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

Μάθηση διατάξεων από δείγματα με θόρυβο

Διπλωματική Εργασία

Καλαβασης Αλβερτος

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

Εργαστηριο Λογικης και Επιστημης Υπολογιστων

Αθήνα, Σεπτέμβριος 2019

Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Εργαστηριο Λογικης και Επιστημης Υπολογιστων (Co.Re.Lab.)

On Learning Rankings from Noisy Samples

Διπλωματική Εργασία

του

ΚΑΛΑΒΑΣΗ ΑΛΒΕΡΤΟΥ

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

Εγκρίθηκε από την τριμελή επιτροπή την 19η Σεπτεμβρίου 2019.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Δημήτριος Φωτάκης	Αριστείδης Παγουρτζής	Μιχαήλ Λουλάκης
Αν. Καθηγητής	Αν. Καθηγητής	Αν. Καθηγητής
Ε.Μ.Π.	Ε.Μ.Π.	Ε.Μ.Π.

(Υπογραφή) .....
Καλαβασης Αλβερτος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

First release, September 2019

Περίληψη

Σε αυτή την διπλωματική εργασία, μελετάμε το πρόβλημα εκμάθησης διατάξεων από δείγματα με θόρυβο. Αυτό το πεδίο στατιστικής μάθησης είναι εξαιρετικά χρήσιμο στους τομείς της εκμάθησης προτιμήσεων και της ανάκτησης πληροφοριών. Σε αυτό το πλαίσιο εργασίας υποθέτουμε ότι κάποιος λαμβάνει ανεξάρτητα δείγματα, τα οποία μοντελοποιούνται ως μεταθέσεις n αντικειμένων, που παράγονται από μια κατανομή, που αντιστοιχεί σε ένα θορυβώδες πιθανοτικό μοντέλο. Τέτοια γνωστά πιθανοτικά μοντέλα είναι το μοντέλο Mallows και το μοντέλο Plackett-Luce. Έτσι, θέτουμε ερωτήματα σχετικά με το πόσα δείγματα είναι απαραίτητα προκειμένου να μάθουμε τις παραμέτρους των κατανομών αυτών, το κατά πόσο είναι δυνατό να μάθουμε την ίδια την κατανομή μοντελοποιώντας το σφάλμα με διάφορες f-αποκλίσεις, όπως η TV απόσταση και η KL απόκλιση, και, τέλος, ασχολούμαστε με την έννοια του εκτιμητή μέγιστης πιθανοφάνειας. Αρχικά, παρουσιάζουμε αποτελέσματα από την εκτεταμένη ερευνητική βιβλιογραφία πάνω στο μοντέλο Mallows συνδυάζοντας μερικά κλασικά αποτελέσματα της έρευνας όπως και ορισμένα πολύ πρόσφατα. Στη συνέχεια, παρουσιάζουμε τη δική μας πρωτότυπη εργασία, όπου επιλέξαμε να μειώσουμε τις πληροφορίες που παρέχονται από τα δείγματα μας και να αντιμετωπίσουμε παρόμοια ερωτήματα, όπως εκείνα που τέθηκαν παραπάνω. Σε αυτό το πλαίσιο, εισάγουμε και μελετάμε το k-Set sampling setting για τα μοντέλα Mallows και Plackett-Luce, επεκτείνοντας τα προηγούμενα ερευνητικά αποτελέσματα. Ταυτόχρονα, εισάγουμε και ένα άλλο μοντέλο δειγματοληψίας με θόρυβο, το μοντέλο k-Gap Filling Mallows.

Λέξεις - Κλειδιά

Στατιστική Μάθηση, Μηχανική Μάθηση, Θεωρία Μάθησης, Θεωρία Πιθανοτήτων, Θεωρία Πληροφορίας, Θεωρία Ψηφοφορίας, Θεωρία Κοινωνικής Επιλογής, Αλγόριθμοι και Πολυπλοκότητα

Abstract

In this thesis, we study the problem of learning rankings using noisy samples. This statistical learning field is extremely useful in the areas of Preference Learning and Information Retrieval. The working setting implies that one is given independent samples, which are permutations of n alternatives, generated by a distribution, that corresponds to a noisy probabilistic model. Such known probabilistic models are the Mallows Model and the Plackett-Luce Model. Having drawn the samples, one could ask questions concerning the sample complexity in order to learn the parameters of the generating distribution, the ability to learn the generating distribution itself in various f-divergence metrics, such as the TV distance and the KL divergence, and the notion of maximum likelihood estimation. At first, we present the extended work on that framework for the Mallows model combining some classical research results with some seminal work. Afterwards, we present our own work where we chose to reduce the information provided by our samples and cope to answer similar questions as the ones mentioned above. Hence, we introduce and study the k-Set sampling framework for both Mallows and Plackett-Luce models, extending the previous research results. At the same time, we introduce another novel sampling model, namely the k-Gap Filling Mallows model.

Keywords

Statistical Learning, Machine Learning, Learning Theory, Probability Theory, Information Theory, Voting Theory, Social Choice, Algorithms and Complexity

Acknowledgements

. . .

A.K.

Contents

List of Figures

1. Εκτεταμένη Ελληνική Περίληψη

Δίνουμε μία εκτεταμένη ελληνική περίληψη που συνοψίζει το περιεχόμενο αυτής της διπλωματικής εργασίας. Θα παρουσιαστούν συνοπτικά τα περιεχόμενα κάθε κεφαλαίου, χωρίς αποδείξεις και τεχνικές λεπτομέρειες.

1.1 Εισαγωγή

Οι διατάξεις- permutations είναι συνδυαστικά αντικείμενα που χρησιμοποιούνται καθημερινά από τους ανθρώπους. Από την λεξικογραφική διάταξη των λέξεων μίας γλώσσας και την κατάταξη αθλητικών ομάδων σε ένα πρωτάθλημα μέχρι τις προτιμήσεις ενός χρήστη στο YouTube και τις απαντήσεις μίας αναζήτησης στο Google, είναι εύκολο κανείς να παρατηρήσει πως η έννοια της διάταξης ή της κατάταξης-ranking εμφανίζεται σε ένα ευρύ φάσμα κατηγοριών με ποικίλες αναπαραστάσεις.

Ταυτόχρονα, ο σύγχρονος κόσμος -επιστημονικός και μη- βιώνει μία άνθηση της Επιστήμης των Υπολογιστών και, συγκεκριμένα, μία έκρηξη γύρω από την Επιστήμη της Μάθησης. Από την περίοδο που ο Alan Turing πρότεινε την ομώνυμη δοκιμή - Turing Test [[Tur50], 1950]- , συσχετίζοντας την έννοια της μηχανής με αυτήν της γνώσης και την περίοδο που ο Arthur Samuel όριζε τη μηχανική μάθηση ως Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μάθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί' [[Sam59], 1959], έχουμε φτάσει στο σημείο οι υπολογιστές να γίνονται ψηφιακοί προσωπικοί βοηθοί [TD18], να παράγουν πρωτότυπα δομημένα κείμενα [DP18], και μουσικά τραγούδια [DP16], να δημιουργούν πίνακες ζωγραφικής [AE17], απλά παρατηρώντας δεδομένα και περνώντας μία φάση εκπαίδευσης (training phase),προσομοιώνοντας της ανθρώπινη μάθηση.

Αναπόφευκτα, ο κόσμος των διατάξεων δεν θα μπορούσε να μην απασχολήσει εκείνον της Επιστήμης της Μάθησης. Έτσι, γεννήθηκε ο τομέας του Machine-Learning Ranking (MLR). Η φιλοσοφία του "Learning to rank" πεδίου είναι η κατασκευή μοντέλων διατάξεων για συστήματα ανάκτησης πληροφορίας. Το ranking model εκπαιδεύεται με

δεδομένα, τα οποία είναι λίστες από αντικείμενα τα οποία κατατάσσονται με κάποιο κριτήριο και ¨μαθαίνει' να διατάσσει νέες λίστες από αντικείμενα σύμφωνα με τον τρόπο με τον οποίο εκπαιδεύτηκε.

Εφαρμογές τέτοιων μοντέλων μπορεί κανείς να παρατηρήσει, για παράδειγμα, σε recommendation systems. Αλγόριθμοι μάθησης αναλύουν το ιστορικό των αγορών ενός πελάτη με σκοπό να ¨μάθουν' ένα preference ranking και, έπειτα, να προτίνουν παρόμοια προϊόντα. Εν γένει, μοντέλα MLR μπορούν να χρησιμοποιηθούν σε πληθώρα τομέων όπως το διαδίκτυο (μηχανές αναζήτησης), η υπολογιστική βιολογία (protein structure prediction problem), η επεξεργασία φυσικής γλώσσας και το Data Mining.

Στην παρούσα εργασία, θα προσπαθήσουμε να μελετήσουμε πολλές οπτικές και μοντέλα αυτού του πεδίου της Επιστήμης της Μάθησης. Συγκεκριμένα, θα παρατηρήσουμε τρόπους με τους οποίους η Θεωρία Πιθανοτήτων και Στατιστικής, καθώς και η Θεωρία Πληροφορίας, εντάχθηκαν στον κόσμο των Αλγορίθμων και της Πολυπλοκότητας, επεκτείνοντας τα όρια της Θεωρίας Στατιστικής Μάθησης.

Κύριο Πρόβλημα

Έστω ένα σύνολο με n αντικείμενα $\{a_i\}_{i=1}^n$, τα οποία μπορούν να διαταχθούν σύμφωνα με μία μετρική. Για παράδειγμα, έστω ένα πρωτάθλημα n ομάδων, όπου η κάθε μία παίζει με τις άλλες $(n-1)$. Τότε, στο τέλος του πρωταθλήματος, κάποια ομάδα a_{i_1} θα είναι πρώτη, κάποια a_{i_2} δεύτερη, κοκ. Εδώ η μετρική σύγκρισης είναι το πλήθος νικών της κάθε ομάδας. Η διάταξη αυτή, έστω $\pi_0 = (a_{i_1} \succ a_{i_2} \succ ... \succ a_{i_n})$, μας είναι κρυφή και δεν έχουμε άμεση πρόσβαση σε αυτή. Εμείς, όμως, επιθυμούμε να την ανακαλύψουμε. Αυτό που μπορούμε να κάνουμε είναι να παίρνουμε noisy samples από αυτή την κρυφή διάταξη. Δηλαδή, κάθε δείγμα μας είναι μία από τις $n!$ διατάξεις και η πιθανότητα να δειγματολειπτήσουμε κάποια διάταξη συσχετίζεται (is correlated) με την κρυφή διάταξη. Έτσι, μπορούμε να κάνουμε sampling κάθε φορά μία διάταξη των n αντικειμένων, η οποία όμως θα έχει θόρυβο, από το πιθανοτικό μοντέλο, το οποίο ακολουθεί κάποια κατανομή την οποία θα θέλαμε να ξέρουμε.

Βασικά ερωτήματα

- Ποιά είναι η κατανομή που ακολουθεί το πιθανοτικό μοντέλο; Τί μάζα πιθανότητας ανατίθεται σε κάθε μία από τις $n!$ υποψήφιες διατάξεις-δείγματα;

  → Υπάρχουν πολλά μοντέλα στο Learning to rank setting. Εμείς θα ασχοληθούμε κυρίως με το μοντέλο Mallows και το Plackett-Luce model, τα οποία θα μελετηθούν στα επόμενα κεφάλαια.

- Στην προηγούμενη παράγραφο αναφέραμε πως υπάρχει μία κρυφή διάταξη που επιθυμούμε να ανακαλύψουμε. Μπορούμε να μάθουμε την κρυφή διάταξη και, αν ναι, πόσα δείγματα θα χρειαστούμε ώστε να την μάθουμε με μεγάλη πιθανότητα;

  → Σε κάθε πιθανοτικό μοντέλο που ορίζεται πάνω στο σύνολο των διατάξεων, αντιστοιχούν κάποιες παράμετροι που το προσδιορίζουν. Στα περισσότερα

μοντέλα, μία από τις παραμέτρους είναι η κρυφή διάταξη που καλούμαστε να μάθουμε. Στην παρούσα εργασία, ασχολούμαστε ενδελεχώς με το ερώτημα Πόσα δείγματα θα χρειαστούμε ώστε να μάθουμε τις κρυφές παραμέτρους του μοντέλου, με μεγάλη πιθανότητα;΄

- Πώς σχετίζεται το παραπάνω πρόβλημα με την θεωρία ψηφοφορίας;

  → Διαβάζοντας το πρόβλημα που αναφέρουμε παραπάνω, μπορεί κανείς να κατασκεύασει μία αντιστοίχιση μεταξύ του προβλήματος και μίας διαδικασίας ψηφοφορίας. Η κρυφή διάταξη αναλογεί σε μία κρυφή από κοινού αλήθεια, μία διάταξη των υποψηφίων μίας εκλογικής διαδικασίας. Οι κοινωνιολόγοι μοντελοποιούν τον κάθε ψηφοφόρο ως ένα θόρυβο γύρω από αυτήν. Κάθε ψηφοφόρος έχει ως στόχο να μάθει αυτή την κρυφή αλήθεια και έτσι η ψήφος αποτελεί μία τυχαία μεταβλητή στον χώρο των πιθανών διατάξεων. Η πιθανοτική κατανομή της ψήφου έχει ως κέντρο την κρυφή αλήθεια και αναθέτει περισσότερη μάζα πιθανότητας σε διατάξεις-ψήφους που είναι κοντά στην κεντρική διάταξη από ότι σε διατάξεις που απέχουν από αυτή. Ποιά πιθανοτική κατανομή σας θυμίζει αυτή η συμπεριφορά;

1.2 Μαθηματικά Θεμέλια I, II, III

1.2.1 Άλγεβρα

Η δομική βάση των πιθανοτικών μοντέλων που θα αναλύσουμε είναι οι διατάξεις-μεταθέσεις (permutations). Η έννοια της μετάθεσης αποτελεί μία από τις πιο θεμελειώδεις οντότητες του κόσμου της Άλγεβρας. Μία διάταξη των αντικειμένων ενός συνόλου A είναι μια αντιστοιχία από το A στο A. Έστω A ένα μη κενό σύνολο και έστω $\mathbb{S}_A$ η συλλογή όλων των μεταθέσεων του A. Τότε η $\mathbb{S}_A$ είναι ομάδα με πράξη τον πολλαπλασιασμό μεταθέσεων. Η ομάδα όλων των μεταθέσεων του A ονομάζεται συμμετρική ομάδα για τους n χαρακτήρες και συμβολίζεται με $\mathbb{S}_n$.

Αυτό που μας ενδιαφέρει και μας είναι απαραίτητο ώστε να περιγράψουμε το πιθανοτικό μοντέλο από το οποίο θα παράγουμε δείγματα, είναι να ανάγουμε τον χώρο των μεταθέσεων που παρουσιάσαμε παραπάνω σε μετρικό χώρο. Δηλαδή, χρειάζεται να ορίσουμε μία έννοια απόστασης μεταξύ δύο αντικειμένω της $\mathbb{S}_n$.

- Έστω δύο μεταθέσεις $\sigma, \pi \in \mathbb{S}_n$. Πόσο απέχουν οι δύο μεταθέσεις;

Η απάντηση στο ερώτημα αυτό δεν είναι μοναδική. Υπάρχουν πάρα πολλοί τρόποι να μοντελοποιήσει κανείς την απόσταση δύο μεταθέσεων. Πριν περάσουμε στην περιγραφή των αποστάσεων, αξίζει να αναφέρουμε πως πλέον μας συμφέρει να σκεφτόμαστε την κάθε μετάθεση στο $\mathbb{S}_n$ ως μία διάταξη των n στοιχείων. Έτσι, για παράδειγμα, η μετάθεση $\sigma = \left(\begin{smallmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 5 & 1 & 4 \end{smallmatrix}\right)$, που σημαίνει πως η σ ικανοποιεί τις συνθήκες $\sigma(1) = 2, \sigma(2) = 3, \sigma(3) = 5, \sigma(4) = 1, \sigma(5) = 4$. Η μετάθεση αυτή αντιστοιχεί με μοναδικό τρόπο στην διάταξη $4 \succ 1 \succ 2 \succ 5 \succ 3$ των 5 στοιχείων.

Θα λέμε ότι $a \succ_\sigma b$, όταν a υπερέχει του b στην διάταξη σ, δηλαδή όταν η θέση του a είναι μικρότερη από αυτή του b :

$$a \succ_\sigma b \iff \sigma(a) < \sigma(b)$$

1.2.2 Θεωρία Πιθανοτήτων

Το πιο χρήσιμο εργαλείο που χρειάζεται κανείς από το κεφάλαιο αυτό είναι οι ανισότητες συγκέντρωσης και συγκεκριμένα η ανισότητα Hoeffding :

Ανισότητα του Hoeffding

Έστω $X_1, ..., X_n$ ανεξάρτητες τυχαίες μεταβλητές τ.ω. $\mathbb{P}[X_i \in [a_i, b_i]] = 1$. Αν είναι $S_n = \sum_{i=1}^n X_i$. Τότε για κάθε $\zeta > 0$, έχουμε:

$$\mathbb{P}[S_n - \mathbb{E}S_n \geq \zeta] \leq exp(\frac{-2\zeta^2}{\sum_{i=1}^n (b_i - a_i)^2})$$

και

$$\mathbb{P}[S_n - \mathbb{E}S_n \leq -\zeta] \leq exp(\frac{-2\zeta^2}{\sum_{i=1}^n (b_i - a_i)^2})$$

Από τον συνδυασμό αυτών των δύο ανισοτήτων, παίρνουμε:

$$\mathbb{P}[|S_n - \mathbb{E}S_n| \geq \zeta] \leq 2exp(\frac{-2\zeta^2}{\sum_{i=1}^n (b_i - a_i)^2})$$

Θα χρησιμοποιούμε συχνά αυτήν την ανισότητα για να αποκτήσουμε φράγματα για την δειγματική πολυπλοκότητα για τα προβλήματα μάθησης που θα ασχοληθούμε. Μια ευρεία συλλογή άλλων ανισοτήτων συγκέντρωσης μπορεί να βρεθεί στο [BS16] .

Επίσης, κομβική είναι η έννοια της απόκλισης μεταξύ δύο μέτρων πιθανότητας. Συγκεκριμένα, αναφέρουμε δύο αποκλίσεις :

Total Variation Distance

Η πρώτη μετρική απόκλισης είναι η ακόλουθη μετρική απόστασης, η οποία σχετίζεται με την l_1 νόρμα στον χώρο που ζουν τα μέτρα πιθανότητας.

$$d_{TV}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|$$

Ισοδύναμα, ισχύει :

$$d_{TV}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2} \sum_{x \in \Omega} |\mathbb{P}(x) - \mathbb{Q}(x)|$$

KL Divergence

Η δεύτερη μετρική δεν είναι μια συνάρτηση απόστασης, επειδή δεν είναι συμμετρική και παραβιάζει την τριγωνική ανισότητα. Για δύο διακριτά μέτρα πιθανότητας $\mathbb{P}, \mathbb{Q}$

$$D_{KL}(\mathbb{P} \parallel \mathbb{Q}) = \sum_{x \in \mathcal{X}} p(x) log \frac{p(x)}{q(x)}$$

1.2.3 Θεωρία Πληροφορίας

Η ανισότητα του Fano είναι ένα δημοφιλές θεωρητικό αποτέλεσμα του τομέας της Θεωρίας Πληροφορίας που παρέχει ένα κάτω φράγμα στην αναμενόμενη τιμή της TV distance μεταξύ του εκτιμητή μας και της πραγματικής κατανομής. Πολλές παραλλαγές της ανισότητας του Fano έχουν προκύψει στη βιβλιογραφία. Σε αυτή την διπλωματική εργασία, θα χρησιμοποιήσουμε την ακόλουθη εκδοχή.

Το ακόλουθο αποτέλεσμα οφείλεται στον Yu.

Ανισότητα Fano

Έστω $\mathcal{F}$ μία πεπερασμένη οικογένεια κατανομών τ.ω.

$$1. \quad \inf_{f,g\in\mathcal{F},f\neq g} d_{TV}(f,g) \geq a$$

$$2. \quad \sup_{f,g\in\mathcal{F},f\neq g} D_{KL}(f \parallel g) \leq b$$

Τότε είναι :

$$R_m(\mathcal{F}) \geq \frac{a}{2}(1 - \frac{mb + ln2}{ln|\mathcal{F}|})$$

Το $R_m(\mathcal{F})$ αντιπροσωπεύει το ελάχιστο αναμενόμενο σφάλμα οποιουδήποτε αλγόριθμου μάθησης όταν εκτελείται στη χειρότερη δυνατή κατανομή από την κλάση $\mathcal{F}$.

1.3 Θεωρία Ψηφοφορίας και Κοινωνικής Επιλογής

Η θεωρία κοινωνικής επιλογής ασχολείται με την συνάθροιση γνώμεων-προτιμήσεων με στόχο την εξαγωγή μίας ΄κοινής΄ απόφασης - από κοινού προτίμησης. Η ανάγκη για συνάθροιση προτιμήσεων και η εξαγωγή μίας καθολικής προτίμησης αναδεικνύεται σε τομείς όπως τα οικονομικά, την θεωρία αποφάσεων και τα εκλογικά συστήματα.

Πώς ξεκίνησαν όλα ;

Ένας από τους πρωτοπόρους αυτού του κλάδου και, συγκεκριμένα, της εφαρμογής μαθηματικών στο τομέα των κοινωνικών επιστημών, ήταν ο Γάλλος μαθηματικός και φιλόσοφος Marquis de Condorcet. Το 1785, ο Condorcet δημοσίευσε το έργο του με τίτλο 'Essay on the Application of Analysis to the Probability of Majority Decisions'. Στην εργασία του αναφέρει περίφημα αποτελέσματα, τα οποία αναφέρονται ακόμα και σήμερα ως το Παράδοξο του Condorcet και το θεώρημα των ενόρκων του Condorcet.

Condorcet's paradox. Έστω μία εκλογική διαδικασία με δύο υποψηφίους, όπου κάθε ψηφοφόρος έχει μία προτίμηση σε έναν εκ των δύο. Εάν η κοινωνία επιθυμεί να διαλέξει από κοινού έναν από τους δύο υποψηφίους, η επιλογή πλειοψηφικής ψήφου φαντάζει εύλογη και σωστή. Το ζήτημα που διέγνωσε ο Condorcet είναι το εξής :

Τί γίνεται αν οι υποψήφιοι είναι τρεις ή παραπάνω; Υπάρχουν προβλήματα με την πλειοψηφική ψήφο;

Ο Condorcet έδωσε το εξής παράδειγμα : Συμβολίζουμε με $a \succ_i b$ ότι ο ψηφοφόρος i προτιμά τον υποψήφιο a έναντι του b. Έστω τρεις υποψήφιοι a, b, c και τρεις ψηφοφόροι με τις ακόλουθες προτιμήσεις :

- $a \succ_1 b \succ_1 c$
- $b \succ_2 c \succ_2 a$
- $c \succ_3 a \succ_3 b$

Παρατηρούμε εύκολα πως η πλειοψηφία προτιμά τον a έναντι του b, τον b έναντι του c και τον c έναντι του a. Έτσι, η απο κοινού πλειοψηφική επιλογή είναι η $a \succ b \succ c \succ a$, η οποία δεν είναι συνεπής. Όποιος και να εκλεγεί, θα υπάρχει πλειοψηφία ατόμων που θα διαφωνεί με το αποτέλεσμα. Ισοδύναμα το γράφημα που θα μπορούσαμε να σχεδιάσουμε θα έχει κύκλο.

Έτσι, ο Condorcet διαπίστωσε πως ο πλειοψηφικός κανόνας εκλογής είναι μια αξιόλογη μέθοδος για την λήψη αποφάσεων σε συλλογικό επίπεδο, λόγω της απλότητας του, αλλά παρουσιάζει ένα πλήθος από σοβαρά προβλήματα. Συνεπώς, έκανε σαφές πως χρειάζεται να σχεδιασθούν μέθοδοι ψηφοφορίας αρκετά πιο σύνθετες, οι οποίες είτε θα επιλύουν ή θα παρακάμπτουν προβλήματα όπως τα παραπάνω.

Condorcet's jury theorem. Έστω μία ομάδα ενόρκων, οποία καλείται να αποφασίσει αν ένας κατηγορούμενος είναι αθώος ή ένοχος. Έστω ότι κάθε μέλος της επιτροπής έχει μία ίση και ανεξάρτητη πιθανότητα ορθής απόφασης $p \in (\frac{1}{2}, 1)$. Τότε η πλειοψηφία των ενόρκων είναι πιο πιθανό να είναι ορθή από κάθε ένορκο ξεχωριστά. Ταυτόχρονα, όσο το πλήθος των ενόρκων αυξάνει, η πιθανότητα ορθής απόφαση τείνει στο 1. Μαθηματικά, αυτό εκφράζεται ως ένα άθροισμα διωνυμικών της μορφής : $Maj(p,n) = \sum_{i=\lfloor n/2 \rfloor + 1}^{n} \binom{n}{i} p^i (1-p)^{n-i} \longrightarrow 1$, όσο το $n \to \infty$. Η $Maj(p,n)$ εκφράζει την πιθανότητα η πλεοψηφία να πάρει την σωστή απόφαση με n ενόρκους και πιθανότητα σωστής απόφασης p. Έτσι, υπό αυτές τις προϋποθέσεις, ο κανόνας της πλειοψηφίας είναι καλός. Από την άλλη, αν ήταν $p \in [0, \frac{1}{2}]$, τα αποτελέσματα αντιστρέφονται και η καλύτερη επιλογή να ήταν κανείς να διαλέξει έναν ένορκο στην τύχη και να δικάσει με βάση την απόφαση του τυχαία επιλεχθέντος ενόρκου.

Με την πάροδο του χρόνου, ερωτήσεις για την συνάθροιση προτιμήσεων συνεχώς ανέρχονταν στην επιφάνεια. Η πιο λογική μορφή συνάθροισης προτιμήσεων (διατάξεων) είναι η ακόλουθη :

Ο κανόνας του Kemeny

Λαμβάνοντας ένα προφίλ - διάνυσμα ψήφων $\vec{\sigma} = (\sigma_1, ..., \sigma_n) \in \mathcal{L}(A)^n$, ο κανόνας του Kemeny επιλέγει την κατάταξη τ που ελαχιστοποιεί την απόσταση KT από τις n δεδομένες ψήφους, δηλαδή :

$$\tau = arg \min_{\tau \in \mathcal{L}(A)} \sum_{i=1}^{n} d_{KT}(\tau, \sigma_i)$$

Βλέποντας αναλυτικά την παραπάνω εξίσωση, παρατηρούμε πως αυτό που επιθυμούμε είναι να ελάχιστοποιήσουμε την l_1 νόρμα πάνω στον μετρικό χώρο (S_n, d_{KT})των διατάξεων του S_n με απόσταση την $Kendall - Tau$ Στο ακόλουθο λήμμα, θα δείξουμε πως η επιλογή να ελαχιστοποιήσουμε την l_1 νόρμα αντιστοιχεί στο να βρούμε την

διάμεσο του μετρικού χώρου.

Ελαχιστοποίηση l_1 νόρμας

Δοθέντων των σημείων $p_1, ..., p_n \in \mathbb{R}$, η l_1 νόρμα $l_1(x) = \sum_{i=1}^{n} ||x - p_i||_1$ ελαχιστοποιείται από την διάμεσο των σημείων.

Σημειώστε πως το πρόβλημα αυτό είναι γνωστό NP-Hard πρόβλημα.

1.4 Πιθανοτικά Μοντέλα πάνω σε Διατάξεις

Όπως θα δείξουμε ότι το μοντέλο θορύβου του Condorcet αντιστοιχεί στο μοντέλο Mallows, που ορίζεται αργότερα. Έτσι, θα αναφερθούμε στην παραπάνω διαδικασία ως διαδικασία θορυβώδους ταξινόμησης Condorcet-Mallows, η οποία περιγράφεται ως εξής:

Algorithm 1 Condorcet-Mallows noisy ranking process

1. Let π_0 be the objective ranking and let $0 \leq p < \frac{1}{2}$.
2. **Initialization :** $\sigma \leftarrow \emptyset$.
3. For each pair of alternatives $a, b \in A$, s.t. $a \succ_{\pi_0} b$,
  3a. with probability $1 - p$, add $a \succ b$ to σ,
  3b. otherwise, add $b \succ a$ to σ.
if σ *is intransitive* **then**
  GOTO step (*2*).
else
  RETURN σ.
end

Ο παραπάνω αλοριθμος ήταν η βασική ιδέα που τροφοδότησε το κίνητρο για την δημιουργία πιθανοτικών μοντέλων πάνω σε διατάξεις. Το πιο διάσημο μοντέλο είναι το μοντέλο Mallows, που ορίζεται ακολούθως :

$$\mathbb{P}[\pi|\pi_0] = \frac{1}{Z(\phi, \pi_0)} e^{-\beta d_{KT}(\pi, \pi_0)} \tag{1.1}$$

Αξίζει να παρατηρήσει κανείς την ομοιότητα του μοντέλου με την δομή της κανονικής κατανομής.

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{|x-\mu|^2}{2\sigma^2}}$$

Η σταθερά κανονικοποίησης είναι:

$$Z(\phi, \pi_0) = Z(\phi) = \prod_{i=1}^{n-1} \sum_{j=0}^{i} \phi^j \tag{1.2}$$

Η διάταξη π_0 παίζει ακριβώς τον ίδιο ρόλο με το μ στην κανονική κατανομή.

1.5 Μάθηση διατάξεων από πληροφορία με θόρυβο

Το βασικό απότέλεσμα αυτού του κεφαλαίου είναι η δειγματική πολυπλοκότητα για την μάθηση της κρυφής διάταξης π_0 του απλού μοντέλου Mallows.

Μάθηση της διάταξης π_0

Για κάθε $\pi_0 \in \mathbb{S}_m$ *και κάθε* $\phi \in [0, \gamma)$, *υπάρχει ένας πολυωνυμικού χρόνου εκτιμητής* $\hat{\pi}$ *τ.ω. δοθέντων* $n = \Theta(\frac{1}{\gamma} log(\frac{m}{\delta}))$ *i.i.d. δειγμάτων* $\pi_1, ..., \pi_n \sim \mathcal{P}_{\phi,\pi_0}$ *ικανοποιεί* $\mathbb{P}[\hat{\pi} \neq \pi_0] \leq \delta$. *Επίσης, αν* $n = o(log(\frac{m}{\delta}))$, *τότε για κάθε εκτιμητή υπάρχει κατανομή* $\mathcal{P}_{\phi,\pi_0}$ *τ.ω.* $\mathbb{P}[\hat{\pi} \neq \pi_0] > \delta$.

Επίσης, ένα πολύ σημαντικό αποτέλεσμα αφορά την (αν)ικανότητα μας στο να μάθουμε την κατανομή $\mathcal{P}_{\phi,\pi_0}$ δοθέντων m δειγμάτων. Συγκεκριμένα, με χρήση της ανισότητας Fano, παίρνουμε το ακόλουθο inapproximity αποτέλεσμα σχετικά με την μάθηση κατανομών υπό την TV απόσταση :

Έστω $\phi^* = \frac{1}{2}$. *Τότε* $\exists \pi_0 \in \mathbb{S}_n$, *τ.ω. εάν δειγματολειπτήσουμε το προφίλ ψήφων* $\pi = (\sigma_1, ..., \sigma_m) \sim \mathcal{P}^m_{\phi^*,\pi_0}$, *όπου* σ_i *είναι i.i.d. δείγματα και εάν* $m = o(logn)$, *τότε κάθε κατανομή* $\mathcal{P}(\pi)$ *οφείλει να ικανοποιεί την ακόλουθη ανισότητα :*

$$\mathbb{P}_{\pi \sim \mathcal{P}^m_{\phi^*,\pi_0}}[d_{TV}(\mathcal{P}(\pi), \mathcal{P}_{\phi^*,\pi_0}) \geq \frac{1}{16}] \geq \frac{1}{3}$$

Συνεπώς, η ανισότητα Fano μας εξασφαλίζει ότι, αν δωθούν 'λίγα' δείγματα, ο εκτιμητής μας θα απέχει πάντα από την πραγματική κατανομή. Η απόσταση θα μπορούσε να παραμετροποιηθεί από μία ακτίνα ϵ, η οποία θα εμφανιζόταν στον παρονομαστή του αριθμού των δειγμάτων. Όσο το ϵ θα μειωνόταν, τα δείγματα θα αυξάνονταν και άρα η απόσταση της εκτίμησης μας από την πραγματική θα μειωνόταν.

1.6 Αναζητώντας τον Εκτιμητή Μέγιστης Πιθανοφάνειας (ΕΜΠ)

Ας υποθέσουμε ότι μας δίνονται r i.i.d. δείγματααπό μία κατανομή $\mathcal{P}_{\phi,\pi_0}$. Στόχος μας είναι να βρούμε τον ΕΜΠ - την διάταξη μέγιστης πιθανοφάνειας $\hat{\pi}^*$ από τα δείγματα που παρατηρούμε :

$$\hat{\pi}^* = arg\max_{\pi^*} \prod_{i=1}^{r} \mathbb{P}[\pi_i | \pi^*] = arg\max_{\pi^*} \prod_{i=1}^{r} \frac{e^{-\beta d_{KT}(\pi_i, \pi^*)}}{Z(\beta)}$$

Από την εκθετική δομή του μοντέλου μας, παίρνουμε :

$$\hat{\pi}^* = arg\max_{\pi^*} \prod_{i=1}^{r} \frac{e^{-\beta d_{KT}(\pi_i, \pi^*)}}{Z(\beta)} = arg\max_{\pi^*} \frac{1}{Z(\beta)^r} exp(-\beta \cdot \sum_{i=1}^{r} d_{KT}(\pi_i, \pi^*))$$

Έτσι :

$$\hat{\pi}^* = arg\max_{\pi^*} lne^{-\beta\cdot\sum_{i=1}^{r} d_{KT}(\pi_i,\pi^*)} = arg\max_{\pi^*}(-\beta\cdot\sum_{i=1}^{r} d_{KT}(\pi_i,\pi^*))$$

Τέλος, αφού $\beta > 0$,

$$\hat{\pi}^* = arg\min_{\pi^*}\sum_{i=1}^{r} d_{KT}(\pi_i,\pi^*)$$

Ο κανόνας του Kemeny είναι ο ΕΜΠ $\hat{\pi}^*$ για την κρυφή κεντρική διάταξη του μοντέλου Mallows.Όμως, αυτό το πρόβλημα είναι NP-Hard, όπως έχουμε ήδη παρατηρήσει. Έτσι, θα πρέπει να σχεδιάσουμε έναν αλγόριθμο, ο οποίος με μεγάλη πιθανότητα να βρίσκει τον ΕΜΠ.

Θεώρημα

Υπάρχει ένας πιθανοτικός αλγόριθμος τ.ω. εάν $\{\pi_i\}_{i=1}^{r}$ είναι διατάξεις πάνω σε n αντικείμενα και αποτελούν ανεξάρτητα δείγματα ενός μοντέλου Mallows με παράμετρο $\beta > 0$, και αν είναι $\alpha > 0$, τότε η διάταξη μέγιστης πιθανοφάνειας π^m μπορεί να υπολογιστεί σε χρόνο :

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

και με πιθανότητα σφάλματος $< n^{-\alpha}$.

1.7 k-Set Sampling

Ας είναι $A = \{a_1, ..., a_n\}$ ένα σύνολο αντικειμένων. Εισάγουμε το ακόλουθο μοντέλο δειγματοληψίας. Τα δείγματα μας εξακολουθούν να προέρχονται από μία κατανομή Mallows $\mathcal{M}_1(\pi_0, \phi)$, όμως πλέον δεν έχουμε πλήρη πρόσβαση στην διάταξη που προέκυψε από το μοντέλο.

Η δειγματοληψία μας παραμετροποιείται από μία παράμετρο $0 < k < n$. Μέχρι τώρα, παρατηρούσαμε διατάξεις $\pi_j \sim \mathcal{M}_1(\pi_0, \phi)$ των n αντικειμένων. Πλέον, από ένα δείγμα $\pi_j = a_{i_1} \succ a_{i_2} \succ ...a_{i_k} \succ a_{i_{k+1}} \succ ...a_{i_n}$, μπορούμε να παρατηρήσουμε μόνο τα k κορυφαία αντικείμενα της διάταξης αλλά δίχως να γνωρίζουμε την κατάταξή τους. Δηλαδή παρατηρούμε ένα σύνολο S_j μεγέθους k των k κορυφαίων αντικειμένων :

$$S_j = \{a_{i_1}, a_{i_2}, ..., a_{i_k}\}$$

Η μάζα που ανατείθεται στο σύνολο S από το μοντέλο SM είναι :

$$\mathbb{P}_{SM}[S|\pi_0] = \sum_{\pi_S\in g(S)}\sum_{\pi_R\in g(R)}\mathbb{P}_{MM}[\pi_S \uplus \pi_R|\pi_0] = \sum_{\pi_S\in g(S)}\sum_{\pi_R\in g(R)}\frac{\phi^{d_{KT}(\pi_S\uplus\pi_R,\pi_0)}}{Z(\phi)}$$

Προτείνουμε το ακόλουθο πρόβλημα :

MLE-MM-K-SET

Είσοδος : r ανεξάρτητα σύνολα $S_1, ..., S_r$ μεγέθους k.
Έξοδος : $\pi^* = arg\max_{\pi} \mathbb{P}_{SM}[S_1, ..., S_r|\pi]$

Αποδεικνύουμε ότι :

Θεώρημα 1

Η λύση του MLE-MM-K-SET *είναι η διάταξη* $argsort_{i\in[n]}\{v_1, ..., v_n\}$.

Παρόμοια, εισάγουμε την Plackett-Luce εκδοχή του k-Set Sampling, όπου δοθέντος ενός διανλυσματος αξιών $\vec{w} \in W$, το μοντέλο αναθέτει μάζα στο σύνολο S ίση με :

$$\mathbb{P}_{PL}[S|\vec{w}] = \sum_{\sigma\in g(S)} (\prod_{i\in[k]} w_{\sigma^{-1}(i)})(\prod_{i=1}^{k} \frac{1}{\sum_{j=i}^{n} w_{\sigma^{-1}(j)}})$$

Προτείνουμε το ακόλουθο πρόβλημα :

MLE-PL-K-SET

Περιβάλλον Εργασίας : Υπάρχουν n αντικείμενα $\{o_i\}_{i=1}^n$με άγνωστες αξίες $\{w_i\}_{i=1}^n$. Παράγουμε σύνολα από το μοντέλο PL-K-SET και επιθυμούμε να καθορίσουμε την διάταξη των αξιών των αντικειμένων
Είσοδος : r ανεξάρτητα σύνολα $S_1, ..., S_r$ μεγέθους k.
Έξοδος : $\vec{w}^*_{\pi} = arg\max_{\vec{w}_{\pi}} \mathbb{P}_{PL}[S_1, ..., S_r|\vec{w}_{\pi}]$

Αποδεικνύουμε ότι :

Θεώρημα 2

Η λύση του MLE-PL-K-SET *είναι η διάταξη* $\vec{w}^* = w_{\pi^{-1}(1)} \geq w_{\pi^{-1}(2)} \geq ... \geq w_{\pi^{-1}(n)}$ *όπου* $\pi = argsort_{i\in[n]}\{v_1, ..., v_n\}$.

2. Introduction

Permutations are combinatorial objects, used by humans on a daily basis. From the lexicographical order of the words of a language and the ranking of sport teams in a league to a user's preferences on YouTube and the results to a Google query, it is easy to notice how the notion of rankings arises in a wide range of fields with various representations.

At the same time, modern world -scientific or not- is experiencing a bloom of Computer Science and, in particular, an upsurge of Learning Science. Since the 50's, when Alan Turing proposed the famous Turing Test [[Tur50], 1950], associating the concept of the machine with that of knowledge and of learning and the period in which Arthur Samuel defined the machine learning as a "Field of study that gives computers the ability to learn without being explicitly programmed [[Sam59], 1959], we have reached a point where computers/machine become digital personal assistants [TD18], are generating structured texts [DP18], are composing music songs [DP16], are creating art [AE17], just by observing data and being trained and tested, simulating the human learning process.

Inevitably, the space of rankings could not be neglected by the Science of Learning. Hence, Machine-Learning Ranking (MLR) field arose. The philosophy of "Learning to rank" area is to develop probabilistic models over rankings/preferences for information retrieval systems. The ranking model is trained with data, which are lists of objects that are ranked by some criterion/metric, and "learns" to order new lists of objects.

Statisticians traditionally studied the problem of ranking data and designed methods and tools which have been applied in various fields. More recently, applications in information retrieval and machine learning have reanimated the interest in the analysis of rankings and in the value of related statistical tools such as probability distributions on rankings and correlation statistics.

Applications of such models can be observed, for instance, in recommendation systems. Learning algorithms analyze the history of a customer's purchases in order to "learn" a preference ranking and, afterwards, propose similar products. In general, MLR models can be used in a variety of fields such as the internet (search engines), computational biology (protein structure prediction problem), natural language processing and Data Mining.

In this thesis, we will try to study many perspectives and models on this field of Learning Science. Specifically, we will observe the ways in which Probability and Statistics Theory, as well as Information Theory, have been integrated into the world of Algorithms and Complexity, expanding the boundaries of the field of Statistical Learning theory.

In the following chapter, we are going to present some mathematical foundations, beginning from the main concept of permutations, continuing with the ideas of measure theory and of probability theory as a branch of that field and, finally, providing an information theoretic perspective.

Afterwards, we will try to connect our learning to rank problem with the fields of voting and social choice theory. As we will see, the idea of ranking over items, given a collection of preferences, is completely similar to the idea of electing a ranking over candidates, given a collection of votes.

In the third part of the thesis, the essential probabilistic models on permutation spaces are presented and emphasis is given on the notion of learning parameters of these models using noisy information. In the following two chapters, we analyze the sample complexity of the learning problem and we present the maximum likelihood estimator approach.

In the final chapter, we present our work concerning sampling from noisy samples. We study the behavior of the MLE by reducing the information provided by our samples. The way that the information provided is reduced will be presented shortly.

Before proceeding to the mathematical foundations of the 'Learning to Rank' field, it is useful to present the crucial problem that we are going to deal with in this thesis.

Main Problem

Consider a set of n objects/alternatives $\{a_i\}_{i=1}^n$, that can be ranked with respect to a metric. For instance, if the n objects are teams in a basketball championship, the ranking is created with respect to the number of wins of each team. Similarly, this ranking can express preferences among the n alternatives. Suppose that there is a true hidden preference among n objects $\{a_i\}_{i=1}^n$, that is expressed by a permutation over these alternatives $a_{i_1} \succ a_{i_2} \succ ... \succ a_{i_n}$. This true ranking is locked to us, we have no direct access to it and we want to learn it. In our setting, we generate from a probabilistic model, parametrized by the true hidden ranking, noisy samples that are rankings of the n objects.

Main Questions

- What is the probabilistic model that generates our samples? How are these samples/rankings distributed? What mass is attained to each of the $n!$ possible rankings?

  → There are multiple probabilistic models in the 'Learning to Rank' setting. We are going to deal with the Mallows model and the Plackett-Luce model, that will be presented in the upcoming chapters.

- In the previous section, we mentioned that there is a hidden ranking that we wish to discover. Can we learn the hidden ranking and, if so, how many samples will be needed in order to learn it with high probability?

  → Each probabilistic model, that is defined on the set of rankings, is determined by a set of parameters. In most models, one of the parameters is the hidden ranking we are have to learn. In this thesis, we deal in depth with the question 'How many samples will we need to learn the hidden parameters of the model, with high probability?'

- Is this problem connected with voting theory?

  → By reading the problem mentioned above, one can link our problem with a voting process. The hidden ordering corresponds to a ground truth, a socially accepted ranking over the candidates in an electoral process. Sociologists model each voter as a noise around this hidden ground truth. Every voter seeks to learn this hidden truth, so her vote is a random variable that takes values on the space of rankings over the voting alternatives. The probability distribution of the vote is centered on the hidden truth and assigns more probability mass to voting arrangements that are close to the central ordering than to rankings that are far from it. What known probability distribution does this behavior remind you?

An example

An everyday life application is the following. Alice watches the same n videos on YouTube daily. Thus, each day, she watches a sequence of these n videos. On the other side, Susan, working on YouTube, wants to learn Alice's video preferences and propose her similar videos that she will like in order to continue using the application. In this case, the hidden central ranking is Alice's video preferences, that exists in her mind but Susan has no access to it. Alice, watching YouTube, provides to Susan each day a sample video sequence, that is a noisy version of her inner video preferences. Of course, each sequence of Alice's videos is sorted by Susan according to some metrics that show the preferences of Alice. For instance, a video that Alice watched without skipping parts is ranked higher than a video where Alice skipped parts or did not

finish. Thus, each day, Susan gets a ranking of Alice's video preferences. Also, note that it is more likely more Alice to watch videos that she likes and, thus, the probability that Susan gets a ranking that is closer to Alice's preferences is higher than the probability that she gets a ranking that will differ a lot. Can Susan, given these daily samples, create a ranking that will be close to Alice's video preference list? How many samples will she need? These type of questions we will try to answer. But first we need to set the necessary mathematical foundations.

