

بسم الله الرحمن الرحيم

جلسه يازدهم

خلاصه سازی برای مهداده



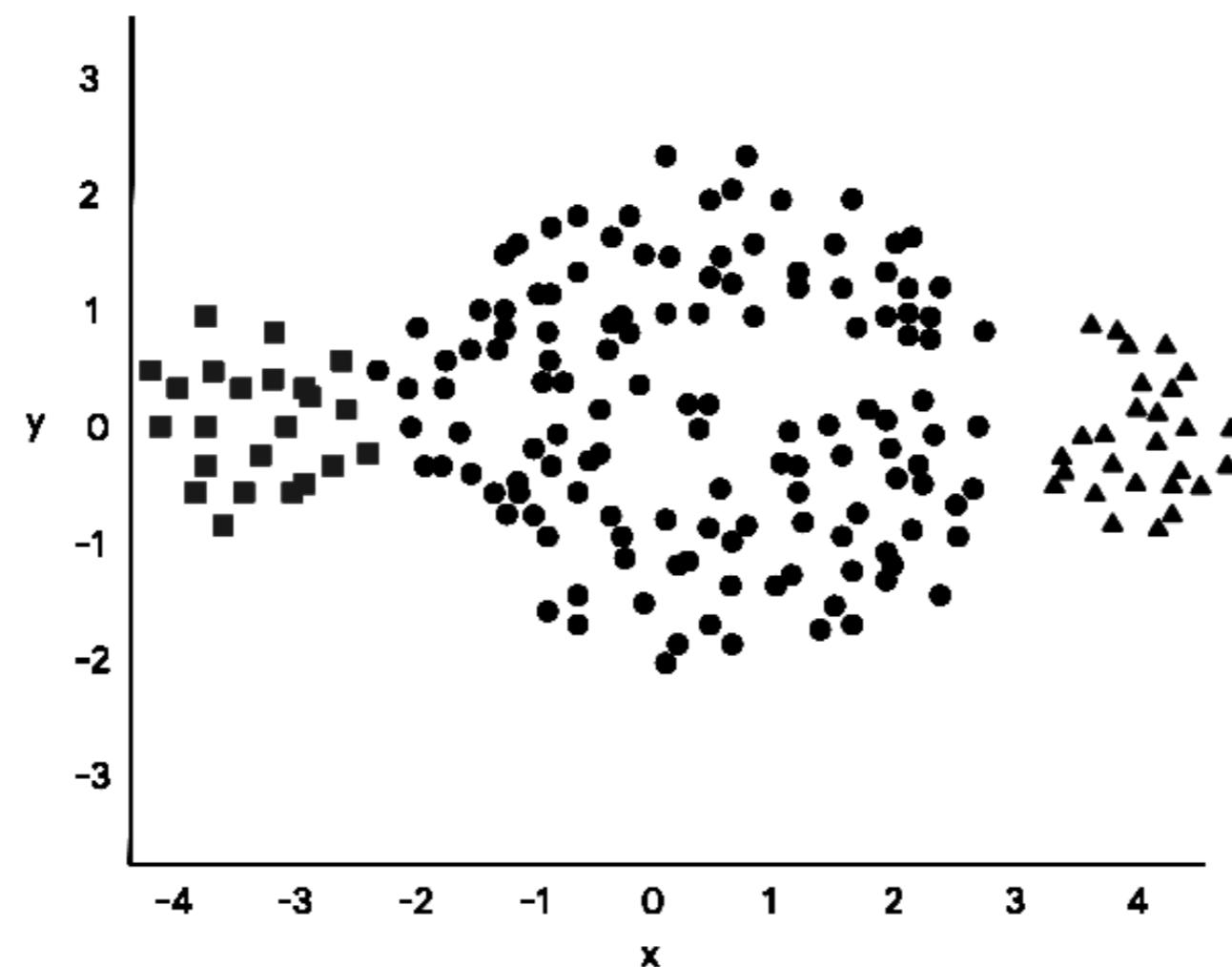
تبدیل JL

Johnson-Lindenstrauss Transforms

مثال: خوشبندی

