67824 - אלגוריתמים מתקדם

מרצה: פרופ' יובל רבני

מסכם: מאור מזרחי

סמסטר א' תשפ"ב (2021-2022)

עדכון אחרון: 28 בפברואר 2022

להערות ותיקונים: maor.mizrachi@mail.huji.ac.il
למרצה אין אף קשר ו/או אחריות על הנכתב בסיכום זה.

תוכן העניינים

הקדמה

הרצאה 1

הקורס הוא קורס המשך לקורס "אלגוריתמים" בתואר ראשון. הנושאים בקורס יהיו מתקדמים יותר לכן צריך ללמוד לפני את קורס אלגוריתמים וקורס חישוביות.
הנושא הראשון שנעסוק בו יהיה תכנון לינארי ואופטימיזציה קמורה. בהמשך הקורס ייתכן שנדבר על בעיות זרימה ברשתות, אלגוריתמי קירוב ונושאים אחרים. הסילבוס לא קבוע ועשוי להשתנות בהתאם למה שמחליט המרצה.

- ציון הקורס יתבסס על 100% תרגילי בית.

  אין הבטחה למספר התרגילים, כנראה זה יהיה בערך פעם ב-3-4 שבועות ככה שזה יוצא 4-5 תרגילים.
- אין חובת נוכחות בכיתה.
- ההרצאות מוקלטות (מהשיעור הראשון זה נראה שלא ניתן לצפות בלייב).
- לעיתים יעלה חומר כתוב לאתר הקורס לפני השיעור.
- **ספר הקורס:** אין ספר מומלץ, אבל אפשר למצוא את החומר די בקלות באינטרנט.
- החל מההרצאה השניה הקורס נלמד באנגלית (הסיכום כפי שאתם רואים בעברית).

בקורס נציג בעיות ונפתור אותן בעזרת אלגוריתמים וגם נראה קשרים שונים בין הנושאים שאנחנו נלמד.

סימונים מוסכמים בסיכום

לאורך הסיכום השתמשתי בכמה סימונים מוסכמים שכדאי להכיר.
השתדלתי לאכוף ולהקפיד על הסימונים, אבל אם אתם מוצאים טעות/משהו לא ברור אני אשמח לשמוע על כך ולתקן.

וקטור עמודה	אותיות קטנות **דגושות** באנגלית $(\boldsymbol{a}, \boldsymbol{u})$
וקטור שורה	יצויין מראש מתי וקטור הוא וקטור שורה
השחלוף (transpose) של וקטור	$\boldsymbol{v}^T$
סקלר	אותיות קטנות באנגלית או ביוונית (a, b, α)
מטריצה	אות גדולה באנגלית (A, B, I)
מטריצת היחידה $n \times n$	I_n, לעיתים רק I (מימד לפי הקשר)
מרחב וקטורי העמודה ממימד n מעל $\mathbb{F}$	$\mathbb{F}^n$
מרחב המטריצות $m \times n$ מעל $\mathbb{F}$	$\mathbb{F}^{m\times n}$
מכפלה פנימית סטנדרטית (סקלרית)	$\boldsymbol{u}^T\boldsymbol{v}$ או $\langle \boldsymbol{u},\ \boldsymbol{v}\rangle$
נורמת p	$\|\|\cdot\|\|_p$
$\mathbb{R}_{\geq 0}$ או $\mathbb{R}_+$	$[0, \infty)$

חלק I

תכנון לינארי

1 מבוא

הגדרה 1.1. **פונקציה לינארית** מעל $\mathbb{R}^n$ היא פונקציה מהצורה $f(\boldsymbol{x}) = \boldsymbol{c}^T\boldsymbol{x} - b$ כאשר $\boldsymbol{c} \in \mathbb{R}^n$ ו-$b \in \mathbb{R}$ (הכפל הוא כפל מטריצות).

הגדרה 1.2. בהינתן f לינארית:

- $f(\boldsymbol{x}) = 0$ נקראת **משוואה לינארית**.
- $f(\boldsymbol{x}) \geq 0$ נקרא **אי שוויון לינארי**.

הערה 1.3. מספר הערות לגבי ההגדרות:

- שימו לב שמבחינה טכנית $\boldsymbol{c}^T\boldsymbol{x}$ הוא "מטריצה חד-מימדית", אבל אנחנו נתייחס לערך של הביטוי בתור מספר.
- נשים לב ש-$\langle \boldsymbol{c}, \boldsymbol{x} \rangle = \sum_{i=1}^{n} c_i x_i = \boldsymbol{c}^T\boldsymbol{x}$ (כאשר c_i ו-x_i הן הכניסות של $\boldsymbol{c}$ ו-$\boldsymbol{x}$ בהתאמה ולפי סדר).
- הביטוי $f(\boldsymbol{x}) = 0$ שקול אל $\boldsymbol{c}^T\boldsymbol{x} = b$, ו-$f(\boldsymbol{x}) \geq 0$ אל $\boldsymbol{c}^T\boldsymbol{x} \geq b$.

הגדרה 1.4. בהינתן מטריצה $A \in \mathbb{R}^{m\times n}$ ו-$\boldsymbol{b} \in \mathbb{R}^m$:

- **מערכת משוואות לינארית** (מעל $\mathbb{R}^n$) היא מהצורה $A\boldsymbol{x} = \boldsymbol{b}$.
- **מערכת אי שוויונים לינארית** (מעל $\mathbb{R}^n$) היא מהצורה $A\boldsymbol{x} \geq \boldsymbol{b}$, כאשר זה סימון מקובל אל $\boldsymbol{A_i}\boldsymbol{x} \geq b_i$ (כאשר $\{\boldsymbol{A_i}\}_{1\leq i\leq m}$ שורות A).

הגדרה 1.5. ב**בעיית תכנון לינארי** רוצים למצוא $\boldsymbol{x} \in \mathbb{R}^n$ הפותר את הבעיה
$$\begin{aligned} &\min_{\boldsymbol{x}} \ \boldsymbol{c}^T\boldsymbol{x} \\ &\text{subject to } A\boldsymbol{x} \geq \boldsymbol{b} \end{aligned}$$
- כאשר $\boldsymbol{c} \in \mathbb{R}^n$, $\boldsymbol{b} \in \mathbb{R}^m$ ו-$A \in \mathbb{R}^{m\times n}$.

הפונקציה $\boldsymbol{x} \mapsto \boldsymbol{c}^T\boldsymbol{x}$ נקראת **פונקציית המטרה** (כי היא המדד להשגת המינימום), ו-$A\boldsymbol{x} \geq \boldsymbol{b}$ **האילוצים**.

ההצגה הנ"ל נקראת **ההצגה הקנונית** (The Canonical Form)

הערה.

- נשים לב שהפונקציה שרוצים למזער, $\boldsymbol{c}^T\boldsymbol{x}$, היא לא פונקציה לינארית כללית (היה לנו $-b$ שם) - הוקטור שממזער את $\boldsymbol{c}^T\boldsymbol{x}$ זהה לזה שממזער את $\boldsymbol{c}^T\boldsymbol{x} + \text{constant}$ לכל קבוע constant, ולכן הקבוע מיותר.
- דרך אי השוויונים אפשר גם להגדיר אילוצי שוויון. למשל את האילוץ $7x = 2$ אפשר לרשום בתור $\begin{aligned} 7x &\geq 2 \\ -7x &\geq -2 \end{aligned}$.
- **מינימזציה היא גם מקסימיזציה:** אם היינו רוצים למקסם את $\boldsymbol{c}^T\boldsymbol{x}$, אפשר למזער את $(-\boldsymbol{c})^T\boldsymbol{x}$.
- בעיית התכנון הלינארי ניתנת לכתיבה מתמטית בתור $\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{argmin}} \left\{\boldsymbol{c}^T\boldsymbol{x} \mid A\boldsymbol{x} \geq \boldsymbol{b}\right\}$.

אם לא היו לנו אילוצים, במקרה בו $\boldsymbol{c} \neq \boldsymbol{0}$, היה אפשר להשיג ערך קטן כרצוננו (אם יש כניסה שלילית ב-$\boldsymbol{c}$ אפשר להגדיל את אותה הכניסה ב-$\boldsymbol{x}$, ואם חיובית אז להקטין), ואז לבעיה לא היה פתרון. במקרה זה אומרים ש**התוכנית לא חסומה**.

במקרה שאין אילוצים וגם $\boldsymbol{c}=\boldsymbol{0}$, יש ערך מינימום והוא 0 (זה הערך היחיד של $\boldsymbol{c}^T\boldsymbol{x}$).

באופן כללי ייתכן שהאילוצים סותרים זה את זה, למשל האילוצים $\substack{x\geq 4\\ -x\geq -3}$ (לא ייתכן ש-$x\geq 4$ וגם $x\leq 3$), ואז אומרים ש**התוכנית לא פיזיבילית** (infeasible).

כשאנחנו מדברים על תוכנית מחשב, המספרים צריכים להיות רציונליים: $A\in\mathbb{Q}^{m\times n}$, $\boldsymbol{b}\in\mathbb{Q}^m$ ו-$\boldsymbol{c}\in\mathbb{Q}^n$. נרצה לקבל כפלט וקטור $\boldsymbol{x}\in\mathbb{Q}^n$.
אפילו שהפונקציה והאילוצים כולם לינאריים תחת הרציונליים, לא ברור למה הממזער של הפונקציה תחת האילוצים הוא רציונלי. לבעיה הבאה למשל, **שאינה בעיית תכנון לינארי**, אין פתרון (שימו לב שזה כי הדרישה היא פתרון ב-$\mathbb{Q}$, אפילו שיש פתרון מעל $\mathbb{R}$): $\begin{array}{ll}\min\limits_{x\in\mathbb{Q}} & x^2-2\\ \text{s.t.} & x^2-2\geq 0\end{array}$.
בבעיות תכנון לינארי ברציונליים תמיד יהיה פתרון רציונלי.

נוסף על כך שימו לב שהקלט לתוכנית הוא מספרים בהצגתם הבינארית - זה ישפיע לנו על מושג היעילות בהמשך.

1.1 הצגה סטנדרטית

הגדרה 1.6. תוכנית לינארית ב**הצגה סטנדרטית** (The Standard Form) היא מהצורה
$$\begin{array}{ll}\text{minimize} & \boldsymbol{c}^T\boldsymbol{x}\\ \text{subject to} & \begin{cases}A\boldsymbol{x}=\boldsymbol{b}\\ \boldsymbol{x}\geq\boldsymbol{0}\end{cases}\end{array}.$$

ההצגה הסטנדרטית שקולה להצגה הקנונית - **כל תוכנית לינארית שאפשר לכתוב בהצגה הקנונית, אפשר בסטנדרטית, ולהיפך**.
אפילו שבהצגה הסטנדרטית אנחנו דורשים אילוצי שוויון, $A\boldsymbol{x}=\boldsymbol{b}$, השקילות היא בזכות אילוץ הסימן $\boldsymbol{x}\geq\boldsymbol{0}$ - אפשר להיעזר בו להמרת אילוצי אי השוויון.

הרצאה 2

טענה 1.7. כל תוכנית בהצגה הקנונית אפשר להציג בצורה הסטנדרטית.

הוכחה. תהא התוכנית $\begin{array}{ll}\min & \boldsymbol{c}^T\boldsymbol{x}\\ \text{subject to} & A\boldsymbol{x}\geq\boldsymbol{b}\end{array}$.
כדי להוכיח שאפשר לכתוב אותה בצורה הסטנדרטית, נשים לב שכל מספר, בפרט שלילי, אפשר לרשום כהפרש שני מספרים אי שליליים. כך, אנחנו נסמן $x_i=x_i^+-x_i^-$, כאשר $x_i^+,x_i^-\geq 0$. בצורה הזו, בה נחליף כל משתנה בהפרש של שניים חיוביים, אנחנו נוכל לקיים את האילוץ $\boldsymbol{x}\geq\boldsymbol{0}$ בתוכנית הסטנדרטית.
האתגר הבא שלנו יהיה להפוך את אילוצי אי השוויון שלנו לאילוצי שוויון.
נסמן את השורה ה-i של A בתור $\boldsymbol{A_i}$, כוקטור שורה. האילוץ ה-i שלנו הוא $\boldsymbol{A_i}\boldsymbol{x}\geq b_i$ (יש פה כפל של וקטור השורה $\boldsymbol{A_i}$ בוקטור העמודה $\boldsymbol{x}$, b_i הוא סקלר).
את אי השוויון הנ"ל אפשר לכתוב גם בתור $(-\boldsymbol{A_i})\,\boldsymbol{x}\leq -b_i$. אגף שמאל קטן מאגף ימין - אז אפשר להוסיף לו איזשהו מספר חיובי $s_i\geq 0$ ככה ששני האגפים יהיו שווים: $(-\boldsymbol{A_i})\,\boldsymbol{x}+s_i=-b_i$.
כך, המספר s_i יהווה למעשה משתנה נוסף בתוכנית הלינארית החדשה (לו אין משקל בפונקציית המינימיזציה כמובן).
שימו לב ש-s_i הוא מספר לכל אילוץ; לכל אילוץ יש s_i משלו. **בסך הכל אנחנו נוסיף** m s_i**ים.**
בעזרת שתי הטרנספורמציות האלה (הפיכת ה-xים להפרש של שניים, והוספת ה-sים), נצטרך כמובן לשנות את $A,\boldsymbol{b},\boldsymbol{c}$ בהתאם ואז לקבל תוכנית חדשה מהצורה $\begin{array}{ll}\text{minimize} & \boldsymbol{c}^T\boldsymbol{x}\\ \text{subject to} & \begin{cases}A\boldsymbol{x}=\boldsymbol{b}\\ \boldsymbol{x}\geq\boldsymbol{0}\end{cases}\end{array}$, שזו הצורה הסטנדרטית. □

הערה 1.8. נבין איך נראית התוכנית החדשה פחות יותר.

נפצל את הוקטור $\boldsymbol{x}$ לוקטור $\boldsymbol{x}^+$ים וקטור $\boldsymbol{x}^-$ים: בהתאמה כל משתני ה-x^+ים וה-x^-ים שאנחנו מוסיפים.
כך למשל $\min c^T x$ הופך אל $\min \boldsymbol{c}^T\left(\boldsymbol{x}^+ - \boldsymbol{x}^-\right)$, ו-$A\boldsymbol{x} \geq \boldsymbol{b}$ הופך תחילה אל $A\boldsymbol{x}^+ - A\boldsymbol{x}^- \geq \boldsymbol{b}$, ואז אחרי הוספת ה-$s$-ים (נסמן ב-$\boldsymbol{s}$ את הוקטור שלהם), הוא הופך אל $-A\boldsymbol{x}^+ + A\boldsymbol{x}^- + \boldsymbol{s} = -\boldsymbol{b}$ (שימו לב לסימנים שהתהפכו פה, ודאו שאתם מבינים מדוע מההוכחה). בנוסף $\boldsymbol{x}^+, \boldsymbol{x}^-, \boldsymbol{s}$ כולם וקטורים אשר גדולים או שווים 0, כפי שהסברנו בפתרון.

כך, התוכנית השקולה היא למעשה משהו כמו:

$$\begin{aligned} \min_{\boldsymbol{x}^+, \boldsymbol{x}^- \in \mathbb{R}^n \text{ and } \boldsymbol{s} \in \mathbb{R}^m} \quad & \left(\boldsymbol{c}^T\boldsymbol{x}^+ - \boldsymbol{c}^T\boldsymbol{x}^-\right) \\ \text{subject to} \quad & -A\boldsymbol{x}^+ + A\boldsymbol{x}^- + \boldsymbol{s} = -\boldsymbol{b} \\ \text{and} \quad & \boldsymbol{x}^+, \boldsymbol{x}^-, \boldsymbol{s} \geq \boldsymbol{0} \end{aligned} \cdot$$

1.2 הלמה של פרקש

לבנתיים אנחנו נסתכל על תוכניות בצורה הסטנדרטית שלהן.

נניח שנתונה לנו תוכנית בצורה הסטנדרטית

$$\begin{aligned} &\text{minimize } \boldsymbol{c}^T\boldsymbol{x} \\ &\text{subject to } \begin{cases} A\boldsymbol{x} = \boldsymbol{b} \\ \boldsymbol{x} \geq \boldsymbol{0} \end{cases} \end{aligned} .$$

איך יודעים שיש בכלל $\boldsymbol{0} \leq \boldsymbol{x} \in \mathbb{R}^n$ אשר $A\boldsymbol{x} = \boldsymbol{b}$?

דוגמה. אם נבחר $A = I_n$, $\boldsymbol{b} = \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix}$, **אין** $\boldsymbol{0} \leq \boldsymbol{x} \in \mathbb{R}^n$ אשר $A\boldsymbol{x} = \boldsymbol{b}$.

מאלגברה לינארית אנחנו מכירים שאלה דומה: "האם קיים $\boldsymbol{x}$ אשר $A\boldsymbol{x} = \boldsymbol{b}$?" (כמובן ש-$A$ לא בהכרח ריבועית). שימו לב שהיא לא זהה לשאלה המקורית שלנו, שכן פה נדרש ש-$\boldsymbol{x} \geq \boldsymbol{0}$.
מתמטית, התשובה היא שאכן יש פתרון אם"ם $\boldsymbol{b} \in \operatorname{span}\left(\left\{\boldsymbol{A^1}, \ldots, \boldsymbol{A^n}\right\}\right)$, כלומר אם"ם $\boldsymbol{b}$ הוא צירוף לינארי של **העמודות** של A. במקרה שכזה, $\boldsymbol{x}$ הוא המקדמים של הצירוף הלינארי הזה.

למה 1.9 (הלמה של פרקש)**.** נתונים $A \in \mathbb{R}^{m \times n}$ ו-$\boldsymbol{b} \in \mathbb{R}^m$. אזי <u>בדיוק אחד</u> מהבאים מתקיים:

1. קיים $\boldsymbol{x} \in \mathbb{R}^n$ כך ש-$A\boldsymbol{x} = \boldsymbol{b}$ ו-$\boldsymbol{x} \geq \boldsymbol{0}$ (למערכת יש פתרון).
2. קיים $\boldsymbol{y} \in \mathbb{R}^m$ כך ש-$\boldsymbol{y}^T A \geq \boldsymbol{0}$ וגם $\boldsymbol{y}^T\boldsymbol{b} < 0$ (שימו לב ש-$\boldsymbol{y}^T A$ הוא קומבינציה **של השורות** של A לפי המקדמים של $\boldsymbol{y}$).

הוכחה. להגיד שבדיוק אחד מתקיים אומר שלכל הפחות אחד מתקיים, ולכל היותר אחד מתקיים.
<u>לכל היותר אחד מתקיים (לא ייתכן ששניהם מתקיימים בו זמנית):</u> נניח בשלילה שכן, ואז ניקח את ה-$\boldsymbol{0} \leq \boldsymbol{x}$ מהסעיף הראשון, וה-$\boldsymbol{y}$ מהסעיף השני שמקיים $\boldsymbol{y}^T A \geq \boldsymbol{0}$, נכפיל מטריציונית אגף באגף ונקבל $\boldsymbol{y}^T A\boldsymbol{x} \geq 0$. מהסעיף הראשון $A\boldsymbol{x} = \boldsymbol{b}$ לכן נקבל $\boldsymbol{y}^T\boldsymbol{b} = \boldsymbol{y}^T A\boldsymbol{x} \geq 0$, אבל מהסעיף השני $\boldsymbol{y}^T\boldsymbol{b} < 0$, סתירה.
<u>לפחות אחד מתקיים:</u> אנחנו נוכיח שבמקרה שבו <u>אין</u> $\boldsymbol{x}$ אשר $A\boldsymbol{x} = \boldsymbol{b}$ ו-$\boldsymbol{x} \geq \boldsymbol{0}$, אז יש $\boldsymbol{y}$ עבורו $\boldsymbol{y}^T A \geq \boldsymbol{0}$ ו-$\boldsymbol{y}^T\boldsymbol{b} = -1$ (למעשה הטענה דורשת ש-$\boldsymbol{y}^T\boldsymbol{b}$ יהיה קטן מ-0, אנחנו נראה שזה אפילו יוצא -1).
נגדיר את $C = \{A\boldsymbol{x} \mid \boldsymbol{0} \leq \boldsymbol{x} \in \mathbb{R}^n\} \subseteq \mathbb{R}^m$. הקבוצה הנ"ל היא "**חרוט**" (cone), ההגדרה היא אוסף צירופים לינאריים עם מקדמים אי שליליים - נרשום אותה אחר כך) ב-$\mathbb{R}^m$. נשים לב ש-$\boldsymbol{b} \notin C$ שכן אחרת ($\boldsymbol{x} \in C$) היה $\boldsymbol{0} \leq \boldsymbol{x}$ עבורו $A\boldsymbol{x} = \boldsymbol{b}$, ואנחנו לא מטפלים במקרה הזה.

אינטואיציה

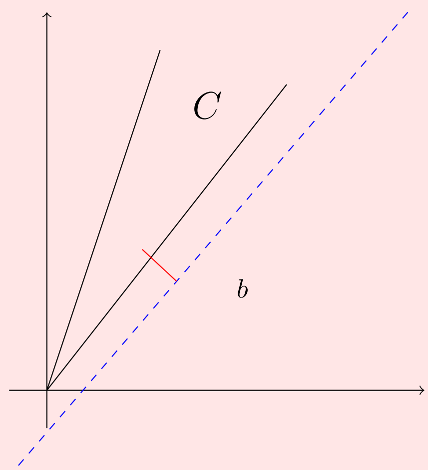
נציג אינטואיציה דו מימדית למה שאנחנו רוצים לעשות.

בציור מופיע החרוט שלנו הוא C (הוא אינסופי, שימו לב), ונזכור ש-$\boldsymbol{b}$ מחוצה לו. נוסף על כך החרוט C הוא קבוצה קמורה.

אנחנו נראה שיש על-מישור מפריד בין הקבוצה C והנקודה $\boldsymbol{b}$. במקרה הזה ($\mathbb{R}^2$), זה אומר שיש ישר ביניהם (מסורטט בכחול).

ניקח וקטור שמאונך לוקטור היוצר את הישר הנ"ל (מסורטט באדום). כעת בהינתן הוקטור האדום הנ"ל שנקרא לו $\boldsymbol{y}$, המכפלה הפנימית שלו עם כל איבר ב-C היא חיובית (כי C מעל הישר), והמכפלה הפנימית עם $\boldsymbol{b}$, שלילית.

על כן, $\boldsymbol{y}$ יהיה הוקטור שאנחנו נבחר.



ניקח כדור סגור B מסביב ל-$\boldsymbol{b}$, מרדיוס $||\boldsymbol{b}||_2$ (נורמת 2, האורך של $\boldsymbol{b}$). מתקיים ש-$B\cap C\neq\emptyset$, שכן $\mathbf{0}\in B$ וגם $\mathbf{0}\in C$. הכדור הסגור B הוא קבוצה סגורה וחסומה. הקבוצה C היא סגורה, אבל לא חסומה. מכאן, הקבוצה $C\cap B$ היא קבוצה חסומה (כי היא מוכלת ב-B) וגם סגורה (כחיתוך של 2 קבוצות סגורות). על כן, היא קבוצה **קומפקטית**.

נתבונן בפונקציה הרציפה המוגדרת על ידי $f(\boldsymbol{x})=||\boldsymbol{b}-\boldsymbol{x}||_2$, לפי משפט באינפי היא חייבת להשיג מינימום בקבוצה הקומפקטית $B\cap C$[1]. נסמן את הנקודה שבה מתקבל המינימום ב-$\boldsymbol{v}$. בפרט משום ש-$\mathbf{0}\in B\cap C$ מתקיים $f(\boldsymbol{v})\leq f(\mathbf{0})$, כלומר $||\boldsymbol{b}-\boldsymbol{v}||_2\leq||\boldsymbol{b}||_2$. לכל $\boldsymbol{u}\in C\setminus B$ מתקיים $||\boldsymbol{b}-\boldsymbol{u}||_2\geq||\boldsymbol{b}||_2$ (אחרת $||\boldsymbol{b}-\boldsymbol{u}||_2<||\boldsymbol{b}||_2$ ואז הנקודה $\boldsymbol{u}$ הייתה בכדור B).

מפה, הנקודה $\boldsymbol{v}$ היא הנקודה הקרובה ביותר ל-b ב-C (כל נקודה אחרת ב-C שאינה ב-B, נמצאת במרחק גדול מ-$\boldsymbol{b}$, ומבין הנקודות שב-C וגם ב-B, המרחק של $\boldsymbol{v}$ מ-$\boldsymbol{b}$ מינימלי מהגדרה).

- נראה שלכל $\boldsymbol{u}\in C$ $\langle\boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}-\boldsymbol{u}\rangle\geq 0$:

  ניקח כל $\boldsymbol{u}\in C$. הקטע בין $\boldsymbol{u}$ ו-$\boldsymbol{v}$[2], כולו נמצא ב-C, וזה מפני ש-C קמורה (זו ההגדרה של קמירות). כלומר, לכל $\alpha\in[0,1]$, מתקיים $\alpha\boldsymbol{u}+(1-\alpha)\boldsymbol{v}\in C$.

  ראינו ש-$||\boldsymbol{b}-\boldsymbol{v}||^2\leq||\boldsymbol{b}-(\alpha\boldsymbol{u}+(1-\alpha)\boldsymbol{v})||^2$ (שכן $(\alpha\boldsymbol{u}+(1-\alpha)\boldsymbol{v})$ נקודה ב-C ואמרנו ש-$\boldsymbol{v}$ היא הקרובה ביותר ל-$\boldsymbol{b}$ מבין הנקודות ב-C).

  אבל $\boldsymbol{b}-(\alpha\boldsymbol{u}+(1-\alpha)\boldsymbol{v})=(\boldsymbol{b}-\boldsymbol{v})+\alpha(\boldsymbol{v}-\boldsymbol{u})$.
  נקבל ש-$||\boldsymbol{b}-(\alpha\boldsymbol{u}+(1-\alpha)\boldsymbol{v})||^2=||(\boldsymbol{b}-\boldsymbol{v})+\alpha(\boldsymbol{v}-\boldsymbol{u})||^2$.

  אבל $||(\boldsymbol{b}-\boldsymbol{v})+\alpha(\boldsymbol{v}-\boldsymbol{u})||^2=\langle\boldsymbol{b}-\boldsymbol{v}+\alpha(\boldsymbol{v}-\boldsymbol{u}),\ (\boldsymbol{b}-\boldsymbol{v})+\alpha(\boldsymbol{v}-\boldsymbol{u})\rangle$ (מהגדרת נורמה בריבוע),
  נשתמש בלינאריות המכפלה הפנימית ונקבל שזה $||\boldsymbol{b}-\boldsymbol{v}||^2+2\alpha\langle\boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}-\boldsymbol{u}\rangle+\alpha^2||\boldsymbol{v}-\boldsymbol{u}||^2$,
  קרי $||\boldsymbol{b}-\boldsymbol{v}||^2+\alpha\left(2\langle\boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}-\boldsymbol{u}\rangle+\alpha||\boldsymbol{v}-\boldsymbol{u}||^2\right)$.

  אז ראינו שמתקיים $||\boldsymbol{b}-(\alpha\boldsymbol{u}+(1-\alpha)\boldsymbol{v})||^2=||\boldsymbol{b}-\boldsymbol{v}||^2+\alpha\left(2\langle\boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}-\boldsymbol{u}\rangle+\alpha||\boldsymbol{v}-\boldsymbol{u}||^2\right)$.

[1] למי שלא מכיר, דומה למשפט ויירשטראס בפונקציות חד מימדיות: "בקבוצה קומפקטית כל פונקציה רציפה מקבלת מינימום ומקסימום".
[2] אשר מוגדר להיות $\{\alpha\boldsymbol{u}+(1-\alpha)\boldsymbol{v}\mid\alpha\in[0,1]\}$.

לכן מתקיים $\left|\left|\boldsymbol{b}-\left(\alpha\boldsymbol{u}+\left(1-\alpha\right)\boldsymbol{v}\right)\right|\right|^{2}-\left|\left|\boldsymbol{b}-\boldsymbol{v}\right|\right|^{2}\geq0$, כלומר $\alpha\left(2\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}-\boldsymbol{u}\right\rangle +\alpha\left|\left|\boldsymbol{v}-\boldsymbol{u}\right|\right|^{2}\right)\geq0$,
כלומר $2\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}-\boldsymbol{u}\right\rangle +\alpha\left|\left|\boldsymbol{v}-\boldsymbol{u}\right|\right|^{2}\geq0$ (כי $\alpha\geq0$).

נשים לב שמתקיים $\boldsymbol{0}=\lim_{\alpha\to\boldsymbol{0}}\boldsymbol{0}\leq\lim_{\alpha\to\boldsymbol{0}}2\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}-\boldsymbol{u}\right\rangle +\alpha\left|\left|\boldsymbol{v}-\boldsymbol{u}\right|\right|^{2}=2\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}-\boldsymbol{u}\right\rangle$.
ראינו שלכל $\boldsymbol{u}\in C$, מתקיים $\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}-\boldsymbol{u}\right\rangle \geq0$.

- נראה ש-$\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}\right\rangle \geq0$: כעת נציב ב- $\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}-\boldsymbol{u}\right\rangle \geq0$ את $\boldsymbol{u}=\boldsymbol{0}$ ($\boldsymbol{0}\in C$) ונקבל $\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}\right\rangle \geq0$.
- נראה $\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{b}\right\rangle >\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}\right\rangle$:

  מלינאריות מתקיים $\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{b}\right\rangle =\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{b}-\boldsymbol{v}+\boldsymbol{v}\right\rangle =\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{b}-\boldsymbol{v}\right\rangle +\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}\right\rangle$.

  מתקיים $\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{b}-\boldsymbol{v}\right\rangle >0$ (זו נורמה בריבוע של הוקטור $\boldsymbol{b}-\boldsymbol{v}$ ששונה מאפס), ולכן $\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{b}\right\rangle >\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}\right\rangle$.
- ראינו שלכל $\boldsymbol{u}\in C$, מתקיים $\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}-\boldsymbol{u}\right\rangle \geq0$, כלומר (מלינאריות) $\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}\right\rangle \geq\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{u}\right\rangle$ לכל $\boldsymbol{u}\in C$.

  מסקנה: לכל $\boldsymbol{u}\in C$ מתקיים $\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{u}\right\rangle \leq\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{v}\right\rangle <\left\langle \boldsymbol{b}-\boldsymbol{v},\ \boldsymbol{b}\right\rangle$. על כן יש איזשהו $\gamma>0$, ככה שלכל $\boldsymbol{u}\in C$, מתקיים $\left\langle \boldsymbol{u},\ \boldsymbol{b}-\boldsymbol{v}\right\rangle \leq\left\langle \boldsymbol{v},\ \boldsymbol{b}-\boldsymbol{v}\right\rangle <\gamma<\left\langle \boldsymbol{b},\ \boldsymbol{b}-\boldsymbol{v}\right\rangle$

  נבחר $\boldsymbol{u}\in C$. לכל $\lambda\geq0$ מתקיים ש-$\lambda\boldsymbol{u}\in C$ מהגדרת cone. מפה, $\left\langle \lambda\boldsymbol{u},\ \boldsymbol{b}-\boldsymbol{v}\right\rangle <\gamma$. בפרט לכל $\lambda>0$ (גדול ממש מ-0) מלינאריות מכפלה פנימית מתקיים $\left\langle \boldsymbol{u},\ \boldsymbol{b}-\boldsymbol{v}\right\rangle <\frac{\gamma}{\lambda}$.
  כאשר $\lambda\to\infty$ נקבל $\lim_{\lambda\to\infty}\left\langle \boldsymbol{u},\ \boldsymbol{b}-\boldsymbol{v}\right\rangle \leq\lim_{\lambda\to\infty}\frac{\gamma}{\lambda}=0$, כלומר לכל $\boldsymbol{u}\in C$ מתקיים $\left\langle \boldsymbol{u},\ \boldsymbol{b}-\boldsymbol{v}\right\rangle \leq0$.
  שוב מלינאריות זה שקול אל $\left\langle \boldsymbol{u},\ \boldsymbol{v}-\boldsymbol{b}\right\rangle \geq0$ (מכפילים ב-(-1)).
- נבחר את הוקטור $\boldsymbol{y}=\frac{1}{\left\langle \boldsymbol{b},\ \boldsymbol{b}-\boldsymbol{v}\right\rangle }\cdot\left(\boldsymbol{v}-\boldsymbol{b}\right)$. מתקיימים הבאים:
  - $\boldsymbol{y}^{T}\boldsymbol{b}=\left\langle \boldsymbol{y},\ \boldsymbol{b}\right\rangle =\frac{1}{\left\langle \boldsymbol{b},\ \boldsymbol{b}-\boldsymbol{v}\right\rangle }\left\langle \boldsymbol{b},\ \boldsymbol{v}-\boldsymbol{b}\right\rangle =-1\cdot\frac{1}{\left\langle \boldsymbol{b},\ \boldsymbol{b}-\boldsymbol{v}\right\rangle }\cdot\left\langle \boldsymbol{b},\ \boldsymbol{b}-\boldsymbol{v}\right\rangle =-1$.
  - $\boldsymbol{y}^{T}\boldsymbol{A_{j}}=\frac{1}{\left\langle \boldsymbol{b},\ \boldsymbol{b}-\boldsymbol{v}\right\rangle }\left\langle \boldsymbol{A_{j}},\ \boldsymbol{v}-\boldsymbol{b}\right\rangle \geq0$ כאשר $\boldsymbol{A_{j}}$ היא וקטור (העמודה) המתאים ל**שורה** ה-j-ית של A, והיא למעשה וקטור ב-C (השתמשנו בזה ש-$\left\langle \boldsymbol{b},\ \boldsymbol{b}-\boldsymbol{v}\right\rangle >0$ ו-$\left\langle \boldsymbol{A_{j}},\ \boldsymbol{v}-\boldsymbol{b}\right\rangle \geq0$ כי $\boldsymbol{A_{j}}\in C$).

כלומר מצאנו וקטור שעונה על דרישותינו, כנדרש.

הערת המסכם

היינו צריכים את הכדור B רק כדי לטעון שיש מינימום לפונקציית המרחק בהתחלה (קומפקטיות).

$\square$

2 הגיאומטריה של תוכניות לינאריות

הגדרה 2.1. תת קבוצה $P \subseteq \mathbb{R}^n$ נקראת $\mathcal{H}$-**פאון** או $\mathcal{H}$-**פוליהדרון** ($\mathcal{H}$-Polyhedron וברבים $\mathcal{H}$-Polyhedra) אם ורק אם אפשר לרשום:

$$P = P\left(A, \boldsymbol{b}\right) = \left\{\boldsymbol{x} \in \mathbb{R}^n \mid A\boldsymbol{x} \geq \boldsymbol{b}\right\}$$

עבור איזשהם $A \in \mathbb{R}^{m \times n}$ ו-$\boldsymbol{b} \in \mathbb{R}^m$.

הגדרה 2.2. $C \subseteq \mathbb{R}^n$ הוא $\mathcal{H}$-**חרוט** ($\mathcal{H}$-cone) אם הוא $\mathcal{H}$-פוליהדרון מהצורה $C = P\left(A, \boldsymbol{0}\right)$.

הערה. הקבוצה הפיזיבילית של תוכנית לינארית (כל מי שמקיים את האילוצים) היא $\mathcal{H}$-פוליהדרון.
אילוץ אחד, כלומר כזה מהצורה $\boldsymbol{A_i}\boldsymbol{x} \geq b_i$ ($\boldsymbol{A_i}$ זו שורה), הוא **חצי מרחב** (Half Space). לכן, $\mathcal{H}$ פוליהדרון, הוא חיתוך של חצאי מרחבים.

הערה. $\mathcal{H}$-**חרוט סגור למכפלה בסקלר:** בהינתן $\boldsymbol{x} \in C$ ו-$0 \leq \lambda \in \mathbb{R}$, מתקיים $\lambda\boldsymbol{x} \in C$, כי $A\left(\lambda\boldsymbol{x}\right) \geq \boldsymbol{0}$ (מפני ש-$A\boldsymbol{x} \geq \boldsymbol{0}$).

הרצאה 3

הגדרה 2.3. תהא $S \subseteq \mathbb{R}^n$. ה**קָמוֹר** של S (Convex Hull) או **הסגור הקמור של** S הוא:

$$\operatorname{conv}\left(S\right) = \left\{\boldsymbol{x} \in \mathbb{R}^n \mid \exists\lambda : S \to \mathbb{R}_+ \text{ s.t. } \underset{\boldsymbol{y} \in S}{\exists}\lambda\left(\boldsymbol{y}\right) = 1 \ \text{ and } \boldsymbol{x} = \sum_{\boldsymbol{y} \in S}\lambda\left(\boldsymbol{y}\right)\boldsymbol{y}\right\}$$

במילים אחרות, כל הוקטורים במרחב שאפשר לכתוב אותם כסכום של מקדמים כפול וקטורים בקבוצה S, כך שהמקדמים מסתכמים ל-1.

הגדרה 2.4. תהא $S \subseteq \mathbb{R}^n$. ה**חרוט שנוצר על ידי** S או **הסגור החרוטי של** S (Conic Hull) הוא:

$$\operatorname{cone}\left(S\right) = \left\{\boldsymbol{x} \in \mathbb{R}^n \mid \exists\lambda : S \to \mathbb{R}_+ \text{ s.t. } \boldsymbol{x} = \sum_{\boldsymbol{y} \in S}\lambda\left(\boldsymbol{y}\right)\boldsymbol{y}\right\}$$

כלומר כמעט כמו הקמור של S, רק שהמקדמים לא חייבים להיסכם ל-1. כך למעשה מקבלים שחרוט הוא צורה אינסופית (בהינתן $\boldsymbol{x} \in \mathbb{R}^n$ לכל $0 \leq \lambda \in \mathbb{R}$ מתקיים $\lambda\boldsymbol{x} \in C$ - אפשר להאריך וקטורים כרצוננו).

הגדרה 2.5. $P \subseteq \mathbb{R}^n$ היא $\mathcal{V}$-**פאון** ($\mathcal{V}$-Polyhedron וברבים $\mathcal{V}$-Polyhedra) אם קיימות <u>**קבוצות סופיות**</u> $V, Y \subseteq \mathbb{R}^n$ ככה ש-$P = \operatorname{conv}\left(V\right) + \operatorname{cone}\left(Y\right)$ כאשר $+$ הוא סכום מנקובסקי (כלומר, סכום בין קבוצות, כל הסכומים של $x + y$ כך ש-x בקבוצה הראשונה ו-y בשניה).

הגדרה 2.6. $\mathcal{V}$-פאון $P \subseteq \mathbb{R}^n$ הוא $\mathcal{V}$-**פוליטופ**, או $\mathcal{V}$-**פאון חסום** ($\mathcal{V}$-Polytope) אם $P = \operatorname{conv}\left(V\right)$. כלומר אם Y בהגדרה של $\mathcal{V}$-פאון היא $Y = \emptyset$.

הגדרה 2.7. $\mathcal{V}$-פאון $P \subseteq \mathbb{R}^n$ הוא $\mathcal{V}$-**חרוט** ($\mathcal{V}$-Cone) אם $P = \operatorname{cone}\left(Y\right)$. כלומר אם V בהגדרה של $\mathcal{V}$-פאון הוא $V = \emptyset$.

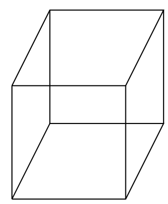
2.1 המשפט היסודי עבור פאונים

משפט 2.8 (המשפט היסודי עבור פאונים). $P \subseteq \mathbb{R}^n$ היא $\mathcal{V}$-פאון אם ורק אם P היא $\mathcal{H}$-פאון.

משפט דומה תקף גם לגבי $\mathcal{H}$-חרוט ו-$\mathcal{V}$-חרוט.
במילים אחרות, המשפט אומר שאפשר להסתכל על פאונים ב-2 דרכים: או כחיתוך של חצאי מרחב (שיטת $\mathcal{H}$), או כסכום של קמור של קבוצה וחרוט שנוצר ע"י קבוצה (שיטת $\mathcal{V}$).

מעתה אנחנו נכתוב "פאון" במקום $\mathcal{H}$-פאון או $\mathcal{V}$-פאון, כי זה אותו דבר, וגם "פוליטופ"/"פאון חסום" במקום $\mathcal{V}$-פאון חסום.

דוגמה 2.9. נסתכל על הקוביה התלת מימדית:



הקוביה היא פאון חסום (פוליטופ) ב-$\mathbb{R}^3$. אפשר לקבל את הקוביה ב-2 דרכים שקולות גיאומטרית (והשקילות נובעת מהמשפט היסודי):

$\mathcal{V}$: הקמור (convex hull) של שמונת הקודקודים של הקוביה.

$\mathcal{H}$: כל אחת מ-6 פאות (צדדי) הקוביה מגדירות חצאי מרחבים אינסופיים ("כל מה שנמצא משמאל/מימין למישור הפאה") והקוביה היא חיתוך של חצאי המרחב האלה.

אפשר גם להכליל. כדי להגדיר את הקוביה ה-n-מימדית אפשר:

$\mathcal{V}$: להגדיר אותה כקמור של 2^n וקטורים (כמספר הקודקודים).

$\mathcal{H}$: או להגדיר אותה כחיתוך של $2n$ אי שוויונות לינאריים (כמספר הפאות, א"ש לכל פאה).

2.2 פאות וקודקודים

כעת נגדיר פורמלית מהי פאה ומהו קודקוד (זה אמנם ברור כשהמרחב שלנו תלת מימדי, אבל במרחבים ממימד גבוה יותר זה עלול קצת להסתבך).

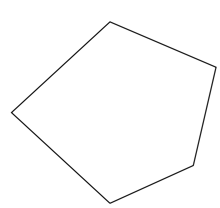
הגדרה 2.10. וקטורים $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ..., \boldsymbol{x}^{(k)} \in \mathbb{R}^n$ ייקראו **בלתי תלויים אפינית** אם הפתרון היחידי ($\boldsymbol{\alpha}$) ל- $\sum_{i=1}^{k} \alpha_i \boldsymbol{x}^{(i)} = 0$ וגם $\sum_{i=1}^{k} \alpha_i = 0$ הוא $\boldsymbol{\alpha} = \mathbf{0}$.

הערה 2.11. זה מקרה חלש יותר מבלתי תלות לינארית בגלל שפה אנחנו דורשים שהשוויון $\sum_{i=1}^{k} \alpha_i \boldsymbol{x}^{(i)} = 0$ יגרור $\boldsymbol{\alpha} = \mathbf{0}$ <u>רק במקרה ש-$\sum_{i=1}^{k} \alpha_i = 0$</u>. כלומר כדי שקבוצה תיקרא "בלתי תלויה אפינית" היא צריכה לקיים את "הבלתי התלות הלינארית" (אותה אנחנו מכירים) <u>**רק עבור**</u> מקדמים שנסכמים ל-0, לא תמיד.

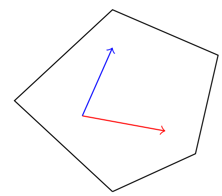
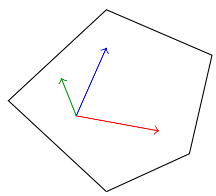
הערה 2.12. נשים לב ש-$\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ..., \boldsymbol{x}^{(k)} \in \mathbb{R}^n$ הם ב"ת אפינית, זה אם ורק אם $\boldsymbol{x}^{(2)} - \boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(1)}$ הם ב"ת לינארית. כלומר אפשר להזיז את כל הוקטורים לכיוון $\boldsymbol{x}^{(1)}$ ואז לקבל בלתי תלות לינארית רגילה.

הגדרה 2.13. בהינתן פאון P, נגדיר את **מימד הפאון** $\dim(P)$ להיות המספר d, אם המספר המקסימלי של נקודות בלתי תלויות אפינית ב-P הוא $d+1$.

דוגמה 2.14. נתון הפוליטופ (פאון חסום) הבא:



ניקח 3 נקודות **ספציפיות** בתוך הפאון, $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}$. אם היינו מזיזים את הוקטורים לכיוון הנקודה $\boldsymbol{a}$, נשים לב שהוקטורים $\boldsymbol{b}-\boldsymbol{a},\ \boldsymbol{c}-\boldsymbol{a}$ היו בלתי תלויים לינארית (כלומר אילו $\boldsymbol{a}$ היה הראשית אז $\boldsymbol{b},\boldsymbol{c}$ היו ב"ת לינארית). אז נקבל בסך הכל ש-$\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}$ הם בלתי תלויים אפינית.
אבל, אם נבחר **כל** 4 נקודות בתוך הפאון, $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{d}$ - נקבל שהזזה ב-$\boldsymbol{a}$ גוררת ש-$\boldsymbol{b}-\boldsymbol{a},\ \boldsymbol{c}-\boldsymbol{a},\ \boldsymbol{d}-\boldsymbol{a}$ הם **כן תלויים לינארית** (בתור 3 וקטורים במימד 2) ולכן $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{d}$ הם תלויים אפינית.
על כן, המספר המקסימלי של נקודות ב"ת אפינית ב-P הוא 3, ו-3 הוא $2+1$, ולכן $\dim(P)$ הוא 2.

אפשר לקבוע ראשית צירים חדשה בתוך הפוליטופ, כמו גם 2 נקודות נוספות, ולקבל 2 וקטורים בת"ל	**אי אפשר** לקבוע <u>אף</u> ראשית חדשה ו-3 נקודות ולקבל 3 וקטורים בת"ל
	

הגדרה 2.15. אי שוויון לינארי $\boldsymbol{c}^T\boldsymbol{x}\leq b_0$ נקרא **תקף** (valid) **עבור פאון** P אם כל $x\in P$ מקיים אותו.

הגדרה 2.16. אם $\boldsymbol{c}^T\boldsymbol{x}\leq b_0$ תקף עבור P, אז הקבוצה $F_{\boldsymbol{c},b_0}=P\cap\left\{\boldsymbol{x}\in\mathbb{R}^n \mid \boldsymbol{c}^T\boldsymbol{x}=b_0\right\}$ (הנקודות בהן יש ממש שוויון) נקראת **שפה** (face) של P.
אם $F_{\boldsymbol{c},b_0}\neq\emptyset$ הוא נקרא **שפה תומכת** (supporting face) ואם $F\neq P$ הוא נקרא **שפה ממש של** P (strict face).

הערה. כמה הערות:

- כמו שרשום בהגדרה ייתכן שהשפה של P (עבור $\boldsymbol{c},b_0$ כלשהם) הוא הקבוצה הריקה, אם אין כאלה ב-P שמקיימים את השוויון הנ"ל.
- אי השוויון $\boldsymbol{0}^T\boldsymbol{x}\leq 0$ תקף לכל פאון P, ולמעשה לכל נקודה ב-P מתקיים שוויון אז במקרה הזה $F_{\boldsymbol{0},0}=P$: כל P הוא השפה של P.

לכל פאון חסום (פוליטופ) P, השפה של P (לכל $\boldsymbol{c},b_0$) הוא פוליטופ בעצמו. לכן יש לו מימד. זה מביא אותנו להגדרה הבאה:

הגדרה 2.17. בהינתן פאון P ואי שוויון לינארי תקף ב-P, נסמן $F=F_{\boldsymbol{c},b_0}$. נגיד ש:

- F היא **קודקוד** (vertex וברבים vertices) אם $\dim F=0$ (בגדול, נגיד שהאיבר היחיד ב-F הוא קודקוד, במקום שהקבוצה היא קודקוד).
- F היא **פאה** (facet) אם $\dim F=\dim P-1$.

הגדרה 2.18. נסמן ב-$\operatorname{vert}(P)$ את קבוצת כל הקודקודים של P.

מעתה והלאה נדבר על פוליטופים, כלומר על פאונים חסומים.

הרצאה 4

משפט 2.19. תהא הקבוצה $P\subseteq\mathbb{R}^n$ פוליטופ ($P=\operatorname{conv}(V)$). אזי:

1. $P=\operatorname{conv}(\operatorname{vert}(P))$.
2. $\operatorname{vert}(P)\subseteq V$.

המשפט לעיל אומר שכדי לתאר את P, שהיא קמור של קבוצה V, מספיק לתאר אותה כקמור של **הקודקודים של** P. בנוסף, שהקבוצה V הנ"ל, צריכה להכיל את כל הקודקודים של P.

הוכחה. 1. ידוע ש-$P=\operatorname{conv}(V)$ לאיזושהי קבוצה סופית $V\subseteq\mathbb{R}^n$. אפשר להניח בלי הגבלת הכלליות שכל $\boldsymbol{v}\in V$ הוא **"הכרחי"**, כלומר בלעדיו ה-conv הוא לא אותו דבר. זה בלי להגביל את הכלליות כי פשוט אפשר לזרוק את $\boldsymbol{v}$ במצב כזה. מתמטית, זה אומר להניח שמתקיים ש-$\boldsymbol{v}\notin\operatorname{conv}(V\setminus\{\boldsymbol{v}\})$.

אנחנו נוכיח ש-$v\in\operatorname{vert}(P)$, ולכן נקבל ש-$P=\operatorname{conv}(V)=\operatorname{conv}(\operatorname{vert}(P))$.

למעשה כדי להראות ש-v הוא קודקוד עלינו להראות שיש איזשהם $\boldsymbol{c},b_0$ ככה ש-

נרשום $V=\left\{\boldsymbol{v},\boldsymbol{u}^{(1)},...,\boldsymbol{u}^{(m)}\right\}$. הנחנו ש-$\operatorname{conv}\left(\left\{\boldsymbol{u}^{(1)},...,\boldsymbol{u}^{(m)}\right\}\right)\notin\boldsymbol{v}$. במילים אחרות, <u>**לא קיימים**</u> $\alpha_1,....,\alpha_m\geq 0$ אשר $\sum_{i=1}^m\alpha_i=1$ ו-$\boldsymbol{v}=\sum_{i=1}^m\alpha_i\boldsymbol{u}^{(i)}$.

