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

Τεχνικές Εκμάθησης Κατανομών Κατάταξης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάργυρος-Γεώργιος Μουζάκης

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

Αθήνα, 12/11/2019

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Τεχνικές Εκμάθησης Κατανομών Κατάταξης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάργυρος-Γεώργιος Μουζάκης

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

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 12/11/2019.

.....	.....	.....
Δημήτριος Φωτάκης	Αριστείδης Παγουρτζής	Μιχαήλ Λουλάκης
Αναπληρωτής Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Ε.Μ.Π.

.....
Ανάργυρος-Γεώργιος Μουζάκης
(Διπλωματούχος Ηλεκτρολόγος Μηχανικός & Μηχανικός Υπολογιστών Ε.Μ.Π.)

Περίληψη

Οι κατανομές κατάταξης είναι ένα πεδίο που παραδοσιακά έχει προσελκύσει το ενδιαφέρον τόσο της κοινότητας των στατιστικολόγων, όσο και των ακαδημαϊκών που εργάζονται στο πεδίο της θεωρίας κοινωνικής επιλογής. Τα τελευταία χρόνια, έχουν επίσης τραβήξει την προσοχή αυτών που εργάζονται στους τομείς της θεωρητικής πληροφορικής και της μηχανικής μάθησης. Σε αυτή τη διπλωματική εργασία, εξετάζουμε προβλήματα μάθησης κατανομών στο πεδίο των κατανομών κατάταξης και, συγκεκριμένα, στο μοντέλο του Mallows. Ξεκινάμε εισάγοντας το τυπικό πλαίσιο της μάθησης κατανομών, καθώς και το απαραίτητο υπόβαθρο για την κατανόηση των θεμελιωδών τεχνικών της μάθησης κατανομών. Ακολουθεί μία εισαγωγή στη θεωρία των μεταθέσεων και στα μοντέλα κατατάξεων, με έμφαση στο μοντέλο του Mallows. Ύστερα, παρουσιάζουμε τις εργασίες των Καραγιάννη et. al. και Busa-Fekete et. al., που παρείχαν βέλτιστα ως προς την δειγματική πολυπλοκότητα αποτελέσματα για την εκτίμηση παραμέτρων και τη μάθηση κατανομών στο μοντέλο Kendall-Mallows. Μετά, προσαρμόζουμε αυτές τις τεχνικές προκειμένου να πάρουμε έναν αλγόριθμο πολυωνυμικού χρόνου που ανακτά την κεντρική κατάταξη στο μοντέλο Cayley-Mallows με μεγάλη πιθανότητα. Τέλος, εξετάζουμε πιθανές κατευθύνσεις έρευνας.

Λέξεις κλειδιά: Μάθηση Κατανομών, Κοινωνική Επιλογή, Θεωρία Πληροφορίας, Κατανομές Κατάταξης, Μοντέλο Kendall-Mallows, Μοντέλο Cayley-Mallows

Abstract

Ranking distributions are a field that has traditionally drawn the interest of the statistics community, as well as that of scholar working in the field of social choice theory. In recent years, they have also drawn the attention of the theoretical computer science and machine learning communities. In this thesis, we examine distribution learning problems in the context of ranking distributions and, in particular, the Mallows model. We start by introducing the formal framework of distribution learning, along with the necessary background for understanding the fundamental techniques of distribution learning. This is followed by an introduction to the theory of permutations and ranking models with emphasis on the Mallows model. Subsequently, we present the works of Caragiannis et. al. and Busa-Fekete et. al., which provided sample optimal results about parameter estimation and distribution learning in the Kendall-Mallows model. Then, we adjust those techniques to obtain a polynomial time algorithm that recovers the central ranking in the Cayley-Mallows model with high probability. Finally, we examine possible research directions.

Keywords: Distribution Learning, Social Choice, Information Theory, Ranking Distributions, Kendall-Mallows model, Cayley-Mallows model

Ευχαριστίες

Ολοκληρώνοντας αυτή τη διπλωματική εργασία και, κατ' επέκταση, τις σπουδές μου στο ΕΜΠ, θα ήθελα να ευχαριστήσω τους ανθρώπους που με βοήθησαν να φτάσω ως εδώ.
Αρχικά, θα ήθελα να ευχαριστήσω τα μέλη της τριμελούς επιτροπής. Πρώτα τον κύριο Φωτάκη, που ήταν ο επιβλέποντας αυτής της εργασίας. Η δουλειά αυτή δεν θα μπορούσε να είχε διεκπεραιωθεί χωρίς την καθοδήγησή του. Τον ευχαριστώ, λοιπόν, για τις συμβουλές που μου έδωσε, τόσο πάνω στο επιστημονικό μέρος αυτής της εργασίας, όσο και πέρα από αυτό, καθώς ήταν καθοριστικής σημασίας για τις αποφάσεις μου στο θέμα των αιτήσεων. Μετά, τον κύριο Παγουρτζή, για την δουλειά που έκανε στα μαθήματα των αλγορίθμων κατά το ακαδημαϊκό έτος $2017-2018$, οπότε και ανέλαβε να τα διδάξει. Τέλος, τον κύριο Λουλάκη, που, αν και δεν συνεργαστήκαμε σε επίπεδο διπλωματικής εργασίας, με έκανε μέσω των μαθημάτων του να αγαπήσω τις Πιθανότητες.
Ακόμη, θα ήθελα να ευχαριστήσω τα μέλη του Εργαστηρίου Λογικής και Επιστήμης Υπολογισμών για το ευχάριστο κλίμα συνεργασίας που δημιουργούν. Ιδιαίτερα, θα ήθελα να ευχαριστήσω τον κύριο Ζάχο που με στηρίζει από το 1ο έτος των σπουδών μου. Επίσης, θα ήθελα να ευχαριστήσω όλα τα παιδιά από το shmmy.ntua για τη συνεργασία και τη φιλία τους όλα αυτά τα χρόνια.
Τέλος, θα ήθελα να ευχαριστήσω τους φίλους μου και την οικογένειά μου που με στηρίζουν σε ό,τι κάνω.

Αφιερώνεται στη μνήμη της γιαγιάς μου.

Contents

1 Εκτεταμένη Ελληνική Περίληψη **1**
- 1.1 Εισαγωγή στη Μάθηση Κατανομών . . . 1
  - 1.1.1 PAC-learning . . . 1
  - 1.1.2 Μάθηση Κατανομών . . . 2
  - 1.1.3 Θεωρία Minimax . . . 2
- 1.2 Κατανομές Κατάταξης . . . 3
  - 1.2.1 Το Μοντέλο του Mallows . . . 3
  - 1.2.2 Εκτίμηση Παραμέτρων στο Μοντέλο του Mallows . . . 5
- 1.3 Μάθηση στο Μοντέλο Kendall-Mallows . . . 5
  - 1.3.1 Η Προσέγγιση των Καραγιάννη et. al. . . . 6
  - 1.3.2 Η Προσέγγιση των Busa-Fekete et. al. . . . 6
- 1.4 Μάθηση στο Μοντέλο Cayley-Mallows . . . 6

2 Introduction **9**
- 2.1 Machine Learning . . . 9
- 2.2 Social Choice . . . 10
- 2.3 The Mallows Model . . . 10
- 2.4 Learning the Central Ranking under the Mallows Model . . . 10
- 2.5 Learning the Spread Parameters under the Mallows Model . . . 11
- 2.6 The Cayley-Mallows Model . . . 11
- 2.7 Organization of this Thesis . . . 12

3 Introduction to Distribution Learning **13**
- 3.1 PAC-learning . . . 13
- 3.2 Maximum Likelihood Estimation . . . 14
  - 3.2.1 The Technique . . . 15
  - 3.2.2 Examples . . . 15
  - 3.2.3 MLE vs ERM . . . 16
- 3.3 Concentration Inequalities . . . 17
  - 3.3.1 Markov's Inequality . . . 17
  - 3.3.2 Chebyshev's Inequality . . . 18
  - 3.3.3 Chernoff Bounds . . . 18
  - 3.3.4 Hoeffding's Inequality . . . 20
- 3.4 Information Theory and Statistics . . . 21
  - 3.4.1 Fundamental Information Theoretic Measures . . . 21
  - 3.4.2 f-divergences . . . 25
  - 3.4.3 Inequalities of Information Theory . . . 30

3.4.3.1 Data Processing Inequality . . . 30
3.4.3.2 Le Cam's Inequality . . . 31
3.4.3.3 Fano's Inequality . . . 32
3.5 Distribution Learning and Lower Bounds . . . 33
3.5.1 The Framework . . . 34
3.5.2 Minimax Lower Bounds . . . 34
3.5.2.1 Minimax Risk . . . 35
3.5.2.2 Le Cam's and Fano's Methods . . . 35
3.5.2.3 The Gilbert-Varshamov Bound . . . 36
3.6 Exponential Families . . . 37

4 Permutations and Rankings **41**
4.1 Permutations . . . 41
4.1.1 Permutations as Groups . . . 41
4.1.2 Cyclic Permutations . . . 45
4.2 Permutation Distances . . . 45
4.2.1 Kendall Tau Distance . . . 46
4.2.2 Cayley Distance . . . 48
4.2.3 Other Distances . . . 49
4.3 Ranking Distributions . . . 50
4.3.1 The Mallows Model . . . 50
4.3.2 The Plackett-Luce Model . . . 51
4.4 Parameter Estimation in the Mallows Model . . . 52
4.4.1 The MLE for the Central Ranking . . . 52
4.4.2 The MLE for the Spread Parameters . . . 52

5 Learning in the Kendall-Mallows Model **53**
5.1 Recovering the Central Ranking . . . 53
5.1.1 Concentration in Kendall-Mallows . . . 54
5.1.2 Approximating the MLE . . . 56
5.1.3 Sample Complexity Analysis . . . 56
5.2 Estimating the Spread Parameters . . . 58
5.2.1 From Mallows to Sums of Truncated Geometrics . . . 58
5.2.2 Parameter Estimation in Truncated Geometrics . . . 59
5.2.3 The Block Model . . . 60
5.2.4 Parameter Estimation in the Kendall-Mallows Block Model . . . 61
5.2.5 Distribution Learning in the Kendall-Mallows Block Model . . . 62

6 Learning in the Cayley-Mallows Model **65**
6.1 Recovering the Central Ranking . . . 65
6.1.1 Concentration in Cayley-Mallows . . . 66
6.1.2 Approximating the MLE . . . 68
6.1.3 Sample Complexity Analysis . . . 68

7 Conclusions and Future Work **71**

List of Figures

3.1 Venn diagram of fundamental measures of information theory. 25

5.1 Graph representation of the identity element of S_3. 54

List of Algorithms

1 Condorcet Sample Generation 54
2 MLE approximation via PM-c rules 56
3 Spread Parameter Estimation 61

4 MLE approximation via position majority-consistent rules 68

Chapter 1

Εκτεταμένη Ελληνική Περίληψη

Σε αυτό το κεφάλαιο παρουσιάζουμε περιληπτικά τα περιεχόμενα αυτή της διπλωματικής εργασίας στα ελληνικά. Εισάγουμε όλες τις βασικές έννοιες που εμφανίζονται στο αγγλικό κείμενο. Ωστόσο, δεν δίνουμε ούτε αποδείξεις ούτε τεχνικές λεπτομέρειες. Αυτές δίνονται εκτενώς στα επόμενα κεφάλαια.

1.1 Εισαγωγή στη Μάθηση Κατανομών

Σε αυτή την ενότητα παρουσιάζονται κάποιες γενικές ιδέες που διέπουν την περιοχή της *μάθησης κατανομών*. Αρχικά, πρέπει να τονιστεί ότι η μάθηση κατανομών (ή αλλιώς *εκτίμηση πυκνότητας*) είναι πιο ισχυρή από την *εκτίμηση παραμέτρων*. Συγκεκριμένα, η εκτίμηση παραμέτρων συνίσταται στο να βρει κανείς καλές προσεγγίσεις για τις παραμέτρους μίας κατανομής δοθέντων κάποιων δειγμάτων. Από την άλλη, στην εκτίμηση πυκνότητας, ο στόχος είναι η εύρεση μίας κατανομής που να βρίσκεται κοντά στην πραγματική με βάση κάποια στατιστική απόσταση. Συνεπώς, το να μάθει κανείς μία κατανομή είναι, εν γένει, μία πιο ισχυρή (άρα δυσκολότερη να επιτευχθεί) απαίτηση από το να μάθει απλά κάποιες τιμές που την περιγράφουν.
Στα επόμενα, ξεκινάμε δίνοντας τον ορισμό του *PAC-learning* και ύστερα εξηγούμε τη σχέση του με το πλαίσιο για μάθηση κατανομών που ορίστηκε από τους Kearns et. al. στο [29]. Τέλος, δίνουμε κάποια βασικά στοιχεία της θεωρίας *minimax κάτω φραγμάτων*, τα οποία χρησιμοποιούνται για την αξιολόγηση των διαφόρων αλγορίθμων μάθησης.

1.1.1 PAC-learning

Το PAC-learning είναι ένα πλαίσιο για την θεωρητική μελέτη της *μηχανικής μάθησης* το οποίο ορίστηκε το 1984 από τον Valiant στο [47]. Η ιδέα είναι πως δουλεύουμε σε ένα σύνολο $\mathcal{X}$ στα στοιχεία του οποίου αντιστοιχούν τιμές από ένα σύνολο $\mathcal{Y}$. Η αντιστοίχιση δεν είναι μονοσήμαντη. Αντιθέτως, υπάρχει κάποια κατανομή $\mathcal{D}$ επί του $Z = \mathcal{X} \times \mathcal{Y}$ η οποία μας είναι άγνωστη. Θα θέλαμε να βρούμε κάποια υπόθεση $h : \mathcal{X} \to \mathcal{Y}$, που να περιγράφει όσο γίνεται καλύτερα τη σχέση ανάμεσα στα $x \in \mathcal{X}$ και στα $y \in \mathcal{Y}$, όπως αυτή ορίζεται από την κατανομή $\mathcal{D}$. Η αξιολόγηση των διαφόρων h γίνεται βάσει μίας συνάρτησης σφάλματος ℓ. Θα θέλαμε οι αλγόριθμοι με βάση τους οποίους παράγονται οι υποθέσεις h να δίνουν καλά αποτελέσματα ανεξαρτήτως της υποκείμενης κατανομής. Βάσει των προηγουμένων, ο τυπικός ορισμός του PAC-learning είναι ο εξής:

Definition 1.1.1 (Valiant (1984))**.** Μία κλάση υποθέσεων $\mathcal{H}$ λέμε πως είναι be PAC-learnable ως προς κάποιο πεδίο Z και μία συνάρτηση σφάλματος $\ell : \mathcal{H} \times Z \to \mathbb{R}_{\geq 0}$ αν υπάρχει συνάρτηση $n_{\mathcal{H}} : \mathbb{R}_{\geq 0} \times (0,1) \to \mathbb{N}$ και αλγόριθμος μάθησης $\mathcal{A}$ τέτοιοι ώστε: για οποιαδήποτε $\epsilon > 0, \delta \in (0,1)$ και για οποιαδήποτε κατανομή $\mathcal{D}$ over Z, εκτελώντας τον $\mathcal{A}$ με είσοδο $\boldsymbol{z} \sim \mathcal{D}^n$ όπου $n \geq n_{\mathcal{H}}(\epsilon, \delta)$, παίρνουμε $\hat{h} = \mathcal{A}(\boldsymbol{z})$ τέτοια ώστε:

$$\mathop{\mathbb{P}}_{\boldsymbol{z} \sim \mathcal{D}^n} \left[L_{\mathcal{D}}\left(\hat{h}\right) - \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \geq \epsilon \right] < \delta$$

όπου το $L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}}\left[\ell(h, z)\right]$ είναι το ρίσκο.

Επειδή η $\mathcal{D}$ είναι άγνωστη και λειτουργούμε ανεξαρτήτως αυτής, ο ακριβής υπολογισμός του ρίσκου είναι αδύνατος. Για αυτό, υπολογίζεται ο αντίστοιχος δειγματικός μέσος ως εκτίμηση. Στα πλαίσια της σχεδίασης αλγορίθμων μάθησης, επιχειρούμε συνήθως να ελαχιστοποιήσουμε αυτόν τον δειγματικό μέσο. Αυτή είναι η *αρχή Ελαχιστοποίησης Εμπειρικού Ρίσκου*. Το ενδιαφέρον αυτής της αρχής είναι ότι στην πραγματικότητα συνιστά γενίκευση μίας θεμελιώδους για τη στατιστική τεχνικής: της *Εκτίμησης Μέγιστης Πιθανοφάνειας*.

1.1.2 Μάθηση Κατανομών

Οι Kearns et. al. γενίκευσαν το προηγούμενο πλαίσιο στο [29] ώστε να μπορεί να εκφράσει και προβλήματα μάθησης κατανομών. Συγκεκριμένα, έδωσαν τον ακόλουθο ορισμό:

Definition 1.1.2 (Kearns et. al. (1994))**.** Μία οικογένεια κατανομών $\mathcal{F}$ λέμε πως μπορεί να μαθευτεί αποδοτικά ως προς κάποιο μέτρο απόκλισης d όταν, για κάθε $\epsilon > 0, \delta \in (0,1)$, έχοντας πρόσβαση σε δείγματα από μία άγνωστη κατανομή $P \in \mathcal{F}$, υπάρχει αλγόριθμος πολυωνυμικού χρόνου $\mathcal{A}$ που δίνει μία κατανομή $\hat{P}$ τέτοια ώστε:

$$\mathbb{P}\left[d\left(\hat{P}, P\right) \geq \epsilon \right] < \delta$$

όπου η πιθανότητα υπολογίζεται ως προς τα δειγματα.
Αν $\hat{P} \in \mathcal{F}$, τότε ο $\mathcal{A}$ λέγεται proper αλγόριθμος. Αλλιώς, λέγεται improper αλγόριθμος.

Το μέτρο απόκλισης d είναι συνήθως είτε η TV-distance είτε η KL-divergence (ή, σπανιότερα, η απόσταση Kolmogorov). Σημειώνεται πως, στην περίπτωση που μας ενδιαφέρουν proper προβλήματα μάθησης (εν αντιθέσει πχ με το [13]) και δεν έχουμε αλλοιωμένα δείγματα (εν αντιθέσει πχ με το [18]), υπάρχει μία πολύ απλή γενίκευση της αρχής Ελαχιστοποίησης Εμπειρικού Ρίσκου: πάρε την κατανομή αντιστοιχεί στη λύση μέγιστης πιθανοφάνειας για τις παραμέτρους της κατανομής. Η θεωρητική θεμελίωση αυτής της ιδέας δίνεται στο κεφάλαιο 24 του [6].

1.1.3 Θεωρία Minimax

Προκειμένου να αξιολογήσουμε έναν αλγόριθμο μάθησης, μας ενδιαφέρει ποια είναι η κατανομή για την οποία έχει τη χειρότερη απόδοση. Μεταξύ των διαφορών αλγορίθμων, μας ενδιαφέρει

εκείνος που, σε αυτή την περίπτωση, έχει την καλύτερη απόδοση. Αυτό αποτελεί το κίνητρο για την εισαγωγή του minimax ρίσκου, το οποίο ορίζεται ως εξής:

$$\mathcal{R}_n(\mathcal{F}) = \inf_{\hat{f}\in\Omega} \sup_{f\in\mathcal{F}} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{F}^n} \left[d_{TV}\left(\hat{f}(\boldsymbol{x}), f\right)\right]$$

Το να γίνει ακριβής υπολογισμός του minimax ρίσκου συχνά είναι δύσκολο. Γι' αυτό, συνήθως επιλέγεται ένα πεπερασμένο υποσύνολο της $\mathcal{F}$ που να αναδεικνύει τη δυσκολία του προβλήματος και βάσει αυτού υπολογίζεται ένα κάτω φράγμα για το minimax ρίσκο. Οι κατανομές αυτού του υποσυνόλου πρέπει να έχουν άνω φραγμένο KL-divergence και κάτω φραγμένη TV-distance. Αν το υποσύνολο που επιλέγεται έχει 2, έχουμε:

Proposition 1.1.1 (Le Cam, Pinsker)**.** *Δοθέντος ενός ζεύγους κατανομών $P_1, P_2 \in \mathcal{F}$ με φορέα $\mathcal{X}$ που ικανοποιούν $d_{TV}(P_1, P_2) \geq a$ και $D_{KL}(P_1||P_2), D_{KL}(P_2||P_1) \leq b$, έχουμε:*

$$\mathcal{R}_n(\mathcal{F}) \geq \frac{a}{4}\left(1 - ||\mathcal{F}_1^n - \mathcal{F}_2^n||_{TV}\right) \geq \frac{a}{4}\left(1 - \sqrt{\frac{n}{2}b}\right)$$

Αν έχει περισσότερα των 2 στοιχείων, έχουμε:

Proposition 1.1.2 (Yu (1997))**.** *Έστω $\mathcal{F}$ πεπερασμένη οικογένεια πυκνοτήτων με:*

$$\inf_{f,g\in\mathcal{F}:f\neq g} d_{TV}(f,g) \geq a, \quad \sup_{f,g\in\mathcal{F}:f\neq g} D_{KL}(f||g) \leq b$$

τότε ισχύει ότι:

$$\mathcal{R}_n(\mathcal{F}) \geq \frac{a}{2}\left(1 - \frac{nb + \ln(2)}{\ln(|\mathcal{F}|)}\right)$$

1.2 Κατανομές Κατάταξης

Σε αυτή την ενότητα εισάγονται οι *κατανομές κατάταξης*. Η ιδέα πίσω από αυτές προήλθε από την *θεωρία κοινωνικής επιλογής*. Συγκεκριμένα, στην θεωρία κοινωνικής επιλογής, υποτίθεται πως, όταν υπάρχουν κάποιες διαθέσιμες επιλογές, υπάρχει ένα αντικειμενικός τρόπος να τις κατατάξει κανείς που όμως είναι άγνωστος. Πέρα από αυτό, υπάρχουν οι υποκειμενικές πεποιθήσεις του καθενός, που μοντελοποιούνται ως θορυβώδεις εκτιμήσεις της προηγούμενης κατάταξης. Στόχος αποτελεί εν γένει η ανάκτηση της προηγούμενης "αντικειμενικής" κατάταξης.
Στα πλαίσια του προηγούμενου στόχου, έχουν αναπτυχθεί διάφορα θορυβοποιά μοντέλα. Αυτό που μας ενδιαφέρει κατά βάση είναι το μοντέλο του Mallows, το οποίο και εισάγουμε παρακάτω.

1.2.1 Το Μοντέλο του Mallows

Η ιδεά πίσω από το μοντέλο του Mallows ήταν να οριστεί ως ένα ανάλογο της κανονικής κατανομής για ranking. Έτσι, επιλέχθηκε η σμπ:

$$\mathbb{P}[\pi = \sigma] = \frac{1}{Z(\phi)} \phi^{d(\sigma, \pi_0)}$$

όπου $\pi_0 \in S_m$, $\phi \in (0, 1)$, d: απόσταση μεταξύ μεταθέσεων. Παρατηρήστε ότι η σταθερά κανονικοποίησης $Z(\phi)$ είναι ανεξάρτητη του π_0. Έχουν προταθεί διάφορες αποστάσεις για το ρόλο της d. Κάποιες από αυτές είναι:

- Kendall's tau (αριθμός αντιμεταθέσεων ανάμεσα σε γειτονικά στοιχεία για να μετατραπεί η π στην π_0).
- Cayley distance (αριθμός αντιμεταθέσεων για να μετατραπεί η π στην $\pi_0 = m$ - # of cycles in $\pi\pi_0^{-1}$).
- Spearman's footrule και Spearman's rank correlation (απόσταση ℓ_1 και τετράγωνο της απόστασης ℓ_2 των π, π_0 όταν αναπαρίστανται ως διανύσματα).
- Hamming distance (αριθμός στοιχείων που βρίσκονται σε λάθος θέση).

Το ενδιαφέρον μας εστιάζεται στις 2 πρώτες. Ο λόγος είναι ότι για αυτές υπάρχει μία κομψή παραγοντοποίηση για τη σταθερά κανονικοποίησης που προκύπτει, ενώ ειδικά η 1η είναι εκείνη που χρησιμοποιείται κυρίως στον τομέα της κοινωνικής επιλογής. Συγκεκριμένα, έχουμε:

- KT-distance:

$$Z(\phi) = \prod_{i=1}^{m} Z_i(\phi) = \prod_{i=1}^{m} \left(\sum_{j=0}^{i-1} \phi^j \right)$$

- Cayley distance:

$$Z(\phi) = \prod_{i=1}^{m} Z_i(\phi) = \prod_{i=1}^{m} [1 + (m - i)\phi]$$

Με βάση τα παραπάνω, οι Flinger και Verducci πρότειναν στο [22] να γενικευθεί το μοντέλο για αυτές τις 2 αποστάσεις αντιστοιχώντας διαφορετικό spread parameter σε κάθε εναλλακτική. Έτσι, προέκυψε το γενικευμένο μοντέλο του Mallows με σμπ:

$$\mathbb{P}[\pi = \sigma] = \prod_{i=1}^{m} \frac{\phi_i^{V_i(\pi, \pi_0)}}{Z_i(\phi_i)}$$

όπου τα Z_i είναι όπως παραπάνω και έχουμε:

- KT-distance: $V_i(\pi, \pi_0) = \sum_{j=0}^{i-1} \mathbb{1}\{(\pi(i) - \pi(j))(\pi_0(i) - \pi_0(j)) < 0\}$.
- Cayley distance: το $V_i(\pi, \pi_0)$ γίνεται 0 όταν το στοιχείο i είναι εκείνο με το μεγαλύτερο δείκτη στον κύκλο όπου ανήκει στην $\pi\pi_0^{-1}$.

Τέλος, αν για κάποια από τα στοιχεία γνωρίζουμε εκ των προτέρων πως τους αντιστοιχεί το ίδιο spread parameter, ορίζουμε το Mallows block model. Εκεί, το σύνολο των στοιχείων διαμερίζεται σε d blocks, όπου τα στοιχεία του ιδίου block μοιράζονται το ίδιο spread parameter. Έτσι, η σμπ γίνεται:

$$\mathbb{P}\left[\pi = \sigma\right] = \prod_{i=1}^{d} \frac{\phi_i^{T_i(\pi, \pi_0, \boldsymbol{B})}}{Z^i\left(\phi_i, \boldsymbol{B}\right)}$$

όπου:

$$T_i\left(\pi, \pi_0, \boldsymbol{B}\right) = \sum_{j \in B_i} V_j\left(\pi, \pi_0\right)$$

$$Z^i\left(\phi_i\right) = \prod_{j \in B_i} Z_j\left(\phi_i\right)$$

1.2.2 Εκτίμηση Παραμέτρων στο Μοντέλο του Mallows

Με βάση τα προηγούμενα, είναι σαφές πως θέλουμε να εστιάσουμε σε προβλήματα μάθησης. Έτσι, ένα σημαντικό πρώτο βήμα είναι να γίνει ο υπολογισμός των εκτιμητριών μέγιστης πιθανοφάνειας για το μοντέλο του Mallows. Εδώ δίνουμε τις εκτιμήτριες για το απλό μοντέλο.
Για την κεντρική κατάταξη, έχουμε:

$$\hat{\pi}_0 = \underset{\pi_0}{argmin} \left\{ \sum_{i=1}^{n} d\left(\pi_i, \pi_0\right) \right\}$$

που είναι η διάμεσος των $\pi_1, \ldots, \pi_n$ ως προς την απόσταση d. Για το παραπάνω ξέρουμε πως:

- έχει αποδειχθεί πως είναι NP-Hard για την KT-distance (βλ. [5]).
- πιστεύεται πως είναι NP-Hard για την Cayley distance (βλ. [44]).

Για το spread parameter, έχουμε:

$$\hat{\phi} \frac{Z'\left(\hat{\phi}\right)}{Z\left(\hat{\phi}\right)} = \frac{1}{m} \left(\sum_{i=1}^{n} d\left(\pi_i, \pi_0\right) \right)$$

που είναι αδύνατο να λυθεί ακριβώς ως προς $\hat{\phi}$.
Σύμφωνα με τα προηγούμενα, καθίσταται σαφές πως η εκτίμηση των παραμέτρων του μοντέλου του Mallows δεν είναι καθόλου τετριμμένο πρόβλημα.

