אופטימיזציה קמורה (67731) -- דו"ח פרוייקט

9 בנובמבר 2021

שם
גיא לוי
דניאל סגל
מתן עצמוני

1 ניסוח שקול לשאלה

- טענה 1: מתקיים

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

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

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

בעוד ש- $\mathrm{Tr}\left(SK\right)$ סופי ולכן

$$\mathrm{Tr}\left(SK\right)-\log\left(|K|\right)=\infty$$

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

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

כדרוש. $\square$

- טענה 2: מתקיים

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

כדי להוכיח את טענה זו נוכיח טענת עזר:

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

הוכחת טענת עזר 1:

  - $\Longrightarrow$: נניח כי L הינה מטריצה $k-\text{band}$. מתקיים $K=LL^T$ ולכן

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

ולכן

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

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

- $\Longleftarrow$: נוכיח זאת באינודקציה על העמודות של K החל מהעמודה הימנית עד לעמודה הכי שמאלית.
<u>בסיס</u>: 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־$L_{nn}\neq0$. כלומר, העמודה ה־n של L מקיימת את תכונת ה$k-\mathrm{band}$.
<u>צעד</u>: כעת, נניח כי העמודות ה־$j+1,j+2,...,n$ של L מקיימת את תכונת ה־$k-\mathrm{band}$:

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

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$L_{jj}\neq0$ נקבל ש־$L_{ij}=0$. כלומר העמודה ה־j מקיימת את הנחת ה$k-\mathrm{band}$.
מעיקרון האינדוקציה נסיק כי L היא מטריצה משולשית עליונה $k-\mathrm{band}$. כדרוש.

ונסיק את נכונות הטענה כדרוש. $\square$

- הוכחת טענה 2: מטענת עזר 1 נסיק כי $K\succ0$ $k-\mathrm{band}$ אם"ם קיימת מטריצה R משולשית עליונה בעלת אלכסון חיובי וגם כן $k-\mathrm{band}$ כך ש־ $K=RR^{T}$. נשים לב כי

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

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

$$f\left(K\right):=\mathrm{Tr}\left(SK\right)-\log\left(|K|\right)$$

נוכל להגדיר

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

וממה שכתבנו לעיל לכל K יהיה קיים R כך ש־ $K=RR^{T}$ ולכן $g\left(R\right)=f\left(K\right)$ ולכל R יהיה מתקיים מהגדרה $g\left(R\right)=f\left(RR^{T}\right)$ ונסיק את השקילות כדרוש. $\square$

2 הפתרון

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

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

- פתרון:
- נסמן את המרחב מעליו אנחנו מחפשים

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

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

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

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

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

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

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

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

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

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

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

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

נוכיח שהפונקציה קמורה ב־ R: נסמן

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

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

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

נגזור:

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

נגזור שוב:

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

טריוואלי 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[1] נסיק כי גם

$$\mathrm{Tr}\left((A-B)(A-B)^TS\right)\geq 0$$

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 [2] נסיק כי f קמורה ב־ R כדרוש. □

- עכשיו לאחר שהוכחנו כי הפונקציה קמורה, נסיק ששימוש ב־ 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 ו־PSD, שהיא גם למעשה הפירוק צ'ולסקי של K:

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

$$=\mathrm{Tr}\left(R^TSR\right)-2\sum_{i=1}^{n}\log R_{ii}$$

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

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

ולעשות GD באופן הבא:

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

כאשר mask היא מטריצה המורכבת מאפסים ואחדים, כך שהכפל שלה איבר־איבר בSR מאפס את הערכים המתאימים לפי ה$k-\text{band}$ 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 50×50 ומעל, ולכן ניסינו גישה שונה המתבססת על פתרון של מערכת משוואות, שהתגלתה כיעילה יותר.

[1] טענה: אם $A,B\in \mathbb{S}^n_+$ אזי $\mathrm{Tr}(AB)\geq 0$

[2] משפט: פונקציה $f(x)$ קמורה ב־ x אם"ם $g(t)=f(ty+(1-t)x)$ 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:

$$\nabla_A\left(\mathrm{Tr}\left(AA^TB\right)\right)=\left(B+B^T\right)A$$

במקרה שלנו S סימטרית ולכן נקבל

$$\nabla_R\left(\mathrm{Tr}\left(RR^TS\right)\right)=2SR$$

וכן מתקיים

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

ולכן נקבל

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

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

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

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

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

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

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

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:

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

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

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

נמצא את הפתרון ל־ $R_{j-i,j}$ על־ידי

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

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

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

כדרוש.

- <u>סיבוכיות זמן ריצה</u>: 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 $O\left(i^2\right)$ (עלות יצירת המטריצת המקדמים), ולכן עלות הפרוצדורה תהיה

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

סך הכל נפעיל את הפרודצורה כ־ n פעמים ונקבל את זמן הריצה

$$O\left(nk^3\right)$$

ולכן האלגוריתם אכן מתחשב במימד S (רלוונטי לסעיף 5 בדיווח התרגיל)
<u>סיבוכיות מקום</u>: 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 $O\left(k^2\right)$, ואם מתחשבים בסיבוכיות מקום הפלט נקבל $O\left(nk\right)$ בסך הכל.

2.3 אופטימיזציה של הקוד

אחרי מספר נסיונות של מימוש טריוויאלי לאלגוריתם לעיל, ראינו כי הוא מהיר יותר מGD, 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 מימושו בקוד. עשינו מספר דברים על מנת לייעל אותו:

השיפור המשמעותי ביותר נבע מוקטוריזציה של קוד במקום לולאות - לדוגמה, במקום לעבוד על איבר־איבר במטריצה כאשר מחשבים מקדמים, לכתוב את הפעולות בשורת קוד אחת בצורה וקטורית על מנת לחשב את מטריצת המקדמים. השינוי הזה שיפר את זמן הריצה על הדוגמה החמישית(מטריצה 500×500) 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, אלה בפתרון $K=S^{-1}$, שהוא מהיר יותר, פיזיבילי ואופטימלי למקרה של $n=k$.