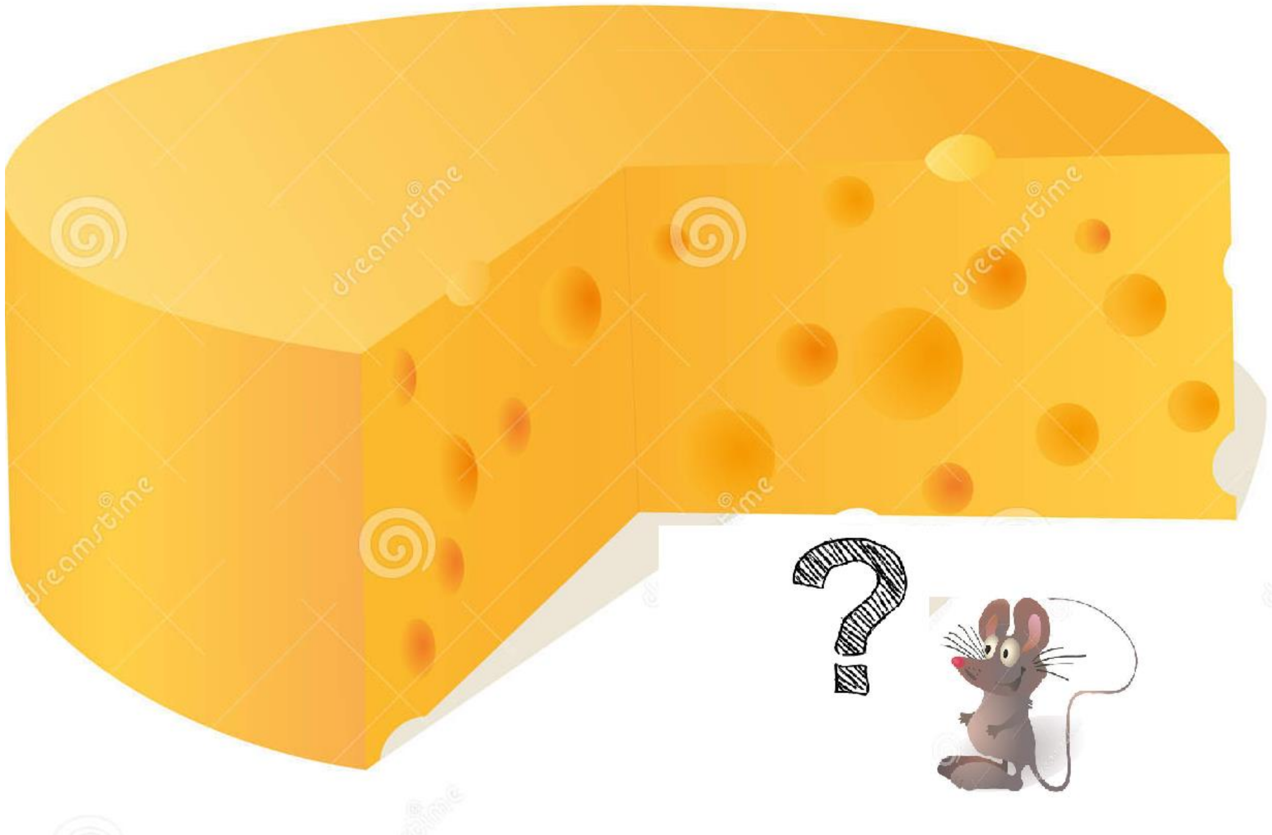
Big Data Class



LECTURER: DAN FELDMAN

TEACHING ASSISTANTS:

IBRAHIM JUBRAN

ALAA MAALOUF

אוניברסיטת חיפה
University of Haifa
جامعة حيفا

Department of Computer Science, University of Haifa.

k-Lines problem

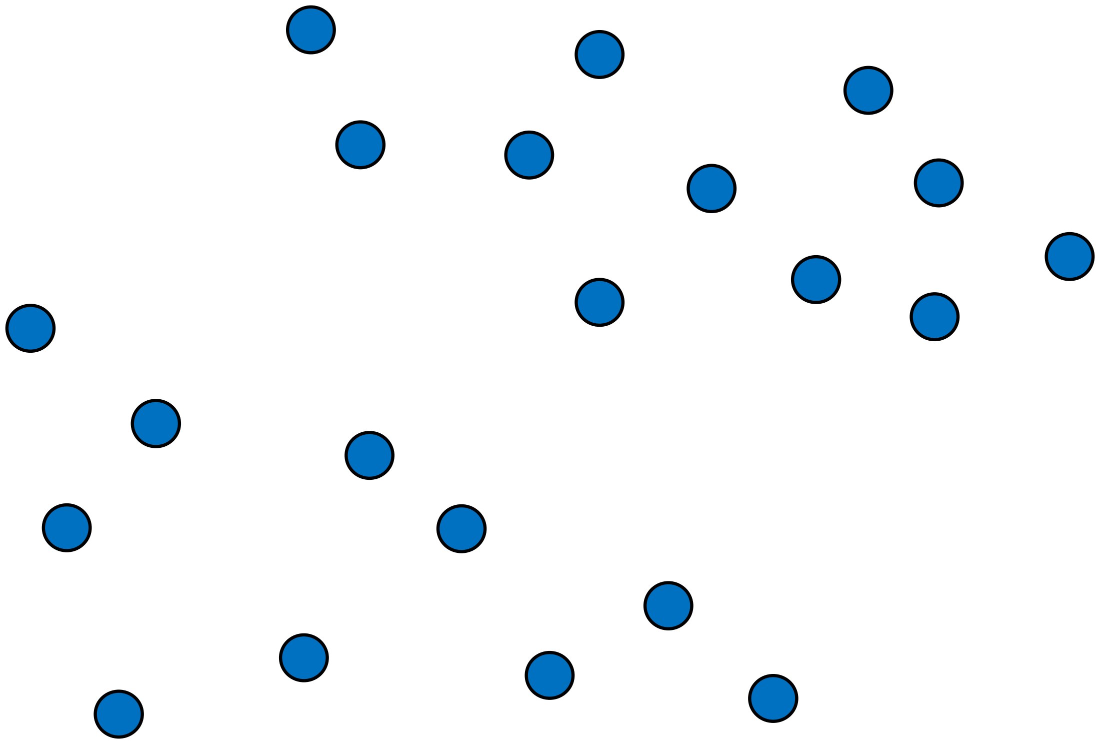
- Input: $P \subseteq R^d$
- Query space: $Q = \{\{\ell_1, \ldots, \ell_k\} \mid \ell_i \text{ is a line in } R^d\}\}$
- Cost function: $\forall L \in Q$:

$$dist(p, L) = \min_{\ell \in L} dist(p, \ell) = \min_{\ell \in L} \min_{x \in \ell} \|p - x\|_2$$

- $\text{OPT} = \min_{L \in Q} dist(P, L)$

4-approximation for k-Lines problem

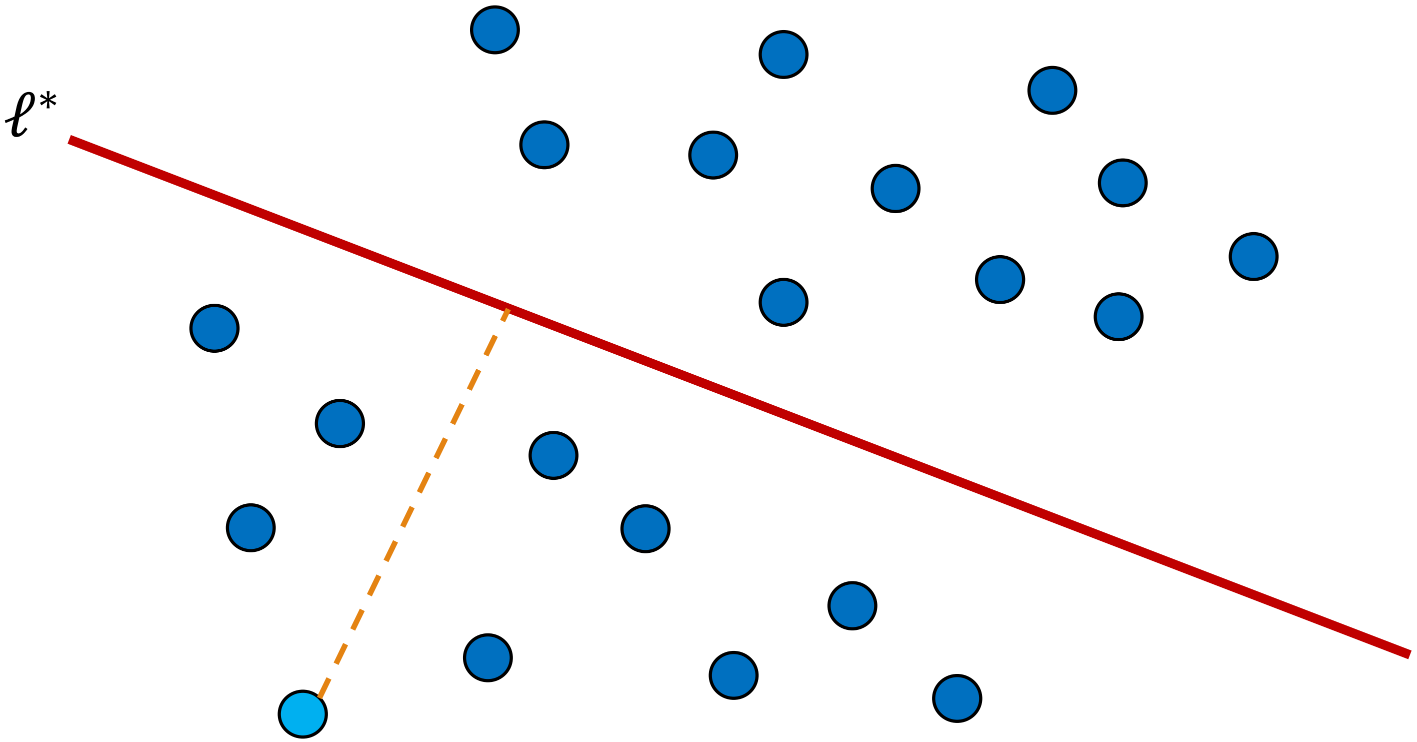
(k=1 , d=2)



4-approximation for k-Lines problem

(k=1 , d=2)

ℓ^* is the line that minimizes $\max_{p \in P} dist(p, \ell)$



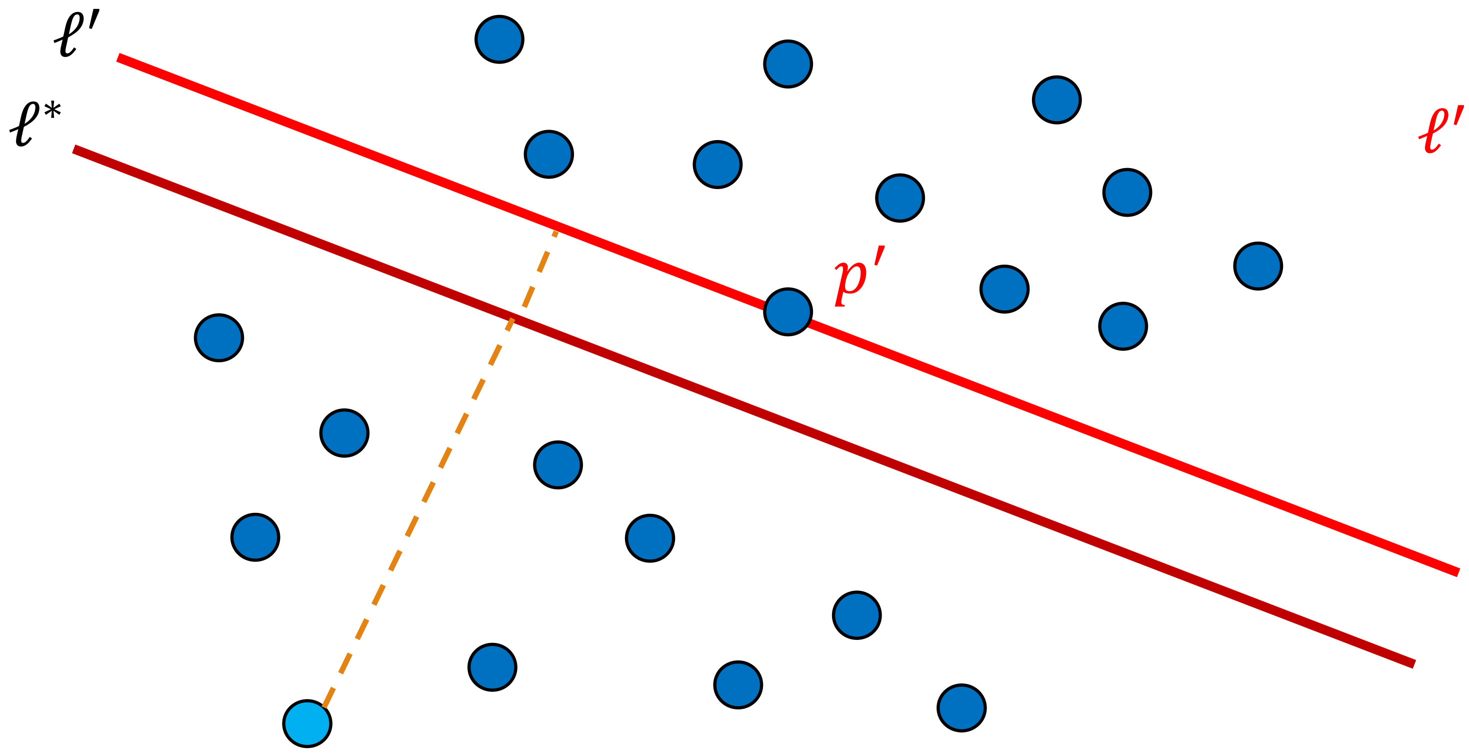
$$p^* = arg \max_{p \in P} dist(p, \ell^*)$$

4-approximation for k-Lines problem

(k=1 , d=2)

ℓ^* is the line that minimizes $\max_{p \in P} dist(p, \ell)$

ℓ' is the translation of ℓ^* to $\ell^{*\prime}s$ closest point p'