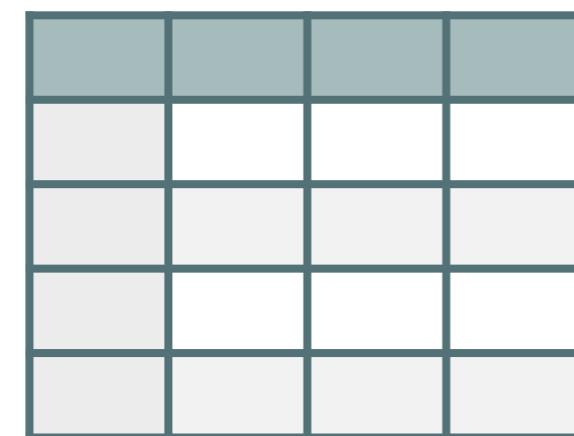
مثال: بیماران MS

= مشخصات بیماران

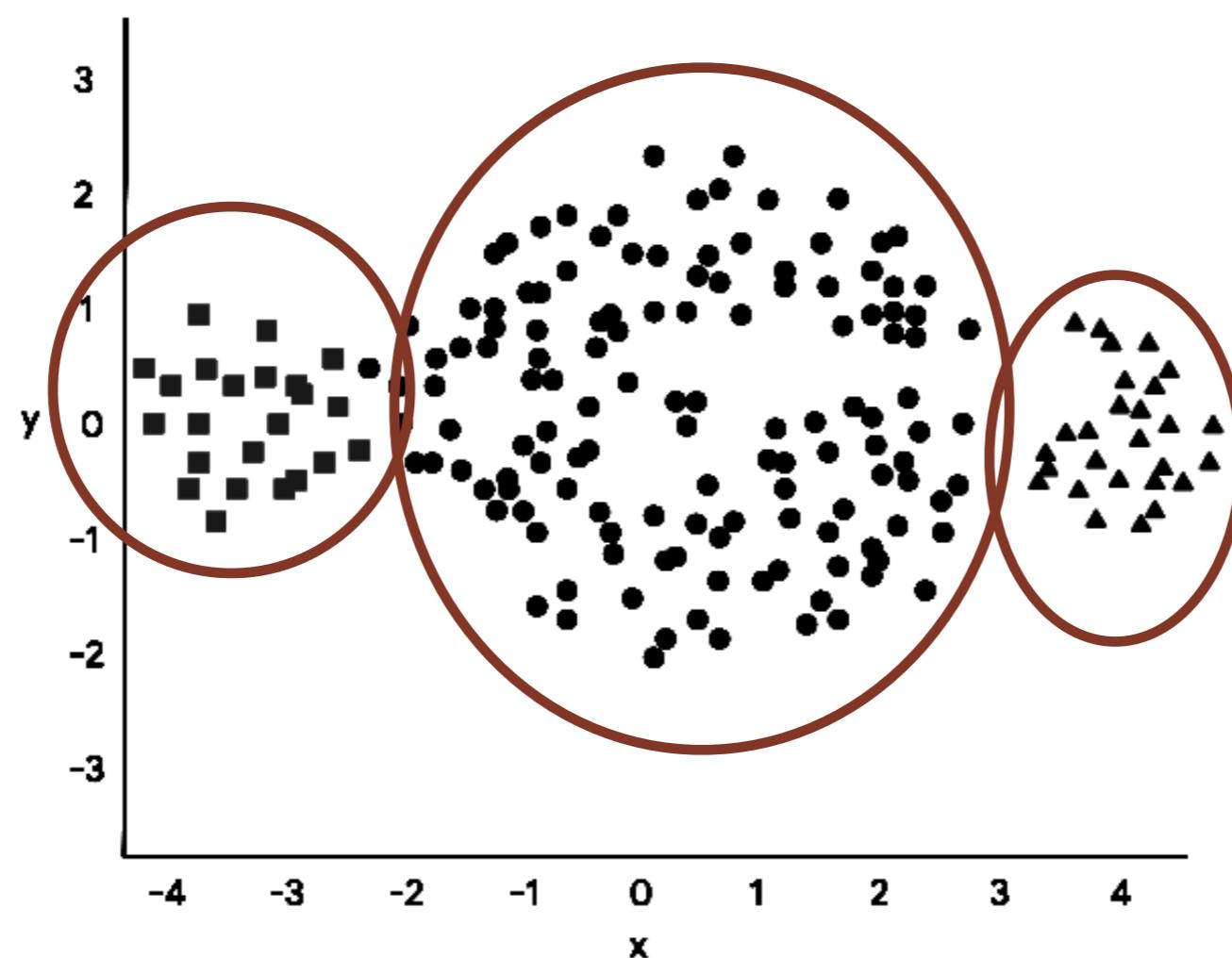


مثال: خوشبندی

= مشخصات بیماران



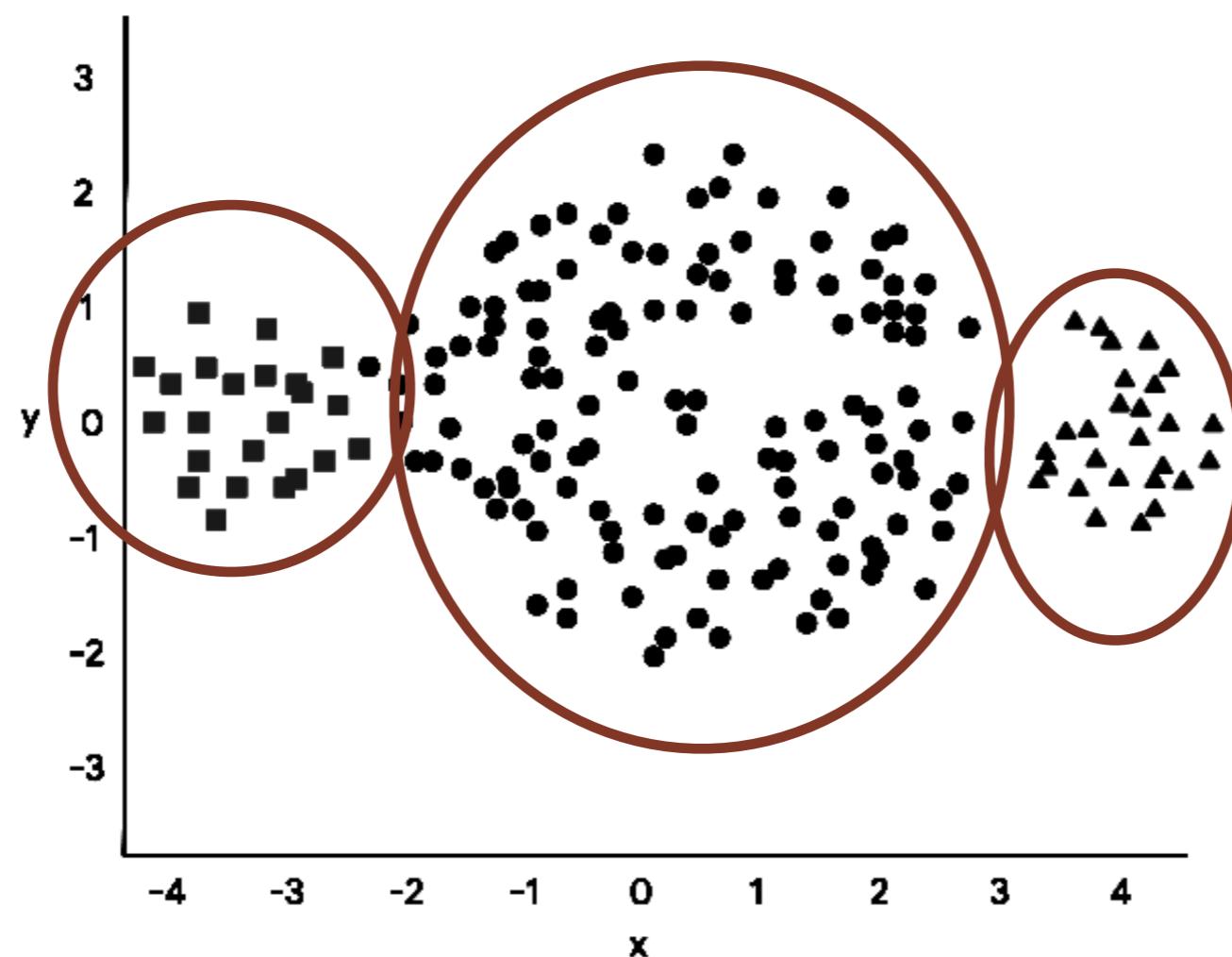
