נושאים מתקדמים באלגוריתמים – מבוא לבדיקת תכונות

אלדר פישר, חדר 625, טלפון 3967, eldar@cs.technion.ac.il

28 ביולי 2021

החוברת הזו תפורסם ותעודכן (כולל תוספת של פרקים חדשים) מדי פעם באתר הקורס. עקב המצב, כרגע אפשר להשיג אותי רק באופצית הדוא"ל, או בעת השיעורים הוירטואלים בתוכנת Zoom.

כעיקרון חומר הקורס מתבסס בעיקר על המאמרים הרלוונטים המוזכרים בחוברת (יש נסיון לדאוג שהחוברת תכיל במידת האפשר את כל החומר הרלוונטי). עבור מידע נוסף על טענות בסיסיות, "פולקלור" וכו, אפשר גם לפנות לספר הבא:

Oded Goldreich, Introduction to Property Testing, Cambridge University Press, 2017

אפשר למצוא גרסה שלו באופן מקוון: http://www.wisdom.weizmann.ac.il/~oded/pt-intro.html

בנוסף, הקורס ישתמש בשיטות הסתברותיות בסיסיות, ועל כן לפעמים יהיה צורך לעיין במקומות המתאימים בקורס "שיטות השתברותיות ואלגוריתמים". אלו מכם שלא למדו את הקורס ידרשו לבצע קריאה מקדימה של החומר הרלוונטי. אתם יכולים למצוא את חוברות הקורס באתר הבא (באופן פתוח לכולם): https://eldar.cswp.cs.technion.ac.il/courses/archive/

ציון הקורס יתבסס על שיעורי בית שינתנו במהלכו.

מבוא לבדיקת תכונות

המטרה של אלגוריתם בדיקת תכונות היא לברר האם הקלט הנתון מקיים תכונה מסויימת בזמן קצר מאוד, פחות מהזמן שלוקח לקרוא אותו. באופן מדוייק זה אינו אפשרי אפילו עבור התכונה שערכי הפונקציה $f:\{1,\ldots,n\}\to\{0,1\}$ הם כולם "0": ברור שהאלגוריתם צריך להיות הסתברותי (צריך "לשקול" לקרוא גם את הביטים שלא קוראים, אחרת התכונה בעצם אינה תלויה בהם), וגם אז צריך הרבה קריאות למציאת 1 בודד במקום שרירותי (אח"כ נראה שיטה לחסמים תחתונים למספר הקריאות של אלגוריתם כזה).

במקום זאת נחפש אלגוריתמים שיבדילו בין המקרה "f מקיימת את התכונה" לבין המקרה "כל עוד נשנה את f בפחות מ־ϵn מקומות היא עדיין לא תקיים את התכונה". במקרה למעלה זה כמעט טריביאלי: מגרילים $2/\epsilon$ אינדכסים באופן מקרי, יוניפורמי וב"ת, מתוך תחום הפונקציה, ובודקים שהקלט שווה ל־0 בהם. אם יש לפחות ϵn מקומות שבהם f היא 1, הסיכוי לטעות חסום ע"י $(1-\epsilon)^{2/\epsilon}<e^{-2}<\frac{1}{3}$. במקרה הזה, האלגוריתם הוא גם חד־כיווני (קלט שכן מקיים את התכונה יתקבל בהסתברות 1), ולא־אדפטיבי (אפשר לכתוב מראש את כל השאילתות שהאלגוריתם יבצע ללא תלות בקלט, ורק ההחלטה אם לקבל או לדחות את הקלט תלויה בתשובות שיתקבלו).

באופן כללי יש תכונות "מוגדרות גלובלית" (כגון צביעת גראפים) עבורן הוכחת בדיקה אינה טריביאלית, ויש תכונות עבורן אין בדיקות יעילות. בד"כ נתמקד במספר השאילתות מהקלט ולא בזמן הריצה, ונשאף באופן אידאלי למספר שאילתות שתלוי רק ב־ϵ. בהרבה מקרים (לא כולם) החסם על מספר השאילתות ייתן גם את החסם הידוע על זמן הריצה. המקרה הכי גרוע מבחינת בדיקת תכונות הוא כאשר קיים ϵ ספציפי

שעבורו ϵ־בדיקה תדרוש $\Theta(n)$ שאילתות (ז"א שהיא לא יותר טובה מקריאה של כל הקלט כמו אלגוריתם קלאסי). יש מקרים שבהם ניתן להוכיח חסם תחתון כזה.

פרט לקירוב מהיר ולתיאור מקרים של קריאות יקרות (אז מודל מספר השאילתות הוא אכן המתאים ביותר), בדיקת תכונות מתאימה גם לקלטים ארוכים ולא מפורשים (אחת המוטיבציות הראשונות היתה בדיקת תוכנה). ישנם גם קשרים בין בדיקת תכונות לבין תורת הלמידה ול־PCP.

עוד דוגמה קלה – בדיקת דרגה נמוכה

נניח שנתון לנו שדה סופי $\mathbb{F}$ (ידוע מראש – לא נדון בזמן החישוב של פעולות כפל וחילוק בתוך השדה), ואנחנו רוצים לבדוק האם הפונקציה $f:\mathbb{F}\to\mathbb{F}$ היא פולינום מדרגה חסומה ע"י k. אפשר לעשות את זה ע"י $O(k+1/\epsilon)$ שאילתות: קודם בודקים את ערך f על קבוצה נתונה S מגודל $k+1$. אם f היה פולינום כנדרש, הרי שעכשיו היינו יכולים לחשב את כל ערכי f על $\mathbb{F}\setminus S$. אם כן אז כל שנותר לעשות הוא לדגום $2/\epsilon$ אינדכסים מקבוצה זו, ולבדוק אותם מול החישוב. בעצם אנחנו עברנו חזרה למקרה דומה לתכונה של "הכל אפסים".

בהמשך נראה דוגמאות פחות טריביאליות לטכניקה הזו של "חישוב לפי קבוצת בסיס ובדיקת זהות לפי דגימת השאר". בחלק מאלו הבדיקה לא תהיה מול מועמד יחיד – השיטה עובדת כל עוד אפשר לעשות רדוקציה לקבוצת מועעמדים קטנה מספיק (מספר השאילתות ממנה בסוף יהיה לוגריתמי בגודל קבוצת המועמדים).

עכשיו אבל ננתח גרסה של בדיקת הדרגה עם תכונות "יוניפורמיות" טובות לשימוש בהוכחות אחרות. נניח שאנחנו פשוט מגרילים קבוצה $Q\subset\mathbb{F}$ בת $k+2$ איברים, באופן יוניפורמי מתוך משפחת הקבוצות האפשריות, ובודקים ש־f מתאימה לפולינום מדרגה חסומה ע"י k על קבוצה זו (זה אומר שאם לוקחים $q\in Q$ שרירותית, אז $f(q)$ שווה לערך הנקבע ע"י אינטרפולציה מ־$f|_{Q\setminus\{q\}}$). מהו הסיכוי לגלות הפרה עבור f שהוא ϵ־רחוק מלהיות פולינום? נסתכל על בחירת Q כעל בחירה (יוניפורמית) של $S\subset\mathbb{F}$ מגודל $k+1$, שלאחר מכן מוסיפים לה איבר $q\notin S$ שנבחר יוניפורמית מהאיברים הנותרים.

אנחנו יודעים שלכל $S\subset\mathbb{F}$ מגודל $k+1$ (בין אם נבחרה יוניפורמית ובין אם לא), חייבים להיות לפחות $\epsilon|\mathbb{F}|$ איברים ב־$\mathbb{F}\setminus S$ שעבורם ערך f יהיה שונה מהערך הניתן ע"י חישוב הפולינום (אחרת היינו יכולים לתקן את f במקומות שהערך אינו שווה, ולקבל פולינום שהיה סותר את ההנחה ש־f היא ϵ־רחוקה מכל פולינום כזה). אפשר להסתכל על Q הנבחרת יוניפורמית כעל התוצאה של בחירה יוניפורמית של S ואז תוספת של איבר $q\in\mathbb{F}\setminus\mathbb{S}$ שגם הוא נבחר יוניפורמית מהאיברים הנותרים. מכאן שהסיכוי לגלות הפרה עבור Q שנבחרת יוניפורמית הוא לפחות $\epsilon|\mathbb{F}|/|\mathbb{F}\setminus S|>\epsilon$.

זוהי דוגמה לאלגוריתם לא תלוי מרחק proximity oblivious, מכיוון שאנחנו לא צריכים להשתמש באסטרטגיות שונות ל־ϵ שונים, אלא רק (אם רוצים להגדיל את הסיכוי לתשובה נכונה) להריץ את אותו אלגוריתם יותר פעמים. אלגוריתמים עם תכונה כזו הם יותר נדירים מאלגוריתמי בדיקת תכונה רגילים. במקרה כאן אנחנו "ברי מזל" במיוחד, כי האלגוריתם גם מקיים שלכל ערך של תחום הפונקציה יש סיכוי זהה להיבחר לשאילתה.

מצד שני, אם רוצים מספר שאילתות מינימלי עבור גילוי של קלטים ϵ־רחוקים בהסתברות לפחות $\frac{2}{3}$, אז עדיף לבצע "במכה אחת" את כל השאילתות, על מנת להידרש ל־$O(k+1/\epsilon)$ שאילתות, במקום $O(k/\epsilon)$.

חסמים תחתונים – ההתחלה

ראשית נראה איך מוכיחים חסם תחתון לדוגמה הכי קלה של "הכל אפסים". החסם הצפוי הוא $\Omega(1/\epsilon)$, שיתאים לאלגוריתם הפשוט שהוצג למעלה, אבל צריך להוכיח את זה. בהינתן אלגוריתם בדיקה בעל q שאילתות, נבדוק מה קורה עבור הפונקציה $f:D\to\{0,1\}$ (כאשר $|D|=n$), במקרה שכל ערכיה שווים ל־0. שימו לב שאפילו אם האלגוריתם אדפטיבי (ז"א שמותר לו לבסס את השאילתות על תשובות קודמות), מכיוון שכאן אנחנו יודעים מראש את כל התשובות, קבוצת השאילתות Q תהיה תת־קבוצה מקרית של D שגודלה חסום ע"י q (כאשר ההתפלגות תלויה באלגוריתם עצמו).

עתה נשים לב שאפשר להניח את ההנחה הבאה על האלגוריתם: כאשר האלגוריתם מגיע לערך שונה מ־0, הוא תמיד ידחה את הקלט (כי זאת לא יכולה להיות טעות עבור התכונה "הכל אפסים"). על כן, אפשר להניח שגם במקרה של קבלת ערך של "1", האלגוריתם ימשיך לקרוא את שאר השאילתות כאילו קיבל ערך "0", רק שבמקרה כזה בסוף הריצה הוא ידחה את הקלט. על כן ההתפלגות על הקבוצה Q היא כמעט כל המידע שצריך בשביל לנתח את האלגוריתם.

לשם המשך הניתוח, נסמן לכל $i \in D$ ב־p_i את ההסתברות שיתקיים $i \in Q$. נשים לב שמתקיים $\sum_{i\in D} p_i \leq q$. ההוכחה משתמשת בשיטה ההסתברותית של לינאריות התוחלת: נסמן ב־X_i את המ"מ שמקבל 1 אם $i \in Q$ ומקבל 0 אחרת (ז"א את משתנה האינדיקטור עבור "$i \in Q$"). מתקיים $|Q| = \sum_{i\in D} X_i$, ולכן לפי לינאריות התוחלת מתקיים $q \geq \mathrm{E}[|Q|] = \sum_{i\in D}\mathrm{E}[X_i] = \sum_{i\in D} p_i$. אם $q < 1/4\epsilon$, ו־n גדול מספיק בשביל שיתקיים $\lceil\epsilon n\rceil \leq \frac{4}{3}\epsilon n$, אז קיימת קבוצה $B \subset D$ מגודל לפחות ϵn שמתקיים עבורה $\sum_{i\in B} p_i < \frac{1}{3}$ (למשל קבוצת האינדכסים של $\lceil\epsilon n\rceil$ ה־p_i הקטנים ביותר), כפי שנראה עתה.

הסיבה שהקבוצה B קיימת היא מקרה פרטי של הטענה הבאה: אם $\sum_{i\in D} p_i = p$ ו־$|D| = n$, אז לכל $0 \leq m \leq n$ קיימת קבוצה B בגודל m כך שמתקיים $\sum_{i\in B} p_i \leq \frac{mp}{n}$ (ואצלנו נציב $m = \lceil\epsilon n\rceil$). יש שתי אפשרויות להוכיח זו. השיטה הראשונה היא שוב ע"י שיטה הסתברותית – מגרילים את הקבוצה B באופן יוניפורמי מבין כל הקבוצות האפשריות מגודל m בדיוק, נגדיר את Y_i להיות משתנה האידיקטור עבור "$i \in B$", ונגדיר את $Z = \sum_{i\in D} Y_i p_i$. מצד אחד מתקיים $Z = \sum_{i\in B} p_i$, מצד שני לפי לינאריות התוחלת מתקיים $\mathrm{E}[Z] = \frac{m}{n}\sum_{i\in D} p_i = \frac{mp}{n}$. על כן יש בחירה ספציפית של B שעבורה הסכום הנ"ל לא יעלה על התוחלת $\frac{mp}{n}$.

נראה גם שיטה שניה להוכחת הקיום של B, באינדוקציה על m. הבסיס $m = 0$ ברור. בשביל המעבר, יותר נוח להראות את התנאי השקול $\sum_{i\in D\setminus B} p_i \geq \frac{(n-m)p}{n}$. אם אנחנו יודעים על קבוצה B' שמקיימת את התנאי עבור $m-1$, אז נבחר את $j \in D\setminus B'$ שעבורו יש ל־p_j ערך מינימלי, ובפרט מתקיים $p_j \leq \frac{1}{n+1-m}\sum_{i\in D\setminus B'} p_i$. נגדיר $B = B' \cup \{j\}$, ואז נקבל $\sum_{i\in D\setminus B} p_i \geq (1-\frac{1}{n+1-m})\sum_{i\in D\setminus B'} p_i \geq \frac{n-m}{n+1-m}\cdot\frac{(n+1-m)p}{n} = \frac{(n-m)p}{n}$.

נבחר את f שתהיה שווה ל־1 מעל B ושווה ל־0 מעל $D\setminus B$. כאשר מריצים את האלגוריתם מעל פונקציה זו, ההסתברות שיתקבל איזה שהוא ערך מ־B היא פחות מ־$\frac{1}{3}$, ואחרת האלגוריתם יקרא רק אפסים. על כן, ההבדל בין ההסתברות של האלגוריתם לדחות את הקלט עבור פונקציה שכולה אפסים (שאמורה להתקבל בהסתברות לפחות $\frac{2}{3}$) לבין ההסתברות לדחות את הפונקציה f (שאמורה להידחות בהסתברות לפחות $\frac{2}{3}$) היא פחות מ־$\frac{1}{3}$, ולכן האלגוריתם חייב להיות שגוי לפחות באחד מהמקרים האלו. על כן כל אלגוריתם בדיקה עבור התכונה "הכל אפסים" חייב לבצע לפחות $1/4\epsilon = \Omega(1/\epsilon)$ שאילתות.

נחזור עתה לשאלה של בדיקת דרגה נמוכה של פונקציה, ונראה חסם תחתון של $\Omega(k)$ עבור בדיקה שהדרגה של $f : \mathbb{F} \to \mathbb{F}$ חסומה ע"י k. נראה שזה נכון אפילו עבור $\frac{1}{2}$־בדיקה, כל עוד $4 \leq 2(k+1) < |\mathbb{F}|$. נניח שהאלגוריתם הנתון $\mathcal{A}$ מבצע לא יותר מ־k שאילתות, ונשווה את ההתנהגות שלו על קלט f שהוא פולינום מקרי ממעלה k שנבחר יוניפורמית מכל האפשריים (מגרילים באופן מקרי וב"ת מקדמים $\alpha_0,\ldots,\alpha_k$ ומגדירים $f(x) = \sum_{i=0}^{k}\alpha_i x^i$), וקלט g שהוא פולינום מקרי ממעלה $k+1$ שנבחר יוניפורמית.

הדבר הראשון לשים לב הוא שמכיוון שהאלגוריתם לא שאל יותר מ־k שאילתות, בכל שאילתה הוא קיבל ערך מקרי יוניפורמי מ־$\mathbb{F}$. זוהי התכונה של פולינומים מדרגה k: לכל קבוצה $Q \subset \mathbb{F}$ מגודל $k+1$ ולכל סדרת ערכים אפשרית עבורה יש בדיוק פולינום יחיד שמסכים עם ההצבה. על כן כל סדרות הערכים האפשריות מתקבלות באותה הסתברות, ולכן, לכל i, בהינתן סדרת הערכים ש־$\mathcal{A}$ קיבל ב־$i-1$ השאילתות הקודמות שלו, התשובה לשאילתה ה־i עדיין תהיה ערך מקרי יוניפורמי מ־$\mathbb{F}$. על כן תהיה לאלגוריתם אותה התנהגות, ואותו סיכוי לדחות, גם עבור f וגם עבור g.

מצד שני, בהסתברות לפחות $\frac{3}{4} > \frac{|\mathbb{F}|-1}{|\mathbb{F}|}$ הפונקציה g לא תהיה פולינום מדרגה k (זה הסיכוי שהמקדם של x^{k+1} יהיה שונה מ־0), ואז המרחק שלה מפולינום כל שהוא מדרגה k יהיה לפחות $|\mathbb{F}| - (k+1) > \frac{1}{2}|\mathbb{F}|$ (עבור פולינום p מדרגה k מסתכלים על פולינום ההפרש $g-p$, ואם הוא שונה מ־0 אז אין לו יותר מ־$k+1$ אפסים). על כן $\mathcal{A}$ צריך לדחות את הפונציה g בהסתברות לפחות $\frac{1}{2} = \frac{3}{4}\cdot\frac{2}{3}$, אולם את הפונקציה f עליו לקבל בהסתברות לפחות $\frac{2}{3}$, ולא יכול להיות שהוא מקיים את שני התנאים בו זמנית.

דוגמה לשיטת התיקון העצמי – בדיקת לינאריות

הדוגמה הזו היא הסטורית התוצאה הראשונה שכונתה "בדיקת תכונות", מהמאמר Blum, Luby, Rubinfeld: Self-testing/correcting with applications to numerical problems. **נניח שיש לנו פונקציה** $f: V \to \mathbb{F}$, כאשר $\mathbb{F}$ הוא השדה $\mathbb{Z}_p$ עבור מספר ראשוני p כל שהוא, V הוא מרחב לינארי סופי מעל $\mathbb{F}$, ואנחנו רוצים לבדוק האם זוהי פונקציה לינארית. מכיוון שהמדובר ב־$\mathbb{Z}_p$, מספיק לדעת שלכל $x, y \in V$ מתקיים $f(x) + f(y) = f(x + y)$. למעשה האלגוריתם שנראה כאן יעבוד גם אם נתון רק שהתחום והטווח הם חבורות חילופיות סופיות, וצריך לבדוק האם f היא הומומורפיזם (פונקציה "שומרת חיבור").

גם כאן האלגוריתם המוצע הוא אלגוריתם לא תלוי מרחק: אנחנו נגריל את $x, y \in V$ באופן יוניפורמי וב"ת, נשאל את שלושת הערכים $f(x), f(y), f(x+y)$, ונבדוק האם תנאי החיבור מתקיים. כדאי אבל לשם הניתוח לנסח את זה בצורה אחרת, שקולה: אנחנו נגריל באופן יוניפורמי ערך z שאותו נרצה "לאמת", ולשם כך נגריל באופן יוניפורמי ערך x שעבורו נבדוק שמתקיים $f(z) = f(z + x) - f(x)$. במילים אחרות, נראה ש־$f(z)$ אכן שווה לערך המתקבל מהדרך האלטרנטיבית לחשב אותו. צורה זו של ניתוח נקראה "תיקון עצמי" self correction, והיא אחת משיטות הניתוח המוקדמות עבור בדיקת תכונות.

אנחנו נראה שההסתברות להפרה עבור פונקציה ϵ־רחוקה מלינאריות היא $\Omega(\epsilon)$. מכיוון שהניתוח ה"רגיל" עובד רק עבור ϵ קטן מקבוע גלובלי כל שהוא, נראה בנפרד שעבור מרחקים גדולים יותר יש חסם תחתון גלובלי קבוע על ההסתברות להפרה. כמו כן (באופן שקורה הרבה בבדיקת תכונות) ננסח את ההוכחה בדרך השלילה: נראה שאם הסיכוי לדחיה קטן, אז f אינה רחוקה מפונקציה לינארית.

כרגע נניח שההסתברות לגילוי הפרה קטנה מ־$\frac{2}{9}$. זה יהיה הקבוע הגלובלי עבור "ϵ גדולים". עבור $z \in V$ כל שהוא, נסמן ב־ $f_x(z)$ את "החישוב האלטרנטיבי" $f(z+x) - f(x)$. נראה עתה שבמקרה זה לכל $z \in V$ קיים ערך, שנסמנו בסימון $g(z)$, שעבורו מתקיים $\Pr_{x\in V}[f_x(z) = g(z)] > \frac{2}{3}$, כאשר ההסתברות היא עבור בחירה יוניפורמית של x מתוך המרחב V. שימו לב שלא בהכרח מתקיים $g(z) = f(z)$.

עבור z נתון, נגדיר וקטור p של מספרים ממשיים עם אינדכסים מתוך $\mathbb{F}$, להיות וקטור ההסתברויות: $p_\alpha = \Pr_{x\in V}[f_x(z) = \alpha]$. ננתח את הנורמה שלו $\|p\|_2^2 = \sum_{\alpha\in\mathbb{F}}(p_\alpha)^2$. מתקיים

$$\sum_{\alpha\in\mathbb{F}}(p_\alpha)^2 = \sum_{\alpha\in\mathbb{F}}\Pr_{x,y\in V}[f_x(z)=f_y(z)=\alpha] = \Pr_{x,y\in V}[f_x(z)=f_y(z)] = \Pr_{x,y\in V}[f_x(z) - f_y(z)=0]$$

כאשר x ו־y מוגרלים יוניפורמית באופן ב"ת. ההפרש $f_x(z) - f_y(z)$ הוא:

$$f(z+x)-f(x)-f(z+y)+f(y) = (f(z+x)+f(y)-f(z+y+x))-(f(z+y)+f(x)-f(z+y+x))$$

עתה נשים לב ש־$z + x$ ו־y (כזוג משתנים מקריים) מתפלגים באופן ב"ת זה בזה ויוניפורמית מתוך V, ולכן מההנחה על f מתקיים $f(z + x) + f(y) - f(z + y + x) = 0$ בהסתברות גדולה מ־$\frac{7}{9}$. באופן דומה מתקיים $f(z + y) + f(x) - f(z + y + x) = 0$ בהסתברות גדולה מ־$\frac{7}{9}$. מכאן (לפי איחוד מאורעות) שמתקיים $\|p\|_2^2 = \Pr_{x,y\in V}[f_x(z) = f_y(z)] > \frac{5}{9}$. חסם כזה על הנורמה של וקטור הסתברות p (שאיבריו אי־שליליים וסכומם הוא 1) יכול להתקיים רק אם יש איבר p_α שערכו גדול מ־$\frac{2}{3}$. נסמן אם כן $g(z) = \alpha$.

השלב הבא הוא להראות שה"פונקציה האידאלית" שלנו g היא באמת לינארית בתנאים האלו. עבור x ו־y קבועים נתונים נגריל באופן יוניפורמי את $z \in V$. אנחנו יודעים בשלב זה שבסיכוי גדול מ־$\frac{2}{3}$ מתקיים $g(x + y) = f_z(x + y)$. כמו כן, בסיכוי גדול מ־$\frac{2}{3}$ מתקיים $g(x) = f_z(x)$, ובסיכוי גדול מ־$\frac{2}{3}$ מתקיים $g(y) = f_{z+x}(y)$, מכיוון שגם $z + x$ מתפלג יוניפורמית מעל V. מכאן, לפי איחוד מאורעות, בסיכוי גדול מ־0 כל שלושת המאורעות מתקיים, ולכן בפרט קיים z ספציפי שעבורו שלושת השוויונות מתקיימים בו זמנית. כל שנותר הוא להציב ולוודא שאכן מתקיים $f_z(x) + f_{z+x}(y) = f_z(x + y)$ לפי איך שאלו מוגדרות. מכאן שלכל x ו־y מתקיים $g(x) + g(y) = g(x + y)$.

מכאן אפשר להראות שאם הסיכוי לגילוי הפרה קטן מ־$\min\{\frac{2}{9},\frac{2}{3}\epsilon\}$, אז הפונקציה f היא ϵ־קרובה לפונקציה לינארית: אנחנו יודעים שבמקרה זה הפונקציה g שהוגדרה למעלה היא לינארית. כמו כן, אם הגרלנו x שעבורו $g(x) \neq f(x)$, אז בהסתברות לפחות $\frac{2}{3}$ על ההגרלה של y מתקיים $f(x) \neq f_y(x) = f(x+y) - f(y)$, ז"א שבדיקת הסכום שלנו תביא לדחיה. על כן יש פחות מ־$\epsilon|V|$ מקומות שעבורם $f(x) \neq g(x)$, ז"א שהפונקציה f היא ϵ־קרובה לפונקציה הלינארית g.

המשמעות היא שאם הפונקציה f היא ϵ־רחוקה מלינאריות, אז האלגוריתם ידחה בהסתברות לפחות $\min\{\frac{2}{9},\frac{2}{3}\epsilon\}$, כנדרש (זה נהיה שווה ל־$\frac{2}{3}\epsilon$ כאשר $\epsilon \leq \frac{1}{3}$).

תכונות חלוקה של גרפים צפופים

המאמר שפרץ דרך לבדיקת תכונות של מבנים קומבינטורים כמו גרפים הוא Goldreich, Goldwasser, Ron: Property testing and its connection to learning and approximation. זוהי גם דוגמה שבה הגדרת המודל עצמו (איך הגרפים מיוצגים) משנה את התשובה לאיזה תכונות ניתן לבדוק. במאמר הזה המודל הנידון הוא מודל הגרפים הצפוף, זה המתקבל מייצוג הגרף ע"י מטריצת סמיכויות (בניגוד לרשימת שכנויות שמתאימה למודל הדליל והמודל ה"כללי"). הבחירה הזו משפיעה גם על מהן השאילתות המותרות וגם על מושג המרחק מתכונות.

קבוצת הצמתים של הגרף $V = \{1,\ldots,n\}$ נתונה לאלגוריתם הבדיקה מראש. שאילתה בודדת היא בירור (עבור $u,v \in V$ שהאלגוריתם שואל עליהם) האם $uv \in E$ או לא. זה אומר שאנחנו מתייחסים לגרף כעל פונקציה שהתחום שלה הוא קבוצה כל הזוגות של איברים מ־V, והטווח שלה הוא $\{0,1\}$ שמייצגים את התשובות האפשריות "לא־קשת" ו־"קשת". בהתאם לכך, המרחק בין $G = (V,E)$ ל־$G' = (V,E')$ מוגדר לפי $|E \Delta E'|/n^2$, מספר הזוגות שהם קשת באחד הגרפים ולא השני מחולק ב־n^2 (הסיבה לחלק ב־n^2 במקום ב־$\binom{n}{2}$ היא מטעמי נוחות, ההשפעה על מושג המרחק היא בפקטור השואף ל־$\frac{1}{2}$).

המודל הזה "גס" למדי. למשל, ניתן לבדוק 3־צביעות במספר שאילתות פולינומי ב־$1/\epsilon$ ולא תלוי ב־n (כאשר זמן החישוב הוא אקספוננציאלי ב־$1/\epsilon$). ההוכחה שמראה שסיבוכיות פתרון 3־צביעות היא NP־קשה מפיקה גרפים עם $o(n^2)$ קשתות, כך שמבחינת המודל הצפוף כל אלו קרובים לגרפים חסרי קשתות ולכן מותר לקבלם.

לבסוף, חשוב להדגיש שגם משפחת התכונות ה"מותרות לדיון" נקבעת ע"י המודל. אנחנו מעוניינים בתכונות גרפים, ז"א שנתון מראש שהתכונה P (כתת־קבוצה של קבוצת כל הגרפים מעל V) היא אינווריאנטית בפרמוטציות של הצמתים: אם G' מתקבל מ־G ע"י הפעלת פרמוטציה $\sigma : V \to V$ (ז"א ש־$uv \in E$ אם ורק אם $\sigma(u)\sigma(v) \in E'$), אז $G \in P$ אם ורק אם $G' \in P$.

בדיקת דו־צדדיות

ראשית נראה איך אפשר לבדוק את התכונה ש־G ניתן לצביעה בשני צבעים. מסתבר שגם כאן משתמשים ברעיון של תיקון עצמי, יחד עם רעיון נוסף שנפרט בקרוב.

ראשית נראה איך אפשר לעשות מעין "תיקון עצמי" לצביעה קיימת שאנחנו יכולים לשאול רק על חלק קטן ממנה. הנקודה לשים לב היא שאם יש לצומת הרבה שכנים, אז ניתן ע"י דגימה קטנה של צמתים למצוא את הצביעה של אחד השכנים שלו, ומכך ניתן להסיק את צבעו (הצבע שאינו שייך לשכנו). לעומת זאת, אם לצומת אין הרבה שכנים, אז הצבע שלו לא משנה הרבה במודל הגרפים הצפופים, כי הוא משפיע על כמות קטנה יחסית של קשתות שאולי נצטרך להסיר.

נניח אם כן שנתון לנו גרף G וכן צביעה $c : V \to \{1,2\}$, ונרצה להיות מסוגלים לבדוק האם זו צביעה טובה, עד כדי ϵn^2 קשתות מפרות (קשתות עם שני צמתים באותו צבע). עם זאת, נרצה לשאול על הצביעה עצמה במספר מקומות קטן ככל האפשר. בשלב ראשון, נבחר קבוצת צמתים $S = \{u_1,\ldots,u_s\}$ שעבורם נשאל את הצביעה, כאשר כל u_i נבחר באופן יוניפורמי וב"ת באחרים, עם חזרות (מותר שיהיו $i < j$ עם $u_i = u_j$, אם

כי הסיכוי לכך שואף ל־0 עבור n גדול). בחירת הצמתים באופן שמתיר חזרות תפשט את הניתוח בהמשך. אנחנו נרצה שבסיכוי גבוה יהיו לא יותר מ־$\epsilon n/4$ צמתים עם דרגה גבוהה מ־$\epsilon n/4$ שלא תפסנו שכן שלהם.

הסיבה לתנאי: אם הגענו למצב כזה, אפשר לצבוע כל צומת שיש לו שכן $u \in S$ ב"צבע השני" $3 - c(u)$. אם הצביעה היתה טובה, אז לא עשינו כאן טעויות. עבור צמתים ללא שכנים פשוט נתעלם מהם ומהקשתות שנוגעות בהם. המדובר בלא יותר מ־$\epsilon n^2/2$ קשתות סה"כ (עד $\epsilon n^2/4$ קשתות של צמתים מדרגה נמוכה, אפילו אם אלו כל הצמתים בגרף, ועד $\epsilon n^2/4$ קשתות של צמתים "מפוספסים" מדרגה גבוהה, שאין יותר מ־$\epsilon n/4$ מהם). עכשיו אפשר לבדוק את הצביעה בגרף: נגריל באופן יוניפורמי וב"ת זוגות $w_1v_1, \ldots, w_tv_t$, ולכל זוג כזה נבדוק האם הוא קשת מפרה (נבדוק האם אנחנו יכולים להסיק את הצבעים של w_i ו־v_i לפי S, ואם אנחנו יכולים אז נבדוק האם זהו אותו צבע והאם הזוג הוא קשת של G). אם היו לפחות $\epsilon n^2/2$ קשתות מפרות נוסף על אלו שאולי התעלמנו מהן (אלו שמכילות צמתים שאי אפשר להסיק את צבעם), אז זוג מקרי יהיה קשת כזו בהסתברות לפחות ϵ. הסיכוי שלא גילינו קשת כזו ב־t דגימות חסום ע"י $(1-\epsilon)^t < e^{-\epsilon t}$. הערך $t = 2/\epsilon$ למשל ייתן לנו הסתברות גדולה מ־$\frac{5}{6}$ שנגלה קשת כזו ונוכל לפסול את הצביעה בהצלחה.

איך ממלאים את התנאי על S: עבור צומת ספציפי מדרגה לפחות $\epsilon n/4$, הסיכוי שלא יהיה לו שכן ב־S חסום ע"י $(1-\epsilon/4)^s < e^{-\epsilon s/4}$. נציב $s = 16/\epsilon - 4\ln(\epsilon)/\epsilon$ ונקבל שתוחלת מספר הצמתים עם דרגה גבוהה ללא שכנים ב־S חסומה ע"י $n \cdot (1-\frac{\epsilon}{4})^s \leq n \cdot e^{\ln(\epsilon)-4} < \epsilon n/24$ (כאן "ln" מתייחס לבסיס טבעי, ו־"log" מתייחס לבסיס 2). לפי אי שוויון מרקוב, בהסתברות לפחות $\frac{5}{6}$ לא יהיו לנו יותר מ־$\epsilon n/4$ צמתים כאלה.

עכשיו שאנחנו יודעים לבדוק צביעה בודדת במספר שאילתות לא גדול, איך נבדוק את הקיום של צביעה כל שהיא? כאן תעזור לנו עוד טכניקה בסיסית של בדיקת תכונות, שהיא הסיבה האמיתית שניסינו מראש לצמצם את גודל S. מכיוון שעכשיו איננו יכולים לשאול את ערכי הצביעה של S, אנחנו פשוט נסתכל על כל 2^s הצביעות האפשריות שלה, ונבדוק כל אחת מהן (אם נמצא צומת מחובר לצמתים משני הצבעים ב־S אפשר לפסול את הצביעה הזו מיידית, או פשוט לצבוע אותו שרירותית ב־"1" – בכל מקרה אח"כ נפסול את הצביעה המתקבלת אם יש בה הרבה קשתות מפירות).

עם זאת, אנחנו לא נרצה לעשות את הבדיקות אחת אחרי השניה, כי זה כבר מספר שאילתות אקספוננציאלי ב־$1/\epsilon$, ולמרות שזה גם לא תלוי ב־n, נרצה חסם פולינומי ב־$1/\epsilon$. על כן נעשה "בדיקה במקביל" של אפשרויות ה"ניחוש": כאשר נבחר את t למעלה, במקום חסם של $\frac{1}{6}$ על הסיכוי לפספס קשת מפרה של הצביעה הנתונה, נרצה שהסיכוי יהיה חסום ע"י $\frac{1}{6}2^{-s}$. במצב כזה, נבחר את $w_1v_1, \ldots, w_tv_t$ פעם בודדת, ואז לפי איחוד מאורעות יהיה לנו חסם של $\frac{1}{6}$ על הסיכוי שפספסנו קשת מפרה של איזו שהיא מהצביעות שיש להן עודף קשתות מפרות (ואם יש צביעה ללא עודף של קשתות מפרות, אז זה בסדר לקבל את הגרף). נשים לב שגם אם בודקים את כל הצביעות האפשריות בבת אחת, מספר השאילתות הכולל אינו עולה על $2st + t$: לכל i ולכל $u \in S$ שואלים את הזוג w_iu ואת הזוג v_iu (זה מספיק כדי למצוא שכן ב־S, שלפיו "נחשב" את הצבעים של v_i ו־w_i לכל צביעה של S), וכן שואלים את הזוג v_iw_i (כדי לדעת האם הוא בכלל יכול להיות מפר – בשביל זה הוא קודם כל חייב להיות קשת של G).

על מנת לקבל את החסם הנדרש, אפשר לבחור $t = 2/\epsilon + \ln(2)s/\epsilon = O(s/\epsilon) = O(\log(1/\epsilon)/\epsilon^2)$. הרבה פעמים נשתמש בסימון ש"בולע" מקדמים שהם חזקה קבועה של לוגריתם הביטוי, ונכתוב $t = \tilde{O}(1/\epsilon^2)$. מספר השאילתות הכולל כאן יהיה $O(st) = O((\log(1/\epsilon))^2/\epsilon^3) = \tilde{O}(1/\epsilon^3)$.

כדאי גם לסכם שוב את הטיעון לנכונות האלגוריתם: אם יש צביעה חוקית $c : V \to \{1,2\}$ עבור הגרף, אז בכל מקרה אחת הצביעות שנבדוק עבור S היא $c|_S$, וצביעה ספציפית זו לא תגרום למציאת קשתות מפרות כאשר נבדוק אותה מול הזוגות w_iv_i (אם יש צמתים ללא שכנים מ־S אז אנחנו בכל מקרה לא מייחסים אותם לזוגות מפרים, אז גם אם S "מקולקלת" לא נטעה לכיוון השלילי).

אם מצד שני לכל צביעה אפשרית יש לפחות ϵn^2 קשתות מפרות, אז נגלה את זה בהסתברות לפחות $\frac{2}{3}$: בהסתברות לפחות $\frac{5}{6}$ הקבוצה S כוללת שכנים לצמתי הגרף מדרגה גבוהה מספיק, למעט מספר קטן מהם. במקרה כזה, לכל צביעה של הצמתים שיש להם שכנים ב־S יהיו לפחות $\epsilon n^2/2$ קשתות מפרות (כי יש לא יותר מ־$\epsilon n^2/2$ קשתות עם צמתים שמתעלמים מהן). לפי איחוד מאורעות, בסיכוי כולל של לפחות $\frac{5}{6}$, בכל אחת ואחת מ־$2^{|S|}$ הצביעות שנבדוק נגלה קשת מפרה בין הזוגות $w_1v_1, \ldots, w_tv_t$, ולכן נצליח לפסול את כולן ולדחות את הגרף.

לסיום, נראה כאן איך במקרה הספציפי של 2־צביעה (זה כבר לא נכון עבור תכונות כמו 3־צביעה או חתך מקסימלי) אפשר לדאוג שגם זמן הריצה מעבר ללקיחת השאילתות יהיה פולינומי ב־$1/\epsilon$. במקום לרוץ באופן מפורש על כל הצביעות של S, נסתכל על תת־הגרף המורכב מכל הקשתות שגילינו במהלך כל השאילתות שלנו, וננסה למצוא 2־צביעה שלו (למשל באמצעות אלגוריתם חיפוש לעומק). אם אין צביעה כזו, אז מצאנו תת־גרף לא צביע של הגרף המקורי, ולכן אפשר לדחות את G. אם יש צביעה כזו, אז בפרט הצמצום שלה ל־S היא צביעה שלא היינו פוסלים בשלב בדיקת הצביעות, ולכן זה לגיטימי לקבל את G.

כמה מילים על בדיקת צביעות במספר יותר גדול של צבעים

כאשר רוצים לבדוק k־צביעות עבור $k \geq 3$ קבוע, הבעיה היא שצביעת שכן של צומת v לא קובעת את הצבע של v, אלא יכולה רק "לפסול" צבע אחד אפשרי שלו. לא ניכנס כאן להוכחה מלאה של אלגוריתם בדיקה חד־כיווני עבור k־צביעות (אתם מוזמנים לקרוא אותה במאמר המקורי), אבל הרעיון הוא כזה: חושבים על הבחירה של S צומת־צומת. אם יש צביעה חוקית נוכחית של S כך שרוב הצמתים שנותרו ניתנים לצביעה בצורה שלא תפסול צביעות להרבה צמתים, אז אפשר לעשות את החשבון שניתן לצבוע את הגרף עם מעט קשתות מפרות. אם אין צביעה כזו, אז תוך כדי בחירת צמתים נוספים נמצא לכל צביעה של S צומת שכל צביעה שלו תפסול צבע להרבה צמתים (בטיעון הפורמלי בונים "עץ צביעות חלקיות אפשריות" עבור תתי־קבוצה מתאימים של S). לבסוף נפסול כל כך הרבה צבעים כך שיהיו מספיק צמתים שאין להם צבע חוקי כלל, וכך נוכל למצוא צמתים שיפסלו כל צביעה חוקית של S (באופן פורמלי נמצא עלים לעץ הצביעות החלקיות עד שלא יישאר ענף "פתוח" שם). אפשר להישאר עם מספר שאילתות פולינומיאלי ב־$1/\epsilon$, אבל זמן הריצה יהיה אקפוננציאלי ב־$1/\epsilon$, כי גם אחרי שנעביר את החישוב לכזה של מציאת k־צביעה בודדת של תת־הגרף של השאילתות שלנו, עדיין נצטרך זמן חישוב אקספוננציאלי בגודל תת־הגרף הנ"ל.

מסתבר שיש גם הכללה להיפרגרפים, ולתכונות יותר כלליות – ההכללה האולטימטיבית היא עבור CSP (Constraint Satisfaction Programs) מעל אלף־בית ("מספר צבעים") קבוע ופסוקיות מאורך קבוע. יש אותה במאמר Alon, Shapira: Testing satisfiability.

בדיקת חתך מקסימלי

נראה כאן את הדוגמה הראשונה שלנו שבה האלגוריתמים חייבים להיות עם שגיאה דו־צדדית. חתך בגרף $G = (V, E)$ הוא חלוקה של של קבוצת הצמתים V לשתי קבוצות זרות U ו־$W = V \setminus U$. הצפיפות של החתך מוגדרת כמספר הקשתות בין U ל־W, מחולק ב־n^2 (במקרה הזה אנחנו נתעניין בצפיפות "אבסולוטית", ולא בהגדרה המקובלת של צפיפות "יחסית" שבה מחלקים במספר הקשתות המקסימלי $|U| \cdot |W|$). במושגים של בדיקת תכונות, נרצה לבדוק את התכונה שיש לגרף חתך בעל לפחות αn^2 קשתות. במקרה הזה יותר נוח לתאר את בעיית הקירוב, של מציאת ערך η כך שבהסתברות לפחות $\frac{2}{3}$ צפיפות החתך המקסימלי תהיה בין $\eta + \epsilon$ ל־$\eta - \epsilon$. על מנת לפתור את בדיקת התכונה, נבצע את הקירוב עם $\epsilon/2$, ונקבל את הגרף אם קיבלנו ערך η של לפחות $\alpha - \epsilon/2$.