1.3 Μάθηση στο Μοντέλο Kendall-Mallows

Σε αυτή την ενότητα εστιάζουμε σε προβλήματα μάθησης διατυπωμένα στην περίπτωση που η απόσταση μεταξύ των μεταθέσεων είναι η KT-distance. Για τη συγκεκριμένη απόσταση, έχουν δοθεί βέλτιστες λύσεις τόσο για το πρόβλημα της εκτίμησης της κεντρικής κατάταξης, όσο και για το πρόβλημα της εκτίμησης των spread parameters. Το μεν πρώτο λύθηκε από τους Καραγιάννη et. al. στο [10], το δε δεύτερο από τους Busa-Fekete et. al. στο [9].

1.3.1 Η Προσέγγιση των Καραγιάννη et. al.

Οι Καραγιάννης et. al. βασίστηκαν για να λύσουν το πρόβλημα σε ένα μοντέλο προγενέστερο του Mallows. Αυτό ήταν το μοντέλο των θορυβωδών συγκρίσεων του Condorcet, το οποίο αποδείχθηκε πως είναι ισοδύναμο με το μοντέλο του Mallows. Αυτή η ισοδυναμία τους ώθησε να ορίσουν την οικογένεια των pairwise majority consistent rules (PM-c rules). Συγκεκριμένα, πρόκειται για μία οικογένεια αλγορίθμων που επιχειρεί να ανακτήσει την κεντρική κατάταξη θεωρώντας μία εκτίμηση όπου επιχειρείται, στο βαθμό που γίνεται, οι συγκρίσεις των στοιχείων ανά 2 να είναι όπως στην πλειοψηφία των δειγμάτων.

Theorem 1.3.1 (CPS13). *Για κάθε $\epsilon \in (0,1]$, οποιοσδήποτε PM-c rule προσδιορίζει την πραγματική κατάταξη με πιθανότητα τουλάχιστον $1-\epsilon$ δοθέντων $\mathcal{O}\left(\log\left(\frac{m}{\epsilon}\right)\right)$ δειγμάτων από ένα απλό μοντέλο Kendall-Mallows.*

Ύστερα, έδειξαν πως η παραπάνω δειγματική πολυπλοκότητα είναι βέλτιστη.

1.3.2 Η Προσέγγιση των Busa-Fekete et. al.

Οι Busa-Fekete et. al. αντιμετώπισαν το ζήτημα της εκτίμησης των spread parameters λύνοντας ένα φαινομενικά δυσκολότερο πρόβλημα. Συγκεκριμένα, αντί να δουλέψουν στο απλό μοντέλο του Mallows, δούλεψαν στο γενικευμένο. Αν και το μοντέλο αυτό είναι εν γένει πιο σύνθετο, το να συλλάβει κανείς τη λύση σε αυτή την περίπτωση είναι στην πραγματικότητα πιο εύκολο. Συγκεκριμένα, θεωρούμε τις τυχαίες μεταβλητές $Y_i = V_i(\pi, \pi_0)$. Η περιθώρια κατανομή καθεμίας εξ' αυτών είναι:

$$\mathop{\mathbb{P}}_{\pi\sim\mathcal{P}_{\boldsymbol{\phi},\pi_0}}[Y_i = k_i] = \frac{\phi_i^{k_i}}{Z_i(\phi_i)}, k_i \in \{0,1,\dots i-1\}$$

Η παραπάνω είναι στην πραγματικότητα παραλλαγή μίας γνωστής κατανομής. Συγκεκριμένα, πρόκειται για την κατανομή που προκύπτει αν πάρουμε μία γεωμετρική και απαιτήσουμε οι τιμές της να μην είναι μεγαλύτερες από $i-1$ (συμβολίζεται $\mathcal{TG}(\phi_i, i-1)$). Εφόσον έχουμε βρει μία κατανομή με απλούστερη μορφή που παραμετροποιείται από την ϕ_i, πλέον η εκτίμηση της ϕ_i είναι αρκετά απλούστερη υπόθεση.
Το ερώτημα που γεννάται τώρα είναι τι γίνεται αν γνωρίζουμε πως κάποια από τα ϕ_i ταυτίζονται (όπως συμβαίνει στην περίπτωση του block model). Η απάντηση είναι ότι τότε μπορούμε να συμψηφίσουμε δειγματικές τιμές που αντιστοιχούν κανονικά σε διαφορετικά Y_i και να πάρουμε έτσι καλύτερα αποτελέσματα και με λιγότερα δείγματα.
Οι παραπάνω ιδέες οδηγούν σε βέλτιστη πολυπλοκότητα δείγματος, τόσο για την εκτίμηση των spread parameters, όσο και για τη μάθηση της αντίστοιχης κατανομής. Επιπλέον, αν το μέγεθος του μικρότερου block τείνει στο άπειρο, καθίσταται εφικτή η εκτίμηση ακόμη και από ένα μεμονωμένο δείγμα (όπως με την εκτιμήτρια που δίνεται στο [41]).

1.4 Μάθηση στο Μοντέλο Cayley-Mallows

Η παρούσα εργασία συμπληρώνει την προηγούμενη δουλειά εξετάζοντας την περίπτωση που η KT-distance αντικατασταθεί με την Cayley distance. Συγκεκριμένα, η Cayley distance θεωρείται πιο ιδιόρρυθμη διότι, αντί να εκφράζει μία απλή λογική (όπως η KT-distance, που εστιάζει στις

συγκρίσεις μεταξύ ζευγών), εκφράζει μαθηματικές έννοιες σχετικές με την κυκλική δομή των μεταθέσεων. Παρά τις ιδιαιτερότητες αυτές, δείχνουμε πως, το μοντέλο του Mallows εφοδιασμένο με την Cayley distance, επιδεικνύει και αυτό συγκέντρωση γύρω από την κεντρική κατάταξη, αλλά αρκετά ασθενέστερη σε σχέση με αυτή στο Kendall-Mallows. Συγκεκριμένα, δείχνουμε πως, εξετάζοντας την θέση όπου εμφανίζεται συχνότερα ένα στοιχείο, μπορούμε, με κατάλληλο αριθμό δειγμάτων, να ανακτήσουμε την κεντρική κατάταξη με μεγάλη πιθανότητα. Αυτό διατυπώνεται ως εξής:

Theorem 1.4.1 (Ανεπίσημο)**.** *Αρκούν* $\mathcal{O}\left(m^2 \log(m)\right)$ *προκειμένου να ανακτήσουμε την κεντρική κατάταξη στο απλό μοντέλο Cayley-Mallows με μεγάλη πιθανότητα.*

Chapter 2

Introduction

The aim of this chapter is to motivate the next by providing an overview of the areas that have inspired the topic of this thesis. Specifically, for the most part, our work focuses on issues such as parameter and density estimation, which are topics that fall into the domain of statistics and machine learning theory. However, the probabilistic models associated with the problems we examine are inspired from social choice theory. For that reason, we will make a short introduction to the above subjects and then refer to previous work in this area.

2.1 Machine Learning

Machine learning is a sub-field of artificial intelligence that has attracted a great deal of interest over the course of the last 10 years. Its aim is to provide a mathematical framework that adequately explains the processes based on which the human brain assimilates knowledge. Once this has been achieved, it will be possible to construct algorithms based on those processes, thus rendering possible the creation of computers that are able to think and adapt to change. The reason this area has attracted such a great deal of interest recently has to do with the impressive (and theoretically inexplicable) performance of neural networks on a number of learning tasks. However, this makes it quite easy to forget that there was an extensive body of work on the topic long before the hardware that made possible the use of neural networks was developed. A lot of that work focused on the mathematical foundations of machine learning. Specifically, the mathematical subjects on which machine learning theory mostly relies on are statistics and mathematical optimization. In particular, statistics offers the necessary framework to describe the learning process mathematically. However, the various statistical models are parameterized by specific quantities. Having chosen a specific model, it is necessary to chose the parameter values that best describe the ground truth, which is where the techniques offered by optimization come in.
Apart from the aforementioned reliance on established mathematical disciplines, machine learning theory also relies on the computational framework defined by Valiant in [47]. In particular, Valiant defined the *PAC (Probably Approximately Correct)* framework for *supervised* learning problems. In that context, each element of some instance space is associated with a value based on a rule that is unknown and learning algorithms are fed with samples that consist of elements with their associated values and attempt to determine that association rule. The evaluation of algorithms is based on the number of samples they require to estimate the target rule within a given error bound and with a given maximum probability of error, regardless of the underlying distribution according to which the samples are generated. A similar framework

was proposed 10 years later in [29] for *distribution learning* problems, which are prime examples of *unsupervised* learning problems. This last framework is the one we are most interested in. In particular, we focus on distributions defined over permutation groups, which are sets that are characterized by a *combinatorial structure*. Apart from their mathematical importance, they boast an interpretation as rankings in social choice theory.

2.2 Social Choice

Social choice theory is field that is preoccupied to the study of voting rules and electoral systems. Its origins can be traced back to the late 18th and early 19th centuries, during which a number of scholars developed an interest in the theoretical study of the aforementioned subjects, in an attempt to construct electoral systems which would result in the aggregation of individual opinions in a manner that best represents society's views as a whole. For more information concerning the origins of social choice, a good resource is [51].
Various approaches have been suggested in order to tackle the above issues (see [3]). The one we are most interested in is that which views the social optimum as a ranking of the various alternatives and the individual opinions as noisy estimates of the previous ranking that are generated by some probabilistic model. In that context, voting rules essentially serve as estimators in the statistical sense of the term and the best voting rule that can be conceived for a specific problem is the one corresponding to the maximum likelihood estimator.
Work in the area continues to this day, with greater emphasis on the computational aspects of the theory. The focus on those issues is motivated by the progress in theoretical computer science, the main objective of which is the development of efficient procedures to solve computational problems. Applying the techniques of computer science to the problems studied by social choice theory, it is possible to get good approximations of the results of various voting rules in a time and sample efficient way, which is one of the issues examined in later chapters.

2.3 The Mallows Model

The *Mallows model* (defined in [36]) is one of the most popular probabilistic models on rankings. It is a distance based ranking model, where rankings that are closer to some ranking with respect to some distance metric tend to be favored. The distance is usually the *Kendall Tau distance*, which measures the number of inversions in a pair of rankings, thus being suitable to express notions in the context of social choice theory. The random behavior exhibited by the model is expressed through one or more spread parameters. Learning problems in this model have attracted a great deal of interest over the years, both from the statistics and computer science communities (due to the challenges posed by the combinatorial nature of the model) and from the social choice community.

2.4 Learning the Central Ranking under the Mallows Model

In the case of the central ranking, it was shown in [5] that it is *NP-hard* to accurately compute the maximum likelihood solution with reduction from the *Feedback Arc Set problem*. Due to this negative result, attempts have been made over the years to work around the computational intractability of the MLE. Some of those approaches were experimental in nature, such as the the branch and bound technique presented in [38], which relied on the empirical observation

that the probability mass tends to concentrate around the central ranking. On a more theoretical note, it was shown in [1] that the problem admits a $\frac{11}{7}$ polynomial time approximation algorithm and a PTAS in [30].
One of the most interesting approaches was given in [8]. The algorithm of Braverman and Mossel relied on the locality induced by the KT distance. Specifically, they showed that, given a sample from the model, the position where an element appears is close to its original position with high probability. As a result, taking an average of each element's position in the samples (and breaking ties) results in a ranking where each element should be at distance $\Theta\left(\frac{\log(n)}{r}\right)$ with high probability (where n denotes the number of alternatives and r denotes the number of samples). Finally, they propose a dynamic programming algorithm which exploits the previous situation, thus outputting a ranking that is close to the MLE and, consequently, to the true ranking. This approach is notable due to the fact that, unlike most learning algorithms, the runtime decreases as the number of available samples increases, because the increase in the number of the samples results in a decrease in the area that has to be searched.
The approach that has the greatest influence to our work is that of Caragiannis et. al. in [10]. There, the authors exploit the fact that each comparison is preserved with probability at least $\frac{1}{2}$ and propose a technique to reconstruct the central ranking by observing the way each pair of elements compares in the majority of samples. Then, they show that this technique indeed results in the recovery of the central ranking with high probability.

2.5 Learning the Spread Parameters under the Mallows Model

The estimation of the spread parameters poses an equally big challenge, this time due to the fact that it is impossible to obtain a closed form for the maximum likelihood solution. However, the work regarding the estimation of the spread parameters is less extensive than that involving the estimation of the central ranking. That is because the spread parameters are not of similar importance for social choice theory. Some have attempted to address the problem from a practical angle, not providing any theoretical guarantees for their methods (see [38]).
One important theoretically oriented work is that of Mukherjee in [41]. There, an approximation is given for the normalizing constant of the simple Kendall-Mallows model which holds when the number of alternatives tend to infinity. This facilitates the computation of the MLE, provided that the central ranking is known. This approximation is shown to be consistent, while it is shown that this estimator makes it possible to estimate the spread parameter even from a single sample as the number of alternatives tend to infinity.
The first optimal result involving the estimation of the spread parameters of the Mallows model was given in [9]. There, the authors introduced the *Mallows Block model* and managed to reduce the problem of estimating the spread parameters to estimating the parameters of *truncated geometric distributions*. Moreover, they showed that, if the minimum block size tends to infinity, it is possible to estimate the spread parameters even from a single sample.

2.6 The Cayley-Mallows Model

A less frequently examined version of the Mallows model is the one where the distance metric used to measure the distance between permutations is the *Cayley distance*. Whereas the KT distance focuses on pair-wise comparisons, the Cayley distance focuses on the cyclic structure of permutations, which makes it more difficult to study and less suitable applications in social

choice theory, though an application in computational biology was given in [27]. Another interesting thing about the Cayley-Mallows model (which serves as the motivation for our work) is the difference noted in the results of [34] (which extended the previous work of [4]) and [14], which examine the problems of learning mixtures of Kendall-Mallows and Cayley-Mallows models, respectively.

2.7 Organization of this Thesis

In chapter 3, we make a lengthy introduction to distribution learning theory, introducing the basic framework, along with all the necessary background from probability, statistics and information theory.
In chapter 4, we introduce the notion of permutations and explain both their theoretical value (in group theory) and their practical value (in modelling rankings). Moreover, we introduce the concept of permutation distances and then we present the Mallows model. The chapter closes with a section that presents the maximum likelihood solutions for parameter estimation problems under the Mallows model, thus setting the tone for the chapters that follow.
In chapter 5, we present the techniques of Caragiannis et. al. and Busa-Fekete et. al. from [10] and [9], respectively. These techniques are the ones that have been the most influential to our work.
In chapter 6, we present our work involving parameter estimation under the Mallows model equipped with the Cayley distance. We give a polynomial time algorithm that provably learns the central ranking with high probability. We perform the sample complexity analysis of the algorithm, but we do not prove its optimality. Moreover, we show how we could

Chapter 3

Introduction to Distribution Learning

In this chapter, which will be the lengthiest of this thesis, we will make an introduction to distribution learning, along with all the necessary background. We will start with an introduction to PAC-learning and refer to its connection with maximum likelihood estimation. Then we will proceed with concentration inequalities. After that, we will talk about information theory and statistical distances with applications to distribution learning and lower bounds. The chapter will close with the definition and the basic properties of exponential families.

3.1 PAC-learning

PAC-learning (Probably Approximately Correct learning) is a framework used to study machine learning problems within the context of theoretical computer science. It was introduced by Valiant in [47] in order to help formalize the study of machine learning algorithms. Here, we will make a short introduction to the topic. For a more lengthy examination, the reader should turn to chapter 3 of [6].
Suppose that we have a pair of domains $\mathcal{X}, \mathcal{Y}$ and an unknown distribution $\mathcal{D}$ defined on $Z = \mathcal{X} \times \mathcal{Y}$. The distribution $\mathcal{D}$ can be written in the form $\mathcal{D}_x \cdot \mathcal{D}\left((x,y)\,|x\right)$ where $\mathcal{D}_x$ denotes the marginal distribution over the elements of $\mathcal{X}$ and $\mathcal{D}\left((x,y)\,|x\right)$ denotes the conditional distribution of the elements of $\mathcal{Y}$ given $x \in \mathcal{X}$. Based on the way the distribution $\mathcal{D}\left((x,y)\,|x\right)$ concentrates around values of $\mathcal{Y}$ for the various $x \in \mathcal{X}$, we would like to come up with an association rule $h : \mathcal{X} \to \mathcal{Y}$ which captures best the way the elements of the 2 domains are associated based on $\mathcal{D}\left((x,y)\,|x\right)$. To do this, we are given a number of samples $z_i \in Z$ generated by the unknown $\mathcal{D}$. If we do not restrict ourselves to some subset of $\{f | f : \mathcal{X} \to \mathcal{Y}\}$, we are bound to be faced with the issue of *overfitting*, resulting in bad performance outside the training set. For that reason, we restrict ourselves to some $\mathcal{H} \subset \{f | f : \mathcal{X} \to \mathcal{Y}\}$ which is referred to as a *hypothesis class*. Given some loss function $\ell : \mathcal{H} \times Z \to \mathbb{R}_{\geq 0}$, we would like to find an algorithm that finds the hypothesis $h \in \mathcal{H}$ which, given the available samples, seems to fit best.
The previous setting closely resembles that of an optimization problem. This is to be expected, since optimization is an important tool in machine learning theory. However, our description fails to take into account the inherent randomness of the problem, which is due to the sample generation process. For that reason, instead of simply demanding that ℓ is minimized, our aim will be to come up with an algorithm finds the $h \in \mathcal{H}$ that minimizes the expected loss

(hence referred to as *risk*) given $z \sim \mathcal{D}$. However, this too is way too strict, again due to the sample generation process. Indeed, we cannot guarantee that the available samples will be representative of the underlying distribution, meaning that it is impossible to rule out the possibility of a sub-optimal solution. Relaxing our demand, we would like to come up with an algorithm which, given an adequate number of samples, can compute a hypothesis whose risk is at most $\epsilon > 0$ greater than the optimal (ϵ is referred to as an *accuracy parameter*) with probability at least $1 - \delta$ (*confidence parameter*). If the previous is possible for a hypothesis class, we say that it is PAC-learnable. Formally:

Definition 3.1.1 (Valiant (1984))**.** A hypothesis class $\mathcal{H}$ is said to be PAC-learnable with respect to some domain Z and a loss function $\ell : \mathcal{H} \times Z \to \mathbb{R}_{\geq 0}$ if there exists a function $n_{\mathcal{H}} : \mathbb{R}_{\geq 0} \times (0,1) \to \mathbb{N}$ and a learning algorithm $\mathcal{A}$ such that: given any $\epsilon > 0, \delta \in (0,1)$ and for any distribution $\mathcal{D}$ over Z, running $\mathcal{A}$ with input $\boldsymbol{z} \sim \mathcal{D}^n$ with $n \geq n_{\mathcal{H}}(\epsilon, \delta)$, we get $\hat{h} = \mathcal{A}(\boldsymbol{z})$ such that:

$$\mathop{\mathbb{P}}_{\boldsymbol{z} \sim \mathcal{D}^n} \left[L_{\mathcal{D}}\left(\hat{h}\right) - \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \geq \epsilon \right] < \delta$$

where $L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$ is the risk.

If a hypothesis class is PAC-learnable, it means that we can attain arbitrary precision levels (approximately) with arbitrarily high probability (probably). This explains the choice of name. The function $n_{\mathcal{H}}$ is referred to as the *sample complexity* of the class. Problems with high sample complexity are considered harder from a computational viewpoint. Note that this is referred to in [6] as *Agnostic PAC-learning*, due to the fact that generally $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \neq 0$, which complicates the definition. Moreover, when the concept was originally defined in [47], there were restrictions on the running time of the learning algorithm. In [6], that version of the definition is referred to as *efficient PAC-learnability*. It is obvious from the previous definition that we want the performance of the learning algorithm to be independent of the underlying distribution. However, due to that fact, it is impossible to compute the risk function. For that reason, we define the *empirical risk*, which is nothing but the mean of the values of the loss function evaluated on the samples. Relying on the *Law of Large Numbers (LLN)*, we expect that to be a good approximation of the true mean. For that reason, we pick one of the hypotheses that minimize the empirical risk and then prove that the desired results are achieved. This reasoning describes a fundamental principle known as *Empirical Risk Minimization (ERM)*.

3.2 Maximum Likelihood Estimation

We now move away from machine learning and examine more traditional topics in statistics. A fundamental problem in that area is *parameter estimation*, which involves being given samples from a distribution that belongs to a known family but whose parameters are unknown. The aim is to use a number samples drawn from the distribution to produce an estimate that is as close as possible to the parameter and with as high probability as possible. There are various estimators which are suitable to different versions of the problem. For example, in cases where a percentage of the available samples may have been corrupted, one would wish to use estimators that do not depend on outliers (samples whose values greatly deviate from

the rest). The previous idea lies at the foundation of *robust statistics* (see for example [18]). However, such cases will not be examined in this thesis. Instead, we focus on the classical version of the problem, where the samples can be considered to be independent and generated by the distribution without any sort of interference. In this setting, the most intuitive approach is none other than *maximum likelihood estimation*.

3.2.1 The Technique

Let $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$ be idd samples generated according to a distribution $\mathcal{P}$ described by a parameter vector $\boldsymbol{\theta}$ and let $f(\boldsymbol{x}|\boldsymbol{\theta})$ be the corresponding pdf. We define the function $L(\boldsymbol{\theta}|\boldsymbol{X_1^n}) = \prod_{i=1}^{m} f(\boldsymbol{X_i}|\boldsymbol{\theta})$, which can be considered to be a measure of the likelihood of the occurrence of the available samples (hence the name likelihood function). At this point, a natural approach to determine the value of $\boldsymbol{\theta}$ is by maximizing the likelihood function. Consequently, the problem we wish to solve is $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{argmax}\{L(\boldsymbol{\theta}|\boldsymbol{X_1^n})\}$ which is achieved by computing the points for which $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{X_1^n}) = \boldsymbol{0}$ and verifying that they are indeed maximizers of the function.
We will now mention a number of properties which we would like an estimator to have and explain which hold for maximum likelihood estimators.

Let θ be some parameter and $\hat{\theta}_n$ be an estimator for it which uses n samples.

- **Unbiasedness:** The bias of $\hat{\theta}_n$ is $\mathbb{E}\left[\hat{\theta}_n\right] - \theta$. If the bias of an estimator is 0, we say that it is *unbiased*. Some estimators are unbiased only as $n \to \infty$. Such estimators are referred to as *asymptotically unbiased*. We usually prefer unbiased estimators over the rest, though that is not always the case.
- **Consistency:** We say that $\hat{\theta}_n$ is *consistent* if the sequence $\left\{\hat{\theta}_n\right\}_{n\in\mathbb{N}}$ converges in probability to θ [1]. Unlike unbiasedness, which is merely desirable, consistency is considered to be essential.
- **Mean Square Error:** The *Mean Square Error (MSE)* is defined as $\mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right]$. It holds that $\mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right] = \mathrm{Var}\left(\hat{\theta}_n\right) + \left(\mathbb{E}\left[\hat{\theta}_n\right] - \theta\right)^2$, so if an estimator is unbiased, its MSE is equal to its variance. The MSE of unbiased estimators is lower bounded by the *Cramér–Rao bound* (see chapter 12 of [11]). The closer the MSE of an estimator is to the bound, the better.

When it comes to MLEs, they may not always be unbiased, but they are asymptotically unbiased. Also, they are consistent and, as $n \to \infty$, they attain the Cramér–Rao bound. These properties justify why they are generally preferred over other estimators.

3.2.2 Examples

We will now examine a number of examples involving the computation of MLEs.

[1] A sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ *converges in probability* to some random variable X if $\mathbb{P}[|X_n - X| \geq \epsilon] \to 0, \forall \epsilon > 0$. This is denoted $X_n \xrightarrow{P} X$.

Example 3.2.2.1. Let $X \sim Be(p), p \in (0,1)$ be a Bernoulli random variable. Its pmf is $f(x|p) = p^x (1-p)^{1-x}, x \in \{0,1\}$. Given idd samples $X_1, \ldots, X_n$ from the distribution, we have:

$$\underset{p}{argmax} \{L(p|X_1^n)\} = \underset{p}{argmax} \prod_{i=1}^{n} p^{X_i} (1-p)^{1-X_i} =$$

$$= \underset{p}{argmax} \left(p^{\sum_{i=1}^n X_i} (1-p)^{n-\sum_{i=1}^n X_i} \right) = \underset{p}{argmax} \left(\ln(p) \sum_{i=1}^{n} X_i + \ln(1-p) \left(n - \sum_{i=1}^{n} X_i \right) \right)$$

Demanding the derivative with respect to p to be equal to 0, we have:

$$\frac{\sum_{i=1}^n X_i}{\hat{p}} - \frac{n - \sum_{i=1}^n X_i}{1-\hat{p}} = 0 \iff \boxed{\hat{p} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}_n}$$

The second order derivative is negative, so $\hat{p}$ is indeed a maximizer of the likelihood function. Remark that the parameter we estimated is equal to the mean and the MLE is equal to the sample mean, meaning that our result is consistent with the LLN.

Example 3.2.2.2. Let $X \sim Ge(p), p \in (0,1)$ be a geometric random variable. Given idd samples $X_1, \ldots, X_n$ from the distribution, we have:

$$\underset{p}{argmax} \{L(p|X_1^n)\} = \underset{p}{argmax} \prod_{i=1}^{n} (1-p)^{X_i - 1} p = \underset{p}{argmax} \left((1-p)^{\sum_{i=1}^n X_i - n} p^n \right) =$$

$$= \underset{p}{argmax} \left(\ln(p) n + \ln(1-p) \left(\sum_{i=1}^{n} X_i - n \right) \right)$$

Demanding the derivative with respect to p to be equal to 0, we have:

$$\frac{n}{\hat{p}} - \frac{\sum_{i=1}^n X_i - n}{1-\hat{p}} = 0 \iff \boxed{\hat{p} = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\bar{X}_n}}$$

The second order derivative is negative, so $\hat{p}$ is indeed a maximizer of the likelihood function. Moreover, if we wanted the MLE for the mean, it holds that $\hat{\mu} = \widehat{\left(\frac{1}{p}\right)} = \frac{1}{\hat{p}} = \bar{X}_n$ (which again is consistent with the LLN). This is known as *functional invariance* and can be used for any quantity that can be expressed as a function of the parameter that is estimated.
Finally, there is an equivalent definition of the geometric distribution where the pmf is $f(x|\phi) = \phi^x (1-\phi), x \in \mathbb{N}$. In that case, the parameter is the probability of failure while x denotes the number of failed Bernoulli trials before the first successful one. The expression of the mean then becomes $\mu = \frac{1}{1-\phi} - 1 = \frac{\phi}{1-\phi} \iff \phi = \frac{\mu}{1+\mu}$ and we have $\hat{\phi} = \frac{\bar{X}_n}{1+\bar{X}_n}$.

3.2.3 MLE vs ERM

Though we did not highlight it in the previous paragraphs, MLE is in fact nothing but a special case of ERM. To understand that, we need to formulate maximum likelihood estimation as a learning problem in a modified version of the PAC framework (the PAC framework as defined in Section 3.1 works for supervised learning tasks while parameter estimation is unsupervised). Let $\mathcal{P}$ be a family of distributions where each is parameterized by some $\boldsymbol{\theta} : \mathcal{P} \to \Theta \subseteq \mathbb{R}^k$ with

$P \mapsto \boldsymbol{\theta} = \boldsymbol{\theta}(P)$. Specifically, ignore the domain $\mathcal{Y}$ from the previous definition and let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{H} = \Theta$ instead of a set of functions. Suppose that $\mathcal{D} = P \in \mathcal{P}$ and that we are given samples $\boldsymbol{X_i} \sim \mathcal{D}$. Indeed, suppose that we pick the *negative log-likelihood* $\ell(\boldsymbol{\theta}, \boldsymbol{x}) = -\log(f(\boldsymbol{x}|\boldsymbol{\theta}))$ as a loss function. Remark that:

$$\hat{\boldsymbol{\theta}} = argmax\{L(\boldsymbol{\theta}|\boldsymbol{X_1^n})\} = argmax\left\{\frac{1}{n}\log(L(\boldsymbol{\theta}|\boldsymbol{X_1^n}))\right\} = argmax\left\{\frac{1}{n}\sum_{i=1}^{n}\log(f(\boldsymbol{X_i}|\boldsymbol{\theta}))\right\} =$$

$$= argmin\left\{\frac{1}{n}\sum_{i=1}^{n}(-\log(f(\boldsymbol{X_i}|\boldsymbol{\theta})))\right\} = argmin\left\{\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, \boldsymbol{X_i})\right\}$$

Based on the above, maximizing the likelihood function is equivalent to minimizing the empirical risk when we work in the PAC-like framework defined above.

