

iiiiii HEAD  
=====

```
»»»»> master
```

# VNUHCM - UNIVERSITY OF SCIENCE FACULTY OF INFORMATION TECHNOLOGY

KNOWLEDGE ENGINEERING DEPARTMENT

---

## Report Lab 3

Project 2: INTRODUCTION TO OPENSLL

---

Course: Introduction to Cryptography

*Students:*

Lê Trường Thịnh (23127018)

Trần Lý Nhật Hào (23127187)

*Instructor:*

Trịnh Văn Minh

Ngày 14 tháng 1 năm 2026



# Mục lục

<b>1</b>	<b>Kiến thức cơ sở</b>	<b>1</b>
1.1	Các công cụ (toán) . . . . .	3
1.2	Các thuật toán con . . . . .	4
1.2.1	Chọn phần tử hạng thứ $i$ trong dãy . . . . .	4
<b>2</b>	<b>Fast-Estimation</b>	<b>4</b>
<b>3</b>	<b>Fast-Filtering</b>	<b>17</b>
	<b>References</b>	<b>28</b>

# 1 Kiến thức cơ sở

Gọi  $P \subset \mathbb{R}^d$  và  $k$  lần lượt là tập dữ liệu và số lượng cụm. Gọi  $m$  là kích thước của tập dữ liệu. Đối với hai điểm  $p, q \in \mathbb{R}^d$  bất kỳ, ký hiệu  $\delta(p, q)$  và  $\delta^2(p, q)$  lần lượt là khoảng cách và bình phương khoảng cách giữa chúng. Cho một điểm  $p \in \mathbb{R}^d$  và một tập các tâm  $C = \{c_1, c_2, \dots, c_k\}$ , gọi  $\delta(p, C) = \min_{c \in C} \delta(p, c)$  là khoảng cách từ  $p$  đến tâm gần nhất trong  $C$ .

Để ý có thể có nhiều cách phân cụm tối ưu nhưng ở đây tác giả chọn 1 cách cố định để phân tích.

Gọi  $C^* = \{c_1^*, \dots, c_k^*\}$  và  $P(C^*) = \{P_1^*, \dots, P_k^*\}$  là tập các tâm tối ưu và phân hoạch phân cụm tối ưu tương ứng. Mỗi tâm tối ưu  $c_i^* \in C^*$  được biểu diễn bởi  $d$  tọa độ, tức là  $c_i^* = (c_{i1}^*, c_{i2}^*, \dots, c_{id}^*)$ . Chi phí phân cụm của tập  $P$  đối với tập tâm  $C$  được định nghĩa là:

$$\delta^2(P, C) = \sum_{x \in P} \delta^2(x, C) \quad (1)$$

Cho một tập hợp  $L(P) = \{P_1, P_2, \dots, P_k\}$  đóng vai trò là bộ dự đoán, gọi  $Q_i = P_i \cap P_i^*$  là tập các điểm dữ liệu trong cụm dự đoán  $P_i$  thuộc cụm tối ưu  $P_i^*$ . Gọi tọa độ chiều của các điểm dữ liệu trong  $P_i$  và  $Q_i$  lên chiều thứ  $j$  lần lượt là  $P_{ij}$  và  $Q_{ij}$ . Gọi  $P_{ij}^*$  là tọa độ chiều của các điểm trong  $P_i^*$  lên chiều thứ  $j$ . Gọi  $m_i$  và  $m$  lần lượt là kích thước của  $P_i$  và  $P$ . Với một tập điểm dữ liệu  $V \subset \mathbb{R}^d$ , gọi  $\bar{V}$  là tâm hình học của tập  $V$ . Gọi  $P(j)$  là tọa độ chiều của toàn bộ các điểm trong  $P$  lên chiều thứ  $j$ . Gọi  $\Delta_{max}$  là tỷ lệ chiều tối đa của các điểm dữ liệu được chiếu, được xác định bởi:

$$\Delta_{max} = \max_{1 \leq j \leq d} \frac{\max_{x, y \in P(j)} \delta(x, y)}{\min_{x, y \in P(j), x \neq y} \delta(x, y)} \quad (2)$$

Cụm từ “Aspect Ratio” được tác giả đề cập khi dịch về tiếng Việt để quen thuộc nhất thì chúng em sẽ gọi là **tỷ lệ khung hình**. Về bản chất thì cũng chỉ là tỉ lệ giữa khoảng cách lớn nhất của 2 điểm và khoảng cách nhỏ nhất của 2 điểm.

Với một số nguyên dương  $t$ , gọi  $[t]$  là tập hợp các số nguyên từ 1 đến  $t$ .

**Bài toán  $k$ -means hỗ trợ học:** Cho tập dữ liệu  $P \subset \mathbb{R}^d$  gồm  $m$  điểm, gọi  $C^*$  và  $P(C^*) = \{P_1^*, P_2^*, \dots, P_k^*\}$  lần lượt là một lời giải tối ưu và phân hoạch tương ứng. Trong thiết lập có hỗ trợ học, giả định rằng có quyền truy cập vào một bộ dự đoán dưới dạng phân hoạch nhân  $L(P) = \{P_1, P_2, \dots, P_k\}$  được tham số hóa bởi tỷ lệ lỗi nhân  $\alpha \in [0, 1]$ , thỏa mãn điều kiện  $|P_i \cap P_i^*| \geq (1 - \alpha) \max\{|P_i|, |P_i^*|\}$ . Mục tiêu của bài toán là tìm tập  $C \subset \mathbb{R}^d$  các tâm sao cho  $\delta^2(P, C)$  đạt giá trị nhỏ nhất.

Các hệ quả toán học trong phần này dựa trên tính chất cơ bản của không gian Euclid, trọng tâm  $\bar{V}$  là điểm duy nhất tối thiểu hóa tổng bình phương khoảng cách (SSE) tới mọi điểm trong tập  $V$ . Logic của các bổ đề dưới đây cho phép phân rã chi phí phân cụm thành hai thành phần: chi phí trong cụm (độ liên kết - cohesion) và chi phí do khoảng cách từ tâm dự đoán đến tâm tối ưu.

Các bổ đề dưới đây là "dân gian truyền miệng"(folklore) rất phổ biến liên quan bài toán  $k$ -means clustering

**Lemma 1.** Cho tập  $X \subset \mathbb{R}^d$  có kích thước  $m$  và một điểm dữ liệu bất kỳ  $c \in \mathbb{R}^d$ , ta luôn có:

$$\delta^2(X, c) = \delta^2(X, \bar{X}) + m \cdot \delta^2(c, \bar{X}) \quad (3)$$

[1]

Bổ đề trên xuất phát từ quan sát trọng tâm  $\bar{P}_i$  của các cụm dự đoán không đủ là nghiệm của bài toán. Vì dự đoán không đúng, có thể tồn tại 1 số điểm trong  $P_i$  nằm ngoài  $P_i^*$ . Nếu các điểm trong  $P_i \setminus P_i^*$  nằm **rất xa**  $\bar{P}_i^*$ , trọng tâm cụm dự đoán sẽ bị lệch tùy ý dẫn đến chi phí tăng cụm tăng lên tùy ý.

**Lemma 2.** Cho tập  $J \subset \mathbb{R}$ , gọi  $J_1 \subseteq J$  với  $|J_1| \geq (1 - \zeta)|J|$ , trong đó  $0 \leq \zeta < 1$ . Khi đó, mối liên hệ giữa chi phí của tập con và tập tổng thể được chặn bởi:

$$\delta^2(\bar{J}, \bar{J}_1) \leq \frac{\zeta}{(1 - \zeta)|J|} \delta^2(J, \bar{J}) \quad (4)$$

[3]

Bổ đề trên cũng đúng với  $J \subset \mathbb{R}^d$ , mặc dù tác giả chỉ ghi trên  $\mathbb{R}$ , xem chứng minh ở .

Bổ đề trên xuất phát từ quan sát liên hệ chi phí và kích thước tập con của cụm tối ưu. Ta muốn tìm  $Q_i = P_i \cap P_i^*$  và lấy  $\bar{Q}_i$  làm đáp án cho cụm  $i$ , điều này tự nhiên xuất phát từ dữ kiện  $Q_i$  của bài toán có hỗ trợ học.

$$\begin{aligned} |Q_i| &\geq (1 - \alpha) \max\{|P_i|, |P_i^*|\} \geq (1 - \alpha)|P_i^*| \\ \Rightarrow |P_i^* \setminus Q_i| &\leq \alpha m_i^* \end{aligned}$$

Dùng bổ đề trên, ta có chặn trên chi phí phân cụm:

$$\begin{aligned} \delta^2(P_i^*, \bar{Q}_i) &= \delta^2(P_i^*, \bar{P}_i^*) + m_i^* \delta^2(\bar{P}_i^*, \bar{Q}_i) \quad (\text{bổ đề 1}) \\ &\leq \delta^2(P_i^*, \bar{P}_i^*) + m_i^* \frac{\alpha}{1 - \alpha} \frac{\delta^2(P_i^*, \bar{P}_i^*)}{m_i^*} \quad (\text{bổ đề 2}) \\ &= \left(1 + \frac{\alpha}{1 - \alpha}\right) \delta^2(P_i^*, \bar{P}_i^*) \end{aligned}$$

Như vậy ta có xấp xỉ  $\left(1 + \frac{\alpha}{1 - \alpha}\right)$  cho 1 cụm và cũng như cho bài toán. Cũng chính là chặn dưới và lí do xuất hiện của nó trong tỷ lệ xấp xỉ trong ??.

Khó khăn là  $Q_i$  chưa biết, vì vậy 3 thuật toán chính dưới đây tập trung vào việc loại bỏ các điểm outlier trong  $P_i$ , tìm một trọng tâm gần  $\bar{Q}_i$ , như vậy đồng thời giảm được khoảng cách đến  $\bar{P}_i^*$  và

chi phí.

**Lemma 3.** Cho tập  $X \subset \mathbb{R}^d$  và một giá trị  $\alpha \in (0, 1]$ , gọi  $X' = \arg \min_{X'' \subseteq X, |X''|=\alpha|X|} \delta^2(X'', \overline{X''})$ . Khi đó, ta có:

$$\delta^2(X', \overline{X'}) \leq \alpha \cdot \delta^2(X, \overline{X}) \quad (5)$$

[3]

## 1.1 Các công cụ (toán)

**Background Theorem 1** (Bất đẳng thức tam giác nổi lồi). Với mọi số thực  $a, b \in \mathbb{R}$  và một tham số dương  $\lambda > 0$ , bất đẳng thức sau luôn thỏa mãn:

$$(a + b)^2 \leq \left(1 + \frac{1}{\lambda}\right) a^2 + (1 + \lambda) b^2$$

Trong không gian vector  $\mathbb{R}^d$  với chuẩn Euclid  $\|\cdot\|$ , bất đẳng thức này tương đương với:

$$\|u + v\|^2 \leq \left(1 + \frac{1}{\lambda}\right) \|u\|^2 + (1 + \lambda) \|v\|^2$$

*Chứng minh.* Ở đây nhóm em chứng minh cho 1 chiều, còn lại cũng tương tự. Ta bắt đầu bằng việc khai triển vế trái của bất đẳng thức:

$$(a + b)^2 = a^2 + 2ab + b^2$$

Ta áp dụng bất đẳng thức AM-GM. Với hai số thực dương  $x, y$ , ta luôn có  $x^2 + y^2 \geq 2xy$ . Chọn  $x = \frac{a}{\sqrt{\lambda}}$  và  $y = b\sqrt{\lambda}$ . Khi đó:

$$\left(\frac{a}{\sqrt{\lambda}}\right)^2 + (b\sqrt{\lambda})^2 \geq 2 \left(\frac{a}{\sqrt{\lambda}}\right) (b\sqrt{\lambda})$$

$$\frac{a^2}{\lambda} + \lambda b^2 \geq 2ab$$

Thay thế chặn trên của  $2ab$  vào khai triển ban đầu của  $(a + b)^2$ :

$$\begin{aligned}(a + b)^2 &= a^2 + 2ab + b^2 \\ &\leq a^2 + \left(\frac{a^2}{\lambda} + \lambda b^2\right) + b^2 \\ &= \left(a^2 + \frac{a^2}{\lambda}\right) + (b^2 + \lambda b^2) \\ &= \left(1 + \frac{1}{\lambda}\right) a^2 + (1 + \lambda) b^2\end{aligned}$$

□

## 1.2 Các thuật toán con

### 1.2.1 Chọn phần tử hạng thứ $i$ trong dãy

Hay chọn trung vị trong  $O(m)$ . Thay vì  $O(m \log m)$

## 2 Fast-Estimation

Mặc dù thuật toán Fast-Sampling có thời gian chạy tuyến tính trong khi vẫn duy trì các đảm bảo về mặt xấp xỉ, nhưng vẫn có  $O(\log(kd))$  khi thực hiện chặn hội tụ xác suất, có thể ảnh hưởng trong thực tế của thuật toán khi xử lý các tập dữ liệu quy mô cực lớn. Để giải quyết vấn đề này, trong phần này, tác giả đề xuất một thuật toán dựa trên lấy mẫu nhanh hơn mang tên Fast-Estimation. Thuật toán Fast-Estimation có thể xấp xỉ hiệu quả tọa độ của từng cụm dự đoán trong thời gian chạy tuyến tính, với một sự đánh đổi nhỏ trong các đảm bảo về chất lượng phân cụm.

Ý tưởng chính: trước tiên tạo ra các tọa độ ứng viên có khả năng xấp xỉ chặt chẽ tọa độ của các tâm tối ưu. Sau đó, trong mỗi chiều của từng cụm dự đoán, một bộ ước lượng (estimator) được xây dựng bằng cách lấy mẫu theo phân phối đều. Bộ ước lượng này được thiết kế để cung cấp các ước tính chi phí phân cụm chính xác cho các tập con tọa độ có kích thước  $(1 - \alpha)m_i$ . Cụ thể, đối với mỗi chiều của từng cụm dự đoán, bộ ước lượng được xây dựng bằng cách chọn ngẫu nhiên một tập  $S_{ij}$  từ  $P_{ij}$ . Mỗi tọa độ được lấy mẫu sau đó được gán một trọng số bằng nhau, vì vậy xấp xỉ chi phí phân cụm thông qua các mẫu trọng số thay vì tính toán bộ cụm dự đoán. Với các bộ ước lượng đã xây dựng, việc tìm kiếm tập hợp các tọa độ có chi phí phân cụm tối thiểu có thể được thực hiện trong thời gian hạ tuyến tính (sub-linear), loại bỏ nhân với  $O(\log(kd))$  khỏi thời gian chạy của thuật toán Fast-Sampling.

---

**Thuật toán 1** Fast-Estimation

---

**Đầu vào:** Một bài toán  $k$ -means  $(P, k, d)$ , một tập các phân vùng  $(P_1, P_2, \dots, P_k)$  với tỷ lệ lỗi  $\alpha$ , và tham số  $0 < \epsilon < 0.5$ .

