אוניברסיטת תל אביב

סמסטר א' תשפ"ב

מבני נתונים – פרויקט מספר 2 – הערה

להלן פתרון עבור הגרסה המקורית של שאלה-2 שפורסמה. בגרסה המקורית האיברים הוכנסו בסדר יורד, בעוד שהכוונה הייתה, כפי שפורסם בתיקון – להכניס את האיברים בסדר עולה.

הגרסה ה"שגויה" קשה יותר בסדר-גודל, ואינה ברמה הנדרשת בקורס. למרות זאת, לטובת העשרה ולטובת מי שניסו לפתור את התרגיל בגרסה המקורית שפורסמה – מצורף פתרון לשאלה.

שאלה 2 – הגרסה הקשה

בגרסה המקורית שבה פורסם התרגיל, סדרת ההכנסות הייתה הפוכה. כלומר:

```
for k=m, m-1, ..., 0: insert(k)
for i=1, ..., 3m/4: deleteMin()
```

להבדל אין השפעה על חסם זמן הריצה האסימפטוטי (סעיף ב'), ואין שינוי במספר החיתוכים (אפס) והפוטנציאל הסופי (כמספר העצים שנותרים בסוף). **ההבדל היחיד הוא במספר החיבורים (links) שמתבצעים לאורך הריצה, וכעת ננתח אותם בתרחיש זה.**

הבהרה: הניתוח שלהלן אינו ניתוח פורמלי מלא, וחלק מהטענות דורשות פרמול או מושארות ללא-הוכחה (תרגיל לקורא/ת) לצורך תמציתיות. כמו-כן, הביטוי שנקבל פשוט ונקי, אך אינו מדויק. ניתן לבצע חישוב מדויק תמורת סרבול הביטוי, כפי שיובהר.

הערת **טרמינולוגיה**: נאמר שעץ "שמן/רזה" כדי לתאר את מספר האיברים בו, ונאמר שעץ "גדול/קטן" מעץ אחר כדי לציין שהאיברים בו גדולים/קטנים מעץ אחר. כמו-כן נסמן $T_1 < T_2$ אם $\forall t_1 \in T_1, t_2 \in T_2 : t_1 < t_2$ ללא-קשר למספר האיברים בעצים.

א. אבחנה: המחיקה הראשונה מניבה את אותם $m - ones(m)$ חיבורים, כמו בפתרון של הגרסה הקלה. ההבדל הוא שהעצים ה"רזים" מכילים את האיברים הגדולים, בעוד שהעצים ה"שמנים" מכילים את האיברים הקטנים. נקרא לעצים הללו "עצים מקוריים".
לכן, כעת בעת מחיקת איברים קטנים, העצים השמנים יתפרקו, ונקבל עוד חיבורים בתהליך ה-consolidation. בהמשך הדיון, ננתח רק את החיבורים ה"עודפים", ביחס ל- $d = \frac{3m}{4} - 1$ צעדים נוספים של מחיקת מינימום.

ב. בתרגיל הגדרנו סדר-קדימויות כך שכאשר מחברים בין עצים, השורש-שאינו-שורש הופך לבן השמאלי ביותר של השורש-שנשאר-שורש. **טענה**: סדר הבנים של קודקוד הוא בסדר דרגות יורד. נדגיש שמכיוון שבכל הריצה מתבצעות רק פעולות delete-min, כל העצים בדיון הם עצים בינומיים.

ג. לטענה הנ"ל השלכות מרחיקות-לכת: כאשר העץ השמן ביותר מתפרק, סדר ה-consolidation קורה מעצים שמנים לרזים, ולכן הסדר בין העצים שמכילים את האיברים הקטנים ביותר אינו משתנה, כך שהעצים הרזים ביותר (שאינם מקוריים) נשארים רזים.
לדוגמה: נניח ש- $m = 16 + 8 + 4$. אז לאחר המחיקה הראשונה (תזכורת: מוכנסים $m + 1$ איברים) נקבל שלושה עצים A_4, B_8, C_{16} (אינדקס: מספר איברים בעץ), כך ש- $A < B < C$.
לאחר המחיקה הבאה, העץ C מתפרק לעצים C_8, C_4, C_2, C_1 (לפי הסדר) ולכן תוצאת ה-consolidation תהיה:

$$C_1, C_2, \{A_4 \cup C_4\}, \{B_8 \cup C_8\}$$

מכיוון ש- C עץ בינומי במקור, מתקיים $C_1 < C_2 < C_4 < C_8$.