مثال: بیماران MS



مثال: خوشبندی

مثال: بیماران MS

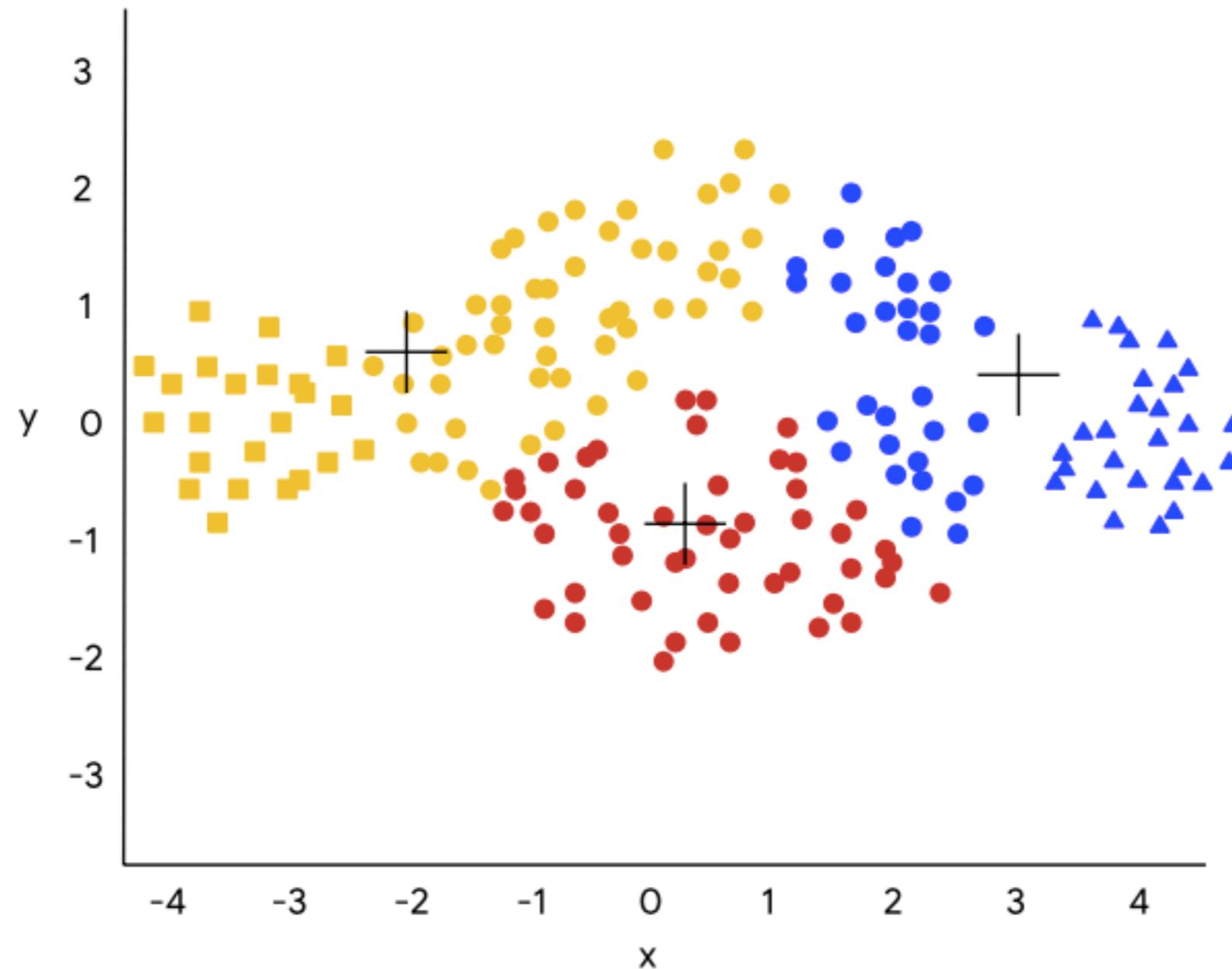
= مشخصات بیماران



خوشبندی

مسئله K-میانگین‌ها (روشی برای خوشه‌بندی)

