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

# Αλγοριθμικές Τεχνικές Μάθησης Πιθανοτικών Κατανομών και Εφαρμογές τους σε Προβλήματα Κοινωνικής Επιλογής

Διπλωματική Εργασία

του

Βλατάκη-Γκαραγκούνη Εμμανουήλ-Βασιλείου

**Επιβλέπων:** Δημήτρης Φωτάκης
Επίκουρος Καθηγητής Ε.Μ.Π.

II

Εργαστήριο Λογικής και Επιστήμης Υπολογισμών
Αθήνα, Ιούλιος 2016

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

# Αλγοριθμικές Τεχνικές Μάθησης Πιθανοτικών Κατανομών και Εφαρμογές τους σε Προβλήματα Κοινωνικής Επιλογής

Διπλωματική Εργασία

του

Βλατάκη-Γκαραγκούνη Εμμανουήλ-Βασιλείου

**Επιβλέπων:** Δημήτρης Φωτάκης
Επίκουρος Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 19η Ιουλίου 2016.

| ..... | ..... | ..... |
|---|---|---|
| Δημήτρης Φωτάκης | Μιχάλης Λουλάκης | Άρης Παγουρτζής |
| Επίκουρος Καθηγητής Ε.Μ.Π. | Επίκουρος Καθηγητής Ε.Μ.Π. | Επίκουρος Καθηγητής Ε.Μ.Π. |

Εργαστήριο Λογικής και Επιστήμης Υπολογισμών
Αθήνα, Ιούλιος 2016

-

.....
Εμμανουήλ-Βασίλειος Βλατάκης-Γκαραγκούνης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

# Περίληψη

Σε αυτή την διπλωματική εργασία, μελετούμε προβλήματα μάθησης πιθανοτικών κατανομών κάτω από το πρίσμα αλγοριθμικών τεχνικών. Μελετούμε μια φυσική γενίκευση του PAC μοντέλου μάθησης μιας άγνωστης διακριτής κατανομής πιθανότητας. Στο πλαίσιο αυτό, ο αλγόριθμος μάθησης λαμβάνει ως είσοδο $n$ ανεξάρτητα δείγματα από μια άγνωστη κατανομή $X$. Χρησιμοποιώντας αυτά τα δείγματα, ο αλγόριθμος προτείνει με πιθανότητα τουλάχιστον $1-\delta$ μια κατανομή-υπόθεση $X^*$ τέτοια ώστε η στατιστική απόσταση των δύο κατανομών (TV ): $d_{\mathrm{TV}}(X, X^*)$ να είναι το πολύ $\epsilon$, όπου $\epsilon, \delta > 0$ είναι οι παράμετροι ακρίβειας και εμπιστοσύνης, τα οποία παρέχονται επίσης ως είσοδο στον χρήστη. Παρουσιάζουμε εκτενώς την προηγούμενη επιστημονική εργασία πάνω σε αυτό το πλαίσιο για τις κλασσικές οικογένειες κατανομών π.χ μονότονες, λογαριθμικά κυρτές, μόνολοφες, παρέχοντας ακριβή άνω και κάτω φράγματα. Εστιάζουμε πάνω σε μια νέα αλγοριθμική τεχνική, στην κατασκευή αραιώς πυκνών καλυμμάτων , τα οποία είναι αραιά ως προς των πληθάριθμο αλλά πυκνά ως προς την μετρικότητα του χώρου της κλάσης των κατανομών που μελετούμε. Η μέθοδος εκμεταλλεύεται την δομή των κατανομών αυτών για να σχεδιάσει αποδοτικά χρονικώς και δειγματικώς αλγορίθμους. Επιπλέον, μελετούμε το πρωτόλειο έργο των [ΔΠ13] και [ΔΔΣ15] στις Poisson Binomial Distribution, το άθροισμα $n$ ανεξάρτητων Bernoulli. Στην συνέχεια αναπτύσσουμε μια παρόμοια προσέγγιση στην αυξανόμενα αναπτυσσόμενη περιοχή της υπολογιστικής θεωρίας κοινωνικής επιλογής και πιο συγκεκριμένα στο τομέα του χροωδοτινγ. Βασιζόμενοι στο διάσημο θορυβοποιό μοντελό του Mallow στο οποίο κάθε ψηφοφόρος είναι ένας εκτιμητής της κοινωνικής πραγματικότητας, παρουσιάζουμε την δουλειά μας πάνω σε εκλογικούς κανόνες όπως ο Kemeny/Plurality επεκτείνοντας την προηγούμενη ερευνητική εργασία.

**Λέξεις Κλειδιά**

Θεωρία Μάθησης, Στατιστική Μάθηση, Μηχανική Μάθηση, Εφαρμοσμένες Πιθανότητες, Θεωρία Πληροφορίας, Υπολογιστική Θεωρία Κοινωνικής Επιλογής,Αλγόριθμοι και Πολυπλοκότητα

# Abstract

In this thesis, we study probability distribution learning problems from a computational algorithmic perspective. We work in a natural PAC-style model of learning an unknown discrete probability distribution. In this framework, the learner is provided with the value of n and with independent samples drawn from the unknown distribution $X$. Using these samples, the learner must with probability at least $1-\delta$ output a hypothesis distribution $X^*$ such that the total variation distance $d_{\mathrm{TV}}(X, X^*)$ is at most $\epsilon$, where $\epsilon, \delta > 0$ are accuracy and confidence parameters that are provided to the learner. We present the previous work on that framework for classical classes of probability distributions i.e monotone, log-concave, unimodal distributions giving tight upper and lower sample bounds. We focus on a new algorithmic technique in distribution learning, the construction of covers, that are sparse in cardinality and dense in metric space of a class of distributions. The method exploits that structure in order to design efficient in time and sample complexity learning algorithms. We study the seminal work of [DP13] and [DDS15] on the Poisson Binomial Distribution, the sum of n independent Bernoulli. We develop then a similar approach on an gradually emerging field of computational social choice, crowdvoting. Based on the famous Mallow noise Model in which every voter is an estimator of the social ground truth, we present our work for the Kemeny rule extending the previous research results.



**Keywords**

Learning Theory, Statistical Learning, Machine Learning, Applied Probabilities, Information Theory, Compuatational Social Choice, Algorithms and Complexity

# Ευχαριστίες

...

Εμμανουήλ-Βασίλειος Βλατάκης-Γκαραγκούνης

# Περιεχόμενα

# Κατάλογος Σχημάτων

# Κεφάλαιο 1

# PAC LEARNING

## 1.1 Εισαγωγή

"Τι σημαίνει ότι ο υπολογιστής μπορεί να μαθαίνει ;"
"Τι σημαίνει ότι ένας αλγόριθμος επιδιώκει να κατανοήσει την απλούστερη μορφή του κρυφού αντικειμένου που προσπαθεί να μάθει;"
"Γιατί υπάρχει πάντα μια απλή περιγραφή που εξηγεί τους νόμους ενός φαινομένου;"
και τέλος
"Πώς ο υπολογιστής μπορεί να αντιληφθεί την απλούστερη μορφή ενός αντικειμένου ενώ αδυνατεί να την αντιληφθεί ο δημιουργός του ;"

Όταν ο Turing πρότεινε τη δοκιμή που φέρει το όνομα του [Tυρ50] για να ορίσει τους στόχους της τεχνητής νοημοσύνης, πιθανόν να μην μπορούσε να φανταστεί ότι σε μόλις 50 χρόνια από τότε, το επιστημονικό πεδίο το οποίο ο ίδιος δημιούργησε με τις ευρηματικές καθολικές του μηχανές, όχι μόνο θα διατύπωνε, αλλά θα προσπαθούσε και να απαντήσει τόσο βαθιά φιλοσοφικές ερωτήσεις όπως οι παραπάνω.

Σήμερα, οι υπολογιστές είναι σε θέση να αυτοματοποιήσουν διαδικασίες, όπως αναγνώριση χαρακτήρων [ΘΚΕ14], ψηφίων [Ψ98] και εν γένει γραπτού λόγου [ΒΒ08], να κατανοήσουν την δόμηση της φυσικής γλώσσας [ΠΙΚΠ15], να αναπαραστήσουν και να παράγουν νέα μοντέλα ήχου [ΜΠ97] και εικόνας [ΜΒ96], να αξιολογήσουν και να δημιουργήσουν μορφές τέχνης [ΛΝ13], έχοντας ως γνωσιακή τους βάση απλώς ένα όγκο δεδομένων και παρατηρήσεων. Μέσω αυτών οι ερευνητές είναι πλέον σε θέση να εξάγουν γενικεύσεις, να ανιχνεύσουν πρότυπα από μια σειρά κλινικών εξετάσεων και να προστατεύσουν ή ακόμα και να θεραπεύσουν μια σειρά από χρόνιες και θανατηφόρες ασθένειες με βάση την ανάλυση των αλυσίδων ενός γονιδίου [ΗΛ04]. Έχοντας, λοιπόν, όλα αυτά τα επιστημονικά και κοινωνικά επιτεύγματα στην πλάτη της, στόχος της Επιστήμης της Μάθησης είναι η κοινότητα της Θεωρητικής Πληροφορικής να αποκτήσει επιτέλους μια συνεκτική θεωρητική βάση που θα θεμελιώνει τα μέχρι τώρα πεπραγμένα της, θα εξάγει καινούργια συμπεράσματα και θα ορίσει με σαφήνεια τους επόμενους στόχους της.

Σε καθεμία από αυτές τις ερωτήσεις οι θεωρητικοί της Μάθησης επιδίωξαν να παρουσιάσουν διαφορετικά μοντέλα [Bpo93], διαφορετικές μαθηματικές οντότητες που έχοντας οριστεί με σαφήνεια και ακρίβεια, μέσω αξιωμάτων, θεωρημάτων και αποδείξεων θα μπορούσαν να εξηγήσουν τον τρόπο λειτουργίας των περισσότερων αλγορίθμων που είχαν εφευρεθεί μέχρι τότε. Το καθένα από αυτά τα μοντέλα σίγουρα περιλαμβάνει ατέλειες, αφού στοχεύει να ερμηνεύσει διαφορετικές οντολογίες και πραγματεύεται μια διαφορετική θεώρηση για τον τρόπο με τον οποίο ο αλγόριθμος παραλαμβάνει, αξιοποιεί και εξάγει δεδομένα. Ακόμη και σήμερα ενδιαφέρον ερώτημα στον χώρο της Θεωρίας της Μάθησης αποτελεί αν τα διαφορετικά μοντέλα που έχουν προταθεί μέχρι σήμερα μπορούν να επιτύχουν να περιγράψουν και να επιλύσουν αποδοτικά το σύνολο των προβλημάτων που εντοπίζονται σε αυτό το πεδίο.

Όπως σε κάθε πρόβλημα που έχει να αντιμετωπίσει ένας επιστήμονας, έτσι και εδώ, πριν προσπαθήσει κανείς να κατανοήσει το πλήρες βάθος των δυνατοτήτων αυτού του σύνθετου πεδίου, στο οποίο η Σχεδίαση Αλγορίθμων και η Θεωρία Πολυπλοκότητας παντρεύεται με την Θεωρία Πιθανοτήτων, την Εκτιμητική Στατιστική και την Θεωρία Πληροφορίας, αξίζει να κάνει πολλά βήματα πίσω και να δει πως η Θεωρία της Μάθησης αντιμετώπισε τα πιο θεμελιακά ερωτήματα της.

## 1.2 Μαθαίνοντας διαστήματα

Η καρδιά των αλγορίθμων μάθησης είναι αρκετά εύκολη. Το πρόβλημα που θα μελετήσουμε είναι αυτό της εκμάθησης ενός διαστήματος, πρόβλημα που αποτελεί το εισαγωγικό ζήτημα στην πλειοψηφία της βιβλιογραφίας που προσεγγίζει το συγκεκριμένο πεδίο.

### 1.2.1 Ο γρίφος

Το πρόβλημα μπορεί να μοντελοποιηθεί με το ακόλουθο παιχνίδι δύο παικτών.

- Ο παίκτης 1, ο αντίπαλος, επιλέγει αρχικά ένα διάστημα στην νοητή γραμμή των πραγματικών αριθμών, στο εξής απλώς $\mathbb{R}$, καθώς και μια κατανομή $D$ με βάση την οποία θα εξάγει αριθμούς. Αυτές οι δύο αποφάσεις του θα παραμείνουν σταθερές κατά την διάρκεια του παιχνιδιού όμως δεν θα τις αποκαλύψει σε εμάς.

- Ο παίκτης 2, εμείς, έχει δικαίωμα να ζητήσει από τον παίκτη 1 να του δώσει τυχαία δείγματα με βάση την κατανομή που επέλεξε και να ενημερωθεί αν ο αριθμός που του δόθηκε βρίσκεται εντός του διαστήματος που διάλεξε στην αρχή του παιχνιδιού ο παίκτης 1.
  Π. χ. το $i$-οστό δείγμα είναι της μορφής:

  $$sample\ s_i =< number, label > \text{ όπως } < 2, \text{ΕΝΤΟΣ} >, < 3.22, \text{ΕΚΤΟΣ} > .$$

  Στόχος του παιχνιδιού είναι ο παίκτης 2 να ανακαλύψει όσο γίνεται ακριβέστερα το διάστημα που επέλεξε ο παίκτης 1 έχοντας ζητήσει τα λιγότερα δυνατά δείγματα.

- Το παιχνίδι τερματίζεται, όταν αποφασίσει ο παίκτης 2 ότι μπορεί να προτείνει το καλύτερο δυνατό διάστημα.

Είναι προφανές από μαθηματικής θεώρησης ότι ο παίκτης 2 δεν μπορεί να μαντέψει ακριβώς τα άκρα που έχει θέσει ο παίκτης 1, αφού είναι πραγματικοί αριθμοί και ο παίκτης 2 θα λάβει πεπερασμένα δείγματα[1].

Ποιος μπορεί, λοιπόν, να είναι ο στόχος μας σε αυτό το παιχνίδι;

Στο τέλος του παιχνιδιού θα υπάρχουν δύο διαστήματα:

- Αυτό που έχει ο παίκτης 1 αρχικά υποθέσει.
- Αυτό που ο παίκτης 2 πρότεινε ως λύση στο γρίφο μας.

Για να μετρήσουμε την αποδοτικότητα της εικασίας που πρότεινε ο παίκτης 2 δεν έχουμε παρά να ζητήσουμε από τον παίκτη 1 να μας δώσει ακόμη 1 δείγμα. Όπως επισημάνθηκε προηγουμένως, το δείγμα θα συνοδεύεται με μια επιπλέον υπογραφή που θα το χαρακτηρίζει ως προς την θέση του, για το εάν βρίσκεται εντός ή εκτός του διαστήματος που έχει υποθέσει ο παίκτης 1. Ταυτόχρονα ο παίκτης 2 με βάση την γνώση που συνέλεξε έχοντας διαβάσει μόνο τον αριθμό από το δοκιμαστικό δείγμα, προτείνει ποια είναι η θέση του, εντός ή εκτός του διαστήματος.

Ο αλγόριθμος που ακολούθησε ο παίκτης 2 θεωρείται επιτυχημένος όταν οι δύο αυτές υπογραφές συμφωνούν κατά πολύ μεγάλη πιθανότητα ή εναλλακτικά, αν η πιθανότητα αυτές οι δύο υπογραφές να διαφέρουν είναι αρκετά μικρή. Σε αυτή την περίπτωση λέμε ότι ο παίκτης 2 κατόρθωσε να "μάθει"το διάστημα. Παρατηρείστε ότι είναι σημαντικό, για να διατηρηθεί η δικαιοσύνη στο παιχνίδι, το τελικό δοκιμαστικό δείγμα να είναι ένα απλό δείγμα, όπως όλα τα προηγούμενα, ώστε ο παίκτης 2 να κριθεί ισότιμα με βάση την σταθερή κατανομή που αρχικά προσπαθούσε να αντιμετωπίσει.

### 1.2.2 Η λύση

Ο αλγόριθμος για το συγκεκριμένο πρόβλημα είναι αρκετά απλός.

- Αντλούμε $m$ δείγματα από τον παίκτη 1: $S = \{s_1, s_2, \cdots, s_m\}$
- Περιοριζόμαστε στα στοιχεία $InInterval = \{s_i.state = \text{`ΕΝΤΟΣ'}\}$
- Επιλέγουμε το μικρότερο και το μεγαλύτερο στοιχείο που βρίσκεται εντός του διαστήματος. $a, b = \min[S \cap InInterval], \max[S \cap InInterval]$
- Πρότεινουμε για διάστημα το $[a, b]$

Αυτό που δεν έχουμε όμως απαντήσει ακόμα είναι και το κυριότερο ερώτημα που θα μας οδηγήσει και στον τυπικό ορισμό του μοντέλου.

Πόσα δείγματα επιθυμούμε να λάβουμε δεδομένου ότι θα απαιτήσουμε το σφάλμα μας να είναι αρκετά μικρό με πολύ μεγάλη πιθανότητα;

[1] Ειδικά στην περίπτωση των συνεχών κατανομών, όπως ξέρουμε από την Θεωρία Πιθανοτήτων, η πιθανότητα να λάβουμε έναν συγκεκριμένο αριθμό είναι μηδέν, συνεπώς η πιθανότητα να λάβουμε ακριβώς τους αριθμούς που υπάρχουν στα άκρα είναι μηδενική.

## 1.3 Ο ορισμός του PAC-Learning κατά Valiant

### 1.3.1 Βασικές Έννοιες

**Ορισμός 1.1.** ***Καθολικός χώρος (domain)*** $\mathbb{X}$. *(Καθολικό) χώρο (domain)* $\mathbb{X}$ *ορίζουμε κόσμο των συναρτήσεων ή των αντικειμένων που ζουν στο μοντέλο μας.*

Π.χ $\mathbb{X} =$ { Όλα τα αυτοκίνητα στην γη }, $\mathbb{X} =$ { Όλα τα δυαδικά αρχεία που μπορούν να υπάρξουν σε ένα υπολογιστή }, $\mathbb{X} =$ { Όλα τα καταστήματα στην Νέα Υόρκη }. Συνήθως χρησιμοποιούμε κωδικοποιήσεις για τα αντικείμενα που μελετούμε. Την έννοια αυτοκίνητο, για παράδειγμα, μπορούμε να την προσδιορίσουμε με βάση διαφορετικά χαρακτηριστικά: $car[\vec{x}] = (x_1, x_2, x_3, x..., \cdots, x_n)$, όπου $x_1, x_2, x_3, x..., \cdots, x_n$, το χρώμα του αυτοκινήτου, τα χιλιόμετρα που έχει καταναλώσει και η μέση κατανάλωση βενζίνης. Επίσης μπορούμε να κάνουμε αναγωγές μεταξύ των αναπαραστάσεων. Για παράδειγμα ακόμα και αν το χαρακτηριστικό είναι κόκκινο ή κίτρινο ή μπλε μπορούμε να του δώσουμε μια αναπαράσταση στο $\mathbb{X} = \{0, 1\}^n$.

**Ορισμός 1.2.** ***Σύλληψη (concept)*** $c$. *Μια σύλληψη (concept)* $c$ *είναι ένα υποσύνολο του χώρου μας* $\mathbb{X}$.

Για παράδειγμα: $c =$ { Αυτοκίνητα με μέση κατανάλωση κάτω από 5 λίτρα/χλμ}. Η έννοια της σύλληψης μπορεί να μοντελοποιηθεί, όπως και κάθε σύνολο αυτού του χώρου, σε μια αντίστοιχη δείκτρια δυαδική συνάρτηση:

$$\hat{c} : \mathbb{X} \to \{0, 1\} \quad \hat{c}(x) = \begin{cases} 1, x \in c \\ 0, otherwise \end{cases}$$

Στο εξής θα χρησιμοποιούμε αυτή την δεύτερη σε σειρά σημασιολογία. Επίσης, για να μην επιβαρύνουμε τον συμβολισμό θα χρησιμοποιούμε τον συμβολισμό $c(x)$ ως συνάρτηση σύλληψης. Σε κάθε άλλη περίπτωση θα διευκρινίζεται με σαφή τρόπο.

**Ορισμός 1.3.** ***Κλάση συλλήψεων (concept class)*** $\mathbb{C}$. *Κλάση συλλήψεων είναι ένα σύνολο από συλλήψεις, από διαφορετικές δηλαδή δυαδικές συναρτήσεις πάνω στο χώρο* $\mathbb{X}$.

Στόχος κάθε σχεδιαστή αλγορίθμων μάθησης αποτελεί έχοντας γνωστή την μορφολογία της κλάσης των συλλήψεων που μελετά, να προτείνει μια δυαδική συνάρτηση που θα συμπίπτει, έστω και προσεγγιστικά, στο βέλτιστο δυνατό βαθμό, με την σύλληψη από την οποία αντλεί πληροφορία. Προσοχή όμως, αναλόγως του μοντέλου, ο αλγόριθμος μπορεί να είναι ή να μην είναι ικανός να προτείνει μια σύλληψη που να αναφέρεται ακριβώς στο $\mathbb{X}$, αλλά μια σύλληψη από ένα συγγενές σύνολο π.χ $\mathbb{Y}$.

Αν για τον αναγνώστη φαίνεται περίεργο αυτό, ας αναλογιστεί διάφορα συνδυαστικά μαθηματικά αντικείμενα για τα οποία έχουμε έναν πολύ σαφή τρόπο να τα αναγνωρίζουμε, αλλά όχι να τα αναπαράγουμε. Π.χ οι γλώσσες των NP προβλημάτων. Μπορούμε γρήγορα να αναγνωρίζουμε με σαφή τρόπο αν ένα πρόβλημα ανήκει στο σύνολο των NP προβλημάτων, αλλά δεν είναι τετριμμένος ο τρόπος με τον οποίο μπορούμε να αναπαράγουμε όλον τον χώρο. Συνεπώς, αναλόγως του προβλήματος και των ικανοτήτων περιγραφής του, ο σχεδιαστής μπορεί να διαλέγει από ένα διαφορετικό καθολικό χώρο, έστω $\mathbb{Y}$. Ασφαλώς, όσο

πιο ικανός και αποδοτικός είναι ο αλγόριθμος τόσο περισσότερο θα ταιριάζουν σε ιδιότητες και αντικείμενα οι δύο χώροι $\mathbb{X}, \mathbb{Y}$.

**Ορισμός 1.4.** ***Υπόθεση (hypothesis)*** *h. Μια υπόθεση (hypothesis) h είναι ένα υποσύνολο του χώρου* $\mathbb{Y}$*.*

Ομοίως με πριν, η έννοια της υπόθεσης μπορεί να μοντελοποιηθεί, όπως και κάθε σύνολο και αυτού του χώρου, σε μια αντίστοιχη δείκτρια δυαδική συνάρτηση:

$$\hat{h} : \mathbb{Y} \to \{0,1\} \quad \hat{h}(x) = \begin{cases} 1, x \in h \\ 0, otherwise \end{cases}$$

Στο εξής θα χρησιμοποιούμε αυτή την δεύτερη σε σειρά σημασιολογία. Επίσης, για να μην επιβαρύνουμε τον συμβολισμό θα χρησιμοποιούμε τον συμβολισμό $h(x)$ ως συνάρτηση υπόθεσης. Σε κάθε άλλη περίπτωση θα διευκρινίζεται με σαφή τρόπο.

Αντιστοίχως με την κλάση συλλήψεων μπορούμε να ορίσουμε και την κλάση υποθέσεων.

**Ορισμός 1.5.** ***Κλάση υποθέσεων (hypothesis class)*** $\mathbb{H}$*. Το σύνολο των πιθανών συναρτήσεων υπόθεσης που μπορεί να προτείνει ένας αλγόριθμος μάθησης αναλόγως των δεδομένων που τελικώς θα λάβει καλείται κλάση υποθέσεων (hypothesis class).*

**Ορισμός 1.6.** ***Το Μαντείο των Δεδομένων****. Ο μηχανισμός πληροφόρησης - διεπαφής από όπου ο αλγόριθμος θα ζητά περισσότερα δείγματα του χώρου* $\mathbb{X}$ *με βάση την κατανομή D, μπορεί να προσομοιωθεί με ένα κλασσικό Oracle* $(EX(c, D))$*.*

Σε αυτό το μαντείο, ο μηχανισμός μάθησης μπορεί να θέτει αιτήματα για καινούργια δείγματα και να λαμβάνει τα προσεσημασμένα στοιχεία του χώρου $\mathbb{X}$ που επιλέγονται τυχαία με βάση την κατανομή $D$.

Κλείνοντας την πρώτη παρουσίαση των ορισμών, προκύπτει η πρώτη διαχωριστική γραμμή που εμφανίζεται μεταξύ των διαφορετικών αλγορίθμων μάθησης.

**Ορισμός 1.7.** ***(Μη) Κανονικοί αλγόριθμοι μάθησης ((Im)Proper Learning Algorithm)*** *(Μη) Κανονικοί αλγόριθμοι μάθησης ονομάζονται οι αλγόριθμοι για τους οποίους, η κλάση υποθέσεων (δεν) ταυτίζεται με την κλάση συλλήψεων*

$$\{\mathbb{Y}(\neq) = \mathbb{X}\} \text{ ή } \{\mathbb{H}(\neq) = \mathbb{C}\}$$

### 1.3.2 PAC Learning Model

Στο μοντέλο μάθησης που πρότεινε ο Leslie Valiant και έγινε ο γνωστό ως PAC [ἄλ84] ο τρόπος με τον οποίο ένας αλγόριθμος μάθησης είναι ικανός να εκπαιδευτεί γύρω από μια ομάδα - κλάση συλλήψεων $\mathbb{C}$ χρησιμοποιώντας μια ομάδα - κλάση υποθέσεων $\mathbb{H}$ είναι ο ακόλουθος:

- Υπάρχει μια συγκεκριμένη, άγνωστη στον εκπαιδευόμενο, σύλληψη $c$ η οποία σε όλη την διαδικασία μάθησης παραμένει σταθερή και κρυφή.

- Ο εκπαιδευόμενος δεν γνωρίζει την $c$, γνωρίζει όμως το συνολικό χώρο στο οποίο ανήκει $\mathbb{C}$.
- Υπάρχει μια συγκεκριμένη, άγνωστη στον εκπαιδευόμενο, κατανομή $D$ των στοιχείων του $\mathbb{X}$ η οποία σε όλη την διαδικασία μάθησης παραμένει σταθερή και κρυφή.
- Ο εκπαιδευόμενος λαμβάνει $m$ προ - σεσημασμένα δείγματα $< x, c(x) >$ όπου το κάθε ένα δείγμα θα εξάγεται ανεξαρτήτως των άλλων από τον χώρο $\mathbb{X}$ και τυχαία με βάση την κατανομή $D$. Εναλλακτικά μπορούμε να πούμε ότι ο εκπαιδευόμενος έχει $m$ προσβάσεις στο Oracle ($EX(c, D)$).
- Ο εκπαιδευόμενος εξάγει μια συνάρτηση $h \in \mathbb{H} : \mathbb{Y} \to \{0, 1\}$ ως τελική προσέγγιση της άγνωστης σε αυτόν $c$.
- Για απλότητα του μοντέλου, ορίζουμε ότι $\mathbb{Y} \equiv \mathbb{X}$. Μπορεί κανείς να το υποθέσει χωρίς βλάβη της γενικότητας, αφού μπορεί να θεωρήσει ότι $\mathbb{X}' \equiv \mathbb{Y}' \equiv \mathbb{X} \cup \mathbb{Y}$. Όμως και πάλι δεν είναι υποχρεωτικό ότι $\mathbb{H} \equiv \mathbb{C}$.

**Ορισμός 1.8.** ***Σφάλμα αλγορίθμου****. Έστω δύο συναρτήσεις $h, c : \mathbb{X} \to \{0, 1\}$, οι συναρτήσεις υπόθεσης και σύλληψης και έστω μια κατανομή $D$ πάνω στα αντικείμενα του χώρου $\mathbb{X}$. Ορίζουμε ως σφάλμα της $h$ ως προσέγγιση για την $c$ με δεδομένη την κατανομή $D$ στον υποκείμενο χώρο $\mathbb{X}$ την ποσότητα*

$$er_D(c, h) = \Pr_{x \sim D}[h(x) \neq c(x)]$$

1. Ας τονίσουμε πάλι ότι δεν θα πρέπει να περιμένει κανείς ότι ο αλγόριθμος θα έχει εξασφαλισμένα μηδενικό λάθος. Αν η κατανομή σταθμίζει με πολύ μικρή πυκνότητα πιθανότητας, μη μηδενική, μια περιοχή του χώρο $\mathbb{X}$, είναι πολύ φυσικό κατά την διαδικασία μάθησης ο αλγόριθμος να μην έχει συναντήσει αυτή την πληροφορία και συνεπώς κατά την διάρκεια δοκιμής να υπάρχει πιθανότητα να αποτύχει στο δοκιμαστικό στοιχείο.
2. Επίσης, δεν είναι φυσικό να υπάρχει 100% βεβαιότητα ότι ο αλγόριθμος δεν θα ξεφύγει του αλγοριθμικού σφάλματος που πιθανοτικώς εγγυάται. Ο λόγος είναι ότι υπάρχει πάντα η πιθανότητα το σύνολο δεδομένων που μας δόθηκε (data set) να είναι αρκετά άσχημο, ώστε να αυξήσει την πιθανότητα λάθους.
3. Επίσης, δεν είναι φυσικό να υπάρχει 100% βεβαιότητα ότι ο αλγόριθμος δεν θα ξεφύγει του αλγοριθμικού σφάλματος που πιθανοτικώς εγγυάται. Ο λόγος είναι ότι υπάρχει πάντα η πιθανότητα το σύνολο δεδομένων που μας δόθηκε (data set) να είναι αρκετά άσχημο, ώστε να αυξήσει την πιθανότητα λάθους.

Πριν προχωρήσει ο αναγνώστης στον επόμενο ορισμό θα πρέπει να γίνει σαφές ότι η ποσότητα $er_D(c, h)$ δεν είναι ντετερμινιστική αλλά τυχαία, αφού η $h$ ως προς τα πλήθος των δειγμάτων αποτελεί, μια τυχαία μεταβλητή στο χώρο των συναρτήσεων υπόθεσης. Φυσικά για δεδομένο πλήθος δειγμάτων και αλγορίθμου μάθησης η $h$ προσδιορίζεται ντετερμινιστικά.

Με άλλα λόγια, το σφάλμα της κατανομής προσδιορίζεται από το σύνολο των δεδομένων που θα δοθούν στον αλγόριθμο. Αφού τα δεδομένα που έρχονται στον αλγόριθμο είναι τυχαία — με κατανομή μάλιστα σταθερά επιλεγμένη την $D$ — συνεπώς και η επιλογή της συνάρτησης υπόθεσης που θα εξορύξει ο αλγόριθμος μάθησης για κάποια δεδομένη συνάρτηση σύλληψης παραμένει επίσης τυχαία.

Με βάση τα παραπάνω μπορεί να δώσει κανείς τον πρώτο ορισμό της κατά PAC εκμαθησιμότητας (PAC Learnability).

**Ορισμός 1.9.** ***PAC Learning Αλγόριθμος****. Ένας αλγόριθμος $A$ μπορεί να* ***μάθει κατά PAC*** *μια κλάση συλλήψεων $\mathbb{C}$ χρησιμοποιώντας μια κλάση υποθέσεων $\mathbb{H}$, εάν δεδομένου $\epsilon, \delta$ και έχοντας πρόσβαση στο μαντείο $EX(c, d)$ ο αλγόριθμος $A$ για οποιαδήποτε σύλληψη $c \in \mathbb{C}$ είναι σε θέση να μας προτείνει μια υπόθεση $h \in \mathbb{H}$, ώστε με πιθανότητα τουλάχιστον $1 - \delta$ το σφάλμα να είναι το πολύ $\epsilon$:*

$$\Pr_{Samples \sim D^m}[er_D(c, h^{(Samples)}) < \epsilon] > 1 - \delta$$

- *Σημείωση:*
  - *$\epsilon$: Ακρίβεια σφάλματος*
  - *$\delta$: Αβεβαιότητα επιτυχίας του αλγορίθμους*
  - *$m$: Συνολικός αριθμός στοιχείων εκμάθησης.*
  - *Η τυχαιότητα βρίσκεται πάνω στα τυχαία $m$ δείγματα που εξάγει ο αλγόριθμος από το Oracle και προσδιορίζουν την $h$ και οποιαδήποτε άλλη πιθανή εσωτερική τυχαιότητα της διαδικασίας του αλγορίθμου*

Βλέποντας ,όμως, κανείς τον ορισμό δεν αντιλαμβάνεται την συμμετοχή της Θεωρητικής Πληροφορικής, αφού τα περισσότερα προβλήματα στα πλαίσια της ασυμπτωτικής στατιστικής πράγματα έχουν μελετηθεί και έχουν όλες τις απαιτούμενες καλές ιδιότητες. Συνεπώς οφείλουμε ενισχύσουμε τον προηγούμενο ορισμό.

**Ορισμός 1.10.** ***PAC Learnability.*** *Ένας αλγόριθμός $A$ μπορεί να* ***μάθει κατά PAC αποδοτικά (efficiently)*** *μια κλάση συλλήψεων $\mathbb{C}$ χρησιμοποιώντας μια κλάση υποθέσεων $\mathbb{H}$, εάν μπορεί να την μάθει κατά PAC (Ορισμός 1.9) και ταυτόχρονα η χρονική πολυπλοκότητα και το μέγεθος του δείγματος (time & sample complexity) είναι συνάρτηση $Poly(1/\epsilon, 1/\delta)$. Επίσης αν η κλάση συλλήψεων είναι παραμετρική στο μέγεθος της από κάποια μεταβλητή $n$, τότε θα πρέπει να είναι πολυωνυμική και ως προς αυτή την παράμετρο, (time & sample complexity=$Poly(1/\epsilon, 1/\delta, n)$)*

#### 1.3.2.1 Πίσω στο γρίφο

Ας δούμε και τυπικά τώρα γιατί το πρόβλημα ανακάλυψης του διαστήματος είναι και PAC αποδοτικά εκμαθήσιμο (PAC efficiently learnable). Για αρχή έχουμε ότι μας είναι γνωστή η μορφή της κλάσης των συλλήψεων. $\mathbb{C} = \{[a, b], \forall a, b : a < b\}$. Συνεπώς η συνάρτηση σύλληψης και τα πιθανά δείγματα από κάποια κατανομή είναι τα ακόλουθα:

$$c = [a, b] = [.43, .67].\ c(.52) = 1,\ c(.85) = 0$$

Τα δείγματα έχουν επιλεχθεί τυχαία από μια σταθερά άγνωστη και επιλεγμένη κατανομή $D$. Αν ο αλγόριθμος μας προτείνει το διάστημα $[a', b']$, τότε το σφάλμα μπορεί να υπολογιστεί ως[2] :

$$er_D(c, h) = \Pr_{x \sim D}[x \in [a, a'] \cup [b', b]]$$

Από Union-Bound είναι φανερό ότι :

$$er_D(c, h) = \Pr_{x \sim D}[x \in [a, a'] \cup [b', b]] \leq \Pr_{x \sim D}[x \in [a, a']] + \Pr_{x \sim D}[x \in [b', b]] \quad (\star)$$

Με βάση την αρχική θεώρηση του PAC εμείς επιτρέπουμε στον αλγόριθμο μας να έχει σφάλμα το πολύ $\epsilon$ με πολύ μεγάλη πιθανότητα.

$$\Leftrightarrow$$

Αρκεί να δείξουμε ότι το να συμβεί τουλάχιστον $\epsilon$ λάθος έχει πολύ μικρή πιθανότητα.

- Για να αποδείξουμε ότι είναι εφικτό αυτό, θα δείξουμε ότι:

Η πιθανότητα έστω μια περιοχή από τις δύο $([a, a'], [b', b])$ να έχει μέγεθος μεγαλύτερο από $\epsilon/2$ είναι πολύ μικρή, π.χ φραγμένη από $\leq \delta$.

$$\Leftrightarrow$$

Η πιθανότητα και οι δύο περιοχές να είναι μικρότερες του $\epsilon/2$ είναι πολύ μεγάλη $\geq 1 - \delta$. Από το τελευταίο μέσω της $\star$, προκύπτει ότι το $er_D(c, h) \leq \epsilon$ συμβαίνει με πολύ μεγάλη πιθανότητα.

Για να μην επιβαρύνουμε τον συμβολισμό θα ονομάσουμε με $\alpha, \beta$, τους αριθμούς για τους οποίους ισχύει ότι $\Pr_{x \sim D}[x \in [a, \alpha]] = \Pr_{x \sim D}[x \in [\beta, b]] = \epsilon/2$.

Συνεπώς:

- Ποια είναι η πιθανότητα ο αλγόριθμος να έχει ανιχνεύσει ως κάτω άκρο έναν αριθμό $a' > \alpha$; *Απάντηση*:

$$\Pr[\min s_i \notin [a, \alpha]] = \Pr[\cap\{s_i > \alpha\}] = \prod \Pr[\{s_i > \alpha\}] = (1 - \epsilon/2)^m$$

[2] *Σημείωση*: Η εκτίμηση του αλγορίθμου αποτελεί μια μόνιμη υποεκτίμηση του πραγματικού διαστήματος, υπό την έννοια ότι $[a', b'] \subset [a, b]$.

- Ομοίως για το άνω όριο:

$$\Pr[\max s_i \notin [\beta, b]] = \Pr[\cap\{s_i < \beta\}] = \prod \Pr[\{s_i < \beta\}] = (1-\epsilon/2)^m$$

Άρα η πιθανότητα να έχουμε ανακαλύψει το πολύ μέχρι το $[\alpha, \beta]$, αντί ολόκληρης της περιοχής είναι:

$$\Pr[\ [\alpha,\beta] \not\subset [a',b']] \leq 2(1-\epsilon)^m \leq 2e^{-\epsilon m} \leq \delta, \text{ αφού } (1-x) \leq e^{-x}$$

Η παραπάνω πιθανότητα θέλουμε να είναι αρκετά μικρή γιατί εκφράζει την πιθανότητα το λάθος να είναι μεγαλύτερο του $\epsilon$. Άρα

$$2e^{\epsilon m} \geq \frac{1}{\delta} \Leftrightarrow \boxed{m \geq \frac{1}{\epsilon}\log(\frac{1}{\delta})}$$

Συνεπώς το πρόβλημα ανακάλυψης του διαστήματος στον χώρο των πραγματικών αριθμών είναι PAC αποδοτικά εκμαθήσιμο (PAC efficiently learnable), αφού ο αλγόριθμος εμφανίζει τόσο πολυωνυμική δειγματοληπτική, όσο και πολυωνυμική υπολογιστική πολυπλοκότητα .

**Παρατήρηση 1.1.**

1. *$\delta$: Η παράμετρος αισιοδοξίας "πιέζει"το μέγεθος του δείγματος προς τα πάνω, ώστε να αποφύγει η πλειοψηφία του δείγματος να είναι κατά μέση περίπτωση αρκετά συχνά εμφανιζόμενη και όχι σπάνια*

2. *$\epsilon$: Η παράμετρος ακρίβειας "πιέζει"το μέγεθος του δείγματος προς τα πάνω, ώστε να έχει εξασφαλιστεί ότι δεδομένης της ποιότητας του δείγματος το μέγεθος του είναι αρκετό, ώστε η προταθείσα συνάρτηση υπόθεσης να είναι αρκετά κοντά στην συνάρτηση σύλληψης.*

Πριν ολοκληρώσει κανείς την πρώτη μελέτη με το PAC ως μοντέλο, θα πρέπει να σχολιάσει δύο πολύ σημαντικά τεχνικά ζητήματα που αφορούν την μορφή & περιγραφή της συνάρτησης σύλληψης και υπόθεσης.

Ζήτημα 1ο: Το 'μέγεθος' της πληροφορίας που απαιτεί η ίδια η συνάρτηση σύλληψης για να αναπαρασταθεί.

Μια πρώτη προσέγγιση θα μπορούσε να είναι ότι :

$$size(c) = \#\ bits\ required\ to\ describe\ c \in \mathbb{C},$$

δηλαδή το μέγεθος της συνάρτησης σύλληψης να αντιστοιχεί απλά στην πληροφορία που χρειάζεται για να την προσδιορίσει κανείς εντός της κλάσης. Χαρακτηριστικό παράδειγμα αποτελούν οι DNF μορφές της Μαθηματικής Λογικής. Θα μπορούσε κάποιος να μετρήσει το μέγεθος τους διαισθητικά, για παράδειγμα, με βάση το πλήθος

των όρων που περιέχει μια τέτοια έκφραση. Θα περίμενε κανείς ότι για να μάθει μια DNF φόρμουλα των 10 και των 1000 όρων θα όφειλε να θυσιάσει διαφορετική ποσότητα υπολογιστικών πόρων. Πράγμα, όμως, που δεν είναι αληθές, γιατί κάθε 3-DNF φόρμουλα μπορεί κανείς να την περιγράψει ακόμα και με 1000 όρους.

Συνεπώς για να ορίσουμε με ακρίβεια τι σημαίνει μαθαίνω μια συνάρτηση σύλληψης, θα πρέπει να ορίσουμε σε ποια μορφή επιθυμούμε να την μαθαίνουμε.

**Ορισμός 1.11.** ***Representation Function Size.*** *Μέγεθος μια συνάρτησης σύλληψης $c$ ορίζεται το πλήθος bits πληροφορίας που χρειάζεται για να αναπαρασταθεί μικρότερη δυνατή μορφή ισοδύναμης συνάρτηση εντός στην κλάση $\mathbb{C}$.*

$$size(c) = \min_{\hat{c} \in \mathbb{C}: \hat{c} \equiv c} [representation\ bits(\hat{c})]$$

**Διόρθωση #1** Για να είναι ένα αλγόριθμος PAC αποδοτικός θα πρέπει να είναι πολυωνυμικός σε χρόνο και ως προς το πλήθος την παράμετρο $size(c)$.

Ζήτημα 2ο: Η μορφή της συνάρτησης υπόθεσης που θα παράξει ο αλγόριθμος και η υπολογιστική ευχρηστία.

Σε πολλές περιπτώσεις η παρουσία του μεγέθους της συνάρτησης σύλληψης στην πολυπλοκότητα δεν σηματοδοτεί κάτι σημαντικό, αφού συνήθως είναι φραγμένο από κάποια παράμετρο. Για παράδειγμα για να μάθει κανείς αποδοτικά τις CNF μορφές θα πρέπει να έχει έναν αλγόριθμο αποδοτικό στο $poly(n, size(c), 1/\epsilon, 1/\delta)$, αλλά το $size(c) \leq n$.

Όμως, ένα σημαντικότερο ζήτημα αποτελεί η συνάρτηση υπόθεσης $h(x)$ και η μορφή με την οποία ο αλγόριθμός μας θα την αποθηκεύσει για χρήση από τον χρήστη. Για παράδειγμα, αν η συνάρτηση υπόθεσης απαιτεί εκθετικό χώρο για να περιγραφεί ή έχει συμπαγή περιγραφή αλλά η ταχύτητα υπολογισμού της είναι υπερπολυωνυμική, τότε το αποτέλεσμα του αλγορίθμου μας είναι πρακτικώς άχρηστη. [3] Συνεπώς:
**Διόρθωση #2** Για να είναι ένα αλγόριθμος PAC αποδοτικός θα πρέπει να είναι πολυωνυμικά υπολογιστή, δηλαδή:

$\forall x \in \mathbb{X}, h \in \mathbb{H}\ h(x)$ πρέπει να είναι πολυωνυμικά υπολογίσιμη και περιγράψιμη

[3] Αυτό σημαίνει ότι η συνάρτηση υπόθεσης δεν είναι υποχρεωτικό να αποτελεί με τη στενή μαθηματική έννοια μια κλειστή φόρμουλα, αλλά με την ευρύτερη έννοια της πληροφορικής να υπάρχει αποδοτικός και σύντομα περιγράψιμος αλγόριθμος που μπορεί σε κάθε $x$ να υπολογίσει το $h(x)$.

## 1.4 Ανακεφαλαιώνοντας

Το PAC ή αναλυτικότερα το Probably Approximate Correct Model, όπως επεξηγούν και τα αρχικά του, αποτελεί ένα από τα βασικά μοντέλα τυπικής περιγραφής των αλγορίθμων μάθησης. Είναι ένα μοντέλο που προτάθηκε από τον Leslie Valiant το 1984 και για το οποίο βραβεύτηκε λίγα χρόνια αργότερα με Turing Award. Το όνομα του μοντέλου εκφράζει τις δύο βασικές αρχές που πρέπει να ακολουθήσει κάθε αλγόριθμος του μοντέλου. Ο αλγόριθμος προσπαθεί να μάθει Approximate Correct λύση, δηλαδή προσεγγιστικά όσο γίνεται σωστότερα το κρυφό αντικείμενο του χώρου, ενώ του δίνεται και η ελαστικότητα να είναι Probably Correct, αφού το μοντέλο επιτρέπει στον αλγόριθμο μάθησης μία μικρή πιθανότητα αστοχίας. Στο επόμενο κεφάλαιο θα δούμε πως επακτάθηκαν αυτοί οι ορισμοί και το γενικό πλαίσιο του μοντέλου, ώστε να μάθει κανείς την συνάρτηση κατανομής που ακολουθεί ένα πλήθος δεδομένων.

# Κεφάλαιο 2

# Μαθαίνοντας κατανομές

## 2.1 Εισαγωγή

Στην επιστημονική μελέτη ένα από τα βασικότερα στάδια αποτελεί η διαμόρφωση μοντέλων που ικανοποιούν το πλήθος των δεδομένων με τα οποία πειραματίζεται ο ερευνητής. Η ανακάλυψη αυτής της κρυφής δομής που υπάρχει συνήθως στο μεγαλύτερο πλήθος των δεδομένων αποτελεί έναν από τους ακρογωνιαίους λίθους της σύγχρονης θεωρίας ανάλυσης δεδομένων. Το πρόβλημα πρόκληση του 21ου αιώνα έχει όνομα, αιτία και μέχρι τώρα όχι ευκρινή λύση. Το όνομα του: Big Data[ΛΛ̃Ζ13, ΛΒΩ$^{+}$13, Λ̃Λ̃13].

Όσο και αν οι υπολογιστικοί πόροι αυξάνονται, όσο και αν η πρόοδος των υλικών και της αρχιτεκτονικής υπολογιστών καθιστά την ίδια την ικανότητα αποθήκευσης μεγάλου όγκου δεδομένων σε πολύ μικρότερους χώρους εφικτή, η πρόκληση παραμένει. Ο λόγος απλός: Οι ερωτήσεις που καλούνται τώρα να απαντήσουν οι επιστήμονες είναι ακόμη δυσκολότερες. Χαρακτηριστικό παράδειγμα, η επένδυση της αυστραλιανής κυβέρνησης στην κατασκευή του μεγαλύτερου κέντρου παρατηρήσεων αστρικών φαινομένων με πάνω από 1000 τηλεσκόπια που το δευτερόλεπτο θα καταγράφουν πάνω από 1 PetaByte. Αυτό είναι και το σημείο στο οποίο καλείται η σύγχρονη Θεωρητική Πληροφορική να παίξει σημαίνοντα ρόλο. Ακόμη και αν κανείς επιθυμούσε να συμπιέσει τα δεδομένα ώστε στο μέλλον, όταν η πρόοδος της επιστήμης θα είναι ικανή να διαχειριστεί αυτόν τον όγκο δεδομένων με καλύτερο τρόπο και να καταλάβει τότε την κρυφή δομή των υπαρχόντων δεδομένων, το πρόβλημα ανεύρεσης της πραγματικής μορφής τους επανέρχεται αυτοαναφορικά. Ο λόγος είναι απλός: ο καλύτερος τρόπος να συμπιέσει κανείς δεδομένα είναι να γνωρίζει το μοντέλο και την κατανομή που ακολουθούν. [Ρυβ12]

Συνεπώς, στο πυρήνα των ερωτημάτων που καλείται να απαντήσει η Θεωρία της Μάθησης βρίσκεται η ανεύρεση κατανομών από το σύνολο των δεδομένων. Το πρόβλημα αυτό έχει μελετηθεί ευρέως από το σύνολο της στατιστικής και εν γένει μαθηματικής κοινότητας από τις αρχές του 19ου αιώνα με πρωτοπόρο τον Pearson[Πεα95], ο οποίος πρώτος μελέτησε το απλούστερο αλγόριθμο της στατιστικής θεωρίας, την κατασκευή ιστογράμματος και την εξαγωγή της εμπειρικής κατανομής.

Υπάρχει κάτι καλύτερο που μπορεί κανείς να κάνει για να μάθει μια κατανομή από ένα Ιστόγραμμα;
&
Ποια είναι τα ελάχιστα δείγματα που χρειάζεται κανείς να αποκτήσει, ώστε ένα Ιστόγραμμα να είναι αντιπροσωπευτικό της πραγματικότητας;

Και στα δύο ερωτήματα η απάντηση δόθηκε λίγα χρόνια αργότερα, στα πρώτα χρόνια της Θεωρίας Πληροφορίας και θα την δούμε στην συνέχεια του κεφαλαίου.

Κατά την διάρκεια των τελευταίων δύο δεκαετιών υπάρχει ένας μεγάλος όγκος αποτελεσμάτων στην Θεωρητική Πληροφορική που έχουν ως αμιγή στόχο να μελετήσουν την υπολογιστική αποδοτικότητα των αλγοριθμικών τεχνικών που εφαρμόζει η Εκτιμητική Στατιστική στα μοντέλα της.

Στην πραγματικότητα ξεκλειδώνοντας κανείς τα μυστικά αυτής της επιστήμης δίνει πρόσφορο έδαφος σε μια πλειάδα άλλων επιστημών που στηρίζουν την γνώση τους στην κατανόηση κατανομών, όπως η Υπολογιστική Γλωσσολογία[ΑΡ$^{\sim+}$15, ΠΙΚΠ15], η Υπολογιστική Θεωρία Κοινωνικών Επιλογών[ΛΔΤΞ16, ΓΜΜ$^{+}$15, ΖΠΞ16] και η Στατιστική Φυσική [ΣΚΕ$^{+}$14] και την Αλγοριθμική Θεωρία Παιγνίων [ΛΣΤ16, ΣΚΣ16, ΣΑΛΣ15]

Το κεφάλαιο αυτό επιδιώκει να φιλοξενήσει κάποια από τα κλασσικότερα αποτελέσματα του προηγούμενου αιώνα, καθώς και μια ομάδα από αποτελέσματα της τελευταίας δεκαετίας που έρχονται να συμπληρώσουν την προηγούμενη θεωρία με σαφώς ισχυρότερες υπολογιστικές τεχνικές.

## 2.2 Βασικές Έννοιες & Προαπαιτούμενα

### 2.2.1 Ορισμοί από την Στοιχειώδη Θεωρία Πιθανοτήτων

Υποθέτουμε ότι ο αναγνώστης είναι γνώστης των βασικών εργαλείων Θεωρίας Πιθανοτήτων. Για λόγους πληρότητας θα αναφέρουμε τους βασικούς ορισμούς ώστε στην συνέχεια να είναι ευκολότερο να εμβαθύνουμε σε λιγότερο τετριμμένες δομές-εργαλεία.

**Ορισμός 2.1.** ***Χώρος πιθανότητας****. Με τον όρο χώρος πιθανότητας εννοούμε την τριπλέτα* $(\Omega, \mathcal{F}, \Pr)$ *όπου* $\Omega$ *είναι το σύνολο όλων των πιθανών ενδεχομένων,* $\mathcal{F}$ *είναι μια σ-άλγεβρα υποσυνόλων του* $\Omega$ *και* $\Pr : \mathcal{F} \to \mathbb{R}$ *είναι ένα μέτρο πιθανότητας.*

Η σ-άλγεβρα περιέχει όλα τα γεγονότα που 'μπορούμε' να παρατηρήσουμε. Για παράδειγμα, η σ-άλγεβρα $\{\emptyset, \Omega\}$ δηλώνει ότι το μόνο που μπορούμε να παρατηρήσουμε είναι αν κάτι συμβαίνει ή δε συμβαίνει τίποτε. Η σ-άλγεβρα $\{\emptyset, \Omega, A, A^c\}$ δηλώνει ότι μπορούμε να παρατηρήσουμε αν το γεγονός A συνέβη ή δε συνέβη ή αλλιώς αν συνέβη « κάτι » ή τίποτε. Όσο περισσότερα στοιχεία περιλαμβάνει η $\mathcal{F}$, τόσο πιο ευαίσθητη μπορεί να είναι η παρατήρησή μας σχετικά με τα γεγονότα που λαμβάνουν χώρα.

**Ορισμός 2.2.** ***Μέτρο πιθανότητας****. Μέτρο πιθανότητας* $\Pr$ *ορίζεται μια συνολοσυνάρτηση από μια σ-άλγεβρα* $\mathcal{F}$ *στο* $\mathbb{R}$ *όταν ικανοποιούνται τα παρακάτω αξιώματα:*

1. $\Pr(A) \geq 0, \forall A \in \mathcal{F}$

2. $\Pr(\Omega) = 1$

3. Αν $A_1, A_2, A_3, \ldots \in \mathcal{F}$ με $A_i \bigcap A_j = \emptyset, \forall i \neq j$ τότε $\Pr(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \Pr(A_n)$

**Ορισμός 2.3.** ***Τυχαία μεταβλητή****. Μια τυχαία μεταβλητή είναι μια πραγματική συνάρτηση που ορίζεται σε ένα δειγματικό χώρο $\Omega$, δηλαδή μια συνάρτηση της μορφής $X : \Omega \to \mathbb{R}$ ή $X : \Omega \to A$, όπου $A$ είναι ένα υποσύνολο των πραγματικών αριθμών ή γενικότερα κάποιου μετρήσιμου χώρου.*

Η τυχαία μεταβλητή εκφράζει συνήθως κάποια μέτρηση που γίνεται στο τυχαίο αποτέλεσμα. Για παράδειγμα αν ο δειγματικός χώρος είναι ένα σύνολο ατόμων, η τυχαία μεταβλητή μπορεί να εκφράζει το ύψος του τυχαία επιλεγμένου ατόμου.

