Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων
Εργαστηριο Λογικης και Επιστημης Υπολογιστων

Μάθηση διατάξεων από δείγματα με ελλιπή πληροφορία

Διπλωματικη Εργασια

του

κωνσταντινου Σ. σταυροπουλου

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

Αθήνα, Ιούλιος 2020

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

Learning rankings from incomplete samples

Διπλωματικη Εργασια

του

ΚΩΝΣΤΑΝΤΙΝΟΥ Σ. ΣΤΑΥΡΟΠΟΥΛΟΥ

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

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 13η Ιουλίου 2020.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Δημήτριος Φωτάκης	Αριστείδης Παγουρτζής	Αντώνιος Συμβώνης
Αναπληρωτής Καθηγητής	Αναπληρωτής Καθηγητής	Καθηγητής
Ε.Μ.Π.	Ε.Μ.Π.	Ε.Μ.Π.

Αθήνα, Ιούλιος 2020

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

(Υπογραφή)

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

Περίληψη

Αυτή η εργασία αποτελεί μία μελέτη πάνω στην εκμάθηση διατάξεων από θορυβώδη δείγματα τα οποία πιθανώς δεν περιέχουν όλα τα στοιχεία προς ταξινόμηση. Η εκμάθηση διατάξεων είναι ένα επίκαιρο πρόβλημα που συνδέεται στενά με τη Θεωρία Κοινωνικής Επιλογής, τα συστήματα Ψηφοφορίας και, γενικότερα, τη Μηχανική Μάθηση. Το βασικό πρόβλημα είναι η κατασκευή μίας διάταξης που είναι η πλέον ευρέως αποδεκτή, μέσω αξιοποίησης της πληροφορίας που περιέχεται σε ένα σύνολο διατάξεων εισόδου που γίνονται αντιληπτές ως ψήφοι ή δείγματα. Η έννοια της καθολικής αποδοχής αποκτά νόημα όταν εισάγεται κάποιο στατιστικό μοντέλο παραγωγής των διατάξεων εισόδου, ως ανεξάρτητα δείγματα. Συγκεκριμένα, ανάμεσα σε πολλά μοντέλα παραγωγής διατάξεων, επικεντρωνόμαστε στο μοντέλο Mallows, που στηρίζεται στην ιδέα της ύπαρξης μίας κεντρικής διάταξης που χαρατηρίζει την κατανομή της πιθανότητας εμφάνισης μίας διάταξης, μέσω μίας συνάρτησης απόστασης μεταξύ διατάξεων. Η πιθανότητα εμφάνισης μίας διάταξης μειώνεται εκθετικά στην απόσταση της διάταξης από την κεντρική διάταξη. Έχουν αναπτυχθεί διάφοροι αλγόριθμοι για την ανακατασκευή της κεντρικής διάταξης ή κάποιας εκτίμησής της από πλήρεις διατάξεις που αποτελούν ανεξάρτητα δείγματα της κατανομής Mallows. Ωστόσο, δεν είναι πάντοτε ρεαλιστική υπόθεση ότι μπορεί κανείς να έχει πρόσβαση σε δείγματα που είναι πλήρεις διατάξεις, καθώς συνήθως το πλήθος των εναλλακτικών είναι πολύ μεγάλο. Στην εργασία αυτή, εκτός από την παρουσίαση βασικών θεωρητικών θεμελίων και ορισμένων αποτελεσμάτων σχετιζόμενων με την μάθηση στο μοντέλο Mallows, προτείνεται ένα γενικευμένο μοντέλο για δείγματα που δεν είναι απαραίτητα πλήρεις διατάξεις, αλλά διατηρεί την έννοια της κεντρικής διάταξης. Επίσης, παρέχονται αυστηρά φράγματα για τη δειγματική πολυπλοκότητα ανακατασκευής της κεντρικής διάταξης σε ορισμένες παραλλαγές του γενικευμένου μοντέλου και παρουσιάζεται ένας αλγόριθμος για τον αποδοτικό υπολογισμό της εκτίμησης μέγιστης πιθανοφάνειας της κεντρικής διάταξης από θορυβώδη δείγματα με ελλιπή πληροφορία.

Λέξεις Κλειδιά

Στατιστική Μάθηση, Μάθηση Κατανομών, Θεωρία Μάθησης, Θεωρία Πιθανοτήτων, Θεωρία Κοινωνικής Επιλογής, Κατανομές Διατάξεων, Μοντέλο Mallows

Abstract

This thesis constitutes a study of learning rankings from noisy and incomplete samples, in the sense that they may not contain all the alternatives to be ranked. Learning on rankings is an emergent problem that is closely liked to Social Choice Theory, Voting Systems and, generally, Machine Learning. The fundamental problem is the construction of a ranking that is the most widely acceptable, using information provided by a set of input rankings, that are considered to be votes or samples. In particular, while many models exist describing the sampling process of rankings, this study focuses on Mallows model, which is based on the idea of a central ranking that determines the probability of sampling a ranking through a notion of distance between rankings. The probability that a specific ranking is sampled diminishes exponentially to its distance from the central ranking. Many learning algorithms have been developed for estimating the central ranking under the Mallows model. However, it is often unrealistic to assume that the samples consist of full rankings, especially when the number of alternatives to be ranked is large. In this work, except from presenting the fundamental theoretical background and some of the main results concerning learning the central ranking under the existing models, a new model is proposed, namely the Selective Mallows model, in which sampling incomplete rankings is possible, but also the concept of the central ranking is preserved. Furthermore, strong bounds for sample complexity of learning the central permutation under some variants of the Selective Mallows model are established, as well as an efficient algorithm for estimating the maximum likelihood estimation of central ranking under the selective Mallows model.

Keywords

Statistical Learning, Learning Theory, Distribution Learning, Probability Theory, Computational Learning, Social Choice Theory, Voting Theory, Mallows Model, Ranking Distributions

στην οικογένειά μου

Ευχαριστίες

Κλείνοντας με την εργασία αυτή τον κύκλο των προπτυχιακών σπουδών μου, δεν μπορώ παρά να είμαι ευγνώμων σε όλους εκείνους που συνέβαλαν έμμεσα ή άμεσα στην ολοκλήρωση του κύκλου αυτού, αλλά και που έκαναν την πορεία μου σε αυτόν ουσιώδη και διασκεδαστική.

Θα ήθελα αρχικά να ευχαριστήσω τον κύριο Φωτάκη, για όλη την υποστήριξη, την εμπιστοσύνη, αλλά και την έμπνευση που μου παρείχε ακόμη και πριν από τη συνεργασία μας, στο αμφιθέατρο. Επίσης, θα ήθελα να ευχαριστήσω θερμά τον Άλκη Καλαβάση για την διαρκή υποστήριξή του κατά τη διάρκεια εκπόνησης αυτής της εργασίας, την έμπρακτη συμβολή του σε αυτήν αλλά και για όλες τις ενδιαφέρουσες συζητήσεις μας. Ευχαριστώ τα μέλη της εξεταστικής επιτροπής, τον κύριο Παγουρτζή και τον κύριο Συμβώνη, όχι μόνο ως εξεταστές αλλά και για όσα μου προσέφεραν ως καθηγητές μου. Γενικότερα, είμαι ευγνώμων σε κάθε καθηγητή μου, διότι από τον καθένα αποκόμισα ένα διαφορετικό τρόπο σκέψης, ιδέες, γνώση και κίνητρο για προσπάθεια.

Έπειτα, θέλω να ευχαριστήσω τους φίλους μου για τη διαχρονική τους παρουσία και το ενδιαφέρον τους καθώς και όλους τους συμφοιτητές μου και ιδιαίτερα εκείνους με τους οποίους είχα την ευκαιρία να συνεργαστώ, να συζητήσω ή να γνωρίσω προσωπικά. Κώστα, Μαρίνα, Πέτρο, Δημήτρη Χ., Δημήτρη Ξ., Σέβη, Πάνο Κ., Πάνο Σ., Μάριε, Αργύρη, Δημήτρη Κ., Ιωάννα, Μιχαήλ, Γιάννη Μ., Βασίλη, Αλέξανδρε Μ., Γιάννη Δ., Γιάννη Α., Ηλία, Αλέξανδρε Σ. σας ευχαριστώ!

Ίσως, όμως, λίγο περισσότερο, θα ήθελα να ευχαριστήσω την οικογένειά μου, για την απεριόριστη, ανιδιοτελή αγάπη τους.

Contents

List of Figures

List of Tables

Κεφάλαιο 1

Εκτεταμένη Ελληνική Περίληψη

Στο κεφάλαιο αυτό παρουσιάζονται τα βασικότερα σημεία της παρούσας εργασίας εν συντομία.

1.1 Εισαγωγή

Οι κατανομές κατάταξης αποτελούν αντικείμενο εντατικής μελέτης τα τελευταία χρόνια. Χρησιμοποιούνται για την μοντελοποίηση διαφόρων προβλημάτων όπως η συνάθροιση προτιμήσεων και οι ψηφοφορίες. Επομένως, συνδέονται με φυσικό τρόπο με τη θεωρία κοινωνικής επιλογής, παρότι διαθέτουν επιπλέον ιδιότητες. Η θεωρία της κοινωνικής επιλογής στοχεύει στην καθιέρωση κανόνων συνάθροισης που ικανοποιούν ορισμένα επιθυμητά αξιώματα και οι οποίοι μπορούν να υπολογιστούν αποδοτικά. Οι προτιμήσεις γίνονται αντιληπτές ως διατάξεις πάνω σε κάποιο σύνολο εναλλακτικών. Ωστόσο, λόγω ορισμένων αποτελεσμάτων ανεφικτότητας, όπως για παράδειγμα το θεμελιώδες αποτέλεσμα του Arrow (Arrow [1951]), που καταλήγουν στο ότι δεν είναι δυνατόν να ικανοποιούνται ταυτόχρονα ορισμένα σύνολα αξιωμάτων από κανέναν κανόνα συνάθροισης, κέρδισε έδαφος η τάση να ιδωθεί το πρόβλημα από μία διαφορετική οπτική γωνία. Οι προτιμήσεις θεωρούνται πλέον δείγματα από κάποια κατανομή κατάταξης και το πρόβλημα συνάθροισης ανάγεται στο πρόβλημα μάθησης αγνώστων παραμέτρων της κατανομής μοντελοποίησης.

Μία από τις πλέον μελετημένες κατανομές κατάταξης είναι η κατανομή Mallows. Στο μοντέλο Mallows υπάρχει η έννοια της κεντρικής διάταξης, που αποτελεί την επικρατούσα τιμή της κατανομής και η πιθανότητα εμφάνισης κάθε άλλης διάταξης μειώνεται εκθετικά στην απόστασή της από την κεντρική διάταξη. Τέτοιες κατανομές καλούνται κατανομές βασισμένες σε αποστάσεις. Υπάρχουν διάφορες αποστάσεις μεταξύ διατάξεων. Ωστόσο, μία από τις πιο χρήσιμες και αυτή που θα χρησιμοποιήσουμε σε αυτήν την εργασία, είναι η απόσταση Kendall tau, η οποία ισούται με το πλήθος των ζευγών εναλλακτικών στα οποία δύο διατάξεις διαφωνούν. Το μοντέλο Mallows συνδέεται επίσης με τον κανόνα του Kemeny για συνάθροιση προτιμήσεων, που έχει ιδιαίτερη σημασία στο πλαίσιο της θεωρίας κοινωνικής επιλογής, καθώς είναι συνδεδεμένος με κάποια έννοια βελτιστότητας: αντιστοιχεί στην εύρεση μίας διάταξης για την οποία ο συνολικός αριθμός ζευγο-διαφορών με τις διατάξεις εισόδου είναι ελάχιστος. Ωστόσο, ο υπολογισμός αυτού του κανόνα έχει αποδειχθεί ότι είναι ένα NP-δύσκολο πρόβλημα στην χειρότερη περίπτωση. Είναι ενδιαφέρον να αναφέρουμε ότι υπό το μοντέλο Mallows , το πρόβλημα γίνεται εύκολο, πράγμα που είναι εφικτό μιας και βρισκόμαστε στην μέση περίπτωση.

Ωστόσο, η υπόθεση ότι οι διατάξεις εισόδου είναι πλήρεις δεν είναι πάντα ρεαλιστική. Ιδιαίτερα σε περιπτώσεις όπου ο αριθμός των εναλλακτικών είναι πολύ μεγάλος, το να υποθέσουμε ότι

έχουμε πρόσβαση σε πλήρεις διατάξεις είναι υπερβολικά αισιόδοξο. Ο στόχος μας, πάντως, παραμένει να μάθουμε την πλήρη κεντρική διάταξη, χρήση δειγμάτων που είναι ελλιπείς διατάξεις. Γι΄ αυτόν το λόγο, προτείνουμε το Selective Mallows μοντέλο, που γενικεύει το μοντέλο Mallows, επιτρέποντας δειγματοληψία ελλιπών διατάξεων. Και σε αυτήν την περίπτωση, υπάρχει μία κεντρική διάταξη, αλλά το νόημά της είναι ελαφρώς διαφορετικό: Αφού επιλέξουμε ένα σύνολο εναλλακτικών προς διάταξη σε ένα δείγμα, η κεντρική διάταξη περιορίζεται στο σύνολο αυτό και ένα δείγμα λαμβάνεται από την κατανομή Mallows με την περιορισμένη διάταξη ως επικρατούσα τιμή. Ένα Selective Mallows δείγμα θεωρείται, εν γένει, ανεξάρτητο από τα υπόλοιπα, δεδομένων των συνόλων επιλογής του προφίλ δειγμάτων. Διαφορετικές διαδικασίες επιλογής συνόλων αντιστοιχούν σε διαφορετικά περιβάλλοντα εφαρμογής. Για παράδειγμα, τα σύνολα μπορεί να επιλέγονται από έναν αντίπαλο, τυχαία ή προσαρμοστικά, αν μάς δίνεται ένας Selective Mallows δειγματολήπτης, που λαμβάνει στην είσοδο υποσύνολα του συνόλου όλων των εναλλακτικών και επιστρέφει ένα ανεξάρτητο δείγμα της αντίστοιχης (περιορισμένης) Mallows κατανομής.

Το selective Mallows μοντέλο λαμβάνει υπόψιν αυτό που καλούμε *μεροληψία λόγω άγνοιας*: Αν ο δειγματολήπτης δεν διαθέτει γνώση για κάποια εναλλακτική, τότε η πιθανότητα να ταξινομήσει λανθασμένα ένα ζευγάρι στοιχείων είναι μεγαλύτερη αν η θέση της άγνωστης εναλλακτικής είναι ανάμεσα στο ζευγάρι. Αυτή η υπόθεση είναι χρήσιμη σε περιπτώσεις που οι εναλλακτικές δεν έχουν ατομικές αξίες, ή οι αξίες τους είναι γενικώς απροσδιόριστες, οπότε κατατάσσονται μόνο με βάση συγκρίσεις ζευγών.

Επιπλέον, το πρόβλημα υπολογισμού μίας εκτίμησης μέγιστης πιθανοφάνειας για την κεντρική διάταξη στο γενικευμένο μοντέλο αποτελεί γενίκευση του προβλήματος Kemeny. Επομένως, η γενίκευση που προτείνουμε είναι κατά κάποιο τρόπο φυσική, αφού τα προβλήματα εκτίμησης μέγιστης πιθανοφάνειας έχουν κατ΄ ουσίαν την ίδια δομή.

Επίσης, το selective Mallows μοντέλο είναι ελαχιστικό, γιατί παρέχει ελάχιστη πληροφορία, με την έννοια ότι η θέση των μη επιλεγμένων εναλλακτικών στην κεντρική διάταξη, δεν επηρεάζει την πιθανότητα παρατήρησης καμίας πιθανής (περιορισμένης) διάταξης.

Σε αυτήν την διπλωματική εργασία, θεωρούμε το πρόβλημα ανεύρεσης της κεντρικής διάταξης από ελλιπή δείγματα καθώς και το πρόβλημα ανακατασκευής μίας εκτίμησης μέγιστης πιθανοφάνειας για την κεντρική διάταξη, για οποιοδήποτε πλήθος δειγμάτων εισόδου. Παρέχουμε αυστηρά φράγματα για την περίπτωση αντιπάλου και την τυχαία περίπτωση και γενικεύουμε τον αλγόριθμο που προτείνεται από τους Braverman and Mossel [2009], για την περίπτωση ελλιπών δειγμάτων. Αφήνουμε ανοικτό το πρόβλημα εύρεσης αυστηρών φραγμάτων δειγματικής πολυπλοκότητας για την προσαρμοστική περίπτωση.

1.2 Θεωρητικό υπόβαθρο

Σε αυτήν την ενότητα, παρουσιάζουμε θεμελιώδεις έννοιες που αποτελούν τη βάση των αποτελεσμάτων που παρουσιάζουμε στις επόμενες ενότητες. Συγκεκριμένα, παρουσιάζουμε αφενός ορισμένα εργαλεία από τη θεωρία πιθανοτήτων και αφετέρου, στοιχεία από τη θεωρία υπολογιστικής μάθησης.

Θεωρία πιθανοτήτων. Η θεωρία πιθανοτήτων είναι το εργαλείο με το οποίο γίνεται δυνατή η ανάλυση και η χρήση της τυχαιότητας, που αποτελεί μία έννοια που έχει απασχολήσει τον άνθρωπο από τις αρχές της ιστορίας.

Ένα από τα βασικότερα εργαλεία της θεωρίας πιθανοτήτων είναι οι ανισότητες συγκέντρωσης. Πρόκειται για εργαλεία που χρησιμοποιούνται για να αποδείξει κανείς ότι ορισμένες τυχαίες μεταβλητές έχουν ικανοποιητικά προβλέψιμη συμπεριφορά. Γενικά, η προβλεψιμότητα συνδέεται με την επανάληψη ανεξαρτήτων πειραμάτων. Παραθέτουμε, λοιπόν, το φράγμα Chernoff-Hoeffding:

Theorem 1.2.1: Chernoff-Hoeffding φράγμα

Έστω $X_1, X_2, \ldots, X_n$, $n \in \mathbb{N}$ *ανεξάρτητες τυχαίες μεταβλητές Bernoulli. Αν* $X = \sum_{i\in[n]} X_i$ *και* $\mu = \mathbb{E}[X]$, *τότε:*

$$Pr[X \geq \mu + a] \leq e^{-nD_{KL}(\frac{\mu+a}{n}||\frac{\mu}{n})}, 0 < a < n - \mu$$

και

$$Pr[X \leq \mu - a] \leq e^{-nD_{KL}(1-\frac{\mu+a}{n}||1-\frac{\mu}{n})}, 0 < a < \mu,$$

όπου $D_{KL}(p||q) = p\log(p/q) + (1-p)\log(\frac{1-p}{1-q}), \forall p, q \in (0,1)$.

Ένα ακόμη σημαντικό εργαλείο είναι η πιθανοτική μέθοδος. Πρόκειται για μία μέθοδο απόδειξης της ύπαρξης ενός αντικειμένου με κάποιες επιθυμητές ιδιότητες. Συγκεκριμένα, αν μπορεί να οριστεί ένας πιθανοτικός χώρος όπου η πιθανότητα να επιλέξουμε ένα στοιχείο που έχει τις επιθυμητές ιδιότητες είναι αυστηρά θετική, τότε είναι βέβαιο πως ένα τέτοιο αντικείμενο υπάρχει.

Θεωρία υπολογιστικής μάθησης. Η μάθηση είναι η διαδικασία μετατροπής της εμπειρίας σε γνώση. Ενώ αποτελεί γενικά ένα διεπιστημονικό πεδίο έρευνας, η μελέτη της μάθησης από μαθηματικής σκοπιάς κατέστη δυνατή μέσω του μοντέλου που πρότεινε ο Valiant [1984]. Σήμερα, υπάρχουν διάφορα μαθηματικά πλαίσια που χρησιμοποιούνται για την μελέτη αντίστοιχων προβλημάτων μάθησης. Σε αυτήν την εργασία, επικεντρωνόμαστε στο πλαίσιο της Μάθησης Κατανομών.

Θεωρούμε ένα σύνολο $\mathcal{X}$ και μία κατανομή $\mathcal{D}$ πάνω στο $\mathcal{X}$. Λαμβάνουμε ανεξάρτητα δείγματα από την $\mathcal{D}$. Το πρόβλημα εκμάθησης κατανομών, όπως ορίστηκε από Kearns et al. [1994a], είναι το ακόλουθο:

Definition 1.2.1 (Μάθηση κατανομών): *Έστω* $\mathfrak{D}$ *μία κλάση κατανομών στον* $\mathcal{X}$. *Τότε η* $\mathfrak{D}$ *καλείται επαρκώς κατάλληλη για μάθηση ως προς κάποια μετρική* d *ανάμεσα σε κατανομές πιθανότητας, αν για κάθε* $\epsilon, \delta \in (0,1)$, *υπάρχει ένας αλγόριθμος πολυωνυμικού χρόνου που, δοσμένης πρόσβασης σε ένα δειγματολήπτη οποιασδήποτε συγκεκριμένης αλλά άγνωστης κατανομής* $\mathcal{D} \in \mathfrak{D}$, *επιστρέφει μία κατανομή* $\mathcal{D}'$ *η οποία ικανοποιεί την ανισότητα:*

$$Pr[d(\mathcal{D}, \mathcal{D}') \geq \epsilon] \leq \delta.$$

Αν $\mathcal{D}' \in \mathfrak{D}$, *τότε ο αλγόριθμος λέγεται κατάλληλος (αλλιώς ακατάλληλος).*

Είναι συνήθης πρακτική να παραμετροποιούμε την κλάση υποθέσεων $\mathfrak{D}$ με ορισμένες παραμέτρους. Τα δείγματα εισόδου χρησιμοποιούνται για να εκτιμήσουμε τις τιμές των παραμέτρων αυτών, χωρίς να χρειάζεται απαραίτητα να βρούμε τις ακριβείς τιμές τους. Γενικά, όσο οι εκτιμήσεις των παραμέτρων πλησιάζουν τις σωστές τιμές, τόσο πλησιάζει και η εκτιμώμενη κατανομή την αρχική. Μία βέλτιστη μέθοδος εκτίμησης παραμέτρων είναι η μέθοδος μέγιστης

πιθανοφάνειας, όπου η εκτίμηση της παραμέτρου επιλέγεται έτσι ώστε η πιθανότητα να έρθουν τα δείγματα που παρατηρούμε να είναι η μέγιστη δυνατή, πάνω σε όλες τις δυνατές επιλογές για την τιμή της παραμέτρου.

1.3 Κατανομές διατάξεων

Υποθέτουμε ότι A είναι ένα σύνολο εναλλακτικών που περιέχει $n \in \mathbb{N}$ στοιχεία. Μία διάταξη είναι μία ένα προς ένα και επί συνάρτηση $\pi : A \to A$. Συμβολισμός: $\pi(i) < \pi(j) \Leftrightarrow i \succ_\pi j$. Το σύνολο $\mathfrak{S}_A$ είναι το σύνολο όλων των διατάξεων στοιχείων του A. Περιορισμένη διάταξη: $\pi|_B$ είναι το στοιχείο του $B \subseteq A$ για το οποίο: $sgn(\pi|_B(i) - \pi|_B(j)) = sgn(\pi(i) - \pi(j)), \forall i, j \in B$.

Μία συνάρτηση απόστασης μεταξύ διατάξεων που είναι ευρέως χρησιμοποιούμενη, είναι η απόσταση Kendall tau. Ισούται με το πλήθος των ζευγαριών που είναι διαφορετικά ταξινομημένα στις δύο διατάξεις. Τυπικά, έχουμε:

$$d_{KT}(\pi, \pi') = |\{i < j : (\pi(i) - \pi(j))(\pi'(i) - \pi'(j)) < 0\}|$$

Κατανομή Mallows. Η κατανομή διατάξεων όπου επικεντρωνόμαστε σε αυτή την εργασία, είναι η κατανομή Mallows. Βασίζεται στην έννοια της κεντρικής διάταξης $\pi_0 \in \mathbb{N}$, που αποτελεί την επικρατούσα τιμή της κατανομής και σε μία παράμετρο εξάπλωσης $\beta > 0$, ως ακολούθως:

$$\Pr[\pi] = \frac{1}{Z} e^{-\beta d_{KT}(\pi_0, \pi)},$$

όπου Z είναι μία σταθερά κανονικοποίησης. Αν διαθέτουμε r ανεξάρτητα δείγματα μίας κατανομής Μαλλοως, τότε η εκτίμηση μέγιστης πιθανοφάνειας για την κεντρική διάταξη συμπίπτει με την συνάθροιση του Kemeny:

$$\pi^* = \arg\max_{\pi \in \mathfrak{S}_n} \sum_{\ell \in [r]} d_{KT}(\pi_0, \pi_\ell)$$

Ο υπολογισμός της συνάθροισης Kemeny είναι γενικά ένα NP-δύσκολο πρόβλημα.

Μοντέλο selective Mallows. Γενικεύουμε το μοντέλο Mallows ώστε να μπορούμε να λάβουμε μη πλήρεις διατάξεις:

$$\Pr[\pi | \pi_0, \beta, S] = \frac{1}{Z(S)} e^{-\beta d_{KT}(\pi_0|_S, \pi)}, \forall \pi \in \mathfrak{S}_S$$

Καλούμε το σύνολο $S \subseteq [n]$ σύνολο επιλογής. Γενικά, σε ένα δειγματικό προφίλ που αποτελείται από περισσότερες της μίας μη πλήρεις διατάξεις, θεωρούμε τους ακόλουθους τρόπους παραγωγής των συνόλων επιλογής:

1. *Ανταγωνιστικά:* Σε αυτήν την περίπτωση, τα σύνολα επιλέγονται από κάποιον αντίπαλο. Ενδεχομένως να έχουν επιλεγεί με τέτοιο τρόπο ώστε να εξαρτώνται το ένα από το άλλο. Συμβολίζουμε το αντίστοιχο μοντέλο με $\mathcal{M}^{\text{ADV}}_{\pi_0,\beta}$.
2. *Τυχαία:* Θεωρούμε ότι τα σύνολα επιλογής είναι ανεξάρτητα δείγματα από κάποια κατανομή $\mathcal{D}$ επί του $2^{[n]}$. Συμβολίζουμε το μοντέλο με $\mathcal{M}^{\text{RND}(\mathcal{D})}_{\pi_0,\beta}$.

3. *Προσαρμοστικά:* Σε αυτήν την περίπτωση, έχουμε πρόσβαση σε ένα δειγματολήπτη selective Mallows δειγμάτων, στην είσοδο του οποίου τοποθετούμε σύνολα επιλογής μεγέθους $m \leq n$ και στην έξοδο λαμβάνουμε δείγματα της αντίστοιχης (περιορισμένης) Mallows κατανομής. Συμβολισμός: $\mathcal{M}_{\pi_0,\beta}^{\text{ADP}(m)}$.

Δεδομένων των συνόλων επιλογής, και σε αυτήν την περίπτωση, η εκτίμηση μέγιστης πιθανοφάνειας για την κεντρική διάταξη, προκύπτει από την επίλυση του γενικευμένου κανόνα του Kemeny:

$$\pi^* = \arg\min_{\pi \in \mathfrak{S}_n} \sum_{\ell \in [r]} d_{KT}(\pi_0|_{S_\ell}, \pi_\ell)$$

1.4 Μαθαίνοντας μία κρυμμένη διάταξη

Στην ενότητα αυτή παρουσιάζουμε αποτελέσματα που αφορούν την δειγματική πολυπλοκότητα του προβλήματος εύρεσης μίας κρυφής διάταξης από θορυβώδη δείγματα. Αρχικά, παρουσιάζουμε αποτελέσματα από τη βιβλιογραφία αναφορικά με την περίπτωση που τα δείγματα είναι πλήρεις διατάξεις αλλά και την περίπτωση που είναι συγκρίσεις ζευγών. Στη συνέχεια, παρουσιάζουμε τα δικά μας αποτελέσματα, τα οποία αφορούν το selective Mallows μοντέλο, σε τρεις διαφορετικές τροποποιήσεις του: την ανταγωνιστική, την τυχαία και την προσαρμοστική επιλογή συνόλων. Για την ανταγωνιστική και την τυχαία περίπτωση, παρέχουμε αυστηρά φράγματα για τη δειγματική πολυπλοκότητα, ενώ για την προσαρμοστική περίπτωση παρέχουμε ένα άνω φράγμα το οποίο είναι απόρροια των αποτελεσμάτων για την περίπτωση που τα δείγματα είναι συγκρίσεις ζευγών.

Mallows model. Στην εργασία των Caragiannis et al. [2013], παρέχονται αυστηρά φράγματα για τη δειγματική πολυπλοκότητα του προβλήματος εύρεσης της κεντρικής διάταξης από ανεξάρτητα δείγματα της κατανομής Mallows. Συγκεκριμένα, έχουμε το ακόλουθο άτυπο θεώρημα:

Informal theorem 1.4.1

Έστω $\mathcal{M}_{\pi_0,\beta}$ μία κατανομή Mallows με κεντρική διάταξη $\pi_0 \in \mathfrak{S}_n$ και παράμετρο εξάπλωσης $\beta > 0$. Για κάθε $\epsilon > 0$, υπάρχει ένας αλγόριθμος που, δοσμένου ενός δειγματικού προφίλ από την $(\mathcal{M}_{\pi_0,\beta})^r$ για κάθε r τουλάχιστον ίσο με κάποια τιμή $O((1-e^{-\beta})^{-2}\log(n/\epsilon))$, ανακαλύπτει την κεντρική διάταξη π_0 με πιθανότητα τουλάχιστον $1-\epsilon$.[α΄] Επιπλέον, αν $r = o(\frac{1}{\beta}\log(n/\epsilon))$, τότε κανένας αλγόριθμος δεν μπορεί να εγγυηθεί πιθανότητα επιτυχίας $1-\epsilon$.