3.3 Concentration Inequalities

Despite arguing about the importance of MLEs, so far, we have not proven any guarantees about the quality of the approximations they produce. For that reason, we talk about concentration inequalities. These are tools used to show that the probability mass of some random variable tends to concentrate around some value (usually its mean). Such inequalities have always been of interest to the statistics community. Recently, the have also attracted the interest of the computer science community, due to the fact that they offer techniques to prove guarantees about the performance of randomized algorithms. Some very good texts on the subject are [7, 21, 45]. We will start from simpler inequalities (which usually yield results that are not tight) and then move on to more stronger ones.

3.3.1 Markov's Inequality

The simplest concentration bound is *Markov's inequality*. It involves the tails of non-negative random variables. Specifically, given a random variable X with cdf $F_X(t) = \mathbb{P}[X \leq x]$, its *(right) tail distribution* is the one with cdf $1 - F_X(t) = \mathbb{P}[X > t]$. It is known that $\lim_{t\to\infty} F(t) = 1 \iff \lim_{t\to\infty} \mathbb{P}[X > t] = 0$. However, we have no knowledge of the rate of convergence of $\mathbb{P}[X > t]$. What Markov's inequality does is that it offers a simple (but not necessarily tight) view of that rate, given that $X \geq 0$ and $\mathbb{E}[X] < \infty$.

Proposition 3.3.1 (Markov)**.** *Let X be a non-negative random variable with finite mean. We have:*

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}, \forall t > 0$$

Proof. We define the random variable $Y = X\mathbb{1}\{X \geq t\}$. We have:

$$X \geq Y \geq t\mathbb{1}\{X \geq t\} \implies \mathbb{E}[X] \geq t\mathbb{E}[\mathbb{1}\{X \geq t\}] = t\mathbb{P}[X \geq t] \iff \mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$$

■

The above implies that, for the random variables described above, we have $\mathbb{P}[X \geq t] = o\left(\frac{1}{t}\right)$. This, is not generally tight, since it was proven under rather weak assumptions. Note that some tend to replace t in the previous inequalities with $t\mathbb{E}[X]$. This results in a description of the tail with respect to the mean value and the elimination of $\mathbb{E}[X]$ from the RHS.
We now present an extension of Markov's inequality which will be useful for the next parts.

Corollary 3.3.1. *Let X be a non-negative random variable with finite mean. For any increasing function $\Phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, we have:*

$$\mathbb{P}[X \geq t] \leq \mathbb{P}[\Phi(X) \geq \Phi(t)] \leq \frac{\mathbb{E}[\Phi(X)]}{\Phi(t)}, \forall t > 0$$

where the first inequality becomes an equality for Φ : strictly increasing.

The previous results can be extended for random variables that take negative values too, simply by replacing X with $|X|$.
In the next parts of our treatment of concentration inequalities, this will be used as a building block for a number of other bounds that involve concentration around the mean.

3.3.2 Chebyshev's Inequality

Chebyshev's inequality is the most elementary inequality involving concentration around the mean. Specifically, we want to upper bound the probability $\mathbb{P}[|X - \mathbb{E}[X]| \geq t]$. A direct application of Markov's inequality results in $\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|]}{t}$, which involves the expected absolute deviation of the random variable from its mean. However, this quantity is not generally used. On the other hand, by applying Proposition 3.3.1 for $\Phi(x) = x^2$ we get:

Proposition 3.3.2 (Chebyshev)**.** *Let X be a random variable with finite mean and variance. We have:*

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathrm{Var}(X)}{t^2}, \forall t > 0$$

The above implies that, $\mathbb{P}[|X - \mathbb{E}[X]| \geq t] = o\left(\frac{1}{t^2}\right)$, for the random variables described above. This result is not tight, since information involving only the first 2 moments of a random variable were taken into account [2]. Instead, it only demonstrates that random variables with finite mean and variance may tend to concentrate around their mean value. Note that some tend to replace t in the previous inequalities with $t\sqrt{\mathrm{Var}(X)}$. This results in the elimination of $\mathrm{Var}(X)$ from the RHS.

3.3.3 Chernoff Bounds

From this point on, we will present concentration inequalities that are tight, thus allowing us to achieve better approximations with fewer samples than Markov's and Chebyshev's inequalities.

[2] The *p-th moment* of a random variable X is the value $\mathbb{E}[X^p], p \in \mathbb{N}$.

That is not say that those 2 inequalities are not useful. It's just that their value is greater from a theoretical standpoint, as a means to create other inequalities that yield better results.
First, we will present the method of *Chernoff bounds*, which relies heavily on Markov's inequality. Specifically, there were 2 problems that we encountered when we first introduced the inequality. The first had to do with the lack of tight results, which we remarked back then. The second had to do with the fact that the inequality applies only for non-negative random variables ($X \to |X|$ is a solution to this, but it suffers from the inequality's inherent weakness). Now we will present a technique which relies on Markov's inequality but works regardless of the involved random variable's values and produces tighter results. Specifically, suppose we have a random variable X and some $t \in \mathbb{R}$ and we want to upper bound $\mathbb{P}[X \geq t]$. By Corollary 3.3.1 for $\Phi(x) = e^{\lambda x}, \lambda > 0$, we have:

$$\mathbb{P}[X \geq t] = \mathbb{P}\left[e^{\lambda X} \geq e^{\lambda t}\right] \leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda t}}, \forall \lambda > 0$$

Since this holds for any $\lambda > 0$, we can choose the one which minimizes the RHS, so we have:

$$\mathbb{P}[X \geq t] \leq \inf_{\lambda > 0} \left\{ \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda t}} \right\}$$

The previous method is quite general and is named after Herman Chernoff. A similar bound can be obtained for $\mathbb{P}[X \leq t]$, simply by using $\Phi(x) = e^{-\lambda x}, \lambda > 0$. To proceed, further knowledge about X is required. In particular, if X is the sum of independent (but not necessarily identically distributed) random variables $X_1, X_2, \ldots, X_n$, we have:

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i \geq t\right] \leq \inf_{\lambda > 0} \left\{ \frac{\mathbb{E}\left[e^{\lambda \sum_{i=1}^{n} X_i}\right]}{e^{\lambda t}} \right\} = \inf_{\lambda > 0} \left\{ \frac{\prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right]}{e^{\lambda t}} \right\}$$

Remark that, for the final transition to be correct, it is necessary that X_i are independent (simply being pairwise uncorrelated would not suffice). Moreover, the quantity $\mathbb{E}\left[e^{\lambda X_i}\right]$ is the moment generating function of X_i. This term is used to describe it because $\mathbb{E}\left[e^{\lambda X_i}\right] = \sum_{k=0}^{\infty} \frac{\lambda^k \mathbb{E}\left[X_i^k\right]}{k!}$, meaning that all moments are involved. Finally, after computing the value of λ we get that $\mathbb{P}[X \geq t] = o\left(\frac{1}{e^{\lambda t}}\right)$, which is as good as one could hope for.
We will now use the above for sums of independent Bernoulli random variables.

Proposition 3.3.3. *Given independent $X_i \sim Be(p_i), i \in [n]$ such that $\sum_{i=1}^{n} p_i = \mu$, for any $\delta \geq 0$ we have:*

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i \geq (1+\delta)\mu\right] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$$

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i \leq (1-\delta)\mu\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}$$

Proof. We will prove only the first of the 2 bounds. Setting $t = (1+\delta)\mu$ to the generic bound we proved above, we get:

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i \geq (1+\delta)\,\mu\right] \leq \inf_{\lambda>0}\left\{e^{-\lambda(1+\delta)\mu}\prod_{i=1}^{n}\mathbb{E}\left[e^{\lambda X_i}\right]\right\}$$

The moment generating function of a Bernoulli random variable is $\mathbb{E}\left[e^{\lambda X_i}\right] = p_i e^{\lambda} + (1-p_i) = 1 + p_i\left(e^{\lambda}-1\right)$. The minimization of the product $e^{-\lambda(1+\delta)\mu}\prod_{i=1}^{n}\left[1+p_i\left(e^{\lambda}-1\right)\right]]$ is by no means an easy task. So, instead of minimizing it, we will take advantage of the fact that $1+p_i\left(e^{\lambda}-1\right) \leq e^{p_i\left(e^{\lambda}-1\right)}$ and instead minimize:

$$e^{-\lambda(1+\delta)\mu+\sum_{i=1}^{n} p_i\left(e^{\lambda}-1\right)} = e^{-\lambda(1+\delta)\mu+\mu\left(e^{\lambda}-1\right)}$$

which is equivalent to minimizing the exponent. This yields $\lambda_{min} = \ln(1+\delta)$ which leads to the desired result. ■

Due to the fact that the above 2 expressions are not practical, the following approximations are commonly used (see [39]):

Corollary 3.3.2. *Given independent $X_i \sim Be(p_i), i \in [1,\ldots,n]$ such that $\sum_{i=1}^{n} p_i = \mu$, for any $\delta \in [0,1]$ we have:*

$$\mathbb{P}[X \geq (1+\delta)\,\mu] \leq e^{-\frac{\delta^2\mu}{3}}$$

$$\mathbb{P}[X \leq (1-\delta)\,\mu] \leq e^{-\frac{\delta^2\mu}{2}}$$

3.3.4 Hoeffding's Inequality

Hoeffding's Inequality is the last concentration bound we will see for now (one more will be added after we talk about exponential families). It is a consequence of the application of Chernoff bounds to sums of independent random variables with bounded support. Specifically, it can be shown that bounded random variables satisfy the following lemma.

Lemma 3.3.1 (Hoeffding (1963))**.** *Let X be a random variable such that $\mathbb{E}[X] = 0$ and $X \in [a,b]$. Then, for any $s \in \mathbb{R}$, its moment generating function is upper bounded by $e^{\frac{s^2(b-a)^2}{8}}$.*

We will not present the proof of the lemma. To fully comprehend it, an introduction to the concept of sub-Gaussian random variables would be necessary. For the sake of brevity, we avoid that. For a better understanding of the issue, the reader should turn to one of the books referenced at the start of the section. Combining the above with Chernoff's method, we get:

Proposition 3.3.4 (Hoeffding (1963)). *Let $X_i, i \in [n]$ be n independent random variables such that $X_i \in [a_i, b_i], \forall i$ and $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$. For any $t > 0$:*

$$\mathbb{P}\left[\bar{X}_n - \mathbb{E}\left[\bar{X}_n\right] \geq t\right] \leq e^{-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

$$\mathbb{P}\left[\bar{X}_n - \mathbb{E}\left[\bar{X}_n\right] \leq -t\right] \leq e^{-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

$$\mathbb{P}\left[\left|\bar{X}_n - \mathbb{E}\left[\bar{X}_n\right]\right| \geq t\right] \leq 2e^{-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

Both the previous 2 results where first given in [26]. Applying Hoeffding's bound to Example 3.2.2.1, we get $\mathbb{P}\left[|\hat{p} - p| \geq \epsilon\right] \leq 2e^{-2n\epsilon^2} \leq \delta \implies n \geq \frac{1}{2\epsilon^2} \ln\left(\frac{2}{\delta}\right)$. That way, we have a tight lower bound on the number of samples required to have an approximation that is within acceptable limits (which we set by choosing $\epsilon, \delta \in (0, 1)$). Another application (this time in the context of machine learning theory) can be found in chapter 4 of [6]. There, Hoeffding's bound is exploited to show that all finite hypothesis classes are Agnostic PAC-learnable.

3.4 Information Theory and Statistics

Information Theory is a mathematical subject that was first developed during the mid 20th century by Shannon in [46]. Its aim is to present a framework for the theoretical analysis of communications. Due to the presence of noise in communication, tools from probability theory and statistics were employed since its early days. However, after some point, information theory begun influencing those fields, due to the fact that some of the tools developed for it proved quite useful in statistics. It is those tools that we wish to emphasize. A good reference on the topic is [20]. Should someone be interested in a more general introduction with greater emphasis on communications, the authoritative text is [11].

3.4.1 Fundamental Information Theoretic Measures

At the foundation of information theory lies the concept of *entropy*. The idea behind it was to define a measure whose value increases when a random variable exhibits more "random" behavior, which is to say that it tends to concentrate less on specific values, thus making it harder to predict what value it will take. For that reason, the proposed definition was:

Definition 3.4.1 (Shannon (1948)). Let X be a random variable taking values in some discrete set $\mathcal{X}$ with pmf $p : \mathcal{X} \to [0, 1]$. Its entropy is defined as:

$$H(X) = \mathbb{E}\left[\log\left(\frac{1}{p(X)}\right)\right] = -\sum_{x \in \mathcal{X}} p(x) \log\left(p(x)\right)$$

The logarithm in the above definition is usually with base 2, so entropy is measured in bits. It is obvious from the above definition it cannot take negative values since $p(x) \in [0, 1]$.

In particular, the only case when the entropy is equal to 0 is when all probability mass is concentrated on a single value, resulting in a constant distribution [3]. Moreover, it is possible to show that, if a random variable has finite support (say $|X| = m \in \mathbb{N}$), its entropy is upper bounded by $\log(m)$, which is attained only by the uniform distribution. The proof relies on the fact that the function $\log(\cdot)$ is strictly concave and Jensen's inequality. At this point, we should remark that $\log(m)$ is equal to the number of bits required to index a set with m elements. If any distribution other than the uniform is used, due to the bias towards specific values, it is possible to encode it using fewer bits, which is the idea that lies at the core of data compression. The minimum number of bits required for such an encoding is given approximately by the value of the entropy (see chapter 5 of [11]). This motivates the previous definition.
Also, note that the entropy is independent of the support $\mathcal{X}$. Indeed, it is the values of the pmf p that determine it. Consequently, if we consider the random variable $Y = f(X)$ with $f : 1 - 1$, its entropy would be equal to that of X. On the other hand, if f is not $1 - 1$, the entropy of Y would be less. Intuitively, this holds because the support of Y is smaller than that of X, resulting in stronger bias towards specific values.
We now proceed to define the *relative entropy* of a pair of random variables, more commonly referred to as Kullback-Leibler divergence (*KL-divergence*). The definition was given in [32].

Definition 3.4.2 (Kullback, Leibler (1951))**.** Let $X \sim P, Y \sim Q$ be random variables with discrete supports $\mathcal{X}, \mathcal{Y}$ and pmfs $p : \mathcal{X} \to [0,1], q : \mathcal{Y} \to [0,1]$, respectively. The KL-divergence between P and Q is:

$$D_{KL}(P||Q) = \mathbb{E}\left[\log\left(\frac{p(X)}{q(X)}\right)\right] = \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right) = -\sum_{x\in\mathcal{X}} p(x)\log\left(\frac{q(x)}{p(x)}\right)$$

The KL-divergence is commonly used as distance metric between distributions. However, it does not satisfy all the properties commonly associated with distance metrics. Specifically, the KL-divergence is non-negative and becomes 0 only when the 2 distributions involved are the same. This can be proven using the fact that $-\log(\cdot)$ is strictly convex and Jensen's inequality. However, it is neither symmetric, nor does it satisfy the triangle inequality.
Another interesting aspect of the KL-divergence is its behavior when distributions not having the same support are involved. Suppose, for example, that we have 2 distributions P and Q, such that $supp(P) \subset supp(Q)$. In that case, when computing $D_{KL}(Q||P)$, terms involving $x \in supp(Q) \setminus supp(P)$ will result in $D_{KL}(Q||P) = +\infty$ while similar terms do not cause the same in $D_{KL}(P||Q)$. This is another peculiarity of the KL-divergence.
Note that, in case $P \equiv P_1 \times \cdots \times P_n$ and $Q \equiv Q_1 \times \cdots \times Q_n$ (joint distributions of an equal number of independent random variables), we have the following *tensorization identity*:

$$\boxed{D_{KL}(P||Q) = \sum_{i=1}^{n} D_{KL}(P_i||Q_i)}$$

Now, the next step is to define *conditional entropy*:

[3] When computing the entropy of such distributions, terms of the form $0\log(0)$ are quite common. To compute such terms, it is necessary to use the limit $\lim_{p\to 0} p\log(p) = 0$.

Definition 3.4.3 (Shannon (1948))**.** Let X, Y be random variables taking values in $\mathcal{X}, \mathcal{Y}$, respectively. The conditional entropy of X given that $Y = y$ is:

$$H(X|Y=y) = -\sum_{x \in X} p_{X|Y}(x|y) \log\left(p_{X|Y}(x|y)\right)$$

Furthermore, we have:

$$H(X|Y) = \sum_{y \in \mathcal{Y}} p_Y(y) H(X|Y=y)$$

Another important definition is that of the *joint entropy* of a random vector. It is simply the entropy that corresponds to the joint pmf of its components. Formally, we have:

Definition 3.4.4 (Shannon (1948))**.** Let $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_d = D$ be a random vector whose components have joint pmf $p : D \to [0,1]$. Its entropy is defined as $H(\boldsymbol{X}) = \mathbb{E}\left[\log\left(\frac{1}{p(\boldsymbol{X})}\right)\right] = -\sum_{\boldsymbol{x} \in D} p(\boldsymbol{x}) \log(p(\boldsymbol{x}))$.

Both conditional entropy and joint entropy are non-negative, just like entropy. Moreover, the 2 previously defined measures are connected by the following *chain rule*:

$$\boxed{H(\boldsymbol{X}) = \sum_{i=1}^{n} H\left(X_i | \boldsymbol{X_1^{i-1}}\right)}$$

where $\boldsymbol{X_i^j}$ denotes $(X_i, \ldots, X_j)$. We will prove the result for $n = 2$ and the rest can be done by induction. We have:

$$H(X_1, X_2) = -\sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} p(x_1, x_2) \log(p(x_1, x_2)) =$$

$$= -\sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} p(x_1) p(x_2|x_1) \left(\log(p(x_1)) + \log(p(x_2|x_1))\right) =$$

$$= -\sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} p(x_1) p(x_2|x_1) \log(p(x_1)) - \sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} p(x_1) p(x_2|x_1) \log(p(x_2|x_1)) =$$

$$= -\sum_{x_1 \in \mathcal{X}_1} p(x_1) \log(p(x_1)) \left(\sum_{x_2 \in \mathcal{X}_2} p(x_2|x_1)\right) + H(X_2|X_1) = H(X_1) + H(X_2|X_1)$$

The above result should be compared with the fundamental property of probability measures which states that, given any events A, B and a probability measure $\mathbb{P}[\cdot]$, we have $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B \setminus A]$. Taking this analogy into account, entropy has a role similar to probability. However, instead of expressing the likelihood of an outcome occurring, entropy measures how

random is the behavior exhibited by a random variable. Extending this to more random variables, we get joint entropy, which is similar to the probability of a union of events. Finally, by computing the conditional entropy, we get a view of how the perceived randomness of a random variable changes when we know the value of another random variable. This should be compared with the probability of $\mathbb{P}[B \setminus A]$. Note that $\mathbb{P}[B \setminus A] \leq \mathbb{P}[B]$, so for the analogy to be correct, we would expect that $H(X_2|X_1) \leq H(X_2)$. This is true, but the necessary tools to prove this have not been introduced yet.
Having said the above, the analogy between entropy and probability measures is not complete, due to the fact that we have not defined a measure that is similar to $\mathbb{P}[A \cap B]$. For that reason, we define *mutual information*.

Definition 3.4.5 (Shannon (1948))**.** Let X, Y be random variables taking values in $\mathcal{X}, \mathcal{Y}$, respectively with joint pmf p. Their mutual information is defined as:

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \left(\frac{p(x,y)}{p_X(x)\, p_Y(y)} \right)$$

Remark that mutual information can be expressed as the KL-divergence of 2 distributions: the first is the joint distribution of X, Y (denoted $P_{X,Y}$), while the second is their product distribution (denoted $P_X \times P_Y$), which is the joint distribution of 2 independent random variables which follow P_X and P_Y, respectively. Formally, we have $I(X;Y) = D_{KL}(P_{X,Y} || P_X \times P_Y)$ which implies that mutual information is non-negative, since the KL-divergence is non-negative. Moreover, $I(X;Y) = 0 \iff P_{X,Y} \equiv P_X \times P_Y$, meaning that X, Y are independent.
From the definition, it is easy to see that $I(X;X) = H(X)$ (which why entropy is sometimes referred to as self information) and $I(X;Y) = H(X) - H(X|Y) \geq 0 \implies H(X) \geq H(X|Y)$. This has a natural interpretation, which is that giving information about the value of some random variable cannot make it harder for us to guess the value of another one. Substituting according to this to $H(X,Y) = H(X) + H(Y|X)$ we get $H(X,Y) = H(X) + H(Y) - I(X;Y)$, which should be compared with $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$. Also, mutual information satisfies the following chain rule:

$$\boxed{I(X_1, \ldots, X_n; Y) = \sum_{i=1}^{n} I\left(X_i; Y | \boldsymbol{X_1^{i-1}}\right)}$$

For $n = 2$ we have:

$$I(X_1, X_2; Y) = H(X_1, X_2) - H(X_1, X_2|Y) =$$

$$= (H(X_1) + H(X_2|X_1)) - (H(X_1|Y) + H(X_2|X_1, Y)) = I(X_1;Y) + I(X_2;Y|X_1)$$

For $n > 2$ the result can be obtained by induction.
The way the measures we defined are related to each other is summarized in the following image, which is taken from [11]. This makes the analogy with elementary probability obvious. We are done with the definitions of the basic information theoretic quantities. Before proceeding, we have to make one final remark. The previous definitions can be extended to continuous random variables, simply by replacing the sums in the previous definitions with integrals. This

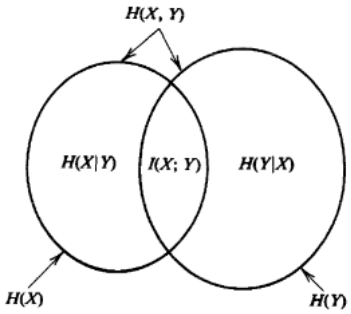

Figure 3.1: Venn diagram of fundamental measures of information theory.

results in the definition of *differential entropy* and is denoted by $h(\cdot)$. Unlike the previously defined measure, it may also take negative values, which is why it lacks a similar interpretation. However, all the other properties satisfied by it and the other measures we defined do not change. For the rest of the text, the definitions and proofs that will be presented will for the most part use differential entropy notation.

3.4.2 f-divergences

We will now make an introduction to the concept of *statistical distances* and, in particular, *f-divergences*. f-divergences offer a unified way of describing measures of dissimilarity between probability distributions. Some of those divergences satisfy all the properties of distance metrics (see Definition 4.2.1).
Below, we will give the definition of f-divergences in the case of continuous distributions, which is the one that will draw our attention for the rest of the text. In order to give a more general definition, it would be necessary to introduce some measure-theoretic concepts and notation, which is outside the scope of this thesis. Should the reader be interested in such an approach, they should turn to chapter 2 of [20] (or, as a matter of fact, [2, 12], where they were independently introduced).

Definition 3.4.6 (Ali, Silvey (1966) and Csiszár (1967))**.** Let P, Q be 2 continuous distributions with pdfs p, q, respectively. Also, let $f : \mathbb{R}_+ \to \mathbb{R}$ be a continuous convex function satisfying $f(1) = 0$. The f-divergence of P, Q is defined as:

$$D_f(P||Q) = \int_{\mathbb{R}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

This definition covers a wide category of statistical distances, all of which satisfy the following property:

Proposition 3.4.1. *Let P_1, P_2, Q_1, Q_2 be distributions with support $\mathcal{X}$ and $f : \mathbb{R}_+ \to R$ be convex with $f(1) = 0$. Then, D_f is jointly convex in its arguments, meaning that for any $\lambda \in [0, 1]$:*

$$D_f\left(\lambda P_1 + (1-\lambda) P_2 || \lambda Q_1 + (1-\lambda) Q_2\right) \leq \lambda D_f\left(P_1||Q_1\right) + (1-\lambda) D_f\left(P_2||Q_2\right)$$

Below, we will present the 3 instances of f-divergences that are of greatest interest to us:

- **KL-divergence:** Choosing $f(x) = x\log(x), x > 0$ results in:

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

  The properties of this divergence measure have been examined in detail in the previous section, so we will not repeat them here.

- **Hellinger distance:** Setting $f(x) = \frac{1}{2}\left(\sqrt{x}-1\right)^2, x > 0$ yields:

$$\boxed{d_{hel}(P,Q)^2 = \frac{1}{2}\int_{\mathbb{R}} q(x)\left(\sqrt{\frac{p(x)}{q(x)}} - 1\right)^2 dx = \frac{1}{2}\int_{\mathbb{R}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx}$$

  which is commonly referred to as the squared Hellinger distance. The Hellinger distance is the ℓ_2 distance of $\sqrt{p}$ and $\sqrt{q}$ multiplied by a normalizing constant equal to $\frac{1}{\sqrt{2}}$. The previous formula can be written in the form:

$$\boxed{d_{hel}(P,Q)^2 = 1 - \int_{\mathbb{R}} \sqrt{p(x)}\sqrt{q(x)} dx}$$

  by using the identity about the square of a difference and the fact that p and q are density functions. Taking advantage of that, we get a very elegant tensorization identity in case $P \equiv P_1 \times \cdots \times P_n, Q \equiv Q_1 \times \cdots \times Q_n$:

$$\boxed{d_{hel}(P,Q)^2 = 1 - \prod_{i=1}^{n} \int_{\mathbb{R}} \sqrt{p_i(x)}\sqrt{q_i(x)} dx = 1 - \prod_{i=1}^{n}\left(1 - d_{hel}(P_i,Q_i)^2\right)}$$

- **Total Variation distance (TV-distance):** Setting $f(x) = \frac{1}{2}|x-1|, x > 0$, we get:

$$\boxed{d_{TV}(P,Q) = \frac{1}{2}\int_{\mathbb{R}} |p(x) - q(x)| dx = \frac{1}{2}||p-q||_1}$$

  Note that the above is not the way the TV-distance is usually defined. Instead, given 2 probability measures P, Q with the same support $\mathcal{X}$, we have:

$$d_{TV}(P,Q) = \sup_{A \subseteq \mathcal{X}} |P(A) - Q(A)|$$

  The 2 definitions are equivalent. Indeed, remark that the absolute value in the above definition is redundant. Indeed, for any $A \subseteq \mathcal{X}$, we have:

$$|P(A) - Q(A)| = |1 - P(A^c) - 1 + Q(A^c)| = |P(A^c) - Q(A^c)|$$

If for some $A \subseteq \mathcal{X}$ we have $Q(A) \geq P(A) \iff P(A^c) \geq Q(A^c)$. Since for every $A \subseteq \mathcal{X}$ we have $A^c \subseteq \mathcal{X}$, dropping the absolute value does not change the value of the TV-distance. Instead, for any set $A \subseteq \mathcal{X}$, it forces us to choose between A and A^c the one for which the difference $P - Q$ is non-negative. As a result, we have:

$$d_{TV}(P, Q) = \sup_{A \subseteq \mathcal{X}} \{P(A) - Q(A)\}$$

Let A be a maximizer of the previous quantity. This implies that $p(x) \geq q(x), \forall x \in A$. Indeed, if for some $x \in A$, we have $p(x) < q(x)$, there would be an interval I where $p(x) < q(x)$ (due to the continuity of p, q). However, this implies that $P(A \setminus I) - Q(A \setminus I) > P(A) - Q(A)$, which contradicts the assumption that A is a maximizer. Similarly, we get that for any interval I where $p(x) \geq q(x)$, $I \subseteq A$. Consequently, we have that $A = \{x \in \mathbb{R} : p(x) \geq q(x)\}$:

$$d_{TV}(P, Q) = \int_A (p(x) - q(x))\, dx = \frac{1}{2}\left(\int_A (p(x) - q(x))\, dx + \int_{A^c} (q(x) - p(x))\, dx\right)$$

$$= \frac{1}{2}\left(\int_A |p(x) - q(x)|\, dx + \int_{A^c} |p(x) - q(x)|\, dx\right) = \frac{1}{2}\int_{\mathbb{R}} |p(x) - q(x)|\, dx$$

Based on the previous, it is easy to see a geometric interpretation of the TV-distance: it is nothing more than the area between p and q where $p(x) \geq q(x)$.

