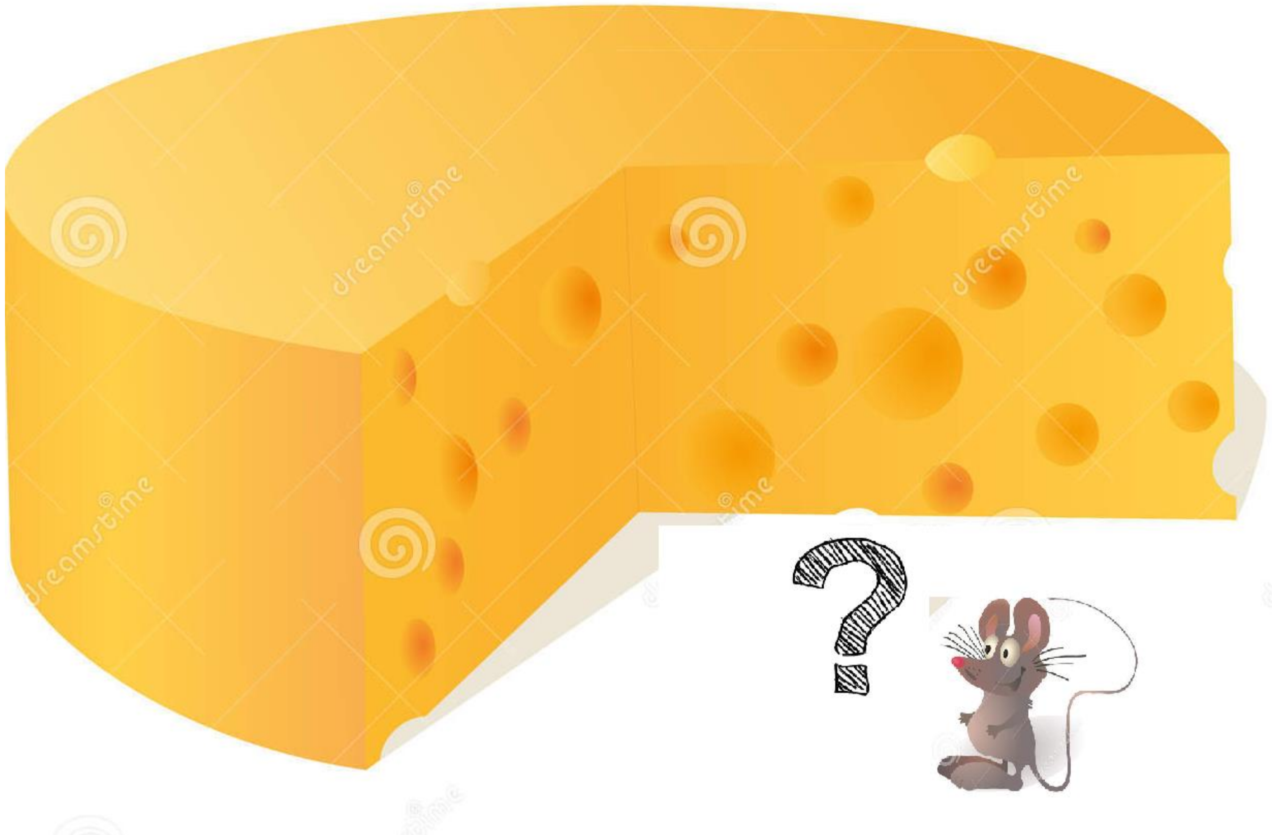
Big Data Class



LECTURER: DAN FELDMAN

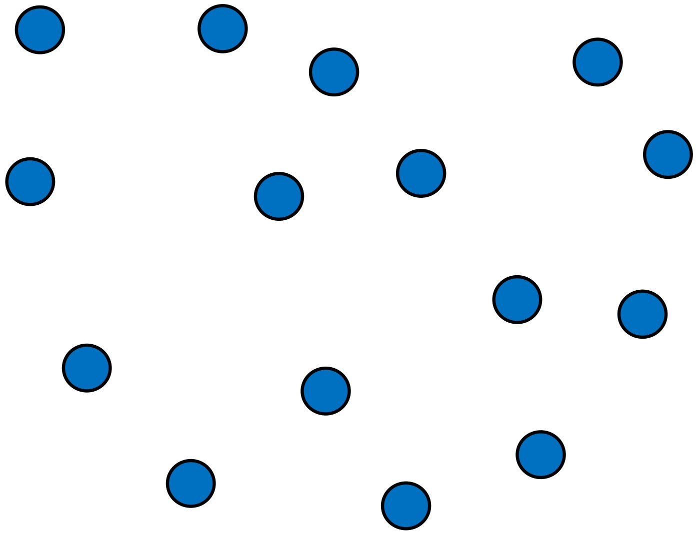
TEACHING ASSISTANTS:

IBRAHIM JUBRAN

ALAA MAALOUF

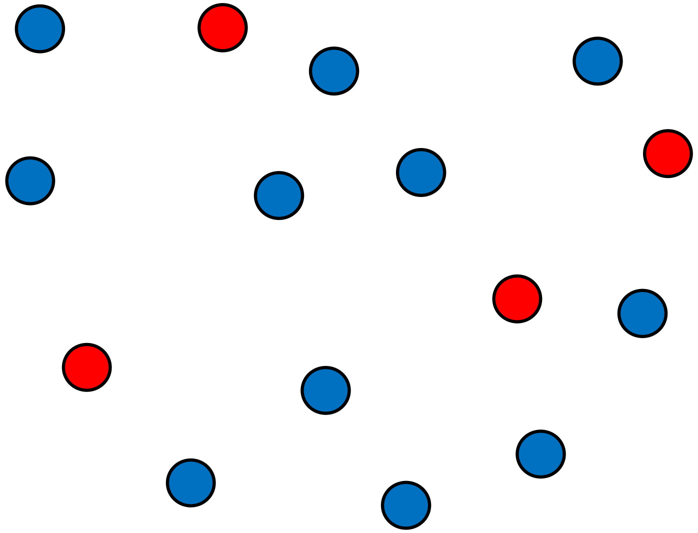
Coreset for 1-Line in R^2

- Input: $P \subseteq R^2$
- Query space: $Q = \{\ell \mid \ell \text{ is a line in } R^2\}$
- Cost function: $dist(p, \ell) = \min_{x \in \ell} \|p - x\|_2$
- Output: $C \subseteq P \ s.t. \ \forall \ell \in Q: \max_{p \in P} dist(p, \ell) - \max_{c \in C} dist(c, \ell) \leq \epsilon \cdot \max_{p \in P} dist(p, \ell)$



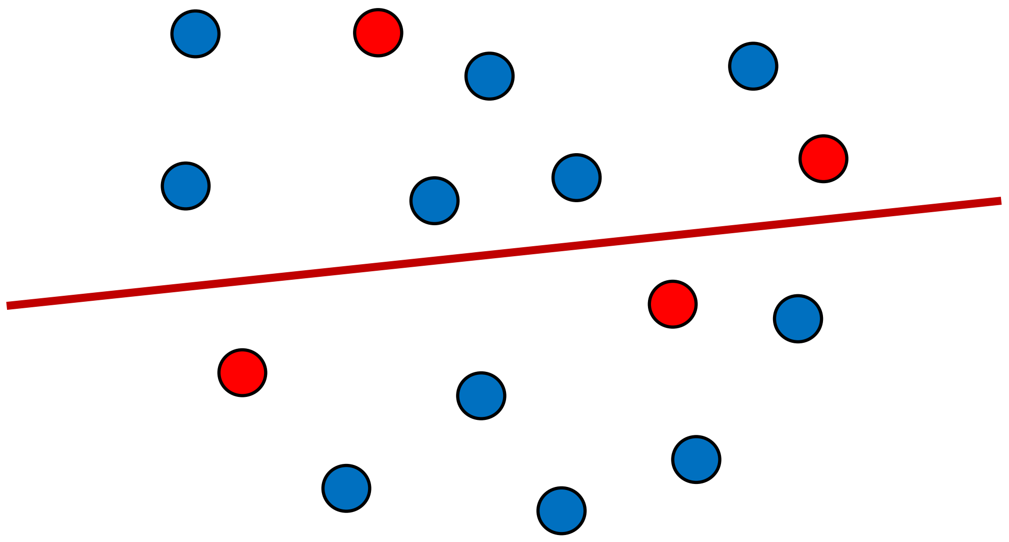
Coreset for 1-Line in R^2

- Input: $P \subseteq R^2$
- Query space: $Q = \{\ell \mid \ell \text{ is a line in } R^2\}$
- Cost function: $dist(p,\ell) = \min_{x\in\ell} \|p-x\|_2$
- Output: $C \subseteq P$ s.t. $\forall \ell \in Q: \max_{p\in P} dist(p,\ell) - \max_{c\in C} dist(c,\ell) \le \epsilon \cdot \max_{p\in P} dist(p,\ell)$



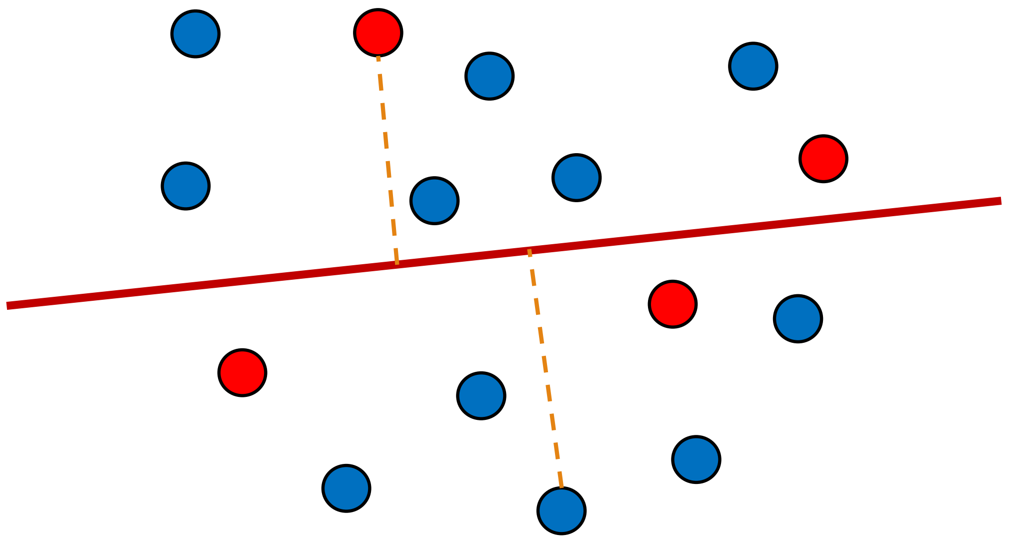
Coreset for 1-Line in R^2

- Input: $P \subseteq R^2$
- Query space: $Q = \{\ell \mid \ell \text{ is a line in } R^2\}$
- Cost function: $dist(p,\ell) = \min_{x\in\ell} \|p - x\|_2$
- Output: $C \subseteq P \text{ s.t. } \forall \ell \in Q: \max_{p\in P} dist(p,\ell) - \max_{c\in C} dist(c,\ell) \le \epsilon \cdot \max_{p\in P} dist(p,\ell)$

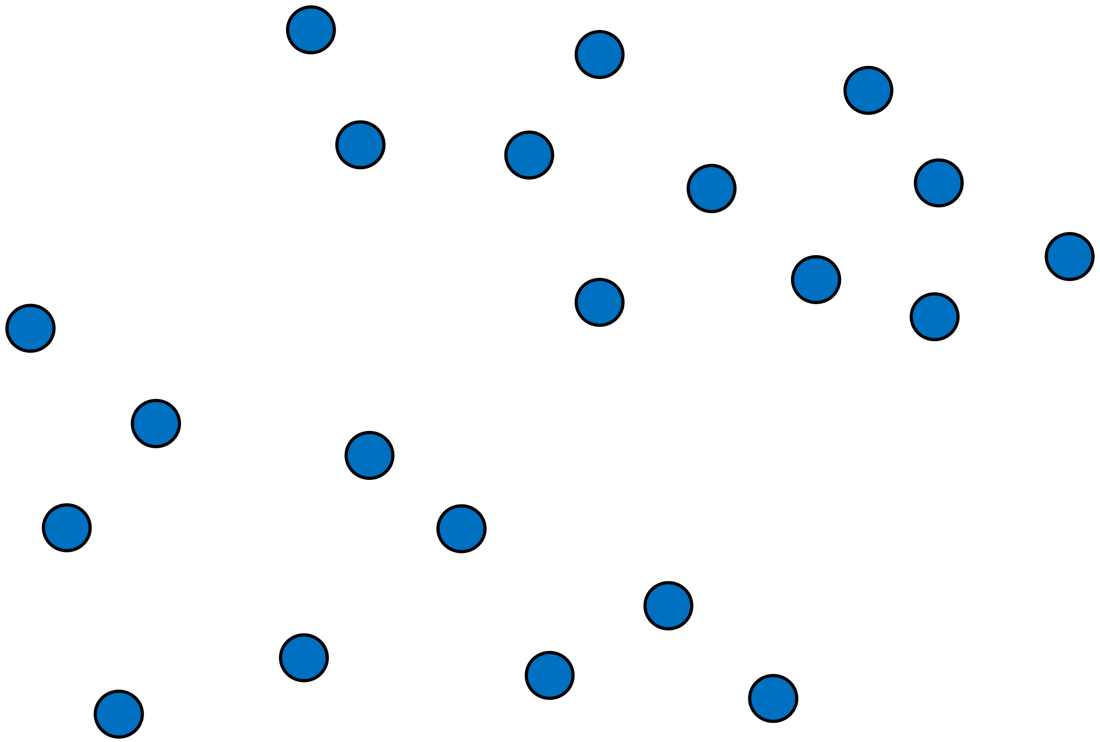


Coreset for 1-Line in R^2

- Input: $P \subseteq R^2$
- Query space: $Q = \{\ell \mid \ell \text{ is a line in } R^2\}$
- Cost function: $dist(p, \ell) = \min_{x \in \ell} \|p - x\|_2$
- Output: $C \subseteq P \text{ s.t. } \forall \ell \in Q: \max_{p \in P} dist(p, \ell) - \max_{c \in C} dist(c, \ell) \leq \epsilon \cdot \max_{p \in P} dist(p, \ell)$



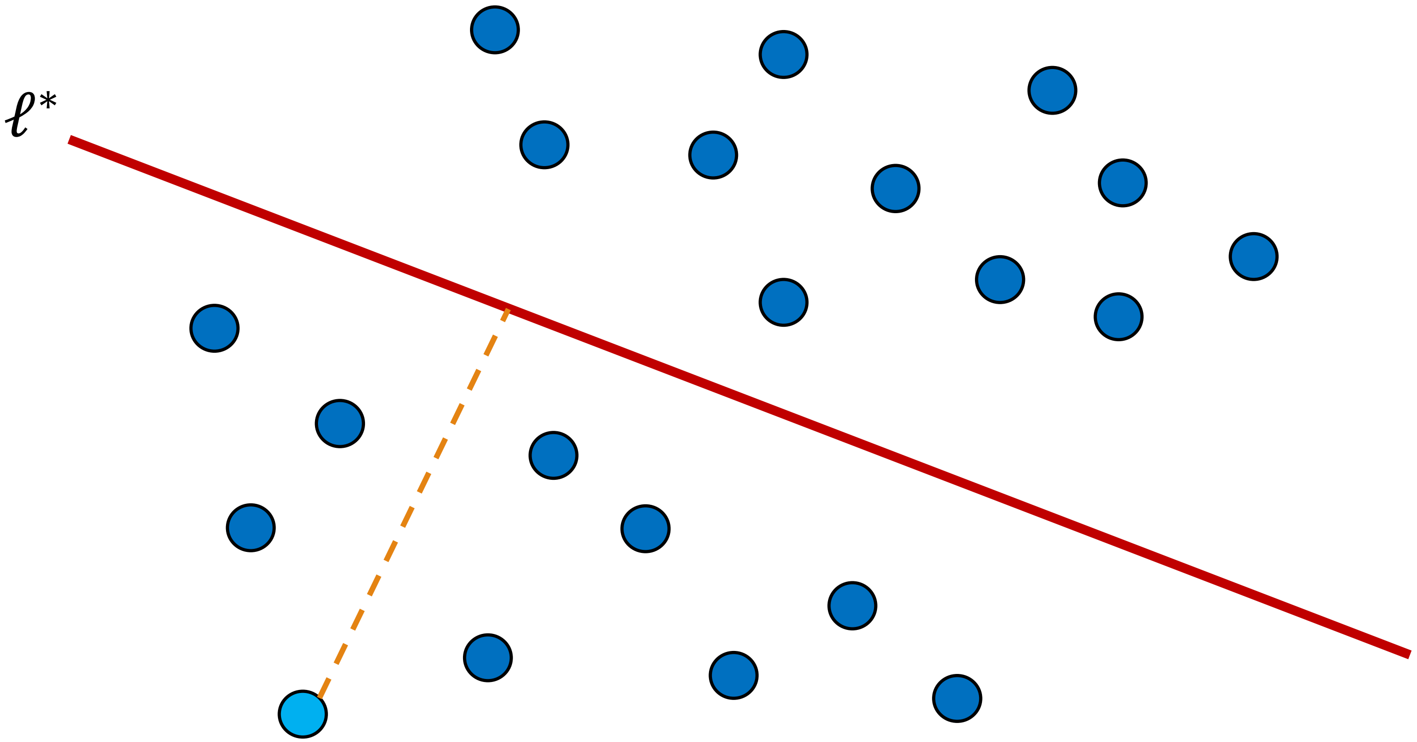
Coreset for 1-Line in R^2



Coreset for 1-Line in R^2

ℓ^* is the line that minimizes

$$\max_{p \in P} dist(p, \ell)$$



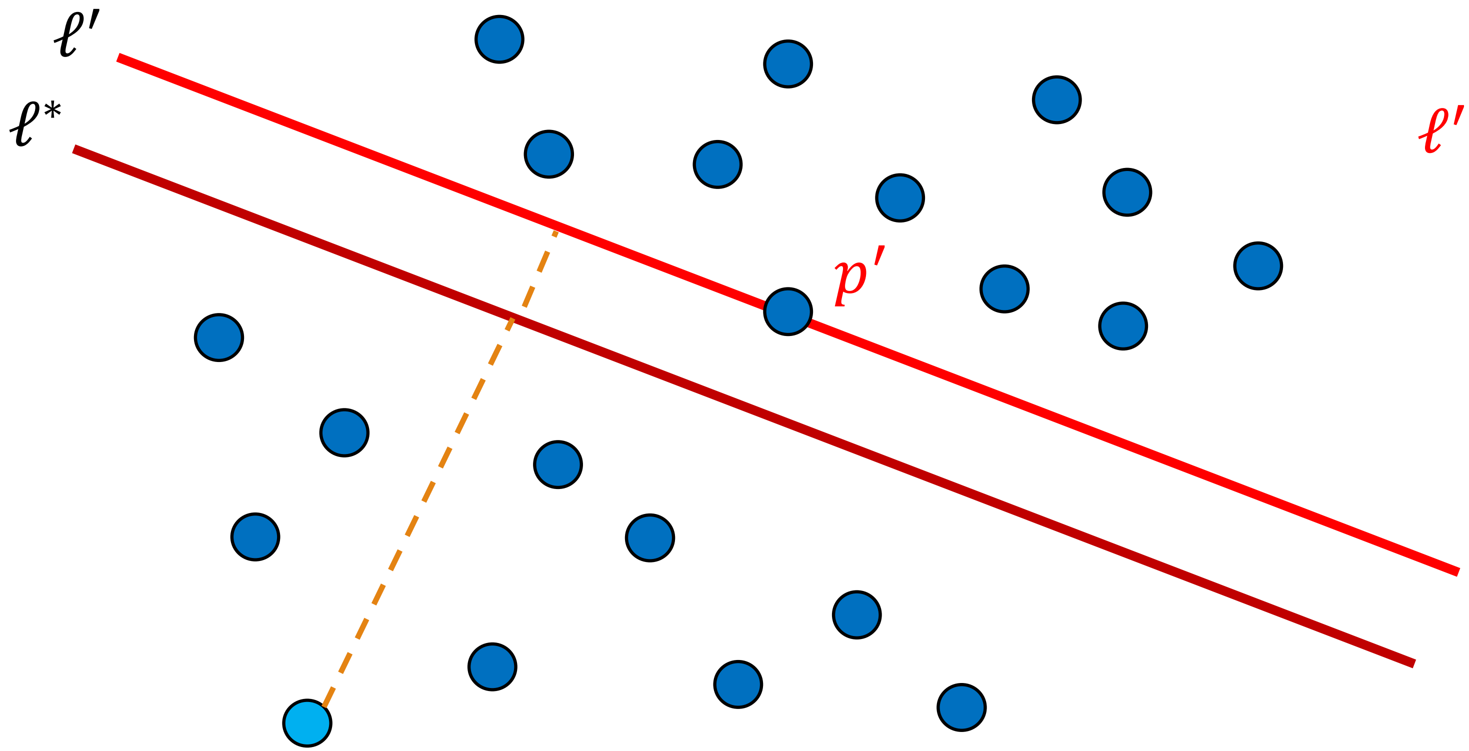
$$p^* = arg \max_{p \in P} dist(p, \ell^*)$$

Coreset for 1-Line in R^2

ℓ^* is the line that minimizes

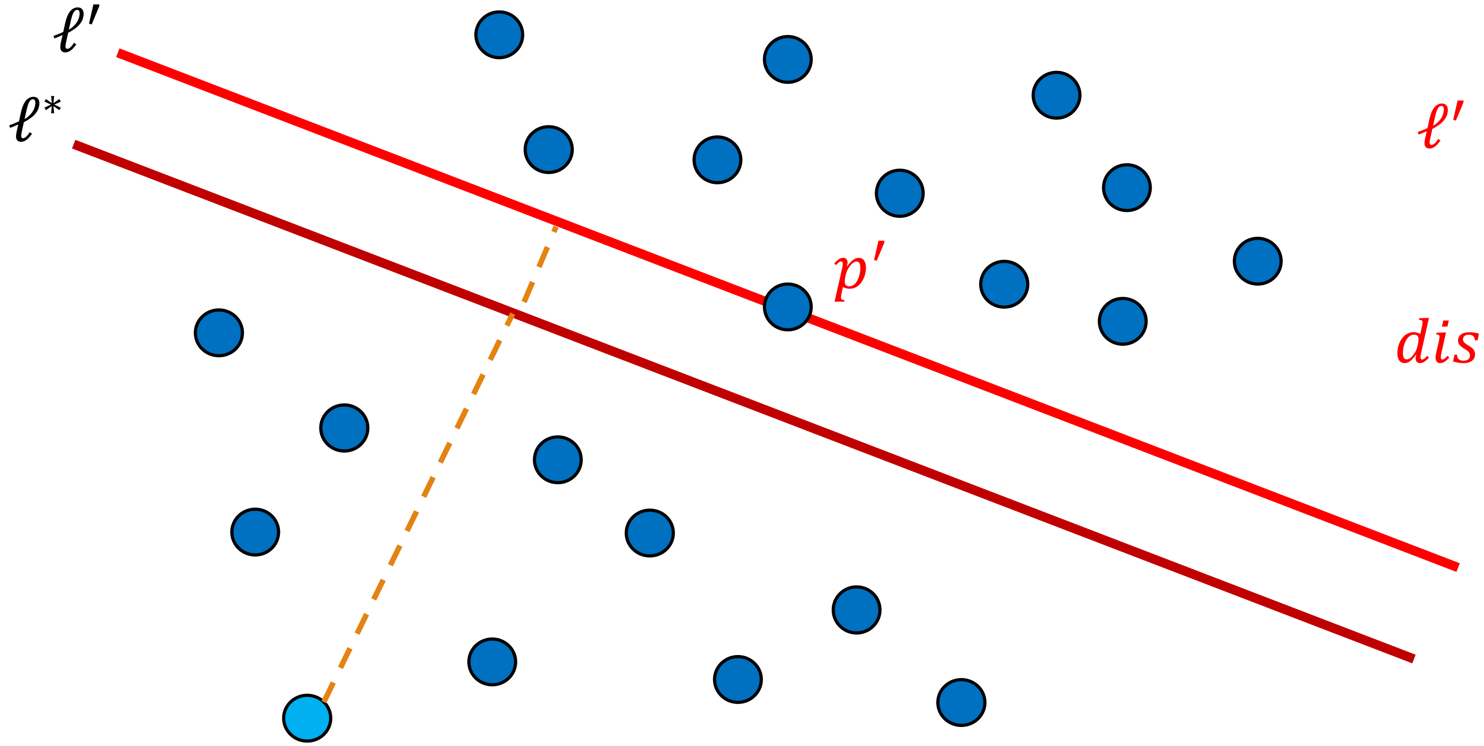
$\max_{p \in P} dist(p, \ell)$

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



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

Coreset for 1-Line in R^2



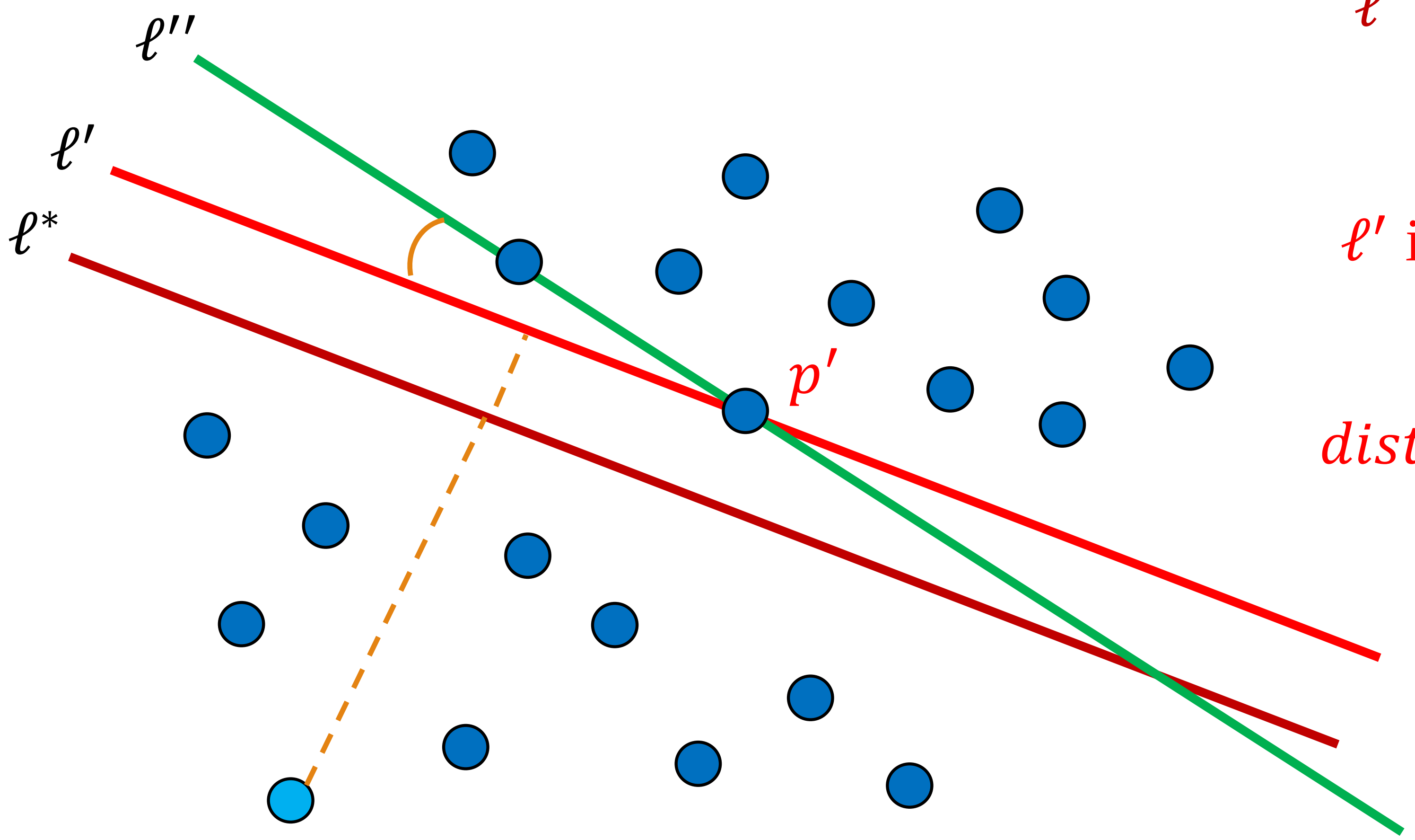
$p^* = arg \max_{p \in P} dist(p, \ell^*)$