- پیدا کردن ۳ مرکز
- نقاط هر دسته نزدیک مرکزشان باشند



مسئله K-میانگین‌ها (روشی برای خوشه‌بندی)

• ورودی: $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ ◉

مسئله K-میانگین‌ها (روشی برای خوشه‌بندی)

• ورودی: $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ ◉

• هدف: یافتن ◉

$y_1, \dots, y_k \in \mathbb{R}^d$ ۱ - مراکز ◉

$\mathcal{P} = (\mathcal{P}_1, \dots, \mathcal{P}_k)$ ۲ - انتساب نقاط به مراکز ◉

مسئله K-میانگین‌ها (روشی برای خوشه‌بندی)

• ورودی: $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ ◉

• هدف: یافتن ◉

$y_1, \dots, y_k \in \mathbb{R}^d$ ۱ - مراکز ◉

$\mathcal{P} = (\mathcal{P}_1, \dots, \mathcal{P}_k)$ ۲ - انتساب نقاط به مراکز ◉

• تابع هدف: $\min \sum_{i=1}^n \|x_i - y_{\pi_{\mathcal{P}}(i)}\|_2^2$

مسئله K-میانگین‌ها (روشی برای خوشه‌بندی)

• ورودی: $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ ◉

• هدف: یافتن ◉

$y_1, \dots, y_k \in \mathbb{R}^d$ ۱ - مراکز ◉

$\mathcal{P} = (\mathcal{P}_1, \dots, \mathcal{P}_k)$ ۲ - انتساب نقاط به مراکز ◉

• تابع هدف: $\min \sum_{i=1}^n \|x_i - y_{\pi_{\mathcal{P}}(i)}\|_2^2$

$\min_{k-\text{partitions } \mathcal{P}} \min_{y_1, \dots, y_k} \sum_{j=1}^k \sum_{i \in \mathcal{P}_j} \|x_i - y_j\|_2^2$

مسئله K-میانگین‌ها (روشی برای خوشه‌بندی)

• ورودی: $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ ◉

• هدف: یافتن ◉

- $y_1, \dots, y_k \in \mathbb{R}^d$ ◉ ۱ - مراکز
- $\mathcal{P} = (\mathcal{P}_1, \dots, \mathcal{P}_k)$ ◉ ۲ - انتساب نقاط به مراکز

• تابع هدف: ◉

$$\min \sum_{i=1}^n \|x_i - y_{\pi_{\mathcal{P}}(i)}\|_2^2$$

$$\min_{k-\text{partitions } \mathcal{P}} \min_{y_1, \dots, y_k} \sum_{j=1}^k \sum_{i \in \mathcal{P}_j} \|x_i - y_j\|_2^2$$

مسئله K-میانگین‌ها (روشی برای خوشه‌بندی)

$$\min_{k\text{-partitions } \mathcal{P}} \min_{y_1, \dots, y_k} \sum_{j=1}^k \sum_{i \in \mathcal{P}_j} \|x_i - y_j\|_2^2$$

مسئله:

مسئله K-میانگین‌ها (روشی برای خوشه‌بندی)

$$\min_{k\text{-partitions } \mathcal{P}} \min_{y_1, \dots, y_k} \sum_{j=1}^k \sum_{i \in \mathcal{P}_j} \|x_i - y_j\|_2^2$$

مسئله:

$$\min_{y_1, \dots, y_k} \sum_{i=1}^n \min_{1 \leq j \leq k} \|x_i - y_j\|_2^2$$

متغیرها: مراکز

مسئله K-میانگین‌ها (روشی برای خوشه‌بندی)

$$\min_{k\text{-partitions } \mathcal{P}} \min_{y_1, \dots, y_k} \sum_{j=1}^k \sum_{i \in \mathcal{P}_j} \|x_i - y_j\|_2^2$$

مسئله:

$$\min_{y_1, \dots, y_k} \sum_{i=1}^n \min_{1 \leq j \leq k} \|x_i - y_j\|_2^2$$

متغیرها: مراکز

$$\mu_j := (1/|\mathcal{P}_j|) \sum_{i \in \mathcal{P}_j} x_i$$

$$\min_{k\text{-partitions } \mathcal{P}} \sum_{j=1}^k \sum_{i \in \mathcal{P}_j} \|x_i - \mu_j\|_2^2$$

متغیرها: افزایش

مسئله K-میانگین‌ها (روشی برای خوشه‌بندی)

$$\min_{k\text{-partitions } \mathcal{P}} \min_{y_1, \dots, y_k} \sum_{j=1}^k \sum_{i \in \mathcal{P}_j} \|x_i - y_j\|_2^2$$

مسئله:

$$\min_{y_1, \dots, y_k} \sum_{i=1}^n \min_{1 \leq j \leq k} \|x_i - y_j\|_2^2$$

متغیرها: مراکز

$$\mu_j := (1/|\mathcal{P}_j|) \sum_{i \in \mathcal{P}_j} x_i$$

$$\min_{k\text{-partitions } \mathcal{P}} \sum_{j=1}^k \sum_{i \in \mathcal{P}_j} \|x_i - \mu_j\|_2^2$$

متغیرها: افزایش

$$= \min_{k\text{-partitions } \mathcal{P}} \sum_{j=1}^k \frac{1}{|\mathcal{P}_j|} \sum_{i < i' \in \mathcal{P}_j} \|x_i - x_{i'}\|_2^2$$