The TV-distance is commonly regarded as the main statistical distance. Unlike the previous ones, it does not have a tensorization identity. Instead, the closest thing we have to that is the following proposition, which is given for discrete distributions, but can be easily generalized for continuous as well:

Proposition 3.4.2. *Let $\{P_i\}_{i\in[n]}, \{Q_i\}_{i\in[n]}$ be distributions where, for each i, the distributions P_i, Q_i have the same support $\mathcal{X}_i$. Suppose there is some non-empty set $I \subseteq [n]$, such that $P_i \equiv Q_i, \forall i \in I$ and let I' denote its complement. Consider the product distributions $P = \bigotimes_{i\in[n]} P_i, Q = \bigotimes_{i\in[n]} Q_i$ and $P' = \bigotimes_{i\in I'} P_i, Q' = \bigotimes_{i\in I'} Q_i$. We have:*

$$d_{TV}(P, Q) = d_{TV}(P', Q')$$

Proof. Let $\mathcal{X} = \bigotimes_{i\in[n]} \mathcal{X}_i$ and $\mathcal{X}_A = \bigotimes_{i\in I} \mathcal{X}_i, \mathcal{X}_B = \bigotimes_{i\in I'} \mathcal{X}_i$. Bold letters used below, based on the context, will denote an element of one of the 3 previous sets. By the definition of the TV-distance, we have that:

$$d_{TV}(P, Q) = \frac{1}{2}\sum_{\boldsymbol{x}\in\mathcal{X}} |P(\boldsymbol{x}) - Q(\boldsymbol{x})| = \frac{1}{2}\sum_{\boldsymbol{x}\in\mathcal{X}} \left|\prod_{i\in[n]} P_i(x_i) - \prod_{i\in[n]} Q_i(x_i)\right|$$

Given that, for all $i \in I$, the corresponding P_i, Q_i are essentially the same distribution, the above can be written in the form:

$$d_{TV}(P,Q) = \frac{1}{2}\sum_{\boldsymbol{x}\in\mathcal{X}}\left[\left(\prod_{i\in I}P_i(x_i)\right)\cdot\left|\prod_{i\in I'}P_i(x_i)-\prod_{i\in I'}Q_i(x_i)\right|\right] \quad (1)$$

Now, the TV-distance of P', Q' is equal to:

$$d_{TV}(P',Q') = \frac{1}{2}\sum_{\boldsymbol{x}\in\mathcal{X}_B}\left|\prod_{i\in I'}P_i(x_i)-\prod_{i\in I'}Q_i(x_i)\right| \quad (2)$$

Our aim now is to group the elements of the sum in the first expression in a way that will help us reach our conclusion. To that end, we write (1) in the form:

$$d_{TV}(P,Q) = \frac{1}{2}\sum_{\boldsymbol{x}\in\mathcal{X}_B}\left[\left|\prod_{i\in I'}P_i(x_i)-\prod_{i\in I'}Q_i(x_i)\right|\cdot\left(\sum_{\boldsymbol{y}\in\mathcal{X}_A}\prod_{i\in I}P_i(y_i)\right)\right] =$$

$$= \frac{1}{2}\sum_{\boldsymbol{x}\in\mathcal{X}_B}\left|\prod_{i\in I'}P_i(x_i)-\prod_{i\in I'}Q_i(x_i)\right| \overset{(2)}{=} d_{TV}(P',Q') \implies$$

$$\implies \boxed{d_{TV}(P,Q) = d_{TV}(P',Q')}$$

where the transition from the first line to the second is based on the fact that in the parenthesis we add over all elements of $\mathcal{X}_A$. ■

There is a simple analogy that helps understand the previous proposition. Suppose we are given 2 points on the real line x_1 and x_2 with $x_1 < x_2$. Their distance is equal to $x_2 - x_1$. Now, suppose that we consider vectors of the form $(x_1, x, \ldots, x), (x_2, x, \ldots, x)$. No matter how many components we add, the distance of the 2 vectors will still be $x_2 - x_1$. The previous proposition states the exact same thing for the TV-distance.

We will now provide another proposition which involves distributions defined on discrete number sets. We slightly abuse notation, since we refer to TV-distance of random variables, when we actually mean the TV-distance between of their respective distributions.

Proposition 3.4.3. *Let $\{X_i\}_{i\in[n]}, \{Y_i\}_{i\in[n]}$ be independent discrete random variables with $X_i \sim P_i, Y_i \sim Q_i$, where, for each i, both P_i and Q_i are defined over $\mathcal{X}_i \subset \mathbb{R}$. Also, let $P = \bigotimes_{i\in[n]} P_i, Q = \bigotimes_{i\in[n]} Q_i$ be the corresponding product distributions. We have:*

$$d_{TV}(P,Q) \geq d_{TV}\left(\sum_{i\in[n]}X_i, \sum_{i\in[n]}Y_i\right)$$

Proof. Let $\mathcal{X} = \bigotimes_{i\in[n]} \mathcal{X}_i$ and let $\mathcal{Y}$ be the (common) support of $X = \sum_{i\in[n]} X_i$ and $Y = \sum_{i\in[n]} Y_i$. We have:

$$d_{TV}(P,Q) = \frac{1}{2}\sum_{\boldsymbol{x}\in\mathcal{X}} |P(\boldsymbol{x}) - Q(\boldsymbol{x})| = \frac{1}{2}\sum_{\boldsymbol{x}\in\mathcal{X}} \left|\prod_{i\in[n]} P_i(x_i) - \prod_{i\in[n]} Q_i(x_i)\right| \quad (1)$$

We write (1) by grouping the terms according to the value of $\sum_{i\in[n]} x_i$. Applying the triangle inequality, we get:

$$d_{TV}(P,Q) = \frac{1}{2}\sum_{y\in\mathcal{Y}}\sum_{\boldsymbol{x}\in\mathcal{X}} \left|\prod_{i\in[n]} P_i(x_i) - \prod_{i\in[n]} Q_i(x_i)\right| \mathbb{1}\left\{\sum_{i\in[n]} x_i = y\right\} \geq$$

$$\geq \frac{1}{2}\sum_{y\in\mathcal{Y}} \left|\sum_{\boldsymbol{x}\in\mathcal{X}} \left[\prod_{i\in[n]} P_i(x_i) - \prod_{i\in[n]} Q_i(x_i)\right] \mathbb{1}\left\{\sum_{i\in[n]} x_i = y\right\}\right| =$$

$$= \frac{1}{2}\sum_{y\in\mathcal{Y}} |\mathbb{P}[X=y] - \mathbb{P}[Y=y]| = d_{TV}(X,Y) \implies$$

$$\implies \boxed{d_{TV}(P,Q) \geq d_{TV}(X,Y)}$$

■

The way the triangle inequality was used in the previous proof is quite general and can be used to prove similar results for any function of the random vectors $\boldsymbol{X_1^n}, \boldsymbol{Y_1^n}$. Intutively, this holds because the support of the resulting distributions is generally smaller, thus making them look more similar than the initial joint distributions.

By combining the 2 previous propositions, we get:

Corollary 3.4.1. *Let $\{X_i\}_{i\in[n]}, \{Y_i\}_{i\in[n]}$ be independent discrete random variables with $X_i \sim P_i, Y_i \sim Q_i$, where, for each i, both P_i and Q_i are defined over $\mathcal{X}_i \subset \mathbb{R}$. Also, let $P = \bigotimes_{i\in[n]} P_i, Q = \bigotimes_{i\in[n]} Q_i$ be the corresponding product distributions. Suppose there is some non-empty set $I \subseteq [n]$, such that $P_i \equiv Q_i, \forall i \in I$ and let I' denote its complement. We have:*

$$d_{TV}(P,Q) \geq d_{TV}\left(\sum_{i\in I'} X_i, \sum_{i\in I'} Y_i\right)$$

Before moving on, we should point out that the 3 divergence measures we introduced above are connected by the following inequalities:

Proposition 3.4.4. *Let P, Q be 2 distributions with the same support:*

- *Using the Hellinger distance, we have:*

$$\frac{1}{2}d_{hel}\left(P,Q\right)^2 \leq d_{TV}\left(P,Q\right) \leq d_{hel}\left(P,Q\right)\sqrt{1-\frac{d_{hel}\left(P,Q\right)^2}{4}}$$

- *Using the KL-divergence, we have Pinsker's inequality:*

$$d_{TV}\left(P,Q\right)^2 \leq \frac{1}{2}D_{KL}\left(P,Q\right)$$

The above inequalities usually help in bounding the TV-distance, which is generally more difficult to compute compared to the other 2. In some cases where the above 2 inequalities do not yield tight results (especially the first one), Corollary 3.4.1 comes in handy.

3.4.3 Inequalities of Information Theory

So far, we have not made any reference to the fundamental topic of information theoretic inequalities. No treatment of the subject is possible without making references to the issue, due to the fact that such inequalities expose inherent limitations of various information processing procedures and statistical methods. Our examination will start from the *Data Processing Inequality* and then move on to *Le Cam's* and *Fano's* inequalities, which will evoked in later sections where we talk about lower bounds.

3.4.3.1 Data Processing Inequality

The Data Processing Inequality is formal statement of a remark that seems intuitively obvious. Specifically, suppose that we have a source of information which is processed by various procedures. As we process the information, the results become less similar to their original form, provided that the source is not involved in any of the processing stages. To express the previous mathematically, it is necessary to introduce the concept of *Markov Chains*.
Markov chains are one of the most common examples of stochastic processes. We do not intend to make a detailed introduction to the topic (for a proper exposition, see [33]). Instead, we focus on their defining feature which is that each state depends solely on the previous one, meaning that, given 3 consecutive states $X \to Y \to Z$, Z and X are conditionally independent given Y. Having said that, we can now state and prove the Data Processing Inequality:

Proposition 3.4.5. *Let $X \to Y \to Z$ be a Markov chain. We have $I\left(X;Y\right) \geq I\left(X;Z\right)$.*

Proof. Consider the quantity $I\left(X,Y;Z\right)$. Based on the chain rule about mutual information we proved previously, there 2 ways to compute it. First, we have:

$$I\left(X;Y,Z\right) = I\left(X;Y\right) + I\left(X;Z|Y\right) = I\left(X;Y\right) \quad (1)$$

due to the conditional independence of X, Z given Y. Secondly, we have:

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) \quad (2)$$

Combining $(1), (2)$, we get:

$$I(X;Y) = I(X;Z) + I(X;Y|Z) \geq I(X;Z)$$

due to the fact that mutual information is non-negative. The equality holds when $I(X;Y|Z) = 0$, which implies that X, Y are conditionally independent given Z, meaning that $X \to Z \to Y$ is also a Markov chain. ■

As we can see, the more we distance ourselves from the source (the random variable X), the less there is in common between it and the random variables that we get in the next stages of the process, judging by the decrease in the values of the mutual information.

3.4.3.2 Le Cam's Inequality

Le Cam's inequality is related to *hypothesis testing*, which is a fundamental problem in statistics. In particular, it involves being given a finite family of distributions along with samples from one of them and being able to discriminate the underlying distribution from the rest. More formally, let $\{P_1, \ldots, P_k\}$ be the candidate distributions and let $P^* \in \{P_1, \ldots, P_k\}$ be the ground truth. All distributions are defined over the same sample space $\mathcal{X}$. Intuitively, the closer the distributions are to each other, the harder it should be to discriminate between them (provided that the same distribution does not appear twice in the family, which would render discrimination impossible). These ideas are expressed through Le Cam's inequality in the case of binary hypothesis testing and through Fano's inequality in the case of multiple hypothesis testing (see [50] as a reference). Below, we will examine the former.
Let $\{P_1, P_2\}$ be the family of possible distributions and let $V \in \{1, 2\}$ be the index corresponding to the ground truth. Since we are not biased towards any of the 2 hypotheses, we view V as a uniform random variable. Supposing that the number of samples we are given is n and that they are denoted X_1^n, we refer to any function $\Psi : \mathcal{X}^n \to \{1, 2\}$ as a *statistical test* or a *testing function*. The probability of error of a testing function is given by:

$$\mathbb{P}[\Psi(X_1^n) \neq V] = \mathbb{P}[V = 1]\,\mathbb{P}[\Psi(X_1^n) \neq V | V = 1] + \mathbb{P}[V = 2]\,\mathbb{P}[\Psi(X_1^n) \neq V | V = 2] =$$

$$= \frac{1}{2} P_1(\Psi(X_1^n) \neq 1) + \frac{1}{2} P_2(\Psi(X_1^n) \neq 2)$$

This can be lower bounded using Le Cam's inequality, which we will now state and prove:

Proposition 3.4.6 (Le Cam). *Let P_1, P_2 be a pair of distributions defined over the same sample space $\mathcal{X}$ and let $\Psi : \mathcal{X} \to \{1, 2\}$ be some testing function. We have:*

$$\inf_{\Psi} \{P_1(\Psi(X) \neq 1) + P_2(\Psi(X) \neq 2)\} = 1 - ||P_1 - P_2||_{TV}$$

Proof. By the definition of the TV-distance, we have:

$$||P_1 - P_2||_{TV} = \sup_{A \subseteq \mathcal{X}} \{P_1(A) - P_2(A)\} \iff 1 - ||P_1 - P_2||_{TV} = 1 - \sup_{A \subseteq \mathcal{X}} \{P_1(A) - P_2(A)\}$$

Remark that $-\sup_{A \subseteq \mathcal{X}} \{P_1(A) - P_2(A)\} = \inf_{A \subseteq \mathcal{X}} \{P_2(A) - P_1(A)\}$. Suppose now that we have found a set A that minimizes the previous quantity. Let Ψ be a testing function such that $\Psi(x) = 1, \forall x \in A$ and $\Psi(x) = 2, \forall x \in A^c$. Consequently, we have $P_1(A) = P_1(\Psi(X) = 1)$ and $P_2(A) = P_2(\Psi(X) = 1) = P_2(\Psi(X) \neq 2)$. This yields:

$$1 - ||P_1 - P_2||_{TV} = \inf_{A \subseteq \mathcal{X}} \{P_2(\Psi(X) \neq 2) + 1 - P_1(\Psi(X) = 1)\} =$$

$$= \inf_{A \subseteq \mathcal{X}} \{P_1(\Psi(X) \neq 1) + P_2(\Psi(X) \neq 2)\}$$

■

Applying this to our previous problem, we get:

$$\frac{1}{2} P_1(\Psi(X_1^n) \neq 1) + \frac{1}{2} P_2(\Psi(X_1^n) \neq 2) \geq \frac{1}{2}(1 - ||P_1^n - P_2^n||_{TV})$$

By Pinsker's inequality, we get $||P_1^n - P_2^n||_{TV} \leq \sqrt{\frac{1}{2} D_{KL}(P_1^n || P_2^n)} = \sqrt{\frac{n}{2} D_{KL}(P_1 || P_2)}$ where the last transition is due to the tensorization identity of the KL-divergence. This results in:

$$\boxed{\mathbb{P}[\Psi(X_1^n) \neq V] = \frac{1}{2} P_1(\Psi(X_1^n) \neq 1) + \frac{1}{2} P_2(\Psi(X_1^n) \neq 2) \geq \frac{1}{2}\left(1 - \sqrt{\frac{n}{2} D_{KL}(P_1 || P_2)}\right)}$$

We have thus lower bounded the error probability of any testing function in binary hypothesis testing problems.

3.4.3.3 Fano's Inequality

Fano's inequality (introduced in [23]) offers a similar lower bound in the setting of multiple hypothesis testing. We now view the whole process as a Markov chain, in a way similar to the data processing inequality. Specifically, let V be a random variable taking values in some finite set $\mathcal{V}$ (which will generally be considered to be $[k]$). Suppose that we are given some random variable X whose value depends on that of V and which we use to guess the value of V. This results in another random variable, denoted $\hat{V}$, which is our guess for the value of V. This results in the Markov chain $V \to X \to \hat{V}$. We define the binary random variable $E = \mathbb{1}\left\{V \neq \hat{V}\right\}$, which becomes 1 when an error is made. Fano's inequality states that:

Proposition 3.4.7 (Fano (1968))**.** *Given the Markov chain $V \to X \to \hat{V}$ where $V, \hat{V} \in \mathcal{V}, |\mathcal{V}| < \infty$ and the random variable E defined as above, we have:*

$$H(E) + \mathbb{P}\left[V \neq \hat{V}\right] \log(|\mathcal{V}| - 1) \geq H\left(V | \hat{V}\right)$$

Proof. We will compute the joint entropy of V (the value we wish to guess) and E (the indicator of error) given $\hat{V}$ (our guess, which is the only of the 3 variables we know). Based on the chain rule introduced earlier, we have:

$$H\left(V,E|\hat{V}\right)=H\left(V|\hat{V}\right)+H\left(E|V,\hat{V}\right)=H\left(V|\hat{V}\right)\ (1)$$

where the last transition is based on the fact that $V,\hat{V}$ fully determine E. Applying the chain rule differently, we get:

$$H\left(V,E|\hat{V}\right)=H\left(E|\hat{V}\right)+H\left(V|E,\hat{V}\right)\leq H\left(E\right)+H\left(V|E,\hat{V}\right)\ (2)$$

The second term in the previous expression can be written in the form:

$$H\left(V|E,\hat{V}\right)=\mathbb{P}\left[V\neq\hat{V}\right]H\left(V|E=1,\hat{V}\right)+\left(1-\mathbb{P}\left[V\neq\hat{V}\right]\right)H\left(V|E=0,\hat{V}\right)\leq$$

$$\leq\mathbb{P}\left[V\neq\hat{V}\right]\log\left(|V|-1\right)\ (3)$$

where the final transition is based on the fact that given $\hat{V}$ and that $E=0$ the value of V is fully determined while given $\hat{V}$ and that $E=1$ the support of V is $\mathcal{V}\setminus\left\{\hat{V}\right\}$ and its entropy is upper bounded by that of the uniform distribution on that set. Combining $(1),(2)$ and (3) we get the desired result. ■

So far, we have not made any assumption regarding the distribution of V. Like in the case of binary hypothesis testing, we will assume that the distribution of V is uniform over the set $\mathcal{V}$. As a result, we have:

Corollary 3.4.2. *Given the Markov chain $V\rightarrow X\rightarrow\hat{V}$ where $V,\hat{V}\in\mathcal{V},|\mathcal{V}|<\infty$ and assuming that V is uniform over $\mathcal{V}$, we have:*

$$\mathbb{P}\left[\hat{V}\neq V\right]\geq1-\frac{\log\left(2\right)+I\left(V;X\right)}{\log\left(|\mathcal{V}|\right)}$$

Proof. In Proposition 3.4.7, the LHS is upper bounded by $\log\left(2\right)+\mathbb{P}\left[V\neq\hat{V}\right]\log\left(|\mathcal{V}|\right)$. The RHS can be written in the form $H\left(V|\hat{V}\right)=H\left(V\right)-I\left(V;\hat{V}\right)=\log\left(m\right)-I\left(V;\hat{V}\right)\geq\log\left(m\right)-I\left(V;X\right)$, where the last transition is due to Proposition 3.4.5. Combining the above leads to the desired version of the inequality. ■

Using the notation introduced previously about hypothesis testing, we get:

$$\boxed{\inf_{\Psi}\mathbb{P}\left[\Psi\left(X\right)\neq V\right]\geq1-\frac{\log\left(2\right)+I\left(V;X\right)}{\log\left(|\mathcal{V}|\right)}}$$

3.5 Distribution Learning and Lower Bounds

Despite being the main issue of this chapter, we have not so far made any references to distribution learning. However, we have now introduced all the necessary background to examine the topic.

3.5.1 The Framework

The framework to study distribution learning problems is analogous to the PAC-learning framework. It was introduced 10 years after PAC-learning in [29]. The main difference between the 2 frameworks is due to the fact that distribution learning problems are examples of unsupervised learning tasks, since all we are given are samples from a distribution and we want to compute a distribution with that is close to the underlying with high probability with respect to some statistical distance. The statistical distance is usually either the KL-divergence or the TV-distance, while there is also Kolmogorov's distance (which is a "weaker" version of the TV-distance where the supremum is taken over all intervals of the form $(-\infty, a]$ instead of all sets $A \subseteq \mathbb{R}$). The KL-divergence is the "strongest" of the 3. That is because, due to Pinsker's inequality (see Proposition 3.4.4), having $D_{KL}(P||Q) \leq \epsilon^2$ implies $d_{TV}(P, Q) \leq \epsilon$. On the other hand, Kolmogorov's distance is the weakest and we will not use it in the rest of the text. We now present the definition of learnability given in [29] (though slightly simplified):

Definition 3.5.1 (Kearns et. al. (1994))**.** A family of distributions $\mathcal{F}$ is said to be efficiently learnable with respect to some divergence measure d when, for any $\epsilon > 0, \delta \in (0, 1)$, given oracle access to samples from an unknown distribution $P \in \mathcal{F}$, there exists a polynomial time algorithm $\mathcal{A}$ that outputs a distribution $\hat{P}$ such that:

$$\mathbb{P}\left[d\left(\hat{P}, P\right) \geq \epsilon\right] < \delta$$

where the probability is computed with respect to the samples.
If $\hat{P} \in \mathcal{F}$, then $\mathcal{A}$ is said to be proper. Otherwise, it is improper.

What we need to do now is to describe a principle similar to ERM, which will help us approach distribution learning problems. We should stress that the distribution learning problems we will examine are proper (unlike [13] for example) and we have no corrupted samples (unlike [18]). It turns out that, in this context, the best approach is simply to output the distribution corresponding to the parameter vector computed using MLE (provided that computing or approximating the MLE is possible, which is true for the problems examined in subsequent chapters of this thesis). The previous description is quite informal, but a proper exposition of those ideas can be found in chapter 24 of [6], where it is shown that minimizing the expected negative log-likelihood is equivalent to minimizing the KL-divergence between the true distribution and the estimate.

3.5.2 Minimax Lower Bounds

Having defined a framework for distribution learning, we need to come up with techniques that verify the optimality of our algorithms. For that reason, we need to think which instances are the ones that are bound to make our algorithms struggle. Specifically, suppose that we have 2 distributions $\mathcal{P}_1, \mathcal{P}_2$ belonging in the same family. If the 2 distributions are well-separated with respect to some of the statistical distances we defined, our algorithms should not have a hard time telling them apart, even with a smaller number of samples. The hard instances are the ones where the distributions are close to each other, resulting in the samples tending to concentrate around the same values. Such instances help us describe the sample complexity of various classes of distributions.

3.5.2.1 Minimax Risk

To express the previous thoughts formally, we need to define the minimax risk. Our presentation will be quite concise, focusing on its definition in the context of distribution learning, using the notation presented in [9], which we will also use in later chapters. Should the reader interested in a more general approach, they should turn to chapter 7 of [20] or [50].
Suppose that we are given a family of distributions $\mathcal{F}$ with support $\mathcal{X}$ and there is some unknown $f \in \mathcal{F}$ which we want to estimate using n samples $\boldsymbol{x} = (x_1, \ldots, x_n) \sim f^n$. Let $\Delta_{\mathcal{X}}$ denote the set of all distributions on the set $\mathcal{X}$ and let $\hat{f} : \mathcal{X}^n \to \Delta_{\mathcal{X}}$ be an estimator which uses n samples and outputs a distribution in $\Delta_{\mathcal{X}}$. Then, the *maximum risk* of $\hat{f}$ is defined as:

$$\mathcal{R}_n\left(\hat{f}, \mathcal{F}\right) = \sup_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{F}^n} \left[d_{TV}\left(\hat{f}\left(\boldsymbol{x}\right), f\right)\right]$$

Now suppose that we have the set of all possible estimators $\Omega = \left\{\hat{f} : \hat{f} : \mathcal{X}^n \to \Delta_{\mathcal{X}}\right\}$. The difficulty of estimating distributions in $\mathcal{F}$ can be determined by the maximum risk of the best possible estimator. For that reason, we define the minimax risk of $\mathcal{F}$:

$$\mathcal{R}_n\left(\mathcal{F}\right) = \inf_{\hat{f} \in \Omega} \mathcal{R}_n\left(\hat{f}, \mathcal{F}\right)$$

There are some remarks that we should make regarding the nature of the previous definition. The first involves the choice of the TV-distance. Specifically, one could ask why we did not use the KL-divergence instead of it. The answer is that most results of minimax theory that we will use require the employed distance measure to be at least a semi-metric (meaning that all conditions of Definition 4.2.1 are satisfied apart (possibly) from the triangle inequality). Conversely, the KL-divergence is not symmetric, which renders it unusable for such tasks. Our second remark is about the definition of $\hat{f}$ and Ω, which are quite general, since they refer to all possible distributions defined on $\mathcal{X}$. The classes of estimators that we will examine will be far more restricted, because the output distributions will be similar to those in $\mathcal{F}$, due to the fact that we are interested in proper learning tasks.
Before moving on, we would like to explain how the version of the minimax risk that we introduced can be derived from the more general version given in chapter 7 of [20]. Specifically, the definition given there is:

$$\mathfrak{M}_n\left(\theta\left(\mathcal{P}\right), \Phi \circ \rho\right) = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\left(\Phi \circ \rho\right)\left(\hat{\theta}\left(X_1^n\right), \theta\left(P\right)\right)\right]$$

where $\mathcal{P}$ is a family of distributions, $\theta : \mathcal{P} \to \Theta$ with $P \mapsto \theta = \theta\left(P\right)$, $\rho : \Theta \times \Theta \to \mathbb{R}_{\geq 0}$ is a (semi)metric and $\Phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is non-decreasing with $\Phi\left(0\right) = 0$. Setting $\mathcal{P} = \mathcal{F}$, $\Theta = \mathcal{F}$ and substituting $\theta\left(f\right) = f$, $\rho = d_{TV}$ and $\Phi\left(x\right) = x$ yields our version of the minimax risk.

3.5.2.2 Le Cam's and Fano's Methods

Generally, the minimax risk can be hard to compute exactly, due to the fact that the structure of the family of distributions involved may be quite complex. For that reason, the common approach is to lower bound it by reducing it to an instance of hypothesis testing. This results in the minimax risk being lower bounded by an expression involving the probability of error in the testing problem. The hypotheses must be chosen in a way that captures the problem's inherent hardness. As we described previously, to do that, it is necessary to chose distributions that are hard to separate. This need is expressed through an upper bound on the KL-divergence (which, as we mentioned, is the "strongest" of the statistical distances) of any pair of hypotheses.

However, on its own, this is not sufficient, because it does not rule out the possibility of having distributions that are way too close to each other (or multiple instances of the same distribution) among the hypotheses, which could lead to a degenerate case. For the previous to be ensured, a lower bound on the TV-distance of any pair of distinct hypotheses is also required.
The previous reasoning is reflected in Le Cam's and Fano's methods. Both are inspired from the inequalities of the same name, which were stated and proven in Section 3.4.3. By adjusting (7.3.3) from [20] to our setting, we get Le Cam's method:

Proposition 3.5.1 (Le Cam, Pinsker)**.** *Given a pair of distributions $P_1, P_2 \in \mathcal{F}$ with support $\mathcal{X}$ satisfying $d_{TV}(P_1, P_2) \geq a$ and $D_{KL}(P_1||P_2), D_{KL}(P_2||P_1) \leq b$, we have:*

$$\mathcal{R}_n(\mathcal{F}) \geq \frac{a}{4}\left(1 - ||\mathcal{F}_1^n - \mathcal{F}_2^n||_{TV}\right) \geq \frac{a}{4}\left(1 - \sqrt{\frac{n}{2}b}\right)$$

The second part of the inequality is a consequence of Pinsker's inequality. It is not included in the version of the statement presented in [20], but it is used in the examples that follow it, so we integrated it to the proposition.
We now proceed with Fano's inequality, as it is stated in [50].