Our contribution

The purpose of this thesis was to study an innovative approach on the concept of noisy sampling. Specifically, we worked on the k-set sampling case. In our work, we chose to reduce the information provided by our samples and try to answer questions concerning the maximum likelihood estimation and the sample complexity. Let $A = \{a_1, ..., a_n\}$ be the set of our objects that we are ranking. There is still a hidden ranking π_0, that we want to learn. We still sample a ranking $\pi_j = a_{i_1} \succ a_{i_2} \succ ...a_{i_k} \succ a_{i_{k+1}} \succ ...a_{i_n}$, but we cannot access the sampled ranking. We can only access the k top ranked items in an unordered way, that is, our sample is a set S_j of size k with the top k alternatives :

$$S_j = \{a_{i_1}, a_{i_2}, ..., a_{i_k}\}$$

Now it should be clear why we named it k-set sampling. Afterwards, given those samples, we have to answer questions similar to the ones listed two sections before. Responses to some of these questions correspond to our contribution.

The study of top k lists was already researched in various works such as [FS03]. The innovative part appears in the set theoretic version of our sampling.

A real-life application of this sampling method is the classical voting (with a cross † next to the names) of our preferred k out of n alternatives in a voting procedure. Each vote is just a set of our k top preferred alternatives, without specifying the order of our preferences.

In our work, we provide the MLE of the k-Set Mallows Model, the MLE of the k-Set Plackett Luce Model. Also, we introduce the k-Gap Filling Mallows Model and provide a geometric perspective of sampling from that distribution.

As mentioned in the introductory chapter, the mathematical perspective of the 'Learning to Rank' field consists of a combination of pure mathematical ideas, that lie in the intersection of Abstract Algebra, Probability Theory & Statistics and Information Theory. Hence, before proceeding to the algorithmic extensions and applications, we consider that it would be prudent to delve into each of these three mathematical branches mentioned above in order to discover concepts and tools that will be useful in the upcoming chapters.

Chapter Contents

3. Mathematical Foundations I : Abstract Algebra

3.1 Abstract Algebra

In this thesis, we focus on learning rankings. Rankings can be modelled as combinatorial objects known as permutations. The theory of permutations is widely studied as a part of Abstract Algebra [Fra03], [Lan05] and, here, we are going to depict a general framework on how to use permutations.

3.1.1 Permutations

The probabilistic models - distributions we will analyze and use in the following chapters are based on the notion of rankings-permutations. The concept of permutation constitutes one of the most fundamental ideas in the area of Abstract Algebra. While the first references concerning the notion of permutations were reported in the 8th century, the fundamentals of permutation theory were developed by A.L. Cauchy (1789-1857). The classical definition follows :

Definition 3.1.1 *Permutation of a set A is called a bijection from A to itself.*

3.1.2 Symmetric group $\mathbb{S}_n$

Let A be a nonempty set of objects $\{o_1, ..., o_n\}$ and let $\mathbb{S}_A$ be the set of all possible permutations of the elements of A. We will show that the composition binary operator between a pair of functions is a well defined operator on the set $\mathbb{S}_A$.

Let σ, π be two permutations of the set A. The function $\sigma\pi$ is a mapping from A to itself and it defined by :

$$\sigma\pi : A \rightarrow_\pi A \rightarrow_\sigma A$$

For any $a \in A$, the function $\sigma\pi$ operates as follows : a is mapped by π and, afterwards, the element $\pi(a) \in A$ is mapped by σ. Now, we will show that $\sigma\pi$ is a

proper permutation. Hence, we have to show that it is a bijection from A to A. We show that in two steps :

- *injective* : Let $a, b \in A$. If $(\sigma\pi)(a) = (\sigma\pi)(b) \Rightarrow \sigma(\pi(a)) = \sigma(\pi(b))$. But, σ is injective since it is a permutation and, hence, $\pi(a) = \pi(b)$. Similarly, π is a one-to-one mapping, that implies $a = b$.
- *surjective* : Let $a \in A$. Since σ is a permutation, it is surjective. Thus, there exists an element $b \in A$ s.t. $\sigma(b) = a$. Additionally, since π is a permutation, it is surjective too and there is an element $c \in A$ s.t. $\pi(c) = b$. Hence, $a = \sigma(b) = \sigma(\pi(c))$ and $\sigma\pi$ is onto A.

In the literature, the permutations' composition is often referred as multiplication.

Theorem 3.1.1 *Let A a nonempty set and let $\mathbb{S}_A$ a collection of all permutations of A. Then. $\mathbb{S}_A$ is a group with operation the permutations' multiplication.*

In this thesis, we will denote the set $\{1, 2, ..., n\}$ with $[n]$.

Definition 3.1.2 *Let A be the finite set $[n]$. The collection of all permutations of A is called the symmetric group of the n characters and will be denoted with $\mathbb{S}_n$.*

The cardinality of $\mathbb{S}_n$ is $n!$.

3.1.3 Metric Space $(\mathbb{S}_n, d)$

In order to describe the probabilistic models from which we are going to generate permutations, we have to define appropriate distance metrics between permutations and, thus, work on a metric space $(\mathbb{S}_n, d)$ whose elements are rankings.

Let $\sigma, \pi \in \mathbb{S}_n$ be two permutations. What is the distance between the two rankings?

The answer to that question is not unique. There are multiple ways to describe the notion of distance between two elements of the symmetric group. Before proceeding to some useful descriptions, it is worth mentioning that we have to think of a permutation in $\mathbb{S}_n$ as a ranking of the n elements. For instance, consider the permutation $\sigma = \left(\begin{smallmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 5 & 1 & 4 \end{smallmatrix}\right)$, which means that σ satisfies the conditions : $\sigma(1) = 2, \sigma(2) = 3, \sigma(3) = 5, \sigma(4) = 1, \sigma(5) = 4$. There is a unique correspondence between this permutation and the ranking $4 \succ 1 \succ 2 \succ 5 \succ 3$ of these 5 elements.

In general, we will say that $a \succ_\sigma b$, when a beats b in the ranking induced by permutation σ, that is when the position where a is mapped by σ is less than the position where b is mapped :

$$a \succ_\sigma b \iff \sigma(a) < \sigma(b)$$

Kendall's Tau ranking distance

We need to define a metric that measures the distance between elements of the symmetric group $\mathbb{S}_n$. Kendall's tau distance is a measure of the similarity of a pair of rankings. It is named after Maurice Kendall, who developed this distance measure in 1938. The Kendall's Tau ranking distance $d_{KT} : \mathbb{S}_n \times \mathbb{S}_n \to \mathbb{Z}_{\geq 0}$ is defined as :

$$d_{KT}(\pi,\sigma) = \sum_{1\leq i<j\leq n} \mathbb{1}\{(\pi(i)-\pi(j))(\sigma(i)-\sigma(j)) < 0)\} \tag{3.1}$$

This function measures the number of pairwise disagreements of a pair of rankings. Clearly, the minimum value of that metric is 0, when we compare a permutation with itself and the maximum value is attained when comparing a permutation with its inverse. For a permutation $\pi \in \mathbb{S}_n$, its inverse permutation maximizes the Kendall tau distance, that is $\pi^{-1} = arg\max_{\sigma} d_{KT}(\pi,\sigma)$ with value $\binom{n}{2}$. From an algorithmic point of view, KT distance counts the number of steps needed for Bubble sort to return the sorted list. Kendall tau distance can be computed in $O(nlogn)$ with a modification of the Merge sort algorithm. We note that there exists a faster $O(n\sqrt{logn})$ algorithm [Cha10] for computing the Kendall tau distance, using the Van Emde Boas tree data structure.

Kendall's Tau distance properties

Kendall's Tau distance is a valid metric function and, thus, satisfies all the classical distance-metric properties :

1. $d_{KT}(\pi,\sigma) \geq 0$.
2. $d_{KT}(\pi,\sigma) = 0 \iff \pi = \sigma$.
3. $d_{KT}(\pi,\sigma) = d_{KT}(\sigma,\pi)$.
4. $d_{KT}(\pi,\sigma) \leq d_{KT}(\pi,\tau) + d_{KT}(\tau,\sigma)$.

Relabeling

It is worth mentioning that when computing the Kendall Tau distance between two permutations, it is always equivalent to compute the distance between the identity element of the symmetric group and another permutation. The Kendall Tau distance is invariant under relabeling.

Let π be a permutation of $\mathbb{S}_n$. Then,

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

Specifically, by taking the inverse of a permutation, :
$d_{KT}(\pi,\sigma) = d_{KT}(id,\sigma\pi^{-1}) = d_{KT}(\pi\sigma^{-1},id)$

Thus, we can always consider the identity permutation as the reference one π_0. Hence, we can assume that that $\pi_0 = id$. We will write $d_{KT}(\pi,id) = d_{KT}(\pi)$.

Major index and Mahonian number

An interesting question arising from the previous analysis is the following :

Consider a central permutation $\sigma \in \mathbb{S}_n$, that wlog can be the id, and let $S_d = \{\tau : d_{KT}(\tau) = d\}$ for $d \in \{0, 1, ..., \binom{n}{2}\}$. What is the cardinality of S_d?

Geometrically, we could think of the central permutation as the center of $\binom{n}{2} + 1$ circles with increasing radii and the question is how many permutations lie on the circle of radius d.

Unfortunately, the cardinality of this set cannot be expressed in closed form.

Suppose that we have a permutation $\pi \in \mathbb{S}_n$. Alexander MacMahon defined the *major index* statistic of a permutation, as follows :

$$MAJ(\pi) = \sum_{\pi(i)>\pi(i+1)} i \tag{3.2}$$

The majority index records the positions ($1 \leq i \leq n-1$,) where we have descents, and returns their sum. For instance, $MAJ(4 \succ 2 \succ 3 \succ 1) = 1 + 3 = 4$. Also, informally, let an inversion be the occurrence of a larger number before a smaller one (considering that our reference is the identity permutation). Formally, $INV(\pi) = |\{(i, j) : 1 \leq i < j \leq n, \pi(i) > \pi(j)\}| = d_{KT}(\pi)$.

MacMahon showed that the number of permutations of $\mathbb{S}_n$ with major index k equals to the number of permutations of $\mathbb{S}_n$ with k inversions. This number is called the Mahonian number $M(n, k)$. Equivalently, the distributions of MAJ and INV over $\mathbb{S}_n$ are the same, i.e., there is equality of the generating functions. For a positive integer n, define

$$[n]_x = \frac{1 - x^n}{1 - x} = 1 + x + ... + x^{n-1}$$

MacMahon showed that :

$$\sum_{\pi \in \mathbb{S}_n} x^{MAJ(\pi)} = \sum_{\pi \in \mathbb{S}_n} x^{INV(\pi)} = [1]_x [2]_x \dots [n]_x = \prod_{i=0}^{n-1} \sum_{j=0}^{i} x^j$$

In the symmetric group $\mathbb{S}_3$, we have the following table. For example, for the permutation 231, the inversions statistic equals to $INV(231) = 2$, since 2 and 3 occur before 1 and the major index is $MAJ(231) = 2$, since the descent occurs at position 2.

Inversions and Major index statistics for $\mathbb{S}_3$.

Permutations of $\mathbb{S}_3$	Inversions	Major Index
123	0	0
132	1	2
213	1	1
231	2	2
312	2	1
321	3	3

Thus, we can see that, for both metrics, there are one permutation with value 0, two with value 1, two with value 2 and one with value 3.

The Mahonian numbers can be expressed as a change of polynomial basis as follows :

Suppose that we have the polynomial :

$$\prod_{i=0}^{n-1}\sum_{j=0}^{i} x^j = 1(1+x)\dots(1+x+...+x^{n-1})$$

and we want to convert it to the monomial basis x^k. The coefficients of this convertion will be called Mahonian numbers, that is :

$$\prod_{i=0}^{n-1}\sum_{j=0}^{i} x^j = \sum_{k=0}^{\infty} M(n,k)x^k \tag{3.3}$$

Notice that in the above equation (3.3), the RHS sum needs not to run up to infinity. What is the range in which k runs in the RHS sum? We will show that $M(n,k)$ is equal to the number of elements of $\mathbb{S}_n$ with k inversions. Let $I(n,k)$ denote the number of permutations of length n with k inversions. Clearly, $0 \leq k \leq \binom{n}{2}$.

Theorem 3.1.2 *The generating function of the numbers $I(n,k)$ is*

$$G(x;n) = \sum_{k=0}^{\binom{n}{2}} I(n,k)x^k = \prod_{i=0}^{n-1}\sum_{j=0}^{i} x^j = \frac{1}{(1-x)^n}\prod_{j=1}^{n}(1-x^j)$$

Proof. We will work inductively. Firstly, $I(n,0) = 1 = G(1;n)$ for all n. Suppose that the formula holds for $n-1$ elements. Hence,

$$G(x;n-1) = \frac{1}{(1-x)^{n-1}}\prod_{j=1}^{n-1}(1-x^j)$$

We insert the n-th element at position $j \in [n]$ randomly. Since the n-th element is larger that the other $n-1$ elements, its insertion at position j will generate $n-j$ additional inversions. The previous inversions do not change. Since, each number of additional inversions is equally likely to be occur, the generating function is $1 + x + x^2 + ... + x^{n-1}$. The new inversions that are added are independent from the inversions in the permutation of length $n-1$. Thus, the generating function is simply the product :

$$G(x;n) = (1+x+x^2+...+x^{n-1})G(x;n-1)$$

The result follows.

■

Thus, we proved that $M(n,k) = I(n,k)\ \forall n, \forall\ 0 \leq k \leq \binom{n}{2}$ and there is no closed form for the answer of the question posed in the beginning of the section. In conclusion, we have that $|S_d| = |\{\tau \in \mathbb{S}_n : d_{KT}(\tau) = d\}|$ equals $I(n,d)$. The Mahonian numbers sequence is the sequence $\boxed{OEIS - A008302}$. The number of permutations at each possible Kendall tau distance d for n elements $S(n,d)$ can be computed recursively :

$$S(n,d) = \begin{cases} 1, & \text{if } d \leq 0 \\ S(n,d-1) + S(n-1,d) - S(n-1,d-n) & otherwise \end{cases}$$

Decomposition vector

For $\pi \in \mathbb{S}_n$, the Kendall tau distance $d_{KT}(\pi) = d_{KT}(\pi, id)$ can be decomposed uniquely to a $(n-1)-$ dimensional vector $\vec{V}(\pi)$, where :

$$\vec{V}(\pi) = (V_1(\pi), ..., V_{n-1}(\pi))$$

where $V_i(\pi)$ counts the number of elements smaller that $\pi(i)$ in the tail of the permutation (so the index runs from $i+1$ up to n.) Formally,

$$V_i(\pi) = \sum_{j=i+1}^{n} \mathbb{1}_{\pi(j)<\pi(i)}$$

This decomposition seems clear if one reconsiders the definition of KT distance. In the definition, the sum $\sum_{1\leq i<j\leq n}$ can be decomposed into two sums $\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}$. The first sum corresponds to the $n-1$ positions of the vector and the second sum is hidden inside the definition of V_i for $i \in [n-1]$.

Thus, it is clear that :

$$d_{KT}(\pi) = \sum_{i=1}^{n-1} V_i(\pi), \forall \pi \in \mathbb{S}_n \tag{3.4}$$

■ **Example 3.1** For the permutation $\pi = 53124$, we have that :

$$d_{KT}(53124) = \sum_{i=1}^{4} V_i(53124) = 4 + 2 + 0 + 0 = 6$$

■

Note that there is a bijection between permutations $\pi \in \mathbb{S}_n$ and decomposition vectors $\vec{V}(\pi)$.

Thus, a ranking of n objects can be represented as a data point located in Euclidean space $\mathbb{R}_{n-1}$. Therefore, only rankings of three or four objects can be represented in a two-dimensional or three-dimensional graph without losing any information. For instance, ranking data with three objects can be displayed on a hexagon, in which each vertex represents a ranking and each edge connects two rankings that

differ by swapping two objects (not neccesarily adjacent) the values of which differ by one. Hence, each edge has length $\sqrt{2}$. In general, a *permutohedron* of order n is a $(n-1)$-dimensional polytope, whose vertices are the elements of the symmetric group $\mathbb{S}_n$.

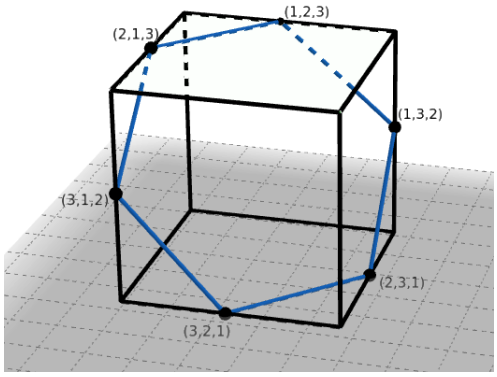

Figure 3.1: Permutohedron of order 3

In the above figure, we can see the 2-dimensional polytope, generated by the elements of $\mathbb{S}_3$. The other visualizable permutohedron is the one generated by the 24 permutations of $\mathbb{S}_4$ and is provided below.

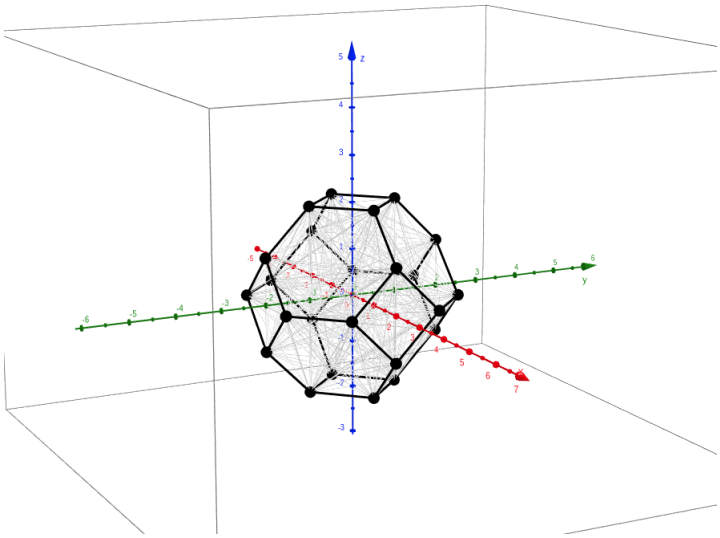

Figure 3.2: Permutohedron of order 4

Swap Increasingness

Consider two permutations σ, π such that $a \succ_\sigma b$ and $a \succ_\pi b$. These permutations contain the elements of a set A with size n, that is isomorphic to $\mathbb{Z}_n$. Then, the set of these permutations generated by the set A is defined as $\mathcal{L}(A)$ and is isomorphic to

$\mathbb{S}_n$. Suppose that we swap objects a, b in π. Then, the swapped permutation will be denoted by $\pi_{a\leftrightarrow b}$. What is the connection between $d_{KT}(\sigma, \pi)$ and $d_{KT}(\sigma, \pi_{a\leftrightarrow b})$?

We can define a swap monotonicity notion. Specifically, an integer-valued distance function d, defined on the symmetric group $\mathbb{S}_n$, is called *swap-increasing* if :

1. $\forall \sigma, \pi \in \mathcal{L}(A)$ and $a, b \in A$ s.t. $a \succ_\sigma b \wedge a \succ_\pi b$ implies that $d(\sigma, \pi_{a\leftrightarrow b}) \geq d(\sigma, \pi) + 1$,

2. and, if a, b are adjacent in π, then $d(\sigma, \pi_{a\leftrightarrow b}) = d(\sigma, \pi) + 1$.

We claim that the Kendall tau distance is swap-increasing. This property will be useful in the later chapters, but we choose to present it now.

Lemma 3.1.3 *The Kendall tau distance d_{KT} is swap-increasing.*

Proof. We recall that :

$$d_{KT}(\sigma, \pi) = \sum_{1\leq i<j\leq n} \mathbb{1}\{(\sigma(i) - \sigma(j))(\pi(i) - \pi(j)) < 0)\}$$

Let $\sigma, \pi \in \mathcal{L}(A)$ and $a, b \in A$ s.t. $a \succ_\sigma b \wedge a \succ_\pi b$. Suppose that $\pi(a) = i, \pi(b) = j$. Then, $i < j$.

We consider the set of elements of A that are between a, b in π, that is :

$$B = \{e \in A : i < \pi(e) < j\}.$$

Since, $a \succ_\sigma b$, it follows that $\sigma(a) < \sigma(b)$.
Obviously, we have that $\mathbb{1}\{\sigma(a) < \sigma(b)\} = 1, \mathbb{1}\{\sigma(b) < \sigma(a)\} = 0$.

Consider a random element e of the set B.

1. If $\sigma(e) < \sigma(a)$, then $\sigma(e) < \sigma(b)$. Adding over the elements of B :

$$\sum_{e\in B} \mathbb{1}\{\sigma(e) < \sigma(a)\} \leq \sum_{e\in B} \mathbb{1}\{\sigma(e) < \sigma(b)\}$$

2. Similarly, if $\sigma(b) < \sigma(e)$, then $\sigma(a) < \sigma(e)$. Adding over the elements of B :

$$\sum_{e\in B} \mathbb{1}\{\sigma(b) < \sigma(e)\} \leq \sum_{e\in B} \mathbb{1}\{\sigma(a) < \sigma(e)\}$$

Now we study the difference $d_{KT}(\sigma, \pi_{a\leftrightarrow b}) - d_{KT}(\sigma, \pi)$.
On the one hand, we have that :

$$d_{KT}(\sigma, \pi_{a\leftrightarrow b}) = \sum_{e\in B} \mathbb{1}\{\sigma(e) < \sigma(b)\} + \sum_{e\in B} \mathbb{1}\{\sigma(a) < \sigma(e)\} + \mathbb{1}\{\sigma(a) < \sigma(b)\}$$

On the other hand :

$$d_{KT}(\sigma, \pi) = \sum_{e\in B} \mathbb{1}\{\sigma(e) < \sigma(a)\} + \sum_{e\in B} \mathbb{1}\{\sigma(b) < \sigma(e)\} + \mathbb{1}\{\sigma(b) < \sigma(a)\}$$

Subtracting these two terms, we get that :

$$d_{KT}(\sigma, \pi_{a \leftrightarrow b}) - d_{KT}(\sigma, \pi) \geq 1$$

Now for the second property, when a, b area adjacent in σ, the above observations *1,2* are simply equalities and thus the last inequality reduces to equality because equal terms are getting canceled out. ■

Dislocation distance - Spearman's Footrule

Another distance function between permutations is the dislocation distance or usually mentioned as the Spearman's Footrule.

We define the distance of the permutations $\sigma, \pi \in \mathbb{S}_n$ as :

$$d_{SF}(\sigma, \pi) = \sum_{i=1}^{n} |\sigma(i) - \pi(i)| \tag{3.5}$$

The reader can think of the SF distance as the l_1 norm embedded in the symmetric group.

$$d_{SF}(\sigma, \pi) = ||\sigma - \pi||_1 \tag{3.6}$$

Common properties with KT distance

The three distance properties and the relabeling still hold for the Spearman's Footrule.

A main difference with KT distance

In the previous section, we observed that KT distance is swap increasing. Is the Spearman's Footrule swap increasing? The answer is no. We provide a simple counterexample. Let $\sigma = a \succ b \succ c$ and $\pi = b \succ c \succ a$. Then, $d_{SF}(\sigma, \pi) = 4$ and, for the swapped $\pi_{b \leftrightarrow c} = c \succ b \succ a$, we have that $d_{SF}(\sigma, \pi_{b \leftrightarrow c}) = 4$.

Comparing KT distance with Spearman's Footrule

It would be useful to obtain a result that shows the comparison between Kendall's tau distance and the Spearman's Footrule. Diaconis and Graham, in their joint work [DP77] , provided the following result. We mention that $d_*(\tau, id) = d_*(\tau)$.

Lemma 3.1.4 $\forall \tau, \frac{1}{2} d_{SF}(\tau) \leq d_{KT}(\tau) \leq d_{SF}(\tau)$

4. Mathematical Foundations II : Probability Theory

4.1 Probability Theory through Measure Theory

Learning rankings using noisy information requires the introduction of appropriate probabilistic models. The use of probability theory is essential in this thesis and so we will emphasize on that topic extensively. Modern probability theory consists, on a technical point of view, a branch of measure theory and so, our exposition of the subject will begin with some elementary measure-theoretic ideas. In order to acquire a profound knowledge on the topic of probability theory, we consider crucial to report some historical background and some fundamental concepts of the extensively studied topic of measure theory. Some excellent sources for the interested reader are [Tao13], [Kor07] and [Kal02].

Euclidean Geometry

The roots of measure theory can be found in Euclidean geometry, where one of the most significant concepts is that of the measure $m(E)$ of a solid body E in d dimensions, $d \geq 1$. When $d = 1, 2, 3$, we question "What is the *length, area* and *volume* respectively of E"? Back then, the idea of computing $m(E)$ was to partition the body, using translations or rotations, into finitely many simpler components, whose measure was possibly known. Archimedes was the one who also tried to obtain lower and upper bounds on $m(E)$ computing the measure of some inscribed or circumscribed body in E. As we will see, this intuition was crucial for the design of modern measure theoretic ideas.

Analytic Geometry

Persian mathematicians ($11th$ century) gave a direction to what René Descartes and Pierre de Fermat independently invented and called Analytic geometry ($17th$ century). Thus, Euclidean geometry was reconsidered as the study of the space

$\mathbb{R}^d = \underbrace{\mathbb{R} \times ... \times \mathbb{R}}_{\text{d times}}$ and rigid bodies of Euclidean geometry began to be thought as subsets E of $\mathbb{R}^d$. After that change, it was not clear how to define anymore, with a mathematical rigor, the measure $m(E)$. Some obvious issues were the following. Firstly, it was logical to think that the measure of a 'point', of a 'particle', of a quantity that cannot be divided into smaller parts should be considered to be 0. But, a rigid body/ set consists of an infinite number of particles/points, and, thus, its measure should result to $0 \cdot \infty$. Uncountable sets constitute another serious issue. Two sets that have the exact number of points, need not have necessarily the same measure. For example, trying to measure the intervals $E_1 = [0, 1]$ and $E_2 = [0, 2]$, one can obviously note that the length of the second interval is twice the length of the first. But, these two sets are in $'1 - 1'$ correspondence, thanks to $x \mapsto 2x$, and, thus, have the exact same number of elements.

Good and bad subsets of $\mathbb{R}^d$

Reading the above paragraph, one could think that the root of the problem is the infinite (and uncountable) number of components, partitioning our subset-body. What if we take only a finite number of partitions? One could think that this would be the solution to the problem. But, we still run into issues. The most famous way to prove our intuition wrong was given in 1924 by Stefan Banach and Alfred Tarski, expanding the works of Giuseppe Vitali and Felix Hausdorff. The so-called Banach-Tarski paradox shows that the unit ball in three dimensions $\mathbb{S}_2 = \{x \in \mathbb{R}^3 : ||x||_2^2 \leq 1\}$ can be decomposed into five pieces, which can be 'glued' back together, using rigid motions, to form two disjoint copies of the initial ball.

This decomposition is not simple and trivial and requires the use of the axiom of choice, that was formulated in 1904 by Ernst Zermelo in order to formalize his proof of the well-ordering theorem. The axiom of choice is necessary for the decomposition. There are models of set theory without the axiom of choice in which the Banach-Tarski paradox does not occur. To recall the axiom, consider a collection $\mathcal{C}$ of nonempty sets. We will say that a function f is a choice function on $\mathcal{C}$ if $\forall A \in \mathcal{C}, f(A)$ is an element of A. The axiom of choice states that for any collection $\mathcal{C}$ of nonempty sets, there exists a choice function f on $\mathcal{C}$.

The idea of measuring the 'right' way is to abandon trying to measure every subset of $\mathbb{R}^d$ and measure only the 'good' subsets of $\mathbb{R}^d$. We will refer to these sets as the *measurable sets*.

Measuring elementary sets

A box in $\mathbb{R}^d$ is a Cartesian product $B = I_1 \times ... \times I_d$ of d intervals. The volume $|B|$ of a box is simply $|B| = \prod_{i=1}^d |I_i| = \prod_{i=1}^d (b_i - a_i)$, where $I_i =< a_i, b_i >$ and $<= \{(, [\}, >= \{),]\}$. An elementary set is any subset of $\mathbb{R}^d$ which is the union of a finite number of boxes.

Let E be an elementary set. Then, E can be decomposed to a finite union of k disjoint boxes $\bigcup_{i=1}^k B_i$ Then, we can define the measure of an elementary set E as

the quantity

$$m(E) = \sum_{i=1}^{k} |B_i|, E \text{ elementary},$$

where the sum is independent of the partition. For the elementary measure $m(E)$ of an elementary set E, one can see the following :

- $m(E) \in \mathbb{R}_{\geq 0}$ *(Non-negativity)..*
- If $E_1, ..., E_k$ are disjoint elementary sets, then $m(E \cup ... \cup E_k) = m(E_1) + ... + m(E_k)$ *(Finite additivity).*
- $m(\emptyset) = 0$.
- $m(B) = |B|$ for all boxes B.
- If $E \subset F$, then $m(E) \leq m(F)$ *(Monotonicity).*
- If $E_1, ..., E_k$ are elementary sets, then $m(E \cup ... \cup E_k) \leq m(E_1) + ... + m(E_k)$ *(Finite subaddivity).*
- For all elementary sets E and $x \in \mathbb{R}^d$, $m(E + x) = m(E)$ *(Translation invariance).*

Jordan measure, Riemann-Darboux integral

Towards the end of the $19th$ century, the French mathematician Camille Jordan came up with the idea of Jordan measure. Jordan expanded the restricted class of elementary sets and introduced an approximation scheme inspired by the one Archimedes used, as we mentioned in the first section. Consider a bounded set $E \subset \mathbb{R}^d$. We say $elem(A)$ if the set A is elementary. In order to provide a further intuition, we try to link the ideas behind Jordan measure with the thoughts of Riemann and Darboux on the integrability concept. The classical Riemann-Darboux integral is closely related to Jordan measure. Firstly, the construction behind the Riemann integral, using Darboux lower and upper sums just like Jordan inner and outer measures and using piecewise constant functions just like elementary sets is completely similar to the Jordan meausure.

In parallel with the definition of measurability on elementary sets, one can define completely similarly the notion of integrability of piecewise constant functions. A piecewise constant function $f : [a, b] \to \mathbb{R}$ is a function for which there exists a partition of $[a, b]$ into finitely many intervals $I_1, ..., I_n$ s.t. f equals to a constant c_i in each one of then. Then, we define the piecewise constant integral of f on $[a, b]$ as :

$$p.c. \int_a^b f(x)dx = \sum_{i=1}^{n} c_i |I_i| \tag{4.1}$$

It is worth mentioning that this sum is independent of the partition of the $[a, b]$. The definition is completely similar to the measure of an elementary set.

Jordan defined the *Jordan inner measure of* E as :

$$m_{*,(J)}(E) = \sup_{A \subset E, elem(A)} m(A) \tag{4.2}$$

That is the biggest elementary set that fills E *from the inside.*

Similarly, Riemann used the lower Darboux integral, that is the best from below approximation of f using a piecewise constant function. If g is a piecewise constant function, we will write $pc(g)$

$$\underline{\int_a^b} f(x)dx = \sup_{g \leq f, pc(g)} p.c. \int_a^b g(x)dx \tag{4.3}$$

On the other side, Jordan introduced the *Jordan outer measure of* E as :

$$m^{*,(J)}(E) = \inf_{A \supset E, elem(A)} m(A) \tag{4.4}$$

That is the smallest elementary set that covers E *from the outside.*

Alongside with the outer Jordan measure, the upper Darboux integral is just the best from above approximation of f using a piecewise constant function, that is

$$\overline{\int_a^b} f(x)dx = \inf_{g \geq f, pc(g)} p.c. \int_a^b g(x)dx \tag{4.5}$$

It is well known that, in the Riemann-Darboux integrability concept, if these two integrals are equal, we say that f is Darboux integrable. A function is Riemann integrable if the Riemann sum with respect to a partition $\mathcal{P}$, $\sum_{i=1}^{n} f(x_i^*)\delta x_i$ converges to a real number as the $sup_{i \in [n]} \delta x_i$ goes to 0. This real number is called the Riemann integral of f. It is known that a function is Darboux integrable if and only if it is Riemann integrable.

Similarly, if the inner and the outer Jordan measures of E are equal then we say that E is Jordan measurable and set the Jordan measure to be equal to $m(E) = m_{*,(J)}(E) = m^{*,(J)}(E)$. It is important to note that elementary sets are Jordan measurable and their elementary measure coincides with the Jordan measure. Jordan only worked with bounded sets and did not consider unbounded sets to be Jordan measurable (they would have infinite measure). Jordan measurable sets are those sets that are 'almost elementary' with respect to Jordan outer measure.

Thus, the question arising is the following : *Is the Jordan measure enough?*

Lebesgue measure

The theory of Jordan measure works well when one works with Jordan measurable sets. Nevertheless, there are sets that are not Jordan measurable. One could show that the countable union or intersection of Jordan measurable sets $E_1, E_2... \subset \mathbb{R}$ need not to be Jordan measurable, even when bounded. Lebesgue extended the Jordan measures in order to tackle that issue. He tried to solve the problems by converting Jordan outer measure to a better upper estimator as follows.

One can use the finite additivity property and subadditivity of elementary measure to rewrite the Jordan outer measure. The outer Jordan measure just uses one elementary set to circumscribe the set E. We could instead cover it with a finite

collection of boxes and define the Jordan outer measure as the infimal cost required for the boxes cover.

$$m^{*,(J)}(E) = \inf_{\bigcup_{i=1}^{k} B_i \supset E, B_i boxes} |B_1| + ... + |B_k| \tag{4.6}$$

Lebesgue then proposed, instead of covering with a finite union of boxes, to cover E with a countable union of boxes and, thus, defined the Lebesgue outer measure of E :

$$m^{*}(E) = \inf_{\bigcup_{i=1}^{\infty} B_i \supset E, B_i boxes} \sum_{i=1}^{\infty} |B_i| \tag{4.7}$$

Note that $m^*(E)$ may be equal to $+\infty$. But, in most cases, this is a better approximation. Clearly, $m^*(E) \leq m^{*,(J)}(E)$.

Finally, Lebesgue introduced the measurability concept with respect to his measure if, given a set $E \subset \mathbb{R}^d$, we say that E is Lebesgue measurable if, for every $\epsilon > 0$, there exists an open set $U \subset \mathbb{R}^d$ that contains E s.t. $m^*(U) \leq \epsilon$. In that case, the Lebesgue measure is defined as $m(E)$ and is equal to the Lebesgue measure of E.

It is useful to note that

$$m_{*,(J)}(E) \leq m^{*}(E) \leq m^{*,(J)}(E), \forall E \subset \mathbb{R}^d \tag{4.8}$$

ans that Lebesgue measure extends Jordan measure,in the sense that every Jordan measurable set is Lebesgue measurable.

Abstracting measure spaces

While defining the Lebesgue measure, we only worked on subsets of $\mathbb{R}^d$. The Lebesgue measure m is the standard way of assigning a measure to subsets of the $n-$dimensional Euclidean space. Usually, it is necessary to work with more general spaces X, whose structure differs from the Euclidean space. Thus, abstraction of the notion of measurability is crucial.

Suppose that we want to work to a general space X and define a proper notion of measure. It is not enough to specify the set X. One needs to define, also, a collection $\mathcal{B}$ of subsets of X, where the measure will work well and a measure $\mu(A)$ that assigns to every set $A \in \mathcal{B}$ a value in $[0, +\infty]$.

Some questions that can easily arise from the above setting are the following :

What does the collection $\mathcal{B}$ consists of?
Does the measure function have to satisfy some axioms?

In this abstract setting, we will build our probability theory concept trying to answer these two elementary questions.

Probability theory as a measure theory branch

After this essential presentation of the main ideas behind measure theory, we are able to work on a specific branch of this widely studied topic, especially probability theory. We begin by recalling some fundamental definitions.

The first object one deals with in probability theory is the space of elementary outcomes, usually denoted by Ω. The elements of this non-empty space are the elementary outcomes $\omega \in \Omega$. In probability theory, we are interested in the probability measure (or simply probability) which maps some of subsets of Ω to the interval $[0, 1]$. But, these mappings cannot be random. For instance, when one says that event 'A happens' with probability $0 \leq p \leq 1$, then the probability measure should also attain value $(1 - p)$ to the event 'A does not happen'. This event is usually called the complement of A, denoted by A^c. Hence, it is crucial to create a 'good' notion of 'collection of subsets'.

Definition 4.1.1 — Boolean algebra. *A collection $\mathcal{B}$ of subsets of Ω is called a Boolean algebra if it has the following properties :*
1. $\Omega \in \mathcal{B}$.
2. $A \in \mathcal{B}$, implies that $A^c = \Omega \setminus A \in \mathcal{B}$ (stable under complement).
3. $A_1, ..., A_n \in \mathcal{B}$, then the union $\bigcup_{i=1}^{n} A_i \in \mathcal{B}$ (stable under finite union).

The way we defined the Boolean algebra is too minimal. We only assumed that it will be closed under two of the basic Boolean operations, the complement and the finite union.

In order to obtain an well defined measure notion, the finite union axiom of a Boolean algebra is not enough. The reason that is hidden behind this issue is the need of good behavior of the measure with limits. The intuition is completely similar to the finite box covering used by Jordan and the need to extend this notion to countable coverings by Lebesgue. The idea is the same. We require our Boolean algebra to be closed under countable unions. Countable union is usually assumed when one uses the greek letter σ as a prefix and thus we introduce the notion of σ-algebra.

Definition 4.1.2 — σ-algebra. *A σ-algebra $\mathcal{F}$ on Ω is a Boolean algebra that is closed under countable unions, i.e. if $(A_i)_{i\geq 1}$ is a sequence of sets in $\mathcal{F}$, then the union $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ (stable under countable unions). The elements of $\mathcal{F}$ are called measurable sets, or events.*

■ **Example 4.1** From the above definition, we have that $\emptyset, \Omega \in \mathcal{F}$. Two basic examples of a σ-algebra are the trivial σ-algebra $\underline{\mathcal{F}} = \{\emptyset, \Omega\}$ and the powerset of Ω, $\overline{\mathcal{F}} = \mathcal{P}(\Omega) = 2^{\Omega}$. In addition, it is obvious that any $\sigma-$algebra is also a Boolean algebra since one can think of any finite sequence of k sets as a countable sequence of these k sets and an infinite sequence of the empty set as the tail of the sequence. ■

Furthermore, it follows that if $(A_i)_{i\geq 1}$ is a sequence of sets in $\mathcal{F}$, then $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$.

■ **Example 4.2** Borel $\sigma-$ algebra. *The Borel $\sigma-$algebra of Ω is the $\sigma-$algebra $\sigma(A)$, where A is the family of open subsets of Ω.* ■

Definition 4.1.3 — Measurable space. *We refer to the pair $(\Omega, \mathcal{F})$ as a measurable space, where Ω is a space of elementary outcomes and $\mathcal{F}$ is a $\sigma-$algebra of subsets of Ω.*

Now, we represent the definition of measurable functions that will be useful in the introduction of random variables.

Definition 4.1.4 — Measurable function. *Let $(\Omega, \mathcal{F})$ be a measurable space. A function $\xi : \Omega \to \mathbb{R}$ is said to be $\mathcal{F}-$measurable (or measurable) if*

$$\{\omega : \alpha \leq \xi(\omega) < \beta\} \in \mathcal{F}, \ \forall \alpha, \beta \in \mathbb{R} \tag{4.9}$$

Right now, it will not be completely clear why we defined measurable functions this way. But, we promise that issues will be resolved when defining the notion of Lebesgue integrability in the next section.

■ **Example 4.3** Let E be an arbitrary subset of Ω. Define the indicator function $\mathbb{1}_E$ on Ω by

$$\mathbb{1}_E(x) = \begin{cases} 1, & x \in E \\ 0, & otherwise \end{cases}$$

We claim that E is measurable iff the indicator function $\mathbb{1}_E$ is measurable. It is not difficult to see that :

$$\{\mathbb{1}_E(x) \leq \beta\} = \begin{cases} \emptyset, & \beta < 0 \\ E^c, & 0 \leq \beta < 1 \\ \Omega, & \beta \geq 1 \end{cases}$$

The sets $\emptyset, \Omega$ are trivially measurable and E^c is measurable iff E is measurable, which implies our claim. ■

Remark 4.1.1 Continuous functions, monotone functions, step functions, Riemann-integrable functions are all Lebesgue measurable.

Before proceeding to the standard presentation of the probability measure, we provide the general definition of the measure function.

Definition 4.1.5 — Finite non-negative measure. *Let $(\Omega, \mathcal{F})$ be a measurable space. A function $\mu : \mathcal{F} \to [0, +\infty)$ is said to be a finite non-negative measure (or measure) if, whenever $\{A_i\}_{i \geq 1} \in \mathcal{F}$ are pairwise disjoint*

$$\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i) \tag{4.10}$$

A measure is a $\sigma-$additive function on a $\sigma-$algebra of subsets of Ω with values on the non-negative real axis.

If the measure takes values to $\mathbb{R}_{\geq 0} \cup \{+\infty\}$, then it will be called a $\sigma-$finite measure and if the measure maps to the whole real axis, it will be called signed measure.

4.1.1 Probability Measure

In modern probability theory, it is usual to link our objects of interest to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. This space is just a normalized measure space. Consequently, random variables could then be considered as measurable functions ξ on Ω and their expectation as Lebesgue integrals with respect to measure $\mathbb{P}$. It is of crucial importance to underline that the reference space Ω is used only for technical convenience. The choice of Ω plays no role and the interest focuses on the multiple induced distributions $\mathbb{P} \circ \xi^{-1}$.

Definition 4.1.6 — Probability measure. *A measure $\mathbb{P}$ on a measurable space $(\Omega, \mathcal{F})$ is called a probability measure (or probability distribution) if $\mathbb{P}[\Omega] = 1$.*

Definition 4.1.7 — Probability space. *The triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space, when $(\Omega, \mathcal{F})$ is a measurable space and $\mathbb{P}$ is a probability measure.*

■ **Example 4.4** (Normalized measure) Given any measure space $(X, \mathcal{B}, \mu)$ with $0 < \mu(X) < +\infty$, the space $(X, \mathcal{B}, \frac{1}{\mu(X)}\mu)$ is a probability space.

What does it mean to draw a sample uniformly at random from Ω?

If Ω is a non-empty finite set with the discrete $\sigma-$ algebra $\mathcal{P}(\Omega) = 2^{\Omega}$ and the counting measure $\#$, then the normalized counting measure $\frac{1}{\#\Omega}\#$ is a probability measure, that is known as the discrete uniform probability measure on Ω and the triplet $(\Omega, 2^{\Omega}, \frac{1}{\#\Omega}\#)$ is a probability space, that captures the 'drawing uniformly at random' notion. ■

Random elements, distribution functions and expectation

Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and some measurable space $(M, \mathcal{M})$. Any measurable mapping ξ of Ω into $(M, \mathcal{M})$ is called a random element in M.

A random element in M is called a random variable whenever $M = \mathbb{R}$, a random vector whenever $M = \mathbb{R}^d$, a stochastic process whenever M is a function space. In the strong majority of the situations that follow, we will refer to random variables.

If $S \in M$, then $\{\xi \in S\} = \xi^{-1}S \in \mathcal{F}$, and we consider the probabilities :

$$\mathbb{P}\{\xi \in S\} = \mathbb{P}[\xi^{-1}S] = (\mathbb{P} \circ \xi^{-1})S, S \in \mathcal{M} \tag{4.11}$$

The quantity $(\mathbb{P} \circ \xi^{-1})$ is a set function and is again a probability measure, defined on the range of the space M and called the probability distribution of the random element ξ.

For a random variable ξ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the function

$$F_{\xi}(x) = \mathbb{P}[\xi \leq x] = \mathbb{P}[\{\omega : \xi(\omega) \leq x\}] \tag{4.12}$$

is called the distribution function of the random variable ξ.

■ **Example 4.5 — Uniform distribution.** The function

$$F(x) = \begin{cases} 0, & x \leq 0 \\ x, & 0 \leq x < 1 \\ 1, & 1 \leq x < \infty \end{cases}$$

is the distribution function for a measure μ on the Borel $\sigma-$algebra $\mathcal{B}_{\mathbb{R}}$, that is concentrated on $(0, 1]$. The measure μ is called the *uniform distribution* on $(0, 1]$. ■

Some known, but useful properties are the following :

- F_ξ *is non-decreasing :* $x \leq y$ implies that $F_\xi(x) \leq F_\xi(y)$.
- $\lim_{x\to-\infty} F_\xi(x) = 0$ and $\lim_{x\to+\infty} F_\xi(x) = 1$.
- F_ξ *is continuous from the right* $\forall x : \lim_{x\downarrow t} F_\xi(x) = F_\xi(t)$.