مسئله:

$$\min_{k-\text{partitions } \mathcal{P}} \sum_{j=1}^k \sum_{i \in \mathcal{P}_j} \|x_i - \mu_j\|_2^2$$

$$= \min_{k-\text{partitions } \mathcal{P}} \sum_{j=1}^k \frac{1}{|\mathcal{P}_j|} \sum_{i < i' \in \mathcal{P}_j} \|x_i - x_{i'}\|_2^2$$

فقط فاصله‌ها مهمند

مسئله:

$$\min_{k\text{-partitions } \mathcal{P}} \sum_{j=1}^k \sum_{i \in \mathcal{P}_j} \|x_i - \mu_j\|_2^2$$

$$= \min_{k\text{-partitions } \mathcal{P}} \sum_{j=1}^k \frac{1}{|\mathcal{P}_j|} \sum_{i < i' \in \mathcal{P}_j} \|x_i - x_{i'}\|_2^2$$

فقط فاصله‌ها مهمند

الگوریتم حساس به بعد d

$$X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$$

مسئله:

$$\min_{k-\text{partitions } \mathcal{P}} \sum_{j=1}^k \sum_{i \in \mathcal{P}_j} \|x_i - \mu_j\|_2^2$$

$$= \min_{k-\text{partitions } \mathcal{P}} \sum_{j=1}^k \frac{1}{|\mathcal{P}_j|} \sum_{i < i' \in \mathcal{P}_j} \|x_i - x_{i'}\|_2^2$$

فقط فاصله‌ها مهمند

الگوریتم حساس به بعد d

$$X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$$

چه خوب می‌شد

فاصله‌ها را حفظ کند

$$f : X \rightarrow \mathbb{R}^m$$

$m \ll d$

$$= \min_{k-\text{partitions } \mathcal{P}} \sum_{j=1}^k \frac{1}{|\mathcal{P}_j|} \sum_{i < i' \in \mathcal{P}_j} \|x_i - x_{i'}\|_2^2$$

چه خوب می‌شد

فاصله‌ها را حفظ کند

$$f : X \rightarrow \mathbb{R}^m \quad m \ll d$$

نسبتاً خوب می‌شد: فاصله‌ها را تقریباً حفظ کند

$$\forall x, y \in X, \quad (1 - \varepsilon) \|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \varepsilon) \|x - y\|_2^2$$

$$= \min_{k-\text{partitions } \mathcal{P}} \sum_{j=1}^k \frac{1}{|\mathcal{P}_j|} \sum_{i < i' \in \mathcal{P}_j} \|x_i - x_{i'}\|_2^2$$

چه خوب می‌شد

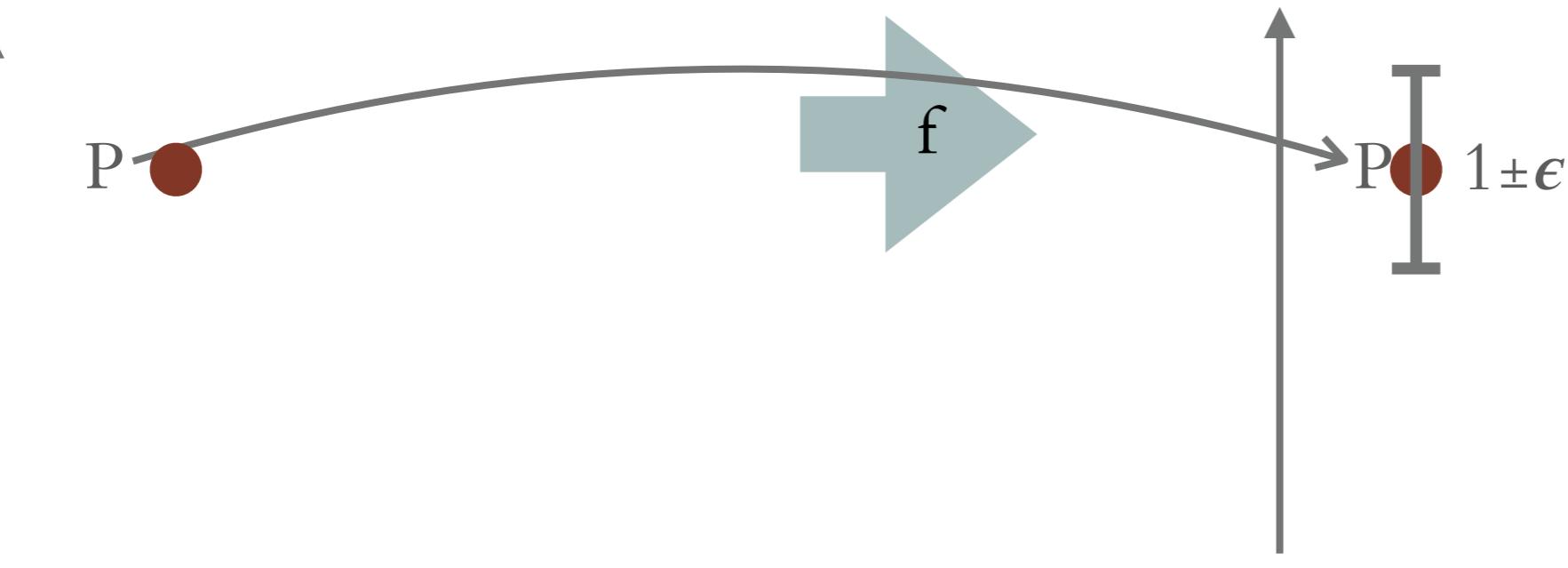
فاصله‌ها را حفظ کند

$$f : X \rightarrow \mathbb{R}^m$$

$m \ll d$

نسبتاً خوب می‌شد: فاصله‌ها را تقریباً حفظ کند

$$\forall x, y \in X, (1 - \varepsilon) \|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \varepsilon) \|x - y\|_2^2$$



$$= \min_{k-\text{partitions } \mathcal{P}} \sum_{j=1}^k \frac{1}{|\mathcal{P}_j|} \sum_{i < i' \in \mathcal{P}_j} \|x_i - x_{i'}\|_2^2$$