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'

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

Coreset for 1-Line in R^2



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

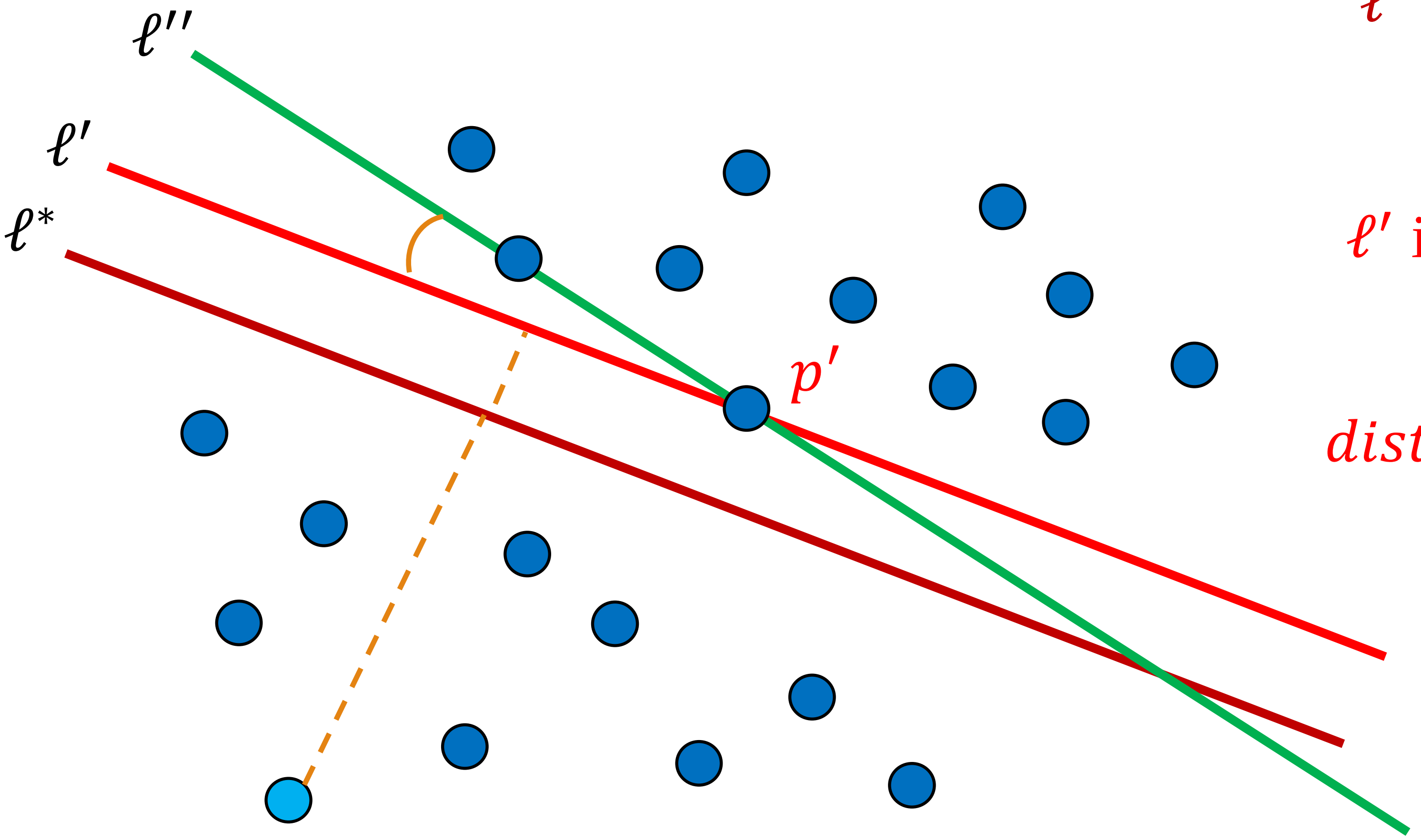
ℓ' is the translation of ℓ^* to $\ell^{*\prime}s$ closest point p'

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

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

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

Coreset for 1-Line in R^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'

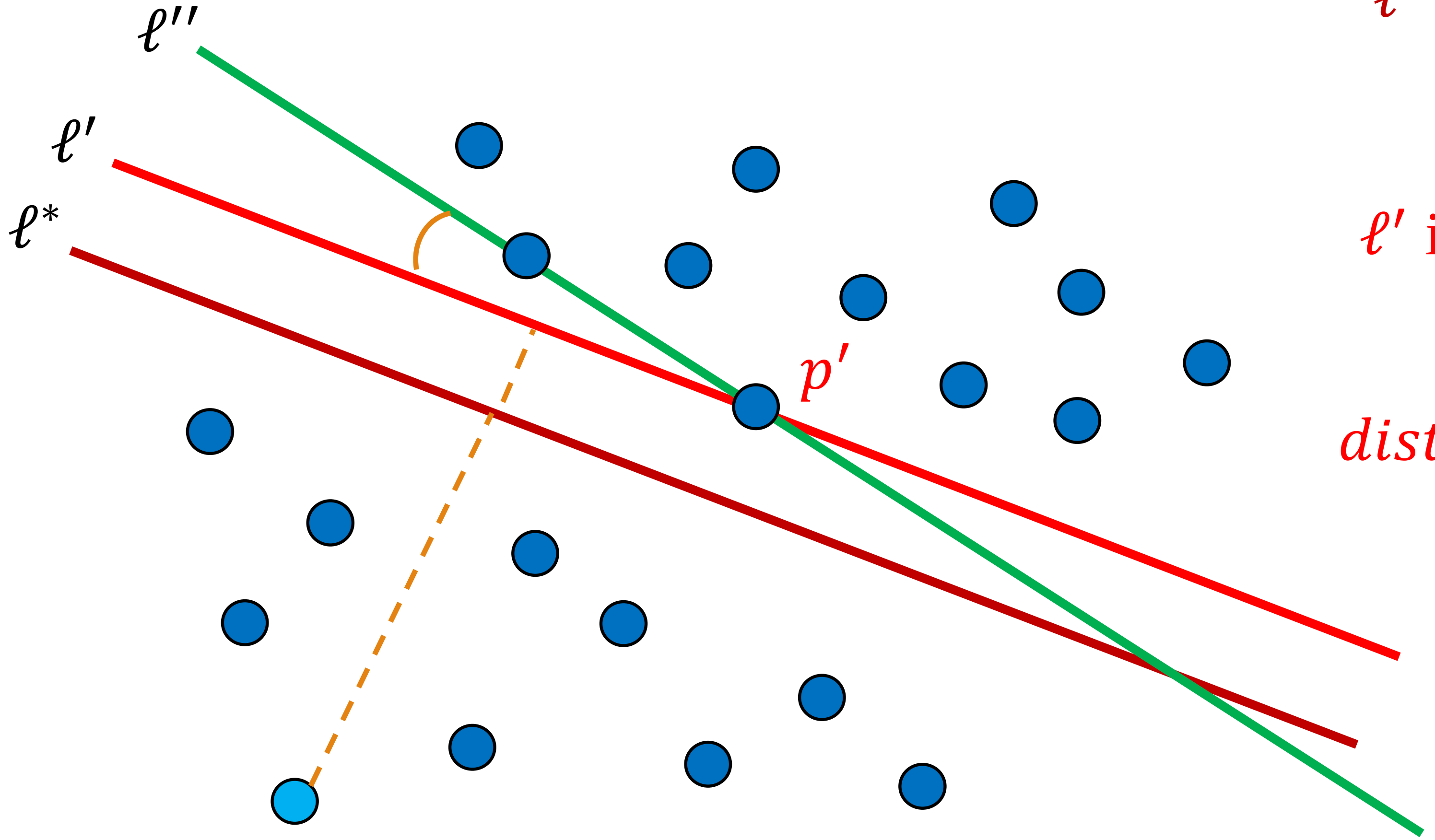
$dist(p, \ell') \leq 2 \cdot dist(p, \ell^*)$

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

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

Coreset for 1-Line in R^2

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



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

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'

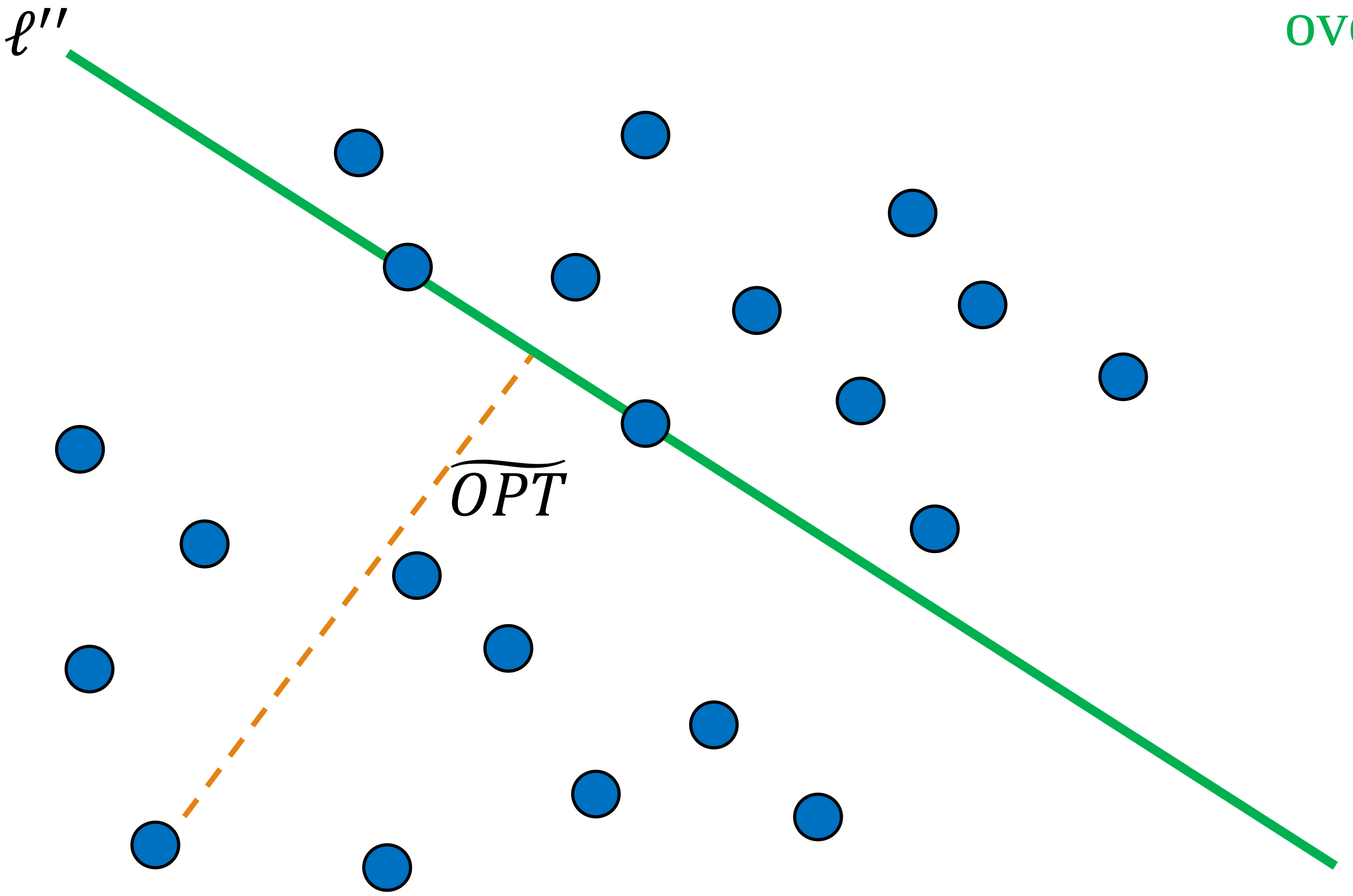
$dist(p, \ell') \leq 2 \cdot dist(p, \ell^*)$

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

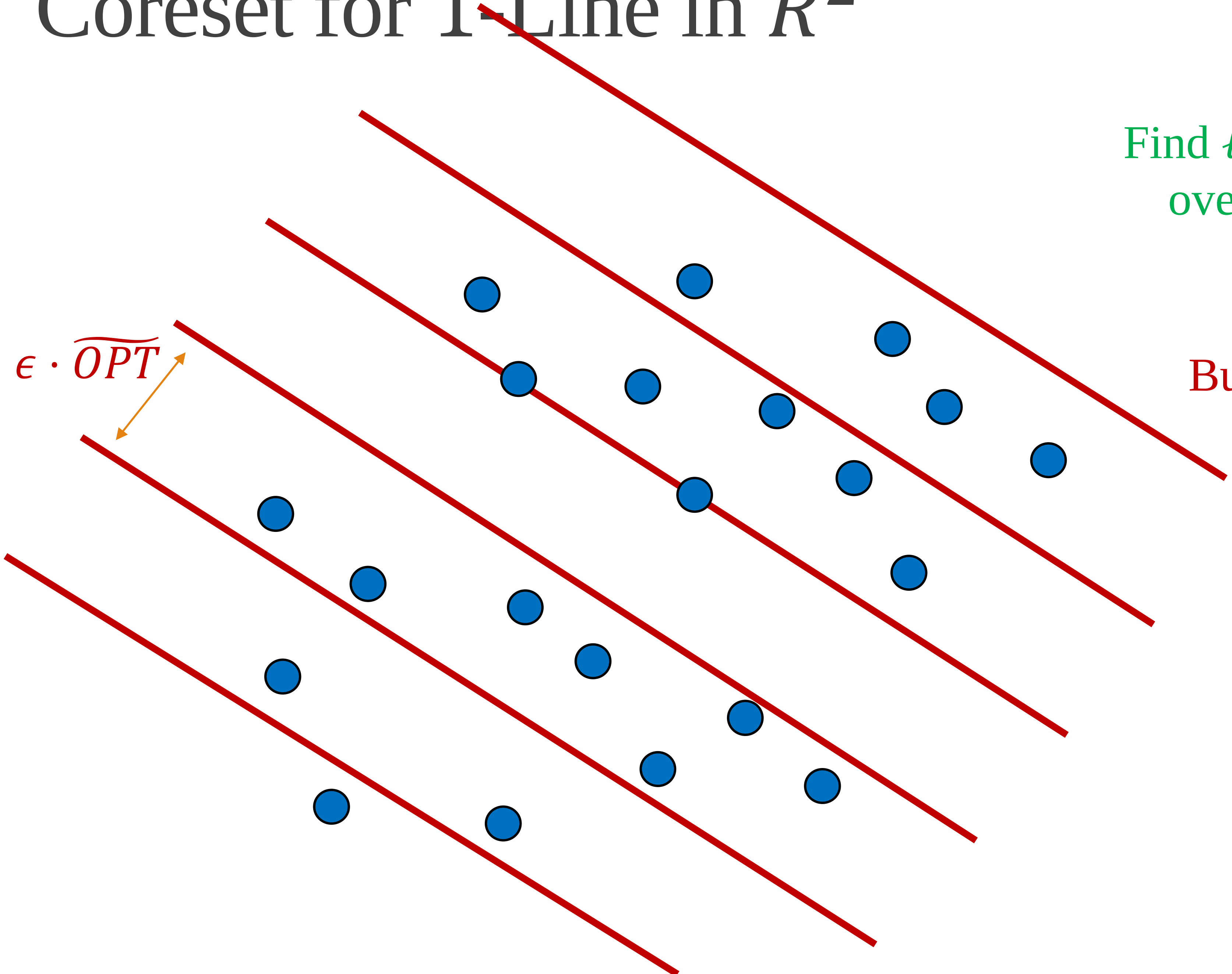
$dist(p, \ell'') \leq 2 \cdot dist(p, \ell')$

Coreset for 1-Line in R^2

Find ℓ'' by exhaustive search over every pair of points.
$\boldsymbol{O(n^2)}$



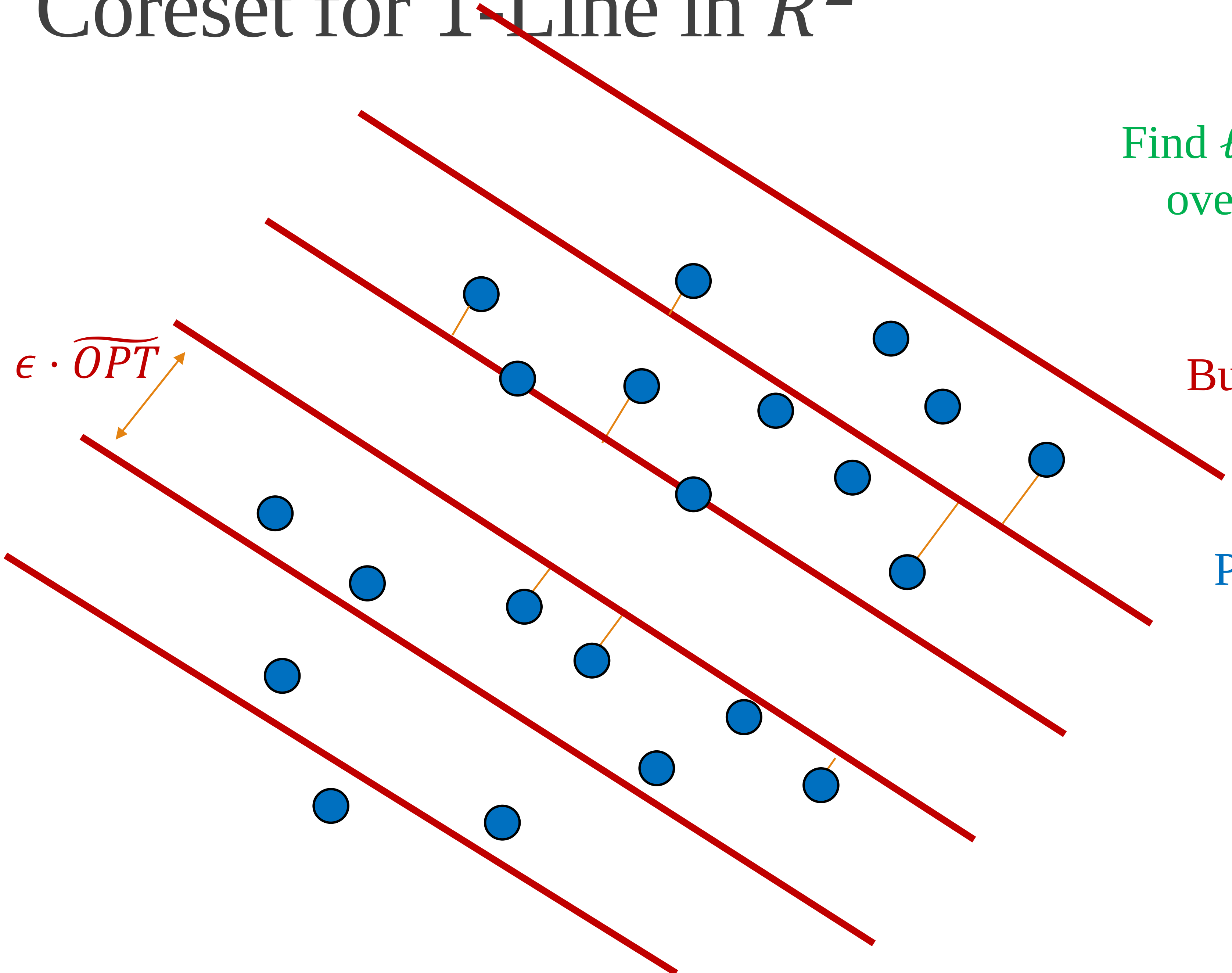
Coreset for 1-Line in R^2



Find ℓ'' by exhaustive search over every pair of points.
$\boldsymbol{O(n^2)}$

Build a grid of lines with $\epsilon \cdot \widetilde{OPT}$ distance

Coreset for 1-Line in R^2

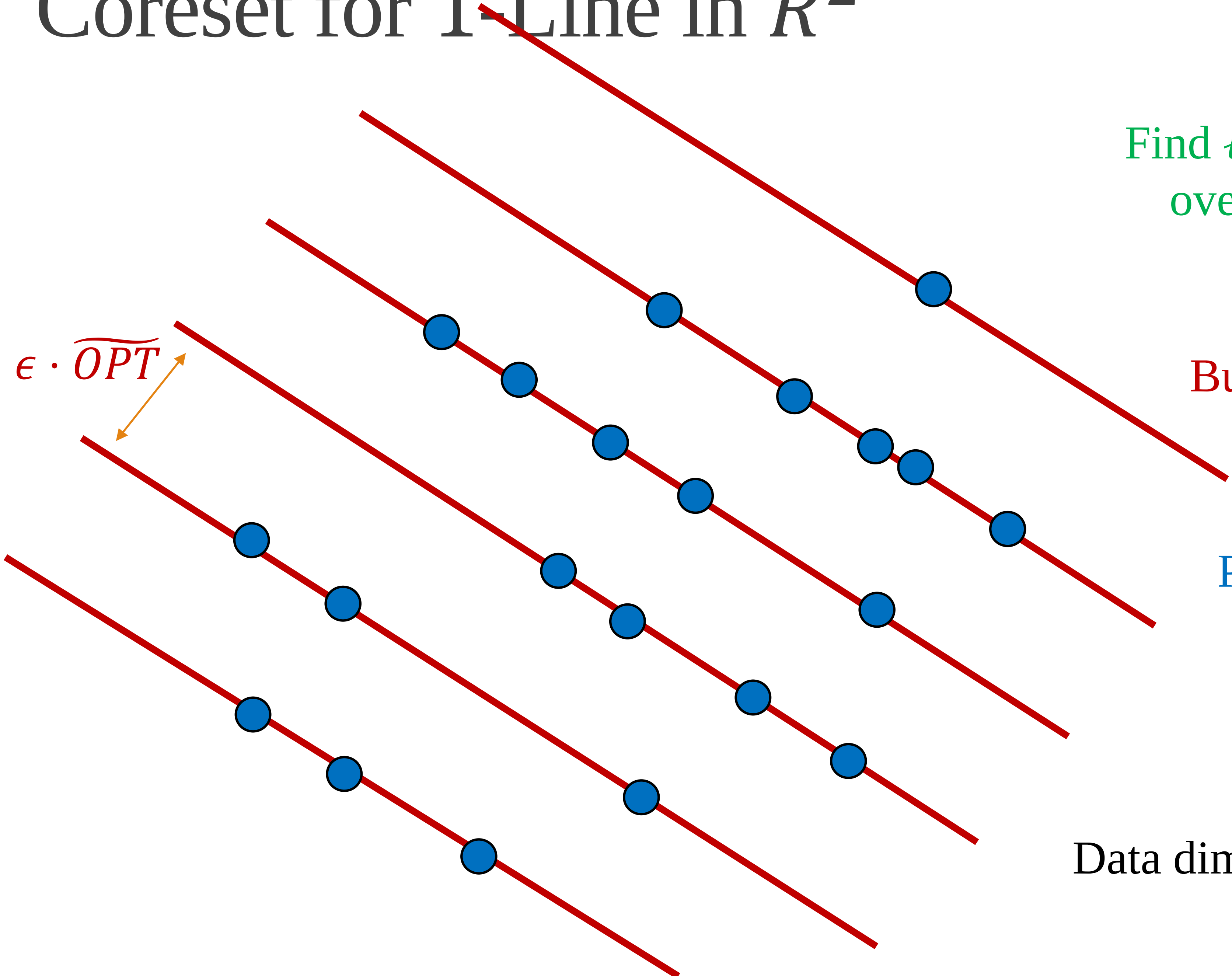


Find ℓ'' by exhaustive search over every pair of points.
$\boldsymbol{O(n^2)}$

Build a grid of lines with $\epsilon \cdot \widetilde{OPT}$ distance

Project each point onto it's closest line

Coreset for 1-Line in R^2



Find ℓ'' by exhaustive search over every pair of points.
$\boldsymbol{O(n^2)}$

Build a grid of lines with $\epsilon \cdot \widetilde{OPT}$ distance

Project each point onto it's closest line

Data dimension is now reduced.

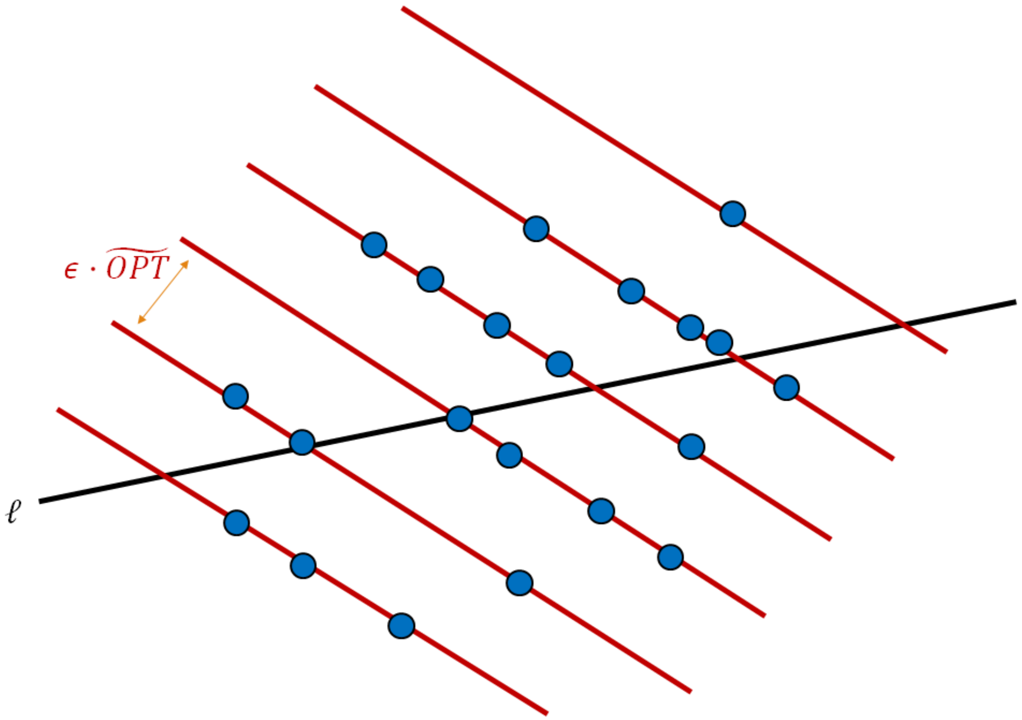
Coreset for 1-Line in R^2

<u>Claim:</u> The projected n points P'
are a "coreset" (not part of the input data)
for any line query:

$$\max_{p\in P} dist(p,\ell) - \max_{p\in P'} dist(p,\ell) \le \epsilon \cdot \widetilde{OPT}$$

