بسم الله الرحمن الرحیم

خلاصه‌سازی برای مه‌داده

ترم پاییز ۱۳۹۹-۱۴۰۰

بسم الله الرحمن الرحیم

جلسه بیست و سوم

خلاصه‌سازی برای مه‌داده

خلاصه‌سازی ضرب‌های داخلی

توصیف مسئله

Definition 89. *An* ε-inner-product sketch *for points* $X \subseteq \mathbb{R}^k$, $|X| = n$, *such that* $\forall x : ||x||_2 \leq 1$, *is a data structure that enables one to compute the inner-product*

$$\langle x, x' \rangle \ \forall x, x' \in X$$

up to additive error ε.

توصیف مسئله

Definition 89. *An* ε-inner-product sketch *for points* $X \subseteq \mathbb{R}^k$, $|X| = n$, *such that* $\forall x : ||x||_2 \leq 1$, *is a data structure that enables one to compute the inner-product*

$$\langle x, x' \rangle \ \ \forall x, x' \in X$$

up to additive error ε.

(تعریف)

$f(n,k,\varepsilon)$ =: کم‌ترین تعداد بیت

توصیف مسئله

Definition 89. *An* ε-inner-product sketch *for points* $X \subseteq \mathbb{R}^k$, $|X| = n$, *such that* $\forall x : ||x||_2 \leq 1$, *is a data structure that enables one to compute the inner-product*

$$\langle x, x' \rangle \ \forall x, x' \in X$$

up to additive error ε.

(تعریف)

$f(n, k, \varepsilon)$:= کم‌ترین تعداد بیت

قبلا (JL):

ـ کاهش بعد: حفظ اندازه

ـ کاهش بعد: حفظ فاصله

دست‌گرمی: f(n,n,e)

(تعریف)

$f(n, k, \varepsilon)$:= کم‌ترین تعداد بیت

Definition 89. *An* ε-inner-product sketch *for points* $X \subseteq \mathbb{R}^k$, $|X| = n$, *such that* $\forall x : ||x||_2 \leq 1$, *is a data structure that enables one to compute the inner-product*

$$\langle x, x' \rangle \ \forall x, x' \in X$$

up to additive error ε.

دست‌گرمی: f(n,n,e)

(تعریف)

$f(n,k,\varepsilon)$ =: کم‌ترین تعداد بیت

Definition 89. *An* ε-inner-product sketch *for points* $X \subseteq \mathbb{R}^k$, $|X| = n$, *such that* $\forall x : ||x||_2 \leq 1$, *is a data structure that enables one to compute the inner-product*

$$\langle x, x' \rangle \ \forall x, x' \in X$$

up to additive error ε.

$$f(n,n,\varepsilon) \leq \ ?$$

دست‌گرمی: f(n,n,e)

(تعریف)

$f(n,k,\varepsilon)$ =: کم‌ترین تعداد بیت

Definition 89. *An* ε-inner-product sketch *for points* $X \subseteq \mathbb{R}^k$, $|X| = n$, *such that* $\forall x : ||x||_2 \leq 1$, *is a data structure that enables one to compute the inner-product*

$$\langle x, x' \rangle \ \forall x, x' \in X$$

up to additive error ε.

$$f(n,n,\varepsilon) \leq \ ?$$

ایده اول:

کاهش ابعاد

دست‌گرمی: f(n,n,e)

(تعریف)

$f(n,k,\varepsilon)$ =: کم‌ترین تعداد بیت

Definition 89. *An* ε-inner-product sketch *for points* $X \subseteq \mathbb{R}^k$, $|X| = n$, *such that* $\forall x : ||x||_2 \leq 1$, *is a data structure that enables one to compute the inner-product*

$$\langle x, x' \rangle \;\; \forall x, x' \in X$$

up to additive error ε.

$$f(n,n,\varepsilon) \leq \;?$$

ایده اول:

کاهش ابعاد

۱- JL به $m := O(1/\epsilon^2 \log n)$ بعد

۲ـ گرد کردن یکی از نقاط گرید با ضلع $1/\epsilon$

دست‌گرمی: f(n,n,e)

(تعریف)

$f(n,k,\varepsilon)$ =: کم‌ترین تعداد بیت

Definition 89. *An* ε-inner-product sketch *for points* $X \subseteq \mathbb{R}^k$, $|X| = n$, *such that* $\forall x : ||x||_2 \leq 1$, *is a data structure that enables one to compute the inner-product*

$$\langle x, x' \rangle \;\; \forall x, x' \in X$$

up to additive error ε.

