دانشکده علوم ریاضی
دانشگاه صنعتی شریف

# الگوریتم‌های خلاصه‌سازی برای مه‌داده

**محمدهادی فروغمنداعرابی**
پاییز ۱۳۹۹

# احساس فشردگی تنک

جلسه هجدهم

**نگارنده: نسترن بهروزنیا**

## ۱ مروری بر مباحث گذشته

در جلسه قبل، با مفهوم احساس فشردگی آشنا شدیم. مسئله بدین صورت است که خلاصه سازی خطی بردار $x \in \mathbb{R}^n$ را به صورت $\pi x$ ذخیره کرده ایم که $\pi \in \mathbb{R}^{m\times n}$ یک ماتریس خلاصه سازی است و میخواهیم فقط با دسترسی به بردار $\pi x$ و ماتریس $\pi$، بردار $x$ و یا تقریبی از آن را بازیابی کنیم.
تاکنون، در الگوریتم های بررسی شده برای حل مسئله احساس فشردگی در حالتی که $x$ یک بردار $k$ـ تنک یا تقریبا $k$ـ تنک است، از جمله الگوریتم کمینه نرم ۱، از ماتریس $RIP$ استفاده کرده ایم و نشان دادیم که اگر $\pi$ یک ماتریس $RIP - (\epsilon, Ck)$ باشد، آنگاه تقریب $\hat{x}$ را میتوان به دست آورد، به طوری که:

$$\|\hat{x} - x\|_1 \leq c(k) \cdot \min_{\text{y is k-sparse}} \|y - x\|_1$$

## ۲ مقدمه

حال ممکن است این سوال مطرح شود که چه ماتریسی برای خلاصه سازی، بهترین عملکرد را در احساس فشردگی دارد؟ برای رسیدن به پاسخ این سوال، بایستی سعی کنیم بین تنک بودن و غیرتنک بودن ماتریس به حالتی متعادل از لحاظ زمان محاسبه و حافظه موردنیاز برای ذخیره $\pi x$ و $\pi$برسیم:
ـ استفاده از ماتریس تنک در الگوریتم، زمان محاسبه‌ی کم از مرتبه $O(k^2 nnz(A))$ دارد؛ اما خلاصه‌سازی آن طولانی است، چون $m$ بردار ذخیره شده بالا و از مرتبه $O(k^2)$ است.

ــ استفاده از ماتریس غیرتنک در الگوریتم، زمان محاسبه‌ی زیاد از مرتبه $O(kn^2 \log n)$ دارد؛ اما خلاصه‌سازی آن کوتاه است، چون $m$ بردار ذخیره شده کم و از مرتبه $O(k \log n)$ است.

# ۳ ماتریس $RIP_1$

برای رسیدن به یک ماتریس خلاصه سازی تنک که منجر به حافظه ذخیره سازی نسبتا کم و زمان محاسبه‌ی نسبتا مناسب در مسئله احساس فشردگی شود، ماتریس $RIP$ را کنار میگذاریم و با تغییر رویکرد از نرم۲ به نرم۱، از ماتریس $RIP_1$ استفاده میکنیم.

**تعریف ۱** (ماتریس $RIP_1$). *ماتریس $A$ را $(\epsilon, k) - RIP_1$ می‌نامیم، هرگاه به ازای هر بردار $k$ــ تنک $v$ داشته باشیم:*