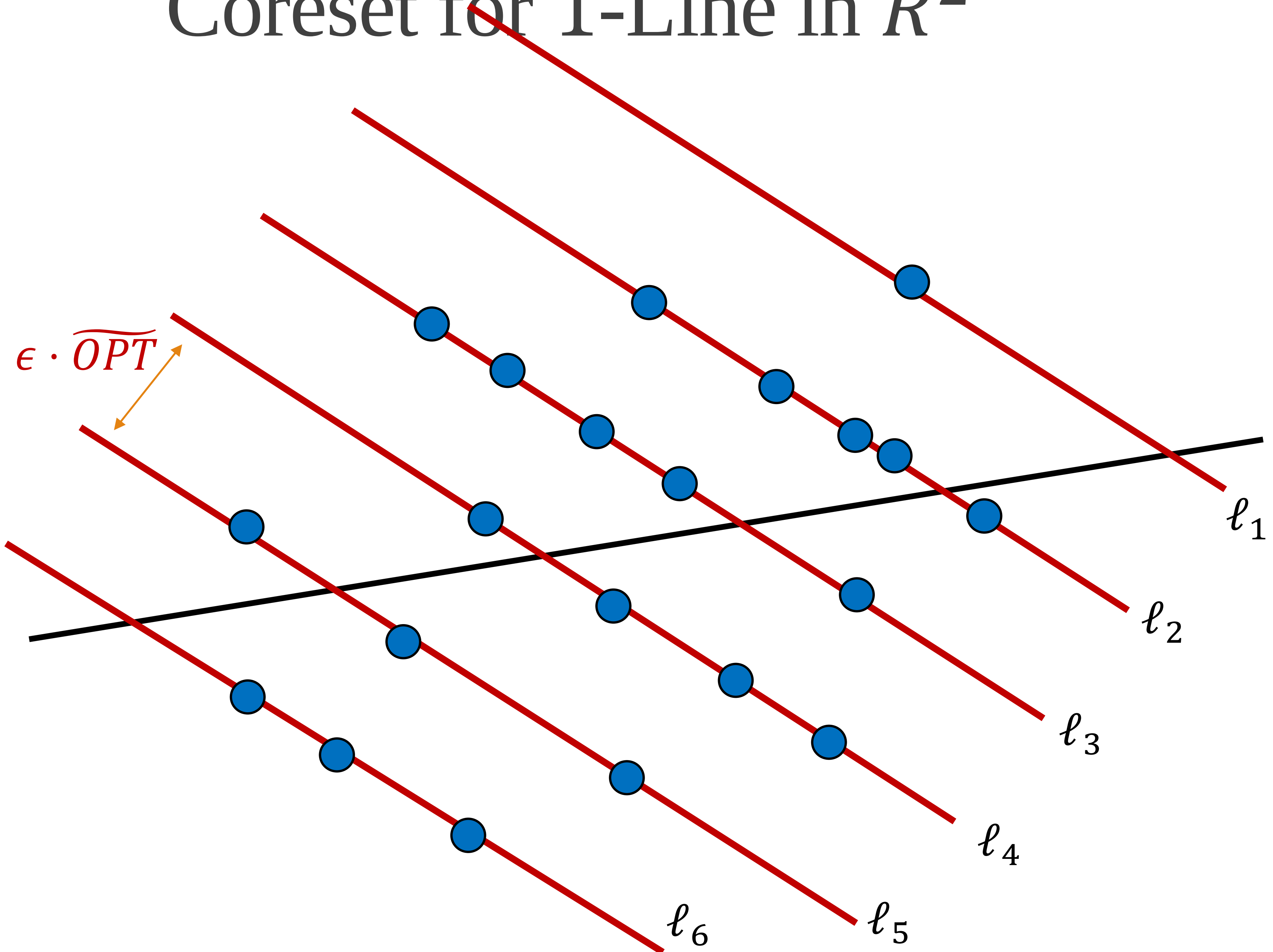
$$\le 4\epsilon \cdot OPT$$

$$\le 4\epsilon \cdot \max_{p\in P} dist(p,\ell)$$

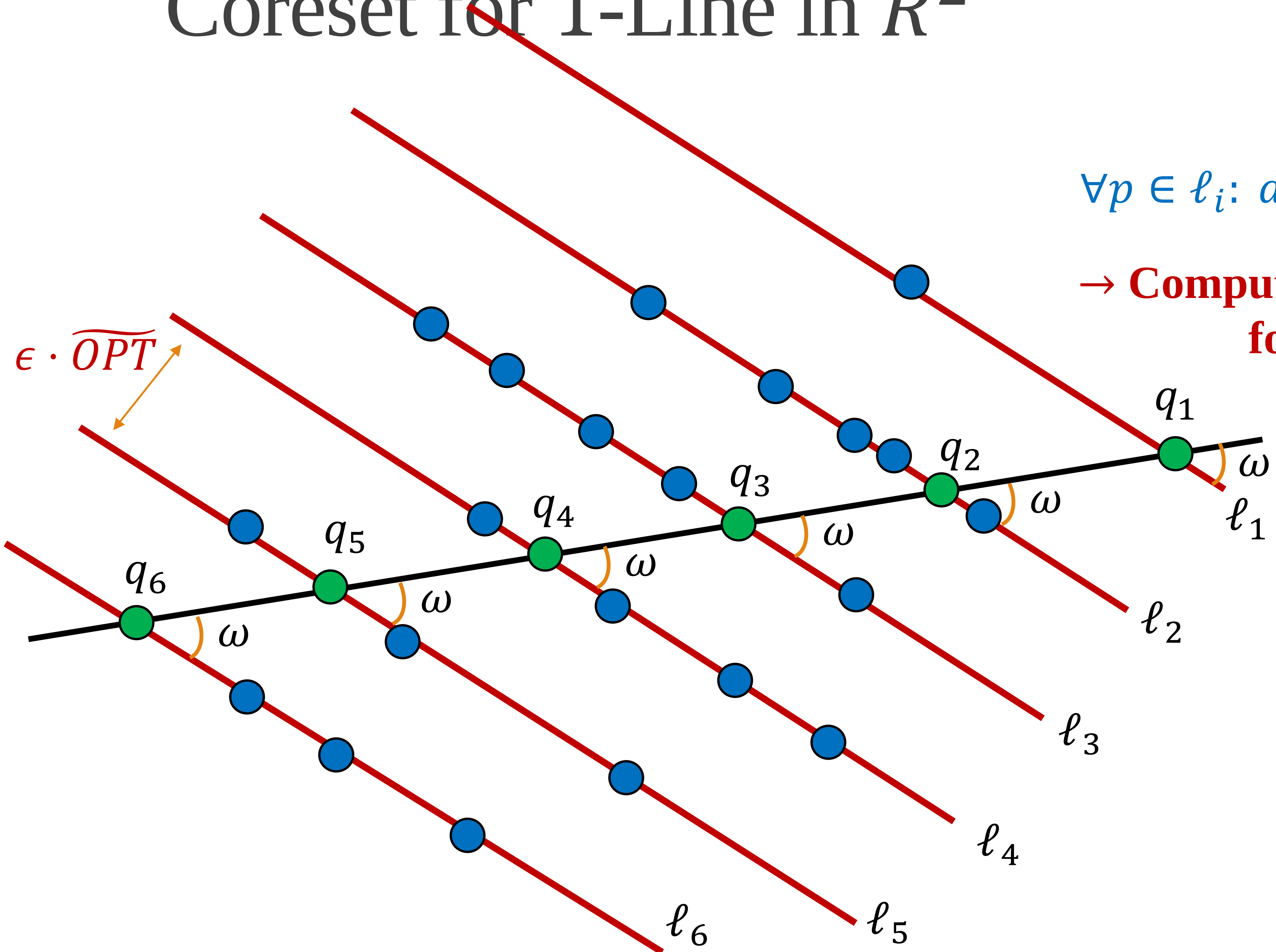


$\rightarrow$ Run with $\epsilon' = \frac{\epsilon}{4}$

Coreset for 1-Line in R^2



Coreset for 1-Line in R^2

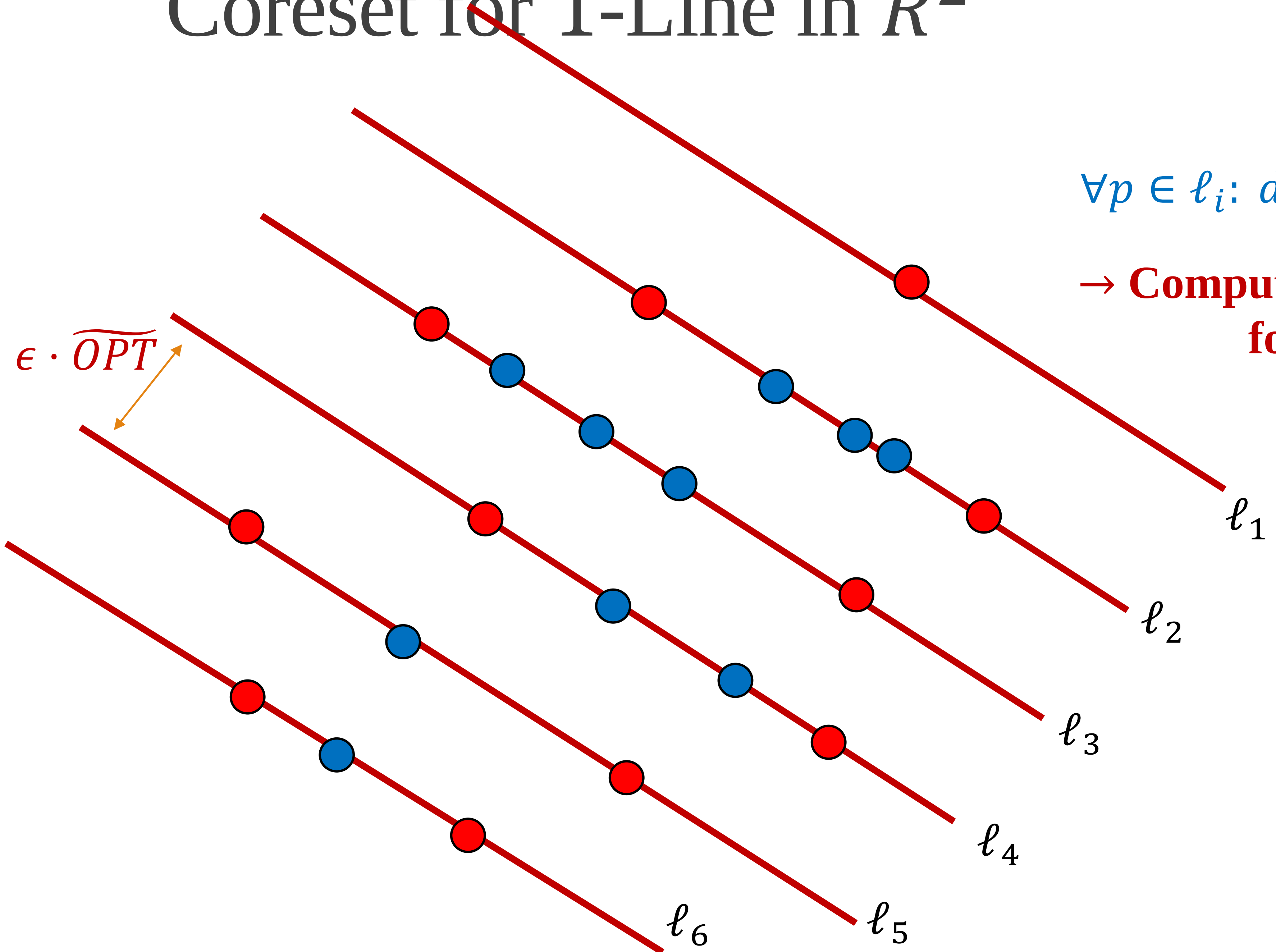


Has no effect since it is the same weight for all points

$\forall p \in \ell_i$: $dist(p, \ell) = \omega \cdot dist(p, q_i)$

→ **Compute a 1-Center coreset C_i for each line ℓ_i!**

Coreset for 1-Line in R^2



Has no effect since it is the same weight for all points

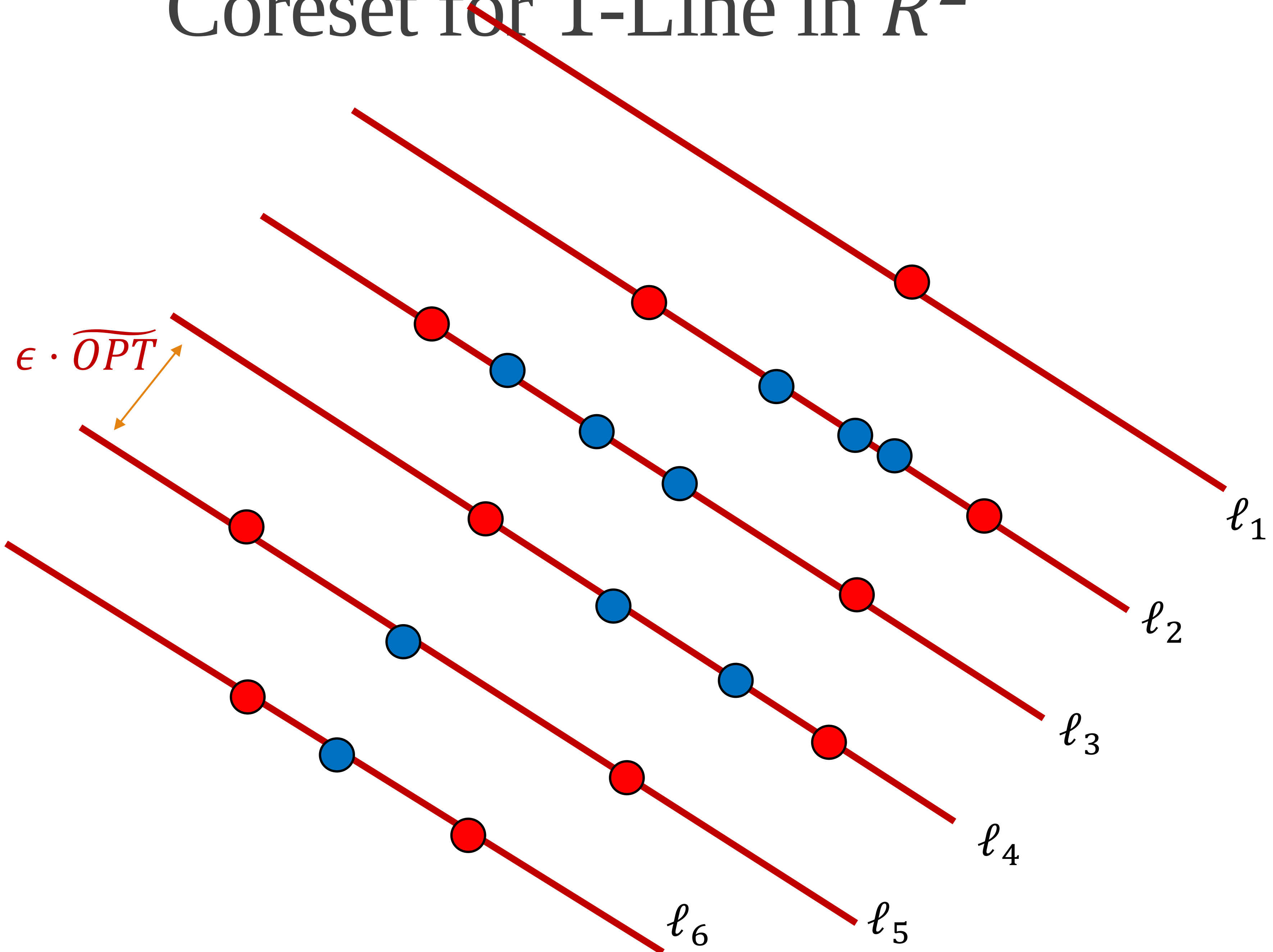
$\forall p \in \ell_i$: $dist(p, \ell) = \omega \cdot dist(p, q_i)$

→ Compute a 1-Center coreset C_i for each line ℓ_i!

$$C = \bigcup C_i$$

since a union of two coresets is a coreset.

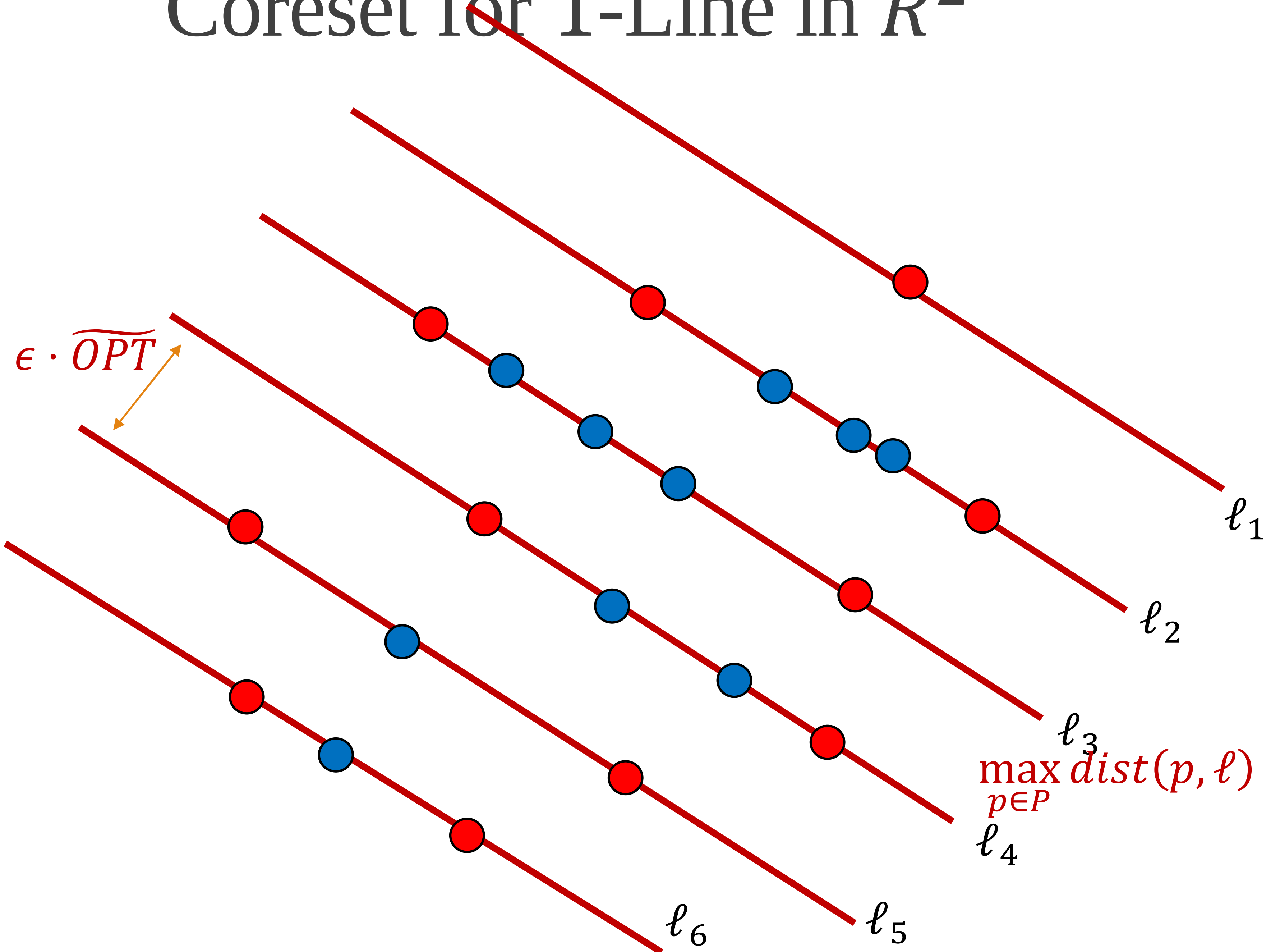
Coreset for 1-Line in R^2



<u>Problem:</u>
The coreset is not part of the input data.

<u>Solution:</u>
Pick the closest points in the input data to the points of C.

Coreset for 1-Line in R^2



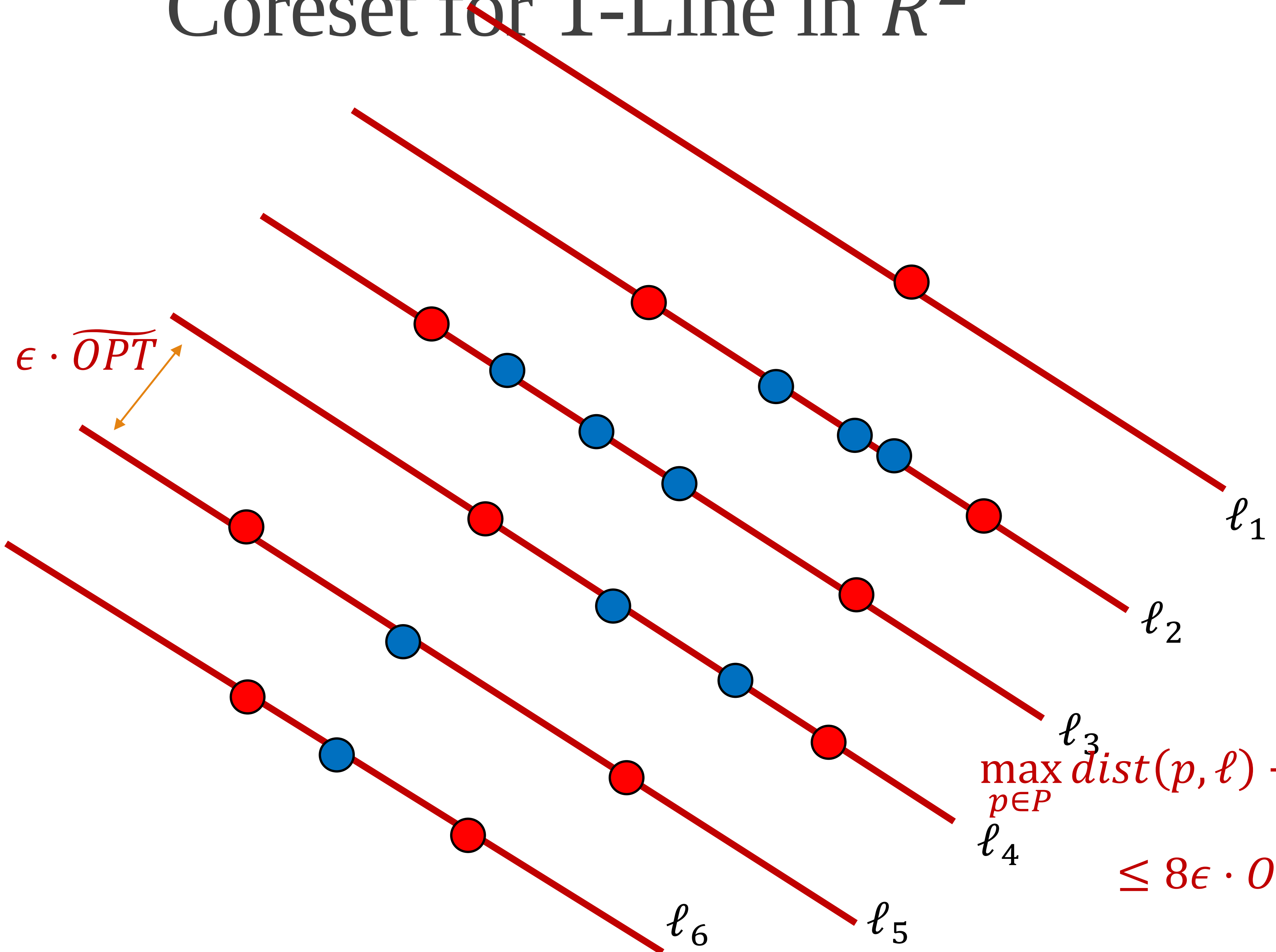
Problem:
The coreset is not part of the input data.

Solution:
Pick the closest points in the input data to the points of C.

→ **This adds another error of $\boldsymbol{\epsilon \cdot \widetilde{OPT}}$**

$$\max_{p \in P} dist(p, \ell) \leq \max_{p \in P'} dist(p, \ell) + 2\epsilon \cdot \widetilde{OPT}$$
$$\leq (1 + 8\epsilon) \cdot \max_{p \in P'} dist(p, \ell)$$

Coreset for 1-Line in R^2



Problem:

The coreset is not part of the input data.

Solution:

Pick the closest points in the input data to the points of C.

→ **This adds another error of $\boldsymbol{\epsilon \cdot \widetilde{OPT}}$**

$$\max_{p \in P} dist(p, \ell) - \max_{p \in P'} dist(p, \ell) \leq 2\epsilon \cdot \widetilde{OPT}$$

$$\leq 8\epsilon \cdot OPT \leq 8\epsilon \cdot \max_{p \in P} dist(p, \ell)$$

Coreset for 1-Line in R^2

<u>Total time:</u>

$O(n^2)$.

<u>Coreset size:</u>

$|C| \leq 2 \cdot \#lines = 2 \cdot \frac{2}{\epsilon} = \frac{4}{\epsilon}$.

Coreset for 1-Line in R^2

Total time:

$O(n^2)$.

Coreset size:

$$|C| \leq 2 \cdot \#lines = 2 \cdot \frac{2}{\epsilon} = \frac{4}{\epsilon}.$$

Improvement:

Run the above algorithm using the streaming tree.

Run on batches of size $2 \cdot |C| = \frac{8}{\epsilon}$.

Total time:

$$\mathrm{O}(\mathrm{n} \cdot TimeForBatch) = \mathrm{O}\left(n \cdot \left(\frac{8}{\epsilon}\right)^2\right).$$

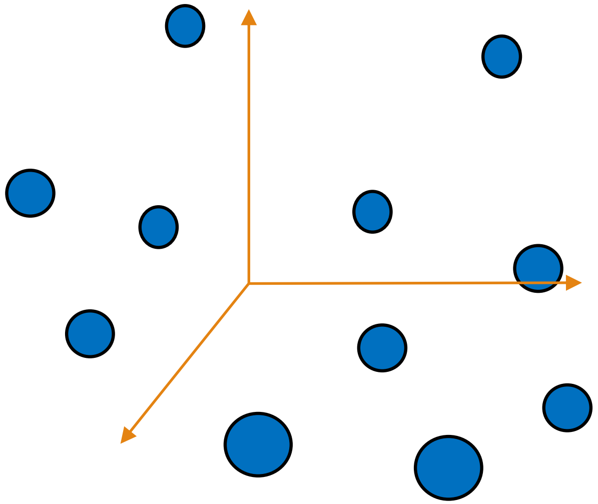
Error for streaming tree:

The error increases to $(1+\epsilon)^{\log n} \sim (1+\epsilon \log n)$

$\rightarrow$ Run with $\epsilon' = \frac{\epsilon}{\log n}$.