Proposition 3.5.2 (Yu (1997))**.** *Let $\mathcal{F}$ be a finite family of densities such that:*

$$\inf_{f,g \in \mathcal{F}: f \neq g} d_{TV}(f, g) \geq a, \quad \sup_{f,g \in \mathcal{F}: f \neq g} D_{KL}(f||g) \leq b$$

then it holds that:

$$\mathcal{R}_n(\mathcal{F}) \geq \frac{a}{2}\left(1 - \frac{nb + \ln(2)}{\ln(|\mathcal{F}|)}\right)$$

An interesting aspect of both the above propositions is the fact that the number of samples required to be able to discriminate between the distributions in the family is depends only on b and not on a. That should be expected, since the hardness of telling apart the distributions is determined by how close they are, which is expressed through b.

3.5.2.3 The Gilbert-Varshamov Bound

The Gilbert-Varshamov bound is a result in coding theory that was proven independently by the people it is named after (see [25, 48]). Despite that, it also has applications in minimax theory when combined with Fano's inequality, since it can be exploited to lower bound the cardinality of families of distributions. We state and prove a special version of the bound which we will use later.

Lemma 3.5.1 (Gilbert (1952), Varshamov (1957)). *Let $d \geq 1$ and $\mathcal{H} = \{0,1\}^d$. There exists a subset $\mathcal{V}$ of $\mathcal{H}$ with $|\mathcal{V}| \geq 2^{\frac{d}{8}}$ where any pair of distinct elements of $\mathcal{V}$ has Hamming distance at least $\frac{d}{8}$.*

Proof. Let $\mathcal{V}$ be a maximal subset satisfying the required property. This means that, for any $\boldsymbol{u} \in \mathcal{H} \setminus \mathcal{V}$, the set $\mathcal{V} \cup \{\boldsymbol{u}\}$ cannot have the desired property. Equivalently, this means that, for any $\boldsymbol{u} \in \mathcal{H}$, there is some $\boldsymbol{v} \in \mathcal{V}$ such that $\boldsymbol{u} \in B\left(\boldsymbol{v}, \frac{d}{8}\right)$, where $B\left(\boldsymbol{v}, \frac{d}{8}\right)$ denotes the set of all the elements that have Hamming distance from $\boldsymbol{v}$ less than $\frac{d}{8}$. Consequently, we have:

$$\bigcup_{\boldsymbol{v} \in \mathcal{V}} B\left(\boldsymbol{v}, \frac{d}{8}\right) = \mathcal{H} \implies \sum_{\boldsymbol{v} \in \mathcal{V}} \left|B\left(\boldsymbol{v}, \frac{d}{8}\right)\right| = |\mathcal{V}| \left|B\left(\boldsymbol{0}, \frac{d}{8}\right)\right| \geq |\mathcal{H}| = 2^d$$

We now have to upper bound $\left|B\left(\boldsymbol{0}, \frac{d}{8}\right)\right|$. To do that, let $X_i \sim Be\left(\frac{1}{2}\right)$ and consider the vector $\boldsymbol{X} = (X_1, \ldots, X_d)$. We have $\boldsymbol{X} \in B\left(\boldsymbol{0}, \frac{d}{8}\right)$ if $\sum_{i=1}^d X_i \leq \frac{d}{8}$, so we have:

$$\frac{\left|B\left(\boldsymbol{0}, \frac{d}{8}\right)\right|}{2^d} = \mathbb{P}\left[\sum_{i=1}^d X_i \leq \frac{d}{8}\right] = \mathbb{P}\left[\sum_{i=1}^d X_i \geq \frac{7d}{8}\right]$$

where the last transition is due to the symmetry resulting from the fact that X_i are uniform over $\{0,1\}$. By using the Chernoff bound we introduced in Corollary 3.3.2 with $\mu = \frac{d}{2}$ and $\delta = \frac{3}{4}$, we get:

$$\frac{1}{|\mathcal{V}|} = \frac{\left|B\left(\boldsymbol{0}, \frac{d}{8}\right)\right|}{2^d} \leq e^{-\frac{3d}{32}} \iff |\mathcal{V}| \geq e^{\frac{3}{4}\frac{d}{8}} > e^{\ln(2)\frac{d}{8}} = 2^{\frac{d}{8}}$$

■

3.6 Exponential Families

This chapter will close with an introduction to *exponential families*. We will use the notation of exponential families in the following parts of this thesis due to its expressive power. Specifically, every distribution whose density can be written in the form:

$$\boxed{p_{\boldsymbol{\theta}}(\boldsymbol{x}) = h(\boldsymbol{x}) \exp\left(\boldsymbol{\theta}^T \boldsymbol{T}(\boldsymbol{x}) - a(\boldsymbol{\theta})\right), \boldsymbol{x} \in \mathbb{R}^d}$$

with $h : \mathbb{R}^d \to \mathbb{R}_+$ (carrier measure), $\boldsymbol{T} : \mathbb{R}^d \to \mathbb{R}^k$ (sufficient statistics) and $\boldsymbol{\theta} \in \mathbb{R}^k$ (natural parameters) belongs to some exponential family. Moreover, distributions described by the same $\boldsymbol{T}, h$ belong to the same exponential family (denoted $\mathcal{E}(\boldsymbol{T}, h)$). Finally, the function $a : \mathbb{R}^k \to \mathbb{R}$ (logarithmic partition function) is equal to:

$$\boxed{a(\boldsymbol{\theta}) = \ln\left(\int h(\boldsymbol{x}) \exp\left(\boldsymbol{\theta}^T \boldsymbol{T}(\boldsymbol{x})\right) d\boldsymbol{x}\right), a(\boldsymbol{\theta}) < \infty}$$

supposing $\boldsymbol{\theta} \in \mathcal{H}$ (range of natural parameters).
Most, if not all the distributions one encounters in introductory probability courses are exponential families. We give some examples below:

Example 3.6.0.1. The Bernoulli distribution has pmf $f(x|p) = p^x(1-p)^{1-x}, x \in \{0,1\}$. This can be written in the form $(1-p)\left(\frac{p}{1-p}\right)^x = e^{\ln\left(\frac{1-p}{p}\right)x + \ln(1-p)}$. Setting $h(x) = 1, \theta = \ln\left(\frac{1-p}{p}\right), T(x) = x, a(\theta) = \ln\left(1+e^{\theta}\right)$ we have an exponential family.

Example 3.6.0.2. The geometric distribution has pmf $f(x|\phi) = \phi^x(1-\phi), x \in \mathbb{N} \cup \{0\}$. This can be written in the form $e^{\ln(\phi)x + \ln(1-\phi)}$. Setting $h(x) = 1, \theta = \ln(\phi), T(x) = x, a(\theta) = \ln\left(\frac{1}{1-e^{\theta}}\right)$ we have an exponential family.

The following theorem gives a general way of computing various quantities about exponential families, thus justifying their usefulness:

Proposition 3.6.1. *Let $\mathcal{E}(\boldsymbol{T}, h)$ be an exponential family parameterized by $\boldsymbol{\theta} \in \mathbb{R}^k$. Then, the following hold:*

- *For all $\boldsymbol{\theta} \in \mathcal{H}$, it holds that $\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{\theta}}}[\boldsymbol{T}(\boldsymbol{x})] = \nabla a(\boldsymbol{\theta})$.*
- *For all $\boldsymbol{\theta} \in \mathcal{H}$, it holds that $\mathop{\mathrm{Var}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{\theta}}}(\boldsymbol{T}(\boldsymbol{x})) = \nabla^2 a(\boldsymbol{\theta})$.*
- *For all $\boldsymbol{\theta} \in \mathcal{H}$, $\boldsymbol{s} \in \mathbb{R}^k$, it holds that:*

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\exp\left(\boldsymbol{s}^T\boldsymbol{T}(\boldsymbol{x})\right)\right] = \exp\left(a(\boldsymbol{\theta}+\boldsymbol{s}) - a(\boldsymbol{\theta})\right)$$

- *For all $\boldsymbol{\theta}, \boldsymbol{\theta'} \in \mathcal{H}$, and for some $\boldsymbol{\xi} \in L(\boldsymbol{\theta}, \boldsymbol{\theta'})$ [4] it holds that:*

$$D_{KL}(\mathcal{P}_{\boldsymbol{\theta'}} || \mathcal{P}_{\boldsymbol{\theta}}) = -\left(\boldsymbol{\theta'} - \boldsymbol{\theta}\right)^T \nabla a(\boldsymbol{\theta}) + \left(a\left(\boldsymbol{\theta'}\right) - a(\boldsymbol{\theta})\right) = \left(\boldsymbol{\theta'} - \boldsymbol{\theta}\right)^T \nabla^2 a(\boldsymbol{\xi})\left(\boldsymbol{\theta'} - \boldsymbol{\theta}\right)$$

- *For all $\boldsymbol{\theta}, \boldsymbol{\theta'} \in \mathcal{H}$, and for some $\boldsymbol{\xi} \in L(\boldsymbol{\theta}, \boldsymbol{\theta'})$ it holds that:*

$$d_{TV}(\mathcal{P}_{\boldsymbol{\theta}}, \mathcal{P}_{\boldsymbol{\theta'}}) = \frac{1}{2} \cdot \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{\xi}}}\left[sign\left(\mathcal{P}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathcal{P}_{\boldsymbol{\theta'}}(\boldsymbol{x})\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta'}\right)^T\left(\boldsymbol{T}(\boldsymbol{x}) - \mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathcal{P}_{\boldsymbol{\xi}}}[\boldsymbol{T}(\boldsymbol{y})]\right)\right]$$

We will use a simpler version of the expression for the TV-distance. Specifically, in the single parameter case, for $\theta' \to \theta^-$ we get:

$$\boxed{d_{TV}(\mathcal{P}_{\theta}, \mathcal{P}_{\theta'}) = \frac{1}{2}(\theta - \theta') \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\xi}}\left[\left|T(\boldsymbol{x}) - \mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathcal{P}_{\xi}}[T(\boldsymbol{y})]\right|\right]}$$

Put in simple terms, the above says that, if θ, θ' are close and $\theta > \theta'$, the TV-distance of 2 distributions in the same exponential family with natural parameters θ, θ' is equal to half the expected absolute deviation from the mean of some distribution in the family whose natural parameter lies between θ, θ' multiplied by $\theta - \theta'$.

[4] $L(\boldsymbol{x}, \boldsymbol{y})$ denotes the line segment defined by the points corresponding to vectors $\boldsymbol{x}, \boldsymbol{y}$.

The terms involving $\boldsymbol{\xi} \in L(\boldsymbol{\theta}, \boldsymbol{\theta'})$ result from the application of the multidimensional version of the *mean value theorem*.
Just as we found general expressions for the mean and the variance of exponential families, it is possible to find a Chernoff bound for single dimensional/single parameter exponential families.

Proposition 3.6.2 (Busa-Fekete et. al. (2019)). *Let $\mathcal{E}(T, h)$ be an exponential family with natural parameter $\theta \in \mathbb{R}$, logarithmic partition function a and range of parameters $\mathcal{H}$. Then, the following concentration inequality holds for all $\theta, \theta' \in \mathcal{H}$:*

$$\mathop{\mathbb{P}}_{\boldsymbol{x} \sim \mathcal{P}_\theta^n} \left[\left(\frac{1}{n} \sum_{i=1}^{n} T(x_i) \right) (\theta' - \theta) \geq \mathop{\mathbb{E}}_{y \sim \mathcal{P}_{\theta'}} [T(y)] (\theta' - \theta) \right] \leq$$

$$\leq \exp\left(-D_{KL}\left(\mathcal{P}_{\theta'} || \mathcal{P}_\theta\right) n\right)$$

In Proposition 3.6.1, we gave an expression of the KL-divergence which in the single dimensional/single parameter case can be written as $D_{KL}(\mathcal{P}_{\theta'} || \mathcal{P}_\theta) = (\theta' - \theta)^2 \mathop{\mathrm{Var}}_{x \sim \mathcal{P}_\xi} (T(x))$ for some $\xi \in L(\theta, \theta')$. This can be used to compute the KL-divergence in the above case. Finally, using Pinsker's inequality, we can get a similar bound using the TV-distance.
The last result about exponential families that we will need involves the relationship between the TV-distance and the KL-divergence of any pair of distributions belonging in the same exponential family and those of their corresponding sufficient statistics.

Proposition 3.6.3 (Busa-Fekete et. al. (2019)). *Let $\mathcal{E}(\boldsymbol{T}, h)$ be an exponential family with sufficient statistics $\boldsymbol{T}$ and carrier measure h. For any $\mathcal{P}_{\boldsymbol{\theta}} \in \mathcal{E}(\boldsymbol{T}, h)$ let $\mathcal{D}_{\boldsymbol{\theta}}$ be the distribution of the corresponding sufficient statistics, i.e $\mathcal{D}_{\boldsymbol{\theta}}$ is the distribution of $\boldsymbol{T}(\boldsymbol{x})$ when $\boldsymbol{x} \sim \mathcal{P}_{\boldsymbol{\theta}}$. Then, for all $\boldsymbol{\theta}, \boldsymbol{\theta'} \in \mathcal{H}$:*

$$d_{TV}(\mathcal{P}_{\boldsymbol{\theta}}, \mathcal{P}_{\boldsymbol{\theta'}}) = d_{TV}(\mathcal{D}_{\boldsymbol{\theta}}, \mathcal{D}_{\boldsymbol{\theta'}}) \text{ and } D_{KL}(\mathcal{P}_{\boldsymbol{\theta}} || \mathcal{P}_{\boldsymbol{\theta'}}) = D_{KL}(\mathcal{D}_{\boldsymbol{\theta}} || \mathcal{D}_{\boldsymbol{\theta'}})$$

Our exposition of the topic ends here. For a more detailed introduction to exponential families, a good resource is [28]. The proofs for the first 3 expressions of Proposition 3.6.1 can be found there. The last 2 can be found in [9] (the last expression is given there without the normalization constant $\frac{1}{2}$) and so can Propositions 3.6.2 and 3.6.3.

Chapter 4

Permutations and Rankings

In this chapter, we will start with the theoretical aspects of permutations and then move on to their more practical value. Specifically, we will introduce some important properties of permutations as mathematical objects. Moreover, we will explain their use in modelling rankings. Finally, we will present a number of probabilistic models on rankings, most notably the Mallows model, where we will focus our attention.

4.1 Permutations

4.1.1 Permutations as Groups

We start by giving the definition of permutations. Our presentation will include some references to group theory, though our exposition of the subject will be rather restricted. For more details, the reader should turn to abstract algebra textbooks (for example [24]).

Definition 4.1.1. Given $A \neq \emptyset$, all functions $\pi : A \to A$ that are bijective ($1-1$ and onto) are referred to as *permutations* of A.

For the rest of the text, given any non-empty set A, the set of all its permutations will be denoted S_A. If A is equal to some initial segment of the set of natural numbers, e.g. $A = \{1, 2, \ldots, m\} = [m]$, its set of permutations will be denoted S_m. Moreover, if $|A| = m \in \mathbb{N}$, it is known that $|S_A| = m!$.
We now proceed to define the algebraic structure known as a group. Groups are important because they allow us to study the properties of a large number of sets. After stating the definition, we will show that permutation sets can be equipped with a binary operation, resulting in them having the structure of a group.

Definition 4.1.2. Let $G \neq \emptyset$ and $* : G \times G \to G$ be a binary operation. We say that $(G, *)$ is *group* if the following hold:

- $*$ is associative $\iff a * (b * c) = (a * b) * c, \forall a, b, c \in G$.

- G has an identity element with respect to $* \iff \exists e \in G : a * e = e * a = a, \forall a \in G$.
- every element of G has an inverse $\iff \forall a \in G : \exists a^{-1} \in G : a * a^{-1} = a^{-1} * a = e$.

The uniqueness of both the identity element as well as each element's inverse are direct consequences of the above definition. Another important consequence of it is that, for any a, b belonging to a group with operation $*$, the equation $a * x = b$ has a unique solution, which is $x = a^{-1} * b$ (the same holds for $x * a = b \iff x = b * a^{-1}$). Furthermore, given some group $(G, *)$ and a non-empty subset G' of G which closed with respect to both the group operation and the inverse operation, we say that $(G', *)$ is a *subgroup* of $(G, *)$ (denoted $(G', *) \leq (G, *)$). We now need to equip S_A with some binary operation $*$ which will result in a structure $(S_A, *)$ that satisfies the above definition. That operation will be function composition, which will henceforth be referred to as permutation multiplication. The usual notation for function composition is $\circ$, though we will, for the most part, avoid it, instead opting to simply write compositions as products. All the previous reasoning is summarized in the following theorem.

Proposition 4.1.1. *Let $A \neq \emptyset$ and S_A be the set of all its permutations. The structure $(S_A, \circ)$ is a group.*

Proof. We first have to prove that permutation multiplication is an internal operation. To that end, we consider 2 permutations $\pi, \sigma \in S_A$. The function $\pi\sigma : A \to A$ has to be bijective. Suppose that we have $x_1, x_2 \in A$ such that $\pi\sigma(x_1) = \pi\sigma(x_2)$. Since both π and σ are $1-1$, we have:

$$\pi(\sigma(x_1)) = \pi(\sigma(x_2)) \overset{\pi:1-1}{\iff} \sigma(x_1) = \sigma(x_2) \overset{\sigma:1-1}{\iff} x_1 = x_2 \iff \pi\sigma : 1-1$$

Moreover, given any $y \in A$, it is possible to find $x \in A$ such that $\pi\sigma(x) = y$, simply by setting $x = \sigma^{-1}\pi^{-1}(y)$. Consequently, $\pi\sigma$ is indeed bijective, so $\pi\sigma \in S_A$.
The remaining conditions are easier to prove. Specifically, associativity follows directly from the fact that function composition is associative, while the identity element is none other than the identity function $id(x) = x, \forall x \in A$. Finally, the fact that permutations are bijective guarantees the existence of inverse functions. ■

Henceforth, we will now use the term *symmetric group* to refer to S_A. However, so far, we have not given a satisfactory explanation of the importance of permutations. To do that, it is necessary to define the concept of homomorphisms. This notion is fundamental in abstract algebra and will help us highlight the ubiquity of permutations.

Definition 4.1.3. Let $(G, *)$ and $(G', *')$ be 2 groups and let $f : G \to G'$ be a mapping. If f is such that $f(x_1 * x_2) = f(x_1) *' f(x_2)$, we refer to f as a *homomorphism*. Based on whether a number of other properties are satisfied, we may also use the following terms:

- if f is $1-1$, it is a *monomorphism*.

- if f is surjective, it is an *epimorphism*.
- if both the above properties are satisfied, it is an *isomorphism* (denoted $(G,*) \cong (G',*')$).

From the above definition, it is obvious that $f(G) \subseteq G'$. We will prove that $(f(G),*')$ is a subgroup of $(G',*')$. The proof will be facilitated by the next lemma.

Lemma 4.1.1. *Let $(G,*)$ and $(G',*')$ be groups and $f: G \to G'$ be a homomorphism. Then, the following hold:*

- $f(e) = e'$.
- $f\left(a^{-1}\right) = f(a)^{-1}, \forall a \in G$.

Proof.
- We have $f(e) = f(e*e) = f(e) *' f(e) \iff f(e) = e'$.
- We have $e' = f(e) = f\left(a * a^{-1}\right) = f(a) *' f\left(a^{-1}\right) \iff (f(a))^{-1} = f\left(a^{-1}\right)$.

■

Using the above lemma, we can now show the following proposition:

Proposition 4.1.2. *Let $(G,*), (G',*')$ be groups and $f: G \to G'$ be a homomorphism. The image of G under f is a subgroup of $(G',*')$.*

Proof. First, we know that $f(G) \neq \emptyset$, since $f(e) = e' \implies e' \in f(G)$. Moreover, given any $y_1, y_2 \in f(G)$, there exist $x_1, x_2 \in G$ such that $y_1 = f(x_1), y_2 = f(x_2)$, so $y_1 *' y_2 = f(x_1) *' f(x_2) = f(x_1 * x_2) \in f(G)$, so $f(G)$ is closed with respect to $*'$. Finally, given $y \in f(G)$, we have $y^{-1} = f(x)^{-1} = f\left(x^{-1}\right) \in f(G)$, so $f(G)$ is closed with respect to the inverse operation. As a result $(f(G),*') \leq (G',*')$. ■

Though the previous propositions may have seemed to be of little relevance to the topic of this thesis, we can now combine all the above to prove Cayley's theorem, which is fundamental in group theory and offers a glimpse of the importance of permutations from an algebraic standpoint. This result and the next will be the last ones of purely theoretical interest.

Theorem 4.1.2 (Cayley)**.** *Given any group $(G,*)$, it is isomorphic to a subgroup of its symmetric group $(S_G, \circ)$.*

Proof. The proof relies on the observation that, given any $a, b \in G$, the equation $a * x = b$ has a unique solution in G, which we mentioned above as a direct consequence of the definition of groups. Specifically, given any $a \in G$, we define the function $f_a : G \to G$ with $f_a(x) =$

$a * x, \forall x \in G$. Based on the previous remark, f_a is bijective, so it is a permutation of the elements of G. We now define the mapping $\phi : G \to S_G$ with $\phi(a) = f_a, \forall a \in G$. We will show that this mapping is a monomorphism from G to S_G.
First, let $a, b \in G$ and let f_a, f_b be their images through ϕ. We have $\phi(a * b) = f_{a*b}$ where $f_{a*b}(x) = (a * b) * x = a * (b * x) = f_a(f_b(x)) = f_a f_b(x), \forall x \in G$, so ϕ is a homomorphism.
Second, let $a, b \in G$ such that $f_a \equiv f_b$. In that case, given any $x \in G$, we have $f_a(x) = f_b(x) \iff a * x = b * x \iff a = b$, so ϕ is also a monomorphism.
To complete the proof, we exploit the fact that the mapping is onto $\phi(G)$ which we know is a subgroup of S_G. Therefore, we have $(G, *) \cong (\phi(G), \circ) \leq (S_G, \circ)$. ■

Cayley's theorem is notable for demonstrating that the study of the properties of any group is essentially equivalent to studying some symmetric subgroup (though it does not offer a general way of determining which subgroup it is). We can take this one step further and show that the study of the symmetric group of any finite set A is equivalent to studying the properties of the symmetric group S_m where $|A| = m \in \mathbb{N}$. Intuitively, that is obvious because each permutation specifies an ordering of some elements. Having 2 sets with an equal number of elements means that the possible orderings are the same. This is the essence of the following theorem, which expresses the previous reasoning in a formal manner.

Proposition 4.1.3. *Let $A \neq \emptyset$ be a set with $|A| = m \in \mathbb{N}$. We have $(S_A, \circ) \cong (S_m, \circ)$.*

Proof. We know that the set A has the same number of elements as $[m]$, so there is some function $f : A \to [m]$ that is bijective. We now have to define an isomorphic mapping $\phi : S_A \to S_m$. For that reason, we define $\phi(\pi) = f\pi f^{-1}, \forall \pi \in S_A$. This choice may seem strange, but there is a very simple intuition behind it. We want the image of any $\pi \in S_A$ to be some $\pi' \in S_m$, so we exploit f to create a function that takes any element of $[m]$ and maps it to its corresponding element in A, then applies π to it and then maps the result back to $[m]$. We now have to verify the above formally and show that ϕ is indeed an isomorphism.
We first show that the images produced by the previously defined mapping are indeed members of S_m. Let $\pi \in S_A$ and $x_1, x_2 \in [m]$ such that $(\phi(\pi))(x_1) = (\phi(\pi))(x_2) \iff f\left(\pi\left(f^{-1}(x_1)\right)\right) = f\left(\pi\left(f^{-1}(x_2)\right)\right)$. Since f, f^{-1}, π are all $1-1$, the previous implies that $x_1 = x_2$, so $\phi(\pi)$ is $1-1$ for all $\pi \in S_A$. Additionally, for any $y \in [m]$ choosing $x = \left(f\pi^{-1}f^{-1}\right)(y)$ results in $(\phi(\pi))(x) = y$, so so $\phi(\pi)$ is onto for all $\pi \in S_A$. Consequently, $\phi(\pi) \in S_m, \forall \pi \in S_A$.
Now it remains to show that ϕ is an isomorphism. Given $\pi_1, \pi_2 \in S_A$, we have $\phi(\pi_1\pi_2) = f\pi_1\pi_2 f^{-1} = \left(f\pi_1 f^{-1}\right)\left(f\pi_2 f^{-1}\right) = \phi(\pi_1)\phi(\pi_2)$, so we know ϕ is a homomorphism. Apart from that, having $\pi_1, \pi_2 \in S_A$ such that $\phi(\pi_1) = \phi(\pi_2)$ yields $f\pi_1 f^{-1} = f\pi_2 f^{-1} \iff \pi_1 = \pi_2$. Finally, given any $\sigma \in S_m$, by choosing $\pi = f^{-1}\sigma f$, we get $\phi(\pi) = \sigma$. Thus, ϕ is indeed an isomorphism and the proof is complete. ■

As we leave group theory behind us, we should make a review of what we have shown. This can summarized in the following informal statement, which serves as the bottom line (both figuratively and literally) for our treatment of the subject:

All groups are permutation groups. All finite permutation groups are permutation groups of initial segments of $\mathbb{N}$.

4.1.2 Cyclic Permutations

We will now introduce the notion of cyclic permutations which, as we will see, are necessary to understand one of the ranking distances that will be examined in the next section. The simplest example of a cyclic permutation that comes to mind is one where each element has been moved to the position of the next and the first and last elements have switched places. However, this does not fully capture the concept. To do that, we need to define what the orbit of an element under a permutation is.

Definition 4.1.4. Let $m \in \mathbb{N}$ and $\sigma \in S_m$. Given some $a \in [m]$, its orbit under σ is the set of positions where a can be moved if σ or σ^{-1} are applied repeatedly to it. This set is denoted $\mathcal{O}_{a,\sigma} = \left\{b \in [m] \,|\exists i \in \mathbb{Z} : \sigma^i(a) = b\right\}$.

Suppose now that we define a binary relation $\sim$ in $[m]$ where $a \sim b \iff b \in \mathcal{O}_{a,\sigma}$. It is easy to see that this is an equivalence relation. Indeed, we have:

- $a \in \mathcal{O}_{a,\sigma}$, since $\sigma^0(a) = id(a) = a$, so it is reflexive.
- $b \in \mathcal{O}_{a,\sigma} \implies \exists i \in \mathbb{Z} : \sigma^i(a) = b \iff \sigma^{-i}(b) = a \iff a \in \mathcal{O}_{b,\sigma}$, so it is symmetric.
- $b \in \mathcal{O}_{a,\sigma}, c \in \mathcal{O}_{b,\sigma} \iff \exists i,j \in \mathbb{Z} : \sigma^i(a) = b, \sigma^j(b) = c \implies \sigma^{i+j}(a) = c \iff c \in \mathcal{O}_{a,\sigma}$, so it is transitive.

The above implies that, given some $\sigma \in S_m$, the set $[m]$ is partitioned into sets of elements that share the same orbit under σ. These sets are the *cycles* of σ. The number of elements in a cycle are referred to as its *length*. Elements that are not moved by σ (referred to as *fixed points*) belong in cycles of length 1. If σ has at most 1 cycle with more than 1 elements, we say that it's a cyclic permutation. However, even if a permutation is not cyclic, it can be written as a product of its cycles, which are disjoint. Indeed, when multiplying disjoint cycles, the elements belonging to each cycle are not affected by the rest, so the elements of different cycles can be examined independently.
We have introduced all the theoretical background about cyclic permutations that is necessary for the rest of text. Before moving on, however, we should also give a bit of notation commonly used to represent a cycle. Specifically, if we write $\sigma = (i_1, i_2, \ldots, i_k)$, it means that $\sigma(i_1) = i_2, \sigma(i_2) = i_3, \ldots, \sigma(i_k) = i_1$ and the rest of the elements are not affected by σ (equivalently, we may write $i_1 \to i_2 \to \cdots \to i_k \to i_1$). Now, we are indeed ready to proceed.

4.2 Permutation Distances

We are almost ready to start dealing with the ranking models that will be the main topic of this thesis. The last notion that we have to introduce is that of *permutations distances*. Specifically, the Mallows model, which is the model we intend to focus on, favors permutations that are "close" to a given permutation, which is a parameter of the model. For the previous sentence to be mathematically meaningful, we need to define distance measures between permutations. In this section, we will present 5 such distances, though only 2 will be heavily referenced later on. We should remark that the authoritative text on the issue is [15].
Before moving on to specific distance measures, we would like to remind the reader the definition of distance metrics.