$$(1-\epsilon)\|v\|_1 \le \|Av\|_1 \le (1+\epsilon)\|v\|_1$$

# ۴ ساخت ماتریس $RIP_1$

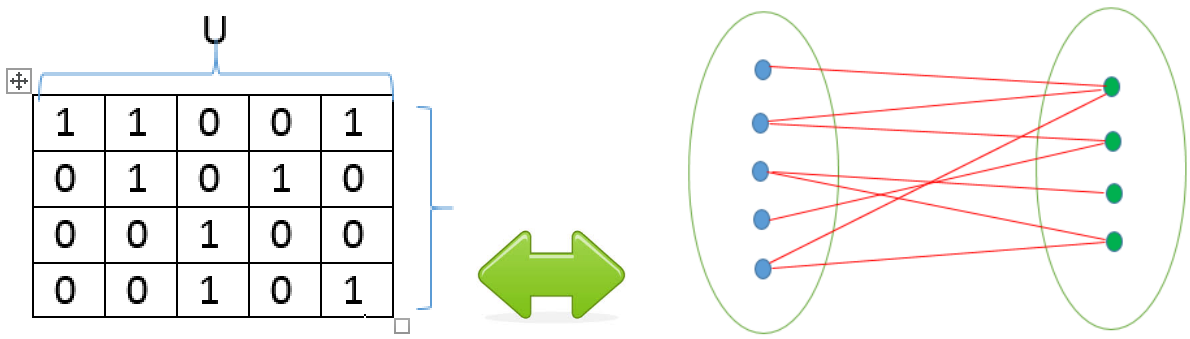
در این بخش، میخواهیم یک ماتریس به دست آوریم که $RIP_1$ بوده و بعلاوه، باینری و تنک نیز است.

**تعریف ۲** (گسترگراف). *گسترگراف نامتوازن* [1] *$(l, \epsilon)$، یک گراف دوبخشی ساده به فرم $G = (U, V, E)$ است که $|U| = n$ ،$|V| = m$ و هر یک از رئوس $U$ از درجه $d$ هستند؛ همچنین، در هر $X \subset U$ که $|X| \le l$ باشد، اندازه مجموعه همسایه‌های $N(X)$ به این صورت است که:*
$|N(X)| \ge (1-\epsilon)d|X|$

ابتدا به این نکته دقت کنید که از روی هر ماتریس مجاورت
$A \in \{0,1\}^{m \times n}$ می‌توان گراف دوبخشی ساده $G = (U, V, E)$ را به این صورت ساخت که:

$$U = [n], V = [m], E = \{(i,j) : i \in U, j \in V, A_{i,j} = 1\}$$

همچنین، عکس این کار نیز امکانپذیر است؛ از روی هر گراف دو بخشی ساده میتوان ماتریس مجاورت متناظر با آن را به همین صورت تشکیل داد.




در تناظر بین گراف و ماتریس شکل فوق، دقت کنید. میتوانیم برای بیان گراف از خلاصه سازی خطی مبتنی بر ماتریس مجاورت استفاده کنیم. بدین صورت که اگر $u$ را یک بردار $n$تایی متشکل از رئوس بخش اول گراف و $v$ را یک بردار $m$تایی در نظر بگیریم، آنگاه میتوانیم ضرب ماتریسی $Au = v$ را به گراف فوق نسبت دهیم که به هر راس از بخش دوم گراف، مجموع رئوس همسایه با آن از بخش اول گراف را نظیر میکند. همچنین، از روی چنین خلاصه سازی خطی با ماتریس باینری میتوان گراف دوبخشی ساده متناظر با ان را به دست آورد.

ماتریس مجاورت متناظر با گسترگراف نامتوازن، در هر ستون دقیقا $d$ تا درایه برابر ۱ دارد و سایر درایه ها صفر هستند. از طرفی چون همسایه های مشترک هر دو راس از مجموعه $U$ کم است، پس درایه های ۱ تا حدی در ماتریس پراکنده هستند و ماتریس مجاورت گراف، تنک و باینری است. حال میخواهیم نشان دهیم که ماتریس مجاورت متناظر با گسترگراف نامتوازن، خاصیت $RIP_1$ نیز دارد.

**قضیه ۳.** *اگر گراف دوبخشی متناظر با ماتریس $A \in \{0,1\}^{m \times n}$، یک گسترگراف $(k, d(1-\frac{\epsilon}{2}))$ باشد، آنگاه به ازای هر بردار $k$ــ تنک $v$ داریم:*

$$d(1-\epsilon)\|v\|_1 \le \|Av\|_1 \le d\|v\|_1$$

*(عکس این قضیه نیز درست است؛ گراف متناظر با یک ماتریس باینری، تنک و $RIP_1$ یک گسترگراف است.)*

