Randomized Algorithms

Ioannis Caragiannis (this time) and Kasper Green Larsen

Multi-dimensional data

- **Documents** as bag of words: # of occurences of word w in a document
- **Network traffic**: number of packets sent by node i to node j
- **User ratings**: rating of user i for service/product/business/etc j

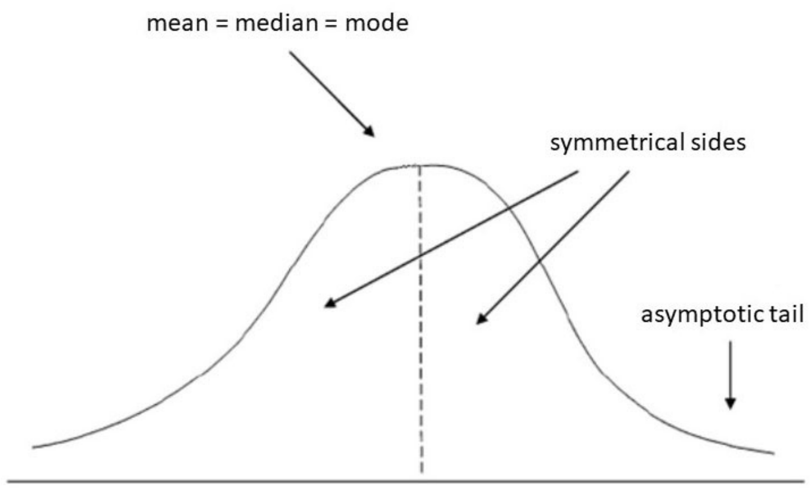
How can we compare documents?

- **Similarity** between two documents is given by the **distance** of their "vectors"
- Claim: **projecting** the document vector in a **smaller space** preserves the similarity between documents
- How? E.g., using the **Johnson-Lindenstrauss** transform

Useful tools

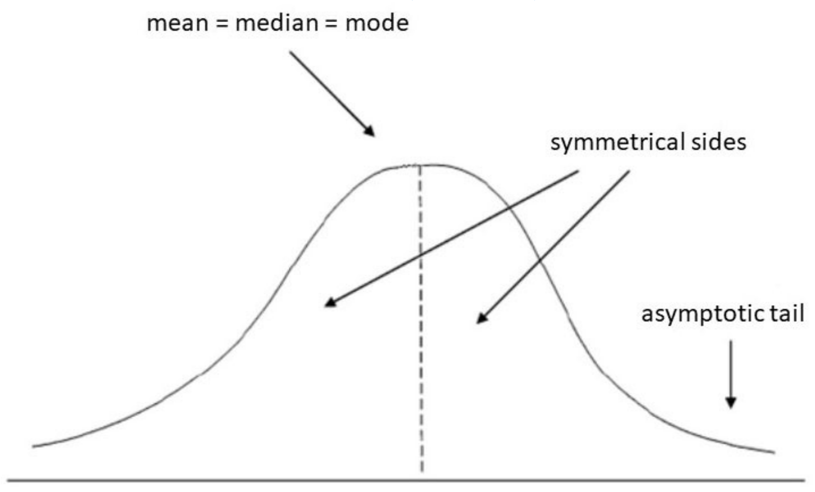
- **Normal/Gaussian** probability distributions
- A random variable that follows the normal distribution $\mathcal{N}(\mu, \sigma^2)$ with **expectation μ** and **standard deviation σ** has probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Useful tools

- **Normal/Gaussian** probability distributions
- A random variable that follows the normal distribution $\mathcal{N}(\mu, \sigma^2)$ with **expectation μ** and **standard deviation σ** has probability density function
  $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
- Today, we will use extensively random variables from $\boldsymbol{\mathcal{N}(0, 1)}$, with pdf
  $$\boxed{f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)}$$



The Johnson-Lindenstrauss transform

The Johnson-Lindenstrauss lemma

- For any $\varepsilon \in (0,1/2)$ and any integer m, then for integer $k = O\left(\frac{1}{\varepsilon^2} \ln m\right)$ and any points $x_1, x_2, \ldots, x_m \in \mathbb{R}^d$, there exists a **linear map** (matrix) $L: \mathbb{R}^d \to \mathbb{R}^k$ such that for any $1 \le i < j \le m$, it holds

$$(1-\varepsilon)\|x_i - x_j\|_2^2 \le \|Lx_i - Lx_j\|_2^2 \le (1+\varepsilon)\|x_i - x_j\|_2^2$$

The Johnson-Lindenstrauss lemma

- For any $\varepsilon \in (0,1/2)$ and any integer m, then for integer $k = O\left(\frac{1}{\varepsilon^2}\ln m\right)$ and any points $x_1, x_2, \ldots, x_m \in \mathbb{R}^d$, there exists a **linear map** (matrix) $L: \mathbb{R}^d \to \mathbb{R}^k$ such that for any $1 \le i < j \le m$, it holds
$$(1-\varepsilon)\left\|x_i - x_j\right\|_2^2 \le \left\|Lx_i - Lx_j\right\|_2^2 \le (1+\varepsilon)\left\|x_i - x_j\right\|_2^2$$
- The linear transformation L is simply **multiplication** by a matrix whose entries are **sampled independently from a standard Gaussian**, **scaled appropriately**

The Johnson-Lindenstrauss lemma

- For any $\varepsilon \in (0,1/2)$ and any integer m, then for integer $k = O\left(\frac{1}{\varepsilon^2} \ln m\right)$ and any points $x_1, x_2, \ldots, x_m \in \mathbb{R}^d$, there exists a **linear map** (matrix) $L: \mathbb{R}^d \to \mathbb{R}^k$ such that for any $1 \leq i < j \leq m$, it holds
$$(1-\varepsilon)\left\|x_i - x_j\right\|_2^2 \leq \left\|Lx_i - Lx_j\right\|_2^2 \leq (1+\varepsilon)\left\|x_i - x_j\right\|_2^2$$
- The linear transformation L is simply **multiplication** by a matrix whose entries are **sampled independently from a standard Gaussian**, **scaled appropriately**
- Let A be random $k \times d$ matrix with $A_{i,j} \sim \mathcal{N}(0,1)$, independently from the other entries
- Set $L = \frac{1}{\sqrt{k}} A$

Useful properties

- Let $X \sim \mathcal{N}(0, \sigma_1^2)$ and $Y \sim \mathcal{N}(0, \sigma_2^2)$ and a, b are any constants
- Then, a$X + bY \sim \mathcal{N}(0, a^2\sigma_1^2 + b^2\sigma_2^2)$

Lemma: For unit vector v, $||Av||_2^2$ is distributed as a sum of i.i.d. squared standard Gaussians

- Let $v \in \mathbb{R}^d$ be a **unit vector**
- Let A be a **random $\boldsymbol{k \times d}$ matrix** with $\boldsymbol{A_{i,j} \sim \mathcal{N}(0,1)}$ independently of the other entries
- Then, the **squared norm $\boldsymbol{||Av||_2^2}$** behaves as a sum of k squared standard Gaussians

Proof

- Observe that $||Av||_2^2 = \sum_{i=1}^{k}\left(\sum_{j=1}^{d} A_{i,j} v_j\right)^2$
- By the **properties of the Gaussian** p.d., $\sum_{j=1}^{d} A_{i,j} v_j \sim \mathcal{N}(0, \sum_{j=1}^{d} v_j^2)$
- But $\sum_{j=1}^{d} v_j^2 = 1$ since x is a **unit vector**
- Hence, $||Av||_2^2$ is the sum of k squared standard Gaussian i.i.d r.v.'s

Lemma: Sums of i.i.d. squared gaussians are sharply concentrated around their expectation

- Let $Z_1, Z_2, \ldots, Z_k$ be k independent and identically distributed guassian random variables with zero mean and standard deviation 1, i.e., $Z_i \sim \mathcal{N}(0,1)$ for $i = 1, .., k$.
- Define $Q = \sum_{i=1}^{k} Z_i^2$
- Then, $\mathbb{E}[Q] = k$, and for $\eta \in [0,1/2]$, $\Pr[|Q - k| \geq \eta k] \leq 2\exp(-\eta^2 k/8)$

The easy part of the proof: $\mathbb{E}[Q] = k$

- By the definition of the **variance** σ^2 of the normal r.v. $Z_i \sim \mathcal{N}(\mu, \sigma^2)$, we have $\mathbb{E}[(Z_i - \mu)^2] = \sigma^2$
- Hence, when $Z_i \sim \mathcal{N}(0,1)$, we have $\mathbb{E}\left[Z_i^2\right] = 1$
- By **linearity of expectation**: $\mathbb{E}[Q] = \mathbb{E}\left[\sum_{i=1}^{k} Z_i^2\right] = \sum_{i=1}^{k} \mathbb{E}\left[Z_i^2\right] = k$

The difficult part of the proof:

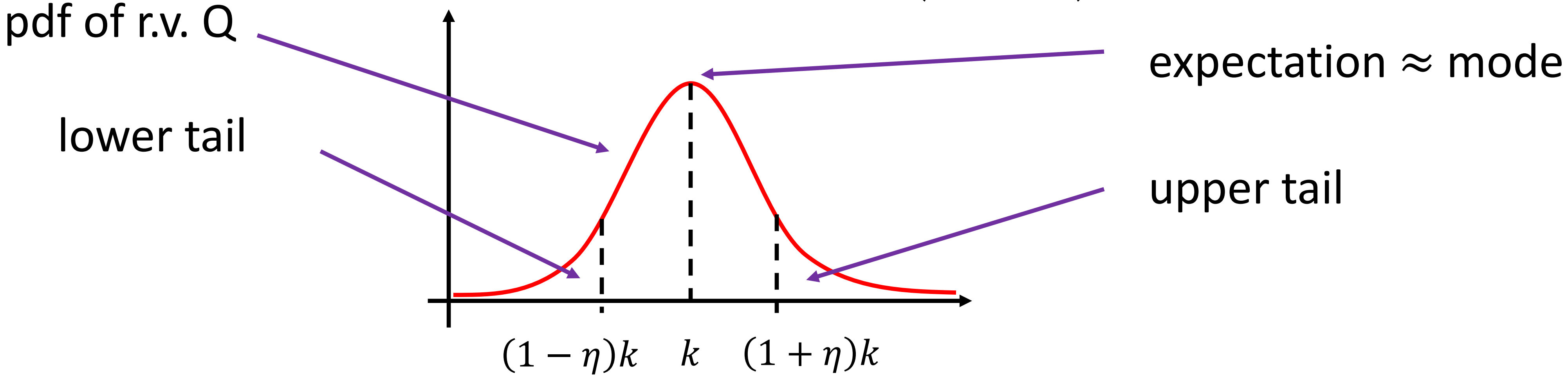
$$\Pr[|Q-k| \geq \eta k] \leq 2\exp(-\eta^2 k/8)$$