ایده اول:
کاهش ابعاد

$$f(n,n,\varepsilon) \leq \; ?$$

تعداد نقاط گرید

۱- JL به $m := O(1/\epsilon^2 \log n)$ بعد

۲ـ گرد کردن یکی از نقاط گرید با ضلع $1/\epsilon$

تعداد بیت =$n \cdot \log((1/\epsilon)^{O(m)})$

دست‌گرمی: f(n,n,e)

(تعریف)

$f(n,k,\varepsilon)$:= کم‌ترین تعداد بیت

Definition 89. *An* ε-inner-product sketch *for points $X \subseteq \mathbb{R}^k$, $|X| = n$, such that $\forall x : ||x||_2 \leq 1$, is a data structure that enables one to compute the inner-product*

$$\langle x, x' \rangle \ \forall x, x' \in X$$

up to additive error ε.

ایده اول:
کاهش ابعاد

$$f(n,n,\varepsilon) \leq \ ?$$

۱– JL به $m := O(1/\epsilon^2 \log n)$ بعد
۲ـ گرد کردن یکی از نقاط گرید با ضلع $1/\epsilon$

تعداد بیت $= n \cdot \log((1/\epsilon)^{O(m)})$

تعداد نقاط گرید

$= O(n/\epsilon^2 \log n \log(1/\epsilon))$

دست‌گرمی: f(n,n,e)

(تعریف)

$f(n,k,\varepsilon)$ =: کم‌ترین تعداد بیت

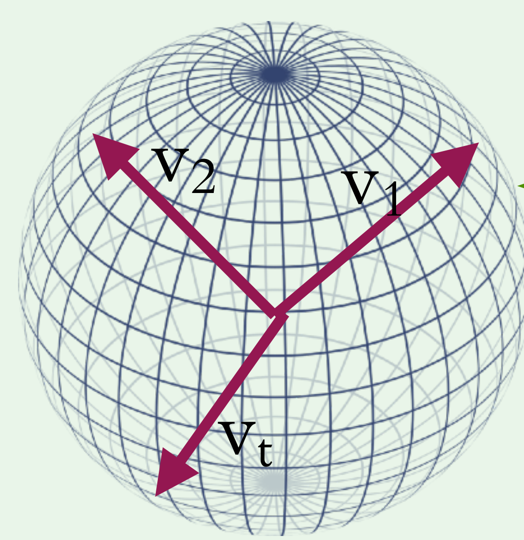
Definition 89. *An* ε-inner-product sketch *for points* $X \subseteq \mathbb{R}^k$, $|X| = n$, *such that* $\forall x : ||x||_2 \leq 1$, *is a data structure that enables one to compute the inner-product*

$$\langle x, x' \rangle \;\; \forall x, x' \in X$$

up to additive error ε.

$$f(n,n,\varepsilon) \leq \;?$$

ایده دوم:



$t := O(1/\epsilon^2 \log n)$

بردار تصادفی در

R^n

دست‌گرمی: f(n,n,e)

(تعریف)

$f(n,k,\varepsilon)$:= کم‌ترین تعداد بیت

Definition 89. *An* ε-inner-product sketch *for points* $X \subseteq \mathbb{R}^k$, $|X| = n$, *such that* $\forall x : ||x||_2 \leq 1$, *is a data structure that enables one to compute the inner-product*

$$\langle x, x' \rangle \ \ \forall x, x' \in X$$

up to additive error ε.