$p^* = arg \max_{p \in P} dist(p, \ell^*)$

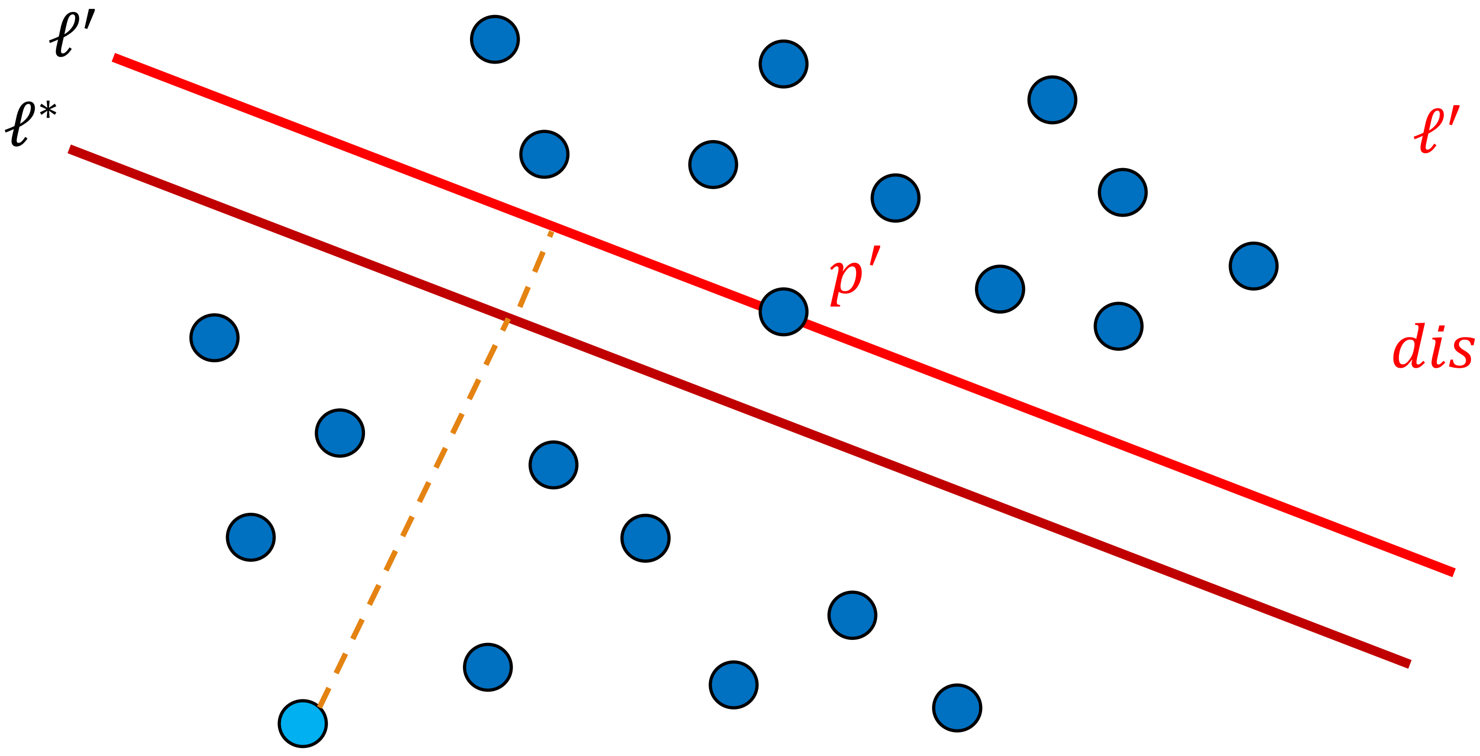
4-approximation for k-Lines problem

(k=1 , d=2)

ℓ^* is the line that minimizes $\max_{p\in P} dist(p,\ell)$

ℓ' is the translation of ℓ^* to $\ell^{*\prime}s$ closest point p'

$dist(p,\ell') \leq 2 \cdot dist(p,\ell^*)$

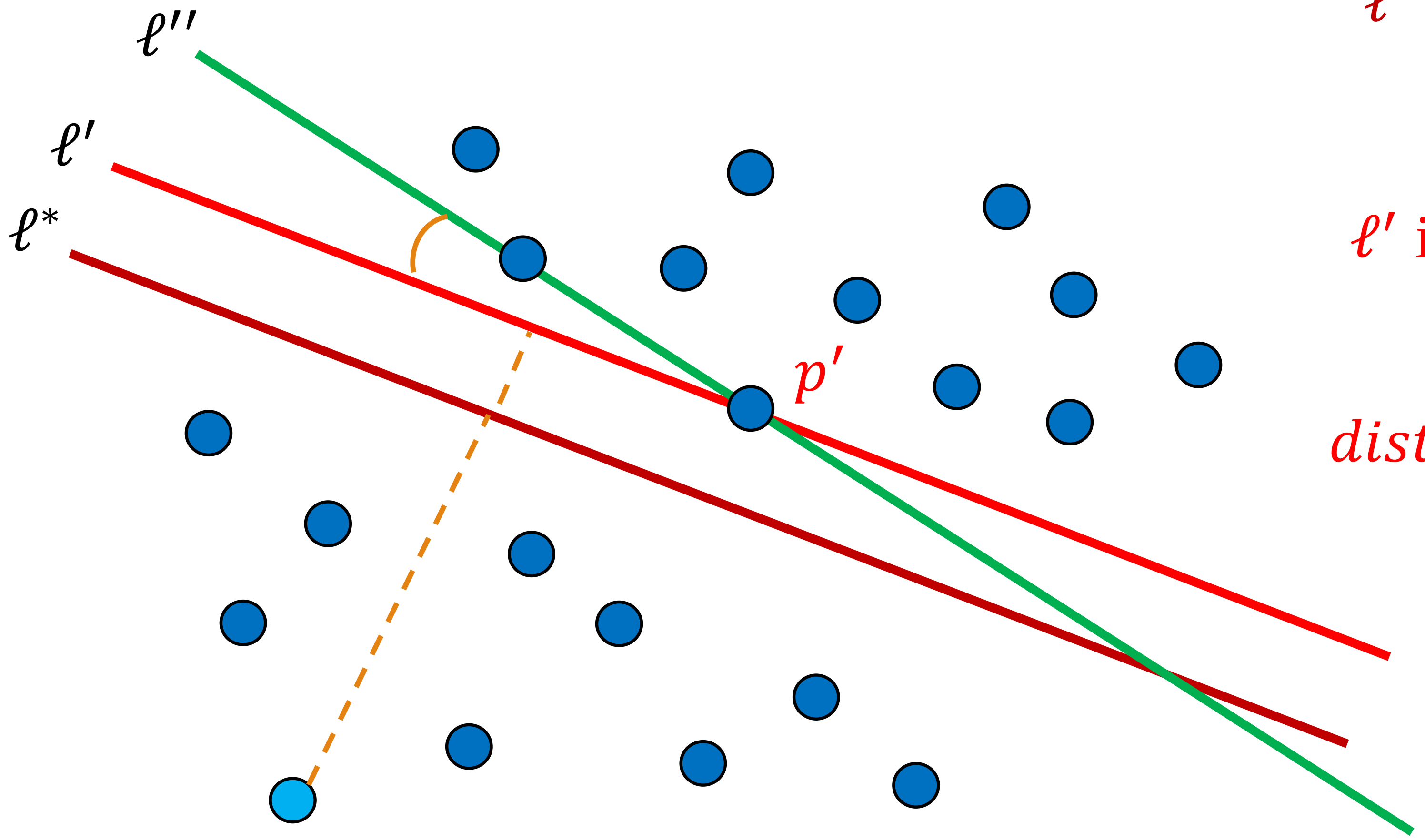


$p^* = arg \max_{p\in P} dist(p,\ell^*)$

4-approximation for k-Lines problem

(k=1 , d=2)

ℓ^* is the line that minimizes $\max_{p \in P} dist(p, \ell)$



$p^* = arg \max_{p \in P} dist(p, \ell^*)$

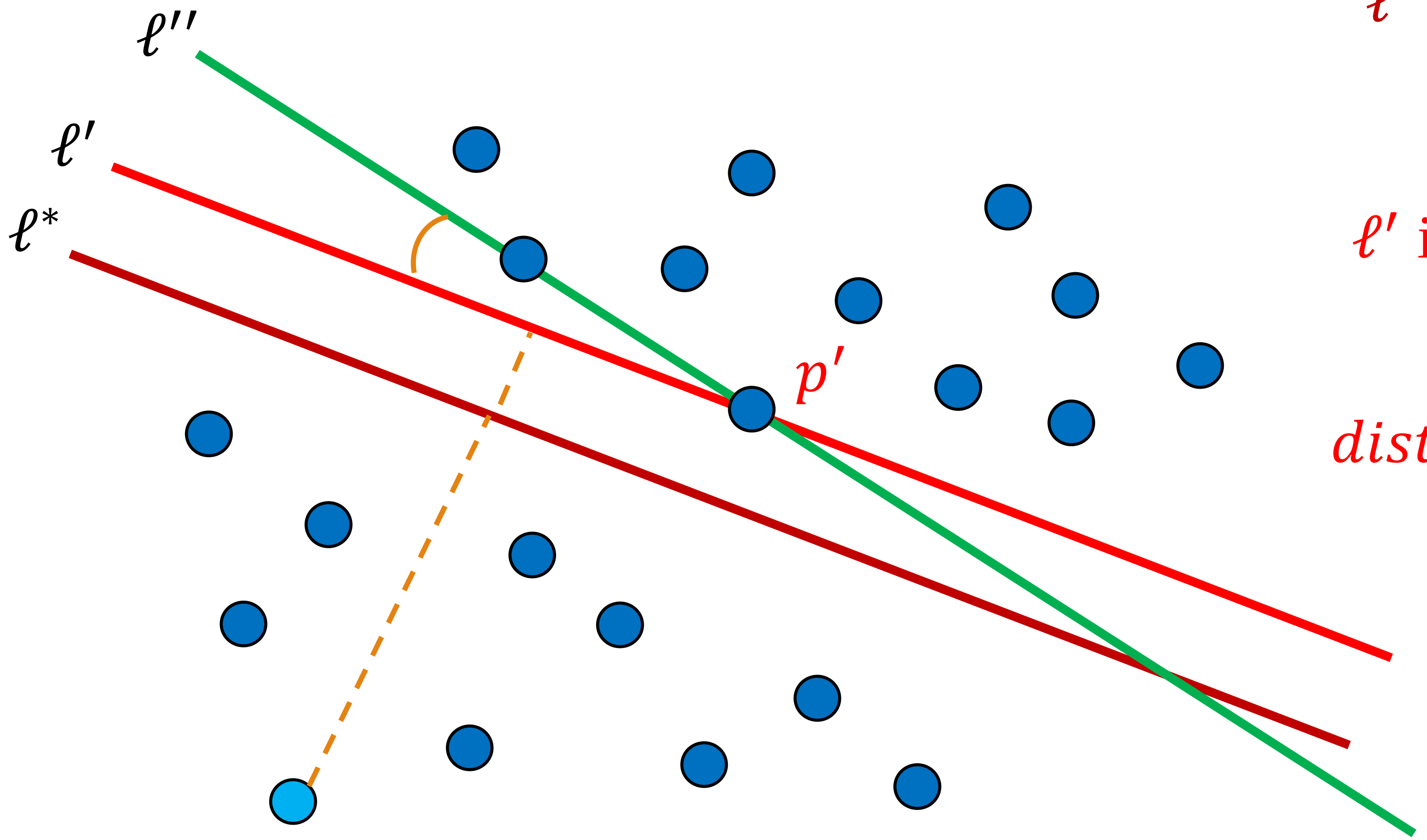
ℓ' is the translation of ℓ^* to $\ell^{*\prime}s$ closest point p'

$dist(p, \ell') \leq 2 \cdot dist(p, \ell^*)$

ℓ'' is the rotation of ℓ' around p' to $\ell''s$ closest point

4-approximation for k-Lines problem

(k=1 , d=2)



$p^* = arg \max_{p \in P} dist(p, \ell^*)$

ℓ^* is the line that minimizes $\max_{p \in P} dist(p, \ell)$

ℓ' is the translation of ℓ^* to $\ell^{*\prime}s$ closest point p'

$dist(p, \ell') \leq 2 \cdot dist(p, \ell^*)$

ℓ'' is the rotation of ℓ' around p' to $\ell''s$ closest point

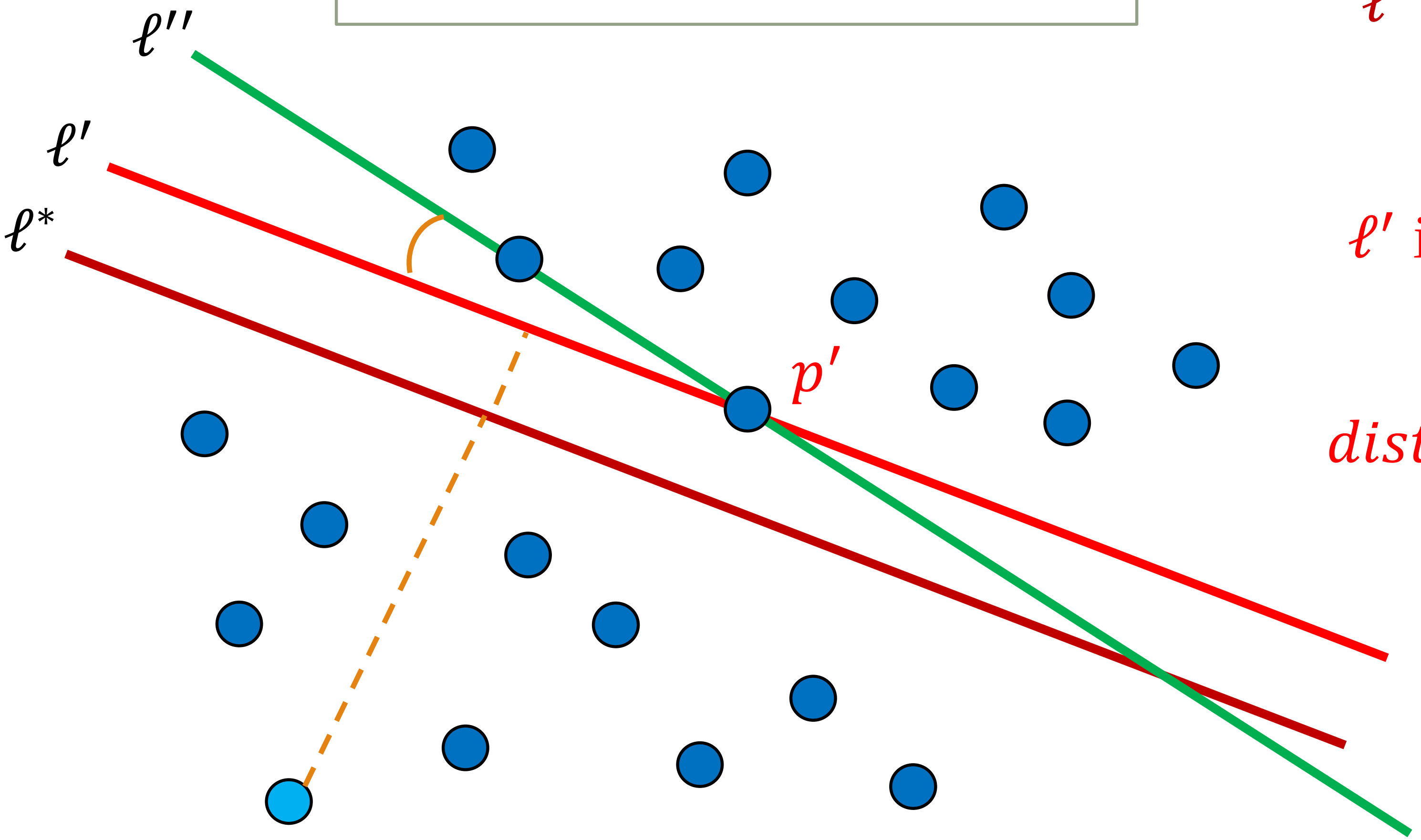
$dist(p, \ell'') \leq 2 \cdot dist(p, \ell')$

4-approximation for k-Lines problem

$$\boldsymbol{dist(p, \ell'') \le 4 \cdot dist(p, \ell^*)}$$

(k=1 , d=2)

