ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Αποδοτικοί Αλγόριθμοι Μάθησης Δυνάμεων Πουασσόν Διωνυμικών Κατανομών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Βασίλης Κοντονής

Επιβλέπων: Δημήτρης Φωτάκης
Επίκουρος Καθηγητής ΕΜΠ

Αθήνα, 11/10/2017

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Αποδοτικοί Αλγόριθμοι Μάθησης Δυνάμεων Πουασσόν Διωνυμικών Κατανομών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Βασίλης Κοντονής

Επιβλέπων: Δημήτρης Φωτάκης
Επικουρος Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 11/10/2017

.....	.....	.....
Δημήτρης Φωτάκης Επίκουρος Καθηγητής ΕΜΠ	Νικόλαος Παπασπύρου Αναπληρωτής Καθηγητής ΕΜΠ	Αριστείδης Παγουρτζής Επίκουρος Καθηγητής ΕΜΠ

Αθήνα, 11/10/2017

.....
Βασίλης Κοντονής
(Διπλωματούχος Ηλεκτρολόγος Μηχανικός & Μηχανικός Υπολογιστών Ε.Μ.Π.)

Οι απόψεις που εκφράζονται σε αυτό το κείμενο είναι αποκλειστικά του συγγραφέα και δεν αντιπροσωπεύουν απαραίτητα την επίσημη θέση του Εθνικού Μετσόβιου Πολυτεχνείου.

Το περιεχόμενο της εργασίας διατίθεται υπό την άδεια Creative Commons Attribution-NonCommercial 4.0. Απαγορεύεται η χρήση του περιεχομένου για εμπορικούς σκοπούς.

Περίληψη

Σε αυτήν την διπλωματική εργασία εισάγουμε το πρόβλημα της μάθησης όλων των δυνάμεων μιας Poυασσόν Διωνυμικής Κατανομής (Poisson Binomial Distribution, PBD). Μια n-PBD είναι η κατανομή του αθροίσματος $X = \sum_{i=1}^{n} X_i$, n ανεξάρτητων 0/1 κατανομών Bernoulli, όπου $\mathbb{E}_{X_i}[=] p_i$. Η k-οστή *δύναμη* της X, για ένα $k \in [m]$, είναι η κατανομή $P_k = \sum_{i=1}^{n} X_i^{(k)}$, όπου $\mathbb{E}_{X_i^{(k)}}[=] (p_i)^k$. Ο αλγόριθμος μάθησης μπορεί να τραβήξει δείγματα από οποιαδήποτε δύναμη P_k και λέμε ότι πετυχαίνει στην ταυτόχρονη εκμάθηση όλων των δυνάμεων στο εύρος $[m]$ αν με πιθανότητα το λιγότερο $1 - \delta$: δεδομένου ενός $k \in [m]$ επιστρέφει μια κατανομή Q_k τέτοια ώστε $d_{\mathrm{tv}}(P_k, Q_k) \leq \varepsilon$.

Δείχνουμε αρχικά ένα πληροφοριό-θεωρητικό κάτω φράγμα για την δειγματική πολυπλοκότητα του παραπάνω προβλήματος στην περίπτωση που οι παράμετροι p_i-s της PBD είναι αρκετά διαχωρισμένες. Το κάτω φράγμα αυτό δείχνει ότι χρειαζόμαστε περίπου έναν σταθερό αριθμό δειγμάτων για κάθε ξεχωριστή παράμετρο p_i, δηλαδή συνολικά $\Omega(n)$ δείγματα για τις δυνάμεις μιας n-PBD. Στη συνέχεια γίνουμε ένα σχεδόν βέλτιστο πάνω φράγμα για την δειγματική πολυπλοκότητα στην περίπτωση που η PBD έχει μορφή παρόμοια με αυτή του κάτω φράγματός μας.

Επεκτείνουμε το κλασικό ορισμό του minimax risk της Στατιστικής και επεκτείνοντας τις τεχνικές μελέτης της δειγματικής πολυπλοκότητας για την προσέγγιση συναρτήσεων ακολουθιών κατανομών. Συγκεκριμένα επεκτείνουμε τις κλασικές μεθόδους των Le Cam και Fano για να παράγουμε κάτω φράγματα στο δικό μας μοντέλο μάθησης ακολουθιών κατανομών.

Μελετούμε το βασικό πρόβλημα της μάθησης των δυνάμεων μιας Διωνυμικής κατανομής και παρέχουμε έναν βέλτιστο αλγόριθμο μάθησης των δυνάμεων χρησιμοποιώντας $O(1/eps^2)$ δείγματα από τις δυνάμεις της Διωνυμικής κατανομής. Αποδεικνύουμε ότι ο αλγόριθμός είναι βέλτιστος δείχνοντας ένα κάτω φράγμα $\Omega(1/\varepsilon^2)$ χρησιμοποιώντας το νέο minimax framework μας.

Η μάθηση των παραμέτρων p_i μιας PBD είναι ένα γνωστό δύσκολο πρόβλημα. Οι Διακονικόλας, Kane, και Stewart [COLT'16] έδειξαν ένα εκθετικό κάτω φράγμα $\Omega(2^{1/\varepsilon})$ δειγμάτων για την μάθηση των p_i με αθροιστικό σφάλμα το πολύ ε. Ένα φυσικό ερώτημα, λοιπόν, είναι αν η δυνατότητα δειγματοληψίας τόσο από την ίδια την PBD όσο και από τις δυνάμεις της βοηθάει στο να μειωθεί η δειγματική πολυπλοκότητα του προβλήματος αυτού. Δίνουμε αρνητική απάντηση σε αυτό το ερώτημα δείχνοντας το ίδιο κάτω φράγμα στο δικό μας μοντέλο των δυνάμεων. Τέλος, παρέχουμε έναν σχεδόν βέλτιστο αλγόριθμο παραμετρικής μάθησης των p_i χρησιμοποιώντας δείγματα από τις δυνάμεις μιας PBD.

Λέξεις-κλειδιά: Poυασσόν Διωνυμική Κατανομή · Μηχανική Μάθηση · Θεωρία Στατιστικής

Ευχαριστίες

Πρώτα από όλα θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κ. Φωτάκη που με βοήθησε και με κατεύθυνε κατά την διάρκεια αυτής της διπλωματικής. Χωρίς τη βοήθειά η διπλωματική αυτή εργασία δεν θα μπορούσε να πραγματοποιηθεί. Επίσης θα ήθελα να ευχαριστήσω τον Piotr Krysta ο οποίος μας επισκέφτηκε εδώ στην Αθήνα και είχαμε μια άψογη συνεργασία πάνω στο παρόν πρόβλημα. Θα ήθελα επίσης να ευχαριστήσω την οικογένειά μου και τους φίλους μου τους οποίους θα παραθέσω με τυχαία σειρά για αποφυγή παρεξηγήσεων: Νίκος Καρ., Κώστης Α., Μελίνα Π., Γιάννης Κ., Νίκος Γ. Μάριος Κ., Γιάννης Σ., Βαγγέλης Σ., Βασίλης Μ., Ελένη Μ., Τάκης Κ., Βασίλης Χ., Νίκος Κορ., Ορέστης Π., Κορίνα Τ.

Βασίλης Κοντονής

Αθήνα, 11/10/2017

Περιεχόμενα

Κατάλογος σχημάτων

Κεφάλαιο 1

Μάθηση Κατανομών

1.1 Εισαγωγή

Λαμβάνοντας υπ' όψιν τον τεράστιο όγκο των δεδομένων τόσο σε εμπορικές όσο και σε επιστημονικές εφαρμογές, είναι φανερή η ανάγκη αποδοτικών μεθόδων για να ανακαλύπτουμε την δομή τους. Ας φανταστούμε ότι έχουμε μια βάση όπου διατηρούμε δεδομένα για τους ασθενείς ενός νοσοκομείου. Η ηλικία εμφάνισης μιας συγκεκριμένης ασθένειας μπορεί να μοντελοποιηθεί σαν μία τυχαία μεταβλητή. Μπορεί να θέλουμε να μάθουμε απλά την μέση τιμή αυτής της τυχαίας μεταβλητής αλλά ένα πιο ενδιαφέρον ερώτημα που θα μας έδινε πολύ περισσότερη πληροφορία είναι να μάθουμε την συνάρτηση πυκνότητας πιθανότητας της κατανομής που ακολουθεί, δηλαδή να *προσεγγίσουμε* την κατανομή της. Η μάθηση κατανομών είναι ένα πρόβλημα *μη-εποπτευόμενης (unsupervised)* μάθησης αφού ένα μεγάλο σύνολο δεδομένων χωρίς *ετικέτες (labels)* μπορεί να μοντελοποιηθεί σαν ένα σύνολο από δείγματα μιας κατανομής. Τα δεδομένα εισόδου κάθε αλγορίθμου προσέγγισης κατανομών είναι δείγματα από την κατανομή. Στο προηγούμενο παράδειγμα τα δείγματα αντιστοιχούν σε εγγραφές της βάσης. Προφανώς για να μάθουμε ακριβώς την κατανομή πρέπει να εξετάσουμε ολόκληρο το σύνολο των δεδομένων της βάσης. Επίσης, είναι διαισθητικά φανερό ότι όσο περισσότερα δείγματα παίρνουμε από τη βάση τόσο καλύτερη θα είναι και η προσέγγιση της κατανομής. Ένας βασικός στόχος του κλάδου προσέγγισης κατανομών είναι η ποσοτικοποίηση της σχέσης σφάλματος και αριθμού δειγμάτων. Αν η βάση δεδομένων είναι τεράστια, ένας αποδοτικός αλγόριθμος μάθησης θα πρέπει να εξετάζει ένα πολύ μικρό της μέρος και να είναι σε θέση να δίνει ως έξοδο μια προσέγγιση της πραγματικής κατανομής. Επιπλέον, πέρα από το να χρησιμοποιεί μικρό αριθμό δειγμάτων, ο αλγόριθμος θα πρέπει να τα επεξεργάζεται σε εύλογο χρονικό διάστημα. Μας ενδιαφέρει, λοιπόν, να μελετάμε τόσο την *δειγματική* όσο και την *υπολογιστική ή χρονική* πολυπλοκότητα των αλγορίθμων μάθησης κατανομών.

Η μάθηση κατανομών αποτελεί ένα πρόβλημα μη-εποπτευόμενης μάθησης που τα τελευταία χρόνια απασχολεί πέρα την κοινότητα της στατιστικής και την κοινότητα της Θεωρητικής Επιστήμης Υπολογιστών κυρίως από την πλευρά της υπολογιστικής πολυπλοκότητας των αλγορίθμων μάθησης. Ας υποθέσουμε ότι η άγνωστη κατανομή ανήκει σε μία κλάση κατανομών $\mathcal{D}$. Η μάθηση κατανομών χωρίζεται σε 3 βασικές κατηγορίες:

- Στην περίπτωση της *non-proper* μάθησης ο στόχος μας είναι να υπολογίσουμε μία προσεγγιστική συνάρτηση πυκνότητας χωρίς κανέναν επιπλέον περιορισμό, δηλαδή η προσεγγιστική κατανομή δεν είναι κατ' ανάγκη μέλος της κλάσης $\mathcal{D}$.
- Στην περίπτωση της *proper* μάθησης περιοριζόμαστε σε προσεγγιστικές κατανομές που ανήκουν στην κλάση $\mathcal{D}$ της άγνωστης κατανομής.
- Στην περίπτωση της *parameter* μάθησης θέλουμε να βρούμε τις παραμέτρους που ορίζουν την άγνωστη κατανομή.

Αξίζει σε αυτό το σημείο να σημειωθεί ότι από πλευράς δειγματικής πολυπλοκότητας non-proper και proper μάθηση ταυτίζονται καθώς μπορούμε πάντα σε πρώτη φάση να χρησιμοποιούμε έναν non-proper learning αλγόριθμο για να παράγουμε μία προσεγγιστική κατανομή και μετά να βρούμε με εξαντλητική αναζήτηση την κοντινότερη σε αυτήν κατανομή της οικογένειας $\mathcal{D}$. Από την άλλη είναι φανερό ότι η υπολογιστική πολυ-

πλοκότητα proper και non-proper μάθησης μπορεί να διαφέρει δραματικά καθώς η εύρεση της κοντινότερης κατανομής μέσα στην $\mathcal{D}$ μπορεί να είναι ιδιαίτερα δαπανηρή υπολογιστικά.

Ο πιο απλός (non-proper) αλγόριθμος προσέγγισης μιας κατανομής είναι αυτός του ιστογράμματος ο οποίος προτάθηκε το 1895 από τον Karl Pearson [25] και πλέον διδάσκεται στους μαθητές του Λυκείου. Στον αλγόριθμο του ιστογράμματος διαμερίζουμε το στήριγμα της κατανομής σε διαστήματα και τραβώντας δείγματα από την κατανομή και για κάθε διάστημα υπολογίζουμε το ποσοστό των δειγμάτων που "έπεσαν" μέσα σε αυτό προς το συνολικό πλήθος των δειγμάτων (σχετική συχνότητα). Σε κάθε διάστημα της διαμέρισης προσεγγίζουμε την άγνωστη συνάρτηση πυκνότητας με μία σταθερή συνάρτηση που ισούται με την σχετική συχνότητα του συγκεκριμένου διαστήματος και ως έξοδο δίνουμε μία κατά-τμήματα σταθερή συνάρτηση. Να σημειωθεί ότι ακόμα και σε αυτήν την φαινομενικά απλή μέθοδο η κατασκευή της διαμέρισης επηρεάζει δραματικά την απόδοση του αλγορίθμου και οι επιλογή του συνολικού πλήθος των διαστημάτων, και του πλάτους κάθε διαστήματος εξαρτάται από το πρόβλημα και είναι μη-τετριμμένη.

Κεφάλαιο 2

Τεχνικό Υπόβαθρο

2.1 Ανισότητες Συγκέντρωσης Μέτρου

Οι ιδέες της συγκέντρωσης μέτρου αναπτύχθηκαν κυρίως τον προηγούμενο αιώνα σε διάφορους τομείς των μαθηματικών όπως η συναρτησιακή ανάλυση, η θεωρία πιθανοτήτων και η στατιστική. Οι ίδιες κεντρικές ιδέες και ανισότητες έχουν χρησιμοποιηθεί σε ένα τεράστιο εύρος εφαρμογών όπως στην ανάλυση των τυχαιοκρατικών αλγορίθμων, στην μηχανική μάθηση και στη κβαντική θεωρία πληροφορίας. Ξεκινάμε με την πασίγνωστη ανισότητα του Markov

Πρόταση 1 (Ανισότητα Markov)**.** Έστω X μια μη-αρνητική τυχαία μεταβλητή και $t > 0$. Τότε

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$$

Η ανισότητα του Chebyshev προκύπτει άμεσα από την ανισότητα του Markov

Πρόταση 2 (Ανισότητα Chebyshev)**.** Έστω X μια τυχαία μεταβλητή και $t > 0$. Τότε

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathrm{Var}[X]}{t}$$

Απόδειξη.

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t] = \mathbb{P}\left[|X - \mathbb{E}[X]|^2 \geq t^2|\right] \leq \frac{\mathbb{E}\left[|X - \mathbb{E}[X]|^2\right]}{t^2} = \frac{\mathrm{Var}[X]}{t^2}$$

□

Συνεχίζουμε δίνοντας δύο εύχρηστες εκδοχές του γνωστού φράγματος του Chernoff.

Πρόταση 3 (Φράγμα Chernoff)**.** Έστω $X = X_1 + \cdots + X_n$, $X_i \in [0,1]$, $\mu = \mathbb{E}[X]$ και $\sigma^2 = \mathrm{Var}[X]$. Τότε, για κάθε $\lambda \in (0, 2\sigma)$, $\mathbb{P}[X > \mu + \lambda\sigma] < e^{-\lambda^2/4}$ και $\mathbb{P}[X < \mu - \lambda\sigma] < e^{-\lambda^2/4}$.

Απόδειξη. Βλέπε σελίδα 8 στο βιβλίο [16]. □

Η επόμενη πρόταση μας δίνει την πολλαπλασιαστική εκδοχή του φράγματος του Chernoff.

Πρόταση 4. Έστω $X = X_1 + \cdots + X_n$ το άθροισμα από n ανεξάρτητες τυχαίες μεταβλητές με τιμές στο $\{0, 1\}$ και έστω $\mu = \mathbb{E}[X]$. Τότε για $0 \leq \delta \leq 1$ έχουμε

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

Ας δούμε σε αυτό το σημείο μία πολύ χρήσιμη εφαρμογή του φράγματος του Chernoff για την αύξηση της πιθανότητας επιτυχίας ενός randomized αλγορίθμου του οποίου η έξοδος είναι ένας πραγματικός αριθμός. Έχοντας στη διάθεσή μας έναν αλγόριθμο με σταθερή πιθανότητα επιτυχίας μπορούμε ουσιαστικά να τον τρέξουμε περισσότερες φορές και στο τέλος να κρατήσουμε την διάμεσο των απαντήσεων που θα μας δώσει. Με αυτό το κόλπο (median trick) μπορούμε να πάρουμε εκτιμήσεις με πιθανότητα επιτυχίας πολύ κοντά στη μονάδα τρέχοντας τον βασικό αλγόριθμο μόνο λογαριθμικές σε πλήθος φορές.

Πρόταση 5. Έστω $\mathcal{A}$ ένας τυχαιοκρατικός αλγόριθμος ο οποίος δίνει σαν έξοδο πραγματικούς αριθμούς. Θα λέμε ότι ο $\mathcal{A}$ πετυχαίνει με πιθανότητα p όταν η έξοδος του βρίσκεται σε ένα διάστημα $[a, b]$ με πιθανότητα p. Τότε υπάρχει αλγόριθμος $\mathcal{B}$ ο οποίος πετυχαίνει με πιθανότητα $1 - \delta$ και τρέχει τον $\mathcal{A}$ $(\log(1/\delta)$ ανεξάρτητες φορές.

Απόδειξη. Έστω $x_1, \ldots, x_N$ τα αποτελέσματα των t ανεξάρτητων εκτελέσεων του αλγορίθμου $\mathcal{A}$ και $\bar{x}$ η διάμεσός τους. Αφού ο $\mathcal{A}$ έχει πιθανότητα επιτυχίας p κάθε ένα από τα αποτελέσματα x_i βρίσκεται στο εύρος επιτυχίας με πιθανότητα p. Για να είναι η διάμεσος $\bar{x}$ εκτός του εύρους επιτυχίας πρέπει τουλάχιστον τα μισά αποτελέσματα να είναι εκτός του του εύρους επιτυχίας. Αυτό που ζητάμε δηλαδή είναι η πιθανότητα να έχουμε περισσότερες από $N/2$ αποτυχίας σε μία ακολουθία n δοκιμών Bernoulli X_i με πιθανότητα επιτυχίας p. Συμβολίζοντας τον αριθμό των επιτυχιών ως $\sum_{i=1}^N X_i$ έχουμε $\mu = \mathbb{E}\left[\sum_{i=1}^N X_i\right] = pN$. Θέλουμε να πάρουμε ένα πάνω φράγμα στην πιθανότητα η διάμεσος να είναι εκτός του διαστήματος επιτυχίας, δηλαδή

$$\mathbb{P}\left[\sum_{i=1}^N X_i \leq n/2\right] = \mathbb{P}\left[\sum_{i=1}^N X_i - \mu \leq (0.5-p)m\right] = \mathbb{P}\left[\sum_{i=1}^N X_i - \mu \leq \frac{(0.5-p)}{p}\mu\right] \leq \mathrm{e}^{-\frac{N}{2p}\left(p-\frac{1}{2}\right)^2},$$

όπου χρησιμοποιήσαμε το πολλαπλασιαστικό φράγμα Chernoff 4. Είναι πλέον φανερό ότι με λογαριθμικό πλήθος επαναλήψεων $N = O(\log(1/\delta))$ μπορούμε να μικρύνουμε την πιθανότητα αποτυχίας σε δ. □

Για την κανονική κατανομή υπάρχουν τα επόμενα πολύ ισχυρά φράγματα. Θυμίζουμε ότι συμβολίζουμε με $\mathcal{N}(\mu, \sigma)$ την κανονική κατανομή με μέση τιμή μ και διακύμανση σ^2. Η πυκνότητα της $\mathcal{N}(\mu, \sigma)$ είναι $f(x) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Συμβολίζουμε με $\operatorname{erf}(x)$ την συνάρτηση σφάλματος Gauss, δηλαδή $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \mathrm{e}^{-t^2}\,\mathrm{d}t$, και με $\operatorname{erfc}(x)$ the συμπληρωματική συνάρτηση σφάλματος, $\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_x^{+\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t$. Ο Chu [7] έδειξε την επόμενη ανισότητα για την συνάρτηση σφάλματος

Πρόταση 6 (Ανισότητες του Chu)**.** Για κάθε $x \geq 0$:

$$\sqrt{1 - \mathrm{e}^{-ax^2}} \leq \operatorname{erf}(x) \leq \sqrt{1 - \mathrm{e}^{-bx^2}},$$

όπου $a = 1$ και $b = 4/\pi$.

Το επόμενο πόρισμα της Πρότασης 6 δίνει ένα ελαφρώς χειρότερο κάτω φράγμα για την $\operatorname{erf}(x)$.

Πόρισμα 1. Αν $0 \leq x \leq 1$, τότε έχουμε $\operatorname{erf}(x) \geq x/c$, όπου c είναι μιά σταθερά τέτοια ώστε $c \geq \sqrt{\mathrm{e}/(\mathrm{e}-1)}$.

Απόδειξη. Χρησιμοποιώντας το κάτω φράγμα της ανισότητας της Πρότασης 6 θέλουμε να δείξουμε ότι

$$\frac{x}{c} \leq \sqrt{1 - \frac{1}{\mathrm{e}^{x^2}}},$$

για κάθε $x \in [0,1]$. Αυτή η ανισότητα είναι ισοδύναμη με

$$f(x) := \mathrm{e}^{x^2}(c^2 - x^2) - c^2 \geq 0.$$

Έχουμε $f'(x) = 2x\mathrm{e}^{x^2}(c^2 - 1 - x^2)$, οπότε βλέπουμε ότι $f'(x) \geq 0$ αν $x \leq \sqrt{c^2-1}$ και $f'(x) \leq 0$ αν $x \geq \sqrt{c^2-1}$. Αυτό σημαίνει ότι η συνάρτηση f μεγιστοποιείται στο $x_0 = \sqrt{c^2-1}$ και επομένως το ελάχιστό της στο $[0,1]$ είναι $\min\{f(0), f(1)\} = \{0, (\mathrm{e}-1)c^2 - \mathrm{e}\}$. Απαιτώντας να ισχύει $(\mathrm{e}-1)c^2 - \mathrm{e} \geq 0$, έχουμε $c \geq \sqrt{\mathrm{e}/(\mathrm{e}-1)}$ και με αυτή τη συνθήκη έχουμε $f(x) \geq 0$ για κάθε $x \in [0,1]$. □

Για να φράξουμε την $\mathrm{erfc}\,(z)$ θα χρησιμοποιήσουμε τις ανισότητες του Komatsu σελίδα 17 στο [21]. Για περισσότερα τέτοια αποτελέσματα βλέπε [32].

Πρόταση 7 (Ανισότητες του Komatsu)**.** Για κάθε $a \geq 0$ ισχύει

$$\frac{\mathrm{e}^{-a^2/2}}{2\sqrt{a^2+4}+a} \leq \int_a^{+\infty} \mathrm{e}^{-t^2/2}\mathrm{d}t \leq \frac{\mathrm{e}^{-a^2/2}}{2\sqrt{a^2+2}+a}$$

2.2 f-Αποκλίσεις Κατανομών

Για να προσδιορίσουμε την ποιότητα της προσέγγισης που παράγει ένας αλγόριθμος μάθησης κατανομών πρέπει να ορίσουμε μετρικές ώστε να μπορούμε να συγκρίνουμε το προσεγγιστικό μέτρο που υπολογίζει με ο αλγόριθμος μάθησης με το άγνωστο πραγματικό μέτρο πιθανότητας. Παρόλο που οι βασικότερες μετρικές για μέτρα πιθανότητας όπως και η απόκλιση Kullback-Leibler ήταν ήδη γνωστές οι Ali και Silvey [2] και ανεξάρτητα ο Csiszar [9] δώσανε έναν ενοποιημένο χαρακτηρισμό των διάφορων *αποκλίσεων* που ποσοτικοποιούν το πόσο "κοντά" είναι δύο μέτρα πιθανότητας. Στην συνέχεια θα προσπαθήσουμε να σκιαγραφήσουμε τις βασικές ιδέες πίσω από την κατασκευή των f-αποκλίσεων αποφεύγοντας μιας μακροσκελή μετροθεωρητική ανάλυση. Ο ενδιαφερόμενος για τις αυστηρές αποδείξεις αναγνώστης μπορεί να δει τα [9], [2], [22], [29], και [1]. Τονίζουμε ότι ο όρος απόκλιση δεν ταυτίζεται με τον όρο μετρική μιας και όπως θα δούμε υπάρχουν αποκλίσεις, όπως η απόκλιση Kullback-Leibler οι οποίες δεν ικανοποιούν τον ορισμό της μετρικής. Στην μέθοδο κατασκευής αποκλίσεων κυρίαρχο ρόλο παίζει η παράγωγος Radon-Nikodym $\mathrm{d}P/\mathrm{d}Q$ των δύο μέτρων. Ένας ιδιαίτερα απλοϊκός τρόπος να σκεφτόμαστε την παράγωγο Radon-Nikodym είναι ως το όριο $\lim_{\varepsilon\to 0} P(\mathcal{B}(x,\varepsilon))/Q(\mathcal{B}(x,\varepsilon))$ δηλαδή ως τον λόγο των μέτρων πολύ μικρών συνόλων. Στην περίπτωση που έχουμε δύο μέτρα απολύτως συνεχή με το μέτρο Lebesgue και μεταξύ τους, δηλαδή μηδενίζονται ακριβώς στα ίδια στοιχεία της σ-άλγεβρας $\mathcal{A}$, έχουμε $\phi(x) = \mathrm{d}Q/\mathrm{d}P = p(x)/q(x)$ όπου p, q είναι οι πυκνότητες των P, Q αντίστοιχα. Αν P και Q ταυτίζονται έχουμε $\phi(x) = 1$. Ας δούμε τώρα τι συμβαίνει στον λόγο $\phi(x)$ καθώς "απομακρύνουμε" την P από την Q. Τί σημαίνει όμως απομακρύνουμε δύο μέτρα πιθανότητας; Αν δοκιμάσουμε να δώσουμε μικρότερη P-πιθανότητα σε κάποια σύνολα τότε αναγκαστικά την μάζα πιθανότητας που αφαιρούμε από αυτά θα πρέπει να την τοποθετήσουμε σε κάποια άλλα σύνολα της σ-αλγεβρας $\mathcal{A}$ οπότε δημιουργούνται δύο ομάδες συνόλων. Αυτά τα οποία έχουν μεγαλύτερη P-πιθανότητα από Q-πιθανότητα και αυτά τα οποία έχουν μεγαλύτερη Q-πιθανότητα. Συνεπώς ο λόγος $\phi(x)$ παίρνει τιμές μεγαλύτερες της μονάδας στα σύνολα που αυξάνεται η P-πιθανότητα και μικρότερες της μονάδας στα σύνολα που αυξάνεται η Q-πιθανότητα. Δεδομένου ότι $\int \phi(x)\mathrm{d}P = \int \mathrm{d}Q = 1$, δηλαδή η μέση τιμή του λόγου $\phi(x)$ ως προς το μέτρο P είναι σταθερή και ίση με 1 αντιλαμβανόμαστε ότι η P-διασπορά (δηλαδή η διασπορά μετρημένη ως προς το μέτρο P) του $\phi(x)$ αυξάνεται καθώς απομακρύνουμε τα μέτρα P και Q αφού η $\phi(x)$ έχει P-μέση τιμή 1 και τιμές που απομακρύνονται από την μονάδα όσο απομακρύνουμε τα δύο μέτρα. Συνεπώς φαίνεται ότι μια ποσότητα που μετρά την P-διασπορά της παραγώγου $\mathrm{d}p/\mathrm{d}Q$ είναι μια καλή επιλογή για να μετράμε την απόκλιση των δύο μέτρων. Η P-μέση τιμή μιας κυρτής συνάρτησης μιας τυχαίας μεταβλητής μετρά, σε μικρότερο ή μεγαλύτερο βαθμό, την P-διασπορά της. Για να το καταλάβουμε αυτό είναι καλύτερο να σκεφτούμε δύο τυχαίες μεταβλητές X, Y. Οι τιμές της X είναι -2

με πιθανότητα $1/2$ και 2 με πιθανότητα $1/2$, οι τιμές της Y είναι -1 με πιθανότητα $1/2$ και 1 με πιθανότητα $1/2$. Προφανώς η διασπορά της X είναι μεγαλύτερη από την διασπορά της X. Έστω τώρα η κυρτή συνάρτηση $x \mapsto \mathrm{e}^x$, τότε η μέση τιμή $\mathbb{E}\left[\mathrm{e}^X\right] > \mathbb{E}\left[\mathrm{e}^Y\right]$ άρα βλέπουμε ότι η μέση τιμή μιας κυρτής συνάρτησης είναι ένας δείκτης της διασποράς μιας τυχαίας μεταβλητής. Όσο μεγαλύτερη είναι η κυρτότητα τόσο πιο ευαίσθητη είναι η μέση τιμή της κυρτής συνάρτησης στην διασπορά της τυχαίας μεταβλητής. Μετά την παραπάνω συζήτηση ο επόμενος ορισμός της απόκλισης δύο μέτρων πιθανότητας πρέπει να φαίνεται εντελώς φυσικός.

Ορισμός 1. Έστω δύο μέτρα πιθανότητας P,Q σε ίδιο μετρήσιμο χώρο $(\Omega, \mathcal{A})$ και $f : \mathbb{R}_+ \to \mathbb{R}$ συνεχής κυρτή συνάρτηση. Ορίζουμε σαν f-απόκλιση

$$D_f\left(Q\|P\right) \coloneqq \int f\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)\mathrm{d}P.$$

Συνεχίζουμε δίνοντας μια βασική ιδιότητα που ικανοποιεί ο παραπάνω ορισμός. Αν $y = t(x)$ είναι ένας μετρήσιμος μετασχηματισμός από τον χώρο $(\Omega, \mathcal{A})$ στον $(Y, \mathcal{G})$ έχουμε

$$D_f\left(Q\|P\right) \geq D_f\left(Qt^{-1}\|Pt^{-1}\right)$$

Από αυτόν τον ορισμό προκύπτουν οι βασικές μετρικές και αποκλίσεις μεταξύ δύο μέτρων πιθανότητας P, Q ορισμένων στον ίδιο μετρήσιμο χώρο $(\Omega, \mathcal{A})$. Στα παρακάτω υποθέτουμε ότι τα μέτρα P και και Q είναι απολύτως συνεχή ως προς ένα κοινό μέτρο v. Από το γνωστό θεώρημα Radon-Nikodym αυτό σημαίνει ότι τα δύο αυτά μέτρα έχουν πυκνότητες $p(x), q(x)$ ως προς το κοινό μέτρο v.

1. Με $f(x) = x\log(x)$ παίρνουμε την απόκλιση Kullback-Leibler των P, Q,

$$D_{\mathrm{kl}}\left(Q\|P\right) = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P.$$

2. Με $f(x) = \frac{1}{2}|x-1|$ παίρνουμε την απόσταση ολικής κύμανσης

$$d_{\mathrm{tv}}\left(Q, P\right) = \sup_{A\in\mathcal{A}} |P(A) - Q(A)|.$$

Πράγματι βλέπουμε

$$\begin{aligned} D_f\left(Q\|P\right) &= \frac{1}{2}\int \left|\frac{p(x)}{q(x)} - 1\right| q(x)\mathrm{d}v \\ &= \frac{1}{2}\int_{p(x)>q(x)} p(x) - q(x)\mathrm{d}v + \frac{1}{2}\int_{p(x)<q(x)} q(x) - p(x)\mathrm{d}v \\ &= \frac{1}{2}\sup_{A\in\mathcal{A}}\left(P(A) - Q(A)\right) + \frac{1}{2}\sup_{A\in\mathcal{A}}\left(Q(A) - P(A)\right) \\ &= d_{\mathrm{tv}}\left(P, Q\right) \end{aligned}$$

3. Με $f(x) = \frac{1}{2}(\sqrt{(}x) - 1)^2$ παίρνουμε την απόσταση Hellinger δύο κατανομών

$$d_{\mathrm{hel}}\left(Q, P\right)^2 = \frac{1}{2}\int\left(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q}\right)^2.$$

4. Με $f(x) = \frac{1}{2}(t-1)^2$ παίρνουμε την απόκλιση χ^2 των P, Q δηλαδή

$$D_{\chi^2}\left(Q\|P\right) = \frac{1}{2}\int\left(\frac{p(x)}{q(x)} - 1\right)^2 q(x)\mathrm{d}v.$$

2.2.1 Ανισότητες για PBDs

Μιας και στην παρούσα διπλωματική ασχολούμαστε κυρίως με Poisson Binomial κατανομές είναι χρήσιμο να συγκεντρώσουμε γνωστές ανισότητες για τις αποστάσεις τους.