- Αν το πεδίο τιμών Α της τυχαίας μεταβλητής είναι ένα διακριτό σύνολο, για παράδειγμα ένα πεπερασμένο σύνολο ή οι ακέραιοι αριθμοί, τότε η τυχαία μεταβλητή ονομάζεται διακριτή.

- Αν το πεδίο τιμών είναι ένα ή περισσότερα διαστήματα πραγματικών αριθμών, τότε η τυχαία μεταβλητή ονομάζεται συνεχής.

Θα ολοκληρώσουμε με τις τρεις βασικές συναρτήσεις που ορίζουν την τυχαία συμπεριφορά μιας τυχαίας μεταβλητής και τον ορισμό της μέσης τιμής και διασποράς.

**Ορισμός 2.4.** ***Συνάρτηση κατανομής****. Έστω ένας χώρος πιθανότητας $(\Omega, \mathcal{F}, \Pr)$ και μια πραγματική τυχαία μεταβλητή $X : \Omega \to \mathbb{R}$ πάνω σε αυτόν. Η συνάρτηση $F_X : \mathbb{R} \to [0, 1]$ με*

$$F_X(x) = \Pr(X \leq x) = \Pr\left(\{\omega \in \Omega \mid X(\omega) \leq x\}\right)$$

*ονομάζεται συνάρτηση κατανομής (σ.κ.) της τυχαίας μεταβλητής $X$.*

**Ορισμός 2.5.** ***Συνάρτηση μάζας πιθανότητας****. Έστω ένας χώρος πιθανότητας $(\Omega, \mathcal{F}, \Pr)$ και μια διακριτή τυχαία μεταβλητή $X : \Omega \to D$ πάνω σε αυτόν. Η συνάρτηση $p_X : D \to [0, 1]$ με*

$$p_X(x) = \Pr(X = x)$$

*ονομάζεται συνάρτηση μάζας πιθανότητας (σ.μ.π) της τυχαίας μεταβλητής $X$.*

Για μια διακριτή τυχαία μεταβλητή που παίρνει τιμές $x_1, x_2, \cdots, x_n$ με συνάρτηση μάζας πιθανότητας $p_X(x_i) = \Pr(X = x_i)$ η αντίστοιχη συνάρτηση κατανομής ισούται με

$$F_X(x) = \Pr(X \leq x) = \sum_{x_i \leq x} \Pr(X = x_i) = \sum_{x_i \leq x} p_X(x_i)$$

Επίσης επειδή $\Pr(\Omega) = 1 \Rightarrow \lim_{x \to +\infty} F_X(x) = 1 = \sum_{x_i \in D} \Pr(X = x_i)$
Στην συνεχή περίπτωση είναι λίγο πιο ιδιαίτερη η αντιστοίχιση.

**Ορισμός 2.6.** ***Συνάρτηση πυκνότητας πιθανότητας****. Αν η συνάρτηση κατανομής μίας τυχαίας μεταβλητής είναι συνεχώς διαφορίσιμη , τότε η συνάρτηση πυκνότητας πιθανότητας ορίζεται ως η παράγωγος της αθροιστικής συνάρτησης κατανομής:*

$$f = F' = \frac{dF(x)}{dx}.$$

Μία συνάρτηση πυκνότητας πιθανότητας έχει τις εξής ιδιότητες:

$$\begin{cases} f(x) \geq 0, \text{ σχεδόν παντού} \\ \int_{-\infty}^{\infty} f(x)dx = 1 \end{cases}$$

Αντιστρόφως αν μία συνάρτηση $f : \mathbb{R} \to \mathbb{R}$ ικανοποιεί τις δύο παραπάνω σχέσεις, τότε ορίζει ένα μέτρο πιθανότητας σύμφωνα με

$$\int_a^b f(x)dx = P(a < X \leq b)$$

**Ορισμός 2.7.** ***Στήριγμα κατανομής****. Στήριγμα ονομάζεται η περιοχή στην οποία η συνάρτηση μάζας/πυκνότητας πιθανότητας είναι θετική στην διακριτή/συνεχή περίπτωση.*

$$S_{\mathbb{P}} = \{x \in \mathbb{X} : f(x) > 0/\mathbb{P}[X = x] > 0\}$$

**Ορισμός 2.8.** ***Μέση (Αναμενόμενη) Τιμή****. Έστω μια τυχαία μεταβλητή $X$. Ορίζουμε ως μέση τιμή ή αναμενόμενη τιμή της τυχαίας μεταβλητής*

$$\mathbb{E}[X] = \begin{cases} \sum_{x_i \in D} x_i \Pr[X = x_i], \text{ στην διακριτή περίπτωση} \\ \int x_i f(x_i)dx_i, \text{ στην συνεχή περίπτωση} \end{cases}$$

**Ορισμός 2.9.** ***Διασπορά-Διακύμανση****. Έστω μία τυχαία μεταβλητή $X$ με μέση τιμή $\mu = \mathbb{E}[X]$ και συνάρτηση κατανομής $F$. Η διακύμανση ορίζεται ως:*
*$Var[X] = \int_{-\infty}^{\infty}(x-\mu)^2 dF(x) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$,*
*όταν το ολοκλήρωμα συγκλίνει.*
*Η θετική τετραγωνική ρίζα της διακύμανσης ονομάζεται* ***τυπική απόκλιση*** *και συμβολίζεται με $\sigma$.*

**Ορισμός 2.10.** ***Bernoulli(p) Τυχαία μεταβλητή.*** *Bernoulli(p) ορίζουμε την μεταβλητή που λαμβάνει την τιμή 1 με πιθανότητα $p$ και την τιμή 0 με πιθανότητα $1-p$*

**Παρατήρηση 2.1.** *Αν $X = Bernoulli(p)$ τότε $\mathbb{E}[X] = p$*

**Ορισμός 2.11.** ***Coupling*** *δύο κατανομών. Ας υποθέσουμε $X_1$ και $X_2$ δύο τυχαίες μεταβλητές ορισμένες στους χώρους πιθανότητας $(\Omega_1, \mathcal{F}_1, \Pr_1)$ και $(\Omega_2, \mathcal{F}_2, \Pr_2)$. Coupling των $X_1$ και $X_2$ αποτελεί ένας καινούργιος χώρος πιθανότητας $\Omega, \mathcal{F}, \Pr$ στον οποίο ορίζεται ένα ζεύγος μεταβλητών $(Y_1, Y_2)$ τέτοιο ώστε αν αναζητήσουμε την επιμέρους κατανομή του $Y_1$ ταυτίζεται με την κατανομή της $X_1$ και αντίστοιχα αν αναζητήσουμε την επιμέρους κατανομή της $Y_2$ ταυτίζεται με την κατανομή της $X_2$.*

**Παρατήρηση 2.2.** *Το απλούστερο Coupling δύο μεταβλητών μπορεί να κατασκευαστεί με το μέτρο γινόμενο, όταν δηλαδή είναι ανεξάρτητες.*

### 2.2.2 Απόσταση μεταξύ κατανομών

Δεδομένου ότι στόχος μας είναι να 'μάθουμε" κατανομές, δηλαδή, να βρούμε προσεγγιστικά τις βέλτιστες δυνατές εκτιμήσεις για την κατανομή που ψάχνουμε θα πρέπει να είμαστε σε θέση να μετρήσουμε την απόσταση μεταξύ δύο κατανομών.

Το πιο κλασσικό μέτρο απόστασης που έχει μελετηθεί στην Θεωρητική Πληροφορική είναι η *στατιστική απόσταση* (statistical distance) ή *απόσταση πλήρους μεταβολής* (total variation distance) [1]

**Ορισμός 2.12.** ***Total Variation Distance.*** *Αν $\mathbb{P}, \mathbb{Q}$, δύο διαφορετικά μέτρα πιθανότητας πάνω στον χώρο $\mathbb{X}$ τότε η TV ορίζεται ως:*

$$d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \max_{A \subseteq \mathbb{X}} |\mathbb{P}(A) - \mathbb{Q}(A)|$$

Το παρακάτω λήμμα προσφέρει μια πιο διαισθητική χρήση της νόρμας TV .

**Λήμμα 2.1.** $d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2} \sum_{x \in D} |\mathbb{P}(x) - \mathbb{Q}(x)|$

*Απόδειξη.*

$$d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \max_{A \subseteq \mathbb{X}} \|\mathbb{P}(A) - \mathbb{Q}(A)\|$$

Υποθέστε ότι αυτό το σύνολο έχει όνομα $A^*$ και χωρίς βλάβη της γενικότητας $\mathbb{P}(A^*) \geq \mathbb{Q}(A^*)$ Τότε:

$$d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \mathbb{P}(A^*) - \mathbb{Q}(A^*)$$

$$d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = 1 - \mathbb{P}((A^*)^c) - 1 + \mathbb{Q}((A^*)^c) = \mathbb{Q}(A^*) - \mathbb{P}(A^*)$$

Αφού η TV είναι θετική ποσότητα μπορούμε να γράψουμε χωρίς βλάβη των ισχυρισμών μας ότι

$$d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = |\mathbb{P}(A^*) - \mathbb{Q}(A^*)|$$

Από αυτό βλέπουμε ότι

$$\begin{aligned} d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) &= \frac{1}{2}\|\mathbb{P}(A^*) - \mathbb{Q}(A^*)\|_1 + \frac{1}{2}\|\mathbb{P}((A^*)^c) - \mathbb{Q}((A^*)^c)\|_1 \\ &= \frac{1}{2} \sum_{x_i \in A^*} |\mathbb{P}(x_i) - \mathbb{Q}(x_i)| + \frac{1}{2} \sum_{x_i \in (A^*)^c} |\mathbb{P}(x_i) - \mathbb{Q}(x_i)| \\ &= \frac{1}{2} \sum_{x_i \in \Omega} |\mathbb{P}(x_i) - \mathbb{Q}(x_i)| \end{aligned}$$

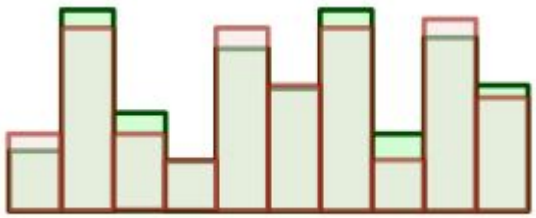
□

Δηλαδή η TV είναι ίση με το μισό της $\mathcal{L}_1$ νόρμα μεταξύ των $\mathbb{P}, \mathbb{Q}$.

**Διαισθητικά**, σε ένα πεπερασμένο χώρο απλώς συγκρίνουμε τα δύο ιστογράμματα των δυο διαφορετικών κατανομών , βρίσκουμε τις απόλυτες διαφορές και στην συνέχεια τις αθροίζουμε. Ο λόγος που διαιρούμε με 2 είναι επίσης διαισθητικά σαφής, αφού όντας και οι

[1] στο εξής θα συμβολίζεται $d_{\mathrm{TV}}$ και θα αναφέρεται συντόμως ως TV

δύο συναρτήσεις μάζας πιθανότητας ξέρουμε ότι η μάζα πιθανότητας στην οποία π.χ υπολειπόταν η $\mathbb{P}$ της $\mathbb{Q}$ θα βρίσκεται κάπου κατανεμημένη στα σημεία όπου η $\mathbb{Q}$ υπολείπεται της $\mathbb{P}$. Συνεπώς, κάθε διαφοροποίηση μεταξύ των ιστογραμμάτων των κατανομών θα διπλομετρηθεί και με την διαίρεση απαλείφουμε αυτόν τον πλεονασμό.



Σχήμα 2.1: Total Variation Distance

Επίσης είναι εύκολο κανείς να δείξει ότι

$$d_{\mathrm{TV}}(\mathbb{P},\mathbb{Q}) \leq \Pr_{X\sim\mathbb{P},Y\sim\mathbb{Q}}[X \neq Y]$$

Ενώ ένα από τα βασικά θεωρήματα γύρω από αυτή την μετρική απόσταση αποδεικνύει ότι

$$\exists \text{Coupling among } \mathbb{P},\mathbb{Q} : d_{\mathrm{TV}}(\mathbb{P},\mathbb{Q}) = \Pr[X \neq Y]$$

Εύκολα βλέπει κανείς ότι για την $d_{\mathrm{TV}}$ όπως και για τις υπόλοιπες μετρικές απόστασης μεταξύ δύο κατανομών ισχύουν οι τρεις βασικές ιδιότητες που ορίζουν μια νόρμα:

1. $d(X,X) = 0 \Leftrightarrow X \equiv Y$
2. $d(X,Y) = d(Y,X)$
3. $d(X,Y) \leq d(X,Z) + d(Z,Y)$

Όπου $X \equiv Y$ σημαίνει ότι οι $X, Y$ ακολουθούν ακριβώς την ίδια κατανομή.

#### 2.2.2.1 Σχέση της TV και του Hypothesis Testing

Σε αυτό το σημείο θα σταματήσουμε για λίγο την παρουσίαση μαθηματικών εργαλείων για να μελετήσουμε λίγο βαθύτερα την σχέση αυτής της μετρικής απόστασης και του προβλήματος της διάκρισης μεταξύ δύο κατανομών

**Ορισμός 2.13.** ***Αλγόριθμος διάκρισης****. Αλγόριθμο διάκρισης δύο κατανομών $\mathbb{P}_1, \mathbb{P}_2$ που ορίζονται στο χώρο $\mathbb{X}$ ονομάζουμε μια διαδικασία $T$ η οποία θα λαμβάνει δείγματα από δύο διαφορετικές κατανομές και θα μπορεί να διακρίνει από ποια κατανομή προήλθε το κάθε δείγμα.*

**Ορισμός 2.14.** *$\delta, k-$**αλγόριθμος διάκρισης**. $\delta, k-$αλγόριθμος διάκρισης $T$ ονομάζεται ένας αλγόριθμος που λαμβάνοντας $k$ δείγματα από μια από τις δύο κατανομές επιτυγχάνει να διακρίνει σε ποια από τις δύο ανήκει με πιθανότητα λάθους μικρότερη από $\delta$*

$$\text{Για } b = 1, 2 : \Pr_{X \sim \mathbb{P}_b^k}(T(\vec{X}_k) = b) > 1 - \delta$$

**Θεώρημα 2.1.** *Έστω δύο κατανομές $\mathbb{P}_1, \mathbb{P}_2$ ορισμένες στον $\mathbb{X}$. Τότε αν $T^*$ ο αλγόριθμος με το ελάχιστο σφάλμα διάκρισης μεταξύ των $\delta, 1$-αλγορίθμων διάκρισης, το σφάλμα του $T^*$ είναι ίσο με $\delta^* = (1 - d_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2))/2$*

*Απόδειξη.* Θεωρούμε το σύνολο $A = \{x \in \mathbb{X} | T^*(x) = 1\}$. Το σύνολο $A$ είναι στο σύνολο όπου ο αλγόριθμος αποφαίνεται $\mathbb{P}_1$. Από τον ορισμό του $\delta^*$-αλγόριθμου διάκρισης προκύπτει ότι $\mathbb{P}_1(A) \geq 1 - \delta^*, \mathbb{P}_2(A) \leq \delta^*$ Από τον αρχικό ορισμό της TV έπεται ότι

$$d_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2) \geq |\mathbb{P}_1(A) - \mathbb{P}_2(A)| \geq 1 - 2\delta^*$$

Συνεπώς το σφάλμα είναι τουλάχιστον $\delta^* \geq (1 - d_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2))/2$

Τώρα αν επιλέξουμε ως $A'$ το σύνολο που ορίζει η TV , δηλαδή $d_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2) = |\mathbb{P}_1(A') - \mathbb{P}_2(A')|$ και $\mathbb{P}_1(A') \geq \mathbb{P}_2(A')$[2] τότε ο αλγόριθμος $T'$ ο οποίος διαλέγει την $\mathbb{P}_1$ όταν και μόνο όταν λαμβάνει στοιχείο που ανήκει στην $A'$ εμφανίζει την ακόλουθη συμπεριφορά σφάλματος

$$\begin{cases} \Pr_{X \sim \mathbb{P}_2}(T(X) = 1) = \mathbb{P}_2(A') \\ \Pr_{X \sim \mathbb{P}_1}(T(X) = 2) = 1 - \mathbb{P}_1(A') \end{cases} \Rightarrow \delta' = \max(1 - \mathbb{P}_1(A'), \mathbb{P}_2(A'))$$

Όμως λόγω της βελτιστότητας

$$\delta^* \leq \delta \Rightarrow 2\delta^* \leq 1 - \mathbb{P}_1(A') + \mathbb{P}_2(A') = (1 - d_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2))$$

και άρα

$$\delta^* \leq (1 - d_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2))/2$$

□

Κλείνουμε την αναφορά μας στην μετρική TV , παρουσιάζοντας μια πολύ γνωστή ιδιότητα:

$$d_{\mathrm{TV}}(\mathbb{P}_1 \times \mathbb{P}_2, \mathbb{Q}_1 \times \mathbb{Q}_2) \leq d_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2) + d_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_2)$$

όπου $\mathbb{P}_1, \mathbb{Q}_1 \in \mathbb{X}_1, \mathbb{P}_2, \mathbb{Q}_2 \in \mathbb{X}_2$

Θα παρουσιάσουμε ακόμα δύο μέτρα απόστασης μεταξύ δύο κατανομών. Το πρώτο εμφανίστηκε κυρίως για να αντιμετωπίσει την αδυναμία του παραπάνω bound που τις περισσότερες φορές είναι αρκετά αδύναμο.

[2] Αυτό γίνεται χωρίς βλάβη της γενικότητας, αφού αν θέλουμε την αντίθετη φορά, αρκεί να ορίσουμε το $A'$ ως το συμπληρωματικό σύνολο

**Ορισμός 2.15.** ***Hellinger distance.*** *Η απόσταση Hellinger*

$$h^2(\mathbb{P},\mathbb{Q}) = \frac{1}{2}\sum_{x\in\mathbb{X}}(\sqrt{\mathbb{P}(x)} - \sqrt{\mathbb{Q}(x)})^2 = 1 - \sum_{x\in\mathbb{X}}\sqrt{(\mathbb{P}(x)\mathbb{Q}(x))}$$

Οι τρεις βασικές ιδιότητες της Hellinger είναι οι ακόλουθες:

1. $1 - h^2(\mathbb{P}_1 \times \mathbb{P}_2, \mathbb{Q}_1 \times \mathbb{Q}_2) = (1 - h^2(\mathbb{P}_1, \mathbb{Q}_1)(1 - h^2(\mathbb{P}_2, \mathbb{Q}_2)$ . Η ιδιότητα αυτή μπορεί να επεκταθεί για πολυδιάστατες κατανομές οποιασδήποτε διάστασης.
2. $h^2(\mathbb{P},\mathbb{Q}) \leq d_{\mathrm{TV}}(\mathbb{P},\mathbb{Q}) \leq h(\mathbb{P},\mathbb{Q})\sqrt{2 - h^2(\mathbb{P},\mathbb{Q})}$
3. $1 - \sqrt{1 - d_{\mathrm{TV}}(\mathbb{P},\mathbb{Q})} \leq h^2(\mathbb{P},\mathbb{Q}) \leq d_{\mathrm{TV}}(\mathbb{P},\mathbb{Q})$

Παρατηρείστε ότι από τις (2,3) προκύπτει μια σχετικά έντονη ισοδυναμία στην ποσότητα μεταξύ των δύο δομών.

**Ορισμός 2.16.** ***Kolmogorov distance.*** *Kolmogorov απόσταση δύο κατανομών ονομάζουμε την supremum νόρμα μεταξύ των συναρτήσεων κατανομής των μέτρων πιθανότητας, δηλαδή:*

$$d_K(X \sim \mathbb{P}, Y \sim \mathbb{Q}) = \max|F_X(x) - F_Y(x)|$$

$$d_K(\mathbb{P},\mathbb{Q}) = \max_{A:(-\infty,x]}|\mathbb{P}(A) - \mathbb{Q}(A)|$$

Από τον ορισμό της Kolmogorov προκύπτει ότι $d_K \leq d_{\mathrm{TV}}$, αφού η TV είναι ορισμένη σε μεγαλύτερη ομάδα συνόλων.

Θα κλείσουμε την μελέτη μας στις μετρικές βλέποντας ένα από τα πιο βασικά Lower Bound, στην Θεωρία διάκρισης των κατανομών.

**Διαισθητικά**, η απόσταση Hellinger αντιστοιχεί σε μια μορφή $L_2$ νόρμας. Η χρήση της Kolmogorov στην βιβλιογραφία εμφανίζεται είτε όταν το πρόβλημα απαιτεί την γνώση των συναρτήσεων κατανομών και όχι των συναρτήσεων πυκνότητας, αλλά κυρίως για να μπορεί κανείς να συνδέσει διακριτές και συνεχείς κατανομές. Αξίζει να παρατηρηθεί ότι είναι εύκολο να κατασκευάσει κανείς μια συνάρτηση κατανομής που να βρίσκεται π.χ μεταξύ του $\mathbb{P} \in (0.25, 0.75)$ και μια συνάρτηση δυική που παίρνει τιμές $\mathbb{Q} \in \{0, 1\}$ και οι οποίες κατά συνέπεια θα έχουν απόσταση $d(\mathbb{P},\mathbb{Q}) = 1$ μεταξύ τους.

**Θεώρημα 2.2 ((Κάτω φράγμα στο γενικό αλγόριθμο διάκρισης)).**
*$\forall\delta : 0 \leq \delta \leq 1/4$ και για κάθε αλγόριθμο $\delta, k$-αλγόριθμο διάκρισης δύο κατανομών $\mathbb{P}_1, \mathbb{P}_2$ των οποίων η $h^2(\mathbb{P}_1, \mathbb{P}_2) \leq 1/2$, το πλήθος των δειγμάτων που χρειάζεται να έχει ώστε το λάθος απόκρισης να είναι το πολύ $\delta$ είναι τουλάχιστον:*

$$k > \ln(\frac{1}{4\delta})\frac{1}{h^2(\mathbb{P}_1, \mathbb{P}_2)}$$

*Απόδειξη.* Αρχικώς κάθε αλγόριθμο $\delta, k$-διάκρισης των κατανομών $\mathbb{P}_1, \mathbb{P}_2$ μπορεί να θεωρηθεί σαν αλγόριθμος διάκρισης $\delta$, 1-διάκρισης των κατανομών $\mathbb{P}_1^k, \mathbb{P}_2^k$.Συνεπώς, από το προηγούμενο θεώρημα, από το πρώτο τμήμα της ανάλυσης μπορούμε να έχουμε ότι $d_{\mathrm{TV}}(\mathbb{P}_1^k, \mathbb{P}_2^k) > 1 - 2\delta$.

Αν προσπαθούσαμε όμως να συνδέσουμε την $d_{\mathrm{TV}}(\mathbb{P}_1^k, \mathbb{P}_2^k)$ με την $d_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2)$, λόγω της αδύναμης ανισότητας :

$$d_{\mathrm{TV}}(\mathbb{P}_1 \times \mathbb{P}_2, \mathbb{Q}_1 \times \mathbb{Q}_2) \leq d_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2) + d_{\mathrm{TV}}(\mathbb{Q}_1, \mathbb{Q}_2)$$

θα είχαμε γραμμική εξάρτηση από το $k$ επιπλέον.

Συνεπώς, θα χρησιμοποιήσουμε το κάτω φράγμα της Hellinger και θα έχουμε:

$$\begin{aligned} h^2(\mathbb{P}_1^k, \mathbb{P}_2^k) &\geq 1 - \sqrt{(1 - d_{\mathrm{TV}}(\mathbb{P}_1^k, \mathbb{P}_2^k)^2)} \\ &\geq 1 - \sqrt{1 - (1 - 2\delta)^2} \\ &> 1 - \sqrt{4\delta} \Rightarrow \end{aligned}$$

$$h^2(\mathbb{P}_1^k, \mathbb{P}_2^k) = 1 - (1 - h^2(\mathbb{P}_1^k, \mathbb{P}_2^k)) = 1 - (1 - h^2(\mathbb{P}_1, \mathbb{P}_2))^k, \text{από ιδιότητα 1}$$

Άρα

$$1 - (1 - h^2(\mathbb{P}_1, \mathbb{P}_2))^k > 1 - \sqrt{4\delta} \Rightarrow \boxed{k > \frac{1}{4h^2(\mathbb{P}_1, \mathbb{P}_2)} \ln \frac{1}{4\delta}}$$

□

**Πόρισμα 2.1.** *Για να διαχωρίσει κανείς δύο Bernoulli χρειάζεται με πιθανότητα επιτυχίας* $9/10$ *τουλάχιστον* $\Omega(1/\epsilon^2)$.

*Απόδειξη.*

- Έστω $X \sim Bernoulli(1/2), Y \sim Bernoulli(1/2 + epsilon)$.
- Η $h^2(X, Y) = 1 - \sqrt{1/2}\sqrt{1/2 + epsilon} - \sqrt{1/2}\sqrt{1/2 - epsilon} < \epsilon$
- Εύκολα βλέπει κανείς ότι $h^2(X, Y) < \epsilon^2$
- Άρα για να έχουμε πιθανότητα τουλάχιστον πχ. $7/8$ θα πρέπει να έχουμε τουλάχιστον $\Omega(1/\epsilon^2)$ δείγματα

□

### 2.2.3 Μέτρα συγκέντρωσης πιθανότητας

Στην συνέχεια όλης αυτής της μελέτης είναι απαραίτητο να έχουμε αναφέρει τα βασικά εργαλεία συγκέντρωσης πιθανότητας. Το βασικό στοιχείο αυτών των εργαλείων είναι η μελέτη της πιθανότητας "η τυχαία μεταβλητή που μελετούμε να απέχει αισθητά από την αναμενόμενη τιμή ".

**Θεώρημα 2.3** (**Ανισότητα Markov**).
*Αν η $X$ είναι μια μη αρνητική τυχαία μεταβλητή και $a > 0$, τότε ισχύει*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

**Θεώρημα 2.4** (**Ανισότητα Chebyshev**).
*Αν η $X$ είναι μια τυχαία μεταβλητή με φραγμένη μέση τιμή $\mu = \mathbb{E}[X]$ και μη μηδενική φραγμένη $Var[X]$, τότε*

$$\forall k > 0,\ \Pr(|X - \mu| \geq k\sqrt{Var[X]}) \leq \frac{1}{k^2}.$$

**Θεώρημα 2.5** (**Ανισότητα Hoeffding**).
*Αν $\overline{X} = \sum_{i=1}^{n} X_i/n$, $X_i \in [a_i, b_i]$ και $X_i$ ανεξάρτητα, τότε $\forall t > 0$:*

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

*ή διαφορετικά αν $S_n = X_1 + \cdots + X_n$*

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

Note that the inequalities also hold when the Xi have been obtained using sampling without replacement; in this case the random variables are not independent anymore. A proof of this statement can be found in Hoeffding's paper. For slightly better bounds in the case of sampling without replacement, see for instance the paper by Serfling (1974).

**Θεώρημα 2.6** (**Ανισότητες Chernoff**).

***Πολλαπλασιαστική μορφή ( Multiplicative Chernoff Bound)***
*Έστω $X_1, \cdots, X_n$ τυχαίες ανεξάρτητες μεταβλητές στο [0,1]. Έστω ότι $X = \sum_{i=1}^{N} X_i$ και $\mu = \mathbb{E}[X]$, τότε $\forall \delta > 0$:*

$$\Pr(X > (1+\delta)\mu) < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu} = e^{-D((1+\delta)p\|p)n}.$$

*&*

$$\Pr(X < (1-\delta)\mu) < \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} = e^{-D((1-\delta)p\|p)n}.$$

*Πιο χρήσιμες μορφές ( αλλά λήγο πιο ασθενείς μορφές ) αποτελούν:*

$$\Pr(X \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2\mu}{3}}, \qquad 0 < \delta < 1,$$

$$\Pr(X \geq (1+\delta)\mu) \leq e^{-\frac{\delta\mu}{3}}, \qquad 1 < \delta,$$

$$\Pr(X \leq (1-\delta)\mu) \leq e^{-\frac{\delta^2\mu}{2}}, \qquad 0 < \delta < 1.$$

Η παραπάνω σειρά ανισοτήτων μπορεί να γενικευτεί και για τιμές $\mu^-, \mu^+$, που φράσσουν την μέση τιμή $\mu$

**Ορισμός 2.17.** ***Εμπειρική Κατανομή***. *Έστω ότι τα στοιχειώδη ενδεχόμενα $\omega_i$ του χώρου $\Omega$. Η πιθανότητα να συμβεί το κάθε ενδεχόμενο με βάση την κατανομή $\mathbb{P}$ είναι $\Pr[X = \omega_i] = \mathbb{P}(\omega_i)$. Η εμπειρική κατανομή $\hat{\mathbb{P}}_m$ του ενδεχομένου αφού έχουμε συλλέξει $m$ δείγματα είναι:*

$$\hat{\mathbb{P}}_m(\omega_i) = \frac{\sum_{i=1}^{m} \mathbb{1}\{X = \omega_i\}}{m}$$

Η ακόλουθη ανισότητα αποτελεί μια από τις κρισιμότερες ανισότητες στο Learning. Εμείς θα μελετήσουμε μια πολύ ειδική μορφή της και θα δούμε και την απόδειξη της σε αυτή την μορφή.

**Θεώρημα 2.7 (VC Inequality).**

$$\forall \mathbb{P} with\ domain\ [n] = \{1, 2, 3, \cdots, n\} : \mathbb{E}[d_{\mathrm{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m)] \leq O(\frac{\sqrt{n}}{\sqrt{m}})$$

*Απόδειξη.* Ας παρατηρήσουμε για αρχή την ποσότητα: $Y_m = |\mathbb{P}(\omega_i) - \hat{\mathbb{P}}_m(\omega_i)|$. Η $Y_m$ είναι μια τυχαία μεταβλητή που εξαρτάται από τα δείγματα της κατανομής και θα θέλαμε να υπολογίσουμε την μέση τιμή της:

$$\begin{aligned} \mathbb{E}[Y_m] &= \mathbb{E}[\sqrt{Y_m^2}] \\ &\leq \sqrt{\mathbb{E}[Y_m^2]} \end{aligned}$$

Όμως είναι εμφανές ότι :$\mathbb{E}[\hat{\mathbb{P}}_m(\omega_i)] = \mathbb{P}(\omega_i)$. Αλλά $\mathbb{E}[Y_m^2] = Var[\hat{\mathbb{P}}_m(\omega_i)] \Rightarrow$

$$\begin{aligned}
\mathbb{E}[Y_m] &\leq \sqrt{Var[\hat{\mathbb{P}}_m(\omega_i)]} \\
&\leq \frac{1}{m}\sqrt{\sum_{i=1}^{m} Var[\mathbb{1}\{X = \omega_i\}]} \\
&\leq \frac{1}{m}\sqrt{m \times Var[\mathbb{1}\{X = \omega_i\}]} \\
&\leq \frac{1}{\sqrt{m}}\sqrt{Var[\mathbb{1}\{X = \omega_i\}]} \\
&\leq \frac{1}{\sqrt{m}}\sqrt{\mathbb{P}(\omega_i) - \mathbb{P}(\omega_i)^2}
\end{aligned}$$

Χρησιμοποιούμε την γραμμικότητα της μέσης τιμής και έχουμε ότι

$$\mathbb{E}[d_{\mathrm{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m)] \leq \frac{1}{2}\sum_{i=1}^{n} \frac{1}{\sqrt{m}}\sqrt{\mathbb{P}(\omega_i) - \mathbb{P}(\omega_i)^2}$$

Εδώ θα χρησιμοποιήσουμε την Cauchy-Schwarz

$$\begin{aligned}
\frac{1}{\sqrt{m}}\sum_{i=1}^{n}\sqrt{\mathbb{P}(\omega_i) - \mathbb{P}(\omega_i)^2} &= \frac{1}{\sqrt{m}}\sum_{i=1}^{n}\sqrt{\mathbb{P}(\omega_i)(1 - \mathbb{P}(\omega_i))} \\
&\leq \frac{1}{\sqrt{m}}\sqrt{\sum_{i=1}^{n}\mathbb{P}(\omega_i)\sum_{i=1}^{n}(1 - \mathbb{P}(\omega_i))} \\
&\leq \frac{1}{\sqrt{m}}\sqrt{1 \times [\sum_{i=1}^{n}(1) - \sum_{i=1}^{n}\mathbb{P}(\omega_i))]} \\
&\leq \frac{1}{\sqrt{m}}\sqrt{n-1}
\end{aligned}$$

Άρα

$$\boxed{\mathbb{E}[d_{\mathrm{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m)] \leq \frac{\sqrt{n-1}}{2\sqrt{m}} \leq \frac{\sqrt{n}}{\sqrt{m}}}$$

□

Η τελευταία ανισότητα αποτελεί ένα από τα πιο σημαντικά εργαλεία σε συναρτήσεις που προκαλούν aggregation της πληροφορίας των περισσοτέρων τυχαίων μεταβλητών.

**Ορισμός 2.18.** ***Bounded Difference Assumption.*** *Έστω συνάρτηση $X_1, X_2, \cdots, X_n \in \mathbb{X}$ και επίσης $f : X^n \to \mathbb{R}$ και ισχύει ότι για $\forall i \in [n]$ :*

$$\sup_{x_1, x_2, \cdots, x_n, x_i'} |f(x_1, x_2, \cdots, x_i, \cdots, x_n) - f(x_1, x_2, \cdots, x_i, \cdots, x_n)|| \leq c_i$$

*Δηλαδή αν αλλάξουμε κάθε μεταβλητή σε κάποια άλλη τιμή διατηρώντας τις υπόλοιπες σταθερές, τότε για κάθε μεταβλητή $x_i$ υπάρχει ένα άνω όριο $c_i$.*

**Θεώρημα 2.8** (**Bounded Difference Inequality**).
*Έστω συνάρτηση $X_1, X_2, \cdots, X_n \in \mathbb{X}$ και επίσης $f : X^n \to \mathbb{R}$ και ισχύει το Bounded Difference Assumption. Τότε*

$$\Pr[|f(X_1, X_2, \cdots, X_n) - \mathbb{E}[f(X_1, X_2, \cdots, X_n)]| \geq \epsilon] \leq 2\exp(-2\epsilon^2 / \sum_{i=1}^{n} c_i^2)$$

Θα κλείσουμε αυτή την μελέτη σε μια σημαντική εφαρμογή του παραπάνω θεωρήματος Έστω $Y_i = \frac{\sum_{i=1}^{m} \mathbb{1}\{\$\text{ample}_i = \omega_i\}}{m}$ και $f(Sample_1, Sample_2, \cdots, Sample_m) = d_{\text{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m)$.

**Πόρισμα 2.2.**

$$\Pr[|d_{\text{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m) - \mathbb{E}[d_{\text{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m)]| \geq \epsilon] \leq 2\exp(-2m\epsilon^2)$$

*Απόδειξη.* Αρκεί να δει κανείς ότι $f(Sample_1, Sample_2, \cdots, Sample_m) = d_{\text{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m)$, αν αλλάξουμε ένα δείγμα, η μεγαλύτερη μεταβολή $(1/m)$ συνεπώς $c_i = \frac{1}{m}$. Άρα $\sum_{i=1}^{m} c_i^2 = \sum_{i=1}^{m} \frac{1}{m^2} = \frac{1}{m}$ και άρα:

$$\Pr[|d_{\text{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m) - \mathbb{E}[d_{\text{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m)]| \geq \epsilon] \leq 2\exp(-2m\epsilon^2)$$

□

### 2.2.4 Κάτω φράγματα σε αλγορίθμους Μάθησης και Διάκρισης Lower Bounds on Learning & Testing Algorithms

#### 2.2.4.1 Βασική Μεθοδολογία

Το πρώτο μας βήμα είναι να κατασκευάσουμε αρχικά ένα γενικό framework για τον υπολογισμό κάτω φραγμάτων. Όταν μελετάμε τα κλασσικά προβλήματα εκτίμησης, χρησιμοποιούμε σχεδόν πάντα την τυπική εκδοχή του minimax risk. Στην ενότητα αυτή θα δούμε πως μπορούμε να χρησιμοποιήσουμε τις τεχνικές για να εξάγουμε πληροφοριο-θεωρητικά κάτω φράγματα στην δειγματική πολυπλοκότητα των αλγορίθμων μάθησης.

Ας ξεκινήσουμε ορίζοντας την τυπική εκδοχή του minimax risk. Στο εξής ας υποθέσουμε ότι $\mathcal{P}$ είναι μία οικογένεια διαφορετικών κατανομών πάνω στο $\mathbb{X}$. Ας υποθέσουμε ότι υπάρχει κάποια παράμετρος $\theta$ που θέλουμε να εκτιμήσουμε επειδή χαρακτηρίζει ή χαρακτηρίζεται από την κατανομή που επιδιώκουμε να μάθουμε. Ας υποθέσουμε ότι η $\theta$ λαμβάνει τιμές στο $\Theta$.

Για παράδειγμα, αν $\theta = \mathbb{E}[X], X(\omega) \in \mathbb{R}$, τότε $\Theta \subset \mathbb{R}$.

Τέλος ας υποθέσουμε ότι υπάρχει μια συνάρτηση $\hat{\theta}_n : \mathcal{X}^n \to \Theta$. Η τελευταία συνάρτηση πορίζει την μέθοδο εκτίμησης που προτείνει ο αλγόριθμος μας βασισμένος στα δείγματα αυτής της κατανομής.

Για παράδειγμα, αν ορίσουμε ότι $\mathcal{P} = \{\mathcal{N}(\theta, \sigma^2) : \theta \in \mathbb{R}\}$, όπου $\sigma^2$ είναι κάποια γνωστή θετική πραγματική τιμή τότε η $\theta(P) = \mathbb{E}_P[X]$ και π.χ:

$$\hat{\theta}_n = \frac{\sum_n X_i}{n}, X_i \sim P, \text{ όπου η } P \in \mathcal{P}$$

Μπορεί όμως και $\theta(P) = \int_0^1 (p\prime(t))^2 \mathrm{d}t$, όπου $p$ είναι πυκνότητα της κατανομής $P$. Σε αυτή την περίπτωση η $\theta$ δεν παραμετροποιεί ή ορίζει την $P$, αλλά η γνώση της μας δίνει μια πιο ευρεία γνώση για την κατανομή.

Από τα παραπάνω γίνεται εμφανές ότι ο χώρος $\Theta$ και η συνάρτηση $\theta$, προσδιορίζουν και το στατιστικό πρόβλημα που θέλουμε να επιλύσουμε. Για να υπολογίσουμε την ποιότητα ενός εκτιμητή $\hat{\theta}$, ορίζουμε μια συνάρτηση $\rho : \Theta \times \Theta \to \mathbb{R}_+$. Αυτή η (ημί)μετρική συνάρτηση στο χώρο $\Theta$, θα μας προσδιορίζει το σφάλμα του προβλέπτη/εκτιμητή. Επίσης ορίζουμε $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$, μια μη-φθίνουσα θετική συνάρτηση ώστε $\Phi(0) = 0, (\Phi(t) = t^2)$. Αυτή η συνάρτηση ορίζει το κόστος που θεωρούμε

Για μια κατανομή $P \in \mathcal{P}$, υποθέτουμε ότι λαμβάνουμε ισόνομα,ανεξάρτητα δείγματα $X_i$ που ακολουθούν την $P$. Στόχος μας είναι να προσεγγίσουμε/εκτιμήσουμε $\theta(P)$ μέσω μιας μεθόδου $\hat{\theta}(P)$.

Ορίζουμε ως minimax risk:

$$\inf_{\hat{\theta}} \sup_{X_1, X_2, \cdots, X_n} \mathbb{E}[\Phi(\rho(\hat{\theta}(X_1, X_2, \cdots, X_n), \theta(P)))]$$

Ας δούμε τι εκφράζεις αυτή η ποσότητα. Αν υποθέσουμε ότι έχουμε κάποιον προβλέπτη, εκτιμητή τότε υπάρχει σίγουρα ένα input από κάποια τυχαία δείγματα του dataset που οδηγούν τον αλγόριθμο στην χειρότερη απόδοση του, από άποψη λάθους. Έστω ότι έχουμε στα χέρια μας τον βέλτιστο αλγόριθμο ως προς αυτό το κριτήριο ταξινόμησης. Ποια είναι η μέση τιμή σφάλματος αυτού του αλγορίθμου; Ακριβώς αυτή η ποσότητα εκφράζεται από αυτό το minimax risk.

Υπάρχουν πολλοί διαφορετικοί τρόποι να αποδείξει κανείς κάτω φράγματα στο πλήθος των δειγμάτων ώστε ένας αλγόριθμος να είναι PAC αποδοτικός. Θα παρουσιάσουμε δύο διαφορετικά λήμματα, το καθένα από αυτά στην πραγματικότητα μετατρέπει το πρόβλημα του maximum risk σε ένα πρόβλημα Bayesian και αντί να φράσσουν από κάτω την ποσότητα του risk, φράσσουν την απόδοση του εκτιμητή σε αυτό το Bayesian πρόβλημα. Πιο συγκεκριμένα, έστω $\Pi = \{P_i\} \subset \mathcal{P}$, μια οικογένεια από συναρτήσεις όπου επιλέγουμε τυχαία από αυτές με κάποια κατανομή $\pi$. Σε αυτή την περίπτωση προκύπτουν τα εξής:

$$\sup_{P \in \mathcal{P}} \mathbb{E}[\Phi(\rho(\hat{\theta}(X_1, X_2, \cdots, X_n), \theta(P)))] \geq \sup_{P \in \Pi} \mathbb{E}[\Phi(\rho(\hat{\theta}(X_1, X_2, \cdots, X_n), \theta(P)))]$$

Επίσης επειδή επιλέγουμε τυχαία κάποια από αυτές με κατανομή $\pi$ έχουμε ότι :

$$\sup_{P \in \Pi} \mathbb{E}[\Phi(\rho(\hat{\theta}(X_1, X_2, \cdots, X_n), \theta(P)))] \geq \sum_i \pi(i) \sup_{P \in P_i} \mathbb{E}[\Phi(\rho(\hat{\theta}(X_1, X_2, \cdots, X_n), \theta(P)))]$$

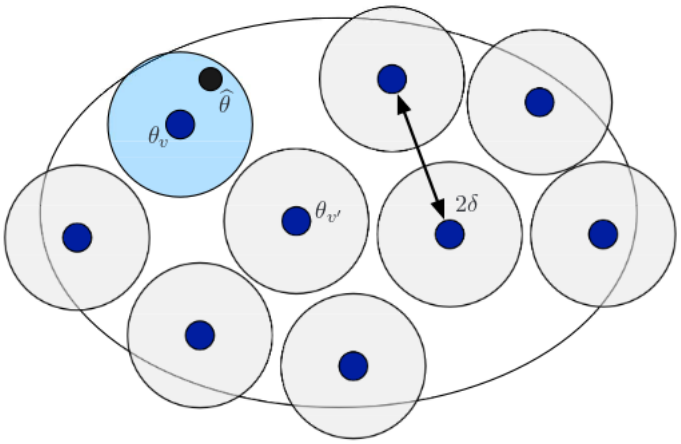
Όπως προαναφέραμε, το πρώτο βήμα στο καθένα από τα δύο λήμματα είναι η αναγωγή του προβλήματος εκτίμησης σε ένα πρόβλημα διάκρισης. Η κεντρική ιδέα είναι να αποδείξει κανείς ότι το risk φράσσεται σχεδόν πάντα από τα κάτω από ένα πρόβλημα διάκρισης των κατανομών της 'περίεργης' ομάδας $\Pi$ που ορίσαμε εμείς.

Ας υποθέσουμε ότι έχουμε $P_i, P_j \in \Pi$.

**Ορισμός 2.19.** $2\delta-$***πακετάρισμα.*** *Η* $\Pi$ *ονομάζεται* $2\delta-$*πακετάρισμα κάτω από* $\rho$*-ημι μετρική αν*

$$\rho(\theta(P_i), \theta(P_j)) \geq 2\delta, \forall i \neq j$$

Παρατηρείστε ότι το $|\Pi|$ σίγουρα κάποια στιγμή λαμβάνει κάποια μέγιστη τιμή, αφού κάποια στιγμή ο χώρος θα 'μπουκώσει'.




Παρατηρήστε ότι

$$\mathbb{E}[\Phi(\rho(\theta, \hat{\theta}))] \geq \mathbb{E}[\Phi(\delta)\mathbb{1}\{\rho(\theta, \hat{\theta}) \geq \delta\}] = \Phi(\delta)\Pr[\rho(\theta, \hat{\theta}) \geq \delta],$$

αφού η $\Phi$ είναι συνάρτηση ποινής και άρα μη-φθίνουσα.

- Ας πάρουμε τώρα κάποιον αλγόριθμο εκτίμησης $\hat{\theta}$ και ας υποθέσουμε ότι η $\rho(\cdot, \cdot)$ είναι μετρική.
- Ας πάρουμε την πλησιέστερη συνάρτηση στον estimator από την ομάδα $\Pi$, $I = \arg\min_{P_i \in \Pi} \rho(\hat{\theta}, \theta(P_i))$.

Αξίζει να δει κανείς ότι αν βρούμε κάποια $P_i : \rho(\hat{\theta}, \theta(P_i)) < \delta$, τότε $I = i$, εξαιτίας της τριγωνικής ανισότητας. Πράγματι οποιαδήποτε άλλη $P_j : \rho(\hat{\theta}, \theta(P_j)) \geq \rho(\theta(P_i), \theta(P_j)) - \rho(\hat{\theta}, \theta(P_i)) \geq 2\delta - \delta = \delta$.

Συνεπώς:

$$\boxed{\text{Αν } P_i : \rho(\hat{\theta}, \theta(P_i)) < \delta, \text{ τότε } I = i. \Leftrightarrow \text{ Αν } I \neq i \text{ τότε } P_i : \rho(\hat{\theta}, \theta(P_i)) \geq \delta}$$

Αν πάρουμε $\pi = \frac{1}{|\Pi|}$ τότε μπορούμε να εξάγουμε:

$$\begin{aligned}\inf_{\hat{\theta}} \sup_{P\in\Pi} \mathbb{E}[\Phi(\rho(\theta(P),\hat{\theta}))] &\geq \Phi(\delta)\Pr[\rho(\theta(P),\hat{\theta}) \\ &\geq \delta] \geq \frac{1}{|\Pi|}\Phi(\delta)\Pr[\rho(\theta(P_i),\hat{\theta}) \geq \delta] \\ &\geq \frac{1}{|\Pi|}\Phi(\delta)\Pr[I \neq i]\end{aligned}$$

**Θεώρημα 2.9.**

$$\inf_{\hat{\theta}} \sup_{P\in\Pi} \mathbb{E}[\Phi(\rho(\theta(P),\hat{\theta}))] \geq \Phi(\delta)\Pr[I(X_1, X_2, \cdots, X_n) \neq V], V \in Uniform([|\Pi|])$$

Συνεπώς καταλήγουμε ότι το πρόβλημα του minimax risk μπορεί να αναχθεί στο πρόβλημα διάκρισης που ορίσαμε.

Πριν παραθέσουμε τα δύο πολύ σημαντικά λήμματα (Le Cam,Assaoud), θα παραθέσουμε μια βασική αρχή των αλγορίθμων που λειτουργούν με τυχαιότητα, την αρχή του Yao

#### 2.2.4.2 Αρχή του Yao

Στην θεωρία Υπολογιστικής Πολυπλοκότητας, η αρχή του Yao μας επιτρέπει να μελετήσουμε το μέσο κόστος ενός τυχαιοκρατικού αλγορίθμου. Συγκεκριμένα ο Yao πρότεινε και απέδειξε ότι στην περίπτωση της χειρότερης εισόδου το κόστος ενός τυχαιοκρατικού αλγορίθμου δεν είναι καλύτερο από το μέσο κόστος οποιαδήποτε ντετερμινιστικού αλγορίθμου κάτω από την χειρότερη δυνατή κατανομή ως προς διαφορετικές εισόδους, ακόμη και του βέλτιστου αλγορίθμου πάνω στο πρόβλημα. Με αυτό τον τρόπο, για να κατασκευάσει κανείς ένα κάτω φράγμα στην απόδοση ενός τυχαιοκρατικού αλγορίθμου, αρκεί να βρει μια κατάλληλη κατανομή πάνω σε δύσκολες εισόδους και να αποδείξει ότι οποιοσδήποτε ντετερμινιστικός αλγόριθμος δεν είναι ικανός να αποδώσει αρκετά καλά εναντίον αυτής της κατανομής των εισόδων.

Για όσους γνωρίζουν την βασική θεωρία παιγνίων και την μικτή ισορροπία του Nash. Η αρχή του Yao, λοιπόν, μπορεί να περιγραφεί και σε παιγνιοθεωριτικούς όρους δύο παικτών ενός παιχνιδιού μηδενικού αθροίσματος. Ας υποθέσουμε ότι η Αλίκη, 1ος παίκτης, επιλέγει κάποιον ντετερμινιστικό αλγόριθμο και ο Βασίλης, 2ος παίκτης, επιλέγει μια είσοδο. Το αντάλλαγμα στο παιχνίδι είναι το κόστος-η απόδοση του αντίστοιχου αλγορίθμου στην επιλεγμένη είσοδο.

Πριν συνεχίσουμε αξίζει να δούμε την βασική μοντελοποίηση που ισχύει στην θεωρία της Υπολογιστικής Πολυπλοκότητας. Κάθε τυχαιοκρατικός αλγόριθμος $R$ μπορεί να θεωρηθεί ως μια τυχαία επιλογή πάνω στους διαφορετικούς ντετερμινιστικούς αλγόριθμους. Έτσι μπορούμε να θεωρήσουμε ότι κάθε κατανομή πάνω σε αυτούς τους αλγόριθμους είναι μια μικτή στρατηγική της Αλίκης. Αντίστοιχα στρατηγική του Βασίλη μπορεί να θεωρηθεί κάποια κατανομή πάνω στις εισόδους.

**Θεώρημα 2.10 (Αρχή του Yao).**
*Η χειρότερη περίπτωση της μέσης απόδοσης ενός τυχαιοκρατικού αλγόριθμου $\mathcal{A}$ είναι καλύτερη από τον καλύτερο ντετερμινιστικό αλγόριθμο όταν του δίνεται μια τυχαία είσοδος.*

*Απόδειξη.* Έστω ότι ο $\mathcal{A}$ είναι μια κατανομή $\mu$ πάνω σε όλους τους $a \in Algos$, όπου $Algos$, όλοι οι ντετερμινιστικοί αλγόριθμοι που επιλύουν αυτό το πρόβλημα. Έστω επίσης μια τυχαία είσοδος, δηλαδή μια κατανομή $\lambda$ πάνω σε όλες τις δυνατές εισόδους $\mathcal{X}$. Έστω $C = \max_{x \in \mathcal{X}} \mathbb{E}[c(\mathcal{A}, x)]$, η χειρότερη περίπτωση της μέσης απόδοσης του τυχαιοκρατικού μας αλγορίθμου. Από τον ορισμό του $C$, προκύπτει ότι για οποιαδήποτε είσοδο έχουμε ισχύει:

$$\forall x \in \mathcal{X} : \; C \geq \mathbb{E}[c(\mathcal{A}, x)] \Rightarrow C \geq \sum_{a \in Algos} c(a, x)\mu(a)$$

Έστω ότι για κάθε διαφορετικό $x \in \mathcal{X}$ πολλαπλασιάζουμε την προηγούμενη σχέση με $\lambda(x)$. Συνεπώς:

$$\forall x \in \mathcal{X} : \; \lambda(x)C \geq \lambda(x) \sum_{a \in Algos} c(a, x)\mu(a)$$

Θα αθροίσουμε όλες τις παραπάνω σχέσεις:

$$\sum_{x \in \mathcal{X}} \lambda(x)C = C \sum_{x \in \mathcal{X}} \lambda(x) = C \geq \sum_{x \in \mathcal{X}} \lambda(x) \sum_{a \in Algos} c(a, x)\mu(a)$$

Τέλος έχουμε:

$$C \geq \sum_{x \in \mathcal{X}} \sum_{a \in Algos} \mu(a)\lambda(x)c(a, x)$$

Απο περιστεροφωλιά, υπάρχει ένας αλγόριθμος $a$ για τον οποίον σίγουρα $C \geq \sum_{x \in \mathcal{X}} \lambda(x)c(a, x)$, γιατί διαφορετικά η παραπάνω ισότητα θα όφειλε να έχει διαφορετική φορά. Συνεπώς $C \geq \sum_{x \in \mathcal{X}} \lambda(x)c(a, x) = \mathbb{E}[c(a, X)]$, όπου $X$ η τυχαία μεταβλητή της εισόδου με κατανομή $\lambda$. Επίσης σίγουρα $\mathbb{E}[c(a, X)] \geq \min_{a \in Algos} \mathbb{E}[c(a, X)]$. Συνδέοντας όλες τις σχέσεις:

$$\max_{x \in \mathcal{X}} \mathbb{E}[c(\mathcal{A}, x)] \geq \min_{a \in Algos} \mathbb{E}[c(a, X)]$$

□

**Θεώρημα 2.11 (Αρχή του Yao στο Property Testing).**
*Έστω $P$ όλες οι συναρτήσεις που έχουν την ιδιότητα $p$ που εξετάζουμε. Έστω ότι υπάρχει μια κατανομή $\lambda$ πάνω σε συναρτήσεις $\mathcal{X} = P \cup \{f : dist(f, P) > \epsilon\}$ τέτοιο ώστε κάθε ντετερμινιστικός αλγόριθμος $a$ που χρησιμοποιεί $m$ δείγματα είναι σωστός με πιθανότητα αυστηρά μικρότερη του $2/3$. Τότε για οποιονδήποτε (τυχαίο ή μη) αλγόριθμο $\mathcal{A}$ που εξακριβώνει την ιδιότητα $p$ στο σύνολο $\mathcal{X}$ και αντλεί $m$ δείγματα από αυτό, υπάρχει μια συνάρτηση εισόδου $f_{\mathcal{A}}$ τέτοια ώστε $\Pr[\mathcal{A}$ είναι λάθος$] > \frac{1}{3}$. Συνεπώς χρειάζεται $\omega(m)$ δείγματα ώστε ο $\mathcal{A}$ να είναι $\delta < 1/3$ αλγόριθμος διάκρισης.*

*Απόδειξη.* Θα εφαρμόσουμε την γενική αρχή του Yao.