It is interesting to point out that each function defined on $\mathbb{R}$, that has the three properties mentioned above, is a distribution function.

In some cases, there exists a non-negative integrable function f_ξ on $\mathbb{R}$ s.t.

$$F_\xi(x) = \int_{-\infty}^{x} f_\xi(y)dy \ \ \forall x$$

Then, f_ξ is called the probability density function of F_ξ.

One of the most important notions of probability theory is that of the expected value or expectation of a random variable. In order to remain to a measure theoretic setting, we shall underline that the concept of expectation is identical to the notion of the Lebesgue integral.

Intuition between Riemann and Lebesgue integration

Given a set X, a $\sigma-$algebra $\mathcal{B}$ and a measure μ on $\mathcal{B}$, we would like to define the integral

$$\int_X f d\mu$$

of any function f on X of an appropriate class of functions.

If X is a bounded closed interval $[a, b]$ of the real line, then the integral

$$\int_a^b f(x)dx$$

is well defined for the class of Riemann integrable functions. We remind that a bounded function on a compact interval $[a, b]$ is Riemann integrable if and only of it is continuous almost everywhere (the discontinuity set is of Lebesgue measure zero).

In order to compute the Riemann integral of f on $[a, b]$, we partition the interval into subintervals and the integral is the limit of the Riemann sum

$$\sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}), \text{where } \{x_i\}_{i=0}^{n} \text{ partition the interval } [a, b] \text{ and } \xi_i \in [x_{i-1}, x_i], i \in [n].$$

One could try to arrange the partition of the set X into a sequence of sets $\{E_i\}$ and define the integral sums :

$$\sum_i f(\xi_i)\mu(E_i), \text{where } \xi_i \in E_i.$$

Now, we have to clarify how one should take the limit at which the above sum will give the desired integral $\int_X f d\mu$.

What is the appropriate notion of limit?
What is the class of functions for which that limit exists?

Riemann's idea is to partition the interval into very small intervals, say Δ_i, on which, thanks to the continuity of the function, the range of the function f restricted in Δ_i is small (the values of f do not change much) and, hence, the restriction of f in the partition can be well approximated by $f(\xi_i)$. We should choose E_i's that respect that property.

Lebesgue approached the problem considering the following sets :

$$E_i = \{x \in X : t_{i-1} \leq f(x) < t_i\}$$

where $\{t_i\}$ is an increasing sequence of reals that partitions the range of f, say $Im(f)$.

Observe that this choice permits to avoid the use of continuity of f. But, this requirement is replaced by the necessity that the value $\mu(E_i)$ is well defined. Thus, our measure, that is defined on the $\sigma-$algebra $\mathcal{B}$ has to be well defined on a large domain that contains E_i. Similarly, we have to restrict functions f for which the sets of the form $\{\alpha \leq f(x) < \beta\}$ live in the domain of the measure μ. Now, we propose the reader to return to the definition of measurable functions 4.1.4. It should be clear why we defined measurable functions class this way.

In order to compute the Lebesgue integral of a one dimensional function f, we partition on the range of f. Thus, the integral should run over each value $t \in Im(f)$ and sum each elementary area contained between $y = t$ and $y = t - dt$. This area equals to $\mu(\{x|f(x) > t\})dt$. Then, the Lebesgue integral of f is defined by $\int f d\mu = \int_{\mathbb{R}} \mu(\{x|f(x) > t\})dt$.

Lebesgue integral

As the piecewise constant functions were the fundamental basis of the Darboux integrability, we will use the notion of simple functions for the Lebesgue integrability.

Let $(\Omega, \mathcal{F}, \mu)$ be a measurable space with a finite measure.

A measurable function is said to be *simple* if it takes a finite or countable number of values. Thus, if f is an (unsigned) simple function taking the non-negative values $c_1, c_2, ...$ and we define the sets $E_i = \{\omega : f(\omega) = c_i\}$, we can express f as a (finite or countable) sum of indicator functions, i.e.

$$f = \sum_i c_i \mathbb{1}_{E_i}$$

Definition 4.1.8 — Lebesgue integral for simple functions. *If the series $\sum_{i=1}^{\infty} c_i \mathbb{1}_{E_i}$ converges, then the sum of the series is called the (simple) Lebesgue integral of the simple function f and is denoted by $\int_\Omega f d\mu$. If the series diverges, then it is said that the integral equals $+\infty$.*

Lemma 4.1.2 *The integral of a simple non-negative function has the following properties*

- *Non-negativity, that is $\int_\Omega f d\mu \geq 0$*
- *Full measure, that is $\int_\Omega \mathbb{1}_\Omega d\mu = \mu(\Omega)$*
- *Linearity, that is $\int_\Omega (\alpha f + \beta g) d\mu = \alpha \int_\Omega f d\mu + \beta \int_\Omega g d\mu, \forall a, b > 0$.*
- *Non-decreasingness, that is $f \geq g \geq 0$ implies that $\int_\Omega f d\mu \geq \int_\Omega g d\mu$.*

In the measure theory literature, it is usually underlined that measurability behaves quite well with limits. We proceed by presenting a very important theorem that binds simple functions with general measurable functions.

Theorem 4.1.3 *Any non-negative measurable function f is a monotone limit from below of non-negative simple functions, that is $f(\omega) = \lim_{n\to\infty} f_n(\omega)\ \forall \omega \in \Omega$, where f_n are non-negative simple functions s.t. $f_n(\omega) \leq f_{n+1}(\omega) \forall \omega$.*

Proof. The proof is based on constructing this sequence of functions. We choose to define f_k by rounding down f to the nearest integer multiple of $\frac{1}{2^{k-1}}$.

For instance, for $k = 1$, we define

$$f_1(\omega) = \begin{cases} 0, & 0 \leq f(\omega) < 1 \\ 1, & 1 \leq f(\omega) \end{cases}$$

The next term of the sequence will be f_2 with values the integer multiples of $\frac{1}{2}$, that are less than $k = 2$, and, thus, will take values $0, \frac{1}{2}, 1, \frac{3}{2}$ and 2. Specifically,

$$f_2(\omega) = \begin{cases} 0, & 0 \leq f(\omega) < \frac{1}{2} \\ 1/2, & \frac{1}{2} \leq f(\omega) < 1 \\ 1, & 1 \leq f(\omega) < \frac{3}{2} \\ 3/2, & \frac{3}{2} \leq f(\omega) < 2 \\ 2, & 2 \leq f(\omega) \end{cases}$$

In general, we define f_k as :

$$f_k(\omega) = \begin{cases} \frac{j-1}{2^{k-1}}, & \frac{j-1}{2^{k-1}} \leq f(\omega) < \frac{j}{2^{k-1}}, j = 1, ..., k2^{k-1} \\ k, & k \leq f(\omega) \end{cases}$$

Notice that, since f is measurable, the sets $\{\omega : \frac{j-1}{2^{k-1}} \leq f(\omega) < \frac{j}{2^{k-1}}\}$ and $\{\omega : f(\omega) \geq k\}$ are measurable too. Thus, f_k is a measurable function for each $k \in \mathbb{N}$. The construction of the sequence implies that $f_k(\omega) \leq f_{k+1}(\omega)$ for each ω. Also, it holds that :

$$f(\omega) \leq k \Rightarrow |f(\omega) - f_k(\omega)| \leq \frac{1}{2^{k-1}}$$

Suppose that f is finite. Then, as k grows, it will eventually exceed f. Let $f(\omega) \leq M < \infty$ for all $\omega \in \Omega$. For all $k \geq M$, we have that

$$sup_{\omega\in\Omega}|f(\omega) - f_k(\omega)| \leq \frac{1}{2^{k-1}}$$

Thus, f_k converges uniformly to f in Ω.

Otherwise, if $f(\omega) = \infty$, then $f_k(\omega) = k, \forall k$, so, as $k \to \infty$, $f_k(\omega) \to f(\omega)$. Thus, we have pointwise convergence. ■

Remark 4.1.4 The above theorem is an equivalent form of the definition 4.1.4.

An additional theorem claims that measurability is preserved under limits. That is the limit of measurable function is measurable.

Theorem 4.1.5 *If a function f is a limit of measurable functions for all ω, then f is measurable.*

Let f be a measurable function taking non-negative values and consider the sequence f_n of non-negative simple functions that converges monotonically from below to f.

Hence, the sequence $\int_\Omega f_n d\mu$ is non-decreasing and there exists the limit :

$$\lim_{n\to\infty} \int_\Omega f_n d\mu$$

which may be even $+\infty$.

Of course, there is not a single sequence of simple functions f_n that converges to f. But, it can be proven that the value of the above limit is independent of the choice of this sequence. This result is similar to the fact that Riemann sum is independent of the partition of the real line.

Thus, one can define the Lebesgue integral as follows :

Definition 4.1.9 — Lebesgue integral. Let f be a non-negative measurable function and consider the sequence f_n of non-negative simple functions that converges monotonically from below to f. The limit

$$\lim_{n\to\infty} \int_\Omega f_n d\mu := \int_\Omega f d\mu$$

is called the Lebesgue integral of the function f.

The definition can be expanded to all measurable function by introducing the indicator functions, just like in the Riemann integral case :

$$\mathbb{1}_{\{+,-\}}(\omega) = \begin{cases} 1, & f(\omega)\{\geq, <\}0 \\ 0, & otherwise \end{cases} \tag{4.13}$$

and working with $f_+ = f_{\mathbb{1}_+}$ and with $f_- = f_{\mathbb{1}_-}$.

Additionally, if f is a measurable function on $(\Omega, \mathcal{F}, \mu)$ and $E \in \mathcal{F}$, the integral of f over E is

$$\int_E f d\mu = \int_\Omega f_{\mathbb{1}_E} d\mu$$

The mathematical expectation is the same as the Lebesgue integral over a probability space. When ξ is a measurable function and the measure is a probability measure, we refer to the integral as the expectation of the random variable, and denote it by $\mathbb{E}[\xi]$ or, if the context is clear, by $\mathbb{E}\xi$.

Before proceeding to the definition, it is important to see the following result :

Lemma 4.1.6 — Substitution. *Consider the real-valued measurable function f on $(\Omega, \mathcal{F}, \mu)$. Let $\mu f = \int f d\mu$. Let $(M, \mathcal{M})$ be a measurable space and $f : \Omega \to M, g : M \to \mathbb{R}$ be two measurable mappings. Then,*

$$\mu(g \circ f) = (\mu \circ f^{-1})g$$

whenever either side exists.

Proof. Notice that, in the LHS, $g \circ f$ is a function via the composition function, whereas, in the RHS, $\mu \circ f^{-1}$ is a measure, similar to 4.11 The way of proving this lemma is quite typical in measure theory proofs and, thus, we consider it would be usuful for the interested reader. Firstly, we begin with the simpler kind of measurable mappings, the indicator functions. Then, using linearity, we expand the result for simple functions. Afterwards, using the monotone convergence property, we get that it holds for any non-negative measurable functions. Finally, we extend the result to all real-valued measurable functions.

- If g is an indicator function, then the formula is just the definition of $\mu \circ f^{-1}$, since :

$$(\mu \circ f^{-1})g = \int g d(\mu \circ f^{-1}) = \int \mathbb{1}_B d(\mu \circ f^{-1}) = (\mu \circ f^{-1})B$$

$$= \mu(f^{-1}B) = \int_B f d\mu = \int (g \circ f) d\mu = \mu(g \circ f)$$

- Then, we can extend the formula to any measurable $g \geq 0$, thanks to linearity and monotone convergence.
- For general g, it is $\mu|g \circ f| = (\mu \circ f^{-1})|g|$. So, the integrals exist simultaneously. If they do, we get the desired equation by taking differences on both sides.

■

The expectation of a random variable ξ is defined as

$$\mathbb{E}\xi = \int_\Omega \xi d\mathbb{P} = \int_{\mathbb{R}} x(\mathbb{P} \circ \xi^{-1})(dx) \tag{4.14}$$

whenever either integral exists. The second equality follows from the above lemma. The quantity $\mathbb{P} \circ \xi^{-1}$ is the probability distribution of the random variable ξ.

In terms of the cumulative distribution function, we have that :

$$\mathbb{E}\xi = \int_{\mathbb{R}} x dF_\xi(x) \tag{4.15}$$

and when the probability density is well defined :

$$\mathbb{E}\xi = \int_{\mathbb{R}} x f_\xi(x) dx \tag{4.16}$$

Intuitively, the quantity $f_\xi(x)dx$ is the probability that the random variable ξ falls within the interval $[x, x + dx]$.

In general, for any random element ξ is some measurable space M and for an arbitrary function $g : M \to \mathbb{R}$,

$$\mathbb{E}g(\xi) = \int_\Omega g(\xi) d\mathbb{P} = \int_M g(t)(\mathbb{P} \circ \xi^{-1})(dt) = \int_{\mathbb{R}} x(\mathbb{P} \circ (g \circ \xi)^{-1})(dx) \tag{4.17}$$

whenever at least one of the three integrals exists.

For any random variable ξ and constant $p > 0$, the integral $\mathbb{E}|\xi|^p$ is called the $p-$th absolute moment of ξ.

Consider a random variable $\xi \geq 0$. Then :

$$\mathbb{E}\xi^p = \mathbb{E}\int_0^\infty \mathbb{1}\{\xi^p > t\}dt = \mathbb{E}\int_0^\infty \mathbb{1}\{\xi > t^{1/p}\}dt = p\mathbb{E}\int_0^\infty \mathbb{1}\{\xi > t\}t^{p-1}dt$$

Now, using the well-known Fubini's theorem, we can interchange the expectation and the improper integral and get :

$$\mathbb{E}\xi^p = p\int_0^\infty \mathbb{P}[\xi > t]t^{p-1}dt \tag{4.18}$$

This is a useful result connecting the moments of a random variable and the tail probabilities.

We note that for $p = 1$,

$$\mathbb{E}\xi = \int_0^\infty \mathbb{P}[\xi \geq t]dt$$

Another crucially important result follows from the convexity of functions. Recall that a function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be convex if

$$f(px + (1-p)y) \leq pf(x) + (1-p)f(y), x, y \in \mathbb{R}^d, 0 \leq p \leq 1 \tag{4.19}$$

A nice way to see that inequality is the following. Consider a random vector ξ in $\mathbb{R}^d$ with $\mathbb{P}[\xi = x] = p = 1 - \mathbb{P}[\xi = y]$. Then the above result can be written as :

$$f(\mathbb{E}\xi) \leq \mathbb{E}f(\xi)$$

The inequality can be extended to arbitrary integrable random vectors giving the probabilistic Jensen's inequality.

Theorem 4.1.7 — Jensen's Inequality. *Let ξ be an integrable random vector in $\mathbb{R}^d$, and fix any convex function $f : \mathbb{R}^d \to \mathbb{R}$. Then,*

$$f(\mathbb{E}\xi) \leq \mathbb{E}f(\xi) \tag{4.20}$$

Proof. The proof is based on a functional analysis theorem, the Hahn–Banach theorem. The convexity condition is equivalent to the existence of a supporting affine function $h_s = ax + b, \forall s \in \mathbb{R}^d$ s.t.

$$f \geq h_s, f(s) = h_s(s).$$

This can be easily seen geometrically but the formal proof requires a version of Hahn-Banach theorem.

Choosing $s = \mathbb{E}\xi$,

$$\mathbb{E}f(\xi) \geq \mathbb{E}h_s(\xi) = h_s(\mathbb{E}\xi) = f(\mathbb{E}\xi)$$

■

Remark 4.1.8 The Hahn-Banach theorem is an extension theorem for linear functionals. It states that for a linear functional on a subspace Y of a real vector space X, that satisfies $f(y) \leq p(y) \forall y \in Y$, where p is a sublinear functional [a] on X. Then, f has a linear extension $\overline{F}$ from Y to X that satisfies $\overline{F}(x) \leq p(x) \forall x \in X$.

[a] A sublinear functional p on X is a real-valued functional p which is subadditive $p(x+y) \leq p(x)+p(y)$ and positive-homogeneous $p(ax) = ap(x)$

In the following sections, random variables will usually be denoted either by X or by ξ. The expectation of the random variable X will be either be denoted by $\mathbb{E}[X]$ or simply by $\mathbb{E}X$.

Types of measure and decomposition

Let μ be a finite measure on the Borel $\sigma-$algebra of the real line. There are three typical types of measures.

- *Discrete measure :* Suppose that there exists a countable set $C = \{c_1, c_2, ...\}$ (which could also be finite) that is a set of full measure i.e. $\mu((-\infty, +\infty)) = \mu(C)$. Then, μ is a measure of discrete type.
- *Absolutely continuous measure :* The measure μ will be called absolutely continuous if for every set of Lebesgue measure zero $m(A) = 0$, we have also that $\mu(A) = 0$. Thus, the collection of Lebesgue zero sets is a subset of $\mu-$measure zero.
- *Singular continuous measure :* The measure μ will be called singular continuous if $\mu(c) = 0$ for every point $c \in \mathbb{R}$ and there is a Borel set B of Lebesgue measure zero and of full $\mu-$ measure.

Theorem 4.1.9 *Given any finite measure μ on the real line, $\exists$ uniquely a discrete measure μ_{disc}, an absolutely continuous measure μ_{ac} and a singular continuous measure μ_{sc} s.t. for any Borel set A of $\mathbb{R}$, we have that $\mu(A) = \mu_{disc}(A) + \mu_{ac}(A) + \mu_{sc}(A)$.*

4.1.2 $f-$ divergence

One of our most significant goals is to learn distributions, that is, to try to approximate with the best possible estimators the distribution we are looking for. This idea is closely related to the classical concept of distribution learning, that can be informally described as follows :

Remark 4.1.10 *A class of distributions $\mathcal{P}$ is called efficiently learnable if for every $\epsilon > 0, 0 < \delta \leq 1$, given access to an oracle $\mathcal{O}(\mathcal{D})$ for an unknown distribution $\mathcal{D} \in \mathcal{P}$, there exists a polynomial time learning algorithm (of $\mathcal{P}$), that outputs an estimator of a distribution $\hat{\mathcal{D}}$ s.t.*

$$\mathbb{P}[d(\mathcal{D}, \hat{\mathcal{D}}) \leq \epsilon] \geq 1 - \delta$$

Note that the oracle $\mathcal{O}(\mathcal{D})$ is just a mechanism that is able to return a sample from the unknown distribution $\mathcal{D}$.

It is important to obtain some tools in order to answer questions such as :

If $\mathbb{P}, \mathbb{Q}$ are two probability distributions. Are these two distributions close to each other?

Thus, we are obligated to define a notion of distance between two distributions or, in general, a measure of how two distributions differ. In order to define the notion of $f-$ divergence, it is crucial to refer to the concept of Lebesgue decomposition from the field of measure theory.

The Lebesgue Decomposition and the Radon-Nikodym Theorem

Previously, we saw a way to decompose any finite measure on a Borel $\sigma-$ algebra on $\mathbb{R}$. Now, we will try to introduce a decomposition a given measure into two measures that are connected (in some proper way) to another given measure. Let μ, ν be $\sigma-$ definite measures on a measurable space $(\Omega, \mathcal{F})$. We would like to decompose ν with respect to μ, that is to break ν into (two) components that are somehow connected to the measure μ.

Firstly, we will expand the definitions presented in the introductory section about the absolutely continuous and singular types of measure.

Absolute continuity with respect to a measure

Before we proceed to the formal definition, it is useful to introduce an intuitive way to think of absolute continuity. One could think of absolute continuity as a stronger version of continuity and of uniform continuity and thus the class of functions being absolutely continuous as a sub-class of continuous functions. The significance of absolute continuity is that it is the largest class of functions for which the fundamental

theorem of calculus holds (using the classical derivative, and the Lebesgue integral). For instance, a counterexample for continuous functions is the Cantor staircase function.

A function f defined on an interval $I = [a, b]$ of the real line is absolutely continuous if $\forall \epsilon > 0, \exists \delta > 0$, s.t. given N pairwise disjoint sub-intervals $\{(a_i, b_i)\} \subset I$ with $\sum_{i=1}^{N} |b_i - a_i| < \delta$, then $\sum_{i=1}^{N} |f(b_i) - f(a_i)| < \epsilon$. In practice, f is absolutely continuous $\iff$ it has a derivate f' almost everywhere $\iff$ there exists a Lebesgue integrable function g on $I = [a, b]$ s.t. $f(x) = f(a) + \int_a^x g(t)dt \; \forall x \in I$ (then $g = f'$).

Expanding this notion to measures, substituting intuitively sub-intervals with elements of the $\sigma-$algebra and the $\epsilon - \delta$ scheme with the vanishing of measures, we get that :

Definition 4.1.10 — Absolute Continuity. *Let $(\Omega, \mathcal{F})$ be a measurable space with a finite non-negative measure μ. A signed measure $\nu : \mathcal{F} \to \mathbb{R}$ is called absolutely continuous with respect to μ, that is $\nu \ll \mu$ if for every $A \in \mathcal{F}$ for which $\mu(A) = 0$, we get that $\nu(A) = 0$.*

From a set theoretic point of view, we could define the measure's zeros set as $\mathcal{Z}_\mu = \{A \in \mathcal{M} | \mu(A) = 0\}$. Then, using the definition, it follows that $\nu \ll \mu \iff \mathcal{Z}_\mu \subseteq \mathcal{Z}_\nu$.

Remark 4.1.11 An equivalent definition could be given as follows : $\nu \ll \mu$ if $\forall \epsilon > 0, \exists \delta > 0$ s.t. $\mu(A) < \delta \Rightarrow |\nu(A)| < \epsilon$.

Theorem 4.1.12 — Radon-Nikodym Theorem. *Let $(\Omega, \mathcal{F})$ be a measurable space with a finite non-negative measure μ and ν be a signed measure s.t. $\nu \ll \mu$. Then, there is an integrable function f such that, for all $A \in \mathcal{F}$,*

$$\nu(A) = \int_A f d\mu$$

Any two functions that have this property differ on at most a set of $\mu-$measure zero.

The function f is called the density or the Radon-Nikodym derivative of ν with respect to μ. We often write $f = \frac{d\nu}{d\mu} = \frac{d\nu_{ac}}{d\mu}$.

■ **Example 4.6** The normal distribution $\nu = \mathcal{N}(0, 1)$ has the density $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ with respect to the Lebesgue measure $\mu = m$ on $\mathbb{R}$. ■

■ **Example 4.7** When the probability distribution $\mathbb{P} \circ \xi^{-1}$ of the random variable ξ is absolutely continuous, it has a well defined density f_ξ, that is $F_\xi(x) = \int_{-\infty}^{x} f_\xi(t)dt$. ■

Singularity

Definition 4.1.11 — Concentrated Measure. *The measure μ is said to be concentrated on $E \in \mathcal{F}$ if $\mu = \mu^E$, where $\mu^E(A) := \mu(E \cap A)$ for every $A \in \mathcal{F}$.*

Definition 4.1.12 — Singular Measure. *Two measures μ_1 and μ_2 are said to be mutually singular, that is $\mu_1 \perp \mu_2$, if $\exists$ disjoint measurable sets E_1 and E_2 in $\mathcal{F}$ s.t. μ_1 is concentrated on E_1 and μ_2 is concentrated on E_2.*

One can think of singularity as a distribution property. Two distributions are singular if there are two disjoint sets A, B on which the first distribution assigns the whole mass on A and the other on B.

■ **Example 4.8** If $\mu = \mathcal{N}(0,1)$ is the Gaussian distributions and $\nu = Poi(\lambda)$ is the Poisson distribution with parameter λ. Then, we can notice that $\mu(\mathbb{N}_0) = 0$ and $\nu(\mathbb{R} \setminus \mathbb{N}_0) = 0$. Thus, the two measures are singular. ■

Lebesgue decomposition

Theorem 4.1.13 — Lebesgue decomposition of μ with respect to ν.. *Let ν be a signed σ-finite measure and μ an unsigned σ-finite measure on $\mathcal{M}$. Then, there exists a unique pair of real measures ν_{ac} and ν_{sing} on $\mathcal{M}$ s.t.*

$$\nu = \nu_{ac} + \nu_{sing}, \text{ where } \nu_{ac} \ll \mu, \nu_{sing} \perp \mu \tag{4.21}$$

Especially, for distributions-probability measures $\mathbb{P}$ and $\mathbb{Q}$ defined on the same probability space $(\Omega, \mathcal{F})$, the Lebesgue decomposition of $\mathbb{P}$ with respect to $\mathbb{Q}$ is defined as $\mathbb{P} = \mathbb{P}_{ac} + \mathbb{P}_{sing}$, where $\mathbb{P}_{ac} \ll \mathbb{Q}$ and $\mathbb{P}_{sing} \perp \mathbb{Q}$.

It is worth mentioning that $\mathbb{P}_{ac}, \mathbb{P}_{sing}$ are sub-probabilities, that is positive measures with total mass less or equal to 1, and, by definition :

$$\frac{d\mathbb{P}}{d\mathbb{Q}} = \frac{d\mathbb{P}_{ac}}{d\mathbb{Q}}$$

Maximal Slope

Let $f : (0, +\infty) \to \mathbb{R}$ be a convex function s.t, $f(1) = 0$. Thanks to the convexity of f, one can well-define : $f(0) := \lim_{t \downarrow 0^+} f(t) \in \mathbb{R} \cup \{+\infty\}$ and extend the function to $f : [0, +\infty) \to \mathbb{R} \cup \{+\infty\}$.

Consider such a convex function f and a point $x > 0$. Then, as t increases, the slope $\Delta f_x(t) = \frac{f(t)-f(x)}{t-x}$ is non decreasing (by convexity).

Thus, the limit

$$\lim_{t \to \infty} \Delta f_x(t) = sup_{t>0} \Delta f_x(t) \in [0, +\infty]. \tag{4.22}$$

exists and is independent on x. Hence, it equals the maximal slope of f :

$$M_f := \lim_{t\to\infty} \Delta f_1(t) = \lim_{t\to\infty} \frac{f(t)}{t} \in \mathbb{R} \cup \{+\infty\}. \tag{4.23}$$

Setting $t = x + y$, we get the following inequality :

$$\frac{f(x+y) - f(x)}{y} \leq M_f \; \forall x > 0, y > 0. \tag{4.24}$$

or equivalently :

$$f(x+y) \leq f(x) + yM_f \; \forall x > 0, y > 0. \tag{4.25}$$

The convexity of f implies continuity on $(0, \infty)$ and thus the above inequality can be extended, by taking $x \downarrow 0$ and $y \downarrow 0$, giving the next inequality that will be used as lemma for the following definitions :

$$f(x+y) \leq f(x) + yM_f \; \forall x \geq 0, y \geq 0. \tag{4.26}$$

The $f-$divergence functional was introduced by Csiszar as a generalized measure of information.

Definition 4.1.13 — f-divergence. *Let $f : (0, +\infty) \to \mathbb{R}$ be a convex function that vanishes at the point $x = 1$. The $f-$divergence $Div_f(\mathbb{P} \parallel \mathbb{Q})$ between a pair of probability distributions in the same probability space $(\Omega, \mathcal{F})$ is defined as :*

$$Div_f(\mathbb{P} \parallel \mathbb{Q}) = \int_\Omega f(\frac{d\mathbb{P}}{d\mathbb{Q}}) d\mathbb{Q} + \mathbb{P}_{sing}(\Omega) M_f \tag{4.27}$$

For the discrete case, the $f-$divergence on the set of probability distributions $\mathbb{P}^n$ is defined as :

$$Div_f(\vec{p} \parallel \vec{q}) = \sum_{i=1}^{n} q_i f(\frac{p_i}{q_i})$$

for convex functions $f : (0, \infty) \to$ with $f(1) = 0$.

Why do we need $f(1) = 0$?

In order to get the non-negativity property for the f-divergence. Firstly, using the Jensen's inequality for the convex f :

$$\int_\Omega f(\frac{d\mathbb{P}}{d\mathbb{Q}}) d\mathbb{Q} \geq f(\int_\Omega \frac{d\mathbb{P}}{d\mathbb{Q}} d\mathbb{Q}) = f(\mathbb{P}_{ac}(\Omega))$$

Secondly, for the pair $(x, y) = (\mathbb{P}_{ac}(\Omega), \mathbb{P}_{sing}(\Omega))$, from the maximal slope inequality :

$$f(\mathbb{P}_{ac}(\Omega) + \mathbb{P}_{sing}(\Omega)) \leq f(\mathbb{P}_{ac}(\Omega)) + \mathbb{P}_{sing}(\Omega) M_f$$

Using the above results :

$$Div_f(\mathbb{P} \parallel \mathbb{Q}) \geq f(\mathbb{P}_{ac}(\Omega)) + \mathbb{P}_{sing}(\Omega) M_f \geq f(\mathbb{P}_{ac}(\Omega) + \mathbb{P}_{sing}(\Omega)) = f(1) = 0.$$

Thus,

$$Div_f(\mathbb{P} \parallel \mathbb{Q}) \geq 0$$

Property : Joint Convexity

The mapping $(p, q) \mapsto qf(\frac{p}{q})$ is convex on $\mathbb{R}^2_{>0}$:

$$Div_f(\beta\mathbb{P}_1+(1-\beta)\mathbb{P}_2||\beta\mathbb{Q}_1+(1-\beta)\mathbb{Q}_2) \leq \beta Div_f(\mathbb{P}_1||\mathbb{Q}_1)+(1-\beta)Div_f(\mathbb{P}_2||\mathbb{Q}_2), \forall\beta \in [0,1].$$

4.1.3 TV Distance & KL Divergence

The two main metrics on how two distributions differ, that we mostly use in this work, are the Total Variation distance and the Kullback–Leibler divergence. Each one of these can be expressed as a $f-$divergence by choosing an appropriate f function.

Total Variation distance

The first metric is a distance function between a pair of distributions $\mathbb{P}, \mathbb{Q}$ on a $\sigma-$algebra $\mathcal{F}$ of subsets of a space Ω.

Reading the relative literature, there are two different definitions of the total variation distance, a general one and a normalized one. We are going to use different symbols for the two versions.

The general version is given by :

$$||\mathbb{P} - \mathbb{Q}||_{TV} = 2 \sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)| \tag{4.28}$$

Lemma 4.1.14 $||\mathbb{P} - \mathbb{Q}||_{TV} = \sum_{x \in \Omega} |\mathbb{P}(x) - \mathbb{Q}(x)|$

The lemma will be proved below for the normalized version.

Notice that the minimum value of $||\mathbb{P} - \mathbb{Q}||_{TV}$ is obviously 0 and the maximum value is 2. The maximum is achieved when $supp(\mathbb{P}) \cap supp(\mathbb{Q}) = \emptyset$, and, hence the total area equals $1 + 1 = 2$. We remind that $supp(\mathbb{P}) = \{x|\mathbb{P}x > 0\}$ is the support of the probability distribution $\mathbb{P}$.

The normalized version takes values only on $[0, 1]$:

$$d_{TV}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)| \tag{4.29}$$

Lemma 4.1.15 $d_{TV}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2} \sum_{x \in \Omega} |\mathbb{P}(x) - \mathbb{Q}(x)|$

Proof. From the definition of TV distance, we have that $d_{TV}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|$. Suppose that the supremum is attained at the set $A_* \in \mathcal{F}$ and, wlog, let $\mathbb{P}[A_*] \geq \mathbb{Q}[A_*]$.

Note that $d_{TV}(\mathbb{P}, \mathbb{Q}) = \mathbb{P}[A_*] - \mathbb{Q}[A_*] = (1 - \mathbb{P}[A_*^c]) - (1 - \mathbb{Q}[A_*^c]) = \mathbb{Q}[A_*^c] - \mathbb{P}[A_*^c]$.

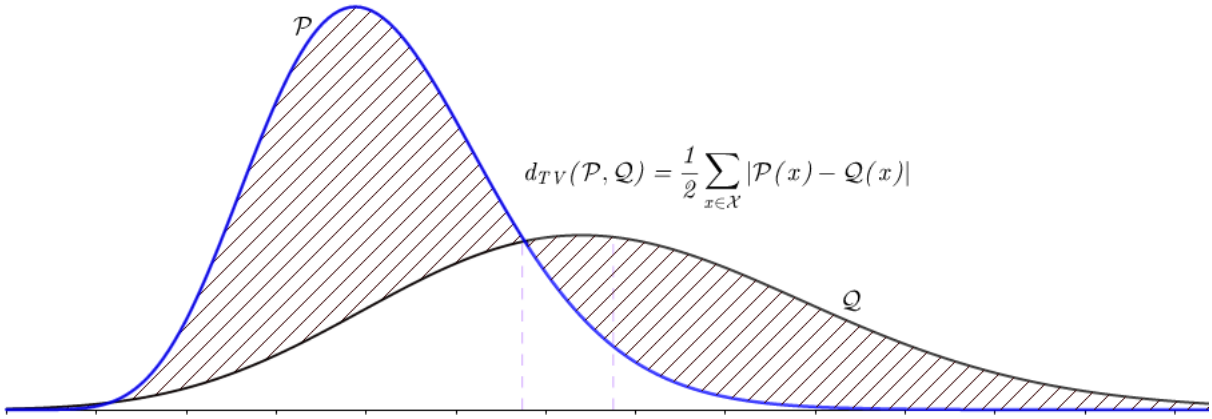

Figure 4.1: TV distance between two probability measures $\mathcal{P}$ and $\mathcal{Q}$.

The space Ω is partioned by A_* and its complement A_*^c. Thus,

$$\frac{1}{2}\sum_{x\in\Omega} |\mathbb{P}(x) - \mathbb{Q}(x)| = \frac{1}{2}\sum_{x\in A_*} |\mathbb{P}(x) - \mathbb{Q}(x)| + \frac{1}{2}\sum_{x\in A_*^c} |\mathbb{P}(x) - \mathbb{Q}(x)| =$$

$$= \frac{1}{2}||\mathbb{P}(A_*) - \mathbb{Q}(A_*)||_1 + \frac{1}{2}||\mathbb{P}(A_*^c) - \mathbb{Q}(A_*^c)||_1 = d_{TV}(\mathbb{P},\mathbb{Q})$$

■

From $f-$ divergence definition, we can get the total variation distance by choosing $f(t) = \frac{1}{2}|t-1|$. Note that $f(1) = 0$ and that f is convex. Intuitively, TV distance equals to the half of l_1 norm of the measures $\mathbb{P}$ and $\mathbb{Q}$. One can think that TV distance is the largest difference of mass assignment of the two measures among all possible subsets of Ω, belonging to the $\sigma-$algebra. It is worth mentioning that d_{TV} is a valid distance metric, that satisfies the three classical distance properties, referred to the d_{KT} section.

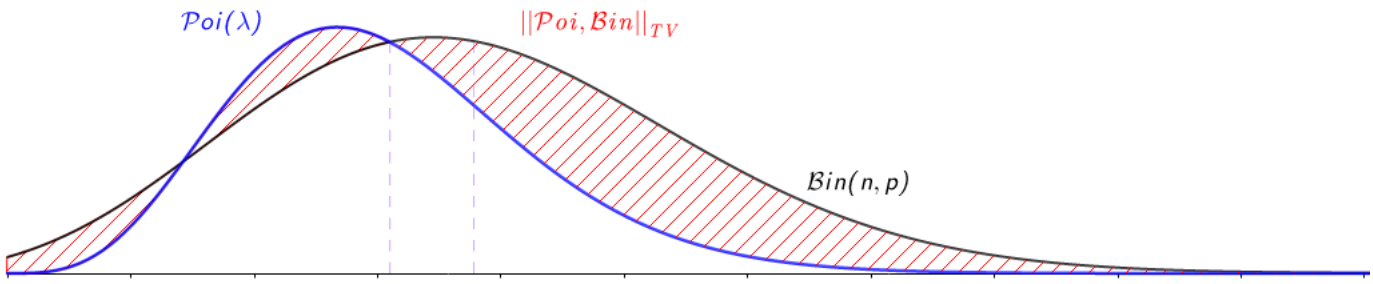

Figure 4.2: TV distance between the Poisson distribution $Poi(\lambda)$ and the Binomial distribution $Bin(n,p)$.

■ **Example 4.9** Up to now, we have defined a way to compute the distance between two probability measures-distributions. We will present an example on how one could upper bound total variation distance in order to provide a Poisson approximation, using the useful technique of coupling. Suppose that π_1 and π_2 are the two projection functions on a space $A \times A$, where :

$$\pi_1(a,b) = a, \quad \pi_2(a,b) = b, \quad (a,b) \in A \times A.$$

Definition 4.1.14 — Coupling. *A coupling of two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on the same probability space $(\Omega, \mathcal{F})$ is any probability measure $\mathbb{C}$ on the product space $(\Omega \times \Omega, \mathcal{F} \otimes \mathcal{F})$ (where $\mathcal{F} \otimes \mathcal{F}$ is the smallest $\sigma-$algebra containing $\mathcal{F} \times \mathcal{F}$) whose marginals are $\mathbb{P}$ and $\mathbb{Q}$:*

$$\mathbb{P} = \mathbb{C} \circ \pi_1^{-1} \quad \textit{and} \quad \mathbb{Q} = \mathbb{C} \circ \pi_2^{-1}$$

It is well known that Poisson distribution $Poi(\lambda)$ with parameter $\lambda > 0$ has probability mass :

$$p_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!}, k \in \mathbb{N}_0 \tag{4.30}$$

Suppose that we have n independent Bernoulli random variables $X_i, i \in [n]$, where $\mathbb{P}[X_i = 1] = p_i$, that is $X_i \sim Be(p_i)$. Consider the sum $X = \sum_{i=1}^n X_i$. It is well known in the literature that, if all the p_i's are small, X is approximately Poisson distributed with parameter $\sum_{i=1}^n p_i$.

We will ask how well X approximates a Poisson distributed random variable. Here is where the total variation distance arises. We will study the TV distance between the distribution of X and the Poisson distribution. The smaller the TV distance is, the better the approximation.

Suppose that $Y \sim Poi(\lambda)$ and fix $k \in \mathbb{N}_0$. So, we have that :

$$\mathbb{P}[X = k] - p_\lambda(k) = \mathbb{P}[X = k] - \mathbb{P}[Y = k]$$

Now, we partition each probability with the complementary events $\{X = Y\}$ and $\{X \neq Y\}$:

$$\mathbb{P}[X = k] - p_\lambda(k) =$$

$$= \mathbb{P}[\{X = k\} \cap \{X = Y\}] + \mathbb{P}[\{X = k\} \cap \{X \neq Y\}] - (\mathbb{P}[\{Y = k\} \cap \{Y = X\}] + \mathbb{P}[\{Y = k\} \cap \{Y \neq X\}])$$

But, when $\{X = Y\}$ holds, the first and the third terms cancel out. Thus :

$$\mathbb{P}[X = k] - p_\lambda(k) = \mathbb{P}[\{X = k\} \cap \{X \neq Y\}] - \mathbb{P}[\{Y = k\} \cap \{Y \neq X\}]$$

Hence, using the definition of TV distance and that $|x - y| \leq |x| + |y|$, we get that :

$$||\mathbb{P}[X \in *] - p_\lambda(*)||_{TV} = \sum_{k \in \mathbb{N}_0} |\mathbb{P}[X = k] - p_\lambda(k)| \leq$$

$$\leq \sum_{k \in \mathbb{N}_0} (\mathbb{P}[\{X = k\} \cap \{X \neq Y\}] + \mathbb{P}[\{Y = k\} \cap \{Y \neq X\}]) = 2\mathbb{P}[X \neq Y]$$

Thus, we have that :

$$||\mathbb{P}[X \in *] - p_\lambda(*)||_{TV} \leq 2\mathbb{P}[X \neq Y] \tag{4.31}$$

Observing the above inequality, we get that in order to have a good approximation, it is crucial to obtain an tight upper bound for $\mathbb{P}[X \neq Y]$. This is where the coupling method gets involved.

Consider n independent random variables (X_i, Y_i) with values on $\mathbb{Z}_2 \times \mathbb{N}_0$ with distribution :

$$\mathbb{P}[(X_i, Y_i) = (x, y)] = \begin{cases} 1 - p_i, & \text{if } x = 0, y = 0 \\ e^{-p_i} - (1 - p_i) & \text{if } x = 1, y = 0 \\ 0, & \text{if } x = 0, y \in \mathbb{N} \\ e^{-p_i} \frac{p_i^y}{y!} & \text{if } x = 1, y \in \mathbb{N} \end{cases}$$

where $i \in [n]$.
Is this a valid coupling? We have to study the marginal distributions.

1. We firstly compute the distribution of X_i's.
$\mathbb{P}[X_i = x] = \sum_{y=0}^{\infty} \mathbb{P}[\{X_i = x\} \cap \{Y_i = y\}]$. For the two values of x :
- $\mathbb{P}[X_i = 0] = (1 - p_i) + 0 + 0 + ... = 1 - p_i$
- $\mathbb{P}[X_i = 1] = (e^{-p_i} - (1 - p_i)) + e^{-p_i}(e^{p_i} - 1) = p_i$

$$\mathbb{P}[X_i = x] = \begin{cases} 1 - p_i, & \text{if } x = 0 \\ p_i & \text{if } x = 1 \end{cases}$$

Thus, $X_i \sim Be(p_i)$.
2. Similarly, we compute the distribution of Y_i's.
$\mathbb{P}[Y_i = y] = \mathbb{P}[\{Y_i = y\} \cap \{X_i = 0\}] + \mathbb{P}[\{Y_i = y\} \cap \{X_i = 1\}] = e^{-p_i} \frac{p_i^y}{y!}$.
Thus, $Y_i \sim Poi(p_i)$.
Now,

$$\mathbb{P}[X \neq Y] = \mathbb{P}[\sum_{i=1}^{n} X_i \neq \sum_{i=1}^{n} Y_i] \leq \mathbb{P}[\exists k \in [n] : X_k \neq Y_k]$$

Using the union bound,

$$\mathbb{P}[X \neq Y] \leq \sum_{i=1}^{n} \mathbb{P}[X_i \neq Y_i] = \sum_{i=1}^{n} (e^{-p_i} - (1 - p_i) + \sum_{y=2}^{\infty} e^{-p_i} \frac{p_i^y}{y!}) = \sum_{i=1}^{n} p_i(1 - e^{-p_i})$$

But, using the convexity of the exponential function (or simply omitting some terms of the Taylor expansion), we have that $e^x \geq 1 + x$ and thus :

$$\mathbb{P}[X \neq Y] \leq \sum_{i=1}^{n} p_i^2$$

Now, set $P = \max_{i \in [n]} p_i$. Then, since :

$$\sum_{i=1}^{n} p_i^2 \leq \sum_{i=1}^{n} (p_i P) = P \sum_{i=1}^{n} p_i = \lambda P$$

we have that :

$$||\mathbb{P}[X \in *] - p_\lambda(*)||_{TV} \leq 2\lambda P$$

■

Remark 4.1.16 We can easily see that when p_i's are small, P will be small too and thus the approximation will be good.

Kullback–Leibler divergence

The second divergence is not a distance function, since it is not symmetric and it violates the triangle inequality. For two discrete probability measures $\mathbb{P}, \mathbb{Q}$:

$$D_{KL}(\mathbb{P} \parallel \mathbb{Q}) = \sum_{x \in \mathcal{X}} p(x) log \frac{p(x)}{q(x)} \tag{4.32}$$

KL divergence is an $f-$divergence metric by choosing $f(x) = xlogx$, that is convex and $f(1) = 0$.

We can expand the RHS sum into two parts :

$$D_{KL}(\mathbb{P} \parallel \mathbb{Q}) = \sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} p(x) \log q(x) = -\sum_{x \in \mathcal{X}} p(x) \log q(x) - (-\sum_{x \in \mathcal{X}} p(x) \log p(x))$$

In the next chapter, we will introduce the notion of entropy in the information theory setting. We will denote by $H(\mathbb{P}) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$ and thus express the KL divergence as a difference of entropies.

Informally, this metric depicts the information gain, that one succeeds if she uses the distribution $\mathbb{Q}$ instead of $\mathbb{P}$.

■ **Example 4.10** For two Bernoulli probability measures with parameters p and q, we have that : $kl(p, q) := D_{KL}(Be(p) \parallel Be(q)) = pln(\frac{p}{q}) + (1-p)ln(\frac{1-p}{1-q})$. ■

Remark 4.1.17

- In order to convert KL divergence to a symmetric measure, we can consider the metric

$$D_{KL}(P, Q) = D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P). \tag{4.33}$$

- Since KL divergence is an appropriate $f-$divergence, from the Jensen inequality, we get the Gibbs inequality :

$$D_{KL}(P \parallel Q) \geq 0 \tag{4.34}$$

Now, that we have presented the total variation distance and the KL divergence, we can present the very useful Pinsker inequality :

$$d_{TV}(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{Q})} \tag{4.35}$$

4.1.4 Concentration Inequalities

Statistical learning is closely related to the notion of concentration of a random variable. Concentration inequalities play a crucial role to the concept of learning, by bounding the deviation of a random variable. The main issue in the topic of concentration inequalities is to study the probability :

$$\mathbb{P}[\xi > t] \text{ or } \mathbb{P}[|\xi - \mathbb{E}\xi| > t]$$

In our learning setting, we will be really interested in proving that our result will not deviate from the expected value. Thus, we would like to having upper bounds for probabilities like the above. Ideally, these bounds should be of exponential order, that is $\mathbb{P}[\xi > t] \leq O(exp(-t))$. So, we prefer bounds for which the probability decreases exponentially, as the deviation grows linearly.