[1] unbalanced expander

*اثبات.* (اثبات $\|Av\|_1 \leq d\|v\|_1$) به تناظر بین خلاصه سازی خطی برمبنای یک ماتریس باینری با یک گراف دوبخشی ساده دقت کنید؛ اگر بخش اول گراف را متشکل از $n$ راس متناظر با بردار $v$ و بخش دوم گراف را متشکل از $m$ راس متناظر با $Av$ بگیریم، در این صورت همسایه های هر راس از بخش دوم گراف برمبنای درایه متناظر از $Av$ مشخص میشود. مجموعه $E$ را مجموعه یال های گراف در نظر بگیرید. در این صورت:

$$\|Av\|_1 = \sum_i |(Av)_i| \leq \sum_i |\sum_{j:(i,j)\in E} v_j| \leq \sum_{(i,j)\in E} |v_i| = d\|v\|_1$$

(اثبات $d(1-\epsilon)\|v\|_1 \leq \|Av\|_1$) درایه دلخواه $j$ام از $Av$ را که متناظر با راسی در بخش دوم گراف است، در نظر بگیرید. اگر فقط یک یال به $j$ وارد شود و درایه تنها همسایه آن از بخش اول گراف $i$ باشد، آنگاه $(i,j) \in E$ بوده و تنها یال با همین $j$ در $E$ است؛ یعنی $|(Av)_j| = |v_i|$. در غیر این صورت، $a_j = argmax\{i : i \in N(j)\}$ را در نظر میگیریم. واضح است که

$$|(Av)_j| \geq |v_{a_j}| - \sum_{i\in N(j), i\neq a} |v_i|$$

بنابراین، اگر $r(e)$ برای $e \in E$ را به این صورت تعریف کنیم که اگر $e = (i,j)$ که $i$ بزرگترین اندیس است که از آن به $j$ یال وجود دارد، $r(e) = 1$ و در غیر این صورت $r(e) = -1$، آنگاه خواهیم داشت:

$$|(Av)_j| \geq \sum_{i\in N(j)} r(e)|v_i|$$

در نتیجه،

$$\|Av\|_1 \geq \sum_{(i,j)=e\in E} r(e)|v_i|$$

از طرفی در گسترگراف متناظر با ماتریس $A$، برای هر $i, i' \in \{1, ..., k\}$ داریم: $|N(i) \cap N(i')| \leq 2d\frac{\epsilon}{2} = \epsilon d$ بنابراین،

$$\sum_{(i,j)=e\in E} r(e)|v_i| \geq d\|v\|_1 - \epsilon d|v\|_1 = d(1-\epsilon)\|v\|_1$$

و درنتیجه،

$$\|Av\|_1 \geq d(1-\epsilon)\|v\|_1$$

□

## ۵ کارآیی ماتریس $RIP_1$ در الگوریتم کمینه نرم ۱

در جلسه‌ی گذشته، این نکته ثابت شد که اگر ماتریس $A$ خاصیت فضای پوچ [2] داشته باشد، آنگاه الگوریتم کمینه نرم ۱ تضمین خوبی از خطای تقریب دارد. همچنین، اثبات شد که ماتریس $RIP$ دارای خاصیت فضای پوچ است. اکنون، میخواهیم نشان دهیم که ماتریس $RIP_1$ نیز خاصیت فضای پوچ را دارد.

**ادعا ۴.** *برای هر $1 \leq l \leq \frac{n}{2}$ و $\epsilon > 0$ ، یک گسترگراف نامتوازن ـ $(l,\epsilon)$ وجود دارد که درجه رئوس بخش اول گراف، $d = O(\log(n/l)/\epsilon)$ و تعداد رئوس بخش دوم گراف، $m = O(ld/\epsilon) = O(l\log(n/l)/\epsilon^2)$ است.*

(ادعای فوق به کمک کران چرنوف اثبات میشود.)