[α΄] $(1-e^{-\beta})^{-2} = O(1/\beta^2)$ όταν $\beta \to 0$

Ο αλγόριθμος που χρησιμοποιείται για το άνω φράγμα, είναι η δημιουργία ενός κατευθυνόμενου γραφήματος όπου κάθε κορυφή αντιστοιχεί σε μία εναλλακτική και η κατεύθυνση κάθε ακμής αντιστοιχεί στην διάταξη για τις αντίστοιχες εναλλακτικές που υποδεικνύει η πλειοψηφία των δειγμάτων. Αποδεικνύεται, λόγω του ότι η πιθανότητα ένα ζεύγος να ταξινομηθεί λάθος είναι άνω φραγμένη από την τιμή $e^{-\beta}/(1-e^{-\beta})$ και το αποτέλεσμα προκύπτει με εφαρμογή της ανισότητας ένωσης και του φράγματος Hoeffding.

Το αποτέλεσμα είναι ασυμπτωτικά σφιχτό, γιατί τα ζεύγη στοιχείων που είναι διπλανά στην κεντρική διάταξη, έχουν πιθανότητα λάθος ταξινόμησης ακριβώς $e^{-\beta}/(1-e^{-\beta})$ και το πλήθος τους είναι n. Προκειμένου να εξασφαλιστεί ότι κάθε τέτοιο ζεύγος θα ταξινομηθεί σωστά, πρέπει το πλήθος δειγμάτων να είναι αρκετά μεγάλο ώστε να δούμε τα ζεύγη αυτά αρκετές φορές. Σημειώνουμε ότι εντός της συνάρτησης λογαρίθμου, το n δεν ξεχωρίζει ασυμπτωτικά από το n^2, που είναι και ο λόγος που το φράγμα είναι σφιχτό.

Θορυβώδεις συγκρίσεις. Η περίπτωση που κάθε δείγμα αποτελείται από μοναδικό ζεύγος στοιχείων αντιστοιχεί στο μοντέλο θορυβωδών συγκρίσεων. Στο πλαίσιο επίλυσης συγκριτικών προβλημάτων με χρήση συγκρίσεων με θόρυβο (κάθε σύγκριση μπορεί να είναι λανθασμένη με πιθανότητα $1/2-\gamma$, όπου $\gamma \in (0,1/2)$), οι Feige et al. [1994], έδειξαν το ακόλουθο αποτέλεσμα:

Informal theorem 1.4.2

Επαρκούν $O(n\log(n/\epsilon))$ θορυβώδεις συγκρίσεις ώστε να μάθουμε την κρυμμένη διάταξη με πιθανότητα τουλάχιστον $1-\epsilon$.

Προφανώς, το παραπάνω άνω φράγμα για τη δειγματική πολυπλοκότητα είναι αυστηρό ως προς n, καθώς η συγκριτική ταξινόμηση χρειάζεται $\Omega(n\log n)$ συγκρίσεις.

Ο τρόπος με τον οποίο επιτυγχάνεται το παραπάνω αποτέλεσμα είναι με εκμετάλλευση του γεγονότος ότι η δυαδική αναζήτηση, ακόμη και στην περίπτωση που έχουμε θορυβώδεις συγκρίσεις, μπορεί να υλοποιηθεί με τέτοιο τρόπο ώστε να χρειάζονται $O(\log(n/\epsilon))$ συγκρίσεις για να ολοκληρωθεί με πιθανότητα επιτυχίας τουλάχιστον $1-\epsilon$. Συγκεκριμένα, υλοποιούμε την δυαδική αναζήτηση ως έναν τυχαίο περίπατο στο δέντρο της δυαδικής αναζήτησης, όπου κάθε κόμβος αντιστοιχεί σε ένα διάστημα. Οι συγκρίσεις γίνονται όχι μόνο για να αποφασίσουμε σε ποιο υποδιάστημα του τρέχοντος διαστήματος θα μεταβούμε (σύγκριση με μεσσαίο στοιχείο) αλλά και για να ελέγξουμε, κάθε φορά που φτάνουμε σε έναν κόμβο, αν το στοιχείο προς αναζήτηση ανήκει στο αντίστοιχο διάστημα (σύγκριση με άκρα του διαστήματος): διαφορετικά, επιστρέφουμε στον παραπάνω κόμβο. Έτσι, τα σφάλματα διορθώνονται και γλιτώνουμε την λογαριθμική ασυμπτωτική επιβάρυνση. Μένει να εκμεταλλευτούμε την ιδιότητα αυτή, κάνοντας κάποια προεπεξεργασία και κάποια επεξεργασία εκ των υστέρων.

Selective Mallows model. Στο γενικευμένο μοντέλο αποδεικνύουμε τα ακόλουθα φράγματα.

Αρχικά, παρέχουμε σφιχτά ασυμπτωτικά φράγματα για την δειγματική πολυπλοκότητα στην ανταγωνιστική περίπτωση, συναρτήσει της παραμέτρου συχνότητας, που ισούται με τον ελάχιστο λόγο του πλήθους εμφανίσεων κάποιου ζεύγους προς το πλήθος των δειγμάτων:

Informal theorem 1.4.3

Η δειγματική πολυπλοκότητα για την εύρεση της κεντρικής διάταξης στο μοντέλο $\mathcal{M}^{\mathrm{ADV}}_{\pi_0,\beta}$ με πιθανότητα τουλάχιστον $1-\epsilon$ είναι poly$(1/\beta)\Theta(\frac{1}{p}\log(n/\epsilon))$, όπου $p \in (0,1]$ είναι η παράμετρος συχνότητας, $\pi_0 \in \mathfrak{S}_n$, $\beta > 0$ και $\epsilon \in (0,1)$.

Το άνω φράγμα προκύπτει με τρόπο εντελώς παρόμοιο με αυτόν που χρησιμοποιήθηκε για την περίπτωση που τα δείγματα είναι πλήρη. Η μόνη διαφορά είναι ότι η εγγύηση που έχουμε σε αυτήν την περίπτωση είναι συναρτήσει του pr και όχι του r.

Το κάτω φράγμα προκύπτει από την παρατήρηση ότι αν το πλήθος των διαθέσιμων συγκρίσεων είναι $o(n^2 \log(n/\epsilon))$, τότε υπάρχει ένα σύνολο από $n/2$ ζεύγη τα οποία δεν έχουν κοινά άκρα και το πλήθος συγκρίσεών τους είναι $o(\log(n/\epsilon))$. Τότε, αν η κεντρική διάταξη είναι τέτοια ώστε καθένα από τα ζεύγη αυτά είναι διαδοχικά τοποθετημένα (που είναι εφικτό), κανένας αλγόριθμος δεν μπορεί να εγγυηθεί πιθανότητα επιτυχίας τουλάχιστον $1 - \epsilon$. Μπορούμε, τέλος, να κατασκευάσουμε ένα διάνυσμα συνόλων επιλογής με συχνότητα p το οποίο δεν περιέχει πάνω από $2prn^2$ συγκρίσεις.

Αντίστοιχα φράγματα αποκτούμε για την τυχαία περίπτωση, όπου η παράμετρος συχνότητας αντιστοιχεί στην ελάχιστη πιθανότητα εμφάνισης ενός ζεύγους στοιχείων.

Informal theorem 1.4.4

Η δειγματική πολυπλοκότητα για την εύρεση της κεντρικής διάταξης στο μοντέλο $\mathcal{M}_{\pi_0,\beta}^{\mathrm{RND}(\mathcal{D})}$ *με πιθανότητα τουλάχιστον* $1 - \epsilon$ *είναι* $poly(1/\beta)\Theta(\frac{1}{p}\log(n/\epsilon))$*, όπου* $p \in (0, 1]$ *είναι η παράμετρος συχνότητας,* $\pi_0 \in \mathfrak{S}_n$*,* $\beta > 0$ *και* $\epsilon \in (0, 1)$*.*

Το άνω φράγμα προκύπτει και σε αυτήν την περίπτωση με αντίστοιχο τρόπο, με την προσθήκη ότι χρησιμοποιούμε για κάθε ζεύγος τον νόμο ολικής πιθανότητας για το πλήθος εμφανίσεών του.

Το κάτω φράγμα προκύπτει αν επιλέξουμε την κατανομή επιλογής $\mathcal{D}$ που επιλέγει κάθε στοιχείο ανεξάρτητα, έτσι ώστε κάθε ζεύγος στοιχείων να επιλέγεται με πιθανότητα p. Γενικά, αν ένα ζεύγος που αποτελείται από διαδοχικά στοιχεία της κεντρικής διάταξης δεν εμφανιστεί ποτέ, δεν μπορούμε παρά να το ταξινομήσουμε τυχαία. Με βάση αυτήν την παρατήρηση και κατάλληλο τεχνικό χειρισμό, προκύπτει το ως άνω αποτέλεσμα.

Τέλος, για την ανταγωνιστική περίπτωση, προκύπτει ως συμπέρασμα των αποτελεσμάτων που παρουσιάζονται στο Feige et al. [1994] το ακόλουθο αποτέλεσμα:

Informal theorem 1.4.5

Η δειγματική πολυπλοκότητα για την εύρεση της κεντρικής διάταξης στο μοντέλο $\mathcal{M}_{\pi_0,\beta}^{\mathrm{ADP}(m)}$ *με πιθανότητα τουλάχιστον* $1-\epsilon$ *είναι* $O(\frac{n}{m}\log(n/\epsilon))$*, όπου* $\pi_0 \in \mathfrak{S}_n$*,* $\beta > 0$ *και* $\epsilon \in (0, 1)$*.*

Συγκεκριμένα, ομαδοποιούμε τις συγκρίσεις ζευγών σε πακέτα μεγέθους $m/2$ ζευγών και για κάθε πακέτο χρησιμοποιούμε ένα δείγμα. Αυτό είναι δυνατόν, διότι ο αλγόριθμος θορυβώδους ταξινόμησης χρησιμοποιεί $O(\log(n/\epsilon))$ παράλληλα βήματα, οπότε μπορούμε να ομαδοποιήσουμε τις συγκρίσεις.

Ωστόσο, στην προσαρμοστική περίπτωση είναι ανοιχτό το πρόβλημα εύρεσης ασυμπτωτικά σφιχτών φραγμάτων για τη δειγματική πολυπλοκότητα.

1.5 Προβλήματα ανακατασκευής Mallows

Στην προσπάθεια να εκτιμήσουμε την κεντρική διάταξη της κατανομής Mallows, η βέλτιστη στρατηγική είναι να χρησιμοποιήσουμε τη μέθοδο μέγιστης πιθανοφάνειας. Όπως, ωστόσο ήδη αναφέραμε, το πρόβλημα αυτό είναι, στην συγκεκριμένη περίπτωση, NP-δύσκολο. Ωστόσο, επειδή τα δείγματα εισόδου προέρχονται από την κατανομή Mallows, δεν βρισκόμαστε απαραίτητα στη χειρότερη περίπτωση. Για την ακρίβεια, επειδή γνωρίζουμε ότι τα δείγματα προέρχονται από την κατανομή Mallows, μπορούμε, όπως έδειξαν και οι Braverman and Mossel [2009], να υπολογίσουμε μία εκτίμηση μέγιστης πιθανοφάνειας σε χρόνο πολυωνυμικό ως προς n:

Informal theorem 1.5.1

Για κάθε $\alpha > 0$, υπάρχει ένας αλγόριθμος ο οποίος υπολογίζει με πιθανότητα τουλάχιστον $1 - n^{-\alpha}$ μία εκτίμηση μέγιστης πιθανοφάνειας, δοσμένων ανεξαρτήτων δειγμάτων της κατανομής Mallows, ο οποίος τρέχει σε χρόνο:

$$T = O(n^{1+\frac{2+\alpha}{\beta r}} 2^{O(\frac{\alpha}{\beta}+\frac{1}{\beta^2})} \log^2 n)$$

Ο αλγόριθμος που επιτυγχάνει τα παραπάνω αποτελείται από δύο φάσεις. Στην πρώτη, υπολογίζει μία εκτίμηση της κεντρικής διάταξης, η οποία ταξινομεί κάθε εναλλακτική σε μία θέση που βρίσκεται κοντά στην θέση του στην κεντρική διάταξη. Αυτό είναι δυνατόν, λόγω της συγκέντρωσης που εμφανίζουν οι τυχαίες τοποθετήσεις των στοιχείων στα δείγματα γύρω από τις αρχικές τους θέσεις. Η αρχική εκτίμηση που χρησιμοποιούμε είναι η μέση θέση εμφάνισης κάθε στοιχείου. Έπειτα, επειδή και η εκτίμηση μέγιστης πιθανοφάνειας έχει την ιδιότητα να ταξινομεί κάθε στοιχείο κοντά στην αρχική του θέση, αντί να ψάξουμε ολόκληρο τον χώρο $\mathfrak{S}_n$, ψάχνουμε μέσα σε ένα υποσύνολό του, όπου γνωρίζουμε ότι περιλαμβάνει με μεγάλη πιθανότητα κάθε εκτίμηση μέγιστης πιθανοφάνειας και μπορούμε να τον σχηματίσουμε με βάση την εκτίμηση που βρήκαμε στην πρώτη φάση. Για την τοπική αναζήτηση, χρησιμοποιούμε έναν αλγόριθμο δυναμικού προγραμματισμού, πράγμα που είναι δυνατόν λόγω και της δομής του προβλήματος βελτιστοποίησης στο οποίο αντιστοιχεί η μεγιστοποίηση της πιθανοφάνειας στο μοντέλο Mallows.

Γενικεύοντας τον παραπάνω αλγόριθμο στην περίπτωση του selective Mallows μοντέλου, λαμβάνουμε ότι:

Informal theorem 1.5.2

Για κάθε $\alpha > 0$, υπάρχει ένας αλγόριθμος ο οποίος υπολογίζει με πιθανότητα τουλάχιστον $1 - n^{-\alpha}$ μία εκτίμηση μέγιστης πιθανοφάνειας, δοσμένων ανεξαρτήτων δειγμάτων της κατανομής selective Mallows, ο οποίος τρέχει σε χρόνο:

$$T = O(n^2 + n^{1+\frac{2+\alpha}{\beta r p^4}} 2^{O(\frac{1}{\beta^2 p^4})} \log^2 n),$$

όπου $p \in (0, 1]$ είναι η παράμετρος συχνότητας.

Ο αλγόριθμος που χρησιμοποιούμε σε αυτήν την περίπτωση έχει αντίστοιχη δομή με αυτό που χρησιμοποιείται στην περίπτωση όπου τα δείγματα είναι πλήρεις διατάξεις. Συγκεκριμένα, αποτελείται από δύο φάσεις:

1. Εύρεση σημειακού εκτιμητή της κεντρικής διάταξης, δηλαδή μίας διάταξης η οποία το-

ποθετεί κάθε εναλλακτική σε μία θέση που είναι κοντά στην θέσης της στην κεντρική διάταξη.

2. Αναζήτηση γύρω από μία περιοχή του σημειακού εκτιμητή στον χώρο $\mathfrak{S}_n$.

Για την πρώτη φάση, είδαμε ότι στην περίπτωση που τα δείγματα είναι πλήρη, μπορεί να χρησιμοποιηθεί ο εκτιμητής μέσης θέσης. Ωστόσο, όταν τα δείγματα δεν είναι πλήρη, ο εκτιμητής αυτός δεν δουλεύει, επειδή η θέση όπου εμφανίζεται κάποια εναλλακτική σε μία περιορισμένη διάταξη βρίσκεται σε διαφορετικό χώρο από τη θέση της στην κεντρική διάταξη. Επομένως, κάνουμε χρήση του σημειακού εκτιμητή: Η θέση κάθε εναλλακτικής καθορίζεται από το πλήθος των άλλων εναλλακτικών που εμφανίζονται ταξινομημένες σε μικρότερες θέσεις από αυτήν στην πλειοψηφία των δειγμάτων που αμφότερες εμφανίζονται. Ο λόγος που δουλεύει αυτός ο εκτιμητής, είναι γιατί αν δύο εναλλακτικές είναι ταξινομημένες μακριά η μία από την άλλη στην (ενδεχομένως περιορισμένη) κεντρική διάταξη, τότε η πιθανότητα να εμφανιστούν με λανθασμένη σειρά είναι μικρή.

Για την δεύτερη φάση, χρησιμοποιούμε το γεγονός ότι το πρόβλημα μεγιστοποίησης της πιθανοφάνειας του εκτιμητή της κεντρικής διάταξης στο selective Mallows μοντέλο έχει παρόμοια δομή με αυτή του πλήρους μοντέλου. Έτσι, μπορούμε να χρησιμοποιήσουμε τον ίδιο αλγόριθμο δυναμικού προγραμματισμού για την τοπική αναζήτηση. Μάλιστα, μπορούμε χαλαρώνοντας τον στόχο μας, ώστε αντί για μία εκτίμηση μέγιστης πιθανοφάνειας να αρκούμαστε σε μία εκτίμηση της οποίας η πιθανοφάνεια είναι τουλάχιστον ίση με αυτήν της κεντρικής διάταξης (μιας και γνωρίζουμε ότι ο σημειακός εκτιμητής είναι σημειακά κοντά στην κεντρική διάταξη, με μεγάλη πιθανότητα), ιδέα που προτάθηκε από τους Rubinstein and Vardi [2017], να επιλύσουμε το πρόβλημα χωρίς να δείξουμε ότι η εκτίμηση μέγιστης πιθανοφάνειας είναι σημειακά κόντά στην κεντρική διάταξη, λαμβάνοντας το ακόλουθο αποτέλεσμα:

Informal theorem 1.5.3

Για κάθε $\alpha > 0$, υπάρχει ένας αλγόριθμος ο οποίος υπολογίζει με πιθανότητα τουλάχιστον $1 - n^{-\alpha}$ μία εκτίμηση με πιθανοφάνεια τουλάχιστον ίση με αυτή της κεντρικής διάταξης, δοσμένων ανεξαρτήτων δειγμάτων της κατανομής selective Mallows, ο οποίος τρέχει σε χρόνο:

$$T = O(n^2 + n^{1+\frac{2+\alpha}{\beta r p^2}} 2^{O(\frac{1}{\beta^2 p^2})} \log^2 n),$$

όπου $p \in (0, 1]$ είναι η παράμετρος συχνότητας.

Ολοκληρώνουμε την ανάλυσή μας αποδεικνύοντας ότι και η εκτίμηση μέγιστης πιθανοφάνειας είναι σημειακά κοντά στην κεντρική διάταξη (άρα και σημειακά κοντά στον σημειακό εκτιμητή), αλλά λαμβάνουμε μία επιβάρυνση της τάξης του $1/p^2$, όπου p είναι η παράμετρος συχνότητας.

1.6 Συμπεράσματα και μελλοντική δουλειά

Ορίσαμε το μοντέλο selective Mallows ως μία παρεμβολή ανάμεσα στο μοντέλο Mallows και το μοντέλο θορυβωδών συγκρίσεων. Πράγματι, σε κάθε μία από τις τρεις αυτές περιπτώσεις, η εκτίμηση μέγιστης πιθανοφάνειας για την κεντρική διάταξη έχει στην ουσία την ίδια δομή.

Αφού ορίσαμε τρεις παράλλαγές του μοντέλου: την ανταγωνιστική, την τυχαία και την προσαρμοστική, αποδείξαμε ασυμπτωτικά σφιχτά φράγματα για την ανταγωνιστική και την τυχαία περίπτωση, τα οποία υποδεικνύουν ότι όταν δεν επιλέγουμε εμείς τα σύνολα επιλογής, τότε το καλύτερο που μπορούμε να κάνουμε είναι να δούμε την είσοδο σαν ένα σύνολο από συγκρίσεις ζευγών.

Τέλος, δείξαμε ότι μπορούμε να υπολογίσουμε μία εκτίμηση μέγιστης πιθανοφάνειας για την κεντρική διάταξη υπολογιστικά αποδοτικά, όταν η παράμετρος συχνότητας δεν παίρνει πολύ μικρές τιμές.

Ωστόσο, υπάρχουν δύο κατευθύνσεις στις οποίες τα αποτελέσματά μας μπορούν να επεκταθούν. Αρχικά, είναι ανοιχτό το πρόβλημα εύρεσης ασυμπτωτικά σφιχτών φραγμάτων για τη δειγματική πολυπλοκότητα του προβλήματος εύρεσης της κεντρικής διάταξης στην προσαρμοστική περίπτωση.

Η δεύτερη κατεύθυνση αντιστοιχεί στην βελτίωση της χρονικής πολυπλοκότητας των αλγορίθμων εύρεσης εκτιμήσεων μέγιστης ή μεγιστικής πιθανοφάνειας, ακόμη και στην περίπτωση που η παράμετρος συχνότητας είναι μικρή. Η ιδέα είναι ότι όταν η παράμετρος συχνότητας μικραίνει, τότε και το πρόβλημα που έχουμε να επιλύσουμε χαλαρώνει, αφού η εκτίμηση μέγιστης πιθανοφάνειας γίνεται λιγότερο ακριβής. Στην ακραία περίπτωση που μία εναλλακτική δεν εμφανίζεται ποτέ στα δείγματα, για παράδειγμα, μπορούμε να την τοποθετήσουμε σε οποιαδήποτε θέση σε μία εκτίμηση μέγιστης πιθανοφάνειας.

Chapter 2

Introduction

2.1 Problem statement and motivation

Ranking distributions have been thoroughly studied during the last years. They are used to model many different problems like preference aggregation and voting. Hence, they are naturally linked to social choice theory, although they go beyond it. Social choice theory (Brandt et al. [2016]) aims to establish aggregation rules that satisfy some specific sets of axioms that they ideally should have and which can be computed efficiently. The preferences are thought of as rankings over a set of possible alternatives. However, due to some impossibility results, prevalent in which is Arrow's impossibility theorem (Arrow [1951]), that state that no aggregation rule can satisfy at the same time some specific sets of axioms, viewing the problem from a different perspective became motivated. The preferences are viewed as samples from some ranking distribution and the problem of preference aggregation was reduced to learning the parameters of the ranking distribution with which the problem was modelled.

One of the most widely studied ranking distributions is Mallows distribution, which was introduced by Mallows [1957]. Under Mallows model, there is the notion of central ranking, which is the mode of the ranking distribution and the probability of sampling each other possible permutation, diminishes exponentially to the distance of the permutation of the central ranking. This is the reason why Mallows distributions are called Distance based ranking distributions. Several metrics between rankings have been proposed. However, one of the most useful ones, is the Kendall tau distance, which counts the number of inversions between two permutations. That is, the number of discordant pairs of alternatives in the two permutations. Mallows model is also linked to Kemeny's rule for ranking aggregation (Kemeny [1959]), which, in the context of social choice, exhibits some notion of optimality, since it corresponds to finding a ranking for which the total number of pairwise disagreements with the input rankings is minimized. However, computing Kemeny's ranking is shown to be an NP-hard problem in the worst case. Interestingly, under Mallows model, the Kemeny's rule, which in this case corresponds to finding the maximum likelihood estimation of a profile of independent Mallows samples, can be computed efficiently, as shown by Braverman and Mossel [2009], since Mallows model corresponds to an average case of the Kemeny's aggregation problem, where it is, with high probability, easy.

However, assuming that the input samples are complete rankings is not always realistic. Especially in cases when the number of alternatives is very large, assuming that we have

access in complete samples is overly optimistic. Nevertheless, our goal is to estimate the complete central ranking, using incomplete sample rankings. For that reason, we propose the Selective Mallows model which generalizes the Mallows model, by enabling the sampling of incomplete permutations. Again, there exists a central ranking, but its meaning is slightly different: After selecting the set of alternatives to be ranked in a sample, the central ranking is restricted to the selected set and the sample is drawn from a Mallows distribution with the restricted central ranking as the distribution's mode. A Selective Mallows sample profile is, in general considered independent, conditioned on the selection sets of each sample it contains. Different selection procedures correspond to different settings. For example, the sets might be selected by an adversary, randomly or adaptively, if we are given access to a Selective Mallows sampler that inputs a set of alternatives and outputs a sample of the corresponding Mallows distribution.

Selective Mallows model takes into account what we call *ignorance bias*. That is, the probability of swap of two alternatives in a selective Mallows sample diminishes exponentially to the number of alternatives that are ranked between them in the central ranking **and** are included in the corresponding selection set. This applies to cases where the alternatives do not have an individual value, or their value is completely unknown, and their ranking is produced by comparing one with another. Clearly, if one does not know an alternative, they cannot compare it with another. For ignorance biased sampling agents, the unknown alternatives' positions in the (complete) central ranking do not influence the probability of swap of any pair of selected alternatives.

Furthermore, the problem of computing the maximum likelihood estimation of the central ranking under the selective Mallows model, conditioned on the selection sets, is shown to be a generalization of Kemeny's aggregation problem. Therefore, the Selective Mallows model can be considered as a natural generalization of the Mallows model, since the maximum likelihood estimation problems in each case have virtually the same structure.

Also, it is interesting to point out that Selective Mallows model is minimal, not only because of its simplicity, but also because the information that each sample provides is minimal, in the sense that the position of the alternatives that are not selected in the central ranking does not influence the probability of sampling a reduced permutation according to selective Mallows model. In contrast, under a model where the samples are produced by initially sampling a complete ranking and then projecting it into a reduced subset of alternatives, the positions of the alternatives that do not appear in the sample, generally, do influence the probability of observing a specific reduced ranking.

In this thesis, we consider the problem of retrieving the central ranking from selective Mallows samples and also reconstructing a maximum likelihood estimation of the central ranking, for any number of input samples. In other words, we study the statistical and computational complexity of learning the central ranking under the selective Mallows model. We establish tight sample complexity bounds in the adversarial and random settings and we generalize the algorithm proposed by Braverman and Mossel [2009] in order to efficiently compute the maximum likelihood estimation of the central ranking given selective Mallows samples. We leave open the problem of establishing tight sample complexity bounds for retrieving the central ranking under the adversarial model.

2.2 Related work

In statistical analysis of ranking data, the first milestone was the introduction of parametric models such as Mallows [1957], Plackett [1975] and Fligner and Verducci [1986]. Those models have been widely studied and many generalizations have been proposed. Another line of research is nonparametric approaches like Lebanon and Mao [2008]. We are interested in the problems of modeling and inference on models for incomplete rankings. This direction has several branches.

First, in works like Huang et al. [2012] and Kakarala [2012], the problem of aggregating partial rankings is considered, which corresponds to the case when the input consists of partial rankings, namely partial relations on the alternatives space.

Another branch corresponds to pairwise queries as input. For instance, in the work of Braverman and Mossel [2007] the Noisy Comparisons model is considered and an efficient algorithm for computing the maximum likelihood estimation of the underlying ranking is proposed, while in Feige et al. [1994], the query complexity of retrieving the underlying ranking under the same model is settled, in the context of providing parallel algorithms for solving various problems using noisy comparisons between pairs of alternatives. Other relevant work in this direction includes Wauthier et al. [2013] and Busa-Fekete et al. [2014].

Incomplete rankings can also be viewed as censored data (Lebanon and Mao [2008]). Taking into consideration the process that projects complete rankings into incomplete ones, projective models emerge, for example in the work of Fahandar et al. [2017]. Such approaches are linked to notions like coarse data (Heitjan and Rubin [1991], Gill et al. [1997]), which in the context of statistics describes data that consist of units that correspond to sets of extensions. For example, an incomplete ranking corresponds to the set of complete rankings that order the alternatives that appear in the incomplete ranking in the same order. A particular line of work regards modeling and inference from top-k lists. That is, the input consists of incomplete rankings that correspond to the highest ranked alternatives. Relevant results can be found in Busse et al. [2007], Meila and Bao [2010], Meila and Chen [2012], Tang [2018] and Chierichetti et al. [2018].

Instead of projecting rankings into incomplete rankings (based on the positions) it could be assumed that the incomplete rankings are created by projections in the alternatives space. For example we refer to Rajkumar and Agarwal [2014] and Sibony et al. [2015].

In Lu and Boutilier [2011], a generalized sampling method that can describe arbitrary ranking distributions is proposed, and a method for inferring in Mallows model from incomplete rankings is provided.

Our selective Mallows model is a different formulation of the noisy choice model introduced in Procaccia et al. [2012], which interpolates between Mallows and Noisy comparisons model. While the two models have similar structures, their formulation serves different purposes. In Procaccia et al. [2012], several interesting questions regarding inference from incomplete rankings are addressed, while we address the problem of retrieving the complete central ranking, under assumptions that correspond to the structure of the selection sets.

On the technical part, our work is mostly related to Caragiannis et al. [2013], Feige et al. [1994] and Braverman and Mossel [2009]. In the works of Feige et al. [1994] and Caragiannis et al. [2013], the sample complexity of retrieving the hidden ranking under Noisy comparisons and Mallows model, respectively, is settled. In the work of Braverman and Mossel [2009], an algorithm for computing the maximum likelihood estimation of the cen-

tral ranking under the Mallows model that runs in polynomial time with respect to the number of alternatives is proposed. The algorithm is generalized in Rubinstein and Vardi [2017], where a relaxed solution concept (likelier than nature estimation) is introduced.

2.3 Related results and our contribution

We examine the problem from two different aspects. First, we propose three different settings and acquire tight sample complexity bounds for retrieving the central ranking in two of them. Second, we extend the algorithm presented by Braverman and Mossel [2009] to solve the maximum likelihood estimation of the central ranking problem from incomplete samples in the case that each pair of alternatives appears frequently in the samples.

Learning a hidden ranking. In the complete Mallows case, according to the work of Caragiannis et al. [2013], the central ranking can be retrieved using $\text{poly}(1/\beta)\Theta(\log(n/\epsilon))$ samples, where β is the spread parameter, n is number of alternatives and ϵ is the accepted margin in the probability of error. In general, the important dependence is on the number of alternatives, since we concentrate on cases where the spread parameter does not take significantly small values. The bound is achieved by using a pairwise majority estimator, namely creating a ranking where each pair is ordered according to the majority order of the pair's comparisons in the samples. This works, because the probability of swap of any two alternatives is upper bounded by $e^{-\beta}/(1-e^{-\beta})$ and hence the result follows by applying the union and Hoeffding bounds on the probability of error. Also, the bound is tight because there are $n-1$ pairs that have probability of swap equal to $e^{-\beta}/(1-e^{-\beta})$ and enough samples are needed in order to be sure about each one of them.

