

# אופטימיזציה קמורה (67731) -- דו"ח פרויקט

שם  
גיא לוי  
דניאל סגל  
מתן עצמוני

## 1 ניסוח שקול לשאלה

- טענה 1: מתקיים

$$\min_{K \succeq 0: K \text{ is } k\text{-band}} \text{Tr}(SK) - \log(|K|) \iff \min_{K \succ 0: K \text{ is } k\text{-band}} \text{Tr}(SK) - \log(|K|)$$

הוכחה: ראשית נשים לב כי בהינתן הגדרות השאלה אם ל- $K$  יש ע"ע השווה ל-0 הרי כי מתקיים

$$-\log(|K|) = -\log(0) = \infty$$

בעוד ש- $\text{Tr}(SK)$  סופי ולכן

$$\text{Tr}(SK) - \log(|K|) = \infty$$

מאחר ועבור  $K \succ 0$  כלשהו פונקציית המטרה משוערכת לערך סופי נסיק כי  $K$  עם ע"ע 0 אינם מהווים פתרון לבעיית מזעור הפונקציה ולכן נסיק שקילות של הבעיה המקורית לבעיה

$$\min_{K \succ 0: K \text{ is } k\text{-band}} \text{Tr}(SK) - \log(|K|)$$

□

כדרוש.

- טענה 2: מתקיים

$$\min_{K \succ 0: K \text{ is } k\text{-band}} \text{Tr}(SK) - \log(|K|) \iff \min_{\substack{R \in \mathbb{R}^{n \times n}: R \text{ is} \\ \text{upper triangular} \\ k\text{-band} \\ \text{positive diagonal}}} \left\{ \text{Tr}(RR^T S) - 2 \sum_{i=1}^n \log R_{ii} \right\}$$

כדי להוכיח את טענה זו נוכיח טענת עזר:

- טענת עזר 1: תהי  $K \in \mathbb{R}^{n \times n}$  מטריצה כך ש- $K \succ 0$  והי  $K = LL^T$  פירוק Cholesky של  $K$  כאשר  $L$  משולשית עליונה אזי  $K$  מטריצה  $k$ -band  $\iff L$  בעלת אלכסון חיובי ומטריצה  $k$ -band עם אותו ה- $k$ .  
הוכחת טענת עזר 1:

-  $\implies$ : נניח כי  $L$  הינה מטריצה  $k$ -band. מתקיים  $K = LL^T$  ולכן

$$\forall i, j \in [n]: K_{ij} = \sum_{t=1}^n L_{it} \cdot L_{jt}^T = \sum_{t=1}^n L_{it} \cdot L_{jt}$$

ולכן

$$\begin{aligned} L_{it} \neq 0 &\implies -k \leq i - t \leq k \implies -k \leq i - t \leq 0 \\ L_{jt} \neq 0 &\implies -k \leq t - j \leq k \implies 0 \leq t - j \leq k \end{aligned}$$

ולכן נקבל שקיים  $t$  עבורו  $L_{it} \neq 0$  וגם  $L_{jt} \neq 0$  אם  $-k \leq i - j \leq k$  ולכן נסיק כי  $K$  הינה מטריצה  $k$ -band.  
כדרוש.

–  $\Leftarrow$ : נוכיח זאת באינדוקציה על העמודות של  $K$  החל מהעמודה הימנית עד לעמודה הכי שמאלית.  
בסיס: יהי  $i$  כך ש- $|i - n| > k$ , כך ש- $K_{in} = 0$  אזי:

$$0 = K_{in} = \sum_{j=1}^n L_{ij} L_{jn}^T = \sum_{j=1}^n L_{ij} L_{nj} = L_{in} L_{nn}$$