Markov's Inequality

The most trivial, but yet strong, way to bound tails of probabilities is based on the Markov's inequality.

For any nonnegative random variable ξ, and for all $t > 0$, we have the following inequality

$$\xi \geq t \mathbb{1}\{\xi \geq t\}$$

One can see why this is true by considering two cases, one for $\xi \geq t$ and one for $\xi < t$ and the fact that ξ is nonnegative.

By taking the expectations on both sides, and thanks to the linearity of the expectation operator (technically we integrate both sides under the probability distribution measure), we will get the Markov's inequality :

$$\mathbb{E}\xi \geq \mathbb{E}[t \mathbb{1}\{\xi \geq t\}] \Rightarrow \mathbb{E}\xi \geq t(1 \cdot \mathbb{P}[\xi \geq t] + 0) \Rightarrow \mathbb{P}[\xi \geq t] \leq \frac{E\xi}{t}$$

Another way of proving Markov's inequality (in the continuous case) is the following. For $t > 0$,

$$\mathbb{E}\xi = \int_0^\infty \mathbb{P}[\xi \geq y] dy \geq \int_0^t \mathbb{P}[\xi \geq y] dy \geq t\mathbb{P}[\xi \geq t] \Rightarrow \mathbb{P}[\xi \geq t] \leq \frac{E\xi}{t}$$

The second inequality follows from the fact that the tail probability decreases as y grows. Note that $\min_{y \in [0,t]} \mathbb{P}[\xi \geq y] = \mathbb{P}[\xi \geq t]$. The result follows. The discrete case is similar.

Hence,

$$\mathbb{P}[\xi \geq t] \leq \frac{E\xi}{t} \tag{4.36}$$

This upper bound is interesting only if $\mathbb{E}\xi < \infty$, that is when ξ is integrable. Markov's inequality is the easiest concentration inequality we can get. Nevertheless, the upper bound's decrease rate $\frac{1}{t}$ is slow. The reason why this bound is bad, is that we only have information about the first moment (the expectation of the random variable).

For signed random variables, Markov's inequality becomes

$$\mathbb{P}[|\xi| \geq t] \leq \frac{E|\xi|}{t} \tag{4.37}$$

However, it is not difficult to expand this result. We can consider the extended version, that is, if ϕ is a nondecreasing nonnegative function on a interval $I \subset \mathbb{R}$, then for any random variable ξ taking values on $\mathbb{R}$ and real number $t \in I$ with $\phi(t) > 0$:

$$\mathbb{P}(\xi \geq t) = \mathbb{P}(\phi(\xi) \geq \phi(t)) \leq \frac{\mathbb{E}\phi(\xi)}{\phi(t)} \tag{4.38}$$

Now, we can choose appropriate ϕ functions to enhance the upper bound.

Chebyshev's Inequality

If we choose $\phi(x) = x^2$ over $I = (0, \infty)$, we can get Chebyshev's inequality. By replacing ξ with the nonnegative random variable $|\xi - \mathbb{E}\xi|$ and $t > 0$,

$$\mathbb{P}(|\xi - \mathbb{E}\xi| \geq t) \leq \frac{\mathbb{E}[|\xi - \mathbb{E}\xi|^2]}{t^2} = \frac{Var(\xi)}{t^2} \tag{4.39}$$

Chebyshev's inequality is a little better than Markov's inequality but the rate $\frac{1}{t^2}$ remains slow. It is worth seeing that the bound is better since we know more information (the first two moments of the random variable).

More generally taking $\phi(x) = x^m (x \geq 0)$, for any $m > 0$, we have

$$\mathbb{P}(|\xi - \mathbb{E}\xi| \geq t) \leq \frac{\mathbb{E}[|\xi - \mathbb{E}\xi|^m]}{t^m} \tag{4.40}$$

A better choice would be to select a function ϕ that will include all the moments of the random variable.

Chernoff's bounds

Another application of Markov's Inequality is choosing $\phi(x) = e^{sx}$, where $s > 0$. The reason behind this choice is not only the fact that we will get an exponential bound. The reason why we prefer exponential functions is because mathematicians back then were interested in studying sums of independent random variables $Z = \sum_{i=1}^{n} X_i$. Thus, exponential function behave well with sums since they convert them to products.

For any random variable X and any $t > 0$, we have :

$$\mathbb{P}[\xi \geq t] = \mathbb{P}[e^{s\xi} \geq e^{st}] \leq \frac{\mathbb{E}e^{s\xi}}{e^{st}} \tag{4.41}$$

where $M_\xi(s) = \mathbb{E}[e^{s\xi}]$ is the moment generating function of the random variable ξ.

The importance of the moment generating function is crucial. It captures all the moments of the random variable and thus all the information contained in it. It can be seen as equivalent to the notion of a Taylor series expansion. The Taylor series of a infinitely differentiable complex-valued function f at c is the power series :

$$\sum_{i=0}^{\infty} \frac{f^{(i)}(c)}{i!}(x-c)^i = f(c) + \frac{f'(c)}{1!}(x-c) + \frac{f''(c)}{2!}(x-c)^2 + ... \tag{4.42}$$

This expansion contains all the information of the function f. So, just like all the secrets of a function are hidden in its derivatives, the information of a random variable is hidden in its moments.

$$M_\xi(t) = \mathbb{E}e^{t\xi} = 1 + t\mathbb{E}\xi + \frac{t^2}{2!}\mathbb{E}[\xi^2] + ... \tag{4.43}$$

■ **Example 4.11 — or why Normal distribution is so important.** Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then $Z = X - \mu \sim \mathcal{N}(0, \sigma^2)$. We have :

$$M_Z(t) = \int_{-\infty}^{\infty} e^{tz} f_Z(z)dz = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z^2}{2\sigma^2}} dz$$

Notice that :

$$-\frac{1}{2}(\frac{z}{\sigma} - \sigma t)^2 = -\frac{z^2}{2\sigma^2} + zt - \frac{\sigma^2 t^2}{2}$$

So :

$$e^{tz} e^{-\frac{z^2}{2\sigma^2}} = e^{-\frac{1}{2}(\frac{z}{\sigma} - \sigma t)^2} e^{\frac{\sigma^2 t^2}{2}}$$

Thus,

$$M_Z(t) = e^{\frac{\sigma^2 t^2}{2}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{z}{\sigma} - \sigma t)^2} dz$$

But the expression under the integral is the probability density function of the distribution $\mathcal{N}(\sigma^2 t, \sigma^2)$, which integrates to 1.

Hence,

$$M_Z(t) = e^{\frac{\sigma^2 t^2}{2}}$$

■

Thus, we can see that the form of the moment generating function is exponential and completely similar to the density function. This reveals the importance and the value of a normal random variable. Using the characteristic function, we can see that

a normal random variable is the identity of the 'Fourier' frequency transformation and, this crucial remark is used in the proof of the celebrated Central Limit Theorem.

Another crucial application is the following :

Theorem 4.1.18 *Let X and Y be two random variables s.t. $M_X(t) = M_Y(t) \forall t \in (-\delta, \delta)$ for some $\delta > 0$. Then $X =_D Y$, that is, they have the same distribution.*

This reveals the intuition that the moment generating function contains all the information a random variables hides.

Returning to the Chernoff's bound, in order to optimize the upper bound, we need to find $s > 0$ that minimizes the RHS of the inequality. Thus,

$$\mathbb{P}[X \geq t] \leq \sup_{s>0} \frac{\mathbb{E}[e^{sX}]}{e^{st}} \tag{4.44}$$

■ **Example 4.12 — Sums of independent random variables.** Consider the random variable $Z = X_1 + ... + X_n$, where X_i are i.i.d. real-valued random variables.

Then, using the Chernoff's bound, for $s > 0$, we get that :

$$\mathbb{P}[Z > t] \leq e^{-st}\mathbb{E}e^{sZ}$$

Consider the logarithm of the moment generating function as

$$\psi_Z(s) = log\mathbb{E}e^{sZ}, s > 0.$$

Then,

$$\mathbb{E}e^{sZ} = \mathbb{E}e^{s\sum_i X_i} = \prod_{i=1}^{n} \mathbb{E}e^{sX_i}$$

So, if we denote by

$$\psi_X(s) = log\mathbb{E}e^{sX_i}, i \in [n]$$

then,

$$\psi_Z(s) = \sum_{i=1}^{n} log\mathbb{E}e^{sX_i} = \sum_{i=1}^{n} \psi_{X_i}(s) =_{iid} n\psi_X(s)$$

For the random variable Z, define :

$$\psi_Z^*(t) = \sup_{s>0}(st - \psi_Z(s))$$

in order to obtain the optimal Chernoff bound :

$$\mathbb{P}[Z > t] \leq e^{-\psi_Z(t)}$$

The upper bound is determined by the distribution of the i.i.d. random variables X_i. ■

Hoeffding's Inequality

Hoeffding's inequality is one of the most important techniques, as far as concentration inequalities are concerned, in learning theory. It is used for bounding the probability that sums of bounded random variables deviate from their expected mean. As we have already seen, we are interested in random variable that have tail probabilities decreasing exponentially as the tail grows linearly.

Chernoff bounds are closely related to the moment generating function. We have seen that the moment generating function of a centered normal random variable Y with variance v is $M_Y(t) = \mathbb{E}e^{tY} = exp(\frac{t^2v}{2})$. Thus, we need to focus on random variables whose moment generating function behaves similarly to Gaussian random variables.

This idea provides some intuition behind the definition of sub-Gaussian random variables. Hence, we formalize the concept of sub-Gaussian random variables by setting :

Definition 4.1.15 — Sub-Gaussian r.v.. *Let $\psi_X(t) = log\mathbb{E}[e^{tX}]$. A centered random variable X, that is $\mathbb{E}X = 0$, is said to be sub-Gaussian with variance factor v if :*

$$\psi_X(t) \leq \frac{t^2v}{2} \forall t \in \mathbb{R} \tag{4.45}$$

Let $\mathcal{G}(v)$ be the class of sub-Gaussian random variables, parameterized by the variance v.

From Chernoff bounding, we have seen how important it is to upper bound the moment generating function. This is exactly what motivates us to define the above class of random variables exactly like that. A random variable X will belong to the class $\mathcal{G}(v)$ if its moment generating function is dominated by the moment generating function of a centered normal random variable Y with variance v, that is $Y \sim \mathcal{N}(0, v)$.

Remark 4.1.19 If $\{X_i\}_{i=1}^n$ are independent with $X_i \in \mathcal{G}(v_i)$, then the sum $\sum_{i=1}^n X_i \in \mathcal{G}(\sum_{i=1}^n v_i)$.

In the next lemma, we establish that bounded random variables belong to appropriate sub-Gaussian classes.

Lemma 4.1.20 — Hoeffding's Lemma. *Let X be a centered random variable, taking values in a bounded interval $[a, b]$. Then, $\psi''_X(t) \leq \frac{(b-a)^2}{4} = v$ and $X \in \mathcal{G}(v)$.*

Proof. Since the random variable X lives in the interval $[a, b]$, one gets that :

$$a \leq X \leq b \Rightarrow \frac{a-b}{2} \leq X - \frac{a+b}{2} \leq \frac{b-a}{2} \Rightarrow |X - \frac{a+b}{2}| \leq \frac{b-a}{2}$$

Hence,

$$Var(X) = Var(X - \frac{a+b}{2}) \leq (\frac{b-a}{2})^2 = \frac{(b-a)^2}{4}$$

We have to upper bound the second derivate of the logarithm of the moment generating function of the bounded random variable X.

$$\psi''_X(t) = \frac{d^2(log\mathbb{E}e^{tX})}{dt^2} = e^{-\psi_X(t)}\mathbb{E}[X^2e^{tX}] - e^{-2\psi_X(t)}(\mathbb{E}[Xe^{tX}])^2$$

Notice that the RHS looks like a variance of some random variable. Let $X \sim \mathcal{P}$ and let $\mathcal{P}_t$ be the probability distribution with density

$$x \to e^{-\psi_X(t)} \cdot xe^{tx}$$

with respect to $\mathcal{P}$.

The density is well defined since

$$\int_{[a,b]} e^{-\psi_X(t)}xe^{tx}dx = e^{-\psi_X(t)}\int_{[a,b]} xe^{tx}dx = e^{-\psi_X(t)}\mathbb{E}e^{tX} = e^{-\psi_X(t)}e^{log\mathbb{E}e^{tX}} = 1.$$

Then, $\mathcal{P}_t$ remains concentrated on $[a, b]$ and, thus, the random variable $Y \sim \mathcal{P}_t$ has a variance upper bounded by $\frac{(b-a)^2}{4}$.

Now, observe that :

$$\mathbb{E}Y = \int_{[a,b]} xe^{-\psi_X(t)}xe^{tx}dx = e^{-\psi_X(t)}\mathbb{E}[Xe^{tX}]$$

and

$$\mathbb{E}[Y^2] = \int_{[a,b]} x^2e^{-\psi_X(t)}xe^{tx}dx = e^{-\psi_X(t)}\mathbb{E}[X^2e^{tX}]$$

Hence, we get that :

$$\psi''_X(t) = \mathbb{E}[Y^2] - (\mathbb{E}Y)^2 = Var(Y) \leq \frac{(b-a)^2}{4}$$

The fact that X belongs to the sub-Gaussian class $\mathcal{G}(\frac{(b-a)^2}{4})$ follows from Taylor's expansion theorem since $\psi_X(0) = \psi'_X(0) = 0$. Specifically, there exists a $\xi \in [0, t]$ such that :

$$\psi_X(t) = \psi_X(0) + \psi'_X(0)t + \psi_X(\xi)\frac{t^2}{2} \leq \frac{t^2(b-a)^2}{8}.$$

■

Now, we are able to deduce the Hoeffding's inequality. Consider n independent random variables $X_1, ..., X_n$ where X_i takes values in a bounded interval $[a_i, b_i]$.

Then, for $S = \sum_{i=1}^n (X_i - \mathbb{E}X_i)$, we know that :

$$\psi_S(t) = \sum_{i=1}^n log\mathbb{E}e^{t(X_i - \mathbb{E}X_i)}$$

By the independence condition and the boundness assumption, one gets, using the Hoeffding's lemma,

$$\psi_S(t) \leq \frac{t^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2$$

Now, one can take a simple Chernoff's bound and get :

$$\mathbb{P}[S \geq \zeta] = \mathbb{P}[e^{\lambda \sum_{i=1}^{n} (X_i - \mathbb{E}X_i)} \geq e^{\lambda \zeta}] \leq e^{-\lambda \zeta} e^{\frac{\lambda^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2} = exp(-\lambda \zeta + \frac{\lambda^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2)$$

Now, optimizing on the parameter $\lambda > 0$, one gets :

$$\lambda^* = \frac{4\zeta}{\sum_{i=1}^{n} (b_i - a_i)^2}$$

Thus, we get the following fundamental inequality :

Theorem 4.1.21 — Hoeffding's Inequality. *Let $X_1, ..., X_n$ be independent bounded random variables s.t. $\mathbb{P}[X_i \in [a_i, b_i]] = 1$. Let $S_n = \sum_{i=1}^{n} X_i$. Then for any $\zeta > 0$, we have that :*

$$\mathbb{P}[S_n - \mathbb{E}S_n \geq \zeta] \leq exp(\frac{-2\zeta^2}{\sum_{i=1}^{n} (b_i - a_i)^2})$$

and

$$\mathbb{P}[S_n - \mathbb{E}S_n \leq -\zeta] \leq exp(\frac{-2\zeta^2}{\sum_{i=1}^{n} (b_i - a_i)^2})$$

By mixing these two inequalities, one gets :

$$\mathbb{P}[|S_n - \mathbb{E}S_n| \geq \zeta] \leq 2exp(\frac{-2\zeta^2}{\sum_{i=1}^{n} (b_i - a_i)^2)})$$

Note that one could use the empirical mean $\bar{X} = \frac{X_1 + ... + X_n}{n}$, where the random variables X_i are strictly bounded in $[a_i, b_i]$, and get :

$$\mathbb{P}[\overline{X} - \mathbb{E}\overline{X} \geq \zeta] \leq exp(-\frac{2n^2\zeta^2}{\sum_{i=1}^{n} (b_i - a_i)^2}) \tag{4.46}$$

We will often use this inequality to obtain good sample complexity bounds for our learning problems. A broad collection of other concentration inequalities can be found in [BS16].

5. Mathematical Foundations III : Information Theory

5.1 Information Theory

Statistical learning theory is based on the idea of discovering knowledge using statistics and functional analysis. The discovery of this hidden knowledge is done through the classical procedure of sampling (for instance, in our problem, we try to discover a hidden permutation via noisy samples). Samples offer information. Often it is important to question whether a new sample offers information. Thus, it is crucial to define a way to measure how much information a sample provides. This measure is provided through the field of information theory.

The field of information theory lies in the intersection of mathematics, computer science and statistics. Concepts like entropy, mutual information, codes and sufficient statistics are broadly studied and applied in statistical learning theory and, hence, they will be presented in the following sections.

5.1.1 Entropy

The information theoretic notion of entropy was introduced by Claude Shannon in his classic paper *'A Mathematical Theory of Communication'* [Sha48] in 1948. Entropy was firstly appeared in the field of statistical thermodynamics through the works of Ludwig Boltzmann (1872) and of J.Willard Gibbs (1878).

They considered a collection of classical particles, a *system*, with a discrete set of microstates X and, for each microstate $i \in X$, with energy E_i, a corresponding probability p_i, that is the probability the system occupy that specific microscopic configuration during thermal fluctuations.

Specifically, Gibbs defined the measure of entropy as :

$$S = -k_B \sum_{i \in X} p_i ln p_i$$

where k_B denoted the Boltzmann's constant. In each step, the system is distributed over a set of $|X|$ microstates, each one with probability p_i and energy E_i. Adding heat to the system, its thermodynamic entropy increases because it increases the number of possible microscopic states that it could be in, thus making any complete state description longer. This observation is important, since as we will see, entropy is strongly connected to the length of the description of a random source, in the information theoretic setting that follows.

In the field of information theory, entropy was introduced by Shannon for a discrete random variable X with range $\mathcal{X}$ and probability mass function $p(X)$ as :

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)} = \sum_x \mathbb{P}[X = x] \log_2 \frac{1}{\mathbb{P}[X = x]} = \mathbb{E}[-\log_2 p(X)] \quad (5.1)$$

where x takes values in the essential range of X (that is to say, those values of X for which $\mathbb{P}[X = x] > 0$).

Shannon expressed the notion of information that is contained in a discrete source via a functional that quantifies the uncertainty of this discrete random variable. The essence of uncertainty is hidden inside the probability mass function and the mass it assigns to the possible output values of the random variable. Shannon proposed the following function to measure the information of each event A in a discrete source :

$$I(A) = log(\frac{1}{P(A)}) \quad (5.2)$$

Intuitively, the notion of entropy gives a way to measure the uncertainty of a random variable, the amount of information it carries. Thus, the smaller the value of the entropy, the more a priori information one has for the random variable. At the same time, entropy is linked to idea of the amount of 'space' one needs to store the information of a random variable. In classical computer science, data are stored in bits. Entropy preserves this notion. Entropy is expressed in bits when one uses the logarithm to the base 2 and in nats when one uses the natural logarithm. In the information theory literature, the logarithm to base 2 is often used to define entropy, rather than the natural logarithm, in which case $H(X)$ can be interpreted as the number of bits needed to describe X on the average.

■ **Example 5.1** Some examples follow in order to understand better Shannon's entropy.

- A discrete random variable taking uniformly M different values has entropy $log_2 M$.
- A fair coin can be seen as a random variable, taking two possible values (Heads or Tails) with equal probability. Hence, its entropy is 1 bit.
- When throwing N fair coins, the number of all possible outcomes is 2^N and so the entropy is N bits.
- A random variable $X \sim Be(p)$ has entropy $H_X = H(p) = -plog_2 p - (1 - p)log_2(1 - p)$. The entropy vanishes when p equals 0 or 1 (since the uncertainty 'vanishes') and is maximal when $p = \frac{1}{2}$ (since the uncertainty is maximal).

■

Usually, we are interested in the joint entropy $H(X,Y)$ of the random variable (X,Y). That is :

$$H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X) \tag{5.3}$$

where the conditional entropy $H(X|Y)$ is defined as :

$$H(X|Y) = \sum_y \mathbb{P}[Y=y]H(X|Y=y) \tag{5.4}$$

where $H(X|Y=y) = \sum_x \mathbb{P}[X=x|Y=y] log_2 \frac{1}{\mathbb{P}[X=x|Y=y]}$ and for the above formulae y is ranging over the essential range of Y and x is running over the essential range of X conditioned to $Y=y$.

Using Jensen's inequality and the concativity of $x \mapsto xlog\frac{1}{x}$, we get that :

$$H(X|Y) = \sum_y \mathbb{P}[Y=y]H(X|Y=y) = \sum_y \mathbb{P}[Y=y] \sum_x \mathbb{P}[X=x|Y=y] log \frac{1}{\mathbb{P}[X=x|Y=y]}$$

By Jensen's inequality,

$$H(X|Y) \leq \sum_x (\sum_y \mathbb{P}[X=x|Y=y]\mathbb{P}[Y=y]) log \frac{1}{\sum_y \mathbb{P}[X=x|Y=y]\mathbb{P}[Y=y]}$$

The RHS can be written as :

$$\sum_x \mathbb{P}[X=x] log \frac{1}{\mathbb{P}[X=x]} = H(X)$$

Hence,

$$H(X|Y) \leq H(X) \tag{5.5}$$

This is intuitively obvious since information (knowledge of Y) decreases the uncertainty and consequently decreases the entropy of X. The conditional entropy $H(X|Y)$ is a measure of the amount of new information carried by X, given that we already know the value of Y. When does equality hold?

From the above inequality, we conclude the subadditivity of entropy :

$$H(X,Y) \leq H(X) + H(Y) \tag{5.6}$$

■ **Example 5.2** As we saw in the previous chapter, KL divergence can be expressed in terms of entropy. Specifically,

$$D_{KL}(\mathbb{P} \parallel \mathbb{Q}) = -\sum_{x \in X} p(x) log q(x) + \sum_{x \in X} p(x) log p(x)$$

In the RHS, the second term is simply the negative entropy of the measure $\mathbb{P}$.

The first term is called the cross entropy of the distributions $\mathbb{P}$ and $\mathbb{Q}$, that is :

$$H(\mathbb{P} \parallel \mathbb{Q}) = \mathbb{E}_{\mathbb{P}}[log\frac{1}{\mathbb{Q}}] = \sum_{x \in X} p(x) log\frac{1}{q(x)}$$

It is important not to confuse the notion of cross entropy with that of joint entropy. Cross entropy is defined as :

$$H(\mathbb{P} \parallel \mathbb{Q}) = H(\mathbb{P}) + D_{KL}(\mathbb{P} \parallel \mathbb{Q})$$

and thus, $D_{KL}(\mathbb{P} \parallel \mathbb{Q}) = H(\mathbb{P} \parallel \mathbb{Q}) - H(\mathbb{P})$.

KL divergence is usually called *information gain* achieved if the distribution $\mathbb{Q}$ is used instead of $\mathbb{P}$.

From Gibb's inequality, we have that :

$$H(\mathbb{P} \parallel \mathbb{Q}) \geq H(\mathbb{P})$$

This is obvious from a information theoretic point of view. The expected number of bits required to code samples from distribution from $\mathbb{P}$ using a code optimized for $\mathbb{Q}$ is larger than the number of bits required to code samples from distribution from $\mathbb{P}$ using a code optimized for $\mathbb{P}$.

■

We define the notion of mutual information between two discrete random variables by :

$$I(X,Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (5.7)$$

Hence, $I(X,Y) \geq 0$. We could consider $I(X,Y)$ as a measure of the extend to which X, Y are not independent. It expresses the amount of information that we get for the one random variable by observing the other one. Thus, it is a symmetric measure.

Remark 5.1.1 In the literature, the sums $\sum_{x \in X}$ usually do not run over the essential range of the random variable X but in the whole range of X, including those x s.t. $\mathbb{P}[X = x] = 0$. Thus, it is, in general, a common knowledge that these terms offer 0 to the sum since $\lim_{a \downarrow 0^+} a \log a = 0$.

How one could deduce Shannon's entropy formula?

We would like to define a useful measure for quantifying the information that we gain by observing an event of probability p. Let $I(p)$ be that information measure. $I(p)$ should satisfy the following properties :

- Information is non-negative : $\boxed{I(p) \geq 0}$ (1)
- Events with probability $p = 1$, provide zero information : $\boxed{I(1) = 0}$ (2)
- Two independent events, whose joint probability is the product of the two measures, the information gained is the sum of the information measures : $\boxed{I(p_1 \cdot p_2) = I(p_1) + I(p_2)}$ (3)

- The measure is continuous with respect to p and is decreasing as p increases (since the more probable the event is, the less information is provided) : $I(p) \downarrow p$

Property (3) offers the intuition that the information measure will have a logarithmic structure. Specifically :

- By (3), $I(p^2) = I(p * p) = 2I(p)$.
- Inductively, $I(p^n) = nI(p)$.
- $I(p) = I((p^{\frac{1}{m}})^m) = mI(p^{\frac{1}{m}}) \Rightarrow I(p^{\frac{1}{m}}) = \frac{1}{m}I(p)$.
- Hence, $I(p^{\frac{n}{m}}) = \frac{n}{m}I(p)$.
- Since the measure is continuous, for $0 < p \leq 1$ and $r \in \mathbb{R}_{>0}$, $I(p^r) = rI(p)$
- Thus, for some base b, $I(p) = log_b(\frac{1}{p}) = -log_b(p)$.

Coding Theory

Another interesting way to link entropy with the term $-\sum p_i log p_i$ is via convex optimization and coding theory using Kraft's inequality. But, firstly, we need to introduce the fundamentals of Coding theory.

Let X be a random variable with range $\mathcal{X}$ and let $\mathcal{D}$ be a $D-$ary alphabet. Without loss of generality, suppose that $\mathcal{D} = \{0, 1, ..., D-1\}$. Also, let $\mathcal{D}^*$ be the set of finite-length strings of symbols from $\mathcal{D}$.

Definition 5.1.1 — Source Code. *A source code for the random variable X to be a mapping C from the range $\mathcal{X}$ of X to $\mathcal{D}^*$.*

For instance, if $\mathcal{D} = \{0, 1\}$ and $\mathcal{X} = \{green\}$, then $C(green) = 01$ with length $l(green) = 2$.

Definition 5.1.2 — Expected Length. *For a source code C of a random variable X with probability mass p the expected length $L(C)$ is given by*

$$L(C) = \mathbb{E}l(X) = \sum_{x \in \mathcal{X}} p(x)l(x). \tag{5.8}$$

Suppose that Alice wants to send Bob an encoded stream of her color preferences. Let $\mathcal{D} = \{0, 1\}$ and $\mathcal{X} = \{green, red, blue, purple\}$.

At first, Alice uses the following encoding $C_1 : \forall x \in \mathcal{X}, C_1(x) = 0$.

It is obvious that Bob, given a symbol 0, cannot decode it in a clear way. This code is said to be singular. Hence, Alice's description-encoding is ambiguous. Non singularity suffices for an unambiguous encoding of the range of the random variable.

Definition 5.1.3 — Non singular Code. *A code is said to be non singular if every element of the range of X maps into a different string in $\mathcal{D}^*$. That is $x \neq x' \Rightarrow C(x) \neq C(x')$.*

Alice wants to send Bob a stream of her color preferences. Thus, the code C needs

to be extended to a code $C^* : \mathcal{X}^* \to \mathcal{D}^*$, that is, for $x = x_1...x_n \in \mathcal{X}^*$:

$$C^*(x_1...x_n) = C(x_1)...C(x_n), \tag{5.9}$$

where $C(x_1)...C(x_n)$ denotes codewords concatenation.

Then, Alice uses the following code C_2 that maps $\{green, red, blue, purple\} \mapsto \{0, 010, 01, 10\}$. This code is non singular since each element of the range $\mathcal{X}$ is encoded using a different codeword. Suppose that Alice sends the stream 010. Then Bob cannot understand if Alice sent 'red' or 'green,purple' or 'blue,green'. This is because the code is not uniquely decodable. In a uniquely decodable code one has only one possible source string generating it.

Definition 5.1.4 — Uniquely Decodable Code. *A code is called uniquely decodable if its extension is non singular.*

Alice uses the following code C_3 mapping $\{green, red, blue, purple\} \mapsto \{10, 00, 11, 110\}$. This code is uniquely decodable.

Suppose that Alice streams the sequence 1100. Then Bob, after seeing the whole sequence will deduce that Alice sent the message 'blue,red'. This sequence cannot be decoded in any other valid way. But, suppose that Bob wanted to parse the stream and do not need to look at the entire string to determine the codewords. Then, this code would fail since Bob will read '11' and then he cannot decide if this means 'blue' or he has to proceed to the next character and read 'purple'. The problem is that the codeword 11 is prefix for the codeword 110. Only after reading the whole stream, he will be totally sure what the unique decoding is.

Thus, in order Bob's desire to be fulfilled, Alice needs to design a new code.

Definition 5.1.5 — Instantaneous Code. *A code is called an instantaneous code (or prefix code) if no codeword is a prefix of any other codeword.*

Finally, Alice uses the following code C_4 that maps $\{green, red, blue, purple\} \mapsto \{0, 10, 110, 111\}$. This code is instantaneous and Bob can decode each symbol as soon as he has read the whole codeword corresponding to it. Now, whatever sequence Alice streams, it will be easily and instantly decoded by Bob.

In general :

$$\text{Instantaneous} \subset \text{Uniquely Decodable} \subset \text{Non Singular} \subset \text{All codes}$$

Kraft's inequality

Consider the problem of finding the instantaneous code with the minimum expected length. Equivalently, we should find the lengths $l_1, ..., l_n$ that satisfy Kraft's inequality and whose expected length is less than any other's instantaneous code.

Theorem 5.1.2 — Kraft's inequality. *For any instantaneous code over a $D-$ary alphabet, the codeword lengths $l_1, l_2, ..., l_n$ must satisfy the inequality :*

$$\sum_{i=1}^{n} D^{-l_i} \leq 1 \tag{5.10}$$

Conversely, given a set of codeword lengths that satisfy this inequality, there exists an instantaneous code with these word lengths.

The theorem's proof can be found in the classical information theory book [CT06].

■ **Example 5.3** Now, we will see how entropy arises from the minimization of the expected length of any prefix code. Suppose that we want to find the instantaneous code over a $D-$ary alphabet with the minimum expected length.

Optimal Prefix Code
Input : *A set of codeword lengths $\{l_i\}$ of size n.*
Output : $L^* :=$ minimum expected length

Thus, the optimization problem can be expressed as follows.

$$(\text{Primal}) : \min_{(l_1,...,l_n)\in\mathbb{Z}^n_{\geq 0}} \sum_{i=1}^{n} l_i p_i$$

$$s.t. \sum_{i=1}^{n} D^{-l_i} \leq 1$$

This is a constraint optimization problem. Using the Lagrange multipliers, we can work on the minimization of the Lagrangian function :

$$J = \sum_{i=1}^{n} l_i p_i - \lambda(1 - \sum_{i=1}^{n} D^{-l_i})$$

Differentiating with respect to l_i, we get that :

$$\frac{\partial J}{\partial l_i} = p_i - \lambda D^{-l_i} lnD$$

Thanks to the convexity of the problem, we can set the derivative to 0 and obtain

$$D^{-l_i} = \frac{p_i}{\lambda lnD}$$

Thus, the constraint now gives for λ :

$$\lambda = \frac{1}{lnD} \text{and } p_i = D^{-l_i}.$$

So, the optimal expected length is :

$$L^* = \sum p_i l_i^* = \sum p_i log_D \frac{1}{p_i} = H_D(X)$$

■

Remark 5.1.3 Since l_i should be integers, in some case a rounding technique will be needed that will round the lengths near the non-integer optimal lengths set.

Thus, the entropy controls the optimal codeword length of the prefix code.

5.1.2 Sufficient statistics

We continue by presenting in short the notion of sufficient statistics, that was introduced by Sir Ronald Fisher in 1920.

Let $\{X_i\}_{i=1}^n$ be samples of a distribution $\mathcal{D}$ with an unknown parameter θ.

We say that the statistic $Y = u(X_1, ..., X_n)$ is *sufficient* if the probability $\mathbb{P}[X_1 = x_1, ..., X_n = x_n | Y = y]$ does not depend on the parameter θ.

That is, if one knows the sufficient statistic, there is no other function of the samples that could offer more information for the unknown parameter.

Theorem 5.1.4 — Fisher–Neyman factorization. *Let f_θ be the density function of a distribution with unknown parameter θ. Then, the statistic T is sufficient for the parameter θ iff there are non-negative functions h, g s.t. $f_\theta(x) = h(x)g_\theta(T(x))$.*

5.1.3 Fano's Inequality

Fano's inequality is a popular information-theoretical result that provides a lower bound on worst-case error probabilities in multiple-hypotheses testing problems. Multiple variants of Fano's inequality have been derived in the literature. In this thesis, we will use the following version.

Let $(\Omega, \mathcal{F})$ be a measurable space, and $\mathcal{D}(\Omega)$ be the set of all probability distributions on it. Consider the set $\mathcal{A}_m = \{A : \mathcal{X}^n \to \mathbb{R}^{\mathcal{X}}\}$ is the set of deterministic learning algorithms A that take m samples and output a hypothesis distribution D_A.

Let $\mathcal{F} \subseteq \Delta(\Omega)$ be a family of distributions and assume that we have access to m i.i.d. samples drawn from a distribution $x = (x_1, ..., x_m) \sim D^m \in \mathcal{F}$. Let $\tilde{D}$ be an estimator of D, given the m samples and define the risk of the estimator $\tilde{D}$:

$$\mathcal{R}_m(\tilde{D}, \mathcal{F}) = \sup_{D \in \mathcal{F}} \mathbb{E}_{x \sim D^m}[d_{TV}(\tilde{D}, D)]$$

We will introduce the notion of the minimax risk of a family of distributions $\mathcal{F} \subseteq \Delta(\Omega)$ and $m > 0$:

$$R_m(\mathcal{F}) = \inf_{A \in \mathcal{A}_m} \mathcal{R}_m(\tilde{D_A}, \mathcal{F}) = \inf_{A \in \mathcal{A}_m} \sup_{D \in \mathcal{F}} \mathbb{E}_{x \sim \mathcal{D}^n}[d_{TV}(\tilde{D}_A, D)]$$

$R_m(\mathcal{F})$ represents the minimum expected error of any $m-$samples learning algorithm A when running on the worst possible target distribution from the class $\mathcal{F}$.

The following result is due to Yu.

Theorem 5.1.5 — Fano's inequality. *Let $\mathcal{F}$ be a finite family of densities s.t.*

$$1. \quad \inf_{f,g\in\mathcal{F},f\neq g} d_{TV}(f,g) \geq a$$

$$2. \quad \sup_{f,g\in\mathcal{F},f\neq g} D_{KL}(f \parallel g) \leq b$$

then it holds that :

$$R_m(\mathcal{F}) \geq \frac{a}{2}(1 - \frac{mb + ln2}{ln|\mathcal{F}|})$$

Thus, in order to lower bound the minimax risk of a family of distributions, and essentially the total variation distance, we just need to consider a finite family of densities and bound their TV distance and their KL divergence from below and from above respectively. Variations of the Fano's inequality and techniques in deriving Fano-type inequalities can be found in [SGS17].

6. On Voting & Social Choice Theory

6.1 Foundations of Voting Theory

For further understanding the problem we want to deal with, it is crucial to discover the foundations of modern voting theory. Thus, in order create the links between that domain and our problem concerning preference learning will begin by answering some crucial questions that could easily arise.

How is our problem related with Voting Theory and Social Choice Theory?

Our goal is to discover a hidden true ordering among m alternatives from a set A, given a collection of samples, where each sample is a permutation of the elements of A. One could easily think of a voting process the exact same way. Each voter proposes her own perspective on how she believes this hidden global preference list is ordered and, afterwards, an appropriate voting rule is applied, that aggregates each vote (noisy sample) and outputs the final result, a global 'socially-acceptable' ranking. Hence, intuition and tools offered by the voting theory field will be useful to our analysis.

Reformulating the problem.

There is a true hidden preference among m objects $\{a_i\}_{i=1}^m$, that is expressed by a permutation over these alternatives $a_{i_1} \succ a_{i_2} \succ ... \succ a_{i_m}$. For instance, these m objects can be candidates of an election process and the hidden ordering could be the 'socially-acceptable' ranking. Each voter offers in the election process a noisy sample, her own point of view considering the true ordering of the m alternatives. We play the role of a voting rule that has to output a ranking of the m alternatives.

What are the points of interest of the social choice field?

Social choice area is interested in the aggregation of preferences-perspectives in order to output a 'common decision', a 'social preference'. The need of techniques and results from that field appears other closely related areas, such as economics, management and voting systems.

How dit it all start?

One of the pioneers of that field, and specifically in the forefront of the application of mathematics in the area of social sciences, was the French mathematician and philosopher Marquis de Condorcet. In 1785, Condorcer published his work *'Essay on the Application of Analysis to the Probability of Majority Decisions'.* In that work, he mentions two crucial results, that are nowadays known as the *Condorcet's Paradox* and the *Condorcet's Jury Theorem.*

Condorcet's paradox. Consider a voting procedure with two candidates, where each voter has a preference to one of them. If the society, as a whole, wants to choose in common one of the two candidates, based on social acceptance, the majority voting rule seems to be a logical and correct choice. Condorcet questioned the following :

What happens if there are more than two candidates? Is the majority voting still a good choice?

Condorcet proposed the following (counter)example : We denote by $a \succ_i b$, when the voter i prefers a over b. Consider a setting with three candidates a, b, c and three voters with the following preferences :

- $a \succ_1 b \succ_1 c$
- $b \succ_2 c \succ_2 a$
- $c \succ_3 a \succ_3 b$

We can easily observe that the majority of the voters prefers a over b, b over c and c over a. Hence, the socially acceptable preference, according to the majority rule is $a \succ b \succ c \succ a$, which is obviously inconsistent. Whoever candidate will get elected, there will be a majority of citizens that will disagree with that choice. Equivalently, the graph generated by this ranking will have a cycle.

Hence, Condorcet deduced that the majority voting rule is a valuable technique for social decision making, thank to its simplicity, but it deals with a considerable amount of issues. Consequently, Condorcet clarified the necessity of designing methods in the field of voting and social choice theory, that will encounter issues like the one mentioned above.

Condorcet's Jury Theorem. Consider a group of juries, that is called to decide if the defendant is innocent or guilty. Suppose that each member of the group has a common and independent probability $p \in (\frac{1}{2}, 1)$ of making the right choice. Then, the majority of juries is more likely to make the correct choice than each jury individually. Additionally, as the size of the group increases, the probability of

making the right choice tends to one. In a mathematical perspective, the probability is expressed as a sum of binomial random variables :

$$Maj(p,n) = \sum_{i=\lfloor n/2 \rfloor + 1}^{n} \binom{n}{i} p^i (1-p)^{n-i} \longrightarrow 1, \text{ as } n \to \infty$$

$Maj(p,n)$ expresses the probability that the majority of the n juries makes the right choice, when each member decides correctly with probability p. Under these circumstances, the majority rule works well. On the other side, if $p \in [0, \frac{1}{2}]$, the results get reversed and the best choice would be to choose a jury at random and judge according to the decision of the randomly chosen jury.

Is there a fair voting method?

One of the most important results concerning the fairness of voting methods is the famous Arrow's impossibility theorem. The economist Kenneth Arrow demonstrated the theorem in his doctoral thesis and popularized it in his 1951 book 'Social Choice and Individual Values'.

No voting method is fair, every ranked voting method is flawed.
The only voting method that isn't flawed is a dictatorship.

Arrow states that, when the number of alternatives is grater that 3, there is no ranked voting electoral system that, given the ranked preferences of voters, can deduce a community-wide ranking (complete and transitive) which will be 'fair' in a sense of satisfying a set of logical criteria.

The voting setting will be formally presented in the following section. Nevertheless, we are going to present the main setting here in order to refer to Arrow's impossibility theorem.

We will consider a set of n voters and a set of alternatives A (the candidates). We denote the set of permutations on A with $\mathcal{L}(A)$. It is worth mentioning that if $\succ \in \mathcal{L}(A)$, then $\succ$ is transitive ($a \succ b, b \succ c \Rightarrow a \succ c$) and antisymmetric ($a \succ b \wedge b \succ a \Rightarrow a = b$). Hence, $\succ$ is a total order on A. The preference of voter j is denoted by $\succ_j$ and j prefers a over b if $a \succ_j b$.

A voting method or voting rule is just a function that aggregates the preferences of all voters into a total social order on the alternatives. This function $f : \mathcal{L}(A)^n \to \mathcal{L}(A)$ is usually called a *social welfare function.*

From our experience in our everyday lives, it seems natural to think that such a function, in order to be 'good', should satisfy some criteria.

- ☐ *A 'good' voting method f should satisfy unanimity* : If each voter has the same identical preference ranking $\succ^*$, then the socially acceptable order should be that exact ranking. Formally, unanimity can be stated as :

$$\forall \succ^* \in \mathcal{L}(A), f(\succ^*, ..., \succ^*) = \succ^*$$

- ☐ *A 'good' voting method f should not be a dictatorship* : A voter is a dictator when the final resulting preference is independent of the other $n-1$ voters and only the dictator chooses the social preference. Specifically, a voter j is a dictator in f if for all voter preferences $\succ_1, ..., \succ_n \in \mathcal{L}(A), f(\succ_1, ..., \succ_n) = \succ_j$. Hence, f is not a dictatorship if no dictator exists.
- ☐ *A 'good' voting method f should satisfy independence of irrelevant alternatives.* That is the social preference between any pair of alternatives $x, y \in A$, depends only on the preferences expressed by the voters' only between these two candidates. Hence, $\forall x, y \in A, \forall \succ_1, ..., \succ_n, \succ_1^*, ..., \succ_n^* \in \mathcal{L}(A)$ where $f(\succ_1, ..., \succ_n) = \succ, f(\succ_1^*, ..., \succ_n^*) = \succ^*$, we have that if :

$$(x \succ_j y \iff x \succ_j^* y \ \forall j) \Rightarrow (x \succ y \iff x \succ^* y)$$

  A simpler way to develop an intuition with that rule is the following question : *Why should a voter's preferences about candidate $z \neq x, y$, influence the social ordering between candidates x and y?*

These three rules would be crucial to hold in order to have a 'good' voting method in common sense. Arrows' theorem clarifies that it is impossible for a voting method with more than 3 candidates to satisfy these three rules at the same time.

Theorem 6.1.1 — Arrow's Theorem. *Every social welfare function over a set of at least 3 alternatives ($|A| \geq 3$) that satisfies unanimity and independence of irrelevant alternatives is a dictatorship.*

Arrow's theorem states that for any non trivial voting procedure, there is no votes aggregation algorithm that can output a ranking that will successfully aggregate the individual voting preferences of each voter to a common socially optimal order of preferences, without violating a collection of axioms. These criteria, presented above, correspond to some properties that a 'good' voting scheme should satisfy. [NN07]

In conclusion, one could think that the impossibility theorem of Arrow could be a terminating point to the field of social choice and of voting theory. This is exactly the point where computer science and statistics arise in order to expand the framework of social choice theory. In order to deal with the weaknesses of ballots and voting rules, one could think that voting rules act like estimators. In this modern voting scheme, we consider that there is a hidden global truth, that each voter is coping to estimate. Thus, each vote corresponds to a noisy version of that underlying truth. Hence, voting rules, trying to output a social acceptable ranking, are connected to the notion of maximum likelihood estimator. This idea will be presented in the following section. But, before proceeding to the statistical perspective of voting rules, we present the main framework of statistical voting theory.

6.2 Statistical Foundations of Virtual Social Choice

After revisiting the foundations and the main results of voting theory, we are able to introduce a voting setting with a statistical point of view. This setting will be closely connected to our ranking learning framework. One could think of these two setting

in parallel, since our ranking models can be seen completely as voting procedures the way defined as follows.

6.2.1 Voting Setting

The main setting of voting theory consists of whom we vote (alternatives) and how we vote (votes).

Definition 6.2.1 — Alternatives. *A set $A = \{a_1, ..., a_m\}$ of alternatives - options, that voters want to rank.*

A vote is just a ranking among the alternatives.

Definition 6.2.2 — Vote. *The vote of each voter is a bijective function $\sigma : A \to \{1, 2, ..., m\}$, that is each vote is a permutation of the elements of A.*

For alternatives $a, b \in A$, if $\sigma(a) < \sigma(b)$, then the voter prefers a over b. This is denoted by $a \succ_\sigma b$.