Η επόμενη πρόταση αποτελεί έναν ακριβή υπολογισμό της Kullback-Leibler απόκλισης στην περίπτωση των διωνυμικών Κατανομών.

Πρόταση 8 (Binomial KL-Divergence)**.** Έστω $X = B(n,p),\ Y = B(n,q)$ δύο διωνυμικές κατανομές. Τότε

$$D_{kl}(X\|Y) = -n\left((1-p)\log\left(\frac{1-q}{1-p}\right) + p\log\left(\frac{q}{p}\right)\right)$$

Απόδειξη. Έχουμε

$$\begin{aligned}
D_{\mathrm{kl}}(X\|Y) &= \sum_{k=0}^{n} X(k)\ln\frac{X(k)}{Y(k)} \\
&= \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}\ln\left(\frac{p^k(1-p)^{n-k}}{q^k(1-q)^{n-k}}\right) \\
&= \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}\left(k\ln\left(\frac{p}{q}\right) + (n-k)\ln\left(\frac{1-p}{1-q}\right)\right) \\
&= \ln\left(\frac{p}{q}\right)\sum_{k=0}^{n}k\binom{n}{k}p^k(1-p)^{n-k} + \ln\left(\frac{1-p}{1-q}\right)\sum_{k=0}^{n}(n-k)\binom{n}{k}p^k(1-p)^{n-k} \\
&= np\ln\left(\frac{p}{q}\right) + n(1-p)\ln\left(\frac{1-p}{1-q}\right).
\end{aligned}$$

□

Στην επόμενη απλή πρόταση αποδεικνύουμε το διαισθητικά αναμενόμενο γεγονός ότι η απόκλιση Kullback-Leibler ανάμεσα σε δύο Διωνυμικές είναι μεγαλύτερη όταν οι παράμετροί τους p, q είναι μακριά.

Πρόταση 9. Έστω $X \sim B(n,p),\ Y \sim B(n,q)$. Τότε $D_{\mathrm{kl}}(X\|Y)$ και $D_{\mathrm{kl}}(Y\|X)$ είναι και οι δύο αύξουσες συναρτήσεις της απόστασης των παραμέτρων, $|p-q|$.

Απόδειξη. Γράφουμε $q = p + x$. Ξεκινάμε με την $D_{\mathrm{kl}}(B(n,p)\|B(n,p+x))$. Η παράγωγος της συνάρτησης $h(x) := D_{\mathrm{kl}}(B(n,p)\|B(n,p+x))$ ως προς x είναι

$$h'(x) = \frac{n(1-p)}{-p-x+1} - \frac{np}{p+x}.$$

Είναι εύκολο να δούμε ότι $h'(x) > 0$ για κάθε $x \in (0, 1-p)$ και $h'(x) < 0$ για κάθε $x \in (-p, 0)$. Συνεπώς η $h(x)$ ελαχιστοποιείται στη θέση $x = 0$. Με όμοιο τρόπο αν θέσουμε $g(x) = D_{\mathrm{kl}}(B(n,p+x)\|B(n,p))$ έχουμε

$$g'(x) = n\ln\left(\frac{p+x}{p}\right) - n\ln\left(\frac{-p-x+1}{1-p}\right).$$

Και πάλι ισχύει ότι $g'(x) < 0$ για κάθε $x \in (-p, 0)$ και $g'(x) > 0$ για κάθε $x \in (0, 1-p)$. Έτσι, η $g(x)$ ελαχιστοποιείται και αυτή για $x = 0$. □

Αν προσπαθήσει κανείς να υπολογίσει την απόσταση ολικής κύμανσης δύο PBD θα παρατηρήσει ότι δεν υπάρχει ελπίδα να προκύψει κάποιος εύχρηστος ακριβής τύπος. Σε αυτές τις περιπτώσεις δοκιμάζουμε να προσεγγίσουμε πρώτα τις PBD με κάποιες πιο φιλικές κατανομές και μετά να πάρουμε πάνω και κάτω φράγματα

για την απόσταση των PBDs χρησιμοποιώντας τα αντίστοιχα των "ευκολότερων" κατανομών. Η προσέγγιση, για παράδειγμα, μιας Διωνυμικής με την προϋπόθεση το variance της να μην είναι πολύ μικρό (συνήθως ζητάμε $np \geq 5, n(1-p) \geq 5$). Ας εξηγήσουμε όμως γιατί δουλεύουν τέτοιου είδους προσεγγίσεις από την αρχή.

Πίσω από αυτήν την ιδέα κρύβεται το κεντρικό οριακό θεώρημα (central limit theorem, CLT) το οποίο μας λέει ότι Σε αυτή τη κατεύθυνση είναι και το επόμενο πολύ γνωστό αποτέλεσμα των Berry και Esseen το οποίο μας δείχνει ότι μπορούμε να προσεγγίσουμε μία PBD με κανονική και αποτελεί μία ποσοτικοποίηση του ρυθμού σύγκλισης του κεντρικού οριακού θεωρήματος. Εμείς θα διατυπώσουμε την εκδοχή όπου στο άθροισμα των τυχαίων μεταβλητών $S_n = \sum X_i$ οι X_i δεν ακολουθούν απαραίτητα την ίδια κατανομή μιας και οι PBDs ανήκουν σε αυτήν την κατηγορία.

Θεώρημα 1 (Θεώρημα Berry-Esseen [4])**.** Έστω $X_1, X_2, \ldots, X_n$ ανεξάρτητες φραγμένες τυχαίες μεταβλητές. Τότε

$$d_{\text{kol}}\left(\sum_{i=1}^{n} X_i, \mathcal{N}\left(\mu, \sigma^2\right)\right) \leq C\frac{\sum_{i=1}^{n} \mathbb{E}\left[|X_i|^3\right]}{\sigma^3},$$

όπου

$$\mu = \sum_{i=1}^{n} \mathbb{E}\left[X_i\right], \ \sigma^2 = \sum_{i=1}^{n} \text{Var}\left[X_i\right], \text{ και } \frac{\sqrt{10}+3}{6\sqrt{2\pi}} \leq C \leq 0.4784$$

Η εκτίμηση $C < 0.4784$ είναι η καλύτερη μέχρι στιγμής και δόθηκε από την Shevtsova το 2011, [28]. Το lower bound $\frac{\sqrt{10}+3}{6\sqrt{2\pi}} \leq C$ αποδείχτηκε από τον Esseen. Στην περίπτωση που οι X_i του Θεωρήματος 1 είναι 0/1 δοκιμές Bernoulli με μέση τιμή $\mathbb{E}\left[X_i\right] = p_i$ έχουμε ότι $|X_i|^3 = X_i$ και συνεπώς $\sum_{i=1}^{n} \mathbb{E}\left[|X_i|^3\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] = \sum_{i=1}^{n} p_i$. Έχουμε λοιπόν το επόμενο πόρισμα του Θεωρήματος 1 για τις PBD. Από αυτό το πόρισμα είναι φανερό πως προκύπτει και ο εμπειρικός κανόνας για την προσέγγιση Διωνυμικών κατανομών με κανονικές.

Πόρισμα 2. Έστω $S = \sum_{i=1}^{n} X_i$ μία n-PBD. Τότε

$$d_{\text{kol}}\left(S, \mathcal{N}\left(\mu, \sigma^2\right)\right) \leq C\frac{\mu}{\sigma^3},$$

όπου

$$\mu = \sum_{i=1}^{n} p_i, \ \sigma^2 = \sum_{i=1}^{n} p_i(1-p_i).$$

Η απόσταση ολικής κύμανσης σπάνια μπορεί να υπολογιστεί ακριβώς, συνήθως μπορούμε να δώσουμε μόνο πάνω ή κάτω φράγματά της. Στην περίπτωση 2 κανονικών κατανομών με την ίδια τυπική απόκλιση το επόμενο απλό επιχείρημα δίνει μια ακριβή έκφραση για την απόστασή τους.

Πρόταση 10. Έστω $X \sim \mathcal{N}(\mu_1, \sigma)$, $Y \sim \mathcal{N}(\mu_2, \sigma)$. Τότε

$$d_{\text{tv}}(X, Y) = \text{erf}\left(\frac{|\mu_1 - \mu_2|}{2\sqrt{2}\sigma}\right)$$

Απόδειξη. Έστω f_X και f_Y να είναι οι συναρτήσεις πυκνότητας πιθανότητας των X και Y αντίστοιχα. Χωρίς βλάβη της γενικότητας υποθέτουμε ότι $\mu_1 < \mu_2$ αφού η απόδειξη για την άλλη περίπτωση είναι ακριβώς ίδια Τότε έχουμε

$$f_X(x) \geq f_Y(x) \Leftrightarrow \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma^2}} \geq \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu_2)^2}{2\sigma^2}} \Leftrightarrow |x-\mu_1| \leq |x-\mu_2| \Leftrightarrow x \leq \frac{\mu_1+\mu_2}{2}$$

Συνεπώς,

$$d_{\text{tv}}(X, Y) = \int_{-\infty}^{\frac{\mu_1+\mu_2}{2}} \left(f_X(x) - f_Y(x)\right) \, \mathrm{d}x = \text{erf}\left(\frac{\mu_2 - \mu_1}{2\sqrt{2}\sigma}\right)$$

□

Το επόμενο πρόσφατο αποτέλεσμα των Chen and Leong [23](Θεώρημα 7.1) δείχνει ότι μπορούμε να χρησιμοποιήσουμε μια διακριτοποιημένη Κανονική κατανομή για να προσεγγίσουμε με μεγάλη ακρίβεια μια PBD της οποίας η διασπορά δεν είναι πολύ μικρή. Η διακριτοποίηση που χρησιμοποιείται είναι φυσική σαν ιδέα, ουσιαστικά σε κάθε δυνατή τιμή $i \in \{0, \ldots, n\}$ της PBD αντιστοιχούμε μάζα πιθανότητας ίση με την μάζα της κανονικής από $i - 1/2$ μέχρι $i + 1/2$. Συγκεκριμένα, έστω $X \sim \mathcal{N}(\mu, \sigma)$. Θα συμβολίζουμε με $\mathrm{DN}(\mu, \sigma)$ την διακριτοποιημένη κανονική κατανομή, δηλαδή αν $X_d \sim \mathrm{DN}(\mu, \sigma)$ τότε η X_d είναι μία διακριτή τυχαία μεταβλητή με συνάρτηση μάζας πιθανότητας

$$\mathbb{P}[X_d = k] = \mathbb{P}\left[k - \frac{1}{2} < X \leq k + \frac{1}{2}\right],$$

όπου $k \in \mathbb{Z}$.

Θεώρημα 2 (Θεώρημα 7.1 [23])**.** Έστω X μία PBD και έστω $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathrm{Var}[X]$. Έστω $Y \sim \mathrm{DN}(\mu, \sigma)$. Τότε

$$d_{\mathrm{tv}}(X, Y) \leq \frac{7.6}{\sigma}.$$

Όταν οι διακυμάνσεις δύο κανονικών κατανομών διαφέρουν η επόμενη πρόταση από το [10] δίνει ένα πάνω φράγμα στην απόσταση total variation των δύο κατανομών.

Πρόταση 11 (Πρόταση B.4, [10])**.** Έστω $\mu_1, \mu_2 \in \mathbb{R}$ και $0 < \sigma_1 \leq \sigma_2$. Τότε

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

Το επόμενο Λήμμα δείχνει ότι 2 διακριτοποιημένες Κανονικές κατανομές είναι κοντά αν και μόνο αν οι αντίστοιχες κανονικές κατανομές είναι κοντά. Η απαραίτητη συνθήκη για να ισχύει αυτό είναι η διακύμανση των δύο κανονικών κατανομών να μην είναι πολύ μικρή.

Πρόταση 12. Έστω $X = \mathcal{N}(\mu_1, \sigma_1)$, $Y = \mathcal{N}(\mu_2, \sigma_2)$ δύο κανονικές κατανομές τέτοιες ώστε $d_{\mathrm{tv}}(X, Y) = \varepsilon$, όπου $\varepsilon > 0$. Έστω $X_d \sim \mathrm{DN}(\mu_1, \sigma_1)$, $Y_d \sim \mathrm{DN}(\mu_2, \sigma_2)$. Τότε

$$\varepsilon - u \leq d_{\mathrm{tv}}(X_d, Y_d) \leq \varepsilon \tag{2.1}$$

όπου

$$u = \frac{1}{2}\left(\mathrm{erf}\left(\frac{1}{\sqrt{2}\sigma_1}\right) + \mathrm{erf}\left(\frac{1}{\sqrt{2}\sigma_2}\right)\right)$$

Απόδειξη. Έστω f_X, f_Y οι συναρτήσεις πυκνότητας πιθανότητας των , αντίστοιχα. Έχουμε ότι $\int_{-\infty}^{+\infty} |f_X(x) - f_Y(x)| \mathrm{d}x = 2d_{\mathrm{tv}}(X, Y) = 2\varepsilon$. Αφού η πυκνότητα μιας κανονική κατανομής με μέση τιμή μ είναι αύξουσα στο $(-\infty, \mu]$ και φθίνουσα στο $[\mu, +\infty)$ υπάρχουν το πολύ δύο σημεία r_1, r_2 όπου το πρόσημο της διαφοράς $d(x) := f_X(x) - f_Y(x)$ αλλάζει. Χωρίς βλάβη της γενικότητας υποθέτουμε ότι η $d(x)$ είναι θετική στο $(-\infty, r_1)$ και στο $(r_2, +\infty)$ και αρνητική στο $[r_1, r_2]$. Θέτουμε τώρα $k_1 = \lfloor r_1 \rfloor$ και $k_2 = \lceil r_2 \rceil$. Έτσι μπορούμε να φράξουμε από κάτω την απόσταση ολικής κύμανσης των X_d, Y_d

$$\begin{aligned} 2d_{\mathrm{tv}}(X_d, Y_d) &= \sum_{k=-\infty}^{+\infty} \left| \mathbb{P}\left[k - \frac{1}{2} < X_d \leq k + \frac{1}{2}\right] - \mathbb{P}\left[k - \frac{1}{2} < Y_d \leq k + \frac{1}{2}\right] \right| \\ &= \sum_{k=-\infty}^{+\infty} \left| \int_{k-1/2}^{k+1/2} f_X(x) - f_Y(x)\, \mathrm{d}x \right| \end{aligned} \tag{2.2}$$

$$\geq \sum_{k=-\infty}^{k_1} \int_{k-1/2}^{k+1/2} d(x)\ \mathrm{d}x + \sum_{k=k_1+1}^{k_2} \int_{k-1/2}^{k+1/2} -d(x)\ \mathrm{d}x + \sum_{k=k_2+1}^{+\infty} \int_{k-1/2}^{k+1/2} d(x)\ \mathrm{d}x$$
$$\geq \int_{-\infty}^{k_1} d(x)\ \mathrm{d}x + \int_{k_1+1}^{k_2-1} -d(x)\ \mathrm{d}x + \int_{k_2}^{+\infty} d(x)\ \mathrm{d}x.$$

Αφού $d_{\mathrm{tv}}(X,Y) = \int_{-\infty}^{\infty} |f_X(X) - f_Y(x)|\ \mathrm{d}x$ μας μένει να βρούμε πάνω φράγματα για τα ολοκληρώματα που λείπουν από την παραπάνω έκφραση.

$$\int_{k_1}^{k_1+1} |d(x)|\ \mathrm{d}x + \int_{k_2-1}^{k_2} |d(x)|\ \mathrm{d}x \leq \int_{\mu_1-1}^{\mu_1+1} f_X(x)\mathrm{d}x + \int_{\mu_2-1}^{\mu_2+1} f_Y(x)\mathrm{d}x$$
$$\leq \mathrm{erf}\left(\frac{1}{\sqrt{2}\sigma_1}\right) + \mathrm{erf}\left(\frac{1}{\sqrt{2}\sigma_2}\right)$$

Για να δείξουμε το πάνω φράγμα της ανισότητας (2.1) παρατηρούμε ότι χρησιμοποιώντας την (2.2) έχουμε

$$\begin{aligned} 2d_{\mathrm{tv}}(X_d, Y_d) &= \sum_{k=-\infty}^{+\infty} \left| \int_{k-1/2}^{k+1/2} f_X(x) - f_Y(x)\ \mathrm{d}x \right| \\ &\leq \sum_{k=-\infty}^{+\infty} \int_{k-1/2}^{k+1/2} |f_X(x) - f_Y(x)\ |\,\mathrm{d}x \\ &\leq \int_{-\infty}^{+\infty} |f_X(x) - f_Y(x)\ |\,\mathrm{d}x \\ &= 2\varepsilon \end{aligned}$$

□

Επιστρέφουμε τώρα στο αρχικό πρόβλημα της εκτίμησης της απόστασης ολικής κύμανσης ανάμεσα σε δύο PBDs. Χρησιμοποιώντας την προσέγγιση με διακριτοποιημένη κανονική κατανομή και την Πρόταση 12 μπορούμε να πάρουμε ένα πάνω φράγμα στην απόσταση των PBDs.

Λήμμα 1. Έστω X και Y PBDs με μέση τιμή μ_X, μ_Y αντίστοιχα και διασπορά σ_X^2, σ_Y^2 αντίστοιχα έτσι ώστε $0 < \sigma_X \leq \sigma_Y$. Τότε

$$d_{\mathrm{tv}}(X,Y) \leq \frac{1}{2}\left(\frac{|\mu_1 - \mu_2|}{\sigma_1} + \frac{\sigma_2^2 - \sigma_1^2}{\sigma_1^2}\right) + 7.6\left(\frac{1}{\sigma_X} + \frac{1}{\sigma_Y}\right)$$

Απόδειξη. Ξεκινάμε προσεγγίζοντας τις δύο PBDs χρησιμοποιώντας διακριτοποιημένες κανονικές κατανομές $D_X = \mathrm{DN}(\mu_X, \sigma_Y)$, $D_Y = \mathrm{DN}(\mu_Y, \sigma_Y)$. Από το Θεώρημα 2 έχουμε $d_{\mathrm{tv}}(X, D_X) \leq 7.6/\sigma_X$ και $d_{\mathrm{tv}}(Y, D_Y) \leq 7.6/\sigma_Y$. Έστω τώρα οι συνεχείς κανονικές κατανομές $\mathcal{N}(\mu_X, \sigma_X)$, $\mathcal{N}(\mu_Y, \sigma_Y)$. Από την Πρόταση 11 έχουμε ότι

$$d_{\mathrm{tv}}(\mathcal{N}(\mu_X, \sigma_Y), \mathcal{N}(\mu_Y, \sigma_Y)) \leq \frac{1}{2}\left(\frac{|\mu_X - \mu_Y|}{\sigma_Y} + \frac{\sigma_Y^2 - \sigma_X^2}{\sigma_X^2}\right)$$

Για να περάσουμε τώρα στις διακριτοποιημένες D_X και D_Y θα χρησιμοποιήσουμε την Πρόταση 12 για να πάρουμε

$$d_{\mathrm{tv}}(D_X, D_Y) \leq \frac{1}{2}\left(\frac{|\mu_X - \mu_Y|}{\sigma_Y} + \frac{\sigma_Y^2 - \sigma_X^2}{\sigma_X^2}\right)$$

Και τελικά, χρησιμοποιώντας την τριγωνική ανισότητα της απόστασης ολικής κύμανσης,

$$d_{\mathrm{tv}}(X,Y) \leq \frac{1}{2}\left(\frac{|\mu_X - \mu_Y|}{\sigma_Y} + \frac{\sigma_Y^2 - \sigma_X^2}{\sigma_X^2}\right) + 7.6\left(\frac{1}{\sigma_X} + \frac{1}{\sigma_Y}\right)$$

□

Με την ίδια λογική και χρησιμοποιώντας την προσέγγιση με κανονική κατανομή μπορούμε να πάρουμε και ένα καλό κάτω φράγμα για την απόσταση ολικής κύμανσης ανάμεσα σε δύο PBDs. Τα κάτω φράγματα στην απόσταση δύο κατανομών, παρόλο που επισκιάζονται συνήθως από τα πάνω φράγματα που μας χρειάζονται στα πάνω φράγματα της δειγματικής πολυπλοκότητας, είναι πολύ χρήσιμα για την κατασκευή κάτω φραγμάτων δειγματικής πολυπλοκότητας όπως θα δούμε στα επόμενα κεφάλαια.

Πρόταση 13. Έστω και Y δύο PBDs με $\mathbb{E}[X] = \mu_X$, $\mathbb{E}[Y] = \mu_Y$, $\mathrm{Var}[X] = \sigma_X$, $\mathrm{Var}[Y] = \sigma_Y$, τέτοιες ώστε $\sigma_Y^2 \geq \sigma_X^2$. Τότε,

$$d_{\mathrm{tv}}(X,Y) \geq \mathrm{erf}\left(\frac{|\mu_X - \mu_Y|}{2\sqrt{2}\sigma}\right) - \frac{\sigma_Y^2 - \sigma_X^2}{\sigma_X^2} - \frac{1}{2}\left(\mathrm{erf}\left(\frac{1}{\sqrt{2}\sigma_X}\right) + \mathrm{erf}\left(\frac{1}{\sqrt{2}\sigma_Y}\right)\right) - 7.6\left(\frac{1}{\sigma_X} + \frac{1}{\sigma_Y}\right)$$

Απόδειξη. Θεωρούμε τις συνεχείς κανονικές κατανομές $\mathcal{N}(\mu_X, \sigma_X)$, $\mathcal{N}(\mu_Y, \sigma_Y)$. Για να χρησιμοποιήσουμε τον ακριβή τύπο για την απόστασή τους που δίνει η Πρόταση 10 πρέπει οι δύο κανονικές κατανομές να έχουμε την ίδια διακύμανση. Χρησιμοποιώντας όμως το πάνω φράγμα της Πρότασης 11 μπορούμε να αντικαταστήσουμε την $\mathcal{N}(\mu_Y, \sigma_Y)$ με την $\mathcal{N}(\mu_X, \sigma_X)$ χωρίς να απομακρυνθούμε πολύ από την αρχική (στην περίπτωση που οι δύο διασπορές σ_X^2 και σ_Y^2 δεν διαφέρουν σημαντικά. Συγκεκριμένα,

$$d_{\mathrm{tv}}(\mathcal{N}(\mu_Y, \sigma_Y), \mathcal{N}(\mu_X, \sigma_X)) \leq \frac{\sigma_Y^2 - \sigma_X^2}{\sigma_X^2}$$

Χρησιμοποιώντας τώρα την πρόταση 10 έχουμε ότι

$$d_{\mathrm{tv}}(\mathcal{N}(\mu_Y, \sigma), \mathcal{N}(\mu_X, \sigma_X)) = \mathrm{erf}\left(\frac{|\mu - \mu|}{2\sqrt{2}\sigma}\right)$$

και από την τριγωνική ανισότητα έχουμε

$$d_{\mathrm{tv}}(\mathcal{N}(\mu_Y, \sigma), \mathcal{N}(\mu_X, \sigma_X)) \geq \mathrm{erf}\left(\frac{|\mu - \mu|}{2\sqrt{2}\sigma}\right) - \frac{\sigma_Y^2 - \sigma_X^2}{\sigma_X^2}$$

Έχουμε πλέον το κάτω φράγμα για τις συνεχείς κανονικές κατανομές και μας μένει να το "μεταφράσουμε" πρώτα σε διακριτές κανονικές κατανομές και στις συνέχεια σε PBDs. Από την Πρόταση 12 έχουμε ότι

$$d_{\mathrm{tv}}(\mathrm{DN}(\mu_Y, \sigma_Y), \mathrm{DN}(\mu_X, \sigma_X)) \geq \mathrm{erf}\left(\frac{|\mu_X - \mu_Y|}{2\sqrt{2}\sigma}\right) - \frac{\sigma_Y^2 - \sigma_X^2}{\sigma_X^2} - \frac{1}{2}\left(\mathrm{erf}\left(\frac{1}{\sqrt{2}\sigma_X}\right) + \mathrm{erf}\left(\frac{1}{\sqrt{2}\sigma_Y}\right)\right)$$

Χρησιμοποιώντας, τέλος, το Θεώρημα 2 και την τριγωνική ανισότητα αποδεικνύουμε το ζητούμενο κάτω φράγμα. □

Δείχνουμε τώρα ένα ακόμα κάτω φράγμα στην απόσταση ολικής κύμανσης ανάμεσα σε δύο PBDs βασισμένο στην απόσταση που έχουν οι μέσες τιμές τους. Σε αντίθεση με το κάτω φράγμα της Πρότασης 13 το επόμενο κάτω φράγμα δεν χρειάζεται οι δύο PBDs να έχουν κοντινές διασπορές. Το μόνο που χρειάζεται είναι να έχουν διασπορές μεγαλύτερες από μία σταθερά.

Λήμμα 2. Έστω X και Y δύο PBDs με μέσες τιμές $\mu_X = \mathbb{E}[X]$, $\mu_Y = \mathbb{E}[Y]$ και διασπορές $\sigma_X^2 = \mathrm{Var}[X]$, $\sigma_Y^2 = \mathrm{Var}[Y]$. Τότε, για κάθε $\varepsilon > 0$ τέτοιο ώστε $\sigma_X^2, \sigma_Y^2 \geq \ln(\frac{2}{1-\varepsilon})$, αν $|\mu_Y - \mu_X| > 2\sqrt{\ln(\frac{2}{1-\varepsilon})}(\sigma_X + \sigma_Y)$, τότε $d_{\mathrm{tv}}(X,Y) > \varepsilon$.

Απόδειξη. Για απλότητα, θέτουμε $\lambda \equiv 2\sqrt{\ln(\frac{2}{1-\varepsilon})}$ και χωρίς βλάβη της γενικότητας υποθέτουμε ότι $\mu_Y > \mu_X + \lambda(\sigma_X + \sigma_Y)$. Από τον ορισμό της απόστασης ολικής κύμανσης, έχουμε

$$2d_{\mathrm{tv}}(X,Y) = \sum_{i=0}^{\infty} |\mathbb{P}[X = i] - \mathbb{P}[Y = i]|$$

$$
\begin{aligned}
&\geq \sum_{i=0}^{\mu_X+\lambda\sigma_X} (\mathbb{P}[X=i] - \mathbb{P}[Y=i]) + \sum_{i=\mu_Y-\lambda\sigma_Y}^{\infty} (\mathbb{P}[Y=i] - \mathbb{P}[X=i]) \\
&= (\mathbb{P}[X \leq \mu_X + \lambda\sigma_X] - \mathbb{P}[Y \leq \mu_X + \lambda\sigma_X]) + \\
&\quad + (\mathbb{P}[Y \geq \mu_Y - \lambda\sigma_Y] - \mathbb{P}[X \geq \mu_Y - \lambda\sigma_Y]) \\
&\geq (1 - \mathbb{P}[X > \mu_X + \lambda\sigma_X] - \mathbb{P}[Y < \mu_Y - \lambda\sigma_Y]) + \\
&\quad + (1 - \mathbb{P}[Y < \mu_Y - \lambda\sigma_Y] - \mathbb{P}[X > \mu_X + \lambda\sigma_X]) \\
&> (1 - (1-\varepsilon)/2 - (1-\varepsilon)/2) + (1 - (1-\varepsilon)/2 - (1-\varepsilon)/2) = 2\varepsilon
\end{aligned}
$$

Για την δεύτερη ανισότητα, χρησιμοποιούμε το γεγονός ότι $\mu_X + \lambda\sigma_X < \mu_Y - \lambda\sigma_Y$. Συνεπώς, $\mathbb{P}[Y \leq \mu_X + \lambda\sigma_X] \leq \mathbb{P}[Y < \mu_Y - \lambda\sigma_Y]$ and $\mathbb{P}[X \geq \mu_Y - \lambda\sigma_Y] \leq \mathbb{P}[X > \mu_X + \lambda\sigma_X]$. Για την τελευταία ανισότητα, χρησιμοποιούμε την Πρόταση 3 με $\lambda = 2\sqrt{\ln(\frac{2}{1-\varepsilon})}$ για να πάρουμε $\mathbb{P}[X > \mu_X + \lambda\sigma_X] < (1-\varepsilon)/2$ και $\mathbb{P}[Y < \mu_Y - \lambda\sigma_Y] < (1-\varepsilon)/2$. □

Αξίζει να επισημάνουμε ότι παρόλο που το κάτω φράγμα της Πρότασης 13 έχει αρκετές συνθήκες για να δουλέψει, είναι ασυμπτωτικά καλύτερο από το γενικότερο κάτω φράγμα του Λήμματος 2 στις περιπτώσεις αυτές. Ας δούμε ένα παράδειγμα για να καταλάβουμε την διαφορά τους. Ας θεωρήσουμε τις διωνυμικές $X = B(n, 1/2)$ και $Y = B(n, 1/2 - 1/\sqrt{n})$. Έχουμε $\mu_X = n/2$, $\sigma_X^2 = n/4$, $\mu_Y = n/2 - \sqrt{n}$, $\sigma_Y = n/4 - 1$. Το Λήμμα 2 λύνοντας ως προς ε σε αυτή δίνει το κάτω φράγμα

$$d_{\text{tv}}(X, Y) \geq 1 - 2\exp\left(\left(\frac{\mu_Y - \mu_X}{2(\sigma_X + \sigma_Y)}\right)^2\right)$$

Αντικαθιστώντας βλέπουμε ότι

$$\frac{\mu_Y - \mu_X}{2(\sigma_X + \sigma_Y)} = O(\varepsilon)$$

Συνεπώς χρησιμοποιώντας την προσέγγιση $1 - \mathrm{e}^x \approx x$ για x κοντά στο 0 παίρνουμε ότι το κάτω φράγμα για την απόσταση ολικής κύμανσης που δίνει το Λήμμα 2 σε αυτή τη περίπτωση είναι $\Omega(\varepsilon^2)$. Ας δούμε τώρα το κάτω φράγμα που παίρνουμε αν χρησιμοποιήσουμε την Πρόταση 13. Οι όροι $1/\sigma_X$ και $1/\sigma_Y$ είναι $O(1/\sqrt{n})$ και κατά συνέπεια αμελητέοι για αρκετά μεγάλο n. Αντίστοιχα χρησιμοποιώντας το πάνω φράγμα του Chu 6 για την συνάρτηση σφάλματος της κανονικής κατανομής παίρνουμε $\operatorname{erf}(1/\sqrt{n}) = O(1/\sqrt{n})$ και κατά συνέπεια μπορούμε να αγνοήσουμε και αυτούς τους όρους ως αμελητέους. Όμοια μπορούμε να αγνοήσουμε τον όρο $\frac{\sigma_Y^2 - \sigma_X^2}{\sigma_X}$. Χρησιμοποιώντας τώρα το κάτω φράγμα του Chu 6 και συγκεκριμένα το Πόρισμα 1 παίρνουμε ότι το κάτω φράγμα για την απόσταση ολικής κύμανσης είναι $\Omega(\epsilon)$. Η διαφορά λοιπόν των δύο αποτελεσμάτων σε αυτήν την περίπτωση ήταν πολύ μεγάλη. Στο κεφάλαιο 4 θα χρησιμοποιήσουμε αυτό το κάτω φράγμα της απόστασης ολικής κύμανσης για να κατασκευάσουμε ένα κάτω φράγμα για την δειγματική πολυπλοκότητα της μάθησης των δυνάμεων μιας διωνυμικής κατανομής.

Στα προηγούμενα φράγματα χρησιμοποιήσαμε κατά κύριο λόγο την μέση τιμή και την διακύμανση για να μετρήσουμε την απόσταση δύο PBDs. Για να πάρουμε πιο ισχυρές εκτιμήσεις πρέπει να υπολογίσουμε την συνεισφορά και των υπολοίπων ροπών στην απόσταση ολικής κύμανσης. Το επόμενο θεώρημα αποδείχτηκε από τους Διακονικόλα, Kane και Stewart και ποσοτικοποιεί την εξάρτηση της απόστασης ολικής κύμανσης από την απόσταση των αθροισμάτων των δυνάμεων των παραμέτρων $\sum_{i=1}^{n} p_i^k$.

Μία εύλογη και φυσική επιλογή προσέγγισης μίας PBD είναι να χρησιμοποιήσουμε μια κατάλληλα επιλεγμένη διωνυμική κατανομή. Για να φράξουμε την απόσταση ολικής κύμανσης ανάμεσα σε μία PBD και μία διωνυμική κατανομή μπορούμε να χρησιμοποιήσουμε το επόμενο αποτέλεσμα του Roos.