- Θα ορίσουμε ως κόστος απλώς $c(a, x) = \begin{cases} 0 & \text{, Ο αλγόριθμος } a \text{ απαντάει ορθά στην είσοδο } x \\ 1 & \text{, αλλιώς} \end{cases}$

- Επίσης *Algos* ορίζουμε το σύνολο όλων των ντετερμινιστικών αλγόριθμων διάκρισης $m$ δειγμάτων

Από την υπόθεση έστω ότι έχουμε βρει μια κατανομή $\lambda$ στις συναρτήσεις εισόδου ώστε:

$$\forall a \in Algos : \mathop{\mathbb{E}}_{f\in X}[c(a,f)] = \Pr[a \text{ είναι λάθος για είσοδο την } f] > \frac{1}{3}$$

Συνεπώς: $\min_{a\in Algos} \mathbb{E}[c(a,X)] > 1/3$ και άρα από την αρχή του Yao: $\max_{f\in\mathcal{X}} \mathbb{E}[c(\mathcal{A},f)] > 1/3$. Όμως επειδή η $c(\cdot,\cdot)$ είναι δείκτρια συνάρτηση αυτό ισοδυναμεί με

$$\max_{f\in\mathcal{X}} \Pr[\mathcal{A} \text{ είναι λάθος στην} f] > \frac{1}{3}$$

Άρα υπάρχει κάποια είσοδος $f_{\mathcal{A}}$ που προφανώς προκαλεί:

$$\Pr[\mathcal{A} \text{ είναι λάθος στην } f_{\mathcal{A}}] > \frac{1}{3}$$

□

#### 2.2.4.3 Le Cam Lemma

Ας υποθέσουμε ότι υπάρχουν μόνο δύο κατανομές στην Π, η $P_1$ & $P_2$. Όπως έχουμε ήδη αναφέρει στην περίπτωση της TV , αν θέλουμε να τις διακρίνουμε σαν κατανομές το σφάλμα είναι τουλάχιστον $1 - d_{\mathrm{TV}}(P_1,P_2)$ και δεδομένου ότι επιλέγουμε κάποια από τις δύο κάθε φορά με τυχαίο τρόπο το σφάλμα διάκρισης τους είναι $\frac{1-d_{\mathrm{TV}}(P_1,P_2)}{2}$. Τότε το minimax risk φράσσεται από :

$$\inf_{\hat\theta}\sup_{P\in\Pi} \mathbb{E}[\Phi(\rho(\theta(P),\hat\theta))] \geq \Phi(\delta)[\frac{1-d_{\mathrm{TV}}(P_1,P_2)}{2}]$$

Αντίστοιχα στην περίπτωση των πολλών δειγμάτων

$$\inf_{\hat\theta}\sup_{P\in\Pi^n} \mathbb{E}[\Phi(\rho(\theta(P),\hat\theta))] \geq \Phi(\delta)[\frac{1-d_{\mathrm{TV}}(P_1^n,P_2^n)}{2}]$$

Αν $d_{\mathrm{TV}}(P_1^n,P_2^n) \leq \epsilon$ προκύπτει:

**Θεώρημα 2.12 (Πρώτη μορφή Le Cam Lemma).**

$$\boxed{\inf_{\hat\theta}\sup_{P\in\Pi^n} \mathbb{E}[\Phi(\rho(\theta(P),\hat\theta))] \geq \Phi(\delta)[\frac{(1-\epsilon)}{2}]}$$

Ας δούμε την γενίκευση αυτού του λήμματος.

- Από εδώ και στο εξής ας υποθέτουμε ότι έχουμε έναν χώρο μετρήσιμο χώρο $\Omega$
- Έστω $\Delta(\Omega)$, το σύνολο όλων των κατανομών που μπορούν να οριστούν σε αυτόν τον χώρο.

- Έστω ότι προσπαθούμε να μάθουμε το σύνολο των κατανομών που ανήκουν στο σύνολο $\mathcal{C}$ (η αντίστοιχη concept κλάση του PAC μοντέλου).

- Ας πάρουμε ένα οποιοδήποτε μείγμα από τις διαφορετικές κατανομές της οικογένειας $\mathcal{C}$.
  Δηλαδή,
  $$conv_m(\mathcal{C}) = \{\sum_{k=1}^{|\mathcal{C}|} a_k D_k : D_k \in \mathcal{C}^{\otimes m}, a_i \geq 0, \sum_{k=1}^{|\mathcal{C}|} a_k = 1\}$$

  Για να μην επιβαρύνουμε τον συμβολισμό, επεξηγούμε ότι με τον όρο $\mathcal{C}^{\otimes m}$ εννοούμε το ότι έχουμε επιλέξει $m$ δείγματα από κάποια κατανομή $D \in \mathcal{C}$

**Θεώρημα 2.13** (Δεύτερη μορφή Le Cam lemma)**.**
*Αν:*

- *Έστω ότι* $\Theta = [0,1]$ *&* $\exists A_1, A_2 \subset [0,1], \gamma \in [0,1]$*:*
  $$d(A_1, A_2) = \inf_{a_1 \in A_1, a_2 \in A_2} |a_1 - a_2| \geq \gamma$$

- *Έστω* $\mathcal{D}_1, \mathcal{D}_2 \subset \mathcal{C}$

- *Επίσης* $\forall D \in \mathcal{D}_1 : \theta(D) \in A_1$ *&* $\forall D \in \mathcal{D}_2 : \theta(D) \in A_2$

*Τότε :*

$$minimax\ risk\ for\ m\ samples \geq \frac{\gamma}{2}(1 - \inf_{\begin{cases} p_1 \in conv_m(\mathcal{D}_1) \\ p_2 \in conv_m(\mathcal{D}_2) \end{cases}} [d_{\mathrm{TV}}(p_1, p_2)])$$

Άμεση συνέπεια αυτού του θεωρήματος είναι:

**Πόρισμα 2.3.** *Έστω* $\epsilon \in (0,1)$ *και* $G \subset \Delta(\Omega)$*, η οικογένεια των κατανομών που επιθυμούμε να ανιχνεύσουμε.Ας υποθέσουμε ένα υποσύνολο αυτών, έστω* $\mathcal{D} \subset G$ *και έστω και ένα "κακό σύνολο"* $M \subset \Delta(\Omega) : \forall m, g \in M \times G : d_{\mathrm{TV}}(m, g) > \epsilon$*.*
*Τότε για κάθε* $m \geq 1$*:*

$$minimax\ risk\ for\ m\ samples \geq \frac{1}{2}(1 - \inf_{\begin{cases} p_1 \in conv_m(\mathcal{M}) \\ p_2 \in conv_m(\mathcal{G}) \end{cases}} [d_{\mathrm{TV}}(p_1, p_2)])$$

*Απόδειξη.* Πρόκειται για απλή εφαρμογή των τύπων. Συγκεκριμένα, ας υποθέσουμε ότι οι αλγόριθμοι βγάζουν είτε $A_1 = \{0\}$ είτε $A_2 = \{1\}$ και άρα ας θέσουμε $\gamma = 1$ και να θέσουμε σαν $\mathcal{D}_1 = M, \mathcal{D}_2 = G$. □

Από το τελευταίο πόρισμα και την αρχή του Yao προκύπτει ότι αν έχουμε λάβει $m$ δείγματα και ισχύει ότι:

$$\epsilon \leq \inf_{\begin{cases} p_1 \in conv_m(\mathcal{M}) \\ p_2 \in conv_m(\mathcal{G}) \end{cases}} [d_{\mathrm{TV}}(p_1, p_2)] \leq \frac{1}{3} \Rightarrow \text{minimax risk for m samples} \geq 1/3$$

Συνεπώς χρειάζονται $\Omega(m)$ δείγματα για να καταφέρουμε να μειώσουμε την επίδοση του αλγόριθμου

#### 2.2.4.4 Assaoud Lemma

Το λήμμα του Assaoud βασίζεται σε παρόμοια απλή ιδέα. Αν θέλουμε να μπορούμε να μάθουμε μια κατανομή μέσα σε ένα σύνολο $\mathbb{C}$, θα πρέπει να μπορούμε παράλληλα να ξεχωρίζουμε εύκολα δύο διαφορετικές κατανομές εντός του συνόλου χρησιμοποιώντας το ίδιο πλήθος δειγμάτων που χρησιμοποιούμε στους αλγορίθμους μάθησης.

**Θεώρημα 2.14** (Λήμμα του Assaoud)**.** *Ας υποθέσουμε ότι έχουμε ένα σύνολο* $H = \{D_z\}_{z \in \{0,1\}^r}$*. Αν ισχύει ότι:*

1. $\forall x, y \in \{0,1\}^r$ *η απόσταση μεταξύ* $D_x, D_y$ *είναι τουλάχιστον ανάλογη της Hamming απόσταση:*
$$d_{\mathrm{TV}}(D_x, D_y) \geq a\|x - y\|_1$$
2. $\forall x, y \in \{0,1\}^r$ *και ισχύει* $\|x - y\|_1 = 1$*, για όλες δηλαδή τις διαδοχικές κατανομές στην αρίθμηση, η τετραγωνική απόσταση Hellinger είναι σχετικά μικρή:*
$$d_H(D_x, D_y)^2 \leq \beta$$

*τότε:*

$$MiniMaxRisk \geq \frac{1}{4} ar(1-\beta)^{2m} = \Omega(are^{-\beta m})$$

*και συγκεκριμένα για να επιτύχει κανείς σφάλμα το πολύ* $\epsilon$ *χρειαζόμαστε τουλάχιστον* $\Omega(\frac{1}{\beta} \log \frac{ar}{\epsilon})$

**Παρατήρηση 2.3.** *Αξίζει να παρατηρηθεί ότι επειδή :*

$$1 - \sqrt{1 - d_{\mathrm{TV}}(p,q)^2} \leq d_H(p,q) \leq d_{\mathrm{TV}}(p,q)$$

*αρκεί να δείξει κανείς ότι* $d_{\mathrm{TV}}(D_x, D_y) \leq \beta$*, όπου* $\alpha \leq \beta$*. Δουλεύοντας λίγο όμως με την τελευταία σχέση αρκεί να βρεθεί* $\beta$ *τέτοιο ώστε* $a^2 \leq 2\beta - \beta^2$*.*

## 2.3 PAC Learning σε Κατανομές

### 2.3.1 Back to the Basics: Ο αλγόριθμος του Ιστογράμματος

#### 2.3.1.1 Upper Bound

Σε αυτήν την ενότητα θα μελετήσουμε την πολυπλοκότητα του αλγορίθμου του ιστογράμματος. Το ιστόγραμμα είναι η πρώτη και κλασσικότερη μέθοδος που εφαρμόζεται στην στατιστική.

**Θεώρημα 2.15.** *Υπάρχει αλγόριθμος $A$ με χρονική πολυπλοκότητα $\Theta(n)$ και δειγματοληπτική πολυπλοκότητα $O((n+\ln(1/\delta))\frac{1}{\epsilon^2})$ που να υπολογίζει εμπειρική κατανομή $\hat{\mathbb{P}}_m$ ε κοντά στην αρχική $\mathbb{P}$ κάτω από νόρμα $TV$ .*

*Απόδειξη.* Για να υπολογίσει κανείς την κατανομή πιθανότητας, αρκεί να συλλέξει ένα αρκετά μεγάλο δείγμα και στην συνέχεια να τα ομαδοποιήσει ώστε να προκύψει η εμπειρική κατανομή.

- Η χρονική πολυπλοκότητα είναι γραμμική ως προς το στήριγμα, αφού στην πραγματικότητα εφαρμόζει CountingSort που είναι και το βέλτιστο δυνατό ως προς την πληροφορία που θέλουμε να δώσει ο αλγόριθμος.

- Το μεγάλο ερώτημα είναι πόσο δείγμα χρειαζόμαστε ώστε ο αλγόριθμος να μας δώσει μια εμπειρική κατανομή με $d_{\mathrm{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m) \leq \epsilon$.

- Από το θεώρημα 7 ξέρουμε ότι αν έχουμε $\mathbb{E}[d_{\mathrm{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m)] \leq \epsilon/2$ χρειαζόμαστε $\boxed{m \geq (4n/\epsilon^2)}$

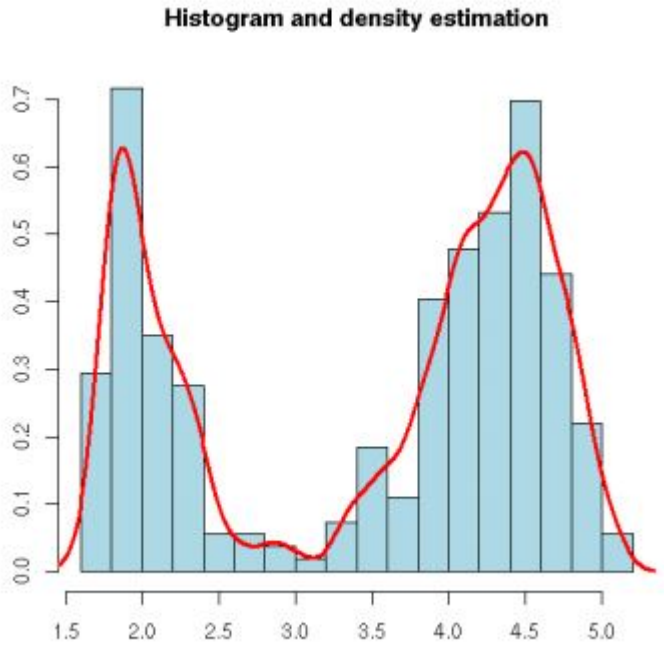
- Για να πετύχουμε ισχυρή συγκέντρωση πιθανότητας, θα επικαλεστούμε το πόρισμα 2.2.
$$\Pr[|d_{\mathrm{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m) - \mathbb{E}[d_{\mathrm{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m)]| \geq \epsilon/2] \leq 2\exp(-m\epsilon^2/2)$$

- Για $\delta \geq 2\exp(-m\epsilon^2/2) \Rightarrow \ln\delta/2 \geq -m\epsilon^2/2 \Rightarrow \boxed{m \geq \ln(1/\delta)\frac{1}{\epsilon^2}}$

- Συνεπώς, για $m = \Omega((n+\ln(1/\delta))\frac{1}{\epsilon^2}))$ έχουμε ότι:

$$\begin{cases} \mathbb{E}[d_{\mathrm{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m)] \leq \epsilon/2 \\ \Pr[|d_{\mathrm{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m) - \mathbb{E}[d_{\mathrm{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m)]| \leq \epsilon/2] \geq 1-\delta \end{cases} \Rightarrow \boxed{\Pr[d_{\mathrm{TV}}(\mathbb{P}, \hat{\mathbb{P}}_m) \leq \epsilon] \geq 1-\delta}$$

□




Σχήμα 2.2: Αλγόριθμος ιστογράμματος

#### 2.3.1.2 Lower Bound

Σε αυτή την ενότητα θα δείξουμε ότι κάτω από πληροφοριο-θεωρητικές παραδοχές, ο καλύτερος αλγόριθμος που υπάρχει για να μάθει κανείς μια άγνωστη πλήρως κατανομή είναι ο αλγόριθμος του ιστογράμματος.

**Θεώρημα 2.16.** *Κάθε αλγόριθμος $A$ που επιθυμεί να "μάθει "μια κατανομή $\mathbb{P}$, δηλαδή να προτείνει μια $\mathbb{Q} : d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$, χρειάζεται $\Omega((n + \ln(1/\delta))\frac{1}{\epsilon^2}))$.*

*Απόδειξη.*

- Ήδη είδαμε ότι χρειαζόμαστε τουλάχιστον αν θέλουμε να πετύχουμε την απαραίτητη συγκέντρωση πιθανότητας $\Omega((\ln(1/\delta))\frac{1}{\epsilon^2}))$.[3]
- Για το δεύτερο σκέλος θα επικαλεστούμε την 2η μορφή του Λήμματος του Le Cam 2.13. Συγκεκριμένα, αν θέλουμε να μπορούμε να προτείνουμε μια $\epsilon$-κοντά κατανομή σε μια ομοιόμορφη κατανομή, θα πρέπει πρώτα να μπορούμε να την ξεχωρίσουμε από κάποιες ιδιαιτέρως κοντινές της. Έτσι έστω ότι από την μια πλευρά έχουμε μόνο

[3] Στο δεύτερο μέρος της απόδειξης χρησιμοποιήσαμε απόδειξη, η οποία μπορεί να γενικευτεί για οποιονδήποτε αλγόριθμο επιδιώκει να βρει κατανομή που να είναι κατά TV $\epsilon$-κοντά στην ζητούσα.

την $Uniform(2n)$, την ομοιόμορφη στους αριθμούς $[1, 2, 3, \cdots, 2n]$ και από την άλλη έχουμε κατανομές που σε κάθε ζεύγος θέσεων $2i, 2i-1$ : $(\frac{1+\epsilon}{n}, \frac{1-\epsilon}{n})$ ή $(\frac{1-\epsilon}{n}, \frac{1+\epsilon}{n})$. Συνεπώς υπάρχουν $2^n$ διαφορετικές κατανομές, τέτοιες "εχθρικές ". Αν εφαρμόσουμε το λήμμα για αυτές τις δύο ομάδες. Κάνοντας πράξεις [Πολ] μπορεί να βρει κανείς ότι η απόσταση των δύο ομάδων κατανομών, στην πραγματικότητα η απόσταση των $m$ δειγμάτων από την ομοιόμορφη και των $m$ δειγμάτων από την συέλιξη όλων των εχθρικών κατανομών προκύπτει ότι $d_{\mathrm{TV}}(Uniform(2n)^{m\ samples}, p_{\mathrm{enemy}}) \leq \frac{1}{2}\sqrt{e^{m^2\frac{\epsilon^4}{2n}} - 1}$. Συνεπώς μπορούμε να παρατηρήσουμε ότι για κατάλληλο $m \leq (\sqrt{n}/\epsilon^2)$ η απόσταση είναι μικρότερη του $1/3$. Από το πόρισμα του Le Cam προκύπτει ότι το minimax risk $> 1/3$. Άρα αν θέλουμε να πετύχουμε πιθανότητα αστοχίας $\delta < 1/3$, θα χρειαστούμε τουλάχιστον $\Omega(\sqrt{n}/\epsilon^2)$

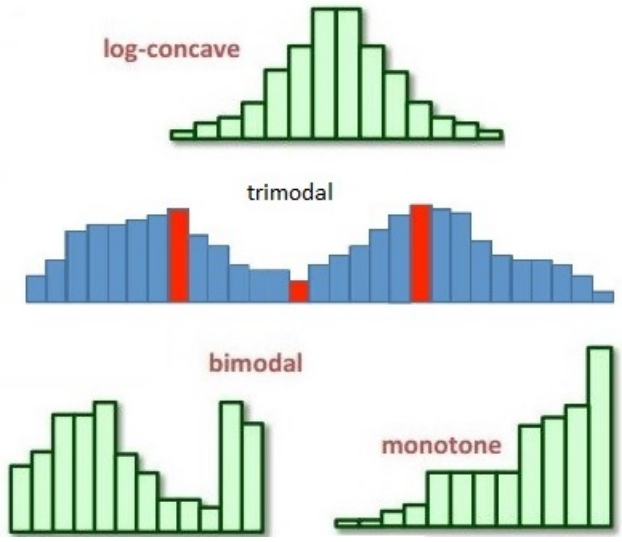
□

#### 2.3.1.3 Σύνοψη

Συνεπώς το κεφάλαιο του Learning κατανομών έκλεισε;;;

Όχι

Ο λόγος είναι απλός. Στην θεωρία πράγματι είναι αρκετές φορές που δεν υπάρχουν τρόποι να γνωρίζεις τίποτα πιο συγκεκριμένο από την κατανομή σου. Μάλιστα μπορεί να μην υπάρχουν καν σαφή μοντέλα για να πάρεις ένα ακέραιο δείγμα, δηλαδή ακόμα και τα ίδια τα δείγματα της κατανομής που ο αλγόριθμος επιδιώκει να μάθει να είναι εκ προοιμίου θορυβημένα και εσφαλμένα. Ο λόγος που η κοινότητα δεν έκλεισε αλλά αντίθετα άνοιξε το ζήτημα, είναι η ύπαρξη του cognition, της εσωτερικής διαίσθησης. Στην πλειοψηφία των περιπτώσεων, υπάρχει κάποια εγγενής ιδιότητα του συστήματος που μας περιορίζει τις επιλογές των κατανομών. Για παράδειγμα τι γίνεται αν γνωρίζεις ότι η κατανομή είναι φθίνουσα, αύξουσα, μονότονη; Τι συμβαίνει αν γνωρίζεις ότι η καμπύλη της συνάρτησης έχεις διάφορες γραφικές ιδιότητες (π.χ μηδενισμοί, κυρτότητα, πλήθος ακρότατων);




Σχήμα 2.3: Διαφορετικές περιπτώσεις κατανομών

Σε όλες αυτές τις περιπτώσεις η πλεονάζουσα αυτή πληροφορία καθιστά πιο εύκολο, τις περισσότερες φορές τουλάχιστον, το πρόβλημα, γιατί υπάρχει ένα σαφέστερο και μικρότερο σύνολο διαφορετικών επιλογών. Συνεπώς αν και πράγματι αν δεν γνωρίζεις τίποτα για την κατανομή-στόχο και τις ιδιότητες το καλύτερο που μπορείς ως επιστήμονας πληροφορικής να προτείνεις είναι ο αλγόριθμος του ιστογράμματος, μόλις αποδεχθείς να γνωρίζεις κάτι περισσότερο για τον κόσμο των κατανομών σου, ανοίγεται ένας ολόκληρος κόσμος.

### 2.3.2 PAC Boolean Learning vs Distribution Learning

Έχοντας δει ήδη μια μεγάλη ποικιλία από εργαλεία και Learning κατανομών, θα προσπαθήσουμε να ορίσουμε με ακρίβεια και λεπτομέρεια την έννοια του Distribution Learning. Με στόχο όμως την ελάφρυνση του κειμένου από πολλούς ορισμούς και για να γίνει εμφανής η βαθύτερη ομοιότητα των ορισμών, δεν θα υπερφορτώσουμε προ-υπάρχοντες ορισμούς σε διαφορετικό setting, εκτός αν είναι απαραίτητο. Αντίθετα θα επιδιώξουμε απλά να αναφέρουμε το τρόπο με τον οποίο κανείς μπορεί να κάνει αναγωγή από το ένα μοντέλο στο άλλο.

Έστω $S$ το στήριγμα των κατανομών για τις οποίες ενδιαφερόμαστε να μελετήσουμε. Όπως και στην αρχική δουλειά του Kearns[KMP$^+$94] αν $S$ είναι πεπερασμένο τότε μπορεί να υποθέσει κανείς χωρίς βλάβη της γενικότητας ότι $S = \{0,1\}^n$, όπου $n$ ο αριθμός των bits τα οποία χρειάζονται για να αναπαρασταθεί οποιοσδήποτε αριθμός $s \in S$.

Όπως αναφέραμε και στο προηγούμενο κεφάλαιο ένα σημαντικό ζήτημα αποτελεί η αναπαράσταση του αντικειμένου που σκοπεύουμε να "μάθουμε ". Υπάρχουν δύο διαφορετικοί τρόποι αναπαράστασης μιας συνάρτησης κατανομής πιθανότητας $D$ πάνω στο $S$.

- **Evaluator**
  Ένας Εκτιμήτορας/ Evaluator $E_D$ για την $D$ είναι ένας μηχανισμός ο οποίος δέχεται

ως είσοδο οποιοδήποτε στοιχείο $s \in S$ και εξάγει ως αποτέλεσμα ένα πραγματικό αριθμό $Ev_D[s]$ το οποίο αντικατοπτρίζει την πιθανότητα του $s$ σύμφωνα την $D$. Δηλαδή $Ev_D[s] = \Pr_{X \sim D}[X = s]$

- **Generator**
  Ένας Γεννήτορας/ Generator $G_D$ για την $D$ είναι ένας μηχανισμός ο οποίος δέχεται ως είσοδο ένα string $str$ από τελείως τυχαία bits και εξάγει ως έξοδο την τιμή $G_D[str] \in S$. Ο Γεννήτορας εν γένει αντικατοπτρίζει μια κλειστή ρουτίνα που προσομοιώνει την διαδικασία της δειγματοληψίας με βάση την κατανομή $D$, δεδομένης μιας σειράς από δίκαια νομίσματα.

**Παρατήρηση 2.4.** *Η κατανομή $D$ θα θεωρείται ότι έχει έναν πολυωνυμικό γεννήτορα/εκτιμήτορα αν υπάρχει αυτός ο μηχανισμός και το αποτέλεσμα το εξάγει σε πολυωνυμικό χρόνος*

Αντίστοιχα με την περίπτωση του προηγούμενου κεφαλαίου θα ορίσουμε ως $\mathbb{C}$ την κλάση κατανομών που επιθυμούμε να μελετήσουμε. Πριν ορίσει κανείς την εκμαθησιμότητα μιας κλάσης κατανομών είναι απαραίτητο να ορίσει κανείς ξεκάθαρα την έννοια της μετρικότητας μεταξύ των κατανομών στην $\mathbb{C}$. Οποιοσδήποτε ορισμός που ακολουθεί μπορεί να εφαρμοστεί για οποιαδήποτε απόσταση από αυτές που αναφέραμε ή και άλλες (Kullback,Total Variation,Kolmogorov,Wasserstein,Hellinger), όμως κάθε φορά η επιλογή που κάνει ο σχεδιαστής του αλγορίθμου αναδεικνύει τον βαθύτερο στόχο του.

Η βασική είσοδος όπως και στην περίπτωση των boolean functions είναι μια ποσότητα δειγμάτων από την κατανομή. Από υπολογιστικής άποψης, μπορεί κανείς να υποθέτει ότι ο αλγόριθμος λαμβάνει ένα δείγμα *sample* σε $O(1)$ χρόνο. Μπορούμε να υποθέσουμε ότι κανείς έχει πρόσβαση σε ένα $G_D[]$ ο οποίος μας επιστρέφει όπως εξηγήσαμε ένα δείγμα από την κατανομή $D$. Όπως ήδη αναφέραμε, εκτός από την υπολογιστική πολυπλοκότητα του χρόνου ή του χώρου, σημαντικό κριτήριο ταξινόμησης της ποιότητας των αλγορίθμων είναι ο αριθμός των δειγμάτων που χρησιμοποιούνται από αυτόν για να προσδιορίσει από ποια κατανομή $D$ από την $\mathbb{C}$ λαμβάνουμε δείγματα.

Για τους αναγνώστες που είναι συγγενείς περισσότερο με τους συμβολισμούς της στατιστικής το πρόβλημα της μάθησης μιας κατανομής μπορεί να περιγραφεί ως εξής. Ας υποθέσουμε ένα training set από ζευγάρια $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ και ας υποθέσουμε (θα χρησιμοποιήσουμε το συμβολισμό που είχαμε στο λήμμα του Le Cam) ότι στόχος μας είναι να βρεθεί μια συνάρτηση $f : X \to Y$ και για τον λόγο αυτό ο αλγόριθμος μας επιδιώκει να εξάγει μια συνάρτηση $g$ που ελαχιστοποιεί το $\mathbb{E}[\Phi(\rho(f, g))]$ . Αν π.χ $\Phi(error) = error^2$ τότε ελαχιστοποιούμε το γνωστό μέσο τετραγωνικό σφάλμα.

Μπορούμε πλέον να ορίσουμε σαφώς την έννοια της αποδοτικής εκμαθησιμότητας και στο σύνολο των συναρτήσεων κατανομής.

**Ορισμός 2.20.** ***Αποδοτική Εκμαθησιμότητα***. *Μια κλάση $\mathbb{C}$ θεωρείται αποδοτικά εκμαθήσιμη (efficiently learnable) αν $\forall \epsilon > 0$ και $\delta \in (0, 1)$ και δοθείσας πρόσβασης στο $G_D[]$, όπου $D \in \mathbb{C}$ η άγνωστη κατανομή, υπάρχει πολυωνυμικός αλγόριθμος $A$ ως προς το πλήθος των δειγμάτων και του χρόνου ώστε να παράγει ως έξοδο έναν Εκτιμήτορα ή έναν*

*Γεννήτορα κάποια κατανομής $D'$ ώστε:*

$$\Pr[d(D, D') \leq \epsilon] \geq 1 - \delta$$

**Παρατήρηση 2.5.** *Αξίζει να παρατηρήσει κανείς μια ουσιαστική διαφορά μεταξύ δύο επιστημών. Στην περίπτωση της Θεωρητικής Πληροφορικής ένας αλγόριθμος είναι αποδοτικός αν το πλήθος των δειγμάτων είναι πολυωνυμικό στο μέγεθος και αντίστοιχα ο χρόνος επεξεργασίας τους επίσης αποδοτικό. Στην περίπτωση όμως της στατιστικής, ένας αλγόριθμος θεωρείται αποδοτικός όταν η δειγματική πολυπλοκότητα φτάνει όσο γίνεται πιο κοντά στο πληροφοριο-θεωρητικό βέλτιστο σφάλμα. Το Ιερόν Δισκοπότηρον της Θεωρίας Μάθησης είναι ακριβώς στην τομή αυτών των δύο αιτημάτων. Να κατασκευάσει κανείς αλγόριθμους που βρίσκονται δειγματικά ακριβώς στα όρια της θεωρίας πληροφορίας και ταυτόχρονα η υπολογιστική ισχύς να είναι γραμμική σε αυτά.*

### 2.3.3 Κριτική στο μοντέλο και στον ορισμό

Σε αυτό το εδάφιο, στόχος μας είναι να δούμε μια αντιπαραβολή κάποιων εννοιών που είχαμε δει μέχρι τώρα στο προηγούμενο κεφάλαιο όπως και να ασχοληθούμε λίγο με δύο βασικές υποθέσεις που βρίσκονται στο πυρήνα του ορισμού της αποδοτικής εκμαθησιμότητας.

Ας ξεκινήσουμε πηγαίνοντας από το Boolean PAC στο Distribution PAC.

#### 2.3.3.1 Μέτρηση Σφάλματος: Απόσταση ή Πιθανότητα

- Αρχικά αξίζει κανείς να σχολιάσει τον διαισθητικό τρόπο μέτρησης απόστασης μεταξύ δύο δυαδικών συναρτήσεων. Μία μέθοδος θα μπορούσε να είναι να μετρήσει κανείς το πλήθος των instances που διαφέρουν ή το ποσοστό επί το συνόλου των όλων δυνατών περιπτώσεων.

- Πρέπει να παρατηρήσει κανείς όμως ότι αυτό δεν είναι εφικτό, επειδή αυτό θα σήμαινε ότι θα μπορούσαμε να γνωρίζουμε ολόκληρη την συνάρτηση $f$ από πριν ή θα έπρεπε να περιμένουμε να λάβουμε όλα τα δυνατά configurations.

> Η μέθοδος που χρησιμοποιείται όμως στο PAC είναι μια ενδιαφέρουσα γενίκευση.

- Ορίζουμε ως απόσταση των δύο κατανομών την πιθανότητα να διαφέρουν σε ένα τυχαίο δείγμα από τον χώρο όλων των instances. Με απλά λόγια, ποια είναι η πιθανότητα να διαφέρουν δύο κατανομές σε ένα στοιχείο που θα επιλεχθεί τυχαία με κάποιο κανόνα τυχαιότητας $D$; Παρατηρείστε ότι αν $D$ είναι η ομοιόμορφη, τότε απλώς υπολογίζουμε το ποσοστό επί του συνόλου όλων των δυνατών περιπτώσεων.

Ας προσπαθήσουμε τώρα να μεταφέρουμε αυτόν τον ορισμό στις κατανομές.

- Ας υποθέσουμε ότι ως σφάλμα θα ορίσουμε πάλι την πιθανότητα $\Pr[X \neq Y]$ όπου :

$$\begin{cases} X \text{ είναι ένα τυχαίο δείγμα της άγνωστης κατανομής} \\ Y \text{ είναι ένα τυχαίο δείγμα της προταθείσας από τον αλγόριθμο κατανομή .} \end{cases}$$

- Αν αυτά ταυτίζονται με μεγάλη πιθανότητα, τότε μπορούμε να πούμε ότι αυτές οι δύο κατανομές μοιάζουν αρκετά.

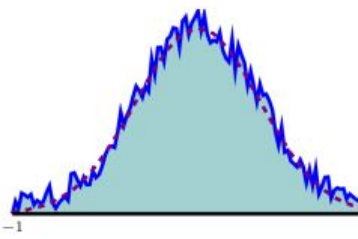
- Από τα μαθηματικά όμως όπως είδαμε ισχύει ότι $\Pr[X \neq Y] \geq d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$, όπου $\mathbb{P}$ η άγνωστη κατανομή και $\mathbb{Q}$ η κατανομή προταθείσα κατανομή. Αξίζει να σημειωθεί ότι ο τρόπος με τον οποίο μπλέκονται οι δύο τυχαίες μεταβλητές σε ένα κοινό χώρο τυχαιότητας δεν είναι πάντα μοναδικός. Αν ορίσουμε όμως αυτός να είναι ο optimal τότε $\Pr[X \neq Y] = d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$. Σίγουρα υπάρχουν διαφορετικά coupling και το καθένα από αυτά μας προσφέρει μια διαφορετική μέτρηση της ποσότητας σφάλματος, αλλά η διαίσθηση που ώθησε αρκετούς στατιστικούς στην TV ήταν ακριβώς ο ίδιος παραλληλισμός με την πρόταση του Leslie. Η μέτρηση του σφάλματος, η πιθανότητα να μην ταυτίζονται ο στόχος και η υπόθεση, η απόσταση του στόχου και της απόστασης να είναι ταυτόσημες έννοιες.

- Ασφαλώς υπάρχουν όμως και προβλήματα όπου η διαίσθηση ή σωστότερα η απαίτηση του προβλήματος μας οδηγεί στο να μάθουμε κατανομές με ισχυρότερες ή και ασθενέστερες μετρικές αποστάσεις όπως την Kolmogorov ή Kullback

Ας επανέλθουμε όμως λίγο στην συζήτηση των υποθέσεων που λαμβάνει κανείς όταν επικαλείται τον ορισμό. Η πρώτη αρκετά ισχυρή υπόθεση είναι ότι στα χέρια του αλγόριθμου βρίσκεται πάντα ένας μηχανισμός άντλησης δειγμάτων από την κατανομή χωρίς κανένα πρόβλημα. Στα πραγματικά προβλήματα όμως αυτή η μοντελοποίηση είναι αρκετά ψευδής.

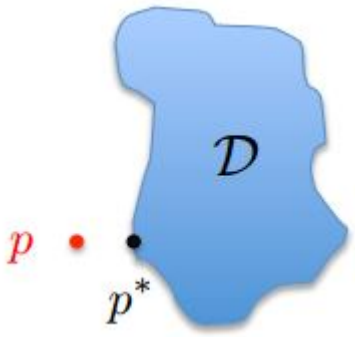
#### 2.3.3.2 Agnostic or Not Μάθηση

Στην πλειοψηφία των περιπτώσεων τα δείγματα που έχουμε για την κατανομή μπορεί να είναι από παραχαραγμένα, αριθμητικώς επηρεασμένα στην ακρίβεια έως και τελείως ψευδή. Αυτή την περίπτωση προσπαθεί να συλλάβει το θορυβημένο ή όπως λέγεται στην βιβλιογραφία το αγνωστικιστικό μοντέλο(agnostic model). Σε αυτή την περίπτωση θεωρούμε ένα αντίπαλο θόρυβο (adversarial noise) μέσα στα δεδομένα. Στην περίπτωση αυτή δεν κάνουμε καμία υπόθεση για την συνάρτηση κατανομής στόχο $p$ και μοναδικός μας σκοπός είναι να μπορέσουμε με πιθανότητα $1-\delta$ να βρούμε κάποια κατανομή $h$ ώστε να ισχύει :



Σχήμα 2.4: Agnostic Model

$$d(p,h) \leq \epsilon + a \inf_{q \in \mathbb{C}} d(q,p), a > 1$$




Σχήμα 2.5: Agnostic Learning

- Ποιος είναι στην πραγματικότητα ο στόχος μας εδώ;
- Επειδή κατά πάσα πιθανότητα ο θόρυβος πάνω στα δεδομένα έχει διαμορφώσει μια συνάρτηση κατανομής $p$ που βρίσκεται εκτός του $\mathbb{C}$, στόχος μας είναι να πετύχουμε σφάλμα $\epsilon$ και ταυτόχρονα να βρεθούμε όσο πιο κοντά γίνεται και εντός της πάλι της κατανομής $\mathbb{C}$.

#### 2.3.3.3 Proper or Not Μάθηση

Όπως και στην περίπτωση των Boolean συναρτήσεων και εδώ μπορούμε να ορίσουμε την έννοια του proper learning. Αν λοιπόν η κατανομή που προτίνει ο αλγόριθμος προέρχεται από την ίδια κλάση από την οποία προέρχονται και οι συναρτήσεις κατανομής στόχοι, τότε θεωρούμε ότι ο αλγόριθμος κάνει proper learning. Στην περίπτωση των κατανομών ο λόγος για τον οποίο ο σχεδιαστής θα αποφασίσει να έχει ή όχι proper μάθηση μπορεί να ποικίλει. Στις πιθανότητες για παράδειγμα ο νόμος των μεγάλων αριθμών μας δίνει μια καλή προσέγγιση μιας πλειάδας κατανομών από την Κανονική Gaussian κατανομή, ή αντίστοιχα ο νόμος των μικρών αριθμών μας δίνει την κατανομή Poisson στο όριο των διωνυμικών κατανομών. Αν κανείς αναζητεί απλά την αριθμητική προσέγγιση της κατανομής είναι πολύ πιθανόν ότι τέτοιο non proper τρόποι να αρκούν. Αν όμως ο αλγόριθμος στην συνέχεια σκοπεύει να επεξεργαστεί αυτή την έξοδο σε κάποιο μετέπειτα στάδιο, τότε η ομαλότητα ή τα απότομα σημεία της συνάρτησης που προέκυψε με non proper τρόπο να παίζουν βαρύνουσα σημασία.

Ένα επίσης σημαντικό σχόλιο που αφορά το proper ή non-proper learning, είναι ότι στην βάση τους έχουν την ίδια δειγματική πολυπλοκότητα, όχι όμως απαραίτητα την ίδια υπολογιστική πολυπλοκότητα. Αντίστοιχη οπτική βρίσκεται και στο καθαρό ή θορυβημένο μοντέλο. Γιατί; Αρκεί κανείς να εφαρμόσει έναν non-proper αλγόριθμο και ύστερα να αναζητήσει μία-προς-μία όλες τις κατανομές εντός της κλάσης και να υπολογίσει ποια είναι η πλέον συγγενική της. Όμως σε αυτή την περίπτωση μπορεί να χρειαστεί να ασχοληθεί σε εκθετικά το πλήθος μεγάλη ομάδα διαφορετικών κατανομών.

#### 2.3.3.4 Parameter or not Μάθηση

Σε μερικές περιπτώσεις όμως κλάσεων, η ίδια η κλάση περιγράφεται συμπυκνωμένα από ένα σύνολο παραμέτρων.

Για παράδειγμα όλες οι κατανομές Gauss:

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

Είναι σαφές ότι αρκεί κανείς να προσδιορίσει τις παραμέτρους $\sigma, \mu$ και θα έχει προσδιορίσει απολύτως την κατανομή. Σε αυτή την περίπτωση ο αλγόριθμος μάθησης θα λέγεται ότι πραγματοποιεί parameter learning algorithm. Σαφώς στην περίπτωση των απλών κλάσεων το ζήτημα της παραμετρικής εκμάθησης είναι ήδη πολύ καλά μελετημένο από τον κλάδο των στατιστικών. Η υπολογιστική θεωρία μάθησης στόχος της είναι να διευρύνει τα εργαλεία αυτά και να εξετάσει στην περίπτωση των πιο περίπλοκων παραμετρικά κλάσεων να προτείνει αλγορίθμους μάθησης.

### 2.3.4 Cover Method. Από το VC-dimension στην Kolmogorov Method .

Σε αυτή την ενότητα θα περιγράψουμε διαισθητικά μια από τις σημαντικότερες και ισχυρότερες μεθόδους που αναπτύχθηκαν και αποτέλεσε και το βασικότερο εργαλείο της εργασίας που παρουσιάζεται σε αυτή την διπλωματική εργασία.

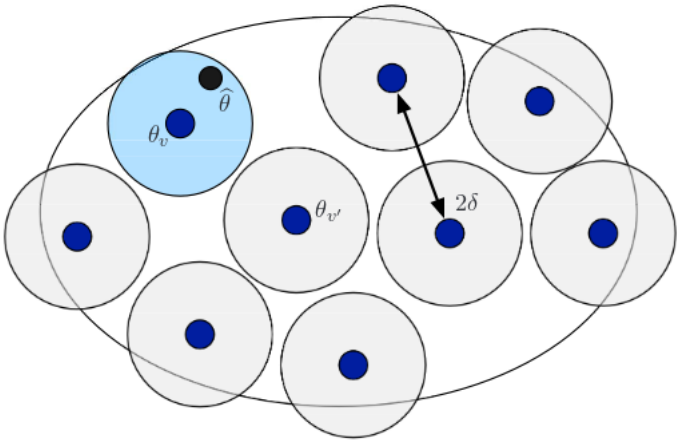
Η μέχρι τώρα συζήτηση μας έχει οδηγήσει στην εξής θεμελιώδη ερώτηση.

Ποιος είναι ο ικανός και αναγκαίος αριθμός δειγμάτων που χρειάζεται κανείς να ώστε να μάθει μια άγνωστη κατανομή $p \in \mathbb{C}$;

Αν και αυτή η ερώτηση έχει μελετηθεί επί ένα περίπου αιώνα στο τομέα της θεωρίας πληροφορίας, της θεωρητικής στατιστικής και της θεωρητικής πληροφορικής ακόμα και για τις απλούστερες κλάσεις πολυπλοκότητες πολλές φόρες τα ακριβή βέλτιστα φράγματα είναι αρκετά δυσδιάκριτα. Είναι φανερό ότι η βέλτιστη δειγματική πολυπλοκότητα εν τέλει εξαρτάται από την μορφολογία που εμφανίζουν οι κατανομές μεταξύ τους. Σε αυτή την μέθοδο βασίστηκε ο Kolmogorov το 1950 και έχει τα σπάργανα της στην στοιχειώδη μαθηματική ανάλυση. Ας θυμηθούμε αρχικώς την έννοια της πυκνότητας ενός συνόλου σε ένα υπερσύνολο κάτω από έναν μετρικό χώρο. Ας υποθέσουμε ότι υπάρχει ένας μετρικός χώρος $(X, d)$ και ας επιτρέψουμε μια παράμετρο $\delta$.

**Ορισμός 2.21.** *$\epsilon-$**Κάλυμμα**$_d$. $\epsilon-$Κάλυμμα$_d$ καλείται ένα σύνολο $Y$ για το οποίο ισχύει ότι $\forall x \in X \; \exists y \in Y d(x, y) < \epsilon$. Δηλαδή για οποιαδήποτε στοιχείο του συνόλου $X$ μπορείς να βρεις ένα στοιχείο που να είναι $\epsilon$ αρκετά κοντά του.*

Στην περίπτωση όπου $Y \subseteq Q$ τότε το κάλυμμα καλείται **Κανονικό** ή **Proper** Μπορεί κανείς να δει μια άμεση ομοιότητα με την έννοια του πακεταρίσματος που ορίσαμε στην μελέτη των εργαλείων εξαγωγής κάτω φραγμάτων σε προηγούμενη ενότητα.




Μπορεί κανείς να παρατηρήσει ότι υπάρχουν πολλά διαφορετικά καλύμματα. Εμάς όμως μας ενδιαφέρουν αυτά που ο πληθάριθμος τους είναι ο ελάχιστος δυνατός. Αξίζει κανείς να δει για παράδειγμα την περίπτωση των ρητών αριθμών $\mathbb{Q}$. Το σύνολο των ρητών αριθμών, όπως είναι γνωστό είναι πυκνό στο σύνολο των πραγματικών αριθμών έχει μέτρο 0 ενώ το $\mathbb{R}$ έχει μέτρο $+\infty$, έχει πληθάριθμο $\aleph_0$ ενώ οι πραγματικοί αριθμοί έχουν $\aleph_1$.

**Ορισμός 2.22.** ***Αριθμός κάλυψης(Covering number).*** *Ως αριθμό κάλυψης ορίζουμε τον ελάχιστο πληθάριθμο που μπορεί να έχει οποιοδήποτε $\epsilon$ κάλυμμα*

Παρατηρείστε ότι ο αριθμός κάλυψης περιγράφει μια πολύ ιδιαίτερη και ουσιώδη πληροφορία για τον μετρικό χώρο που μελετάμε. Εκφράζει την έκταση, το span που φέρει αυτός ο χώρος. Για τους αριθμούς κάλυψης και τους λογαρίθμους τους, οι οποίοι είναι γνωστοί ως αριθμοί μετρικής εντροπίας ο αναγνώστης παραπέμπεται σε μια σειρά από βιβλία [Δυδ74, Ψ.Μ86, ρον17, Κολ93, Τσψ09, δ΄Ω96, ΗΙ90, ΗΟ97, ΄Σ90, ΕΤ96].

Η σημασία του καλύμματος αποκαλύπτεται από το παρακάτω πολύ σημαντικό θεώρημα. Θα περιοριστούμε για λόγους απλότητας στο σύνολο των διακριτών κατανομών γύρω από το $S = [n]$.

**Θεώρημα 2.17.** *Έστω $\mathbb{C}$ μια τυχαία οικογένεια κατανομών και $\epsilon > 0$. Έστω $C$ ένα $\epsilon-$κάλυμμα πληθάριθμου $N$. Τότε υπάρχει ένας αλγόριθμος ο οποίος χρησιμοποιεί $\frac{\log N}{\epsilon^2} \log \frac{1}{\delta}$ δείγματα από μια άγνωστη κατανομή $p \in \mathbb{C}$ και με πιθανότητα $1 - \delta$ επιτυγχάνει να βρίσκει μια συνάρτηση $h$ ώστε $d_{\mathrm{TV}}(h, p) < \Theta(\epsilon)$*

*Απόδειξη.* Για να αποδείξουμε το σημαντικότατο αυτό θεώρημα θα αποδείξουμε πρώτα ένα βασικό λήμμα, την ρουτίνα Choose-Hypothesis$^X$. Η ρουτίνα αυτή χρησιμοποιεί δείγματα από την άγνωστη κατανομή $X$ και στην πραγματικότητα τρέχει έναν διαγωνισμό, ένα τουρνουά μεταξύ δύο υποψήφιων κατανομών υπόθεσης $H_1, H_2$. Θα δείξουμε ότι αν τουλάχιστον μια κατανομή από τις δύο βρίσκεται αρκετά κοντά στην κατανομή $X$, τότε με πολύ μεγάλη πιθανότητα πάνω στα δείγματα που ζητήσαμε από τον γεννήτορα της κατανομής $X$, ο διαγωνισμός θα την προτείνει ως νικήτρια. Παρόμοιες τεχνικές εμφανίζονται και στις [ΔΓ01, Ψατ85, Τσψ09].

**Λήμμα 2.2.** *Υπάρχει ένας αλγόριθμος Choose-Hypothesis$^X(H_1, H_2, \epsilon', \delta')$ ο οποίος έχοντας πρόσβαση σε ένα γεννήτορα δειγμάτων από μια άγνωστη κατανομή $X$ σε μια εκτιμήτορα για δύο γνωστές κατανομές $H_1, H_2$ μια παράμετρο ακρίβειας $\epsilon'$ και τέλος μια παράμετρο εμπιστοσύνης $\delta'$, ο οποίος χρησιμοποιώντας*

$$m = O(\log(1/\delta')/\epsilon'^2)$$

*δείγματα από την $X$ επιστρέφει μια κατανομή $H \in \{H_1, H_2\}$. Αν υπάρχει κάποια κατανομή $d_{\mathrm{TV}}(H_i, X) \leq \epsilon'$ για $i \in \{1, 2\}$, τότε με πιθανότητα τουλάχιστον $1 - \delta'$ η κατανομή $H$ που ο αλγόριθμος προτείνει θα έχει απόσταση το πολύ $d_{\mathrm{TV}}(H, X) \leq 6\epsilon'$.*

*Απόδειξη.* Ας δούμε αρχικά τον αλγόριθμο Choose-Hypothesis.

Choose-Hypothesis$(H_1, H_2, \epsilon', \delta')$
Είσοδος: $G_X[]\cdot$ Ένα ζευγάρι από κατανομές $:(H_1, H_2)\cdot$ Παράμετρος ακρίβειας και εμπιστοσύνης $\epsilon', \delta' > 0$.

Έστω $\mathcal{W}$ το στήριγμα της $X$, $\mathcal{W}_1 = \mathcal{W}_1(H_1, H_2) := \{w \in \mathcal{W} \mid H_1(w) > H_2(w)\}$, και $p_1 = H_1(\mathcal{W}_1)$, $p_2 = H_2(\mathcal{W}_1)$. /* *Σήγουρα, $p_1 > p_2$ και $d_{\mathrm{TV}}(H_1, H_2) = p_1 - p_2$.* */

1. Αν $p_1 - p_2 \leq 5\epsilon'$, πρότεινε ισοπαλία και πρότεινε τυχαία $H_i$. Διαφορετικά:
2. Ζήτα $m = 2\frac{\log(1/\delta')}{\epsilon'^2}$ δείγματα $s_1, \ldots, s_m$ από το $X$, και έστω $\tau = \frac{1}{m}|\{i \mid s_i \in \mathcal{W}_1\}|$ είναι το μέρος των δειγμάτων που βρίσκονται μέσα στο $\mathcal{W}_1$.
3. Αν $\tau > p_1 - \frac{3}{2}\epsilon'$, πρότεινε $H_1$ ως νικητή. Διαφορετικά:
4. Αν $\tau < p_2 + \frac{3}{2}\epsilon'$, πρότεινε $H_2$ ως νικητή. Διαφορετικά:
5. πρότεινε ισοπαλία και διάλεξε τυχαία $H_i$.

Σχήμα 2.6: Choose-Hypothesis$(H_1, H_2, \epsilon', \delta')$

Ας υποθέσουμε λοιπόν ότι πράγματι για κάποιο $i \in \{1, 2\}$ ισχύει όντως κάποια κατανομή $d_{\mathrm{TV}}(H_i, X) \leq \epsilon$. Τότε:

(i) Αν $d_{\mathrm{TV}}(X, H_{3-i}) > 6\epsilon'$, η πιθανότητα η ρουτίνα Choose-Hypothesis$^X(H_1, H_2, \epsilon', \delta')$ να μην προτείνει νικητή την $H_i$ είναι το πολύ $2e^{-m\epsilon'^2/2}$, όπου $m$ διαλέγεται όπως ορίστηκε πριν στην ρουτίνα.(Διαισθητικά, αν η $H_{3-i}$ είναι αρκετά άσχημη τότε είναι υπερβολικά πιθανό η $H_i$ να ανακυρηχθεί νικήτρια.)

(ii) Αν $d_{\mathrm{TV}}(X, H_{3-i}) > 4\epsilon'$, η πιθανότητα η ρουτίνα Choose-Hypothesis$^X(H_1, H_2, \epsilon', \delta')$ ορίζει ως νικητή την $H_{3-i}$ είναι το πολύ $2e^{-m\epsilon'^2/2}$. (Διαισθητικά, αν η $H_{3-i}$ είναι μετριοπαθώς άσχημα τότε η ισοπαλία είναι αρκετά πιθανή αλλά είναι επίσης και αρκετά απίθανο να ανακηρυχθεί ως νικητής η $H_{3-i}$.)

Πράγματι έστω ότι $r = X(\mathcal{W}_1)$. Από τον ορισμό της total variation συνεπάγεται ότι $|r - p_i| \leq \epsilon'$. Ας υποθέσουμε ανεξάρτητες δείκτριες $\{Z_j\}_{j=1}^m$ τέτοιες ώστε, για όλα τα $j$, $Z_j = 1$ αν και μόνο αν $s_j \in \mathcal{W}_1$. Σίγουρα, $\tau = \frac{1}{m}\sum_{j=1}^m Z_j$ και $\mathbb{E}[\tau] = \mathbb{E}[Z_j] = r$. Αφού $Z_j{}'$ είναι αμοιβαία ανεξάρτητα, επάγεται από τα Chernoff Bounds ότι η $\Pr[|\tau - r| \geq \epsilon'/2] \leq 2e^{-m\epsilon'^2/2}$. Χρησιμοποιώντας $|r - p_i| \leq \epsilon'$ παίρνουμε ότι $\Pr[|\tau - p_i| \geq 3\epsilon'/2] \leq 2e^{-m\epsilon'^2/2}$. Συνεπώς :

- Για το πρώτο τμήμα (i): Αν $d_{\mathrm{TV}}(X, H_{3-i}) > 6\epsilon'$, από τριγωνική ανισότητα προκύπτει ότι $p_1 - p_2 = d_{\mathrm{TV}}(H_1, H_2) > 5\epsilon'$. Συνεπώς ο αλγόριθμός θα προχωρήσει μακρύτερα της ισοπαλίας και με πιθανότητα τουλάχιστον $1 - 2e^{-m\epsilon'^2/2}$, και θα καθορίσει νικητή στο 3ο ή στο 4ο βήμα.

