's Inequality*:

$$\Pr\left[Y \geq t\right] \leq \frac{\mathbb{E}\left[Y\right]}{t}\,.$$

Taking $Y = |Z - \mathbb{E}[Z]|$ gives us a first bound on the distance of a random variable from its expected value. Notice that the initial inequality holds for every function nonnegative, nondecreasing function ϕ of Y. Then $Y \geq t$ implies $\phi(Y) \geq \phi(t)$ and thus we get:

$$\Pr\left[Y \geq t\right] \leq \Pr\left[\phi(Y) \geq \phi(t)\right] \leq \frac{\mathbb{E}\left[\phi(Y)\right]}{\phi(t)}\,.$$

For $Y = |Z - \mathbb{E}[Z]|$ and $\phi(t) = t^2$, the above result becomes *Chebyshev's Inequality*:

$$\Pr\left[|Z - \mathbb{E}[Z]| \geq t\right] \leq \frac{\mathrm{Var}[Z]}{t^2}\,.$$

Cramér-Chernoff Method

We will now present the Cramér-Chernoff method for bounding the probability that a random variable is away from its expected value. This method is a simple application of the above inequality using a specific function ϕ. It is extensively used and results in some sharp bounds.

Let $\phi(t) = e^{\lambda t}$. The parameter $\lambda > 0$ will set so as we acquire the best possible bound. This will become clear shortly. Markov's inequality implies:

$$\Pr\left[Y \geq t\right] \leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda Y}\right]\,.$$

Denote $\psi_Y(\lambda) = \log \mathbb{E}[e^{\lambda Y}]$. The above inequality becomes:

$$\Pr\left[Y \geq t\right] \leq \exp\left(-(\lambda t - \psi_Y(\lambda))\right)\,.$$

Thus to get the tighter bound possible for this ϕ we must minimize $\lambda t - \psi_Y(\lambda)$. This is the Cramér-Chernoff method. We denote the $\psi_Y^*(t) = \sup_{\lambda \geq 0}(\lambda t - \psi_Y(\lambda))$ the desired value.

Example 3.2.1 (Normal Distribution)**.** *Let $Z \sim N(0, \sigma^2)$. Then it holds that:*

$$\psi_Z(\lambda) = \log\left(\int_{\mathbb{R}} e^{\lambda z} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma^2}} dz\right) = \log\left(e^{\lambda^2\sigma^2/2} \int_{\mathbb{R}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\lambda\sigma^2)^2}{2\sigma^2}} dz\right) = \frac{\lambda^2\sigma^2}{2},$$

where the second equality follows by completing the square. Then the function $f(\lambda) = \lambda t - \frac{\lambda^2\sigma^2}{2}$ is differentiable. It takes its maximum value for $\lambda = t/\sigma^2$. Thus $\psi_Z^*(t) = \frac{t^2}{2\sigma^2}$ and

$$\Pr[Z \geq t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Example 3.2.1 implies that for $Y \sim N(\mu, \sigma^2)$ it holds that

$$\Pr[Y - \mathbb{E}[Y] \geq t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

That is, the probability for a normal random variable to differ from its expected value falls exponentially with its variance. This behaviour is desirable as it implies a strong concentration around the mean value. What is more, many random variables, not normally distributed, share this property as can be indicated by their graph.

Sub-Gaussian Random Variable

We introduce the notion of a sub-Gaussian random variable to formalize the above. Consider a random variable X such that $\mathbb{E}[X] = 0$. Then X is called *sub-Gaussian* with variance factor u if

$$\psi_X(\lambda) \leq \frac{\lambda^2 u}{2}, \forall \lambda \in \mathbb{R}.$$

We denote the collection of such random variable $\mathcal{G}(u)$. So we write $X \in \mathcal{G}(u)$. Following the same procedure as in Example 3.2.1 we derive that for a sub-Gaussian random variable X it holds that $\Pr[X \geq t] \leq e^{-t^2/(2u)}$, $t > 0$.

The Cramér-Chernoff method is generally popular due to its elegant application on sums of random variables.

Lemma 3.2.1. *Consider a sequence $(X_i)_{i=1}^n$ of independent sub-Gaussian random variables, i.e. $X_i \in \mathcal{G}(u_i)$. Then their sum $X = \sum_{i=1}^n X_i$ is also a sub-Gaussian random variable $X \in \mathcal{G}(\sum_{i=1}^n u_i)$.*

Proof. The lemma follows by simply applying independence

$$\psi_X(\lambda) = \log\left(\mathbb{E}\left[e^{\lambda \sum_{i=1}^n X_i}\right]\right) = \sum_{i=1}^n \log\left(\mathbb{E}\left[e^{\lambda X_i}\right]\right) \leq \frac{\lambda^2 \sum_{i=1}^n u_i}{2}.$$

□

Thus we can get a characterization over a sum of independent r.v. by information on each one of them.

As motivation for defining 'sub-Gaussianity' we claimed that many random variables exhibit concentration similar to the Normal distribution. As a matter of fact, every random variable taking values in a bounded interval is sub-Gaussian. This is Hoeffding's lemma presented right after.

Lemma 3.2.2 (Hoeffding's Lemma [BLM13])**.** *Let Y be a random variable with $\mathbb{E}[Y] = 0$, taking values in a bounded interval $[\alpha, \beta]$ and let $\psi_Y(\lambda) = \log \mathbb{E}[e^{\lambda Y}]$. Then*

$$\psi''_Y(\lambda) \leq (\beta - \alpha)^2/4 \quad \textit{and} \quad Y \in \mathcal{G}\left((\beta - \alpha)^2/4\right) .$$

The following theorem combines the two previous lemmas 3.2.1 and 3.2.2 to get a powerful result.

Theorem 3.2.1 (Hoeffding's Inequality [BLM13])**.** *Let $X_1, \ldots, X_n$ be independent random variables such that X_i takes its value in $[\alpha_i, \beta_i]$ for all $i \leq n$. Let*

$$S = \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])) .$$

Then, for every $t > 0$,

$$\Pr[S \geq t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(\beta_i - \alpha_i)^2}\right) .$$

Applying the above theorem for the Binomial distribution is straight forward. We give this result as a corollary because it will be used very often it the following chapters.

Corollary 3.2.1.1 (Hoeffding's Inequality - Binomial Distribution)**.** *Consider the random variable $X \sim \text{Bin}(n, p)$ following the Binomial distribution. Then*

$$\Pr[|X - np| \geq t] \leq 2\exp\left(-2t^2/n\right), t > 0 .$$

Proof. Consider the sequence of random variables $(X_i)_{i=1}^{n}$ such that $X_i \sim \text{Be}(p)$. By definition it holds that $X = \sum_{i=1}^{n} X_i$ and $X_i \in [0, 1]$. Theorem 3.2.1 implies:

$$\Pr[X - np \geq t] \leq \exp\left(-2\frac{t^2}{n}\right) .$$

Then consider the distribution of

$$-X = \sum_{i=1}^{n} -X_i = \sum_{i=1}^{n} X'_i ,$$

where $X'_i \in [-1, 0]$. The same theorem gives:

$$\Pr[-(X - np) \geq t] \leq \exp\left(-2\frac{t^2}{n}\right) .$$

So by union bound we get:

$$\Pr[|X - np| \geq t] = \Pr[\{X - np \geq t\} \cup \{-(X - np) \geq t\}] \leq 2\exp\left(-2\frac{t^2}{n}\right) .$$

□

Sub-Gamma Random Variable

Surely, sub-Gaussianity is a highly convenient tool. However, strong concentration results hold for random variables that have slightly 'heavier' tails. The notion of sub-Gaussianity inspires us to define similar notions comparing to the tails of other distributions. One of them is Gamma distribution. This was not introduced in the introductory tools since it is not important for our purposes. The only demand from the reader is to consider it as a probability distribution with p.d.f. presented in figure 3.1. Thus its tail converges slower than the Gaussian's but still fast enough.

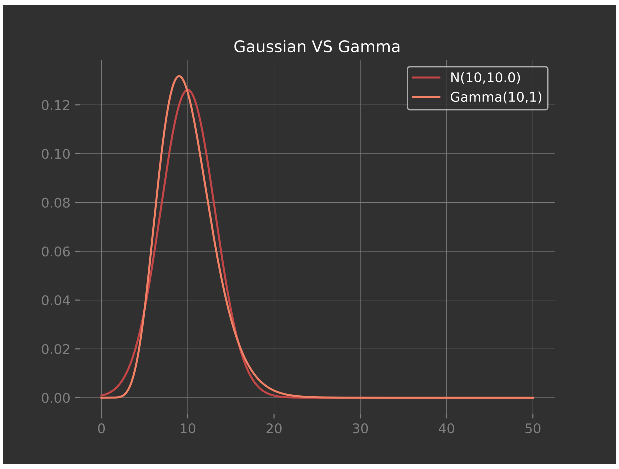

Figure 3.1: Gaussian VS Gamma Tails

Formally, consider a centered random variable X such that

$$\psi_X(\lambda) \leq \frac{\lambda^2 u}{2(1-c\lambda)}, 0 < \lambda < 1/c.$$

Then X is said to be *sub-gamma on the right tail with variance factor u and scale parameter c.* If the above property is true for $-X$ the random variable is said to be *sub-gamma on the left tail with variance factor u and scale parameter c.* Notice that since one of the tails is 'heavier' than the Gaussian tail the other must be sub-Gaussian. This is immediate by the demand for full probability over the domain equal to 1.

Recall the initial motivation for defining sub-gamma characterization. We proceed to a characterization of the sub-gamma tail.

Lemma 3.2.3 (Sub-Gamma Random Variable)**.** *Consider a centered random variable X that is right or left sub-gamma with parameters u and c. Then*

$$\Pr\left[X > \sqrt{2ut} + ct\right] \vee \Pr\left[-X > \sqrt{2ut} + ct\right] \leq e^{-t}$$

Proof. Assume X is right sub-gamma random variable so it holds

$$\psi_X(\lambda) = \frac{\lambda^2 u}{2(1-c\lambda)}, 0 < \lambda < 1/c.$$

Apply Cramér-Chernoff method. Let

$$f(\lambda) = \lambda t - \frac{\lambda^2 u}{2(1-c\lambda)}$$

be the function to be maximized. Recall we ask for

$$\psi_X^*(t) = \sup_{\lambda \in (0,1/c)} (\lambda t - \psi_X(\lambda)) = \sup_{\lambda \in (0,1/c)} f(\lambda).$$

For $\lambda \in (0, 1/c)$ the function $f(\lambda)$ is differentiable. So we can calculate its maximum by just setting the derivative to zero. This happens for $\lambda = \frac{1}{c} - \frac{1}{c}\sqrt{\frac{u}{2ct+u}}$ and we get

$$\psi_X^*(t) = \frac{u}{c^2}\left(1 + \frac{ct}{u} - \sqrt{1 + 2\frac{ct}{u}}\right).$$

So the tail bound is

$$\Pr[X > t] \leq \exp\left(-\frac{u}{c^2}\left(1 + \frac{ct}{u} - \sqrt{1 + 2\frac{ct}{u}}\right)\right).$$

Notice that the same bound holds for the left tail as well, since it must be sub-Gaussian. Thus

$$\Pr\left[X > t\right] \vee \Pr\left[-X > t\right] \leq \exp\left(-\frac{u}{c^2}\left(1+\frac{ct}{u}-\sqrt{1+2\frac{ct}{u}}\right)\right).$$

Getting the inverse of ψ_X^* we get the result

$$\Pr\left[X > \sqrt{2ut}+ct\right] \vee \Pr\left[-X > \sqrt{2ut}+ct\right] \leq e^{-t}.$$

Equivalent manipulations show the same result for a left sub-gamma random variable and the lemma is proven. □

Finally we want to highlight the gap between the sub-Gaussian and sub-gamma notions. We present a random variable that is sub-gamma but not sub-Gaussian, customizing the need for the newly defined notion.

Example 3.2.2 (sub-gamma but not sub-Gaussian)**.** *Let* $X \sim N(0,1)$ *and* $Y = X^2$. *Then* $\mathbb{E}[Y] = 1$ *and*

$$\mathbb{E}\left[e^{\lambda(Y-1)}\right] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\lambda(x^2-1)}e^{-x^2/2}dx = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}.$$

Thus for $\lambda \geq 1/2$ *the integral does not exists and* Y *is not sub-Gaussian. However notice that*

$$\psi_Y\left(\lambda\right) = -\lambda - \frac{1}{2}\log\left(1-2\lambda\right) \leq \frac{1}{2}\frac{(2\lambda)^2}{2(1-2\lambda)} = \frac{\lambda^2}{(1-2\lambda)},$$

where the inequality follows from some calculus since $2\lambda \in (0,1)$. *So* Y *is sub-gamma with parameters* $u = c = 2$.

Thus by lemma 3.2.3 we have

$$\Pr\left[|X| > 2\sqrt{t}+2t\right] \leq \Pr\left[|X| > 2t\right] \leq 2e^{-t},$$

and $\Pr[|X| > t] \leq 2e^{-t/2}$.

Anti-concentration

Undoubtedly, the concentration of distributions is a very useful property. It permits safe predictions for the behaviour of phenomena. However, anti-concentration is also an important trait. When the domain of a distribution is $[n]$ an anti-concentration result guarantees that many of the values in the domain have some mass. That is, the distribution is actually a random phenomenon not just a noisy signal around a constant and, thus, a trivial situation. In the next lemma we give an anti-concentration result about the Binomial distribution which will be useful in what follows.

Lemma 3.2.4 (Binomial Anti-concentration [FKS21])**.** *Consider a random variable* X *following the binomial distribution, i.e.* $X \sim Bin(n,p)$. *Then,*

$$\Pr\left[X = x\right] = O\left(\frac{1}{\sigma}\right), \forall x \in [n],$$

where σ *is the deviation of the binomial distribution. Moreover, the above implies that:*

$$\Pr\left[x \leq X \leq x+t\right] = O\left(\frac{t+1}{\sigma}\right), \forall x, t \in [n]$$

3.3 Probability Approximation

As is common in analysis, studying sequences and their convergence is of fundamental importance. In the case of random variables this practice is imperative. If we wish to understand a non-deterministic experiment we are forced to analyze it through the frequency of its outcomes. Thus arguing about the asymptotic behaviour of a sequence of random variables lies at the very heart of probability theory.

Convergence of Random Variables

There is a number of choices to approach the convergence of a random sequence X_n. First, recall that a random variable is a function over a probability space. The usual convergence of $X_n(\omega)$ gives almost sure convergence $X_n \to_{a.s.} X$, i.e.

$$\Pr\left[\lim_{n\to\infty} X_n = X\right] = 1\,.$$

Moreover, we can consider that X_n converges to X if the probability that they differ goes to zero as n goes to infinity. This is called convergence in probability. We write $X_n \to_{\Pr} X$ and it holds

$$\lim_{n\to\infty} \Pr\left[|X_n - X| > \varepsilon\right] = 0\,.$$

The concept that will mostly interest us here is convergence in distribution. It is also called weak convergence.

Definition 3.3.1 (Convergence in Distribution)**.** *Consider $(I, \mathcal{B}_I, \mathcal{D})$ a probability space. Let X_n a sequence of random variables and X be random variable over the space. Let F_n, F be their respective c.d.f. We say that X_n converges in distribution to X and write $X_n \to_D X$ if*

$$\lim_{n\to\infty} F_n\left(x\right) = F\left(x\right)\,,$$

for all continuity points of F.

The three notions of convergence presented above are given in order of strength. That means that $X_n \to_{a.s.} X$ implies $X_n \to_{\Pr} X$ which in turn implies $X_n \to_D X$. Convergence in distribution, though the weaker one, depends solely on the distribution of random variables. This could actually be an advantage since it allows us to depart from the properties of a specific mapping between spaces. Then the results of our analysis are applicable to any phenomenon whose behaviour is well modeled by this distribution. For more information about the notion of weak convergence we refer to [Bil13].

Central Limit Theorem

A mere application of distribution convergence is enough to remove any doubt about its 'strength'. This is no other than the celebrated Central Limit Theorem. Recalling the statement we have:

Theorem 3.3.1 (CLT)**.** *Let $X_1, \ldots, X_n$ be independent, identically distributed random variables, with mean value $\mathbb{E}[X_i] = \mu$ and variance $\mathrm{Var}[X_i] = \sigma^2$. Denote*

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}\,.$$

Then the c.d.f. of Z_n converges to the c.d.f. of the normal distribution i.e.

$$\lim_{n\to\infty} \Pr\left[Z_n \leq z\right] = \Phi\left(z\right) ,$$

for all $z \in \mathbb{R}$.

Needless to say, CLT refers to the convergence in distribution of the sum of random variables to the Normal distribution. It is considered the most fundamental result of probability theory with myriads of implications and uses in a variety of fields. Thus the importance of convergence in distribution is without question.

Probability Distances

To establish weak convergence, one would think of classical analytic proofs through $F_n(x)$ or equivalent conditions. For instance,

$$\lim_{n\to\infty} \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n = e^{-\frac{t^2}{2}} .$$

As very neatly stated in [Zol76], an alternative and more attractive way to acquire such results is by studying the distance between distributions. More attractive, since it also provides with some quantitative information about their proximity. On a more involved motivation, an asymptotic argument cannot be of much use when creating algorithms.

Thus we proceed to define two notions of distance for probability distributions. Apart from their leading role in general theory, they will be our main tools in all the following chapters. For an extensive presentation of probability metrics and their relation we refer to [GS02].

Definition 3.3.2 (Total Variation Distance)**.** *Consider P, Q probability distributions over $(\mathbb{R}, \mathcal{B})$. The total variation distance between P, Q is defined as:*

$$\mathrm{TV}\left(P, Q\right) = \sup_{B\in\mathcal{B}} \left|P\left(B\right) - Q\left(B\right)\right| .$$

An equivalent definition holds for probability distributions over discrete spaces. In fact, for a discrete space F we get:

$$\mathrm{TV}\left(P, Q\right) = \frac{1}{2} \sum_{k\in F} \left|p\left(k\right) - q\left(k\right)\right| .$$

The total variation distance is an important distribution metric. This is partly because it implies weak convergence, i.e. if $\mathrm{TV}(P_n, P) \to 0$ then $P_n \to P$. Upper bounding it for some classes of distributions is the central aim of this thesis. However, it is often hard to compute as it compares the mass on every subset of the domain. An easier to manipulate notion of distance is the KL-Divergence which is defined right away.

Definition 3.3.3 (KL-Divergence)**.** *Consider P, Q probability distributions over $(\mathbb{R}, \mathcal{B})$. Let p, q denote the p.m.f./p.d.f. of P, Q respectively. The Kullback-Leibler divergence between P, Q is defined as:*

$$\mathrm{KL}\left(P \| Q\right) = \int_{\mathbb{R}} p(x) \log \frac{p(x)}{q(x)} dx .$$

One can easily check that KL-Divergence is not an actual metric. It does not satisfy the symmetry property nor the triangle inequality. Still it formalizes an important comparison between the information encoded by different distributions. Moreover, as the next Theorem demonstrates, it offers an upper bound for their TV distance.

Theorem 3.3.2 (Pinsker's Inequality). *Consider P, Q probability distributions over $(\mathbb{R}, \mathcal{B})$. Then*

$$\mathrm{TV}\left(P, Q\right) \leq \sqrt{\frac{1}{2}\mathrm{KL}\left(P \| Q\right)}\,.$$

Note that, thanks to its multiplicative form, KL-divergence is easier to compute. Thus it consists a powerful tool for several approximation tasks involving the TV distance.

As an application, the following lemma calculates the upper bounds of the two metrics for the Binomial Distribution.

Lemma 3.3.1. *Let $\mathrm{Bin}(n,p)$, $\mathrm{Bin}(n,q)$ be two binomial distributions with natural parameters θ_p, θ_q respectively. Then, the following hold:*

1. $\mathrm{KL}\left(\mathrm{Bin}(n,p) \| \mathrm{Bin}(n,q)\right) \leq n \cdot |\theta_p - \theta_q|^2\,.$
2. $\mathrm{TV}\left(\mathrm{Bin}(n,p), \mathrm{Bin}(n,q)\right) \leq \frac{\sqrt{2n}}{2}|\theta_p - \theta_q|\,.$

Proof. We will prove bound 1 for the KL-Divergence and bound 2 follows from Pinsker's Inequality 3.3.2. By definition of the KL-Divergence we get:

$$\mathrm{KL}\left(\mathrm{Bin}(n,p) \| \mathrm{Bin}(n,q)\right) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^x \log \frac{\binom{n}{x} p^x (1-p)^x}{\binom{n}{x} q^x (1-q)^x} = n\left(p \log \frac{p}{q} + (1-p)\log\frac{1-p}{1-q}\right)\,,$$

after some elementary calculations. We will now prove the following inequality

$$\left(p \log \frac{p}{q} + (1-p)\log\frac{1-p}{1-q}\right) \leq \left(\log\frac{p}{1-p} - \log\frac{q}{1-q}\right)^2\,.$$

Note that $\theta_p = \log p/1-p$ as given in table 3.1. Thus showing this inequality gives the result and the proof is concluded. We refer to Proposition 17 of [FKT20] for the inequality. □

Moreover, we will prove the bound for KL-Divergence of the Gaussian distribution as it will be useful in subsequent chapters.

Lemma 3.3.2. *Let $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ be two Normal distributions. Then*

$$\mathrm{KL}\left(N\left(\mu_1, \sigma_1^2\right) \| N\left(\mu_2, \sigma_2^2\right)\right) \leq \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2} - 1 + 2\log\frac{\sigma_2}{\sigma_1}\right)\,.$$

Proof. We get this result by the definition of KL-Divergence.

$$\mathrm{KL}\left(N\left(\mu_1, \sigma_1^2\right) \| N\left(\mu_2, \sigma_2^2\right)\right) = \int_{\mathbb{R}} \frac{\exp\left(-(x-\mu_1)^2/2\sigma_1^2\right)}{\sigma_1\sqrt{2\pi}} \log\left(\frac{\sigma_2 \exp\left(-(x-\mu_1)^2/2\sigma_1^2\right)}{\sigma_1 \exp\left(-(x-\mu_2)^2/2\sigma_2^2\right)}\right) dx\,.$$

Then the logarithms gives three quantities. First

$$\log \sigma_2/\sigma_1\,,$$

which is independent of x so the integral equals one. Then

$$-\frac{(x-\mu_1)^2}{2\sigma_1^2}\,,$$

where $(x-\mu_1)^2$ in the integral is the definition of $N(\mu_1,\sigma_1^2)$ variance. Thus the result is $-1/2$. The third is

$$-\frac{(x-\mu_2)^2}{2\sigma_2^2}\,.$$

Here we have to develop $(x-\mu_2)^2$ and work with powers of x. Putting everything together we get the result

$$\mathrm{KL}\left(N\left(\mu_1,\sigma_1^2\right)\|N\left(\mu_2,\sigma_2^2\right)\right)\leq\frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2}+\frac{(\mu_1-\mu_2)^2}{\sigma_2^2}-1+2\log\frac{\sigma_2}{\sigma_1}\right)\,,$$

and the proof is complete. □

Chapter 4

Learning Theory

In this chapter the second fundament of this thesis, learning theory, is set. If probability theory presents us with the problem, learning theory will be the perspective through which we address this problem. In the first section, a formalized model of the learning objective is given. At the same time we introduce the famous Stochastic Gradient Descent algorithm and explore at which extend it achieves this objective. In sequence, the merging between probabilities and learning occurs. Finally, we define and give some of the background of the specific problem studied here.

4.1 PAC Learning Model and Stochastic Gradient Descent

Today, Machine Learning consists the state of the art tool for automated problem solving. Creating a system that can *learn* is a very attractive and exciting goal. Still the exact meaning of it can be tricky and hard to formalize. In what follows, we give a mathematical foundation for learning and present the most widespread algorithm of the field, Stochastic Gradient Descend.

Probably Approximately Correct

In practice, what does it mean to learn? Suppose you are trying to learn a new game. Learning demands a subject, since you always learn *something.* Second it usually consists of the ability to process this subject and reach a conclusion. In our example, given a state of the game you must decide on your next move. Finally, to consider you learnt something your conclusions must be 'correct', for some notion of correctness. Notice that, given the game's state, there are good, bad and even unreasonable moves.

One can recognise three structures in the previous description. We proceed to formalize them. Let $\mathcal{Z}$ denote the domain of the learning task. That is, objects in $\mathcal{Z}$ represent instances of the task which we must process. Usually $\mathcal{Z}$ consists of two other sets $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Then $\mathcal{X}$ represents the instances of the task (state of the game) and $\mathcal{Y}$ is a labeling set (possible moves). This is actually a vast category of learning problem called supervised learning. In general, $\mathcal{Z}$ can be any set.

Let $\mathcal{H}$ denote the hypothesis class. This is the set of functions that process the input from $\mathcal{Z}$ and reach a conclusion useful for the problem. In the supervised case, it contains functions $h : \mathcal{X} \to \mathcal{Y}$ that characterize the task instances by some label (suggest the next move). Again, $\mathcal{H}$ can be an arbitrary set, even one that just *represents* some function. Consider $h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$, $\mathbf{w}, \mathbf{x} \in \mathbb{R}^d$. These functions are called hyperplanes in $\mathbb{R}^d$ since they divide the space. Notice that we can represent such a function just by the vector $\mathbf{w}$. Thus, in this case, we set $\mathcal{H} = \mathbb{R}^d$.