Definition 6.2.3 — Voting Profile. *The set of all votes, that is the set of all possible bijective functions (permutations) σ, is denoted by $\mathcal{L}(A)$. A voting profile of n votes is denoted by $\pi \in \mathcal{L}(A)^n$.*

6.2.2 Voting Rules

We can think of a voting rule as a function that takes as input a vote profile, that is a list of the preferences of voters, and outputs a ranking. These rules can output either deterministically or randomly.

Definition 6.2.4 — Deterministic voting rule of n votes. *A deterministic voting rule of n votes is a function, that takes as input a voting profile of n votes and outputs a winner vote.*

$$r_n^{Det} : \mathcal{L}(A)^n \to \mathcal{L}(A) \tag{6.1}$$

Collecting all the n-votes deterministic voting rules for all $n \in \mathbb{N}$, we can define the deterministic voting rule as the union of all the n-votes deterministic voting rules.

Definition 6.2.5 — Deterministic voting rule. *A deterministic voting rule is a function*

$$r^{Det} : \cup_{n \geq 1} \mathcal{L}(A)^n \to \mathcal{L}(A), \tag{6.2}$$

which operates on a vote profile and outputs a ranking.

Note that we define the voting rule to output a ranking over alternatives rather than a single alternative; such functions are also known as social welfare functions in the literature.

The models that we are going to introduce in the following chapters are probabilistic, so it is unavoidable not to introduce a probabilistic notion of a voting rule.

Definition 6.2.6 — Randomized voting rule. *A randomized voting rule is a function*

$$r^{Rand} : \cup_{n\geq 1}\mathcal{L}(A)^n \to D(\mathcal{L}(A)), \tag{6.3}$$

where D is the set of all distributions over an outcome space.

Given a profile π, the probability of a (randomized) rule r to return a ranking σ is denoted by $\mathbb{P}[r(\pi) = \sigma]$.

Thinking of voting rules as estimators

In 1959, John Kemeny developped the following rule in order to answer to the question

If I am given a preference of each voter on a set of alternatives, what is the 'socially-wide' acceptable preference?

Suppose that we are given a set of m rankings. We have to choose a permutation in the symmetric group that will represent the socially acceptable order of preferences. If one thinks of each permutation as a point on a metric space, we have to choose the point that minimizes the distance between the m given points. But, as we saw in the introductory chapter, we can define appropriate distance metrics in order to convert $\mathbb{S}_n$ into a metric space $(\mathbb{S}_n, d_*)$, where d_* is a valid distance metric on the symmetric groups. From now on, we will work with the Kendall's tau distance.

Definition 6.2.7 — Kemeny's Rule. *Given a voting profile - voting vector $\vec{\sigma} = (\sigma_1, ..., \sigma_n) \in \mathcal{L}(A)^n$, Kemeny's rule choose the ranking τ that is the closer under Kendall-Tau distance to the n given votes, that is :*

$$\tau = arg \min_{\tau\in\mathcal{L}(A)} \sum_{i=1}^{n} d_{KT}(\tau, \sigma_i) \tag{6.4}$$

During 1980-90, Peyton Young developped a technique for the study of preferences aggregation. Specifically, he considered that there exists a 'true' but 'hidden' ranking-preference among the alternatives. We get noisy signals-samples from that true and locked ranking. Young, using a probabilistic model on that exact idea, proved that the Kemeny rule is the maximum likelihood estimator of the true ranking given i.i.d. noisy samples generated by the model. Thus, in the literature, Kemeny's rule is usually referred as the Kemeny-Young method.

If we observe the above equation, it should be clear that we want to minimize the l_1 norm of the given elements on the metric space $(\mathbb{S}_m, d_{KT})$ of the permutations in $\mathbb{S}_m$ with the Kendall's tau metric. In the following lemma, we are going to show that the choice of minimizing the l_1 norm is equivalent to the choice of finding the median of the elements of the metric space.

Lemma 6.2.1 — l_1 minimization. *Given the points $p_1, ..., p_n \in \mathbb{R}$, the l_1 norm defined as $l_1(x) = \sum_{i=1}^{n} ||x - p_i||_1$ is minimzed by the median of the given points.*

Proof.
- Let $\epsilon > 0$ and let $x \in [p_k, p_{k+1}]$. Then, $p_k \leq x \leq x + \epsilon \leq p_{k+1}$.
- The l_1 norm's sum becomes

$$l_1(x) = \sum_{i=1}^{n} ||x - p_i||_1 = \sum_{i=1}^{k} (x - p_i) + \sum_{i=k+1}^{n} (p_i - x)$$

- We transpose x to the point $x + \epsilon$ and we observe how the sum is changed :

$$l_1(x+\epsilon) = \sum_{i=1}^{n} ||(x+\epsilon)-p_i||_1 = \sum_{i=1}^{k} (x+\epsilon-p_i) + \sum_{i=k+1}^{n} (p_i-x-\epsilon) = \epsilon(k-n+k)+l_1(x)$$

- Hence, the discrete difference equals to $\frac{l_1(x+\epsilon)-l_1(x)}{\epsilon} = 2k - n$.
- We let $\epsilon \downarrow 0$, and we get, for $x \in [p_k, p_{k+1}]$, $l_1'(x) = 2(k - \frac{n}{2})$.
- The monotonicity of the l_1 norm can be simply derived : For $k < \frac{n}{2}$, the function is decreasing, for $k = \frac{n}{2}$ is constant and for $k > \frac{n}{2}$, is an increasing function of k.
- Thus, the minimum is attained at $k = \frac{n}{2}$ and hence the element chosen corresponds to the median of the collection of the given points.

■

Remark 6.2.2 It is useful to note how Kemeny's rule works when the solution is not unique. In that case, the set $T = arg\min_{\tau} \sum_{i=1}^{n} d_{KT}(\tau, \sigma_i)$ will have more than one element and, hence, the rule chooses uniformly at random an element from T.

An algorithm for computing a ranking according to the Kemeny rule in polynomial time in the number of candidates is not known, and unlikely to exist since the problem is NP-hard even if there are just 4 voters.

An *election* (V, C) consists of a set V of n votes and a set C of m candidates. The score of a ranking σ with respect to election (V, C) is defined as $\sum_{v \in V} d_{KT}(v, \sigma)$. A permutation σ^* that attains the minimum score is usually called *Kemeny consensus* of (V, C) and the corresponding score $\sum_{v \in V} d_{KT}(v, \sigma^*)$ is called the *Kemeny score* of (V, C). The problem is defined as follows :

KEMENY SCORE
Input : An election (V, C) and a positive integer k.
Question : Is the Kemeny score of $(V, C) < k$?

This consensus ranking problem is known to be NP-hard ([BT89]). From a graph theoretic point of view, the NP-hardness is expressed as follows : The election can be seen as a tournament problem. The tournament requires each alternative to play

every other of the $n-1$ alternatives, but there is one more round (set of pairwise contests) to be played. We ask whether there exists a set of outcomes for the final round that will guarantee tournament victory for a particular competitor.

Tournament Outcome
Instance : A simple clique K_n where each edge can be either directed (i beats j) or undirected (the pairwise winner between i and j is not decided) and a distinguished player x.
Question : Is there a way of assigning directions to the undirected edges so that x wins the tournament?

The Tournament Outcome under the second-order Copeland is NP-complete and the reduction is via the celebrated NP-complete problem 3,4-SAT.

As we will see in the following chapter, when the rankings are i.i.d. samples from a special noisy model, namely the Mallows distribution, consensus ranking arises during the computation of the maximum likelihood ranking.

7. On Probabilistic Models of Permutations

7.1 Prelude

Data ranking appears in a wide variety of applications, as we have already referred to the introductory chapter, but remains too difficult to model, learn from, and predict. Working with ranking data creates significant computational challenges that stem from the structural complexity of the symmetric group $\mathbb{S}_n$, the space of permutations on n elements. Models of ranking data are mainly parametric families of distributions in the symmetric group.

There are many distributional models of rankings that have been developed in order to explain choice behavior. Two of the more popular in the machine learning community are the Mallows model and the Plackett-Luce model. The Mallows model is a distance-based ranking model and was firstly introduced by C.L. Mallows in 1957 in his paper 'Non-Null Ranking Models'. The Plackett-Luce distribution derives its name from the independent work by Plackett (1975) and Luce (1959).

Despite the fact that these two seminal ranking models were first develloped in the 20th century, a probabilistic perspective of preferences was studied two centuries earlier (1785) by Condorcet, while he was questioning the issue of political decision making. Thus, in order to link the past with the present, we choose to present the first recorded attempt to draw a ranking sample.

7.2 Condorcet's Decision Problem

In the previous chapter, we presented a review of some fundamental results from the theory of voting and social choice. One of the most influential people, working on that field of science, was the French mathematician and philosopher Marquis of Condorcet. During his life, Condorcet was questioning constantly the political decisions, that were deviating from the social benefit. In 1785, Condorcet worked on a probabilistic view of making the 'right decisions', where one chooses from a set of

policies and deduces a ranking that maximizes social welfare.

He considered a set of choices, from which the members of the society, the voters, express their opinion, their preference, that is a ranking over this set of choices. He considered that there is an underlying objective ranking that is the most beneficial to the society. Condorcet's model has the following properties :

- Each player votes independently from each other.
- The comparison between any pair of alternatives is independent.
- If, in the objective ranking, the choice a is preferred over choice b, that is $a \succ b$, then a voter ranks a over b with probability $1 - p > \frac{1}{2}$, (and hence the error probability $p < \frac{1}{2}$).

It is easy to think of the ranking generation process as a directed graph, whose vertices are the set of alternatives A and, in each step, we add a directed edge between alternatives a, b. The direction of the edge is determined by the error probability $p < \frac{1}{2}$. If the objective ranking is π_0 and for a, b, we have that $a \succ_{\pi_0} b$, then we add the edge $a \to b$, with probability $1 - p$, otherwise we add the edge $b \to a$.

Equivalently, we can think that the directed graph is initially the tournament graph induced by the objective ranking π. Afterwards, we iterate over each edge and with probability p, we flip the direction of that edge.

As we will show the Condorcet's noise model corresponds to the Mallows model, defined later. Thus, we will refer to the above procedure as the Condorcet-Mallows noisy ranking process, that is decribed as follows :

Algorithm 2 Condorcet-Mallows noisy ranking process

1. Let π_0 be the objective ranking and let $0 \leq p < \frac{1}{2}$.
2. **Initialization :** $\sigma \leftarrow \emptyset$.
3. For each pair of alternatives $a, b \in A$, s.t. $a \succ_{\pi_0} b$,
  3a. with probability $1 - p$, add $a \succ b$ to σ,
  3b. otherwise, add $b \succ a$ to σ.
if σ *is intransitive* **then**
  GOTO step (*2*).
else
  RETURN σ.
end

It is clear that the way we generate the ranking σ, there is a significant probability that the generated directed graph will have a cycle and, thus, the permutation will be intransitive. But, this is unacceptable. We cannot output, for instance, the permutation $a \succ b \succ c \succ a$. Hence, we have to restart the generative process.

Now, we will try to deduce the Mallows probabilistic model by analyzing the Condorcet-Mallows process.

The Condorcet-Mallows process independently decides for each pair $a, b \in A$ by flipping a random $p-$biased coin. Thus, with probability p, the objective ranking π_0 and the generated π will have a pairwise disagreement on a, b and with probability $1 - p$, they will agree on that pair.

Consider the probability measure $\mathbb{P}_{CM}$ over rankings σ :

$$\mathbb{P}_{CM}[\sigma|\pi_0, p] = \frac{1}{Z_{CM}} \prod_{a \succ_{\pi_0} b} \begin{cases} p & \text{if } \sigma, \pi_0 \text{ disagree on } a, b \\ 1-p & \text{otherwise} \end{cases} \tag{7.1}$$

The probability that one generates, using the Condorcet-Mallows process, a ranking σ, given π_0, p is exactly equal to the $\mathbb{P}_{CM}[\sigma|\pi_0, p]$.

We note that Z_{CM} is a normalization constant that will be computed later.

We have already described a notion of distance between permutations that counts the number of pairwise disagreements. Thus, the Kendall's tau distance naturally arises to the context of Mallows model.

Set the number of alternatives to be $|A| = m$. It is already known that the number of pairwise agreements equals $\binom{|A|}{2} - d$, where d equals the number of pairwise disagreements. The number $\binom{m}{2}$ equals the number of edges of the clique K_m.

The measure $\mathbb{P}_{CM}$ can be expressed as follows :

$$\mathbb{P}_{CM}[\sigma|\pi_0, p] = \frac{1}{Z_{CM}} p^{d_{KT}(\sigma,\pi_0)} (1-p)^{\binom{m}{2} - d_{KT}(\sigma,\pi_0)} = \frac{1}{Z_{CM}} (1-p)^{\binom{m}{2}} (\frac{p}{1-p})^{d_{KT}(\sigma,\pi_0)} \tag{7.2}$$

Set $\phi = \frac{p}{1-p}$. Since $0 \leq p < \frac{1}{2}$, we get that $0 \leq \phi < 1$. In the Mallows model, we will set the normalization constant Z_{MM} to be :

$$\frac{1}{Z_{MM}} = \frac{1}{Z_{CM}} (1-p)^{\binom{m}{2}}$$

It can be shown that :

$$Z_{MM} = 1 \cdot (1+\phi) \cdot ... \cdot (1+\phi+...+\phi^{m-1}) \tag{7.3}$$

and thus :

$$Z_{CM} = (1-p)^{\binom{m}{2}} (1+\frac{p}{1-p})(1+\frac{p}{1-p}+(\frac{p}{1-p})^2)...(1+\frac{p}{1-p}+...+(\frac{p}{1-p})^{m-1}) \tag{7.4}$$

7.3 The Mallows Model

The importance of the Mallows Model for permutations is equivalent to the importance of the normal distribution on the real line. In order to understand the notion of this noisy model, it is worth reminding the reader the problem we want to solve.

Main Problem: There is a true hidden preference among m objects $\{a_i\}_{i=1}^m$, that is expressed by a permutation over these alternatives $a_{i_1} \succ a_{i_2} \succ ... \succ a_{i_m}$. Our goal is to learn this true hidden ordering, given noisy samples in the sense that each sample is an element of the symmetric group generated by these m elements. Thus, each samples is one of the $m!$ possible ranking and is drawn by a distribution parameterized by the true hidden ranking.

How many samples will be needed in order to learn the true hidden ranking with high probability?

This question is of significant importance but we firstly need to clarify our setting. Hence, the most useful question is the following :

What is the distribution of our model?

7.3.1 The Mallows model $\mathcal{M}(\pi_0, \phi)$

The simplest form for the Mallows model is the $\mathcal{M}(\pi_0, \phi)$. Let π_0 be the 'underlying truth', the hidden central ranking parameter of the Mallows models and let $\phi \in [0, 1]$ be the spread parameter or the dispersion of the model. One could think of ϕ as the swapping probability of each pair of adjacent elements. The higher the value of the dispersion, the more noisy are the samples generated by the model, since swaps are more likely to occur in the initial central ranking. The lower the value of ϕ, the more stable our samples will be, since the sampling procedure will not cause, with high probability, vaste of swaps.

In the previous section, we showed that the probability of drawing a ranking π of size n, given the true order π_0 is proportional to :

$$(1-p)^{\binom{n}{2}-d_{KT}(\pi,\pi_0)}p^{d_{KT}(\pi,\pi_0)}$$

Thus, using a standard normalization and setting $\phi = \frac{p}{1-p} < 1$, we get that :

$$\mathbb{P}[\pi|\pi_0]^1 = \frac{1}{Z(\phi,\pi_0)}\phi^{d_{KT}(\pi,\pi_0)} \tag{7.5}$$

where $d_{KT} : \mathbb{S}_n \times \mathbb{S}_n \to \mathbb{Z}_{\geq 0}$ is the known Kendall's tau ranking distance

$$d_{KT}(\pi,\sigma) = \sum_{1\leq i<j\leq n} \mathbb{1}\{(\pi(i)-\pi(j))(\sigma(i)-\sigma(j)) < 0)\} \tag{7.6}$$

The probability $\mathbb{P}[\pi|\pi_0]$ corresponds to the probability of drawing π as sample from a Mallows model $\mathcal{M}(\pi_0, \phi)$. Equivalently, in the corresponding literature, we use a parameter β instead of ϕ such that $\phi = e^{-\beta}$. Thus :

$$\mathbb{P}[\pi|\pi_0] = \frac{1}{Z(\phi,\pi_0)}e^{-\beta d_{KT}(\pi,\pi_0)} \tag{7.7}$$

In this way, it is easier to see the connection between the Mallows model and the normal distribution.

The normal distribution $\mathcal{N}(\mu, \sigma^2)$ has density

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{|x-\mu|^2}{2\sigma^2}}$$

[1]The correct notation would be $\mathbb{P}[\pi|\pi_0, \phi]$, but for simplicity reasons, we omit ϕ. The majority of the related literature makes this assumption too.

The support of the above density is $\mathbb{R}$. It is not difficult to notice the similarities between the two probability measures. The normal distribution assigns mass to each $x \in \mathbb{R}$, that decreases exponentially to a distance measure $x \to \frac{|x-\mu|^2}{2\sigma^2}$ with center the mean μ, whereas the Mallows measure assigns mass that decreases exponentially with the Kendall Tau distance from the central ranking π_0. One can think of the Mallows distribution as an embedding of the normal distribution to the symmetric group. This dimension reduction via the embedding alters some properties of the normal distribution, like the symmetry property and the existence of a turning point.

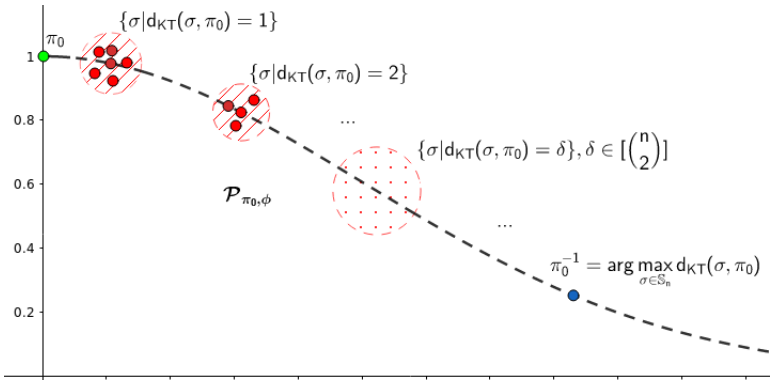

Figure 7.1: Informally, the Mallows model can be seen as a discrete version of an one-sided normal distribution. Intuitively, each point in the discrete x-axis is a set $S_d = \{\sigma|d_{KT}(\sigma,\pi_0)=\delta\}$ for $\delta = \{0\} \cup [\binom{n}{2}]$.

The more a permutation deviates from the central ranking in d_{KT}, the less is the probability of being chosen. In the same notion, the more a value deviates from the mean of the normal distribution, the probability of being drawn falls exponentially in the square of the distance.

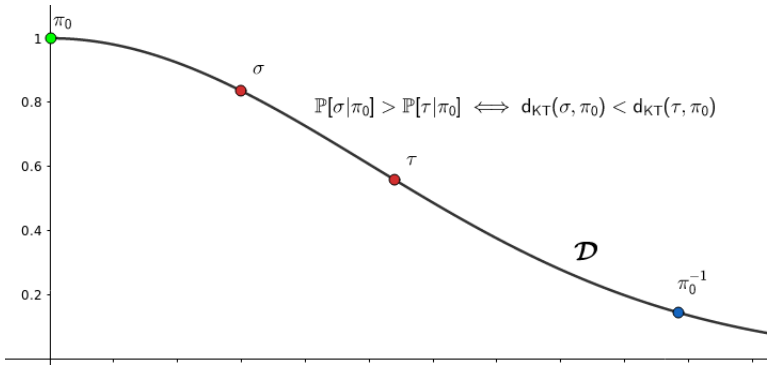

Figure 7.2: A distribution $\mathcal{D}$ that follows the monotonicity property. Such an example is the Mallows measure, where one can observe an exponential decay as the KT distance grows.

This is an important characteristic of the Mallows model. We refer to that property as monotonicity. Mallows model is monotonic since, given two samples σ, τ, one

has :

$$d_{KT}(\sigma,\pi_0) < d_{KT}(\tau,\pi_0) \iff \mathbb{P}[\sigma|\pi_0] > \mathbb{P}[\tau|\pi_0]$$

We remind that KT distance counts the number of pairwise disagreements of a pair of permutations. Considering a fixed ranking π, KT distance is maximized by comparing π with its inverse permutation $\pi^{-1} = arg\max_{\sigma} d_{KT}(\pi,\sigma)$ and the value attained is $\binom{n}{2}$. This value is the maximum number of swaps needed when one sorts with bubblesort.

The normalization constant is :

$$Z(\phi,\pi_0) = Z(\phi) = \prod_{i=1}^{n-1}\sum_{j=0}^{i}\phi^j \tag{7.8}$$

This expression was firstly observed in the Mahonian numbers $M(n,k)$ section, as this function is the generating function $\sum_{k=0}^{\infty} M(n,k)x^k$.

Lemma 7.3.1 $\mathrm{Z}(\phi,\pi_0) = Z(\phi) = \prod_{i=1}^{n-1}\sum_{j=0}^{i}\phi^j$

Proof. Informally, since we can express the normalization constant as,

$$1(1+\phi)(1+\phi+\phi^2)\dots(1+\phi+...+\phi^{n-1}) = \sum_{k=0}^{\binom{n}{2}} M(n,k)\phi^k$$

one can easily notice that summing the Mallows probability measure over all possible permutations of size n will produce the exact same sum and this sum will be equal to :

$$1 = \sum_{\pi\in\mathbb{S}_n}\mathbb{P}[\pi|\pi_0] \Rightarrow Z(\phi,\pi_0) = Z_\phi = \sum_{k=0}^{\binom{n}{2}} M(n,k)\phi^k$$

Notice that k runs over the possible number of inversions needed. We now proceed to a more formal proof.

We remind the reader the decomposition vector defined in Chapter 2 for the KT distance. There is '1-1' correspondence between every permutation $\sigma\in\mathbb{S}_n$ and the vector of numbers $(V_1(\sigma,\pi_0),...,V_n(\sigma,\pi_0))$, where $V_j(\sigma,\pi_0)\in[0,j-1]$. Let $\Omega_n^k = [k]\times[k+1]\times...\times[n-1]$. This '1-1' correspondence allows us to write the partition function $Z(\phi)$ in the following way

$$Z(\phi) = \sum_{y\in\Omega_n^0}\prod_{j=1}^{n}\phi^{y_j} = \sum_{y_1\in[0]}\phi^{y_1}\Big(\sum_{y\in\Omega_n^1}\prod_{j=2}^{n}\phi^{y_j}\Big) = \Big(\sum_{y_1\in[0]}\phi^{y_1}\Big)\Big(\sum_{y\in\Omega_n^1}\prod_{j=2}^{n}\phi^{y_j}\Big)$$

Continuing this process recursively, the lemma follows. ∎

Remark 7.3.2 For each pair of alternatives, let $p(i,j) = \mathbb{P}[i \succ j] = \sum_{\pi \in \mathcal{L}(\pi(i)<\pi(j))} \mathbb{P}[\pi]$ the probability that i beats j in a ranking randomly drawn according to the Mallows measure $\mathbb{P}$. Notice that the probability matrix $\mathbf{P} = [p(i,j)]$ is Toeplitz.

Afterwards, we will provide a way to deduce why one chooses this normalization constant and how the Kendall tau distance naturally arises for the definition of Mallows model. These observations come from the so-called RIM process. But, firstly, we present a different way to think of the Mallows model.

7.3.2 A different point of view

Kernelization has been proved a remarkably useful technique for the machine learning community. The main notion behind kernel-based methods is to define a positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ over the input space $\mathcal{X}$. Thus, for two input vectors $x, y \in \mathcal{X}$, $K(x,y)$ can be seen as a measure of similarity. Our purpose should be to design an embedding $\Phi : \mathcal{X} \to \mathcal{H}$ of the input space to a Hilbert space $\mathcal{H}$ (space with a well-defined inner product) in which the kernel reduces to an inner product, that is:

$$K(x,y) =< \Phi(x), \Phi(y) >_{\mathcal{H}}, \forall x, y \in \mathcal{X} \tag{7.9}$$

Probably, the most famous kernel is the $N-$dimensional Gaussian kernel defined as :

$$G(x;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}^N} exp(-\frac{||x||^2}{2\sigma^2}) \tag{7.10}$$

and

$$K_G(x,y) = exp(-\frac{||x-y||^2}{2\sigma^2}) \tag{7.11}$$

It is well mentioning that the Fourier transform of the Gaussian function is again a Gaussian function on the frequency domain. The Gaussian function is the only function with this property.

The Mallows kernel plays a role on the symmetric group similar to the Gaussian kernel on Euclidean space.

The Mallows kernel is defined for any $\beta \geq 0$ by :

$$K_{M,\beta}(\sigma,\pi) = exp(-\beta d_K T\sigma, \pi) \tag{7.12}$$

We can show that $K_{M,\beta}$ is positive definite for any $\beta \geq 0$. Define the mapping :

$$\Phi : \mathbb{S}_n \to \mathbb{R}^{\binom{n}{2}}$$

$$\sigma \xrightarrow{\Phi} (sgn(\sigma(i) - \sigma(j)))_{1 \leq i < j \leq n}$$

Mallows kernel corresponds to a Gaussian kernel on a $\binom{n}{2}-$dimensional embedding of $\mathbb{S}_n$. For more extensive results, we refer the reader to [JV15]. Thus, we can see how complex is the structure of the symmetric group. In order to project $\sigma \in \mathbb{S}_n$ to

a $\{\pm 1\}-$valued vector, one needs to map to the $\binom{n}{2}-$dimensional space. Thus, the increase is $\Theta(n^2)$.

For instance, for $n = 2$, $\Phi(1 \succ 2) = -1$ and $\Phi(2 \succ 1) = +1$. For $n = 3$, we have an embedding in $\mathbb{R}_3$ such that $\Phi(a \succ b \succ c) = (sgn(a-b), sgn(a-c), sgn(b-c))$.

7.4 The Repeated Insertion Model

The Condorcet/Mallows sampling procedure for drawing rankings from the Mallows distribution can be very inefficient in a computational point of view. For instance, the sample $a \succ b \succ c \succ a$ should be rejected. In general, it is inefficient since it relies on the rejection of partially constructed rankings as soon as a single circular or non-transitive sample is drawn (once the directed graph of the permutation has a cycle).

Efficient sampling is essential for a variety of inference and learning tasks. A computational perspective of the efficiency of sampling concerning the Mallows Models can be found in [LC14]. The main question that one could have is the following : Suppose we have a Mallows model and we want to draw a sample permutation.

Is there a process that offers a computationally efficient way to sample rankings?

Doignon proposed a generative process, called Repeated Insertion Model (RIM) that gives rise to a family of distributions over rankings and provides a practical way to sample rankings from a Mallows model.

The main idea is that we create a ranking by inserting each alternative one after another. The process in completed after n steps where n is the number of alternatives.

The model assumes a reference ranking $\pi = a_1 a_2 ... a_n$ and insertion probabilities $p_{i,j}$ for each $i \leq n, j \leq i$. RIM generates a new output permutation using the following procedure. We remind that we denote with $i \succ j$ when the alternative i is ranked above j.

- At step 1, the alternative a_1 is added to the output ranking.
- At step 2, the alternative a_2 is inserted either before or after a_1. The item a_2 is inserted above a_1 with probability $p_{2,1}$, generating the permutation $a_2 \succ a_1$ and below a_1 with probability $p_{2,2} = 1 - p_{2,1}$, generating the permutation $a_1 \succ a_2$.
- At step i, a permutation of alternatives $a_1, a_2, ..., a_{i-1}$ will be created and the alternative a_i will eventually be inserted in position $j \leq i$ with probability $p_{i,j}$.
- After n steps, the output ranking on the n alternatives will be a (valid) permutation $\in \mathbb{S}_n$.

Remark 7.4.1 The insertion probabilities are independent of the ordering of the previously inserted alternatives.

Doignon showed that one could choose the $p_{i,j}$ appropriately in order to create a generative process that corresponds to the Mallows model.

Definition 7.4.1 — Repeated insertion function. *Let $\pi = a_1...a_n$ be a reference ranking. Let an insertion vector be any positive integer vector $\vec{j} = (j_1, ..., j_n)$ s.t. $j_i \leq i \forall i \in [n]$. Consider the set J of all possible such vectors. A repeated insertion function $F_\pi : J \to \mathcal{L}(A)$ (where $\mathcal{L}(A)$ is isomorphic to $\mathbb{S}_n$) maps an insertion vector $\vec{j}$ into a ranking $F_\pi(\vec{j})$ by placing each a_i, in turn, into rank j_i for all $i \leq n$.*

■ **Example 7.1** For instance, consider the reference ranking $\pi = a_1 \succ a_2 \succ a_3 \succ a_4$. For the insertion vector $(1, 2, 3, 4)$ we get that $F_\pi(1, 2, 3, 4) = a_1 \succ a_2 \succ a_3 \succ a_4$ and for $(1, 1, 2, 3)$ we get that $F_\pi(1, 1, 2, 3) = a_2 \succ a_3 \succ a_4 \succ a_1$. ■

Given the reference ranking π, there is a '1-1' correspondence between rankings and insertion vectors. That is F_π is a bijection between J and $\mathbb{S}_n$.

Suppose we are given an insertion vector $\vec{j}$. What is the dislocation it creates?

It is easy to observe that whenever a_i is inserted at position j_i, it creates $(i - j_i)$ pairwise misorderings with respect to alternatives $a_1, ..., a_{i-1}$. All pairwise misorderings can be accounted this way. Thus, summing over all $i \leq n$ gives the Kendall tau distance.

Lemma 7.4.2 *For any insertion vector $\vec{j} = (j_1, ..., j_n) \in J$, we have that :*

$$\sum_{i=1}^{n} (i - j_i) = d_{KT}(F_\pi(\vec{j}, \pi)). \tag{7.13}$$

Suppose we are given an insertion vector $\vec{j}$ that is mapped to a ranking r under F_π. What is the probability of generating r under RIM?

Let $F_\pi(\vec{j}) = F_\pi(j_1, ..., j_n) = r$. Then the probability to generate ranking r under RIM is $\prod_{i=1}^{n} p_{i,j_i}$.

Theorem 7.4.3 — *RIM $\sim \mathcal{M}(\pi, \phi)$. By setting the insertion probabilities $p_{i,j} = \frac{\phi^{i-j}}{1+\phi+...+\phi^{i-1}}$ for $i \leq n, j \leq i$, the distribution induced by RIM with insertion function F_π is identical to that of the Mallows model $\mathcal{M}(\pi, \phi)$.*

Proof. Let $\mathcal{M}(\pi, \phi)$ be the Mallows model and r be any ranking in $\mathbb{S}_n$. Let $(j_1, ..., j_n)$ be the insertion ranks s.t. $F_\pi(j_1, ..., j_n) = r$. If we multiply the factors ϕ^{i-j_i} across $i \leq n$, we get $\phi^{\sum_{i=1}^{n} i-j_i} = \phi^{d_{KT}(r,\pi)}$. This term is exactly the proportional probability to draw r in Mallows model. The denominator of $\prod_{i=1}^{n} p_{i,j} = (1+\phi)(1+\phi+\phi^2)...(1+\phi+...+\phi^{m-1})$, that is independent of r. This is exactly the normalization constant $Z(\phi, \pi) = Z(\phi)$ of the Mallows model. ■

7.5 Generalized Mallows Model

In 1986, Fligner and Verducci introduced a generalization of the simple Mallows model. In the KT distance section, we mentioned that each permutation is in '1-1'

correspondence with a vector of numbers, namely the decomposition vector. The idea behind the Generalized Mallows Model exploits this correspondence.

Let $\sigma, \pi \in \mathbb{S}_n$. We define $V_j(\sigma, \pi)$ to be the number of discordant alternative pairs involving alternatives $i < j$ and alternative j, that is, for $j \in [n]$,

$$V_j(\sigma, \pi) = \sum_{1 \leq i < j} \mathbb{1}\{(\sigma(i) - \sigma(j))(\pi(i) - \pi(j)) < 0\}$$

The generalized Mallows family of distribution is

$$\mathcal{M}_n = \{\mathcal{P}_{\vec{\phi},\pi_0} : \vec{\phi} \in [0,1]^n, \pi_0 \in \mathbb{S}_n\}$$

parametrized with central ranking $\pi_0 \in \mathbb{S}_n$ and n-dimensional dispersion vector $\vec{\phi} = (\phi_1, ..., \phi_n) \in [0,1]^n$.

The probability mass function is defined as :

$$p_{\vec{\phi},\pi_0}(\sigma) = \frac{1}{Z(\vec{\phi}, \pi_0)} \prod_{i=1}^{n} \phi_i^{V_i(\sigma,\pi_0)}$$

It is clear that if $\phi_i = \phi$, we get the single parameter Mallows Model since $d_{KT}(\sigma, \pi) = \sum_{i=1}^{n} V_i(\sigma, \pi)$.

Another important property of the generalized Mallows Model is that, when the distance metric is the Kendall tau distance, the random variable $Y_j = V_j(\xi, \pi_0)$ where $\xi \sim \mathcal{P}_{\vec{\phi},\pi_0}$ are independent.

This follows from the following decomposition lemma of the partition function $Z(\vec{\phi})$:

Lemma 7.5.1 $Z(\vec{\phi}, \pi_0) = Z(\vec{\phi}) = \prod_{i=1}^{n} Z_i(\phi_i) = \prod_{i=1}^{n} \sum_{j=0}^{i-1} \phi_i^j$

The proof is completely similar to the normalization constant's proof in the single parameter Mallows Model by substituting each ϕ with a ϕ_j.

Hence, one can write :

$$p_{\vec{\phi},\pi_0}(\sigma) = \prod_{i=1}^{n} \frac{\phi_i^{V_i(\sigma,\pi_0)}}{Z_i(\phi_i)}$$

Studying the random variables Y_j

The random variables $Y_j = V_j(\xi, \pi_0)$ where $\xi \sim \mathcal{P}_{\vec{\phi},\pi_0}$ are the sufficient statistics for the distribution $\mathcal{P}_{\vec{\phi},\pi_0}$ when the central ranking is known. A very important question that arises is how each variable Y_j is distributed. Observe that the probability mass of the random vector $(Y_1, ..., Y_n)$ equals :

$$\mathbb{P}[Y_1 = d_1, ..., Y_n = d_n] = (\frac{\phi_1^{d_1}}{Z_1(\phi_1)}) \cdots (\frac{\phi_n^{d_n}}{Z_n(\phi_n)}) = \prod_{i=1}^{n} \mathbb{P}[Y_i = d_i]$$

Hence, it is clear how the above lemma decomposed the normalization constant in order to provide us with the desired independence of the random variables. If we

isolate the term Y_j, we can observe that the distribution of the random variable is quite similar with a geometric distribution. The main difference is that the geometric distribution has an infinite tail, while the tail of the distribution of Y_j is finite. This distribution is known and is called a *truncated geometric distribution.*

Definition 7.5.1 — Truncated geometric distribution. *A random variable ξ follows the truncated geometric distribution $\mathcal{TG}(\phi, k)$ with parameters $\phi \in [0,1]$ and $k \in \mathbb{N} \cup \{\infty\}$ if it has the following probability mass function*

$$p_\xi(i) = \frac{\phi^i}{\sum_{j=0}^{k} \phi^j}, i \in \{0, 1, ..., k\}$$

with support $\{i : i \in \{0, 1, ..., k\}\}$.

■ **Example 7.2** For $k = 1$, $\mathcal{TG}(\phi, 1) =_D Be(\frac{\phi}{1+\phi})$ where the success probability $\frac{\phi}{1+\phi} = 1 - p$ if $\phi = \frac{1-p}{p}$. For $k = \infty$ and $\phi < 1$, $\mathcal{TG}(\phi, \infty) =_D Geo(\phi)$. Note that if we fix k, then $\mathcal{E}_k = \{\mathcal{TG}(\phi, k) : \phi \in [0,1]\}$ is an exponential family with natural parameter $\theta = ln\phi$. ■

So, it is not difficult to see that :

$$Y_j = V_j(\xi, \pi_0) \sim \mathcal{TG}(\phi_j, j-1), \text{ where } \xi \sim \mathcal{P}_{\vec{\phi},\pi_0}$$

A useful lemma

Consider the distribution $\mathcal{P}_{\vec{\phi}}$ to be the multivariate distribution $(Y_1, ..., Y_n)$, where $Y_j \sim \mathcal{TG}(\phi_j, j-1)$. We need to link this distribution with the initial distribution $\mathcal{P}_{\vec{\phi},\pi_0}$ when the central ranking π_0 is known.

Lemma 7.5.2 *Let $\pi_0 \in \mathbb{S}_n$ and $\vec{\phi} \in [0,1]^n$. Let $R_{\vec{\phi}}$ be the support of the distribution $\mathcal{P}_{\vec{\phi}}$ and let $R_{\vec{\phi},\pi_0}$ the support of $\mathcal{P}_{\vec{\phi},\pi_0}$. Then, there exists a bijective map $g : R_{\vec{\phi},\pi_0} \to R_{\vec{\phi}}$ s.t. for any $\sigma \in R_{\vec{\phi},\pi_0}$, it holds that :*

$$\mathbb{P}_{\tau \sim \mathcal{P}_{\vec{\phi},\pi_0}}[\tau = \sigma] = \mathbb{P}_{\vec{y} \sim \mathcal{P}_{\vec{\phi}}}[\vec{y} = g(\sigma)]$$

Also, it holds that :

$$d_{TV}(\mathcal{P}_{\vec{\phi_1},\pi_0}, \mathcal{P}_{\vec{\phi_2},\pi_0}) = d_{TV}(\mathcal{P}_{\vec{\phi_1}}, \mathcal{P}_{\vec{\phi_2}})$$

and

$$D_{KL}(\mathcal{P}_{\vec{\phi_1},\pi_0} \parallel \mathcal{P}_{\vec{\phi_2},\pi_0}) = D_{KL}(\mathcal{P}_{\vec{\phi_1}} \parallel \mathcal{P}_{\vec{\phi_2}})$$

Proof. The bijective mapping is given by the structure of the Generalized Mallows model, since one can define, for any $\sigma \in R_{\vec{\phi},\pi_0}$:

$$g(\sigma) = (V_1(\sigma, \pi_0), ..., V_n(\sigma, \pi_0))$$

We have already seen that $(Y_1, ..., Y_n)$, where $Y_j = V_j(\xi, \pi_0)$, are independent if $\xi \sim \mathcal{M}(\pi_0, \vec{\phi})$. Thus, the joint distribution of the n-dimensional random vector is equivalent to the probability mass of the Generalized Mallows model, that is :

$$\mathbb{P}_{(y_1,...,y_n)\sim\mathcal{P}_{\vec{\phi}}}[\vec{y} = g(\sigma)] = \mathbb{P}[Y_1 = y_1, ..., Y_n = y_n] = \prod_{i=1}^{n} \frac{\phi_i^{V_i(\sigma,\pi_0)}}{Z_i(\phi_i)} = \mathbb{P}_{\tau\sim\mathcal{P}_{\vec{\phi},\pi_0}}[\tau = \sigma]$$

This bijective mapping preserves the mass and, hence, we get the other two equalities. ■

Remark 7.5.3 Note that if we consider the case where $\phi_j = \phi \ \forall j \in [n]$, the above results still hold for the single parameter Mallows model.

7.6 The Plackett - Luce Model

The Plackett - Luce probabilistic model is another model for permutations. The difference between the Mallows and the PL model is the way one creates the permutation. The intuition behind Mallows model could be the idea of the construction of a tournament as we referred to previous sections. The idea of PL model looks like a generalized version of repeated insertion model.

Firstly, we provide an examples of how a permutation will be generated by the PL model.

Assume that we have a box with m balls. Each ball a_i values w_i. We can normalize the values of the balls in order to get $\sum_{i=1}^{m} w_i = 1$. We create a permutation of the m balls in m steps, as follows :

In each step, we choose a ball and we pick it out of the box. The probability for each remaining ball to be chosen equals its value over the sum of values of the remaining balls inside the box. The way to pick the m balls induces a unique ranking of the items. We note that, in the first step, the probability to draw ball i equals its value $0 \leq w_i \leq 1$.

Before formalizing the PL model, we introduce the fundamental idea behind Plackett-Luce model, that is due to the work of Duncan Luce in 1959.

The Luce's choice axiom

Luce's choice axiom consists of two parts. Let C be a choice set.

Luce's Axiom 1

Assume that C contains two elements x, y such that x is never chosen over y when the choice is restricted to only x and y. Without affecting any of the choice probabilities, x can be deleted from C.

Luce's Axiom 2

Assume that $S \subset C$. Then, the choice probabilities for the choice set S are considered to be identical to the choice probabilities for the choice set C conditional on S having been chosen

$$\mathbb{P}_S(\alpha) = \mathbb{P}_C(\alpha|S),\ \alpha \in S$$

The axiom can be restated as, if we assign masses w_i on the items, the probability of selecting item i from a pool S of j items is :

$$\mathbb{P}_S[i] = \frac{w_i}{\sum_{j \in S} w_j}$$

This formula is completely similar to the known softmax function.

PL model

Let $\mathcal{A} = \{a_i | i \in [m]\}$ the set of m alternatives and let $W = \{\vec{w} = \{w_i | i \in [m], w_i \in [0,1], \sum_{i=1}^{m} w_i = 1\}\}$ be the set of all possible values of these alternatives. The m values could be represented by a $m-$dimensional vector.

Remark 7.6.1 Since we require that the sum of values is fixed, we only need $m-1$ values, but for simplicity, we prefer to have a $m-$dimensional vector.

Given a value vector $\vec{w} \in W$, the probability to generate the ranking

$$\sigma = a_{i_1} \succ a_{i_2} \succ ... \succ a_{i_m}$$

equals to :

$$\mathbb{P}[\sigma|\vec{w}] = w_{a_{i_1}} \cdot \frac{w_{a_{i_2}}}{\sum_{p>1} w_{i_p}} \cdot ... \cdot \frac{w_{a_{i_{m-1}}}}{w_{a_{i_{m-1}}} + w_{a_{i_m}}}$$

Alternatives with a higher weight tend to occupy higher positions in the induced ranking. The most probable ranking can be obtained by sorting the alternatives in decreasing order with respect to their weight :

$$\sigma^* = arg \max_{\sigma \in \mathbb{S}_n} \mathbb{P}[\sigma|\vec{w}] = argsort_{i \in [n]}\{w_1, ..., w_n\}$$

Note that PL model is more flexible that the simple Mallows model since the parameter size is linear to the number of alternatives. There is a vaste collection of probabilistic models over rankings. In this chapter, we presented the two more fundamental noisy models, the Mallows model and the Plackett-Luce model. These two models are those that one first encounters when working on that topic. For sake of completeness, in the following section, we present some other probabilistic noisy models.

7.7 Other noisy models

In the previous sections, we have seen three noisy models-distributions : The single parameter Mallows model (Mallows, 1957, [Mal57]), the generalized Mallows model (Fligner & Verducci, 1989, [Fli86]) and the PL model (Plackett, 1975, [Pla75]- Luce, 1959, [Luc59]). In the current section, we shortly introduce some other more rare noisy models for sake of completeness ([AEMP18], [YR08], [LDSR16]).

The Babington Smith Distribution

The Babington Smith (BS) model was introduced in 1950 by B. Babington-Smith [BS50]. Consider a collection of n alternatives $A = \{a_1, ..., a_n\}$. The probability of sampling the permutation $\sigma \in \mathcal{L}(A)$ equals :

$$\mathbb{P}_\theta(\sigma) = \frac{1}{C(\theta)} \prod_{1 \leq i < j \leq n} p_{\sigma^{-1}(i),\sigma^{-1}(j)}$$

where $p_{i,j}$ is the probability of observing the preference $a_i \succ a_j$ when comparing the two alternatives. The quantity $C(\theta)$ is just a normalization constant. The BS model is parametrized by θ, which consists of all pairwise probabilities $p_{i,j} = 1 - p_{j,i}$. The parameter θ consists of $\binom{n}{2}$ values.

The BS distribution results from a process quite similar to the Condorcet-Mallows' model. The order of each pair of alternatives a_i and a_j is determined independently at random by flipping a $p_{i,j}$-biased coin $\sim Be(p_{i,j})$. If these comparisons generate a valid and consistent ranking, the BS model outputs the induced permutation. Otherwise, we repeat the generating process.

It is not difficult to observe that BS model has a much richer parametrization, compared to the Mallows model and to the PL model. The parameter size of the single parameter Mallows model is far more restricted since it contains only two parameters π_0, ϕ. The PL model is more flexible since the parameter size grows linearly to the size of alternatives. BS model is even more flexible since the parameter size grows quadratically with the number of elements in A.