**Đầu ra:** Một tập  $C \subset \mathbb{R}^d$  các tâm với  $|C| = k$ .

```

1: for  $i \in [k]$  do
2:   for  $j \in [d]$  do
3:     Lấy mẫu ngẫu nhiên và độc lập một tập  $U_{ij}$  từ  $P_{ij}$  với kích thước  $O(\log(kd))$ , sau đó
       khởi tạo  $U'_{ij} = \emptyset$  và  $\epsilon_1 = \frac{\epsilon}{126}$ .
4:     for  $q = 0$  to  $O(\log(m\Delta_{max}^2))$  do
5:        $l_{ij} = \sqrt{\frac{2^{q-1}}{(1-\alpha)m_i}}$ .
6:       for  $u \in U_{ij}$  do
7:          $s(u) = \{u + \epsilon_2 \lambda l_{ij} : \lambda \in [-\frac{1}{\epsilon_2}, \frac{1}{\epsilon_2}] \cap \mathbb{Z}\}$ , với  $\epsilon_2 = \sqrt{\frac{\epsilon_1}{32}}$ .
8:          $U'_{ij} = U'_{ij} \cup s(u)$ .
9:     Lấy mẫu ngẫu nhiên và độc lập một tập  $S_{ij}$  từ  $P_{ij}$  với kích thước
        $O\left(\frac{\log(m^3 d \log^3(m\Delta_{max}^2)/\epsilon_1^2) \log(m\Delta_{max}^2)}{\alpha \epsilon_1^4}\right)$ , gán cho mỗi điểm trong  $S_{ij}$  một trọng số  $\frac{m_i}{|S_{ij}|}$ .
10:    Xây dựng bộ ước lượng  $\omega$  sao cho  $\forall u \in U'_{ij}$ ,  $\omega(u) = \sum_{p \in S_{ij} \setminus F(u)} \frac{m_i}{|S_{ij}|} \delta^2(p, u)$ , trong đó
        $F(u)$  là tập hợp  $(1 + 3\epsilon_1)\alpha|S_{ij}|$  điểm xa  $u$  nhất trong  $S_{ij}$ .
11:     $c_{ij} = \arg \min_{u \in U'_{ij}} \omega(u)$ .
12:    Gọi  $I_{ij}$  là tập hợp  $(1 - 2\alpha - \alpha\epsilon)m_i$  tọa độ gần  $c_{ij}$  nhất từ  $P_{ij}$ .
13:     $\hat{c}_i = (I_{ij})_{j \in [d]}$ .
14: return  $\{\hat{c}_1, \hat{c}_2, \dots, \hat{c}_k\}$ .
```

---

**Phân tích thuật toán:**

Trong bước 3, đối với mỗi chiều của cụm dự đoán, thuật toán chọn một mẫu ngẫu nhiên  $U_{ij}$  để xấp xỉ tọa độ của các tâm tối ưu. Theo Lemma 4, với xác suất hằng số, tồn tại ít nhất một tọa độ được lấy mẫu  $u \in U_{ij}$  sao cho  $\delta(u, Q'_{ij}) \leq \sqrt{2\delta^2(Q'_{ij}, \overline{Q'_{ij}})/|Q'_{ij}|}$ . Sau đó, từ bước 4 liệt kê tất cả các độ dài khoảng ứng viên để xây dựng tập hợp các tọa độ ứng viên. Không mất tính tổng quát, có thể giả sử khoảng cách cặp tối thiểu giữa các tọa độ trong  $P_{ij}$  là 1 và khoảng cách cặp tối đa là  $\Delta_{max}$ . Do đó, trong bước 5, tồn tại ít nhất một lần đoán  $q$  cho độ dài thỏa  $\sqrt{\frac{2\delta^2(Q'_{ij}, \overline{Q'_{ij}})}{(1-\alpha)m_i}} \leq l_{ij} \leq \sqrt{\frac{4\delta^2(Q'_{ij}, \overline{Q'_{ij}})}{(1-\alpha)m_i}}$ . Tiếp theo, trong các bước 7-8, theo Lemma 5, cũng tồn tại ít nhất một tọa độ  $u' \in U'_{ij}$  sao cho  $u'$  đủ gần với trọng tâm của  $Q'_{ij}$ , tức là  $\delta(u', Q'_{ij}) \leq \sqrt{\epsilon_1 \delta^2(Q'_{ij}, \overline{Q'_{ij}})/|Q'_{ij}|}$ .

Đối với mỗi  $u \in U'_{ij}$ , gọi  $\mathcal{N}_{ij}(u)$  là tập hợp  $(1 - \alpha)m_i$  tọa độ gần nhất từ  $P_{ij}$  đến  $u$ . Gọi  $O(u) = P_{ij} \setminus \mathcal{N}_{ij}(u)$  là tập hợp  $\alpha m_i$  tọa độ xa nhất từ  $P_{ij}$  đến  $u$ . Trước khi xây dựng bộ ước lượng  $\omega$  (bước 9-10), tác giả bắt đầu bằng cách chia  $\mathcal{N}_{ij}(u)$  thành  $\gamma = \frac{(1+\epsilon_1)\log(m\Delta_{max}^2)}{\epsilon_1}$  khối. Cụ thể, đối với mỗi  $u \in U'_{ij}$ ,  $\mathcal{N}_{ij}(u)$  được phân rã thành  $\gamma$  khối (ký hiệu là  $\mathcal{B}_u^1, \mathcal{B}_u^2, \dots, \mathcal{B}_u^\gamma$ ) dựa trên khoảng cách từ các tọa độ trong  $\mathcal{N}_{ij}(u)$  đến  $u$ , trong đó  $\mathcal{B}_u^l = \{x \in \mathcal{N}_{ij}(u) : (1 + \epsilon_1)^l \leq \delta^2(x, u) < (1 + \epsilon_1)^{l+1}\}$ .

Các "khối" này có thể hình dung là các phần tiếp nối giữa khối cầu có bán kính  $(1 + \epsilon_1)^l$  và  $(1 + \epsilon_1)^{l+1}$

trong  $\mathbb{R}^d$

Sau đó, các khối này được chia tiếp thành hai nhóm dựa trên kích thước:  $\mathcal{L}(u) = \{\mathcal{B}_u^l : |\mathcal{B}_u^l| \geq \frac{\epsilon_1^2 \alpha m_i}{(1+\epsilon_1) \log(m_i \Delta_{max}^2)}, l \in [\gamma]\}$  là nhóm các khối lớn và  $\mathcal{S}(u) = \{\mathcal{B}_u^1, \dots, \mathcal{B}_u^\gamma\} \setminus \mathcal{L}(u)$  là nhóm các khối nhỏ.

### Sự hội tụ xác suất:

Mục tiêu là xấp xỉ tốt từng khối lớn trong  $\mathcal{L}(u)$  đồng thời cho phép bỏ qua các tọa độ trong các khối nhỏ.

1. **Biến ngẫu nhiên:** Đối với mỗi mẫu  $p \in S_{ij}$ , xét biến ngẫu nhiên chỉ thị cho việc  $p$  rơi vào một khối cụ thể.
2. **Áp dụng Bất đẳng thức Chernoff:** Với kích thước mẫu  $|S_{ij}|$  được chọn, kỳ vọng số điểm rơi vào mỗi khối lớn đủ lớn để xác suất sai lệch quá  $\epsilon_1$  lần kỳ vọng bị chặn bởi một hàm mũ âm. Cụ thể,  $Pr(|X - \mathbb{E}[X]| \geq \epsilon_1 \mathbb{E}[X]) \leq 2e^{-\epsilon_1^2 \mathbb{E}[X]/3}$ .
3. **Chặn hội tụ (Union Bound):** Bằng cách lấy tổng xác suất lỗi trên tất cả các khối và các tọa độ ứng viên, tác giả đảm bảo rằng bộ ước lượng  $\omega$  hoạt động chính xác với xác suất cao trên toàn không gian ứng viên.

Với bộ ước lượng đã được chứng minh là hội tụ về giá trị thực, việc tìm  $c_{ij}$  tại bước 11 nhanh hơn vì số lượng ứng viên  $|U'_{ij}|$  chỉ phụ thuộc logarit vào  $\Delta_{max}$  và  $m$ , trong khi việc tính toán mỗi giá trị  $\omega(u)$  chỉ tốn thời gian phụ thuộc vào kích thước mẫu  $|S_{ij}|$  thay vì kích thước toàn bộ dữ liệu  $m_i$ . Cuối cùng, bằng cách sử dụng Lemma 7, Theorem 2 có thể được chứng minh để độ phức tạp thời gian tuyến tính  $O(md) + \tilde{O}(\epsilon^{-5}kd/\alpha)$  cho bài toán có hỗ trợ học.

**Lemma 4.** Giả sử  $S_{ij}$  là một mẫu được lấy ngẫu nhiên từ cụm dự đoán  $P_{ij}$  với kích thước mẫu  $|S_{ij}| = \tilde{O}(1/\alpha \epsilon_1^4)$ . Với xác suất ít nhất  $1 - \frac{\epsilon_1}{m^3 d \log^2(m \Delta_{max}^2)}$ , các bất đẳng thức sau đây đồng thời xảy ra cho mọi khối lớn  $\mathcal{B}_u^l \in \mathcal{L}(u)$  và tập các điểm xa nhất  $\mathcal{O}(u)$ :

$$(1 - \epsilon_1) \mathbb{E}[|\mathcal{B}_u^l \cap S_{ij}|] \leq |\mathcal{B}_u^l \cap S_{ij}| \leq (1 + \epsilon_1) \mathbb{E}[|\mathcal{B}_u^l \cap S_{ij}|]$$

$$(1 - \epsilon_1) \mathbb{E}[|\mathcal{O}(u) \cap S_{ij}|] \leq |\mathcal{O}(u) \cap S_{ij}| \leq (1 + \epsilon_1) \mathbb{E}[|\mathcal{O}(u) \cap S_{ij}|]$$

*Chứng minh.* Chúng ta sẽ phân tích chi tiết cho một khối lớn bất kỳ  $\mathcal{B}_u^l \in \mathcal{L}(u)$ . Quy trình tương tự cũng áp dụng cho tập  $\mathcal{O}(u)$ .

#### 1. Kỳ vọng

Các tọa độ trong  $P_{ij}$  được lấy mẫu độc lập và phân phối đều. Xác suất để một mẫu đơn lẻ rơi vào khối  $\mathcal{B}_u^l$  là tỷ lệ kích thước  $|\mathcal{B}_u^l|/|P_{ij}|$ . Với tập mẫu kích thước  $|S_{ij}|$ , giá trị kỳ vọng số điểm rơi vào khối là:

$$\mathbb{E}[|\mathcal{B}_u^l \cap S_{ij}|] = |S_{ij}| \cdot \frac{|\mathcal{B}_u^l|}{m_i}$$

do tuyến tính của kỳ vọng.

Theo định nghĩa của thuật toán, kích thước mẫu  $|S_{ij}|$  được là:

$$|S_{ij}| = \frac{c \log(m^3 d \log^3(m \Delta_{\max}^2)/\epsilon_1^2) \log(m \Delta_{\max}^2)}{\alpha \epsilon_1^4}$$

trong đó  $c$  là một hằng số đủ lớn. Theo định nghĩa của tập hợp các khối lớn  $\mathcal{L}(u)$ , kích thước của khối  $\mathcal{B}_u^l$  phải thỏa mãn chặn dưới:

$$|\mathcal{B}_u^l| \geq \frac{\epsilon_1^2 \alpha m_i}{(1 + \epsilon_1) \log(m_i \Delta_{\max}^2)}$$

$$\begin{aligned} \mathbb{E}[|\mathcal{B}_u^l \cap S_{ij}|] &= \left( \frac{c \log(m^3 d \dots) \log(m \Delta_{\max}^2)}{\alpha \epsilon_1^4} \right) \cdot \left( \frac{|\mathcal{B}_u^l|}{m_i} \right) \\ &\geq \left( \frac{c \log(m^3 d \dots) \log(m \Delta_{\max}^2)}{\alpha \epsilon_1^4} \right) \cdot \left( \frac{\epsilon_1^2 \alpha m_i}{(1 + \epsilon_1) \log(m_i \Delta_{\max}^2) m_i} \right) \end{aligned}$$

Ta thu được:

$$\mathbb{E}[|\mathcal{B}_u^l \cap S_{ij}|] \geq \frac{c \log(m^3 d \log^3(m \Delta_{\max}^2)/\epsilon_1^2)}{(1 + \epsilon_1) \epsilon_1^2} \quad (6)$$

## 2. Áp dụng Bất đẳng thức Chernoff

- (a) **Bất đẳng thức:** Gọi  $X$  là tổng các biến ngẫu nhiên Bernoulli  $X_1, \dots, X_{m_i}$ ,  $X_i = 1$  nếu điểm  $i$  thuộc  $\mathcal{B}_u^l \cap S_{ij}$ . Áp dụng Bất đẳng thức Chernoff dạng nhân cho tổng các biến Bernoulli độc lập với độ lệch tương đối  $\epsilon_1 \in (0, 1)$ :

$$\Pr(|X - \mathbb{E}[X]| \geq \epsilon_1 \mathbb{E}[X]) \leq 2e^{-\frac{\epsilon_1^2 \mathbb{E}[X]}{3}}$$

- (b) **Thay thế cận dưới của kỳ vọng:**

$$\begin{aligned} \text{Số mũ} &= -\frac{\epsilon_1^2}{3} \cdot \mathbb{E}[X] \\ &\leq -\frac{\epsilon_1^2}{3} \cdot \frac{c \ln \left( \frac{m^3 d \log^3(m \Delta_{\max}^2)}{\epsilon_1^2} \right)}{(1 + \epsilon_1) \epsilon_1^2} \\ &= -\frac{c}{3(1 + \epsilon_1)} \ln \left( \frac{m^3 d \log^3(m \Delta_{\max}^2)}{\epsilon_1^2} \right) \end{aligned}$$

(c) **Biến đổi:** Đặt  $\Lambda = \frac{m^3 d \log^3(m\Delta_{\max}^2)}{\epsilon_1^2}$ . Khi đó, vế phải Chernoff:

$$2e^{-\frac{c}{3(1+\epsilon_1)} \ln(\Lambda)} = 2\Lambda^{-\frac{c}{3(1+\epsilon_1)}}$$

Để đảm bảo xác suất thất bại đủ nhỏ, ta chọn hằng số  $c$  đủ lớn sao cho số mũ  $\frac{c}{3(1+\epsilon_1)} \geq 1$ . Khi đó:

$$\begin{aligned} \Pr(\text{Thất bại tại } \mathcal{B}_u^l) &\leq 2\Lambda^{-\frac{c}{3(1+\epsilon_1)}} \\ &\leq 2\Lambda^{-1} \\ &= 2 \left( \frac{m^3 d \log^3(m\Delta_{\max}^2)}{\epsilon_1^2} \right)^{-1} \\ &= \frac{2\epsilon_1^2}{m^3 d \log^3(m\Delta_{\max}^2)} \end{aligned}$$

$$\Pr(|X - \mathbb{E}[X]| \geq \epsilon_1 \mathbb{E}[X]) \leq O\left(\frac{\epsilon_1^2}{m^3 d \log^3(m\Delta_{\max}^2)}\right)$$

### 3. Chặn Union

Bổ đề yêu cầu bất đẳng thức đúng cho *tất cả* các khối lớn. Số lượng khối lớn  $\gamma$  bị chặn bởi  $O(\log(m\Delta_{\max}^2)/\epsilon_1)$ . Áp dụng Bất đẳng thức Union Bound để tính tổng xác suất thất bại:

$$\begin{aligned} \Pr(\exists \mathcal{B}_u^l \text{ vi phạm}) &\leq \sum_{l=1}^{\gamma} \Pr(\text{Thất bại tại } \mathcal{B}_u^l) \\ &\leq \gamma \cdot O\left(\frac{\epsilon_1^2}{m^3 d \log^3(m\Delta_{\max}^2)}\right) \\ &\leq \frac{\epsilon_1}{m^3 d \log^2(m\Delta_{\max}^2)} \end{aligned}$$

Đối với tập ngoại lai  $\mathcal{O}(u)$ , vì kích thước  $|\mathcal{O}(u)| = \alpha m_i$  lớn hơn kích thước tối thiểu của khối lớn, kết quả tương tự cũng được áp dụng.

□

**Lemma 5.** Gọi  $\mathcal{J}(u)$  là tập hợp các tọa độ nằm trong các khối nhỏ đối với một tọa độ ứng viên  $u$ . Với xác suất ít nhất  $1 - \frac{\epsilon_1}{m^3 d \log^2(m\Delta_{\max}^2)}$ , giao của tập mẫu  $S_{ij}$  và  $\mathcal{J}(u)$  bị chặn như sau:

$$|\mathcal{J}(u) \cap S_{ij}| \leq 2\epsilon_1 \alpha |S_{ij}|$$

*Chứng minh.* Chứng minh này dựa trên việc áp dụng Bất đẳng thức Chernoff để giới hạn độ lệch của biến ngẫu nhiên so với kỳ vọng của nó.

1. Gọi biến ngẫu nhiên  $X = |\mathcal{J}(u) \cap S_{ij}|$ . Vì  $S_{ij}$  được lấy mẫu ngẫu nhiên đều từ  $P_{ij}$ , giá trị kỳ vọng của  $X$  được tính bằng tỷ lệ kích thước:

$$\mathbb{E}[X] = |S_{ij}| \cdot \frac{|\mathcal{J}(u)|}{m_i}$$

2. **Chuẩn bị áp dụng Bất đẳng thức Chernoff** Chúng ta muốn chứng minh rằng  $X$  không vượt quá ngưỡng  $2\epsilon_1\alpha|S_{ij}|$ . Để làm điều này, ta biểu diễn ngưỡng này dưới dạng độ lệch so với kỳ vọng  $(1 + \lambda')\mathbb{E}[X]$ . Ta cần tìm  $\lambda'$  sao cho:

$$(1 + \lambda')\mathbb{E}[X] = 2\epsilon_1\alpha|S_{ij}|$$

Thay thế  $\mathbb{E}[X]$  vào phương trình trên:

$$(1 + \lambda') \left( |S_{ij}| \frac{|\mathcal{J}(u)|}{m_i} \right) = 2\epsilon_1\alpha|S_{ij}|$$

Giải phương trình tìm  $\lambda'$ :

$$1 + \lambda' = \frac{2\epsilon_1\alpha m_i}{|\mathcal{J}(u)|} \Rightarrow \lambda' = \frac{2\epsilon_1\alpha m_i}{|\mathcal{J}(u)|} - 1$$

vì  $|\mathcal{J}(u)| \leq \epsilon_1\alpha m_i$ , ta có tỷ số  $\frac{\epsilon_1\alpha m_i}{|\mathcal{J}(u)|} \geq 1$ , suy ra  $\frac{2\epsilon_1\alpha m_i}{|\mathcal{J}(u)|} \geq 2$ , do đó  $\lambda' \geq 1$ .