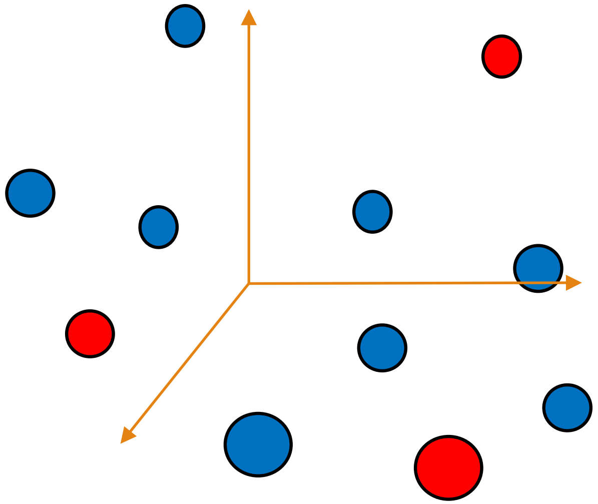
Coreset for 1-Plane in R^3

- <u>Input:</u> $P \subseteq R^3$
- <u>Query space:</u> $Q = \{\pi \mid \pi \text{ is a plane in } R^3\}$
- <u>Cost function:</u> $dist(p,\pi) = \min_{x\in\pi} \|p - x\|_2$
- <u>Output:</u> $\boldsymbol{C} \subseteq P \text{ s.t. } \forall \pi \in Q: \max_{p\in P} dist(p,\pi) - \max_{c\in C} dist(c,\pi) \leq \epsilon \cdot \max_{p\in P} dist(p,\pi)$



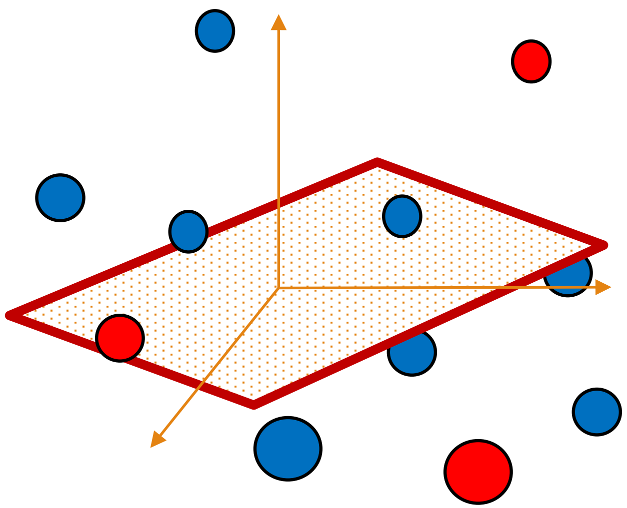
Coreset for 1-Plane in R^3

- Input: $P \subseteq R^3$
- Query space: $Q = \{\pi \mid \pi \text{ is a plane in } R^3\}$
- Cost function: $dist(p,\pi) = \min_{x\in\pi} \|p - x\|_2$
- Output: $C \subseteq P$ s.t. $\forall \pi \in Q: \max_{p\in P} dist(p,\pi) - \max_{c\in C} dist(c,\pi) \le \epsilon \cdot \max_{p\in P} dist(p,\pi)$



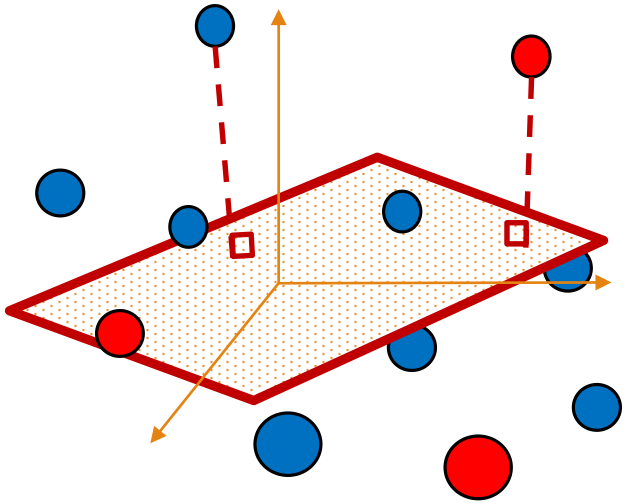
Coreset for 1-Plane in R^3

- Input: $P \subseteq R^3$
- Query space: $Q = \{\pi \mid \pi \text{ is a plane in } R^3\}$
- Cost function: $dist(p,\pi) = \min_{x\in\pi} \|p - x\|_2$
- Output: $C \subseteq P \text{ s.t. } \forall \pi \in Q: \max_{p\in P} dist(p,\pi) - \max_{c\in C} dist(c,\pi) \leq \epsilon \cdot \max_{p\in P} dist(p,\pi)$

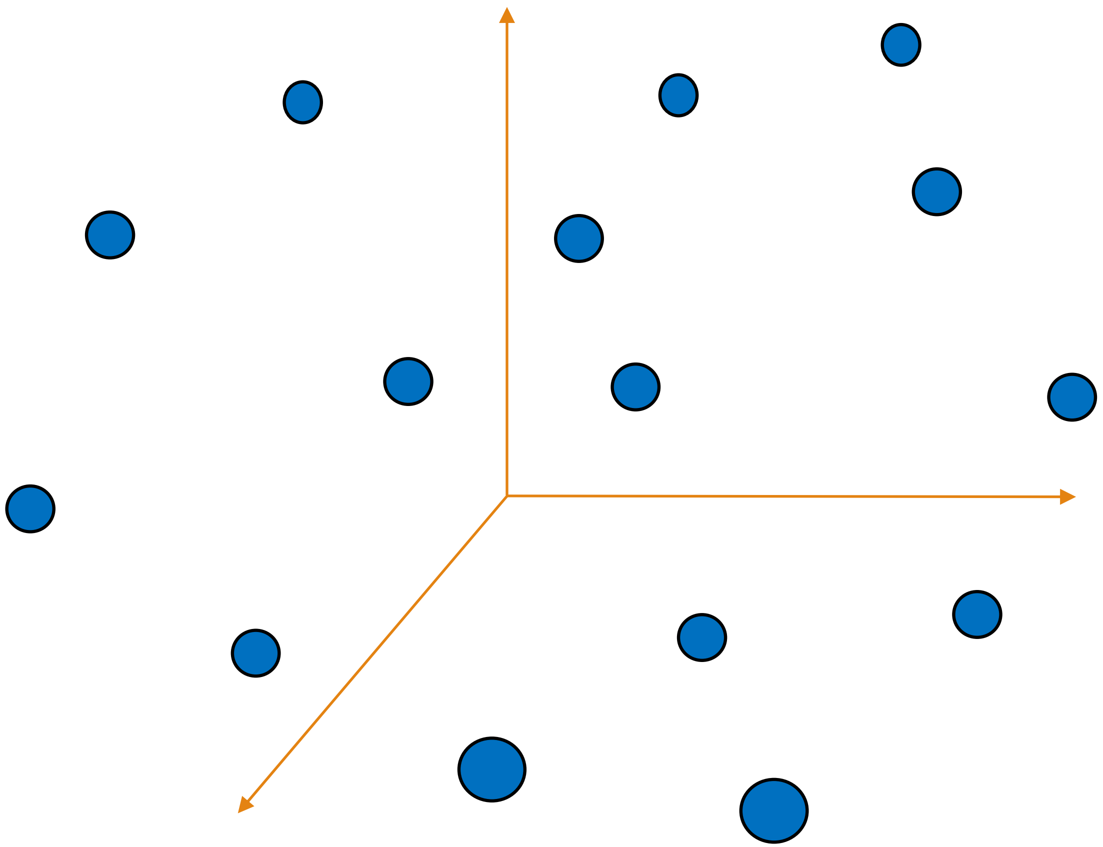


Coreset for 1-Plane in R^3

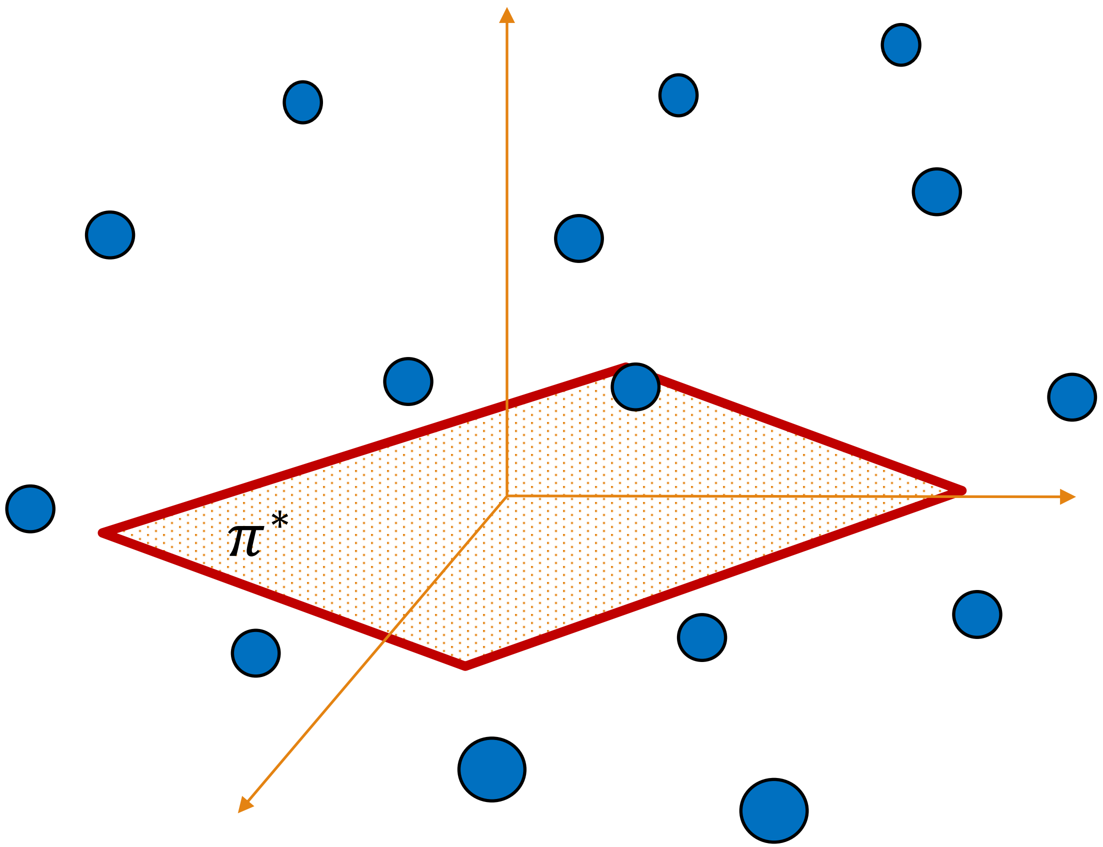
- Input: $P \subseteq R^3$
- Query space: $Q = \{\pi \mid \pi \text{ is a plane in } R^3\}$
- Cost function: $dist(p,\pi) = \min_{x\in\pi} \|p - x\|_2$
- Output: $C \subseteq P \text{ s.t. } \forall \pi \in Q: \max_{p\in P} dist(p,\pi) - \max_{c\in C} dist(c,\pi) \leq \epsilon \cdot \max_{p\in P} dist(p,\pi)$



Coreset for 1-Plane in R^3



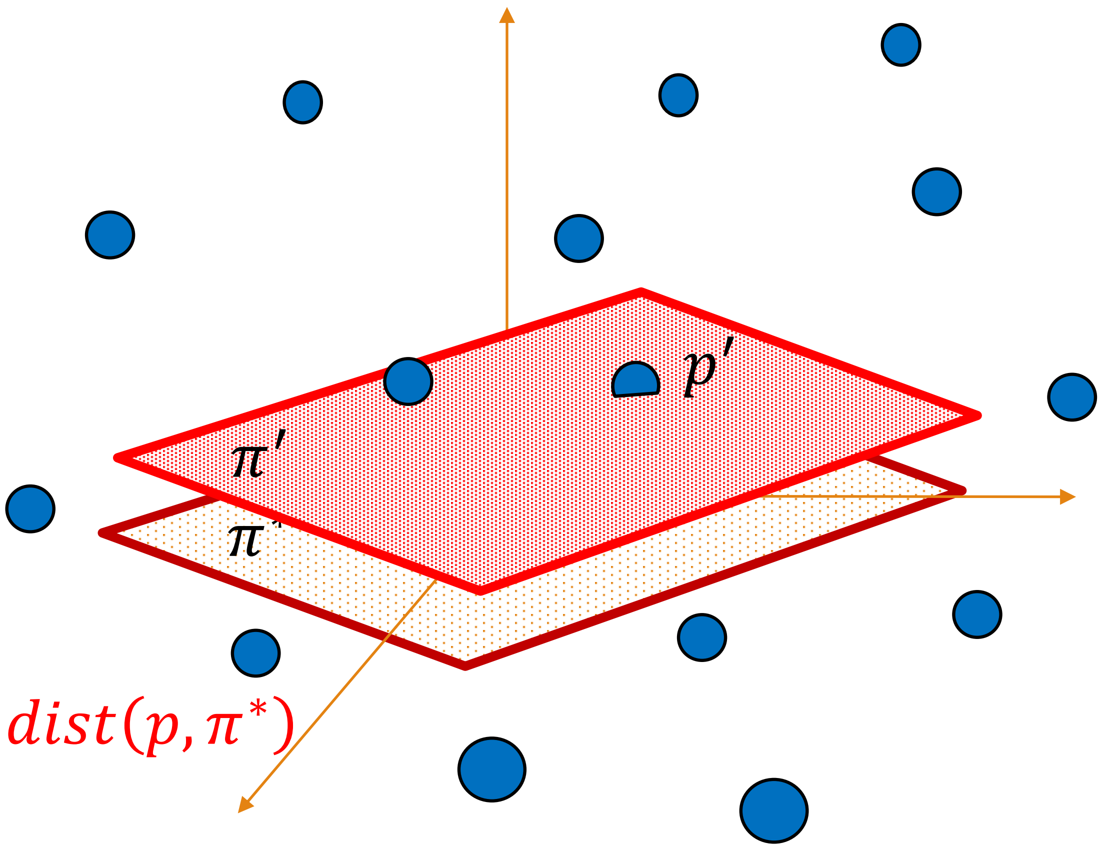
Coreset for 1-Plane in R^3



π^* is the plane that minimizes

$$\max_{p \in P} dist(p, \pi)$$

Coreset for 1-Plane in R^3

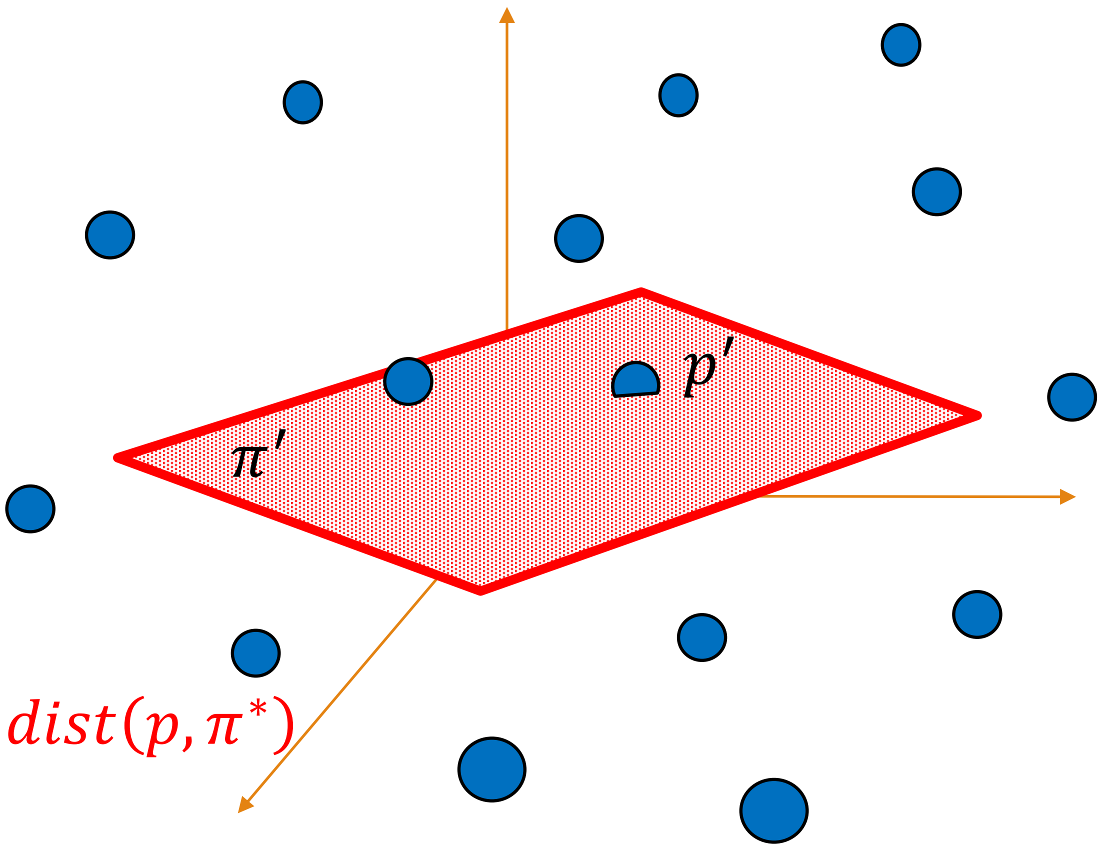


$dist(p, \pi') \leq 2 \cdot dist(p, \pi^*)$

π^* is the plane that minimizes
$\max_{p \in P} dist(p, \pi)$

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

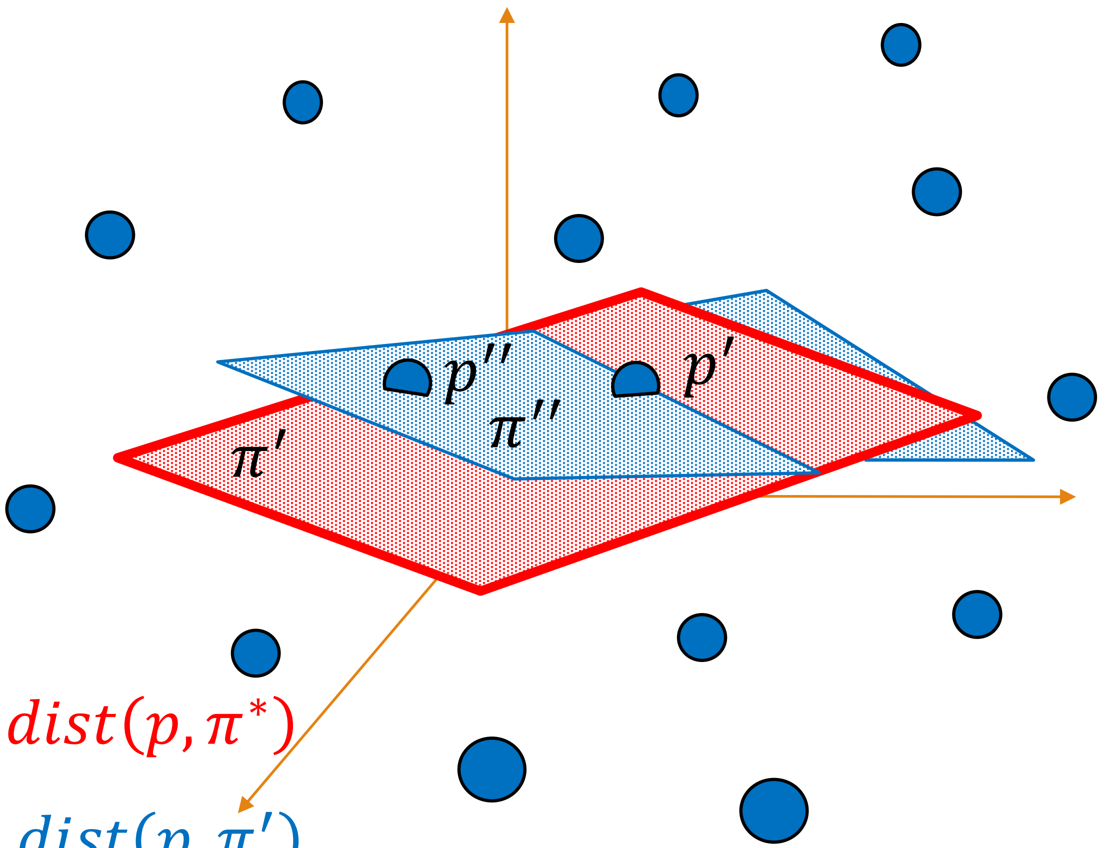
Coreset for 1-Plane in R^3



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

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

Coreset for 1-Plane in R^3



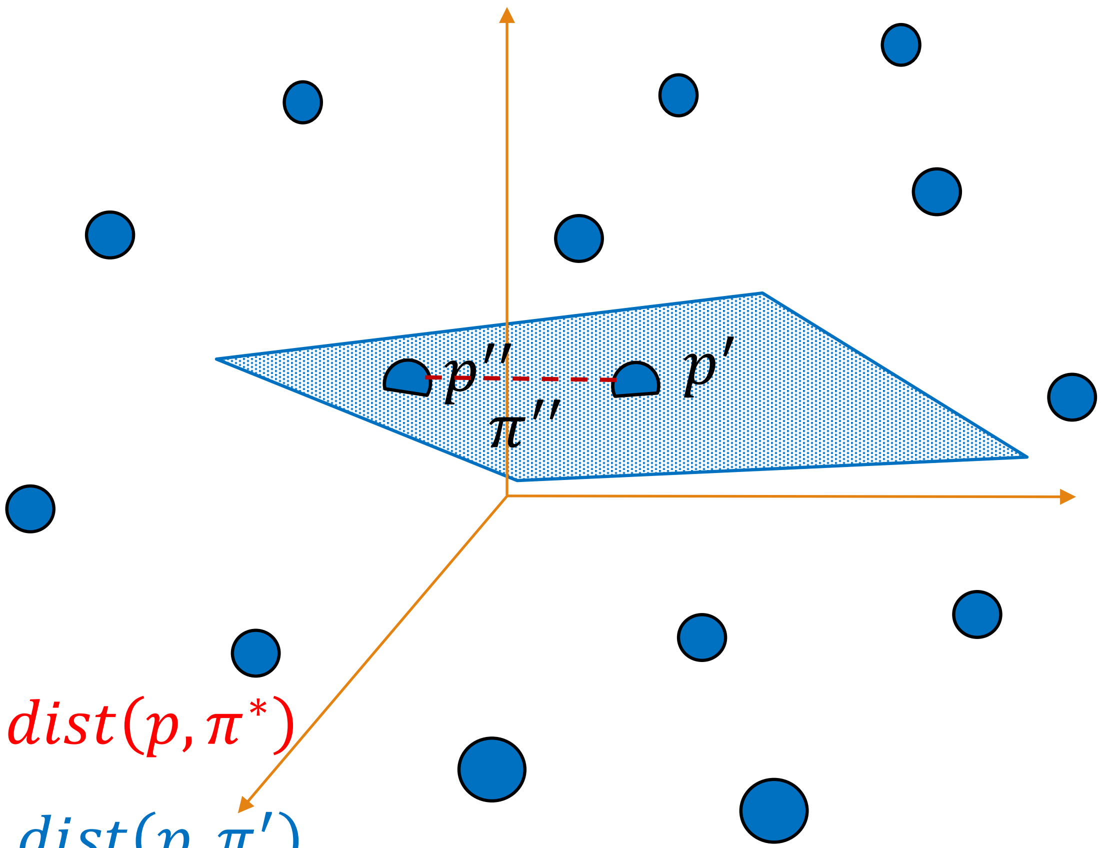
π' is the translation of π^* to $\pi^{*\prime}s$ closest point p'

π'' is the rotation of π' around p' to $\pi' s$ closest point p''

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

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

Coreset for 1-Plane in R^3

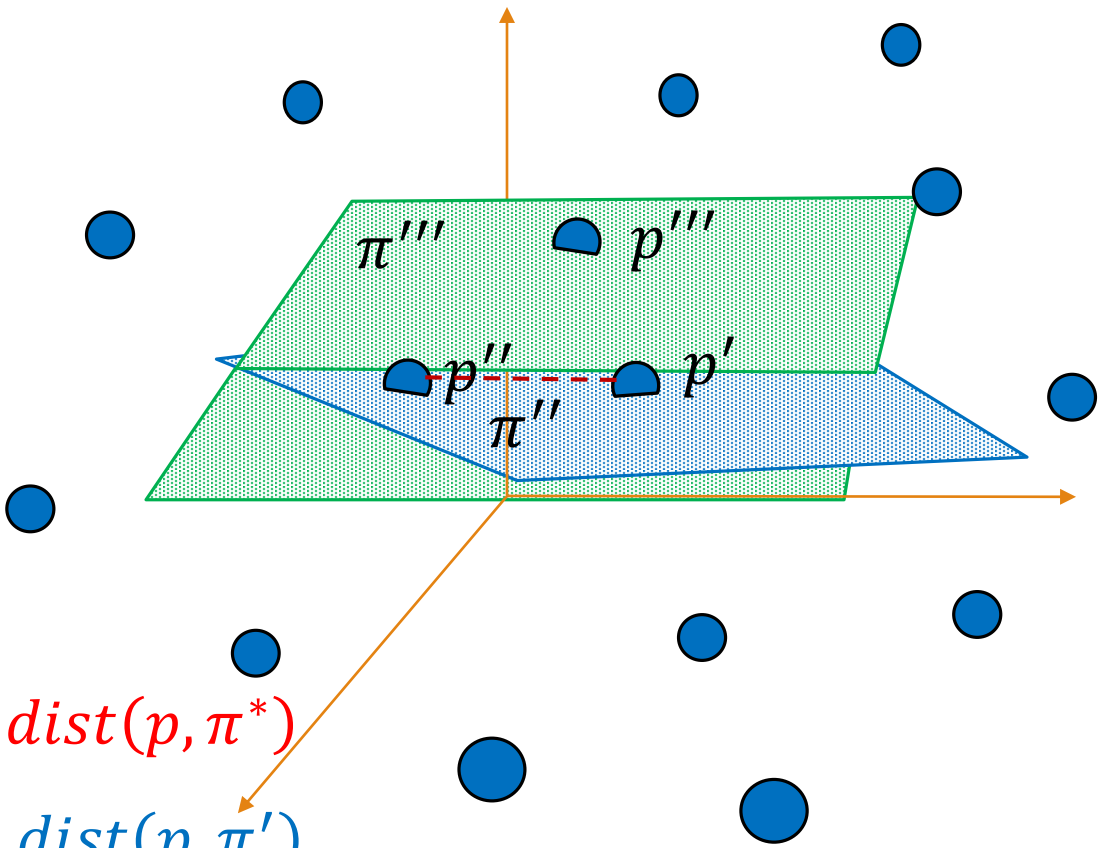


$dist(p, \pi') \leq 2 \cdot dist(p, \pi^*)$

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

π'' is the rotation of π' around p' to $\pi's$ closest point p''

Coreset for 1-Plane in R^3



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

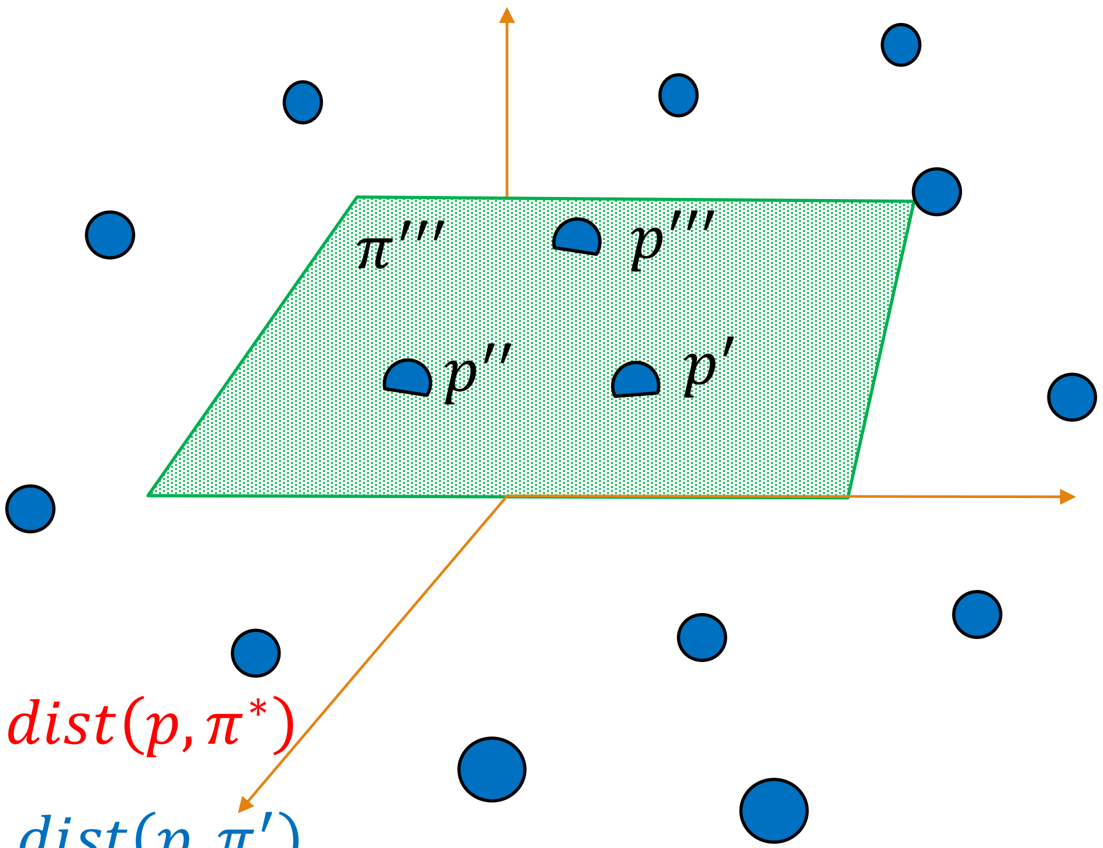
$dist(p,\pi'') \leq 2 \cdot dist(p,\pi')$

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

π'' is the rotation of π' around p' to π's closest point p''

π''' is the rotation of π'' around $p' - p''$ to π''s closest point p'''