Learning a task means reaching *correct* conclusions. We define a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$ to quantify the *correctness* of a hypothesis. The larger the values of $\ell(h, z)$, the greater the loss h induces by its result on z. The aim of learning is to find an $h \in \mathcal{H}$ that minimizes the loss. We will denote a learning problem by the triplet $(\mathcal{H}, \mathcal{Z}, \ell)$.

As for the learning process, our framework is as follows. The aim is to construct a system that learns automatically, i.e. *a learning algorithm.* Consider algorithm A for $(\mathcal{H}, \mathcal{Z}, \ell)$. Then A outputs a hypothesis $h \in \mathcal{H}$. The algorithm's input is a *training set* $\mathcal{S} \in \mathcal{Z}^m$, $m \in \mathbb{N}$. That means $\mathcal{S}$ is a list of m instances of the problem that the algorithm processes to decide on a 'good', with respect to ℓ, hypothesis h. These instances are sampled independently according to a *distribution* $\mathcal{D}$. Thus the assessment of the output is based on the *risk function* $L_\mathcal{D}(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h, z)]$. This is to model the fact that, a good hypothesis does not have to be correct over all possible tasks, but over the probable ones.

Definition 4.1.1 (Agnostic PAC Learnability [SSBD14])**.** *A hypothesis class $\mathcal{H}$ is agnostic PAC learnable with respect to a set $\mathcal{Z}$ and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$, if there exist a function $m_\mathcal{H} : (0, 1)^2 \to \mathbb{N}$ and a learning algorithm A such that: For every $\varepsilon, \delta \in (0, 1)$, for every distribution $\mathcal{D}$ over $\mathcal{Z}$, given a training set $\mathcal{S} \sim \mathcal{D}^m$ with $m \geq m_\mathcal{H}(\varepsilon, \delta)$ i.i.d. samples from $\mathcal{D}$, the algorithm returns a hypothesis $h = A(\mathcal{S})$ so as*

$$L_\mathcal{D}(h) \leq \min_{h' \in \mathcal{H}} L_\mathcal{D}(h') + \varepsilon \,,$$

where $L_\mathcal{D}(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h, z)]$, with probability at least $1 - \delta$.

Notice that the notion of learnability is defined with respect to a hypothesis class. That is, we wish to know whether a task can be modeled by a function in this set. This immediately implies that, even if a learning task is not learnable in $\mathcal{H}$, it could be in another hypothesis class $\mathcal{H}'$. Thinking about it though, it would be a lost cause to search over all possible functions for the one to model our specific problem. In many cases, this might also lead to wrong results (e.g. *overfitting*). This definition allows us to exploit the *prior knowledge* we have for the problem.

To consider a learning algorithm successful it must satisfy two requirements, based on the above definition. It must be efficient in terms of *sample complexity* as well as regular *bit-operation complexity.* In a vast number of problems, there is a complete characterization for the classes $\mathcal{H}$ that satisfy the first condition: their VC-dimension. We do not introduce this notion here as these problems belong mostly to the supervised learning framework and will not interest us. Refer to [SSBD14] for an extensive presentation of the topic.

Convex Learning Problem

The regular complexity of the learning algorithm is another story. Indeed, to guarantee efficient learning we must restrict ourselves to a specific class of learning problems: *convex learning problems.* The definition is given in the following. But first, let us recall the basics about convexity.

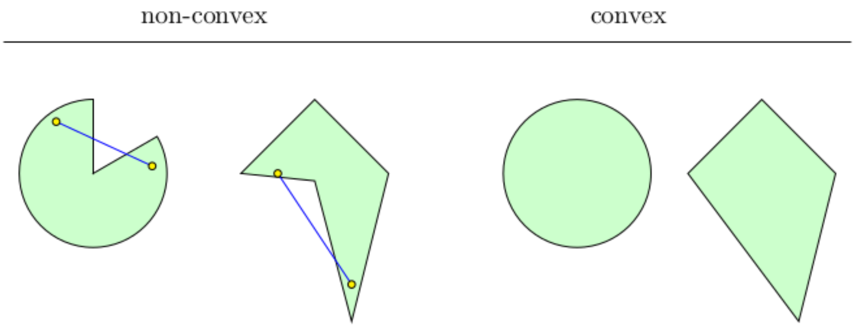

Figure 4.1: Convex Sets [SSBD14]

Definition 4.1.2 (Convex Set)**.** *A set C in a vector space is convex if for any two vectors $\mathbf{u}$, $\mathbf{v} \in C$, the line segment between $\mathbf{u}$ and $\mathbf{v}$ is contained in C. That is, for any $\alpha \in [0,1]$ we have that:*

$$\alpha\mathbf{u} + (1-\alpha)\mathbf{v} \in C\,.$$

The notion of convexity is better understood through an image, see 4.1. We proceed to define a convex function.

Definition 4.1.3 (Convex Function)**.** *Let C be a convex set. A function $f : C \to \mathbb{R}$ is convex if for every $\mathbf{u}$, $\mathbf{v} \in C$ and $\alpha \in [0,1]$,*

$$f\left(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}\right) \leq \alpha f\left(\mathbf{u}\right) + (1-\alpha) f\left(\mathbf{v}\right)\,.$$

A very important property of a convex function is that every local minimum is also a global one. Thus, if it differentiable and its derivative has a zero point, it follows that it is a minimum. Recall that our aim is to locate such a minimum for the risk function.

Again we illustrate a convex function in figure 4.2. Notice that the tangent in any point of the domain lies below the function graph. This is no coincidence. In fact, convexity is equivalent to an even more general property that bypasses the strict tangent notion (i.e. the demand for differentiability).

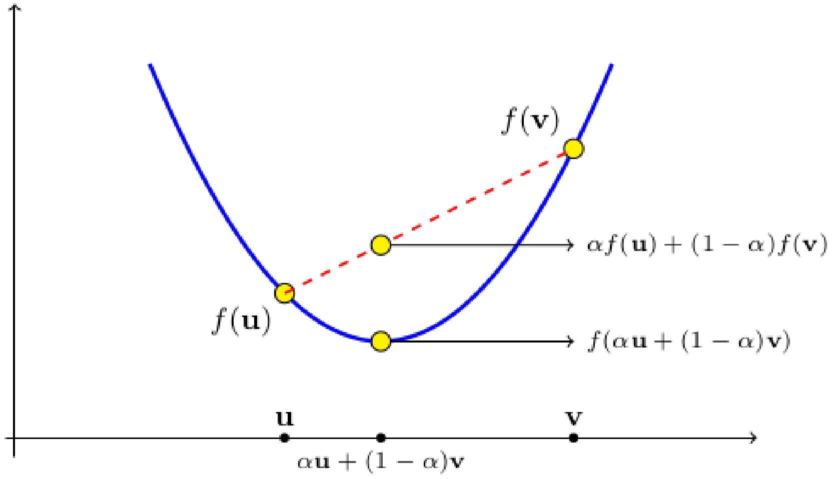
$f(\mathbf{v})$
$\alpha f(\mathbf{u}) + (1-\alpha) f(\mathbf{v})$
$f(\mathbf{u})$
$f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v})$
$\mathbf{u}$
$\mathbf{v}$
$\alpha\mathbf{u} + (1-\alpha)\mathbf{v}$

Figure 4.2: Convex Function [SSBD14]

Lemma 4.1.1. *Let C be an open convex set. A function $f : C \to \mathbb{R}$ is convex iff for every $\mathbf{w} \in C$ there exists $\mathbf{v}$ such that*

$$f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \mathbf{v} \rangle, \forall \mathbf{u} \in C .$$

The above vector $\mathbf{v}$ specifies a line, passing from $(\mathbf{w}, f(\mathbf{w}))$, while leaving the rest of the function points 'above' it. When f is differentiable on $\mathbf{w}$ this line becomes the tangent. We define a set that contains all these lines, since they have a very useful for our purpose property: they point in the direction of the function's minimum.

Definition 4.1.4 (Subgradient)**.** *A vector $\mathbf{v}$ that satisfies lemma 4.1.1 is called a subgradient of f at $\mathbf{w}$. The set of subgradients of f at $\mathbf{w}$ is called the differential set and denoted $\partial f(\mathbf{w})$.*

We are now ready to define the learning problem that interests us here.

Definition 4.1.5 (Convex Learning Problem)**.** *A learning problem, $(\mathcal{H}, Z, \ell)$, is called convex if the hypothesis class $\mathcal{H}$ is a convex set and for all $z \in Z$, the loss function, $\ell(\cdot, z)$, is a convex function.*

Since $\mathcal{H}$ is a convex set it is a set in a vector space. Thus, in what follows, we can consider hypotheses as vectors in $\mathbb{R}^d$.

Stochastic Gradient Descend

For a convex learning problem to be learnable, we need an algorithm that minimizes the risk function $\mathbb{E}_{z\sim\mathcal{D}}[\ell(\mathbf{w}, z)]$. The vector $\mathbf{w} \in R^d$ denotes the hypothesis. Since the loss function for this problem is convex the risk function will also be convex[1]. Thus if it has a

[1] This is a result of Jensen's Inequality defined as follows: Consider a convex function ϕ on the real line and a function g over a probability space with measure P. Then

$$\phi\left(\int g dP\right) \leq \int (\phi \circ g) dP .$$

minimum, we can accurately calculate it.

Of course, this claim is not really straightforward. How will we find this minimum? A first approach would derive from high school analysis. Calculating the minimum of a function is just a matter of setting its derivative to zero. Still, there is no guarantee that the risk function is differentiable. As a result, we turn to an iterative method, Gradient Descend, to reach the minimum. That is, based on the subgradient on subsequent points of the function, we move our guess for the minimum towards the real one. In more detail:

- Denote $\mathbf{w}_t$ our guess for the minimum of a convex function and initialize it at zero, i.e. $\mathbf{w}_1 = 0$.
- Update its value according to the rule

  $$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{v}_t \,, \mathbf{v}_t \in \partial f(\mathbf{w}_t)\,.$$

  where η is a constant to control the step size.
- Return $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$ after T iterations.

Intuitively, this process should lead us to the minimum. However, the purpose is to get there 'fast'. The analysis of the algorithm indicates whether it efficiently solves a convex learning problem.

Formally, the function to be minimized is $f(\mathbf{w}) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(\mathbf{w}, z)]$. Let $\mathbf{w}^*$ denote the minimum value of the function. We want to bound $f(\bar{\mathbf{w}}) - f(\mathbf{w}^*)$ by a decreasing, in terms of the required samples, function. We proceed in two steps.

Step 1: (Convexity) The following relation is just a combination of Jensen's Inequality (for a sum instead of an integral) and the definition of convexity.

$$f\left(\frac{1}{T}\sum_{t=1}^{T} w_t\right) - f\left(\mathbf{w}^*\right) \leq \frac{1}{T}\sum_{t=1}^{T}\langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{v}_t\rangle$$

Step 2: (A Lemma) Combining the previous relation with the following lemma gives a bound on the error of Gradient Descend after T iterations.

Lemma 4.1.2 (Lemma 14.1 [SSBD14])**.** *Let* $\mathbf{v}_1, \ldots, \mathbf{v}_T$ *be an arbitrary sequence of vectors. Any algorithm with initialization* $\mathbf{w}_1 = 0$ *and an update rule of the form*

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{v}_t$$

satisfies

$$\sum_{t=1}^{T}\langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{v}_t\rangle \leq \frac{\|\mathbf{w}^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2\,.$$

Notice that there is a tradeoff on the value of η. If the norm of $\mathbf{w}^*$ is large we should set η to be large as well. But this would increase the value of the other term in the bound. It becomes obvious that some further restrictions on the norms of $\mathbf{w}^*$ and $\mathbf{v}_t$ must be placed to acquire convergence. Still we will first address a more immediate problem that arises by using Gradient Descend.

This is a very useful tool in probability theory since the expectation of a quantity is, in fact, an integral.

Gradient Descend reduces the problem of convex learning to finding the subgradient of the risk function $\mathbb{E}_{z\sim\mathcal{D}}[\ell(\mathbf{w}, z)]$, which is simple enough, had we known the function. The case is, distribution $\mathcal{D}$ along with the risk function is unknown to the learning algorithm. The idea of Stochastic Gradient Descend (SGD) is to specify an estimate of the subgradient. Specifying the *expected* direction of the subgradient suffices for the algorithm to behave as the original Gradient Descend in expectation.

Formally, we will now bound $\mathbb{E}[f(\bar{\mathbf{w}}) - f(\mathbf{w}^*)]$. Again Jensen's Inequality implies that

$$\mathbb{E}\left[f\left(\bar{\mathbf{w}}\right) - f\left(\mathbf{w}^*\right)\right] \leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(f\left(\mathbf{w}_t\right) - f\left(\mathbf{w}^*\right)\right)\right].$$

Then it is a matter of manipulating the expectation over $\mathbf{v}_t$ to gain

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(f\left(\mathbf{w}_t\right) - f\left(\mathbf{w}^*\right)\right)\right] \leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{v}_t\rangle\right].$$

Working as in *Step 2*, it is immediate to bound the last quantity. The formal steps for SGD are given in algorithm 1.

Algorithm 1 SGD

```
procedure StochasticGradientDescent($\eta$, T)    ▷ $\eta > 0$: scalar, $T > 0$: integer
    $\mathbf{w}_1 \leftarrow 0$
    for $t \in [1 : T]$ do
        choose $\mathbf{v}_t$ such that $\mathbb{E}[\mathbf{v}_t|\mathbf{w}_t] \in \partial f(\mathbf{w}_t)$    ▷ problem depended
        $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta\mathbf{v}_t$
    return $\bar{\mathbf{w}} \leftarrow \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$
```

Notice that the bound on the error of Gradient Descend remains unchanged in the stochastic case, since it is based on lemma 4.1.2. We highlight that the inability to efficiently learn some convex problems is not a trait of SGD. Indeed there are convex problems that cannot be learnt by any deterministic algorithm as illustrated by the following example.

Example 4.1.1. *Consider a supervised learning problem with $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$ and $\mathcal{H} = \mathbb{R}$. Let the loss function be $\ell(w, x) = (wx - y)^2$ where $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $w \in \mathcal{H}$. This is called a regression problem, since the aim (controlled by the loss function) is to specify a linear relation between x, y. Thus we can consider the samples as points on the plane (x, y).*

Now let $\varepsilon = 1/100$, $\delta = 1/2$ and $\mu = \frac{\log(100/99)}{2m}$ where $m > m(\varepsilon, \delta)$ is the samples demanded by the algorithm. Since the learning algorithm R is deterministic m is fixed given ε and δ. Consider two points on the plane $A = (1, 0)$ and $B = (\mu, -1)$. Assume we want to distinguish between two distributions over A, B: $\mathcal{D}_1 : p(A) = \mu, p(B) = 1 - \mu$ and $\mathcal{D}_2 : p(A) = 0, p(B) = 1$.

Notice that μ is a very small quantity. So the probability of seeing point A is very small for both distributions. Indeed, the probability that $\mathcal{D}_1$ creates independent samples just of B is $\prod_{i=1}^{m} p(B) = (1 - \mu)^m \geq e^{-2m\mu} = 0.99$. Then the risk function for each of them is $L_{\mathcal{D}_1}(w) = \mu w^2 + (1 - \mu)(w\mu + 1)^2$ and $L_{\mathcal{D}_2}(w) = (w\mu + 1)^2$

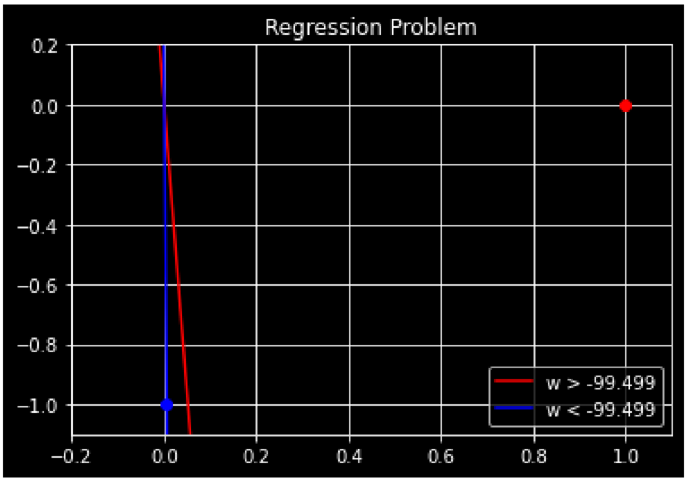

Figure 4.3: Example

As a result, whatever value R decides on w it will be suboptimal with respect to at least one of $\mathcal{D}_1$, $\mathcal{D}_2$. This is illustrated in figure 4.3. If $w < -1/(2\mu)$ the line will be too far from A. Thus $L_{\mathcal{D}_1}(w) \geq 1/(4\mu)$ while a better value is $L_{\mathcal{D}_1}(0) = (1-\mu)$. It follows that $L_{\mathcal{D}_1}(w) - \min_{w'} L_{\mathcal{D}_1}(w') > 1/100 = \varepsilon$. On the other hand, if $w \geq -1/(2\mu)$ the error on $\mathcal{D}_2$ will become larger than ε. Notice that $min_{w'} L_{\mathcal{D}_2}(w') = 0$ while $L_{\mathcal{D}_2}(w) \geq 1/4$.

Thus we have a convex learning problem that is not learnable by any deterministic algorithm R.

Thus we need to further restrict ourselves. The good news is we already know the assumption needed for a convex problem to be learnable: bounding the norms of $\mathbf{w}^*$ and $\mathbf{v}_t$. There is a number of ways to define such problems, such as *convex-Lipschitz-bounded* learning problem and *convex-smooth-bounded* learning problems (see [SSBD14]). We will focus on a specific one which will also concern us in what follows.

Projected SGD and Strong Convexity

It is important to understand the actual difficulty in learning a convex problem. Firstly, convexity does not guarantee the existence of a minimum. Secondly, even if the minimum exists it might not be unique. As a result, it is not possible to distinguish between two minimizers which is the actual solution to the problem. There is a stronger condition, strong convexity that does guarantee both the above properties.

Definition 4.1.6 (Strong Convexity)**.** *Let C be a convex set. A function $f : C \to \mathbb{R}$ is λ-strongly convex if for all $\mathbf{w}$, $\mathbf{v}$ and $\alpha \in (0,1)$ we have*

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

Again, we can acquire some useful bounds involving the subgradient of the λ-strongly convex function. That is

$$\langle \mathbf{w}-\mathbf{u}, \mathbf{v}\rangle \geq f(\mathbf{w}) - f(\mathbf{u}) + \frac{\lambda}{2}\|\mathbf{w}-\mathbf{u}\| .$$

Thus strong convexity implies a a gap between the 'tangent' on a specific point $\mathbf{w}$ and the line through the $f(\mathbf{w})$ and any other point on the function. Notice that figure 4.2

actually represents a strongly convex function. The distance between $f(\alpha \mathbf{u}) + (1-\alpha)\mathbf{v})$ and $\alpha f(\mathbf{u}) + (1-\alpha)f(\mathbf{v})$ is at least $\frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{w}-\mathbf{v}\|^2$. After some thought, we gain the intuition that this property distinguishes convex functions from the line. Thus the function always has a minimum, which is unique, and it reaches it in a λ-controllable rate.

Before stating the theorem for learnability of strongly convex problems, we introduce a variation of SGD. This mild modification does not affect the convergence of the algorithm and is quite useful in a variety of applications (including the one that interests this thesis). That is, a projection step. Recall the set of our hypotheses $\mathcal{H}$. Notice that the iteration process can get $\mathbf{w}$ outside $\mathcal{H}$. This, in general, does not affect learnability of a problem (definition 4.1.1 does not demand $h = A(\mathcal{S}) \in \mathcal{H}$). However, it is usually convenient if $\mathbf{w} \in \mathcal{H}$. What is more, in the case of SGD, projection does not affect the analysis of the algorithm. This is implied by the following lemma.

Lemma 4.1.3 (Projection Lemma [SSBD14]). *Let $\mathcal{H}$ be a closed convex set and let $\mathbf{v}$ be the projection of $\mathbf{w}$ onto $\mathcal{H}$, namely,*

$$\mathbf{v} = \min_{\mathbf{x} \in \mathcal{H}} \|\mathbf{x} - \mathbf{w}\|^2 .$$

Then, for every $\mathbf{u} \ni \mathcal{H}$,

$$\|\mathbf{w} - \mathbf{u}\|^2 - \|\mathbf{v} - \mathbf{u}\|^2 \geq 0 .$$

Then projected SGD algorithm for the strongly convex case is as follows.

Algorithm 2 PSGD λ-strongly Convex

```
procedure ProjectedSGDStronglyConvex(T, λ)            ▷ $T > 0$: integer
    $\mathbf{w}_1 \leftarrow 0$
    for $t \in [1 : T]$ do
        choose $\mathbf{v}_t$ such that $\mathbb{E}[\mathbf{v}_t|\mathbf{w}_t] \in \partial f(\mathbf{w}_t)$     ▷ problem depended
        $\eta_t \leftarrow 1/(\lambda t)$
        $\mathbf{w}_{t+1/2} \leftarrow \mathbf{w}_t - \eta_t \mathbf{v}_t$
        $\mathbf{w}_{t+1} \leftarrow argmin_{\mathbf{w} \in \mathcal{H}} \|\mathbf{w} - \mathbf{w}_{t+1/2}\|^2$
    return $\bar{\mathbf{w}} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$
```

Notice that the value of η is now specifically defined. This proves helpful for the convergence of the algorithm. Moreover, it is consistent to intuition. While t increases we approach closer to the minimum. Thus the step size becomes smaller to avoid bypassing the minimum.

Finally, theorem 4.1.1 gives the desired convergence rate which implies efficient learnability both in terms of samples and computation steps.

Theorem 4.1.1 (SGD with Strongly Convex Objective). *Assume f is λ-strongly convex and $\mathbb{E}[\|\mathbf{v}_t\|^2] \leq \rho^2$. Let $\mathbf{w}^* = argmin_{\mathbf{w} \in \mathcal{H}} f(\mathbf{w})$ be an optimal solution and $\mathbf{w}$ the output of SGD algorithm. Then,*

$$\mathbb{E}\left[f(\mathbf{w})\right] - f(\mathbf{w}^*) \leq \frac{\rho^2}{2\lambda T}(1 + \log T) .$$

Subsequent work has removed the $\log T$ term from the bound by slight modifications of the algorithm (see [SZ13](averaging), [HK14](batch method)). Then, setting $\varepsilon = \frac{\rho^2}{2\lambda T}$, we deduce that after $T > \frac{\rho^2}{2\lambda\varepsilon}$ iterations (and, in consequence, samples) arbitrary precision is achieved.

Given the previous theorem one could still argue that learnability is not proven. Surely, definition 4.1.1 requires $f(\mathbf{w}) - f(\mathbf{w}^*)$ to be bounded. The last lemma of this section suggests a routine that achieves this result. To elaborate, given an algorithm that guarantees convergence in expectation, running it a number of times and carefully choosing an outcome we can guarantee true convergence. This is procedure is called *boosting* and is a thoroughly studied area (see [SF12]).

Lemma 4.1.4. *Assume f is λ-strongly convex and there exists an algorithm* A *such that, with $m \geq m(\varepsilon/3)$:*

$$\mathbb{E}\left[f(\mathbf{w})\right] \leq f(\mathbf{w}^*) + \varepsilon/3\,,$$

where $\mathbf{w}$ is the output of algorithm A *and $\mathbf{w}^*$ the minimum of f. Then, there exists an algorithm that uses $m \geq \log(1/\delta)m(\varepsilon)$ and finds a w' such that*

$$f(\mathbf{w}') \leq f(\mathbf{w}^*) + \varepsilon\,,$$

with probability at least $1-\delta$.

Proof. Consider the random variable $W = f(\mathbf{w}) - f(\mathbf{w}^*)$ with probability over $\mathbf{w}$. Note that W is non negative since $\mathbf{w}^*$ is the minimizer of f by assumption. Then, Markov's inequality implies:

$$\Pr\left[f(\mathbf{w}) - f(\mathbf{w}^*) \geq t\right] \leq \frac{\varepsilon}{3t}\,.$$

Setting $t = \varepsilon$ we get that:

$$f(\mathbf{w}) \geq f(\mathbf{w}^*) + \varepsilon$$

with probability at most $\frac{1}{3}$. Running algorithm A for $n = 72\log(1/\delta)$ times, it follows that, with probability at least $1-\delta$, over half of the output values will get f ε-close to its optimal value. Formally, consider the random variable $Y = \frac{1}{n}\sum_{i\in[n]} Y_i$ where Y_i equals 1 if $f(\mathbf{w}_i) > f(\mathbf{w}^*) + \varepsilon$, 0 otherwise. It follows that $\mathbb{E}[Y] \leq \frac{1}{3}$. By Hoeffding's inequality 3.2.1.1:

$$\Pr\left[Y - \mathbb{E}[Y] \geq 1/12\right] \leq \exp\left(-n/72\right) \leq \delta\,.$$

Thus, with probability at least $1-\delta$ it holds that $Y \leq \mathbb{E}[Y] + \frac{1}{12} \leq \frac{1}{3} + \frac{1}{12}$. So, $1 - Y > \frac{1}{2}$ and more than half of the estimations give $f(\mathbf{w}_i) < f(\mathbf{w}^*) + \varepsilon$. Since f is λ-strongly convex it is true that:

$$\frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}^*\|^2 \leq f(\mathbf{w}) - f(\mathbf{w}^*) \leq \varepsilon\,,$$

for at least half of $\mathbf{w}_i$. Thus, the following procedure gives a good $\mathbf{w}$: find the distances between all the $72\log(1/\delta)$ estimators and choose $\mathbf{w}'$ that is $\frac{2\varepsilon}{\lambda}$-close to at least half of the others. Then, with probability $1-\delta$ it holds:

$$f(\mathbf{w}') - f(\mathbf{w}^*) \leq \varepsilon$$

and the proof is complete. □

4.2 Distribution Learning from Truncated Samples

In this section, we define the actual problem addressed in this thesis. A specific goal is set: learning parametric probability distributions. That is, given i.i.d. samples that are known/assumed to follow a probability distribution, we ask for a function that gives the 'same' probabilities. This problem lies in the heart of statistical learning. Recall the actual problem in PAC framework: the distribution over the samples is unknown. Assuming that the distribution belongs to a known, parametric family and retrieving it, we can optimize any objective for the problem. What is more, we further challenge ourselves, assuming partial access to samples from the distribution, i.e. we sample from a truncated version of it.