### 3. Bất đẳng thức

Đặt biến ngẫu nhiên  $X = |\mathcal{J}(u) \cap S_{ij}|$ . Ta muốn chặn trên xác suất  $X$  vượt quá ngưỡng  $2\epsilon_1\alpha|S_{ij}|$ . Đặt độ lệch  $\lambda' = \frac{2\epsilon_1\alpha m_i}{|\mathcal{J}(u)|} - 1$ . Khi đó, ngưỡng cần chặn chính là  $(1 + \lambda')\mathbb{E}[X]$ .

Áp dụng Bất đẳng thức Chernoff dạng nhân:

$$\Pr(X \geq (1 + \lambda')\mathbb{E}[X]) \leq e^{-\frac{\mathbb{E}[X](\lambda')^2}{3}}$$

Ta xét số mũ  $\mathcal{E} = \frac{\mathbb{E}[X](\lambda')^2}{3}$ . Thay thế  $\mathbb{E}[X] = \frac{|S_{ij}||\mathcal{J}(u)|}{m_i}$  và giá trị của  $\lambda'$ :

$$\mathcal{E} = \frac{|S_{ij}||\mathcal{J}(u)|}{3m_i} \left( \frac{2\epsilon_1\alpha m_i}{|\mathcal{J}(u)|} - 1 \right)^2$$

Để tìm chặn dưới cho số mũ  $\mathcal{E}$ , ta thực hiện biến đổi đại số sau. Đặt  $A = \frac{\epsilon_1\alpha m_i}{|\mathcal{J}(u)|}$ . Theo định nghĩa khối nhỏ,  $|\mathcal{J}(u)| \leq \epsilon_1\alpha m_i$ , suy ra  $A \geq 1$ . Ta có:  $(2A - 1)^2 \geq A^2 \Leftrightarrow A \geq 1$ .

Áp dụng vào biểu thức của  $\mathcal{E}$ :

$$\begin{aligned}\mathcal{E} &\geq \frac{|S_{ij}||\mathcal{J}(u)|}{3m_i} \left( \frac{\epsilon_1 \alpha m_i}{|\mathcal{J}(u)|} \right)^2 \\ &= \frac{|S_{ij}||\mathcal{J}(u)|}{3m_i} \cdot \frac{\epsilon_1^2 \alpha^2 m_i^2}{|\mathcal{J}(u)|^2} \\ &= \frac{\epsilon_1^2 \alpha^2 m_i |S_{ij}|}{3|\mathcal{J}(u)|}\end{aligned}$$

Để  $\mathcal{E}$  nhỏ nhất, ta thay  $|\mathcal{J}(u)|$  bằng giá trị lớn nhất:

$$\mathcal{E} \geq \frac{\epsilon_1^2 \alpha^2 m_i |S_{ij}|}{3(\epsilon_1 \alpha m_i)} = \frac{\epsilon_1 \alpha |S_{ij}|}{3}$$

4. Theo thuật toán, kích thước mẫu  $|S_{ij}|$  được chọn là:

$$|S_{ij}| = \Omega \left( \frac{\log(m^3 d \log^3(m \Delta_{\max}^2) / \epsilon_1^2) \log(m \Delta_{\max}^2)}{\alpha \epsilon_1^4} \right)$$

Thay thế  $|S_{ij}|$  vào chặn dưới của số mũ  $\mathcal{E}$  tìm được ở Bước 3:

$$\mathcal{E} \geq \frac{\epsilon_1 \alpha}{3} \cdot \frac{C \cdot \ln(\dots)}{\alpha \epsilon_1^4} = \frac{C \cdot \ln(\dots)}{3 \epsilon_1^3}$$

Vì  $\epsilon_1 < 1$  và  $C$  là hằng số đủ lớn, ta có:

$$e^{-\mathcal{E}} \leq \frac{\epsilon_1}{m^3 d \log^2(m \Delta_{\max}^2)}$$

Do đó:

$$\Pr(|\mathcal{J}(u) \cap S_{ij}| \geq 2\epsilon_1 \alpha |S_{ij}|) \leq \frac{\epsilon_1}{m^3 d \log^2(m \Delta_{\max}^2)}$$

Lấy phần bù, ta có điều phải chứng minh. □

**Lemma 6.** Cho một tọa độ ứng viên bất kỳ  $u \in U'_{ij}$ . Với xác suất cao (xác suất hằng số), ước lượng  $\omega(u)$  thỏa mãn các chặn sau:

$$\frac{\delta^2(P_{ij} \setminus \mathcal{F}^\dagger(u), u)}{1 + 7\epsilon_1} \leq \omega(u) \leq (1 + \epsilon_1)^2 \delta^2(\mathcal{N}_{ij}(u), u)$$

trong đó:

- $\mathcal{F}^\dagger(u)$  là tập hợp gồm  $(2 + 20\epsilon_1)\alpha m_i$  tọa độ xa nhất từ  $P_{ij}$  đến  $u$ .

- $\mathcal{N}_{ij}(u)$  là tập hợp gồm  $(1 - \alpha)m_i$  tọa độ gần nhất trong  $P_{ij}$  đến  $u$ .

*Chứng minh.* Theo Bổ đề 8 và 9, với xác suất ít nhất  $1 - \frac{\epsilon_1}{m^2 d \log^2(m \Delta_{\max}^2)}$ , các điều kiện sau đây đồng thời xảy ra đối với tập mẫu ngẫu nhiên  $S_{ij}$ :

1. Số lượng phần tử thuộc các khối nhỏ trong mẫu:  $|\mathcal{J}(u) \cap S_{ij}| \leq 2\epsilon_1 \alpha |S_{ij}|$ .
2. Số lượng phần tử ngoại lai trong mẫu:  $|\mathcal{O}(u) \cap S_{ij}| \leq (1 + \epsilon_1) \alpha |S_{ij}|$ .
3. Với mọi khối lớn  $\mathcal{B}_u^l \in \mathcal{L}(u)$ , số lượng phần tử trong mẫu xấp xỉ giá trị kỳ vọng:

$$(1 - \epsilon_1) \frac{|S_{ij}|}{m_i} |\mathcal{B}_u^l| \leq |\mathcal{B}_u^l \cap S_{ij}| \leq (1 + \epsilon_1) \frac{|S_{ij}|}{m_i} |\mathcal{B}_u^l|$$

Chúng ta áp dụng Bất đẳng thức Union Bound để đảm bảo các điều kiện này đúng cho mọi  $u \in U'_{ij}$  với xác suất hằng số.

### 1. Chặn Trên

Mục tiêu là chứng minh  $\omega(u) \leq (1 + \epsilon_1)^2 \delta^2(\mathcal{N}_{ij}(u), u)$ .

Gọi  $\mathcal{F}'(u) = (\mathcal{J}(u) \cup \mathcal{O}(u)) \cap S_{ij}$  là tập hợp các điểm thuộc khối nhỏ và các điểm ngoại lai nằm trong mẫu. Kích thước của tập này bị chặn bởi:

$$|\mathcal{F}'(u)| = |\mathcal{J}(u) \cap S_{ij}| + |\mathcal{O}(u) \cap S_{ij}| \leq 2\epsilon_1 \alpha |S_{ij}| + (1 + \epsilon_1) \alpha |S_{ij}| = (1 + 3\epsilon_1) \alpha |S_{ij}|$$

Theo định nghĩa trong thuật toán,  $\mathcal{F}(u)$  là tập hợp gồm  $(1 + 3\epsilon_1) \alpha |S_{ij}|$  điểm xa nhất từ  $S_{ij}$  đến  $u$ . Do đó,  $|\mathcal{F}(u)| \geq |\mathcal{F}'(u)|$ . Vì  $\omega(u)$  tính tổng chi phí sau khi loại bỏ những điểm xa nhất ( $\mathcal{F}(u)$ ), giá trị này sẽ nhỏ hơn hoặc bằng chi phí khi loại bỏ tập  $\mathcal{F}'(u)$ :

$$\omega(u) = \frac{m_i}{|S_{ij}|} \delta^2(S_{ij} \setminus \mathcal{F}(u), u) \leq \frac{m_i}{|S_{ij}|} \delta^2(S_{ij} \setminus \mathcal{F}'(u), u)$$

Khi loại bỏ  $\mathcal{F}'(u)$ , phần còn lại của mẫu  $S_{ij}$  chỉ chứa các điểm thuộc các khối lớn  $\mathcal{L}(u)$ . Ta có:

$$\delta^2(S_{ij} \setminus \mathcal{F}'(u), u) = \sum_{\mathcal{B}_u^l \in \mathcal{L}(u)} \delta^2(\mathcal{B}_u^l \cap S_{ij}, u)$$

$$\begin{aligned}
\delta^2(\mathcal{B}_u^l \cap S_{ij}, u) &< |\mathcal{B}_u^l \cap S_{ij}| \cdot (1 + \epsilon_1)^{l+1} \\
&\leq \left( (1 + \epsilon_1) \frac{|S_{ij}|}{m_i} |\mathcal{B}_u^l| \right) \cdot (1 + \epsilon_1)^{l+1} \quad (\text{từ Bổ đề 8}) \\
&= \frac{|S_{ij}|}{m_i} (1 + \epsilon_1)^2 (|\mathcal{B}_u^l| (1 + \epsilon_1)^l) \\
&\leq \frac{|S_{ij}|}{m_i} (1 + \epsilon_1)^2 \delta^2(\mathcal{B}_u^l, u)
\end{aligned}$$

$$\omega(u) \leq \frac{m_i}{|S_{ij}|} \sum_{\mathcal{B}_u^l \in \mathcal{L}(u)} \frac{|S_{ij}|}{m_i} (1 + \epsilon_1)^2 \delta^2(\mathcal{B}_u^l, u) \leq (1 + \epsilon_1)^2 \delta^2(\mathcal{N}_{ij}(u), u)$$

## 2. Chặn Dưới

Với mỗi khối lớn  $\mathcal{B}_u^l \in \mathcal{L}(u)$ , gọi  $\mathcal{Z}_u^l = \mathcal{F}(u) \cap \mathcal{B}_u^l$  là các điểm thuộc khối này bị loại bỏ trong mẫu. Gọi  $\mathcal{H}_u^l$  là tập con (tùy ý) trong tập  $\mathcal{B}_u^l$  sao cho:

$$|\mathcal{H}_u^l| = \left\lceil (1 + 3\epsilon_1) \frac{m_i}{|S_{ij}|} |\mathcal{Z}_u^l| \right\rceil$$

Đặt  $\mathcal{F}''(u)$  là tập hợp các điểm "bị loại bỏ" trên toàn bộ dữ liệu, bao gồm các điểm ngoại lai, các khối nhỏ và các phần tỉ lệ từ khối lớn:

$$\mathcal{F}''(u) = \mathcal{O}(u) \cup \mathcal{J}(u) \cup \left( \bigcup_{\mathcal{B}_u^l \in \mathcal{L}(u)} \mathcal{H}_u^l \right)$$

Ta ước tính kích thước của  $\mathcal{F}''(u)$ :

$$\begin{aligned}
|\mathcal{F}''(u)| &\leq |\mathcal{O}(u)| + |\mathcal{J}(u)| + \sum_{\mathcal{B}_u^l} |\mathcal{H}_u^l| \\
&\leq \alpha m_i + \epsilon_1 \alpha m_i + (1 + 3\epsilon_1) \frac{m_i}{|S_{ij}|} \sum_{\mathcal{B}_u^l} |\mathcal{Z}_u^l|
\end{aligned}$$

$\sum |\mathcal{Z}_u^l| \leq |\mathcal{F}(u)| \leq (1 + 3\epsilon_1) \alpha |S_{ij}|$ . Do đó:

$$\begin{aligned}
|\mathcal{F}''(u)| &\leq \alpha m_i (1 + \epsilon_1) + (1 + 3\epsilon_1)^2 \alpha m_i \\
&\leq \alpha m_i (2 + 20\epsilon_1)
\end{aligned}$$

Theo định nghĩa,  $\mathcal{F}^\dagger(u)$  là tập hợp gồm  $(2 + 20\epsilon_1) \alpha m_i$  điểm xa nhất trong  $P_{ij}$ . Do đó, việc loại bỏ  $\mathcal{F}^\dagger(u)$  sẽ làm giảm chi phí nhiều hơn hoặc bằng việc loại bỏ  $\mathcal{F}''(u)$ :

$$\delta^2(P_{ij} \setminus \mathcal{F}^\dagger(u), u) \leq \delta^2(P_{ij} \setminus \mathcal{F}''(u), u)$$

Chi phí còn lại sau khi loại bỏ  $\mathcal{F}''(u)$  là tổng chi phí của các khối lớn sau khi trừ đi  $\mathcal{H}_u^l$ . Sử dụng chặn trên khoảng cách  $(1 + \epsilon_1)^{l+1}$  trong khối  $\mathcal{B}_u^l$ :

$$\delta^2(P_{ij} \setminus \mathcal{F}''(u), u) = \sum_{\mathcal{B}_u^l} \delta^2(\mathcal{B}_u^l \setminus \mathcal{H}_u^l, u) \leq \sum_{\mathcal{B}_u^l} (1 + \epsilon_1)^{l+1} (|\mathcal{B}_u^l| - |\mathcal{H}_u^l|)$$

Từ Bổ đề 8, ta có  $|\mathcal{B}_u^l| \leq \frac{m_i}{|S_{ij}|(1-\epsilon_1)} |\mathcal{B}_u^l \cap S_{ij}|$ . Thay thế vào bất đẳng thức:

$$\begin{aligned} |\mathcal{B}_u^l| - |\mathcal{H}_u^l| &\leq \frac{m_i}{|S_{ij}|(1-\epsilon_1)} |\mathcal{B}_u^l \cap S_{ij}| - (1 + 3\epsilon_1) \frac{m_i}{|S_{ij}|} |\mathcal{Z}_u^l| \\ &= \frac{m_i}{|S_{ij}|} \left( \frac{1}{1-\epsilon_1} |\mathcal{B}_u^l \cap S_{ij}| - (1 + 3\epsilon_1) |\mathcal{Z}_u^l| \right) \end{aligned}$$

Với  $\epsilon_1 < 0.5$ , ta có  $\frac{1}{1-\epsilon_1} \leq 1 + 3\epsilon_1$ .

$$|\mathcal{B}_u^l| - |\mathcal{H}_u^l| \leq (1 + 3\epsilon_1) \frac{m_i}{|S_{ij}|} (|\mathcal{B}_u^l \cap S_{ij}| - |\mathcal{Z}_u^l|)$$

Thay thế trở lại công thức tổng chi phí:

$$\begin{aligned} \delta^2(P_{ij} \setminus \mathcal{F}^\dagger(u), u) &\leq \sum_{\mathcal{B}_u^l} (1 + \epsilon_1)^{l+1} (1 + 3\epsilon_1) \frac{m_i}{|S_{ij}|} (|\mathcal{B}_u^l \cap S_{ij}| - |\mathcal{Z}_u^l|) \\ &= (1 + \epsilon_1)(1 + 3\epsilon_1) \frac{m_i}{|S_{ij}|} \sum_{\mathcal{B}_u^l} (1 + \epsilon_1)^l (|\mathcal{B}_u^l \cap S_{ij}| - |\mathcal{Z}_u^l|) \\ &\leq (1 + 7\epsilon_1) \frac{m_i}{|S_{ij}|} \sum_{\mathcal{B}_u^l} \delta^2((\mathcal{B}_u^l \cap S_{ij}) \setminus \mathcal{Z}_u^l, u) \end{aligned}$$

Do đó:

$$\delta^2(P_{ij} \setminus \mathcal{F}^\dagger(u), u) \leq (1 + 7\epsilon_1) \omega(u)$$

□

**Lemma 7.** Với tập hợp các tọa độ  $I_{ij}$  được xác định bởi thuật toán *Fast-Estimation*, chặn sau đây luôn thỏa mãn:

$$\delta^2(\overline{I_{ij}}, \overline{Q_{ij}}) \leq \frac{13\alpha - 15\alpha^2}{(1 - 3\alpha - \epsilon)(1 - 2\alpha - \epsilon)} \frac{\delta^2(Q_{ij}, \overline{Q_{ij}})}{|Q_{ij}|}$$

*Chứng minh.* Chứng minh được chia thành ba giai đoạn chính: xác định sự tồn tại của ứng viên tốt, giới hạn chi phí của ứng viên được chọn, và sử dụng kỹ thuật cầu nối để giới hạn khoảng cách giữa các tâm.