נשים לב שאנחנו מתארים פה מערכת משוואות שאין לה פתרון, אשר המשתנים שלה הם $\alpha_1,...,\alpha_m$. המטריצה שמייצגת את מערכת המשוואות הנ"ל היא $A=\begin{bmatrix}1&1&\cdots&1\\ |&|&&|\\ \boldsymbol{u}^{(1)}&\boldsymbol{u}^{(2)}&&\boldsymbol{u}^{(m)}\\ |&|&&|\end{bmatrix}\in\mathbb{R}^{(n+1)\times m}$. כך מערכת המשוואות היא $A\boldsymbol{\alpha}=\begin{pmatrix}1\\ |\\ \boldsymbol{v}\\ |\end{pmatrix}$, ו-$\boldsymbol{\alpha}\geq 0$. **זו תוכנית ליניארית בצורתה הסטנדרטית.**

אז למערכת הזו אין פתרון, ולכן לפי הלמה של פרקש יש $\boldsymbol{c}\in\mathbb{R}^{n+1}$, ככה ש-$\boldsymbol{c}^TA\geq\boldsymbol{0}$ וגם $\boldsymbol{c}^T\begin{pmatrix}1\\ |\\ \boldsymbol{v}\\ |\end{pmatrix}<0$.

נגדיר $\beta,\boldsymbol{b}$ כך: $\boldsymbol{c}=\begin{pmatrix}\beta\\ |\\ -\boldsymbol{b}\\ |\end{pmatrix}$. אז $\boldsymbol{c}^TA=\left(\beta-\boldsymbol{b}^T\boldsymbol{u}^{(1)},...,\beta-\boldsymbol{b}^T\boldsymbol{u}^{(m)}\right)$. ידוע כי $\boldsymbol{c}^TA\geq\boldsymbol{0}$ ואז נקבל שלכל $1\leq i\leq m$ מתקיים $\boldsymbol{b}^T\boldsymbol{u}^{(i)}\leq\beta$.

מנגד, נסתכל על $\boldsymbol{c}^T\begin{pmatrix}1\\ |\\ \boldsymbol{v}\\ |\end{pmatrix}<0$, כלומר $\begin{pmatrix}\beta & - & (-\boldsymbol{b}) & -\end{pmatrix}\begin{pmatrix}1\\ |\\ \boldsymbol{v}\\ |\end{pmatrix}<0$, שנותן לנו $\begin{matrix}\beta<0\\ (-\boldsymbol{b})^T\boldsymbol{v}<0\end{matrix}$.

כעת $(-\boldsymbol{b})^T\boldsymbol{v}<0$ ו-$\boldsymbol{b}^T\boldsymbol{u}^{(i)}\leq\beta$ יתנו לנו את $\boldsymbol{b}^T\boldsymbol{u}^{(i)}-\boldsymbol{b}^T\boldsymbol{v}\leq\beta$, ו-$\beta<0$ לכן $\boldsymbol{b}^T\boldsymbol{u}^{(i)}-\boldsymbol{b}^T\boldsymbol{v}<0$.

כעת נסתכל על אי השוויון $\boldsymbol{b}^T\boldsymbol{x}\leq\boldsymbol{b}^T\boldsymbol{v}$. אם $\boldsymbol{x}=\boldsymbol{v}$ מבון שיש פה שוויון, ואם $\boldsymbol{x}=\boldsymbol{u}^{(i)}$ ראינו שורה מעל שיש אי שוויון ממש: $\boldsymbol{b}^T\boldsymbol{u}^{(i)}<\boldsymbol{b}^T\boldsymbol{v}$.

אז כל $\boldsymbol{x}\in\left\{\boldsymbol{v},\boldsymbol{u}^{(1)},...,\boldsymbol{u}^{(m)}\right\}$ מקיים את אי השוויון. מה לגבי צירופים ליניאריים שלהם (שסכומם 1)?

בהינתן $\boldsymbol{x}=\lambda_0\boldsymbol{v}+\sum_{i=1}^m\lambda_i\boldsymbol{u}^{(i)}$ ו-$\lambda_0+\sum_{i=1}^m\lambda_i=1$ (כולם אי שליליים), כלומר בהינתן $\boldsymbol{x}\in P$,

יתקיים $\boldsymbol{b}^T\boldsymbol{x}=\boldsymbol{b}^T\left(\lambda_0\boldsymbol{v}+\sum_{i=1}^m\lambda_i\boldsymbol{u}^{(i)}\right)=\lambda_0\boldsymbol{b}^T\boldsymbol{v}+\sum_{i=1}^m\lambda_i\boldsymbol{b}^T\boldsymbol{u}^{(i)}\leq\underbrace{\left(\lambda_0+\sum_{i=1}^m\lambda_i\right)}_{1}\boldsymbol{b}^T\boldsymbol{v}=\boldsymbol{b}^T\boldsymbol{v}$.

בנוסף, שימו לב שכאשר $\sum_{i=1}^m\lambda_i>0$, אי השוויון הוא חזק:

$$\lambda_0\boldsymbol{b}^T\boldsymbol{v}+\sum_{i=1}^m\lambda_i\boldsymbol{b}^T\boldsymbol{u}^{(i)}<\underbrace{\left(\lambda_0+\sum_{i=1}^m\lambda_i\right)}_{1}\boldsymbol{b}^T\boldsymbol{v}=\boldsymbol{b}^T\boldsymbol{v}\quad(\text{כי } \boldsymbol{b}^T\boldsymbol{u}^{(i)}<\boldsymbol{b}^T\boldsymbol{v})$$

אז הראנו ש-$\boldsymbol{b}^T\boldsymbol{x} \le \boldsymbol{b}^T\boldsymbol{v}$ הוא אי שוויון תקף עבור $P = \operatorname{conv}(V) = \operatorname{conv}\left(\left\{\boldsymbol{v}, \boldsymbol{u}^{(1)}, ..., \boldsymbol{u}^{(m)}\right\}\right)$, כי הוא מתקיים לכל הנקודות ב-P. בנוסף, יש שוויון ממש רק ב-$\boldsymbol{v}$, כלומר השפה F היא רק $\boldsymbol{v}$, כלומר $\dim F = 0$, ומפה $\boldsymbol{v}$ הוא קודקוד של P.

2. להוכחת החלק השני של הטענה נניח בשלילה ש-$\boldsymbol{v}$ הוא קודקוד של P ולא ב-V. נסמן $V = \left\{\boldsymbol{u}^{(1)}, ..., \boldsymbol{u}^{(m)}\right\}$. בגלל ש-$\boldsymbol{v}$ הוא קודקוד, מהגדרה יש אי שוויון תקף ב-P, $\boldsymbol{c}^T\boldsymbol{x} \le \boldsymbol{b}$, ששפתו היא רק $\boldsymbol{v}$, כלומר $\left\{\boldsymbol{b} \in P \mid \boldsymbol{c}^T\boldsymbol{x} = \boldsymbol{b}\right\} = \{\boldsymbol{v}\}$. בפרט $\boldsymbol{c}^T\boldsymbol{u}^{(i)} < \boldsymbol{b}$ לכל $1 \le i \le m$.

הנחנו כי $\boldsymbol{v} \in P$ כלומר יש מקדמים ככה ש-$\boldsymbol{v} = \sum_{i=1}^{m} \lambda_i \boldsymbol{u}^{(i)}$ עבור $\boldsymbol{\lambda} \ge \boldsymbol{0}$, ו-$\sum_{i=1}^{m} \lambda_i = 1$.

נקבל $\boldsymbol{b} = \sum_{i=1}^{m} \lambda_i \boldsymbol{b} \cdot \boldsymbol{b} = \sum_{i=1}^{m} \lambda_i \boldsymbol{b} > \sum_{i=1}^{m} \lambda_i \boldsymbol{c}^T\boldsymbol{u}^{(i)} = \boldsymbol{c}^T\boldsymbol{v}$ - זו סתירה ($\boldsymbol{c}^T\boldsymbol{v} = \boldsymbol{b}$ - שווה ולא קטן ממש).

כלומר, $\operatorname{vert}(P) \subseteq V$ כמתבקש. $\square$

הגדרה 2.20. $P \subseteq \mathbb{R}^n$ פוליהדרון. $\boldsymbol{x} \in P$ נקרא **נקודה קיצונית של** P (extremal point) אם ורק אם לא קיים $\boldsymbol{0} \neq \boldsymbol{y} \in \mathbb{R}^n$ ככה ש-$\boldsymbol{x} + \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y} \in P$.

הפירוש הגיאומטרי הוא שלא משנה איזה כיוון $\boldsymbol{y}$ נבחר מהנקודה $\boldsymbol{x}$, אם נלך בו **או** לכיוון הנגדי שלו, נצא מהפאון.

מסקנה 2.21 (בשיעורי הבית). $\boldsymbol{x}$ היא נקודה קיצונית אם ורק אם היא קודקוד.

3 קבוצת הפתרונות

משפט 3.1. יהא $P=\{\boldsymbol{x}\in\mathbb{R}^n \mid A\boldsymbol{x}=\boldsymbol{b} \ \wedge \ \boldsymbol{x}\geq\boldsymbol{0}\}$ פוליהדרון. ויהא $\boldsymbol{c}\in\mathbb{R}^n$. נניח כי יש מינימום לתוכנית הלינארית $\begin{array}{c}\text{minimize } \boldsymbol{c}^T\boldsymbol{x}\\ \text{subject to } \boldsymbol{x}\in P\end{array}$ (במקרה זה אומרים ש**התוכנית חסומה**, אפילו שלא בהכרח מלמעלה). אז לכל $\boldsymbol{x}\in P$ יש $\boldsymbol{v}\in\text{vert}(P)$ ככה ש-$\boldsymbol{c}^T\boldsymbol{v}\leq\boldsymbol{c}^T\boldsymbol{x}$.

מסקנה 3.2. נניח שיש פתרון. נבחר את ה-$\boldsymbol{x}^*$ אשר הוא המינימום, יש קודקוד $\boldsymbol{v}$ אשר $\boldsymbol{c}^T\boldsymbol{v}\leq\boldsymbol{c}^T\boldsymbol{x}^*$, אבל $\boldsymbol{c}^T\boldsymbol{x}^*$ מינימלי ו-$v\in P$, **ולכן נסיק שכשיש פתרון אופטימלי - תמיד גם יש פתרון אופטימלי שהוא קודקוד.**

המשפט הנ"ל חשוב מאוד לפתירה של תוכניות לינאריות ואנחנו נראה את זה.

הוכחה למשפט 3.1. יהא $\boldsymbol{x}\in P$. אם $\boldsymbol{x}$ הוא קודקוד אז אין מה להוכיח (הוא מקיים את אי השוויון $\boldsymbol{c}^T\boldsymbol{x}\leq\boldsymbol{c}^T\boldsymbol{x}$).
כעת למקרה בו $\boldsymbol{x}$ איננו קודקוד. אנחנו יודעים ממסקנה 2.21 שיש $\boldsymbol{y}\in\mathbb{R}^n\setminus\{\boldsymbol{0}\}$ ככה ש-$\boldsymbol{x}+\boldsymbol{y}$ וגם $\boldsymbol{x}-\boldsymbol{y}$ ב-P. על כן מאפיון התוכנית הלינארית מתקיים $A(\boldsymbol{x}+\boldsymbol{y})=\boldsymbol{b}$, וגם $A\boldsymbol{x}=\boldsymbol{b}$, ואז נקבל $A\boldsymbol{y}=\boldsymbol{0}$.
מתקיים גם ש-$\boldsymbol{x}-\boldsymbol{y},\boldsymbol{x}+\boldsymbol{y}\geq\boldsymbol{0}$, וזה גורר שאם הקורדינטה ה-j של $\boldsymbol{x}$, x_j, היא 0, אז $y_j=0$ (כי $\underbrace{x_j}_{0}+y_j\geq 0$ וגם $\underbrace{x_j}_{0}-y_j\geq 0$).

אפשר להניח בלי הגבלת הכלליות ש-$\boldsymbol{c}^T\boldsymbol{y}\leq 0$ (אחרת במקום $\boldsymbol{y}$ נסמן מחדש $\boldsymbol{y}\leftarrow-\boldsymbol{y}$, הוא גם יקיים אותן תכונות שתיארנו לעיל).
נשים לב שאם $\boldsymbol{c}^T\boldsymbol{y}=0$ אז בלי הגבלת הכלליות, יש j כלשהו ככה ש-$y_j<0$. זאת משום ששוב, אם זה לא נכון, אפשר לסמן $\boldsymbol{y}\leftarrow-\boldsymbol{y}$ ואז תהיה קורדינטה כזאת (יודעים בוודאות ש-$\boldsymbol{y}\neq\boldsymbol{0}$ אז יש קורדינטה שונה מ-0).

- **אם** $\boldsymbol{y}\geq\boldsymbol{0}$ אז בדיוק בגלל זה $\boldsymbol{c}^T\boldsymbol{y}<0$ (אם היה שווה 0, הייתה לנו קורדינטה קטנה ממש מ-0 ממה שהסברנו הרגע).

  בהינתן $\lambda>0$, $\boldsymbol{x}+\lambda\boldsymbol{y}\geq\boldsymbol{0}$ (כי $\boldsymbol{x},\boldsymbol{y}\geq\boldsymbol{0}$), וגם $A(\boldsymbol{x}+\lambda\boldsymbol{y})=A\boldsymbol{x}+\lambda\underbrace{A\boldsymbol{y}}_{0}=A\boldsymbol{x}=\boldsymbol{b}$. לכן $\boldsymbol{x}+\lambda\boldsymbol{y}\in P$.

  אבל $\boldsymbol{c}^T(\boldsymbol{x}+\lambda\boldsymbol{y})=\boldsymbol{c}^T\boldsymbol{x}+\underbrace{\lambda\boldsymbol{c}^T\boldsymbol{y}}_{<0}<\boldsymbol{c}^T\boldsymbol{x}$.

  אז קיבלנו סתירה: לכאורה $\boldsymbol{x}+\lambda\boldsymbol{y}$ היא נקודה ב-P (פשוט מקבעים איזשהו $\lambda>0$) - אבל הערך של פונקציית המטרה בה **קטן ממש** מהערך של פונקציית המטרה ב-x, שהנחנו שזה הערך המינימלי ב-P.

  נזכור ש-$\boldsymbol{c}^T\boldsymbol{y}<0$, ו-$\boldsymbol{c}^T\boldsymbol{x}$ - ושניהם קבועים. ככל שנגדיל את λ נקבל ערכים קטנים כרצוננו, כלומר נקטין כמה שנרצה את פונקציית המטרה - זה בסתירה לכך שהתוכנית חסומה.

- **מפה, לא ייתכן $\boldsymbol{y}\geq\boldsymbol{0}$.** כלומר אכן ישנו איזשהו j עבורו $y_j<0$. נגדיר $\lambda=\min\limits_{j\text{ s.t. }y_j<0}\left\{\frac{x_j}{-y_j}\right\}$ (המינימום מוגדר כי יש קורדינטה שבה $y_j<0$, ולא מחלקים ב-0 במכנה).

  מתקיים $\lambda\geq 0$ שכן $x_j\geq 0$ (כי $\boldsymbol{x}\geq\boldsymbol{0}$) ו-$(-y_j)>0$. נטען כי $\boldsymbol{x}+\lambda\boldsymbol{y}\in P$. לשם כך נראה:

  - <u>$A(\boldsymbol{x}+\lambda\boldsymbol{y})=\boldsymbol{b}$</u>: מתקיים $A(\boldsymbol{x}+\lambda\boldsymbol{y})=A\boldsymbol{x}+\lambda\underbrace{A\boldsymbol{y}}_{0}=A\boldsymbol{x}=\boldsymbol{b}$.
  - <u>נראה כי $\boldsymbol{x}+\lambda\boldsymbol{y}\geq\boldsymbol{0}$</u>: נזכור ש-$\boldsymbol{x}\geq\boldsymbol{0}$ (כי $\boldsymbol{x}\in P$). תהא קורדינטה j ועלינו להראות ש-$x_j+\lambda y_j\geq 0$.
    מקודם אמרנו שאם $x_j=0$, אז בהכרח $y_j=0$ ומפה אי השוויון מתקיים: $0\geq 0$.
    נניח כי $x_j>0$. במקרה בו $y_j\geq 0$ מובן כי $x_j+\lambda y_j\geq 0$. במקרה שבו $y_j<0$, אז אנחנו יודעים להגיד:
    $$x_j+\lambda y_j=x_i+\min_{j\text{ s.t. }y_j<0}\left\{\frac{x_j}{-y_j}\right\}y_j\geq x_i+\frac{x_i}{-y_i}y_i=0\geq 0$$
    כאשר אי השוויון הוא לכיוון $\geq$ כי $y_j<0$ (אז הוא הופך את הסימן שלו, במקום $\leq$ שצריך להיות).
    בסך הכל אכן תמיד $\boldsymbol{x}+\lambda\boldsymbol{y}\geq\boldsymbol{0}$, כי תמיד הקורדינטות אי שליליות.

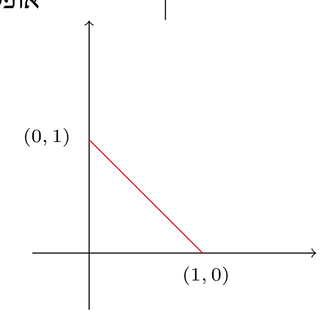
אז אכן מתקיימים התנאים לכך ש-$\boldsymbol{x}+\lambda\boldsymbol{y}\in P$.
הראנו שכל הקורדינטות גדולות-שוות 0, אבל עבור הקורדינטה j שהיא המינימום ב-λ מתקיים $x_j+\lambda y_j=0$.

למעשה מה שעשינו היה להגדיל את מספר הקורדינטות ב-$\boldsymbol{x}$ ששוות ממש ל-0 (כי אם קורדינטה ב-$\boldsymbol{x}$ הייתה שווה 0, הראנו שגם אותה קורדינטה ב-$\boldsymbol{y}$, ולכן גם הקורדינטה ב-$\boldsymbol{x}+\lambda\boldsymbol{y}$). נסמן מחדש $\boldsymbol{x}\leftarrow\boldsymbol{x}+\lambda\boldsymbol{y}$. אם הגענו במקרה לקודקוד - מצוין, סיימנו. אבל אם לא, אפשר לחזור על התהליך שעשינו שוב (שכן ההנחה היחידה על $\boldsymbol{x}$ הייתה שהוא לא קודקוד). ואז שוב, ושוב. כאמור, בכל פעם הוספנו קורדינטת אפס חדשה ל-$\boldsymbol{x}$ (ושמרנו על הקיימות).
כאמור, $\boldsymbol{x}$ מכיל n קורדינטות לכן ההליך יחזור על עצמו לכל היותר n פעמים (אם לא נעצור לפני, כשנגיע לקודקוד). בסוף נגיע אל $\boldsymbol{x}=\boldsymbol{0}$ ובמקרה שכזה $\boldsymbol{0}$ חייב להיות קודקוד (שכן אפשר ללכת לכיוון שלילי ולקבל וקטור שאיננו ב-P כי הוא לא חיובי).[3]

למה הקודקוד שאליו הגענו הוא פתרון מינימלי? נשים לב ש-$\boldsymbol{c}^T\boldsymbol{x}$ $\boldsymbol{c}^T(\boldsymbol{x}+\lambda\boldsymbol{y})=\boldsymbol{c}^T\boldsymbol{x}+\underbrace{\lambda\boldsymbol{c}^T\boldsymbol{y}}_{\leq 0}\leq\boldsymbol{c}^T\boldsymbol{x}$ כי $\boldsymbol{c}^T\boldsymbol{y}\leq 0$ (הנחנו את זה בהתחלה). כלומר כל איטרציה כשמשיגים $\boldsymbol{x}+\lambda\boldsymbol{y}$ חדש אנחנו מקטינים את הערך של פונקציית המטרה שלנו ביחס למה שהוא היה ב-$\boldsymbol{x}$, וממילא $\boldsymbol{c}^T\boldsymbol{x}$ מינימלי - ולכן אנחנו ממשיכים לשמור על המינימליות מאיטרציה לאיטרציה. $\square$

הערה .3.3 זה לא אומר שכל הפתרונות האופטימליים הם קודקודים, ולא אומר שכל הקודקודים הם פתרונות אופטימליים.

לא כל הקודקודים הם פתרונות אופטימליים	**לא כל הפתרונות האופטימליים הם קודקודים**
התוכנית הלינארית שלנו היא $\begin{aligned}&\min x_1\\ &\text{s.t. } x_1+x_2=1\\ &x_1,x_2\geq 0\end{aligned}$. $(0,1)$ הוא קודקוד, אבל לא פתרון אופטימלי. $(1,0)$ למשל כן.	התוכנית הלינארית שלנו היא $\begin{aligned}&\min x_1+x_2\\ &\text{s.t. } x_1+x_2=1\\ &x_1,x_2\geq 0\end{aligned}$. $(1,0)$ ו-$(0,1)$ קודקודים ופתרונות אופטימליים. אבל כל נקודה שמקיימת את האילוצים (יכולה גם לא להיות קודקוד אלא סתם על הישר) היא פתרון אופטימלי (האופטימום הוא 1).



משפט .3.4 תהא $P=\{\boldsymbol{x}\in\mathbb{R}^n\mid A\boldsymbol{x}=\boldsymbol{b}\text{ וגם }\boldsymbol{x}\geq\boldsymbol{0}\}$. עבור $\boldsymbol{x}\in P$ נסמן ב-$A_{\boldsymbol{x}}$ את המטריצה המתקבלת על ידי לקיחת העמודות של A המתאימות לכניסות ב-$\boldsymbol{x}$ שהן חיוביות ממש (למשל אם $x_2>0$, נכלול את העמודה השניה).
אזי, $\boldsymbol{x}$ הוא קודקוד אם ורק אם העמודות של $A_{\boldsymbol{x}}$ הן בלתי תלויות לינארית.

דוגמה .3.5 אם $A=\begin{bmatrix}3&0&-4&1&4&0\\-4&1&1&0&2&0\\5&0&-2&0&-1&1\end{bmatrix}$ ו-$\boldsymbol{b}=\begin{bmatrix}6\\1\\1\end{bmatrix}$. אם נגדיר את $\boldsymbol{x}=\begin{bmatrix}0\\1\\0\\6\\0\\1\end{bmatrix}$, אז $A_{\boldsymbol{x}}=$

[3] זה לא אומר ש-$\boldsymbol{0}$ הוא קודקוד של כל פוליהדרון! זה נכון אך ורק אם חזרנו על התהליך n פעמים ולא עצרנו עוד לפני. הפוליהדרון יכול בכלל לא להכיל את $\boldsymbol{0}$.

$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ - העמודות של $A_{\boldsymbol{x}}$ הן בלתי תלויות, ומפה $\boldsymbol{x}$ קודקוד.

לא נוכיח את המשפט הנ"ל, כן נשתמש בו. הרצאה 5

נניח ש-$P = \{\boldsymbol{x} \in \mathbb{R}^n \mid A\boldsymbol{x} = \boldsymbol{b} \text{ וגם } \boldsymbol{x} \geq \boldsymbol{0}\}$, ככה ש-$A \in \mathbb{R}^{m \times n}$. מובן ש-$\text{rank}\,(A) \leq m$.

אם $\text{rank}\,(A) < m$ יש איזשהו אילוץ שמתקבל מאילוצים אחרים (= שורות המטריצה תלויות לינארית) ואז:

- או שלמערכת אין פתרון (ייתכן שהשורה היא צירוף לינארי של שורות אחרות, אבל ה-b המתאים לה הוא לא אותו הצירוף של ה-bים).
- אם למערכת יש פתרון, אפשר לוותר על השורה הזו שתלויה לינארית בשורות האחרות.

  ככה נגיע למטריצה שהדרגה שלה היא בדיוק דרגת השורות - ו-P שלנו לא ישתנה (הורדת אילוצים מיותרים לא תשנה לנו את הפוליהדרון).

על כן אין בעיה להניח מעתה כי $\text{rank}\,(A) = m$.

נניח שנתון קודקוד $\boldsymbol{x}$ כלשהו. נגדיר $B = \{j \mid x_j > 0\}$ - קבוצת האינדקסים של $\boldsymbol{x}$ ככה שהקורדינטה באינדקס גדולה מ-0.

אם היינו לוקחים את A ומורידים את העמודות j כך ש-$x_j = 0$, אנחנו נישאר עם מטריצה בלתי תלויה לינארית (ממשפט 3.4).

מפה, מתקיים $|B| \leq \text{rank}\,(A) = m$ (שכן אחרת אם $|B| > m$, היינו מקבלים יותר מ-m עמודות בת"ל של A (העמודות של האינדקסים j ככה ש-$x_j > 0$) - אבל הדרגה של A היא m).

בהינתן מטריצה או וקטור, נסמן באינדקס תחתון $_{(B)}$ את העמודות המתאימות לאינדקסים ב-B (או לכל קבוצה אחרת שנבחר).

- <u>המקרה בו $|B| = m$</u>: עמודות $A_{(B)}$ הן בלתי תלויות לינארית, ו-$A_{(B)}$ היא מטריצה ריבועית $m \times m$ - מפה $A_{(B)}$ היא מטריצה הפיכה.
- <u>אם $|B| < m$</u>: אפשר להוסיף עמודות בלתי תלויות לינארית כלשהן ל-A, כדי ש-B תהיה מגודל m: $|B| = m$.
  ואז $\boldsymbol{x}_{(B)} = A_{(B)}^{-1}\boldsymbol{b}$ שלנו תכיל כניסות שהן 0 (ושאר הכניסות גדולות ממש מ-0).

נסמן $N = \{1, ..., n\} \setminus B$. נשים לב כי $A_{(B)}\boldsymbol{x}_{(\boldsymbol{B})} = A\boldsymbol{x} = \boldsymbol{b}$ (שכן כל כניסה אחרת בוקטור $\boldsymbol{x}$ היא 0, לכן לא תורמת כלום לסכום, ולכן הסכום יוצא בדיוק $A\boldsymbol{x}$: $A\boldsymbol{x} = A_{(B)}\boldsymbol{x}_{(\boldsymbol{B})} + A_{(N)}\underbrace{\boldsymbol{x}_{(\boldsymbol{N})}}_{\boldsymbol{0}} = A_{(B)}\boldsymbol{x}_{(\boldsymbol{B})}$).

על כן מתקיים $\boldsymbol{x}_{(\boldsymbol{B})} = \left(A_{(B)}\right)^{-1}\boldsymbol{b}$, נזכור כי $\boldsymbol{x}_{(\boldsymbol{B})} = A_{(B)}^{-1}\boldsymbol{b} > \boldsymbol{0}$ (שכן $\boldsymbol{x}_{(\boldsymbol{B})}$ הוא הוקטור $\boldsymbol{x}$ בקורדינטות החיוביות ממש שלו), ו-$\boldsymbol{x}_{(\boldsymbol{N})} = \boldsymbol{0}$.

נקרא ל-B **בסיס** ו-$\boldsymbol{x}$-ש שנתון על ידי הנוסחה $\boldsymbol{x}_{(\boldsymbol{B})} = A_{(B)}^{-1}\boldsymbol{b}$ ו-$\boldsymbol{x}_{(\boldsymbol{N})} = \boldsymbol{0}$ (בקורדינטות המתאימות) נקרא **פתרון אפשרי בסיסי** (basic feasible solution).

כמו שאמרנו, לא מתקיים תמיד ש-$\boldsymbol{x}_{(\boldsymbol{B})} > \boldsymbol{0}$ (אם הוא קודקוד, אז כן).

4 אלגוריתם הסימפלקס

שיטת הסימפלקס (Dantzig, 1947) היא אלגוריתם לפתירה של תוכניות לינאריות.

נתונה תוכנית לינארית בהצגתה הסטנדרטית: $\begin{cases} \min\limits_{\boldsymbol{x}} & \boldsymbol{c}^T\boldsymbol{x} \\ \text{subject to} & A\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0} \end{cases}$.

תזכורת

נזכיר את הסימונים שהשתמשנו בהם:

נתונה קבוצת אינדקסים $I \subseteq \{1, ..., n\}$. בהינתן וקטור $\boldsymbol{v}$ כלשהו, $\boldsymbol{v_{(I)}}$ הוא הוקטור שנוצר מ-$\boldsymbol{v}$ על ידי לקיחת כל הקורדינטות שמספרן ב-I (לפי הסדר). בהינתן מטריצה M כלשהי, $M_{(I)}$ היא המטריצה של לקיחת כל העמודות שמספרן ב-I.

בהינתן קבוצה B אשר $|B| = m$, (נגדיר $N = \{1, ..., n\} \setminus B$), שמקיימת ש-$A_{(B)}$ הפיכה ($m \times m$), וגם וקטור $\boldsymbol{x}$ שהוא פתרון לתוכנית לינארית, ככה ש-$\boldsymbol{x_{(B)}} \geq \boldsymbol{0}$ ו-$\boldsymbol{x_{(N)}} = \boldsymbol{0}$, קוראים ל-$B$ **בסיס** ול-$\boldsymbol{x}$ **פתרון אפשרי בסיסי**.

ראינו שמתקיים $A_{(B)}\boldsymbol{x_{(B)}} = A_{(B)}\boldsymbol{x_{(B)}} + A_{(N)}\underbrace{\boldsymbol{x_{(N)}}}_{\boldsymbol{0}} = A\boldsymbol{x} = \boldsymbol{b}$.

ראינו משפט (3.4) שאומר שאם $\boldsymbol{x_{(B)}} > \boldsymbol{0}$ (גדול ממש), אז $\boldsymbol{x}$ קודקוד.

4.1 הרעיון הכללי

נניח שמשמיים מישהו זורק לנו בסיס B ופתרון אפשרי בסיסי $\boldsymbol{x}$ אשר $\boldsymbol{x_{(B)}} = A_{(B)}^{-1}\boldsymbol{b}$ ו-$\boldsymbol{x_{(N)}} = \boldsymbol{0}$ (בהמשך נדון בשאלה איך משיגים בסיס B ופתרון אפשרי בסיסי $\boldsymbol{x}$). כמו שהזכרנו קודם, לא מחייב שכל הכניסות ב-$\boldsymbol{x_{(B)}}$ גדולות מ-0, אבל כן מה שבטוח הוא ש**כל הכניסות ב-**$\boldsymbol{x_{(N)}}$ שוות 0.

נזכור כי השוויון $A\boldsymbol{x} = \boldsymbol{b}$ במקרה שלנו הוא גם $A_{(B)}\boldsymbol{x_{(B)}} = \boldsymbol{b}$. אפשר להניח שמתקיים ש-$A_{(B)} = I_m$ (מטריצת היחידה $m \times m$) שכן אחרת אפשר לבצע הליך אלימינציית גאוס ל-$[A \mid b]$, ולקבל שהעמודות שהאינדקס שלהן ב-B, יוצרות את מטריצת היחידה: $A_{(B)} = I_m$. **(שימו לב שאלימינציית גאוס לא משנה את מרחב הפתרונות!)**

בנוסף ניתן להניח ש-$\boldsymbol{b} \geq \boldsymbol{0}$, אחרת אם יש כניסה של $\boldsymbol{b}$ שהיא שלילית, נכפיל את כל השורה המתאימה במטריצה A וגם את הכניסה המתאימה ב-$\boldsymbol{b}$ ב-(-1) - אנחנו עדיין נשמור על השוויון.

כעת ננסה להבין איך נראית התוכנית הלינארית בהינתן ש-x הוא פתרון (לא בהכרח אופטימלי) שלה. החקר הזה יאפשר לנו לקבוע "עד כמה טוב" x וגם "איך ניתן להתקדם ממנו".

המטרה הכללית הולכת להיות להוציא איבר מהבסיס, ולהחליף אותו באיבר אחר. כלומר נשנה את הבסיס, ולכן גם נשנה את הפתרון הבסיסי (ולמעשה, נזוז לקודקוד אחר).

את התוכנית הלינארית שלנו $\begin{cases} \min\limits_{\boldsymbol{x}} & \boldsymbol{c}^T\boldsymbol{x} \\ \text{subject to} & A\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0} \end{cases}$, אפשר לכתוב גם: (נפצל ל-B ו-N)

$$\begin{cases} \min\limits_{\boldsymbol{x}} & \boldsymbol{c_{(B)}}^T\boldsymbol{x_{(B)}} + \boldsymbol{c_{(N)}}^T\boldsymbol{x_{(N)}} \\ \text{subject to} & A_{(B)}\boldsymbol{x_{(B)}} + A_{(N)}\boldsymbol{x_{(N)}} = \boldsymbol{b} \\ & \boldsymbol{x_{(B)}}, \boldsymbol{x_{(N)}} \geq \boldsymbol{0} \end{cases}$$

נכון ש-$\boldsymbol{x_{(N)}}$ אפילו שווה ל-$\boldsymbol{0}$, ושאפשר להוריד חלק מהדברים: אנחנו בכוונה משאירים אותם.

$A_{(B)}\boldsymbol{x}_{(\boldsymbol{B})} + A_{(N)}\boldsymbol{x}_{(\boldsymbol{N})} = \boldsymbol{b}$ במעבר אגפים שקול אל $A_{(B)}\boldsymbol{x}_{(\boldsymbol{B})} = \boldsymbol{b} - A_{(N)}\boldsymbol{x}_{(\boldsymbol{N})}$, נכפיל בהופכי של $A_{(B)}$ (היא הפיכה והנחנו שאפילו מטריצת היחידה, אבל נגיע לזה תכף) ונקבל למעשה:

$$\boldsymbol{x}_{(\boldsymbol{B})} = A_{(B)}^{-1}\left(\boldsymbol{b} - A_{(N)}\boldsymbol{x}_{(\boldsymbol{N})}\right)$$

אז:

$$\begin{aligned}
{\boldsymbol{c}_{(\boldsymbol{B})}}^T\boldsymbol{x}_{(\boldsymbol{B})} + {\boldsymbol{c}_{(\boldsymbol{N})}}^T\boldsymbol{x}_{(\boldsymbol{N})} = \quad & {\boldsymbol{c}_{(\boldsymbol{B})}}^T\left(A_{(B)}^{-1}\left(\boldsymbol{b} - A_{(N)}\boldsymbol{x}_{(\boldsymbol{N})}\right)\right) + {\boldsymbol{c}_{(\boldsymbol{N})}}^T\boldsymbol{x}_{(\boldsymbol{N})} = \\
= \quad & {\boldsymbol{c}_{(\boldsymbol{B})}}^T A_{(B)}^{-1}\boldsymbol{b} - {\boldsymbol{c}_{(\boldsymbol{B})}}^T A_{(B)}^{-1}A_{(N)}\boldsymbol{x}_{(\boldsymbol{N})} + {\boldsymbol{c}_{(\boldsymbol{N})}}^T\boldsymbol{x}_{(\boldsymbol{N})} \\
= \quad & {\boldsymbol{c}_{(\boldsymbol{B})}}^T A_{(B)}^{-1}\boldsymbol{b} + \left({\boldsymbol{c}_{(\boldsymbol{N})}}^T - {\boldsymbol{c}_{(\boldsymbol{B})}}^T A_{(B)}^{-1}A_{(N)}\right)\boldsymbol{x}_{(\boldsymbol{N})}
\end{aligned}$$

שימו לב ש-${\boldsymbol{c}_{(\boldsymbol{B})}}^T A_{(B)}^{-1}\boldsymbol{b}$ הוא קבוע: הוא אמנם תלוי ב-$\boldsymbol{b}$ וב-B, אבל הוא לא תלוי ב-$\boldsymbol{x}$. הוספה/החסרה של קבועים לפונקציית המטרה זה דבר טוב לנו, כי הם לא משנים את הפתרון.
נסמן $-\gamma = {\boldsymbol{c}_{(\boldsymbol{B})}}^T A_{(B)}^{-1}\boldsymbol{b}$ ($\gamma \in \mathbb{R}$), ו-$\widetilde{\boldsymbol{c}} = \left({\boldsymbol{c}_{(\boldsymbol{N})}}^T - {\boldsymbol{c}_{(\boldsymbol{B})}}^T A_{(B)}^{-1}A_{(N)}\right)$ (**$\widetilde{\boldsymbol{c}}$ הוא וקטור שורה**).
כעת, התוכנית שלנו היא:

$$\begin{cases}
\min\limits_{\boldsymbol{x}} & \widetilde{\boldsymbol{c}}\boldsymbol{x}_{(\boldsymbol{N})} - \gamma \\
\text{subject to} & A_{(B)}\boldsymbol{x}_{(\boldsymbol{B})} + A_{(N)}\boldsymbol{x}_{(\boldsymbol{N})} = \boldsymbol{b} \\
& \boldsymbol{x}_{(\boldsymbol{B})}, \boldsymbol{x}_{(\boldsymbol{N})} \geq \boldsymbol{0}
\end{cases}$$

- אם לכל $j \in N$ מתקיים $\widetilde{c}_j \geq 0$, אז (זכרו ש-$\boldsymbol{x}_{(\boldsymbol{N})} \geq \boldsymbol{0}$ זה אילוץ) מתקיים ש-$-\gamma = 0 - \gamma \leq \widetilde{\boldsymbol{c}}\boldsymbol{x}_{(\boldsymbol{N})} - \gamma$. מנגד, $-\gamma$ זה הערך של $\boldsymbol{x}$ שנתון לנו כפתרון אפשרי בסיסי (שכן $\boldsymbol{x}_{(\boldsymbol{N})} = 0$).

  כלומר למעשה $\boldsymbol{x}$ שלנו הוא פתרון אופטימלי (לא משנה מה, אי אפשר להוריד יותר את ערך פונקציית המטרה).

- נניח שיש כניסה $j \in N$ ככה ש-$\widetilde{c}_j < 0$. **במקרה זה אנחנו כן יכולים להוריד ערך מפונקציית המטרה:** נעשה את זה על ידי הגדלה של הכניסה המתאימה ב-$\boldsymbol{x}_{(\boldsymbol{N})}$, כלומר את $x_{(N)j}$. שימו לב שכרגע אנחנו לא בהכרח נשארים עם פתרון חוקי, כי צריך לאכוף את זה ש-$A_{(B)}\boldsymbol{x}_{(\boldsymbol{B})} + A_{(N)}\boldsymbol{x}_{(\boldsymbol{N})} = \boldsymbol{b}$ (והתנאי השני $\boldsymbol{x}_{(\boldsymbol{B})}, \boldsymbol{x}_{(\boldsymbol{N})} \geq \boldsymbol{0}$ בטוח מתקיים).

  זה לא קשה להבטיח ש-$A_{(B)}\boldsymbol{x}_{(\boldsymbol{B})} + A_{(N)}\boldsymbol{x}_{(\boldsymbol{N})} = \boldsymbol{b}$, כי הנחנו ש-$A_{(B)} = I_m$, ואז רק צריך לעשות התאמה ב-$\boldsymbol{x}_{(\boldsymbol{B})}$ כדי לשמר את השוויון $\boldsymbol{x}_{(\boldsymbol{B})} + A_{(N)}\boldsymbol{x}_{(\boldsymbol{N})} = \boldsymbol{b}$. למעשה זה יהיה להקטין איזושהי קורדינטה ב-$\boldsymbol{x}_{(\boldsymbol{B})}$, בתור פיצוי על ההגדלה שעשינו ("מבחינת עלות" אין לנו בעיה עם זה, כי הקורדינטות של $\boldsymbol{x}_{(\boldsymbol{B})}$ לא משפיעות בכלל על פונקציית המטרה!).

במקרה שבו $\boldsymbol{x}$ לא פתרון אופטימלי (המקרה השני), שימו לב שהשאיפה שלנו היא שהשינוי של $\boldsymbol{x}_{(\boldsymbol{N})}$ יהיה שינוי מקסימלי (ראינו שככל שהוא גדל, ערך פונקציית המטרה קטן). מנגד, יש לנו את מגבלת החיוביות של $\boldsymbol{x}_{(\boldsymbol{B})}$, אז אפשר להגדיל את $\boldsymbol{x}_{(\boldsymbol{N})}$ עד גבול מסוים (צריך להבין מהו).
שינינו את $\boldsymbol{x}_{(\boldsymbol{N})}$ ו-$\boldsymbol{x}_{(\boldsymbol{B})}$ בהתאם, על הדרך הקטנו את ערך פונקציית המטרה. אז כעת ל-$\boldsymbol{x}$ ערכים חדשים, ולכן צריך לעדכן בהתאם את B ואת N (אם איפסנו את הכניסה x_i כאשר $i \in B$, והגדלנו את x_j כאשר $j \in N$, אנחנו נכניס את j ל-B ואת i ל-N). כעת נמשיך לבצע אותו תהליך.
התהליך שביצענו היה למעשה **להחליף איבר בבסיס**. למעשה משום שלהיות קודקוד זה שקול לקיים $\boldsymbol{x}_{(B)} > \boldsymbol{0}$, אנחנו עברנו מקודקוד אחד לקודקוד אחר.

רעיון אלגוריתם הסימפלקס יהיה לחשב בכמה אפשר להגדיל את $x_{(N)j}$ עד שיש קורדינטה מתאימה ב-$\boldsymbol{x}_{(\boldsymbol{B})}$ שמתאפסת (כאמור, השאיפה היא להגדיל כמה שיותר, אבל אנחנו מוגבלים על ידי אילוץ הסימן).
ואז במקרה כזה, הורדנו את ערך פונקציית המטרה, ואיפסנו כניסה ב-$\boldsymbol{x}_{(\boldsymbol{B})}$. כעת קיבלנו וקטור $\boldsymbol{x}$ חדש: אנחנו נגדיר מחדש את הבסיס B (נחליף בין האינדקסים שהקטנו והעלינו) והפתרון הפיזיבילי הבסיסי $\boldsymbol{x}$ בהתאם. עלינו לדאוג כעת

ש-$A_{(B)}$ (זו מטריצה חדשה כעת, לא מה שהייתה פעם) תהיה מטריצת היחידה - אפשר לעשות את זה על ידי דירוג, וכן $\boldsymbol{b} \geq \boldsymbol{0}$, ולאחר מכן נמשיך לאיטרציה נוספת.

בעמוד הבא דוגמת הרצה מפורטת.

4.2 דוגמת הרצה

הרצאה 6

נסתכל על התוכנית
$$\begin{aligned} \min\ & 19x_1 - 13x_3 + x_4 + 3x_6 \\ \text{s.t. }\ & 3x_1 - 4x_3 + x_4 + 4x_5 = 6 \\ & -4x_1 + x_2 + x_3 + 2x_5 = 1 \\ & 5x_1 - 2x_3 - x_5 + x_6 = 1 \\ & \boldsymbol{x} \geq \boldsymbol{0} \end{aligned}$$
ונדגים את אלגוריתם הסימפלקס עליה.

מקובל לרשום את התוכנית בטבלה הבאה:

בשורה הראשונה מקדמי פונקציית המטרה, ובשורות הבאות האילוצים ככה שבעמודה הימנית מופיע $\boldsymbol{b}$.

באיבר הימני עליון יופיע מקדם חופשי $-\gamma$ של פונקציית המטרה, כרגע הוא 0 (בתוכניות לינאריות המקדם לא משנה למינימיזציה, כמו שכבר ראינו, אבל הוא ישחק תפקיד באלגוריתם שלנו כי בכל שלב נחשב מחדש את פונקציית המטרה, כפי שהצגנו מקודם).

19	0	-13	1	0	3	0
3	0	-4	1	4	0	6
-4	1	1	0	2	0	1
5	0	-2	0	-1	1	1

1. בשלב הראשון נבחר בסיס ופתרון בסיסי. במקרה הזה זה קל (אם כי נצטרך לתת את הדעת לגבי איך עושים את זה במקרה הכללי): הבסיס שנבחר הוא $B = \{2,4,6\}$ וזה רק בגלל שהמטריצה $A_{(B)} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ היא מטריצה הפיכה. בואו נניח שהבסיס הוא $B = \{4,2,6\}$ ושהסדר משנה ואז $A_{(B)}$ היא ממש מטריצת היחידה I_3.
מי הפתרון האפשרי הבסיסי שלנו? ע"פ הגדרה הקורדינטות שלו שלא בבסיס (מה שכינינו N) צריכות להיות 0, ולכן הוא מהצורה $(0,?,0,?,0,?)$. הוא אמור לקיים את מערכת המשוואות שמתוארת בטבלה ($A\boldsymbol{x} = \boldsymbol{b}$) ואפשר לראות שהאופציה היחידה היא שזהו $\boldsymbol{x} = (0,1,0,6,0,1)$.

נזכיר שהמטרה היא לאפס את מקדמי פונקציית המטרה בבסיס, כלומר את c_2, c_4, c_6 שהם $0,1,3$.

נשים לב שאם אנחנו מוסיפים לשורה הראשונה כפולה של שורה אחרת (למשל נחסיר מהשורה הראשונה את השורה השניה), אנחנו לא משנים את מרחב הפתרונות **כל עוד אנחנו מעדכנים את המקדם החופשי**. כלומר כן משנים את פונקציית המטרה, אבל לא את מרחב הפתרונות וגם לא את הערך המינימלי של הפונקציה.

תחילה המטרה היא לאפס את $c_4 = 1$ ולכן נחסיר מהשורה הראשונה את השורה השניה:

16	0	-9	0	-4	3	-6
3	0	-4	1	4	0	6
-4	1	1	0	2	0	1
5	0	-2	0	-1	1	1

עכשיו צריך לאפס את $c_6 = 3$ ולשם כך נחסיר מהשורה הראשונה 3 פעמים את השורה השלישית:

1	0	-3	0	-1	0	-9
3	0	-4	1	4	0	6
-4	1	1	0	2	0	1
5	0	-2	0	-1	1	1

כל מקדמי פונקציית המטרה בבסיס עכשיו הגיעו ל-0 ($c_2 = c_4 = c_6 = 0$).

שימו לב שהסיבה שהיה לנו קל לטפל בכל אחד ממקדמי הבסיס בנפרד, הוא העובדה ש-$A_{(B)} = I_3$.

מה שאנחנו עשינו זה בעצם להגיד שלכל $\boldsymbol{x}$ ככה ש-$A\boldsymbol{x} = \boldsymbol{b}$ ו-$\boldsymbol{x} \geq \boldsymbol{0}$ מתקיים:

$$\underbrace{19x_1 - 13x_3 + x_4 + 3x_6}_{\text{פונקציית המטרה הקודמת}} = \underbrace{x_1 - 3x_3 - x_5 - (-9)}_{\text{פונקציית המטרה החדשה}}$$

(שימו לב למינוס המקדם החופשי).

2. נבחר כניסה שלילית בפונקציית המטרה הנוכחית, למשל את $c_5 = -1$. מה שאומר שאם נגדיל את הכניסה החמישית בוקטור הפתרון $\boldsymbol{x}$ שלנו, x_5, פונקציית המטרה תקבל ערכים **קטנים** יותר וזה מה שאנחנו רוצים. שוב נזכור שכדי להגדיל את x_5 עלינו לשמור על האילוצים שיתקיימו.
כדי לשמור שהאילוצים יתקיימו, נשנה את ערכי משתני הבסיס (x_2, x_4, x_6):

- כשמגדילים את x_5 ומשאירים את המשתנים x_3, x_1 קבועים (כי הם לא בבסיס), האילוץ הראשון, זה שמופיע בשורה השניה של הטבלה ($3x_1 - 4x_3 + x_4 + 4x_5 = 6$), אומר שאנחנו צריכים להקטין גם את x_4. אם x_5 גדל ב-ε, x_4 חייב לרדת ב-4ε, כדי לשמור את המשוואה $3x_1 - 4x_3 + x_4 + 4x_5 = 6$.
אנחנו רוצים לאפס את x_4 לגמרי (אם היה אפשר, היינו אפילו רוצים אותו עוד יותר קטן, אבל אילוץ הסימן $\boldsymbol{x} \geq \boldsymbol{0}$ מגביל אותנו), כרגע $x_4 = 6$ - אז מהו ה-ε שצריך להגדיל את x_5 בו? נפתור את $6 - 4\varepsilon = 0$ ונקבל $\varepsilon = 1.5$.
במילים אחרות: אפשר להקטין את x_4 ב-1.5 לכל היותר.

חישוב דומה ייעשה בשורות האחרות.

- למשל בשורה השלישית, שמתאימה למשוואה $-4x_1 + x_2 + x_3 + 2x_5 = 1$, הגדלה של x_5 ב-ε תגרור הקטנה של x_2 ב-2ε. כרגע $x_2 = 1$ ולכן אפשר להקטין את x_2 ב-$\varepsilon = \frac{1}{2}$ לכל היותר ($1 - 2\varepsilon = 0$).
- בשורה הרביעית שמתאימה למשוואה $5x_1 - 2x_3 - x_5 + x_6 = 1$, הגדלה של x_5 ב-ε לא תקטין לנו כלל את x_6 אלא רק תגדיל אותו - **וזה בגלל שהמקדם של x_5, שהוא -1, שלילי.** מפה, אין לנו מגבלה לגבי ε, אבל כן נשים לב שאם נגדיל את x_5 ב-ε, חובה עלינו יהיה להגדיל גם את x_6 ב-ε.