Θεώρημα 3 (Θεώρημα 2, [26]). Έστω S μία n-PBD με διάνυσμα πιθανοτήτων $\boldsymbol{p} = (p_i)_{i=1}^n$ και έστω $p \in (0,1)$. Τότε

$$d_{\text{tv}}\left(X, B(n,p)\right) \leq \frac{\sqrt{\text{e}}}{2} \frac{\sqrt{\tau(p)}}{\left(1-\sqrt{\tau(p)}\right)^2},$$

όπου

$$\tau(p) = \frac{\gamma_1(p)^2 + 2\gamma_2(p)}{2np(1-p)}, \quad \gamma_j(p) = \sum_{i=1}^n (p-p_i)^j$$

Αν θέλει κανείς να προσεγγίσει μία n-PBD με διάνυσμα πιθανοτήτων $\boldsymbol{p}$ με μια διωνυμική $B(n,p)$ μια λογική επιλογή για το p είναι να χρησιμοποιήσει τον μέσο όρο των p_i, $\bar{p} = (1/n)\sum_{i=1}^n p_i$. Χρησιμοποιώντας την μέθοδο του Stein ο Ehm έδειξε το επόμενο αποτέλεσμα προσέγγισης PBD με διωνυμική.

Θεώρημα 4 (Θεώρημα 1, [18]). Έστω S μία n-PBD με διάνυσμα πιθανοτήτων $\boldsymbol{p}$, και $\bar{p} = (1/n)\sum_{i=1} p_i, \bar{q} = 1-\bar{p}$. Τότε

$$C \min\left(\frac{1}{npq}, 1\right) \gamma_2(\bar{p}) \leq d_{\text{tv}}\left(S, B(n,\bar{p})\right) \leq \frac{1-p^{n+1}-q^{n+1}}{(n+1)\bar{p}\bar{q}} \gamma_2(\bar{p})$$

όπου

$$\gamma_2(p) = \sum_{i=1}^n (p_i - p)^2$$

Στο Κεφάλαιο 3 θα δείξουμε πως μπορεί κανείς να μάθει σε TVD δυνάμεις διωνυμικών κατανομών. Το επόμενο πολύ χρήσιμο πόρισμα του Λήμματος 3 μας δείχνει πόσο καλά πρέπει να μάθουμε την παράμετρο μιας διωνυμικής κατανομής ώστε να πετύχουμε μικρή απόσταση ολικής κύμανσης. Σε αυτήν την περίπτωση λοιπόν το πρόβλημα της μάθησης σε TVD είναι ισοδύναμο με την προσέγγιση της παραμέτρου p της διωνυμικής κατανομής, κάτι το οποίο δεν ισχύει όπως θα δούμε στην γενική περίπτωση των PBD κατανομών.

Πόρισμα 3. Έστω $\varepsilon < 1/2$, $n \geq 1$. Έστω $B(n,p)$, $B(n,q)$ δύο διωνυμικές κατανομές τέτοιες ώστε $|p-q| \leq \varepsilon\sqrt{\frac{p(1-p)}{n}}$ τότε $d_{\text{tv}}\left(B(n,q), B(n,p)\right) \leq 2\sqrt{\text{e}}\varepsilon$.

Απόδειξη. Ακολουθώντας τον συμβολισμό του Λήμματος 3 έχουμε $\gamma_1(p) \leq \varepsilon\sqrt{n}p(1-p)$, $\gamma_2(p) \leq \varepsilon^2 p(1-p)$, $\tau(p) \leq \varepsilon^2/2 + \varepsilon^2/(2n) \leq \varepsilon^2$. Συνεπώς, $d_{\text{tv}}\left(B(n,q), B(n,p)\right) \leq \frac{\sqrt{\text{e}}\varepsilon}{2(1-\varepsilon)^2} \leq 2\sqrt{\text{e}}\varepsilon$ όταν $\varepsilon < 1/2$. □

2.3 Έννοιες από την Θεωρία Πληροφορίας

Οι ιδέας της θεωρίας πληροφορίας είναι γνωστό ότι βρίσκουν εφαρμογή σε ένα ευρύ φάσμα κλάδων των μαθηματικών και της επιστήμης υπολογιστών. Για την ακρίβεια λίγα είναι τα κομμάτια αυτών των επιστημών όπου η θεωρία πληροφορίας δεν έχει κάτι να προσφέρει. Στον Κεφάλαιο 3 θα παρουσιάσουμε τεχνικές κατασκευής κάτω φραγμάτων του minimax risk, δηλαδή κάτω φράγματα στην δειγματική πολυπλοκότητα των αλγορίθμων εκτίμησης συναρτήσεων ακολουθιών κατανομών. Οι τεχνικές που θα παρουσιάσουμε είναι στη βάση τους πληροφοριοθεωρητικές. Το γεγονός ότι μπορούμε να χρησιμοποιήσουμε τέτοιες τεχνικές της θεωρίας πληροφορίας για την κατασκευή κάτω φραγμάτων δεν είναι τυχαίο. Τα προβλήματα μάθησης (εκτίμησης και ελέγχου) είναι πολύ συγγενικά με τα προβλήματα συμπίεσης και μεταφοράς δεδομένων. Ένα πρόβλημα μεταφοράς δεδομένα αποτελείται από τρεις φάσεις: συμπίεση, μεταφορά και αποσυμπίεση των δεδομένων. Το κανάλι μεταφοράς εισάγει θόρυβο με συνέπεια να πρέπει να μεταφέρουμε μεγαλύτερη ποσότητα δεδομένων από αυτά που πραγματικά θέλουμε να φτάσουν στον δέκτη καθώς πρέπει να υπάρχει κάποια ποσότητα επικάλυψης

ώστε να μπορεί να γίνει η ανακατασκευή του μηνύματος.

$$\text{Source} \longrightarrow \text{Encoder} \xrightarrow{X} \overset{p(y|x)}{\text{Channel}} \xrightarrow{Y} \text{Decoder} \xrightarrow{\hat{f}} \text{Receiver}$$

Η θεωρία πληροφορίας μελετά την ακριβή ποσότητα αυτής της επικάλυψης ώστε η ανακατασκευή να πετυχαίνει με μεγάλη πιθανότητα. Ας παρατηρήσουμε τώρα τη δομή ενός προβλήματος εκτίμησης. Σε ένα πρόβλημα παλινδρόμησης(regression) έχουμε

$$\xrightarrow{f(X_1),f(X_2),\dots} \overset{p(y|f(x))}{\text{Channel}} \xrightarrow{Y_1,Y_2,\dots} \text{Decoder} \xrightarrow{\hat{f}}$$

Σε αυτή την περίπτωση μπορούμε να φανταστούμε τον αλγόριθμο μάθησης ως τον Αποκωδικοποιητή και τον θόρυβο του Καναλιού ως τον θόρυβο που έχουν οι τιμές της άγνωστης συνάρτησης f που καλούμαστε να προσεγγίσουμε σε ένα γενικό πρόβλημα παλινδρόμησης. Ένα άλλο παράδειγμα αποτελεί η περίπτωση του density estimation όπου ο decoder βλέπει απλά δείγματα (χωρίς ταμπέλες) από μια άγνωστη κατανομή.

$$\theta \longrightarrow \overset{p_\theta(x)}{\text{Channel}} \xrightarrow{X_1,X_2,\dots} \text{Decoder} \xrightarrow{\hat{p}}$$

Έχοντας παρακινήσει λοιπόν την ανάγκη της προχωράμε στην εισαγωγή μερικών βασικών εννοιών και αποτελεσμάτων της θεωρίας πληροφόριας που θα μας χρειαστούν στα επόμενα κεφάλαια.

Έστω μία διακριτή τυχαία μεταβλητή X με τιμές στο σύνολο $\mathcal{X}$ και με συνάρτηση μάζας πιθανότητας f. Ξεκινάμε με την βασική έννοια της εντροπίας () της τυχαίας μεταβλητής X η οποία ορίζεται ως

$$H(X) = -\sum_{x\in\mathcal{X}} f(x)\ln f(x),$$

και μετρά την αβεβαιότητα μιας τυχαίας μεταβλητής. Είναι προφανές ότι η εντροπία είναι θετική ποσότητα. Σημειώνουμε σε αυτό το σημείο ότι όλα τα πληροφοριοθεωρητικά μεγέθη μπορούν να οριστούν είτε χρησιμοποιώντας φυσικούς λογάριθμους και τότε η πληροφορία μετριέται σε nats(1 nat $\approx$ 1.44 bits) είτε με λογάριθμους με βάση 2 οπότε η πληροφορία μετριέται σε bits. Μπορούμε τώρα να ορίσουμε την δεσμευμένη εντροπία $H(X|Y)$ ως την εντροπία της X δεσμευμένη στη γνώση της τυχαίας μεταβλητής Y. Έχουμε

$$H(X|Y=y) = -\sum_x p(x|y)\log p(x|y) \text{ και } H(X|Y) = \sum_y p(y)H(X|Y=y),$$

όπου $p(x|y)$ είναι η πυκνότητα μάζας πιθανότητας της X δεδομένου ότι $Y=y$. Είναι φανερό από το Σχήμα 2.1 ότι $H(X,Y) = () + H(Y|X)$. Εύκολα επαληθεύουμε ότι

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x,y)\ln p(x,y) \\ &= -\sum_{x,y} p(x)p(y|x)\ln p(x)p(y|x) \\ &= -\sum_x p(x)\ln p(x)\sum_y p(y|x) - \sum_x p(x)\sum_y p(y|x)\ln p(y|x) \\ &= H(X) + H(Y|X) \end{aligned}$$

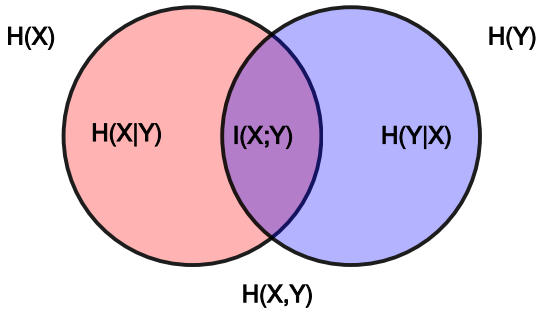
Επαγωγικά προκύπτει ότι

$$H(X_1,\dots,X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_2,X_1) + \dots + H(X_n|X_{n-1},\dots X_1) \tag{2.3}$$

Χρησιμοποιώντας τον γενικό ορισμό της απόστασης Kullback-Leibler της Ενότητας 2.2 ορίζουμε την από κοινού πληροφορία δύο τυχαίων μεταβλητών X,Y (όχι απαραίτητα διακριτών) με από κοινού κατανομή P_{XY} και περιθώριες P_X P_Y ως

$$I(X;Y) = D_{\text{kl}}\left(P_{XY}\|P_X\times P_Y\right).$$

Ο μετροθεωρητικός ορισμός της απόστασης Kullback-Leibler και κατ' επέκταση της από κοινού πληροφορίας $I(X;Y)$ μπορεί να γενικεύσει τον ορισμό της εντροπίας ως $H(X) = I(X;X)$. Με την διαίσθηση που



Σχήμα 2.1: Διάγραμμα των $H(X), H(Y), H(X|Y), I(X;Y)$

αποκτήσαμε στην Ενότητα 2.2 για τις f-αποκλίσεις καταλαβαίνουμε ότι η από κοινού πληροφορία μετράει την απόσταση της από κοινού κατανομής με την κατανομή γινόμενο. Δηλαδή η από κοινού πληροφορία είναι μία ποσότητα που μετράει την εξάρτηση δύο μεταβλητών. Όταν λοιπόν η κατανομή γινόμενο $P_X \times P_Y$ είναι κοντά στην από κοινού P_{XY} τότε οι X, Y έχουν μικρή εξάρτηση και η από κοινού πληροφορία είναι μικρή. Από τις ιδιότητες της απόκλισης Kullback-Leibler έχουμε ότι η $I(X;Y) \geq 0$ με την ισότητα να ισχύει αν και μόνο αν οι , είναι ανεξάρτητες.

Σε αυτό το σημείο πρέπει να αναρωτηθούμε ποια είναι η διαφορά της από κοινού πληροφορίας με ποσότητες όπως π.χ. η συνδιακύμανση η οποία μετρά και αυτή την εξάρτηση δύο τυχαίων μεταβλητών. Ήδη από το γεγονός ότι η $I(X;Y)$ μηδενίζεται αν και μόνο αν οι , είναι ανεξάρτητες δείχνει ότι είναι ισχυρότερο μέτρο της εξάρτησης από τη συνδυακύμανση. Για παραδείγμα ας θεωρήσουμε την ομοιόμορφα κατανεμημένη $X \sim U[-1,1]$ και την X^2. Αυτές οι τυχαίες μεταβλητές προφανώς δεν είναι ανεξάρτητες και συνεπώς $I(X;X^2) > 0$. Από την άλλη Η συνδιακύμανσή τους είναι

$$\operatorname{cov}(X, X^2) = \mathbb{E}\left[XX^2\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[X^2\right] = \mathbb{E}\left[X^3\right] - 0 = 0.$$

Αυτό συμβαίνει επειδή η από κοινού πληροφορία λαμβάνει υπ' όψιν της *όλες* τις εξαρτήσεις των τυχαίων μεταβλητών και όχι μόνο τις δευτεροτάξιες όπως η συνδιακύμανση. Ξεκινώντας από τον ορισμό της από κοινού πληροφορίας έχουμε

$$\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \ln \frac{p(x|y)}{p(x)} \\
&= \sum_{x,y} p(x,y) \ln p(x|y) - \sum_{x,y} p(x,y) \ln p(x) \\
&= \sum_{x,y} p(x,y) \ln p(x|y) - \sum_{x} p(x) \ln p(x) \\
&= \sum_{x,y} p(y)p(x|y) \ln p(x|y) - \sum_{x} p(x) \ln p(x) \\
&= \sum_{y} p(y) \left(\sum_{x} p(x|y) \ln p(x|y) \right) - \sum_{x} p(x) \ln p(x) \\
&= H(X) - H(X|Y)
\end{aligned}$$

Γνωρίζουμε όμως ότι $I(X;Y) \geq 0$ άρα παίρνουμε την γνωστή ανισότητα

$$H(X|Y) \leq H(X), \tag{2.4}$$

δηλαδή η γνώση μιας τυχαίας μεταβλητής μπορεί μόνο να μειώσει την εντροπία μιας άλλης, κάτι που είναι εντελώς σύμφωνο με την διαίσθησή μας και το διάγραμμα του Σχήματος 2.1. Αξίζει να δούμε ένα παράδειγμα (σελίδα 30, [8]) για να κατανοήσουμε καλύτερα την ανισότητα 2.4. Ας υποθέσουμε ότι η από κοινού κατανομή των X, Y είναι

Y \ X	1	2
1	0	3/4
2	1/8	1/8

Όπως βλέπουμε η περιθώρια πυκνότητα της είναι $(1/8, 7/8)$ ενώ της Y είναι $(3/4, 1/4)$. Συνεπώς, $H(X) = -1/8 \log 1/8 - 7/8 \log 7/8 = 0.544$ bits, $H(Y) = -3/4 \log 3/4 - 1/4 \log 1/4 = 0.811$ bits. Κατ' αρχάς παρατηρούμε ότι η Y έχει μεγαλύτερη εντροπία από την X που είναι αναμενόμενο αφού η Y είναι πιο κοντά στην ομοιόμορφη $(1/2, 1/2)$ που γνωρίζουμε ότι μεγιστοποιεί την εντροπία. Ας δούμε τώρα πως επηρεάζει η γνώση της Y την εντροπία της X. Έχουμε $(| = 1) = 0$ bits αφού σε αυτή τη περίπτωση η $p(X|Y = 1)$ είναι $(0, 1)$. Από την άλλη $H(X|Y = 2) = 1$ bits αφού η $p(X|Y = 2)$ είναι η $(1/2, 1/2)$. Παρατηρούμε λοιπόν ότι συγκεκριμένες παρατηρήσεις μπορεί να αυξήσουν την εντροπία μιας τυχαίας μεταβλητής αλλά *η μέση τιμή της* μετά την παρατήρηση μιας άλλης τυχαίας μεταβλητής πάντα μειώνεται. Στο παράδειγμα $H(X|Y) = 3/4 \times 0 + 1/4 \times 1 = 0.25 < 0.544 = H(X)$.

Το γεγονός ότι η απόκλιση Kullback-Leibler είναι μη αρνητική ποσότητα έχει την πολύ ενδιαφέρουσα συνέπεια ότι η ομοιόμορφη κατανομή είναι η κατανομή μέγιστης εντροπίας πάνω σε ένα πεπερασμένο σύνολο $\mathcal{X}$. Για να το δούμε αυτό ας υπολογίσουμε την απόκλιση Kullback-Leibler της ομοιόμορφης U η οποία έχει πυκνότητα $u(x) = 1/|\mathcal{X}|$ με μια άλλη τυχαία μεταβλητή X με πυκνότητα $p(x)$.

$$D_{\text{kl}}(X\|U) = \sum_{x\in\mathcal{X}} p(x) \ln \frac{p(x)}{u(x)} = \sum_{x\in\mathcal{X}} p(x) \ln p(x) - \sum_{x\in\mathcal{X}} p(x) \ln u(x) = \ln |\mathcal{X}| - H(X).$$

Συνεπώς $H(X) \leq \ln |\mathcal{X}| = H(U)$.

Συνεχίζουμε με τον φυσικό ορισμό της δεσμευμένης από κοινού πληροφορίας

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = D_{\text{kl}}(p(X,Y|Z)\|p(X|Z)p(Y|Z))$$

η οποία έχει και αυτή έναν κανόνα αλυσίδας αντίστοιχο με τον κανόνα αλυσίδας της εντροπίας (2.3)

$$I(X_1, X_2, \ldots X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, \ldots, X_1). \tag{2.5}$$

Κλείνουμε αυτήν την ενότητα με μία πολύ σημαντική και χρήσιμη ανισότητα. Ας υποθέσουμε ότι έχουμε μία αλυσίδα Markov $\to Y \to Z$. Αυτό σημαίνει ότι οι Z, X είναι ανεξάρτητες δεδομένης της Y. Τέτοιες αλυσίδες θα δούμε στο Κεφάλαιο 3 μοντελοποιούν την διαδικασία της εκτίμησης ενός μεγέθους από παρατηρήσεις του. Σε αυτή τη περίπτωση η τυχαία μεταβλητή είναι η άγνωστη ποσότητα που θέλουμε να εκτιμήσουμε, η μεταβλητή Y είναι το δείγμα (οι παρατηρήσεις), και η Z είναι μία στατιστική, δηλαδή $Z = f(Y)$. Η αλυσίδα $X \to Y \to f(Y)$ είναι Markov αφού

$$\mathbb{P}[f(Y) = w|X = x, Y = y] = \frac{\mathbb{P}[f(Y) = w, X = x, Y = y]}{\mathbb{P}[X = x, Y = y]} = \begin{cases} 0 \text{ if } y \in f^{-1}(w) \\ 1 \text{ else} \end{cases}.$$

Αντίστοιχα

$$\mathbb{P}[f(Y) = w|Y = y] = \frac{\mathbb{P}[f(Y) = w, = y]}{\mathbb{P}[Y = y]} = \begin{cases} 0 \text{ if } y \in f^{-1}(w) \\ 1 \text{ else} \end{cases}.$$

Άρα οι δύο πιθανότητες είναι πάντα ίσες και κατά συνέπεια οι $f(Y), X$ είναι ανεξάρτητες δεδομένης της Y. Η στατιστική f επιτρέπεται να μαθαίνει την X *μόνο* μέσω του δείγματος Y. Η επόμενη πρόταση είναι ιδιαίτερα ενδιαφέρουσα και χρήσιμη αφού μας λέει ότι αφού τραβήξουμε ένα δείγμα από μια κατανομή δεν μπορούμε να

εφαρμόσουμε κάποιον "έξυπνο" μετασχηματισμό σε αυτό f ώστε να αυξήσουμε την πληροφορία μας για την τυχαία μεταβλητή

Πρόταση 14 (Ανισότητα Data-processing)**.** Αν $X \to Y \to Z$ τότε $I(X;Y) \geq I(X;Z)$.

Απόδειξη. Από τον κανόνα της αλυσίδας για την από κοινού πληροφορία

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

Αντίστοιχα

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

Άρα $(;)+(;|) = I(X;Z)+I(X;Y|Z)$ και αφού οι , είναι ανεξάρτητες δεδομένης της Y έχουμε $I(X;Z|Y) = 0$. Τέλος από το γεγονός ότι η από κοινού πληροφορία είναι μη αρνητική παίρνουμε $(;Y) \geq I(X;Z)$. □

Στην προηγούμενη απόδειξη καθοριστικό ρόλο έπαιξε το γεγονός ότι η αλυσίδα $X \to Y \to Z$ είναι Markov.

Κεφάλαιο 3

Κάτω Φράγματα για το Minimax Risk

Ας ξεκινήσουμε εισάγοντας κάποιον απαραίτητο συμβολισμό για τον ορισμό του minimax risk για την εκτίμηση συναρτήσεων ακολουθιών κατανομών. Σημειώνουμε σε αυτό το σημείο πως θα δώσουμε την βασική ανάλυση στο γενικότερο πλαίσιο των συναρτήσεων ακολουθιών κατανομών μιας και τα κλασικά θεωρήματα για την εκτίμηση συναρτήσεων μίας κατανομής προκύπτουν άμεσα από τα γενικότερα. Χρειαζόμαστε μία ιεραρχία στον συμβολισμό οπότε συνήθως θα χρησιμοποιούμε απλά κεφαλαία γράμματα για να συμβολίσουμε μία κατανομή, καλλιγραφικά γράμματα για τις ακολουθίες κατανομών, και γοτθικά γράμματα για να συμβολίσουμε μία οικογένεια ακολουθιών κατανομών. Ας συμβολίσουμε τώρα με $\mathfrak{P}$ μια οικογένεια ακολουθιών κατανομών, όπου κάθε ακολουθία $\mathcal{P}$ δεικτοδοτείται από το ίδιο σύνολο I. Δηλαδή αν $\mathcal{P}, \mathcal{Q} \in \mathfrak{P}$ τότε $\mathcal{P} = (P_i)_{i\in I}$ και $\mathcal{Q} = (Q_i)_{i\in I}$. Υπενθυμίζουμε τώρα το μοντέλο μάθησης για ακολουθίες κατανομών. Έστω μία ακολουθία $\mathcal{P}$. Ο αλγόριθμος μάθησης μπορεί να τραβήξει δείγματα από *οποιαδήποτε* κατανομή P_i της ακολουθίας $\mathcal{P}$. Συνεπώς η είσοδος του αλγορίθμου θα είναι ένα διάνυσμα από δείγματα

$$X^m = (X_{1,1}, \ldots, X_{1,m_1},\ \ldots,\ X_{k,1}, \ldots, X_{k,m_k})$$

όπου η i-οστή ομάδα δειγμάτων του m_i προέρχεται από την κατανομή P_i.

Για έναν πιο κομψό συμβολισμό θα ορίσουμε τον πολυδείκτη $m = (m_1, \ldots, m_k)$. Όλα τα δείγματα είναι ανεξάρτητα και κατά συνέπεια το δείγμα X^m ακολουθεί την $|m|$-fold κατανομή γινόμενο $P^m = P_1^{m_1} \times P_2^{m_2} \times \ldots \times P_k^{m_k}$. Περνάμε τώρα στον συμβολισμό για την εκτίμηση συναρτήσεων ακολουθιών. Έστω $\theta : \mathfrak{P} \to \times$ μία συνάρτηση ακολουθιών του $\mathfrak{P}$ που θέλουμε να εκτιμήσουμε, έστω $\hat{\theta} : \mathcal{X}^m \to \Theta$ μία εκτιμήτρια του θ, και $\rho : \Theta \times \Theta \to \mathbb{R}_+$ μία ημιμετρική στον χώρο Θ. Ένα παράδειγμα εκτίμησης σε αυτό το μοντέλο είναι η περίπτωση που θέλουμε να εκτιμήσουμε την μέση τιμή όλων των κατανομών μίας ακολουθίας. Σε αυτή τη περίπτωση έχουμε $\theta(\mathcal{P}) = (\mathbb{E}[P_i])_{i\in I}$ και ο χώρος των παραμέτρων, Θ, είναι οι ακολουθίες πραγματικών αριθμών. Για να μετρήσουμε την απόσταση ανάμεσα σε 2 μέσες τιμές μπορούμε να χρησιμοποιήσουμε την απόλυτη διαφορά τους και για να μετρήσουμε την απόσταση ανάμεσα σε δύο ακολουθίες από μέσες τιμές μία εύλογη μετρική είναι η μέγιστη διαφορά των αντίστοιχων στοιχείων των δύο ακολουθιών, δηλαδή $\rho(a, b) = \sup_{i\in I} |a_i - b_i|$ όπου υπενθυμίζουμε ότι τα a, b είναι ακολουθίες πραγματικών αριθμών (των αναμενόμενων τιμών). Γενικότερα, αν υποθέσουμε ότι θέλουμε να εκτιμήσουμε μία συγκεκριμένη συνάρτηση των στοιχείων μίας ακολουθίας κατανομών τότε παρατηρούμε ότι το supremum των επιμέρους αποστάσεων των συναρτήσεων των αντίστοιχων κατανομών αποτελεί μία φυσική γενίκευση της απόστασης στην περίπτωση των ακολουθιών. Ένα άλλο παράδειγμα είναι η περίπτωση που θέλουμε να εκτιμήσουμε την πυκνότητα της κάθε P_i μίας ακολουθίας $\mathcal{P}$. Η ημιμετρική μας στον χώρο των συναρτήσεων πυκνότητας πιθανότητας θα μπορούσε να η απόσταση total variation και κατά συνέπεια στον χώρο των ακολουθιών των συναρτήσεων πυκνότητας η μετρική μας θα είναι η $\sup_{i\in I} d_{\text{tv}}(P_i, Q_i)$. Δεδομένης μίας μετρικής για πυκνότητες (total variation, Hellinger) d θα χρησιμοποιούμε το ίδιο σύμβολο d για την αντίστοιχη μετρική ανάμεσα σε ακολουθίες. Για παράδειγμα $d_{\text{tv}}(\mathcal{P}, \mathcal{Q}) = \sup_{i\in I} d_{\text{tv}}(P_i, Q_i)$.

Ορισμός 2. Χρησιμοποιώντας τον παραπάνω συμβολισμό ορίζουμε το minimax risk να είναι

$$\mathfrak{M}_N\left(\theta(\mathfrak{P}), \rho\right) \coloneqq \inf_{\hat{\theta}} \sup_{\mathcal{P}\in\mathfrak{P}} \inf_{|m|=N} \mathbb{E}_{P^m}\left[\rho\left(\hat{\theta}(X^m),\ \theta(\mathcal{P})\right)\right]. \tag{3.1}$$

Στον Ορισμό 2 το infimum πάνω σε όλους τους πολυδείκτες m τέτοιους ώστε $|m| = N$ αντιστοιχεί στην διαδικασία επιλογής των βέλτιστων κατανομών της ακολουθίας από τις οποίες θα τραβήξουμε δείγματα. Αυτό γίνεται γιατί όπως θα δούμε υπάρχουν περιπτώσεις όπου τα δείγματα από κάποιες συγκεκριμένες κατανομές μιας ακολουθίας μπορούν να μας βοηθήσουν στο να εκτιμήσουμε τις παραμέτρους των υπόλοιπων κατανομών χωρίς να χρειαστεί να τραβήξουμε δείγματα από αυτές. Σημειώνουμε ότι σε περίπτωση που οι ακολουθίες των κατανομών αποτελούνται από μία μοναδική κατανομή το minimax risk του Ορισμού 2 γίνεται

$$\mathfrak{M}_N\left(\theta(\mathfrak{P}), \rho\right) \coloneqq \inf_{\hat{\theta}} \sup_{P\in\mathfrak{P}} \mathbb{E}_{P^m}\left[\rho\left(\hat{\theta}(X^m),\ \theta(P)\right)\right] \tag{3.2}$$

και επομένως συμπίπτει με τον κλασσικό ορισμό του minimax risk της στατιστικής. Η διαφορά των δύο ορισμών είναι ότι στην κλασική περίπτωση όλα τα δείγματα προέρχονται αναγκαστικά από μία κατανομή ενώ στην περίπτωση των ακολουθιών από περισσότερες. Ο Ορισμός 2 φανερώνει την δομή που έχουν οι αλγόριθμοι επίλυσης των προβλημάτων εκτίμησης συναρτήσεων ακολουθιών κατανομών. Έστω ο βέλτιστος αλγόριθμος $\hat{\theta}$. Αφού επιλεχθεί μία "δύσκολη" για τον βέλτιστο αλγόριθμο ακολουθία κατανομών $\mathcal{P} \in \mathfrak{P}$ τότε αυτός μπορεί να διαλέξει πόσα δείγματα θα τραβήξει από την κάθε κατανομή της ακολουθίας με τον μόνο περιορισμό το συνολικό τους πλήθος να είναι N. Αυτή η ιδέα θα γίνει ξεκάθαρη αργότερα στον Αλγόριθμο 1. Αξίζει τον κόπο όμως να σκιαγραφήσουμε τώρα την βασική ιδέα. Ας συμβολίσουμε με $B(n,p)$ μία διωνυμική κατανομή και ας υποθέσουμε ότι η οικογένεια $\mathfrak{P}$ είναι οι ακολουθίες $(B(n,p^i)_{i\in\mathbb{N}})$. Θα αναφερόμαστε στην κατανομή $B(n,p^i)$ ως την i-οστή δύναμη της $B(n,p)$. Ένα ενδιαφέρον πρόβλημα είναι η προσέγγιση των πιθανοτήτων επιτυχίας όλων των δυνάμεων μιας διωνυμικής κατανομής, δηλαδή $\theta(B(n,p^i))_{i\in\mathbb{N}} = (p^i)_{i\in\mathbb{N}}$. Μια εύλογη μετρική για αυτό το πρόβλημα εκτίμησης είναι το supremum της απολύτων διαφορών δηλαδή $d(\theta(B(n,p^i)_{i\in\mathbb{N}}, \theta(B(n,q^i)_{i\in N})_{i\in\mathbb{N}}) = \sup_{i\in\mathbb{N}} |p^i - q^i|$. Σε αυτή τη περίπτωση βλέπουμε καθαρά ότι μπορούμε να εκτιμήσουμε περισσότερες από μία τιμές της ζητούμενης ακολουθίας χρησιμοποιώντας δείγματα από μία κατανομή. Αν δηλαδή τραβήξουμε N δείγματα από την πρώτη δύναμη $B(n,p)$ τότε μπορούμε να εκτιμήσουμε το p με το $\hat{p} = \sum_{i=1}^{N} X_i/(Nn)$. Χρησιμοποιώντας μια τέτοια προσέγγιση μπορούμε να εκτιμήσουμε το p^2 με το $\hat{p}^2$ χωρίς να κάνουμε μεγάλο λάθος. Δεν ισχύει όμως το ίδιο απαραίτητα για μεγαλύτερες δυνάμεις καθώς το σφάλμα μπορεί να μεγαλώνει απαγορευτικά. Στο Ενότητα 4.2 θα παρουσιάσουμε έναν βέλτιστο αλγόριθμο για αυτό το πρόβλημα ο οποίος τραβάει δείγματα *μόνο* από δύο δυνάμεις της ακολουθίας και είναι προσαρμοστικός με την έννοια ότι αποφασίζει για το ποια θα είναι η δεύτερη κατανομή από την οποία θα τραβήξει δείγματα αφού πρώτα τραβήξει κάποια δείγματα από την πρώτη δύναμη. Αυτήν ακριβώς την ελευθερία των αλγορίθμων να είναι προσαρμοστικοί ενσωματώνουμε στον ορισμό του κλασικού minimax risk χρησιμοποιώντας το εσωτερικό infimum πάνω στα δυνατά διανύσματα από N δείγματα. Θα δούμε επίσης ότι στην περίπτωση όπου το p είναι πολύ κοντά στο 1 δηλαδή $p \approx 1 - 1/n^d$ (βλέπε Υποενότητα 4.2.2) η αναζήτηση των σωστών κατανομών είναι αρκετά δυσκολότερη και χρειαζόμαστε $O(\log d \log\log d/\varepsilon^2)$ δείγματα δοκιμάζοντας $\log d$ κατανομές κάνοντας δυαδική αναζήτηση για να βρούμε την καλύτερη δυνατή κατανομή για να εκτιμήσουμε τις δυνάμεις του p.