On the other hand, according to Feige et al. [1994], in the Noisy comparisons model, namely when we have access to noisy pairwise queries (each query may be wrong with some probability, say $e^{-\beta}/(1-e^{-\beta})(< 1/2)$), the number of samples needed to retrieve the hidden ranking (when we pick the queries during the runtime - noisy sorting) is $\text{poly}(1/\beta)\Theta(n\log(n/\epsilon))$. Interestingly, the sample (query) complexity's dependence on n is the same as in the case of noiseless sorting. Observe that in the complete Mallows case, there is a logarithmic blow up: In the noiseless setting, a single complete ranking is sufficient. The reason that the logarithmic blow up can be prevented in the noisy comparisons case is that noisy binary search can be implemented in a way that enables error correction: It can be viewed as a random walk on the binary search tree, where each time the search reaches a node, we (noisy) check whether the searched element falls into the corresponding (to the node) interval (see figure 2.1). The algorithm proposed by Feige et al. [1994] consists of three parts, the first and last of which are designed in order to exploit the central part (noisy binary search), without having excessive query cost.

Our constribution: We define three different settings, corresponding to ways of picking r selection sets:

- *Adversarially* ($\mathcal{M}^{\text{ADV}}_{\pi_0,\beta}$): Picked by an adversary. In this case, we consider a frequency parameter $p \in (0,1]$: Each pair of alternatives appears in at least pr sets. In this setting, the sample complexity's dependence on p is crucial, since it might take very small values. We prove that the sample complexity for retrieving the central ranking is $\text{poly}(1/\beta)\Theta(\frac{1}{p}\log(n/\epsilon))$. We use pairwise majority estimators similarly to the

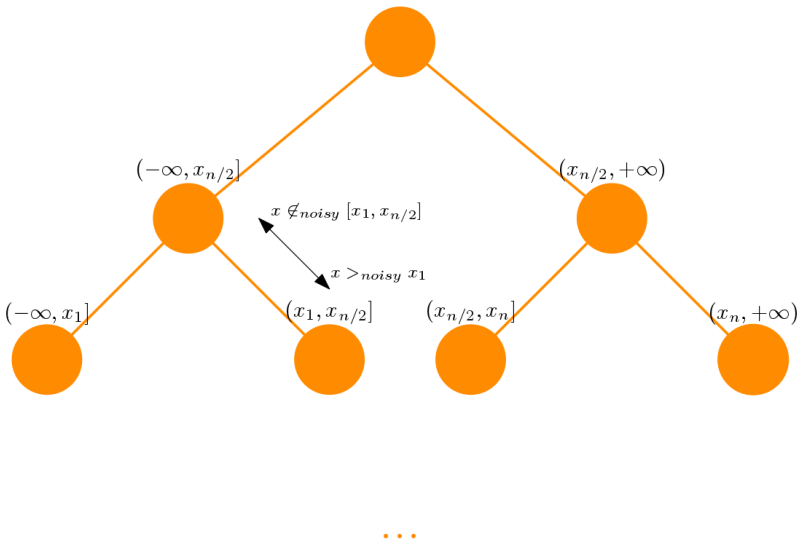

Figure 2.1: Noisy binary search

complete Mallows case to get the upper bound. The lower bound is a consequence of the fact that if we are given a fixed selection sets vector with $o((n^2/\beta)\log(n/\epsilon))$ pairwise comparisons, then we cannot guarantee to find the central ranking with probability of error less than ϵ. Then, we just show that one can create a selection sets vector that is $p-$frequent, but has no more than $2prn^2$ comparisons. The reason why the number of pairwise comparisons must be $\Omega((n^2/\beta)\log(n/\epsilon))$ is illustrated by figure 2.2: Each row contains $n/2$ disjoint pairs for which any algorithm needs sufficient information to rank them ($\Omega(n\log(n/\epsilon))$ queries) and there are $n/2$ rows.

- *Randomly* ($\mathcal{M}^{\text{RND}(\mathcal{D})}_{\pi_0,\beta}$): Picked according to some selective distribution $\mathcal{D}$, independently. We consider a frequency parameter $p \in (0, 1]$:

$$\mathcal{D}(i \text{ and } j \text{ selected}) \geq p$$

In this case, we get similar bounds to the adversarial case: $\text{poly}(1/\beta)\Theta(\frac{1}{p}\log(n/\epsilon))$. The upper bound is established using the same method, additionally applying the law of total probability on the number of appearances of a pair. As for the lower bound, the idea is that if $n/2$ pairs are selected independently with probability p in a sample (according to $\mathcal{D}$), then it is likely that some of these pairs never appears. If we pick a central ranking with those pairs adjacent. If an adjacent pair does not appear we rank it at random at best. The outline of the proof follows:

  - Pick any row of figure 2.2. It corresponds to a set R of $2^{n/2}$ possible rankings. (Randomness has negated "vertical attacks" (according to figure 2.2)).
  - Pick a central ranking uniformly at random from R.
  - Then, observe that if the samples do not contain some of the pairs of the row we picked: each such pair can be swapped in the central ranking without changing the probability of observing the samples.
  - With a careful handling, we get a bound for the expected success probability of any algorithm (over uniform distribution in R).
  - *Probabilistic method:* There exists a central ranking satisfying the bound.

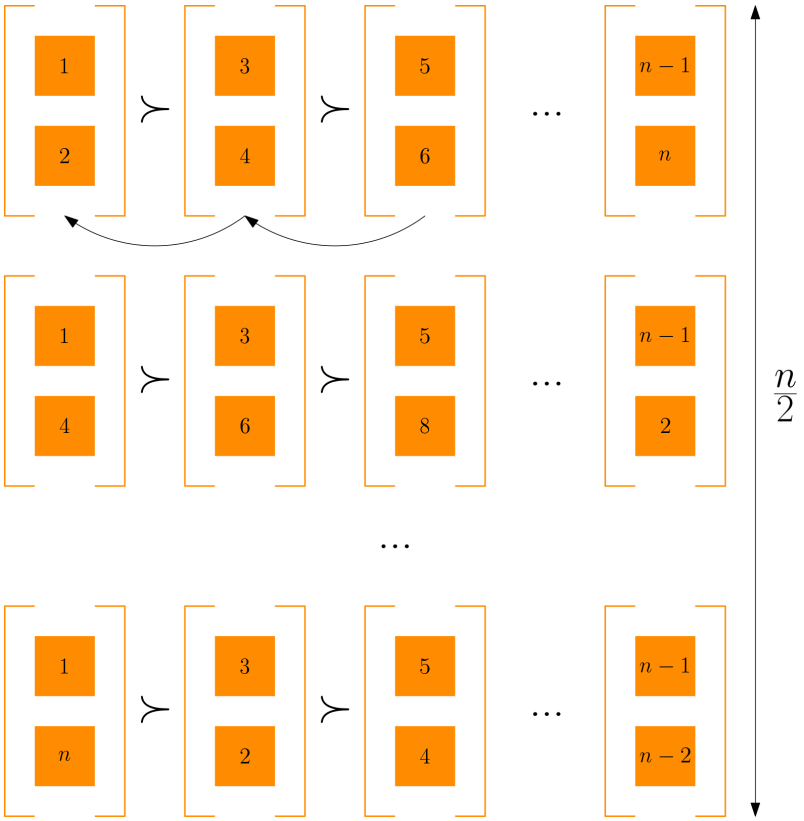

Figure 2.2: Possible selections for a "difficult" central ranking

- *Adaptively* ($\mathcal{M}^{\mathrm{ADP}(m)}_{\pi_0,\beta}$): Picked in the runtime of an algorithm having access to a selective Mallows sampler. In this case, we define a capacity parameter $m \leq n$: $|S| \leq m, \forall S$ picked. A trivial consequence of the algorithm for noisy sorting proposed by Feige et al. [1994], is that the sample complexity in the adversarial model is $\mathrm{poly}(1/\beta)O(\frac{n}{m}\log(n/\epsilon))$. However, finding tight sample complexity bounds for this model remains open.

Mallows reconstruction problems. Assume we are given r iid Mallows samples. The problem of finding the maximum likelihood estimation of the central ranking from these samples coincides with Kemeny's aggregation problem, which is known to be NP-hard, even if $r = 4$. However, as shown by Braverman and Mossel [2009], there exists an algorithm that outputs a maximum likelihood estimation with probability at least $1 - n^{-\alpha}$, for any $\alpha > 0$ that runs in polynomial time to n. This is done by exploiting the structure of the average case to which Mallows reconstruction problem coincides: the input rankings are samples from a Mallows distribution. The algorithm consists of two main parts. In the first part, an estimation of the central ranking that with high probability ranks each alternative close to its location in the central ranking is computed. This is possible since the alternatives' positions are concentrated exponentially around their initial positions, under Mallows model: the average position is used. In the second part, since it happens that the maximum likelihood estimation of the central ranking has the same property regarding the pointwise proximity (every alternative is placed close to its initial position) to the central ranking, the search space is reduced to a subspace of $\mathfrak{S}_n$ which includes the rankings that place each alternative close to its position in the estimator calculated in the

first part. Additionally, restricted in this subspace, the structure of the problem enables one more manipulation: use of dynamic programming to move from the initial estimation to a maximum likelihood one.

Our contribution: We generalize the algorithm described above to solve the maximum likelihood estimation problem from incomplete samples. First, we need to establish a new initial estimator, since the average estimation does not work in the case of incomplete samples, since the positions space of central ranking and that of a sample are different. Instead, we use the positional estimator, which ranks an alternative according to the number of other alternatives which are ranked before it in the majority of samples. We prove that the positional estimator is pointwise close to the central ranking when the samples are $p-$frequent for any $p \in (0, 1]$, for which: $1/p \ll \log n$. Our proof is based on a notion of neighborhood defined for each alternative $i \in [n]$ that includes all other alternatives that are ranked close to i in many samples. The neighborhood is defined by two parameters: the first specifies when two positions are considered close, while the second one specifies the threshold of close appearances in order to include an alternative in the neighborhood of i. On the one hand, the neighborhood's length is bounded and no the other hand the alternatives not appearing in the neighborhood of an element i are in many samples most likely ranked correctly with respect to i. Therefore, with a careful handling we prove that the positional estimator ranks each alternative close to its position in the central ranking.

2.4 Organization

We organize this thesis into this introductory chapter, four main chapters (3, 4, 5, 6) and a conclusions chapter (7).

Chapter 3: Theoretical Background. In this chapter, we introduce the theoretical framework of our work. We begin by presenting elements of probability theory and some useful tools that it provides. Probability theory is deeply connected with computer science, since it provides the tools to manage and analyze randomness.

In the second part of the chapter we present the fundamental elements of computational learning theory. Learning theory is a very active field of research that is relatively young, having been formally introduced from a mathematical perspective by Valiant [1984].

Chapter 4: Probability and Permutations. In this chapter we introduce the notation that is used throughout the rest of the work and after mentioning some ranking distributions, focusing in particular on Mallows distribution, as well as several relevant results of the existing literature, we present, in section 4.2.3, the model we propose, namely the Selective Mallows model.

Chapter 5: Learning a Hidden Ranking. This chapter concerns the problem of retrieving the central ranking from complete or incomplete rankings. After presenting the extant results on statistical complexity of retrieving the central ranking from complete noisy rankings, under Mallows model (Caragiannis et al. [2013]), as well as from noisy pairwise comparisons (Feige et al. [1994]), we establish, in section 5.3, our own tight sample complexity bounds for the adversarial and random settings of selective Mallows model.

Chapter 6: Mallows reconstruction problems. In this chapter, we encounter the problem of finding maximum (or maximal) likelihood estimations for the central ranking. We initially present the work of Braverman and Mossel [2009], in which a polynomial time algorithm for solving the problem with high probability under Mallows model is introduced. In section 6.3, building on the work just presented, we acquire a generalized efficient algorithm for the selective Mallows case.

Chapter 7: Conclusions and further work. In the final chapter we summarize our results and identify some of the relevant problems that are left open.

Chapter 3

Theoretical Background

In this chapter, we present the fundamental concepts that constitute the basis of the work we present in the following chapters. First, we discuss about probability theory and its general connections to computer science, meanwhile presenting some of the tools we will use throughout the rest of this work. Consequently, we describe elements of computational learning theory and its history, which constitutes the framework of our work. As will become clear, probability theory and computational learning are closely related, since computational learning's development requires the tools that probability theory provides.

3.1 Probability theory

Randomness is a concept that has drawn the interest of humanity from the beginning of its history. However, a solid theory providing the necessary tools to explore it in a consistent way has been established relatively recently. As presented by Kolmogorov [1950], Probability theory, although undoubtedly possessing individual interest, appears to be a branch of the more general field of Measure theory (Halmos [2013]), in whose development, prevalent were the contributions of Lebesgue [1918] and Borel [1919], among others.

In this section, after providing some context on measure theory, we focus on describing some of the strong tools that probability theory offers.

3.1.1 Measure theory

Measure theory emerged as a result from the endeavor of mathematicians to generalize the Riemann integral (Riemann [1868]) in order to be able to integrate more complex functions and, generally, acquire a deeper understanding on the limitations of integrability. In a nutshell, it is known that each function that does not have too many incontinuities is Riemann integrable. However, the class of functions that are Lebesgue integrable is strictly wider that those that are Riemann integrable.

The answer to the question of how to increase the width of integrability lied to reversing the calculation method: the integral was no longer perceived as a sum of the products of the lengths (area or volume) of infinitesimal intervals with the corresponding approximations of the value of the function, but as a sum of the products of the values y of the function with the total length (area or volume) of the subset of the domain that corresponds to

points where the function takes values close (defined via some partition of the value axis) to y. The problem reduces to generalizing the concept of length (area or volume) to sets that are as complex as possible, while the notion of length keeps some of its basic intuitive properties and also, the length (area or volume) of intervals remains unchanged. This is, in fact, the core of difficulty of integration and the generalized length (area or volume) corresponds to what is known as Lebesgue measure.

What turns out to be really surprising, is that the family of the subsets of $\mathbb{R}$ that are Lebesgue measurable is so wide that in order to prove that there exists a subset of $\mathbb{R}$ that is not Lebesgue measurable, one has to use the axiom of choice (for example we refer to Jech [2013]). This means that one cannot construct a set that is not Lebesgue measurable.

To measure is to project subsets of a set to non negative values (possibly infinite), such that adding the measures of countable families of independent subsets is equivalent to measuring their countable union, while the measure of an empty subset cannot be positive. Length, area and volume are measures. In fact, the way these concepts are intuitively perceived led to the development of measure theory, in order to examine them in an abstract way. However, probability is also a measure, with the additional property that the values it assigns to sets are finite and normalized. It measures the likelihood of subsets of a set of possible outcomes, which, in some sense, corresponds to the attempt to predict the unknown. Inheriting all the properties of measure theory, as well as developing some of its own, it serves to provide tools for analyzing and predicting the behavior of extremely complex or inherently random systems or even designing such systems in order to solve problems.

3.1.2 Probabilistic tools

In Section 3.1.1, we presented the foundations of probability theory in the context of measure theory. Indeed, measure theory is the right tool to study probability theory from a generic point of view. However, it is often useful to concentrate on specific fields of probability theory, without the formality of measure theory. In this section, we will present some useful probabilistic tools as well as some of their applications.

Computer science and randomness are closely related. For one thing, randomness is frequently the most efficient way to model or analyze a problem, but also because designing algorithms that use randomness often results into simple and elegant solutions.

In general, randomness has the property to negate adversarial input choices. For example, the algorithm of Quicksort, developed by Hoare [1961], uses randomness to eliminate the possibility of adversarial selection of the initial permutation of the elements to be sorted. Another application where this property can be taken advantage of is dimension reduction, which is the problem of reducing the dimensionality of a set of points in $\mathbb{R}^d, d \in \mathbb{N}$ to $k < d$, so that the distance of any two points is not affected significantly by the projection. In particular, the random projection algorithm of Johnson and Lindenstrauss [1984], which is virtually as simple as picking a random $k-$dimensional hyperplane and providing the projections of the points on that plane to the output, is, remarkably, competitive to preexisting sophisticated techniques, like Principal Component Analysis (Pearson [1901] and Hotelling [1933]). Intuitively, what randomness offers in this case is, similarly to the example of Quicksort, the debilitation of the adversary (who in this case picks the points), by making it impossible for her to pick a direction along which the elements are stretched (and therefore this direction's contribution to the distance of two points is significant) and

the random projection is probable to ignore this direction's contribution, because picking a random hyperplane is the same as making a random rotation. Furthermore, the computation of the random embedding can become more efficient by calculating a projection matrix with the use of binary coins, which was proposed by Achlioptas [2003].

Another example of the power of randomness is its application to solving the primality testing problem. Algorithms like the Miller-Rabin test (Miller [1976] and Rabin [1980]) are probabilistic algorithms that use Fermat primality test (we refer to Thomas H. Cormen [2001]) in order to determine with high probability whether a (large) number is prime. The high level intuition behind this algorithm is that when repeating independent random trials (coin flips or Bernoulli trials) that have a non negligible probability of success, the first successful trial will most probably not appear with much delay.

Furthermore, randomness appears to be useful for solving mathematical problems or analyzing deterministic algorithms. A method that is frequently used to solve combinatorial problems is the Probabilistic Method, which was initially introduced by Erdös [1947].

For all these reasons, it is useful to present some of the tools that probability theory provides. For further reading, we refer to Mitzenmacher and Upfal [2017], Alon and Spencer [2004] and Motwani and Raghavan [2010]. In this work, we focus on presenting some basic notions before deriving some methods to prove that sometimes random variables behave in a highly foreseeable way (also known as concentration inequalities) and conclude by presenting some elements of the Probabilistic Method.

Moments and moment methods. For the following, let X be a random variable that is either discrete or (absolutely) continuous. If X is discrete, we will denote with p its probability mass function and with $\mathcal{S}$ its support, while if it is continuous, we will denote with f its probability density function. The mathematical expectation of a random variable will be denoted with:

$$\mathbb{E}[X] = \begin{cases} \sum_{x\in\mathcal{S}} x \cdot p(x), & \text{if } X \text{ is discrete,} \\ \int_{\mathbb{R}} x \cdot f(x)\, dx, & \text{if } X \text{ is continuous.} \end{cases}$$

For any function g, it holds that:

$$\mathbb{E}[g(X)] = \begin{cases} \sum_{x\in\mathcal{S}} g(x)p(x), & \text{if } X \text{ is discrete,} \\ \int_{x\in\mathbb{R}} g(x)f(x)\, dx, & \text{if } X \text{ is continuous.} \end{cases}$$

In the specific case when $g(x) = x^r$ for some $r \in \mathbb{N}$, $\mathbb{E}[g(X)]$ is called the $r-$th moment of X. To acquire an intuition on why the moments of a random variable are important, we use the following example:

Pilot example 1. Let X be a discrete random variable and let E be a set of events that determine X as follows: $X = \sum_{e\in E} \mathbb{1}\{e \text{ happened}\}$. (Observe that this representation is in fact very general.) Then, $\mathbb{E}[X] = \sum_{e\in E} \Pr[e]$ and for the second moment we have:

$$\mathbb{E}[X^2] = \sum_{e,e'\in E} \Pr[\{e \text{ happened}\} \cap \{e' \text{ happened}\}] = \sum_{e\in E} \Pr[e] + \sum_{e\neq e'} \Pr[e'|e]\Pr[e]$$

Observe that while $\mathbb{E}[X]$ only depends on $\mathbb{P}[e], e \in E$, the second moment requires knowledge about the probability that an event $e' \in E$ happens, conditioned on the event $e \in E \setminus \{e'\}$. Therefore, second moment includes more detail about X.

We could say that second moment measures the inward pairwise correlations of the random variable X. In that sense, the importance of variance ($\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$) can be better understood: It is a comparison between a quantity that includes the inward pairwise correlations of X ($\mathbb{E}[X^2]$) and a quantity that is, conceptually, the closest to the first one, but which is conditioned to ignore inward correlations ($(\mathbb{E}[X])^2$). Hence, the quantity $\mathbb{E}[X^2]/(\mathbb{E}[X])^2$ would be an equally useful way to describe a concept like variance.

In the specific case when the number equivalence classes of $E \times E$ that correspond to the equivalence relation: $R = \{((e_1, e_1'), (e_2, e_2')) : \Pr[e_1'|e_1] = \Pr[e_2'|e_2]\}$ is small and the corresponding value $\Pr[e'|e]$ of each class is available in a closed form, then $\mathbb{E}[X^2]$ can be calculated analytically, which enables the use of the second moment method.

Similar, generalized arguments can be made for the moments of order higher than 2.

As illustrated by Example 1, the moments of a random variable contain information about it. Therefore, we have managed to partially express a random variable with numerical values that correspond to its distribution. These values can be used to quantify some properties of the random variables and, most practically, their concentration.

A basic inequality that involves the first moment of a non negative random variable, is Markov's inequality, which is easily derived from the following observation:

$$\mathbb{1}\{X \geq a\} \leq X/a,$$

where X is a non negative random variable and $a > 0$. Taking the expectation of both sides, with respect to X, we get the following theorem:

Theorem 3.1.1: Markov's Inequality

For any random variable $X \geq 0$ and any $a > 0$:

$$Pr[X \geq a] \leq \mathbb{E}[X]/a$$

As a corollary: $\Pr[X \geq a\mathbb{E}[X]] \leq 1/a$. Indeed, when the only known parameter of the distribution of X is its expectation, then Markov's inequality is the best bound we can hope for. But, influenced by the discussion up to this point, it is natural to conjecture that knowing the variance of X would yield a better bound. In fact, this bound can be easily acquired by applying Markov's inequality to the random variable $Y = (X - \mathbb{E}[X])^2$. This time, X need not be non negative and we get the following theorem:

Theorem 3.1.2: Chebyshev's Inequality

For any random variable X and any $a > 0$:

$$Pr[|X - \mathbb{E}[X]| \geq a] \leq Var(X)/a^2$$

As a corollary: $\Pr[|X - \mathbb{E}[X]| \geq a\mathbb{E}[X]] \leq \frac{1}{a^2}\frac{\text{Var}(X)}{(\mathbb{E}[X])^2}$. Using these bounds, we are often able to obtain interesting results. This process is usually referred to as the moment method.

Application 1. We consider a $\mathcal{G}_{n,p}$ graph G, which is a random graph with n vertices that includes each possible edge independently with probability p (we toss a coin with probability of heads p for every pair of vertices). We are interested in deciding whether we should bet all of our money (real estate included) on one of the two following statements:

1. G contains a 4−clique.
2. G does not contain a 4−clique.

One might think that since G is produced randomly, we should be reluctant in accepting the bet. However, using the second moment method, we can conclude that in some cases, i.e. when $p \ll n^{-2/3}$ or $p \gg n^{-2/3}$ and n is very large, we should accept the bet with no second thoughts.

Let $X = X(G)$ be a random variable that equals to the number of 4−cliques contained in G. Also let E be the set of events $e(v_1, v_2, v_3, v_4) : G[\{v_1, v_2, v_3, v_4\}]$ is isomorphic to K_4 for all $(v_1, v_2, v_3, v_4) \in (V(G))^4$, where $G[S]$ is the subgraph of G induced by $S \subseteq V$ and K_4 is the 4−clique. Then:

$$X = \sum_{e \in E} \mathbb{1}\{e \text{ happened}\}$$

Therefore, similarly similarly to Example 1, we get:

$$\mathbb{E}[X] = \sum_{e \in E} \Pr[e] = \sum_{e \in E} p^6 = \binom{n}{4} p^6 = \Theta(n^4 p^6)$$

- If $p \ll n^{-2/3}$, then as $n \to \infty$, $\mathbb{E}[X]$ tends to become zero. Hence, from Markov's Inequality: $\Pr[X \geq 1] \to 0$. In this case, we should bet our money on statement 2. - If $p \gg n^{-2/3}$, then $\mathbb{E}[X] \to +\infty$. But this does not give us any guarantee. The intuitive reason why this happens is due to lottery effect: the expectation of a random variable can be arbitrarily high, if the probability of it being zero is not exactly equal to 1. We will have to use the second moment method.

We omit details and claim that: Using the method described in Example 1, we can show that:

$$\mathrm{Var}(X) = O(n^4 p^6) + O(n^6 p^{11}) + O(n^5 p^9)$$

Therefore, since $\mathbb{E}[X] = \Theta(n^4 p^6)$, we have: $\mathrm{Var}(X)/(\mathbb{E}[X])^2 \to 0$ as $n \to +\infty$. Applying Chebyshev's inequality: $\Pr[|X - \mathbb{E}[X]| \geq \mathbb{E}[X]] \to 0$, which gives that $\Pr[X = 0] \to 0$. Therefore, in this case, we should bet on choice 1.

- In case $p \approx n^{-2/3}$, we should decline the bet, because we do not have a guarantee.

Chernoff-Hoeffding bounds. We saw that the higher the order of a moment, the more the information it provides for the random variable, which leads to stronger bounds. However, while the order increases, the calculation of the corresponding moment becomes more difficult, because we have to take into consideration the dependencies between longer sequences of events. However, when $X = \sum_{e \in E} \mathbb{1}\{e \text{ happened}\}$ and E consists of independent events, then it should not be very difficult to increase the order of the moment that we examine arbitrarily. We also know that the function $g(x) = e^x$ contains, in some sense, the monomials of any order, due to its Taylor expansion: $e^x = 1 + x + x^2/2! + x^3/3! + \ldots$. Therefore, when E's events are independent, we are driven to compare, instead of $\mathbb{E}[X^r]$ and $(\mathbb{E}[X])^r$ for some $r \in \mathbb{N}$, the exponential moment $\mathbb{E}[e^X]$ and $e^{\mathbb{E}[X]}$. This leads us to the Chernoff-Hoeffding bounds, initially presented by Chernoff [1952] and brought to their general form by Hoeffding [1963]. However, there is a less strict property than independence that can be used instead of it to obtain similar bounds. This property is refered to as Martingale sequences' property and the corresponding bound (Azuma's Inequality) is attributed to Azuma [1967] and Hoeffding [1963].

Another way to view Chernoff-Hoeffding bounds is, among others, as a tightened version of Markov's Inequality, using knowledge of the structure of X: $X = \sum_{i\in[n]} X_i$, where $(X_i)_{i\in[n]}$ are independent. Then, from Markov's inequality, for any $\lambda > 0$, we have:

$$\Pr[X > x] = \Pr[e^{\lambda X} > e^{\lambda x}] \leq e^{-\lambda x}\mathbb{E}\left[\prod_{i\in[n]} e^{\lambda X_i}\right]$$

Assuming independence between X_i:

$$\Pr[X > x] \leq e^{-\lambda x} \prod_{i\in[n]} \mathbb{E}[e^{\lambda X_i}]$$

Then λ is selected so that the right-hand side of the equation is minimized. The result is stated in the following theorem:

Theorem 3.1.3: Chernoff-Hoeffding Bound

Let $X_1, X_2, \ldots, X_n$, where $n \in \mathbb{N}$ be independent Bernoulli random variables, $X = \sum_{i\in[n]} X_i$ and $\mu = \mathbb{E}[X]$. Then:

$$Pr[X \geq \mu + a] \leq e^{-nD_{KL}(\frac{\mu+a}{n}||\frac{\mu}{n})}, \text{ where } 0 < a < n - \mu$$

and:

$$Pr[X \leq \mu - a] \leq e^{-nD_{KL}(1-\frac{\mu-a}{n}||1-\frac{\mu}{n})}, \text{ where } 0 < a < \mu,$$

where $D_{KL}(p||q) = p\log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q}), \forall p, q \in (0,1)$.

Remark 3.1.1: *There are several variations of the Chernoff-Hoeffding bound. For example, bounds can be derived from the relative distance of X and its expectation.*

The Chernoff-Hoeffding bound implies exponential concentration. That is, although the sum of independent random variables is itself random, its behavior is more or less predictable for sufficiently large number of summands.

Probabilistic Method. Finally, we present a method that is useful for proving the existence of objects within a class that satisfy some properties, when explicitly constructing them is complicated. The main idea is to define a class of objects and select one of its elements at random. Given that the probability that the random object satisfies the desired property can be proven to be strictly positive, then one such object must exist within the class. Interestingly, there exist methods of derandomization (e.g. the method of conditional expectations) that, in some cases, find an element with the desired property deterministically.

Probabilistic method should already be familiar to the reader, since it has been in fact used already in 1. However, it is important to present it separately as a general method which makes use of various techniques, second moment method included.

The goal of probabilistic method is clear: prove that an object exists. The most obvious approach is to make use of counting arguments. That is, define a class of objects, measure its size and compare it with the probability that a random element has the desired property. Another technique is derived from the following theorem:

Theorem 3.1.4

Let X be a random variable. Then:

$$Pr[X \geq \mathbb{E}[X]] > 0 \text{ and } Pr[X \leq \mathbb{E}[X]] > 0$$

Moment methods can also be used in order to apply probabilistic method. However, a technique of particular interest is the one introduced by ErdHos and Lovász [1975], commonly refered to as the Lovász Local lemma. The high level intuition of this result is that if a family of "bad" events cannot cooperate (through dependencies) and also each event is not very probable, then the probability that none of them happens should be significant.

Theorem 3.1.5: Lovász Local lemma (symmetric version)

Let $E = e_1, e_2, \ldots, e_n$ be a family of events such that:

1. $\mathbb{P}[e] \leq p, \forall e \in E$ *for some* $p \in (0, 1)$.
2. *For any $e \in E$ there exist a subset I of E of size at least $n - d - 1$, for some $d \leq n$, such that:*

$$Pr[e | \cap_{e' \in I} e'] = Pr[e] \text{ (mutual independence).}$$

Then, if $4pd \leq 1$, it holds that: $\mathbb{P}[\cap_{e \in E}\{e \text{ did not happen}\}] > 0$.

3.2 Computational Learning Theory

Learning is the process of transforming experience into expertise. Its study has inspired scientists from various fields like psychology (for example, we refer to Gross [2015], chapter 11), biology and medicine. However, in the recent years, learning is also examined from a formal, mathematical perspective, which enabled its emergence as a branch of computer science. Formalization of learning had been thought as an impossible task, but, surprisingly, a formal framework was eventually proposed by Valiant [1984]: the Probably Approximately Correct Learning model (PAC learning). It was one of the cases when the important part of a mathematical study is the definition rather than the theorems or their proofs: The PAC learning model enabled the development of computational learning theory which served as a way to use a machine not only as a tool to apply human expertise in order to solve problems efficiently, but also as a producer of expertise. There are currently various frameworks of computational learning theory that share the same principles but serve different purposes. For example, some of the frameworks of computational learning theory are: PAC Learning, Distribution Learning, Online Learning, among others.