4.2.1 Parametric Distribution Learning

We want to study parametric distribution learning in the way introduced in [KMR+94]. This model was inspired from the PAC framework. First, we consider a parametric distribution family such as the Gaussian $N(\mu,\sigma^2)$ or the Binomial $\text{Bin}(n,p)$. The assumed distribution over the samples is the hypotheses class $\mathcal{H}$. The domain of the distribution also specifies the domain of the learning problem $\mathcal{Z}$. Finally, we have to set our goal. In this thesis we consider we learnt a distribution if we can arbitrarily approximate it in TV-distance. Here we depart from the standard PAC framework since there is no loss function leading to TV-distance objective. Also, notice that, in this model, there is always an $h \in \mathcal{H}$ with zero cost.

A quite attractive methodology in parametric distribution learning is specifying the parameters of a distribution and consider our goal achieved. Still, there is often a great distance from learning a distribution to learning its parameters. Consider that a non-parametric function h might be very close to the actual parametric function of the distribution. However, for several reasons, our algorithm is able to specify that h but not the actual distribution. Thus our algorithm has no idea as for the parameters. On the other hand, getting arbitrarily close to the parameters does not always guarantee we are close enough to the distribution. Note that a probability distribution is usually a non-linear function of its parameters. As a result, a small difference on the parameters' values may induce a prohibitively large loss value.

Maximum Likelihood Estimator

Having said that, the common practice when learning a parametric distribution is to estimate its parameters. The most widespread tool to achieve this is the Maximum Likelihood Estimator (M.L.E.). Given a sample set $\mathcal{S}$ we want to specify those parameters $\boldsymbol{\vartheta}$ that give us a probability distribution maximizing the probability that we saw $\mathcal{S}$. In fact, given $\mathcal{S} = \{s_i | 1 \leq i \leq m\} \sim \mathcal{D}^m$,

$$\hat{\boldsymbol{\vartheta}} = argmax_{\boldsymbol{\vartheta}} \left(p \left(\bigcap_{i=1}^{m} s_i ; \boldsymbol{\vartheta} \right) \right) = argmax_{\boldsymbol{\vartheta}} \left(\prod_{i=1}^{m} p\left(s_i ; \boldsymbol{\vartheta}\right) \right) .$$

Notice that getting the logarithm of the objective returns the same maximizer, since it is an increasing function. Moreover, the logarithm turns the product into sum over the probabilities. This is understood to be convenient as maximizing a function will, at some point, demand a differentiation. Moreover, in accordance to the general learning framework, we consider the negative logarithm to get a minimization problem. So we equivalently demand

$$\hat{\boldsymbol{\vartheta}} = argmin_{\boldsymbol{\vartheta}} \left(-\sum_{i=1}^{m} \log\left(p\left(s_i; \boldsymbol{\vartheta}\right)\right) \right) .$$

Thus we give a formal definition for M.L.E.

Definition 4.2.1 (Maximum Likelihood Estimator)**.** *Consider a parametric probability distribution $\mathcal{D}$ over some domain $\mathcal{Z}$ with p.m.f./p.d.f. $p(x; \boldsymbol{\vartheta})$, $x \in \mathcal{Z}$ and $\boldsymbol{\vartheta} \in \mathcal{H}$. Let $\mathcal{S} \sim \mathcal{D}^m$ be an i.i.d. sample set of size m. Then the Maximum Likelihood Estimator is*

$$\hat{\boldsymbol{\vartheta}} = argmin_{\boldsymbol{\vartheta} \in \mathcal{H}} \left(-\sum_{s \in \mathcal{S}} \log\left(p\left(s; \boldsymbol{\vartheta}\right)\right) \right) .$$

We can formalize the parametric distribution learning problem as follows. Let $\mathcal{H} = \{\boldsymbol{\vartheta} | \boldsymbol{\vartheta}\text{is a valid parameter for the distribution}\}$ be the family of distributions and $\mathcal{Z}$ be the domain of this distribution. Also, let

$$\ell\left(\boldsymbol{\vartheta}, x\right) = -\log\left(p\left(x; \boldsymbol{\vartheta}\right)\right), x \in \mathcal{Z}\,, \boldsymbol{\vartheta} \in \mathcal{H}\,.$$

This is called *the negative log-likelihood objective.* Then M.L.E. defines a learning algorithm. It consists of minimizing the loss function over the sample set, i.e.

$$\hat{\boldsymbol{\vartheta}} = argmin_{\boldsymbol{\vartheta}} \left(\sum_{i=1}^{m} \ell\left(\boldsymbol{\vartheta}, s_i\right) \right) = argmin_{\boldsymbol{\vartheta}} \left(-\sum_{i=1}^{m} \log\left(p\left(s_i; \boldsymbol{\vartheta}\right)\right) \right) .$$

This a very common algorithmic technique in learning called Empirical Risk Minimization (E.R.M.). Note that if $\mathcal{H}$ is a convex set and ℓ is a convex function the learning problem is convex (since $-\log(\cdot)$ is convex).

Example 4.2.1 (Normal Distribution M.L.E.)**.** *Assume we are given samples from a normal distribution $N(\mu, \sigma^2)$ defined over $\mathbb{R}$. Let $\mathcal{S} = \{x_i | 1 \le i \le m\} \sim N(\mu, \sigma^2)^m$ be the sample set. Denote $\hat{\mu}$ and $\hat{\sigma}^2$ the M.L.E. for μ and σ^2. Then, in accordance to the M.L.E. definition we have $\boldsymbol{\vartheta} = [\mu, \sigma^2]$ and we want*

$$\hat{\boldsymbol{\vartheta}} = [\hat{\mu}, \hat{\sigma}^2] = argmin_{\boldsymbol{\vartheta} \in \mathbb{R} \times \mathbb{R}_+} \left(-\sum_{i=1}^{m} \log\left(p\left(x_i; \boldsymbol{\vartheta}\right)\right) \right) .$$

Denote $L(\mathcal{S}; \boldsymbol{\vartheta})$ the function to be minimized for convenience. Since the p.d.f. is that of the normal distribution it becomes

$$L\left(\mathcal{S}; [\mu, \sigma^2]\right) = -\sum_{i=1}^{m} \left(\log\left(1/(\sqrt{2\pi\sigma^2})\right) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) .$$

Notice that L is a two-parameter, differentiable, convex function. So its minimizing points can be found setting the partial derivatives to zero.

$$\frac{\partial L(\mathcal{S}; [\hat{\mu}, \hat{\sigma}^2])}{\partial \mu} = \sum_{i=1}^{m} \frac{(x_i - \mu)}{\sigma^2} = 0\,,$$

$$\frac{\partial L(\mathcal{S};[\hat{\mu},\hat{\sigma}^2])}{\partial\sigma^2}=\sum_{i=1}^{m}\left(-\frac{1}{\sigma}+\frac{(x_i-\mu)}{\sigma^3}\right)=0\,.$$

Thus

$$\hat{\mu}=\frac{1}{m}\sum_{i=1}^{m}x_i \quad \textit{and} \quad \hat{\sigma}^2=\frac{1}{m}\sum_{i=1}^{m}(x_i-\hat{\mu})^2\ .$$

Notice that the M.L.E. is not always a *consistent* estimator. In statistics, an estimator is consistent if its limit, given an infinite number of samples, it reaches its expected value, i.e.

$$\lim_{n\to\infty}\hat{\mu}=\mu\,.$$

The common correction that makes the variance estimator of a Gaussian consistent is to divide by one less sample. That is

$$\hat{\sigma}^2=\frac{1}{m-1}\sum_{i=1}^{m}(x_i-\hat{\mu})^2\ .$$

Learning Gaussian Distribution in TV-distance

We are now ready to present an example in which learnability of parameters implies learnability of distribution. The next lemma along with its proof indicates that, if the KL-divergence is determined by a 'good' function of these parameters this approach works. Moreover, it consists an example for M.L.E.'s usefulness.

Lemma 4.2.1 (Learning Gaussian Distribution in TV-distance)**.** *Consider the Gaussian distribution $N(\mu,\sigma^2)$. Let $\varepsilon\leq 2/3$. Then there exists an efficient algorithm that given $m=\Theta(\frac{\log(1/\delta)}{\varepsilon^2})$ samples from $N(\mu,\sigma^2)$ calculates estimators $\hat{\mu}$, $\hat{\sigma}^2$ such that*

$$\mathrm{TV}\left(N\left(\mu,\sigma^2\right),N\left(\hat{\mu},\hat{\sigma}^2\right)\right)\leq\varepsilon$$

with probability at least $1-\delta$.

Proof. Consider $N(\mu,\sigma^2)$ the unknown distribution. Let $\mathcal{S}=\{x_i|1\leq i\leq m\}\sim N(\mu,\sigma^2)^m$ denote the sample set. The promised learning algorithm is M.L.E. So the estimators are given by example 4.2.1. We will use the consistent estimator for the variance, so

$$\hat{\mu}=\frac{1}{m}\sum_{i=1}^{m}x_i \quad \text{and} \quad \hat{\sigma}^2=\frac{1}{m-1}\sum_{i=1}^{m}(x_i-\hat{\mu})^2\ .$$

We will shortly prove how these estimators give good approximations for μ and σ^2 i.e. we will prove that, given $m=\Theta(\frac{2\log(2/\delta)}{\varepsilon^2})$ samples,

$$|\hat{\mu}-\mu|\leq\sigma\varepsilon \quad \text{and} \quad |\frac{\hat{\sigma}^2}{\sigma^2}-1|\leq\varepsilon^2\,.$$

Let X_i, $1\leq i\leq m$ be random variables representing the samples drawn from the distribution. Then $X_i\sim N(\mu,\sigma^2)$ and is a sub-Gaussian random variable, $X_i\in\mathcal{G}(\sigma^2)$. By lemma 3.2.1 we get that

$$\Pr\left[|m\hat{\mu}-m\mu|>t\right]\leq e^{-t^2/(2m\sigma^2)}\,.$$

So setting $\varepsilon = t/(m\sigma)$ it follows that

$$\Pr\left[|\hat{\mu} - \mu| > \sigma\varepsilon\right] \leq 2e^{-m\varepsilon^2/2}\,.$$

Given $m = \frac{2\log(2/\delta)}{\varepsilon^2}$ we get

$$|\hat{\mu} - \mu| \leq \sigma\varepsilon\,,$$

with probability at least $1-\delta$, as promised.

As for the variance, consider $Y_i = \frac{X_i - \hat{\mu}}{\sigma}$. So $Y_i \sim N(0,1)$ and Y_i^2 is a sub-gamma random variable by example 3.2.2. Thus

$$\Pr\left[\left|\frac{(X_i - \hat{\mu})^2}{\sigma^2} - 1\right| > t\right] \leq 2e^{-t/2}\,.$$

By manipulations similar to lemma 3.2.1 it is straight-forward to show for the normalized sum of Y_i^2

$$\mathbb{E}\left[e^{\lambda\left(\sum_{i=1}^m \frac{Y_i^2}{m-1} - 1\right)}\right] = \left(\frac{e^{-\lambda/m-1}}{\sqrt{1 - 2\frac{\lambda}{m-1}}}\right)^m\,,$$

and thus get

$$\psi_{\left(\frac{1}{m-1}\sum_{i=1}^m Y_i^2\right)}(\lambda) \leq \frac{2\lambda^2}{(m-1)(1 - 2\frac{\lambda}{m-1})}\,.$$

So the normalized sum is also a sub-gamma random variable with $u = 4/m$, $c = 2/m$. The sub-gamma tail bound gives:

$$\Pr\left[\left|\frac{\hat{\sigma}^2}{\sigma^2} - 1\right| > \varepsilon^2\right] \leq 2e^{-(m-1)\varepsilon^2/2}\,.$$

Again, given $m = 1 + \frac{2\log(2/\delta)}{\varepsilon^2}$ samples the desired relation follows and

$$\left|\frac{\hat{\sigma}^2}{\sigma^2} - 1\right| \leq \varepsilon^2\,,$$

with probability at least $1-\delta$.

Given the above approximations for the Gaussian parameters, lemma 3.3.2 implies that

$$\mathrm{KL}\left(N\left(\hat{\mu}, \hat{\sigma}^2\right), N\left(\mu, \sigma^2\right)\right) \leq \frac{1}{2}\left(\frac{\sigma^2\varepsilon^2}{\sigma^2} + \left(\frac{\hat{\sigma}}{\sigma}\right)^2 - 1 - \log\left(\frac{\hat{\sigma}}{\sigma}\right)^2\right) \leq 2\varepsilon^2\,,$$

since $x^2 - 1 - \log x^2 < 3(x-1)^2$ and $(\hat{\sigma}/\sigma - 1)^2 \leq \varepsilon^2$ is implied by the previous result[2]. Thus, by Pinsker's Inequality 3.3.2 we get the desired result

$$\mathrm{TV}\left(N\left(\hat{\mu}, \hat{\sigma}^2\right), N\left(\mu, \sigma^2\right)\right) \leq \varepsilon\,.$$

□

Note that by tighter bounds on TV-distance we can remove the demand for $\varepsilon \leq 2/3$ (see Theorem 1.3 [DMR18]).

[2] see Lemma 2.11 in [ABDH$^+$20].

4.2.2 Truncated Samples

In this subsection a greater challenge for parametric distribution learning is raised. To sum up our current framework, we assume sample access to the unknown distribution. Solely based on this information, we want to retrieve the actual distribution. This is possible under a wide range of conditions, enabling the formalization of strong prediction techniques. This sample oracle is, in practice, data, gathered from real world 'experiments', imposed to all the imperfections this might dictate.

Specifically, assume a part of the distribution's domain is 'hidden' from our oracle. This is not due to the low probability of the values there, but because of some dysfunctionality of the experiment. For instance, if samples are given by a measurement device and the device has some sensitivity limit, we might lose samples below it.

Thus the new framework assumes oracle access to a subset of the distribution's domain. So we truncated the domain to include only this set. We can think of this procedure as follows: There is a visible set, subset of the distribution's domain, $S \subseteq \mathcal{Z}$. This is called the truncation set. Now we sample from our complete oracle and reject any sample that does not belong to S. Returning all the samples that do to the algorithm, we have constructed our sample set. So we get every sample in the truncation set with probability proportional to its original probability. All other values are given probability zero. This defines a new probability distribution called *truncated distribution.* Note that for the newly defined function to be a distribution, it must assign probability 1 to its domain, i.e. the truncation set. Thus, the probability of a value in S in the truncated case is not exactly equal to the original one, but they have a fixed fraction.

Definition 4.2.2 (Truncated Distribution)**.** *Let $\mathcal{D}$ be a probability distribution over $\mathcal{Z}$. Consider the truncation set $S \subseteq \mathcal{Z}$. The truncated on S probability distribution $\mathcal{D}$, denoted $\mathcal{D}_S$, is defined as*

$$\mathcal{D}_S\left(x\right) = \frac{\mathbf{1}_{\{x \in S\}}}{\mathcal{D}(S)} \mathcal{D}\left(x\right)\,, x \in \mathcal{Z}\,.$$

A general problem

We will insist a little longer on the motivation for defining this framework. This is an effort to connect it with a more general requirement in problem solving.

Let us recall the fundamental problem of solving an equation system. This is probably the first problem one encounters in his undergraduate studies and lies at the heart of linear algebra (see [Str]). An equation system models the behaviour of a real life system. In principle, we decide on a set of variables $\mathbf{x}$ and wish to understand how they affect system's behaviour. This behaviour is defined through some characteristic quantities of the system. These can be measured and their values consist another vector $\mathbf{b}$. The connection between the two is given by system's dynamics, represented by an array $\mathbf{A}$. So, we finally get the well known

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

In general, our aim is to specify the values of chosen parameters $\mathbf{x}$ so as to achieve a certain behaviour $\mathbf{b}$, i.e.

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}\,.$$

One could consider the inverse matrix $\mathbf{A}^{-1}$ as a *procedure* that takes us from *observations* in $\mathbf{b}$ to a *decision* for the appropriate values for $\mathbf{x}$. Notice we are faced with the same challenge: how does a small error in measurement of $\mathbf{b}$ affect our decision? The answer depends on the *procedure* $\mathbf{A}$ to go from one to the other.

It is now clear that the problem of *sensitivity* or *stability* or *robustness* of an algorithm is of grave importance. Back to learning theory, robustness of learning algorithms is highly desired trait. There is a vast, still developing, literature concerned with this topic. Note that robust algorithms should work even in the case that no errors exists. Thus studying them must offer a better understanding of the problem. Needless to say, the kind of 'error' learning theory regards refers to the quality of the training set, since this is the only input in our algorithms. One such, relatively easy to be honest, 'error' is our newly defined truncation over distribution's domain.

Results so far

As highlighted above, truncated samples is a phenomenon arising in many real world problems we might need to model. Thus, regardless of its great connections, its study was early necessitated by application demand.

The generally accepted as first to deal with truncated samples is Galton ([Gal98]) in 1898. Apparently, Galton wished to model race-horses' speed. In fact, we wanted to examine whether the Gaussian distribution is a good model for the average race-horse speed. However, the organisers of the races only registered those horses that achieved a time record over $2'$ and $30''$. Thus, the data available for Galton's model was truncated. Still he went on and, taking into account that one tail of the distribution must be missing, fit a truncated Gaussian into the data he had.

In the following years the study of truncated distributions continued, though in a framework quite different of what we presented above. That is, specific cases of truncation sets were considered, usually removing the left or the right tail or a combination of them. We refer the reader to [Hug62], [Coh91] for an extensive presentation of the early approach on parameter estimation from the truncated samples.

The modern formulation we introduced was actually given in [DGTZ18]. This is also where the problem of learning a Gaussian distribution, in d dimensions, from an arbitrary truncation set is settled. In theorem 2.0.1 we state this result in the case of $d = 1$ for simplicity (and consistency to our definitions).

Algorithm 3 PSGD for Truncated distributions

```
function GRADIENTESTIMATION($\mathbf{p}, S$)    ▷ Assume oracle access to samples
    $x \leftarrow$ Oracle.
    while $y \leftarrow \mathcal{D}(\mathbf{p}) \notin S$ do.
                                                ▷ $\mathcal{D}$ denotes the truncated distribution
    return $-x + y$.
procedure TRUNCATEDPSGD($M, \lambda, S$)
    $\hat{\mathbf{p}} \leftarrow$ EmpiricalEstimator.    ▷ The parameters are initialized as their empirical estimators
    $\bar{\mathbf{p}} \leftarrow$ PROJECTEDSGDSTRONGLYCONVEX($M, \lambda$).    ▷ Using $\hat{\mathbf{p}}$, projection on $S$ and gradientEstimation
    return $\bar{\mathbf{p}}$.
```

The learning algorithm is actually familiar to the reader: it is Projected SGD for a λ-strongly convex learning problem! The objective function is the negative log-likelihood, thus providing us with the m.l. estimators. This can be shown to be strongly convex. In fact, in the next lemma, we prove that every distribution in exponential family has a convex 'truncated' negative log-likelihood objective. Recall that the key for working with SGD is to specify a good gradient estimation. There lies the beauty of the previous theorem. Taking a sample x from the sample's oracle (that is, the true truncated distribution) and a sample y from the truncated current guess of the distribution. Their difference indicates the expected direction of the gradient.

We will give the, in essence, exact proof of the above theorem when proving our main result. In preparation, we prove the claimed convexity of negative log-likelihood of any distribution belonging to an exponential family. Observe the strength of this result. Had we shown strong convexity, it would imply that *every distribution in exponential family* is learnable by its truncation.

Lemma 4.2.2 (Application: Exponential Families and M.L.E.)**.** *Let D be a distribution in the exponential family set and D_S the truncation of D on the set S. Then, the negative log-likelihood objective for D_S is convex with respect to its natural parameters.*

Proof. Let $\boldsymbol{\vartheta}$ be the natural parameters vector of D and, consequently, D_S. Assume D is a discrete distribution and denote N the support set of D. Then, for $x \in N$:

$$D_S\left(\boldsymbol{\vartheta}; x\right) = \frac{\exp\left(\boldsymbol{\vartheta}^T \mathbf{T}(x) - A(\boldsymbol{\vartheta}) + B(x)\right)}{D\left(\boldsymbol{\vartheta}; S\right)}.$$

Thus, the negative log-likelihood objective is

$$\ell\left(\boldsymbol{\vartheta}; x\right) = -\boldsymbol{\vartheta}^T \mathbf{T}(x) + A(\boldsymbol{\vartheta}) - B(x) + \ln D\left(\boldsymbol{\vartheta}; S\right).$$

For $\ell(\boldsymbol{\vartheta}; x)$ to be convex with respect to $\boldsymbol{\vartheta}$ it suffices to show that its Hessian with respect to $\boldsymbol{\vartheta}$ is non-negative.
First, we compute the gradient of the negative log-likelihood as

$$\nabla_\theta \ell(\boldsymbol{\vartheta}; x) = \frac{\nabla_{\boldsymbol{\vartheta}} D\left(\boldsymbol{\vartheta}; S\right)}{D\left(\boldsymbol{\vartheta}; S\right)} - \mathbf{T}(x) + \nabla_{\boldsymbol{\vartheta}} A(\boldsymbol{\vartheta}).$$

Note that for $A(\boldsymbol{\vartheta})$ we have that

$$A(\boldsymbol{\vartheta}) = -\log D\left(\boldsymbol{\vartheta}; S\right) + \log \sum_{x \in S} e^{\left(\boldsymbol{\vartheta}^T \mathbf{T}(x) + B(x)\right)}$$

for normalization to hold. Then

$$\nabla_{\boldsymbol{\vartheta}} A(\boldsymbol{\vartheta}) = -\frac{\nabla_{\boldsymbol{\vartheta}} D\left(\boldsymbol{\vartheta}; S\right)}{D\left(\boldsymbol{\vartheta}; S\right)} + \frac{\sum_{x \in S} \mathbf{T}(x) \cdot e^{\left(\boldsymbol{\vartheta}^T \mathbf{T}(x) + B(x)\right)}}{\sum_{x \in S} e^{\left(\boldsymbol{\vartheta}^T \mathbf{T}(x) + B(x)\right)}}$$

which multiplying the nominator and denominator by $\exp\left(-A(\boldsymbol{\vartheta})\right)$ gives

$$\nabla_{\boldsymbol{\vartheta}} A(\boldsymbol{\vartheta}) = -\frac{\nabla_{\boldsymbol{\vartheta}} D\left(\boldsymbol{\vartheta}; S\right)}{D\left(\boldsymbol{\vartheta}; S\right)} + \mathbb{E}_{x \sim D_S(\boldsymbol{\vartheta})}\left[\mathbf{T}(x)\right] .$$

Then the gradient of the negative log-likelihood becomes

$$\nabla_{\theta} \ell(\boldsymbol{\vartheta}; x) = -\mathbf{T}(x) + \mathbb{E}_{x \sim D_S(\boldsymbol{\vartheta})}\left[\mathbf{T}(x)\right] .$$

Finally, we calculate the Hessian of the negative log-likelihood

$$H_{\ell}(\boldsymbol{\vartheta}) = \frac{\sum_{x \in S} \mathbf{T}(x) \cdot \left[\mathbf{T}(x) - \nabla_{\boldsymbol{\vartheta}} A(\boldsymbol{\vartheta})\right] \cdot e^{\left(\boldsymbol{\vartheta}^T \mathbf{T}(x) - A(\boldsymbol{\vartheta})\right)}}{D\left(\boldsymbol{\vartheta}; S\right)} - \frac{\nabla_{\boldsymbol{\vartheta}} D\left(\boldsymbol{\vartheta}; S\right)}{D^2\left(\boldsymbol{\vartheta}; S\right)} \cdot \sum_{x \in S} \mathbf{T}(x) \cdot e^{\left(\boldsymbol{\vartheta}^T \mathbf{T}(x) - A(\boldsymbol{\vartheta})\right)} .$$

Recalling the results about the gradient of $A(\boldsymbol{\vartheta})$ we get

$$H_{\ell}(\boldsymbol{\vartheta}) = \mathbb{E}_{x \sim D_S(\boldsymbol{\vartheta})}\left[\mathbf{T}(x)^T \mathbf{T}(x)\right] - \mathbb{E}_{x \sim D_S(\boldsymbol{\vartheta})}\left[\mathbf{T}(x)\right] \cdot \mathbb{E}_{x \sim D_S(\boldsymbol{\vartheta})}\left[\mathbf{T}(x)\right] ,$$

which follows from the definition of the covariance matrix of $\mathbf{T}(x)$

$$H_{\ell}(\boldsymbol{\vartheta}) = \mathrm{Cov}_{x \sim D_S(\boldsymbol{\vartheta})}\left[\mathbf{T}(x), \mathbf{T}(x)\right] .$$

Thus, the Hessian is positive semi-definite and the negative log-likelihood is convex. □

Chapter 5

Poisson Binomial Distribution

In this chapter we present the family of Poisson Binomial Distributions. This is the first of the two distributions studied in this thesis. We refer to it as a *family* since a particular Poisson Binomial distribution is defined with respect to the number of its parameters n. Moreover, n controls the distribution's domain which is $[n] = \{0, 1, \ldots, n\}$. This work studies whether we can learn a PBD with n parameters in TV distance based on truncated samples.