על מנת לראות מדוע אנחנו צריכים שגיאה דו צדדית, אפשר למשל להסתכל על גרף שהוא איחוד זר של קליק בעל $\lceil n/2 \rceil$ צמתים יחד עם עוד $\lfloor n/2 \rfloor$ צמתים חסרי קשתות. אפילו אם אנחנו מבצעים $\lfloor n/4 \rfloor$ שאילתות, יש סיכוי חיובי (אומנם קטן) שלא נגלה קשתות כלל, כמו שיש סיכוי שכל השאילתות שלנו יחזירו קשתות. במקרה כזה לא נוכל להבחין בין המקרה ש־G הוא הגרף המלא לבין המקרה ש־G חסר קשתות כלל, מה שאומר שלא נוכל לבצע בדיקה אפילו עבור למשל $\epsilon = \frac{1}{3}$ (או קירוב עבור למשל $\epsilon = \frac{1}{6}$). כעיקרון, בכל תכונה של ספירה, אפילו עבור התכונה "יש בגרף לפחות $n^2/4$ קשתות" (או במקרה של מחרוזות, תכונה הסופרת את מספר ה־1 במחרוזת), נצטרך בדיקה עם שגיאה דו־צדדית.

אנחנו נשתמש בשיטה דומה לזו של בדיקת דו־צביעות. בבדיקת דו־צביעות אנחנו מסתמכים על קבוצה S כך שלרב הצמתים ה"משמעותיים" יש שכנים ב־S, ולכן מחלוקה של S אפשר להסיק על חלוקה מקורבת של כל הגרף. אח"כ אנחנו מנתחים את כל החלוקות האפשריות של S. במקרה שלנו לא מספיק לדעת על שכן בודד של צומת v על מנת לשייך אותו לחלוקה, אבל כן אפשר להסיק משהו אם יודעים את מספר

השכנים שלו בכל אחד מהצדדים. אם עבור חתך מקסימלי (U,W) יש ל־v למשל יותר שכנים ב־U מאשר ב־W, אז הוא חייב להיות משוייך ל־W (אחרת ניתן להגדיל את החתך ע"י שינוי השיוך של v). אם יש ל־v אותו מספר שכנים ב־U ו־W, ניתן לשייך אותו לכל אחת מהקבוצות ונקבל חתך מאותו גודל.

אי אפשר לדעת את מספר השכנים במדוייק, אבל נראה מה קורה אם נסתכל על החלוקה ה"נכונה" של קבוצה S בגודל $s = 500\log(1/\epsilon)/\epsilon^2$ (אנחנו לא ננסה להגיע לקבועים האופטימלים כאן) שנבחרה מקרית. ליתר דיוק, S תהיה סידרה של צמתים שכל אחד מהם נבחר באופן יוניפורמי וב"ת, ואם צומת נבחר יותר מפעם אחת אז הוא גם ייספר יותר מפעם אחת במניינים שנגדיר. עבור צומת v וקבוצה R נסמן ב־$n_{v,R}$ את מספר השכנים של v ב־R. במקרה שלנו נבדוק את $\alpha_{v,U} = \frac{1}{s}n_{v,S\cap U}$ (כאשר נספור כפילויות אפשריות ב־S אם יש כאלה) ונשווה אותו ל־$\beta_{v,U} = \frac{1}{n}n_{v,U}$, ובאופן דומה נשווה את $\alpha_{v,W}$ ל־$\beta_{v,W}$. התוחלת של $\alpha_{v,U}$ שווה ל־$\beta_{v,U}$. יתרה מזו, המספר $n_{v,S\cap U}$ הוא (עבור S שנבחר יוניפורמית באופן ב"ת עם חזרות) סכום של s משתנים מקריים ב"ת שכל אחד מהם שווה ל־1 בהסתברות $\beta_{v,U}$ בדיוק ושווה ל־0 בהסתברות $1-\beta_{v,U}$. ניזכר שלפי חסימת סטיות גדולות $\Pr[n_{v,S\cap U} > s\beta_{v,U}+a] < e^{-2a^2/s}$, עם חסמים דומים עבור המאורע $n_{v,S\cap U} < s\beta_{v,U}-a$ והמאורעות הדומים עבור W. מציבים $a = \frac{\epsilon}{16}s$, ומאיחוד על ארבעת המאורעות נקבל שהסיכוי שלא מתקיימת רביעיית המאורעות $\alpha_{v,U} = \beta_{v,U} \pm \frac{\epsilon}{16}$ ו־$\alpha_{v,W} = \beta_{v,W} \pm \frac{\epsilon}{16}$ חסום ע"י $\epsilon^2/50$.

בשלב זה הדבר הנאיבי לעשות הוא לנסות את כל החלוקות האפשריות של S (אחת מהם מובטחת להיות החלוקה ה"נכונה" ל־$S\cap U$ ו־$S\cap W$), לקטלג לפיה את צמתי הגרף, ואז לקרב את מספר קשתות החלוקה באמצעות דגימה (ולבסוף לקחת את המקסימום של האפשרויות). יש אבל בעיה עם זה: אם מעבירים צומת בודד מצד לצד, אכן האינטואיציה שלא איבדנו הרבה קשתות מהחתך עובדת. אבל אם מעבירים הרבה צמתים בבת אחת, אז הקשתות בתוך קבוצת הצמתים המועברים משפיעות יותר מדי. כעיקרון, אחרי שמעבירים יותר מדי צמתים, החלוקה שלפיה חילקנו את S בתחילת הקטלוג מפסיקה להיות החלוקה הנכונה עבור קטלוג הצומת הבא.

נחשב אבל את "אי־הדיוק המירבי" אם בשלב הראשון נקטלג רק קבוצת צמתים Y שגודלה חסום ע"י $\lceil\frac{\epsilon}{4}n\rceil$: נניח שהגרלנו את S כפי שמתואר למעלה. עבור $v \in Y$ שמקיים את $\alpha_{v,U} = \beta_{v,U} \pm \frac{\epsilon}{16}$ ו־$\alpha_{v,W} = \beta_{v,W} \pm \frac{\epsilon}{16}$, אם (בפעם שבחנו את החלוקה ה"נכונה" של S) הזזנו את v יחסית לחלוקה המקורית אז איבדנו לא יותר מ־$\frac{\epsilon}{8}n$ קשתות חתך בין v ל־$V\setminus Y$. אם למשל העברנו אותו מ־U ל־W (בגלל שהתקיים $\alpha_{v,W} \le \alpha_{v,U}$), אז מתקיים $\beta_{v,U} \ge \alpha_{v,U} - \frac{\epsilon}{16} \ge \alpha_{v,W} - \frac{\epsilon}{16} \ge \beta_{v,W} - \frac{\epsilon}{8}$, ובמקרה הכי גרוע כל $\frac{\epsilon}{8}n$ הקשתות האפשריות בהבדל נמצאות בין v ל־$W\setminus Y$. בנוסף לכך, יכול להיות שאיבדנו את כל הקשתות בין v לצמתים אחרים ב־Y, ומספר קשתות אלו לא עולה על $\frac{\epsilon}{4}n$.

מכיוון שלכל $v \in Y$ ההסתברות שלא יקיים את החסמים הדרושים על $\alpha_{v,U}$ ו־$\alpha_{v,W}$ חסומה ע"י $\epsilon^2/50$, לפי אי שוויון מרקוב, בהסתברות לפחות $1-\epsilon/25$ לא יהיו יותר מ־$\frac{\epsilon}{2}|Y|$ צמתים רעים כאלו. במקרה כזה, האיבוד הכולל כתוצאה מהקטלוג של Y הוא של לכל היותר $(\frac{\epsilon}{8}+\frac{\epsilon}{4})n|Y| + \frac{\epsilon}{2}|Y|n = \frac{7}{8}\epsilon|Y|n$ קשתות.

נחשוב עתה על התהליך הבא: מחלקים את V באופן שרירותי ל־$l = \lceil\frac{4}{\epsilon}\rceil$ קבוצות $Y_1,\ldots,Y_l$ שוות ככל שניתן, ז"א שבפרט כל קבוצה היא מגודל לכל היותר $\lceil\frac{\epsilon}{4}n\rceil$. לכל $1 \le i \le l$ נגריל לפי הסדר קבוצה S_i בת $s = 500\log(1/\epsilon)/\epsilon^2$ צמתים, ולפיה נקטלג את Y_i. צריך לחשוב על זה שאת Y_1 מקטלגים ביחס לחלוקה המקורית (U,W), את Y_2 מקטלגים לפי החלוקה לאחר הקטלוג מחדש של Y_1, וכו. מכיוון שאנחנו בעצם לא יודעים את החלוקה המקורית, או אפילו את הקטלוגים מחדש במלואם, פשוט ננסה את כל החלוקות האפשריות של $S_1,\ldots,S_l$, ז"א שיש לנו $2^{sl} = 2^{O(\log(1/\epsilon)/\epsilon^3)}$ קטלוגים אפשריים שנצטרך לבדוק.

במידה וכל $S_1,\ldots,S_l$ מקיימים את תנאי הקירוב הדרושים ביחס ל־$Y_1,\ldots,Y_l$, הקטלוג לפי האפשרות ה"נכונה" בין 2^{sl} האפשרויות יהיה עם איבוד של לא יותר מ־$\frac{7}{8}\sum_{i=1}^{l}\epsilon n|Y_i| = \frac{7}{8}\epsilon n^2$ קשתות. הסיכוי שזה יקרה, לפי איחוד מאורעות, הוא לפחות $1-\frac{\epsilon}{25}l$, ואם נניח שמתקיים $\epsilon \le \frac{1}{6}$ אז ערך זה יהיה לפחות $\frac{5}{6}$ (העיגול למעלה של l הוא זה שמצריך את ההנחה). עבור $\epsilon > \frac{1}{6}$ פשוט נבצע $\frac{1}{6}$־בדיקה במקום ϵ־בדיקה.

אם היינו יכולים לחשב את מספר הקשתות החוצות עבור כל קטלוג אפשרי בשלב זה אז היינו מסיימים, אבל כמובן שאי אפשר לעשות כזה דבר. אפשר אבל לעשות את הדבר הבא: נגריל $t = 100sl/\epsilon^2$ זוגות $(u_1,v_1),\ldots,(u_t,v_t)$ של צמתים מתוך G באופן יוניפורמי וב"ת. לכל חלוקה אפשרית של $S_1,\ldots,S_l$ נספור את מספר הזוגות (u_i,v_i) המהווים קשת של החתך המתאים (ז"א שהם קשת של G, ובנוסף u_i ו־v_i לא

משוייכים לאותה קבוצה של החתך). הספירות האלו דורשות שנשאל את כל הקשתות u_i, v_i, כל זוג של u_i עם צומת כל שהוא של S_j כאשר j הוא האינדכס כך ש־$u_i \in Y_j$, וכל זוג של v_i עם צומת כל שהוא של S_j כאשר j הוא האינדכס כך ש־$v_i \in Y_j$. סה"כ מספר השאילתות יהיה $t + 2ts = O(s^2 l/\epsilon^2) = O((\log(1/\epsilon))^2/\epsilon^7)$, או בקיצור $\tilde{O}(\epsilon^{-7})$ (כזכור זהו הסימון כשלא אכפת לנו מתוספת של פקטור $\log$ בחזקה קבועה כל שהיא). נסמן ב־m את מספר הזוגות מבין $(u_1, v_1), \ldots, (u_t, v_t)$ שהתגלו להיות קשתות חתך.

עם הגרלה וספירה כמו למעלה, תוחלת הביטוי $m/2t$ היא בדיוק M/n^2, כאשר M יסמן את גודל החתך שאנחנו דוגמים. מחסימת סטיות גדולות, בסיכוי לפחות $\frac{5}{6}$ הספירה שלנו תתן קירוב של גודל כל החתכים שאנחנו בודקים עד כדי תוספת או חיסור של $\frac{1}{8}\epsilon n^2$ קשתות. בסיכוי לפחות $\frac{2}{3}$, גם $S_1, \ldots, S_l$ מקיימות את תנאי הקירוב וגם ההערכות לגדלי החתך מדוייקות מספיק, ולכן לקיחת המקסימום מבין כל החתכים המועמדים תתן לנו את הקירוב הדרוש לגודל החתך המקסימלי.

באופן יותר מדוייק, אם גודל החתך המקסימלי הוא ηn^2, ומתקיימים כל התנאים (כל המאורעות למעלה בנוגע לקירובים), אז בפרט לאף חתך לא נקבל קירוב העולה על $(\eta + \frac{1}{8}\epsilon)n^2$. כמו כן, לפחות אחד החתכים שנבדוק (זה שנובע מהחלוקות ה"נכונות" של $S_1, \ldots, S_l$) הוא בעל לפחות $(\eta - \frac{7}{8}\epsilon)n^2$ קשתות, והקירוב שנקבל עליו הוא לפחות $(\eta - \epsilon)n^2$. סה"כ קיבלנו ערך שנמצא בתחום $(\eta \pm \epsilon)n^2$, כנדרש.

כמה מילים על בדיקת חלוקה כללית

המקרה הכי כללי הוא זה שמספקים תחומים מותרים עבור גודל כל קבוצה בחלוקה של V ל־$V_1, \ldots, V_k$ (שוב עבור k קבוע), וכן תחומים מותרים עבור מספר הקשתות בין V_i ו־V_j עבור $1 \leq i \leq j \leq k$ (שימו לב שזה כולל אפשרות ל־$i = j$). גם כאן הניתוח נעשה דרך חלוקה שרירותית למספר קבוע של "פרוסות". בשביל להשלים את הניתוח אבל מסתכלים על תכונות יותר "מפורטות", ועבור בדיקת תכונת החלוקה מסתכלים על משפחה של תכונות המפורטות שגוררות קירוב שלה. תכונה מפורטת טיפוסית כוללת לא רק את מספר הקשתות בין כל V_i ו־V_j, אלא לכל V_i אנחנו נדרוש הגבלה על מספר הצמתים מכל "סוג" אפשרי, כאשר סוג הצומת נקבע ע"י (קירוב של) מספר הקשתות ממנו לכל V_j בחלוקה.

גם כאן ניתן להכליל לחלוקות של היפרגרפים, וההכללה נמצאת במאמר Fischer, Matsliah, Shapira: Approximate hypergraph partitioning and applications.

בדיקות קנוניות

אפשר להפוך כל אלגוריתם בדיקה במודל הצפוף לאלגוריתם בדיקה לא־אדפיטיבי בעל מבנה פשוט ביותר. זה נעשה לראשונה במאמר Goldreich, Trevisan: Three theorems regarding testing graph properties. המחיר במספר השאילתות יהיה ריבועי: אם לאלגוריתם המקורי היו q שאילתות לכל היותר, לאלגוריתם החדש יהיו $\binom{2q}{2}$ שאילתות.

בהינתן אלגוריתם $\mathcal{A}$, בשלב ראשון נהפוך את השאילתות שלו ל"שאילתות צמתים". נתחזק קבוצת צמתים Q, ובתחילת האלגוריתם נקבע $Q = \emptyset$. עתה, בכל פעם שהאלגוריתם $\mathcal{A}$ שואל שאילתה על זוג u, v, נכניס את שני הצמתים האלו ל־Q אחד אחרי השני, למעט צמתים שכבר היו ב־Q קודם. בכל פעם שנכניס צומת חדש ל־Q, נשאל את כל השאילתות האפשריות בינו לבין כל צומת אחר ב־Q. בסוף התהליך קבוצת השאילתות ששאלנו תכלול גם את השאילתה על u, v שאותה נזין לאלגוריתם המקורי (קבוצת השאילתות גם יכולה לכלול הרבה שאילתות "מיותרות"). רגע לפני שהאלגוריתם עוצר, אם $|Q| < 2q$, אז נכניס שרירותית עוד צמתים ל־Q (ונשאל את השאילתות המתאימות) על מנת שגודל הקבוצה תמיד יהיה בסוף $2q$ בדיוק.

נסמן את האלגוריתם החדש ב־$\mathcal{A}'$. שימו לב שהוא זהה לאלגוריתם המקורי למעט העובדה שקבוצת השאילתות שלו תכלול שאילתות נוספות על אלו של $\mathcal{A}$. עתה ננתח את ההתנהגות של $\mathcal{A}'$ כאשר רגע לפני ההרצה, מעבירים את קבוצת הצמתים V של הגרף דרך פרמוטציה מקרית $\sigma : V \to V$ שנבחרה יוניפורמית מבין $|V|!$ הפרמוטציות האפשריות. במקרה כזה, בכל פעם שהאלגוריתם מוסיף צומת חדש לקבוצה Q, זה יהיה צומת שנבחר יוניפורמית מבין קבוצת כל הצמתים שעוד לא הוכנסו ל־Q קודם לכן. מכאן שאפשר לכתוב

את $\mathcal{A}'$ כאלגוריתם לא אדפטיבי בצורה הבאה: בוחרים מראש את קבוצת הצמתים שתוכנס ל־Q ואת סדר ההכנסה, שנסמן ב־$v_1, \ldots, v_{2q}$, באופן יוניפורמי מבין כל הסדרות ללא חזרות מאורך $2q$. אח"כ שואלים את כל הזוגות האפשריים v_i, v_j עבור $1 \leq i < j \leq 2q$, ולבסוף מבצעים סימולציה של ההרצה המתאימה של $\mathcal{A}'$, שכל השאילתות שלו מכוסות עתה ע"י שאילתות מתאימות מבין אלו שבצענו.

כדאי לציין שכאשר העברנו את קבוצת הצמתים דרך פרמוטציה מקרית, עשינו שימוש בכך שהתכונה הנבדקת היא תכונה של גרפים: זה מה שמבטיח שלגרף לפני הפרמוטציה ולגרף אחרי הפרמוטציה יהיה אותו מרחק מהתכונה (כולל שגרף מקיים יועבר לגרף מקיים).

הצורה הזו של אלגוריתם בדיקה במודל הצפוף נקראת צורה קנונית. בעבר נעשה בה שימוש כחלק מהוכחה של חסמים תחתונים על אלגורימי בדיקה. זהו גם קדימון טוב לטכניקות נוספות על חסמים תחתונים, כולל אלו שישמשו אותנו בהמשך הקורס בהוכחת חסם תחתון על בדיקת דו־צדדיות במודל הגרפים הדליל.

בדיקת מונוטוניות

דוגמה ראשונה

נבדוק את התכונה של היותה של של פונקציה $f : \{1, \ldots, n\} \to \mathbb{N}$ מונוטונית לא־יורדת. זו ניתנת לבדיקה ע"י $O(\log n)$ שאילתות לכל ϵ קבוע. זה הוכח ע"י Ergün, Kannan, Kumar, Rubinfeld, Viswanathan: Spot checkers (כיום גם ידוע שאי אפשר לבדוק בפחות שאילתות). בהמשך נראה את האלגוריתם שלהם, אבל עתה נראה אלגוריתם אחר (של Rubinfeld) עם הוכחה פשוטה.

הרעיון הוא שאם הסידרה היא אכן מונוטונית עולה, אז אפשר למצוא כל איבר בה ע"י חיפוש בינארי. לכל איבר נוכל לבדוק האם נסיון לחיפוש בינארי אכן מגיע אליו, ב־$\log n$ שאילתות. במקרה של שוויון, נלך לכיוון ה"נכון". בעצם אפשר לחשוב על זה שאנחנו "מכריחים" את כל האיברים להיות שונים זה מזה (אנחנו "טיפה מגדילים" את האיבר המאוחר יותר), באופן שאם הסדרה היתה מונוטונית לא־יורדת קודם, עכשיו היא תהיה מונוטונית עולה ממש (ומצד שני, לא "נמחק" ככה שום הפרה ישנה למונוטוניות).

ע"י בחירה יוניפורמית ב"ת של $2/\epsilon$ אינדקסים וביצוע החיפוש עבורם, נוכל להבדיל ב־$O(\epsilon^{-1} \log n)$ שאילתות בין המקרה שכל הערכים $\{f(1), \ldots, f(n)\}$ ניתנים למציאה כזו, לבין המקרה שלפחות ϵn מתוכם אינם ניתנים לכך. נראה עתה שתת־הסדרה של האיברים הניתנים למציאה היא מונוטונית לא יורדת, ובכך נסיים: אם ההסתברות לגילוי של איבר שאינו ניתן למציאה ב־$2/\epsilon$ נסיונות קטנה מ־$\frac{2}{3}$, אז זה אומר שלפחות $(1-\epsilon)n$ מהערכים הם ניתנים למציאה, ולכן הם בסדר מונוטוני בין עצמם. את המקומות ה"חסרים" אפשר למלא בעותקים של הערך הראשון הניתן למציאה שאחריהם (אם אין כזה, אז לוקחים את המקסימום מבין שאר הערכים).

ההוכחה שכל זוג של ערכים ניתנים למציאה הוא בסדר הנכון, היא ע"י הסתכלות בחיפושים הבינארים שלהם בנקודה המשותפת האחרונה ששני החיפושים עברו דרכה. עבור הערך של האיבר הגדול יותר עברנו "ימינה", ז"א שערכו הוא לפחות כמו ערך הנקודה המשותפת האחרונה (וזה נכון גם במקרה שהאיבר עצמו הוא הנקודה המשותפת האחרונה של שני החיפושים), בעוד שעבור האיבר בעל האינדכס הקטן יותר ערכו הוא לכל היותר ערך האיבר המשותף האחרון בחיפוש.

לבסוף, שימו לב לכך שהאלגוריתם אינו אדפטיבי, למרות שבמבט ראשון הוא נראה ככזה. אפשר כל פעם לשאול מראש את "רצף השאילתות עבור החיפוש הבינארי שמגיע לאיבר המבוקש" ולדחות את הקלט מיידית ברגע שמגלים חריגה ממנו. האלגוריתם הוא גם חד־כיווני, כי דחיה מייד נותנת דוגמה נגדית למונוטוניות (הזוג המפר מורכב מנקודת החריגה מהחיפוש הבינארי מול האיבר ש"מחפשים").

לפני שממשיכים – כמה הגדרות

במקרה הכללי נרצה לבדוק מונוטוניות עבור פונקציה $f : D \to R$, כאשר D, תחום הפונקציה, הוא קבוצה (לרב סופית) שמוגדר עליה סדר חלקי, ו־R, הטווח, הוא קבוצה עם סדר לינארי (לרב זו תהיה הקבוצה

$\{1,\ldots,k\}$ עבור k כל שהוא או אפילו רק $\{0,1\}$, אבל כעיקרון אפשר אפילו לקחת את קבוצת כל המספרים הממשיים). כדאי להזכיר: סדר חלקי הוא יחס דו־מקומי "$\leq$" המוגדר מעל D, כך שלכל x,y מתקיים $x=y$ אם ורק אם $x \leq y$ ו־$y \leq x$ (רפלקסיביות ואנטי־סימטריה), וכן לכל x,y,z אם $x \leq y$ וגם $y \leq z$ אז $x \leq z$ (טרנזיטיביות). פרט ליחס שאנחנו מכירים על המספרים, יש למשל את היחס של "הכלה" על משפחת כל תתי־הקבוצה של קבוצה נתונה, או יחס סדר המכפלה מעל $\{1,\ldots,a\}^d$, כאשר $u \leq v$ אם מתקיים $u_i \leq v_i$ לכל $1 \leq i \leq d$ (עבור $a=2$, שבד"כ מסמנים $\{0,1\}^d$, זה שקול ליחס ההכלה על תתי־קבוצה).

זוג מפר הוא זוג $x,y \in D$ שעבורם מתקיים $x < y$ (ללא שוויון) אולם $f(x) > f(y)$. במילים אחרות, זוהי הוכחה לכך ש־f אינה מונוטונית. בהתאם לכך מגדירים גם את גרף ההפרות, שקבוצת הצמתים שלו היא D וקבוצת הקשתות שלו היא קבוצת הזוגות המפירים של f.

בהרבה מקרים, אלגוריתם לבדיקת מונוטוניות יינתן כ"דגימת זוגות": מגדירים מרחב הסתברות μ מעל קבוצת הזוגות $\{x,y \in D : x < y\}$, ומראים שאם f היא ϵ־רחוקה ממונוטוניות, אז בהסתברות לפחות δ (שתלוי ב־ϵ ולרב גם ב־$|D|$), זוג שנבחר לפי μ יהיה זוג מפר. אפשר להפוך אלגוריתם דגימת זוגות לבדיקה "רגילה" (שדוחה בהסתברות לפחות $2/3$ פונקציה ϵ־רחוקה ממונוטוניות) ע"י כך שמגרילים באופן ב"ת $2/\delta$ זוגות לפי μ, ודוחים אם לפחות אחד מהם הוא זוג מפר.

את האלגוריתם שתואר למעלה עבור $D=\{1,\ldots,n\}$ אפשר להפוך (באופן קצת לא טבעי) לאלגוריתם זוגות: במקום לבדוק את כל מסלול החיפוש הבינארי עבור $f(i)$, בוחרים באופן יוניפורמי נקודה עליו (מתוך $\lceil \log(n) \rceil$ הנקודות האפשריות) ובודקים את הערך הזה בלבד מול $f(i)$. כאן יתקיים $\delta(\epsilon,n)=\Omega(\epsilon/\log(n))$.

מונוטוניות של פונקציה בוליאנית

כאן אנחנו נתמקד בבדיקה שפונקציה $f:\{0,1\}^n \to \{0,1\}$ היא מונוטונית. התוצאה הזו הופיעה לראשונה במאמר Goldreich, Goldwasser, Lehman, Ron, Samorodnitsky: Testing monotonicity. הבדיקה עצמה היא באמצעות דגימת זוגות: אנחנו נגריל (באופן יוניפורמי וב"ת מכל הזוגות האפשריים) זוג $x < y$ שעבורו קיים i יחיד עם $x_i < y_i$, כשבשאר הקורדינטות יש שוויון. שיטה אחרת להסתכל על זה: מגרילים באופן יוניפורמי קורדינטה $1 \leq i \leq n$ ווקטור $z \in \{0,1\}^n$, ואז x ו־y יהיו הווקטורים המתקבלים מהחלפת הקורדינטה ה־i של z ב־0 ו־1 בהתאמה. הטענה היא שאם f היא ϵ־רחוקה ממונוטוניות, אז הזוג $x < y$ יהיה מפר בהסתברות שמקיימת $\delta \geq \epsilon/n$. מכיוון שגודל הקלט כולו הוא 2^n, זוהי הסתברות גדולה יחסית.

על מנת להוכיח שהבדיקה הזו עובדת, נסתכל על כמות "חוסר המונוטוניות" בכל קורדינטה לחוד. נסמן ב־$n_i(f)$ את מספר הזוגות $x < y$ שנבדלים אך ורק על הקורדינטה ה־i שעבורם $f(x)=1$ בעוד $f(y)=0$, ונסמן $\delta_i(f)=n_i(f)/2^{n-1}$ (החלוקה היא במספר כל הזוגות הנבדלים על הקורדינטה ה־i, מפרים או לא). כאשר ברור על איזו פונקציה אנחנו מדברים נשמיט ונסמן פשוט δ_i. בפרט מתקיים $\delta=\frac{1}{n}\sum_{i=1}^n \delta_i$, כאשר נסמן ב־$\delta$ את ההסתברות לדחיה ע"י הרצה בודדת של אלגוריתם הזוגות.

הדבר הראשון שנוכיח הוא שאם $\delta=0$ (ז"א שכל ה־δ_i שווים ל־0), אז אין זוגות מפרים כלל, גם כאלה שנבדלים ביותר מקורדינטה אחת: אם $x < y$, אז נגדיר לכל $0 \leq j \leq n$ את הווקטור $x^{(j)}$ כך ש־$x_i^{(j)}=x_i$ אם $i \leq j$, ו־$x_i^{(j)}=y_i$ אם $i > j$. מתקיים תמיד $x=x^{(0)} \leq x^{(1)} \leq \cdots \leq x^{(n)}=y$, ולכל j עבורו $x^{(j-1)} < x^{(j)}$ המדובר בזוג שנבדל רק על קורדינטה בודדת. על כן, אם $\delta=0$, מתקיים גם $f(x)=f(x^{(0)}) \leq f(x^{(1)}) \leq \cdots \leq f(x^{(n)})=f(y)$, כנדרש.

עתה נסתכל על פעולת ההזזה shift בכל קורדינטה: נסמן ב־$S_j f$ את הפונקציה הבאה. מחלקים את $\{0,1\}^n$ ל־2^{n-1} הזוגות של ווקטורים הנבדלים על הקורדינטה ה־j. נסמן את קבוצת הזוגות הזו ב־P_j. לכל $(x,y) \in P_j$ (כאשר $x < y$), אם $f(x)=1$ ו־$f(y)=0$ אז נגדיר את $(S_j f)(x)=0$ ואת $(S_j f)(y)=1$. בכל מקרה אחר נגדיר עבור הזוג הזה את $(S_j f)(x)=f(x)$ ואת $(S_j f)(y)=f(y)$. אפשר לחשוב על זה כעל "הפעלת גרביטציה במימד ה־i", כאשר כל ערך של 1 הנמצא "מעל" ערך של 0 "שוקע למטה".

הטענה המרכזית היא שלכל פונקציה g ולכל $i \neq j$ מתקיים $\delta_i(S_j g) \leq \delta_i(g)$. על מנת להראות זאת, ראשית נסתכל על המקרה $n=2$, ועל $i=1$ ו־$j=2$. בעצם אנחנו רוצים להראות ש"מיון" העמודות של מטריצת אפס/אחד ריבועית מגודל 2 לא מגדיל את מספר השורות הלא־ממויינות. פשוט עוברים על כל המקרים של מיון עמודות של מטריצה כזו – זה אינו מספר גדול של מקרים.

עתה ננתח את המקרה הכללי: נחלק שוב את $\{0,1\}^n$, הפעם ל־2^{n-2} רביעיות, חלוקה שנסמן ב־Q_{ij}. עבור כל $(w,x,y,z)\in Q_{ij}$ המדובר יהיה בארבעה וקטורים אשר מסכימים ביניהם על כל הקורדינטות פרט לקורדינטות ה־i וה־j (יש ארבעה וקטורים סה"כ כי לכל קורדינטה כזו יש שני ערכים אפשריים). האבחנה המרכזית היא שניתן לבצע את הפעולה S_j על כל רביעיה מתוך Q_{ij} לחוד (היא כוללת שני זוגות מתוך P_j שאפשר לחשוב עליהם כעל "עמודות" של מטריצה), וכן ניתן לבצע את הספירה עבור $n_i(g)$ ו־$n_i(S_jg)$ לחוד על כל רביעיה כזו (שכוללת שני זוגות מ־P_i כ"שורות" של המטריצה). על כן אפשר להחיל את המקרה הפרטי של $n=2$ על כל רביעיה כזו, ולקבל שמתקיים $n_i(S_jg)\le n_i(g)$ מהסכום המתאים מעל Q_{ij}. חלוקת אי השוויון ב־2^{n-1} תתן את הטענה עבור δ_i.

מכאן אפשר להוכיח את נכונות אלגוריתם בדיקת הזוגות באמצעות סידרה של טענות על הפעלת כל n פעולות ההזזה ע"פ סידרן על פונקצית הקלט המקורית f. ראשית נשים לב שהפונקציה $h=S_nS_{n-1}\cdots S_1f$ היא בעצמה מונוטונית: לכל g מתקיים $\delta_i(S_ig)=0$, ומכיוון שהפעולה S_j עבור $j>i$ לא תגדיל חזרה את δ_i אם הוא היה כבר אפס, מתקיים $\delta_i(h)=0$ לכל $1\le i\le n$, ז"א ש־h היא מונוטונית.

כמו כן, המרחק בין h ל־f הוא לכל היותר $\sum_{i=1}^n\delta_i(f)$, ולכן זהו חסם עבור המרחק של f ממונוטוניות. הסיבה לכך היא שלכל פונקציה g, המרחק בינה לבין S_jg הוא $\delta_j(g)$ בדיוק, לפי הגדרת פעולת ההזזה. במקרה שלנו המרחק בין $S_{i-1}\cdots S_1f$ לבין $S_iS_{i-1}\cdots S_1f$ הוא $\delta_i(S_{i-1}\cdots S_1f)$, ולפי הטענה מקודם על אי ההגדלה של δ_i, מרחק זה חסום ע"י $\delta_i(f)$. לבסוף, מאי שוויון המשולש, המרחק מ־f ל־h חסום ע"י סכום המרחקים $\sum_{i=1}^n d(S_iS_{i-1}\cdots S_1f,S_{i-1}\cdots S_1f)\le\sum_{i=1}^n\delta_i$.

קיבלנו את החסם $\epsilon\le\sum_{i=1}^n\delta_i=n\delta$, ז"א שההסתברות לגלות הפרה בפונקציה שהיא ϵ־רחוקה ממונוטוניות היא לפחות ϵ/n. קיימת דוגמה שבה לאלגוריתם הזה יש הסתברות שגיאה של $\Omega(\epsilon/n)$. עבור $\epsilon=\frac12$ הדוגמה הזו היא הפונקציה $f(x)=\neg x_i$ עבור i קבוע שרירותי, ולא קשה להתאים אותה לדוגמה עבור $\epsilon\le\frac12$ כל שהוא.

לאחר הרבה שנים נמצא אלגוריתם יותר מתוחכם שמבצע את בדיקת המונוטוניות הזו ב־$\tilde O(\sqrt n/\epsilon^2)$ שאילתות, במאמר Khot, Minzer, Safra: On monotonicity testing and Boolean isoperimetric type theorems. יש גם (עבור ϵ קבוע) חסם תחתון של $\tilde\Omega(n^{1/3})$ שאילתות (כאן הסימון מעיד שיש חלוקה בחזקה קבועה כל שהיא של $\log(n)$), במאמר Chen, Waingarten, Xie: Beyond Talagrand functions: New lower bounds for testing monotonicity and unateness.

ניתוח מקרים כלליים לפי גרף ההפרות

עבור $f:D\to R$, כאשר D הוא סדר חלקי סופי ו־R הוא סדר לינארי, נגדיר את גרף ההפרות G_f כגרף שקבוצה הצמתים שלו היא D וקבוצת הקשתות שלו היא הקבוצה $\{x,y\in D:x<y\wedge f(x)>f(y)\}$ של כל הזוגות המפרים את המונוטוניות של f. לרב נסתכל על זה בעל גרף לא מכוון (ממילא ה"כיוון" של כל קשת נקבע ע"י D ואינו תלוי ב־f).

המרחק של f ממונוטוניות כרוך בגרף ההפרות שלו. ליתר דיוק, הוא זהה לגודל כיסוי הצמתים המינימלי של G_f, מחולק ב־$|D|$ (כיסוי צמתים הוא קבוצת צמתים שכוללת את לפחות אחד הצמתים של כל קשת בגרף). נוכיח את זה עתה.

כיוון ראשון: אם $f':D\to R$ היא פונקציה מונוטונית, נראה ש־$D'=\{x:f'(x)\neq f(x)\}$ הוא בפרט כיסוי צמתים של G_f, ולכן מספר השינויים הדרוש להפוך את f למונוטונית הוא לפחות מספר הכיסוי של הגרף. אם xy קשת בגרף ההפרות, אז לא יתכן ש־f' זהה ל־f על שני הצמתים האלו, כי אז היינו מוצאים זוג מפר ל־f' והנחנו שאין כזה. על כן לפחות אחד מהצמתים האלו נמצא ב־D'.

כיוון שני: נניח ש־D' כיסוי לגרף ההפרות של f. נגדיר $f'=f|_{D\setminus D'}$. זוהי פונקציה מונוטונית על התחום שלה (לא יכולים להיות לה זוגות מפרים שאינם מפרים את f, אולם אין זוגות כאלו שמוכלים בתחום של f'). נראה עתה איך אפשר להרחיב אותה לאיבר נוסף מ־D' ולשמור על המונוטוניות. מכך נובע (באינדוקציה על מספר האיברים ב־D שאינם בתחום של f') שאפשר לבסוף להרחיב את f' לפונקציה מונוטונית מעל כל D, והיא יכולה להיבדל מ־f רק על הכיסוי D' (או תת־קבוצה שלו).

על מנת להוכיח זאת, נבחר $x \in D'$ שהוא איבר מינימלי שם (אין $z \in D'$ המקיים $z < x$). אם x מינימלי גם ב־D, אז נבחר את $f'(x)$ להיות המינימום $\min_{z \in D \setminus D'} f'(z)$ (אפשר להניח שמתקיים $D' \neq D$ כי אפילו קליק ניתן לכיסוי ע"י קבוצת כל הצמתים שלו פרט לאחד מהם). ברור עתה ש־x לא יכול להיות האיבר הנמוך בזוג מפר ($x < y$ לפי הסדר של D), כי הערך שלו אינו גדול מאף ערך אחר של f', והוא לא יכול להיות גם האיבר הגבוה בזוג מפר ($y < x$ לפי הסדר של D), פשוט כי אין איבר נמוך ממנו ב־D.

אם x אינו מינימלי ב־D, אז נבחר את $f'(x)$ להיות המקסימום של ערכי הפונקציה f עבור האיברים שמתחתיו, $\max_{\{z \in D \setminus D' : z < x\}} f'(z)$. האיבר x לא יכול להיות האיבר הגבוה בזוג מפר כי הוא נבחר להיות המקסימלי מכל הערכים הרלוונטים. אם לעומת זאת x היה האיבר הנמוך בזוג מפר xy (כאשר y נמצא ב־$D \setminus D'$), אז מכיוון שערכו זהה לאחד מהערכים בביטוי המקסימום, קיים $z \in D \setminus D'$ עבורו $z < x < y$ וגם $f'(z) = f'(x) > y$. מכאן ש־zy הוא זוג מפר עבור f' המקורית, בסתירה להנחות.

הקשר בין המרחק לבין הכיסוי של גרף ההפרות נותן לנו כלי חזק לניתוח. בפרט, שימו לב לקשר לגודל הזיווג (קבוצת קשתות זרות צמתים) המקסימלי ב־G_f: גודל הכיסוי המינימלי של גרף הוא תמיד לפחות גודל הזיווג המקסימלי, כי בפרט הוא חייב להכיל לפחות צומת אחד מכל קשת של הזיווג. מצד שני, גודל הכיסוי המינימלי הוא גם לא יותר מכפליים גודל הזיווג המקסימלי, כי אם לוקחים את צמתי הזיווג המקסימלי כולם, בפרט יש לנו כיסוי צמתים של הגרף (המקסימליות של הזיווג אומרת שאין בגרף קשת שזרה לכל צמתיו).

זה אומר שלכל סדר חלקי D באשר הוא, אפשר לכתוב אלגוריתם בדיקה למונוטוניות בעל $O(\sqrt{|D|/\epsilon})$ שאילתות, באופן הבא: נבחר קבוצה Q של שאילתות ע"י כך שכל צומת נבחר להיות ב־Q בהסתברות $2\sqrt{1/\epsilon|D|}$, באופן ב"ת (זה מאפשר ניתוח קל יותר מאשר בחירה יותר "ישירה" של Q). אם בחרנו קבוצה בעלת יותר מ־$12\sqrt{|D|/\epsilon}$ צמתים, נוותר על השאילתות ונקבל את הקלט f מיידית. זה קורה בהסתברות שאינה עולה על $\frac{1}{6}$ לפי אי־שוויון מרקוב (בעצם זה קורה בהסתברות קטנה בהרבה לפי חסימת סטיות גדולות). אם Q אינה גדולה מדי, אז נשאל את כל הערכים של $f|_Q$ ונבדוק אם יש זוג מפר בכל אלו.

אם f היא ϵ־רחוקה ממונוטניות, אז לפי הדיון למעלה על גודל הזיווג המקסימלי, יש ב־D קבוצה של לפחות $\epsilon|D|/2$ זוגות מפרים זרים זה לזה. הסיכוי שהקבוצה Q שבחרנו אינה כוללת זוג כזה חסום ע"י $(1 - (2\sqrt{1/\epsilon|D|})^2)^{\epsilon|D|/2} < e^{-(4/\epsilon|D|)\epsilon|D|/2} < 1/6$. מאיחוד מאורעות, הסיכוי שלא הצלחנו לגלות זוג מפר (כי Q היתה גדולה מדי או כי Q לא הכילה זוג מהקבוצה הזו) חסום ע"י $\frac{1}{3}$.

ובחזרה לדוגמה הראשונה

נראה עתה שיטה אלטרנטיבית לבדיקת מונוטוניות של פונקציה $f : \{1, \ldots, n\} \to \mathbb{N}$, המבוססת על ניתוח גרף ההפרות. ההוכחה יותר "מסובכת" (היא דורשת את ההגדרות מלמעלה), אבל השיטות כאן יותר נוחות להכללות. האינטואיציה המרכזית היא שעבור כל זוג מפר, לפחות לאחד הצמתים שלו יש סביבה עם ריכוז גבוה של "בני זוג מפרים". על כן הצמתים עם "סביבת ריכוז גבוה" יהוו כיסוי לגרף ההפרות, ולכן צומת מוגרל מקרית יהיה כזה בהסתברות לפחות ϵ. נעבור להוכחות פורמליות.

נניח ש־$i < j$ הוא זוג מפר, ז"א $f(i) > f(j)$. לכל $i < k < j$ (אם יש k כזה), חייב אם כן להתקיים או $f(i) > f(k)$ או $f(k) > f(j)$ (או שניהם), מטרנזיטיביות. על כן, לפחות אחד מ־i או j יפר את המונוטוניות עם לפחות חצי מערכי k האפשריים. מכאן נובע שעבור $s = i$ או $s = j$ קיים $l \in \mathbb{N}$, כך שלפחות $l/2$ מהצמתים של הסביבה שלו $\{s-l, \ldots, s+l\} \setminus \{s\}$ הם מפרים עם s (לא נדאג לאינדכסים "מחוץ לתחום" של $\{1, \ldots, n\}$, אם יש כאלה אז נניח ש"שאילתה" מתוכם מוציאה ערך שאינו חלק מזוג מפר).

לפני שנוכל לדגום, נשים לב שיש יותר מדי ערכים אפשריים עבור l. אפשר אבל לצמצם אותם ללוגריתם המוכר באמצעות בדיקת חזקות של 2 בלבד. נשים לב שאם נחליף את l מלמעלה ב־r הקטן ביותר שעבורו $2^r \geq l$, אז יהיו לפחות $l/2$ צמתים מפרים עם s ב־$\{s - 2^r, \ldots, s + 2^r\} \setminus \{s\}$, ומכיוון שגודל הקבוצה הוא לכל היותר $4l$, בחירה מקרית יוניפורמית של צומת מהקבוצה תתן צומת מפר בהסתברות לפחות $\frac{1}{8}$.