רוצים לפתור את ה-$\varepsilon > 0$ ההדוק ביותר שמקיים $\begin{cases} 6 - 4\varepsilon \geq 0 & \text{עבור } x_4 \\ 1 - 2\varepsilon \geq 0 & \text{עבור } x_2 \\ 1 + \varepsilon \geq 0 & \text{עבור } x_6 \end{cases}$. והאילוץ המגביל ביותר (נכנה אותו "האילוץ הקריטי") הוא האילוץ השני, של x_2: אז נקבל שנוכל להגדיל את x_5 רק עד כדי $\frac{1}{2}$. במקרה שכזה כאמור, x_2 יתאפס.
האיבר בשורה השלישית (זו שמתאימה לאילוץ השני), ובעמודה החמישית, זה האיבר שבחרנו לאפס. בכל איטרציה הוא ייקרא **איבר הציר** (pivot).

נחשב את הכניסות החדשות של הוקטור $\boldsymbol{x} = (0, 1, 0, 6, 0, 1)$ כדי להבין מיהו הפתרון הבסיסי החדש. ב-x_1, x_3 לא נגענו (קיבענו אותם).
מתקיים $x_2 = 1 - 2 \cdot \frac{1}{2} = 0$, $x_4 = 6 - 4 \cdot \frac{1}{2} = 4$, $x_5 = 0 + \frac{1}{2} = \frac{1}{2}$, ו-$x_6 = 1 + \frac{1}{2} = \frac{3}{2}$.

נבחר בסיס חדש וזהו יהיה כמעט אותו בסיס מקודם, רק שנחליף בין 2 ו-5 (בגלל שאיפסנו את 2, כש-5 גדל על חשבונו) ולכן 2 שהיה בבסיס יוחלף ב-5 ועתה הבסיס יהיה $B = \{4, 5, 6\}$.

כלומר כעת $\boldsymbol{x} = \left(0, 0, 0, 4, \frac{1}{2}, \frac{3}{2}\right)$ - זהו קודקוד חדש של הפוליהדרון $P = \{\boldsymbol{x} \mid A\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0}\}$ (שימו לב שמדובר בקודקוד, כי $A_{(B)}$ היא הפיכה (משפט 3.4)).

3. כעת יש לנו בסיס חדש - $B = \{4,5,6\}$. נזכור שהמטרה היא תמיד לאפס את מקדמי הבסיס בפונקציית המטרה. $c_4 = 0$ ו-$c_6 = 0$, אבל $c_5 = -1$.

רוצים לאפס את c_5 ולשם כך נעשה דבר כזה: נוסיף לפונקציית המטרה $\frac{1}{2}$ מהשורה השלישית (זו שמתאימה לאילוץ השני). כך נקבל:

1	0	−3	0	−1	0	−9
3	0	−4	1	4	0	6
−4	1	1	0	2	0	1
5	0	−2	0	−1	1	1

$\Leftarrow$

-1	$\frac{1}{2}$	$-\frac{5}{2}$	0	0	0	$-\frac{17}{2}$
3	0	−4	1	4	0	6
−4	1	1	0	2	0	1
5	0	−2	0	−1	1	1

4. לפני שנמשיך, אנחנו רוצים לדאוג ש-$A_{(B)}$ תהיה מטריצת היחידה, כי כרגע $A_{(B)} = \begin{bmatrix} 1 & 4 & 0 \\ 0 & 2 & 0 \\ 0 & -1 & 1 \end{bmatrix}$.

כפי שניתן לראות הבעיה היא העמודה השניה. כדי לטפל בה נפעיל תהליך של דירוג. ראינו כבר שהדירוג לא משנה את מרחב הפתרונות (אנחנו שומרים על השוויונות, כמובן ש-$\boldsymbol{b}$ משתתף יחד בדירוג).

- נכפיל את השורה שמתאימה לאילוץ השני בחצי, כעת נקבל:

-1	$\frac{1}{2}$	$-\frac{5}{2}$	0	0	0	$-\frac{17}{2}$
3	0	−4	1	4	0	6
−2	$\frac{1}{2}$	$\frac{1}{2}$	0	1	0	$\frac{1}{2}$
5	0	−2	0	−1	1	1

- נוסיף לשורה השניה (אילוץ ראשון) -4 פעמים השורה השלישית (אילוץ שני), ולשורה הרביעית (אילוץ שלישי) נוסיף בדיוק את השורה השלישית (אילוץ שני). כעת נקבל:

-1	$\frac{1}{2}$	$-\frac{5}{2}$	0	0	0	$-\frac{17}{2}$
11	−2	−6	1	0	0	2
−2	$\frac{1}{2}$	$\frac{1}{2}$	0	1	0	$\frac{1}{2}$
3	$\frac{1}{2}$	$-\frac{3}{2}$	0	0	1	$\frac{3}{2}$

שימו לב - קיבלנו **אותה תוכנית לינארית** פשוט מיוצגת בדרך אחרת ועם פונקציית מטרה שונה.

5. נמשיך - נבחר כניסה שלילית כלשהי של $\boldsymbol{c}$, למשל הכניסה c_1, כי $c_1 = -1$. המטרה היא להעביר את 1 לבסיס, השאלה היא על חשבון מי.

שוב עוברים אילוץ אילוץ ומנסים לראות עד כמה הגדלה של x_1 ב-ε תגרור הגדלה של x_4, x_5, x_6 (משתני הבסיס). למשל, האילוץ הראשון יגיד לנו שהגדלה של x_1 ב-ε, גורר **הקטנה** של x_4 ב-11ε, האילוץ השני יגיד לנו שהגדלה של x_1 ב-ε יגרור **הגדלה** של x_5 ב-2ε, והשלישי יגיד שהגדלה של x_1 ב-ε יגרור **הקטנה** של x_6 ב-3ε.

משום ש-$\boldsymbol{x} = \left(0,0,0,4,\frac{1}{2},\frac{3}{2}\right)$, כלומר $x_4 = 4$, $x_5 = \frac{1}{2}$ ו-$x_6 = \frac{3}{2}$,

עלינו למצוא את ה-$\varepsilon > 0$ ההדוק ביותר שמקיים $\begin{cases} \text{עבור } x_4 & 4 - 11\varepsilon \geq 0 \\ \text{עבור } x_5 & \frac{1}{2} + 2\varepsilon \geq 0 \\ \text{עבור } x_6 & \frac{3}{2} - 3\varepsilon \geq 0 \end{cases}$. נקבל ש"האילוץ הקריטי" הוא של x_4, $\varepsilon = \frac{4}{11}$. כך איבר הציר (pivot) הוא 11 (השורה השניה, זו של האילוץ הראשון, והעמודה הראשונה).

כלומר נקטין כל כניסה בהתאם, ונקבל את הוקטור $\boldsymbol{x} = \left(\frac{4}{11},0,0,0,\frac{27}{22},\frac{9}{22}\right)$ ($x_1 = 0 + \frac{4}{11} = \frac{4}{11}$).

כעת הבסיס שלנו יהיה $B = \{1,5,6\}$.

עכשיו המטרה תהיה לאפס את מקדמי פונקציית המטרה באינדקסים של הבסיס, כלומר c_1, c_5, c_6. $c_5 = c_6 = 1$

אבל $c_1 = -1$, נאפס אותו על ידי הוספה של $\frac{1}{11}$ כפול השורה השניה לפונקציית המטרה, לאחר מכן נדרג את הטבלה כדי ש-$A_{(B)} = I_3$ ונחזור על התהליך.

6. **עצירה:** כאשר מתקיים $\boldsymbol{c} \geq \boldsymbol{0}$, וקטור הפתרון הבסיסי $\boldsymbol{x}$ שנישאר איתו הוא הפתרון האופטימלי.

הערה 4.1. ה-pivot rule:

- איך בוחרים את הכניסה של $\boldsymbol{c}$ שאנחנו רוצים לאפס? - מה שעשינו בשלבים 2 ו-5.
כלל הבחירה הזאת נקרא ה-pivot rule. ה-pivot rule המקובל הוא בדר"כ לבחור את העמודה עם הערך הכי שלילי; שזה לא מה שעשינו במקרה שלנו (אם היינו עוקבים אחרי הכלל הזה, למשל, היינו בוחרים את c_3 בשלב 2 במקום c_5, כי הוא קטן יותר).

- כלל הפיבוט של בחירת הערך הכי שלילי יכול הרבה פעמים להוביל אותנו ללולאות אינסופיות, בהן אנחנו בוחרים את j ומקבלים במקום את k, אבל אז כשרוצים לאפס את k מגלים שצריך לשנות את j, וכן הלאה וכן הלאה.
זו בעיה נפוצה ברוב כללי הפיבוט.

- למרות זאת, Bland's rule הוא כלל פיבוט שמונע לולאות. בפועל מדובר בכלל נוראי ואיטי שמפחית מהביצועים של האלגוריתם.
הכלל הוא: נבחר את העמודה עם האינדקס הקטן ביותר j כך ש-c_j שלילי, והאיבר שייכנס לבסיס הוא האינדקס של השורה בה נמצא האיבר החיובי הקטן ביותר (מתוך האיברים של אותה העמודה).

- אפשר להוכיח לרוב כללי הפיבוט שהם יכולם להגיע לכל היותר, במקרה הגרוע, מספר אקספוננציאלי של צעדי פיבוט.
מה שאומר שאלגוריתם הסימפלקס הוא בגדול אלגוריתם אקספוננציאלי. אנחנו לא יודעים עדיין אם יש כלל פיבוט שמבטיח לנו אלגוריתם שירוץ בזמן פולינומיאלי תמיד. כן יש כללי פיבוט שיכולים לתת זמן ריצה שהוא תת-אקספוננציאלי (עדיין לא פולינומי).
בפועל זמן ריצת האלגוריתם של הסימפלקס הוא זמן טוב בפועל.

תיאור האלגוריתם עדיין לוקה בחסר, שכן עלינו לטפל בשאלה "איך מקבלים בסיס התחלתי?" (וקודקוד שמתאים לו).
כדי להשיג קודקוד, אנחנו הולכים להעלות את הפוליהדרון למימד גבוה יותר. לכל אחד מ-m האילוצים שלנו נגדיר משתנה חדש x', כלומר נגדיר את המשתנים $x'_1, \dots, x'_m$.

כעת הבעיה
$$\begin{aligned} \min \ & \sum_{i=1}^{m} x'_i \\ \text{s.t. } \ & A\boldsymbol{x} + I_m \boldsymbol{x'} = \boldsymbol{b} \\ & \boldsymbol{x}, \boldsymbol{x'} \geq \boldsymbol{0} \end{aligned}$$
(כאשר $\boldsymbol{x'} = \begin{pmatrix} x'_1 \\ \vdots \\ x'_m \end{pmatrix}$) - היא בעיה שיש לה קודקוד טריויאלי וזה הקודקוד שבו $x_i = b_i$ ו-$x'_i = 0$ (שימו לב שמדובר בהשמה חוקית עבור $\boldsymbol{x}$ שכן $\boldsymbol{b} \geq \boldsymbol{0}$).

ואז **אם הפוליהדרון המקורי שלנו לא ריק,** אז מובן כי המינימום $\min \sum_{i=1}^{m} x'_i$ הוא 0 (שכן אפשר לבחור $\boldsymbol{x}$ שמקיים את המשוואה $A\boldsymbol{x} = \boldsymbol{b}$, $\boldsymbol{x} \geq \boldsymbol{0}$, ו-$\boldsymbol{x'} = \boldsymbol{0}$, ככה המינימום יוצא $\sum_{i=1}^{m} x'_i = 0$). אם המינימום של $\sum_{i=1}^{m} x'_i$ גדול מ-0, סימן שהפוליהדרון המקורי שלנו ריק.

כלומר יהיה עלינו להריץ את אלגוריתם הסימפלקס פעמיים, פעם אחת על הבעיה החדשה (אפשר להתחיל להריץ את הסימפלקס על הבעיה הזו, כי יש לנו קודקוד התחלתי בשביל התוכנית, כפי שראינו), ואז נקבל זוג $(\boldsymbol{x}, \boldsymbol{x'})$ שהוא הפתרון האופטימלי לבעיה החדשה. אם $\boldsymbol{x'} = \boldsymbol{0}$ סימן שאפשר להשתמש ב-$\boldsymbol{x}$ בתור קודקוד התחלתי לבעיה המקורית, ומשם נריץ סימפלקס. אחרת, כאמור, התוכנית לא פיזיבילית.

5 דואליות

תהא התוכנית $\begin{aligned} &\min \boldsymbol{c}^T\boldsymbol{x} \\ \text{s.t. } &A\boldsymbol{x}=\boldsymbol{b} \\ &\boldsymbol{x}\geq \boldsymbol{0}\end{aligned}$ בצורתה הסטנדרטית. איך אפשר להעריך את המינימום $z=\min \boldsymbol{c}^T\boldsymbol{x}$?

אפשר לחשוב על כך כבעיית החלטה: בהינתן $z_0\in\mathbb{R}$ האם הערך האופטימלי z מקיים $z\leq z_0$ או $z>z_0$?

ניקח קומבינציות לינאריות של שורות האילוצים. עושים את זה על ידי הכפלת A משמאל ב-$\boldsymbol{y}^T$ כאשר $\boldsymbol{y}$ הוא וקטור המקדמים של הקומבינציה.

נשים לב כי בהינתן $\boldsymbol{y}\in\mathbb{R}^m$ מתקיים $\boldsymbol{y}^TA\boldsymbol{x}=\boldsymbol{y}^T\boldsymbol{b}$.

אז נרצה לבחור $\boldsymbol{y}\in\mathbb{R}^m$ ככה ש-$\boldsymbol{y}^TA\leq \boldsymbol{c}^T$, ואז $\boldsymbol{y}^T\boldsymbol{b}=\boldsymbol{y}^TA\boldsymbol{x}\leq \boldsymbol{c}^T\boldsymbol{x}$ (כי $\boldsymbol{x}\geq\boldsymbol{0}$).

באופן ספציפי, נוכל לבחור מספר w שהוא $w=\max\limits_{\boldsymbol{y} \text{ s.t. } \boldsymbol{y}^TA\leq \boldsymbol{c}^T}\boldsymbol{y}^T\boldsymbol{b}$, אז מתקיים $w=\boldsymbol{y}^T\boldsymbol{b}\leq\boldsymbol{c}^T\boldsymbol{x}=z$.

אופן בחירת w הוא **תוכנית לינארית** בעצמו (אמנם לא בהצגה הסטנדרטית או הקנונית, אבל זו עדיין תוכנית).

הגדרה 5.1. התוכנית $\begin{aligned} &\max_{\boldsymbol{y}\in\mathbb{R}^m} \boldsymbol{y}^T\boldsymbol{b} \\ \text{s.t. } &\boldsymbol{y}^TA\leq \boldsymbol{c}\end{aligned}$ נקראת **התוכנית הדואלית** (dual program) לתוכנית $\begin{aligned} &\min \boldsymbol{c}^T\boldsymbol{x} \\ \text{s.t. } &A\boldsymbol{x}=\boldsymbol{b} \\ &\boldsymbol{x}\geq \boldsymbol{0}\end{aligned}$.

אז אם אנחנו יכולים לפתור את התוכנית הדואלית, יש לנו w שמקיים $w\leq z$, ועל כן חסם תחתון ל-z. אי השוויון $w\leq z$ מביא אותנו ל**משפט הדואליות החלשה**.

משפט 5.2 (דואליות חלשה)**.** לכל פתרון פיזיבילי $\boldsymbol{x}$ לבעיה המקורית ו-$\boldsymbol{y}$ לבעיה הדואלית, $\boldsymbol{b}^T\boldsymbol{y}\leq\boldsymbol{c}^T\boldsymbol{x}$.

הוכחה. ראינו ש-$w\leq z$, אבל $w=\max\limits_{\boldsymbol{y} \text{ is feasible}}\boldsymbol{y}^T\boldsymbol{b}$ ו-$z=\min\limits_{\boldsymbol{x} \text{ is feasible}}\boldsymbol{c}^T\boldsymbol{x}$.

לכן לכל פתרון פיזיבילי $\boldsymbol{x}$ לבעיה המקורית ו-$\boldsymbol{y}$ לבעיה הדואלית, מתקיים $\boldsymbol{y}^T\boldsymbol{b}\leq w\leq z\leq\boldsymbol{c}^T\boldsymbol{x}$ כנדרש. □

אז כל ערך אפשרי של התוכנית הדואלית חוסם מלמטה את האופטימום של התוכנית המקורית - וזה מה שאנחנו רצינו (רצינו חסם מלמעלה, אבל זה שקול).

למעשה משפט חזק יותר נותן לנו אפילו שוויון בין w ל-z:

משפט 5.3 (דואליות חזקה)**.** אם אחת מהתוכניות הלינאריות היא פיזיבילית (לא ריקה), אז $w=z$.

משפט הדואליות החזקה הוא אחד הדברים החשובים שיש לנו בניתוח בעיות לינאריות. למשל, למדנו בקורס אלגוריתמים הראשון שבעיית החתך המינימלי ברשת זרימה ובעיית הזרימה המקסימלית הן בעיות לינאריות דואליות - ומשפט הדואליות החזקה מאפשר לנו להסיק משפט בשם "משפט השטף והחתך" (max-flow min-cut) והוא שהקיבול של החתך המינימלי, שווה לשטף הזרימה המקסימלית בכל רשת זרימה.

הוכחה. ראינו כי $w\geq z$ מהדואליות החלשה ולכן נותר להראות כי $z\leq w$.

נניח בלי הגבלת הכלליות, שמתקיים שבעיית המינימיזציה היא פיזיבילית (הסיבה שזה בה"כ היא שנראה אחר כך שהבעיה הדואלית לבעיה הדואלית היא הבעיה המקורית, אז אפשר להחליף את התפקידים).

אם הבעיה המקורית (המינימיזציה) איננה חסומה, אין מה להוכיח: המינימום הוא $-\infty$, ואז הדואליות החלשה שגורסת כי $z\leq w$, תגיד לנו שבהכרח $w=-\infty$. הרצאה 7

אחרת נסמן ב-$\boldsymbol{x}^*$ פתרון אופטימלי לבעיית המינימיזציה: $A\boldsymbol{x}^*=\boldsymbol{b}$, $\boldsymbol{x}^*\geq\boldsymbol{0}$ ו-$\boldsymbol{c}^T\boldsymbol{x}^*=z$.

עלינו למצוא $\boldsymbol{y}$ אשר $A^T\boldsymbol{y}\leq\boldsymbol{c}$ (מקיים את אילוץ התוכנית הדואלית) ו-$\boldsymbol{b}^T\boldsymbol{y}\geq z$ (ערך הפתרון $\boldsymbol{y}$ גדול שווה z), קרי $-\boldsymbol{b}^T\boldsymbol{y}\leq -z$.

נניח בשלילה שלא קיים $\boldsymbol{y}$ כזה, כלומר לאי השוויון $\begin{pmatrix} A^T \\ \hline -\boldsymbol{b}^T \end{pmatrix} \boldsymbol{y} \leq \begin{pmatrix} \boldsymbol{c} \\ -z \end{pmatrix}$ (באגף שמאל יש מטריצה, בימין וקטור) **אין פתרון**.

לפי הלמה של פרקש (או גרסה שלה, שלא ראינו), יש $\lambda \geq 0$, וקטור $\boldsymbol{x} \geq \boldsymbol{0}$, כלומר $\begin{pmatrix} \boldsymbol{x} \\ \lambda \end{pmatrix} \geq \boldsymbol{0}$, ככה ש-

$\begin{pmatrix} A^T \\ \hline -\boldsymbol{b}^T \end{pmatrix}^T \begin{pmatrix} \boldsymbol{x} \\ \lambda \end{pmatrix} = \boldsymbol{0}$,[4] אבל $\begin{pmatrix} \boldsymbol{c} \\ -z \end{pmatrix}^T \begin{pmatrix} \boldsymbol{x} \\ \lambda \end{pmatrix} < 0$.

כלומר $\begin{pmatrix} A & | & -\boldsymbol{b} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \lambda \end{pmatrix} = \boldsymbol{0}$, ו-$\begin{pmatrix} \boldsymbol{c}^T & | & -z \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \lambda \end{pmatrix} < 0$.

אז קיים $\boldsymbol{x} \geq \boldsymbol{0}$ ו-$\lambda \geq 0$ ככה ש-$A\boldsymbol{x} - \lambda\boldsymbol{b} = \boldsymbol{0}$, ו-$\boldsymbol{c}^T\boldsymbol{x} - \lambda z < 0$, כלומר $A\boldsymbol{x} = \lambda\boldsymbol{b}$ ו-$\boldsymbol{c}^T\boldsymbol{x} < \lambda z$.

- אם $\lambda > 0$: במקרה זה $\frac{1}{\lambda}\boldsymbol{x}$ הוא פתרון לבעיה המקורית, כי $\frac{1}{\lambda}A\boldsymbol{x} = \boldsymbol{b}$, ובנוסף $\boldsymbol{c}^T\left(\frac{1}{\lambda}\boldsymbol{x}\right) < z$ - זו סתירה, כי z הוא המינימום של התוכנית הלינארית המקורית.
- אם $\lambda = 0$ נקבל $A\boldsymbol{x} = \boldsymbol{0}$ ו-$\boldsymbol{c}^T\boldsymbol{x} < 0$, ובנוסף $\boldsymbol{x} \geq \boldsymbol{0}$. כלומר לכל $\alpha > 0$ מתקיים ש-$\boldsymbol{x}^* + \alpha\boldsymbol{x} \geq \boldsymbol{0}$, ובנוסף $A(\boldsymbol{x}^* + \alpha\boldsymbol{x}) = A\boldsymbol{x}^* + \alpha A\boldsymbol{x} = A\boldsymbol{x}^* = \boldsymbol{b}$,
  וגם $c^T(\boldsymbol{x}^* + \alpha\boldsymbol{x}) = c^T\boldsymbol{x}^* + \alpha\boldsymbol{c}^T\boldsymbol{x} = z + \alpha(-\varepsilon)$ עבור $\varepsilon = \boldsymbol{c}^T\boldsymbol{x} > 0$, ולכן התוכנית המקורית שלנו אינה חסומה (אפשר להגדיל את α כרצוננו ולקבל ערכים קטנים של פונקציית המטרה כרצוננו): סתירה.

הגענו לסתירה בשני המקרים, אז $z \leq w$. יחד עם $z \geq w$ של דואליות חלשה, $z = w$ כנדרש. □

הגדרה 5.4 אם $\begin{aligned} \max_{\boldsymbol{y} \in \mathbb{R}^m} & \ \boldsymbol{y}^T\boldsymbol{b} \\ \text{s.t.} & \ \boldsymbol{y}^T A \leq \boldsymbol{c} \end{aligned}$ היא הדואלית של $\begin{aligned} \min & \ \boldsymbol{c}^T\boldsymbol{x} \\ \text{s.t.} & \ A\boldsymbol{x} = \boldsymbol{b} \\ & \ \boldsymbol{x} \geq \boldsymbol{0} \end{aligned}$, לתוכנית $\begin{aligned} \min & \ \boldsymbol{c}^T\boldsymbol{x} \\ \text{s.t.} & \ A\boldsymbol{x} = \boldsymbol{b} \\ & \ \boldsymbol{x} \geq \boldsymbol{0} \end{aligned}$ קוראים **התוכנית הפרימאלית** (primal program).

טענה 5.5 (כנראה בשיעורי הבית). התוכנית הדואלית לתוכנית הדואלית היא התוכנית הפרימאלית.

הגדרה 5.6 נניח כי $\boldsymbol{x}$ הוא פתרון חוקי לבעיה הפרימאלית, ו-$\boldsymbol{y}$ לבעיה הדואלית. מדואליות חלשה, $\boldsymbol{y}^T\boldsymbol{b} \leq \boldsymbol{c}^T\boldsymbol{x}$. **פער הדואליות (duality gap) בין $\boldsymbol{x}$ ו-$\boldsymbol{y}$** הוא הגודל $\boldsymbol{c}^T\boldsymbol{x} - \boldsymbol{y}^T\boldsymbol{b}$, ו-$\boldsymbol{s}^T = \boldsymbol{c}^T - \boldsymbol{y}^T A$ נקרא ה-dual slack.

נשים לב שתמיד ה-slack מקיים $\boldsymbol{s} \geq \boldsymbol{0}$.

אנחנו יודעים מדואליות חזקה ש-$\boldsymbol{c}^T\boldsymbol{x} - \boldsymbol{y}^T\boldsymbol{b} = 0$ אם ורק אם $\boldsymbol{x}, \boldsymbol{y}$ שניהם הם פתרונות אופטימליים.

את פער הדואליות $\boldsymbol{c}^T\boldsymbol{x} - \boldsymbol{y}^T\boldsymbol{b}$ אפשר לרשום גם: $\boldsymbol{c}^T\boldsymbol{x} - \boldsymbol{y}^T\boldsymbol{b} = \boldsymbol{c}^T\boldsymbol{x} - \boldsymbol{y}^T A\boldsymbol{x} = \left(\boldsymbol{c}^T - \boldsymbol{y}^T A\right)\boldsymbol{x} = \boldsymbol{s}^T\boldsymbol{x}$.

משפט 5.7 (complementary slackness)**.** התנאים הבאים הם שקולים:

1. $\boldsymbol{x}, \boldsymbol{y}$ אופטימליים (כל אחד בבעיה שלו).
2. $\boldsymbol{s}^T\boldsymbol{x} = 0$.
3. לכל i מתקיים $s_i x_i = 0$.
4. לכל i, אם $s_i > 0$ אז $x_i = 0$.

הוכחה. ראינו את השקילות של 1,2 לפני שדיברנו על המשפט. 2,3 שקולים - ברור. גם 3,4 שקולים. לכן הכל שקול. □

[4] בגרסה שאנחנו ראינו ללמה של פרקש לא ראינו את השוויון ל-$\boldsymbol{0}$

6 אלגוריתם האליפסואיד

בשנת 1979 Khachiyan הציג אלגוריתם בשם **אלגוריתם האליפסואיד** שמסוגל לפתור תוכניות לינאריות בזמן אסימפטוטי פולינומי. אף על פי שהוא די גרוע בפועל, האלגוריתם היווה פריצת דרך בעולם האלגוריתמים. זהו האלגוריתם הראשון הפולינומי לתכנון לינארי.

בקצרה

האלגוריתם מקבל תוכנית לינארית לא בצורה הקנונית ולא בצורה הסטנדרטית (אין $\boldsymbol{c}$).
נתונה תוכנית $A\boldsymbol{x} \leq \boldsymbol{b}$ כאשר $\boldsymbol{b} \in \mathbb{Q}^m$ ו-$A \in \mathbb{Q}^{m\times n}$. אנחנו רוצים $\boldsymbol{x} \in \mathbb{Q}^n$.
הנחות אלגוריתם האליפסואיד:

- $P = \{\boldsymbol{x} \in \mathbb{R}^n \mid A\boldsymbol{x} \leq \boldsymbol{b}\}$ הוא פוליטופ (כלומר פוליהדרון **חסום** כקבוצה במרחב) ונתון גם $0 < R \in \mathbb{Q}$ כך ש-$P \subset B(\boldsymbol{0}, R)$.
- $\dim(P) = n$.
- יש $0 < r < R$ שמגיע כקלט לאלגוריתם, ככה ש-$\mathrm{vol}(P) \geq \mathrm{vol}(B(\boldsymbol{0}, r))$ (vol = volume: הנפח של P, צריך להיות גדול מהנפח של כדור כלשהו ברדיוס r סביב $\boldsymbol{0}$).

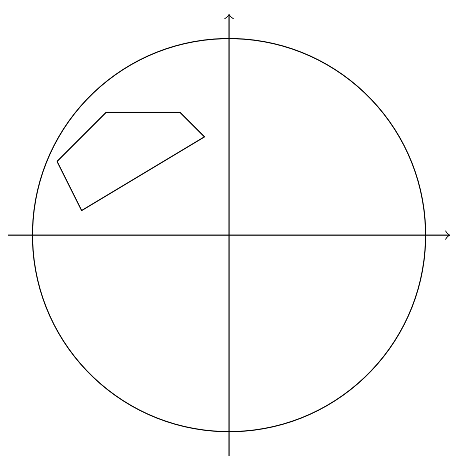
הקלט לאלגוריתם הוא $A, \boldsymbol{b}, R, r$ (נדבר לאן $\boldsymbol{c}$ נעלם בתיאור האלגוריתם).
נסמן ב-$P_< = \{\boldsymbol{x} \in \mathbb{R}^n \mid A\boldsymbol{x} < \boldsymbol{b}\}$ את פנים (interior) הפוליטופ. הפלט הוא אחד מהבאים: $\boldsymbol{x} \in P_< \cap \mathbb{Q}^n$ או (פתרון רציונלי ב-$P_<$, לא בהכרח אופטימלי), או הוכחה שהאלגוריתם נכשל: $\mathrm{vol}(P) < \mathrm{vol}(B(\boldsymbol{0}, r))$.

6.1 האלגוריתם

הרצאה 8

בגדול, מה שהאלגוריתם הולך לעשות הוא:

- נתון פוליטופ P. יודעים שיש כדור כלשהו מרדיוס R שמכיל אותו:



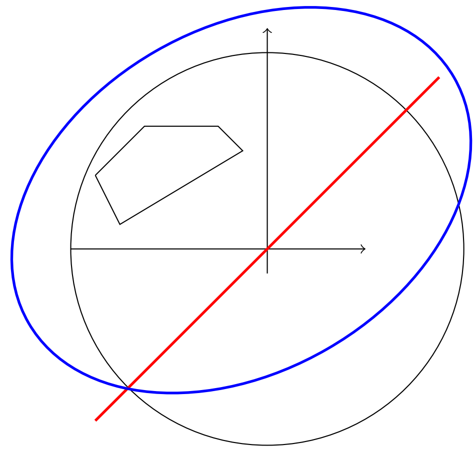
- בוחרים נקודה $\boldsymbol{x_0}$ בכדור, למשל הכי קל את הראשית, ובודקים האם $\boldsymbol{x_0} \notin P_<$ (כלומר $A_i\boldsymbol{x_0} \geq b_i$ לאיזשהו $1 \leq i \leq m$). אם אכן לא, אז יש על-מישור מפריד בין הנקודה ל-$P_<$ (ראו מטה באדום).

לאלגוריתם האליפסואיד לא אכפת איך אנחנו מוצאים אותו. אבל אפשר למצוא אותו בצורה הנ"ל: יודעים ש-$\boldsymbol{A_i x_0} \geq b_i$ לאיזשהו i, ולכן $\boldsymbol{A_i x_0} \geq b_i > \boldsymbol{A_i x}$ כאשר $\boldsymbol{x} \in P_<$ (וזה מההגדרה של $P_<$).

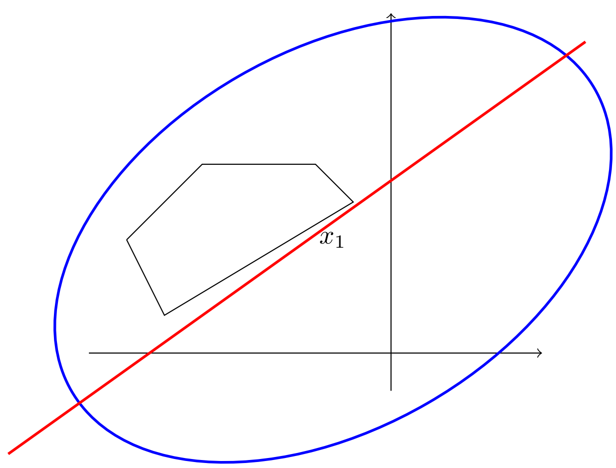
לכן אפשר לבחור את על המישור $\boldsymbol{A_i x} = \boldsymbol{A_i x_0}$, ויש לו את תכונת ההפרדה.

- כעת נחתוך את המישור שלנו לחצי שבו נמצא $P_{<}$. כבר אין לנו כדור שלם, אלא מחצית ממנו (אפילו פחות, אם לא בחרנו את $\boldsymbol{x_0}$ להיות הראשית). מה עושים?

כאן מגיע שמו של האלגוריתם: נמצא אליפסואיד שמכיל חצי של הכדור הזה. אליפסואיד הוא כיווץ/הרחבה של כדור (במישור זו אליפסה).



- **שוב נעשה את אותו דבר:** גם הפעם נבחר את המרכז של האליפסואיד, $\boldsymbol{x_1}$, אם הוא בתוך הפוליטופ - סיימנו, ואחרת נמצא על מישור מפריד (באדום), "נחתוך" את החצי שמתאים לפוליטופ, ושוב נמצא אליפסואיד שמכיל את החצי שלנו.



- **עצירה:** כאשר $\boldsymbol{x} \in P_{<} \cap \mathbb{Q}$ כמו שרצינו. נדבר על מתי מחזירים כישלון אחר כך.

אנחנו נוכל להמשיך כל עוד מתקיים **תנאי האליפסואיד**: אם E הוא אליפסואיד, ו-E' הוא אליפסואיד מנפח מינימלי שמכיל איזשהו חצי של E, אז מתקיים $\operatorname{vol}(E') \leq (1-\varepsilon)\operatorname{vol}(E)$ עבור איזשהו $0 < \varepsilon < 1$ שאנחנו נבחר בהמשך וכל עוד הוא יהיה גדול מספיק.
כלומר, נפחי האליפסואידים קטנים בקצב שהוא לפחות אקספוננציאלי.
<u>נוכיח את התנאי הזה בהמשך</u>.

כאמור, אלגוריתם האליפסואיד מסוגל רק למצוא פתרון פיזיבילי, אם כי לא בהכרח אופטימלי - הוא אפילו לא מקבל את $\boldsymbol{c}$.

אנחנו נשתמש בהצגה מיוחדת שנקראת **ההצגה האליפסואידית**: $\begin{array}{l}\max \boldsymbol{c}^T\boldsymbol{x}\\ \text{s.t. } A\boldsymbol{x}\le \boldsymbol{b}\end{array}$ (לא נוכיח אבל זו הצגה אוניברסלית, כלומר כל תוכנית אפשר להמיר להצגה כזו). נסמן את התוכנית הלינארית ב-LP.

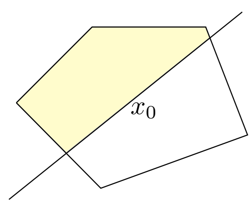
מה אנחנו מניחים באלגוריתם האליפסואיד?

- נצטרך ש-P יהיה מוכל בכדור $B(\mathbf{0},R)$ עבור R נתון התחלתי.
- יש לתוכנית הלינארית LP יש פתרון **ייחודי אופטימלי** שהוא **קודקוד**.
- שמתקיים $\text{vol}(B(\mathbf{0},r))<\text{vol}(P)$.

באלגוריתם האליפסואיד כאמור המטרה היא <u>למקסם את פונקציית המטרה</u> ולא למזער אותה (אנחנו יודעים שזה כמובן שקול).

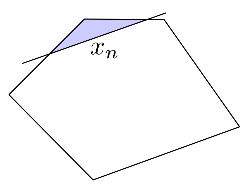
איך בכל זאת ממקסמים את פונקציית המטרה אם האלגוריתם לא מקבל כקלט את c?

אפשר למצוא $\boldsymbol{x_0}\in P_<$ ואז להוסיף את האילוץ $-\boldsymbol{c}^T\boldsymbol{x}\le -\boldsymbol{c}^T\boldsymbol{x_0}$, השקול אל $\boldsymbol{c}^T\boldsymbol{x}\ge \boldsymbol{c}^T\boldsymbol{x_0}$. כך נוכל להריץ שוב את האלגוריתם ולקבל $\boldsymbol{x_1}\in P_<\cap\left\{\boldsymbol{x}\mid -\boldsymbol{c}^T\boldsymbol{x}<-\boldsymbol{c}^T\boldsymbol{x_0}\right\}$ שמקיים אי שוויון ממש: $\boldsymbol{c}^T\boldsymbol{x_1}>\boldsymbol{c}^T\boldsymbol{x_0}$. כעת נוסיף את האילוץ $-\boldsymbol{c}^T\boldsymbol{x}\le -\boldsymbol{c}^T\boldsymbol{x_1}$, השקול אל $\boldsymbol{c}^T\boldsymbol{x}\ge \boldsymbol{c}^T\boldsymbol{x_1}$ וכך נמקסם עוד יותר את פונקציית המטרה. נמשיך הלאה והלאה, עד שהאלגוריתם נכשל.



הוספת האילוץ $-\boldsymbol{c}^T\boldsymbol{x}\le -\boldsymbol{c}^T\boldsymbol{x_0}$ היא למעשה "חיתוך" של הפוליטופ לחלק מתאים.
לאחר מכן נמצא נקודה $\boldsymbol{x_1}$ בפוליהדרון הצהוב, ונחתוך שוב. נחזור על התהליך עוד ועוד,
ונפסיק אותו כאשר הנפח של הפוליטופ החתוך שנקבל הוא קטן מספיק.

כשאנחנו מפסיקים, כשהנפח של הפוליטופ החתוך שנקבל קטן מספיק, נקבל משהו כזה (בדו מימד זה משולש):



כעת, אפשר להשתמש במתודה אחרת שמסוגלת למצוא את הקודקוד הנ"ל - **ומובטח לנו שהוא הקודקוד האופטימלי** (כי אנחנו הנחנו שיש קודקוד אופטימלי ויחיד).

6.2 הפרמול

נגדיר כמה מושגים כדי לפרמל את האלגוריתם וכדי להוכיח אותו.

הגדרה 6.1. נגדיר את **הכדור ברדיוס** $\rho>0$ סביב $\boldsymbol{y}$ להיות:

$$B(\boldsymbol{y},\rho)=\left\{\boldsymbol{x}\in\mathbb{R}^n\mid(\boldsymbol{x}-\boldsymbol{y})^T D^{-1}(\boldsymbol{x}-\boldsymbol{y})\le 1 \text{ where } D=\rho^2 I_n\right\}$$

הערה 6.2. למה אכן מוגדר לנו כדור כפי שאנחנו מכירים אותו? שימו לב כי $D^{-1}=\frac{1}{\rho^2}I_n$, ואז לרשום $(\boldsymbol{x}-\boldsymbol{y})^T D^{-1}(\boldsymbol{x}-\boldsymbol{y})$ זה כמו לרשום:

$$\begin{aligned}(\boldsymbol{x}-\boldsymbol{y})^T D^{-1}(\boldsymbol{x}-\boldsymbol{y}) &= (\boldsymbol{x}-\boldsymbol{y})^T \begin{bmatrix} \frac{1}{\rho^2} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{1}{\rho^2} \end{bmatrix} (\boldsymbol{x}-\boldsymbol{y}) = \\ &= \frac{1}{\rho^2}(\boldsymbol{x}-\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y}) = \\ &= \frac{1}{\rho^2}\sum_{n=1}^{n}(x_i-y_i)^2\end{aligned}$$

ולכן $(\boldsymbol{x}-\boldsymbol{y})^T D^{-1}(\boldsymbol{x}-\boldsymbol{y}) \le 1$ זה אם ורק אם $\sum_{n=1}^{n}(x_i-y_i)^2 \le \rho^2$, כלומר $||x-y|| \le \rho$.

הגדרה 6.3. מטריצה סימטרית $D \in \mathbb{R}^{n\times n}$ נקראת **מוגדרת חיובית** אם לכל $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ מתקיים $\boldsymbol{x}^T D\boldsymbol{x} > 0$.

הערה 6.4. הגדרה שקולה היא שכל הערכים העצמיים של D הם חיוביים ממש.

הגדרה 6.5. אליפסואיד E הוא הכללה של כדור, שם D צריכה להיות מטריצה סימטרית מוגדרת חיובית, לא בהכרח סקלרית. $\boldsymbol{y}$ נקרא **המרכז של האליפסואיד**.

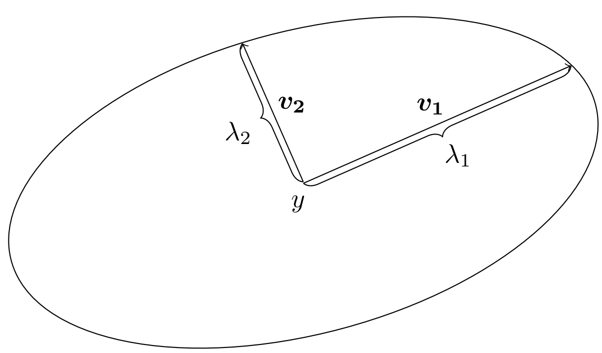
מהמשפט הספקטרלי באלגברה לינארית כל מטריצה סימטרית היא לכסינה אורתוגונלית. כלומר אפשר לרשום את D (מטריצה סימטרית מוגדרת חיובית כלשהי) בתור $D = V\Lambda V^T$ עבור $V \in \mathbb{R}^{n\times n}$ אורתוגונלית (עמודותיה הם בסיס אורגונלי) ולא סינגולרית,[5] ו-Λ הוא מטריצה של הערכים העצמיים של D, ובגלל ש-D היא מטריצה מוגדרת חיובית כולם חיוביים ממש, כלומר הם ריבועים של מספרים: $\Lambda = \begin{bmatrix} \lambda_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_n^2 \end{bmatrix} \in \mathbb{R}^{n\times n}$ (ומשום ש-λ_i כולם חיוביים ממש, היא גם הפיכה).

ולכן אפשר לרשום $D = V\Lambda V^T = \lambda_1 \boldsymbol{v_1}\boldsymbol{v_1}^T + ... + \lambda_n \boldsymbol{v_n}\boldsymbol{v_n}^T$ כאשר $\boldsymbol{v_i}$ היא העמודה ה-i של V (שימו לב ש-$\boldsymbol{v_i}\boldsymbol{v_i}^T$ זו מטריצה).

איך זה נראה גיאומטרית?

נניח שנתונה לנו $D = V\Lambda V^T \in \mathbb{R}^2$ מטריצה מוגדרת חיובית כאשר $V = \begin{bmatrix} | & | \\ v_1 & v_2 \\ | & | \end{bmatrix}$ ו-$\Lambda = \begin{bmatrix} \lambda_1^2 & 0 \\ 0 & \lambda_2^2 \end{bmatrix}$.

האליפסואיד המוגדר על ידי $E = \left\{\boldsymbol{x} \in \mathbb{R}^2 \mid (\boldsymbol{x}-\boldsymbol{y})^T D^{-1}(\boldsymbol{x}-\boldsymbol{y}) \le 1\right\}$ נראה כך:



[5] מטריצה סינגולרית היא מטריצה לא הפיכה.

באופן דומה אם יהיו לנו 3 מימדים, האליפסואיד "ייפרש" על ידי 3 וקטורי הכיוון $\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}$ שהם העמודות של V, והאורכים שלהם יהיו $\lambda_1, \lambda_2, \lambda_3$ בהתאמה.

הגדרה 6.6. תהא $V \in \mathbb{R}^{n\times n}$ מטריצה לא סינגולרית ו-$\boldsymbol{y} \in \mathbb{R}^n$. נגדיר את הטרנספורמציה $T_{V,\boldsymbol{y}} : \mathbb{R}^n \to \mathbb{R}^n$ על ידי $T_{V,\boldsymbol{y}}(\boldsymbol{x}) = V\boldsymbol{x} + \boldsymbol{y}$. טרנספורמציה כזו נקראת **טרנספורמציה (העתקה) אפינית**.

הערה 6.7. נשים לב שלהעתקה זו יש הופכית: $T^{-1}_{V,\boldsymbol{y}}(\boldsymbol{x}) = V^{-1}(\boldsymbol{x} - \boldsymbol{y}) = V^{-1}\boldsymbol{x} - V^{-1}\boldsymbol{y}$. נשים לב ש-$V^{-1}\boldsymbol{y}$ הוא וקטור קבוע, ואז נקבל ש-$T^{-1}_{V,\boldsymbol{y}} = T_{V^{-1},-V^{-1}\boldsymbol{y}}$.
המסקנה היא שהופכית של העתקה אפינית היא העתקה אפינית.

טענה 6.8. יהיו X ו-$T_{V,\boldsymbol{y}}$ העתקה אפינית. אז:

- אם $X \subseteq X' \subseteq \mathbb{R}^n$ אז $T_{V,\boldsymbol{y}}(X) \subseteq T_{V,\boldsymbol{y}}(X')$.
- אם X תת מרחב, ו-$\dim X = n$ ו-X קמור, אז $\mathrm{vol}\left(T_{V,\boldsymbol{y}}(X)\right) = |\det V| \,\mathrm{vol}(X)$ (כלומר העתקות לינאריות משמרות נפחים עד כדי מכפלה בקבוע שהוא הערך המוחלט של הדטרמיננטה של המטריצה בהגדרת ההעתקה האפינית).

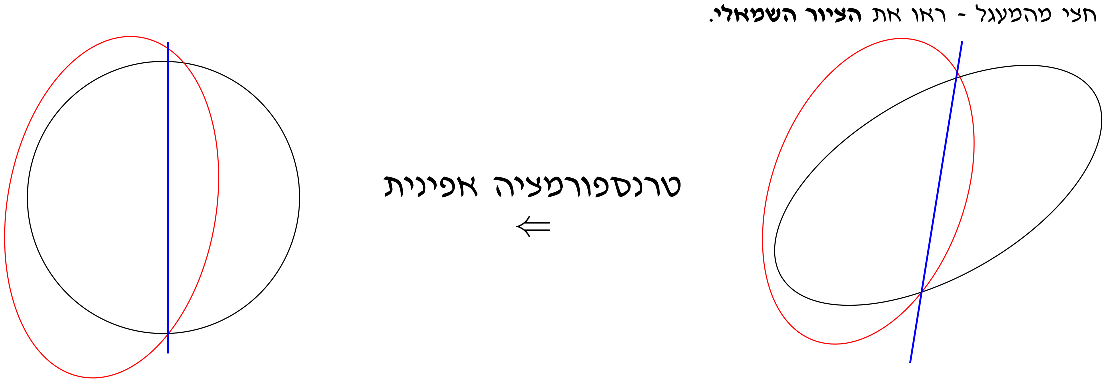
הערה. **סימון:**

- לעיתים נסמן עבור $\Lambda = \begin{bmatrix} \lambda_1^2 & 0 & 0 \\ & \ddots & \\ 0 & & 0 \\ 0 & 0 & \lambda_n^2 \end{bmatrix}$ את $\sqrt{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ & \ddots & \\ 0 & & 0 \\ 0 & 0 & \lambda_n \end{bmatrix}$.
- נסמן ב-$\left\{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x}^T\boldsymbol{x} \le 1\right\} = \left\{\boldsymbol{x} \in \mathbb{R}^n \mid (\boldsymbol{x}-\boldsymbol{0})^T I^{-1}(\boldsymbol{x}-\boldsymbol{0}) \le 1\right\} = B(\boldsymbol{0},1) = B_n$ את **כדור היחידה הסגור** ממימד n.

טענה 6.9. תהא $E = E(D,\boldsymbol{y}) = \left\{\boldsymbol{x} \in \mathbb{R}^n \mid (\boldsymbol{x}-\boldsymbol{y})^T D^{-1}(\boldsymbol{x}-\boldsymbol{y}) \le 1\right\}$, עבור $D = V\Lambda V^T$, $\Lambda = \begin{bmatrix} \lambda_1^2 & 0 & 0 \\ & \ddots & \\ 0 & & 0 \\ 0 & 0 & \lambda_n^2 \end{bmatrix}$.
אזי מתקיים ש-$E = T_{W,\boldsymbol{y}}(B_n)$ עבור $W = V\sqrt{\Lambda}$ (במילים אחרות $D = WW^T$).

כלומר כל אליפסואיד יכול להיות מוגדר על ידי טרנספורמציה אפינית של כדור היחידה הסגור.

מסקנה 6.10. **בציור הימני** נתון האליפסואיד השחור E. חתכנו אותו (בקו הכחול) לחצי, וציירנו אליפסואיד אדום שמכיל את החצי השמאלי. מה אפשר להגיד על הנפח שלו? (ביחס לנפח האליפסואיד השחור)
מהטענה האחרונה ומִמה שהראנו שכל טרנספורמציה אפינית היא הפיכה, אפשר להפעיל טרנספורמציה אפינית הופכית ל-$T_{W,\boldsymbol{y}}$ ($T_{W^{-1},W^{-1}\boldsymbol{y}}$) שתתן לנו את כדור היחידה. האליפסואיד האדום יומר גם הוא לאיזשהו אליפסואיד חדש, שעדיין מכיל חצי מהמעגל - ראו את **הציור השמאלי**.

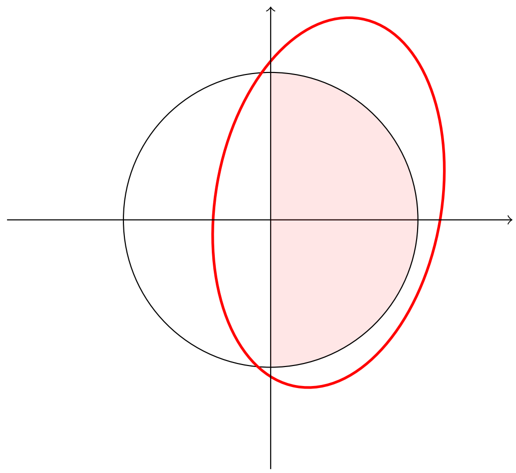


בגלל שהטרנספורמציה אפינית, אנחנו נקבל **שהיחס בין הנפחים** נשמר. אז, המסקנה היא שמספיק להוכיח את הטענה לגבי הנפחים של האליפסואידים ($\mathrm{vol}(\cdots) \le (1-\varepsilon)\,\mathrm{vol}(\cdots)$) כאשר מניחים בה"כ ש-$E$ הוא כדור היחידה.

בנוסף, כל החיתוכים של כדור היחידה לחצי הם זהים עד כדי סיבוב (בניגוד למשל לאליפסואיד כללי) ולכן מספיק להוכיח את הטענה לגבי חיתוך אחד ספציפי של כדור היחידה.