In this section we will focus on two frameworks: PAC learning and Distribution learning, defining their main goals and describing some of the techniques often used in each case, meanwhile providing some examples. For further reading, we refer to Kearns et al. [1994b], Shalev-Shwartz and Ben-David [2014] and Blum et al. [2020].

3.2.1 Probably Approximately Correct learning

In order to examine a problem of the real world from a mathematical perspective, it is necessary to be able to express it in a formal, abstract way. The PAC learning model is the first such expression for the learning problem.

The context. Assume that $\mathcal{X}$ is a set of objects for which one wants to be able to have some knowledge about and $\mathcal{Y}$ a set that quantifies this knowledge, which we call labels set. In particular, we assume that there is a rule $f : \mathcal{X} \to \mathcal{Y}$ that matches the elements of $\mathcal{X}$ to elements of $\mathcal{Y}$. The problem we need to address is approximating f using a finite number of input data. Note that if nothing is assumed of f, then the best thing that can be inferred for f is nothing more than what data directly suggest. However, in reality, each problem has a specific structure, which corresponds to what is called *prior knowledge*. Exploiting the structure of the problem and combining it with the input data, one may be able to sufficiently approximate f.

Definition. In PAC learning model, the input data consist of pairs of the form: $(x, y), x \in \mathcal{X}, y = f(x) \in \mathcal{Y}$. In this case, we say that the input data are labelled, for obvious reasons. We also assume that there is an unknown distribution $\mathcal{D}$ over $\mathcal{X}$ from which an independent sample is drawn each time an input pair is formed. The goal is to find some $h : \mathcal{X} \to \mathcal{Y}$ for which the following quantity:

$$L_{\mathcal{D},f}(h) = \mathrm{Pr}_{x\sim\mathcal{D}}[h(x) \neq f(x)]$$

is small. The quantity $L_{\mathcal{D},f}(h)$ is called the loss function. Observe that calculating the value of the loss function is impossible, since f and $\mathcal{D}$ are unknown. However, making some assumptions about f, we can show that for a specified selection of h, the loss function can be bounded.

Making assumptions about f is in fact unavoidable, if one is interested in deriving useful results. The following theorem illustrates this observation:

Theorem 3.2.1: No free lunch

Let $\delta \in (0, 1)$ and $\epsilon \in (0, 1/2)$. Then, if $\mathcal{X}$ is not finite, there exist $\mathcal{D}, f$ such that any estimator h of f has loss: $L_{\mathcal{D},f}(h) \geq \epsilon$, with probability at least δ, for any finite sized input.

Therefore, it is crucial to possess some prior knowledge. The prior knowledge we assume we have is in the following form: *There exists some known class of functions $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ whose minimal loss element's loss is adequately small.* We call such a class of functions a *hypothesis class*. The particular case where we assume $f \in \mathcal{H}$ is called *realizability assumption*. We are now ready to formally define the PAC learning framework. For the following, we focus on the cases when $\mathcal{Y} = \{0, 1\}$.

Definition 3.2.1 (PAC learnability): *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$. Then $\mathcal{H}$ is called PAC learnable, if there exists a function $r = r(\mathcal{H}, \epsilon, \delta)$, where $\epsilon, \delta \in (0,1)$ such that for any $\epsilon, \delta \in (0,1)$, every distribution $\mathcal{D}$ over $\mathcal{X}$ and every function $f : \{0,1\} \to \mathcal{Y}$, there exists an algorithm that given an input of size at least r, returns with probability at least $1-\delta$ an element h of $\mathcal{H}$ with: $L_{\mathcal{D},f}(h) \leq \epsilon$.*

Empirical Risk Minimization. After defining the PAC learning model, one is interested in designing algorithms that find a hypothesis that minimizes the loss. The fundamental tool that helps the design of such algorithms is the Empirical Risk Minimization (ERM). More specifically, assume that:

- $I : (x_1, y_1), (x_2, y_2), \ldots, (x_r, y_r)$ is the input.
- $\hat{h}$ denotes an algorithm that inputs I and outputs an element of $\mathcal{H}$.

Then, the algorithm:

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \left\{ \frac{1}{r} \sum_{i \in [r]} \mathbb{1}\{h(x_i) \neq y_1\} \right\}$$

is called the empirical risk minimization algorithm. Note that $\hat{h}$ might be a subset of $\mathcal{H}$. ERM returns one of the elements of $\hat{h}$ at random.

Example: Finite hypotheses classes. In the specific case when $\mathcal{H}$ is a finite hypothesis class and $f \in \mathcal{H}$, then the size of input required to find a probably approximately correct hypothesis h can be bounded from above. In particular, if $|I| = r \geq \frac{1}{\epsilon} \log(|\mathcal{H}|/\delta)$, then ERM returns with probability at least $1-\delta$ an element h of $\mathcal{H}$ with:

$$L_{\mathcal{D},f}(h) \leq \epsilon$$

Let $L(I,h) = \frac{1}{r} \sum_{i \in [r]} \mathbb{1}\{h(x_i) \neq y_i\}$ (empirical risk). Since $f \in \mathcal{H}$ and $L(I,f) = 0$: $\min_{h \in \mathcal{H}} L(I,h) = 0$. Hence, the "bad" samples are those that assign zero empirical risk to hypotheses that have a loss greater than ϵ. Therefore:

$$\Pr[\text{error}] \leq \Pr_{I \sim \mathcal{D}^r}[\exists h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon \wedge L(I,h) = 0]$$

From the union bound, we get that:

$$\Pr[\text{error}] \leq |\{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}| \max\{\Pr_{I \sim \mathcal{D}^r}[L(I,h) = 0] | h : L_{\mathcal{D},f}(h) > \epsilon\}$$

However: $|\{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}| \leq |\mathcal{H}|$ and $\Pr_{I \sim \mathcal{D}^r}[L(I,h) = 0] \leq (1 - L_{\mathcal{D},f}(h))^r$, therefore:

$$\Pr[\text{error}] \leq |\mathcal{H}|(1-\epsilon)^r < |\mathcal{H}| e^{-\epsilon r},$$

from which we conclude that if $r \geq \frac{1}{\epsilon} \log(|\mathcal{H}|/\delta)$, then $\Pr[\text{error}] < \delta$.

Vapnik-Chervonenkis dimension. In many interesting problems, the hypotheses class must be infinite in order to have meaning. A graphic example is the problem of finding a hyperplane that separates points of $\mathbb{R}^d$ that are labelled 1 with those labelled 0. Assuming that there exists such a hyperplane, it would be interesting if we could approximate it satisfyingly using only a finite number of samples. It turns out that we can and, furthermore, there is a systematic way of measuring how "difficult" is a hypotheses class $\mathcal{H}$ with respect to PAC learning. This systematic way is referred to as Vapnik-Chervonenkis dimnesion (VC dimension) and had already been introduced by Vapnik and Chervonenkis [1968], before the definition of PAC learning model.

In order to define the VC dimension of a class $\mathcal{H}$, we have to insert some notation. Let $C \subseteq \mathcal{X}$. Then, for any $h \in \mathcal{H}$, we define h_C to be a function defined on C that has the same values with h on their common domain. Observe that it is possible that there exist $h, h' \in \mathcal{H}$ such that $h \neq h'$ but $h_C = h'_C$. We define the set $\mathcal{H}_C = \{h_C : h \in \mathcal{H}\}$. Then: $|\mathcal{H}_C| \leq 2^C$.

Definition 3.2.2: *Let $\mathcal{H}$ be a hypotheses class and $C \subseteq \mathcal{X}$. Then, we say that $\mathcal{H}$ shatters C if:*

$$|\mathcal{H}_C| = 2^{|C|}$$

A hypotheses class $\mathcal{H}$ shatters a set C if $\mathcal{H}$ for each labeling of the elements of C contains at least one hypothesis that labels the elements of C accordingly. For each hypotheses class, there exist a maximum number of elements that can be grouped in a set C that $\mathcal{H}$ shatters. This is the VC-dimension of $\mathcal{H}$:

Definition 3.2.3 (VC dimension): *Let $\mathcal{H}$ be a hypotheses class. Then, the VC dimension of $\mathcal{H}$ is defined as follows:*

$$VC(\mathcal{H}) = \sup_{C \subseteq \mathcal{X}} \{|C| : \mathcal{H} \text{ shatters } C\}$$

The interesting fact is that VC dimension can be shown to provide almost tight bounds for the sample complexity of learning a hypotheses class $\mathcal{H}$, according to the fundamental theorem of PAC learning:

Theorem 3.2.2: Fundamental Theorem of PAC learning

Let $\mathcal{H}$ be a hypotheses class and $VC(\mathcal{H}) = d \in \mathbb{N}$. Then, for the sample complexity $r = r(\mathcal{H}, \delta, \epsilon)$, where $\epsilon, \delta \in (0, 1)$, of learning $\mathcal{H}$ in the PAC setting, it holds:

$$r = \Omega\left(\frac{d + \log(1/d)}{\epsilon}\right) \text{ and } r = O\left(\frac{d\log(1/\epsilon) + \log(1/d)}{\epsilon}\right)$$

Remark 3.2.1: *The sample complexity of Theorem 3.2.2 can be achieved from ERM.*

Therefore, ERM has been proven to be an almost optimal learner, despite its generality. However, the efficient computation of ERM is not always an easy task and for that reason various techniques have emerged. Nonetheless, description of such methods exceeds the goals of this study.

3.2.2 Distribution learning

Another, slightly different framework of learning is the Distribution learning setting. The main difference with PAC learning is that instead of learning a rule for labeling data, we want to learn a rule for predicting which data will arrive, that is, their distribution, previously denoted with $\mathcal{D}$. In this case, the input data are not labelled.

In the general case, when no information about the distribution is known a priori, we encounter the same problem as in the PAC learning setting: there is not much we can learn about $\mathcal{D}$. Therefore, it is a common practice to hypothesize that $\mathcal{D}$ belongs to some family of distributions that is parameterized by some parameters that might or might not be numerical. For example, in the context of learning ranking distributions, a commonly used parameter is a central ranking, which is the most probable element of the support of $\mathcal{D}$ (mode).

Another example is assuming that $\mathcal{D}$ is a normal distribution. It is widely known that any multivariate normal distribution can be determined by a vector consisting of the expected values per dimension and a covariance matrix. Therefore, if we are able to determine these two structures within some small margin, then we will also have a good estimate of $\mathcal{D}$.

Definition. In order to measure how good an estimate distribution is, one may use some metric between distributions. One such metric is, for example, Total Variation distance, which equals to the maximum absolute difference over all possible events between the measures that each of two probability measures assign to an event. In fact, the distribution learning setting as introduced by Kearns et al. [1994a] uses the concept of distance between probability measures:

> ***Definition 3.2.4 (Distribution learning):*** *Let $\mathfrak{D}$ be a class of distributions over $\mathcal{X}$. Then $\mathfrak{D}$ is called efficiently learnable with respect to some metric between probability measures d, if for any $\epsilon, \delta \in (0,1)$ there exists a polynomial time algorithm that, given access to a sampler of any fixed but unknown $\mathcal{D} \in \mathfrak{D}$, outputs a distribution $\mathcal{D}'$ which satisfies:*
>
> $$Pr[d(\mathcal{D}, \mathcal{D}') \geq \epsilon] < \delta.$$
>
> *If $\mathcal{D}' \in \mathfrak{D}$ then the algorithm is said to be proper (otherwise: improper).*

Definition 3.2.4 is similar to the definition of PAC learning in many ways. In fact, considering a more general definition of PAC learning setting (which we did not provide), Distribution learning can be described as a specific case.

Parameter estimation. As we already mentioned, it is common to determine the class of distributions $\mathfrak{D}$ using some parameters. That is, the hardness of learning $\mathfrak{D}$ is concentrated on some specific quantities that usually live in a discrete or a continuous space. However, learning a specific parameter might be useful in itself, which is another motive for pursuing parameter estimation. Hence, we only have to design algorithms for estimating these parameters, as well as to show that the particular parameterization of $\mathfrak{D}$ that we use is good in the sense that small variations of a parameter imply small variations to the distance between elements of $\mathfrak{D}$. Furthermore, in the case that a parameter lives in a discrete space, we can be hopeful that we can design an estimation algorithm that retrieves the unknown parameter with high probability. In this case, the distance between

the estimated distribution and the actual one might also be zero, with high probability.

In general, we want to design estimators of the parameters of $\mathfrak{D}$ that are:

1. *Unbiased:* That means that if θ is the unknown parameter, $\hat{\theta}$ is its estimation and S is a sample of size r, then:
$$\mathbb{E}_{S\sim\mathcal{D}^r}[\hat{\theta}(S)] = \theta$$
2. *Consistent (with respect to some metric d'):* This means that when $r \to +\infty$, given a metric d' defined on the space where θ lives, the probability that $d'(\hat{\theta}(S), \theta) \geq \epsilon$ tends to zero for any $\epsilon > 0$.

However, the properties described above do not give any guarantee for the number of samples that might be needed in order to get an accurate approximation. For this reason, we are usually interested in designing estimation algorithms with bounded sample complexity for learning a parameter within a small margin of its true value, or even finding it exactly, with high probability. In that context, concentration inequalities are used, as well as probabilistic analysis of the specific model, namely the family of distributions $\mathfrak{D}$.

Maximum Likelihood Estimation. Like in the case of PAC learning, there is a general approach to the problem of estimating a parameter. This general approach is called Maximum Likelihood Estimation and is a special case of the Empirical Risk Minimization. The concept of Maximum Likelihood Estimation was formally introduced in the first decades of 20^{th} century, by Ronald Fisher (Pfanzagl [2011]). However, in the context of Distribution learning, it gained one more interpretation: it is linked to an algorithm (ERM) that provides nearly optimal sample complexity for a slightly different model.

Assume that $s = (x_1, x_2, \ldots, x_r) \in \mathcal{X}^r$. Also, let $\mathcal{D}(\theta)$ be the distribution in $\mathfrak{D}$ that corresponds to selecting the value of the unknown parameter to be θ. Then, the maximum likelihood estimation of θ is defined as follows:

$$\theta^* = \arg\max_{\theta} \Pr_{S\sim\mathcal{D}(\theta)^r}[S = s]$$

In other words, the maximum likelihood estimation of an unknown parameter is the selection that maximizes the probability to observe the sample that is actually observed. Therefore, there is some sense of optimality associated with it: It is the best excuse that the sample can give in order to justify its appearance.

However, there are cases when the maximum likelihood estimation corresponds to a problem that is believed to be unable to be solved efficiently (that is, in polynomial time). Nonetheless, we should be aware of the following important fact: We should never forget the structure of the problem we initially tried to solve - even if the maximum likelihood estimation corresponds to an NP complete problem, there is the chance that, since the input of our algorithm is created from some specific family of distributions, we are interested in analyzing the maximum likelihood estimation problem in a corresponding average case where it might be, in fact, easy.

Chapter 4

Probability and Permutations

In this chapter, we introduce the concept of permutations, identify the domains where they are useful and present some probabilistic models that involve them either as parameters or as elements of the support of the corresponding distribution. The reason why such models are useful is illustrated by what we mentioned in the previous sections: While, in the context of social choice theory, they provide an alternative point of view that helps surpass problems encountered in voting, techniques from the learning theory can be used in order to exploit them as prior information. For further studying, we refer to Marden [1996].

4.1 Permutations

4.1.1 Definitions and notation

Imagine we have a set A that contains $n \in \mathbb{N}$ distinct elements. Without loss of generality, assume that $A = \{1, 2, \ldots, n\} = [n]$. We call A the set of alternatives (its elements).

> ***Definition 4.1.1:*** *The function $\pi : A \to A$ is called a permutation of the elements of A, if π is a bijection.*

In other words, a permutation is a shuffling of the elements of A. The following statements are equivalent for any $i, j \in A$:

1. $\pi(i) < \pi(j)$
2. $i \succ_\pi j$

For any set A, we denote with $\mathfrak{S}_A$ or $\mathrm{Sym}(A)$ the set of all permutations of A. In the particular case that $A = [n]$: $\mathfrak{S}_A = \mathfrak{S}_n$. For any $B \subseteq A$ and any $\pi \in \mathfrak{S}_A$, we define the permutation of the elements of B which is induced be π as the element γ of $\mathfrak{S}_B$ for which:

$$i \succ_\gamma j \Rightarrow i \succ_\pi j$$

and we denote it with: $\pi|_B$.

It is interesting to point out that the set $\mathfrak{S}_n$ equipped with the operator of function composition is a group and in particular it is called the symmetric group. The identity

element of $\mathfrak{S}_n$ is the permutation $\pi_{id} : i \mapsto \pi_{id}(i) = i$. For further study on group theory and abstract algebra we refer to classic literature like Dummit and Foote [2004], as well as Fraleigh [2003].

For the following, if $i, j \in [n]$ we denote with $\pi_{i \leftrightarrow j}$ the permutation that we get if we swap the positions of i and j in π.

4.1.2 Distances between permutations

Since permutations are not numerical values, the notion of distance between them does not appear naturally. However, there are many ways to define a metric in $\mathfrak{S}_n$. For the following, let $\pi, \pi' \in \mathfrak{S}_n$.

Hamming distance. The hamming distance between two permutations is defined as the number of positions at which the two permutations differ. In particular:

$$d_{Ham}(\pi, \pi') = |\{i \in [n] : \pi^{-1}(i) \neq \pi'^{-1}(i)\}|$$

For example, if $\pi = 1 \succ 2 \succ 3 \succ 4$ and $\pi' = 3 \succ 2 \succ 1 \succ 4$ then: $d_{Ham}(\pi, \pi') = 3$. This distance is not very interesting, because it hides the structure of permutations. For example, rotating all elements to the right (the last element becomes first) would have the same distance from the initial permutation with that of its inverse.

Spearman's footrule. A more interesting metric between permutations was introduced by Spearman [1906] and it corresponds to the absolute dislocation of elements between two permutations as defined by Diaconis and Graham [1977]:

$$d_{Sf}(\pi, \pi') = \sum_{i \in [n]} |\pi(i) - \pi'(i)|$$

Spearman's footrule is indeed a more illustrative distance, since it takes into account the proximity between the positions of elements.

Kendall tau distance. The distance which we will use in the biggest part of the rest of this study is the Kendall tau distance, which corresponds to the number of swaps that the algorithm of Bubblesort performs in order to sort π into π' (or reversely). Equivalently, it equals the number of discordant pairs between the two permutations. Formally, we define:

$$d_{KT}(\pi, \pi') = |\{i < j : (\pi(i) - \pi(j))(\pi'(i) - \pi'(j)) < 0\}|$$

In contrast to the aforementioned metrics, the computation of Kendall tau distance is not straightforward. In fact, it can be computed in $O(n \log(n))$ time using Mergesort and in $O(n\sqrt{\log(n)})$ using more advanced techniques (Chan and Pătraşcu [2010]).

Although the time complexity of computing Kendall tau distance indicates that it is a generally complex metric and hence many of its properties are not obvious, it provides a meaningful interpretation of the distance between rankings, a statement which we hope we will justify throughout the rest of this work.

We now provide some properties of the Kendall tau distance. Of course, it can be easily seen that it satisfies all the properties of a metric: It is symmetric, it satisfies the triangle inequality and it becomes zero only when the two inputs coincide.

• *Relabeling:* Another useful property of Kendall tau distance is that it is independent of relabeling. That is, for any $\pi, \pi', \sigma \in \mathfrak{S}_n$:

$$d_{KT}(\pi, \pi') = d_{KT}(\pi\sigma, \pi'\sigma),$$

where $\pi\sigma(i) = \pi(\sigma(i)), \forall i \in [n]$. In particular, picking $\sigma = \pi'^{-1}$, we conclude that: $d_{KT}(\pi, \pi') = d_{KT}(\pi\pi'^{-1}, \pi_{id})$.

• *Swap increasingness:* An interesting question is how a distance between permutations behaves in relation to the swapping of elements in a permutation. More specifically, it would be interesting to be able to compare the quantities: $d_{KT}(\pi, \pi')$ and $d_{KT}(\pi_{i\leftrightarrow j}, \pi')$ for some $i, j \in [n]$, where we assume that the pair (i, j) is in the same order in π as in π'. Without loss of generality, we may assume that $i \succ_\pi j$ (and $i \succ_{\pi'} j$).

Imagine, first, that i and j are adjacent in π. That is: $|\pi(i) - \pi(j)| = 1$. Then, what is the Kendall tau distance of $\pi_{i\leftrightarrow j}$ and π'? Recall that Kendall tau distance is the number of discordant pairs between two permutations. Then, it is clear that: $d_{KT}(\pi_{i\leftrightarrow j}, \pi') = 1 + d_{KT}(\pi, \pi')$, since all other pairs preserve their order in each permutation and the pair (i, j) becomes discordant.

We now concentrate on the case that i, j are not adjacent in π. Of course, the pair i, j now becomes discordant. However, the pairs that involve either i or j and some alternative that is ordered between them in π, might also become concordant (with respect to π'), previously being discordant. Let $k \in [n]$ such that: $i \succ_\pi k \succ_\pi j$. Then, if we swap i and j and (i, k) becomes concordant, that means that: $k \succ_{\pi'} i$ and since $i \succ_{\pi'} j$, this implies that $k \succ_{\pi'} j$. Therefore, (j, k) becomes discordant after the swap, while previously it was concordant. Observe that the symmetric argument is equivalent. Hence:

$$d_{KT}(\pi_{i\leftrightarrow j}, \pi') \geq 1 + d_{KT}(\pi, \pi')$$

In other words, Kendall tau distance is *swap increasing*.

• *Aggregation:* Suppose that $\pi_1, \pi_2, \ldots, \pi_r \in \mathfrak{S}_n$ where $r \in \mathbb{N}$. Then, it holds that:

$$d_{KT}(\pi, \pi_\ell) = \sum_{i \succ_\pi j} \mathbb{1}\{j \succ_{\pi_\ell} i\}$$

Therefore, if for any $i, j \in [n]$, we define $q(i \succ j)$ as: $q(i \succ j) = \sum_{\ell \in [r]} \mathbb{1}\{i \succ_{\pi_\ell} j\}$, we have:

$$\sum_{\ell \in [n]} d_{KT}(\pi, \pi_\ell) = \sum_{i \succ_\pi j} q(j \succ i) = \binom{n}{2} r - \sum_{i \succ_\pi j} q(i \succ j) \tag{4.1}$$

• *Incomplete permutations:* Suppose that $S \subseteq A$ and $\gamma \in \mathfrak{S}_S$. Then, we define the quantity $d_{KT}(\pi, \gamma)$ as follows:

$$d_{KT}(\pi, \gamma) = d_{KT}(\pi|_S, \gamma),$$

which is the Kendall tau distance between the induced subpermutation of π on $\mathfrak{S}_S$ and γ. Therefore:

$$d_{KT}(\pi, \gamma) = \sum_{i \succ_\pi j} \mathbb{1}\{j \succ_\gamma i\},$$

where $j \succ_\gamma i \Leftrightarrow i, j \in S \wedge \gamma(j) < \gamma(i)$. Therefore, if $\mathcal{S} = (S_1, S_2, \ldots, S_r)$, where $S_1, S_2, \ldots, S_r \subseteq A$ and $\gamma_\ell \in \mathfrak{S}_{S_\ell}, \forall \ell \in [r]$ and also: $q(i \succ j) = \sum_{\ell \in [r]} \mathbb{1}\{i \succ_{\gamma_\ell} j\}$, we have:

$$\sum_{\ell \in [r]} d_{KT}(\pi, \gamma_\ell) = \sum_{i \succ_\pi j} q(j \succ i)$$

We denote with $W_{ij} = W_{ij}(\mathcal{S})$ the number of indexes ℓ in $[r]$ for which $i, j \in S_\ell$. Then:

$$\sum_{\ell \in [r]} d_{KT}(\pi, \gamma_\ell) = \sum_{i \succ_\pi j} W_{ij} - \sum_{i \succ_\pi j} q(i \succ j)$$

However, it holds that: $W_{ij} = W_{ji}$, therefore:

$$\sum_{\ell \in [r]} d_{KT}(\pi, \gamma_\ell) = \sum_{i < j} W_{ij} - \sum_{i \succ_\pi j} q(i \succ j) \tag{4.2}$$

4.2 Probabilistic models of permutations

For the following, we will use the term ranking in order to refer to a permutation. The term ranking implies that the order of the alternatives in a permutation corresponds to a preference over them. This is what links permutations to social choice theory. Social choice theory aims to define properties that the rules used to aggregate public opinion should ideally satisfy, as well as establish such rules and study them from a computational point of view. Probabilistic models of permutations have a rather similar goal, but use a different approach: They hypothesize that each input ranking, instead of a vote, corresponds to a sample drawn independently from some ranking distribution that belongs to a parametric family of distributions and its parameters' values are unknown. The goal is to learn these parameters, using a small sample. These parameters typically correspond directly or indirectly to an underlying ranking of the alternatives which is the ideal output of the aggregation algorithm in terms of social approval. The voters become samplers (or noisy voters) and vote aggregation reduces to learning a distribution.

4.2.1 Important ranking models

There are many probabilistic ranking models. Perhaps some of the most thoroughly studied are:

1. The *Placket-Luce model*, which was introduced independently by Plackett [1975] and Luce [1959]. In this model, we assume that each alternative $i \in [n]$ corresponds to an individual value w_i, which expresses how "valuable" it is, in the sense that higher w_i (relatively to the values of the other alternatives) implies that it is likelier to rank i higher. In particular, the sampling process is the following:

   -Pick an alternative at random, where each alternative i is picked with probability $w_i / \sum_{j \in [n]} w_j$ and place it in the highest available position.

-Restrict on the rest alternatives and repeat the process until no alternative remains. Note that the sum of weights in the denominator of the probabilities of selections includes each time one less summand.

2. *Models induced by pairwise comparisons.* In this case, for each pair of alternatives $i, j \in [n]$ we define a quantity $p_{ij} = 1 - p_{ji} \in [0, 1]$. In order to sample a ranking, we use the rejection sampling method: First, we create a tournament graph, where each vertex corresponds to an alternative and the direction of the edge between each pair of alternatives $\{i, j\} \in [n]$ is selected independently and is from i to j with probability p_{ij}. If the resulting tournament graph is not acyclic, we reject it and restart the process. When we eventually create an acyclic tournament graph, we return the ranking that corresponds to its unique topological ordering.

3. *Mallows models* or distance based models. The idea behind such models is that there exists a central permutation which is typically denoted with π_0 and the probability of sampling a permutation $\pi \in \mathfrak{S}_n$ is linked with the value of some notion of distance between π_0 and π. In particular, π_0 is the mode of the distribution and the probabilities of sampling diminish exponentially to the distance between π_0 and π. The Mallows model was introduced by Mallows [1957]. We will focus on the case when the distance used is the Kendall tau distance and whenever we refer to Mallows model, we will assume that it uses this particular metric. One might think the Mallows distribution as the Gaussian-equivalent distribution on $\mathfrak{S}_n$.

4.2.2 Mallows model

Definition and notation. The probabilistic model for the ranking generation with which we will work is the Mallows model. The Mallows model is associated with two parameters:

1. The central ranking $\pi_0 \in \mathfrak{S}_n$.
2. The spread parameter $\phi = e^{-\beta}$, where $\beta > 0$.

We denote with $\mathcal{M}_{\pi_0,\beta}$ the Mallows distribution with central ranking π_0 and spread parameter $e^{-\beta}$. For any $\pi \in \mathfrak{S}_n$ we denote with $\Pr[\pi|\pi_0, \beta]$ or, when the corresponding parameters are clear from the context: $\Pr[\pi]$ or $\Pr[\pi|\pi_0]$, the probability that we sample π from $\mathcal{M}_{\pi_0,\beta}$. Then, it holds that:

$$\Pr[\pi] = \frac{1}{Z} e^{-\beta d_{KT}(\pi_0,\pi)} \tag{4.3}$$

Note that Z is a normalization constant that depends on β and n and is equal to:

$$Z = \prod_{k=0}^{n-1} \sum_{t=0}^{k} e^{-\beta t} \tag{4.4}$$

Sampling method. It can be shown that Mallows model is in fact a specific example of a model induced by pairwise comparisons, as defined in 2. Therefore, a possible sampling process is based on rejection sampling, where $p_{ij} = \frac{e^{-\beta}}{1+e^{-\beta}}$ for any $i \succ_{\pi_0} j$, which is attributed to the French mathematician and philosopher Marquis de Condorcet, during

the Age of Enlightenment. However, this method is not computationally efficient and also hides some of the structure of Mallows model. For this reason, the so called Repeated Insertion Model was proposed by Doignon et al. [2004].

The Repeated Insertion Model (RIM) is based on the idea that a permutation can be constructed by iteratively inserting the alternatives in a way that enables each time to select the position of the inserted element independently with probabilities that can be analytically determined a priori.

More specifically, the output ranking is created as follows. We insert the alternatives according to their order in π_0. For simplicity, without loss of generality, assume that $\pi_0 = \pi_{id}$. For any $i \in [n]$ and any $j \leq i$, we define:

$$p_i(j) = \frac{e^{-\beta(i-j)}}{\sum_{k=0}^{i-1} e^{-\beta k}} \tag{4.5}$$

Then, we run the following procedure:

Algorithm 1: Repeated Insertion Algorithm

Result: π
$\pi(1) = 1$;
$\pi(i) = -1, \forall i > 1$;
for $i = 2, \ldots, n$ **do**
  Pick $j \in [i]$ at random, where: $\Pr[j] = p_i(j), \forall j \in [i]$;
  for *any* $i' \in [i-1]$ *for which* $\pi(i') \geq j$ **do**
    $\pi(i')$++;
  end
  $\pi(i) = j$;
end

It can be easily seen that the output of Algorithm 1 is always a ranking. Also, it can be proven that for each $\pi \in \mathfrak{S}_n$, the probability that Algorithm 1 outputs π equals the probability that π is sampled from $\mathcal{M}_{\pi_0,\beta}$.