עכשיו אפשר לכתוב את האלגוריתם לבחירת זוג עבור בדיקת מונוטוניות.

- ראשית, נבחר צומת $s \in \{1, \ldots, n\}$ באופן מקרי ויוניפורמי. אם f היא ϵ־רחוקה ממונוטוניות, אז בסיכוי לפחות ϵ בחרנו את הצומת ה"נכון" (צומת עם סביבה מפרה) מתוך זוג מפר. הסיבה היא שלכל

זוג מפר יש לפחות צומת אחד כזה, ולכן קבוצת הצמתים הרצויים מהווה כיסוי לגרף ההפרות. כזכור, המרחק של פונקציה ממונוטונית שווה לגודל כיסוי הצמתים המינימלי של גרף ההפרות, ובפרט מהווה חסם תחתון לגודל כאן.

- עתה נבחר באופן מקרי ויוניפורמי $r \in \{0, \ldots, \lceil \log n \rceil\}$. אם s היה צומת עם סביבה מפרה בגודל l, בהסתברות לפחות $1/\log(n)$ בחרנו את ה־r המינימלי כך ש־$2^r \geq l$.
- לבסוף נבחר $k \in \{s - 2^r, \ldots, s + 2^r\} \setminus \{s\}$ באופן מקרי ויוניפורמי. אם s ו־r נבחרו טוב, אז בהסתברות לפחות $\frac{1}{8}$ קיבלנו ש־k ו־s מהווים זוג מפר.

סה"כ, אם f היא פונקציה ϵ־רחוקה, אז (לפי הסתברויות מותנות) יש לנו סיכוי של לפחות $\epsilon/8\log(n)$ לקבל בצורה זו זוג מפר. בפרט אפשר לבצע ϵ־בדיקה (עם הסתברות הצלחה $\frac{2}{3}$) באמצעות $O(\log(n)/\epsilon)$ שאילתות (ע"י ביצוע של $16\log(n)/\epsilon$ סבבי דגימה כאלו באופן ב"ת).

הכללה של השיטה הזו לבדיקת מושג של "כמעט מונוטוניות" נמצאת במאמר הבא (זו לא טעות שיש שני "פישר" ברשימת המחברים – השני הוא אחי): Ben Moshe, Fischer, Fischer, Kanza, Matsliah Staelin: Detecting and exploiting near-sortedness for efficient relational query evaluation.

מבוא לשיטת יאו לחסמים תחתונים

חסמים תחתונים רבים על אלגוריתמי בדיקה נעשים באמצעות השיטה של יאו Yao, שמאפשרת לעבור לניתוח של אלגוריתמים דטרמיניסטים. על מנת לראות את זה, צריך לחשוב על אלגוריתם הסתברותי כמרחב הסתברות שבו מגרילים אלגוריתם דטרמיניסטי (לאו דווקא אחד בעל "תיאור" קצר): בכל רגע נתון (לקראת ביצוע שאילתה או החלטה אם לקבל או לדחות את הקלט) האלגוריתם מבצע את ההחלטה הבאה בהסתמך על מרחב הסתברות מתאים. אפשר להניח שההגרלה על ההחלטה הבאה תלויה רק במקומות שנשאלו עד עכשיו ובתשובות עליהם: אם באיזה שהוא שלב יש תלות בהגרלה שבוצעה קודם שלא השפיעה כבר על הבחירה של השאילתות הקודמות, אפשר לבצע אותה בשלב שבו היא משפיעה (באופן פורמלי, מבצעים הגרלה לפי ההתפלגות המותנה המתאימה). אנחנו מניחים כאן שהתחום והטווח של הפונקציה f הם קבוצות בדידות, על מנת לא להזקק לכלים "כבדים" מתורת ההסתברות.

עתה נניח שמבצעים מראש את כל ההגרלות לכל מצבי הביניים האפשריים (סדרת שאילתות מאורך חסום ע"י q וסדרת התשובות המתאימות), ורק אז מבצעים את האלגוריתם לפיהם. עבור כל סידרת הגרלות אפשרית שאנחנו יכולים לקבוע, האלגוריתם יהיה עתה דטרמיניסטי (הוא יהיה תלוי רק בתשובות לשאילתות). על כן יש לנו מרחב הסתברות מעל אלגוריתמים דטרמיניסטים.

בשלב זה נניח שיש לנו מרחב הסתברות מעל קלטים אפשריים (גם כאלו שמקיימים את התכונה וגם כאלו שלא), כך שלכל אלגוריתם דטרמיניסטי אפשרי, ההסתברות לטעות גדולה מ־$\frac{1}{3}$ (אצלינו "טעות" היא קבלה של קלט ϵ־רחוק מהתכונה, או דחיה של קלט מקיים; במידה ומוגרל קלט שאינו בתכונה אבל אינו ϵ־רחוק ממנה, כל תשובה של האלגוריתם תחשב לנכונה). במקרה כזה, גם אם ניקח אלגוריתם הסתברותי $\mathcal{A}$ (כמרחב הסתברות מעל אלגוריתמים דטרמיניסטים) ונזין לו קלט ממרחב ההסתברות שלנו, ההסתברות לטעות תהיה גדולה מ־$\frac{1}{3}$. על כן, לכל $\mathcal{A}$ כזה יהיה קלט ספציפי כך שהוא יטעה בהסתברות לפחות $\frac{1}{3}$: מסתכלים על ההסתברות לטעות של $\mathcal{A}$ כמשתנה מקרי התלוי בקלט שהוגרל, ומשתמשים בכך שחייב להיות איבר ממרחב ההסתברות שעבורו ערך המשנה המקרי הוא לפחות ערך התוחלת. כך קיבלנו את הוכחת אי ההתכנות שהיינו צריכים עבור אלגוריתמים בעלי q שאילתות.

כדאי להעיר כאן שזהו הצד ה"קל" של שיטת יאו. במאמר המקורי הוא הראה (כמסקנה ממשפט הדואליות בתכנות לינארי) גם את הטענה ההפוכה, שאם אין מרחב הסתברות מעל קלטים ש"מביס" את כל האלגוריתמים הדטרמיניסטים המתאימים (המאמר המקורי לא כוון ספציפית לבדיקת תכונות), אז יש אלגוריתם הסתברותי (שאולי אנחנו לא יודעים לרשום) שפותר את הבעיה המתאימה.

שיטה כללית עבור אלגורתמים לא־אדפטיבים

אלגוריתם בדיקה לא־אדפטיבי הוא אלגוריתם שחייב לבצע את כל השאילתות שלו לפני שהוא מקבל ערכים כל שהם. רק את ההחלטה הסופית אם לקבל או לדחות את הקלט האלגוריתם קובע בהסתמך על המידע שקיבל. הרבה מהאלגוריתמים שראינו עד עכשיו הם לא אדפטיבים. באופן פורמלי, עבור קלט $f : D \to R$, אלגוריתם כזה מתואר ע"י מרחב הסתברות על תתי־קבוצה $Q \subset D$ (כאשר D הוא תחום הקלט), שגודלן הוא q (בלי הגבלת הכלליות אפשר להניח שהגודל תמיד שווה למספר השאילתת המקסימלי q, אחרת מוסיפים שאילתות "מיותרות" ופשוט מתעלמים מהתשובות עליהן), בתוספת פונקצית החלטה $\alpha_Q : R^q \to [0,1]$ לכל Q אפשרי. ריצה טיפוסית של האלגוריתם מורכבת מהגרלה של הקבוצה Q, שאילת כל הערכים של f על Q, ולבסוף קבלה של הקלט בהסתברות $\alpha_Q(f|_Q)$.

הגרסה הדטרמיניסטית של אלגוריתם כזה מורכבת מקבוצה קבועה $Q \subset D$ מגודל q, וקבוצה $A \subseteq R^q$ שמתארת את כל המקרים שבהם האלגוריתם יקבל על סמך ערכי f מעל Q. על מנת לנתח אלגוריתמים כאלו נשתמש במושג של מרחק בין התפלגויות (variation distance). באופן כללי, עבור התפלגויות μ ו־ν מעל קבוצה בדידה של תוצאות אפשריות S, מגדירים $d(\mu,\nu) = \frac{1}{2}\sum_{s\in S}|\mu(s)-\nu(s)|$ (אנחנו נשתמש בסימון מקוצר "$\mu(s)$" עבור $\Pr_\mu[s]$). זה שווה בדיוק למקסימום על ההבדל בהסתברות למאורעות $d(\mu,\nu) = \max_{B\subseteq S}|\Pr_\mu[B]-\Pr_\nu[B]|$ (יש הרחבות למרחבי הסתברות לא בדידים שדורשות יותר ידע מתמטי בשביל הפורמליזם, לא נטפל באלו כעת).

עבור התפלגות μ מעל פונקציות מהצורה $f : D \to R$ ועבור ת"ק $Q \subseteq D$, נסמן ב־$\mu|_Q$ את ההתפלגות מעל פונקציות מהצורה $g : Q \to R$, המתקבלת מהתהליך של בחירת פונקציה f לפי μ ומעבר ל־$g = f|_Q$. עבור אלגוריתמים לא־אדפטיבים בעלי q שאילתות, הוכחה של חסם תחתון שקולה, עד כדי שינוי בקבועים, למציאת שתי התפלגויות עם הפרמטרים הבאים:

- ההתפלגות τ תהיה מעל קלטים $f : D \to R$ שכולם מקיימים את התכונה.
- ההתפלגות ν תהיה מעל קלטים $f : D \to R$ שכולם ϵ־רחוקים מלקיים את התכונה.
- לכל $Q \subset D$ מגודל q, מתקיים $d(\tau|_Q, \nu|_Q) < \frac{1}{3}$.

בהינתן ההתפלגויות האלו, נגדיר את ההתפלגות μ להיות התוצאה של בחירה בהסתברות $\frac{1}{2}$ האם לוקחים קלט מקיים לפי τ או האם לוקחים קלט ϵ־רחוק לפי ν, ואז בחירת הקלט לפי ההתפלגות המתאימה. בסימון מקוצר, $\mu = \frac{1}{2}(\tau+\nu)$ (זהו חישוב ההסתברויות אם מתייחסים לפונקציות המתאימות כאל ווקטורים).

בהינתן אלגוריתם לא־אדפטיבי דטרמיניסטי, המתואר ע"י קבוצת שאילתות Q וקבוצת קבלה $A \subseteq R^q$, שנתייחס אליה כאל מאורע במרחבים $\tau|_Q$ ו־$\nu|_Q$, מהנתון על ההתפלגויות נקבל $|\Pr_\tau[A]-\Pr_\nu[A]| < \frac{1}{3}$. ההסתברות לטעות של האלגוריתם מעל μ היא $\frac{1}{2}(\Pr_\nu[A]+1-\Pr_\tau[A]) > \frac{1}{2}(\Pr_\nu[A]+\frac{2}{3}-\Pr_\nu[A]) = \frac{1}{3}$, מה שמעיד שאין אלגוריתם ϵ־בדיקה לא־אדפטיבי בעל q שאילתות במקרה הזה.

ניתן (והרבה פעמים יהיה נוח) להשתמש בזוג התפלגויות כאשר ν נותנת קלט ϵ־רחוק בהסתברות גבוהה, אבל לא בהסתברות מלאה. כאן משתמשים בפרמטר $\alpha < \frac{1}{3}$.

- ההתפלגות τ תהיה מעל קלטים $f : D \to R$ שכולם מקיימים את התכונה.
- עבור ההתפלגות ν, הסיכוי שיתקבל $f : D \to R$ שאינו ϵ־רחוק מהתכונה הוא לכל היותר α.
- לכל $Q \subset D$ מגודל q, מתקיים $d(\tau|_Q, \nu|_Q) < \frac{1}{3} - \alpha$.

תחת ההתפלגות $\mu = \frac{1}{2}(\tau+\nu)$, הסיכוי שהאלגוריתם יטעה הוא לפחות $\frac{1}{2}(\Pr_\nu[A]-\alpha+1-\Pr_\tau[A]) > \frac{1}{3}$.

עתה נראה ישום: נסתכל על התכונה של כל המילים מעל האלפבית $\{0,1\}$ שהן שרשור של שני פלינדרומים (מותר שאחד מהם יהיה "פלינדרום" מאורך אפס). נראה שאלגוריתם לא־אדפטיבי שמבצע $\frac{1}{5}$־בדיקה יצטרך $\Omega(\sqrt{n})$ שאילתות לפחות. נניח ש־n גדול מקבוע מתאים, ונסתכל על שתי ההתפלגויות הבאות מעל מילים $w = w_1,\ldots,w_n \in \{0,1\}^n$:

- בהתפלגות τ אנחנו בוחרים באופן מקרי ויוניפורמי $1 \le k \le n$, בוחרים את u להיות פלינדרום מקרי ויוניפורמי באורך k (מתוך $2^{\lceil k/2 \rceil}$ האפשרויות), את v להיות פלינדרום מקרי ויוניפורמי באורך $n-k$, ומגדירים את הקלט להיות השרשור $w = uv$.
- בהתפלגות ν אנחנו בוחרים את המילה $w \in \{0,1\}^n$ באופן מקרי ויוניפורמי (זה כמו לבחור כל אות $w_i \in \{0,1\}$ באופן יוניפורמי וב"ת בכל הבחירות האחרות). ההתפלגות הזו נותנת קלט שהוא $\frac{1}{5}$־רחוק מהתכונה בהסתברות $1-o(1)$: לכל $1 \le k \le n$ קבוע, מספר השינויים שצריך כדי להפוך את w לשרשור של פלינדרום מאורך k ופלינדרום מאורך $n-k$ הוא סכום של לפחות $\lfloor \frac{n-1}{2} \rfloor$ משתנים יוניפורמים ב"ת מ־$\{0,1\}$ (סופרים את מספר הזוגות שצריכים להיות "תואמים"; בפלינדרום מאורך אי־זוגי האיבר האמצעי הוא ללא בן־זוג). לפי חסימת סטיות גדולות, ההסתברות שהסכום יהיה קטן מ־$\frac{n}{5}$ חסומה ע"י $e^{-(1-o(1))n/100}$. יש n ערכי k אפשריים, ולכן לפי איחוד מאורעות הסיכוי שקיים k כל שהוא המתאים למרחק קטן מ־$\frac{1}{5}$ הוא $o(1)$.

נותר אם כן לנתח את $d(\tau|_Q, \nu|_Q)$ עבור קבוצה Q שגודלה הוא $q \le \frac{1}{2}\sqrt{n}$. נשים לב שההתפלגות $\nu|_Q$ היא פשוט ההתפלגות היוניפורמית על מילים מאורך $|Q|$. עבור ניתוח $\tau|_Q$, ראשית נתבונן בזוג אינדכסים $i < j$. הסיכוי שיוגרל k שעבורו חייב להתקיים $w_i = w_j$ הוא $\frac{1}{n}$: אם $i + j \le n + 1$, אז ה־k היחידי שיגרום לתיאום הוא זה שעבורו $i = k + 1 - j$, ז"א $k = i + j - 1$. אם $i + j > n + 1$, אז צריך להתקיים $(n + 1 - i) = (n + 1 - k) - (n + 1 - j)$, ז"א $k = i + j - n - 1$. בכל מקרה יש k יחיד שמצריך את התיאום, כך שהסיכוי לבחור דווקא אותו הוא $\frac{1}{n}$.

הסיכוי שנבחר k כך שיש $i < j$ כל שהם בתוך Q שחייבים להיות מתואמים (לפי τ) הוא לכל היותר $\binom{q}{2}/n < \frac{1}{8}$. במידה והמאורע הזה לא קרה, הרי שההתפלגות המותנה המתאימה של $\tau|_Q$ (לכל k אחר) זהה לזו של $\nu|_Q$. על כן מתקיים $d(\tau|_Q, \nu|_Q) < \frac{1}{8}$, ובפרט מתקיימים כל התנאים על ההתפלגויות שלנו על מנת להוכיח ש־$\frac{1}{2}\sqrt{n}$ שאילתות אינן מספיקות עבור $\frac{1}{3}$־בדיקה של התכונה הנ"ל.

ראוי להעיר כאן שעבור התכונה הזו קיים אלגוריתם ϵ־בדיקה לא־אדפטיבי שמבצע $O(\sqrt{n \log n/\epsilon})$ שאילתות, כך שהחסם התחתון הנ"ל קרוב לאמת.

בהמשך הקורס נראה איך אפשר להרחיב את הטיעונים בחלק מהמקרים, כולל המקרה הזה, לאלגוריתמים אדפטיבים. החשיבות בתכונה הספציפית הזו היא שזהו מקרה של דקדוק חסר הקשר. לעומת זאת, כל השפות הרגולריות כן ניתנות לבדיקה ע"י מספר שאילתות התלוי רק ב־ϵ (ובשפה עצמה), לפי המאמר Alon, Krivelevich, Newman, Szegedy: Regular languages are testable with a constant number of queries.

חסם תחתון על בדיקת מונוטוניות

בבדיקה של פונקציה $f : \{1, \ldots, n\} \to \mathbb{N}$ למונוטוניות חייבים $\Omega(\log(n))$ שאילתות עבור $\frac{1}{4}$־בדיקה. במאמר Fischer: On the strength of comparisons in property testing יש רדוקציה של בדיקת מונוטוניות של פונקציה עם טווח בלתי מוגבל לאלגוריתמים שהם "מבוססי סדר" בלבד. אלו אלגוריתמים שמבססים את ההחלטות הבאות שלהם אך ורק על סמך הסדר בין הערכים שהתקבלו עד כה (מי גדול ממי, ומי שווה למי), ולא על הערכים עצמם. כאן לא נדון ברדוקציה עצמה (אתם מוזמנים לקרוא עליה – המדובר בשימוש לא צפוי במשפט רמזי), אלא בחסם התחתון שאפשר להשיג לאלגוריתמים מבוססי סדר עבור בעיית הבדיקה.

החסם התחתון הופיע בצורה מסויימת כבר במאמר הראשון שדן בבדיקת מונוטוניות (Spot checkers). על מנת להוכיח את החסם, נתמקד אך ורק בפונקציות שכל ערכיהן שונים זה מזה. זה אומר שאנחנו לא דורשים כלום מהאלגוריתם במקרה שהוא נתקל בשוויון, ולכן נתעלם מההתנהגות שלו במקרים כאלה (במילים אחרות, החסם יהיה תקף אפילו אם אנחנו מנמיכים את הדרישות שהאלגוריתם חייב לקיים). זה מאפשר לנו להניח הנחה מפשטת שמזכירה את זו של הבדיקה של התכונה "הכל אפסים": ברגע שהאלגוריתם מגלה זוג אינדכסים $i < j$ עבורם $f(i) > f(j)$, הוא יכול לדחות מיידית. על כן מעניינת רק האפשרות היחידה הנותרת, והיא שאם האלגוריתם עד כה שאל את הערכים עבור הקבוצה Q שאיבריה הם $i_1 < \cdots < i_r$, אז הסדר שהוא קיבל הוא $f(i_1) < \cdots < f(i_r)$. במילים אחרות, אפשר להתייחס לאלגוריתם כאל אלגוריתם לא־אדפטיבי.

על כן אפשר להסתכל על האלגוריתם כעל מרחב הסתברות מעל ת"ק $Q \subset \{1,\ldots,n\}$ מגודל חסום ע"י מספר השאילתות המקסימלי q. אלגוריתם דטרמיניסטי אם כן יתואר פשוט ע"י קבוצה אחת $Q \subset \{1,\ldots,n\}$, כאשר האלגוריתם דוחה אם סדרת הערכים של f מעל Q אינה עולה. במקרה היחידי שנותר, כאשר סדרת הערכים אכן עולה, יש שתי אפשרויות עבור האלגוריתם, קבלה או דחיה. דחיה אבל לא תהיה אפשרית אצלנו, כי זה יהיה המצב כאשר האלגוריתם יקבל קלט מונוטוני־עולה, ובהתפלגות שלנו זה יתקיים בהסתברות $\frac{1}{2}$: באופן אנלוגי לשיטה המשתמשת בזוג התפלגויות, אצלנו τ תהיה "התפלגות" שתמיד בוחרת בפונקציה מונוטונית עולה ממש, בעוד ש־ν תהיה התפלגות מעל קלטים רחוקים עם התכונה ש־$f|_Q$ תהיה בהסתברות גבוהה תת־סדרה מונוטונית עולה ממש. ההבדל מתת־הפרק הקודם הוא שאצלנו מנתחים רק את יחסי הסדר בין ערכי $f|_Q$ (כהתפלגות מעל $q!$ הסדרים האפשריים ללא שוויונים), ולא את הערכים עצמם, כי הנחנו שהאלגוריתם הוא מבוסס סדר.

לסיכום, עבורנו אלגוריתם מבוסס סדר דטרמיניסטי עבור המונוטוניות של $f : \{1,\ldots,n\} \to \mathbb{N}$ יתואר ע"י קבוצת השאילתות $Q \subset \{1,\ldots,n\}$, כאשר הקלט יתקבל אם ורק אם $f|_Q$ נותנת סדרת ערכים עולה.

ההתפלגות המקבילה ל־ν תשתמש בפונקציות "מסור". נסתכל על הפונקציה הבאה: נתחיל מפונקצית הזהות מעל קבוצת הטבעיים $\mathbb{N}$, ואז עבור פרמטר טבעי k, לכל קטע מהצורה $\{2l2^k+1,\ldots,2(l+1)2^k\}$, נחליף בהתאמה בין הערכים של $\{2l2^k+1,\ldots,(2l+1)2^k\}$ ואלו של $\{(2l+1)2^k+1,\ldots,2(l+1)2^k\}$. נקרא לפונקציה הזו s_k. באופן יותר מדוייק, עבור $1 \le r \le 2^k$ ועבור כל l, נגדיר $s_k(2l2^k+r) = (2l+1)2^k+r$ וכן $s_k((2l+1)2^k+r) = 2l2^k+r$.

ננתח את הפונקציה: הפונקציה $s_k : \mathbb{N} \to \mathbb{N}$ אינה מונוטונית, ולמעשה גרף ההפרות שלה כולל זיווג מושלם: קבוצת כל הזוגות $\{2l2^k+r,(2l+1)2^k+r\}$ לכל $l \ge 0$ ולכל $1 \le r \le 2^k$. מצד שני, במובן מסויים אין הרבה "מרווחים" שבהם ניתן לגלות חוסר מונוטוניות. עבור $i < j$, אם $j-i \ge 2^{k+1}$ אז תמיד $s_k(i) < s_k(j)$. אם $j-i < 2^{k+1}$, אז יתקיים $s_k(i) > s_k(j)$ רק אם קיים l שעבורו $2l2^k < i \le (2l+1)2^k < j \le 2(l+1)2^k$.

עתה נגדיר את התפלגות שלנו מעל פונקציות $f : \{1,\ldots,n\} \to \mathbb{N}$. אנחנו מניחים כאן ש־$n$ גדול מקבוע מתאים (הקבוע 100 יספיק). ההתפלגות תוגדר לפי $\frac{1}{2}(\tau+\nu)$ עבור τ ו־ν מתאימים, בדומה לתת־הפרק הקודם.

- בהתפלגות τ, ניקח (בהסתברות 1) את פונקצית הזהות $f(x) = x$ (שהיא מונוטונית). כזכור, זה גורם לכל אלגוריתם דטרמיניסטי שלא מקבל במקרה שהתשובות לשאילתות שלו מהוות סידרה מונוטונית להכשל בהסתברות $\frac{1}{2}$ (זאת ההסתברות שנבחר קלט לפי τ), ולכן מעתה עלינו לנתח רק אלגוריתמים שמקבלים במקרה זה.
- בהתפלגות ν, נגריל באופן יוניפורמי $1 \le k \le \log(n)-2$, נגריל באופן יוניפורמי $0 \le t \le 2^{k+1}-1$, ונקבע $f(x) = s_k(x+t)$. נשים לב שגרף ההפרות של f כולל זיווג בגודל $\frac{1}{2}n$ לפחות (הוא מושרה מהזיווג המושלם של הפרות s_k, שהזוגות בו הם עם הפרש אינדכסים חסום ע"י $\frac{n}{4}$, ולכן לפחות לחצי מהאינדכסים בתחום של f יש בן־זוג שגם הוא בתחום). לכן תמיד נקבל כאן פונקציה $\frac{1}{4}$־רחוקה ממונוטוניות.

עתה ננתח את ההסתברות שהצמצום של f לקבוצה Q אינו מונוטוני (כזכור אפשר כבר להניח שהאלגוריתם הדטרמיניסטי מקבל במידה והצמצום כן מונוטוני). נראה שאם $q < \frac{1}{24}\log(n)$, אז עבור בחירה מקרית של פונקציה מהצורה $f(x) = s_k(x+t)$ ההסתברות הזו קטנה מ־$\frac{1}{3}$, ולכן האלגוריתם הדטרמיניסטי יטעה (הפעם בגלל המקרים עם הפונקציות הרחוקות) בהסתברות גדולה מ־$\frac{1}{3} = \frac{1}{2} \cdot (1-\frac{1}{3})$.

נסמן $Q = \{i_1,\ldots,i_q\}$ כאשר $i_1 < \cdots < i_q$, ונשים לב שעל מנת שהצמצום של f ל־Q לא יהיה מונוטוני, חייב להיות $1 \le j < q$ שעבורו $f(i_j) > f(i_{j+1})$. נגדיר את המאורע B_j שאכן זה מתקיים. יש לנו סה"כ $q-1$ מאורעות כאלו, ונחסום את הסיכוי לקיום של מי מהם באמצעות החסם על איחוד מאורעות (אם אף מאורע לא מתקיים, אז הצמצום של f הוא אכן מונוטוני).

על מנת לחסום את הסיכוי למאורע בודד, ראשית "נפטר" לשם הבהירות מעודף האינדכסים: נקבע $1 \le i < j \le n$, נגריל k ו־t כפי שעשינו למעלה, ונגדיר את B להיות המאורע $s_k(i+t) > s_k(j+t)$. נחסום

את הסיכוי ש־B מתקיים. נסמן $k' = \lceil \log(j-i) \rceil$. כאשר מתנים על בחירה ספציפית של k, אם $k' > k+1$ אז יתקיים $i - j > 2^{k+1}$ והמאורע B בטוח לא יתקיים. אם $k' \leq k+1$, אז נשים לב לבחירה היוניפורמית של t. ההסתברות שהמאורע B יתקיים תהיה לכל היותר $2^{k'-k} \leq (j-i)2^{-k-1}$. עתה אפשר לחסום את ההסתברות הלא־מותנה ש־B יתקיים ע"י $\frac{1}{\lfloor \log(n) \rfloor - 2} \sum_{k=k'-1}^{\lfloor \log(n) \rfloor - 2} 2^{k'-k} \leq \frac{2}{\log(n)} \sum_{r=-1}^{\infty} 2^{-r} = \frac{8}{\log(n)}$, כשמניחים $n \geq 64$ בשביל שיתקיים $\lfloor \log(n) \rfloor - 2 \geq \frac{1}{2}\log(n)$.

אם חוזרים לאלגוריתם הדטרמיניסטי השלם, בפרט כאשר $q < \frac{1}{24}\log(n)$, בהסתברות לפחות $1 - \frac{8(q-1)}{\log(n)} > \frac{2}{3}$ אף אחד מהמאורעות $B_1, \ldots, B_{q-1}$ לא יתקיים, ז"א שהצמצום של $s_k(x+t)$ ל־Q יהיה מונוטוני, והוכחת אי ההתכנות שלנו הושלמה.

מודל הגרפים הדליל

מודל הגרפים הדליל הוגדר במאמר Goldreich, Ron: Property testing in bounded degree graphs. המודל הזה מגביל אותנו לגרפים שהדרגה שלהם חסומה ע"י פרמטר d (בד"כ מתייחסים ל־d כאל קבוע). עבור צומת v ואינדכס $1 \leq i \leq d$ ניתן לשאול מהו השכן ה־i של v; אם ל־v יש פחות מ־d שכנים, התשובה עבור ערכי ה־i הגדולים ממספר השכנים בפועל תהיה "$\perp$", סימן מיוחד עבור "אין שכן כזה". גרף יחשב לרחוק מתכונה מסויימת כאשר הוספה ו/או מחיקה של עד ϵdn קשתות לא תהפוך אותו לגרף שמקיים את התכונה (ועדיין יש לו דרגה חסומה ע"י d). אנחנו נחקור רק ערכים של d גדולים מ־2, כי עבור $d = 2$ הגרף חייב להיות איחוד של מסלולים ומעגלים, מה שמאפשר "ללמוד" אותם בקלות יחסית (אם ישאר זמן לקראת סוף הקורס נפרט יותר על "בדיקה באמצעות למידה").

המודל הזה מתאים לייצוג של גרף ע"י רשימת שכנויות עם "אורך קבוע" לכל הצמתים. באופן פורמלי, עבור קבוצת צמתים V, אנחנו בודקים את הפונקציה $f : V \times \{1, \ldots, d\} \to V \cup \{\perp\}$. גם כאן מגבילים את הדיון לתכונות שלא משתנות כאשר מעבירים את הגרף דרך איזומורפיזם, ודורשים בנוסף שהן לא משתנות גם כאשר ממספרים מחדש את השכנים (למשל כאשר מחליפים את ערכי $f(v,1)$ ו־$f(v,2)$). אנחנו נגביל את עצמנו לתכונות של גרפים לא מכוונים, ז"א שאם קיים i עבורו $f(v,i) = w$ אז קיים j עבורו $f(w,j) = v$.

התכונות הניתנות לבדיקה במודל זה שונות מאלו של המודל הצפוף. למשל, בדיקת 2־צביעות כאן דורשת $\tilde{\Theta}(\sqrt{n})$ שאילתות עבור ϵ קבוע (התלות של המקדמים ב־ϵ היא פולינומית). אנחו נראה בהמשך את החסם התחתון. בדיקת 3־צביעות כבר אינה אפשרית במספר תת־לינארי של שאילתות.

מודל נוסף שהוגדר הוא המודל הכללי. במודל זה אין חסם קבוע על דרגת הצמתים, ולכן מאפשרים גם שאילתה שאומרת עבור צומת v את מספר השכנים שלו (ובהתאמה מובטח שעבור $i \leq d(v)$ השאילתה עבור השכן ה־i של v לא תחזיר "$\perp$"). לפעמים גם מתירים שאילתות על זוגות צמתים. המושג של להיות ϵ־רחוק מוגדר יחסית למספר הקשתות המקורי בקלט. בדיקת תכונות במודל הזה בד"כ תיקח מספר לא קבוע של שאילתות. אפשר לראות את הדוגמה של גרף בעל n צמתים המורכב מקליק בעל $\lceil \sqrt{n} \rceil$ צמתים בתוספת צמתים מבודדים - כאן יקח מספר לא קבוע של שאילתות אפילו לגלות את העובדה שיש קשתות בגרף זה. המודל הכללי הוגדר לראשונה במאמר Kaufman, Krivelevich, Ron: Tight bounds for testing bipartiteness in general graphs.

בדיקת קשירות של גרף דליל

במודל הגרפים הצפוף אין בעיה לכתוב "אלגוריתם" בדיקה עבור התכונה שהגרף קשיר: אם הגרף הוא עם יותר מ־$1/\epsilon$ צמתים ואינו קשיר, אפשר פשוט להוסיף עץ שרירותי לקבוצת הקשתות וכך להפוך אותו לקשיר בפחות מ־ϵn^2 שינויים. האלגוריתם המלא לבדיקת קשירות במודל הצפוף ישאל את כל זוגות הצמתים בגרף אם יש פחות מ־$1/\epsilon$ צמתים (ואז יבצע BFS למשל), ואם יש יותר מ־$1/\epsilon$ צמתים אז הוא פשוט ידפיס "כן" בלי לשאול שאילתות כלל.

במודל הדליל בדיקת קשירות אינה קשה במיוחד, אבל גם לא טריביאלית, בעיקר אם רוצים לצמצם את מספר השאילתות ככל שניתן. הדבר העיקרי לשים לב הוא שמספר רכיבי הקשירות קובע את המרחק של

הגרף מלהיות קשיר: גרף עם k רכיבי קשירות ניתן להפוך לקשיר ע"י תוספת של $k-1$ קשתות בין הרכיבים – מסדרים אותם בסדר שרירותי ומוסיפים קשת בין כל שני רכיבים עוקבים. זהו אבל לא סוף הסיפור, כי עלינו גם לשמור על התנאי שהדרגה המקסימלית היא d. ברכיב חסר מעגלים אין בעיה, הוא בהכרח או יהיה מורכב מצומת בודד או יהיו לו לפחות שני צמתים מדרגה 1, ובשני המקרים אפשר לחבר אותו בקשתות לשני רכיבים אחרים בלי לעבור את דרגת המקסימום. ברכיב קשירות שמכיל מעגל אפשר להסיר את אחת מקשתות המעגל בלי "לשבור" את הרכיב, מה ש"מפנה" לנו שני צמתים שדרגתם תהיה עכשיו קטנה מ־d, שאפשר לחבר בקשתות לרכיבים אחרים. סה"כ מתקבל שעדיין צריך לכל היותר $2k-1$ שינויים בשביל להפוך גרף עם k רכיבי קשירות לקשיר, מה שאומר שלגרף ϵ־רחוק מקשירות חייבים להיות לפחות $\epsilon dn/2 = \Omega(\epsilon dn)$ רכיבי קשירות.

הדבר הבא לשים לב הוא שאם יש k רכיבי קשירות, אז לא יכולים להיות יותר מ־$k/2$ מהם שהם בעלי יותר מ־$2n/k$ צמתים, ולכן יש לפחות $k/2$ רכיבי קשירות בעלי פחות מ־$2n/k$ צמתים. אם נדגום $4n/k$ צמתים באופן יוניפורמי וב"ת, בהסתברות לפחות $\frac{2}{3}$ לפחות אחד מהם יהיה ברכיב קשירות בעל לא יותר מ־$2n/k$ צמתים. אם נבצע חיפוש לרוחב (BFS) או לעומק (DFS) מצומת כזה נוכל לגלות לאחר לא יותר מ־$2dn/k$ שאילתות את כל רכיב הקשירות המכיל אותו, ובפרט נגלה שהגרף בכללותו אינו קשיר (ההכפלה ב־d בחסם על מספר השאילתות הוא בגלל הצורך לשאול לכל צומת את כל שכניו, גם אם בדיעבד מתברר שאלו צמתים שהחיפוש כבר עבר בהם).

אלגוריתם ϵ־בדיקה יתבצע אם כן באופן הבא: דוגמים באופן יוניפורמי וב"ת $\lceil 8/\epsilon d\rceil$ צמתים, ועבור כל אחד מהם מבצעים חיפוש לרוחב או חיפוש לעומק עד שמגלים רכיב קשירות או עד שמגיעים ממנו ל־$\lceil 4/\epsilon d\rceil$ צמתים (ואז מסיקים שהוא לא ברכיב קשירות קטן מספיק). אם במהלך ההרצה מגלים רכיב קשירות אז דוחים את הגרף, ואחרת מקבלים אותו. סה"כ אנחנו מבצעים כאן $O(1/\epsilon^2 d)$ שאילתות על שכנים של צמתים.

נרצה עתה להוריד את החזקה של ϵ בביטוי הזה. על מנת לקבל תובנה למקור הבזבוז נסתכל על מקרי הקצה: אם כל $k/2$ הרכיבים הם בני צומת בודד, אז באמת צריך $O(k/n)$ שאילתות כדי למצוא צומת מתוך רכיב כזה, אבל רק שאילתה בודדת על צומת כזה בשביל להבין שהוא מהווה רכיב. מצד שני, אם כל הרכיבים הנ"ל הם למשל בני $2n/k$ צמתים, אז צריך רק $O(1)$ דגימות בשביל למצוא צומת ברכיב כזה, שעבור הווידוא שלו באמת נצטרך $O(dk/n)$ שאילתות. במילים אחרות, אם היינו יכולים לדעת את יותר במדוייק את גודל הרכיבים הקטנים ולהתאים אליו את האסטרטגיה שלנו, אז היינו יכולים לחסוך שאילתות.

במקום זאת, ננסה את ההתאמה עבור כל הגדלים האפשריים – עבור ϵ־בדיקה צריך לחשוב על הגדלים בין 1 ל־$4/\epsilon d$. מספיק אבל לקבץ את הגדלים לפי חזקות של 2: נסמן $t = \lfloor\log(4/\epsilon d)\rfloor + 1$, ולכל $1 \leq r \leq t$ נסמן ב־k_r את מספר הרכיבים שגודלם בין 2^{r-1} לבין $2^r - 1$. אם הגרף הוא ϵ־רחוק, אז יש לפחות $\epsilon dn/4$ רכיבי קשירות מגודל קטן מ־2^t, ז"א שמתקיים $\sum_{r=1}^{t} k_r \geq \epsilon dn/4$. על כן (עבור ϵ קטן מקבוע גלובלי מתאים) קיים r ספציפי שעבורו $k_r \geq \epsilon dn/10\log(1/\epsilon d)$.

אלגוריתם הבדיקה ינסה את כל ה־r האפשריים. לכל r, האלגוריתם ידגום $20\log(1/\epsilon d)/2^{r-1}\epsilon d$ צמתים, ולכל אחד מהם יבדוק האם הוא ברכיב קשירות בגודל קטן מ־2^r באמצעות $d2^r$ שאילתות. עבור ה־r שעבורו מתקיים $k_r \geq \epsilon dn/10\log(1/\epsilon d)$, בהסתברות גדולה מ־$\frac{2}{3}$ האלגוריתם ימצא צומת בתוך רכיב קשירות בגודל לפחות 2^{r-1} אך פחות מ־2^r, ואז יאמת שאכן זהו המצב. סה"כ מספר השאילתות לסיבוב ה־r הוא $O(\log(1/\epsilon d)/\epsilon)$, ובכל הסיבובים יחדיו הסה"כ הוא $O(\log(1/\epsilon d)^2/\epsilon) = \tilde{O}(1/\epsilon)$.

חסם תחתון עבור בדיקת דו־צדדיות

נראה עתה שקיים ϵ קבוע שעבורו ϵ־בדיקה של דו־צביעות של גרף במודל הדליל דורשת $\Omega(\sqrt{n})$ שאילתות, אפילו עבור $d = 3$. לשם פשטות נאפשר בבניה שלנו קשתות כפולות (מצב שבו יש שתי קשתות בין אותו זוג צמתים, אשר מתבטא בכך שלכל אחד מצמתי הקשת הכפולה יופיע בן הזוג השני שלו פעמיים ברשימה). אח"כ נראה איך אפשר (די בקלות) להיפטר ממצב זה.

במודל הדליל, כפי שראינו בדוגמה הקודמת, אדפטיביות היא בד"כ מאוד חיונית עבור האלגוריתם. אלגוריתם טיפוסי יבצע סידרה של חיפושים (לצורך העניין גם הילוך מקרי הוא סוג של חיפוש). בדוגמה הנגדית שלנו נדאג בעצם שהאלגוריתם לא ימצא מעגל. ליתר דיוק, נרצה להגיע למצב שבהסתברות גבוהה, בכל שאילתה

שהאלגוריתם יבצע, התשובה תהיה צומת חדש שלא הופיע בשום שאילתה קודמת (לא בשאילתה עצמה ולא בתשובה לשאילתה). מכיוון שאפשר להניח שלפני הרצת האלגוריתם העברנו את צמתי הגרף דרך פרמוטציה שנבחרה באופן יוניפורמי, בכל פעם שהתשובה לשאילתה היא צומת שלא הופיע קודם, אפשר להניח שהתווית שלו תהיה איבר שנבחר יוניפורמית מהאיברים הנותרים בקבוצת הצמתים V.

נגדיר שתי התפלגויות מעל קבוצת הגרפים בעלי $2n$ צמתים ועם דרגה מקסימלית 3. בשתי ההתפלגויות נתחיל ממעגל המילטוני מעל קבוצת הצמתים $\{1,\ldots,2n\}$, כאשר הצומת ה־i מחובר לצומת ה־$i-1$ וה־$i+1$ (ו־1 מחובר ל־$2n$). בשלב הבא, עבור ההתפלגות ν, פשוט נוסיף לגרף זיווג מושלם שנבחר מקרית באופן יוניפורמי מבין כל הזיווגים המושלמים האפשריים. עבור ההתפלגות τ, נגריל באופן יוניפורמי זיווג מכל הצמתים עם מזהה זוגי לכל הצמתים עם מזהה אי־זוגי (מבין $n!$ האפשרויות), ונוסיף אותו לגרף. מכיוון שלא דרשנו שהזיווגים לא יכילו קשתות מהמעגל המקורי, יכולות לצאת לנו מכך קשתות כפולות. לבסוף, בשתי ההתפלגויות, נמספר מחדש את קבוצת הצמתים דרך פרמוטציה מקרית σ שנבחרה יוניפורמית מ־S_{2n}, על מנת שהאלגוריתם לא יוכל לקבל מידע על "מרחק יחסי על המעגל ההמילטוני" בין הצמתים ששאל עליהם (אלא אם כן הצליח לגלות את כל המסלול בין הצמתים). בפרט אנחנו רוצים שיהיה קשה לאלגוריתם לברר את הזוגיות של המרחק הנ"ל.

דבר אחד שלא נטרח לעשות זה להגריל מחדש את סדר רשימת השכנים עבור כל צומת. אנחנו אפילו נניח שהשאילתה $(v,1)$ תמיד תחזיר את הצומת שהיה "אחרי" v על המעגל המקורי, השאילתה $(v,2)$ תמיד תחזיר את הצומת "לפני" v, והשאילתה $(v,3)$ תחזיר את הצומת שזווג ל־v בשלב שבו הוספנו את הזיווג המושלם למעגל המקורי.