### 1. Tọa độ ứng viên tốt

Theo Bổ đề 4 và Bổ đề 5, với xác suất hằng số, tồn tại ít nhất một tọa độ  $u_1 \in U'_{ij}$  nằm rất gần

trọng tâm của tập  $Q'_{ij}$  (tập con của  $Q_{ij}$  có chi phí nhỏ nhất với kích thước  $(1 - \alpha)m_i$ ). Cụ thể:

$$\delta^2(u_1, \overline{Q'_{ij}}) \leq \frac{\epsilon_1 \delta^2(Q'_{ij}, \overline{Q'_{ij}})}{|Q'_{ij}|}$$

Sử dụng Bổ đề 1, ta liên hệ chi phí của tập các điểm lân cận  $\mathcal{N}_{ij}(u_1)$  với chi phí tối ưu:

$$\delta^2(\mathcal{N}_{ij}(u_1), u_1) \leq \delta^2(Q'_{ij}, u_1) = \delta^2(Q'_{ij}, \overline{Q'_{ij}}) + |Q'_{ij}| \delta^2(u_1, \overline{Q'_{ij}})$$

Thay thế chặn của  $u_1$  vào, ta có:

$$\delta^2(\mathcal{N}_{ij}(u_1), u_1) \leq (1 + \epsilon_1) \delta^2(Q'_{ij}, \overline{Q'_{ij}})$$

## 2. Giới hạn chi phí của tập được chọn $I_{ij}$

Gọi  $c_{ij}$  là tọa độ được bộ ước lượng  $\omega$  chọn ở Bước 11 của Thuật toán 2. Do  $c_{ij}$  tối thiểu hóa  $\omega$  trên  $U'_{ij}$ , ta có  $\omega(c_{ij}) \leq \omega(u_1)$ . Kết hợp với các chặn của bộ ước lượng từ Bổ đề 10:

$$\frac{\delta^2(P_{ij} \setminus \mathcal{F}^\dagger(c_{ij}), c_{ij})}{1 + 7\epsilon_1} \leq \omega(c_{ij}) \leq \omega(u_1) \leq (1 + \epsilon_1)^2 \delta^2(\mathcal{N}_{ij}(u_1), u_1)$$

Từ đó suy ra chặn trên cho chi phí thực tế của  $c_{ij}$ :

$$\delta^2(I_{ij}, c_{ij}) \leq (1 + 7\epsilon_1) \omega(c_{ij}) \leq (1 + \epsilon_1)^3 (1 + 7\epsilon_1) \delta^2(Q'_{ij}, \overline{Q'_{ij}})$$

Bằng cách chọn  $\epsilon_1 = \epsilon/126$ , và  $\delta^2(I_{ij}, \overline{I_{ij}}) \leq \delta^2(I_{ij}, c_{ij})$ , ta thu được:

$$\delta^2(I_{ij}, \overline{I_{ij}}) \leq (1 + \epsilon/2) \delta^2(Q'_{ij}, \overline{Q'_{ij}})$$

## 3. Giới hạn khoảng cách tâm bằng bắc cầu

Để giới hạn  $\delta^2(\overline{I_{ij}}, \overline{Q_{ij}})$ , ta sử dụng giao tập hợp  $S = I_{ij} \cap Q_{ij}$  làm cầu nối và áp dụng Bất đẳng thức tam giác cho khoảng cách Euclid:

$$\delta(\overline{I_{ij}}, \overline{Q_{ij}}) \leq \delta(\overline{I_{ij}}, \overline{S}) + \delta(\overline{S}, \overline{Q_{ij}})$$

Áp dụng Bổ đề 6 (đã được chứng minh cho Fast-Sampling và mở rộng cho Fast-Estimation), ta có các chặn sau cho từng thành phần khoảng cách:

$$\delta^2(\overline{I_{ij}}, \overline{S}) \leq \frac{(2\alpha + \alpha\epsilon)(1 + \epsilon)}{(1 - 3\alpha - \epsilon)} \frac{|Q'_{ij}|}{|I_{ij}||Q_{ij}|} \delta^2(Q_{ij}, \overline{Q_{ij}})$$

Do  $|I_{ij}| = |Q'_{ij}|$ :

$$\delta^2(\overline{I_{ij}}, \overline{S}) \leq \frac{(2\alpha + \alpha\epsilon)(1 + \epsilon)(1 - \alpha)}{(1 - 3\alpha - \epsilon)(1 - 2\alpha - \epsilon)} \frac{\delta^2(Q_{ij}, \overline{Q_{ij}})}{|Q_{ij}|}$$

Tương tự:

$$\delta^2(\overline{Q_{ij}}, \overline{S}) \leq \frac{2\alpha + \alpha\epsilon}{1 - 3\alpha - \epsilon} \frac{\delta^2(Q_{ij}, \overline{Q_{ij}})}{|Q_{ij}|}$$

Kết hợp lại, bình phương tổng các khoảng cách và thực hiện các phép biến đổi đại số với điều kiện  $\epsilon < 0.5$  và  $\alpha < 1/3$ , ta thu được chặn cuối cùng:

$$\begin{aligned} \delta^2(\overline{I_{ij}}, \overline{Q_{ij}}) &\leq \left( \sqrt{\delta^2(\overline{I_{ij}}, \overline{S})} + \sqrt{\delta^2(\overline{S}, \overline{Q_{ij}})} \right)^2 \\ &\leq \frac{13\alpha - 15\alpha^2}{(1 - 3\alpha - \epsilon)(1 - 2\alpha - \epsilon)} \frac{\delta^2(Q_{ij}, \overline{Q_{ij}})}{|Q_{ij}|} \end{aligned}$$

□

**Theorem 1.** Thuật toán *Fast-Estimation* xấp xỉ  $(1 + O(\alpha))$  cho bài toán *k-means* có hỗ trợ học (learning-augmented) trong thời gian  $O(md) + \tilde{O}(\epsilon^{-5}kd/\alpha)$  với xác suất hằng số, với tỷ lệ lỗi nhãn  $\alpha \in (0, 1/3 - \epsilon)$ .

*Chứng minh.* Chứng minh này đánh giá chất lượng của tập tâm được trả về bởi thuật toán và độ phức tạp của quy trình ước lượng dưới tuyến tính.

## 1. Chất lượng Phân cụm

Giả sử  $C = \{\hat{c}_1, \hat{c}_2, \dots, \hat{c}_k\}$  là tập hợp các tâm được thuật toán trả về, trong đó mỗi tâm  $\hat{c}_i$  được cấu thành từ các tọa độ trên từng chiều  $j$ , ký hiệu là  $c_{ij}$  (trong thuật toán được xác định là  $\overline{I_{ij}}$ ).

Tổng chi phí phân cụm  $\delta^2(P, C)$  có thể được phân rã theo từng cụm tối ưu  $P_i^*$  và từng chiều  $j$ :

$$\delta^2(P, C) \leq \sum_{i=1}^k \sum_{j=1}^d \delta^2(P_{ij}^*, c_{ij})$$

$$\delta^2(P_{ij}^*, c_{ij}) = \delta^2(P_{ij}^*, \overline{P_{ij}^*}) + |P_{ij}^*| \delta^2(\overline{P_{ij}^*}, c_{ij})$$

Dựa vào Bổ đề 7 (được chứng minh dựa trên kết quả của Bổ đề 11 về khoảng cách giữa  $\overline{I_{ij}}$  và  $\overline{Q_{ij}}$ ), ta có chặn trên cho khoảng cách giữa các tâm:

$$\delta^2(\overline{P_{ij}^*}, c_{ij}) \leq \left( \frac{\alpha}{1 - \alpha} + \frac{13\alpha - 15\alpha^2}{(1 - 3\alpha - \epsilon)(1 - 2\alpha - \epsilon)} \right) \frac{\delta^2(P_{ij}^*, \overline{P_{ij}^*})}{|P_{ij}^*|}$$

$$\begin{aligned}\delta^2(P_{ij}^*, c_{ij}) &\leq \delta^2(P_{ij}^*, \overline{P_{ij}^*}) + |P_{ij}^*| \left[ \left( \frac{\alpha}{1-\alpha} + \frac{13\alpha - 15\alpha^2}{(1-3\alpha-\epsilon)(1-2\alpha-\epsilon)} \right) \frac{\delta^2(P_{ij}^*, \overline{P_{ij}^*})}{|P_{ij}^*|} \right] \\ &= \left( 1 + \frac{\alpha}{1-\alpha} + \frac{13\alpha - 15\alpha^2}{(1-3\alpha-\epsilon)(1-2\alpha-\epsilon)} \right) \delta^2(P_{ij}^*, \overline{P_{ij}^*})\end{aligned}$$

Đặt  $\mathcal{K}(\alpha) = \frac{\alpha}{1-\alpha} + \frac{13\alpha-15\alpha^2}{(1-3\alpha-\epsilon)(1-2\alpha-\epsilon)}$ . Vì  $\alpha < 1/3$ ,  $\mathcal{K}(\alpha) = O(\alpha)$ . Lấy tổng trên tất cả các cụm  $i$  và các chiều  $j$ :

$$\delta^2(P, C) \leq (1 + \mathcal{K}(\alpha)) \sum_{i,j} \delta^2(P_{ij}^*, \overline{P_{ij}^*}) = (1 + O(\alpha)) \delta^2(P, C^*)$$

## 2. Xác suất thành công

Sự thành công của Fast-Estimation phụ thuộc vào độ chính xác của bộ ước lượng  $\omega(u)$ . Điều này được đảm bảo bởi Bổ đề 8 và Bổ đề 9 thông qua việc lấy mẫu ngẫu nhiên.

1. **Biến ngẫu nhiên:** Xét quá trình lấy mẫu  $S_{ij}$  từ  $P_{ij}$ . Gọi biến ngẫu nhiên  $X_p$ , bằng 1 nếu điểm  $p \in P_{ij}$  được chọn vào  $S_{ij}$  và 0 nếu ngược lại. Tổng số điểm thuộc một tập con bất kỳ  $A \subseteq P_{ij}$  rơi vào mẫu là  $X = \sum_{p \in A} X_p$ .
2. **Kỳ vọng:**  $\mathbb{E}[X] = \frac{|S_{ij}|}{|P_{ij}|} |A|$ . Thuật toán kích thước mẫu  $|S_{ij}|$  đủ lớn sao cho kỳ vọng số điểm trong các "khối lớn" thỏa mãn  $\mathbb{E}[X] \geq \Omega(\frac{\log m}{\epsilon^2})$ .
3. **Áp dụng Chặn Chernoff:** Để chứng minh độ tập trung của giá trị ước lượng quanh giá trị kỳ vọng, ta sử dụng Bất đẳng thức Chernoff dạng nhân:

$$\Pr(|X - \mathbb{E}[X]| \geq \epsilon_1 \mathbb{E}[X]) \leq 2e^{-\frac{\epsilon_1^2 \mathbb{E}[X]}{3}}$$

Với  $\mathbb{E}[X] \approx \log m$ , số mũ trở thành  $-\Omega(\log m)$ , dẫn đến xác suất sai lệch là nghịch đảo đa thức của  $m$  (xấp xỉ  $1/m^3$ ).

4. **Chặn Union (Union Bound):** Để đảm bảo thuật toán hoạt động đúng trên toàn cục, ta lấy tổng xác suất thất bại trên tất cả các chiều  $d$ , tất cả các cụm  $k$ , và tất cả các ứng viên  $u$ . Do xác suất thất bại cá lẻ rất nhỏ, tổng xác suất thất bại vẫn được giữ ở mức hằng số nhỏ.

## 3. Thời gian chạy

Thời gian chạy của thuật toán được phân tích theo từng giai đoạn xử lý:

- **Ứng viên (Bước 3-7):** Việc lấy mẫu  $U_{ij}$  mất thời gian  $O(\log kd)$ . Thuật toán lặp qua  $O(\log(m\Delta_{\max}))$  giá trị độ dài khoảng, mỗi lần tạo ra tập ứng viên. Tổng số lượng ứng viên được tạo ra là  $O(\epsilon^{-1} \log(m\Delta_{\max}) \log(kd))$ .
- **Ước lượng chi phí (Bước 10):** Kích thước của bộ ước lượng (số lượng mẫu  $S_{ij}$ ) là  $\tilde{O}(\frac{1}{\alpha \epsilon^4})$ .

Việc tính toán  $\omega(u)$  cho tất cả các ứng viên đòi hỏi thời gian tỷ lệ thuận với số lượng ứng viên nhân với kích thước bộ ước lượng. Tổng thời gian cho bước này là:

$$O(\text{số ứng viên}) \times O(\text{kích thước ước lượng}) = \tilde{O}\left(\frac{1}{\alpha\epsilon^5}\right)$$

bước này độc lập với kích thước dữ liệu  $m$  (sublinear).

- **Lựa chọn (Bước 12):** Sau khi chọn được tâm xấp xỉ  $c_{ij}$ , thuật toán cần tìm  $(1 - 2\alpha - \alpha\epsilon)m_i$  điểm lân cận nhất trong  $P_{ij}$ . Sử dụng thuật toán lựa chọn tuyến tính (linear selection algorithm - Blum et al., 1973), bước này mất thời gian  $O(m_i)$  cho mỗi chiều của mỗi cụm.
- Cộng gộp thời gian trên tất cả  $k$  cụm và  $d$  chiều:

$$\sum_{i=1}^k \sum_{j=1}^d O(m_i) + k \cdot d \cdot \tilde{O}\left(\frac{1}{\alpha\epsilon^5}\right) = O(md) + \tilde{O}\left(\frac{kd}{\alpha\epsilon^5}\right)$$

□

### 3 Fast-Filtering

Đối với Fast-Sampling và Fast-Estimation, các tâm được tạo ra bằng cách tìm xấp xỉ tọa độ trong từng chiều không gian. Tuy nhiên, quy trình lấy mẫu này có thể làm phát sinh các sai số tích lũy, dẫn đến sự suy giảm chất lượng phân cụm tổng thể. Trong phần này, dựa trên các thuật toán Fast-Sampling và Fast-Estimation, tác giả đề xuất một thuật toán heuristic thực tiễn hơn mang tên Fast-Filtering nhằm bảo toàn tốt hơn chất lượng phân cụm trong khi vẫn duy trì được thời gian chạy hiệu quả.

Thuật toán đề xuất được trình bày trong Thuật toán 2, với ý tưởng chủ đạo là trực tiếp tìm kiếm các xấp xỉ tâm cho từng cụm dự đoán thay vì xấp xỉ từng chiều độc lập. Tại bước 2, một tập hợp các mẫu được rút ra một cách ngẫu nhiên và độc lập từ mỗi cụm dự đoán để đóng vai trò là các tâm ứng viên. Sau đó, trong các bước 3-4, các bộ ước lượng được xây dựng dựa trên những ý tưởng tương tự từ thuật toán Fast-Estimation. Dựa trên các bộ ước lượng này, tâm ứng viên có chi phí phân cụm tối thiểu được lựa chọn tại bước 5 để xác định các khoảng chứa  $(1 - \alpha)m_i$  điểm gần nhất. Cuối cùng, tại bước 7, các trọng tâm của các tập điểm đã xác định được chọn làm các tâm cuối cùng. Trong Phụ lục A.4, tác giả cung cấp phân tích lý thuyết cho thuật toán Fast-Filtering và chỉ ra rằng, với việc điều chỉnh số lượng lân cận gần nhất cùng kích thước mẫu  $R_1$  và  $R_2$ , thuật toán này có thể đưa ra một nghiệm xấp xỉ  $(1 + O(\sqrt{\alpha}))$ .

#### Giải thích thuật toán:

Thuật toán này giải quyết vấn đề "sai số tích lũy" bằng cách làm trực tiếp trên vecto thay vì gộp

**Thuật toán 2** Fast-Filtering

**Đầu vào:** Bài toán  $k$ -means  $(P, k, d)$ , tập các phân vùng  $(P_1, P_2, \dots, P_k)$  với tỷ lệ lỗi  $\alpha$ , các tham số  $R_1 > 0, R_2 > 0$  và  $0 < \epsilon < 1$ .