- Για το δεύτερο τμήμα (ii): Αν $p_1 - p_2 \leq 5\epsilon'$ τότε ο διαγωνισμός θα ανακηρύξει ισοπαλία, άρα ο $H_{3-i}$ δεν θα είναι νικητής. Διαφορετικά αν $p_1 - p_2 > 5\epsilon'$ τότε θα ανακηρυχθεί νικητής ο $H_{3-i}$ με πιθανότητα το πολύ $2e^{-m\epsilon'^2/2}$.

□

Tournament$(\epsilon', \delta')$
Είσοδος: $G_X[]$· Κάλυμμα από κατανομές :$\Sigma_\epsilon$,$|\Sigma_\epsilon| = N$· Παράμετρος ακρίβειας και εμπιστοσύνης $\epsilon', \delta' > 0$.

1. Εφαρμόζουμε ανα δύο μεταξύ τους $\binom{n}{2}$ την ρουτίνα Choose-Hypothesis για $\delta = \delta/4N$. Αν βρεθεί καθολικός νικητής, δηλαδή κατανομή που να φέρνει ισοπαλίες ή νίκες με όλες τις υπόλοιπες κατανομές προτείνεται νικητής. Διαφορετικά:
2. Έχουμε αποτυχία και απλά διάλεξε μια στην τύχη και πρότεινε την ως νικητή.

Σχήμα 2.7: Tournament$(\epsilon', \delta')$

Αφού έχουμε ένα κάλυμμα $\mathcal{S}_\epsilon$ για το σύνολο των κατανομών μας $\mathcal{S}$, σίγουρα υπάρχει κάποια κατανομή $Y \in \mathcal{S}_\epsilon$ τέτοια ώστε $d_{\mathrm{TV}}(X, Y) \leq \epsilon$.

Αρχικά θα επιχειρηματολογήσουμε με μεγάλη πιθανότητα ότι η κατανομή $Y$ δεν χάνει σε διαγωνισμό με οποιοδήποτε άλλη $Y' \in \mathcal{S}_\epsilon$. Δηλαδή δεν θα έχουμε ποτέ αποτυχία στο τουρνουά. Πράγματι, θεωρείστε οποιαδήποτε $Y' \in \mathcal{S}_\epsilon$. Αν $d_{\mathrm{TV}}(X, Y') > 4\epsilon$, όπως είδαμε πριν η πιθανότητα ο $Y$ από το $Y'$ είναι το πολύ $2e^{-m\epsilon^2/2} \leq \frac{\delta}{2N}$. Από την άλλη πλευρά αν $d_{\mathrm{TV}}(X, Y') \leq 4\epsilon$, από την τριγωνική ανισότητα ισχύει ότι $d_{\mathrm{TV}}(Y, Y') \leq 5\epsilon$ και έτσι στην χειρότερη $Y$ και $Y'$ θα έρθουν σε ισοπαλία.

Χρησιμοποιώντας ένα απλό union bound πάνω στις υπόλοιπες $N - 1$ κατανομές $\in \mathcal{S}_\epsilon \setminus \{Y\}$ μπορεί να δείξει κανείς ότι με πιθανότητα τουλάχιστον $1 - \delta/2$, η κατανομή $Y$ δεν χάνει ποτέ..

Επίσης μπορούμε να επιχειρηματολογήσουμε ότι με πιθανότητα τουλάχιστον $1-\delta/2$, κάθε άλλη κατανομή $Y' \in \mathcal{S}_\epsilon$ η οποία δεν χάνει επίσης βρίσκεται αρκετά κοντά στην $X$ επίσης. Υποθέστε ότι έχετε μια κατανομή $Y'$ τέτοια ώστε $d_{\mathrm{TV}}(Y', X) > 6\epsilon$. Από την προηγούμενη μελέτη μας ξέρουμε ότι η $Y'$ χάνει από την $Y$ με πιθανότητα τουλάχιστον $1 - 2e^{-m\epsilon^2/2} \geq 1 - \delta/(2N)$. Με union bound έχουμε ότι με πιθανότητα τουλάχιστον $1-\delta/2$, κάθε κατανομή $Y'$ για την οποία ισχύει $d_{\mathrm{TV}}(Y', X) > 6\epsilon$ θα χάσει κάποια στιγμή.

Συνεπώς με πιθανότητα $1-\delta$, το τουρνουά δεν αποτυγχάνει στο να εξάγει μια κατανομή $Y^\star$ για την οποία ισχύει ότι $d_{\mathrm{TV}}(X, Y^\star) \leq 6\epsilon$. □

Η προφανής υλοποίηση του τουρνούα προφανώς χρειάζεται χρόνο $\Omega(\frac{N^2}{\epsilon^2})$. Πρόσφατη αλγοριθμική δουλεία [ΔΚ14β, ΣΟΑΘ14] βελτίωσε σε σχεδόν γραμμικό το χρόνο διεξαγωγής του τουρνουά σε $O(N \log N/\epsilon^2)$. Παρ' όλα αυτά ο χρόνος εκτέλεσης είναι εκθετικός ως προς τα δείγματα που λαμβάνει.[4].

Όπως φαίνεται και από την απόδειξη μπορεί κανείς να δει την ρουτίνα Choose-Hypothesis ως μια ρουτίνα που επιλέγει από δύο υποψήφιες κατανομές. Αν τουλάχιστον μια από αυτές τις κατανομές βρίσκεται κονα στην κατανομή στόχο, τότε με μεγάλη πιθανότητα η ρουτίνα θα την επιλέξει και ως νικήτρια στον διαγωνισμό που τρέχει. Αξίζει να παρατηρήσει κανείς ότι η βασική ανάλυση του αλγορίθμου είναι αρκετά απλή και στηρίζεται μόνο σε Chernoff Bounds.

Κάτι ακόμα σημαντικό με την μέθοδο του καλύμματος είναι ότι είναι noise-tolerance-θορυβικά ανθεκτική. Ο αναγνώστης για περισσότερες λεπτομέρειες παραπέμπεται στο κεφάλαιο 7.3 του [ΔΓ01]

Αξίζει επίσης να επισημανθεί ότι στην γενική περίπτωση η δειγματική πολυπλοκότητα δεν μπορεί να βελτιωθεί, τουλάχιστον υπό την έννοια ότι υπάρχουν οικογένειες κατανομών όπου απαιτούν αυτόν τον αριθμό δειγμάτων ή ισοδύναμο αυτού αν αφαιρέσουμε του λογαριθμικού όρους.

Ο Yang και ο Barron στο [ΨΒ99] έδειξαν ότι πλήθος από ομαλές μη παραμετρικές οικογένειες κατανομών χαρακτηρίζονται στην δυσκολία τους ως προς την εκμάθηση αυστηρά μέσω των αριθμών μετρικής εντροπίας, δηλαδή, το Cover αποτελεί την βέλτιστη αντιμετώπιση για αυτά.

Παρ'όλα αυτά υπάρχουν και κατανομές όπου τα πράγματα είναι σαφώς ευκολότερα στην ειδική τους περίπτωση. Απλό παράδειγμα ας πάρουμε τις Dirac κατανομές που δίνουν όλη την μάζα πιθανότητας τους σε ένα ακριβώς στοιχείο σε ένα στήριγμα από το $[1 \cdots n]$. Στην περίπτωση αυτή το δειγματικό φράγμα του τουρνουά είναι πολυωνυμικό αφού το στήριγμα έχει μέγεθος $n$ και η πολυπλοκότητα του αλγόριθμου είναι $\frac{\log n}{\epsilon^2}$.

Εδώ τίθεται και το σπουδαιότερο ζήτημα στην περίπτωση του Distribution learning σε σχέση με το κλασσικό PAC Boolean μοντέλο. Υπάρχει μέτρο πολυπλοκότητας που μπορεί να χαρακτηρίσει την δειγματική πολυπλοκότητα μιας κλάσης κατανομών; Στην περίπτωση του PAC υπάρχει και λέγεται VC-dimension και φράσσει από τα κάτω στην δειγματική πολυπλοκότητα οποιοδήποτε πρόβλημα PAC μελετάται.

[4]Για σταθερό $\epsilon = 0.01$ χρησιμοποιεί $10^4 \log N$ δείγματα αλλά τρέχει σε χρόνο $10^4 2^{\log N} \log N$

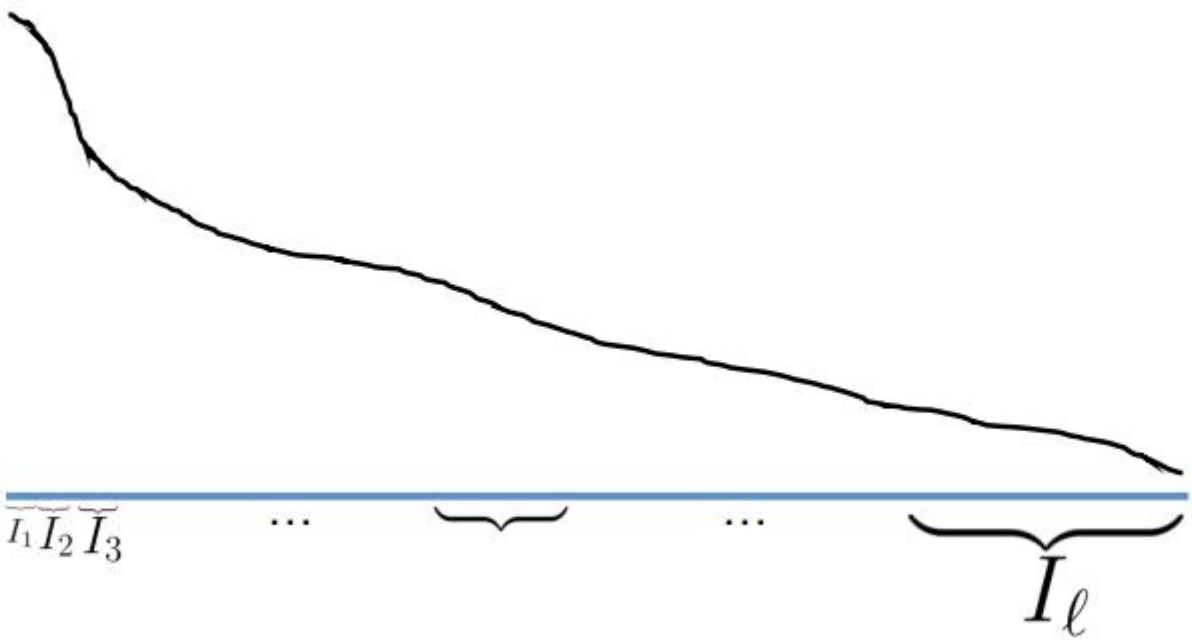
### 2.3.5 Μονότονες κατανομές: Ο αλγόριθμος του Birgé

Σε αυτή την τελευταία ενότητα θα δούμε ένα από τα πλέον θεμελιώδη αποτελέσματα που κατακτήθηκαν τον προηγούμενο αιώνα πάνω στις μονότονες και μονόλοφες κατανομές.[Birg86, Birg87α, Birg87β] Όπως σχολιάσαμε σε προηγούμενη ενότητα κάθε φορά που θα μπορούμε να είμαστε όσο πιο συγκεκριμένοι είμαστε πάνω στο χώρο αναζήτησης θα πρέπει να μπορούμε να μειώσουμε την πολυπλοκότητα τόσο στον αριθμό των δειγμάτων όσο και στον χρόνο.

#### 2.3.5.1 Upper Bound

Για να υπολογίσουμε την κατανομή θα αποδείξουμε ένα βασικό θεώρημα, το οποίο προκαλεί και ιδιαίτερη έκπληξη αφού είναι τυφλό - oblivious ως προς τα δεδομένα της συνάρτησης[5].

**Θεώρημα 2.18.** ***Μονότονη αποσύνθεση*** *Έστω οποιοσδήποτε $n \in \mathbb{Z}^+$ και $\epsilon > 0$. Για κάθε $p$ φθίνουσα κατανομή, υπάρχει $I$ μια διαμέριση του χώρου $[n] = \{1, 2, \cdots, n\}$, $\mathcal{I} := \{I_i\}_{i=1}^{\ell}$ στην οποία το $j$-στο διάστημα έχει μήκος $\lfloor (1+\epsilon)^j \rfloor$ και συνολικό μέγεθος $\ell = O\left((1/\epsilon) \cdot \log(\epsilon \cdot n + 1)\right)$ ώστε η $\mathcal{I}$ να ορίζει μια $(p, O(\epsilon), \ell)$-κλιμακωτή αποσύνθεση της κατανομής $p$ πάνω στο $[n]$.*




Σχήμα 2.8: Μονότονη Αποσύνθεση

[5]Στην περίπτωση του Machine Learning θα έλεγε κανείς ότι είναι μη-προσαρμοστικός

Με την έννοια φθίνουσα κλιμακωτή αποσύνθεση εννοούμε ότι υπάρχει μια διαμέριση σε συνεχόμενες ομάδες φυσικών-διαστήματα $\{1, 2, \cdots, k_1\}, \{k_1 + 1, k_1 + 2, \cdots, k_1 + k_2\}, \cdots, \{k_1 + k_2 + \cdots + k_{\ell-1}, k_1 + k_2 + \cdots + k_{\ell-1} + 1, k_1 + k_2 + \cdots + k_{\ell-1} + 2, \cdots k_1 + k_2 + \cdots + k_{\ell-1} + k_\ell\}$, στα οποία η κατανομή είναι σταθερή αλλά καθώς μετακινούμαστε από την πρώτη προς την τελευταία ομάδα η συνάρτηση διαμορφώνεται φθίνουσα σε σκαλοπάτια. Το ότι είναι μια $(p, O(\epsilon), \ell)$-κλιμακωτή αποσύνθεση της κατανομής $p$ πάνω στο $[n]$ σημαίνει ότι η κατανομή $\hat{p}$ που προκύπτει από την κλιμακωτή αποσύνθεση έχει απόσταση από την αρχική το πολύ $O(\epsilon)$ χρησιμοποιώντας $\ell$ σκαλοπάτια. Συνεπώς το θεώρημα διατείνεται ότι μπορεί κανείς πάντα να κατασκευάσει με κατάλληλο αριθμό δειγμάτων από την άγνωστη κατανομή $p$ μια κατανομή όπου θα παράγει $\hat{p} : d_{\mathrm{TV}}(p, \hat{p}) \leq O(\epsilon)$.

*Απόδειξη.* Αρχικά μπορούμε να παρατηρήσουμε ότι αν το σφάλμα είναι μικρότερο του $\frac{1}{n}$ τότε η αποσύνθεση είναι προφανής αφού μπορούμε να πάρουμε ως σκαλοπάτι την μετακίνηση από το σημείο $k$ στο σημείο $k + 1$.

Ας περιγράψουμε πρώτα την μέθοδο αποσύνθεσης και στην συνέχεια θα δείξουμε την ορθότητα του θεωρήματος. Έστω η αποσύνθεση $I$ του $[n]$ σε $\ell$ μη κενά συνεχόμενα διαστήματα $I_1, \ldots, I_\ell$. Συγκεκριμένα, για $j \in \{1, 2, \cdots, \ell\}$, εμείς έχουμε $I_j = [n_{j-1} + 1, n_j]$ όπου $n_0 = 0$ και $n_\ell = n$. Για να διατηρήσουμε την διαίσθηση της συνεχής περίπτωσης θα θεωρήσουμε ως μήκος του διαστήματος $I_i$, στο εξής συμβολίζεται $l_i$, τον πληθάριθμο του συνόλου $l_i = |I_i|$ [6].

Μπορούμε να υποθέσουμε χωρίς βλάβη της γενικότητα ότι $n, \epsilon^{-1}$ είναι αρκούντως μεγάλα. Τα μήκη των διαστημάτων ορίζονται ως ακολούθως: Έστω ότι $\ell \in \mathbb{Z}^+$ ο μικρότερος ακέραιος αριθμός ώστε $\sum_{i=1}^{\ell} \lfloor (1+\epsilon)^i \rfloor \geq n$. Για $i = 1, 2, \ldots, \ell - 1$ ορίζουμε $l_i := \lfloor (1+\epsilon)^i \rfloor$. την περίπτωση του τελευταίου διαστήματος $\ell-$οστού διαστήματος, θέτουμε $l_\ell := n - \sum_{i=1}^{\ell-1} l_i$.

Από τον παραπάνω ορισμό μπορεί κανείς να δει ότι ο αριθμός $\ell$ των διαστημάτων στην αποσύνθεση είναι το πολύ $O((1/\epsilon) \cdot \log n)$, για την ακρίβεια μετά από πράξεις έχουμε ότι $\ell = O\left((1/\epsilon) \cdot \log(1 + \epsilon \cdot n)\right)$.

Έστω λοιπόν ότι $p$ είναι μια οποιαδήποτε φθίνουσα κατανομή με στήριγμα το $[n]$. Θα δείξουμε ότι υπάρχει μια απλή κλιμακωτή προσέγγιση που ικανοποιεί την βασική ιδιότητα

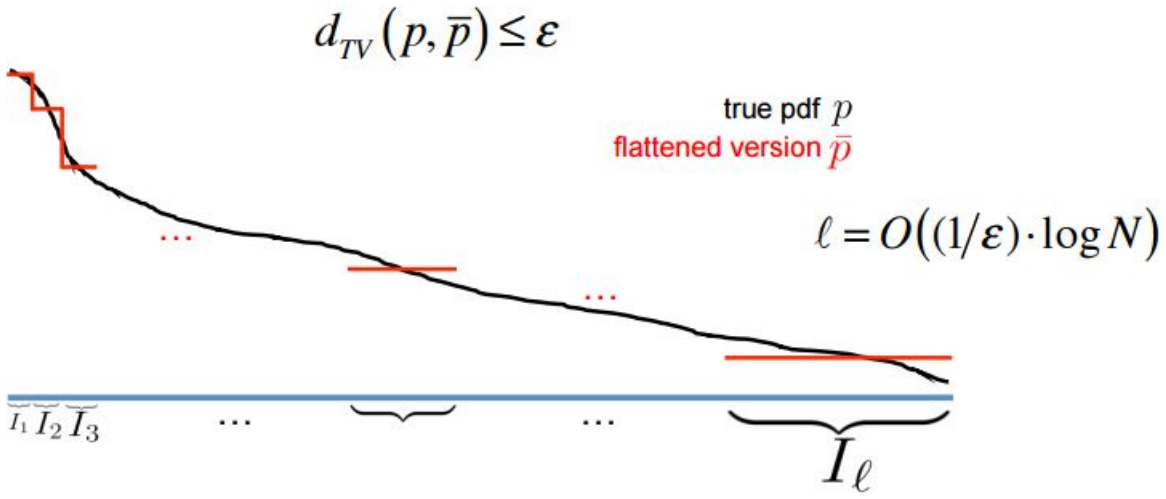
$$d_{\mathrm{TV}}(p_f, p) = O(\epsilon)$$

Συγκεκριμένα ας υποθέσουμε ότι η $p_f$ είναι συνάρτηση σκάλα πάνω $\mathcal{I} = \{I_i\}_{i=1}^{\ell}$. Η $p_f$ ορίζεται ως εξής:

$$p_f(x) = \frac{\sum_{e \in I_j} p(e)}{|I_j|} \forall x \in I_j$$

Παρατηρείστε ότι η $p_f$ απλώς σε κάθε διάστημα της αποσύνθεσης μεσοσταθμίζει ολόκληρο το διάστημα.

[6] Δεδομένου ότι τα διαστήματα είναι ξένα και συνεχόμενα, μπορεί κανείς να ορίσει τα σύνολα έχοντας προσδιορίσει μόνο τα μήκη.




Σχήμα 2.9: Αλγόριθμος μάθησης μονότονων συναρτήσεων

Με βάση αυτόν τον ορισμό έχουμε ότι

$$d_{\mathrm{TV}}(p_f, p) = (1/2) \cdot \sum_{i=1}^{n} |p_f(i) - p(i)| = \sum_{j=1}^{\ell} d_{\mathrm{TV}}\left((p_f)^{I_j}, p^{I_j}\right)$$

όπου $p^I$ ο περιορισμός της κατανομής $p$ πάνω στο $I$.

Έστω $I_j = [n_{j-1}+1, n_j]$ με $l_j = |I_j| = n_j - n_{j-1}$. Τότε έχουμε

$$d_{\mathrm{TV}}\left((p_f)^{I_j}, p^{I_j}\right) = (1/2) \cdot \sum_{i=n_{j-1}+1}^{n_j} |p_f(i) - p(i)| \,.$$

Η κατανομή $p_f$ όπως είδαμε είναι σταθερή σε κάθε διάστημα $I_j$ και συγκεκριμένα

$$\bar{p}_f^j = \sum_{i=n_{j-1}+1}^{n_j} p(i)/l_j$$

Επίσης αφού η $p$ είναι φθίνουσα, τότε

$$p(n_{j-1}) \ge p(n_{j-1}+1) \ge \bar{p}_f^j \ge p(n_j)$$

Συνεπώς μπορεί κανείς να φράξει την TV στο $I_j$ ως ακολούθως:

$$d_{\mathrm{TV}}\left((p_f)^{I_j}, p^{I_j}\right) \le l_j \cdot (p(n_{j-1}+1) - p(n_j)) \le l_j \cdot (p(n_{j-1}) - p(n_j)) \,.$$

Έχουμε λοιπόν:

$$d_{\mathrm{TV}}(p_f, p) \le \sum_{j=1}^{\ell} l_j \cdot (p(n_{j-1}) - p(n_j)) \,. \tag{2.1}$$

Για να φράξουμε την παραπάνω απόσταση θα αναλύσουμε τους προσθετέους σε δύο τμήματα $l_j < 1/\epsilon$ και $l_j \geq 1/\epsilon$ χωριστά.

Θα καλούμε:

- την πρώτη κατηγορία ως *μικρά* διαστήματα
- την δεύτερη κατηγορία ως *μεγάλα* διαστήματα

**Παρατήρηση 2.6.** *Αν υπάρχει ένα διάστημα $I_j$ που ικανοποιεί την ανισότητα $l_j \geq 1/\epsilon$, τότε ας θέσουμε $j_0 \in \mathbb{Z}^+$ να είναι ο μεγαλύτερος ακέραιος έτσι ώστε $l_{j_0} < 1/\epsilon$· Διαφορετικά όλα τα διαστήματα είναι υπερβολικά μικρά και τότε θα θέσουμε ως $j_0 = \ell$. Αν $j_0 < \ell$ τότε έχουμε ότι $j_0 = \Theta((1/\epsilon) \cdot \log_2(1/\epsilon))$. Έστω λοιπόν $S = \{I_i\}_{i=1}^{j_0}$, το σύνολο όλων των μικρών διαστημάτων και έστω το συμπληρωματικό σύνολο όλων των μεγάλων διαστημάτων $L = \mathcal{I} \setminus S$.*

- *Μικρά διαστήματα*

  Ας ομαδοποιήσουμε τα διαστήματα σε *groups* σύμφωνα με το μήκος τους. Κάθε group $G_i$ θα περιλαμβάνει τα διαστήματα του $S$ που έχουν το ίδιο μήκος $i$. Επίσης μπορεί κανείς να δει ότι εξαιτίας της εκθετικής αύξησης που συμβαίνει στο μέγεθος των διαστημάτων, για να έχουν ίδιο μέγεθος σημαίνει ότι η ποσότητα $\lfloor (1+e)^x \rfloor$ δεν έχει ακόμη μεγαλώσει αρκετά. Συνεπώς εύκολα βλέπει κανείς ότι στο $G_i$ ανήκουν σύνολα συνεχόμενα μεταξύ τους. Για να μην έχουμε προβλήματα με τους ορισμούς θα λέμε ότι:

  – Ο πληθάριθμος ενός group ( θα συμβολίζεται ως $|\cdot|$) είναι ο αριθμός των διαστημάτων που περιλαμβάνει

  – Το μήκος ενός group είναι ο αριθμός όλων των στοιχείων που περιέχει ή διαφορετικά το άθροισμα των μηκών των διαστημάτων που περιλαμβάνει

  Ας πάρουμε για παράδειγμα το $G_1$ (το group που περιλαμβάνει μονοσύνολα). Εύκολα βλέπουμε ότι $|G_1| = \Omega(1/\epsilon)$ αφού υποθέσαμε ότι $1/\epsilon < n$. Τότε ξέρουμε ότι $G_1$ έχει μήκος τουλάχιστον $\Omega(1/\epsilon)$. Έστω $j^* < 1/\epsilon$ το μέγιστο μήκος κάποιου παρατηρηθέντος μικρού διαστήματος στο $S$. Είναι εύκολο να παρατηρήσει κανείς ότι οποιοδήποτε group $G_j$ για $j \leq j^*$ επίσης θα είναι μη κενό, και ότι για κάθε $j \leq j^* - 1$, θα έχουμε $|G_j| = \Omega\left((1/\epsilon) \cdot (1/j)\right)$, που σημαίνει ότι $G_j$ είναι της τάξης $\Omega(1/\epsilon)$.

  Για να φράξουμε την συμμετοχή των μικρών διαστημάτων στην (2.1), θα μελετήσουμε την συμβολή του κάθε group. Για παράδειγμα μπορούμε να χρησιμοποιήσουμε το γεγονός ότι το $G_1$ προσφέρει μηδενικό σφάλμα. Συγκεκριμένα για τα υπόλοιπα group έχουμε:

  $$\sum_{l=2}^{j^*} l \cdot \left(p_l^- - p_l^+\right) \tag{2.2}$$

  όπου $p_l^-$ (ρεσπ. $p_l^+$) η πιθανότητα του αριστερότερου και του δεξιότερου στοιχείου στο $G_l$. Δεδομένο ότι η $p$ είναι φθίνουσα, έχουμε ότι $p_l^+ \geq p_{l+1}^-$. Συνεπώς μπορούμε

να φράξουμε από πάνω την ζητούμενη ποσότητα με την (2.2) με

$$2 \cdot p_1^+ + \sum_{l=2}^{j^*-1} p_l^+ - j^* \cdot p_{j^*}^+.$$

Σημειώστε ότι $p_1^+ = O(\epsilon) \cdot p(G_1)$, αφού $G_1$ έχει μήκος (συνολικά στοιχεία πάνω στο στήριγμα) $\Omega(1/\epsilon)$ και $p$ είναι φθίνουσα. Συνεπώς, για $l < j^*$, έχουμε ότι $p_l^+ = O(\epsilon) \cdot p(G_l)$, αφού το μήκος του $G_l$ είναι $\Omega(1/\epsilon)$. Συνεπώς μπορούμε να φράξουμε την παραπάνω ποσότητα με:

$$O(\epsilon) \cdot p(G_1) + O(\epsilon) \cdot \sum_{l=2}^{j^*-1} p(G_l) - j^* \cdot p_{j^*}^+ = O(\epsilon) \cdot p(S) - j^* \cdot p_{j^*}^+. \tag{2.3}$$

- *Μεγάλα διαστήματα*

  Τώρα θα θεωρήσουμε δύο περιπτώσεις:

  - $L = \emptyset$. Σε αυτή την περίπτωση τελειώσαμε αφού η ποσότητα (2.3) είναι της τάξεως $O(\epsilon)$.
  - Έστω $L \neq \emptyset$. Ας σημειώσει κανείς ότι σε αυτή την περίπτωση τα μικρά διαστήματα είναι σε πλήθος $\Omega(1/\epsilon^2)$, που σημαίνει ότι $\epsilon = \Omega(1/\sqrt{n})$. Για την συμμετοχή των μεγάλων διαστημάτων στην TV έχουμε:

$$\sum_{j=j_0+1}^{\ell} l_j \cdot (p(n_{j-1}) - p(n_j)) \leq \begin{cases} l_{j_0} p(n_{j_0}) \\ + \sum_{j=j_0+1}^{\ell-1} (l_{j+1} - l_j) \cdot p(n_j) \\ - p(n_\ell) l_\ell \end{cases} \tag{2.4}$$

  Συνεπώς:

$$\sum_{j=j_0+1}^{\ell} l_j \cdot (p(n_{j-1}) - p(n_j)) \leq\leq \begin{cases} (j^*+1) \cdot p_{j^*}^+ + \\ \sum_{j=j_0+1}^{\ell-1} (l_{j+1} - l_j) \cdot p(n_j) \end{cases} \tag{2.5}$$

Δεδομένου ότι $l_{j+1} - l_j \leq (2\epsilon) \cdot l_j$ και $\sum_j l_j \cdot p(n_j) \leq p(L)$, έπεται ότι ο δεύτερος μεγάλος προσθεταίος (2.5) είναι της τάξεως του $O(\epsilon) \cdot p(L)$. Συνεπώς η TV μεταξύ των $p$ , $p_f$ είναι το πολύ (2.3) + (2.5), δηλαδή

$$O(\epsilon) \cdot p(S) + O(\epsilon) \cdot p(L) + p_{j^*}^+. \tag{2.6}$$

αφού $p(L) + p(S) = 1$ και $p_{j^*}^+ = O(\epsilon)$[7]

Από την τελευταία ισότητα αποδεικνύεται και το ζητούμενο κλείνοντας την απόδειξη. □

[7]Το τελευταίο ισχύει γιατι $p_{j^*}^+$ είναι το δεξιότερο διάστημα εντός του $S$ αφού το $S$ είναι τουλάχιστον μήκους $1/\epsilon$ και η $p$ είναι φθίνουσα.

Μπορεί να δείξει κανείς πλέον επικαλούμενος την παραπάνω απόδειξη ότι αν χρησιμοποιήσει $\ell = O(m^{1/3}\log^{2/3} n)$ διαστήματα, η κλιμακωτή συνάρτηση θα έχει απόσταση της τάξης $O((\log n/m)^{1/3})$. Ας προσπαθήσουμε να μάθουμε τώρα αυτή την κλιμακωτή συνάρτηση στηρίγματος $\ell$ χρησιμοποιώντας τον αλγόριθμο του ιστογράμματος. Από την VC ανισότητα έχουμε ότι $\mathbb{E}[d_{\mathrm{TV}}(p, p_f)] = O(\sqrt{\frac{l}{m}})$. Συνεπώς $\mathbb{E}[d_{\mathrm{TV}}(p, p_f)] = O(\log^{1/3} n/m^{1/3})$ και άρα καταλήγουμε ότι για $m = O(\log n/\epsilon^3)$ έχουμε σφάλμα το πολύ $\epsilon$.

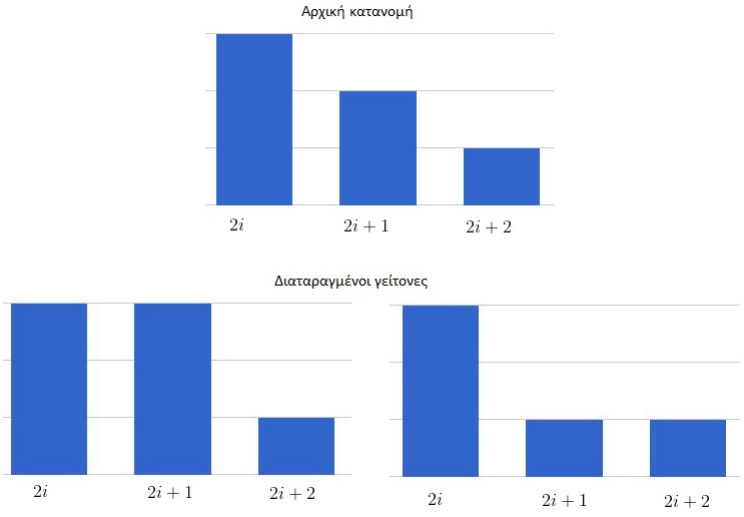
#### 2.3.5.2 Lower Bound

Για την απόδειξη του θεωρήματος αυτού θα χρησιμοποιήσουμε το λήμμα του Assaoud, 2.14. Στο συμβολισμό του λήμματος θα χρησιμοποιήσουμε τις εξής απαιτήσεις/σταθερές:

$$\begin{cases} r = \Theta(\frac{\log n}{\epsilon}) \\ \alpha = \Theta(\epsilon/r) \\ \beta = \Theta(\epsilon^2/r) \end{cases}$$

Για την κατασκευή αυτών των εναλλακτικών συναρτήσεων θα χρησιμοποιήσουμε μια τακτική παρόμοια με αυτή του ιστογράμματος. Θα αποφύγουμε τις τεχνικές λεπτομέρειες παρουσιάζοντας την διαίσθηση πίσω από την απόδειξη και θα παραπέμψουμε τον αναγνώστη στην [Βιρ86, Βιρ87α, Βιρ87β].

Ας δούμε την διαισθητική κατασκευή αυτής της δύσκολης ομάδας κατανομών. Ας πάρουμε μια φθίνουσα κατανομή, για παράδειγμα μια υπερβολή η οποία δειγματοληπτείται πάνω σε ένα μεγάλο διακριτό ακέραιο στήριγμα $[2n]$, για πολύ μεγάλο $n$. Στην συνέχεια, από κάθε διαδοχική τριάδα σημείων $(2i, 2i+1, 2i+2)$, εφαρμόζουμε την εξής τεχνική διαταραχής.




Σχήμα 2.10: Μέθοδος κατασκευής δύσκολα διακρίσιμων μονότονων συναρτήσεων.

Κατασκευάζοντας παρόμοια με την περίπτωση του ιστογράμματος και της μη-παραμετρικής

μάθησης $2^r$ διαφορετικές συναρτήσεις λαμβάνουμε άμεσα από το λήμμα του Assaoud το ζητούμενο κάτω φράγμα.

#### 2.3.5.3 Από Μονότονες σε Unimodal Κατανομές

Για τις τεχνικές λεπτομέρειες ο αναγνώστης παραπέμπεται στο [Bιρ97]. Διαισθητικά έχουμε ότι:

- Στην περίπτωση του Lower Bound εφαρμόζουμε ακριβώς την ίδια τεχνική αριστερόπλευρα και δεξιό-πλευρα του 0 και θα πάρουμε το ίδιο επιθυμητό αποτέλεσμα.
- Στην περίπτωση του Upper Bound μπορούμε πρώτα να εφαρμόσουμε έναν Chernoff εκτιμητή μιας και ξέρουμε ότι υπάρχει μοναδική κορυφή και θα έχουμε το επιθυμητό αποτέλεσμα. Στην συνέχεια μπορούμε να μάθουμε την δεξιά και την αριστερή κατανομή χρησιμοποιώντας τον αλγόριθμο του Birgé. Μια διαφορετική πρόταση βρίσκεται στο [Bιρ97]. Μπορεί κανείς να υπολογίσει δύο συναρτήσεις $d^+, d^-$ το convex minorant και το convex majorant. Οι συναρτήσεις αυτές έχουν την ιδιότητα να μετρούν την καμπυλότητα της συνάρτησης κατανομής. Μπορεί κανείς δειγματοληπτικά να υπολογίσει αυτές τις δύο συναρτήσεις και μέσω δυαδικής αναζήτησης να υπολογίσει το σημείο στο οποίο μεγιστοποιούνται αυτές οι δύο, το οποίο αποτελεί και το mode της κατανομής.

## 2.4 Επίλογος

Ανακεφαλαιωτικά, σε αυτό το κεφάλαιο ο αναγνώστης είχε την ευκαιρία να μελετήσει τα βασικά εργαλεία πιθανοτήτων, τεχνικές εξαγωγής κάτω φραγμάτων και των βασικότερων αλγορίθμων μάθησης του προηγούμενου αιώνα. Το ταξίδι μας θα συνεχίσει με μια πολύ ιδιαίτερη κατηγορία κατανομών, την Poisson Binomial. Η κατανομή αυτή αποτελεί την προφανή γενίκευση πολλών διακριτών κατανομών που χρησιμοποιούνται και έχει την τύχη να έχει όλες τις καλές ιδιότητες που αναφέραμε όπως μονόλοφη, λογαριθμικά κοίλη κ.λ.π. Φεύγοντας κανείς από αυτό το κεφάλαιο σίγουρα θα έπρεπε να κρατήσει στην σκέψη του το βασικά αλγοριθμικά μας ερωτήματα.

- Υπάρχει κάποιο μέτρο πολυπλοκότητας για να υπολογίσουμε ή να φράξουμε την χρονική και την δειγματική πολυπλοκότητα της κάθε κλάσης;
- Ποίο είναι το αντάλλαγμα πληροφορίας που πληρώνουμε κάθε φορά που περιορίζεται το σύνολο των κατανομών στόχος;
- Υπάρχει κάποια σύνδεση μεταξύ των δύο προηγούμενων ερωτήσεων;
- Υπάρχει κάποιος μετρικός μηχανισμός που μπορεί να προσδιορίσει πως αλλάζει η πολυπλοκότητα οποιουδήποτε αλγορίθμου προσπαθεί να μάθει μια κλάση κατανομών όσο αυξάνει αυτή η κλάση;

# Κεφάλαιο 3

# Poisson Binomial Distribution

## 3.1 Εισαγωγή

Όπως αναφέραμε και στο προηγούμενο κεφάλαιο μια από τις πιο βασικές τεχνικές στην κατασκευή εκμάθησης δύσκολων μαθηματικών οντοτήτων είναι η κατασκευή ενός όσο πιο γίνεται αραιού καλύμματος.

Εδώ πρέπει να κάνουμε μια στάση και να σιγουρευτούμε ότι είναι σαφής η έννοια αραιό. Αραιό θεωρείται ένα κάλυμμα όταν ο πληθάριθμος του είναι σχετικά μικρός σε σχέση με το μέγεθος του συνόλου που καλύπτει. Αυτό δεν σημαίνει ότι το κάλυμμα δεν είναι πυκνό. Χαρακτηριστικό παράδειγμα ύπαρξης τέτοιου συνόλου είναι οι ρητοί αριθμοί. Ως σύνολο είναι πυκνό, δηλαδή για οποιοδήποτε $x \in \mathbb{R}, \epsilon > 0$, σίγουρα $\exists q \in \mathbb{Q} : q \in (x-\epsilon, x+\epsilon)$. Ο λόγος του πληθάριθμου όμως των ρητών ως προς τους πραγματικούς ορίζει ακριβώς την διαφορά μεταξύ του αριθμήσιμου από του υπεραριθμήσιμου απείρου $(\aleph_0, \aleph_1)$.

Αν και η Εκτιμητική Στατιστική έχει αναδείξει πολλά αποτελέσματα στο χώρο της ασυμπτωτικής ανάλυσης, σε ένα χώρο που θα φαινόταν ιεραρχικά πιο κύριος, στον χώρο των πολυπαραμετρικών κατανομών η μαθηματική έρευνα είχε μείνει εξαιρετικά πίσω για αρκετά χρόνια, ειδικά για την οικογένεια κατανομών που θα μελετήσουμε σε αυτό το κεφάλαιο.

### 3.1.1 Poisson Binomial Distribution-Ορισμοί

**Ορισμός 3.1.** ***Poisson Binomial Distribution*** *Μια Poisson Binomial Distribution τάξης $n$ στον $\mathbb{N}$ ορίζεται η διακριτή κατανομή του αθροίσματος $\sum_{i=1}^{n} X_i$ των $n$ αμοιβαία ανεξάρτητων Bernoulli τυχαίων μεταβλητών $X_1, X_2, \cdots, X_n$.*

**Ορισμός 3.2.** *Το σύνολο όλων των PBD της τάξης $n$ θα το συμβολίζουμε με $S_n$.*

**Παρατηρήσεις:**

1. Η $S_n$ αποτελεί πολυπαραμετρική γενίκευση της Binomial Distribution , $\mathcal{B}(n,p) = \sum_{i=1}^{n} X_i$ , όπου τα $\mathbb{E}[X_i] = p \ \forall i \in [n]$, έχουν την ίδια παράμετρο $p$.

2. Επίσης, όταν στο κείμενο θα αναφέρεται ότι μια κατανομή $D \in S_n$, αυτό σημαίνει ότι υπάρχει ένα διάνυσμα από $(p_i)_1^n \in [0,1]^n$ που θα αντιστοιχεί στις παραμέτρους των Bernoulli δείκτριων $X_i$ , δηλαδή $\mathbb{E}[X_i] = p_i$ και $\sum_{i=1}^n X_i$ είναι τυχαία μεταβλητή που ακολουθεί την κατανομή $D$.

Ένα ενδιαφέρον ερώτημα που πρέπει να απαντήσει κανείς όταν θέλει να υπολογίσει το κάλυμμα μιας οικογένειας κατανομών είναι αν οι παράμετροι προσδιορίζουν μοναδικά την κατανομή. Σε διαφορετική περίπτωση κινδυνεύει κανείς ο αλγόριθμος του να επεξεργάζεται περισσότερες κατανομές από τις πραγματικά διαφορετικές.

Για αυτόν τον λόγο πριν προχωρήσουμε παρακάτω, θα παραθέσουμε ένα πολύ σημαντικό λήμμα που αποδεικνύει την παραπάνω απαίτηση για την κλάση των PBDs .

**Λήμμα 3.1.** *Αν ταξινομήσουμε αριθμητικά τις παραμέτρους σε αύξουσα σειρά , δηλαδή $p_1 \leq p_2 \leq p_3 \cdots \leq p_n$ και εξαφανίσουμε τις συγκρούσεις στις ισότητες με λεξικογραφική σειρά τότε η κατανομή που προκύπτει έχει μοναδική περιγραφή. Δηλαδή αν*

$$Αν: \begin{cases} X, Y \in S_n \\ X = \sum_{i=1}^n X_i \\ Y = \sum_{i=1}^n Y_i \\ \mathbb{E}[X_i] = p_i \\ \mathbb{E}[Y_i] = q_i \end{cases} \quad \text{τότε :} X \equiv Y \Leftrightarrow (p_1, \ldots, p_n) = (q_1, \ldots, q_n)$$

*όπου $X \equiv Y$ σημαίνει ότι οι δύο τυχαίες μεταβλητές ακολουθούν την ίδια ακριβώς κατανομή.*

*Απόδειξη.* Είναι προφανές ότι:
Αν $(p_1, \ldots, p_n) = (q_1, \ldots, q_n)$, τότε $X \equiv Y$, αφού είναι οι ίδιες.
Θα αποδείξουμε την αντίθετη κατεύθυνση:
Αν $X \equiv Y$ τότε $(p_1, \ldots, p_n) = (q_1, \ldots, q_n)$.

Θεωρείστε τα πολυώνυμα

$$g_X(s) = \mathbb{E}\left[(1+s)^X\right] = \prod_{i=1}^n \mathbb{E}\left[(1+s)^{X_i}\right] = \prod_{i=1}^n (1 + p_i s);$$
$$g_Y(s) = \mathbb{E}\left[(1+s)^Y\right] = \prod_{i=1}^n \mathbb{E}\left[(1+s)^{Y_i}\right] = \prod_{i=1}^n (1 + q_i s).$$

Αφού $X, Y$ ακολουθούν την ίδια κατανομή αναγκαστικά τα δύο πολυώνυμα $g_X, g_Y$ είναι ίσα, συνεπώς έχουν και τον ίδιο βαθμό και τις ίδιες ρίζες. Παρατηρείστε ότι:

- $g_X$ έχει βαθμό $n - |\{i \mid p_i = 0\}|$ και ρίζες $\{-\frac{1}{p_i} \mid p_i \neq 0\}$.
- $g_Y$ έχει βαθμό $n - |\{i \mid q_i = 0\}|$ και ρίζες $\{-\frac{1}{q_i} \mid q_i \neq 0\}$.

Συνεπώς αναγκαστικά $(p_1, \ldots, p_n) = (q_1, \ldots, q_n)$. □

Έτσι κλείνει το πρώτο μας βήμα για την κατανόηση των PBDs .

### 3.1.2 Η προϊστορία αυτού του καλύμματος

Την μέθοδο και την κατασκευή του καλύμματος που θα περιγράψουμε σε αυτό το κεφάλαιο την οφείλουμε κατα κύριο λόγω στους Παπαδημητρίου και Δασκαλάκη [ΔΠ13] και πιο συγκεκριμένα σε μια σειρά από δημοσιεύσεις τους στο κλάδο των "Ανώνυμων Παιχνιδιών" στο πεδίο της Αλγοριθμικής Θεωρίας Παιγνίων [ΔΠ15, Δασ08, ΔΠ07, ΔΠ11].

Ο στόχος της μεθοδολογίας, δηλαδή η κατασκευή ενός αραιού καλύμματος κατανομών για Παιγνιοθεωρητικούς λόγους, είχε επαναληφθεί στο παρελθόν στους Lipton, Markakis, Mehta με το διάσημο [ΛΜΜ03] και πολύ παλαιότερα με το διάσημο για την εποχή του paper του Althofer[Αλτ94]. Στόχος όμως τότε αποτελούσε το γενικότερο πρόβλημα του περιορισμού των περίπλοκων στρατηγικών που πιθανώς να είναι βέλτιστες, ως προς την Ισορροπία Nash και πιο συγκεκριμένα αν είναι εφικτό απομακρυνόμενος $\epsilon$ από την βέλτιστη λύση να μπορεί κανείς να έχει μαι πολύ πιο μικρή στρατηγική σε πραγματικά περίπλοκα settings.

Στην μελέτη αυτών των δύο σημαντικών έργων η βασική παρατήρηση είναι το sampling. Όμως και στις δύο περιπτώσεις οι δημιουργοί τους δείχνουν ότι για ένα αρκετά μικρότερο πλήθος δειγμάτων από το αναμενόμενο το ιστόγραμμα, η εμπειρική κατανομή μπορεί να αποτελέσει πρόκληση και απάντηση της κρυφής κατανομής.

Η ιδέα του Παπαδημητρίου-Δασκαλάκη βρίσκεται στην τομή της τεχνικής του καλύμματος και των προηγούμενων. Ας δούμε όμως πρώτα την αφετηρία του μοντέλου τους. Στο μοντέλο του Παπαδημητρίου-Δασκαλάκη θεωρούμε το ακόλουθο 2-στρατηγικών ανώνυμο παίγνιο $n$ παικτών με ακριβώς τις δύο ίδιες επιλογές για στρατηγικές[1] $0, 1$ για κάθε παίκτη.

Το πρόβλημα χαρακτηρίζεται ανώνυμο γιατί αν και η συνάρτηση χρησιμότητας του παίκτη $i$ χαρακτηρίζεται από τον τρόπο με τον οποίο παίζουν οι υπόλοιποι παίκτες στον τελικό απολογισμό αποτελεί μια συνάρτηση $u_i : \{0,1\} \times [n-1] \to [0,1]$. Αυτό σημαίνει ότι δεν εξαρτάται από το ποιος παίκτης παίζει την εκάστοτε στρατηγική αλλά από τα στατιστικά στοιχεία των αντιπάλων, από τιμές δηλαδή που έχουν εφαρμόσει κάποια συνάθροιση κάποια ομαδοποίηση της πληροφορίας, μετατρέποντας την σε ανώνυμη.

Συγκεκριμένα, στο μοντέλο τους, η $u_i$ εξαρτάται από την στρατηγική του παίκτη $i'$ και τον αριθμό των παικτών που θα παίξουν επίσης την στρατηγική 1.

Για να εφαρμόσουμε το υπαρξιακό θεώρημα του Nash θα χρησιμοποιήσουμε τυχαιοποιημένες στρατηγικές, δηλαδή την στρατηγική του παίκτη $i$ να την συμβολίσουμε με μια δείκτρια τυχαία μεταβλητή, μια Bernoulli $X_i$, η οποία τυχαιοποιεί την επιλογή του παίκτη $i$ να παίξει την στρατηγική 1.

Μόλις τυχαιοποίησει κανείς το portfolio των στρατηγικών, από μια αναπαράσταση συγκεκριμένων επιλογών, μετατρέπεται σε μια κατανομή και η στρατηγική που θα επιλέξει εκείνος είναι μια τυχαία κατανομή που προσπαθεί να είναι βέλτιστη ως προς ένα σύνολο άλλων κατανομών. Επίσης παρατηρείστε ότι λόγω του μοντέλου δεν είναι στόχος μας να είμαστε βέλτιστοι σε ένα διάνυσμα άλλων κατανομών αλλά σε μια κατανομή μίας μεταβλητής, αυτής του αθροίσματος των μεταβλητών[2]

[1] Π.χ Πάω/Δεν Πάω στο Συμπόσιο , Αγοράζω/ Δεν αγοράζω το εισιτήριο για την συναυλία

[2] Για αυτούς που είναι φιλικοί στην διαχείριση πράξεων μεταξύ κατανομών, ο στόχος του μοντέλου είναι να βρεί μια βέλτιστη κατανομή στρατηγικής απέναντι στην συνέλιξη των στρατηγικών των υπολοίπων

Η σπουδαία παρατήρηση που οδήγησε στην κατασκευή αυτών των καλυμμάτων είναι ότι αν η $X_i$ είναι βέλτιστη στρατηγική σε σχέση με την $X_{-i}$, τότε είναι σχεδόν βέλτιστη σε κάθε $\epsilon$-στρατηγική $X'_{-i}$. Εδώ ακριβώς έρχεται και η σπουδαία ιδέα του καλύμματος.

Ας υποθέσουμε λοιπόν ότι ο χώρος των κατανομών έχουν την όμορφη ιδιότητα να μπορούν να σπάσουν σε πλακάκια, σε σφαίρες και από κάθε μπάλα υπάρχει ένας εκπρόσωπος αυτής της περιοχής με την εξής ιδιότητα.

Η τιμή της συνάρτησης χρησιμότητας σε αυτό το κέντρο της μπάλας είναι αρκετά κοντά στη τιμή της συνάρτησης χρησιμότητας σε κάθε άλλο σημείο στην μπάλα.

Αυτό σημαίνει ότι αν ο χώρος έχει λίγα κέντρα δηλαδή μπορεί να διασπαστεί σε λίγες αλλά ουσιαστικές σφαίρες τότε μπορεί κανείς να αποσύρει το δύσκολο αίτημα υπολογισμού της βέλτιστης τιμής της συνάρτησης και να υπολογίσει αποδοτικά μια $\epsilon$-προσέγγιση βρίσκοντας την βέλτιστη τιμή μεταξύ των κέντρων.

Ακριβώς αυτή η ιδέα οδήγησε στην κατασκευή ενός πολυωνυμικού καλύμματος πάνω στο σύνολο των $X_{-i}$, δηλαδή του χώρου των PBDs $(n-1)$ μεταβλητών.

### 3.1.3 Χρήσιμες ανισότητες-προσεγγίσεις μίας PBD

Για την κατασκευή ενός τέτοιο καλύμματος, το βασικότερο εργαλείο μας είναι ανισότητες προσέγγισης. Με τον όρο αυτό αναφερόμαστε στην βιβλιογραφία τις ανισότητες που φράσσουν από τα πάνω την απόσταση δύο κατανομών. Αν κάτω από κάποιον μετρικό χώρο δύο κατανομές είναι κοντά δηλαδή οι κατανομές τους εμφανίζουν μικρή απόσταση ή η απόσταση αυτών των κατανομών φράσσεται από ένα εξαιρετικά μικρό αριθμό τότε μπορούμε να χρησιμοποιήσουμε την μια ως προσέγγιση της άλλης.

Σε όλες τις ακόλουθες ανισότητες θα υποθέτουμε μια PBD ως $\mathbb{V} = \sum_{i=1}^{n} X_i, \mathbb{E}[X_i] = p_i$ και θα υποθέτουμε ότι ο συμβολισμός -αν δεν οριστεί κάτι άλλο- αντιστοιχεί σε

$$\begin{cases} \mathbb{E}[\mathbb{V}] = \mu \\ Var[\mathbb{V}] = \sigma^2 \end{cases}$$

#### 3.1.3.1 Οι ανισότητες

- **Berry-Essen**[Βερ41]

  Ας είμαστε ειλικρινείς. Η πρώτη προσέγγιση που θα ακολουθούσε κανείς για να προσεγγίσει μια PBD είναι το κεντρικό οριακό θεώρημα. Η PBD παραμένει ένα ανεξάρτητο άθροισμα ανεξάρτητων κατανομών όχι ισόνομων αλλά το κλασσικό εργαλείο του Berry-Essen, από το οποίο αποδείχτηκε για πρώτη φορά το κεντρικό οριακό θεώρημα παραμένει μια πρώτη απάντηση.

---

παικτών.

Η προσέγγιση του γνωστού και ως θέωρημα Berry-Essen μας δίνει την εξής πολύ βασική ανισότητα:

$$d_K(\sum_i X_i, \mathcal{N}(\mu, \sigma^2)) \leq C\frac{\sum \mathbb{E}[|X_i|^3]}{\sigma^3}$$

Μάλιστα γίνεται μελέτη ακόμα στην μαθηματική κοινότητα για να προσδιοριστεί με ακρίβεια η σταθερά η οποία αυτή την στιγμή έχει αποδειχθεί ότι κυμαίνεται μεταξύ του $0.4 \leq C \leq 0.52$. Αν εξειδικεύσουμε τον υπολογισμό αυτής της ανισότητας στις PBDs έχουμε:

$$d_K(\sum_i X_i, \mathcal{N}(\mu, \sigma^2)) \leq C\frac{\mu}{\sigma^3}$$

Ας εξετάσουμε την απλούστερη δυνατή περίπτωση της Binomial κατανομής που βρίσκεται στο σύνολο των PBDs .

$$d_K(Binom(n,p), \mathcal{N}(np, np(1-p))) \leq C\frac{1}{\sqrt{np}(1-p)^{1.5}}$$

Σε αυτή την περίπτωση όμως είναι εμφανές ότι αν $p = \frac{1}{n}$ έχουμε ένα εξαιρετικά άσχημο φράγμα $O(\frac{1}{\sqrt[2]{(1-\frac{1}{n})^3}})$, το οποίο εξαρτάται με εξαιρετικά άσχημο τρόπο από το πλήθος των μεταβλητών που αθροίζεται.