$$\Rightarrow L_{in} L_{nn} = 0 \Rightarrow L_{in} = 0$$

כיוון ש- $L_{nn} \neq 0$ . כלומר, העמודה ה- $n$  של  $L$  מקיימת את תכונת ה- $k$ -band.  
צעד: כעת, נניח כי העמודות ה- $n, n-1, n-2, \dots, j+1, j$  של  $L$  מקיימות את תכונת ה- $k$ -band:

$$|(j+m) - i| > k \Rightarrow L_{i,j+m} = 0$$

יהי  $i$  כך ש- $|i - j| > k$ , כלומר  $K_{ij} = 0$  אזי:

$$0 = K_{ij} = \sum_{t=1}^n L_{it} L_{jt}$$

כיוון ש- $L$  משולשית עליונה, עבור  $t < i$  מתקיים  $L_{it} = 0$ , לכן:

$$= \sum_{t=i}^n L_{it} L_{jt}$$

מהנחת האינדוקציה, אם  $t > j$  מתקיים  $L_{it} = 0$  (כיוון שאם  $j - i > k$  אז גם  $t - i > k$  עבור  $t > i$ ). לכן:

$$= \sum_{t=i}^j L_{it} L_{jt}$$

עבור  $t < j$  מתקיים  $L_{jt} = 0$  כיוון ש- $L$  משולשית עליונה, לכן:

$$= L_{ij} L_{jj}$$

וכיוון ש- $L_{jj} \neq 0$  נקבל ש- $L_{ij} = 0$ . כלומר העמודה ה- $j$  מקיימת את הנחת ה- $k$ -band.  
 מעיקרון האינדוקציה נסיק כי  $L$  היא מטריצה משולשית עליונה  $k$ -band. כדרוש.

□

ונסיק את נכונות הטענה כדרוש.

- הוכחת טענה 2: מטענת עזר 1 נסיק כי  $K > 0$   $k$ -band אם "קיימת מטריצה  $R$  משולשית עליונה בעלת אלכסון חיובי וגם כן  $k$ -band כך ש- $K = RR^T$ . נשים לב כי

$$|K| = |RR^T| = |R| \cdot |R^T| = |R|^2 = \left( \prod_{i=1}^n R_{ii} \right)^2$$

כאשר השוויון האחרון נובע מכך שדטרמיננטה של מטריצה משולשית שווה למכפלת האיברים באלכסון ולכן אם  $f$  פונקצית המטרה שלנו, כלומר

$$f(K) := \text{Tr}(SK) - \log(|K|)$$

נוכל להגדיר

$$g(R) := f(RR^T) = \text{Tr}(RR^T S) - 2 \sum_{i=1}^n \log R_{ii}$$

וממה שכתבנו לעיל לכל  $K$  יהיה קיים  $R$  כך ש- $K = RR^T$  ולכן  $g(R) = f(K)$  ולכל  $R$  יהיה מתקיים מהגדרה  $g(R) = f(RR^T)$  ונסיק את השקילות כדרוש.  
 □

## 2 הפתרון

- מטרה: בהינתן  $S \in S_+^n$  ופרמטר  $k \in \mathbb{N}$  למצוא

$$\arg \min_{\substack{R \in \mathbb{R}^{n \times n}: R \text{ is} \\ \text{upper triangular} \\ k\text{-band} \\ \text{positive diagonal}}} \left\{ \text{Tr} (RR^T S) - 2 \sum_{i=1}^n \log R_{ii} \right\}$$

- פתרון:

- נסמן את המרחב מעליו אנחנו מחפשים

$$C = \{R \in \mathbb{R}^n \mid R \text{ is upper triangular, } k\text{-band, positive diagonal}\}$$

נוכיח שזוהי קבוצה קמורה: יהיו  $A, B \in C$  ו-  $\theta \in [0, 1]$  אז:

– משולשית עליונה: יהיו  $1 \leq j < i \leq n$  הרי כי

$$[\theta A + (1 - \theta) B]_{ij} = \theta [A]_{ij} + (1 - \theta) [B]_{ij} = \theta \cdot 0 + (1 - \theta) \cdot 0 = 0$$

כאשר השוויון הלפני אחרון נובע מכך ש-  $A, B$  משולשיות עליונות ולכן  $\theta A + (1 - \theta) B$  משולשית עליונה.