משפט 6.11. יהא $E=E\left(D,\boldsymbol{y}\right)$ אליפסואיד, ו-$F=E\cap\left\{ \boldsymbol{x}\in\mathbb{R}^{n}\mid\boldsymbol{a}^{T}\boldsymbol{x}\leq\boldsymbol{a}^{T}\boldsymbol{y}\right\}$ לאיזשהו $\boldsymbol{a}\in\mathbb{R}^{n}\setminus\left\{ \boldsymbol{0}\right\}$ (כלומר F הוא חצי אליפסואיד).
אזי קיים אליפסואיד E' ככה ש-$E'\supseteq F$, וגם $\frac{\mathrm{vol}\left(E'\right)}{\mathrm{vol}\left(E\right)}\leq e^{-\frac{1}{2(n+1)}}$ (מתקיים $e^{-\frac{1}{2(n+1)}}\approx1-\frac{1}{2(n+1)}$ ולכן ε שרצינו מקודם יכול להיות בערך $\frac{1}{2(n+1)}$).

הוכחה. כאמור במסקנה 6.10 מספיק להוכיח עבור $E=B_{n}$. לא צריך להסתכל על כל ה-$\boldsymbol{a}$-ים האפשריים לחיתוך - הסברנו את זה גם במסקנה 6.10.
אז אנחנו נבחר את החיתוך לפי הציר הראשון: $F=B_{n}\cap\left\{ \boldsymbol{x}\in\mathbb{R}^{n}\mid x_{1}\geq0\right\}$. צריך למצוא אליפסואיד E' שמכיל את F. המחשה ב-$\mathbb{R}^{2}$ (האליפסואיד שצריך למצוא באדום, F צבועה באדום בהיר):



ההוכחה קונסטרוקטיבית: נביא נוסחה ל-E' ונראה שהיא מקיימת את הנדרש.
נגדיר את המרכז שלו $\boldsymbol{y}$ להיות $\frac{1}{n+1}\boldsymbol{e_{1}}=\begin{pmatrix}\frac{1}{n+1}\\0\\\vdots\\0\end{pmatrix}$ (כאשר $\boldsymbol{e_{1}}=\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}$), ו-$D$ היא $D=\frac{n^{2}}{n^{2}-1}\left(I-\frac{2}{n+1}\boldsymbol{e_{1}}\boldsymbol{e_{1}}^{T}\right)$ (שימו לב ש-$\boldsymbol{e_{1}}\boldsymbol{e_{1}}^{T}$ זה מטריצה).
כלומר D היא המטריצה:

$$\begin{aligned}D & =\frac{n^{2}}{n^{2}-1}\left(I-\frac{2}{n+1}\boldsymbol{e_{1}}\boldsymbol{e_{1}}^{T}\right)=\\
 & =\frac{n^{2}}{n^{2}-1}\left(I-\begin{bmatrix}\frac{2}{n+1} & 0 & 0 & 0\\
0 & 0 & 0 & 0\\
0 & 0 & \ddots & 0\\
0 & 0 & 0 & 0
\end{bmatrix}\right)=\\
 & =\frac{n^{2}}{n^{2}-1}\cdot\begin{bmatrix}1-\frac{2}{n+1} & 0 & 0 & 0\\
0 & 1 & 0 & 0\\
0 & 0 & \ddots & 0\\
0 & 0 & 0 & 1
\end{bmatrix}=\\
 & =\begin{bmatrix}\left(\frac{n}{n+1}\right)^{2} & 0 & 0 & 0\\
0 & \frac{n^{2}}{n^{2}-1} & 0 & 0\\
0 & 0 & \ddots & 0\\
0 & 0 & 0 & \frac{n^{2}}{n^{2}-1}
\end{bmatrix}
\end{aligned}$$

קל לראות שזו מטריצה סימטרית מוגדרת חיובית (הערכים העצמיים, שהם גם האיברים על האלכסון במקרה הזה, כולם חיוביים).

עלינו להראות תחילה ש-E' המוגדרת על ידי $\left\{\boldsymbol{x} \in \mathbb{R}^n \mid (\boldsymbol{x}-\boldsymbol{y})^T D^{-1} (\boldsymbol{x}-\boldsymbol{y}) \le 1\right\}$, אכן מקיימת $F \subseteq E'$. כלומר שכל נקודה בכדור היחידה הסגור B_n, עם קורדינטה ראשונה אי שלילית, מקיימת את המשוואה שמאפיינת את E'. הרצאה 9

יהא $\boldsymbol{x} \in F$. עלינו להראות ש-$\left(\boldsymbol{x}-\frac{1}{n+1}\boldsymbol{e_1}\right)^T D^{-1}\left(\boldsymbol{x}-\frac{1}{n+1}\boldsymbol{e_1}\right) \le 1$.

נבחין כי $D^{-1} = \begin{bmatrix} \left(\frac{n+1}{n}\right)^2 & 0 & 0 & 0 \\ 0 & \frac{n^2-1}{n^2} & 0 & 0 \\ & & \ddots & \\ 0 & 0 & & 0 \\ 0 & 0 & 0 & \frac{n^2-1}{n^2} \end{bmatrix}$. בגלל שפרט לאלכסון כל האיברים של D^{-1} הם אפסים, מתקיים:

$$\begin{aligned}
& \left(\boldsymbol{x}-\tfrac{1}{n+1}\boldsymbol{e_1}\right)^T D^{-1}\left(\boldsymbol{x}-\tfrac{1}{n+1}\boldsymbol{e_1}\right) = \\
&= \sum_{i=1}^{n} \left(x_i - (\boldsymbol{e_1})_i\right)^2 \left(D^{-1}\right)_{i,i} = \\
&= \left(x_1 - \tfrac{1}{n+1}\right)^2 \left(\tfrac{n+1}{n}\right)^2 + \sum_{i=2}^{n} x_i^2 \tfrac{n^2-1}{n^2} = \\
&= \left(x_1 - \tfrac{1}{n+1}\right)^2 \left(\tfrac{n+1}{n}\right)^2 - x_1^2\tfrac{n^2-1}{n^2} + \sum_{i=1}^{n} x_i^2 \tfrac{n^2-1}{n^2} = \\
&= \left(\tfrac{n+1}{n}\right)^2 x_1^2 - \tfrac{2}{n+1}\cdot\left(\tfrac{n+1}{n}\right)^2 \cdot x_1 + \left(\tfrac{1}{n+1}\right)^2 \cdot \left(\tfrac{n+1}{n}\right)^2 - x_1^2\tfrac{n^2-1}{n^2} + \sum_{i=1}^{n} x_i^2 \tfrac{n^2-1}{n^2} = \\
&= \tfrac{1}{n^2} - \tfrac{2(n+1)}{n^2}x_1 + \left(\tfrac{n^2+2n+1}{n^2} - \tfrac{n^2-1}{n^2}\right)x_1^2 + \sum_{i=1}^{n} x_i^2 \tfrac{n^2-1}{n^2} = \\
&= -\tfrac{2(n+1)}{n^2}x_1 + \left(\tfrac{n^2+2n+1-n^2+1}{n^2}\right)x_1^2 + \sum_{i=1}^{n} x_i^2 \tfrac{n^2-1}{n^2} + \tfrac{1}{n^2} = \\
&= -\tfrac{2(n+1)}{n^2}x_1 + \tfrac{2(n+1)}{n^2}x_1^2 + \tfrac{n^2-1}{n^2}\underbrace{\sum_{i=1}^{n} x_i^2}_{\le 1,\ \boldsymbol{x}\in B_n} + \tfrac{1}{n^2} \le \\
&\le \underbrace{\frac{2(n+1)}{n^2}\overbrace{\left(x_1^2 - x_1\right)}^{\le 0}}_{\boldsymbol{x}\in B_n \text{ כי } x_i \in [0,1]} + \underbrace{\frac{n^2-1}{n^2} + \frac{1}{n^2}}_{1} = \\
\left(\forall 0 \le x \le 1,\ x \ge x^2\right) \quad & \le 1
\end{aligned}$$

כפי שרצינו להראות.

כעת נראה ש-$\mathrm{vol}\left(E'\right) \le e^{-\frac{1}{2(n+1)}} \cdot \mathrm{vol}\left(B_n\right)$. ידוע לנו כי:

$$\begin{aligned}
|\det W| &= \left|\det\left(V\sqrt{\Lambda}\right)\right| = \overbrace{|\det V|}^{1}\cdot\left|\det\sqrt{\Lambda}\right| = \sqrt{\det\Lambda} = \sqrt{\det D} = \\
&= \sqrt{\det\begin{bmatrix} \left(\frac{n}{n+1}\right)^2 & 0 & 0 & 0 \\ 0 & \frac{n^2}{n^2-1} & 0 & 0 \\ & & \ddots & \\ 0 & 0 & & 0 \\ 0 & 0 & 0 & \frac{n^2}{n^2-1} \end{bmatrix}} = \\
&= \sqrt{\left(\frac{n}{n+1}\right)^2 \cdot \left(\frac{n^2}{n^2-1}\right)^{n-1}} = \\
&= \sqrt{\left(\left(\frac{n}{n+1}\right)^1 \cdot \left(\frac{n^2}{n^2-1}\right)^{\frac{n-1}{2}}\right)^2} =
\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{n}{n+1}\right)^{1}\cdot\left(\frac{n^2}{n^2-1}\right)^{\frac{n-1}{2}} =\\
&= \left(1-\frac{1}{n+1}\right)^{n+1\cdot\frac{1}{n+1}}\cdot\left(1+\frac{1}{n^2-1}\right)^{\left(n^2-1\right)\frac{n-1}{2\left(n^2-1\right)}} =\\
&= \underbrace{\left(\left(1+\frac{-1}{n+1}\right)^{n+1}\right)^{\frac{1}{n+1}}}_{\leq e^{-1}}\cdot\underbrace{\left(\left(1+\frac{1}{n^2-1}\right)^{\left(n^2-1\right)}\right)^{\frac{n-1}{2\left(n^2-1\right)}}}_{\leq e} =\\
&\leq e^{-\frac{1}{n+1}+\frac{n-1}{2\left(n^2-1\right)}} =\\
&= e^{-\frac{1}{n+1}+\frac{1}{2(n+1)}} =\\
&= e^{-\frac{1}{2(n+1)}}
\end{aligned}
$$

כמו שרצינו להראות. $\square$

מסקנה 6.12. בכל איטרציה של אלגוריתם האליפסואיד, הנפח יורד בפקטור של $e^{-\frac{1}{2(n+1)}}$. למשל אחרי $2\left(n+1\right)$ איטרציות, הנפח ירד פי e^{-1} סך הכל. נקודת ההתחלה שלנו היא $B\left(\mathbf{0},R\right)$, ומשום שאנחנו מניחים שהנפח של P מקיים $\text{vol}\left(P\right)\geq\text{vol}\left(B\left(\mathbf{0},r\right)\right)$ - נקודת העצירה היא $\text{vol}\left(B\left(\mathbf{0},r\right)\right)$.

אבל $\left(\frac{R}{r}\right)^{n}=\frac{\text{vol}(B(\mathbf{0},R))}{\text{vol}(B(\mathbf{0},r))}$. אז כמה איטרציות זה יקח לנו?

i צריך לקיים $\underbrace{\text{vol}\left(B\left(\mathbf{0},r\right)\right)}_{\text{נפח סופי מקסימלי}}\leq\underbrace{\text{vol}\left(B\left(\mathbf{0},R\right)\right)}_{\text{נפח התחלתי}}\cdot\left(e^{-\frac{1}{2(n+1)}}\right)^{i}$, כלומר $\frac{\text{vol}(B(\mathbf{0},r))}{\text{vol}(B(\mathbf{0},R))}\leq e^{-\frac{i}{2(n+1)}}$.

דרך הפיכת שני האגפים זה שקול אל $\frac{\text{vol}(B(\mathbf{0},R))}{\text{vol}(B(\mathbf{0},r))}\geq e^{\frac{i}{2(n+1)}}$. נקבל שמתקיים $i\geq2\left(n+1\right)\cdot\ln\left(\frac{\text{vol}(B(\mathbf{0},R))}{\text{vol}(B(\mathbf{0},r))}\right)$.

כלומר, $i\geq2\left(n+1\right)\ln\left(\left(\frac{R}{r}\right)^{n}\right)=2n\left(n+1\right)\ln\left(\frac{R}{r}\right)$ - דרושות לכל הפחות $2n\left(n+1\right)\ln\left(\frac{R}{r}\right)$ איטרציות כדי לעצור.

לכן: אם לא הגענו לערך הזה תוך i איטרציות, זו הוכחה שלא מתקיים $\text{vol}\left(P\right)\geq\text{vol}\left(B\left(\mathbf{0},r\right)\right)$, ואז האלגוריתם נכשל.

אז איך נקבל פתרון מקסימלי?

טענה (ללא הוכחה). לכל $0<\varepsilon$, אם $B\left(\boldsymbol{x_0},r\right)\subseteq P_{<}\subseteq B\left(\mathbf{0},R\right)$, אז לכל היותר אחרי $N=2n\left(n+1\right)\left\lceil\ln\frac{2R^2||\boldsymbol{c}||}{r\cdot\varepsilon}\right\rceil$ איטרציות, אפשר למצוא פתרון פיזיבילי $\widehat{\boldsymbol{x}}\in P_{<}$ ככה שמתקיים $\boldsymbol{c}^{T}\widehat{\boldsymbol{x}}\geq\text{OPT}-\varepsilon$.

כרגע אפשר לקבל פתרון שקרוב לאופטימום כמה שאנחנו רק רוצים (עד כדי ε).

אנחנו נרצה להבטיח שאכן קיים פתרון אופטימלי **״יחודי״** ככה שהוא פתרון בסיסי, כלומר קודקוד. ואז נניח שניקח מספר ε קטן מספיק, נוכל למצוא את הקודקוד הזה. צריכים להיות מספיק קרובים לקודקוד (ולא לקודקודים אחרים) כדי לעשות את זה (זה גם דורש לתת את הדעת לגבי הייצוג של ε בתוכנית מחשב: הגודל שלו צריך להיות פולינומי בגודל הקלט). **נתן את הדעת על כך בהמשך.**

בגדול, אפשר לעשות את זה בכל מיני שיטות (נזכור שאם $A,\boldsymbol{b}$ רציונליים אז גם הקודקודים רציונליים, כי הם מפגשים בין אילוצים).

איך מבטיחים את זה? נבצע טריקים לתוכנית המקורית: בגלל שכל פתרון אופטימלי מגדיר בסיס (תת קבוצה של מספרי הקורדינטות), כפי שראינו בעבר, אנחנו הולכים למשקל את הקורדינטות בבסיס ולהגדיר $\boldsymbol{c}$ חדש:

נקבע $q>1$ קבוע רציונלי (הוא תלוי בגודל התוכנית הלינארית, ברמת העיקרון) - אבל בכל זאת מספר הביטים שלו צריך להיות פולינומי בגודל הקלט. אנחנו נחליף את $\boldsymbol{c}$ ב-$\widetilde{\boldsymbol{c}}=\boldsymbol{c}\cdot q^{n}+\left(1,q,q^2,...,q^{n-1}\right)^{T}$ (נשים לב שמספר הביטים של q הוא בקירוב $\log\left(q\right)$ לכן מספר הביטים של q^{n} למשל הוא $n\log\left(q\right)$ - כלומר אנחנו עדיין שומרים על הפולינומיות של הקלט).

כעת הבעיה היא $\begin{array}{ll} \max & \tilde{\boldsymbol{c}}^T \boldsymbol{x} \\ \text{s.t.} & A\boldsymbol{x} \le \boldsymbol{b} \end{array}$. נשים לב שקודקוד שלא אופטימלי בבעיה המקורית, לא הופך לאופטימלי בבעיה החדשה. ונוסף על כך, אילו היו כמה קודקודים אופטימליים כשהבעיה הייתה מקסימיזציה תחת $\boldsymbol{c}$, עכשיו המשקול של הקורדינטות הופך אותם ללא אופטימליים. מצד שני, יש לנו הבטחה שעדיין יש איזשהו קודקוד אופטימלי (ראינו שלכל בעיה חסומה יש פתרון אופטימלי שהוא קודקוד) - על כן בהכרח הקודקוד הזה היה אופטימלי גם בבעיה תחת $\boldsymbol{c}$, וכעת הוא היחיד בבעיה תחת $\tilde{\boldsymbol{c}}$.

6.3 סיבוכיות אלגוריתם האליפסואיד

הרצאה 10

תהא תוכנית לינארית בהצגתה הסטנדרטית $\begin{smallmatrix} \min \boldsymbol{c}^T\boldsymbol{x} \\ A\boldsymbol{x}=\boldsymbol{b} \\ \boldsymbol{x}\ge\boldsymbol{0} \end{smallmatrix}$, כך ש-$A \in \mathbb{Q}^{m\times n}$, $\boldsymbol{b} \in \mathbb{Q}^m$ ו-$\boldsymbol{c} \in \mathbb{Z}^n$.

משום ש-A מטריצה רציונלית, כל איבר בה ניתן לרשום כמנה של שני שלמים. כלומר ישנן $A^{(1)}, A^{(2)} \in \mathbb{Z}^{m\times n}$ (במטריצה $A^{(2)}$ אין אפסים) כך שמתקיים $A_{i,j} = \frac{A^{(1)}_{i,j}}{A^{(2)}_{i,j}}$ לכל i,j.

אותו דבר לגבי $\boldsymbol{b}$ ו-$\boldsymbol{c}$. אפשר למצוא וקטורים $\boldsymbol{b^{(1)}}, \boldsymbol{b^{(2)}} \in \mathbb{Z}^m$ ו-$\boldsymbol{c^{(1)}}, \boldsymbol{c^{(2)}} \in \mathbb{Z}^n$ כך ש-$c_i = \frac{c^{(1)}_i}{c^{(2)}_i}$ ו-$b_i = \frac{b^{(1)}_i}{b^{(2)}_i}$.

נשים לב שכל הוקטורים מכילים מספרים וגודל הקלט יצטרך להיות גודל הקלט **הבינארי**. כדי לייצג מספר $a \in \mathbb{Z}$, אנחנו צריכים $\lceil \log_2 |a| \rceil + 1$ ביטים. **נקרא לגודל הזה ה-**$\text{size}\,(a)$**, גודל הייצוג של מספר שלם.**

גודל הייצוג של רציונלי $x = \frac{p}{q}$ הולך להיות $\text{size}\,(x) = \text{size}\,(p) + \text{size}\,(q)$.

עבור וקטור $\boldsymbol{v} \in \mathbb{Z}^d$ אנחנו נגדיר $\text{size}\,(\boldsymbol{v}) = \sum_{i=1}^{d} \text{size}\,(v_i)$ ונקרא לזה **גודל הייצוג של הוקטור**.

עבור מטריצה $M \in \mathbb{Z}^{d\times f}$ אנחנו נגדיר $\text{size}\,(M) = \sum_{i=1}^{d}\sum_{j=1}^{f} \text{size}\,(M_{i.j})$. שימו לב שבפועל צריך גם לייצג את מימדי המטריצה (הם באים יחד עם הקלט) d ו-f.

עבור A שלנו, שמפוצלת ל-2, נגדיר $\text{size}\,(A) = \text{size}\left(A^{(1)}\right) + \text{size}\left(A^{(2)}\right)$. באופן דומה נגדיר כך עבור $\boldsymbol{b}, \boldsymbol{c}$ גם (סכום שני ה-sizeים).

בסך הכל בהינתן תוכנית לינארית $\text{LP} = \begin{smallmatrix} \min \boldsymbol{c}^T\boldsymbol{x} \\ A\boldsymbol{x}=\boldsymbol{b} \\ \boldsymbol{x}\ge\boldsymbol{0} \end{smallmatrix}$ גודל הקלט שלה היא $\text{size}\,(\text{LP}) \ge \text{size}\,(A) + \text{size}\,(\boldsymbol{b}) + \text{size}\,(\boldsymbol{c})$ (כאמור יש גם ייצוגים שאנחנו צריכים למימדים כמו m, n ועוד כמה פרמטרים לכן זה $\ge$ ולא $=$).

אפשר לנתח אלגוריתמים על סמך הגודל $\text{size}\,(A) + \text{size}\,(b) + \text{size}\,(c)$. אבל זה גודל מעצבן, אז מה שנעשה יהיה להגדיר גודל אחר לתוכנית לינארית.

הגדרה 6.13. תהא תוכנית לינארית $\text{LP} = \begin{smallmatrix} \min \boldsymbol{c}^T\boldsymbol{x} \\ A\boldsymbol{x}=\boldsymbol{b} \\ \boldsymbol{x}\ge\boldsymbol{0} \end{smallmatrix}$. נגדיר $\det_{\max} = \max_{\substack{A' \text{ תת מטריצה} \\ \text{ריבועית של } A}} |\det A'|$, $b_{\max} = \max_{1\le i\le m} |b_i|$ ו-$c_{\max} = \max_{1\le i\le n} |c_i|$.

הגודל של התוכנית הלינארית מוגדר להיות $L = \text{size}\,(\det_{\max}) + \text{size}\,(b_{\max}) + \text{size}\,(c_{\max}) + m + n$.

מעתה נשתמש ב-L כדי לסמן גודל תוכנית לינארית.

הערה 6.14. לגבי גודל של תוכניות לינאריות:

- ○ טענה שאפשר להוכיח היא ש-$L \le \text{size}\,(A) + \text{size}\,(\boldsymbol{b}) + \text{size}\,(\boldsymbol{c})$.
- ○ נשים לב שמההגדרה כל מספר בקלט (בערכו המוחלט) הוא מהצורה $\frac{p}{q}$ כאשר $p, q \in \mathbb{Z}$, כך ש-$0 \le p < 2^L$ ו-$1 \le q < 2^L$.

משפט 6.15. יהא $\boldsymbol{x}$ קודקוד של $P = \{\boldsymbol{x} \in \mathbb{R}^n \mid A\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \ge \boldsymbol{0}\}$. אזי $\boldsymbol{x}$ הוא רציונלי, ובפרט אפשר לרשום $\boldsymbol{x}^T = \left(\frac{p_1}{q}, ..., \frac{p_n}{q}\right)$ (מכנה משותף) ככה ש-$p_1, ..., p_n, q \in \mathbb{Z}$ וגם $0 \le p_1 < 2^L$ ו-$1 \le q < 2^L$.

שימו לב שהמשפט הזה חשוב: הוא מבטיח לנו שגודל הפלט של אלגוריתם האליפסואיד (הפלט הוא תמיד קודקוד) הוא פולינומי בגודל הקלט. זו כמובן אחת מהדרישות ההכרחיות לכך שהאלגוריתם יעבוד בזמן פולינומי (אלגוריתם לא יכול

לעבוד בזמן פולינומי אם הפלט שלו לא פולינומי...).

דיברנו על זה שאלגוריתם האליפסואיד מקרב אותנו לפתרון אופטימלי, וכשהוא מספיק קרוב - הוא יודע למצוא אותו. ובדיוק בעזרת המשפט הנ"ל הוא יודע למצוא את הקודקוד (רק לקודקוד יש את הייצוג הנ"ל): $\boldsymbol{x}$ קודקוד לכן הוא נתון על בסיס כלשהו $B \subseteq \{1, ..., n\}$ (כלומר $A_{(B)}\boldsymbol{x}_{(\boldsymbol{B})} = \boldsymbol{b}$. ובה"כ B בגודל m אחרת נוסיף לו קורדינטות), ומפה אפשר לחשבו על ידי $\boldsymbol{x}_{(\boldsymbol{B})} = A_{(B)}^{-1}\boldsymbol{b}$ (ככה מחשבים את הקורדינטות ב-B, כל שאר הקורדינטות שאינן בבסיס הן 0).

רעיון ההוכחה הוא שאם יש לנו את הקודקוד $\boldsymbol{x}$, אפשר לקבל בסיס B, ואז יתקיים $\boldsymbol{x}_{(\boldsymbol{B})} = A_{(B)}^{-1}\boldsymbol{b}$ ואז עם כל מיני נוסחאות של מטריצות הופיכות, נקבל משהו כמו $\boldsymbol{x}_{(\boldsymbol{B})} = \frac{1}{\left|\det A_{(B)}\right|} \cdot ....$.
ה-q שלנו יהיה בדיוק $\left|\det A_{(B)}\right|$.

הערה 6.16. נדבר על הבעיה של חסימת האופטימום של פתרון תוכניות לינאריות (LP): "בהינתן תוכנית לינארית וערך m, האם אפשר למצוא פתרון שערך פונקציית המטרה עליו לכל היותר m?". כמובן שאנחנו עובדים עם מחשב (או מכונת טיורינג) לכן אפשר להניח שכל הקלטים הם רציונליים.
מה שאומר לנו משפט 6.15 הוא למעשה שלבעיות לינאריות יש פתרון אופטימלי **שהוא פולינומי** באורך הקלט. כלומר תיאורטית מישהו יכול להביא לי $\boldsymbol{x}$, ואני רק צריך לבדוק האם $A\boldsymbol{x} = \boldsymbol{b}$ וגם $\boldsymbol{x} \geq \boldsymbol{0}$ (שאלה דברים שאפשר לבדוק בזמן פולינומי) כדי להכריע האם $\boldsymbol{x}$ הוא אכן פתרון טוב. במילים האחרות לבעיה הנתונה שלנו יש עד פולינומי - ולכן הבעיה הזו היא בעייה ב-NP.
בנוסף בגלל שהבעיה הדואלית היא בעיית תכנון לינארי (שהיא, מהגדרתה, באותו גודל של הבעיה המקורית) שתאפשר לנו לבדוק בדיוק ההפך (בגלל שהיא בעיית מקסימום): "האם יש פתרון שערך פונקציית המטרה עליו שהוא לכל הפחות m?" - נוכל להסיק שבעיית מציאת החסם היא בעייה ב-coNP.
אז הבעיה LP שייכת ל-$\mathtt{coNP} \cap \mathtt{NP}$. למעשה אלגוריתם האליפסואיד מראה שהיא אפילו שייכת ל-P.

טענה 6.17. יהא $P \subseteq \mathbb{R}^n$ פוליטופ (פאון חסום) לא ריק, ככה שקודקודיו ניתנים להצגה בעזרת מספרים רציונליים מהצורה $\frac{p}{q}$, עם $p, q \in \mathbb{Z}$ ו-$|p|, |q| < 2^L$.
אזי קיים $\boldsymbol{a_0} \in P$ עבורו $B(\boldsymbol{a_0}, r) \subseteq P \subseteq B(\boldsymbol{a_0}, R)$ כאשר $r = \left(n2^L\right)^{-2n^2-2n} = \frac{1}{2^{O(n^2L)}}$ ו-$R = 2n2^L$.

מסקנה 6.18. ראינו בעבר (במסקנה 6.12) שמספר האיטרציות הדרושות יהיה המנה $\frac{B(\boldsymbol{a_0},R)}{B(\boldsymbol{a_0},r)}$.
במקרה הזה, זה יוצא $\left(\frac{R}{r}\right)^n$, שזה $\leq \left(2 \cdot \left(n2^L\right)^{1+2n^2+2n}\right)^n = \left(2n^{1+2n^2+2n}2^{L\left(1+2n^2+2n\right)}\right)^n = \left(n2^L\right)^{O\left(n^3\right)}$ $\left(2 \cdot 2^L\right)^{O\left(n^3\right)}$ (מתקיים $n \leq 2^L$, כי בין היתר n מיוצג ב-L).
ראינו שמספר האיטרציות הוא בערך ln של זה, ולכן זה יוצא משהו כמו $O\left(n^3L\right)$.
כמו שהזכרנו הרבה קודם זה מספר האיטרציות שצריך לעשות רק כדי להשיג פתרון פיזיבילי. צריך להוסיף אילוץ שיאפשר לנו לשפר את ערך פונקציית המטרה, וככה הלאה והלאה עד שנהיה קרובים מספיק לקודקוד.

המסקנה החשובה ביותר: אלגוריתם האליפסואיד עובד בזמן פולינומי בגודל הקלט (L, ו-n שכלול גם).

בעיה. בעיה פתוחה: האם קיים אלגוריתם לפתרון תוכניות לינאריות שהוא פולינומי חזק, כלומר לא תלוי בגודל של המספרים בקלט עצמו (L)?

7 אופטימיזציה קמורה

הגדרה 7.1. הקבוצה $D \subseteq \mathbb{R}^n$ תיקרא **קמורה** אם לכל $\boldsymbol{x}, \boldsymbol{y} \in D$ הקטע $[\boldsymbol{x}, \boldsymbol{y}] = \{t\boldsymbol{x} + (1-t)\,\boldsymbol{y} \mid t \in [0,1]\}$ מוכל ב-D.

הגדרה 7.2. אם D קמורה, אז פונקציה $f : D \to \mathbb{R}$ תיקרא **קמורה** אם לכל $\boldsymbol{x}, \boldsymbol{y} \in D$ מתקיים $f\left(\frac{\boldsymbol{x}+\boldsymbol{y}}{2}\right) \le \frac{1}{2}\left(f(\boldsymbol{x}) + f(\boldsymbol{y})\right)$. f נקראת **קמורה ממש** אם אי השוויון מתקיים אך ורק כאשר $\boldsymbol{x} = \boldsymbol{y}$.

בעיה 7.3. תהא $D \subseteq \mathbb{R}^n$ קבוצה חסומה וקמורה. תהא $f : D \to \mathbb{R}^2$ קמורה ממש. אז יש $\boldsymbol{x} \in D$ יחיד ככה ש-$f(\boldsymbol{x}) = \min_{\boldsymbol{y} \in D} f(\boldsymbol{y})$. איך נוכל למצוא את $\boldsymbol{x}$?

נסמן ב-∂D את **השפה (boundary)** של D. אנחנו נרצה להתעניין במקרים בהם $\boldsymbol{x} \in D \setminus \partial D$. זה גורר ש-$\dim D = n$ (dim הוא המימד האפיני כפי שהגדרנו בעבר).

הגדרה 7.4. בהינתן פונקציה $f : D \to \mathbb{R}$, שהיא גזירה ברציפות, מגדירים את ∇f, **הגרדיאנט של** f, להיות פונקציה $D \setminus \partial D \to \mathbb{R}^n$ המוגדרת על ידי:

$$\nabla f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \end{pmatrix}$$

הערה 7.5. אנחנו נניח ש-∇f היא גם L-ליפשיץ עבור איזשהו קבוע L. כלומר לכל $\boldsymbol{x}, \boldsymbol{y} \in D \setminus \partial D$, מתקיים ש-$||\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})||_2 \le L \cdot ||\boldsymbol{x} - \boldsymbol{y}||_2$. האינטואיציה היא ש"הנגזרות לא גדלות מהר מדי": שינוי של ε במרחק לא יגרור שינוי של יותר מ-$L\varepsilon$ בגודל הנגזרת.

הגדרה 7.6. בהינתן פונקציה $f : D \to \mathbb{R}$, שהיא גזירה פעמיים ברציפות, מגדירים את $\nabla^2 f$, **ההסיאן של** f, להיות פונקציה $D \setminus \partial D \to \mathbb{R}^{n \times n}$ (כלומר היא מחזירה מטריצות), ככה ש-$\left(\nabla^2 f(\boldsymbol{x})\right)_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x})$.

הערה. משפט במתמטיקה (משפט קלרו, Clairaut) מבטיח לנו שמתקיים $\frac{\partial^2 f(\boldsymbol{x})}{\partial x_j \partial x_i} = \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}$: כלומר אפשר לגזור לפי x_i ואז לפי x_j או ההפך ולקבל אותה תוצאה. כמסקנה מכך, לכל $\boldsymbol{x} \in D \setminus \partial D$ המטריצה $\nabla^2 f(\boldsymbol{x})$ היא מטריצה סימטרית $n \times n$.

נוסף על כך בגלל ש-f קמורה, מתקיים ש-$\nabla^2 f(\boldsymbol{x})$ מוגדרת חיובית (ראו הגדרה 6.3).

טענה 7.7 (משפט ערך הביניים). תהא $f : D \to \mathbb{R}$ גזירה פעמיים ברציפות. אז לכל $\boldsymbol{x}, \boldsymbol{y} \in D \setminus \partial D$ קיים $\boldsymbol{\xi} \in [\boldsymbol{x}, \boldsymbol{y}] \subseteq D \setminus \partial D$ כך ש-$\nabla f(\boldsymbol{x}) = \nabla f(\boldsymbol{y}) + \nabla^2 f(\boldsymbol{\xi})(\boldsymbol{x} - \boldsymbol{y})$.

מסקנה 7.8. $||\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})||_2 = \left|\left|\nabla^2 f(\boldsymbol{\xi})(\boldsymbol{x} - \boldsymbol{y})\right|\right|_2$.

הנחנו ש-∇f היא L-ליפשיץ לכן $||\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})||_2 \le L \cdot ||\boldsymbol{x} - \boldsymbol{y}||_2$.

כלומר, $\left|\left|\nabla^2 f(\boldsymbol{\xi})(\boldsymbol{x} - \boldsymbol{y})\right|\right|_2 \le L \cdot ||\boldsymbol{x} - \boldsymbol{y}||$. במילים אחרות, **הערכים העצמיים של המטריצה** $\nabla^2 f(\boldsymbol{\xi})$ (שכאמור יש כאלה כי היא סימטרית), **קטנים מ-L.**

7.1 מציאת מינימום

בהינתן פונקציה קמורה f, איך נמצא את $\boldsymbol{x} = \underset{\boldsymbol{y} \in D \setminus \partial D}{\operatorname{argmin}} f(\boldsymbol{y})$?

שימוש: בלמידת מכונה המטרות הן למזער פונקציית loss שאומרות לנו "כמה הפתרון שלנו טוב". שיטות המזעור לפונקציות קמורות משתמשות ב-Gradient Descent או בשיטה כמו Stochastic Gradient Descent.

הפונקציה קמורה, אנחנו נניח שהמינימום שלה בתוך התחום (וגם שאין כל מיני נקודות בהן הנגזרת מתאפסת אבל הנקודה היא לא מינימום, כמו נקודות אוכף). מהקמירות נובע שיש מינימום אחד.

ממשפט באינפי, למעשה אנחנו צריכים למצוא $\boldsymbol{x}$ ככה ש-$\nabla f(\boldsymbol{x}) = \boldsymbol{0}$.

לשם כך אפשר לקחת נקודה $\boldsymbol{x}$ ולזוז בכיוון של $-\nabla f\left(\boldsymbol{x}\right)$ (גם פה אנחנו מסתמכים על משפט באינפי שאומר שבנקודה מסויימת, הגרדיאנט הוא וקטור שכיוונו הוא הכיוון בו משתנה הפונקציה בקצב המהיר ביותר. אז אנחנו נזוז בדיוק לכיוון הנגדי לו, כדי למזער את הפונקציה).
הבעיה היא שאנחנו יכולים לזוז יותר מדי. אז צריך לזוז לכיוון $-\nabla f\left(\boldsymbol{x}\right)$ אבל לא יותר מדי, כלומר לזוז ב-$-\varepsilon\nabla f\left(\boldsymbol{x}\right)$. כך הנקודה החדשה שלנו אליה נזוז תהיה $\boldsymbol{x}-\varepsilon\nabla f\left(\boldsymbol{x}\right)$. כאמור, צריך לדעת איזה ε לבחור, כדי שלא נזוז יותר מדי.

הנחה מקלה: נניח כי f היא קמורה ממש, אז יש $\mu>0$ כך ש-$\forall\boldsymbol{x}\in D\setminus\partial D$, כל הערכים העצמיים של $\nabla f^{2}\left(\boldsymbol{x}\right)$ הם בקטע $\left[\mu,L\right]$ (זה שונה ממה שהיה לנו עכשיו, כי הם היו בקטע $\left[0,L\right]$).
שימו לב שכעת $\nabla f^{2}\left(\boldsymbol{x}\right)$ היא מטריצה מוגדרת חיובית ממש.

טענה 7.9 (פיתוח טיילור). תהא $f:D\rightarrow\mathbb{R}$ גזירה. אז לכל $\boldsymbol{x}\in D$ ולכל $\boldsymbol{y}\in D$ יש $\boldsymbol{\xi}\in\left[\boldsymbol{x},\boldsymbol{y}\right]$ כך ש:

$$f\left(\boldsymbol{y}\right)=f\left(\boldsymbol{x}\right)+\left(\boldsymbol{y}-\boldsymbol{x}\right)^{T}\nabla f\left(\boldsymbol{x}\right)+\tfrac{1}{2}\left(\boldsymbol{y}-\boldsymbol{x}\right)^{T}\nabla^{2}f\left(\boldsymbol{\xi}\right)\left(\boldsymbol{y}-\boldsymbol{x}\right)$$

ישנו משפט בשם "עיקרון ריילי" שאומר ש-$\frac{1}{2}\boldsymbol{v}^{T}\nabla^{2}f\left(\boldsymbol{t}\right)\left(\boldsymbol{v}\right)$, חסום מלמעלה על ידי הערך העצמי הגדול ביותר של $\nabla^{2}f\left(\boldsymbol{t}\right)$ כפול $||\boldsymbol{v}||$, וחסומה מלמטה אותו דבר לגבי הערך העצמי הקטן ביותר. במקרה שלנו כל הערכים העצמיים הם בין μ ו-L, אז נקבל:

$$\begin{aligned}
f\left(\boldsymbol{y}\right) & = f\left(\boldsymbol{x}\right)+\left(\boldsymbol{y}-\boldsymbol{x}\right)^{T}\nabla f\left(\boldsymbol{x}\right)+\tfrac{1}{2}\left(\boldsymbol{y}-\boldsymbol{x}\right)^{T}\nabla^{2}f\left(\boldsymbol{\xi}\right)\left(\boldsymbol{y}-\boldsymbol{x}\right)\leq\\
 & \leq f\left(\boldsymbol{x}\right)+\left(\boldsymbol{y}-\boldsymbol{x}\right)^{T}\nabla f\left(\boldsymbol{x}\right)+\tfrac{L}{2}\left|\left|\boldsymbol{y}-\boldsymbol{x}\right|\right|^{2}\\
f\left(\boldsymbol{y}\right) & \geq f\left(\boldsymbol{x}\right)+\left(\boldsymbol{y}-\boldsymbol{x}\right)^{T}\nabla f\left(\boldsymbol{x}\right)+\tfrac{\mu}{2}\left|\left|\boldsymbol{y}-\boldsymbol{x}\right|\right|^{2}
\end{aligned}$$

נניח שאנחנו ב-$\boldsymbol{x}$ ורוצים לזוז ל-$\boldsymbol{y}$. נובע שהשינוי האפשרי, **חסום מלמטה** על ידי

$$f\left(\boldsymbol{y}\right)-f\left(\boldsymbol{x}\right)\quad\geq\quad\left(\boldsymbol{y}-\boldsymbol{x}\right)^{T}\nabla f\left(\boldsymbol{x}\right)+\tfrac{\mu}{2}\left|\left|\boldsymbol{y}-\boldsymbol{x}\right|\right|^{2}$$

רוצים שהשינוי יהיה קטן בכל האפשר באגף שמאל (שיהיה שלילי כמה שיותר). ננסה למזער את אגף ימין.
אם כך מיהו $\underset{\boldsymbol{y}}{\operatorname{argmin}}\left\{ \left(\boldsymbol{y}-\boldsymbol{x}\right)^{T}\nabla f\left(\boldsymbol{x}\right)+\frac{\mu}{2}\left|\left|\boldsymbol{y}-\boldsymbol{x}\right|\right|^{2}\right\}$? צריך לגזור את הפונקציה ולמצוא מינימום כך.
הנגזרת יוצאת $\nabla f\left(\boldsymbol{x}\right)+\mu\left(\boldsymbol{y}-\boldsymbol{x}\right)$. נשווה אותה ל-$0$ ונקבל $\boldsymbol{y}=\boldsymbol{x}-\frac{1}{\mu}\nabla f\left(\boldsymbol{x}\right)$ - וזה ה-$\boldsymbol{y}$ שממזער את הביטוי.
החסם על הערך שנקבל ב-$\boldsymbol{y}$ יצא:

$$\begin{aligned}
f\left(\boldsymbol{y}\right) & \geq f\left(\boldsymbol{x}\right)+\left(\boldsymbol{y}-\boldsymbol{x}\right)^{T}\nabla f\left(\boldsymbol{x}\right)+\tfrac{\mu}{2}\left|\left|\boldsymbol{y}-\boldsymbol{x}\right|\right|^{2}=\\
 & = f\left(\boldsymbol{x}\right)+\left(\boldsymbol{x}-\tfrac{1}{\mu}\nabla f\left(\boldsymbol{x}\right)-\boldsymbol{x}\right)^{T}\nabla f\left(\boldsymbol{x}\right)+\tfrac{\mu}{2}\left|\left|\boldsymbol{x}-\tfrac{1}{\mu}\nabla f\left(\boldsymbol{x}\right)-\boldsymbol{x}\right|\right|^{2}=\\
 & = f\left(\boldsymbol{x}\right)+\tfrac{1}{\mu}\left(\nabla f\left(\boldsymbol{x}\right)\right)^{T}\nabla f\left(\boldsymbol{x}\right)+\tfrac{\mu}{2}\cdot\tfrac{1}{\mu^{2}}\left|\left|\nabla f\left(\boldsymbol{x}\right)\right|\right|^{2}=\\
 & = f\left(\boldsymbol{x}\right)-\tfrac{1}{\mu}\left|\left|\nabla f\left(\boldsymbol{x}\right)\right|\right|^{2}+\tfrac{1}{2\mu}\left|\left|\nabla f\left(\boldsymbol{x}\right)\right|\right|^{2}=\\
 & = f\left(\boldsymbol{x}\right)-\tfrac{1}{2\mu}\left|\left|\nabla f\left(\boldsymbol{x}\right)\right|\right|^{2}
\end{aligned}$$

אם היינו ב-$\boldsymbol{x^{(t)}}$ ואנחנו זזים אל $\boldsymbol{x^{(t+1)}}=\boldsymbol{x^{(t)}}-\frac{1}{L}\nabla f\left(\boldsymbol{x^{(t)}}\right)$, נשים לב שמתקיים (ממה שהראנו מקודם):

$$\begin{aligned}
f\left(\boldsymbol{x^{(t+1)}}\right) & \leq f\left(\boldsymbol{x^{(t)}}\right)+\left(\boldsymbol{x^{(t)}}-\tfrac{1}{L}\nabla f\left(\boldsymbol{x^{(t)}}\right)-\boldsymbol{x^{(t)}}\right)^{T}\nabla f\left(\boldsymbol{x^{(t)}}\right)+\tfrac{L}{2}\left|\left|\boldsymbol{x^{(t)}}-\tfrac{1}{L}\nabla f\left(\boldsymbol{x^{(t)}}\right)-\boldsymbol{x^{(t)}}\right|\right|^{2}=\\
 & = f\left(\boldsymbol{x^{(t)}}\right)-\tfrac{1}{L}\left|\left|\nabla f\left(\boldsymbol{x^{(t)}}\right)\right|\right|+\tfrac{L}{2}\cdot\tfrac{1}{L^{2}}\left|\left|\nabla f\left(\boldsymbol{x^{(t)}}\right)\right|\right|^{2}=\\
 & = f\left(\boldsymbol{x^{(t)}}\right)-\tfrac{1}{2L}\left|\left|\nabla f\left(\boldsymbol{x^{(t)}}\right)\right|\right|
\end{aligned}$$

המסקנה היא שהשיפור של פונקציית המטרה שלנו, כלומר $f\left(\boldsymbol{x^{(t)}}\right)-f\left(\boldsymbol{x^{(t+1)}}\right)$, מקיים $f\left(\boldsymbol{x^{(t)}}\right)-f\left(\boldsymbol{x^{(t+1)}}\right)\geq\frac{1}{2L}\left|\left|\nabla f\left(\boldsymbol{x^{(t)}}\right)\right|\right|$.
מצד שני אם נסמן $\boldsymbol{y_{\min}}=\underset{\boldsymbol{y}\in D\setminus\partial D}{\operatorname{argmin}}f\left(\boldsymbol{y}\right)$, ממה שראינו קודם $f\left(\boldsymbol{x^{(t)}}\right)-f\left(\boldsymbol{y_{\min}}\right)\leq\frac{1}{2\mu}\left|\left|\nabla f\left(\boldsymbol{x}\right)\right|\right|^{2}$.
אם כך:

$$\begin{aligned} f\left(\boldsymbol{x}^{(t)}\right)-f\left(\boldsymbol{x}^{(t+1)}\right) &\geq \frac{1}{2L}\left|\left|\nabla f\left(\boldsymbol{x}^{(t)}\right)\right|\right|= \\ (\mu>0) \quad &= \frac{\mu}{L}\cdot\frac{1}{2\mu}\left|\left|\nabla f\left(\boldsymbol{x}^{(t)}\right)\right|\right|\geq \\ &\geq \frac{\mu}{L}\cdot\left(f\left(\boldsymbol{x}^{(t)}\right)-f\left(\boldsymbol{y}_{\min}\right)\right) \end{aligned}$$

כלומר בכל איטרציה אנחנו **משתפרים** בלכל היותר מכפלה בקבוע $\frac{\mu}{L}$, של השיפור הכי טוב שאפשר.
מתקיים $f\left(\boldsymbol{x}^{(t)}\right)-f\left(\boldsymbol{y}_{\min}\right)+f\left(\boldsymbol{y}_{\min}\right)-f\left(\boldsymbol{x}^{(t+1)}\right)=f\left(\boldsymbol{x}^{(t)}\right)-f\left(\boldsymbol{x}^{(t+1)}\right)\geq\frac{\mu}{L}\cdot\left(f\left(\boldsymbol{x}^{(t)}\right)-f\left(\boldsymbol{y}_{\min}\right)\right)$.
כלומר $f\left(\boldsymbol{y}_{\min}\right)-f\left(\boldsymbol{x}^{(t+1)}\right)\geq\frac{\mu}{L}\cdot\left(f\left(\boldsymbol{x}^{(t)}\right)-f\left(\boldsymbol{y}_{\min}\right)\right)-\left(f\left(\boldsymbol{x}^{(t)}\right)-f\left(\boldsymbol{y}_{\min}\right)\right)$.
קיבלנו $f\left(\boldsymbol{y}_{\min}\right)-f\left(\boldsymbol{x}^{(t+1)}\right)\geq\left(\frac{\mu}{L}-1\right)\left(f\left(\boldsymbol{x}^{(t)}\right)-f\left(\boldsymbol{y}_{\min}\right)\right)$.
או, אחרי הכפלת שני האגפים ב-(-1) הפיכת הא"ש $f\left(\boldsymbol{y}_{\min}\right)-f\left(\boldsymbol{x}^{(t+1)}\right)\leq\left(1-\frac{\mu}{L}\right)\left(f\left(\boldsymbol{x}^{(t)}\right)-f\left(\boldsymbol{y}_{\min}\right)\right)$.

הרצאה 11

הוכחנו שעבור המינימום f^* מתקיים $f\left(\boldsymbol{x}^{(t+1)}\right)-f^*\leq\left(1-\frac{\mu}{L}\right)\left(f\left(\boldsymbol{x}^{(t)}\right)-f^*\right)$.
נסמן $\kappa=\frac{L}{\mu}$ ואז $\frac{1}{\kappa}=\frac{\mu}{L}$. ברקורסיה נסיק שמתקיים $f\left(\boldsymbol{x}^{(t)}\right)-f^*\leq\left(1-\frac{1}{\kappa}\right)^t\left(f\left(\boldsymbol{x}^{(0)}\right)-f^*\right)$.
וכך, עבור $\varepsilon>0$ אחרי $t=\kappa\ln\left(\varepsilon^{-1}\right)$ צעדים, אנחנו נקבל $f\left(\boldsymbol{x}^{(t)}\right)-f^*\leq\varepsilon\left(f\left(\boldsymbol{x}^{(0)}\right)-f^*\right)$ (בגלל ש- $\left(1-\frac{1}{\kappa}\right)^t$ -> $e^{-\kappa}$ לכל t).

במילים אחרות, אפשר להתקרב למקסימום כרצוננו (עד כדי ε) אחרי מספר כלשהו של צעדים (שאפשר לחסום אותו מלמטה).

מה קורה אם אין את μ? כלומר יש ערך עצמי שהוא 0?

נראה בתרגיל הבית כנראה (עם רמז). הרעיון דומה. אנחנו נקבל קצב גידול שהוא לינארי (משהו כמו $\frac{1}{\varepsilon}$).

8 אלגוריתם הנקודה הפנימית (interior point method)

אנחנו ראינו אלגוריתם פולינומי לפתרון תוכניות לינארית: אלגוריתם האליפסואיד. למעשה בפועל זה אלגוריתם נורא גרוע, ומאוחר יותר התגלה אלגוריתם חדש לפתרון תוכניות לינאריות. האלגוריתם הזה הוא אלגוריתם פולינומי שרץ בזמן טוב.

אפשר לחשוב על אלגוריתם האליפסואיד כ"אלגוריתם נקודה חיצונית" - הוא כל פעם זז מסתמך על נקודות מחוץ לפוליטופ, עד שהוא מוצא נקודה שנמצאת בפנים, וכל פעם הוא זז מחדש (על ידי ציור אליפסואיד אחר). אלגוריתם הסימפלקס זז בין הקודקודים למציאת הפתרון האופטימלי.

אלגוריתם הנקודה הפנימית ימצא את הפתרון האופטימלי (שעל השפה), <u>בעזרת עבודה עם נקודות בתוך הפוליטופ</u>.

נזכיר שאנחנו רוצים למצוא את הפתרון לתוכנית $\begin{array}{ll} \min \boldsymbol{c}^T\boldsymbol{x} \\ \text{s.t. } \boldsymbol{x}\in P \end{array}$ כאשר P היא קבוצה חסומה קמורה ב-$\mathbb{R}^n$ ממימד אפיני מלא (הפעם נכליל ונדבר על P שהיא לא בהכרח פוליטופ, אבל יכולה להיות).