چه خوب می‌شد

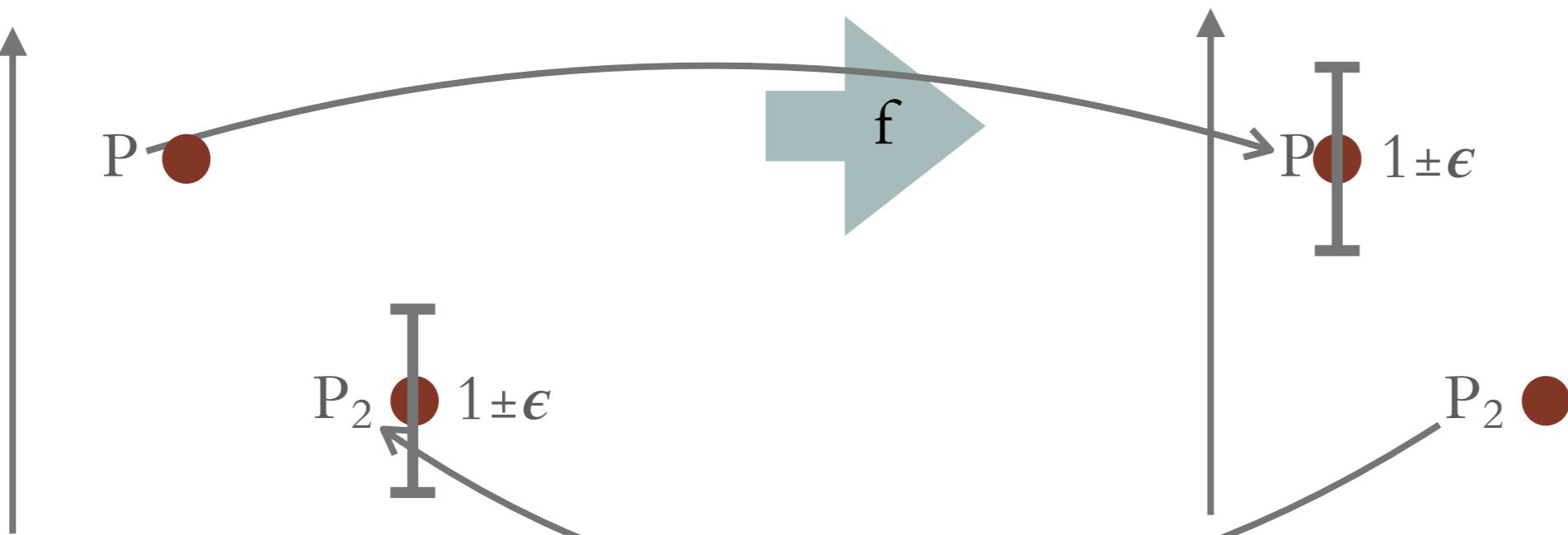
فاصله‌ها را حفظ کند

$$f : X \rightarrow \mathbb{R}^m$$

$$m \ll d$$

نسبتاً خوب می‌شد: فاصله‌ها را تقریباً حفظ کند

$$\forall x, y \in X, (1 - \varepsilon) \|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \varepsilon) \|x - y\|_2^2$$



Theorem 5.0.1 (JL lemma [JL84]). *For any $\varepsilon \in (0, 1)$ and any $X \subset \mathbb{R}^d$ for $|X| = n$ finite, there exists an embedding $f : X \rightarrow \mathbb{R}^m$ for $m = O(\varepsilon^{-2} \log n)$ such that*

$$\forall x, y \in X, \quad (1 - \varepsilon) \|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \varepsilon) \|x - y\|_2^2. \quad (5.1)$$

Theorem 5.0.1 (JL lemma [JL84]). *For any $\varepsilon \in (0, 1)$ and any $X \subset \mathbb{R}^d$ for $|X| = n$ finite, there exists an embedding $f : X \rightarrow \mathbb{R}^m$ for $m = O(\varepsilon^{-2} \log n)$ such that*

$$\forall x, y \in X, \quad (1 - \varepsilon) \|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \varepsilon) \|x - y\|_2^2. \quad (5.1)$$

$$\forall x, y \in X \quad c \cdot d_X(x, y) \leq d_Y(f(x), f(y)) \leq \rho \cdot c \cdot d_X(x, y) \quad \text{تبدیل}$$

اعوجاج: ρ

تبدیل :JL

Theorem 5.0.1 (JL lemma [JL84]). *For any $\varepsilon \in (0, 1)$ and any $X \subset \mathbb{R}^d$ for $|X| = n$ finite, there exists an embedding $f : X \rightarrow \mathbb{R}^m$ for $m = O(\varepsilon^{-2} \log n)$ such that*

$$\forall x, y \in X, \quad (1 - \varepsilon) \|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \varepsilon) \|x - y\|_2^2. \quad (5.1)$$

$$\forall x, y \in X \quad c \cdot d_X(x, y) \leq d_Y(f(x), f(y)) \leq \rho \cdot c \cdot d_X(x, y) \quad \text{تبدیل جاوج: } \rho$$

$$1 + O(\varepsilon) = \sqrt{(1 + \varepsilon)/(1 - \varepsilon)} \quad \text{جاوج JL: } \rho$$