–  $k$ -band: יהיו  $i, j \in [n]$  כך ש-  $|i - j| > k$  הרי כי

$$[\theta A + (1 - \theta) B]_{ij} = \theta [A]_{ij} + (1 - \theta) [B]_{ij} = \theta \cdot 0 + (1 - \theta) \cdot 0 = 0$$

כאשר השוויון הלפני אחרון נובע מכך ש-  $A, B$  הן מטריצות  $k$ -band ולכן  $\theta A + (1 - \theta) B$  מטריצה  $k$ -band.

– אלכסון חיובי: יהי  $i \in [n]$  הרי כי

$$[\theta A + (1 - \theta) B]_{ii} = \theta [A]_{ii} + (1 - \theta) [B]_{ii}$$

נשים לב כי מתקיים

$$(\theta \geq 0 \wedge 1 - \theta > 0) \vee (\theta > 0 \wedge 1 - \theta \geq 0), \quad [A]_{ii} > 0, \quad [B]_{ii} > 0$$

ולכן סך הכל קיבלנו

$$\theta [A]_{ii} + (1 - \theta) [B]_{ii} > 0$$

ולכן  $\theta A + (1 - \theta) B$  בעלת אלכסון חיובי.

ולכן נסיק כי  $\theta A + (1 - \theta) B \in C$  כדרוש.

- נסמן את פונקציית המטרה שלנו

$$f(R) = \text{Tr} (RR^T S) - 2 \sum_{i=1}^n \log R_{ii}$$

נוכיח שהפונקציה קמורה ב-  $R$ : נסמן

$$\forall A, B \in C : g(t) := f(tA + (1 - t)B) = f(B + t(A - B))$$

ונוכיח כי  $g$  קמורה ב-  $t$  לכל  $A, B \in C$ : נפתח את ההגדרה של  $g$

$$\begin{aligned} g(t) &= \text{Tr} \left( (B + t(A - B))(B + t(A - B))^T S \right) - 2 \sum_{i=1}^n \log ([B + t(A - B)]_{ii}) \\ &= \text{Tr} (BB^T S) + t \cdot \left( \text{Tr} (B(A - B)^T S) + \text{Tr} ((A - B)B^T S) \right) + t^2 \cdot \text{Tr} ((A - B)(A - B)^T S) \\ &\quad - 2 \sum_{i=1}^n \log ([B]_{ii} + t \cdot ([A]_{ii} - [B]_{ii})) \end{aligned}$$

נגזור:

$$\frac{\partial g(t)}{\partial t} = \left( \text{Tr} \left( B (A - B)^T S \right) + \text{Tr} \left( (A - B) B^T S \right) \right) + 2t \cdot \text{Tr} \left( (A - B) (A - B)^T S \right) - 2 \sum_{i=1}^n \frac{[A]_{ii} - [B]_{ii}}{[B]_{ii} + t \cdot ([A]_{ii} - [B]_{ii})}$$

נגזור שוב:

$$\frac{\partial^2 g(t)}{\partial t^2} = 2 \cdot \text{Tr} \left( (A - B) (A - B)^T S \right) + 2 \cdot \sum_{i=1}^n \left( \frac{[A]_{ii} - [B]_{ii}}{[B]_{ii} + t \cdot ([A]_{ii} - [B]_{ii})} \right)^2$$

טרינומי כי

$$2 \cdot \sum_{i=1}^n \left( \frac{[A]_{ii} - [B]_{ii}}{[B]_{ii} + t \cdot ([A]_{ii} - [B]_{ii})} \right)^2 \geq 0$$