**Đầu ra:** Một tập  $C \subset \mathbb{R}^d$  các tâm với  $|C| \leq k$ .

- 1: **for**  $i = 1..k$  **do**
- 2:     Lấy mẫu ngẫu nhiên và độc lập một tập  $U_i$  từ  $P_i$  với kích thước  $R_1$ .
- 3:     Lấy mẫu ngẫu nhiên và độc lập một tập  $S_i$  từ  $P_i$  với kích thước  $R_2$ , và gán cho mỗi điểm trong  $S_i$  một trọng số  $\frac{m_i}{|S_i|}$ .
- 4:     Xây dựng bộ ước lượng  $\omega$  sao cho  $\forall u \in U_i, \omega(u) = \sum_{p \in S_i \setminus F(u)} \frac{m_i}{|S_i|} \delta^2(p, u)$ , trong đó  $F(u)$  là tập hợp  $(1 + \epsilon)\alpha|S_i|$  điểm trong  $S_i$  có khoảng cách xa nhất đối với  $u$ .
- 5:      $c_i = \arg \min_{u \in U_i} \omega(u)$ .
- 6:     Gọi  $I_i$  là tập hợp  $(1 - \alpha)m_i$  điểm trong  $P_i$  gần  $c_i$  nhất.
- 7:      $\hat{c}_i = \bar{I}_i$ .
- return**  $\{\hat{c}_1, \hat{c}_2, \dots, \hat{c}_k\}$ .

kết quả từ  $d$  bài toán đơn chiều.

- **Ước lượng nhanh:** Ý tưởng giống Fast-Estimation. Tại bước 4, thay vì tính toán tổng bình phương khoảng cách  $\delta^2$  trên toàn bộ tập dữ liệu  $P_i$  (vốn tốn thời gian  $O(m_i d)$ ), tác giả sử dụng tập mẫu  $S_i$  có kích thước  $R_2$  nhỏ hơn nhiều. Trọng số  $\frac{m_i}{|S_i|}$  đảm bảo rằng kỳ vọng của bộ ước lượng  $\omega(u)$  sẽ hội tụ về giá trị chi phí thực tế của cụm.
- **Loại bỏ nhiễu (Filtering):** Một đóng góp quan trọng của tác giả là việc định nghĩa tập  $F(u)$  gồm các điểm xa nhất. Trong bài toán có hỗ trợ học, cụm dự đoán  $P_i$  có thể chứa tới  $\alpha m_i$  điểm âm tính giả (nhiều). Nếu các điểm nhiễu này nằm rất xa tâm thực, chúng sẽ kéo trọng tâm lệch khỏi vị trí tối ưu. Việc loại bỏ  $F(u)$  trong quá trình ước lượng giúp "cô lập" ảnh hưởng của các điểm ngoại lệ này, giúp việc chọn  $c_i$  trở nên bền bỉ (robust) hơn.
- **Trọng tâm:** Sau khi đã xác định được một tâm ứng viên tốt  $c_i$ , thuật toán thực hiện một bước tinh chỉnh tại bước 6 và 7. Tập  $I_i$  đại diện cho phần "lỗi" sạch nhất của cụm. Theo Lemma 1, trọng tâm  $\bar{I}_i$  là điểm duy nhất tối thiểu hóa tổng bình phương khoảng cách tới tất cả các điểm trong tập đó. Do đó,  $\hat{c}_i$  chính là nghiệm tối ưu địa phương cho tập điểm đã được lọc nhiễu.

Sự kết hợp giữa lấy mẫu ngẫu nhiên để tìm ứng viên và bộ lọc thống kê để đánh giá chi phí cho phép Fast-Filtering đạt được sự cân bằng giữa tốc độ tính toán và độ chính xác phân cụm.

**Theorem 2.** Cho  $R_1 = O\left(\frac{\log k}{1-2\alpha}\right)$  và  $R_2 = O\left(\frac{\log(m^3 d \log^3(m\Delta^2)/\epsilon^2) \log(m\Delta^2)}{\alpha \epsilon^4}\right)$ , trong đó  $\Delta$  là tỷ lệ chiều của tập dữ liệu. Với xác suất hằng số, Thuật toán 4 (Fast-Filtering) trả về nghiệm xấp xỉ  $(1 + O(\sqrt{\alpha}))$  cho bài toán  $k$ -means có hỗ trợ học trong thời gian  $O(md) + \tilde{O}\left(\frac{kd}{\epsilon^4(1-2\alpha)\alpha}\right)$  với  $\alpha \in (0, 1/3 - \epsilon)$ .

*Chứng minh.* Chứng minh được chia thành ba giai đoạn chính: phân tích thành công của việc lấy mẫu ứng viên, độ tin cậy của bộ ước lượng chi phí, và tổng hợp chi phí phân cụm cuối cùng.

## 1. Xác suất lấy mẫu thành công

Mục tiêu là đảm bảo tập ứng viên  $U_i$  chứa ít nhất một điểm "tốt" nằm gần tâm tối ưu thực sự  $c_i^*$  (ký hiệu là  $\overline{P_i^*}$  trong các phần trước, ở đây ta dùng  $c_i^*$  để đồng nhất với ký hiệu trong bài báo cho Fast-Filtering).

Định nghĩa tập  $G_2(P_i^*) = \{x \in P_i^* : \delta^2(x, c_i^*) \leq 2\delta^2(P_i^*, c_i^*)/|P_i^*|\}$ . Theo Bổ đề 4,  $|P_i \cap P_i^*| \geq (1 - \alpha) \max(|P_i|, |P_i^*|)$ , ta suy ra:

$$|P_i \cap G_2(P_i^*)| \geq |P_i \cap P_i^*| - |P_i^* \setminus G_2(P_i^*)| \geq (1 - \alpha)|P_i^*| - \frac{|P_i^*|}{2} = \left(\frac{1}{2} - \alpha\right) |P_i^*|$$

Tỷ lệ điểm tốt trong  $P_i$  là  $\zeta_i = \frac{|P_i \cap G_2(P_i^*)|}{|P_i|} \geq (1 - \alpha)(\frac{1}{2} - \alpha)$ . Với kích thước mẫu  $R_1 = \Theta(\frac{1}{1-2\alpha} \log(\frac{k}{\eta}))$ , xác suất để tập  $U_i$  chứa ít nhất một điểm tốt  $u_i \in G_2(P_i^*)$  là rất cao.

## 2. Độ tin cậy của Bộ ước lượng

Tiếp theo, ta cần đảm bảo bộ ước lượng  $\omega(u)$  chọn ra được tâm  $c_i$  tốt từ tập  $U_i$ . Do tồn tại  $u_i \in G_2(P_i^*)$  trong tập ứng viên, chi phí của nó bị chặn bởi:

$$\delta^2(H_i(u_i), u_i) \leq \delta^2(Q_i, u_i) \leq 3\delta^2(P_i^*, c_i^*)$$

Vì thuật toán chọn  $c_i$  để tối thiểu hóa  $\omega$ , ta có kết quả quan trọng:

$$\delta^2(P_i \setminus \mathcal{Z}^\dagger(c_i), c_i) \leq 4\delta^2(P_i^*, c_i^*)$$

Điều này đảm bảo rằng tâm được chọn  $c_i$  (và tập hợp sau lọc  $I_i$ ) có chất lượng tốt, làm tiền đề cho Bổ đề 12.

## 3. Tổng chi phí

Ta đánh giá tổng chi phí của giải pháp cuối cùng  $C = \{\overline{I_1}, \dots, \overline{I_k}\}$ . Tổng chi phí là tổng chi phí của từng cụm tối ưu  $P_i^*$  được gán cho tâm tương ứng  $\overline{I_i}$ :

$$\delta^2(P, C) \leq \sum_{i=1}^k \delta^2(P_i^*, \overline{I_i})$$

Sử dụng kết quả trực tiếp từ Bổ đề 14 (Lemma 14), ta có chặn trên cho từng cụm:

$$\delta^2(P_i^*, \overline{I_i}) \leq \left(1 + \frac{O(\sqrt{\alpha})}{(1 - \alpha)(1 - (3 + \epsilon)\alpha)}\right) \delta^2(P_i^*, c_i^*)$$

Lấy tổng trên tất cả  $k$  cụm:

$$\delta^2(P, C) \leq \left(1 + \frac{O(\sqrt{\alpha})}{(1 - \alpha)(1 - (3 + \epsilon)\alpha)}\right) \sum_{i=1}^k \delta^2(P_i^*, c_i^*)$$

Biểu thức trong ngoặc có thể được đơn giản hóa thành  $(1 + O(\sqrt{\alpha}))$  khi  $\alpha$  nhỏ và  $\epsilon$  là hằng số. Vậy thuật toán đạt tỷ lệ xấp xỉ  $(1 + O(\sqrt{\alpha}))$ .

#### 4. Thời gian chạy

- **Lấy mẫu (Bước 2 & 3):** Việc lấy mẫu  $U_i$  và  $S_i$  mất thời gian  $O(1)$  cho mỗi cụm (hoặc phụ thuộc kích thước mẫu nhưng độc lập với  $m$ ).
- **Ước lượng (Bước 4 & 5):** Tính toán  $\omega(u)$  cho tất cả  $u \in U_i$  đòi hỏi tính khoảng cách giữa các cặp điểm trong  $U_i$  và  $S_i$ . Thời gian cho mỗi cụm là  $O(R_1 \cdot R_2 \cdot d)$ . Tổng thời gian ước lượng là:

$$k \cdot O\left(\frac{\log k}{1 - 2\alpha} \cdot \frac{\text{polylog}(m)}{\alpha \epsilon^4} \cdot d\right) = \tilde{O}\left(\frac{kd}{\epsilon^4(1 - 2\alpha)\alpha}\right)$$

- **Lọc và tính tâm (Bước 6 & 7):** Tìm  $(1 - \alpha)m_i$  lân cận gần nhất cho tâm  $c_i$  đã chọn đòi hỏi quét qua  $P_i$ . Sử dụng thuật toán chọn tuyến tính (Linear Selection), bước này mất  $O(m_i d)$ . Tổng thời gian cho  $k$  cụm là  $\sum O(m_i d) = O(md)$ .

Tổng hợp lại, độ phức tạp thời gian là  $O(md) + \tilde{O}\left(\frac{kd}{\epsilon^4(1 - 2\alpha)\alpha}\right)$ . □

**Corollary 2.1.** Cho kích thước mẫu  $R_1 = \Theta\left(\frac{\log k}{1 - 2\alpha}\right)$ . Với mỗi cụm dự đoán  $i \in [k]$ , với xác suất hằng số, tồn tại ít nhất một điểm dữ liệu  $u$  trong tập mẫu  $U_i$  sao cho  $u \in G_2(P_i^*)$ , trong đó  $G_2(P_i^*)$  là tập hợp các điểm nằm gần tâm tối ưu.

*Chứng minh.* Hệ quả này tương tự Corollary 1, chứng minh tương tự.

Gọi  $\zeta_i$  là xác suất chọn được một điểm thuộc  $G_2(P_i^*)$  khi lấy mẫu ngẫu nhiên đều từ  $P_i$ :

$$\begin{aligned} \zeta_i &= \frac{|P_i \cap G_2(P_i^*)|}{|P_i|} \\ &\geq \frac{(\frac{1}{2} - \alpha)|P_i^*|}{|P_i|} \end{aligned}$$

Ta cần chặn trên cho  $|P_i|$ . Từ giả thiết  $|Q_i| \geq (1 - \alpha)|P_i|$  và  $Q_i \subseteq P_i^*$ , ta suy ra  $|P_i| \leq \frac{|P_i^*|}{1 - \alpha}$ .

$$\begin{aligned}\zeta_i &\geq \left(\frac{1}{2} - \alpha\right) \frac{|P_i^*|}{\frac{|P_i^*|}{1-\alpha}} \\ &= \left(\frac{1}{2} - \alpha\right) (1 - \alpha) \\ &= \frac{(1 - 2\alpha)(1 - \alpha)}{2}\end{aligned}$$

Với  $\alpha < 1/2$ , giá trị  $\zeta_i$  luôn dương.

Giả sử ta lấy mẫu độc lập. Xác suất để *tất cả* các mẫu đều không thuộc tập điểm tốt là:

$$\Pr(\text{Thất bại tại cụm } i) = (1 - \zeta_i)^{R_1} \leq e^{-\zeta_i R_1}$$

Để xác suất này nhỏ hơn  $\frac{\eta}{k}$ :

$$e^{-\zeta_i R_1} \leq \frac{\eta}{k} \iff -\zeta_i R_1 \leq \ln\left(\frac{\eta}{k}\right) \iff R_1 \geq \frac{1}{\zeta_i} \ln\left(\frac{k}{\eta}\right)$$

Thay thế chặn dưới của  $\zeta_i$ :

$$\begin{aligned}R_1 &\geq \frac{2}{(1 - 2\alpha)(1 - \alpha)} \ln\left(\frac{k}{\eta}\right) \\ &= \left(\frac{4}{1 - 2\alpha} - \frac{2}{1 - \alpha}\right) \ln\left(\frac{k}{\eta}\right)\end{aligned}$$

Ta chỉ cần lấy mẫu vừa đủ, như vậy phù hợp với  $R_1 = \Theta\left(\frac{\log k}{1 - 2\alpha}\right)$ .

Để đảm bảo thành công trên tất cả  $k$  cụm:

$$\Pr(\exists i \in [k] : U_i \cap G_2(P_i^*) = \emptyset) \leq \sum_{i=1}^k \frac{\eta}{k} = \eta$$

Như vậy, với xác suất ít nhất  $1 - \eta$  (xác suất hằng số), thuật toán tìm được ít nhất một ứng viên tốt cho mọi cụm  $i \in [k]$ .  $\square$

**Corollary 2.2.** *Cho*

$$R_2 = O\left(\frac{\log(m^3 d \log^3(m \Delta^2)/\epsilon_1^2) \log(m \Delta^2)}{\alpha \epsilon_1^4}\right)$$

Với một điểm dữ liệu bất kỳ  $u \in U_i$ , với xác suất cao, bộ ước lượng  $\omega(u)$  thỏa mãn:

$$\frac{\delta^2(P_i \setminus \mathcal{Z}^\dagger(u), u)}{1 + 7\epsilon_1} \leq \omega(u) \leq (1 + \epsilon_1)^2 \delta^2(H_i(u), u)$$

trong đó  $H_i(u)$  là tập hợp  $(1 - \alpha)m_i$  điểm gần  $u$  nhất trong  $P_i$ , và  $\mathcal{Z}^\dagger(u)$  là tập hợp  $(2 + 20\epsilon_1)\alpha m_i$  điểm xa  $u$  nhất.

*Chứng minh.* Hệ quả này tương tự Lemma 10, chứng minh tương tự.