Coreset for 1-Plane in R^3



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

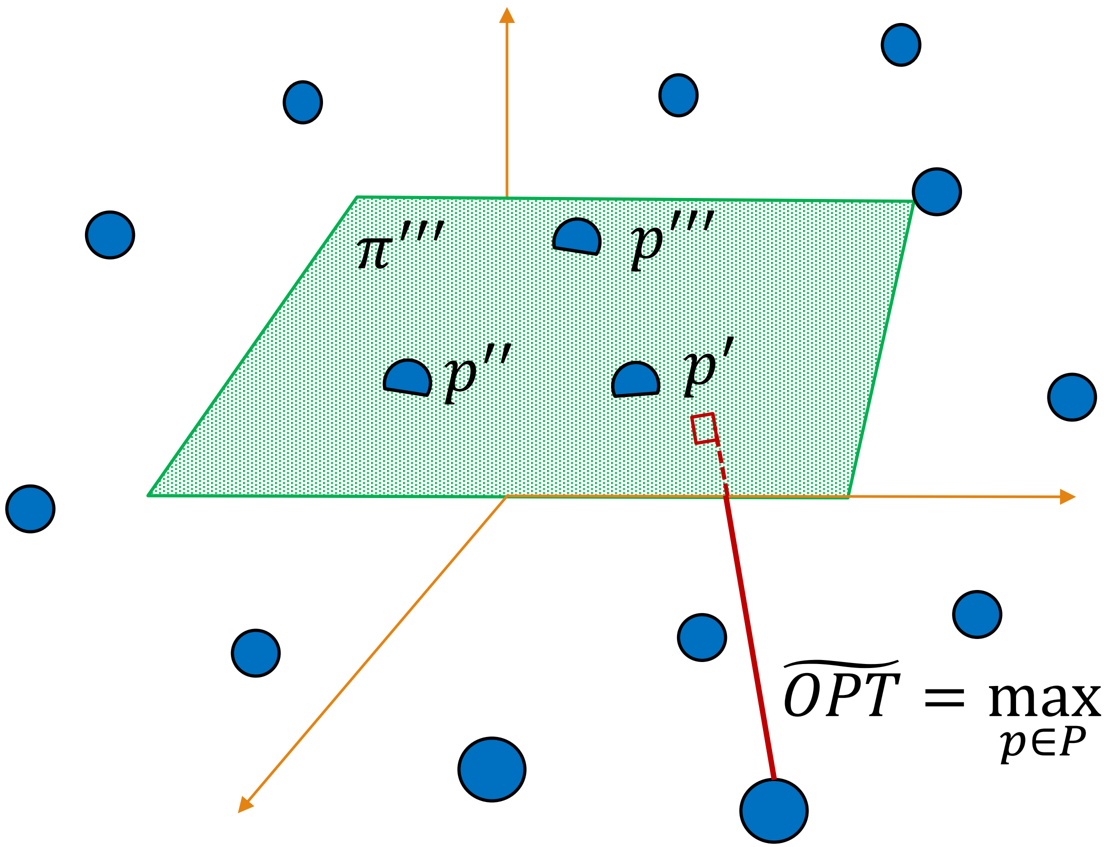
$dist(p,\pi'') \leq 2 \cdot dist(p,\pi')$

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

$dist(p,\pi''') \leq 8 \cdot dist(p,\pi^*)$

π''' is the rotation of π'' around $p' - p''$ to $\pi''s$ closest point p'''

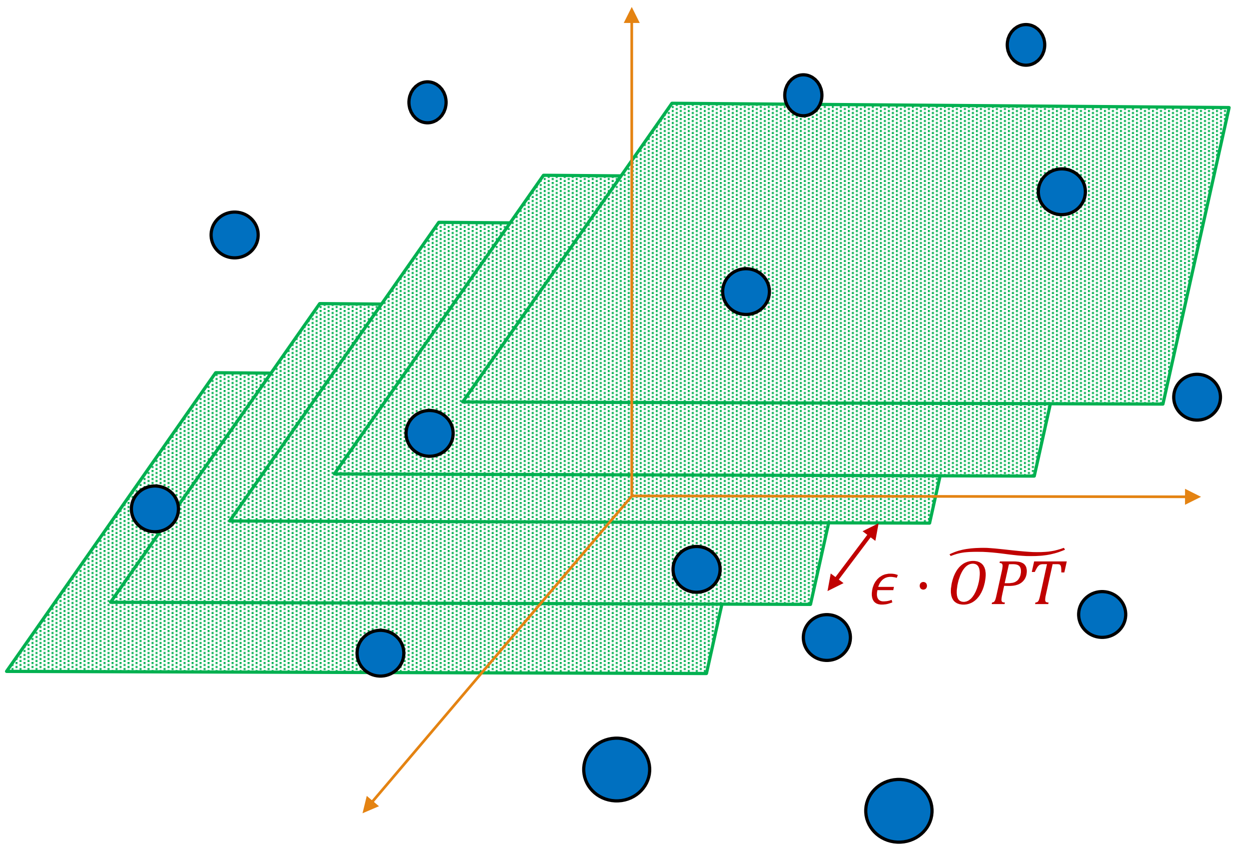
Coreset for 1-Plane in R^3



Find π''' by exhaustive search over every triplet of points.
$\boldsymbol{O(n^3)}$

$$\widetilde{OPT} = \max_{p \in P} dist(p, \pi''')$$

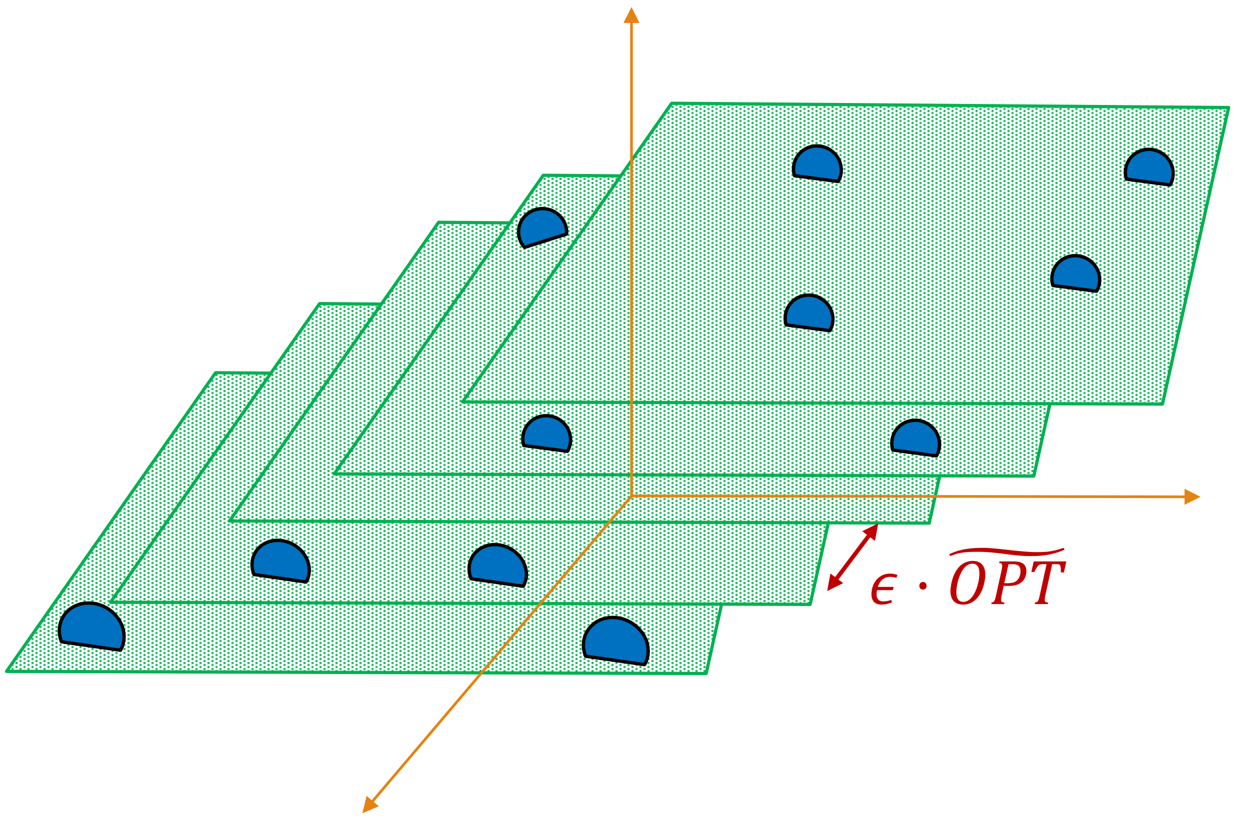
Coreset for 1-Plane in R^3



Find π''' by exhaustive search over every triplet of points.
$\boldsymbol{O}(\boldsymbol{n}^3)$

Build a grid of planes with $\epsilon \cdot \widetilde{OPT}$ distance

Coreset for 1-Plane in R^3



Find π''' by exhaustive search over every triplet of points.
$\boldsymbol{O(n^3)}$

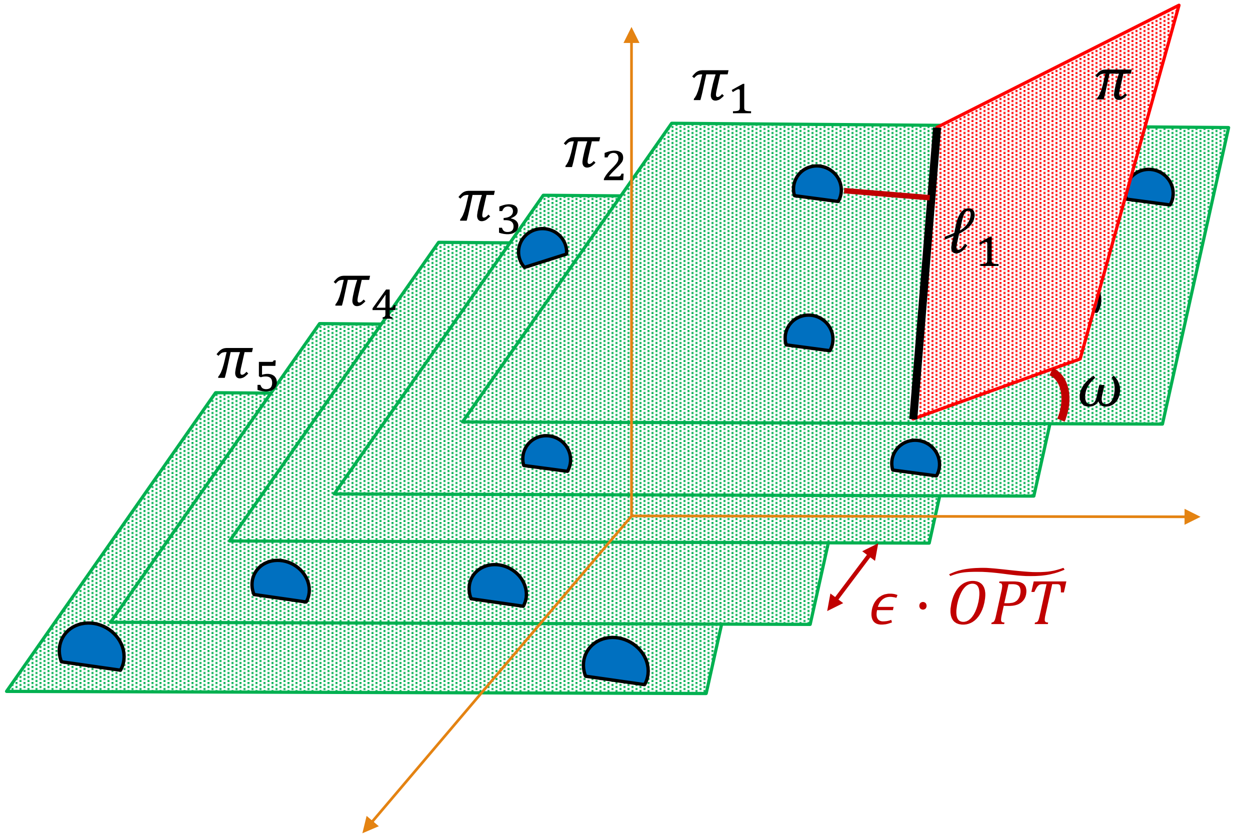
Build a grid of planes with $\epsilon \cdot \widetilde{OPT}$ distance

Project each point onto it's closest plane

Coreset for 1-Plane in R^3

$$\forall p \in \pi_i:\ dist(p, \pi) = \omega \cdot dist(p, \ell_i)$$

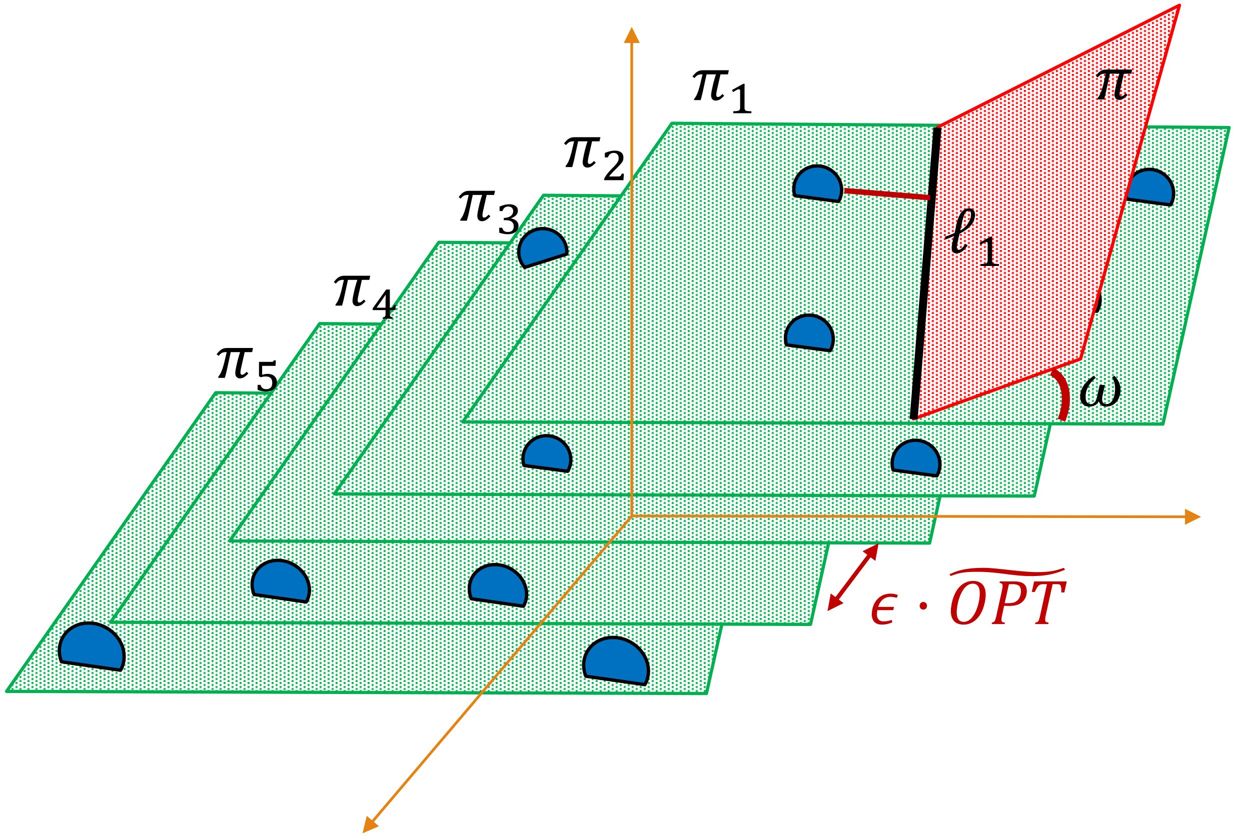
$$\ell_i = \pi_i \cap \pi$$



Coreset for 1-Plane in R^3

$\forall p \in \pi_i: \, dist(p, \pi) = \omega \cdot dist(p, \ell_i)$

$\ell_i = \pi_i \cap \pi$



→ Compute a **1-Line** coreset $\boldsymbol{C_i}$ for each **plane $\boldsymbol{\pi_i}$**!

$$\boldsymbol{C} = \bigcup \boldsymbol{C_i}$$

since a union of two coresets is a coreset.

Coreset for $Hyperplane$ in R^d

$\underline{\boldsymbol{HyperplaneCoreset}(\boldsymbol{P}, \boldsymbol{d}):}$

Coreset for $Hyperplane$ in R^d

$\underline{\boldsymbol{HyperplaneCoreset(P, d)}}$**:**

- $h' \leftarrow$ an α-approximation for the optimal hyperplane of P.

Coreset for $Hyperplane$ in R^d

<u>$\boldsymbol{HyperplaneCoreset(P, d)}$:</u>

- $h' \leftarrow$ an α-approximation for the optimal hyperplane of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.

Coreset for $Hyperplane$ in R^d

<u>$\boldsymbol{HyperplaneCoreset(P, d)}$**:**</u>

- $h' \leftarrow$ an α-approximation for the optimal hyperplane of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.
- $h^{\perp} \leftarrow$the vector that is orthogonal to h'.

Coreset for $Hyperplane$ in R^d

<u>$\boldsymbol{HyperplaneCoreset(P, d)}$:</u>

- $h' \leftarrow$ an α-approximation for the optimal hyperplane of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.
- $h^{\perp} \leftarrow$the vector that is orthogonal to h'.
- Construct a grid on $h^{\perp}$ whose cell length is $\epsilon \cdot \widetilde{OPT}$.

Coreset for $Hyperplane$ in R^d

<u>$\boldsymbol{HyperplaneCoreset}(\boldsymbol{P}, \boldsymbol{d})$:</u>

- $h' \leftarrow$ an α-approximation for the optimal hyperplane of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.
- $h^{\perp} \leftarrow$the vector that is orthogonal to h'.
- Construct a grid on $h^{\perp}$ whose cell length is $\epsilon \cdot \widetilde{OPT}$.
- Through each grid point construct a hyperplane parallel to h'. $\left(\#Hyperplanes = \frac{2}{\epsilon}\right)$

Coreset for $Hyperplane$ in R^d

$\underline{HyperplaneCoreset(P, d)}$:

- $h' \leftarrow$ an α-approximation for the optimal hyperplane of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.
- $h^{\perp} \leftarrow$the vector that is orthogonal to h'.
- Construct a grid on $h^{\perp}$ whose cell length is $\epsilon \cdot \widetilde{OPT}$.
- Through each grid point construct a hyperplane parallel to h'. $\left(\#Hyperplanes = \frac{2}{\epsilon}\right)$
- Compute the projection p' of each point $p \in P$ onto it's closest hyperplane h_p.

Coreset for $Hyperplane$ in R^d

<u>$\boldsymbol{HyperplaneCoreset(P, d)}$:</u>

- $h' \leftarrow$ an α-approximation for the optimal hyperplane of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.
- $h^{\perp} \leftarrow$the vector that is orthogonal to h'.
- Construct a grid on $h^{\perp}$ whose cell length is $\epsilon \cdot \widetilde{OPT}$.
- Through each grid point construct a hyperplane parallel to h'. $\left(\#Hyperplanes = \frac{2}{\epsilon}\right)$
- Compute the projection p' of each point $p \in P$ onto it's closest hyperplane h_p.
- $H_p \leftarrow$an $R^{d \times d-1}$ matrix whose columns span h_p and $H_p^T H_p = I$.

Coreset for $Hyperplane$ in R^d

$\underline{\boldsymbol{HyperplaneCoreset(P, d):}}$

- $h' \leftarrow$ an α-approximation for the optimal hyperplane of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.
- $h^{\perp} \leftarrow$the vector that is orthogonal to h'.
- Construct a grid on $h^{\perp}$ whose cell length is $\epsilon \cdot \widetilde{OPT}$.
- Through each grid point construct a hyperplane parallel to h'. $\left(\#Hyperplanes = \frac{2}{\epsilon}\right)$
- Compute the projection p' of each point $p \in P$ onto it's closest hyperplane h_p.
- $H_p \leftarrow$an $R^{d \times d-1}$ matrix whose columns span h_p and $H_p^T H_p = I$.
- $P' = \{H_p p' | p \in P\}$.

Coreset for $Hyperplane$ in R^d

$\underline{\boldsymbol{HyperplaneCoreset(P, d):}}$

- $h' \leftarrow$ an α-approximation for the optimal hyperplane of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.
- $h^{\perp} \leftarrow$the vector that is orthogonal to h'.
- Construct a grid on $h^{\perp}$ whose cell length is $\epsilon \cdot \widetilde{OPT}$.
- Through each grid point construct a hyperplane parallel to h'. $\left(\#Hyperplanes = \frac{2}{\epsilon}\right)$
- Compute the projection p' of each point $p \in P$ onto it's closest hyperplane h_p.
- $H_p \leftarrow$an $R^{d \times d-1}$ matrix whose columns span h_p and $H_p^T H_p = I$.
- $P' = \{H_p p' | p \in P\}$
- Call $\boldsymbol{HyperplaneCoreset(P', d-1)}$.

Coreset for *Hyperplane* in R^d

<u>Total time:</u>

$O(n^d)$

<u>Coreset size:</u>

$$|C| \leq \left(\frac{1}{\epsilon}\right)^{O(d)}$$

Coreset for $Hyperplane$ in R^d

<u>Total time:</u>

$O(n^d)$

<u>Coreset size:</u>

$|C| \leq \left(\frac{1}{\epsilon}\right)^{O(d)}$

<u>Improvement:</u>

Run the above algorithm using the streaming tree.

Run on batches of size $2 \cdot |C| \leq 2\left(\frac{1}{\epsilon}\right)^{O(d)}$.

<u>Total time:</u>

$$\mathrm{O}(\mathrm{n} \cdot TimeForBatch) = \mathrm{O}\left(n \cdot \left(\frac{1}{\epsilon}\right)^{O(d^2)}\right).$$

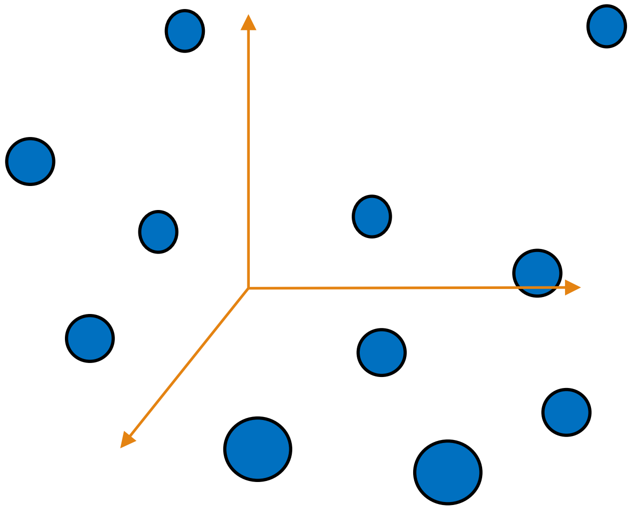
<u>Error for streaming tree:</u>

The error increases to $(1+\epsilon)^{\log n} \sim (1+\epsilon \log n)$

→ Run with $\epsilon' = \frac{\epsilon}{\log n}$.

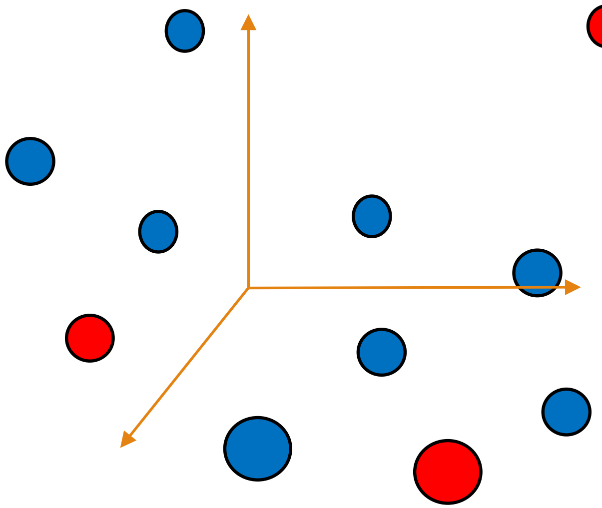
Coreset for 1-Line parallel to z-axis in R^3

- Input: $P \subseteq R^3$
- Query space: $Q = \{\ell \mid \ell \textit{ is a line in } R^3 \textit{ parallel to the z-axis}\}$
- Cost function: $dist(p, \ell) = \min_{x \in \ell} \|p - x\|_2$
- Output: $\boldsymbol{C} \subseteq P \ s.t.\ \forall \ell \in Q: \max_{p \in P} dist(p, \ell) - \max_{c \in C} dist(c, \ell) \leq \epsilon \cdot \max_{p \in P} dist(p, \ell)$



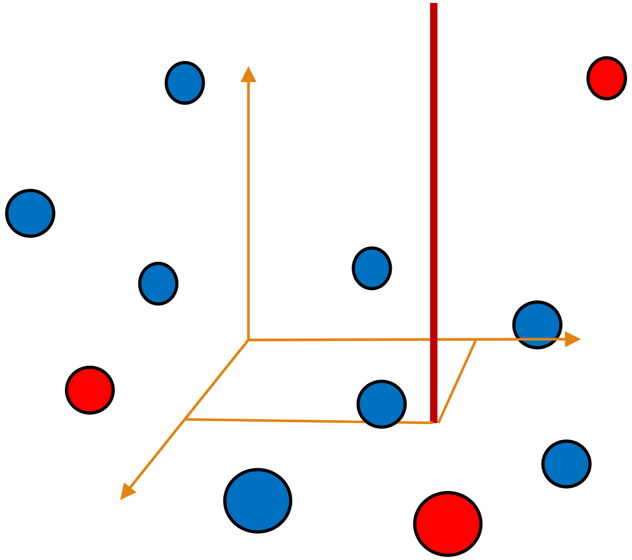
Coreset for 1-Line parallel to z-axis in R^3

- <u>Input:</u> $P \subseteq R^3$
- <u>Query space:</u> $Q = \{\ell \mid \ell \textit{ is a line in } R^3 \textit{ parallel to the z-axis}\}$
- <u>Cost function:</u> $dist(p, \ell) = \min_{x \in \ell} \|p - x\|_2$
- <u>Output:</u> $\boldsymbol{C} \subseteq P \; s.t. \; \forall \ell \in Q: \max_{p \in P} dist(p, \ell) - \max_{c \in C} dist(c, \ell) \leq \epsilon \cdot \max_{p \in P} dist(p, \ell)$



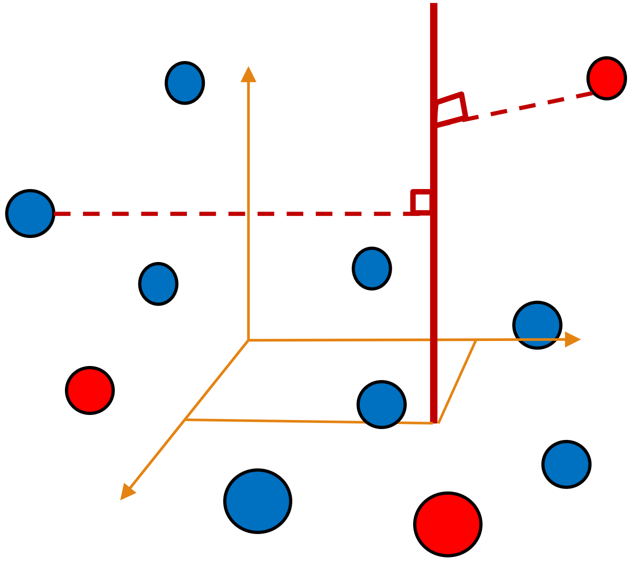
Coreset for 1-Line parallel to z-axis in R^3

- Input: $P \subseteq R^3$
- Query space: $Q = \{\ell \mid \ell \textit{ is a line in } R^3 \textit{ parallel to the z-axis}\}$
- Cost function: $dist(p,\ell) = \min_{x \in \ell} \|p - x\|_2$
- Output: $\boldsymbol{C} \subseteq P \; s.t. \; \forall \ell \in Q: \max_{p \in P} dist(p,\ell) - \max_{c \in C} dist(c,\ell) \leq \epsilon \cdot \max_{p \in P} dist(p,\ell)$