- The proof will follow by proving

$$\Pr[Q \geq (1+\eta)k] \leq \exp\left(-\frac{\eta^2 k}{8}\right)$$

and

$$\Pr[Q \leq (1-\eta)k] \leq \exp\left(-\frac{\eta^2 k}{8}\right)$$



Proof of part 1: $\Pr[Q \geq (1+\eta)k] \leq \exp\left(-\frac{\eta^2 k}{8}\right)$

- Let $\lambda > 0$

$$\Pr[Q \geq (1+\eta)k] = \Pr[\exp(\lambda Q) \geq \exp(\lambda(1+\eta)k)]$$

Proof of part 1: $\Pr[Q \geq (1+\eta)k] \leq \exp\left(-\frac{\eta^2 k}{8}\right)$

- Let $\lambda > 0$. Then, using **Markov inequality**, we get

$$\Pr[Q \geq (1+\eta)k] = \Pr[\exp(\lambda Q) \geq \exp(\lambda(1+\eta)k)] \leq \frac{\mathbb{E}[\exp(\lambda Q)]}{\exp(\lambda(1+\eta)k)}$$

Proof of part 1: $\Pr[Q \geq (1+\eta)k] \leq \exp\left(-\frac{\eta^2 k}{8}\right)$

- Let $\lambda > 0$. Then, using **Markov inequality**, we get

$$\Pr[Q \geq (1+\eta)k] = \Pr[\exp(\lambda Q) \geq \exp(\lambda(1+\eta)k)] \leq \frac{\mathbb{E}[\exp(\lambda Q)]}{\exp(\lambda(1+\eta)k)}$$

- The numerator becomes

$$\mathbb{E}[\exp(\lambda Q)] = \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{k} Z_i^2\right)\right] = \mathbb{E}\left[\prod_{i=1}^{k} \exp\left(\lambda Z_i^2\right)\right]$$

$Z_1, Z_2, \dots, Z_k$ are **independent** $\longrightarrow$

$$= \prod_{i=1}^{k} \mathbb{E}\left[\exp\left(\lambda Z_i^2\right)\right] = \left(\mathbb{E}[\exp(\lambda Z_1^2)]\right)^k$$

Proof of part 1: $\Pr[Q \geq (1+\eta)k] \leq \exp\left(-\frac{\eta^2 k}{8}\right)$

- Let $\lambda > 0$. Then, using **Markov inequality**, we get

$$\Pr[Q \geq (1+\eta)k] = \Pr[\exp(\lambda Q) \geq \exp(\lambda(1+\eta)k)] \leq \frac{\mathbb{E}[\exp(\lambda Q)]}{\exp(\lambda(1+\eta)k)}$$

- The numerator becomes

$$\mathbb{E}[\exp(\lambda Q)] = \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{k} Z_i^2\right)\right] = \mathbb{E}\left[\prod_{i=1}^{k} \exp\left(\lambda Z_i^2\right)\right]$$

$Z_1, Z_2, \ldots, Z_k$ are **independent** $\longrightarrow$ $= \prod_{i=1}^{k} \mathbb{E}\left[\exp\left(\lambda Z_i^2\right)\right] = \left(\mathbb{E}\left[\exp\left(\lambda Z_1^2\right)\right]\right)^k$

- Using the definition of the expectation,

$$\mathbb{E}[\exp(\lambda Z_1^2)] = \int_{-\infty}^{+\infty} f(t)\exp(\lambda t^2)\, dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{t^2}{2}(1-2\lambda)\right) dt = \frac{1}{\sqrt{1-2\lambda}}$$

Proof of part 1: $\Pr[Q \geq (1+\eta)k] \leq \exp\left(-\frac{\eta^2 k}{8}\right)$

- So,

$$\Pr[Q \geq (1+\eta)k] \leq (1-2\lambda)^{-k/2}\exp(-\lambda(1+\eta)k)$$

Proof of part 1: $\Pr[Q \geq (1+\eta)k] \leq \exp\left(-\frac{\eta^2 k}{8}\right)$

- So,

$$\Pr[Q \geq (1+\eta)k] \leq (1-2\lambda)^{-k/2}\exp(-\lambda(1+\eta)k)$$

- **Selecting** $\lambda = \frac{\eta}{2(1+\eta)}$ (this is the value of λ that minimizes the RHS above), we have

$$\Pr[Q \geq (1+\eta)k] \leq (1+\eta)^{k/2}\exp(-\eta k/2)$$

Proof of part 1: $\Pr[Q \geq (1+\eta)k] \leq \exp\left(-\frac{\eta^2 k}{8}\right)$

- So,

$$\Pr[Q \geq (1+\eta)k] \leq (1-2\lambda)^{-k/2} \exp(-\lambda(1+\eta)k)$$

- **Selecting** $\lambda = \frac{\eta}{2(1+\eta)}$ (this is the value of λ that minimizes the RHS above), we have

$$\Pr[Q \geq (1+\eta)k] \leq (1+\eta)^{k/2} \exp(-\eta k/2)$$

- Note that $1+\eta \leq \exp\left(\eta - \frac{\eta^2}{4}\right)$ for $\eta \in [0,1/2]$. Hence,

$$\Pr[Q \geq (1+\eta)k] \leq \exp(-\eta^2 k/8)$$

Proof of part 1: $\Pr[Q \geq (1+\eta)k] \leq \exp\left(-\frac{\eta^2 k}{8}\right)$

- So,

$$\Pr[Q \geq (1+\eta)k] \leq (1-2\lambda)^{-k/2}\exp(-\lambda(1+\eta)k)$$

- **Selecting** $\lambda = \frac{\eta}{2(1+\eta)}$ (this is the value of λ that minimizes the RHS above), we have

$$\Pr[Q \geq (1+\eta)k] \leq (1+\eta)^{k/2}\exp(-\eta k/2)$$

- Note that $1+\eta \leq \exp\left(\eta - \frac{\eta^2}{4}\right)$ for $\eta \in [0,1/2]$. Hence,

$$\Pr[Q \geq (1+\eta)k] \leq \exp(-\eta^2 k/8)$$

- The proof of part 2 is similar

QED

Summarizing up to now

- For unit vector v, $\|Av\|_2^2$ is distributed as a sum of k i.i.d. squared standard Gaussians
- Hence, $\Pr\left[\left|\|Av\|_2^2 - k\right| \geq \eta k\right] \leq 2\exp(-\eta^2 k/8)$
- Since $L = \frac{1}{\sqrt{k}} A$, this is equivalent to
$$\Pr\left[\left|\|Lv\|_2^2 - 1\right| \geq \eta\right] \leq 2\exp(-\eta^2 k/8)$$
- I.e., ***L* does not distort the squared norm of the unit vector *v*** by much

Lemma: It suffices to focus on unit vectors

- For $1 \leq i < j \leq m$, denote by v_{ij} the unit vector $v_{ij} = \frac{x_i - x_j}{\|x_i - x_j\|}$
- Assume that matrix L is such that $1 - \varepsilon \leq \|Lv_{ij}\|_2^2 \leq 1 + \varepsilon$, for $1 \leq i < j \leq m$
- Then, $(1 - \varepsilon)\|x_i - x_j\|_2^2 \leq \|Lx_i - Lx_j\|_2^2 \leq (1 + \varepsilon)\|x_i - x_j\|_2^2$, for $1 \leq i < j \leq m$

Lemma: It suffices to focus on unit vectors

- For $1 \leq i < j \leq m$, denote by v_{ij} the unit vector $v_{ij} = \frac{x_i - x_j}{\|x_i - x_j\|}$
- Assume that matrix L is such that $1 - \varepsilon \leq \|Lv_{ij}\|_2^2 \leq 1 + \varepsilon$, for $1 \leq i < j \leq m$
- Then, $(1 - \varepsilon)\|x_i - x_j\|_2^2 \leq \|Lx_i - Lx_j\|_2^2 \leq (1 + \varepsilon)\|x_i - x_j\|_2^2$, for $1 \leq i < j \leq m$

- Proof: Notice that

$$\|Lx_i - Lx_j\|_2^2 = \|L(x_i - x_j)\|_2^2 = \left\| \|x_i - x_j\| L \frac{x_i - x_j}{\|x_i - x_j\|} \right\|_2^2 = \|x_i - x_j\|_2^2 \cdot \|Lv_{ij}\|_2^2$$

Lemma: It suffices to focus on unit vectors

- For $1 \leq i < j \leq m$, denote by v_{ij} the unit vector $v_{ij} = \frac{x_i - x_j}{\|x_i - x_j\|}$
- Assume that matrix L is such that $1 - \varepsilon \leq \left\|Lv_{ij}\right\|_2^2 \leq 1 + \varepsilon$, for $1 \leq i < j \leq m$
- Then, $(1-\varepsilon)\left\|x_i - x_j\right\|_2^2 \leq \left\|Lx_i - Lx_j\right\|_2^2 \leq (1+\varepsilon)\left\|x_i - x_j\right\|_2^2$, for $1 \leq i < j \leq m$

- Proof: Notice that

$$\left\|Lx_i - Lx_j\right\|_2^2 = \left\|L(x_i - x_j)\right\|_2^2 = \left\|\left\|x_i - x_j\right\| L \frac{x_i - x_j}{\left\|x_i - x_j\right\|}\right\|_2^2 = \left\|x_i - x_j\right\|_2^2 \cdot \left\|Lv_{ij}\right\|_2^2$$

- Hence $(1-\varepsilon)\left\|x_i - x_j\right\|_2^2 \leq \left\|Lx_i - Lx_j\right\|_2^2 \leq (1+\varepsilon)\left\|x_i - x_j\right\|_2^2$