We postpone dealing with our learnability issue until the next chapter. For now, we first define the distribution formally and analyze some basic properties. Next a series of approximations for PBDs is given to gain an insight on the actual structure of a PBD. Right after we present a theorem that fully characterizes this structure and will act as the cornerstone of our solution. Finally, to acquire a complete perspective of the learning problem, the learning algorithm for the non-truncated case is outlined.

5.1 Poisson Binomial Distribution

The definition of Poisson Binomial Distribution is, in fact, a very basic and natural one. Recall the Binomial distribution defined in section 3.1. It refers to the sum of n independent Bernoulli random variables with the same parameter p. The Poisson Binomial distribution is defined the same way, but the Bernoulli's are allowed to have different parameters p_i. Poisson was the first to consider this kind of distribution, thus its name. One can see [TT19] for a full survey on its properties and the literature around it.

Definition 5.1.1 (PBD)**.** *Let $X_1, \ldots, X_n$ be mutually independent random variables that follow the Bernoulli distribution, i.e. $X_i \sim \mathrm{Be}(p_i)$, $1 \leq i \leq n$. Then, the sum $X = \sum_{i=1}^{n} X_i$ is said to follow the Poisson Binomial Distribution of order n. We will denote $X \sim \mathrm{PBD}(p_1, \ldots, p_n)$ and write PBD_n for the set of PBDs with n parameters.*

The common practice for working with this distribution is to manipulate it as a sum. Still, we can explicitly express its p.m.f. as follows. First, refer to the Binomial for the idea. The probability of a Binomial random variable $X \sim \mathrm{Bin}(n, p)$ to equal k is

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k} \,.$$

The term $\binom{n}{k}$ counts all the ways for *exactly* k Bernoullis to equal 1. The rest of the expression follows from the independence of X_i-s. In consequence, the p.m.f. of a PBD is given

$$p\left(k\right) = \sum_{A \subseteq [n] : |A| = k} \left(\prod_{i \in A} p_i \cdot \prod_{i \notin A} \left(1 - p_i\right) \right) ,$$

where we again sum over all the ways for exactly k of the Bernoullis to give 1 (subsets A of Bernoullis).

A very important property of the PBD, not straight-forward by the looks of it, is unimodality. That means that a PBD of order n has a unique maximum point. The graph increases at the right of it and decreases at its left. In other words, it is a concave function. See [Wan93] for an elegant combinatorial proof. We demonstrate an example of a PBD graph in fig. 5.1.

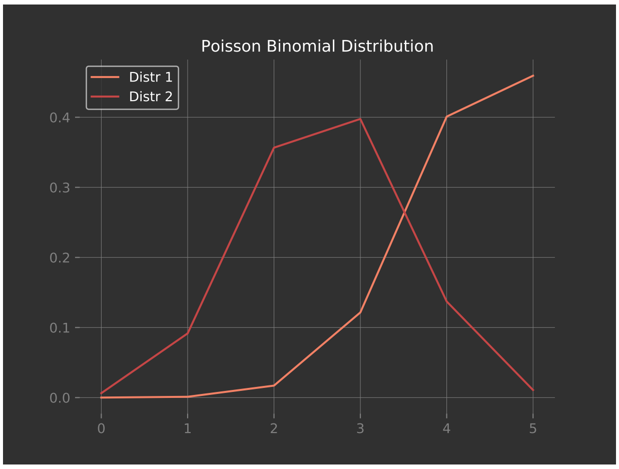

Figure 5.1: Poisson Binomial Distribution Graph.

What is more, PBDs exhibit strong concentration properties. An immediate application of theorem 3.2.1 for $X \sim \mathrm{PBD}\left(p_1, \ldots, p_n\right)$ gives

$$\Pr\left[\left|X - \bar{p}\right| \geq t\right] \leq 2e^{-2t^2} ,$$

where $\bar{p} = \sum_{i=1}^{n} p_i$.

As a final remark, note that we have already encounter this kind of distribution in this thesis. Recall that the Central Limit Theorem 3.3.1 states that the sum of n independent, identically distributed random variables converges to the Normal distribution. Of course, the PBD consists of independent but not identically distributed random variables. Still, there are variants of CLT that remove the 'identically distributed' assumption (e.g. Lindeberg and Lyapunov CLT [Bil08]).

Thus there is a close, asymptotic connection between PBDs and the Normal distribution. This observation promises a simplified way to manipulate PBDs through their Normal counterparts, thus bypassing the complicated, combinatorial expression of their p.m.f. However the asymptotic nature of CLT makes it hard for an algorithm to use it. In the following section we quantify this similarity giving a number of approximations for PBDs.

5.2 Approximation of PBDs

Approximations of PBDs by other distributions have been long studied in the literature. Here, we present three such approximations that will be useful in the following. But first, we should elaborate on the kind of approximation we want. We aim to bound the TV-distance between PBDs and each one of the distributions: Poisson, Binomial and Normal. Recall that this a quite strict requirement because of the proximity TV distance demands on every subset of the domain. Moreover, it is a very useful one, since it implies weak convergence of the distributions (see section 3.3).

Before proceeding to the approximation statements, we should work on the intuition behind them. This is actually a quite simple one. To get good approximations we try to match the moments of the distributions. That is $\mathbb{E}\left[|X|^p\right]$. This is called *method of moments* and is a widespread approach in statistics. Note that a probability distribution is fully specified by its moments, thus the method's success. In what follows, our main guideline will be to match the PBD's mean value and variance (the first two moments) with those of the approximating distribution's.

One of the first approximations given for the PBD is the Poisson distribution.

Theorem 5.2.1 (Poisson Approximation [CX02]). *Let $J_1, \ldots, J_n$ be mutually independent indicators with $\mathbb{E}[J_i] = p_i$. Then*

$$\mathrm{TV}\left(\sum_{i=1}^{n} J_i, Poisson\left(\sum_{i=1}^{n} p_i\right)\right) \leq \frac{\sum_{i=1}^{n} p_i^2}{\sum_{i=1}^{n} p_i}.$$

Notice that the Poisson distribution has the same mean and variance. Thus, matching the mean values as in the above allows for large difference in variances. To get better approximation, we need to match the second moment as well. We define a distribution, based on the Poisson, so as we can control its variance as follows.

Definition 5.2.1 (Translated Poisson [TT19]). *An integer-valued random variable X is said to be translated Poisson distributed with parameters (μ, σ^2), denoted as $TP(\mu, \sigma^2)$, if $X - \mu + \sigma^2 + \{\mu - \sigma^2\} \sim Poisson(\sigma^2 + \{\mu - \sigma^2\})$, where $\{\cdot\}$ is the fraction part of a positive number.*

The resulting bound is given by the following theorem.

Theorem 5.2.2 (Translated Poisson Approximation [DP15]). *Let $J_1, \ldots, J_n$ be mutually independent indicators with $\mathbb{E}[J_i] = t_i$. Then*

$$\mathrm{TV}\left(\sum_{i=1}^{n} J_i, TP\left(\mu, \sigma^2\right)\right) \leq \frac{\sqrt{\sum_{i=1}^{n} t_i^3(1-t_i)} + 2}{\sum_{i=1}^{n} t_i(1-t_i)},$$

where $\mu = \sum_{i=1}^{n} t_i$ and $\sigma^2 = \sum_{i=1}^{n} t_i(1-t_i)$

Next, we will approximate the PBD with a Binomial. This is another improvement on the approximation using the the Poisson distribution. We can note two reasons for this. First, the Binomial distribution allows us to control both its mean value and its variance. Second, recall that the Poisson can be considered the limiting distribution of the Binomial (while $n \to \infty$). In [CX02], [TT19] we learn that there is a monotonicity in the TV-distance between a PBD and a Binomial with the same mean value. In fact, the TV-distance increases with n, until it is maximized for $n \to \infty$ i.e. the Poisson distribution. The actual bound is given below.

Theorem 5.2.3 (Binomial Approximation [CX02]). *Let $J_1, \dots, J_n$ be mutually independent indicators with $\mathbb{E}[J_i] = t_i$, and $\bar{t} = \frac{\sum_{i=1}^{n} t_i}{n}$. Then*

$$\mathrm{TV}\left(\sum_{i=1}^{n} J_i, \mathrm{Bin}\left(n, \bar{t}\right)\right) \leq \frac{\sum_{i=1}^{n}(t_i - \bar{t})^2}{(n+1)\bar{t}(1-\bar{t})}\,.$$

Finally, we want to acquire a bound for the TV-distance between a PBD and the Normal distribution. In this case, we can also get an exact value! The TV-distance between a discrete and a continuous distribution always equals 1. Thus, to compare between the two, we will need a discretization of the Normal distribution and find the TV-distance between this and a PBD. Let us define a commonly used discretization of the Normal distribution.

Definition 5.2.2 (Discretized Normal Distribution). *Let Y be a gaussian random variable with mean value μ and variance σ^2, i.e. $Y \sim N(\mu, \sigma^2)$. The discretized normal random variable $X \sim N^d(\mu, \sigma^2)$ is defined as $X = \lfloor Y \rceil$. Then, the probability mass function of X is:*

$$\Pr\left[X = k\right] = \Pr\left[k - \frac{1}{2} \leq Y < k + \frac{1}{2}\right], \forall k \in \mathbb{Z}\,.$$

Now, we can bound the TV-distance between a PBD and a 'Normal' distribution as follows.

Theorem 5.2.4 ([CGS10] Theorem 7.1). *Consider the random variables $X_1, \dots, X_n$ such that $X_i \sim \mathrm{Be}(p_i)$. Let $X = \sum_{i=1}^{n} X_i$, $\mu = \sum_{i=1}^{n} p_i$ and $\sigma^2 = \sum_{i=1}^{n} p_i(1-p_i)$. Then,*

$$\mathrm{TV}\left(\mathcal{L}(X), N^d(\mu, \sigma^2)\right) \leq \frac{7.6}{\sigma}\,.$$

Thus, for the binomial distribution $\mathrm{Bin}(n,p)$ it holds:

$$\mathrm{TV}\left(\mathrm{Bin}(n,p), N^d(np, np(1-p))\right) \leq \frac{7.6}{\sqrt{np(1-p)}}\,.$$

Before we close this subsection, it is important to define another discretization of the Normal distribution. Moreover, we show that it is very close to the previous one and, thus, a good approximation for the PBD as well.

Definition 5.2.3. *Consider the normal distribution $N(\mu, \sigma^2)$ and denote $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ the probability mass on $x \in \mathbb{R}$. The $\mathbb{Z}$-discretized normal distribution $N_{\mathbb{Z}}^d(\mu, \sigma^2)$ is defined as:*

$$\Pr\left[X = k\right] = \frac{1}{\sum_{i \in \mathbb{Z}} p(i)} \cdot p(k) = \frac{p(k)}{p(\mathbb{Z})}\,,$$

for all $k \in \mathbb{Z}$. Notice that this is the truncated normal distribution on $\mathbb{Z}$.

Notice that the pmf of the above has the same expression as the actual Normal distribution. This will be proven very convenient for the aim of this thesis. Its proximity to the PBDs is also important and follows from the Mean Value Theorem.

Lemma 5.2.1. *Consider the random variables $X \sim N^d(\mu, \sigma^2)$ and $Y \sim N_{\mathbb{Z}}^d(\mu, \sigma^2)$. Then, it holds that:*

$$\mathrm{TV}\left(X, Y\right) < O\left(\frac{1}{\sigma^3}\right)\,.$$

Proof. Let $X \sim N^d(\mu, \sigma^2)$, $Y \sim N^d_{\mathbb{Z}}(\mu, \sigma^2)$ and $Z \sim N(\mu, \sigma^2)$. Denote $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$ the mass on $x \in \mathbb{R}$ by the normal distribution. Notice that, by the mean value theorem, there exists $c \in (k - \frac{1}{2}, k + \frac{1}{2})$ for all $k \in \mathbb{Z}$ such that:

$$p(c) = \Phi'(c) = \frac{\Phi(k + \frac{1}{2}) - \Phi(k - \frac{1}{2})}{1} = \Pr\left[X = k\right] .$$

Thus, it suffices to show that the mass on c is close to the mass on the integer closest to c, i.e. $\lfloor c \rceil$. Then,

$$\mathrm{TV}\left(X, Y\right) = \frac{1}{2} \sum_{y \in \mathbb{Z}} |p(c) - \frac{p(y)}{p(\mathbb{Z})}| .$$

Noticing that $\sum_{y \in \mathbb{Z}} p(c) = \sum_{y \in \mathbb{Z}} \frac{p(y)}{p(\mathbb{Z})} = 1$ and by the triangle inequality it follows that:

$$\mathrm{TV}\left(X, Y\right) \leq \sum_{y \in \mathbb{Z}} |p(y) - p(c)| .$$

Using the mean value theorem again, we can derive that for some c_i between y, c it holds that:

$$|p(y) - p(c)| \leq |p'(c_i)| \cdot |y - c| \leq |p'(c_i)| .$$

Since the derivative of the normal p.d.f. equals $p'(x) = -\frac{(x-\mu)}{\sigma^2} p(x)$, it follows that:

$$\mathrm{TV}\left(X, Y\right) \leq \frac{1}{\sigma^3\sqrt{2\pi}} \sum_{y \in \mathbb{Z}} |(c_i - \mu) \exp\left(-\frac{(c_i - \mu)^2}{2\sigma^2}\right)| .$$

The above series converges by the root criterion and the result follows. □

5.3 Sparse Covers for PBDs

So far the precision of a PBD approximation depended on the variance of the distribution. In this section we construct an ε-cover for PBD_n. That is a set of distributions ε-close to any PBD of order n. As one can now guess, a Normal distribution can cover any PBD with ε-large variance. See, for example, Theorem 5.2.4. For the rest, a significantly small number of parameters, say $l \ll n$, suffices to describe them up to ε accuracy. Note that ε does not depend on the variance and can be arbitrary.

The original aim of this cover, introduced in [DP15], was to make the set of PBDs easier to explore and understand. Here, we focus on one of the many implications of this construction: there is a clear separation between the close-to-continuous and the strictly discrete PBDs of order n, with respect to a degree ε of closeness. This gives us a transition from continuous to discrete that makes this family of distributions especially interesting for the problem studied in this thesis.

In what follows, we will state the main covering theorem and give a brief description of its proof.

Theorem 5.3.1 (Theorem 2 [DP15]). *Let $X_1, \ldots, X_n$ be arbitrary mutually independent indicators, and $k \in \mathbb{N}$. Then there exist mutually independent indicators $Y_1, \ldots, Y_n$ satisfying the following:*

- $\mathrm{TV}\left(\sum_i X_i, \sum_i Y_i\right) \leq 41/k$;
- *at least one of the following is true:*
  - *((n,k)-heavy Binomial form) there is some $l \in \{1,\dots,n\}$ and $q \in \{\frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n}\}$ such that, for all $i \leq l$, $\mathbb{E}[Y_i] = q_i$ and, for all $i > l$, $\mathbb{E}[Y_i] = 0$; moreover, l and q satisfy $lq \geq k^2$ and $lq(1-q) \geq k^2 - k - 1$; or*
  - *(k-sparse form) there exists some $l \leq k^3$ such that, for all $i \leq l$, $\mathbb{E}[Y_i] \in \{\frac{1}{k^2}, \frac{2}{k^2}, \dots, \frac{k^2-1}{k^2}\}$ and, for all $i > l$, $\mathbb{E}[Y_i] \in \{0, 1\}$.*

The approximating distribution (Y_i) takes two forms. The heavy Binomial form serves as the Normal approximation while the k-sparse form accounts for the rest of the distributions.

The proof proceeds in two stages. In stage 1, the parameters ε-close to 0 or 1 are eliminated. We need this step to specify the actual support of the distribution, as will be explained shortly. On the next stage, we decide on the approximation form. This will depend on the variance of the distribution, as made clear in section 5.2.

The following proof sketch aims to present a primitive and simplified intuition on why the steps taken are the ones that should. With this in mind, it is important to insist on the actual meaning of a p_i being close to 0 or 1. An event happening with probability almost 0 practically does not happen. Thus the corresponding random variables X_i never adds on the sum. This means that the maximum value of this sum is smaller than n. In the same spirit, a random variable that occurs with probability 1 always adds on the sum, initiating the support of the distribution at a value larger than 0. Thus the number of p_i-s away from 0 and 1 specify the size of the support of the distribution as well as its variance.

Proof. (sketch) Let $X = \sum_{i=1}^n X_i$ be the PBD to be approximated. Let $\varepsilon = 1/k$ specify the intended accuracy of the approximation.

A key lemma that is constantly used in the following is:

Lemma 5.3.1 ([DP15])**.** *Let $X_1, \dots, X_n$ be mutually independent random variables, and let $Y_1, \dots, Y_n$ be mutually independent random variables. Then*

$$\mathrm{TV}\left(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i\right) \leq \sum_{i=1}^n \mathrm{TV}\left(X_i, Y_i\right) .$$

Stage 1

Consider $Z_i, 1 \leq i \leq n$, a set of indicators with success probability z_i and denote $Z = \sum_{i=1}^n Z_i$ their sum. We want to properly define the z_i-s such that

$$\mathrm{TV}\left(X, Z\right) \leq c/k\,,$$

where c a constant.

Let $L_k = \{i | p_i \in (0, \frac{1}{k})\}$ denote the set of p_i-s ε-close to 0. Then map each $p_i, i \in L_k$ to a value $z_i \in \{0, 1/k\}$ such as the mean value within the set does not change much i.e.

$$\left|\sum_{i\in L_k} p_i - \sum_{i \in L_k} z_i\right| < 1/k\,.$$

By theorem 5.2.1 the sum of X_i-s on L_k is ε-close to a Poisson with the same mean. Formally:

$$\mathrm{TV}\left(\sum_{i\in L_k} X_i, \mathit{Poisson}\left(\sum_{i\in L_k} p_i\right)\right) \leq \frac{\sum_{i\in L_k} p_i^2}{\sum_{i\in L_k} p_i} \leq \frac{\frac{1}{k}\sum_{i\in L_k} p_i}{\sum_{i\in L_k} p_i} \leq 1/k\,.$$

The same is true for the Z_i-s on L_k that were just defined. Since, the mean values of the Poisson distributions are ε-close, by the definition of z_i, a known bound on the TV distance of Poisson distributions gives us a final (3.5ε)-closeness of the distributions.

Lemma 5.3.2 ([DP15]). *It holds that*

$$\mathrm{TV}\left(\mathit{Poisson}\left(\lambda_1\right), \mathit{Poisson}\left(\lambda_2\right)\right) \leq \frac{1}{2}\left(e^{|\lambda_1-\lambda_2|} - e^{-|\lambda_1-\lambda_2|}\right)\,.$$

We apply the same procedure for the p_i-s that are ε-close to 1 i.e. $H_k = \left\{i|p_i \in \left(\frac{1}{k}, 1\right)\right\}$. Finally, set the rest of z_i-s equal to the p_i-s. By lemma 5.3.1 we get that

$$\mathrm{TV}\left(X, Z\right) \leq \frac{7}{k}\,.$$

Notice that the mapping $p_i \to z_i$ increases the variance while not departing much from the original distribution. This observation follows from [TT19], Theorem 2.4., which states that the more similar the parameters of the PBD, the more variance it has. Increasing the variance brings closer to the Normal approximation which decreases the parameters from n to 2, thus, is a desired goal.

Stage 2

In this stage we must decide which approximation is suitable for the PBD in question. This is based on the variance of the distribution. We use a rough estimation that gives a criterion with respect to the size of $m = |\{i|z_i \notin \{0,1\}\}|$. As noted above, this is the size of the actual support of the distribution. Then the variance of the distribution is given as

$$\sum_{z_i\notin\{0,1\}} z_i(1-z_i) \leq \sum_{z_i\notin\{0,1\}} (1-\frac{1}{k})^2 = m(1-\frac{1}{k})^2\,.$$

For the variance to be greater than $1/\varepsilon^2 = k^2$ it suffices $m > k^3$. Thus we take the following two cases.

- $m > k^3$: As already mentioned, it is not possible to approximate well a discrete distribution with the actual Normal distribution. Thus we turn to discretizations of it. Here we use the Binomial distribution $\mathrm{Bin}(m', q)$ which for large variances is close to Normal. Since the Binomial is a special case of a PBD, we end up with a covering set within the space PBD_n. Studying theorems 5.2.2 and 5.2.3 more closely it is obvious that the Translated Poisson gives a better approximation than the Binomial because of the square root. So we match Z and $Y \sim \mathrm{Bin}(q, m')$ with their respective Translated Poisson distribution first. Then we get an overall bound from the closeness of the Translated Poissons.

Lemma 5.3.3 ([DP15]). *Let $\lfloor \mu_1 - \sigma_1^2 \rfloor \leq \lfloor \mu_2 - \sigma_2^2 \rfloor$. Then it holds that*

$$\mathrm{TV}\left(TP\left(\mu_1, \sigma_1^2\right), TP\left(\mu_2, \sigma_2^2\right)\right) \leq \frac{|\mu_1 - \mu_2|}{\sigma_1} + \frac{|\sigma_1^2 - \sigma_2^2| + 1}{\sigma_1^2} \,.$$

Notice the similarity of this bound to lemma 3.3.2.

Now the problem reduces to appropriately defining the parameters q and m'. For the resulting Binomial to be close to Z we should match their mean value and variance. This is expected to work since the bounds in TV-distance depend only on these two moments. Thus we set

$$m' = \left\lceil \frac{(\sum_{i=1}^{n} z_i)^2}{\sum_{i=1}^{n} z_i^2} \right\rceil$$

and

$$q = \frac{\ell}{n} \,,$$

where ℓ is such that $\frac{\sum_{i=1}^{n} z_i}{m'} \in [\frac{\ell-1}{n}, \frac{\ell}{n}]$. It can be shown that these parameters are well defined and result in good approximations for the mean value and variance of Z. In fact we get the following lemma.

Lemma 5.3.4 (Lemma 4 [DP15]). *Let $\mu = \sum_{z_i \notin \{0,1\}} z_i$, $\mu' = m'q$, $\sigma^2 = \sum_{z_i \notin \{0,1\}} z_i(1 - z_i)$ and $\sigma'^2 = m'q(1-q)$. Then it holds that:*

$$\mu \leq \mu' \leq \mu + 1 \,, \quad \mu \geq k^2 \,,$$

$$\sigma^2 - 1 \leq \sigma'^2 \leq \sigma^2 + 2 \,, \quad \sigma^2 \geq k^2(1 - \frac{1}{k}) \,.$$

Now applying theorem 5.2.2 and lemma 5.3.3, we can derive:

$$\mathrm{TV}\left(Z, Y\right) \leq 9/k \,.$$

Thus the distance between X and its final approximation Y in this case is 16ε at most.

- $m \leq k^3$: In this case the variance of the distribution is small and a Normal approximation would be sub-optimal. Thus, none of the above approximations works for the low variance case. We have already reduced the number of parameters from n to $m = O\left(()\, k^3\right)$ since we have m non-trivial p_i values. The aim is to end up with a set of $O(k)$ different parameters.

  This will be achieved by a 'local' approximation of z_i-s by Binomial distributions, using theorem 5.2.3. That is, we group z_i-s that have similar values. Recall a previous observation based on [TT19]. We noted that the variance of a PBD increases as the parameters become similar. Thus these groups give PBDs with enough variance to be approximated by a Binomial distribution.

  Since the purpose is to reduce the parameters to $O(k)$ we will create $O(k)$ such groups. In fact, we partition the interval $(0, 1)$ (actually $[1/l, 1 - 1/k]$) into $O(k)$ subintervals since our parameters take their values in it. Then the $z - i$-s in each interval are bound to be close. What is more, we can afford a $O(1/k^2)$ error in our 'sub-approximation' by lemma 5.3.1. If we have k groups with $1/k^2$ error each, the final error is bound to be at most $1/k$ by the triangle inequality.

So far so good, but how do we divide $[1/k, 1-1/k]$? Is a trivial, equidistant partition enough? In truth, it is not! To see this, we must return to fig. 5.1 and the property next to it. A PBD has strong concentration resulting in a slow change of values around the mean and a large one in the tails. As must be clear by now, the mass of the tails of the distribution is controlled by the z_i-s close to 0 or 1 (recall our discussion in Stage 1 on how the number of 0 and 1 p_i-s determines the support). Putting all together, the z_i-s close to the border of the interval $[1/k, 1-1/k]$ will have large variance PBDs and the partition can be denser close to them (include less values as they will give the desired variance to approximate a Binomial easily). Thus, the subinterval's length increases around 1/2.

We will not be bothered by the technical details of the partition here. Our goal was to give an overall understanding of how the k-sparse form takes this description as well as why this 'rounding' works. The interested reader can refer to [DP15] for the full analysis.