ונשים לב כי  $(A - B) (A - B)^T$  הוא מכפלה של משולשית עליונה בטרנספוז שלה ולכן PSD ולכן משאלה 16 בתרגיל 0<sup>1</sup> נסיק כי גם

$$\text{Tr} \left( (A - B) (A - B)^T S \right) \geq 0$$

ולכן קיבלנו כי

$$\frac{\partial^2 g(t)}{\partial t^2} \geq 0$$

ולכן  $g$  קמורה ב- $t$  לכל  $A, B$  ולכן ממשפט 5 מסיכומי ההרצאה<sup>2</sup> נסיק כי  $f$  קמורה ב- $R$  כדרוש. □

- עכשיו לאחר שהוכחנו כי הפונקציה קמורה, נסיק ששימוש ב- $\text{gd}$  עם גודל צעד מספיק קטן תמיד יקרב אותנו לפתרון האופטימלי.

## 2.1 ניסיון עם Gradient Descent

ניסינו להשתמש באלגוריתם gradient descent בשביל האופטימיזציה באופן הבא - נתחיל במטריצה משולשית עליונה  $R$  שהיא  $k$ -banded PSD, שהיא גם למעשה הפירוק צ'ולסקי של  $K$ :

$$f(K) = \text{Tr}(SK) - \log |K| = \text{Tr}(SRR^T) - \log |RR^T|$$

$$= \text{Tr}(R^T SR) - 2 \sum_{i=1}^n \log R_{ii}$$

כעת ניתן לגזור לפי  $R$  (להסבר הגזירה ראה את הפרק "הפתרון האופטימלי"):

$$\nabla_R (f(RR^T)) = 2SR - 2 \begin{bmatrix} \frac{1}{R_{11}} & 0 & \dots & 0 \\ \cdot & \frac{1}{R_{22}} & \dots & 0 \\ \cdot & 0 & \dots & 0 \\ 0 & \dots & 0 & \frac{1}{R_{nn}} \end{bmatrix} = 2SR - 2 \text{diag} \left( \frac{1}{R} \right)$$

ולעשות GD באופן הבא:

$$R_{t+1} = R_t - (\text{step size}) 2 \left[ \text{mask} \odot SR - \text{diag} \left( \frac{1}{R} \right) \right]$$

כאשר  $\text{mask}$  היא מטריצה המורכבת מאפסים ואחדים, כך שהכפל שלה איבר-איבר ב- $SR$  מאפס את הערכים המתאימים לפי  $k$ -band והמשולשיות של המטריצה  $R$ .

ביחד עם line-search (בשביל למצוא גודל צעד טוב) האלגוריתם הנ"ל עבד טוב למטריצות קטנות, אך לקח לו זמן רב מדי להתכנס עבור מטריצות בגודל  $50 \times 50$  ומעל, ולכן ניסינו גישה שונה המתבססת על פתרון של מערכת משוואות, שהתגלתה כיעילה יותר.

<sup>1</sup>טענה: אם  $A, B \in S_+^n$  אזי  $\text{Tr}(AB) \geq 0$   
<sup>2</sup>משפט: פונקציה  $f(x)$  קמורה ב- $x$  אם  $f(ty + (1-t)x) \geq f(ty) + t(f(x) - f(ty))$  קמורה ב- $t$  לכל  $x, y$

## 2.2 הפתרון האופטימלי

- מצד שני מאחר והפונקציה קמורה נוכל לגזור ולהשוות ל-0 ואם נמצא פתרון נסיק כי זהו המינימום של הפונקציה. נגזור את פונקצית המטרה לפי  $R$ , נשתמש בזהות גזירה הבאה:

$$\nabla_A (\text{Tr}(AA^T B)) = (B + B^T) A$$

במקרה שלנו  $S$  סימטרית ולכן נקבל

$$\nabla_R (\text{Tr}(RR^T S)) = 2SR$$

וכן מתקיים

$$\nabla_R \left( 2 \sum_{i=1}^n \log R_{ii} \right) = 2 \cdot \text{diag} \left( \frac{1}{R_{11}}, \dots, \frac{1}{R_{nn}} \right)$$