The Average-Precision Distribution

The Average-Precision (AP) model was introduces by Yilmaz, Aslam, and Robertson in 2008 [YR08]. In the area of information retrieval (IR), a fundamental part of the ongoing research is crucially linked to ranked lists of items. For instance, the output of search engines is a ranked list of documents. Thus, it is important to be able to compute the correlation between two ranked lists. One of the most commonly used statistics is the Kendall's tau statistic-distance. However, in the field of IR, it is quite common that inconsistencies among alternatives having higher rankings are far more important that those between low ranked items. A ranking mistake at high positions of a Google search should be penalized more than a ranking error in the third page of the same Google search. The Kendall's tau statistic is 'blind' in this property, since it make no distinction in the position of the mistake but just in the mistake

itself. Thus, it would be useful to introduce a statistic that penalizes in a different way errors at high rankings and at low rankings.

This is exactly why Yilmaz et al. introduced the AP distance. Consider a reference ranking $\pi \in \mathbb{S}_n$. The AP distance of a ranking σ from the ranking π is :

$$d_{AP}(\pi, \sigma) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E_{ij} \frac{n}{2(j-1)}$$

where $E_{ij} = 1 \iff \pi(i)$ is ranked after $\pi(j)$ in σ. Otherwise, $E_{ij} = 0$ ($\iff \sigma(\pi(i)) < \sigma(\pi(j))$).

Remark 7.7.1 AP distance is not symmetric since it is computed with respect to a central ranking π.

Remark 7.7.2 AP distance constitutes a generalization of KT distance since if one replaces the weights $\frac{n}{2(j-1)}$ with 1, one gets the KT distance (that is symmetric),

$$d_{KT}(\pi, \sigma) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E_{ij} = d_{KT}(\sigma, \pi)$$

It is not difficult to observe that AP distance assigns weights to the inverted pairs in σ with respect to π which are dependent on their positions in π.. An inversion in σ for the two alternatives $\pi(i)$ and $\pi(j)$ for $i < j$ costs $\frac{n}{2(j-1)}$.

The cost can be seen as $\frac{\frac{n}{2}}{j-1}$. Thus, the cost assigned by AP for $j < \frac{n}{2} + 1$ is higher than 1 and the cost for $j > \frac{n}{2} + 1$ is less than 1.

The AP model corresponds to a probability distribution over the symmetric group $\mathbb{S}_n$, parametrized by a central ranking π_0 and a dispersion parameter $\beta > 0$. The probability of drawing a ranking σ is :

$$\mathbb{P}[\sigma|\pi_0] = \frac{1}{Z(\beta)} e^{-\beta d_{AP}(\pi_0, \sigma)}$$

where $Z(\beta)$ is a normalization constant.

Remark 7.7.3 The MLE of the simple Mallows model is NP-hard. The same holds for the AP case.

AP-MLE PROBLEM

Given a multiset R of elements of $\mathbb{S}_n$, find the permutation $\pi^*_{AP} \in \mathbb{S}_n$ such that :

$$\pi^*_{AP} = arg \max_{\pi \in \mathbb{S}_n} \prod_{\sigma \in R} \mathbb{P}[\sigma|\pi]$$

The AP-MLE PROBLEM is NP-hard too. The proof can be found to the complete version of [LDSR16].

8. Learning to rank from noisy information

As referred previously, a noise model defines the probability measure $\mathbb{P}$ of observing a ranking π given an underlying true ranking π_0, that is $\mathbb{P}[\pi|\pi_0]$ for all $\pi, \pi_0 \in \mathcal{L}(A)$. In this section, we will focus on the Mallows noise model.

8.1 Sample Complexity in Mallows Models

The main question in this section will be the following :

How many samples are needed by different voting rules in order to determine the true (hidden) ranking of a Mallows model with high probability?

Firstly, we need to describe a metric of 'counting samples'. We use as criterion the *sample complexity* to distinguish between voting rules.

Definition 8.1.1 — Accuracy of rule r. *For any randomized voting rule r, true ranking $\pi_0 \in \mathcal{L}(A)$, and m samples $\in \mathbb{N}$, let :*

$$ACC_r(m, \pi_0) = \sum_{\pi \in \mathcal{L}(A)^m} \mathbb{P}[\pi|\pi_0]\mathbb{P}[r(\pi) = \pi_0] \tag{8.1}$$

Accuracy of rule r is the probability that rule r returns π_0 given m samples from Mallows model with true ranking π_0.

In order to let $ACC_r(m, \pi_0)$ be independent of the true ranking π_0, we consider the worst case scenario that is :

$$ACC_r(m) = \min_{\pi_0 \in \mathcal{L}(A)} ACC_r(m, \pi_0) \tag{8.2}$$

That is rule r returns the underlying true ranking with probability at least $ACC_r(m)$.

Let $\epsilon > 0$, small enough. We need to define a quantity that denotes the required number of samples in order to get the true ranking from rule r with high probability (that is at least $1 - \epsilon$) :

$$N_r(\epsilon) = \min\{m | ACC_r(m) \geq 1 - \epsilon\}. \tag{8.3}$$

We call $N_r(\epsilon)$ the *sample complexity of rule r.*

Claim : Kemeny rule (where ties are broken uniformly) requires the minimum number of samples from Mallows model to determine the true ranking with high probability.

This claim is not completely random. There is a profound reason why Kemeny rule is that 'strong'. Before we proceed, it is worth reminding that Kemeny rule is the maximum likelihood estimator for the true ranking given samples from Mallows model.

Given a profile $\pi = (\pi_1, ..., \pi_m) \in \mathcal{L}(A)^m$, where each sample is drawn independently from a Mallows distribution $\pi_i \sim \mathcal{P}_{\pi_0,\phi}$, the MLE τ of the true ranking is that ranking that maximizes the probability of drawing the profile π :

$$arg \max_{\tau \in \mathcal{L}(A)} \mathbb{P}[\pi|\tau] = arg \max_{\tau \in \mathcal{L}(A)} \frac{1}{Z^m} \prod_{1 \leq i \leq m} \phi^{d_{KT}(\pi_i, \tau)} = arg \min_{\tau \in \mathcal{L}(A)} \sum_{1 \leq i \leq m} d_{KT}(\pi_i, \tau)$$

The first equality follows from the independence of our samples and the second follows since $0 < \phi < 1$ and $x \mapsto logx$ is an increasing function with $log\phi < 0$. This result is proved in detail in the beginning of the next chapter. We show that the Kemeny's rule is optimal as far as sample complexity is concerned for the Mallows model.

Theorem 8.1.1 — Optimality of Kemeny's rule. *The Kemeny rule with uniform tie-breaking has the optimal sample complexity in Mallows model, that is, for any $\epsilon > 0$, any number of alternatives n and any randomized voting rule r, $N_{KEMENY}(\epsilon) \leq N_r(\epsilon)$.*

The above theorem arises two natural questions :

Are there any other rules that have the same asymptotic sample complexity as that of Kemeny's rule?

What is the value $N_{KEMENY}(\epsilon)$? That is how many samples Kemeny's rule requires?

Both questions will be answered in the following section.

8.2 PM-c Rules

Kemeny rule belongs to a large family of voting rule with optimal samples complexity. This family relies on the concept of pairwise-majority graph (PM graph) and of pairwise-majority consistent rules (PM-c rules).

Firstly, we should expand our intuition of a voting profile $\pi \in \mathcal{L}(A)^m$ to a graph theoretic concept. Consider a graph $G = (V, E)$ such that each vertex is just an alternative, that is $V = A$ and there is an edge from alternative a to b if a is preferred to b in a strong majority of the votes of π. That is $(a, b) \in E$ iff $|\sigma \in \pi : a \succ_\sigma b| > |\sigma \in \pi : b \succ_\sigma a|$. In case of ties between alternatives, there is no edge between them. Note that there can never be an edge in both directions, but a PM graph can have directed cycles.

Thus, each voting profile π generates a directed graph G_π, which is called the pairwise-majority graph (PMg).

Given a PM graph G, can we deduce a unique ranking from G?

When a PM graph is complete and acyclic, there exists a unique $\sigma \in \mathcal{L}(A)$ such that there is an edge $a \to b$ iff $a \succ_\sigma b$.

Then, we say that G reduces to σ. Thus, there exists an isomorphism between a subset $C \cap A$ of PM graphs and the symmetric group $\mathbb{S}_n$, where $C \cap A = \{G : E(G) = \binom{n}{2}, acyclic, G \in PMg\}$.

Hence, a 'nice' voting rule r would be one that agrees with the PM graph. That is given a voting profile π, which generates a PM graph that reduces to a ranking σ, then the 'nice' voting rule would give the same ranking σ. This is exactly the notion of PM-c rules.

Definition 8.2.1 — Pairwise-Majority Consistent Rules. *Consider a profile π. A deterministic voting rule r is called PM-c if $r(\pi) = \sigma$ whenever the PM graph of π reduces to σ. A randomized voting rule is similarly called PM-c whenever $\mathbb{P}[r(\pi) = \sigma] = 1$.*

The main result of this section follows. We claim that PM-c rules have logarithmic sample complexity.

Theorem 8.2.1 — Sample Complexity of PM-c rules. *For any given $\epsilon > 0$, any PM-c rule determines the true ranking of the n alternatives with probability at least $1 - \epsilon$ given $O(log(\frac{n}{\epsilon}))$ samples generated from the Mallows model.*

Before proving the theorem, we should note that the number of samples behaves naturally with n and ϵ. The increase of alternatives (increase of n) and the increase of the probability being correct (decrease of ϵ) require more samples. As we will see, the *log* factor appears due to the inversion of the exponential generated by Hoeffding's inequality and n appears when using the union bound technique.

Proof. Let π_0 be the hidden true ranking of the n alternatives. We will show that, given $m = O(log(\frac{n}{\epsilon}))$ votes generated from the Mallows model, the corresponding PM graph reduces to π_0 with high probability (at least $1 - \epsilon$).

Let m samples drawn from Mallows models and let $VP \in \mathcal{L}(A)^m$ the voting profile. For any two alternatives a, b, we will count the number of votes in which a beats b and denote that counter with n_{ab}. Thus, $n_{ab} = |\{\sigma : a \succ_\sigma b, \sigma \in VP\}|$. Obviously, $n_{ab} + n_{ba} = m, \forall a, b \in A$.

The PM graph of the voting profile VP reduces to $\pi_0 \leftrightarrow \forall a, b \in A$ for which $a \succ_{\pi_0} b$, we have that $n_{ab} - n_{ba} \geq 1$.

In order to learn π_0 using m samples, we need to provide an upper bound on m s.t.

$$\mathbb{P}[a \succ_{\pi_0} b \Rightarrow n_{ab} - n_{ba} \geq 1, \forall a, b \in A] \geq 1 - \epsilon$$

Equivalently, we have to upper bound the probability of the 'bad' event and then use the classical union bound technique.

For any $a, b \in A$ with $a \succ_{\pi_0} b$, we have that :

$$\mathbb{P}[n_{ab} - n_{ba} \leq 0] = \mathbb{P}[\frac{n_{ab} - n_{ba}}{m} \leq 0]$$

Set $X_{ab} = \frac{n_{ab} - n_{ba}}{m}$. Observe that if $p_{a \succ b}$ be the probability that $a \succ_\pi b$ in a random sample π, then $\mathbb{E}[X_{ab}] = p_{a \succ b} - p_{b \succ a}$, using the linearity of the expectation.

So,

$$\mathbb{P}[n_{ab} - n_{ba} \leq 0] = \mathbb{P}[X_{ab} \leq 0] = \mathbb{P}[X_{ab} - \mathbb{E}X_{ab} \leq -\mathbb{E}X_{ab}] \leq \mathbb{P}[|X_{ab} - \mathbb{E}X_{ab}| \geq \mathbb{E}X_{ab}]$$

The last inequality follows from the properties of the absolute value function.

Now, we have a classical tail inequality. We could use any known inequality that we have seen in the concentration inequalities section. We will use the Hoeffding's inequality 4.46 and get :

$$\mathbb{P}[n_{ab} - n_{ba} \leq 0] \leq 2exp(-2(\mathbb{E}X_{ab})^2 m)$$

In order to take a general upper bound, we pick the minimum value that expectation can take, setting $\delta_{min} = \min_{a,b \in A : a \succ_{\pi_0} b} \mathbb{E}X_{ab}$ and getting :

$$\mathbb{P}[n_{ab} - n_{ba} \leq 0] \leq 2exp(-2\delta_{min}^2 m)$$

We can use the union bound technique to upper bound the probability that 'bad' events happen :

$$\mathbb{P}[\exists a, b \in A : \{a \succ_{\pi_0} b\} \cap \{n_{ab} - n_{ba} \leq 0\}] \leq \binom{n}{2} 2exp(-2\delta_{min}^2 m) \leq n^2 e^{-2\delta_{min}^2 m}$$

We want that this probability is at most ϵ and, thus,

$$\mathbb{P}[\exists a, b \in A : \{a \succ_{\pi_0} b\} \cap \{n_{ab} - n_{ba} \leq 0\}] \leq n^2 e^{-2\delta_{min}^2 m} < \epsilon \Rightarrow m \geq \frac{1}{2\delta_{min}^2} log(\frac{n^2}{\epsilon})$$

Hence,

$$m \geq \frac{1}{2\delta_{min}^2} log(\frac{n^2}{\epsilon}) \tag{8.4}$$

Additionally, we have to show that $\delta_{min} = \Omega(1)$. Thus, it is lower bounded by a constant independent of n.

We have that, for any $a, b \in A$, s.t. $a \succ_{\pi_0} b$:

$$\delta_{ab} = p_{a\succ b} - p_{b\succ a} = \sum_{\pi \in \mathcal{L}(A): a \succ_\pi b} \mathbb{P}[\pi|\pi_0] - \sum_{\pi \in \mathcal{L}(A): b \succ_\pi a} \mathbb{P}[\pi|\pi_0]$$

Now, recall the swap increasingness section from the introductory chapter 3.1.3. We can unify this two sums into one using the $\pi_{a\leftrightarrow b}$ permutation by simply observing that :

$$\delta_{ab} = \sum_{\pi \in \mathcal{L}(A): a \succ_\pi b} (\mathbb{P}[\pi|\pi_0] - \mathbb{P}[\pi_{a\leftrightarrow b}|\pi_0])$$

By the swap increasing property,

$$\delta_{ab} \geq (1 - \phi) p_{a\succ b}$$

But, since $\delta_{ab} = p_{a\succ b} - p_{b\succ a}$ and $p_{a\succ b} + p_{b\succ a} = 1$, we get that :

$$\delta_{ab} \geq \frac{1-\phi}{1+\phi}$$

and this holds for all $a, b \in A$ with $a \succ_{\pi_0} b$. Thus, it holds for δ_{min} too, and this completes the proof. Notice that the equality holds when a, b are adjacent. ■

The following result states that no randomized voting rule can do better. Hence, we provide a matching lower bound and, so, PM-c rules can learn the central ranking with $\Theta(log(\frac{n}{\epsilon}))$ samples with high probability.

Theorem 8.2.2 — Matching lower bound. *For any $\epsilon \in (0, \frac{1}{2}]$, any randomized voting rule requires $\Omega(log(\frac{n}{\epsilon}))$ samples generated from the Mallows model to determine the true central ranking with probability at least $1 - \epsilon$.*

Proof. Consider any voting rule r. Assume that for some $n \in \mathbb{N}$, $ACC_r(m) \geq 1-\epsilon$. We would like to show that $m = \Omega(log(\frac{n}{\epsilon}))$. Obviously, for any $\sigma \in \mathcal{L}(A)$, $ACC_r(m, \sigma) \geq 1 - \epsilon$ since accuracy considers the worst case scenario.

Let $\sigma \in \mathcal{L}(A)$. We will call a permutation π a neighboor of σ if $d_{KT}(\pi, \sigma) = 1$. Let $\mathcal{N}(\sigma)$ be the set of all neighboors of σ. Firstly, from the triangle inequality, we have that, if $\sigma' \in \mathcal{N}(\sigma)$ and $\pi \in \mathcal{L}(A)$:

$$d_{KT}(\pi, \sigma) \leq d_{KT}(\pi, \sigma') + d_{KT}(\sigma', \sigma) = d_{KT}(\pi, \sigma') + 1$$

Hence, $\phi^{d_{KT}(\pi,\sigma)} \geq \phi^{d_{KT}(\pi,\sigma')+1}$, since $\phi < 1$.

Thus, for any $\sigma' \in \mathcal{N}(\sigma)$ and a voting profile $VP = (\pi_1, ..., \pi_m) \in \mathcal{L}(A)^m$ of m i.i.d. votes, we get that :

$$\mathbb{P}[VP|\sigma] = \prod_{i=1}^{m} \frac{\phi^{d_{KT}(\pi_i,\sigma)}}{Z} \geq \prod_{i=1}^{m} \frac{\phi^{d_{KT}(\pi,\sigma')+1}}{Z} = \phi^m \mathbb{P}[VP|\sigma']$$

Now,

$$ACC_r(m,\sigma) = \sum_{\pi\in\mathcal{L}(A)^m} \mathbb{P}[\pi|\sigma]\mathbb{P}[r(\pi)=\sigma] = \sum_{\pi\in\mathcal{L}(A)^m} \mathbb{P}[\pi|\sigma](1-\mathbb{P}[r(\pi)\neq\sigma])$$

$$= 1 - \sum_{\pi\in\mathcal{L}(A)^m} \mathbb{P}[\pi|\sigma]\mathbb{P}[r(\pi)\neq\sigma]$$

The probability that $r(\pi)\neq\sigma$ is less than the probability that $r(\pi)$ returns one of the neighbors of σ.

Thus,

$$ACC_r(m,\sigma) \leq 1 - \sum_{\pi\in\mathcal{L}(A)^m} \mathbb{P}[\pi|\sigma](\sum_{\sigma'\in\mathcal{N}(\sigma)} \mathbb{P}[r(\pi)=\sigma'])$$

Then, using the derived inequality,

$$ACC_r(m,\sigma) \leq 1 - \sum_{\sigma'\in\mathcal{N}(\sigma)} \sum_{\pi\in\mathcal{L}(A)^m} \phi^m \mathbb{P}[\pi|\sigma']\mathbb{P}[r(\pi)=\sigma']$$

$$= 1 - \phi^m \sum_{\sigma'\in\mathcal{N}(\sigma)} ACC_r(m,\sigma') \leq 1 - \phi^m(n-1)(1-\epsilon)$$

The last inequality follows because $ACC_r(m)\geq 1-\epsilon$ and $|\mathcal{N}(\sigma)| = n-1$.

Hence, in order to obtain an accuracy $\geq 1-\epsilon$, we need :

$$1-\phi^m(n-1)(1-\epsilon)\geq 1-\epsilon \Rightarrow m = \Omega(log(\frac{n}{\epsilon})).$$

■

8.3 Non-Robustness of PM-c Rules

In this section, we will study how robust are PM-c voting rules under noise.

Informal Theorem

There exist profiles in which the PM-c graph is acyclic and all edge weights are large (the difference between pairwise preferences is large), but the noisy profile will, with high probability, have an acyclic PM-c graph too, that reduces to a different ranking.

We firstly introduce the setting to the reader. As always, suppose that we have a set of n alternatives A and that preferences over this set are permutations on A, that is each sample will be a ranking $\sigma \in \mathcal{L}(A)$. Given a preference profile $\pi = (\sigma_1, ..., \sigma_m) \in \mathcal{L}(A)^m$ of m votes, we say that $a\in A$ beats $b\in A$, that is $a \succ b$ when :

$$|\{i\in[m] : a\succ_{\sigma_i} b\}| > \frac{m}{2}$$

Hence, a beats b in pairwise majority comparison, when the strong majority of voters prefers a over b.

In previous sections, we referred to the concept of PM-c graph. Now, we will expand that notion to the weighted PM-c graph. The weight of the directed edge $a \to b$ is just the difference of votes where $a \succ b$ and the votes where $b \succ a$.

Definition 8.3.1 — Weighted PM-c Graph. *The profile $\pi = (\sigma_1, ..., \sigma_m) \in \mathcal{L}(A)^m$ induces the weighted pairwise majority graph $G_\pi = (V, E, w)$, where :*

- $V = A$
- $a \to b \in E \iff a \succ b$, $\forall a, b \in A, a \neq b$ *and the weight of the edge will be equal to*

$$w_{(a,b)}(\pi) = |\{i \in [m] : a \succ_{\sigma_i} b\}| - |\{i \in [m] : b \succ_{\sigma_i} a\}|$$

Under a PM-c voting rule, given a voting profile π, when the weighted PM-c graph is a tournament and, furthermore, is acyclic, it reduces uniquely to a ranking τ. This output ranking is simply the topological ordering of the PM-c graph G_π.

We claim that there exist profiles in which the PM-c graph is acyclic and all edge weights are large (the difference between pairwise preferences is large), but, with high probability, the noisy profile has an acyclic PM-c graph too, that reduces to a different ranking. This implies that any PM-c rule is not robust under noise, since it would return a different ranking when applied to the true and to the noisy profiles.

Theorem 8.3.1 *For all $\delta > 0, \phi \in (0, 1)$ and $m \in \mathbb{N}$ with $n \geq 3$, $\exists n_0 \in \mathbb{N}$ s.t. $\forall n \geq n_0$, $\exists$ a voting profile $\pi^* \in \mathcal{L}(A)^n$ s.t. G_{π^*} is acyclic and all edges have weight $\Omega(n)$, but, with probability at least $1 - \delta$, one could sample a noisy profile π, where G_π is acyclic and there is a pair of alternatives on which the unique rankings induces by G_{π^*} and G_π disagree.*

The intuition of the theorem is that even if a PM-c rule provides big gaps between alternatives, that is the pairwise preference differences (which are expressed via graph weights) are significant, some alternatives will possibly flip under that rule with high probability. Extended results concerning the robustness of PM-c rules and of other rules, such as Borda's count, can be found in [AK19].

8.4 Learning the parameters of Mallows model

The Mallowss model can be parametrized by the set of distributions

$$\mathcal{M}_1 = \{\mathcal{P}_{\phi,\pi_0} | \phi \in [0, 1], \pi_0 \in \mathbb{S}_n\}$$

with probability mass function $p_{\phi,\pi_0}(\pi) = \frac{\phi^{d_{KT}(\pi,\pi_0)}}{Z}$, where ϕ and π_0 are the parameters of the model. If we fix the permutation parameter π_0, the family $\mathcal{M}_1(\pi_0) = \{\mathcal{P}_{\phi,\pi_0} | \phi \in [0, 1]\}$.

In this section, we will try to answer the following question :

What is the sample complexity for learning the parameters ϕ, π_0 of a single distribution $\mathcal{P}_{\phi,\pi_0} \in \mathcal{M}_1$, given i.i.d. samples $\pi_1, ..., \pi_m$?

Learning the central ranking π_0

In section 8.2, we have shown that given $\Theta(log\frac{n}{\epsilon})$ samples, one can learn the hidden central ranking with high probability.

We restate the central ranking learning theorem :

Theorem 8.4.1 *For any $\pi_0 \in \mathbb{S}_n$ and any $\phi \in [0, \gamma)$, there exists a polynomial time estimator π^* s.t. given $m = \Theta(\frac{1}{\gamma} log\frac{n}{\epsilon})$ i.i.d. samples $\pi_1, ..., \pi_m \sim \mathcal{P}_{\phi,\pi_0}$ satisfies*

$$\mathbb{P}_{\vec{\pi}\sim\mathcal{P}^m_{\phi,\pi_0}}[\pi^* \neq \pi_0] \leq \epsilon$$

Moreover, if $m = o(log\frac{n}{\epsilon})$ then for any estimator π^* these exists a distribution $\mathcal{P}_{\phi,\pi_0}$ s.t.

$$\mathbb{P}_{\vec{\pi}\sim\mathcal{P}^m_{\phi,\pi_0}}[\pi^* \neq \pi_0] > \epsilon.$$

It is worth noticing that the sample complexity is a function of the error parameter ϵ (where the smaller the probability of being mistaken, the larger the samples needed) and of the size of the permutation n.

Now, it remains to estimate the parameter ϕ.

Learning the spread parameter ϕ

Having learned the central ranking, one wants to further discover what the spread parameter is. But, since in general $\phi \in [0, 1]$, the probability to learn exactly its value is 0. So, we introduce an additional estimation error ϵ, that controls the interval in which the predictor will be correct. So, for the learning of the spread parameter, our learning algorithm is an ϵ, δ algorithm, similar to the classical PAC learning concept, and the sample complexity is a function of these two parameters and of the size of the permutation.

Theorem 8.4.2 *In the case where π_0 is known, for $\phi \in [0, \gamma), \epsilon, \delta > 0$ there exists an estimator ϕ^* that can be computed in polynomial time s.t. given m i.i.d. samples $\vec{\pi} \sim \mathcal{P}^m_{\phi,\pi_0}$ with $m \geq \Omega(\frac{1}{n\epsilon^2} log\frac{1}{\delta})$*

$$\mathbb{P}_{\vec{\pi}\sim\mathcal{P}^m_{\phi,\pi_0}}[|\phi - \phi^*| \leq \epsilon] = \mathbb{P}_{\vec{\pi}\sim\mathcal{P}^m_{\phi,\pi_0}}[\phi^* \in [\phi - \epsilon, \phi + \epsilon]] \geq 1 - \delta$$

- ϵ controls the boundaries of the accuracy of the ϕ estimator. The higher the ϵ, the larger the accepted deviation from the correct values and thus the less the required samples.
- δ controls the error probability. As δ grows, the probability of being mistaken grows and thus the number of samples required drops down.

Proof. The proof uses similar techniques to the proof of theorem 8.6.1. Hence, for the complete proof, we refer the reader to [RBF19]. ■

Thus, combining the above results :

Theorem 8.4.3 *For any $\pi_0 \in \mathbb{S}_n$, $\phi \in [0, \gamma), \epsilon, \delta > 0$ there exist estimators π^*, ϕ^* that can be computed in polynomial time s.t. given m i.i.d. samples $\vec{\pi} \sim \mathcal{P}^m_{\phi,\pi_0}$ with $m \geq \Omega(\frac{logn}{\gamma} + \frac{1}{n\epsilon^2} log\frac{1}{\delta})$, then*

$$\mathbb{P}_{\vec{\pi} \sim \mathcal{P}^m_{\phi,\pi_0}}[(\pi^* = \pi_0) \wedge (\phi^* \in [\phi - \epsilon, \phi + \epsilon])] \geq 1 - \delta$$

Learning with 1 sample

How well one would estimate the spread parameter given only one sample? From the above theorem, requiring that $m = 1$, one could let the probability δ of being mistaken free and lock the boundary of the estimation :

$$\frac{1}{n\epsilon^2} log\frac{1}{\delta} = 1 \Rightarrow \epsilon = \sqrt{\frac{1}{n} log\frac{1}{\delta}}$$

Thus, one gets the following :

Theorem 8.4.4 *In the case where π_0 is known, any $\phi \in [0, 1], \delta > 0$ there exists an estimator ϕ^* that can be computed in polynomial time from one sample $\pi \sim \mathcal{P}_{\phi,\pi_0}$ s.t.*

$$\mathbb{P}_{\pi \sim \mathcal{P}_{\phi,\pi_0}}[\phi^* \in [\phi - \epsilon, \phi + \epsilon]] \geq 1 - \delta, \text{ where } \epsilon = O(\sqrt{\frac{1}{n} log\frac{1}{\delta}})$$

8.5 Learning Mallows model in TV Distance

In this section, we provide a lower bound for learning in TV distance in the setting of the simple single parameter Mallows model.

We will use the Fano's inequality mentioned in the information theory section 5.1.3. We will show that, fixing a bad spread parameter ϕ^*, there is a central ranking π_0, such that, whatever we sample from the distribution $\mathcal{P}_{\phi^*,\pi_0}$, given that the number of samples is small, then the distribution that we will construct cannot be close to the initial in total variation distance.

Theorem 8.5.1 Let $\phi^* = \frac{1}{2}$. Then $\exists \pi_0 \in \mathbb{S}_n$, s.t. if one samples a voting profile $\pi = (\sigma_1, ..., \sigma_m) \sim \mathcal{P}^m_{\phi^*,\pi_0}$, where σ_i are i.i.d. samples and if $m = o(logn)$, then any distribution $\mathcal{P}(\pi)$ has to satisfy :

$$\mathbb{P}_{\pi \sim \mathcal{P}^m_{\phi^*,\pi_0}}[d_{TV}(\mathcal{P}(\pi), \mathcal{P}_{\phi^*,\pi_0}) \geq \frac{1}{16}] \geq \frac{1}{3}$$

Proof. We have to consider an appropriate family of distributions in order to apply the Fano's inequality. Pick $\phi = 1/2$. Consider the following collection of permutations with $l = \lfloor n/2 \rfloor$.

$$\pi_1 = (1\ 2), \pi_2 = (3\ 4), ..., \pi_l = ((n-1)\ n)$$

Above we used the cycle notation for the permutation. Informally, the first ordering swaps the two first elements, the second swaps the third and the fourth, etc. Thus,

we study the family $\mathcal{F}$ of distributions parameterized by these permutations and dispersion ϕ^*,

$$\mathcal{F} = \{\mathcal{P}_{\phi,\pi_i}\}_{i=1}^{l}$$

whose size is obviously $\lfloor n/2 \rfloor$.

Now, we have to upper bound KL divergence and to lower bound the TV distance.

For any pair of the above family,

$$D_{KL}(\mathcal{P}_{\phi,\pi_i} \parallel \mathcal{P}_{\phi,\pi_j}) = \sum_{\sigma\in\mathbb{S}_n} \frac{\phi^{d_{KT}(\sigma,\pi_i)}}{Z} ln \frac{\phi^{d_{KT}(\sigma,\pi_i)}}{\phi^{d_{KT}(\sigma,\pi_j)}} = ln(\phi) \sum_{\sigma\in\mathbb{S}_n} \frac{\phi^{d_{KT}(\sigma,\pi_i)}}{Z} (d_{KT}(\sigma,\pi_i) - d_{KT}(\sigma,\pi_j))$$

Hence,

$$D_{KL}(\mathcal{P}_{\phi,\pi_i} \parallel \mathcal{P}_{\phi,\pi_j}) = ln(\frac{1}{\phi}) \mathbb{E}_{\sigma\sim\mathcal{P}_{\phi,\pi_i}} [d_{KT}(\sigma,\pi_j) - d_{KT}(\sigma,\pi_i)]$$

Applying the triangle inequality, we get that

$$d_{KT}(\sigma,\pi_j) \leq d_{KT}(\sigma,\pi_i) + d_{KT}(\pi_i,\pi_j)$$

But, $d_{KT}(\pi_i,\pi_j) = 2$, which indicates the reason we chose that collection of permutations. Thus,

$$D_{KL}(\mathcal{P}_{\phi,\pi_i} \parallel \mathcal{P}_{\phi,\pi_j}) \leq 2ln2$$

since we chose $\phi = \frac{1}{2}$.

In order to lower bound the TV distance, we use the following result from Liu and Moitra, [LM18] :

Lemma 8.5.2 *For any $\pi \neq \sigma \in \mathbb{S}_n$ and any $\phi_1, \phi_2 \in [0, 1-\gamma]$, we have that*

$$d_{TV}(\mathcal{P}_{\phi_1,\pi}, \mathcal{P}_{\phi_2,\sigma}) \geq \frac{\gamma}{2}$$

Proof. Let $\phi_2 \geq \phi_1$. Since $\pi \neq \sigma$, there is at least one pair of elements a, b that $a \succ_\pi b$ and $b \succ_\sigma a$. Hence, the total variation distance is at least the difference between the probabilities that a is ranked higher than b, say this event A, that is :

$$d_{TV}(\mathcal{P}_{\phi_1,\pi}, \mathcal{P}_{\phi_2,\sigma}) \geq |\mathcal{P}_{\phi_1,\pi}(A) - \mathcal{P}_{\phi_2,\sigma}(A)| = \frac{1}{1+\phi_1} - \frac{\phi_2}{1+\phi_2} \geq \frac{\gamma}{2}$$

where the event $A = \{\tau : a \succ_\tau b\}$ is an element of the σ-algebra $\mathcal{F}$ of our probability space and, by definition, the TV distance equals the supremum over the elements of $\mathcal{F}$. ■

Hence, it follows that :

$$d_{TV}(\mathcal{P}_{\phi,\pi_i}, \mathcal{P}_{\phi,\pi_j}) \geq 1/4$$

Also, notice that :

$$ln|\mathcal{F}| = ln(n) - ln2$$

Hence, by Fano's inequality 5.1.3,

$$R_n(\mathcal{F}) \geq \frac{1}{8}(1 - \frac{m2ln2 + ln2}{ln(n) - ln2})$$

If $m = o(logn)$, it follows that $R_n(\mathcal{F}) \geq \frac{1}{16}$ and, consequently, we cannot learn $\mathcal{P}_{\phi,\pi_0}$ ϵ-close in TV distance unless $m = O(logn)$. ■

Lemma 8.5.3 A result similar to the above lemma is the following : *Consider two Mallows models $M_1 = \mathcal{M}(\phi_1, \pi)$ and $M_2 = \mathcal{M}(\phi_2, \pi)$ with the same central ranking on $n \geq 2$ alternatives. If $|\phi_1 - \phi_2| \leq \frac{\sigma^2}{10n^3}$, then $d_{TV}(M_1, M_2) \leq \sigma$.*

8.6 Learning Mallows model in KL Divergence

Finally, we provide a sufficient sample complexity lower bound in order to learn in KL divergence (and as we see in TV distance) in the setting of simple Mallows model. We claim that if enough samples are given, where the samples complexity depends on the permutation size n, on the accuracy parameter ϵ and on the confidence parameter $1 - \delta$, we can learn the distribution of the generating noisy model, with high probability/confidence $1 - \delta$, and with small error both in KL divergence and in TV distance.

Theorem 8.6.1 *For any $\pi_0 \in \mathbb{S}_n$, $\phi \in [0, 1], \epsilon, \delta > 0$ there exist estimators π^*, ϕ^* that can be computed in polynomial time from m i.i.d. samples $\vec{\pi} \sim \mathcal{P}^m_{\phi,\pi_0}$ such that if $m \geq \Omega(\frac{1}{\epsilon^2} log\frac{1}{\delta} + logn)$, then*

$$\mathbb{P}_{\vec{\pi} \sim \mathcal{P}^m_{\phi,\pi_0}}[D_{KL}(\mathcal{P}_{\phi^*,\pi^*} \parallel \mathcal{P}_{\phi,\pi_0}) \leq \epsilon^2] \geq 1 - \delta$$

and hence

$$\mathbb{P}_{\vec{\pi} \sim \mathcal{P}^m_{\phi,\pi_0}}[d_{TV}(\mathcal{P}_{\phi^*,\pi^*}, \mathcal{P}_{\phi,\pi_0}) \leq \epsilon] \geq 1 - \delta$$

Firstly, it is easy to notice that the TV distance result follows from the Pinsker's inequality, mentined in the mathematical foundations chapter 2.2.3. For the sake of completeness, we remind it below :

$$d_{TV}(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{Q})}$$

Hence, if $D_{KL}(\mathbb{P} \parallel \mathbb{Q}) \leq 2\epsilon^2$, it implies that $d_{TV}(\mathbb{P} \parallel \mathbb{Q}) \leq \epsilon$. This is why Pinsker's inequality is very significant in learning theory. Learning well in KL divergence, implies that one also learns well in TV distance. This link is thanks to the above inequality. At the same time, if one cannot learn in TV distance ($d_{TV} \geq \epsilon$), then she cannot even learn in KL divergence ($D_{KL} \geq 2\epsilon^2$). This can be applied to the previous section 8.5.

In order to prove theorem 8.6.1, we have to introduce the notion of exponential families.

Exponential families

Let μ be a measure of $\mathbb{R}^d$, $h : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ and $\mathbf{T} : \mathbb{R}^d \to \mathbb{R}^k$. Suppose that both functions are measurable.

Definition 8.6.1 — Logarithmic partition function. *The logarithmic partition function with parameters $h, \mathbf{T}$ is a mapping $\alpha_{\mathbf{T},h} : \mathbb{R}^k \to \mathbb{R}_{\geq 0}$ where :*

$$\alpha_{\mathbf{T},h}(\eta) = ln \int h(x) exp(\eta^T \mathbf{T}(x)) d\mu(x)$$

The variable η is usually referred as *natural parameters* and is a vector in $\mathbb{R}^k$. We are interested for the space where the logarithmic partition function exists (is finite). Thus, we define the following space :

Definition 8.6.2 — Range of natural parameters. The range of natural parameters for the logarithmic partition function $\alpha_{\mathbf{T},h}$ is the space :

$$\mathcal{H}_{\mathbf{T},h} = \{\eta \in \mathbb{R}^k : \alpha_{\mathbf{T},h}(\eta) < \infty\}$$

Thus, we can define a family of distributions parametrized by η. (that is why η is called natural parameters). This kind of family will be called an exponential family.

Definition 8.6.3 — Exponential family. *The exponential family $\mathcal{E}(\mathbf{T}, h)$ with sufficient statistics $\mathbf{T}$, carrier measure h and natural parameters η is the family of distributions :*

$$\mathcal{E}(\mathbf{T}, h) = \{\mathcal{P}_\eta : \eta \in \mathcal{H}_{\mathbf{T},h}\}$$

where the probability distribution $\mathcal{P}_\eta$ has density :

$$p_\eta(x) = h(x) exp(\eta^T \mathbf{T}(x) - \alpha(\eta))$$

■ **Example 8.1 — Single parameter Mallows Model.** The Mallows $\phi-$distribution is a parametrized distance-based probability distribution that belongs to the family :

$$\mathcal{M}_1 = \{\mathcal{P}_{\phi,\pi_0} : \phi \in [0, 1], \pi_0 \in \mathbb{S}_n\}$$

with probability mass function

$$p_{\phi,\pi_0}(\pi) = \frac{1}{Z(\phi)} \phi^{d_{KT}(\pi,\pi_0)} = \frac{1}{Z(\phi)} e^{d_{KT}(\pi,\pi_0) ln\phi}$$

The parameters correspond to a two dimensional vector (ϕ, π_0). Observe that the family of distributions as stated is not an exponential family because of the central ranking parameter π_0.

If we fix the central ranking, then the family

$$\mathcal{M}_1(\pi_0) = \{\mathcal{P}_\phi : \phi \in [0, 1]\}$$

is an exponential family with natural parameter $\theta = ln\phi$. Then, the sufficient statistic is $T(\pi) = d_{KT}(\pi, \pi_0)$ and the logarithmic partition function is $a(\theta) = lnZ(e^\theta) = lnZ(\phi)$. ■

Lemma 8.6.2 *Let $\mathcal{E}(\mathbf{T}, h)$ be an exponential family with sufficient statistics $\mathbf{T}$ and carrier measure h. For any $\mathcal{P}_\eta \in \mathcal{E}(\mathbf{T}, h)$, let $\mathcal{D}_\eta$ be the distribution of the corresponding sufficient statistics $\mathbf{T}(x)$ when $x \sim \mathcal{P}_\eta$. Then for all $\eta, \eta' \in \mathcal{H}_{\mathbf{T},h}$*

$$d_{TV}(\mathcal{P}_\eta, \mathcal{P}_{\eta'}) = d_{TV}(\mathcal{D}_\eta, \mathcal{D}_{\eta'})$$

$$\&$$

$$D_{KL}(\mathcal{P}_\eta \parallel \mathcal{P}_{\eta'}) = D_{KL}(\mathcal{D}_\eta \parallel \mathcal{D}_{\eta'})$$

Proof. We postpone the proof for the end of the chapter *Proof – 8.6.2*. ■

For any $\alpha, \beta \in \mathbb{R}^d$, let $\mathcal{L}(\alpha, \beta) = \{c \in \mathbb{R}^d : c = p\alpha + (1-p)\beta, p \in [0,1]\}$. For the one dimensional case, this space corresponds to the closed interval $[min(\alpha, \beta), max(\alpha, \beta)]$ and for $d = 2$, this definition corresponds to the parametric representation of a line.

Lemma 8.6.3 *Let $\mathcal{E}(\mathbf{T}, h)$ be an exponential family parametrized by $\eta \in \mathbb{R}^k$ with sufficient statistics $\mathbf{T}$ and carrier measure h. Let α be the logarithmic partition function of the family. For all $\eta \in \mathcal{H}_{\mathbf{T},h}$, it holds that :*

$$\mathbb{E}_{x \sim \mathcal{P}_\eta}[\mathbf{T}(x)] = \nabla\alpha(\eta)$$

$$Var_{x \sim \mathcal{P}_\eta}[\mathbf{T}(x)] = \nabla^2\alpha(\eta)$$

$$\mathbb{E}_{x \sim \mathcal{P}_\eta}[exp(s^T\mathbf{T}(x))] = exp(\alpha(\eta + s) - \alpha(\eta)), \ s \in \mathbb{R}^d$$

Also, for all $\eta, \eta' \in \mathcal{H}_{\mathbf{T},h}$ and for some $\xi \in \mathcal{L}(\eta, \eta')$, it holds that :

$$D_{KL}(\mathcal{P}_\eta \parallel \mathcal{P}_{\eta'}) = -(\eta' - \eta)^T\nabla\alpha(\eta) + \alpha(\eta') - \alpha(\eta) = (\eta' - \eta)^T\nabla^2\alpha(\xi)(\eta' - \eta)$$

Proof. We postpone the proof for the end of the chapter *Proof – 8.6.3* ■

We continue with the proof of the main theorem.

Proof. 1. Observe that we can use $O(log\frac{n}{\delta})$ samples to learn the central ranking π_0.

2. We remind that ϕ is the true unknown dispersion, that we want to estimate. Once we know the central ranking, we can assume that our samples are coming from the distribution $\mathcal{P}_\phi$ and that we want to learn $\mathcal{P}_\phi$ in KL divergence. This is due to the lemma 7.5.2

3. Applying the lemma 8.6.2, we can assume sample access to the distribution $\mathcal{D}_\phi$ of the sufficient statistics of $\mathcal{P}_\phi$.

4. We have that $\mathcal{D}_\phi$ is a distribution in a single parameter exponential family with natural parameter $\theta = ln\phi$. Let α be the logarithmic partition function of the family.

5. From lemma 8.6.3, the KL divergence between a distribution $\mathcal{D}_{\phi'}$ parametrized by a dispersion parameter ϕ' and the true distribution $\mathcal{D}_\phi$ equals :

$$D_{KL}(\mathcal{D}_{\phi'} \parallel \mathcal{D}_\phi) = -(\theta' - \theta)\dot{\alpha}(\theta) + \alpha(\theta') - \alpha(\theta)$$

This KL divergence can be seen as a function of θ'. Consider the function

$$f(x) = -(x - \theta)\dot{\alpha}(\theta) + \alpha(x) - \alpha(\theta)$$

6. Analyzing the function f, it is easy to see that f is convex with minimum at $x = \theta$. Hence, $f \downarrow \{x \leq \theta\}$ and $f \uparrow \{x \geq \theta\}$.

7. Note that $\alpha(\theta) = lnZ(e^\theta) = lnZ(\phi) \geq 0$, $\forall \theta \in (-\infty, 0]$, since $Z(\phi) \geq 1$. Then it is easy to observe that :

$$\lim_{x \to -\infty} f(x) = +\infty = \lim_{\phi' \to 0} D_{KL}(\mathcal{D}_{\phi'} \parallel \mathcal{D}_\phi)$$

and

$$\lim_{\phi' \to \infty} D_{KL}(\mathcal{D}_{\phi'} \parallel \mathcal{D}_\phi) = +\infty$$

when $\phi < \infty$.

8. Define the set

$$Q = \{\theta \in \mathbb{R} : D_{KL}(\mathcal{D}_{\phi'} \parallel \mathcal{D}_\phi) \leq \epsilon\}$$

Since f is convex, the space Q is an interval s.t. $\theta = ln\phi \in Q$. Define

$$\theta^- = infQ,\ \theta^+ = supQ$$

Because of the limits observed before, Q is a closed interval with $Q = [\theta^-, \theta^+] \subset (-\infty, +\infty)$. Define $\phi^{\{-,+\}} = exp(\theta^{\{-,+\}})$.

9. From the above, we have that :

$$D_{KL}(\mathcal{D}_{\phi^{\{-,+\}}} \parallel \mathcal{D}_\phi) = \epsilon$$

10. Once the central ranking is known the distribution is an exponential family and let $T(\pi)$ be its sufficient statistics. From the lemma 8.6.3, we know that $h(\theta) = \mathbb{E}_{\pi \sim \mathcal{P}_{\phi', \pi_0}}[T(\pi)] = \dot{\alpha}(\theta)$. Thus, h is an increasing function with respect to θ. and the better we estimate $\mathbb{E}_{\pi \sim \mathcal{P}_{\phi', \pi_0}}[T(\pi)]$, the better we estimate ϕ. So, the estimation of the true parameter $\theta = ln\phi$ is equivalent to the estimation of the image of the function h.

11. Since h is injective, given any real number r in the image of h, we can find θ^* s.t. $|\theta(r) - \theta^*| \leq \epsilon$, where $\theta(r)$ is well defined from the equation $h(\theta(r)) = r$.