□

5.4 Learning PBDs

In this last section, the learnability of PBDs in TV-distance is established. As already revealed, our aim is to study the learnability of PBDs from truncated samples. Thus the traditional case should be the first step. For what follows, consider $X \sim PBD(p_1, \dots, p_n)$.

Recall that that a PBD is a discrete distribution, supported on $[n]$. There is an obvious way for learning such distributions: estimate every $p_i, 0 \leq i \leq n$. This demands $m = \Theta(n/\varepsilon^2)$ samples for ε-closeness in TV-distance. The formal proof should be an exercise by now. We refer the lazy reader to [Can20].

However this n dependent sample complexity need not be optimal. The specific structure of a PBD entices us to hope for more. Thinking about this structure, recall that a PBD is unimodal. It is another well-established fact, that discrete, unimodal distributions are learnable given

$$O\left(\frac{\log n}{\varepsilon^3}\log\left(1/\delta\right) + \frac{1}{\varepsilon^2}\log\left(1/\delta\right)\log\log\left(1/\delta\right)\right)$$

samples. This is proven in Birgé's [Bir97] and expressed in this modern form in [DDS15]. Birgé's approach for distribution learning differs from what we encountered so far in this thesis. Indeed Birgé and the related literature offer us a glimpse of non-parametric distribution estimation, where other properties of distribution functions, such as unimodality, as exploited. Back to our problem, notice the improvement in dependence over n from linear to logarithmic. This was gained employing our knowledge for the unimodal structure of the distribution.

Still unimodality does not exhaust the information about PBDs. Thus we can do even better. In this next step, the dependence over n for the sample complexity is completely removed. The result was given by Daskalakis, Diakonikolas and Servedio in [DDS15] and will be presented in the following. As for a little spoiler, let us note that the best characterization for a PBD's structure is no other from its ε-cover.

Theorem 5.4.1 (Learning PBDs [DDS15]). *Let $X = \sum_{i=1}^{n} X_i$ be an unknown PBD. There is an efficient algorithm that given*

$$m = \tilde{O}\left(1/\varepsilon^2\right) \log 1/\delta$$

independent samples from X returns a vector $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_n)$ such that $\hat{X} \sim PBD(\hat{p})$ satisfies:

$$\mathrm{TV}\left(X, \hat{X}\right) < \varepsilon\,,$$

with probability at least $1 - \delta$.

In the following we present that algorithm and sketch the proof of theorem 5.4.1. Fully exploiting our knowledge for the PBD structure we reduce the problem to learning a sparse PBD, a heavy PBD (Binomial) and choosing between the two. This is demonstrated in the following algorithm.

Algorithm 4 LearnPBD

```
procedure LearnPBD(n, ε, δ)
    $H_S$ ← LearnSparse(n, ε, δ/3).
    $H_P$ ← LearnHeavy(n, ε, δ/3).
    return ChooseHypothesis($H_S, H_P, \varepsilon, \delta/3$).
```

That is, if the unknown PBD is close to a sparse, the subroutine for learning a sparse PBD -LearnSparse- will succeed. If it is not, it must be close to a heavy PBD. Thus the correct answer is given by LearnHeavy. Finally, we decide on which is actually the case, choosing between the two suggested hypotheses.

The rest of this section aims to give a high-level understanding of how these subroutines work. We study each in a subsection below.

5.4.1 Learn Sparse PBD - Birgé's algorithm

The sparse case will be settled by unimodality. Note that this, and the sparsity property, is the only structural information concerning this case. Sparsity is also deployed to remove the n dependence. So we consider this an 'optimal' approach.

Birgé's algorithm, as given in Theorem 5 of [DDS15], guarantees that with

$$O\left(\frac{\log n}{\varepsilon^3} \log\left(1/\delta\right) + \frac{1}{\varepsilon^2} \log\left(1/\delta\right) \log\log\left(1/\delta\right)\right)$$

samples it returns a distribution H over $[n]$ such that $\mathrm{TV}(X, H) \leq \varepsilon$. However, for the sparse case, the effective mass of a PBD lies in an interval $[a, b]$ such that $|b - a| \leq \frac{1}{\varepsilon^3}$. This it immediate by *Stage 2* of the cover proof. Thus, Birgé's algorithm's sample complexity is now independent of n, i.e.

$$m = O\left(\frac{1}{\varepsilon^3} \log\left(1/\varepsilon\right) \log\left(1/\delta\right) + \frac{1}{\varepsilon^2} \log\left(1/\delta\right) \log\log\left(1/\delta\right)\right)\,.$$

Still a subtle detail is missing. In theorem's 5.3.1 proof, there is also a *Stage 1*, the massage step. That is, to work with the 'effective' mass of the distribution that is supported on $1/\varepsilon^3$ elements, we must remove ε-mass of the tails. Consider that given sample access to a sparse PBD X we run Birgé's algorithm. Then there is an ε probability that we get samples outside $[a, b]$. Then the guaranteed, independent of n, sample complexity is spoiled.

To overcome this issue, we truncate the distribution tails. In other words, we give Birgé's algorithm access to a conditioned version of X on the interval $[a, b]$, denote $X_{[a,b]}$. In contrast to *Stage 1*, we cannot explicitly specify a, b by mere sample access of the distribution. Thus, we estimate them by the following procedure:

- Draw $M = 32 \log{(8/\delta)}/\varepsilon^2$ samples.
- Set a to be *close to* the minimum and b *close to* the maximum of these samples.

We will specifically define *close to* right away. The procedure will be formulated for a since b can be handled similarly. Note that we cannot estimate an a that leaves exactly ε-mass out, so we bound the mass it excludes.

Claim 5.4.1. *Consider $X \sim PBD(p_1, \ldots, p_n)$ and sample access of X. Let $\mathcal{S} = \{s_1, \ldots, s_M\}$ be the sorted set of $M = \frac{32\log{(8/\delta)}}{\varepsilon^2}$ i.i.d. samples of X. Also, let $\hat{a} = s_{\lceil 2\varepsilon M \rceil}$. Then*

$$\Pr\left[X \leq \hat{a}\right] \in \left[3\varepsilon/2, 5\varepsilon/2\right] \; ,$$

with probability at least $1 - \delta/4$

Proof. The above claim follows by a simple concentration observation. Let $a_l = \max\{i | \Pr[X \leq i] \leq 3\varepsilon/2\}$, thus $\Pr[X \leq a_l] \leq \frac{3\varepsilon}{2}$. Note that we cannot calculate this a_l. Assuming we know it, however, consider the number of samples in $\mathcal{S}$ below a_l, i.e.

$$\sum_{i=1}^{M} \mathbf{1}_{s_i \leq a_l} \, .$$

Notice that

$$\mathbb{E}\left[\sum_{i=1}^{M} \mathbf{1}_{s_i \leq a_l}\right] \leq 3\varepsilon M/2 \, .$$

By Hoeffding's Inequality 3.2.1 and for

$$t = \frac{7\varepsilon M}{4} - \mathbb{E}\left[\sum_{i=1}^{M} \mathbf{1}_{s_i \leq a_l}\right] > \varepsilon M/4$$

we get

$$\Pr\left[\sum_{i=1}^{M} \mathbf{1}_{s_i \leq a_l} > \frac{7\varepsilon M}{4}\right] \leq \delta/8 \, .$$

Thus, with high probability, the $2\varepsilon M > \frac{7\varepsilon M}{4}$ element is larger than a_l. So $\Pr[X \leq s_{\lceil 2\varepsilon M \rceil}] \geq 3\varepsilon/2$.

An equivalent procedure is followed for the upper bound. Consider $a_u = \min\{i | \Pr[X \leq i] > 5\varepsilon/2\}$ so $\Pr[X \leq a_u] > 5\varepsilon/2$. We will now work with $-\sum_{i=1}^{M} \mathbf{1}_{s_i \leq a_u}$ for which it holds

$$-\mathbb{E}\left[-\sum_{i=1}^{M} \mathbf{1}_{s_i \leq a_u}\right] > 5\varepsilon M/2\,.$$

Thus, for

$$t = -\frac{9\varepsilon M}{4} + \mathbb{E}\left[\sum_{i=1}^{M} \mathbf{1}_{s_i \leq a_l}\right] > \varepsilon M/4\,,$$

Hoeffding's inequality implies

$$\Pr\left[\sum_{i=1}^{M} \mathbf{1}_{s_i \leq a_u} < \frac{9\varepsilon M}{4}\right] \leq \delta/8\,.$$

So $2\varepsilon M < \frac{9\varepsilon M}{4}$ element is smaller than a_u and $\Pr[X \leq s_{\lceil 2\varepsilon M\rceil}] \leq 5\varepsilon/2$ as required. □

Algorithm 5 Learn Sparse PBD

```
procedure LearnSparse(n, ε, δ)                    ▷ Assume oracle access in X's distribution
    $M \leftarrow 32\log(8/\delta)/\varepsilon^2$
    $\mathcal{S} \leftarrow \{s_1, \dots, s_M\} \sim_{sort} X^M$.    ▷ M i.i.d. sorted samples from X's distribution
    $a \leftarrow s_{\lceil 2\varepsilon M\rceil}$ and $b \leftarrow s_{\lfloor (1-2\varepsilon)M\rfloor}$.
    if $b - a > (C/\varepsilon)^3$ then                          ▷ C consant
        return $\hat{\mathbf{p}} = \mathbf{0}$.
    else
        return Birgé($X_{[a,b]}$).                ▷ Birgé's algorithm for X conditional on [a, b]
    return ChooseHypothesis($H_S, H_P, \varepsilon, \delta/3$).
```

Conditioning on estimator $\hat{a}$ we can guarantee we excluded approximately ε-mass of the PBD distribution. Moreover, recall that we do not really know X is a sparse PBD. That means, if it is not, Birgé's algorithm will demand samples that depend on n. The estimators $\hat{a}$, $\hat{b}$ can, therefore, be used to reject PBDs that are not close to sparse. Algorithm 5 puts everything together.

5.4.2 Learn Heavy PBD

We will now assume that X is ε-close to a heavy Binomial distribution. As emphasized in section 4.2 this result derives from the proximity of PBD to a Gaussian distribution. The exact argument in *Stage 2* states that a PBD is ε-close to a Translated Poisson of the same mean μ and variance σ^2 (and through this it is ε-close to the Binomial). Thus it suffices to estimate the mean $\hat{\mu}$ and the variance $\hat{\sigma}^2$ of the PBD. Then the Translated Poisson with the estimated parameters must be close to the original Translated Poisson.

The problem reduces to finding the correct estimators. It is no surprise that they should be

$$\hat{\mu} = \frac{1}{m}\sum Z_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{m-1}\left(Z_i - \hat{\mu}\right)^2\,,$$

where Z_i denotes the samples and m the number of samples drawn.

Algorithm 6 LearnHeavy

```
function EmpiricalEstimators(n, ε, δ) ▷ Assume oracle access in $X$'s distribution
    $r \leftarrow O(\log(1/\delta))$.
    for $i \in [1:r]$ do
        $m \leftarrow \lceil 3/\varepsilon^2 \rceil$.
        $\mathcal{S} = \{Z_1, \dots, Z_m\} \sim X^m$.    ▷ $m$ i.i.d. samples from $X$'s distribution
        $\hat{\mu}_i \leftarrow \frac{\sum Z_j}{m}$ and $\hat{\sigma}_i^2 \leftarrow \frac{\sum (Z_j - \hat{\mu}_i)^2}{m-1}$.
    $\hat{\mu} \leftarrow \text{median}(\hat{\mu}_1, \dots, \hat{\mu}_r)$ and $\hat{\sigma}^2 \leftarrow \text{median}(\hat{\sigma}_1^2, \dots, \hat{\sigma}_r^2)$.
    return $(\hat{\mu}, \hat{\sigma}^2)$.
procedure LearnHeavy(n, ε, δ)
    $\varepsilon' \leftarrow \varepsilon / \sqrt{4 + \frac{1}{\theta^2}}$.
    $(\hat{\mu}, \hat{\sigma}^2) \leftarrow$ EmpiricalEstimators(n, ε', δ).
    return $TP(\hat{\mu}, \hat{\sigma}^2)$.
```

However, these estimators are not exactly correct. The problem here is that Z_i-s do not really follow a Gaussian distribution. Thus the sub-Gaussianity property and concentration results are not quite right. Note that Translated Poisson is a Poisson distribution. Thus, it has a heavier tail than the Gaussian.

The aforementioned estimators are 'weak'. That is, their probability of being away from their mean is not negligible, but is small. It can be shown that

$$\Pr\left[|\hat{\mu} - \mu| > \sigma\varepsilon\right] \le 1/3$$

and

$$\Pr\left[\left|\frac{\hat{\sigma}^2}{\sigma^2} - 1\right| > \varepsilon^2\right] \le 1/3\,.$$

To put it right, there is a *bias* towards getting estimators that are close to the true value. This is enough. Applying a boosting technique, such as in lemma 4.1.4, we end up with a precise estimator. This technique consists of repeating the estimation a number of times and choosing a common outcome. This must work, since most of the time the estimator is good.

Algorithm 6 employs the described procedure. We will not proceed into the formal analysis. It is given at section 2.2 of [DDS15] in enough detail. Note, however, that despite our understanding for the Gaussian distribution and its proximity to a heavy PBD, proving the same results for the later can be tricky.

5.4.3 PBD Hypothesis Testing

We must now decide which distribution is indeed close to X. We need a criterion to compare the two. Notice that, if the two distributions are close in TV-distance and since one of them must be close to X, any output is correct. Thus, the actual need to choose occurs when the two distributions are away from each other. That means hypotheses H_1, H_2 differ in TV-distance and so, they assign different masses to some subsets ofthe domain. As can be seen in algorithm 7, this subset is specified (we actually know the distributions so we can calculate their p.m.f.). Then the true mass on this subset is estimated. The distribution closest to this estimation wins and is returned.

Algorithm 7 ChooseHypothesis

```
procedure ChooseHypothesis($H_1$, $H_2$, ε, δ)    ▷ Assume oracle access in $X$'s distribution
    $W \leftarrow$ support($X$), $W_1 = \{w \in W | H_1(w) > H_2(w)\}$.
    $p_1 = H_1(W_1)$ and $p_2 = H_2(W_1)$.    ▷ $p_1 > p_2$ and $TV(H_1, H_2) = p_1 - p_2$
    if $p_1 - p_2 \leq 5\varepsilon$ then    ▷ Draw, return either
        return $H_1$.
    else
        $m \leftarrow 2\frac{\log(1/\delta)}{\varepsilon^2}$
        $\mathcal{S} = \{s_1, \ldots, s_m\} \sim X^m$.    ▷ $m$ i.i.d. samples from $X$'s distrbution
        $\tau \leftarrow \frac{1}{m}|\{i | s_i \in W_1\}|$.
        if $\tau > p_1 - \frac{3}{2}\varepsilon$ then
            return $H_1$.
        if $\tau < p_2 + \frac{3}{2}\varepsilon$ then
            return $H_2$.
    return $H_2$.    ▷ If not any of the above we have a draw and return either
```

Chapter 6

Learning PBD from Truncated Samples

In this chapter we study the learnability of PBDs from samples truncated on a set S. This is one of the two distributions that are studied in this thesis. In agreement with the non-truncated case, we exploit the information given by a PBD's cover. So the problem is reduced in two subproblems: learning a heavy binomial distribution and learning a sparse PBD from truncated samples. In what follows, we study these subproblems in depth and try to specify those traits that will make efficient learnability possible.

6.1 Identifiability

In this section we study the identifiability of truncated PBDs. Recall that a PBD is fully characterized by its parameters $\mathbf{p}$ (for a specific number of parameters n). That is, every vector $\mathbf{p}$ creates a unique PBD. When truncating the distribution, however, this is no longer guaranteed. The first step for studying learnability from truncated samples is identifiability. That means, no two different distributions truncated on a set result in the same truncated distribution. Thus, we need to give a condition for the truncation set S so as any PBD truncated on it is unique. For example, the Binomial distribution truncated on a set with more than three elements is unique, as will be proven in the next section.

In the case of a general PBD this cause is quite challenging. The following lemma shows that the set of PBDs of any order will include at least two distributions that are not identifiable by their truncation on a set S. What is more, this S only hides one point of the distributions' support and it has at least $1/2$ mass. Thus it is impossible to construct an algorithm that guarantees to learn any PBD truncated on any set S.

Lemma 6.1.1. *There exist a truncation set S and at least two distributions in the set of* PBD_n *with non trivial mass on S such that their truncations on S are identical.*

Proof. It suffices to show that there exist a set S and two PBDs Y, Z such that $\mathrm{TV}(Y, Z) > \varepsilon$ and $\mathrm{TV}(Y_S, Z_S) < \varepsilon$.

We assume $n > 5$. Then let

$$Y \sim \mathrm{PBD}(0.9, 0.9, 0.9, 0.9, 0.7, 0, \ldots)\,,$$

$$Z \sim \mathrm{PBD}(0.95, 0.95, 0.8, 0.3, 0.95, 0, \ldots)\,,$$

and $S = [5]$. In fig. 2.1 we can see that although S removes just one point from the distribution and it has at least 1/2 mass the truncation of the distributions on it are identical. □

Note that the previous lemma 6.1.1 does not prohibit the learnability of some subset of the set of PBDs. For example, the set of 'heavy' PBDs is learnable from truncated samples as argued in section 6.2. Moreover, in [DKS16] a robust algorithm for learning PBDs is given. Thus a truncation set S that removes only an ε-mass of the PBD does not affect its learnability.

In the sparse case, however, lemma 6.1.1 guarantees that there always be at least two distributions and a set S on which the supposing learning algorithm will always fail. In the next section, we restrict ourselves to PBDs that are close to heavy Binomials. That is, PBDs that are close to a Gaussian-continuous structure.

6.2 Learning a Heavy Binomial Distribution from Truncated Samples

As has been highlighted in the previous chapter, a heavy PBD is, in principle, close to a Gaussian distribution. In theorem 2.0.1 we give an algorithm that efficiently learns a Gaussian from truncated samples. So, it would appear that the first problem is already solved.

However, the two distributions, a heavy PBD and its ε-close Gaussian, have a TV distance. Thus, sampling from a PBD is not equivalent to sampling from a Gaussian, however close they might be. Recalling that algorithm 3 is, in principle, the SGD, it is not straight-forward that it should work given samples with ε-error. Indeed, the key characteristic of SGD is that it minimizes the objective given a suitable estimation of its derivative. Here our objective is the Gaussian negative log-likelihood. So there will be an error in this derivative estimation, since our samples come from a PBD distribution. In Approach One we study this technique and give experimental results on its performance.

On the other hand, theorem 2.0.1 is a powerful technique designed to address the problem in many dimensions. One should reasonably think, that a single dimensional Gaussian, close to a heavy Binomial distribution, should be easier to learn, even from truncated samples. With this in mind, we study the learnability of a Binomial distribution from truncated samples in Approaches Two and Three.

6.2.1 Approach One: Discretized Gaussians

First, we want to demonstrate the equivalence of learning a heavy PBD to learning a discretized Gaussian distribution. This follows directly from the closeness of a heavy PBD to a heavy Binomial and the closeness of a heavy Binomial to a discretized Gaussian. This is an observation emphasized throughout the thesis. In lemma 6.2.1 it is formally stated.

Lemma 6.2.1. *Let $X_1, X_2, \ldots, X_n$ be arbitrary mutually independent indicators and $\epsilon \in (0, 1/2)$. Assume there exists a binomial distribution $\mathrm{Bin}(k, p)$ with $k \in [1 : n]$ and $p \in$*

$\{\frac{1}{n}, \dots, \frac{n}{n}\}$ *such that* $kp \geq \frac{1}{\epsilon^2}$ *and* $lp(1-p) \geq \frac{\epsilon^2}{1-\epsilon-\epsilon^2}$ *and for* $X \sim \mathrm{Bin}(k,p)$ *it holds that:*

$$\mathrm{TV}\left(\sum_{i=1}^{n} X_i, X\right) \leq 41\epsilon\,.$$

Then, for $Y \sim N^d_{\mathbb{Z}}(kp, kp(1-p))$ *it holds that:*

$$\mathrm{TV}\left(\sum_{i=1}^{n} X_i, Y\right) \leq O\left(\epsilon\right)\,.$$

Proof. Consider the random variable $Z \sim N^d(kp, kp(1-p))$. Notice that $\epsilon < 1/2$ results in $\sqrt{1-\epsilon-\epsilon^2} > 1/2$. From theorem 5.2.4, it follows that:

$$\mathrm{TV}\left(\sum_{i=1}^{n} X_i, Z\right) \leq 15.2\epsilon\,.$$

An application of lemma 5.2.1 and the triangle inequality gives:

$$\mathrm{TV}\left(\sum_{i=1}^{n} X_i, Y\right) \leq O\left(\epsilon\right)\,.$$

□

By definition, the $\mathbb{Z}$-discretized normal distribution is the truncation of the normal distribution on the set of all integers $\mathbb{Z}$. So, the truncated $\mathbb{Z}$-discretized normal distribution on a set $S \subseteq \mathbb{Z}$ is the truncation of the normal distribution on S.

Thus, the learnability of the $\mathbb{Z}$-discretized normal distribution, and, consequently, of the heavy PBD, from its truncation on a set $S \subseteq \mathbb{Z}$ directly follows from theorem 2.0.1.

As noted before, this closeness does not guarantee the learnability of the heavy PBD. This would imply robustness of the behaviour of SGD. We leave a formal such statement open for future work. For now, we suffice with studying this conjecture in practice. That is, we produce samples from a heavy Binomial distribution and treat them as Gaussian samples. Note that this technique applies to another problem as well: learning Binomial distributions from truncated samples.

Experiments

For the experiments we have chosen three kinds of truncation sets. The first consists of samples around the mean of the distribution. The second consists of one of the tails. And the third rejects all the mass around the mean. We study these sets leaving only α mass on them for three different mass values. So, we test our distribution on nine different settings.

Our aim is a quite demanding one: we want to retrieve the original distribution in $\varepsilon = 0.08$ TV distance. In the case of the Gaussians, we use $m = 50/(\alpha\varepsilon^2)$ samples. This matches theorem 2.0.1 up to a constant factor, revealed through the experiments. The results are presented in fig. 6.1.

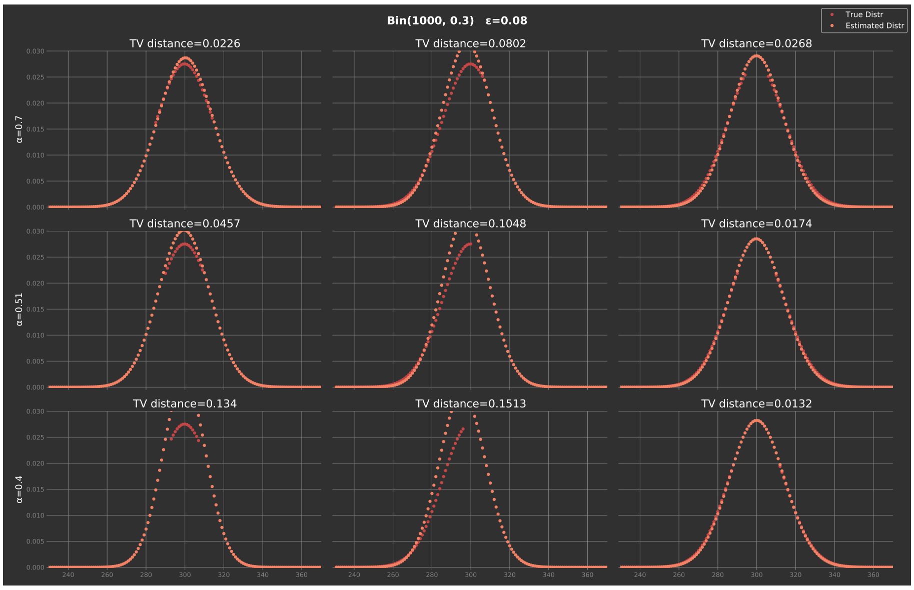

Figure 6.1: PSGD for Gaussians run with Binomial samples using $50/\alpha\varepsilon^2$. The right-hand side inscription α denotes the mass of the truncation set.

The conclusion from fig. 6.1 is encouraging. We can see that in most cases the goal is attained. Even when $\varepsilon = 0.08$ distance is not achieved, the error is not much larger. We ran the same experiments for a different Binomial distribution, as well. The results remain positive as can be seen in fig. 6.2

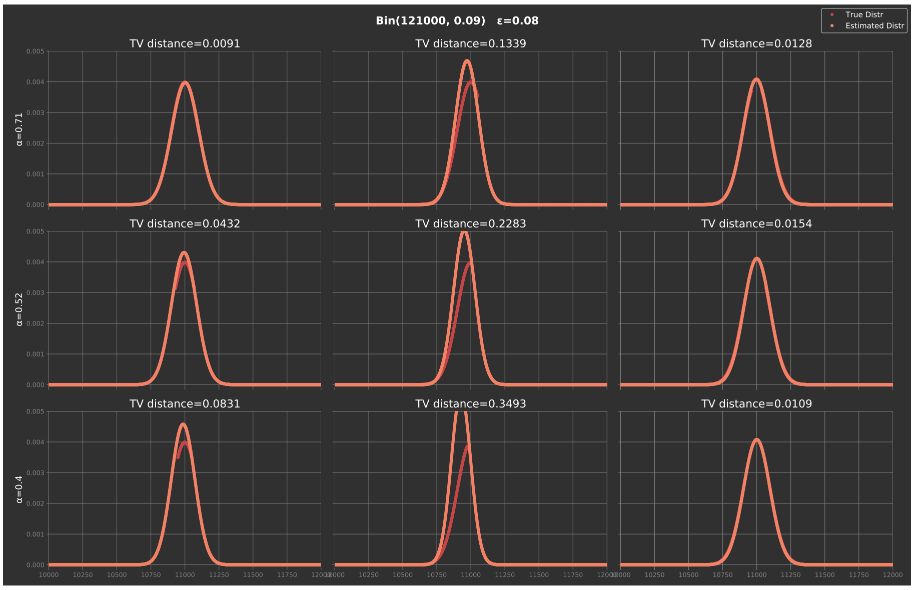

Figure 6.2: PSGD for Gaussians run with Binomial samples using $50/\alpha\varepsilon^2$. The right-hand side inscription α denotes the mass of the truncation set.