ראשית, נשים לב שגרף שנבחר לפי τ הוא תמיד דו־צדדי. לשם כך נסתכל על תוויות הצמתים לפני שמספרנו אותן מחדש, ונראה שנוכל לצבוע את כל בעלי התווית הזוגית בצבע אחד, ואת כל בעלי התווית האי־זוגית בצבע השני. לעומת זאת, גרף שנבחר לפי ν יהיה $\frac{1}{180}$-רחוק מדו־צדדיות בהסתברות $1-o(1)$. נוכיח זאת בשלבים. אנחנו צריכים להוכיח שבהסתברות $1-o(1)$, המצב הוא שלכל צביעה של הגרף ב־2 צבעים יהיו יותר מ־$\frac{1}{30}n$ קשתות מפרות (קשתות בין שני צמתים מאותו צבע). בשביל להבין את החישוב של המרחק "$\frac{1}{180}$" שימו לב ש־$d=3$, ושמספר הצמתים כאן הוא $2n$ (לא n).

נוכיח עתה שעבור צביעה קבועה מראש של קבוצת הצמתים $\{1,\ldots,2n\}$, לזיווג מושלם מקרי יהיו יותר מ־$\frac{1}{30}n$ קשתות מפרות בהסתברות מאוד גבוהה. נשים לב שאפשר להגריל זיווג מושלם באופן יוניפורמי תוך שימוש בתהליך הבא: בכל פעם נבחר באופן שרירותי צומת v שעוד אין לו בן־זוג, ונבחר עבורו בן־זוג באופן יוניפורמי מבין כל הצמים הלא־מזווגים הנותרים. התובנה החשובה כאן היא שמותר לבחור את v באופן שרירותי לחלוטין, אפילו באופן שתלוי בקשתות הזיווג שכבר נבחרו. הסיבה לכך היא שההתפלגות של הזיווג האקראי, גם כשמתנים אותה על הקשתות שכבר קיימות, תמיד תהיה שווה להתפלגות של זיווג מושלם שנבחר יוניפורמית מבין הזיווגים על הצמתים שעוד לא זווגו (שאותו מוסיפים לקשתות שהיו קיימות קודם).

עתה נשתמש בתהליך הבא: בכל שלב נבחר את v להיות מקבוצת הצבע שיש לה יותר צמתים בלתי מזווגים. כל עוד יש לפחות 4 צמתים לא מזווגים, הסיכוי של v להיות מזווגת לצומת מאותו צבע הוא לפחות $\frac{1}{3}$ (זוהי ההסתברות כשיש בדיוק 4 צמתים נותרים, ומתוכם יש בדיוק 2 צמתים מכל צבע). המדובר אם כן ב־$n-1$ שלבים, שניתן לחסום את התוצאה שלהם מלמטה ע"י סכום של $n-1$ משתנים ב"ת שכל אחד מהם הוא 1 בהסתברות $\frac{1}{3}$ ו־0 בהסתברות $\frac{2}{3}$. לפי חסימת סטיות גדולות, הסיכוי שיהיו לא יותר מ־$\frac{1}{30}n$ קשתות מפרות חסום ע"י $e^{-2(1/3-1/30)^2(n-1)}=o(2^{-n/4})$.

עבור השלב הבא נשתמש בחסם הבינום $\frac{1}{n+1}2^{nH(k/n)}\leq\binom{n}{k}\leq 2^{nH(k/n)}$, כאשר $H:[0,1]\to[0,1]$ היא הפונקציה המוגדרת ע"י $H(x)=x\log\frac{1}{x}+(1-x)\log\frac{1}{1-x}$ כאשר $H(0)=H(1)=0$. הצביעות שנצטרך להראות שיש עבורן יותר מ־$\frac{1}{30}n$ קשתות זיווג מפרות הן הצביעות שאין להן יותר מ־$\frac{1}{30}n$ קשתות מפרות כבר מהמעגל עצמו. כל צביעה כזו נקבעת (עד כדי החלפת שני הצבעים) אך ורק לפי קבוצת הקשתות המפרות בגרף. על כן, מספרן חסום ע"י $\sum_{k=0}^{\lfloor n/30\rfloor}\binom{2n}{k}\leq\frac{n}{30}2^{nH(1/30)}=o(2^{n/4})$. מכאן שאפשר להשתמש בחסם איחוד מאורעות, ולקבל שההסתברות שיש צביעה כל שהיא (עבור גרף שנבחר לפי ν) שבה יש לכל היותר $\frac{1}{30}n$ קשתות מפרות היא $o(1)$, כנדרש.

עתה כשיש לנו את ההתפלגות τ ואת ההתפלגות ν, נרצה להראות שעבור אלגוריתם $\mathcal{A}$ שמבצע פחות מ־$\frac{1}{3}\sqrt{n}$ שאילתות, ההבדל בין ההתנהגות של $\mathcal{A}$ מעל שתי ההתפלגויות חסום ע"י $\frac{1}{4}$ (ב"התנהגות" הכוונה להתפלגות על סדרה של שאילתות של האלגוריתם והתשובות עליהן, כולל הקבלה או הדחיה בסוף). מכאן נובע שלא יתכן שהאלגוריתם נותן את התשובה הנכונה בהסתברות לפחות $\frac{2}{3}$ מעל ההתפלגות $\mu = \frac{1}{2}(\tau + \nu)$.

האדפטיביות החיונית לאלגוריתמים במודל הדליל יוצרת לנו בעיה בזה, שמחמירה עוד יותר מכיוון שמספר התשובות האפשריות לכל שאילתה הוא כמספר הצמתים הכולל בגרף. אנחנו נראה שיש מאורע שכאשר הוא מתקיים, אפשר להתייחס לאלגוריתם כאילו במובן מסויים הוא לא אדפטיבי – עדיין תהיה אדפטיביות במובן שבכל שלב האלגוריתם יבדוק שכן של צומת שאולי הוא קיבל כתוצאה משאילתה קודמת, אבל תהיה רשימה קבועה מראש של שאילתות בסגנון "מהו השכן ה־i של הצומת ה־j מבין אלו שראינו קודם".

לפני שנדון באדפטיביות, ראשית נחלק את השאילתות האפשריות בשלב ה־k לשלושה סוגים אפשריים.

- השכן ה־i של v, כאשר v הוא צומת ששאלנו עליו כבר בשלב ה־l עבור $l < k$ כל שהוא.
- השכן ה־i של v, כאשר v הוא צומת התשובה לשאילתה בשלב ה־l עבור $l < k$ כל שהוא.
- השכן ה־i של v, כאשר v הוא צומת שלא הופיע כלל בשלב קודם. מכיוון שבהתפלגויות שלנו אנחנו בסוף מעבירים את קבוצת הצמתים דרך פרמוטציה מקרית שנבחרה יוניפורמית, אפשר להניח שבמקרה הזה v נבחר יוניפורמית מבין קבוצת כל הצמתים שלא הופיעו בשאילתות או תשובות קודמות.

עתה נציין עוד תוצאה של העברת קבוצת הצמתים דרך פרמוטציה מקרית רגע לפני הרצת האלגוריתם: בכל פעם שהאלגוריתם מקבל כתשובה לשאילתה צומת שלא הופיע קודם כחלק משאילתה או כתשובה לשאילתה, התווית של הצומת המוחזר תתפלג יוניפורמית מעל כל תוויות הצמתים שעוד לא נראו. על כן, כל ההחלטות של האלגוריתם $\mathcal{A}$ יהיו על בסיס שוויונים ואי־שוויונים בין התוויות, ללא תלות כל שהיא בתוויות עצמן.

ההנחה הבאה על האלגוריתם היא שהוא לא מבצע שאילתות מיותרות. זה אומר שהוא לא חוזר פעמיים על אותה שאילתה, וספציפית עבור ההתפלגויות שלנו, אפשר גם להניח שהוא לא שואל את $f(v,3)$ אם כבר ידוע שמתקיים $v = f(w,3)$ (ולכן $w = f(v,3)$), או שואל את $f(v,3-i)$ עבור $i \in \{1,2\}$ אם כבר ידוע שמתקיים $v = f(w,i)$ (ולכן $w = f(v,3-i)$).

אנחנו נראה, גם עבור τ וגם עבור ν, שאם האלגוריתם מבצע $q < \frac{1}{3}\sqrt{n}$ שאילתות (ולא מבצע שאילתות מיותרות), אז בהסתברות כוללת של לפחות $\frac{7}{8}$, כל התשובות לשאילתות יהיו צמתים שלא הופיעו קודם. נשתמש כאן בגרסה קצת אחרת של השיטה של יאו: על מנת להראות שכל אלגוריתם הסתברותי לא יקבל כתשובה צומת שהופיע קודם תחת התפלגות כל שהיא μ, מספיק להראות שכל אלגוריתם דטרמיניסטי לא יקבל כתשובה צומת שהופיע קודם (שינינו את "תנאי הניצחון" של האלגוריתם לזה שהוא מוצא צומת שהופיע בעבר, במקום התנאי המקורי שהוא עונה נכון על בעיית הבדיקה).

מכיוון שאנחנו רוצים רק לחסום את ההסתברות שהאלגוריתם לא מקבל כתשובה צומת שהופיע בעבר, נבצע הקלה שתאפשר לנו אח"כ לסדר מחדש את השאילתות: בשאילתה מהסוג השלישי למעלה (כזו שבה v לא הופיע בשאילתה קודמת), v ייבחר באופן יוניפורמי מקבוצת כל הצמתים של הגרף, לא רק אלו שעוד לא הופיעו. במידה וזה יהיה צומת שכבר הופיע, גם זה יחשב כאילו האלגוריתם קיבל כתשובה צומת שכבר הופיע בעבר. השינוי הזה יכול רק להגדיל את ההסתברות למצוא צומת שהופיע בעבר, ולכן מספיק לחסום את ההסתברות הנ"ל תחת ההקלה הזו.

ועכשיו לחלק המכריע: כל עוד האלגוריתם לא קיבל צומת שהופיע קודם, יש רק אפשרות אחת לשוויונים בין הצמתים שהופיעו עד כה. על כן אלגוריתם דטרמיניסטי יהיה מאוד דומה לאלגוריתם לא אדפטיבי. ליתר דיוק, האלגוריתם יתואר כסידרה שנקבעה מראש של שאילתות, כאשר לכל $1 \leq k \leq q$, השאילתה ה־k היא מאחת משלושת הצורות שתוארו קודם. נסמן ב־w_k את תוצאת השאילתה ה־k. כמו כן, נסמן ב־$N \subseteq \{1,\ldots,q\}$ את האיברים שבהם השאילתה מתחילה מצומת שלא הוזכר קודם (ולכן זהותו נבחרת יוניפורמית מקבוצת כל הצמתים). עבור $k \in N$, נסמן ב־x_k את הצומת שנבחר עבור השאילתה ה־k (כאשר התשובה עדיין תסומן ב־w_k). שימו לב שבפרט תמיד $1 \in N$ (בעת השאילתה הראשונה אין צמתים קודמים שאפשר להתייחס אליהם).

לסיכום, עבור כל $k \in N$ השאילתה תהיה מהצורה $f(x_k, i_k)$, ועבור כל $k \in \{1,\ldots,q\} \setminus N$ השאילתה תהיה או מהצורה $f(w_l, i_k)$ עבור $1 \le l < k$ או מהצורה $f(x_l, i_k)$ עבור $l \in \{1,\ldots,k-1\} \cap N$.

סדר השאילתות עצמו אינו משנה בשלב זה, כל עוד שומרים על זה שהצומת שממנו השאילתה יוצאת אינו נובע משאילתה שמוקמה אחריה. אפשר לייצג את האלגוריתם אם כן בצורת יער מכוון עם תוויות על הקשתות: השורשים של היער יהיו כל הצמתים x_k עבור $k \in N$. אם השאילתה ה־k היתה $f(v, i_k)$ (כאשר v יכול להיות או x_k או אחד ה־v_l או ה־x_l הקודמים) אז w_k יהיה בן של v עם קשת עם תווית i_k. מספר העצים ביער הזה יהיה בדיוק $|N|$, ומספר הצמתים הכולל יהיה $q + |N| \le 2q$.

עבור צומת v מעץ השאילתות, נסמן ב־$r(v)$ את האב הקדמון הגבוה ביותר שיש ממנו מסלול ל־v ללא קשתות עם תווית "3". עבור שורשים יתקיים $r(v) = v$, וכן זה יתקיים עבור צמתים בעצמם ניתנו כתשובה לשאילתה מהצורה "$f(u,3)$". לפי ההנחה ש־$\mathcal{A}$ לא מבצע שאילתות מיותרות, המסלול מ־$r(v)$ ל־v יהיה כולו מקשתות עם אותה תווית (כולן "1" או כולן "2").

נסדר את שאילתות האלגוריתם, ונחשב הצבות מתאימות של צמתי הגרף לצמתי העץ, באופן הבא: בכל שלב או נגריל הצבה של שורש (צומת מסוג x_k) באופן יוניפורמי מבין צמתי הגרף, או נבצע הצבה של צומת שהאב שלו כבר הוצב ומחובר אליו בקשת מסוג "3" לפי גרף הקלט (ע"י מעבר על קשת הזיווג המתאימה). ברגע שעשינו את ההצבה לצומת u הנ"ל, נציב מיידית את כל הצמתים v עבורם $r(v) = u$. ההצבה של אלו נקבעת דטרמיניסטית ע"י ההצבה ל־u, כי המדובר בשאילתות על קשתות המעגל ההמילטוני שהתחלנו ממנו בעת בניית גרף הקלט (גם אם בנינו לפי τ וגם אם לפי ν).

עבור שני צמתים u ו־v כל שהם בעץ, נחסום את הסיכוי שהוצב בהם אותו צומת של גרף הקלט (כשזה קורה, האלגוריתם נתקל בצומת שהופיע בעבר). אם $r(v) = r(u)$ אז לעולם לא יוצבו עבורם צמתים זהים (אורכי המסלולים אליהם מ־$r(v) = r(u)$ יקבעו את המרחק ביניהם על המעגל ההמילטוני שהשתמשנו בו בבניית גרף הקלט, והוא יהיה שונה מ־0 אם $u \neq v$). נניח ש־v קיבל את ההצבה שלו אחרי u. נסתכל על $r(v)$ (כזכור v מקבל את ההצבה שלו מייד אחרי $r(v)$). אם זהו שורש, אז $r(v)$ יקבל צומת שנבחר יוניפורמית מתוך צמתי הגרף, ולכן גם v יקבל צומת שנבחר יוניפורמית (ה"כיוון" והמרחק על המעגל ההמילטוני בין $r(v)$ ל־v קבועים מראש ע"י קשתות העץ). על כן ההסתברות לשוויון עם הצומת שהוצב ל־u היא $\frac{1}{2n}$ בדיוק.

עתה נניח ש־$r(v)$ הוא לא שורש, ז"א שהוא מחובר באמצעות קשת מסוג "3" לצומת שכבר יש בו הצבה. נסמן את האב הזה ב־w. המדובר בקשת של הזיווג שמוסיפים למעגל ההמילטוני. עבור ההתפלגות ν, נזכור איך אפשר לנתח את הזיווג המקרי: אפשר להניח שבכל שלב לוקחים צומת שרירותי שעוד לא זווג, ובוחרים לו בן־זוג באופן יוניפורמי מהצמתים הלא מזווגים האחרים. אנחנו נניח שעוקבים אחרי שאילתות האלגוריתם, ומבצעים בחירה כזו כל פעם שיש שאילתה על קשת מסוג "3". כשהגענו לשאילתה על הצומת שנבחר עבור w, יש לנו לפחות $2n - 2q \ge \frac{3}{2}n$ צמתים לא־מזווגים. על כן $r(v)$ נבחר יוניפורמית מבין לפחות $\frac{3}{2}n$ צמתים אפשריים, ולכן v נבחר יוניפורמית מבין $\frac{3}{2}n$ צמתים אפשריים (לפי הכיוון והמרחק על המעגל ההמילטוני בין $r(v)$ ל־v). על כן ההסתברות לשוויון בין הצומת המוצב ל־v לבין זה של u חסומה ע"י $\frac{2}{3n}$.

לבסוף ננתח את המקרה ש־$r(v)$ אינו שורש עבור τ. גם כאן אפשר לטעון שההצבה ל־$r(v)$ תתפלג יוניפורמית מעל קבוצת צמתים מתאימה, אבל הפעם יש גם את הנתון שהזוגיות של האינדכס של הצומת של $r(v)$ שונה מזו של w (כזכור ב־τ בוחרים את הזיווג ככה שצביעת המעגל ההמילטוני בשני צבעים לא תופר על ידו). על כן החסם התחתון על גודל קבוצת הצמתים המתאימה יהיה $n - 2q \ge \frac{1}{2}n$, ז"א ההסתברות לשוויון בין v ל־u חסומה ע"י $\frac{2}{n}$.

לסיום, עושים איחוד מאורעות על כל זוגות הצמתים שבעץ. מכיוון שמספר הצמתים בעץ אינו עולה על $2q$, מקבלים את החסם $\frac{1}{8} > \frac{(\sqrt{n}/3)^2}{2}\frac{2}{n} > \binom{2q}{2}\frac{2}{n}$. בזאת הראנו שגם עבור ν וגם עבור τ, בהסתברות לפחות $\frac{7}{8}$ האלגוריתם לא יקבל כתשובה לשאילתה צומת שנתקל בו בעבר באף שלב של האלגוריתם (הראינו את זה עבור אלגוריתמים דטרמיניסטים, ולפי שיטת יאו זה נכון גם עבור אלגוריתמים הסתברותיים).

עתה נסמן שלושה פרמטרים עבור האלגוריתם $\mathcal{A}$. הפרמטר p_τ הוא ההסתברות לקבל גרף שנבחר לפי τ, הפרמטר p_ν הוא ההסתברות לקבל גרף שנבחר לפי ν, והפרמטר p_σ הוא ההסתברות לקבל כאשר במקום לתת לאלגוריתם גרף כל שהוא, עונים לכל אחת מהשאילתות שלו בצומת מקרי שנבחר יוניפורמית מהצמתים שעוד לא הופיעו בשלבים הקודמים (זאת "ההסתברות של $\mathcal{A}$ לקבל תחת סימולציה של כישלון"). לפי

מה שהראינו למעלה מתקיים $\frac{1}{8} > |p_\tau - p_\sigma|$ וכן מתקיים $\frac{1}{8} > |p_\nu - p_\sigma|$, ולכן מתקיים $\frac{1}{4} > |p_\tau - p_\nu|$. אבל זה אומר ש־$\mathcal{A}$ אינו יכול להיות אלגוריתם בדיקה, כי אלגוריתם בדיקה היה חייב לקיים $p_\tau \geq \frac{2}{3}$ וגם $p_\nu \leq \frac{1}{3} + o(1)$, שזו סתירה.

לבסוף, נסקור בקצרה איך אפשר להיפטר מהאפשרות לקשתות כפולות. נגדיר את מרחבי ההסתברות ν ו־τ בדיוק כמו קודם, רק שהפעם בסוף ההגרלה נסיר את הקשת של הזיווג המושלם מכל זוג בעל קשת כפולה. חישוב מהיר יגלה שתוחלת מספר הקשתות הכפולות, גם עבור ν וגם עבור τ, היא $O(1)$. על כן (אפשר להשתמש באי־שוויון מרקוב) בהסתברות $1 - o(1)$ יהיו פחות מ־$n^{1/4}$ קשתות כפולות. הדבר אומר שגם אחרי ההסרה, גרף שנבחר לפי ν יהיה רחוק מדו־צביעות בהסתברות $1 - o(1)$. בנוסף, זה אומר שבניתוח האלגוריתם שלנו, ההסתברות ש־$\mathcal{A}$ יגיע לצומת שמחקנו ממנו קשת גם תהיה $o(1)$. עבור n גדול דיו, נקבל שבהסתברות גדולה ב־$\frac{7}{8}$ האלגוריתם $\mathcal{A}$ גם לא יגיע לצומת שהופיע בשלב קודם וגם לא יגיע לצומת מדרגה קטנה מ־3, ומשם אפשר להפעיל את השיקולים עבור p_τ ו־p_ν מול p_σ כמו קודם.

תוצאות נוספות שלא הספקנו לעבור עליהן

רב החומר עד כאן הגיע מהמאמר המקורי על המודל הדליל. מאמר זה הכיל גם תוצאות נוספות, כגון בדיקה עבור k־קשירות בקשתות ל־k גדול מ־1, עם עקרון דומה (אבל יותר מסובך) לזה שהוצג כאן. כעיקרון מסתכלים על מבנה עץ שמתאר רכיבי k־קשירות (זה דורש $k-1$ קשירות של הגרף, אז בעצם יהיה צריך לבדוק l־קשירות לכל $1 \leq l \leq k$). אם הגרף רחוק מלהיות k־קשיר אז יהיו הרבה עלים במבנה הנ"ל, ואת אלו אפשר יהיה למצוא דרך דגימת צמתים, שאת שייכותם לעלה ניתן לבדוק ע"י אלגוריתם חיפוש מתאים.

יש חסם עליון של $\tilde{O}(\sqrt{n}/\mathrm{poly}(\epsilon))$ על בדיקת דו־צדדיות שתואם את החסם התחתון שראינו. הוא הוכח במאמר Goldreich, Ron: A sublinear bipartite tester for bounded degree graphs. ההוכחה די מסובכת ומסתמכת על חלוקה של הגרף למעין "רכיבי אקספנדר", שבתוכם אפשר לגלות הפרה לדו־צדדיות באמצעות הילוכים מקריים. עבודה יותר מאוחרת השתמשה בטכניקות דומות על מנת להוכיח בדיקה חד־כיוונית של אי הכלת מינורים אסורים. בדיקה דו־כיוונית של מינורים אסורים אפשרית עם מספר שאילתות לא תלוי ב־n לפי המאמר Newman, Sohler: Every property of hyperfinite graphs is testable.

חסמים תחתונים על אלגוריתמים אדפטיבים

עד עכשיו ראינו שיטה כללית לחסימת אלגוריתמים לא־אדפטיבים (יישום של שיטת יאו), אבל החסמים נגד אלגוריתמים אדפטיבים שראינו היו בשיטות אד־הוק. רק עבור בדיקה של גרפים צפופים ראינו שאפשר לתרגם כל אלגוריתם אדפטיבי לאלגוריתם קנוני (שהוא בפרט לא־אדפטיבי) במחיר ריבועי במספר השאילתות. כאן נראה מספר טכניקות שעובדות נגד אלגוריתמים אדפטיבים.

פריסה של עץ ההחלטות

ההסתכלות על אלגוריתם הסתברותי כעל התפלגות מעל אלגוריתמים דטרמיניסטים תעזור לנו גם כאן. באופן פורמלי, אלגוריתם אדפטיבי דטרמיניסטי עבור $f: D \to R$ מיוצג על ידי עץ מכוון: כל צומת v בעץ שאינו עלה (החל מהשורש) יהיה מתוייג בזיהוי של איבר בתחום $a(v) \in D$ שלגביו תתבצע שאילתה, וכל עלה יהיה מתוייג בהחלטה "לקבל" או "לדחות". לכל צומת שאינו עלה תהיה בדיוק קשת יוצאת אחת לכל ערך אפשרי בטווח, מתוייגת בערך זה. אם למשל $R = \{0,1\}$, אז העץ יהיה עץ בינארי מלא.

בהינתן הקלט f, הרצה של האלגוריתם מאופיינת ע"י מסלול משורש העץ לאחד העלים, שמחושב באופן אינדוקטיבי. אם כבר חישבנו את $r = v_0, \ldots, v_i$ ו־v_i אינו עלה, אז הצומת v_{i+1} יהיה הבן של v_i דרך הקשת עם התגית $f(v_i)$. מספר השאילתות המקסימלי של האלגוריתם הוא אורך המסלול המקסימלי, ז"א גובה העץ.

אפשר לראות עתה שאלגוריתם אדפטיבי בעל q שאילתות עבור פונקציות עם טווח בינארי ניתן לתרגום לאלגוריתם לא־אדפטיבי בעל לכל היותר $2^q - 1$ שאילתות: עבור אלגוריתם דטרמיניסטי, שואלים מראש

את כל איברי D שמופיעים בכל התגיות של כל צמתי העץ שאינם עלים. מספר כל הצמתים האלו חסום ע"י $2^q - 1$. לאחר שכל השאילתות נשאלו, אפשר לחשב את המסלול המתקבל במעבר מהשורש לעלה תוך כדי שימוש בערכים שכבר התקבלו. משראינו איך התרגום נעשה עבור אלגוריתמים דטרמיניסטים, המעבר לאלגוריתמים הסתברותיים הוא באמצעות ההסתכלות על אלגוריתמים כאלה כעל התפלגויות מעל אלגוריתמים דטרמיניסטים.

ישנן תכונות מעל אלפבית בגודל קבוע שבהן באמת יש פער אקספוננציאלי בין אלגוריתמים אדפטיבים לבין אלגוריתמים לא־אדפטיבים. נסתכל על התכונה של כל המילים מעל $\{0,1,2,3\}$ שהן שרשור של פלינדרום מעל $\{0,1\}$ עם פלינדרום מעל $\{2,3\}$. האלפבית אומנם אינו $\{0,1\}$, אבל זה כמעט שקול – אפשר למשל לקודד כל אות כאן ע"י שתי אותיות מ־$\{0,1\}$ (המרחק החדש יהיה לא יותר מהמרחק הישן ולא פחות ממחציתו). אותה שיטה שראינו עבור שרשור פלינדרומים בפרק השיטה של יאו עובדת כאן, ונותנת חסם תחתון של $\Omega(\sqrt{n})$ שאילתות עבור $\frac{1}{5}$־בדיקה לא אדפטיבית של התכונה, באמצעות שתי ההתפלגויות הבאות.

- בהתפלגות τ אנחנו בוחרים באופן מקרי ויוניפורמי $1 \le k \le n$, בוחרים את u להיות פלינדרום מקרי ויוניפורמי באורך k מעל $\{0,1\}$, את v להיות פלינדרום מקרי ויוניפורמי באורך $n-k$ מעל $\{2,3\}$, ומגדירים את הקלט להיות השרשור $w = uv$.
- בהתפלגות ν אנחנו בוחרים באופן מקרי ויוניפורמי $1 \le k \le n$, בוחרים את u להיות מילה מקרית ויוניפורמית באורך k מעל $\{0,1\}$ (מתוך כל 2^k האפשרויות), את v להיות מילה מקרית ויוניפורמית באורך $n-k$ מעל $\{2,3\}$, ומגדירים את הקלט להיות השרשור $w = uv$.

ההוכחה ש־ν נותנת בהסתברות גבוהה מילה רחוקה מהתכונה כמעט זהה להוכחה המקבילה מהפרק על השיטה של יאו (הדרישה הנוספת שהפלינדרום השני הוא מעל אלפבית שונה מ־$\{0,1\}$ יכולה רק להגדיל את המרחק). גם ההוכחה שלכל קבוצה $Q \subset \{1,\ldots,n\}$ מגודל קטן מ־$\frac{1}{2}\sqrt{n}$ מתקיים $d(\tau|_Q, \nu|_Q) < \frac{1}{8}$ דומה להוכחה המקבילה משם, רק שכאן צריך לבדוק לחוד את ההתניה של שתי ההתפלגוית עבור כל ערך אפשרי של k. עבור ערכים של k שלא מביאים לקורלציה בין איזה שהם w_i ו־w_j עם $i,j \in Q$, נקבל שוויון בין ההתפלגויות המותנות של τ ו־ν.

זה נותן חסם תחתון של $\Omega(\log(n))$ עבור אלגוריתמים אדפטיבים: אלגוריתם אדפטיבי בעל $q < \frac{1}{5}\log(n)$ שאילתות היה ניתן לתרגום לאלגוריתם לא־אדפטיבי בעל לכל היותר $\frac{1}{3}(4^q - 1) = o(\sqrt{n})$ שאילתות, ואנחנו יודעים שאין אלגוריתם כזה.

מול החסם התחתון של $\Omega(\sqrt{n})$ עבור אלגוריתמים לא־אדפטיבים, באמת ניתן לכל ϵ קבוע לבצע ϵ־בדיקה אדפטיבית ב־$O(\log(n))$ שאילתות. האלגוריתם יפעל באופן הבא עבור מילה $w_1,\ldots,w_n$.

- ראשית, מוצאים k כך ש־$w_k \in \{0,1\}$ ו־$w_{k+1} \in \{2,3\}$, כאשר ייתכן גם המקרה $k=0$ ו־$w_1 \in \{2,3\}$ או המקרה $k=n$ ו־$w_n \in \{0,1\}$. לשם כך קודם כל שואלים את w_1 ואת w_n; אם $w_1 \in \{0,1\}$ ו־$w_n \in \{2,3\}$ אז משתמשים בטכניקה של חיפוש בינארי על מנת למצוא את k ב־$O(\log(n))$ שאילתות (בכל שלב "פונים ימינה" אם מוצאים ערך ב־$\{0,1\}$ ו"פונים שמאלה" אם מוצאים ערך ב־$\{2,3\}$).
- באמצעות $O(1/\epsilon)$ שאילתות מוודאים עכשיו ש־$w_1,\ldots,w_n$ קרובה להיות שרשור של פלינדרום מאורך k מעל $\{0,1\}$ עם פלינדרום מאורך $n-k$ מעל $\{2,3\}$, ע"י ביצוע התהליך הבא $2/\epsilon$ פעמים: בוחרים את $1 \le i \le n$ באופן יוניפורמי (וב"ת בסבבים הקודמים); אם $i \le k$, בודקים שמתקיים $w_i = w_{k+1-i} \in \{0,1\}$, ואם $i > k$, בודקים שמתקיים $w_i = w_{n+k+1-i} \in \{2,3\}$. מקבלים את הקלט אם ורק אם כל הבדיקות הנ"ל יצאו תקינות.

לא קשה לראות שהאלגוריתם יקבל כל מילה שמקיימת את התכונה בהסתברות 1. אם המילה $w_1,\ldots,w_n$ היא ϵ־רחוקה מלהיות שרשור של פלינדרום מעל $\{0,1\}$ עם פלינדרום מעל $\{2,3\}$, אז לכל k שהשלב הראשון יכול להעביר לשלב השני, השלב השני יקבל בהסתברות לכל היותר $(1-\epsilon)^{2/\epsilon} < \frac{1}{3}$.

שימוש בשיטת יאו לאלגוריתמים אדפטיביים

הראיה של אלגוריתם אדפטיבי דטרמיניסטי כעץ החלטות מאפשרת שימוש בשיטה דומה לשיטה שראינו עבור אלגוריתמים לא אדפטיביים לעבוד נגד אלגוריתמים אדפטיביים. הדבר מצריך תנאי יותר חזק על הדמיון בין שתי ההתפלגויות τ ו־ν (במקרה הזה התנאי אינו שקול לאי קיום אלגוריתם אדפטיבי – לפעמים צריך להשמש בשיטות אחרות). אם ההתפלגויות מקיימות:

- ההתפלגות τ תהיה מעל קלטים $f : D \to R$ שכולם מקיימים את התכונה.
- ההתפלגות ν תהיה מעל קלטים $f : D \to R$ שכולם ϵ־רחוקים מלקיים את התכונה.
- לכל $Q \subset D$ מגודל q, ולכל פונקציה $h : Q \to R$, מתקיים $\Pr_\tau[f|_Q = h] > \frac{2}{3}\Pr_\nu[f|_Q = h]$.

אז אין אלגוריתם אדפטיבי בעל q שאילתות אשר בודק את התכונה. היה אפשר להחליף את התנאי השלישי באחד שבו $\Pr_\nu[f|_Q = h] > \frac{2}{3}\Pr_\tau[f|_Q = h]$, אבל האופציה למעלה היא זו השימושית יותר.

על מנת להוכיח את את הטענה, ראשית נסמן ב־N את קבוצת העלים על עץ ההחלטה שבהגעה אליהם האלגוריתם דוחה. לכל $v \in N$, נסמן ב־Q_v את קבוצת איברי D שנמצאים על הצמתים במסלול מהשורש לעלה v, וב־$h_v : Q \to R$ את הפונקציה שמתקבלת ע"י הצבת התווית של כל קשת על המסלול מהשורש ל־v לצומת המקור של אותה קשת. זה אומר שהרצה על עץ ההחלטות תגיע לעלה v אם ורק אם $f|_{Q_v} = h_v$. על כן, עבור התפלגות כל שהיא מעל קבוצת הקלטים האפשריים, ההסתברות לקבלת הקלט המוגרל תהיה $\Pr[N] = \sum_{v\in N}\Pr[f|_{Q_v} = h_v]$.

במקרה שלנו מתקבל $\Pr_\tau[N] = \sum_{v\in N}\Pr_\tau[f|_{Q_v} = h_v] > \frac{2}{3}\sum_{v\in N}\Pr_\nu[f|_{Q_v} = h_v] = \frac{2}{3}\Pr_\nu[N]$. ניקח עתה את ההתפלגות $\mu = \frac{1}{2}(\tau + \nu)$, ונקבל שההסתברות לשגיאה של האלגוריתם חסומה מלמטה ע"י $\frac{1}{2}(\Pr_\tau[N] + 1 - \Pr_\nu[N]) > \frac{1}{2}(\frac{2}{3}\Pr_\nu[N] + 1 - \Pr_\nu[N]) = \frac{1}{2}(1 - \frac{1}{3}\Pr_\nu[N]) \ge \frac{1}{3}$. מכיוון שהוכחנו את זה עבור כל אלגוריתם דטרמיניסטי, הדבר יהיה נכון גם עבור הסתברות השגיאה של אלגוריתם הסתברותי.

גם כאן, אפשר לשנות את ההוכחה לכזו שתעבוד עם גרסה אלטרנטיבית, שבה ההתפלגות ν יכולה להוציא קלטים שאינו ϵ־רחוקים מהתכונה בהסתברות קטנה, עם פרמטר $\alpha < \frac{1}{3}$ מתאים. אם מתקיים:

- ההתפלגות τ תהיה מעל קלטים $f : D \to R$ שכולם מקיימים את התכונה.
- עבור ההתפלגות ν, הסיכוי שיתקבל $f : D \to R$ שאינו ϵ־רחוק מהתכונה הוא לכל היותר α.
- לכל $Q \subset D$ מגודל q, ולכל פונקציה $h : Q \to R$, מתקיים $\Pr_\tau[f|_Q = h] > (\frac{2}{3} + \alpha)\Pr_\nu[f|_Q = h]$.

אז גם במקרה זה אין אלגוריתם אדפטיבי שבודק את התכונה ב־q שאילתות. עבור הוכחת החסם כותבים $\frac{1}{2}(\Pr_\tau[N]+1-\alpha-\Pr_\nu[N]) > \frac{1}{2}((\frac{2}{3}+\alpha)\Pr_\nu[N]+1-\alpha-\Pr_\nu[N]) = \frac{1}{2}(1-\alpha-(\frac{1}{3}-\alpha)\Pr_\nu[N]) \ge \frac{1}{3}$ כאשר עבור אי־השוויון האחרון, מ־$\frac{1}{3} - \alpha > 0$ נובע שהמינימום ביחס ל־$\Pr_\nu[N]$ מתקבל כאשר הוא שווה ל־1, ואז יש שוויון.

להמחשת השיטה, נחזור לתכונה של המילים מעל $\{0,1\}$ שהן שרשור של שני פלינדרומים. זוג ההתפלגויות שהשתמשנו בו נגד אלגוריתמים לא־אדפטיביים עובד גם נגד אלגוריתמים אדפטיביים. נזכיר אותן כאן.

- בהתפלגות τ אנחנו בוחרים באופן מקרי ויוניפורמי $1 \le k \le n$, בוחרים את u להיות פלינדרום מקרי ויוניפורמי באורך k, את v להיות פלינדרום מקרי ויוניפורמי באורך $n - k$, ומגדירים את הקלט להיות השרשור $w = uv$.
- בהתפלגות ν אנחנו בוחרים את המילה $w \in \{0,1\}^n$ באופן מקרי ויוניפורמי. כזכור ההתפלגות הזו נותנת קלט שהוא $\frac{1}{5}$־רחוק מהתכונה בהסתברות $1 - o(1)$.

נניח עתה ש־$Q \subset D$ היא קבוצה בת $q \le \frac{1}{2}\sqrt{n}$ איברים, ו־$h : Q \to \{0,1\}$ פונקציה כל שהיא. ישירות מההגדרה מתקיים $\Pr_\nu[f|_Q = h] = 2^{-q}$. בהוכחה המקורית עבור אלגוריתמים לא־אדפטיביים, ניתחנו את המאורע שיש $i < j \in Q$ כך שעבור ה־k שהוגרל חייב להתקיים $w_i = w_j$. נסמן את המאורע הזה ב־B. כזכור מתקיים $\Pr_\tau[B] \le \frac{1}{8}$. כמו כן, כאשר B לא מתקיים, כל הערכים w_i עבור $i \in Q$ מוגרלים באופן ב"ת. על כן $\Pr_\tau[f|_Q = h] \ge \Pr_\tau[f|_Q = h \wedge \neg B] = \Pr_\tau[\neg B] \cdot \Pr_\tau[f|_Q = h | \neg B] \ge \frac{7}{8} 2^{-q}$. מכאן שעבור n גדול דיו מתקיימים התנאים שלפיהם זוג ההתפלגויות הזה מראה שלא קיים אלגוריתם אדפטיבי, שמבצע לא יותר מ־$\frac{1}{2}\sqrt{n}$ שאילתות ומצליח לבצע $\frac{1}{5}$־בדיקה של התכונה (עם הסתברות הצלחה לפחות $\frac{2}{3}$).

רדוקציה לסיבוכיות תקשורת

כאן נראה את השיטה הכללית הראשונה שפותחה עבור בדיקת תכונות שאינה מסתמכת על שיטת יאו. הרעיון הכללי הוא להראות שאם תכונה מסויימת היא ניתנת לבדיקה, אז אפשר באמצעות הבדיקה הזו לפתור בעיית סיבוכיות תקשורת אשר ידועה כדורשת כמות גבוהה של תקשורת. המודל המקובל של סיבוכיות תקשורת מאפשר הרבה סבבים, וזה "מתרגם" ברדוקציה לאדפטיביות של אלגוריתם הבדיקה. שיטה זו פותחה במאמר Blais, Brody, Matulef: Property Testing lower bounds via Communication Complexity.

נסקור בקצרה את המודל של סיבוכיות תקשורת: קיימים שני "שחקנים", לשחקן הראשון יש קלט $x \in \{0,1\}^n$ ולשני יש קלט $y \in \{0,1\}^n$. המטרה היא לחשב את הערך של פונקציה משותפת $f(x,y)$ (בד"כ זו פונקציה בוליאנית, ז"א שצריך לקבל או לדחות את (x,y)). במודל הזה השחקנים אמינים, ז"א שמניחים ששניהם יריצו את האלגוריתם האופטימלי. בכל סיבוב תקשורת כל אחד מהשחקנים שולח מחרוזת של ביטים לשני. ההרצה מסתיימת כאשר אחד השחקנים מחליט לקבל או לדחות את הקלט (או במקרה היותר כללי, מכריז על ערך $f(x,y)$). סיבוכיות התקשורת היא מספר הביטים הכולל שנשלח ע"י השחקנים.

במקרה הכי גרוע אפשר לפתור את הבעיה בסיבוכיות תקשורת $O(n)$: השחקן הראשון שולח את כל x בסבב התקשורת הראשון, ואז השחקן השני מחשב את $f(x,y)$ ופולט אותו. מטרת המחקר בתחום היא לברר אלו בעיות ניתנות לפתרון בפחות תקשורת. אצלנו בד"כ לא נגביל את מספר סבבי התקשורת עצמו, ז"א שזה בסדר גם לשלוח ביט בודד בכל סבב.

במודל שלנו נאפשר אלגוריתמים הסתברותיים, וליתר דיוק נשתמש במודל (היותר מקל) של מטבעות פומביים: זה אומר שכאשר אחד השחקנים צריך לבצע החלטה הסתברותית, שני השחקנים יודעים את תוצאות ההגרלה ללא צורך בתקשורת. משמעות השם "מטבעות פומביים" היא שאפשר להסתכל על המודל כאילו שני השחקנים מקבלים עבור ההרצה גישה משותפת למחרוזת אקראית ארוכה מספיק, שממנה הם קוראים כל אימת שמי מהם צריך לבצע הגרלה. במילים אחרות, אלגוריתם תקשורת עם מטבעות פומביים ניתן לתיאור כמרחב הסתברות מעל אלגוריתמי תקשורת דטרמיניסטים.

דוגמה לאלגוריתם תקשורת עם מטבעות פומביים שמשתמש רק ב־$O(1)$ תקשורת הוא האלגוריתם הבא עבור בעיית השוויון של המחרוזות x ו־y: משתמשים במטבעות הפומביים לבחור באופן מקרי, יוניפורמי וב"ת, את $a = a_1, \ldots, a_n \in \{0,1\}^n$. השחקן הראשון מדווח לשחקן השני את $\bigoplus_{i=1}^n a_i x_i$ (במילים אחרות, את זוגיות מספר ה־i שעבורם $a_i = x_i = 1$), והשחקן השני דוחה את הקלט אם הערך הזה שונה מ־$\bigoplus_{i=1}^n a_i y_i$. אם שתי המחרוזות שוות זו לזו, השחקן השני לעולם לא ידחה. אם המחרוזות שונות זו מזו, אז השחקן השני ידחה אם $\bigoplus_{i=1}^n a_i |x_i - y_i| = 1$, וזה קורה בהסתברות $\frac{1}{2}$. אפשר ע"י חזרה על התהליך (עם $b \in \{0,1\}^n$ שנבחר באופן ב"ת ב־a) להגדיל את ההסתברות לדחיית מחרוזות לא־שוות ל־$\frac{3}{4} > \frac{2}{3}$.

נחזור אלינו: אנחנו בד"כ נשתמש בזה שיש תכונה קשה ידועה בסיבוכיות תקשורת, זו של זרות קבוצות. נגיד ש־(x,y) מקיימים את הזרות אם לא קיים i שעבורו $x_i = y_i = 1$ (ניתן לחשוב על x ועל y כעל פונקציות אופייניות של ת"ק של $\{1, \ldots, n\}$). תכונה זו דורשת $\Omega(n)$ תקשורת להכרעה, אפילו כאשר מאפשרים מטבעות פומביים ושגיאה דו־צדדית של עד $\frac{1}{3}$. יתרה מזו, ידוע חסם חזק (ואופטימלי) של $\Omega(k)$ על הבעיה הבאה: נתון מראש שיש בדיוק $\lfloor \frac{k}{2} \rfloor$ אינדכסים עבורם $x_i = 1$, בדיוק $\lceil \frac{k}{2} \rceil$ אינדכסים עבורם $y_i = 1$, וכן ידוע שיש לכל היותר אינדכס יחיד שעבורו $x_i = y_i = 1$. הדרישה היא שנוכל להבחין בין המקרה שיש i עבורו $x_i = y_i = 1$, לבין המקרה שבו אין אף אינדכס כזה (לכל זוג (x,y) שאינו מקיים את הנתונים כאן מותר לתת כל תשובה שהיא).