ℓ^* is the line that minimizes $\max_{p \in P} dist(p, \ell)$



ℓ' is the translation of ℓ^* to $\ell^{*\prime}s$ closest point p'

$$dist(p, \ell') \le 2 \cdot dist(p, \ell^*)$$

ℓ'' is the rotation of ℓ' around p' to $\ell'^{\prime}s$ closest point

$$dist(p, \ell'') \le 2 \cdot dist(p, \ell')$$

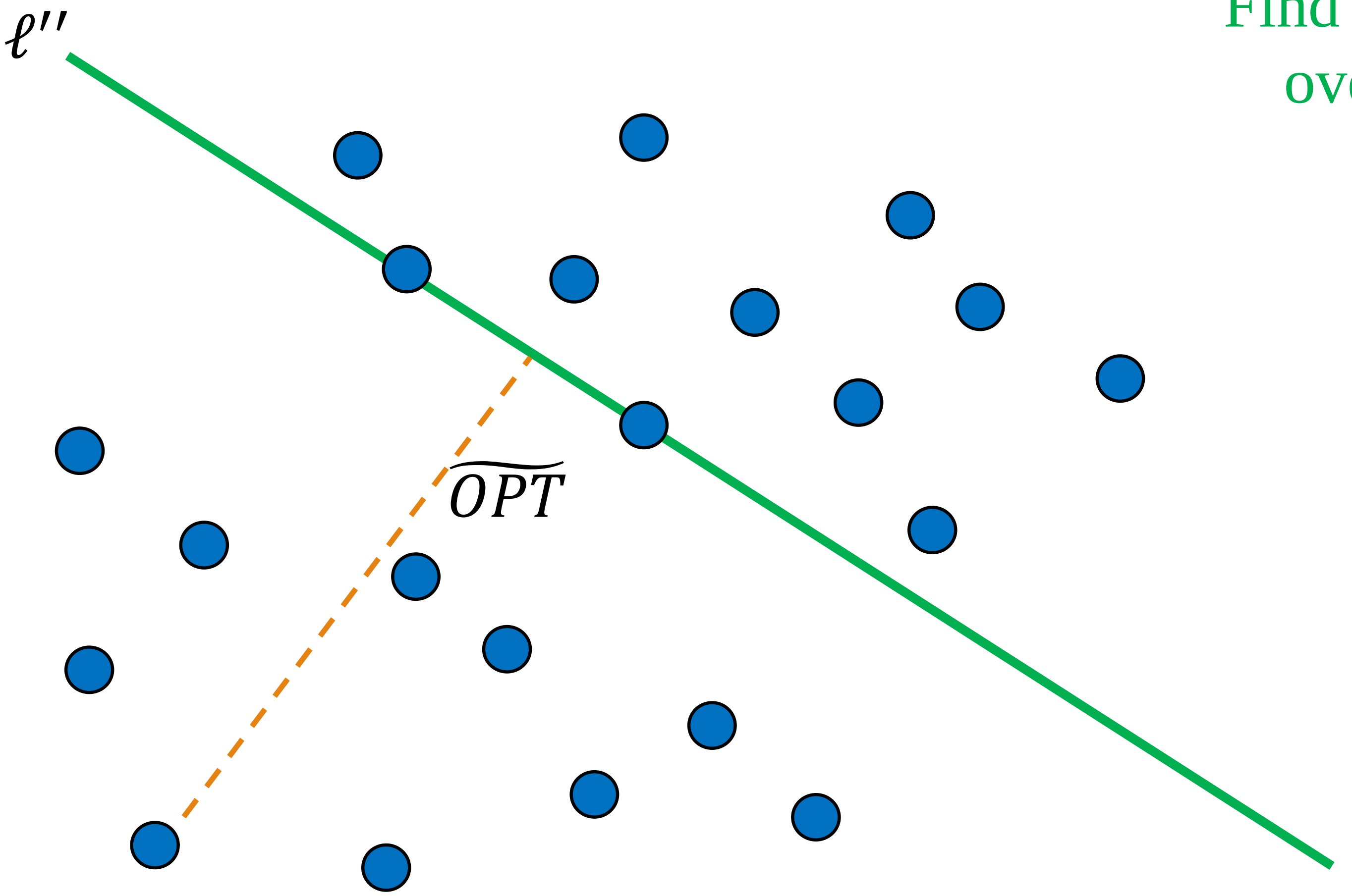
$$p^* = arg \max_{p \in P} dist(p, \ell^*)$$

4-approximation for k-Lines problem

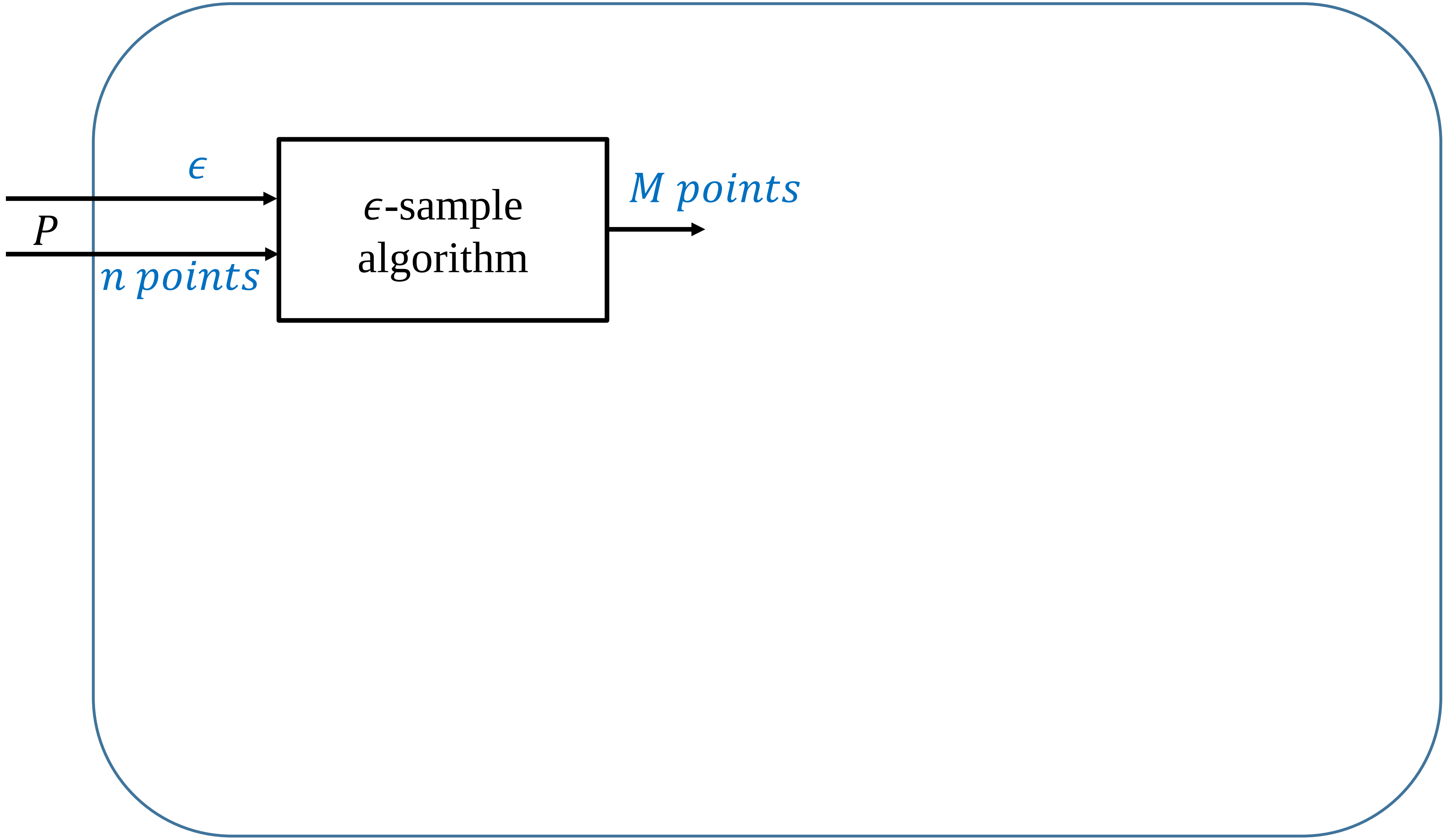
(k=1 , d=2)

Find ℓ'' by exhaustive search over every pair of points.