Maximum likelihood estimation of central ranking. The central ranking of a Mallows distribution has a specific meaning: It is the "common truth" around which the sampled rankings fluctuate. Therefore, under Mallows model, the central ranking is a parameter that we would like to learn. As we analyzed in Section 3.2.2, a common technique is to compute the maximum likelihood estimation (MLE) of the parameter of interest, in order to estimate it in an optimal way, given a number of independent samples. It turns out that finding the MLE of the central ranking given r independent samples of $\mathcal{M}_{\pi_0,\beta}$ is equivalent to applying Kemeny's rule (Kemeny [1959]) on the samples. We denote with π^* the MLE of the central ranking.

Let $\pi_1, \pi_2, \ldots, \pi_r \sim \mathcal{M}_{\pi_0,\beta}$, independent. Then, for the MLE of π_0 we have:

$$\pi^* = \arg\max_{\pi \in \mathfrak{S}_n} \prod_{\ell \in [r]} \Pr[\pi_\ell | \pi_0]$$

Due to Eq.(4.3), we get that:

$$\pi^* = \arg\min_{\pi \in \mathfrak{S}_n} \sum_{\ell \in [r]} d_{KT}(\pi_0, \pi_\ell) \tag{4.6}$$

Unfortunately, the problem of finding π^* as described in Eq.(4.6) is shown to be NP-hard (Bartholdi et al. [1989]). Even in the particular case that $r = 4$, the problem remains NP-hard (Dwork et al. [2001a]). There are constant factor approximation algorithms for this problem. In Diaconis and Graham [1977] and Dwork et al. [2001b] 2-approximation algorithms are presented, while in Ailon et al. [2008] a simple algorithm (KwikSort) that works for a family of related problems is introduced and combined with another simple algorithm, yields an $11/7-$approximation. More complicated techniques can be used in order to get a PTAS for this problem, as presented by Kenyon-Mathieu and Schudy [2007].

However, we are interested in finding MLE exactly, under the Mallows model. Towards this direction, there have been proposed various heuristics, for example by Fligner and Verducci [1990], Cohen et al. [1998] and Meila et al. [2007]. Nevertheless, the first algorithm that computes π^* in polynomial $n-$time with high probability was introduced by Braverman and Mossel [2009], whose work we will present in a following chapter.

Retrieving the central ranking. The maximum likelihood estimation of π_0 is one way to estimate it. Arguably, among all the possible estimators of π_0, the maximum likelihood estimation achieves the optimum sample complexity for retrieving π_0. However, as shown by Caragiannis et al. [2013], there exists a wide family of estimators that includes polynomial time ones, that can achieve optimum sample complexity. In fact, the sample complexity of retrieving the central ranking of $\mathcal{M}_{\pi_0,\beta}$ has been shown to be $\Theta(\log(n/\epsilon))$, where $\epsilon \in (0, 1/2]$ is the threshold of error probability. The techniques that achieve this sample complexity will be discussed in the following chapter.

Estimating the spread parameter. Although our own work is focused on central ranking estimations under a model that we propose and is a generalized version of Mallows model, we shall present results concerning the estimation of the spread parameter. The spread parameter monitors the amount of uncertainty of the model. In fact, when $\phi = e^{-\beta}$ tends to become 1, the model degenerates into a state of maximum entropy: the uniform distribution on $\mathfrak{S}_n$, while when $\phi \to 0$, the model becomes deterministic and always outputs the central ranking π_0. The technique presented by Mukherjee et al. [2016] can be used in order to estimate the spread parameter, given a single sample, when the central ranking is known. As the number of alternatives grows, the estimation of the spread parameter becomes more accurate. However, the sample complexity for estimating the spread parameter was settled in the work of Busa-Fekete et al. [2019], where a more general model was considered, namely Mallows Block model. In both works, the parametric family of Mallows distributions with the same central ranking and unknown spread parameters, was viewed as an exponential family, whose properties were exploited in order to acquire the aforementioned results. In particular, the following theorem holds:

Theorem 4.2.1: (Busa-Fekete et al.)

Let $\pi_0 \in \mathfrak{S}_n$, $\phi = e^{-\beta} \in [0, 1]$ and $\epsilon, \delta \in (0, 1)$. If π_0 is known, then there exists an estimator $\hat{\phi}$ of ϕ that can be computed in polynomial time from r i.i.d. samples from $\mathcal{M}_{\pi_0,\beta}$ such that if r is at least equal to some value that is $O(\frac{1}{n\epsilon^2}\log(1/\delta))$, then:

$$Pr_{\Pi\sim\mathcal{M}^r_{\pi_0,\beta}}[\hat{\phi}(\Pi) \in [\phi - \epsilon, \phi + \epsilon]] \geq 1 - \delta$$

Generalizations. Several generalizations of the Mallows model have been proposed.

There is the possibility of considering a different distance metric between permutations, as proposed by Fligner and Verducci [1986]. For example, a widely used alternative metric is Cayley distance. The resulting model is called Cayley-Mallows model and some of its properties are examined for example in Irurozki et al. [2018]. However, depending on the selection of the metric, the structure of the model varies significantly. Furthermore, the selection of metric considerably influences the performance of the model in modeling different problems.

Another generalization that one might naturally consider is the case of mixture models. As we already mentioned, Mallows distribution is unimodal. However, in many settings, unimodality is unrealistic. Mixtures of two Mallows distributions have been examined by Awasthi et al. [2014], while mixtures of any constant number of Mallows models have been studied by Liu and Moitra [2018]. In De et al. [2018] mixtures of Cayley-Mallows distributions were studied, among other ranking models, and an algorithm that runs in quasi-polynomial time to the number of components of the mixture was introduced.

A classic generalization of Mallows model is the Generalized Mallows model, which was considered in the works of Fligner and Verducci [1986] and Doignon et al. [2004], among others. Recall that in the repeated insertion model, each time an alternative $i \in [n]$ is inserted, we calculate the values $p_i(j), \forall j \in [i]$, according to Eq.(4.5). Therefore, there is the possibility to assign each alternative a different spread parameter $\phi_i = e^{-\beta_i}$, and acquire a different ranking model, which is called Generalized Mallows model. Intuitively, this model corresponds to cases where each alternative might be more or less agile relatively to the others, that is its position in a sample might be more or less uncertain. In Busa-Fekete et al. [2019], a model that interpolates between Mallows and Generalized Mallows is introduced: the Mallows Block model. In this case, there are groups (blocks) of alternatives that have the same corresponding spread parameter. This model fills the gap between Mallows and Generalized Mallows models and the tight bounds presented in the same work provide a deeper understanding of the structure of the space between them. The model we present in this work aims to achieve a similar goal, but for a different kind of generalization.

Finally, a natural type of generalizations of Mallows model is the one that takes into account the possibility that the sampled rankings do not include all of the alternatives, primarily due to the large number of alternatives in many applications. Instead of complete rankings, we consider incomplete ones: They do not include all the alternatives. A possible way to generalize the Mallows model in this direction is to assume that only a number of highest ranked elements are important, while the others are secondary. These models are referred to as top-t models and are considered in the works of Fligner and Verducci [1986], Busse et al. [2007], Meila and Bao [2010], Meila and Chen [2012] and Tang [2018], among others. In the work of Chierichetti et al. [2018], a model in which each sample includes only a small number of rankings that correspond to the most preferable ones is introduced and analyzed. From another point of view, one might consider models that correspond to projecting Mallows samples to smaller sets of alternatives. That is, for each incomplete sample observed, there is some underlying complete sample that corresponds to it. However, each of these perspectives fails to exploit the following possibility: The agents that actually generate the samples (which we try to model) do not have any information about some of the alternatives. That is, they are constrained to a specific subset of alternatives which does not necessarily coincide with the set of their top (or bottom) preferences, but

with those that they can rank. For example, say that there are three movies A, B, C. Bob has only watched A and B. Therefore, he is unable to rank C. It might be the case that if Bob watched C, it would become either his favorite or his least favorite movie. We are now ready to introduce our model.

4.2.3 Selective Mallows model

As we already mentioned, the agents that generate the samples might be aware only of some of the alternatives. In our model, we, again, assume that there exists a central ranking $\pi_0 \in \mathfrak{S}_n$ which is the most socially accepted. However, each sample π has access to a restricted version of π_0, according to some set $S \subseteq [n]$ for which: $\pi \in \mathfrak{S}_S$.

Definition and notation. Given the set S, which we call the selection set, we define the Selective Mallows distribution $\mathcal{M}^S_{\pi_0,\beta}$ for which the probability of observing $\pi \in \mathfrak{S}_S$ is denoted with $\Pr[\pi|\pi_0, \beta, S]$ (or without explicit declaration of any parameter that is clear by the context) and equal to:

$$\Pr[\pi|\pi_0, \beta, S] = \frac{1}{Z(S)} e^{-\beta d_{KT}(\pi_0|_S, \pi)}, \tag{4.7}$$

where $Z(S)$ is a normalization constant such that: $\sum_{\pi \in \mathfrak{S}_S} \Pr[\pi|\pi_0, \beta, S] = 1$, which turns out to be, according to Eq.(4.4):

$$Z(S) = Z(|S|) = \prod_{k=0}^{|S|-1} \sum_{t=0}^{k} e^{-\beta t} \tag{4.8}$$

Observe that under the Selective Mallows model, the alternatives that are not included in the selection set do not influence the probabilities of appearance of each possible incomplete permutation. That is, for example, if $\pi_0 = \pi_{id}$, $n = 5$ and $S = \{1, 5\}$, the probability that the alternatives 1, 5 swap is equal to the probability of swap of adjacent alternatives in classic Mallows model. Observe that although 1 and 5 are distant in π_0, without knowing the intermediate alternatives, they do not behave as if they are distant. We attribute this property to what we call *ignorance bias*: Due to lack or ignorance of a reliable measure of the value of the alternatives, the only way to rank them is by comparing one with another. Our thesis is that: *Everything is relative, even relative distance.*

Multiple samples. Under this model, the notion of independent samples must be slightly generalized. In particular, picking a fixed $S \subseteq [n]$ and drawing independent samples from $\mathcal{M}^S_{\pi_0,\beta}$ is rather pointless: This is identical to sampling a Mallows distribution on $\mathfrak{S}_S$. Therefore, in order to get a sample of size $r \in \mathbb{N}$, we consider a vector of selection sets $\mathcal{S} = (S_1, S_2, \ldots, S_r)$, where $S_\ell \subseteq [n], \forall \ell \in [r]$. We define the set $\mathfrak{S}^{\mathcal{S}} = \times_{\ell \in [r]} \mathfrak{S}_{S_\ell}$ and the constrained product selective Mallows distribution $\mathcal{M}^{\mathcal{S}}_{\pi_0,\beta}$, for which the probability of observing $\Pi = (\pi_\ell)_{\ell \in [r]} \in \mathfrak{S}^{\mathcal{S}}$ is denoted with $\Pr[\Pi|\pi_0, \beta, \mathcal{S}]$ and equals:

$$\Pr[\Pi|\pi_0, \beta, \mathcal{S}] = \prod_{\ell \in [r]} \Pr[\pi_\ell|\pi_0, \beta, S_\ell] \tag{4.9}$$

In other words, the samples $(\pi_\ell)_{\ell \in [r]}$ are independent conditioned on the selection sets $\mathcal{S}$.

Selecting the sets. There are three different ways to view the process that generates the selection sets:

1. *Explicitly:* This is the general case, where $\mathcal{S}$ is defined explicitly. There is no assumption that can be made for the procedure that generates the selection sets. This means that they might have been picked dependently on one another. Since in this case the selection sets can be picked adversarially, we denote the corresponding model as $\mathcal{M}_{\pi_0,\beta}^{\text{ADV}}$ or as $\mathcal{M}_{\pi_0,\beta}^{\text{ADV}\to\mathcal{S}}$ if we want to fix the selection sets vector.

2. *Randomly:* In this case, we assume that there exists some distribution $\mathcal{D}$ over $2^{[n]}$ (which we call selection distribution) that generates selection sets. We denote the corresponding model as: $\mathcal{M}_{\pi_0,\beta}^{\text{RND}(\mathcal{D})}$. The concept of independent samples is now clearer:
$$\Pr[\Pi] = \prod_{\ell\in[r]} \mathcal{D}(S_\ell)\Pr[\pi_\ell] \tag{4.10}$$
The product distribution is denoted with $\mathcal{M}_{\pi_0,\beta}^{\text{RND}(\mathcal{D},r)}$.

3. *Adaptively:* In this case, one is given access to a Selective Mallows sampler, namely a random generator that inputs a selection set $S \subseteq [n]$ and outputs an element of $\mathfrak{S}_S$ according to $\mathcal{M}_{\pi_0,\beta}^{S}$. This implies that there is the possibility of picking the selection sets in the runtime of an algorithm that aims, for example, to retrieve the central ranking. However, there is a constraint: Each selection set must be of size no more than $m \leq n$. We denote the corresponding model with $\mathcal{M}_{\pi_0,\beta}^{\text{ADP}(m)}$.

Maximum likelihood estimation of central ranking. Suppose we are given a sample $\Pi = (\pi_\ell)_{\ell\in[r]}$ drawn from $\mathcal{M}_{\pi_0,\beta}^{\mathcal{S}}$, where $\mathcal{S} = (S_\ell)_{\ell\in[r]}$, $S_\ell \subseteq [n], \forall \ell \in [r]$. Then, the maximum likelihood estimation of π_0 from Π is:

$$\pi^* = \arg\max_{\pi\in\mathfrak{S}_n} \Pr[\Pi|\pi,\beta,\mathcal{S}]$$

However, due to Eq.(4.7) and (4.9), we get that:

$$\pi^* = \arg\min_{\pi\in\mathfrak{S}_n} \sum_{\ell\in[r]} d_{KT}(\pi_0|_{S_\ell},\pi_\ell) \tag{4.11}$$

Comparing Eq.(4.6) to Eq.(4.11), it becomes clear that the structure of the problem of finding the maximum likelihood estimation of central ranking under Mallows model is pretty similar to that of the structure of the problem of finding the maximum likelihood estimation of central ranking under the Selective Mallows model. We now provide formal definitions of these problems and a relaxation for each one.

Definition 4.2.1 (MRP): *The Mallows Reconstruction Problem (*MRP*) is the problem of finding a ranking $\pi \in \mathfrak{S}_n$, given a vector Π of r independent samples from $\mathcal{M}_{\pi_0,\beta}$, where $\pi_0 \in \mathfrak{S}_n, \beta > 0$ for which:*

$$Pr[\Pi|\pi,\beta] \geq Pr[\Pi|\pi_0,\beta]$$

That is, MRP corresponds to finding a ranking that is at least as likely as the central ranking. This solution concept was introduced by Rubinstein and Vardi [2017] and is, arguably, at least as useful as the maximum likelihood estimation of the central ranking.

Definition 4.2.2 (max-MRP): *The maximum-Mallows Reconstruction Problem (*MAX-MRP*) is the problem of finding a ranking $\pi \in \mathfrak{S}_n$, given a vector Π of r independent samples from $\mathcal{M}_{\pi_0,\beta}$, where $\pi_0 \in \mathfrak{S}_n, \beta > 0$ for which:*

$$Pr[\Pi|\pi,\beta] \geq Pr[\Pi|\pi',\beta], \forall \pi' \in \mathfrak{S}_n$$

Clearly, the MAX-MRP corresponds to finding the maximum likelihood estimation of the central ranking.

Similarly, we define the corresponding problems for the case of Selective Mallows model.

Definition 4.2.3 (SMRP): *The Selective Mallows Reconstruction Problem (*SMRP*) is the problem of finding a ranking $\pi \in \mathfrak{S}_n$, given a vector Π of a sample of size r drawn from $\mathcal{M}^{\mathcal{S}}_{\pi_0,\beta}$, where $\mathcal{S} = (S_\ell)_{\ell\in[r]}$, $S_\ell \subseteq [n], \forall \ell \in [r]$ and for every pair of alternatives $i, j \in [n]$, the number of sets in which: $i, j \in S_\ell$ is at least pr, for some $p \in (0, 1]$, $\pi_0 \in \mathfrak{S}_n$ and $\beta > 0$ for which:*

$$Pr[\Pi|\pi,\beta,\mathcal{S}] \geq Pr[\Pi|\pi_0,\beta,\mathcal{S}]$$

The maximum likelihood estimation problem is equivalent to the following:

Definition 4.2.4 (max-SMRP): *The maximum-Selective Mallows Reconstruction Problem (*MAX-SMRP*) is the problem of finding a ranking $\pi \in \mathfrak{S}_n$, given a vector Π of a sample of size r drawn from $\mathcal{M}^{\mathcal{S}}_{\pi_0,\beta}$, where $\mathcal{S} = (S_\ell)_{\ell\in[r]}$, $S_\ell \subseteq [n], \forall \ell \in [r]$ and for every pair of alternatives $i, j \in [n]$, the number of sets in which: $i, j \in S_\ell$ is at least pr, for some $p \in (0, 1]$, $\pi_0 \in \mathfrak{S}_n$ and $\beta > 0$ for which:*

$$Pr[\Pi|\pi,\beta,\mathcal{S}] \geq Pr[\Pi|\pi',\beta,\mathcal{S}], \forall \pi' \in \mathfrak{S}_n$$

Interpolating between two models. A way to view the Selective Mallows model is like an interpolation between the classic Mallows model and the Noisy Comparisons model. While the former is already familiar to the reader, the latter can be thought of as follows:

• *Noisy Comparisons model:* There exists a central ranking $\pi_0 \in \mathfrak{S}_n$ and some $\theta \in (0, 1/2)$ which are the parameters of the model. The output of a Noisy Comparisons generator is an ordered pair of alternatives. Fix $i, j \in [n], i < j$ and suppose that $\pi_0 = \pi_{id}$. Then:

$$\Pr[i \succ j|\pi_0] = 1 - \Pr[j \succ i|\pi_0] = \frac{1}{2} + \theta,$$

under the noisy comparisons model.

Observe that while the Mallows model outputs complete rankings, the Noisy Comparisons model outputs ordered pairs. Their fundamental difference is that under the Noisy Comparisons model, each pair of alternatives swaps with the same probability. However, one could consider a model that is similar to the Noisy Comparisons model, but for which there exist different parameters $(\theta_{ij})_{i<j}$ that determine the probabilities of swap. Furthermore,

$(\theta_{ij})_{i<j}$ can be picked in order to correspond to the probabilities of swap of the pair (i, j) under the Mallows model (which decreases to the distance of i, j in π_0). However, in either case, the parameters of swap probability have to be fixed. We argue that this is not accurate: The swap probabilities must be determined by the selection set.

Revisiting the Noisy Comparisons model, suppose that we are given r ordered pairs $o_1, o_2, \ldots, o_r$ whose order was determined independently from the Noisy Comparisons model. Then, the maximum likelihood estimation of π_0 is:

$$\pi^* = \arg\max_{\pi \in \mathfrak{S}_n} (1/2+\theta)^{\sum_{i \succ_\pi j} q(i \succ j)} (1/2-\theta)^{\sum_{i \succ_\pi j} q(j \succ i)}$$

Recall that $q(i \succ j)$ denotes the number of samples where i, j are compared and i is ranked before j. From arguments similar to those used in order to get Eq.(4.2), we get that:

$$\pi^* = \arg\max_{\pi \in \mathfrak{S}_n} \left(\frac{1/2+\theta}{1/2-\theta}\right)^{\sum_{i \succ_\pi j} q(i \succ j)} (1/2-\theta)^r = \arg\max_{\pi \in \mathfrak{S}_n} \sum_{i \succ_\pi j} q(i \succ j)$$

Therefore, due to Eq.(4.2):

$$\pi^* = \arg\min_{\pi \in \mathfrak{S}_n} \sum_{\ell \in [r]} d_{KT}(\pi_0, o_\ell) \tag{4.12}$$

Hence, we have concluded that Eq.(4.11) includes the expressions of MLE of the central ranking for all three models: Mallows, Selective Mallows and Noisy Comparisons. That is, the reconstruction problem in all three cases is the same and Selective Mallows model includes the other models, by appropriately selecting the parameter β and the selection sets vector $\mathcal{S}$: The Selective Mallows model is the natural generalization of the Mallows model.

Chapter 5

Learning a Hidden Ranking

In this chapter we consider the problem of retrieving the central ranking exactly, from Selective Mallows samples. In particular, we provide some asymptotic sample complexity bounds for the problem of retrieving the central ranking with high probability, under the models $\mathcal{M}_{\pi_0,\beta}^{\mathrm{RND}(\mathcal{D})}$ and $\mathcal{M}_{\pi_0,\beta}^{\mathrm{ADV}}$, where $\pi_0 \in \mathfrak{S}_n, \beta > 0$ and $\mathcal{D}$ is a selection distribution. We also refer to a classic result (by Feige et al. [1994]) that implies a tight sample complexity bound for the problem under the model $\mathcal{M}_{\pi_0,\beta}^{\mathrm{ADP}(m)}$, when $m = 2$. In the case that $m > 2$, the problem remains open, as we will discuss in a following chapter.

Comment on notation. For the following, Π will be used to represent either a fixed vector of rankings (complete or incomplete) or a random variable (vector), when we use the following notation: $\Pr_{\Pi}[\cdot]$. When Π denotes a random variable (vector), then its distribution will be clear by the context or it will be explicitly declared. For example:

1. $\Pr[f(\Pi) = x]$ refers to the probability that the output of (possibly random) process f equals x, when the (fixed) vector Π is provided to its input. In this case the randomness is included exclusively on f and/or x.

2. $\Pr_{\Pi}[f(\Pi) = x]$ takes into account the randomness involved in the selection of Π. Here, Π is a random variable (vector).

5.1 Learning under Mallows model

In this section we present the work of Caragiannis et al. [2013], in which asymptotically optimal bounds for the sample complexity of learning the central ranking under Mallows model were established, by using methods that are also useful for the analysis of learning under Selective Mallows model.

Upper Bound. The main idea for solving this problem is to consider a tournament graph where the direction of each edge is determined as the one that is found in the majority of samples we draw. If this graph turns out to be acyclic, we return its unique topological ordering as the estimation of central ranking. This defines a family of estimators that are called pairwise majority consistent estimators. Although Theorem 5.1.1 works for

any pairwise majority consistent (PM-c) estimator, we will focus on the following PM-c estimator, which we call *positional estimator* and denote with $\hat{\pi}$, for reasons that will become clear in the following chapter.

$$\hat{\pi}(i) = 1 + \sum_{j \in [n] \setminus \{i\}} \mathbb{1}\{q(j \succ i) > q(i \succ j)\}, \forall i \in [n] \tag{5.1}$$

Assume that in any possible ties are broken uniformly from left to right. That is, if $\hat{\pi}(1) = \hat{\pi}(2) = \hat{\pi}(3) = 1$ and $\hat{\pi}(i) > 1, \forall i > 3$, we pick a uniform permutation of $\{1, 2, 3\}$ which is put on the first 3 positions of $\hat{\pi}$ and shift any element $i > 3$ with $\hat{\pi}(i) \in \{2, 3\}$ to the position 4. We then repeat the process until $\hat{\pi}$ becomes a ranking. For the following, when we refer to $\hat{\pi}$, we will clarify whether we consider it to be the function defined in Eq.(5.1) before or after breaking the ties.

Theorem 5.1.1: Caragiannis et al. [2013]

Let $\mathcal{M}_{\pi_0,\beta}$ be a Mallows distribution with central ranking $\pi_0 \in \mathfrak{S}_n$ and spread parameter $\beta > 0$. For any $\epsilon > 0$, there exists an algorithm that, given a Mallows profile drawn from $(\mathcal{M}_{\pi_0,\beta})^r$ for any r at least equal to some value $O((1-e^{-\beta})^{-2} \log(n/\epsilon))$, retrieves the central ranking π_0 with probability at least $1 - \epsilon$.[a]

[a] $(1 - e^{-\beta})^{-2} = O(1/\beta^2)$ when $\beta \to 0$

Proof. Assume that we draw a sample profile $\Pi \in \mathfrak{S}_n^r$ from the product Mallows distribution $(\mathcal{M}_{\pi_0,\beta})^r$. Recall that for each pair of alternatives i, j, we let $q(i \succ j)$ be the number of rankings in profile Π for which i is placed before j. Let $\hat{\pi}$ the central ranking estimator using the pairwise statistics $q(i \succ j)$, defined in Eq.(5.1). We will upper bound the probability of the event $\hat{\pi}(\Pi) \neq \pi_0$. Without loss of generality, assume that $\pi_0 \equiv \pi_{id}$.

For $\Pi \sim (\mathcal{M}_{\pi_0,\beta})^r$, it holds that:

$$\Pr_{\Pi}[\hat{\pi}(\Pi) \neq \pi_0] \leq \Pr[\exists i < j : q(i \succ j) \leq q(j \succ i)] \leq \sum_{i<j} \Pr[q(i \succ j) \leq q(j \succ i)]$$

where the second inequality follows from the union bound.

The value of $\Pr[q(i \succ j) \leq q(j \succ i)]$ depends on the probabilities of removal of each subset of $[n]$ from a sample. However, since the probability of swapping two alternatives i and j in a Mallows sample is maximized when the alternatives are adjacent in the corresponding central ranking, taking the value $\frac{e^{-\beta}}{1+e^{-\beta}}$, if we set $X_\ell \sim \text{Be}(\frac{e^{-\beta}}{1+e^{-\beta}}), \forall \ell \in [r]$ and $Y_\ell = 1 - X_\ell$, then we get:

$$\Pr[q(i \succ j) \leq q(j \succ i)] \leq \Pr\left[\sum_{\ell \in [r]} (X_\ell - Y_\ell) \geq 0\right] =$$

$$\Pr\left[\frac{1}{r} \sum_{\ell \in [r]} (X_\ell - Y_\ell) - \frac{e^{-\beta} - 1}{1 + e^{-\beta}} \geq \frac{1 - e^{-\beta}}{1 + e^{-\beta}}\right]$$

Using Hoeffding's inequality, we get:

$$\Pr[q(i \succ j) \leq q(j \succ i)] \leq \exp\left(-2r \left(\frac{1 - e^{-\beta}}{1 + e^{-\beta}}\right)^2\right)$$

For simplicity, let $\zeta \coloneqq (\frac{1-e^{-\beta}}{1+e^{-\beta}})^2$. Therefore:

$$\Pr_{\Pi}[\hat{\pi}(\Pi) \neq \pi_0] \leq n^2 \exp(-2r\zeta)$$

Demanding $n^2 \exp(-2r\zeta) \leq \epsilon$ and solving for r, the result follows. □

Lower Bound. It turns out that the upper bound of the sample complexity that Theorem 5.1.1 establishes is tight. This is associated with the fact that in order to learn the central ranking, one has to learn the order of $n/2$ adjacent pairs in the central ranking and in order to do that with high probability, lagarithmic to the number of alternatives samples are needed.

Theorem 5.1.2: Caragiannis et al. [2013]

For any $\epsilon \in (0, 1/2]$ and any central ranking estimator, there exists a central ranking $\pi_0 \in \mathfrak{S}_n$ such that, for any $\beta > 0$, the estimator, given a sample profile drawn from $(\mathcal{M}_{\pi_0,\beta})^r$, retrieves π_0 with probability at least $1-\epsilon$, only if $r = \Omega(\frac{1}{\beta}\log(n/\epsilon))$.

Proof. Let $\tilde{\pi}$ be any (possibly randomized) estimator of π_0. Assume that:

$$\Pr_{\Pi\sim(\mathcal{M}_{\pi,\beta})^r}[\tilde{\pi}(\Pi) = \pi] \geq 1-\epsilon, \forall \pi \in \mathfrak{S}_n$$

Fix $\sigma \in \mathfrak{S}_n$. Then, we have:

$$\Pr_{\Pi\sim(\mathcal{M}_{\sigma,\beta})^r}[\tilde{\pi}(\Pi) = \sigma] = \sum_{\Pi\in\mathfrak{S}_n^r} \Pr[\Pi|\sigma]\Pr[\tilde{\pi}(\Pi) = \sigma]$$

Let $\mathcal{N}(\sigma) = \{\pi \in \mathfrak{S}_n : d_{KT}(\sigma,\pi) = 1\}$ ($|\mathcal{N}(\sigma)| = n-1$). Observe that for any $\Pi \in \mathfrak{S}_n^r$, it holds:

$$\sum_{\pi\in\mathcal{N}(\sigma)} \Pr[\tilde{\pi}(\Pi) = \pi] \leq 1 \tag{5.2}$$

Also, it is true that for any $\pi \in \mathcal{N}(\sigma)$: $\Pr[\Pi|\sigma] \geq e^{-\beta r}\Pr[\Pi|\pi]$, from triangle inequality.

Therefore, if we multiply the parts of Eq.(5.2) with $\Pr[\Pi|\sigma]$ and sum over all possible Π, we get that:

$$1 - \epsilon + (1-\epsilon)(n-1)e^{-\beta r} \leq 1$$

Solving for r, the result follows.

□

Therefore, the sample complexity of learning the central ranking with probability at least $1-\epsilon$, under the Mallows model is settled to $\Theta(\text{poly}(\frac{1}{\beta})\log(n/\epsilon))$. However, it is interesting to point out that the corresponding number of pairwise comparisons is $\Theta(\text{poly}(\frac{1}{\beta})n^2\log(n/\epsilon))$, since each sample contains n^2 comparisons. That is, the Mallows model is "too rigid": we cannot choose the pairwise comparisons therefore some of them are wasted, in the sense that even without knowing them, we could retrieve π_0 with high

probability. Also, observe that for proving the upper bound we did not need to use the concentration property of Mallows model: two elements that are for example $\Omega(\log(n))$ positions away in π_0 are very unlikely to be swapped. We identify two properties here:

1. *Flexibility:* This property corresponds to the ability to choose which pairwise comparisons will be made. In Mallows model, there is now flexibility, while in Noisy Comparisons model there is complete flexibility. In Selective Mallows model and specifically, in the model $\mathcal{M}^{\mathrm{ADP}(m)}_{\pi_0,\beta}$, the flexibility depends on m: when m is high, there is less flexibility and reversely.
2. *Concentration:* This property corresponds to the reduced uncertainty of the apparent ordering of a pair of alternatives in a sample. In Mallows model, two alternatives that appear many places away in a sample are more likely in the correct order. In Noisy Comparisons model, each pair of alternatives has the same probability of swap. In Selective Mallows model $\mathcal{M}^{\mathrm{ADP}(m)}_{\pi_0,\beta}$, the concentration property is more intense when m is high.