Εδώ εμφανίζεται και η πρώτη βασική ιδιότητα που πρέπει να αποκτήσουμε για να κατασκευάσουμε ένα κάλυμμα τέτοιου είδους για μια πολυπαραμετρική κατανομή. Οι κατανομές που θα προσπαθούν να την προσεγγίσουν θα πρέπει να μην συνδέονται ισχυρά με το πλήθος των παραμέτρων που χαρακτηρίζουν την περίπλοκη κατανομή μας.

- **Poisson Approximation**[ΒΗ84, ΒΗΘ92]

Ας δοκιμάσουμε να εφαρμόσουμε την προσέγγιση που χρησιμοποιούμε συνήθως στην περίπτωση όπου $n = \frac{1}{p}$, την προσέγγιση της Poisson όπου έχει μέση τιμή ίδια με την PBD που εξετάζουμε $\lambda_{Poisson} = \mathbb{E}[\mathbb{V}] = \mu$.

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

Η ανισότητα αυτή οφείλεται σε έναν πολύ διάσημο μαθηματικό που αναφέραμε και στο προηγούμενο κεφάλαιο τον Le Cam. Όπως αναμένεται όταν $p = \frac{1}{n}$ η προσέγγιση αυτή είναι αρκετά καλή $O(1/n)$.

- **Binomial Approximation**[Εημ91]
Σε αυτή την περίπτωση θα προσπαθήσουμε να εφαρμόσουμε την τεχνική του Berry-Essen και της Poisson Approximation για να έχουμε μια πιο tight προσέγγιση. Ας

υποθέσουμε ότι $\bar{\mu} = \frac{\sum_{i=1}^{n} p_i}{n}$ τότε έχουμε:

$$d_{\mathrm{TV}}(\sum_{i=1}^{n} X_i, \mathcal{B}inom\,(n, \bar{\mu})) \leq \frac{\sum_{i=1}^{n}(p_i - \bar{\mu})^2}{(n+1)\bar{\mu}(1-\bar{\mu})}$$

Εμπνεόμενοι από το αποτέλεσμα του Berry-Essen εμφανίστηκαν ακόμα δύο βασικά εργαλεία τα οποία μας δείχνουν την σχέση δύο PBDs που έχουν πολύ μικρή διακύμανση.

- **Rounded Normal**[ΓΣ10]
  Όπως βλέπετε στην περίπτωση του Berry το φράγμα μεταξύ των δύο κατανομών χρησιμοποιεί την απόσταση Kolmogorov γιατί έχουν διαφορετικό τελείως πεδίο ορισμού, αφού η PBD είναι διακριτή ενώ η κανονική κατανομή είναι συνεχής κατανομή. Αν προσπαθήσουμε να εφαρμόσουμε την μέθοδο του Chen-Stein για την εξαγωγή του δεδομένου αποτελέσματος αλλά αντικαταστήσουμε την κανονική κατανομή με την στρογγυλοποίηση της μπορεί κανείς να αποκτήσει το ακόλουθο bound.

$$d_{\mathrm{TV}}(\sum_{i=1}^{n} X_i, Z\left(\mu, \sigma^2\right)) \leq \frac{\sum_{i=1}^{n}(p_i - \bar{\mu})^2}{(n+1)\bar{\mu}(1-\bar{\mu})}$$

- **Translated Poisson Bounds**[Pö07]
  Σε αυτό το σημείο θα χρησιμοποιήσουμε μια ακόμα πιο ενδιαφέρουσα προσέγγιση χρησιμοποιώντας μια Translated Poisson κατανομή.

  **Ορισμός 3.3** (**Translated Poisson**.). *Μια ακέραια τυχαία μεταβλητή $Y$ με παραμέτρους $\mu, \sigma^2$ καλείται Translated Poisson και συμβολίζεται $TP(\mu, \sigma^2)$ αν και μόνο αν $(Y - \lfloor \mu - \sigma^2 \rfloor) = \mathrm{Poisson}(\sigma^2 + \{\mu - \sigma^2\})$, όπου $\{\mu - \sigma^2\}$ αποτελεί αναπαράσταση για το δεκαδικό μέρος του $\mu - \sigma^2$.*

  Ο Adrian Röllin το 2007 έδειξε ότι:

$$d_{\mathrm{TV}}(\sum_{i=1}^{n} X_i, TP(\mu, \sigma^2)) \leq \frac{\sqrt{\sum_{i=1}^{n} p_i^3(1-p_i)} + 2}{\sum_{i=1}^{n} p_i(1-p_i)} \Rightarrow$$

$$d_{\mathrm{TV}}(\sum_{i=1}^{n} X_i, TP(\mu, \sigma^2)) \leq \frac{1}{\sigma} + \frac{2}{\sigma^2}$$

και παράλληλα ότι :

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

#### 3.1.3.2 Επιμύθιο πίσω από τις ανισότητες

Παρατηρείστε ότι από τις παραπάνω ανισότητες οδηγούμαστε στο έξης μεγάλο επιμύθιο:

1. Υπάρχουν καλές ανισότητες με απροσδιόριστα μικρό $\epsilon$;
2. Υπάρχουν ανισότητες μεταξύ PBDs ώστε να έχουμε proper προσεγγίσεις;
3. Ποία είναι η απόσταση δύο PBDs όταν έχουν ίδιες τις πρώτες δύο ροπές τους;

#### 3.1.3.3 Αν είναι ίδιες οι πρώτες ροπές, πόσο διαφέρουν οι επόμενες;

Θα δώσουμε από τώρα την απάντηση στο τελευταίο ερώτημα. Για να απαντήσουμε αυτό το ερώτημα θα επικαλεστούμε μια από τις ισχυρές μαθηματικές προσεγγίσεις των PBDs από τον Roos.

**Θεώρημα 3.1** ([Poo00]). *Έστω $\mathcal{P} := (p_i)_{i=1}^n \in [0,1]^n$, $X_1, \ldots, X_n$ ανεξάρτητες δείκτριες μεταβλητές με μέσες τιμές $p_1, \ldots, p_n$ και $X = \sum_i X_i$. Τότε $\forall m \in \{0, \ldots, n\}$ και $p \in [0,1]$ έχουμε:*

$$\Pr[X = m] = \sum_{\ell=0}^{n} \alpha_\ell(\mathcal{P}, p) \cdot \delta^\ell \mathcal{B}_{n,p}(m)$$

*όπου*

- $\alpha_0(\mathcal{P}, p) := 1$ *και για* $\ell \in [n]$ :

$$\alpha_\ell(\mathcal{P}, p) := \sum_{1 \le k(1) < \ldots < k(\ell) \le n} \prod_{r=1}^{\ell} (p_{k(r)} - p);$$

- $\forall \ell \in \{0, \ldots, n\}$ :

$$\delta^\ell \mathcal{B}_{n,p}(m) := \frac{(n-\ell)!}{n!} \frac{d^\ell}{dp^\ell} \mathcal{B}_{n,p}(m),$$

*όπου στο τελευταίο ορισμό θεωρούμε το $\mathcal{B}_{n,p}(m) \equiv \binom{n}{m} p^m (1-p)^{n-m}$ ως μαθηματική συνάρτηση του $p$.*

Μέσω αυτού του θεωρήματος μπορεί κανείς να βρει πολύ στενότερες προσεγγίσεις για την κλάση των Poisson Binomial απλά ρυθμίζοντας κατάλληλα των αριθμό των όρων που θα επιλέξουμε να διατηρήσουμε από το κράτημα. Η ακόλουθη σημαντική πρόταση βρίσκεται στην απόδειξη του [;], φράσσει το προσεγγιστικό σφάλμα χρησιμοποιώντας μόλις τους πρώτους $d+1$ όρους. Το σφάλμα πέφτει εκθετικά στο $d$ όσο η ποσότητα $\theta(\mathcal{P}, p)$ είναι μικρότερη της μονάδας.

**Λήμμα 3.2** ([Poo00]). *Έστω $\mathcal{P} = (p_i)_{i=1}^n \in [0,1]^n$, $p \in [0,1]$, $\alpha_\ell(\cdot,\cdot)$ και $\delta^\ell \mathcal{B}_{n,p}(\cdot)$ όπως στο θεώρημα. Τότε :*

$$\theta(\mathcal{P},p) = \frac{2\sum_{i=1}^n (p_i-p)^2 + (\sum_{i=1}^n (p_i-p))^2}{2np(1-p)}.$$

*Αν $\theta(\mathcal{P},p) < 1$, τότε για όλα τα $\forall d \geq 0$:*

$$\sum_{\ell=d+1}^{n} |\alpha_\ell(\mathcal{P},p)| \cdot \|\delta^\ell \mathcal{B}_{n,p}(\cdot)\|_1 \leq \sqrt{e}(d+1)^{1/4}\theta(\mathcal{P},p)^{(d+1)/2}\frac{1-\frac{d}{d+1}\sqrt{\theta(\mathcal{P},p)}}{(1-\sqrt{\theta(\mathcal{P},p)})^2},$$

*όπου $\|\delta^\ell \mathcal{B}_{n,p}(\cdot)\|_1 := \sum_{m=0}^n |\delta^\ell \mathcal{B}_{n,p}(m)|$.*

Ας χρησιμοποιήσουμε τα παραπάνω μαθηματικά αποτελέσματα ώστε να εξάγουμε ένα συμπέρασμα για το ερώτημα που θέσαμε.

**Θεώρημα 3.2.** *Έστω $\mathcal{P} := (p_i)_{i=1}^n$ και $\mathcal{Q} := (q_i)_{i=1}^n$ δύο $n$-άδες που ανήκουν στο $[0,1/2]^n$. Επίσης έστω δύο συλλογές τυχαίων μεταβλητών $\mathcal{X} := (X_i)_{i=1}^n$ , $\mathcal{Y} := (Y_i)_{i=1}^n$ όπου $\mathbb{E}[X_i] = p_i$ και $\mathbb{E}[Y_i] = q_i \forall i \in [n]$. Αν για κάποια $d \in [n]$ ισχύει ότι*

$$(C_d): \quad \sum_{i=1}^n p_i^\ell = \sum_{i=1}^n q_i^\ell, \quad \forall \ell \in \{1,\ldots,d\}$$

*Τότε:*

$$d_{\mathrm{TV}}(\sum_i X_i, \sum_i Y_i) \leq 13(d+1)^{1/4}2^{-(d+1)/2}$$

**Παρατήρηση 3.1.** *Η συνθήκη $(C_d)$ περιορίζει τις πρώτες $d$ δυνάμεις του αθροίσματος των αναμενόμενων τιμών των δεικτριών που εμφανίζονται εντός μιας PBD . Για να συνδέσει κανείς αυτό το αποτέλεσμα με τις ροπές μιας PBD θα πρέπει να χρησιμοποιήσει την θεωρία των συμμετρικών πολυωνύμων, μια εξαιρετικά ισχυρή ομάδα πολυωνύμων, στην οποία ανήκει η κλάση που μελετάμε, μιας και με όποιον τρόπο κι αν αλλάξουμε την θέση των μεταβλητών/παικτών το αποτέλεσμα θα παραμένει το ίδιο.*

*Συγκεκριμένα μπορεί να δείξει κανείς ότι:*

$$(V_d): \quad \mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^\ell\right] = \mathbb{E}\left[\left(\sum_{i=1}^n Y_i\right)^\ell\right], \ \forall \ell \in [d].$$

*με την προηγούμενη ιδιότητα $C_d$ είναι ισοδύναμες*

$$(C_d) \Leftrightarrow (V_d)$$

**Παρατήρηση 3.2** ( *Τι λέει λοιπόν το προηγούμενο θεώρημα ;)*.

1. "Αν δύο αθροίσματα $n$ ανεξάρτητων δεικτριών με μέσες τιμές στο $[0, 1/2]$ έχουν ίσες τις πρώτες $d$ ροπές, τότε η TV τους είναι $2^{-\Omega(d)}$."

2. *Το εξαιρετικά σημαντικό στο προηγούμενο αποτέλεσμα είναι ότι το τελευταίο αποτέλεσμα είναι ότι είναι πλήρως ανεξάρτητο των του αριθμού των μεταβλητών $n$, και δεν στηρίζεται σε κάποιο άθροισμα κάποιων γιγαντιαίου αριθμού μεταβλητών. Επίσης εν γένει δεν εμφανίζεται κανένας περιορισμός ούτε στην μέση τιμή ούτε στην διακύμανση της τελικής PBD κάνοντας το παραπάνω αποτέλεσμα των Δασκαλάκη και Παπαδημητρίου εξαιρετικά πιο χρήσιμο από αυτό του Berry-Essen. Τέλος αξίζει να σημειωθεί ότι η απόδειξη παραμένει ίδια και στην κατοπτρική περίπτωση του* $[1/2, 1]$

Ας δούμε τώρα την απόδειξη του σημαντικού αυτού θεωρήματος.

*Απόδειξη.* Έστω $\mathcal{X}$ και $\mathcal{Y}$ δύο $n-$άδες τυχαίων μεταβλητών όπως ορίζει το θεώρημα Για τα $\alpha_\ell(\cdot,\cdot)$ όπως όρισε ο Roos στο [Poo00] έχουμε ότι :

**Λήμμα 3.3.** *Αν $\mathcal{P}, \mathcal{Q} \in [0,1]^n$ ικανοποιούν ως συλλογές την $(C_d)$ τότε για όλα τα $p$, $\ell \in \{0, \ldots, d\}$ :*

$$\alpha_\ell(\mathcal{P}, p) = \alpha_\ell(\mathcal{Q}, p).$$

*Απόδειξη.* Αρχικά $\alpha_0(\mathcal{P}, p) = 1 = \alpha_0(\mathcal{Q}, p)$ εξ' ορισμού. Τώρα ας πάρουμε ένα σταθερό $\ell \in \{1, \ldots, d\}$ και έστω η συνάρτηση $f(\vec{x}) := \alpha_\ell((x_1, \ldots, x_n), p)$ με μεταβλητές τιες $x_1, \ldots, x_n \in \mathbb{R}$. Εύκολα βλέπει κανείς ότι η $f$ είναι συμμετρικό πολυώνυμο βαθμού $\ell$ στις $x_1, \ldots, x_n$. Συνεπώς από την θεωρία των συμμετρικών πολυωνύμων έχουμε ότι η $f$ μπορεί να γραφτεί ως πολυωνυμική συνάρτηση των δυναμοσειρών-συμμετρικών πολυωνύμων $\pi_1, \ldots, \pi_\ell$, όπου

$$\pi_j(x_1, \ldots, x_n) := \sum_{i=1}^{n} x_i^j, \text{ για κάθε } j \in [\ell],$$

αφού τα στοιχειώδη συμμετρικά πολυώνυμα τάξης $j \in [n]$ μπορούν να γραφτούν ως πολυωνυμική συνάρτηστη των δυναμοσειρών-συμμετρικών πολυωνύμων $\pi_1, \ldots, \pi_j$ (βλέπε [Ζολ87]).[3] Όμως τώρα η $(C_d)$ συνεπάγει ότι $\pi_j(\mathcal{P}) = \pi_j(\mathcal{Q})$, για κάθε $j \leq \ell$. Συνεπώς η $f(\mathcal{P}) = f(\mathcal{Q})$, δηλαδή $\alpha_\ell(\mathcal{P}, p) = \alpha_\ell(\mathcal{Q}, p)$. □

[3] Αν υποθέσουμε ότι έχουμε μια επιλογή από $n$ μεταβλητές $X_1, \cdots, X_n$ είναι τα εξής:

$$e_0(X_1, X_2, \ldots, X_n) = 1, \tag{3.1}$$
$$e_1(X_1, X_2, \ldots, X_n) = \textstyle\sum_{1 \leq j \leq n} X_j, \tag{3.2}$$
$$e_2(X_1, X_2, \ldots, X_n) = \textstyle\sum_{1 \leq j < k \leq n} X_j X_k, \tag{3.3}$$
$$e_3(X_1, X_2, \ldots, X_n) = \textstyle\sum_{1 \leq j < k < l \leq n} X_j X_k X_l, \tag{3.4}$$
$$e_k(X_1, \ldots, X_n) = \sum_{1 \leq j_1 < j_2 < \cdots < j_k \leq n} X_{j_1} \cdots X_{j_k}, \tag{3.5}$$
$$e_n(X_1, X_2, \ldots, X_n) = X_1 X_2 \cdots X_n \tag{3.6}$$

και αποτελούν την βάση του χώρου όλων των συμμετρικών πολυωνύμων $n$ μεταβλητών.

Τώρα για όλα τα $p \in [0,1]$, συνδιάζοντας το θεώρημα του [Poo00] και το προηγούμενο λήμμα έχουμε ότι

$$\Pr[X = m] - \Pr[Y = m] = \sum_{\ell=d+1}^{n} (\alpha_\ell(\mathcal{P}, p) - \alpha_\ell(\mathcal{Q}, p)) \cdot \delta^\ell \mathcal{B}_{n,p}(m), \forall m \in \{0, \ldots, n\}.$$

Συνεπώς $p$:

$$\begin{aligned} d_{\mathrm{TV}}(X, Y) &= \frac{1}{2} \sum_{m=0}^{n} |\Pr[X = m] - \Pr[Y = m]| \\ &\leq \frac{1}{2} \sum_{\ell=d+1}^{n} |\alpha_\ell(\mathcal{P}, p) - \alpha_\ell(\mathcal{Q}, p)| \cdot \|\delta^\ell \mathcal{B}_{n,p}(\cdot)\|_1 \\ &\leq \frac{1}{2} \sum_{\ell=d+1}^{n} (|\alpha_\ell(\mathcal{P}, p)| + |\alpha_\ell(\mathcal{Q}, p)|) \cdot \|\delta^\ell \mathcal{B}_{n,p}(\cdot)\|_1. \end{aligned} \tag{3.7}$$

Εφαρμόζοντας για $p = \bar{p} := \frac{1}{n}\sum_i p_i$ στο λήμμα που υπάρχει στην απόδειξη του Roos στο [Poo00] έχουμε ότι

$$\theta(\mathcal{P}, \bar{p}) = \frac{\sum_{i=1}^{n} (p_i - \bar{p})^2}{n\bar{p}(1-\bar{p})} \leq \left| \max_i \{p_i\} - \min_i \{p_i\} \right| \leq \frac{1}{2} \qquad (\text{βλέπε [Poo00]})$$

και τότε

$$\begin{aligned} \frac{1}{2} \sum_{\ell=d+1}^{n} |\alpha_\ell(\mathcal{P}, \bar{p})| \cdot \|\delta^\ell \mathcal{B}_{n,\bar{p}}(\cdot)\|_1 &\leq \sqrt{e}(d+1)^{1/4} 2^{-(d+1)/2} \frac{1 - \frac{1}{\sqrt{2}} \frac{d}{d+1}}{(\sqrt{2}-1)^2} \\ &\leq 6.5(d+1)^{1/4} 2^{-(d+1)/2}. \end{aligned}$$

Αλλά από την $(C_d)$ έπεται ότι $\sum_i q_i = \sum_i p_i = \bar{p}$. Συνεπώς παίρνουμε ομοίως ότι :

$$\frac{1}{2} \sum_{\ell=d+1}^{n} |\alpha_\ell(\mathcal{Q}, \bar{p})| \cdot \|\delta^\ell \mathcal{B}_{n,\bar{p}}(\cdot)\|_1 \leq 6.5(d+1)^{1/4} 2^{-(d+1)/2}.$$

Και από αυτό έπεται πλέον άμεσα ότι:

$$d_{\mathrm{TV}}(X, Y) \leq 13(d+1)^{1/4} 2^{-(d+1)/2}.$$

□

### 3.1.4 Διαίσθηση πίσω από την κατασκευή

Σε αυτό το σημείο θέλουμε να κρατήσουμε ένα σκιαγράφημα της απόδειξης. Αν κανείς προσπαθεί να κατανοήσει ένα συμπαγή συνεχή χώρο και να τον κωδικοποιήσει με κάποιο τρόπο, αλγοριθμικά η πρώτη του σκέψη είναι να τον διακριτοποιήσει. Για παράδειγμα ως προσέγγιση του $\mathbb{R}^2$ κανείς μπορεί να πάρει τον $\epsilon\mathbb{Z}^2$. Αυτός ακριβώς θα είναι και ο δικός μας στόχος.

Το πρώτο πρόβλημα που έχουμε να διαχειριστούμε είναι η τεχνική με την οποία θα μεταχειριστούμε τα δύο σύνολα, το προσεγγιζόμενο και το προσεγγίζον σύνολο. Σε αυτό το σημείο υπάρχουν δύο επιλογές.

- Είτε κάθε στοιχείο του συνόλου θα βρούμε ένα συγκεκριμένο αλγόριθμο διακριτοποίησης και οι τελικές διαφορετικές επιλογές θα είναι το κάλυμμα που κατασκευάσαμε.
- Είτε εξ αρχής θα κατασκευάζουμε ομαδοποιημένα στοιχεία τα οποία θα προσεγγίζονται αυστηρά από ένα συγκεκριμένο εκπρόσωπο.

Σε αυτή την δουλειά θα εφαρμόσουμε εναλλάξ και τις δύο τεχνικές. Θα ξεκινήσουμε παίρνοντας μια τυχαία κατανομή PBD και θα επιδιώξουμε να την προσεγγίσουμε. Αντί να αντιμετωπίσουμε κάθε PBD ξεχωριστά θα μετράμε πόσες διαφορετικές προσεγγίσεις μπορεί να εμφανίσει μια τυχαία εκδοχή της.

Ας δούμε λοιπόν διαισθητικά ποια είναι τα πρώτα βήματα που θα εφαρμόζαμε.

Φάση 1η Ας υποθέσουμε ότι υπάρχουν κάποια $p_i$ τα οποία βρίσκονται $\epsilon$-κοντά στο 0 ή στο 1. Αυτά μπορούμε να τα στείλουμε απευθείας είτε στο 0 είτε στο 1 είτε εναλλακτικά σε μια τιμή όπως $\epsilon$ ή $1-\epsilon$. Τα κρίσιμα ερωτήματα είναι πόσα θα στρογγυλοποιήσουμε και προς ποια κατεύθυνση ώστε το σφάλμα μεταξύ αυτής της προσέγγισης να μην υπερβεί το $\epsilon$.

Η τεχνική που θα ακολουθήσουμε σε πρώτο στάδιο είναι να προσπαθήσουμε μετά την στρογγυλοποίηση η πρώτη ροπή της προσεγγιστικής κατανομής να είναι αρκετά κοντά στην αρχική. Παρατηρείστε ότι αυτό διαισθητικά έχει αρκετό νόημα. Ξέρουμε από την κλασσική θεωρία πιθανοτήτων ότι η μέση τιμή διακρίνεται ως ένας στατιστικός όρος επιρρεπής στις ακραίες τιμές. Όμως συρρικνώνοντας τις ακραίες τιμές κατά λίγο αντί τις κεντρικές προς τα έξω ξέρουμε ότι η πρώτη ροπή, δηλαδή η μέση τιμή δεν θα διαταραχθεί.

Φάση 2η Ο τρόπος με τον οποίο κανείς θα επιλέξει να διακριτοποιήσει την κατανομή μετά από αυτό το πρώτο ξεσκαρτάρισμα είναι διαφέρει ανάλογα το στήριγμα της κατανομής. Αν η κατανομή εμφανίζει πολύ μικρό στήριγμα ξέρουμε ότι σχεδόν οι περισσότερες είναι σταθερά κάποια τιμή και υπάρχουν ελάχιστες που εμφανίζουν κάποια κλιμάκωση. Επίσης αν το στήριγμα είναι μεγάλο, δηλαδή υπάρχει μεγάλη κλιμάκωση τότε θα πρέπει να προσεγγίσουμε την κατανομή με μια αντίστοιχης μέσης τιμής αλλά και διακύμανσης κατανομή.

Συγκεκριμένα στόχος μας περιληπτικά είναι να εφαρμόσουμε την εξής τακτικής

1. Αν το πολύ $k^3$ $p_i$΄ς δεν είναι $\{0,1\}$ μετά την πρώτη στρογγυλοποίηση, τότε θα χρησιμοποιήσουμε μια λεπτή μέθοδο προσέγγισης των $p_i$ σε ακέραιες στοιβάδες διαστήματος $\epsilon^2$.
2. Αν το στήριγμα είναι $\omega(k^3)$, τότε θα δείξουμε ότι ένα κάλο aggregation σε μια διωνυμική με αντίστοιχη μέση τιμή με την αρχική.

Ας δούμε μερικά θέματα που πρέπει να διαλευκάνουμε στο τοπίο αυτής της τακτικής.
**Ερωτήσεις:**

- Επηρεάζεται η θέση και το μέγεθος του στηρίγματος $0, 1X_i s$;
- Γιατί να μην διακριτοποιήσουμε τα $p_i$ στην κοντινότερη $\frac{\epsilon}{n}$ τιμή ;
- Μεταξύ δύο $\epsilon$-διακριτοποιήσεων από ακέραια πολλαπλάσια του $\epsilon^a$ και $\epsilon^b$, ποίο είναι το καταλληλότερο;
- Αν κατασκευάσει κανείς ένα κάλυμμα για PBDs με $(p_i)_1^n \in [0, 1/2]^n$ έχει ολοκληρωθεί η απόδειξη;

**Απαντήσεις:**

- Αν οι δείκτριες είναι 0 ή 1 αυστηρά τότε το μέγεθος του στηρίγματος δεν επηρεάζεται παρά μόνο την θέση του στηρίγματος.
- Από τριγωνική ανισότητα έχουμε ότι :

$$d_{\mathrm{TV}}(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i) \leq \sum_{i=1}^{n} d_{\mathrm{TV}}(X_i, Y_i) = \sum_{i=1}^{n} |p_i - q_i| \leq \sum_{i=1}^{n} \frac{\epsilon}{n} \leq \epsilon$$

  Πράγματι θα μπορούσαμε λοιπόν να διακριτοποιήσουμε την ευθεία $[0,1]$ σε $\frac{n}{\epsilon}$ σημεία $Q = \{q_i = k\frac{\epsilon}{n}k, k = \{0, 1, \cdots, \frac{n}{\epsilon}\}\}$ και να στρογγυλοποιήσουμε κάθε $p_i$ στο κοντινότερο στοιχείο αυτής της ομάδας.

  Για να κρίνουμε όμως την αποδοτικότητα αυτής της διακριτοποίησης/καλύμματος θα πρέπει να δούμε το συνολικό μέγεθος του καλύμματος, δηλαδή όλα τα διαφορετικά στοιχεία που μπορεί να περιλαμβάνει το κάλυμμα. Το πρόβλημα είναι ότι αν πάμε να μετρήσουμε όλους τους δυνατούς συνδυασμούς είναι $\frac{n}{\epsilon} \times \frac{n}{\epsilon} \times \frac{n}{\epsilon} \times \frac{n}{\epsilon} \times \cdots \times \frac{n}{\epsilon} = (\frac{n}{\epsilon})^n$, δηλαδή ένα υπερ-εκθετικά στο μέγεθος κάλυμμα.
- Προφανώς αυτό που έχει το μικρότερο μέγεθος δηλαδή αυτό με την μικρότερη δύναμη.
- Προφανώς αφού $d_{\mathrm{TV}}(K_1, K_2) = |p_i - q_i| = |(1-p_i)-(1-q_i)| = d_{\mathrm{TV}}(1 - K_1, 1 - K_2)$

## 3.2 Η Κατασκευή

Ας ακολουθήσουμε και μελετήσουμε την τεχνική των Δασκαλάκη και Παπαδημητρίου [ΔΠ13].

**Θεώρημα 3.3.** $\forall n \in \mathbb{N},\ \forall \epsilon > 0\ \exists$ *proper* $\epsilon$*-κάλυμμα* $S_\epsilon^{(n)}$ *τέτοιο ώστε:*

- $|S_\epsilon^{(n)}| \leq n^2 + n(\frac{1}{\epsilon})^{O(\log^2(1/\epsilon))}$
- $|S_\epsilon^{(n)}|$ *είναι κατασκευάσιμο* $O(n^2 \log n) + O(n \log n)(\frac{1}{\epsilon})^{O(\log^2(1/\epsilon))}$
- *Τα στοιχεία του* $S_\epsilon^{(n)}$ *μπορούν να διαχωριστούν ξεκάθαρα σε δύο βασικές κατηγορίες:*
  1. *Heavy Περίπτωση: Διωνυμικές κατανομές:* $\mathcal{B}(m, q), m \leq n$
  2. *Sparse Περίπτωση: PBDs με στήριγμα* $O(1/\epsilon^3)$ *όπου τα* $p_i$ *είναι διακριτοποιημένα σε ακέραια πολλαπλάσια του* $1/\epsilon^2$

**Παρατήρηση 3.3.** *Για ευκολία στην απόδειξη θα χρησιμοποιήσουμε μια αντίστροφη μεταβλητή* $k = k(\epsilon) = O(1/\epsilon)$ *και μάλιστα θα υποθέσουμε ότι είναι κάποιος φυσικός που ικανοποιεί αυτή την ιδιότητα.*

### 3.2.1 Βήμα 1ο: Poisson Approximation

Ας δούμε την πρώτη μέθοδο προσέγγισης. Θα κατασκευάσουμε έναν αλγόριθμο στρογγυλοποίησης που θα διατηρεί την μέση τιμή $\epsilon$-κοντά.

∗ Ας υποθέσουμε ότι έχουμε $t$ $p_i = \mathbb{E}[X_i] \in (0, 1/k = \epsilon)$ και ότι θέλουμε στρογγυλοποιήσουμε τα $p_i$ σε 0 και τα υπόλοιπα σε $1/k = \epsilon$ .

∗ Ποία είναι η σωστή αναλογία ώστε το σφάλμα να είναι μικρότερο του $1/k = \epsilon$;

$$D = |\mathbb{E}[\sum_{i=1}^{n} X_i] - \mathbb{E}[\sum_{i=1}^{n} Y_i]| = |\mathbb{E}[\sum_{i=1}^{n} X_i - Y_i]| = |\sum_{i=1}^{n} \mathbb{E}[X_i - Y_i]| = |0(t-r) + r\epsilon - \sum_{i=1}^{n} \mathbb{E}[X_i]|$$

∗ Εύκολα βλέπουμε ότι θα πρέπει $r = \lfloor k \sum \mathbb{E}[X_i] \rfloor$ ώστε $D \leq \frac{1}{k} = \epsilon$.

Μάλιστα μπορεί να δείξει κανείς εύκολα ότι μπορούμε να επεκτείνουμε αυτόν τον μηχανισμό και περισσότερο.

∗ Ας υποθέσουμε ότι $\mathbb{E}[X_i] \in [0, 1/k), [1/k, 2/k), \ldots, [k^{(a)} - 1/k, k^a/k] \forall a \in (0, 1)$ και ας υποθέσουμε ότι προσεγγίζουμε το κάθε $p_i$ στην κοντινότερη τιμή του εντός αυτού του διαστήματος.

∗ Θα εφαρμόσουμε έναν robust αλγόριθμο. Σε κάθε διάστημα $[m/k, (m+1)/k)$ προσεγγίζουμε προς τα πάνω τα $r = \lfloor k(PreviousError + \sum(\mathbb{E}[X_i] - m/k))\rfloor$

Για να χρησιμοποιήσουμε τον αλγόριθμο αυτόν θα πρέπει με κάποιο τρόπο να ποσοτικοποιήσουμε την απόκλιση. Για αυτό θα επικαλεστούμε το ακόλουθο λήμμα:

**Λήμμα 3.4** (Total Variation μεταξύ δύο $P(\lambda_1), P(\lambda_2)$)**.**

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

*Απόδειξη.* Χωρίς βλάβη της γενικότητας ας υποθέσουμε ότι $0 < \lambda_1 \leq \lambda_2$ και έστω $\delta = \lambda_2 - \lambda_1$. Τότε για κάθε $i \in \{0, 1, \ldots\}$, ορίζουμε

$$p_i = e^{-\lambda_1}\frac{\lambda_1^i}{i!} \qquad \text{και} \qquad q_i = e^{-\lambda_2}\frac{\lambda_2^i}{i!}.$$

Επίσης έστω : $\mathcal{I}^* = \{i : p_i \geq q_i\}$.

Συνεπώς έχουμε :

$$\begin{aligned}\sum_{i\in\mathcal{I}^*} |p_i - q_i| = \sum_{i\in\mathcal{I}^*} (p_i - q_i) &\leq \sum_{i\in\mathcal{I}^*} \frac{1}{i!}(e^{-\lambda_1}\lambda_1^i - e^{-\lambda_1-\delta}\lambda_1^i) \\ &= \sum_{i\in\mathcal{I}^*} \frac{1}{i!}e^{-\lambda_1}\lambda_1^i(1 - e^{-\delta}) \\ &\leq (1 - e^{-\delta})\sum_{i=0}^{+\infty} \frac{1}{i!}e^{-\lambda_1}\lambda_1^i = 1 - e^{-\delta}.\end{aligned}$$

Από την άλλη έχουμε ότι :

$$\begin{aligned}\sum_{i\notin\mathcal{I}^*} |p_i - q_i| = \sum_{i\notin\mathcal{I}^*} (q_i - p_i) &\leq \sum_{i\notin\mathcal{I}^*} \frac{1}{i!}(e^{-\lambda_1}(\lambda_1+\delta)^i - e^{-\lambda_1}\lambda_1^i) \\ &= \sum_{i\notin\mathcal{I}^*} \frac{1}{i!}e^{-\lambda_1}((\lambda_1+\delta)^i - \lambda_1^i) \\ &\leq \sum_{i=0}^{+\infty} \frac{1}{i!}e^{-\lambda_1}((\lambda_1+\delta)^i - \lambda_1^i) \\ &= e^{\delta}\sum_{i=0}^{+\infty} \frac{1}{i!}e^{-(\lambda_1+\delta)}(\lambda_1+\delta)^i - \sum_{i=0}^{+\infty} \frac{1}{i!}e^{-\lambda_1}\lambda_1^i \\ &= e^{\delta} - 1.\end{aligned}$$

Συνδυάζοντας τα προηγούμενα δύο αποκτούμε το τελικό αποτέλεσμα. □

Ας υποθέσουμε ότι έχουμε τα $p_i$ των $X_i$ και τα στρογγυλοποιημένα $p_i'$ των $X_i'$, για $p_i \in (m/k, m + 1/k)$

1. $d_{\mathrm{TV}}(\sum_{i=1}^{n} X_i, Poisson(\sum_{i=1}^{n} p_i)) \leq \frac{\sum_{i=1}^{n} p_i^2}{\sum_{i=1}^{n} p_i} \leq \max_i\{p_i\}$
2. $d_{\mathrm{TV}}(\sum_{i=1}^{n} X_i', Poisson(\sum_{i=1}^{n} p_i')) \leq \frac{\sum_{i=1}^{n} p_i'^2}{\sum_{i=1}^{n} p_i'} \leq \max_i\{p_i'\}$
3. $d_{\mathrm{TV}}(Poisson(\sum_{i=1}^{n} p_i), Poisson(\sum_{i=1}^{n} p_i')) \leq |\sinh(|(\sum_{i=1}^{n} p_i' - p_i)|)|$

Επίσης για μικρές τιμές ισχύει ότι $\sinh(x) \approx 3/2x$ Για το τρίτο ξέρουμε ότι $|(\sum_{i=1}^{n} p_i' - p_i)| \leq 1/k = \epsilon$, αφού ο αλγόριθμος μας στοχεύει στο να διατηρήσει την νέα μέση τιμή $\epsilon$ κοντά στην αρχική. Αν αθροίσουμε τις (1)+(2)+(3) έχουμε ότι :

$$d_{\mathrm{TV}}(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i') \leq \max_i\{p_i\} + \max_i\{p_i'\} + |(\sum_{i=1}^{n} p_i' - p_i)|$$

Αυτό σημαίνει ότι αν αυτή η διαδικασία σταματάει σε κάποιο σύνολο $[m/k, m + 1/k)$ το φράγμα είναι της τάξεως του $O(m/k) = O(\epsilon)$, για $m = O(1)$.

Συνεπώς:

- Όλα τα $X_i$ που έχουν μέση τιμή $\mathbb{E}[X_i] \in \{0, 1\} \cup (1/k, 1 - 1/k)$ αντιστοιχούν σε $X_i'$ς με ίδιο $p_i$, με πριν $\mathbb{E}[X_i] = \mathbb{E}[X_i']$
- Αντίθετα τα $X_i$ $\mathbb{E}[X_i] \in (0, 1/k) \cup (1 - 1/k, 1)$ χρησιμοποιούν την προαναφερθείσα διαδικασία.
- Αν θέσουμε $m = 0$ έχουμε ότι:

$$d_{\mathrm{TV}}(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i') \leq \max_i\{p_i\} + \max_i\{p_i'\} + |(\sum_{i=1}^{n} p_i' - p_i)| \leq 2 \times \frac{3.5}{k} = 7\epsilon$$

Παρατηρείστε ότι η $X' = \sum_{i=1}^{n} X_i'$ είναι απλώς μια στρογγυλοποιημένη PBD .

### 3.2.2 Βήμα 2ο: Binomial or Transported Poisson

#### 3.2.2.1 Μικρό Στήριγμα: $|\mathcal{M}| \leq k^3$

- Έστω ότι $\mathcal{M}$ το στήριγμα της PBD .
- Έστω ότι η $X'$ έχει στήριγμα $\leq k^3$ και ας θεωρήσουμε ότι μελετούμε την κάτω συμμετρική περίπτωση $\mathbb{E}[X_i'] \in [0, 1/2]$
- Παρατηρείστε ότι αν στρογγυλοποιήσουμε τις $X_i'$ σε κάποια $Y_i$ με πιθανότητες στο κοντινότερο πολλαπλάσιο $q_i = 1/k^4 = \epsilon^4$ έχουμε

$$d_{\mathrm{TV}}(\sum X_i', \sum Y_i) \leq \sum d_{\mathrm{TV}}(X_i', Y_i) \leq |support| \times max(|\mathbb{E}[X_i'] - \mathbb{E}[Y_i]|) \leq k^3/k^4 = \frac{1}{k} = \epsilon$$

- Και όμως μπορούμε να διακριτοποιήσουμε σε κάποια $Y_i$ με πολλαπλάσια $q_i = 1/k^2 = \epsilon^2$.

  Αρχικά θέτουμε ως $q_i = p'_i,\ \forall i \in [n] \setminus \mathcal{M}$. Προκύπτει ότι :

  $$d_{\mathrm{TV}}(\sum_{i\in[n]\setminus\mathcal{M}} X'_i, \sum_{i\in[n]\setminus\mathcal{M}} Y_i) = 0.$$

  Για να υπολογίσουμε τα $(q_i)_{i\in\mathcal{M}}$, θα χρησιμοποιήσουμε την προσέγγιση του Ehm μέσω των Binomial που αναφέραμε στην προηγούμενη ενότητα. Αρχικά διαμελίζουμε το $\mathcal{M}$ σε $\mathcal{M} = \mathcal{M}_l \sqcup \mathcal{M}_h$,όπου $\mathcal{M}_l = \{i \in \mathcal{M} \mid p'_i \leq 1/2\}$. Θα στρογγυλοποιήσουμε τα $(q_i)_{i\in\mathcal{M}_l}$ ώστε να ισχύουν οι εξής προϋποθέσεις:

  1. $d_{\mathrm{TV}}(\sum_{i\in\mathcal{M}_l} X'_i, \sum_{i\in\mathcal{M}_l} Y_i) \leq 17/k$.
  2. $\forall i \in \mathcal{M}_l$, τα $q_i$ είναι ένα ακέραια πολλαπλάσια του $1/k^2$.

  Για να ορίσουμε τα $(q_i)_{i\in\mathcal{M}_h}$, για τις μεταβλητές με $p'_i > 1/2$ εφαρμόζουμε την ίδια διαδικασία για τις τιμές $(1-p'_i)_{i\in\mathcal{M}_h}$ για να κερδίσουμε τις στρογγυλοποιημένες τιμές $(1-q_i)_{i\in\mathcal{M}_h}$. Υποθέτοντας την ορθότητα της διαδικασίας για πιθανότητες $\leq 1/2$ τότε και για την συμμετρική πάνω περίπτωση που έχουμε:

  1. $d_{\mathrm{TV}}(\sum_{i\in\mathcal{M}_h} X'_i, \sum_{i\in\mathcal{M}_h} Y_i) \leq 17/k$
  2. $\forall i \in \mathcal{M}_h$, τα $q_i$ είναι ένα ακέραια πολλαπλάσια του$1/k^2$.

  Από τριγωνική ανισότητα προκύπτει ότι:

  $$d_{\mathrm{TV}}(\sum_{i\in\mathcal{M}} X'_i, \sum_{i\in\mathcal{M}} Y_i) \leq 34/k$$

  Αρκεί λοιπόν να διαλέξουμε τα $(q_i)_{i\in\mathcal{M}_l}$ κατάλληλα.

  Συνεπώς διαμερίζουμε το $\mathcal{M}_l = \mathcal{M}_{l,1} \sqcup \mathcal{M}_{l,2} \sqcup \ldots \sqcup \mathcal{M}_{l,k-1}$ ώστε $\forall j$:

  $$\mathcal{M}_{l,j} = \left\{ i \;\middle|\; p'_i \in \left[ \frac{1}{k} + \frac{(j-1)j}{2}\frac{1}{k^2}, \frac{1}{k} + \frac{(j+1)j}{2}\frac{1}{k^2} \right) \right\}.$$

  (Παρατηρήστε ότι το μήκος του κάθε διαστήματος $|\mathcal{M}_{l,j}|$ είναι $\frac{j}{k^2}$.)

  Τώρα για κάθε $j = 1, \ldots, k-1$ τέτοιο ώστε $\mathcal{M}_{l,j} \neq \emptyset$, ορίζουμε τα $(q_i)_{i\in\mathcal{M}_{l,j}}$ μέσω αυτής της διαδικασίας :

  1. Θέτουμε $p_{j,\mathcal{M}in} := \frac{1}{k} + \frac{(j-1)j}{2}\frac{1}{k^2}$, $p_{j,\mathcal{M}ax} := \frac{1}{k} + \frac{(j+1)j}{2}\frac{1}{k^2}$, $n_j = |\mathcal{M}_{l,j}|$, $\bar{p}_j = \frac{\sum_{i\in\mathcal{M}_{l,j}} p'_i}{n_j}$.
  2. Θέτουμε $r = \left\lfloor \frac{n_j(\bar{p}_j - p_{j,\mathcal{M}in})}{j/k^2} \right\rfloor$
  3. Έστω επίσης $\mathcal{M}'_{l,j} \subseteq \mathcal{M}_{l,j}$ ένα τυχαίο υποσύνολο με $r$ στοιχεία.

4. Σετ $q_i = p_{j,\mathcal{M}ax}, \forall i \in \mathcal{M}'_{l,j}$.
5. Για κάποιο τυχαίο δείκτη $i^*_j \in \mathcal{M}_{l,j} \setminus \mathcal{M}'_{l,j}$, θέτουμε $q_{i^*_j} = n_j\bar{p}_j - (rp_{j,\mathcal{M}ax} + (n_j - r - 1)p_{j,\mathcal{M}in})$.
6. Τέλος, θέτουμε $q_i = p_{j,\mathcal{M}in}, \forall i \in \mathcal{M}_{l,j} \setminus \mathcal{M}'_{l,j} \setminus \{i^*_j\}$.

Είναι εύκολο να δει κανείς ότι

1. $\sum_{i\in\mathcal{M}_{l,j}} p'_i = \sum_{i\in\mathcal{M}_{l,j}} q_i \equiv n_j\bar{p}_j$.
2. $\forall i \in \mathcal{M}_{l,j} \setminus \{i^*_j\}$, όπου τα $q_i$ είναι ακέραια πολλαπλάσια του $1/k^2$.

Χρησιμοποιώντας την προσέγγιση του Ehm έχουμε:

$$d_{\mathrm{TV}}(\sum_{i\in\mathcal{M}_{l,j}} X'_i, \mathcal{B}inomial\,(n_j, \bar{p}_j)) \leq \frac{\sum_{i\in\mathcal{M}_{l,j}}(p'_i - \bar{p}_j)^2}{(n_j+1)\bar{p}_j(1-\bar{p}_j)} \leq \begin{cases} \frac{n_j(j\frac{1}{k^2})^2}{(n_j+1)p_{j,\min}(1-p_{j,\min})}, \text{όταν } j < k-1 \\ \frac{n_j(j\frac{1}{k^2})^2}{(n_j+1)p_{j,\max}(1-p_{j,\max})}, \text{όταν } j = k-1 \end{cases}$$
$$\leq \frac{8}{k^2}.$$

Αντίστοιχα προκύπτει ότι η $d_{\mathrm{TV}}(\sum_{i\in\mathcal{M}_{l,j}} Y_i, \mathcal{B}inomial\,(n_j, \bar{p}_j)) \leq \frac{8}{k^2}$. Από τριγωνική λοιπόν έχουμε:

$$d_{\mathrm{TV}}(\sum_{i\in\mathcal{M}_{l,j}} X'_i, \sum_{i\in\mathcal{M}_{l,j}} Y_i) \leq \frac{16}{k^2}.$$

Και πάλι από τριγωνική ανισότητα έχουμε ότι $\forall j = 1, \ldots, k-1$,

$$d_{\mathrm{TV}}(\sum_{i\in\mathcal{M}_l} X'_i, \sum_{i\in\mathcal{M}_l} Y_i) \leq \sum_{j=1}^{k-1} d_{\mathrm{TV}}(\sum_{i\in\mathcal{M}_{l,j}} X'_i, \sum_{i\in\mathcal{M}_{l,j}} Y_i) \leq \frac{16}{k}.$$

Με την τεχνική αυτή έχουμε όλα τα $q_i$ να είναι ακέραια πολλαπλάσια της ποσότητας του $1/k^2 = \epsilon^2$, εκτός από κάποια στοιχεία $Q^* = \{q_{i^*_1}, \ldots, q_{i^*_{k-1}}\}$. Αν προσεγγίσουμε κάθε στοιχείο του $Q^*$ με το κοντινότερο ακέραιο πολλαπλάσιο του $1/k^2 = \epsilon^2$, τότε συνολικά έχουμε ένα επιπλέον σφάλμα της τάξεως $\leq (k-1) \times \frac{1}{k^2} = O(\epsilon)$.

Ανακεφαλαιώνοντας, θα μπορούσε κανείς να παρατηρήσει ότι όταν το στήριγμα είναι αρκετά μικρό, τότε οι περισσότερες μεταβλητές είναι $0, 1$. Για τις υπόλοιπες μεταβλητές μπορούμε να επιχειρήσουμε είτε ένα χοντρικό $\epsilon^4$ είτε ένα ακόμα πιο λεπτό $\epsilon^2$ μηχανισμό στρογγυλοποίησης και το κόστος αυτής πιο προσεκτική στρογγυλοποίησης είναι ανεκτό αφού είναι σχετικά λίγες.

Αξίζει κανείς να παρατηρήσει ότι η τελική μεταβλητή $Y = \sum_i Y_i$ αποτελεί μια επίσης κλασσική PBD .

#### 3.2.2.2 Μεγάλο Στήριγμα: $|\mathcal{M}| > k^3$

Σε αυτή την περίπτωση σκοπεύουμε να εφαρμόσουμε μια πολύ πιο επιθετική αντιμετώπιση. Στόχος μας είναι να βρούμε μια συμπαγής κατανομή, όπως μια διωνυμική κατανομή που θα καταφέρνει να καλύπτει όλο το στήριγμα της κατανομής.

Σε αυτό το σημείο θα χρησιμοποιήσουμε την Translated Poisson ως ενδιάμεσο κρίκο για να υπολογίσουμε την ποιότητα της προσέγγισης. Για το σκοπό αυτό θα προσπαθήσουμε η μέση τιμή και η διακύμανση της αρχικής και της τελικής να είναι όσο πιο κοντά γίνεται. Παρατηρείστε ότι :

$$\begin{cases} \mu = \sum_i p'_i \geq (\min_i p'_i) \times |support| = k^2 \\ \sigma^2 = \sum_i p'_i(1-p'_i) \geq |support| \times (\min_i p'_i)(1-\min_i p'_i) = k^2(1-\frac{1}{k} \end{cases}$$

$$\begin{aligned} d_{\mathrm{TV}}(\sum_i X'_i, TP(\mu,\sigma^2)) &\leq \frac{\sqrt{\sum_i p'^3_i(1-p'_i)}+2}{\sum_i p'_i(1-p'_i)} \leq \frac{\sqrt{\sum_i p'_i(1-p'_i)}+2}{\sum_i p'_i(1-p'_i)} \\ &\leq \frac{1}{\sqrt{\sum_i p'_i(1-p'_i)}} + \frac{2}{\sum_i p'_i(1-p'_i)} = \frac{1}{\sigma} + \frac{2}{\sigma^2} \end{aligned}$$

Όπως προαναφέραμε, θέλουμε να υπολογίσουμε τα $m', q$ για την $\mathcal{B}inomial(m', q)$ που θα προσεγγίσει την PBD μας. Ας εφαρμόσουμε την αντίστροφη διαδικασία και ας καταγράψουμε τα μέχρι τώρα εργαλεία.

1. $d_{\mathrm{TV}}\left(\mathcal{B}inomial(m',q), TP(m'q, m'q(1-q))\right) \leq \frac{1}{\sqrt{m'q(1-q)}} + \frac{2}{m'q(1-q)}$
2. $d_{\mathrm{TV}}\left(\sum_i X'_i, TP(\mu,\sigma^2)\right) \leq \frac{1}{\sigma} + \frac{2}{\sigma^2}$
3. $d_{\mathrm{TV}}\left(TP(\mu_1,\sigma_1^2), TP(\mu_2,\sigma_2^2)\right) \leq \frac{|\mu_1-\mu_2|}{min(\sigma_1,\sigma_2)} + \frac{|\sigma_1^2-\sigma_2^2|+1}{min(\sigma_1^2,\sigma_2^2)}$

Άρα

$$d_{\mathrm{TV}}\left(\mathcal{B}inomial(m',q), \sum_i X'_i\right) \leq \begin{cases} \frac{|\mu-m'q|}{min(\sigma', \sqrt{m'q(1-q)})} + \\ \frac{|\sigma'^2-m'q(1-q)|+1}{min(\sigma^2, m'q(1-q))} + \\ \frac{1}{\sigma} + \\ \frac{2}{\sigma^2} + \\ \frac{1}{\sqrt{m'q(1-q)}} + \\ \frac{2}{m'q(1-q)} \end{cases}$$

- Στόχος μας είναι να πετύχουμε $|\sigma^2_{old} - \sigma^2_{new}|, |\mu_{old} - \mu_{new}| \in \Theta(1)$.
- Παρατηρείστε ότι αρκεί το $m' \leq |\#\mathbb{E}[X'_i] > 0| \leq n$.

- Αφού προσδιορίσουμε το $m'$ το $q$ είναι εύκολο να επιλεγεί αφού θα πρέπει:

$$|\mu_{old} - \mu_{new}| \in \Theta(1)$$

Όντως:

$$\sum \mathbb{E}[X_i] \approx m'q, q = \mathbb{Z} \times (1/n) \Rightarrow q = \frac{\lceil n * \sum_i \mathbb{E}[X_i]/m \rceil}{n}$$

- Για να πετύχουμε $|\sigma^2_{old} - \sigma^2_{new}| \in \Theta(1)$ μετά από πράξεις προκύπτει η απαίτηση ότι $m \in \Theta(\lceil \frac{(\sum \mathbb{E}[X_i])^2}{\sum (\mathbb{E}[X_i])^2} \rceil)$

**Παρατήρηση 3.4.** *Ο λόγος που απαιτούμε να προσδιορίσουμε ένα ακέραιο πολλαπλάσιο του $\frac{1}{n}$ είναι αλγοριθμικός, αφού θέλουμε να είναι εύκολη η διακριτοποίηση των δυνατών επιλογών.*

Για περισσότερες λεπτομέρειες μπορεί κανείς να δει τις αντίστοιχες ενότητες στο [ΔΠ13]. Κρίσιμα στοιχεία αυτής της διαδικασίας αποτελούν ότι αν το στήριγμα είναι αρκετά μεγάλο τότε η αντίστροφη τιμή της διακύμανση φράσσεται από το επιτρεπτό σφάλμα $\epsilon$ και άρα τα bounds είναι αρκετά μικρά ώστε να μην παραβιάζεται η αντοχή του σφάλματος. Συγκεκριμένα:

$$\sigma^2 \geq k^2 \left(1 - \frac{1}{k}\right) \Rightarrow \frac{1}{\sigma} = O(\frac{1}{k}) = O(\epsilon)$$

Έχοντας εξασφαλίσει με τις παραπάνω τιμές ότι $|\sigma^2_{old} - \sigma^2_{new}|, |\mu_{old} - \mu_{new}| \in O(1)$ και $\mu \geq k^2, \sigma \geq k^2(1 - \frac{1}{k})$, εύκολα προκύπτει ότι $d_{\mathrm{TV}}\left(\sum_i X_i, \sum_i Y_i\right) \leq \frac{9}{k} = O(\epsilon)$.

Συναθροίζοντας όλες τις μέχρι τώρα προσεγγίσεις έχουμε ότι για κάθε τυχαία PBD $X = \sum_{i=1}^n X_i$ μπορούμε να βρούμε μια PBD $Y = \sum_{i=1}^n Y_i$ δύο συγκεκριμένων μορφών που θα απέχει από την αρχική το πολύ $d_{\mathrm{TV}}(X, Y) \geq 7/k + 2 \times \max\{17/k, 9/k\} = \frac{41}{k} = O(\epsilon)$.

### 3.2.3 Ανακεφαλαιώνοντας

Ας ανακεφαλαιώσουμε και να δούμε την κατασκευή του καλύμματος που έχουμε μέχρι στιγμής. Ας επιβεβαιώσουμε πρώτα την ύπαρξη του $\epsilon$-καλύμματος $\mathcal{S}'_{n,\epsilon}$ των PBDs με μέγεθος το πολύ $n^2 + n \cdot \left(\frac{1}{\epsilon}\right)^{O(1/\epsilon^2)}$.