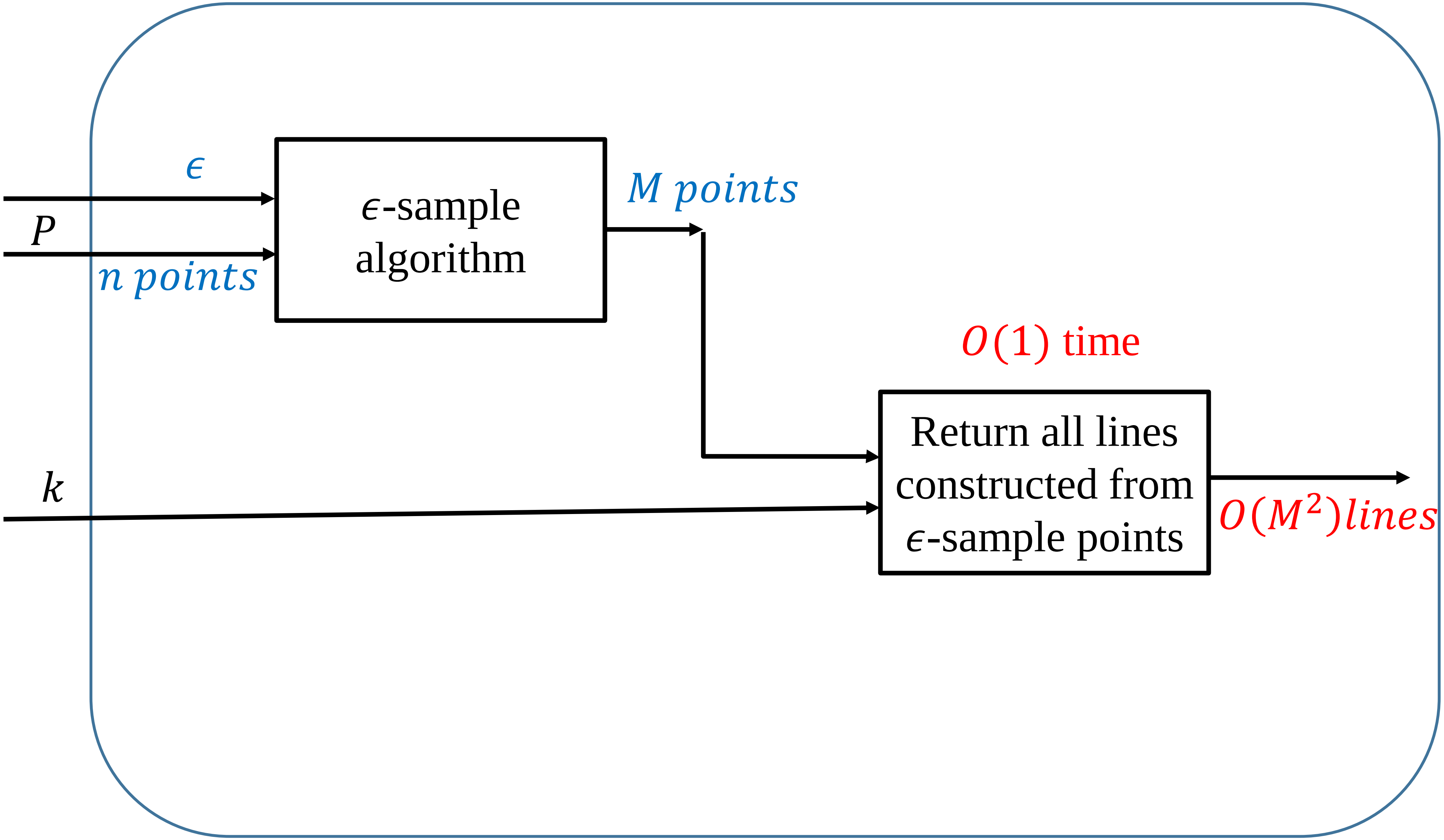
$O(n^2)$



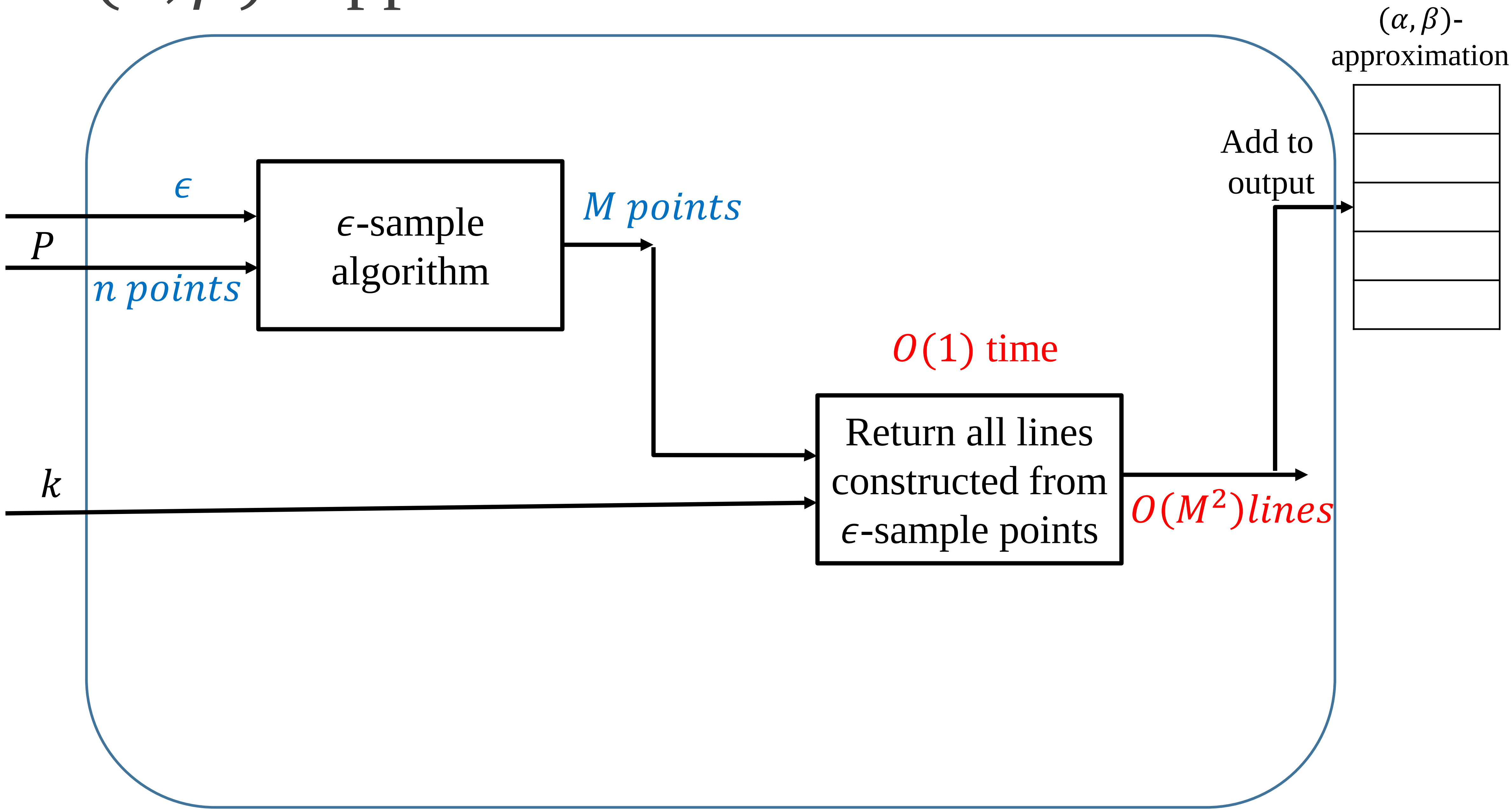
(α, β)-Approximation for k-Lines



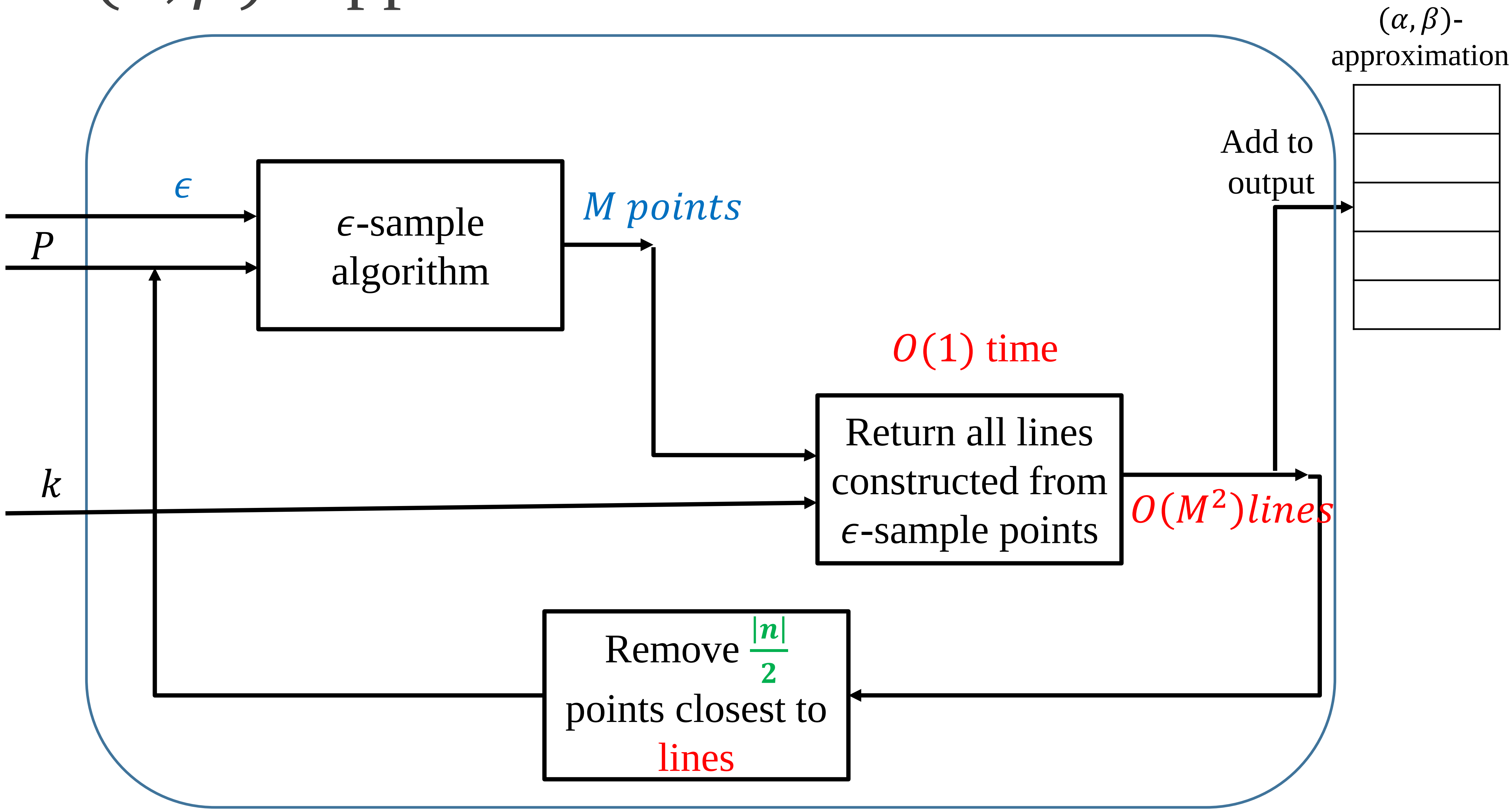
(α, β)-Approximation for k-Lines



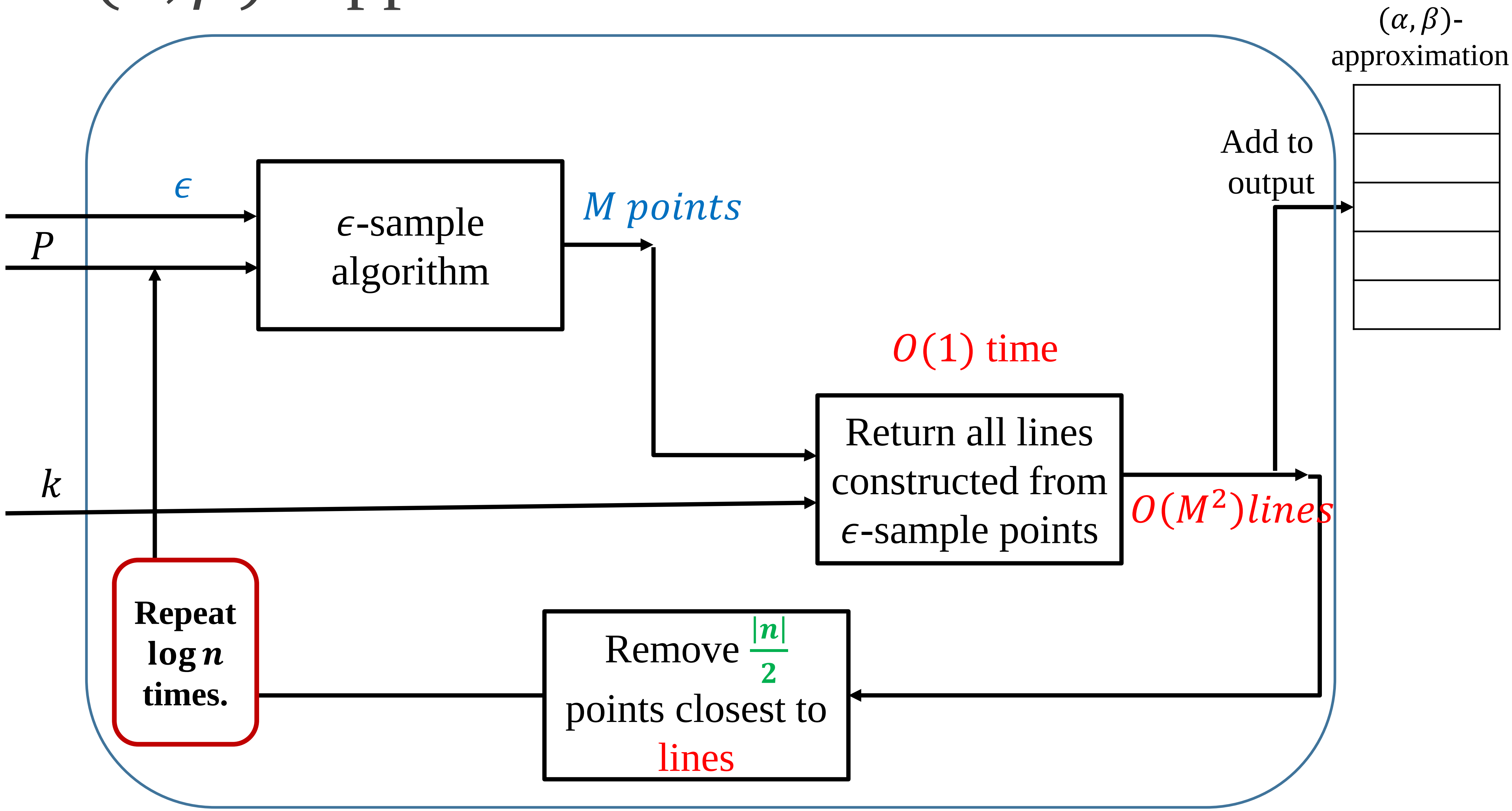
(α, β)-Approximation for k-Lines



(α, β)-Approximation for k-Lines



(α, β)-Approximation for k-Lines



(α, β)-Approximation for k-Lines

Analysis:

- $M =$ number of points returned by the ϵ-sample algorithm
- $\beta = O(M^2 \log n)$.
- $\alpha = 4$ since the ϵ-sample points is an 4-approximation.

Coreset for k-lines mean

- Input: $P \subseteq R^d$
- Query space: $Q = \{\{\ell_1, \dots, \ell_k\} \mid \ell_i \text{ is a line in } R^d\}$
- Cost function: $\forall L \in Q: dist(p, L) = \min_{\ell \in L} \min_{x \in \ell} \|p - x\|_2, \ f(p, L) = dist(p, L)^2$
- Output: $C \subseteq P \ s.t. \ \forall L \in Q:$

$$\left| \sum_{p \in P} f(p, L) - \sum_{c \in C} f(c, L) \right| \le \epsilon \cdot \sum_{p \in P} f(p, L)$$

Coreset for k-lines mean

- Input: $P \subseteq R^d$
- Query space: $Q = \{\{\ell_1, \ldots, \ell_k\} \mid \ell_i \textit{ is a line in } R^d\}$
- Cost function: $\forall L \in Q: dist(p, L) = \min_{\ell \in L} \min_{x \in \ell} \|p - x\|_2,\ f(p, L) = dist(p, L)^2$
- Output: $C \subseteq P\ s.t.\ \forall L \in Q:$

$$\left| \sum_{p \in P} f(p, L) - \sum_{c \in C} f(c, L) \right| \leq \epsilon \cdot \sum_{p \in P} f(p, L)$$

→ Need to compute sensitivity $s(p)$ for the problem above.

Coreset for k-lines mean

- Output: $\boldsymbol{C} \subseteq P\ s.t.\ \forall L \in Q$:

$$\left|\sum_{p \in P} f(p, L) - \sum_{c \in C} f(c, L)\right| \leq \epsilon \cdot \sum_{p \in P} f(p, L)$$

→ Need to compute sensitivity $s(p)$ for the problem above.

By the sensitivity Lemma:

$$\sum_{p \in P} s(p) \leq \rho\alpha + \rho^2(1+\alpha) \sum_{p' \in P'} \max_{L \in Q} \frac{f(p', L)}{f(P', L)}$$

Coreset for k-lines mean

- Output: $\boldsymbol{C} \subseteq P \; s.t. \; \forall L \in Q$:

$$\left| \sum_{p \in P} dist^2(p, L) - \sum_{c \in C} dist^2(c, L) \right| \leq \epsilon \cdot \sum_{p \in P} dist^2 (p, L)$$

→ Need to compute sensitivity $s(p)$ for the problem above.

The sensitivity of the desired problem

By the sensitivity Lemma:

$$\sum_{p \in P} s(p) \leq \rho\alpha + \rho^2(1+\alpha) \sum_{p' \in P'} \max_{L \in Q} \frac{f(p', L)}{f(P', L)}$$

Sensitivity of the projected points

Projection of P onto the Bicreteria

Coreset for k-lines mean

By the sensitivity Lemma:

$$\sum_{p \in P} s(p) \leq \rho\alpha + \rho^2(1+\alpha) \sum_{p' \in P'} \max_{L \in Q} \frac{f(p', L)}{f(P', L)}$$

✓ → Compute an (α, β)-approximation B for the k-lines mean problem as previously described.

Coreset for k-lines mean

By the sensitivity Lemma:

$$\sum_{p \in P} s(p) \leq \rho\alpha + \rho^2(1+\alpha) \sum_{p' \in P'} \max_{L \in Q} \frac{f(p', L)}{f(P', L)}$$

✓ → Compute an (α, β)-approximation B for the k-lines mean problem as previously described.

✓ → Compute P' = projection of P onto B.

Coreset for k-lines mean

By the sensitivity Lemma:

$$\sum_{p \in P} s(p) \leq \rho\alpha + \rho^2(1+\alpha) \sum_{p' \in P'} \max_{L \in Q} \frac{f(p', L)}{f(P', L)}$$

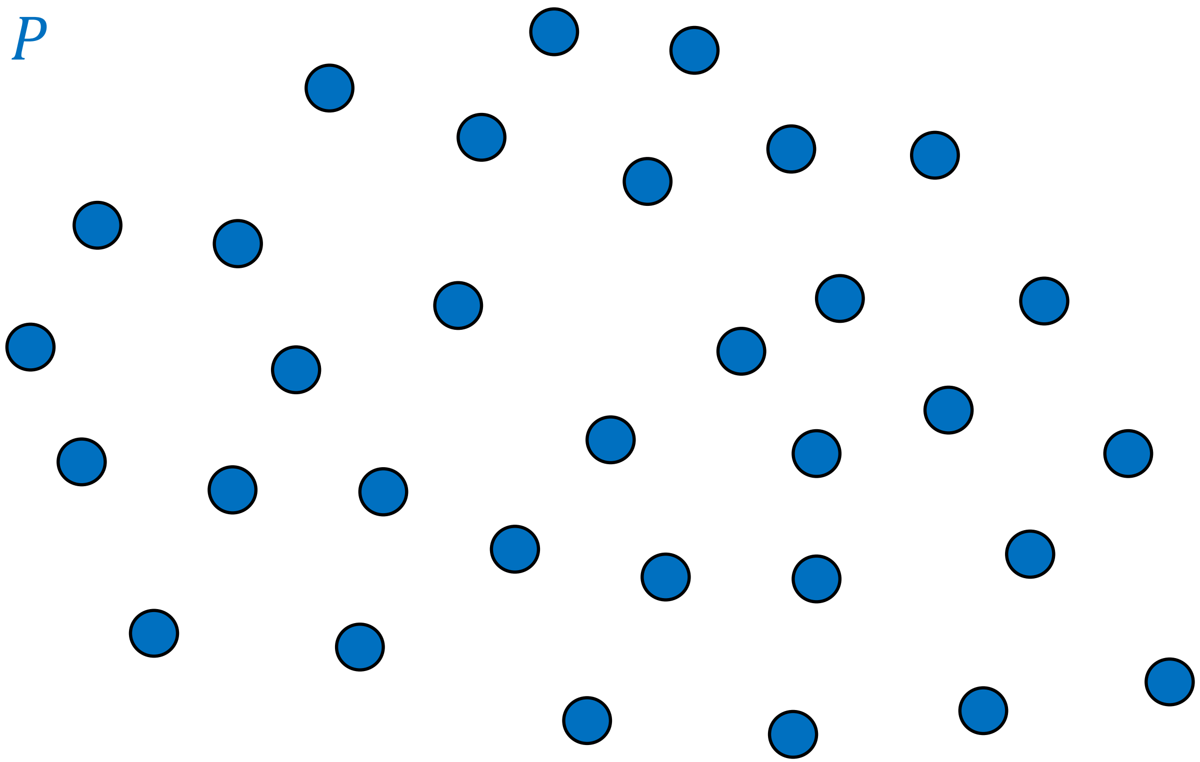
✓ → Compute an (α, β)-approximation B for the k-lines mean problem as previously described.

✓ → Compute P' = projection of P onto B.

→ Need to compute sensitivity $s(p')$ for the projected points.