There is a trade-off between flexibility and concentration, which Selective Mallows model is trying to take advantage of, since complete lack of flexibility renders Mallows model incapable of exploiting its concentration property.

5.2 Learning from Noisy Comparisons

In this section we will briefly discuss in high level the problem of retrieving a hidden ranking under the Noisy Comparisons model. The problem can be thought of as the problem of sorting a list with a comparative algorithm, where each comparison has some probability of failure. Studied among other similar problems, this particular problem was solved by Feige et al. [1994]. In our presentation, we ignore any other parameter except the number of alternatives in the central ranking, $n \in \mathbb{N}$.

A trivial lower bound for the sample complexity comes from the lower bound of comparisons in order to sort n elements: $\Omega(n \log(n))$. According to the following theorem, surprisingly, this bound is tight with respect to n (of course there is some dependence on the parameter of error but the remarkable result is that an initially expected logarithmic to n blow up, needed for ensuring with high probability that every pair of elements whose order we assume known is in fact correctly ordered is avoided).

Informal theorem 5.2.1: Feige et al. [1994]

The number of samples needed in order to learn the central ranking with probability at least $1-\epsilon$, under the Noisy Comparisons model is $O(n \log(n/\epsilon))$.

Proof (Sketch). The proof is based on the following observation: Noisy binary search (that is binary search under noisy comparisons) can be viewed as a walk on the binary search tree, where each node corresponds to an interval. Then, each time the search reaches a node, it can compare the element under search with the interval's limits and if it turns out that it does not belong in the interval, the search moves to the father of the node. Interestingly, the number of steps until noisy binary search is completed is $O(\log(n))$, since the errors cancel out.

In order to take advantage of this property, the algorithm is separated in three parts:

1. Pick $O(n/\log(n))$ elements at random and sort them using any algorithm of sorting that uses $O(N\log(N))$ comparisons in the noiseless case (N is the size of input), repeat each comparison $\log(N)$ times and take the majority order for the pair, in order to ensure that with high probability each pair is ordered correctly. Since $N = n/\log(n)$, the number of comparisons in this step is $O(n\log(n))$.

2. This step is the core of the algorithm. Consider each interval between two consecutive elements of the previous step a bucket and put each of the $O(n)$ other elements in the correct bucket using noisy binary search. This requires also $O(n\log(n))$ comparisons.

3. It remains to sort each bucket. Since the pivot elements of the first step are randomly chosen, it holds with high probability that the size of each bucket is $O(\log^2(n))$. With this observation, and a careful handling, this step also makes $O(n\log(n))$ comparisons.

□

The reason we presented these results is to underline the difference in the query complexity of Mallows model and Noisy Comparisons model. While in Mallows model the query complexity is $\Theta(n^2\log(n))$, in Noisy Comparisons model it is $\Theta(n\log(n))$. Therefore, there is a gap of order $\Theta(n)$. This is justified by the lack of flexibility in the Mallows model, as we described it above. In Selective Mallows model $\mathcal{M}^{\text{ADP}(m)}_{\pi_0,\beta}$ the query complexity would at first glance be expected to be $\Theta(nm\log(n))$ (although it might not be true): In each of the corner cases ($m = 2$ and $m = n$) this assumption seems accurate. In fact, we can, informally, provide an upper bound of order $O(nm\log(n))$:

> ***Informal corollary 5.2.1:*** *The sample complexity of retrieving the central ranking with probability at least $1-\epsilon$, under the adaptive Selective Mallows model $\mathcal{M}^{\text{ADP}(m)}_{\pi_0,\beta}$, where $\pi_0 \in \mathfrak{S}_n$, $\beta > 0$ and $m \leq n$ is: $O(\frac{n}{m}\log(n/\epsilon))$*

Proof (Sketch). Observe that every ranking of size m contains $m/2$ disjoint pairs (with no common elements). Also, the algorithm described in Theorem 5.2.1 can be executed in $\log n$ parallel time (with n processors), as shown in Feige et al. [1994]. Therefore, we can group the comparisons in batches of size $\Theta(m)$, each of which corresponds to a Selective Mallows sample, picking the appropriate selection set. □

Therefore, the corresponding query complexity is $O(nm\log(n))$, since each sample contains $\Theta(m^2)$ comparisons. The interesting open question is whether this bound is tight for any value of m. The motivation behind this question is that there are two possibilities:

1. The flexibility of adaptive Selective Mallows model can be exploited more efficiently: for the upper bound we informally provided, we used only $\Theta(m)$ of the $\Theta(m^2)$ comparisons in each sample.

2. The concentration of adaptive Selective Mallows model might be somehow exploited.

5.3 Learning under selective Mallows model

In this section we provide our own original results on sample complexity of learning under the selective Mallows model. Consider the problem of retrieving the central ranking from Selective Mallows samples, where the selection sets cannot be picked, but come from either an adversary, or a random procedure. In particular, the applications of this setting are the following:

1. In the case where we consider an adversary, we essentially search for a parameter that ensures that we can retrieve the central ranking with optimal sample complexity. The way we think it is as if an adversary picks the selection sets in order to deprive us from the ability to retrieve the central ranking. However, we restrict their behavior by selecting a parameter $p \in (0, 1)$ which is interpreted as follows: The selection sets vector picked by the adversary has to include each pair of alternatives in at least pr sets, where r is the size of the sample (and of the selection sets vector). Recall that this model is denoted with $\mathcal{M}^{\text{ADV}}_{\pi_0,\beta}$.

2. The case where the selection sets are created by a random process ($\mathcal{M}^{\text{RND}(\mathcal{D})}_{\pi_0,\beta}$) corresponds to applications like voting. For example, we might consider the problem of inferring the optimal ordering of n movies according to a population where each individual has seen some of the movies and therefore can only rank them. The movies that somebody has seen is assumed to be an independent random variable drawn from a selective distribution $\mathcal{D}$ over $2^{[n]}$. Also, each individual ranks the set $S \subseteq [n]$ of movies they have watched according to the selective Mallows distribution $\mathcal{M}^{S}_{\pi_0,\beta}$ (that is the population is characterized by the ignorance bias), where $\pi_0 \in \mathfrak{S}_n$ and $\beta > 0$. In this case, the parameter of interest is called *selectivity* of the distribution $\mathcal{D}$ and is a value $p \in (0, 1)$ equal to:

$$p = \min_{i<j} \mathcal{D}(i, j \text{ are both selected}) \tag{5.3}$$

The results for each case are summarized in Table 5.1.

Model	Sample Complexity		
$\mathcal{M}_{\pi_0,\beta}$	$O(\frac{1}{\beta^2}\log(n/\epsilon))$	$\Omega(\frac{1}{\beta}\log(n/\epsilon))$	$\Theta(\log(n/\epsilon))$
$\mathcal{M}^{\text{ADV}}_{\pi_0,\beta}$	$O(\frac{1}{\beta^2 p}\log(n/\epsilon))$	$\Omega(\frac{1}{\beta p}\log(n/\epsilon))$	$\Theta(\frac{1}{p}\log(n/\epsilon))$
$\mathcal{M}^{\text{RND}(\mathcal{D})}_{\pi_0,\beta}$	$O(\frac{1}{\beta^2 p}\log(n/\epsilon))$	$\Omega((\frac{1}{\beta}+\frac{1}{p})\log(n/\epsilon))$	$\Theta(\frac{1}{p}\log(n/\epsilon))$

Table 5.1: Each of the examined cases for the Selective Mallows model and the corresponding sample complexity. The upper bounds refer to the performance of the positional estimator (and the dependence on β refers to the case when $\beta \to 0$). The lower bounds provide qualitative information about each problem. We also illustrate that the bounds are tight, when the spread parameter is considered constant.

The interesting aspect of these results is that the positional estimator $\hat{\pi}$ (defined in Eq.(5.1)) is optimal for each of the considered cases. This is only true for the Selective Mallows model, since, for example, considering some model that corresponds to projecting a Mallows sample on a subset of $[n]$, we can understand that there is more information contained in the samples. In particular, even if an alternative i is not selected in any sample, due to the fact that being in between other elements influences the probability of their

swap, its position could be estimated. In our model this is not possible and we conclude that: All the valuable information included in a Selective Mallows sample is the relative ordering of between the elements of the corresponding selection set. The key observation is that if: $S \subseteq [n]$ such that $i \notin S \vee j \notin S$ for some $i, j \in [n], i \neq j$, which are adjacent in $\sigma \in \mathfrak{S}_n$, $\beta > 0$ and $\pi \in \mathfrak{S}_S$, then:

$$\Pr[\pi|\sigma, \beta, S] = \Pr[\pi|\sigma_{i \leftrightarrow j}, \beta, S] \tag{5.4}$$

In other words, when a pair of adjacent alternatives does not appear, no information about its order is provided. This is not the case, however, for other generalizations of Mallows model on reduced size rankings.

5.3.1 Learning from adversarially incomplete rankings

In this section we establish asymptotically tight sample complexity bounds for the adversarial Selective Mallows model ($\mathcal{M}^{\text{ADV}}_{\pi_0,\beta}$). We call an adversarial Mallows sample profile $p-$frequent if any pair of alternatives is contained in at least a $p-$fraction of the samples.

Upper Bound. We first provide an upper bound for the sample complexity.

Theorem 5.3.1

Let $\mathcal{M}_{\pi_0,\beta}$ be a Mallows distribution with central ranking $\pi_0 \in \mathfrak{S}_n$ and spread parameter $\beta > 0$. For any $\epsilon > 0$, there exists an algorithm that, given an p-frequent adversarial Mallows profile induced by $\mathcal{M}_{\pi_0,\beta}$ of size r which is at least equal to some $O(\frac{1}{(1-e^{-\beta})^2 p} \log(n/\epsilon))$ value, retrieves the central ranking π_0 with probability at least $1 - \epsilon$.

Proof. The estimator we use is the positional estimator $\hat{\pi}$. The proof is almost the same as that of Theorem 5.1.1. The difference is that the guarantee we have is that each pair appears in pr samples, instead of r. Which gives the wanted result. □

Lower Bound. Although the lower bound of sample complexity can be established by picking a "difficult" $p-$frequent selection sets vector, as for example one in which each pair of alternatives appears in at most $2p-$fraction of the samples and repeating the proof of Theorem 5.1.2, we derive it from a more general result that we prove in the following lemma:

Lemma 5.3.1

Let $\mathcal{S}$ be any fixed[a] selection sets profile. Then, for any $\epsilon \in (0, 1/2]$ and any estimator of the central ranking, there exists $\pi_0 \in \mathfrak{S}_n$ such that, for any $\beta > 0$, the estimator retrieves π_0 with probability at least $1 - \epsilon$ from a sample profile drawn from $\mathcal{M}^{\mathcal{S}}_{\pi_0,\beta}$ only if $\mathcal{S}$ includes $\Omega(\frac{n^2}{\beta} \log(n/\epsilon))$ comparisons between pairs of alternatives.

[a] It is important to point out that $\mathcal{S}$ is fixed because if it is selected randomly or during the runtime of the algorithm of estimation, the proof techniques we use do not work: We cannot find a ranking that is "difficult" in each execution of the algorithm.

Proof. Let $\tilde{\pi}$ be any estimator of π_0. Fix $r \in \mathbb{N}$. Let $\mathcal{S} = (S_1, S_2, \ldots, S_r) \in (2^{[n]})^r$. For every pair of alternatives $i, j \in [n]$, let $W_{ij}(\mathcal{S})$ be the number of sets of $\mathcal{S}$ where both i and j appear. Assume, for simplicity that $n/4 \in \mathbb{N}$.

It suffices to show that if $\sum_{i<j} W_{ij}(\mathcal{S}) < \frac{n^2}{8\beta} \log(n(1-\epsilon)/(4\epsilon))$, then there exists a ranking $\pi_0 \in \mathfrak{S}_n$ which cannot be retrieved by $\tilde{\pi}$ with probability at least $1-\epsilon$.

There is a family $\{P_k\}_{k\in[n/2]}$ of perfect matchings of the set of alternatives, that does not contain any pair of alternatives twice: $(i,j) \in P_k \Rightarrow (i,j) \notin P_{k'}, \forall k, k' \in [n/2], k \neq k'$. We can construct such a family inductively by picking $P_1 = \{(1,2),(3,4),\ldots,(n-1,n)\}$, $P_2 = \{(1,4),(3,6),\ldots,(n-1,2)\}$ and $P_{n/2} = \{(1,n),(3,2),\ldots,(n-1,n-2)\}$.

Observe that $\sum_{k\in[n/2]} \sum_{(i,j)\in P_k} W_{ij}(\mathcal{S}) \leq \sum_{i<j} W_{ij}(\mathcal{S})$, since we skip pairs of the same parity. Assuming that $\sum_{i<j} W_{ij}(\mathcal{S}) < \frac{n^2}{8\beta} \log(n(1-\epsilon)/(4\epsilon))$, there exists $k \in [n/2]$ such that: $\sum_{\{i,j\}\in P_k} W_{ij}(\mathcal{S}) < \frac{n}{4\beta} \log(n(1-\epsilon)/(4\epsilon))$. Since $|P_k| = n/2$, there exist at least $n/4$ pairs of alternatives $(i,j) \in P_k$ for which:

$$W_{ij}(\mathcal{S}) < \frac{1}{\beta} \log(n(1-\epsilon)/(4\epsilon)) \tag{5.5}$$

We will show that, for such $\mathcal{S}$, it cannot be the case that: $\Pr_{\Pi\sim\mathcal{M}^{\mathcal{S}}_{\pi_0,\beta}}[\tilde{\pi}(\Pi) = \pi] \geq 1-\epsilon, \forall \pi \in \mathfrak{S}_n$. Without loss of generality, assume that $P_k = P_1$ and denote $P = P_1$ for simplicity. Let $R \subset \mathfrak{S}_n$ such that for any $\pi \in R$ the alternatives $\{1,2\}$ are placed in positions $\{1,2\}$, the alternatives $\{3,4\}$ are placed in positions $\{3,4\}$ and so on: $\{\pi(2\ell-1), \pi(2\ell)\} = \{2\ell, 2\ell-1\}, \forall \ell \in [n/2]$. For example, if $n=4$, then $R = \{(1 \succ 2 \succ 3 \succ 4), (1 \succ 2 \succ 4 \succ 3), (2 \succ 1 \succ 3 \succ 4), (2 \succ 1 \succ 4 \succ 3)\}$.

Fix $\pi_0 \in R$. For any $\pi \in R$, let $D(\pi)$ be the set of pairs in P in which π, π_0 disagree: $D(\pi) = \{\{i,j\} \in P : (\pi(i) - \pi(j))(\pi_0(i) - \pi_0(j)) < 0\}$. Then, for any $\pi \in R$ and $\Pi \in \mathrm{Sym}(\mathcal{S}) = \times_{\ell=1}^{r} \mathrm{Sym}(S_\ell)$, the probability to observe Π conditional on the central ranking π_0 is at least:

$$\Pr[\Pi|\pi_0] \geq e^{-\beta \sum_{\{i,j\}\in D(\pi)} W_{ij}(\mathcal{S})} \Pr[\Pi|\pi] \tag{5.6}$$

The above is a consequence of the triangle inequality property of Kendall tau distance, applied inductively to the pairs in $D(\pi)$, since each pair in $D(\pi)$ is adjacent in π, π_0.

For a fixed sample profile $\Pi \in \mathrm{Sym}(\mathcal{S})$, it holds:

$$\sum_{\pi\in R} \Pr[\tilde{\pi}(\Pi) = \pi] \leq 1 \tag{5.7}$$

We multiply each term of Ineq. (5.7) with $\Pr[\Pi|\pi_0]$, apply Ineq. (5.6) and summing over $\Pi \in \mathrm{Sym}(\mathcal{S})$ to get:

$$\sum_{\pi\in R} e^{-\beta \sum_{\{i,j\}\in D(\pi)} W_{ij}(\mathcal{S})} \sum_{\Pi\in\mathrm{Sym}(\mathcal{S})} \Pr[\tilde{\pi}(\Pi) = \pi] \Pr[\Pi|\pi] \leq \sum_{\Pi\in\mathrm{Sym}(\mathcal{S})} \Pr[\Pi|\pi_0] = 1$$

Assume, for contradiction, that $\Pr_{\Pi\sim\mathcal{M}^{\mathcal{S}}_{\pi_0,\beta}}[\tilde{\pi} = \pi] \geq 1-\epsilon, \forall \pi \in R$. Then:

$$(1-\epsilon) \sum_{\pi\in R} e^{-\beta \sum_{\{i,j\}\in D(\pi)} W_{ij}(\mathcal{S})} \leq 1$$

However, from Ineq. (5.5), it turns out that: $\sum_{\pi\in R} e^{-\beta \sum_{\{i,j\} D(\pi)} W_{ij}(\mathcal{S})} > 1 + \frac{n}{4} \frac{4\epsilon}{n(1-\epsilon)} = 1 + \frac{1}{1-\epsilon}$. Therefore: $1 - \epsilon + \epsilon < 1$, contradiction. □

The lower sample complexity bound for the adversarial Mallows model follows.

Theorem 5.3.2

For any $p \in (0,1)$, there exists a p-frequent adversarial set profile $\mathcal{S}$ of size r, such that for any $\epsilon \in (0, 1/2]$ and any central ranking estimator, there exists a central ranking $\pi_0 \in \mathfrak{S}_n$ such that, for any $\beta > 0$, the estimator, given the corresponding adversarial sample profile drawn from $\mathcal{M}_{\pi_0,\beta}^{\text{ADV}\to\mathcal{S}}$, retrieves π_0 with probability at least $1-\epsilon$, only if $r = \Omega(\frac{1}{\beta p}\log(n/\epsilon))$.

Proof. Consider a selection sets profile $\mathcal{S}$ that is $p-$frequent and also, the number of pairwise comparisons between alternatives is no more than $2prn^2$ samples. We can pick such a profile as follows: $\mathcal{S}$ consists of pr complete sets and $(1-p)r$ sets of length $m \leq n\sqrt{p/(1-p)}$.

Then, the number of queries Q we make are: $Q \leq 2prn^2$. However, from Lemma 5.3.1, Q must be $\Omega(\frac{n^2}{\beta}\log(n/\epsilon))$. Therefore: $r = \Omega(\frac{1}{\beta p}\log(n/\epsilon))$.

□

Note that when $p \ll 1/n^2$, then, according to our analysis, many sets might have to be chosen to be empty. However, the interesting case is when $p = (m'/n)^2$, for some $m' \in [n]$. In this case, after choosing pr complete sets (in order to ensure $p-$frequency), we can arbitrarily pick $(1-p)r$ sets of length $m \leq m'$.

5.3.2 Learning from randomly incomplete rankings

In this section we provide more interesting results that consider the case that the selection sets are formed randomly, under some selection distribution $\mathcal{D}$ that is $p-$frequent, that is, it satisfies Eq.(5.3).

An interesting observation of the upcoming Theorems 5.3.3 and 5.3.4 is that incompleteness ($\propto p$) and noisiness ($\propto \beta$) affect the hardness of estimating the central ranking independently. In particular, the absence of one of these factors that influence the quality of the sample profile does not necessarily imply a collapse of the sample complexity. Although it is not clear from the upper bounds, when $\beta \to +\infty$, then the number of samples that are sufficient for the positional estimator to retrieve the central ranking tend to 1, for there are no swaps expected, when we consider the classic Mallows model and the Adversarial Selective Mallows model. However, in the random case, even if there is no swap, we have to draw enough samples to see enough pairs of alternatives.

Upper bound. The upper bound is again derived using the positional estimator $\hat{\pi}$.

Theorem 5.3.3

Let $\mathcal{M}_{\pi_0,\beta}$ be a Mallows distribution with central ranking $\pi_0 \in \mathfrak{S}_n$ and spread parameter $\beta > 0$. For any $\epsilon > 0$, there exists an algorithm that, given a p-frequent randomized Mallows profile induced by $\mathcal{M}_{\pi_0,\beta}$ of size r which is at least equal to some value in $O(\frac{1}{(1-e^{-\beta})^2 p}\log(n/\epsilon))$, retrieves the central ranking π_0 with probability at least $1-\epsilon$.

Proof. Assume that we draw a sample profile $\Pi \in \mathcal{L}^r$, where $\mathcal{L} = \cup_{S\subseteq[n]}\mathfrak{S}_S$, from the selective Mallows model $\mathcal{M}^{\text{RND}(\mathcal{D})}_{\pi_0,\beta}$. For each pair of alternatives i, j, we let $q(i \succ j)$ be the number of rankings in profile Π for which i is placed before j. Let $\hat{\pi}$ the central ranking estimator using the pairwise statistics $q(i \succ j)$, defined in Eq. (5.1). We will upper bound the probability of the event $\hat{\pi}(\Pi) \neq \pi_0$. Without loss of generality, assume that $\pi_0 \equiv id$.

For $\Pi \sim \mathcal{M}^{\text{RND}(\mathcal{D},r)}_{\pi_0,\beta}$, it holds that:

$$\Pr_\Pi[\hat{\pi}(\Pi) \neq \pi_0] \leq \Pr[\exists i < j : q(i \succ j) \leq q(j \succ i)] \leq \Pr[q(i \succ j) \leq q(j \succ i)]$$

where the second inequality follows from the union bound.

Consider the random variable $W_{i,j}$ that refers to the number of samples in the profile Π where both i and j are selected. By the law of total probability, the above sum its equal to:

$$\sum_{k=0}^{r} \Pr[W_{i,j} = k]\Pr[q(i \succ j) \leq q(j \succ i)|W_{i,j} = k] \tag{5.8}$$

The value of $\Pr[q(i \succ j) \leq q(j \succ i)|W_{i,j} = k]$ depends on the probabilities of removal of each subset of $[n]$ from a sample. However, since the probability of swapping two alternatives i and j in a Mallows sample is maximized when the alternatives are adjacent in the corresponding central ranking, taking the value $\frac{e^{-\beta}}{1+e^{-\beta}}$, if we set $X_\ell \sim \text{Be}(\frac{e^{-\beta}}{1+e^{-\beta}}), \forall \ell \in [k]$ and $Y_\ell = 1 - X_\ell$, then we get:

$$\Pr[q(i \succ j) \leq q(j \succ i)|W_{i,j} = k] \leq \Pr\left[\sum_{\ell\in[k]}(X_\ell - Y_\ell) \geq 0\right] =$$

$$\Pr\left[\frac{1}{k}\sum_{\ell\in[k]}(X_\ell - Y_\ell) - \frac{e^{-\beta}-1}{1+e^{-\beta}} \geq \frac{1-e^{-\beta}}{1+e^{-\beta}}\right]$$

Using Hoeffding's inequality, we get:

$$\Pr[q(i \succ j) \leq q(j \succ i)|W_{i,j} = k] \leq \exp(-2k(\frac{1-e^{-\beta}}{1+e^{-\beta}})^2)$$

For simplicity, let $\zeta := (\frac{1-e^{-\beta}}{1+e^{-\beta}})^2$. Therefore, returning to the sum of Eq.(5.8), it is sufficient to bound:

$$\mathbb{E}_{W_{i,j}}[\exp(-2\zeta W_{i,j})]$$

Let us denote the probability that both i and j are selected under the measure $\mathcal{D}$ by p_{ij}. Observe that $W_{i,j} \sim (r, p_{ij})$ and hence:

$$\mathbb{E}_{W_{i,j}}[\exp(-2\zeta)^{W_{i,j}}] = ((1-p_{ij})(1-\exp(-2\zeta)) + \exp(-2\zeta))^r \leq ((1-p) + p\exp(-2\zeta))^r$$

Consequently:

$$\Pr[\hat{\pi} \neq \pi_0] \leq \binom{n}{2}((1-p) + p\exp(-2\zeta))^r$$

Demanding $\binom{n}{2}((1-p) + p\exp(-2\zeta))^r \leq \epsilon$ and solving for r, the result follows. □

Lower bound. The next result derives the bound $\Omega((\frac{1}{\beta}+\frac{1}{p})\log(n/\epsilon))$ of the randomized setting under a p-selective distribution. The qualitative difference with the bound provided by Theorem 5.3.2 originates to the possibility of total absence of comparison between some pairs of alternatives in the sample profile. If, additionally, the alternatives are adjacent in the central ranking, then their order cannot be determined better than randomly.

Theorem 5.3.4

For any $p \in (0,1]$, there exists a p-frequent selection distribution $\mathcal{D}$, such that for any $\epsilon \in (0,1/2]$ and any central ranking estimator, there exists $\pi_0 \in \mathfrak{S}_n$ such that, for any $\beta > 0$, the estimator, given a sample profile drawn from $\mathcal{M}_{\pi_0,\beta}^{\mathrm{RND}(\mathcal{D},r)}$ of size r, retrieves π_0 with probability at least $1-\epsilon$, only if:

$$r = \Omega\left(\left(\frac{1}{\beta}+\frac{1-p}{p}\right)\log(n/\epsilon)\right)$$

Proof. Let $\delta = 1-p$. Let $\tilde{\pi}$ be any estimator of the hidden ranking. More specifically, $\tilde{\pi}$ inputs a sampling profile $\Pi \in \mathcal{L}^r$, where $\mathcal{L} = \cup_{S\subseteq[n]}\mathfrak{S}_S$, from the selective Mallows distribution $\mathcal{M}_{\pi_0,\beta}^{\mathrm{RND}(\mathcal{D})}$, with unknown central ranking $\pi_0 \in \mathfrak{S}_n$ which will be determined later, spread parameter $\beta > 0$ and selection distribution $\mathcal{D}$ and outputs an element of $\mathfrak{S}_n$ either deterministically or randomly.

The proof consists of two parts. In Part **I**, we get the term $\Omega(\frac{1}{\beta}\log(n/\epsilon))$, while in Part **II**, we get the term $\Omega(\frac{1}{\log(1/\delta)}\log(n/\epsilon)) = \Omega(\frac{1-p}{p}\log(n/\epsilon))$. Combining them, we obtain the desired lower bound.

Let $q \in (0,1)$. We consider $\mathcal{D}$ as follows: Every element $i \in [n]$ is selected independently with probability $1-q$. Set $\delta = 2q-q^2$. Obviously, $\mathcal{D}$ is $(1-\delta)-$selective. Furthermore, in every sequence of disjoint pairs, each pair is selected by $\mathcal{D}$ independently from the others, with probability $1-\delta = p$. Assume, for simplicity, that $n/2 \in \mathbb{N}$.

Part I. This is a trivial consequence of the lower bound for complete Mallows samples.

Part II. It remains to show that for the selective distribution $\mathcal{D}$ any estimator of the central ranking requires $\Omega(\frac{\log(n/\epsilon)}{\log(1/\delta)})$ samples.

We define the set R as follows:

$$R = \{\pi \in \mathfrak{S}_n : \{\pi(2i), \pi(2i-1)\} = \{2i, 2i-1\}, \forall i \in [n/2]\}.$$

The set R can be described as follows: Consider the transpositions $(1\ \ 2), (3\ \ 4), ..., (n-1\ \ n)$. Then, any composition of any number of these permutations with the identity belongs to R and, obviously, $|R| = 2^{n/2}$.

The idea is that selecting a random element of R to be the central ranking, the probability of success of $\tilde{\pi}$ cannot be higher than that of an estimator that randomly selects an element of R, due to the structure of the selection distribution $\mathcal{D}$.

Our goal is to compute the expected value of the probability $\Pr_{\Pi\sim\mathcal{M}_{\pi,\beta}^{\mathrm{RND}(\mathcal{D},r)}}[\tilde{\pi}(\Pi) = \pi]$ over the family $\{\pi \in R\}$. Afterwards, using the probabilistic method, we will get the desired bound.

$$\mathbb{E}_{\pi\sim\mathrm{Uni}(R)}\Big[\Pr_{\Pi}[\tilde{\pi}(\Pi)=\pi]\Big]=\sum_{\pi\in R}\Pr[\pi]\Pr_{\Pi\sim\mathcal{M}_{\pi,\beta}^{\mathrm{RND}(\mathcal{D},r)}}[\tilde{\pi}(\Pi)=\pi]=$$

$$\sum_{\pi\in R}\Pr[\pi]\sum_{\Pi\in\mathcal{L}^r}\Pr[\Pi|\pi]\Pr[\tilde{\pi}(\Pi)=\pi]$$

Since we choose uniformly at random from the class R and since the sums are finitely many:

$$\mathbb{E}_{\pi\sim\mathrm{Uni}(R)}\Big[\Pr_{\Pi}[\tilde{\pi}(\Pi)=\pi]\Big]=2^{-n/2}\sum_{\Pi\in\mathcal{L}^r}\sum_{\pi\in R}\Pr[\Pi|\pi]\Pr[\tilde{\pi}(\Pi)=\pi]$$

We partition the sum $\sum_{\Pi\in\mathcal{L}^r}$ as follows:

1. Let $T=\{(1,2),(3,4),...,(n-1,n)\}$.
2. Select $t\in\{0,1,...,n/2\}$.
3. Select t distinct pairs $p_1,...,p_t$ from the set T.
4. Create the set of profiles $\mathcal{L}^r(p_1,...,p_t)$, where for each $\Pi\in\mathcal{L}^r(p_1,...,p_t)$, the only pairs that are never observed are the chosen $p_1,...,p_t$.

Hence, the sum $\sum_{\Pi\in\mathcal{L}^r}$ is equivalent to the following

$$\sum_{t=0}^{n/2}\sum_{p_1,...,p_t}\sum_{\Pi\in\mathcal{L}^r(p_1,...,p_t)}$$

Let us fix such a collection of pairs that never appear in a profile $P_t=(p_1,...,p_t)$. Now, we can partition the set of permutations R with respect to that fixed collection in the following steps:

1. Fistly, there are $\frac{n}{2}-t$ pairs from T not chosen. Let $Q_t=Q(P_t)=Q(p_1,...,p_t)=T\setminus\{p_1,...,p_t\}$.
2. Viewing each such pair $q\in Q_t$ as a transposition, there are $2^{\frac{n}{2}-t}$ possible choices to create an ordering among these pairs (for each such $q=(i,i+1)$, we either let $(i\ \ i+1)$ or $(i+1\ \ i)$).
3. Denote $\mathrm{Sym}((i,i+1))=\{(i,i+1),(i+1,i)\}$. More generally, $\mathrm{Sym}((p_1,...,p_t))=\times_{i=1}^{t}\mathrm{Sym}(p_i)$.
4. For $\tau\in\mathrm{Sym}((p_1,...,p_t))$, the set $[R:\tau]$ is the 'stabilizer' group of τ, that is all other transpositions can be permuted and the cycles of τ are fixed.