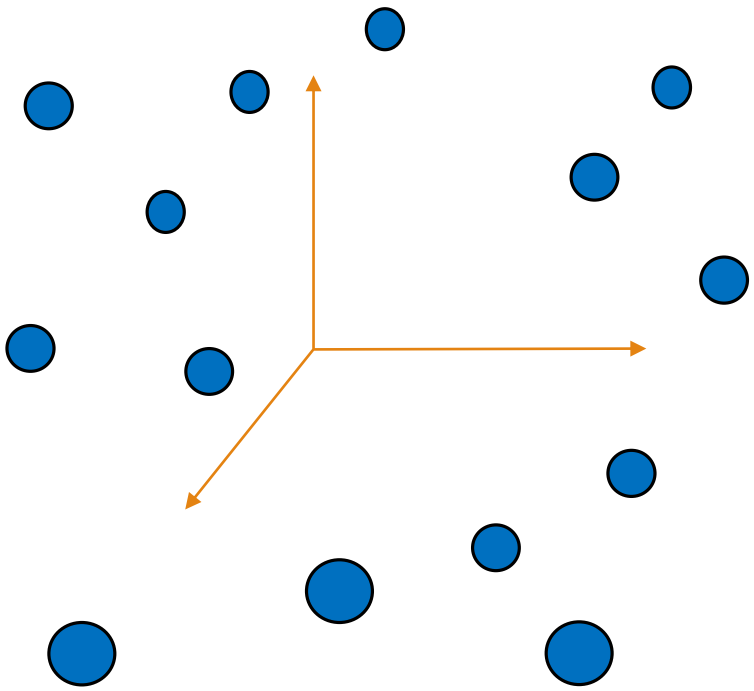


Coreset for 1-Line parallel to z-axis in R^3

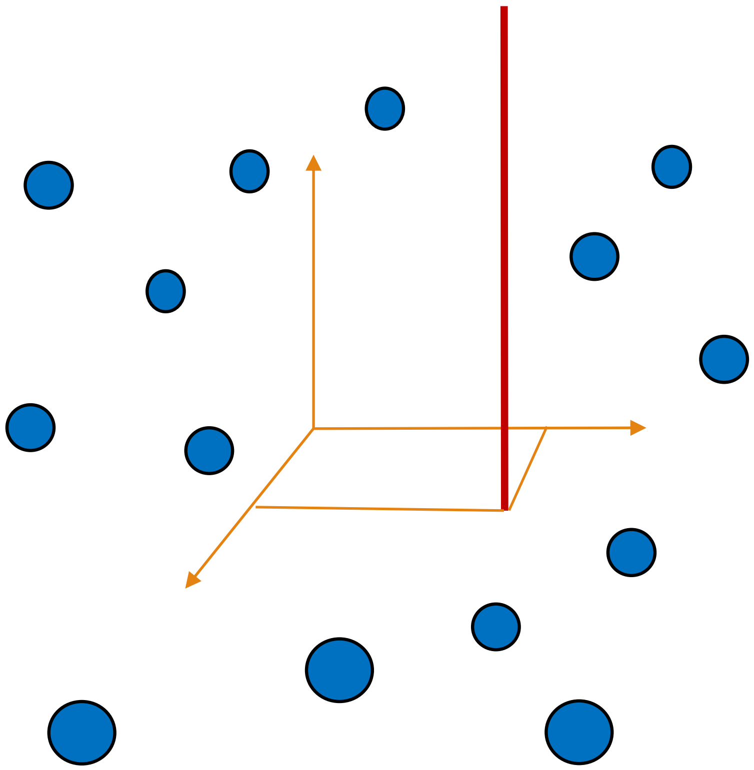
- Input: $P \subseteq R^3$
- Query space: $Q = \{\ell \mid \ell \textit{ is a line in } R^3 \textit{ parallel to the z-axis}\}$
- Cost function: $dist(p, \ell) = \min_{x \in \ell} \|p - x\|_2$
- Output: $\boldsymbol{C} \subseteq P \text{ s.t. } \forall \ell \in Q\colon \max_{p \in P} dist(p, \ell) - \max_{c \in C} dist(c, \ell) \leq \epsilon \cdot \max_{p \in P} dist(p, \ell)$



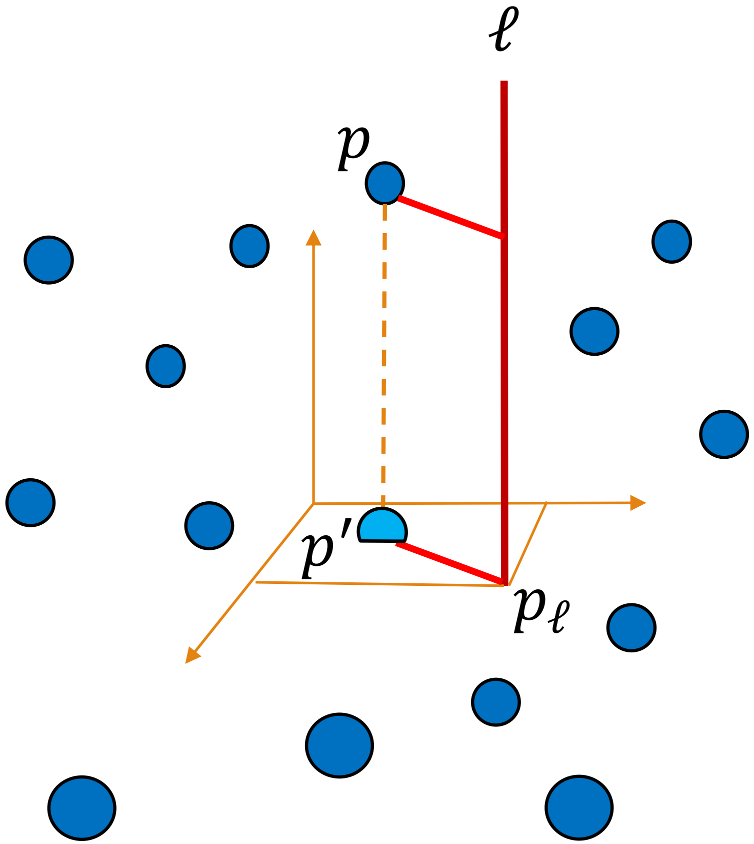
Coreset for 1-Line parallel to z-axis in R^3



Coreset for 1-Line parallel to z-axis in R^3



Coreset for 1-Line parallel to z-axis in R^3



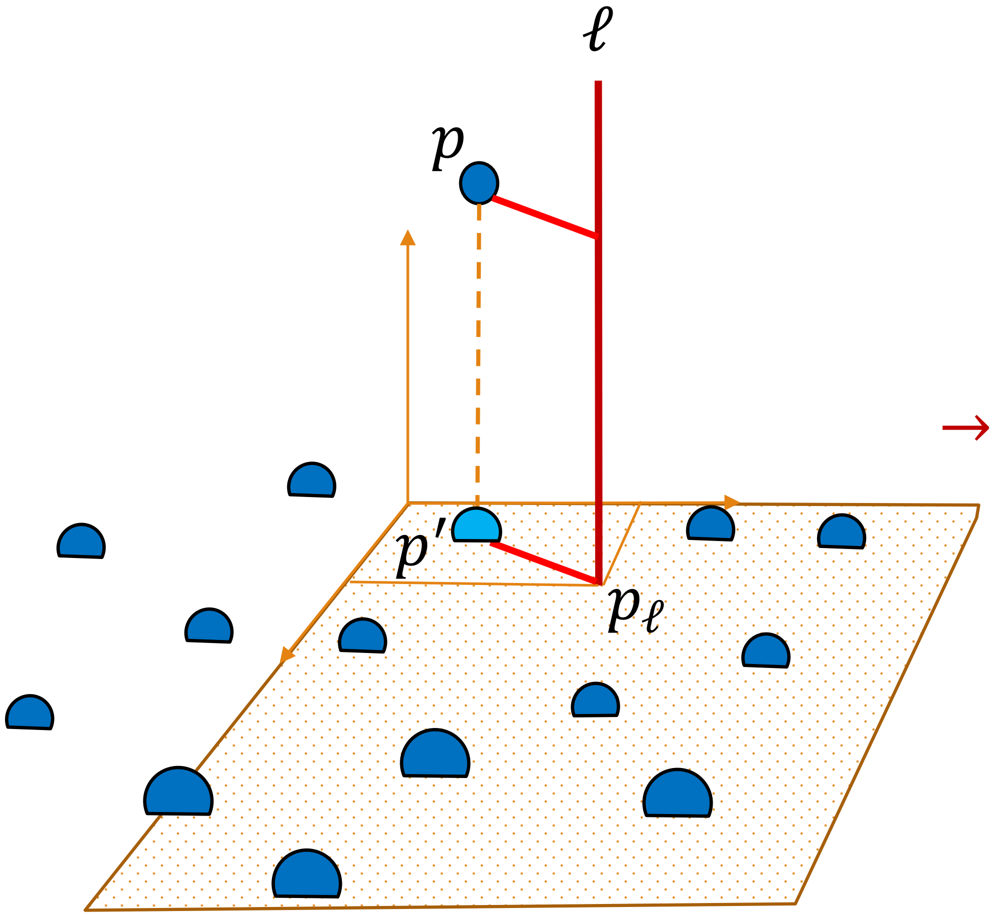
$proj(p, xy)$

Notice that:

$dist(p, \ell) = dist(p', p_\ell)$

$\ell \cap xy$-plane

Coreset for 1-Line parallel to z-axis in R^3



Notice that:

$proj(p, xy)$

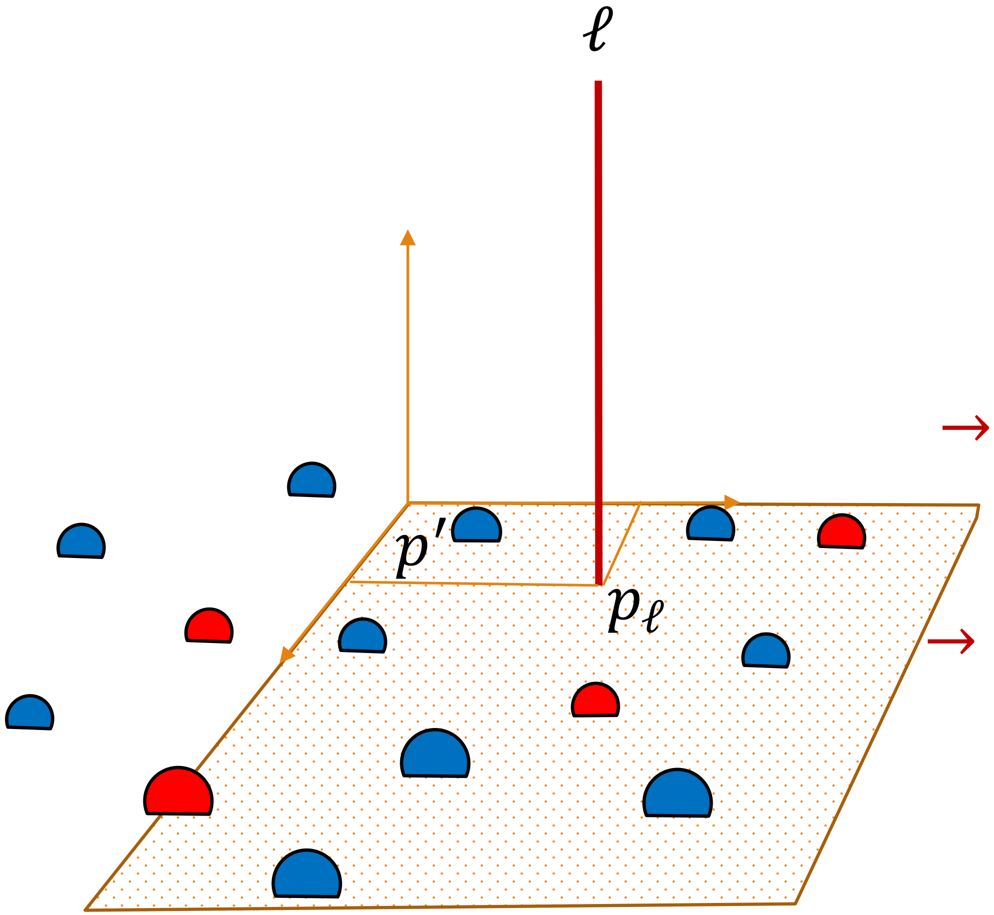
$dist(p, \ell) = dist(p', p_\ell)$

$\ell \cap xy$-plane

→ Project points onto xy-plane

$$\boldsymbol{P' = \{\, p' \mid p \in P \,\}}$$

Coreset for 1-Line parallel to z-axis in R^3



$proj(p, xy)$

Notice that:

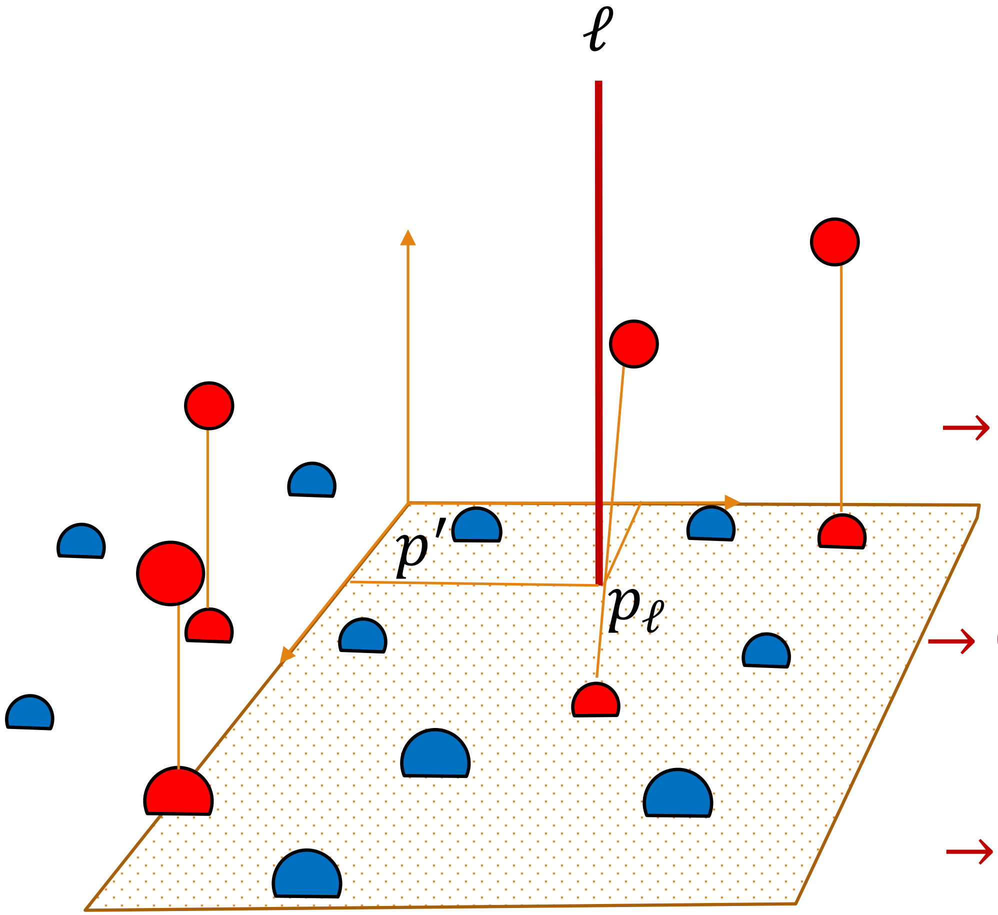
$dist(p, \ell) = dist(p', p_\ell)$

$\ell \cap xy$-plane

→ Project points onto $\boldsymbol{xy}$-plane
$\boldsymbol{P' = \{\, p' \mid p \in P \,\}}$

→ Compute a 1-Center coreset $\boldsymbol{C'}$ for $\boldsymbol{P'}$

Coreset for 1-Line parallel to z-axis in R^3



$proj(p, xy)$

Notice that:

$dist(p, \ell) = dist(p', p_\ell)$

$\ell \cap xy$-plane

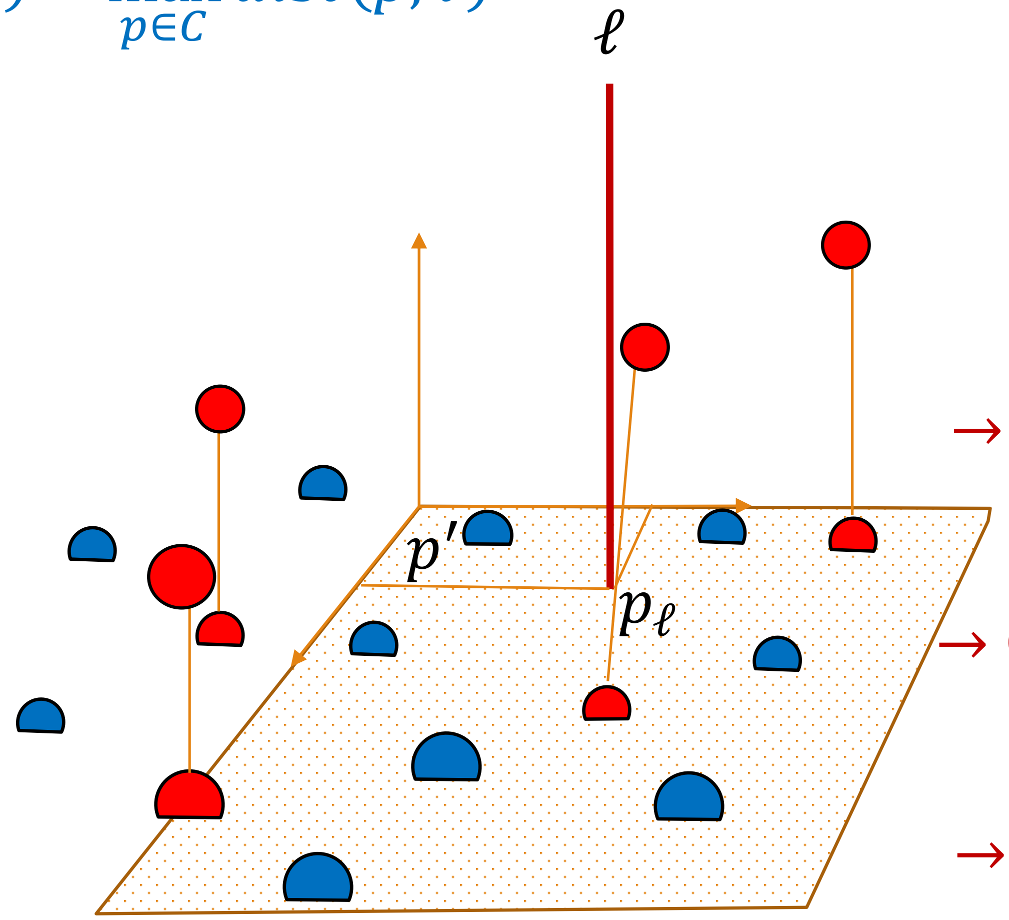
→ Project points onto xy-plane
$P' = \{\, p' \mid p \in P \,\}$

→ Compute a 1-Center coreset C' for P'

→ Return to original data
$C' \to C$

Coreset for 1-Line parallel to z-axis in R^3

$$\forall \ell \in Q: \max_{p \in P} dist(p,\ell) - \max_{p \in C} dist(p,\ell)$$



Notice that:

$proj(p,xy)$

$$dist(p,\ell) = dist(p', p_\ell)$$

$\ell \cap xy$-plane

→ Project points onto xy-plane
$P' = \{ p' \mid p \in P \}$

→ Compute a 1-Center coreset C' for P'

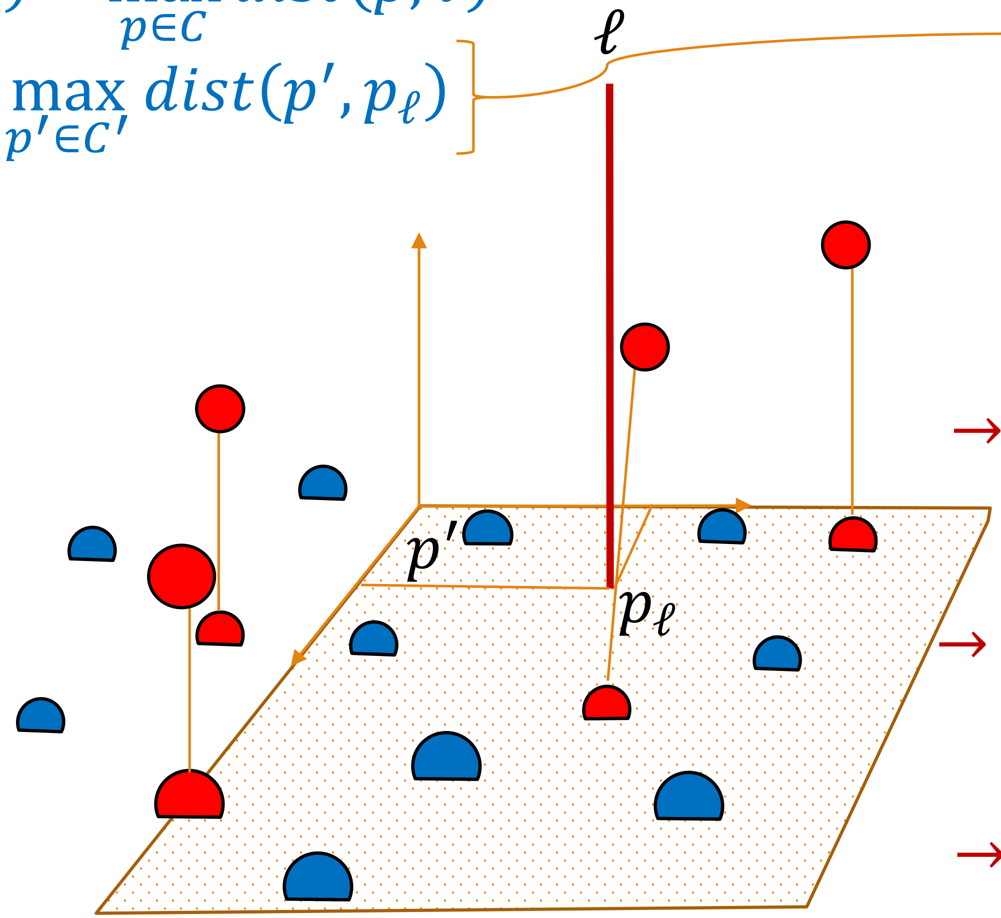
→ Return to original data
$C' \to C$

Coreset for 1-Line parallel to z-axis in R^3

$$\forall \ell \in Q: \max_{p \in P} dist(p,\ell) - \max_{p \in C} dist(p,\ell)$$
$$= \max_{p' \in P'} dist(p',p_\ell) - \max_{p' \in C'} dist(p',p_\ell)$$

Notice that:

$$dist(p,\ell) = dist(\underbrace{p'}_{proj(p,xy)},\underbrace{p_\ell}_{\ell \,\cap\, xy\text{-plane}})$$



→ **Project points onto xy-plane**
$\boldsymbol{P' = \{\, p' \mid p \in P \,\}}$

→ **Compute a 1-Center coreset C' for P'**

→ **Return to original data**
$\boldsymbol{C' \to C}$

Coreset for 1-Line parallel to z-axis in R^3

$$\forall \ell \in Q: \max_{p \in P} dist(p,\ell) - \max_{p \in C} dist(p,\ell)$$
$$= \max_{p' \in P'} dist(p',p_\ell) - \max_{p' \in C'} dist(p',p_\ell)$$
$$\leq \epsilon \cdot \max_{p' \in P'} dist(p',p_\ell)$$

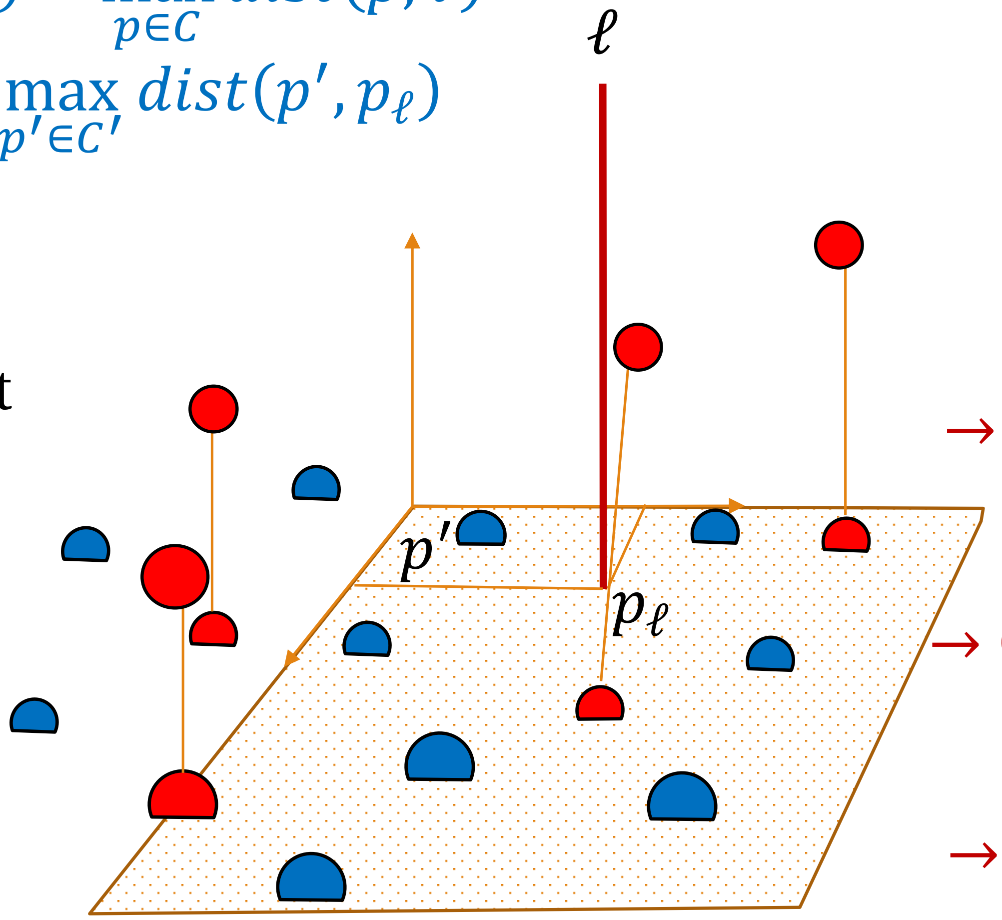
C' is a 1-Center coreset

Notice that:

$$\underset{\ell \,\cap\, xy\text{-plane}}{\overset{proj(p,xy)}{dist(p,\ell) = dist(p',p_\ell)}}$$

$proj(p,xy)$

$\ell \cap xy$-plane



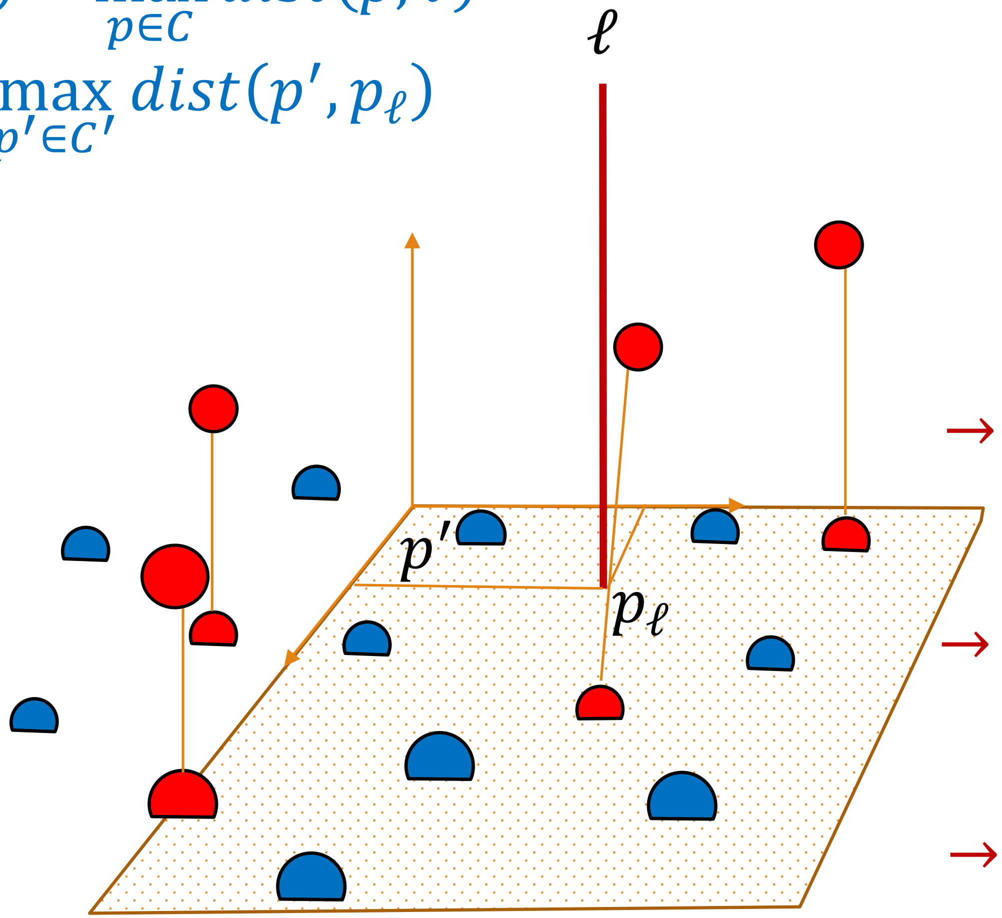
→ **Project points onto xy-plane**
$P' = \{ p' \mid p \in P \}$