ד. מכיוון שאיברי C הם הקטנים ביותר, הצעדים הבאים (של מחיקות מינימום) מחסלים אותם לפי הסדר, בסדר ה"טוב" שמחסל תתי-עצים מדרגה נמוכה קודם. בפרט, בדוגמה שלעיל, מחיקת כל איברי $C_1 \cup C_2 \cup C_4$ תתרחש בלי חיבורים. ייווצרו חיבורים רק כאשר נצטרך למחוק את $\min(C_8)$, בגלל אינטראקציה עם A_4.

באופן כללי יותר: העצים המקוריים "מפריעים" להתפרקות ה"טבעית" של העצים, ובכל-פעם כשעץ מתפרק "מעליהם" (מבחינת הביטים שמייצגים העצים, שכן גדליהם חזקות-2), הם גורמים לחיבור. אם אין עץ מקורי מתחת לעץ שמתפרק, אז לא יהיו חיבורים (כמו שקורה אם $m = 2^k$ עבור k שלם, או בגרסה ה"קלה" של שאלה זו.)

ה. על-סמך האבחנה קודמת, נספור את החיבורים שמתרחשים כך: בכל קבוצה שמתחברת, מוכרח להיות עץ מקורי כלשהו. מספר החיבורים הוא מספר העצים שהתחברו פחות אחד. לדוגמה, $\{A_4 \cup C_4\}$ הוא חיוב של חיבור בודד על A_4. בהמשך, כשנמשיך למחוק ונקבל את העץ $\{A_4 \cup B_8 \cup C_4\}$, נחייב את A_4 על שני חיבורים (B_8 לא יחויב).
נדגיש: ספירת החיבורים מחויבת ביחס לעצים המקוריים, ותמיד ביחס לעץ אחד מסוים.

ו. כאמור, העצים המקוריים יוצרים חיבור כאשר מתפרק מעליהם עץ ששמן מהם. מכיוון שה-consolidation קורה מהעצים השמנים לרזים, ניתן להפריד את הדיון לכל עץ בנפרד, ולראות כמה חיבורים ספרנו עבורו בסך-הכל.
יהי עץ מקורי T שגודלו 2^k. אזי, מדי 2^{k+1} מחיקות מינימום נבצע חיבור שלו. ניתן לראות זאת משום שכאשר קורה חיבור, העץ T התחבר עם עץ בגודל 2^k ש"נולד" מהפירוק של עץ שמן ממנו, ומלבד זאת "נולדו" עוד k עצים רזים וגם קטנים ממנו שסך-מספר האיברים בהם הוא $2^k - 1$ שמצאו את עצמם בדרגות נמוכות יותר (בין אם התאחדו עם עצים מקוריים רזים יותר, או לא). במשך $2^k - 1$ הצעדים הבאים כל הפירוקים/איחודים שיקרו אינם נוגעים ב- T, וב- 2^k הצעדים הבאים המחיקות תהיינה מהעץ שהתחבר עם T, לכן עדיין לא נקבל אף חיבור.