6.2.2 Approach Two: Binomial System Solution

The motivation for the following discussion comes from this simple example. Assume we want to learn a Binomial distribution truncated on a set S with only one element.

However, the truncated distribution on S will put all the mass in just one element, making it impossible to distinguish between distributions. This introduces the problem of identifiability. For a distribution to be learnable, given its truncation on S, it is necessary to be identifiable from this truncation. That is, its truncation on the set must be unique.

It is now clear that the first step when studying the learnability of a distribution if to specify whether it is identifiable truncated on the specific set. In the non-truncated case this is usually not necessary since distributions are uniquely defined. In lemma 6.2.2 we prove that, for the Binomial distribution to be identifiable, S must contain at least 3 elements.

Lemma 6.2.2. *A binomial distribution* $\mathrm{Bin}(n, p)$ *is identifiable by its truncation on a set* $S \subseteq [n]$ *iff* $|S| \geq 3$.

Proof. Consider two binomials $\mathrm{Bin}(n, p)$, $\mathrm{Bin}(m, q)$ and let S be the truncation set. Let $\alpha = \Pr_{X\sim\mathrm{Bin}(n,p)}[X \in S]$ be the mass of S in $\mathrm{Bin}(n, p)$. Denote $p_x = \Pr_{X\sim\mathrm{Bin}(n,p)}[X = x]$ the probability mass on a point $x \in S$ by $\mathrm{Bin}(n, p)$ and q_x the mass assigned by $\mathrm{Bin}(m, q)$.

First, we will prove that for $|S| = 2$ the binomial distribution is not identifiable. Set $p = \frac{1}{2}$, $q = \frac{1}{3}$ and $m = 2n - x$. Assume $S = \{x, x + 1\}$ such that $S \subset [min(n, m)]$.

Notice that:

$$\frac{p_x}{p_{x+1}} = \frac{\binom{n}{x}(\frac{1}{2})^x(1-\frac{1}{2})^{n-x}}{\binom{n}{x+1}(\frac{1}{2})^{x+1}(1-\frac{1}{2})^{n-x-1}} = \frac{x+1}{n-x} = b\,,$$

and

$$\frac{q_x}{q_{x+1}} = \frac{\binom{m}{x}(\frac{1}{3})^x(1-\frac{1}{3})^{m-x}}{\binom{m}{x+1}(\frac{1}{3})^{x+1}(1-\frac{1}{3})^{m-x-1}} = 2\frac{x+1}{2n-2x} = \frac{x+1}{n-x} = b\,.$$

Thus,

$$\Pr_{X\sim\mathrm{Bin}_S(n,p)}[X = x] = \frac{p_X(x)}{p_X(x) + p_X(x+1)} = \frac{\alpha}{1+\alpha} = \frac{p_Y(x)}{p_Y(x) + p_Y(x+1)} = \Pr_{Y\sim\mathrm{Bin}_S(m,q)}[Y = x]\ ,$$

and, similarly, for $x + 1$. Thus, for $\mathrm{Bin}(n, p)$ to be identifiable, the cardinality of S must be at least 2.

Now, we will prove that the probability mass of the truncated distribution on three elements can uniquely determine the original distribution. Let $S = \{x, y, z\}$. Assume, by way of contradiction, that the mass assigned on each point of S by the truncated distributions $\mathrm{Bin}_S(n, p)$ and $\mathrm{Bin}_S(m, q)$ are equal, i.e.

$$\frac{p_x}{p_x + p_y + p_z} = \frac{q_x}{q_x + q_y + q_z} \quad\text{and}\quad \frac{p_y}{p_x + p_y + p_z} = \frac{q_y}{q_x + q_y + q_z} \quad\text{and}\quad \frac{p_z}{p_x + p_y + p_z} = \frac{q_z}{q_x + q_y + q_z}\,.$$

This implies:

$$\frac{p_x}{p_y} = \frac{q_x}{q_y} \quad\text{and}\quad \frac{p_x}{p_z} = \frac{q_x}{q_z}\,.$$

By definition, we get:

$$\frac{\binom{n}{x}}{\binom{n}{y}}\left(\frac{p}{1-p}\right)^{x-y} = \frac{\binom{m}{x}}{\binom{m}{y}}\left(\frac{q}{1-q}\right)^{x-y} \quad\text{and}\quad \frac{\binom{n}{x}}{\binom{n}{z}}\left(\frac{p}{1-p}\right)^{x-z} = \frac{\binom{m}{x}}{\binom{m}{z}}\left(\frac{q}{1-q}\right)^{x-z}\,,$$

and, thus,

$$\left(\frac{q(1-p)}{p(1-q)}\right)^{x-y} = \frac{\binom{n}{x}\binom{m}{y}}{\binom{n}{y}\binom{m}{x}} \quad\text{and}\quad \left[\frac{\binom{n}{x}\binom{m}{z}}{\binom{n}{z}\binom{m}{x}}\right]^{x-y} = \left[\frac{\binom{n}{x}\binom{m}{y}}{\binom{n}{y}\binom{m}{x}}\right]^{x-z}\,.$$

Setting $c_1 = x - y > 0$ and $c_2 = x - z > 0$ and after some calculations the second relation becomes:

$$\frac{[(n-z)\dots(n-z-c_2+1)]^{c_1}}{[(n-z-c_1+c_2)\dots(n-z-c_2+1)]^{c_2}} = \frac{[(m-z)\dots(m-z-c_2+1)]^{c_1}}{[(m-z-c_1+c_2)\dots(m-z-c_2+1)]^{c_2}} .$$

Notice that the nominator has c_2 terms while the denominator has $2c_2 - c_1$. If $c_2 > 2c_2 - c_1$ it follows that $c1 > c_2$ and fraction increases with n. On the other hand, when $c_2 < 2c_2 - c_1$ the opposite holds. Note that, since x, y, z are different points it is always $c_1 \neq c_2$. In either case, the equality holds iff $n = m$ and we reached a contradiction. □

Based on the above analysis, we develop an algorithm to learn the parameter p given the probability mass on just two points of the domain. Let $x, y \in S$ and p_x, p_y the masses on x, y respectively. Then we have:

$$\frac{p_x}{p_y} = \frac{\binom{n}{x} p^x (1-p)^{n-x} / p(S)}{\binom{n}{y} p^y (1-p)^{n-y} / p(S)} = \frac{\binom{n}{x}}{\binom{n}{y}} \left(\frac{p}{1-p}\right)^{x-y} .$$

So we get for p:

$$p = \left(1 + \left(\frac{p_x\binom{n}{y}}{p_y\binom{n}{x}}\right)^{x-y}\right) .$$

Thus algorithm 8 should give a good estimation for p.

Algorithm 8 System Solution

```
procedure ESTIMATEPSYSTEMSOLVING(M)                 ▷ Assume oracle access to S
    $x, y \leftarrow$ ChoosePoints($S$).             ▷ Decide on estimation points
    $\hat{p}_x, \hat{p}_y \leftarrow$ EmpiricalEstimation($M$).
    $\hat{p} \leftarrow [1 + [\frac{\hat{p}_x\binom{n}{y}}{\hat{p}_y\binom{n}{x}}]^{x-y}]^{-1}$.
    return $\hat{p}$
```

We make two observations about algorithm 8. The implementation of function 'ChoosePoints' is not specified. The following analysis gives information about the importance of this function. Thus based on it we decide how it would be best implemented. The 'EmpiricalEstimation' function simply counts the x-samples in sample set of size M.

Moreover, we insist on the following remark. The above algorithm is a quite simple algorithm. It depends on the mass of only two points to calculate p. Therefore, it ignores a lot of information in a sample set, referring to the rest of the domain.

Study function ESTIMATEPSYSTEMSOLVING

In lemma 6.2.3 we formally analyze the sample complexity of algorithm 8.

Lemma 6.2.3 (Learnability of Truncated Binomial Distribution - System Solving)**.** *Consider a binomial distribution* $\mathrm{Bin}(n, p)$*, where n is fixed, and a set $S \subseteq [n]$, such that $|S| \geq 2$. Denote b be the minimum probability mass assigned on a point in S by the truncated binomial distribution* $\mathrm{Bin}_S(n, p)$*, i.e.*

$$b = min_{x \in S} \Pr_{X \sim \mathrm{Bin}_S(n,p)} [X = x] .$$

Then, there exists an algorithm that given $m = \Theta(\frac{n \log{(1/\delta)}}{b^2\varepsilon^2})$ *i.i.d. samples from* $\mathrm{Bin}_S(n,p)$*, computes an estimate* $\hat{p}$ *such that:*

$$\mathrm{TV}\left(\mathrm{Bin}(n,p), \mathrm{Bin}(n,\hat{p})\right) \leq \varepsilon\,,$$

with probability at least $1-\delta$ *and for* $\varepsilon < b$*.*

Proof. Consider the set $S \subset [n]$ and the truncated binomial distribution $\mathrm{Bin}_S(n,p)$.

Denote $p_x = \Pr_{X\sim \mathrm{Bin}_S(n,p)}[X = x]$ the probability mass on a point $x \in S$ and $\hat{p}_x$ its estimate. The algorithm computes the estimates for two points, $x, y \in S$.

Define $\hat{p}_x = \frac{1}{m}\sum_{i=1}^m X_i$ where $X_i \sim Be(p_x)$, i.e. X_i is 1 when sample i equals x, 0 otherwise. Similarly, $\hat{p}_y = \frac{1}{m}\sum_{i=1}^m Y_i$. Then, $\mathbb{E}[X] = p_x$ and $\mathbb{E}[Y] = p_y$.

By Hoeffding's inequality we can derive:

$$\Pr\left[|\hat{p}_x - \mathbb{E}[X]| \geq \varepsilon\right] \leq 2\exp\left(-2m\varepsilon^2\right).$$

It follows that given

$$m = \frac{\log{(4/\delta)}}{2\varepsilon^2}$$

samples it holds that:

$$|\hat{p}_x - p_x| < \varepsilon \quad \text{and} \quad |\hat{p}_y - p_y| < \varepsilon$$

with probability at least $1-\delta$.

Considering $\mathrm{Bin}(n,p)$ as an exponential family, denote θ its natural parameter. Then, θ is given in terms of the probabilities p_x, p_y:

$$\theta = \frac{1}{x-y}\cdot \log\left(\frac{p_x\binom{n}{y}}{p_y\binom{n}{x}}\right).$$

Then, it holds:

$$|\theta - \hat{\theta}| = \frac{1}{|x-y|}\cdot\left|\log\left(\frac{p_x}{p_y}\right) - \log\left(\frac{\hat{p}_x}{\hat{p}_y}\right)\right| = \frac{1}{|x-y|}\cdot\left|\log{(p_x)} - \log{(\hat{p}_x)} + \log{(\hat{p}_y)} - \log{(p_y)}\right|.$$

Notice that $f(x) = \log x$ is $\frac{1}{C}-$Lipschitz if $x \geq C$. Since b denotes the minimum probability assigned by distribution $\mathrm{Bin}_S(n,p)$ we get that:

$$p_x \geq b \quad \text{and} \quad \hat{p}_x \geq b - \varepsilon\,.$$

Also, it holds that $|x-y| \geq 1$, so it follows:

$$|\theta - \hat{\theta}| \leq \frac{1}{b-\varepsilon}\cdot\left(|p_x - \hat{p}_x| + |p_y - \hat{p}_y|\right) \leq \frac{2\varepsilon}{b-\varepsilon}\,.$$

Thus, by 3.3.1, we get that:

$$\mathrm{TV}\left(\mathrm{Bin}(n,p), \mathrm{Bin}(n,\hat{p})\right) \leq \frac{\sqrt{2n}}{2}\cdot|\theta - \hat{\theta}| \leq \frac{\varepsilon\sqrt{2n}}{b-\varepsilon}\,.$$

So, for

$$m = \Theta(\frac{n\log{(1/\delta)}}{b^2\varepsilon^2})$$

it follows that:

$$\mathrm{TV}\left(\mathrm{Bin}(n,p), \mathrm{Bin}(n,\hat{p})\right) \leq \varepsilon$$

with probability at least $1-\delta$ and the proof is concluded. □

Notice that the sample complexity for algorithm 8 depends on n. This is not an expected result. In Approach One we imply that a Binomial distribution is learnable given just $\Theta(1/\varepsilon^2)$ samples. Here, simply retrieving p costs samples that increase linearly with n. We want to study whether this is a byproduct of our analysis or a real characteristic of the algorithm. Thus we proceed on the following experiments.

Experiments

First, we specify the exact implementation of the algorithm. In algorithm 8 'Choose-Points' function remains 'open'. In lemma 6.2.3 the sample complexity depends on the least mass of these points. Thus we should choose points that will likely have important mass. We use a small portion of the samples (1/7) to get estimations on the mass of every point in S. Then we choose the points with highest estimations. Although these estimations might be very bad, we just want a guidance for our choice.

The experiments are identical to the Gaussian case. We use $m = n/(\alpha\varepsilon^2)$ samples for our estimation. We get fig. 6.3.

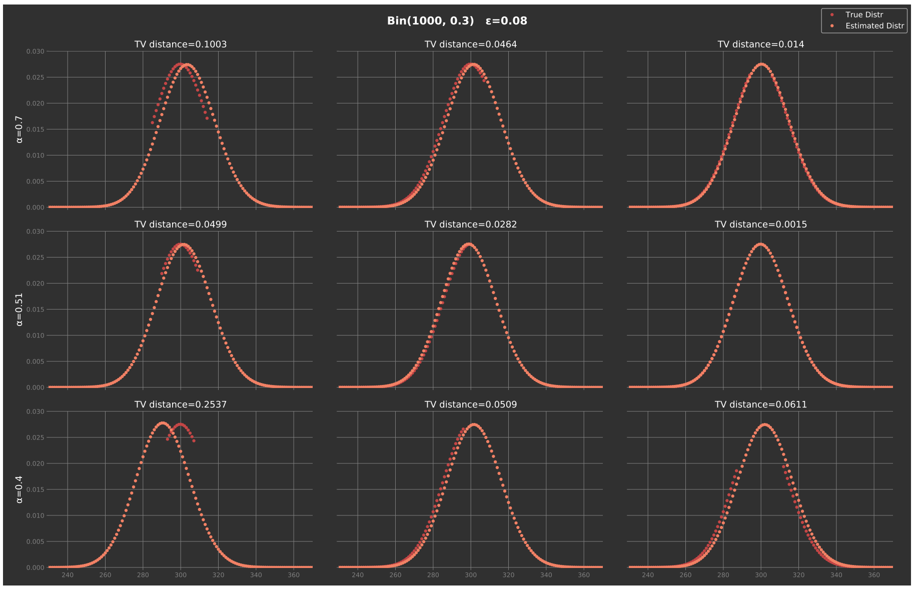

Figure 6.3: Learn parameter p of Binomial using $n/(\alpha\varepsilon^2)$ samples. The right-hand side inscription α denotes the mass of the truncation set.

Notice that even using n times more samples than in the Gaussian case, algorithm 8 fails to retrieve p close enough.

We repeat the experiment allowing for only $m = 1/(\alpha\varepsilon^2)$ samples. This is to emphasize on the sub-optimality of this algorithm compared to Approach One. At the same time, we are reassured for the analysis of lemma 6.2.3. The results are given in fig. 6.4.

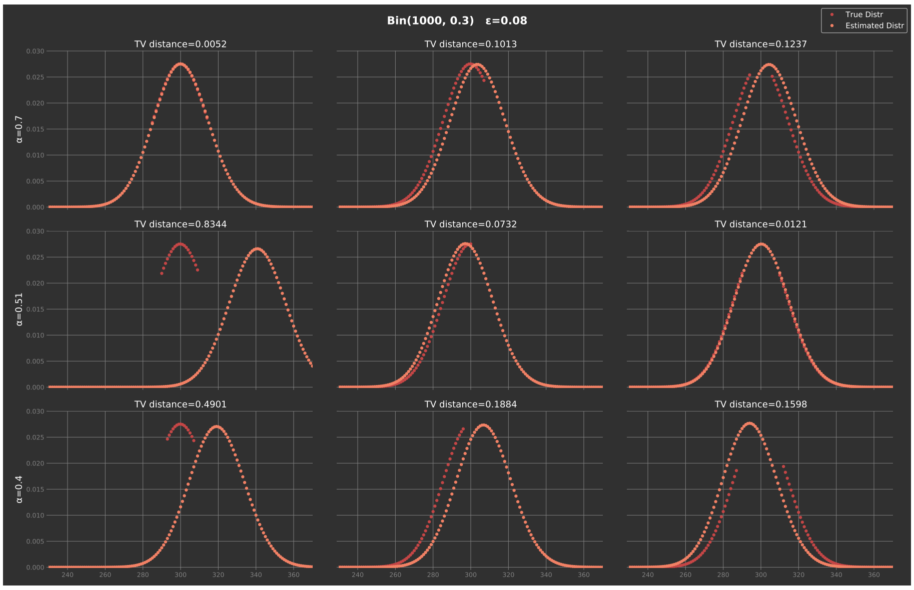

Figure 6.4: Learn parameter p of Binomial using $1/(\alpha\varepsilon^2)$ samples. The right-hand side inscription α denotes the mass of the truncation set.

It is a safe conclusion that, ignoring information, the algorithm demands more samples to guarantee small error. Thus, lemma 6.2.3 applies in practice and a 'simpler' approach to learning the Binomial distribution seems inadequate.

6.2.3 Approach Three: PSGD for Binomial

In this final Approach we take advantage of all the information available to estimate p. Note that the Binomial distribution with known parameter n is an exponential family (see table 3.1). Thus it has a convex negative log-likelihood objective. We can prove that algorithm 3, that is basedon learning a Gaussian from truncated samples, retrieves p. This is shown in lemma 6.2.4 that follows.

Lemma 6.2.4 (Learnability of Truncated Binomial Distribution - PSGD)**.** *Consider a binomial distribution* $\mathrm{Bin}(n, p)$*, where* n *is fixed, and a set* $S \subset [n]$*, such that* $|S| \geq 2$*. Assume we have membership oracle access to* S *and that* $\alpha = \Pr_{X\sim \mathrm{Bin}(n,p)}[X \in S] > 0$*. Then, there exists an algorithm that, given* $m \geq \frac{n \log(1/\delta)}{poly(\alpha)\varepsilon^2}$ *i.i.d. samples from the truncated binomial distribution* $Bin_S(n, p)$*, with* $p \in (c, 1-c)$*,* c *constant, computes an estimate* $\bar{p}$ *such that:*

$$\mathrm{TV}\left(\mathrm{Bin}(n,p), \mathrm{Bin}(n,\bar{p})\right) < O\left(\varepsilon\right) ,$$

with probability at least $1 - \delta$*.*

Proof. Consider the set $S \subset [n]$ and the truncated binomial distribution $\mathrm{Bin}_S(n, p)$. The estimator $\bar{\theta}$ returned by Projected SGD defines a binomial distribution that is $O(\varepsilon)$-close to the original one. We prove this in six steps.

Denote $\mathrm{Bin}(\theta)$ a binomial distribution such that its natural parameter is θ. We write $\mathrm{Bin}(\theta; x)$ to refer to the mass assigned to point x by the distribution. We will also denote with θ^* the natural parameter of true distribution $\mathrm{Bin}(n, p)$.

Step 1: Convexity of the objective
We will run Projected SGD with the maximum likelihood objective. For SGD to converge, we need the objective to be convex. The binomial distribution, for n fixed, defines an exponential family such that:

$$\Pr_{X\sim\mathrm{Bin}(\theta)}[X=x]=\binom{n}{x}\exp\left(\theta x-n\log\left(1+e^{\theta}\right)\right),$$

where $\theta=\log\frac{p}{1-p}$ is the natural parameter. Thus, by 4.2.2, the negative log-likelihood objective is convex.

Step 2: Initial Feasible Point
Consider the M.L. estimator of the binomial distribution

$$\hat{p}_S=\frac{1}{m}\sum_{i=1}^{m}\frac{X_i}{n}$$

where $X_i\sim\mathrm{Bin}_S(n,p)$ denotes a sample. We will show that:

$$|\hat{p}_S-p|\le O\left(\sqrt{\frac{1}{n}\log\left(1/\alpha\right)}\right),$$

that is, the M.L. estimator is a good initialization for the algorithm.

Let μ_S denote the mean value of $\mathrm{Bin}_S(n,p)$. We will first show that $n\hat{p}_S$ is close to μ_S and that, in turn, μ_S is close to the true mean, np.

Hoeffding's inequality on $\hat{p}_S$ implies that, using $\Theta(\frac{\log(1/\delta)}{\varepsilon^2})$ samples, we get:

$$|\hat{p}_S-\frac{\mu_S}{n}|<\varepsilon$$

with probability $1-\delta$.

Now, applying Hoeffding's inequality for the original distribution $\mathrm{Bin}(n,p)$, we get that it has an exponential tail, i.e.

$$\Pr_{X\sim\mathrm{Bin}(n,p)}[X-np\ge\varepsilon]\le\exp\left(-2\varepsilon^2/n\right).$$

Notice that the distance between the true mean, np, and μ_S, will be maximized when all the mass of S is assigned as far away from np. So, to get an upper bound on $|np-\mu_S|$, we assign all the mass of S, α, on the tail, i.e.

$$\Pr_{X\sim\mathrm{Bin}(n,p)}\left[X\ge np+\varepsilon'\right]=\frac{\alpha}{2}$$

so as $\mu_S=np+\varepsilon'$.

Using the tail bound, we get that:

$$|np-\mu_S|=\varepsilon'\le\sqrt{\frac{n}{2}\cdot\log\left(2/\alpha\right)}.$$

Thus,

$$|n\hat{p}_S-np|<n\varepsilon+\sqrt{\frac{n}{2}\cdot\log\left(2/\alpha\right)},$$

and, with probability at least $1-\delta$:

$$|\hat{p}_S - p| = O\left(\sqrt{\frac{1}{n}\log\left(1/\alpha\right)}\right)$$

as claimed.

Step 3: Feasible Region

Note that the algorithm iterates over θ. Thus, the actual initial point of the algorithm is

$$\hat{\theta}_S = \log\frac{\hat{p}_S}{1-\hat{p}_S}\,.$$

We need to show that this is also a good initial point i.e.

$$|\hat{\theta}_S - \theta^*| \leq O\left(\sqrt{\frac{1}{n}\log\left(1/\alpha\right)}\right)$$

What is more, we will prove the existence of a convex set

$$\mathcal{B} = \{\theta : |\hat{\theta}_S - \theta| < O\left(\sqrt{\frac{1}{n}\log\left(1/\alpha\right)}\right)\}\,,$$

such that every distribution with parameter $\theta \in \mathcal{B}$ assigns $poly(\alpha)$-mass on S. Thus, PSGD iterates over the convex set $\mathcal{B}$ which guarantees the functionality of the algorithm.

Note that $p, p_S \in (c, 1-c)$ by assumption. Then, $f(x) = \log\left(\frac{x}{1-x}\right)$ is C-Lipschitz for some constant C depending on c and, consequently,

$$|\hat{\theta}_S - \theta^*| \leq C|\hat{p}_S - p| \leq O\left(\sqrt{\frac{1}{n}\log\left(1/\alpha\right)}\right)\,.$$

It follows that $\theta^* \in \mathcal{B}$ and $\hat{\theta}_S$ is indeed a good initial point.

It remains to show that a distribution with parameter $\theta \in \mathcal{B}$ assigns non-negligible mass on S.

Recall that $\text{Bin}(\theta; S)$ denotes the mass assigned to S by the binomial with natural parameter θ. Also, notice that:

$$\text{Bin}(\theta; S) = \mathbb{E}_{x\sim\text{Bin}(\theta^*)}\left[\frac{\text{Bin}\left(\theta; x\right)}{\text{Bin}\left(\theta^*; x\right)}\cdot \mathbf{1}_{x\in S}\right]\,,$$

which can be written:

$$\text{Bin}(\theta; S) = \mathbb{E}_{x\sim\text{Bin}(\theta^*)}\left[\exp\left(-\log\frac{\text{Bin}\left(\theta^*; x\right)}{\text{Bin}\left(\theta; x\right)}\right)\cdot \mathbf{1}_{x\in S}\right]\,.$$

Then, by definition:

$$\text{Bin}(\theta; x) = \binom{n}{x}\exp\left(\theta x - A(\theta)\right)\,,$$

from which we get that:

$$\log\frac{\text{Bin}(\theta^*; x)}{\text{Bin}(\theta; x)} = (\theta^* - \theta)\cdot x + C\,,$$

$C = A(\theta) - A(\theta^*)$ independent from x.