Definition 4.2.1. Let $S \neq \emptyset$ and $d : S \times S \to \mathbb{R}_{\geq 0}$. We say that (S, d) is a metric space and that d is a distance metric if the following hold:

- $d(x, y) \geq 0, \forall x, y \in S$ and $d(x, y) = 0 \iff x = y$.
- $d(x, y) = d(y, x), \forall x, y \in S$ (*symmetric*).
- $d(x, y) + d(y, z) \geq d(x, z), \forall x, y, z \in S$ (*triangle inequality*).

In our setting, distances will take values in the set $\mathbb{N} \cup \{0\}$ instead of $\mathbb{R}_{\geq 0}$. As we will see, the above properties are satisfied by all the distances that we will present except 1. Additionally, there exists a property referred to as *right-invariance* that is satisfied by all ranking distances. Specifically, given any permutations $\pi, \sigma, \tau \in S_m$, we have $d(\pi, \sigma) = d(\pi\tau, \sigma\tau)$ for any of the ranking distances that we will encounter.

4.2.1 Kendall Tau Distance

The Kendall Tau (KT) distance is the distance metric most commonly used in the bibliography. Given 2 permutations $\pi, \pi_0 \in S_m$, their KT distance $d_{KT}(\pi, \pi_0)$ is equal to the minimum number of swaps between adjacent elements required to convert π^{-1} to π_0^{-1}. To make this clear, we will examine the following example:

Example 4.2.1.1. Let $\pi, \pi_0 \in S_3$ with $\pi(1) = 1, \pi(2) = 2, \pi(3) = 3$ and $\pi_0(1) = 3, \pi_0(2) = 2, \pi_0(3) = 1$. To simplify the notation, we write: $\pi^{-1} : 1, 2, 3$ and $\pi_0^{-1} : 3, 2, 1$. We have $d_{KT}(\pi, \pi_0) = 3$. Indeed, the process to convert π^{-1} to π_0^{-1} is $1, 2, 3 \to 2, 1, 3 \to 2, 3, 1 \to 3, 2, 1$.

Due to the similarity of the previous process with the bubble-sort algorithm, this distance is sometimes referred to as bubble-sort distance. Another similarity with bubble-sort stems from the fact that the maximum value the KT distance can have for elements of S_m is $\frac{m(m-1)}{2}$, which is number of steps required by bubble sort in the worst case.
Apart from the previous definition we gave, there is another way to view this distance. In particular, it can be shown that it is equal to the number of discordant pairs between π and π_0. To express that formally, we write:

$$\boxed{d_{KT}(\pi, \pi_0) = \sum_{i=1}^{m} \sum_{j=0}^{i-1} \mathbb{1}\{(\pi(i) - \pi(j))(\pi_0(i) - \pi_0(j)) < 0\}}$$

Using this definition and the previous one, it is possible to verify that all the aforementioned properties are satisfied, so equipping S_m with this distance function results in a metric space. Another property that is quite important for this distance is that it is *swap-increasing*. This means that, given 2 permutations $\pi, \pi_0 \in S_m$ that agree in the way a pair of elements a, b compare, the KT distance between the permutation $\pi_{a \leftrightarrow b}$ (which is the same as π with the exception of the pair a, b which has been swapped) and π_0 is greater than the distance between π and π_0 (their difference must be exactly 1 if a, b are adjacent in π). This may seem obvious intuitively, since we have:

$$\mathbb{1}\{(\pi_{a \leftrightarrow b}(a) - \pi_{a \leftrightarrow b}(b))(\pi_0(a) - \pi_0(b)) < 0\} > \mathbb{1}\{(\pi(a) - \pi(b))(\pi_0(a) - \pi_0(b)) < 0\}$$

However, this fails to take into consideration the contribution of terms corresponding to elements that are placed between a and b by π. For that reason, we will present a formal proof that the property holds. Note that our proof will be slightly different than the one presented in [10], where the notion of *swap-increasingness* was first introduced.

Lemma 4.2.1 (Caragiannis et. al. (2013))**.** *The KT distance is swap-increasing.*

Proof. Let $\pi, \pi_0 \in S_m$. Assume, without loss of generality that, for some $a, b \in [m]$, we have $\pi_0(a) < \pi_0(b)$ and $\pi(a) < \pi(b)$. In $\pi_{a \leftrightarrow b}$, they only comparisons that may change are those involving either a or b. This motivates us to partition the elements of $[m]$ into 3 sets based on their position in π with respect to a and b. The elements that do not lie between a and b are of no interest to us. Indeed, swapping a and b does not affect the way they compare with them. Now, let $Y = \{y \in [m] : \pi(a) < \pi(y) < \pi(b)\}$. Based on the above, we can write:

$$d_{KT}(\pi_{a \leftrightarrow b}, \pi_0) - d_{KT}(\pi, \pi_0) = 1 + \sum_{y \in Y} A(y)$$

where:

$$A(y) = \mathbb{1}\{(\pi_{a \leftrightarrow b}(a) - \pi_{a \leftrightarrow b}(y))(\pi_0(a) - \pi_0(y)) < 0\} +$$

$$+\mathbb{1}\{(\pi_{a \leftrightarrow b}(b) - \pi_{a \leftrightarrow b}(y))(\pi_0(b) - \pi_0(y)) < 0\} - \mathbb{1}\{(\pi(a) - \pi(y))(\pi_0(a) - \pi_0(y)) < 0\} -$$

$$-\mathbb{1}\{(\pi(b) - \pi(y))(\pi_0(b) - \pi_0(y)) < 0\}$$

The above expression can be simplified based on the facts that $\pi(a) < \pi(y) < \pi(b), \pi_0(a) < \pi_0(b)$ and $\pi_{a \leftrightarrow b}(a) = \pi(b), \pi_{a \leftrightarrow b}(b) = \pi(a), \pi_{a \leftrightarrow b}(y) = \pi(y)$. Specifically, we have:

$$A(y) = (\mathbb{1}\{\pi_0(a) < \pi_0(y)\} + \mathbb{1}\{\pi_0(b) > \pi_0(y)\}) -$$

$$-(\mathbb{1}\{\pi_0(a) > \pi_0(y)\} + \mathbb{1}\{\pi_0(b) < \pi_0(y)\})$$

Suppose that for some $y \in Y$ we have $A(y) < 0$. Based on the above expression, $A(y) \in \{0, \pm 1, \pm 2\}$. If $A(y) = -2$, we would have $\pi_0(b) < \pi_0(y) < \pi_0(a)$, which violates the assumption that $\pi_0(a) < \pi_0(b)$. If $A(y) = -1$, both terms in the second line should be 1 while exactly one term in the first line should be 1. If $\mathbb{1}\{\pi_0(a) < \pi_0(y)\} = 1$, then $\mathbb{1}\{\pi_0(a) > \pi_0(y)\} = 0$. The same is true for the other term, so $A(y) \neq -1$. This leads to a contradiction, so $A(y) \geq 0, \forall y \in Y$, which yields $d_{KT}(\pi_{a \leftrightarrow b}, \pi_0) - d_{KT}(\pi, \pi_0) \geq 1$. The equality holds either when $Y = \emptyset$ (a and b are adjacent in π) or when $\forall y \in Y : (\pi_0(y) > \pi_0(b)) \vee (\pi_0(y) < \pi_0(a))$. ■

We will now present a way to decompose the KT distance in terms that offer us a different way of representing permutations. The starting point for all this is that the second definition we gave can be written in the form:

$$d_{KT}(\pi, \pi_0) = \sum_{i=1}^{m} V_i(\pi, \pi_0)$$

where:

$$V_i(\pi, \pi_0) = \sum_{j=0}^{i-1} \mathbb{1}\{(\pi(i) - \pi(j))(\pi_0(i) - \pi_0(j)) < 0\}$$

Note that $V_i(\pi, \pi_0) \leq i - 1$. This implies that the vector $(V_1(\pi, \pi_0), \ldots, V_m(\pi, \pi_0))$ takes values in the set $\{0\} \times \{0, 1\} \times \cdots \times \{0, \ldots, m-1\}$, which has $m! = |S_m|$ elements. This leads to the following representation, pioneered by Hall in 1956 (as is mentioned in [31]):

Proposition 4.2.1 (Hall). *The mapping $\pi \mapsto (V_1(\pi, \pi_0), \ldots, V_m(\pi, \pi_0))$ is a bijection from S_m to the set $\{0\} \times \{0, 1\} \times \cdots \times \{0, \ldots, m-1\}$.*

This last property, along with swap-increasingness, will prove to be indispensable to prove the results that are referenced in chapter 5.

4.2.2 Cayley Distance

The other distance that will draw our attention is the Cayley distance. It is similar to the KT distance in the sense that it involves the minimum number of swaps required to convert one permutation to another. However, unlike the previous distance, the swaps do not need to be between adjacent elements. The properties of distance metrics as well as right-invariance are satisfied by this distance.

We will now prove a formula that gives us an easier way to calculate the Cayley distance between some $\pi, \pi_0 \in S_m$. Specifically, it is possible to compute the value $d_{ca}(\pi, \pi_0) = d_{ca}\left(\pi\pi_0^{-1}, id\right)$ based on the remark that each permutation can be written as a product of disjoint cycles. Suppose that $\pi\pi_0^{-1}$ can be written as a product of k cycles, each comprising of $x_1, x_2, \ldots, x_k \geq 1$ elements with $x_1 + x_2 + \cdots + x_k = m$. The key observation here is that it is not necessary to transpose elements belonging in different cycles. Instead, the optimal way of working to convert $\pi\pi_0^{-1}$ into the identity permutation is by treating the elements of each cycle separately. Now, to order the elements of a cycle with x_i elements, it is necessary to perform $x_i - 1$ swaps, so overall we have $d_{ca}(\pi, \pi_0) = d_{ca}\left(\pi\pi_0^{-1}, id\right) = \sum_{i=1}^{k}(x_i - 1) = m - k$ (note that it is possible to perform the computation directly for π instead of $\pi\pi_0^{-1}$ simply by relabelling the elements of id according to π_0). A consequence of this is that the Cayley distance can be at most $m - 1$, since there has to be at least one cycle.

Furthermore, we will present a property of the Cayley distance that is similar to swap-increasingness and will be crucial for our work in chapter 6. We should note that, despite the property's simplicity, we have not found any reference to it in the bibliography.

Lemma 4.2.2. *Suppose we are given permutations $\pi, \pi_0 \in S_m$ such that, for some element $i \in [m]$, we have $\pi(i) = \pi_0(i)$. For any permutation π' that is produced by taking π and transposing i with some other element we have $d_{ca}(\pi', \pi_0) = d_{ca}(\pi, \pi_0) + 1$.*

Proof. The relation we want to prove means that π' has exactly one cycle less than π with respect to the positions of the elements in π_0. The result is obvious if the element with which

i was swapped is in its correct position. Indeed, all the other elements are not affected while the 2 elements that are transposed form a cycle of length 2 (while at first each belonged in a cycle of length 1). Consequently, the number of cycles decreases by exactly 1.
The statement is equally easy to prove when the element with which i was swapped belongs in a cycle of length $k \geq 2$. Suppose that the elements that are part of that cycle are $i_1, i_2, \ldots, i_k \in [m]$. Using notation from abstract algebra, we write the cycle in the form $(\pi_0(i_1), \pi_0(i_2), \ldots, \pi_0(i_k))$, which implies that i_1 is moved in position $\pi_0(i_2)$, i_2 is moved in position $\pi_0(i_3)$ etc by π. Suppose that the element participating in the swap was $i_l, 1 \leq l \leq k$. Due to the cycle's nature, i_l occupies the position $\pi_0(l \pmod k + 1)$. As a result of the swap, element i_l ends up occupying the position $\pi_0(i)$ while i is moved to the position $\pi_0(l \pmod k + 1)$. Consequently, the cycle becomes:

$$(\pi_0(i_1), \pi_0(i_2), \ldots, \pi_0(i_l), \pi_0(i), \pi_0(l \pmod k + 1), \ldots, \pi_0(i_k))$$

while the rest of the cycles are unaffected, meaning that we have exactly 1 cycle less. ■

Finally, we will present a useful decomposition of the Cayley distance. We write $d_{ca} = \sum_{i=1}^{m} V_i(\pi, \pi_0)$ where $V_i(\pi, \pi_0) \in \{0, 1\}$ where $V_i(\pi, \pi_0)$ becomes 0 when the element i is the one with the biggest index in the cycle where it belongs in π (having π_0 as a reference and not the identity permutation, as is usually the case). With the above definition, we manage to have as many 0s in the sum as the number of cycles, which verifies its correctness (note that $V_m(\pi, \pi_0) \equiv 0$, which commonly results in terms involving alternative m being omitted). Obviously, in this case, the mapping $\pi \mapsto (V_1(\pi, \pi_0), \ldots, V_m(\pi, \pi_0))$ is not a bijection, since there are $m!$ permutations, while there are only 2^{m-1} vectors that satisfy the previous conditions. Instead, we have the following proposition, which can be proved using counting arguments:

Proposition 4.2.2. *Given a vector $(v_1, \ldots, v_m) \in \{0, 1\}^{m-1} \times \{0\}$ and a permutation $\pi_0 \in S_m$ that serves as a reference, there are $\prod_{i=1}^{m-1} (m - i)^{v_i}$ permutations $\pi \in S_m$ such that $(V_1(\pi, \pi_0), \ldots, V_m(\pi, \pi_0)) = (v_1, \ldots, v_m)$.*

4.2.3 Other Distances

We will now make a short reference to other distances between permutations which will not be used in the rest of the text. These are Spearman's measures and the Hamming distance.
There are 2 measures attributed to Spearman, namely Spearman's footrule and Spearman's rank correlation. They both rely on the fact that permutations can be represented as vectors in the form $(\pi(1), \ldots, \pi(m))$. Given that representation, Spearman's footrule is the ℓ_1 distance (and thus satisfies all properties of distance metrics), while Spearman's rank correlation is the square of the ℓ_2 distance (which is the reason that this distance does not satisfy the triangle inequality). Both measures are right-invariant. The most important work involving Spearman's footrule is [16], where many of the metric's properties were established.
The Hamming distance of 2 permutations $\pi, \pi_0 \in S_m$ is equal to the number of elements that are assigned to different positions. Equivalently, we can say that it is equal to $m-$ (the number of fixed points in $\pi\pi_0^{-1}$). It satisfies all the aforementioned properties.

4.3 Ranking Distributions

In this section we will introduce the concept of ranking distributions. As we mentioned back in chapter 2, these models are inspired by social choice theory. By Proposition 4.1.3, given any set of m alternatives, its permutation group is isomorphic to S_m. For that reason, the support of all the distribution models we will study will be S_m. We now define the Mallows model, which is the one on which we will focus. After that, we will, for the sake of completeness, make a short presentation of the Plackett-Luce model.

4.3.1 The Mallows Model

The Mallows model with m alternatives is a probability distribution defined over the symmetric group S_m. It is parameterized by a central ranking $\pi_0 \in S_m$ and a spread parameter $\phi \in (0, 1)$ (the cases where $\phi = 1$ or $\phi = 0$ are degenerate and correspond to the uniform distribution and the constant one, respectively). It was introduced in 1957 in [36]. Specifically, the model was conceived as an analogue to the normal distribution for ranking distributions. In particular, the probability of a ranking occurring decreases exponentially according to its distance from the central ranking while the spread parameter plays a role similar to that of the variance. The pmf of the model is:

$$\boxed{\mathbb{P}\left[\pi = \sigma\right] = \frac{1}{Z\left(\phi\right)}\phi^{d(\sigma,\pi_0)}}$$

where d is a ranking distance (usually the KT distance) and $Z\left(\phi\right)$ is a normalizing constant that depends on ϕ and the ranking distance (but not the central ranking). In the case of the KT distance, the normalizing constant has the form:

$$\boxed{Z\left(\phi\right) = \prod_{i=1}^{m} Z_i\left(\phi\right) = \prod_{i=1}^{m}\left(\sum_{j=0}^{i-1}\phi^j\right) = \frac{1}{\left(1-\phi\right)^{m-1}}\prod_{i=2}^{m}\left(1-\phi^i\right)}$$

while in that of the Cayley distance we have:

$$\boxed{Z\left(\phi\right) = \prod_{i=1}^{m} Z_i\left(\phi\right) = \prod_{i=1}^{m}\left[1 + \left(m-i\right)\phi\right]}$$

A systematic way of computing those constants using generating functions can be found in [22]. However, the only cases where the result can be expressed in known expression are those above and that of the Hamming distance. Indeed, for Spearman's metrics, only approximations are known which are valid for $m \to \infty$ (see [41]). Moreover, the 2 above examples are the only ones where it is possible to factorize the constant in a manner that can be used to construct an iterative process that generates samples from the model (see [19]). For that reason, Flinger and Verducci proposed a generalization in [22], which eventually became known as the Generalized Mallows model and argued that it is meaningful only in the cases of the KT and the Cayley distances. In that generalization, a different spread parameter is assigned to each alternative. As a result, the pmf has the form:

$$\boxed{\mathbb{P}\left[\pi = \sigma\right] = \prod_{i=1}^{m}\frac{\phi_i^{V_i(\pi,\pi_0)}}{Z_i\left(\phi_i\right)}}$$

where $V_i(\pi, \pi_0)$ are the terms to which the distance metric we use is decomposed to (see Propositions 4.2.1 and 4.2.2). For more information on the above, a good reference is [37]. Finally, there is one last generalization to consider, which was recently introduced in [9]. That is the Mallows block model, where the set of alternatives is partitioned into d blocks and a spread parameter is assigned to the alternatives of each block. The block structure is denoted $\boldsymbol{B} = (B_1, B_2, \ldots, B_d)$, we have $\boldsymbol{\phi} \in [0,1]^d$ and the pmf becomes:

$$\mathbb{P}[\pi = \sigma] = \prod_{i=1}^{d} \frac{\phi_i^{T_i(\pi, \pi_0, \boldsymbol{B})}}{Z^i(\phi_i, \boldsymbol{B})}$$

where:

$$T_i(\pi, \pi_0, \boldsymbol{B}) = \sum_{j \in B_i} V_j(\pi, \pi_0)$$

$$Z^i(\phi_i) = \prod_{j \in B_i} Z_j(\phi_i)$$

For the rest of the text, instances of the Mallows block distribution will be denoted $\mathcal{P}_{\boldsymbol{\phi}, \pi_0, \boldsymbol{B}}$. The family of all d-block Mallows distributions with block structure $\boldsymbol{B}$ is denoted $\mathcal{M}_d(\boldsymbol{B}) = \left\{\mathcal{P}_{\boldsymbol{\phi}, \pi_0, \boldsymbol{B}} : \boldsymbol{\phi} \in [0,1]^d, \pi_0 \in S_m\right\}$. In the cases of the simple Mallows model and the Generalized Mallows model, the same notation will be used but the subscript $\boldsymbol{B}$ will be dropped.

4.3.2 The Plackett-Luce Model

The Plackett-Luce model was introduced independently by the people it is named after (see [43, 35]). It is different from the Mallows model in the sense that, instead of being parameterized by a central ranking and one or more spread parameters, it is described using a weight vector $\boldsymbol{w} = (w_1, \ldots, w_m) \in [0,1]^m$ with $\sum_{i=1}^m w_i = 1$. Specifically, the greater the value of the weight corresponding to alternative i, the more likely it is to be preferred over the rest in the samples generated by the model. The sample generation process is performed in m rounds, where in round i the alternative that will be placed in position i is picked with probability that is proportional to its weight. The pmf of the model is:

$$\boxed{\mathbb{P}[\pi = \sigma] = \prod_{i=1}^{m} \frac{w_{\sigma^{-1}(i)}}{\sum_{j=i}^{m} w_{\sigma^{-1}(j)}}}$$

It can be shown that, given some $i, j \in [m]$, the probability that i is placed before j is:

$$\boxed{\mathbb{P}[\pi(i) < \pi(j)] = \frac{w_i}{w_i + w_j}}$$

It is interesting to compare this to the Mallows model. As mentioned previously, the concept of a central ranking does not exist in this model. Instead, the closest thing there is to one can be obtained by sorting the elements in decreasing weight order. However, there is the issue of how ties are resolved in cases where a number of elements share the same weight. Additionally, there is no parameter that determines the variance of the model. Instead, the model exhibits greater variance when the weight vector is close to a vector whose coordinates are equal to $\frac{1}{m}$, which results in a distribution that is close to a uniform over S_m.

4.4 Parameter Estimation in the Mallows Model

In this short section, we will attempt to connect the ideas presented in chapter 3 with the material of this chapter and pave the way for the next. Specifically, we will examine parameter estimation under the Mallows model (the form of the MLE is the same for any distance metric).

4.4.1 The MLE for the Central Ranking

Let $\mathcal{P}_{\phi,\pi_0}, \phi \in (0,1)$ be a simple Mallows distribution and $\boldsymbol{\pi} \sim \mathcal{P}^n_{\phi,\pi_0}$. We have:

$$\hat{\pi} = \underset{\pi_0}{argmax} \left\{ L\left(\pi_0 | \boldsymbol{\pi}\right) \right\} = \underset{\pi_0}{argmax} \left\{ \prod_{i=1}^{n} \frac{\phi^{d(\pi_i,\pi_0)}}{Z\left(\phi\right)} \right\} = \underset{\pi_0}{argmax} \left\{ \phi^{\sum_{i=1}^{n} d(\pi_i,\pi_0)} \right\} =$$

$$= \underset{\pi_0}{argmin} \left\{ \sum_{i=1}^{n} d\left(\pi_i, \pi_0\right) \right\}$$

Remark that the above is the median of $\pi_1, \ldots, \pi_n$ with respect to the distance d. However, the support of the distribution is S_m and computing the median of the samples would require us to perform discrete search over a set with $m!$ elements. Intuitively, this appears to be impossible in polynomial time. Indeed, in the case of the KT distance, it was proven in [5] that computing the MLE is NP-Hard with reduction from the Feedback Arc Set problem. For the Cayley distance the problem has not been shown to be NP-hard, though it believed to be so (see [44]). We believe this to be reasonable, due to the fact that, as we will see in chapter 6, this version of the model exhibits weaker concentration than the one with the KT distance.

4.4.2 The MLE for the Spread Parameters

As before, let $\boldsymbol{\pi} \sim \mathcal{P}^n_{\phi,\pi_0}$. Suppose, for simplicity, that π_0 is known. We have:

$$\hat{\phi} = \underset{\phi}{argmax} \left\{ L\left(\phi | \boldsymbol{\pi}\right) \right\} = \underset{\phi}{argmax} \left\{ \prod_{i=1}^{n} \frac{\phi^{d(\pi_i,\pi_0)}}{Z\left(\phi\right)} \right\} = \underset{\phi}{argmax} \left\{ \frac{\phi^{\sum_{i=1}^{n} d(\pi_i,\pi_0)}}{\left(Z\left(\phi\right)\right)^m} \right\} =$$

$$= \underset{\phi}{argmax} \left\{ \left(\sum_{i=1}^{n} d\left(\pi_i, \pi_0\right) \right) \ln\left(\phi\right) - m \ln\left(Z\left(\phi\right)\right) \right\}$$

Taking the derivative with respect to ϕ we get:

$$\left(\sum_{i=1}^{n} d\left(\pi_i, \pi_0\right) \right) \frac{1}{\hat{\phi}} - m \frac{Z'\left(\hat{\phi}\right)}{Z\left(\hat{\phi}\right)} = 0 \iff \hat{\phi} \frac{Z'\left(\hat{\phi}\right)}{Z\left(\hat{\phi}\right)} = \frac{1}{m} \left(\sum_{i=1}^{n} d\left(\pi_i, \pi_0\right) \right)$$

However, the normalizing constant has a known expression only in the cases of the KT, Cayley and Hamming distances. Even then, it is impossible to get a closed form for the solution. Moreover, the distribution of $d\left(\pi_i, \pi_0\right)$ is generally not a known one, so we cannot prove any guarantees about the estimates produced by the MLE. This constitutes another example where exploiting the MLE is not as straightforward as we could have hoped for, though for entirely different reasons than before.

Chapter 5

Learning in the Kendall-Mallows Model

The Mallows model equipped with the Kendall tau distance (henceforth referred to as the *Kendall-Mallows* model) is generally regarded as the most "natural" version of the model. That is because this distance metric imbues it with a strong concentration around the central ranking, resulting in a behavior similar to that of the normal distribution. Moreover, as we explained in chapter 4, the model is also important in social choice theory and this distance is the most suitable for that setting. In this chapter, we will present the works that have mostly influenced ours. We will focus only in problems involving the central ranking and the spread parameters. We will not examine the problem of learning mixtures of Kendall-Mallows models (though the reader should turn to [4] and [34] if they are interested).
In the next 2 sections, we will make a survey of past work concerning the estimation of the central ranking and the spread parameters, respectively, with greater focus on [10] and [9], whose approaches have mostly influenced ours. The proofs of the results of [9] will not be presented in detail, due to the fact that most of them involve lengthy computations. Instead, we will restrict ourselves to sketches of proofs.

5.1 Recovering the Central Ranking

The work where the sample complexity of recovering the central ranking was essentially settled is that of Caragiannis et. al. in [10]. There, they approach the problem in the context of computational social choice, motivated by the noisy comparisons model defined in the 18th by the Marquis de Condorcet, where a ranking is given and samples are generated by inverting the way elements compare with probability $1-p, p > \frac{1}{2}$. This model is known to be equivalent to the Kendall-Mallows model. To better understand that, consider a representation of each ranking as an *acyclic tournament graph*, where vertices correspond to alternatives and there is a single edge for any pair of vertices the direction of which expresses the way the corresponding alternatives compare. For example, the graph corresponding to $\pi^{-1} : 1, 2, 3$ is:
The question now is why choose this representation instead of a path for example. The answer is given by the fact that such a representation facilitates the sample generation process of Condorcet's model, which is described by the following algorithm:

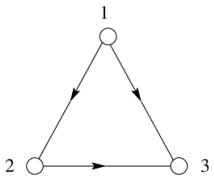

Figure 5.1: Graph representation of the identity element of S_3.

Algorithm 1: Condorcet Sample Generation

Data: $\pi_0 \in S_m, p \in \left(\frac{1}{2}, 1\right)$
Result: $\pi \in S_m$
construct the corresponding acyclic tournament graph;
for *each pair of vertices in* $[m]$ **do**
  preserve the direction of the edge between them with independent probability p;
end
if *the resulting graph contains cycles* **then**
  restart the process;
else
  return the corresponding ranking π;

Given $\pi \in S_m$, since the KT distance is equal to the number of discordant pairs, the probability of π being generated in a **single iteration of the algorithm**:

$$p^{\binom{m}{2}-d_{KT}(\pi,\pi_0)}(1-p)^{d_{KT}(\pi,\pi_0)} = p^{\binom{m}{2}}\left(\frac{1-p}{p}\right)^{d_{KT}(\pi,\pi_0)} =$$

$$= p^{\binom{m}{2}}\phi^{d_{KT}(\pi,\pi_0)}$$

where $\phi = \phi(p) = \frac{1-p}{p} \in (0,1)$ is a strictly decreasing function of $p \in \left(\frac{1}{2}, 1\right)$ (its inverse is $p = p(\phi) = \frac{1}{1+\phi}$). If the probability of an iteration outputting a valid ranking is denoted a, then the probability of π being generated after an arbitrary number of iterations is:

$$\boxed{\sum_{i=0}^{\infty}(1-a)^i\left(\frac{1}{1+\phi}\right)^{\binom{m}{2}}\phi^{d_{KT}(\pi,\pi_0)} = \frac{1}{a}\left(\frac{1}{1+\phi}\right)^{\binom{m}{2}}\phi^{d_{KT}(\pi,\pi_0)} = \frac{\phi^{d_{KT}(\pi,\pi_0)}}{Z(\phi)}}$$

where $Z(\phi) = a(1+\phi)^{\binom{m}{2}}$.
The above verify the equivalence of the 2 models and give a sample generation process for the Mallows model. However, it is rather impractical, since the average number of iterations required for it to produce a sample are $\frac{1}{a} = \frac{(1+\phi)^{\binom{m}{2}}}{Z(\phi)}$, so the method of [19] is used instead.