Final push: Proof of JL lemma

- So, we know that if $\left|\|Lv_{ij}\|_2^2 - 1\right| \le \varepsilon$ for the $m(m-1)/2$ unit vectors v_{ij}, then $(1-\varepsilon)\|x_i - x_j\|_2^2 \le \|Lx_i - Lx_j\|_2^2 \le (1+\varepsilon)\|x_i - x_j\|_2^2$, for $1 \le i < j \le m$

Final push: Proof of JL lemma

- So, we know that if $\left|\left\|Lv_{ij}\right\|_2^2 - 1\right| \leq \varepsilon$ for the $m(m-1)/2$ unit vectors v_{ij}, then $(1-\varepsilon)\left\|x_i - x_j\right\|_2^2 \leq \left\|Lx_i - Lx_j\right\|_2^2 \leq (1+\varepsilon)\left\|x_i - x_j\right\|_2^2$, for $1 \leq i < j \leq m$
- We have shown that $\Pr\left[\left|\left\|Lv_{ij}\right\|_2^2 - 1\right| \geq \varepsilon\right] \leq 2\exp(-\varepsilon^2 k/8)$

Final push: Proof of JL lemma

- So, we know that if $\left|\|Lv_{ij}\|_2^2 - 1\right| \leq \varepsilon$ for the $m(m-1)/2$ unit vectors v_{ij}, then $(1-\varepsilon)\|x_i - x_j\|_2^2 \leq \|Lx_i - Lx_j\|_2^2 \leq (1+\varepsilon)\|x_i - x_j\|_2^2$, for $1 \leq i < j \leq m$
- We have shown that $\Pr\left[\left|\|Lv_{ij}\|_2^2 - 1\right| \geq \varepsilon\right] \leq 2\exp(-\varepsilon^2 k/8)$
- Selecting $k = 24\varepsilon^{-2} \ln m$, we have that $\Pr\left[\left|\|Lv_{ij}\|_2^2 - 1\right| \geq \varepsilon\right] \leq \frac{2}{m^3}$

Final push: Proof of JL lemma

- So, we know that if $\left|\left\|Lv_{ij}\right\|_2^2 - 1\right| \leq \varepsilon$ for the $m(m-1)/2$ unit vectors v_{ij}, then $(1-\varepsilon)\left\|x_i - x_j\right\|_2^2 \leq \left\|Lx_i - Lx_j\right\|_2^2 \leq (1+\varepsilon)\left\|x_i - x_j\right\|_2^2$, for $1 \leq i < j \leq m$
- We have shown that $\Pr\left[\left|\left\|Lv_{ij}\right\|_2^2 - 1\right| \geq \varepsilon\right] \leq 2\exp(-\varepsilon^2 k/8)$
- Selecting $k = 24\varepsilon^{-2} \ln m$, we have that $\Pr\left[\left|\left\|Lv_{ij}\right\|_2^2 - 1\right| \geq \varepsilon\right] \leq \frac{2}{m^3}$
- Using the union bound, we have $\Pr\left[\exists i, j\colon \left|\left\|Lv_{ij}\right\|_2^2 - 1\right| \geq \varepsilon\right] \leq \frac{1}{m}$

Final push: Proof of JL lemma

- So, we know that if $\left|\|Lv_{ij}\|_2^2 - 1\right| \leq \varepsilon$ for the $m(m-1)/2$ unit vectors v_{ij}, then $(1-\varepsilon)\|x_i - x_j\|_2^2 \leq \|Lx_i - Lx_j\|_2^2 \leq (1+\varepsilon)\|x_i - x_j\|_2^2$, for $1 \leq i < j \leq m$
- We have shown that $\Pr\left[\left|\|Lv_{ij}\|_2^2 - 1\right| \geq \varepsilon\right] \leq 2\exp(-\varepsilon^2 k/8)$
- Selecting $k = 24\varepsilon^{-2} \ln m$, we have that $\Pr\left[\left|\|Lv_{ij}\|_2^2 - 1\right| \geq \varepsilon\right] \leq \frac{2}{m^3}$
- Using the union bound, we have $\Pr\left[\exists i,j\colon \left|\|Lv_{ij}\|_2^2 - 1\right| \geq \varepsilon\right] \leq \frac{1}{m}$
- Equivalently, $\Pr\left[\forall i,j\colon \left|\|Lv_{ij}\|_2^2 - 1\right| < \varepsilon\right] \geq 1 - \frac{1}{m}$

Final push: Proof of JL lemma

- So, we know that if $\left|\left\|Lv_{ij}\right\|_2^2 - 1\right| \leq \varepsilon$ for the $m(m-1)/2$ unit vectors v_{ij}, then $(1-\varepsilon)\left\|x_i - x_j\right\|_2^2 \leq \left\|Lx_i - Lx_j\right\|_2^2 \leq (1+\varepsilon)\left\|x_i - x_j\right\|_2^2$, for $1 \leq i < j \leq m$
- We have shown that $\Pr\left[\left|\left\|Lv_{ij}\right\|_2^2 - 1\right| \geq \varepsilon\right] \leq 2\exp(-\varepsilon^2 k/8)$
- Selecting $k = 24\varepsilon^{-2} \ln m$, we have that $\Pr\left[\left|\left\|Lv_{ij}\right\|_2^2 - 1\right| \geq \varepsilon\right] \leq \frac{2}{m^3}$
- Using the union bound, we have $\Pr\left[\exists i,j\colon \left|\left\|Lv_{ij}\right\|_2^2 - 1\right| \geq \varepsilon\right] \leq \frac{1}{m}$
- Equivalently, $\Pr\left[\forall i,j\colon \left|\left\|Lv_{ij}\right\|_2^2 - 1\right| < \varepsilon\right] \geq 1 - \frac{1}{m}$
- Hence, with probability at least $1 - 1/m$, we get that, for $1 \leq i < j \leq m$,

$$(1-\varepsilon)\left\|x_i - x_j\right\|_2^2 \leq \left\|Lx_i - Lx_j\right\|_2^2 \leq (1+\varepsilon)\left\|x_i - x_j\right\|_2^2$$

QED

An application of JL lemma

k-means clustering

- Input: An integer k and n points $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$
- Objective: Select k cluster centers $c_1, c_2, \ldots, c_k$ so that the sum of squared distances of the points to their nearest center

$$\sum_{i=1}^{n} \min_{j} \left\| x_i - c_j \right\|_2^2$$

is minimized

k-means clustering

- Input: An integer k and n points $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$
- Objective: Select k cluster centers $c_1, c_2, \ldots, c_k$ so that the sum of squared distances of the points to their nearest center