אם יש לנו חסם תחתון עבור בעיית תקשורת, על מנת להוכיח חסם תחתון עבור בדיקת התכונה $\mathcal{P}$ של פונקציות $h : D \to \{0,1\}$, אנחנו נרצה לבנות "סכימת רדוקציה" שלכל $x, y \in \{0,1\}^n$ מתאימה "פונקציה משולבת" $f_{x,y} : D \to \{0,1\}$, שמקיימת את התנאים הבאים:

- לכל $a \in D$, אפשר בתקשורת של לכל היותר r לחשב את $f_{x,y}(a)$.
- אם צריך לקבל את (x,y), אז הפונקציה $f_{x,y}$ מקיימת את $\mathcal{P}$.
- אם צריך לדחות את (x,y), אז הפונקציה $f_{x,y}$ היא ϵ־רחוקה מלקיים את $\mathcal{P}$.

אם יש לנו רדוקציה כזו, ויש חסם תחתון של $\Omega(k)$ על בעיית התקשורת, אז יש חסם תחתון של $\Omega(k/r)$ על ϵ־בדיקה של $\mathcal{P}$. אם היה אפשר לבדוק את $\mathcal{P}$ ב־q שאילתות, אז היינו יכולים לכתוב אלגוריתם עבור בעיית התקשורת בסיבוכיות $O(qr)$ באופן הבא: האלגוריתם היה מבצע הרצה של אלגוריתם הבדיקה מעל $f_{x,y}$. כל פעם שצריך לבצע שאילתה מהצורה $f_{x,y}(a)$, השחקנים יחשבו אותו בתקשורת של לכל היותר r, וימשיכו את ההרצה של הבדיקה לפי תוצאת החישוב. מכיוון שיש מטבעות פומביים, שני השחקנים יכולים לעקוב אחרי אלגוריתם הבדיקה ולדעת מהי ה"שאילתה" הבאה שלו, גם אם הוא הסתברותי.

הדוגמה הכי טובה ליישום קשורה בפונקציות לינאריות: כזכור צריך $O(1/\epsilon)$ שאילתות בשביל לבדוק האם פונקציה $f : \{0,1\}^n \to \{0,1\}$ היא לינארית (עם הזיהוי $\{0,1\} = \mathbb{Z}_2$), או במילים אחרות, האם קיימים $a_1, \ldots, a_n \in \{0,1\}$ כך שמתקיים $f(x_1, \ldots, x_n) = \bigoplus_{i=1}^n a_i x_i$. כמה שאילתות צריך בשביל לבדוק האם הפונקציה f היא לינארית עם בדיוק k מקדמים שונים מ־0, ז"א פונקציה מהצורה $\bigoplus_{j=1}^k x_{i_j}$ עבור $i_1 < \cdots < i_k$ מתאימים?

התשובה היא שצריך $\tilde{\Theta}(k)$ שאילתות. נראה כאן את החסם התחתון של $\Omega(k)$, כאשר החסם העליון של $\tilde{O}(k)$ מתקבל משילוב בדיקת לינאריות עם בדיקת חונטות (שמוזכרת בהמשך החוברת). החסם התחתון הוא באמצעות רדוקציה לבעייה הקשה של זרות קבוצות שתארנו למעלה (כאשר ל־x יש $\lfloor \frac{k}{2} \rfloor$ אינדכסים עם 1 ול־y יש $\lceil \frac{k}{2} \rceil$ אינדכסים כאלה). הרדוקציה תעשה באופן הבא.

- עבור $x, y \in \{0,1\}^n$ נגדיר את $f_{x,y} : \{0,1\}^n \to \{0,1\}$ לפי $f_{x,y}(a_1, \ldots, a_n) = \bigoplus_{i=1}^n a_i(x_i \oplus y_i)$.
- על מנת לחשב את $f_{x,y}(a_1, \ldots, a_n)$, השחקן הראשון שולח את $\bigoplus_{i=1}^n a_i x_i$ והשחקן השני שולח את $\bigoplus_{i=1}^n a_i y_i$, ואז שניהם יכולים לחשב את $\bigoplus_{i=1}^n a_i(x_i \oplus y_i) = (\bigoplus_{i=1}^n a_i x_i) \oplus (\bigoplus_{i=1}^n a_i y_i)$.
- אם x ו־y מקיימים את תכונת הזרות, אז $f_{x,y}$ היא פונקציה לינארית עם k מקדמים שונים מ־0.
- אם ל־x ו־y יש אינדכס i יחיד עבורו $x_i = y_i = 1$, אז $f_{x,y}$ היא פונקציה לינארית עם $k-2$ מקדמים שונים מ־0 (כל המקדמים שהם 1 באחד מ־x ו־y אבל לא בשניהם). מכיוון שכל שתי פונקציות לינאריות נבדלות זו מזו ב־2^{n-1} מקומות בדיוק, זה אומר שהמרחק של $f_{x,y}$ מפונקציה לינארית עם k מקדמים בפרט גדול מ־$\frac{1}{3}$.

מהבניה הזו ומהחסם התחתון של $\Omega(k)$ על בעיית התקשורת של קבוצות זרות, אנחנו מקבלים את החסם הנדרש של $\Omega(k)$ על $\frac{1}{3}$־בדיקה של התכונה של להיות פונקציה לינארית עם k מקדמים שונים מ־0.

בדיקת התפלגויות

נקדיש עתה זמן למודל בדיקה שהוא הרבה יותר חלש מהמודל הרגיל, אבל בעל שימושים רבים, גם כחלק מאלגוריתמי בדיקה במודלים יותר חזקים וגם בתחום אלגוריתמי למידה.

כאן ה"קלט" שלנו תהיה התפלגות מעל קבוצת בסיס S, בד"כ $S = \{1, \ldots, n\}$. אפשר לתאר התפלגות כפונקציה $\mu : S \to [0,1]$ שמקיימת $\sum_{a \in S} \mu(a) = 1$. גם כאן, החלק החשוב במודל הוא מהי השאילתה המותרת ומה מושג המרחק (מתי μ נחשבת "ϵ־רחוקה" מתכונה מסויימת).

במקום שאילתות, האלגוריתם יכול לקבל דגימות. דגימה פירושה קבלה של ערך $a \in S$ שנבחר (לא ע"י האלגוריתם) לפי μ. אלגוריתם שמבצע q דגימות הוא בעצם אלגוריתם עם גישה ל־q משתנים מקריים $A_1, \ldots, A_q$, כך שכולם ב"ת (לחלוטין) זה בזה וכל A_i מתפלג לפי μ. אין כאן שום מקום לאדפטיביות – אפילו שלעיתים יהיה נוח לתאר אלגוריתם כזה באופן "אינטראקטיבי" (למשל אם האלגוריתם מתעלם מחלק מהדגימות במהלך החישוב), בעצם המדובר יהיה בפונקציה שמתארת לכל סדרה של ערכים $a_1, \ldots, a_q$ עבור $A_1, \ldots, A_q$ את ההסתברות לקבל אותה.

המרחק של התפלגות מהתכונה הנבדקת יוגדר ע"י מרחק ההתפלגויות (variation distance). כזכור המדובר ב־$d(\nu,\mu) = \frac{1}{2}\sum_{a\in S}|\mu(a)-\nu(a)|$. ההתפלגות μ תיקרא ϵ־רחוקה מהתכונה, אם אין התפלגות ν שמקיימת את התכונה ושעבורה $d(\nu,\mu) \le \epsilon$.

בדיקת יוניפורמיות

התכונה הכי פשוטה של התפלגות שאפשר לבדוק היא שהמדובר בהתפלגות יוניפורמית מעל S. מסתבר שאפילו תכונה זו דורשת מספר לא קבוע של דגימות, $\Theta(\sqrt{n})$ לכל ϵ קבוע קטן מ־1 (אנחנו נוכיח את החסם התחתון עבור $\epsilon = \frac{1}{3}$), כאשר נסמן $n = |S|$. תוצאה זו, הראשונה בתחום של בדיקה הסתברויות, הופיעה כחלק מפרוצדורת בדיקה במודל הגרפים הדליל, במאמר Goldreich, Ron: On testing expansion in bounded-degree graphs.

נתחיל מהחסם התחתון: גם כאן משתמשים בשיטת יאו, אבל כדאי (בגלל שזו פעם ראשונה) להסביר בדיוק למה הכוונה. הקלט כאן הוא התפלגות מעל S. התפלגות מעל קבוצת הקלטים היא התפלגות מעל התפלגויות מעל S. אנחנו נגדיר שתי התפלגויות כאלו. נניח ש־n מספר זוגי.

- בהתפלגות τ אנחו תמיד נבחר את ההתפלגות היוניפורמית π_S מעל S. ז"א שאם נסמן את התפלגות הקלט ב־μ, אז בעצם מתקיים $\Pr_\tau[\mu = \pi_S] = 1$.
- עבור ההתפלגות ν, אנחנו ראשית נבחר תת־קבוצה $S' \subset S$ מגודל $\frac{n}{2}$ בדיוק, ונעשה את זה יוניפורמית מהמשפחה של כל תתי־הקבוצה של S מגודל $\frac{n}{2}$. לאחר בחירה זו נגדיר $\mu = \pi_{S'}$, ההתפלגות היוניפורמית מעל S' (עם הסתברות 0 לקבלת איבר ב־$S \setminus S'$). נשים לב שהתפלגות זו מקיימת $d(\pi_S, \pi_{S'}) = \frac{1}{2}$, ולכן היא (למשל) $\frac{1}{3}$־רחוקה מההתפלגות היוניפורמית על כל S.

עתה ננתח מה קורה כאשר לוקחים פחות מ־$\frac{1}{3}\sqrt{n}$ דגימות. נסמן את המ"מ של הדגימות ב־$A_1, \ldots, A_q$. כאשר אנחנו תחת ההתפלגות τ, ההסתברות שאותו ערך יופיע יותר מפעם אחת חסום (באמצעות איחוד מאורעות) ע"י $\binom{q}{2}\frac{1}{n} < \frac{1}{18}$. ההתפלגות של $A_1, \ldots, A_q$, תחת התניה על המאורע שאין חזרות על ערכים, היא ההתפלגות היוניפורמית מעל כל סדרות הערכים האפשריות ללא חזרות של q ערכים מ־S. באופן יותר פורמלי: אם נסמן ב־U את המאורע שאין חזרות על הערכים הנדגמים, אז אם $\alpha_1, \ldots, \alpha_q \in S$ כולם שונים זה מזה נקבל $\Pr_\tau[A_1 = \alpha_1 \wedge \cdots \wedge A_q = \alpha_q \,|\, U] = (n-q)!/n!$, ואם יש חזרות ב־$\alpha_1, \ldots, \alpha_q$ נקבל $\Pr_\tau[A_1 = \alpha_1 \wedge \cdots \wedge A_q = \alpha_q \,|\, U] = 0$.

נסתכל עתה על ההתפלגות ν. במקרה זה, $A_1, \ldots, A_q$ יהיו דגימות ב"ת מההתפלגות S' עבור הקבוצה S' שנבחרה בתיאור של ν (שימו לב שאלו אינם ב"ת תחת ההתפלגות מעל התפלגויות ν – הם ב"ת רק תחת התניה על S' ספציפית). ההסתברות שתהיה חזרה על ערך חסומה ע"י $\binom{q}{2}\frac{2}{n} < \frac{1}{9}$. נסתכל עתה על ההתפלגות של $A_1, \ldots, A_q$ כאשר אין חזרה: זאת תהיה סדרה ללא חזרות שנבחרת יוניפורמית מכל הסדרות האפשריות מעל S'. אולם S' עצמה נבחרה באופן יוניפורמי מתוך כל תתי־הקבוצה המתאימים של S, ולכן (כאשר לוקחים בחשבון את ההתפלגות ν מעל ההתפלגויות), תחת התניה על המאורע שאין חזרות על ערכים במהלך הדגימה, $A_1, \ldots, A_q$ היא סדרת ערכים שנבחרת יוניפורמית מכל הסדרות האפשריות מעל S.

מכל אלו נובע שאם ננתח את ההתפלגות של סדרת המ"מ $A_1, \ldots, A_q$ תחת τ ותחת ν, נקבל שההבדל בין אלו חסום ע"י סכום ההבדלים מההתפלגות היוניפורמית ללא חזרות, שחסום ע"י $\frac{1}{18} + \frac{1}{9} < \frac{1}{3}$ (עם קצת יותר מאמץ היה אפשר להקטין את החסם ל־$\frac{1}{9}$). על כן אי אפשר לבצע $\frac{1}{3}$־בדיקה עבור יוניפורמיות במספר דגימות כזה – האלגוריתם יטעה בהסתברות גדולה מ־$\frac{1}{3}$ כאשר מזינים לו התפלגות מעל S שנבחרת לפי $\frac{1}{2}(\tau + \nu)$.

עתה נוכיח את החסם העליון. הרעיון הוא להתייחס להתפלגות μ כאל וקטור מעל n קורדינטות, ולנסות לשערך את הנורמה $\|\mu\|_2^2 = \sum_{a\in D}|\mu(a)|^2$. תזכורת: באופן כללי, עבור $1 \le \alpha \le \infty$ הנורמה מוגדרת ע"י $\|\mu\|_\alpha = (\sum_{a\in S}|\mu(a)|^\alpha)^{1/\alpha}$, כאשר עבור "$\infty$" מגדירים $\|\mu\|_\infty = \lim_{\alpha\to\infty}\|\mu\|_\alpha = \max_{a\in S}|\mu(a)|$.

התועלת בשיערוך הנורמה היא שעבור ההתפלגות היוניפורמית מתקיים $\|\pi_S\|_2^2 = \frac{1}{n}$. לעומת זאת, עבור μ כללי נסמן $\mu(a) = \frac{1}{n} + \delta_a$ ונקבל $\sum_{a\in S}\delta_a = 0$ ו־$\sum_{a\in S}|\delta_a| = 2d(\mu,\pi_S)$. כמסקנה מכך נקבל $\|\mu\|_2^2 = \frac{1}{n} + \sum_{a\in S}\frac{2}{n}\delta_a + \sum_{a\in S}\delta_a^2 \ge \frac{1}{n} + \frac{1}{n}(2d(\mu,\pi_S))^2$. הסבר: המחובר האמצעי מתאפס, ועבור המחובר הימני כותבים לפי משפט קושי־שווארץ $(2d(\mu,\pi_S))^2 = (\sum_{a\in S}1\cdot|\delta_a|)^2 \le (\sum_{a\in S}1^2)(\sum_{a\in S}\delta_a^2)$.

השיערוך של $\|\mu\|_2^2$ יהיה לפי ספירת חזרות. לכל $1 \le i < j \le q$ נסמן ב־X_{ij} את משתנה האינדיקטור עבור המאורע $A_i = A_j$, ונסמן ב־$X = \sum_{1\le i<j\le q}X_{ij}$ את מספר החזרות הכולל. נשים לב שהתוחלת מקיימת $\mathrm{E}[X_{ij}] = \Pr[A_i = A_j] = \sum_{a\in S}\Pr[A_i = a \wedge A_j = a] = \|\mu\|_2^2$, ולכן $\mathrm{E}[X] = \binom{q}{2}\|\mu\|_2^2$.

ננסה לשערך את הנורמה ע"י $X/\binom{q}{2}$, אבל בשביל זה צריך גם לחסום את הסיכוי לסטיה מהותית מהתוחלת. כאן נשתמש בשיטת המומנט השני, ונחסום על כן את $\mathrm{V}[X] = \sum_{1\le i<j\le q\,,\,1\le i'<j'\le q}\mathrm{Cov}[X_{ij}, X_{i'j'}]$. הסכום הזה מחושב בצורה הבאה:

- אם אין איברים משותפים ל־$\{i,j\}$ ו־$\{i',j'\}$, אז X_{ij} ו־$X_{i'j'}$ הם ב"ת ולכן $\mathrm{Cov}[X_{ij}, X_{i'j'}] = 0$.
- אם יש איבר משותף יחיד בין $\{i,j\}$ ו־$\{i',j'\}$, למשל אם $i = i'$ אבל $j \ne j'$, אז ראשית מחשבים את $\mathrm{E}[X_{ij}X_{ij'}] = \sum_{a\in S}\Pr[A_i = A_j = A_{j'} = a] = \sum_{a\in S}|\mu(s)|^3 = \|\mu\|_3^3 \le \|\mu\|_2^3$. שימו לב לשימוש באי־שוויון הנורמות, שקובע ש־$\|v\|_\beta \le \|v\|_\alpha$ לכל $1 \le \alpha < \beta \le \infty$. חישוב הקווריאנס עצמו יתן $\mathrm{Cov}[X_{ij}, X_{ij'}] = \mathrm{E}[X_{ij}X_{ij'}] - \mathrm{E}[X_{ij}]\mathrm{E}[X_{ij'}] \le \mathrm{E}[X_{ij}X_{ij'}] \le \|\mu\|_2^3$. החישוב יצא זהה גם למקרים האחרים של איבר משותף, כמו למשל $i < j = i' < j'$.
- אם $\{i,j\} = \{i',j'\}$, אז $\mathrm{Cov}[X_{ij}, X_{ij}] = \mathrm{V}[X_{ij}] = \mathrm{E}[X_{ij}^2] - (\mathrm{E}[X_{ij}])^2 \le \mathrm{E}[X_{ij}^2] = \|\mu\|_2^2$ (שימו לב שמכיוון שהמשתנה מקבל ערכים מ־$\{0,1\}$ בלבד, מתקיים $X_{ij} = X_{ij}^2$).

עתה אפשר לחסום את $\mathrm{V}[X] = \sum_{1\le i<j\le q\,,\,1\le i'<j'\le q}\mathrm{Cov}[X_{ij}, X_{i'j'}] \le \binom{q}{2}\|\mu\|_2^2 + 6\binom{q}{3}\|\mu\|_2^3$. לכל $0 < \alpha \le 1$, אם $q \ge 24/\alpha^2\|\mu\|_2$ אז מתקיים $\mathrm{V}[X]/(\alpha\binom{q}{2}\|\mu\|_2^2)^2 < \frac{1}{3}$, ואז לפי אי־שוויון צ'בישב מתקיים $\Pr[|X - \binom{q}{2}\|\mu\|_2^2| > \alpha\binom{q}{2}\|\mu\|_2^2] < \frac{1}{3}$. זה אומר שעם מספר דגימות של $q = \lceil 24\sqrt{n}/\alpha\rceil \ge 24/\alpha^2\|\mu\|_2$, בהסתברות לפחות $\frac{2}{3}$ ההערכה $r = X/\binom{q}{2}$ תקיים $r = (1\pm\alpha)\|\mu\|_2^2$.

במקרה שלנו נניח שמתקיים $\epsilon \le \frac{1}{2}$ (אחרת פשוט נעשה $\frac{1}{2}$־בדיקה במקום ϵ־בדיקה), ונבחר $\alpha = \epsilon^2$. אנחנו נבצע אם כן $q = O(\sqrt{n}/\epsilon^4)$ דגימות, ונקבל את μ אם ההערכה שלנו תקיים $r \le (1+\epsilon^2)\frac{1}{n}$. אם ההתפלגות יוניפורמית אז $\|\mu\|_2^2 = \frac{1}{n}$ ואז נקבל את הקלט בהסתברות לפחות $\frac{2}{3}$. לעומת זאת, אם ההתפלגות היא ϵ־רחוקה מיוניפורמיות אז מתקיים $\|\mu\|_2^2 \ge (1+4\epsilon^2)\frac{1}{n}$. בהסתברות לפחות $\frac{2}{3}$ השיערוך יקיים $r \ge (1+4\epsilon^2)(1-\epsilon^2)\frac{1}{n} > (1+\epsilon^2)\frac{1}{n}$ ואכן נדחה את הקלט.

כהכנה להמשך, נראה האם אפשר להבטיח קבלה (בהסתברות $\frac{2}{3}$) גם של קלטים שאינם בדיוק יוניפורמים, אלא רק קרובים ליוניפורמיות. עם קירבה במושגים של מרחק התפלגויות זה לא יעבוד. לדוגמה, עבור כל η קבוע, ההתפלגות המוגדרת ע"י $\mu(1) = \frac{1}{n} + \eta$ ו־$\mu(i) = \frac{1}{n} - \frac{\eta}{n-1}$ ל־$2 \le i \le n$ (שהיא η־קרובה ליוניפורמיות) תקיים $\|\mu\|_2^2 > \eta^2$, נורמה גדולה מהסף $(1+4\epsilon^2)\frac{1}{n}$ שעבורו חייבים לדחות. גרוע מכך, כיום כבר ידוע חסם תחתון של $n^{1-o(1)}$ עבור כל בדיקה שחייבת גם לקבל התפלגויות קרובות ליוניפורמיות, לפי המאמר Valiant: Testing symmetric properties of distributions.

המצב יותר טוב אם ההתפלגות קרובה ליוניפורמיות במובן יותר חזק. נגיד ש־μ היא η־יוניפורמית אם לכל a ו־b מתקיים $\mu(a) \le (1+\eta)\mu(b)$. במקרה כזה בפרט מתקיים $\mu(a) \le (1+\eta)\frac{1}{n}$ לכל a (בגלל שחייב להיות אינדכס עבורו ההסתברות אינה עולה על $\frac{1}{n}$), וגם $\mu(a) \ge \frac{1}{n}/(1+\eta) \ge (1-\eta)\frac{1}{n}$. שוב נסמן $\mu(a) = \frac{1}{n} + \delta_a$, ונשתמש ב־$|\delta_a| \le \frac{\eta}{n}$ לקבלת $\|\mu\|_2^2 = \frac{1}{n} + \sum_{a\in S}\frac{2}{n}\delta_a + \sum_{a\in S}\delta_a^2 \le (1+\eta^2)\frac{1}{n}$.

אם נחזור אלינו, נניח ש־$\epsilon \le \frac{1}{3}$ (אנחנו לא נצטרך את זה אח"כ ל־ϵ גדולים יותר). נבצע את הבדיקה כפי שבצענו אותה קודם, רק שעתה נקבל את הקלט אם השערוך שלנו מקיים $X/\binom{q}{2} \le (1+\frac{5}{2}\epsilon^2)\frac{1}{n}$. חישוב ישיר יראה שכאשר מתקיים $X/\binom{q}{2} = (1\pm\epsilon^2)\|\mu\|_2^2$ (שכזכור קורה בהסתברות לפחות $\frac{2}{3}$), אנחנו אכן נקבל את μ אם היא ϵ־יוניפורמית ונדחה אותה אם היא ϵ־רחוקה (במרחק התפלגויות) מיוניפורמיות.

לסיום, נעיר שהניתוח שלנו אינו מיטבי. אפשר להסתפק ב־$O(\sqrt{n}/\epsilon^2)$ בדיקות בלבד, כפי שמוסבר במאמר Paninski: A coincidence-based test for uniformity given very sparsely sampled discrete data.

חלוקה לדליים ובדיקה מול התפלגות קיימת

כל עוד אנחנו נמצאים במודל של בדיקת תכונות סימטריות של התפלגויות ע"י דגימות בלבד, ספירת התנגשויות (על מנת לבדוק את הנורמה $\|\mu\|_2^2$, ולפעמים גם נורמות $\|\mu\|_3^3$ וכו) היא פחות או יותר הדבר היחידי שאפשר לעשות. בהרבה מקרים זה מועיל להעביר את הבעיה המקורית לכזו של בדיקת יוניפורמיות, עם האופציה לקבלת קלטים ϵ־יוניפורמים גם.

טכניקה שימושית לכך היא חלוקה לדליים (bucketing). נדגים אותה עבור בדיקה של μ עבור שוויון להתפלגות ידועה τ מעל S. אנחנו נרצה לחלק את S ל"איזורי יוניפורמיות" של τ. אנחנו יודעים שערכי τ הם בין 0 ל־1. כמו כן, אנחנו נתעלם מערכים "קטנים מדי", כאלו שגם אם סוכמים על כולם הסכום יהיה קטן מ־ϵ. על כן ה"דלי" הראשון שלנו יהיה $S_0 = \{a\in S : \tau(a) < \frac{\epsilon}{n}\}$.

לכל $j \ge 1$, נגדיר את הדלי $S_j = \{a \in S : \frac{\epsilon}{n}(1+\epsilon)^{j-1} \le \tau(a) < \frac{\epsilon}{n}(1+\epsilon)^j\}$. הדבר הראשון לשים לב הוא שההתפלגות המותנה $\tau|_{S_j}$ היא ϵ־יוניפורמית, לפי הגדרה. כמו כן, נשים לב שעבור $j > \lceil\log_{1+\epsilon}(n/\epsilon)\rceil = O(\log(n/\epsilon)/\epsilon)$ מתקיים $S_j = \emptyset$ (כי לא יכולים להיות ל־τ ערכים גדולים מ־1), ולכן יש לנו חסם על גודל החלוקה. נסמן את החלוקה ב־$\mathcal{B} = \{S_0,\dots,S_r\}$, כאשר $r = \lceil\log_{1+\epsilon}(n/\epsilon)\rceil$.

ננתח את האלגוריתם הבא, שאמור לבדוק האם התפלגות הקלט μ זהה להתפלגות τ שידועה לנו מראש. לשם הנוחות נסמן עבור כל $S' \subseteq S$ את ההסתברות למאורע המתאים ב־$\mu(S') = \Pr_\mu[S'] = \sum_{a\in S}\mu(a)$. כמו כן נסמן ב־$\mu_{\mathcal{B}}$ את ההתפלגות מעל $\{0,\dots,r\}$ שמקיימת $\mu_{\mathcal{B}}(j) = \mu(S_j)$, ונשתמש בסימונים דומים עבור ההתפלגות τ. נשים לב שעבור $a \in S_j$ מתקיים $\mu(a) = \mu_{\mathcal{B}}(j)\cdot\mu|_{S_j}(a)$.

הקבוע C יהיה זה שעבורו אפשר להבחין בין התפלגות ϵ־יוניפורמית מעל $\{1,\dots,n\}$ לבין התפלגות ϵ־רחוקה מיוניפורמיות באמצעות לא יותר מ־$C\sqrt{n}/\epsilon^2$ דגימות. כזכור ציינו שבדיקה כזו אפשרית, למרות שהוכחנו כאן רק אחת עם מספר דגימות גבוה יותר.

- מבצעים q דגימות, שנסמן אותן ב־$A_1,\dots,A_q$.
- לכל $0 \le j \le r$ נגדיר את $Q_j = \{i : A_i \in S_j\}$ (כזכור $S_0,\dots,S_r$ זו החלוקה לדליים של τ, שאותה אנחנו יודעים מראש).
- אם קיים $0 \le j \le r$ שעבורו $|Q_j|/q < \tau(S_j) - \epsilon/(r+1)$ אז דוחים את הקלט; אחרת ממשיכים.
- לכל $1 \le j \le r$ שעבורו $|Q_j| \ge 40C\sqrt{|S_j|}\lceil\log(r)\rceil/\epsilon^2$, מתייחסים לדגימות A_i עם $i \in Q_j$ כאל דגימות מ־$\mu|_{S_j}$, ומשתמשים באלו עבור $\lceil 40\log(r)\rceil$ הרצות ב"ת של בדיקת יוניפורמיות. דוחים את הקלט מיידית אם יותר מחצי מההרצות הנ"ל דחו את $\mu|_{S_j}$.
- אם לא הייתה דחיה עד כאן, מקבלים את הקלט.

אנחנו נוכיח שעבור $q = \tilde{O}(\sqrt{n}/\epsilon^3)$ מתאים, האלגוריתם הזה הוא אלגוריתם בדיקה עבור זהות עם τ (אפשר לשפר את החלק של התלות ב־$\log(n)$ ו־$\log(1/\epsilon)$ שחבוי בסימון למעלה, אבל לא נעשה את זה כאן). ראשית נראה שאם $\tau = \mu$ ו־$\epsilon \le \frac{1}{3}$ אז האלגוריתם יקבל בהסתברות לפחות $\frac{2}{3}$:

- אם $q \geq 2(r+1)^2\log(r)/\epsilon^2 = \tilde{O}((\log(n/\epsilon))^2/\epsilon^2) = o(\sqrt{n}/\epsilon^3)$, אז לכל $0 \leq j \leq r$, כאשר לוקחים q דגימות ב"ת מתוך $\mu = \tau$, ההסתברות שיתקיים $|Q_j|/q < \mu(S_j) - \epsilon/(r+1)$ היא $o(1/r)$, ולכן קטנה מ־$1/6(r+1)$ אם מניחים ש־n ולכן r גדולים מספיק. על כן ההסתברות שתהיה דחיה של הקלט בגלל הגודל של Q_j עבור j כל שהוא היא קטנה מ־$\frac{1}{6}$.

- לכל $1 \leq j \leq r$, כל דגימה שהאינדכס שלה ב־Q_j מתפלגת כדגימה מתוך $\mu|_{S_j} = \tau|_{S_j}$. על כן, מכיוון ש־$\tau|_{S_j}$ היא ϵ־יוניפורמית (ו־$\epsilon \leq \frac{1}{3}$), כל הרצה של בדיקת היוניפורמית תדחה בהסתברות לכל היותר $\frac{2}{3}$. על כן לכל j שעבורו Q_j גדול מספיק לביצוע $40\log(r)$ הרצות ב"ת של הבדיקה, הסיכוי לדחות בגלל j קטן מ־$1/9r$ (ולכל j שעבורו Q_j אינו גדול מספיק אנחנו פשוט לא עושים בדיקות ולא דוחים בגללו). על כן סה"כ הסיכוי לדחות עבור איזה שהוא j בגלל בדיקות היוניפורמיות קטן מ־$\frac{1}{6}$.

- סה"כ ההסתברות לדחיה של הקלט מסיבה כל שהיא חסומה ע"י $\frac{2}{6} = \frac{1}{3}$, כנדרש.

עתה נלך בכיוון ההפוך. אנחנו נראה שאם אף אחד מהשלבים אינו דוחה בהסתברות גבוהה, ומספר הדגימות גדול מספיק, אז הקלט μ בהכרח יהיה 7ϵ־קרוב ל־τ. עבור ϵ־בדיקה, פשוט מבצעים את התהליך עם $\epsilon' = \epsilon/7$ במקום עם ϵ, וזה גם יבטיח שמתקיים $\epsilon' \leq \frac{1}{3}$.

- אם $d(\mu_{\mathcal{B}}, \tau_{\mathcal{B}}) > 2\epsilon$, אז זה אומר שקיים j שעבורו $\mu(S_j) < \tau(S_j) - 2\epsilon/(r+1)$. הסיבה לכך היא שעבור כל שתי התפלגויות α ו־β מעל T כל שהיא מתקיים $d(\alpha,\beta) = \sum_{t\in T:\beta(t)>\alpha(t)}(\beta(t)-\alpha(t))$. אם $q \geq 2(r+1)^2/\epsilon^2 = \tilde{O}((\log(n/\epsilon))^2/\epsilon^2) = o(\sqrt{n}/\epsilon^3)$, אז בפרט עבור ה־$j$ הספציפי הזה, בהסתברות שעולה על $\frac{5}{6}$ יתקיים $|Q_j|/q \leq \mu(S_j) + \epsilon/(r+1) < \tau(S_j) - \epsilon/(r+1)$ והקלט ידחה.

- אם קיים j שעבורו $d(\mu|_{S_j}, \tau|_{S_j}) > 2\epsilon$, אז מכיוון ש־$\tau|_{S_j}$ היא בפרט ϵ־יוניפורמית ולכן היא גם ϵ־קרובה ליוניפורמיות, לפי אי שוויון המשולש $\mu|_{S_j}$ היא ϵ־רחוקה מיוניפורמיות. אם בנוסף $\tau(S_j) \geq 2\epsilon/(r+1)$, אז אחד הדברים הבאים יקרה אם בוחרים $q \geq 40C\sqrt{|S_j|}(r+1)\lceil\log(r)\rceil/\epsilon^3 = \tilde{O}(\sqrt{n}/\epsilon^3)$: האפשרות הראשונה הוא שיתקיים $|Q_j| < 40C\sqrt{|S_j|}\lceil\log(r)\rceil/\epsilon^2$ ואז הקלט ממילא ידחה לפי הסעיף הקודם. האפשרות השניה היא שמתקיים $|Q_j| \geq 40C\sqrt{|S_j|}\lceil\log(r)\rceil/\epsilon^2$, וכאשר זה קורה, בהסתברות שעולה על $\frac{5}{6}$ בדיקת היוניפורמיות תגלה את החריגה בהתפלגות $\mu|_{S_j}$, והקלט ידחה.

נסכם כאן: אם μ מתקבל בהתסברות לפחות $\frac{2}{3}$ כאשר בחרנו למשל $q = \lceil 40C\sqrt{n}(r+1)\lceil\log(r)\rceil/\epsilon^3\rceil$ (זה חוסם את כל הדרישות שכתבנו על q), אז מתקיים גם $d(\mu_{\mathcal{B}}, \tau_{\mathcal{B}}) \leq 2\epsilon$, וגם לכל $1 \leq j$ שעבורו $\tau(S_j) \geq 2\epsilon/(r+1)$ מתקיים $d(\mu|_{S_j}, \tau|_{S_j}) \leq 2\epsilon$. נראה שנובע מזה חסם על $d(\mu,\tau)$. רושמים:

$$\begin{aligned}
d(\mu,\tau) &= \frac{1}{2}\sum_{a\in S}|\mu(a)-\tau(a)| = \frac{1}{2}\sum_{j=0}^{r}\sum_{a\in S_j}\left|\mu(S_j)\mu|_{S_j}(a) - \tau(S_j)\tau|_{S_j}(a)\right| \\
&= \frac{1}{2}\sum_{j=0}^{r}\sum_{a\in S_j}\left|\mu(S_j)\mu|_{S_j}(a) - \tau(S_j)\mu|_{S_j}(s) + \tau(S_j)\mu|_{S_j}(s) - \tau(S_j)\tau|_{S_j}(a)\right| \\
&\leq \frac{1}{2}\sum_{j=0}^{r}\sum_{a\in S_j}\left|\mu(S_j)-\tau(S_j)\right|\mu|_{S_j}(a) + \frac{1}{2}\sum_{j=0}^{r}\sum_{a\in S_j}\left|\mu|_{S_j}(a) - \tau|_{S_j}(a)\right|\tau(S_j) \\
&= d(\mu_{\mathcal{B}}, \tau_{\mathcal{B}}) + \sum_{j=0}^{r} d(\mu|_{S_j}, \tau|_{S_j})\cdot\tau(S_j)
\end{aligned}$$

במקרה שלנו: כזכור $d(\mu_{\mathcal{B}}, \tau_{\mathcal{B}}) \leq 2\epsilon$. עבור הסכום השני, נפצל אותו ל־$j=0$, ל־$1 \leq j \leq r$ שעבורם $\tau(S_j) < 2\epsilon/(r+1)$, ול־$1 \leq j \leq r$ שעבורם $\tau(S_j) \geq 2\epsilon/(r+1)$. עבור שני המקרים הראשונים, אפילו

אם לכולם מתקיים $d(\mu|_{S_j},\tau|_{S_j}) = 1$, הסכום יהיה חסום ע"י 3ϵ (חסום ע"י ϵ עבור $j=0$ וע"י 2ϵ עבור כל השאר). עבור כל j שנותר לנו אנחנו יודעים שמתקיים $d(\mu|_{S_j},\tau|_{S_j}) \le 2\epsilon$, ולכן הסכום של המחוברים הנ"ל גם חסום ע"י 2ϵ (כי מתקיים $\sum_{j=0}^{r}\tau(S_j) = 1$). סה"כ נקבל $d(\mu,\tau) \le 7\epsilon$, כנדרש.

בדיקה באמצעות למידה של התפלגויות

נראה דוגמה ראשונה לטכניקה שנקראת "בדיקה באמצעות למידה". הרעיון הכללי הוא לפי הסכימה הבאה:

- מגדירים "מאפיין", בעצם תכונה עם פרמטרים, שיהיה כללי ככל שניתן. הפרמטרים יכולים להיות תלויים בפרמטר דיוק הלמידה ϵ שמופיע בסעיף הבא.
- בונים "אלגוריתם למידה" – מראים קיום אלגוריתם שמבצע מספר שאילתות קטן יחסית (תלוי בפרמטרים של המאפיין ולפעמים גם ב־$n = |D|$), ועבור כל קלט $f : D \to R$ מחזיר פונקציה $g : D \to R$ או סימן מיוחד לדחיה "$\perp$". אם f מקיימת את המאפיין, אז האלגוריתם חייב בהסתברות $\frac{2}{3}$ לפחות להחזיר פונקציה g שהיא ϵ־קרובה ל־f (הסכימה כאן היא לבדיקה עם שגיאה דו־צדדית). כמו כן, האלגוריתם לא יחזיר פונקציה g שאינה ϵ־קרובה ל־f בהסתברות גבוהה מ־$\frac{1}{3}$ גם אם f אינה מקיימת את המאפיין, אבל במקרה כזה מותר להחזיר "$\perp$" בכל הסתברות כל שהיא.
- עבור התכונה שרוצים לבדוק, מוכיחים שכל קלט שמקיים את התכונה מקיים את המאפיין (עם פרמטרים מתאימים – הם בד"כ יהיו תלויים ב־ϵ שעבורו רוצים לבצע את הלמידה, ולפעמים תלויים ב־n – אבל תלות חזקה מדי ב־n תסכל את אפשרות הבדיקה).
- לבניית אלגוריתם ϵ־בדיקה, מריצים את אלגוריתם הלמידה עם פרמטר $\frac{\epsilon}{2}$ לדיוק הלמידה (וזה גם משפיע על הפרמטרים שצריך למאפיין התכונה). אם אלגוריתם הלמידה החזיר "$\perp$" או פונקציה g שהיא $\frac{\epsilon}{2}$־רחוקה מהתכונה דוחים את f, ואם הוא החזיר פונקציה $\frac{\epsilon}{2}$־קרובה לתכונה מקבלים את f.

עבור קלט שמקיים את התכונה (ולכן מקיים גם את המאפיין המתאים), בהסתברות $\frac{2}{3}$ לפחות אלגוריתם הלמידה יחזיר פונקציה שהיא $\frac{\epsilon}{2}$־קרובה לקלט (ולכן גם לתכונה), והקלט יתקבל. עבור קלט שהוא ϵ־רחוק מהתכונה, בהסתברות לפחות $\frac{2}{3}$ או שיוחזר "$\perp$" או שתוחזר פונקציה $\frac{\epsilon}{2}$־קרובה לקלט (ולכן לפי אי־שוויון המשולש $\frac{\epsilon}{2}$־רחוקה מהתכונה), ובשני מקרים אלו הקלט יידחה.

מאפיין יוניפורמיות למקוטעין של התפלגויות ותכונת המונוטוניות

הדוגמה שלנו תהיה בתחום של בדיקת התפלגויות מעל $S = \{1,\dots,n\}$. זה אומר שמרחק ימדד במושגים של מרחק התפלגויות, ו"שאילתה" תהיה קבלת דגימה שמתפלגת לפי התפלגות הקלט μ. אם הקלט מקיים את המאפיין, אז אלגוריתם הלמידה צריך בהסתברות גבוהה להחזיר התפלגות ν קרובה ל־μ.

המאפיין אצלנו יהיה זה של יוניפורמיות למקוטעין. אנחנו נגיד ש־μ היא (ϵ,k)־יוניפורמית למקוטעין אם קיימים $0 = l_0 < l_1 < \cdots < l_k = n$, כך שלכל $1 \le i \le k$ מתקיים ש־$\mu|_{\{l_{i-1}+1,\dots,l_i\}}$ היא ϵ־יוניפורמית (במובן "הקירוב הכפלי" שהוגדר בפרק הקודם).

התכונה של מונוטוניות, ז"א שמתקיים $\mu(i) \le \mu(j)$ לכל $1 \le i < j \le n$, תהיה (ϵ,k)־יוניפורמית למקוטעין לכל ϵ עבור $k = O(\log(n)/\epsilon)$. עד מנת להראות את זה, מחלקים את S לדליים $S_0,\dots,S_r$ עבור $r = \lceil \log_{1+\epsilon}(n/\epsilon) \rceil$ כפי שנעשה בפרק הקודם. הצמצום לכל דלי יהיה כזכור ϵ־יוניפורמי, ומכיוון שהתהפלגות מונוטונית, כל דלי S_i יכלול את כל האינדכסים בין הנמוך ביותר והגבוה ביותר שהוכנסו אליו, ז"א שהוא יהיה קטע מתאים מהצורה $\{l_{i-1}+1,\dots,l_i\}$.

בדיקה של מונוטוניות ב־$\tilde{O}(\sqrt{n})$ דגימות (עבור ϵ קבוע) הוכחה לראשונה במאמר Batu, Kumar, Rubinfeld: Sublinear algorithms for testing monotone and unimodal distributions. אלגוריתם למידה עבור

התפלגויות (ϵ,k)־יוניפורמיות ב־$\tilde{O}(\sqrt{kn}/\epsilon^3)$ דגימות פותח לראשונה במאמר Canonne, Diakonikolas, Gouleakis, Rubinfeld: Testing shape restrictions of discrete distributions. מקיום אלגוריתם הלמידה בפרט נובע אלגוריתם הבדיקה עבור מונוטוניות, ותכונות אחרות שיש להם אפיון של יוניפורמיות למקוטעין.

אנחנו נראה את אלגוריתם הלמידה המשופר שפותח בעקבות הקודם במאמר Fischer, Lachish, Vasudev: Improving and extending the testing of distributions for shape-restricted properties. עם זאת לא נתרכז בהשגת היעילות המירבית, ונראה אלגוריתם שמבצע $\tilde{O}(k\sqrt{n})$ דגימות עבור כל ϵ קבוע.