Οι αποδείξεις που θα δούμε σε αυτό το κεφάλαιο είναι ουσιαστικά γενικεύσεις των γνωστών αποδείξεων για κάτω φράγματα του κλασικού minimax risk στην περίπτωση της εκτίμησης συναρτήσεων ακολουθιών κατανομών. Οι βασικές αποδείξεις για την περίπτωση του κλασικού minimax risk μπορούν να βρεθούν στα [33], [30], και επίσης στις πολύ καλές σημειώσεις του John Duchi [17].

3.1 Αναγωγή της Εκτίμησης σε Έλεγχο

Σε αυτήν την ενότητα θα δείξουμε πως μπορούμε να πάρουμε κάτω φράγματα της δειγματικής πολυπλοκότητας για προβλήματα εκτίμησης ανάγοντάς τα σε προβλήματα ελέγχου.

Ένα βασικό και ενδιαφέρον στατιστικό πρόβλημα είναι το επόμενο. Έστω μία οικογένεια κατανομών από την οποία επιλέγεται στην τύχη μια από αυτές. Εμείς μπορούμε να τραβήξουμε δείγματα από την άγνωστη κατανομή και αφού τραβήξουμε αρκετά πρέπει να "μαντέψουμε" από ποια κατανομή της οικογένειας μας ήρθαν

αυτά τα δείγματα, να βρούμε δηλαδή σε ποια κατανομή της οικογένειας αντιστοιχεί η άγνωστη και τυχαία επιλεγμένη κατανομή.

Έστω $\mathcal{V}$ ένα πεπερασμένο σύνολο δεικτών και έστω $\mathfrak{F}_\mathcal{V} \subseteq \mathfrak{P}$ ένα σύνολο $|\mathcal{V}|$ ακολουθιών κατανομών δεικτοδοτούμενο από το σύνολο $\mathcal{V}$. Έστω V μία τυχαία μεταβλητή η οποία αναπαριστά την ομοιόμορφα τυχαία επιλογή μιας ακολουθίας του χώρου $\mathfrak{F}_\mathcal{V}$. Δεσμεύοντας πάνω στην επιλογή $V = v$, το δείγμα X^m προέρχεται από την κατανομή γινόμενο P_v^m. Θα συμβολίζουμε με ν^m το μέτρο της από κοινού κατανομής των V, X^m. Έστω $\Psi : \mathcal{X}^m \to \mathcal{V}$ μία συνάρτηση ελέγχου (testing function), η οποία τραβά δείγματα από την άγνωστη ακολουθία $\mathcal{P}_V$ και επιστρέφει έναν δείκτη $u \in \mathcal{V}$ ο οποίος αντιστοιχεί στην ακολουθία των κατανομών από την οποία η testing function *ισχυρίζεται* ότι προέρχεται το δείγμα.

Θα δείξουμε τώρα την κλασική αναγωγή από εκτίμηση σε έλεγχο χρησιμοποιώντας τον ορισμό του minimax risk για ακολουθίες κατανομών.

Πρόταση 15. Έστω $\mathfrak{F}_\mathcal{V} \subseteq \mathfrak{P}$ μία οικογένεια ακολουθιών κατανομών δεικτοδοτούμενη από τα $v \in \mathcal{V}$ τέτοια ώστε $\rho\left(\theta(\mathcal{P}_v, \mathcal{P}_u)\right) \geq 2\delta$ για κάθε $\mathcal{P}_v,\ \mathcal{P}_u \in \mathfrak{F}_\mathcal{V}$, όπου, $v \neq u \in \mathcal{V}$ και $\delta > 0$. Τό minimax risk του Ορισμού 2 έχει κάτω φράγμα

$$\mathfrak{M}_N\left(\theta(\mathfrak{P}), \rho\right) \geq \delta \inf_{m=|N|} \inf_{\Psi} \nu^m \left(\Psi(X^m) \neq V\right).$$

Απόδειξη. Θα χρησιμοποιήσουμε τον ίδιο συμβολισμό με αυτόν του Ορισμού 2 Έστω ένας αλγόριθμος εκτίμησης $\hat{\theta}$. Για να απλοποιήσουμε τον συμβολισμό θα συμβολίζουμε με

θ το $\theta(\mathcal{P})$ όταν η ακολουθία κατανομών $\mathcal{P}$ είναι εμφανής από το κείμενο. Θα γράφουμε επίσης θ_v αντί για $\theta(\mathcal{P}_v)$. Από την ανισότητα του Markov 1 έχουμε

$$\mathbb{E}_{P_v^m}\left[\rho(\hat{\theta},\, \theta)\right] \geq \delta\, P_v^m\left(\rho(\hat{\theta},\, \theta) \geq \delta\right) = \delta\, \nu^m\left(\rho(\hat{\theta},\, \theta) \geq \delta | V = v\right) \tag{3.3}$$

Συνεχίζουμε τώρα ορίζοντας της testing function $\Psi(X^m) \coloneqq \operatorname{argmin}_{v\in\mathcal{V}}\{\rho(\hat{\theta}, \theta_v)\}$. Χρησιμοποιώντας το γεγονός ότι $\rho(\theta_v, \theta_u) \geq 2\delta$ για κάθε $v \neq u \in \mathcal{V}$ έχουμε ότι $\rho(\hat{\theta}, \theta_v) \leq \delta \Leftrightarrow \Psi(\hat{\theta}) = v$. Τώρα μπορούμε να πάρουμε ένα κάτω φράγμα στο minimax risk ως εξής

$$\begin{aligned}
\mathfrak{M}_N\left(\theta(\mathfrak{P}), \rho\right) &= \inf_{\hat{\theta}} \sup_{\mathcal{P}\in\mathfrak{P}} \inf_{|m|=N} \mathbb{E}_{P^m}\left[\rho\left(\hat{\theta}(X^m),\ \theta(\mathcal{P})\right)\right] \\
&\geq \inf_{\hat{\theta}} \sum_{v\in\mathcal{V}} \left(\frac{1}{|\mathcal{V}|} \inf_{|m|=N} \mathbb{E}_{P_v^m}\left[\rho(\hat{\theta},\, \theta_v)\right]\right) \\
&\geq \delta \inf_{\hat{\theta}} \sum_{v\in\mathcal{V}} \left(\frac{1}{|\mathcal{V}|} \inf_{|m|=N} \nu^m\left(\rho(\hat{\theta},\, \theta_v) \geq \delta \mid V = v\right)\right) \\
&= \delta \inf_{|m|=N} \inf_{\hat{\theta}} \sum_{v\in\mathcal{V}} \left(\frac{1}{|\mathcal{V}|} \nu^m\left(\rho(\hat{\theta},\, \theta_v) \geq \delta \mid V = v\right)\right) \\
&= \delta \inf_{|m|=N} \inf_{\Psi} \nu^m\left(\Psi(X^m) \neq V\right),
\end{aligned}$$

όπου για την πρώτη ανισότητα χρησιμοποιούμε το γεγονός ότι το supremum ενός συνόλου είναι μεγαλύτερο από την μέση τιμή ενός υποσυνόλου του αρχικού συνόλου, για την δεύτερη ανισότητα χρησιμοποιούμε την ανισότητα του Markov1 και για την δεύτερη ισότητα το γεγονός ότι για δύο μη κενά σύνολα $A,\ B$ ισχύει $\inf(A + B) = \inf(A) + \inf(B)$ for any nonempty sets $A,\ B$. Η τελευταία ισότητα προκύπτει από το θεώρημα του Bayes. □

3.2 Le Cam

3.2.1 Η Ανισότητα του Le Cam

Θα θεωρήσουμε τώρα ένα παρόμοιο testing πρόβλημα με αυτό της ενότητας 3. Έστω δύο κατανομές P και Q. Διαλέγεται κρυφά μία από αυτές και εμείς τραβάμε δείγματα από αυτήν. Ο στόχος μας είναι να καταλάβουμε

αν η κατανομή που διαλέχτηκε είναι η P ή η Q. Το πρόβλημα αυτό είναι ένα πρόβλημα δυαδικού ελέγχου υποθέσεως (binary hypothesis testing) και σκοπός μας σε αυτήν την ενότητα είναι να παρουσιάσουμε ένα βασικό κάτω φράγμα που χαρακτηρίζει την δυσκολία αυτού του προβλήματος. Σε αυτό το σημείο ας πάρουμε λίγο χρόνο για να σκέφτουμε πότε ένα τέτοιο πρόβλημα είναι δύσκολο. Ας θεωρήσουμε ότι βλέπουμε μόνο ένα δείγμα από την άγνωστη κατανομή αφού η ιδέα και η απόδειξη παραμένουν ίδιες και για την περίπτωση που τραβάμε περισσότερα δείγματα. Διαισθητικά καταλαβαίνουμε ότι θα δυσκολευτούμε να ξεχωρίσουμε τις δύο κατανομές αν αυτές "μοιάζουν" πολύ. Η δυσκολία του προβλήματος λοιπόν εξαρτάται από την *στατιστική απόσταση* των P_1, Q_1. Το επόμενο αποτέλεσμα του Le Cam [6] ποσοτικοποιεί ακριβώς αυτή την διαίσθηση

Πρόταση 16 (Ανισότητα Le Cam)**.** Έστω $\mathcal{X}$ ένα σύνολο. Για κάθε κατανομές P_1, P_2 στο $\mathcal{X}$ έχουμε

$$\inf_{\Psi} \{P_1(\Psi(X) \neq 1) + P_2(\Psi(X) \neq 2)\} = 1 - d_{\text{tv}}(P_1, P_2)$$

όπου στο infimum θεωρούμε όλες τις συναρτήσεις ελέγχου $\Psi : \mathcal{X} \to \{1, 2\}$.

Απόδειξη. Κάθε συνάρτηση ελέγχου $\Psi : \mathcal{X} \to \{1, 2\}$ ταυτίζεται με μία διαμέριση του χώρου $\mathcal{X}$ σε δύο περιοχές $, ^c = \mathcal{X}$ A. Αν το δείγμα X που τραβάμε από την άγνωστη κατανομή πέσει στην περιοχή A τότε χωρίς βλάβη της γενικότητας υποθέτουμε ότι η testing συνάρτηση Ψ απαντάει 1, δηλαδή προτείνει ως υποψήφια κατανομή την P_1 αλλιώς αν το δείγμα πέσει στην A^c η Ψ απαντάει 2. Η πιθανότητα $P_1(\Psi \neq 1)$ είναι η πιθανότητα η πραγματική κατανομή να είναι η P_1 και η Ψ να απαντήσει 2, δηλαδή είναι η πιθανότητα του λάθους αν το δείγμα προέρχεται από την P_1. Αντίστοιχα η $P_2(\Psi \neq 2)$ είναι η πιθανότητα λάθους αν δείγμα προέρχεται από την P_2. Γράφουμε $P_1(\Psi \neq 1) = P_1(A^c)$ και $P_2(\Psi \neq 2) = P_2(A)$. Έτσι έχουμε

$$\begin{aligned}
\inf_{\Psi} \{P_1(\Psi(X) \neq 1) + P_2(\Psi(X) \neq 2)\} &= \inf_{A \subseteq \mathcal{X}} \{P_1(A^c) + P_2(A)\} \\
&= \inf_{A \subseteq \mathcal{X}} \{1 - P_1(A) + P_2(A)\} \\
&= 1 - \sup_{A \subseteq \mathcal{X}} \{P_2(A) - P_1(A)\} \\
&= 1 - d_{\text{tv}}(P_1, P_2),
\end{aligned}$$

□

Σε αυτό το σημείο έχουμε αποδείξει το βασικό εργαλείο για να κατασκευάζουμε κάτω φράγματα για testing προβλήματα. Ας δούμε όμως κάτι ακόμα, σχετικό με το binary hypothesis testing πρόβλημα. Από την απόδειξη είδαμε ότι το να κατασκευάσουμε έναν "αλγόριθμο ελέγχου" Ψ σημαίνει να βρούμε κατάλληλη διαμέριση του χώρου $\mathcal{X}$ σε περιοχές αποδοχής και απόρριψης A, A^c. Θα κατασκευάσουμε τώρα έναν απλό και βέλτιστο τρόπο για να διαλέγουμε ανάμεσα στις P_1, P_2. Ας συμβολίσουμε με p_1, p_2 τις αντίστοιχες συναρτήσεις πυκνότητας πιθανότητας των P_1, P_2. Η ιδέα για την επιλογή είναι ίδια με αυτήν της εκτιμήτριας μέγιστης πιθανοφάνειας (Maximum Likelihood Estimator, MLE). Ουσιαστικά διαλέγουμε την κατανομή η οποία μεγιστοποιεί την πιθανότητα της παρατήρησης του δείγματος X που έχουμε ήδη τραβήξει. Ουσιαστικά δηλαδή ελέγχουμε ποια από τις συναρτήσεις πυκνότητας έχει μεγαλύτερη τιμή για την τιμή x του δείγματος X. Η επόμενη συνάρτηση ελέγχου Ψ καθώς και η απόδειξη ότι είναι η βέλτιστη αποτελούν το διάσημο Λήμμα Neyman-Pearson.

$$\Psi(X) = \begin{cases} 1 & \text{if } p_1(x) \geq p_2(x) \\ 2 & \text{if } p_1(x) < p_2(x) \end{cases} \tag{3.4}$$

Σημειώνουμε σε αυτό το σημείο ότι η άγνωστη συνάρτηση επιλέγεται με πιθανότητα $1/2$ να είναι η P_1 και με $1/2$ να είναι η P_2. Έτσι λοιπόν διαλέγουμε περιοχή αποδοχής $A = \{x \in \mathcal{X} : p_1(x) \geq p_2(x)\}$. Ας υπολογίσουμε την πιθανότητα να κάνει λάθος η Ψ. Από τον τύπο ολικής πιθανότητας του Bayes έχουμε ότι η πιθανότητα λάθος απάντησης είναι

$$\frac{1}{2}P_1(A^c) + \frac{1}{2}P_2(A) = \frac{1}{2}\left(\int_{A^c} \mathrm{d}P_1 + \int_A \mathrm{d}P_2\right)$$

$$
\begin{aligned}
&= \frac{1}{2}\left(1 + \int_A (p_2 - p_1)\right) \\
&= \frac{1}{2}\left(1 - \int_A (p_1 - p_2)\right) \\
&= \frac{1}{2}\left(1 - d_{\mathrm{tv}}\left(P_1, P_2\right)\right)
\end{aligned}
$$

Για την τελευταία ισότητα χρησιμοποιήσουμε την ταυτότητα του Scheffe

Πρόταση 17 (Ταυτότητα Scheffe)**.** Έστω δύο κατανομές P_1 και P_2 με πυκνότητες p_1, p_2 αντίστοιχα, ορισμένες στον $\mathbb{R}^d$. Έστω $\mathcal{B}$ η οικογένεια των Borel συνόλων του $\mathbb{R}^d$. Τότε

$$
d_{\mathrm{tv}}\left(P_1, P_2\right) = \sup_{B \in \mathcal{B}} \left| \int_B p_1 - \int_B p_2 \right| = \int_{p_1 \geq p_2} (p_1 - p_2) = \frac{1}{2} \int |p_1 - p_2|
$$

Απόδειξη. Η απόδειξη βασίζεται στο γεγονός ότι έχουμε να κάνουμε με πυκνότητες και κατά συνέπεια $\int p_1 = \int p_2 = 1$. Παρατηρούμε κατ' αρχάς ότι αφού οι p_1, p_2 είναι μετρήσιμες το σύνολο $\{x \in \mathbb{R}^d : p_1(x) \geq p_2(x)\}$ είναι Borel και ανήκει στην $\mathcal{B}$. Συνεπώς $d_{\mathrm{tv}}\left(P_1, P_2\right) = \sup_{B \in \mathcal{B}} \left| \int_B p_1 - \int_B p_2 \right| = \int_{p_1 \geq p_2} (p_1 - p_2)$ Παρατηρούμε ότι

$$
\begin{aligned}
\int_{p_1 \geq p_2} (p_1 - p_2) &= \int_{p_1 < p_2} (p_2 - p_1) \\
\int_{p_1 \geq p_2} p_1 + \int_{p_1 < p_2} p_1 &= \int_{p_1 < p_2} p_2 + \int_{p_1 \geq p_1} p_2 \\
\int p_1 &= \int p_2 \\
1 &= 1
\end{aligned}
$$

Άρα

$$
\int |p_1 - p_2| = \int_{p_1 \geq p_2} (p_1 - p_2) + \int_{p_1 < p_2} (p_2 - p_1) = 2 \int_{p_1 \geq p_2} (p_1 - p_2)
$$

□

Δείξαμε λοιπόν ότι μπορούμε να "πιάσουμε" το κάτω φράγμα της ανισότητας του Le Cam. Για μία εκτενέστερη και γενικότερη ανάλυση του Λήμματος Neyman-Pearson παραπέμπουμε στο Κεφάλαιο 5 του [27].

3.2.2 Η Μέθοδος του Le Cam

Θα δείξουμε τώρα πως μπορούμε να χρησιμοποιήσουμε την ανισότητα της υποενότητας 3.2.1 και την αναγωγή της εκτίμησης σε έλεγχο για να πάρουμε μία χρήσιμη μέθοδο κατασκευής κάτω φραγμάτων για προβλήματα εκτίμησης συναρτήσεων ακολουθιών κατανομών.

Λήμμα 3. Έστω μία οικογένεια ακολουθιών κατανομών $\mathfrak{P}$. Έστω $\mathcal{P}, \mathcal{Q} \in \mathfrak{P}$ και $\delta > 0$ τέτοια ώστε $\rho\left(\theta(\mathcal{P}), \theta(\mathcal{Q})\right) \geq 2\delta$ τότε μετά από N παρατηρήσεις(δείγματα) το minimax risk έχει κάτω φράγμα

$$\mathfrak{M}_N\left(\theta(\mathfrak{P}), \rho\right) \geq \frac{\delta}{2}\left(1 - \sup_{|m|=N} d_{\mathrm{tv}}\left(P^m, Q^m\right)\right).$$

Ισχύει επίσης,

$$\mathfrak{M}_N\left(\theta(\mathfrak{P}), \rho\right) \geq \frac{\delta}{2}(1 - \sqrt{2}\sqrt{1 - \left(1 - d_{\mathrm{tv}}\left(\mathcal{P}, \mathcal{Q}\right)\right)^N}).$$

Αν επίσης $d_{\mathrm{tv}}\left(\mathcal{P}, \mathcal{Q}\right) \leq 1/(16N)$ τότε

$$\mathfrak{M}_N\left(\theta(\mathfrak{P}), \rho\right) \geq \frac{\delta}{4}$$

Απόδειξη. Επανερχόμενοι στον συμβολισμό που δώσαμε στην αρχή της ενότητας θυμίζουμε ότι αφού έχουμε μόνο δύο υποθέσεις να ελέγξουμε η τυχαία μεταβλητή V αντιστοιχεί στην ομοιόμορφα τυχαία επιλογή ανάμεσα στις P_1, P_2. Ορίζουμε τώρα το μέτρο πιθανότητας μ της από κοινού κατανομής του δείγματος X^m και της επιλογής V. Η πιθανότητα να απαντήσει λάθος η συνάρτηση ελέγχου Ψ είναι $\mu(\Psi(X^m) \neq V) = \frac{1}{2}P^m(\Psi(X^m) \neq 1) + \frac{1}{2}Q^m(\Psi(X) \neq 2)$. Από την ανισότητα του Le Cam έχουμε ότι

$$\inf_{\Psi}\{P^m(\Psi(X^m) \neq 1) + Q^m(\Psi(X^m) \neq 2)\} = 1 - d_{\mathrm{tv}}\left(P^m, Q^m\right) \tag{3.5}$$

Χρησιμοποιώντας την (3.5) καθώς και την Πρόταση 15 παίρνουμε

$$\mathfrak{M}_N\left(\theta(\mathfrak{P}), \rho\right) \geq \frac{\delta}{2}\inf_{|m|=N}\left(1 - d_{\mathrm{tv}}\left(P^m, Q^m\right)\right) = \frac{\delta}{2}\left(1 - \sup_{|m|=N} d_{\mathrm{tv}}\left(P^m, Q^m\right)\right)$$

Παρατηρήστε ότι

$$\begin{aligned}
\sup_{|m|=N} d_{\mathrm{tv}}\left(P^m, Q^m\right) \leq & \sqrt{2}\sup_{|m|=N}\sqrt{1 - \prod_{i=1}^{N}\left(1 - d_{\mathrm{hel}}\left(P_i, Q_i\right)^2\right)} \\
\leq & \sqrt{2}\sqrt{1 - \left(1 - \sup_{i \in I} d_{\mathrm{hel}}\left(P_i, Q_i\right)^2\right)^N} \\
\leq & \sqrt{2}\sqrt{1 - \left(1 - \sup_{i \in I} d_{\mathrm{tv}}\left(P_i, Q_i\right)\right)^N} \\
= & \sqrt{2}\sqrt{1 - \left(1 - d_{\mathrm{tv}}\left(\mathcal{P}, \mathcal{Q}\right)\right)^N}
\end{aligned} \tag{3.6}$$

όπου χρησιμοποιήσαμε την ανισότητα $d_{\mathrm{hel}}\left(P, Q\right)^2 \leq d_{\mathrm{tv}}\left(P, Q\right) \leq \sqrt{2}\, d_{\mathrm{hel}}\left(P, Q\right)$ και την Πρόταση **??**. Για να πάρουμε την τρίτη χρήσιμη εκδοχή απαιτούμε

$$\begin{aligned}
1 - \sqrt{2}\sqrt{1 - \left(1 - d_{\mathrm{tv}}\left(\mathcal{P}, \mathcal{Q}\right)\right)^N} &\geq 1/2 \\
\sqrt{1 - \left(1 - d_{\mathrm{tv}}\left(\mathcal{P}, \mathcal{Q}\right)\right)^N} &\leq 1/(\sqrt{2}2) \\
1 - \left(1 - d_{\mathrm{tv}}\left(\mathcal{P}, \mathcal{Q}\right)\right)^N &\leq 1/8 \\
\left(1 - d_{\mathrm{tv}}\left(\mathcal{P}, \mathcal{Q}\right)\right)^N &\geq 7/8
\end{aligned}$$

Χρησιμοποιώντας το γεγονός ότι $d_{\mathrm{tv}}\left(\mathcal{P}, \mathcal{Q}\right) \leq 1/(16N)$ έχουμε

$$\left(1 - d_{\mathrm{tv}}\left(\mathcal{P}, \mathcal{Q}\right)\right)^N \geq \left(1 - \frac{1}{16N}\right)^N \geq \mathrm{e}^{-1/8} \geq 7/8,$$

όπου χρησιμοποιήσαμε την ανισότητα $\mathrm{e}^{-x} \leq 1 - x/2$ για κάθε $x \in [0, 3/2]$. □

Η γενίκευση του Λήμματος του Le Cam για ακολουθίες κατανομών είναι φυσική και διαισθητικά εμφανής καθώς για να ξεχωρίσουμε δύο *ακολουθίες* κατανομών $\mathcal{P}$, $\mathcal{Q}$ αρκεί να βρούμε έναν δείκτη $i \in$ έτσι ώστε τα αντίστοιχα στοιχεία των δύο ακολουθιών να διαφέρουν αρκετά, δηλαδή η $d_{\text{tv}}(P_i, Q_i)$ να είναι μεγάλη. Ο ορισμός του minimax risk για ακολουθίες κατανομών επιτρέπει στον προσαρμοστικό βέλτιστο να διαλέξει τον δείκτη i και να τραβήξει δείγματα από την i-οστή κατανομή της ακολουθίας. Άρα λοιπόν για να έχουμε ένα ισχυρό κάτω φράγμα πρέπει να "προστατευθούμε" απέναντι στην δύναμη του βέλτιστου αλγορίθμου να διαλέγει από που θα τραβήξει δείγματα και κατά συνέπεια να βρούμε ένα ζεύγος ακολουθιών των οποίων όλα τα αντίστοιχα στοιχεία έχουν μικρή στατιστική απόσταση. Πρέπει δηλαδή η $d_{\text{tv}}(P_i, Q_i)$ να είναι μικρή για κάθε $i \in I$ και κατά συνέπεια η $d_{\text{tv}}(\mathcal{P}, \mathcal{Q})$ να είναι μικρή παρόλο που οι παράμετροι τους (τις οποίες θέλουμε να εκτιμήσουμε) είναι μακριά.

Στη συνέχεια θα δείξουμε πως μπορούμε να εφαρμόσουμε την μέθοδο του Le Cam για να πάρουμε ένα βέλτιστο κάτω φράγμα για την εκτίμηση της παραμέτρου p μιας διωνυμικής $B(n,p)$. Διατηρώντας τον γνωστό συμβολισμό έχουμε ότι ο χώρος των κατανομών $\mathfrak{P}$ σε αυτή την περίπτωση είναι ο χώρος των διωνυμικών κατανομών, δηλαδή $\mathfrak{P} = \{B(n,p), :, p \in [0,1]\}$ και η συνάρτηση που θέλουμε να εκτιμήσουμε είναι η $\theta(B(n,p)) = p$. Ξεκινάμε δίνοντας ένα πάνω φράγμα στο minimax rate. Έστω η εκτιμήτρια $\hat{p} = \sum_{i=1}^{N} X_i/(Nn)$ όπου $X_i \sim B(n,p)$. Οι παράμετρος p που θέλουμε να εκτιμήσουμε βρίσκεται στον χώρο $[0,1]$ και η μετρική που θα χρησιμοποιήσουμε είναι η $\rho(q,p) = |q-p|$. Ξεκινάμε δίνοντας ένα πάνω φράγμα για το minimax rate αυτού του προβλήματος χρησιμοποιώντας την εκτιμήτρια $\hat{p}$. Ας συμβολίσουμε για ευκολία $\sigma = \sqrt{\text{Var}[\hat{p}]}$. Από το φράγμα του Chernoff 3 έχουμε ότι

$$\mathbb{P}[|\hat{p} - p| > x] \leq 2\mathrm{e}^{-\frac{x^2}{4\sigma^2}}$$

Φράσσουμε τώρα την μέση τιμή του σφάλματος ως εξής

$$\mathfrak{M}_N \leq \mathbb{E}[|\hat{p} - p|] \leq 2\int_0^1 \mathrm{e}^{-\frac{x^2}{4\sigma^2}}\,\mathrm{d}x = 2\sqrt{\pi}\,\sigma\,\text{erf}\left(\frac{1}{2\sigma}\right) \leq 2\sqrt{\pi}\sigma = 2\sqrt{\pi}\sqrt{\frac{p(1-p)}{Nn}} \leq \sqrt{\frac{\pi}{Nn}}.$$

Περνάμε τώρα στο κάτω φράγμα. Έστω οι διωνυμικές $B(n, 1/2-\delta)$ και $B(n, 1/2+\delta)$. Η απόσταση των παραμέτρων τους είναι ακριβώς 2δ άρα ικανοποιείται η συνθήκη της μεθόδου του Le Cam, οι παράμετροί τους είναι "αρκετά" μακριά. Θα δείξουμε τώρα ότι η απόσταση Kullback-Leibler των δύο κατανομών είναι πολύ μικρή. Έστω $p = 1/2 - \delta$, $q = 1/2 + \delta$, τότε από την Πρόταση 8 έχουμε

$$\begin{aligned} D_{\text{kl}}(B(n,p)\|B(n,q)) &= n\left(p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}\right) = n\left(p\ln\frac{p}{q} + q\ln\frac{q}{p}\right) \\ &= n(q-p)\ln\frac{q}{p} = n2\delta\ln\frac{1/2+\delta}{1/2-\delta} \leq 10n\delta^2 \end{aligned}$$

όπου χρησιμοποιήσαμε την ανισότητα $\ln\frac{1/2+\delta}{1/2-\delta} \leq 5\delta^2$ για κάθε $\delta \in (0, 1/3)$. Χρησιμοποιώντας την ανισότητα του Pinsker έχουμε ότι

$$d_{\text{tv}}\left(P^N, Q^N\right)^2 \leq \frac{1}{2}D_{\text{kl}}\left(P^N\|Q^N\right) = \frac{N}{2}D_{\text{kl}}(P\|Q) \leq 5Nn\delta^2$$

Από το Λήμμα 3 έχουμε

$$\mathfrak{M}_N \geq \frac{\delta}{2}(1 - d_{\text{tv}}\left(P^N, Q^N\right)) \geq \frac{\delta}{2}(1 - 5Nn\delta^2)$$

Επιλέγοντας $\delta = 1/(\sqrt{10Nn})$ έχουμε ότι $\delta < 1/3$ για κάθε

$$\mathfrak{M}_N \geq \frac{1}{4\sqrt{10Nn}}$$

Παρόλο που οι σταθερές του πάνω και του κάτω φράγματος δεν είναι ίσες οι δύο ρυθμοί σύγκλισης του minimax risk ταυτίζονται ασυμπτωτικά και κατά συνέπεια η εκτιμήτρια που δώσαμε για αυτό το πρόβλημα είναι βέλτιστη από πλευράς δειγματικής πολυπλοκότητας. Συγκεντρωτικά λοιπόν για το minimax risk του προβλήματος της εκτίμησης της παραμέτρου επιτυχίας μιας διωνυικής κατανομής χρησιμοποιώντας το μέσο απόλυτο σφάλμα έχουμε

$$\frac{1}{4\sqrt{10Nn}} \leq \mathfrak{M}_N(\theta(\mathfrak{P}), \rho) \leq \sqrt{\frac{\pi}{Nn}}. \tag{3.7}$$

Παρατηρούμε ότι το σφάλμα είναι ανάλογο της ποσότητας $1/\sqrt{N}$. Αυτός ο ρυθμός μείωσης του σφάλματος είναι χαρακτηριστικός των παραμετρικών προβλημάτων. Επίσης μεγαλύτερες τιμές του n βοηθούν στο να μένει το σφάλμα της εκτίμησης μικρό το οποίο είναι διαισθητικά αναμενόμενο αφού αφού η διακύμανση της εκτιμήτριάς μας $\hat{p}$ είναι φθίνουσα συνάρτηση του n.