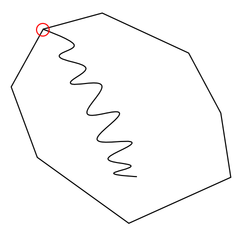
נניח שיש לנו איזושהי **פונקציית מחסום** (barrier function), כלומר פונקציה קמורה $\phi : P\setminus\partial P\to\mathbb{R}$, כך שהמטרה היא שאם נתקרב לקצוות ϕ תקבל ערכים אינסופיים: $\lim_{\boldsymbol{x}\to\partial P}\phi\left(x\right)=\infty$ (כאשר $x\to\partial P$ משמעותו $\boldsymbol{y}\to\boldsymbol{x}$ כאשר $\boldsymbol{y}\in\partial P$).

יהא $t\in[0,\infty)$. נגדיר $\boldsymbol{x}^{(\boldsymbol{t})}=\underset{\boldsymbol{x}}{\operatorname{argmin}}\left(t\cdot\boldsymbol{c}^T\boldsymbol{x}+\phi\left(\boldsymbol{x}\right)\right)$. ככל ש-$t$ גדל, אנחנו ממשקלים יותר את פונקציית המטרה שלנו $\boldsymbol{c}^T\boldsymbol{x}$, במקום את $\phi\left(\boldsymbol{x}\right)$. לכן ככל ש-$t$ גדל אנחנו מרשים לפתרונות של בעיית האופטימיזציה $t\cdot\boldsymbol{c}^T\boldsymbol{x}+\phi\left(\boldsymbol{x}\right)$ להיות קרובים עוד ועוד לשפה ∂P. הפונקציה $t\cdot\boldsymbol{c}^T\boldsymbol{x}+\phi\left(\boldsymbol{x}\right)$ תהיה קמורה, לכן אפשר למצוא לה מינימום (ראינו איך בעזרת אופטימיזציה קמורה).

בכל מקרה בגלל ש-ϕ מקבלת ערך אינסופי בשפה, מתקיים שלכל t, $\boldsymbol{x}^{(\boldsymbol{t})}\in P\setminus\partial P$.

למסלול (מסילה) $t\mapsto\boldsymbol{x}^{(\boldsymbol{t})}$ (נסמן את הפונקציה הזו ב-$x:[0,\infty)\to\mathbb{R}^n$) קוראים **המסלול המרכזי** (central path).

למשל כאשר $t=0$ מסתכלים על $\boldsymbol{x}^{(\boldsymbol{0})}=\operatorname{argmin}\phi\left(\boldsymbol{x}\right)$ (**נקודת התחלה**). ככל שאנחנו מגדילים את ערך t כך אנחנו מגיעים קרוב יותר לשפה ∂P ולמעשה מתקרבים לפתרון של התוכנית הלינארית (מי שממזער את $\boldsymbol{c}^T\boldsymbol{x}$).



ממשיכים להתקדם לאורך מסלול עד שמגיעים לנקודה אופטימלית (בשפה של ∂P), המוקפת באדום

אז המטרה תהיה ללכת בעקבות "המסלול" הזה, אך כמובן שאנחנו נצטרך לבחור צעד מתאים ל-t (כי אם גודל הצעד ממש קטן נצטרך לעשות הרבה צעדים כדי לזוז במסלול, ואם גודל הצעד גדול מדי זה עלול להקשות לנו על החישובים).

זה הולך להיות הרעיון מאחורי אלגוריתם הנקודה הפנימית.

נניח שאנחנו ב-$\boldsymbol{x}^{(\boldsymbol{t})}$ ואנחנו צריכים לבחור איזשהו $t'>t$ ולזוז אל $\boldsymbol{x}^{(\boldsymbol{t'})}$. המטרה היא ש-$t'$ יהיה גדול כמה שיותר.

האסטרטגיה תהיה לבחור את t', רק אם קיים אליפסואיד E שמרכזו $\boldsymbol{x}^{(\boldsymbol{t})}$ והוא מקיים:

1. $\boldsymbol{x}^{(\boldsymbol{t'})}\in E$.
2. $\forall\boldsymbol{y}\in E,\ \nabla^2\phi\left(\boldsymbol{y}\right)\approx\nabla^2\phi\left(\boldsymbol{x}^{(\boldsymbol{t})}\right)$ ("ההסיאן לא משתנה הרבה").

   כלומר יש μ,L כך ש-$\nabla^2\phi\left(\boldsymbol{y}\right)\preceq L\cdot\nabla^2\phi\left(\boldsymbol{x}^{(\boldsymbol{t})}\right)$ ו-$\nabla^2\phi\left(\boldsymbol{y}\right)\succeq\mu\nabla^2\phi\left(\boldsymbol{x}^{(\boldsymbol{t})}\right)$ לכל $\boldsymbol{y}$.

   כאשר המשמעות של הסימן $\preceq$ הוא שהמטריצה היא מוגדרת חיובית.

   כלומר, רוצים ש-$L\cdot\nabla^2\phi\left(\boldsymbol{x}^{(\boldsymbol{t})}\right)-\nabla^2\phi\left(\boldsymbol{y}\right)$ ו-$\nabla^2\phi\left(\boldsymbol{y}\right)-\mu\nabla^2\phi\left(\boldsymbol{x}^{(\boldsymbol{t})}\right)$ מוגדרות חיובית לכל $\boldsymbol{y}$.

נשים לב שבתנאי השני אם $\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)$ היא מטריצת היחידה, למשל, אז למעשה הערכים μ, L הם בדיוק הערך העצמי הגדול ביותר והקטן ביותר (בהתאמה).

אפשר לחשוב על זה בתור לעשות טרנספורמציה לכל המרחב $\mathbb{R}^n$ בעזרת המטריצה $\left(\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)\right)^{-1}$. התמונה של האליפסואיד E תהיה אליפסואיד (ראינו את זה בעבר כבר), ואז ההסיאן בכל נקודה באליפסואיד החדש, יקיים את תנאי הערכים העצמיים (שהערכים העצמיים שלו בין μ ל-L).

קרמרקר (Karmarkar) הראה בשנת 1984 שכדי להשיג את $\boldsymbol{x}^{(t')}$, כלומר למזער את $f\left(\boldsymbol{x}\right)=t'\boldsymbol{c}^T\boldsymbol{x}+\phi\left(\boldsymbol{x}\right)$ (זה ההגדרה של $\boldsymbol{x}^{(t')}$), די למזער את $g\left(\boldsymbol{x}\right)=f\left(\left(\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)\right)^{-1/2}\boldsymbol{x}\right)$.

מי זו ϕ ואיך בוחרים אותה?

נזכור שרצינו ש-$\nabla^2\phi\left(\boldsymbol{y}\right)\approx\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)$ לכל $\boldsymbol{y}$. אז המטרה היא שההסיאן של ϕ ישתנה בקצב איטי ככל שאפשר. כלומר יש פה תנאי על הנגזרת השלישית של ϕ. נפרמל את זה:

הגדרה 8.1. בהינתן פונקציית מחסום ϕ נגדיר את **הנגזרת השניה בכיוון $\boldsymbol{y}$, בנקודה $\boldsymbol{x}$** להיות:

$$\nabla^2\phi\left(\boldsymbol{x}\right)\left[\boldsymbol{y},\boldsymbol{y}\right]=\left(\frac{\partial^2}{\partial\varepsilon_1\partial\varepsilon_2}\phi\left(\boldsymbol{x}+\left(\varepsilon_1+\varepsilon_2\right)\boldsymbol{y}\right)\right)\bigg|_{\varepsilon_1,\varepsilon_2=0}$$

כלומר אנחנו מחשבים את הנגזרת הנ"ל לפי $\varepsilon_1,\varepsilon_2$, מקבלים ביטוי שתלוי ב-$\varepsilon_1,\varepsilon_2$ - ואז נציב $\varepsilon_1,\varepsilon_2=0$.

באופן דומה **הנגזרת השלישית בכיוון $\boldsymbol{y}$ בנקודה $\boldsymbol{x}$** היא:

$$\nabla^3\phi\left(\boldsymbol{x}\right)\left[\boldsymbol{y},\boldsymbol{y},\boldsymbol{y}\right]=\left(\frac{\partial^3}{\partial\varepsilon_1\partial\varepsilon_2\partial\varepsilon_3}\phi\left(\boldsymbol{x}+\left(\varepsilon_1+\varepsilon_2+\varepsilon_3\right)\boldsymbol{y}\right)\right)\bigg|_{\varepsilon_1,\varepsilon_2,\varepsilon_3=0}$$

אלה בעצם הכללות להגדרות של נגזרת שניה ושלישית של פונקציות מעל $\mathbb{R}$.

הערה 8.2. אפשר להראות שהנגזרת השניה היא תבנית ריבועית, שגם שווה ל:

$$\nabla^2\phi\left(\boldsymbol{x}\right)\left[\boldsymbol{y},\boldsymbol{y}\right]=\boldsymbol{y}^T\nabla^2\phi\left(\boldsymbol{x}\right)\boldsymbol{y}$$

הנגזרת השלישית היא למעשה תבנית משולשת, אפשר לרשום אותה כטנזור, זה קצת מסובך ולא הכרחי אז לא ניכנס לזה.

נובע מכך ש-$\nabla^2\phi\left(\boldsymbol{x}\right)$ **מגדירה נורמה** במרחב $\mathbb{R}^n$. נגדיר אותה ע"י $||\boldsymbol{y}||_{\nabla^2\phi(\boldsymbol{x})}=\sqrt{\nabla^2\phi\left(\boldsymbol{x}\right)\left[\boldsymbol{y},\boldsymbol{y}\right]}=\sqrt{\boldsymbol{y}^T\nabla^2\phi\left(\boldsymbol{x}\right)\boldsymbol{y}}$.

הגדרה 8.3. פונקציית מחסום ϕ נקראת self-concordant **(מתואמת עצמית?)** אם ורק אם $\left|\nabla^3\phi\left(\boldsymbol{x}\right)\left[\boldsymbol{y},\boldsymbol{y},\boldsymbol{y}\right]\right|\leq 2\,||\boldsymbol{y}||^3_{\nabla^2\phi(\boldsymbol{x})}$ לכל $\boldsymbol{x},\boldsymbol{y}\in P\setminus\partial P$. מההגדרה זה שקול אל $\left|\nabla^3\phi\left(\boldsymbol{x}\right)\left[\boldsymbol{y},\boldsymbol{y},\boldsymbol{y}\right]\right|\leq 2\left|\nabla^3\phi\left(\boldsymbol{x}\right)\left[\boldsymbol{y},\boldsymbol{y}\right]\right|^{3/2}$.

דוגמה 8.4. (שימו לב שאנחנו במקרה חד מימדי אז אין לנו פה משמעות ל-y. המשמעות ל-y היא רק במימדים גבוהים יותר)

נסתכל על הפונקציה $\phi\left(x\right)=-\ln x$ (ב-$[0,\infty)$, ב-0 נגדיר אותה להיות ∞). ה-boundary שלנו פה הוא הנקודה 0 ואכן כשמתקרבים ל-0 הפונקציה שואפת ל-∞.

הנגזרת השלישית של ϕ היא $\phi'''\left(x\right)=\left(-\frac{1}{x}\right)''=\left(\frac{1}{x^2}\right)'=-\frac{2}{x^3}$. כלומר $|\phi'''\left(x\right)|=\frac{2}{x^3}$.

הנגזרת השניה היא $\frac{1}{x^2}$. לכן אנחנו רוצים שיתקיים $\left|\frac{2}{x^3}\right|\leq 2\left|\frac{1}{x^2}\right|^{3/2}$ - במקרה הזה אפילו יש שוויון, ולכן הפונקציה ϕ היא self-concordant.

הערה 8.5. הסיבה שבחרנו את המקדם 2 בהגדרה של self concordant היא רק בגלל הפונקציה $-\ln\left(x\right)$ (אם המקדם הוא 2 היא נותנת שוויון כמו שראינו), היה אפשר לקבוע קבוע אחר.

בהינתן $\boldsymbol{x}^{(t)}$ נגדיר את האליפסואיד $E=\left\{\boldsymbol{x}^{(t)}+\boldsymbol{y}\mid||\boldsymbol{y}||_{\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)}\leq\frac{1}{2}\right\}$ (ודאו שאתם מבינים למה אכן מדובר באליפסואיד). בשיעורי הבית אנחנו נראה שבחירה נכונה של ϕ תביא לנו שהאליפסואיד הזה מקיים את תכונה 2 שרצינו

$\forall \boldsymbol{y}\in E,\ \nabla^2\phi(\boldsymbol{y})\approx\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)$), כלומר שההסיאן לא משתנה הרבה.

אם כך כעת השאלה היא מי ה-t' הכי גדול שמקיים $\boldsymbol{x}^{(t')}\in E$?
נזכור כי $\boldsymbol{x}^{(t)}=\underset{\zeta}{\operatorname{argmin}}\left\{t\boldsymbol{c}^T\zeta+\phi(\zeta)\right\}$, ולכן הנגזרת של הפונקציה הזו מתאפסת בנקודה $\boldsymbol{x}^{(t)}$. הנגזרת היא $t\boldsymbol{c}+\nabla\phi\left(\boldsymbol{x}^{(t)}\right)$, ולכן נקבל $\nabla\phi\left(\boldsymbol{x}^{(t)}\right)=-t\boldsymbol{c}$.
ובנוסף **בהינתן** t' **כלשהו**, בהתחשב בזה ש-$\boldsymbol{x}^{(t')}=\underset{\zeta}{\operatorname{argmin}}\left\{t'\boldsymbol{c}^T\zeta+\phi(\zeta)\right\}$ אנחנו נקבל $\nabla\phi\left(\boldsymbol{x}^{(t')}\right)=-t'\boldsymbol{c}$.
משתי המשוואות הכחולות $\nabla\phi\left(\boldsymbol{x}^{(t')}\right)=-t'\boldsymbol{c}=\frac{t'}{t}\nabla\phi\left(\boldsymbol{x}^{(t)}\right)$.

נסתכל על **הנורמה הדואלית** לנורמה $||\cdot||_{\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)}$. הנורמה הזו היא הנורמה $||\cdot||_{\left(\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)\right)^{-1}}$ שנסמן אותה $||\cdot||_*$.
כאמור (בהערה 8.2) היא מוגדרת על ידי $||\boldsymbol{y}||_*=\boldsymbol{y}^T\left(\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)\right)^{-1}\boldsymbol{y}$. כלומר מתקיים:
$$\left|\left|\nabla\phi\left(\boldsymbol{x}^{(t')}\right)-\nabla\phi\left(\boldsymbol{x}^{(t)}\right)\right|\right|_*=\sqrt{\left(\nabla\phi\left(\boldsymbol{x}^{(t')}\right)-\nabla\phi\left(\boldsymbol{x}^{(t)}\right)\right)^T\left(\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)\right)^{-1}\left(\nabla\phi\left(\boldsymbol{x}^{(t')}\right)-\nabla\phi\left(\boldsymbol{x}^{(t)}\right)\right)}.$$
את הגרדיאנט אפשר לפתח לפי טיילור מסדר ראשון, $\nabla\phi\left(\boldsymbol{x}^{(t')}\right)=\nabla\phi\left(\boldsymbol{x}^{(t)}\right)+\nabla^2\phi(\boldsymbol{z})\left(\boldsymbol{x}^{(t')}-\boldsymbol{x}^{(t)}\right)$ כאשר $\boldsymbol{z}$ היא נקודת אינטרפולציה כלשהי בין $\boldsymbol{x}^{(t)}$ ו-$\boldsymbol{x}^{(t')}$ (כלומר בקטע בין שניהם) (בפרט בגלל ש-E קבוצה קמורה, כי היא אליפסואיד, $\boldsymbol{z}$ באליפסואיד).
בגלל שכל ההסיאנים באליפסואידים הם דומים, כלומר $\nabla^2\phi(\boldsymbol{z})\approx\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)$, נקבל למעשה שמתקיים
$$\left|\left|\nabla\phi\left(\boldsymbol{x}^{(t')}\right)-\nabla\phi\left(\boldsymbol{x}^{(t)}\right)\right|\right|_*=\left|\left|\nabla^2\phi(\boldsymbol{z})\left(\boldsymbol{x}^{(t')}-\boldsymbol{x}^{(t)}\right)\right|\right|_*\approx\left|\left|\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)\left(\boldsymbol{x}^{(t')}-\boldsymbol{x}^{(t)}\right)\right|\right|_*.$$
לכן $\left|\left|\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)\left(\boldsymbol{x}^{(t')}-\boldsymbol{x}^{(t)}\right)\right|\right|_*\approx\sqrt{\left(\boldsymbol{x}^{(t')}-\boldsymbol{x}^{(t)}\right)^T\left(\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)\right)^T\cdot\left(\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)\right)^{-1}\cdot\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)\left(\boldsymbol{x}^{(t')}-\boldsymbol{x}^{(t)}\right)}$.
המטריצה $\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)$ כמובן סימטרית: $\left(\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)\right)^T=\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)$.
לכן נקבל $\left|\left|\left(\boldsymbol{x}^{(t')}-\boldsymbol{x}^{(t)}\right)\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)\right|\right|_*\approx\sqrt{\left(\boldsymbol{x}^{(t')}-\boldsymbol{x}^{(t)}\right)^T\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)\left(\boldsymbol{x}^{(t')}-\boldsymbol{x}^{(t)}\right)}$.
אבל זה $\left|\left|\boldsymbol{x}^{(t')}-\boldsymbol{x}^{(t)}\right|\right|_{\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)}$. אז בסך הכל קיבלנו $\left|\left|\boldsymbol{x}^{(t')}-\boldsymbol{x}^{(t)}\right|\right|_{\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)}\approx\left|\left|\nabla\phi\left(\boldsymbol{x}^{(t')}\right)-\nabla\phi\left(\boldsymbol{x}^{(t)}\right)\right|\right|_*$.
אבל $\nabla\phi\left(\boldsymbol{x}^{(t')}\right)-\nabla\phi\left(\boldsymbol{x}^{(t)}\right)=\frac{t'}{t}\nabla\phi\left(\boldsymbol{x}^{(t)}\right)-\nabla\phi\left(\boldsymbol{x}^{(t)}\right)=\left(\frac{t'}{t}-1\right)\nabla\phi\left(\boldsymbol{x}^{(t)}\right)$.
ולכן $\left|\left|\boldsymbol{x}^{(t')}-\boldsymbol{x}^{(t)}\right|\right|_{\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)}\approx\left(\frac{t'}{t}-1\right)\cdot\left|\left|\nabla\phi\left(\boldsymbol{x}^{(t)}\right)\right|\right|_*$.
אמרנו שאנחנו רוצים $\boldsymbol{x}^{(t')}\in E$, כלומר $\left|\left|\boldsymbol{x}^{(t')}-\boldsymbol{x}^{(t)}\right|\right|_{\nabla^2\phi\left(\boldsymbol{x}^{(t)}\right)}\leq\frac{1}{2}$.
אז נבחר t' המקיים $\left(\frac{t'}{t}-1\right)\cdot\left|\left|\nabla\phi\left(\boldsymbol{x}^{(t)}\right)\right|\right|_*\leq\frac{1}{2}$, כלומר $t'=\left(1+\frac{1}{2\left|\left|\nabla\phi\left(\boldsymbol{x}^{(t)}\right)\right|\right|_*}\right)t$.
עבדנו עם שוויונות קירוב $\approx$, בפועל יהיו שוויונות עד כדי קבועים, ואז t' לא יהיה שווה בדיוק לביטוי שרשמנו מעל, אלא לקבוע כלשהו כפול זה.

הגדרה 8.6. פונקציית מחסום ϕ נקראת ν-self-concordant אם לכל $\boldsymbol{x}$ ולכל $\boldsymbol{x}^{(t)}$ מתקיים $\left(||\nabla\phi(\boldsymbol{x})||_*\right)^2\leq\nu$ כאשר הנורמה הדואלית היא לנורמה $||\cdot||_{\nabla\phi\left(\boldsymbol{x}^{(t)}\right)}$.

טענה 8.7 (לא נוכיח בקורס זה). **הקרבה לפתרון האופטימלי:** עבור $0<\varepsilon$, אם $t=\frac{\nu}{\varepsilon}$, $\boldsymbol{c}^T\boldsymbol{x}^{(t)}\leq\left(\min_{\boldsymbol{x}\in P}\boldsymbol{c}^T\boldsymbol{x}\right)+\varepsilon$.

כעת אם ϕ היא ν-self-concordant, אז $t'=\left(1+\frac{1}{2\left|\left|\nabla\phi\left(\boldsymbol{x}^{(t)}\right)\right|\right|_*}\right)t\leq\left(1+\frac{1}{2\sqrt{\nu}}\right)t$.
אז כמה איטרציות אנחנו צריכים? אם נסמן ב-k את מספר האיטרציות שאנחנו עושים, ואת ε הקִרבה שאנחנו רוצים לפתרון,
נרצה שיתקיים $\left(1+\frac{1}{2\sqrt{\nu}}\right)^k\geq\frac{\nu}{\varepsilon}$, כלומר אפשר לבחור $k=2\sqrt{\nu}\ln\frac{\nu}{\varepsilon}=O\left(\sqrt{\nu}\ln\frac{\nu}{\varepsilon}\right)$.

זה בערך הפתרון הטוב ביותר שאנחנו יודעים לספק היום לתוכניות לינאריות.

עכשיו נצטרך לתת את הדעת לגבי מי זו ϕ וגם מי ה-ν - האם בכלל קיימות פונקציות ν-self-concordant?

משפט 8.8 (Nesterov-Nemirovski 1994)**.** לכל קבוצה קמורה $P \subseteq \mathbb{R}^n$ יש פונקציית מחסום ϕ אשר היא $O(n)$-self-concordant.

אבל, לא קל לחשב את ההסיאן של הפונקציה במשפט הקודם (קשה לחשב אותו לפחות כמו לפתור תוכניות לינאריות).
בנוסף עבור $P = [-1,1]^n$ (הקוביה ה-n מימדית) מתקיים $\nu \geq n$.

טענה 8.9. תהא $f : (0,\infty)^n \to \mathbb{R}$ המוגדרת על ידי $f(\boldsymbol{x}) = -\sum_{i=1}^{n} \log(x_i)$. אז f היא פונקציית מחסום n-self-concordant בקבוצה $[0,\infty)^n$.

הוכחה. קודם כל כדי שהיא תהיה פונקציית מחסום צריך להראות שהיא קמורה, והיא אכן קמורה (אפילו ממש) בתור סכום של הפונקציות קמורות $-\log(x_i)$.
הנקודות ב-boundary אלה נקודות שיש 0 באחת מהקורדינטות שלהן, ולכן אם מתקרבים לאחת מהנקודות האלה, $-\log(x_i)$ (כאשר i של הקורדינטה המתאימה) שואפת ל-∞ כמו שרוצים.
לכן זו אכן פונקציית מחסום. כעת נראה שהיא n-self-concordant. נחשב את הנגזרות. מתקיים $\nabla_i f(\boldsymbol{x}) = -\frac{1}{x_i}$.
עבור ההסיאן $\nabla^2 f(\boldsymbol{x})$, הקורדינטה במקום ה-$i,j$ תהיה $\begin{cases} 0 & i \neq j \\ \frac{1}{x_i^2} & i = j \end{cases}$ (שכן זה לגזור את $-\frac{1}{x_i}$ לפי x_j).
אז המטריצה ההופכית היא $\left(\nabla^2 f(\boldsymbol{x})\right)^{-1} = \begin{bmatrix} x_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & x_n^2 \end{bmatrix}$.
כעת נחשב את $||\nabla f(\boldsymbol{x})||_*^2$ (כאשר $||\cdot||_*$ היא הנורמה הדואלית של $\nabla^2 f(\boldsymbol{x})$).

$$||\nabla f(\boldsymbol{x})||_*^2 = (\nabla f(\boldsymbol{x}))^T \left(\nabla^2 f(\boldsymbol{x})\right)^{-1} \cdot \nabla f(\boldsymbol{x}) = \begin{bmatrix} -\frac{1}{x_1} & \cdots & -\frac{1}{x_n} \end{bmatrix} \begin{bmatrix} x_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & x_n^2 \end{bmatrix} \begin{bmatrix} -\frac{1}{x_1} \\ \vdots \\ -\frac{1}{x_n} \end{bmatrix}$$

נקבל שזה יוצא $\sum_{i=1}^{\infty} \left(-\frac{1}{x_i}\right) x_i^2 \cdot \left(-\frac{1}{x_i}\right) = n$. ולכן הפונקציה היא אכן n-self-concordant כנדרש. $\square$

התוכנית הלינארית שלנו היא מהצורה $\begin{matrix} \min_{\boldsymbol{x}} \boldsymbol{c}^T \boldsymbol{x} \\ A\boldsymbol{x} \leq \boldsymbol{b} \end{matrix}$. אנחנו צריכים פונקציית מחסום n-self-concordant לפוליטופ $A\boldsymbol{x} \leq \boldsymbol{b}$.
המשפט הבא אומר שפונקציות לינאריות שומרות על פונקציות מחסום וגם על ה-concordance שלהן.

משפט 8.10. תהא $g : \mathbb{R}^n \to \mathbb{R}^m$ פונקציה לינארית, ו-f פונקציית מחסום ν-self-concordant על $P \subseteq \mathbb{R}^m$. אז $f \circ g$ היא פונקציה ν-self-concordant ל-$g^{-1}(P) = \{\boldsymbol{x} \in \mathbb{R}^n \mid g(\boldsymbol{x}) \in P\}$.

נוכל להשתמש במשפט הנ"ל: נגדיר $g(\boldsymbol{x}) = \boldsymbol{b} - A\boldsymbol{x}$. ואז $g^{-1}([0,\infty)^m) = \{\boldsymbol{x} \in \mathbb{R} \mid \boldsymbol{b} - A\boldsymbol{x} \geq 0\}$.
כלומר $g^{-1}([0,\infty)^m)$ היא הפוליטופ שלנו. נגדיר את f שראינו מקודם ($-\sum_{i=1}^{m} \log(x_i)$), וכך נוכל לקבל פונקציה m-self-concordant עבור הפוליטופ.
במקרה שלנו, $(f \circ g)(\boldsymbol{x}) = f(g(\boldsymbol{x})) = f(\boldsymbol{b} - A\boldsymbol{x}) = -\sum_{i=1}^{m} \log(b_i - \boldsymbol{A_i} x_i)$.
אז כפי שראינו מקודם אנחנו צריכים $O\left(\sqrt{m} \log\left(\frac{m}{\varepsilon}\right)\right)$ איטרציות כדי להגיע לדיוק ε בפתרון.
אבל, אנחנו צריכים לחשב גם את ההסיאן של $f \circ g$ כחלק מהפתרון. זה יקח לנו $O\left(m \cdot n^2\right)$ (חישוב הנגזרות החלקיות, וכו'), ולהפוך אותו יכול לקחת לנו $O\left(n^3\right)$ (אולי אפילו יותר טוב, תלוי באלגוריתמים שאנחנו משתמשים).
בסך הכל אלגוריתם הנקודה הפנימית יכול לעבוד בזמן $O\left((n+m)^3\right)$.

חלק II

רשתות זרימה

1 מבוא לרשתות זרימה

הרצאה 13

אנחנו נגדיר את המושגים של רשתות זרימה וגם את המושגים הקשורים לרשתות רב מוצריות, עליהן נדבר בעתיד.

הגדרה 1.1. רשת זרימה היא חמישיה $\left(V,E,c,\{s_i\}_{i=1}^{n},\{t_i\}_{i=1}^{\infty}\right)$, כאשר $G=(V,E)$ הוא גרף לא מכוון, $c:E\to\mathbb{N}$ **פונקציית קיבול** על הצלעות, ו-$\{(s_i,t_i)\}_{i=1}^{\infty}$ הם זוגות של **קודקודי מקור** ו**קודקודי בור**. כל זוג (s_i,t_i) נקרא **מוצר** (commodity). לעיתים אנחנו גם נדרוש פונקציית דרישה (demand function) $d:\{1,...,k\}\to\mathbb{N}$ (שמשמעותה כמה זרימה רוצים לשלוח מ-s_i ל-t_i).

רשת זרימה בה $n>1$ נקראת **רשת זרימה רב מוצרית**, אחרת תיקרא **רשת זרימה חד מוצרית**.

הגדרה 1.2. זרימה (מצומצמת) **ברשת חד מוצרית** ברשת זרימה עם קודקודי מקור ובור s,t (בהתאמה) היא פונקציה אי שלילית $f:V\times V\to\mathbb{R}_{\geq 0}$ (הזרימה היא פונקציה **מכוונת** על הקודקודים) כך ש:

- **אילוץ הקיבול:** לכל $(x,y)\in E$ מתקיים $f(x,y)\leq c(x,y)$.
- **חוק שימור החומר:** $\sum_{y\in V}f(y,x)=\sum_{y\in V}f(x,y)$ לכל $x\in V\setminus\{s,t\}$.

הגדרה 1.3. זרימה רב מוצרית ברשת זרימה עם n מוצרים היא אוסף פונקציות זרימה $f^{(1)},...,f^{(n)}$ כש-$f^{(i)}$ מקיימת את דרישות הזרימה החד מוצרית עבור s_i ו-t_i.

הגדרה 1.4. שטף של זרימה f מוגדר להיות $|f|=\sum_{v\in V}f(s,v)$ במקרה של זרימה חד מוצרית, ו-$|f|=\sum_{v\in V}f^{(i)}(s_i,v)$ במקרה של זרימה רב מוצרית.

בעיה (זרימה מקסימלית (maximum flow))**.** איך מוצאים את הזרימה המקסימלית ברשת זרימה? כלומר זו שהשטף שלה הגדול ביותר.

במקרה שהרשת היא רשת זרימה רב מוצרית, הבעיה נקראת maximum multicommodity flow.

לא תמיד אפשר לספק את פונקציית הדרישה d (מטעמי קיבול או חוקיות זרימה). לכן אנחנו דנים גם בבעיה הבאה:

בעיה (זרימה נדרשת מקסימלית (maximum demand flow))**.** רוצים את ה-α המקסימלי כך שאפשר לספק את פונקציית הדרישה $\alpha\cdot d$ (ייתכן אפילו $\alpha\geq 1$).

בעיה (זרימה בו זמנית מקסימלית (maximum concurrent flow))**.** רוצים את ה-u המקסימלי, ככה שניתן להזרים זרימה חוקית של לפחות u **מכל** המקורות.

במקרה שיש לנו commodity אחד (כמו באלגו מתואר ראשון) - כל הבעיות האלה שקולות. כשהזרימה היא רב מוצרית הבעיות האלה הן שונות.

נשים לב שבעיית זרימה מקסימלית היא בעיית תכנון לינארי שמשתניה הם הערכים של הזרימה (לכל צלע). כל האילוצים הם אילוצים לינאריים, ופונקציית המטרה גם היא לינארית.

לכן אפשר לפתור את בעיית הזרימה המקסימלית בעזרת תכנון לינארי. למרות זאת, לבעיה הזו יש אפילו פתרון טוב יותר.

נתחיל בלדבר על הבעיה הראשונה, כשהרשת היא **חד מוצרית**.

הגדרה 1.5. פונקציית זרימה מורחבת ברשת זרימה חד מוצרית עם גרף **מכוון** $G = (V, A)$ ופונקציית קיבול $c : A \to \mathbb{N}$ היא פונקציה $f : V \times V \to \mathbb{R}$ שמקיימת:

- **אילוץ הקיבול:** $f(u, v) \leq c(u, v)$, כאשר אם $(u, v) \notin E$ אז $c(u, v) = 0$.
- **אילוץ האנטי סימטריה:** $f(u, v) = -f(v, u)$ לכל $u, v \in V$.
- **חוק שימור החומר:** $\sum_{u \in V} f(u, v) = 0$ לכל $v \in V$.

הגדרה 1.6. שטף זרימה מורחבת הוא $|f| = \sum_{u \in V \setminus \{s\}} f(s, u)$.

2 אלגוריתם push-relabel

ראינו בתואר ראשון את אלגוריתם פורד-פלקרסון למציאת זרימה אופטימלית. האלגוריתם הזה מתחיל מזרימת אפס, כל פעם מוצא מסלול שיורי (מסלול שאפשר להזרים בו זרימה), ומחשב את הרשת והזרימה השיורית. בכל האיטרציות אנחנו נשמרים עם זרימה חוקית, **אם** האלגוריתם מסיים את ריצתו זה יהיה כשאין איך להזרים זרימה נוספת.
אפשר לממש את אלגוריתם פורד-פלקרסון בצורה טובה יותר עם דרכים מתחכמות, למשל אלגוריתם אדמונדס-קארפ שכל פעם בוחר את המסלול הקצר ביותר או אלגוריתם דיניץ שעובד עם מספר מסלולים במקביל ולכן גם מהיר יותר.

עכשיו נראה גישה חדשה לגמרי. מה ששונה פה הוא שהאלגוריתם לא תמיד משמר זרימה חוקית בין האיטרציות, אלא כל פעם משמר קדם-זרימה (pre-flow). כלומר זרימה שיכולה "לדלוף" בדרך מ-s אל t. ואז נצטרך "לסגור את הדליפות", אז האלגוריתם למעשה מקטין את הזרימה הכוללת (בניגוד לפורד-פלקרסון למשל).

הגדרה 2.1. פונקציית קדם-זרימה (pre-flow) ברשת זרימה חד מוצרית עם גרף **מכוון** $G = (V, A)$ ופונקציית קיבול $c : A \to \mathbb{N}$ היא פונקציה $f : V \times V \to \mathbb{R}$ שמקיימת:

- **אילוץ הקיבול:** $f(u, v) \leq c(u, v)$, כאשר אם $(u, v) \notin E$ אז $c(u, v) = 0$.
- **אילוץ האנטי סימטריה:** $f(u, v) = -f(v, u)$ לכל $u, v \in V$.
- **דליפה אפשרית:** $\sum_{u \in V} f(u, v) \geq 0$ לכל $v \in V$.
  הגודל $e_v = \sum_{u \in V} f(u, v)$ נקרא **עודף הזרימה** (excess flow) ב-v.

מסקנה. אם לכל $v \in V$ זרימת היתר e_v היא 0, אז f היא פונקציית זרימה.

הגדרה 2.2. תיוג מרחקים (distance labeling) המתאימה לקדם-זרימה $f : V \times V \to \mathbb{R}$ היא פונקציה $d : V \to \{0, 1, 2, ..., n\}$ (כאשר $|V| = n$) ככה ש:

- $d(t) = 0$.
- $d(s) = n$.
- לכל צלע (u, v) ברשת הזרימה השיורית G_f(כמעט אותה רשת רק שפונקציית הקיבול היא $c_f = c(u, v) - f(u, v)$, ואנחנו מורידים צלעות עם קיבול 0) מתקיים $d(u) \leq d(v) + 1$.

הפונקציה d יכולה לתת הערכה של המרחק של כל קודקוד מקודקוד הבור:

טענה 2.3. לכל $s \neq u$, $d(u)$ הוא לכל היותר המרחק של u ל-t ב-G_f.

הוכחה. נסתכל על המסלול הקצר ביותר $(x_0,...,x_m)$ מ-$u = x_0$ אל $x_m = t$ (המרחק בין u ל-t הוא m).
מתקיים $d(x_m) = d(t) = 0$ ובנוסף לכל $0 \le i \le m-1$, מהאילוץ השלישי מתקיים $d(x_i) \le d(x_{i+1}) + 1$.
באינדוקציה $d(x_j) \le d(x_{i+1}) + 1 + i - j$ לכל $0 \le j \le i+1$.
נקבל עבור $j = 0$ ש-$d(x_0) \le 1 + i$, נבחר $i = m-1$ ונקבל $d(x_0) \le m$, כלומר $d(u) \le m$ כנדרש. □

הרצאה 14

נתחיל בתיוג המרחקים d המוגדר ע"י $d(s) \leftarrow n$ ו-$d(x) \leftarrow 0$ לכל $x \in V \setminus \{s\}$. בנוסף נתחיל עם הזרימה המוגדרת ע"י $f(s,x) \leftarrow c(s,x)$ ו-$f(x,s) \leftarrow -f(s,x)$ (בכל צלע אחרת - 0). נשים לב שנובע מכך ש-$e_x = f(s,x)$.
נסמן ב-$E(G_f)$ את קבוצת הצלעות של הרשת השיורית (הרשת החדשה כשמורידים מהקיבולים על הצלעות את f, ואת כל הצלעות עם קיבול 0).

בהינתן עודף זרימה $e_x > 0$ ($x \in V$) היינו רוצים להזיז את עודף הזרימה e_x אל קודקוד כלשהו y כך ש-$(x,y) \in E(G_f)$, כאשר $d(x) = d(y) + 1$. כלומר להזיז את הזרימה העודפת כל פעם קודקוד אחד לקראת קודקוד הבור. הצעד הזה נקרא "push" והזרימה שאנחנו מזיזים היא $\min(\{e_x,\ c_f(x,y)\})$.
אם אין לנו צעד push לעשות (זכרו כי עושים push רק אם $d(x) = d(y) + 1$) אנחנו נעשה צעד אחר.
אנחנו במצב בו לכל $x \in V$ ולכל $y \in V$ אשר $(x,y) \in E(G_f)$, מתקיים $d(x) \le d(y)$. אז מה שנעשה יהיה לשנות את $d(x)$ (כאמור, עבור x כלשהו שיש לו עודף זרימה). נגדיר את $d(x)$ החדש להיות $d(x) \leftarrow \min(\{d(y) + 1 \mid y \in V \ \ (x,y) \in E(G_f)\})$. הצעד הזה נקרא "relabel".
כל עוד קיים $x \ne s,t$ אשר $e_x > 0$ - אנחנו לא נעצור (נעשה push או relabel).

אלגוריתם 1 push relabel

---

1. **נאתחל** $f \leftarrow 0$, $e \leftarrow 0$.
2. <u>אתחול:</u> לכל קודקוד $x \in V$:
   (א) $f(s,x) \leftarrow c(s,x)$, $f(x,s) \leftarrow -f(s,x)$.
   (ב) $e_x \leftarrow f(s,x)$.
   (ג) אם $x = s$: נגדיר $d(x) \leftarrow n$, ואחרת $d(x) \leftarrow 0$.
3. <u>כל עוד:</u> קיים $x \in V \setminus \{s,t\}$, אשר $e_x > 0$ (עודף זרימה חיובי):
   (א) **אם** יש שכן y של x אשר $c_f(x,y) > 0$, וגם $d(x) = d(y) + 1$:
      i. נגדיר $\delta \leftarrow \min\{e_x, c_f(x,y)\}$.
      ii. $f(x,y) \leftarrow f(x,y) + \delta$, $f(y,x) \leftarrow f(y,x) - \delta$.
      iii. $e_x \leftarrow e_x - \delta$, $e_y \leftarrow e_y + \delta$.
      push
   (ב) **אחרת:**
      i. תהא A_f הרשת השיורית המתאימה לזרימה f.
      ii. $d(x) \leftarrow \min_{y \text{ s.t. } (x,y) \in A_f} \{d(y) + 1\}$.
      relabel

עלינו להוכיח שהאלגוריתם הזה חוקי ואכן מחשב זרימה מקסימלית בזמן ריצה טוב.

למה 2.4. בכל איטרציה d היא תיוג מרחקים תקינה ברשת השיורית G_f.

הוכחה. באינדוקציה על מספר האיטרציות של האלגוריתם.
הבסיס: לפני האיטרציה הראשונה, אם $(x,y) \in E$ (כאשר f היא הזרימה ההתחלתית), אז מתקיים $d(x) = 0$ ו-$d(y) \le n$ מההגדרה של אתחול פונקציית תיוג המרחקים (שכן $d(y) = 0$ אם $y \ne s$ ו-$d(s) = n$). ולכן $d(x) \le d(y) + 1$.

כעת להנחה ולצעד האינדוקציה: נניח שבתחילת האיטרציה התחלנו עם פונקציית תיוג מרחקים d חוקית בהתחשב בפונקציית הזרימה f, ונוכיח שגם בסיומה.

צעד ה-relabel לכשעצמו לא מהווה בעיה.

זה בגלל שמשנים את $d(x)$ להיות המינימלי מבין $\{d(y)+1 \mid y\in V \ \ (x,y)\in E(G_f)\}$ (ואז אכן $d(x)\le d(y)+1$ לכל $(x,y)\in E(G_f)$). קודקודי המקור והבור לא משנים את התיוג שלהם, אז אנחנו נשארים עם תיוג מרחקים חוקי.

ה"בעיה" היא בצעד ה-push: אם (x,y) נעלם בחישוב הרשת השיורית, זה מעלים לנו אילוץ אחד לחוקיות של d ($d(x)\le d(y)+1$) - ואז אין בעיה. אבל ייתכן ש-(y,x) לא הייתה ברשת ולפתאום הופיעה ברשת השיורית, ואז אנחנו צריכים לתת את הדעת על התיוג שם. נזכור שהתנאי לעשות push מ-x ל-y הוא ש-$d(x)=d(y)+1$. כלומר $d(y)=d(x)-1$ ולכן בפרט גם יתקיים $d(y)\le d(x)+1$ - כמו שרוצים. □

משפט 2.5. אם האלגוריתם עוצר, והמרחקים d סופיים, אז f היא זרימה מקסימלית.

הוכחה. נשים לב ש-f זרימה חוקית, כי תנאי העצירה שלנו היה רק כאשר אין זרימה עודפת (ועל תנאי אילוץ הקיבול והאנטי סימטריה אנחנו נשארים).

נניח בשלילה ש-f איננה זרימה מקסימלית. אז ב-G_f יש מסלול פשוט (בלי מעגלים) מ-s ל-t:

$$s=x_0\longrightarrow x_1\longrightarrow x_2\longrightarrow\cdots\longrightarrow x_{\ell-1}\longrightarrow x_\ell$$

מתקיים $d(x_\ell)=d(t)=0$, ו-$d(x_{\ell-1})\le d(x_\ell)+1\le 1$, ולכן באינדוקציה $d(x_{\ell-k})\le k$, נקבל $d(x_0)=d(s)\le \ell$.

אבל $\ell\le n-1$ - המסלול עובר לכל היותר בכל הקודקודים בדרך. נקבל $d(s)\le n-1$ - סתירה, כי באלגוריתם שלנו תמיד $d(s)=n$ (זה נשמר בין האיטרציות). □

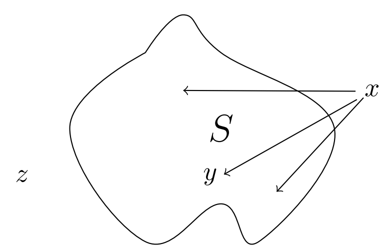
אז הקושי העיקרי הוא להוכיח שהאלגוריתם עוצר.

נקרא לקודקוד x **פעיל** (active) אם $e_x>0$.

למה 2.6. אם f היא קדם-זרימה ו-x פעיל, אז יש ב-G_f מסלול מ-x ל-s (כל הצלעות בכיוונים ההפוכים צריכות להופיע).

הוכחה. נסמן ב-S את קבוצת הקודקודים שאפשר להגיע אליהם מ-x ב-G_f. נניח בשלילה ש-$s\notin S$.

ניקח איזשהו $y\in S$ ו-$z\notin S$. זה נראה משהו כזה:



הצלע (y,z) לא קיימת שכן אחרת $z\in S$ (היה אפשר להגיע מ-x ל-y ומ-y ל-z). זה גורר שהקדם-זרימה מ-z ל-y איננה חיובית: $f(z,y)\le 0$ (אחרת הצלע הייתה מופיעה ב-G_f). זה אומר ש-$f(y,z)=c(y,z)\ge 0$.

זה נכון לכל $y\in S$. נסתכל על $\sum_{y\in S} e_y$. כל הזרימות שבאות ל-y מקודקודים שנמצאים בתוך s מתבטלים בסכימה הגדולה - כי איפשהו יש לנו $f(x,y)$ ואז $f(y,x)$ (המנוגדים זה לזה).

ולכן כל הזרימות שמשנות הן רק "הזרימות בגבול", בדיוק כמו בציור מעלה. $\sum_{y\in S} e_y\le\sum_{y\in S}\ \sum_{z\ (z,y)\in E(G_f)} f(z,y)$.

אבל $f(z,y)\le 0$ לכל $y\in S$ ו-$z\in S$ $(z,y)\in E(G_f)$ ולכן $\sum_{y\in S} e_y\le 0$. אבל עודף זרימה לא יכול להיות קטן מ-0.

נקבל $\sum_{y\in S} e_y=0$. כלומר $e_y=0$ לכל $y\in S$.

מפה, $e_x=0$ בין היתר (שכן $x\in S$ - אפשר להגיע מ-x לעצמו) - זו סתירה לכך ש-x פעיל. □

למה 2.7. לכל $x\in V$ ובכל שלב $d(x)\le 2n-1$.

הוכחה. אם $x=s$ או $x=t$ זה טריויאלי, כי $d(s)=n$ ו-$d(t)=0$ והם לא משתנים.
עבור x אחר, $d(x)$ משתנה רק אם x הוא פעיל (עושים relabel ל-x רק אם $e_x>0$), אבל מטענה קודמת יש מסלול ב-G_f מ-x ל-s. בפרט יש מסלול פשוט (ללא מעגלים), אז יש לו לכל היותר $n-1$ קודקודים. לפני השינוי d היא פונקציית labeling תקינה, אז $d(x)\le d(s)+\text{path length}$ (כל צלע במסלול יכולה להגדיל את ה-labeling ב-1 לכל היותר), נקבל $d(x)\le n+n-1=2n-1$. $\square$

אנחנו רוצים לחסום את מספר האיטרציות. עכשיו כשיש לנו את כל הלמות אפשר לעשות את זה:

למה 2.8. מספר צעדי ה-relabel באלגוריתם הוא לכל היותר $2n^2$.

הוכחה. לכל x מתקיים $0\le d(x)\le 2n-1$ (מהלמה הקודמת) ואז relabel ל-$x\in V\setminus\{s,t\}$ יגדיל את $d(x)$ בלכל היותר 1. יש לנו $n-2$ קודקודים שאפשר לעשות להם relabel (כולם חוץ מהבור והמקור). אז מספר ה-relabelים הוא לכל היותר $(n-2)\cdot(2n-1)\le 2n^2$. $\square$

עכשיו נרצה לחסום את מספר צעדי ה-push. אנחנו נבדיל בין 2 סוגים של צעדי push: מרווים ולא מרווים.
נגיד ש-push על (x,y) **מרווה** (saturates) את (x,y) אם אנחנו עושים push ל-$c_f(x,y)$ זרימה מ-x ל-y. במילים אחרות, הצלע (x,y) נעלמת בגרף השיורי.

למה 2.9. push-relabel מבצע לכל היותר mn צעדי push מרווים (כאשר $m=2|E(G)|$ ו-$n=|V|$).

הוכחה. בהינתן $(x,y)\in E(G_f)$ אם אנחנו עושים push מרווה על (x,y), באיטרציה ה-i, חייב להתקיים (מתנאי ה-push) $d^{(i)}(x)=d^{(i)}(y)+1$. בגלל שהזרימה מרווה, (x,y) נעלמת. אז כדי לעשות push שוב לצלע, הצלע חייבת להופיע מחדש בהמשך. אז באיזושהי איטרציה בהמשך $j>i$ עושים push מ-y ל-x, אז מתקיים $d^{(j)}(y)=d^{(j)}(x)+1$, ואיפשהו עושים push בחזרה מ-x ל-y באיטרציה $k>j$ - אז $d^{(k)}(x)=d^{(k)}(y)+1$. מתקיים $d^{(k)}(y)\ge d^{(j)}(y)$, ו-$d^{(j)}(x)\ge d^{(i)}(x)$ (אנחנו רק מגדילים את ערך d בין האיטרציות).
נקבל $d^{(k)}(x)=d^{(k)}(y)+1\ge d^{(j)}(y)+1=d^{(j)}(x)+2\ge d^{(i)}(x)+2$.
אז בין ה-pushים המרווים, $d(x)$ משתנה על ידי לכל הפחות 2. הערך המקסימלי של $d(x)$ הוא $2n-1$ לכן מספר ה-pushים הוא לכל היותר $\frac{2n-1}{2}+1=n+\frac{1}{2}$, כלומר הוא לכל היותר n (כי מספר pushים זה מספר שלם).
זה נכון לכל צלע $(x,y)\in E(G_f)$ ב-G_f ולכן מספר ה-pushים המרווים הוא לכל היותר $2mn$. $\square$

מה לגבי מספר ה-pushים הלא מרווים?

למה 2.10. push-relabel מבצע לכל היותר $6n^2m$ צעדי push לא מרווים.