חלוקות עדינות

חלוקה של $\{1,\ldots,n\}$ לקטעים לפי $t_0,\ldots,t_r$ תיקרא η־עדינה ביחס ל־μ, אם לכל $1\le i\le r$ שעבורו $t_{i-1}<t_i-1$ מתקיים $\mu(\{t_{i-1}+1,\ldots,t_i\})\le\eta$ (עבור "קטעים" של אינדכס בודד אנחנו לא דורשים כלום, כי יכולים להיות בהתפלגות אינדכסים j שעבורם $\mu(j)>\eta$).

אם μ היא (ϵ,k)־יוניפורמית למקוטעין, אז במקום לחפש את $\{l_0,\ldots,l_k\}$ המקוריים שמדגימים את זה, מספיק לקחת חלוקה ϵ/k־עדינה שרירותית כל שהיא: יש לכל היותר k ערכים של i שעבורם קיים j (אחד או יותר) עם $t_{i-1}+1<l_j<t_i$. לכל i אחר (כולל כל אלה שעבורם $t_{i-1}=t_i-1$), הצמצום $\mu|_{\{l_{i-1}+1,\ldots,l_i\}}$ הוא ϵ־יוניפורמי. בפרט, המשקל הכולל של קטעים בחלוקה העדינה שעבורם הצמצום של μ אינו ϵ־יוניפורמי חסום ע"י ϵ.

על מנת למצוא חלוקה η־עדינה, נסתכל על הפרוצדורה הבאה: נדגום $s=\lceil\frac{3}{\eta}\ln(\frac{3}{\eta\delta})\rceil$ אינדכסים לפי μ, ולכל אינדכס $\{j\}$ שנדגם נהפוך אותו לקטע של החלוקה. באופן פורמלי – לפני התחלת הדגימה נגדיר $T=\{0,n\}$, ולכל אינדכס i שנדגם נוסיף את $i-1$ ואת i ל־T. בסוף נמיין את T (ללא כפילויות) ונרשום $T=\{t_0,\ldots,t_r\}$ כאשר $0=t_0<\cdots<t_r=n$. נשים לב שעבור δ קבוע מתקיים $r=\tilde{O}(1/\eta)$.

החלוקה המתקבלת תהיה η־עדינה בהסתברות לפחות $1-\delta$. על מנת לראות את זה, ראשית נשים לב שבהסתברות לפחות $1-\frac{\delta}{2}$ אנחנו נדגום את כל האינדכסים i שעבורם $\mu(i)>\frac{\eta}{3}$, ע"י חסם איחוד מאורעות (יש לא יותר מ־$\frac{3}{\eta}$ אינדכסים כאלה, וההסתברות לא לדגום כל i כזה היא $(1-\mu(i))^s<\frac{\eta\delta}{3}$).

נרחיב עתה את השיקול: נחשוב על כל האפשרויות לקטעים $I=\{i,\ldots,j\}\subseteq\{1,\ldots,n\}$ שעבורם $\mu(I)\ge\frac{\eta}{3}$, ומתוך אלו נסתכל על קטעים מאורך מינימלי (ה"אורך" של I הוא $|I|=j+1-i$). נסמן קבוצה זו ב־$\mathcal{I}$. בפרט לכל i שעבורו $\mu(i)\ge\frac{\eta}{3}$ מתקיים $\{i\}\in\mathcal{I}$. נבחר תת־קבוצה $\mathcal{J}\subseteq\mathcal{I}$ מקסימלית של קטעים זרים זה לזה (מספיק אפילו מקסימלית בהכלה – ואם יש כמה אפשרויות אז נבחר אחת מהן שרירותית). בפרט מתקיים $\sum_{I\in\mathcal{J}}\mu(I)\le\mu(\{1,\ldots,n\})=1$ ולכן $|\mathcal{J}|\le\frac{\eta}{3}$. וגם כאן, לפי איחוד מאורעות, בהסתברות לפחות $1-\delta$ הדגימה שלנו תכלול לפחות נקודה אחת מכל $I\in\mathcal{J}$.

נטען שכאשר זה קורה, החלוקה תהיה עדינה: עלינו לבחון את קטעי החלוקה שאינם מורכבים מנקודה בודדת. כזכור, כל נקודה שדגמנו הפכנו לקטע של נקודה בודדת בחלוקה. לכל $I\in\mathcal{J}$ נסמן ב־a_I את אחת הנקודות שדגמנו מתוכו. קטע החלוקה שלנו על כן צריך להיות מוכל בקטע $\{a_I+1,\ldots,a_J-1\}$, כאשר $I,J\in\mathcal{J}$ ואין ביניהם קטע אחר ב־$\mathcal{J}$. אם נסמן ב־$K=\{a_I+1,\ldots,a_J-1\}\setminus(I\cup J)$ את המרווח בין I ל־J, אז מתקיים $\mu(K)<\frac{\eta}{3}$, כי אחרת היינו יכולים להוסיף ל־$\mathcal{J}$ גם תת־קטע של המרווח הנ"ל. כמו כן מתקיים $\mu(I\cap\{a_I+1,\ldots,a_J-1\})<\frac{\eta}{3}$ (אחרת I לא היה מאורך מינימלי), וכן $\mu(J\cap\{a_I+1,\ldots,a_J-1\})<\frac{\eta}{3}$ (אחרת J לא היה מאורך מינימלי). על כן $\mu(\{a_I+1,\ldots,a_J-1\})<\eta$, כנדרש.

אימות יוניפורמיות למקוטעין ולמידת ההתפלגות

עבור ההמשך, נניח שיש בידינו חלוקה $\frac{\epsilon}{k}$־עדינה לפי $0=t_0<\cdots<t_r=n$. נסמן את החלוקה המתאימה לקטעים ב־$\mathcal{B}=\{K_1,\ldots,K_r\}$, כאשר $K_i=\{t_{i-1}+1,\ldots,t_i\}$ לכל $1\le i\le r$.

אם μ היא (ϵ,k)־יוניפורמית למקוטעין, ננסה ללמוד את ההתפלגות בצורה הבאה: נבצע $O(r/\epsilon^2)$ דגימות על מנת שבהסתברות $1-\delta$ לכל $1\le i\le r$ נדע קירוב $\tilde{\mu}(K_i)$, כך שיתקיים $\frac{1}{2}\sum_{i=1}^{r}|\mu(K_i)-\tilde{\mu}(K_i)|\le\epsilon$.

בעצם מקרבים את ההתפלגות $\mu_{\mathcal{B}}$ שמוגדרת מעל $\{1,\ldots,r\}$, דבר שניתן להעשות במספר השאילתות הנ"ל. הקירוב המלא עבור μ יוגדר עבור $j \in K_i$ לפי $\tilde{\mu}(j) = \tilde{\mu}(K_i)/|K_i|$, ז"א שנניח שההתפלגות המותנה על כל קטע K_i היא יוניפורמית.

זה נותן 3ϵ־קירוב של μ: חוסמים את ההפרש בדומה לחסימה שנעשתה ביחס לחלוקה לדליים, לפי אי־השוויון $d(\mu,\tilde{\mu}) \leq d(\mu_{\mathcal{B}},\tilde{\mu}_{\mathcal{B}}) + \sum_{i=1}^{r}\mu(K_i)d(\mu|_{K_i},\tilde{\mu}|_{K_i})$. לכל i שעבורו $\mu|_{K_i}$ היא ϵ־יוניפורמית, המרחק שלה מההתפלגות היוניפורמית $\pi_{K_i} = \tilde{\mu}|_{K_i}$ חסום ע"י ϵ. אם סוכמים את ההפרשים על כל הקטעים ש־μ אינה ϵ־יוניפורמית מעליהם, שמשקלם הכולל הוא ϵ, יתווסף לכל היותר עוד ϵ לסכום אפילו אם לכל i כזה מתקיים $d(\mu|_{K_i},\tilde{\mu}|_{K_i}) = 1$. לבסוף, המרחק בין $\mu_{\mathcal{B}}$ לבין $\tilde{\mu}_{\mathcal{B}}$ גם חסום ע"י ϵ.

בכל הדיון הזה חסרה עדיין דרישה אחת חשובה של אלגוריתם הלמידה: אסור לפלוט התפלגות רחוקה מ־μ אפילו אם μ אינה (ϵ,k)־יוניפורמית למקוטעין. על כן נוסיף חלק שיוודא שבאמת יש יוניפורמיות ברב קטעי החלוקה, ובמידה וזה אינו המצב, נפלוט "$\perp$" ולא את $\tilde{\mu}$. זהו החלק שדורש את מירב הדגימות (עד עכשיו עבור ϵ קבוע השתמשנו ב־$\tilde{O}(k)$ דגימות בלבד).

אנחנו ננסח עכשיו אלגוריתם שדוגם את ממרחב ההתפלגות μ, ובצירוף החלוקה $\mathcal{B}$ והקירוב המוצע $\tilde{\mu}$ שנבנה קודם, בודק שאכן μ היא יוניפורמית עבור מרבית הקטעים $K_i \in \mathcal{B}$ (לפי משקל). הטכניקה כאן תהיה דומה לטכניקה של בדיקה מול התפלגות ידועה, רק שכאן נשתמש בחלוקה העדינה ולא בחלוקה לדליים. גם כאן C יסמן את הקבוע כך ש־$C\sqrt{n}/\epsilon^2$ דגימות מספיקות לבדיקת יוניפורמיות מעל $\{1,\ldots,n\}$.

- מבצעים q דגימות, שנסמן אותן ב־$A_1,\ldots,A_q$.
- לכל $0 \leq i \leq r$ מגדירים את $Q_i = \{j : A_j \in K_i\}$.
- לכל $1 \leq i \leq r$ שעבורו $|Q_i| \geq 40C\sqrt{|K_i|}\log(r/\delta)/\epsilon^2$, מתייחסים לדגימות A_j עם $j \in Q_i$ כאל דגימות מ־$\mu|_{K_i}$, ומשתמשים באלו עבור $40\log(r/\delta)$ הרצות ב"ת של בדיקת ϵ־יוניפורמיות. אם יותר מ־$\frac{1}{2}$ מההרצות דחו, מסמנים את i כ"דחוי" (אבל עוד לא דוחים מיידית את הקלט).
- נסמן ב־N את קבוצת ה־i שדחינו, ונשתמש בקירוב $\tilde{\mu}$ שנבנה קודם עבור אלגוריתם הלמידה. אם מתקיים $\tilde{\mu}_{\mathcal{B}}(N) > 2\epsilon$ אז דוחים את הקלט עם החלוקה המוצעת, ואחרת מקבלים.

עבור $q = \tilde{O}(r\sqrt{n}\log(1/\delta)/\epsilon^3)$ מתאים, בהסתברות כוללת של לפחות $1-\frac{\delta}{2}$, לכל i שעבורו מתקיים $\mu(K_i) \geq \frac{\epsilon}{r}$ נבצע את בדיקות היוניפורמיות. בהסתברות כוללת של לפחות $1-\frac{\delta}{2}$, לכל i שבצענו עבורו את הבדיקות נקבל תשובה נכונה, ז"א שהוא יסומן כדחוי אם $\mu|_{K_i}$ היא ϵ־רחוקה מיוניפורמיות, ולא יסומן כדחוי אם $\mu|_{K_i}$ היא ϵ־יוניפורמית (במקרה שאף אחד מהדברים אינו מתקיים, זה לא משנה איך i יסומן). בהסתברות לפחות $1-\delta$ יתקיימו שני הדברים, ז"א שתהיה בידנו תשובה נכונה לכל i שעבורו $\mu(K_i) \geq \frac{\epsilon}{r}$.

אם הקלט הוא (k,ϵ)־יוניפורמי למקוטעין, $\tilde{\mu}_{\mathcal{B}}$ הוא ϵ־קירוב של $\mu_{\mathcal{B}}$, והחלוקה שלנו היא $\frac{\epsilon}{k}$־עדינה, אז (בהסתברות לפחות $1-\delta$) יתקיים $\mu_{\mathcal{B}}(N) \leq \epsilon$, ולכן $\tilde{\mu}_{\mathcal{B}}(N) \leq 2\epsilon$, והקלט יתקבל. מצד שני, אם סך המשקל לפי μ של הקטעים ה־ϵ־רחוקים מיוניפורמיות עולה על 4ϵ, אז הקלט ידחה: סך המשקל של הקטעים K_i עבורם $\mu(K_i) < \frac{\epsilon}{r}$ קטן מ־ϵ, כל הקטעים הרחוקים האחרים ידחו ולכן יתקיים $\mu_{\mathcal{B}}(N) > 3\epsilon$, וכתוצאה מכך יתקיים $\tilde{\mu}_{\mathcal{B}}(N) > 2\epsilon$.

עתה נוכל להרכיב את אלגוריתם הלמידה המלא שלנו מהאלגוריתמים למעלה. זה יהיה אלגוריתם 6ϵ־למידה. עבור אלגוריתם ϵ־למידה עוברים לפרמטר $\epsilon' = \epsilon/6$ (כולל שימוש ב־k' כך שקלטים שמקיימים את התכונה שלנו יהיו (k',ϵ')־יוניפורמים למקוטעים).

- מבצעים את האלגוריתם למציאת חלוקה $\frac{\epsilon}{k}$־עדינה עבור הקלט μ בהסתברות $1-\delta$, עם $\delta = \frac{1}{9}$.
- עבור החלוקה $\mathcal{B}$ שהתקבלה (כאשר כזכור $|\mathcal{B}| = r = \tilde{O}(k/\epsilon)$), מוצאים (בהסתברות $1-\delta$ עם $\delta = \frac{1}{9}$) קירוב $\tilde{\mu}_{\mathcal{B}}$ של $\mu_{\mathcal{B}}$ עם דיוק ϵ, ומגדירים את $\tilde{\mu}$ מעל $\{1,\ldots,n\}$, כמו למעלה.
- מבצעים את אלגוריתם הווידוא עבור הקלט μ, החלוקה $\mathcal{B}$ והקירוב $\tilde{\mu}$ (עם $\delta = \frac{1}{9}$). אם הווידוא דחה אז פולטים "$\perp$", ואחרת פולטים את $\tilde{\mu}$ כקירוב להתפלגות הקלט.

מספר הדגימות הכולל יהיה $\tilde{O}(k\sqrt{n}/\epsilon^4)$ לפי השלב השלישי שהוא הכי "יקר". עבור הוכחת הנכונות צריך להוכיח שני דברים: שעבור קלט שהוא (k,ϵ)-יוניפורמי למקוטעין לא יוחזר "$\perp$" בהסתברות גדולה מ-$\frac{1}{3}$, ושעבור קלט μ כל שהוא לא יוחזר $\tilde{\mu}$ המקיים $d(\mu,\tilde{\mu}) > 6\epsilon$ בהסתברות גדולה מ-$\frac{1}{3}$. נשים לב שבהסתברות לפחות $\frac{2}{3}$ לא קוראת "תקלה" באף אחד מהשלבים (ז"א שגם $\mathcal{B}$ היא $\frac{\epsilon}{k}$-עדינה, גם $\tilde{\mu}_{\mathcal{B}}$ מקרב את $\mu_{\mathcal{B}}$, וגם אלגוריתם הווידוא מקבל או דוחה לפי התנאים שנכתבו למעלה). על כן נגביל את עצמנו למקרה "חסר התקלות", ונראה שבמקרה זה שתי הדרישות יתקיימו.

- אם μ היא (k,ϵ)-יוניפורמית למקוטעין, אז לפי הדיון על אלגוריתם הווידוא (שמשתמש בהנחות על $\mathcal{B}$ ועל $\tilde{\mu}_{\mathcal{B}}$), אלגוריתם הווידוא יקבל ולכן לא יוחזר "$\perp$".
- לכל μ, אם אלגוריתם הווידוא לא דחה (ז"א שהחזרנו את $\tilde{\mu}$), אז לפי הדיון על אלגוריתם הווידוא בהכרח מתקיים שסה"כ המשקלים על קטעים שבהם μ אינה ϵ-קרובה להתפלגות יוניפורמית חסום ע"י 4ϵ. נחזור לחסם המרחק $d(\mu,\tilde{\mu}) \leq d(\mu_{\mathcal{B}},\tilde{\mu}_{\mathcal{B}}) + \sum_{i=1}^{r}\mu(K_i)d(\mu|_{K_i},\tilde{\mu}|_{K_i})$. המרחק בין $\tilde{\mu}_{\mathcal{B}}$ ל-$\mu_{\mathcal{B}}$ חסום ע"י ϵ, הסכום הממושקל של מרחקי ההתפלגויות המותנות על קטעים רחוקים מיוניפורמיות חסום ע"י 4ϵ (בגלל החסם על המשקל הכולל של קטעים אלו), והסכום הממושקל של מרחקי ההתפלגויות המותנות על קטעים לא רחוקים מיוניפורמיות חסום ע"י ϵ. סה"כ נקבל $d(\tilde{\mu},\mu) < 6\epsilon$.

לסיום, נראה בקווים כלליים איך היה אפשר ליעל את מספר השאילתו בפקטור של $\sqrt{k}$: בשלב שבו חישבנו כמה דגימות צריך בשביל שנוכל לבדוק כל I_j ליוניפורמיות, חסמנו לפי $n \leq |I_j|$. אם היינו מתעלמים מקטעים שעבורם $|I_j| \geq \frac{n}{k}$, אז מספר השאילתות הדרוש (עבור ϵ קבוע) היה מתחלק ב-$O(\sqrt{k})$. מכיוון שאין יותר מ-k קטעים כאלו (והחלוקה היא עדינה), זה היה גורם לכל היותר לאיבוד של ϵ נוסף בקירובים, והיינו מקבלים חסם מרחק של 7ϵ.

בדיקת חונטות

בפרק זה נתמקד בפונקציות $f:\{0,1\}^n \to \{-1,1\}$, כי החישובים המתמטים שלנו יהיו יותר נוחים עבור טווח זה מאשר עבור ה-"$\{0,1\}$" שהשתמשנו בו במקומות אחרים. עבור קבוצה $J \subseteq \{1,\ldots,n\}$ ו-$x = (x_1,\ldots,x_n) \in \{0,1\}^n$, נסמן ב-$x_J \in \{0,1\}^{|J|}$ את הווקטור $(x_{j_1},\ldots,x_{j_{|J|}})$, כאשר מגדירים $J = \{j_1,\ldots,j_{|J|}\}$ כך ש-$j_1 < j_2 < \cdots < j_{|J|}$. אנחנו נגיד שהפונקציה $f:\{0,1\}^n \to \{-1,1\}$ היא k-חונטה, אם היא תלויה ב-k קורדינטות ("משתנים") בלבד, ז"א שקיימת קבוצה $J \subset \{1,\ldots,n\}$ מגודל k ופונקציה $h:\{0,1\}^k \to \{-1,1\}$, כך שלכל $x \in \{0,1\}^n$ מתקיים $f(x) = h(x_J)$.

אנחנו נרצה לבדוק את התכונה שהפונקציה f היא k-חונטה. בדיקה כזו, עם מספר שאילתות פולינומי ב-k ו-ϵ, פותחה לראשונה במאמר Fischer, Kindler, Ron, Safra, Samorodnitsky: Testing juntas. מחקרים יותר מאוחרים שיפרו את מספר השאילתות, עד למאמר Blais: Testing juntas nearly optimally שהשיג מספר שאילתות של $O(k\log(k) + k/\epsilon)$. אנחנו נראה את ההוכחה של המאמר הראשון כי היא יותר פשוטה, ויש לה גם יתרון נוסף עבורנו, שהיא משתמשת ב"תחליף" לאנליזת פוריה המקובלת עבור בדיקות מסוג זה.

מידת השתנות של פונקציה מעל קבוצות משתנים

עבור קבוצה $J \subset \{1,\ldots,n\}$, יהיה חשוב לנו לדעת עד כמה שינוי של $x \in \{0,1\}^n$ בתוך קבוצה זו עלול לגרום לשינוי בערך של $f(x)$. אנחנו נגדיר מידה פורמלית שקשורה בזה, ונוכיח מספר תכונות "אלגבריות" שלה, כמו למשל מונוטוניות (ז"א שההשתנות מעל קבוצה היא לפחות ההשתנות מעל תת-קבוצה שלה).

ההגדרה שלנו תתבסס על מושגים כמו שונות של מ"מ. מרחב ההסתברות הבסיסי שלנו יהיה זה של הגרלה יוניפורמית של x מתוך $\{0,1\}^n$. עבור קבוצה J ועבור וקטור $y \in \{0,1\}^{n-|J|}$, נבחן את השונות במרחב ההתפלגות המותנה, $\mathrm{V}[f(x)|x_{\{1,\ldots,n\}\setminus J} = y]$. מכיוון ש-$f(x)$ מקבל ערכים ב-$\{-1,1\}$, מתקיים $\mathrm{V}[f(x)|x_{\{1,\ldots,n\}\setminus J} = y] = 2\mathrm{Pr}_{x,x'}[f(x) \neq f(x')|x_{\{1,\ldots,n\}\setminus J} = x'_{\{1,\ldots,n\}\setminus J} = y]$.

יהיה לנו בהמשך יותר נוח להשתמש בסימון של "שרשור וקטורים". עבור J נתונה, $y \in \{0,1\}^{|J|}$ ו־$z \in \{0,1\}^{n-|J|}$, נסמן ב־$y \sqcup z$ את הווקטור x שעבורו $x|_J = y$ ו־$x|_{\{1,\ldots,n\}\setminus J} = z$. בסימונים כאלו השוויון למעלה ייכתב $\mathrm{V}_y[f(y \sqcup z)] = 2\mathrm{Pr}_{y,y'}[f(y \sqcup z) \neq f(y' \sqcup z)]$. הסימון "$\mathrm{V}_y$" משמעו השונות תחת הגרלה יוניפורמית של $y \in \{0,1\}^{|J|}$, והסימון "$\mathrm{Pr}_{y,y'}$" משמעו ההסתברות תחת הגרלה יוניפורמית ב"ת של y ו־y'.

ההשתנות של f מעל J מוגדרת כתוחלת של השונות הנ"ל, כאשר מגרילים את y עצמו יוניפורמית מתוך $\{0,1\}^{n-|J|}$. באופן פורמלי, $V_f(J) = \mathrm{E}_y\big[\mathrm{V}_x[f(x \sqcup y)]\big]$. בהתאמה לדיון בשונות למעלה, ההשתנות מעל J תהיה שווה ל־$2\mathrm{Pr}_{y,y',z}[f(y \sqcup z) = f(y' \sqcup z)]$, כאשר z מוגרל יוניפורמית מ־$\{0,1\}^{n-|J|}$, והפעולה של ה"שרשור" מחשיבה את z כקובע הערכים מעל $\{1,\ldots,n\} \setminus J$.

בהמשך נרצה להבחין, עבור קבוצה J, בין המקרה ש־f כלל אינה תלויה בקורדינטות של J לבין המקרה שיש לקבוצה השתנות גדולה מ־α. לשם כך, מבצעים $O(\log(1/\delta)/\alpha)$ הרצות ב"ת של הגרלה של y, y', z כמו למעלה ובדיקה האם מתקיים $f(y \sqcup z) = f(y' \sqcup z)$ (שתי שאילתות לכל הרצה). ככה נוכל בהמשך לקבל בהסתברות 1 קבוצה J ש־f אינה תלויה בה, ו"לסמן" (לא בהכרח נדחה את הקלט במקרה כזה) בהסתברות לפחות $1-\delta$ קבוצה J עם השתנות גדולה מ־α.

אפשר להשתמש בהשתנות על־מנת לחסום את המרחק של הפונקציה מאי־תלות ב־J: עבור $z \in \{0,1\}^{n-|J|}$ קבוע, אם המרחק של $g(y) = f(y \sqcup z)$ מפונקציה קבועה הוא $\eta \leq \frac{1}{2}$, ז"א שהערך הכי נפוץ שלה מתקבל עבור $(1-\eta)2^{|J|}$ מהמחרוזות האפשריות עבור z, אז הגרלה של y, y' ב"ת תתן ערכים שונים של g בהסתברות לפחות $\eta \geq 2\eta - 2\eta^2 = 1-(1-\eta)^2-\eta^2$. זה אומר (לאחר שלוקחים תוחלת עבור z שמוגרל יוניפורמית) שהמרחק של f מהפונקציה הקרובה ביותר שאינה תלויה ב־J חסום ע"י $\frac{1}{2}V_f(J) = \mathrm{Pr}_{y,y',z}[f(y \sqcup z) = f(y' \sqcup z)]$.

לפני שנמשיך, נראה תכונות שימושיות של השונות. עבור קבוצות זרות I ו־J, ועבור $z \in \{0,1\}^{n-|I|-|J|}$, מתקיים $\mathrm{V}_{x,y}[f(x \sqcup y \sqcup z)] = \mathrm{E}_x\mathrm{V}_y[f(x \sqcup y \sqcup z)] + \mathrm{V}_x\mathrm{E}_y[f(x \sqcup y \sqcup z)]$, כאשר x מוגרל מתוך $\{0,1\}^{|I|}$ ו־y מוגרל מתוך $\{0,1\}^{|J|}$, והשרשור הוא לפי הקבוצות המתאימות. מוכיחים את זה מההגדרה של שונות:

$$\begin{aligned}
\mathrm{V}_{x,y}[f(xyz)] &= \mathrm{E}_{x,y}[(f(xyz))^2] - (\mathrm{E}_{x,y}[f(xyz)])^2 \\
&= \mathrm{E}_{x,y}[(f(xyz))^2] - \mathrm{E}_x[(\mathrm{E}_y[f(xyz)])^2] + \mathrm{E}_x[(\mathrm{E}_y[f(xyz)])^2] - (\mathrm{E}_{x,y}[f(xyz)])^2 \\
&= \mathrm{V}_x\mathrm{E}_y[f(xyz)] + \mathrm{E}_x\mathrm{V}_y[f(xyz)]
\end{aligned}$$

למעלה רשמנו "xyz" במקום "$x \sqcup y \sqcup z$" על מנת שיהיה מקום למשוואות.

התכונה השימושית השניה שנצטרך היא אי השוויון $\mathrm{V}_x\mathrm{E}_y[f(x \sqcup y \sqcup z)] \leq \mathrm{E}_y\mathrm{V}_x[f(x \sqcup y \sqcup z)]$. ההוכחה שלו היא גם לפי ההגדרות, בתוספת שימוש באי־שוויון קושי־שוורץ (או לחילופין אי־שוויון ינסן Jensen), ואתם מוזמנים לקרוא אותה במאמר המקורי.

עתה אפשר להוכיח מספר תכונות שימושיות של מידת ההשתנות.

- מונוטוניות – לכל $I, J \subseteq \{1,\ldots,n\}$ מתקיים $V_f(I) \leq V_f(I \cup J)$. מספיק להוכיח את זה במקרה שהמדובר בת"ק זרות. לשם כך רושמים (כאשר x הוא מעל I, y הוא מעל J, ו־z הוא מעל השאר):

$$\begin{aligned}
V_f(I \cup J) = \mathrm{E}_z\mathrm{V}_{x,y}[f(x \sqcup y \sqcup z)] &= \mathrm{E}_z\big[\mathrm{E}_x\mathrm{V}_y[f(x \sqcup y \sqcup z)] + \mathrm{V}_x\mathrm{E}_y[f(x \sqcup y \sqcup z)]\big] \\
&\geq \mathrm{E}_{z,x}\mathrm{V}_y[f(x \sqcup y \sqcup z)] = V_f(I)
\end{aligned}$$

- תת־חיבוריות (סאב־אדיטיביות) – לכל $I, J \subseteq \{1,\ldots,n\}$ מתקיים $V_f(I \cup J) \leq V_f(I) + V_f(J)$. לאחר שהוכחנו מונוטוניות מספיק להוכיח את זה לת"ק זרות. רושמים:

$$\begin{aligned}
V_f(I \cup J) &= \mathrm{E}_z\big[\mathrm{E}_x\mathrm{V}_y[f(x \sqcup y \sqcup z)] + \mathrm{V}_x\mathrm{E}_y[f(x \sqcup y \sqcup z)]\big] \\
&\leq \mathrm{E}_z\big[\mathrm{E}_x\mathrm{V}_y[f(x \sqcup y \sqcup z)] + \mathrm{E}_y V_x[f(x \sqcup y \sqcup z)]\big] = V_f(J) + V_f(I)
\end{aligned}$$

- השתנות שולית פוחתת – לכל $I,J,K \subseteq \{1,\ldots,n\}$ זרות זו לזו מתקיים אי־השוויון של "האיחוד עם K מוסיף פחות אם זה לקבוצה מכילה": $V_f(I\cup J\cup K)-V_f(I\cup J)\le V_f(I\cup K)-V_f(I)$. שימו לב שזאת הכללה של תת־חיבוריות (מציבים $I=\emptyset$ ומעבירים אגפים). גם ההוכחה דומה, ואתם מוזמנים לקרוא אותה במאמר המקורי.

לפני שנעבור לאלגוריתם, נגדיר מידה נוספת, "השתנות שפה", שתחסום את ההשתנות מלמטה, ותהיה נוחה לניתוח "חיבורי". המידה תוגדר ביחס לפונקציה f, וקבוצה J שנרצה להוציא מהניתוח הזה. נגדיר לכל קורדינטה $1\le i\le n$ את המידה $U_{f,J}(i)=V_f(\{1,\ldots,i\}\setminus J)-V_f(\{1,\ldots,i-1\}\setminus J)$ (כאשר $V_f(\emptyset)=0$ בכל פעם שזה מופיע בחישוב), ועבור קבוצות פשוט נגדיר את הסכום $U_{f,J}(I)=\sum_{i\in I}U_{f,J}(i)$.

מצד אחד, לפי חישוב הסכום הטלסקופי, מתקיים $U_{f,J}(\{1,\ldots,n\})=V_f(\{1,\ldots,n\}\setminus J)$. מצד שני אפשר להוכיח באינדוקציה על $|I|$ שלכל קבוצה I זרה ל־J מתקיים $V_f(I)\ge U_{f,J}(I)$. הבסיס, $|I|=0$, ברור (שני האגפים שווים ל־0). בצעד האינדוקציה משתמשים בתכונת ההשתנות השולית הפוחתת. עבור האיבר i הכי גבוה ב־I מתקיים:

$$\begin{aligned}U_{f,J}(I)-U_{f,J}(I\setminus\{i\})=U_{f,J}(i) &= V_f(\{1,\ldots,i\}\setminus J)-V_f(\{1,\ldots,i-1\}\setminus J)\\ &\le V_f(I)-V_f(I\setminus\{i\})\end{aligned}$$

ממונוטוניות מתקיים $V_f(I)\ge V_f(I\setminus J)\ge U_{f,J}(I)$ גם אם I אינה זרה ל־J.

אלגוריתם הבדיקה

ננסח כאן את האלגוריתם הכי פשוט לניתוח, ונסתפק במספר שאילתות לא אופטימלי, אבל עדיין פולינומי ב־k ו־$1/\epsilon$. הרעיון הוא לנסות "להפריד" בין קורדינטות שגורמות לתלות גבוהה ע"י חלוקה מקרית של קבוצת הקורדינטות $\{1,\ldots,n\}$. אם הפונקציה היא k־חונטה, אז היא תהיה תלויה בלא יותר מ־k מהקבוצות בחלוקה. לעומת זאת, נטען שעבור פונקציה רחוקה מחונטה יהיו יותר קבוצות עם השתנות ברת גילוי. נניח ש־$\epsilon\le\frac{1}{2}$ (בכל מקרה כל פונקציה עם הטווח $\{-1,1\}$ היא $\frac{1}{2}$־קרובה להיות קבועה). האלגוריתם:

- מחלקים את $\{1,\ldots,n\}$ ל־$r=16k^2$ קבוצות $I_1,\ldots,I_r$, ע"י כך שעבור כל i נבחר באופן יוניפורמי וב"ת $1\le j\le r$ ונקבע $i\in I_j$.
- לכל $1\le j\le r$, נשתמש בבדיקה מתת־הפרק הקודם על מנת להבחין בין המקרה $V_f(I_j)=0$ (הפונקציה f אינה תלויה באיברי I_j) והמקרה $V_f(I_j)\ge\frac{\epsilon}{2er}$, זאת בהסתברות $1-\frac{1}{12r}$ לכל j.
- אם לפחות $k+1$ מהקבוצות I_j סומנו כבעלות תלות, דוחים את f. אחרת מקבלים אותה.

זהו אלגוריתם לא־אדפטיבי. כפי שהוזכר למעלה, אם f היא k־חונטה, ז"א תלויה בלא יותר מ־k קורדינטות, אז רק הקבוצות המכילות אותן יוכלו להיות מסומנות כבעלות תלות, ולכן הקלט יתקבל בהסתברות 1. להמשך הניתוח, נסמן ב־$J=\{1\le i\le n: V_f(\{i\})>\frac{\epsilon}{2er}\}$ את קבוצת הקורדינטות עם השתנות גדולה מ־$\frac{\epsilon}{2er}$.

אם מתקיים $|J|>k$, אז בהסתברות לפחות $\frac{3}{4}$ יהיו לפחות $k+1$ קבוצות I_j שמכילות קורדינטה מ־J. אפשר למשל להוכיח את זה בצורה הבאה: הסיכוי עבור המאורע ש־i ו־i' הם באותו I_j הוא $1/r$. אם לוקחים $J'\subseteq J$ מגודל $k+1$ בדיוק, ועושים חסם איחוד מאורעות עבור כל הזוגות $i,i'\in J'$, מקבלים הסתברות קטנה מ־$\frac{1}{4}$ שאיברי J' לא יהיו ב־$k+1$ קבוצות שונות.

בגלל המונוטוניות של ההשתנות, לכל הקבוצות המכילות איברים מ־J יש השתנות גדולה מ־$\frac{\epsilon}{2er}$. מכיוון שבשלב השני של האלגוריתם בהסתברות לפחות $\frac{11}{12}$ כל הקבוצות עם השתנות כזו יסומנו, זה אומר שסה"כ אם מתקיים $|J|>k$ אז הקלט ידחה בהסתברות לפחות $\frac{2}{3}$.

עתה ננתח את המקרים שבהם $|J|\le k$. אם $U_{f,J}(\{1,\ldots,n\})=V_f(\{1,\ldots,n\}\setminus J)\le 2\epsilon$, אז (כפי שראינו בתת־הפרק הקודם) הפונקציה היא ϵ־קרובה לפונקציה שתלויה באיברי J בלבד, ז"א k־חונטה, ואז זה לגיטימי לחלוטין לקבל אותה. המקרה האחרון שצריך לנתח הוא זה שבו $|J|\le k$ וגם $U_{f,J}(\{1,\ldots,n\})>2\epsilon$.

עבור ניתוח זה, ראשית ננתח את ההתפלגות של $U_{f,J}(I)$ כאשר בוחרים את I ע"י כך שכל קורדינטה $1 \leq i \leq n$ תיכלל ב־I בהסתברות $\frac{1}{r}$, באופן ב"ת לכל קורדינטה. במקרה זה $U_{f,J}(I) = \sum_{i \in I} U_{f,J}(i)$ הוא סכום של משתנים מקריים ב"ת (אחד לכל $1 \leq i \leq n$, שמקבל $U_{f,J}(I)$ אם $i \in I$ ומקבל 0 אם $i \notin I$), כאשר התוחלת של הסכום היא $\frac{1}{r}U_{f,J}(\{1,\ldots,n\}) > \frac{2\epsilon}{r}$, וכל משתנה לכשעצמו הוא אי־שלילי וערכו חסום ע"י $\frac{\epsilon}{2er}$. יש חסמים של סטיות גדולות למצבים כמו זה (לא ניכנס להוכחה שלהם), ונובע מאלו שמתקיים $\Pr[U_{f,J}(I) < \frac{\epsilon}{er}] < e^{-2} < \frac{1}{6}$.

נחזור לחלוקה המקרית שלנו, $I_1,\ldots,I_r$. כאן, כל I_j מתפלג בדיוק כמו הקבוצה המקרית I מהדיון למעלה, ולכן תוחלת מספר האינדכסים j עם $U_{f,J}(I_j) \geq \frac{\epsilon}{er}$ היא גדולה מ־$\frac{5}{6}r$. מצד שני, מספר האינדכסים j עם התכונה הזו כמובן יהיה חסום ע"י r (בהסתברות 1). זה אומר שבהסתברות לפחות $\frac{3}{4}$ מספר האינדכסים הנ"ל עולה על $\frac{1}{3}r > k$. על כן, גם במקרה שבו $|J| \leq k$ ו־$U_{f,J}(\{1,\ldots,n\}) > 2\epsilon$, האלגוריתם ידחה בהסתברות לפחות $\frac{2}{3}$.

בדיקה באמצעות למידה של תכונות של פונקציות

הרעיון של בדיקה באמצעות למידה כטכניקה כללית נוסח לראשונה דרך המאפיין של קרבה לחונטה, במאמר Diakonikolas, Lee, Matulef, Onak, Rubinfeld, Servedio, Wan: Testing for concise representations. אנחנו נראה איך אפשר ללמוד, עד כדי פרמוטציה של המשתנים, פונקציה עם מאפיין זה (גרסה יותר מוגבלת של הטענה הוכחה לראשונה במאמר החונטות המקורי, בהקשר של בדיקת איזומורפיזם לפונקציה נתונה).

באמצעות הלמידה אפשר לבדוק מספר תכונות שהפונקציות המקיימות אותן הן קרובות לחונטות. דוגמה אחת היא התכונה של להיות תוצאה של DNF עם מספר נוסחאות חסום. למשל, DNF עם נוסחה אחת הוא בעצם פונקציה מהצורה $\bigwedge_{i \in I} x_i$ עבור $I \subseteq \{1,\ldots,n\}$ כל שהוא. למרות שהתלות היא לא במספר משתנים חסום, פונקציית $\wedge$ של k משתנים היא 2^{-k}־קרובה להיות פונקציית ה־0, כך שבכל מקרה, לכל r, פונקציית $\wedge$ תהיה 2^{-r}־קרובה להיות חונטה של r משתנים. במאמר המקורי היו דוגמאות נוספות, חלקן דרך הכללות של בדיקת חונטות עבור קבוצת טווח יותר גדולות מ־$\{-1,1\}$. דוגמה מרכזית במאמר היא של בדיקת פולינומיות (עם ריבוי משתנים) מדרגה חסומה.

לפני שנמשיך נצטרך מספר הגדרות. עבור וקטור $x \in \{0,1\}^n$ ופרמוטציה $\sigma : \{1,\ldots,n\} \to \{1,\ldots,n\}$, נסמן ב־$y = x_\sigma$ את הווקטור עבורו $y_i = x_{\sigma(i)}$ לכל $1 \leq i \leq n$. פונקציה $f : \{1,\ldots,n\} \to \{-1,1\}$ תיקרא איזומורפית לפונקציה $g : \{1,\ldots,n\} \to \{-1,1\}$ אם קיימת פרמוטציה σ כך שמתקיים $g(x) = f(x_\sigma)$ לכל $x \in \{0,1\}^n$.

ישנם שני הבדלים מהותיים בין הלמידה כאן לבין הלמידה שהגדרנו עבור התפלגויות:

- האלגוריתם יהיה חייב לאפשר טולרנטיות מסויימת – גם אם הקלט f הוא רק η־קרוב להיות k־חונטה, עבור η מתאים שיהיה פולינומי ב־ϵ ו־$1/k$, על האלגוריתם ללמוד בהצלחה את f. זה יאפשר בדיקת תכונות כמו זו של ה־DNF למעלה, שבה אין שוויון ממש לחונטה. לשם כך נוסיף את הפרמטר η להגדרת מאפיין הלמידה.
- כאשר האלגוריתם לא פולט "$\perp$", ההבטחה על הפונקציה הנפלטת (בהסתברות גבוהה) היא לא שהיא קרובה ל־f, אלא רק שהיא קרובה לפונקציה איזומורפית ל־f. זה אומר שיהיה אפשר להשתמש בלמידה רק עבור בדיקה של תכונות אינווריאנטיות תחת איזומורפיזם (אם h ו־g איזומורפיות, אז או שתיהן מקיימות את התכונה או שתיהן לא מקיימות אותה).

אלגוריתם הלמידה יהיה עם שגיאה דו־צדדית. למעשה בלתי אפשרי לתת אלגוריתמים עם שגיאה חד־צדדית ומספר שאילתות קבוע עבור בדיקת חלק מהתכונות שניתן לפתור ע"י אלגוריתם הלמידה. כדוגמה אפשר לנתח את האפשרות להבדיל בין פונקציות מהצורה $f(x) = x_i \wedge x_j$ לבין פונקציות מהצורה $g(x) = x_k$: כל עוד אנחנו עושים פחות מ־$\log(n)$ שאילתות, עבור פונקציה מהצורה $x_i \wedge x_j$ כאשר בוחרים את $1 \leq i < j \leq n$ באופן יוניפורמי, בסיכוי גדול מ־0 (אפילו אם הוא קטן) המצב יהיה שכל השאילתות יהיו על $x \in \{0,1\}^n$

המקיימים $x_i = x_j$. על כן אי אפשר לנסח אלגוריתם שמקבל בהסתברות 1 פונקציה מהצורה $f(x) = x_i \wedge x_j$, אלא אם כן הוא מקבל גם פונקציה מהצורה $g(x) = x_k$.

מספר השאילתות יהיה אקספוננציאלי ב־k, וזה סביר – אנחנו בסופו של דבר רוצים ללמוד את כל ערכי החונטה $h : \{0,1\}^k \to \{-1,1\}$ אשר קובעת את הפונקציה הקרובה ל־f.

לקראת אלגוריתם הלמידה – דגימה בודדת מחונטה

מעתה ועד סוף הפרק נניח ש־ϵ קטן מקבוע מתאים, מספיק להניח למשל $\epsilon < \frac{1}{5}$.

נחזור לצורה שבה אלגוריתם ϵ־הבדיקה עבור k־חונטות פועל: הקורדינטות $\{1,\ldots,n\}$ מחולקות (באמצעות הגרלה יוניפורמית) לקבוצות $I_1,\ldots,I_r$ כאשר $r = 16k^2$. הראינו שאם הקלט f הוא ϵ־רחוק מלהיות k־חונטה, אז בהסתברות לפחות $\frac{3}{4}$ יש יותר מ־k קבוצות I_j שמקיימות $V_f(I_j) > \frac{\epsilon}{2er}$.