Ας δοκιμάσουμε τώρα να κατασκευάσουμε ένα κάτω φράγμα για ένα πιο γενικό πρόβλημα. Αυτή τη φορά θέλουμε να υπολογίσουμε τις παραμέτρους από μία PBD ειδικής μορφής. Συγκεκριμένα θεωρούμε την κλάση κατανομών $\mathfrak{F} = \{B(n_1, p_1) + B(n_2, p_2) : n_1, n_2 \in \mathbb{N}, p_1, p_2 \in [0, 1]\}$. Δηλαδή τα αθροίσματα δύο διωνυμικών. Αυτές οι κατανομές αν και ειδική περίπτωση των PBD είναι αρκετά περίπλοκες στην περιγραφή τους και αυτό μπορεί να το δει κανείς προσπαθώντας να κατασκευάσει την συνάρτηση πυκνότητας μιας τέτοιας κατανομής X

$$\mathbb{P}[X = k] = \sum_{i+j=k} \binom{n_1}{i} \binom{n_2}{j} p_1^i (1-p_1)^{n_1-i} p_2^j (1-p_2)^{n_2-j}.$$

Το ερώτημα τώρα είναι εάν το να εκτιμήσουμε τις παραμέτρους p_1, p_2 από μία τέτοια κατανομή είναι δυσκολότερο από το πρόβλημα της εκτίμησης της παραμέτρου μιας διωνυμικής κατανομής. Θα δείξουμε ότι το πρόβλημα γίνεται σαφώς δυσκολότερο με μια αρκετά χειρότερη εξάρτηση από το n. Έστω $X \in \mathfrak{F}$. Στο παραπάνω πλαίσιο έχουμε $\theta(X) = (p_1, p_2)$ όπου $p_1 \leq p_2$, και $\rho((p_1, p_2), (q_1, q_2)) = \max_{i\in\{1,2\}} |p_i - q_i|$ αφού είναι λογικό να θέλουμε οι εκτιμήσεις και για τις δύο παραμέτρους να είναι καλές. Η μεγάλη διαφορά σε σχέση με την περίπτωση των διωνυμικών είναι ότι τώρα μπορούμε να βρούμε ένα ζεύγος κατανομών του οποίου οι παράμετροι θα απέχουν περίπου το ίδιο με την περίπτωση των διωνυμικών αλλά η απόσταση των ίδιων των κατανομών θα είναι πολύ μικρότερη από ότι στην προηγούμενη περίπτωση. Η σημαντική παρατήρηση για την κατασκευή ενός κάτω φράγματος για αυτή την κλάση κατανομών είναι η συμμετρία που έχουν. Υποθέτουμε για απλοποίηση των πράξεων ότι το n είναι άρτιος αριθμός και θεωρούμε το ζεύγος των κατανομών $P = B(n, 1/2)$ και $Q = B(n/2, p) + B(n/2, q)$ όπου $p = 1/2 - \delta$ και $q = 1/2 + \delta$. Παρατηρούμε ότι οι P, Q έχουν ίση μέση τιμή και παραπλήσια διασπορά ενώ οι παράμετροί τους βρίσκονται σε απόσταση 2δ. Το γεγονός ότι καταφέραμε να ταιριάξουμε πλήρως την μέση τιμή δύο κατανομών με αρκετά διαφορετικές παραμέτρους είναι η διαφορά που κάνει αυτό το πρόβλημα αρκετά δυσκολότερο από την περίπτωση των διωνυμικών όπου αυτό δεν ήταν δυνατόν. Περιμένουμε λοιπόν η απόσταση των παραπάνω κατανομών να είναι αρκετά μικρότερη από αυτή των διωνυμικών του προηγούμενου παραδείγματος. Αυτή τη φορά θα χρησιμοποιήσουμε έναν διαφορετικό τρόπο να φράξουμε την απόσταση ολικής κύμανσης των κατανομών γινομένων P^N, Q^N χρησιμοποιώντας την απόσταση ολικής κύμανσης των P, Q. Παρατηρούμε ότι η P είναι διωνυμική ενώ η Q είναι PBD οπότε μπορούμε να χρησιμοποιήσουμε το Θεώρημα 3. Χρησιμοποιώντας τον συμβολισμό του Θεωρήματος 3 έχουμε

$$\gamma_1(1/2) = \sum_{i=1}^{n/2}(1/2 - p) + \sum_{i=n/2}^{n}(1/2 - q) = \frac{n}{2}\delta - \frac{n}{2}\delta = 0$$

$$\gamma_2(1/2) = \sum_{i=1}^{n/2}(1/2 - p)^2 + \sum_{i=n/2}^{n}(1/2 - q)^2 = n\delta^2$$

$$\tau(1/2) = \frac{(\gamma_1(1/2))^2 + 2\gamma_2(1/2)}{2n(1/2)(1/2)} = 4\delta^2.$$

Συνεπώς η απόσταση ολικής κύμανσης των P, Q φράσσεται από πάνω από το Θεώρημα του Roos

$$d_{\mathrm{tv}}(P, Q) \leq \sqrt{\mathrm{e}}\frac{\delta}{(1-2\delta)^2} \leq \frac{64\sqrt{\mathrm{e}}}{49}\delta, \tag{3.8}$$

όπου υποθέτουμε ότι $\delta \leq 1/(16N)$ οπότε $1 - 2\delta \geq 7/8$. Χρησιμοποιώντας την τρίτη εκδοχή του Λήμματος 3 αρκεί να διαλέξουμε το δ έτσι ώστε η $d_{\mathrm{tv}}(P, Q)$ να είναι μικρότερη από $1/(16N)$. Πράγματι, διαλέγοντας $\delta = \frac{49}{1024\sqrt{\mathrm{e}}N}$ στην εξίσωση 3.8 έχουμε ότι $d_{\mathrm{tv}}(P, Q) \leq 1/(16N)$ και τελικά

$$\mathfrak{M}_N(\theta(\mathfrak{F}), \rho) \geq \frac{49}{4096\sqrt{\mathrm{e}}N}$$

Αν και ήδη το παραπάνω κάτω φράγμα δείχνει ότι το πρόβλημα είναι αρκετά δυσκολότερο, μιας και δεν υπάρχει πλέον ο καθοριστικός παράγοντας $\sqrt{n}$ στον παρονομαστή, το κάτω φράγμα αυτό δεν είναι το καλύτερο δυνατό

καθώς έχουμε χάσει την σωστή εξάρτηση από τον αριθμό των δειγμάτων N. Θα θέλαμε να πάρουμε ένα κάτω φράγμα της μορφής $C/\sqrt{N}$ όπου C είναι μια σταθερά. Για να το πετύχουμε αυτό θα χρησιμοποιήσουμε προσεγγίσεις με κανονικές κατανομές για τις PBDs αφού είδαμε ότι αυτές οι προσεγγίσεις είναι πολύ ακριβείς όταν η διασπορά των PBDs είναι μεγάλη. Επειδή θέλουμε η απόσταση ολικής κύμανσης των δύο να είναι συνάρτηση του δ και η προσέγγιση με κανονικές εισάγει σφάλμα $7.6/\sigma$ ελέγχουμε το σφάλμα χρησιμοποιώντας PBDs με μεγαλύτερο μήκος για μικρές τιμές του δ. Έχουμε δηλαδή

$$\begin{aligned} P &= B(n/\delta^4, 1/2) \\ Q &= B\left(\frac{n}{2\delta^4}, 1/2-\delta\right) + B\left(\frac{n}{2\delta^4}, 1/2+\delta\right) \end{aligned}$$

Η διασπορά της P είναι $\sigma_P^2 = n/(4\delta^4)$ ενώ της Q είναι $\sigma_Q^2 = n/(2\delta^4)(1/2-\delta)(1/2+\delta) + n/(2\delta^4)(1/2-\delta)(1/2+\delta) = (n/\delta^4)(1/4-\delta^2)$. Το σφάλμα λοιπόν της κανονικής προσέγγισης της P είναι το πολύ $7.6\delta^2/\sqrt{n} \leq 16\delta^2$ ενω της Q είναι $20\delta^2$ για κάθε $\delta < 1/4$. Από το Λήμμα 1 έχουμε ότι η απόσταση ολικής κύμανσης των P, Q είναι

$$\begin{aligned} d_{\mathrm{tv}}(P,Q) &\leq \frac{\sigma_P^2 - \sigma_Q^2}{\sigma_Q^2} + 7.6\left(\frac{1}{\sigma_P} + \frac{1}{\sigma_Q}\right) \\ &= \frac{\delta^2}{1/4-\delta^2} + 36\delta^2 \\ &\leq 42\delta^2 \end{aligned}$$

όπου χρησιμοποιήσαμε την ανισότητα $\delta^2/(1/4-\delta^2) \leq 6\delta^2$ για κάθε $\delta < 1/4$. Άρα για να πετύχουμε $d_{\mathrm{tv}}(P,Q) \leq 1/(16N)$ αρκεί τώρα να να θέσουμε $\delta = \frac{1}{4\sqrt{42}\sqrt{N}}$. Τελικά έχουμε το ζητούμενο κάτω φράγμα

$$\mathfrak{M}_N(\theta(\mathfrak{F}), \rho) \geq \frac{1}{16\sqrt{42}\sqrt{N}} \tag{3.9}$$

Το παραπάνω κάτω φράγμα αποτελεί κάτω φράγμα και για το γενικότερο πρόβλημα της προσέγγισης των παραμέτρων $\boldsymbol{p}$ μιας γενικής PBD αλλά σε αυτή την περίπτωση θα δούμε στο Ενότητα 4.3 ότι μπορούμε να κατασκευάσουμε ένα πολύ ισχυρότερο κάτω φράγμα.

3.3 Fano

3.3.1 Η Ανισότητα του Fano

Στρέφουμε τώρα την προσοχή στο αρχικό γενικότερο πρόβλημα του ελέγχου πολλών υποθέσεων. Σε αντίθεση δηλαδή με την ανισότητα του Le Cam θέλουμε τώρα ένα αντίστοιχο κάτω φράγμα για την πιθανότητα ο αλγόριθμος ελέγχου Psi να αποτύχει να αναγνωρίσει την πραγματική κατανομή από όπου προήλθε το δείγμα που παρατήρησε. Είναι χρήσιμο να σκεφτόμαστε την παραπάνω διαδικασία επιλογής μίας κατανομής από την οικογένεια, την δειγματοληψία από αυτήν και τέλος την "μαντεψιά" $\Psi(X) = \hat{V}$ σαν μια αλυσίδα Markov $V \to X \to \hat{V}$ αφού $\mathbb{P}\left[\hat{V} = u | X = x, V = v\right] = \mathbb{P}\left[\hat{V} = u | X = x\right]$ αφού η η συνάρτηση $\Psi(X)$ εξαρτάται άμεσα μόνο από το δείγμα X και μόνο έμμεσα από την επιλογή V. Ξεχνώντας προς στιγμήν το πρόβλημα της επιλογής μίας κατανομής από ένα σύνολο κατανομών ας υποθέσουμε ότι έχουμε μία τυχαία μεταβλητή με τιμές στον χώρο $\mathcal{X}$ και την αλυσίδα Markov $X \to Y \to Z$. Θα δείξουμε την επόμενη βασική ανισότητα

Πρόταση 18 (Ανισότητα Fano)**.** Έστω μια τυχαία μεταβλητή X με τιμές στο πεπερασμένο σύνολο $\mathcal{X}$, μία τυχαία μεταβλητή Y με τιμές στο $\mathcal{Y}$ και μία τυχαία μεταβλητή $Z : \mathcal{Y} \to \mathcal{X}$. Για κάθε αλυσίδα Markov $X \to Y \to Z$ έχουμε

$$h(\mathbb{P}[X \neq Z]) + \mathbb{P}[X \neq Z](\ln(|\mathcal{X}| - 1) \geq H($$

όπου $h(x) = -x\ln x - (1-x)\ln(1-x)$.

Απόδειξη. Θα χρησιμοποιήσουμε την δείκτρια τυχαία μεταβλητή

$$E = \begin{cases} 1 & \text{if } Z \neq X \\ 0 & \text{else} \end{cases}$$

Από τον κανόνα αλυσίδας για την εντροπία έχουμε Έχουμε $H(X, E|Z) = H(|Z) + (E|Z, X)$. Αν γνωρίζουμε όμως τις μεταβλητές , η εντροπία της E είναι εξ' ορισμού 0. Άρα $H(X, E|Z) = H(X|Z)$. Αντίστοιχα γράφουμε

$$H(X, E|Z) = H(E|Z) + H(X|Z, E) = H(E|Z) + \mathbb{P}[E = 1] H(X|E = 1, Z)\mathbb{P}[E = 0] H(X|E = 0, Z),$$

όμως $H(X|E = 0, Z)$ είναι μηδέν αφού γνωρίζοντας ότι $X = Z$ και την τιμή της Z γνωρίζουμε την τιμή της X, η έχει μηδενική εντροπία. Έχουμε λοιπόν

$$H(X|Z) = H(E|Z) + \mathbb{P}[E = 1] H(X|E = 1, Z)$$

Γνωρίζουμε όμως ότι $H(E|Z) \leq H(E) = h(\mathbb{P}[E = 1])$ αφού είδαμε ότι η δέσμευση μόνο να μειώσει την εντροπία μπορεί. Επίσης έχουμε $H(X|E = 1, Z) \leq \ln(|\mathcal{X} - 1)$ αφού γνωρίζουμε ότι από τις δυνατές τιμές $|\mathcal{X}|$ που μπορεί να πάρει η X η (γνωστή) τιμή της Z δεν είναι επιτρεπτή επιλογή αφού $E = 1$. □

Αν γνωρίζουμε ότι η X είναι ομοιόμορφα κατανεμημένη στο $\mathcal{X}$ μπορούμε να πάρουμε μια πιο εύχρηστη εκδοχή της ανισότητας του Fano.

Πόρισμα 4. Έστω μια τυχαία μεταβλητή X με ομοιόμορφα κατανεμημένη πάνω στο πεπερασμένο σύνολο $\mathcal{X}$, μία τυχαία μεταβλητή Y με τιμές στο $\mathcal{Y}$ και μία τυχαία μεταβλητή $Z : \mathcal{Y} \to \mathcal{X}$. Για κάθε αλυσίδα Markov $X \to Y \to Z$ έχουμε

$$\mathbb{P}[Z \neq X] \geq 1 - \frac{I(X;Y) + \ln 2}{\log(|\mathcal{X}|)},$$

όπου $h(x) = -x \ln x - (1 - x) \ln(1 - x)$.

Απόδειξη. Ξεκινώντας από την εκδοχή της Πρότασης 18 παρατηρούμε ότι η συνάρτηση binary εντροπίας $h(p) \leq h(1/2) = -1/2 \ln(1/2) - 1/2 \ln(1/2) = \ln 2$. Επίσης μπορούμε να γράψουμε $H(X|Z) = H(X) - I(X;Z) = \ln|\mathcal{X}| - I(X;Z)$. Χρησιμοποιώντας την ανισότητα Data Processing 14 έχουμε $I(X;Z) \leq I(X;Y)$ και τελικά $H(X|Z) \geq \ln|\mathcal{X}| - I(X;Y)$. Αντικαθιστώντας στην ανισότητα της Πρότασης 18 έχουμε

$$\ln 2 + \mathbb{P}[X \neq Z] \ln(|\mathcal{X}| - 1) \geq \ln|\mathcal{X}| - I(X;Y)$$
$$\mathbb{P}[X \neq Z] \geq 1 - \frac{I(X;Y) + \ln 2}{\ln|\mathcal{X}|}$$

□

3.3.2 Η Μέθοδος του Fano

Σε αυτήν την ενότητα θα δείξουμε μια γενικευμένη εκδοχή της μεθόδου του Fano στην περίπτωση που θέλουμε να εκτιμήσουμε μια συνάρτηση μιας ακολουθίας κατανομών, δηλαδή να κατασκευάσουμε ένα κάτω φράγμα για το minimax risk του Ορισμού 2.

Λήμμα 4. Έστω $\mathfrak{P}$ ένα σύνολο ακολουθιών κατανομών. Έστω $\mathfrak{F}_\mathcal{V} \subseteq \mathfrak{P}$ ένα υποσύνολο του $\mathfrak{P}$ δεικτοδοτούμενο από τα $v \in \mathcal{V}$ έτσι ώστε $\rho(\theta(\mathcal{P}_v), \theta(\mathcal{P}_u)) \geq 2\delta$ για κάθε $\mathcal{P}_v, \mathcal{P}_u \in \mathfrak{F}_\mathcal{V}$, όπου, $v \neq u \in \mathcal{V}$ και $\delta > 0$. Το minimax risk του Ορισμού 2 έχει κάτω φράγμα

$$\mathfrak{M}_N(\theta(\mathfrak{P}), \rho) \geq \delta \left(1 - \frac{1}{\ln|\mathcal{V}|} \left(N \sup_{v,u \in \mathcal{V}} D_{\text{kl}}(\mathcal{P}_v \| \mathcal{P}_u) + \ln 2\right)\right).$$

Απόδειξη. Χρησιμοποιώντας την Πρόταση 15 και την ανισότητα του Fano (Πόρισμα 4 έχουμε ένα κάτω φράγμα για το $\inf_{\Psi} \nu^m \left(\Psi(X^m) \neq V\right)$, και κατά συνέπεια

$$\mathfrak{M}_N\left(\theta(\mathfrak{P}), \rho\right) \geq \delta \inf_{|m|=N} \left(1 - \frac{I(V; X^m) + \ln 2}{\ln |\mathcal{V}|}\right) = \delta \left(1 - \frac{\sup_{|m|=N} I(V; X^m) + \ln 2}{\ln |\mathcal{V}|}\right),$$

όπου $I(V; X^m)$ είναι η από κοινού πληροφορία των V, X^m. Για να πάρουμε ένα πάνω φράγμα για την από κοινού πληροφορία $I(V; X^m)$ θα χρησιμοποιήσουμε την ανισότητα

$$I(V; X^m) \leq \frac{1}{|\mathcal{V}|^2} \sum_{v,u \in \mathcal{V}} D_{\mathrm{kl}}\left(P_v^m \| P_u^m\right) \leq \sup_{v,u \in \mathcal{V}} D_{\mathrm{kl}}\left(P_v^m \| P_u^m\right),$$

η οποία μπορεί να βρεθεί στα [5], [33], σελίδα 149 του [17]. Έχουμε

$$\begin{aligned}
\sup_{|m|=N} I(V; X^m) &\leq \sup_{|m|=N} \sup_{v,u \in \mathcal{V}} D_{\mathrm{kl}}\left(P_v^m \| P_u^m\right) \\
&= \sup_{|m|=N} \sup_{v,u \in \mathcal{V}} D_{\mathrm{kl}}\left(P_{v,1}^{m_1} \times \ldots \times P_{v,k}^{m_k} \| P_{u,1}^{m_1} \times \ldots \times P_{u,k}^{m_k}\right) \\
&= \sup_{v,u \in \mathcal{V}} \sup_{m=|N|} \sum_{i=1}^{k} m_i D_{\mathrm{kl}}\left(P_{v,i} \| P_{u,i}\right) \\
&\leq N \sup_{v,u \in \mathcal{V}, i \in I} D_{\mathrm{kl}}\left(P_{v,i} \| P_{u,i}\right) \\
&= N \sup_{v,u \in \mathcal{V}} D_{\mathrm{kl}}\left(\mathcal{P}_v \| \mathcal{P}_u\right),
\end{aligned}$$

όπου για να πάρουμε την δεύτερη ισότητα χρησιμοποιήσαμε την Πρόταση **??**. □

Κεφάλαιο 4

Μάθηση των Δυνάμεων μιας PBD

4.1 Εισαγωγή

4.1.1 Δυνάμεις μιας PBD

Υπενθυμίζουμε ότι μία n-PBD με διάνυσμα πιθανοτήτων $\boldsymbol{p} = (p_i)_{i\in[n]}$ είναι η τυχαία μεταβλητή $X = \sum_{i=1}^{n} X_i$, όπου κάθε X_i είναι μία 0/1 Bernoulli τυχαία μεταβλητή τέτοια ώστε $\mathbb{E}_{X_i}[=] p_i$. Ορίζουμε σαν την k-οστή δύναμη μιας PBD την PBD που αντιστοιχεί στο διάνυσμα πιθανότήτων $\boldsymbol{p}^k$. Εύκολα βλέπει κανείς ότι οι δυνάμεις μιας PBD σχετίζονται με τις ροπές της. Έστω τώρα το εξής πρόβλημα μη-εποπτευόμενης μάθησης: Σαν είσοδος μας δίνονται μία παράμετρος ακρίβειας $\epsilon \in (0,1)$ και μία παράμετρος εμπιστοσύνης $\delta \in (0,1)$ καθώς και ένας ακέραιος αριθμός m οποίος μπορεί να ξεπερνάει ακόμα και το n της PBD. Ο αλγόριθμος μάθησης μπορεί να τραβήξει ανεξάρτητα δείγματα από οποιαδήποτε PBD δύναμη και σαν έξοδο δίνει μία ακολουθία προσεγγιστικών κατανομών $\mathcal{Q} = (Q_i)_{i\in[m]}$ των δυνάμεων στο εύρος $1,\ldots,m$ τέτοια ώστε $\mathbb{P}\left[d_{\mathrm{tv}}\left(\mathcal{P},\mathcal{Q}\right) \le \varepsilon\right] \ge 1-\delta$. Σημειώνουμε ότι σε περίπτωση που θέλουμε να κάνουμε non-proper μάθηση οι κατανομές Q_i δεν είναι απαραίτητα PBDs. Η δειγματική πολυπλοκότητα των αλγορίθμων μάθησης δυνάμεων είναι μια συνάρτηση των $n, m, 1/\delta, 1/\varepsilon$. Το παραπάνω πρόβλημα μπορεί να λυθεί χρησιμοποιώντας m φορές κάποιον από τους υπάρχοντες αλγορίθμους μάθησης PBDs ώστε να μάθουμε κάθε δύναμη στο εύρος $[m]$ ξεχωριστά. Η δειγματική πολυπλοκότητα αυτής της μεθόδου είναι $O(m\sqrt{\log(1/\varepsilon)}/\varepsilon^2)$, αφού $O(\sqrt{\log(1/\varepsilon)}/\varepsilon^2)$ χρειάζονται για να μάθουμε προσεγγιστικά μία PBD. (βλέπε [15]). Μπορούμε να εκμεταλλευτούμε την σχέση που έχουν μεταξύ τους οι κατανομές σε μία ακολουθία δυνάμεων ώστε να πετύχουμε κάτι καλύτερο από πλευράς δειγματικής πολυπλοκότητας;

Πριν περάσουμε στα αποτελέσματα κρίνουμε σκόπιμο να δώσουμε κάποιες εφαρμογές του προβλήματος των PBD δυνάμεων ώστε να παρακινήσουμε την μελέτη του.

4.1.2 Τυχαίες Αποτιμήσεις Κάλυψης

Υποθέτουμε ότι θέλουμε να μοιράσουμε ένα σύνολο αντικειμένων U με $|U| = m$ σε n άτομα. Κάθε άτομο δίνει μία θετική τιμή σε κάθε υποσύνολο αντικειμένων $S \subseteq U$ η οποία ποσοτικοποιεί την προτίμησή του για το σύνολο S. Σε κάθε άτομο λοιπόν αντιστοιχεί μία *συνάρτηση αποτίμησης (valuation function)* $v_i : 2^U \to R_+$. Έστω ότι αποδίδουμε το σύνολο S_i στο i-οστό άτομο. Ψάχνουμε την *διαμέριση* $\{S_i : i \in n\}$ του συνόλου U έτσι ώστε να μεγιστοποιείται η συνολική ευημερία (welfare) των n ατόμων $\sum_{i=1}^{n} v_i(S_i)$. Το πρόβλημα αυτό είναι γνωστό ως μεγιστοποίηση της ευημερίας ("welfare maximization", "combinatorial auctions", ή "allocation problem").

Μία συνάρτηση *αποτίμησης* $v : 2^U \to \mathbb{R}_+$ είναι *μονότονη (monotone)* αν για κάθε $S \subseteq T \subseteq U$ ισχύει $v(S) \le v(T)$ και *υποτμηματική (submodular)* αν για κάθε $S \subseteq T \subseteq U$ ισχύει και για κάθε στοιχείο $e \in U$

$v(T \cap e) - v(T) \leq v(T \cap e) - v(S)$, δηλαδή η τιμή της συνάρτησης επηρεάζεται λιγότερο αν προσθέσουμε ένα στοιχείο e σε ένα μεγαλύτερο σύνολο.

Ορισμός 3. Έστω δύο σύνολα X, U. Μία συνάρτηση αποτίμησης $v : 2^U \to R_+$ λέγεται αποτίμηση κάλυψης (coverage valuation) αν υπάρχει μία οικογένεια συνόλων $\mathcal{A} = \{A_j : j \in U,\ A_j \subseteq X\}$ τέτοια ώστε για κάθε $S \subseteq U$

$$v(S) = \left| \bigcup_{j \in S} A_j \right|.$$

Για παράδειγμα έστω U ένα σύνολο υπαλλήλων και ένα σύνολο δεξιοτήτων. Έστω τώρα $A_j \subseteq U$ το υποσύνολο δεξιοτήτων του j-οστού υπαλλήλου. Τότε η αποτίμηση $v(S)$ που μετράει το πλήθος των διαφορετικών δεξιοτήτων των υπαλλήλων του συνόλου S είναι μία αποτίμηση κάλυψης.

Είναι εύκολο να δούμε ότι οι συναρτήσεις αποτίμησης κάλυψης είναι μονότονες. Είναι επίσης submodular αφού για ένα σύνολο $G \subseteq U$ η διαφορά

$$v(S \cup \{e\}) - v(S) = \left| \bigcup_{i \in S} A_i \cup A_e \right| - \left| \bigcup_{i \in S} A_i \right|$$

ισούται με το πλήθος των στοιχείων του A_e που δεν περιέχονται στο $\bigcup_{i \in S} A_i$. Το πλήθος αυτών των στοιχείων είναι προφανώς μικρότερο για ένα σύνολο $T \supseteq S$, δηλαδή είναι φθίνουσα συνάρτηση του μεγέθους των συνόλων. Οι συναρτήσεις κάλυψης έχουν εφαρμογές εκτός από την αλγοριθμική θεωρία παιγνίων στη βελτιστοποίηση, την θεωρία μάθησης, και στον αλγοριθμικό σχεδιασμό μηχανισμών (algorithmic mechanism design) [3, 13, 19].

Περνάμε τώρα στο μοντέλο των τυχαίων αποτιμήσεων κάλυψης. Αντιστοιχίζουμε τώρα σε κάθε στοιχείο $x_i \in X$ μία πιθανότητα $p_i \in [0, 1]$. Κατασκευάζουμε τα m σύνολα A_j τοποθετώντας το κάθε στοιχείο x_i ανεξάρτητα με πιθανότητα p_i στα m σύνολα A_j. Πλέον η *τυχαία* αποτίμηση κάλυψης v είναι μία τυχαία μεταβλητή που παίρνει τιμές στο εύρος $0, \ldots, n$. Μας ενδιαφέρει να μάθουμε την κατανομή πυκνότητας πιθανότητας D_k για κάθε $k \in [m]$ της αποτίμησης κάλυψης v. Στην περίπτωση όπου $k = 1$ η D_1 αντιστοιχεί στην PBD με διάνυσμα πιθανοτήτων $\boldsymbol{p} = (p_i)_{i \in [n]}$ αφού έχουμε μόνο ένα σύνολο A_1 και κάθε στοιχείο συμμετέχει στο σύνολο A_1 με πιθανότητα p_i. Συνεπώς ο πληθάριθμος του A_1 είναι μια τυχαία μεταβλητή που ακολουθεί PBD με διάνυσμα πιθανοτήτων $\boldsymbol{p} = (p_i)_{i \in [n]}$. Γενικότερα αν έχουμε $k \in [m]$ σύνολα A_i παρατηρούμε ότι η πιθανότητα ένα στοιχείο να μην περιέχεται σε κανένα από τα k A_i είναι $(1 - p_i)^k$ άρα η πιθανότητα το στοιχείο x_i να συμμετέχει στην ένωση $\bigcup_{i=1}^{k} A_i$ είναι $1 - (1 - p_i)^k$. Η κατανομή D_k λοιπόν είναι μία n-PBD με διάνυσμα πιθανοτήτων $(1 - (1 - p_i)^k)_{i \in [n]}$. Ένα φυσικό μοντέλο μάθησης για αυτό το πρόβλημα είναι να υποθέσουμε ότι μπορούμε να τραβάμε δείγματα από τις κατανομές των τυχαίων μεταβλητών $X_k = \bigcup_{i \in [k]}$ το οποίο αντιστοιχεί στο μοντέλο της μάθησης των δυνάμεων PBD που περιγράψαμε στην Ενότητα 4.1.1.

4.1.3 Ταυτότητες Newton-Girard

Οι ταυτότητες Newton-Girard σχετίζουν τα αθροίσματα των δυνάμεων των ριζών $\mu_i = \sum_{i=1}^{n} z_i^k$, $k = 1, \ldots, n$ με τους συντελεστές ενός πολυωνύμου $P(x) = \prod_{i=1}^{n}(x - z_i)$. Έτσι αν ξέρουμε τα αθροίσματα μ_i *ακριβώς* μπορούμε από τις ταυτότητες αυτές να βρούμε τους συντελεστές του πολυωνύμου $P(x)$ και στη συνέχεια να χρησιμοποιήσουμε κάποιον από τους γνωστούς αλγορίθμους της υπολογιστικής άλγεβρας ώστε να βρούμε τις ρίζες του z_i σε οσοδήποτε καλή ακρίβεια. Στο δικό πρόβλημα των PBD δυνάμεων το άθροισμα μ_i ισούται με την μέση τιμή της i-οστής PBD δύναμης. Παρατηρούμε επίσης ότι οι ρίζες του πολυωνύμου P πλέον αντιστοιχούν στις παραμέτρους p_i της PBD. Το πρόβλημα είναι ότι πλέον μπορούμε να μάθουμε μόνο προσεγγίσεις των μ_i μέσω των δειγμάτων των PBD δυνάμεων και κατ' επέκταση προσεγγίσεις των συντελεστών του πολυωνύμου P. Το ερώτημα εδώ είναι κατά πόσο μπορούμε να χρησιμοποιήσουμε αυτήν την "noisy" εκδοχή των ταυτοτήτων του Newton ώστε να μάθουμε τα p_i. Απαντάμε αυτό το ενδιαφέρον ερώτημα στην Ενότητα 4.3.

4.2 Μάθηση των Δυνάμεων μιας Διωνυμικής Κατανομής

Σε αυτήν την ενότητα μελετάμε το βασικό πρόβλημα της μάθησης των δυνάμεων μίας Διωνυμικής κατανομής. Η ιδέα εδώ είναι ότι μπορούμε να χρησιμοποιήσουμε την γνωστή από τη στατιστική εκτιμήτρια $\hat{p} = \frac{\sum_{i=1}^{m} X_i}{nm}$ ώστε να μάθουμε την παράμετρο p της Διωνυμικής. Η ακρίβεια που μας δίνεται χρησιμοποιώντας $O(1/\varepsilon^2)$ αριθμό δειγμάτων δεν επαρκεί για να προσεγγίσουμε όλα τα p_i αλλά δείχνουμε ότι τραβώντας δείγματα από προσεχτικά επιλεγμένες δυνάμεις μπορούμε να αποκτήσουμε ποιοτικές εκτιμήσεις του p, ικανές να προσεγγίσουν όλες τις δυνάμεις μιας Διωνυμικής κατανομής. Ξεκινάμε δείχνοντας την ποιότητα της εκτίμησης που μας παρέχεται από την εκτιμήτρια $\hat{p} = \frac{\sum_{i=1}^{m} X_i}{nm}$.

Πρόταση 19. Για κάθε $\varepsilon, \delta \in (0, 1/2)$, και $\psi > 0$, έστω $m = \lceil 4\ln(1/\delta)/(\varepsilon^2\psi^2) \rceil$ και $\hat{p} = (s_1 + \cdots + s_m)/(mn)$, όπου $s_1, \ldots, s_m$ είναι m ανεξάρτητα δείγματα από μια Διωνυμική κατανομή $B(n,p)$. Τότε, $\mathbb{P}\left[\hat{p} < p + \psi \mathrm{err}\left(n,p,\varepsilon\right)\right] \geq 1-\delta$, $\mathbb{P}\left[\hat{p} > p - \psi \mathrm{err}\left(n,p,\varepsilon\right)\right] \geq 1-\delta$.

Απόδειξη. Έστω $X = \sum_{i=1}^{m} s_i/n$. Τότε $s_i/n \in [0,1]$ και $\mathbb{E}\left[X\right] = mp$, $\mathrm{Var}\left[X\right] = \frac{m}{n}p(1-p)$, αφού τα δείγματα είναι ανεξάρτητα και προέρχονται από την $B(n,p)$. Δείχνουμε μόνο ότι $\mathbb{P}\left[\hat{p} - p > \psi \ \mathrm{err}\left(n,p,\varepsilon\right)\right] \leq \delta$ αφού η άλλη περίπτωση είναι παρόμοια. Από την Πρόταση 3 με $t = m\psi \ \mathrm{err}\left(n,p,\varepsilon\right)$ έχουμε

$$\begin{aligned}
\mathbb{P}\left[\hat{p} - p > \psi \mathrm{err}\left(n,p,\varepsilon\right)\right] =& \mathbb{P}\left[\hat{p} - p > t/m\right] \\
=& \mathbb{P}\left[X - \mathbb{E}\left[X\right] > \sqrt{m}\psi\varepsilon\sqrt{\mathrm{Var}\left[X\right]}\right] \\
\leq& \exp\left(-m\varepsilon^2\psi^2/4\right) \\
\leq& \delta,
\end{aligned}$$

όπου, για την τελευταία ανισότητα, χρησιμοποιούμε ότι $m = \lceil 4\ln(1/\delta)/(\varepsilon^2\psi^2) \rceil$. □

Πρόταση 20. Έστω $p \in [0,1]$, $\varepsilon, \delta \in (0,1/2)$, $\psi > 0$, $k = \lceil \ln(4/\delta)/\ln(2) \rceil$, $m = \lceil 4\ln(\lceil 2k/\delta \rceil)/(\varepsilon^2\psi^2) \rceil$. Για κάθε $i \in [k]$ έστω $w_i = \sum_{i=1}^{m} s_i/(nm)$, με $s_1, \ldots, s_m$ m ανεξάρτητα δείγματα από την $B(n,p)$. Αν $\hat{q}_1 = \min_{1 \leq i \leq k} w_i$, $\hat{q}_2 = \max_{1 \leq i \leq k} w_i$, τότε $\mathbb{P}\left[p - \psi \mathrm{err}\left(n,p,\varepsilon\right) < \hat{q}_1 < p\right] \geq 1-\delta$, $\mathbb{P}\left[p < \hat{q}_2 < p + \psi \mathrm{err}\left(n,p,\varepsilon\right)\right] \geq 1-\delta$. Ο συνολικός αριθμός δειγμάτων για τις εκτιμήσεις $\hat{q}_1, \hat{q}_2$ είναι $km = O\left(\ln(1/\delta)^2/(\varepsilon^2\psi^2)\right)$.