$$\sum_{i=1}^{n} \min_j \left\| x_i - c_j \right\|_2^2$$

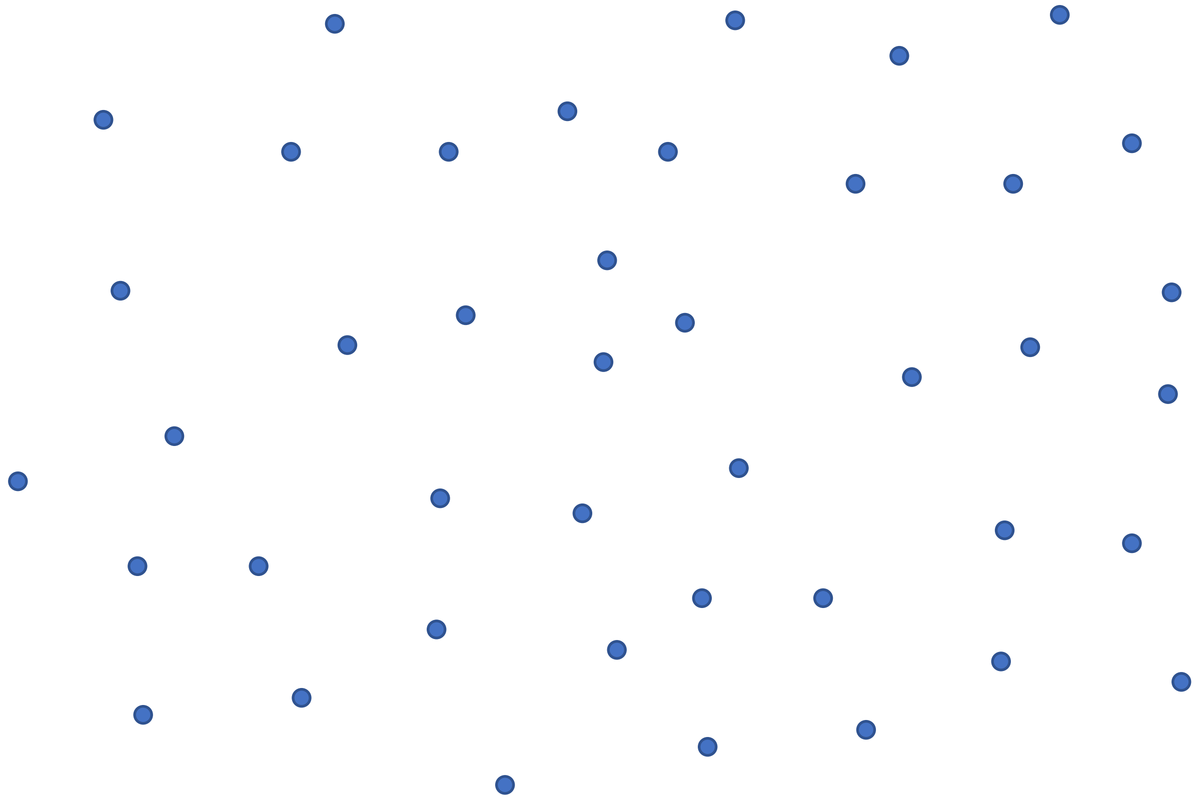
is minimized

sum over all points

minimum squared distance of the point from the closest cluster center

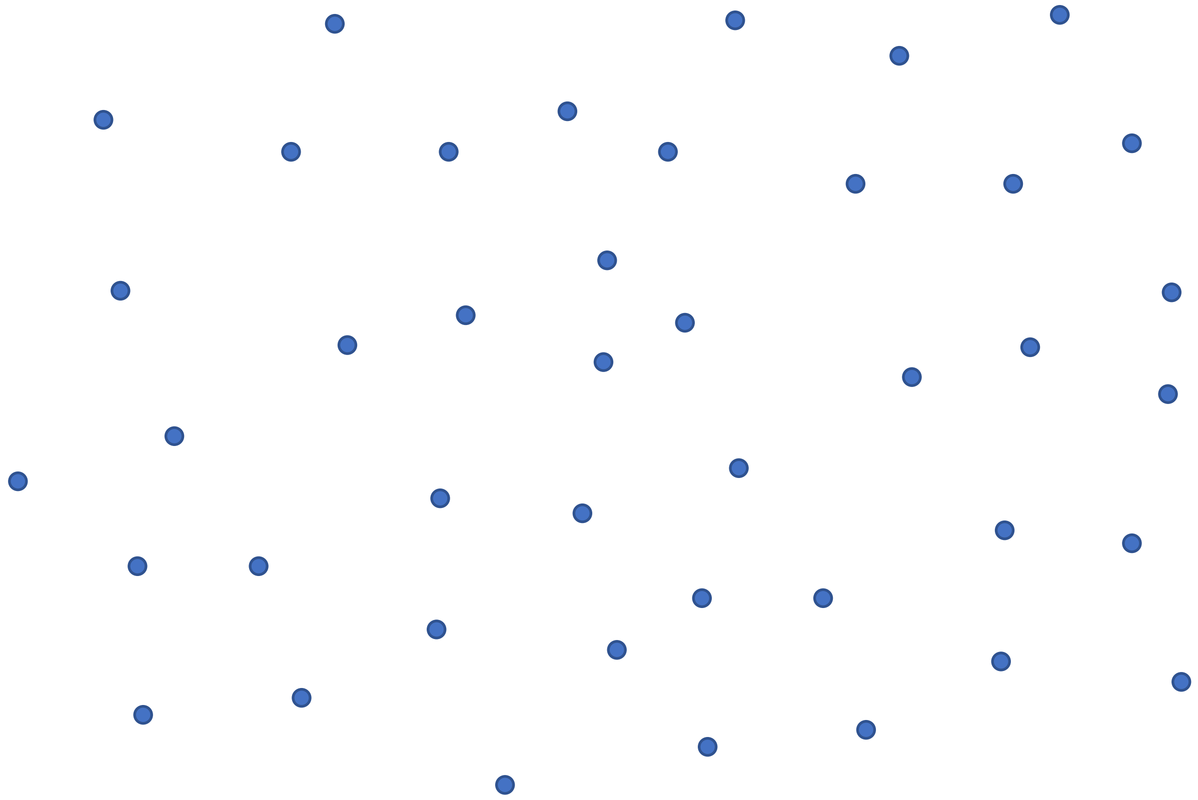
Example

- Points in $\mathbb{R}^2$
- $k = 5$



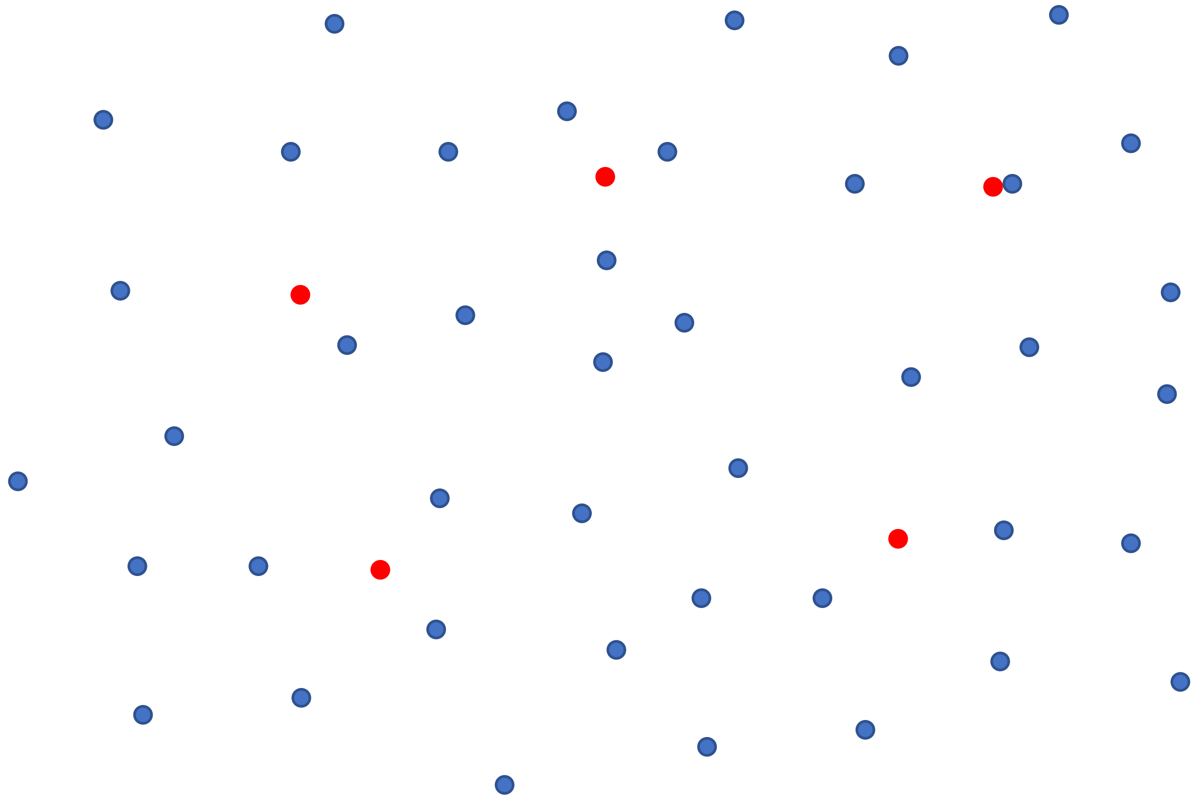
Example

- Points in $\mathbb{R}^2$
- $k = 5$
- Solution?