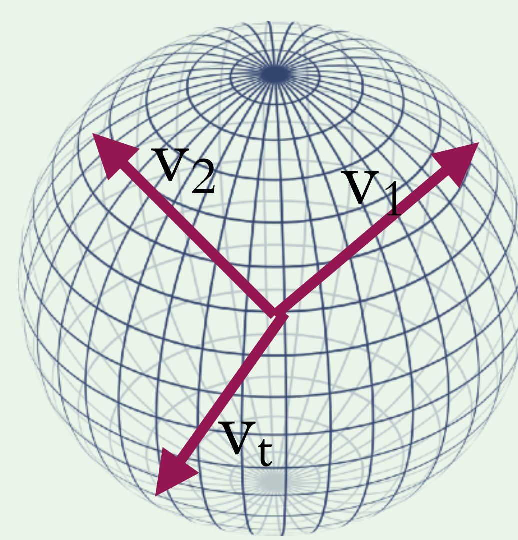
$$f(n,n,\varepsilon) \leq \ ?$$

نگه‌داری

$$\{\text{sign}(\langle x, v_i \rangle)\}_{i \in [t]}$$

به جای x_i، مقدار

ایده دوم:



$t := O(1/\epsilon^2 \log n)$

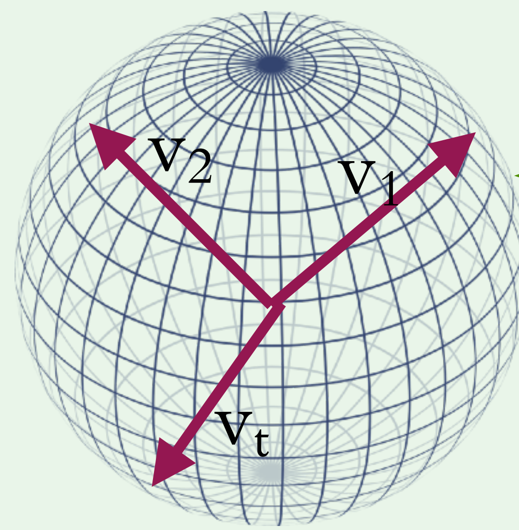
بردار تصادفی در

R^n

دست‌گرمی: f(n,n,e)

تعداد بیت = $O(n/\epsilon^2 \log n)$

ایده دوم:



$t := O(1/\epsilon^2 \log n)$

بردار تصادفی یکه در

R^m

$f(n, n, \varepsilon) \leq \ ?$

نگه‌داری

$\{\text{sign}(\langle x, v_i \rangle)\}_{i \in [t]}$

به جای x_i، مقدار

ایده اثبات:

* برای v تصادفی:
  * احتمال $sign(\langle x, v_i \rangle) \neq sign(\langle x', v_i \rangle)$ متناسب با زاویه x و x’
* ضرب داخلی = وارون کسینوس زاویه

قضیه اصلی

Theorem 90 ([Alon17]). *Let $f(n,k,\varepsilon)$ be the minimum size (in bits) of an ε-inner-product sketch for n points in dimension k.*

Then, $\forall n \geq k, \varepsilon \geq \frac{1}{n^{0.49}}$*:*

(A) *For* $\frac{\log n}{\varepsilon^2} \leq k \leq n : f(n,k,\varepsilon) = \Theta(\frac{n \log n}{\varepsilon^2})$

(B) *For* $\log n \leq k \leq \frac{\log n}{\varepsilon^2} : f(n,k,\varepsilon) = \Theta(nk \log(2 + \frac{\log n}{\varepsilon^2 k}))$

(C) *For* $1 \leq k \leq \log n : f(n,k,\varepsilon) = \Theta(nk \log(1/\varepsilon))$.