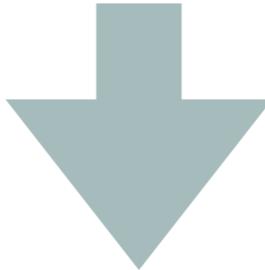
برای اثبات لم JL:

Lemma 5.0.3 (DJL lemma). *For any $\varepsilon, \delta \in (0, 1/2)$ and integer $d > 1$, there exists a distribution $\mathcal{D}_{\varepsilon, \delta}$ over matrices $\Pi \in \mathbb{R}^{m \times d}$ for $m = O(\varepsilon^{-2} \log(1/\delta))$ such that for any fixed $z \in \mathbb{R}^d$ with $\|z\|_2 = 1$,*

$$\mathbb{P}_{\Pi \sim \mathcal{D}_{\varepsilon, \delta}} (|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta.$$

Lemma 5.0.3 (DJL lemma). *For any $\varepsilon, \delta \in (0, 1/2)$ and integer $d > 1$, there exists a distribution $\mathcal{D}_{\varepsilon, \delta}$ over matrices $\Pi \in \mathbb{R}^{m \times d}$ for $m = O(\varepsilon^{-2} \log(1/\delta))$ such that for any fixed $z \in \mathbb{R}^d$ with $\|z\|_2 = 1$,*

$$\mathbb{P}_{\Pi \sim \mathcal{D}_{\varepsilon, \delta}} (|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta.$$



Theorem 5.0.1 (JL lemma [JL84]). *For any $\varepsilon \in (0, 1)$ and any $X \subset \mathbb{R}^d$ for $|X| = n$ finite, there exists an embedding $f : X \rightarrow \mathbb{R}^m$ for $m = O(\varepsilon^{-2} \log n)$ such that*

$$\forall x, y \in X, \quad (1 - \varepsilon)\|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \varepsilon)\|x - y\|_2^2. \quad (5.1)$$

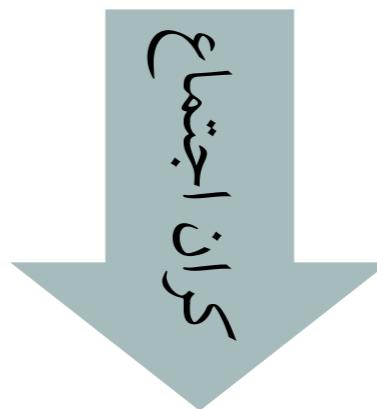
$$\delta < 1/n^2$$

$$\epsilon$$

$$d$$



$$z_{x,y} := (x - y) / \|x - y\|_2 \quad \mathbb{P}(|\|\Pi z_{x,y}\|_2^2 - 1| > \varepsilon) < \delta$$



كاربرد
الگوريتمي؟

احتمال بد بودن حداقل يك جفت < $\binom{n}{2} \delta < 1$

اثبات DJL

$$\mathbb{P}_{\Pi \sim \mathcal{D}_{\varepsilon, \delta}} (|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta.$$

تعريف $\Pi_{r,i} = \sigma_{r,i}/\sqrt{m}$: $\sigma_{r,i}$ با احتمال $1/2$ یا -1

اثبات DJL

$$\mathbb{P}_{\Pi \sim \mathcal{D}_{\varepsilon, \delta}} (|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta.$$

$$\Pi_{r,i} = \sigma_{r,i} / \sqrt{m}$$

تعريف $\sigma_{r,i}$ با احتمال $1/2$ یا -1

حکم: Π ها شرایط را دارند

اثبات DJL

$$\mathbb{P}_{\Pi \sim \mathcal{D}_{\varepsilon, \delta}} (|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta.$$

$$\Pi_{r,i} = \sigma_{r,i}/\sqrt{m}$$

تعريف $\sigma_{r,i}$ با احتمال $1/2$ یا 1

تعريف

حکم: Π ها شرایط را دارند

$$\sigma = (\sigma_{1,1}, \sigma_{1,2}, \dots, \sigma_{1,d}, \dots, \sigma_{m,1}, \dots, \sigma_{m,d})$$

$$B_z = \frac{1}{\sqrt{m}} \cdot \begin{bmatrix} -z^\top - & 0 & \cdots & 0 \\ 0 & -z^\top - & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & -z^\top - \end{bmatrix} \quad \rightarrow \quad \Pi z = B_z \sigma$$

اثبات DJL

$$\mathbb{P}_{\Pi \sim \mathcal{D}_{\varepsilon, \delta}} (|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta.$$

$$\Pi_{r,i} = \sigma_{r,i}/\sqrt{m}$$

یا $1 - 1/2$ با احتمال $\sigma_{r,i}$

تعريف

حکم: Π ها شرایط را دارند

$$\sigma = (\sigma_{1,1}, \sigma_{1,2}, \dots, \sigma_{1,d}, \dots, \sigma_{m,1}, \dots, \sigma_{m,d})$$

$$B_z = \frac{1}{\sqrt{m}} \cdot \begin{bmatrix} -z^\top - & 0 & \cdots & 0 \\ 0 & -z^\top - & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & -z^\top - \end{bmatrix} \quad \rightarrow \quad \Pi z = B_z \sigma$$

$$\rightarrow \|\Pi z\|_2^2 - 1 = \|B_z \sigma\|_2^2 - 1 = \sigma^\top B_z^\top B_z \sigma - \mathbb{E} \sigma^\top B_z^\top B_z \sigma$$