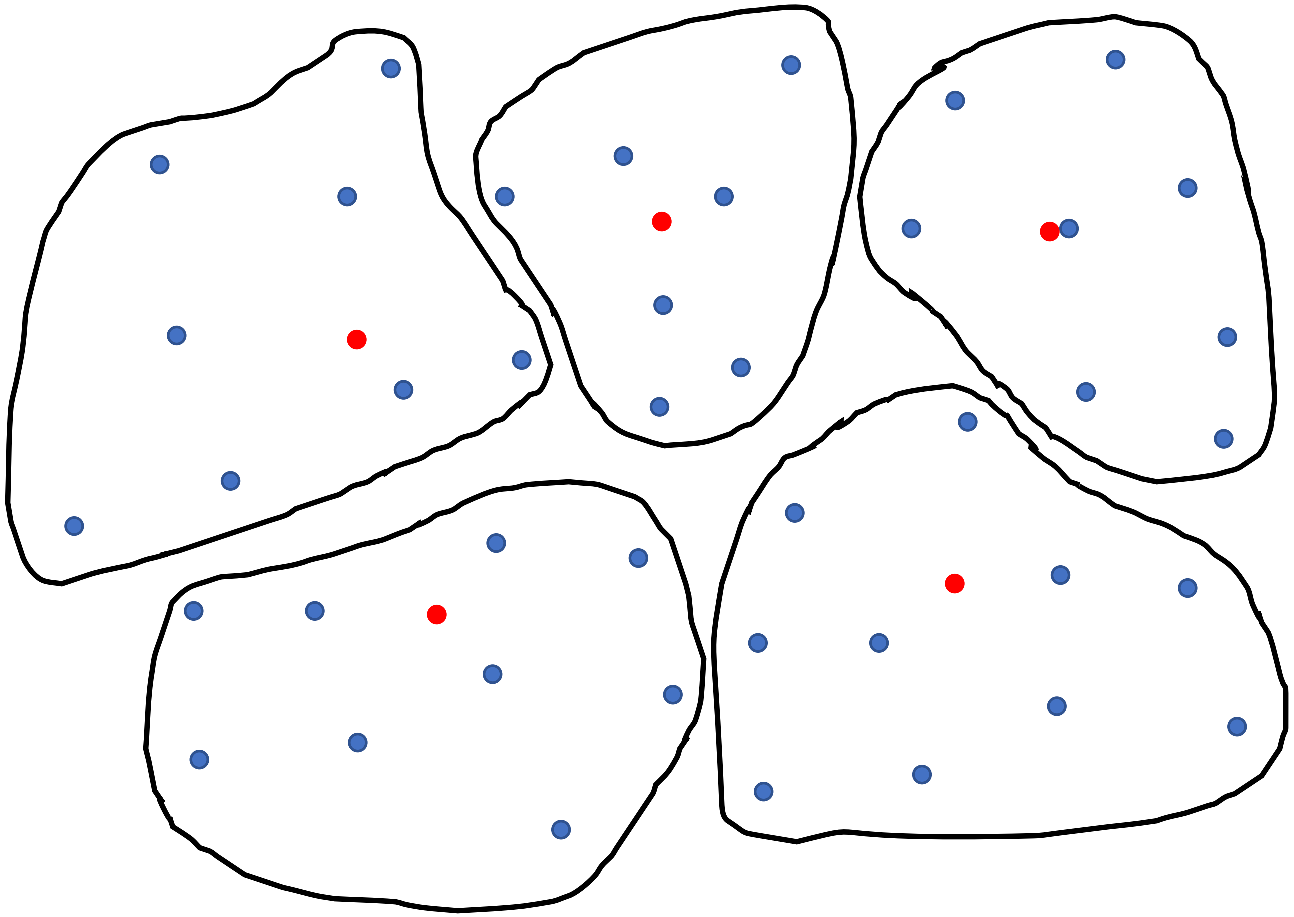
Example

- Points in $\mathbb{R}^2$
- $k = 5$
- Solution?
- Spread cluster centers



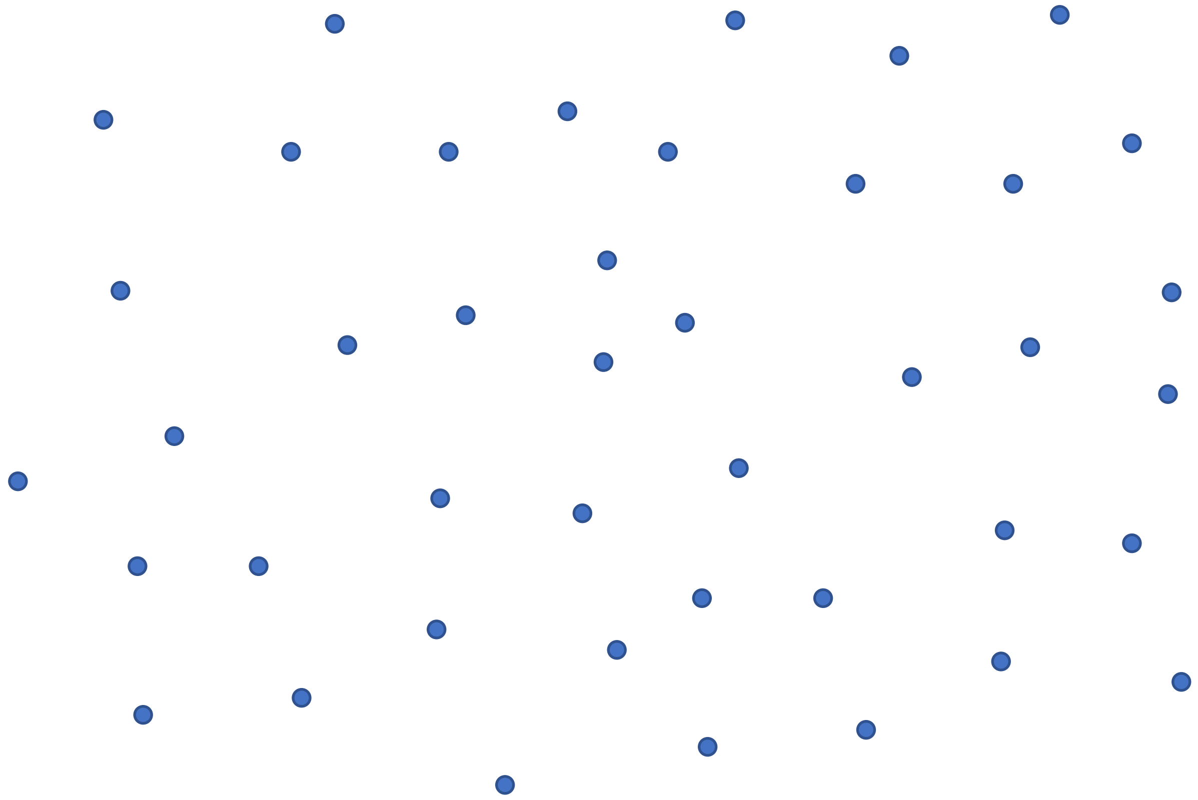
Example

- Points in $\mathbb{R}^2$
- $k = 5$
- Solution?
- Spread cluster centers
- Connect points to the closest cluster center



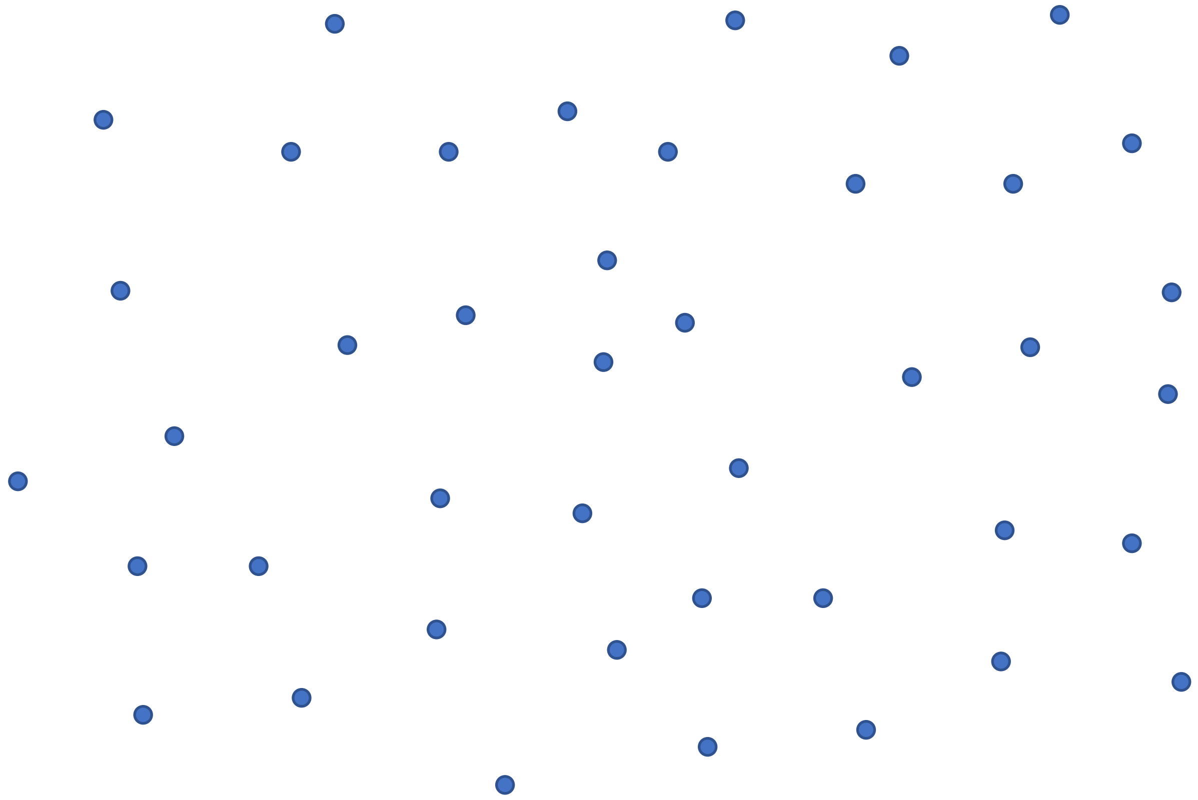
Example

- Points in $\mathbb{R}^2$
- $k = 5$
- Solution?



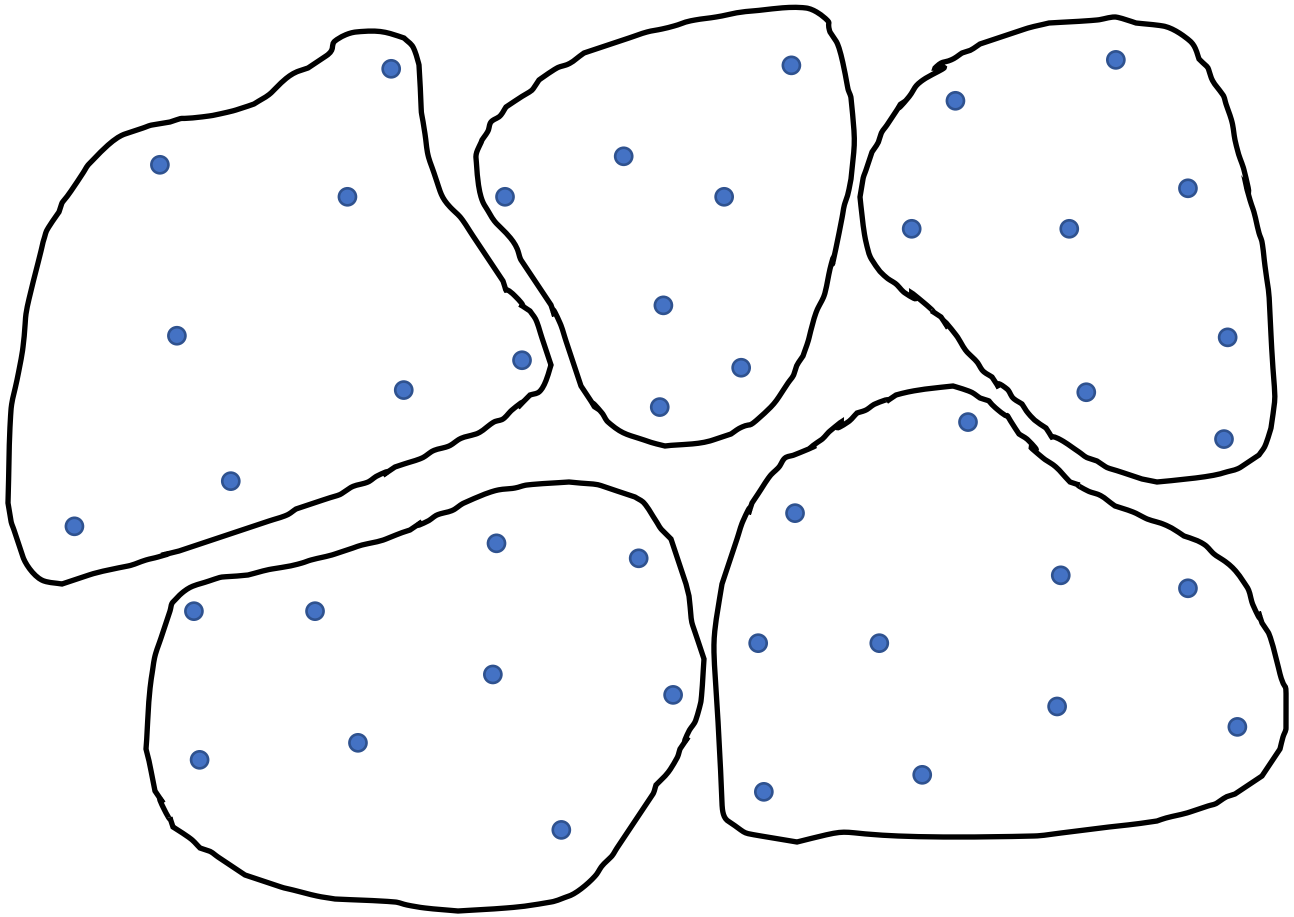
Example

- Points in $\mathbb{R}^2$
- $k = 5$
- Solution?
- Better idea: define the clusters of points first