Gọi  $\mathcal{F}'(u)$  là tập hợp các điểm thuộc mẫu  $S_i$  nằm trong các khối nhỏ hoặc là điểm ngoại lai. Ta đã biết  $|\mathcal{F}'(u)| \leq (1 + 3\epsilon_1)\alpha|S_i|$ . Khi tính  $\omega(u)$ , thuật toán loại bỏ một lượng điểm tương ứng, do đó chi phí chỉ còn phụ thuộc vào các khối lớn. Xét tổng chi phí trên mẫu:

$$\omega(u) \leq \frac{m_i}{|S_i|} \sum_{\mathcal{B}_u^l \in \mathcal{L}(u)} \delta^2(\mathcal{B}_u^l \cap S_i, u)$$

Sử dụng tính chất khoảng cách trong khối  $\delta^2(x, u) \leq (1 + \epsilon_1)^{l+1}$  và chặn trên của số lượng mẫu :

$$\begin{aligned} \delta^2(\mathcal{B}_u^l \cap S_i, u) &\leq (1 + \epsilon_1)^{l+1} |\mathcal{B}_u^l \cap S_i| \\ &\leq (1 + \epsilon_1)^{l+1} (1 + \epsilon_1) \frac{|S_i|}{m_i} |\mathcal{B}_u^l| \\ &= \frac{|S_i|}{m_i} (1 + \epsilon_1)^2 ((1 + \epsilon_1)^l |\mathcal{B}_u^l|) \end{aligned}$$

Lưu ý rằng  $(1 + \epsilon_1)^l |\mathcal{B}_u^l| \approx \delta^2(\mathcal{B}_u^l, u)$ . Tổng hợp lại trên các khối lớn (là tập con của  $H_i(u)$ ):

$$\omega(u) \leq (1 + \epsilon_1)^2 \delta^2(H_i(u), u)$$

Ta sử dụng bất đẳng thức đại số: với  $\epsilon_1$  nhỏ,  $\frac{1}{1 - \epsilon_1} \leq 1 + 3\epsilon_1$ . Từ kết quả tập trung ở Bước 1, ta suy ra kích thước thực tế của khối lớn trong  $P_i$ :

$$|\mathcal{B}_u^l| \leq \frac{m_i}{|S_i|(1 - \epsilon_1)} |\mathcal{B}_u^l \cap S_i| \leq (1 + 3\epsilon_1) \frac{m_i}{|S_i|} |\mathcal{B}_u^l \cap S_i|$$

Nhân cả hai vế với bình phương khoảng cách (xấp xỉ  $(1 + \epsilon_1)^l$ ) và lấy tổng trên các khối lớn (lưu ý rằng việc loại bỏ  $\mathcal{Z}^\dagger(u)$  tương ứng với việc giữ lại các khối này):

$$\begin{aligned} \delta^2(P_i \setminus \mathcal{Z}^\dagger(u), u) &\leq \sum (1 + \epsilon_1)^{l+1} (1 + 3\epsilon_1) \frac{m_i}{|S_i|} |\mathcal{B}_u^l \cap S_i| \\ &\leq (1 + 3\epsilon_1)(1 + \epsilon_1) \frac{m_i}{|S_i|} \sum (1 + \epsilon_1)^l |\mathcal{B}_u^l \cap S_i| \\ &\approx (1 + 4\epsilon_1) \omega(u) \end{aligned}$$

Để đảm bảo tính chặt chẽ cho mọi số hạng bậc cao, bài báo sử dụng hệ số an toàn là  $1 + 7\epsilon_1$ :

$$\delta^2(P_i \setminus \mathcal{Z}^\dagger(u), u) \leq (1 + 7\epsilon_1)\omega(u)$$

Sắp xếp lại bất đẳng thức ta thu được chặn dưới cần chứng minh .  $\square$

**Lemma 8.** *Khoảng cách giữa trọng tâm của tập hợp đã lọc  $\overline{I}_i$  và tâm tối ưu  $c_i^*$  bị chặn như sau:*

$$\delta^2(\overline{I}_i, c_i^*) \leq \frac{9\delta^2(P_i^*, c_i^*)}{(1 - (3 + \epsilon)\alpha)m_i}$$

### *Chứng minh.* 1. Kích thước các tập hợp điểm

Theo định nghĩa của quy trình lọc trong thuật toán 2, tập  $I_i$  được tạo thành bằng cách loại bỏ tập  $\mathcal{Z}^\dagger(c_i)$  gồm các điểm xa nhất từ tâm ứng viên  $c_i$ . Kích thước của phần bị loại bỏ là  $|\mathcal{Z}^\dagger(c_i)| = (2 + 20\epsilon_1)\alpha m_i$ . Do đó, kích thước của tập hợp giữ lại là:

$$|I_i| = m_i - (2 + 20\epsilon_1)\alpha m_i = (1 - (2 + 20\epsilon_1)\alpha)m_i$$

Tiếp theo, ta xét phần giao giữa tập đã lọc  $I_i$  và cụm tối ưu  $P_i^*$ . Ta biết rằng số lượng điểm "sai nhãn"(nhiều) trong cụm dự đoán  $P_i$  tối đa là  $|P_i \setminus P_i^*| \leq \alpha m_i$ . Trong trường hợp xấu nhất, toàn bộ các điểm nhiễu này vẫn nằm trong  $I_i$ . Do đó, số lượng điểm thuộc cụm tối ưu thực sự nằm trong  $I_i$  bị chặn dưới bởi:

$$|I_i \cap P_i^*| \geq |I_i| - |P_i \setminus P_i^*|$$

Thay thế kích thước của  $|I_i|$  vào:

$$|I_i \cap P_i^*| \geq (1 - (2 + 20\epsilon_1)\alpha)m_i - \alpha m_i = (1 - (3 + 20\epsilon_1)\alpha)m_i$$

### 2. Kẹp tập đã lọc

Dựa trên Hệ quả 4 (Corollary 4) trong bài báo , tâm  $c_i$  được chọn bởi bộ ước lượng thỏa mãn điều kiện về chi phí với xác suất cao:

$$\delta^2(I_i, c_i) \leq 4\delta^2(P_i^*, c_i^*)$$

Theo tính chất của trọng tâm, tổng bình phương khoảng cách từ các điểm trong một tập hợp đến trọng tâm của nó ( $\overline{I}_i$ ) luôn nhỏ hơn hoặc bằng tổng bình phương khoảng cách đến bất kỳ điểm nào khác ( $c_i$ ). Do đó:

$$\delta^2(I_i, \overline{I}_i) \leq \delta^2(I_i, c_i) \leq 4\delta^2(P_i^*, c_i^*)$$

### 3. Áp dụng Bất đẳng thức tam giác nổi lờng

Để chặn khoảng cách  $\delta^2(\overline{I}_i, c_i^*)$ , ta xét tổng khoảng cách trên các điểm trung gian  $p$  thuộc giao tập

$I_i \cap P_i^*$ . Ta có đẳng thức trung bình:

$$\delta^2(\bar{I}_i, c_i^*) = \frac{1}{|I_i \cap P_i^*|} \sum_{p \in I_i \cap P_i^*} \delta^2(\bar{I}_i, c_i^*)$$

Áp dụng bất đẳng thức tam giác nới lỏng (relaxed triangle inequality) dạng  $(a+b)^2 \leq (1+\frac{1}{\lambda})a^2 + (1+\lambda)b^2$ . Ở đây ta chọn  $\lambda = 2$  để tối ưu hóa các hệ số theo bài báo:

$$\delta^2(\bar{I}_i, c_i^*) \leq (1+0.5)\delta^2(\bar{I}_i, p) + (1+2)\delta^2(p, c_i^*)$$

Thay thế vào công thức tổng:

$$\delta^2(\bar{I}_i, c_i^*) \leq \frac{1}{|I_i \cap P_i^*|} \sum_{p \in I_i \cap P_i^*} [1.5\delta^2(\bar{I}_i, p) + 3\delta^2(p, c_i^*)]$$

Ta thực hiện chặn trên cho từng thành phần của tử số:

- Tổng khoảng cách từ  $p$  đến  $\bar{I}_i$ : Vì  $p \in I_i$ , tổng này nhỏ hơn tổng trên toàn bộ tập  $I_i$ :

$$\sum_{p \in I_i \cap P_i^*} \delta^2(\bar{I}_i, p) \leq \delta^2(I_i, \bar{I}_i)$$

- Tổng khoảng cách từ  $p$  đến  $c_i^*$ : Vì  $p \in P_i^*$ , tổng này nhỏ hơn tổng chi phí của cụm tối ưu:

$$\sum_{p \in I_i \cap P_i^*} \delta^2(p, c_i^*) \leq \delta^2(P_i^*, c_i^*)$$

Thay thế các bất đẳng thức này vào biểu thức chính:

$$\delta^2(\bar{I}_i, c_i^*) \leq \frac{1.5\delta^2(I_i, \bar{I}_i) + 3\delta^2(P_i^*, c_i^*)}{|I_i \cap P_i^*|}$$

Sử dụng kết quả từ Bước 2 ( $\delta^2(I_i, \bar{I}_i) \leq 4\delta^2(P_i^*, c_i^*)$ ) và Bước 1 cho mẫu số:

$$\begin{aligned} \delta^2(\bar{I}_i, c_i^*) &\leq \frac{1.5(4\delta^2(P_i^*, c_i^*)) + 3\delta^2(P_i^*, c_i^*)}{(1 - (3 + 20\epsilon_1)\alpha)m_i} \\ &= \frac{(6 + 3)\delta^2(P_i^*, c_i^*)}{(1 - (3 + 20\epsilon_1)\alpha)m_i} \\ &= \frac{9\delta^2(P_i^*, c_i^*)}{(1 - (3 + 20\epsilon_1)\alpha)m_i} \end{aligned}$$

Cuối cùng, dựa vào điều kiện thiết lập tham số trong thuật toán là  $20\epsilon_1 \leq \epsilon$  (với  $\epsilon_1 = \epsilon/126$ ), ta có chặn cuối cùng:

$$\delta^2(\bar{I}_i, c_i^*) \leq \frac{9\delta^2(P_i^*, c_i^*)}{(1 - (3 + \epsilon)\alpha)m_i}$$

□

**Lemma 9.** *Chi phí phân cụm của tập  $Q_i$  đối với tâm của tập hợp đã lọc  $\bar{I}_i$  thỏa mãn chặn cụ thể sau:*

$$\delta^2(Q_i, \bar{I}_i) \leq \delta^2(Q_i, c_i^*) + \left(5\sqrt{\alpha} + \frac{36(\sqrt{\alpha} + \alpha)}{1 - (3 + \epsilon)\alpha}\right) \delta^2(P_i^*, c_i^*)$$

*Chứng minh.* Chúng ta phân tích sự chênh lệch chi phí bằng cách phân rã tập hợp dựa trên sai số của bộ lọc (Filter).

**1. Phân rã tập hợp và chi phí** Gọi  $A_i = Q_i \setminus I_i$  là tập các điểm thuộc  $Q_i$  bị loại bỏ (âm tính giả của bộ lọc). Gọi  $B_i = I_i \setminus Q_i$  là tập các điểm nhiễu được giữ lại (dương tính giả của bộ lọc).

Ta có đẳng thức phân rã chi phí như sau:

$$\delta^2(Q_i, \bar{I}_i) - \delta^2(Q_i, c_i^*) = \underbrace{[\delta^2(I_i, \bar{I}_i) - \delta^2(I_i, c_i^*)]}_{\leq 0} + [\delta^2(A_i, \bar{I}_i) - \delta^2(A_i, c_i^*)] + [\delta^2(B_i, c_i^*) - \delta^2(B_i, \bar{I}_i)]$$

Vì  $\bar{I}_i$  là trọng tâm của  $I_i$ , số hạng đầu tiên luôn  $\leq 0$ . Ta tập trung chặn hai số hạng còn lại.

**2. Áp dụng bất đẳng thức tam giác nổi lỏng** Sử dụng bất đẳng thức  $\delta^2(x, y) \leq (1 + \lambda)\delta^2(x, z) + (1 + \frac{1}{\lambda})\delta^2(z, y)$  với  $\lambda = \sqrt{\alpha}$ .

Đối với tập  $A_i$  (tương tự cho  $B_i$ ):

$$\begin{aligned} \delta^2(A_i, \bar{I}_i) - \delta^2(A_i, c_i^*) &\leq \sum_{a \in A_i} \left( (1 + \sqrt{\alpha})\delta^2(a, c_i^*) + (1 + \frac{1}{\sqrt{\alpha}})\delta^2(c_i^*, \bar{I}_i) - \delta^2(a, c_i^*) \right) \\ &= \sqrt{\alpha}\delta^2(A_i, c_i^*) + |A_i| \left( 1 + \frac{1}{\sqrt{\alpha}} \right) \delta^2(c_i^*, \bar{I}_i) \end{aligned}$$

Tương tự cho  $B_i$ :

$$\delta^2(B_i, c_i^*) - \delta^2(B_i, \bar{I}_i) \leq \sqrt{\alpha}\delta^2(B_i, \bar{I}_i) + |B_i| \left( 1 + \frac{1}{\sqrt{\alpha}} \right) \delta^2(c_i^*, \bar{I}_i)$$

Tổng hợp lại:

$$\delta^2(Q_i, \bar{I}_i) - \delta^2(Q_i, c_i^*) \leq \sqrt{\alpha}[\delta^2(A_i, c_i^*) + \delta^2(B_i, \bar{I}_i)] \quad (7)$$

$$+ \left( 1 + \frac{1}{\sqrt{\alpha}} \right) (|A_i| + |B_i|) \delta^2(\bar{I}_i, c_i^*) \quad (8)$$

### 3. Tổng hợp

Theo giả thiết bài toán và Bổ đề 12:

- Tổng kích thước sai số:  $|A_i| + |B_i| \leq 3\alpha m_i + \alpha m_i = 4\alpha m_i$ .
- Hệ số khoảng cách tâm:

$$\left(1 + \frac{1}{\sqrt{\alpha}}\right) (|A_i| + |B_i|) \leq \frac{\sqrt{\alpha} + 1}{\sqrt{\alpha}} \cdot 4\alpha m_i = 4\sqrt{\alpha}(1 + \sqrt{\alpha})m_i$$

- Khoảng cách giữa các tâm (từ Lemma 12):  $\delta^2(\bar{I}_i, c_i^*) \leq \frac{9\delta^2(P_i^*, c_i^*)}{(1-(3+\epsilon)\alpha)m_i}$ .
- Chi phí trong cụm:  $\delta^2(A_i, c_i^*) \leq \delta^2(P_i^*, c_i^*)$  và  $\delta^2(B_i, \bar{I}_i) \leq 4\delta^2(P_i^*, c_i^*)$ .

Thay thế các giá trị này vào phương trình Lemma 13:

$$\begin{aligned} \delta^2(Q_i, \bar{I}_i) - \delta^2(Q_i, c_i^*) &\leq \sqrt{\alpha}[\delta^2(P_i^*, c_i^*) + 4\delta^2(P_i^*, c_i^*)] + 4\sqrt{\alpha}(1 + \sqrt{\alpha})m_i \cdot \frac{9\delta^2(P_i^*, c_i^*)}{(1 - (3 + \epsilon)\alpha)m_i} \\ &= 5\sqrt{\alpha}\delta^2(P_i^*, c_i^*) + \frac{36(\sqrt{\alpha} + \alpha)}{1 - (3 + \epsilon)\alpha}\delta^2(P_i^*, c_i^*) \end{aligned}$$

Kết luận:

$$\delta^2(Q_i, \bar{I}_i) \leq \delta^2(Q_i, c_i^*) + \left(5\sqrt{\alpha} + \frac{36(\sqrt{\alpha} + \alpha)}{1 - (3 + \epsilon)\alpha}\right) \delta^2(P_i^*, c_i^*)$$

□

**Lemma 10.** Tổng chi phí phân cụm của cụm tối ưu  $P_i^*$  đối với trọng tâm  $\bar{I}_i$  bị chặn bởi:

$$\delta^2(P_i^*, \bar{I}_i) \leq \left(1 + 6\sqrt{\alpha} + \frac{36(\sqrt{\alpha} + \alpha)}{1 - (3 + \epsilon)\alpha} + \frac{9(\sqrt{\alpha} + \alpha)}{(1 - \alpha)(1 - (3 + \epsilon)\alpha)}\right) \delta^2(P_i^*, c_i^*)$$

Chứng minh.

$$\delta^2(P_i^*, \bar{I}_i) = \delta^2(Q_i, \bar{I}_i) + \delta^2(R_i, \bar{I}_i)$$

1. **Chặn trên cho  $Q_i$**  Sử dụng kết quả từ Bổ đề 13:

$$\delta^2(Q_i, \bar{I}_i) \leq \delta^2(Q_i, c_i^*) + \left(5\sqrt{\alpha} + \frac{36(\sqrt{\alpha} + \alpha)}{1 - (3 + \epsilon)\alpha}\right) \delta^2(P_i^*, c_i^*)$$

2. **Chặn trên cho  $R_i$  (âm tính giả của dự đoán)** Áp dụng bất đẳng thức tam giác nói lỏng

với  $\lambda = \sqrt{\alpha}$  cho mỗi  $p \in R_i$ :

$$\begin{aligned}\delta^2(R_i, \bar{I}_i) &\leq (1 + \sqrt{\alpha})\delta^2(R_i, c_i^*) + |R_i| \left(1 + \frac{1}{\sqrt{\alpha}}\right) \delta^2(c_i^*, \bar{I}_i) \\ &= \delta^2(R_i, c_i^*) + \sqrt{\alpha}\delta^2(R_i, c_i^*) + |R_i| \frac{\sqrt{\alpha} + 1}{\sqrt{\alpha}} \delta^2(c_i^*, \bar{I}_i)\end{aligned}$$

Ta có các chặn kích thước và chi phí:

- $|R_i| \leq \frac{\alpha m_i}{1-\alpha}$  (do điều kiện  $P_i$  chứa ít nhất  $(1-\alpha)$  phần tử của  $P_i^*$ ).
- $\delta^2(R_i, c_i^*) \leq \delta^2(P_i^*, c_i^*)$ .

