18: Min Set Cover Linear Programming Algorithms

עבור היפר-גרף $H := (V, \mathcal{E})$, פונקציית משקל $c: \mathcal{E} \to \mathbb{R}^+$. נמצא תת-קבוצה של צלעות $C \subseteq \mathcal{E}$ שמכסה את כל הקודקודים, שממזערת את $\sum_{e \in C} c(e)$.

הגדרת בעיית ה-*IP*

יהיה לנו וקטור x שמייצג את הפתרון. כל אינדקס בווקטור מייצג צלע:

$$x_e := \begin{cases} 1, & e \in C \\ 0, & else \end{cases}$$

ואנחנו נרצה למזער את $\sum_{e \in \mathcal{E}} c(e) x_e$, תחת האילוצים:

נגדיר לכל $v \in V$ את קבוצת הצלעות שמכסות אותו: $C_v := \{e \in \mathcal{E} : v \in e\}$.

נדרוש שיתקיים $\sum_{e \in C_v} x_e \geq 1$ לכל $v \in V$. כלומר, לכל קודקוד, לפחות צלע אחת שמכסה אותה תהיה חלק מהכיסוי.

בעיית ה-*IP* היא *NPC*. נבצע *relaxation* ל-*LP*:

הגדרת בעיית ה-*LP*

הכל אותו דבר, חוץ מזה ש- $0 \leq x_e \leq 1$. כרגיל, אפשר פשוט לדרוש $x_e \geq 0$, כי אם יש $x_e > 1$, היינו מקבלים פתרון טוב יותר עם $x_e = 1$.

אז דרישת המינימום תיתן לנו את ההגבלה $x_e \leq 1$.

עיגול דטרמיניסטי של ה-*LP* הראשונית (*Deterministic Rounding of the Primal LP*)

עבור $v \in V$, נסמן את התדירות (*frequency*) שלו – גודל **הכוכב** שמכיל אותו. כלומר מספר הצלעות שמכסות אותו: $f_v := |C_v|$.

ונגדיר $f := \max_{v \in V} f_v$ את התדירות המקסימלית.

אז האלגוריתם לבעיית *SC*:

1. נמצא פתרון אופטימלי לבעיית ה-*LP*. נסמן את הפיתרון $(x_e^*)_{e \in \mathcal{E}}$.
2. נחזיר את: $A := \{e \in \mathcal{E} : x_e^* \geq 1/f\}$

טענה: האלגוריתם הוא *f*-מקרב.

הוכחה: ראשית, נראה ש-*A* היא אכן כיסוי.

יהי $v \in V$ כלשהו. לפי הפיתרון שקיבלנו מה-*LP*, מתקיים: $\sum_{e \in C_v} x_e^* \geq 1$.

כלומר קיימת $e \in C_v$ כך ש:

$$x_e^* \geq^{\text{א}} \frac{1}{|C_v|} =^{\text{ב}} \frac{1}{f_v} \geq^{\text{ג}} \frac{1}{f}$$

א. שובך היונים.
ב. הגדרת f_v.
ג. כי $f \geq f_v$ לכל *v*.

אזי לפי הגדרת *A*, הצלע *e* שמכסה את *v* תהיה בכיסוי.

נוכיח את איכות הקירוב

נגדיר ווקטור *z* שמזכיר את הווקטור האופייני של *A*. לכל $e \in \mathcal{E}$, נגדיר:

$$z_e := \begin{cases} 1, & e \in A \\ x_e^*, & e \notin A \end{cases}$$

נשים לב שלכל $e \in \mathcal{E}$, מתקיים:

$$z_e \leq f \cdot x_e^*$$

עבור $e \in A$, מתקיים $x_e^* \geq 1/f$ (כי זה מה שגרם ל-*e* להיכלל ב-*A*). אז מתקיים ש:

$$f \cdot x_e^* \geq f \cdot (1/f) = 1 = z_e$$

עבור $e \notin A$, ניזכר שמתקיים $f \geq 1$ (באופן כללי, כי ברור שלכל קודקוד יש לפחות צלע אחת שמכסה אותו כי אחרת אין כיסוי). ונקבל:

$$z_e = x_e^* \leq f \cdot x_e^*$$

נוכל עכשיו להעריך את עלות הפתרון שחזר מהאלגוריתם. נסמן אותו $c(A)$. מתקיים:

$$c(A) \leq^{א} \sum_{e \in \mathcal{E}} z_e \cdot c(e) \leq^{ב} f \cdot \sum_{e \in \mathcal{E}} x_e^* \cdot c(e) =^{ג} f \cdot \mathrm{OPT}_f \leq^{ד} f \cdot OPT$$

א. כי אם $e \in A$, אז $z_e = 1$, וזה מוסיף לסכום בדיוק את המשקל של הצלע.
אם $e \notin A$, אז היא מוסיפה 0 לסכום, וזה בוודאי פחות מ- $x_e^* \cdot c(e)$.
ב. לפי הטענה לעיל, $z_e \leq f \cdot x_i^*$.
ג. הסכום זה בדיוק העלות של הפתרון האופטימלי השברי.
ד. עלות הפתרון השברי היא לכל היותר עלות הפתרון בשלמים (כי בשברי $x_e \leq 1$)

כנדרש.

עיגול אקראי של ה-*LP* הראשונית (*Randomized Rounding of the Primal LP*)

באלגוריתם הקודם, ביצענו עיגול דטרמיניסטי כלפי מעלה: לקחנו כל צלע שמקיימת $x_e^* \geq 1/f$, כלומר בעצם עבורן קבענו ש- $x_e^* \geq 1$.

נשתמש בערכים של הפיתרון האופטימלי בתור הסתברויות:

1. נמצא פתרון אופטימלי לבעיית ה-*LP*. נסמן את הפיתרון $(x_e^*)_{e \in \mathcal{E}}$.
2. לכל צלע $e \in \mathcal{E}$, נצרף אותה לכיסוי A בהסתברות x_e^*.

נראה ש-A מהווה כיסוי חוקי. זאת תהיה טענה הסתברותית.

למה: (בהוכחה משתמשים בקבוע אוילר, e. נסמן אותו $\bar{e}$ כדי לא להתבלבל עם הצלע e).

לכל $v \in V$, מתקיים:

$$\mathbb{P}[v \in \cup A] \geq 1 - \frac{1}{\bar{e}}$$

הוכחה: ניזכר ב- $C_v \coloneqq \{e \in \mathcal{E} : v \in e\}$. אזי:

$$\mathbb{P}[v \notin \cup A] = \mathbb{P}[no\ e \in C_v\ was\ chosen] =^{א} \prod_{e \in C_v} (1 - x_e^*) \leq^{ב} \prod_{e \in C_v} \exp(-x_e^*) =^{ג} \exp\left(-\sum_{e \in C_v} x_e^*\right) \leq \exp(-1) = \frac{1}{\bar{e}}$$

א. כי כל בחירה של צלע היא בת"ל.
ב. כרגיל, עבור $x \in (0,1]$ מתקיים $1 - x \leq \exp(-x)$.
ג. נהפוך מחפלה של חזקות לחזקה של סכום.
ד. ניזכר ש $\sum_{e \in C_v} x_e^* \geq 1$, לפי האילוץ של הבעיה.

מסקנה: ההסתברות ש-A לא מהווה כיסוי חוקי, לפי חסם איחוד:

$$\mathbb{P}[A\ is\ not\ a\ valid\ SC] \leq \sum_{v \in V} \mathbb{P}[v \notin \cup A] \leq \sum_{v \in V} \frac{1}{\bar{e}} = \frac{n}{\bar{e}}$$

נצטרך להגביר (*amplify*) את ההסתברות להצלחה. כרגיל עם אלגוריתמים הסתברותיים, נחזור על הפעולה הפשוטה הרבה פעמים.