הוכחה. נגדיר פונקציית פוטנציאל $\Phi=\sum_{x\neq s,t\ \ e_x>0} d(x)$. בהתחלה $\Phi=0$ כי כל ה-d-ים של קודקודים שאינם s,t הם 0. גם בסוף, $\Phi=0$, כי אין קודקוד עם זרימה עודפת. אז Φ גדל וקטן במהלך ההרצה.
כל צעדי ה-relabel רק מגדילים את ה-d-ים ולכן יגדילו את Φ. הם יגדילו את Φ בלכל היותר $2n^2$ (בכל צעד relabel איזשהו d גדל בלכל היותר 1, ויש לכל היותר $2n^2$ צעדי relabel).
צעדי push מרווים גורמים ל-x כלשהו לאפס את עודף הזרימה שלו ולצלע נוספת להופיע בגרף השיורי. לכן, הם מגדילים את Φ רק כשהם מגדילים את **קבוצת הקודקודים** הפעילים. צעד push מרווה יכול להוסיף לנו קודקוד אחד נוסף לקבוצת הקודקודים. במקרה הזה, אנחנו סוכמים מחובר חדש, $d(y)$, אבל ערכו לכל היותר $2n-1$. כלומר כל צעד push מרווה מגדיל את Φ בלכל היותר $2n-1$ **פר קודקוד ופר צלע**. אז ביחס זה $(2n-1)\cdot 2nm$ (כאשר $2m$ מספר הצלעות בגרף ו-n מספר הקודקודים).
צעד push לא מרווה לא יכול להגדיל את Φ - אמנם בצעד push לא מרווה (מ-x ל-y) קודקוד כלשהו נפטר מהזרימה העודפת שלו, כלומר אנחנו לא סוכמים אותו יותר ב-Φ, אבל מצטרף קודקוד חדש לסכימה y, אבל $d(x)=d(y)+1$. כלומר $d(y)=d(x)-1$ - במילים אחרות, רק הורדנו את הערך של Φ **ב-1**!
בגלל ש-Φ מתחילה מ-0 וחוזרת ל-0, וכל צעד push רק מוריד את הערך של Φ - בהתחשב בכך שהעלנו את Φ בלכל היותר $(2n-1)2mn+2n^2$. זה קטן מ-$2n^2+4mn^2$ שקטן מ-$6mn^2$. אז מספר צעדי ה-push הלא מרווים הוא לכל היותר $6mn^2$, כנדרש. $\square$

מסקנה 2.11. זמן הריצה של אלגוריתם push-relabel הוא $6mn^2+mn+2n^2$ לכל היותר, כלומר $O\left(mn^2\right)$. לשם השוואה, אלגוריתם אדמונדס-קארפ רץ ב-$O\left(m^2n\right)$. יש מימושים דומים של push-relabel שיכולים לרוץ ב-$O\left(n^3\right)$ (יכול להיות קצת יותר טוב, תלוי מי זה m).

זה הכל לזרימה חד מוצרית. ראוי להזכיר שיש דרכים נוספות וחדשות לגמרי לפתור בעיות זרימה. עד היום הגישות האלה לא הובילו לפרצת דרך משמעותית.

חלק III

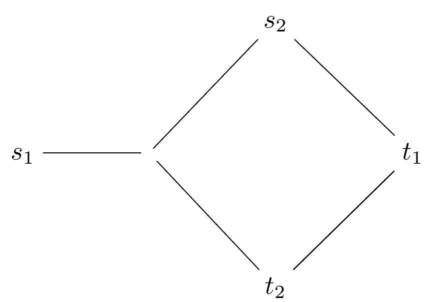
זרימה רב מוצרית וקירובים

בפרק הזה של הקורס נסקור את הרעיון של זרימה רב מוצרית, ונראה עד כמה היא שונה ודומה למושגים שראינו בזרימה חד מוצרית בתואר ראשון.

1 קירוב לזרימת מקסימום בשלמים

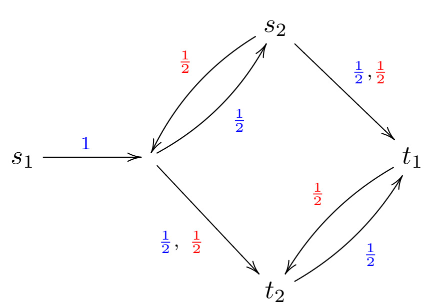
בבעיות זרימה רב מוצרית הכל מסתבך. בעיית הזרימה המקסימלית הרב מוצרית **<u>בשלמים</u>** היא בעיה NP-קשה (גם כשהקיבולים על הצלעות הם שלמים כפי שמיד נראה).

דוגמה 1.1. נתבונן ברשת הזרימה הרב מוצרית הבאה (כל הקיבולים הם 1):



מה הזרימה המקסימלית?

להזרים זרימה מקסימלית מ-s_1 ל-t_1 ואז (ברשת השיורית) מ-s_2 ל-t_2 תתן לנו זרימה מקסימלית של 1. נדמה שלא משנה מה נעשה, הזרמה של זרימה מ-s_1 ל-t_1 תחסום זרימה מ-s_2 ל-t_2 (ובאופן סימטרי ההפך). למרות זאת, אם "לא עובדים בצעדים" אפשר להזרים זרימה מקסימלית של 2 (1 לכל commodity): בצבע אדום זרימה מוצרית אחת ובצבע כחול שניה:



נובע מכך ההבדל הראשון בין זרימה רב וחד מוצרית: אם הקיבולים שלמים - בזרימה חד מוצרית בהכרח יש זרימה אופטימלית בשלמים (והמון שימושים של זרימה משתמשים בתכונה הזו), אבל בזרימה רב מוצרית זה לא המצב.

הרצאה 15

שימו לב שבהינתן רשת זרימה רב מוצרית עם k מוצרים, אפשר לתת אלגוריתם של $\frac{1}{k}$-קירוב: אם נריץ k פעמים אלגוריתם לחישוב זרימה מקסימלית של רשת זרימה **חד** מוצרית מ-s_i ל-t_i (כאמור לכל $1 \leq i \leq k$), כלומר אם נתעלם משאר המוצרים, ונבחר את הזרימה המקסימלית.

את בעיית הזרימה הרב מוצרית אפשר לנסח כבעיית תכנון לינארי. קודם נדבר על בעיית זרימת המקסימום.
בהינתן מוצרים $(s_1, t_1), ..., (s_k, t_k)$, נסמן ב-$P_i$ את קבוצת המסלולים הפשוטים (ללא מעגלים) מ-s_i ל-t_i. שימו לב שייתכן ש-P_i בגודל אקספוננציאלי ביחס לגודל הגרף.

המשתנים שלנו יהיו $f_P^{(i)}$ לכל $P \in P_i$ וזה לכל $1 \le i \le k$. זו כמות הזרימה שאנחנו מזרימים למוצר i דרך המסלול P.
לכן אפשר לכתוב את בעיית התכנון הלינארי של מקסום הזרימה הכוללת כך:

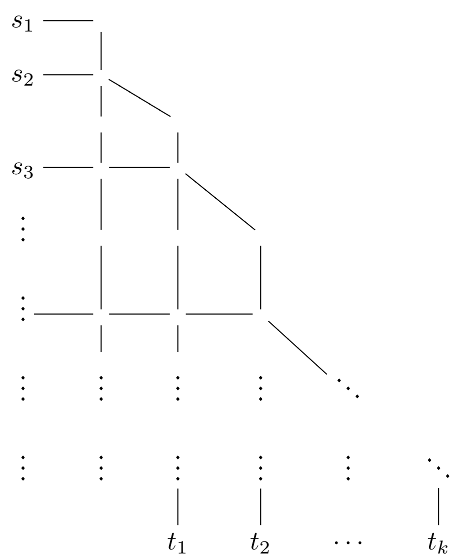
$$\begin{aligned} &\max \sum_{i=1}^{k} \sum_{P \in P_i} f_P^{(i)} \\ &\text{s.t.} \quad \forall (u,v) \in E \qquad \sum_{i=1}^{k} \sum_{P \in P_i} \sum_{\text{s.t. } (u,v) \in P} f_P^{(i)} \le c(u,v) \\ &f_P^{(i)} \ge 0 \end{aligned}$$

שימו לב שהתוכנית הלינארית הזו היא לא בגודל פולינומי בקלט, כי כאמור מספר המסלולים בגרף הוא לא פולינומי בגודל הקלט.

ולמרות זאת - מספר המשתנים שצריך להחזיק בזיכרון של המחשב, לדוגמה, הוא פולינומי. אפשר לדוגמה להריץ את אלגוריתם האליפסואיד על התוכנית הדואלית[6] (צריך לכסות כמה פרטים טכניים כדי להשתמש באליפסואיד, אבל הם טכניים ואפשריים), ולהסיק שהאלגוריתם בוחן מספר אילוצים פולינומי בלבד במהלך כל ריצתו.

אפשר גם לזרוק את כל האילוצים המיותרים בתוכנית הדואלית, כלומר אלה שלא נגענו בהם, "לחזור חזרה" לתוכנית הפרימאלית, ולקבל הפעם תוכנית שהיא פולינומית. נובע מכך שמספר פולינומי בלבד של משתנים בתוכנית הפרימאלית מחזיקים בערך שונה מ-0.

שימו לב שלמרות שכל הקיבולים שלמים - הזרימה המקסימלית אינה שלמה, שלא כמו במקרה של מוצר יחיד - ראינו דוגמה לכך בשיעור הקודם.
הינה דוגמה נוספת (כל הקיבולים הם 1):



זרימת המקסימום **בשלמים** תתן לכל צלע זרימה של 0 או 1. ולכן לא משנה איזה מוצר נבחר, ולא משנה איזה מסלול נבחר מ-s_i ל-t_i - אנחנו נחסום את כל הזרימה למוצרים האחרים (בגלל הצלעות "האלכסוניות" ו"ההסטה" של הבורות לעומת המקורות).

בזרימה **השברית** מזרימים $\frac{1}{2}$ במסלול "הברור": מכל s_i הולכים הכי ימינה שאפשר ואז הכי למטה שאפשר. הזרימה המקסימלית יוצאת כך $\frac{1}{2}k$ (כי זה $\frac{1}{2}$ לכל מוצר).

[6] אנחנו נדבר בהמשך על מהי התוכנית הדואלית לתוכנית הנ"ל, אבל התוכנית הדואלית תחזיק מספר אילוצים השווה למספר המסלולים בגרף (כי זו כמות המשתנים).

נסכם מה שראינו עד עכשיו: אם אנחנו מתעקשים להזרים זרימה בשלמים, צריך לשלם המון על חשבון הזרימה (כלומר לקבל זרימה שיכולה להיות מאוד מאוד קטנה ביחס לזרימה האופטימלית). אבל, מה אם נבחר לשנות את הרשת עצמה - בפרט את הקיבולים שלה?
בדוגמה שהובאה לעיל, אפשר להגדיל את הקיבולים של הרשת פי 2, וככה יתאפשר לנו להזרים זרימה **שלמה** של 1 לכל מוצר, לקבל k, ולהסיק (כי הכפלנו פי 2) שיש זרימה שברית של $\frac{k}{2}$ ברשת המקורית.

אז נניח שיש לנו פתרון שברי לתוכנית הלינארית. בכמה לפחות צריך להכפיל את c כדי שנוכל להזרים אותה כמות זרימה לפחות, רק בשלמים?
האינטואיציה של "ככל שנכפיל יותר את c נוכל להעביר יותר זרימה" היא אינטואיציה פחות או יותר נכונה, אבל היא לא כזו תמימה. הבעיה הזו עדיין פתוחה במדעי המחשב, אבל נראה פתרון עבור מקרה פרטי שלה.

נזכור שנתון לנו פתרון שברי $\left\{f_P^{(i)}\right\}$.
נניח כי $c\left(u,v\right)=1$ לכל $\{u,v\}\in E$, וגם כי $\sum\limits_{P\in P_i} f_P^{(i)}\leq 1$ לכל i (כלומר לכל מוצר אפשר להזרים זרימה של לכל היותר 1) - זה למשל מה שהיה לנו בדוגמה הקודמת.
כעת אפשר לחשוב על זה מנקודת גישה הסתברותית: בהסתברות $f_P^{(i)}$ אנחנו נזרים במסלול P זרימה (של 1), ולכן בהסתברות $1-\sum\limits_{p\in P_i} f_P^{(i)}$ לא תוזרם אף זרימה מ-s_i ל-t_i. נניח שעושים את הניסוי ההסתברותי הזה לכל מוצר בנפרד, באופן בלתי תלוי. ברגע שמזרימים ממסלול כלשהו ממוצר - לא מזרימים בכלל משאר המסלולים האחרים.
זה נותן לנו איזשהו פתרון של זרימה בשלמים $\widetilde{f}$, אבל ייתכן ש-2 מוצרים שונים יזרימו זרימה של 1 דרך צלע כלשהי e, והקיבול שלה הוא 1. זה מפר את אילוץ הקיבול: מיד נטפל בזה.
אך תחילה, תוחלת השטף של $\widetilde{f}$ היא $\mathbb{E}\left[\left|\widetilde{f}\right|\right]=\mathbb{E}\left[\sum\limits_{i=1}^{k}\sum\limits_{P\in P_i}\textbf{Pr}\left[P_i \text{ יש זרימה במסלול}\right]\right]$. אך $\textbf{Pr}\left[P_i \text{ יש זרימה במסלול}\right]=f_P^{(i)}$.

נקבל מלינאריות התוחלת, ש-$\mathbb{E}\left[\left|\widetilde{f}\right|\right]=\sum\limits_{i=1}^{k}\sum\limits_{P\in P_i} f_P^{(i)}$ - וזה בדיוק הערך שאותו ממקסמת התוכנית הלינארית השברית.

ננסה לטפל באילוץ הקיבול. נגדיר לכל $e\in E$ ולכל i את $f_e^{(i)}=\sum\limits_{P\in P_i \text{ s.t. } e\in P} f_P^{(i)}$, סך כל הזרימה בצלע e מהמוצר ה-i.
משום שהמאורעות "להזרים ב-P_1 זרימה וב-P_2" הם מאורעות זרים לשני מסלולים שונים P_1,P_2 תחת אותו מוצר (זכרו כי ברגע שאנחנו מזרימים זרימה דרך מסלול אחד, לא נזרים דרך אף מסלול אחר),
יתקיים $f_e^{(i)}=\sum\limits_{P\in P_i \text{ s.t. } e\in P} f_P^{(i)}=\textbf{Pr}\left[e \text{ זרימה של 1 מ-} s_i \text{ ל-} t_i \text{ עוברת דרך}\right]$.
על כן $\textbf{Pr}\left[e \text{ זרימה של 1 מ-} s_i \text{ ל-} t_i \text{ עוברת דרך}\right]=f_e^{(i)}=\sum\limits_{P\in P_i \text{ s.t. } e\in P} f_P^{(i)}\leq\sum\limits_{P\in P_i} f_P^{(i)}\leq 1$.

ומאילוץ הקיבול של התוכנית הלינארית השברית (אנחנו מניחים כי $\left\{f_P^{(i)}\right\}$ הוא פתרון שלה) נקבל $\sum\limits_{i=1}^{k} f_e^{(i)}\leq c\left(e\right)=1$.
בהינתן איזושהי צלע $e\in E$, יש לנו פה בעצם סכום של משתנים מקריים $X_1,X_2,...,X_k$ כך ש-$X_i\in\{0,1\}$ (אינדיקטור):
X_i מקבל את הערך "1" אם אנחנו בוחרים להזרים זרימה של 1 מ-s_i ל-t_i דרך e (באיזשהו מסלול), אחרת יקבל את הערך "0".
המשתנים X_i הם בלתי תלויים בזוגות, ו-$\textbf{Pr}\left[X_i=1\right]=\mathbb{E}\left[X_i\right]=f_e^{(i)}$.
נגדיר $X=\sum\limits_{i=1}^{k} X_i$. מלינאריות התוחלת $\mathbb{E}\left[X\right]=\sum\limits_{i=1}^{k} f_e^{(i)}\leq 1$ (ראינו).

משפט 1.2 (אי שוויון צ'רנוף Multiplicative Chernoff bound)**.** יהא $X=\sum\limits_{i=1}^{k} X_i$ כאשר $X_i\in\{0,1\}$ בלתי תלויים בזוגות, ו-$\textbf{Pr}\left[X_i=1\right]=q_i$. נסמן ב-$\mu=\mathbb{E}\left[X\right]=\sum\limits_{i=1}^{k} q_i$. אזי:

1. לכל $\delta > 0$ $\Pr\left[X \geq (1+\delta)\mu\right] \leq e^{-\frac{\delta^2}{2+\delta}\mu}$.
2. לכל $0 < \delta < 1$ מתקיים $\Pr\left[X \geq (1-\delta)\mu\right] \leq e^{-\frac{\delta^2}{2}\mu}$.
3. לכל $0 < \delta < 1$ מתקיים $\Pr\left[|X-\mu| \geq \delta\mu\right] \leq 2e^{-\frac{\delta^2}{3}\mu}$.

אנחנו נראה שאי שוויון צ'רנוף יעזור לנו להבין כמה בערך צריכה להיות δ כדי שההסתברות ש-2 מוצרים שונים יזרימו 1 תחת אותה הצלע, תהיה ממש זניחה.

הרצאה 16

הוכחה. נוכיח את 1 (2 אולי בשיעורי בית) נשים לב שהטענה השלישית נובעת מ-1 ומ-2 (כי $e^{-\frac{\delta^2}{2+\delta}\mu}, e^{-\frac{\delta^2}{2}\mu}$ קטנים מ-$e^{-\frac{\delta^2}{3}\mu}$).

תזכורת

פונקציה יוצרת מומנטים של משתנה מקרי X היא פונקציה $\mathbb{R} \to \mathbb{R}$ המוגדרת על ידי: $\lambda \mapsto \mathbb{E}\left[e^{\lambda X}\right]$.
הנגזרת ה-n-ית של הפונקציה הזו בנקודה 0 היא $\mathbb{E}\left[X^n\right]$.
אי שוויון מרקוב: אם X משתנה מקרי אי שלילי עם תוחלת סופית, לכל $0 < a$ מתקיים $\Pr\left[X \geq a\right] \leq \frac{\mathbb{E}[X]}{a}$.

נסתכל על הפונקציה יוצרת המומנטים של X. מתקיים:
לכל $\lambda > 0$, נזכור כי $\Pr\left[X \geq (1+\delta)\mu\right] = \Pr\left[e^{\lambda X} \geq e^{\lambda(1+\delta)\mu}\right]$. המשתנה המקרי $e^{\lambda X}$ הוא משתנה מקרי חיובי וגם $e^{\lambda(1+\delta)\mu}$ חיובי, ולכן אפשר להשתמש באי שוויון מרקוב.
נקבל $\Pr\left[e^{\lambda X} \geq e^{\lambda(1+\delta)\mu}\right] \leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda(1+\delta)\mu}}$. אבל $\mathbb{E}\left[e^{\lambda X}\right] = \mathbb{E}\left[e^{\lambda\sum_{i=1}^{n} X_i}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda X_i}\right]$. משום ש-$\{X_i\}$ בלתי תלויים, גם $\left\{e^{X_i}\right\}$, ולכן $\mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda X_i}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right]$.
X_i הם משתנים מקריים ברנוליים: ערכיהם 1 בהסתברות q_i ואחרת 0. לכן $\mathbb{E}\left[e^{\lambda X_i}\right] = q_i e^{\lambda \cdot 1} + (1-q)\, e^{\lambda \cdot 0}$.
נקבל $\mathbb{E}\left[e^{\lambda X}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right] = \prod_{i=1}^{n} \left(q_i e^{\lambda} + 1 - q_i\right) = \prod_{i=1}^{n} \left(1 + q_i\left(e^{\lambda} - 1\right)\right)$.
אבל $(1+x) \leq e^x$ לכל x, כך נקבל $\mathbb{E}\left[e^{\lambda X}\right] \leq \prod_{i=1}^{n} e^{q_i\left(e^{\lambda}-1\right)} = e^{\sum_{i=1}^{n} q_i\left(e^{\lambda}-1\right)} = e^{\left(e^{\lambda}-1\right)\sum_{i=1}^{n} q_i} = e^{\left(e^{\lambda}-1\right)\mu}$.
בסך הכל קיבלנו את הנוסחה הבאה: $\Pr\left[X \geq (1+\delta)\mu\right] \leq \frac{\exp\left(\left(e^{\lambda}-1\right)\mu\right)}{\exp(\lambda(1+\delta)\mu)}$.
נזכור שאי השוויון הנ"ל נכון לכל $0 < \lambda$. המטרה תהיה למזער את אגף ימין כמה שיותר כדי לחסום כמה שיותר טוב. אפשר לגזור את הפונקציה הזו, לבדוק איפה יש מינימום וכולי.
נבחר $\lambda = \ln(1+\delta)$. נקבל $\Pr\left[X \geq (1+\delta)\mu\right] \leq \frac{\exp((1+\delta-1)\mu)}{\exp(\ln(1+\delta)(1+\delta)\mu)} = \frac{e^{\delta\mu}}{\left(e^{\ln(1+\delta)}\right)^{(1+\delta)\mu}} = \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$.
כדי להראות ש-$\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}} \leq e^{-\frac{\delta^2}{2+\delta}}$ אפשר להראות א"ש שקול (מהוצאת log משני האגפים): $\delta - (1+\delta)\ln(1+\delta) \leq -\frac{\delta^2}{2+\delta}$. וזה שקול אל $\frac{2\delta^2+2\delta}{2+\delta} \leq (1+\delta)\log(1+\delta)$, כלומר $\frac{2\delta}{2+\delta} \leq \log(1+\delta)$ - וזה אי שוויון נכון.
הראנו שמתקיים $\Pr\left[X \geq (1+\delta)\mu\right] \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu} \leq \left(e^{-\frac{\delta^2}{2+\delta}}\right)^{\mu}$ כנדרש. □

בחזרה לבעיה המקורית שלנו: נתונה זרימה שברית ואנחנו רוצים להראות שיש זרימה בשלמים, ככה שמובטח לנו שהזרימה בכל צלע e היא לכל היותר $\alpha \cdot c(e)$ לאיזשהו α (כלומר המטרה היא להגדיל את c פי קבוע α).
אילו היה β כך שמתקיים ש-$\Pr\left[e\text{-ב זרימה} \geq \alpha \cdot c(e)\right] \leq \frac{1}{\beta|E|}$ (זה "מאורע רע"), אז מחסם האיחוד ההסתברות שיש צלע עם זרימה $\geq \alpha \cdot c(e)$ היא לכל היותר $\frac{1}{\beta}$.

נזכיר שראינו שלכל צלע e, אפשר לסמן ב-$X_1, \ldots, X_k$ משתנים מקריים ברנוליים, כך ש-$X_i = 1$ אם"ם יש זרימה מ-s_i ל-t_i שעוברת דרך i. נסמן משתנה מקרי לזרימה הכוללת $X = \sum_{i=1}^{k} X$ (כלומר הזרימה דרך e מיוצגת על ידי המשתנה X) ו-$\mathbb{E}[X] = \mu_e$. מקודם הראנו כי $\mu_e \leq 1$.

הבחירה הנכונה של α הוא $\alpha = c \cdot \frac{\log|E|}{\log(\log|E|)}$ לאיזשהו קבוע $0 < c$ (אין קשר לקיבול). זו למעשה הבחירה הטובה ביותר היום.

נסמן ב-μ_e את תוחלת הזרימה דרך e. נבחר δ שמקיימת $\alpha \cdot c(e) = (1+\delta) \cdot \mu_e$.

על כן מתקיים $\Pr\left[\mathbb{E}[X] \geq \alpha \cdot c(e)\right] = \Pr\left[\mathbb{E}[X] \geq (1+\delta) \cdot \mu_e\right]$.

נשתמש באי שוויון שראינו שהוא נכון במהלך הוכחת אי שוויון צ'רנוף:

$\Pr\left[\mathbb{E}[X] \geq (1+\delta) \cdot \mu_e\right] = \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu_e} \leq \left(\frac{e^{1+\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu_e} \leq \left(\frac{e}{1+\delta}\right)^{(1+\delta)\mu_e} = \left(\frac{e}{1+\delta}\right)^{\alpha}$.

ראינו כי $\mu_e \leq 1 = c(e)$. ומשום ש-$\alpha \cdot c(e) = (1+\delta) \cdot \mu_e$, נובע כי $\alpha \leq (1+\delta)$.

על כן נקבל $\Pr\left[e\text{-זרימה ב} \geq \alpha \cdot c(e)\right] \leq \left(\frac{e}{1+\delta}\right)^{\alpha} \leq \left(\frac{e}{\alpha}\right)^{\alpha}$.

נציב כעת $\alpha = c \cdot \frac{\log|E|}{\log(\log|E|)}$, ונקבל:

$\Pr\left[e\text{-זרימה ב} \geq \alpha \cdot c(e)\right] \leq \left(\frac{1}{\left(\frac{c}{e}\right) \cdot \frac{\log|E|}{\log(\log|E|)}}\right)^{\left(\frac{c}{e}\right) \cdot \frac{\log|E|}{\log(\log|E|)}} \leq \left(\frac{1}{\left(\frac{c}{e}\right) \cdot \frac{\log|E|}{\log(\log|E|)}}\right)^{c \cdot \frac{\log|E|}{\log(\log|E|)}}$.

נניח כי $c > e$, ואז $\Pr\left[\mathbb{E}[X] \geq \alpha \cdot c(e)\right] \leq \left(\frac{c \cdot \log|E|}{\log(\log|E|)}\right)^{-c \cdot \frac{\log|E|}{\log(\log|E|)}}$. נסמן $m = |E|$. ניקח log על אגף ימין.

זה יוצא $-\left(\frac{c}{e}\right) \cdot \frac{\log m}{\log(\log(m))} \cdot \log\left(\frac{c \log m}{\log(\log(m))}\right) = -c \cdot \frac{\log m}{\log(\log(m))} \cdot \left(\log(\log m) - \log(\log(\log(m))) + \log(c)\right)$.

זה שווה אל $-c\log(m) + c \cdot \frac{\log(\log(\log(m)))}{\log(\log(m))} - \frac{c\log(c)\log(m)}{\log(\log(m))}$, שקטן מ-$-c\log(m) + c \cdot \frac{\log(\log(\log(m)))}{\log(\log(m))}$.

מה שבאדום שואף ל-0, אז עבור m גדול מספיק, כל הביטוי הזה יוצא קטן מ-$-\frac{c}{2}\log(m)$.

אז ה-log של $\left(\frac{c \cdot \log|E|}{\log(\log|E|)}\right)^{-c \cdot \frac{\log|E|}{\log(\log|E|)}}$ הוא קטן מ-$-\frac{c}{2}\log(m)$. כלומר $\Pr\left[e\text{-זרימה ב} \geq \alpha \cdot c(e)\right] \leq e^{-\frac{c}{2}\log(m)} = \frac{1}{m^{c/2}}$.

וזה מה שרצינו (רצינו $\frac{1}{\beta|E|}$ וקיבלנו אפילו טוב יותר: המכנה פולינומי ב-$m = |E|$).

לסיכום, ראינו שבהינתן רשת זרימה רב מוצרית, אם ננפח את הקיבולים פי $\Theta\left(\frac{\log|E|}{\log(\log|E|)}\right)$, בהסתברות טובה יהיה ניתן להזרים זרימה בשלמים ברשת החדשה, השווה לזרימה המקסימלית בשברים ברשת המקורית.

2 קירוב לזרימת מקסימום ברשת רב מוצרית

כעת נדון במטרה אחרת והיא השגת זרימת מקסימום שברית ברשת רב מוצרית. כבר הזכרנו שהבעיה הזו **בשלמים** היא NP-קשה, אם כי הבעיה השברית לא - זו בעיית תכנון לינארי פשוטה.[7]
בפרק זה נציג אלגוריתם שמסוגל למצוא פתרון מקורב כרצוננו, בזמן ריצה טוב (אנחנו מתכננים אלגוריתם קירוב למרות שיש לנו דרך יעילה להשיג את האופטימום).

ניזכר בתוכנית הלינארית שמייצגת את בעיית הזרימה המקסימלית ברשת רב מוצרית (כאשר נעבוד בשיטת המסלולים להצגת התוכנית):

$$\begin{aligned}&\max \sum_{i=1}^{k}\sum_{P\in P_i} f_P^{(i)}\\ &\text{subject to}\quad \forall\{u,v\}\in E \qquad \sum_{i=1}^{k}\sum_{P\in P_i\ \text{s.t.}\ (u,v)\in P} f_P^{(i)}\le c(u,v)\\ &f_P^{(i)}\ge 0\end{aligned}$$

מה התוכנית הדואלית?

התוכנית הדואלית היא תוכנית מינימזציה. אנחנו נקרא למשתנים בה $\ell(e)$ (משום שהם הולכים להיות אורכים של צלעות).
בתוכנית הנוכחית וקטור ה-$\boldsymbol{b}$ שלנו הוא הוקטור של ה-$c(e)$ים. נזכור שבתוכנית דואלית ה-$\boldsymbol{b}$ הופך להיות פונקציית המטרה.
מבחינת האילוצים: יהיו לנו אילוצים לכל $1\le i\le k$ ו-P ב-P_i. פונקציית המטרה בתוכנית המקורית ($\boldsymbol{c}^T$) היא וקטור אחדים, ולכן היא ה-$\boldsymbol{b}$ החדשה.
כלומר התוכנית הדואלית היא למעשה הבאה:

$$\begin{aligned}&\min \sum_{e\in E} c(e)\,\ell(e)\\ &\text{s.t.}\forall 1\le i\le k \text{ and } \forall P\in P_i \qquad \sum_{e\in P}\ell(e)\ge 1\\ &\ell(e)\ge 0\end{aligned}$$

נניח שיש לנו פתרון $f_{\max}$ לתוכנית הלינארית של זרימה מקסימלית (כלומר קיבול של זרימה רב מוצרית מקסימלית).
בעזרת התוכנית הדואלית, אנחנו נראה שלכל $\varepsilon>0$ יש אלגוריתם מהיר לחישוב זרימה f, כך ש-$|f|\ge(1-2\varepsilon)f_{\max}$ - כלומר קירוב כרצוננו ($(1-2\varepsilon)$-קירוב).
את השיטה לעשות זאת נדגים בעזרת האלגוריתם של Garg ו-Konemann, אליו הם הגיעו לאחר שורה ארוכה של עבודות (קרוב ל-20 שנה).

2.1 אלגוריתם Garg & Konemann

האלגוריתם משתמש בעקיפין בתוכנית הלינארית הזו. הוא יקרב את הפתרון, באופן לא ישיר, בעזרת חישוב של פתרון **פרימאלי** ופתרון **דואלי** בו זמנית (ואז, בגלל הדואליות החלשה, מספיק להגיע לקירבה בין שתי פתרונות לפרימאלי ודואלי, כדי להבין שאנחנו קרובים לאופטימום).
אינטואיטיבית: בהינתן זוג פתרונות אופטימליים, אחד לתוכנית הפרימאלית ואחד לתוכנית הדואלית, רק מסלולים P בהם יש זרימה ($\sum_{P\in P_i\ \text{s.t.}\ (u,v)\in P} f_P^{(i)}>0$) הם מסלולים בהם מתקיים $\sum_{e\in P}\ell(e)=1$ (תמיד $\sum_{e\in P}\ell(e)\ge 1$), השאלה המעניינת מתי יש שוויון ממש), **וההפך**.
אנחנו נחשב זרימה f שתדרוש מאיתנו לעשות scaling ל-c, ככה ש-$\sum_{e\in P}\ell(e)$ יהיה קטן ככל היותר (ככה נגיע לפתרון כמה שיותר טוב לתוכנית הדואלית). הרצאה 17

[7] שימו לב לא לבלבל את הבעיה הזו עם בעיה אחרת שנקראת "maximum multicommodity flow **over time**" שהיא אכן בעיה NP-קשה.

אלגוריתם 2 אלגוריתם Garg & Konemann ל-$(1-2\varepsilon)$-קירוב של זרימה רב מוצרית מקסימלית

במהלך ריצת האלגוריתם אין צורך להחזיק את ערכי כל המשתנים - רק את אלה ששונים מ-0.

1. אתחול: $f_P^{(i)} \leftarrow 0$ לכל $i=1,...,k$ ו-$P\in P_i$.
$\ell(e)\leftarrow\delta$ עבור δ שנגדיר בהמשך (זה יהיה $\frac{1+\varepsilon}{((1+\varepsilon)n)^{1/\varepsilon}}$).
2. **כל עוד** יש i ו-$P\in P_i$ כך ש-$\sum\limits_{e\in P}\ell(e)<1$: (כלומר התוכנית הדואלית היא לא פיזיבילית)
   (א) נגדיר $c_{\min}\leftarrow\min\limits_{e\in P}c(e)$.
   (ב) שינוי התוכנית הפרימאלית: $f_P^{(i)}\leftarrow f_P^{(i)}+c_{\min}$.
   (ג) שינוי התוכנית הדואלית: לכל $e\in P$ נגדיר $\ell(e)\leftarrow\left(1+\varepsilon\cdot\frac{c_{\min}}{c(e)}\right)\ell(e)$.
3. הפיכת הזרימה לפיזיבילית: נגדיר $M\leftarrow\max\limits_{e\in E}\frac{\sum\limits_i\sum\limits_{P\in P_i}f_P^{(i)}}{c(e)}$.
4. נחזיר את $\frac{1}{M}f$.

שימו לב שכדי לבדוק את התנאי ב-2 לא צריך לבדוק כל מסלול! זה יקח לנו זמן אקספוננציאלי. לחילופין, נריץ אלגוריתם למציאת מסלולים קצרים ביותר (עם משקלים ℓ).

טענה 2.1. נסמן $m=|E|$. אז מספר האיטרציות של לולאת ה-while הוא לכל היותר $m\log_{1+\varepsilon}\frac{1+\varepsilon}{\delta}$, ובפרט סופי.

הוכחה. בכל איטרציה (נניח שעל המוצר ה-i והמסלול $P\in P_i$), לכל $e\in P_i$, $\ell(e)$ גדל בפקטור של $1+\varepsilon\cdot\frac{C_{\min}}{c(e)}$, שהוא לכל היותר $1+\varepsilon$, מההגדרה. לפני העדכון, $\ell(e)\le\sum\limits_{e\in P}\ell(e)<1$, וכך לכל $e\in P_i$ מתקיים $\ell(e)\le1+\varepsilon$.
מתקיים בהתחלה $\ell(e)=\delta$, ובכל איטרציה יש לפחות צלע אחת e עבורה אנו מגדילים את $\ell(e)$ **בדיוק פי** $1+\varepsilon$ (יש צלע e שיוצאת המינימום ב-$c_{\min}=\min\limits_{e\in P}c(e)$).
אז בכל איטרציה יש לפחות צלע אחת e כך ש-$\ell(e)$ גדל בפקטור $1+\varepsilon$ בדיוק שנכנה אותה "הצלע הקריטית"[8]. כמה פעמים כל צלע e יכולה להיות "הצלע הקריטית" לכל היותר? אם אחרי t פעמים שהצלע היא צלע קריטית, מתקיים $\ell(e)\ge\delta(1+\varepsilon)^t$, אבל ראינו גם כי תמיד $\ell(e)\le1+\varepsilon$. לכן, $(1+\varepsilon)^t\delta\le1+\varepsilon$, נקבל $t\le\log_{1+\varepsilon}\frac{1+\varepsilon}{\delta}$.
בכל איטרציה יש צלע קריטית, כל צלע יכולה לשמש לכל היותר $\log_{1+\varepsilon}\frac{1+\varepsilon}{\delta}$ פעמים צלע קריטית, ויש m צלעות, ולכן מספר האיטרציות חסום על ידי $m\log_{1+\varepsilon}\frac{1+\varepsilon}{\delta}$ כנדרש. □

הלמה הבאה אומרת ש-M לא גדול מדי:

למה 2.2. $M\le\log_{1+\varepsilon}\frac{1+\varepsilon}{\delta}$.

הוכחה. נסמן ב-P_r את המסלול שנבחר באיטרציה ה-r של לולאת ה-while, ו-c_r הוא ערך ה-$c_{\min}$ באותה איטרציה.
בהינתן $e\in P$, $\ell(e)$ גדל בפקטור של $1+\varepsilon\cdot\frac{c_r}{c(e)}$, ומאותחל להיות δ, ולכן $\ell(e)=\delta\cdot\prod\limits_{r\text{ s.t. }e\in P_r}\left(1+\varepsilon\cdot\frac{c_r}{c(e)}\right)$.
אבל $\frac{c_r}{c(e)}\le1$ ולכן מאי שוויון ברנולי נקבל $(1+\varepsilon)^{\frac{c_r}{c(e)}}\ge1+\varepsilon\cdot\frac{c_r}{c(e)}$.
כלומר $\ell(e)=\delta\cdot\prod\limits_{r\text{ s.t. }e\in P_r}\left(1+\varepsilon\cdot\frac{c_r}{c(e)}\right)\ge\ell(e)=\delta\cdot\prod\limits_{r\text{ s.t. }e\in P_r}(1+\varepsilon)^{\frac{c_{\min}}{c(e)}}=\delta(1+\varepsilon)^{\sum\limits_{r\text{ s.t. }e\in P_r}\frac{c_r}{c(e)}}$.
בנוסף $\ell(e)\le1+\varepsilon$. קיבלנו $\delta(1+\varepsilon)^{\sum\limits_{r\text{ s.t. }e\in P_r}\frac{c_r}{c(e)}}\le1+\varepsilon$.
כלומר $(1+\varepsilon)^{\sum\limits_{r\text{ s.t. }e\in P_r}\frac{c_r}{c(e)}}\le\frac{1+\varepsilon}{\delta}$ ומלקיחת $\log_{1+\varepsilon}(\cdot)$ על שני האגפים $\sum\limits_{r\text{ s.t. }e\in P_r}\frac{c_r}{c(e)}\le\log_{1+\varepsilon}\left(\frac{1+\varepsilon}{\delta}\right)$.

[8] המונח לא הוזכר בהרצאה, אני הוספתי אותו כדי שהוכחה תהיה ברורה.

המכנה $c(e)$ משותף, אז קיבלנו $\frac{\sum_{r \text{ s.t. } e\in P_r} c_r}{c(e)} \le \log_{1+\varepsilon}\left(\frac{1+\varepsilon}{\delta}\right)$. אבל, בגלל שורה 2ב' והעובדה שהזרימה מאותחלת תמיד ל-0, יתקיים ש-$\sum_{r \text{ s.t. } e\in P_r} c_r$ היא סך כל הזרימה דרך e.

קיבלנו כי $\frac{\text{סך כל הזרימה דרך } e}{c(e)} \le \log_{1+\varepsilon}\left(\frac{1+\varepsilon}{\delta}\right)$ ולכן $\max_{e\in E}\frac{\text{סך כל הזרימה דרך } e}{c(e)} \le \log_{1+\varepsilon}\left(\frac{1+\varepsilon}{\delta}\right)$, וכך $M \le \log_{1+\varepsilon}\left(\frac{1+\varepsilon}{\delta}\right)$. $\square$

מסקנה 2.3. אם $M = \log_{1+\varepsilon}\frac{1+\varepsilon}{\delta}$, אז הזרימה $\frac{1}{M}f$ עומדת באילוץ הקיבול, כי היחס בין זרימה לקיבול בכל צלע הוא לכל הפחות M. הרצאה 18

משפט 2.4. לבחירה טובה של δ, האלגוריתם מחשב זרימה מגודל לכל הפחות $(1-2\varepsilon)\,\text{OPT}$.

הוכחה. נסמן ב-ℓ את פונקציית האורך על הצלעות. זה הפתרון הדואלי שאנחנו מחשבים במהלך ריצת האלגוריתם (זכרו שעד שהאלגוריתם עוצר, לא מדובר בפתרון פיזיבילי לתוכנית הדואלית). נסמן ב-$D(\ell)$ את פונקציית המטרה הדואלית, $D(\ell) = \sum_{e\in E} c(e)\,\ell(e)$, וב-$\alpha(\ell)$ את $\alpha(\ell) = \min_{\substack{1\le i\le k\\ P\in P_i}} \sum_{e\in P}\ell(e)$. נשים לב שתנאי העצירה באלגוריתם הוא כש-$\alpha(\ell) > 1$.

נסמן ב-$\ell^{(r)}$ את ערך הפונקציה ℓ בסוף האיטרציה ה-r, ו-$f^{(r)}$ את ערך הזרימה f בסוף האיטרציה ה-r. ב-T נסמן את מספר האיטרציה האחרונה.
ההוכחה תהיה בחלקים.

- נסמן ב-$\beta = \min_{\text{dual feasible } \ell} D(\ell)$, הערך האופטימלי של התוכנית הדואלית, וב-ℓ^* את ה-argmin. ומדואליות חזקה, β הוא ערך הזרימה הרב מוצרית המקסימלית (OPT).

  נטען כי $\beta = \min_{\boldsymbol{\ell}\ge\mathbf{0} \text{ s.t. } \alpha(\boldsymbol{\ell})\neq 0}\frac{D(\boldsymbol{\ell})}{\alpha(\boldsymbol{\ell})}$ (כאשר $\boldsymbol{\ell}$ הוא וקטור שמייצג פונקציית אורכי צלעות). נסמן ב-ℓ^{**} את $\operatorname{argmin}_{\boldsymbol{\ell}\ge\mathbf{0} \text{ s.t. } \alpha(\boldsymbol{\ell})\neq 0}\frac{D(\boldsymbol{\ell})}{\alpha(\boldsymbol{\ell})}$.

  עבור ℓ^* אנחנו יודעים כי $\alpha(\ell^*) \ge 1$ (שכן ℓ^* הוא פתרון לתוכנית הדואלית, ו-$\alpha(\ell^*) \ge 1$ הוא אילוץ בה), ולכן גם $\alpha(\ell^*) \neq 0$. בנוסף אפילו יש שוויון: $\alpha(\ell^*) = 1$, אחרת $\frac{1}{\alpha(\ell^*)}\ell^*$ הוא פתרון לתוכנית הדואלית, והוא טוב יותר מ-ℓ^* (ואילו ℓ^* פתרון אופטימלי לתוכנית הדואלית).

  כלומר $D(\ell^*) = \frac{D(\ell^*)}{1} = \frac{D(\ell^*)}{\alpha(\ell^*)}$, אבל $\ell^* \ge \mathbf{0}$, ובנוסף $\alpha(\ell^*) \neq 0$. ניזכר כי $\ell^{**} = \operatorname{argmin}_{\boldsymbol{\ell}\ge\mathbf{0} \text{ s.t. } \alpha(\boldsymbol{\ell})\neq 0}\frac{D(\boldsymbol{\ell})}{\alpha(\boldsymbol{\ell})}$, ולכן נקבל כי $\frac{D(\ell^{**})}{\alpha(\ell^{**})} \le D(\ell^*)$.

  מצד שני, ניתן לראות כי $\frac{1}{\alpha(\ell^{**})}\ell^{**}$ הוא פתרון פיזיבילי לתוכנית הדואלית, ולכן $D\left(\frac{1}{\alpha(\ell^{**})}\ell^{**}\right) \ge D(\ell^*)$ (כי ℓ^* פתרון אופטימלי), אבל $D\left(\frac{1}{\alpha(\ell^{**})}\ell^{**}\right) = \frac{1}{\alpha(\ell^{**})}D(\ell^{**})$, ולכן קיבלנו $\frac{D(\ell^{**})}{\alpha(\ell^{**})} \ge D(\ell^*)$.

  על כן $\frac{D(\ell^{**})}{\alpha(\ell^{**})} = D(\ell^*)$, ולכן מתקיים ש-$\beta = \min_{\boldsymbol{\ell}\ge\mathbf{0} \text{ s.t. } \alpha(\boldsymbol{\ell})\neq 0}\frac{D(\boldsymbol{\ell})}{\alpha(\boldsymbol{\ell})}$.

- נזכור כי $\ell^{(T)}$ הוא ערך הפונקציה ℓ שהאלגוריתם מחשב בסוף האיטרציה האחרונה, ואנחנו יודעים מתנאי העצירה כי $\alpha\left(\ell^{(T)}\right) \ge 1$. אנחנו נראה ש-$\alpha\left(\ell^{(T)}\right) \le \delta\cdot n\cdot e^{\frac{\varepsilon\left|f^{(T)}\right|}{\text{OPT}}}$. נשים לב שברגע שאנחנו מוכיחים את זה, נקבל $1 \le \delta\cdot n\cdot e^{\frac{\varepsilon\left|f^{(T)}\right|}{\text{OPT}}}$, ואז $e^{\frac{\varepsilon\left|f^{(T)}\right|}{\text{OPT}}} \ge \frac{1}{\delta n}$. מלקיחת log על שני האגפים $\frac{\varepsilon\left|f^{(T)}\right|}{\text{OPT}} \ge \log\left(\frac{1}{\delta n}\right)$, ואז $\left|f^{(T)}\right| \ge \frac{\log\left(\frac{1}{\delta n}\right)\text{OPT}}{\varepsilon}$, ומחלוקה של שני האגפים ב-M, נקבל $\frac{\left|f^{(T)}\right|}{M} \ge \frac{\log\left(\frac{1}{\delta n}\right)}{\varepsilon M}\cdot\text{OPT}$. באגף שמאל - השטף של הזרימה שאנחנו מחזירים, ובאגף ימין מופיע לנו β שהוא ערך הזרימה האופטימלית. לכן כל מה שאנחנו צריכים לעשות זה לדאוג ש-$\frac{\log\left(\frac{1}{\delta n}\right)}{\varepsilon M} \ge 1-2\varepsilon$. ואז נקבל $|f| \ge (1-2\varepsilon)\,\text{OPT}$ כמו שרצינו.

  נבחר את δ להיות $\delta = \frac{1+\varepsilon}{((1+\varepsilon)n)^{1/\varepsilon}}$, מתקיים $\log\left(\frac{1}{\delta n}\right) = \log\left(((1+\varepsilon)n)^{1/\varepsilon-1}\right) = \left(\frac{1}{\varepsilon}-1\right)\log((1+\varepsilon)n)$.

  מנגד, $M \le \log_{1+\varepsilon}\frac{1+\varepsilon}{\delta} = \frac{\log\left(\frac{1+\varepsilon}{\delta}\right)}{\log(1+\varepsilon)} = \frac{\log\left(((1+\varepsilon)n)^{1/\varepsilon}\right)}{\log(1+\varepsilon)} = \frac{\frac{1}{\varepsilon}\cdot(\log(1+\varepsilon)n)}{\log(1+\varepsilon)}$.

עלינו להראות ש- $\frac{\log\left(\frac{1}{\delta n}\right)}{\varepsilon M} \geq 1-2\varepsilon$. ממה שראינו מתקיים $\left(\frac{1}{\varepsilon}-1\right)\log\left(1+\varepsilon\right) = \frac{\left(\frac{1}{\varepsilon}-1\right)\log((1+\varepsilon)n)}{\varepsilon\cdot\frac{\frac{1}{\varepsilon}\cdot(\log(1+\varepsilon)n)}{\log(1+\varepsilon)}} \leq \frac{\log\left(\frac{1}{\delta n}\right)}{\varepsilon M}$.

אבל $\log\left(1+\varepsilon\right) \geq \varepsilon-\varepsilon^2$, אפשר לראות את זה מפיתוח טיילור, וכך נקבל $\left(\frac{1}{\varepsilon}-1\right)\left(\varepsilon-\varepsilon^2\right) \leq \frac{\log\left(\frac{1}{\delta n}\right)}{\varepsilon M}$.

קיבלנו כי $1-2\varepsilon \leq 1-\varepsilon-\varepsilon+\varepsilon^2 \leq \frac{\log\left(\frac{1}{\delta n}\right)}{\varepsilon M}$.

- כעת נוכיח את אי השוויון $\alpha\left(\ell^{(T)}\right) \leq \delta\cdot n\cdot e^{\frac{\varepsilon\left|f^{(T)}\right|}{\text{OPT}}}$, שכמו שראינו הרגע - מסיים לנו את הניתוח.

  – נראה איך הערך $D\left(\ell^{(r)}\right)$ (כפונקציה של r) משתנה עם האיטרציות (r). מתקיים $D\left(\ell^{(r)}\right) = \sum_{e\in E} c\left(e\right)\ell\left(e\right)$.

באיטרציה ה-r, נסמן את המסלול הנבחר ב-P_r ואת ה-$c_{\min}$ ב-c_r. הצלעות של P_r הצלעות שאנחנו משנים את ערך ה-ℓ שלהן ($\ell^{(r)}\left(e\right) = \left(1+\varepsilon\cdot\frac{c_r}{c(e)}\right)\ell^{(r-1)}\left(e\right)$), וערך ה-$\ell$ של כל השאר נשאר זהה, $\ell^{(r)}\left(e\right) = \ell^{(r-1)}\left(e\right)$.

אז $D\left(\ell^{(r)}\right) = \sum_{e\notin P_r} c\left(e\right)\ell^{(r-1)}\left(e\right) + \sum_{e\in P_r} c\left(e\right)\left(1+\varepsilon\cdot\frac{c_r}{c(e)}\right)\ell^{(r-1)}\left(e\right)$.

מפתיחת הסוגריים $D\left(\ell^{(r)}\right) = \underbrace{\sum_{e\notin P_r} c\left(e\right)\ell^{(r-1)}\left(e\right) + \sum_{e\in P_r} c\left(e\right)\ell^{(r-1)}\left(e\right)}_{D\left(\ell^{(r-1)}\right)} + \varepsilon c_r\sum_{e\in P_r}\ell^{(r-1)}\left(e\right)$.

אבל c_r הוא הזרימה הנוספת שהזרמנו באיטרציה ה-r, מהגדרה (2ב' באלגוריתם), אז זה השינוי בזרימה: $\left|f^{(r)}\right|-\left|f^{(r-1)}\right|$.

אפשר לבחור את P_r להיות המסלול בעל האורך המינימלי לפי פונקציות המרחקים ℓ (זה גם רשום מתחת לאלגוריתם). ואז (נזכיר כי $\alpha\left(\ell\right) = \min_{\substack{1\leq i\leq k\\ P\in P_i}}\sum_{e\in P}\ell\left(e\right)$) נקבל $\sum_{e\in P_r}\ell^{(r-1)}\left(e\right) = \alpha\left(\ell^{(r-1)}\right)$.

בסך הכל קיבלנו כי $D\left(\ell^{(r)}\right) = D\left(\ell^{(r-1)}\right) + \varepsilon\cdot\left(\left|f^{(r)}\right|-\left|f^{(r-1)}\right|\right)\cdot\alpha\left(\ell^{(r-1)}\right)$.