Thay thế vào biểu thức của  $R_i$ :

$$\begin{aligned}\delta^2(R_i, \bar{I}_i) &\leq (1 + \sqrt{\alpha})\delta^2(R_i, c_i^*) + \left(\frac{\alpha m_i}{1-\alpha}\right) \frac{\sqrt{\alpha} + 1}{\sqrt{\alpha}} \cdot \frac{9\delta^2(P_i^*, c_i^*)}{(1 - (3 + \epsilon)\alpha)m_i} \\ &\leq (1 + \sqrt{\alpha})\delta^2(R_i, c_i^*) + \frac{9(\sqrt{\alpha} + \alpha)}{(1 - \alpha)(1 - (3 + \epsilon)\alpha)} \delta^2(P_i^*, c_i^*)\end{aligned}$$

3. **Tổng hợp** Cộng bước 1 + 2, và  $\delta^2(Q_i, c_i^*) + \delta^2(R_i, c_i^*) = \delta^2(P_i^*, c_i^*)$ .

$$\begin{aligned}\delta^2(P_i^*, \bar{I}_i) &\leq \underbrace{\delta^2(Q_i, c_i^*) + \delta^2(R_i, c_i^*)}_{\delta^2(P_i^*, c_i^*)} + \underbrace{\sqrt{\alpha}\delta^2(R_i, c_i^*)}_{\leq \sqrt{\alpha}\delta^2(P_i^*, c_i^*)} \\ &\quad + \left(5\sqrt{\alpha} + \frac{36(\sqrt{\alpha} + \alpha)}{1 - (3 + \epsilon)\alpha}\right) \delta^2(P_i^*, c_i^*) \\ &\quad + \frac{9(\sqrt{\alpha} + \alpha)}{(1 - \alpha)(1 - (3 + \epsilon)\alpha)} \delta^2(P_i^*, c_i^*)\end{aligned}$$

Gộp các hệ số chứa  $\sqrt{\alpha}$ :  $1\sqrt{\alpha} + 5\sqrt{\alpha} = 6\sqrt{\alpha}$ . Ta thu được bất đẳng thức cuối cùng:

$$\delta^2(P_i^*, \bar{I}_i) \leq \left(1 + 6\sqrt{\alpha} + \frac{36(\sqrt{\alpha} + \alpha)}{1 - (3 + \epsilon)\alpha} + \frac{9(\sqrt{\alpha} + \alpha)}{(1 - \alpha)(1 - (3 + \epsilon)\alpha)}\right) \delta^2(P_i^*, c_i^*)$$

hay:

$$\delta^2(P_i^*, \bar{I}_i) \leq \left(1 + \frac{O(\sqrt{\alpha})}{(1 - \alpha)(1 - (3 + \epsilon)\alpha)}\right) \delta^2(P_i^*, c_i^*)$$

với  $\alpha \in [0, 1)$

□

## 4 Fast-Sampling (k-median)

Trong mục này, tác giả trình bày cách mở rộng các phương pháp dựa trên lấy mẫu được đề xuất cho bài toán  $k$ -median có hỗ trợ học. Thách thức chính ở đây nảy sinh từ sự khác biệt trong các mục tiêu tối ưu hóa. Cụ thể, đối với một tập hợp tọa độ  $S \subset \mathbb{R}^d$  bất kỳ, tâm hình học của  $S$  không còn đóng vai trò là tâm phân cụm tối ưu cho  $S$  theo mục tiêu  $k$ -median, khiến việc xác định các tọa độ hoặc tâm ứng viên chất lượng cao trở nên khó khăn. Kết quả là, các thuật toán  $k$ -median có hỗ trợ học hiện có thường gặp khó khăn trong việc đạt được các đảm bảo xấp xỉ chất lượng cao.

Để vượt qua thách thức này, mục tiêu của tác giả là sử dụng các chiến lược dựa trên lấy mẫu để xây dựng một tập  $U_i$  các tâm nằm gần các tâm phân cụm tối ưu cho mỗi cụm dự đoán  $P_i$ . Sau đó, bằng cách rời rạc hóa lưới, tác giả có thể tạo ra các tâm ứng viên có khả năng xấp xỉ tốt các tâm phân cụm tối ưu. Cuối cùng, bằng cách liệt kê các tâm ứng viên đã xây dựng, tác giả chứng minh rằng chi phí phân cụm của mỗi cụm tối ưu có thể được xấp xỉ tốt bằng cách sử dụng tâm tốt nhất được chọn từ quá trình liệt kê.

Bảng 1: Kết quả so sánh các thuật toán  $k$ -median có hỗ trợ học

Phương pháp và Tài liệu tham khảo	Tỷ lệ xấp xỉ	Khoảng lỗi nhân $\alpha$	Độ phức tạp thời gian
Phân vùng và Sắp xếp [2]	$1 + \tilde{O}((k\alpha)^{1/4})$	Hằng số nhỏ	$O(md \log^3 m + \text{poly}(k, \log m))$
Sắp xếp [3]	$1 + \frac{\alpha(7+10\alpha-10\alpha^2)}{(1-\alpha)(1-2\alpha)}$	$[0, 1/2)$	$O\left(\frac{md \log^3 m \log^2(k/\delta)}{1-2\alpha}\right)$
<b>Fast-Sampling (Tác giả)</b>	$1 + \frac{\alpha(6+4\epsilon-4\alpha-3\epsilon\alpha)}{(1-\alpha)(1-2\alpha)}$	$(0, 1/2)$	$O\left(\frac{md \log(kd) \log(m\Delta)}{1-2\alpha} \cdot \left(\frac{\sqrt{d}}{\epsilon\alpha}\right)^{O(d)}\right)$

Bảng 3 cung cấp một so sánh chi tiết các kết quả cho bài toán  $k$ -median có hỗ trợ học. Tác giả cũng đưa ra một biểu đồ (Hình 2) về tỷ lệ xấp xỉ so với tỷ lệ lỗi  $\alpha$ . Có thể thấy từ bảng rằng kết quả tốt nhất hiện nay đạt được xấp xỉ  $(1 + O(\alpha))$  với  $\alpha \in [0, 1/2)$  [3]. So với các kết quả tiên tiến nhất, thuật toán Fast-Sampling có thể đạt được các đảm bảo chất lượng phân cụm tốt hơn với thời gian chạy kém hơn một chút đối với số chiều  $d$  cố định.

Mô tả cụ thể cho thuật toán  $k$ -median có hỗ trợ học được trình bày trong Thuật toán 5. Ý tưởng chung đằng sau thuật toán là trước tiên tạo ra các tâm ứng viên có thể xấp xỉ chặt chẽ các tâm phân cụm tối ưu cho mỗi cụm dự đoán. Sau đó, bằng cách chọn tâm tốt nhất với chi phí  $k$ -median nhỏ nhất, tác giả chứng minh rằng thuật toán đề xuất có thể đưa ra các đảm bảo xấp xỉ tốt hơn cho bài toán  $k$ -median có hỗ trợ học. Dưới đây, tác giả đưa ra phân tích cụ thể cho thuật toán được đề xuất.

Không mất tính tổng quát, tác giả có thể giả định rằng khoảng cách từng đôi tối thiểu giữa các điểm dữ liệu trong  $P$  là 1 trong khi khoảng cách từng đôi tối đa là  $\Delta$ . Lưu ý rằng điều này có thể thực hiện được bằng các kỹ thuật tỉ lệ chuẩn. Theo Bổ đề 4, trong mỗi bước 2 của Thuật toán 5, với xác suất ít nhất  $1 - 1/k$ , có thể tìm thấy ít nhất một tâm  $u \in U_i$  sao cho  $\delta(u, c_i^*) \leq 2\delta(P_i^*, c_i^*)/|P_i^*| \leq \frac{2\delta(P_i^*, c_i^*)}{(1-\alpha)m_i}$ , trong đó bước cuối cùng tuân theo thực tế là  $|P_i^*| \geq |Q_i| \geq (1-\alpha)m_i$ . Sau đó, trong bước 3 của Thuật toán 5, vì thuật toán liệt kê tất cả các giá trị có thể giữa 1 và

---

**Thuật toán 3** Fast-Sampling ( $k$ -median)

---

**Đầu vào:** Bài toán  $k$ -median  $(P, k, d)$ , tập các phân vùng  $(P_1, \dots, P_k)$  với tỷ lệ lỗi  $\alpha$ , tham số  $\epsilon \in (0, 1]$ .

**Đầu ra:** Tập  $C \subset \mathbb{R}^d$  gồm  $k$  tâm sao cho  $|C| = k$ .

- 1: **for** mỗi  $i \in [k]$  **do**
  - 2:     Lấy mẫu ngẫu nhiên và độc lập tập  $U_i$  từ  $P_i$  với kích thước  $O\left(\frac{\log(kd)}{1-2\alpha}\right)$ , sau đó khởi tạo  $U'_i = \emptyset$ .
  - 3:     **for**  $q = 0$  đến  $O(\log(m\Delta))$  **do**
  - 4:          $l_i = 2^{q-1}/(1-\alpha)m_i$ .
  - 5:         **for** mỗi  $u \in U_i$  **do**
  - 6:             Gọi  $G(u)$  là lưới tâm  $u$  với độ dài cạnh  $2l_i$ .
  - 7:             Phân rã  $G(u)$  thành các lưới con nhỏ hơn với độ dài cạnh  $(1-\alpha)\alpha\epsilon_1 l_i/\sqrt{d}$ , và gọi  $s(u)$  là tập các tâm của các lưới con này, với  $\epsilon_1 < \epsilon/4$ .
  - 8:              $U'_i = U'_i \cup s(u)$ .
  - 9:      $u_i = \arg \min_{u \in U'_i} \delta(\mathcal{N}_i(u), u)$ , trong đó  $\mathcal{N}_i(u)$  là tập  $(1-\alpha)m_i$  điểm trong  $P_i$  gần  $u$  nhất.
  - 10:      $\hat{c}_i = u_i$ .
  - 11: **return**  $\{\hat{c}_1, \hat{c}_2, \dots, \hat{c}_k\}$ .
- 

$\log(m\Delta)$ , tồn tại ít nhất một dự đoán cho bán kính phân cụm (bước 4 của Thuật toán 5) sao cho  $\delta(P_i^*, c_i^*)/(1-\alpha)m_i \leq l_i \leq 2\delta(P_i^*, c_i^*)/(1-\alpha)m_i$ . Do đó, trong bước 6 của Thuật toán 5, lưới có tâm tại  $u$  với độ dài cạnh  $2l_i$  ( $G(u)$ ) sẽ chứa tâm phân cụm tối ưu  $c_i^*$ . Sau đó, trong bước 7 của Thuật toán 5, bằng cách phân rã lưới  $G(u)$  thành các lưới con nhỏ hơn với độ dài cạnh  $(1-\alpha)\alpha\epsilon_1 l_i/\sqrt{d}$  cho một số  $\epsilon_1 < \epsilon/4$ , tâm phân cụm tối ưu  $c_i^*$  cũng phải thuộc về một trong các lưới con. Vì lưới con có độ dài cạnh  $(1-\alpha)\alpha\epsilon_1 l_i/\sqrt{d}$ , cũng tồn tại ít nhất một  $u' \in U'_i$  sao cho  $u'$  đủ gần với  $c_i^*$ , tức là  $\delta(u', c_i^*) \leq (1-\alpha)\alpha\epsilon_1 l_i \leq \alpha\epsilon\delta(P_i^*, c_i^*)/m_i$ . Gọi  $u_i$  là điểm được chọn trong bước 9 của Thuật toán 5. Đối với bất kỳ điểm dữ liệu nào  $u \in U'_i$ , gọi  $\mathcal{N}_i(u)$  là tập hợp các điểm gần nhất  $(1-\alpha)m_i$  trong  $P_i$  tới  $u$ . Do đó, ta có

$$\begin{aligned}
 \delta(\mathcal{N}_i(u_i), u_i) &\leq \delta(\mathcal{N}_i(u'), u') \\
 &\leq \delta(\mathcal{N}_i(u'), c_i^*) + |\mathcal{N}_i(u')| \delta(u', c_i^*) \\
 &\leq \delta(\mathcal{N}_i(u_i), c_i^*) + m_i \cdot \left( \frac{\alpha\epsilon\delta(P_i^*, c_i^*)}{m_i} \right) \\
 &\leq \delta(\mathcal{N}_i(u_i), c_i^*) + \alpha\epsilon\delta(P_i^*, c_i^*)
 \end{aligned}$$

Trong đó:

- Bất đẳng thức đầu tiên tuân theo tính tối tiểu của  $u_i$  trong tập  $U'_i$ .
- Bất đẳng thức thứ hai áp dụng bất đẳng thức tam giác cho từng điểm trong  $\mathcal{N}_i(u')$ .

- Bước cuối cùng sử dụng giới hạn  $|\mathcal{N}_i(u_i)| \leq m_i$  và khoảng cách tâm  $\delta(u', c_i^*)$  đã thiết lập.

**Corollary 2.3.** *Đối với một cụm dự đoán  $P_i$ , với xác suất ít nhất  $1 - 1/k$ , tâm  $u_i$  được chọn bởi thuật toán Fast-Sampling cho mục tiêu  $k$ -median thỏa mãn:*

$$\delta(\mathcal{N}_i(u_i), u_i) \leq \delta(\mathcal{N}_i(u_i), c_i^*) + \alpha \epsilon \delta(P_i^*, c_i^*)$$

trong đó  $\mathcal{N}_i(u_i)$  là tập hợp  $(1 - \alpha)m_i$  điểm gần nhất trong  $P_i$  đến  $u_i$ , và  $c_i^*$  là tâm phân cụm tối ưu cho cụm thứ  $i$ .

*Chứng minh.* Chứng minh dựa trên sự tồn tại của một ứng viên "tốt" trong lưới (grid) và tính chất tối ưu của quy trình lựa chọn tâm.

**1. Sự tồn tại của ứng viên tốt trong lưới** Theo Bước 6 và 7 của Thuật toán 5, không gian xung quanh các mẫu được phân rã thành các lưới con. Với xác suất cao, tồn tại một điểm ứng viên  $u' \in U'_i$  nằm rất gần tâm tối ưu  $c_i^*$ . Cụ thể, dựa trên kích thước lưới con  $(1 - \alpha)\alpha\epsilon_1 l_i / \sqrt{d}$ , khoảng cách này được chặn bởi:

$$\delta(u', c_i^*) \leq \frac{\alpha \epsilon \delta(P_i^*, c_i^*)}{m_i}$$

**2. Tính tối ưu của tâm được chọn ( $u_i$ )** Trong Bước 9 của thuật toán,  $u_i$  được chọn để tối thiểu hóa chi phí  $k$ -median đối với  $(1 - \alpha)m_i$  điểm lân cận nhất của nó. Do đó, với mọi ứng viên khác  $u' \in U'_i$ , ta có:

$$\delta(\mathcal{N}_i(u_i), u_i) \leq \delta(\mathcal{N}_i(u'), u')$$

**3. Áp dụng Bất đẳng thức Tam giác** Ta cần chặn trên vế phải  $\delta(\mathcal{N}_i(u'), u')$ .

Chọn tập so sánh là  $\mathcal{N}_i(u_i)$  (tập lân cận của  $u_i$ ), ta có:

$$\delta(\mathcal{N}_i(u'), u') \leq \delta(\mathcal{N}_i(u_i), u')$$

$$\delta(x, u') \leq \delta(x, c_i^*) + \delta(c_i^*, u')$$

Lấy tổng trên toàn bộ tập  $\mathcal{N}_i(u_i)$ :

$$\begin{aligned} \delta(\mathcal{N}_i(u_i), u') &= \sum_{x \in \mathcal{N}_i(u_i)} \delta(x, u') \\ &\leq \sum_{x \in \mathcal{N}_i(u_i)} \delta(x, c_i^*) + \sum_{x \in \mathcal{N}_i(u_i)} \delta(c_i^*, u') \\ &= \delta(\mathcal{N}_i(u_i), c_i^*) + |\mathcal{N}_i(u_i)| \cdot \delta(u', c_i^*) \end{aligned}$$

#### 4. Tổng hợp

Kết hợp các bất đẳng thức từ Bước 2 và Bước 3:

$$\delta(\mathcal{N}_i(u_i), u_i) \leq \delta(\mathcal{N}_i(u_i), c_i^*) + |\mathcal{N}_i(u_i)| \cdot \delta(u', c_i^*)$$

$$\begin{aligned} \delta(\mathcal{N}_i(u_i), u_i) &\leq \delta(\mathcal{N}_i(u_i), c_i^*) + m_i \left( \frac{\alpha \epsilon \delta(P_i^*, c_i^*)}{m_i} \right) \\ &= \delta(\mathcal{N}_i(u_i), c_i^*) + \alpha \epsilon \delta(P_i^*, c_i^*) \end{aligned}$$

□

**Lemma 11.** *Đối với hàm mục tiêu k-median, khoảng cách giữa tâm thuật toán  $u_i$  và tâm phân cụm tối ưu  $c_i^*$  thỏa mãn:*

$$\delta(u_i, c_i^*) \leq \frac{(2 + \alpha \epsilon) \delta(P_i^*, c_i^*)}{(1 - 2\alpha)m_i}$$

*Chứng minh.* 1. **Kích thước tập giao**

Thuật toán chọn tập  $\mathcal{N}_i(u_i)$  gồm  $(1 - \alpha)m_i$  điểm gần  $u_i$  nhất trong cụm dự đoán  $P_i$ . Theo giả thiết của mô hình, số lượng điểm trong  $P_i$  không thuộc cụm tối ưu  $P_i^*$  (dương tính giả) tối đa là  $\alpha m_i$  (vì  $|P_i \setminus P_i^*| \leq \alpha m_i$ ). Do đó, số lượng điểm thuộc  $P_i^*$  nằm trong tập  $\mathcal{N}_i(u_i)$  bị chặn dưới bởi:

$$|\mathcal{N}_i(u_i) \cap P_i^*| \geq |\mathcal{N}_i(u_i)| - |P_i \setminus P_i^*|$$

Thay thế các giá trị kích thước vào:

$$|\mathcal{N}_i(u_i) \cap P_i^*| \geq (1 - \alpha)m_i - \alpha m_i = (1 - 2\alpha)m_i$$

Điều này đảm bảo rằng phần giao chứa một lượng điểm đáng kể .