הגברת הסתברות – *Probability Amplification*

עבור $k \in \mathbb{N}$, נריץ את האלגוריתם k פעמים בת"ל ונקבל $A_1, A_2, \ldots, A_k$ מועמדים. נגדיר את הפיתרון המאוחד, כל הצלעות שנבחרות:

$$\cup_k \coloneqq \bigcup_{i=1}^{k} A_i$$

ועכשיו לכל $v \in V$ מתקיים:

$$\mathbb{P}[v \notin \cup_k] \leq \exp(-k)$$

טענה: קיים קבוע $D > 0$ כך שאם $k \geq D \cdot \ln n$ (כלומר אם נבצע $\Omega(\ln n)$ הרצות), אז:

$$\mathbb{P}[\cup_k \ \ is\ not\ a\ valid\ SC] \leq 1/4$$

הוכחה: עבור $v \in V$ מתקיים:

$$\mathbb{P}[v \notin \cup_k] \leq \exp(-k) \leq \exp(-D \cdot \ln n) = \exp(\ln n^{-D}) = \exp\left(\ln \frac{1}{n^D}\right) = \frac{1}{n^D} \leq^{א} \frac{1}{4n}$$

א. אם D מספיק גדול.

אזי ההסתברות ש- $\cup_k$ לא מהווה כיסוי, לפי חסם איחוד היא:

$$\mathbb{P}[\cup_k \ \ is\ not\ a\ valid\ SC] \leq \sum_{v \in V} \mathbb{P}[v \notin \cup_k] \leq \frac{1}{4}$$

יחס הקירוב

בשביל ההוכחה, נשתמש באי"ש מרקוב. יהי X מ"מ ממשי אי-שלילי, ויהי $t \in \mathbb{R}^+$, אזי מתקיים:

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

למה: יהי OPT המחיר האופטימלי, ויהי k טבעי. אזי,

$$\mathbb{P}[c(\cup_k) \geq 4 \cdot k \cdot OPT] \leq 1/4$$

הוכחה: נמצא את התוחלת של העלות. בריצה אחת של האלגוריתם, עבור $e \in \mathcal{E}$, נגדיר:

$$X_e \coloneqq \begin{cases} 0, & e \notin A \\ c(e), & e \in A \end{cases}$$

אזי,

$$\mathbb{E}[X_e] = c(e) \cdot \mathbb{P}[X_e \neq 0] = c(e) \cdot x_e^*$$

ומכיוון ש: $c(A) = \sum_{e \in \mathcal{E}} X_e$, נוכל לרשום (לפי לינאריות התוחלת):

$$\mathbb{E}[c(A)] = \mathbb{E}\left[\sum_{e \in \mathcal{E}} X_e\right] = \sum_{e \in \mathcal{E}} \mathbb{E}[X_e] = \sum_{e \in \mathcal{E}} c(e) \cdot x_e^* = \mathrm{OPT}_f$$

כלומר,

$$\mathbb{E}[c(\cup_k)] \leq k \cdot \mathrm{OPT}_f$$

אז לפי אי"ש מרקוב:

$$\mathbb{P}[c(\cup_k) \geq 4 \cdot k \cdot OPT] \leq \frac{k \cdot \mathrm{OPT}_f}{4 \cdot k \cdot OPT} \leq \frac{k \cdot OPT}{4 \cdot k \cdot OPT} \leq \frac{1}{4}$$

מסקנה: אם נחזור על האלגוריתם $\Omega(\ln n)$ פעמים, נקבל כיסוי בעלות לכל היותר $O(\ln n \cdot OPT)$, בהסתברות לפחות חצי.

למה? כי עבור $k = \Omega(\ln n)$, מתקיים:

$$\mathbb{P}[\cup_k \ \ is\ a\ valid\ SC \quad \wedge \quad c(\cup_k) \leq 4 \cdot k \cdot OPT] \geq 1 - \mathbb{P}[c(\cup_k) > 4 \cdot k \cdot OPT] - \mathbb{P}[\cup_k \ \ is\ not\ a\ valid\ SC] \geq 1/2$$

אז יש לנו שני אלגוריתמים: ההסתברותי שראינו עכשיו, ודטרמיניסטי שנותן (בוודאות) קירוב $f \cdot OPT$. מה עדיף?