סיכום ביניים: עץ שגודלו 2^k סופר חיבורים מדי 2^{k+1} צעדים. כעת נשים לב שמספר החיבורים משתנה, משום שלפעמים יש חיבור אחד ולפעמים יותר:
חצי מהחיבורים מחברים את T עם עץ אחד בלבד.
רבע מהחיבורים מחברים את T עם שני עצים (כגון $\{A_4 \cup B_8 \cup C_4\}$ בדוגמה).
באופן דומה, $\frac{1}{2^i}$ מהחיבורים של T מבצעים i חיבורים.
נדגיש: לעץ T "לא אכפת" אילו עצים מתפרקים מעליו, משום שמשמעות הפירוק היא שכל העצים שנוצרים מהפירוק מכילים איברים קטנים ממנו, וכל העצים הרזים יותר יסיימו את ה-consolidation "מתחתיו". לכן התבנית של מספר החיבורים שספרנו עבורו מתקיימת לאורך כל המחיקות.
סימנו את מספר המחיקות ב- $d + 1$, כלומר $d = \frac{3m}{4} - 1$. בסך-הכל, אם גודל העץ הוא 2^k, אז מספר החיבורים שנחייב אותו בהם, **בקירוב**, הוא:

$$d \cdot \frac{1}{2^{k+1}} \cdot \left(\frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \cdots\right) \approx \frac{d}{2^{k+1}} \cdot \sum_{i=1}^{\infty} \frac{i}{2^i} = \frac{d}{2^k}$$

נדגיש: מדובר בחסם-עליון כי המחזורים אינם בהכרח מלאים, והטור קטום ולא אינסופי.

ז. נרחיב את הניתוח לכל העצים ונסמן $m = \sum_{s_i \in S} 2^{s_i}$ כאשר S קבוצת השלמים שהם הביטים הדלוקים בייצוג של m. אז מספר החיבורים הנוסף הוא בקירוב:

$$\sum_{s_i \in S} \frac{d}{2^{s_i}} = \frac{d}{2^{\lfloor \lg m \rfloor}} \sum_{s_i \in S} 2^{\lfloor \lg m \rfloor - s_i} = \frac{d \cdot rev(m)}{2^{\lfloor \lg m \rfloor}}$$

כאשר $rev(m)$ הוא המספר שנקבל מהיפוך הייצוג הביטי של m.

לסיכום, קיבלנו שמספר החיבורים בתרחיש ההכנסה-ההפוכה הוא בקירוב:

$$m + ones(m) + \frac{\left(\frac{3m}{4} - 1\right) \cdot rev(m)}{2^{\lfloor \lg m \rfloor}}$$

ניתן לדייק את ההערכה אם במקום לחשב את הטורים במספרים ממשיים נחשב לכל $s_i \in S$ תרומה מדויקת יותר. בפרט, התרומה צריכה להיות מספר שלם, וכתלות ב-m ובמספר המחיקות d, ניתן לחשב בדיוק כמה צעדים יש ובהתאם לשקלל לכל s_i את התרומה המדויקת. **מכיוון שזה לא ייתן ביטוי נקי, נסתפק בהצגת חסמים "הדוקים יחסית" על הביטוי.**

עבור שני החסמים ביחד: נשים לב שמלכתחילה הביטוי $\frac{d}{2^{s_i}}$ הוא לא מדויק, משום שהוא מניח "אחידות" של מספר החיבורים שמתווספים בכל צעד, אך כמובן שבכל צעד מתווסף מספר שלם, ושונה. נבחין שבצעד הראשון קורים $s_{i+1} - s_i$ חיבורים. מקיפול טלסקופי, נקבל שבצעד הזה יש עודף של (כמעט) $\max_j s_j - \min_j s_j$ חיבורים.

לאחר הצעד הראשון, בכל ביט מתחיל "מחזור" של בלוקים של צעדים, כך שבכל בלוק $2^{s_i+1} - 1$ הצעדים הראשונים אינם מבצעים חיבור, והצעד הבא מבצע, לפי התבנית +1,+2,+1,+3,... (באופן כללי: לפי אינדקס הביט הדלוק הנמוך ביותר, ועוד 1).

באופן כללי אם נתעלם ממספר החיבורים ההתחלתי של $s_{i+1} - s_i$, הערך $\frac{d}{2^{s_i}}$ מקדים את המספר האמיתי של חיבורים משום שערך ℓ מסוים "תורם" לתבנית רק מדי 2^{ℓ} צעדים בה, ולכן התרומה תמיד בפיגור. סך כל הפיגורים חסום ב- $\max_j s_j - \min_j s_j + 1$ כי זה מספר החיבורים המקסימלי שיכול לקרות בצעד בודד. (נדרשת הוכחה שה"עקיבה" של הביטוי הממשי אכן לא סוטה יותר ממספר החיבורים האפשריים בצעד בודד).