ולכן נקבל

$$\nabla_R \left( \text{Tr}(RR^T S) - 2 \sum_{i=1}^n \log R_{ii} \right) = 2SR - 2 \cdot \text{diag} \left( \frac{1}{R_{11}}, \dots, \frac{1}{R_{nn}} \right)$$

נשים לב כי מה שמעניין אותנו זה רק הנגזרות לפי המשתנים הלא מאופסים ב- $R$  ולכן לאחר השוואת הגרדיאנט ל-0 נקבל את המשוואות

$$i \in [n], \quad i < j \leq \min\{i+k, n\} : \frac{\partial f(R)}{\partial R_{ij}} = 2 \cdot \sum_{t=\max\{j-k, 1\}}^j S_{i,t} \cdot R_{t,j} = 0$$

$$i \in [n] : \frac{\partial f(R)}{\partial R_{ii}} = 2 \cdot \sum_{t=\max\{i-k, 1\}}^i S_{i,t} \cdot R_{t,i} - \frac{2}{R_{ii}} = 0$$

עתה נראה כי אפשר לחלק את המשוואות לקבוצות ומכל קבוצה למצוא את המשתנים של המשוואות בה:  
לכל  $j \in [n]$  נסמן  $d = \max\{j-k, 1\}$  ונשים לב כי עבור המשתנים

$$R_{d,j}, R_{d+1,j}, \dots, R_{j,j}$$

(המשתנים בעמודה ה- $j$  של  $R$ ) יש בדיוק  $j-d+1$  המשוואות

$$\forall i \in \{d, \dots, j-1\} : \sum_{t=d}^j S_{i,t} \cdot R_{t,j} = 0$$

והמשוואה

$$\sum_{t=d}^j S_{j,t} \cdot R_{t,j} = \frac{1}{R_{jj}}$$

נראה פרוצדורה לפתירת מערכת המשוואות הזו לכל  $j$ : נגדיר את מטריצת המקדמים של כל המשוואות (כאשר נתעלם מהאיבר  $\frac{1}{R_{jj}}$  במשוואה האחרונה)  $A^{j-d+1} \in \mathbb{R}^{(j-d+1) \times (j-d+1)}$ , כלומר השורה ה- $i$  של מטריצה זו מייצגת את המקדמים של המשתנים במשוואה ה- $i$ , כלומר

$$[A^{j-d+1}]_{i,r} = S_{i,d+r}$$

בהינתן שהגדרנו את המטריצה  $A^m \in \mathbb{R}^{m \times m}$  נרצה להגדיר את  $A^{m-1} \in \mathbb{R}^{(m-1) \times (m-1)}$  עבור  $1 < m \leq j-d+1$  באופן הבא:

$$\forall i, j \in [m-1] : [A^{m-1}]_{i,j} = [A^m]_{i+1,j+1} - \frac{[A^m]_{1,j+1} \cdot [A^m]_{i+1,1}}{[A^m]_{1,1}}$$

כאשר המשמעות של ביטוי זה זה המקדם של המשתנה ה- $j+1$  במערכת המשוואות המתאימה ל- $A^m$  לאחר הצבה של המשתנה ה-1 כשמחלצים אותו כביטוי של שאר המשתנים מהמשוואה הראשונה, וב- $A^{m-1}$  מקטינים ב-1 גם את כמות המשוואות וגם את כמות המשתנים ב-1 אז זה הופך למקדם של המשתנה ה- $j$ .  
 עתה נשים לב כי  $A^1 \in \mathbb{R}^{1 \times 1}$  והמשוואה האחרונה היא מהצורה

$$c \cdot R_{jj} = \frac{1}{R_{jj}} \implies R_{jj} = \frac{1}{\sqrt{c}}$$

ולכן

$$R_{jj} = \frac{1}{\sqrt{[A^1]_{11}}}$$

עבור  $i \in \{d, \dots, j-1\}$  בהינתן פתרונות למשתנים

$$R_{j-i+1,j}, R_{j-i+2}, \dots, R_{j,j}$$

נמצא את הפתרון ל- $R_{j-i,j}$  על-ידי

$$R_{j-i,j} = - \sum_{t=1}^{i-1} R_{j-i+t,j} \cdot \frac{[A^{i+1}]_{1,i+1}}{[A^{i+1}]_{1,1}}$$