נשים לב ש-f (התדירות המקסימלית) ניתנת לחישוב בזמן לינארי עבור H נתון.

אז אם $f = o(\ln n)$, עדיף את הדטרמיניסטי. אם הדרגות גבוהות, עדיף את ההסתברותי.

הוכחת קירוב של האלגוריתם החמדן ע"י ה-*dual*

הבעיה הדואלית של גרסת ה-*LP* של *min-SC*: יהיה וקטור $y \geq 0$ שיש לו אינדקס לכל $v \in V$. אנחנו נרצה למקסם אותו, תחת ההגבלה:

$$\sum_{v\in e} y_v \leq c(e), \qquad \forall e \in \mathcal{E}$$

נשים לב שזו בעצם בעיית אריזה (*packing*). זה נפוץ מאוד שבעיות כיסוי ובעיות אריזה הן דואליות. לדוגמה *min-VC* ובעיית *max-match*.

ניזכר באלגוריתם החמדן:

1. נאתחל $A \coloneqq \emptyset$.
2. כל עוד $\bigcup A \neq V$:
   a. $e \coloneqq argmin\{\mathrm{eff}_{\bigcup A}(e) : e \in E(H)\}$.
   b. עבור כל קודקוד $v \in (e \setminus \bigcup A)$, (הקודקודים החדשים ש-*e* מכסה):
      i. נגדיר $ac(v) \coloneqq \mathrm{eff}_{\bigcup A}(e)$. (*ac = amortized cost*)
   c. $A \coloneqq A \cup \{e\}$.

טענה: יהי $n \in \mathbb{N}$. נסמן את הטור ההרמוני של *n*: $\mathcal{H}_n \coloneqq \sum_{i=1}^{n} 1/i$. נגדיר את הווקטור:

$$y \coloneqq \left(y_v \coloneqq \frac{\mathrm{ac}(v)}{\mathcal{H}_n}\right)_{v\in V}$$

כלומר, יש אינדקס עבור כל קודקוד, והערך בכל אחד הוא ה-*ac* חלקי הטור ההרמוני של *n*.

אזי, הווקטור הזה הוא פתרון ישים עבור הבעיה הדואלית.

הוכחה:

נסמן את הסדר שבו הקודקודים כוסו ע"י האלגוריתם החמדן: $\{v_1, v_2, \ldots, v_n\}$.

כבר הוכחנו שלכל $i \in [n]$. מתקיים:

$$\mathrm{ac}(v_i) \leq \frac{OPT}{n-i+1}$$

בהינתן $e \in \mathcal{E}$ בגודל *k*, נגדיר את הסדר שבו הקודקודים של *e* כוסו: $\{v_1, v_2, \ldots, v_k\}$.

(לא בהכרח אותו סדר שיש למעלה! לא אמרנו ש-*e* היא בחירה של החמדן).

נקבע קודקוד כלשהו ב-*e*: $v_i \in e = \{v_1, v_2, \ldots, v_k\}$. ונתבונן ברגע לפני שהוא מכוסה.

אם *A* הוא הפתרון החמדן הנוכחי, אז מספר הקודקודים ב-*e* שעוד לא כוסו הוא:

$$|e \setminus \bigcup A| = |\{v_i, v_{i+1}, \ldots, v_k\}| = |e| - i + 1 = k - i + 1$$

ננתח את ה-*amortized cost* של v_i. בגלל שאנחנו עומדים לכסות אותו, זה אומר שיש צלע כלשהי e' שהחמדן בחר לכסות, ו- $v_i \in e'$:

$$\mathrm{ac}(v_i) = \mathrm{eff}_{\bigcup A}(e') \leq^{\text{א}} \mathrm{eff}_{\bigcup A}(e) =^{\text{ב}} \frac{c(e)}{|e \setminus \bigcup A|} =^{\text{ג}} \frac{c(e)}{k-i+1}$$

א. כי בחרנו את הצלע שממזערת את *eff*.
ב. הגדרת *eff*.
ג. לפי השוויון למעלה.