5.1.1 Concentration in Kendall-Mallows

The previous method motivates us to compute the probability of a comparison being preserved in a sample. This results in the following lemma:

Lemma 5.1.1. *In the simple Kendall-Mallows model, given $a, b \in [m]$ such that $\pi_0(a) < \pi_0(b)$, we have $p_{\pi_0(a)<\pi_0(b)} \geq \frac{1}{1+\phi} = p > \frac{1}{2}$.*

Proof. Given $\mathcal{P}_{\phi,\pi_0}$ and $a, b \in [m]$ such that $\pi_0(a) < \pi_0(b)$. We have:

$$p_{\pi_0(a)<\pi_0(b)} = \mathop{\mathbb{P}}_{\pi\sim\mathcal{P}_{\phi,\pi_0}}[\pi(a) < \pi(b)] = \sum_{\substack{\sigma\in S_m \\ \sigma(a)<\sigma(b)}} \mathop{\mathbb{P}}_{\pi\sim\mathcal{P}_{\phi,\pi_0}}[\pi = \sigma] = 1 - \sum_{\substack{\sigma\in S_m \\ \sigma(a)>\sigma(b)}} \mathop{\mathbb{P}}_{\pi\sim\mathcal{P}_{\phi,\pi_0}}[\pi = \sigma] \quad (1)$$

Remark that, for any ranking that preserves the comparison of a, b, there is exactly one ranking were that pair is inverted and the rest are the same. By Lemma 4.2.1, given $\sigma \in S_m$ such that $\sigma(a) < \sigma(b)$, we have:

$$d_{KT}(\sigma_{a\leftrightarrow b}, \pi_0) \geq d_{KT}(\sigma, \pi_0) + 1 \iff \phi^{d_{KT}(\sigma_{a\leftrightarrow b},\pi_0)} \leq \phi^{d_{KT}(\sigma,\pi_0)+1} \iff$$

$$\iff \mathop{\mathbb{P}}_{\pi\sim\mathcal{P}_{\phi,\pi_0}}[\pi = \sigma_{a\leftrightarrow b}] \leq \phi \mathop{\mathbb{P}}_{\pi\sim\mathcal{P}_{\phi,\pi_0}}[\pi = \sigma] \quad (2)$$

Combining (1) and (2), we get:

$$p_{\pi_0(a)<\pi_0(b)} = 1 - \sum_{\substack{\sigma\in S_m \\ \sigma(a)<\sigma(b)}} \mathop{\mathbb{P}}_{\pi\sim\mathcal{P}_{\phi,\pi_0}}[\pi = \sigma_{a\leftrightarrow b}] \geq 1 - \phi \sum_{\substack{\sigma\in S_m \\ \sigma(a)<\sigma(b)}} \mathop{\mathbb{P}}_{\pi\sim\mathcal{P}_{\phi,\pi_0}}[\pi = \sigma] =$$

$$= 1 - \phi p_{\pi_0(a)<\pi_0(b)} \implies \boxed{p_{\pi_0(a)<\pi_0(b)} \geq \frac{1}{1+\phi} = p > \frac{1}{2}}$$

■

Remark that repetitions do not cause a decrease in the probability of a comparison being preserved. This has a very simple interpretation. Suppose that we have an instance where $p \to 1^-$. In that case, we start from a valid ranking and few inversions take place, resulting in what will probably be another valid ranking. On the other hand, if $p \to \frac{1}{2}^+$, we compensate for having a non-negligible probability of an invalid result by performing multiple iterations. Additionally, despite choosing independently each edge's direction in each iteration, the comparisons in the resulting sample are not independent (and they shouldn't be- if we take $\pi_0^{-1} : 1, 2, 3$ and swap 1 and 3, this inadvertently causes the comparisons of those elements with 2 to change as well, so they can't, by any means, be independent). Finally, not all swaps have the same probability of occurring at the end.
Exploiting the fact that $p_{\pi_0(a)<\pi_0(b)} + p_{\pi_0(a)>\pi_0(b)} = 1$, we get the following corollary:

Corollary 5.1.1. *In the simple Kendall-Mallows model, given $a, b \in [m]$ such that $\pi_0(a) < \pi_0(b)$, we have $p_{\pi_0(a)>\pi_0(b)} \leq \frac{\phi}{1+\phi} < 1 - p < \frac{1}{2}$ and $\delta_{ab} \geq \frac{1-\phi}{1+\phi}$ where $\delta_{ab} = p_{\pi_0(a)<\pi_0(b)} - p_{\pi_0(a)>\pi_0(b)}$.*

5.1.2 Approximating the MLE

The previous results motivate us to estimate the central ranking by focusing on pair-wise comparisons. The idea is to construct a ranking by examining each pair and seeing how they compare in the majority of samples. However, the previous approach does not define a single algorithm, since we have to resolve the issues of tie-breaking and cycle formation. Solving these defines an algorithm (or *voting rule*, as it would be referred in the context of social choice). This approach defines a family of rules referred to as pairwise-majority consistent (PM-c). We give a high level description of the algorithms of that family:

Algorithm 2: MLE approximation via PM-c rules

Data: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n) \sim \mathcal{P}^n_{\phi,\pi_0}$
Result: $\hat{\pi}$
for *each* $a \in [m]$ **do**
  for *each* $b \in [m] \setminus \{a\}$ **do**
    determine the way $\hat{\pi}(a), \hat{\pi}(b)$ compare based on the majority of the samples (break ties arbitrarily);
  end
end
if *the resulting graph is not acyclic* **then**
  apply rule to convert it to an acyclic tournament graph;
convert the graph to a ranking $\hat{\pi}$;
return $\hat{\pi}$;

The time complexity of the above family is $\mathcal{O}\left(nm^2\right)$ (supposing that tie-breaking takes negligible time compared to the other steps).

5.1.3 Sample Complexity Analysis

Having introduced PM-c rules, we can now present the main results of Caragiannis et. al.:

Theorem 5.1.2 (Caragiannis et. al. (2013))**.** *For any $\epsilon \in (0, 1]$, any PM-c rule determines the true ranking with probability at least $1-\epsilon$ given $\mathcal{O}\left(\log\left(\frac{m}{\epsilon}\right)\right)$ samples from a simple Kendall-Mallows model.*

Proof. Let $\mathcal{P}_{\phi,\pi_0}$ be an instance of the model and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n) \sim \mathcal{P}^n_{\phi,\pi_0}$ be n iid samples from it. We write $n_{ab} = \sum_{i=1}^{n} \mathbb{1}\{\pi_i(a) < \pi_i(b)\}$ for any distinct $a, b \in [a, b]$. The probability of error of any PM-c rule whose result is denoted $\hat{\pi}$ is:

$$\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}[\hat{\pi} \neq \pi_0] = \mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}[\exists a, b \in [m] : \pi_0(a) < \pi_0(b) \wedge n_{ab} \leq n_{ba}] \leq$$

$$\leq \sum_{a\in[m]} \sum_{b\in[m]\setminus\{a\}} \mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}[\pi_0(a) < \pi_0(b) \wedge n_{ab} \leq n_{ba}]$$

Given a pair of a, b such that $\pi_0(a) < \pi_0(b)$, we have:

$$\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}[n_{ab}\leq n_{ba}] = \mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}\left[\frac{n_{ab}-n_{ba}}{n}\leq 0\right]\leq$$

$$\leq \mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}\left[\left|\frac{n_{ab}-n_{ba}}{n}-\mathop{\mathbb{E}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}\left[\frac{n_{ab}-n_{ba}}{n}\right]\right|\geq \mathop{\mathbb{E}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}\left[\frac{n_{ab}-n_{ba}}{n}\right]\right]$$

where $\mathop{\mathbb{E}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}\left[\frac{n_{ab}-n_{ba}}{n}\right]=\delta_{ab}\geq\frac{1-\phi}{1+\phi}=\delta$ (by Corollary 5.1.1). By the Hoeffding bound, we get:

$$\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}[n_{ab}\leq n_{ba}]\leq 2e^{-2\delta^2 n}\implies \mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}[\hat{\pi}\neq\pi_0]\leq m^2e^{-2\delta^2 n}$$

Demanding that to be less than ϵ, we get $n\geq\frac{1}{2\delta^2}\ln\left(\frac{m^2}{\epsilon}\right)$. Since we have $\delta=\Theta(1)$, we get that $n=\Omega\left(\log\left(\frac{m}{\epsilon}\right)\right)$ samples suffice for the probability of error to be less than ϵ. ■

The time complexity of PM-c rules is $\Theta\left(nm^2\right)=\Theta\left(m^2\log\left(\frac{m}{\epsilon}\right)\right)$, assuming that the time required by the tie-breaking steps and post-processing of the result are negligible compared to the main procedure (which is a reasonable hypothesis).

We remark that the previous result holds even if the samples are drawn from simple Kendall-Mallows distributions with the same central ranking but unequal spread parameters. Indeed, the Hoeffding bound does not require the involved random variables to identically distributed, just independent. The only thing that would change is δ, which would have to be defined as $\delta=\min_{i\in[n]}\frac{1-\phi_i}{1+\phi_i}$. This was first remarked in [49].

This result comes with a matching lower bound. In order to present the proof, we need to define the *accuracy* of a (randomized) voting rule r given n samples. Given an instance of the model where the central ranking is π_0 the accuracy of the voting rule r given n samples $\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}$ is defined as $Acc^r(n,\pi_0)=\sum_{\boldsymbol{\sigma}\in S_m^n}\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}[\boldsymbol{\pi}=\boldsymbol{\sigma}]\,\mathbb{P}[r(\boldsymbol{\sigma})=\pi_0]$ (the second probability is computed with respect to r). The accuracy of r given n samples is $Acc^r(n)=\min_{\pi_0\in S_m}Acc^r(n,\pi_0)$.

Theorem 5.1.3 (Caragiannis et. al. (2013))**.** *For any $\epsilon\in\left(0,\frac{1}{2}\right]$, any (randomized) voting rule requires $\Omega\left(\log\left(\frac{m}{\epsilon}\right)\right)$ samples from a simple Kendall-Mallows model to determine the true ranking with probability at least $1-\epsilon$.*

Proof. Though we do not have to resort to minimax theory, the idea of the proof is similar to that of the lower bounds presented in chapter 3: we need to find the instances that would make any algorithm struggle. For that reason, suppose that we have an instance of the model where the central ranking is π_0. We define the set $\mathcal{N}(\pi_0)=\{\sigma\in S_m : d_{KT}(\sigma,\pi_0)=1\}$. These are the rankings that can be constructed by swapping a single pair of adjacent elements of π_0 (the "neighbors" of π_0). We have $|\mathcal{N}(\pi_0)|=m-1$. The instances of the model that correspond to the elements of $\mathcal{N}(\pi_0)$ are those that, intuitively, should be harder for any voting rule to tell apart. Given some $\sigma_0\in\mathcal{N}(\pi_0)$, the triangle inequality yields $d_{KT}(\pi,\pi_0)\leq d_{KT}(\pi,\sigma_0)+d_{KT}(\sigma_0,\pi_0)=d_{KT}(\pi,\sigma_0)+1,\forall\pi\in S_m$. For any $\boldsymbol{\sigma}\in S_m^n$, we have:

$$\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}[\boldsymbol{\pi}=\boldsymbol{\sigma}]=\prod_{i=1}^n\frac{\phi^{d_{KT}(\sigma_i,\pi_0)}}{Z(\phi)}\geq\prod_{i=1}^n\frac{\phi^{d_{KT}(\sigma_i,\sigma_0)+1}}{Z(\phi)}=\phi^n\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\sigma_0}}[\boldsymbol{\pi}=\boldsymbol{\sigma}]$$

Let r be a voting rule such that $Acc^r(n) \geq 1-\epsilon \implies Acc^r(n,\pi_0) \geq 1-\epsilon$. We have:

$$1-\epsilon \leq Acc^r(n,\pi_0) = \sum_{\boldsymbol{\sigma}\in S_m^n} \mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}} [\boldsymbol{\pi}=\boldsymbol{\sigma}]\,\mathbb{P}[r(\boldsymbol{\sigma})=\pi_0] =$$

$$= \sum_{\boldsymbol{\sigma}\in S_m^n} \mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}} [\boldsymbol{\pi}=\boldsymbol{\sigma}]\,(1-\mathbb{P}[r(\boldsymbol{\sigma})\neq\pi_0]) = 1-\sum_{\boldsymbol{\sigma}\in S_m^n} \mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}} [\boldsymbol{\pi}=\boldsymbol{\sigma}]\,\mathbb{P}[r(\boldsymbol{\sigma})\neq\pi_0] \leq$$

$$\leq 1-\sum_{\sigma_0\in\mathcal{N}(\pi_0)}\sum_{\boldsymbol{\sigma}\in S_m^n} \mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}} [\boldsymbol{\pi}=\boldsymbol{\sigma}]\,\mathbb{P}[r(\boldsymbol{\sigma})=\sigma_0] \leq$$

$$\leq 1-\phi^n\sum_{\sigma_0\in\mathcal{N}(\pi_0)}\sum_{\boldsymbol{\sigma}\in S_m^n} \mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\sigma_0}} [\boldsymbol{\pi}=\boldsymbol{\sigma}]\,\mathbb{P}[r(\boldsymbol{\sigma})=\sigma_0] = 1-\phi^n\sum_{\sigma_0\in\mathcal{N}(\pi_0)} Acc^r(n,\sigma_0) \leq$$

$$\leq 1-\phi^n(m-1)(1-\epsilon) \implies \phi^n(m-1)(1-\epsilon)\leq\epsilon$$

The above yields $n=\Omega\left(\log\left(\frac{m}{\epsilon}\right)\right)$. ■

Our presentation of [10]. will stop here, though we have presented about half of the work of Caragiannis et. al. Their other results are important in the context of social choice theory, which is not where we want to focus.

5.2 Estimating the Spread Parameters

The sample complexity of estimating the spread parameters of the Kendall-Mallows model was essentially settled in [9]. We will now present their results.

5.2.1 From Mallows to Sums of Truncated Geometrics

First, the authors show that the approach of Caragiannis et. al. can be applied to the Mallows Block model, as well as that $\log(m)$ samples are necessary to learn in TV-distance an instance of the simple Mallows model with unknown central ranking. Having done that, the authors tackle the problem of estimating the spread parameter of the Kendall-Mallows model. There, they follow a counter-intuitive approach, by solving a seemingly more complicated problem. Specifically, they observe that the pmf of the Generalized Kendall-Mallows model can be factorized as:

$$\mathop{\mathbb{P}}_{\pi\sim\mathcal{P}_{\boldsymbol{\phi}},\pi_0}[\pi=\sigma] = \prod_{i=1}^{m}\frac{\phi_i^{V_i(\sigma,\pi_0)}}{Z_i(\phi_i)}$$

Computing the marginals of the random variables $Y_i = V_i(\pi,\pi_0)$, we get:

$$\mathop{\mathbb{P}}_{\pi\sim\mathcal{P}_{\boldsymbol{\phi}},\pi_0}[Y_i=k_i] = \frac{\phi_i^{k_i}}{Z_i(\phi_i)}, k_i\in\{0,1,\dots i-1\}$$

Based on that, we conclude that the random variables Y_i are independent. Moreover, the above distribution is in fact a variation of the geometric distribution, where ϕ_i denotes the probability of failure and we consider the support to be the set $\{0\}\cup[i-1]$. This is referred to as the truncated geometric distribution with truncation parameter $i-1$ (denoted $\mathcal{TG}(\phi_i,i-1)$). We write $\mathcal{P}_{\boldsymbol{\phi}} = \bigotimes_{i\in[n]}\mathcal{TG}(\phi_i,i-1)$.

Based on the above, the logical step is to show that learning in the Generalized Kendall-Mallows model reduces to learning the spread parameters independently (which follow a known distribution). Busa-Fekete et. al. show that, if the central ranking is fixed, we essentially have $\mathcal{P}_{\boldsymbol{\phi},\pi_0} \equiv \mathcal{P}_{\boldsymbol{\phi}}$. To prove that, they exploit the fact that it is possible to define a bijective mapping between the elements of the support of $\mathcal{P}_{\boldsymbol{\phi},\pi_0}$ and those of the support of $\mathcal{P}_{\boldsymbol{\phi}}$ based on Proposition 4.2.1.
The other key remark at this point is that both the Generalized Kendall-Mallows model and the truncated geometric distribution are exponential families. Indeed, we have:

$$p_{\boldsymbol{\theta}}(\pi) = \exp\left(\boldsymbol{\theta}^T \boldsymbol{T}(\boldsymbol{x}) - a(\boldsymbol{\theta})\right)$$

$$\boldsymbol{\theta} = (\ln(\phi_1), \ldots, \ln(\phi_m))$$
$$\boldsymbol{T}(\pi) = (V_1(\pi, \pi_0), \ldots, V_m(\pi, \pi_0))$$

$$a(\boldsymbol{\theta}) = \sum_{i=1}^{m} a_i(\theta_i) = \sum_{i=1}^{m} \ln\left(Z_i\left(e^{\theta_i}\right)\right)$$

and:

$$p_{\theta_i}(x) = \exp(\theta_i T(x) - a(\theta)), x \in \{0, 1, \ldots i-1\}$$
$$\theta_i = \ln(\phi_i)$$
$$T(x) = x$$

$$a(\theta_i) = \ln\left(Z_i\left(e^{\theta_i}\right)\right)$$

provided that i is fixed.
The above imply that Proposition 3.6.3 can be used, so the problem of learning a Generalized Kendall-Mallows model in KL-divergence and TV-distance can indeed be reduced to learning the joint distribution of its sufficient statistics.

5.2.2 Parameter Estimation in Truncated Geometrics

Since we showed that the problem of learning Generalized Mallows reduces to learning truncated geometrics, we will now write down the MLE for the distribution $\mathcal{TG}(\phi, i-1)$, provided that i is known. Suppose we have $\boldsymbol{X} \sim \mathcal{TG}(\phi, i-1)^n$. The MLE is:

$$\hat{\phi} = \underset{\phi}{argmax}\{L(\phi|\boldsymbol{X})\} = \underset{\phi}{argmax}\left\{\prod_{i=1}^{n} \frac{\phi^{X_i}}{Z_i(\phi)}\right\} = \underset{\phi}{argmax}\left\{\frac{\phi^{\sum_{i=1}^{n} X_i}}{Z_i(\phi)^n}\right\} =$$
$$= \underset{\phi}{argmax}\left\{\left(\sum_{i=1}^{n} X_i\right)\ln(\phi) - n\ln Z_i(\phi)\right\}$$

Taking the derivative with respect to ϕ, we get:

$$\left(\sum_{i=1}^{n} X_i\right)\frac{1}{\hat{\phi}} - n\frac{Z_i'\left(\hat{\phi}\right)}{Z_i\left(\hat{\phi}\right)} = 0 \iff \hat{\phi}\frac{Z_i'\left(\hat{\phi}\right)}{Z_i\left(\hat{\phi}\right)} = \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right) = \bar{X}_n$$

The form of the above equation is similar to that of Subsection 4.4.2. However, this time, we can actually face the issue, using the fact that truncated geometrics with given truncation parameter form an exponential family. Indeed, remark that the RHS is equal to the mean value of $\mathcal{TG}\left(\hat{\phi}, i-1\right)$. We have:

$$\mathop{\mathbb{E}}_{X\sim\mathcal{TG}\left(\hat{\phi},i-1\right)}\left[T\left(X\right)\right]=\mathop{\mathbb{E}}_{X\sim\mathcal{TG}\left(\hat{\phi},i-1\right)}\left[X\right]\overset{\hat{\theta}=\ln\left(\hat{\phi}\right)}{=}\frac{d}{d\hat{\theta}}\left(a\left(\hat{\theta}\right)\right)=\hat{\phi}\frac{Z_i'\left(\hat{\phi}\right)}{Z_i\left(\hat{\phi}\right)}$$

Consequently, the problem is reduced to finding $\hat{\phi}$ such that:

$$h\left(\hat{\phi}\right)=\mathop{\mathbb{E}}_{X\sim\mathcal{TG}\left(\hat{\phi},i-1\right)}\left[X\right]=\frac{1}{m}\left(\sum_{i=1}^{n}X_i\right)$$

This is greatly facilitated by the fact that h is strictly increasing. Indeed:

$$h'\left(\hat{\phi}\right)=\frac{d^2}{d\hat{\theta}}\left(a\left(\hat{\theta}\right)\right)=\mathop{\mathrm{Var}}_{X\sim\mathcal{TG}\left(\hat{\phi},i-1\right)}\left(X\right)>0$$

As a result, despite being unable to compute the inverse function, we can find an approximate solution that is $\gamma-$close to the desired one in $\mathcal{O}\left(\log\left(\frac{1}{\gamma}\right)\right)$ time using binary search. The exact expressions for the mean and the variance of $\mathcal{TG}\left(\phi,k\right)$ are (the derivation can be found in [9]):

$$\boxed{\mathop{\mathbb{E}}_{X\sim\mathcal{TG}(\phi,k)}\left[X\right]=\frac{\phi}{1-\phi}-\left(k+1\right)\frac{\phi^{k+1}}{1-\phi^{k+1}}}$$

$$\boxed{\mathop{\mathrm{Var}}_{X\sim\mathcal{TG}(\phi,k)}\left[X\right]=\frac{\phi}{\left(1-\phi\right)^2}-\left(k+1\right)^2\frac{\phi^{k+1}}{\left(1-\phi^{k+1}\right)^2}}$$

The above results should be compared with the case of the standard geometric distribution, which we examined in Example 3.2.2.2. In both examples, the MLE is computed by calculating the sample mean and inverting the function that gives the mean as a function of the parameter. The only difference is that for the truncated geometric it is not possible to compute the inverse function so we have to estimate it. Moreover, if we take $k\to\infty$, the second term vanishes in both the above expressions and we get the ones for the simple geometric distribution.

5.2.3 The Block Model

It appears that everything is in place. We know that each sample $\pi\sim\mathcal{P}_{\boldsymbol{\phi},\pi_0}$ "breaks" into m independent values $Y_i\sim\mathcal{TG}\left(\phi_i,i-1\right)$, each of which can be used to estimate the corresponding spread parameter. However, Busa-Fekete et. al. make one last important remark. Specifically, they consider the case where we know from the start that some of the alternatives share the same spread parameter value. This motivates the introduction of the Mallows Block model, which is also an exponential family (provided that the block structure is fixed). This makes it possible to aggregate the values Y_i that correspond to alternatives belonging in the same block in order to produce better results with fewer samples. A special case of this is the simple Kendall-Mallows model, where all alternatives are in the same block, thus making it possible to estimate ϕ even from a single sample as $m\to\infty$. This manages to incorporate one of the most important attributes of the estimator given by Mukherjee in [41].

5.2.4 Parameter Estimation in the Kendall-Mallows Block Model

We now proceed with the results of the paper related to parameter estimation. They are stated informally, followed by proof sketches. Note that, for the following, $m^\star = \min_{i\in[d]} m_i$.
The learning algorithm whose correctness and optimality we wish to prove is the following:

Algorithm 3: Spread Parameter Estimation

Data: $\pi_1, \dots, \pi_n \sim \mathcal{P}_{\boldsymbol{\phi},\pi_0,\boldsymbol{B}}, \pi_0$
Result: $\hat{\boldsymbol{\phi}}$
for *each* $i \in [d]$ **do**
  compute $r = \frac{1}{n}\sum_{k=1}^{n} T_i(\pi_k, \pi_0, \boldsymbol{B})$;
  use binary search to find $\hat{\theta}_i$ such that $a_i'\left(\hat{\theta}_i\right) \approx r$;
  $\hat{\phi}_i = e^{\hat{\theta}_i}$;
end
return $\hat{\boldsymbol{\phi}}$;

We first state the main result involving parameter estimation.

Theorem 5.2.1 (Busa-Fekete et. al (2019) (Informal))**.** *Given* $n = \tilde{\Omega}\left(\frac{d}{m^\star\epsilon^2} + \log(m)\right)$ *samples from a Kendall-Mallows d-block distribution* $\mathcal{P}_{\boldsymbol{\phi}^\star,\pi_0,\boldsymbol{B}}$*, we can find estimates* $\hat{\pi}$ *and* $\hat{\boldsymbol{\phi}}$ *such that:*

$$\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\boldsymbol{\phi}^\star,\pi_0,\boldsymbol{B}}}\left[(\hat{\pi}=\pi_0) \wedge \left(\left|\left|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^\star\right|\right|_2 \le \epsilon\right)\right] \ge 0.99$$

If π_0 *is known, then with* $n = \tilde{\Omega}\left(\frac{d}{m^\star\epsilon^2}\right)$ *we have:*

$$\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\boldsymbol{\phi}^\star,\pi_0,\boldsymbol{B}}}\left[\left|\left|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^\star\right|\right|_2 \le \epsilon\right] \ge 0.99$$

Proof Sketch. If the central ranking is not known, it can be learned with high probability using $\log(m)$ samples (hence the corresponding term in the first expression). For the rest, we consider that $\hat{\pi} = \pi_0$.
We know that, given $\pi \sim \mathcal{P}_{\boldsymbol{\phi}^\star,\pi_0,\boldsymbol{B}}$, the random variables $V_i(\pi,\pi_0)$ follow a truncated geometric distribution and are independent. Consequently, the sufficient statistics $T_i(\pi,\pi_0,\boldsymbol{B}) = \sum_{j\in B_i} V_j(\pi,\pi_0)$ are also independent, as sums of independent random variables with no common terms. Consequently, each spread parameter can be estimated separately.
Since the central ranking is known, the target distribution belongs to an exponential family, so their properties can be exploited. For that reason, instead of the spread parameters ϕ_i, we will estimate their logarithms θ_i (which are the natural parameters of the exponential family). By the properties of exponential families, we have:

$$h_i(\theta_i) = \mathop{\mathbb{E}}_{\pi\sim\mathcal{P}_{\boldsymbol{\phi}^\star,\pi_\prime,\boldsymbol{B}}}[T_i(\pi,\pi_0,\boldsymbol{B})] = \sum_{j\in B_i} \mathop{\mathbb{E}}_{\pi\sim\mathcal{P}_{\boldsymbol{\phi}^\star,\pi_\prime,\boldsymbol{B}}}[V_j(\pi_0)]$$

All the summands are strictly increasing functions (based on what we showed previously about the truncated geometric distribution), so h_i is strictly increasing. This verifies that the algorithm presented previously is correct. It remains to determine the sample complexity. To do

that, the concentration bound of Proposition 3.6.2. From that point on, the proof's difficulty is to lower bound the KL-divergence of 2 distributions in the family, which is equivalent to lower bounding the variance of a distribution whose parameter is in the interval defined by those of the previous 2. ■

Suppose now that the central ranking is known. Setting $n = 1$ and solving for ϵ gives us a view of the error rate with respect to the minimum block size. Specifically, we have:

Corollary 5.2.1 (Busa-Fekete et. al (2019) (Informal))**.** *Given a **single sample** from Kendall-Mallows d-block distribution $\mathcal{P}_{\boldsymbol{\phi}^\star,\pi_0,\boldsymbol{B}}$ with known central ranking π_0 and unknown spread parameters $\boldsymbol{\phi}^\star$, we can estimate $\hat{\boldsymbol{\phi}}$ so that:*

$$\mathop{\mathbb{P}}_{\pi\sim\mathcal{P}_{\boldsymbol{\phi}^\star,\pi_0,\boldsymbol{B}}}\left[\left|\left|\hat{\boldsymbol{\phi}}-\boldsymbol{\phi}^\star\right|\right|_2 \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{m^\star}}\right)\right] \geq 0.99$$

Observe that, as $m^\star \to \infty$, the error goes to 0. This is due to the aggregation effect we described previously. In the case of the simple Kendall-Mallows model, we have $m^\star = m$, so a single sample ends up behaving like m independent samples, all of which are used to estimate the same value. This verifies that estimation from a single sample is possible, just as with the estimator given in [41].