اثبات DJL

$$\mathbb{P}_{\Pi \sim \mathcal{D}_{\varepsilon, \delta}} (|\|\Pi z\|_2^2 - 1| > \varepsilon) < \delta.$$

$$\Pi_{r,i} = \sigma_{r,i}/\sqrt{m}$$

تعريف $\sigma_{r,i}$ با احتمال $1/2$ یا -1

تعريف

حکم: Π ها شرایط را دارند

$$\sigma = (\sigma_{1,1}, \sigma_{1,2}, \dots, \sigma_{1,d}, \dots, \sigma_{m,1}, \dots, \sigma_{m,d})$$

$$B_z = \frac{1}{\sqrt{m}} \cdot \begin{bmatrix} -z^\top - & 0 & \cdots & 0 \\ 0 & -z^\top - & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & -z^\top - \end{bmatrix} \quad \rightarrow \quad \Pi z = B_z \sigma$$

$$\rightarrow \|\Pi z\|_2^2 - 1 = \|B_z \sigma\|_2^2 - 1 = \sigma^\top B_z^\top B_z \sigma - \mathbb{E} \sigma^\top B_z^\top B_z \sigma$$

$$A_z := B_z^\top B_z$$

قضیه

$$\mathbb{P}(|\sigma^\top A_z \sigma - \mathbb{E} \sigma^\top A_z \sigma| > \varepsilon) \lesssim e^{-C\varepsilon^2/\|A_z\|_F^2} + e^{-C\varepsilon/\|A\|} \leq ?$$

$$\mathbb{P}(|\sigma^\top A_z \sigma - \mathbb{E} \sigma^\top A_z \sigma| > \varepsilon) \lesssim e^{-C\varepsilon^2/\|A_z\|_F^2} + e^{-C\varepsilon/\|A\|} \leq ?$$

$$B_z = \frac{1}{\sqrt{m}} \cdot \begin{bmatrix} -z^\top - & 0 & \cdots & 0 \\ 0 & -z^\top - & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & -z^\top - \end{bmatrix}$$

$$A_z := B_z^\top B_z$$

$$A_z =$$

$(1/m)zz^\top$			
0	$(1/m)zz^\top$	0	
	0	\dots	
			$(1/m)zz^\top$

$$\mathbb{P}(|\sigma^\top A_z \sigma - \mathbb{E} \sigma^\top A_z \sigma| > \varepsilon) \lesssim e^{-C\varepsilon^2/\|A_z\|_F^2} + e^{-C\varepsilon/\|A\|} \leq ?$$

$$B_z = \frac{1}{\sqrt{m}} \cdot \begin{bmatrix} -z^\top - & 0 & \cdots & 0 \\ 0 & -z^\top - & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & -z^\top - \end{bmatrix}$$

$$A_z := B_z^\top B_z$$

$$A_z =$$

$(1/m)zz^\top$			
0	$(1/m)zz^\top$	0	
	0	\dots	
			$(1/m)zz^\top$

$$\|A_z\|_F =$$

$$(1/m^2) \sum_{r=1}^m \|zz^\top\|_F^2 = (1/m^2) \sum_{r=1}^m \sum_{i,j} z_i^2 z_j^2 = (1/m) \|z\|_2^4 = 1/m$$

$$\mathbb{P}(|\sigma^\top A_z \sigma - \mathbb{E} \sigma^\top A_z \sigma| > \varepsilon) \lesssim e^{-C\varepsilon^2/\|A_z\|_F^2} + e^{-C\varepsilon/\|A\|} \leq ?$$

$$B_z = \frac{1}{\sqrt{m}} \cdot \begin{bmatrix} -z^\top - & 0 & \cdots & 0 \\ 0 & -z^\top - & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & -z^\top - \end{bmatrix}$$

$$A_z := B_z^\top B_z$$

$$A_z =$$

$(1/m)zz^\top$			
0	$(1/m)zz^\top$	0	
	0	\dots	
			$(1/m)zz^\top$

$$\|A_z\|_F =$$

$$(1/m^2) \sum_{r=1}^m \|zz^\top\|_F^2 = (1/m^2) \sum_{r=1}^m \sum_{i,j} z_i^2 z_j^2 = (1/m) \|z\|_2^4 = 1/m$$

$\|A_z\|$ largest eigenvalue of $(1/m)zz^\top$ eigenvalue $(1/m)\|z\|_2^2 = 1/m$.

$$\mathbb{P}(|\sigma^\top A_z \sigma - \mathbb{E}\,\sigma^\top A_z \sigma| > \varepsilon) \lesssim e^{-C\varepsilon^2/\|A_z\|_F^2} + e^{-C\varepsilon/\|A\|}$$

$$<=$$

$$e^{-C\varepsilon^2/m} + e^{-C\varepsilon/m}$$

$$\mathbb{P}(|\sigma^\top A_z \sigma - \mathbb{E}\sigma^\top A_z \sigma| > \varepsilon) \lesssim e^{-C\varepsilon^2/\|A_z\|_F^2} + e^{-C\varepsilon/\|A\|}$$

$$<=$$

$$e^{-C\varepsilon^2/m} + e^{-C\varepsilon/m}$$

$$\text{at most } \delta \text{ for } m = \Omega(\varepsilon^{-1} \log(1/\delta) + \varepsilon^{-2} \log(1/\delta)) = \Omega(\varepsilon^{-2} \log(1/\delta))$$