اثبات قسمت (A)

- اول ـ JL به $m := \frac{\log n}{\varepsilon^2}$

اثبات قسمت (A)

- اول‌ـ JL به $m := \frac{\log n}{\varepsilon^2}$

- **تعریف**: ماتریس گرام (برای $w_1, \ldots, w_n$):

$$G_{i,j} := \langle w_i, w_j \rangle$$

اثبات قسمت (A)

- اول_ JL به $m := \frac{\log n}{\varepsilon^2}$

- **تعریف**: ماتریس گرام (برای $w_1, \ldots, w_n$):

$$G_{i,j} := \langle w_i, w_j \rangle$$

- **تعریف**: دو ماتریس گرام ϵ_ متفاوت:

$$\exists i \neq j : \quad |G_{i,j} - G'_{i,j}| > \varepsilon$$

اثبات قسمت (A)

- اول‌ـ JL به $m := \frac{\log n}{\varepsilon^2}$

- **تعریف**: ماتریس گرام (برای $w_1, \ldots, w_n$):

$$G_{i,j} := \langle w_i, w_j \rangle$$

- **تعریف**: دو ماتریس گرام ϵـ متفاوت:

$$\exists i \neq j : \quad |G_{i,j} - G'_{i,j}| > \varepsilon$$

- **تعریف** $\mathcal{G}$: مجموعه‌ای بیشینه از eـ متفاوت‌ها

- <= $f(n, k, \varepsilon) \leq \log |\mathcal{G}|$

★ یک پدیده تصادفی:

★ انتخاب $v_1, \ldots, v_n$ در R^m در گوی با شعاع ۲

- ★ یک پدیده تصادفی:
  - ★ انتخاب $v_1, \ldots, v_n$ در R^m در گوی با شعاع ۲
- ★ رخداد

$$A_G := \{\forall i, j : |\langle v_i, v_j\rangle - \langle w_i, w_j\rangle| \leq \varepsilon/2\}$$

★ یک پدیده تصادفی:

★ انتخاب $v_1, \ldots, v_n$ در R^m درگوی با شعاع ۲

★ رخداد

$$A_G := \{\forall i, j : |\langle v_i, v_j \rangle - \langle w_i, w_j \rangle| \leq \varepsilon/2\}$$

★ رخدادهای A_G مجزا هستند.

$$\sum_{G \in \mathcal{G}} \mathbb{P}[A_G] \leq 1$$

★ یک پدیده تصادفی:

★ انتخاب $v_1, \ldots, v_n$ در R^m در گوی با شعاع ۲

★ رخداد

$$A_G := \{\forall i, j : |\langle v_i, v_j \rangle - \langle w_i, w_j \rangle| \leq \varepsilon/2\}$$

★ رخدادهای A_G مجزا هستند.

$$\sum_{G \in \mathcal{G}} \mathbb{P}[A_G] \leq 1$$

بزرگ

★ یک پدیده تصادفی:

★ انتخاب $v_1, \dots, v_n$ در R^m درگوی با شعاع ۲

★ رخداد

$$A_G := \{\forall i, j : | \langle v_i, v_j \rangle - \langle w_i, w_j \rangle | \leq \varepsilon/2\}$$

★ رخدادهای A_G مجزا هستند.

$$\sum_{G \in \mathcal{G}} \mathbb{P}[A_G] \leq 1$$

بزرگ

پس: $|G|$ کم

★ یک پدیده تصادفی:

★ انتخاب $v_1, \ldots, v_n$ در R^m در گوی با شعاع ۲

★ رخداد

$$A_G := \{\forall i, j : | \langle v_i, v_j \rangle - \langle w_i, w_j \rangle | \leq \varepsilon/2\}$$

★ رخدادهای A_G مجزا هستند.

$$\sum_{G \in \mathcal{G}} \mathbb{P}[A_G] \leq 1$$

پس: $|G|$ کم

بزرگ

کافی است:

$$\{\forall i : ||v_i - w_i||_2 \leq \varepsilon/4\}$$

$$\mathbb{P}[A_G] \geq \Omega(\varepsilon)^{mn}$$

★ یک پدیده تصادفی:

★ انتخاب $v_1, ..., v_n$ در R^m در گوی با شعاع ۲

★ رخداد

$$A_G := \{\forall i, j : |\langle v_i, v_j\rangle - \langle w_i, w_j\rangle| \leq \varepsilon/2\}$$