→ **Compute a 1-Center coreset C' for P'**

→ **Return to original data**
$C' \to C$

Coreset for 1-Line parallel to z-axis in R^3

$$\forall \ell \in Q: \max_{p \in P} dist(p, \ell) - \max_{p \in C} dist(p, \ell)$$
$$= \max_{p' \in P'} dist(p', p_\ell) - \max_{p' \in C'} dist(p', p_\ell)$$
$$\leq \epsilon \cdot \max_{p' \in P'} dist(p', p_\ell)$$
$$= \epsilon \cdot \max_{p \in P} dist(p, \ell)$$



Notice that:

$$dist(p, \ell) = dist(\overset{proj(p,xy)}{p'}, \underset{\ell \cap xy\text{-plane}}{p_\ell})$$

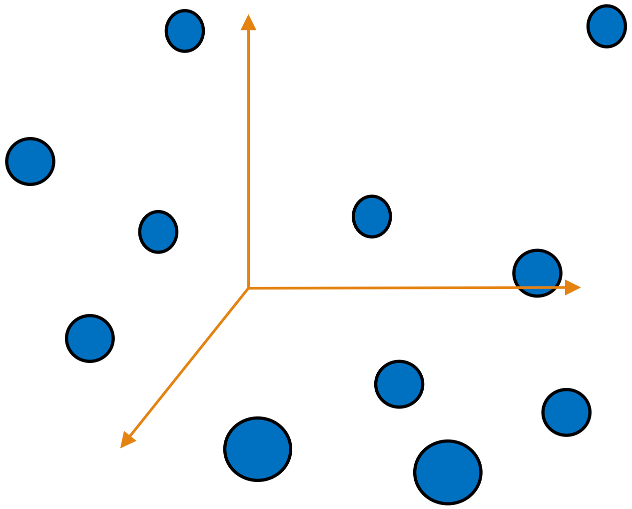
→ Project points onto $\boldsymbol{xy}$-plane
$\boldsymbol{P' = \{ p' \mid p \in P \}}$

→ Compute a 1-Center coreset $\boldsymbol{C}$ for $\boldsymbol{P'}$

→ Return to original data
$\boldsymbol{C' \rightarrow C}$

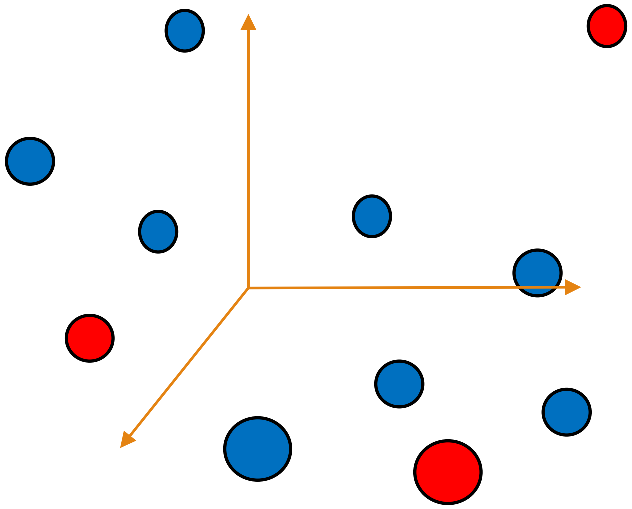
Coreset for 1-Line in R^3

- Input: $P \subseteq R^3$
- Query space: $Q = \{\ell \mid \ell \textit{ is a line in } R^3\}$
- Cost function: $dist(p, \ell) = \min_{x \in \ell} \|p - x\|_2$
- Output: $\boldsymbol{C} \subseteq P \ s.t.\ \forall \ell \in Q: \max_{p \in P} dist(p, \ell) - \max_{c \in C} dist(c, \ell) \leq \epsilon \cdot \max_{p \in P} dist(p, \ell)$



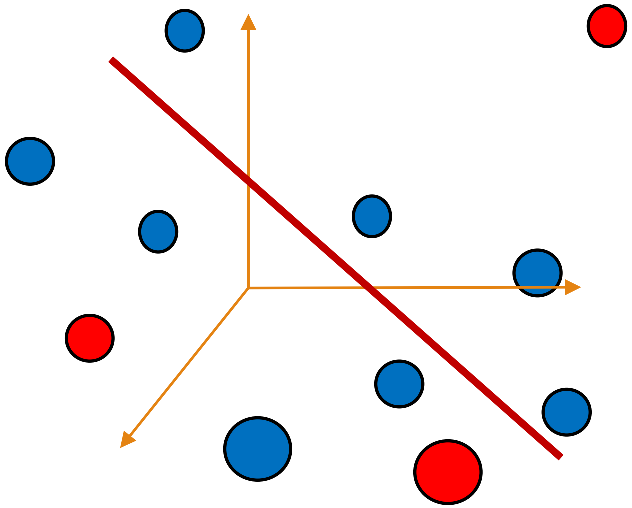
Coreset for 1-Line in R^3

- Input: $P \subseteq R^3$
- Query space: $Q = \{\ell \mid \ell \text{ is a line in } R^3\}$
- Cost function: $dist(p, \ell) = \min_{x \in \ell} \|p - x\|_2$
- Output: $\boldsymbol{C} \subseteq P \ s.t.\ \forall \ell \in Q: \max_{p \in P} dist(p, \ell) - \max_{c \in C} dist(c, \ell) \leq \epsilon \cdot \max_{p \in P} dist(p, \ell)$



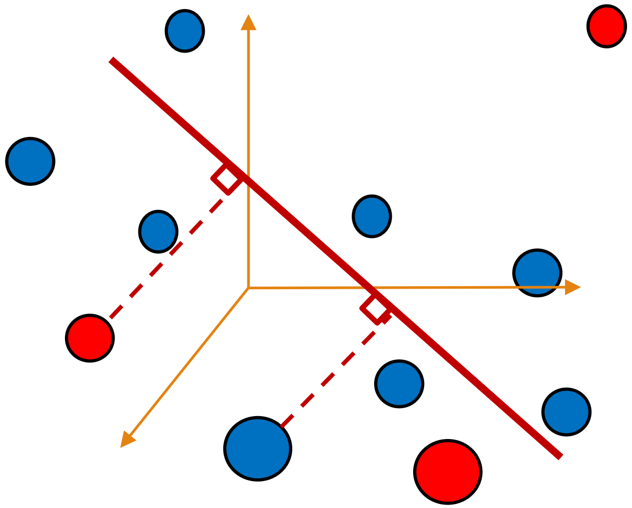
Coreset for 1-Line in R^3

- Input: $P \subseteq R^3$
- Query space: $Q = \{\ell \mid \ell \text{ is a line in } R^3\}$
- Cost function: $dist(p,\ell) = \min_{x\in\ell} \|p - x\|_2$
- Output: $C \subseteq P \ s.t.\ \forall \ell \in Q: \max_{p\in P} dist(p,\ell) - \max_{c\in C} dist(c,\ell) \le \epsilon \cdot \max_{p\in P} dist(p,\ell)$



Coreset for 1-Line in R^3

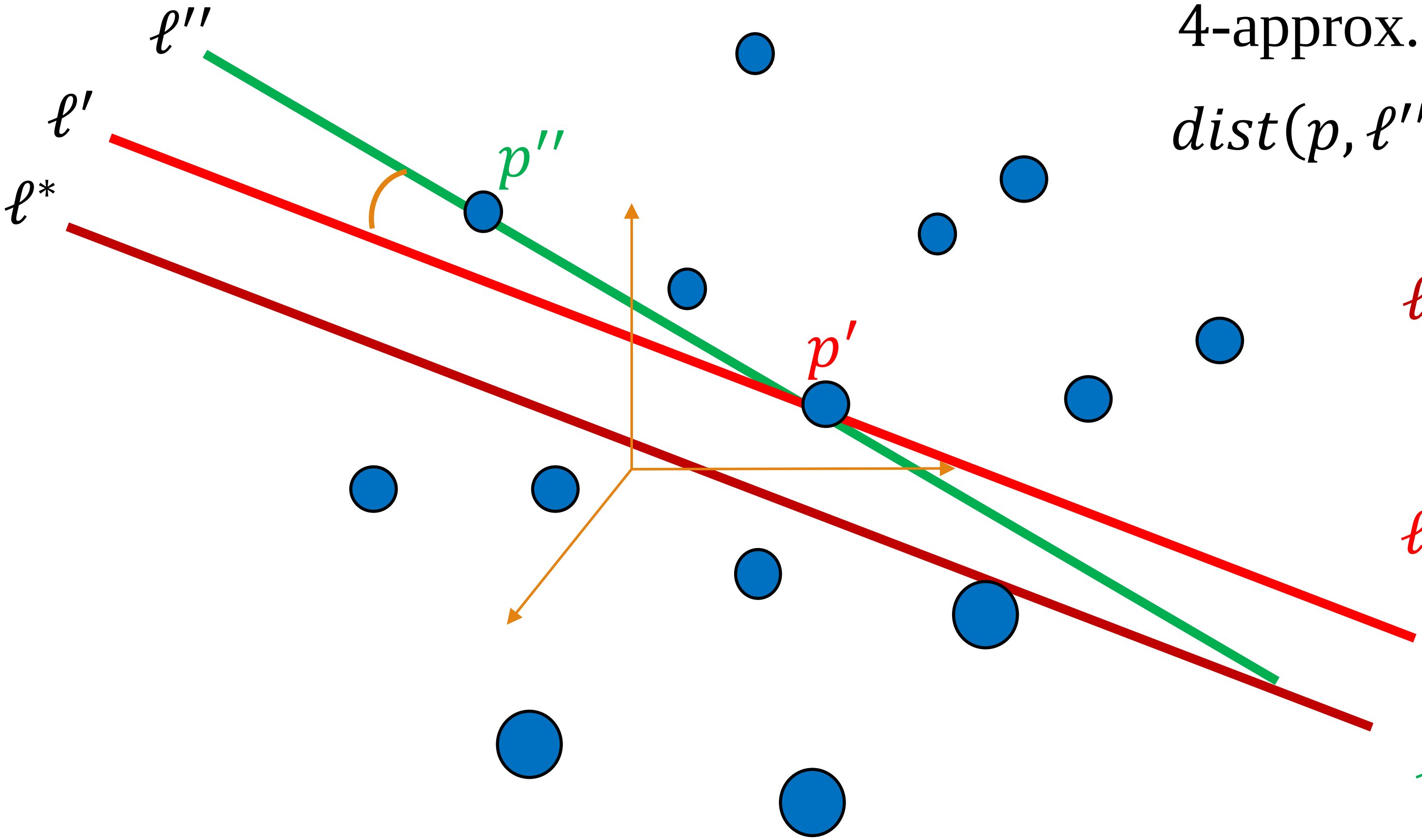
- Input: $P \subseteq R^3$
- Query space: $Q = \{\ell \mid \ell \text{ is a line in } R^3\}$
- Cost function: $dist(p, \ell) = \min_{x \in \ell} \|p - x\|_2$
- Output: $C \subseteq P \text{ s.t. } \forall \ell \in Q: \max_{p \in P} dist(p, \ell) - \max_{c \in C} dist(c, \ell) \le \epsilon \cdot \max_{p \in P} dist(p, \ell)$



Coreset for 1-Line in R^3

Similar to the problem in R^2, there is a line ℓ'' that passes through 2 points of the data and is a 4-approx. to the optimal line ℓ^*.

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



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

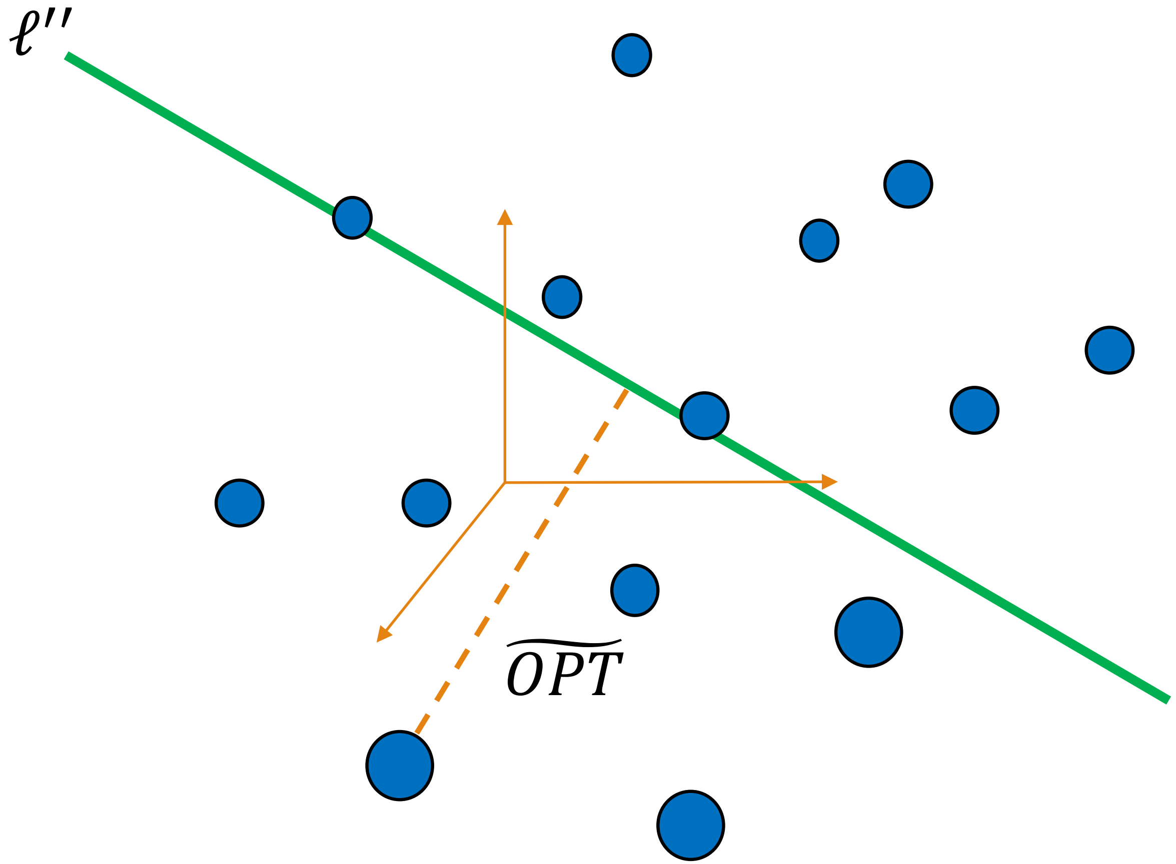
ℓ' is the translation of ℓ^* to $\ell^{*\prime}s$ closest point p'

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

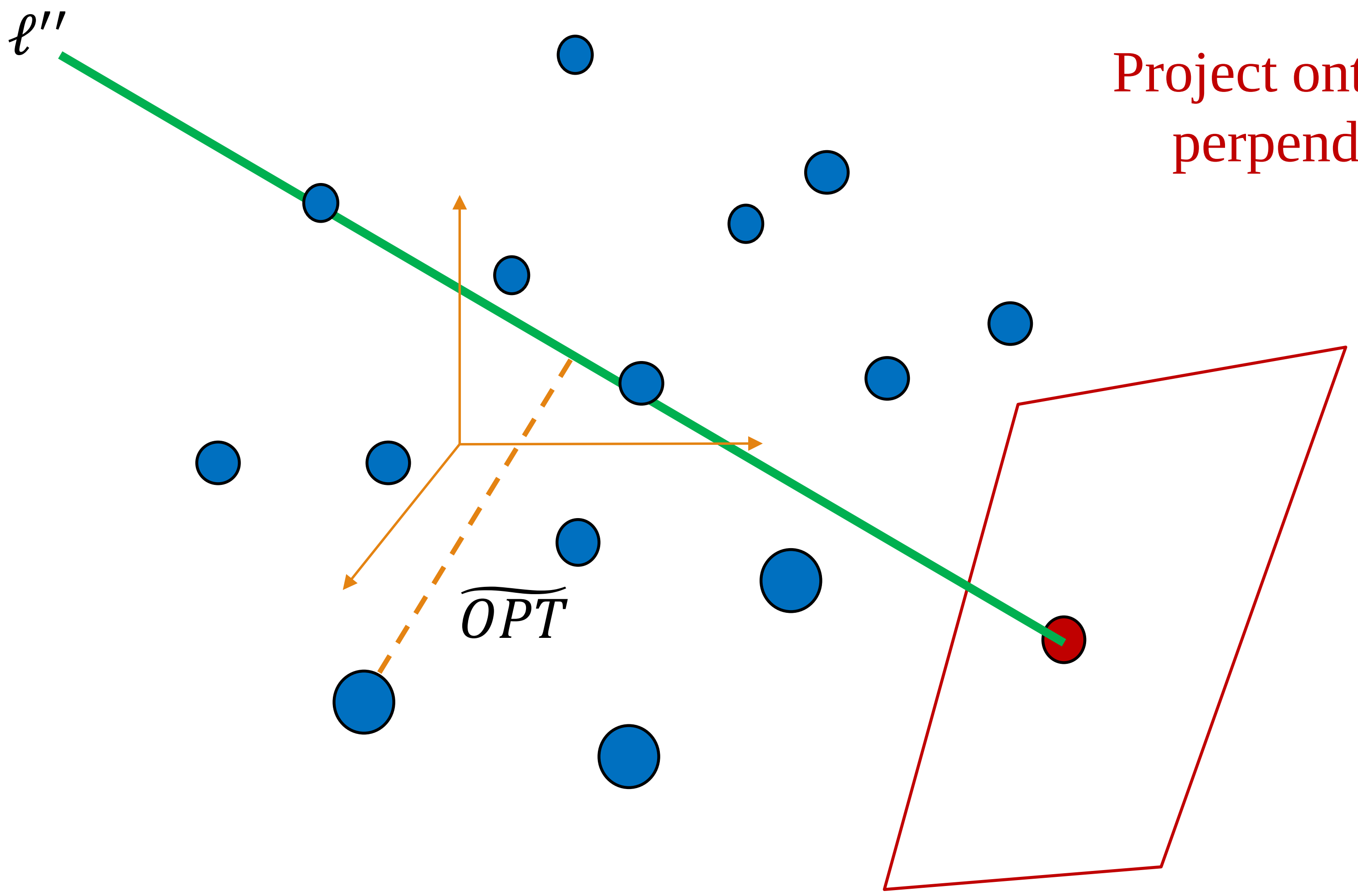
Coreset for 1-Line in R^3

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

$\boldsymbol{O(n^2)}$



Coreset for 1-Line in R^3



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

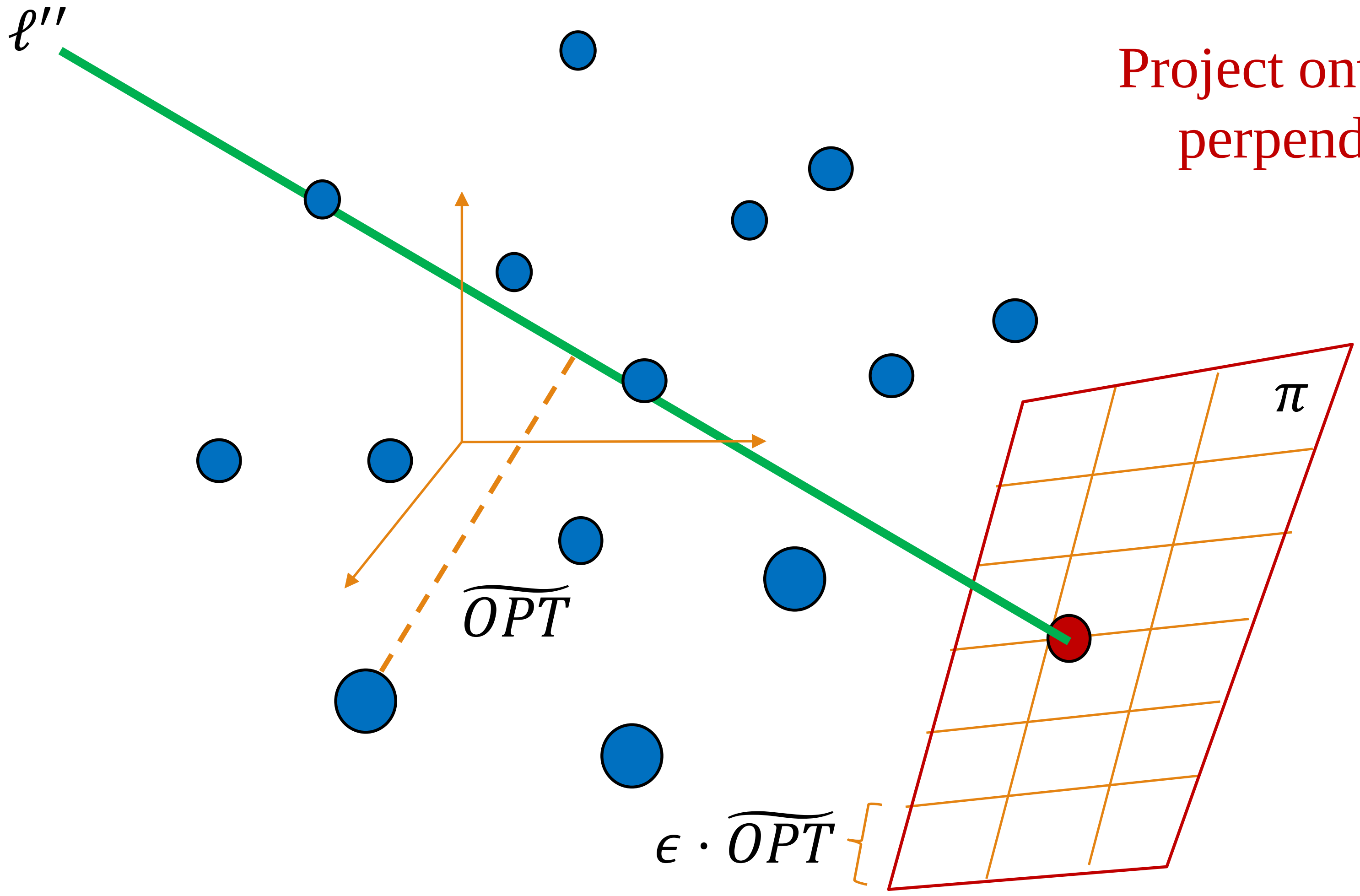
$\boldsymbol{O(n^2)}$

Project onto the plane π perpendicular to ℓ''

Coreset for 1-Line in R^3

Find ℓ'' by exhaustive search over every pair of points.
$\boldsymbol{O(n^2)}$

Project onto the plane π perpendicular to ℓ''



Build a grid with distances $\boldsymbol{\epsilon \cdot \widetilde{OPT}}$

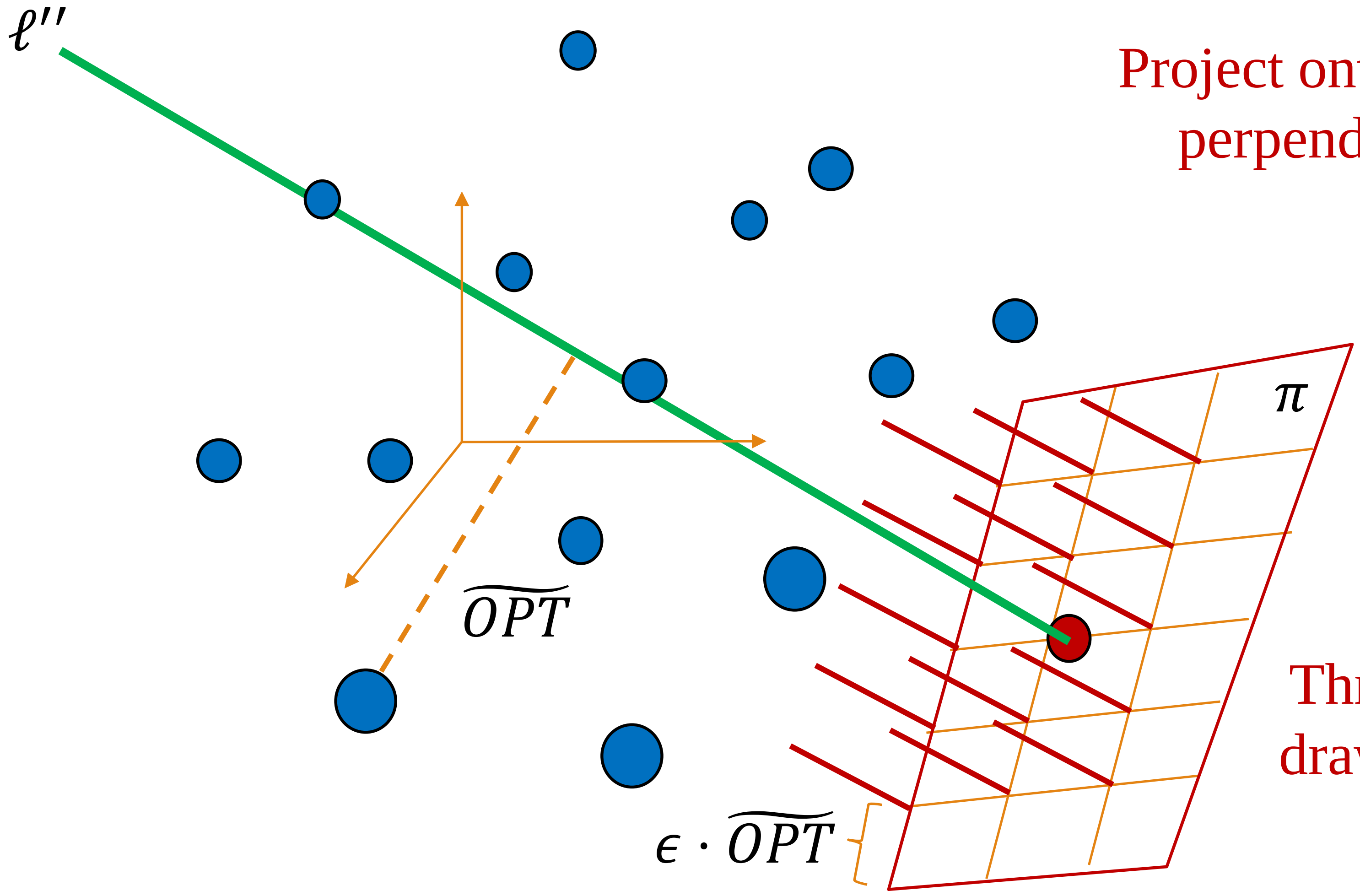
Coreset for 1-Line in R^3

Find ℓ'' by exhaustive search over every pair of points.
$\boldsymbol{O(n^2)}$

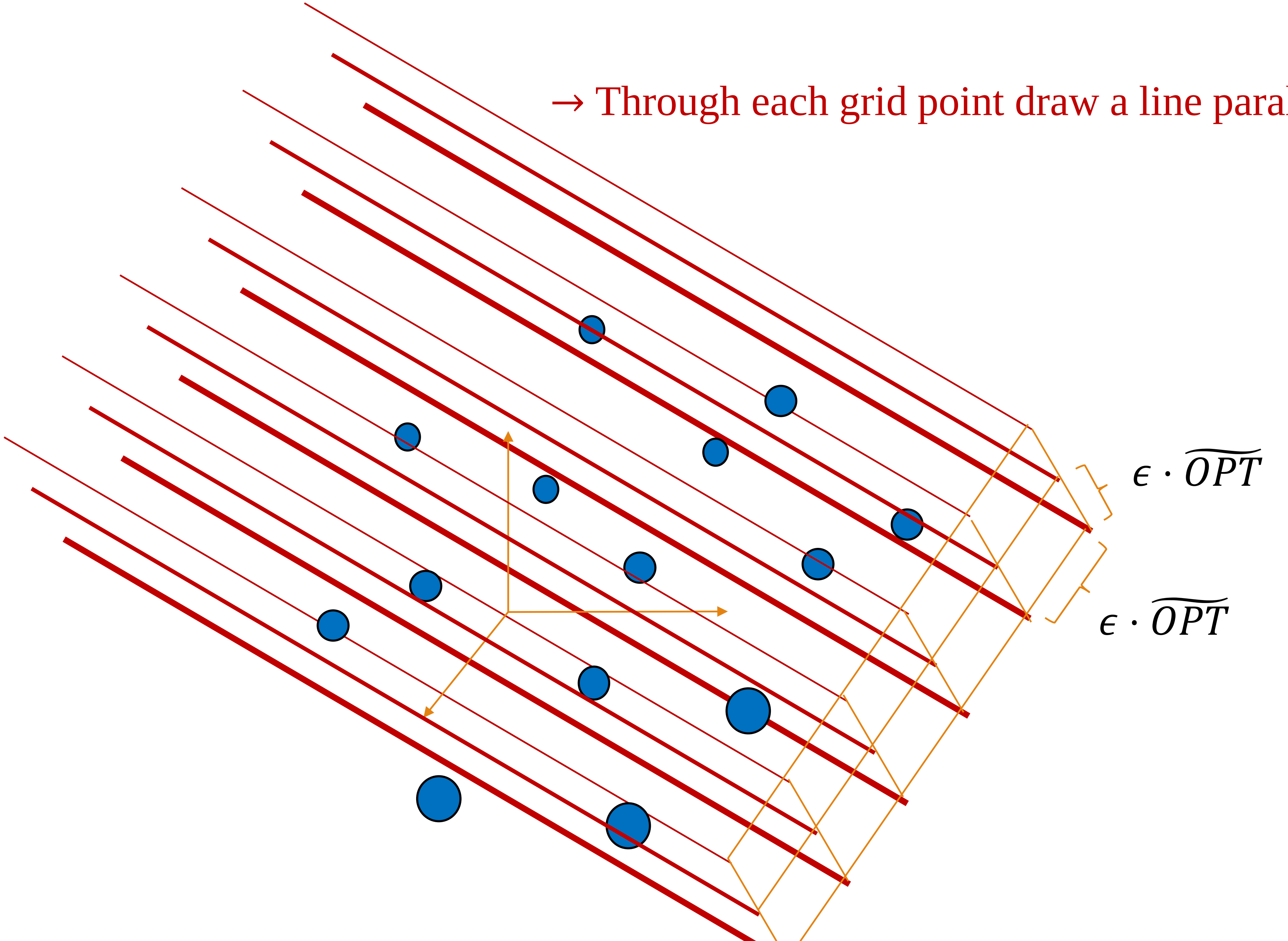
Project onto the plane π perpendicular to ℓ''