5.2.5 Distribution Learning in the Kendall-Mallows Block Model

We now present the results related to distribution learning. The learning algorithm we use is the same as before. However, the sample complexity that we get is higher than the one for parameter estimation. That is because computing a distribution that is close to another one with respect to some statistical distance is generally a more strict demand than having a good estimate for some parameter. The sample complexity that is achieved by the algorithm is:

Theorem 5.2.2 (Busa-Fekete et. al (2019) (Informal))**.** *Given $n = \tilde{\Omega}\left(\frac{d}{\epsilon^2} + \log(m)\right)$ samples from a Mallows d-block distribution $\mathcal{P}_{\boldsymbol{\phi}^\star,\pi_0,\boldsymbol{B}}$, we can learn a distribution $\hat{\mathcal{P}}$ such that:*

$$\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\boldsymbol{\phi}^\star,\pi_0,\boldsymbol{B}}}\left[D_{KL}\left(\mathcal{P}_{\boldsymbol{\phi}^\star,\pi_0,\boldsymbol{B}}||\hat{\mathcal{P}}\right) \leq \epsilon^2\right] \geq 0.99 \implies \mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\boldsymbol{\phi}^\star,\pi_0,\boldsymbol{B}}}\left[d_{TV}\left(\mathcal{P}_{\boldsymbol{\phi}^\star,\pi_0,\boldsymbol{B}},\hat{\mathcal{P}}\right) \leq \epsilon\right] \geq 0.99$$

Proof Sketch. The proof is similar to the previous one, though it simpler this time. Since we are interested in learning the distribution in KL-divergence and we know that the Mallows block model is an exponential family, we exploit Proposition 3.6.3, along with the tensorization identity given about the KL-divergence in chapter 3. After that point, the reasoning mostly the same as in the previous proof. ■

The result comes with a matching lower bound:

Theorem 5.2.3 (Busa-Fekete et. al (2019) (Informal)). *For any distribution $\hat{\mathcal{P}}$ that is based only on $o\left(\frac{d}{\epsilon^2} + \log(m)\right)$ samples from a Mallows d-block distribution there exists some $\mathcal{P}_{\boldsymbol{\phi}^\star,\pi_0,\boldsymbol{B}}$ such that $d_{TV}\left(\mathcal{P}_{\boldsymbol{\phi}^\star,\pi_0,\boldsymbol{B}}, \hat{\mathcal{P}}\right) \geq \epsilon$.*

Proof Sketch. The result is a consequence of Fano's inequality. However, for that to be applied, it is necessary to define a family of distributions such that for any pair of distinct members of the family, both their KL-divergence is upper bounded and their TV-distance is lower bounded. For that reason, the central ranking and the block structure are fixed and the case where the alternatives are partitioned into d blocks with equal number of elements is considered. The spread parameters take values in the set $\left\{\frac{1}{2}, \frac{1}{2} - c\frac{\epsilon}{\sqrt{m}}\right\}$ where ϵ is chosen to be small enough for the parameter value to be $\geq \frac{1}{4}$. That way, 2^d distributions can be defined. However, we are interested only in those that are not too "similar". Since the distributions are determined simply based on the values of the spread parameters, we can represent each distribution in the family as a binary string of length d based on the mapping $\frac{1}{2} - c\frac{\epsilon}{\sqrt{m}} \to 0$ and $\frac{1}{2} \to 1$. Applying the Gilbert-Varshamov bound (Lemma 3.5.1), we get that there are at least $2^{\frac{d}{8}}$ distributions such that, given any pair them, they differ on at least $\frac{d}{8}$ spread parameters. This is the family of the distributions that are used to lower bound the minimax risk.
Based on the above, the KL-divergence of any pair of distributions in the above family is shown to be upper bounded by $32c^2\epsilon^2$. What is less trivial is to lower bound the TV-distance of any pair of distinct distributions in the family. Because Proposition 3.4.4 does not yield a tight lower bound, the authors exploit Corollary 3.4.1. Specifically, they define random variables that are the sums of the sufficient statistics of a pair of distributions in the previously defined family. However, they only include a subset of those sufficient statistics that correspond to spread parameters that do not have the same value in the 2 distributions. The resulting random variables belong in the same exponential family but have different natural parameters, so Proposition 3.6.1 can be applied. The rest of the proof relies on the fact that the TV-distance is lower bounded by a function of the sum of the expected absolute deviation of each random variable from its mean, while each of those terms is lower bounded by $\frac{1}{\sqrt{2}}$.
■

Chapter 6

Learning in the Cayley-Mallows Model

In this chapter, we focus on the Mallows model equipped with the Cayley distance (henceforth referred to as the *Cayley-Mallows model*). While the KT distance focuses on the pair-wise comparisons between elements, thus being suitable for problems related to the modelling of ranking data, the Cayley distance focuses on the cyclic structure of permutations. Moreover, this version of the model exhibits weaker concentration around the central ranking, due to the fact that the Cayley distance is not bound by whether elements that have been swapped are adjacent or not. Consequently, it has generally been harder to come up with applications of this version, though one example can be found in [27], where the authors present an application within the context of computational biology where the Cayley-Mallows model fits better than the Kendall-Mallows model. Despite its influence on our approach, that work is mostly experimental in nature, whereas ours is more theoretically oriented. Specifically, we intend to the consider the problems that were examined in the previous chapter this time for the Cayley-Mallows model, showing how the techniques presented there can be adjusted to this setting. Our work is motivated by the difference noted in the results of [14] where the problem of learning mixtures under the Cayley-Mallows models was examined and those of [34] where the problem of learning mixtures under the Kendall-Mallows model was examined.

6.1 Recovering the Central Ranking

To tackle the problem, we will adjust the approach presented in [10]. Specifically, Caragiannis et. al. noted that, the model with the KT distance favors rankings where the inverted pairs are few and they used this property to efficiently approximate the maximum likelihood solution. In similar fashion, we will show that, when equipped with the Cayley distance, the model favours events where most alternatives are in their original positions. We will exploit that property to show that it is possible to recover the central ranking with high probability by determining which is the position where an element appears most frequently and then we will determine the sample complexity of the above algorithm.

6.1.1 Concentration in Cayley-Mallows

As we noted back in chapter 4, the Cayley distance between 2 permutations π, π_0 is $d_{ca}(\pi, \pi_0) = m - k$, where k is the number of cycles found in π when considering the original positions of the objects in π_0. Therefore, when the Mallows model is equipped with this distance, the events that are favored are those where there are many cycles (ideally, when $k = m$, there is one cycle corresponding to each alternative which means that all cycles are of length 1 and no element is misplaced). This motivates us to compute the probability of an object's position being the same in a sample as in the original permutation.

Lemma 6.1.1. *In the simple Cayley-Mallows model, the probability that an alternative $i \in [m]$ is ranked correctly in sample π is equal to $p_{i\pi_0(i)} = \frac{1}{1+(m-1)\phi}, \forall i \in [m]$.*

Proof. We have:

$$p_{i\pi_0(i)} = \sum_{\substack{\sigma \in S_m \\ \sigma(i) = \pi_0(i)}} \mathop{\mathbb{P}}_{\pi \sim \mathcal{P}_{\phi,\pi_0}} [\pi = \sigma] = 1 - \sum_{\substack{\sigma' \in S_m \\ \sigma'(i) \neq \pi_0(i)}} \mathop{\mathbb{P}}_{\pi \sim \mathcal{P}_{\phi,\pi_0}} [\pi = \sigma'] \quad (1)$$

Note that, when given a permutation where i is ranked correctly, there are $m-1$ permutations where the i has been transposed with another element and the rest are the same. By Lemma 4.2.2, we know that, for any such permutation σ' we have $d_{ca}(\sigma', \pi_0) - d_{ca}(\sigma, \pi_0) = 1$. Based on the above remarks, (1) can be written in the form:

$$p_{i\pi_0(i)} = 1 - (m-1)\phi \left(\sum_{\substack{\sigma \in S_m \\ \sigma(i) = \pi_0(i)}} \mathop{\mathbb{P}}_{\pi \sim \mathcal{P}_{\phi,\pi_0}} [\pi = \sigma] \right) = 1 - (m-1)\phi p_{i\pi_0(i)} \iff$$

$$\iff \boxed{p_{i\pi_0(i)} = \frac{1}{1 + (m-1)\phi}}$$

■

A direct consequence of the above is that the probability of i being misplaced is equal to $\frac{(m-1)\phi}{1+(m-1)\phi}$. However, we can prove something even stronger:

Corollary 6.1.1. *Given some alternative $i \in [m]$ and $j \in [m]$ such that $\pi_0(i) \neq j$, we have $p_{ij} = \frac{\phi}{1+(m-1)\phi}$.*

Proof. We have:

$$p_{ij} = \sum_{\substack{\sigma' \in S_m \\ \sigma'(i) = j}} \mathop{\mathbb{P}}_{\pi \sim \mathcal{P}_{\phi,\pi_0}} [\pi = \sigma']$$

For each permutation σ' in the above sum, there exists exactly one permutation σ where the element i is swapped with the one occupying its correct position and the rest are as in σ'. This, combined with Lemma 4.2.2, gives:

$$\boxed{p_{ij} = \phi \left(\sum_{\substack{\sigma \in S_m \\ \sigma(i) = \pi_0(i)}} \mathop{\mathbb{P}}_{\pi \sim \mathcal{P}_{\phi,\pi_0}} [\pi = \sigma] \right) = \phi p_{i\pi_0(i)} = \frac{\phi}{1 + (m-1)\phi}}$$

■

Another interesting aspect of the model is the way the above probabilities change in case we are given that some alternatives are ranked correctly by a sample. We give the following result, which extends the previous lemmas.

Lemma 6.1.2. *Let $i_1, \ldots, i_{k+1} \in [m]$ be distinct numbers and $j \neq \pi_0(i_{k+1})$. We have:*

$$\mathop{\mathbb{P}}_{\pi \sim \mathcal{P}_{\phi,\pi_0}} [\pi(i_{k+1}) = \pi_0(i_{k+1}) \,|\pi(i_l) = \pi_0(i_l), \forall l \in [k]] = \frac{1}{1 + (m-1-k)\phi}$$

$$\mathop{\mathbb{P}}_{\pi \sim \mathcal{P}_{\phi,\pi_0}} [\pi(i_{k+1}) = j | \pi(i_l) = \pi_0(i_l), \forall l \in [k]] = \frac{\phi}{1 + (m-1-k)\phi}$$

Proof. The proof is essentially the same as the ones we gave for the previous 2 lemmas. However, given some ranking where i_{k+1} is ranked correctly, there are $m-1-k$ rankings where the i_{k+1} has been transposed with another element and the rest are the same (given that $i_1, i_2, \ldots, i_k$ have to be ranked correctly). ■

At this point, we can get a more concrete sense of the way the Cayley-Mallows model functions. Specifically, the model is a variation of the well-known *matching problem* (first defined in [40]). There, a number of letters are addressed to distinct individuals and for each letter there exists a corresponding envelope, but the letters are put into envelopes in a random fashion. In our setting, the letters are the alternatives and the envelopes are their correct positions. However, unlike the classical version of the problem, where all choices are performed uniformly at random, in this version, the parameter ϕ ensures that there is a bias towards the correct permutation (with the exception of when $\phi = 1$, which corresponds to a degenerate case). The previous similarity is mentioned in [17], where the number of fixed points of permutations generated by the model is calculated, which can also be computed using the above result. A similar result about fixed points for the Kendall-Mallows model can be found in [42]. However, that result holds only in the asymptotic regime.
Overall, the above essentially give us a way to determine the central ranking. Specifically, note that $p_{i\pi_0(i)} > p_{ij}, \forall j \neq \pi_0(i)$ (provided that $\phi < 1$), meaning that, given some element and a number of samples generated by the model, the position where it appears most frequently is the one more likely to be its correct position, although the concentration exhibited by the model is rather weak, since $p_{i\pi_0(i)} \to 0$ as $m \to \infty$. Taking all the above into account, in the next section, we will present a family of algorithms that learn the central ranking with high probability. This family, however, is expected to result in a higher sample complexity than the one given in [10] for the Kendall-Mallows model. That is because, in that model, the

probability that a pair-wise comparison is preserved is always greater than $\frac{1}{2}$, regardless of the value of the spread parameter, which is not true about our model.

6.1.2 Approximating the MLE

As explained previously, our approach will rely on the fact that each element's most likely position is the one where it most frequently appears in the available samples. This however describes a family of algorithms instead of a single one. That is because there may be 2 kinds of ties. First, there may be ties involving the appearances of a specific element. Second, there may be a pair of elements whose most frequent appearances correspond to the same position. Specifying tie-breaking rules for the above cases results in the definition of a single algorithm. The above should be compared with the PM-c rules defined in [10]. For the following, we will refer to the previous as position majority-consistent rules.
The above approach is computationally efficient. Indeed, it is reasonable to assume that the time required to break ties will be less than the time required to compute the frequency of each alternative's appearance in various positions. This last task can be performed as the input is read in $\Theta(nm)$ time. Consequently, the running time of the algorithm is linear with respect to the size of the input it receives. We give a high level description of the algorithm:

Algorithm 4: MLE approximation via position majority-consistent rules

Data: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n) \sim \mathcal{P}^n_{\phi,\pi_0}$
Result: $\hat{\pi}$
for *each* $i \in [m]$ **do**
  construct a histogram with the frequencies of its appearances;
  let j be the mode (break ties arbitrarily);
  $\hat{\pi}(i) := j$ (break ties arbitrarily);
end
return $\hat{\pi}$;

6.1.3 Sample Complexity Analysis

Now, we are ready determine the sample complexity of the previously described family of algorithms, provided that ϕ is upper bounded by some constant < 1.

Theorem 6.1.3. *Given any instance of the simple Cayley-Mallows model with central ranking π_0 and spread parameter $\phi \in [0, 1-\gamma], \gamma \in (0,1]$ and any $\epsilon \in (0,1]$, any position majority-consistent algorithm recovers the central ranking with probability at least $1-\epsilon$ using $n = \Theta\left(\left(\frac{m}{\gamma}\right)^2 \log\left(\frac{m}{\epsilon}\right)\right)$ samples with time complexity $\Theta\left(\frac{m^3}{\gamma^2}\log\left(\frac{m}{\epsilon}\right)\right)$.*

Proof. Let $\hat{\pi}$ be the ranking recovered by an algorithm in the aforementioned family using n samples, denoted $\{\pi_k\}_{k\in[n]} \sim \mathcal{P}^n_{\phi,\pi_0}$. The probability of error is equal to:

$$\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}\left[\exists i \in [m] : \hat{\pi}(i) \neq \pi_0(i)\right] \leq \sum_{i=1}^{m} \mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}\left[\hat{\pi}(i) \neq \pi_0(i)\right] \quad (1)$$

where the upper results from a direct application of the union bound. Let n_{ij} denote the number of samples where alternative i is placed in position j. Taking that into account, the terms of the sum on the RHS of (1) can be written in the form:

$$\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}\left[\exists j\in[m]\setminus\{\pi_0(i)\}:n_{ij}\geq n_{i\pi_0(i)}\right]\leq\sum_{j\neq\pi_0(i)}\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}\left[n_{ij}\geq n_{i\pi_0(i)}\right]\quad(2)$$

To proceed from here, we observe that $n_{ij}=\sum_{k=1}^{n}\mathbb{1}\{\pi_k(i)=j\}$ where $\mathbb{1}\{\pi_k(i)=j\}$ are iid Bernoulli random variables with probability of success $\frac{\phi}{1+(m-1)\phi}$ if $j\neq\pi_0(i)$ and $\frac{1}{1+(m-1)\phi}$ if $j=\pi_0(i)$. Therefore, we have:

$$n_{i\pi_0(i)}-n_{ij}=\sum_{k=1}^{n}\left(\mathbb{1}\{\pi_k(i)=\pi_0(i)\}-\mathbb{1}\{\pi_k(i)=j\}\right),j\neq\pi_0(i)\quad(3)$$

Let X be the above sum and let X_k be the summands. The random variables X_k are iid and have support $\{-1,0,1\}$. By linearity of expectation, we have:

$$\mathop{\mathbb{E}}_{\pi_k\sim\mathcal{P}_{\phi,\pi_0}}[X_k]=\mathop{\mathbb{E}}_{\pi_k\sim\mathcal{P}_{\phi,\pi_0}}[\mathbb{1}\{\pi_k(i)=\pi_0(i)\}]-\mathop{\mathbb{E}}_{\pi_k\sim\mathcal{P}_{\phi,\pi_0}}[\mathbb{1}\{\pi_k(i)=j\}]=\frac{1-\phi}{1+(m-1)\phi}>0$$

and $\mathop{\mathbb{E}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}[X]=\frac{n(1-\phi)}{1+(m-1)\phi}$. Consequently, the terms of the sum in the RHS of (2) can be written in the form:

$$\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}\left[X-\mathop{\mathbb{E}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}[X]\leq-\mathop{\mathbb{E}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}[X]\right]\leq\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}\left[\left|X-\mathop{\mathbb{E}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}[X]\right|\geq\mathop{\mathbb{E}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}[X]\right]$$

Applying Hoeffding's inequality to the above, we get:

$$\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}[X\leq0]\leq2\exp\left\{-\frac{n(1-\phi)^2}{2[1+(m-1)\phi]^2}\right\}$$

Combining all the above, we get that the probability of failure is upper bounded by:

$$\mathop{\mathbb{P}}_{\boldsymbol{\pi}\sim\mathcal{P}^n_{\phi,\pi_0}}[\exists i\in[m]:\hat{\pi}(i)\neq\pi_0(i)]\leq2m(m-1)\exp\left\{-\frac{n(1-\phi)^2}{2[1+(m-1)\phi]^2}\right\}$$

We demand the above to be less than ϵ and get:

$$n\geq\frac{2[1+(m-1)\phi]^2}{(1-\phi)^2}\ln\left(\frac{2m(m-1)}{\epsilon}\right)$$

Note that $2[1+(m-1)\phi]^2=\Theta(m^2)$ and that $\frac{1}{\gamma^2}\geq\frac{1}{(1-\phi)^2}$. Consequently, we get that $n=\Theta\left(\left(\frac{m}{\gamma}\right)^2\log\left(\frac{m}{\epsilon}\right)\right)$ samples suffice to recover the central ranking with probability at least $1-\epsilon$ and this yields the desired time complexity.

■

We remark that, the result still holds even if the spread parameters of the distributions where the samples are drawn from were not the same (just as in the Kendall-Mallows model).
So far, we have upper bounded the sample complexity of recovering the central ranking in the Cayley-Mallows model. We will now apply the technique used in chapter 5 to get a lower bound. As we will see, the lower bound that we will get does not match the sample complexity of the family of algorithms described above.

Theorem 6.1.4. *For any $\epsilon \in \left(0, \frac{1}{2}\right]$, any (randomized) algorithm requires $\Omega\left(\log\left(\frac{m}{\epsilon}\right)\right)$ samples from a simple Cayley-Mallows model to determine the true ranking with probability at least $1-\epsilon$.*

Proof. The proof is similar to that we gave for the Kendall-Mallows model in chapter 5 for Theorem 5.1.3. The only thing that changes is the cardinality of the set $\mathcal{N}(\pi_0)$. Specifically, we have $|\mathcal{N}(\pi_0)| = \binom{m}{2}$. Repeating the steps of that proof results in $\phi^n \frac{m(m-1)}{2}(1-\epsilon) \leq \epsilon$ which yields $n = \Omega\left(\log\left(\frac{m}{\epsilon}\right)\right)$. ■

Obviously, this lower bound does not match the upper bound we gave. Applying Fano's method would yield a similar result. This could either due to the fact that algorithms that belong to the family we presented previously are sub-optimal or because the previous technique does not yield tight results for this version of the Mallows model. There is also the case that the sample complexity analysis we gave above is not tight, though this seems rather unlikely. For a further discussion of the issue, see chapter 7.
Another remark that we should make involves the applicability of the previous for the Mallows model for other distance metrics. In particular, in the case of Spearman's footrule, by applying the same technique as in 6.1.1 combined with the triangle inequality, we would get that $p_{i\pi_0(i)} \geq \frac{1}{1+(m-1)\phi}$, so the previous family of algorithms works for this version of the model as well. The same cannot be said about Spearman's rank correlation, due to the fact that it does not satisfy the triangle inequality.

Chapter 7

Conclusions and Future Work

In this thesis, we gave the first algorithm that recovers the central ranking in the Cayley-Mallows model in a provably computationally efficient fashion. The issue of whether the algorithm we gave is optimal is an open question, though one we are actively working on. Moreover, it would be interesting to examine whether the approach of [9] can be applied in the case of the Cayley-Mallows model. We think that this is highly likely, due to the fact that it is possible to show that the sufficient statistics of the Cayley-Mallows model are independent Bernoulli random variables, based on the definition of the model given in chapter 4 and Proposition 4.2.2. This is another issue that we are examining. Furthermore, it would be interesting to examine which other versions of the Mallows model admit similar approaches.
Another possible direction for future work is one motivated by the questions posed in [9]. Specifically, the authors asked what is the minimum number of samples required to recover the block structure of a Kendall-Mallows block distribution. This is a question that has drawn our attention in the past, though none of our work concerning this issue is included in this thesis. Another question the authors posed in the same paper is whether it is possible to estimate the spread parameters from a single sample without knowing the block structure. Though we do not have a definitive answer, we believe this to be unlikely, due to the fact that it seems to be impossible to aggregate samples of the sufficient statistics without knowing that they correspond to alternatives belonging in the same block. However, this is merely an intuitive argument and this line of research should not be dismissed only because of it. Moreover, it would be interesting to examine those questions for the Cayley-Mallows model as well.
Finally, it would be interesting to examine the above questions in the context of other ranking models, such the Plackett-Luce model (which we introduced back in chapter 4).

Bibliography

[1] Ailon N., Charikar M., Newman A. Aggregating inconsistent information: Ranking and clustering. *In Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing, pages 684–693, 2005.*

[2] Ali S.M., Silvey S.D. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society. Series B (Methodological) 28, 1 (1966), 131–142.*

[3] Arrow K. Social Choice and Individual Values. *John Wiley and Sons.*

[4] Awasthi P., Blum A., Sheffet O., Vijayaraghavan A. Learning mixtures of ranking models. *In Advances in Neural Information Processing Systems, pages 2609–2617, 2014.*

[5] Bartholdi J., Tovey C.A., Trick M.A. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare, 6(2):157–165, 1989.*

[6] Ben-David S., Shalev-Schwartz S. Understanding Machine Learning: From Theory to Algorithms. *Cambridge University Press; 1 edition (May 19, 2014).*

[7] Boucheron S., Lugosi G., Massart P. Concentration inequalities: A nonasymptotic theory of independence. *Oxford University Press; 1 edition (April 1, 2016).*

[8] Braverman M., Mossel E. Sorting from noisy information. *CoRR, abs/0910.1191, 2009.*

[9] Busa-Fekete R., Fotakis D., Szörényi B., Zampetakis M. Optimal Learning of Mallows Block Model. *COLT 2019: 529-532.*

[10] Caragiannis I., Procaccia A.D., Shah N. When do noisy votes reveal the truth? *ACM Transactions on Economics and Computation (TEAC), 4(3):15, 2016.*

[11] Cover T.M., Thomas J.A. Elements of Information Theory. *Wiley-Interscience; 2 edition (July 18, 2006).*

[12] Csiszár I. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica 2 (1967), 299–318.*

[13] Daskalakis C., Diakonikolas I., Servedio R. Learning Poisson Binomial Distributions. *CoRR abs/1107.2702 (2015).*

[14] De A., O'Donnell R., Servedio R.A. Learning sparse mixtures of rankings from noisy information. *CoRR abs/1811.01216 (2018).*

[15] Diaconis P. Group Representations in Probability and Statistics. *Institute of Mathematical Statistics, Lecture Notes-Monograph Series.*

[16] Diaconis P., Graham R.L. Spearman's Footrule as a Measure of Disarray. *Journal of the Royal Statistical Society. Series B (Methodological) Vol. 39, No. 2 (1977), pp. 262-268.*

[17] Diaconis P., Holmes S. A Bayesian Peek into Feller Volume 1. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002) Vol. 64, No. 3, In Memory of D. Basu, Part 2 (Oct., 2002), pp. 820-841.*

[18] Diakonikolas I., Kamath G., Kane D., Li J., Moitra A., Stewart A. Robust Estimators in High Dimensions without the Computational Intractability. *in Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on. Institute of Electrical and Electronics Engineers (IEEE), pp. 655-664.*

[19] Doignon J.P., Pekeč A., Regenwetter M. The repeated insertion model for rankings: Missing link between two subset choice models. *in Psychometrika, 69(1):33–54, 2004.*

[20] Duchi J.C. Lecture Notes for Statistics 311/Electrical Engineering 377. *in Stanford Lecture Notes (2019).*

[21] Dubhashi D.P., Panconesi A. Concentration of Measure for the Analysis of Randomised Algorithms. *Cambridge University Press; 1 edition (March 12, 2012).*

[22] Flinger M.A., Verducci J.S. Distance Based Ranking Models. *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 48, No. 3 (1986), pp. 359-369.*

[23] Fano R. Transmission of information: a statistical theory of communications. *Cambridge, Mass: MIT Press (1968).*

[24] Fraleigh J.B. A First Course in Abstract Algebra. *Pearson; 7 edition (November 16, 2002).*

[25] Gilbert E. A comparison of signalling alphabets. *Bell System Technical Journal, 31: 504–522 (1952).*

[26] Hoeffding W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association. 58 (301): 13–30. (1963).*

[27] Irurozki E., Calvo B., Lozano J.A. Sampling and Learning Mallows and Generalized Mallows Models under the Cayley distance. Technical Report, University of the Basque Country, Department of Computer Science and Artificial Intelligence, January, 2014.

[28] Jordan M.I. The exponential family: Basics. Berkeley Lecture Notes (2010).

[29] Kearns M., Mansour Y., Ron D., Rubinfeld R., Schapire R., Sellie L. On the Learnability of Discrete Distributions. ACM Symposium on Theory of Computing, 1994.

[30] Kenyon-Mathieu C., Schudy W. How to rank with few errors. In Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing, pages 95–103, 2007.

[31] Knuth D.E. The Art of Computer Programming. v3, Addison-Wesley, 1973.

[32] Kullback S., Leibler R.A On information and sufficiency. Annals of Mathematical Statistics. 22 (1): 79–86 (1951).

[33] Levin D.A., Peres Y., Wilmer E.L. Markov Chains and Mixing Times. *American Mathematical Society; 1 edition (December 9, 2008).*

[34] Liu A., Moitra A. Efficiently learning mixtures of Mallows models. *In FOCS, pages 627–638. IEEE Computer Society, 2018.*

[35] Luce R.D. Individual Choice Behavior: A Theoretical Analysis. *New York: Wiley, 1975.*

[36] Mallows C.L. Non-null ranking models. *Biometrika, 44(1):114–130, 1957.*

[37] Marden J.I. Analyzing and Modeling Rank Data. *Chapman & Hall, 1995.*

[38] Meila M., Phadnis K., Patterson A., Bilmes J.A. Consensus ranking under the exponential model *arXiv preprint arXiv:1206.5265, 2012.*

[39] Mitzenmacher M., Upfal E. Probability and Computing: Randomized Algorithms and Probabilistic Analysis. *Cambridge University Press (January 31, 2005).*

[40] de Montmort P.R. Essay d'Analyse sur les Jeux de Hazard. *1st edn. (1708), 2nd edn. (1713). Jacques Quillau, Paris. Reprinted 2005 by AMS/Chelsea, New York.*

[41] Mukherjee S. Estimation in exponential families on permutations. *The Annals of Statistics, 44 (2):853–875, 2016.*

[42] Mukherjee S. Fixed points and cycle structure of random permutations. *Electronic Journal of Probability 21 (2016).*

[43] Plackett R.L. The Analysis of Permutations. *Appl. Statist 24 (2): 193–202. 1959.*

[44] Popov V.Y. Multiple genome rearrangement by swaps and by element duplications. *Theoretical computer science, 385(1-3):115–126, 2007..*

[45] Rigollet P., Hutter J.C. High Dimensional Statistics. *MIT Math 18.657 lecture notes.*

[46] Shannon C.E. A Mathematical Theory of Communication. *Bell System Technical Journal. 27 (3): 379–423 (July 1948).).*

[47] Valiant L.G. A theory of the learnable. *Communications of the ACM, 27, 1984.*

[48] Varshamov R.R. Estimate of the number of signals in error correcting codes. *Dokl. Akad. Nauk SSSR, 117: 739–741 (1957).*

[49] Vlatakis-Gkaragkounis E.V. Learning Algorithms on Social Choice Models. *Unpublished Manuscript in Greek.*

[50] Yu B. Assouad, Fano, and Le Cam. *In Festschrift for Lucien Le Cam, pages 423–435. Springer-Verlag, 1997.*

[51] Principal Editor: Zalta E.N. The Stanford Encyclopedia of Philosophy. *The Metaphysics Research Lab Center for the Study of Language and Information Stanford University Stanford, CA 94305-4115.*