Απόδειξη. Δείχνουμε μόνο ότι $\mathbb{P}\left[p < \ \hat{q}_2 \ < p + \mathrm{err}\left(n,p,\varepsilon\right)\right] \geq 1 - \delta$ αφού η απόδειξη για το $\hat{q}_1 = \min_{1 \leq i \leq k} w_i$ είναι παρόμοια.

$$\begin{aligned}
&\mathbb{P}\mathrm{r}\left[(\max_i w_i < p) \bigcup (\max_i w_i > p + \psi \mathrm{err}\left(n,p,\varepsilon\right))\right] \\
&\leq \mathbb{P}\left[\max_i w_i < p\right] + \mathbb{P}\left[\max_i w_i > p + \psi \mathrm{err}\left(n,p,\varepsilon\right)\right] \\
&= \mathbb{P}\left[\bigcap_{i=1}^{k} (w_i < p)\right] + \mathbb{P}\left[\bigcup_{i=1}^{k} (w_i > p + \psi \mathrm{err}\left(n,p,\varepsilon\right))\right] \\
&\leq \left(\frac{1}{2}\right)^k + ku \\
&\leq \delta,
\end{aligned}$$

όπου η τελευταία ανισότητα προκύπτει από το $k = \lceil \ln(2/\delta)/\ln(2) \rceil$ και διαλέγοντας $u \leq \delta/(2k)$. Από την Πρόταση **??** έχουμε ότι το μέγεθος δείγματος

$$m = \lceil 4\ln(1/u)/(\varepsilon^2\psi^2) \rceil = \left\lceil 4\frac{\ln\left(\frac{2\lceil \ln(2/\delta)/\ln(2) \rceil}{\delta}\right)}{\varepsilon^2\psi^2} \right\rceil = O\left(\frac{\ln(1/\delta)}{\varepsilon^2\psi^2}\right)$$

επαρκεί για να διασφαλίσει ότι $\mathbb{P}\left[w_i < p + \psi \mathrm{err}\left(n, p, \varepsilon\right)\right] \leq \delta/(2k)$. □

4.2.1 Πάνω Φράγμα για $p \in [\varepsilon^2/n, 1 - \varepsilon^2/n]$.

Αποδεικνύουμε εδώ ότι $O\left(1/\epsilon^2\right)$ δείγματα είναι επαρκή για να μάθουμε όλες τις δυνάμεις μίας Διωνυμικής κατανομής $B(n,p)$ με σταθερή πιθανότητα επιτυχίας. Από το Πόρισμα **??** προκύπτει ότι για να μάθουμε properly μία Διωνυμική κατανομή $B(n,p)$ σε TVD $O\left(\varepsilon\right)$ αρκεί να προσεγγίσουμε την παράμετρό της p με σφάλμα $\mathrm{err}\left(n,p,\varepsilon\right) = \varepsilon\sqrt{p(1-p)/n}$. Σε αυτήν την ενότητα θα λύσουμε το πρόβλημα υποθέτοντας ότι το p δεν είναι πολύ κοντά στο 0 ή στο 1. Για την γενική περίπτωση επεκτείνουμε τον Αλγόριθμο 1 στην Ενότητα **??**.

Υποθέτοντας αρχικά ότι η άγνωστη παράμετρος p είναι περίπου $1 - 1/n$, δεν είναι ξεκάθαρο ότι τραβώντας δείγματα από έναν *σταθερό* αριθμό δυνάμεων επαρκεί για να προσεγγίσουμε όλες τις δυνάμεις. Θα μπορούσαμε να προσεγγίσουμε το p από την πρώτη δύναμη $B(n,p)$ αλλά όπως θα δούμε η ποιότητα της προσέγγισης δεν επαρκεί για να μάθουμε τις υπόλοιπες δυνάμεις. Από την άλλη, αν το p είναι "μικρό", π.χ. $1/2$, τότε χοντρικά μόνο οι πρώτες $\log(n)$ δυνάμεις έχουν σημασία καθώς όλες οι επόμενες μπορούν να προσεγγιστούν από την $B(n,0)$. Στην πραγματικότητα μπορούμε να δείξουμε ότι πάντα υπάρχει μία "σταθερή"(δηλαδή όχι συνάρτηση του n) δύναμη j, τέτοια ώστε η προσέγγιση $\hat{p}_1$ υψωμένη στην $j' \in \{j+1, j+2, \ldots, \log(n)\}$ προσεγγίζει το $p^{j'}$ αρκετά καλά. Τότε μπορούμε να κάνουμε brute force και να μάθουμε ξεχωριστά όλες τις δυνάμεις πριν από το j ξεχωριστά, εφόσον το πλήθος τους είναι σταθερό.

Συνεχίζουμε όμως να μην μπορούμε να γεφυρώσουμε το χάσμα ανάμεσα στη λύση όταν το p είναι σταθερό και όταν είναι κοντά στο 1.

Αν το p είναι μεγάλο (κοντά στο 1), μια λογική προσέγγιση θα ήταν να χρησιμοποιήσουμε την προσέγγιση του από τα δείγματα της πρώτης δύναμης για να βρούμε μία δύναμη ℓ^*, τέτοια ώστε $\hat{p}_1^{\ell^*} \approx const$. Στην συνέχεια αν τραβήξουμε δείγματα από την $B(n, p^{\ell^*})$ και υπολογίσουμε μία προσέγγιση $\hat{q}_1 \approx p^{\ell^*}$, τότε μπορούμε να δείξουμε ότι οι προσεγγίσεις $\hat{p}_j := \hat{q}_1^{\,j/\ell^*}$ είναι αρκετά καλές για τα p^j, για $j = 2, 3, \ldots, \ell^* - 1$; ουσιαστικά πηγαίνουμε "προς τα πίσω" στην ακολουθία των δυνάμεων. Αντίστοιχα στην περίπτωση όπου το $p \approx const$, είναι εφικτό να δείξουμε ότι υπάρχει πάντα μια σταθερή δύναμη k τέτοια ώστε το $\hat{q}_1^j$ προσεγγίζει το $p^{j\ell^*}$ αρκετά καλά για $j \geq k+1$. Τις υπόλοιπες δυνάμεις $j\ell^* + i$, για $j = 2, \ldots, k$ και $i = 1, \ldots, \ell^* - 1$, μπορούμε να τις προσεγγίσουμε κάθε μία ξεχωριστά με δείγματα από τις $B(n, p^{j\ell^*})$ για $j = 2, \ldots, k$ (υπολογίζοντας $\hat{q}_j \approx p^{j\ell^*}$), και προσεγγίζοντας το $p^{j\ell^*+i}$ με $\hat{q}_j\hat{p}_i$, όπου οι εκτιμήσεις $\hat{p}_i \approx p^i$ βρέθηκαν στην προηγούμενη φάση για $i = 1, 2, \ldots, \ell^*$. Ουσιαστικά αυτή τη φορά πηγαίνουμε "προς τα εμπρός" χρησιμοποιώντας τις προσεγγίσεις $\hat{q}_j$ και "γεμίζοντας" τα κενά μεταξύ των δυνάμεων $j\ell^*$ και $(j+1)\ell^*$ χρησιμοποιώντας τις προσεγγίσεις $\hat{p}_i$'s. Είναι εφικτό να αναλύσει κανείς το σφάλμα αυτής της μεθόδου αλλά επειδή στην ουσία θέλουμε να μελετήσουμε μία συνάρτηση πέντε διαστάσεων (n, p, ε, δ και των δυνάμεων ℓ) η ανάλυση είναι ιδιαίτερα επίπονη.

Η λύση που θα παρουσιάσουμε αποφεύγει όλες αυτές τις δυσκολίες και το τέχνασμα είναι να λύσουμε το πρόβλημα θεωρώντας ότι οι εκθέτες των δυνάμεων είναι θετικοί πραγματικοί αριθμοί. Αρχικά το πρόβλημα αυτό φαίνεται αρκετά δυσκολότερο καθώς πλέον πρέπει να προσεγγίσουμε ένα υπεραριθμήσιμο σύνολο δυνάμεων. Μαθαίνουμε λοιπόν όλες τις δυνάμεις $B(n, p^\ell)$ για κάθε $\ell \in \mathbb{R}_{++}$. Θεωρώντας το συνεχές φάσμα των δυνάμεων $\mathbb{R}_{++}$ αντί για $\mathbb{N}$ κάνει φανερή την συμμετρία του προβλήματος. Αυτό φαίνεται από το γεγονός ότι η $B(n, p^\ell)$ συγκλίνει σε "ντετερμινιστικές" κατανομές αφού $\lim_{\ell\to\infty} B(n, p^\ell) = B(n, 0)$ και $\lim_{\ell\to 0} B(n, p^\ell) = B(n, 1)$. Στη λύση μας είμαστε σε θέση να αντιμετωπίσουμε με τον ίδιο τρόπο τις "προς τα μπροστά" και "προς τα πίσω" περιπτώσεις που περιγράψαμε οι οποίες αντιστοιχούν σε δυνάμεις μικρότερες και μεγαλύτερες από το 1 αντίστοιχα και πλέον δεν υπάρχει η ανάγκη να γεμίσουμε τα "κενά" ανάμεσά στις δυνάμεις. Αυτό μας οδηγεί σε έναν κομψό αλγόριθμο που χρειάζεται να τραβήξει δείγματα μόνο από 2 διαφορετικές δυνάμεις. Από το Πόρισμα **??**, το πρόβλημα της ε-προσέγγισης των Διωνυμικών δυνάμεων $B(n, p^\ell)$ σε TVD ανάγεται στο να προσεγγίσουμε την ακολουθία των p^ℓ για κάθε $\ell \in (0, +\infty)$.

Θα εξηγήσουμε τώρα την βασική ιδέα του αλγορίθμου μας μέσα από ένα συγκεκριμένο παράδειγμα. Υποθέτουμε $p = 1 - 1/n + c$, όπου $c < 1/n$. Χωρίζουμε την δεκαδική αναπαράσταση του p σε 2 μέρη. Το πρώτο

κομμάτι αποτελείται από περίπου $\log n$ ψηφία 9 και καθορίζει το πόσο κοντά είναι το p στην μονάδα. (ή κοντά στο 0 στην συμμετρική περίπτωση όπου $p \approx 0$) και στο δεύτερο μέρος, στο οποίο θα ανεφερόμαστε ως το "σταθερό κομμάτι", , το οποίο αντιστοιχεί στο c. Η δεκαδική αναπαράσταση ενός τέτοιου p θα μπορούσε π.χ. να είναι

$$p = 0.\underbrace{99\ldots9}_{\#\log n}\ \underbrace{458382}_{\text{"constant" part}}\ .$$

Είναι ξεκάθαρο, ότι για τις πρώτες δυνάμεις, τα bits του σταθερού μέρους του p είναι ασήμαντα αλλά σε μεγαλύτερες δυνάμεις αυτά θα παίξουν ρόλο και κατά συνέπεια θα πρέπει να τα μάθουμε για να έχουμε ε-προσεγγίσεις των μεγαλύτερων δυνάμεων της Διωνυμικής. Χρησιμοποιώντας την Πρόταση 19 για να προσεγγίσουμε το $p = 1 - 1/n + c$ με δείγματα της πρώτης δύναμης βλέπουμε ότι μπορούμε να υπολογίσουμε μία εκτίμηση $\hat{p}$ με ακρίβεια χοντρικά $\sqrt{p(1-p)/n} \approx 1/n$. Μαθαίνουμε λοιπόν τα πρώτα $\log n$ 9 της δεκαδικής αναπαράστασης του p. Για να μάθουμε το σταθερό κομμάτι του p πρέπει να τραβήξουμε δείγματα από μία μεγαλύτερη δύναμη ώστε να μπορούμε να ξεχωρίσουμε τα αντίστοιχα bits με δεδομένη την ακρίβεια που μας παρέχει η Πρόταση 19. Για να μάθει κανείς το σταθερό κομμάτι στο συγκεκριμένο παράδειγμα πρέπει να προσεγγίσει το p χρησιμοποιώντας δείγματα από περίπου την n-οστή δύναμη. Η ιδέα των δύο τμημάτων της δεκαδικής αναπαράστασης του p μας δείχνει ότι για να προσεγγίσουμε όλη την ακολουθία των δυνάμεων του p^ℓ, $\ell \in (0, +\infty)$ πρέπει να μάθουμε *δύο* πράγματα

1. το πλήθος των αρχικών μηδενικών ή εννέα της δεκαδικής αναπαράστασης του p.
2. μία προσέγγιση του σταθερού τμήματος c του p.

Ο Αλγόριθμος 1 ακολουθεί στενά αυτή την διαισθητική ιδέα. Η αυστηρή ανάλυση του βασίζεται στα δύο βασικά λήμματα 5 και 6. Το Λήμμα 6 δείχνει ότι το να προσεγγίσουμε το p με δείγματα της πρώτης δύναμης είναι επαρκές για να μάθουμε μια προσέγγιση του $\hat{a} = -1/\log(\hat{p})$ του $a = -1/\log(p)$ που ουσιαστικά είναι η δύναμη που θα μας δώσει το σταθερό κομμάτι του p. Το Λήμμα 5 μας δείχνει την ακρίβεια που χρειαζόμαστε να προσεγγίσουμε το p για να ικανοποιήσουμε την συνθήκη $\text{err}\left(n, p^\ell, \varepsilon\right)$ *για κάθε* $\ell \in (0, +\infty)$, και είναι το κλειδί για να διαχειριστούμε την ανάλυση του πολυδιάστατου προβλήματος που περιγράψαμε. Ο Αλγόριθμος 1 τραβά $O\left(1/\varepsilon^2\right)$ δείγματα από δύο δυνάμεις, οπότε συνολικά η δειγματική αλλά και η χρονική πολυπλοκότητα του είναι $O\left(1/\varepsilon^2\right)$.

Σημειώνουμε εδώ ότι το $\psi(p^{\hat{a}})$ στον Αλγόριθμο 1 είναι μία καθολική (απόλυτη) σταθερά όπως θα δείξουμε παρακάτω.

Algorithm 1 Binomial Powers

Input : $O(\ln(1/\delta)^2/\varepsilon^2)$ δείγματα από τις δυνάμεις της $B(n,p)$.
Output : $\hat{a},\ \hat{q}_1, \hat{q}_2$.

1: Τραβάμε $O(\ln(1/\delta)/\varepsilon^2)$ δείγματα από την $B(n,p)$ για να υπολογίσουμε την προσέγγιση $\hat{p}$ χρησιμοποιώντας την Πρόταση 19.
2: Let $\hat{a} \leftarrow -1/\ln(\hat{p})$.
3: Τραβάμε $O\left(\ln(1/\delta)^2/\left(\varepsilon^2\psi(p^{\hat{a}})^2\right)\right)$ δείγματα από την $B(n, p^{\hat{a}})$ για να υπολογίσουμε προσεγγίσεις $\hat{q}_1,\ \hat{q}_2$ of p, $\hat{q}_1 \le p \le \hat{q}_2$, χρησιμοποιώντας την Πρόταση 20.
4: **return** $\hat{a},\ \hat{q}_1,\ \hat{q}_2$

Λήμμα 5. Έστω $\psi(p) = D\sqrt{\frac{p}{1-p}}\ \ln(1/p)$, όπου $D \approx 1.24263$. Έστω $p, \hat{q}_1, \hat{q}_2 \in (0,1)$ με $\hat{q}_1 < p < \hat{q}_2$. Τότε αν $p - \hat{q}_1 \le \psi(p)\text{err}\left(n,p,\varepsilon\right)$, $\hat{q}_2 - p \le \psi(p)\text{err}\left(n,p,\varepsilon\right)$, ισχύει $p^l - \hat{q}_1^l \le \text{err}\left(n, p^l, \epsilon\right)$ για κάθε $l \in (1, +\infty)$ και $\hat{q}_2^l - p^l \le \text{err}\left(n, p^l, \epsilon\right)$ για κάθε $l \in (0,1)$.

Απόδειξη. Από το Θεώρημα Μέσης Τιμής για την συνάρτηση $x \mapsto x^l$ έχουμε ότι $p^l - \hat{q}_1^l \le lp^{l-1}(p - \hat{q}_1)$ για $l \in (0,1)$ και $\hat{q}_2^l - p^l \le lp^{l-1}(\hat{q}_2 - p)$ για $l \in (1, +\infty)$. Στην συνέχεια θα βρούμε μία συνάρτηση $u(p)$ τέτοια

ώστε για κάθε $l > 0$ να ισχύει

$$\begin{aligned} u(p)lp^{l-1}\mathrm{err}\left(()\, n, p, \varepsilon\right) &\leq \mathrm{err}\left(()\, n, p^l, \varepsilon\right) \\ u(p)lp^{l-1}\sqrt{\frac{p(1-p)}{n}} &\leq \sqrt{\frac{p^l(1-p^l)}{n}} \\ u^2(p)l^2p^{2l-2}p(1-p) &\leq p^l(1-p^l) \\ u^2(p) &\leq \frac{p}{1-p}\frac{p^{-l}-1}{l^2} \end{aligned} \tag{4.1}$$

Έστω $f(l) = \frac{p^{-l}-1}{l^2}$, $g(p) = 6 - 6p^l + 4l\ln p + l^2(\ln p)^2$. Τότε

$$f'(l) = \frac{p^{-l}(-2+2p^l - l\ln p)}{l^3} \qquad f''(l) = \frac{p^{-l}(6-6p^l+4l\ln p + l^2(\ln p)^2)}{l^4}$$
$$g'(p) = \frac{2l(2-3p^l + l\ln(p))}{p}.$$

Θέτοντας $p^l = y$ έχουμε ότι το μέγιστο της κοίλης συνάρτησης $y \mapsto 2 - 3y + \ln(y)$ είναι $1 - \ln(3) < 0$. Συνεπώς η g είναι μία συνεχής, γνησίως φθίνουσα συνάρτηση του p και $\lim_{p\to 1} g(p) = 0$. Έτσι, $g(p) > 0$ για κάθε $p \in (0,1)$. Ως αποτέλεσμα, f είναι κυρτή συνάρτηση του l και ελαχιστοποιείται στο $\bar{l} = -\frac{C}{\ln p}$ (είναι η ρίζα της $f'(l) = 0$), όπου $C = 2 + W_n(-2/\mathrm{e}^2)$[1] ≈ 1.59362. Η ελάχιστη τιμή της f είναι $f(\bar{l}) = \frac{\mathrm{e}^C - 1}{C^2}(\ln p)^2$. Διαλέγοντας $u(p) = D\sqrt{\frac{p}{1-p}}\ \ln(1/p),\ D = \frac{\sqrt{\mathrm{e}^C-1}}{C}$ έχουμε ότι η Ανισότητα 4.1 ισχύει. □

Λήμμα 6. Έστω $\varepsilon \in (0, 1/6),\ n \geq 1$, και $p \in (\tau, \mu)$ όπου $\tau = \frac{1}{2}\left(1 - \sqrt{1 - 36\varepsilon^2/n}\right) \leq \varepsilon^2/n$, $\mu = \frac{1}{2}\left(1 + \sqrt{1 - 36\varepsilon^2/n}\right) \geq 1 - \varepsilon^2/n$. Επιπλέον, έστω $a, \hat{a} \in \mathbb{R}_{++}$ τέτοια ώστε $p^a = \hat{p}^{\hat{a}} = 1/\mathrm{e}$. If $|p - \hat{p}| \leq \mathrm{err}\left(n, p, \varepsilon\right)$, τότε $\frac{1}{\mathrm{e}^2} \leq p^{\hat{a}} \leq \frac{1}{\mathrm{e}^{3/2}}$.

Απόδειξη. Έστω $h = \mathrm{err}\left(n, p, \varepsilon\right)$. Η προσέγγιση Taylor της $f(x) = \ln(x)$ για $x \in (p-h, p+h)$ είναι $\ln(x) = \ln(p) + R_0(x)$. Έχουμε

$$\left|\frac{R_0(x)}{\ln p}\right| \leq \frac{1}{|\ln p|}\frac{h}{|p-h|} \leq \frac{1}{|(1-p)p/h + p - 1|} = \frac{1}{\left|\frac{\sqrt{n}}{\varepsilon}\sqrt{p(1-p)} + p - 1\right|},$$

αφού $|f'(x)| = 1/x \leq 1/|p-h|$, Θα βρούμε τώρα τις τιμές του p για τις οποίες η παραπάνω ποσότητα φράσσεται από πάνω από το $1/2$, δηλαδή το σύνολο των λύσεων της ανισότητας $\frac{\sqrt{n}}{\varepsilon}\sqrt{p(1-p)} \geq 3$ η οποία υποθέτοντας ότι $\varepsilon < 1/6$ δίνει $\frac{1}{2}\left(1 - \sqrt{1-36\varepsilon^2/n}\right) \leq p \leq \frac{1}{2}\left(1 + \sqrt{1-36\varepsilon^2/n}\right)$. Συνεπώς, για κάθε $\hat{p} \in (p-h, p+h)$ έχουμε

$$\frac{1}{2} \leq \frac{\ln\hat{p}}{\ln p} \leq \frac{3}{2} \ \Leftrightarrow\ -2\frac{1}{\ln p} \geq -\frac{1}{\ln\hat{p}} \geq -\frac{2}{3}\frac{1}{\ln p} \ \Leftrightarrow\ 2a \geq \hat{a} \geq \frac{2}{3}a \ \Leftrightarrow\ \frac{1}{\mathrm{e}^2} \leq p^{\hat{a}} \leq \frac{1}{\mathrm{e}^{3/2}}.$$

□ □

Θεώρημα 5. Έστω $\varepsilon \in (0, 1/6), n \in \mathbb{N}$. Τότε, για κάθε $p \in [\varepsilon^2/n, 1 - \varepsilon^2/n]$, ο Αλγόριθμος 1 τραβά $O(\ln(1/\delta)^2/\varepsilon^2)$ δείγματα και υπολογίζει εκτιμήσεις $\hat{a} \in \mathbb{R}_{++}$, $\hat{q}_1, \hat{q}_2 \in (0,1)$ τέτοιες ώστε $d_{\mathrm{tv}}\left(B(n, \hat{q}_1^l), B(n, p^{\hat{a}l})\right) = O(\varepsilon)$ για κάθε $l \in (1, +\infty)$ και $d_{\mathrm{tv}}\left(B(n, \hat{q}_2^l), B(n, p^{\hat{a}l})\right) = O(\varepsilon)$ για $l \in (0,1)$ με πιθανότητα τουλάχιστον $1 - \delta$.

[1] W_n denotes the Lambert W function.

Απόδειξη. Από το Πόρισμα **??** έχουμε ότι για να προσεγγίσουμε την $B(n,p^\ell)$ σε TVD ε χρειαζόμαστε μία προσέγγιση $\hat{p}_\ell$ του p^ℓ τέτοια ώστε $|p^\ell - \hat{p}_\ell| \leq \mathrm{err}\left(n,p^\ell,\varepsilon\right)$. Δείχνουμε ότι ο Αλγόριθμος 1 υπολογίζει προσεγγίσεις $\hat{q}_1,\ \hat{q}_2$ of p οι οποίες ικανοποιούν αυτή την συνθήκη. Χρησιμοποιούμε το Λήμμα 6 για να δείξουμε ότι $1/\mathrm{e}^2 \leq p^{\hat{a}} \leq 1/\mathrm{e}^{3/2}$, και κατά συνέπεια, $\psi(p^{\hat{a}}) \geq \psi(1/\mathrm{e}^2) = 0.983226$. Χρησιμοποιώντας την Πρόταση **??** τραβάμε $O\left(\ln(1/\delta)^2/(\varepsilon^2)\right)$ δείγματα για να πάρουμε τις προσεγγίσεις $\hat{q}_1,\ \hat{q}_2$ τέτοιες ώστε $\mathbb{P}\left[p - \mathrm{err}\left(n,p,\varepsilon\right) < \hat{q}_1 < p\right] \geq 1-\delta/2$, $\mathbb{P}\left[p < \ \hat{q}_2 \ < p + \mathrm{err}\left(n,p,\varepsilon\right)\right] \geq 1-\delta/2$, και επομένως η πιθανότητα επιτυχίας του να ικανοποιούνται και οι δύο περιορισμοί είναι τουλάχιστον $1-\delta$. Έχοντας τις προσεγγίσεις $\hat{q}_1,\ \hat{q}_2$ το αποτέλεσμα προκύπτει απευθείας από το Λήμμα 5. □

Σημείωση 1. Ο αλγόριθμος 1 μπορεί να τροποποιηθεί εύκολα ώστε να καλύψουμε την περίπτωση όπου μπορούμε να τραβάμε δείγματα μόνο από *ακέραιες* θετικές δυνάμεις του p, δηλαδή $\ell \geq 1$. Σε αυτή την περίπτωση παρατηρούμε ότι όταν $p \leq \mathrm{e}^{-C} \leq 0.2$, η συνάρτηση f του Λήμματος 5 ελαχιστοποιείται για $\ell = 1$, αφού η f είναι κυρτή και η θέση του ολικού της ελαχίστου είναι $\bar{\ell} \leq 1$. Αρκεί λοιπόν να διαλέξουμε $u(p) = \frac{p(p^{-1}-1)}{1-p} = 1$ και μπορούμε να μάθουμε όλες τις δυνάμεις $\ell \geq 1$ χρησιμοποιώντας την εκτίμηση $\hat{p}$ με δείγματα από την πρώτη δύναμη σύμφωνα με την Πρόταση **??**. Αν $p \geq 0.2$ τότε μπορούμε απλώς να τρέξουμε τον Αλγόριθμο 1 με $\lceil \hat{a} \rceil$ αντί για $\hat{a}$. Τότε, $0.2/\mathrm{e}^2 \leq p^{\lceil \hat{a} \rceil} \leq 1/\mathrm{e}^{3/2}$, και επομένωςψ$(p) \geq \psi(0.2/\mathrm{e}^2)$, το οποίο σημαίνει ότι ο Αλγόριθμος 1 χρησιμοποιεί $O\left(\ln(1/\delta)/\varepsilon^2\right)$ δείγματα για να μάθουμε όλες τις δυνάμεις $\ell \geq 1$.

Σημείωση 2. Στον Αλγόριθμο 1 θα μπορούσαμε να χρησιμοποιήσουμε τις προσεγγίσεις $\hat{p} \approx p$ της Πρότασης **??** χωρίς καμία διαφορά στην πράξη. Χρησιμοποιούμε τις προσεγγίσεις $\hat{q}_1,\ \hat{q}_2$ του p, ώστε να έχουμε μία ενοποιημένη ανάλυση χρησιμοποιώντας το Θεώρημα Μέσης Τιμής στην απόδειξη του Λήμματος 5.

4.2.2 Ακραίες τιμές του p. $p \in [\varepsilon^2/n^d,\ 1-\varepsilon^2/n^d]$.

Σε αυτήν την ενότητα γενικεύουμε τον Αλγόριθμο 1 και την ανάλυση του στην περίπτωση όπου το p είναι πολύ κοντά στο 0 ή το 1 δηλαδή βρίσκεται στο διάστημα $[\varepsilon^2/n^d, 1-\varepsilon^2/n^d]$, για κάποιον ακέραιο $d \in \mathbb{N}_+$. Ως εκ τούτου καλύπτουμε όλες τις τιμές του p που μπορούν να αναπαρασταθούν με $O(\log n)$ bits.

Θεώρημα 6. Έστω $\varepsilon \in (0,1/6)$ και $d \in \mathbb{N}_+$, $n \in \mathbb{N}$, $n \geq 5$. Για κάθε $p \in [\varepsilon^2/n^d, 1-\varepsilon^2/n^d]$, μία επέκταση του Αλγορίθμου 1 χρησιμοποιεί $O(\log(d)\log(\log(d)/\delta)/\varepsilon^2)$ δείγματα και υπολογίζει $t, \hat{a} \in (0,+\infty)$, $\hat{q}_1, \hat{q}_2 \in (0,1)$ τέτοια ώστε $d_{\mathrm{tv}}\left(B(n,\hat{q}_2^l), B(n,p^{lt\hat{a}})\right) \leq O(\varepsilon)$ για κάθε $l \in (0,1)$ και $d_{\mathrm{tv}}\left(B(n,\hat{q}_1^l), B(n,p^{lt\hat{a}})\right)$ για κάθε $l \in (1,+\infty)$ με πιθανότητα τουλάχιστον $1-\delta$.

Απόδειξη. Ξεκινάμε περιγράφοντας την επέκταση του Αλγορίθμου 1. Προσέξτε ότι, σε αυτή τη περίπτωση αρκεί να βρούμε $t \in (0,+\infty)$ τέτοιο ώστε $p^t \in [\varepsilon^2/n, 1-\varepsilon^2/n]$. Τότε απλώς καλούμε τον Αλγόριθμο 1 χρησιμοποιώντας την $B(n,p^t)$ ως την "πρώτη" δύναμη της ακολουθίας, για να πάρουμε τα $\hat{q}_1,\ \hat{q}_2,\ \hat{a}$ τέτοια ώστε $d_{\mathrm{tv}}\left(B(n,\hat{q}_2^l), B(n,p^{lt\hat{a}})\right) \leq O(\varepsilon)$ for $l \in (0,1)$ και $d_{\mathrm{tv}}\left(B(n,\hat{q}_1^l), B(n,p^{lt\hat{a}})\right)$ για $l \in (1,+\infty)$. Για να βρούμε το ζητούμενο t πρέπει πρώτα να τραβήξουμε δείγματα από την αρχική $B(n,p)$ και έπειτα χρησιμοποιώντας τη Πρόταση **??** παίρνουμε $\hat{q}_{1,1},\ \hat{q}_{1,2}$ τέτοια ώστε $\mathbb{P}\left[p - \mathrm{err}\left(n,p,\varepsilon\right) < \hat{q}_{1,1} < p < \hat{q}_{1,2} < p + \mathrm{err}\left(n,p,\varepsilon\right)\right] \geq 1-\delta$. Διακρίνουμε τις παρακάτω περιπτώσεις.

- $\hat{q}_{1,1} > \varepsilon^2/n$ και $\hat{q}_{1,2} < 1-\varepsilon^2/n$. Σε αυτή την περίπτωση δεν έχουμε κάτι να δείξουμε αφού μπορούμε να χρησιμοποιήσουμε τον Αλγόριθμο 1 απευθείας.
- $\hat{q}_{1,1} < \varepsilon^2/n$. Έστω $I_1 = \{1/(i\ln n) : i \in \{2,\dots,d\}\}$. Χρησιμοποιώντας την Πρόταση 20 τραβάμε $O(d\ln^2(d/\delta)/\varepsilon^2)$ δείγματα από τις δυνάμεις $B(n,p^l)$, $l \in I_1$, και κατασκευάζουμε το σύνολο των προσεγγίσεων $Q_1 = \{\hat{q}_{i,1} : i \in I_1\}$ έτσι ώστε με πιθανότητα $1-\delta/2$ όλα τα $\hat{q}_{i,1} \in Q_1$ ικανοποιούν τους περιορισμούς $p^i - \mathrm{err}\left(n,p^i,\varepsilon\right) < \hat{q}_{i,1} < p^i$. Δείχνουμε πρώτα ότι υπάρχει ένα στοιχείο t του $_1$ τέτοιο ώστε $\hat{q}_{t,1} \geq \varepsilon^2/n$.

  Αρκεί να δείξουμε ότι ένα τέτοιο t υπάρχει όταν το p είναι το μικρότερο δυνατό, δηλαδή $p = \varepsilon^2/n^d$. Τότε $p^{1/(d\ln n)} = \varepsilon^{2/(d\ln n)}/\mathrm{e}$ και $\hat{q}_{1/(d\ln n)} \geq p^{1/(d\ln n)} - \frac{\varepsilon}{2\sqrt{n}} \geq \varepsilon^2/n$, για κάθε $n \geq 7$.

  Διαλέγουμε το t να είναι το μέγιστο στοιχείο του I_1 τέτοιο ώστε $\hat{q}_{t,1} \geq \varepsilon^2/n$. Τότε $p^t > \varepsilon^2/n$ αφού $\hat{q}_{t,1} < p^t$. Επιπλέον, $p^t < 1-\varepsilon^2/n$. Για να το δείξουμε αυτό, γράφουμε $t = 1/(\rho\ln n)$ για κάποιο $\rho \geq 2$ and $t' = 1/((\rho-1)\ln n)$. Τότε, $p^{t'} \leq \hat{q}_{t',1} + \mathrm{err}\left(n,p^{t'},\varepsilon\right) \leq \varepsilon^2/n + \varepsilon/(2\sqrt{n}) \leq \varepsilon/\sqrt{n}$. Οπότε,

$p^t = p^{1/(\rho \log n)} = p^{\frac{1}{(\rho-1)\log n}\frac{\rho-1}{\rho}} = (p^{t'})^{\frac{\rho-1}{\rho}} \leq (\varepsilon/\sqrt{n})^{\frac{\rho-1}{\rho}} = \frac{\varepsilon}{\sqrt{n}}\left(\frac{\sqrt{n}}{\varepsilon}\right)^{1/\rho} \leq \frac{\sqrt{\varepsilon}}{n^{1/4}} \leq 1-\varepsilon^2/n$, όπου η τελευταία ανισότητα ισχύει για $n \geq 2,\ \varepsilon < 1/2$.