לסיכום, הסטייה היא לא יותר מ- $\max_j s_j - \min_j s_j + 1$ כלפי מעלה או מטה, ביחס לפירוק הראשון. אבל, התבניות אינן ממשיכות באופן אינסופי, אלא נגמרות כאשר העץ-המקורי השמן ביותר מסיים להימחק, ואז מתחיל מחזור חדש. למחזור החדש תיתכן סטייה משלו, והיא חסומה ב- $\max_{s \in S'} s - \min_j s_j + 1$ עבור $S' = S \setminus \max(S)$. נעיר שאם לא קיים עץ-מקורי שגודלו חצי מהעץ-המקורי השמן ביותר, אז באופן אפקטיבי לאחר שהעץ השמן נשבר, מחצית ממנו תהיה העץ שיתחיל את המחזור הבא, כאילו זה היה עץ-מקורי (וכן הלאה אם נמשיך במחיקות).

נבחין שבמקרה המיוחד שלנו, מכיוון שמתבצעות $\frac{3m}{4}$ מחיקות, יתבצעו בדיוק שני מחזורים – פירוק של העץ השמן ביותר (תואם לביט העליון בייצוג של m), ופירוק של העץ הבא (שתואם לביט שמתחתיו). מכאן שהחסם לסטייה הוא $2 \cdot (\max(S) - \min(S) + 1) - 1$ (השתמשנו בעובדה ש- $\max(S) > \max(S')$ כדי לחסר 1). זהו בדיוק מספר הביטים בייצוג של m אם נזניח ביטים תחתונים שהם 0. נעיר שבמקרה הכללי, כאשר יש יותר מחיקות, הסטייה גדולה יותר. במקרה הקיצון, כאשר נבצע m מחיקות, החסם לגודל הסטייה ריבועי בהפרש $\max(S) - \min(S)$.

לפני שנסיים, נדרש עוד תיקון נוסף **עבור החסם התחתון**: נעיר שהתרומה מכל עץ היא שלמה, ולכן היה נכון יותר לרשום את הביטוי $\sum_{s_i\in S}\left\lfloor\frac{d}{2^{s_i}}\right\rfloor$ על-פני $\sum_{s_i\in S}\frac{d}{2^{s_i}}$. אבל הביטוי הראשון לא מניב ביטוי קומפקטי כמו השני, ולכן הקומפקטיות אולי הגדילה את הביטוי ב- x כך ש- $x<ones(m)$.

לסיכום:

כאשר מספר המחיקות הוא $d+1$ נקבל:

$$\#_{links}=m-ones(m)+\frac{d\cdot rev(m)}{2^{\lfloor\lg m\rfloor}}+\Delta$$

כאשר עבור d כללי, להלן $b=bits\big(rev(m)\big)$ (bits = מספר הביטים בייצוג הבינארי) ומתקיים:

$$-\left(\frac{b(b+1)}{2}+ones(m)\right)<\Delta\le\frac{b(b+1)}{2}$$

ניתן לשפר עבור $d\le\frac{3m}{4}$, ולהדק את החסמים:

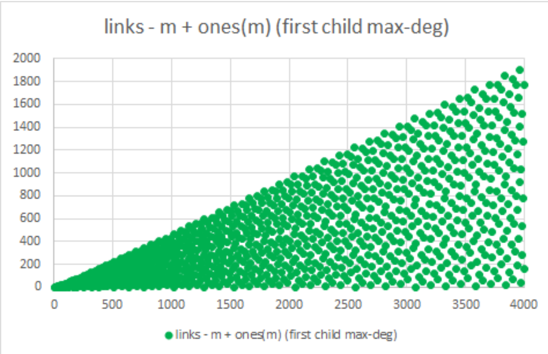
$$-\big(2b-1+ones(m)\big)<\Delta\le 2b-1$$