★ رخدادهای A_G مجزا هستند.

$$\sum_{G \in \mathcal{G}} \mathbb{P}[A_G] \leq 1$$

کافی است:

$\{\forall i : ||v_i - w_i||_2 \leq \varepsilon/4\}$

بزرگ

پس: $|G|$ کم

$\mathbb{P}[A_G] \geq \Omega(\varepsilon)^{mn}$

پس $|G| \leq O(mn \log \epsilon)$

پس $|G| \leq O(n/\epsilon^2 \log n \log \epsilon)$

$$\sum_{G\in\mathcal{G}} \mathbb{P}[A_G] \le 1$$

پس: |G| کم

بزرگ

کافی است:

$\{\forall i : ||v_i - w_i||_2 \le \varepsilon/4\}$

پس $|G| \le O(mn \log \epsilon)$

$\mathbb{P}[A_G] \ \ge \ \Omega(\varepsilon)^{mn}$

پس $|G| \le O(n/\epsilon^2 \log n \log \epsilon)$

$\mathbb{P}[A_G] \ge \Omega(1)^{mn}$

بردار تصادفی واحد

$E := \{\forall i : ||v_i - w_i||_2 \le 1\}$

$\mathbb{P}[E] = (\frac{1}{2})^{nm}$

$$\forall i \ne j : \mathbb{P}[|\langle v_i - w_i, w_j\rangle| \ge \varepsilon/4 \ \ |E] \le 2e^{-\Omega(\varepsilon^2 m)} < \frac{1}{2n^2}$$

$$\forall i \ne j : \mathbb{P}[|\langle v_i, v_j - w_j\rangle| \ge \varepsilon/4 \ \ |E] \le 2e^{-\Omega(\varepsilon^2 m)} < \frac{1}{2n^2}$$

$$\sum_{G\in\mathcal{G}} \mathbb{P}[A_G] \le 1$$

پس: |G| کم

بزرگ

کافی است:
$\{\forall i : ||v_i - w_i||_2 \le \varepsilon/4\}$

پس $|G| \le O(n/\epsilon^2 \log n \log \epsilon)$

$$\mathbb{P}[A_G] \ge \Omega(1)^{mn}$$

$$E := \{\forall i : ||v_i - w_i||_2 \le 1\}$$

$$\mathbb{P}[E] = (\tfrac{1}{2})^{nm}$$

$$|\langle v_i, v_j\rangle - \langle w_i, w_j\rangle| = |\langle v_i - w_i, v_j\rangle + \langle w_i, v_j - w_j\rangle| \le \epsilon/2$$

$$\forall i \ne j : \mathbb{P}[|\langle v_i - w_i, w_j\rangle| \ge \varepsilon/4 \;\; |E] \le 2e^{-\Omega(\varepsilon^2 m)} < \frac{1}{2n^2}$$

$$\forall i \ne j : \mathbb{P}[|\langle v_i, v_j - w_j\rangle| \ge \varepsilon/4 \;\; |E] \le 2e^{-\Omega(\varepsilon^2 m)} < \frac{1}{2n^2}$$

بردار تصادفی واحد

اثبات قسمت (B)

- بردار تصادفی vها

$$E := \{\forall i : ||v_i - w_i||_2 \le \delta/40\}$$

اثبات قسمت (B)

- بردار تصادفی vها

$$E := \{\forall i : ||v_i - w_i||_2 \leq \delta/40\}$$

$$\mathbb{P}[E] = (\delta/80)^{kn}$$

اثبات قسمت (B)

- بردار تصادفی vها

$$E := \{\forall i : ||v_i - w_i||_2 \leq \delta/40\}$$

$$\mathbb{P}[E] = (\delta/80)^{kn}$$

$$\forall i \neq j : \mathbb{P}[|\langle v_i - w_i, w_j \rangle| \geq \varepsilon/4 \;\; |E] \leq 2e^{-\Omega(k\varepsilon^2/\delta^2)} < \frac{1}{2n^2}$$