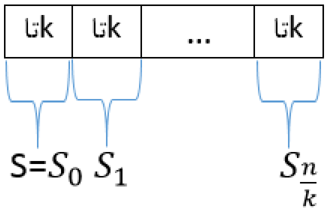
**قضیه ۵.** *فرض کنید $A \in \{0,1\}^{m\times n}$ ماتریس مجاورت متناظر با گسترگراف نامتوازن ـ $(2k,\epsilon)$ بوده که درجه رئوس بخش اول آن، $d$ است. در این صورت، به ازای هر $\eta \in \mathbb{R}^n$ که $A\eta = 0$ و هر زیرمجموعه $k$تایی $S$ از $\eta$ داریم:*

$$\|\eta_S\|_1 \leq \alpha(\epsilon)\|\eta\|_1$$

*که $\alpha(\epsilon) = \frac{2\epsilon}{1-2\epsilon}$ است.*

[2] null-space property

*اثبات.* در جلسه گذشته دیدیم که برای اثبات نامساوی فوق، کافی است درستی آن را در حالتی که $S$ شامل بزرگترین $k$ مقدارهای $\eta$ است، نشان دهیم. بنابراین عناصر $\eta$ را به صورت کاهشی بر حسب مقدار، مرتب میکنیم:




زیرماتریسی از $A$ که شامل سطرهای متناظر با $N(S)$ است را $A'$ می‌نامیم. واضح است که $\|A\eta_s\|_1 = \|A'\eta_s\|_1$.
همچنین طبق اثبات قسمت قبل، گسترگراف با مشخصه $\epsilon$، دارای خاصیت $RIP_1$ با مشخصه $2\epsilon$ است.
ابتدا ایده اثبات را به طور شهودی بیان میکنیم. فرض کنید $\|\eta_S\|_1$ در مقایسه با $\|\eta\|_1$ "خیلی بزرگ" باشد. از آنجایی که $\|A\eta_S\|_1 \geq d(1-2\epsilon)\|\eta_S\|_1$، پس $\|A'\eta_S\|_1$ نیز بایستی مقدار بزرگی داشته باشد. از طرفی، $\|A'\eta\|_1 = \|A'\eta_S + A'\eta_{-S}\|_1 = 0$؛ بنابراین، $\|A'\eta_{-S}\|_1$ نیز برای خنثی کردن اثر بیشتر شدن $\|A'\eta_S\|_1$ بایستی بزرگ بیشتر شود. بدین معنی که تعداد یال های زیادی از $-S$ به $N(S)$ وجود دارد. اما این اتفاق با توجه به گسترگراف بودن گراف موردنظر، یک تناقض است.
اکنون به طور دقیقتر اثبات میکنیم:
برای نمایش مجموعه یال‌های بین مجموعه $X$ و مجموعه $Y$ از نماد $E(X:Y) = E \cap (X \times Y)$ استفاده شده است.
همچنین، دقت کنید که با توجه به خواص گسترگراف، $|N(S \cup S_l)| \geq d(1-\epsilon)|S \cup S_l|$. بنابراین، حداکثر $2k\epsilon d$ یال از $S_l$ که $l \neq 0$، به $N(S)$ وجود دارد.

$$\begin{aligned}
\|A'\eta\|_1 &\geq \|A'\eta_S\|_1 - \sum_{l\geq 1} \sum_{(i,j)\in E, i\in S_l, j\in N(S)} |\eta_i| \\
&\geq d(1-2\epsilon)\|\eta_S\|_1 - \sum_{l\geq 1} |E(S_l : N(S))|.min_{i\in S_{l-1}}\eta_i \\
&\geq d(1-2\epsilon)\|\eta_S\|_1 - \sum_{l\geq 1} |E(S_l : N(S))|.\frac{\|\eta_{S_{l-1}}\|}{k} \\
&\geq d(1-2\epsilon)\|\eta_S\|_1 - 2k\epsilon d \sum_{l\geq 1} \frac{\|\eta_{S_{l-1}}\|}{k} \\
&\geq d(1-2\epsilon)\|\eta_S\|_1 - 2\epsilon d\|\eta\|_1
\end{aligned}$$

از آنجایی که $\|A'\eta\|_1 = 0$ است،

$$d(1-2\epsilon)\|\eta_S\|_1 \leq 2\epsilon d\|\eta\|_1$$

بنابراین،

$$\|\eta_S\|_1 \leq \frac{2\epsilon}{1-2\epsilon}\|\eta\|_1$$

□

# مراجع

[۱] مباحث جلسه هجدهم

[2] The Course of Sketching Algorithms for Big Data, Harvard CS 226/MIT 6.889 - Fall 2017, Lec 15.