12. Suppose that we sample $\vec{\pi} \sim \mathcal{P}^m_{\phi, \pi_0}$. In order to get an estimation for the true θ, it suffices to find a real value $r(\vec{\pi})$ s.t.

$$h(\theta(r(\vec{\pi}))) = r(\vec{\pi}) \wedge |\theta - \theta(r(\vec{\pi}))| \leq \epsilon.$$

13. We have to choose an estimator for r. Notive that $h(\theta)$ is expressed as an expected value. Thus, it seems logical to choose

$$r = \frac{1}{m}\sum_{i=1}^{m} T_i(\pi_i)$$

We need to bound the probability that this estimation is far from the expected value of the sufficient statistic T when drawing a random sample, that is :

$$p = \mathbb{P}_{\vec{\pi}\sim\mathcal{P}^m_{\phi,\pi_0}}[\frac{1}{m}\sum_{i=1}^{m} T_i(\pi_i) \geq \mathbb{E}_{\pi\sim\mathcal{P}_{\phi',\pi_0}}[T(\pi)]]$$

We will try to prove the $\geq$ case. The case $\leq$ can be handled similarly.

14. We will use the Markov's inequality to upper bound p. Choose $s > 0$. Then, we get that :

$$p = \mathbb{P}_{\vec{\pi}\sim\mathcal{P}^m_{\phi,\pi_0}}[exp(s\cdot\sum_{i=1}^{m} T_i(\pi_i)) \geq exp(s\cdot m\cdot\mathbb{E}_{\pi\sim\mathcal{P}_{\phi',\pi_0}}[T(\pi)])] \leq \frac{\mathbb{E}_{\pi_i\sim\mathcal{P}_{\phi,\pi_0}}[exp(s\cdot\sum_{i=1}^{m} T_i(\pi_i))]}{exp(s\cdot m\cdot\mathbb{E}_{\pi\sim\mathcal{P}_{\phi',\pi_0}}[T(\pi)])}$$

By the independence of our samples, we have that :

$$p \leq (\frac{\mathbb{E}_{\sigma\sim\mathcal{P}_{\phi,\pi_0}}[exp(s\cdot T(\sigma))]}{exp(s\cdot\mathbb{E}_{\pi\sim\mathcal{P}_{\phi',\pi_0}}[T(\pi)])})^m$$

From lemma 8.6.3, the RHS becomes :

$$p \leq exp(-m(s\dot{\alpha}(\phi') - \alpha(\phi+s) + \alpha(\phi)))$$

Now, we have to find the minimum value for the RHS by seeing it as a function of s. It is not difficult to see that we get the optimal bound for $s = \phi' - \phi$. Hence, again by lemma 8.6.3,

$$p \leq exp(-m\cdot D_{KL}(\mathcal{P}_{\phi'} \parallel \mathcal{P}_{\phi}))$$

15. Thus, for any upper and lower estimations θ^- and $\theta^+ \leq 0$,

$$\mathbb{P}_{\vec{\pi}\sim\mathcal{P}^m_{\theta,\pi_0}}[r(\vec{\pi}) \notin [h(\theta^-), h(\theta^+)]] \leq 2exp(-m\min_{\theta^*\in\{\theta^-,\theta^+\}} D_{KL}(\mathcal{P}_{\theta^*} \parallel \mathcal{P}_{\theta}))$$

or equivalently, by the monotonicity of h,

$$\mathbb{P}_{\vec{\pi}\sim\mathcal{P}^m_{\theta,\pi_0}}[\theta(r(\vec{\pi})) \notin [\theta^-, \theta^+]] \leq 2exp(-m\min_{\theta^*\in\{\theta^-,\theta^+\}} D_{KL}(\mathcal{P}_{\theta^*} \parallel \mathcal{P}_{\theta}))$$

16. From step *9*,

$$\mathbb{P}_{\vec{\pi}\sim\mathcal{P}^m_{\theta,\pi_0}}[\theta(r(\vec{\pi})) \notin Q] \leq 2exp(-m\cdot\epsilon)$$

Now, we pick the estimation $\phi(r(\vec{\pi})) = exp(\theta(r(\vec{\pi})))$ and we get that :

$$\mathbb{P}_{\vec{\pi}\sim\mathcal{P}^m_{\theta,\pi_0}}[D_{KL}(\mathcal{D}_{\phi(r(\vec{\pi}))} \parallel \mathcal{D}_\phi) \geq \epsilon] \leq 2exp(-m \cdot \epsilon)$$

and going back to the distribution $\mathcal{P}$:

$$\mathbb{P}_{\vec{\pi}\sim\mathcal{P}^m_{\theta,\pi_0}}[D_{KL}(\mathcal{P}_{\phi(r(\vec{\pi})),\pi_0} \parallel \mathcal{P}_{\phi,\pi_0}) \geq \epsilon] \leq 2exp(-m \cdot \epsilon)$$

Hence, we want $2exp(-m \cdot \epsilon) \leq \delta$ and we solve for the number of samples m. To that result, we have to add the samples we need to learn the central ranking.

By reversing the result, we get that with $m \geq \Omega(\frac{1}{\epsilon^2}log\frac{1}{\delta} + logn)$, we get the desired bound :

$$\mathbb{P}_{\vec{\pi}\sim\mathcal{P}^m_{\phi,\pi_0}}[D_{KL}(\mathcal{P}_{\phi^*,\pi^*} \parallel \mathcal{P}_{\phi,\pi_0}) \leq \epsilon^2] \geq 1 - \delta$$

■

8.7 Appendix

Let $\mathcal{E}(\mathbf{T}, h)$ be an exponential family with sufficient statistics $\mathbf{T}$ and carrier measure h. For any $\mathcal{P}_\eta \in \mathcal{E}(\mathbf{T}, h)$, let $\mathcal{D}_\eta$ be the distribution of the corresponding sufficient statistics $\mathbf{T}(x)$ when $x \sim \mathcal{P}_\eta$. Then for all $\eta, \eta' \in \mathcal{H}_{\mathbf{T},h}$

$$d_{TV}(\mathcal{P}_\eta, \mathcal{P}_{\eta'}) = d_{TV}(\mathcal{D}_\eta, \mathcal{D}_{\eta'})$$

$$\&$$

$$D_{KL}(\mathcal{P}_\eta \parallel \mathcal{P}_{\eta'}) = D_{KL}(\mathcal{D}_\eta \parallel \mathcal{D}_{\eta'})$$

Proof. We will only prove for the TV distance for the discrete case. The proof for the KL divergence and for continuous distributions is similar. Let S be the support of the exponential family, let $S_{\mathbf{T}} = \{y|\exists x \in S : \mathbf{T}(x) = y\}$ be the range of sufficient statistics $\mathbf{T}$ and $N_y = \sum_{x\in S} \mathbb{1}\{\mathbf{T}(x) = y\}$.

Then,

$$d_{TV}(\mathcal{P}_\eta, \mathcal{P}_{\eta'}) = \frac{1}{2}\sum_{x\in S}|p_\eta(x) - p_{\eta'}(x)| = \frac{1}{2}\sum_{x\in S}|h(x)exp(\eta^T\mathbf{T}(x) - \alpha(\eta)) - h(x)exp(\eta'^T\mathbf{T}(x) - \alpha(\eta'))|$$

The summation over the support is equivalent to the following :

$$d_{TV}(\mathcal{P}_\eta, \mathcal{P}_{\eta'}) = \frac{1}{2}\sum_{y\in S_{\mathbf{T}}}\sum_{x:\mathbf{T}(x)=y}|h(x)exp(\eta^T\mathbf{T}(x) - \alpha(\eta)) - h(x)exp(\eta'^T\mathbf{T}(x) - \alpha(\eta'))|$$

$$= \frac{1}{2}\sum_{y\in S_{\mathbf{T}}} N_y|h(x)exp(\eta^T\mathbf{T}(x) - \alpha(\eta)) - h(x)exp(\eta'^T\mathbf{T}(x) - \alpha(\eta'))| =$$

$$= \frac{1}{2}\sum_{y\in S_{\mathbf{T}}}|N_y h(x)exp(\eta^T\mathbf{T}(x) - \alpha(\eta)) - N_y h(x)exp(\eta'^T\mathbf{T}(x) - \alpha(\eta'))|$$

But this is exactly the TV distance of the sufficient statistics distributions. Thus,

$$d_{TV}(\mathcal{P}_\eta, \mathcal{P}_{\eta'}) = \frac{1}{2}\sum_{y \in S_{\mathbf{T}}} |d_\eta(y) - d_{\eta'}(y)| = d_{TV}(\mathcal{D}_\eta, \mathcal{D}_{\eta'}).$$

■

Let $\mathcal{E}(\mathbf{T}, h)$ be an exponential family parametrized by $\eta \in \mathbb{R}^k$ with sufficient statistics $\mathbf{T}$ and carrier measure h. Let α be the logarithmic partition function of the family. For all $\eta \in \mathcal{H}_{\mathbf{T},h}$, it holds that :

$$\mathbb{E}_{x \sim \mathcal{P}_\eta}[\mathbf{T}(x)] = \nabla\alpha(\eta) \tag{8.5}$$

Also, for all $\eta, \eta' \in \mathcal{H}_{\mathbf{T},h}$ and for some $\xi \in \mathcal{L}(\eta, \eta')$, it holds that :

$$D_{KL}(\mathcal{P}_\eta \parallel \mathcal{P}_{\eta'}) = -(\eta' - \eta)^T \nabla\alpha(\eta) + \alpha(\eta') - \alpha(\eta) = (\eta' - \eta)^T \nabla^2\alpha(\xi)(\eta' - \eta) \tag{8.6}$$

Proof. The log-partition function is defined as :

$$\alpha(\eta) = log \int h(x) exp(\eta^T \mathbf{T}(x)) d\mu(x)$$

In the range of natural parameters, $g(\eta) = e^{\alpha(\eta)}$ is continuous and has continuous derivatives of all orders, which can be computed under the integration sign. Let $\eta \in \mathbb{R}^k$. Then,

$$e^{\alpha(\eta)} \frac{\partial\alpha(\eta)}{\partial\eta_j} = \int \frac{\partial}{\partial\eta_j} h(x) exp[\sum_{i=1}^{k} \eta_i T_i(x)] d\mu(x) = \int T_j(x) h(x) exp[\sum_{i=1}^{k} \eta_i T_i(x)] d\mu(x)$$

But, we have that $p_\eta(x) = h(x) exp(\eta^T \mathbf{T}(x) - \alpha(\eta))$. Thus, for a single dimension,

$$\frac{\partial\alpha(\eta)}{\partial\eta_j} = \int T_j(x) p_\eta(x) d\mu(x) = \mathbb{E}_{x \sim \mathcal{P}_\eta}[T_j(x)]$$

So,

$$\nabla\alpha(\eta) = \mathbb{E}_{x \sim \mathcal{P}_\eta}[\mathbf{T}(x)]$$

For equation (8.6), one has :

$$D_{KL}(\mathcal{P}_\eta \parallel \mathcal{P}_{\eta'}) = \int p_\eta(x) ln \frac{p_\eta(x)}{p_{\eta'}(x)} d\mu(x)$$

$$= \int p_\eta(x)((\eta - \eta')^T \mathbf{T}(x) + \alpha(\eta') - \alpha(\eta)) d\mu(x) = (\eta - \eta')^T \mathbb{E}_{x \sim \mathcal{P}_\eta} \mathbf{T}(x) + \alpha(\eta') - \alpha(\eta)$$

But, from (8.5), $D_{KL}(\mathcal{P}_\eta \parallel \mathcal{P}_{\eta'}) = -(\eta' - \eta)^T \nabla\alpha(\eta) + \alpha(\eta') - \alpha(\eta)$.

From the multidimensional Taylor's theorem, there is some $\xi \in \mathcal{L}(\eta, \eta')$ s.t.

$$D_{KL}(\mathcal{P}_\eta \parallel \mathcal{P}_{\eta'}) = (\eta' - \eta)^T \nabla^2\alpha(\xi)(\eta' - \eta)$$

which completes the proof. ■

9. Finding the maximum likelihood ranking

In this chapter, we will ignore the sample complexity and focus on the idea of solving the maximum likelihood ranking estimator problem. That is, given r samples drawn from the hidden central ranking, we have to find which element from the symmetric group maximizes the probability of being given those r permutations. We will refer to the solution of the MLE as the maximum likelihood permutation.

9.1 The goal, a technique and a promise

Goal

Suppose that we are given r i.i.d. samples drawn from a Mallows model. Our goal is to find the maximum likelihood permutation $\hat{\pi}^*$ given the samples observed, that is :

$$\hat{\pi}^* = arg\max_{\pi^*} \prod_{i=1}^{r} \mathbb{P}[\pi_i|\pi^*] = arg\max_{\pi^*} \prod_{i=1}^{r} \frac{e^{-\beta d_{KT}(\pi_i,\pi^*)}}{Z(\beta)} \tag{9.1}$$

But, thanks to the exponential structure of the model, the above product can be converted into a sum :

$$\hat{\pi}^* = arg\max_{\pi^*} \prod_{i=1}^{r} \frac{e^{-\beta d_{KT}(\pi_i,\pi^*)}}{Z(\beta)} = arg\max_{\pi^*} \frac{1}{Z(\beta)^r} exp(-\beta \cdot \sum_{i=1}^{r} d_{KT}(\pi_i, \pi^*))$$

Firstly, we can just ignore the normalization constant for the optimization problem. Also, using the fact that $x \mapsto lnx$ is an increasing function, the problem reduces to the following :

$$\hat{\pi}^* = arg\max_{\pi^*} lne^{-\beta \cdot \sum_{i=1}^{r} d_{KT}(\pi_i,\pi^*)} = arg\max_{\pi^*}(-\beta \cdot \sum_{i=1}^{r} d_{KT}(\pi_i, \pi^*))$$

Hence, since $\beta > 0$,

$$\hat{\pi}^* = arg\min_{\pi^*} \sum_{i=1}^{r} d_{KT}(\pi_i, \pi^*) \tag{9.2}$$

The Kemeny's voting rule is the Maximum Likelihood Estimator $\hat{\pi}^*$ for the underlying hidden truth of the Mallows model.

Technique

The problem of finding the MLE given r samples is reduced to the one of finding the median for r permutations in the metric space $(\mathbb{S}_n, d_{KT})$. One could think geometrically that the permutations create a polytope inside that metric space. It is well known that we cannot attack this problem in a straightforward way since the problem is NP-hard, as mentioned in a previous chapter $NP - hard - Kemeny$. Of course, the solution to the problem will be one of the $n!$ permutations in the symmetric group. We have to choose the correct one. Obviously, an exhaustive search approach is not a good choice since, besides the computational inefficiency, we do not exploit the knowledge offered from the r samples. It would be a good idea to somehow deduce a ranking from these r samples that will be 'close' to the one that we are looking for. An obvious idea would be to aggregate those rankings. This is exactly the technique that we will use. Specifically, we are going to create the so-called average permutation $\bar{\pi}$, that will map each alternative to the position she appeared on average in these r samples. We will break ties uniformly. We will show that this average permutation is not completely bad. Afterwards, the idea is that we will find a value ρ and create a ball $\mathcal{B}(\bar{\pi}, \rho)$ of center $\bar{\pi}$ and radius ρ in the metric space $(\mathbb{S}_n, d_{KT})$. This ball will contain with high probability the maximum likelihood permutation. Our exhaustive search will be executed only inside this ball, whose size will be significantly smaller than the order of $n!$ and the search will be quite fast.

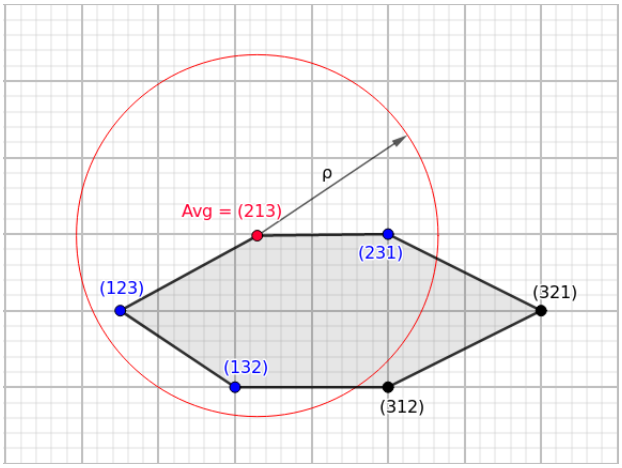

Figure 9.1: Slicing the solution space of the $\mathbb{S}_3-$permutohedron with a ball $\mathcal{B}(\bar{\pi}, \rho)$.

For instance, in order to visualize the above technique, if we sample rankings of size 3, the permutohedron will be the 2-dimensional convex hull of the 6 points.

Suppose that the average permutation is the ranking (213). Then, the solution space will be reduced to the collection of rankings only inside the circle with center the average ranking and appropriate radius ρ, that will be explained in the next sections. The visualization is provided in the above figure.

Promise

The main result, trying to solve the Mallows Reconstruction Problem, is the following. One can compute the maximum likelihood permutation, with high probability, in time $T(n)$, where n is the size of the permutations.

Theorem 9.1.1 *There exists a randomized algorithm such that if $\{\pi_i\}_{i=1}^r$ be rankings on n elements independently generated by Mallows model with parameter $\beta > 0$, and let $\alpha > 0$. Then a maximum probability order π^m can be computed in time :*

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

and error probability $< n^{-\alpha}$.

The parameter α controls the error probability. As α increases, the error probability falls, and, consequently, the time $T(n)$ increases.

As r grows, the algorithm tends to almost linear. This remarkably different than anything one sees in other fields of algorithms. In classical algorithmic theory, as the input grows, the algorithm usually becomes slower. Now, in the theoretical machine learning concept, as the input grows and, thus, the provided information is larger, one can note that the algorithm accelerates as the number of samples grows.

In the next section, we will prove that one can find the MLE in a computationally efficient way. Then, another question naturally arises.

Suppose that we have found the MLE ranking $\hat{\pi}^*$. How close it will be to the original ranking π_0?

9.2 Mallows' Reconstruction Problem

We will begin with some notation :

- Each permutation has size n and the set of elements of the permutation A is isomorphic to $[n]$.
- π_0 is the initial hidden ranking and $\beta = ln\frac{1}{\phi}$, where ϕ is the (unknown) dispersion.
- $\{\pi_i\}_{i=1}^r$ are the r noisy samples drawn from the Mallows model $\mathcal{M}(\pi_0, \beta)$ with distribution $\mathcal{P}_{\pi_0,\beta}$
- $\overline{\pi} = Avg(\pi)$ denotes the 'average' ranking, which we will explain later.
- $\hat{\pi}^*$ is the MLE ranking we want to find.

Under the Mallows probabilistic model $\mathcal{M}(\pi_0, \beta)$, the probability of drawing the ranking π equals :

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

9.2.1 Computing the MLE ordering

As a first step, we show that, under this model, the locations of individual elements $j \in [n]$ are distributed geometrically. That is, the probability that the element j is transposed 'far away' from its original positions decreases exponentially, as the length of the transposition grows linearly.

Thanks to the relabeling property discussed in the introductory chapter, one can assume that $\pi_0 = id$.

Lemma 9.2.1 *Suppose that $\pi_0 = id$ and let $k \in [n]$. Obviously, $\pi_0(k) = k$. Then, for a ranking $\pi \sim \mathcal{P}_{\pi_0,\beta}$, we have that :*

$$\mathbb{P}[|\pi(k) - k| \geq i] < 2\frac{e^{-\beta i}}{(1 - e^{-\beta})}, \forall i.$$

Proof. From the RIM process, it is already known that π can be sampled by inserting the elements $1, ..., n$ into the ordering one-by-one, each time conditioning on the order so far. Hence, for the kth element, suppose we have sampled the relative ranking of the first $(k-1)$ alternatives under π and we want to insert k. By the definition of the Mallows model, the probability that k is mapped to position $k - i$ is bounded by $e^{-\beta i}$. This indicates the truncated geometric distribution we have already mentioned. During the insertion of the elements $k+1, ..., n$, the location of k may only increase. Hence,

$$\mathbb{P}[\pi(k) \leq k - i] < \sum_{j=i}^{\infty} e^{-\beta j} = \frac{e^{-\beta i}}{1 - e^{-\beta}}$$

Hence, by symmetry of the dislocation, we get the factor 2 for the wanted upper bound. ■

Secondly, we construct the 'average' ranking given r samples and create a similar lemma for the locations of the individual elements of the average permutation.

Suppose that the permutations $\pi_1, ..., \pi_r \sim \mathcal{P}^r_{\pi_0,\beta}$. Consider $k \in [n]$ and let $\pi_0 = id$. Let $\overline{\pi(k)}$ be the average index of k under the samples drawn, that is :

$$\overline{\pi(k)} = \frac{1}{r}\sum_{i=1}^{r} \pi_i(k).$$

Lemma 9.2.2 *Suppose that $\pi_0 = id$ and let $k \in [n]$. Then, for the average ranking $\overline{\pi}$, constructed by the aggregation of r samples $\pi_1, ..., \pi_r \sim \mathcal{P}^r_{\pi_0,\beta}$, we have that :*

$$\mathbb{P}[|\overline{\pi(k)} - k| \geq i] \leq 2(\frac{(5i+1)e^{-\beta i}}{1 - e^{-\beta}})^r, \forall i.$$

Proof. We will study the item k. In each one of the r samples, suppose that its dislocation is at most d_i, for $i \in [r]$. We choose $d_i \geq 0$. Consider the dislocation

vector $\vec{d} = (d_1, ..., d_r)$. We consider the event $E(\vec{d})$ that is $\pi_i(k) \leq k - d_i$, for $i \in [r]$. From the previous lemma, we get that :

$$\mathbb{P}[E(\vec{d})] \leq \frac{exp(-\beta \sum_{i=1}^{r} d_i)}{(1 - e^{-\beta})^r}$$

We have that, for the event $D_i = [\overline{\pi(k)} \leq k - i]$:

$$D_i \subset \bigcup_{\sum_{j=1}^{r} d_j = ri} E(\vec{d}) \Rightarrow \mathbb{P}[\overline{\pi(k)} \leq k - i] < \mathbb{P}[\bigcup_{\sum_{j=1}^{r} d_j = ri} E(\vec{d})]$$

By union bound, we get that :

$$\mathbb{P}[\overline{\pi(k)} \leq k - i] < \#\{\vec{d} : \sum_{j=1}^{r} d_j = ri\} \frac{e^{-\beta ri}}{(1 - e^{-\beta})^r}$$

But, the cardinality of this set is exactly the number of ways to place ri balls into r bins, that equals $\binom{ri+r-1}{r-1}$. Hence,

$$\mathbb{P}[\overline{\pi(k)} \leq k - i] < \binom{ri + r - 1}{r - 1} \frac{e^{-\beta ri}}{(1 - e^{-\beta})^r}$$

and, using a known binomial coefficient inequality, we get that :

$$\mathbb{P}[\overline{\pi(k)} \leq k - i] < (5i + 1)^r \frac{e^{-\beta ri}}{(1 - e^{-\beta})^r}$$

Working for the symmetric event $D_i' = [\overline{\pi(k)} \geq k + i]$, we get the desired 2 factor and the desired concentration bound for the average ranking. ■

We assume that r is fixed. From the above lemma, we can easily get the following result, that bounds the probability of the 'bad' event, that is that it will exist some element k that, in the average ranking, it will make a jump of length at least $\Theta(logn)$. Thus, with high probability, each element in the average ranking will be $\Theta(logn)-$close to its original position.

Lemma 9.2.3 *Let $\pi_0 = id$ and let $\alpha > 0$. Fix the number of samples r. Then, for sufficiently large n,*

$$\mathbb{P}[\exists k : |\overline{\pi(k)} - k| \geq \frac{\alpha + 2}{\beta r} logn] < n^{-\alpha} \tag{9.3}$$

Note that the error margin for each element decreases proportionally to r.

We have shown that, with high probability, the average ranking will be close to the original ranking.

Result 1 : The average ranking $\overline{\pi}$ is more likely $\Theta(logn)-$close to the original π_0.

We continue by showing that the maximum likelihood ranking is close to the original. Suppose that we would be able to prove that, with high probability, the MLE $\bar{\pi}^*$ is $\Theta(logn)-$close to the original π_0. Then, it would be $\Theta(logn)-$close to the average permutation too. This is our goal and we will try to prove it afterwards. A good question would be why prove that? How this will help find the MLE? The idea is that we could get the MLE ranking from the average ranking, using as a black box a sorting algorithm. Specifically, there is an dynamic programming algorithm, that given a pre-sorted ranking, can sort it fast to a desired one. The notion of pre-sorting corresponds to the idea that each element in the given ranking will be at most k positions away from the correct position. Note that, if $k = \Theta(logn)$, we could say that the average ranking is a pre-sorted ranking for the MLE goal permutation. This idea is not yet completely clear, but it will be soon. But it must be clear that we should prove that the maximum likelihood ranking is close to the original, and, hence, to the average one.

Before proceeding we introduce some notation. We define the score of a permutation as the value it attains when used on the MLE, as follows :

Definition 9.2.1 — Ranking score metric. *The Mallows reconstruction problem can be restated as follows :*

$$\hat{\pi}^* = arg\min_{\pi^*} \sum_{i=1}^{r} d_{KT}(\pi_i, \pi^*) = arg\min_{\pi^*} \sum_{i<_{\pi^*} j} |\{k : \pi_k(i) > \pi_k(j)\}| \tag{9.4}$$

Thus, we try to minimize the pairwise disagreements. We will denote with :

$$q(i < j) = |\{k : \pi_k(i) < \pi_k(j)\}|$$

Then, MRP is equivalent to :

$$\hat{\pi}^* = arg\max_{\pi^*} \; Score(\pi^*) = arg\max_{\pi^*} s(\pi^*) = arg\max_{\pi^*} \sum_{i<_{\pi^*} j} q(i < j) \tag{9.5}$$

Hence, we try to maximize the pairwise agreements.

For simplicity, we will let $\pi_0 = id$ and let

$$L = max(6\frac{\alpha+2}{\beta r}logn, 6\frac{\alpha+2+\frac{1}{\beta}}{\beta})$$

Intuitively, L controls the margin that we proved before for the length of the jump of an element under the average permutation $\overline{\pi}$. Also, consider a error parameter $a > 0$.

Lemma 9.2.4 *Except with probability n^{-a} we have that for any i, j s.t. $j - i \geq L$,*

$$q(i < j) > \frac{2r}{3}$$

that is, less than $1/3$ *of the permutations* $\pi_1, ..., \pi_r$ *order* i *and* j *incorrectly.*

Proof. Suppose that $j - i \geq L$. We study the probability that a sample ranking $\pi_v, v \in [r]$ swaps this pair.

$$\mathbb{P}[j \succ_{\pi_v} i] = \mathbb{P}[\pi_v(j) \leq \pi_v(i)]$$

Now, notice that in order to swap these elements, it must hold either the event $[\pi_v(j) \leq j - \frac{L}{2}]$ or the event $[\pi_v(i) \geq i + \frac{L}{2}]$. If neither of them holds, the swap is impossible. By union bound, we get :

$$\mathbb{P}[j \succ_{\pi_v} i] \leq \mathbb{P}[\pi_v(j) \leq j - \frac{L}{2}] + \mathbb{P}[\pi_v(i) \geq i + \frac{L}{2}]$$

By the lemma 9.2.1, we can upper bound each term :

$$\mathbb{P}[j \succ_{\pi_v} i] \leq 2\frac{e^{-\beta\frac{L}{2}}}{(1 - e^{-\beta})} \leq n^{-3(a+1)/r}$$

for sufficiently large n. There are two cases :

- If $r \leq logn$, the probability of having at least $\frac{r}{3}$ samples having swapped i, j is bounded by $n^{-(a+1)}2^r < n^{-a}$.
- If $r > logn$, we have that : $\mathbb{P}[j \succ_{\pi_v} i] \leq n^{-3(a+1)}$ and, hence, the probability of having at least $\frac{r}{3}$ samples having swapped i, j is bounded by $n^{-3(a+1)\frac{r}{3}}2^r < e^{-ar} < n^{-a}$.

■

Now, we are capable of analyzing the proximity of the MLE $\hat{\pi}^*$ to the original ranking π_0.

Lemma 9.2.5 *Except with probability* $< 2n^{-a}$*, for any optimal* $\hat{\pi}^*$ *and for all* k*, we have*

$$|\hat{\pi}^*(k) - \pi_0(k)| \leq 32L$$

Proof. Let $p_0 = id$. Firstly, we make the assumption that our samples $\{\pi_i\}_{i=1}^r$ satisfy the previous lemma with probability $< n^{-a}$. Suppose that $\exists k : |\hat{\pi}^*(k) - k| = M > 32L$. We will get a contradiction.

Without loss of generality, let $\hat{\pi}^*(k) = k + M$. Our goal is to find a permutation that scores higher than the MLE optimal and, thus, get a contradiction.

Let $T \geq M/4 - L > 7L$. We will show that there must be at least T i's from below k that are mapped above the k-th position by $\hat{\pi}^*$. That is :

$$|i : i < k \cap \hat{\pi}^*(i) \geq k| \geq T$$

Define the set of items that are mapped between position k and $k + M$ by the MLE optimal, $S := \{j : k \leq \hat{\pi}^*(j) < k + M\}$.

It must hold :

$$\sum_{j \in S}(q(j < k) - q(j > k)) > 0 \tag{9.6}$$

Otherwise, the permutation that maps k to k scores better.

Now, we partition S into three disjoint subsets $S = S_1 \cup S_2 \cup S_3$, where :

- $S_1 \to (j < k), |S_1| < T$
- $S_2 \to (k < j < k + L), |S_2| < L$
- $S_3 \to (j \geq k + L)$

Now, we study the above equation (9.6), by breaking the sum into sums over the three partitions :

$$\sum_{j \in S}(q(j < k) - q(j > k)) = \sum_{j \in S_1 \cup S_2 \cup S_3}(q(j < k) - q(j > k))$$

$$< r|S_1| + r|S_2| - \frac{r}{3}|S_3| < r(T + L) - \frac{r}{3}(M - T - L) \Rightarrow T > 7L$$

So, we get the desired :

$$|i : i < k \cap \hat{\pi}^*(i) \geq k| \geq T \geq M/4 - L > 7L$$

So, there must exist at least T i's with $i \geq k$ and $\hat{\pi}^*(i) < k$. We, then, define the sets $T_1 = \{i < k : \hat{\pi}^*(i) \geq k\}$ and $T_2 = \{i \geq k : \hat{\pi}^*(i) < k\}$.

We now are able to create a ranking π^m obtained by the OPT by concatenating its restriction to $\{1, ..., k-1\}$ with its restriction to $\{k, ..., n\}$. Next, we count the pairs $i < j$ on which the two permutations disagree.

- *Case A :* $|i - j| < L$. To get a disagreement, either i or j has to belong to $T_1 \cup T_2$ and in each case we have at most L choice for the other. So, $|P_1| < 2TL$.
- *Case B :* $|i - j| \geq L$. Note that $q(i < j) > 2r/3$. Each $t \in T_1$ participated in such a pair with each $t' \in T_2$ except $|t - t'| < L$. Thus, $|P_2| \geq T(T - L)$.

Finally, we show that π^m will score higher than the MLE optimal.

$$s(\pi^m) - s(\hat{\pi}^*) = \sum_{P_1}(q(i < j) - q(j < i)) - \sum_{P_2}(q(i < j) - q(j < i)) >$$

$$> (-r)|P_1| + (r/3)|P_2| \geq (-r)(2TL) + (r/3)T(T - L)$$

But, we have that $T > 7L$ and, hence, we get that :

$$s(\pi^m) > s(\hat{\pi}^*)$$

that is a contradiction. ■

Result 2 : The MLE ranking $\hat{\pi}^*$ is more likely $\Theta(logn)-$close to the original π_0.

Thus, combining results 1 and 2, one gets that :

Result 3 : The MLE ranking $\hat{\pi}^*$ is, with high probability, $\Theta(logn)-$close to the average ranking $\overline{\pi}$.

Now, we are able to perform the pre-sorting trick that we mentioned before. Our goal is to find the MLE ranking. We know the average permutation $\overline{\pi}$ and, additionally, we proved something remarkably useful. We have shown that these two permutations are close. That is we can create a ball $\mathcal{B}(\bar{\pi}, \rho)$ of center $\overline{\pi}$ and radius ρ in the metric space $(\mathbb{S}_n, d_{KT})$. We will pick radius $\rho = \Theta(logn)$. This ball will contain with high probability the maximum likelihood permutation $\bar{\pi}^*$, that we want to find.

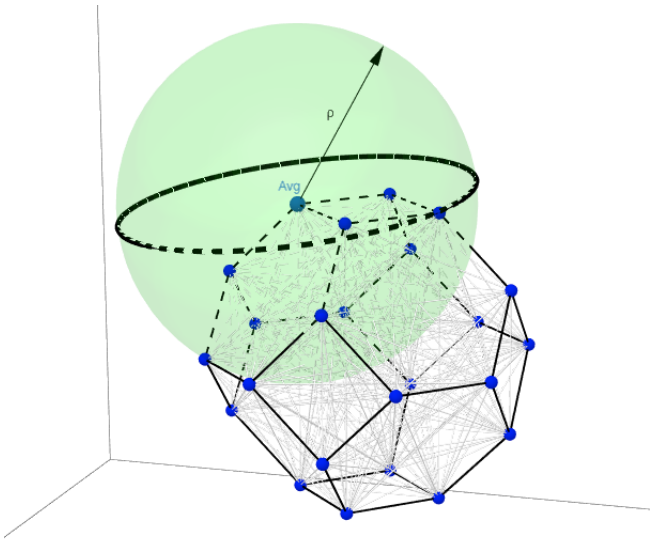

Figure 9.2: Ball $\mathcal{B}(\overline{\pi}, \rho)$ reducing the solution space of the $\mathbb{S}_4$-permutohedron.

A $3D$ visualization of this concept is provided in the above figure. Also, we provide a layout of the figure, that is a projections to the xy plane.

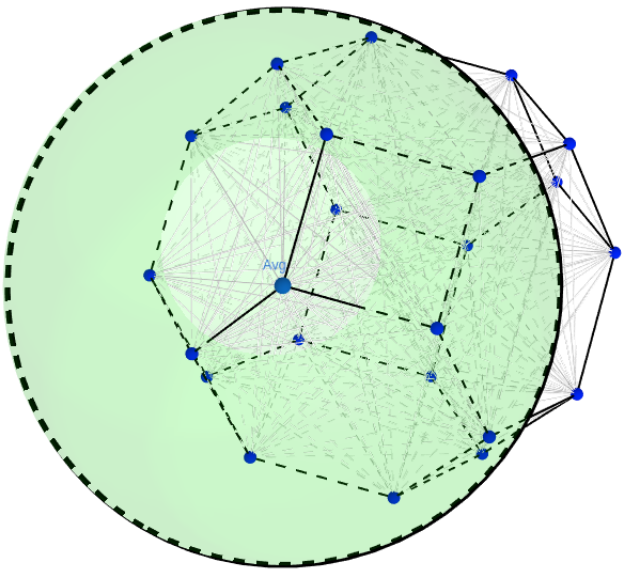

Figure 9.3: xy-projection of the ball $\mathcal{B}(\overline{\pi}, \rho)$ and of the $\mathbb{S}_4$-permutohedron.

Now, we will execute an exhaustive search only inside this ball. Our search will try to maximize the score function defined above. The permutation that maximizes the score is obviously the requested MLE ranking. Thus, we design a sorting algorithm, based on dynamic programming that exploits Result 3.

Sorting an almost sorted list

Lemma 9.2.6 *Let $[n]$ be a set of n elements together with a scoring function q. Suppose that we are given that there is an optimal ordering $\sigma(1), ..., \sigma(n)$, that maximizes the score $s(\sigma) = \sum_{i<_\sigma j} q(i < j)$, such that $|\sigma(i) - i| \leq k$ for all i. Then we can find such an optimal σ in time $O(n \cdot k^2 \cdot 2^{6k})$.*

Remark 9.2.7 A brute force approach over all possible solutions would require time $k^{\Theta(n)}$, whereas a dynamic programming approach reduces the time complexity. Notice that when k is small ($o(logn)$), the algorithm tends to be linear.

Proof. Let $i < j$ be any pair of indices. Then, the optimal ranking σ maps the interval $[i, j]$ into the elements $I = \sigma([i, j]) = \{\sigma(i), ..., \sigma(j)\}$. This set of elements, by the assumption that $|\sigma(i) - i| \leq k$, satisfies the following subset coverings :

$$I^- = [i + k, j - k] \subset I \subset I^+ = [i - k, j + k]$$

All the elements inside I^- are obligated to be contained in I and each element of $[i, j]$ is mapped at most k positions apart by the optimal ordering.

By these two conditions, the set $S_I = \{\sigma(i), ..., \sigma(j)\}$ contains $j - i + 1$ elements. Thus, a possible selection of such a set requires choosing $j - i + 1$ containing the elements of I^- and be contained in I^+. Since the set I^- contains $j - i + 1 - 2k$ elements, it remains to pick $2k$ elements from the collection $I^+ \setminus I^- = [i - k, ..., i + k - 1] \cup [j - k + 1, ..., j + k]$, which contains $4k$ elements. Thus, the number of possible S_I's is at most 2^{4k}.

Let I be an interval and denote by $LH(I)$ and $RH(I)$ the left and right half of the interval. Without loss of generality, we choose the number of elements be $n = 2^m$ for some $m \in \mathbb{N}$. Let I_0 denote the interval containing all the elements, $I_1 = LH(I_0)$, $I_2 = RH(I_0)$, $I_3 = LH(I_1)$, ..., $I_{n-2} = I_{2^m-2} = RH(I_{2^{m-1}-2})$. In total, we have $n - 1 = 2^m - 1$ intervals, where there are 2^{m-1} intervals of length 2, ..., 2^{m-j} of length 2^j and 1 of length $2^m = n = |I_0|$.

For each such interval $I_t = [i..j]$, let S_t be the possible sets of the elements $J_t = [\sigma(i), ..., \sigma(j)]$. We will use dynamic programming to store an optimal ranking σ' of each such $J_t \in S_I$. In total, the number of J_t's is at most $(\#I_t) \cdot (\#S_t) < n \cdot 2^{4k}$. The optimal ordering satisfies the assumption : $|\sigma'(i) - i| \leq k$ for all i. Hence, the score of an optimal ranking σ' and a processed interval J_t :

$$s(J_t, \sigma') = \sum_{\sigma'(i')<\sigma'(j')} q(i' < j') = \sum_{\sigma'(i')<\sigma'(j'), i'<j'<i'+2k} q(i' < j') + \sum_{j' \geq i'+2k} q(i' < j')$$

Now, notice, from the optimality of σ', that the second term in the RHS is independent of σ'. Thus, we can define a score s' that is the sum over pairs $i', j' \in J_t$ that are less than $2k$ apart. These are the only pairs that may get swapped. Hence,

$$s(J_t, \sigma') = s'(J_t, \sigma') + \sum_{j' \geq i'+2k} q(i' < j') \Rightarrow \max_{\sigma'} s(J_t, \sigma') = \max_{\sigma'} s'(J_t, \sigma')$$

We apply the dynamic programming technique from $t = n-1 \rightarrow t = 0$, producing and storing an optimal ordering for each possible J_t.

1. If $n - 1 \leq t \leq \frac{n}{2}$, the length of J_t is 2 and, thus, the optimal ordering can be found in $O(1)$ steps.
2. If $t < \frac{n}{2}$, we have to find an optimal ordering of a given $J_t = [i, i + 2s - 1]$ for some appropriate $s > 0$. In order to achieve this, we study the two halves $LH(J_t)$ and $RH(J_t)$ and sort them independently.
   - For the $LH(J_t)$: It must contain all the elements in J_t that come from $[1, ..., i + s - 1 - k]$ and must be contained in $[1, ..., i + s - 1 + k]$. Thus, there are at most 2^{2k} choice for the elements of $LH(J_t)$.
   - The choice of the elements of $LH(J_t)$ determined uniquely the elements of $RH(J_t)$.
   - For each of the 2^{2k} choices, we search for an optimal ordering for the two halves, that we have already stored in the dynamic programming table. From the possible choices for the left half, we pick the best one. This is done by recomputing the score s' for the joined interval and takes at most $O(k^2)$ time. The only new pairs $(i, j) : |i - j| < 2k$ are along the boundary between $LH(J_t)$ and $RH(J_t)$.

Hence, the total cost is :

$$\sum_{d=1}^{logn} \sum_{j:|I_j|=d} cost_{DP}(I_j) = \sum_{d=1}^{logn} O(\frac{n \cdot 2^{4k}}{2^d} \cdot 2^{2k}k^2) = O(n \cdot k^2 \cdot 2^{6k})$$

■

Sorting the almost sorted average ranking $\overline{\pi}$

We have shown that the known average ranking is pointwise close with $k = 33L$ to the MLE ranking, with high probability. Thus, we can apply the pre-sorting algorithm presented above for appropriate k and get the following theorem.

Theorem 9.2.8 *There exists a randomized algorithm such that if $\{\pi_i\}_{i=1}^r$ be rankings on n elements independently generated by Mallow's model with parameter $\beta > 0$, and let $\alpha > 0$. Then a maximum probability order π^m can be computed in time :*

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

and error probability $< n^{-\alpha}$.

9.2.2 Proximity between the MLE ordering and the original ranking

Right now, we have fulfilled half of our promises. We have shown that we can find the MLE computationally fast via a pre-sorting technique on the average permutation. But, we have talked nothing about how close the MLE is compared to the central ranking π_0. This is our final goal.

→ We have to prove that the l_1 norm between the MLE ranking and the original permutation is of order n. The l_1 norm corresponds to the so-called *Average proximity*.

→ We have to prove that the l_∞ norm between the MLE ranking and the original permutation is of order $logn$. The l_∞ norm corresponds to the so-called *Pointwise proximity*.

Equivalent setting

In order to prove the the two results, we need to modify our setting by viewing our samples as noisy comparisons. Suppose there is a hidden ordering π_0 on n alternatives. Specifically, the input is no more an ordering of n alternatives but a collection of $\binom{n}{2}$ queries $q(i,j)$ for $i<j$. These queries are expressed as binary signals such that, for a constant $\lambda > 0$,

$$q(i,j) = \{+,-\} \text{ with probability } (\frac{1}{2}+\lambda) \text{ if } \{\pi_0(i) > \pi_0(j), \pi_0(i) < \pi_0(j)\}, \quad (9.7)$$

that is, the probability the signal has the correct sign is higher than 50%. It is assumed that the signals are independent. The parameter λ controls the *bias* of our noisy model. The higher the value of λ, the more robust to the true order our signals are. For each pair, the correct order is observed with probability greater than $\frac{1}{2}$. This idea is completely similar to the directed graph idea presented in the Condorcet-Mallows model 7.2.

For each unordered pair $\{x,y\}$, we receive a signal $s_{x,y} = s_{y,x}$. The signal distribution $\mathcal{D}$ for the pair $\{x,y\}$ depends on how these alternatives are ordered in the true hidden ranking π_0. Thus,

$$D = \mathbb{1}(\pi_0(x) < \pi_0(y))D_{x<y} + \mathbb{1}(\pi_0(y) < \pi_0(x))D_{y<x}$$

Signals are independent conditioned on the true order. The mass that the distribution $\mathcal{D}$ is assigned to the signal $s_{x,y}$ depends only on the position of x,y in the true ranking. Choose a set of pair of indices $I = \{(i_1,j_1),...,(i_{|I|},j_{|I|})\}$ such that $(x,y) \notin I$ and define the $|I|$-dimensional signal vector $\vec{s} = (s_{i_1,j_1},...,s_{i_{|I|},j_{|I|}})$. Then,

$$\mathcal{D}[s_{x,y} = *|\pi_0, \vec{s}] = \mathcal{D}[s_{x,y} = *|\mathbb{1}(\pi_0(x) < \pi_0(y))]$$

Definition 9.2.2 — Noisy Signal Aggregation (NSA). *Given the signals $s_{i,j}$ for all pairs $\{i,j\} \in [n]$, the NSA is the maximum likelihood permutation $\hat{\pi}^*$, assuming uniform prior. Thus,*

$$\hat{\pi}^* = arg\max_{\pi^*} \mathbb{P}[\{s_{i,j}\}|\pi^*] = \prod_{i,j:\pi^*(i)<\pi^*(j)} D_{i<j}(s_{i,j})$$

We have already defined the mass assigned by the signal distribution. We define the signal ratio, assuming uniform prior, for the signal $s_{x,y}$ and a ranking σ

$$\frac{D_{x<y}(s_{x,y})}{D_{y<x}(s_{x,y})} = \frac{\mathbb{P}[x <_\sigma y|s_{x,y}]}{\mathbb{P}[y <_\sigma x|s_{x,y}]}$$

Using this ratio, we can define the score $q(x < y)$ with the decision to rank x to below y as the log-ratio :

$$q(x < y) = log\frac{D_{x<y}(s_{x,y})}{D_{y<x}(s_{x,y})}$$

Note that $q(x < y) = -q(y < x)$. Observe that this log-ratio reminds the KL divergence and thus by Gibbs' inequality 4.34, we have that

$$\mathbb{E}[q(x < y)|\sigma(x) < \sigma(y)] \geq 0$$

Definition 9.2.3 — From NSA to Score. *The NSA is equivalent to the problem of finding a ranking σ such that*

$$\sigma = arg\max_{\sigma} s_q(\sigma) = arg\max_{\sigma} \sum_{x,y:\sigma(x)<\sigma(y)} q(x < y) \tag{9.8}$$

Main result

The main task it remains to point out is that the MLE optimal ranking and the true ranking are close in two norms, the l_1 and the l_∞. The following result holds.