اثبات قسمت (C)

- هر نقطه به نزدیک‌ترین نقطه در یک e_شبکه
- تعداد بیت: $O(k\log(1/\varepsilon))$
-

الگوریتمی

$$m := \frac{\log n}{\varepsilon^2}$$

- حالت (A):
  - ۱ـ تبدیل JL به m بعد
  - ۲ـ گرد کردن به عدد صحیح ضریب $1/\sqrt{2m}$
- حالت (B):

$$2\varepsilon \leq \delta < 1/2$$

  - گرد کردن به $\delta/\sqrt{k}$
- حالت (C):
  - گرد کردن به گرید با فاصله ϵ

کران پایین (A)

Theorem 90 ([Alon17]). *Let $f(n,k,\varepsilon)$ be the minimum size (in bits) of an ε-inner-product sketch for n points in dimension k.*

Then, $\forall n \geq k, \varepsilon \geq \frac{1}{n^{0.49}}$:

(A) *For* $\frac{\log n}{\varepsilon^2} \leq k \leq n : f(n,k,\varepsilon) = \Theta(\frac{n \log n}{\varepsilon^2})$

بهتر از LJ نمی‌شود:

(B) *For* $\log n \leq k \leq \frac{\log n}{\varepsilon^2} : f(n,k,\varepsilon) = \Theta(nk \log(2 + \frac{\log n}{\varepsilon^2 k}))$

(C) *For* $1 \leq k \leq \log n : f(n,k,\varepsilon) = \Theta(nk \log(1/\varepsilon))$.

کران پایین (B)

Lemma 91. *If* $k = \frac{\delta^2}{200\varepsilon^2} \log n$, *then*

$$f(n, k, \varepsilon) \geq \Omega(nk \log(1/\delta)).$$

کران پایین (B)

Lemma 91. *If* $k = \frac{\delta^2}{200\varepsilon^2} \log n$, *then*

$$f(n,k,\varepsilon) \geq \Omega(nk \log(1/\delta)).$$

N: گرید با فاصله $1/\delta$ (تعداد = $(1/\delta)^k$)

R: $n/2$ بردار تصادفی در R^k

کران پایین (B)

Lemma 91. *If* $k = \frac{\delta^2}{200\varepsilon^2} \log n$, *then*

$$f(n, k, \varepsilon) \geq \Omega(nk \log(1/\delta)).$$

$$\forall x \neq x' \in N :$$
$$\exists y \in R : |\langle x, y \rangle - \langle x', y \rangle| > \varepsilon$$

N: گرید با فاصله $1/\delta$ (تعداد = $(1/\delta)^k$)

R: n/2 بردار تصادفی در R^k

کران پایین (B)

Lemma 91. *If* $k = \frac{\delta^2}{200\varepsilon^2}\log n$*, then*

$$f(n,k,\varepsilon) \geq \Omega(nk\log(1/\delta)).$$

N: گرید با فاصله $1/\delta$ (تعداد = $(1/\delta)^k$)

R: n/2 بردار تصادفی در R^k

$$\forall x \neq x' \in N:$$
$$\exists y \in R : |\langle x,y\rangle - \langle x',y\rangle| > \varepsilon$$

مجموعه‌های R + n/2 تا از N ==> هر دو متفاوت

کران پایین (B)

Lemma 91. *If* $k = \frac{\delta^2}{200\varepsilon^2}\log n$, *then*

$$f(n,k,\varepsilon) \geq \Omega(nk\log(1/\delta)).$$

N: گرید با فاصله $1/\delta$ (تعداد = $(1/\delta)^k$)

R: n/2 بردار تصادفی در R^k

$$\forall x \neq x' \in N:$$
$$\exists y \in R: |\langle x,y\rangle - \langle x',y\rangle| > \varepsilon$$

مجموعه‌های R + n/2 تا از N <== هر دو متفاوت

پس: $f(n,k,\varepsilon) \geq \log|N|^{n/2} = \Omega(\frac{n}{2}\log|N|) = \Omega(nk\log(1/\delta))$