We then have:

$$\mathbb{E}_{\pi\sim\mathrm{Uni}(R)}\Big[\Pr_{\Pi}[\tilde{\pi}(\Pi)=\pi]\Big]=$$

$$= 2^{-n/2}\Big\{\sum_{t=0}^{n/2}\sum_{P_t}\sum_{\Pi\in\mathcal{L}^r(P_t)}\Big\}\Big\{\sum_{\tau\in\mathrm{Sym}(Q(P_t))}\sum_{\pi\in[R:\tau]}\Big\}\Pr[\Pi|\pi]\Pr[\tilde{\pi}(\Pi)=\pi]$$

$$= 2^{-n/2}\sum_{t=0}^{n/2}\sum_{P_t}\sum_{\tau\in\mathrm{Sym}(Q(P_t))}\sum_{\Pi\in\mathcal{L}^r(P_t)}\sum_{\pi\in[R:\tau]}\Pr[\Pi|\pi]\Pr[\tilde{\pi}(\Pi)=\pi] \tag{5.9}$$

Let $\tau\in\mathrm{Sym}(Q(P_t))$. Now, observe that, if $\pi,\sigma\in[R:\tau]$ and $\Pi\in\mathcal{L}^r(P_t)$, we have that:

$$\Pr[\Pi|\pi]=\Pr[\Pi|\sigma]$$

Also, it holds that $\sum_{\pi\in R'}\Pr[\tilde{\pi}(\Pi)=\pi]\leq 1$, for any fixed $\Pi\in\mathcal{L}^r(P_t)$. Therefore, by the above observation and for any $\pi'\in[R:\tau]$:

$$\sum_{\pi\in[R:\tau]}\Pr[\Pi|\pi]\Pr[\tilde{\pi}(\Pi)=\pi]\leq\Pr[\Pi|\pi']$$

Summing over all possible $\Pi\in\mathcal{L}^r(P_t)$, we get that:

$$\sum_{\Pi\in\mathcal{L}^r(P_t)}\sum_{\pi\in[R:\tau]}\Pr[\Pi|\pi]\Pr[\tilde{\pi}(\Pi)=\pi]\leq\Pr_{\Pi\sim\mathcal{M}^{\mathrm{RND}(\mathcal{D},r)}_{\pi',\beta}}[\Pi\in\mathcal{L}^r(P_t)]. \tag{5.10}$$

Observe now that $\Pr_{\Pi\sim\mathcal{M}^{\mathrm{RND}(\mathcal{D},r)}_{\pi',\beta}}[\Pi\in\mathcal{L}^r(P_t)]$ is independent from the selection of π', since the probability of the event that $\Pi\in\mathcal{L}^r(P_t)$ is determined from the distribution $\mathcal{D}$. Therefore, we get that for any $\pi\in R$:

$$\mathbb{E}_{\pi\sim\mathrm{Uni}(R)}\Big[\Pr_{\Pi}[\tilde{\pi}(\Pi)=\pi]\Big]\leq 2^{-n/2}\sum_{t=0}^{n/2}\sum_{P_t}\sum_{\tau\in\mathrm{Sym}(Q(P_t))}\Pr_{\Pi\sim\mathcal{M}^{\mathrm{RND}(\mathcal{D},r)}_{\pi,\beta}}[\Pi\in\mathcal{L}^r(P_t)]=$$

$$=\sum_{t=0}^{n/2}2^{-t}\sum_{P_t}\Pr_{\Pi\sim\mathcal{M}^{\mathrm{RND}(\mathcal{D},r)}_{\pi,\beta}}[\Pi\in\mathcal{L}^r(P_t)]$$

Let X be a random variable that equals to the number of elements of T that do not appear together in any of r independent samples of $\mathcal{D}$. Due to the structure of $\mathcal{D}$: $X\sim(n/2,\delta^r)$. Observe that: $\Pr[X=t]=\sum_{P_t}\Pr_{\Pi\sim\mathcal{M}^{\mathrm{RND}(\mathcal{D},r)}_{\pi,\beta}}[\Pi\in\mathcal{L}^r(P_t)]$. Therefore:

$$\mathbb{E}_{\pi\sim\mathrm{Uni}(R)}\Big[\Pr_{\Pi}[\tilde{\pi}(\Pi)=\pi]\Big]\leq\sum_{t=0}^{n/2}2^{-t}\Pr[X=t]=\mathbb{E}[2^{-X}]=(1-\delta^r+\delta^r/2)^{n/2}\leq\frac{1}{1+n\delta^r/4}$$

From the above, we conclude that there exists $\pi\in R$ such that: $\Pr_{\Pi}[\tilde{\pi}=\pi]\leq\frac{1}{1+n\delta^r/4}$. Assuming that $\Pr_{\Pi}[\tilde{\pi}=\pi]\geq 1-\epsilon$ and solving for r, we get that: $r=\Omega(\log(n/\epsilon)/\log(1/\delta))$.

□

In order to understand the implications of the bounds we established, we provide the following example:

Example. Consider the case when each sample contains m alternatives on average. Assume also that all the alternatives are selected independently with the same probability q. Then, $q = m/n$. Each pair of alternatives is selected with probability $p = q^2 = (m/n)^2$. Therefore, from Theorem 5.3.4, we need:

$$\Omega\left(\frac{n^2}{m^2}\log(n/\epsilon)\right) \text{ samples from } \mathcal{M}_{\pi_0,\beta}^{\mathrm{RND}(\mathcal{D})},$$

where $\mathcal{D}$ is the distribution we described, in order to learn π_0 with probability at least $1-\epsilon$. Therefore, since each sample contains $\Theta(m^2)$ pairwise comparisons on average, our result indicates that even in the randomized case, if the alternatives are selected independently and with the same probability, then the query complexity lower bound we get is similar to the one for the case of fixed selection sets (Lemma 5.3.1).

Chapter 6

Mallows reconstruction problems

In this chapter, we consider the reconstruction problems that we defined in Definitions 4.2.1, 4.2.2, 4.2.3 and 4.2.4, which correspond to the problems of finding the maximum likelihood, or a likelier that nature (likelier than the central ranking) estimation of the central ranking under Mallows or Selective Mallows model. The MRP and MAX-MRP problems' solutions that we present were introduced by Braverman and Mossel [2009]. Building on this work, we provide efficient algorithms for solving the SMRP and MAX-SMRP problems.

The problem of finding a ranking that minimizes the total Kendall tau distance from each of at least 4 rankings is known to be NP-hard. However, the Mallows reconstruction problems correspond to an average case analysis of the rank aggregation problem, in which it can be solved efficiently. Recall that the rank aggregation problem corresponds to finding a ranking $\pi^* \in \mathfrak{S}_n$, given a set of r rankings $\pi_1, \pi_2, \ldots, \pi_r \in \mathfrak{S}_n$, such that:

$$\pi^* = \arg\min_{\pi \in \mathfrak{S}_n} \sum_{\ell \in [r]} d_{KT}(\pi, \pi_\ell)$$

The key observation to solve the Mallows reconstruction problems in the average case is that under Mallows model, the alternatives' positions are concentrated: the probability of displacement of an alternative from its position in the central ranking diminishes exponentially to the length of the displacement. Therefore, instead of searching for the maximum likelihood estimation exhaustively in $\mathfrak{S}_n$, we could restrict our search in a local subspace of $\mathfrak{S}_n$ consisting of rankings that rank each alternative close to its original location (in the central ranking). However, since the central ranking is unknown, we first have to specify the subspace of interest. This can be done efficiently, since the samples are expected to be concentrated with respect to alternatives' positions, therefore they can provide the information required to construct an approximation ranking of the central ranking, in which each alternative is with high probability close to its original position.

We next describe the algorithm presented by Braverman and Mossel [2009] in high level, before providing a more precise analysis and generalizing it to solve SMRP and MAX-SMRP. The structure of our algorithm is similar to the one for solving MRP ans MAX-MRP, but with some non trivial generalizations. In Table 6.1, we present the time complexity of the algorithms in each case.

Observe that as the number of samples grows, the time tends to become linear for MRP and

Problem	Time complexity
MRP	$O(n^{1+O(\frac{2+\alpha}{\beta r})} \cdot \log^2 n)$
MAX-MRP	$O(n^{1+O(\frac{2+\alpha}{\beta r})} \cdot 2^{O(\frac{\alpha}{\beta}+\frac{1}{\beta^2})} \cdot \log^2 n)$
SMRP	$O(n^2 + n^{1+O(\frac{2+\alpha}{\beta r p^2})} 2^{O(\frac{1}{(\beta p)^2})} \log^2 n)$
MAX-SMRP	$O(n^2 + n^{1+O(\frac{2+\alpha}{\beta r p^4})} 2^{O(\frac{1}{\beta^2 p^4})} \log^2 n)$

Table 6.1: Time complexity of solving Mallows reconstruction problems with probability of failure bounded above by $n^{-\alpha}$, where $\alpha > 0$.

MAX-MRP and quadratic for SMRP and MAX-SMRP. This corresponds to the increase in the quality of the initial estimator of the central ranking, due to the increase in the number of samples. Also, in the SMRP and MAX-MRP, recall that pr is the minimum number of appearances of a pair in the samples, where $p \in (0, 1]$. At first glance, one could suppose that in the selective case pr would substitute r in the expression of time complexity of the corresponding classic case. However there is a catch: The concentration in the alternatives' positions is relaxed in the selective case. In particular, even two elements $i, j \in [n]$ that are distant in the central ranking (that is they are $\Omega(\log n)$ positions away), may swap easily in a sample where the elements that were ranked between them in the central ranking are absent, which coincides with what we called ignorance bias.

6.1 Algorithm description

In each case, the algorithm consists of two phases. In the first phase, we create an estimation of the central ranking, for which, with high probability, each alternative is ranked within a small margin of its original position. In the second phase, we search locally with the use of Dynamic Programming in order to find a ranking that is at least as likely as the central ranking. In the case that the subspace of $\mathfrak{S}_n$ includes a maximum likelihood ranking, then we find it. The two phases are described below with more detail.

Initialization. In this step, we create a polynomial time estimator $\tilde{\pi}$ of the central ranking π_0, taking advantage of the concentration that Mallows (and selective Mallows samples display), which, with high probability satisfies (minus the details) the following property:

$$|\tilde{\pi}(i) - \pi_0(i)| = O(\log n), \forall i \in [n]$$

In the classic Mallows model, we use the average estimator $\overline{\pi}$, which ranks each alternative according to its average position in the samples, breaking ties uniformly. However, in the Selective Mallows case, the average estimator does not work, since each ranking contains, in general, different sets of alternatives, which might also be of different lengths. Hence, we use the positional estimator $\hat{\pi}$, defined in Eq.(5.1). One problem that the positional estimator has to face is the ignorance bias, but we will discuss more on this in a following section. Observe also that the positional estimator $\hat{\pi}$ is computed in time $O(n^2)$, while the average estimator in just $O(n)$, which is why it was prefered in the classic Mallows case.

Local Search. After creating the initial estimator $\tilde{\pi}$, we have to implement a local search on the subspace of $\mathfrak{S}_n$, which corresponds to the set of rankings $\pi \in \mathfrak{S}_n$ for which (minus the details):

$$|\tilde{\pi}(i) - \pi(i)| = O(\log n), \forall i \in [n],$$

for we already know that π_0 lies within this subspace. Since the maximum likelihood estimation π^* can be proven to satisfy the proximity property with high probability:

$$|\pi^*(i) - \pi_0(i)| = O(\log n), \forall i \in [n],$$

adjusting the precise size of the subspace of search, we manage to solve the maximum reconstruction problems.

The only question remaining unanswered is how the local search will be implemented. The answer was given by Braverman and Mossel [2009], in the following lemma, which we use without proof. It uses Dynamic Programming techniques to "sort an almost sorted list".

Lemma 6.1.1: Braverman and Mossel [2009]

Let $f : [n] \times [n] \to \mathbb{N}$ be a scoring function. Supposing that there exists an optimal ordering $\pi \in \mathfrak{S}_n$ that maximizes the score:

$$s(\pi) = \sum_{i \succ_\pi j} f(i, j),$$

such that $|\pi(i) - i| \leq k, \forall i \in [n]$, then π can be computed in time $O(n \cdot k^2 \cdot 2^{6k})$.

Recall that the score in each of the Mallows reconstruction problems we examine is given by Eq.(4.11). Combining Eq.(4.11) with Eq.(4.2), we get that, equivalently, we can consider maximizing the score:

$$s(\pi) = \sum_{i \succ_\pi j} q(i \succ j),$$

which satisfies the conditions for applying Lemma 6.1.1.

6.2 Solving MAX-MRP

In this section we present the algorithm for solving MAX-MRP (and MRP), as proposed by Braverman and Mossel [2009].

Initialization. The initial estimator of π_0 that we use in this case is the average estimator $\overline{\pi}$:

$$\overline{\pi}(i) = \frac{1}{r} \sum_{\ell \in [r]} \pi_\ell(i), \forall i \in [n] \tag{6.1}$$

Clearly, $\overline{\pi}$ is not necessarily an element of $\mathfrak{S}_n$. However, it can be transformed to a ranking by putting the elements in increasing $\overline{\pi}-$order, breaking ties uniformly.

The following lemma is the basis of our analysis. It states that the probability of displacement of an alternative from its original position decays exponentially to the length of the displacement.

Lemma 6.2.1

Let $\pi \sim \mathcal{M}_{\pi_0,\beta}$, where $\pi_0 \in \mathfrak{S}_n$, $\beta > 0$. Then:

$$Pr[|\pi(i) - \pi_0(i)| \geq K] < 2e^{-\beta K}/(1 - e^{-\beta}), \forall i \in [n]$$

Proof. Assume for simplicity that $\pi_0 = \pi_{id}$. According to Algorithm 1, when i is inserted, the probability that i is ranked in position j, $p_i(j)$, is given by Eq.(4.5). In particular, if $j = i - k$, then: $p_i(j) \leq e^{-\beta k}$. Also, after its insertion, i can only be moved to positions high higher index. Therefore:

$$\Pr[\pi(i) \leq i - K] \leq \Pr[\cup_{k \geq K}\{\pi(i) \leq i - k\}] \leq \sum_{k \geq K} e^{-\beta k} = e^{-\beta K}/(1 - e^{-\beta})$$

From the symmetry of the problem: $\Pr[\pi(i) \geq i + K] \leq e^{-\beta K}/(1 - e^{-\beta})$ and the result follows. □

It follows that the average estimator satisfies a proximity constraint:

Lemma 6.2.2: Average estimator proximity

Suppose that $\Pi = (\pi_\ell)_{\ell \in [r]} \sim (\mathcal{M}_{\pi_0,\beta})^r$, where $\pi_0 \in \mathfrak{S}_n$ and $\beta > 0$. Then, for sufficiently large n and fixed r, it holds that for any $\alpha > 0$:

$$Pr\left[\exists i \in [n] : |\overline{\pi}(i) - \pi_0(i)| \geq \frac{\alpha + 2}{\beta r} \log n\right] < n^{-\alpha}$$

Proof. Let $v = (v_1, \ldots, v_r)$ be a vector of non-negative integers. Also, let

$$A_v = \bigcap_{\ell \in [r] : v_\ell > 0} \{\pi_\ell(i) \leq i - v_\ell\}$$

Then, it holds that:

$$\Pr[\overline{\pi}(i) \leq i - K] \leq \Pr\left[\bigcup_{||v||_1 = rK} A_v\right].$$

Also, from Lemma 6.2.1: $\Pr[A_v] < e^{-\beta||v||_1}/(1 - e^{-\beta})^r$. From the union bound, a counting argument and the following inequality:

$$\binom{rK + r - 1}{r - 1} < (5K + 1)^r,$$

we get that:

$$\Pr[\overline{\pi}(i) \leq i - K] < (5K + 1)^r \frac{e^{-\beta r K}}{(1 - e^{-\beta})^r}$$

Clearly, the symmetric argument also holds and for $K = \frac{\alpha+2}{\beta r} \log n$, with $n \gg r$, the result follows. □

Therefore, the first step of our algorithm is designed successfully. It remains to search within a neighborhood of the average estimator, in order to find a likelier than nature estimator.

Local search. Lemma 6.1.1 provides an efficient way to perform the local search.

• MRP: Applying Lemma 6.1.1 on the average estimator, we get the following Theorem, which corresponds to an efficient algorithm for solving the MRP.

Theorem 6.2.1: MRP solution

Let Π *be a Mallows profile consisting of independent samples of* $\mathcal{M}_{\pi_0,\beta}$*, where* $\pi_0 \in \mathfrak{S}_n$ *and* $\beta > 0$*. Then, for any* $\alpha > 0$*, there exists an algorithm that solves* MRP *with input* Π *in time:*

$$T = O(n^{1+O(\frac{2+\alpha}{\beta r})} \cdot \log^2 n)$$

and with probability of failure less than $n^{-\alpha}$*.*

• MAX-MRP: In order to find the maximum likelihood estimation, it remains to show that π^* is also close to the central ranking. This is shown by Lemma 6.2.3. For the proof of this Lemma we refer to Braverman and Mossel [2009], for we will provide the complete proof of its generalization to the selective Mallows case in the following section.

Lemma 6.2.3

Let $J = 6 \cdot \max(\frac{\alpha+2}{\beta r} \log n, \frac{\alpha+2+1/\beta}{\beta})$*. Then, for the maximum likelihood ranking* π^* *that solves the* MAX-MRP *problem when the central ranking is* π_0*, it holds:*

$$Pr[\exists i \in [n] : |\pi^*(i) - \pi_0(i)| > 32J] < 2n^{-\alpha}$$

Therefore, since the property of proximity is transitive (summing the bounds), the average estimator is proximate to the maximum likelihood estimation and using the algorithm of sorting an almost sorted list, we conclude to the following theorem:

Theorem 6.2.2: max-MRP solution

Let Π *be a Mallows profile consisting of independent samples of* $\mathcal{M}_{\pi_0,\beta}$*, where* $\pi_0 \in \mathfrak{S}_n$ *and* $\beta > 0$*. Then, for any* $\alpha > 0$*, there exists an algorithm that solves* MAX-MRP *with input* Π *in time:*

$$T = O(n^{1+O(\frac{2+\alpha}{\beta r})} \cdot 2^{O(\frac{\alpha}{\beta}+\frac{1}{\beta^2})} \cdot \log^2 n)$$

and with probability of failure less than $n^{-\alpha}$*.*

In the following section, we present our contribution to the direction of solving the SMRP and MAX-SMRP problems.

6.3 Solving MAX-SMRP

In this section, we present our own, original results on solving the Selective Mallows reconstruction problem. The structure of the algorithm we present is similar to the one introduced by Braverman and Mossel [2009] for solving the Mallows reconstruction problem. However, in order to design a solution in the selective Mallows case, non trivial generalizations were required.

Initialization. We show that the positional estimator $\hat{\pi}$, defined in Eq.(5.1) is a good initialization with high probability. Note that it is probable that $\hat{\pi}$ is not a bijection over $[n]$. We break the ties uniformly at random, e.g. from left to right and get a valid permutation.

The initialization of the algorithm presented by Braverman and Mossel [2009] is obtained by estimating the position of each alternative $i \in [n]$ by calculating its average position in the sample profile. However, under selective Mallows model, the average position does not have the same meaning, since each sample is drawn according to a Mallows distribution with a different central ranking of smaller size. In contrast, the estimator $\hat{\pi}$, estimates the position of each alternative by comparing it with each of the other alternatives, using information provided by the sample profile. Given that the sample profile contains information about the relative position of each pair of elements, then we show that $\hat{\pi}$ provides a good approximation for the position of each alternative.

The quality of the approximation is connected to the parameter β of the Mallows distribution, the number of alternatives n, the accepted probability of error ϵ and the frequency p of the sample profile. Intuitively, as β diminishes, the alternatives are more easily swapped and therefore, an alternative might appear in a sample far from its position in the central ranking. Increasing the number of alternatives, the conditions that must hold in order for $\hat{\pi}$ to be accurate are more strict, since more alternatives must be ranked within a small margin of their positions in the central ranking. Finally, as p diminishes, some pairs of alternatives are not compared enough times and therefore, their relative position remains unknown. If the number of such pairs is big enough, then there might be an alternative that is ranked by $\hat{\pi}$ far from its position in central ranking.

Assume that $\mathcal{S}$ is a vector consisting of elements of $2^{[n]}$. We remind the reader that $\mathcal{S}$ (and any corresponding sampling profile of the selective Mallows model) is said to be $p-$frequent if for any pair of elements of $[n]$, the ratio of elements of $\mathcal{S}$ where they both appear to the length of $\mathcal{S}$ is at least p, where $p \in [0, 1]$.

Lemma 6.3.1

Consider a Mallows distribution $\mathcal{M}_{\pi_0,\beta}$ with central ranking π_0 and spread parameter β and let Π be an p-frequent adversarial Mallows profile of size r, induced by π_0 and β. Then for any $\epsilon \in (0, 1)$, there exists an algorithm that, computes an estimate $\hat{\pi}$ of the central ranking π_0, such that, for some:

$$N = O\left(\frac{1}{(\beta p)^2} + \frac{1}{\beta p^2 r}\log(n/\epsilon)\right), \tag{6.2}$$

the probability that there exists $i \in [n]$ such that $|\hat{\pi}(i) - \pi_0(i)| > N$ is no more than ϵ.

• *Proof of Lemma 6.3.1:* The basic observation that supports the idea of approximating the true position of each alternative is that the probability of displacement of an alternative decays exponentially to the length of the displacement (which also implies that the probability of swap between two alternatives decays exponentially to their distance in the central ranking, since at least one of them must be displaced at least half their distance in order for them to swap).

In the classic Mallows case, this observation directly implies a notion of neighborhood $\mathcal{N}_i(K)$ (see figure 6.1) for each alternative $i \in [n]$ which includes all the elements that are

ranked no more than $K(\in [n])$ places away from i in π_0. For the elements outside this neighborhood there is a probabilistic guarantee that provides a bound for the probability of their swapping with i, which diminishes exponentially with respect to K. Picking K to be of order $O(\log n)$, we can guarantee that, with high probability, the elements outside the neighborhood of each alternative i will not be swapped with i except within a minority of the samples. This indicates that the positional estimator $\hat{\pi}$ is a good initialization in the classic Mallows case.

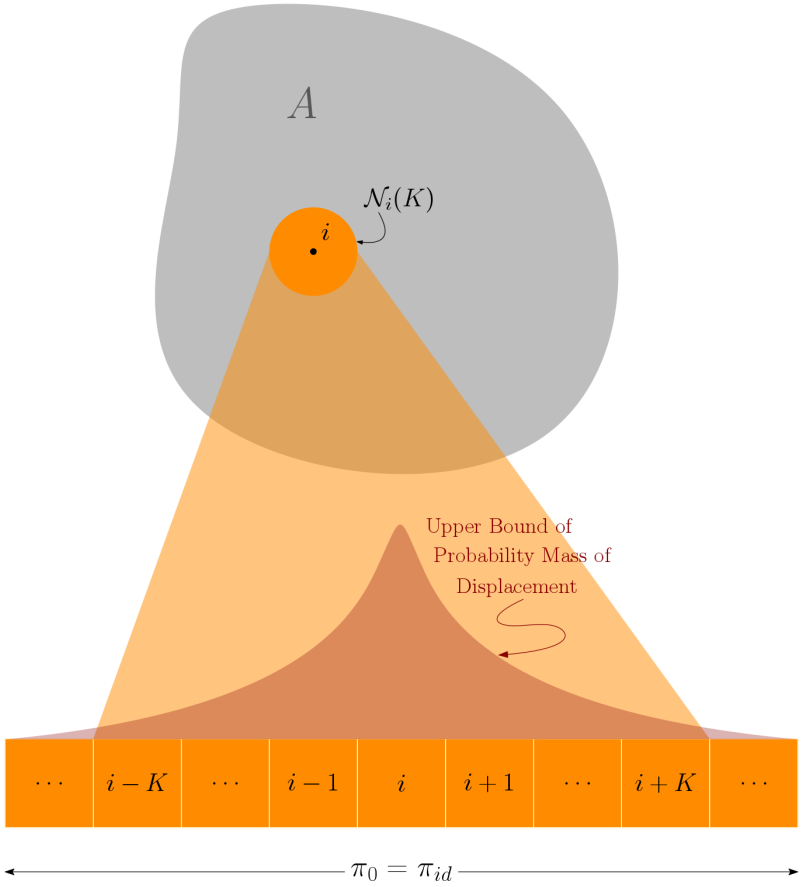

Figure 6.1: Neighborhood in classic Mallows case

However, under the selective Mallows model, the distance of two alternatives in the central ranking might be significantly smaller in a reduced central ranking. Therefore, in order to take into consideration this possibility, for each sample profile, we define a notion of neighborhood for each alternative, as follows:

Definition 6.3.1 (Right Neighborhood): *Assume we are given a sample profile* Π *consisting of* r *independent samples from* $\mathcal{M}_{\pi_0,\beta}^{\text{ADV}\to\mathcal{S}}$, *where* $\pi_0 \in \mathfrak{S}_n$, $\beta > 0$, $\mathcal{S} = (S_1, ..., S_r) \in (2^{[n]})^r$. *For any* $i \in [n]$, $M \in [n-1]$ *and* $\lambda > 1$, *we denote with* $\mathcal{N}_i^R(M, \lambda) \subset [n]$, *each element* j *of which has the following properties:*

1. $\pi_0(i) < \pi_0(j)$
2. $|\{k : \{i, j\} \in S_k \wedge \pi_0|_{S_k}(j) \leq \pi_0|_{S_k}(i) + M\}| \geq r/\lambda$

We, similarly, define the left neighborhood $\mathcal{N}_i^L, \forall i \in [n]$.

In the Selective Mallows case, the neighborhood we defined depends not only on π_0, but also on the selection sets vector $\mathcal{S}$: The neighborhood of i consists of the alternatives j which are ranked $M-$close to i in at least a fraction (determined by parameter λ) of the reduced central rankings of the samples (see figure 6.2).

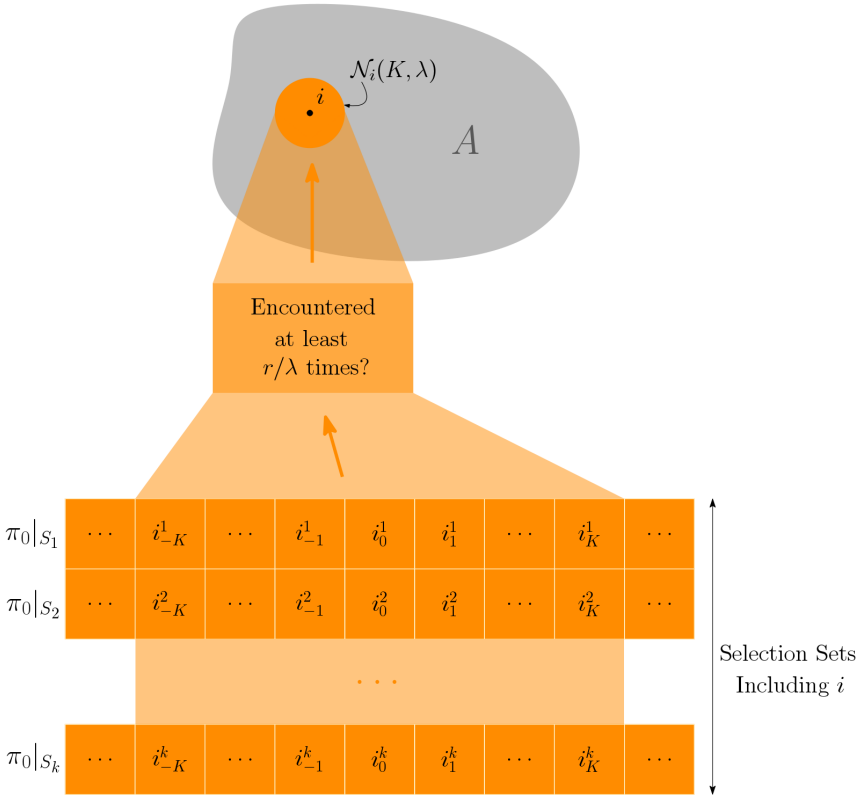

Figure 6.2: Neighborhood in selective Mallows case. Notation: $i \in [n]$ and $i^x_y = (\pi_0|_{S_x})^{-1}(i+y)$.

Observe that if for some $j \in [n]$ with $\pi_0(j) > \pi_0(i)$ it holds that $j \notin \mathcal{N}_i^R(M,\lambda)$, then the number of samples where i and j have a distance less than M is bounded, which means that in the rest of the samples where i and j both appear, they will be distant. The following proposition shows that if we allow the non-neighboring elements to be close in a nontrivial number of samples, then the neighborhood length is not very large.

Proposition 6.3.1: *It holds that:* $|\mathcal{N}_i^R(M,\lambda)| \leq \lambda M, i \in [n]$.

Proof. In each sample where i appears there are M places available for candidate neighbors. Therefore, there are Mr places in total. However, each neighbor takes r/λ places at least. Therefore, the number of neighbors cannot be higher than λM. □

Obviously, the same holds for the left neighborhood $\mathcal{N}_i^L, i \in [n]$.

We continue with a very useful lemma. Lemma 6.3.2 states that selecting appropriate parameters for the neighborhoods that we defined, we ensure that with arbitrary high

probability, for each alternative i, all the alternatives which are not included in its neighborhood are ranked correctly relatively to i in the majority of samples, while the size of the neighborhood remains small.