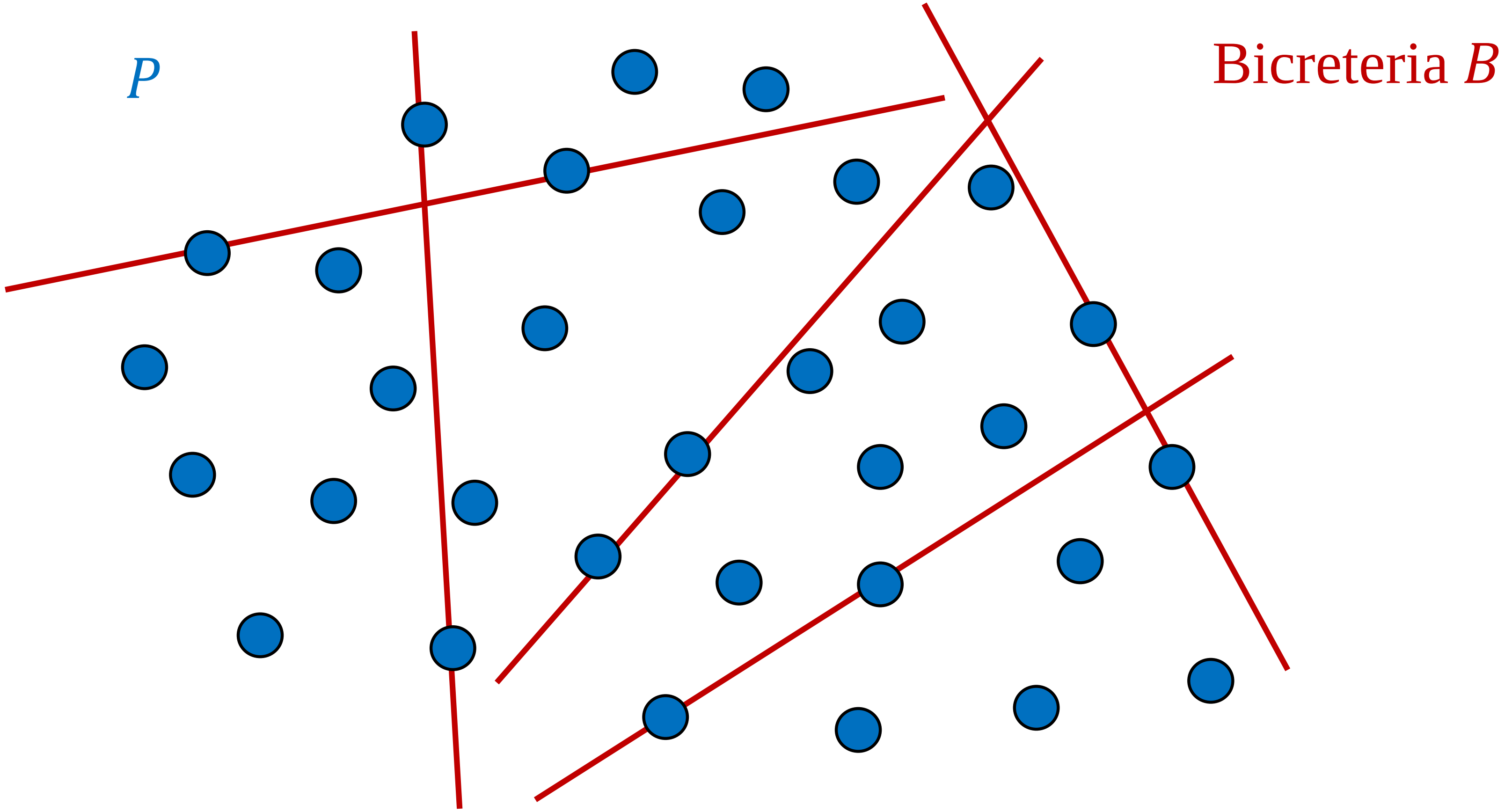
Coreset for k-lines mean

→ Need to compute sensitivity $s(p')$ for the projected points.



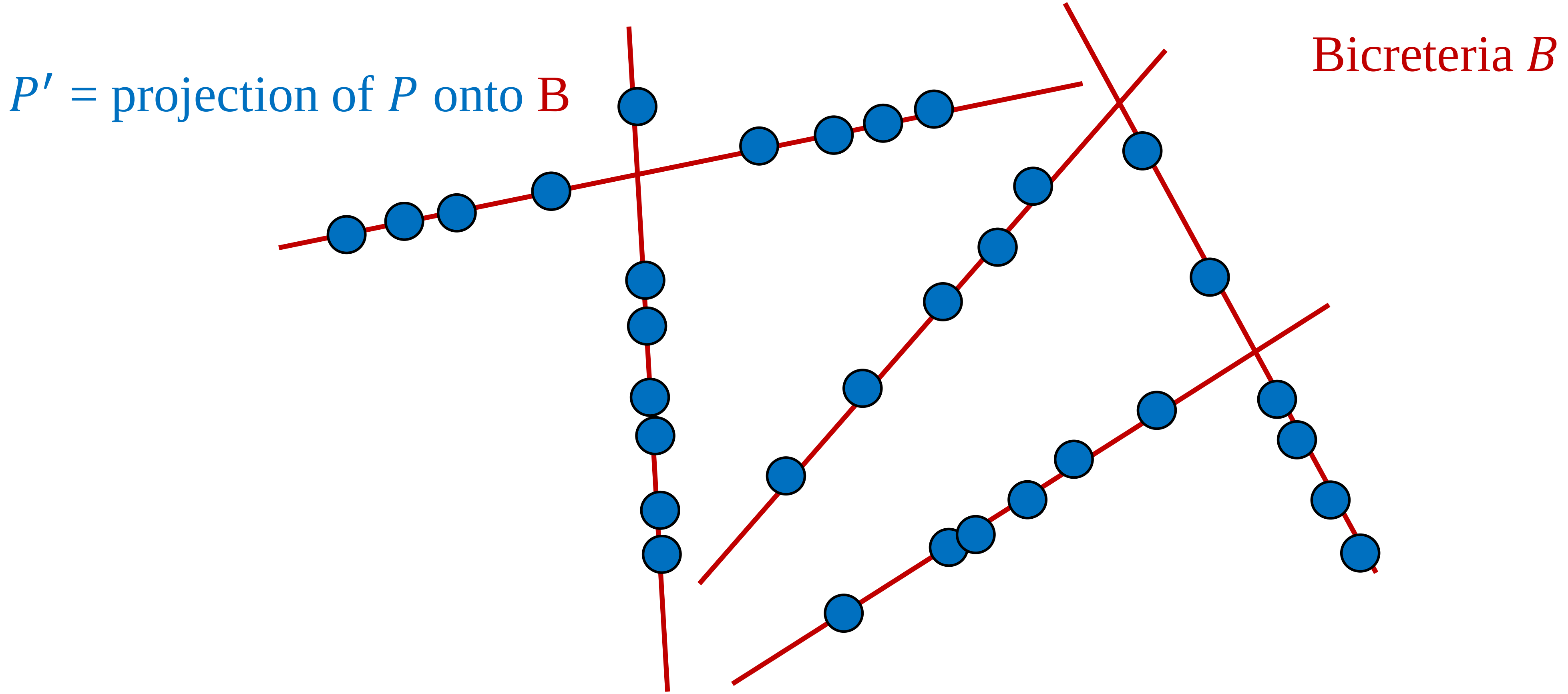
Coreset for k-lines mean

→ Need to compute sensitivity $s(p')$ for the projected points.



Coreset for k-lines mean

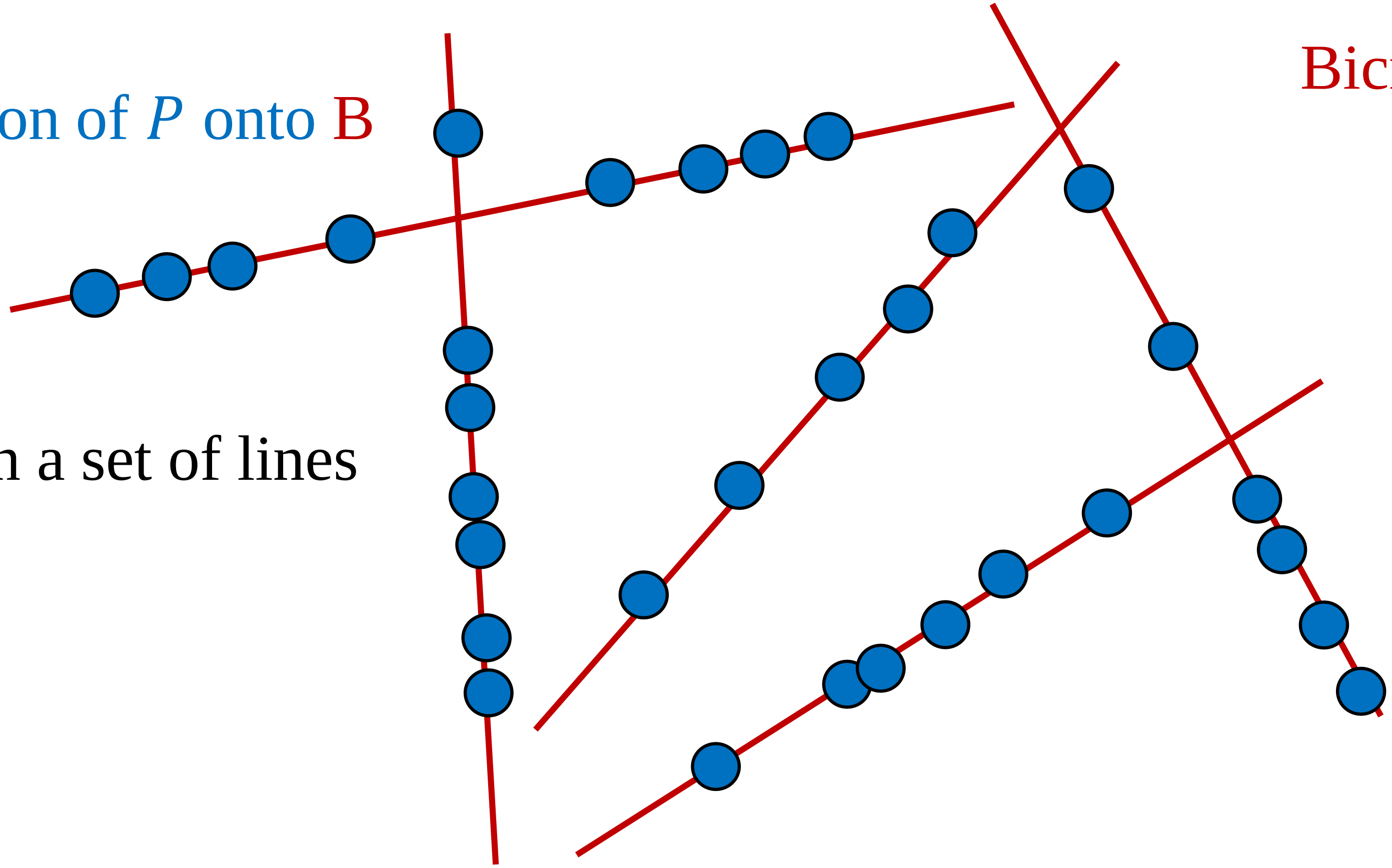
→ Need to compute sensitivity $s(p')$ for the projected points.



Coreset for k-lines mean

→ Need to compute sensitivity $s(p')$ for the projected points.

P' = projection of P onto B

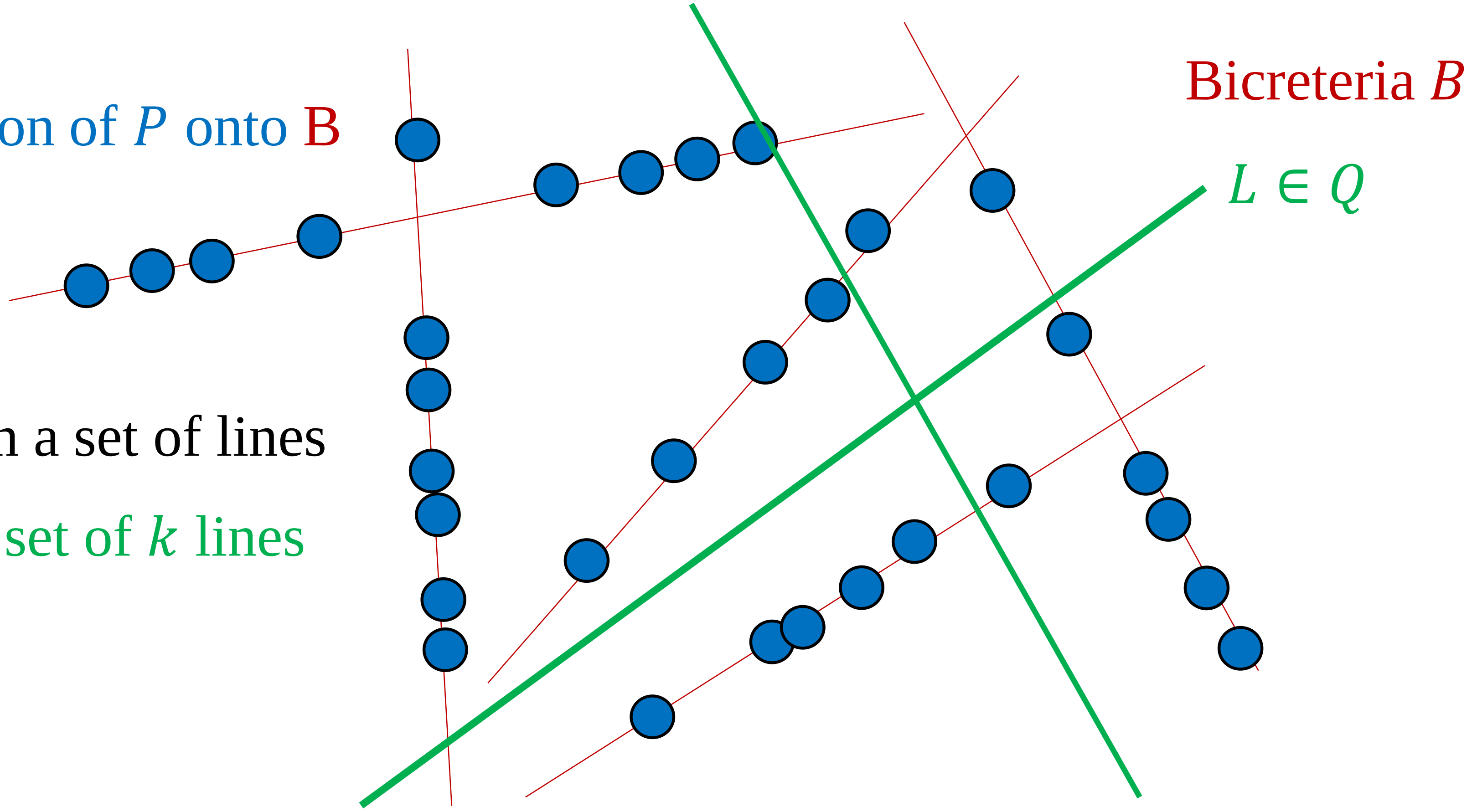


→ Now the points are on a set of lines

Coreset for k-lines mean

→ Need to compute sensitivity $s(p')$ for the projected points.