עתה נניח שאנחנו מקטינים את הסתברות השגיאה לבדיקת אי־התלות של כל I_j ל־$\frac{1}{24r}$ (במקום $\frac{1}{12r}$), כך שהאלגוריתם יגלה קלטים רחוקים מחונטות בהסתברות לפחות $\frac{17}{24}$. נסמן ב־$q(k,\epsilon)$ את מספר השאילתות של האלגוריתם כתלות בפרמטר החונטה k ובמרחק המותר ϵ. כזכור מספר זה פולינומי ב־k וב־$1/\epsilon$.

אלגוריתם הלמידה יתחיל מהרצה בודדת של של אלגוריתם הבדיקה לחונטה (עם הגדלת ההסתברות שציינו), ולאחריו מספר הרצות של אלגוריתם הדגימה שנתאר מייד. עבור אלגוריתם הדגימה נזכור את $I_1,\ldots,I_r$ מאלגוריתם הבדיקה, וכן את הקבוצה K של כל ה־$1 \le j \le r$ שעבורם גילינו תלות ב־I_j (בפרט $|K| \le k$, כי אחרת היינו דוחים את הקלט). אלגוריתם הדגימה אמור לפלוט איבר $y \in \{0,1\}^{|K|}$ וערך $w \in \{-1,1\}$, עם הטענה שהפונקציה $h : \{0,1\}^{|K|} \to \{-1,1\}$ שמגדירה את החונטה הקרובה ל־f תקיים $h(y) = w$. האלגוריתם יפעל בצורה הבאה:

- נגריל את $x = (x_1,\ldots,x_n) \in \{0,1\}^n$ באופן יוניפורמי. אנחנו נגדיר $w = f(x)$, ועכשיו נותר לחשב את y. נסמן את האינדכסים שלו לפי K, ז"א $y = (y_j)_{j\in K} \in \{0,1\}^{|K|}$.
- לכל $j \in K$ נבצע את הדברים הבאים.
  - נגדיר את $I_{j,0} = \{i \in I_j : x_i = 0\}$ ואת $I_{j,1} = \{i \in I_j : x_i = 1\}$, ז"א את קבוצת ה־0 וקבוצה ה־1 של x בתוך I_j.
  - נבצע בדיקת אי־תלות של $I_{j,0}$, שמסמנת אותו בהסתברות לפחות $1 - \frac{1}{2k}\epsilon$ אם $V_f(I_{j,0}) > \frac{2\epsilon}{k}$. נבצע בדיקת אי־תלות כזו גם עבור $I_{j,1}$.
  - אם רק $I_{j,0}$ סומן, נגדיר $y_j = 0$. אם רק $I_{j,1}$ סומן, נגדיר $y_j = 1$. בכל מקרה אחר נגריל את y_j באופן יוניפורמי מתוך $\{0,1\}$.
- נפלוט את y ואת w שהגדרנו.

נסמן ב־$q'(k,\epsilon)$ את מספר השאילתות הכולל של הרצה בודדת של אלגוריתם הדגימה. עתה ננתח מה קורה כאשר הפונקציה f אכן שווה (לא רק קרובה) לחונטה המוגדרת לפי J, ז"א $f(x) = h(x|_J)$ עבור $h : \{0,1\}^k \to \{-1,1\}$ מתאימה. אנחנו נניח ששלב בדיקת החונטה "הצליח", שזה אומר שבפרט לכל j מתקיים $|I_j \cap J| \le 1$ (אין שתי קורדינטות חונטה באותה I_j), וכן לכל $i \in I_j \cap J$ שעבורו $V_f(\{i\}) > \frac{2\epsilon}{k} > \frac{\epsilon}{2er}$ מתקיים $j \in K$, ז"א שהצלחנו "לגלות" את כל קורדינטות החונטה עם תלות גדולה מ־$\frac{2\epsilon}{k}$.

במקרה זה, לפי תת־חיבוריות, מתקיים $V_f(\bigcup_{j\notin K} I_j) \le 2\epsilon$. נגדיר את "החונטה המצומצמת" $h'(y)$ לפי הערך שמופיע יותר פעמים עבור וקטורים שהצמצום שלהם לקורדינטות החונטה שגילינו שווה ל־y: אם $|\{z \in \{0,1\}^{k-|K|} : h(z \sqcup y) = 1\}| > |\{z \in \{0,1\}^{k-|K|} : h(z \sqcup y) = -1\}|$ נגדיר $h'(y) = 1$, ואחרת נגדיר $h'(y) = -1$. נשים לב עתה שהמרחק של f מהפונקציה f' המוגדרת ע"י החונטה $h' : \{0,1\}^{|K|} \to \{-1,1\}$ (במקום $h : \{0,1\}^k \to \{-1,1\}$) חסום ע"י ϵ, מחצית החסם על ההשתנות של קבוצת קורדינטות החונטה שהתעלמנו מהן.

עוד דבר לשים לב הוא שבלי קשר להצלחה של אלגוריתם הדגימה, הערך $y \in \{0,1\}^{|K|}$ יהיה בעל התפלגות יוניפורמית. הסיבה היא שלכל $j \in K$, אם הצלחנו בשלב הבדיקה של $I_{j,0}$ מול $I_{j,1}$ לגלות איזו משתי הקבוצות מכילה את קורדינטת החונטה המתאימה (לקבוצה הזו תהיה השתנות $V_f(I_j)$ ולקבוצה השניה תהיה השתנות 0 כי היא לא מכילה קורדינטת חונטה), אז y_j יקבל את הערך הנכון מתוך x, שכזכור הוגרל יוניפורמית. אם לא הצלחנו בשלב הבדיקה של $I_{j,0}$ מול $I_{j,1}$ אז ממילא הצבנו ב־y_j ערך שנבחר יוניפורמית.

אנחנו נגדיר אי־הצלחה של אלגוריתם הדגימה לפי אי־גילוי של השתנות של $I_{j,0}$ או $I_{j,1}$ שיש לו השתנות גדולה מ־$\frac{2\epsilon}{k}$. הסיכוי לאי־הצלחה מסוג זה (שיכול לגרום לערך w שאין לו שום קשר ל־$h'(y)$) חסום ע"י ϵ.

לבסוף, במקרה של הצלחה, נחסום את הסיכוי של w להיות שווה ל־$h'(y)$. במקרה זה עדיין יכול להיות שקבענו y_j שגוי ל־I_j שמכיל קורדינטת חונטה, אם ההשתנות המתאימה אינה עולה על $\frac{2\epsilon}{k}$. במקרה של אי־גילוי, הערך של y_j הוגרל יוניפורמית באופן ב"ת מהערך שקורדינטת החונטה קיבלה דרך x. בגלל החסם הכולל של 2ϵ על ההשתנות מעל כל קורדינטות החונטה עם השתנות שאינה עולה על $\frac{2\epsilon}{k}$, הסיכוי לערך שגוי כאן גם חסום ע"י ϵ.

סה"כ, יוצא שהסיכוי הכולל לדגימה שגויה של $h'(y)$ (ז"א מקרה שבו $h'(y) \neq w$) חסום ע"י 3ϵ, שזה הסכום על המקרים שעלינו על x שעבורו $f(x) \neq f'(x)$, או שאירע מאורע אי־הצלחה של האלגוריתם עצמו, או שהאלגוריתם טעה בגלל y_j שמתאימים לקורדינטות עם השתנות נמוכה. מאורע הטעות לא בהכרח ב"ת במשתנה המקרי y, יכול להיות שיהיו הסתברויות שונות בהתניה על y ספציפים שונים.

למידת פונקציה קרובה לחונטה

נסמן ב־$q'(k,\epsilon)$ את מספר השאילתות הכולל בהרצה אחת של אלגוריתם הדגימה (השגת זוג z, w בודד, לא כולל בדיקת החונטה בהתחלה). נראה עתה מה קורה אם הפונקציה f אינה זהה לחונטה המוגדרת ע"י h, אלא רק $\eta'(k,\epsilon)$־קרובה לחונטה g אשר היא מוגדרת ע"י h, כאשר $\eta' = \epsilon/q'(k,\epsilon)$. נשים לב שכל שאילתה בודדת באלגוריתם הדגימה מתפלגת באופן יוניפורמי מעל $\{0,1\}^n$, כאשר מתייחסים להתפלגות הלא־מותנה על השאילתות האחרות שבוצעו (התפלגויות השאילתות השונות אינן ב"ת זו בזו). זה אומר שההסתברות, כשמבצעים את השאילתה $f(z)$ עבור ה־z המתאים, לקבל ערך שונה מ־$g(z)$, חסומה ע"י $\eta'(k,\epsilon)$. לפי חסם על איחוד מאורעות, ההסתברות שבמהלך כל ההרצה של אלגוריתם הדגימה נקבל ערכים ששונים מאלו של $g(z)$ חסומה ע"י $\epsilon = q'(k,\epsilon)\eta'(k,\epsilon)$.

על כן, ההסתברות שהרצה של אלגוריתם הדגימה על f תתן זוג z, w שעבורו $h'(z) \neq w$ תהיה חסומה ע"י 4ϵ. נראה עתה את הפרוצדורה הבאה: מבצעים $O(k2^k)$ הרצות של אלגוריתם הדגימה, על מנת שבהסתברות לפחות $1 - \frac{1}{72}$ נקבל לפחות זוג אחד (z,w) לכל $z \in \{0,1\}^{|K|}$ אפשרי. נגדיר את $h'(z) = w$ לפי ה־w שקיבלנו עבור הזוג הראשון עם ה־z המתאים, ונחזיר לבסוף את h' (אם לא הצלחנו לכסות את כל $\{0,1\}^{|K|}$ אז פשוט נחזיר "$\perp$"). בגלל שתוחלת המרחק של g מהפונקציה המוגדרת ע"י h' שנחזיר חסומה ע"י 4ϵ, בהסתברות לפחות $1 - \frac{2}{72}$ אנחנו גם נצליח להחזיר פונקציה h' וגם המרחק של g מהחונטה המתאימה יהיה חסום ע"י 288ϵ (השתמשנו באי־שוויון מרקוב כאן). המרחק של f מהפונקציה שהחזרנו, לפי אי שוויון המשולש, יהיה בפרט קטן מ־289ϵ (מכיוון ש־$\eta' < \epsilon$).

אלגוריתם הלמידה המלא (מייד נוודא אלו פונקציות הוא לומד) ישתמש בשלבים הבאים:

- נבצע את אלגוריתם הבדיקה עבור k־חונטות (ובפרט, אם לא דוחים, אז נגדיר את $I_1, \ldots, I_r$ ואת K), כאשר נשתמש בהסתברות הצלחה $\frac{17}{24}$, אבל נדרוש גם פרמטר מרחק של $\eta'(k,\epsilon)$ במקום ϵ. נשים לב ש־r תלוי רק ב־k ולא בפרמטר המרחק. אם האלגוריתם דחה אז נעצור כאן ונחזיר "$\perp$", ואחרת נמשיך לשלב הבא.
- נבצע את ההרצות של אלגוריתם הדגימה על מנת לפלוט (בהסתברות לפחות $1 - \frac{2}{72}$) פונקציה h' עם הבטחת קירבה מתאימה.

נסמן לבסוף $\eta(k,\epsilon) = 1/72q(k,\eta'(k,\epsilon))$, כאשר q הוא מספר השאילתות של האלגוריתם לבדיקת k־חונטה עם הפרמטרים המתאימים (אם תחשבו את זה, זה עדיין פולינומי ב־ϵ ו־$1/k$). נראה עתה שהאלגוריתם כאן יבצע 289ϵ־למידה עבור פונקציות שמאופיינות ע"י $\eta(k,\epsilon)$־קירבה להיות k־חונטות.

ראשית נראה את התכונה שעבור פונקציה כל שהיא, ההסתברות להחזיר פונקציה בעלת מרחק יותר גדול מ־289ϵ חסום ע"י $\frac{1}{3}$: אם f היא $\eta'(k,\epsilon)$־רחוקה מחונטה, אז ממילא ההסתברות לדחות אותה ולהחזיר "$\perp$" גדולה מ־$\frac{2}{3}$. אחרת, בהסתברות לפחות $\frac{17}{24}$ בשלב בדיקת החונטה הצלחנו "להפריד" בין הקורדינטות של h, וגם להכליל ב־K את כל הקבוצות עם השתנות גדולה מספיק. כשזה קורה, ההסתברות של שלב הדגימה להחזיר פונקציה h' שגויה (עם מרחק גדול מדי) חסום ע"י $\frac{2}{72}$, וסה"כ יש הסתברות של לפחות $\frac{2}{3}$ להחזיר פונקציה נכונה.

עתה נניח ש־f היא $\eta(k,\epsilon)$־קרובה להיות k־חונטה. ראשית נשים לב שגם באלגוריתם בדיקת החונטה, ההתפלגות הלא־מותנה של כל שאילתה היא יוניפורמית מעל $\{0,1\}^n$. על כן, בהסתברות לפחות $\frac{71}{72}$ כל תוצאות השאילתות מ־f יהיו זהות לפונקצית החונטה הקרובה אליה g, ובהסתברות לפחות $\frac{50}{72} = \frac{17}{24} - \frac{1}{72}$ אלגוריתם הבדיקה יקבל את f וגם יחזיר חלוקה $I_1,\ldots,I_r$ ו־K שיקיימו את הנדרש לאלגוריתם הדגימה. בשלב השני נקבל פונקציה עם הקירבה המתאימה בהסתברות לפחות $1-\frac{2}{72}$, ולכן סה"כ תהיה הסתברות של לפחות $\frac{2}{3}$ לקבל תשובה h' עם הבטחת הקירבה המתאימה.

לפני שנסיים, נראה איך מבצעים ϵ־בדיקה עבור התכונה ש־f היא מהצורה $\bigwedge_{i\in I} x_i$ עבור I כל שהוא (לא בהכרח מגודל חסום). כזכור, לכל k, פונקציה כזו תהיה 2^{-k}־קרובה להיות k־חונטה. עבור ϵ־בדיקה אנחנו צריכים למידה עם פרמטר $\frac{\epsilon}{2}$, ולכן נצטרך לבצע את האלגוריתם שתואר למעלה עם פרמטר $\frac{\epsilon}{578}$. זה אומר שאנחנו רוצים לבחור k שמקיים $2^{-k} < \eta(k,\frac{\epsilon}{578})$, ומכיוון ש־$\eta$ הוא פולינומי ב־$1/k$ ו־ϵ, זה יאפשר לנו בחירה של $k = O(\log(1/\epsilon))$. הדבר ייתן לנו אלגוריתם ϵ־בדיקה עבור התכונה (עם שגיאה דו־צדדית) בעל מספר שאילתות פולינומי ב־ϵ (השלב עם הכי הרבה שאילתות הוא זה של $O(k2^k)$ הרצות של אלגוריתם הדגימה הבודדת).

פרטים על שגיאה חד־צדדית

נראה עתה יותר פרטים על הטענה שצריך $\Omega(\log(n))$ שאילתות עבור בדיקה חד־צדדית של התכונות מהסוג שהשתמשנו כאן בלמידה עבורן, וספציפית נתייחס לתכונה "f היא מהצורה $x_i \wedge x_j$ עבור $1 \leq i < j \leq n$ כל שהם". בעצם זו תכונה של איזומורפיזם לחונטה ספציפית, תכונה שהוזכרה במאמר החונטות המקורי.

נסתכל על אלגוריתם בדיקה הסתברותי כעל מרחב הסתברות מעל אלגוריתמים דטרמיניסטים. אנחנו נטען שבכל אלגוריתם דטרמיניסטי בעל q שאילתות, המתואר ע"י עץ החלטות מגובה q המופיע בהסתברות גדולה מ־0, לא יכול להיות עלה v שדוחה את הקלט אם סדרת השאילתות בדרך אליו $Q = \{x^{(1)},\ldots,x^{(q)}\}$ וסדרת התשובות המתאימה $w_1,\ldots,w_q$ יכולה להתאים לקלט אשר מקיים את התכונה. אם הייתה f שמקיימת את התכונה שעבורה $f|_Q = (w_1,\ldots,w_q)$, האלגוריתם ההסתברותי היה דוחה אותה בהסתברות גדולה מ־0.

ניזכר שאצלנו כל שאילתה $x^{(i)}$ היא בעצם איבר ב־$\{0,1\}^n$ שעליו שואלים את f, ועתה נראה מה קורה עבור עלה ספציפי v של עץ ההחלטה: לכל $1 \leq j \leq n$ נגדיר את הווקטור $z_j = (x_j^{(1)},\ldots,x_j^{(q)}) \in \{0,1\}^q$, הווקטור של ערכי הקורדינטה j בסדרת השאילתות. אם קיימים $i < j$ שעבורם $z_i = z_j$, אז עבור הפונקציה $f(x) = x_i \wedge x_j$, והפונקציה $g(x) = x_i$, סדרת התשובות תהיה זהה עבור f ו־g. זה אומר שאם העלה דוחה, פונקציה מהצורה $g(x) = x_i$ יכולה להגיע אליו (מבחינת סדרת הערכים $w_1,\ldots,w_q$) רק אם הערך z_i אינו שווה לאף z_j אחר.

נראה עתה מה קורה כאשר $q < \log(n)/3$. לכל עלה v שדוחה, יש לא יותר מ־$2^q < n^{1/3}$ ערכים אפשריים לווקטור z_j, ולכן זהו חסם על מספר הפונקציות מהצורה $g(x) = x_i$ שמגיעות אליו (כל ווקטור כזה יכול להתאים לכל היותר לפונקציה אחת). כמו כן, מספר העלים הכולל של עץ ההחלטה חסום ע"י $n^{1/3}$, ולכן מספר הפונקציות הכולל מהצורה $g(x) = x_i$ שיכול להגיע לעלים דוחים חסום ע"י $n^{2/3}$. על כן, אם נגריל באופן יוניפורמי את $1 \leq i \leq n$ ונקבע את הקלט להיות הפונקציה $g(x) = x_i$, אז האלגוריתם הדטרמיניסטי ידחה אותו בהסתברות חסומה ע"י $n^{2/3}/n = o(1)$. מכיוון שהדבר נכון לכל עצי ההחלטה הדטרמיניסטים שיכולים להיבחר בהסתברות חיובית ע"י האלגוריתם ההסתברותי, המסקנה היא שעבור $\epsilon < \frac{1}{4}$ אין ϵ־בדיקה חד־כיוונית לתכונה המדוברת בפחות מ־$\Omega(\log(n))$ שאילתות, גם עבור אלגוריתם אדפטיבי.

הערה – היה אפשר לקשר את הטיעון הזה לשיטת יאו. כעיקרון שיטת יאו עובדת גם כאשר במקום "הצלחה או כישלון" נותנים ערכים של "עלויות" (חיוביות או שליליות). כאן העלות של קבלת קלט ϵ־רחוק מהתכונה היא 1, אבל העלות של דחיית קלט מקיים היא "$+\infty$".

לסיכום, נתאר בקצרה איך אפשר לכתוב אלגוריתם ϵ־בדיקה אדפטיבי בעל $\tilde{O}((\log(n))^2)$ שאילתות עבור ϵ ו־k קבועים (בדוגמה למעלה $k=2$). הרעיון מזכיר חיפוש בינארי: בכל שלב אנחנו נתחזק קבוצות זרות $I_1,\ldots,I_l\subseteq\{1,\ldots,n\}$ (לאו דווקא חלוקה, יכולים להיות אינדכסים שלא נמצאים באף קבוצה). בתחילת האלגוריתם $l=1$ ו־$I_1=\{1,\ldots,n\}$. בכל פעם ניקח קבוצה I_j שהגודל שלה גדול מ־1, נחלק אותה לשתי קבוצות בעלות גדלים $\lfloor\frac{1}{2}I_j\rfloor$ ו־$\lceil\frac{1}{2}I_j\rceil$, ונבדוק כל אחת מהן עבור אי־תלות. נסיר את I_j ונוסיף במקומה את הקבוצה שעבורה גילינו תלות אם יש כזו, ואת שתי הקבוצות אם גילינו תלות בשתיהן.

אם באיזה שהוא שלב נקבל $l>k$ אז נדחה את הקלט, ואחרת לאחר $O(\log(n))$ איטרציות נקבל קבוצות מגודל 1, שיתארו את קורדינטות החונטה שעתה נוכל ללמוד. הצורך לגלות גם קבוצות עם השתנות גבוהה מ־$O(1/\log(n))$ בלבד (ע"מ שלאיחוד כל הקבוצות שלא שמרנו לא תהיה השתנות גבוהה מדי) יוסיף עוד פקטור של $O(\log(n))$. הסיבה לתוספת נוספת של פקטור $O(\log\log(n))$ במספר השאילתות היא הצורך להבטיח שבהסתברות גבוהה לא נפספס קבוצה בעלת תלות באף אחד מהשלבים.

בדיקת התפלגויות עם דגימות התפלגות מותנה

במודל של בדיקת התפלגויות, האלגוריתמים אינם מבצעים שום החלטה על הדגימות – הם רק מקבלים סדרה של דגימות ומחליטים לפיהם אם לקבל את הקלט. בנוסף, המידע המתקבל בדגימות מתוך התפלגות הקלט μ הוא מועט למדי, ובפרט כל התכונות המשמעותיות דורשות מספר דגימות של לפחות $\sqrt{n}$, כאשר $n=|S|$ יסמן את גודל קבוצת הבסיס של מרחב ההסתברות.

מספר מודלים עם דגימות או שאילתות חזקות יותר הוצעו לעניין זה, והנחקר ביותר ביניהם הוא זה של דגימות מותנות, מודל שהוצע לראשונה במקביל במאמר Chakraborty, Fischer, Goldhirsh, Matsliah: On the power of conditional samples in distribution Testing ובמאמר Canonne, Ron, Servedio: Testing probability distributions using conditional samples.

במודל זה עדיין מקבלים דגימות הסתברותיות, אבל האלגוריתם יכול לבקש דגימה על תת־מרחב של ההתפלגות: עבור הדגימה ה־i, האלגוריתם מוסר כשאילתה תת־קבוצה $\emptyset\neq S_i\subseteq S$, ומקבל דגימה A_i שנבחרה (באופן ב"ת בדגימות קודמות) לפי מרחב ההתפלגות המותנה $\mu|_{S_i}$. האלגוריתם יכול להיות אדפטיבי – הבחירה של S_i יכולה להסתמך על תשובות לדגימות קודמות (הדרישה ל"אי־תלות" מקודם מתייחסת להבטחה ש־A_i יתפלג לפי $\mu|_{S_i}$ בלי קשר לאיך האלגוריתם חישב את S_i).

בתיאור למעלה יש "חסר" קטן, וזו השאלה מה התשובה שהאלגוריתם מקבל במידה ומבוצעת השאילתה "דגימה מתוך S_i" כאשר $\mu(S_i)=0$. במאמר הראשון למעלה האלגוריתם יקבל דגימה שנבחרה יוניפורמית מ־S_i, ובמאמר השני האלגוריתם יקבל כתשובה את הסימן המיוחד "$\perp$" (ובכך הוא יכול לקבל כמות גדולה יחסית של מידע על μ). אנחנו נשתמש במודל של המאמר הראשון.

כדאי גם לדון בריאליזם של המודל: במקרה הכללי ביותר, תאור השאילתה (פירוט תת־הקבוצה S_i) לוקח n ביטים. כמו כן, "בעולם האמיתי" הרבה יותר קל לתת גישה להתפלגות הקלט מאשר גישה להתפלגויות שיכולות להיות מותנות על קבוצות בעלות הסתברות נמוכה (ואלו בדיוק השאילתות שמוסיפות חוזק למודל – עבור קבוצה S_i עם הסתברות גבוהה, אפשר פשוט לקחת $O(1/\mu(S_i))$ דגימות מההתפלגות הלא־מותנה עד שמקבלים אחת שנוחתת ב־S_i). על כן נוהגים גם לחקור מודלים שבהם הדגימות המותנות יותר מוגבלות, למשל למשפחה קטנה יחסית של S_i אפשריים. כאן בעיקר נתמקד במודל הכללי.

כמו במודל המקורי של בדיקת התפלגויות, האלגוריתמים כאן יהיו בעלי שגיאה דו־צדדית.

בדיקת יוניפורמיות במספר קבוע של שאילתות דגימה מותנית

כאשר מרשים דגימות מותנות, ואדפטיביות של האלגוריתם, ניתן לבצע בדיקת יוניפורמיות במספר דגימות שתלוי ב־ϵ בלבד. החסם הכי טוב הוא זה של המאמר השני מאלו שנזכרו למעלה, של $\tilde{O}(1/\epsilon^2)$ דגימות. אנחנו נראה כאן גרסה פשוטה יותר של הבדיקה, שמבצעת יותר דגימות. אנחנו נדאג להבטיח שהבדיקה הזו תקבל בהסתברות גבוהה גם התפלגויות $\frac{\epsilon}{3}$־יוניפורמיות, דבר שמועיל בהמשך למשל ללמידת התפלגות יוניפורמית למקוטעין.

ננתח מה מתקבל עבור "אלגוריתם הזוג" הבא.

- **לוקחים דגימה $u \in S$ לפי μ (דגימה לא מותנה).**
- **בוחרים באופן יוניפורמי (בלי קשר ל־μ) איבר $v \in S$.**
- **באמצעות $O(\log(1/\delta)/\epsilon^2)$ דגימות מותנות על $\mu|_{\{u,v\}}$, מבדילים בין המקרה שבו $\mu(u) \geq \frac{1}{n}(1+\frac{\epsilon}{2})$ ו־$\mu(v) \leq \frac{1}{n}(1-\frac{\epsilon}{2})$, לבין המקרה שבו $\mu(u) \leq (1+\frac{\epsilon}{3})\mu(v)$. זה אפשרי כי במקרה הראשון מתקיים $\mu|_{\{u,v\}}(u) \geq \frac{1}{2}+\frac{\epsilon}{5}$, ובמקרה השני מתקיים $\mu|_{\{u,v\}}(u) \leq \frac{1}{2}+\frac{\epsilon}{6}$.**

ראשית נשים לב שאם μ היא $\frac{\epsilon}{3}$־יוניפורמית, הזוג שנבחר הוא תמיד יהיה כזה שיקיים $\mu(u) < (1+\frac{\epsilon}{3})\mu(v)$ ואנחנו נקבל את הזוג בהסתברות לפחות $1-\delta$.

עתה נניח ש־μ היא ϵ־רחוקה מיוניפורמיות. נגדיר את הקבוצה $H = \{u \in S : \mu(u) > \frac{1}{n}\}$ ואת הקבוצה $L = \{v \in S : \mu(v) < \frac{1}{n}\}$. ניזכר שמתקיים $d(\mu,\pi) = \sum_{u\in H}(\mu(u)-\frac{1}{n}) = \sum_{v\in L}(\frac{1}{n}-\mu(v))$. כמו כן נגדיר את $H' = \{u \in S : \mu(u) \geq \frac{1}{n}(1+\frac{\epsilon}{2})\}$ ואת הקבוצה $L' = \{v \in S : \mu(v) \leq \frac{1}{n}(1-\frac{\epsilon}{2})\}$.

מאי השוויון $\sum_{u\in H}(\mu(u)-\frac{1}{n}) > \epsilon$ נובע שבפרט מתקיים $\mu(H') > \frac{\epsilon}{2}$, ולכן בהסתברות לפחות $\frac{\epsilon}{2}$ נקבל $u \in H'$. מאי השוויון $\sum_{v\in L}(\frac{1}{n}-\mu(v)) > \epsilon$ נובע שמתקיים $|L'| > \frac{\epsilon}{2}n$, ולכן בהסתברות לפחות $\frac{\epsilon}{2}$ נקבל $v \in L'$. לכן בהסתברות לפחות $\epsilon^2/4$ נקבל זוג שאנחנו הולכים לדחות בהסתברות לפחות $1-\delta$.

על מנת לבנות אלגוריתם שבהסתברות לפחות $1-\delta$ מבחין בין קלט $\frac{\epsilon}{3}$־יוניפורמי לבין קלט ϵ־רחוק מיוניפורמיות, נבחר באופן יוניפורמי וב"ת $r = 4\ln(2/\delta)/\epsilon^2$ זוגות באופן שצויין, ונבדוק כל אחד מהם עם הסתברות שגיאה $\delta' = \delta/2r$. אנחנו נדחה את הקלט אם דחינו לפחות את אחד מהזוגות, ואחרת נקבל. סה"כ נבצע $\tilde{O}(\log(1/\delta)/\epsilon^4)$ דגימות.

לפני שנמשיך, נעיר קצת על בדיקה לשוויון עם התפלגות ידועה מראש: אם מבצעים חלוקה לדליים כמו שעשינו במודל של דגימות בלבד (צריך פרמטר קצת יותר קטן לדליים, כי כאן ההבטחה היא לקבל התפלגויות $\frac{\epsilon}{3}$־יוניפורמיות, לא התפלגויות ϵ־יוניפורמיות), התוצאה תהיה אלגוריתם פולינומי ב־$1/\epsilon$ וב־$\log(n)$. יש תלות ב־n כי עדיין יהיה צריך למשל לבדוק את $\mu_{\mathcal{B}}$ מול $\tau_{\mathcal{B}}$ (כאשר $\mathcal{B}$ היא החלוקה לדליים ומקיימת $|\mathcal{B}| = O(\log(n)/\epsilon)$). אפשר להמשיך "למתוח" את זה עם עוד איטרציות של האלגוריתם היעיל גם לחלוקה לדליים (במקום למידת $\mu_{\mathcal{B}}$ כפי שנעשה במקור), ולהגיע לבסוף לתלות פולינומית ב־$\log^*(n)$. במאמר השני, עם ניתוח יותר זהיר, יש בדיקה פולינומית ב־$1/\epsilon$ ולא תלויה ב־n לתכונה זו. לעומת זאת, תלות מסויימת ב־n היא הכרחית אם רוצים להשוות בין שתי התפלגויות לא־ידועות שדוגמים מהן.

למידה של התפלגויות יוניפורמיות למקוטעין

כאן אנחנו נשתמש בהבטחה שאלגוריתם ϵ־בדיקה מקבל בהסתברות גבוהה גם התפלגויות $\frac{\epsilon}{3}$־יוניפורמיות. באלגוריתם הלמידה במודל הדגימות המקורי, מרבית הדגימות נלקחו בשלב שבו מוודאים שמעל רב הקטעים (לפי משקל) בחלוקה העדינה ההתפלגות היא קרובה ליוניפורמית. כאן אנחנו יכולים לעקוף את זה: אנחנו פשוט נבצע את אלגוריתם בדיקת היוניפורמיות עם התפלגויות מותנות עבור כל קטע בחלוקה העדינה שלנו. ננתח את האלגוריתם הבא.

- מוצאים (באמצעות $\tilde{O}(k/\epsilon)$ דגימות לא מותנות) חלוקה $\frac{\epsilon}{k}$־עדינה, בהסתברות הצלחה לפחות $\frac{8}{9}$. נסמן את החלוקה ב־$\mathcal{B} = \{K_1, \ldots, K_r\}$, כאשר $K_i = \{t_{i-1}+1, \ldots, t_i\}$ לכל $1 \le i \le r$, עבור $0 = t_0 < \cdots < t_r = n$ מתאימים.

- מוצאים (באמצעות $\tilde{O}(k/\epsilon^3)$ דגימות לא מותנות) קירוב $\tilde{\mu}_{\mathcal{B}}$ שיהיה ϵ־קרוב (במרחק התפלגות) ל־$\mu_{\mathcal{B}}$, עם הסתברות הצלחה לפחות $\frac{8}{9}$.

- לכל $0 \le i \le r$, מבצעים בדיקת 3ϵ־יוניפורמיות עבור $\mu|_{K_i}$, בהסתברות הצלחה לפחות $1 - \frac{1}{9r}$. סה"כ נשתמש ב־$\tilde{O}(r/\epsilon^4) = \tilde{O}(k/\epsilon^5)$ דגימות מותנות (אפשר להוריד את חזקת ϵ ל־3 אם משתמשים באלגוריתם בדיקת היוניפורמיות הכי טוב שידוע, בזה שראינו כאן).

- נסמן ב־N את קבוצת ה־i שעבורם בדיקת היוניפורמיות דחתה את $\mu|_{K_i}$. אם מתקיים $\tilde{\mu}_{\mathcal{B}}(N) > 2\epsilon$ אז נפלוט "$\perp$", ואחרת נפלוט את הקירוב $\tilde{\mu}$, אשר מוגדר עבור $j \in K_i$ לפי $\tilde{\mu}(j) = \tilde{\mu}_{\mathcal{B}}(i)/|K_i|$.

נשים לב שבהסתברות לפחות $\frac{2}{3}$, כל השלבים מצליחים: החלוקה $\mathcal{B}$ תהיה $\frac{\epsilon}{k}$־עדינה, יתקיים $d(\mu_{\mathcal{B}}, \tilde{\mu}_{\mathcal{B}}) < \epsilon$, כל הקטעים K_i שעבורם $\mu|_{K_i}$ היא ϵ־יוניפורמית יתקבלו וכל הקטעים K_i שעבורם $\mu|_{K_i}$ היא 3ϵ־רחוקה מיוניפורמיות ידחו (i יוכנס ל־N). ננתח עתה מה קורה כאשר כל השלבים מצליחים.

אם הפלט שונה מ־"$\perp$", אז זה אומר שמתקיים $\mu_{\mathcal{B}}(N) \le \tilde{\mu}_{\mathcal{B}}(N) + \epsilon \le 3\epsilon$. כמו כן, לכל $i \notin N$, מתקיים $d(\mu|_{K_i}, \tilde{\mu}|_{K_i}) \le 3\epsilon$ (כזכור $\tilde{\mu}|_{K_i}$ הוגדרה להיות ההתפלגות היוניפורמית מעל K_i). על כן יתקיים $d(\mu, \tilde{\mu}) \le 6\epsilon$.

בנוסף, אם ההתפלגות μ היא (k,ϵ)־יוניפורמית למקוטעין, אז (מכיוון שהחלוקה היא $\frac{\epsilon}{k}$־עדינה) יתקיים בהכרח $\mu(N) \le \epsilon$, ולכן $\tilde{\mu}(N) \le 2\epsilon$, ז"א שהאלגוריתם לא יחזיר "$\perp$" (ויחזיר התפלגות $\tilde{\mu}$ שהיא 6ϵ־קרובה ל־μ).

מאלו נובע שבנינו אלגוריתם 6ϵ־למידה עבור התפלגויות (k,ϵ)־יוניפורמיות למקוטעין בעל $\tilde{O}(k/\epsilon^5)$ דגימות. ניתן לבנות אלגוריתם דומה לזה בעל $k \cdot \tilde{O}(1/\epsilon^3)$ דגימות בלבד (שימו לב שאין כאן פקטור של $\log(k)$). נסקור בקצרה איך אפשר לשפר את מספר הדגימות.

- מגדירים חלוקות (η,γ)־עדינות – ההבדל בין אלו לבין חלוקות η־עדינות הוא שכאן מאפשרים גם קטעים חריגים בעלי משקל גדול מ־η, כל עוד המשקל הכולל של אלו אינו עולה על γ. אפשר למצוא חלוקה $(\frac{\epsilon}{k},\epsilon)$־עדינה ב־$k \cdot \tilde{O}(1/\epsilon)$ דגימות בלבד, ויתקיים גם $r = k \cdot \tilde{O}(1/\epsilon)$.

- מציאת $\tilde{\mu}_{\mathcal{B}}$ תיקח (בחשבון מדוייק) $k \cdot \tilde{O}(1/\epsilon^3)$ דגימות.

- במקום לבצע בדיקה של $\mu|_{K_i}$ לכל $1 \le i \le r$ עם סיכוי הצלחה $1 - \frac{1}{9r}$, נסתפק בסיכוי הצלחה של $1 - \frac{\epsilon}{9}$ (ונשתמש באלגוריתם הידוע עם $\tilde{O}(1/\epsilon^2)$ דגימות – סה"כ $k \cdot \tilde{O}(1/\epsilon^3)$ דגימות לכל הקטעים). באמצעות שיקול של תוחלת ואי־שוויון מרקוב, בהסתברות לפחות $\frac{8}{9}$ המשקל הכולל של קטעים שטעינו לגביהם חסום ע"י ϵ.

- בקריטריון "$\tilde{\mu}(N) \le 2\epsilon$" נצטרך להחליף את המקדם "2" במקדם קבוע גדול יותר. זה גם ישפיע על המקדם "6" בפרמטר הלמידה המובטחת של האלגוריתם.

למידה עד כדי פרמוטציה של התפלגות כללית

בהינתן התפלגות μ ופרמוטציה $\sigma : S \to S$, נגדיר את ההתפלגות μ_σ לפי $\mu_\sigma(a) = \mu(\sigma(a))$ לכל $a \in S$. ישנן תכונות אינווריאנטות בפרמוטציה שהן קשות לבדיקה במודל הדגימות הלא־מותנות, עם חסם מהצורה $\Omega(n/(\log(n))^c)$ עבור קבוע c מתאים (יש גם חסם עליון תואם מהצורה $O(n/(\log(n))^d)$). למשל, התכונה שקבוצת האיברים עם הסתברות חיובית (התומך של μ) היא מגודל $\frac{1}{2}|S|$ או פחות היא תכונה כזו.

במודל של בדיקות עם דגימות מותנות ניתן, עד כדי פרמוטציה, ללמוד את ההתפלגות μ עם מספר דגימות פולינומי ב־$\log(n)$ (עבור ϵ קבוע), ובפרט ניתן לבדוק ביעילות יחסית את כל התכונות מהצורה הזו. ההוכחה

עוברת דרך ביצוע סימולציה של מודל דגימות אחר, חזק במיוחד: אנחנו נבנה פרוצדורה ש"דוגמת" מתוך התפלגות אחרת $\tilde{\mu}$, קרובה ל־μ, כאשר כל דגימה תגיעה עם ההתפלגות עצמה – במקום להחזיר רק "a" הדוגם יחזיר את הזוג "$(a, \tilde{\mu}(a))$". לפרוצדורת דגימה שמחזירה זוגות כאלו נקרא "דוגם מפורש", ונבנה אותו עכשיו. עבור ההמשך נניח ש־S היא הקבוצה $\{1, \ldots, n\}$, ולמען פשטות הניסוח (שלא יהיו סימנים כמו "$\lfloor \cdots \rfloor$" בכל מקום) נניח שמתקיים $n = 2^k$ עבור k מתאים.

הרעיון של הדגימה יזכיר חיפוש בינארי, רק שכאן אנחנו משתמשים בהסתברויות במקום בהשוואות אינדכס. נבנה עץ בינארי מלא מאוזן מגובה k, כאשר העלים שלו יזוהו עם האיברים של $S = \{1, \ldots, 2^k\}$. הצמתים הפנימיים יזוהו עם תתי־קבוצה של S, ליתר דיוק קטעים. השורש יזוהה עם $\{1, \ldots, 2^k\}$ כולו, וצומת ברמה ה־h יזוהה עם $\{i2^{n-h}+1, \ldots, (i+1)2^{n-h}\}$ עבור $0 \leq i < 2^h$ מתאים. הבנים של צומת זה יהיו זה המזוהה עם תת־הקטע $\{2i2^{n-h-1}+1, \ldots, (2i+1)2^{n-h-1}\}$ וזה המזוהה עם $\{(2i+1)2^{n-h-1}+1, \ldots, (2i+2)2^{n-h-1}\}$. נשים לב ששני תתי הקטעים הנ"ל מהווים חלוקה של הקטע המקורי. לצורך הדיון, כל עלה יהיה מזוהה עם "קטע" בעל איבר בודד.

על מנת להמחיש את ההמשך, נניח שאנחנו מבצעים "חיפוש בינארי" באופן הבא: מתחילים מהשורש. בכל שלב, עבור צומת המזוהה עם קטע I ושני בניו המזוהים עם I_l ו־I_r, נעבור לצומת של I_l בהסתברות $\mu|_I(I_l) = \mu(I_l)/\mu(I)$, ונעבור לצומת של I_r בהסתברות $\mu|_I(I_r) = 1 - \mu|_I(I_l)$. כשנגיע לעלה, נפלוט את האיבר של S המזוהה אתו. אם נבדוק את ההסתברות להגיע לעלה המזוהה עם $a \in S$, ונסמן ב־$S = I_0 \supset I_1 \supset \cdots \supset I_k = \{a\}$ את כל הקטעים המזוהים עם הצמתים שעברנו דרכם, נקבל $\Pr[a] = \prod_{i=1}^{k} \frac{\mu(I_i)}{\mu(I_{i-1})} = \mu(a)$, ז"א שיש בידנו שיטה קצת מסורבלת לדגום לפי μ.

עתה נניח שלכל צומת (לא עלה) המזוהה עם הקטע I והבנים שלו המזוהים עם I_l ו־I_r כפי שהוגדרו למעלה, נשמור ערך α_I, עם ההבטחה שמתקיים $\left|\alpha_I - \mu|_I(I_l)\right| \leq \frac{\epsilon}{k}$. נבצע שוב את התהליך מלמעלה, אבל במקום ההסתברויות $\mu|_I(I_l)$ ו־$1 - \mu|_I(I_l)$ נשתמש בהסתברויות α_I ו־$1 - \alpha_I$ בהתאמה. נסמן את ההתפלגות המתאימה על העלים ב־$\tilde{\mu}$, כך ש־$\tilde{\mu}(a)$ הוא מכפלת ההתפלגויות שהשתמשנו בהן עבור הקטעים המכילים את העלה במסלול משורש העץ לעלה המזוהה עם a. נשים לב שבהינתן a, אפשר לחשב את המכפלה הנ"ל ולפלוט אותה יחד עם הערך a, כך שיש בידנו דוגם מפורש עבור $\tilde{\mu}$.

הטענה המרכזית היא שמתקיים $d(\tilde{\mu}, \mu) \leq \epsilon$. על מנת לראות את זה נשתמש בשיטת הצימוד: על מנת להגריל צמד $(a, b) \in S \times S$, נתחיל משורש העץ, ובכל שלב נעבור מצומת המזוהה עם קטע I לאחד הבנים שלו, I_l או I_r, או לשניהם בו זמנית, באמצעות התהליך הבא.