Το κάλυμμα μας αποτελείται από την ένωση των διωνυμικών κατανομών στην περίπτωση του μεγάλου στηρίγματος και των λεπτότερης ακρίβειας PBDs στην περίπτωση του μικρού στηρίγματος. Στόχος μας είναι να απαριθμίσουμε τα στοιχεία αυτού του καλύμματος. Προς τον σκοπό αυτό θα εργαστούμε αποκλειστικά μέσω της συζυγής ακέραιας μεταβλητής $k = \lceil 41/\epsilon \rceil$.

Ας αριθμήσουμε για αρχή τις διωνυμικές κατανομές. Ο συνολικός αριθμός τέτοιων κατανομών είναι το πολύ $n^2$, γιατί υπάρχουν $n$ επιλογές για το ακέραιο πολλαπλάσιο του $\frac{1}{n}$ που ορίζει την πιθανότητα επιτυχίας σε κάθε ανεξάρτητο πείραμα της διωνυμικής και υπάρχουν και το πολύ $n$ διαφορετικές επιλογές για την τιμή των πειραμάτων $m$.

Πιο μεγάλο ενδιαφέρον αποτελεί ο υπολογισμός των sparse-PBDs στην περίπτωση του μικρού στηρίγματος. Το μέγεθος του είναι $(k^3+1) \cdot k^{3k^2} \cdot (n+1)$ ή ισοδύναμα $n \cdot \left(\frac{1}{\epsilon}\right)^{O(1/\epsilon^2)}$,

αφού υπάρχουν $k^3+1$ για το μήκος του στηρίγματος $\ell$, επίσης υπάρχουν $n+1$ επιλογές για να τοποθετήσουμε την αρχή του στηρίγματος καρφώνοντας κάποιες τυχαίες μεταβλητές στην 1 ενώ υπάρχουν το πολύ $(k^3)^{k^2}$ για τις πιθανότητες $p_1 \leq p_2 \leq \ldots \leq p_\ell$ από το σύνολο των $\left\{\frac{1}{k^2}, \frac{2}{k^2}, \ldots, \frac{k^2-1}{k^2}\right\}$ δυνατών στρογγυλοποιημένων τιμών. Παρατηρείστε επίσης ότι η διαδικασία απαρίθμισης του διαρκεί $O(n^2 \log n) + O(n \log n) \cdot \left(\frac{1}{\epsilon}\right)^{O(1/\epsilon^2)}$, αφού κάθε αριθμός στο $\{0, \ldots, n\}$ και κάθε πιθανότητα στο $\left\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\right\}$ μπορεί να αναπαρασταθεί χρησιμοποιώντας $O(\log n)$ bits, ενώ οι αριθμοί $\{0, \ldots, k^3\}$ και κάθε πιθανότητα στο σύνολο $\left\{\frac{1}{k^2}, \frac{2}{k^2}, \ldots, \frac{k^2-1}{k^2}\right\}$ μπορεί να αναπαρασταθεί με $O(\log k) = O(\log 1/\epsilon)$ bits.

Τέλος αξίζει να σημειωθεί ότι το κάλυμμα είναι proper, αφού περιλαμβάνει PBD κατανομές.

### 3.2.4 Αποβάλλοντας τις $\epsilon$-επαναλήψεις

Ένα από τα πιο ενδιαφέροντα στοιχεία αυτής της κατασκευής είναι η εκθετική συρρίκνωση του καλύμματος. Το βασικό αποτέλεσμα προκύπτει από το θεώρημα 3.2 που μελετήσαμε προηγουμένως. Η ιδέα είναι εξαιρετικά απλή. Έχουμε απαριθμήσει το σύνολο μιας ομάδας κατανομών. Από αυτές υπάρχουν αρκετές οι οποίες εμφανίζουν ίσες τις πρώτες τους ροπές. Όμως το θεώρημα 3.2 μας διδάσκει ότι η απόσταση τους είναι εκθετικά μικρή. Συνεπώς αν κρατήσουμε μια εξ αυτών το σφάλμα δεν θα επηρεαστεί αισθητά.

Ας γίνουμε πιο τυπικοί. Για μια συλλογή $\mathcal{P} = (p_i)_{i\in[n]} \in [0,1]^n$ από διάφορες τιμές που αντιστοιχούν σε πιθανότητες θα συμβολίζουμε ως $\mathcal{L}_\mathcal{P} = \{i \mid p_i \in (0, 1/2]\}$ και με $\mathcal{R}_\mathcal{P} = \{i \mid p_i \in (1/2, 1)\}$. Το θεώρημα 3.2 και τα πορίσματα που το ακολουθούν μας οδηγούν στο ακόλουθο συμπέρασμα. Αν δύο $n-$άδες από πιθανότητες $\mathcal{P} = (p_i)_{i\in[n]}$, $\mathcal{Q} = (q_i)_{i\in[n]}$ ικανοποιούν τις εξής συνθήκες:

$$\sum_{i\in\mathcal{L}_\mathcal{P}} p_i^t = \sum_{i\in\mathcal{L}_\mathcal{Q}} q_i^t, \forall t = 1, \ldots, d$$

$$\sum_{i\in\mathcal{R}_\mathcal{P}} p_i^t = \sum_{i\in\mathcal{R}_\mathcal{Q}} q_i^t, \forall t = 1, \ldots, d$$

$(p_i)_{[n]\setminus(\mathcal{L}_\mathcal{P}\cup\mathcal{R}_\mathcal{P})}$ και $(q_i)_{[n]\setminus(\mathcal{L}_\mathcal{Q}\cup\mathcal{R}_\mathcal{Q})}$

Τότε:

$$d_{\mathrm{TV}}(\mathcal{P}, \mathcal{Q}) \leq 2 \cdot 13(d+1)^{1/4} 2^{-(d+1)/2}$$

Συγκεκριμένα για κάποιο $d(\epsilon) = O(\log 1/\epsilon)$, το φράγμα γίνεται της τάξεως του $O(\epsilon)$. Για κάθε ομάδα θα ορίσουμε ένα χαρτοφυλάκιο ροπών.

Π.χ: Για την $\mathcal{P} = (p_i)_{i\in[n]} \in [0,1]^n$, ορίζουμε το *χαρτοφυλάκιο ροπών* $m_\mathcal{P}$ ως ένα $(2d(\epsilon)+1)$-διάστατο διάνυσμα.

$$m_\mathcal{P} = \left(\sum_{i\in\mathcal{L}_\mathcal{P}} p_i, \sum_{i\in\mathcal{L}_\mathcal{P}} p_i^2, \ldots, \sum_{i\in\mathcal{L}_\mathcal{P}} p_i^{d(\epsilon)}; \sum_{i\in\mathcal{R}_\mathcal{P}} p_i, \ldots, \sum_{i\in\mathcal{R}_\mathcal{P}} p_i^{d(\epsilon)}; |\{i \mid p_i = 1\}|\right).$$

Από την προηγούμενη συζήτηση οι $\mathcal{P}, \mathcal{Q}$, αν $m_{\mathcal{P}} = m_{\mathcal{Q}}$ θα εμφανίσουν απόσταση:

$$d_{\mathrm{TV}}(\mathrm{PBD}(\mathcal{P}), \mathrm{PBD}(\mathcal{Q})) \leq \epsilon$$

Δεδομένου αυτού μπορούμε να καθαρίσουμε το κάλυμμα με τον εξής τρόπο: Για κάθε διαφορετικό χαρτοφυλάκιο ροπών που προκύπτει από αραιές PBDs , εμείς κρατάμε αυστηρά μια εξ αυτών. Το προκύπτον κάλυμμα αποτελεί ένα $2\epsilon$-κάλυμμα, αφού πλέον θα πιθανό να πρέπει να κάνεις δύο άλματα μεγέθους $\epsilon$ ώστε να φτάσεις από μια τυχαία PBD σε μία που θα βρίσκεται εντός του καλύμματος.

Ας φράξουμε τώρα το μέγεθος του καλύμματος με αντίστοιχο τρόπο όπως πριν. Αρκεί να μετρήσουμε τα διαφορετικά χαρτοφυλάκια $k^{O(d(\epsilon)^2)} \cdot (n+1)$. Πράγματι ας θεωρήσουμε μια αραιή $\mathrm{PBD}(\mathcal{P} = (p_i)_{i\in[n]})$. Υπάρχουν $k^3+1$ για να προσδιορίσεις το μέγεθος του κάτω στηρίγματος $|\mathcal{L}_{\mathcal{P}}|$, αντίστοιχα $k^3+1$ για το μέγεθος του άνω στηρίγματος $|\mathcal{R}_{\mathcal{P}}|$ και το πολύ $(n+1)$ επιλογές για $|\{i \mid p_i = 1\}|$.

Επίσης ο συνολικός αριθμός διανυσμάτων της μορφής:

$$\left( \sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i, \sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i^2, \ldots, \sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i^{d(\epsilon)} \right)$$

είναι $k^{O(d(\epsilon)^2)}$. Πράγματι αν $|\mathcal{L}_{\mathcal{P}}| = 0$ υπάρχει ακριβώς ένα τέτοιο διάνυσμα, $\vec{0}$. Αν $|\mathcal{L}_{\mathcal{P}}| > 0$, τότε $\forall t = 1, \ldots, d(\epsilon)$, $\sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i^t \in (0, |\mathcal{L}_{\mathcal{P}}|]$ και θα πρέπει να είναι ακέραια πολλαπλάσια της μορφής $1/k^{2t}$. Συνεπώς ο συνολικός αριθμός των δυνατών τιμών των :

$$\sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i^t$$

είναι το πολύ :

$$k^{2t}|\mathcal{L}_{\mathcal{P}}| \leq k^{2t}k^3$$

και ο συνολικός αριθμός των δυνατών τιμών των

$$\left( \sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i, \sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i^2, \ldots, \sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i^{d(\epsilon)} \right)$$

είναι το πολύ:

$$\prod_{t=1}^{d(\epsilon)} k^{2t}k^3 \leq k^{O(d(\epsilon)^2)}.$$

Το ίδιο άνω φράγμα μπορεί να χρησιμοποιηθεί για να προσεγγίσουμε την

$$\left( \sum_{i\in\mathcal{R}_{\mathcal{P}}} p_i, \sum_{i\in\mathcal{R}_{\mathcal{P}}} p_i^2, \ldots, \sum_{i\in\mathcal{R}_{\mathcal{P}}} p_i^{d(\epsilon)} \right).$$

Συνεπώς υπάρχουν γενικώς το πολύ $k^{O(d(\epsilon)^2)} \cdot (n+1)$ χαρτοφυλάκια ροπών. Σε αυτά βρίσκονται όλα τα χαρτοφυλάκια ροπών των αραιών PBDs . Συνεπώς μπορούμε να αντικαταστήσουμε το ένα τμήμα του καλύμματος με μόλις $k^{O(d(\epsilon)^2)} \cdot (n+1) = n \cdot \left(\frac{1}{\epsilon}\right)^{O(\log^2 1/\epsilon)}$ κατανομές. Ο αριθμός των διωνυμικών κατανομών$(n,k)$ δεν άλλαξε συνεπώς το συνολικό νέο κάλυμμα είναι $n^2 + n \cdot \left(\frac{1}{\epsilon}\right)^{O(\log^2 1/\epsilon)}$.

Τέλος για να ολοκληρώσουμε την αφήγηση μας είναι σημαντικό να επιχειρματολογήσει κανείς ότι δεν χρειάζεται να υπολογίσει το αρχικό κάλυμμα και στην συνέχεια να εφαρμόσει αυτήν την τεχνική της αραίωσης. Αντιθέτως μπορεί κανείς να απαριθμήσει με δυναμικό προγραμματισμό απευθείας τα διαφορετικά χαρτοφυλάκια και όντως σε χρόνο $O(n^2 \log n) + O(n \log n) \cdot \left(\frac{1}{\epsilon}\right)^{O(\log^2 1/\epsilon)}$. Περισσότερες πληροφορίες μπορεί να βρει κανείς στο appendix του [ΔΠ13]

## 3.3 Μαθαίνοντας μια PBD

Ένα από τα σημαντικότερα αποτελέσματα του [ΔΠ13] ήταν ότι επέτρεψε την αποδοτική εκμάθηση μιας τυχαίας PBD . Σε αυτό το τμήμα του κεφαλαίου θα παρουσιάσουμε αυτήν την εξαιρετική εργασία των Rocco Servedio,Costis Daskalakis,Ilias Diakonikolas [ΔΔΣ15].

Ας ανακεφαλαιώσουμε πρώτα για ακόμα μια φορά το θεώρημα του καλύμματος.

**Θεώρημα 3.4.** *Αν $\mathcal{S}$, το σύνολο όλων των PBDs , τότε ένα proper $\forall \epsilon > 0$ $\exists \epsilon$-κάλυμμα $\mathcal{S}_\epsilon \subseteq \mathcal{S}$ του $\mathcal{S}$ τέτοιο ώστε*

1. $|\mathcal{S}_\epsilon| \leq n^2 + n \cdot \left(\frac{1}{\epsilon}\right)^{O(\log^2 1/\epsilon)}$ *και*
2. $\mathcal{S}_\epsilon$ *μπορεί να κατασκευασεί σε γραμμικό χρόνο ως προς το μέγεθος της αναπαράστασης του:* $O(n^2 \log n) + O(n \log n) \cdot \left(\frac{1}{\epsilon}\right)^{O(\log^2 1/\epsilon)}$.

*Επιπλέον αν $\{Y_i\} \in \mathcal{S}_\epsilon$,τότε η $n$-άδα των Bernoulli τυχαίων μεταβλητών $\{Y_i\}_{i=1,\ldots,n}$ έχει μια από τις ακόλουθες μορ΄φες $k = k(\epsilon) \leq C/\epsilon$ όπου $C$ μια ακέραια θετική σταθερά $C = 41 > 0$:*

*(ι) (k-Sparse Form) Υπάρχει $\ell \leq k^3 = O(1/\epsilon^3)$ τέτοιο ώστε*
$\forall i \leq \ell$, $\mathbb{E}[Y_i] \in \left\{\frac{1}{k^2}, \frac{2}{k^2}, \ldots, \frac{k^2-1}{k^2}\right\}$ *και*
$\forall i > \ell$, $\mathbb{E}[Y_i] \in \{0,1\}$.

*(ιι) (k-Heavy Binomial Form) Υπάρχει $\ell \in \{1,\ldots,n\}$ και $q \in \left\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\right\}$ τέτοιο ώστε*
$\forall i \leq \ell$, $\mathbb{E}[Y_i] = q$ *και* $\forall i > \ell$, $\mathbb{E}[Y_i] = 0$. *Επιπλέον $\ell, q$*
$\mu = \ell q \geq k^2 = \Omega(1/\epsilon^2)$ *και* $\sigma^2 = \ell q(1-q) \geq k^2 - k - 1 = \Omega(1/\epsilon^2)$.

**Παρατήρηση 3.5.** *Από την απόδειξη επίσης προκύπτει ότι αν για την $\{X_i\} \in \mathcal{S}$ δεν υπάρχει $\epsilon$-γείτονας στο κάλυμμα $\mathcal{S}_\epsilon$ το οποίο να έχει sparse form, τότε υπάρχει σίγουρα μια $\{Y_i\} \in \mathcal{S}_\epsilon$ k-heavy Binomial form τέτοια ώστε*

*(ιιι)* $d_{\mathrm{TV}}(\sum_i X_i, \sum_i Y_i) \leq \epsilon \cdot$ *και*

*(ι)* *Αν* $\mu = \mathbb{E}[\sum_i X_i]$, $\mu' = \mathbb{E}[\sum_i Y_i]$, $\sigma^2 = \mathbb{V}ar[\sum_i X_i]$ *και* $\sigma'^2 = \mathbb{V}ar[\sum_i Y_i]$, *τότε* $|\mu - \mu'| = O(1)$ *και* $|\sigma^2 - \sigma'^2| = O(1 + \epsilon \cdot (1 + \sigma^2))$.

### 3.3.1 Non Proper Learning

Ας κάνουμε την πρώτη μας προσπάθεια να μάθουμε με έναν non-proper τρόπο την κατανομή.

**Θεώρημα 3.5.** *Έστω* $X = \sum_{i=1}^{n} X_i$ *μια άγνωστη PBD .*

- **[Μαθαίνοντας μια PBD με σταθερό αριθμό δειγμάτων]** *Υπάρχει ένας αλγόριθμος με τις ακόλουθες ιδιότητες δεδομένου των αριθμών* $n, \epsilon, \delta$ *και πρόσβασης σε ανεξάρτητα δείγματα στην* $X$*, με χρήση*
  - $\tilde{O}\left((1/\epsilon^3) \cdot \log(1/\delta)\right)$ *δειγμάτων από το* $X$
  - $\tilde{O}\left((1/\epsilon^3) \cdot \log n \cdot \log^2 \frac{1}{\delta}\right)$ *υπολογιστικών πράξεων*

  *με πιθανότητα τουλάχιστον* $1 - \delta$ *να προτείνει μια κατανομή (σε κάποια τυπική μορφή)* $\hat{X}$ *πάνω στο* $[n] = \{1, 2, \cdots, n\}$ *και η οποία βρίσκεται σε απόσταση* $d_{\mathrm{TV}}(\hat{X}, X) \leq \epsilon$.

**Ο Βασικός Αλγόριθμος Μάθησης.**

`Learn-PBD`$(n, \epsilon, \delta)$

1. Εκτέλεσε τον `Learn-Sparse`$^X(n, \epsilon, \delta/3)$ και αποθήκευσε την έξοδο κατανομή $H_S$.
2. Εκτέλεσε τον `Learn-Poisson`$^X(n, \epsilon, \delta/3)$ και αποθήκευσε την έξοδο κατανομή $H_P$.
3. Χρησιμοποίησε την `Choose-Hypothesis`$^X(H_S, H_P, \epsilon, \delta/3)$.

Σχήμα 3.1: `Learn-PBD`$(n, \epsilon, \delta)$

Θυμίζουμε ότι ο αλγόριθμος `Choose-Hypothesis` αποτελεί την τεχνική 2.2 που παρατηρήσαμε στο προηγούμενο κεφάλαιο. Μέσω αυτής μπορούμε να διαλέξουμε με μεγάλη πιθανότητα ποία από τις δύο κατανομές-υποψήφιες βρίσκονται πιο κοντά στο την κατανομή μας.

Η υπορουτίνα `Learn-Sparse` υπολογίζει χρησιμοποιώντας δείγματα από την $X$ μια $\epsilon$-κοντά κατανομή $H_S$ με πιθανότητα τουλάχιστον $1 - \delta/3$, αν η άγνωστη PBD $X$ είναι $\epsilon$-κοντά σε κάποια sparse form PBD μέσα στο κάλυμμα $\mathcal{S}_\epsilon$. Η υπορουτίνα `Learn-Poisson` υπολογίζει χρησιμοποιώντας δείγματα από την $X$ μια $\epsilon$-κοντά κατανομή $H_P$ σε περίπτωση όπου η $X$ δεν είναι $\epsilon$-κοντά σε καμία sparse form PBD . Σε αυτή την περίπτωση γνωρίζουμε ότι η $X$ βρίσκεται $\epsilon$-κοντά σε κάποια $k(\epsilon)$-heavy Binomial form.

#### 3.3.1.1 Μαθαίνοντας την $X$ όταν είναι κοντά σε μία sparse form PBD .

Αρχική μας παρατήρηση αποτελεί ότι η PBD είναι μονόλοφη (unimodal) κατανομή πάνω στο $[n]$. Συνεπώς μπορούμε να χρησιμοποιήσουμε το θεώρημα του Birgé [Βιρ97] και αναπτύξαμε στο τέλος του προηγούμενου κεφαλαίου.

**Θεώρημα 3.6** ([Βιρ97]Birge unimodal)**.** *Birge unimodal $\forall n, \epsilon, \delta > 0$, υπάρχει ένας αλγόριθμος που χρησιμοποιεί:*

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

*δείγματα από την άγνωστη κατανομή $X$ πάνω στο $[n]$, πραγματοποιεί*

$$\tilde{O}\left(\frac{\log^2 n}{\epsilon^3}\log^2\frac{1}{\delta}\right)$$

*πράξεις, και εξάγει μια κλιμακωτή αύξουσα και κλιμακωτή φθίνουσα κατανομή $H$ με $O(\log n/\epsilon)$ σκαλοπάτια με πιθανότητα $1-\delta$ με βεβαιότητα ότι η $d_{\mathrm{TV}}(X,H)\leq\epsilon$.*

Η βασική ιδέα είναι αρκετά απλή. Αρχικά θα κατασκευάσουμε μια $O(\epsilon')$-κοντά κατανομή του $X$, από τα άκρα της κατανομή λαμβάνοντας την δεσμευμένη κατανομή $X_{[\hat{a},\hat{b}]}$. Στόχος μας με αυτή την ενέργεια είναι να μετρήσουμε το στήριγμα της $X$ και να βρούμε μια $\epsilon$-κοντά κατανομή. Επίσης αφού η $X$ είναι unimodaltότε και η $X_{[\hat{a},\hat{b}]}$. Αν $\hat{b}-\hat{a}$ είναι μεγαλύτερο του $O(1/\epsilon'^3)$ τότε ο αλγόριθμος έχει αποτύχει, αφού είναι σίγουρο ότι δεν είναι κοντά σε sparse form. Διαφορετικά χρησιμοποιούμε τον αλγόριθμο του Birgé για να μάθουμε την κατανομή $X_{[\hat{a},\hat{b}]}$.

`Learn-Sparse`$^X(n,\epsilon',\delta')$

1. Ζήτα $M=32\log(8/\delta')/\epsilon'^2$ δείγματα από την $X$ και ταξινόμησε τα ως: $0\leq s_1\leq\cdots\leq s_M\leq n$.
2. Όρισε $\hat{a}:=s_{\lceil 2\epsilon' M\rceil}$ και $\hat{b}:=s_{\lfloor(1-2\epsilon')M\rfloor}$.
3. Αν $\hat{b}-\hat{a}>(C/\epsilon')^3$ (όπου $C=41$ η σταθερά του θεωρήματος του καλύμματος) και επέστρεψε μια singleton κατανομή στο 0 με πιθανότητα 1.
4. Αλλίως, τρέξε τον αλγόριθμο του Birgé

Σχήμα 3.2: `Learn-Sparse`$^X(n,\epsilon',\delta')$

Ας ξεκινήσουμε την ανάλυση λοιπόν. Διαισθητικά το $\hat{a}$ είναι η εκτίμηση του $a\in[n]$, που είναι το κάτω άκρο του άγνωστου στηρίγματος και αντιστοίχως το $\hat{b}$. Μάλιστα θα δείξουμε ότι με πιθανότητα $1-\delta'/2$, έχουμε $X(\leq\hat{a})\in[3\epsilon'/2,5\epsilon'/2]$ και $X(\leq\hat{b})\in[1-5\epsilon'/2,1-3\epsilon'/2]$.

Αρκεί να δείξει κανείς ότι $X(\leq \hat{a}) \geq 3\epsilon'/2$ με πιθανότητα τουλάχιστον $1-\delta'/8$, μιας κι τα υπόλοιπα επιχειρήματα για τα $X(\leq \hat{a}) \leq 5\epsilon'/2$, $X(\leq \hat{b}) \leq 1-3\epsilon'/2$ και $X(\leq \hat{b}) \geq 1-5\epsilon'/2$ είναι ολόιδια. Η πιθανότητα $1-\delta'/8$ στο κάθε ένα από αυτά τα 4 με χρήση ενός union bound ολοκληρώνει τον ισχυρισμό.

Πράγματι έστω το στοιχείο $\alpha' = \max\{i \mid X(\leq i) < 3\epsilon'/2\}$. Σίγουρα, $X(\leq \alpha') < 3\epsilon'/2$ όπου $X(\leq \alpha'+1) \geq 3\epsilon'/2$.

Δεδομένου αυτού, αν $M$ το πλήθος των δειγμάτων από την $X$ τότε ο αναμενόμενος αριθμός που θα βρίσκονται $\leq \alpha'$ είναι το πολύ $3\epsilon' M/2$. Εφαρμόζοντας τα γνωστά από το 2ο κεφάλαιο Chernoff bounds, η πιθανότητα τουλάχιστον $\frac{7}{4}\epsilon' M$ δείγματα να είναι εντός της περίπτωσης $\leq \alpha'$ είναι το πολύ $e^{-(\epsilon'/4)^2 M/2} \leq \delta'/8$. Συνεπώς πέραν αυτής της πιθανότητας αποτυχίας το $\hat{a} \geq \alpha'+1$, και άρα $X(\leq \hat{a}) \geq 3\epsilon'/2$. Αν αντικαταστήσει κανείς την ποσότητα $M$, όπως την ορίσαμε θα λάβει το επιθυμητό $\delta/8$.

Συνεπώς έχουμε μια κατανομή η οποία είναι $\epsilon$-κοντά στην άγνωστη και από την οποία μπορούμε να έχουμε μια σαφή εκτίμηση του μεγέθους τους στηρίγματος. Αν τώρα $\hat{b}-\hat{a} > (C/\epsilon')^3$, τότε ο αλγόριθμος απέτυχε και εμφανίζει την προαναφερθείσα singleton κατανομή στο 1. Αλλιώς εφαρμόζουμε τον αλγόριθμο του Birgé για να μάθουμε την $X_{[\hat{a},\hat{b}]}$.

Για να εφαρμόσουμε τον αλγόριθμο του Birgé, αντλούμε δείγματα ώσπου να έχουμε $O(\log(1/\delta')\log(1/\epsilon')/\epsilon'^3)$ δείγματα εντός της περιοχής $[\hat{a},\hat{b}]$ ώστε να ικανοποιούνται οι απαιτήσεις του αλγόριθμου για μια κατανομή με στήριγμα $(C/\epsilon')^3$. Το όμορφο είναι ότι με πιθανότητα $1-\delta'/4$ μπορεί κανείς να αποκτήσει τόσα δείγματα εντός της περιοχής $[\hat{a},\hat{b}]$ χρησιμοποιώντας τον ίδιο ασυμπτωτικό αριθμό $O(\log(1/\delta')\log(1/\epsilon')/\epsilon'^3)$ από δείγματα της $X$ . Αυτό μπορεί να το δείξει κανείς και πάλι με Chernoff bounds, αλλά γενικότερα είναι αρκετά σαφές και διαισθητικό αυτό το αίτημα αφού σε αυτήν την περιοχή βρίσκεται το $1-O(\epsilon)$ της μάζας πιθανότητας της κατανομής της $X$.

Ας δούμε όμως τι συμβαίνει αν η άγνωστη PBD $X$ βρίσκεται $\epsilon'$-κοντά στην σε κάποια sparse form $Y$. Ας πούμε ότι το στήριγμα της $Y$ είναι το $\{a', \ldots, b'\}$ όπου $b'-a' \leq (C/\epsilon')^3$. Αφού η $X$ είναι $\epsilon'$-κοντά στο $Y$ ως προς την TV σίγουρα ισχύει $X(\leq a'-1) \leq \epsilon'$. Αφού $X(\leq \hat{a}) \geq 3\epsilon'/2$ ξέρουμε ότι $\hat{a} \geq a'$. Παρόμοια επιχειρήματα οδηγούν $\hat{b} \leq b'$. Συνεπώς το διάστημα $[\hat{a},\hat{b}]$ βρίσκεται εντός $[a',b']$ το οποίο έχει μέγεθος το πολύ $(C/\epsilon')^3$.

Συνεπώς ο αλγόριθμος με πιθανότητα $1-\delta'/2$ θα πετύχει να υπολογίσει μια κατανομή με ένα καλό προσεγγιστικά στήριγμα και με τουλάχιστον πιθανότητα $1-\delta'/2$ ο αλγόριθμος

του Birgé θα βρει μια κατανομή που βρίσκεται $\epsilon'$ κοντά:$d_{\mathrm{TV}}(H_S, X_{[\hat{a},\hat{b}]}) \leq \epsilon'$,

$$\begin{aligned}
2d_{\mathrm{TV}}(X, X_{[\hat{a},\hat{b}]}) = & \sum_{i\in[\hat{a},\hat{b}]} |X_{[\hat{a},\hat{b}]}(i) - X(i)| + \sum_{i\notin[\hat{a},\hat{b}]} |X_{[\hat{a},\hat{b}]}(i) - X(i)| \\
= & \sum_{i\in[\hat{a},\hat{b}]} \Big|\frac{1}{X([\hat{a},\hat{b}])}X(i) - X(i)\Big| + \sum_{i\notin[\hat{a},\hat{b}]} X(i) \\
= & \sum_{i\in[\hat{a},\hat{b}]} \Big|\frac{1}{1-O(\epsilon')}X(i) - X(i)\Big| + O(\epsilon') \\
= & \frac{O(\epsilon')}{1-O(\epsilon')} \sum_{i\in[\hat{a},\hat{b}]} \Big|X(i)\Big| + O(\epsilon') \\
= & O(\epsilon').
\end{aligned}$$

Από τριγωνική λοιπόν ανισότητα έχουμε ότι $d_{\mathrm{TV}}(H_S, X) = O(\epsilon')$.
Με αυτήν την ανάλυση ολοκληρώνεται η απόδειξη του παρακάτω θεωρήματος:

**Λήμμα 3.5.** *$\forall n, \epsilon', \delta' > 0$, υπάρχει ένας αλγόριθμος $A$ $\texttt{Learn-Sparse}^X(n, \epsilon', \delta')$ ο οποίος λαμβάνει*

$$O\left(\frac{1}{\epsilon'^3}\log\frac{1}{\epsilon'}\log\frac{1}{\delta'} + \frac{1}{\epsilon'^2}\log\frac{1}{\delta'}\log\log\frac{1}{\delta'}\right)$$

*δείγματα από την άγνωστη PBD $X$ πάνω στο $[n]$ και ο οποίος πραγματοποιεί*

$$\log n \cdot \tilde{O}\left(\frac{1}{\epsilon'^3}\log^2\frac{1}{\delta'}\right)$$

*πράξεις και εξάγει μια συμπαγή περιγραφή μιας κατανομής $H_S$ με στήριγμα $[a, b] \subseteq [n]$ μεγέθους $O(1/\epsilon'^3)$. Αν η $X$ έχει $\epsilon$-κοντά κάποια sparse formPBD $Y$ τότε ο αλγόριθμος εγγυάται ότι με πιθανότητα $1 - \delta'$ η απόσταση της $H_S$ φράσσεται $d_{\mathrm{TV}}(X, H_S) \leq c_1\epsilon'$, και η $H_S$ έχει στήριγμα εντός της $Y$.*

#### 3.3.1.2 Μαθαίνοντας την $X$ όταν είναι κοντά σε μία $k$-heavy Binomial Form PBD .

Στόχος μας είναι να εκμεταλλευτούμε την δομή του καλύμματος. Συγκεκριμένα, αν $X$ δεν είναι $\epsilon'$-κοντά σε καμία sparse form PBD στο κάλυμμα, πρέπει να είναι $\epsilon'$-κοντά με μια PBD heavy Binomial form με προσεγγιστικά ίδια μέση τιμή και διακύμανση με την $X$. Η στρατηγική είναι απλή αφού μια PBD που έχει heavy Binomial form βρίσκεται $\epsilon'$-κοντά επίσης σε μια Translated Poisson με O(1) κοντινή μέση τιμή και διακύμανση. Η μέθοδος είναι απλώς να υπολογίσουμε τους αμερόληπτους εκτιμητές $\hat{\mu}$ και $\hat{\sigma}^2$ της $X$. Το μόνο που μένει να αποδείξουμε είναι ότι μια Translated Poisson με αυτή την μέση τιμή και αυτή την διακύμανση είναι $\epsilon'$-κοντά στην $X$.

Το βασικό θεωρημά του [ΔΠ13] μας διαβεβαιώνει ότι σε αυτή την περίπτωση ισχύει ότι
:

$$\sigma^2 = \Omega(1/\epsilon'^2) \geq \theta^2 \quad \text{για κάποια σταθερά } \theta.$$

`Learn-Poisson`$^X(n, \epsilon', \delta')$

1. Έστω $\epsilon = \epsilon'/\sqrt{4 + \frac{1}{\theta^2}}$ και $\delta = \delta'$.
2. Χρησιμοποίησε τον αλγόριθμο $\mathcal{A}(n, \epsilon, \delta)$ για να εκτιμήσεις την $\mathbb{E}[X]$ και την $\mathbb{V}\mathrm{ar}[X]$.
3. Πρότεινε ως απάντηση την Translated Poisson $TP(\hat{\mu}, \hat{\sigma}^2)$.

Σχήμα 3.3: `Learn-Poisson`$^X(n, \epsilon', \delta')$.

$\mathcal{A}(n, \epsilon, \delta)$

1. Έστω $r = O(\log 1/\delta)$. Για $i = 1, \ldots, r$ επανάλαβε:
   (α′) Ζήτα $m = \lceil 3/\epsilon^2 \rceil$ ανεξάρτητα δείγματα $Z_{i,1}, \ldots, Z_{i,m}$ από την $X$.
   (β′) Έστω $\hat{\mu}_i = \frac{\sum_j Z_{i,j}}{m}$, $\hat{\sigma}_i^2 = \frac{\sum_j (Z_{i,j} - \frac{1}{m}\sum_k Z_{i,k})^2}{m-1}$.
2. Υπολόγισε την διάμεσο $\hat{\mu}$ των $\hat{\mu}_1, \ldots, \hat{\mu}_r$ και την διάμεσο $\hat{\sigma}^2$ των $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_r^2$.
3. Πρότεινε ως εκτιμήσεις τις $\hat{\mu}$ και $\hat{\sigma}^2$.

Σχήμα 3.4: $\mathcal{A}(n, \epsilon, \delta)$

**Παρατήρηση 3.6** (Median Trick)**.** *Πριν ξεκινήσουμε θα μελετήσουμε ένα από τα βασικότερα trick για να αυξήσει κανείς την πιθανότητα επιτυχίας ενός αλγορίθμου. Ας υποθέσουμε ότι ένας αλγόριθμος* A *μπορεί και εκτιμά μια πραγματική τιμή $F$ μέσω μιας τιμής $\hat{F}$, δηλαδή $|F - \hat{F}| \leq \epsilon$, με σταθερή πιθανότητα $p$. Θα ήθελα κανείς να πετύχει πιθανότητα οσοδήποτε κοντά στην μονάδα $(1-\delta)$, για οποιοδήποτε $\delta > 0$ μικρό. Μια κλασσική μέθοδος είναι να υπολογίσεις τον διάμεσο πολλών δειγμάτων-απαντήσεων του αλγόριθμου* A*.*

*Απόδειξη.* Πράγματι ας ορίσουμε:

$X_i = 1 \Leftrightarrow$ η i-οστή απάντηση του αλγόριθμου είναι εντός της εγγυούμενης περιοχής $\epsilon$

Προφανώς αν τρέξουμε $m$ πειράματα ξέρουμε ότι $\mathbb{E}[X_i] = p$ και $\mathbb{E}[\sum_{i=1}^m X_i] = mp$. Επίσης, η μέθοδος μας είναι σωστή σίγουρα όταν $\sum_i X_i > m/2$. Πράγματι αυτό σημαίνει ότι τουλάχιστον οι μισές περιπτώσεις είναι επιτυχημένες. Άρα η διάμεσος των αποτελεσμάτων

είναι μια τιμή που βρίσκεται εντός της $\epsilon$ περιοχής. Το μόνο που μένει να αποδειχθεί είναι ότι :

$$\Pr[\sum_i X_i \geq 0.5m] \geq 1-\delta$$

$$\begin{aligned}
\Pr[\sum_i X_i < 0.5m] &= \Pr[\sum_i X_i < (0.5+p)m - pm] \\
&= \Pr[\sum_i X_i - \mathbb{E}[\sum_i X_i] < (0.5-p)m] \\
&\leq \Pr[|\sum_i X_i - \mathbb{E}[\sum_i X_i]]| < (0.5-p)m] \\
&\leq \Pr[|\sum_i X_i - \mathbb{E}[\sum_i X_i]]| < (0.5-p)\frac{\mathbb{E}[\sum_i X_i]}{p}] \\
&= \Pr[|\sum_i X_i - \mathbb{E}[\sum_i X_i]]| < \mathbb{E}[\sum_i X_i]\frac{(0.5-p)}{p}] \\
&\leq 2e^{-\frac{(0.5-p)}{p}\mathbb{E}[\sum_i X_i]/3} \\
&= 2e^{-\frac{(0.5-p)}{m}/3} = 2e^{-Cm} \Rightarrow \\
m &= \Omega(\log(1/\delta))
\end{aligned}$$

όπου η τελευταία ανισότητα προκύπτει από εφαρμογή των πολλαπλασιαστικών Chernoff Bounds. □

Συνεπώς τώρα μπορούμε να μελετήσουμε πιο απλούς εκτιμητές που απλώς με πιθανότητα $2/3$ πετυχαίνουν μια καλή προσέγγιση.

**Λήμμα 3.6.** *Για κάθε $n, \epsilon, \delta > 0$, υπάρχει ένας αλγόριθμος $\mathcal{A}(n,\epsilon,\delta)$ με τις εξής ιδιότητες: Δεδομένης πρόσβασης στην άγνωστης PBD $X$ τάξεως $n$, παράγει εκτιμητές $\hat{\mu}$ και $\hat{\sigma}^2$ για τους $\mu = \mathbb{E}[X]$ και $\sigma^2 = \mathbb{V}ar[X]$ και με πιθανότητα $1-\delta$ εγγυάται ότι:*

$$|\mu - \hat{\mu}| \leq \epsilon\cdot\sigma \qquad \text{και} \qquad |\sigma^2 - \hat{\sigma}^2| \leq \epsilon\cdot\sigma^2\sqrt{4+\frac{1}{\sigma^2}}.$$

*Ο αλγόριθμος θα ζητήσει $O(\log(1/\delta))$ φορές*

$$O(1/\epsilon^2)$$

*δείγματα και θα τρέξει σε χρόνο*

$$O(\log n \log(1/\delta)/\epsilon^2).$$

*Απόδειξη.* • *$\hat{\mu}$:* Έστω $Z_1,\ldots,Z_m$ ανεξάρτητα δείγματα της $X$ και έστω $\hat{\mu} = \frac{\sum_i Z_i}{m}$. Τότε

$$\mathbb{E}[\hat{\mu}] = \mu \ \text{ και } \ \mathbb{V}\text{ar}[\hat{\mu}] = \frac{1}{m}\mathbb{V}\text{ar}[X] = \frac{1}{m}\sigma^2.$$

Από την Chebyshev έχουμε

$$\Pr[|\hat{\mu} - \mu| \geq t\sigma/\sqrt{m}] \leq \frac{1}{t^2}.$$

Διαλέγοντας $t = \sqrt{3}$ και $m = \lceil 3/\epsilon^2 \rceil$, η παραπάνω σχέση οδηγεί στο $|\hat{\mu} - \mu| \leq \epsilon\sigma$ με πιθανότητα τουλάχιστον $2/3$.

- $\sigma^2$*:* Έστω $Z_1, \ldots, Z_m$ ανεξάρτητα δείγματα της $X$ και έστω $\hat{\sigma}^2 = \frac{\sum_i (Z_i - \frac{1}{m}\sum_i Z_i)^2}{m-1}$ ο αμερόληπτος εκτιμητής της διακύμανσης. Ξέρουμε ότι :

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2 \quad \text{και} \quad \mathbb{V}\mathrm{ar}[\hat{\sigma}^2] = \sigma^4 \left( \frac{2}{m-1} + \frac{\kappa}{m} \right),$$

όπου $\kappa$ είναι η κύρτωση της κατανομής $X$ (π.χ $\kappa = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4} - 3$). Όμως έχουμε για την κύρτωση ότι:

$$\begin{aligned} \kappa &= \frac{1}{\sigma^4} \sum_i (1 - 6p_i(1-p_i))(1-p_i)p_i \qquad \text{(σεε [;])} \\ &\leq \frac{1}{\sigma^4} \sum_i (1-p_i)p_i = \frac{1}{\sigma^2}. \end{aligned}$$

Συνεπώς: $\mathbb{V}\mathrm{ar}[\hat{\sigma}^2] = \sigma^4 \left( \frac{2}{m-1} + \frac{\kappa}{m} \right) \leq \frac{\sigma^4}{m}(4 + \frac{1}{\sigma^2})$. Συνεπώς η Chebyshev οδηγεί στο ότι:

$$\Pr\left[ |\hat{\sigma}^2 - \sigma^2| \geq t\frac{\sigma^2}{\sqrt{m}} \sqrt{4 + \frac{1}{\sigma^2}} \right] \leq \frac{1}{t^2}.$$

Αν διαλέξει κανείς $t = \sqrt{3}$ και $m = \lceil 3/\epsilon^2 \rceil$, συνεπάγεται ότι $|\hat{\sigma}^2 - \sigma^2| \leq \epsilon\sigma^2\sqrt{4 + \frac{1}{\sigma^2}}$ με πιθανότητα $2/3$.

Ας συνδυάσουμε τις παραπάνω προτάσεις για να δούμε την ορθότητα του αλγορίθμου `Learn-Poisson`$^X(n, \epsilon', \delta')$ ο οποίος εκτελεί τον $\mathcal{A}(n, \epsilon, \delta)$ για κατάλληλο $\epsilon = \epsilon(\epsilon')$ και $\delta = \delta(\delta')$, και εξάγει την translated Poisson $TP(\hat{\mu}, \hat{\sigma}^2)$, όπου $\hat{\mu}$ και $\hat{\sigma}^2$ η μέση τιμή και η διακύμανση της $X$ όπως μας δόθηκε από τον $\mathcal{A}$. Επίσης θα πρέπει να βρούμε τα κατάλληλα $\epsilon, \delta$ που ικανοποιούν το ζητούμενο στόχο.

Αν η $X$ δεν είναι $\epsilon'$-κοντά σε κάποια PBD σε sparse form του καλύμματος, τότε υπάρχει μια PBD $Z$ ($k = O(1/\epsilon')$)-heavy Binomial μορφής μέσα στο κάλυμμα που είναι $\epsilon'$-κοντά στο $X$. Η ύπαρξη αυτής της $Z$ μας βοηθάει στο να φράξουμε την μέση τιμή και την διακύμανση της $X$,$\mu = \mathbb{E}[X]$ και $\sigma^2 = \mathbb{V}\mathrm{ar}[X]$:. Πράγματι η $Z \sim \mathrm{Bin}(\ell, q)$, όπου για τις παραμέτρους $\ell, q$ το [ΔΠ13] μας αποδεικνύει ότι:

(a) $\ell q \geq k^2 \cdot$

(b) $\ell q(1-q) \geq k^2 - k - 1 \cdot$

(c) $|\ell q - \mu| = O(1) \cdot$ και

$(d)$ $|\ell q(1-q) - \sigma^2| = O(1 + \epsilon' \cdot (1+\sigma^2))$.

Από την $(b), (d)$ έχουμε:

$$\sigma^2 = \Omega(k^2) = \Omega(1/\epsilon'^2) \geq \theta^2,$$

για κάποια σταθερά $\theta$. Αν διαλέξουμε $\epsilon = \epsilon'/\sqrt{4 + \frac{1}{\theta^2}}$ και $\delta = \delta'$ και λάβουμε $O(\log(1/\delta')/\epsilon'^2)$ δείγματα από εκτιμητές $\hat{\mu}$ και $\hat{\sigma}^2$ του $\mu$ και $\sigma^2$ έχουμε το επιθυμητό αποτέλεσμα.

Από την επιλογή των παραμέτρων, αν η $X$ δεν ήταν $\epsilon'$-κοντά σε καμία PBD σε sparse form μέσω στο κάλυμμα $\mathcal{S}_{\epsilon'}$, τότε με πιθανότητα τουλάχιστον $1-\delta'$ οι εκτιμητές $\hat{\mu}$ και $\hat{\sigma}^2$ ικανοποιούν:

$$|\mu - \hat{\mu}| \leq \epsilon' \cdot \sigma \quad \text{και} \quad |\sigma^2 - \hat{\sigma}^2| \leq \epsilon' \cdot \sigma^2,$$

□

Θα δείξουμε ότι αν $Y$ ακολουθεί $TP(\hat{\mu}, \hat{\sigma}^2)$, τότε $d_{\mathrm{TV}}(X, Y) \leq O(\epsilon')$

*Απόδειξη.* Έστω η PBD $X = \sum_{i=1}^{n} X_i$, με $\mathbb{E}[X_i] = p_i, \forall i$.

$$\begin{aligned}
d_{\mathrm{TV}}(X, TP(\mu, \sigma^2)) &\leq \frac{\sqrt{\sum_i p_i^3(1-p_i)} + 2}{\sum_i p_i(1-p_i)} \\
&\leq \frac{\sqrt{\sum_i p_i(1-p_i)} + 2}{\sum_i p_i(1-p_i)} \\
&\leq \frac{1}{\sqrt{\sum_i p_i(1-p_i)}} + \frac{2}{\sum_i p_i(1-p_i)} \\
&= \frac{1}{\sigma} + \frac{2}{\sigma^2} \\
&= O(\epsilon'). \qquad (3.8)
\end{aligned}$$

Αρκεί τώρα να φράξουμε την απόσταση μεταξύ των $TP(\mu, \sigma^2)$ και $TP(\hat{\mu}, \hat{\sigma}^2)$. Έχουμε λοιπόν:

$$\begin{aligned}
d_{\mathrm{TV}}(TP(\mu, \sigma^2), TP(\hat{\mu}, \hat{\sigma}^2)) &\leq \frac{|\mu - \hat{\mu}|}{\min(\sigma, \hat{\sigma})} + \frac{|\sigma^2 - \hat{\sigma}^2| + 1}{\min(\sigma^2, \hat{\sigma}^2)} \\
&\leq \frac{\epsilon'\sigma}{\min(\sigma, \hat{\sigma})} + \frac{\epsilon' \cdot \sigma^2 + 1}{\min(\sigma^2, \hat{\sigma}^2)} \\
&\leq \frac{\epsilon'\sigma}{\sigma/\sqrt{1-\epsilon'}} + \frac{\epsilon' \cdot \sigma^2 + 1}{\sigma^2/(1-\epsilon')} \\
&= O(\epsilon') + \frac{O(1-\epsilon')}{\sigma^2} \\
&= O(\epsilon') + O(\epsilon'^2) \\
&= O(\epsilon') \qquad (3.9)
\end{aligned}$$

Εφαρμόζοντας την τριγωνική ανισότητα επιβεβαιώνεται ο τελικός μας ισχυρισμός. □

**Λήμμα 3.7.** *Για κάθε $n, \epsilon', \delta' > 0$, υπάρχει αλγόριθμος* `Learn-Poisson`$^X(n, \epsilon', \delta')$ *η οποία αντλεί*

$$O(\log(1/\delta')/\epsilon'^2)$$

*δείγματα από μια άγνωστη PBD $X$ πάνω $[n]$ και εκτελεί*

$$O(\log n \cdot \log(1/\delta')/\epsilon'^2)$$

*υπολογιστικές πράξεις και επιστρέφει δύο παραμέτρους $\hat{\mu}$ και $\hat{\sigma}^2$. Επίσης ο αλγόριθμος εγγυάται ότι αν η $X$ δεν έχει sparse γείτονα τότε η κατανομή $H_P = TP(\hat{\mu}, \hat{\sigma}^2)$ με πιθανότητα τουλάχιστον $1 - \delta'$ θα έχει απόσταση $d_{\mathrm{TV}}(X, H_P) \leq c_2\epsilon'$ για μια σταθερά $c_2 \geq 1$.*

**Παρατήρηση 3.7.** *Για τυπικούς λόγους θα πρέπει κανείς να βρει ένα τρόπο κωδικοποίησης της translated poisson. Για περισσότερες πληροφορίες παραπέμπουμε στο Appendix C του [ΔΔΣ15].*

#### 3.3.1.3 Ανακεφαλαίωση

Ο αλγόριθμος αποτελείται από τρία μέρη. Στο (α) μέρος ελέγχουμε αν το εκτιμώμενο στήριγμα της κατανομής είναι αρκετά μικρό. Δεδομένου ότι η κατανομή είναι μονόλοφη και με μικρό στήριγμα αρκεί να τρέξουμε τον αλγόριθμο του Birgé και αποθηκεύουμε την απάντηση. Στο (β) μέρος υπολογίζουμε την αμερόληπτη εκτίμηση της μέσης τιμής και της διακύμανσης της άγνωστης κατανομής. Υπολογίζουμε με αυτό τον τρόπο μια κατανομή με αντίστοιχη μέση τιμή και διακύμανση και αποθηκεύουμε την απάντηση. Η καθεμία από τις δύο εκπροσωπεί την καλύτερη κατανομή που μπορεί να προέρχεται από τις δύο ομάδες του καλύμματος. Στο (γ) και τελευταίο μέρος τρέχουμε έναν απλό διαγωνισμό μεταξύ τους για να κρίνουμε ποια είναι η κοντινότερη της.

### 3.3.2 Proper Learning

Για να κάνουμε τον αλγόριθμο να παράγει σίγουρα PBD , θα πάρουμε το προηγούμενο αλγόριθμο και θα εφαρμόσουμε τα εξής modifications. Θα δούμε αρχικώς την πιο μαθηματική περίπτωση αυτή της heavy form

#### 3.3.2.1 Μαθαίνοντας την $X$ όταν είναι κοντά σε μία $k$-heavy Binomial Form PBD properly.

Για αυτή την περίπτωση, θεωρούμε ότι έχουμε ήδη εκτελέσει τον αλγόριθμο `Learn-Poisson`

- Η `Locate-Binomial`$(\hat{\mu}, \hat{\sigma}^2, n)$:

Αυτή η ρουτίνα δέχεται ως εισόδους τις παραμέτρους $(\hat{\mu}, \hat{\sigma}^2)$ της `Learn-Poisson`$^X(n, \epsilon, \delta)$ και υπολογίζει μια Binomial $H_B$, χωρίς επιπλέον δείγματα της $X$. Αυτό που θα δείξουμε

`Locate-Binomial`$(\hat{\mu}, \hat{\sigma}^2, n)$

1. Αν $\hat{\sigma}^2 \leq \frac{n}{4}$, θέτουμε $\sigma_1^2 = \hat{\sigma}^{2\cdot}$ αλλιώς $\sigma_1^2 = \frac{n}{4}$.
2. Αν $\hat{\mu}^2 \leq n(\hat{\mu} - \sigma_1^2)$, θέτουμε $\sigma_2^2 = \sigma_1^{2\cdot}$ αλλιώς $\sigma_2^2 = \frac{n\hat{\mu} - \hat{\mu}^2}{n}$.
3. Επέστρεψε την κατανομή $H_B = \mathrm{Bin}(\hat{n}, \hat{p})$, όπου $\hat{n} = \left\lfloor \hat{\mu}^2/(\hat{\mu} - \sigma_2^2) \right\rfloor$ και $\hat{p} = (\hat{\mu} - \sigma_2^2)/\hat{\mu}$.

Σχήμα 3.5: `Locate-Binomial`$(\hat{\mu}, \hat{\sigma}^2, n)$.

ότι χρησιμοποιώντας το θεώρημα του καλύμματος από το [ΔΠ13] είναι ότι αν η $X$ δεν είναι $\epsilon$-κοντά σε κάποια sparse τότε $d_{\mathrm{TV}}(X, H_B) < O(\epsilon)$.

Πράγματι, έστω $\mu$ και $\sigma^2$ η μέση τιμή και η διακύμανση της $X$ και έστω η $X$ δεν είναι $\epsilon$-κοντά σε κάποια sparse form PBD από το κάλυμμα $S_\epsilon$. Στην προηγούμενη ανάλυση είδαμε ότι με πιθανότητα $1 - \delta$, οι τιμές $(\hat{\mu}, \hat{\sigma}^2)$ της `Learn-Poisson`$^X(n, \epsilon, \delta)$ παράγουν μια translated poisson αρκετά γειτονική στην $X$, δηλαδή $d_{\mathrm{TV}}(X, TP(\hat{\mu}, \hat{\sigma}^2)) = O(\epsilon)$

Η ρουτίνα αυτή έχει τρία βασικά βήματα.

1. **Tweaking $\hat{\sigma}^2$:** Αν $\hat{\sigma}^2 \leq \frac{n}{4}$, θέτουμε $\sigma_1^2 = \hat{\sigma}^{2\cdot}$ αλλιώς $\sigma_1^2 = \frac{n}{4}$. Σαν διαίσθηση πρέπει να κρατήσει κανείς ότι η μεγαλύτερη δυνατή τιμή διακύμανσης για μια διωνυμική κατανομή $\mathrm{Bin}(n, \cdot)$ είναι $n/4$.. Παρ΄ όλα αυτά και στις δύο περιπτώσεις έχουμε:

   $$(1 - \epsilon)\sigma^2 \leq \sigma_1^2 \leq (1 + \epsilon)\sigma^2,$$

   όπου το κάτω φράγμα το αντλούμε από το γεγονός ότι κάθε PBD ικανοποιεί $\sigma^2 \leq \frac{n}{4}$. Στη συνέχεια θα δείξουμε ότι αυτός ο καθορισμός του $\sigma_1^2$ έχει ως αποτέλεσμα

   $$d_{\mathrm{TV}}(TP(\hat{\mu}, \hat{\sigma}^2), TP(\hat{\mu}, \sigma_1^2)) \leq O(\epsilon).$$

   Πράγματι:

   - Αν $\hat{\sigma}^2 \leq \frac{n}{4}$ , τότε αυτή είναι απόσταση είναι μηδέν
   - Διαφορετικά, $(1 + \epsilon)\,\sigma^2 \geq \hat{\sigma}^2 > \sigma_1^2 = \frac{n}{4} \geq \sigma^2$,

   $$\begin{aligned} d_{\mathrm{TV}}(TP(\hat{\mu}, \hat{\sigma}^2), TP(\hat{\mu}, \sigma_1^2)) &\leq \frac{|\hat{\sigma}^2 - \sigma_1^2| + 1}{\hat{\sigma}^2} \\ &\leq \frac{\epsilon\sigma^2 + 1}{\sigma^2} = O(\epsilon), \end{aligned} \tag{3.10}$$

   ενώ χρησιμοποιούμε και πάλι ισχύει $\sigma^2 = \Omega(1/\epsilon^2)$ από το (3.3.1.2).