P' = projection of P onto B

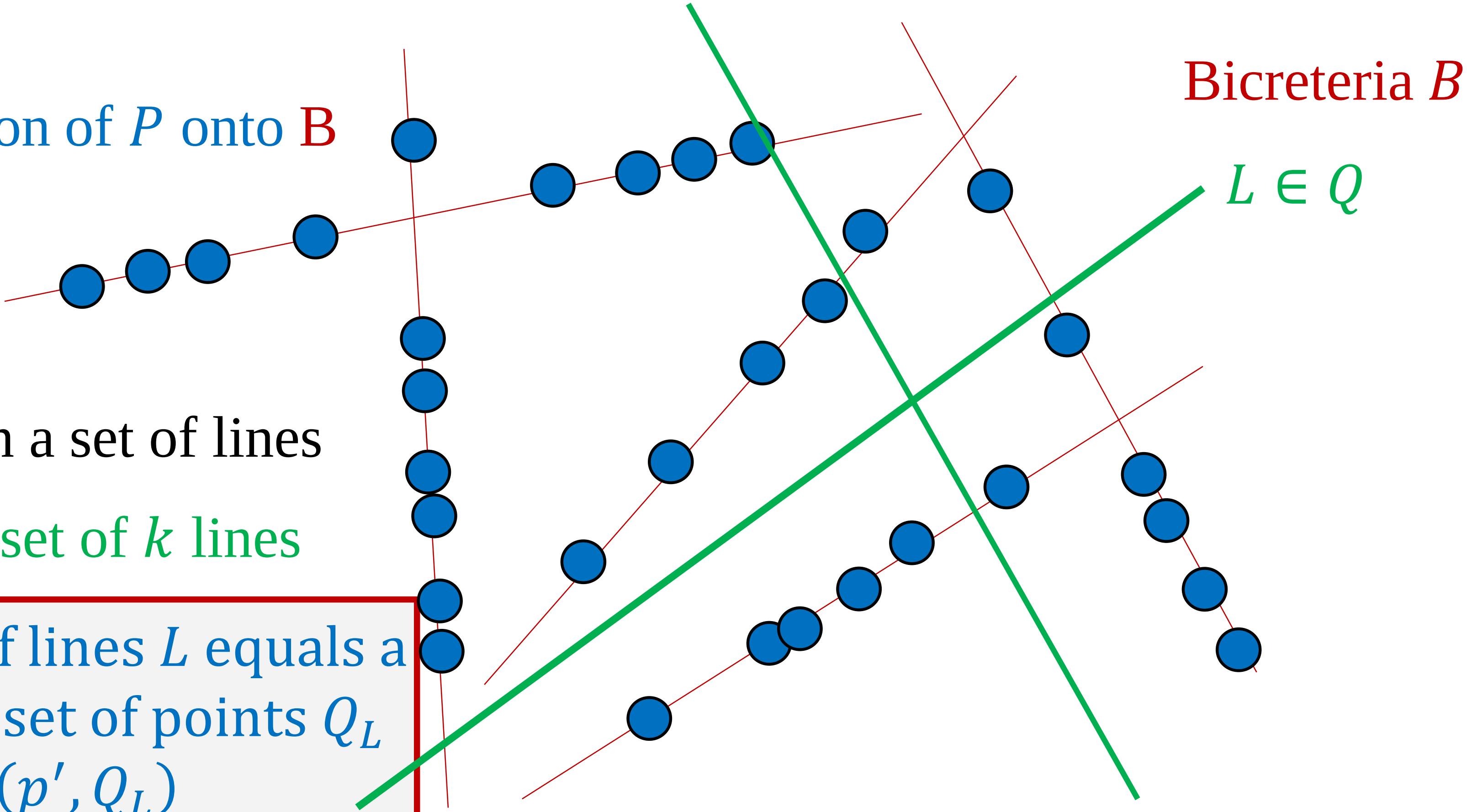


→ Now the points are on a set of lines

→ The query $L \in Q$ is a set of k lines

Coreset for k-lines mean

→ Need to compute sensitivity $s(p')$ for the projected points.



→ Now the points are on a set of lines

→ The query $L \in Q$ is a set of k lines

→ Distance to the set of lines L equals a distance to a weighted set of points Q_L

$$f(p', L) = f_{\omega}(p', Q_L)$$

Coreset for k-lines mean

$$\rightarrow f(p', L) = f_\omega(p', Q_L) = \min_{(q,\omega)\in Q_L} \omega \cdot \|p - q\|_2^2$$

$$\rightarrow s(p') = \max_{L\in Q} \frac{f(p', L)}{f(P', L)} = \max_{Q_L \in R^d} \frac{f_\omega(p', Q_L)}{f_\omega(P', Q_L)}$$

→ Need to compute sensitivity for the **weighted k-means problem**

Weights are
unknown beforehand
(part of the query)

Sensitivity for Weighted k-means

- Input: $P \subseteq R^d$
- Query space: $Q = \{\{(q_1, \omega_1), \dots, (q_k, \omega_k)\} \mid q_i \in R^d, \omega_i \in [0, \infty)\}$
- Cost function: $\forall C \in Q$:

$$f_\omega(p, C) = \min_{(c,\omega) \in C} \omega \cdot f(p, c) = \min_{(c,\omega) \in C} \omega \cdot dist^2(p, c)$$

Sensitivity for Weighted k-means

- Input: $P \subseteq R^d$
- Query space: $Q = \{\{(q_1, \omega_1), \ldots, (q_k, \omega_k)\} \mid q_i \in R^d, \omega_i \in [0, \infty)\}$
- Cost function: $\forall C \in Q$:

$$f_\omega(p, C) = \min_{(c,\omega)\in C} \omega \cdot f(p, c) = \min_{(c,\omega)\in C} \omega \cdot dist^2(p, c)$$

r (Lipschitz)

- The function f satisfies the following two conditions for every $p, q, c \in R^d$:

1) For $\phi = (4r)^r$: $f(p, q) - f(q, c) \leq \phi f(p, q) + \frac{f(p,c)}{4}$.
2) For $\rho = \max\{2^{r-1}, 1\}$: $f(p, q) \leq \rho\big(f(p, c) + f(c, q)\big)$.

Sensitivity for Weighted k-means

- Consider the following algorithm:

Robust-Median($\boldsymbol{P}, \boldsymbol{k}$):

- $Q_0 = P$
- For $i = 1 \rightarrow k$

  Compute a $\left(\frac{1}{k}, \epsilon, \alpha\right)$-approx q_i of Q_{i-1}

  $Q_i = closest\left\{Q_{i-1}, \{q_i\}, \frac{1-\epsilon}{2k}\right\}$
- Return (q_k, Q_k)

Sensitivity for Weighted k-means

- Consider the following algorithm:

Robust-Median$(\boldsymbol{P}, \boldsymbol{k})$:

- $Q_0 = P$
- For $i = 1 \to k$

  Compute a $\left(\frac{1}{k}, \epsilon, \alpha\right)$-approx q_i of Q_{i-1}

  $Q_i = closest\left\{Q_{i-1}, \{q_i\}, \frac{1-\epsilon}{2k}\right\}$
- Return (q_k, Q_k)

Lemma:

Let (q_k, Q_k) be the output of **Robust-Median**$(\boldsymbol{P}, \boldsymbol{k})$.
Then for every $p \in Q_k$:

$$s(p) = \max_{C \in Q} \frac{f_\omega(p, C)}{\sum_{q \in P} f_\omega(q, C)} \leq \frac{O(k)}{|Q_k|}$$

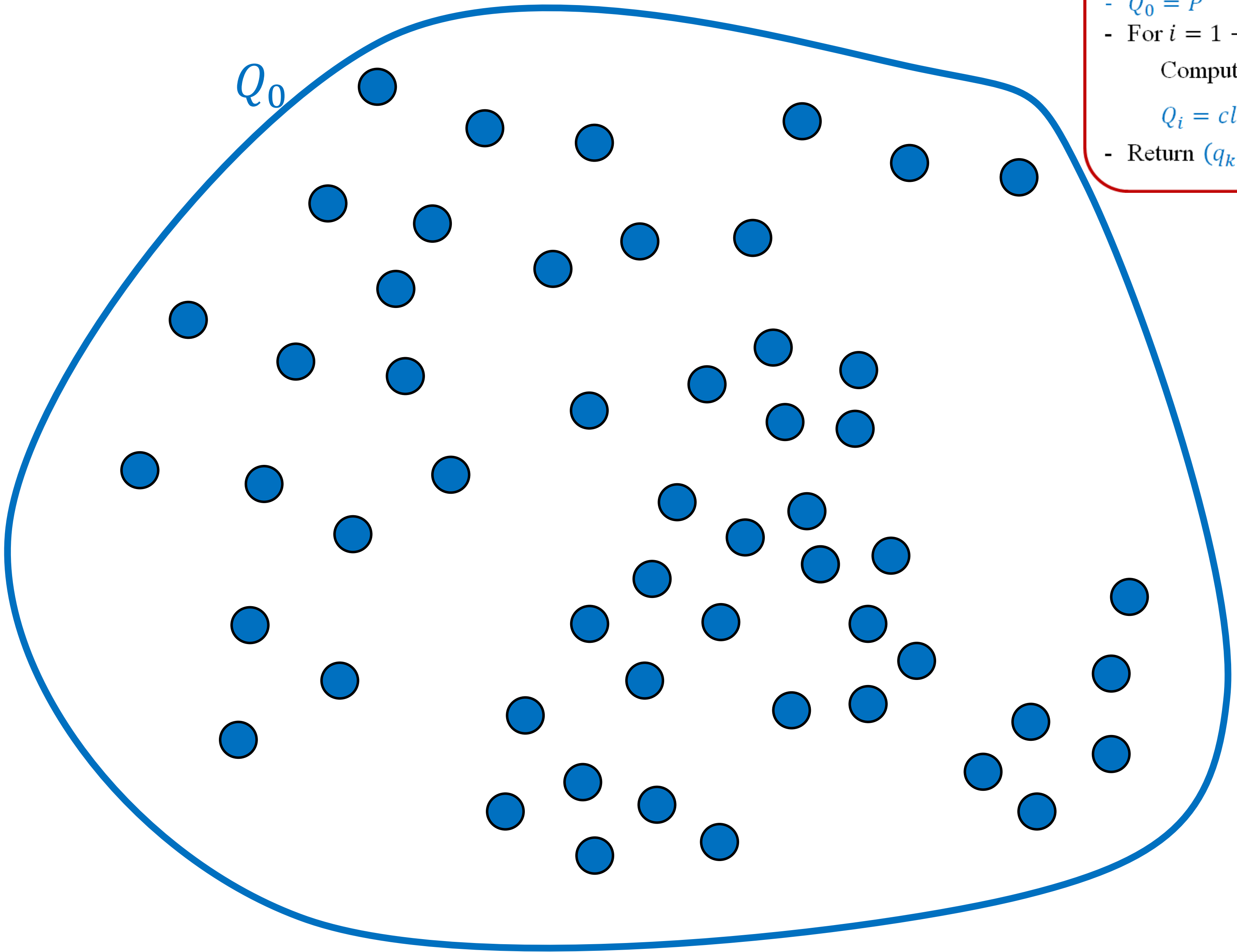
Sensitivity for Weighted k-means

<u>Example:</u>

$k = 2$

$\epsilon = \frac{1}{2}$

Iteration #1



<u>Robust-Median(P, k):</u>

- $Q_0 = P$
- For $i = 1 \rightarrow k$

  Compute a $\left(\frac{1}{k}, \epsilon, \alpha\right)$-approx q_i of Q_{i-1}

  $Q_i = closest\left\{Q_{i-1}, \{q_i\}, \frac{1-\epsilon}{2k}\right\}$
- Return (q_k, Q_k)

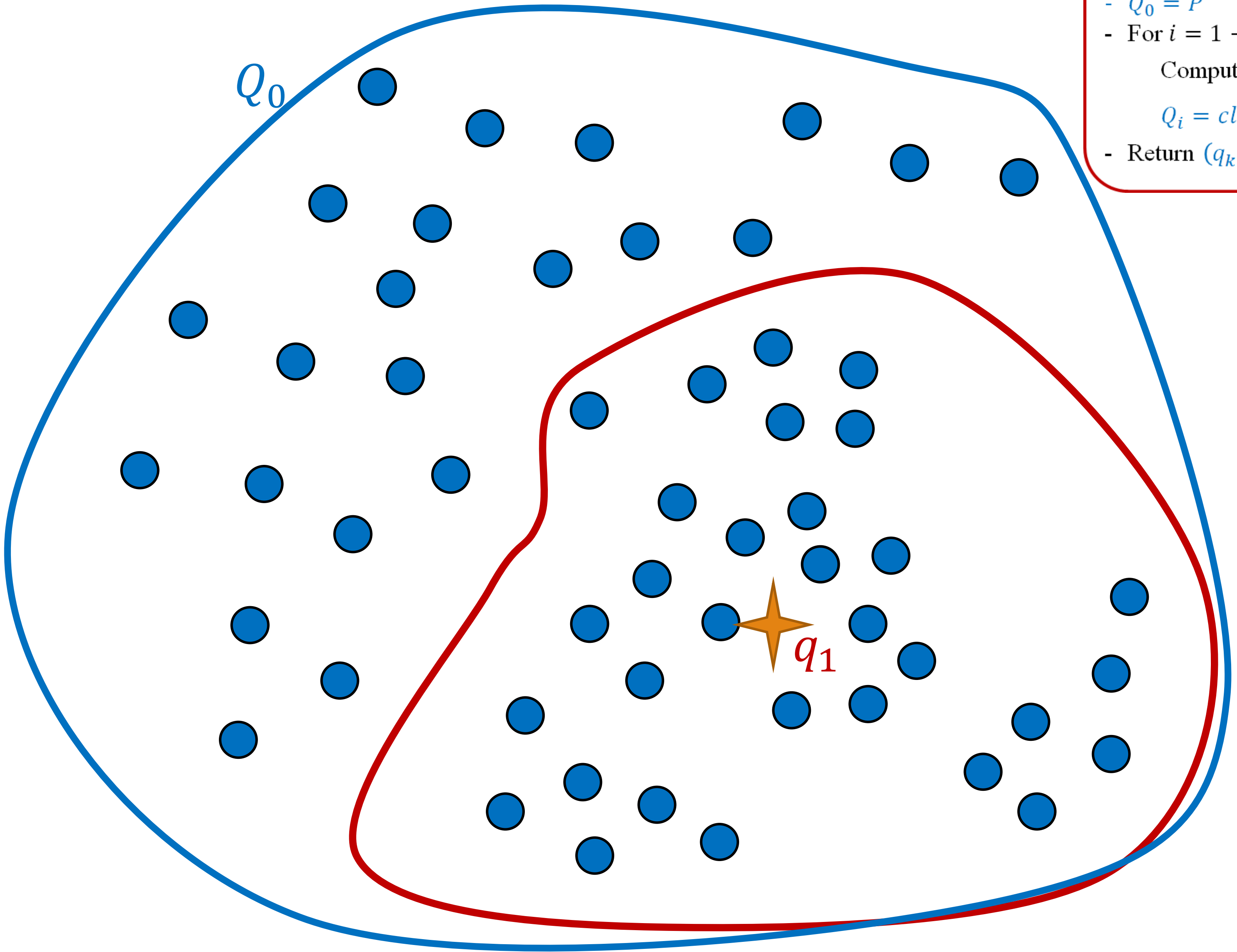
Sensitivity for Weighted k-means

Example:

$k = 2$

$\epsilon = \frac{1}{2}$

Iteration #1



Robust-Median(P, k):

- $Q_0 = P$
- For $i = 1 \rightarrow k$

  Compute a $\left(\frac{1}{k}, \epsilon, \alpha\right)$-approx q_i of Q_{i-1}

  $Q_i = closest\left\{Q_{i-1}, \{q_i\}, \frac{1-\epsilon}{2k}\right\}$
- Return (q_k, Q_k)

