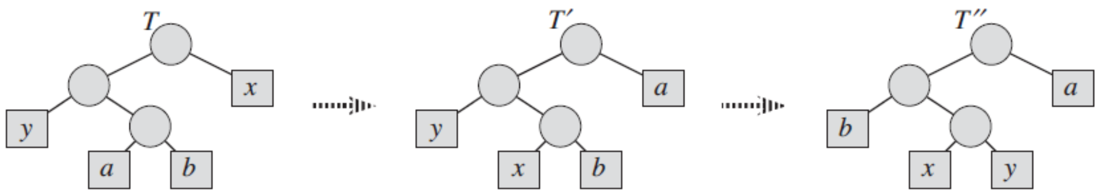
שעור 6 –Huffman coding הוכחת נכונות

נכונות האלגוריתם של האפמן.

למה 1
יהיה C אלפבית, לכל $c \in C$ מוגדרת תדירות $c.freq$. יהיה x,y שני תווים בעלי תדירות מינימאלית. אזי קיים prefix code אופטימאלי ל-C כך שקודים של x,y הם בעלי עומק מקסימאלי ,שווי אורך ונבדלים רק בסיבית אחרון, כלומר עלים סמוכים.

הוכחה. ניקח T עץ שמיצג prefix code אופטימאלי ונשנה את T כך שנקבל prefix code אופטימאלי אחר שבו תווים x,y יהיו עלים בעלי עומק מקסימאלי בעץ חדש.
יהיו a,b שני תווים שהם עלים סמוכים בעלי עומק (מרחק עד השורש) מקסימאלי בעץ T.
ללא אובדן של הכלליות, אנו מניחים כי $a.freq \le b.freq$ ו-$x.freq \le y.freq$. בגלל ש- x,y הם בעלי תדירות מינימאלית אז $x.freq \le a.freq$ וגם $y.freq \le b.freq$.

נניח כי $x.freq \neq b.freq$, כלומר $x \neq b$ (אם $x = b$ הלמה טריוויאלית) . נחליף את a ו-x כמו שרואים באיור 2:



נקבל עץ חדש T', נחליף את y ו- b, נקבל T'' שבו x,y קדקודים סמוכים בעלי עומק מקסימאלי. ההפרש בין העלויות:

$$B(T) - B(T') = \sum_{c \in C} c.freq \cdot d_T(c) - \sum_{c \in C} c.freq \cdot d_{T'}(c) =$$
$$x.freq \cdot d_T(x) + a.freq \cdot d_T(a) - x.freq \cdot d_{T'}(x) - a.freq \cdot d_{T'}(a) =$$
$$x.freq \cdot d_T(x) + a.freq \cdot d_T(a) - x.freq \cdot d_T(a) - a.freq \cdot d_T(x) =$$
$$(a.freq - x.freq) \cdot \left(d_T(a) - d_T(x)\right) \ge 0$$

כאן $a.freq - x.freq \ge 0$ כי x- הוא קדקוד בעל תדירות מינימאלית, ו- $d_T(a) - d_T(x) \ge 0$ כי a הוא בעל עומק מקסימאל.

באופן דומה נחליף y ו-b ונקבל $B(T') - B(T'') \ge 0$, לכן $B(T) - B(T'') \ge 0$
בגלל ש-T הוא אופטימאלי אז ערכך של פונקציה המטרה שלו קטן מערך של כל פונקציה אחרת, כלומר $B(T) \le B(T'')$ ומכאן נובע כי $B(T) = B(T'')$. לכן ו-T'' הוא אופטימאלי ובו x,y הם קדקודים סמוכים בעלי עומק מקסימאלי הנבדלים רק בסיבית אחרון. מש"ל.

למה 2 יהיה C אלפבית, לכל $c \in C$ מוגדרת תדירות $c.freq$. יהיה x,y שני תווים בעלי תדירות מינימאלית. נמחק מ-C את קדקודים x,y ונוסף קדקוד Z כך ש- $z.freq = x.freq + y.freq$, קבלנו C' - אלפבית חדש. יהיה T' עץ אופטימלי עבור C'.
אז $\mathrm{B(T)} = \mathrm{B}(T') + x.freq + y.freq$

הוכחה: לכל תו $c \in C - \{x,y\}$ מתקיים $d_T(c) = d_{T'}(c)$, לכן $d_T(c) \cdot c.freq = d_{T'}(c) \cdot c.freq$. מכוון ש- $d_T(x) = d_T(y) = 1 + d_{T'}(z)$ יש לנו

$$x.freq \cdot d_T(x) + y.freq \cdot d_T(y) = (x.freq + y.freq) \cdot (d_{T'}(z) + 1) =$$
$$z.freq \cdot d_{T'}(z) + (x.freq + y.freq)$$

מכאן מקבלים כי

$$B(T) - B(T') = x.freq \cdot d_T(x) + y.freq \cdot d_T(y) - z.freq \cdot d_{T'}(z) = x.freq + y.freq,$$

או

$$B(T') = B(T) - x.freq - y.freq$$ מש"ל.

משפט עץ T שהתקבל ע"י אלגוריתם האפמן הוא אופטימלי.

הוכחה: באינדוקציה.

א. בסיס האינדוקציה: $n = 2$, אלפבית $C = \{a,b\}$. האלגוריתם נותן קוד 0 ל-a ו-1 ל-b, או הפוך – תלוי בתדירות של האותיות. ברור שקוד המורכב מסיבית אחת הוא אופטימלי.

ב. הנחת אינדוקציה: נניח שקוד של האפמן אופטימאלי עבור $n - 1$ תווים. נוכיח שהוא אופטימאלי עבור n תווים.

ג. נבנה $C_1 = C - \{x,y\} \cup z$, כאשר $z.freq = x.freq + y.freq$, וללא אובדן של כלליות ניתן להניח כי x,y הם עלים סמוכים בעלי עומק מקסימאלי (למה 1). בגלל ש-C_1 מכיל $n - 1$ תווים אלגוריתם של האפמן נותן T_1 עץ אופטימלי עבור אלפבית C_1 (הנחת אינדוקציה).
יהיה T הוא עץ הבנוי לפי האלגוריתם עבור n תווים.
נניח בדרך השלילה שהוא לא מייצג קוד אופטימאלי עבור אלפבית C. לכן קיים עץ אופטימאלי T_2 כך ש- $B(T_2) < B(T)$. בגלל ש- T_2 הוא עץ אופטימלי ,ניתן להניח כי x,y הם עלים סמוכים ב-T_2. יהיה T_3 עץ שבנוי מ-T_2 כך שקדקוד האב של T_2 הוחלף בעלה z, כך ש- $z.freq = x.freq + y.freq$. אז

$$B(T_3) = B(T_2) - x.freq - y.freq < B(T) - x.freq - y.freq = B(T_1)$$

סתירה לעובדה ש- T_1 אופטימאלי. מש"ל.