Thus, setting $g(X) = \log \frac{\mathrm{Bin}(\theta^*;X)}{\mathrm{Bin}(\theta;X)}, X \sim \mathrm{Bin}(\theta^*)$ we can derive:

$$\Pr\left[g(X) - \mathbb{E}\left[g(X)\right] \geq t\right] = \Pr\left[\sum_{i\in[n]} (\theta^* - \theta)\cdot X_i - \mathbb{E}\left[(\theta^* - \theta)X\right] \geq t\right] \leq \exp\left(-\frac{2t^2}{4n|\theta^* - \theta|^2}\right),$$

since $(\theta^* - \theta)X_i \in [-|\theta^* - \theta|, |\theta^* - \theta|]$ and by Hoeffding's inequality.

Setting $t = |\theta^* - \theta|\sqrt{2n\log(2/\alpha)}$ it follows that:

$$\Pr\left[g(X) - \mathbb{E}\left[g(X)\right] \geq \sqrt{2n\log(2/\alpha)|\theta^* - \theta|^2}\right] \leq \frac{\alpha}{2}.$$

Note that:

$$\mathbb{E}\left[g(X)\right] = KL\left(\mathrm{Bin}(\theta^*)\|\mathrm{Bin}(\theta)\right) \leq n|\theta^* - \theta|^2,$$

where the last inequality follows from 3.3.1.

Since $|\theta^* - \theta| \leq C\sqrt{\frac{1}{n}\log(1/\alpha)}$ for some C constant, it follows that, with probability at least $1 - \frac{\alpha}{2} > \frac{\alpha}{2}$:

$$-\log\frac{\mathrm{Bin}(\theta^*;x)}{\mathrm{Bin}(\theta;x)} \geq -\mathbb{E}\left[g(X)\right] - \sqrt{2n\log(2/\alpha)\cdot|\theta^* - \theta|^2} \geq -C^2\log(1/\alpha).$$

Then,

$$\mathrm{Bin}\left(\theta;S\right) \geq \exp\left(C^2\log\alpha\right)\cdot\frac{\alpha}{2} = poly(\alpha),$$

which is the expected result.

Step 4: Unbiased Estimation of the Gradient

We now have all the information to guarantee that PSGD will work. Recall that, on every iteration t, the algorithm computes a vector v_t such that the expected value of the vector is the gradient of the objective function i.e. $\mathbb{E}[v_t|\theta] = \nabla_\theta\ell(\theta;x)$.

In our case, this equals:

$$\mathbb{E}_{x\sim\mathrm{Bin}_S(\theta^*)}\left[\nabla_\theta\ell(\theta;x)\right] = -\mathbb{E}_{x\sim\mathrm{Bin}_S(\theta^*)}\left[x\right] + \mathbb{E}_{y\sim\mathrm{Bin}_S(\theta)}\left[y\right].$$

Notice that for the first expected value we can use a sample from the given distribution. For the second estimate, there is the following procedure. By assumption, we have membership oracle access to S. Since the support of the distribution is n, we can recover set S with just n queries. Thus, to get a sample from $\mathrm{Bin}_S(\theta)$, we draw a sample from $\mathrm{Bin}(\theta)$ and check whether it falls into S. We draw until we access a sample in S. Since the mass of S is $poly(\alpha)$ for every distribution defined by $\theta \in \mathcal{B}$, we will get a sample in $poly(1/\alpha)$ steps.

Step 5: Strong Convexity of the objective

To guarantee efficiency of PSGD, convexity of the objective is not enough. We will show that the negative log-likelihood objective for the truncated binomial distribution is strongly convex in $\mathcal{B}$. It suffices to show that the Hessian of the objective is strictly positive.

Notice that the Hessian in the single parameter case becomes:

$$H_\ell(\theta) = \mathrm{Var}_{X\sim\mathrm{Bin}_S(\theta)}\left[X\right] = \mathbb{E}_{X\sim\mathrm{Bin}_S(\theta)}\left[(X - \mu_S)^2\right],$$

where $\mu_S = \mathbb{E}_{X\sim\mathrm{Bin}_S(\theta)}[X]$ is the mean value of the conditional binomial distribution. By the anticoncentration of the binomial distribution, see lemma 3.2.4, it holds that:

$$\Pr_{X\sim\mathrm{Bin}(\theta)}\left[\mu_S \leq X \leq \mu_S + t\right] \leq C\frac{t}{\sigma},$$

where $\sigma = \sqrt{np(1-p)}$ and C a constant. Choosing $t = poly(\alpha)\sigma/2C$ we get that the mass of the set $\bar{S} = \{x : |x - \mu_S| < t\}$ is at most $poly(\alpha)/2$. Since the truncation set S has mass $poly(\alpha)$ it follows that $|x - \mu_S| > t$ with probability at least $1/2$ for every $x \in S$. Thus,

$$H_\ell(\theta) = \mathrm{Var}_{X \sim \mathrm{Bin}_S(\theta)}\left[X\right] \geq \frac{poly(\alpha)^2 \cdot \sigma^2}{4C} \cdot \frac{1}{2} > 0\,,$$

and the negative log-likelihood objective is $(poly(\alpha)n)$-strongly convex.

Step 6: PSGD

We want to apply theorem 4.1.1 to bound the expected error of our algorithm's estimation. Thus, we need to show that the variance of the gradient estimator is bounded. Recall that:

$$v_t = \nabla_\theta \ell\left(\theta\right) = -x + y\,,$$

where x is the sample drawn from the true distribution and y the sample drawn from the currently estimated distribution. Thus,

$$\mathbb{E}\left[|v_t|^2\right] = \mathbb{E}\left[|y - x|^2\right]\,,$$

and since x, y come from different distributions we get:

$$\mathbb{E}\left[|v_t|^2\right] \leq 2\mathbb{E}_{Y \sim \mathrm{Bin}_S(\theta)}\left[Y^2\right] + 2\mathbb{E}_{X \sim \mathrm{Bin}_S(\theta^*)}\left[X^2\right]\,.$$

Notice that:

$$\mathbb{E}_{Y \sim \mathrm{Bin}_S(\theta)}\left[Y^2\right] = \sum_{y \in S} y^2 p_S(y) \leq n^2\,.$$

From theorem 4.1.1 we get the bound:

$$\mathbb{E}\left[\ell(\theta)\right] - \ell(\theta^*) \leq \frac{n^2}{2 \cdot poly(\alpha)n \cdot M} \cdot \left(1 + \log M\right)\,,$$

where M is the number of iterations.

Next, we apply the procedure used in lemma 4.1.4 and acquire an estimator $\bar{\theta}$ such that:

$$\ell(\bar{\theta}) - \ell(\theta^*) \leq \frac{n}{2poly(\alpha)M} \cdot (1 + \log M)$$

with probability at least $1 - \delta$. Since ℓ is $(poly(\alpha)n)$-strongly convex, we get that:

$$|\bar{\theta} - \theta|^2 \leq \frac{n}{poly(\alpha) \cdot poly(\alpha)n \cdot M} \cdot \left(1 + \log M\right)\,.$$

Thus, for

$$M = O\left(\frac{n \log\left(1/\delta\right)}{poly(\alpha)\varepsilon^2}\right)$$

there exists an algorithm that gives an estimator $\bar{\theta}$ such that:

$$|\bar{\theta} - \theta^*| \leq \varepsilon/\sqrt{n}$$

with probability at least $1 - \delta$.

By lemma 3.3.1 it also holds that:

$$\mathrm{TV}\left(\mathrm{Bin}(n, \bar{p}), \mathrm{Bin}(n, p)\right) \leq \varepsilon$$

where $\bar{p}$ is defined by the natural parameter $\bar{\theta}$. □

It seems that this new approach offers no improvement. The sample complexity increases linearly with n in this case too. However, we should still explore the quality of our analysis through experiments. In this way we can verify whether this amount of samples is indeed necessary to retrieve p. We do so right after.

Experiments

We run the same experiments as in the previous cases. We demonstrate our results in fig. 6.5. In contrast to lemma 6.2.4 we used only $m = 10/(\alpha\varepsilon^2)$ samples. Recall that in Approach One where we used the PSGD with the Gaussian objective we used $50/(\alpha\varepsilon^2)$.

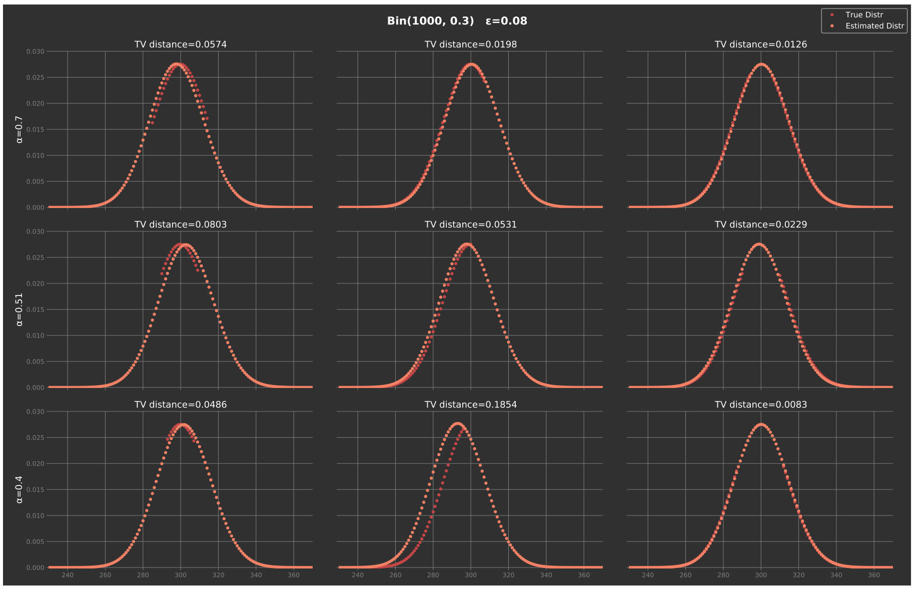

Figure 6.5: Learn parameter p of Binomial using $10/(\alpha\varepsilon^2)$ samples. The right-hand side inscription α denotes the mass of the truncation set.

Despite ignoring the formal result in lemma 6.2.4, the estimations are the best so far. Of course, we only retrieve p here. Thus we cannot compare this behaviour to Approach One. However this is an encouraging result since it estimates p with high precision.

We run the same experiments for a different Binomial distribution as well. This is the same used in Approach One and the results are given in fig. 6.6. We can see that $m = 10/(\alpha\varepsilon^2)$ suffice for retrieving p in every. Thus the analysis in lemma 6.2.4 should be improved to give the experimental sample complexity.

Notice that the last two Approaches only care for learning the parameter p of the Binomial. Approach Two is sub-optimal so it should not concern us further. Approach Three could be used along with a method for retrieving n. A candidate algorithm could be: iterate over n and, on every iteration, calculate p assuming that is the correct n. The problem with this algorithm is that we need a stopping criterion. That is, a way to understand if a couple (n, p) is guarantees small TV distance. Specifying such a criterion is another open problem that arises by this study.

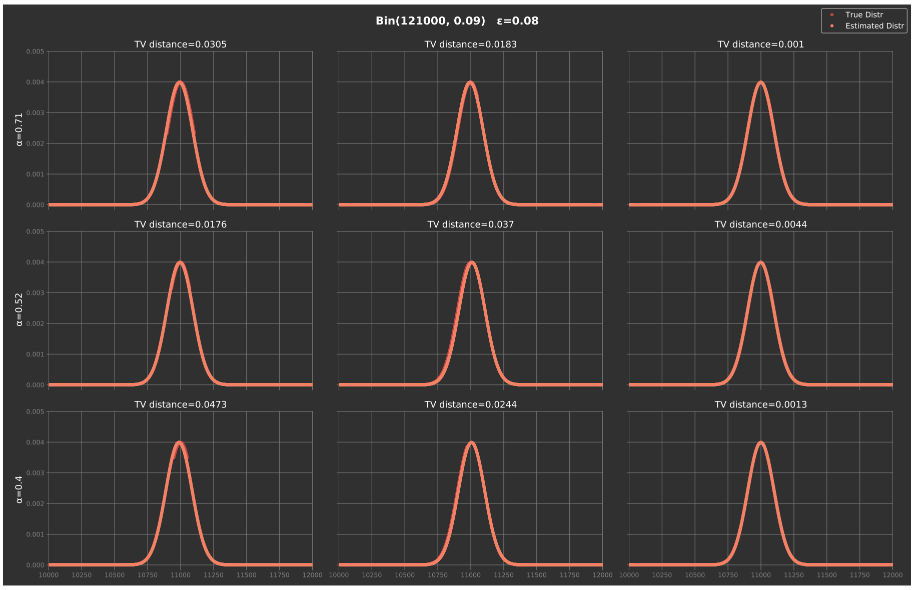

Figure 6.6: Learn parameter p of Binomial using $10/(\alpha\varepsilon^2)$ samples. The right-hand side inscription α denotes the mass of the truncation set.

Chapter 7

Learning Mallows Distribution from Truncated Samples

This chapter is concerned with another probability distribution: the Mallows distribution or Mallows Model. This model plays a fundamental role in ranking aggregation and social choice theory, which are related to finding an optimal ranking of m objects. The Mallows distribution is defined over the set of permutations $\mathbb{S}_m$ (rankings) and has two parameters. Once again, our aim is to recover these parameters accessing samples from a subset of $\mathbb{S}_m$. In fact, this work focuses on retrieving only one of these parameters: the central ranking $\pi_0 \in \mathbb{S}_m$. Note the connection to ranking aggregation theory, since we search for the 'winning' ranking π_0 from a set of rankings sampled according to the Truncated Mallows distribution.

First, a formal definition of the Mallows distribution is given. Then we present how we learn π_0 in the non-truncated case to get an insight on the difficulties that might arise. We proceed to find that condition on the truncation set S which characterizes learnability. As in the preceding chapter, we will first point out a condition for identifiability. What is of interest here is that this condition is not sufficient to learn the central ranking. We will define a sufficient condition, though, and examine whether it holds for some natural truncation sets.

7.1 The Mallows Model

The Mallows distribution, often referred as the Mallows Model, was originally introduced by Mallows in [Mal57]. It is a statistical model created to solve the following scenario: Let A be a set of distinct elements which we want to rank. These elements are also called the *alternatives* of the model. Every person can give an opinion on their ranking that can be based on their preference, sense of justice etc. Aggregating the individual opinions into a correct one -for some definition of correctness- is the subject of ranking aggregation theory. One can think of this procedure as the familiar *voting*.

There are plenty aggregations rules proposed in the literature ([BF02]). One can think of the well-known plurality rule as a common example. Assessing them, however, depends on which ranking of the alternatives we consider correct. There are some natural properties of aggregation rules consistent with the conventional notion of fairness. However, it has been shown that demanding some very few of them to hold simultaneously results in impossibility results (see Arrow's theorem [FH19]). What is more, specifying the ranking that has the minimum disagreement with the set of individual opinions is also shown to be NP-hard ([HSV05]). One solution is to resort in statistical models of rank data and consider the MLE or other model parameters as the correct ranking. It is in this context that the Mallows Model was created. See [Xia19] for a presentation of such models.

Mallows Model

The Mallows Model assumes that there exists an objective, true order of these alternatives, specified by nature. Then, every individual's ranking is a noisy version of the underlying, true ranking which is denoted $\pi_0 \in \mathbb{S}_m$. According to Mallows, this noise is produced in the following way: for every pair of alternatives $a, b \in A$ such that a is preferred over b in π_0, write $a \succ_{\pi_0} b$, an individual will flip their order in her ranking with probability $1 - p < 1/2$. Thus the probability of a ranking $\pi \in \mathbb{S}_m$ equals

$$p^{\binom{m}{2}-d} \cdot (1-p)^d , \tag{1}$$

where d denotes the number of pairs $a, b \in A$ on whose order π_0 and π disagree. Formally, this quantity represents a distance between two rankings, the Kendall-Tau distance.

Definition 7.1.1 (Kendall-Tau Distance). *For any pair of rankings $\pi, \sigma \in \mathcal{L}(A)$ let $\mathrm{KT}(\pi, \sigma)$ denote the Kendall-Tau distance between π and σ, such that:*

$$\mathrm{KT}(\pi, \sigma) = \left| \{\{a, b\} \subseteq A : a \succ_\pi b \quad \textit{and} \quad b \succ_\sigma a\} \right| .$$

For $\phi = (1-p)/p$ and appropriate normalization eq. (1) becomes a probability distribution with p.m.f.

$$p(\pi) = \frac{\phi^{\mathrm{KT}(\pi,\pi_0)}}{Z(\phi)} .$$

It can be shown (see [LB11]) that:

$$Z(\phi) = \sum_{\pi \in \mathcal{L}(A)} p(\pi) = \prod_{i=1}^{|A|-1} \sum_{j=0}^{i} \phi^j .$$

The formal definition of the Mallows distribution is as follows:

Definition 7.1.2 (Mallows Distribution). *Let $\phi \in (0,1)$ and $\pi_0 \in \mathbb{S}_m$. Denote $\mathcal{M}(\pi_0, \phi)$ the Mallows distribution with true ranking π_0 and dispersion ϕ, where, for every $\pi \in \mathbb{S}_m$, the probability mass on π equals:*

$$p(\pi) = \frac{1}{Z(\phi)} \phi^{\mathrm{KT}(\pi,\pi_0)} ,$$

where $Z(\phi)$ is the normalization constant.

We will write $\mathcal{M}(\pi; \pi_0, \phi)$ (or $\mathcal{M}(\pi; \pi_0)$ when ϕ is known) to denote the probability mass of a ranking $\pi \sim \mathcal{M}(\pi_0, \phi)$. Needless to say, $\mathcal{M}(S; \pi_0, \phi)$ is the probability mass on a set $S \subseteq \mathbb{S}_m$.

Truncated Mallows Model

The central problem of this chapter is to reconstruct the true ranking π_0 of the Mallows Model, given access only to samples that belong to a truncation set $S \subseteq \mathbb{S}_m$. Consider that samples come from the truncated Mallows distribution, write $\mathcal{M}_S(\pi_0, \phi)$. This is defined as in definition 4.2.2. We give the formal definition in the Mallows case right after for completeness.

Definition 7.1.3. *Consider the Mallows distribution $\mathcal{M}(\pi_0, \phi)$ with central ranking $\pi_0 \in \mathbb{S}_m$ and dispersion $\phi \in (0,1)$. The truncated Mallows distribution $\mathcal{M}_S(\pi_0, \phi)$ on a set $S \subseteq \mathbb{S}_m$ is defined as:*

$$\mathcal{M}_S\left(\pi\right) = \frac{1}{\mathcal{M}\left(S; \pi_0, \phi\right)} \frac{\phi^{\mathrm{KT}(\pi,\pi_0)}}{Z(\phi)}, \pi \in S$$

Next, we define a notion of a 'good' truncated Mallows distribution, in the sense that the probability of $a \succ_\pi b$, $\pi \sim \mathcal{M}_S(\pi_0, \phi)$ is over a half for every $a, b \in A$ such that $a \succ_{\pi_0} b$. That is, the truncated distribution has a bias to rank pairs of alternatives according to the true ranking. Note that this is always true in the non-truncated case.

Definition 7.1.4 (Consistency)**.** *Consider the truncated Mallows distribution $\mathcal{M}_S(\pi_0, \phi)$, where $\pi_0 \in \mathbb{S}_m$ is the true ranking, $\phi \in (0,1)$ is the dispersion and $S \subseteq \mathbb{S}_m$ the truncation set. We will call $\mathcal{M}_S(\pi_0, \phi)$ consistent if for every pair of alternatives $a, b \in A$ such that $a \succ_{\pi_0} b$ it holds:*

$$p^S_{a \succ b} > p^S_{b \succ a},$$

where $p^S_{a \succ b} = \sum_{a \succ_\pi b} \mathcal{M}_S(\pi)$ is the probability of the truncated distribution to rank two alternatives according to the true ranking.

Note that we will often denote $\delta_{ab} = p^S_{a \succ b} - p^S_{b \succ a}$. We will prove that *consistency* is a sufficient condition to retrieve the true ranking given samples from the truncated distribution. The following section gives an intuition on why this happens.

7.2 Learn the central ranking of Mallows Model

In [CPS13] an optimal, with respect to sample complexity, algorithm is given to acquire the central ranking π_0 of the Mallows Model. This is algorithm 9. We prove that with high probability and given a logarithmic number of samples, this algorithm retrieves the central ranking in theorem 7.2.1.

One can think of this procedure the following way. Create a directed graph with nodes the alternatives in A. An edge goes from $a \in A$ to $b \in A$ if a is ranked above b in the majority of samples. Since the samples are drawn from the Mallows distribution, theorem 7.2.1 guarantees that the resulting graph will be a DAG representing π_0 with high probability.

Theorem 7.2.1 ([CPS13])**.** *For any given $\delta > 0$, there exists a polynomial time algorithm that determines the true ranking with probability at least $1 - \delta$ given $O(\log{(m/\delta)})$ samples from the Mallows Model.*

Algorithm 9 Estimate central ranking of Mallows Model

```
procedure MallowsCentralRanking($\Pi$)          ▷ $\Pi$ the samples' set
    $\pi_0 \leftarrow Id$                        ▷ Id is the id ranking
    for $\{a, b\} \in A$ do                      ▷ For all pairs of alternatives
        $n_{ab} \leftarrow |\{\pi \in \Pi | a \succ_\pi b\}|$
        if $n_{ab} > |\Pi|/2$ then
            $\pi_0 \leftarrow \pi_0(a \succ b)$    ▷ Move $a$ above $b$ in ranking
        else
            $\pi_0 \leftarrow \pi_0(b \succ a)$
    return $\pi_0$
```

Proof. We will prove that algorithm 9 determines the true ranking π_0 with high probability given $O(\log{(m/\delta)})$ samples.

Let $\Pi \in (\mathbb{S}_m)^n$ be the set of n samples drawn from the Mallows distribution. Denote $n_{ab} = \sum_{\pi_i \in \Pi} X_i^{ab}$ the number of samples that rank a above b, where X_i^{ab} equals 1 when $a \succ_{\pi_i} b$ and 0 otherwise. We also write $p_{a \succ b} = \Pr_{\pi \sim \mathcal{M}(\pi_0, \phi)}[a \succ_\pi b]$ for the probability that a pair of alternatives is ranked in a specific order by Mallows. It follows that $X_i^{ab} \sim \text{Be}(p_{a \succ b})$.

It holds that algorithm 9 returns π_0 iff $\forall a, b \in A$ such that $a \succ_{\pi_0} b$ it holds that $n_{ab} - n_{ba} \geq 1$. Let $\delta_{ab} = p_{a \succ b} - p_{b \succ a}$ for every $a, b \in A$ such that $a \succ_{\pi_0} b$. Since $\mathbb{E}[n_{ab} - n_{ba}] = n\delta_{ab}$, it follows that:

$$\Pr\left[n_{ab} - n_{ba} \leq 0\right] = \Pr\left[\frac{n_{ab} - n_{ba}}{n} - \mathbb{E}[\frac{n_{ab} - n_{ba}}{n}] \leq -\delta_{ab}\right] \leq \Pr\left[|\frac{n_{ab} - n_{ba}}{n} - \mathbb{E}[\frac{n_{ab} - n_{ba}}{n}]| \geq \delta_{ab}\right],$$

and by Hoeffding's inequality:

$$\Pr\left[n_{ab} - n_{ba} \leq 0\right] \leq 2\exp{(-2n\delta_{ab}^2)}\,.$$

If $\delta_{min} = \min_{a,b \in A : a \succ_{\pi_0} b} \delta_{ab}$, we get that $\Pr\left[n_{ab} - n_{ba} \leq 0\right] \leq 2\exp{(-2n\delta_{min}^2)}$ for every pair of alternatives $a, b \in A$. Thus, by union bound,

$$\Pr\left[\forall a, b \in A, a \succ_{\pi_0} b : n_{ab} - n_{ba} > 0\right] \geq 1 - \binom{m}{2} 2\exp{(-2n\delta_{min}^2)}\,.$$

Setting $\delta = \binom{m}{2} 2\exp{(-2n\delta_{min}^2)}$ it follows that given $O(\log{(m/\delta)}/\delta_{min}^2)$ samples suffice to get π_0 with probability $1 - \delta$. Thus, we still need to show that δ_{min} is a constant. Notice that, for $a, b \in A$ consecutive in π_0 it holds that:

$$p_{a \succ b} = \sum_{\pi \in \mathbb{S}_m, a \succ_\pi b} \frac{1}{Z} \phi^{d_{KT}(\pi, \pi_0)} = \sum_{\pi \in \mathbb{S}_m, b \succ_\pi a} \frac{1}{Z} \phi^{d_{KT}(\pi, \pi_0) - 1} = \phi^{-1} p_{b \succ a}\,.$$

For $a, b \in A$ not consecutive, quantifying the increase in KT-distance between rankings where only a, b is flipped is not that simple. However, we notice that the distance from π_0 must increase by at least 1 and, thus, it holds that:

$$p_{a \succ b} \geq \phi^{-1} p_{b \succ a}\,.$$

Also, note that $p_{a \succ b} + p_{b \succ a} = 1$ so $p_{b \succ a} = \frac{1-\delta_{ab}}{2}$. It follows from calculations that:

$$\delta_{ab} \geq (\phi^{-1} - 1) \cdot \frac{1 - \delta_{ab}}{2} \geq \frac{1-\phi}{1+\phi},$$

and so it is for δ_{min}. Thus, δ_{min} is greater than a constant that depends on ϕ and so the sample complexity in $O(\log(m/\delta))$ as claimed. □

The above proof attributes the success of algorithm 9 to two facts. First, the concentration of n_{ab} around its mean value $np_{a \succ b}$ and, second, the fact that $p_{a \succ b} > 1/2 > p_{b \succ a}$ for every $a \succ_{\pi_0} b$. Notice that the later is a property of the Mallows Model on $\mathbb{S}_m$ that can be violated when truncation is applied.