קיבלנו נוסחה רקורסיבית ($D\left(\ell^{(r)}\right)$ רשום במונחי $D\left(\ell^{(r-1)}\right)$).

אפשר לפתוח אותה, וככה לקבל $D\left(\ell^{(r)}\right) = D\left(\ell^{(0)}\right) + \varepsilon\cdot\sum_{s=1}^{r}\left(\left|f^{(s)}\right|-\left|f^{(s-1)}\right|\right)\alpha\left(\ell^{(s-1)}\right)$.

  – נגדיר את הפונקציה $\ell^{(r)}-\ell^{(0)}$, בדיוק כמו שהיא נשמעת - האורך של הצלע e תחת הפונקציה הזו, הוא $\ell^{(r)}\left(e\right)-\ell^{(0)}\left(e\right)$ (בפרט מדובר באורך חיובי, שכן ערכי ℓ גדלים בין האיטרציות).

$D\left(\ell^{(r)}-\ell^{(0)}\right) = \sum_{e\in E} c\left(e\right)\left(\ell^{(r)}-\ell^{(0)}\right)\left(e\right) = \sum_{e\in E} c\left(e\right)\ell^{(r)}\left(e\right) - \sum_{e\in E} c\left(e\right)\ell^{(0)}\left(e\right) = D\left(\ell^{(r)}\right) - D\left(\ell^{(0)}\right)$.

כלומר $D\left(\ell^{(r)}-\ell^{(0)}\right) = D\left(\ell^{(r)}\right)-D\left(\ell^{(0)}\right) = \varepsilon\cdot\sum_{s=1}^{r}\left(\left|f^{(s)}\right|-\left|f^{(s-1)}\right|\right)\alpha\left(\ell^{(s-1)}\right)$.

בנוסף $\alpha\left(\ell^{(r)}-\ell^{(0)}\right) = \min_{\substack{1\leq i\leq k\\ P\in P_i}}\sum_{e\in P}\left(\ell^{(r)}-\ell^{(0)}\right)\left(e\right) = \min_{\substack{1\leq i\leq k\\ P\in P_i}}\left(\sum_{e\in P}\ell^{(r)}\left(e\right)-\sum_{e\in P}\ell^{(0)}\left(e\right)\right)$.

בהתחלה הגדרנו $\ell^{(0)}\left(e\right) = \delta$ לכל צלע e, אז $\alpha\left(\ell^{(r)}-\ell^{(0)}\right) = \min_{\substack{1\leq i\leq k\\ P\in P_i}}\left(\sum_{e\in P}\ell^{(r)}\left(e\right)-\sum_{e\in P}\delta\right)$. P הוא מסלול פשוט, אורכו הוא לכל היותר $n = |V|$ (למעשה $|V|-1$, אבל בסדר), וכך נקבל $-\sum_{e\in P}\delta \geq -\delta\cdot n$,

כלומר $\alpha\left(\ell^{(r)}-\ell^{(0)}\right) = \min_{\substack{1\leq i\leq k\\ P\in P_i}}\left(\sum_{e\in P}\ell^{(r)}\left(e\right)-\sum_{e\in P}\delta\right) \geq \min_{\substack{1\leq i\leq k\\ P\in P_i}}\left(\sum_{e\in P}\ell^{(r)}\left(e\right)-\delta n\right) =$
$= -\delta n + \min_{\substack{1\leq i\leq k\\ P\in P_i}}\sum_{e\in P}\ell^{(r)}\left(e\right) = -\delta n + \alpha\left(\ell^{(r)}\right)$.

נסתכל על הפונקציה $\ell^{(T)}-\ell^{(0)}$ (נזכיר כי T היא האיטרציה האחרונה). מתקיים $\alpha\left(\ell^{(T)}-\ell^{(0)}\right) > 0$, כי אין מסלול מ-$s_i$ ל-t_i (לאף i) עם אורך שהוא עם אורך קטן מ-1 (מתנאי העצירה).

מקודם הראנו כי $\mathrm{OPT}=\min_{\boldsymbol{\ell}\geq\mathbf{0}\text{ s.t. }\alpha(\boldsymbol{\ell})\neq 0}\frac{D(\boldsymbol{\ell})}{\alpha(\boldsymbol{\ell})}$, ועכשיו הסברנו מדוע $\alpha\left(\ell^{(T)}-\ell^{(0)}\right)>0$. נשתמש בזה,

ראינו $D\left(\ell^{(r)}-\ell^{(0)}\right)=\varepsilon\cdot\sum_{s=1}^{r}\left(\left|f^{(s)}\right|-\left|f^{(s-1)}\right|\right)\alpha\left(\ell^{(s-1)}\right)$ וגם $\alpha\left(\ell^{(r)}-\ell^{(0)}\right)\geq-\delta n+\alpha\left(\ell^{(r)}\right)$:

נציב את זה עבור $r=T$, ונקבל את אי השוויון $\frac{D\left(\ell^{(T)}-\ell^{(0)}\right)}{\alpha\left(\ell^{(T)}-\ell^{(0)}\right)}\leq\frac{\varepsilon\cdot\sum_{s=1}^{T}\left(\left|f^{(s)}\right|-\left|f^{(s-1)}\right|\right)\alpha\left(\ell^{(s-1)}\right)}{-\delta n+\alpha\left(\ell^{(T)}\right)}$.

נקבל כי $\mathrm{OPT}=\min_{\boldsymbol{\ell}\geq\mathbf{0}\text{ s.t. }\alpha(\boldsymbol{\ell})\neq 0}\frac{D(\boldsymbol{\ell})}{\alpha(\boldsymbol{\ell})}\leq\frac{D\left(\ell^{(T)}-\ell^{(0)}\right)}{\alpha\left(\ell^{(T)}-\ell^{(0)}\right)}\leq\varepsilon\cdot\frac{\sum_{s=1}^{T}\left(\left|f^{(s)}\right|-\left|f^{(s-1)}\right|\right)\alpha\left(\ell^{(s-1)}\right)}{-\delta n+\alpha\left(\ell^{(T)}\right)}$.

מבחירת δ שלנו מתקיים $\delta n=\left((1+\varepsilon)\,n\right)^{1-1/\varepsilon}<1\leq\alpha\left(\ell^{(T)}\right)$, כלומר $-\delta n+\alpha\left(\ell^{(T)}\right)>0$.

לפיכך (כפל שני האגפים במספר חיובי), נקבל $\left(-\delta n+\alpha\left(\ell^{(T)}\right)\right)\leq\frac{\varepsilon\cdot\left(\sum_{s=1}^{T}\left(\left|f^{(s)}\right|-\left|f^{(s-1)}\right|\right)\alpha\left(\ell^{(s-1)}\right)\right)}{\mathrm{OPT}}$.

כלומר, $\alpha\left(\ell^{(T)}\right)\leq\delta n+\frac{\varepsilon}{\mathrm{OPT}}\cdot\left(\sum_{s=1}^{T}\left(\left|f^{(s)}\right|-\left|f^{(s-1)}\right|\right)\alpha\left(\ell^{(s-1)}\right)\right)$.

– עכשיו נבנה נוסחה רקורסיבית שתאמר לנו מה יקרה כשיש בדיוק שוויון בא"ש הנ"ל, ומשום שאין באמת שוויון אלא רק $\leq$ נוכל להיעזר בזה כדי לקבל חסם מלמעלה.

נגדיר רקורסיבית את הפונקציה y, שהיא $y\left(\ell^{(0)}\right)=\delta n$

ו-$y\left(\ell^{(r)}\right)=\delta n+\frac{\varepsilon}{\mathrm{OPT}}\cdot\left(\sum_{s=1}^{r}\left(\left|f^{(s)}\right|-\left|f^{(s-1)}\right|\right)\cdot y\left(\ell^{(s-1)}\right)\right)$. נשים לב שמההגדרה, $\alpha\left(\ell^{(r)}\right)\leq y\left(\ell^{(r)}\right)$ לכל r (אפשר להוכיח את זה אינדוקטיבית).

אפשר לרשום אותה גם (דרך פיצול הסכום לסכום של $r-1$ הראשונים + האחרון):

$$y\left(\ell^{(r)}\right)=\underbrace{\delta n+\frac{\varepsilon}{\mathrm{OPT}}\cdot\left(\sum_{s=1}^{r-1}\left(\left|f^{(s)}\right|-\left|f^{(s-1)}\right|\right)\cdot y\left(\ell^{(s-1)}\right)\right)}_{y\left(\ell^{(r-1)}\right)}+\frac{\varepsilon}{\mathrm{OPT}}\cdot\left(\left(\left|f^{(r)}\right|-\left|f^{(r-1)}\right|\right)\cdot y\left(\ell^{(r-1)}\right)\right).$$

קיבלנו $y\left(\ell^{(r)}\right)=y\left(\ell^{(r-1)}\right)\cdot\left(1+\frac{\varepsilon\left(\left|f^{(r)}\right|-\left|f^{(r-1)}\right|\right)}{\mathrm{OPT}}\right)$.

ניזכר באי השוויון $1+x\leq e^{x}$, ונקבל $y\left(\ell^{(r)}\right)\leq y\left(\ell^{(r-1)}\right)\cdot e^{\frac{\varepsilon\left(\left|f^{(r)}\right|-\left|f^{(r-1)}\right|\right)}{\mathrm{OPT}}}$.

שוב יש לנו פה נוסחה רקורסיבית: נקבל $y\left(\ell^{(r)}\right)\leq y\left(\ell^{(0)}\right)\cdot\prod_{s=1}^{r}e^{\frac{\varepsilon\left(\left|f^{(s)}\right|-\left|f^{(s-1)}\right|\right)}{\mathrm{OPT}}}$.

אבל $\prod_{s=1}^{r}e^{\frac{\varepsilon\left(\left|f^{(s)}\right|-\left|f^{(s-1)}\right|\right)}{\mathrm{OPT}}}=e^{\frac{\sum_{s=1}^{r}\varepsilon\left(\left|f^{(s)}\right|-\left|f^{(s-1)}\right|\right)}{\mathrm{OPT}}}$, והסכום $\sum_{s=1}^{r}\varepsilon\left(\left|f^{(s)}\right|-\left|f^{(s-1)}\right|\right)$ הוא טלסקופי: אפשר לראות שהאיברים בו מתבטלים, כך שנשארים רק עם $\varepsilon\left|f^{(r)}\right|-\varepsilon\left|f^{(0)}\right|$. אבל $\left|f^{(0)}\right|=0$ מהגדרה ($f^{(0)}$ היא זרימת האפס).

קיבלנו כי $y\left(\ell^{(r)}\right)\leq y\left(\ell^{(0)}\right)\cdot e^{\frac{\varepsilon\cdot\left|f^{(r)}\right|}{\mathrm{OPT}}}$, ובפרט $y\left(\ell^{(T)}\right)\leq y\left(\ell^{(0)}\right)\cdot e^{\frac{\varepsilon\cdot\left|f^{(T)}\right|}{\mathrm{OPT}}}$. הגדרנו $y\left(\ell^{(0)}\right)=\delta n$.

ולכן קיבלנו $y\left(\ell^{(T)}\right)\leq\delta n\cdot e^{\frac{\varepsilon\cdot\left|f^{(T)}\right|}{\mathrm{OPT}}}$, אבל הסברנו (בהגדרת y) ש-$\alpha\left(\ell^{(T)}\right)\leq y\left(\ell^{(T)}\right)$. לכן $\alpha\left(\ell^{(T)}\right)\leq\delta\cdot n\cdot e^{\frac{\varepsilon\left|f^{(T)}\right|}{\beta}}$, כמו שרצינו. □

אלגוריתם גארג-קונמן משתמש בגישה נפוצה לפתרון בעיות קירוב: משתמשים בדואליות בין 2 בעיות, כאשר באלגוריתם עצמו כל עוד אילוץ כלשהו בבעיה הדואלית מופר (אצלנו: כל עוד האורך של איזשהו מסלול קטן מ-1) - מגדילים את הערך של פונקציית המטרה הדואלית כדי שהאילוץ לא יופר עוד, ובמקביל משנים גם את הבעיה הפרימאלית. בסוף מנרמלים הכל.

3 רב חתך (multicut)

עד עכשיו ראינו שבבעיות זרימה רב מוצרית, רשתות עם קיבולים שלמים לא מחייבות זרימה אופטימלית שלמה, וראינו אלגוריתם לא רע לחישוב זרימת מקסימום מקורבת.
הבעיה הדואלית לבעיית הזרימה החד מוצרית היא בעיית החתך המינימלי (נזכיר כי הרעיון של חתך הוא שאם מורידים צלע אחת בחתך (כלומר צלע המחברת בין קודקוד שבחתך לקודקוד שאינו בחתך), אי אפשר להזרים זרימה מ-s ל-t).
מה קורה ברשתות רב מוצריות?

האם בעיית הזרימה האופטימלית הרב מוצרית היא דואלית לבעיית הרב-חתך (multi-cut) המינימלי?

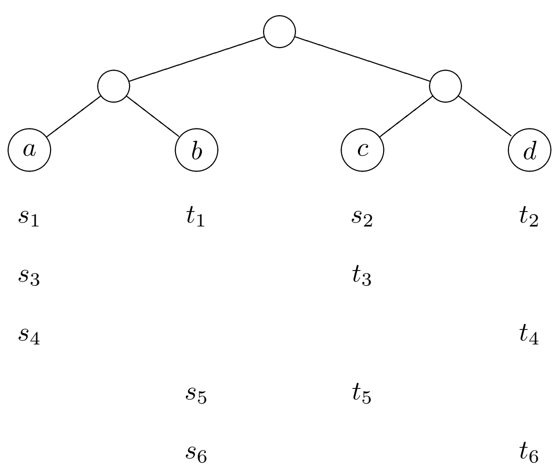
הגדרה 3.1. רב-חתך (multicut) הוא קבוצת צלעות, כך שהסרת כל הצלעות בקבוצה הזו, מביאה לגרף חדש בו מנותקים **כל** זוגות s_i, t_i.

ניזכר מהי הבעיה הדואלית:

$$
\begin{aligned}
&\min \sum_{e\in E} c(e)\,\ell(e) \\
&\text{s.t.}\forall 1\le i\le k \text{ and } \forall P\in P_i \qquad \sum_{e\in P}\ell(e)\ge 1 \\
&\ell(e)\ge 0
\end{aligned}
$$

הבעיה של זרימה אופטימלית דואלית לבעיית הרב-חתך **<u>השברי</u>** המינימלי (minimum fractional multicut), כלומר מה קורה אם אפשר לבחור להסיר, למשל, "רבע צלע". ההסרה של הצלעות מתבטאת על ידי ℓ - כאשר האינטואיציה היא שכל צלע עם $\ell = 1$ מוסרת, וכאמור הדרישה המתבטאת באילוץ היא שבכל מסלול ממקור s_i ל-t_i תהיה צלע מוסרת אחת לפחות, כלומר כזו עם $\ell = 1$.
שימו לב שהבעיה הזו אינה בשלמים, כלומר ℓ יכול לקבל טווח רחב של ערכים, גם בין 0 ו-1, ולכן היא קרויה בעיית הרב-חתך השברי המינימלי.
כאמור, מדואליות חזקה, הרב-זרימה המקסימלית בגרף שווה לקיבול של הרב-חתך השברי המינימלי. אבל מה לגבי רב חתך בשלמים - כלומר כשאנחנו דורשים או להשאיר או להסיר צלע?
האם הקיבול המינימלי של רב-חתך ברשת זרימה רב מוצרית, זהה לזרימה המקסימלית ברשת?
למרבה הצער - לא!

דוגמה 3.2. נתבונן ברשת הבאה שהיא למעשה עץ (לא מכוון) מעומק 2, ובה זרימה של 6 מוצרים. קיבולי כל הצלעות הן 1:



מפני שכל זרימה משתמשת ב-2 צלעות לפחות (צלע אחת כדי לעזוב את המקור, וצלע נוספת כדי "להיכנס" לבור), נקבל שהזרימה המקסימלית היא 2 (מ-s_1 ל-t_1 נזרים 1 ומ-s_2 ל-t_2).

החתך המינימלי, לעומת זאת, הוא מקיבול כולל של לפחות 3, כי הוא חייב לנתק בין **כל** זוגות המקור-בור וזה דורש בחירה של לפחות 3 צלעות להוריד - בדקו.

אז אם התשובה היא לא - האם כן יש איזשהו הבדל בין קיבול מינימלי של רב-חתך לזרימה מקסימלית, עד כדי פער **בקבוע**?

בעצים למשל, הפער תמיד לא יגדל מעל 2 - נראה את זה בתרגיל הבית. אבל במקרה הכללי, התשובה גם לזה היא לא. נראה את זה בעזרת מושג האקספנדר.

3.1 גרפים מרחיבים

הגדרה 3.3. גרף לא מכוון $G=(V,E)$ נקרא (n,d,α)**-מרחיב** ((n,d,α)-expander) או **אקספנדר** אם:

1. $|V|=n$.
2. $\max_{v\in V}\deg(v)\le d$.
3. לכל קבוצת קודקודים $\emptyset\ne U\subseteq V$, כך ש-$|U|\le\frac{1}{2}|V|$ (לא רוב הקודקודים בקבוצה), מספר השכנים של צמתי U שנמצאים ב-$V\setminus U$ (כלומר מהצד השני), הוא לכל הפחות $\alpha|U|$.

הגדרה 3.4. משפחה של גרפים מרחיבים קבועי דרגה (bounded degree expanders family), היא סדרה אינסופית $G_1,G_2,...$ של גרפים לא מכוונים, שיש עבורה קבועים $\alpha>0$ ו-$d\in\mathbb{N}$, כמו גם סדרה $n_1<n_2<...$ כך שלכל i, G_i הוא גרף (n_i,d,α)-מרחיב.

הרצאה 19

עבור $d=3$, הינה טענה שיחסית קל להוכיח אותה מתמטית:

טענה. קיים $\alpha>0$ קבוע כך שלכל $n\in\mathbb{N}$, גרף רנדומי 3-רגולרי על n קודקודים, הוא $(n,3,\alpha)$-מרחיב, בהסתברות גבוהה.

מסובך גם לבנות משפחה של גרפים מרחיבים מדרגה 3 (אבל אפשר).

טענה. אם G הוא אקספנדר, לכל צומת u יש **לפחות** $\frac{n}{2}$ צמתים (חצי ממספר הצמתים) שמרחקם (הקצר ביותר) מ-u הוא לפחות $\frac{1}{2}\log_d\frac{n}{2}$.

לא נוכיח פורמלית, אבל זה מתבסס על זה של-u יש לכל היותר d שכנים, לכל שכן בעצמו יש לכל היותר d כך שיש לכל היותר d^2 שכנים ממרחק 2, לכל שכן של שכן יש לכל היותר d ולכן יש לכל היותר d^3 ממרחק 3 וכן הלאה.
כך, יש לכל היותר (u הוא שכן של עצמו, ממרחק 0) $1+d+d^2+d^3+...+d^{\frac{1}{2}\log_d\left(\frac{n}{2}\right)-1}$ שכנים ממרחק $\frac{1}{2}\log_d\left(\frac{n}{2}\right)$ לכל היותר. אפשר לעשות סכום סדרה הנדסית ולקבל שזה לכל הפחות $\frac{n}{2}$.

מסקנה - מספר זניח (פחות מחצי מספר הקודקודים) של צמתים הוא במרחק **לכל היותר** $\frac{1}{2}\log_d\frac{n}{2}$ מ-u. כמעט כל הקודקודים הם במרחק $\Omega(\log n)$ מ-u.
מלמעלה אפשר גם לחסום על ידי $\log n$:

טענה 3.5. אם G הוא אקספנדר יש $c>0$ כך שלכל $u,v\in V$ מתקיים $d_G(u,v)\le c\cdot\log n$.

הוכחה. **(בערך, נפנוף ידיים)**
נריץ BFS (חיפוש לרוחב), כשמתחילים בקודקוד כלשהו $v\in V$. נראה שכל הקודקודים הם במרחק $O(\log n)$ ממנו, וזה ע"י מציאת כמה איטרציות (**שכבות**) עושה אלגוריתם BFS, בערך.
כל עוד מספר הקודקודים שביקרנו בהם הוא n' אשר מקיים $n'\le\frac{n}{2}$, נוכל לסמן ב-U את קבוצת כל הקודקודים בהם ביקרנו, מתקיים $|U|=n'\le\frac{n}{2}$ ומתכונת ההרחבה, "בשכבה" הבאה נבקר בלכל הפחות $\alpha|U|$ קודקודים (שהם שכנים של קודקודים ב-U).

נסמן ב-n_i את מספר הקודקודים שראינו עד ה"שכבה" ה-i. אם $n_i \leq \frac{n}{2}$, אז כאמור $n_{i+1} \geq n_i + \alpha n_i = (1+\alpha) n_i$.
אז כמה שכבות יש לנו עד שנגיע ל-$\frac{n}{2}$ קודקודים? - נצטרך $\log_{1+\alpha} \frac{n}{2}$ שכבות לפחות כדי לבקר בלפחות $\frac{n}{2}$ קודקודים.
נסמן ב-f את מספר השכבה הראשונה בה $n_f \geq \frac{n}{2}$ (כלומר $n_{f-1} < \frac{n}{2}$). אנחנו יודעים שקבוצת הקודקודים שלא ביקרנו בהם, כשהגענו לאחרי השכבה ה-f, היא קבוצת קודקודים שמכילה לכל היותר $\frac{n}{2}$ קודקודים. נסמנה U. מתקיים $|U| = n - n_f$.

נסמן ב-$\Gamma(U)$ את השכנים של U. אנחנו יודעים שהקודקודים ב-$\Gamma(U)$ בהם ביקרנו, חייבים להיות שייכים לשכבה ה-$f-1$, שכן אחרת היינו נתקלים בהם בשכבה ה-i כלשהי ורצים עליהם בשכבה ה-$i+1$ (ככה שזה לא אפשרי, כי אז היינו מבקרים בו עוד לפני השכבה ה-f). מתכונת ההרחבה, מספרם הוא לכל הפחות $\alpha |U| = \alpha \cdot (n - n_f)$.
כל קודקוד שאנחנו עוברים עליו בשכבה ה-f, מכיל לכל היותר d שכנים מ-$\Gamma(U)$ (כי הגרף הוא (n,d,α)-מרחיב).
אז, מספר הקודקודים שלא ביקרנו בהם, עם שכן שכן ביקרנו בו, הוא לכל הפחות $\frac{\alpha}{d} \cdot (n - n_f)$.
אבל α, d הם קבועים, ו-$(n - n_f) \leq \frac{n}{2}$. קבוצת הקודקודים שלא ביקרנו בהם מאיטרציה לאיטרציה, קטנה בקצב של $1 - \frac{\alpha}{d}$. כלומר, אחרי ℓ צעדים (מהאיטרציה שבה הגענו ל"רוב" הקודקודים, כלומר מ-f) גודל קבוצת הקודקודים שלא ביקרנו בהם היא לכל היותר $\left(1 - \frac{\alpha}{d}\right)^{\ell} \cdot \frac{n}{2}$ (מתחילים עם לכל היותר $\frac{n}{2}$, וכל פעם קופצים ב-$1 - \frac{\alpha}{d}$).
מהו ℓ עבורו $\left(1 - \frac{\alpha}{d}\right)^{\ell} \cdot \frac{n}{2} \leq 1$ (כלומר יישאר קודקוד אחד)? נפתור את אי השוויון $\left(1 - \frac{\alpha}{d}\right)^{\ell} \leq \frac{2}{n}$, בסיס הלוג קטן מ-1, נהפוך את השבר $\left(\frac{1}{1-\frac{\alpha}{d}}\right)^{\ell} \geq \frac{n}{2}$, ונקבל $\ell \geq \log_{\frac{1}{1-\frac{\alpha}{d}}} \left(\frac{n}{2}\right)$.
עד f עשינו $\log_{1+\alpha} \frac{n}{2}$ איטרציות, ומפה עוד לפחות $\log_{\frac{1}{1-\frac{\alpha}{d}}} \left(\frac{n}{2}\right)$. הלוגים הם קבועים, ולכן אנחנו צריכים $\Omega\left(\log\left(\frac{n}{2}\right)\right)$ צעדים. קיבלנו שכל הקודקודים הם במרחק $O(\log n)$ מ-v, כנדרש. $\square$

בטענה הבאה מנוסחות תכונות של אקספנדרים:

טענה 3.6. יהא G אקספנדר. אזי:

- $|E| = O(|V|)$.
- $\operatorname{diam}(G) = O(\log n)$ ($\operatorname{diam}(G)$ הוא המרחק המקסימלי בין שני קודקודים ב-G).
- לכל u, קיימים לפחות $\frac{n}{2}$ קודקודים במרחק $\Omega(\log n)$ מ-u.

נחזור לזרימה רב מוצרית ונראה איך אקספנדרים מביאים לנו דוגמה רעה להבדל (לא קבוע) של קיבול מינימלי של רב חתך וזרימת מקסימום.

טענה 3.7. נניח כי $G = (V,E)$ הוא (n,d,α)-אקספנדר, כל הקיבולים על הצלעות הם 1, ו-(s_i, t_i) הם כל זוגות הקודקודים במרחק לכל הפחות $\lambda \log(n)$ לאיזשהו λ קבוע (שהוא ה-c מטענה 3.5) (יש לכל הפחות $\left(\frac{1}{2}n\right)^2 = \frac{1}{4}n^2$ כאלה).
אזי, גודל קיבול הרב-חתך המינימלי, הוא $\Omega(\log n)$ כפול גודל זרימה מקסימלית רב מוצרית.

הוכחה. כדי להראות את זה, נחסום מלמעלה את הזרימה המקסימלית, ומלמטה את גודל הרב-חתך. ונראה שהפער חלוקה בין החסמים האלה הוא $\Omega(\log n)$.
כדי לחסום מלמעלה את הזרימה המקסימלית נעשה את זה על ידי מציאת פתרון פיזיבילי **לתוכנית הדואלית** (ושימוש בדואליות חלשה). זכרו כי התוכנית הדואלית לאחת מהתצורות של תוכנית לינארית לזרימה רב מוצרית, היא תוכנית אורכים של הצלעות. בפתרון שנביא ניתן אורך $\frac{1}{\lambda \log n}$ לכל צלע. וזה אכן פתרון פיזיבילי (שכן האילוץ היחיד שהיה לנו בתוכנית ההיא, הוא שאורך כל מסלול $s_i \to t_i$ יהיה לפחות 1. בכל מסלול יש לנו לפחות $\lambda \log(n)$ קשתות, מההגדרה של הזוגות (s_i, t_i), מהגדרה, וכך אורך כל מסלול הוא לפחות 1).
הערך בפונקציית המטרה של הפתרון הזה, הוא $\sum_{e \in E} \ell(e) c(e) = \sum_{e \in E} \ell(e) = \frac{|E|}{\lambda \log n} = O\left(\frac{n}{\log n}\right)$ שכן $|E| = O(|V|)$ (מהטענה הקודמת).
אז זה פתרון פיזיבילי דואלי, ולכן מדואליות חלשה הזרימה המקסימלית חסומה מלמעלה על ידי $O\left(\frac{n}{\log(n)}\right)$.
אז הזרימה המקסימלית, היא לכל היותר $O\left(\frac{n}{\log n}\right)$.

נחסום מלמטה גודל חתך מינימלי. עלינו להראות שגודלו הוא לכל הפחות קבוע כפול n (כי רצינו מנה של $(O\left(\log\left(n\right)\right)$).
נבחר רגע $\varepsilon>0$ שנקבע ערכו בהמשך. נניח שנוריד מ-G לכל היותר εn-צלעות, אנחנו נראה שקיבלנו זוג מקור ובור **שעדיין מחוברים** וכך נגיע למסקנה שקבוצת הצלעות שהורדנו איננה חתך. כלומר, גודל חתך מינימום צריך להיות גדול יותר מ-$\varepsilon\cdot n$.
לכל צומת יש לכל הפחות $\frac{n}{2}$ צמתים שרחוקים ממנו מספיק, שהם בורות עבורו. ולכן קבוצה של קודקודים בגודל מעל $\frac{n}{2}$, תכיל לפחות זוג אחד מקור ובור.
נראה שאם מורידים מ-G לכל היותר $\varepsilon\cdot n$ צלעות - לא ייתכן <u>שכל</u> רכיבי הקשירות מגודל קטן מ-$\frac{n}{2}$. כלומר, יש רכיב קשירות מגודל $\frac{n}{2}$ לפחות, וכמו שאמרנו נובע שיש זוג מקור ובור שמחוברים - ואז, הקבוצה הזו של $\varepsilon\cdot n$ לא תהיה חתך מינימום (אפילו לא חתך).
אז נניח בשלילה **שאין** בגרף הנותר (לאחר הסרת הצלעות) רכיב קשירות מגודל $\frac{n}{2}$ לפחות.

- **במקרה א'**, בגרף הנותר לאחר ההסרה יש רכיב קשיר מגודל $\frac{n}{3}$ לפחות. אבל אז, מתכונת ההרחבה של האקספנדר (אנחנו עם קבוצה מגודל לכל היותר $\frac{n}{2}$), יש לרכיב הזה ב-G לפחות $\frac{\alpha}{3}n$ שכנים **שאינם ברכיב**. כלומר כל הקשתות שהובילו לשכנים האלה הוסרו, אז הסרנו לפחות $\frac{\alpha}{3}n$ קשתות. ואם $\varepsilon<\frac{\alpha}{3}$ זו תהיה סתירה (אנחנו נדאג אחר כך שאכן זה יקרה).

הרצאה 20

- **במקרה ב'**, כל הרכיבים הקשירים הם מגודל קטן מ-$\frac{n}{3}$. מה שנעשה זה "לצבור" רכיבי קשירות של קודקודים, עד שנגיע לקבוצת קודקודים S שגודלה בין $\frac{n}{3}$ ל-$\frac{2n}{3}$ (אם באיטרציה כלשהי אנחנו עם קבוצת קודקודים מגודל פחות מ-$\frac{n}{3}$, ואנחנו מוסיפים עוד קבוצת קודקודים מגודל $\frac{n}{3}$ - נסיים עם פחות מ-$\frac{2n}{3}$).
נתבונן בקבוצות S ו-$\overline{S}=V\setminus S$. אחת מהן קטנה יותר, בפרט גודלה לכל היותר $\frac{n}{2}$. כעת יש לנו קבוצה מגודל לכל היותר $\frac{n}{2}$, אפשר להפעיל עליה את תכונת ההרחבה של האקספנדר. מהטיעון של מקרה א', שהסרנו בסך הכל $\frac{\alpha}{3}n$ קשתות לכל הפחות וזו תהיה סתירה.

על כן נבחר $\varepsilon<\frac{\alpha}{3}$. ואז אחרי הסרת לכל היותר εn קשתות יוותר רכיב מגודל לכל היותר $\frac{n}{2}$. ברכיב זה יהיה זוג של קודקוד מקור וקודקוד בור, נקבל ש-G לא חתך.
על כן, גודל חתך צריך להיות לכל הפחות $\varepsilon\cdot n$, כלומר $\Omega\left(n\right)$.
כפי שהסברנו קודם זה מוכיח שגודל קיבול הרב-חתך המינימלי, הוא $\Omega\left(\log n\right)$ כפול גודל זרימה מקסימלית רב מוצרית.
□

3.2 אלגוריתם $O\left(\ln k\right)$-מקרב לרב חתך מינימלי

אם כן, יש פער בין הרב-חתך המינימלי לבין הזרימה המקסימלית. מה הפער הכי גרוע? להלן חסם מלמעלה על ההבדל:

משפט 3.8. לכל רשת זרימה רב מוצרית <u>לא מכוונת</u> שבה k מוצרים, הפער בין קיבול חתך מינימום לבן ערך זרימת מקסימום הוא לכל היותר פקטור של $O\left(\log k\right)$ (כלומר לכל היותר $c\cdot\log k$, עבור c קבוע אוניברסלי כלשהו).

ההוכחה של המשפט הולכת להיות אלגוריתם. האלגוריתם המקרב ייתן רב-חתך עם קיבול C וזרימה עם ערך F, ככה ש-$C\leq O\left(\log k\right)\cdot F$. וכך אם C^* קיבול רב-חתך מינימלי ($C^*\leq C$), F^* זרימה מקסימלית ($F\leq F^*$), נקבל $C^*\leq C\leq O\left(\log k\right)\cdot F\leq O\left(\log k\right)\cdot F^*$, כלומר הרב-חתך המינימלי הוא לכל היותר $O\left(\log k\right)$ הזרימה המקסימלית.

רעיון הוכחת המשפט יהיה להתבונן בתוכנית הדואלית לתוכנית המבטאת זרימת מקסימום, כלומר הרב חתך השברי.
נסמן את הפתרון האופטימלי שלה ב-w^*. כמובן, מדואליות חזקה, זו גם זרימת המקסימום ברשת.
כלומר $w^*=\underset{e\in E}{\mathrm{argmin}}\,c\left(e\right)\ell\left(e\right)$, יחד עם האילוצים $\sum_{e\in P}\ell\left(e\right)\geq 1$ לכל $P\in P_i$ (P_i זו קבוצת המסלולים מ-s_i ל-t_i) ו-$\ell\geq 0$.
הפתרון האופטימלי w^* משרה פונקציית מרחק חלקית (**פסאודו-מטריקה**) d_{ℓ^*} על הצמתים: $d_{\ell^*}\left(u,v\right)$ הוא משקל המסלול הקל ביותר בין u ו-v.

אנחנו יודעים להגיד כי $d_{\ell^*}(s_i,t_i)\geq 1$ בגלל האילוץ. כמו כן, משום שאפשר לקחת חתך מינימלי ולהגדיר לו פתרון פיזיבילי לבעיית האורכים הנ"ל (צלע בחתך מקבלת 1, צלע שלא מקבל אורך 0. ערך פונקציית המטרה יוצאת קיבול החתך. מהגדרת חתך, בכל מסלול מ-s_i ל-t_i תהיה צלע עם אורך 1), ומשום ש-w^* הוא פתרון אופטימלי לבעייה הזו, הוא המינימלי, אז $\sum\limits_{e\in E} c(e)\,d_{\ell^*}(e)$ הוא חסם תחתון על קיבול של חתך מינימלי.

אז, ניקח את d_{ℓ^*} ונבנה בידיים רב-חתך F שהוא לא יותר הרבה יקר ממנו: לכל היותר $O(\log k)$ מהערך הדואלי האופטימלי w^*. כלומר $\sum\limits_{e\in F} f(e)\leq O(\log k)\cdot w^*$.

תיאור האלגוריתם במילים

תחילה נסתכל על s_1 ו-t_1. אנחנו יודעים שהמרחק $\ell^*(s_1,t_1)$ הוא לכל הפחות 1. נסתכל על כדור מרדיוס ρ סביב s_1, אותו נגדיר להיות $B(s,\rho)=\{v\mid d_{\ell^*}(s_1,v)\leq\rho\}$ - כל הקודקודים במרחק לכל היותר ρ מ-s_1. ונניח גם כי $\rho<\frac{1}{2}$ (הוא ייבחר באלגוריתם).

אז אין ב-$B(s,\rho)$ זוג כלשהו של מקור ובור. נניח בשלילה שיש $\{s_i,t_i\}\subseteq B(s,\rho)$. אבל d_{ℓ^*}, מטריקת המסלול הקצר ביותר, מקיימת את אי שוויון המשולש.

נקבל $1<d_{\ell^*}(s_i,t_i)\leq d_{\ell^*}(s_i,s_1)+d_{\ell^*}(t_i,s_1)<\rho+\rho=2\rho<1$ - סתירה.

אז אין זוג מקור ובור שנמצא ב-$B(s,\rho)$. ניתקנו את כולם. מקרה נוסף הוא מקרה של מקור ובור שנמצאים אחד בכדור ואחד לא.

כדי לטפל בו, נוריד (הכוונה היא, נוסיף לחתך F) גם את כל הקשתות $\{u,v\}$ אשר $u\in B(s_1,\rho)$ ו-$v\notin B(s_1,\rho)$.

שימו לב שבתוך הכדור יכולות להיות צלעות.

עשינו את כל תהליך הכדור עבור הזוג (s_1,t_1). נמשיך בתהליך עם הגרף שנשאר, על כל k המוצרים, וככה יובטח לנו שכל הזוגות (s_i,t_i) לא יהיו מחוברים.

כדי לחסום את הקיבול של הרב-חתך הזה, נשתמש בזה שברגע שאנחנו טיפלנו בכדור כלשהו, אפשר לשכוח מכל מי שנמצא בתוכו. בסכום $\sum\limits_{e\in E} c(e)\,d_{\ell^*}(e)$, יש צלעות שהסרנו, אפשר לשכוח מהן בסכום הזה, אבל לזכור **שהצלעות האלה מתווספות לחתך** ולכן הן יתווספו לערך הקיבול של החתך.

אנחנו נראה את הדבר הבא: סכום קיבולי הצלעות שהסרנו מהכדור (אלה שמחברים בין קודקודים בכדור לכאלה שלא בכדור), לא גדול במידה מה מסכום האורכים של הצלעות שבתוך הכדור. לאחר מכן, אפשר לחסום מלמעלה את הקיבול של כל החתך F על ידי סכום (על כל הכדורים) של כל סכומי אורכי הצלעות שבתוך כל כדור.

נוכל גם "לשכוח" מהקיום של הקודקודים בתוך הכדור (הם לא מעניינים אותנו).

האלגוריתם הוא פרי יצירתם של Garg, Vazirani ו-Yannakis מ-1993.

אלגוריתם 3 $O(\ln k)$-מקרב לרב-חתך מינימלי ברשת זרימה רב מוצרית

1. אתחל $F\leftarrow\emptyset$.
2. כל עוד יש זוג מקור-בור מקושר (s_i,t_i):
   (א) בחר $\rho\in\left(0,\frac{1}{2}\right)$.
   (ב) הוסף ל-F את כל הצלעות $\{u,v\}$ בגרף כך ש-$d_{\ell^*}(s_i,u)\leq\rho$ ו-$d_{\ell^*}(s_i,v)>\rho$.
   (ג) הסר את כל הקודקודים ב-$B(s_i,\rho)$ מהגרף.

הוכחת משפט 3.8 - אלגוריתם הקירוב. לצורך ההוכחה, בהינתן כדור $B(s_i,\rho)$, נגדיר מהו הנפח שלו.

לכל $u,v\in V$, נגדיר $\ell'(u,v)=|d_{\ell^*}(s_i,u)-d_{\ell^*}(s_i,v)|$. מאי שוויון המשולש $\ell'(u,v)\leq\ell^*(u,v)$.

עבור צלעות $\{u,v\}$ כך ש-$u\in B(s_i,\rho)$ אבל $v\notin B(s_i,\rho)$, נרצה להגדיר את "השבר של הצלע שבתוך הכדור". נגדיר את זה בתור $\frac{\rho-d_{\ell^*}(s_i,u)}{\ell'(u,v)}$. **אורך הצלע** בתוך הכדור הוא $\rho-d_{\ell^*}(s_i,u)$.

כך נגדיר את **הנפח של הכדור** $B=B\left(s_i,\rho\right)$ להיות:

$$\mathrm{vol}\left(B\right)=\left(\underbrace{\sum_{u,v\in B}c\left(u,v\right)\ell'\left(u,v\right)+\sum_{\substack{e=\{u,v\}\\ u\in B,\ v\notin B}}\frac{\rho-d_{\ell^*}\left(s_i,u\right)}{\ell'\left(u,v\right)}c\left(u,v\right)\ell'\left(u,v\right)}_{(*)}+\frac{w^*}{k}\right)$$

נגדיר את הפונקציה $V\left(\rho\right)=\mathrm{vol}\left(B\left(s_i,\rho\right)\right)$ (לאיזשהו s_i).

נסתכל על הנגזרת, $\frac{\mathbf{d}V(\rho)}{\mathbf{d}\rho}$ (ב-$\rho\in\left(0,\frac{1}{2}\right)$). אולי היא לא רציפה או לא גזירה בכמה נקודות (מספר סופי של נקודות). אבל באופן כללי, כשמסתכלים על הביטוי הנ"ל, רואים שהנגזרת לפי ρ היא סכום כל הקיבולים של הצלעות שמתווספות ל-F באותו הזמן, כלומר אלה שנחתכות ע"י $B\left(s_i,\rho\right)$.

זכרו שגודל קיבול החתך בכל פעם גדל בסכום הזה (הנגזרת), וכפי שהסברנו בהתחלה - נראה שהוא לא גדול מאוד בפער מנפח הכדור. אז נרצה ρ ככה ש-$\frac{\frac{\mathbf{d}V(\rho)}{\mathbf{d}\rho}}{\mathrm{vol}(B(s_i),\rho)}$, כלומר $\frac{\frac{\mathbf{d}V(\rho)}{\mathbf{d}\rho}}{V(\rho)}$, הוא קטן.

נבחר ρ באופן אקראי יוניפורמי מ-$\left(0,\frac{1}{2}\right)$ (האמת שלא חייב, כי הכדור סופי ולכן יש רק מספר סופי של ρים שבאמת משנים לנו. אפשר לבחור את המינימלי. אבל זה כמובן גורם לאלגוריתם לעבוד בזמן פחות טוב). התוחלת של $\frac{\frac{\mathbf{d}V(\rho)}{\mathbf{d}\rho}}{V(\rho)}$, היא (כי ה-pdf של ההסתברות האחידה בקטע הוא 2):

$$\begin{aligned}\mathbb{E}\left[\frac{\frac{\mathbf{d}V(\rho)}{\mathbf{d}\rho}}{V(\rho)}\right] &= 2\int_0^{1/2}\frac{\frac{\mathbf{d}V(\rho)}{\mathbf{d}\rho}}{V(\rho)}\mathbf{d}\rho=\\ &= 2\left(\ln\left(V\left(\rho\right)\right)\right)\Big|_0^{1/2}=\\ &= 2\ln\left(V\left(\tfrac{1}{2}\right)\right)-2\ln\left(V\left(0\right)\right)\end{aligned}$$

מתקיים $V\left(0\right)=\mathrm{vol}\left(B\left(s_i,0\right)\right)=\frac{w^*}{k}$. אנחנו לא יודעים מהו $V\left(\frac{1}{2}\right)$ אבל הוא לכל היותר w^*.

קיבלנו כי $\mathbb{E}\left[\frac{\frac{\mathbf{d}V(\rho)}{\mathbf{d}\rho}}{V(\rho)}\right]\leq 2\left(\ln\left(w^*\right)-\ln\left(\frac{w^*}{k}\right)\right)=2\ln\left(k\right)$.

בפרט בגלל שאנחנו רציפים באמת יש ρ ככה שהיחס $\frac{\frac{\mathbf{d}V(\rho)}{\mathbf{d}\rho}}{V(\rho)}$ יוצא **קטן מ-**$2\ln\left(k\right)$ (התוחלת היא הממוצע, אם לא היה - כל הערכים היו מעל הממוצע - זה לא ייתכן!).

כלומר, יש ρ אשר $\sum_{\substack{e=\{u,v\}\\ u\in B,\ v\notin B}}c\left(u,v\right)\leq 2\ln k\cdot\mathrm{vol}\left(B\left(s_i,\rho\right)\right)$. בכל איטרציה i נבחר ρ_i כזה.

נניח שהאלגוריתם משתמש בכדורים $B\left(s_1,\rho_1\right),...,B\left(s_m,\rho_m\right)$ (כלומר עוצר ב-m). מתקיים $m\leq k$.

נקבל כי $\sum_{e\in F}c\left(e\right)\leq 2\ln\left(k\right)\cdot\sum_{i=1}^{m}\mathrm{vol}\left(B\left(s_i,\rho_i\right)\right)$. כל הכדורים לא נחתכים זה עם זה (כי אנחנו מורידים את כל הקודקודים בכל כדור מהגרף). כלומר, כל הכדורים הם זרים ביחס התרומה שלהם לנפח (אין חיתוך בין קשתות וחלקי הקשתות שמשמשות לחישוב הנפח בין הכדורים). לכן כשסוכמים הכל, הביטוי $(*)$ (למעלה) קטן מהסכום של אורכי כל הצלעות בגרף, וזה יוצא w^*.

קיבלנו $\sum_{i=1}^{m}\mathrm{vol}\left(B\left(s_i,\rho_i\right)\right)\leq w^*+m\cdot\frac{w^*}{k}$ (סוכמים m פעמים את $\frac{w^*}{k}$). אבל $m\leq k$ לכן $\sum_{i=1}^{m}\mathrm{vol}\left(B\left(s_i,\rho_i\right)\right)\leq 2w^*$.

קיבלנו כי $\sum_{e\in F}c\left(e\right)\leq 4\ln k\cdot w^*$ - וזה מסיים לנו את ההוכחה. $\square$

4 קירוב לזרימה בו-זמנית מקסימלית והחתך הדליל ביותר

הרצאה 21

ניזכר בהגדרה של בעיית הזרימה הבו זמנית המקסימלית (maximum concurrent flow):

$$\begin{aligned} & u^* = \max \quad u \\ & \text{subject to:} \\ & \forall 1 \le i \le k \qquad u - \sum_{P \in P_i} f_P^{(i)} \le 0 \\ & \forall e \in E \qquad \sum_{i=1}^{k} \sum_{P \in P_i,\ e \in P} f_P^{(i)} \le c(e) \\ & \boldsymbol{f} \ge \boldsymbol{0} \end{aligned}$$

נגדיר גם את בעיית עומס הזרימה המינימלי (maximum congestion flow). בבעיית עומס הזרימה מנסים למזער את העומס היחסי של כל קשת, כשנתונה זרימה של 1 לכל מוצר. כלומר:

$$\begin{aligned} & z^* = \min \ z \\ & \text{subject to:} \\ & \forall 1 \le i \le k \qquad \sum_{P \in P_i} f_P^{(i)} \le 1 \\ & \forall e \in E \qquad \sum_{i=1}^{k} \sum_{P \in P_i,\ e \in P} f_P^{(i)} \le z \cdot c(e) \\ & \boldsymbol{f} \ge \boldsymbol{0} \end{aligned}$$

למעשה אלה בעיות שקולות, כי אם u^* הוא פתרון אופטימלי לראשונה, אז $\frac{1}{u^*}$ לשניה. כלומר $z^* = \frac{1}{u^*}$. כדי לעבור בין פתרונות, אפשר לחלק פתרון (כלומר לחלק כל זרימה) של הראשונה ב-u, או להכפיל פתרון של השניה ב-z.

מה התוכנית הדואלית לתוכנית הראשונה? זו תהיה בעיית מינימיזציה **על הצלעות** שנסמן את משתניה ב-$\ell(e)$ (משתנה לכל צלע) וב-h_i (משתנה לכל מוצר). המקדמים של $c(e)$ בפונקציית המטרה יצאו בדיוק $\ell(e)$. המקדמים של h_i למעשה יהיו אפסים. התוכנית תיראה כך:

$$\begin{aligned} & \min \sum_{e \in E} c(e)\,\ell(e) \\ & \text{subject to:} \\ & \forall 1 \le i \le k \qquad \sum_{i=1}^{k} h_i \ge 1 \\ & \forall i,\ P \in P_i \qquad \sum_{e \in P} \ell(e) \ge h_i \\ & \boldsymbol{\ell}, \boldsymbol{h} \ge \boldsymbol{0} \end{aligned}$$

בשיעור שעבר ראינו ש**בעיית הזרימה המקסימלית** הייתה אנלוגית **(לא דואלית)** - במובן שיש איזשהו פער מקסימלי בין חתך מינימלי (בשלמים) לזרימה מקסימלית. הפתרון האופטימלי היה יכול להיות בפער של $O(\log k)$.

האם אפשר למצוא בעייה אנלוגית (כלומר להבטיח איזשהו פער בין אופטימומים של 2 בעיות) לבעיית הזרימה הבו-זמנית המקסימלית?

4.1 החתך הדליל ביותר

הגדרה 4.1. יהא $(S, \overline{S})$ חתך ($S \subseteq V$, כאשר $S \ne \emptyset$ ו-$S \ne V$). נניח כי קיים i ככה ש-$s_i \in S$ ו-$t_i \in \overline{S}$ (או ההפך). במצב זה נגיד שהחתך $(S, \overline{S})$ "מפריד לכל הפחות מוצר אחד".

נגדיר את **הדלילות של החתך** $(S, \overline{S})$ (sparse) להיות $\alpha(S, \overline{S}) = \frac{\sum_{e \in E,\ |e \cap S| = 1} c(e)}{|\{i \mid \{s_i, t_i\} \cap S = 1\}|}$.

הערה. שימו לב שאין שום קשר ל**רב חתך** (multicut) מהפרק הקודם. מדובר על חתך רגיל (S-T cut), כזה שמובטח שמפריד לפחות מקור ובור אחד (אולי במקרה גם מקורות-בורות של מוצרים אחרים).

המכנה בביטוי אומר לנו כמה זוגות של מקור-בור נמצאים בתוך החתך. המונה אומר מה הקיבול הכולל של כל מי שעובר דרך החתך.

ולכן הדלילות היא מושג מעניין: זה היחס בין הקיבול הכולל למספר הזוגות מקור-בור ש"נכלאו" בין החתך, והוא אומר לנו מה הקיבול "הממוצע" של כל זוג מקור-בור שעובר דרך החתך.

הגדרה 4.2. החתך הדליל ביותר (sparsest cut) הוא החתך $\left(S,\overline{S}\right)$ שממזער את $\alpha\left(S,\overline{S}\right)$.

נסמן את ערך הדלילות של החתך הדליל ביותר ב-α^*. אנחנו יודעים ש-u^* (הזרימה בו זמנית המקסימלית) מקיים $u^*\leq\alpha^*$.

מה הפער בין הפתרונות u^* ו-α^*?