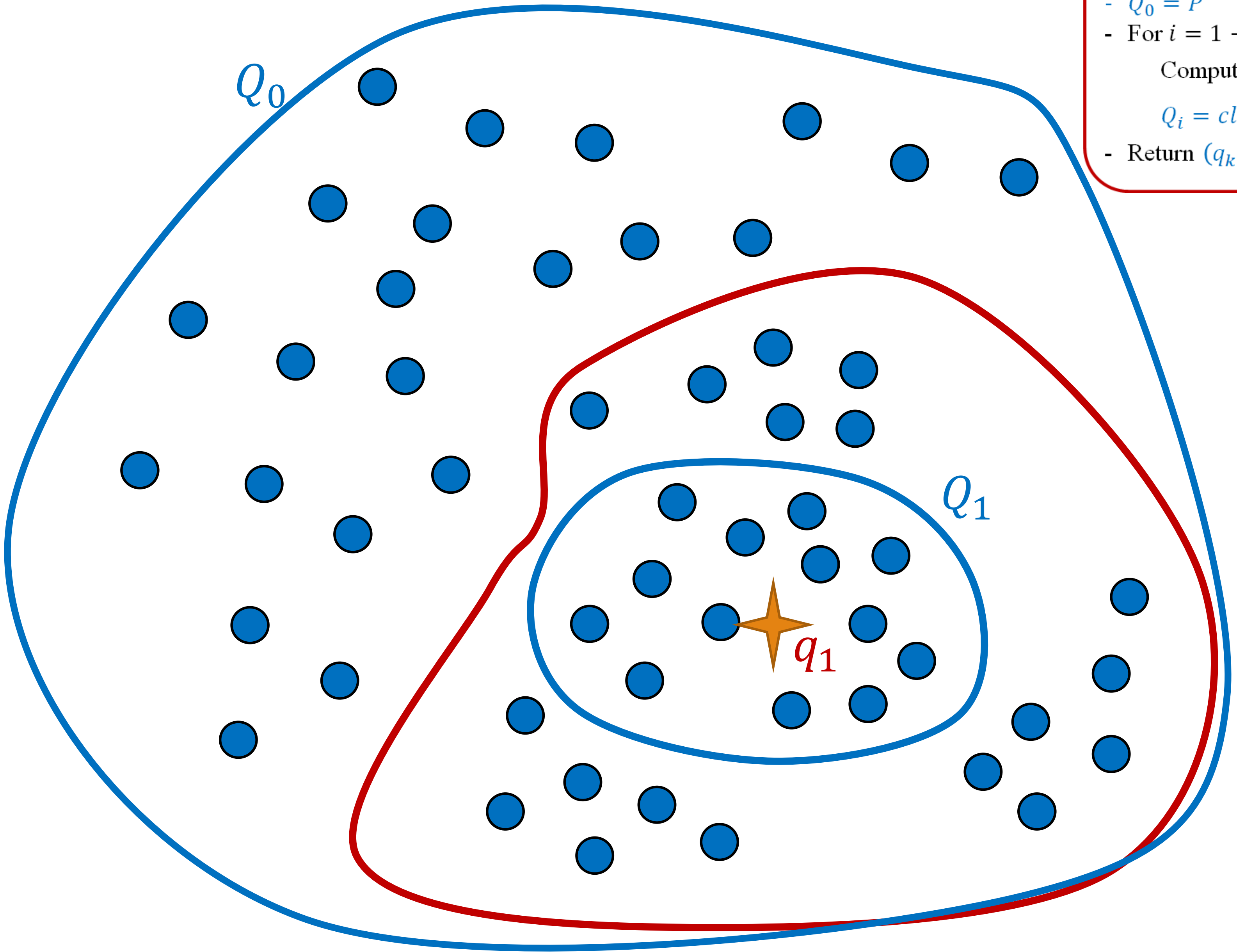
Sensitivity for Weighted k-means

Example:

$k = 2$

$\epsilon = \frac{1}{2}$

Iteration #1



Robust-Median(P, k):

- $Q_0 = P$
- For $i = 1 \to k$

  Compute a $\left(\frac{1}{k}, \epsilon, \alpha\right)$-approx q_i of Q_{i-1}

  $Q_i = closest\left\{Q_{i-1}, \{q_i\}, \frac{1-\epsilon}{2k}\right\}$
- Return (q_k, Q_k)

Sensitivity for Weighted k-means

Robust-Median(P, k):

- $Q_0 = P$
- For $i = 1 \to k$

  Compute a $\left(\frac{1}{k}, \epsilon, \alpha\right)$-approx q_i of Q_{i-1}

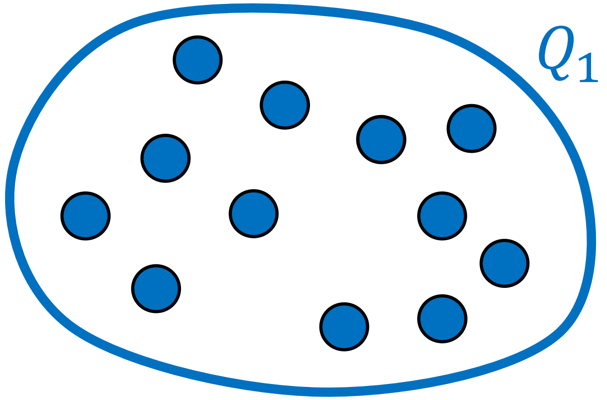
  $Q_i = closest\left\{Q_{i-1}, \{q_i\}, \frac{1-\epsilon}{2k}\right\}$
- Return (q_k, Q_k)

<u>Example:</u>

$k = 2$

$\epsilon = \frac{1}{2}$

Iteration #1



Sensitivity for Weighted k-means

Robust-Median(P, k):

- $Q_0 = P$
- For $i = 1 \to k$

  Compute a $\left(\frac{1}{k}, \epsilon, \alpha\right)$-approx q_i of Q_{i-1}

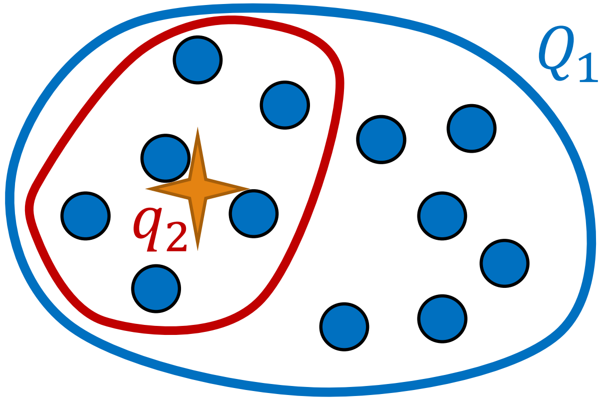
  $Q_i = closest\left\{Q_{i-1}, \{q_i\}, \frac{1-\epsilon}{2k}\right\}$
- Return (q_k, Q_k)

Example:

$k = 2$

$\epsilon = \frac{1}{2}$

Iteration #1



Sensitivity for Weighted k-means

Robust-Median(P, k):

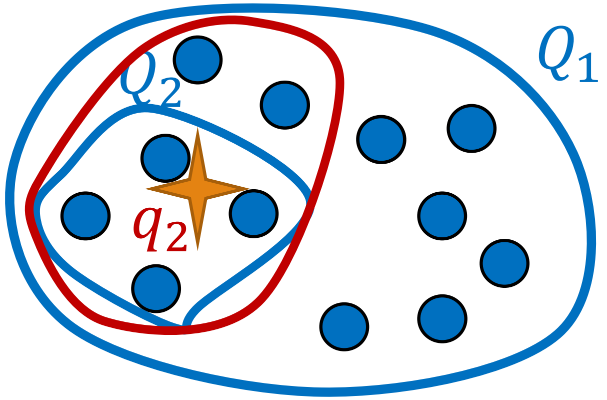
- $Q_0 = P$
- For $i = 1 \to k$
  - Compute a $\left(\frac{1}{k}, \epsilon, \alpha\right)$-approx q_i of Q_{i-1}
  - $Q_i = closest\left(Q_{i-1}, \{q_i\}, \frac{1-\epsilon}{2k}\right)$
- Return (q_k, Q_k)

<u>Example:</u>

$k = 2$

$\epsilon = \frac{1}{2}$

Iteration #1



Sensitivity for Weighted k-means

Robust-Median(P, k):

- $Q_0 = P$
- For $i = 1 \to k$

  Compute a $\left(\frac{1}{k}, \epsilon, \alpha\right)$-approx q_i of Q_{i-1}

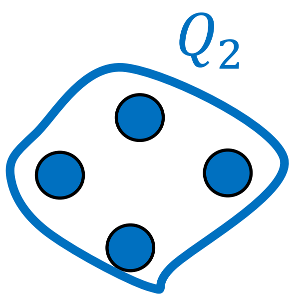
  $Q_i = closest\left\{Q_{i-1}, \{q_i\}, \frac{1-\epsilon}{2k}\right\}$
- Return (q_k, Q_k)

Example:

$k = 2$

$\epsilon = \frac{1}{2}$

Iteration #1



Sensitivity for Weighted k-means

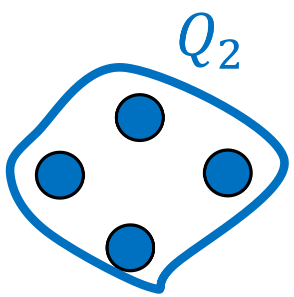
Robust-Median(P, k):
- $Q_0 = P$
- For $i = 1 \to k$
  - Compute a $\left(\frac{1}{k}, \epsilon, \alpha\right)$-approx q_i of Q_{i-1}
  - $Q_i = closest\left\{Q_{i-1}, \{q_i\}, \frac{1-\epsilon}{2k}\right\}$
- Return (q_k, Q_k)

Example:

$k = 2$

$\epsilon = \frac{1}{2}$

Iteration #1



$$\forall p \in Q_2 : s(p) \leq \frac{O(1)}{|Q_2|}$$

Sensitivity for Weighted k-means

Proof:

Consider the variables $Q_0, \dots, Q_k$ and $q_1, \dots, q_k$ that are computed in the algorithm.

- $p \in P$ is served by a weighted center $(c, \omega) \in C$ if $f_\omega(p, C) = \omega \cdot f(p, c)$.
- Let (c_i, ω_i) denote a center that serves at least $\frac{|Q_{i-1}|}{k}$ points from Q_{i-1} for every $i \in [k+1]$.
- Let P_i denote the points of P that are served by (c_i, ω_i).
- Let $Q_i' \coloneqq closest\left(Q_{i-1}, \{q_i\}, \frac{(1-\epsilon)}{k}\right)$, $f_i^* = \sum_{q \in Q_i'} f(q, q_i)$ for every $i \in [k]$.

Sensitivity for Weighted k-means

Proof:

Consider the variables $Q_0, \dots, Q_k$ and $q_1, \dots, q_k$ that are computed in the algorithm.

- $p \in P$ is served by a weighted center $(c, \omega) \in C$ if $f_\omega(p, C) = \omega \cdot f(p, c)$.
- Let (c_i, ω_i) denote a center that serves at least $\frac{|Q_{i-1}|}{k}$ points from $Q_{i-1}$for every $i \in [k+1]$.
- Let P_i denote the points of P that are served by (c_i, ω_i).
- Let $Q'_i \coloneqq closest\left(Q_{i-1}, \{q_i\}, \frac{(1-\epsilon)}{k}\right)$, $f_i^* = \sum_{q \in Q'_i} f(q, q_i)$ for every $i \in [k]$.

It follows that $|P_i \cap Q_{i-1}| \geq \frac{|Q_{i-1}|}{k} \geq |Q'_i|$

Sensitivity for Weighted k-means

Proof:

Consider the variables $Q_0, \ldots, Q_k$ and $q_1, \ldots, q_k$ that are computed in the algorithm.

- $p \in P$ is served by a weighted center $(c, \omega) \in C$ if $f_\omega(p, C) = \omega \cdot f(p, c)$.
- Let (c_i, ω_i) denote a center that serves at least $\frac{|Q_{i-1}|}{k}$ points from Q_{i-1} for every $i \in [k+1]$.
- Let P_i denote the points of P that are served by (c_i, ω_i).
- Let $Q_i' := closest\left(Q_{i-1}, \{q_i\}, \frac{(1-\epsilon)}{k}\right)$, $f_i^* = \sum_{q \in Q_i'} f(q, q_i)$ for every $i \in [k]$.

$$f^*(Q_i, \gamma) = \min_{C \in Q} \sum_{p \in closest(Q_i, C, \gamma)} f(p, C)$$

It follows that $|P_i \cap Q_{i-1}| \geq \frac{|Q_{i-1}|}{k} \geq |Q_i'|$.

$$\rightarrow \sum_{q \in P_i \cap Q_{i-1}} f(q, c_i) \geq f^*\left(Q_{i-1}, \frac{1}{k}\right)$$

Sensitivity for Weighted k-means

Proof:

Case (i):

There is $i \in [k]$ such that: $f(p, c_i) \leq 16\phi\rho\alpha \cdot \frac{f_i^*}{|Q_k'|}$.

Case (ii):

Otherwise.

Sensitivity for Weighted k-means

Proof:

Case (i):

There is $i \in [k]$ such that: $f(p, c_i) \leq 16\phi\rho\alpha \cdot \frac{f_i^*}{|Q_k'|}$.

Case (ii):

Otherwise.

Proof of Case (ii):

By the pigeonhole principle, $c_i = c_j$ for some $i, j \in [k+1]$, i < j.

Put $q \in P_j \cap Q_{j-1}$. Note that $p \in Q_k \subseteq Q_{j-1}$.

Using the Markov inequality,

$$f(q, q_{j-1}), f(p, q_{j-1}) \leq \frac{2f_{j-1}^*}{|Q_{j-1}'|}$$

Sensitivity for Weighted k-means

Proof of Case (ii):

By the pigeonhole principle, $c_i = c_j$ for some $i, j \in [k+1]$, i < j.

Put $q \in P_j \cap Q_{j-1}$. Note that $p \in Q_k \subseteq Q_{j-1}$.

Using the Markov inequality,

$$f(q, q_{j-1}), f(p, q_{j-1}) \leq \frac{2f_{j-1}^*}{|Q'_{j-1}|}$$

Notice that

$$f(p,q) \leq \rho\left(f(p, q_{j-1}) + f(q_{j-1}, q)\right) \leq \rho\left(\frac{2f_{j-1}^*}{|Q'_{j-1}|} + \frac{2f_{j-1}^*}{|Q'_{j-1}|}\right) \leq \frac{4\rho \cdot f_{j-1}^*}{|Q'_{j-1}|}$$

Weak triangle inequality

$$\rightarrow f(p,q) \leq \frac{4\rho \cdot f_{j-1}^*}{|Q'_{j-1}|}$$

Sensitivity for Weighted k-means

<u>**Proof of Case (ii):**</u>

$$\rightarrow f\left(p, c_j\right) - f\left(q, c_j\right) \leq \phi f(p, q) + \frac{f\left(p, c_j\right)}{4}$$

$$f(p, q) - f(q, c) \leq \phi f(p, q) + \frac{f(p, c)}{4}$$

Sensitivity for Weighted k-means

Proof of Case (ii):

$$\rightarrow f(p, c_j) - f(q, c_j) \leq \phi f(p, q) + \frac{f(p, c_j)}{4}$$

$$\leq \frac{4\phi\rho \cdot f_{j-1}^*}{|Q'_{j-1}|} + \frac{f(p, c_j)}{4}$$

Proved in last slide

Sensitivity for Weighted k-means

Proof of Case (ii):

$$\rightarrow f(p, c_j) - f(q, c_j) \leq \phi f(p, q) + \frac{f(p, c_j)}{4}$$

$$\leq \frac{4\phi\rho \cdot f_{j-1}^*}{|Q'_{j-1}|} + \frac{f(p, c_j)}{4}$$

$$\leq \frac{4\phi\rho\alpha \cdot f_i^*}{|Q'_k|} + \frac{f(p, c_j)}{4}$$

$Q_k \subseteq Q_{j-1} \rightarrow |Q_k| \leq |Q_{j-1}| \rightarrow |Q'_k| \leq |Q'_{j-1}|$
and $f_{j-1}^* \leq \alpha f_i^*$.