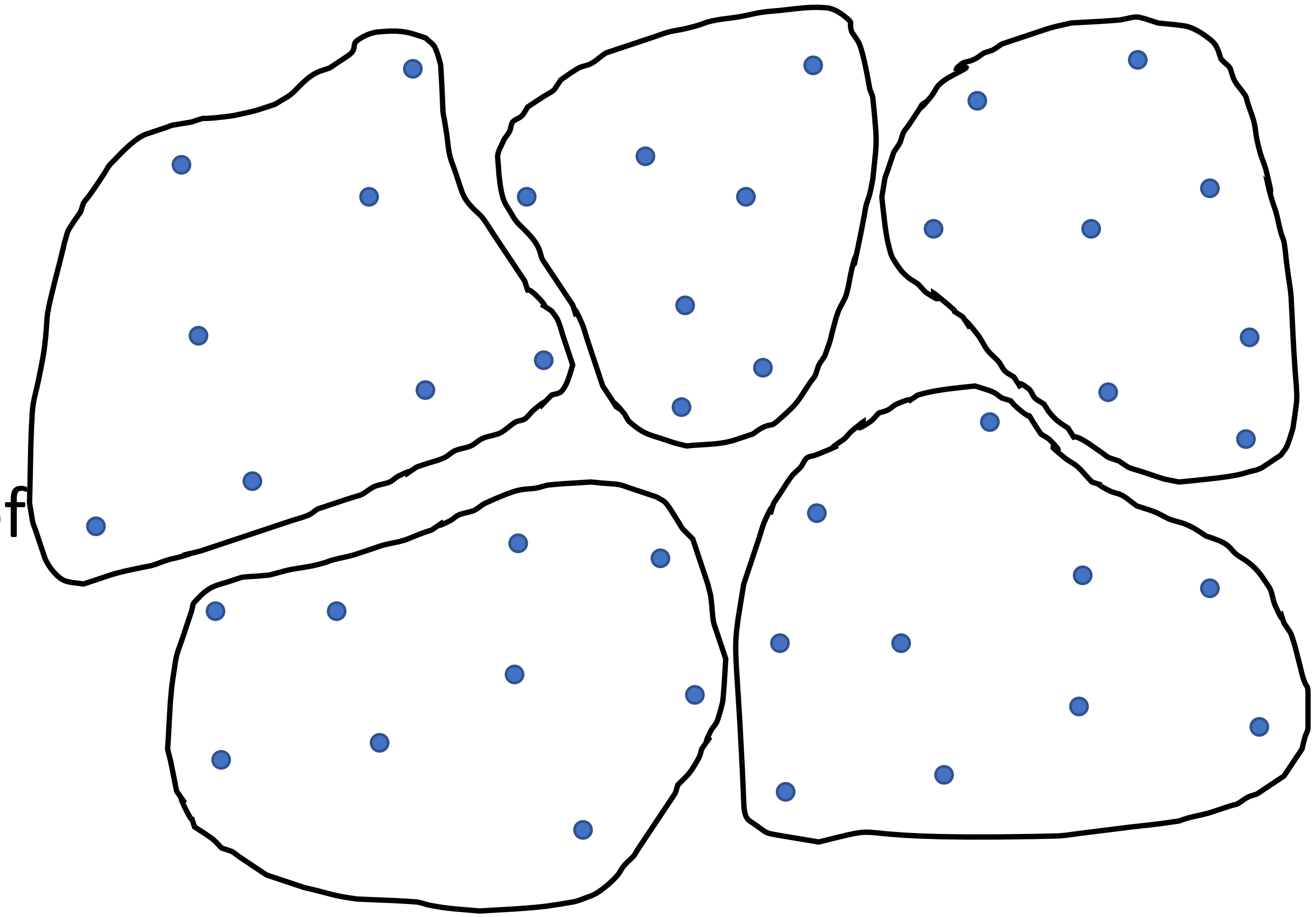
Example

- Points in $\mathbb{R}^2$
- $k = 5$
- Solution?
- Better idea: define the clusters of points first



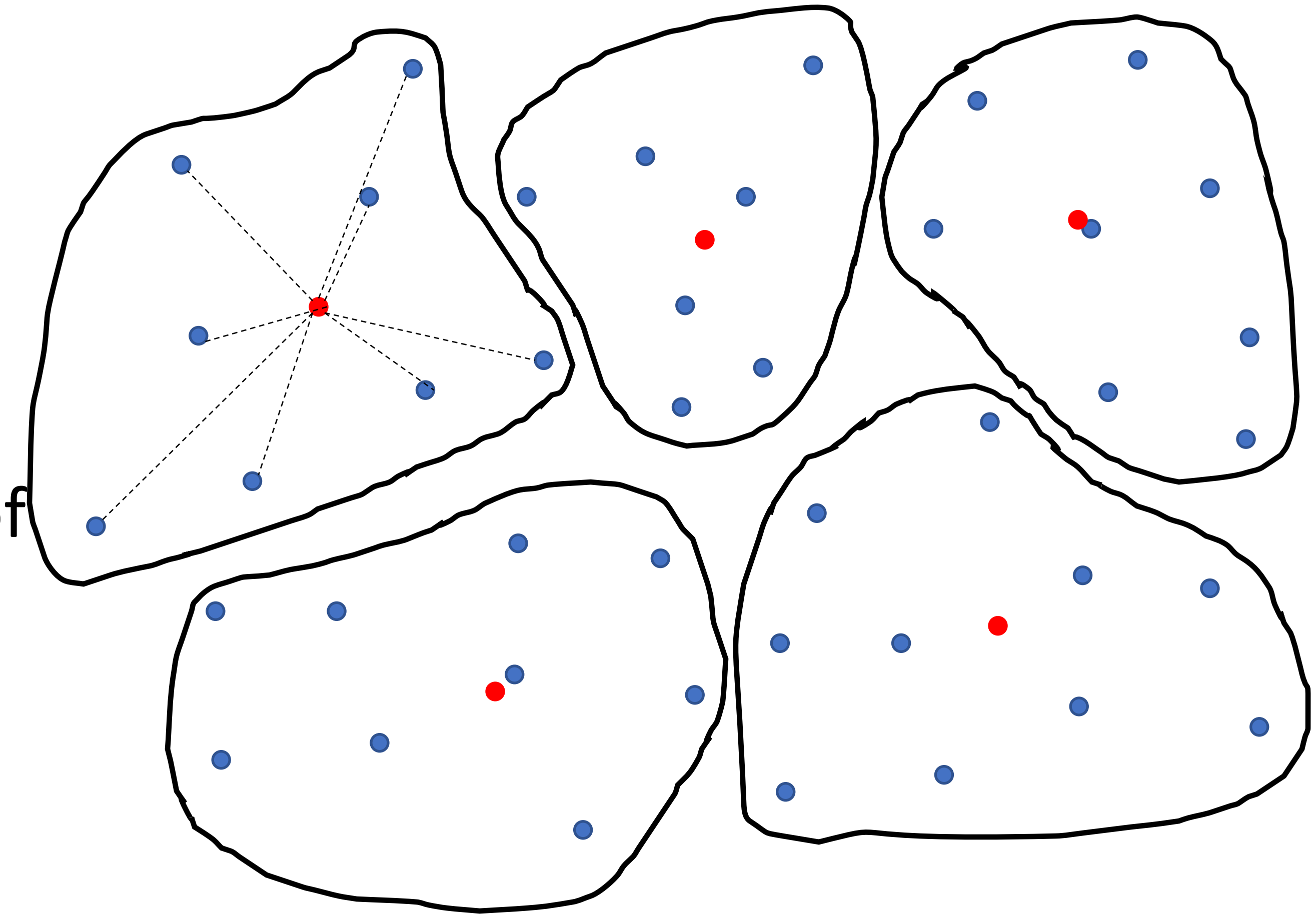
Example

- Points in $\mathbb{R}^2$
- $k = 5$
- Solution?
- Better idea: define the clusters of points first
- Computer the center of each cluster optimally



Example

- Points in $\mathbb{R}^2$
- $k = 5$
- Solution?
- Better idea: define the clusters of points first
- Computer the center of each cluster optimally



Selecting cluster centers optimally

- Cluster with the set X_j of points
- Where should the cluster center c_j be so that $\sum_{i:x_i \in X_j} \left\| x_i - c_j \right\|_2^2$ is minimized?
- Answer: it should be the mean of the points in the cluster, i.e.,

$$c_j = \frac{1}{|X_j|} \sum_{i:x_i \in X_j} x_i$$

Proof? Nullify the derivatives of $\sum_{i:x_i \in X_j} \sum_{t=1}^{d} (x_{i,t} - c_{j,t})^2$ with respect to $c_{j,t}$ for $t = 1, \dots, d$

k-means clustering (alternative definition)

- Input: An integer k and n points $x_1, x_2, \dots, x_n \in \mathbb{R}^d$
- Objective: Select k clusters $X_1, X_2, \dots, X_k$ so that the quantity

$$\sum_{j=1}^{k} \sum_{i : x_i \in X_j} \left\| x_i - \frac{1}{|X_j|} \sum_{h : x_h \in X_j} x_h \right\|_2^2$$

  is minimized

k-means clustering (alternative definition)