2. **Tweaking $\sigma_1^2$:** Αν $\hat{\mu}^2 \leq n(\hat{\mu} - \sigma_1^2)$ ισοδύναμα ($\sigma_1^2 \leq \frac{n\hat{\mu} - \hat{\mu}^2}{n}$), θέτουμε $\sigma_2^2 = \sigma_1^{2\cdot}$ αλλιώς $\sigma_2^2 = \frac{n\hat{\mu} - \hat{\mu}^2}{n}$.

Σαν διαίσθηση πρέπει να κρατήσει κανείς ότι η $\mathrm{Bin}(n,\cdot)$ ως κατανομή με μέση τιμή $\hat{\mu}$ δεν μπορεί σαν διακύμανση να ξεπεράσει $\frac{n\hat{\mu}-\hat{\mu}^2}{n}$.

Θα αποδείξουμε ότι αυτό οδηγεί στην:

$$d_{\mathrm{TV}}(TP(\hat{\mu},\sigma_1^2),TP(\hat{\mu},\sigma_2^2)) \leq O(\epsilon).$$

Πράγματι:

- Αν $\hat{\mu}^2 \leq n(\hat{\mu}-\sigma_1^2)$, η απόσταση είναι μηδενική.
- Διαφορετικά,
  - Αρχικά παρατηρείστε ότι $\sigma_1^2 > \sigma_2^2$ και $\sigma_2^2 \geq 0$, όπου η τελευταία σχέση προκύπτει από την ιδιότητα $\hat{\mu} \leq n$ από κατασκευής.
  - Στην συνέχεια, υποθέστε ότι $X = PBD(p_1,\ldots,p_n)$. Τότε από Cauchy-Schwarz έχουμε:

$$\mu^2 = \left(\sum_{i=1}^n p_i\right)^2 \leq n\left(\sum_{i=1}^n p_i^2\right) = n(\mu-\sigma^2).$$

    Εναλλακτικώς:

$$\frac{\mu(n-\mu)}{n} \geq \sigma^2.$$

    Συνεπώς έχουμε ότι:

$$\begin{aligned}\sigma_2^2 = \frac{n\hat{\mu}-\hat{\mu}^2}{n} &\geq \frac{n(\mu-\epsilon\sigma)-(\mu+\epsilon\sigma)^2}{n}\\ &= \frac{n\mu-\mu^2-\epsilon^2\sigma^2-\epsilon\sigma(n+2\mu)}{n}\\ &\geq \sigma^2-\frac{\epsilon^2}{n}\sigma^2-3\epsilon\sigma\\ &\geq (1-\epsilon^2)\sigma^2-3\epsilon\sigma \geq (1-O(\epsilon))\sigma^2\end{aligned}$$

  - Από τα παραπάνω έχουμε:

$$\begin{aligned}d_{\mathrm{TV}}(TP(\hat{\mu},\sigma_1^2),TP(\hat{\mu},\sigma_2^2)) &\leq \frac{\sigma_1^2-\sigma_2^2+1}{\sigma_1^2}\\ &\leq \frac{(1+\epsilon)\sigma^2-(1-O(\epsilon))\sigma^2+1}{(1-\epsilon)\sigma^2} = O(\epsilon)\end{aligned}$$

3. Θα κατασκευάσουμε μια διωνυμική κατανομή $H_B$ η οποία θα είναι $O(\epsilon)$-κοντά στην $TP(\hat{\mu},\sigma_2^2)$. Αν το επιτύχουμε αυτό, αφού η $d_{\mathrm{TV}}(H_P,X) = O(\epsilon)$ και η $d_{\mathrm{TV}}(H_B,X) = O(\epsilon)$ τελειώσαμε.

   Η $H_B$ είναι της μορφής $\mathrm{Bin}(\hat{n},\hat{p})$, όπου

$$\hat{n} = \left\lfloor \hat{\mu}^2/(\hat{\mu}-\sigma_2^2)\right\rfloor \quad \text{ανδ} \quad \hat{p} = (\hat{\mu}-\sigma_2^2)/\hat{\mu}.$$

Παρατηρείστε ότι όπως επιλέγεται το $\sigma_2^2$ έχουμε ότι $\hat{n} \leq n$ και $\hat{p} \in [0,1]$, όπως ακριβώς στο αρχικό θεώρημα.

Ας μελετήσουμε την $d_{\mathrm{TV}}\left(\mathrm{Bin}(\hat{n},\hat{p}), TP(\hat{\mu},\sigma_2^2)\right)$.

$$\begin{aligned} & d_{\mathrm{TV}}(\mathrm{Bin}(\hat{n},\hat{p}), TP(\hat{n}\hat{p}, \hat{n}\hat{p}(1-\hat{p})) \\ \leq & \frac{1}{\sqrt{\hat{n}\hat{p}(1-\hat{p})}} + \frac{2}{\hat{n}\hat{p}(1-\hat{p})}. \end{aligned} \tag{3.11}$$

Παρατηρείστε ότι:

$$\begin{aligned} \hat{n}\hat{p}(1-\hat{p}) &\geq \left(\frac{\hat{\mu}^2}{\hat{\mu}-\sigma_2^2} - 1\right)\left(\frac{\hat{\mu}-\sigma_2^2}{\hat{\mu}}\right)\left(\frac{\sigma_2^2}{\hat{\mu}}\right) \\ &= \sigma_2^2 - \hat{p}(1-\hat{p}) \geq (1-O(\epsilon))\sigma^2 - 1 \\ &\geq \Omega(1/\epsilon^2), \end{aligned}$$

Συνεπώς:

$$d_{\mathrm{TV}}(\mathrm{Bin}(\hat{n},\hat{p}), TP(\hat{n}\hat{p}, \hat{n}\hat{p}(1-\hat{p})) = O(\epsilon).$$

Ας συγκρίνουμε τώρα τις $TP(\hat{n}\hat{p}, \hat{n}\hat{p}(1-\hat{p}))$, $TP(\hat{\mu},\sigma_2^2)$ κάτω από την TV .

$$\begin{aligned} & d_{\mathrm{TV}}(TP(\hat{n}\hat{p}, \hat{n}\hat{p}(1-\hat{p})), TP(\hat{\mu},\sigma_2^2)) \\ \leq & \frac{|\hat{n}\hat{p}-\hat{\mu}|}{\min(\sqrt{\hat{n}\hat{p}(1-\hat{p})},\sigma_2)} + \frac{|\hat{n}\hat{p}(1-\hat{p})-\sigma_2^2|+1}{\min(\hat{n}\hat{p}(1-\hat{p}),\sigma_2^2)} \\ \leq & \frac{1}{\sqrt{\hat{n}\hat{p}(1-\hat{p})}} + \frac{2}{\hat{n}\hat{p}(1-\hat{p})} \\ = & O(\epsilon). \end{aligned}$$

Από τριγωνική ανισότητα έχουμε το τελικό μας αποτέλεσμα:

$$d_{\mathrm{TV}}(\mathrm{Bin}(\hat{n},\hat{p}), TP(\hat{\mu},\sigma_2^2)) = O(\epsilon)$$

#### 3.3.2.2 Μαθαίνοντας την $X$ όταν είναι κοντά σε μία sparse form PBD properly.

Σε αυτή την περίπτωση θα κάνουμε κάτι αρκετά διαισθητικά σαφές. Θα μιμηθούμε και πάλι την τακτική που ακολουθήσαμε στην non-proper περίπτωση. Θα εφαρμόσουμε την ίδια περικοπή και θα επιδιώξουμε και πάλι να εκτιμήσουμε το στήριγμα όπως και πριν. Αν και πάλι εκτιμήσουμε στήριγμα μεγαλύτερο του $C/\epsilon^3$, θεωρούμε πάλι ότι αστοχήσαμε και εξάγουμε μια τετριμμένη κατανομή. Η στρατηγική μας θα είναι απλή. Κατασκευάζουμε το ένα κομμάτι του καλύμματος αναλυτικά, όλων των $(\frac{1}{\epsilon})^{\log^2(\frac{1}{\epsilon})}$. Στην συνέχεια θα εφαρμόσουμε την στρατηγική του Tournament και θα προτείνουμε την νικήτρια κατανομή, όπως το μελετήσαμε στο 2ο κεφάλαιο.

#### 3.3.2.3 Ανακεφαλαιώνοντας properly.

Συνεπώς ο αλγόριθμος μας είναι ο ακόλουθος. Εφαρμόζουμε διαδοχικά τα βήμα του non-proper learning αλγορίθμου,`Learn-PBD`, με τις εξής αλλαγές: Στην πρώτη περίπτωση αντί του αλγορίθμου του Birgé, εφαρμόζουμε μια μερική κατασκευή του καλύμματος και εφαρμόζουμε ένα Tournament μεταξύ των κατανομών. Ο νικητής αποθηκεύεται ως πιθανός εκπρόσωπος της sparse περίπτωσης. Στην δεύτερη περίπτωση εκτελούμε πλήρως τα βήματα του non-proper learning αλγορίθμου και στην συνέχεια μετατρέπουμε με ασφαλή τρόπο την translated poisson σε heavy Binomial όπως ορίζει το [ΔΠ13]. Τέλος εφαρμόζουμε και πάλι τον αλγόριθμο του Choose-Hypothesis.

### 3.3.3 Ποία ήταν η ιδέα πίσω από όλα;

Τι είναι όμως αυτό που ξεχώρισε αυτήν την αλγοριθμική διαδικασία σε σχέση με ότι παρατηρήσαμε έως τώρα στην πορεία αυτής της διπλωματικής; Ας δούμε τις περιπτώσεις των Birgé και Pearson με τα ευφυή ιστογράμματα που πρότειναν. Η ουσιαστική ιδέα είναι η στοιχειώδης δειγματοληψία και η ευφυής στρογγυλοποίηση. Στην πραγματικότητα στην περίπτωση του ιστογράμματος η στρογγυλοποίηση είναι η ενιαία διακριτοποίηση σε όλο τον χώρο. Στην περίπτωση του Birgé εφαρμόζουμε μια λογαριθμική στρογγυλοποίηση ώστε να τονίσει περιοχές με μεγαλύτερη μάζα πιθανότητας. Το κρίσιμο στην δουλειά των Δασκαλάκη και Παπαδημητρίου και αντιστοίχως των Δασκαλάκη , Servedio , Διακονικόλα είναι ο τρόπος ανάλυσης του μαθηματικού αντικείμενου που επιδιώκει κανείς να μάθει.

Το δομικό θεώρημα του [ΔΠ13] μας επιτρέπει δύο θεμελιακές σκέψεις:

- Ποία είναι τα πραγματικά διαφορετικά clusters του κόσμου που προσπαθούμε να μάθουμε;
- Τι μορφή, τι μέγεθος και ποία είναι η πιο συμπαγείς αναπαράσταση αυτών clusters ;

Ειδικότερα στην περίπτωση των PBDs οι απαντήσεις ήταν οι εξής:

- Υπάρχει μια κατηγορία PBDs που μπορείς να τις αντιμετωπίζεις επιθετικά με κάποια ενιαία ολιγο-παραμετρική κατανομή σε πλήρη αντιστοιχία με το κεντρικό οριακό θεώρημα.
- Επίσης υπάρχει μια κατηγορία PBDs που δεν εμφανίζουν κάποια συγκλητική συμπεριφορά, αλλά το μέγεθος τους είναι αρκετά μικρό ώστε να μπορείς να τις προσεγγίσεις από την εξυπνότερη διακριτοποίηση που μπορείς να κατασκευάσεις για αυτήν.

Το δομικό αυτό θεώρημα μας ανοίγει και τον δρόμο στο πως μπορείς να μάθεις αυτές τις κατανομές.

- Υπολόγισε/Εκτίμησε τις λίγες παραμέτρους και προσδιόρισε τις γενικευμένες κατανομές που προκύπτουν από κάποια παραλλαγή κεντρικού οριακού θεωρήματος.
- Αν οι κατανομές έχουν μικρό στήριγμα , ακόμα και οι brute-force αλγόριθμοι μπορούν να φανούν αποδοτικοί και αν μπορείς να κρατήσεις όσο πιο μικρή το κάλυμμα, τότε οι

επαναλήψεις ακόμη και των πιο απλών αλγορίθμων συναθροίστηκα επιβαρύνουν λίγο την πολυπλοκότητα του αλγορίθμου.

## 3.4 Επεκτάσεις

Το πλαίσιο αυτό είναι που ενέπνευσε τους Δασκαλάκη , Servedio , Διακονικόλα, Paul & Gregory Valiant και αρκετών ακόμα να πετύχουν μια πλειάδα αποτελεσμάτων. Στόχος μας είναι να παρουσιάσουμε σε αυτήν την ενότητα μια επισκόπηση τέτοιων αποτελεσμάτων. Στόχος είναι σε αυτό το σημείο ο αναγνώστης να βρει μια ευρεία αναφορά σε αυτά τα διαφορετικά προβλήματα και βιβλιογραφικά τις κυριότερες δουλειές γύρω από αυτό

### 3.4.1 Weighted-PBDs [ΔΔΣ15]

Έστω ότι θέτουμε $X = \sum_i a_i X_i$. Οι Daskalakis,Servedio,Diakonikolas έδειξαν πως μπορεί να επεκτείνει κανείς με τετριμμένο τρόπο το κάλυμμα προτείνοντας παράλληλα επεκτάσεις των προηγούμενων αλγορίθμων με προφανή τρόπο. Συγκεκριμένα έδειξαν ότι:

**Θεώρημα 3.7 (Learning sums of weighted independent Bernoulli random variables).** *Αν η $X = \sum_{i=1}^{n} a_i X_i$ μια weighted-PBD . Τότε υπάρχει αλγόριθμος δεδομένου $n, \epsilon, \delta$, $a_1, \ldots, a_n$ και πρόσβασης σε ανεξάρτητα δείγματα της $X$, χρησιμοποιεί*

$$\widetilde{O}(k/\epsilon^2) \cdot \log(n) \cdot \log(1/\delta)$$

*δείγματα από την $X$, και τρέχει σε χρόνο*

$$poly\left(n^k \cdot \epsilon^{-k \log^2(1/\epsilon)}\right) \cdot \log(1/\delta),$$

*ο οποίος με πιθανότητα $1-\delta$ εξάγει $n$ μεταβλητές $\hat{X}_i$ με $\mathbb{E}[\hat{X}_i] = \hat{p}_i$ ώστε $d_{\mathrm{TV}}(\hat{X}, X) \leq \epsilon$, όπου $\hat{X} = \sum_{i=1}^{n} a_i \hat{X}_i$.*

### 3.4.2 SIIRV via Cover[ΔΔO$^+$13]

Η επόμενη επέκταση που εμφανίστηκε σε αυτό το μοντέλο υπήρξε το άθροισμα II-RVs, δηλαδή ανεξάρτητων ακέραιων τυχαίων μεταβλητών που μπορούν να πάρουν τιμές στο $[0, k-1]$ Και σε αυτή την περίπτωση υπάρχει ένα δομικό θεώρημα που διακρίνει πάλι σε δύο περιπτώσεις το άθροισμα τέτοιων μεταβλητών.

Σε αυτή την περίπτωση έχουμε ότι:

**Θεώρημα 3.8 (Learning sums of weighted independent Bernoulli random variables).** *Αν η $X = \sum_{i=1}^{n} X_i$ μια Sum of IIRVs. Τότε υπάρχει αλγόριθμος δεδομένου $n, \epsilon, \delta$, $a_1, \ldots, a_n$ και πρόσβασης σε ανεξάρτητα δείγματα της $X$, χρησιμοποιεί $poly(k/\epsilon)$ δείγματα από την $X$, και τρέχει σε χρόνο $poly(k/\epsilon)$ ο οποίος με πιθανότητα $1-\delta$ εξάγει είτε μια τυχαία μεταβλητή με στήριγμα το πολύ $\frac{k^9}{\epsilon^4}$ είτε σε μια τυχαία μεταβλητή της μορφής $cZ + Y$, όπου $c \in [k-1]$ και $U$ μια c-IRV και $Z$ μια διακριτοποιημένη κανονική κατανομή με παραμέτρους $\mu_Z = \mu_X/c$ και $\sigma_Z^2 = \frac{\sigma_X^2}{c^2}$.*

Βλέπουμε ότι και σε αυτή την περίπτωση, είτε έχουμε μια τυχαία μεταβλητή με ένα αναλυτικό στήριγμα είτε έχουμε μια μεταβλητή που είναι σε θέση να εκφράσει την συνολική συνάθροιση.

### 3.4.3 SIIRV via Fourier[ΔΚΣ15β]

Μια από τις πλέον ανατρεπτικές μεθόδους που πρότειναν οι Διακονικόλας , Kane και Stewart ήρθε να δώσει μια εξαιρετικά αποδοτικότερη απάντηση σε αυτό το πρόβλημα.

Σε αυτή την εργασία ο αλγόριθμος φέρεται υπολογιστικά ισχυρότερος αφού από το αδιανόητο από πρακτικής άποψης $poly(k/\epsilon)$ προσδιορίζουν ένα σαφές αραιό κάλυμμα της τάξεως $(k/\epsilon)^3$. Η διαφορά που διακρίνει αυτή την δουλεία είναι η εκμετάλλευση της αραιής δομής που εμφανίζουν αυτές οι οικογένειες κατανομών στο χώρο των συχνοτήτων. Συγκεκριμένα το κάλυμμα που κατασκευάζουν είναι μεγέθους $(\frac{1}{\epsilon})^{k\log(1/\epsilon)}$

Η σημαντική διαφορά σε σχέση με τις υπόλοιπες περιπτώσεις είναι ότι ο αλγόριθμος εγγυάται μια $\epsilon$-προσέγγιση αρχικά στον χώρο του Fourier και στην συνέχεια γίνεται ορθό calibrate ώστε να παραχθεί μια εξίσου καλή κατανομή στο χώρο των κατανομών.

### 3.4.4 PMD via Cover Approximation [ΔΚΤ15]

Μια σπουδαία γενίκευση των Daskalakis,Kamath,Tzamos και De είναι η μελέτη τους πάνω στις PMDs. Οι multinomial κατανομές αποτελούν την γενίκευση από άποψη διάστασης του παρόντος κεφαλαίου.

Μια $(n,k)$-Poisson Multinomial Distribution (PMD) είναι η κατανομή του αθροίσματος $n$ ανεξάρτητων τυχαίων διανυσμάτων τα οποία λαμβάνουν τιμές στην standard βάση $B_k = \{e_1, \ldots, e_k\}$ του $\mathbb{R}^k$. Στην πρώτη τους δουλειά [ΔΚΤ15] έδειξαν ότι για κάθε $\epsilon > 0$, κάθε $(n,k)$-PM τυχαίο διάνυσμα είναι $\epsilon$-κοντά κάτω από TV στο άθροισμα μίας στρογγυλοποιημένες πολυδιάστατης Gaussian κατανομής και ενός ανεξάρτητου μικρού στηρίγματος $(poly(k/\epsilon), k)$ PM διανύσματος. Το σημαντικό σε αυτήν την εργασία είναι ότι στην προσπάθεια κατασκευής ενός καλύμματος από κατανομές όπως αυτές που αναφέρθηκαν επέκτειναν το κεντρικό οριακό θεώρημα που πρότειναν οι αδερφοί Valiant. Η μέχρι τώρα δουλειά ήταν πλήρως εξαρτώμενη από στατιστικά στοιχεία όπως την μικρότερη ιδιοτιμή του πίνακα συνδιακύμανσης. Αυτή η generic εργασία βελτίωσε το ουσιαστικό μέγεθος του καλύμματος για την ειδική περίπτωση των PBDs .

### 3.4.5 PMDs via Fourier Approximation [ΔΚΣ15α]

Παρόμοια εργασία με το SIIRVS και παράλληλα με τους Daskalakis,Kamath,Tzamos,De και πάλι οι Diakonikolas,Stewart και Kane εφάρμοσαν την στρατηγική υπολογισμού και πάλι του διακριτού μετασχηματισμού Fourier διαμορφώνοντας το δικό τους επεκτεταμένο κεντρικό οριακό θεώρημα από το αρχικό των Gregory & Paul Valiant. Η διαίσθηση πίσω από την εργασία βρίσκεται στο ότι το συχνοτικό περιεχόμενο βρίσκεται συγκεντρωμένο σε μεγάλο ποσοστό σε ένα πολύ μικρό σύνολο συχνοτήτων. Συνεπώς μπορεί κανείς να μάθει αναλυτικά αυτό το τμήμα πληροφορίας και να διακριτοποιήσει το υπόλοιπο τμήμα του μετασχηματισμού. Κρίσιμο στοιχείο και στις δύο εργασίες, αποτελεί ο στατιστικός υπολογισμός μέσω δειγμάτων του ίδιου του μετασχηματισμού.

### 3.4.6 a size-Free CLT for PMD [ΔΔΚΤ16]

Οι μέχρι τώρα εργασίες στις PMDs περιορίζονταν πολλές φορές από την διάσταση του χώρου. Σε αυτή την εργασία οι Daskalakis,De,Tzamos,Kamath εκμεταλλεύονται την γνώση της αλγεβρικής γεωμετρίας και του φασματικού περιεχομένου της εργασίας των Diakonikola,Kane,Stewart και κατασκευάζουν ένα πλήρως καινούργιο κεντρικό οριακό θεώρημα το οποίο τους οδηγεί στην κατασκευεί του βέλτιστου δυνατού καλύμματος. Οι τεχνικές που χρησιμοποιούν συνδιάζουν προηγούμενα κεντρικά οριακά θεωρήματα όπως των αδερφών Valiant των Shapley-Folkman και των μοντέρνων τεχνικών sparsification σε Laplacian πίνακες από τους Batson, Spielman, and Srivastava.

### 3.4.7 Mixture of Gaussians [ΔΚ14α]

Ο Δασκαλάκης και ο Kamath μελετώντας την σχέση της Kolmogorov και της TV απόστασης και χρησιμοποιώντας τα Robust Statistics σχεδίασαν αποδοτικούς αλγορίθμους μάθησης για μείγματα Gaussian και μάλιστα για πρώτη φορά απαλλαγμένα από συνθήκες και υποθέσεις διάκρισης των κεντρών των διαφορετικών Gaussian. Σε αυτή την εργασία εμφανίζεται για πρώτη φορά βελτίωση του αλγορίθμου Tournament.

### 3.4.8 k-Modal [ΔΔΣ14]

Καθώς οι Daskalakis,Servedio,Diakonikolas ανέσυραν από το χρονοντούλαπο της επιστημονικής βιβλιογραφίας τον αλγόριθμο του Birgé διαπίστωσαν ότι η εξαιρετικά εύστοχη τεχνική του στον υπολογισμό των modes μέσω δυαδικών αναζητήσεων μπορούσε να δώσει ένα πολύ πιο γενικό αλγόριθμο εντοπισμού ακροτάτων σε μια κατανομή. Σε συνδυασμό με το απλό και optimal τρόπο υπολογισμού μονότονων κατανομών επέτρεψε τον σχεδιασμό ενός αποδοτικού αλγορίθμου μάθησης μιας k-λόφη κατανομή. Στην περίπτωση μάλιστα όπου $k < O(\log n)$ ο αλγόριθμος φαίνεται να είναι πληροφοριο-θεωρητικά βέλτιστος.

### 3.4.9 Κλάσεις Δομημένων Κατανομών [ΔΣΣ13]

Η εργασία αυτή βρίσκεται στο μεταίχμιο της αλγοριθμικής θεωρίας μάθησης και της θεωρίας πολυπλοκότητας. Σε αυτή την εργασία μελετάται η χρήση των tournament και η διαχείριση καλυμμάτων με ένα πλήρως generic τρόπο.

Συγκεκριμένα έστω $C$ μια κλάση πιθανοτικών κατανομών πάνω στο διακριτό χώρο $[n] = 1, ..., n$. Στην δουλειά αυτή αποδεικνύεται ότι αν η $C$ ικανοποιεί μια γενική συνθήκη, συγκεκριμένα αν κάθε κατανομή στην $C$ μπορεί να προσεγγιστεί με μια μεταβλητή με πιο συμπαγές ιστόγραμμα τότε υπάρχει ένας εξαιρετικά αποδοτικός αλγόριθμος από άποψη δειγμάτων και χρόνου ο οποίος μπορεί να μάθει οποιοδήποτε μείγμα $k$ άγνωστων από αυτές τις κατανομές. Παραδείγματα κλάσεων όπως log-concave, monotone hazard rate, unimodal κατανομών έχουν αυτήν την δομική περιγραφή από ιστογράμματα με λίγες θέσεις αποκτώντας αποδοτικούς αλγορίθμους για τα μείγματα τους. Βασικό εργαλείο τους αποτελεί το Tournament του προηγουμένου κεφαλαίου.

# Κεφάλαιο 4

# Learning The Truth

## 4.1 Υπολογιστική Θεωρία Κοινωνικής Επιλογής

Με το πέρασμα από τον 20ο στον 21ο αιώνα, οι αλγόριθμοι από ζήτημα εσωτερικής κατανάλωσης αποτέλεσε κεντρικό ζήτημα εφαρμογής τόσο στις θετικές όσο και στις θεωρητικές επιστήμες. Εκεί τοποθετεί κανείς και την εμφάνιση την υπολογιστικής κοινωνιολογίας ή ακριβέστερα της υπολογιστικής θεωρίας κοινωνικής επιλογής. Για να δούμε όμως πως η επιστήμη της κοινωνιολογίας χρειάζεται πλέον όχι απλά ως εξωτερικό εργαλείο αλλά ως εσωτερικό μηχανισμό της την Πληροφορική, θα ταξιδέψουμε λίγο στο χρόνο και θα δούμε πως ο τρόπος εκλογής μιας απόφασης για ένα κοινωνικό σύνολο αποτέλεσε θέμα ενδιαφέροντος πολλών μαθηματικών και προσφάτως αρκετών computer scientists.

Ας γνωρίσουμε λοιπόν καλύτερα αυτή την θεωρία

### Εισαγωγή

Η θεωρία κοινωνικής επιλογής είναι η μελέτη των συλλογικών διαδικασιών και των διαδικασιών λήψης απόφασης. Δεν είναι μια ενιαία θεωρία, αλλά ένα σύμπλεγμα από τα μοντέλα και αποτελέσματα σχετικά με το άθροισμα των επιμέρους ατομικών εισόδων (π.χ. την ατομική ψήφο, την ατομική προτίμηση, την ατομική απόφαση). Αυτά τα συλλογικά αθροίσματα , τις συλλογικές εξόδους (π.χ. συλλογικές αποφάσεις, συλλογικές προτιμήσεις, συλλογικές αποφάσεις) επιθυμεί μια κοινωνία να αποκτήσει προς το καλύτερο δυνατό στόχο της γενικής πρόνοια.

Κεντρικά ερωτήματα της θεωρίας είναι:

- Πώς μπορεί μια ομάδα ατόμων να επιλέξει συλλογικά το βέλτιστο αποτέλεσμα από ένα δεδομένο σύνολο των επιλογών;
- Ποιες είναι οι ιδιότητες των διαφόρων συστημάτων ψηφοφορίας;
- Πότε είναι ένα σύστημα ψηφοφορίας δημοκρατικό;
- Πώς μπορεί ένα συλλογικό σώμα (π.χ. το εκλογικό σώμα, νομοθέτες, συλλογικό δικαστήριο, ομάδα εμπειρογνωμόνων, ή επιτροπή) να φτάσει με συνεκτικό τρόπο σε

συλλογικές κρίσεις σε ορισμένα θέματα, με βάση τις ατομικές προτιμήσεις ή τις αποφάσεις των μελών της;

- Πώς μπορούμε να κατατάξουν διαφορετικές κοινωνικά εναλλακτικές λύσεις για διαφορετικούς συλλογικούς στόχους;

Η πρώτη σύνδεση της με τα μαθηματικά και τις θετικές επιστήμες ξεκίνησαν τον 18ο αιώνα. Τότε είναι και που πρωτοστάτησε ο Nicolas de Condorcet και ο Jean-Charles de Borda και τον 19ο αιώνα εκφράστηκε από τον Charles Dodgson (επίσης γνωστός ως Lewis Carroll). Είναι η πρώτη περίοδος όπου οι θεωρητικοί της κοινωνικής επιλογής μελετούν τα ζητήματα αυτά όχι μόνο κοιτάζοντας τα παραδείγματα, αλλά με την ανάπτυξη γενικών προτύπων και αποδεικνύουν θεωρήματα. Στον 20ο αιώνα με τα έργα του Kenneth Arrow, Amartya Sen, και Duncan Black η θεωρία της κοινωνικής επιλογής απογειώθηκε. Η επιρροή τους εκτείνεται σε ολόκληρη την οικονομία, την πολιτική επιστήμη, τη φιλοσοφία, τα μαθηματικά, και, πρόσφατα, της επιστήμης των υπολογιστών και της βιολογίας. Εκτός από τη συμβολή στην κατανόηση των διαδικασιών συλλογικής απόφασης, η θεωρία της κοινωνικής επιλογή έχει εφαρμογές στους τομείς του θεσμικού σχεδιασμού, στην οικονομία της κοινωνικής ευημερίας και της κοινωνικής επιστημολογίας.

### 4.1.1 1η Στάση. Ο Condorcet

Ο Condorcet ήταν φιλελεύθερος στοχαστής στην εποχή της Γαλλικής Επανάστασης ο οποίος όμως καταδιωκόταν από τις επαναστατικές αρχές γιατί τους επέκρινε στις λανθασμένες αποφάσεις τους. Μετά από μια περίοδο όπου είχε κρυφτεί, τελικώς συνελήφθη, και πέθανε στη φυλακή. Στο δοκίμιό του για την Εφαρμογή της Μαθηματικής Ανάλυσης στην θεωρία των πιθανοτήτων και στα μοντέλα πλειοψηφικής λήψης αποφάσεων (1785), υποστήριξε ένα συγκεκριμένο σύστημα ψηφοφορίας, ένα σύστημα κατά ζεύγη πλειοψηφικό, και παρουσίασε τις δύο πιο σημαντικές ιδέες του.

Η πρώτη, γνωστή ως το θεώρημα των ενόρκων του Condorcet, είναι ότι αν κάθε μέλος της κριτικής επιτροπής έχει μια ίση και ανεξάρτητη πιθανότητα καλύτερη από τυχαία, αλλά χειρότερη από τέλεια, $\Pr[\text{Correct Desicion}] = p \in (1/2, 1)$ , να λάβει μια σωστή απόφαση σχετικά με το αν ο κατηγορούμενος είναι ένοχος (ή σε κάποια άλλη πραγματική πρόταση), η πλειοψηφία των ενόρκων είναι πιο πιθανό να είναι σωστή από κάθε ένορκο ξεχωριστά, και η πιθανότητα μιας σωστής απόφασης πλειοψηφίας πλησιάζει το 1 όσο το μέγεθος της κριτικής επιτροπής αυξάνει. Έτσι, υπό ορισμένες προϋποθέσεις, ο κανόνας της πλειοψηφίας είναι καλός στη «παρακολούθηση της αλήθειας» . Αντίθετα αν η μεγάλη πλειοψηφία των ψηφοφόρων χαρακτηρίζονται από εσφαλμένη κρίσης, δηλαδή $\Pr[\text{Correct Desicion}] = p \in [0, 1/2]$, η καλύτερη τακτική είναι να διαλέξεις κάποιον ένορκο τυχαία και να δικάζεις με βάση την απόφαση του.

Η δεύτερη διορατικότητα του Condorcet, συχνά ονομάζεται παράδοξο του Condorcet, είναι η παρατήρηση ότι πλειοψηφία των προτιμήσεων μπορεί να είναι «παράλογη» (συγκεκριμένα, αμετάβατη) ακόμα και όταν ατομικές προτιμήσεις είναι «ορθολογικές» (συγκεκριμένα, μεταβατικές). Ας υποθέσουμε, για παράδειγμα, ότι το πρώτο 1/3 από μια ομάδα προτιμά εναλλακτική λύση $x$ μετά το $y$ μετά το $z$, ένα δεύτερο 1/3 προτιμά $y$ μετά το $z$ μετά το $x$,

και ένα τελευταίο 1/3 προτιμά $z$ μετά το $x$ και μετά το $y$. Με βάση την ανα δύο πλειοψηφία, υπάρχουν πλειοψηφίες (2/3) για το $x$ προς $y$, του $y$ προς $z$, και για το $z$ προς το $x$.

Ένας «κύκλος», ο οποία παραβιάζει την μεταβατικότητα του υπέρτατου νικητή. Συνεπώς, δεν υπάρχει νικητής Condorcet, δηλαδή μια επιλογή που να κερδίζει, ή τουλάχιστον να έρχεται σε ισοπαλία με κάθε άλλη επιλογή σε διαγωνισμούς πλειοψηφίας κατά ζεύγη.

Ο Condorcet διαπίστωσε πολύ νωρίς το καίριο θέμα της σύγχρονης θεωρίας κοινωνικής επιλογής:

> Ο πλειοψηφικός κανόνας εκλογής είναι ταυτόχρονα μια εύλογη μέθοδο της συλλογικής λήψης αποφάσεων αλλά υπόκεινται σε ορισμένα εξαιρετικά απροσδόκητα προβλήματα.

Η επίλυση ή η παράκαμψη αυτών των προβλημάτων παραμένει μία από τις βασικές ανησυχίες της σύγχρονης θεωρίας κοινωνικής επιλογής.

### 4.1.2 2η Στάση. Το θεώρημα του Arrow

#### 4.1.2.1 Πρόλογος

Στην θεωρία της κοινωνικής επιλογής σημαίνον ρόλο παίζει επίσης το σπουδαίο θεώρημα του Arrow. Ακόμα και αν σε αυτή η διπλωματική αντιμετωπίζει το πεδίο της κοινωνικής επιλογής ως ένα πεδίο εφαρμογής των υπολογιστικών εργαλείων που προσφέρει η Πληροφορική, θα ήταν λάθος πριν εμβαθύνουμε στο μοντέλο ψηφοφοριών και συνδέσουμε το κεφάλαιο αυτό με τα προηγούμενα να μην αναφερθούμε στις βασικές αρχές αυτού του θεωρήματος. Το θεώρημα του Arrow αποτελεί ένα από τα βασικά θεωρήματα μη ικανοποιησιμότητας περιορισμών στα μοντέλα των εκλογών.

Το θεώρημα του Arrow ορίζει ότι για οποιαδήποτε μη τετριμμένη ψηφοφορία, πέραν των δύο επιλογών, κανένας αλγόριθμος συνάθροισης των ψήφων δεν μπορεί να ανακηρύξει νικητή που να συνδέει τις ατομικές προτιμήσεις των ψηφοφόρων σε μια κοινή κοινωνική διάταξη προτιμήσεων χωρίς να παραβιάσει μια σειρά απο απαιτήσεις-κριτήρια που συνήθως τίθενται ως ιδιότητες μιας καλώς ορισμένης ψηφοφορίας. Με άλλα λόγια, στόχος μας είναι να εξάγουμε μια διάταξη προτίμησης κάποιων δεδομένων επιλογών λαμβάνοντας ως είσοδο την προσωπική επιλογή του κάθε ψηφοφόρου. Στόχος μας είναι να βρούμε έναν αλγόριθμο εκλογής νικητή, έναν μηχανισμό εκλογής, ένα κανόνα συνάθροισης των ψήφων, το οποίο μετατρέπει μια ομάδα από ψήφους-προτιμήσεις σε μια μοναδική καθολική διάταξη προτίμησης. Αξιωματικά ο κανόνας αυτός θα πρέπει για να διατηρεί την έννοια της δικαιοσύνης οφείλει να διατηρήσει κάποιες βασικές αρχές.

#### 4.1.2.2 Αξιώματα του Arrow

- Καθολικότητα (Universality).
  Για οποιοδήποτε σύνολο ατομικών επιλογών προτίμησης, ο εκλογικός κανόνας θα πρέπει κάθε φορά που του δίνεται η ίδια κάλπη να απαντάει οριστικά την ίδια ακριβώς νικήτρια διάταξη επιλογών. Επίσης η διάταξη θα πρέπει να είναι ολική και όχι μερική.

- Δημοκρατικότητα (Non-DictatorShip)
  Ο εκλογικός κανόνας θα πρέπει να στηρίζεται στην προτίμηση όλων των ψηφοφόρων και όχι αποκλειστικά ενός

- Ανεξαρτησία ασυσχέτιστων επιλογών (Independence of Irrelevant Alternatives)
  Η αρχή αυτή αφορά την εσωτερική δομή της νικήτριας διάταξης. Η διάταξη μετάξυ δύο υποψήφιων επιλογών $x, y$ θα πρέπει να εξαρτάται αποκλειστικά από τις ατομικές προτιμήσεις των ψηφοφόρων σε σχέση με αυτά τα δύο στοιχεία $x, y$. Αυτό σημαίνει ότι αν η νικήτρια διάταξη ορίζει ότι $x > y$ [1] θα πρέπει να χρησιμοποιεί μόνο την ατομική πληροφορία από κάθε ψηφοφόρο για το αν $x > y$ ή $y < x$ και να μην χρησιμοποιεί πληροφορίες για άλλα στοιχεία που είναι ασυσχέτιστα με τα $x, y$. Με άλλα λόγια, αν στις ατομικές ψήφους του καθενός ισχύει ότι $x > y$ τότε αν αλλάξουμε την ενδιάμεση διάταξη άλλων επιλογών η τελική διάταξη θα πρέπει και πάλι να διατηρεί την $x > y$. Για παράδειγμα, η είσοδος ενός τρίτου υποψηφίου σε μια εκλογή που συμμετέχουν δύο συμμετέχοντες μέχρι εκείνη την στιγμή, δεν θα πρέπει να επηρεάσει την μεταξύ διάταξη των πρώτων δύο.

- Μονοτονία (Monotonicity)
  Καμία υποψήφια επιλογή δεν γίνεται να βρεθεί ψηλότερα αν έστω και ένα ψηφοφόρος αλλάξει την προτίμηση του σε αυτή την επιλογή προς το χειρότερο. Αντιστρόφως, καμία επιλογή δεν γίνεται να βρεθεί σε χαμηλότερη θέση αν κάποιος ψηφοφόρος αλλάξει την προτίμηση του αρνητικά σε αυτή την επιλογή προς το καλύτερο

- Απροκαταληψία (Non-imposition)
  Ο εκλογικός νόμος για κάθε δυνατή διάταξη θα πρέπει να υπάρχει μια κάλπη που να τον οδηγεί να την ανακηρύσσει νικήτρια.

- Pareto-βελτιστότητα ή Ομοφωνία. (Pareto-optimality or Unanimity)
  Αν όλοι οι ψηφοφόροι προτείνουν την ίδια διάταξη θα πρέπει να είναι και η νικήτρια διάταξη από τον εκλογικό νόμο.

Το θεώρημα του Arrow αποδεικνύει ότι δεν υπάρχει εκλογικός νόμος που να μπορεί να συμβιβάσει όλες τις παραπάνω ιδιότητες.

Σήμερα οι περισσότεροι θεωρητικοί της κοινωνικής επιλογής έχουν προχωρήσει πέραν από τις πρώτες αρνητικές ερμηνείες του θεωρήματος του Αρροω και ενδιαφέρονται για τις αλγοριθμικές συνέπειες για τα πιθανά trade-offs που εμπλέκονται στην εξεύρεση ικανοποιητικής διαδικασίας λήψης αποφάσεων. Ο Sen έχει προωθήσει αρκετά σε αυτή την ερμηνεία

[1] Υπό την έννοια ότι η τελική διάταξη είναι $x > other_1 > other_2 > \cdots > other_n > y$

της θεωρίας κοινωνικής επιλογής (στην δική του ομιλία του για την απονομή του Βραβείου Νόμπελ, το 1998).

Στο πλαίσιο αυτής της προσέγγισης, η αξιωματική μέθοδος του Αρροω έχει ίσως ακόμη μεγαλύτερη επιρροή από ό, τι η ίδια η επιφανειακή αδυναμία που προτείνει το θεώρημα του . Πλέον λοιπόν στόχος είναι να εντοπίσει κανείς μια σειρά από εύλογες αναγκαίες και επαρκείς συνθήκες που μοναδικά χαρακτηρίζουν μια συγκεκριμένη λύση (ή μια κατηγορία λύσεων) σε ένα συγκεκριμένο είδος προβλήματος συλλογικής απόφασης. Ένα πρώιμο παράδειγμα είναι το θεώρημα χαρακτηρισμού του Kenneth το 1952 πάνω στο κανόνα πλειοψηφίας. Το σημαντικό σημείο της προσφοράς του Arrow που κατατάσσει το έργο του στα σπερματικά έργα του πεδίου είναι η τακτική του να μην ασχοληθεί με ένα εκλογικό νόμο-κανόνα αλλά αντίθετα με μια γενική ομάδα κανόνων συνάθροισης προτιμήσεων και των αξιωματικών ηθικών αρχών που θα ήθελε οι κανόνες αυτές να ικανοποιούν.

### 4.1.3 3η Στάση. Ο Borda και οι φίλοι του

Ο Condorcet και ο Arrow δεν είναι οι μόνες ιδρυτικές προσωπικότητες της θεωρίας κοινωνικής επιλογής. Ο σύγχρονος και συμπατριώτης του Condorcet, ο Jean-Charles de Borda (1733-1799) υπερασπίστηκε ένα σύστημα ψηφοφορίας που θεωρείται συχνά ως εξέχων εναλλακτική λύση για την πλειοψηφία. Η καταμέτρηση Borda, τυπικά ορίζεται αργότερα, αποφεύγει το παράδοξο Condorcet, αλλά παραβιάζει μία από τις προϋποθέσεις του Αρροω, την ανεξαρτησία των ασυσχέτιστων επιλογών. Έτσι, η συζήτηση μεταξύ Condorcet και Borda είναι ένας πρόδρομος για ορισμένες σύγχρονες συζητήσεις σχετικά με το πώς θα πρέπει να ανταποκριθούν οι ερευνητές στο θεώρημα του Αρροω.

Η προέλευση αυτής της συζήτησης προηγείται του Condorcet και Borda.

Κατά το Μεσαίωνα, ο Ramon Llull (1235-1315) πρότεινε την πλειοψηφική μέθοδο ψηφοφορίας σε συνάθροιση κατά ζεύγη, ενώ ο Nicolas Cusanus (1401-1464) πρότεινε μια παραλλαγή του κανόνα του Borda. Το 1672, ο Γερμανός πολιτικός και λόγιος Samuel von Pufendorf (1632-1694) πρότεινε σε σύγκριση με την απλή πλειοψηφία, μια ειδική πλειοψηφία, και τους κανόνες της ομοφωνίας και πρόσφερε μια ανάλυση της δομής των προτιμήσεων που μπορεί να θεωρηθεί ως πρόδρομος ύστερων μοντέλων όπως η Single-peakedness.

Τον 19ο αιώνα, ο Βρετανός μαθηματικός και κληρικός Charles Dodgson (1832-1898), περισσότερο γνωστός ως Lewis Carroll, ανεξάρτητα ξανά-ανακάλυψε πολλά από τις ιδέες του Condorcet και του Borda, αλλά και ανέπτυξε μια θεωρία της αναλογικής εκπροσώπησης.

### 4.1.4 Ο Maximum Likelihood Estimator

Μια διαφορετική προσέγγιση για να αντιμετωπίσει κανείς την ενδογενή αδυναμία των ψηφοφοριών είναι να θεωρήσει κανείς ότι οι εκλογικοί κανόνες λειτουργούν ως εκτιμητές. Σε αυτό το μοντέλο υπάρχει μια υποβόσκουσα αλήθεια, "βαθιά χωμένη στην γη". Στόχος του κάθε ψηφοφόρου είναι να προσεγγίσει αυτήν την αλήθεια. Συνεπώς κάθε ψηφοφόρος αποτελεί ένα θορυβημένο εκτιμητή αυτής της αλήθειας. Οι αλγόριθμοι αυτής της φιλοσοφίας ταξινομούνται ποιοτικά όσο αποδίδουν μεγαλύτερη πιθανότητα στην υποβόσκουσα αλήθεια σε νικήτρια. Οι καλύτεροι εκλογικοί νόμοι αποτελούν οι εκτιμητές μέγιστης πιθανοφάνειας

της υποβόσκουσα αλήθειας. Αυτή ακριβώς είναι και η ιδέα του Condorcet στο θεώρημα των ενόρκων.

Ο βασικότερος λόγος όπου όλες αυτές οι τεχνικές απέκτησαν πάλι την δυναμική που είχε η Υπολογιστική Θεωρία κοινωνικής επιλογής τον 18ο αιώνα είναι η καινοτομία του crowd sourcing. Συγκεκριμένα η παρέα του Condorcet στο πέρασμα των χρόνων στόχευαν αποκλειστικά στο να αποδείξουν ότι για το κανόνα που προτείνουν καθώς αυξάνει το πλήθος των ψηφοφόρων συγκλίνει την πιθανότητα της βέλτιστης εκλογής στην μονάδα. Εδώ είναι και που η Θεωρητική Πληροφορική έρχεται να συμβάλει με τον δικό της τρόπο. Όσο κι αν οι μαθηματικοί, οι οικονομολόγοι ή κοινωνικοί μελετητές επιβεβαιώνουν τα solution concept που παρουσιάζουν στέλνοντας το πλήθος των ψηφοφόρων στο άπειρο και αποδεικνύοντας ότι οι εκλογικοί μηχανισμοί που έχουν κατασκευάσει συγκλίνουν, η πραγματικότητα είναι διαφορετική. Γιατί η διαφορετικότητα είναι μη ασυμπτωτική και πεπερασμένη. Συνεπώς θέλουμε να ξέρουμε την ταχύτητα με την οποία ο κάθε μηχανισμός συγκλίνει στην βέλτιστη επιλογή της κοινωνίας.

- Ποία είναι η ποσότητα της πληροφορίας που απαιτείται να παραδώσει ο ψηφοφόρος ώστε ο μηχανισμός μη ασυμπτωτικά να συγκλίνει στηνν βέλτιστη λύση;
- Ποία είναι η ποσότητα των ψηφοφόρων που χρειαζόμαστε για να βρεθούμε $\epsilon$ κοντά στην βέλτιστη κοινωνικά απάντηση;
- Ποία είναι η ποσότητα υπολογιστικών κινήσεων που χρειαζόμαστε για να κάνουμε εφικτό τον στόχο μας;

## 4.2 Ορισμοί - Εκλογικά Μοντέλα

### 4.2.1 Μοντέλα Ψηφοφοριών

**Ορισμός 4.1. *Υποψήφιοι Ψηφοφορίας.*** *Θεωρούμε ένα σύνολο $A = \{a_1, a_2, a_3, \cdots, a_m\}$ από καθήκοντα τα οποία πρέπει να φέρουμε εις πέρας ή υποψήφιους τους οποίος θέλουμε να κατατάξουμε.*

**Ορισμός 4.2. *Ψηφοδέλτιο.*** *Η ψήφος του κάθε συμμετέχοντα αποτελεί μια 1-1 και επί συνάρτηση $\sigma : A \to \{1, 2, \cdots, m\}$.*

Με βάση αυτή την μοντελοποίηση μπορούμε να θεωρήσουμε την κάθε ψήφο ως μια αναδιάταξη των στοιχείων του $A$. Αν έχουμε δύο καθήκοντα $a, b$ και ισχύει $\sigma(a) < \sigma(b)$ τότε το καθήκον $a$ είναι προτιμότερο από το $b$. Στο εξής αυτό θα το συμβολίζουμε ως $a >_\sigma b$ .

$$\sigma(a) < \sigma(b) \Leftrightarrow a >_\sigma b$$

**Ορισμός 4.3. *Σύνολο Ψηφοδελτίων.*** *Το σύνολο όλων των διαφορετικών συναρτήσεων $\sigma$ θα καλείται $\mathcal{L}(A)$.*

**Ορισμός 4.4. *Κάλπη.*** *Μια κάλπη ή ένα χαρτοφυλάκιο $n$ ψήφων θα καλείται $\pi \in \mathcal{L}(A)^n$*

**Ορισμός 4.5.** ***Ντετερμινιστικός Εκλογικός Νόμος*** $n$ ***ψήφων*** *Ένας ντετερμινιστικός κανόνας εκλογής $VotingRule_n$ είναι μια συνάρτηση που δέχεται ως είσοδο μια κάλπη $n$ ψήφων και εξάγει κάποια ψήφο νικήτρια. Συνοπτικά:*

$$VotingRule_n : \mathcal{L}(A)^n \to \mathcal{L}(A)$$

.

**Ορισμός 4.6.** ***Ντετερμινιστικός Εκλογικός Νόμος*** *Ένας πλήρως ορισμένος ντετερμινιστικός κανόνας εκλογής $VotingRule$ αποτελεί την ένωση κανόνων εκλογής ώστε για κάθε διαφορετικό αριθμό ψήφων να ενεργοποιείται διαφορετικός κανόνας και να εξάγεται κάποια νικήτρια.*

$$VotingRule : \cup_{n\geq 1}\mathcal{L}(A)^n \to \mathcal{L}(A)$$

**Παρατήρηση 4.1.** *Παρατηρείστε ότι η έξοδος μιας εκλογής δεν είναι αποκλειστικά ένα αποτέλεσμα, αλλά μια ολόκληρη διάταξη των αντικειμένων.*

**Ορισμός 4.7.** ***Τυχαιοκρατικός Εκλογικός Νόμος*** *Ένας πλήρως ορισμένος τυχαιοκρατικός κανόνας εκλογής $RVotingRule$ αποτελεί την ένωση κανόνων εκλογής ώστε για κάθε διαφορετικό αριθμό ψήφων να ενεργοποιείται διαφορετικός κανόνας και να εξάγεται κάποια νικήτρια κατανομή.*

$$RVotingRule : \cup_{n\geq 1}\mathcal{L}(A)^n \to D(\mathcal{L}(A))$$

*όπου $D(\cdot)$ συμβολίζει το σύνολο όλων των κατανομών πάνω στο σύνολο των ψήφων-αναδιατάξεων*

**Παρατήρηση 4.2.** $\Pr[RVotingRule(\pi) = \sigma]$ *συμβολίζει την πιθανότητα ο κανόνας $RVotingRule$ με είσοδο την κάλπη $pi$ να εξάγει την νικήτρια αναδιάταξη $\sigma$.*

Τώρα προχωρούμε σε έναν ακόμα ορισμό. Ένας βαθμολογικός κανόνας εκλογής βασίζεται σε μια κάλπη και ένα βαθμολογική αντιστοίχιση ανάλογα της θέσης που προτείνει η κάθε ψήφος σε ένα υποψήφιο καθήκον.

**Ορισμός 4.8.** ***Βαθμολογικός Κανόνας.*** *Ένας βαθμολογικός κανόνας καθορίζεται από ένα διάνυσμα $\vec{v} = (v_1, v_2, \cdots, v_m)$ ώστε για μια ψήφο $\sigma$ να αποδίδονται στο $i-$οστό στοιχείο της κατάταξης, $\sigma^{-1}(i)$ $v_i$ πόντοι ψήφου.*

**Ορισμός 4.9.** ***Βαθμολογικός Εκλογικός Νόμος.*** *Εφαρμόζοντας έναν βαθμολογικό κανόνα πάνω σε ένα σύνολο ψήφων και αθροίζοντας τους πόντους της κάθε ψήφου καταλήγουμε σε μια σταθμισμένη κατάταξη καθηκόντων.*

**Παρατήρηση 4.3.** *Παραδείγματα Βαθμολογικών Εκλογικών Νόμων Παραδείγματα τέτοιων συστημάτων είναι :*

1. *Ο εκλογικός νόμος στα περισσότερα κράτη, γνωστός και ως ο κανόνας της πλειοψηφίας, όπου το διάνυσμα βαθμών είναι :*

$$\vec{v} = (1, 0, 0, \cdots, 0)$$

2. *Ο εκλογικός νόμος που χρησιμοποιείται στην βαθμολογική κατάταξη των τραγουδιών της Eurovision, γνωστός και ως Borda, όπου το διάνυσμα βαθμών είναι:*

$$\vec{v} = (m, m-1, m-2, \cdots, 1)$$

3. *Ο εκλογικός νόμος που χρησιμοποιείται στην συνδιάσκεψη του ΟΗΕ για έγκριση πολεμικών επιχειρήσεων, γνωστός και ως Veto, όπου το διάνυσμα βαθμών είναι :*

$$\vec{v} = (1, 1, \cdots, 1, 0)$$

4. *Ο αρμονικός εκλογικός νόμος που αποδίδει διάνυσμα βαθμών είναι :*

$$\vec{v} = (1, 1/2, 1/3, 1/4, \cdots, 1/m)$$

**Ορισμός 4.10.** ***Κανόνας του Kemeny.*** *Δοθείσης μιας κάλπης $\pi = (\sigma_1, \sigma_2, \cdots, \sigma_n) \in \mathcal{L}(A)^n$, ο Kemeny επιλέγει την διάταξη $\sigma$ που ελαχιστοποιεί την $\sum_{i=1}^{n} d_{Kendall-Tau}(\sigma, \sigma_i)$*

Πριν εξηγήσουμε την απόσταση του Kendall-Tau, ας δούμε τι εκφράζει ο κανόνας. Ο κάθε υποψήφιος δείχνει την προτίμηση του καταθέτοντας μία ψήφο-ένα διάνυσμα-μια διάταξη-μια οντότητα. Η επιλογή να ελαχιστοποιήσουμε την $L_1$ νόρμα πάνω σε αυτόν τον μετρικό χώρο αντιστοιχεί στο να βρούμε την διάμεσο του χώρου.