בסה"כ:

$$\mathrm{ac}(v_i) \leq \frac{c(e)}{k-i+1}, \qquad |e| = k$$

עכשיו, אנחנו צריכים להראות ש-*y* הוא פתרון ישים לבעיה הדואלית.

ראשית, כל כניסה בווקטור גדולה מ-0 לפי ההגדרה של *y*. אז מתקיים $y \geq 0$, כנדרש.

צריך רק להוכיח שלכל $e \in \mathcal{E}$: $\sum_{v\in e} y_v \leq c(e)$.

נקבע $e\in\mathcal{E}$ כלשהי, ויהי $\{v_1,v_2,\dots,v_k\}$ הסדר שבו הקודקודים שלו כוסו ע"י החמדן. לפי הגדרת y, מתקיים:

$$\sum_{v\in e} y_v=\sum_{v\in e}\frac{\mathrm{ac}(v)}{\mathcal{H}_n}\overset{\text{א}}{\le}\sum_{i=1}^{k}\frac{c(e)}{\mathcal{H}_n(k-i+1)}\overset{\text{ב}}{=}\frac{c(e)}{\mathcal{H}_n}\sum_{i=1}^{k}\frac{1}{k-i+1}\overset{\text{ג}}{=}\frac{c(e)}{\mathcal{H}_n}\sum_{i=1}^{k}\frac{1}{i}\overset{\text{ד}}{=}\frac{c(e)}{\mathcal{H}_n}\cdot\mathcal{H}_k=\frac{\mathcal{H}_k}{\mathcal{H}_n}\cdot c(e)\overset{\text{ה}}{\le}c(e)$$

א. נסכום לפי סדר הכיסוי, עם הטענה הקודמת שהוכחנו.
ב. נוציא החוצה מה שלא תלוי ב-i.
ג. נשים לב שהסכום הוא פשוט מ-k עד 1.
ד. הסכום הזה הוא הטור ההרמוני של k.
ה. מתקיים $\mathcal{H}_k/\mathcal{H}_n\le 1$, כי $k\le n$.

כנדרש.

ניגש להוכחת הטענה המרכזית: שהאלגוריתם החמדן הוא $\mathcal{H}_n$ מקרב עבור *min-cost SC*. (הטור ההרמוני הוא $O(\ln n)$).

נשתמש בטענה שאומרת ש-y הוא פתרון ישים לבעיה הדואלית. ראשית, אפשר להעביר אגפים בהגדרת y ולקבל:

$$y_v\coloneqq\frac{\mathrm{ac}(v)}{\mathcal{H}_n}\quad\Longrightarrow\quad\mathcal{H}_n\cdot y_v=\mathrm{ac}(v)$$

מתקיים שהעלות של הפתרון החמדן הוא סכום ה-ac, אז:

$$cost\ of\ greedy=\sum_{v\in V}\mathrm{ac}(v)=\sum_{v\in V}\mathcal{H}_n\cdot y_v=\mathcal{H}_n\sum_{v\in V}y_v$$

ואמרנו ש-y הוא פתרון ישים. ניזכר במשפט הדואליות החלש:

$$\max\{c^Tx : Ax\le b\}\le\min\{y^Tb : A^Ty=c,\ \ y\ge 0\}$$

הפתרון המקסימלי של הדואלי (y) חסום בפתרון המינימלי של הראשוני. כלומר,

$$\mathcal{H}_n\sum_{v\in V}y_v\le\mathcal{H}_n\cdot\mathrm{OPT}_f\overset{\text{א}}{\le}\mathcal{H}_n\cdot OPT$$

א. כי האופטימלי של הראשוני (LP) חסום באופטימלי של ה-IP.

כנדרש.