- Input: An integer k and n points $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$
- Objective: Select k clusters $X_1, X_2, \ldots, X_k$ so that the quantity

$$\sum_{j=1}^{k} \sum_{i:x_i \in X_j} \left\| x_i - \underbrace{\frac{1}{|X_j|} \sum_{h:x_h \in X_j} x_h}_{\text{optimal cluster center}} \right\|_2^2$$

squared distance of point from center

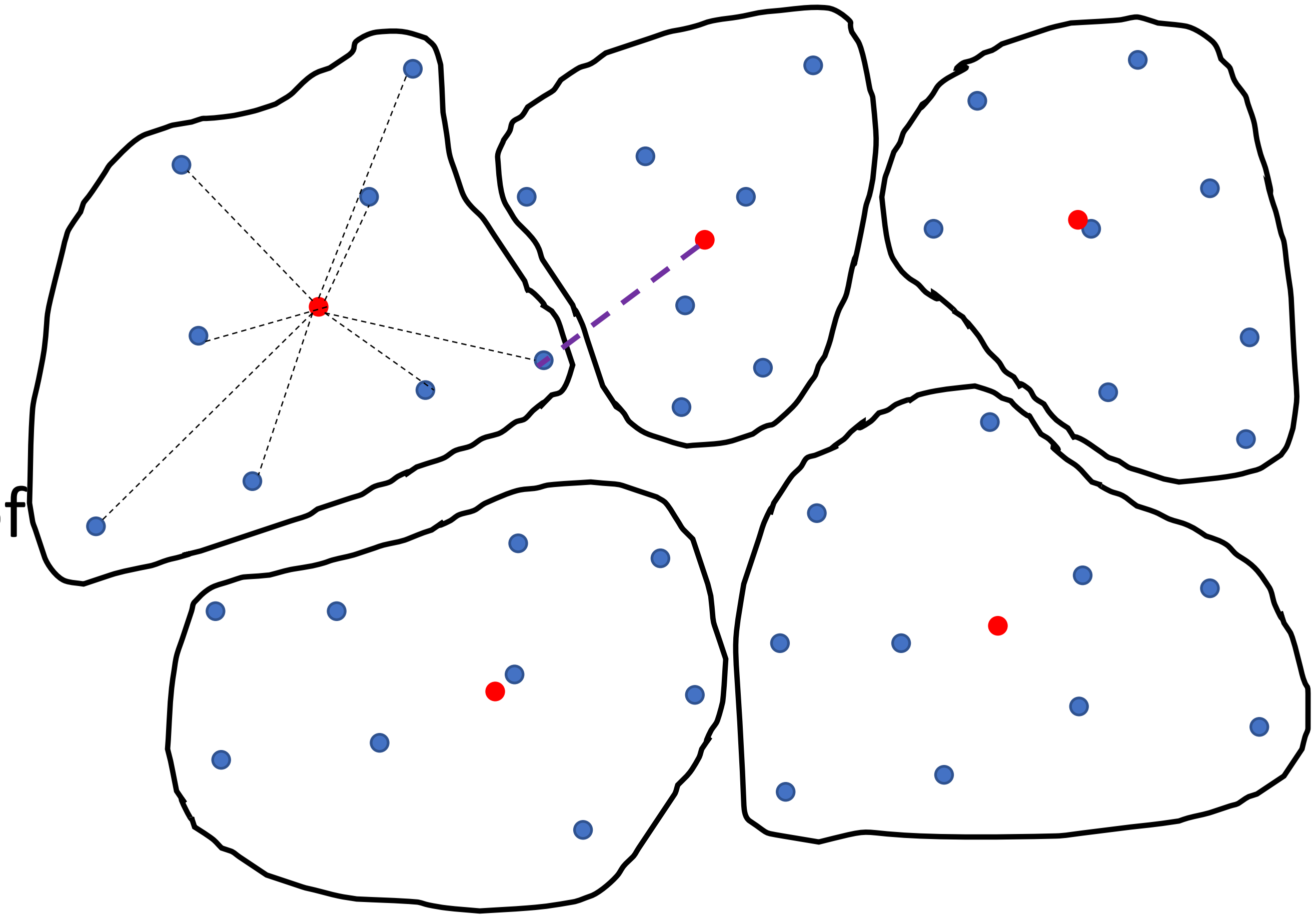
is minimized

sum over all points in the cluster

sum over all clusters

Back to the example

- Points in $\mathbb{R}^2$
- $k = 5$
- Solution?
- Better idea: define the clusters of points first
- Computer the center of each cluster optimally
- Possible **issue**: a point may not belong to the cluster of its closest center



Lloyd's algorithm

- Start by partitioning the points into k clusters
- Repeat
  - Compute optimal centers
  - Reassign points to the cluster of the closest center
- Until no change in clusters

- Running time: $O(tndk)$, where t = #iterations
- Time $O(nd)$ for computing the new centers in each iteration
- Time $O(ndk)$ for reassigning points in each iteration

Speeding up Lloyd’s algorithm using JL transform

- Idea: Reduce dimensionality by applying JL transform and apply Lloyd’s alg
- n points in $\mathbb{R}^d$
- JL uses matrix $A \in \mathbb{R}^{m \times d}$ with $m = O(\varepsilon^{-2} \ln n)$
- Multiplying each point with the random matrix takes $O(nd\varepsilon^{-2} \ln n)$ time
- Time $O(n\varepsilon^{-2} \ln n)$ for computing the new centers in each iteration
- Time $O(nk\varepsilon^{-2} \ln n)$ for reassigning points in each iteration
- Overall running time: $O(n\varepsilon^{-2}(d + kt) \ln n)$
- Can be better, depending on the parameters
- Clustering is almost as good as the one computed on the original data

Lemma: The cost of the clustering can be written as a sum of pairwise distances

- For any set of points $x_1, x_2, \ldots, x_n$ and a partionining of the points into clusters $X_1, X_2, \ldots, X_k$, the cost of the clustering satisfies

$$\sum_{j=1}^{k} \sum_{i:x_i \in X_j} \left\| x_i - \frac{1}{|X_j|} \sum_{h:x_h \in X_j} x_h \right\|_2^2 = \frac{1}{2} \sum_{j=1}^{k} \frac{1}{|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \|x_i - x_h\|_2^2$$

Proof

Proof

$$\sum_{j=1}^{k} \sum_{i:x_i \in X_j} \left\| x_i - \frac{1}{|X_j|} \sum_{h:x_h \in X_j} x_h \right\|_2^2$$

Proof

$$\sum_{j=1}^{k} \sum_{i:x_i \in X_j} \left\| x_i - \frac{1}{|X_j|} \sum_{h:x_h \in X_j} x_h \right\|_2^2$$
$$= \sum_{j=1}^{k} \sum_{i:x_i \in X_j} \left(\|x_i\|_2^2 - \frac{2}{|X_j|} \sum_{h:x_h \in X_j} \langle x_i, x_h \rangle + \frac{1}{|X_j|^2} \left\| \sum_{h:x_h \in X_j} x_h \right\|_2^2 \right)$$

Proof

$$\sum_{j=1}^{k} \sum_{i:x_i \in X_j} \left\| x_i - \frac{1}{|X_j|} \sum_{h:x_h \in X_j} x_h \right\|_2^2$$

$$= \sum_{j=1}^{k} \sum_{i:x_i \in X_j} \left(\|x_i\|_2^2 - \frac{2}{|X_j|} \sum_{h:x_h \in X_j} \langle x_i, x_h \rangle + \frac{1}{|X_j|^2} \left\| \sum_{h:x_h \in X_j} x_h \right\|_2^2 \right)$$

- We will work with each of the three terms in parenthesis separately

Proof

first term $$\sum_{i:x_i \in X_i} \|x_i\|_2^2$$

Proof

first term $$\sum_{i:x_i \in X_j} \|x_i\|_2^2 = \sum_{i:x_i \in X_j} \|x_i\|_2^2 \sum_{h:x_h \in X_j} \frac{1}{|X_j|}$$

Proof

first term $$\sum_{i:x_i \in X_j} \|x_i\|_2^2 = \sum_{i:x_i \in X_j} \|x_i\|_2^2 \sum_{h:x_h \in X_j} \frac{1}{|X_j|} = \frac{1}{|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \|x_i\|_2^2$$

Proof

first term

$$\sum_{i:x_i \in X_j} \|x_i\|_2^2 = \sum_{i:x_i \in X_j} \|x_i\|_2^2 \sum_{h:x_h \in X_j} \frac{1}{|X_j|} = \frac{1}{|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \|x_i\|_2^2$$

$$= \frac{1}{2|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} (\|x_i\|_2^2 + \|x_h\|_2^2)$$

Proof

first term

$$\sum_{i:x_i \in X_j} \|x_i\|_2^2 = \sum_{i:x_i \in X_j} \|x_i\|_2^2 \sum_{h:x_h \in X_j} \frac{1}{|X_j|} = \frac{1}{|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \|x_i\|_2^2$$

$$= \frac{1}{2|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \left(\|x_i\|_2^2 + \|x_h\|_2^2\right)$$

second term

$$- \sum_{i:x_i \in X_j} \frac{2}{|X_j|} \sum_{h:x_h \in X_j} \langle x_i, x_h \rangle$$

Proof

first term
$$\sum_{i:x_i \in X_j} \|x_i\|_2^2 = \sum_{i:x_i \in X_j} \|x_i\|_2^2 \sum_{h:x_h \in X_j} \frac{1}{|X_j|} = \frac{1}{|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \|x_i\|_2^2$$
$$= \frac{1}{2|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \left(\|x_i\|_2^2 + \|x_h\|_2^2\right)$$

second term
$$-\sum_{i:x_i \in X_j} \frac{2}{|X_j|} \sum_{h:x_h \in X_j} \langle x_i, x_h \rangle = -\frac{2}{|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \langle x_i, x_h \rangle$$

Proof

first term
$$\sum_{i:x_i\in X_j}\|x_i\|_2^2=\sum_{i:x_i\in X_j}\|x_i\|_2^2\sum_{h:x_h\in X_j}\frac{1}{|X_j|}=\frac{1}{|X_j|}\sum_{i:x_i\in X_j}\sum_{h:x_h\in X_j}\|x_i\|_2^2$$
$$=\frac{1}{2|X_j|}\sum_{i:x_i\in X_j}\sum_{h:x_h\in X_j}\left(\|x_i\|_2^2+\|x_h\|_2^2\right)$$

second term
$$-\sum_{i:x_i\in X_j}\frac{2}{|X_j|}\sum_{h:x_h\in X_j}\langle x_i,x_h\rangle=-\frac{2}{|X_j|}\sum_{i:x_i\in X_j}\sum_{h:x_h\in X_j}\langle x_i,x_h\rangle$$

third term $$\sum_{i:x_i\in X_j}\frac{1}{|X_j|^2}\left\|\sum_{h:x_h\in X_j}x_h\right\|_2^2$$

Proof

first term
$$\sum_{i:x_i\in X_j} \|x_i\|_2^2 = \sum_{i:x_i\in X_j} \|x_i\|_2^2 \sum_{h:x_h\in X_j} \frac{1}{|X_j|} = \frac{1}{|X_j|} \sum_{i:x_i\in X_j} \sum_{h:x_h\in X_j} \|x_i\|_2^2$$
$$= \frac{1}{2|X_j|} \sum_{i:x_i\in X_j} \sum_{h:x_h\in X_j} \left(\|x_i\|_2^2 + \|x_h\|_2^2\right)$$

second term
$$-\sum_{i:x_i\in X_j} \frac{2}{|X_j|} \sum_{h:x_h\in X_j} \langle x_i, x_h\rangle = -\frac{2}{|X_j|} \sum_{i:x_i\in X_j} \sum_{h:x_h\in X_j} \langle x_i, x_h\rangle$$

third term
$$\sum_{i:x_i\in X_j} \frac{1}{|X_j|^2} \left\| \sum_{h:x_h\in X_j} x_h \right\|_2^2 = \frac{1}{|X_j|} \left\| \sum_{h:x_h\in X_j} x_h \right\|_2^2$$