- $\hat{q}_{1,2} > 1-\varepsilon^2/n$. Σε αυτή τη περίπτωση θεωρούμε το σύνολο $I_2 = \{n^{i/3} : i \in \{0\} \cup [3d]\}$. Χρησιμοποιώντας την Πρόταση 20 τραβάμε $O(d\ \ln^2(d/\delta)/\varepsilon^2)$ δείγματα και υπολογίζουμε το σύνολο των προσεγγίσεων $Q_2 = \{\hat{q}_{i,2} : i \in I_2\}$ τέτοιο ώστε με πιθανότητα $1-\delta/2$ όλα τα $\hat{q}_{i,2} \in Q_2$ ικανοποιούν τους περιορισμούς $p^i < \hat{q}_{i,2} < p^i + \mathrm{err}\left(n, p^i, \varepsilon\right)$. Όπως και στην προηγούμενη περίπτωση δείχνουμε πρώτα ότι υπάρχει ένα $t \in I_2$ τέτοιο ώστε $\hat{q}_{t,2} \leq 1-\varepsilon^2/n$. Αρκεί να το δείξουμε για $p = 1-\varepsilon^2/n^d$. Τότε διαλέγουμε $t = n^d$ και έχουμε $p^t = (1-\varepsilon^2/n^d)^{n^d} \leq \mathrm{e}^{-\varepsilon^2} \leq 1-\varepsilon^2/2 \leq 1-\varepsilon^2/n$ για $\varepsilon < 0.85,\ n \geq 2$. Ξεκινώντας από το 1 βρίσκουμε το μικρότερο στοιχείο t του $_2$ τέτοιο ώστε $\hat{q}_{t,2} < 1-\varepsilon^2/n$. Ισχυριζόμαστε ότι $\varepsilon^2/n < p^t < 1-\varepsilon^2/n$. Προφανώς $p^t < 1-\varepsilon^2/n$ since $p^t < \hat{q}_{t,2}$. Για να δείξουμε την άλλη ανισότητα, γράφουμε $t = n^{\rho/3}$ και $t' = n^{(\rho-1)/3}$ για κάποιο $\rho \in [3d]$. Έχουμε ότι $\hat{q}_{t',2} \geq 1-\varepsilon^2/n$ και επομένως $p^{t'} \geq 1-\varepsilon^2/n - \mathrm{err}\left(n,p,\varepsilon\right) \geq 1-\varepsilon^2/n - \varepsilon/(2\sqrt{n}) \geq 1-\varepsilon/\sqrt{n}$, for $n \geq 4$. Άρα, $p^t = p^{n^{\rho/3}} = p^{n^{(\rho-1)/3+1/3}} = (p^{t'})^{n^{1/3}} \geq (1-\varepsilon/\sqrt{n})^{n^{1/3}} \geq \mathrm{e}^{-2\varepsilon n^{1/3}/n^{1/2}} = \mathrm{e}^{-2\varepsilon/n^{1/6}} \geq \varepsilon^2/n$, όπου για την δεύτερη ανισότητα χρησιμοποιούμε την ανισότητα $1-x \geq \mathrm{e}^{-2x}$ για $x \in [0, 0.75]$ και η τελευταία τελευταία ανισότητα ισχύει για $\varepsilon \leq 1/\mathrm{e},\ n \geq 1$.

Είναι εύκολο να ελέγξει κανείς ότι η παραπάνω διαδικασία μπορεί να γίνει με δυαδική αναζήτηση πάνω sto d δίνοντας έτσι την αρκετά καλύτερη δειγματική πολυπλοκότητα $O(\log(d)\log(\log(d)/\delta)/\varepsilon^2)$. Ουσιαστικά για κάθε ρ που ελέγχουμε, ο αλγόριθμος επιλέγει $O(\log(\log(d)/\delta)/\varepsilon^2)$ ανεξάρτητα δείγματα το οποίο οδηγεί στην τελική συνολική δειγματική πολυπλοκότητα $O(\log(d)\log(\log(d)/\delta)/\varepsilon^2)$. □

4.2.3 Κάτω Φράγμα

Είναι γενικά γνωστό ότι για να ξεχωρίσουμε δύο Διωνυμικές κατανομές χρειαζόμαστε $\Omega(1/\varepsilon^2)$ δείγματα (βλέπε [11]). Συνεπώς το ίδιο κάτω φράγμα ισχύει και για να μάθουμε μία Διωνυμική κατανομή. Αυτό το φράγμα δεν μπορεί να χρησιμοποιηθεί στην δικιά μας περίπτωση καθώς η είσοδος του αλγορίθμου αποτελείται από δείγματα που προέρχονται από όλη την ακολουθία των δυνάμεων της Διωνυμικής. Για να δώσουμε μία αυστηρή απόδειξη του κάτω φράγματος για την μάθηση των Δυνάμεων Διωνυμικών κατανομών θα χρησιμοποιήσουμε την μέθοδο του Fano που περιγράψαμε στο Κεφάλαιο 2. Διατυπώνουμε το κάτω φράγμα για την μάθηση μιας ακολουθίας από δυνάμεις μιας Διωνυμικής κατανομής στο επόμενο θεώρημα

Θεώρημα 7. Έστω A ένας αλγόριθμος που επιστρέφει κατανομές πιθανότητας οι οποίες βρίσκονται σε TVD το πολύ ε από τις αντίστοιχες δυνάμεις $B(n,p^i)$ για κάθε $i \in \{1,2,3,\ldots\}$, χρησιμοποιώντας δείγματα από τις κατανομές $B(n,p^i)$ με πιθανότητα επιτυχίας το λιγότερο $2/3$. Τότε ο A χρησιμοποιεί $\Omega\left(1/\varepsilon^2\right)$ δείγματα από τις δυνάμεις.

Επειδή η απόδειξη του θεωρήματος είναι αρκετά τεχνική θα δώσουμε αρχικά ένα σκαρίφιμά της.

Χρησιμοποιώντας τον συμβολισμό του Λήμματος 4 και αφού μας ενδιαφέρει να προσεγγίσουμε τις πυκνότητες των κατανομών έχουμε $\theta(\mathcal{P}) = (f_i)_{i\in\mathbb{N}}$. Συνεπώς θα χρησιμοποιήσουμε την μετρική $\rho(\theta(\mathcal{P}),\theta(\mathcal{Q})) = d_{\mathrm{tv}}\left(\mathcal{P},\mathcal{Q}\right) = \sup_{i\in\mathbb{N}} d_{\mathrm{tv}}(f_i,\hat{f}_i)$.
Έστω $\delta = \Theta(1/\sqrt{n\,N})$. Έστω $p_1 = 1/2$, $p_2 = 1/2+\delta/4$, $p_3 = 1/2+\delta/2$. Έστω $\mathcal{P}_1 = (B(n,(1/2)^i))_{i\in\mathbb{N}}$, $\mathcal{P}_2 = B(n,(1/2+\delta/4)^i)_{i\in\mathbb{N}}$, $\mathcal{P}_3 = B(n,(1/2+\delta/2)^i)_{i\in\mathbb{N}}$. Η TVD των πρώτων δυνάμεων αυτών των Διωνυμικών είναι $\Omega(1/\sqrt{N})$. Για να το δείτε αυτό παρατηρήστε ότι αφού η διακύμανση των Διωνυμικών είναι $O(n)$ μπορούμε να τις προσεγγίσουμε χρησιμοποιώντας κανονικές κατανομές με αμελητέο σφάλμα. Δύο κανονικές κατανομές οι οποίες έχουν παραπλήσιες διακυμάνσεις έχουν TVD η οποία είναι χοντρικά ανάλογη με την διαφορά των μέσων τιμών τους διαιρεμένη με την "κοινή" τους τυπική απόκλιση η οποία είναι $\Omega(1/\sqrt{N})$. Έτσι παίρνουμε το κάτω φράγμα για την TVD. Στην συνέχεια δείχνουμε ένα πάνω φράγμα για την απόκλιση Kullback-Leibler ανάμεσα στις *ακολουθίες* των δυνάμεων, δηλαδή $D_{\mathrm{kl}}\left(\mathcal{P}_1\|\mathcal{P}_3\right) = O(1/N)$. Είναι εύκολο να διαπιστώσετε ότι αυτό το πάνω φράγμα ισχύει για την πρώτη δύναμη. Για να αποδείξουμε ότι ισχύει για όλες τις δυνάμεις παρατηρούμε ότι η $D_{\mathrm{kl}}\left(B(n,p)\|B(n,q)\right)$ είναι αύξουσα συνάρτηση της απόστασης των παραμέτρων $|p-q|$ (βλέπε Πρόταση 9). Έτσι, αφού οι αποστάσεις των p_i' των τριών Διωνυμικών που θεωρήσαμε μειώνονται σε υψηλότερες δυνάμεις η απόκλιση Kullback-Leibler της πρώτης δύναμης είναι ουσιαστικά ένα πάνω φράγμα της $D_{\mathrm{kl}}\left(\mathcal{P}_1\|\mathcal{P}_3\right)$. Στη συνέχεια, εφαρμόζοντας το Λήμμα 4 έχουμε ότι $\mathfrak{M}_N\left(\theta(\mathfrak{P}),\rho\right) = \Omega(1/\sqrt{N})$, το οποίο

με τη σειρά του σημαίνει ότι για να έχουμε έναν αλγόριθμο που προσεγγίζει όλες τις δυνάμεις σε απόσταση μικρότερο από ϵ πρέπει να ισχύει $\mathfrak{M}_N\left(\theta(\mathfrak{P}),\rho\right) < \epsilon$ και κατά συνέπεια ο αριθμός των δειγμάτων N πρέπει να είναι $\Omega(1/\varepsilon^2)$.

4.2.4 Η πλήρης Απόδειξη

Απόδειξη. Υπενθυμίζουμε ότι αναζητούμε μία οικογένεια κατανομών η οποία να ικανοποιεί την συνθήκη

$$\rho(\theta(\mathcal{P}),\theta(\mathcal{Q})) = \Omega\left(\delta\right).$$

Το να ικανοποιείται αυτή η συνθήκη με μία σταθερά c αντί για την τιμή 2 αλλάζει το κάτω φράγμα μόνο κατά έναν σταθερό παράγοντα. Έτσι, για να απλοποιήσουμε την ανάλυση, δεν θα υπολογίσουμε τις σταθερές σε όσα ακολουθούν. Σημειώνουμε ότι αυτές είναι απόλυτες σταθερές που δεν εξαρτώνται από καμία από τις παραμέτρους, ε, δ, του προβλήματος.

Οι ακολουθίες δυνάμεων που θα χρησιμοποιήσουμε είναι οι εξής για $\delta = \Theta(1/\sqrt{n\,N})$. Έστω $p_1 = 1/2$, $p_2 = 1/2 + \delta/4$, $p_3 = 1/2 + \delta/2$, και $P_{1,1} = B(n,p_1)$, $P_{2,1} = B(n,p_2), P_{1,3} = B(n,p_3)$ 3 διωνυμικές κατανομές με αντίστοιχες ακολουθίες δυνάμεων: $\mathcal{P}_1 = (B(n,p_1^i))_{i\in(1,+\infty)}$, $\mathcal{P}_2 = (B(n,p_2^i))_{i\in(1,+\infty)}$, $\mathcal{P}_3 = (B(n,p_3^i))_{i\in(1,+\infty)}$.

Για την απόσταση ολικής κύμανσης των παραπάνω ζευγών για $i,j \in \{1,2,3\}$, $i \neq j$, έχουμε $d_{\mathrm{tv}}\left(\mathcal{P}_i,\mathcal{P}_j\right) = \Omega\left(1/\sqrt{N}\right)$. Χωρίς βλάβη της γενικότητας αποδεικνύουμε ότι $d_{\mathrm{tv}}\left(\mathcal{P}_1,\mathcal{P}_2\right) = \Omega(1/\sqrt{N})$. Από τον ορισμό της απόστασης ολικής κύμανσης για ακολουθίες κατανομών βλέπουμε ότι για να πάρουμε ένα κάτω φράγμα για την μετρική ρ πρέπει απλώς να δείξουμε ότι η συνολική απόσταση ολικής κύμανσης των $P_{1,1}, P_{2,1}$ είναι $\Omega(1/\sqrt{N})$, δηλαδή αρκεί να θεωρήσουμε μόνο την πρώτη δύναμη των ακολουθιών.

Έστω $\mu_1 = \mathbb{E}\left[P_{1,1}\right]$, $\mu_2 = \mathbb{E}\left[P_{2,1}\right]$, $\sigma_1^2 = \mathrm{Var}\left[P_{1,1}\right]$, $\sigma_2^2 = \mathrm{Var}\left[P_{2,1}\right]$.

Χρησιμοποιούμε πρώτα το Λήμμα **??** για να προσεγγίσουμε τις $P_{1,1}, P_{2,1}$ με διακριτοποιημένες κανονικές κατανομές $\mathrm{DN}\left(\mu_1,\sigma_1\right)$, $\mathrm{DN}\left(\mu_2,\sigma_2\right)$. Αφού τα σ_1,σ_2 είναι και τα δύο $O(\sqrt{n})$, το σφάλμα της προσέγγισης με τις διακριτοποιημένες Κανονικές κατανομές θα είναι $O(1/\sqrt{n})$. Από την Πρόταση 11 παίρνουμε ότι μπορούμε να προσεγγίσουμε $\mathcal{N}\left(\mu_2,\sigma_2\right)$ χρησιμοποιώντας μία Κανονική με την ίδια μέση τιμή αλλά με διακύμανση σ_1^2. Εφαρμόζοντας την Πρόταση 11 παίρνουμε

$$d_{\mathrm{tv}}\left(\mathcal{N}\left(\mu_2,\sigma_2\right),\mathcal{N}\left(\mu_2,\sigma_1\right)\right) \leq \frac{1}{2}\frac{\sigma_1^2-\sigma_2^2}{\sigma_2^2} = \frac{1}{2}\frac{n/4 - n(1/4-\delta^2/16)}{n(1/4-\delta^2/16)} = O\left(\delta^2\right) = O\left(1/n\right)$$

Έστω τώρα ότι το ζεύγος των συνεχών κανονικών κατανομών $\mathcal{N}\left(\mu_1,\sigma_1\right)\mathcal{N}\left(\mu_2,\sigma_1\right)$. Χρησιμοποιώντας την Πρόταση 10 έχουμε ότι

$$d_{\mathrm{tv}}\left(\mathcal{N}\left(\mu_1,\sigma_1\right),\mathcal{N}\left(\mu_2,\sigma_1\right)\right) = \mathrm{erf}\left(\frac{n\delta}{4\sqrt{2}\sqrt{n}}\right) = \mathrm{erf}\left(\frac{\sqrt{n}\delta}{4\sqrt{2}}\right) = \mathrm{erf}\left(\frac{1}{4\sqrt{2}\sqrt{N}}\right) \geq \frac{1}{9\sqrt{N}},$$

από το Πόρισμα 1. Επομένως, από την τριγωνική ανισότητα, έχουμε

$$d_{\mathrm{tv}}\left(\mathcal{N}\left(\mu_1,\sigma_1\right),\mathcal{N}\left(\mu_2,\sigma_2\right)\right) \geq \frac{1}{9\sqrt{N}} - O\left(\frac{1}{n}\right)$$

Εφαρμόζουμε το Λήμμα **??** όπου σ_1,σ_2 are $O\left(\sqrt{n}\right)$ παίρνουμε $\ell+m+u = O\left(\frac{1}{\sqrt{n}}\right)$, αφού από τις ανισότητες του Komatsu (Πρόταση 7) έχουμε $\mathrm{erfc}\left(\sqrt{n}\right) = \Theta\left(\frac{\mathrm{e}^{-n}}{\sqrt{n}}\right)$ και από την Πρόταση 6 έχουμε ότι $\mathrm{erf}\left(1/\sqrt{n}\right) = \Theta\left(\frac{1}{\sqrt{n}}\right)$. Επομένως

$$d_{\mathrm{tv}}\left(\mathrm{DN}\left(\mu_1,\sigma_1\right),\mathrm{DN}\left(\mu_2,\sigma_2\right)\right) \geq \frac{1}{9\sqrt{N}} - O\left(\frac{1}{\sqrt{n}}\right)$$

Συνολικά, χρησιμοποιώντας την τριγωνική ανισότητα και τα παραπάνω φράγματα έχουμε ότι

$$d_{\mathrm{tv}}\left(P_{1,1},P_{2,1}\right) \geq \frac{1}{9\sqrt{N}} - O\left(\frac{1}{\sqrt{n}}\right)$$

Συνεχίζουμε αποδεικνύοντας ένα πάνω φράγμα για την απόκλιση Kullback-Leibler ανάμεσα σε όλες τις δυνάμεις, δηλαδή ένα φράγμα το $\sup_{a\in\mathbb{N}} D_{\mathrm{kl}}\left(P_{1,a}\|P_{3,a}\right)$. Για να εφαρμόσουμε το Θεώρημα 4 αρκεί να δείξουμε ότι ισχύει: $\sup_{i,j\in[3],\, a\in\mathbb{N}} D_{\mathrm{kl}}\left(P_{i,a}\|P_{j,a}\right) = O\left(1/N\right)$. Από την Πρόταση 9 είναι εμφανές ότι χρειάζεται να φράξουμε μόνο την απόκλιση Kullback-Leibler για το ζεύγος των πιο απομακρυσμένων p_i, δηλαδή τις αποστάσεις $\sup_{a\in\mathbb{N}} D_{\mathrm{kl}}\left(P_{1,a}\|P_{3,a}\right)$, και $\sup_{a\in\mathbb{N}} D_{\mathrm{kl}}\left(P_{3,a}\|P_{1,a}\right)$. Σημειώνουμε ότι είναι εύκολο να διαπιστώσει κανείς ότι $D_{\mathrm{kl}}\left(P_{1,a}\|P_{3,a}\right) \approx D_{\mathrm{kl}}\left(P_{3,a}\|P_{1,a}\right)$ για κάθε $a \in \mathbb{N}$ και επομένως θα φράξουμε την $D_{\mathrm{kl}}\left(P_{1,a}\|P_{3,a}\right)$.

Χρησιμοποιώντας την Πρόταση 8 για τις $P_{1,a}$ και $P_{3,a}$ δίνει

$$D_{\mathrm{kl}}\left(P_{1,a}\|P_{3,a}\right) = 2^{-a} n \ln\left(2^{-a}\left(\frac{1}{\frac{\delta}{2}+\frac{1}{2}}\right)^{a}\right) + \left(1-2^{-a}\right) n \ln\left(\frac{1-2^{-a}}{1-\left(\frac{\delta}{2}+\frac{1}{2}\right)^{a}}\right)$$

Έστω $f(\delta) = D_{\mathrm{kl}}\left(P_{1,a}\|P_{3,a}\right)$ που ορίζεται από την παραπάνω παράσταση. Αναπτύσσοντας κατά Taylor την $f(\delta)$ γύρω 0 δίνει $f(\delta) = 0 + R_1(z)$ for a $z \in [-\delta, \delta]$. Για να φράξουμε το σφάλμα της προσέγγισης Taylor φράσσουμε την παράγωγο $f''(\delta)$

$$\begin{aligned}
f''(\delta) &= \frac{an\left(\left(2^a-1\right)a(\delta+1)^a - \left(2^a-(\delta+1)^a\right)\left((\delta+1)^a-1\right)\right)}{(\delta+1)^2\left(2^a-(\delta+1)^a\right)^2} \\
&\leq \frac{an\left(\left(2^a-1\right)a(\delta+1)^a\right)}{(\delta+1)^2\left(2^a-(\delta+1)^a\right)^2} \\
&\leq \frac{na^2(2^a-1)(3/2)^a}{\left(2^a-(3/2)^a\right)^2} \\
&\leq \frac{na^2 3^a}{\left(2^a/4\right)^2} = 16n\, a^2(3/4)^a \leq 105n
\end{aligned}$$

Έτσι, $D_{\mathrm{kl}}\left(P_{1,i}\|P_{3,i}\right) \leq |R_1(z)| \leq 105n\,\delta^2 \leq 105/N$ για κάθε $i \in \mathbb{N}$. □

4.3 Μάθηση των Παραμέτρων μιας PBD

Σε αυτήν την ενότητα θα μελετήσουμε το δύσκολο πρόβλημα της εύρεσης των παραμέτρων που ορίζει μία PBD. Έστω λοιπόν $S = \sum_{i=1}^{n} X_i$ όπου $\mathbb{E}\left[X_i\right] = p_i$ μία PBD. Αρχικά πρέπει να εξετάσουμε αν κάθε διάνυσμα $\boldsymbol{p}$ ορίζει με μοναδικό τρόπο μία PBD. Κατ' αρχάς είναι φανερό ότι η σειρά των p_i στο διάνυσμα $\boldsymbol{p}$ δεν παίζει ρόλο αφού η σειρά που αθροίζουμε τις δοκιμές Bernoulli για να πάρουμε την PBD δεν παίζει ρόλο.Δηλαδή οι PBDs είναι μοναδικές ως προς τις μεταθέσεις των παραμέτρων τους. Υποθέτουμε λοιπόν χωρίς βλάβη της γενικότητας ότι τα διανύσματα παραμέτρων των PBD θα είναι ταξινομημένα σε αύξουσα σειρά. Παρ' όλα αυτά δεν είναι ακόμα ξεκάθαρο αν κάθε PBD ορίζει ένα μοναδικό διάνυσμα παραμέτρων ή μία PBD μπορεί να περιγραφτεί με περισσότερα από ένα διανύσματα πιθανοτήτων. Την απάντηση σε αυτό το ερώτημα δίνει το επόμενο λήμμα του οποίου η απόδειξη μπορεί να βρεθεί στο [12].

Λήμμα 7 (Λήμμα 1 [12])**.** Έστω $= \sum_{i=1}^{n} X_i$, $= \sum_{i=1}^{n} Y_i$ είναι n-PBDs με ταξινομημένα σε αύξουσα σειρά διανύσματα $\boldsymbol{p}$, $\boldsymbol{q}$ αντιστοίχως. Τότε $\mathbb{P}\left[X = i\right] = \mathbb{P}\left[Y = i\right]$ για κάθε $i \in \{0, \ldots, n\}$ αν και μόνο αν $\boldsymbol{p} = \boldsymbol{q}$.

Απόδειξη. Υποθέτοντας ότι $\boldsymbol{p} = \boldsymbol{q}$ παίρνουμε ότι οι κατανομές των X και Y ταυτίζονται άρα αυτή η κατεύθυνση είναι προφανής. Για την άλλη κατεύθυνση θεωρούμε τα πολυώνυμα $p(s) = \mathbb{E}\left[(1+s)^X\right]$, $q(s) = \mathbb{E}\left[(1+s)^Y\right]$. Αφού οι κατανομές X, Y έχουν ίδιες πυκνότητες τα πολυώνυμα αυτά έχουν παίρνουν την ίδια τιμή για κάθε $s \in \mathbb{R}$. Συνεπώς είναι ίσα και αυτό σημαίνει ότι έχουν ίδιο βαθμό και ίδιες ρίζες. Ας βρούμε τώρα τις ρίζες τους. Έχουμε

$$p(s) = \mathbb{E}\left[(1+s)^{\sum_{i=1}^{n} X_i}\right] = \mathbb{E}\left[\prod_{i=1}^{n}(1+s)^{X_i}\right] = \prod_{i=1}^{n}\mathbb{E}\left[(1+s)^{X_i}\right],$$

όπου χρησιμοποιήσαμε το γεγονός ότι οι X_i είναι ανεξάρτητες άρα και οι $(1+s)^{X_i}$ είναι ανεξάρτητες ως συναρτήσεις ανεξάρτητων τυχαίων μεταβλητών. Συνεχίζοντας, έχουμε $\mathbb{E}\left[(1+s)^{X_i}\right] = (1+s)^1 p_i + (1+s)^0(1-p_i) = 1+sp_i$, άρα $p(s) = \prod_{i=1}^{n}(1+sp_i)$ και κατά συνέπεια έχει ρίζες τις $-1/p_1, \ldots -1/p_n$. Αντίστοιχα το πολυώνυμο $q(s)$ έχει ρίζες τις $-1/q_1 \ldots -1/q_n$. Γνωρίζοντας ότι τα $\boldsymbol{p}$ και $\boldsymbol{q}$ είναι ταξινομημένα σε αύξουσα σειρά παίρνουμε ότι οι δύο ακολουθίες των ριζών είναι και αυτές σε αύξουσα σειρά και επειδή γνωρίζουμε ότι πρέπει να ταυτίζονται λόγω της ισότητας των δύο πολυωνύμων έχουμε $p_i = q_i$ για κάθε $i \in [n]$. □

Οι Διακονικόλας, Kane και Stewart δείξανε ότι το πρόβλημα της μάθησης των παραμέτρων είναι δύσκολο και χρειάζονται εκθετικά ως προς την ακρίβεια ε δείγματα για να μάθουμε το διάνυσμα παραμέτρων μιας PBD. Το ενδιαφέρον ερώτημα που απαντάμε σε αυτήν την ενότητα είναι αν έχοντας στην διάθεσή μας δείγματα από όλο το φάσμα των δυνάμεων μιας PBD και όχι μόνο από την πρώτη δύναμη, μπορούμε να πετύχουμε κάτι καλύτερο από πλευράς δειγματικής πολυπλοκότητας; Δίνουμε αρνητική απάντηση σε αυτό το ερώτημα επεκτείνοντας το κάτω φράγμα που έδωσαν οι Διακονικόλας, Kane, και Stewart στην Πρόταση 15 του [14]. Για να γενικεύσουμε αυτό το κάτω φράγμα στην περίπτωση που τα δείγματα προέρχονται εν δυνάμει από κάθε δύναμη θα χρησιμοποιήσουμε την γενίκευση της μεθόδου του Le Cam που αποδείξαμε στο προηγούμενο κεφάλαιο και το instance που χρησιμοποιήθηκε στο [14]. Σε αυτό το σημείο πρέπει να διαχωρίσουμε τις περιπτώσεις της μάθησης των παραμέτρων

Αφού δείξουμε την δυσκολία του προβλήματος με το ισχυρό κάτω φράγμα ακόμα και στο μοντέλο των δυνάμεων προχωράμε δίνοντας ένα πολύ κοντινό πάνω φράγμα όπου χρησιμοποιούμε το πλεονέκτημα ότι μπορούμε να τραβάμε δείγματα από τις δυνάμεις μιας PBD. Δείχνουμε πως χρησιμοποιώντας τις ταυτότητες του Newton μπορεί κανείς από τα δείγματα να πάρει προσεγγίσεις των συντελεστών του πολυωνύμου του οποίου οι ρίζες είναι οι παράμετροι της άγνωστης PBD. Η εύρεση ριζών ενός πολυωνύμου βαθμού n όταν το n είναι είναι μεγάλο είναι γνωστό ill-conditioned πρόβλημα. Οι ρίζες δηλαδή των πολυωνύμων είναι εξαρτώνται με ιδιαίτερα ευαίσθητο τρόπο από τους συντελεστές του και μικρά σφάλματα στους συντελεστές μπορούν να καταστρέψουν τελείως τον υπολογισμό των ριζών. Ο Wilkinson ήταν από τους πρώτους μαθηματικούς που προσπάθησαν να υπολογίσουν αριθμητικά τις ρίζες ενός πολυωνύμου. Θέλοντας να δοκιμάσει ένα πρόγραμμα που υπολογισμού ριζών που είχε γράψει δοκίμασε να βρει τις ρίζες του πολυωνύμου $(x-1)(x-2)\ldots(x-20)$,του οποίου οι ρίζες είναι φαίνονται καλά διαχωρισμένες και εύκολες στον ακριβή προσδιορισμό. Προς έκπληξή του το σφάλμα σε κάποιες από τις ρίζες ήταν μεγάλο και αυτό ήταν ίσως η πρώτη ένδειξη ότι κάποια παραδοσιακά "καλόβολα" μαθηματικά αντικείμενα όπως τα πολυώνυμα μπορεί να έχουν απρόσμενη συμπεριφορά στην πράξη. Για περισσότερα σχετικά με το διάσημο ύπουλο πολυώνυμο του Wilkinson δείτε το [31]. Αξίζει να σημειωθεί πάντως ότι το πολυώνυμο αυτό δεν έχει κάτι το ιδιαίτερο και το πρόβλημα της εύρεσης ριζών είναι ευαίσθητο στον θόρυβο για τα περισσότερα πολυώνυμα μεγάλου βαθμού. Κατά συνέπεια πρέπει να τραβήξουμε εκθετικά σε πλήθος δείγματα ώστε να διασφαλίσουμε ότι αφού λύσουμε το σύστημα των ταυτοτήτων του Newton οι συντελεστές που θα προκύψουν θα έχουν την απαραίτητη ακρίβεια ώστε να δώσουν καλές προσεγγίσεις των ριζών.

4.3.1 Κάτω Φράγμα

Για να δείξουμε το κάτω φράγμα στο πρόβλημα της εκτίμησης του διανύσματος των παραμέτρων μιας PBD θα χρησιμοποιήσουμε την μέθοδο του Le Cam που παρουσιάσαμε στο Κεφάλαιο 3 η οποία ενδείκνυται για την κατασκευή κάτω φραγμάτων σε παραμετρικά προβλήματα εκτίμησης.

Έχουμε ήδη δει στην Ενότητα **??** ότι το πρόβλημα προσέγγισης των παραμέτρων μιας PBD δυσκολεύει όσο το πλήθος των διαφορετικών παραμέτρων γίνεται μεγαλύτερο. Για την ακρίβεια είδαμε ότι είναι πολύ εύκολο στην περίπτωση της προσέγγισης του p διωνυμικών κατανομών και αρκετά πιο δύσκολο στην περίπτωση που προσθέσουμε ακόμα και μία διαφορετική παράμετρο στο άθροισμα, δηλαδή θεωρήσουμε κατανομές της μορφής $B(n_1, p) + B(n_2, q)$. Η ουσία της δυσκολίας του προβλήματος βρίσκεται ακριβώς στο γεγονός ότι ακριβώς όπως στην περίπτωση των δύο παραμέτρων μπορέσαμε να ταιριάξουμε απολύτως την μέση τιμή των δύο κατανομών ενώ οι παράμετροι ήταν αρκετά μακριά έτσι στην γενική περίπτωση των PBD θα κατασκευάσουμε δύο διανύσματα παραμέτρων τα οποία θα είναι αρκετά διαχωρισμένα αλλά θα καταφέρουμε να ταιριάξουμε τις επόμενες n ροπές και όχι μόνο την μέση τιμή κάτι που θα έχει ως αποτέλεσμα οι PBD που θα προκύψουν να είναι σε απόσταση ολικής κύμανσης 2^{-n} δίνοντάς μας το εκθετικό κάτω φράγμα.

Διατηρώντας τον γνωστό συμβολισμό για κατανομές, ακολουθίες, και οικογένειες η συνάρτηση μιας ακολουθίας $\mathcal{P}$ δυνάμεων μιας PBD που προσπαθούμε να εκτιμήσουμε είναι τώρα η $\theta(\mathcal{P}) = \boldsymbol{p} \in (0,1)^n$. Η μετρική

μας στον χώρο της διανυσμάτων $(0,1)^n$ θα είναι η $\rho(p,\hat{p}) = \|p-\hat{p}\|_\infty$ αφού θέλουμε να προσεγγίζουμε το διάνυσμα $\boldsymbol{p}$ των παραμέτρων με αθροιστικό σφάλμα *το πολύ* ε. Ακολουθεί η επέκταση του κάτω φράγματος που δόθηκε στην Πρόταση 15 στο [14] στην περίπτωση που ο αλγόριθμος μάθησης επιτρέπει να τραβήξει δείγματα τόσο από την αρχική PBD όσο και από οποιαδήποτε δύναμή της.

Θεώρημα 8. Αν $n \geq 1/\varepsilon$, τότε κάθε αλγόριθμος μάθησης που τραβά N δείγματα από τις δυνάμεις μίας n-PBD και επιστρέφει προσεγγίσεις των παραμέτρων αυτής της PBD με αθροιστικό σφάλμα το πολύ ε και πιθανότητα επιτυχίας τουλάχιστον $2/3$ πρέπει να έχει $N = 2^{\Omega(1/\varepsilon)}$.

Απόδειξη. Θέτουμε το μήκος n του διανύσματος της PBD να είναι $n = \Theta(\log(N/\varepsilon))$ όπου το N αναπαριστά τον αριθμό των δειγμάτων στον ορισμό του minimax risk. Έχουμε τώρα τα εξής διανύσματα $p_j := (1+\cos(\frac{2\pi j}{n}))/8$, $q_j := (1+\cos(\frac{2\pi j+\pi}{n}))/8$, $j \in [n]$. Έτσι για $j = n/4 + O(1)$ έχουμε $|q_i - p_j| = \Omega(1/\log(N/\varepsilon))$ αφού για όλα τα i, $\frac{2\pi i+\pi}{n}$ είναι σε απόσταση το λιγότερο $\Omega(1/\log(N/\varepsilon))$ από το $\frac{2\pi i}{n}$ και το $\frac{2\pi(n-j)}{n}$.