- בהסתברות $\min\{\alpha_I, \mu|_I(I_l)\}$ נבחר את הצומת I_l. אם זה עלה המזוהה עם $c \in S$, אז נבחר את הערכים $a = b = c$ ונסיים.
- בהסתברות $\min\{1 - \alpha_I, 1 - \mu|_I(I_l)\}$ נבחר את הצומת I_r. אם זה עלה המזוהה עם $c \in S$, אז נבחר את הערכים $a = b = c$ ונסיים.
- בהסתברות הנותרת, $\left|\alpha_I - \mu|_I(I_l)\right|$, נבצע "פיצול". אם $\alpha_I < \mu|_I(I_l)$, אז עבור הערך a נמשיך ללכת במורד תת־העץ של I_l לפי ההסתברויות המתאימות ל־μ, ועבור הערך b נמשיך במורד תת־העץ של I_r לפי ערכי ה־α המתאימים. אם $\alpha_I > \mu|_I(I_l)$, אז עבור הערך a נמשיך ללכת במורד תת־העץ של I_r לפי μ, ועבור הערך b נמשיך במורד תת־העץ של I_l לפי ערכי ה־α.

על מנת לסיים, נשים לב ש־a מתפלג לפי μ (אם מסתכלים רק על החיפוש עבורו אז הסתברויות המעבר כולן קורות לפי ערכי $\mu|_I(I_l)$), ו־b מתפלג לפי $\tilde{\mu}$. ההסתברות עבור $a \neq b$ היא בדיוק ההסתברות לכך שבוצע פיצול בשלב כל שהוא, ולפי איחוד מאורעות הסתברות זו חסומה ע"י ϵ.

כעיקרון אפשר ע"י $O(k^2 \log(1/\delta)/\epsilon^2)$ דגימות מותנות מההתפלגות $\mu|_I$ למצוא, בהסתברות $1 - \delta$ לפחות, ערך α_I המקיים $\left|\alpha_I - \mu|_I(I_l)\right| \leq \frac{\epsilon}{k}$. עם זאת, "לאכלס" את העץ הבינארי המלא ידרוש קירוב של $n - 1$ ערכים כאלה, ולא עשינו כלום (היה אפשר בפחות דגימות פשוט לקרב את μ ישירות). במקום זה אנחנו נכתוב אלגוריתם שיעבוד בצורה "עצלנית": הערך של α_I יחושב רק בפעם הראשונה שנגיע לצומת המזוהה עם I.

עבור ביצוע q דגימות מפורשות ממרחב ההסתברות המקרב את μ, נשתמש באלגוריתם הבא.

- לפני שנתחיל את הדגימות, נבנה את העץ הבינארי עבור תתי הקטעים המתאימים של S, ולכל צומת שאינו עלה נציב ב־α_I את הערך המיוחד "$\perp$".
- כאשר אנחנו נדרשים לדגימה, נבצע את ה"חיפוש" לפי ערכי α_I מהשורש.
  - בכל פעם שהמגיעים לצומת המזוהה עם קטע I, אם הערך של α_I הוא עדיין "$\perp$", נבצע $O(k^2\log(kq/\delta)/\epsilon^2)$ דגימות מותנות, ונחשב ערך α_I שבהסתברות לפחות $1-\frac{\delta}{kq}$ יקיים את הקירוב $\left|\alpha_I-\mu|_I(I_l)\right|\le\frac{\epsilon}{k}$.
  - עתה כשיש ערך מספרי עבור α_I (אם היה קיים מראש או חושב לפי הסעיף הקודם), משתמשים בו: בהסתברות α_I (באופן ב"ת בהגרלות שנעשו עד עכשיו) עוברים ל־I_l, ובהסתברות $1-\alpha_I$ עוברים ל־I_r.
  - כשנגיע לבסוף לעלה המזוהה עם $I=\{a\}$, נפלוט את a ואת החישוב של $\tilde{\mu}(a)$ לפי מכפלת הערכים המתאימה.

נשים לב שלאחר קבלה של q דגימות כאלה, ההסתברות שאלו לא היו כולם לפי ערכי α_I המקיימים את תנאי הקירוב הדרושים חסומה ע"י δ . זה אומר שבהסתברות לפחות $1-\delta$ קיבלנו q דגימות מפורשות מהתפלגות אחת $\tilde{\mu}$ שהיא ϵ־קרובה ל־μ (ההתפלגות $\tilde{\mu}$ עצמה תלויה בתוצאות של תהליכים הסתברותיים, אבל העיקר שזו תהיה אותה $\tilde{\mu}$ לכל q הדגימות שלנו). מספר הדגימות המותנות מ־μ עבור קבלת דגימה מפורשת בודדת חסום ע"י $O(k^3\log(kq/\delta)/\epsilon^2)$.

עתה נראה איך, בהינתן $q=O(\log(n)\log(1/\delta)/\epsilon^3)$ דגימות מפורשות מהתפלגות $\tilde{\mu}$, אפשר למצוא התפלגות $\hat{\mu}$, כך שבהסתברות לפחות $1-\delta$ תהיה פרמוטציה σ שעבורה $d(\hat{\mu}_\sigma,\tilde{\mu})\le 8\epsilon$. זה אומר שבאמצעות הדוגם המפורש ניתן ללמוד בהסתברות לפחות $1-2\delta$, עד כדי פרמוטציה, קירוב של μ עם פרמטר מרחק 9ϵ, תוך ביצוע $O(qk^3\log(kq/\delta)/\epsilon^2)=\tilde{O}((\log(n))^4(\log(1/\delta))^2/\epsilon^4)$ דגימות מותנות.

הרעיון יהיה לבחון את החלוקה לדליים $\mathcal{B}=\{S_0,\ldots,S_r\}$ של $\tilde{\mu}$, לפי $S_0=\{a\in S:\tilde{\mu}(a)<\frac{\epsilon}{n}\}$ ו־$S_j=\{a\in S:\frac{\epsilon}{n}(1+\epsilon)^{j-1}\le\tilde{\mu}(a)<\frac{\epsilon}{n}(1+\epsilon)^j\}$ עבור $1\le j\le r$, כאשר כזכור $r=O(\log(n)/\epsilon)$. נשים לב שכאשר אנחנו מקבלים זוג $(a,\tilde{\mu}(a))$ אנחנו יודעים במדויק את הדלי j שעבורו $a\in S_j$. מכאן שבאמצעות דגימות מפורשות אנחנו יכולים להשיג דגימות מ־$\tilde{\mu}_{\mathcal{B}}$. האלגוריתם פשוט יקרב את $\tilde{\mu}_{\mathcal{B}}$, ויפלוט $\hat{\mu}$ שיהיו לו משקלות דומים של הדליים.

- מבצעים $q=O(\log(n)\log(1/\delta)/\epsilon^3)$ דגימות מפורשות מ־$\tilde{\mu}$, מתרגמים אותם לדגימות מ־$\tilde{\mu}_{\mathcal{B}}$, ומוצאים התפלגות ν מעל $\{0,\ldots,r\}$ כך שבהסתברות לפחות $1-\delta$ מתקיים $d(\nu,\tilde{\mu}_{\mathcal{B}})\le\epsilon$.
- מחשבים התפלגות $\hat{\mu}$ מעל $\{1,\ldots,n\}$ שעבורה מתקיים $d(\nu,\hat{\mu}_{\mathcal{B}})\le\epsilon$, ופולטים אותה (אם אין $\hat{\mu}$ כזו אז פולטים התפלגות שרירותית).

אם הסעיף הראשון מצליח (דבר שקורה בהסתברות לפחות $1-\delta$) אז הסעיף השני לא יפלוט התפלגות שרירותית, מכיוון ש־$\tilde{\mu}$ היא בפרט התפלגות שמקיימת $d(\nu,\tilde{\mu}_{\mathcal{B}})\le\epsilon$. עם זאת לא מובטח שדווקא $\tilde{\mu}_{\mathcal{B}}$ היא ההתפלגות שתיפלט, כל שאנחנו יכולים להבטיח לפי אי־שוויון המשולש הוא שיתקיים $d(\hat{\mu}_{\mathcal{B}},\tilde{\mu}_{\mathcal{B}})\le 2\epsilon$. במודל שלנו אנחנו לא מתחייבים לזמן חישוב מסויים (ולכן לא כותבים איך מחשבים את $\hat{\mu}$), אבל במאמרים המקוריים יש אלגוריתמים למציאת $\hat{\mu}$ בזמן מהיר יחסית, אם מתחילים מקירוב דומה (עם קצת שינויים) לקירוב כאן.

במידה והסעיף הראשון הצליח, אפשר לחסום את המרחק של $\hat{\mu}$ מפרמוטציה מתאימה של $\tilde{\mu}$. לשם כך נחסום את המרחק בין "הגרסאות המקוצצות" של ההתפלגויות. בהינתן התפלגות ν מעל $\{1,\ldots,n\}$, נגדיר את ההתפלגות המקוצצת ν' מעל $\{0,\ldots,n\}$ באופן הבא: נחשב את החלוקה לדליים $\{S_0,\ldots,S_r\}$ של ν. לכל $i\in S_0$ נגדיר $\nu'(i)=0$, ולכל $i\in S_j$ כאשר $1\le j\le r$ נגדיר $\nu'(i)=\frac{\epsilon}{n}(1+\epsilon)^{j-1}$. לבסוף נגדיר $\nu'(0)=1-\sum_{i=1}^n\nu'(i)$. חישוב מהיר מראה שמתקיים $d(\nu,\nu')\le 2\epsilon$ (כאשר מגדירים $\nu(0)=0$), ובפרט $\nu'(0)\le 2\epsilon$.

מתקיים גם $d(\nu_{\mathcal{B}},\nu'_{\mathcal{B}}) \le 2\epsilon$. מכיוון שהשינויים בין ν ל־ν' מתבטאים בהפחתה בלבד של $\nu(1),\ldots,\nu(n)$, נקבל מאלו שמתקיים $d(\hat{\mu}'_{\mathcal{B}},\tilde{\mu}'_{\mathcal{B}}) \le 4\epsilon$ (אם היינו משתמשים ישירות באי שוויון המשולש אז היה מתקבל 6ϵ שם). נסמן עתה ב־$S_0,\ldots,S_r$ את הדליים של $\hat{\mu}$ וב־$T_0,\ldots,T_r$ את הדליים של $\tilde{\mu}$. נגדיר פרמוטציה σ מעל $\{1,\ldots,n\}$, עם ההרחבה $\sigma(0)=0$, לפי כך שלכל $0 \le j \le r$ נתאים $\min\{S_j,T_j\}$ איברים של S_j ל־T_j, ואת שאר האיברים נתאים שרירותית. חישוב מתאים יראה שבהתאמה הזו מתקיים $d(\hat{\mu}'_\sigma,\tilde{\mu}') \le 8\epsilon$ (המרחק יכול להיות מוכפל בגלל שלא חסמנו מראש את הפרש ההסתברויות של האיבר 0), ומכאן ניתן להסיק מאי־שוויון המשולש שמתקיים $d(\hat{\mu}_\sigma,\tilde{\mu}) \le 12\epsilon$.

בדיקה לאיזומורפיזם מול גרף ידוע במודל הצפוף

נחזור למודל בדיקת הגרפים הצפוף. נניח שנתון גרף "ידוע מראש" H בעל n צמתים, ואנחנו רוצים לבדוק גרף G עם אותו מספר צמתים עבור התכונה שהוא איזומורפי ל־H. נראה קודם את אפשרויות הבדיקה כאשר H מקיים פרמטר מתאים של "פשטות" (קצת מזכיר את הגדרת החונטה עבור פונקציות), ואח"כ נראה חסמים עבור גרף כללי H.

איזומורפיזם מול גרף עם סיבוכיות נמוכה

נגיד ש־H הוא מסיבוכיות k אם קבוצת הצמתים של H ניתנת לחלוקה ל־k קבוצות $V_1,\ldots,V_k$, כך שלכל $1 \le i \le j \le n$ מתקיים שבין V_i ל־V_j או שאין קשתות או שיש את כל הקשתות האפשריות (במקרה של $i=j$ זה אומר ש־V_i הוא או קליק או קבוצה נטולת קשתות פנימיות).

עבור H מסיבוכיות k אפשר לבצע ϵ־בדיקה לאיזומורפיזם באמצעות מספר שאילתות שתלוי רק ב־ϵ ו־k: אפשר להשתמש בבודק החלוקות הכללי מהמאמר המקורי של Goldreich, Goldwasser, Ron. כאשר ההגבלות על כל הצפיפויות הן "שווה ל־0" או "שווה ל־1", ויש הגבלות גודל מדוייקות על קבוצות החלוקה, אז הגרף G יכול לקיים את תכונת החלוקה רק אם הוא איזומורפי ל־H. שימו לב אבל שזה אלגוריתם עם שגיאה דו־צדדית, והדבר הכרחי – לא יכול להיות למשל אלגוריתם שיקבל בהסתברות 1 את הגרף ששווה לקליק על $\lceil \frac{n}{2} \rceil$ מהצמתים בדיוק (ללא קשתות המערבות את $\lfloor \frac{n}{2} \rfloor$ הצמתים האחרים), אבל ידחה בהסתברות לפחות $\frac{2}{3}$ את הגרף ששווה לקליק על $\lceil \frac{n}{4} \rceil$ מהצמתים בדיוק.

אפשר לשאול את עצמנו האם אלו כל הגרפים שאפשר לבדוק עם מספר שאילתות שאינו תלוי ב־n. התשובה היא חיובית – נראה עכשיו שאם H הוא ϵ־רחוק מלהיות בעל סיבוכיות k, אז קיים חסם תחתון שתלוי ב־k על מספר השאילתות של ϵ־בדיקה עבור איזומורפיזם ל־H. ההוכחה תהיה באמצעות השיטה של יאו, בתוספת טכניקת "מעיכה" של הקלט.

כזכור, באמצעות מעבר לבדיקה קנונית, אפשר להניח שאלגוריתם הבדיקה הוא לא־אדפטיבי ו"מבוסס צמתים", במחיר ריבועי בלבד במספר השאילתות. הגרסה הדטרמיניסטית של אלגוריתם כזה מתוארת ע"י תת־קבוצה U של קבוצת הצמתים V מגודל q, כאשר קבוצת השאילתות בפועל תהיה קבוצת הזוגות $Q=\binom{U}{2}$, ותנאי הקבלה של האלגוריתם מתואר ע"י תת־קבוצה של קבוצת התשובות האפשריות, $\mathcal{A} \subseteq \{0,1\}^Q$ (בעצם זו משפחה של גרפים בעלי $|U|=q$ צמתים).

נרשום אם כן שתי התפלגויות, τ ו־ν. ההתפלגות τ היא פשוט זו המתקבלת על ידי הפעלת פרמוטציה $\sigma: V \to V$ שנבחרה באופן מקרי ויוניפורמי על H. ברור שזו התפלגות מעל גרפים שמקיימים את התכונה, וגם ברור שההתפלגות המושרה ע"י τ על קבוצת השאילתות $\binom{Q}{2}$ היא של תת־גרף מושרה מקרי של H.

ההתפלגות ν מוגדרת לפי התהליך הבא.

- עבור $r=2q^2$, מגרילים באופן יוניפורמי ובלי חזרות סדרה של צמתים שונים זה מזה $v_1,\ldots,v_r$.
- לכל צומת $w \in V$ נגריל באופן יוניפורמי וב"ת לחלוטין באחרים מספר $1 \le i_w \le r$.

- על מנת להגדיר את $G(V,E)$, לכל $u,w \in V$ שונים זה מזה נגדיר $(u,w) \in E$ אם ורק אם (v_{i_u}, v_{i_w}) היא קשת ב־H.

נתאר במילים את התהליך: אנחנו נחלק את קבוצת הצמתים של G המיועד באופן מקרי ויוניפורמי ל־r תתי־קבוצה, $V_1,\ldots,V_r$, ולכל אחד מהם נתאים "נציג" בתוך קבוצת הצמתים של H. אנחנו נוסיף את כל הקשתות בין V_i ל־V_j אם היתה קשת בין הנציגים שלהם ב־H, ואחרת לא נוסיף ביניהם קשתות כלל. מהתהליך ברור שהסיבוכיות של G תמיד תהיה לכל היותר r, ולכן אם H היה ϵ־רחוק מלהיות בעל סיבוכיות r אז ההתפלגות ν תתן קלט ϵ־רחוק מאיזומורפיזם עם H בהסתברות 1.

נראה עתה שההתפלגויות $\tau|_Q$ ו־$\nu|_Q$ הן קרובות. כזכור $\tau|_Q$ מתוארת ע"י (פרמוטיציה מקרית של) תת־גרף מקרי של H. עבור $\nu|_Q$, נסמן את הצמתים של השאילתות ב־$U=\{u_1,\ldots,u_q\}$, ונבדוק את סדרת האינדכסים $i_{u_1},\ldots,i_{u_q}$. מכיוון שזוהי סדרה שנבחרה יוניפורמית לחלוטין של ערכים ב־$\{1,\ldots,r\}$, ההסתברות שתהיה שם חזרה על אינדכס כל שהוא חסומה ע"י $\frac{1}{8}$. כאשר המאורע הזה אינו קורה, הנציגים $v_{i_{u_1}},\ldots,v_{i_{u_q}}$ יהיו סדרת צמתים שנבחרה יוניפורמית ללא חזרות מתוך הצמתים של H, ולכן המדובר יהיו, בדומה ל־τ, בתת־גרף מושרה מקרי ויוניפורמי של H.

זה אומר שאם מתנים על כך שמאורע החזרה על אינדכסים, שההסתברות שלו חסומה ע"י $\frac{1}{8}$, אינו קורה, אז התפלגות הגרף המושרה מ־ν זהה להתפלגות מ־τ. לכן, ללא התניה, מתקיים $d(\tau|_Q,\nu|_Q) \leq \frac{1}{8}$. מכיוון שקבענו $r=O(q^2)$, הדבר נותן לנו חסם תחתון של $\Omega(\sqrt{k})$ על ϵ־בדיקה מבוססת צמתים עבור התכונה של להיות איזומורפי ל־H.

מכיוון שניתן לתרגם אלגוריתם בדיקה כללי (אפילו אדפטיבי) במודל הגרפים הצפוף עם q שאילתות לאלגוריתם בדיקה מבוסס צמתים עם לכל היותר $2q$ צמתים, אנחנו קיבלנו כאן חסם של $\Omega(\sqrt{k})$ שאילתות על ϵ־בדיקה כל שהיא עבור תכונת האיזומורפיות ל־H.

לסיום, נעיר משהו על ההוכחה כאן: ההוכחה המקורית של חסם תחתון לפי סיבוכיות היתה ארוכה, השתמשה בטכניקות מתקדמות (למת הרגולריות) ונתנה תלות גרועה ב־k (בסגנון $\log^*(k)$). היא הופיעה במאמר Fischer: The difficulty of testing for isomorphism against a graph that is given in advance. שמונה שנים אח"כ נמצאה ההוכחה הפשוטה, במאמר Chakraborty, Fischer, García-Soriano, Matsliah: Junto-symmetric functions, hypergraph isomorphism, and crunching.

חסם תחתון לבדיקה מול גרף כללי

תת־פרק זה והבא מביאים תוצאות מהמאמר Fischer, Matsliah: Testing graph isomorphism. העניין המרכזי בתוצאות שנזכיר הוא הקשר שאפשר למצוא לבדיקת התפלגויות במודל הדגימה הלא־מותנה. אנחנו נראה כאן חסם תחתון של $\Omega(\sqrt{n})$ שאילתות לבדיקת איזומורפיזם מול גרף נתון.

על מנת להוכיח את החסם התחתון, נשתמש בשיטת יאו ונגדיר (כרגיל) שתי התפלגויות על G, אבל ראשית נגדיר את הגרף הנתון H באופן שיתאים לנו. אנחנו נעבוד מעל קבוצת צמתים V מגודל n, ונניח ש־n הוא זוגי וגדול דיו.

עבור גרף נתון $H(V,E)$, ועבור קבוצת צמתים $V' \subset V$ מגודל $\frac{1}{2}n$ בדיוק, נסמן ב־$H_{V'}(V,E')$ את הגרף הבא: נבחר שרירותית פונקציה חח"ע ועל מ־V' ל־$V \setminus V'$ שנסמן ב־α. לכל $v \in V$, נסמן $v'=v$ אם $v \in V'$, ואחרת נסמן $v'=\alpha^{-1}(v) \in V'$. עבור $u,v \in V$, נגדיר $uv \in E'$ אם ורק אם $u'v' \in E$ (ובפרט מתקיים $u' \neq v'$). במילים: $H_{V'}$ הוא הגרף המתקבל מ"הכפלת" קבוצת הצמתים V', עם הכפלה מתאימה של הקשתות המקוריות בתוך V' (כל קשת מקורית תתאים לארבע קשתות ב־$H_{V'}$).

אנחנו נגיד ש־H הוא "עמיד", אם לכל $V' \subset V$ מגודל $\frac{1}{2}n$ הגרף $H_{V'}$ הוא $\frac{1}{32}$־רחוק מלהיות איזומורפי ל־H. עבור כל n גדול מספיק קיימים גרפים עמידים. ניתן לראות זאת באמצעות השיטה ההסתברותית (לאלו מכם שלמדו את הקורס) – מגרילים את H באופן מקרי ויוניפורמי (לכל $u,v \in V$ בוחרים את uv להיות ב־E בהסתברות $\frac{1}{2}$ באופן ב"ת לכל הזוגות). באמצעות חסימת סטיות גדולות ואיחוד מאורעות ניתן

לראות שההסתברות לקיום V' כך ש־$H_{V'}$ יהיה קרוב ל־H היא $o(1)$, ולכן בהסתברות $1-o(1)$ קיבלנו בצורה זו גרף עמיד (ובפרט גרף כזה קיים עבור n גדול דיו).

כאשר יש לנו H עמיד (ידוע מראש), נגדיר שתי התפלגויות על G באופן הבא.

- בהתפלגות τ, נבחר את G להיות איזומורפי ל־H באמצעות פרמוטציה מקרית שנבחרה יוניפורמית.
- בהתפלגות ν, ראשית נבחר $V' \subset V$ מגודל $\frac{1}{2}n$ באופן יוניפורמי מכל הקבוצות המתאימות, ואז נבחר את G להיות איזומורפי ל־$H_{V'}$ באמצעות פרמוטציה מקרית שנבחרה יוניפורמית.

מכיוון ש־H עמיד, ההתפלגות ν תחזיר גרף $\frac{1}{32}$־רחוק מלהיות איזומורפי ל־H בהסתברות 1. נותר רק לראות איך שתי ההתפלגויות האלו מסכלות אלגוריתם בדיקה. כאן אנחנו לא נעבור לאלגוריתם קנוני לא־אדפטיבי, כי כזכור מעבר כזה גובה מחיר ריבועי שאנחנו לא יכולים לשלם הפעם. במקום זאת נשתמש ישירות בקריטריון של שיטת יאו עבור אלגוריתמים אדפטיבים.

נניח ש־Q היא קבוצת שאילתות בת פחות מ־$\sqrt{n}/4$ שאילתות (כל שאילתה מתייחסת כזכור לזוג צמתים), ו־U היא קבוצת כל הצמתים המעורבים בקבוצה Q. בפרט מתקיים $|U| < \sqrt{n}/2$. נסמן $U = \{u_1, \ldots, u_r\}$. עבור גרף G שנבחר לפי τ, נסמן ב־$v_1, \ldots, v_r$ את הצמתים כך שהפרמוטציה המקרית שנבחרה עבור G שולחת את u_i ל־v_i. נשים לב ש־$v_1, \ldots, v_r$ היא סדרה ללא חזרות של r צמתים שנבחרת יוניפורמית מכל הסדרות אפשריות. נשים לב גם שסידרה זו קובעת לחלוטין את כל התשובות ל־Q (אם כי יכול להיות שיש יותר מסדרה אחת שתתן את אותן תשובות).

עתה ננתח את Q ואת $U = \{u_1, \ldots, u_r\}$ עבור גרף הנבחר לפי ν. הפעם נסמן ב־$v_1, \ldots, v_r$ את סדרת הצמתים כך שהפרמוטציה המקרית שולחת את u_i לצומת w_i המקיים $w'_i = v_i$. אם נחזור להגדרות של $H_{V'}$ ושל ν, נראה שכאשר $v_1, \ldots, v_r$ היא ללא חזרות, היא קובעת את התשובות לשאילתות בדיוק באותה צורה כמו $v_1, \ldots, v_r$ בניתוח לפי τ.

נראה שלכל $v_1, \ldots, v_r$ מתקיים $\Pr_\nu[v_1, \ldots, v_r] > \frac{2}{3}\Pr_\tau[v_1, \ldots, v_r]$, ומזה נובע שלכל $h : Q \to \{0,1\}$ מתקיים $\Pr_\nu[h] > \frac{2}{3}\Pr_\tau[h]$. בפרק על יישום שיטת יאו נגד אלגוריתמים אדפטיבים התנאי היה עם ν ו־τ בתפקידים הפוכים, אבל ההוכחה שהתנאי כאן גם מספק חסם תחתון היא כמעט זהה. עבור $v_1, \ldots, v_r$ שמכילים חזרות זה ברור, כי אז מתקיים $\Pr_\tau[v_1, \ldots, v_r] = 0$. תחת ההתפלגות ν, אם מתנים על המאורע שאין חזרות ב־$v_1, \ldots, v_r$, אז מדובר בסדרה שנבחרה יוניפורמית מבין כל האפשריות, בדיוק כמו τ. על כן (על מנת להשתמש בנוסחת הסתברויות מותנות) מספיק להראות שההסתברות שיש חזרות תחת ν בסדרה $v_1, \ldots, v_r$ היא קטנה מ־$\frac{1}{3}$. חישוב ישיר מראה שעבור $1 \leq i < j \leq r$ מתקיים $\Pr_\nu[v_i = v_j] = \frac{1}{n-1}$, ולכן מאיחוד מאורעות הסיכוי לקיום חזרה כל שהיא חסום ע"י $\binom{\sqrt{n}/2}{2}/(n-1) < \frac{1}{8}$.

לסיום, נשים לב לאנלוגיה שיש כאן עם החסם התחתון על בדיקת התפלגות עבור יוניפורמיות: שם ההתפלגות השלילית היתה התפלגות יוניפורמית על קבוצה מקרית של חצי מהאיברים של מרחב ההסתברות, וכאן בחרנו "לנפח" קבוצה מקרית של חצי מהצמתים. עבור החסם העליון נשתמש בבדיקת התפלגויות באופן מפורש.

חסם עליון לבדיקה מול גרף כללי

נראה כאן, לכל ϵ קבוע, חסם עליון של $\tilde{O}(\sqrt{n})$ עבור בדיקת איזומורפיזם מול גרף ידוע H. על מנת שהניתוח יהיה יותר נוח, נראה בדיקה עבור גרפים שיכולות להיות להם לולאות ("קשתות" מצומת לעצמו). כעיקרון צומת v יהיה בקבוצת השכנים של עצמו אם ורק אם קבוצת הקשתות מכילה לולאה על v. מכיוון שאפשור של תוספת לולאות לא מקטין את המרחק בין גרפים שלא היו להם לולאות, ϵ־בדיקה עבור גרפים עם לולאות תתפקד גם כבדיקה עבור גרפים רגילים.

לקראת בניית האלגוריתם המלא, נראה קודם בדיקה עבור פרמוטציה ספציפית σ מעל קבוצת הצמתים V. הבדיקה תשתמש רק ב־$O(\log(n)/\epsilon)$ ערכים של σ, דבר שיאפשר לנו יותר מאוחר לנסות מספר קטן יחסית של אפשרויות, עבורם נשתמש בפרמטר שגיאה δ שיהיה קטן מספיק לאיחוד מאורעות (כזכור, התלות של מספר השאילתות של הבדיקה ב־δ היא $O(\log(1/\delta))$ בלבד). נבצע את הפרוצדורה הבאה עבור σ נתונה.

- נגריל באופן יוניפורמי (עם אפשרות לחזרות) $s = \lceil 10\log(n)\log(1/\delta)/\epsilon \rceil$ צמתים $v_1, \ldots, v_s$.

- לכל צומת $w \in V$, נגדיר את התווית $L(w)$ כווקטור $(a_1, \ldots, a_s) \in \{0,1\}^s$, כאשר $a_i = 1$ אם ורק אם $\sigma(v_i), w$ היא קשת של הגרף הידוע H. כמו כן, נגדיר את התווית $R(w)$ כווקטור $(b_1, \ldots, b_s) \in \{0,1\}^s$ כאשר $b_i = 1$ אם ורק אם v_i, w היא קשת של הגרף G שאנחנו בודקים.

- נגדיר את ההתפלגות μ כתוצאה של בחירה מקרית יוניפורמית של w ולקיחת $L(w)$. זאת התפלגות שאנחנו יכולים לחשב מעל n ערכים לכל היותר (אנחנו מצמצמים את הטווח רק לערכים שיש צומת שמקבל אותם). נגדיר את ההתפלגות ν כתוצאה של בחירה של בחירה מקרית של w ולקיחת $R(w)$, רק שכאשר התווית המתקבלת אינה כזו שיכולה להתקבל לפי μ, נחליף אותה בתווית מיוחדת "$\perp$".

- נבצע ϵ־בדיקה של ν עבור שוויון ל־μ, עם הסתברות שגיאה $\delta/3$. אם הבדיקה דחתה, נדחה את σ עבור G, ואחרת נקבל. נשים לב שכל "שאילתה" של $R(w)$ שאנחנו מבצעים עבור קבלת דגימה מ־ν לוקחת s שאילתות מ־G, כך שסה"כ אנחו מבצעים כאן $\tilde{O}(\sqrt{n}(\log(1/\delta))^2/\epsilon^3)$ שאילתות.

- נאתחל $c = 0$, ונבצע $t = 100\log(1/\delta)/\epsilon$ איטרציות של הבדיקה הבאה.

  – נבחר באופן מקרי ויוניפורמי זוג צמתים $u, w \in V$. בוחרים עבור u צומת מקרי ויוניפורמי u' מבין אלו שמקיימים $R(u) = L(u')$ ולא נבחרו במהלך איטרציה קודמת (אלא אם כן u כבר נבחר באיטרציה קודמת ואז שומרים את u'). אם לא נשאר אף u' כזה אז דוחים את σ מיידית. בוחרים באותו אופן צומת w' עבור w. מגדילים את c ב־1 אם מתקיים ש־uw קשת של G אבל $u'w'$ אינה קשת של H, או ש־uw אינה קשת של G אבל $u'w'$ כן קשת של H. נשים לב שלכל זוג כאן אנחנו מבצעים $2s = O(\log(n)\log(1/\delta)/\epsilon)$ שאילתות מ־G למציאת התוויות שם, בנוסף לשאילתה על uw עצמו (לבחירת u', w' צריך גם לקרוא את כל קשתות H שמכילות צומת מתוך $\sigma(v_1), \ldots, \sigma(v_s)$, אבל H ידוע מראש כך שהדבר לא מצריך שאילתות).

- אם $c > 3\epsilon t$ אז נדחה את σ, ואחרת (אם לא דחינו מסיבה אחרת קודם) נקבל את σ.

לפני שנמשיך, נראה שיטה אלטרנטיבית להסתכל על האיטרציות של בדיקת זוגות הצמתים בשלב האחרון של האלגוריתם. נבחן את קבוצת הפרמוטציות $\tilde{\sigma} : V \to V$ שעבורן $U_{\tilde{\sigma}} = \{v : L(\tilde{\sigma}(w)) \neq R(w)\}$ היא מגודל מינימלי. אנחנו יכולים לחשוב על שלב זה כעל בחירת פרמוטציה $\tilde{\sigma}$ באופן מקרי ויוניפורמי מתוך הקבוצה הנ"ל, בחירה יוניפורמית של הצמתים u, w מתוך הקבוצה $V \setminus U_{\tilde{\sigma}}$, ובדיקת הזוג מול הזוג $\tilde{\sigma}(u), \tilde{\sigma}(w)$. בסעיף כמו שהוא כתוב, קודם בחרנו את u, w באופן יוניפורמי ואח"כ בחרנו את התמונה שלהם לפי $\tilde{\sigma}$ מקרית מבין אלו שממקמות אותם מחוץ ל־$U_{\tilde{\sigma}}$, אולם פרוצדורה זו שקולה (למעט המקרה שבו אנחנו דוחים ממילא).

נראה עתה שבהסתברות $1 - \delta$ לפחות, הבדיקה תקבל אם σ הוא אכן איזומורפיזם מ־G ל־H, ותדחה אם יש לפחות $3\epsilon n^2$ הבדלים בין הקשתות של H ושל כל פרמוטציה אפשרית של G (σ או פרמוטציה אחרת). לשם כך נשים לב שבהסתברות לפחות $1 - \delta$ שלושת המאורעות הבאים קורים.

- לכל $u, v \in V$ שקבוצות השכנים שלהם לפי G נבדלות בלפחות ϵn צמתים (ז"א שיש לפחות ϵn צמתים שכ"א מהם שכן של אחד מ־u ו־v ולא שכן של השני), מתקיים $R(u) \neq R(v)$.

- הבדיקה של ν מול μ אכן קיבלה אם שתי ההתפלגויות זהות, ואכן דחתה אם שתי ההתפלגויות הן ϵ־רחוקות זו מזו.

- בשלב האחרון של האלגוריתם (אם הגענו אליו), קיבלנו אם יש לא יותר מ־ϵn^2 זוגות ($2\epsilon n^2$ זוגות סדורים) של צמתים ב־$V \setminus U_{\tilde{\sigma}}$ שהם קשת ב־G ותמונתם לפי $\tilde{\sigma}$ אינה קשת ב־H, או שהם אינם קשת ב־G ותמונתם היא קשת ב־H. כמו כן דחינו במידה ויש יותר מ־$2\epsilon n^2$ זוגות כאלו ($4\epsilon n^2$ זוגות סדורים). חסם ההסתברות כאן הוא לפי חסימת סטיות גדולות.

אם σ היא אכן איזומורפיזם מ־G ל־H, אז (אפילו בלי קשר למאורע הראשון) לכל $w \in V$ יתקיים $L(\sigma(w)) = R(w)$, ובפרט ההתפלגות μ תהיה זהה ל־ν. נשים לב גם שמתקיים $U_{\tilde{\sigma}} = \emptyset$ לכל $\tilde{\sigma}$ אפשרי.

אם נשווה את $\tilde{\sigma}$ ל־σ, אז בגלל הסעיף הראשון, לכל $u \in V$ יש לכל היותר ϵn צמתים אפשריים $w \in V$ שעבורם הסטטוס של uw (האם זו קשת) ב־G והסטטוס של $w, \sigma^{-1}(\tilde{\sigma}(u))$ ב־G אינם זהים. הסיבה לכך היא המאורע הראשון, כי מתקיים $L(u) = L(\sigma^{-1}(\tilde{\sigma}(u)))$. כמו כן, לכל w יש לכל היותר ϵn צמתים אפשריים $u \in V$ שעבורם הסטטוס של $w, \sigma^{-1}(\tilde{\sigma}(u))$ ב־G והסטטוס של $\sigma^{-1}(\tilde{\sigma}(u)), \sigma^{-1}(\tilde{\sigma}(w))$ ב־G אינם זהים (השתמשנו בזה שההעתקה $\sigma^{-1} \circ \tilde{\sigma}$ היא בפרט פרמוטציה). על כן, בבחירה מקרית של u ו־w, ההסתברות לסטטוס שונה בין u, w לבין $\sigma^{-1}(\tilde{\sigma}(u)), \sigma^{-1}(\tilde{\sigma}(w))$ ב־G חסום ע"י 2ϵ. לבסוף, נשים לב שהסטטוס של $\sigma^{-1}(\tilde{\sigma}(u)), \sigma^{-1}(\tilde{\sigma}(w))$ ב־G זהה לסטטוס של $\tilde{\sigma}(u), \tilde{\sigma}(w)$ ב־H (דרך האיזומורפיזם σ), ולכן אין יותר מ־$2\epsilon n^2$ זוגות סדורים עם סטטוס שונה לתמונה שלהם, וזה אומר שהאלגוריתם יקבל.

עתה נראה את הכיוון השני, שאם המאורעות למעלה התקיימו, אז H קרוב לפרמוטציה כל שהיא של G. מכיוון ש־μ היא ϵ־קרובה ל־ν, מתקיים $|U_{\tilde{\sigma}}| \leq \epsilon n$. אם היו לפחות $2\epsilon n^2$ זוגות לא־סדורים u, w של צמתים ב־$V \setminus U_{\tilde{\sigma}}$ שהסטטוס שלהם ב־G שונה מהסטטוס של $\tilde{\sigma}(u), \tilde{\sigma}(w)$ ב־H אז היינו דוחים. לכן אפשר לחסום את מספר הזוגות שבהם יש הבדלים דרך $\tilde{\sigma}$ ע"י $3\epsilon n^2$ (מחברים ל־$2\epsilon n^2$ את את מספר הזוגות הלא־סדורים שאינם זרים ל־$U_{\tilde{\sigma}}$, אשר חסום ע"י ϵn^2).

עתה נוכל לבנות אלגוריתם ϵ־בדיקה עבור איזומורפיזם מול גרף ידוע. הדבר העיקרי לשים לב הוא שבבדיקה של σ למעלה לא השתמשנו ב־σ כולו, אלא רק בערכים $\sigma(v_1), \ldots, \sigma(v_s)$. כמו כן, את $v_1, \ldots, v_s$ מספיק להגריל פעם אחת, בגלל שכל מה שאנחנו צריכים הוא שיתקיים התנאי על התוויות $R(v)$, שאינן תלויות ב־σ. על כן אנחנו לא צריכים לעבור על כל $n!$ האפשרויות ל־σ, אלא רק על האפשרויות עבור $\sigma|_{\{v_1,\ldots,v_s\}}$ שמספרן חסום ע"י n^s. אנחנו גם נבצע את הבדיקות עבור כל האפשרויות האלו "במקביל". נשתמש באלגוריתם הבא – הצעדים שלו כתובים באופן אנלוגי לצעדים של האלגוריתם עבור σ בודד.

- נגריל את הצמתים $v_1, \ldots, v_s$, עם הפרמטר $\epsilon' = \epsilon/3$, ובשלב זה $\delta = \frac{1}{3}$ (ז"א שההסתברות לקיום שני צמתים עם אותה תווית אבל עם הבדל של יותר מ־$\epsilon n/3$ בקבוצות השכנים לפי G חסומה ע"י $\frac{1}{9}$).
- לכל $w \in V$, התווית $R(w)$ תלויה רק בזהות של $v_1, \ldots, v_s$. נגדיר עבור כל פונקציה חח"ע $\sigma : \{v_1, \ldots, v_s\} \to V$ את התווית $L_\sigma(w)$, כפי שהוגדרה באלגוריתם הפרמוטציה הבודדת (אנחנו לא צריכים "ערכים אחרים של σ").
- בהתאם נגדיר את ההתפלגות ν (שתלויה רק ב־$v_1, \ldots, v_s$) ואת ההתפלגויות μ_σ (שכ"א מהן תלויה גם ב־$\sigma : \{v_1, \ldots, v_s\} \to V$ הספציפית) – בשלב זה לא נחליף ערכים אפשריים של ν בסימן "$\perp$" (זה גם תלוי ב־σ), אלא נעשה את זה אח"כ לכל בדיקת התפלגות בנפרד.
- לכל σ אפשרי, נבצע $\epsilon/3$־בדיקה של ν מול μ_σ, עם הסתברות שגיאה $1/9n^s$ לכל בדיקה ספציפית, כך שההסתברות עבור שגיאה באיזו מהבדיקות חסומה ע"י $\frac{1}{9}$ סה"כ. לכל בדיקה ספציפית נצטרך $\tilde{O}(\sqrt{n}(\log(9n^s)/\epsilon^2)) = \tilde{O}(\sqrt{n}/\epsilon^3)$ שאילתות מ־G.
  - על מנת שלא תהיה לנו הכפלה ב־n^s עבור הבדיקה לכל ה־σ האפשריות, נבצע את הדבר הבא: עבור הדגימות מ־ν, נבחר מראש באופן יוניפורמי את הצמתים $w_1, \ldots, w_r$ כך ש־$R(w_1), \ldots, R(w_r)$ ישמשו כדגימות מתוך ν, ונשתמש באותן דגימות לכל ההתפלגויות μ_σ. השיקול של איחוד מאורעות עדיין נכון, כך שההסתברות לשגיאה עדיין חסומה ע"י $\frac{1}{9}$.
- עבור כל σ שלא נדחתה בסעיף הקודם, עכשיו נבצע את האיטרציות של בדיקת זוג u, w מול הזוג u'_σ, w'_σ (שימו לב לתלות ב־σ), עם $\epsilon' = \epsilon/3$ (גם כאן), ועם $t = O((\log(n)^2/\epsilon)$ שיבטיח שהסיכוי לטעות עבור כל σ ספציפי חסום ע"י $1/9n^s$ (כך שהסיכוי לטעות עבור σ כל שהוא חסום ע"י $\frac{1}{9}$).
  - גם כאן, נבחר מראש את $u_1, w_1, \ldots, u_t w_t$, שבהם נשתמש בבדיקה לכל ה־σ האפשריים (כן יהיו הבדלים ב־$u'_{1,\sigma}, w'_{1,\sigma}, \ldots, u'_{t,\sigma} w'_{t,\sigma}$ עבור σ שונים, אבל אלו גורמים רק לקריאות מתוך H, אשר אינן מצריכות שאילתות).
- אנחנו נקבל את הגרף G אם קיימת σ שעברה את בדיקת μ_σ ושעבורה הספירה מהאיטרציות של הסעיף הקודם מקיימת $c_\sigma \leq \epsilon t$ (אין את המקדם "3" כי בחרנו $\epsilon' = \epsilon/3$). אחרת נדחה את G.

ההוכחה שזהו אלגוריתם בדיקה היא כמעט מיידית בשלב זה: בהסתברות לפחות $\frac{2}{3}$, הצמתים $v_1, \ldots, v_s$ מקיימים את התנאי הקשור לתוויות $R(v)$, וגם כל הבדיקות למעלה נתנו תשובות נכונות לכל האפשרויות עבור $\sigma : \{v_1, \ldots, v_s)$. אם G הוא אכן איזומורפי ל־H, אז בפרט נקבל לפי הצמצום של האיזומורפיזם ביניהם ל־$\{v_1, \ldots, v_s\}$. אם G הוא ϵ־רחוק מאיזומורפיזם כל שהוא ל־H, אז כל האפשרויות עבור $\sigma : \{v_1, \ldots, v_s\} \to V$ ידחו ולכן G ידחה.