כאשר הביטוי האחרון זה בדיק החילוץ של המשתנה ה-1 ב- $A^{i+1}$  כביטוי של שאר המשתנים ולכן בעת סיום התהליך נקבל פתרונות לכל המשתנים

$$R_{d,j}, R_{d+1,j}, \dots, R_{j,j}$$

כדרוש.

- **סיבוכיות זמן ריצה:** לכן מציאת כל המשתנים תתרחש על-ידי הפעלה חוזרת של הפרוצדורה לעיל על קבוצות של כל המשתנים, יש קבוצות משתנים כאלו ככמות העמודות ב- $R$ , כלומר כ- $n$  קבוצות. נשים לב כי עלות האיטרציה ה- $i$  מהסוף עולה  $O(i^2)$  (עלות יצירת המטריצת המקדמים), ולכן עלות הפרוצדורה תהיה

$$\sum_{i=1}^{j-d+1} O(i^2) \leq \sum_{i=1}^k O(i^2) = O(k^3)$$

סך הכל נפעיל את הפרוצדורה כ- $n$  פעמים ונקבל את זמן הריצה

$$O(nk^3)$$

ולכן האלגוריתם אכן מתחשב במימד  $S$  (רלוונטי לסעיף 5 בדיווח התרגיל)  
סיבוכיות מקום: נוכל לשים לב שלא באמת צריך ליצור מטריצה חדשה כל פעם אלא לעדכן אינדקסים במטריצה הראשונית (ראה "אופטימיזציה של הקוד") שניצור ולכן נקבל סיבוכיות מקום של  $O(k^2)$ , ואם מתחשבים בסיבוכיות מקום הפלט נקבל  $O(nk)$  בסך הכל.

## 2.3 אופטימיזציה של הקוד

אחרי מספר נסיונות של מימוש טריוויאלי לאלגוריתם לעיל, ראינו כי הוא מהיר יותר מ-GD, אך למטריצות גדולות עם *bandness* יחסית גבוה הוא עדיין לא היה יעיל מספיק, וזמן הריצה שלו היה מעל 2 דקות. לכן, רצינו לשפר את יעילות האלגוריתם מבחינת מימוש בוד. עשינו מספר דברים על מנת ליעל אותו:

השיפור המשמעותי ביותר נבע מווקטוריזציה של קוד במקום לולאות - לדוגמה, במקום לעבוד על איבר-איבר במטריצה כאשר מחשבים מקדמים, לכתוב את הפעולות בשורת קוד אחת בצורה ווקטורית על מנת לחשב את מטריצת המקדמים. השינוי הזה שיפר את זמן הריצה על הדוגמה החמישית (מטריצה  $500 \times 500$ ) בכ-158 שניות. בנוסף, שינויים נוספים כגון עבודה עם מטריצה קבועה ללא הקצאה מחדש, ומימוש איטרטיבי במקום רקורסיבי שיפרו את זמן הריצה על הדוגמה הזאת בכ-3 שניות בערך.

בנוסף לכך, הוספנו לאלכסון של  $S$  מספר מאוד קטן בשביל יציבות נומרית, ולמקרה שבו  $n = k$  לא השמשתנו באלגוריתם הנ"ל, אלא בפתרון  $K = S^{-1}$ , שהוא מהיר יותר, פיזיבילי ואופטימלי למקרה של  $n = k$ .