Παρατηρούμε ότι τα $p_1,\ldots,p_n$ και αντίστοιχα τα $q_1,\ldots,q_n$ είναι οι ρίζες των Chebyshev πολυωνύμων $(T_n(8x-1)-1)$, $(T_n(8x-1)+1)$ αντίστοιχα όπου το T_n είναι το n-οστό πολυώνυμο Chebyshev. Αφού αυτά τα πολυώνυμα συμφωνούν σε όλες τους συντελεστές εκτός από τον σταθερό τους όρο οι ταυτότητας Newton-Girard δίνουν ότι $\sum_{i=1}^n p_i^l = \sum_{i=1}^n q_i^l$ για κάθε $l \in \{1,2,\ldots,n-1\}$ και επιπλέον, για $l \geq n$ είναι εύκολο να δούμε ότι $3^l(\sum_{i=1}^n (p_i^l - q_i^l)) \leq n(3/4)^n = \log(N/\varepsilon)(3/4)^{\log(N/\varepsilon)}$. Για αρκετά μικρά ε χρησιμοποιώντας το Λήμμα 9 του [14] έχουμε ότι $d_{\mathrm{tv}}(P_1,Q_1) \leq c/N$ για κάποια σταθερά c. Θα δείξουμε ότι αυτό ισχύει για όλες τις δυνάμεις των P και Q δηλαδή θα φράξουμε την απόσταση ολικής κύμανσης $d_{\mathrm{tv}}(\mathcal{P},\mathcal{Q})$. Για να το δείξουμε αυτό θεωρούμε κάποια δύναμη $s \in \{1,2,\ldots,n\}$. Τότε, έχουμε ότι $\sum_{i=1}^n p_i^{sl} = \sum_{i=1}^n q_i^{sl}$, για κάθε $l = 1,2,\ldots,\lfloor (n-1)/s \rfloor$ υποθέτοντας ότι $s \leq n-1$. Επιπλέον, όταν $l \in \{\lfloor (n-1)/s \rfloor + 1, \lfloor (n-1)/s \rfloor + 2, \ldots\}$, έχουμε $3^l(\sum_{i=1}^n (p_i^{sl} - q_i^{sl})) \leq n\frac{3^l}{4^{sl}} \leq n\frac{3^{sl}}{4^{sl}} \leq n(\frac{3}{4})^n$, όπου η τελευταία ανισότητα ισχύει αφού $sl \geq n$. Είναι εύκολο να δούμε ότι αυτό ισχύει όταν $s = n$, και χρησιμοποιώντας ξανά το Λήμμα Lemma 9 του [14], παίρνουμε ότι $d_{\mathrm{tv}}(P_s,Q_s) \leq c/N$. Αφού οι παράμετροι των P,Q είναι σε απόσταση $\Omega(1/\log(N/\varepsilon))$ η απόσταση ολικής κύμανσης των δύο ακολουθιών είναι μικρότερη από c/N. Χρησιμοποιώντας τώρα το Λήμμα 3 παίρνουμε ως ένα κάτω φράγμα για τον ρυθμό minimax το $1/\log(\varepsilon/N)$. Έτσι, αφού πρέπει να προσεγγίσουμε τις παραμέτρους σε αθροιστικό σφάλμα μικρότερο του ε, θέλουμε $\mathfrak{M}_N < \varepsilon$ οπότε παίρνουμε ότι ο αριθμός των δειγμάτων N πρέπει να είναι $\Omega(2^{1/\varepsilon})$. □

4.3.2 Πάνω Φράγμα

Newton's identities, a.k.a the Newton-Girard formulae, give relations between power sums and elementary symmetric polynomials of variables $x_1,...,x_n$. In that setting, the kth power sum is $s_k(x_1,\ldots,x_n) = x_1^k + \cdots + x_n^k$. The kth elementary symmetric polynomial $e_k(x_1,\ldots,x_n)$ is the sum of all distinct products of k distinct variables. Newton's identities allow us to compute the elementary symmetric polynomials if we know the power sums *exactly*. Moreover, the polynomial with roots x_i, i.e., $\prod_{i=1}^n (x-x_i)$, may be expanded as $\sum_{k=0}^n (-1)^{n+k} e_{n-k} x^k$. Thus, if we know the power sums $s_1(x_1,\ldots,x_n),\ldots,s_n(x_1,\ldots,x_n)$ exactly, we can first find the coefficients of the elementary symmetric polynomials and then compute the roots $x_1,\ldots,x_n$ with an arbitrarily good accuracy. A similar approach was used in [12] to derive sparse covers for PBDs.

In this section we provide the analysis of the "noisy" version of Newton's identities. Given query access to PBD powers, we can obtain good estimations of the power sums $s_k(p_1,\ldots,p_n)$ using a reasonable number of samples, since the expectations of PBD powers are the power sums of the unknown probabilities $p_1,\ldots,p_n$. An intriguing question is to which extent these "noisy" power sum estimations can be used to recover the actual values of $p_1,\ldots,p_n$ within sufficiently good accuracy. In this Section we answer this question by providing an upper bound on the sampling complexity of estimating the parameters of a PBD using samples from its powers. This upper bound matches the corresponding lower bound of Theorem 8.

Ο Αλγόριθμος 2 που θα παρουσιάσουμε σε αυτήν την ενότητα χωρίζεται σε τρία μέρη: προσέγγιση των μέσων τιμών των δυνάμεων μιας PBD, λύση ενός γραμμικού συστήματος, και εύρεση των ριζών ενός πολυωνύμου. Παρουσιάζουμε πρώτα τα βασικά εργαλεία που θα χρησιμοποιήσουμε σε καθένα από τα τρία αυτά βήματα.

Ξεκινάμε από το πρόβλημα της προσέγγισης της μέσης τιμής μιας PBD. Εδώ τα πράγματα είναι απλά. Στην επόμενη Πρόταση δείχνουμε πως μπορούμε να πάρουμε μία ε-προσέγγιση της μέσης τιμής μιας PBD χρησιμοποιώντας $O((\log(1/\delta)/\varepsilon^2)$ για πιθανότητα επιτυχίας $1-\delta$. Βασικά η επόμενη πρόταση δείχνει ότι οι φυσική επιλογή της εκτιμήτριας $\hat{\mu} = \sum_{i=1}^{N} X_i/N$ για την μέση τιμή και της $\sum_{i=1}^{N}(X_i-\mu)^2/(m-1)$ για την διακύμανση δουλεύουν καλά στην περίπτωση που τα δείγματα έρχονται από μια PBD.

Πρόταση 21 (Λήμμα 6, [11])**.** Για κάθε $n,\ \varepsilon, \delta > 0$, υπάρχει ένας αλγόριθμος $\mathcal{A}(n, \varepsilon, \delta)$ με τις εξής ιδιότητες: χρησιμοποιώντας δείγματα από μιας PBD X τάξης n, παράγει εκτιμήσεις $\hat{\mu}$, $\hat{\sigma}^2$ για την μέση τιμή $\mu = \mathbb{E}[X]$, και την διακύμανση $\sigma^2 = \mathrm{Var}[X]$ αντίστοιχα έτσι ώστε με πιθανότητα τουλάχιστον $1-\delta$ να ισχύουν

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

Επιπλέον, ο αλγόριθμος $\mathcal{A}$ χρησιμοποιεί $O(\log(1/\delta)/\varepsilon^2)$ δείγματα και τρέχει σε χρόνο $O(\log n \log(1/\delta)/\varepsilon^2)$.

Απόδειξη. Αρκεί να δείξουμε ότι υπάρχουν εκτιμήτριες που δίνουν τις ζητούμενες προσεγγίσεις με σταθερή πιθανότητα επιτυχίας $2/3$ αφού, χρησιμοποιώντας το median trick (5) μπορούμε να αυξήσουμε την πιθανότητα επιτυχίας σε $1-\delta$ με κόστος έναν παράγοντα $\log(1/\delta)$ στην δειγματική πολυπλοκότητα. Ξεκινάμε με την περίπτωση της εκτίμησης της μέσης τιμής. Έστω $X_1, \ldots, X_N$ δείγματα από την PBD X και έστω η εκτιμήτρια $\mu = \sum_{i=1}^{N} X_i$ της μέσης τιμής της X. Εύκολα βλέπουμε ότι $\mathbb{E}[\hat{\mu}] = \mu$ και

$$\mathrm{Var}[\hat{\mu}] = \frac{1}{N}\mathrm{Var}[X] = \frac{\sigma^2}{N}.$$

Συνεπώς χρησιμοποιώντας την ανισότητα του Chebyshev παίρνουμε ότι

$$\mathbb{P}\left[|\hat{\mu} - \mu| \geq \frac{t\sigma}{\sqrt{N}}\right] \leq \frac{1}{t^2},$$

όποτε επιλέγοντας $t = \sqrt{3}$, και $m = \lceil 3/\varepsilon^2 \rceil$ η παραπάνω ανισότητα συνεπάγεται ότι $|\hat{\mu} - \mu| \leq \varepsilon\sigma$ με πιθανότητα τουλάχιστον $2/3$.

Συνεχίζουμε δείχνοντας ότι η εκτιμήτρια για την διακύμανση,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}\left(X_i - \frac{1}{N}\sum_{i=1}^{N} X_i\right)^2}{m-1}$$

μπορεί να δώσει καλή εκτίμηση για την διακύμανση με σταθερή πιθανότητα. Αρχικά παρατηρούμε ότι διαιρώντας με $n-1$ αντί για n η εκτιμήτρια της διακύμανσης είναι αμερόληπτη (unbiased). Το τέχνασμα αυτό ώστε η εκτιμήτρια να είναι αμερόληπτη είναι γνωστό σαν διόρθωση Bessel (Bessel's correction). Συνεπώς έχουμε ότι $\mathbb{E}\left[\hat{\sigma}^2\right] = \sigma^2$. Για να χρησιμοποιήσουμε την ανισότητα του Chebyshev πρέπει τώρα να υπολογίσουμε και την διακύμανση αυτής της εκτιμήτριας. Έχουμε ότι

$$\mathrm{Var}\left[\hat{\sigma}^2\right] = \sigma^4\left(\frac{2}{N-1} + \frac{k}{N}\right)$$

όπου k είναι η υπερβάλλουσα κύρτωση (excess kurtosis) της κατανομής , δηλαδή $k = \mathrm{Kurt}[X] - 3 = \frac{\mathbb{E}\left[(X-\mu)^4\right]}{\sigma^4} - 3$. Για την κύρτωση αθροίσματος ανεξάρτητων τυχαίων μεταβλητών ισχύει

$$\mathrm{Kurt}\left[\sum_{i=1}^{n} X_i\right] - 3 = \frac{1}{(\sum_{i=1}^{n}\mathrm{Var}[X_i])^2}\sum_{i=1}^{n}\mathrm{Var}[X_i]^2\left(\mathrm{Kurt}[X_i] - 3\right),$$

Υποθέτοντας ότι το διάνυσμα πιθανοτήτων της PBD είναι το $\boldsymbol{p}$ από παραπάνω τύπο μπορούμε να υπολογίσουμε την υπερβάλλουσα κύρτωση της PBD X

$$\begin{aligned} k = \text{Kurt}\left[X\right] - 3 &= \frac{1}{\sigma^4}\sum_{i=1}^{n}((1-p_i)p_i)^2\left(\frac{1}{1-p_i}+\frac{1}{p}_i - 6\right) \\ &= \frac{\sum_{i=1}^{n}((1-p_i)p_i)^2\left(\frac{1}{1-p_i}+\frac{1}{p}_i - 6\right)}{\sigma^4} \\ &= \frac{\sum_{i=1}^{n}p_i(1-p_i)(1-6(1-p_i)p_i)}{\sigma^4}. \\ &\leq \frac{\sum_{i=1}^{n}(1-p_i)p_i}{\sigma^4} \\ &= \frac{1}{\sigma^2} \end{aligned}$$

Χρησιμοποιώντας το παραπάνω άνω φράγμα για την υπερβάλλουσα κύρτωση της X έχουμε ότι η διακύμανση της εκτιμήτριας $\hat{\sigma}^2$ φράσσεται ως εξής

$$\text{Var}\left[\hat{\sigma}^2\right] \leq \frac{\sigma^4}{N}\left(4+\frac{1}{\sigma^2}\right).$$

Μπορούμε τώρα να χρησιμοποιήσουμε την ανισότητα του Chebyshev για να πάρουμε

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

Συνεπώς, επιλέγοντας $t=\sqrt{3}$ και $N=\lceil 3/\varepsilon^2\rceil$ έχουμε ότι $|\hat{\sigma}^2-\sigma^2| \leq \varepsilon^2\sqrt{4+\frac{1}{\sigma^2}}$ με πιθανότητα τουλάχιστον $2/3$. □

Έστω $\mathbf{x}\in\mathbb{R}^n$ ένα διάνυσμα και έστω $\mathbf{A}=(\boldsymbol{A}_{ij})_{i,j\in[n]}$ ένας $n\times n$ πίνακας. Τότε $\|\boldsymbol{x}\|_\infty = \max_{i\in[n]}|\boldsymbol{x}_i|$, $\|\boldsymbol{A}\|_\infty = \max_{i\in[n]}\sum_{j=1}a_{ij}$, $|\boldsymbol{x}| = (|\boldsymbol{x}_i|)_{i\in[n]}$, $|\boldsymbol{A}| = (|\boldsymbol{A}_{ij}|)_{i,j\in[n]}$. Θα χρησιμοποιούμε το "$\leq$" στις εκφράσεις $\boldsymbol{A}\leq\boldsymbol{B}$ για να συμβολίζουμε την ανά-στοιχείο ανισότητα των πινάκων $\boldsymbol{A}$, $\boldsymbol{B}$, δηλαδή $\boldsymbol{A}\leq\boldsymbol{B}\Leftrightarrow A_{ij}\leq B_{ij}$ για κάθε $i,j\in[n]$.

Για υπολογίσουμε την ευαισθησία της λύσης ενός γραμμικού συστήματος $\boldsymbol{Ax}=\boldsymbol{b}$ σε διαταραχές της εισόδου $\boldsymbol{A},\boldsymbol{b}$ θα χρησιμοποιήσουμε το επόμενο θεώρημα από το [20].

Θεώρημα 9 (Θεώρημα 7.4, [20]). Έστω $\boldsymbol{Ax}=\boldsymbol{b}$ και $(\boldsymbol{A}+\Delta\boldsymbol{A})\boldsymbol{y}=\boldsymbol{b}+\Delta\boldsymbol{b}$, όπου $|\Delta\boldsymbol{A}|\leq u\,\boldsymbol{E}$ and $|\Delta\boldsymbol{b}|\leq u\,\boldsymbol{f}$, και υποθέτουμε ότι u $\||\boldsymbol{A}^{-1}|\,\boldsymbol{E}\|<1$, όπου $\|\cdot\|$ είναι μια απόλυτη νόρμα. Τότε

$$\frac{\|\boldsymbol{x}-\boldsymbol{y}\|}{\|\boldsymbol{x}\|} \leq \frac{u}{1-u\,\||\boldsymbol{A}^{-1}|\,\boldsymbol{E}\|}\frac{\|\,|\boldsymbol{A}^{-1}|\,(\boldsymbol{E}|\boldsymbol{x}|+\boldsymbol{f})\,\|}{\|\boldsymbol{x}\|}$$

και για την ∞-νόρμα αυτό το φράγμα είναι εφικτό σε πρώτη τάξη ως προς u.

Για να προσεγγίσουμε τις ρίζες ενός πολυωνύμου μίας μεταβλητής $P(x)$ θα χρησιμοποιήσουμε τον σχεδόν βέλτιστο αλγόριθμο εύρεσης ριζών του Pan [24](Theorem 2.1.1).

Θεώρημα 10 (Θεώρημα 2.1.1, [24]). Έστω $P(x)=\sum_{i=0}^{n}c_ix^i=c_n\prod_{i=1}^{n}(x-p_i)$, $c_n\neq 0$, ένα πολυώνυμο βαθμού n τέτοιο ώστε για όλες τις μιγαδικές ρίζες του ισχύει ότι $|p_j|\leq 1$. Έστω b ένας σταθερός πραγματικός αριθμός τέτοιος ώστε Τότε μπορούν να υπολογιστούν μιγαδικοί αριθμοί $\hat{p}_j$ χρησιμοποιώντας $O\left((n\log^2 n)(\log^2 n+\log b)\right)$ αριθμητικές πράξεις εκτελεσμένες με ακρίβεια $O\left(b\right)$ bits ώστε για τις προσεγγίσεις $\hat{p}_j$ να ισχύει $|\hat{p}_j-p_j|<2^{2-b/n}$ για κάθε $j=1,\ldots,n$.

Η επόμενη χρήσιμη πρόταση μας δίνει ένα πάνω φράγμα για το μέγεθος των συντελεστών ενός πολυωνύμου δεδομένου ότι οι ρίζες του είναι πιθανότητες, δηλαδή ανήκουν στο $[0,1]$.

Πρόταση 22. Αν όλες οι ρίζες από ένα μονικό πολυώνυμο $P = x^n + a_{n-1}x^{n-1} + ... + a_0$ βαθμού n βρίσκονται στο διάστημα $[-1,1]$ τότε $|a_k| \leq \binom{n}{k} \leq 2^n$.

Απόδειξη. Χρησιμοποιώντας τους τύπους του Vieta έχουμε

$$a_{n-k} = (-1)^k \sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq n} x_{i_1} x_{i_2} \ldots x_{i_k}$$

Συνεπώς, ο συντελεστής $|a_{n-k}|$ είναι μέγιστος όταν όλες οι ρίζες x_i είναι 1 και επομένως $|a_{n-k}| \leq \binom{n}{k} = \binom{n}{n-k}$. □

Συμβολίζουμε με P_j την j-οστή δύναμη της PBD με διάνυσμα πιθανοτήτων $\boldsymbol{p}$, δηλαδή η P_j είναι η PBD με διάνυσμα πιθανοτήτων $\boldsymbol{p}^j = (p_i^j)_{i=1}^n$. Έστω $P(x) = x^n + c_{n-1}x^{n-1} + \ldots + c_0 = \prod_{i=1}^n (x - p_i)$ το μονικό πολυώνυμο βαθμού n του οποίου οι ρίζες είναι τα p_i. Παρατηρούμε ότι η μέση τιμή της κατανομής P_j, την οποία θα συμβολίζουμε με μ_j ισούται με το j-οστό άθροισμα Newton των ριζών, αφού $\mu_j = \sum_{i=1}^n p_i^j$. Δεδομένου ότι το $P(x)$ είναι μονικό οι συντελεστές του, δεδομένων των μ_j, δίνονται από τις ταυτότητες Newton-Girard. Τα $\mu_1, \ldots, \mu_n$ ικανοποιούν λοιπόν το παρακάτω γραμμικό σύστημα.

$$\mu_j + \sum_{i=1}^{j-1} c_{n-i}\mu_{j-i} + jc_{n-j} = 0, \quad j = 1, 2, \ldots, n$$

Σε μορφή πίνακα φαίνεται η πολύ φιλική κάτω τριγωνική μορφή του συστήματος.

$$\begin{pmatrix} 1 & & & & \\ \mu_1 & 2 & & & \\ \mu_2 & \mu_1 & 3 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \mu_{n-1} & \mu_{n-2} & \ldots & \mu_1 & n \end{pmatrix} \begin{pmatrix} c_{n-1} \\ c_{n-2} \\ c_{n-3} \\ \vdots \\ c_0 \end{pmatrix} = \begin{pmatrix} -\mu_1 \\ -\mu_2 \\ -\mu_3 \\ \vdots \\ -\mu_n \end{pmatrix} \Leftrightarrow \boldsymbol{Ac} = \boldsymbol{b} \tag{4.2}$$

Χρησιμοποιώντας λοιπόν το σύστημα αυτό μπορούμε από τις προσεγγίσεις των μέσων τιμών των πρώτων n δυνάμεων να περάσουμε σε προσεγγίσεις των συντελεστών του πολυωνύμου $P(x)$ και στην συνέχεια να χρησιμοποιήσουμε τον αλγόριθμο του Pan για να βρούμε τις ρίζες. Για να "γεμίσουμε" όμως τους πίνακες $\boldsymbol{A}$ και $\boldsymbol{b}$ του παραπάνω συστήματος χρειαζόμαστε τις μέσες τιμές μ_j οπότε χρειαζόμαστε μια εκτιμήτρια της μέσης τιμής από δείγματα μιας PBD.

Algorithm 2 Εκτίμηση Παραμέτρων

Είσοδος: $2^{O(n \max(\log(1/\varepsilon), \log(n))}$ δείγματα από τις δυνάμεις P_j, $j \in [n]$.
Output: Ένα διάνυσμα πιθανοτήτων που προσεγγίζει σε αθροιστικό σφάλμα ε το $\boldsymbol{p}$.

1: Χρησιμοποιώντας τον αλγόριθμο $\mathcal{A}$ του Λήμματος **??** τραβάμε $2^{O(n \max(\log(1/\varepsilon), \log n))}$ δείγματα από κάθε δύναμη P_j για να πάρουμε τις προσεγγίσεις $\hat{\mu}_j$ των μ_j.
2: Λύνουμε το σύστημα 4.2 και παίρνουμε την λύση $\hat{\boldsymbol{c}}$.
3: Χρησιμοποιώντας τον αλγόριθμο του Pan του Θεωρήματος 10 υπολογίζουμε προσεγγίσεις $\hat{\boldsymbol{p}}_j$ σε όλες τις ρίζες του πολυωνύμου $P(x) = \sum_{i=1}^n c_i x^i$.
4: **return** $\hat{\boldsymbol{p}}$.

Θεώρημα 11. Έστω X μία n-PBD με διάνυσμα πιθανοτήτων $\boldsymbol{p}$. Υπάρχει ένας αλγόριθμος που τραβά $2^{O(n \max(\log(1/\varepsilon), \log n))}$ δείγματα από τις δυνάμεις της X και υπολογίζει ένα διάνυσμα $\hat{\boldsymbol{p}}$ τέτοιο ώστε $\|\boldsymbol{p} - \hat{\boldsymbol{p}}\|_\infty \leq \varepsilon$.

Απόδειξη. Ξεκινώντας από το τελευταίο βήμα του αλγορίθμου 2, δηλαδή την χρήση του αλγορίθμου του Pan για την εύρεση των ριζών, από το Θεώρημα 10, έχουμε ότι για να πάρουμε ε-προσεγγίσεις των ριζών του πολυωνύμου $P(x)$ χρειάζεται να έχουμε ένα διάνυσμα $\hat{\boldsymbol{c}}$ με προσεγγίσεις των συντελεστών του $P(x)$ τέτοιο ώστε

$$\|\boldsymbol{c} - \hat{\boldsymbol{c}}\|_\infty = 2^{O(-n\max(\log(1/\varepsilon),\log(n))}. \tag{4.3}$$

Στη συνέχεια, προχωράμε στον υπολογισμό της ακρίβειας που χρειαζόμαστε στην είσοδο του συστήματος των ταυτοτήτων του Newton (4.2) ώστε η λύση του να έχει την ζητούμενη ακρίβεια (4.3). Αφού στο δικό μας πρόβλημα το σφάλμα της προσέγγισης της j-οστής μέσης τιμής είναι ανάλογο με την τυπική απόκλιση της j-οστής δύναμης P_j, οι πίνακες σφαλμάτων $\boldsymbol{E}$ $\boldsymbol{f}$ του Λήμματος 9 είναι

$$\boldsymbol{E} = \begin{pmatrix} \sigma_1 & & & \\ \sigma_2 & \sigma_1 & & \\ \vdots & \vdots & \ddots & \\ \sigma_n & \sigma_2 & \dots & \sigma_1 \end{pmatrix} \leq \sqrt{n} \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & \vdots & \ddots & \\ 1 & 1 & \dots & 1 \end{pmatrix} \qquad \boldsymbol{f} = \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \vdots \\ \sigma_n \end{pmatrix} \leq \sqrt{n} \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Αφού ισχύει ότι $\mu_j \leq n$ έχουμε ότι $\boldsymbol{A}_{ij} \leq n$. Επίσης παρατηρώντας ότι ο $\boldsymbol{A}$ είναι κάτω τριγωνικός, έχουμε ότι $\det(\boldsymbol{A}) = n!$, και ισχύει ότι $\det(\boldsymbol{A}) \geq M_{ij}$, όπου M_{ij} είναι η ορίζουμε του $(n-1)\times(n-1)$ υποπίνακα του $\boldsymbol{A}$ μετά την διαγραφή της γραμμής i και της στήλης j. Συνεπώς προκύπτει ότι $|\boldsymbol{A}^{-1}|_{ij} \leq 1$. Επιπλέον, αφού το διάνυσμα λύσης $\boldsymbol{c}$ αντιστοιχεί στους συντελεστές του πολυωνύμου $P(x)$ μπορούμε να πάρουμε ένα πάνω φράγμα για αυτό χρησιμοποιώντας την Πρόταση 22. Έχουμε λοιπόν ότι $|x|_i \leq \binom{n}{n-i}$. Χρησιμοποιώντας τις παραπάνω ανισότητες φράσσουμε τώρα τον βαθμό κατάστασης του $\boldsymbol{A}$.

$$|\boldsymbol{A}^{-1}||\boldsymbol{E}| \leq \sqrt{n} \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & \vdots & \ddots & \\ 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & \vdots & \ddots & \\ 1 & 1 & \dots & 1 \end{pmatrix} = \sqrt{n} \begin{pmatrix} 1 & & & \\ 2 & 1 & & \\ \vdots & \vdots & \ddots & \\ n & n-1 & \dots & 1 \end{pmatrix}$$

Επιπλέον, $|\boldsymbol{A}^{-1}||\boldsymbol{f}| \leq (n\ 2n\ \dots\ n^2)^T$. Συνδυάζοντας τις παραπάνω ανισότητες μπορούμε πλέον να βρούμε ένα πάνω φράγμα για το βαθμό κατάστασης (condition number) του πίνακα $\boldsymbol{A}$.

$$\| |\boldsymbol{A}^{-1}|\,|\boldsymbol{A}|\,|\boldsymbol{c}| + |\boldsymbol{A}^{-1}|\,|\boldsymbol{b}| \|_\infty \leq \sqrt{n}\sum_{i=0}^{n}(n-i)\binom{n}{n-i} + \sqrt{n}n = n^{3/2}(2^{n-1}+1) = O\left(n^{3/2}2^n\right)$$
$$\||\boldsymbol{A}^{-1}||\boldsymbol{E}|\|_\infty = O\left(n^{5/2}\right)$$

Έτσι, από το Θεώρημα 9 παίρνουμε το επόμενο πάνω φράγμα για το λάθος στην ακρίβεια

$$\|\boldsymbol{c} - \hat{\boldsymbol{c}}\|_\infty \leq u\ O\left(n^{3/2}2^n\right).$$

Αφού χρειάζεται να τρέξουμε τον αλγόριθμο $\mathcal{A}$ της Πρότασης **??** n φορές για να πάρουμε τις προσεγγίσεις $\hat{\mu_j}$ τέτοιες ώστε $|\mu_j - \hat{\mu_j}| \leq u_j\sigma_j$ για κάθε $j \in [n]$, από το union bound έχουμε ότι με πιθανότητα τουλάχιστον $1-\delta$ πρέπει να τραβήξουμε $O\left(\log(1/n)n/u^2\right)$ από τις δυνάμεις P_j, $j \in [n]$. Κατά συνέπεια, αφού το $uO\left(n^{3/2}2^n\right)$ πρέπει να ικανοποιεί την (4.3) συμπεραίνουμε ότι συνολικά χρειαζόμαστε $2^{O(n\max(\log(1/\varepsilon),\log(n))}$ δείγματα από τις δυνάμεις της PBD. □

Βιβλιογραφία

[1] Ali, S. M., and Silvey, S. D. Association Between Random Variables and the Dispersion of a Radon-Nikodym Derivative. *Journal of the Royal Statistical Society. Series B (Methodological) 27*, 1 (1965), 100–107.

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

[3] Balcan, M., and Harvey, N. Learning submodular functions. In *Proc. of the 43rd ACM Symposium on Theory of Computing (STOC '11)* (2011), pp. 793–802.

[4] Berry, A. C. The Accuracy of the Gaussian Approximation to the Sum of Independent Variates. *Transactions of the American Mathematical Society 49*, 1 (1941), 122–136.

[5] Birgé, L. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 65*, 2 (Dec. 1983), 181–237.

[6] Cam, L. L. *Asymptotic Methods in Statistical Decision Theory*, 1986 edition ed. Springer, New York, Aug. 1986.

[7] Chu, J. On bounds for the normal integral. *Biometrika 42* (1955), 263–265.

[8] Cover, T. M., and Thomas, J. A. *Elements of Information Theory 2nd Edition*, 2 edition ed. Wiley-Interscience, Hoboken, N.J, July 2006.

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

[10] Daskalakis, C., Diakonikolas, I., O'Donnell, R., Servedio, R., and Tan, L. Learning sums of independent integer random variables. In *Proc. of the 54th IEEE Symposium on Foundations of Computer Science (FOCS '13)* (2013), pp. 217–226.

[11] Daskalakis, C., Diakonikolas, I., and Servedio, R. Learning Poisson Binomial Distributions. *Algorithmica 72*, 1 (2015), 316–357.

[12] Daskalakis, C., and Papadimitriou, C. Sparse Covers for Sums of Indicators. *Probability Theory and Related Fields 162*, 3 (2015), 679–705.

[13] Daskalakis, C., and Syrgkanis, V. Learning in Auctions: Regret is Hard, Envy is Easy. In *Proc. of the 57th IEEE Symposium on Foundations of Computer Science (FOCS '16)* (2016).

[14] Diakonikolas, I., Kane, D., and A.Stewart. Properly learning poisson binomial distributions in almost polynomial time. In *Proceedings of the 29th Conference on Learning Theory, (COLT'16)* (2016), pp. 850–878.

[15] Diakonikolas, I., Kane, D., and Stewart, A. Optimal learning via the fourier transform for sums of independent integer random variables. In *Proceedings of the 29th Conference on Learning Theory, (COLT'16)* (2016), pp. 831–849.

[16] Dubhashi, D., and Panconesi, A. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.

[17] Duchi, J. Stats311, Lecture Notes.

[18] Ehm, W. Binomial approximation to the Poisson binomial distribution. *Statistics & Probability Letters 11*, 1 (Jan. 1991), 7–16.

[19] Feldman, V., and Kothari, P. Learning coverage functions and private release of marginals. In *Proc. of the 27th Conference on Learning Theory (COLT 2014)* (2014), vol. 35 of *JMLR Proceedings*, pp. 679–702.

[20] Higham, N. J. *Accuracy and Stability of Numerical Algorithms*, 2nd edition ed. SIAM: Society for Industrial and Applied Mathematics, Philadelphia, Aug. 2002.

[21] Itô, K., and Mckean, H. P. j. *diffusion processes and their sample paths*, 1996 edition ed. springer, berlin ; new york, Feb. 1996.

[22] Liese, F., and Vajda, I. On Divergences and Informations in Statistics and Information Theory. *IEEE Transactions on Information Theory 52*, 10 (Oct. 2006), 4394–4412.

[23] Louis H.Y. Chen, Larry Goldstein, Q.-M. S. *Normal Approximation by Stein's Method*. Springer, 2010.

[24] Pan, V. Univariate polynomials: Nearly optimal algorithms for numerical factorization and rootfinding. *Journal of Symbolic Computation 33(5):701-733* (2002).

[25] Pearson, K. Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society of London. A 186* (1895), 343–414.

[26] Roos, B. Binomial Approximation to the Poisson Binomial Distribution: The Krawtchouk expansion. *Theory of Probability and its Applications 45(2)* (2000), 328–344.

[27] scharf, l. l. *statistical signal processing: detection, estimation, and time series analysis*, 1 edition ed. pearson, reading, mass, July 1991.

[28] Shevtsova, I. On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands. *arXiv:1111.6554 [math]* (Nov. 2011).

[29] Silvey, S. D. On a Measure of Association. *The Annals of Mathematical Statistics 35*, 3 (Sept. 1964), 1157–1166.

[30] Tsybakov, A. B. *Introduction to Nonparametric Estimation*, 1 edition ed. Springer, New York ; London, Nov. 2008.

[31] Wilkinson, J. H. The perfidious polynomial. *Studies in Numerical Analysis, ed. by G. H. Golub, pp. 1–28. (Studies in Mathematics, vol. 24).* (1984).

[32] Yang, Z.-H., and Chu, Y.-M. On approximating Mills ratio. *Journal of Inequalities and Applications 2015* (Sept. 2015), 273.

[33] Yu, B. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*. Springer, New York, NY, 1997, pp. 423–435.