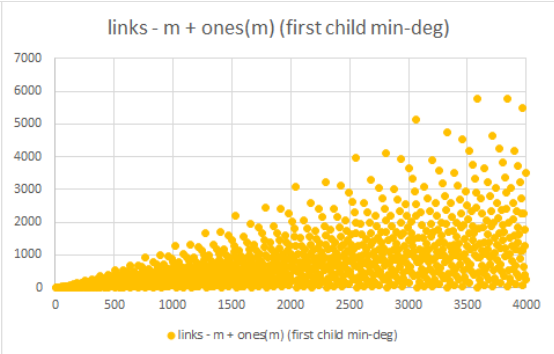
נבחין שבביטויים אסימפטוטיים, $2^{\lfloor\lg m\rfloor}=\Theta(m)$, כמו-כן $d,rev(m)=O(m)$, ולבסוף $b,ones(m)=O(\lg m)$ (הוא O ולא Θ אם יש הרבה אפסים נמוכים בייצוג של m). לכן **קיבלנו ביטוי די הדוק למספר החיבורים: ליניארי ב- m, עם סטיית לוג-בריבוע לכל היותר, לכל d.**

קוד פייתון לחישוב ההערכה והחסמים:

```
def ones(n): return bin(n).count('1')
def bits(n): return len(bin(n)[2:]) # remove the '0b'
def rev(n): return int(bin(n)[2:][::-1],2) # note: we do not reverse with a fixed
width, so the result is always odd (when n>0).
def top_bit(n): return len(bin(n)[2:])-1 # 0-based indexing.

def estimate_extra_reals(m,numDeletes):
    return (numDeletes-1) * (1.0 * rev(m) / 2**top_bit(m))
def estimate_extra_integers(m,numDeletes):
    s,d = 0,1
    while 2*d <= m:
        if m&d == d: s += numDeletes//d
        d *= 2
    return s
def links_bounds(m,numDeletes = None,useReals=True):
    b = bits(rev(m))
    spread = b*(b+1)/2
    if numDeletes == None:
        numDeletes = 3*m//4
        spread = 2*b-1
    lb_fix,func = 0,estimate_extra_integers
    if useReals:
        func = estimate_extra_reals
        lb_fix = ones(m)  # looser due to not rounding separate contributions.
    links = m - ones(m) + func(m,numDeletes)
    return (links-spread-lb_fix,links+spread) # (lower-bound, upper-bound)

def links_bounds_integers(m,numDeletes = None):
    return links_bounds(m,numDeletes,False)

def links_bounds_reals(m,numDeletes = None):
    return links_bounds(m,numDeletes,True)
```

הערה אחרונה בהחלט: היה לנו "מזל" שכאשר מחברים עצים, השורש-שאינו-שורש הופך לבן הראשון, ושסדר הדרגות הוא יורד. אילו היה סדר-עולה, אזי בפירוק של עץ שמן (עם איברים קטנים), תהליך ה-consolidation היה גורם לכך שכל העצים המקוריים יתקפלו ביחד, תוך-כדי טיפוס של עצים רזים-מאוד שיצאו מהפירוק, שמטפסים בחזרה למעלה. בגלל ההתנהגות הזו, האיברים הקטנים היו חוזרים שוב-ושוב להיות חלק מהעץ הגדול ביותר, ויתקבלו הרבה יותר פירוקים-וחיבורים בגלל ההתנהגות הזו. כמה זה יותר? פרופיל הגידול עבור ערכי m מסויימים (שצורתם $m_k = 4 \cdot \left(2^k - 1\right)$) גדל כמו $\propto m \lg m$. להשוואה, להלן שני גרפים עבור מספר החיבורים עבור $m = 4,8,\dots,4000$ לאחר שחיסרנו את הערך $m - ones(m)$ כדי לראות ביתר בהירות רק את החיבורים שהתווספו לאחר בניית היער הראשונית:





שימו לב לסקאלות השונות. בירוק: הניתוח שביצענו. בצהוב: התרחיש האחר, חיבור כבן-ימני.