Sensitivity for Weighted k-means

Proof of Case (ii):

$$\rightarrow f\left(p, c_j\right) - f\left(q, c_j\right) \leq \phi f(p, q) + \frac{f\left(p, c_j\right)}{4}$$

$$\leq \frac{4\phi\rho \cdot f_{j-1}^*}{\left|Q_{j-1}'\right|} + \frac{f\left(p, c_j\right)}{4}$$

$$\leq \frac{4\phi\rho\alpha \cdot f_i^*}{\left|Q_k'\right|} + \frac{f\left(p, c_j\right)}{4}$$

$$< \frac{f(p, c_i)}{4} + \frac{f\left(p, c_j\right)}{4}$$

Since Case (i) doesn't hold:

$$16\phi\rho\alpha \cdot \frac{f_i^*}{\left|Q_k'\right|} < f(p, c_i) \rightarrow 4\phi\rho\alpha \cdot \frac{f_i^*}{\left|Q_k'\right|} < \frac{f(p, c_i)}{4}$$

Sensitivity for Weighted k-means

Proof of Case (ii):

$$\rightarrow f(p, c_j) - f(q, c_j) \leq \phi f(p, q) + \frac{f(p, c_j)}{4}$$

$$\leq \frac{4\phi\rho \cdot f_{j-1}^*}{|Q'_{j-1}|} + \frac{f(p, c_j)}{4}$$

$$\leq \frac{4\phi\rho\alpha \cdot f_i^*}{|Q'_k|} + \frac{f(p, c_j)}{4}$$

$$< \frac{f(p, c_i)}{4} + \frac{f(p, c_j)}{4}$$

$$= \frac{f(p, c_j)}{4} + \frac{f(p, c_j)}{4}$$

$c_i = c_j$

Sensitivity for Weighted k-means

Proof of Case (ii):

$$
\begin{aligned}
\rightarrow f(p, c_j) - f(q, c_j) &\leq \phi f(p, q) + \frac{f(p, c_j)}{4} \\
&\leq \frac{4\phi\rho \cdot f_{j-1}^*}{|Q'_{j-1}|} + \frac{f(p, c_j)}{4} \\
&\leq \frac{4\phi\rho \cdot f_i^*}{|Q'_k|} + \frac{f(p, c_j)}{4} \\
&< \frac{f(p, c_i)}{4} + \frac{f(p, c_j)}{4} \\
&= \frac{f(p, c_j)}{4} + \frac{f(p, c_j)}{4} = \frac{f(p, c_j)}{2}
\end{aligned}
$$

Sensitivity for Weighted k-means

Proof of Case (ii):

$$\rightarrow f(p, c_j) - f(q, c_j) \leq \phi f(p, q) + \frac{f(p, c_j)}{4}$$

$$\leq \frac{4\phi\rho \cdot f_{j-1}^*}{|Q'_{j-1}|} + \frac{f(p, c_j)}{4}$$

$$\leq \frac{4\phi\rho \cdot f_i^*}{|Q'_k|} + \frac{f(p, c_j)}{4}$$

$$< \frac{f(p, c_i)}{4} + \frac{f(p, c_j)}{4}$$

$$= \frac{f(p, c_j)}{4} + \frac{f(p, c_j)}{4} = \frac{f(p, c_j)}{2}$$

$$\rightarrow f(q, c_j) > \frac{f(p, c_j)}{2}$$

Sensitivity for Weighted k-means

Proof:

$$\rightarrow \frac{f_\omega(p,C)}{\sum_{q\in P} f_\omega(q,C)} < \frac{f(p,c_j)}{\sum_{q\in P_j\cap Q_{j-1}} f(q,c_j)}$$

$$< \frac{2\cdot f(p,c_j)}{\sum_{q\in P_j\cap Q_{j-1}} f(p,c_j)}$$

$$= \frac{2\cdot f(p,c_j)}{f(p,c_j)\cdot |P_j\cap Q_{j-1}|}$$

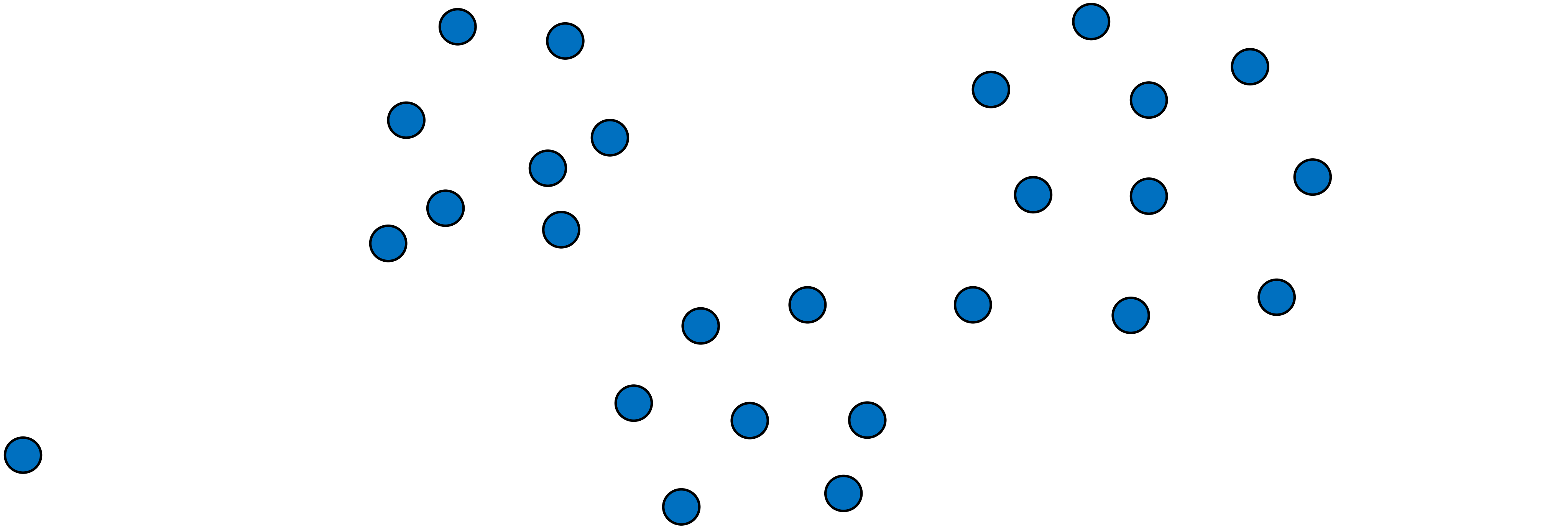
$$\le \frac{2k}{|Q_{j-1}|}$$

$$\le \frac{2k}{|Q_j|}$$

k-means With Outliers

<u>Definition:</u>

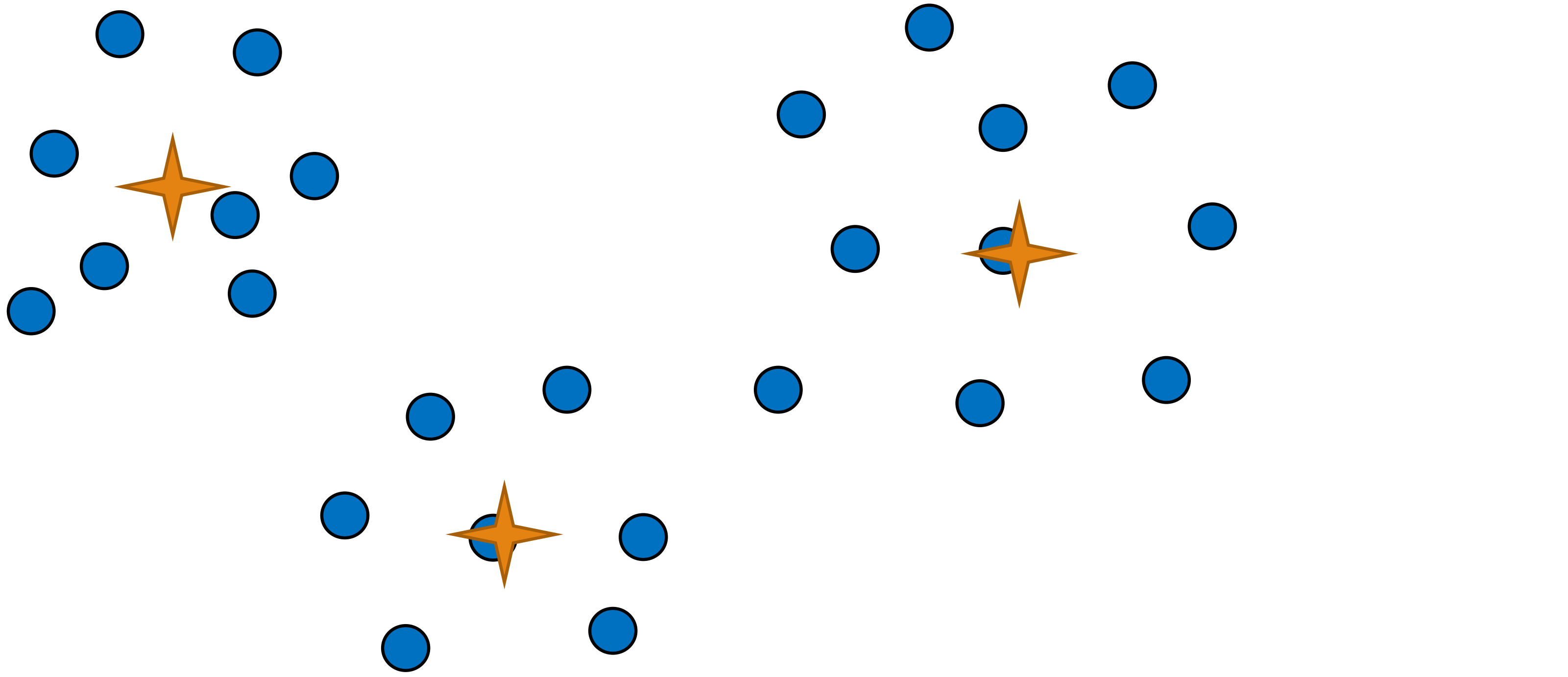
Find k centers that minimize sum of squared distances to the closest $n - m$ points i.e., ignore the farthest m points (outliers).



k-means With Outliers

<u>Definition:</u>
Find k centers that minimize sum of squared distances to the closest $n - m$ points i.e., ignore the farthest m points (outliers).

<u>Example:</u>
$k = 3, m = 3$



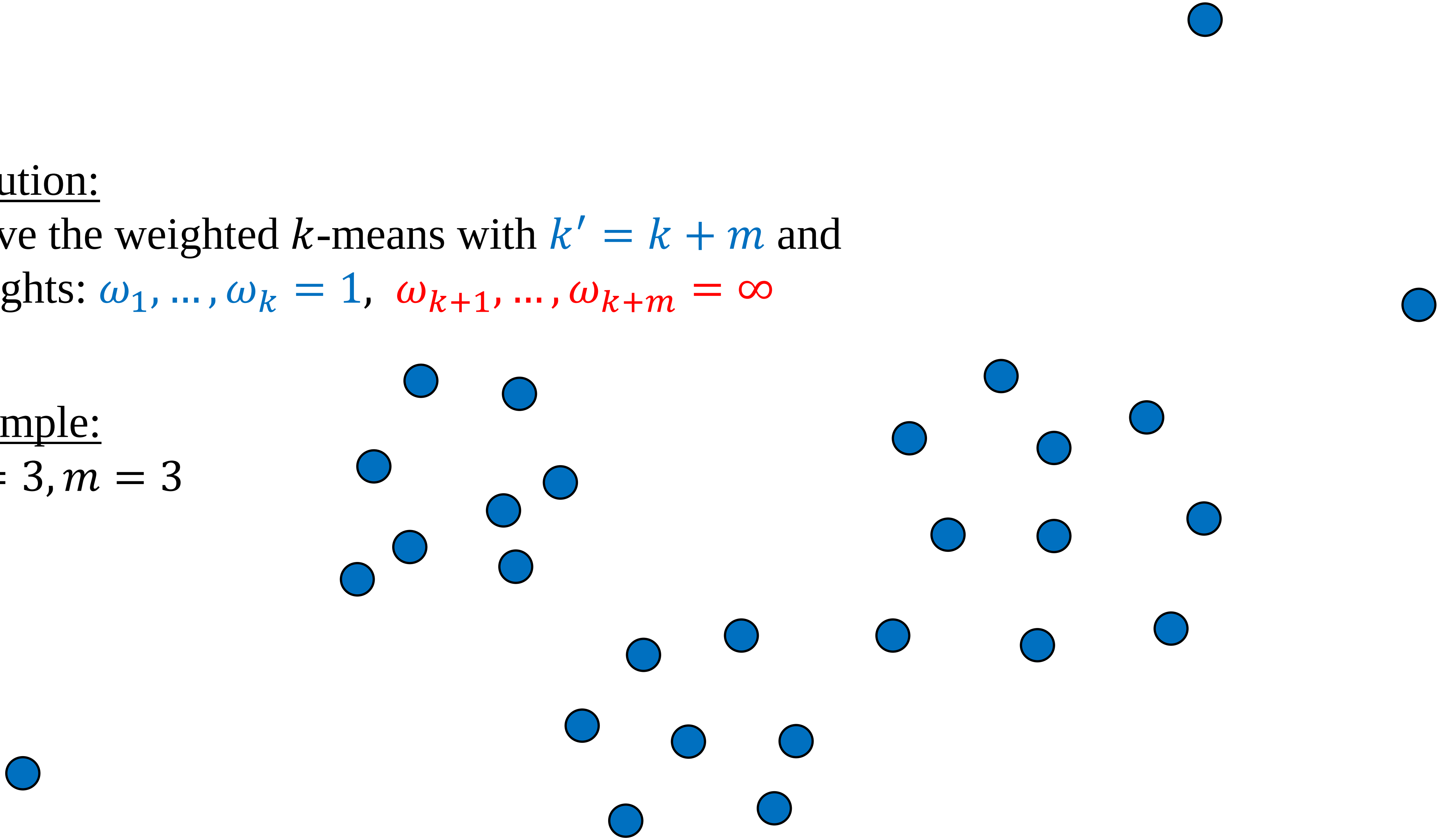
k-means With Outliers

Solution:

Solve the weighted k-means with $k' = k + m$ and weights: $\omega_1, \ldots, \omega_k = 1$, $\omega_{k+1}, \ldots, \omega_{k+m} = \infty$

Example:

$k = 3, m = 3$



k-means With Outliers

<u>Solution:</u>

Solve the weighted k-means with $k' = k + m$ and weights: $\omega_1, \dots, \omega_k = 1$, $\omega_{k+1}, \dots, \omega_{k+m} = \infty$

Will automatically be assigned to the outliers

<u>Example:</u>

$k = 3, m = 3$