Lemma 6.3.2

Assume we are given a sample profile Π *consisting of* r *independent samples from* $\mathcal{M}_{\pi_0,\beta}^{\text{ADV}\to\mathcal{S}}$, *where* $\pi_0 \in \mathfrak{S}_n$, $\beta > 0$ *and* $\mathcal{S} \in (2^{[n]})^r$, *that is* $p-$*frequent for some* $p \in (0,1]$. *Then, for any* $c \in (0, 1/2]$, $\epsilon \in (0,1)$, *considering:*

$$L = L(p) = \frac{2}{\beta^2} + \frac{4}{\beta cp} + \frac{4}{\beta cpr}\log(n^2/\epsilon),$$

with probability at least $1-\epsilon$, *for every* $i \in [n]$, *there exists a set* $\mathcal{C}_i^R \subset R_i = \{j \in [n] : \pi_0(j) > \pi_0(i)\}$ *such that:*

1. *For any* $j \in R_i \setminus \mathcal{C}_i^R$, *it holds that* $q(j \succ i) \leq cW_{ij}$.
2. $|\mathcal{C}_i^R| < \frac{2(1+c)}{cp}L$.

Proof. Suppose that $i \in [n]$ and j is selected so that $\pi_0(j) > \pi_0(i)$ and $j \notin \mathcal{N}_i^R(L,\lambda)$, where $\lambda > 1$ will be defined later.

Then there are the following groups of samples (see figure 6.3):

1. The set of all samples, which has cardinality r.
2. The set of the samples where both i and j appear, which includes $W_{ij} \geq rp$ samples (since $p := \inf_{i,j} W_{ij}/r$).
3. The set of the samples (π_k, S_k) where both i and j appear and, also, the truncated central ranking $\pi_0|_{S_k}$ ranks them at least L positions away. This set includes $W \geq W_{ij} - r/\lambda$ samples, due to the selection of j.

This means that for the chosen $j \notin \mathcal{N}_i^R$, we have that $W \geq W_{ij} - r/\lambda$.

We denote with $q(j \succ i)$ the number of samples where j is ranked before i (we count only the samples where both i and j appear) and with $\tilde{q}(j \succ i)$ the number of samples where j is ranked before i and the corresponding restricted central ranking ranks i and j at least L positions away.

We would like to bound the probability that $q(j \succ i) > cW_{ij}$. Note that provided that W is large enough in relation to W_{ij}, then it would be sufficient to bound the probability that $\tilde{q}(j \succ i) > c'W$, for some appropriate c'. In particular, we would like:

$$q(j \succ i) > cW_{ij} \Rightarrow \tilde{q}(j \succ i) > c'W,$$

We choose $c' \leq c - 1/(\lambda p - 1)$ and get:

$$q(j \succ i) \leq \tilde{q}(j \succ i) + r/\lambda \Rightarrow \tilde{q}(j \succ i) > cW_{ij} - r/\lambda \geq (c' + 1/(\lambda p - 1))W - r/\lambda \geq c'W$$

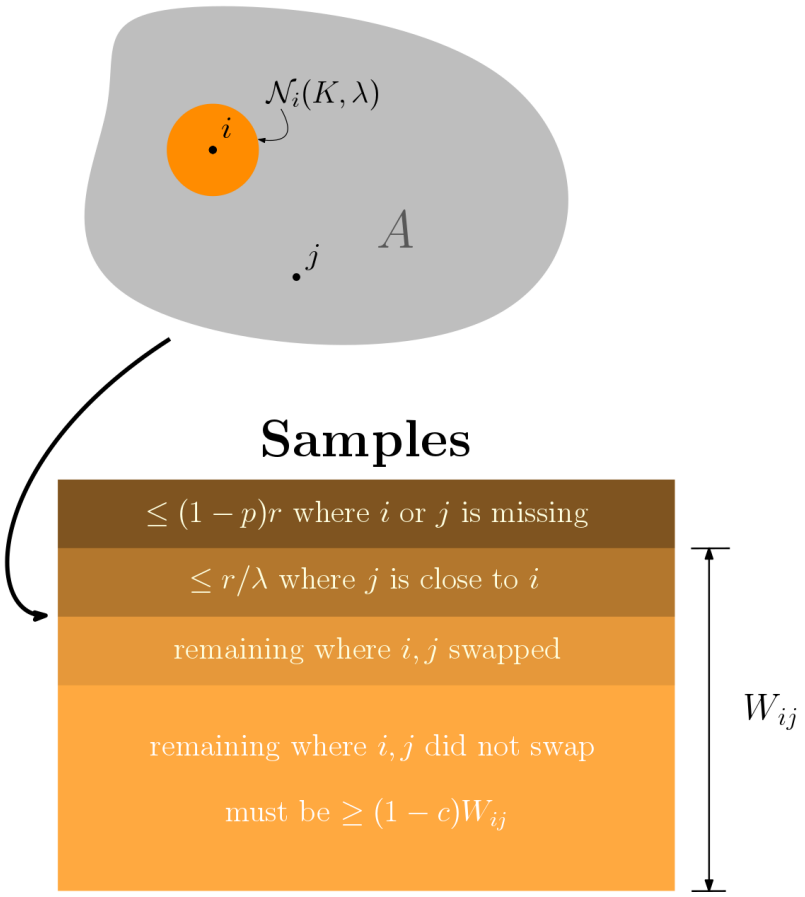

Figure 6.3: Sample grouping in order to pick appropriate parameters for the neighborhood in the selective Mallows case.

It remains to bound the probability that $\tilde{q}(j \succ i) > c'W$.

For a Mallows sample $\pi \sim \mathcal{M}_{\pi_0,\beta}$, the probability of the event $\pi(i) \leq \pi_0(i) - K$ is bounded by $e^{-\beta K}/(1-e^{-\beta}), \forall K$ (Lemma 6.2.1).

Therefore, in each of the W samples where i and j are distant in the central ranking, the probability of their swapping is less than $p = 2e^{-\beta L/2}/(1-e^{-\beta})$, since in case of swapping, at least one of them must be misplaced by at least $L/2$ positions.

We have:

$$\Pr[\tilde{q}(j<i) > c'W] = \sum_{k=(p-1/\lambda)r}^{r} \Pr[\tilde{q}(j<i) > c'k | W = k]\Pr[W = k],$$

since $W \geq W_{i,j} - r\lambda \geq (p - 1/\lambda)r$. In the case when $W = k$, $\tilde{q}(j < i)$ is the sum of k Bernoulli trials, each one of which has a parameter at most p. Therefore:

$$\Pr[\tilde{q}(j<i) > c'k | W = k] \leq \sum_{t>c'k} \binom{k}{t} p^t \leq p^{c'k}2^k \leq p^{c'(p-1/\lambda)r}2^r,$$

since $k \in [(p-1/\lambda)r, r]$. We pick $c' = c - 1/(\lambda p - 1)$. In order for the exponent of p to be positive, the following condition must hold: $\lambda > (1+c)/(cp)$.

Hence:

$$\Pr[\tilde{q}(j<i) > c'W] \leq p^{(cp-c/\lambda-1/\lambda)r}2^r = (2e^{-\beta L/2}/(1-e^{-\beta}))^{(cp-c/\lambda-1/\lambda)r}2^r$$

We demand that the above quantity is less than ϵn^{-2}, which gives that L must be:

$$L \geq \frac{2}{\beta}\left(\log(1/(1-e^{-\beta})) + \frac{1}{cp-c/\lambda-1/\lambda}\log 2 + \frac{1}{(cp-c/\lambda-1/\lambda)r}\log(n^2/\epsilon)\right)$$

It suffices that $L = \frac{2}{\beta^2} + \frac{2}{\beta Z} + \frac{2}{\beta Z r}\log(n^2/\epsilon)$, where $Z = (cp - c/\lambda - 1/\lambda)$. Afterwards, We pick the set $\mathcal{C}_i^R = \mathcal{N}_i^R(L^*, \lambda)$, where λ is to be chosen.

From the union bound over all possible i and j, since $|\mathcal{C}_i^R| \leq \lambda L$, due to Proposition 6.3.1.

Finally, we optimize over λ. In order to approximately minimize the quantity $\lambda L(\lambda)$, subject to $\lambda > (1+c)/(cp)$, we choose: $\lambda^* = 2(c+1)/(cp)$, which minimizes the function $g(\lambda) = \lambda(L(\lambda) - 2/\beta^2)$ and the remaining term $2\lambda^*/\beta^2$ is close to its minimum value, since λ must be greater than $(1+c)/(cp)$. Therefore, we get that $Z^* = cp/2$ and:

$$L^* = \frac{2}{\beta^2} + \frac{2}{\beta Z^*} + \frac{2}{\beta Z^* r}\log(n^2/\epsilon)$$

□

Exactly similar results hold for the symmetric problem (when $\pi_0(j) < \pi_0(i)$). Using the Lemma 6.3.2 we prove that the estimator $\hat{\pi}$, extended by breaking ties uniformly provides an approximation of π_0, which completes the proof of Lemma 6.3.1:

Proof. We apply Lemma 6.3.2, for any $c < 1/2$, as well as the similar one which corresponds to the left neighborhood. Therefore, if $L = \frac{2}{\beta^2} + \frac{4}{\beta cp} + \frac{4}{\beta cpr}\log(n^2/\epsilon)$, then with probability at least $1-\epsilon$, for each $i \in [n]$, there are at most $K = \frac{2(1+c)}{cp}L$ elements $j \in [n]$, with $\pi_0(i) < \pi_0(j)$ ($\pi_0(j) < \pi_0(i)$) which are not ranked correctly in relation to i in the samples. Therefore: $|\hat{\pi}(i) - \pi_0(i)| \leq K$, with probability at least $1-\epsilon$, where $\hat{\pi}(i) - 1$ is the number of elements of $[n]$ that are ranked before i in most samples.

It remains to convert $\hat{\pi}$ into a ranking $\hat{\pi} \in \mathfrak{S}_n$ by breaking ties uniformly. Following the steps presented by Rubinstein and Vardi [2017]: If $\hat{\pi}(j) \leq \hat{\pi}(i)$ then: $\hat{\pi}(j) + K \leq \hat{\pi}(i) + K$. However: $j \leq \hat{\pi}(j) + K$ and $\hat{\pi}(i) + K \leq i + 2K$ therefore: $j \leq i + 2K \Rightarrow j - i \leq 2K$. Therefore: $\hat{\pi}(i) \leq i + 2K$. Symmetrically: $\hat{\pi}(i) \geq i - 2K$. □

Therefore, we have proven that the positional estimator $\hat{\pi}$ is a good initialization of the algorithm, as it satisfies the proximity property. Note that parameter p appears squared in the denominator of the proximity guarantee (Lemma 6.3.1), due to the uncertainty that selectivity inserts into the distance between a pair of alternatives in the reduced central rankings that correspond to input samples.

Local search. We finally present the two main results.

• SMRP: Applying the algorithm of solving an almost sorted list, we immediately get the following result:

Theorem 6.3.1

Consider a Mallows distribution $\mathcal{M}_{\pi_0,\beta}$ with central ranking π_0 and spread parameter β and let Π be an p-frequent adversarial Mallows profile of size r, induced by $\mathcal{M}_{\pi_0,\beta}$. Then for any $\alpha > 0$, there exists an algorithm that computes an estimate π^*, that solves SMRP *with input Π in time:*

$$T = O\left(n^2 + n^{1+O\left(\frac{2+\alpha}{\beta r p^2}\right)} 2^{O\left(\frac{1}{(\beta p)^2}\right)} \log^2 n\right), \tag{6.3}$$

and with probability of failure at most $n^{-\alpha}$.

Proof. First we calculate the estimation $\hat{\pi}$ in time $O(n^2)$, which with probability at least $1 - n^{-\alpha}$ sorts every element at most $N = O(\frac{1}{(\beta p)^2} + \frac{\alpha}{\beta p^2 r}\log(n))$ places away from its position in π_0, by Lemma 6.3.1. Therefore, $\hat{\pi}$ is an almost sorted list. Using the algorithm for sorting an almost sorted list (Lemma 6.1.1) we get the result. □

• MAX-SMRP: It remains to show that the maximum likelihood estimation of the central ranking ranks each alternative close to its position in the central ranking:

Lemma 6.3.3

Assume we are given an $p-$frequent sample profile Π from $\mathcal{M}^{\text{ADV}}_{\pi_0,\beta}$, where $p \in (0,1]$, $\pi_0 \in \mathfrak{S}_n$ and $\beta > 0$. Then, for the maximum likelihood estimator π^* of π_0 from Π, it holds that for any $\epsilon \in (0,1)$, there exists some K, such that:

$$K = O\left(\frac{1}{\beta^2 p^4} + \frac{1}{\beta p^4 r}\log(n^2/\epsilon)\right),$$

and the probability that there exists $i \in [n]$ such that $|\pi^*(i) - \pi_0(i)| > K$ is at most 2ϵ.

Proof. The proof follows the same steps presented by Braverman and Mossel [2009], generalizing them to suit the Selective Mallows model. Assume, without loss of generality, that $\pi_0 \equiv id$. Let $h > 0$ and $c \in (0, 1/2)$, which will be defined later. Assume that for any $i, j \in [n]$, $q(j \succ i)$ is the number of samples where j is ranked before i (that is $i \succ j$) and W_{ij} is the number of samples where both i and j appear. Apparently, it holds that: $q(i \succ j) + q(j \succ i) = W_{ij}$. Then, from Lemma 6.3.2, with probability at least $1 - \epsilon$, for every alternative $i \in [n]$, there exists a set $\mathcal{C}_i^R \subseteq \{i+1, i+2, \ldots, n\}$ such that:

1. $|\mathcal{C}_i^R| \leq N = \frac{2(1+c)}{cp}(\frac{2}{\beta^2} + \frac{4}{\beta cp} + \frac{4}{\beta cpr}\log(n^2/\epsilon))$
2. $j \in \{i+1, i+2, \ldots, n\} \setminus \mathcal{C}_i^R \Rightarrow q(j \succ i) \leq cW_{ij}$ and $W_{ij} \geq pr$

Symmetrically, the same holds for sets $\mathcal{C}_i^L$, for any $i \in [n]$.

Fix $i \in [n]$ such that $|\pi^*(i) - i| = K$, where $K \geq hN$. Without loss of generality, assume $\pi^*(i) = i + K$. It suffices to find values of c and h that contradict the assumption $\pi^*(i) = i + K$.

Let $S = \{j \in [n\} : i \leq \pi^*(j) < i + K]$ and: $S_1 = \{j \in S : j < i\}$, $S_2 = \{j \in S : j \in \mathcal{C}_i^R\}$, $S_3 = \{j \in S : j > i \text{ and } j \notin \mathcal{C}_i^R\}$. Apparently: $S = S_1 \cup S_2 \cup S_3$.

Observe that since π^* maximizes the following score function:

$$\begin{aligned} s: \ & \mathfrak{S}_n \to \mathbb{N} \\ & \pi \to s(\pi) = \sum_{i_1 \succ_\pi i_2} q(i_1 \succ i_2) \end{aligned}$$

It must hold that:

$$0 \leq \sum_{j \in S} (q(j \succ i) - q(i \succ j))$$

For any $j \in S_3$, from the assumption for $\mathcal{C}_i^R$, it holds that:

$$q(j \succ i) \leq cW_{ij} \Rightarrow q(j \succ i) - q(i \succ j) \leq -(1 - 2c)pr$$

Let $|S_1| = T$. Furthermore: $|S_2| \leq N$ and $|S_3| \geq K - N - T \geq (h-1)N - T$. Therefore:

$$\begin{aligned} 0 \leq rT + rN - (1 - 2c)pr((h-1)N - T) \Rightarrow \\ T \geq \frac{(1-2c)p(h-1) - 1}{1 + (1-2c)p} N \end{aligned} \tag{6.4}$$

Observe that, since there are at least T alternatives $j < i$ such that $\pi^*(j) \geq i$, say $T_1 \subset [n]$, there must be at least T alternatives $j \geq i$ such that $\pi^*(j) < i$, say $T_2 \subset [n]$. Let $H_1 = \{1, ..., i-1\}$ and $H_2 = \{i, ..., n\}$. We construct $\pi_1 \in \mathfrak{S}_n$ by concatenating $\pi^*|_{H_1}$ and $\pi^*|_{H_2}$. It remains to select appropriate values for h and c for which $s(\pi_1) > s(\pi^*)$, which is a contradiction.

Create the following sets:

1. P_1: The pairs of elements $i_1, i_2 \in [n], i_1 < i_2$ for which $i_2 \in \mathcal{C}_{i_1}^R$ (or equivalently $i_1 \in \mathcal{C}_{i_2}^L$) and π_1, π^* disagree on their relative ranking. Note that: $|P_1| \leq 2TN$.
2. P_2: The pairs of elements $i_1, i_2 \in [n], i_1 < i_2$ for which π_1, π^* disagree, but $i_2 \notin \mathcal{C}_{i_1}^R$ and $i_1 \notin \mathcal{C}_{i_2}^L$. Note that π_1 has the right answer for this pair and $q(i_1 \succ i_2) - q(i_2 \succ i_1) \geq (1-2c)pr$. Also: $|P_2| \geq T(T-N)$ (select an element of T_1 and an element of T_2 which is not in the first element's neighborhood).

Then: $s(\pi_1) - s(\pi^*) = \sum_{(i_1,i_2) \in P_1} (q(i_1 \succ i_2) - q(i_2 \succ i_1)) + \sum_{(i_1,i_2) \in P_2} (q(i_1 \succ i_2) - q(i_2 \succ i_1)) \geq -2rTN + (1-2c)prT(T-N) = rT((1-2c)pT - ((1-2c)p + 2)N)$

Using Ineq. 6.4, we get that:

$$s(\pi_1) - s(\pi^*) \geq rTN \left[\frac{(1-2c)p((1-2c)p(h-1) - 1)}{1 + (1-2c)p} - (2 + (1-2c)p) \right]$$

We search for values of c and h such that the quantity inside the brackets is positive. After analysis, we choose $c < 1/4$ (constant) and $h = 2 + 8/p + 8/p^2 = O(\frac{1}{p^2})$.

□

We are now ready to prove Theorem 6.3.2:

Theorem 6.3.2: max-SMRP solution

Consider a Mallows distribution $\mathcal{M}_{\pi_0,\beta}$ with central ranking π_0 and spread parameter β and let Π be an p-frequent adversarial Mallows profile of size r, induced by $\mathcal{M}_{\pi_0,\beta}$. Then for any $\alpha > 0$, there exists an algorithm that computes an estimate π^*, that solves MAX-SMRP *with input Π in time:*

$$T = O\left(n^2 + n^{1+O\left(\frac{2+\alpha}{\beta r p^4}\right)} 2^{O\left(\frac{1}{\beta^2 p^4}\right)} \log^2 n\right), \tag{6.5}$$

and with probability of failure at most $n^{-\alpha}$.

Proof. First we calculate the estimation $\hat{\pi}$ in time $O(n^2)$, which with probability at least $1 - n^{-\alpha}$ sorts every element at most $N = O(\frac{1}{(\beta p)^2} + \frac{\alpha}{\beta p^2 r}\log(n))$ places away from its position in π_0, by Lemma 6.3.1. Therefore, $\hat{\pi}$ is an almost sorted list. The solutions of MAX-SMRP are also almost sorted lists. Therefore, $\hat{\pi}$ is also an almost sorted list with respect to the maximum likelihood estimation. Using the algorithm for sorting an almost sorted list (Lemma 6.1.1), we get the result. □

Chapter 7

Conclusions and further work

We introduced the Selective Mallows model as an interpolation between two models: Mallows model and Noisy Comparisons model. Indeed, as we showed in Chapter 4, Selective Mallows Reconstruction problem is a generalized version of the Reconstruction problems that correspond to each of these models; in other words, the problem of finding the maximum likelihood estimation of the central ranking is the same in these models.

Furthermore, in Chapter 5, we established tight sample complexity bounds for retrieving the central ranking in the case that selection sets are picked adversarially or randomly, introducing the notion of frequency of selectivity. These bounds indicate that when one cannot pick the selection sets, an optimal option is to focus only on the information provided by the samples in the form of pairwise comparisons. That is, it is an optimal strategy to view the sample profile as a collection of noisy comparisons.

Finally, in Chapter 6, we showed that under the Selective Mallows model, if a sample profile is rich in pairwise information - that is, each pair of alternatives appears frequently in the samples - then, the maximum likelihood estimation of the central ranking can be reconstructed efficiently. This is a consequence of the fact that rich pairwise information enables estimating each alternative's position in the central ranking and also of the similarity in the structure of Selective Mallows reconstruction problem and the classic Mallows Reconstruction problem, for which Braverman and Mossel [2009] have already provided an efficient algorithm.

However, there are still two directions in which our results can be extended. The first one regards central ranking's retrieval. In particular, although in the case that we do not pick the selection sets, the Selective Mallows model does not have an advantage over the Noisy Comparisons model with respect to retrieving the central ranking, it might be true that in the adversarial model, namely when we pick the selection sets in the runtime of the estimation algorithm, the concentration property of the Selective Mallows model could be exploited. The results we already have are presented in table 7.1.

Queries	Noiseless Setting	Noisy Setting
Pairwise comparisons	$\Theta(n \log n)$	$\Theta(n \log n)$
Incomplete Rankings of length m	$O(\frac{n}{m} \log(n/m))$	$O(\frac{n}{m} \log(n))$
Complete Rankings	1	$\Theta(\log n)$

Table 7.1: Query complexity for sorting and noisy sorting.

However, in the incomplete ranking queries case, it remains open to establish lower bounds of the query complexity or improve the bounds we provided.

The second direction corresponds to improving the time complexity of the algorithms corresponding to Theorems 6.3.1 and 6.3.2, taking advantage of the observation that when the selectivity parameter p is small, then the maximum likelihood estimation is also influenced, by becoming less precise. In the extreme case that an alternative never appears in the samples, picking an arbitrary position for the alternative does not influence the likelihood of the proposed solution.

Bibliography

Kenneth Arrow. Social choice and individual values. 1951.

Mark Braverman and Elchanan Mossel. Sorting from noisy information. *arXiv preprint arXiv:0910.1191*, 2009.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

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, 1994a.

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.

Uriel Feige, Prabhakar Raghavan, David Peleg, and Eli Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018, 1994.

Aviad Rubinstein and Shai Vardi. Sorting from noisier samples. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 960–972. SIAM, 2017.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.

Colin L Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.

John G Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

Michael A Fligner and Joseph S Verducci. Distance based ranking models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(3):359–369, 1986.

Guy Lebanon and Yi Mao. Non-parametric modeling of partially ranked data. *Journal of Machine Learning Research*, 9(Oct):2401–2429, 2008.

Jonathan Huang, Ashish Kapoor, and Carlos E Guestrin. Efficient probabilistic inference with partial ranking queries. *arXiv preprint arXiv:1202.3734*, 2012.

Ramakrishna Kakarala. Interpreting the phase spectrum in fourier analysis of partial ranking data. *Advances in Numerical Analysis*, 2012, 2012.

Mark Braverman and Elchanan Mossel. Noisy sorting without resampling, 2007.

Fabian Wauthier, Michael Jordan, and Nebojsa Jojic. Efficient ranking from pairwise comparisons. In *International Conference on Machine Learning*, pages 109–117, 2013.

Róbert Busa-Fekete, Eyke Hüllermeier, and Balázs Szörényi. Preference-based rank elicitation using statistical models: The case of mallows. 2014.

Mohsen Ahmadi Fahandar, Eyke Hüllermeier, and Inés Couso. Statistical inference for incomplete ranking data: the case of rank-dependent coarsening. *arXiv preprint arXiv:1712.01158*, 2017.

Daniel F Heitjan and Donald B Rubin. Ignorability and coarse data. *The annals of statistics*, pages 2244–2253, 1991.

Richard D Gill, Mark J Van Der Laan, and James M Robins. Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–294. Springer, 1997.

Ludwig M Busse, Peter Orbanz, and Joachim M Buhmann. Cluster analysis of heterogeneous rank data. In *Proceedings of the 24th international conference on Machine learning*, pages 113–120, 2007.

Marina Meila and Le Bao. An exponential model for infinite rankings. *J. Mach. Learn. Res.*, 11:3481–3518, 2010.

Marina Meila and Harr Chen. Dirichlet process mixtures of generalized mallows models. *arXiv preprint arXiv:1203.3496*, 2012.

Wenpin Tang. Mallows ranking models: maximum likelihood estimate and regeneration. *arXiv preprint arXiv:1808.08507*, 2018.

Flavio Chierichetti, Anirban Dasgupta, Shahrzad Haddadan, Ravi Kumar, and Silvio Lattanzi. Mallows models for top-k lists. In *Advances in Neural Information Processing Systems*, pages 4382–4392, 2018.

Arun Rajkumar and Shivani Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *International Conference on Machine Learning*, pages 118–126, 2014.

Eric Sibony, Stéphan Clémençon, and Jérémie Jakubowicz. Mra-based statistical learning from incomplete rankings. 2015.

Tyler Lu and Craig Boutilier. Learning mallows models with pairwise preferences. In *ICML*, 2011.

Ariel D Procaccia, Sashank J Reddi, and Nisarg Shah. A maximum likelihood approach for selecting sets of alternatives. *arXiv preprint arXiv:1210.4882*, 2012.

Andrei Nikolaevics Kolmogorov. Foundations of the theory of probability. 1950.

Paul R Halmos. *Measure theory*, volume 18. Springer, 2013.

Henri Lebesgue. Remarques sur les théories de la mesure et de l'intégration. In *Annales scientifiques de l'École Normale Supérieure*, volume 35, pages 191–250, 1918.

Émile Borel. L'intégration des fonctions non bornées. In *Annales scientifiques de l'École Normale Supérieure*, volume 36, pages 71–92, 1919.

Bernhard Riemann. Ueber die darstellbarkeit einer function durch eine trigonometrische reihe. In *Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen*, volume 13, pages 87–132, 1868.

Thomas Jech. *Set theory*. Springer Science & Business Media, 2013.

Charles Antony Richard Hoare. Algorithm 64: quicksort. *Communications of the ACM*, 4 (7):321, 1961.

William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559–572, 1901.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66(4):671–687, 2003.

Gary L Miller. Riemann's hypothesis and tests for primality. *Journal of computer and system sciences*, 13(3):300–317, 1976.

Michael O Rabin. Probabilistic algorithm for testing primality. *Journal of number theory*, 12(1):128–138, 1980.

Ronald L. Rivest Clifford Stein Thomas H. Cormen, Charles E. Leiserson. *Introduction to Algorithms (Second ed.)*. MIT Press, 2001.

Paul Erdös. Some remarks on the theory of graphs. *Bulletin of the American Mathematical Society*, 53(4):292–294, 1947.

Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

Noga Alon and Joel H Spencer. *The probabilistic method*. John Wiley & Sons, 2004.

Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Chapman & Hall/CRC, 2010.

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal*, pages 13–30, 1963.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Paul ErdHos and László Lovász. Problems and results on 3-chromatic hypergraphs and some related questions. In *Infinite and finite series*, pages 609–628. A. Hajnal et al. eds., North-Holland, 1975.

Richard Gross. *Psychology: The science of mind and behaviour 7th edition*. Hodder Education, 2015.

Michael J Kearns, Umesh Virkumar Vazirani, and Umesh Vazirani. *An introduction to computational learning theory*. MIT press, 1994b.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of data science*. Cambridge University Press, 2020.

Vladimir Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Doklady Akademii Nauk USSR*, volume 181, pages 781–787, 1968.

Johann Pfanzagl. *Parametric statistical theory*. Walter de Gruyter, 2011.

John I Marden. *Analyzing and modeling rank data*. CRC Press, 1996.

David Steven Dummit and Richard M Foote. *Abstract algebra*, volume 3. Wiley Hoboken, 2004.

John B Fraleigh. *A first course in abstract algebra*. Pearson Education India, 2003.

Charles Spearman. 'footrule'for measuring correlation. *British Journal of Psychology, 1904-1920*, 2(1):89–108, 1906.

Persi Diaconis and Ronald L Graham. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):262–268, 1977.

Timothy M Chan and Mihai Pătraşcu. Counting inversions, offline orthogonal range counting, and related problems. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 161–173. SIAM, 2010.

Duncan R Luce. *Individual choice behavior: A theoretical analysis*. Wiley, New York, 1959.

Jean-Paul Doignon, Aleksandar Pekeč, and Michel Regenwetter. The repeated insertion model for rankings: Missing link between two subset choice models. *Psychometrika*, 69 (1):33–54, 2004.

John Bartholdi, Craig A Tovey, and Michael A Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and welfare*, 6(2):157–165, 1989.

Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622, 2001a.

Cynthia Dwork, Ravi Kumar, Moni Naor, and D Sivakumar. Rank aggregation revisited, 2001b.

Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)*, 55(5):1–27, 2008.

Claire Kenyon-Mathieu and Warren Schudy. How to rank with few errors. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 95–103, 2007.

Michael A Fligner and Joseph S Verducci. Posterior probabilities for a consensus ordering. *Psychometrika*, 55(1):53–63, 1990.

William W Cohen, Robert E Schapire, and Yoram Singer. Learning to order things. In *Advances in neural information processing systems*, pages 451–457, 1998.

Marina Meila, Kapil Phadnis, Arthur Patterson, and Jeff Bilmes. Consensus ranking under the exponential model. 2007.

Sumit Mukherjee et al. Estimation in exponential families on permutations. *The Annals of Statistics*, 44(2):853–875, 2016.

Robert Busa-Fekete, Dimitris Fotakis, Balázs Szörényi, and Manolis Zampetakis. Optimal learning of mallows block model. In *Conference on Learning Theory*, pages 529–532, 2019.

Ekhine Irurozki, Borja Calvo, and Jose A Lozano. Sampling and learning mallows and generalized mallows models under the cayley distance. *Methodology and Computing in Applied Probability*, 20(1):1–35, 2018.

Pranjal Awasthi, Avrim Blum, Or Sheffet, and Aravindan Vijayaraghavan. Learning mixtures of ranking models. In *Advances in Neural Information Processing Systems*, pages 2609–2617, 2014.

Allen Liu and Ankur Moitra. Efficiently learning mixtures of mallows models. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 627–638. IEEE, 2018.

Anindya De, Ryan O'Donnell, and Rocco Servedio. Learning sparse mixtures of rankings from noisy information. *arXiv preprint arXiv:1811.01216*, 2018.