Moreover, we note, though not prove, that this algorithm is optimal.

Theorem 7.2.2 ([CPS13]). *For any $\delta \in (0, 1/2]$, any algorithm requires $\Omega(\log(m/\delta))$ samples from the Mallows Model to determine the true ranking with probability at least $1 - \delta$.*

This optimality implies an *if-and-only-if* connection between pairwise comparisons and the Mallows Model. We note this observation. We will shortly present a sufficient condition for learnability from truncated samples. We will claim it is also necessary partly because of this exact behaviour.

7.3 Necessary Condition

The first step to explore the learnability of a truncated distribution is to make sure that distinguishing it from others is even possible. Recall that this is the notion of identifiability. In the following lemma, we give a condition involving the distances between rankings that fully characterizes the identifiability of a truncated Mallows distribution.

Lemma 7.3.1 (Identifiability). *Let $D_\pi(\pi_i, \pi_j) = d_{KT}(\pi_i, \pi) - d_{KT}(\pi_j, \pi)$. The truncated Mallows distribution $\mathcal{M}_S(\pi_0, \phi)$ is non-identifiable iff there exists a ranking π_1 such that, for every $\pi_i, \pi_j \in S$, it holds that:*

$$D_0(\pi_i, \pi_j) = D_1(\pi_i, \pi_j) = 0\,, \quad \textit{if} \quad D_0(\pi_i, \pi_j) = 0 \cup D_1(\pi_i, \pi_j) = 0\,,$$

or

$$\frac{D_0(\pi_i, \pi_j)}{D_1(\pi_i, \pi_j)} = c\,, \quad \textit{if} \quad D_0(\pi_i, \pi_j) \neq 0 \cap D_1(\pi_i, \pi_j) \neq 0\,,$$

where c is a constant.

Proof. Consider the Mallows distributions $M(\pi_0, \phi_0)$ and $M(\pi_1, \phi_1)$. Let $S \subseteq \mathbb{S}_m$ be the truncation set. Denote $p_{S,\pi_i}(\pi)$ the probability mass on π by distribution $\mathcal{M}_S(\pi_i, \phi_i)$, i.e.

$$p_{S,\pi_i}(\pi) = \frac{1}{\mathcal{M}(\pi_i; S)} \frac{1}{Z} \phi_i^{d_{KT}(\pi, \pi_i)}.$$

Assume $\mathcal{M}_S(\pi_0, \phi_0) \equiv \mathcal{M}_S(\pi_1, \phi_1)$, i.e. $\mathcal{M}_S(\pi_0, \phi_0)$ is non-identifiable. Then, for every $\pi \in S$ it must hold that:

$$p_{S,\pi_0}(\pi) = p_{S,\pi_1}(\pi)\ .$$

This is equivalent, by definition, to:

$$\frac{Z_{\phi_1}\phi_0^{d_{KT}(\pi,\pi_0)}}{Z_{\phi_0}\phi_1^{d_{KT}(\pi,\pi_1)}} = \frac{\mathcal{M}(\pi_0;S)}{\mathcal{M}(\pi_1;S)} .$$

Thus, for every $\pi_i, \pi_j \in S$ it holds that:

$$\phi_0^{d_{KT}(\pi_i,\pi_0)-d_{KT}(\pi_j,\pi_0)} = \phi_1^{d_{KT}(\pi,\pi_1)-d_{KT}(\pi_j,\pi_1)} ,$$

which, taking the logarithm on both sides, implies the result. □

In lemma 7.3.1 we characterize the identifiability of a general Mallows distribution. This thesis examines the case that ϕ is known. Thus the condition is given by the following

Corollary 7.3.0.1. *The truncated Mallows distribution $\mathcal{M}_S(\pi_0,\phi)$ for fixed ϕ is non-identifiable iff there exists a ranking π_1 such that, for every $\pi_i \in S$, it holds that:*

$$\mathrm{KT}(\pi_i,\pi_0) - \mathrm{KT}(\pi_i,\pi_1) = c ,$$

where c is a constant.

Proof. The result is given immediately if we set $\phi_0 = \phi_1$ in the proof of lemma 7.3.1. □

None of the above conditions is very informative about the structure of a good or bad truncation set. In fact, it is not straight-forward how to even check the conditions efficiently, given a central ranking and a truncation set. We claim that this is NP-hard. Consider $\mathcal{M}_S(\pi_0)$ is not identifiable because there exists a π_1 such that

$$\mathrm{KT}(\pi_i,\pi_0) = \mathrm{KT}(\pi_i,\pi_1) \ ,\forall \pi_i \in S .$$

Thus $c = 0$ in corollary 7.3.0.1. In the case that π_0 is the median of S, determining whether such a π_1 exists is reduced to determining whether the median of S is unique. Thus the problem resembles known hardness results such as [HL19].

For small m, though, these conditions are easily checked by code and give us some useful observations. First, notice that there is not such a π_1 for every $\mathcal{M}_S(\pi_0,\phi)$. E.g. take $\pi_0 = (1,2,3)$ and $S = \{(1,3,2),(2,1,3),(2,3,1)\}$. So there are truncation sets that allow identifiability.

An important observation for the Mallows distribution is that identifiability does not imply learnability. We illustrate this in example 7.3.1.

Example 7.3.1. *Consider the case that $m = 3$, the central ranking $\pi_0 = (1,2,3)$ and the truncation set is $S = \{(2,1,3),(3,1,2),(3,2,1)\}$. It can be checked that $\mathcal{M}_S(\pi_0,\phi)$ is identifiable for every $\phi \in (0,1)$.*

Notice that the probability of seeing $\pi_0^{-1} = (3,2,1) \in S$ while sampling from $\mathcal{M}_S(\pi_0,\phi)$ is exponentially small. That means that we need exponentially many samples to learn even the truncation set. Indeed, drawing n samples, the probability that we get π_0^{-1} is:

$$\Pr\left[\bigcup_{i=1}^{n} X_i\right] \leq \sum_{i=1}^{n} \frac{C\phi^m}{Z_\phi} = \frac{nC\phi^m}{Z_\phi} .$$

For this event to happen, even with probability $1/2$*, it holds that:* $\frac{1}{2} = \Pr\left[\bigcup_{i=1}^{n} X_i\right] \le \frac{nC\phi^m}{Z_\phi}$ *so*

$$n \ge C Z_\phi \phi^{-m} .$$

Thus, demanding polynomial number of samples means that, with high probability, the truncation set we have access to we will be $S' = \{(2,1,3),(3,1,2)\}$*. However,* $\mathcal{M}_{S'}(\pi_0,\phi)$ *is not identifiable. Taking* $\pi_1 = (2,3,1)$ *it holds that* $\mathcal{M}_{S'}(\pi_0,\phi) \equiv \mathcal{M}_{S'}(\pi_1,\phi)$*.*

7.4 Sufficient Condition

In the previous section, we established a gap between identifying the Mallows distribution and learning its parameters. We have already defined the notion of *consistency* and elaborated on the recoverability of π_0 under this assumption. Thus, the two conditions cannot be equivalent i.e. *non-consistent* distributions can be identifiable. It is still open whether they can be learnable, though.

We will now prove the learnability of π_0 when the truncated distribution is *consistent.* The proof follows almost precisely that of theorem 7.2.1 and is given below.

Lemma 7.4.1. *Consider* $\mathcal{M}_S(\pi_0,\phi)$ *the truncated Mallows distribution, with central ranking* $\pi_0 \in \mathbb{S}_m$ *and dispersion* $\phi \in (0,1)$*. Assume* $\mathcal{M}_S(\pi_0,\phi)$ *is consistent. Let* $\delta_{min} = min_{a \succ_{\pi_0} b}(p^S_{a\succ b} - p^S_{b\succ a})$*, where* $p^S_{a\succ b}$ *as defined in definition 7.1.4. Then, there exists an algorithm that recovers the true ranking* π_0 *with probability at least* $1-\delta$ *given* $O(\log(m/\delta)/\delta^2_{min})$ *samples.*

Proof. Consider the truncated Mallows distribution $M_S(\pi_0,\phi)$ with central ranking π_0 and dispersion ϕ. We will prove that algorithm 9 determines the true ranking π_0 with high probability given $O(\log(m/\delta))$ samples.

Let $\Pi \in S^n$ be the set of n samples drawn from the truncated Mallows distribution. Denote $n_{ab} = \sum_{\pi_i \in \Pi} X^{ab}_i$ the number of samples that rank a above b, where X^{ab}_i equals 1 when $a \succ_{\pi_i} b$ and 0 otherwise. We also write $p^S_{a\succ b} = \Pr_{\pi\sim\mathcal{M}_S(\pi_0,\phi)}[a \succ_\pi b]$ for the probability that a pair of alternatives is ranked in a specific order by Mallows. It follows that $X^{ab}_i \sim \mathrm{Be}(p^S_{a\succ b})$.

It holds that algorithm 9 returns π_0 iff $\forall a,b \in A$ such that $a \succ_{\pi_0} b$ it holds that $n_{ab} - n_{ba} \ge 1$. Let $\delta_{ab} = p^S_{a\succ b} - p^S_{b\succ a}$ for every $a,b \in A$ such that $a \succ_{\pi_0} b$. Since $\mathbb{E}[n_{ab} - n_{ba}] = n\delta_{ab}$, it follows that:

$$\Pr\left[n_{ab} - n_{ba} \le 0\right] = \Pr\left[\frac{n_{ab}-n_{ba}}{n} - \mathbb{E}[\frac{n_{ab}-n_{ba}}{n}] \le -\delta_{ab}\right] \le \Pr\left[|\frac{n_{ab}-n_{ba}}{n} - \mathbb{E}[\frac{n_{ab}-n_{ba}}{n}]| \ge \delta_{ab}\right] ,$$

and by Hoeffding's inequality:

$$\Pr\left[n_{ab} - n_{ba} \le 0\right] \le 2\exp\left(-2n\delta^2_{ab}\right) .$$

If $\delta_{min} = \min_{a,b\in A : a\succ_{\pi_0} b} \delta_{ab}$, we get that $\Pr\left[n_{ab} - n_{ba} \le 0\right] \le 2\exp\left(-2n\delta^2_{min}\right)$ for every pair of alternatives $a,b \in A$. Thus, by union bound,

$$\Pr\left[\forall a,b \in A, a \succ_{\pi_0} b : n_{ab} - n_{ba} > 0\right] \ge 1 - \binom{m}{2} 2\exp\left(-2n\delta^2_{min}\right) .$$

Setting $\delta = \binom{m}{2} 2\exp\left(-2n\delta^2_{min}\right)$ it follows that given $O(\log(m/\delta)/\delta^2_{min})$ samples suffice to get π_0 with probability $1-\delta$. □

In the case of complete Mallows, we insisted on the fact that this algorithm is actually optimal, that is no algorithm can retrieve the true ranking using less samples. Note that a learning algorithm with access to truncated samples should work in the case $S \equiv \mathbb{S}_m$, i.e. when there is not truncation. Constructing an algorithm that retrieves π_0 from a *non-consistent* Mallows would give a different algorithm for learning π_0 in the general case.

7.5 Examples of Truncation Sets

In the previous sections, we have given some formal characterizations for the truncation set S. Here we will be concerned with a more practical aspect of the problem. We want to know how easy it is to actually construct an S that makes learnability hard. We present two lemmas that will hopefully help us form an intuition.

First, we study the case that a truncation set is chosen uniformly at random. Unless ϕ is very close to 1, i.e. the Mallows distribution is close to the uniform distribution, a random S cannot affect our ability to learn the true ranking.

Lemma 7.5.1. *Let $S \sim_u \mathcal{P}(\mathbb{S}_m)$ be drawn uniformly at random from the powerset of permutations. Then, the truncated Mallows distribution $\mathcal{M}_S(\pi_0, \phi)$, where $\phi \in (0, 1 - \sqrt{16 \log(\frac{m}{\delta})/m!})$, is consistent with probability at least $1 - \delta$.*

Proof. Let S be the truncation set. We choose S uniformly at random, meaning that we choose every set of permutations with equal probability $1/|\mathcal{P}(\mathbb{S}_m)|$.

We will prove that every Mallows distribution $\mathcal{M}(\pi_0, \phi)$, where ϕ is quite smaller than 1, truncated on S, is *consistent* with high probability. That is, $\forall a, b \in A$ such that $a \succ_{\pi_0} b$ it must hold that $p^S_{a \succ b} > p^S_{b \succ a}$. Notice that, since S is chosen uniformly at random, and every permutation $\pi \in \mathbb{S}_m$ belongs to exactly half of the sets in the powerset, the probability for a permutation to belong to S equals $1/2$.

Denote $\hat{p}_{a \succ b} = \sum_{\pi \in S : a \succ b} p(\pi)$, where $p(\pi)$ is the probability mass Mallows assigns on π. Notice that:

$$p^S_{a \succ b} > p^S_{b \succ a} \iff \hat{p}_{a \succ b} > \hat{p}_{b \succ a} \,.$$

Let $X_i \sim \mathrm{Be}(1/2)$ indicate whether π_i belongs to S. Then, $\hat{p}_{a \succ b} = \sum_{a \succ_{\pi_i} b} p(\pi_i) X_i$ and we get that:

$$\mathbb{E}\left[\hat{p}_{a \succ b}\right] = \mathbb{E}\left[\sum_{a \succ_{\pi_i} b} p(\pi_i) X_i\right] = \frac{p_{a \succ b}}{2} \,.$$

Thus, by Hoeffding's inequality:

$$\Pr\left[|\hat{p}_{a \succ b} - \frac{p_{a \succ b}}{2}| \geq \varepsilon\right] \leq 2 \exp\left(-m! \varepsilon^2\right) .$$

By similar argument, we can bound the probability of $|\hat{p}_{b \succ a} - \frac{p_{b \succ a}}{2}| \geq \varepsilon$. Combining the two, we can derive:

$$\hat{p}_{a \succ b} - \hat{p}_{b \succ a} \geq \frac{p_{a \succ b} - p_{b \succ a}}{2} - 2\varepsilon \,.$$

Recall from the proof of theorem 7.2.1 that $\delta_{ab} = p_{a \succ b} - p_{b \succ a} \geq \frac{1-\phi}{1+\phi}$. Thus, for $\varepsilon = \frac{1}{4} \frac{1-\phi}{1+\phi}$, it follows that $\hat{p}_{a \succ b} > \hat{p}_{b \succ a}$ for every $a, b \in A$ such that $a \succ_{\pi_0} b$. So we get that:

$$\Pr_{S \sim_u \mathcal{P}(\mathbb{S}_m)}\left[S \ \textit{consistent}\right] \geq 1 - \binom{m}{2} 2 \exp\left(-m! \frac{1}{16} (\frac{1-\phi}{1+\phi})^2\right),$$

and setting $\delta = \binom{m}{2} 2 \exp\left(-m! \frac{1}{16} (\frac{1-\phi}{1+\phi})^2\right)$ the result follows.

□

To make a more quantitative approach, we investigate how robust our algorithm is, when we are allowed to remove only one permutation from the truncation set.

Lemma 7.5.2. *Consider the Mallows distribution $\mathcal{M}(\pi_0, \phi)$. Let $|S| = |\mathbb{S}_m| - 1$. Then, the truncated Mallows distribution $\mathcal{M}_S(\pi_0, \phi)$ is consistent if $\frac{1-\phi}{1+\phi} Z_\phi > 1$, where Z_ϕ is the normalization constant of the Mallows distribution.*

Proof. Consider the truncated Mallows distribution $\mathcal{M}_S(\pi_0, \phi)$, where $|S| = |\mathbb{S}_m| - 1$. That means we are allowed to remove only one permutation from the domain of Mallows distribution.

Recall that *consistency* of the distribution strongly depends on the difference $\delta_{ab} = p_{a \succ b} - p_{b \succ a}$ for $a, b \in A$ such that $a \succ_{\pi_0} b$. Thus, to make $\mathcal{M}_S(\pi_0, \phi)$ *inconsistent* the best choice is to remove the permutation π with the larger mass, such that $a \succ_\pi b$. This is, by definition, π_0 having probability mass $1/Z_\phi$. As a result, we care about the sign of:

$$p_{a \succ b} - \frac{1}{Z_\phi} - p_{b \succ a} \,.$$

Moreover, recall from the proof of theorem 7.2.1 that $\delta_{ab} \geq \frac{1-\phi}{1+\phi}$, where equality holds for a, b consecutive in π_0. Thus, subtracting $1/Z_\phi$, δ_{ab} for a, b consecutive is more liable to change sign. We focus on a, b consecutive. Then

$$\sum_{a \succ_\pi b} p(\pi) = \phi^{-1} \sum_{b \succ_\pi a} p(\pi) \,,$$

and

$$Z_\phi = \sum_{a \succ_\pi b} p(\pi) + \sum_{b \succ_\pi a} p(\pi) = (1 + \phi^{-1}) \sum_{b \succ_\pi a} p(\pi) \,,$$

imply that, for $p_{a \succ b} - \frac{1}{Z_\phi} > p_{b \succ a}$, it must be

$$\frac{1-\phi}{1+\phi} Z_\phi > 1 \,.$$

□

We point out that the above condition depends not only on ϕ, but on m as well. Calculating the ϕ values through code, we make the following remarks. When $m = 3$, removing one permutation instantly makes the distribution *inconsistent* for any ϕ not very close to 0. For $m > 3$, however, ϕ must be very close to 1 for the truncation to affect learnability.

In truth, these two examples only add on a simple observation. Since *consistency* depends on δ_{ab} and this is lower bounded from $\frac{1-\phi}{1+\phi}$, while ϕ increases δ_{ab} becomes smaller. Thus, the truncation of only few permutations can flip its sign.

Chapter 8

Future Work

Each of our future directions evolves either from the PBD or the Mallows problem studied in this thesis. Thus we divide them into two categories.

- The most immediate open issue left by this work is a formal statement for the learnability of a PBD close to a heavy Binomial distribution. This is actually an important result. It implies that, in the one-dimensional case, the Gaussian distribution is learnable from truncated samples even if the mass of the truncation set is zero. This is, of course, provided that the variance of the distribution is not very small.

  On the next step it is an interesting question whether a similar result holds in higher dimensions. That is, we define an appropriate discretization for the multivariate Gaussian distribution. Is there an assumption (e.g. large variance) for it to be learnable and/or learnable from truncated samples?

- In the Mallows case a sufficient and necessary condition for learnability from truncated samples is yet to be specified. Our conjecture is that the algorithm 9 is not optimal in the truncated case. It is compelling to examine whether another algorithm can give a condition for S that fully characterizes the problem. A first direction would be to calculate the average permutation, i.e. rank in the first position the candidate that is mostly in the first position, etc. The work in [LM21] could also give some intuition on the correct approach for the problem.

  Moreover, the Mallows Model has another parameter which did not interest us in this thesis. Studying how ϕ is affected by truncation is imperative for a complete understanding of Mallows' behaviour.

Bibliography

[ABDH$^+$20] Hassan Ashtiani, Shai Ben-David, Nicholas JA Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *Journal of the ACM (JACM)*, 67(6):1–42, 2020.

[BF02] Steven J Brams and Peter C Fishburn. Voting procedures. *Handbook of social choice and welfare*, 1:173–236, 2002.

[Bil08] Patrick Billingsley. *Probability and measure.* John Wiley & Sons, 2008.

[Bil13] Patrick Billingsley. *Convergence of probability measures.* John Wiley & Sons, 2013.

[Bir97] Lucien Birgé. Estimation of unimodal densities without smoothness assumptions. *The Annals of Statistics*, pages 970–981, 1997.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press, 2013.

[Can20] Clément L Canonne. A short note on learning discrete distributions. *arXiv preprint arXiv:2002.11457*, 2020.

[CGS10] Louis HY Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein's method.* Springer Science & Business Media, 2010.

[Coh91] A Clifford Cohen. *Truncated and censored samples: theory and applications.* CRC press, 1991.

[CPS13] Ioannis Caragiannis, Ariel D Procaccia, and Nisarg Shah. When do noisy votes reveal the truth? In *Proceedings of the fourteenth ACM conference on Electronic commerce*, pages 143–160, 2013.

[CX02] KP Choi and Aihua Xia. Approximating the number of successes in independent trials: Binomial versus poisson. *The Annals of Applied Probability*, 12(4):1139–1148, 2002.

[DDS15] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A Servedio. Learning poisson binomial distributions. *Algorithmica*, 72(1):316–357, 2015.

[DGTZ18] Constantinos Daskalakis, Themis Gouleakis, Chistos Tzamos, and Manolis 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.

[DGTZ19] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Computationally and statistically efficient truncated regression. In *Conference on Learning Theory*, pages 955–960. PMLR, 2019.

[DKK$^+$19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

[DKS16] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Efficient robust proper learning of log-concave distributions. *arXiv preprint arXiv:1606.03077*, 2016.

[DKTZ21] Constantinos Daskalakis, Vasilis Kontonis, Christos Tzamos, and Emmanouil Zampetakis. A statistical taylor theorem and extrapolation of truncated densities. In *Conference on Learning Theory*, pages 1395–1398. PMLR, 2021.

[DMR18] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

[DP15] Constantinos Daskalakis and Christos Papadimitriou. Sparse covers for sums of indicators. *Probability Theory and Related Fields*, 162(3):679–705, 2015.

[DSYZ21] Constantinos Daskalakis, Patroklos Stefanou, Rui Yao, and Emmanouil Zampetakis. Efficient truncated linear regression with unknown noise variance. *Advances in Neural Information Processing Systems*, 34, 2021.

[FH19] Frank Feys and Helle Hvid Hansen. Arrow's theorem through a fixpoint argument. *arXiv preprint arXiv:1907.10381*, 2019.

[FKS21] Jacob Fox, Matthew Kwan, and Lisa Sauermann. Combinatorial anti-concentration inequalities, with applications. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 171, pages 227–248. Cambridge University Press, 2021.

[FKT20] Dimitris Fotakis, Alkis Kalavasis, and Christos Tzamos. Efficient parameter estimation of truncated boolean product distributions. In *Conference on Learning Theory*, pages 1586–1600. PMLR, 2020.

[Gal98] Francis 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(379-387):310–315, 1898.

[GS02] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[HK14] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

[HL19] Olivier Hudry and Antoine Lobstein. Unique (optimal) solutions: Complexity results for identifying and locating–dominating codes. *Theoretical Computer Science*, 767:83–102, 2019.

[HSV05] Edith Hemaspaandra, Holger Spakowski, and Jörg Vogel. The complexity of kemeny elections. *Theoretical Computer Science*, 349(3):382–391, 2005.

[Hub65] Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758, 1965.

[Hub92] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

[Hug62] Edwin Joseph Hughes. *Maximum likelihood estimation of distribution parameters from incomplete data.* Iowa State University, 1962.

[IZD20] Andrew Ilyas, Emmanouil Zampetakis, and Constantinos Daskalakis. A theoretical and practical framework for regression and classification from truncated samples. In *International Conference on Artificial Intelligence and Statistics*, pages 4463–4473. PMLR, 2020.

[JO19] Ayush Jain and Alon Orlitsky. Robust learning of discrete distributions from batches. *arXiv preprint arXiv:1911.08532*, 2019.

[KMR$^+$94] Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 273–282, 1994.

[KTZ19] Vasilis Kontonis, Christos Tzamos, and Manolis Zampetakis. Efficient truncated statistics with unknown truncation. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1578–1595. IEEE, 2019.

[LB11] Tyler Lu and Craig Boutilier. Learning mallows models with pairwise preferences. In *ICML*, 2011.

[LM21] Allen Liu and Ankur Moitra. Robust voting rules from algorithmic robust statistics. *arXiv preprint arXiv:2112.06380*, 2021.

[Mal57] Colin L Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.

[NP20] Sai Ganesh Nagarajan and Ioannis Panageas. On the analysis of em for truncated mixtures of two gaussians. In *Algorithmic Learning Theory*, pages 634–659. PMLR, 2020.

[SF12] Robert E Schapire and Yoav Freund. *Boosting: Foundations and Algorithms.* MIT Press, 2012.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

[Str] Gilbert Strang. Linear algebra and its applications. 2014.

[SZ13] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pages 71–79. PMLR, 2013.

[TT19] Wenpin Tang and Fengmin Tang. The poisson binomial distribution–old & new. *arXiv preprint arXiv:1908.10024*, 2019.

[Val84] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vap99] Vladimir Vapnik. *The nature of statistical learning theory.* Springer science & business media, 1999.

[VdMH08] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Wan93] Yuan H Wang. On the number of successes in independent trials. *Statistica Sinica*, pages 295–312, 1993.

[Xia19] Lirong Xia. Learning and decision-making from rank data. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 13(1):1–159, 2019.

[Zol76] Vladimir Mikhailovich Zolotarev. Metric distances in spaces of random variables and their distributions. *Mathematics of the USSR-Sbornik*, 30(3):373, 1976.