בפרק זה של הקורס, נראה שמדובר גם פה על פער של $O\left(\log k\right)$. נבצע זאת בעזרת גישה אחרת לחלוטין (לא על ידי אלגוריתם, כמו ברב חתך) והיא שימוש בשיכונים מטריים.

ראשית נראה את הכיוון ההפוך - **לכל חתך** אפשר למצוא פתרון דואלי ששווה לדלילות החתך. זהו הכיוון ההפוך שכן אנו עומדים לראות שמבין כל החתכים, החתך הדליל ביותר לא רק שנותן איזושהי זרימה בו זמנית u (השווה בדלילותו לחתך הנ"ל), אלא גם קטן פי לכל היותר $O\left(\log n\right)$ מהזרימה הבו זמנית המקסימלית.

טענה 4.3. לכל חתך $\left(S,\overline{S}\right)$ שמפריד לפחות מוצר אחד, יש פתרון דואלי פיזיבילי $\boldsymbol{h},\boldsymbol{\ell}$ ככה ש-$\alpha\left(S,\overline{S}\right)=\sum\limits_{e\in E}c\left(e\right)\ell\left(e\right)$.

הוכחה. הוכחה קונסטרוקטיבית: נבנה פתרון דואלי מתאים.

נניח כי $\left(S,\overline{S}\right)$ חתך שמפריד את המוצרים $i_1,...,i_s$. נגדיר את $h_i=\begin{cases}\frac{1}{s} & j=1,...,s\\ 0 & \text{אחרת}\end{cases}$, ואת $\ell\left(e\right)$ להיות כזו שאורך הצלע הוא 1 אם הצלע חוצה את החתך, ואחרת 0. כלומר $\ell\left(e\right)=\begin{cases}\frac{1}{s} & \left|e\cap S\right|=1\\ 0 & \text{אחרת}\end{cases}$.

זו אכן זרימה פיזיבילית, כי $\boldsymbol{\ell},\boldsymbol{h}$ אי שליליות ומתקיימים האילוצים $\sum\limits_{i=1}^{k}h_i=1\geq$ (שכן פרט ל-s אילוצים פה, כולם 0 והשאר $\frac{1}{s}$), ובמקרה בו $h_i=0$ מתקיים $\sum\limits_{e\in P}\ell\left(e\right)=0\geq h_i$ (אם $h_i=0$ אז המוצר ה-i לא נחצה דרך החתך, אז אף צלע שמחברת בין המוצרים לא עוברת דרך החתך), ואם $h_i=\frac{1}{s}$ בהינתן $P\in P_i$ מתקיים $\sum\limits_{e\in P}\ell\left(e\right)\geq\frac{1}{s}\geq h_i$ (לכל הפחות צלע אחת עוברת דרך החתך, וערך ה-ℓ שלה יהיה $\frac{1}{s}$).

בנוסף $\frac{\sum\limits_{e\in E,\ \left|e\cap S\right|=1}c(e)}{s}=\sum\limits_{e\in E,\ \left|e\cap S\right|=1}c\left(e\right)\cdot\frac{1}{s}=\sum\limits_{e\in E,\ \left|e\cap S\right|=1}c\left(e\right)\ell\left(e\right)=\sum\limits_{e\in E}c\left(e\right)\ell\left(e\right)$.

אבל s הוא מספר המוצרים שעוברים דרך החתך, קיבלנו כי $\alpha\left(S,\overline{S}\right)=\frac{\sum\limits_{e\in E,\ \left|e\cap S\right|=1}c(e)}{\left|\left\{i\ \mid\ \left\{s_i,t_i\right\}\cap S=1\right\}\right|}=\sum\limits_{e\in E}c\left(e\right)\ell\left(e\right)$ כמו שרצינו. $\square$

בפרט, גם החתך הדליל ביותר הוא פתרון פיזיבילי לתוכנית הדואלית לתוכנית ה-concurrent flow (כלומר לתוכנית האורכים). אם כי הוא, בצער רב, לא הפתרון האופטימלי: **מציאת החתך הדליל ביותר הוא בעיה NP-קשה.**

ועל כן, ננסה להבין מה הפער בין הדלילות המינימלית של חתך ובין הערך המקסימלי של הזרימה הבו-זמנית ברשת רב מוצרית (או לחילופין, כי זה אותו דבר מדואליות חזקה, הבעיה לתוכנית הדואלית של max concurrent flow).

4.2 שיכונים מטריים

הרצאה 22

נזכור שהמטרה שלנו היא לקשר בין הערך האופטימלי של התוכנית הדואלית (שזה גם הערך האופטימלי לתוכנית ה-maximum concurrent flow), שהוא $w^*=\sum\limits_{e\in E}c\left(e\right)\ell^*\left(e\right)$ (כאשר $\left(\boldsymbol{\ell}^*,\boldsymbol{h}^*\right)$ פתרון אופטימלי לתוכנית הדואלית), לבין

הדלילות המינימלית של חתך α^*. כלומר, למצוא ρ שמקיים $\rho \cdot w^* \le \alpha^*$.

הגדרה 4.4. מרחב נקודות $X \neq \emptyset$ יחד עם פונקציית מרחק $d : X \times X \to \mathbb{R}_{\geq 0}$ נקראת **מרחב מטרי** (metric space), אם הפונקציה d, שתיקרא **מטריקה** (metric), מקיימת:

- **רפלקסיביות:** $d\left(x,y\right)=0$ אם ורק אם $x=y$.
- **סימטריה:** $d\left(x,y\right)=d\left(y,x\right)$ לכל $x,y\in X$.
- **אי שוויון המשולש:** $d\left(x,y\right)\le d\left(x,z\right)+d\left(z,y\right)$ לכל $x,y,z\in X$.

במקרה בו d מקיימת את כל השניים האחרונים, ו-$d\left(x,x\right)=0$ לכל $x\in X$ (כלומר ייתכנו $x\neq y\in X$ אשר $d\left(x,y\right)=0$), היא תיקרא **פסאודו מטריקה** (semi-metric או pseudo-metric) והמרחב ייקרא **מרחב פסאודו-מטרי**.

הערה 4.5. למעשה משפט מתמטי אומר שמכל מרחב פסאודו-מטרי X ניתן לבנות מרחב מטרי X', דרך מחלקות שקילות: היחס R שמוגדר על ידי $R=\{x,y \mid d\left(x.y\right)=0\}$ הוא יחס שקילות, יש לו מחלקות שקילות: המרחב המטרי החדש שנבנה הוא $X'=\{[x]_R \mid x\in X\}$, כאשר $d'\left([x]_R,[y]_R\right)=d\left(x,y\right)$ - אפשר להוכיח שהפונקציה d' מוגדרת היטב (לא תלויה בבחירת הנציגים של מחלקות השקילות), ואכן מהווה מטריקה על X'.

טענה. יהא V מרחב וקטורי ו-$||\cdot||$ נורמה על V. נגדיר את הפונקציה $d\left(x,y\right)=||x-y||$ שתיקרא "המטריקה המושרית מהנורמה $||\cdot||$" (induced metric). אזי $\left(V,d\right)$ מרחב מטרי.

בנוסף, בדוגמה לעיל, $\left(V,||\cdot||\right)$ ייקרא **מרחב נורמי (norm space)**. הטענה האחרונה אומרת שמרחבים נורמיים הם מקרים פרטיים של מרחבים מטריים, במובן שכל נורמה משרה מטריקה.

נשים לב ש-ℓ^* משרה פסאודו-מטריקה על V, כך שהמרחק בין שני קודקודים u,v מוגדר להיות אורך המסלול הקצר ביותר (בהשתמש במשקלים ℓ) ביניהם. אם נניח שאורך כל צלע הוא גדול ממש מאפס, למעשה ℓ^* תהיה ממש מטריקה, לא נניח את זה כרגע.

אנחנו יודעים שמאחד האילוצים, מתקיים $\ell^*\left(s_i,t_i\right)\ge h_i$ לכל $1\le i\le k$ ו-$\ell^*\left(e\right)\ge\ell^*\left(u,v\right)$ לכל $e=\{u,v\}\in E$.

מעכשיו נתייחס לבעייה הגרפית בתור בעיית מרחבים מטריים. יהא $\left(V,\ell^*\right)$ מרחב פסאודו-מטרי.

נשכן את $\left(V,\ell^*\right)$ אל $\left(\mathbb{R}^d,||\cdot||_2\right)$ (עבור d שנבחר אחר כך), כלומר נגדיר $\varphi:V\to\mathbb{R}^d$, שמקיימת שאם $\ell^*\left(u,v\right)=0$ אז $\varphi\left(u\right)=\varphi\left(v\right)$ (כלומר חח"ע על מחלקות שקילות).

הגדרה 4.6. נגדיר את **נורמת ליפשיץ** על איברי מרחב השיכונים מ-X ל-Y (כלומר על פונקציות), על ידי $||\varphi||_{\text{Lip}}=\max\limits_{\substack{u,v\in V\\ \text{s.t. }\ell^*(u,v)>0}}\frac{||\boldsymbol{\varphi(u)-\varphi(v)}||}{\ell^*(u,v)}$ כאשר φ היא שיכון $X\to Y$. הרבה פעמים היא נקראת **הנורמה האופרטורית**.

נורמת הליפשיץ של φ מודדת לנו "את ההארכה המקסימלית שמבצעת העתקה φ".

משום ש-φ היא חד חד ערכית על מחלקות שקילות, כפי שרצינו מקודם, אפשר לדבר על פונקציה הופכית $\varphi^{-1}:\varphi\left(V\right)\to V$ (לוקחת נקודה מהתמונה, ומעבירה אותה לאיזושהי נציג של מחלקת השקילות המתאימה לכל אלה שעוברים לנקודה). נשים לב שע"פ הגדרה $\left|\left|\varphi^{-1}\right|\right|_{\text{Lip}}\max\limits_{\boldsymbol{x}\neq\boldsymbol{y}\in\varphi(V)}\frac{\ell^*\left(\varphi^{-1}(\boldsymbol{x}),\varphi^{-1}(\boldsymbol{y})\right)}{||\boldsymbol{x}-\boldsymbol{y}||}$. זה למעשה גודל שמודד לנו את **הכיווץ המקסימלי** של φ (לא של φ^{-1}!).

הגדרה 4.7. יהא $\varphi : X \to Y$ שיכון. **העיוות (bi-lipschitz distortion)** של φ מוגדר להיות $\operatorname{dist}(f) = ||\varphi||_{\text{Lip}} \cdot \left|\left|\varphi^{-1}\right|\right|_{\text{Lip}}$.

הגדרה 4.8. יהא $\varphi : X \to Y$ שיכון. הוא ייקרא **איזומטרי (isometric embedding)** אם $d_Y(f(x), f(y)) = d_X(x,y)$ לכל $x, y \in X$.

עיוות של שיכון הוא מדד לא רע ל"עד כמה השיכון משמר **יחסים** בין מרחקים?".

מסקנה (ישיר מההגדרה)**.** עיוות של שיכון $f : X \to Y$ הוא 1 אם ורק אם השיכון משמר את כל המרחקים עד כדי אותו קבוע (כלומר המרחקים ב-Y הם **כולם בדיוק** כפולה בסקלר של המרחקים ב-X). כלומר, אם ורק אם קיים $c \in \mathbb{R}_{\geq 0}$ אשר לכל $x \neq y \in X$ מקיים $\frac{d_Y(f(x),f(y))}{d_X(x,y)} = c$ ו-$\frac{1}{c} = \frac{d_X(x,y)}{d_Y(f(x),f(y))}$.

נסמן ב-ℓ_p^d את המרחב המטרי $\left(\mathbb{R}^d, ||\cdot||_p\right)$ (נורמת p, עבור $1 \leq p \leq \infty$). אם לא נרצה לציין במפורש את המימד, למשל אם הוא לא חשוב, נסמן רק ℓ_p.

משפט 4.9. נניח כי V סופית. אז קיים שיכון איזומטרי $f : X \to \ell_\infty$.

המשפט הנ"ל בתורת השיכונים נקרא "משפט Frechet" (פְּרֵשֶׁה) או "אוניברסליות ℓ_∞".

הוכחה. נסמן $V = \{v_1, ..., v_n\}$. המימד שנבחר יהיה n (למעשה אפשר גם ל-$n-1$). נגדיר את השיכון φ להיות:

$$\boldsymbol{\varphi(v_j)} = \left(\ell^*(v_0, v_j), \ell^*(v_1, v_j), ...\ell^*(v_n, v_j)\right)$$

וכעת עלינו להוכיח ש-φ היא אכן איזומטריה.

- יהיו $v_r, v_s \in V$ שני איברים כלליים מ-V.
  ע"פ הגדרה, $||\boldsymbol{\varphi(v_r)} - \boldsymbol{\varphi(v_s)}||_\infty = \max_{i=1,\ldots,n} \left(|(\boldsymbol{\varphi(v_r)} - \boldsymbol{\varphi(v_s)})_i|\right) = \max_{i=1,\ldots,n} |\ell^*(v_r, v_i) - \ell^*(v_s, v_i)|$.
  כעת, מאי שוויון המשולש $|\ell^*(v_r, v_i) - \ell^*(v_s, v_i)| \leq \ell^*(v_r, v_s)$. ולכן $||\boldsymbol{\varphi(v_r)} - \boldsymbol{\varphi(v_s)}||_\infty \leq \ell^*(v_r, v_s)$.
- מנגד, $||\boldsymbol{\varphi(v_r)} - \boldsymbol{\varphi(v_s)}||_\infty = \max_{i=1,\ldots,n} |\ell^*(v_r, v_i) - \ell^*(v_s, v_i)| \geq \ell^*(v_r, v_s) - \ell^*(v_s, v_s)$.
  ולכן, $||\boldsymbol{\varphi(v_r)} - \boldsymbol{\varphi(v_s)}||_\infty \geq \ell^*(v_r, v_s)$.

הראנו אי שוויון בשני הכיוונים, אז מתקיים $||\boldsymbol{\varphi(v_r)} - \boldsymbol{\varphi(v_s)}||_\infty = \ell^*(v_r, v_s)$ כנדרש. □

אנחנו נתעניין, ספציפית, בנורמה 1, ולא בנורמת ∞. נזכיר כי $||\boldsymbol{x}||_1 = \sum_{i=1}^{\infty} |x_i|$. מסתבר שהיא מאוד שימושית וחשובה לנו, שכן יש קשר הדוק בינה ובין חתכים בגרפים.

בהינתן $(S, \overline{S})$ חתך של G, כאשר $\emptyset \neq S \subsetneq V$, אפשר להגדיר פסאודו-מטריקה $\ell_{S,\overline{S}}$ על V, הנקראת **פסאודו מטריקת חתך**: $\ell_{S,\overline{S}}(u,v) = \begin{cases} 1 & |S \cap \{u,v\}| = 1 \\ 0 & \text{אחרת} \end{cases}$ - כלומר מרחק בין 2 קודקודים הוא 1 אם הצלע ביניהם עוברת דרך החתך, ואחרת 0.

יש שיכון איזומטרי של $\ell_{S,\overline{S}}$ אל ℓ_1, שהוא $\varphi(u) = \begin{cases} 1 & u \notin S \\ 0 & u \in S \end{cases}$. ובנוסף:

טענה 4.10. בכל תת קבוצה סופית X עם n נקודות, של ℓ_1 (מכל מימד סופי d), אפשר לבטא את פונקציית המרחק כצירוף לינארי בעל מקדמים אי שליליים של פסאודו-מטריקות חתך.

כלומר אם נסתכל על הגרף שמתאים ל-n-הנקודות האלה (גרף מלא, שמרחק בין 2 נקודות בו הוא המרחק $||\boldsymbol{x} - \boldsymbol{y}||_1$), אפשר ליצור בו סדרת חתכים $\left(S_1, \overline{S_1}\right), ..., \left(S_t, \overline{S_t}\right)$ כך שלכל $\boldsymbol{x}, \boldsymbol{y} \in X$ אשר $||\boldsymbol{x} - \boldsymbol{y}||_1 = \sum_{i=1}^{t} c_i \ell_{S_i,\overline{S_i}}(\boldsymbol{x}, \boldsymbol{y})$ לאיזשהם קבועים $c_1, ..., c_t \geq 0$.

הוכחה.

הערת המסכם

אחדד בקרוב את כל ההוכחה הזו בדגש על השורות האחרונות.

בלי הגבלת הכלליות, $X \subseteq [a_1, b_1] \times ... \times [a_d, b_d]$, ואת זה אפשר לעשות כי $a_i = \min_{\boldsymbol{x} \in X} x_i$ ו-$b_i = \max_{\boldsymbol{x} \in X} x_i$.

כל $\rho \in (a_i, b_i)$ מגדירה חתך $\left(S_\rho^{(i)}, \overline{S_\rho^{(i)}}\right)$ כך: $S_\rho^{(i)} = \{\boldsymbol{x} \in X \mid x_i \le \rho_i\}$ - כלומר קוטמים את הציר ה-i, כל הנקודות מצד אחד מהוות קבוצה אחת וכל הנקודות מצד שני הקבוצה השניה.

מספר האופציות לבחירת ρ הוא מעוצמת הרצף (אינסופי). אבל למעשה בין 2 ערכים סמוכים בקורדינטה מסויימת, החתכים יהיו זהים (כי אין ערכים בין 2 הקורדינטות). כך מספר האופציות לבחירת ρ אשר יתנו לנו **חתכים שונים**, הוא $n \cdot d$ (אולי $(n-2) \cdot d$ בגלל הקצוות), משום שיש לנו d קורדינטות, ובכל קורדינטה אפשר למיין את כל n הנקודות לפי הקורדינטה הנוכחית.

לכל חתך מהצורה הנ"ל $(S, \overline{S})$ נגדיר מקדם $c_S = \sum_j \text{length}\left(\left\{\rho \mid S_\rho^{(j)} = S\right\}\right)$ כאשר $\text{length}(I)$ הוא אורך הקטע I.

כלומר c_S הוא למעשה סכומי האורכים, של הקטעים שיתנו לנו בדיוק את החתך S (אמרנו שבין 2 נקודות סמוכות, לא משנה איזה ρ נבחר - נקבל אותו חתך. אז נמדוד את אורך הקטע שהתופעה הזו קורית בה).

תהיינה $\boldsymbol{x}, \boldsymbol{y}$ נקודות. אז $||\boldsymbol{x} - \boldsymbol{y}||_1 = \sum_i |x_i - y_i| = \sum_i \text{length}([x_i, y_i])$.

נשים לב שהקטע $[x_i, y_i]$, הוא הקטע של כל ה-ρ-ים, שמקיימים ש-$S_\rho^{(i)}$ מכיל אך ורק אחד מבין x ו-y ולא את שניהם.

כלומר $||\boldsymbol{x} - \boldsymbol{y}||_1 = \sum_i \text{length}([x_i, y_i]) = \sum_i \text{length}\left(\left\{\rho \mid \left|S_\rho^{(i)} \cap \{\boldsymbol{x}, \boldsymbol{y}\}\right| = 1\right\}\right)$.

$\ell_{S_\rho^{(i)}}(\boldsymbol{x}, \boldsymbol{y}) = 1$ במקרה בו $S_\rho^{(i)} \cap \{\boldsymbol{x}, \boldsymbol{y}\}$, שכן הם נמצאים לא באותו צד של החתך.

נקבל $\ell_{S_\rho^{(i)}}(\boldsymbol{x}, \boldsymbol{y}) \left(\left\{\rho \mid \left|S_\rho^{(i)} \cap \{\boldsymbol{x}, \boldsymbol{y}\}\right| = 1\right\}\right) ||\boldsymbol{x} - \boldsymbol{y}||_1 = \sum_i \text{length}$, אבל במקרה ובו $\left|S_\rho^{(i)} \cap \{\boldsymbol{x}, \boldsymbol{y}\}\right| \ne 1$,

כלומר $S_\rho^{(i)}$ לא חותך את הצלע $\{\boldsymbol{x}, \boldsymbol{y}\}$, ממילא $\ell_{S_\rho^{(i)}}(\boldsymbol{x}, \boldsymbol{y}) = 0$.

אז הסכום הזה שווה גם ל-$||\boldsymbol{x} - \boldsymbol{y}||_1 = \sum_{(S, \overline{S})} c_S \ell^*(\boldsymbol{x}, \boldsymbol{y})$. $\square$

וכעת נכניס את השימוש בשיכונים מטריים: הטענה הבאה אומרת שבהינתן איזשהו שיכון φ, אפשר למצוא בעזרתו חסם עליון לדלילות של איזשהו חתך (השיכון ישתלב באיזשהו ביטוי), ולכן זה גם חסם עליון לקיבול של החתך עם הדלילות המינימלית.

טענה 4.11. יהא $G = (V, E)$ גרף לא מכוון, $c : E \to \mathbb{N}$ פונקציית קיבול על הצלעות, $(s_1, t_1), ..., (s_k, t_k)$ מוצרים, ו-$\varphi : V \to \mathbb{R}^d$ שיכון כאשר $\mathbb{R}^d$ מצויד עם נורמת ה-1 ו-V מצוייד עם הפסאודו מטריקה ℓ.

אז קיים חתך $(S, \overline{S})$ ככה ש-$\alpha(S, \overline{S}) = \frac{\sum_{e \in E \text{ s.t. } |e \cap S| = 1} c(e)}{|\{j \mid S \cap \{s_j, t_j\} = 1\}|} \le \frac{\sum_{e = \{u,v\} \in E} c(e) \cdot ||\varphi(u) - \varphi(v)||_1}{\sum_j ||\varphi(s_j) - \varphi(t_j)||_1}$.

הערה. בפרט אם φ איזומטרי, כלומר $||\varphi(a) - \varphi(b)||_1$ הוא הדרך הקצרה ביותר בגרף מ-a ל-b ביחס למשקולות ℓ על הצלעות, אנחנו מקבלים שאגף ימין הוא $\frac{\sum_{e = \{u,v\} \in E} c(e) \cdot \ell(u,v)}{\sum_j \ell(s_j, t_j)}$.

במילים אחרות, אגף ימין הוא ערך התוכנית הלינארית הדואלית חלקי סכום המרחקים של כל זוג מקור-בור.

הוכחה. נסתכל על החתך S מההוכחה הקודמת, כלומר סכום החתכים $S_\rho^{(i)}$ ב-$\mathbb{R}^d$ לכל $1 \le i \le d$ ולכל $\rho \in (a_i, b_i)$. מהגדרת S, מתקיים ש-$\sum_{e \in E \text{ s.t. } |e \cap S| = 1} c(e)$ היא הפונקציה שמהווה הסכום (על כל הקורדינטות iים) של קיבולי כל הצלעות שחוצות את החתך, לכל ρ בין a_i ל-b_i, זוהי $\sum_i \int_{a_i}^{b_i} \sum_{e \in E, \left|e \cap S_\rho^{(i)}\right| = 1} c(e) \, \mathbf{d}\rho$. הפונקציה הפנימית אינטגרבילית שכן אינה רציפה במספר סופי של נקודות (למעשה היא פונקציית מדרגות, ואין רציפות כשזזים בין ערך ρ אחד לאחר שפתאום משנה משהו בצלעות שחותכות את החתך). יש פה סכום סופי אז נשנה סדר סכימה:

$$\sum_i\int_{a_i}^{b_i}\sum_{e\in E,\ \left|e\cap S_\rho^{(i)}\right|=1}c(e)\,\mathbf{d}\rho=\sum_{e\in E}\sum_i\int_{a_i}^{b_i}c(e)\cdot\mathbb{1}_{\left|S_\rho^{(i)}\cap e\right|=1}(\rho)\,\mathbf{d}\rho=$$
$$=\sum_{\{\varphi(u),\varphi(v)\}\in\varphi(E)}c(\{u,v\})\cdot\sum_i\int_{a_i}^{b_i}\mathbb{1}_{\left|S_\rho^{(i)}\cap\{u,v\}\right|=1}(\rho)\,\mathbf{d}\rho$$

כאשר $\mathbb{1}_{\left|S_\rho^{(i)}\cap e\right|=1}$ היא פונקציית האינדיקטור להאם הצלע בחתך או לא.

אבל $\int_{a_i}^{b_i}\mathbb{1}_{\left|S_\rho^{(i)}\cap e\right|=1}$ הוא אינטגרל על פונקציה שהיא 1 או 0. ולכן האינטגרל הזה יהיה שווה לאורך הקטע שבו ρ מקיים $\left|S_\rho^{(i)}\cap e\right|=1$.

ואם $e=\{\varphi(u),\varphi(v)\}$, אז נשים לב שאורך הקטע הזה הוא $|\varphi_i(u)-\varphi_i(v)|$. בסך הכל נקבל:

$$\sum_i\int_{a_i}^{b_i}\sum_{e\in E,\ \left|e\cap S_\rho^{(i)}\right|=1}c(e)\,\mathbf{d}\rho=\sum_{\{\varphi(u),\varphi(v)\}\in\varphi(E)}c(\{u,v\})\cdot\sum_i\int_{a_i}^{b_i}\mathbb{1}_{\left|S_\rho^{(i)}\cap\{u,v\}\right|=1}\mathbf{d}\rho=$$
$$=\sum_{\{\varphi(u),\varphi(v)\}\in\varphi(E)}c(\{u,v\})\cdot\sum_i|\varphi_i(u)-\varphi_i(v)|=$$
$$=\sum_{\{\varphi(u),\varphi(v)\}\in\varphi(E)}c(\{u,v\})\cdot||\varphi(u)-\varphi(v)||_1$$

טיפלנו במונה. כעת נטפל במכנה $|\{j\mid S\cap\{s_j,t_j\}=1\}|$.

המכנה הוא למעשה, באופן דומה, הפונקציה $\sum_i\int_{a_i}^{b_i}\sum_{e\in E,\ \left|e\cap S_\rho^{(i)}\right|=1}1\mathbf{d}\rho$,

כלומר זו הפונקציה $\sum_{e=\{\varphi(u),\varphi(v)\}\in E}||\varphi(u)-\varphi(v)||_1$, והיא גדולה או שווה ל-$\sum_{e=\{s_i,t_i\}\in E}||\varphi(s_i)-\varphi(t_i)||_1$ (כי פה אנחנו סוכמים פחות איברים).

בסך הכל הראנו שקיים חתך S (סכום החתכים) ככה ש-$\frac{\sum_{e\in E\text{ s.t. }|e\cap S|=1}c(e)}{|\{j\mid S\cap\{s_j,t_j\}=1\}|}\le\frac{\sum_{e=\{u,v\}\in E}c(e)\cdot||\varphi(u)-\varphi(v)||_1}{\sum_j||\varphi(s_j)-\varphi(t_j)||_1}$, כנדרש. □

הרצאה 23

המסקנה היא שאם נמצא שיכון טוב (מבחינת עיוות) של (V,ℓ^*) אל $\left(\mathbb{R}^d,||\cdot||_1\right)$ אנחנו נוכל לחסום מלמעלה את גודל החתך הדליל ביותר.

משפט 4.12. יהא (V,ℓ) מרחב מטרי סופי. אז יש $\alpha,\beta>0$ קבועים, ככה שלכל $|V|=n$, ו-$X\subseteq V$ עם $|X|=m$ (יש $\binom{n}{m}$ אפשרויות בחירה כאלה), יש שיכון $\varphi:V\to\mathbb{R}^d$, עם נורמת ℓ_1, כאשר $d=O\left(\log^2 m\right)$, עבורו:

1. לכל $u,v\in V$, מתקיים $||\varphi(u)-\varphi(v)||\le\alpha\cdot\log m\cdot\ell(u,v)$.
2. לכל $u,v\in X$ מתקיים $||\varphi(u)-\varphi(v)||\ge\beta\cdot\ell(u,v)$.

ובנוסף ניתן לחשב את φ בעזרת אלגוריתם הסתברותי, בזמן פולינומי.[9]

הוכחה. לכל $i=0,1,...,\lfloor\log_2 m\rfloor$ נגדיר את ערך השיכון בקורדינטה ה-i.

אנחנו תחילה נגריל קבוצה B של $\frac{m}{2^i}$ נקודות מ-X (יש פה הנחה ש-m הוא חזקה של 2, אבל תמיד אפשר לקחת ערך שלם או לשכפל נקודות. נמשיך עם ההנחה הזו בהוכחה), אשר מוגרלות באופן אחיד ובלתי תלוי. מותרות חזרות. למשל כש-$i=0$ לוקחים את כל m הנקודות מ-X, כש-$i=1$ לוקחים חצי מהנקודות ב-X, $i=2$ רבע וכולי.

נחזור על התהליך $r_i=\left|\left\{B\subseteq X\mid |B|=\frac{m}{2^i}\right\}\right|=\binom{m}{\frac{m}{2^i}}$ פעמים. כלומר שימו לב ייתכן שבאינדקסים i שונים אנחנו ניקח מספר שונה של אינדקסים (r_i).

נסמן את הקבוצות שבחרנו בכל פעם, בהתאמה, (ה-Bים), ב-$B_{i_1},...,B_{i_{r_i}}$.

נגדיר מרחק של קודקוד מקבוצה להיות $\ell(u,B)=\min_{v\in B}\ell(u,v)$.

[9] אפשר גם להגדיל את המימד d ולקבל אלגוריתם דטרמיניסטי. המימד לא משנה לצורכינו.

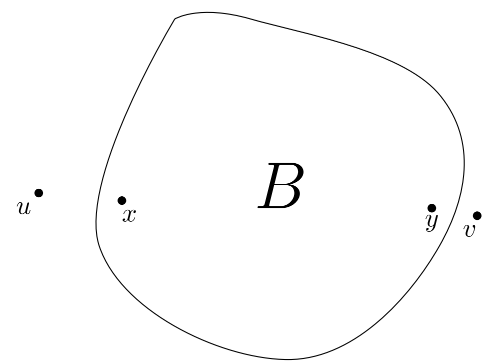
נגדיר את $\varphi_{i_j}(x)=\frac{1}{r_i}\ell\left(x,B_{i_j}\right)$, וכך את הקורדינטה ה-$i$ בשיכון להיות $\varphi_i(x)=\sum_{j=1}^{r_i}\varphi_{i_j}(x)$.
ננתח את השיכון:

1. יהיו $u,v\in V$. תחילה,
$$||\varphi(u)-\varphi(v)||_1=\sum_{i=0}^{\lfloor\log_2 m\rfloor}\sum_{j=1}^{r_i}\left|\varphi_{i_j}(u)-\varphi_{i_j}(v)\right|=\sum_{i=1}^{\lfloor\log_2 m\rfloor}\left|\frac{1}{r_i}\sum_{j=1}^{r_i}\left(\ell\left(u,B_{i_j}\right)-\ell\left(v,B_{i_j}\right)\right)\right|.$$

אבל מאי שוויון המשולש $\left|\frac{1}{r_i}\sum_{j=1}^{r_i}\left(\ell\left(u,B_{i_j}\right)-\ell\left(v,B_{i_j}\right)\right)\right|\le\frac{1}{r_i}\sum_{j=1}^{r_i}\left|\left(\ell\left(u,B_{i_j}\right)-\ell\left(v,B_{i_j}\right)\right)\right|$.

קיבלנו $||\varphi(u)-\varphi(v)||_1\le\sum_{i=1}^{\lfloor\log_2 m\rfloor}\frac{1}{r_i}\sum_{j=1}^{r_i}\left|\left(\ell\left(u,B_{i_j}\right)-\ell\left(v,B_{i_j}\right)\right)\right|$.

נסמן ב-x את הנקודה הקרובה ביותר ב-B_{i_j}, ל-u, ו-y הקרובה ביותר ל-v:



מתקיים $\ell\left(u,B_{i_j}\right)-\ell\left(v,B_{i_j}\right)=\ell(u,x)-\ell(v,y)$ מהתכונה של x,y.

אבל $\ell(u,x)\le\ell(u,y)$ (כי $y\in B_{i_j}$ ו-x הוא הקרוב ביותר ל-u מבין כל הנקודות ב-B_{i_j}).
לכן $\ell(u,x)-\ell(v,y)\le\ell(u,y)-\ell(v,y)$.

מאי שוויון המשולש נקבל $\ell(u,y)-\ell(v,y)\le\ell(u,v)$. קיבלנו $\ell\left(u,B_{i_j}\right)-\ell\left(v,B_{i_j}\right)\le\ell(u,v)$.

ולכן $||\varphi(u)-\varphi(v)||_1\le\sum_{i=0}^{\lfloor\log_2 m\rfloor}\frac{1}{r_i}\cdot\sum_{j=1}^{r_i}\left(\ell(u,v)\right)=\sum_{i=0}^{\lfloor\log_2 m\rfloor}\ell(u,v)=\left(\lfloor\log_2 m\rfloor+1\right)\ell(u,v)$.

הוכחנו שלכל $u,v\in V$ מתקיים $||\varphi(u)-\varphi(v)||\le\left(\lfloor\log_2 m\rfloor+1\right)\cdot\ell(u,v)$.

ניקח α קצת יותר גדולה מ-1 (לא תלויה ב-m) כזו שתקיים $\left(\lfloor\log_2 m\rfloor+1\right)\le\alpha\cdot\log m$ - אפשר גם $\alpha=2$ אם אתם רוצים. ואז $||\varphi(u)-\varphi(v)||\le\alpha\cdot\log m\cdot\ell(u,v)$.

2. עלינו להוכיח שלכל $x,y\in X$ מתקיים $||\varphi(x)-\varphi(y)||\ge\beta\cdot\ell(x,y)$.

נניח גם של-x ול-y, כל אחת, יש רשימה של איברי X הממויינת לפי המרחקים של הקודקודים מ-x או מ-y (בהתאמה) מהקטן לגדול (אם יש שניים במרחק זהה, נבחר אחד מהם שיופיע קודם - לא משנה מי).

בנוסף נגדיר $0=\rho_0\le\rho_1\le\rho_2\le\ldots\le\rho_{i_{\max}}$ באופן הבא:

$$\rho_i=\min\left\{\rho\ \middle|\ \left|B(x,\rho)\ge 2^i\ \text{וגם}\ \left|B(y,\rho)\right|\ge 2^i\right|\right\}$$

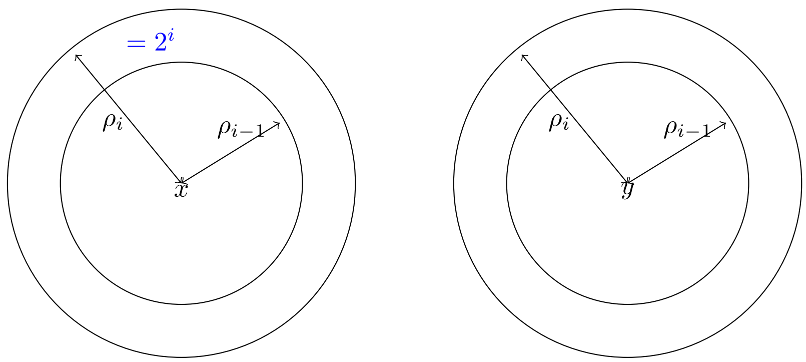
כאשר $i_{\max}$ הוא האינדקס המירבי ככה ש-$\rho_{i_{\max}}\le\frac{1}{3}\ell(x,y)$, ובאופן פרטני גם נגדיר $\rho_{i_{\max}+1}=\frac{1}{3}\ell(x,y)$.

אנחנו יודעים כי $\left|B(x,\rho_i)\right|,\left|B(y,\rho_i)\right|\ge 2^i$ מההגדרה. נגדיר את הכדורים ככה שאחד מהם (לא ידוע מי) הוא בגודל **בדיוק** 2^i (אפשר לעשות את זה בצורה הבאה - נמשיך להגדיל את הרדיוס ρ עוד ועוד, עד שאנחנו מגיעים

למצב שאחד מהכדורים הוא בגודל 2^i לפחות אבל השני עוד לא, ואז כשהשני מגיע לגודל בדיוק 2^i - נעצור. אנחנו יודעים לעשות את כל התהליך הזה ובזמן לינארי ב-$|X| = m$ כי יש לנו רשימה ממויינת של נקודות X ע"פ מרחקן מ-x, y.

נניח בה"כ שהכדור עם בדיוק 2^i נקודות הוא של u.

בנוסף שימו לב כי $B(x,\rho_i) \cap B(y,\rho_j) = \emptyset$ לכל i,j, שכן $\rho_i, \rho_j \le \rho_{i_{\max}} \le \frac{1}{3}\ell(x,y)$, ואז הכדורים האלה לא חופפים כלל - הרדיוסים של כל אחד מהם הוא שליש המרחק בין המרכזים.



נניח שנבחר סדרה אקראית של נקודות מ-X, מגודל $\frac{m}{2^i}$. ולשם הפשטות, נבחר עם חזרות (כלומר נקודות יכולות לחזור כמה פעמים בסדרה) - כי קל יותר לחשב הסתברות ככה (הנקודות בלתי תלויות) ועדיין נקבל חסם עליון.

נסמן את הסדרה הנבחרת ב-$A = \left(z_1, ..., z_{m/2^i}\right)$.

נרצה לחשב את ההסתברות שסדרה כזו מכילה לפחות נקודה אחת מ-$B(y,\rho_{i-1})$ אבל לא מכילה נקודה מ-$B(x,\rho_i)$. כי אז, בשיכון, ההפרש בין 2 הקורדינטות, יצא לפחות $\rho_i - \rho_{i-1}$.

ניקח איזושהי נקודה $z_j \in A$.

מה הסיכוי ש-$\Pr[z_j \in B(y,\rho_{i-1})]$? משום שלפחות 2^{i-1} נקודות מ-m נופלות בכדור $B(y,\rho_{i-1})$ מעצם הגדרתו, זה לכל הפחות $\frac{2^{i-1}}{m}$.

ובנוסף, $\Pr[z_j \in B(x,\rho_i)] \ge \frac{2^i}{m}$ מאותו נימוק ($|B(x,\rho_i)| = 2^i$) אז $\Pr[z_j \notin B(x,\rho_i)] \le 1 - \frac{2^i}{m}$.

עד עכשיו חישבנו על נקודה ספציפית. מה ההסתברות **שקיימת** נקודה שעונה על מה שרצינו?

$\Pr[\exists z\ z \in B(y,\rho_{i-1})] = 1 - \Pr[\forall z\ z \notin B(y,\rho_{i-1})]$ ובגלל שבחירות z הן בחירות בלתי תלויות, נקבל בסך הכל ש-$\Pr[\exists z\ z \in B(y,\rho_{i-1})] \ge 1 - \prod_{j=1}^{m/2^i} \Pr[z_j \notin B(y,\rho_{i-1})]$,

אבל עבור z ספציפי ראינו כי $\Pr[z \notin B(y,\rho_{i-1})] \le 1 - \frac{2^{i-1}}{m}$.

ולכן נקבל $\Pr[\exists z\ z \in B(y,\rho_{i-1})] \ge 1 - \left(1 - \frac{2^{i-1}}{m}\right)^{m/2^i} = 1 - \left(1 + \frac{-\frac{1}{2}}{\frac{m}{2^i}}\right)^{m/2^i}$. בהתחשב בזה ש-$\left(1 + \frac{-\frac{1}{2}}{\frac{m}{2^i}}\right)^{m/2^i} \le e^{-\frac{1}{2}}$ נקבל כעת כי $\Pr[\exists z\ z \in B(y,\rho_{i-1})] \ge 1 - \sqrt{\frac{1}{e}}$, ובפרט $1 - \sqrt{\frac{1}{e}}$ הוא קבוע.

ובנוסף, נסיק כי $\Pr[\forall z\ z \notin B(x,\rho_i)] = \prod_{j=1}^{m/2^i} \Pr[z_j \notin B(x,\rho_i)] \ge \left(1 - \frac{2^i}{m}\right)^{m/2^i} \ge \frac{1}{4}$ (סתם בחרנו $\frac{1}{4}$, העיקר שיהיה קבוע).

אבל אנחנו מתעניינים בחיתוך של שתי המאורעות: $\Pr[\forall z\ z \notin B(x,\rho_i)$ וגם $\exists z \in B(y,\rho_{i-1})]$.

מתקיים $\Pr[\forall z\ z \notin B(x,\rho_i)$ וגם $\exists z \in B(y,\rho_{i-1})] \ge \Pr[\forall z\ z \notin B(x,\rho_i)] \cdot \Pr[\exists z \in B(y,\rho_{i-1})]$.

כלומר נקבל $\Pr[\forall z\ z \notin B(x,\rho_i)$ וגם $\exists z \in B(y,\rho_{i-1})] \ge \frac{1}{4}\left(1 - \frac{1}{\sqrt{e}}\right)$, זה קבוע שנסמנו $\gamma = \frac{1}{4}\left(1 - \frac{1}{\sqrt{e}}\right)$.

אז במקרה כזה, בהסתברות הקבועה γ לפחות, לאיזושהו קורדינטה בשיכון i_j, מתקיים:

$\varphi_{i_j}(x)-\varphi_{i_j}(y)\geq\frac{1}{r_i}(\rho_i-\rho_{i-1})$, שכן $\varphi_{i_j}(x)\geq\frac{1}{r_i}\rho_i$ ו-$\varphi_{i_j}(y)\leq\frac{1}{r_i}\rho_{i-1}$. הרצאה 24

בהסתברות γ קורה מה שאמרנו, אז $\varphi_j(x)-\varphi_j(y)=\sum_{j=1}^{r_i}\varphi_{i_j}(x)-\varphi_{i_j}(y)\geq\gamma(\rho_i-\rho_{i-1})$.

כך, $||\varphi(x)-\varphi(y)||_1=\sum_{i=0}^{\lfloor\log_2 m\rfloor}\varphi_i(x)-\varphi_i(y)\geq\gamma\sum_{i=0}^{i_{\max}+1}(\rho_i-\rho_{i-1})$. זה סכום טלסקופי, אנחנו מקבלים שמתקיים $||\varphi(x)-\varphi(y)||_1\geq\gamma\cdot(\rho_{i_{\max}+1}-\rho_0)=\gamma\cdot\left(\frac{1}{3}\ell(x,y)\right)=\frac{\gamma}{3}\ell(x,y)$.

כלומר קיבלנו מה שרצינו עבור הקבוע $\beta=\frac{\gamma}{3}$. □

מסקנה 4.13. נשלב את טענה 4.11 ומשפט 4.12. מדואליות חזקה יש פתרון לתוכנית הדואלית ℓ,h אשר $\sum_{e=\{u,v\}\in E}c(e)\cdot\ell(u,v)=\text{max concurrent flow}$.

נתבונן בקבוצה $X=\{s_1,t_1,s_2,t_2,...,s_k,t_k\}$, שגודלה הוא $|X|\leq 2k$ (זה לכל היותר שכן ייתכן מקור/בור המשמשים כמה מוצרים). ממשפט 4.12 יש שיכון $\varphi:V\to\mathbb{R}^d$, ככה שמתקיים $||\varphi(u)-\varphi(v)||_1\leq\alpha\cdot\log(2k)\cdot\ell(u,v)$ לכל $u,v\in V$, וגם לכל s_i,t_i מתקיים $||\varphi(s_i)-\varphi(t_i)||_1\geq\beta\cdot\ell(s_i,t_i)$. ניזכר בטענה 4.11 ונקבל שקיים חתך $(S,\overline{S})$ ככה ש:

$$\begin{aligned}
\alpha(S,\overline{S}) &\leq \frac{\sum_{e=\{u,v\}\in E}c(e)\cdot||\varphi(u)-\varphi(v)||_1}{\sum_j||\varphi(s_j)-\varphi(t_j)||_1}\leq\\
&\leq \frac{\sum_{e=\{u,v\}\in E}c(e)\cdot\alpha\log(2k)\cdot\ell(u,v)}{\sum_j\beta\cdot\ell(s_j,t_j)}=\\
&= \frac{\alpha}{\beta}\cdot\log(2k)\cdot\frac{\sum_{e=\{u,v\}\in E}c(e)\cdot\ell(u,v)}{\sum_j\ell(s_j,t_j)}=\\
&= \frac{\alpha}{\beta}\cdot\log(2k)\cdot\frac{\sum_{e=\{u,v\}\in E}c(e)\cdot\ell(u,v)}{\sum_j\ell(s_j,t_j)}\leq\\
(**)\quad &\leq \frac{\alpha}{\beta}\cdot\log(2k)\cdot\sum_{e=\{u,v\}\in E}c(e)\cdot\ell(u,v)=\\
&= \frac{\alpha}{\beta}\cdot\log(2k)\cdot\text{max concurrent flow}=\\
&= O(\log k)\cdot\text{max concurrent flow}
\end{aligned}$$

כאשר המעבר $(**)$ נכון שכן (היזכרו בתוכנית הדואלית):

$$\begin{aligned}
\sum_j\ell(s_j,t_j) &= \sum_{i=1}^{k}\sum_{P\in P_i}\ell(P)\geq\\
&\geq \sum_{i=1}^{k}h_i\geq\\
&\geq 1
\end{aligned}$$

לסיכום אכן קיבלנו שהפער בין חתך דליל ביותר ובין זרימה בו-זמנית מקסימלית הוא $O(\log k)$.

למעשה ידוע היום קירוב טוב יותר לדלילות חתך מינימלי, משהו כמו $O\left(\sqrt{\log k}\log\log k\right)$. גם הוא מסתמך על שיכון ל-ℓ_1 אבל משתמש בכלי חזק יותר מהתוכנית הלינארית הדואלית שהשתמשנו בה (אפשר לנסח את טענה 4.11 ולקבל חסם טוב יותר לאיזשהו מספר מצומצם של מטריקות).

5 חתכים רב צדדיים

נתבונן במקרה מיוחד של זרימה רב מוצרית/רב חתך. נניח שנתונה רשת זרימה k-מוצרית, $X \subseteq V$, אשר $|X| = k$, והמוצרים כולם זוגות (s_i, t_i) המקיימים $s_i, t_i \in X$ (כאשר $s_i \neq t_i$). כלומר שימו לב שכל איבר ב-X הוא מקור/בור של כמה מוצרים (אולי לפעמים מקור ולפעמים בור).

בעיית הרב-חתך המינימלי במקרה הזה נקראת multiway cut - **חתך רב-צדדי** [מינימלי]. בבעיה זו עלינו להוריד צלעות ככה שכל זוגות המקור-בור ינותקו, והקיבול הכולל של הצלעות שהורדנו מינימלי.

התוכנית הדואלית לבעיית הזרימה הרב מוצרית, היא, כמו שאנו זוכרים, בעיית הרב-חתך **השברי** המינימלי:

$$\begin{aligned}
&\min \sum_{e\in E} c(e)\,\ell(e) \\
&\text{s.t.} \quad \forall 1 \le i \le k \text{ and } \forall P \in P_i \qquad \sum_{e\in P} \ell(e) \ge 1 \\
&\ell(e) \ge 0
\end{aligned}$$

נניח שאנחנו מסתכלים על מקור s_1, ומסתכלים על כל הקודקודים סביבו במרחק ρ (המרחקים נמדדים ע"פ ℓ), ככה ש-$\rho < \frac{1}{2}$. נסתכל על מקור אחר s_2. בגלל ש-$|X| = k$ וכל המוצרים נמצאים בקבוצה הזו, ובגלל האילוץ בתוכנית הדואלית, המרחק בין s_1 ו-s_2 חייב להיות לפחות 1. הכדור מרדיוס ρ סביב s_2, לפיכך, לא חותך את הכדור מרדיוס ρ סביב s_1.

אם צריך, אנחנו נפצל צלעות ארוכות מדי (כלומר נוסיף "קודקוד ביניים"), כאלה שמחברות בין 2 קודקודים שנמצאים בכמה כדורים (כלומר הקצה האחד נמצא בכדור מרדיוס ρ סביב s_i כלשהו, והקצה השני בכדור מרדיוס ρ סביב s_j אחר). ככה נבטיח מצב שכל צלע נמצאת בכדור אחד בדיוק.

שימו לב שבבעיית זרימה רב מוצרית פשוטה, לא מובטח לנו ש-s_1, s_2 "לא קרובים" (כמו כאן, שהמרחק ביניהם הוא לפחות 1), ולכן ייתכן שהתהליך כמו שעשינו הוא לא אפשרי. זו הייתה הסיבה שבבעיית רב החתך קיבלנו פקטור קירוב של $O(\log k)$. לכן, אולי פה אפשר לצפות לפקטור קירוב טוב יותר - אולי קבוע?

תהא צלע כלשהי $\{u, v\}$, ככה ש-$u, v \in B(s_1, \rho)$ (הכדור מרדיוס ρ סביב s_1), כאשר $\rho \sim_{\text{Unif}} \left(0, \frac{1}{2}\right)$.

מה ההסתברות ש-$\{u, v\}$ אכן נחתכת על ידי הכדור? (יש שלושה מצבים - הצלע כולה בתוך הכדור, הצלע נחתכת על ידי הכדור - כלומר קודקוד אחד בתוכו והשני מחוצה לו, והצלע כולה מחוץ לכדור. אנחנו מדברים על המקרה השני).

$\Pr[\text{נחתכת } \{u,v\}] = 2\,|\ell(s,u) - \ell(s,v)|$, שכן ρ נבחר באופן אחיד, וכך מא"ש המשולש $\Pr[\text{נחתכת } \{u,v\}] \le 2\ell(u,v)$.

כך ניתן להראות שהאלגוריתם שבוחר ρ באופן אחיד מ-$\left(0, \frac{1}{2}\right)$, הוא טוב בקירוב של 2 (כלומר מחזיר חתך שהוא עם קיבול לכל היותר פי 2 מהחתך המינימלי).

אפשר גם להביא אלגוריתם $2 \cdot \left(1 - \frac{1}{k}\right)$ מקרב עם איזשהו טריק.

האם אפשר יותר טוב?

> הערת המסכם
>
> לא סוכמו הרצאות נוספות.

הגדרות