Let π_0 be the true hidden ranking. Consider the NSA problem on biased signals parametrized by a bias $\lambda > 0$ and let $\hat{\pi}^*$ be any MLE optimal order. Let $\alpha > 0$ be a confidence parameter. Then, there exist two constants $c_i(\alpha, \lambda)$ for $i = 1, 2$ such that except with probability $O(n^{-\alpha})$ the following inequalities hold :

$$||\hat{\pi}^* - \pi_0||_1 = \sum_{i=1}^{n} |\hat{\pi}^*(i) - \pi_0(i)| \leq c_1 n \tag{9.9}$$

$$||\hat{\pi}^* - \pi_0||_\infty = \max_i |\hat{\pi}^*(i) - \pi_0(i)| \leq c_2 logn \tag{9.10}$$

Hence, we can see that the MLE ranking with high probability will be close to the central ranking. The proofs of the two inequalities can be found in [BM09].

10. $k-$ Set Sampling

10.1 Setting & Idea

In this thesis, we have analyzed in depth the field of rankings learning using noisy samples. This well-studied setting implies that one is given (independent) samples that are permutations of n alternatives, generated by a distribution, which corresponds to a noisy probabilistic model such as Mallows Model (MM) and Plackett-Luce Model (PL). Afterwards, one could ask questions concerning the sample complexity to learn the parameters of the model, the ability to learn the generating distributions in various f-divergence metrics (TV distance, KL divergence) and the concept of maximum likelihood estimation.

In our work, we chose to reduce the information provided by our samples and try to answer similar questions. This information reduction idea will be clear shortly. Firstly, we will introduce some helpful notation and, afterwards, we will present our results.

k-Set Sampling

Let $A = \{a_1, ..., a_n\}$ be the set of our alternatives. We are now ready to explain how to choose to reduce the information provided by our samples. We will use the single parameter Mallows Model as an example. Our main question remains to learn the central ranking π_0 in $\mathcal{L}(A)$. Our samples are still generated by a Mallows distribution $\mathcal{M}_1(\pi_0, \phi)$, but we do not have full access to the permutation sampled.

Our sampling will be parameterized by a natural number $0 < k < n$. In the previous chapter, we were observing a ranking $\pi_j \sim \mathcal{M}_1(\pi_0, \phi)$ of the n alternatives. Now, we again sample $\pi_j = a_{i_1} \succ a_{i_2} \succ ...a_{i_k} \succ a_{i_{k+1}} \succ ...a_{i_n}$ but we cannot access the sampled ranking. We can only access the k top ranked alternatives in an unordered way, that is, our sample is a set S_j of size k with the top k alternatives :

$$S_j = \{a_{i_1}, a_{i_2}, ..., a_{i_k}\}$$

The study of top k lists was already researched in various works such as [FS03]. The innovative part appears in the set theoretic version of our sampling.

Thus, our samples will be the sets $S_1, ..., S_r$ and, for instance, we question whether we can learn what the central ranking π_0 is.

A real-life application of this sampling method is the classical voting (with a cross † next to the names) of our preferred k out of n alternatives in a voting procedure. Each vote is just a set of our k top preferred alternatives, without specifying the order of our preferences.

An important remark

The way we have converted the nature of our samples is crucial. In the classical setting, one can easily observe that both input and output live in the symmetric group $\mathbb{S}_n$. They are both permutations. However, in our setting, we have not respected this property. The input consists of a collection of sets and the desired output is a permutation. This problem can be generalized to the quite interesting problem where the input and the output live in different metric spaces and one should create an interconnection between these spaces.

10.2 Notation

Let $A = \{a_1, ..., a_n\}$ be the set of our alternatives. A ranking $\pi \in \mathcal{L}(A)$ will be a bijection from A to itself.

We will denote with r the number of samples drawn from a distribution. Each sample will be a set S of size k and will be generated by a distribution $\mathcal{S}_k\mathcal{P}_{\phi,\pi_0}$, where firstly we draw a sample ranking from the distribution $\mathcal{P}_{\phi,\pi_0}$ and, afterwards, applying a k-set filtering $\mathcal{S}_k$.

In the next section, we will work with two probability measures. We will denote with $\mathbb{P}_{MM}$ the distribution of the single parameter Mallows model $\mathcal{M}_1(\pi_0, \phi)$ and with $\mathbb{P}_{PL}$ the distribution of the Plackett-Luce model. We have to expand the definition of these two measures from permutations to sets. Before that, we introduce the following notation. Each sample is a set of k alternatives. Thus, given r samples, one could aggregate them and get a vote counter random variable for each alternative. This vote-counter will be denoted by

$$v_a = \sum_{i=1}^{r} \mathbb{1}\{a \in S_i\}, \forall a \in A$$

Obviously, $0 \leq v_a \leq r$ and $\sum_{a\in A} v_a = r \cdot k$. Note that, if we define $p_a = \mathbb{P}[a \in S]$, $v_a \sim Bin(r, p_a)$, where S is drawn from $\mathcal{S}_k\mathcal{P}_{\phi,\pi_0}$.

Suppose that we have the sequence of vote-counters $\{v_a\}_{a\in A}$. We will be interested with the ranking of the n alternatives sorted in decreasing order of their vote-counters. This ranking will be denoted by $argsort_{i\in[n]}\{v_1, ..., v_n\}$

Also, given a set S of size k, we will denote by $g(S)$ the set of $k!$ permutations generated by the elements of the set. The set $g(S)$ will be called the generator of S.

Finally, given $A_1 \subset A$ and two permutations $\pi \in \mathcal{L}(A_1), \sigma \in \mathcal{L}(A \setminus A_1)$ of sizes $|A_1|$ and $n - |A_1|$, we will denote by $\pi \uplus \sigma$ the concatenated permutation of size n.

Now, we are ready to expand the definition of the two probability measures defined above for the k-set setting. Let S be a set of size k and let $R = A \setminus S$. We denote the probability measures as follows :

- $\mathbb{P}_{MM}$: Simple Mallows Model (on Rankings)
- $\mathbb{P}_{SM}$: Set Mallows Model (on Sets)
- $\mathbb{P}_{PL}$: Set Plackett-Luce Model

The probability to draw a sample S in the SM setting is :

$$\mathbb{P}_{SM}[S|\pi_0] = \sum_{\pi_S \in g(S)} \sum_{\pi_R \in g(R)} \mathbb{P}_{MM}[\pi_S \uplus \pi_R|\pi_0] = \sum_{\pi_S \in g(S)} \sum_{\pi_R \in g(R)} \frac{\phi^{d_{KT}(\pi_S \uplus \pi_R, \pi_0)}}{Z(\phi)}$$

Notice that the normalization constant is the same since the space of $n!$ possible permutations is decomposed to $\binom{n}{k}$ possible sets, each of which generates a collection of $k!(n-k)!$ permutations and each pair of such collections will be disjoint. Let $\mathcal{S}_k$ be the collection of all possible k-sets among n elements.

$$Z_{SM} = \sum_{S \in \mathcal{S}_k} \sum_{\pi \in g(S)} \sum_{\sigma \in g(R)} \phi^{d_{KT}(\pi \uplus \sigma, \pi_0)} = \sum_{\pi \in \mathcal{L}_A} \phi^{d_{KT}(\pi, \pi_0)} = Z_{MM} = Z(\phi)$$

We remind that we represent a ranking as a bijection $\sigma : [n] \to [n]$, where $\sigma(a)$ is the rank or position of the alternative a in the ranking. For $i \in [n]$, $\sigma^{-1}(i)$ is the alternative that is ranked at position i.

For the PL setting, we have that, given a value vector $\vec{w} \in W$:

$$\mathbb{P}_{PL}[S|\vec{w}] = \sum_{\sigma \in g(S)} \left(\prod_{i \in [k]} w_{\sigma^{-1}(i)}\right)\left(\prod_{i=1}^{k} \frac{1}{\sum_{j=i}^{n} w_{\sigma^{-1}(j)}}\right)$$

Notice that the values product $\prod_{i \in [k]} w_{\sigma(i)}$ are the same in each term of the sum $\sum_{\sigma \in g(S)}$, since it only permutes the elements of the set S and, hence, the above formula can be written :

$$\mathbb{P}_{PL}[S|\vec{w}] = \left(\prod_{i \in S} w_i\right) \sum_{\sigma \in g(S)} \left(\prod_{i=1}^{k} \frac{1}{\sum_{j=i}^{n} w_{\sigma^{-1}(j)}}\right)$$

10.3 MLE Analysis

MLE-MM-k-SET

Input : r independent sets $S_1, ..., S_r$ of size k.
Output : $\pi^* = arg\max_{\pi} \mathbb{P}_{SM}[S_1, ..., S_r|\pi]$

Theorem 10.3.1 The solution to the MLE-MM-K-SET is the $argsort_{i\in[n]}\{v_1, ..., v_n\}$

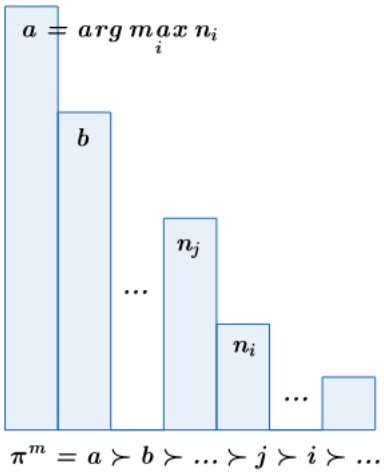

Figure 10.1: Our proposed MLE for the MLE-MM-k-SET problem

Proof. The k-Set Mallows probability measure is defined as:

$$\mathbb{P}_{SM}[S|\pi_0] = \sum_{\pi_S\in g(S)} \sum_{\pi_R\in g(R)} \mathbb{P}_{MM}[\pi_S \uplus \pi_R|\pi_0] = \frac{1}{Z} \sum_{\pi_S\in g(S)} \sum_{\pi_R\in g(R)} \phi^{d_{KT}(\pi_S\uplus\pi_R,\pi_0)}$$

Suppose that we are given r set samples. We define the product of these r terms (where each term contains $k!(n-k)!$ summands) as score of the ranking.

We claim that the permutation $\pi^m = i_1 \succ i_2 \succ ... \succ i_n$, s.t. $i_j = arg \max_{w\in[n]\setminus\{i_1,...,i_{j-1}\}} v_w$ (decreasing sequence of votes) is the MLE optimal ranking.

We proceed via contradiction using the swap-increasingness of the KT distance. Suppose that π^m is not the optimal ranking. Then, there exists another ranking, say OPT, that scores higher than π^m. Then, there must be some indexes i, j such that $i \succ_{OPT} j$ and $v_i < v_j$. We analyze two cases (inductively) :

CASE 1. Say that i, j are adjacent. The MLE score of the OPT solution is then given by :

$$score(OPT) = \prod_{i=1}^{r} \mathbb{P}_{SM}[S_i|OPT]$$

There are 4 distinct cases for the r sets drawn from the k-set Mallows model. There is a collection C_1 of r_1 sets that contain both i and j, a collection C_2 of r_2 sets that contain neither i nor j and collections C_3 and C_4 of r_3 and r_4 sets that contain only i and only j respectively. Obviously, $r_1 + ... + r_4 = r$ and $r_4 - r_3 = v_j - v_i > 0$. Hence,

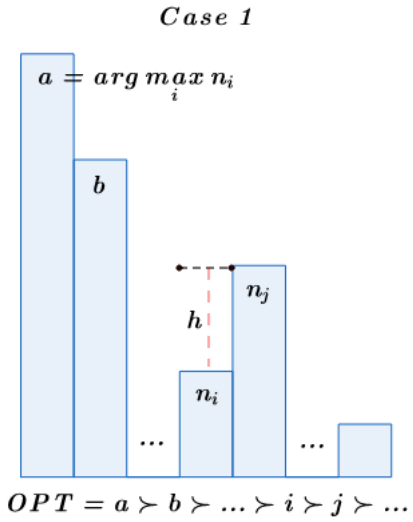

Figure 10.2: Case 1 : A possible OPT MLE for the Mle-MM-k-Set problem.

the score that be expressed as :

$$score(OPT) = \prod_{i=1}^{4} \prod_{S \in C_i} \mathbb{P}_{SM}[S|OPT]$$

Notice that if a set S contains both i and j,

$$\mathbb{P}_{SM}[S|OPT] = \mathbb{P}_{SM}[S|\pi^m]$$

The same holds for sets of the collection C_2.

Intuitively, this means that we cannot deduce preference over the alternatives i and j if we vote in the k-election system both i and j or neither of them.

Now, we study how sets of the collection C_3 behave on the score of the OPT. Notice that

$$OPT = \pi^m_{i \leftrightarrow j}$$

Thus, for $S \in C_3$,

$$\mathbb{P}_{SM}[S|OPT] = \frac{1}{Z} \sum_{\pi_S \in g(S)} \sum_{\pi_R \in g(R)} \phi^{d_{KT}(\pi_S \uplus \pi_R, OPT)}$$

Let $\sum_{\pi \in g(S) \cup g(R)} f(\pi) = \sum_{\pi_S \in g(S)} \sum_{\pi_R \in g(R)} f(\pi_S \uplus \pi_R)$ be the sum of $k!(n-k)!$ summands. Remind that $i \succ_\sigma j \ \forall \sigma \in g(S) \cup g(R)$. Then, by the swap-increasingness property of the KT distance and the fact that i, j are adjacent :

$$d_{KT}(OPT_{i \leftrightarrow j}, \sigma) =_{i \succ_\sigma j} d_{KT}(OPT, \sigma) + 1$$

Hence, since $OPT_{i \leftrightarrow j} = \pi^m$,

$$\mathbb{P}_{SM}[S|\pi^m] = \frac{1}{Z} \sum_{\sigma \in g(S) \cup g(R)} \phi^{d_{KT}(\sigma, \pi^m)} = \frac{1}{Z} \sum_{\sigma \in g(S) \cup g(R)} \phi^{d_{KT}(\sigma, OPT_{i \leftrightarrow j})} =_{i,j \text{ adjacent, } i \succ_\sigma j,\ i \succ_{OPT} j}$$

$$=_{i,j \text{ adjacent, } i\succ_\sigma j,\, i\succ_{OPT} j} \frac{1}{Z} \sum_{\sigma\in g(S)\cup g(R)} \phi^{d_{KT}(\sigma,OPT)+1} = \phi\mathbb{P}_{SM}[S|OPT]$$

Similarly, for $S \in C_4$:

$$\mathbb{P}_{SM}[S|OPT] =_{i,j \text{ adjacent}} \phi\mathbb{P}_{SM}[S|\pi^m]$$

Hence, since $r_4 - r_3 = v_j - v_i > 0$,

$$score(OPT) = score(\pi^m_{i\leftrightarrow j}) = \phi^{v_j - v_i} score(\pi^m) < score(\pi^m)$$

We have reached a contradiction.

Case 2. Suppose that OPT is any ranking with $d_{KT}(OPT, \pi^m) = d > 1$. Then, there is a finite sequence (of length d) of adjacent Case 1 swaps of elements that finally gives OPT. In each swap, from Case 1, the MLE decreases. Let $OPT^{(i)}$ be the candidate optimal ranking after i swaps from π^m, $i = 1, 2, ..., d$. Obviously, $OPT^{(d)} = OPT$ and set $\pi^m = OPT^{(0)}$. The pairs $(\pi^m, OPT^{(1)}), (OPT^{(1)}, OPT^{(2)}), ..., (OPT^{(d-1)}, OPT^{(d)})$ all belong to Case 1.

Hence, if we consider the sequence $(a_i, b_i)_{i=1}^d$ with $v_{b_i} > v_{a_i}$ and a_i adjacent to b_i in the pair of rankings $(OPT^{(i-1)}, OPT^{(i)})$ for all $i \in [d]$, we get that :

$$score(OPT^{(i)}) = score(OPT^{(i-1)}_{a_i\leftrightarrow b_i}) = \phi^{v_{b_i} - v_{a_i}} score(OPT^{(i-1)}), i \in [d]$$

Hence :

$$score(OPT) = \phi^{\sum_{i=1}^d (v_{b_i} - v_{a_i})} score(\pi^m) < score(\pi^m)$$

Cases 1 and 2 provide the optimality of the proposed MLE $argsort_{i\in[n]}\{v_1, ..., v_n\}$. ■

Remark 10.3.2 Notice that we have a closed form for the ratio of how the MLE score is changed for each proposed ranking σ with respect to the optimal solution.

Afterwards, we provide a TV distance result between the measure $\mathcal{S}_k\mathcal{P}_{\phi,\pi}$ and the measure $\mathcal{P}_{\phi,\pi}$. We denote $\sum_{\pi\in g(S)\cup g(R)} f(\pi) = \sum_{\pi_S\in g(S)} \sum_{\pi_R\in g(R)} f(\pi_S \uplus \pi_R)$.

Lemma 10.3.3 For any $\pi_i, \pi_j \in \mathcal{L}(A)$, $d_{TV}(\mathcal{S}_k\mathcal{P}_{\phi,\pi_i}, \mathcal{S}_k\mathcal{P}_{\phi,\pi_j}) \leq d_{TV}(\mathcal{P}_{\phi,\pi_i}, \mathcal{P}_{\phi,\pi_j})$

Proof. Let A_k be the set collection that contains all the possible $\binom{n}{k}$ sets of size k.

$$d_{TV}(\mathcal{S}_k\mathcal{P}_{\phi,\pi_i}, \mathcal{S}_k\mathcal{P}_{\phi,\pi_j}) = \frac{1}{2}\sum_{S\in A_k} |\mathcal{S}_k\mathcal{P}_{\phi,\pi_i}(S) - \mathcal{S}_k\mathcal{P}_{\phi,\pi_j}(S)|$$

$$= \frac{1}{2}\sum_{S\in A_k} |\sum_{\pi\in g(S)\cup g(R)} \frac{\phi^{d_{KT}(\pi,\pi_i)}}{Z_{MM}} - \sum_{\pi\in g(S)\cup g(R)} \frac{\phi^{d_{KT}(\pi,\pi_j)}}{Z_{MM}}| \leq$$

$$\leq \frac{1}{2}\sum_{S\in A_k}\sum_{\pi\in g(S)\cup g(R)} |\frac{\phi^{d_{KT}(\pi,\pi_i)}}{Z_{MM}} - \frac{\phi^{d_{KT}(\pi,\pi_j)}}{Z_{MM}}| = \frac{1}{2}\sum_{\pi\in\mathcal{L}(A)} |\mathbb{P}_{MM}[\pi|\pi_i] - \mathbb{P}_{MM}[\pi|\pi_j]| = d_{TV}(\mathcal{P}_{\phi,\pi_i}, \mathcal{P}_{\phi,\pi_j})$$

■

Remark 10.3.4 When we want to learn the whole ranking π_0, we can just consider $k < \frac{n}{2}$. When solving the problem k-Set with samples $S_1, ..., S_m$ and $k > \frac{n}{2}$, then it is equivalent to solve the problem $(n-k)$-Set with samples $[n] \setminus S_1, ..., [n] \setminus S_m$.

MLE-PL-k-SET

Setting : There are n objects $\{o_i\}_{i=1}^n$ with unknown values $\{w_i\}_{i=1}^n$. We generate samples from the PL-k-SET Model and we want to determine the elements ranking with respect to their value. Hence, our goal is to be able to answer the $\binom{n}{2}$ pairwise comparisons $\{w_i > w_j\}$.
Input : r independent sets $S_1, ..., S_r$ of size k, each one containing k objects.
Output : $\vec{w}_\pi^* = arg\max_{\vec{w}_\pi} \mathbb{P}_{PL}[S_1, ..., S_r | \vec{w}_\pi]$

Theorem 10.3.5 The solution to the MLE-PL-k-SET is the $\vec{w}^* = w_{\pi^{-1}(1)} \geq w_{\pi^{-1}(2)} \geq ... \geq w_{\pi^{-1}(n)}$ where $\pi = argsort_{i\in[n]}\{v_1, ..., v_n\}$

Proof. The likelihood function we want to maximize can be written as :

$$\mathcal{L}(\{S_1, ..., S_r\}|\vec{w}) = \prod_{m=1}^{r} (\prod_{i\in S_m} w_i) \sum_{\sigma\in g(S_m)} (\prod_{i=1}^{k} \frac{1}{\sum_{j=i}^{n} w_{\sigma^{-1}(j)}})$$

The space that our maximization problem is defined is the following :

$$\mathbb{S}_W = \{\vec{w}_\pi = w_{\pi^{-1}(1)} \geq w_{\pi^{-1}(2)} \geq ... \geq w_{\pi^{-1}(n)} | \pi \in \mathbb{S}_n, \sum_i w_i = 1\}$$

We do not care about the value of each object but only to determine between any pair of objects which is the most valuable. Of course, according the PL model, the values have to satisfy the normalization condition $\sum_{i=1}^{n} w_{\pi^{-1}(i)} = 1$.

Hence, we want to solve the optimization problem :

$$\vec{w}_\pi^* = arg\max_{\vec{w}\in\mathbb{S}_W} \mathcal{L}(\{S_1, ..., S_r\}|\vec{w})$$

Firstly, we have to gain some intuition, we can simplify the above expression by using the log-likelihood function and get :

$$log\mathcal{L}(\{S_1, ..., S_r\}|\vec{w}) = \sum_{m=1}^{r} \left\{ log(\prod_{i\in S_m} w_i) + log(\sum_{\sigma\in g(S_m)} (\prod_{i=1}^{k} \frac{1}{\sum_{j=i}^{n} w_{\sigma^{-1}(j)}})) \right\} =$$

$$= \sum_{m=1}^{r} \sum_{i\in S_m} log(w_i) + \sum_{m=1}^{r} log(\sum_{\sigma\in g(S_m)} (\prod_{i=1}^{k} \frac{1}{\sum_{j=i}^{n} w_{\sigma^{-1}(j)}}))$$

We have to maximize this function.

The first term can be rewritten as :

$$RHS_1 = \sum_{m=1}^{r} \sum_{i \in S_m} log(w_i) = \sum_{i=1}^{m} v_i log(w_i)$$

Hence, it seems logical to think that $argsort_{i \in [n]}\{v_1, ..., v_n\}$ is the MLE we are looking for. Return to the first likelihood function. We have to maximize the following function :

$$\mathcal{L}(\{S_1, ..., S_r\}|\vec{w}) = \prod_{m=1}^{r} (\prod_{i \in S_m} w_i) \sum_{\sigma \in g(S_m)} (\prod_{i=1}^{k} \frac{1}{\sum_{j=i}^{n} w_{\sigma^{-1}(j)}})$$

Notice that :

$$\prod_{m=1}^{r} (\prod_{i \in S_m} w_i) = \prod_{i=1}^{m} w_i^{v_i}$$

For the part after that product in the likelihood function, consider the function :

$$f(x_1, ..., x_n) = \sum_{m=1}^{r} log(\sum_{\sigma \in g(S_m)} (\prod_{i=1}^{k} \frac{1}{\sum_{j=i}^{n} x_{\sigma^{-1}(j)}}))$$

$$exp(f(w_1, ..., w_n)) = f_1 \dots f_r = \prod_{m=1}^{r} \sum_{\sigma \in g(S_m)} (\prod_{i=1}^{k} \frac{1}{\sum_{j=i}^{n} w_{\sigma^{-1}(j)}})$$

Suppose that we score $A > 0$ if we order the weights according to the appearances v_i. Now, suppose that this is not optimal. Then, suppose that we change the elements with appearances $v_i > v_j$. We will show that this choice decreases the score. We choose a vector with $w_j > w_i$. We will show that is choice scores less than setting $w_i > w_j$.

Let I, J be the collections of the sets where the elements o_i and o_j appear respectively. Then, $|I| = v_i, |J| = v_j, |I \cap J| = t \leq v_j$. Then, the likelihood function can be partitioned into four disjoint products :

$$f_1 \dots f_r = (\prod_{S_v \in \{S_1, ..., S_r\} \setminus I, J} f_v)(\prod_{S_v \in I \cap J} f_v)(\prod_{S_v \in I \setminus J} f_v)(\prod_{S_v \in J \setminus I} f_v)$$

The first two terms remain the same after the swap. The first term contains neither i nor j and, thus, there is no impact in the score. The second remains the same since we have all the permutations over the elements contained in the sets $S \in I \cap J$ and, hence, each element goes around all possible positions. Thus, there is a symmetry between the appearances of i and j. The third product has $v_i - t$ terms and the last $v_j - t$, and there is the score difference we want to observe,

We can show that :

$$f_1 \dots f_r(v_i > v_j \wedge w_i < w_j) < A$$

Obviously, if $f, g > 0$ and increasing in an interval I, then $f \cdot g$ will also be increasing. We study the function that is the product of k terms :

$$f(x_1, ..., x_k) = \prod_{i=1}^{k} \frac{1}{\sum_{j=i}^{n} x_j} = \prod_{i=1}^{k} \frac{1}{1 - \sum_{j=1}^{i-1} x_j} = \frac{1}{1 - x_1} \frac{1}{1 - x_1 - x_2} \frac{1}{1 - x_1 - ... - x_{k-1}}$$

For $\vec{x} \in [0, 1]^k$, f is increasing. If we fix the $k - 1$ values and let one variable run free, then :

$$0 < x < y < 1 \Rightarrow f(x|w_1, ..., w_{k-1}) < f(y|w_1, ..., w_{k-1})$$

The same holds for $F(x|w_1, ..., w_{k-1}) = \sum_{\sigma \in g(S_m)} f(x|w_1, ..., w_{k-1}, \sigma)$.

For a set $S = \{i, i_1, ..., i_{k-1}\}$, let $\vec{w}_{S\setminus\{i\}} = \{w_{i_1}, ..., w_{i_{k-1}}\}$. Hence, by picking $w_j > w_i$:

$$\underbrace{\left\{ \prod_{S \in I \setminus J} F(w_i|\vec{w}_{S\setminus\{i\}}) \right\}}_{(v_i - t) \text{ terms}} \underbrace{\left\{ \prod_{S \in J \setminus I} F(w_j|\vec{w}_{S\setminus\{j\}}) \right\}}_{(v_j - t) \text{ terms}} <_{v_i > v_j, w_j > w_i}$$

$$< \left\{ \prod_{S \in I \setminus J} F(w_j|\vec{w}_{S\setminus\{i\}}) \right\} \left\{ \prod_{S \in J \setminus I} F(w_i|\vec{w}_{S\setminus\{j\}}) \right\} = A$$

Note that the weights can only be swapped because the sum should remain fixed to 1.

So, any swap that does not respect the relation between v_i and v_j for any pair i, j, will only decrease the score of the MLE. Hence, the MLE optimizer is the values ranking $argsort_{i \in [n]}\{v_1, ..., v_n\}$, that is to assign values in decreasing order of the appearance frequency of the alternatives. ■

10.4 The Mallows k-Gap Filling Model

Finally, we propose another noisy sampling model. Here, we draw a ranking $\pi \sim \mathcal{P}_{\phi,\pi_0}$ and afterwards apply a uniform filtering in order to hide k elements. Thus, in the given sample, we will only access $(n - k)$ elements and in the positions of elements missing we see a $\star$ symbol. Suppose that the given ranking that is missing k elements is drawn from a distribution $\mathcal{U}_k\mathcal{P}_{\phi,\pi_0}$. At first, we will define the appropriate probability measure $\mathbb{P}_{GF}$.

Denote by $\mathbb{S}_{n,k}$ the set of all rankings of size n, that are missing k elements. Given a $\tau \in \mathbb{S}_{n,k}$, we define the set $\mathcal{M}_\tau$ of the missing elements. Obviously, $g(\mathcal{M}_\tau)$ contains all the possible permutations of the k missing elements. Define a filling function $f : \mathbb{S}_{n,k} \times g(\mathcal{M}) \to \mathbb{S}_n$, which fills a partial ranking τ with $(n - k)$ fixed objects and k stars, with the k missing items from $\mathcal{M}_\tau$, according to a ranking from $g(\mathcal{M}_\tau)$. For instance, if $\tau = 1 \succ \star \succ 2 \succ \star$, $\mathcal{M}_\tau = \{3, 4\}$ and $3 \succ 4 \in g(\mathcal{M}_\tau)$. Hence, $f(\tau, 3 \succ 4) = 1 \succ 3 \succ 2 \succ 4$. Now, we have the necessary notation to proceed to the definition of the measure :

$$\mathbb{P}_{GF}(\pi_\star|\pi_0) = \sum_{\sigma \in g(\mathcal{M}_{\pi_\star})} \frac{\phi^{d_{KT}(f(\pi_\star,\sigma),\pi_0)}}{Z_{GF}}$$

For simplicity, let $GF(\pi_\star)$ be the set $\{f(\pi_\star, \sigma)|\sigma \in g(\mathcal{M}_{\pi_\star})\}$.

$$\mathbb{P}_{GF}(\pi_\star|\pi_0) = \sum_{\sigma \in GF(\pi_\star)} \frac{\phi^{d_{KT}(\sigma,\pi_0)}}{Z_{GF}}$$

The set $GF(\pi_\star)$ contains $k!$ permutations and, hence, generates all the possible samples drawn from the distribution $\mathcal{P}_{\phi,\pi_0}$ before applying the filter $\mathcal{U}_k$.

In order to understand this sum, we have to study the cardinality of the $\mathbb{S}_{n,k}$ (the set of all possible samples drawn from the distribution $\mathcal{U}_k\mathcal{P}_{\phi,\pi_0}$).

The size of $\mathbb{S}_{n,k}$ is

$$|\mathbb{S}_{n,k}| = supp(\mathcal{U}_k\mathcal{P}_{\phi,\pi_0}) = \binom{n}{k}\binom{n}{n-k}(n-k)! = \frac{(n!)^2}{(k!)^2(n-k)!}$$

since, at first, we can place k stars and, then, for the remaining $n-k$ positions, choose $n-k$ among the n elements and create all the possible rankings. For instance, for $n = 4$, $|\mathbb{S}_{4,2}| = 72$, whereas $|\mathbb{S}_4| = 24$.

As far as the normalization constant is concerned, in this case, there is no $1-1$ correspondence between the samples and the times each ranking of size n will be appeared. For instance, in the $(n, k) = (4, 2)$ case, the ranking $1 \succ 2 \succ 3 \succ 4$, can be generated from many samples such as $\star \succ 2 \succ 3 \succ \star, 1 \succ \star \succ 3 \succ \star$, etc. In the 2-set case, the only generator was the set $\{1, 2\}$. Now, the normalization constant can be expressed as :

$$Z_{GF} = \sum_{\pi_\star \in \mathbb{S}_{n,k}} \sum_{\sigma \in GF(\pi_\star)} \phi^{d_{KT}(\sigma,\pi_0)}$$

This sum contains $\frac{(n!)^2}{(k!)^2(n-k)!}k! = n!\binom{n}{k}$ summands and, thus, it offers us a hint of how many times each permutation of size n appears in the sum. Notice that each of the $n!$ possible permutations appears in the sum the same number of times (due to symmetry) and the number of appearances is $\binom{n}{k}$. Hence,

$$Z_{GF} = \sum_{\pi_\star \in \mathbb{S}_{n,k}} \sum_{\sigma \in GF(\pi_\star)} \phi^{d_{KT}(\sigma,\pi_0)} = \binom{n}{k} \sum_{\pi \in \mathbb{S}_n} \phi^{d_{KT}(\pi,\pi_0)} = \binom{n}{k} Z_{MM}$$

It is easy to verify the extreme cases $k = 1$ and $k = n$.

Geometric Intuition

In this setting, we would like to apply a generalized version of the techniques applied in Chapter 9. Thus, we consider crucial to obtain a geometric intuition of our samples. Firstly, notice the recursive structure of the permutations and permutohedra. If one fixes a coordinate of a permutation of size n, then one gets a permutation of the

remaining $n-1$ elements. Similarly, for fixing some k elements of the permutation. Thus, the same geometric intuition holds for the $\mathbb{S}_n$-permutohedra. For instance, if $n=4$, the corresponding permutohedron is constructed via the S_1-permutohedra (trivially), the S_2-permutohedra (fixing two elements) and the S_3-permutohedra (fixing one element). Notice, in the following figure, that the edge between (123) and (132) is a $\mathbb{S}_2$-permutohedron, that corresponds to the sample $4 \succ 1 \succ \star \succ \star$.

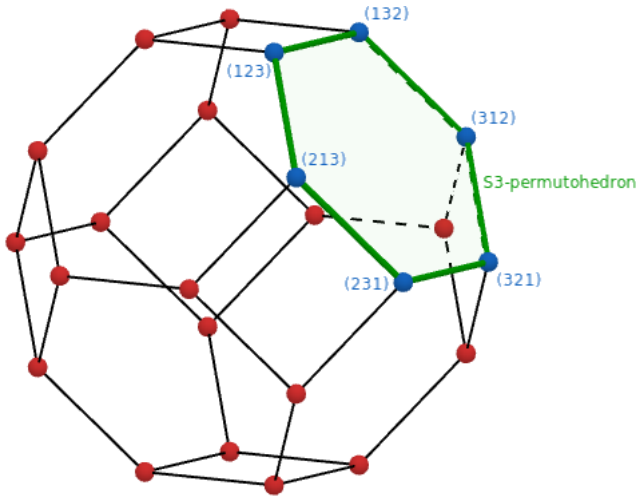

Figure 10.3: For $n=4$, the sample $4 \succ * \succ * \succ *$ corresponds to the green subspace, that is one of the $\mathbb{S}_3$-permutohedron sides of $\mathbb{S}_4$-permutohedron.

Now, suppose that we are given a sample π, drawn by the k-Gap Filling Model $\mathcal{U}_k\mathcal{P}_{\phi,\pi_0}$. This sample contains k stars, put uniformly at random among the n elements. Hence, our sample is just a 'side', a projection of our $\mathbb{S}_n$ polytope. This 'side' is a $\mathbb{S}_k$-permutohedron, lives in the space $\mathbb{R}_{k-1}$ and corresponds to the collection of the $k!$ possible permutations of the missing elements.

Finally, we provide a TV distance result between the measure $\mathcal{U}_k\mathcal{P}_{\phi,\pi}$ and the measure $\mathcal{P}_{\phi,\pi}$.

Lemma 10.4.1 For any $\pi_i, \pi_j \in \mathcal{L}(A)$, $d_{TV}(\mathcal{U}_k\mathcal{P}_{\phi,\pi_i}, \mathcal{U}_k\mathcal{P}_{\phi,\pi_j}) \leq d_{TV}(\mathcal{P}_{\phi,\pi_i}, \mathcal{P}_{\phi,\pi_j})$

Proof.

$$d_{TV}(\mathcal{U}_k\mathcal{P}_{\phi,\pi_i}, \mathcal{U}_k\mathcal{P}_{\phi,\pi_j}) = \frac{1}{2}\sum_{\sigma\in\mathbb{S}_{n,k}} |\mathcal{U}_k\mathcal{P}_{\phi,\pi_i}(\sigma) - \mathcal{U}_k\mathcal{P}_{\phi,\pi_j}(\sigma)|$$

$$= \frac{1}{2}\sum_{\sigma\in\mathbb{S}_{n,k}} |\sum_{\pi\in GF(\sigma)} \frac{\phi^{d_{KT}(\pi,\pi_i)}}{\binom{n}{k} Z_{MM}} - \sum_{\pi\in GF(\sigma)} \frac{\phi^{d_{KT}(\pi,\pi_j)}}{\binom{n}{k} Z_{MM}}| \leq$$

$$\leq \frac{1}{2}\sum_{\sigma\in\mathbb{S}_{n,k}}\sum_{\pi\in GF(\sigma)} |\frac{\phi^{d_{KT}(\pi,\pi_i)}}{\binom{n}{k} Z_{MM}} - \frac{\phi^{d_{KT}(\pi,\pi_j)}}{\binom{n}{k} Z_{MM}}| = \frac{1}{2}\sum_{\pi\in\mathcal{L}(A)} |\mathbb{P}_{MM}[\pi|\pi_i] - \mathbb{P}_{MM}[\pi|\pi_j]| = d_{TV}(\mathcal{P}_{\phi,\pi_i}, \mathcal{P}_{\phi,\pi_j})$$

■

10.5 Future Work

As a future step, we propose the following three different directions. Firstly, it is an interesting question to expand our learning framework concerning problems whose input and output belong to different metric spaces. Secondly, we have thought a connection between a classical NP-hard problem and our k-set sampling setting. The problem is called Min Sum Set Cover (MSSC). Our idea links the learning problems we are interested in with the optimal solution (or a good approximation) of the MSSC problem. The MSSC is a problem related both to the classical min set cover problem and to the linear arrangement problems and is defined as follows :

Min Sum Set Cover

Input : A hypergraph $H(V, E)$ [1], a linear ordering, that is a bijection $f : V \mapsto \{1, ..., |V|\}$. We, then, define for a hyperedge e, the cost $f(e) := \min_{v \in e} f(v)$.

Output : $f^* = arg\min_f \sum_{e \in E} f(e)$

It is well known that the greedy algorithm approximates MSSC within a ratio no worse than 4, and that this is the best possible approximation, that this for every $\epsilon > 0$, it is NP-hard to approximate MSSC within a ratio of $4 - \epsilon$. This result can be found in [UFT02]. Another good source is the [Im16]. How this problem is linked to our k-set sampling? The problem's structure is quite similar to our learning framework. Note that the linear ordering f is just a ranking of the elements of the set V. Thus, we are given sets (of different sizes) and we want to learn a ranking. This is quite similar to our k-set sampling setting if we do not fix k. It would be interesting to see MSSC as a learning problem. However, there are some difficulties one has to deal with. For instance, if one chooses to cover a vertex v, we should afterwards delete all the hyperedges covered by this vertex. Thus, each choice we make, causes a deletion of a subset of our sets.

Another interesting direction would be to be able to answer towards the following kind of questions : Suppose that there are two models, a single parameter Mallows model $\mathcal{M}_1$ and a generalized Mallows model $\mathcal{M}_n$. Let π be a voting profile of size r generated by one of the these models. Can we determine from which distribution we have drawn our samples, and, if so, how many samples are needed? These are some potential directions for study in the field of learning theory for ranking distributions. Finally, another fascinating problem is to bound the fluctuations of the length of the longest increasing subsequence of a sample π, drawn from a Mallows distribution. [CM11], [NB14].

[1] Let S be a set of points and $\mathcal{F} = \{S_1, ..., S_r$ be a collection of subsets of S. The hyperedges of H correspond to the points in the set system and the vertices of H correspond to the subsets. Note that E is a a set of non-empty subsets of V and constitutes a generalization of the classical edge (that is a two-set).

Bibliography

[AE17] Mohamed Elhoseiny Marian Mazzone Ahmed Elgammal, Bingchen Liu. Can: Creative adversarial networksgenerating "art" by learning about styles anddeviating from style norms. *arXiv preprint arXiv:1706.07068*, 2017.

[AEMP18] Robert Busa-Fekete Adil El Mesaoudi-Paul, Eyke Hüllermeier. Ranking distributions based on noisy sorting. *Proceedings of the 35th International Conference on Machine Learning, PMLR 80:3472-3480*, 2018.

[AK19] Ritesh Noothigattu Ariel Procaccia Christos-Alexandros Psomas Anson Kahng, Min Kyung Lee. Statistical foundations of virtual democracy. *Proceedings of the 36th International Conference on Machine Learning, PMLR 97:3173-3182*, 2019.

[BM09] Mark Braverman and Elchanan Mossel. Sorting from noisy information. *CoRR, abs/0910.1191*, 2009.

[BS50] B. Babington-Smith. Discussion of professor ross's paper. *Journal of the Royal Statistical Society B, 12:153–162*, 1950.

[BS16] Massart P. Boucheron S., Lugosi G. *Concentration inequalities : A nonasymptotic theory of independence*. Oxford University Press, 2016.

[BT89] Tovey C.A. Bartholdi, J. and Trick. Voting schemes for which it can be difficult to tell who won the election. *M.A. Soc Choice Welfare (1989) 6: 157*, 1989.

[Cha10] Pătraşcu Mihai Chan, Timothy M. Counting inversions, offline orthogonal range counting, and related problems. *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms. p. 161*, 2010.

[CM11] Shannon Starr Carl Mueller. The length of the longest increasing subsequence of a random mallows permutation. *Journal of Theoretical Probability, pages 1–27*, 2011.

[CT06] Thomas J. Cover T. *Elements of Information Theory*. Wiley-Interscience, 2006.

[DP77] Graham R.L. Diaconis P. Spearman's footrule as a measure of disarray. *J. Roy. Statistics Soc., 39(Ser. B):262–268*, 1977.

[DP16] Mansi Jain Neha Jain Krushi Gada Dipti Pawade, Avani Sakhapara. Deep learning for music. *arXiv preprint arXiv:1606.04930*, 2016.

[DP18] Mansi Jain Neha Jain Krushi Gada Dipti Pawade, Avani Sakhapara. Story scrambler -automatic text generation using word level rnn-lstm. *I.J. Information Technology and Computer Science, 2018, 6, 44-53*, 2018.

[Fli86] Verducci J. S. Fligner, M. A. Distance based ranking models. *Journal of the Royal Statistical Society. Series B (Methodological), pp. 359–369*, 1986.

[Fra03] John B. Fraleigh. *First Course in Abstract Algebra*. Pearson, 2003.

[FS03] Kumar R. Fagin, R. and D. Sivakumar. Comparing top k lists. *In Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms(Baltimore, MD, USA,2003), SIAM, pp. 28–36*, 2003.

[Im16] Sungjin Im. Min-sum set cover and its generalizations. *Encyclopedia of Algorithms. Springer, New York, NY*, 2016.

[JV15] Y. Jiao and J.-P. Vertz. The kendall and mallows kernels for permutations. *In Proceedings of The 32nd International Conference on Machine Learning, volume 37 of JMLR:WCP, pages 1935–1944*, 2015.

[Kal02] Olav Kallenberg. *Foundations of Modern Probability*. Springer-Verlag New York, 2002.

[Kor07] Sinai Yakov G. Koralov, Leonid. *Theory of Probability and Random Processes*. Springer-Verlag Berlin Heidelberg, 2007.

[Lan05] S. Lang. *Algebra (Graduate Texts in Mathematics)*. Springer, 2005.

[LC14] T. Lu and C.Boutilier. Effective sampling and learning for mallows models with pairwise-preference data. *Journal of Machine Learning Research, 15:3783–3829*, 2014.

[LDSR16] E. Upfal L. De Stefani, A. Epasto and F. Vandin. R. Reconstructing hidden permutations using the average-precision (ap) correlation statistic. *AAAI, pages 1526–1532*, 2016.

[LM18] Allen Liu and Ankur Moitra. Efficiently learning mixtures of mallows models. *In FOCS, pages 627–638. IEEE Computer Society*, 2018.

[Luc59] R. D. Luce. Individual choice behavior: A theoretical analysis. *Wiley*, 1959.

[Mal57] C Mallows. Non-null ranking models. *Biometrika, 44(1): 114–130*, 1957.

[NB14] Ron Peled Nayantara Bhatnagar. Lengths of monotone subsequences in a mallows permutation. *Probability Theory and Related Fields, 2015 - Springer*, 2014.

[NN07] Tardos E. Vazirani V. Nisan N., Roughgarden T. *Algorithmic game theory.* Cambridge University Press, 2007.

[Pla75] R. Plackett. The analysis of permutations. *Applied Statistics, 24:193–202*, 1975.

[RBF19] Balázs Szörényi Manolis Zampetakis Róbert Busa-Fekete, Dimitris Fotakis. Optimal learning for mallows block model. *Journal of Machine Learning Research*, 2019.

[Sam59] Arthur L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 1959.

[SGS17] P. Menard S. Gerchinovitz and G. Stoltz. Fano's inequality for random variables. *preprint,arXiv:1702.05985*, 2017.

[Sha48] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 1948.

[Tao13] T. Tao. *An introduction to measure theory.* American Mathematical Society, 2013.

[TD18] Amrita S. Tulshan and Sudhir Namdeorao Dhage. Survey on virtual assistant: Google assistant, siri, cortana, alexa. *International Symposium SIRS 2018, Bangalore, India, September 19–22*, 2018.

[Tur50] A. M. Turing. Computing machinery and intelligence. *Mind 49: 433-460*, 1950.

[UFT02] László Lovász Uriel Feige and Prasad Tetali. Approximating min-sum set cover. *In Proc. of the 5th Intl. Workshop on Approximation Algorithms for Combinatorial Optimization (APPROX), Sept.*, 2002.

[YR08] J. A.; Yilmaz, E.; Aslam and S. Robertson. A new rank correlation coefficient for information retrieval. *In Proceedings of the 31st annual ACM SIGIR conference, 587–594. ACM.*, 2008.