## 2. Chặn trên cho chi phí

Dựa vào tính chất tối ưu trong Bước 9 của Thuật toán 5,  $u_i$  là điểm tối thiểu hóa chi phí k-median đối với tập lân cận của nó. Theo Hệ quả 5 (Corollary 5) đã chứng minh trước đó, với xác suất cao, chi phí này bị chặn bởi:

$$\delta(\mathcal{N}_i(u_i), u_i) \leq (1 + \alpha \epsilon) \delta(P_i^*, c_i^*)$$

## 3. Bất đẳng thức tam giác

Áp dụng bất đẳng thức tam giác cho mọi  $p$ :

$$\delta(u_i, c_i^*) \leq \delta(p, u_i) + \delta(p, c_i^*)$$

Lấy tổng trên tất cả các điểm  $p$  thuộc phần giao  $\mathcal{N}_i(u_i) \cap P_i^*$  và chia cho kích thước của phần giao này:

$$\delta(u_i, c_i^*) \leq \frac{\sum_{p \in \mathcal{N}_i(u_i) \cap P_i^*} \delta(p, u_i) + \sum_{p \in \mathcal{N}_i(u_i) \cap P_i^*} \delta(p, c_i^*)}{|\mathcal{N}_i(u_i) \cap P_i^*|}$$

Ta thực hiện chặn trên cho tử số:

- Tổng thứ nhất  $\sum \delta(p, u_i)$  nhỏ hơn hoặc bằng tổng chi phí của toàn bộ tập  $\mathcal{N}_i(u_i)$  đối với  $u_i$ :

$$\sum_{p \in \mathcal{N}_i(u_i) \cap P_i^*} \delta(p, u_i) \leq \delta(\mathcal{N}_i(u_i), u_i) \leq (1 + \alpha\epsilon)\delta(P_i^*, c_i^*)$$

- Tổng thứ hai  $\sum \delta(p, c_i^*)$  nhỏ hơn hoặc bằng tổng chi phí của toàn bộ cụm tối ưu  $P_i^*$ :

$$\sum_{p \in \mathcal{N}_i(u_i) \cap P_i^*} \delta(p, c_i^*) \leq \delta(P_i^*, c_i^*)$$

Thay thế các chặn trên vào tử số và sử dụng chặn dưới của mẫu số từ Bước 1:

$$\begin{aligned} \delta(u_i, c_i^*) &\leq \frac{(1 + \alpha\epsilon)\delta(P_i^*, c_i^*) + \delta(P_i^*, c_i^*)}{(1 - 2\alpha)m_i} \\ &= \frac{(2 + \alpha\epsilon)\delta(P_i^*, c_i^*)}{(1 - 2\alpha)m_i} \end{aligned}$$

□

**Lemma 12.** *Đối với hàm mục tiêu  $k$ -median, chi phí phân cụm của tập  $Q_i$  đối với tâm được chọn  $u_i$  thỏa mãn chặn sau:*

$$\delta(Q_i, u_i) \leq \delta(Q_i, c_i^*) + \frac{\alpha(4 + 3\epsilon)}{1 - 2\alpha} \delta(P_i^*, c_i^*)$$

trong đó  $c_i^*$  là tâm tối ưu của cụm thứ  $i$ .

*Chứng minh.* Chứng minh được thực hiện bằng cách phân rã sự chênh lệch chi phí giữa tâm thuật toán  $u_i$  và tâm tối ưu  $c_i^*$  thành các thành phần dựa trên tập hợp con.

### 1. Phân rã tập hợp và chênh lệch chi phí

Ta có  $Q_i = P_i \cap P_i^*$ . Dựa trên tập lân cận  $\mathcal{N}_i(u_i)$  được thuật toán chọn, ta định nghĩa các tập sai số:

- $A_i = Q_i \cap \mathcal{Z}^\dagger(u_i)$ : Tập các điểm thuộc  $Q_i$  nhưng bị loại bỏ (âm tính giả).
- $B_i = P_i \setminus (Q_i \cup \mathcal{Z}^\dagger(u_i))$ : Tập các điểm không thuộc  $Q_i$  nhưng được chọn (dương tính giả).

Khi đó  $Q_i = (\mathcal{N}_i(u_i) \setminus B_i) \cup A_i$ . Hiệu chi phí phân cụm được viết lại như sau:

$$\begin{aligned}\delta(Q_i, u_i) - \delta(Q_i, c_i^*) &= [\delta(\mathcal{N}_i(u_i), u_i) - \delta(\mathcal{N}_i(u_i), c_i^*)] \\ &\quad + [\delta(B_i, c_i^*) - \delta(B_i, u_i)] \\ &\quad + [\delta(A_i, u_i) - \delta(A_i, c_i^*)]\end{aligned}$$

## 2. Từng thành phần

Thành phần 1 (Tập được chọn): Theo Hệ quả 5 (Corollary 5), với xác suất cao, chi phí của tập được chọn thỏa mãn:

$$\delta(\mathcal{N}_i(u_i), u_i) - \delta(\mathcal{N}_i(u_i), c_i^*) \leq \alpha \epsilon \delta(P_i^*, c_i^*)$$

Thành phần 2 (Dương tính giả  $B_i$ ): Sử dụng bất đẳng thức tam giác  $\delta(b, c_i^*) \leq \delta(b, u_i) + \delta(u_i, c_i^*)$ . Suy ra  $\delta(b, c_i^*) - \delta(b, u_i) \leq \delta(u_i, c_i^*)$ . Lấy tổng trên  $B_i$ :

$$\delta(B_i, c_i^*) - \delta(B_i, u_i) \leq |B_i| \delta(u_i, c_i^*)$$

Thành phần 3 (Âm tính giả  $A_i$ ): Tương tự, với  $a \in A_i$ , ta có  $\delta(a, u_i) \leq \delta(a, c_i^*) + \delta(c_i^*, u_i)$ . Suy ra  $\delta(a, u_i) - \delta(a, c_i^*) \leq \delta(u_i, c_i^*)$ . Lấy tổng trên  $A_i$ :

$$\delta(A_i, u_i) - \delta(A_i, c_i^*) \leq |A_i| \delta(u_i, c_i^*)$$

## 3. Tập sai số

Theo phân tích kích thước trong bài báo :

$$|Q_i| = |\mathcal{N}_i(u_i)| + |A_i| - |B_i| = (1 - \alpha)m_i + |A_i| - |B_i|$$

Mặt khác  $|Q_i| \geq (1 - \alpha)m_i$  (do định nghĩa mô hình). Suy ra  $|A_i| \geq |B_i|$ . Đồng thời,  $A_i \subseteq \mathcal{Z}^\dagger(u_i)$  (tập các điểm bị loại bỏ), nên  $|A_i| \leq \alpha m_i$ . Do đó, tổng kích thước sai số bị chặn bởi:

$$|A_i| + |B_i| \leq 2|A_i| \leq 2\alpha m_i$$

## 4. Tổng hợp

Thay thế các chặn trên vào phương trình hiệu chi phí:

$$\delta(Q_i, u_i) - \delta(Q_i, c_i^*) \leq \alpha \epsilon \delta(P_i^*, c_i^*) + (|A_i| + |B_i|) \delta(u_i, c_i^*)$$

Thay thế  $|A_i| + |B_i| \leq 2\alpha m_i$  và khoảng cách tâm  $\delta(u_i, c_i^*)$  từ Bổ đề 15:

$$\delta(u_i, c_i^*) \leq \frac{(2 + \alpha\epsilon)\delta(P_i^*, c_i^*)}{(1 - 2\alpha)m_i}$$

Ta có:

$$\begin{aligned} \Delta_{cost} &\leq \alpha\epsilon\delta(P_i^*, c_i^*) + 2\alpha m_i \left( \frac{(2 + \alpha\epsilon)\delta(P_i^*, c_i^*)}{(1 - 2\alpha)m_i} \right) \\ &= \delta(P_i^*, c_i^*) \left[ \alpha\epsilon + \frac{2\alpha(2 + \alpha\epsilon)}{1 - 2\alpha} \right] \end{aligned}$$

Quy đồng mẫu số cho biểu thức trong ngoặc vuông:

$$\begin{aligned} \alpha\epsilon + \frac{4\alpha + 2\alpha^2\epsilon}{1 - 2\alpha} &= \frac{\alpha\epsilon(1 - 2\alpha) + 4\alpha + 2\alpha^2\epsilon}{1 - 2\alpha} \\ &= \frac{\alpha\epsilon - 2\alpha^2\epsilon + 4\alpha + 2\alpha^2\epsilon}{1 - 2\alpha} \\ &= \frac{4\alpha + \alpha\epsilon}{1 - 2\alpha} \\ &= \frac{\alpha(4 + \epsilon)}{1 - 2\alpha} \end{aligned}$$

Nhận thấy rằng  $\frac{\alpha(4+\epsilon)}{1-2\alpha} < \frac{\alpha(4+3\epsilon)}{1-2\alpha}$  (với  $\epsilon > 0$ ). Để đảm bảo tính tổng quát và khớp với chặn đã công bố trong Bổ đề, ta sử dụng chặn rộng hơn:

$$\delta(Q_i, u_i) \leq \delta(Q_i, c_i^*) + \frac{\alpha(4 + 3\epsilon)}{1 - 2\alpha} \delta(P_i^*, c_i^*)$$

□

**Lemma 13.** Với mỗi cụm  $i \in [k]$ , với xác suất ít nhất  $1 - 1/k$ , chi phí phân cụm  $k$ -median của cụm tối ưu  $P_i^*$  đối với tâm thuật toán  $u_i$  bị chặn bởi:

$$\delta(P_i^*, u_i) \leq \left( 1 + \frac{6\alpha + 4\alpha\epsilon - 4\alpha^2 - 3\epsilon\alpha^2}{(1 - \alpha)(1 - 2\alpha)} \right) \delta(P_i^*, c_i^*)$$

trong đó  $c_i^*$  là tâm tối ưu của cụm thứ  $i$ .

*Chứng minh.* Chúng ta phân tích tổng chi phí bằng cách chia cụm tối ưu  $P_i^*$  thành hai phần rời nhau dựa trên kết quả dự đoán của mô hình: phần giao với cụm dự đoán  $(P_i^* \cap P_i)$  và phần bị dự đoán sai  $(P_i^* \setminus P_i)$ .

$$\delta(P_i^*, u_i) = \delta(P_i^* \cap P_i, u_i) + \delta(P_i^* \setminus P_i, u_i)$$

### 1. Giới hạn chi phí của phần giao (dương tính thật)

Xét tập hợp  $Q_i = P_i^* \cap P_i$ . Áp dụng trực tiếp kết quả từ Bổ đề 16, ta có chặn trên cho chi phí của tập này đối với tâm  $u_i$ :

$$\delta(P_i^* \cap P_i, u_i) \leq \delta(P_i^* \cap P_i, c_i^*) + \frac{\alpha(4 + 3\epsilon)}{1 - 2\alpha} \delta(P_i^*, c_i^*)$$

## 2. Giới hạn chi phí của phần sai lệch (âm tính giả)

Với các điểm  $p \in P_i^* \setminus P_i$  (các điểm thuộc cụm tối ưu nhưng không nằm trong cụm dự đoán  $P_i$ ), ta sử dụng bất đẳng thức tam giác qua tâm tối ưu  $c_i^*$ :

$$\delta(p, u_i) \leq \delta(p, c_i^*) + \delta(c_i^*, u_i)$$

Lấy tổng trên toàn bộ tập  $P_i^* \setminus P_i$ :

$$\delta(P_i^* \setminus P_i, u_i) \leq \delta(P_i^* \setminus P_i, c_i^*) + |P_i^* \setminus P_i| \delta(c_i^*, u_i)$$

## 3. Khoảng cách tâm

Ta có

$$\begin{aligned} |P_i^*| &\geq |P_i \cap P_i^*| \geq (1 - \alpha)|P_i^*| \\ \Rightarrow |P_i^*| &\leq \frac{|P_i|}{1 - \alpha} \end{aligned}$$

$$|P_i^* \setminus P_i| \leq \alpha |P_i^*| \quad \text{hoặc theo } m_i : \quad |P_i^* \setminus P_i| \leq \frac{\alpha m_i}{1 - \alpha}$$

Sử dụng kết quả từ Bổ đề 15 cho khoảng cách giữa hai tâm  $\delta(u_i, c_i^*) \leq \frac{(2 + \alpha\epsilon)\delta(P_i^*, c_i^*)}{(1 - 2\alpha)m_i}$ . Thay thế vào biểu thức ở Bước 2:

$$\begin{aligned} \text{Sai số dịch chuyển} &= |P_i^* \setminus P_i| \delta(c_i^*, u_i) \\ &\leq \left( \frac{\alpha m_i}{1 - \alpha} \right) \left( \frac{(2 + \alpha\epsilon)\delta(P_i^*, c_i^*)}{(1 - 2\alpha)m_i} \right) \\ &= \frac{\alpha(2 + \alpha\epsilon)}{(1 - \alpha)(1 - 2\alpha)} \delta(P_i^*, c_i^*) \end{aligned}$$

## 4. Tổng hợp

Cộng gộp kết quả từ Bước 1 và Bước 3:

$$\begin{aligned}\delta(P_i^*, u_i) &\leq \underbrace{\delta(P_i^* \cap P_i, c_i^*) + \delta(P_i^* \setminus P_i, c_i^*)}_{\delta(P_i^*, c_i^*)} \\ &\quad + \left[ \frac{\alpha(4 + 3\epsilon)}{1 - 2\alpha} + \frac{\alpha(2 + \alpha\epsilon)}{(1 - \alpha)(1 - 2\alpha)} \right] \delta(P_i^*, c_i^*)\end{aligned}$$

$$\begin{aligned}\text{Hệ số vế phải} &= \frac{\alpha(4 + 3\epsilon)(1 - \alpha) + \alpha(2 + \alpha\epsilon)}{(1 - \alpha)(1 - 2\alpha)} \\ &= \frac{(4\alpha + 3\alpha\epsilon - 4\alpha^2 - 3\alpha^2\epsilon) + (2\alpha + \alpha^2\epsilon)}{(1 - \alpha)(1 - 2\alpha)} \\ &= \frac{6\alpha - 4\alpha^2 + 3\alpha\epsilon - 2\alpha^2\epsilon}{(1 - \alpha)(1 - 2\alpha)}\end{aligned}$$

Ta nhận thấy rằng  $3\alpha\epsilon - 2\alpha^2\epsilon = \alpha\epsilon(3 - 2\alpha)$ . Biểu thức trong Bỏ đề yêu cầu là  $4\alpha\epsilon - 3\alpha^2\epsilon = \alpha\epsilon(4 - 3\alpha)$ . Vì  $\alpha < 1$ , ta có  $3\alpha\epsilon - 2\alpha^2\epsilon \leq 4\alpha\epsilon - 3\alpha^2\epsilon$  (do hiệu số là  $\alpha\epsilon(1 - \alpha) > 0$ ). Do đó, ta có thể nói lỏng chặn trên để khớp với công thức tổng quát của Bỏ đề:

$$\delta(P_i^*, u_i) \leq \left( 1 + \frac{6\alpha + 4\alpha\epsilon - 4\alpha^2 - 3\epsilon\alpha^2}{(1 - \alpha)(1 - 2\alpha)} \right) \delta(P_i^*, c_i^*)$$

□

## Tài liệu

- [1] David Arthur and Sergei Vassilvitskii.  $k$ -means++: The advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.
- [2] Thy Dinh Nguyen, Anamay Chaturvedi, and Huy Nguyen. Improved learning-augmented algorithms for  $k$ -means and  $k$ -medians clustering. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.