**Λήμμα 4.1** (Ελαχιστοποίηση της $L_1$)**.** *Ας υποθέσουμε ότι έχουμε $n$ στοιχεία $S = s_1, s_2, \cdots, s_n \in \mathbb{R}$, τότε $l_S(x) = \sum_{i=1}^{n} \|x - s_i\|_1$ ελαχιστοποιείται από την* ***διάμεσο*** *των στοιχείων.*

*Απόδειξη.* 1. Έστω $x : s_k \leq x \leq x + \epsilon \leq s_{k+1}$ τότε:

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

2. Επίσης έχουμε $l_S(x+\epsilon) = \sum_{i=1}^{n} \|x - s_i\|_1 = \sum_{i=1}^{k} (x + \epsilon - s_i) + \sum_{i=k+1}^{n} (s_i - x - \epsilon) =$ $\epsilon(2k - n) + \sum_{i=1}^{k} (x - s_i) + \sum_{i=k+1}^{n} (s_i - x)$.

3. Από τα παραπάνω προκύπτει ότι $(l_S(x+\epsilon) - l_S(x))/\epsilon = (2k - n)$

4. Ας πάρουμε το όριο για $\epsilon \to 0$ και έτσι μπορούμε να υπολογίσουμε την παράγωγο της $l'_S(x) = (2k - n), x \in [s_k, s_{k+1}]$

5. Έτσι μπορούμε να μελετήσουμε την μονοτονία της $l_S(x)$. Πράγματι αν $\begin{cases} n > k/2, & \text{Αυξουσα} \\ n < k/2, & \text{Φθίνουσα} \\ n = k/2, & \text{Σταθερή} \end{cases}$

6. Συνεπώς για $k = n/2$ η συνάρτηση εμφανίζει ελάχιστο.

□

Συνεπώς ο κανόνας του Kemeny αναζητεί την διάμεσο στο μετρικό χώρο της Kendall-Tau. Αξίζει να παρατηρήσει κανείς ότι αν υπάρχει απόλυτη πλειοψηφία σε ένα στοιχείο, τότε η διάμεσος θα είναι αυτό το στοιχείο, άρα ο κανόνας του Kemeny υποστηρίζει τις περιπτώσεις της απολύτου πλειοψηφίας.

**Ορισμός 4.11.** *Απόσταση tau Δεδομένων δύο διατάξεων $\sigma_1, \sigma_2$ έχουμε ότι*

$$d_{KT}(\sigma_1, \sigma_2) = |\{(a, b) : ((b >_{\sigma_1} a) AND (a >_{\sigma_2} b)) OR ((a >_{\sigma_1} b) AND (b >_{\sigma_2} a))\}|$$

Με απλά λόγια η $KT$ απόσταση μεταξύ δύο διατάξεων είναι ο αριθμός των ελάχιστων swaps που χρειάζεται να γίνουν ώστε η μια διάταξη να ταυτιστεί με την άλλη. Διαφορετικά η $KT$ υπολογίζει τον αριθμό των ζευγαριών που δεν υπάρχει κοινή προτίμηση[2].

Επειδή υπάρχει περίπτωση ισοπαλίας θα επεκτείνουμε τον κανόνα του Kemeny στην τυχαιοκρατική έκδοση όπου για κάθε φορά που υπάρχει ισοπαλία εφαρμόζουμε μια δίκαιη ομοιόμορφα τυχαία επιλογή, δηλαδή κάθε διάταξη που ανήκει στο σύνολο των διατάξεων $\arg\min_{\sigma \in \mathcal{L}(a)} \sigma_{i=1}^{n} d_{KT}(\sigma, \sigma_i)$ απολαμβάνει την ίδια πιθανότητα εκλογής.

**Παρατήρηση 4.4.** *Η απόσταση $KT$ μπορεί να δει κανείς ότι ικανοποιεί όλους τους όρους μιας νόρμας*

1. $d_{KT}(\sigma_1, \sigma_2) \geq 0$
2. $d_{KT}(\sigma_1, \sigma_2) = 0 \Leftrightarrow \sigma_1 \equiv \sigma_2$
3. $d_{KT}(\sigma_1, \sigma_2) = d_{KT}(\sigma_2, \sigma_1)$
4. $d_{KT}(\sigma_1, \sigma_2) \leq d_{KT}(\sigma_1, \sigma_2) + d_{KT}(\sigma_2, \sigma_3)$

*Μια επιπλέον πολύ σημαντική ιδιότητα είναι η ανώνυμια. Η απόσταση μεταξύ των δύο διατάξεων πάνω στα αντικείμενα είναι πλήρως ανεξάρτητη της μετονομασίας που μπορεί να συμβεί πάνω στα αντικείμενα.*

[2]Για παράδειγμα αν $\sigma_1 = (1, 2, 3, 4), \sigma_2 = (1, 3, 4, 2)$ υπάρχει κοινή προτίμηση στο $1, 4$ και στις δύο ψήφους αφού $a_1 >_{\sigma_1} a_4$ και $a_1 >_{\sigma_2} a_4$ ενώ δεν υπάρχει κοινή προτίμηση στο $2, 3$ και στις δύο ψήφους αφού $a_2 >_{\sigma_1} a_3$ και $a_3 >_{\sigma_2} a_2$

### 4.2.2 Μοντέλα Ψηφοφόρων

Στόχος αυτής της ενότητας είναι να ορίσουμε το μοντέλο συμπεριφοράς και επιλογής των ψηφοφόρων.

**Ορισμός 4.12.** *Ground Truth Ας υποθέσουμε ότι υπάρχει μια κρυφή σωστή σειρά των καθηκόντων $\sigma^* \in \mathcal{L}(A)$. Θα υποθέτουμε χωρίς βλάβη της γενικότητας ότι το $\sigma^*(a_i) = i$.*

Χρησιμοποιώντας ως υπόθεση ότι υπάρχει λοιπόν μια κρυφή αλήθεια, οι κοινωνιολόγοι μοντελοποιούν τον κάθε ψηφοφόρο ως ένα θόρυβο γύρω από αυτήν. Επίσης κάθε ψηφοφόρος έχει ως βασικό στόχο να ανακαλύψει αυτή την κρυφή αλήθεια με βάση την γνώση και μόρφωση που έχει. Η ψήφος του κάθε συμμετέχοντα διαμορφώνεται λοιπόν ως μια τυχαία μεταβλητή. Η κατανομή της τυχαίας αυτής μεταβλητής πάνω στο σύνολο όλων των διατάξεων έχει ως κέντρο την ground truth και ενισχύει με περισσότερη αναλογικά μάζα πιθανότητας τις διατάξεις που ομοιάζουν περισσότερα στην κρυφή αλήθεια.

### 4.2.3 Ορισμός του Mallow Model

` Το μοντέλο στο οποίο θα επενδύσουμε στην μελέτη μας είναι ένα από τα απλούστερα θορυβικά μοντέλα. Ας υποθέσουμε ότι κάθε συμμετέχον χαρακτηρίζεται από μια παράμετρο $p \in [0, 1]$. Έτσι για κάποια δεδομένη $\sigma^*$ διαμορφώνεται μια κατανομή πιθανότητας $\Pr_{\sigma^*}[\sigma]$ από κάθε ψηφοφόρο. Ο αλγόριθμος με τον οποίο διαμορφώνεται η ψήφος του κάθε παίκτη είναι η ακόλουθη:

- Mallow Model
    1. Έστω όλα τα δυνατά $\binom{m}{2}$ ζεύγη καθηκόντων.
    2. Για κάθε ζεύγος $(a, b)$ ρίχνουμε ένα κέρμα και με πιθανότητα $p$ ακολουθούμε την διάταξη που υπάρχει στο $a >^*_\sigma b$ αλλιώς χρησιμοποιούμε την ανάποδη.
    3. Αν στο τέλος προκύπτει ολική διάταξη χωρίς κύκλους και παράδοξα, ορίζεται η διάταξη αυτή ως ψήφος
    4. Διαφορετικά ο αλγόριθμος επανεκκινεί την διαδικασία από την αρχή.
    5. Κάθε επιλογή γίνεται ανεξάρτητη των προηγούμενων.

Μπορεί κανείς να δει ότι αυτό το απλό μοντέλο μας οδηγεί στην ακόλουθη κατανομή πιθανότητας: $\Pr_{\sigma^*}[\sigma] = C * p^{\binom{m}{2} - d_{KT}(\sigma,\sigma^*)}(1-p)^{d_{KT}(\sigma,\sigma^*)}$, όπου $C$ κάποια σταθερά κανονικοποίησης. Ξαναγράφωντας την παραπάνω σχέση θα λέμε ότι:

$$\Pr_{\sigma^*}[\sigma] = \frac{\phi^{d_{KT}(\sigma,\sigma^*)}}{Z^m_\phi}$$

, όπου $Z^m_\phi$ η σταθερά κανονικοποίησης και $\phi = \frac{1-p}{p}$. Μπορεί κανείς εύκολα με επιχειρήματα συμμετρίας ότι η σταθερά κανονικοποίησης είναι ανεξάρτητη της όποιας συγκεκριμένης αλήθειας $\sigma^*$. Για λόγους συμβολισμού επισημαίνουμε δύο σχέσεις-αντίστοιχες της μάζας και της συνάρτησης πιθανότητας

- $p_{i,j} = \sum_{\sigma \in \mathcal{L}(A)} \Pr_{\sigma^*}[\sigma] \times \mathbb{1}\{\sigma(a_i) = j\}$
- $q_{i,j} = \sum_{k=1}^{j} p_{i,k}$

Τέλος η πιθανότητα της κάθε κάλπης, αφού θεωρήσουμε ότι οι ψήφοι είναι ανεξάρτητοι έχουμε

$$\Pr_{\sigma^*}[\pi] = \prod_{i=1}^{n} \Pr_{\sigma^*}[\sigma_i]$$

Αυτό ακριβώς το μοντέλο χρονολογείται πίσω στην 1η Στάση μας στον Condorcet. Ο ίδιος απέδειξε μάλιστα ότι ο εκλογικός κανόνας του Kemeny που προσπαθεί να βρει τον διάμεσο στο μετρικό χώρο των Kendall-Tau αποστάσεων είναι εκτιμητής μέγιστης πιθανοφάνειας της υποβόσκουσας αλήθειας αν ξέρουμε ότι οι ψηφοφόροι ακολουθούν το Mallow Model.

### 4.2.4 Η απόσταση $d_{KT}$ & ο κανόνας του Kemeny

Ας δούμε λοιπόν πως ορίζεται η πιθανοφάνεια μιας κάλπης. Η πιθανοφάνεια μιας κάλπης $\pi$ ορίζεται ως $\mathcal{L}_\pi(\sigma^*) = \Pr_{\sigma^*}[\pi]$. Ο εκτιμητής μέγιστης πιθανοφάνειας είναι η τιμή $\hat{\sigma^*} = \arg\max \mathcal{L}_\pi(\sigma^*)$. Ας δούμε όμως την ανάλυση αυτού του όρου:

$$\begin{aligned}
\hat{\sigma^*} &= \arg\max_{\sigma^*} \mathcal{L}_\pi(\sigma^*) \\
&= \arg\max_{\sigma^*} \Pr_{\sigma^*}[\pi] \\
&= \arg\max_{\sigma^*} \prod_{\sigma \in \pi} \Pr_{\sigma^*}[\sigma] \\
&= \arg\max_{\sigma^*} \prod_{\sigma \in \pi} \frac{\phi^{d_{KT}(\sigma,\sigma^*)}}{Z_\phi^m} \\
&= \arg\max_{\sigma^*} \frac{\phi^{\sum_{\sigma \in \pi} d_{KT}(\sigma,\sigma^*)}}{(Z_\phi^m)^{|\pi|}} \\
&= \arg\max_{\sigma^*} \phi^{\sum_{\sigma \in \pi} d_{KT}(\sigma,\sigma^*)} \\
&= \arg\max_{\sigma^*} \log_2 \phi^{\sum_{\sigma \in \pi} d_{KT}(\sigma,\sigma^*)} \\
&= \arg\max_{\sigma^*} (\log_2 \phi) \Big( \sum_{\sigma \in \pi} d_{KT}(\sigma, \sigma^*) \Big)
\end{aligned}$$

Όμως $\phi(p) = \frac{1-p}{p} = \frac{1}{p} - 1 \Rightarrow \phi'(p) = \frac{-1}{p^2}$. Άρα η $\phi(p)$ είναι φθίνουσα συνάρτηση. Αν θεωρήσουμε ότι $p > 1/2$ έχουμε ότι $\phi \in [0, 1) \Rightarrow \log_2 \phi < 0$

Συνεπώς:

$$\begin{aligned}\hat{\sigma^*} &= \arg\max_{\sigma^*} \mathcal{L}_\pi(\sigma^*) \\ &= \arg\max_{\sigma^*} (\log_2 \phi)\Big(\sum_{\sigma\in\pi} d_{KT}(\sigma,\sigma^*)\Big) \\ &= \arg\min_{\sigma^*} \Big(\sum_{\sigma\in\pi} d_{KT}(\sigma,\sigma^*)\Big)\end{aligned}$$

Η τελευταία σχέση αποδεικνύει ότι ο εκλογικός κανόνας του Kemeny είναι ο Maximum Likelihood Estimator (MLE) της υποβόσκουσας αλήθειας στο Mallow Model.

### 4.2.5 Μέτρα αποδοτικότητας Εκλογικών Κανόνων

**Ορισμός 4.13.** ***Ακρίβεια Εκλογικού Κανόνα.*** *Ακρίβεια ενός εκλογικού κανόνα $r$ για μια ground truth της κοινωνίας $\sigma^*$ και ένα πλήθος ψηφοφόρων $k$ ορίζεται ως:*

$$Accuracy(r,\sigma^*,k) = \sum_{\pi\in\mathcal{L}(A)^k} \Pr_{\sigma^*}[\pi]\Pr[r(\pi)=\sigma^*]$$

Η ακρίβεια ενός εκλογικού κανόνα εκφράζεται από την πιθανότητα για οποιαδήποτε κάλπη μπορεί να παραχθεί ο εκλογικός κανόνας να εξάγει την ground truth.

**Ορισμός 4.14.** ***Οριακή Ακρίβεια Εκλογικού Κανόνα.*** *Οριακή ακρίβεια ενός εκλογικού κανόνα $r$ με πλήθος ψηφοφόρων $k$ ορίζεται ως η χειρότερη περίπτωση ground truth σε ακρίβεια :*

$$Acc(r,k) = \min_{\sigma^*} Accuracy(r,\sigma^*,k)$$

Η οριακή ακρίβεια υπολογίζεται πάνω στην χειρότερη περίπτωση ground truth μπορεί να τύχει στον εκλογικό κανόνα $r$ από άποψη ακρίβειας.

**Ορισμός 4.15.** ***Πλήρης Ακρίβεια Εκλογικού Κανόνα.*** *Πλήρης ακρίβεια ενός εκλογικού κανόνα $r$ με πλήθος ψηφοφόρων $k$ ορίζεται ως το άθροισμα των διαφορετικών ground truth σε ακρίβεια :*

$$TotalAcc(r,k) = \sum_{\sigma^*\in\mathcal{L}(A)} Accuracy(r,\sigma^*,k)$$

Η οριακή ακρίβεια υπολογίζεται πάνω στην χειρότερη περίπτωση ground truth μπορεί να τύχει στον εκλογικό κανόνα $r$ από άποψη ακρίβειας.

**Ορισμός 4.16.** ***$\epsilon-$Αντοχή εκλογικού κανόνα.*** *$\epsilon-$Αντοχή ενός εκλογικού κανόνα ορίζουμε τον ελάχιστο αριθμό ψηφοφόρων που απαιτούνται ώστε η οριακή ακρίβεια ενός εκλογικού κανόνα να είναι τουλάχιστον $1-\epsilon$.*

$$N(r,\epsilon) = \min\{k | Acc(r,k) \geq 1-\epsilon\}$$

Για τους σκοπούς αυτού του κεφαλαίου η προηγούμενη ποσότητα θα αποτελεί την βασική μετρική της δειγματικής πολυπλοκότητας του εκλογικού κανόνα που μελετάμε.

Το επόμενο κρίσιμο σημείο που θα μελετήσουμε είναι η σύγκριση αυτού του κανόνα σε σχέση με οποιονδήποτε άλλον.

**Θεώρημα 4.1.** *Αν οι ισοπαλίες λύνονται με ομοιόμορφο τρόπο, τότε ο κανόνας του Kemeny έχει την μικρότερη δυνατή αντοχή για δεδομένο $\epsilon$ από κάθε τυχαιοκρατικό εκλογικό κανόνα.*

$$\forall r : N(r, \epsilon) \geq N(KEMENY, \epsilon)$$

*Απόδειξη.*

**Λήμμα 4.2.** $Accuracy(KEMENY, \sigma_1^*, k) = Accuracy(KEMENY, \sigma_2^*, k)$

*Απόδειξη.* Έστω σταθερό $k \in \mathbb{N}$ και $\sigma_1^*, \sigma_2^* \in \mathcal{L}(A)$ και έστω ένας 1-1 και επί μετασχηματισμός $bijection$ ώστε να ισχύει ότι $bijection(\sigma_1^*) = \sigma_2^*$.

$$\begin{aligned}
Accuracy(KEMENY, \sigma_2^*, k) &= \sum_{\pi \in \mathcal{L}(A)^k} \Pr_{\sigma_2^*}[\pi] \Pr[KEMENY(\pi) = \sigma_2^*] \\
&= \sum_{bijection(\pi) \in \mathcal{L}(A)^k} \Pr_{\sigma_1^*}[bijection(\pi)] \Pr[KEMENY(bijection(\pi)) = \sigma_2^*] \\
&= \sum_{\pi \in \mathcal{L}(A)^k} \Pr_{bijection^{-1}(\sigma_2^*)}[\pi] \Pr[KEMENY(\pi) = bijection^{-1}(\sigma_2^*)] \\
&= \sum_{\pi \in \mathcal{L}(A)^k} \Pr_{\sigma_1^*}[\pi] \Pr[KEMENY(\pi) = \sigma_1^*] \\
&= Accuracy(KEMENY, \sigma_1^*, k)
\end{aligned}$$

όπου $\omega(\pi) = \omega((\sigma_1, \sigma_2, \cdots, \sigma_k)) = (\omega(\sigma_1), \omega(\sigma_2), \cdots, \omega(\sigma_k))$ *Εξήγηση κατά γραμμή:*

1. Ορισμός της ακρίβειας της $\sigma_2^*$.
2. Κάθε bijection πάνω σε ένα permutation είναι επίσης ένα 1-1 και μετασχηματισμός από το $\mathcal{L}(A) \to \mathcal{L}(A)$.
3. Ισοδυναμία μεταξύ της $d_{KT}(\sigma, \sigma_2^*) = d_{KT}(\sigma, bijection(\sigma_1^*)) = d_{KT}(bijection^{-1}(\sigma), \sigma_1^*)$
4. Ορισμός του bijection.
5. Ορισμός της ακρίβειας της $\sigma_1^*$.

□

Συνεπώς για δεδομένο πλήθος ψηφοφόρων η ακρίβεια του εκλογικού κανόνα του Kemeny παραμένει η ίδια για κάθε διαφορετική ground truth.

Επίσης η πλήρης ακρίβεια ενός εκλογικού κανόνα για $k$ ψηφοφόρους είναι μικρότερη από την πλήρη ακρίβεια του Kemeny. Για την συνέχεια

$$KemenySolution = \{\sigma_{\%} : [\min_{\sigma^*} \sum_{i=1}^{k} d_{KT}(\sigma_i, \sigma^*)] = \sum_{i=1}^{k} d_{KT}(\sigma_i, \sigma_{\%})\}$$

$$\begin{aligned}
TotalAcc(r,k) &= \sum_{\sigma^* \in \mathcal{L}(A)} Accuracy(r, \sigma^*, k) = \sum_{\sigma^* \in \mathcal{L}(A)} \sum_{\pi \in \mathcal{L}(A)^k} \Pr_{\sigma^*}[\pi] \Pr[r(\pi) = \sigma^*] \\
&= \sum_{\pi \in \mathcal{L}(A)^k} \sum_{\sigma^* \in \mathcal{L}(A)} \Pr_{\sigma^*}[\pi] \Pr[r(\pi) = \sigma^*] \\
&= \sum_{\pi \in \mathcal{L}(A)^k} \sum_{\sigma^* \in \mathcal{L}(A)} \Pr_{\sigma^*}[\pi] \Pr[r(\pi) = \sigma^*] \\
&\leq \sum_{\pi \in \mathcal{L}(A)^k} \sum_{\sigma^* \in \mathcal{L}(A)} \max_{\sigma^* \in \mathcal{L}(A)} (\Pr_{\sigma^*}[\pi]) \Pr[r(\pi) = \sigma^*] \\
&= \sum_{\pi \in \mathcal{L}(A)^k} \max_{\sigma^* \in \mathcal{L}(A)} (\Pr_{\sigma^*}[\pi]) \\
&= \sum_{\pi \in \mathcal{L}(A)^k} \max_{\sigma^* \in \mathcal{L}(A)} (\Pr_{\sigma^*}[\pi]) \times 1 \\
&= \sum_{\pi \in \mathcal{L}(A)^k} \max_{\sigma^* \in \mathcal{L}(A)} (\Pr_{\sigma^*}[\pi]) \times \sum_{\sigma \in KemenySolution} \frac{1}{|KemenySolution|} \\
&= \sum_{\pi \in \mathcal{L}(A)^k} \sum_{\sigma \in KemenySolution} \max_{\sigma^* \in \mathcal{L}(A)} (\Pr_{\sigma^*}[\pi]) \frac{1}{|KemenySolution|} \\
&= \sum_{\pi \in \mathcal{L}(A)^k} \sum_{\sigma \in KemenySolution} (\Pr_{\sigma}[\pi]) \Pr[KEMENY(\pi) = \sigma] \\
&= TotalAcc(KEMENY, k)
\end{aligned}$$

Για να ολοκληρώσουμε την απόδειξη έχουμε τα ακόλουθα:

- Έστω ότι οι ψήφοι που ορίζουν την αντοχή του Kemeny είναι $k$, δηλαδή, $N(KEMENU, \epsilon) = k$.

$\Rightarrow$ Αυτό σημαίνει ότι για $k-1$ υπάρχει μια $\hat{\sigma} \in \mathcal{L}(A)$ ώστε

$$Accuracy(KEMENY, \hat{\sigma}, k-1) < 1 - \epsilon$$

$\Rightarrow$ Από το λήμμα όμως αυτό μας οδηγεί στο ότι :

$$\forall \sigma \in \mathcal{L}(A) : Accuracy(KEMENY, \sigma, k-1) < 1 - \epsilon$$

$\Rightarrow$ Έτσι έχουμε ότι $TotalAcc(KEMENY, k-1) < m! \times (1-\epsilon)$ και άρα κάθε άλλος κανόνας $r$ έχει την ιδιότητα $TotalAcc(r, k-1) < m! \times (1-\epsilon)$.

$\Rightarrow$ Από περιστεροφωλία προκύπτει ότι για οποιονδήποτε κανόνα $r$ υπάρχει μια ψήφος $\sigma \in \mathcal{L}(A)$ ώστε $Accuracy(r, \sigma, k-1) < 1 - \epsilon$.

$\Rightarrow$ Άρα για κάθε κανόνα $r$ για $k-1$ έχουμε $Acc(r, k-1) < 1 - \epsilon$.

$\Rightarrow$ Άρα $N(r, \epsilon) \geq k = N(KEMENY, \epsilon)$

□

### 4.2.6 Αλγόριθμος

Ας υποθέσουμε ότι έχουμε μια κάλπη $\pi = (\sigma_1, \sigma_2, \sigma_3, \cdots, \sigma_k)$ Κάθε ψήφος μπορεί να θεωρηθεί σαν ένας γράφος $\sigma_1 \Leftarrow \sigma_2 \Leftarrow \sigma_3 \Leftarrow \cdots \Leftarrow \sigma_k$) Όπως ακριβώς στο προηγούμενο κεφάλαιο στόχος μας είναι να μελετήσουμε τον τρόπο συνάθροισης των γραμμικών αυτών γράφων σε ένα ενιαίο γράφο κάλπης.

**Ορισμός 4.17.** ***Γράφος κάλπης*** *Έστω ένας πλήρης κατευθυνόμενος γράφος $G(V, E)$, όπου $V = A = \{a_1, a_2, \cdots, a_m\}$ και $E = \{(a_1, a_2) : a_i \in A\}$. Για να ορίσουμε το βάρος/κατεύθυνση της κάθε ακμής υπολογίζουμε την συνάρτηση*

$$weight(a_1, a_2) = \Big|\sigma \in \pi | a_1 >_\sigma a_2\Big| - \Big|\sigma \in \pi | a_2 >_\sigma a_1\Big| \Rightarrow \begin{cases} weight(a_1, a_2) < 0, & \exists e := (a_1, a_2) \\ weight(a_1, a_2) > 0, & \exists e := (a_2, a_1) \\ weight(a_1, a_2) = 0, & \nexists e := (a_2, a_1) \end{cases}$$

**Λήμμα 4.3.** *Αν ένας γράφος tournament-πλήρης κατευθυνόμενος γράφος- είναι ακυκλικός τότε έχει μονοπάτι Hamilton*

Αυτό σημαίνει ότι το DAG που ορίζει ο προηγούμενος γράφος έχει ολική διάταξη και όχι μερική. Συνεπώς από τον $G(V, E)$ επάγεται μια μοναδική μάλιστα σειρά στοιχείων $\bar{\sigma} = (a_{i_1}, a_{i_2}, \cdots, a_{i_m})$. Αξίζει να παρατηρήσει κανείς ότι η γραμμή που επάγεται από τον γράφο της κάλπης προσδιορίζει με έναν τρόπο ένα νικητή Condorcet. Ο νικητής Condorcet είναι ένας υποψήφιος ο οποίος νικάει ή έρχεται σε ισοπαλία με οποιονδήποτε άλλον υποψήφιο σε ζευγαρωτούς εκλογικούς αγώνες. Είναι εύκολο να δει κανείς ότι αν υπάρχει νικητής Condorcet τότε θα είναι άκρη του γράφου της κάλπης.

**Ορισμός 4.18.** ***Ανταλλακτικά αύξουσα μετρική*** *Μια ακέραια μετρική καλείται ανταλλακτικά αύξουσα αν για δύο $\sigma_1, \sigma_2 \in \mathcal{L}(A)$ και $a >_{\sigma_1} b, a >_{\sigma_2} b$, τότε αν ανταλλάξουμε τις θέσεις των $a, b$ στην $\sigma_1$, η απόσταση των $\sigma_1, \sigma_2$ αυξάνει τουλάχιστον κατά ένα κι αν είναι μάλιστα γειτονικά ως προς το $\sigma_2$, τότε η αύξηση είναι αυστηρή κατά 1.*

**Θεώρημα 4.2.** *Η $d_{KT}$ είναι ανταλλακτικά αύξουσα μετρική*

*Απόδειξη.* Έστω $\sigma_1, \sigma_2 \in \mathcal{L}(A)$ και $a, b \in A$ τέτοια ώστε $a >_{\sigma_1} b, a >_{\sigma_2} b$. Ας υποθέσουμε ότι $\sigma_1(a) = i, \sigma_1(b) = j, i < j$. Ας δούμε όλα τα στοιχεία που είναι μεταξύ αυτών των δύο στοιχείων $Y = \{y \in A | i < \sigma(y) < j\}$. Αφού ισχύει ότι $a >_{\sigma_2} b \Rightarrow \sigma_2(a) < \sigma_2(b)$ ισχύουν ότι:

1. $\forall y \in Y : \sigma_2(y) < \sigma_2(a) \Rightarrow \sigma_2(y) < \sigma_2(b)$ Συνεπώς:

$$\sum_{y \in Y} \mathbb{1}[\sigma_2(y) < \sigma_2(a)] \leq \sum_{y \in Y} \mathbb{1}[\sigma_2(y) < \sigma_2(a)]$$

2. $\forall y \in Y : \sigma_2(b) < \sigma_2(y) \Rightarrow \sigma_2(a) < \sigma_2(y)$ Συνεπώς:

$$\sum_{y \in Y} \mathbb{1}[\sigma_2(b) < \sigma_2(y)] \leq \sum_{y \in Y} \mathbb{1}[\sigma_2(a) < \sigma_2(y)]$$

3. $\mathbb{1}[\sigma_2(b) < \sigma_2(a)] = 0$

4. $\mathbb{1}[\sigma_2(a) < \sigma_2(b)] = 1$

$$d_{KT}(\sigma_1^{a \Leftrightarrow b}, \sigma_2) - d_{KT}(\sigma_1, \sigma_2) =$$

$$\sum_{y \in Y} \mathbb{1}[\sigma_2(y) < \sigma_2(b)] + \sum_{y \in Y} \mathbb{1}[\sigma_2(a) < \sigma_2(y)] + \mathbb{1}[\sigma_2(a) < \sigma_2(b)]$$

$$-(\sum_{y \in Y} \mathbb{1}[\sigma_2(y) < \sigma_2(a)] + \sum_{y \in Y} \mathbb{1}[\sigma_2(b) < \sigma_2(y)] + \mathbb{1}[\sigma_2(b) < \sigma_2(a)]) \geq 1$$

Είναι προφανές ότι αν $\sigma_2(b) = 1 + \sigma_2(a)$ τότε η παραπάνω ανισότητα οδηγείται σε ισότητα. □

### 4.2.7 Επεκτείνοντας από το μοναδικό $\mathcal{M}(p)$ σε διαφορετικά $\mathcal{M}(p_i)$

Μεχρις αυτού του σημείου έχουμε μελετήσει το μοντέλο του Mallow όπως προτάθηκε και επεξεργάστηκε από τον Procaccia και τον Caragiannis στο [ΠΣ13]. Στο εξής, τα επόμενα εδάφια θα συνεχίσουν με την δουλειά μας πάνω στην γενίκευση αυτής της δουλειάς.

#### 4.2.7.1 Multi-Mallow Model vs Poisson-Binomial Distribution

Στο προηγούμενο κεφάλαιο μελετήσαμε μια πολύ ισχυρή γενίκευση της Binomial distribution. Ας δούμε κάποιες ομοιότητες των δύο μοντέλων. Στο προηγούμενο μοντέλο είχαμε δείκτριες $\mathbb{1}$ μεταβλητές που παραμετροποιούνται από $p_i$ οι οποίες συναθροίζονταν σε μια τυχαία μεταβλητή της οποίας προσπαθούσαμε να μάθουμε την κατανομή. Στο κεφάλαιο αυτό θα μελετήσουμε το εξής μοντέλο. Ο κάθε ψηφοφόρος παραμετροποιείται από μια μεταβλητή $p_i$ και μοντελοποιεί την κατανομή του Mallow όπως είδαμε στις προηγούμενες ενότητες. Ο κάθε ψηφοφόρος ψηφίζει μια permutation και οι ψήφοι συναθροίζονται με τις υπόλοιπες σε μία κάλπη. Στόχος μα είναι να ανακαλύψουμε την κατανομή αυτής της κάλπης και αρχικώς τα δείγματα που απαιτούνται για να ανακαλύψουμε την ground truth.

Ομοιότητες & Διαφορές:

| Κατηγορία | Poisson Binomial Distribution | Multi Mallow Model |
|---|---|---|
| Generator | Bernoulli:$X_i \sim G_{\mathbb{1}_p}[]$ | Vote Distribution $\sigma_i \sim G_{\mathcal{M}(p)}[]$ |
| Συνάθροιση | $Y = \sum_i X_i$ | $G = \bigoplus_i \sigma_i$ |
| Ανεξαρτησία | Ανεξάρτητες δείκτριες | Ανεξάρτητοι Ψηφοφόροι |
| Τύπος μεταβλητής | Ακέραια | Διάνυσμα |
| Domain | $\{0, 1\}$ | $\mathcal{L}(A)$ |

#### 4.2.7.2 Sample Complexity

Ας υποθέσουμε ότι $i$-στος ψηφοφόρος ακολουθεί την κατανομή από το Μοντέλο Mallow με παράμετρο $p_i$.

**Θεώρημα 4.3.** *Για κάθε $\epsilon > 0$, ο κανόνας του Kemeny προσδιορίζει το ground truth με πιθανότητα τουλάχιστον $1 - \epsilon$ έχοντας $O(\log m/\epsilon)$ ψηφοφόρους.*

*Απόδειξη.* Ας υποθέσουμε ότι $\sigma^*$ είναι η ground truth. Θα αποδείξουμε ότι ο γράφος της κάλπης αν έχει κατασκευαστεί από $O(\log m/\epsilon)$ ψηφοφόρους από το Πολυπαραμετρικό Mallow μοντέλο επάγει ολική διάταξη την $\sigma^*$ με πιθανότητα τουλάχιστον $1 - \epsilon$.

Ας υποθέσουμε ότι μας δίνεται μια τυχαία κάλπη $\pi \in \mathcal{L}(A)^n$. Για καθένα από τα $a, b \in A$ ορίζουμε $n_{a,b} = |\sigma \in \pi : a >_\sigma b|$ και αντιστοιχώς $n_{b,a} = |\sigma \in \pi : b >_\sigma a|$ και προφανώς $n_{a,b} + n_{b,a} = n$.

Για να καταφέρει ο γράφος της κάλπης να επάγει την $\sigma^*$ θα πρέπει για κάθε ζευγάρι $a, b \in A$ όπου $a >_{\sigma^*} b$ θα πρέπει $n_{a,b} - n_{b,a} \geq 1$. Άρα :

$$\Pr[G(\pi) \text{ επάγει την } \sigma^*] = \Pr[\forall a, b \in A : a >_{\sigma^*} b \Rightarrow n_{a,b} - n_{b,a} \geq 1] \geq 1 - \epsilon$$

Ας ορίσουμε την τυχαία μεταβλητή το μέσο βάρος ακμής :

$$\bar{W}_{a,b} = \frac{w(a,b)}{n} = \frac{n_{a,b} - n_{b,a}}{n}$$

Παρατηρείστε ότι :

$$\bar{W}_{a,b} = \frac{\sum_i \mathcal{X}_i^{a,b}}{n}$$

όπου $X_i^{a,b} = \begin{cases} +1, & a >_{\sigma^i} b \\ -1, & b >_{\sigma^i} a \end{cases}$ και

Ας υπολογίσουμε αρχικά για δύο συγκεκριμένα υποψήφια στοιχεία

$$a, b : \Pr[n_{a,b} - n_{b,a} \leq 0] = \Pr[\frac{n_{a,b} - n_{b,a}}{n} \leq 0] = \Pr[\bar{W}_{a,b} \leq 0]$$

$$\Pr[\bar{W}_{a,b} \leq 0] \leq \Pr[-\bar{W}_{a,b} \geq 0] = \Pr[\mathbb{E}[\bar{W}_{a,b}] - \bar{W}_{a,b} \geq \mathbb{E}[\bar{W}_{a,b}]]$$

$$\Pr[\mathbb{E}[\bar{W}_{a,b}] - \bar{W}_{a,b} \geq \mathbb{E}[\bar{W}_{a,b}]] \leq \Pr[|\mathbb{E}[\bar{W}_{a,b}] - \bar{W}_{a,b}| \geq \mathbb{E}[\bar{W}_{a,b}]]$$

Έως αυτού του σημείου τα πράγματα είναι σαφή αφού

$$\{\mathbb{E}[\bar{W}_{a,b}] - \bar{W}_{a,b} \geq \mathbb{E}[\bar{W}_{a,b}]\} \subseteq \{|\mathbb{E}[\bar{W}_{a,b}] - \bar{W}_{a,b}| \geq \mathbb{E}[\bar{W}_{a,b}]\}$$

Τώρα θα χρησιμοποιήσουμε το γεγονός ότι η $\bar{W}^{a,b}$ είναι ένας μέσος όρος ανεξαρτήτων μεταβλητών

$$\Pr[\bar{W}_{a,b} \leq 0] \leq \Pr[|\mathbb{E}[\bar{W}_{a,b}] - \bar{W}_{a,b}| \geq \mathbb{E}[\bar{W}_{a,b}]] \leq 2e^{-2\mathbb{E}[\bar{W}_{a,b}]^2 n} \leq 2e^{-2\min \mathbb{E}[\bar{W}_{a,b}]^2 n}$$

$$1 - \Pr[\forall a, b \in A : a >_{\sigma^*} b \Rightarrow n_{a,b} - n_{b,a} \geq 1] \leq \Pr[\forall a, b : \bar{W}_{a,b} \leq 0]$$

Πάλι εξαιτίας της γενίκευσης του ενδοχομένου.

$$\Pr[\forall a, b : \bar{W}_{a,b} \leq 0] \leq \binom{n}{2} 2e^{-2\min \mathbb{E}[\bar{W}_{a,b}]^2 n} \leq \epsilon$$

$$\binom{n}{2} 2e^{-2\min \mathbb{E}[\bar{W}_{a,b}]^2 n} \leq \epsilon \Rightarrow n \geq \frac{\log(m^2/\epsilon)}{2\min \mathbb{E}[\bar{W}_{a,b}]^2}$$

Για να ολοκληρώσουμε την απόδειξη θα πρέπει να δείξουμε $\frac{1}{2\min \mathbb{E}[W_{a,b}]^2} = \Theta(1)$. Έχουμε ότι: $\mathbb{E}[\bar{W}_{a,b}] = \frac{1}{n}\sum_i \mathbb{E}[X_i]$ Όμως κανείς μπορεί να δει ότι ότι το $X_i = \mathbb{1}_{a>_{\sigma_i} b} - \mathbb{1}_{b>_{\sigma_i} a}$. Συνεπώς $\mathbb{E}[\bar{W}_{a,b}] = \frac{1}{n}\sum_i (\Pr_{\sigma^*}[a >_{\sigma_i} b] - \Pr_{\sigma^*}[b >_{\sigma_i} a])$. Όμως $\Pr_{\sigma^*}[a >_{\sigma_i} b] = 1 - \Pr_{\sigma^*}[b >_{\sigma_i} a]$

Ας μελετήσουμε όλα τα στοιχεία $a >_{\sigma_*} b$, μιας και αυτά είναι που μας ενδιαφέρουν κατά κύριο λόγο στον υπολογισμό της προηγούμενης πιθανότητας. Έχουμε:

$$\begin{aligned}
Q = \Pr_{\sigma^*}[a >_{\sigma_i} b] - \Pr_{\sigma^*}[b >_{\sigma_i} a] &= \sum_{\sigma \in \mathcal{L}(A) \cap a >_\sigma b} \Pr_{\sigma^*}[\sigma_i = \sigma] - \sum_{\sigma \in \mathcal{L}(A) \cap b >_\sigma a} \Pr_{\sigma^*}[\sigma_i = \sigma] \\
&= \sum_{\sigma \in \mathcal{L}(A) \cap a >_\sigma b} \Pr_{\sigma^*}[\sigma_i = \sigma] - \sum_{\sigma \in \mathcal{L}(A) \cap a >_\sigma b} \Pr_{\sigma^*}[\sigma_i = \sigma^{a \Leftrightarrow b}] \\
&= \sum_{\sigma \in \mathcal{L}(A) \cap a >_\sigma b} \Pr_{\sigma^*}[\sigma_i = \sigma] - \Pr_{\sigma^*}[\sigma_i = \sigma^{a \Leftrightarrow b}] \\
&= \sum_{\sigma \in \mathcal{L}(A) \cap a >_\sigma b} \frac{\phi_i^{d_{KT}(\sigma,\sigma^*)} - \phi_i^{d_{KT}(\sigma^{a \Leftrightarrow b},\sigma^*)}}{Z_{\phi_i}^m} \\
&\geq \sum_{\sigma \in \mathcal{L}(A) \cap a >_\sigma b} \frac{\phi_i^{d_{KT}(\sigma,\sigma^*)}(1-\phi_i)}{Z_{\phi_i}^m} \\
&(1-\phi_i) \sum_{\sigma \in \mathcal{L}(A) \cap a >_\sigma b} \frac{\phi_i^{d_{KT}(\sigma,\sigma^*)}}{Z_{\phi_i}^m} \\
&(1-\phi_i) \sum_{\sigma \in \mathcal{L}(A) \cap a >_\sigma b} \frac{\phi_i^{d_{KT}(\sigma,\sigma^*)}}{Z_{\phi_i}^m} \\
&(1-\phi_i) \Pr_{\sigma^*}[a >_{\sigma_i} b]
\end{aligned}$$

Όμως $Q = \Pr_{\sigma^*}[a >_{\sigma_i} b] - \Pr_{\sigma^*}[b >_{\sigma_i} a] = 2\Pr_{\sigma^*}[a >_{\sigma_i} b] - 1 \Rightarrow \Pr_{\sigma^*}[a >_{\sigma_i} b] = \frac{1+Q}{2}$
Άρα καταλήγουμε:

$$\begin{aligned} Q &\geq (1-\phi)(1+Q)/2 \\ 2Q &\geq (1-\phi) + (1-\phi)Q \\ Q &\geq \frac{(1-\phi)}{1+\phi} \end{aligned}$$

Άρα:

$$\mathbb{E}[\bar{W}_{a,b}] = \frac{1}{n}\sum_i (\Pr_{\sigma^*}[a >_{\sigma_i} b] - \Pr_{\sigma^*}[b >_{\sigma_i} a]) \geq \frac{1}{n}\sum_i \frac{1-\phi_i}{1+\phi_i}$$

Μέχρις αυτού του σημείου οι ανισότητες μπορούν να είναι tight. Συνεπώς καταλήγουμε ότι για κάθε ομάδα $\{\pi - \phi(\pi)\}_i$ ικανοποιεί την συνθήκη

$$\sum_i \frac{1-\phi_i}{1+\phi_i} = \Omega(n)$$

προκύπτει ότι $\frac{1}{\mathbb{E}[\bar{W}_{a,b}]} = \Omega(1)$. Αν δεχτούμε ότι $1 > \pi > 1/2 \Rightarrow \phi \in (0,1)$, συνεπάγεται ότι

$$\frac{1-\phi_i}{1+\phi_i} = \epsilon_i \in (0,1)$$

Αν διαλέξουμε τώρα το $\varepsilon = \min \epsilon_i$ προκύπτει ότι :

$$\mathbb{E}[\bar{W}_{a,b}] \geq \frac{1}{n}\sum_{i=1}^{n} \frac{1-\phi_i}{1+\phi_i} \geq \frac{1}{n}\sum_{i=1}^{n} \varepsilon = \varepsilon$$

Συνεπώς : $\mathbb{E}[\bar{W}_{a,b}]^{-1} = O(1)$. □

#### 4.2.7.3 Optimality,Lower bounds,Κριτκή

## 4.3 Ανοικτά προβλήματα-Μελλοντικές Επεκτάσεις

# Βιβλιογραφία


[Alt94] Ingo Althöfer. On sparse approximations to randomized strategies and convex combinations. *Linear Algebra and applications*, 1994.

[ARC+15] Annalisa Appice, Pedro Pereira Rodrigues, Vítor Santos Costa, Carlos Soares, João Gama, and Alípio Jorge, editors. *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I*, volume 9284 of *Lecture Notes in Computer Science*. Springer, 2015.

[BB08] Roman Bertolami and Horst Bunke. Hidden markov model-based ensemble methods for offline handwritten text line recognition. *Pattern Recognition*, 41(11):3452–3460, 2008.

[Ber41] Andrew C. Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Mathematical Proceedings of the Cambridge Philosophical Society*, 49:122–136, 1941.

[BH84] Andrew D. Barbour and Peter Hall. On the rate of poisson convergence. *Mathematical Proceedings of the Cambridge Philosophical Society*, 95:473–480, 1984.

[BHJ92] Andrew D. Barbour, Lars Holst, and Svante Janson. *Poisson Approximation*. Cambridge Unvieristy Press, Cambridge, New York, 1992.

[Bir86] L. Birge. On estimating a density using hellinger distance and some other strange facts. *IEEE Trans. Information Theory*, 51(4):271–291, 1986.

[Bir87a] L. Birge. Estimating a density under order restrictions: Nonasymptotic minimax risk. *IEEE Trans. Information Theory*, 51(4):995–1012, 1987.

[Bir87b] L. Birge. On the risk of histograms for estimating decreasing densities. *IEEE Trans. Information Theory*, 51(4):1013–1022, 1987.

[Bir97] L. Birge. Estimation of unimodal densities without smoothness assumptions. *The Annals of Statistics*, 25(3):970–981, 1997.

[Bro93] Jason Brownlee. A tour of machine learning algorithms. `http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/`, 7 1993.

[CDSS13] Siu-on Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Learning mixtures of structured distributions over discrete domains. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 1380–1394, 2013.

[CGS10] Louis H. Y. Chen, Larry Goldstein, and Qi-Man Shao. *Normal Approximation by Stein's Method*. Springer, Cambridge, New York, 2010.

[CPS13] Ioannis Caragiannis, Ariel D. Procaccia, and Nisarg Shah. When do noisy votes reveal the truth? In *ACM Conference on Electronic Commerce, EC '13, Philadelphia, PA, USA, June 16-20, 2013*, pages 143–160, 2013.

[CS90] B. Carl and I. Stephani. *Entropy, compactness and the approximation of operators*, volume 98 of *Cambridge tracts in mathematics*. Cambridge Univieristy Press, Cambridge, New York, 1990.

[CY98] Dahai Cheng and Hong Yan. Recognition of handwritten digits based on contour information. *Pattern Recognition*, 31(3):235–255, 1998.

[Das08] Constantinos Daskalakis. An efficient PTAS for two-strategy anonymous games. *CoRR*, abs/0812.2277, 2008.

[DDKT16] Constantinos Daskalakis, Anindya De, Gautam Kamath, and Christos Tzamos. A size-free CLT for poisson multinomials and its applications. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 1074–1086, 2016.

[DDO+13] Constantinos Daskalakis, Ilias Diakonikolas, Ryan O'Donnell, Rocco A. Servedio, and Li-Yang Tan. Learning sums of independent integer random variables. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 217–226, 2013.

[DDS14] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning $k$-modal distributions via testing. *Theory of Computing*, 10:535–570, 2014.

[DDS15] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning poisson binomial distributions. *Algorithmica*, 72(1):316–357, 2015.

[DG01] Luc Devroye and Lugosi Gábor. *Combinatorial methods in density estimation.* Springer series in statistics. Springer, New York, Berlin, Heidelberg, 2001.

[DK14a] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 1183–1213, 2014.

[DK14b] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 1183–1213, 2014.

[DKS15a] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. The fourier transform of poisson multinomial distributions and its algorithmic applications. *CoRR*, abs/1511.03592, 2015.

[DKS15b] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Nearly optimal learning and sparse covers for sums of independent integer random variables. *CoRR*, abs/1505.00662, 2015.

[DKT15] Constantinos Daskalakis, Gautam Kamath, and Christos Tzamos. On the structure, covering, and learning of poisson multinomial distributions. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 1203–1217, 2015.

[DP07] Constantinos Daskalakis and Christos H. Papadimitriou. Computing equilibria in anonymous games. *CoRR*, abs/0710.5582, 2007.

[DP11] Constantinos Daskalakis and Christos H. Papadimitriou. On oblivious ptas's for nash equilibrium. *CoRR*, abs/1102.2280, 2011.

[DP13] Constantinos Daskalakis and Christos H. Papadimitriou. Sparse covers for sums of indicators. *CoRR*, abs/1306.1265, 2013.

[DP15] Constantinos Daskalakis and Christos H. Papadimitriou. Approximate nash equilibria in anonymous games. *Journal of Economic Theory*, 156:207–245, 2015.

[Dud74] R.M Dudley. Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory*, abs/10.3, 1974.

[Ehm91] Werner Ehm. *Binomial Approximation to the Poisson Binomial Distribution.* Statistics and Probability Letters, 1991.

[ET96] David Eric Edmunds and Hans (mathématicien) Triebel. *Function spaces, entropy numbers, differential operators.* Cambridge tracts in mathematics. Cambridge Unvieristy Press, Cambridge, New York, 1996.

[GMM+15] Ethan Gertler, Erika Mackin, Malik Magdon-Ismail, Lirong Xia, and Yuan Yi. Computing manipulations of ranking systems. pages 685–693, 2015.

[HI90] R. Hasminskii and I. Ibragimov. On density estimation in the view of kolmogorov's ideas in approximation theory. *Annals of Statistics*, 18(3):999–1010, 1990.

[HL04] Ying Huang and Yanda Li. Prediction of protein subcellular locations using fuzzy $k$-nn method. *Bioinformatics*, 20(1):21–28, 2004.

[HO97] D. Haussler and M. Opper. Mutual information, metric entropy and cumulative relative entropy risk. *Annals of Statistics*, 25(6):2451–249, 1997.

[JKE14] Faten Kallel Jaiem, Slim Kanoun, and Véronique Eglin. Arabic font recognition based on a texture analysis. In *14th International Conference on Frontiers in Handwriting Recognition, ICFHR 2014, Crete, Greece, September 1-4, 2014*, pages 673–677, 2014.

[KMR+94] Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E. Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the Twenty-sixth Annual ACM Symposium on Theory of Computing*, STOC '94, pages 273–282, New York, NY, USA, 1994. ACM.

[Kol93] A. N. Kolmogorov. $\epsilon$-entropy and $\epsilon$-capacity of sets in functional spaces. III: Information Theory and the Theory of Algorithms:86–170, 1993.

[LBW+13] Chun-Hsiang Lee, David Birch, Chao Wu, Dilshan Silva, Orestis Tsinalis, Yang Li, Shulin Yan, Moustafa Ghanem, and Yike Guo. Building a generic platform for big sensor data application. In *Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA*, pages 94–102, 2013.

[LCLC13] Jialin Liu, Bradly Crysler, Yin Lu, and Yong Chen. Locality-driven high-level I/O aggregation for processing scientific datasets. In *Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA*, pages 103–111, 2013.

[LDTX16] Tie Luo, Sajal K. Das, Hwee Pink Tan, and Lirong Xia. Incentive mechanism design for crowdsourcing: An all-pay auction approach. *ACM TIST*, 7(3):35, 2016.

[LLCZ13] Tao Luo, Yin Liao, Guoliang Chen, and Yunquan Zhang. P-DOT: A model of computation for big data. In *Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA*, pages 31–37, 2013.

[LMM03] Richard J. Lipton, Evangelos Markakis, and Aranyak Mehta. Playing large games using simple strategies. In *Proceedings 4th ACM Conference on Electronic Commerce (EC-2003), San Diego, California, USA, June 9-12, 2003*, pages 36–41, 2003.

[LN13] Annie Louis and Ani Nenkova. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300, 2013.

[LST16] Thodoris Lykouris, Vasilis Syrgkanis, and Éva Tardos. Learning and efficiency in games with dynamic population. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 120–129, 2016.

[MB96] Petros Maragos and Muhammad Akmal Butt. Partial differential equations in image analysis: continuous modeling, discrete processing. In *Proceedings 1996 International Conference on Image Processing, Lausanne, Switzerland, September 16-19, 1996*, pages 61–64, 1996.

[MP97] Petros Maragos and Alexandros Potamianos. On using fractal features of speech sounds in automatic speech recognition. In *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997*, 1997.

[Pea95] Kerr Pearson. Contributions to the mathematical theory of evolution. ii. skew variation in homogeneous material. *Philosophical Trans.*, pages 186–343,414, 1895.

[PIKP15] Elisavet Palogiannidi, Elias Iosif, Polychronis Koutsakis, and Alexandros Potamianos. Valence, arousal and dominance estimation for english, german, greek, portuguese and spanish lexica using semantic models. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 1527–1531, 2015.

[Pol] David Pollard. Section 14.4. `http://www.stat.yale.edu/~pollard/Books/Asymptopia/Minimax.pdf`.

[RÖ7] Adrian Röllin. Translated poisson approximation using exchangeable pair couplings. *Annals of Applied Probability*, 17:1596–1614, 2007.

[ron17] Metric entropy of high dimensional distributions. In *Proceedings of the Twenty-sixth Annual ACM Symposium on Theory of Computing*, pages 4009–4018. ACM, 2017.

[Roo00] Bero Roos. Binomial approximation to the poisson binomial distribution: The krawtchouk expansion. *Theory of Probability and its Applications*, 45(2):328–344, 2000.

[Rub12] Ronitt Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, September 2012.

[SALS15] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2989–2997, 2015.

[SKE+14] Debdoot Sheet, Athanasios Karamalis, Abouzar Eslami, Peter Noël, Jyotirmoy Chatterjee, Ajoy Kumar Ray, Andrew F. Laine, Stephane G. Carlier, Nassir Navab, and Amin Katouzian. Joint learning of ultrasonic backscattering statistical physics and signal confidence primal for characterizing atherosclerotic plaques using intravascular ultrasound. *Medical Image Analysis*, 18(1):103–117, 2014.

[SKS16] Vasilis Syrgkanis, Akshay Krishnamurthy, and Robert E. Schapire. Efficient algorithms for adversarial contextual learning. *CoRR*, abs/1602.02454, 2016.

[SOAJ14] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1395–1403, 2014.

[Tsy09] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer series in statistics. Springer, 1 edition, 2009.

[Tur50] Alan Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 10 1950.

[Val84] Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

[vdVW96] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer series in statistics. Springer, New York, Berlin, Heidelberg, 1996.

[Yat85] Yannis Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *Annals of Statistics*, (13):768–774, 1985.

[YB99] Y. Yang and Andrew R. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[Y.M86] Y.Makovoz. On the kolmogorov complexity of functions of nite smoothness. *Journal of Complexity*, abs/130, 1986.

[Zol87] Vladimir M. Zolotarev. Random symmetric polynomials. *Journal of Mathematical Sciences*, 38(5):2262–2272, 1987.

[ZPX16] Zhibing Zhao, Peter Piech, and Lirong Xia. Learning mixtures of plackett-luce models. *CoRR*, abs/1603.07323, 2016.