Proof

first term

$$\sum_{i:x_i\in X_j}\|x_i\|_2^2 = \sum_{i:x_i\in X_j}\|x_i\|_2^2 \sum_{h:x_h\in X_j}\frac{1}{|X_j|} = \frac{1}{|X_j|}\sum_{i:x_i\in X_j}\sum_{h:x_h\in X_j}\|x_i\|_2^2$$

$$= \frac{1}{2|X_j|}\sum_{i:x_i\in X_j}\sum_{h:x_h\in X_j}\left(\|x_i\|_2^2 + \|x_h\|_2^2\right)$$

second term

$$-\sum_{i:x_i\in X_j}\frac{2}{|X_j|}\sum_{h:x_h\in X_j}\langle x_i, x_h\rangle = -\frac{2}{|X_j|}\sum_{i:x_i\in X_j}\sum_{h:x_h\in X_j}\langle x_i, x_h\rangle$$

third term

$$\sum_{i:x_i\in X_j}\frac{1}{|X_j|^2}\left\|\sum_{h:x_h\in X_j}x_h\right\|_2^2 = \frac{1}{|X_j|}\left\|\sum_{h:x_h\in X_j}x_h\right\|_2^2 = \frac{1}{|X_j|}\sum_{i:x_i\in X_j}\sum_{h:x_h\in X_j}\langle x_i, x_h\rangle$$

Proof

$$
\sum_{j=1}^{k} \sum_{i:x_i \in X_j} \left\| x_i - \frac{1}{|X_j|} \sum_{h:x_h \in X_j} x_h \right\|_2^2
$$

$$
= \sum_{j=1}^{k} \sum_{i:x_i \in X_j} \left(\|x_i\|_2^2 - \frac{2}{|X_j|} \sum_{h:x_h \in X_j} \langle x_i, x_h \rangle + \frac{1}{|X_j|^2} \left\| \sum_{h:x_h \in X_j} x_h \right\|_2^2 \right)
$$

k

Proof

$$\sum_{j=1}^{k} \sum_{i:x_i \in X_j} \left\| x_i - \frac{1}{|X_j|} \sum_{h:x_h \in X_j} x_h \right\|_2^2$$

$$= \sum_{j=1}^{k} \sum_{i:x_i \in X_j} \left(\|x_i\|_2^2 - \frac{2}{|X_j|} \sum_{h:x_h \in X_j} \langle x_i, x_h \rangle + \frac{1}{|X_j|^2} \left\| \sum_{h:x_h \in X_j} x_h \right\|_2^2 \right)$$

$$= \frac{1}{2} \sum_{j=1}^{k} \frac{1}{|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \left(\|x_i\|_2^2 + \|x_h\|_2^2 - 2\langle x_i, x_h \rangle \right)$$

Proof

$$\sum_{j=1}^{k} \sum_{i:x_i \in X_j} \left\| x_i - \frac{1}{|X_j|} \sum_{h:x_h \in X_j} x_h \right\|_2^2$$

$$= \sum_{j=1}^{k} \sum_{i:x_i \in X_j} \left(\|x_i\|_2^2 - \frac{2}{|X_j|} \sum_{h:x_h \in X_j} \langle x_i, x_h \rangle + \frac{1}{|X_j|^2} \left\| \sum_{h:x_h \in X_j} x_h \right\|_2^2 \right)$$

$$= \frac{1}{2} \sum_{j=1}^{k} \frac{1}{|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} (\|x_i\|_2^2 + \|x_h\|_2^2 - 2\langle x_i, x_h \rangle)$$

$$= \frac{1}{2} \sum_{j=1}^{k} \frac{1}{|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \|x_i - x_h\|_2^2$$

Proof

$$\sum_{j=1}^{k} \sum_{i:x_i \in X_j} \left\| x_i - \frac{1}{|X_j|} \sum_{h:x_h \in X_j} x_h \right\|_2^2$$

$$= \sum_{j=1}^{k} \sum_{i:x_i \in X_j} \left(\|x_i\|_2^2 - \frac{2}{|X_j|} \sum_{h:x_h \in X_j} \langle x_i, x_h \rangle + \frac{1}{|X_j|^2} \left\| \sum_{h:x_h \in X_j} x_h \right\|_2^2 \right)$$

$$= \frac{1}{2} \sum_{j=1}^{k} \frac{1}{|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \left(\|x_i\|_2^2 + \|x_h\|_2^2 - 2\langle x_i, x_h \rangle \right)$$

$$= \frac{1}{2} \sum_{j=1}^{k} \frac{1}{|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \|x_i - x_h\|_2^2$$

QED

Lemma: Approximation is (almost) preserved

- If the solution computed by applying Lloyd's algorithm on the reduced instance is a $(1+\gamma)$-approximation, this is a $(1+\gamma)(1+4\varepsilon)$-approx. for the original instance

Lemma: Approximation is (almost) preserved

- Proof: Consider the clustering X that is returned by Lloyd's algorithm on the reduced instance, and let X^* be the optimal clustering for the initial instance

Lemma: Approximation is (almost) preserved

- Proof: Consider the clustering X that is returned by Lloyd's algorithm on the reduced instance, and let X^* be the optimal clustering for the initial instance
- Let C and R be the cost of a clustering for the initial and reduced points, respectively, i.e.,

$$C(X) = \frac{1}{2}\sum_{j=1}^{k} \frac{1}{|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \|x_i - x_h\|_2^2$$

$$R(X) = \frac{1}{2}\sum_{j=1}^{k} \frac{1}{|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \|Lx_i - Lx_h\|_2^2$$

Lemma: Approximation is (almost) preserved

- Proof: Consider the clustering X that is returned by Lloyd's algorithm on the reduced instance, and let X^* be the optimal clustering for the initial instance
- Let C and R be the cost of a clustering for the initial and reduced points, respectively, i.e.,

$$C(X) = \frac{1}{2}\sum_{j=1}^{k} \frac{1}{|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \boxed{\|x_i - x_h\|_2^2}$$

$$R(X) = \frac{1}{2}\sum_{j=1}^{k} \frac{1}{|X_j|} \sum_{i:x_i \in X_j} \sum_{h:x_h \in X_j} \boxed{\|Lx_i - Lx_h\|_2^2}$$

$1 \pm \varepsilon$ of each other

Lemma: Approximation is (almost) preserved

- By JL lemma, we have

$$(1-\varepsilon)C(X) \leq R(X) \leq (1+\varepsilon)C(X)$$

and

$$(1-\varepsilon)C(X^*) \leq R(X^*) \leq (1+\varepsilon)C(X^*)$$

Lemma: Approximation is (almost) preserved

- By JL lemma, we have
$$(1-\varepsilon)C(X) \le R(X) \ \le (1+\varepsilon)C(X)$$
and
$$(1-\varepsilon)C(X^*) \le R(X^*) \ \le (1+\varepsilon)C(X^*)$$
- Since X is a $(1+\gamma)$-approximation of the optimal clustering in the reduced instance, it is also a $(1+\gamma)$-approximation of clustering X^* in the reduced instance, i.e., $R(X) \ \le (1+\gamma)R(X^*)$

Lemma: Approximation is (almost) preserved

- By JL lemma, we have
$$\boxed{(1-\varepsilon)C(X) \le R(X)} \le (1+\varepsilon)C(X)$$
  and
$$(1-\varepsilon)C(X^*) \le \boxed{R(X^*) \le (1+\varepsilon)C(X^*)}$$
- Since X is a $(1+\gamma)$-approximation of the optimal clustering in the reduced instance, it is also a $(1+\gamma)$-approximation of clustering X^* in the reduced instance, i.e., $\boxed{R(X) \le (1+\gamma)R(X^*)}$

Lemma: Approximation is (almost) preserved

- Putting everything together, we have

$C(X)$

Lemma: Approximation is (almost) preserved

- Putting everything together, we have

$$C(X) \leq \frac{1}{1-\varepsilon} R(X)$$

JL lemma

Lemma: Approximation is (almost) preserved

- Putting everything together, we have

$$C(X) \le \frac{1}{1-\varepsilon} R(X) \le \frac{1+\gamma}{1-\varepsilon} R(X^*)$$

JL lemma

approximation

Lemma: Approximation is (almost) preserved

- Putting everything together, we have

$$C(X) \le \frac{1}{1-\varepsilon} R(X) \le \frac{1+\gamma}{1-\varepsilon} R(X^*) \le \frac{(1+\gamma)(1+\varepsilon)}{1-\varepsilon} C(X^*)$$

JL lemma

approximation

JL lemma

Lemma: Approximation is (almost) preserved

- Putting everything together, we have

$$C(X) \le \frac{1}{1-\varepsilon} R(X) \le \frac{1+\gamma}{1-\varepsilon} R(X^*) \le \frac{(1+\gamma)(1+\varepsilon)}{1-\varepsilon} C(X^*) \le (1+\gamma)(1+4\varepsilon) C(X^*)$$

JL lemma

approximation

JL lemma

$\frac{1+\varepsilon}{1-\varepsilon} \le 1 + 4\varepsilon$ since $\varepsilon \le 1/2$

Lemma: Approximation is (almost) preserved

- Putting everything together, we have

$$C(X) \leq \frac{1}{1-\varepsilon} R(X) \leq \frac{1+\gamma}{1-\varepsilon} R(X^*) \leq \frac{(1+\gamma)(1+\varepsilon)}{1-\varepsilon} C(X^*) \leq (1+\gamma)(1+4\varepsilon) C(X^*)$$

- Hence, clustering X is a $(1+\gamma)(1+4\varepsilon)$-approximation for the original instance

QED

Last slide

- Johnson-Lindenstrauss transform
- Proof of the JL lemma
- Application to k-means clustering