Build a grid with distances $\boldsymbol{\epsilon \cdot \widetilde{OPT}}$

Through each grid point draw a line parallel to ℓ''

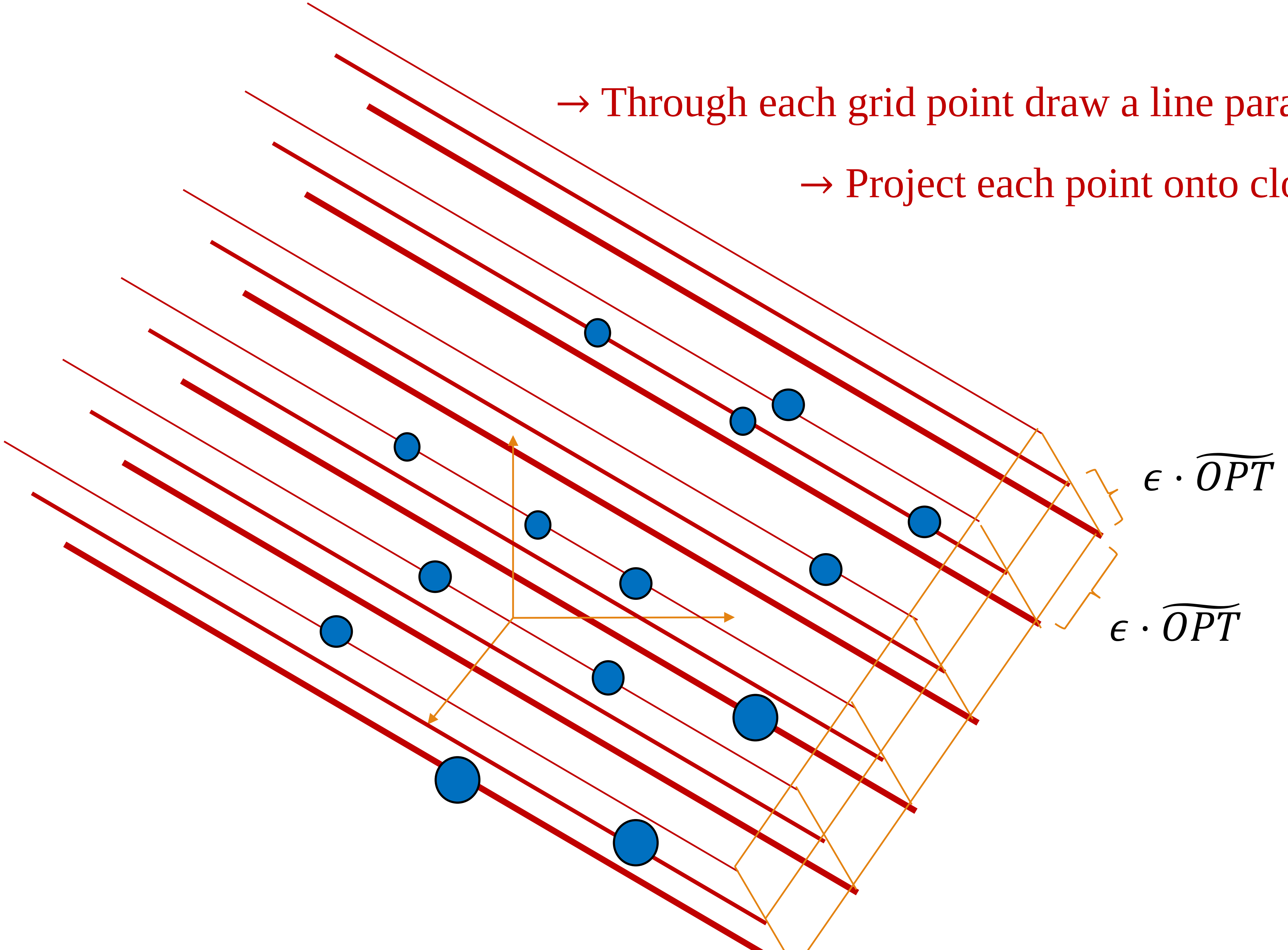


→ Through each grid point draw a line parallel to ℓ''

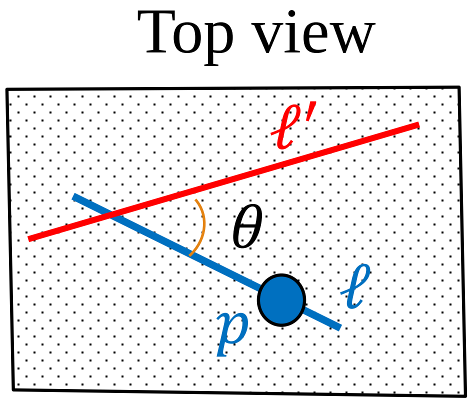


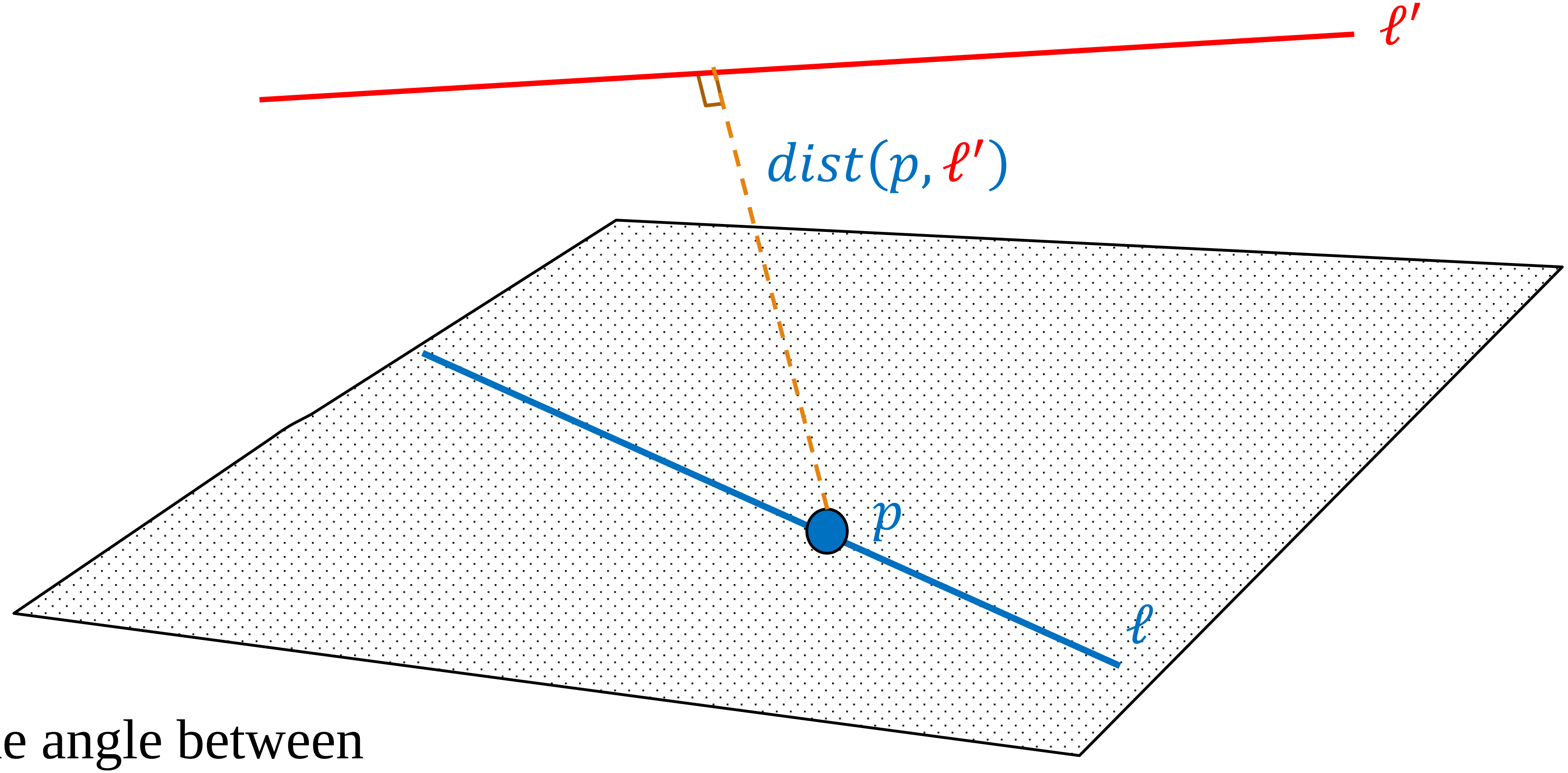
→ Through each grid point draw a line parallel to ℓ''

→ Project each point onto closest line



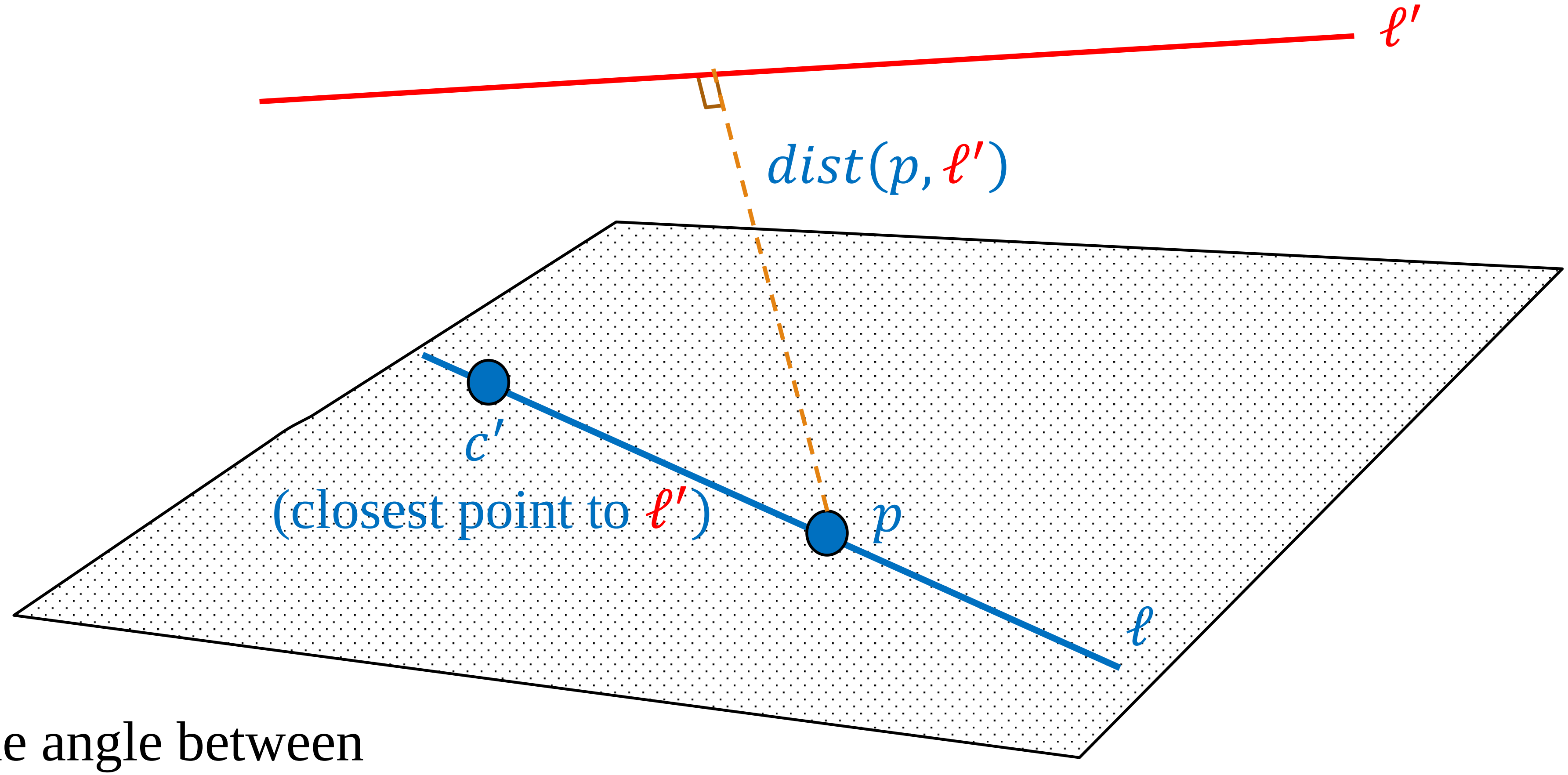
Distance between $p \in \ell$ and ℓ'

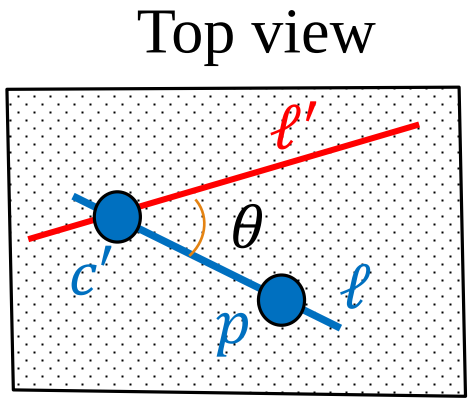




θ is the angle between the line directions

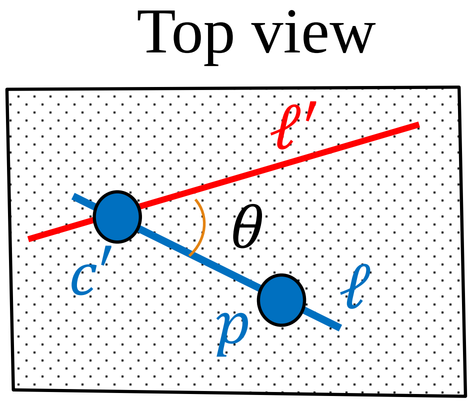
Distance between $p \in \ell$ and ℓ'

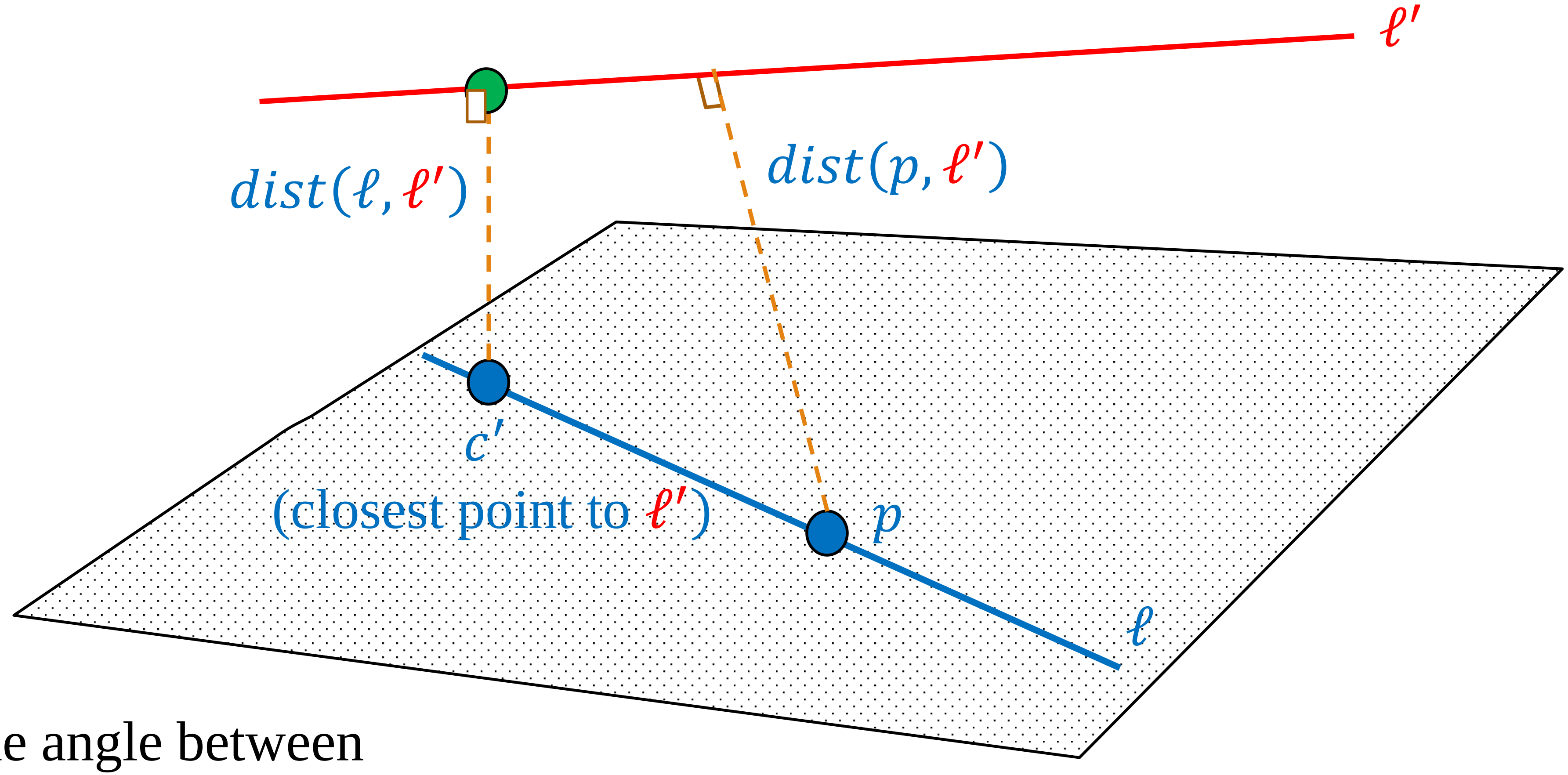




θ is the angle between
the line directions

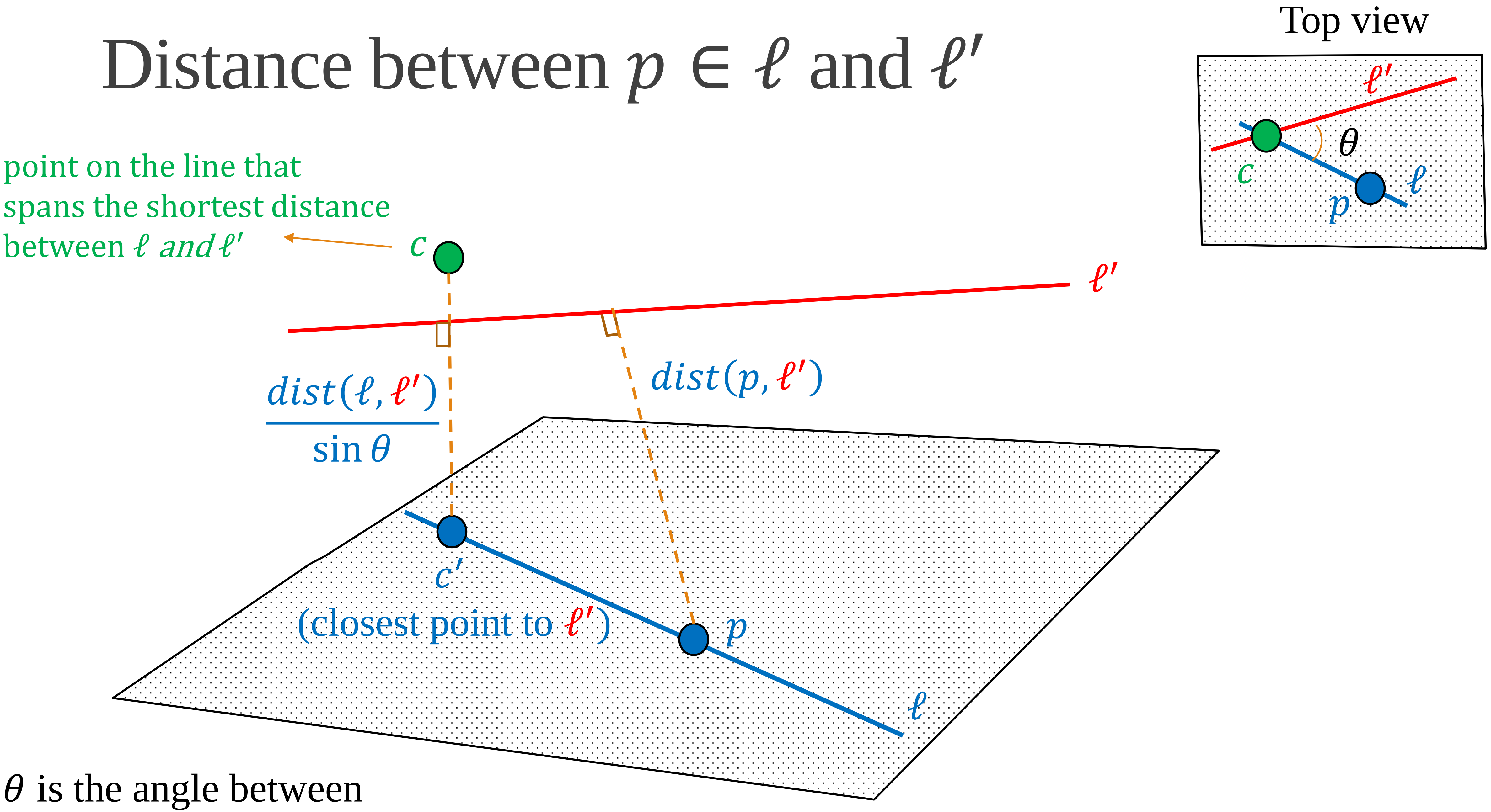
Distance between $p \in \ell$ and ℓ'





θ is the angle between the line directions

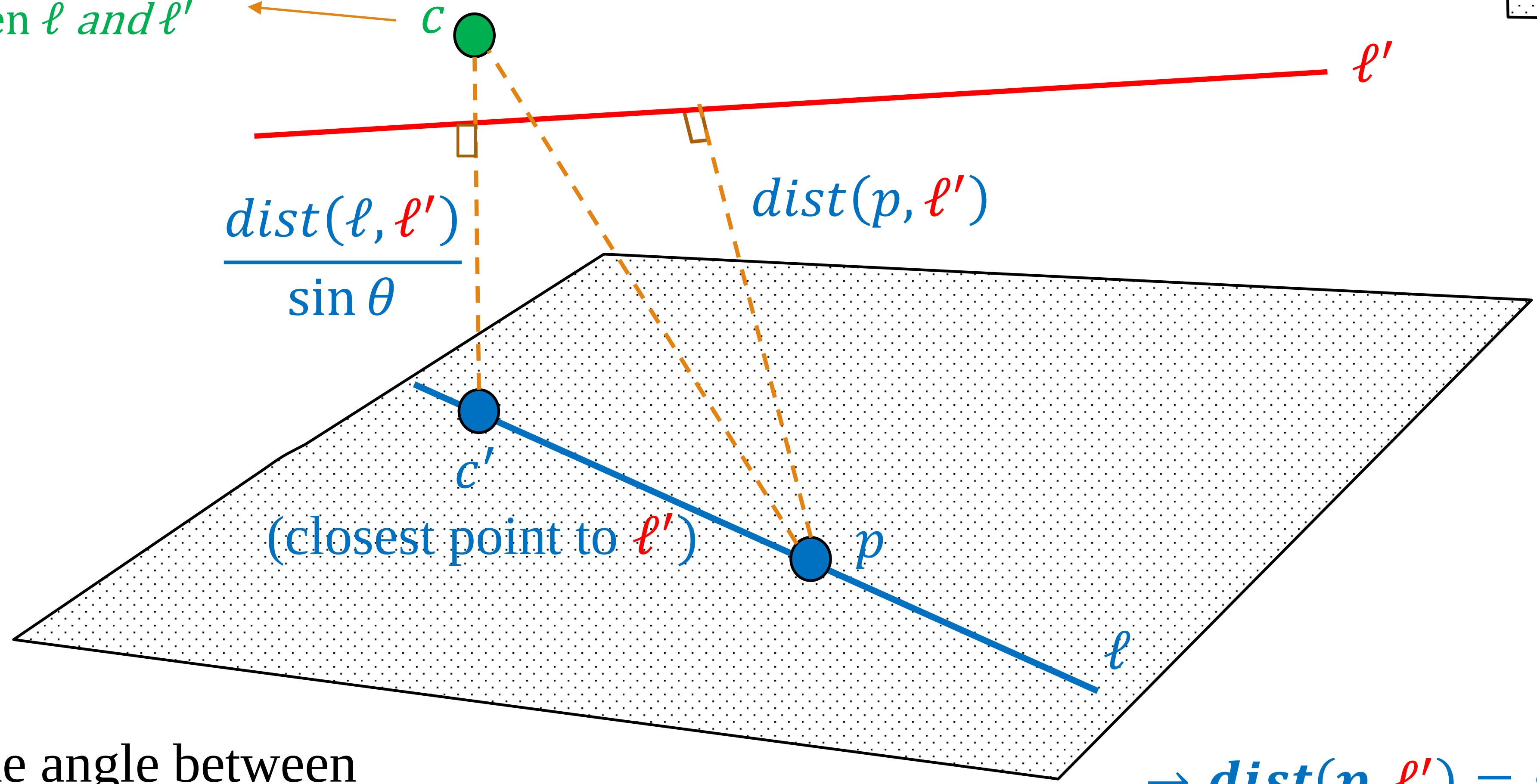
Distance between $p \in \ell$ and ℓ'



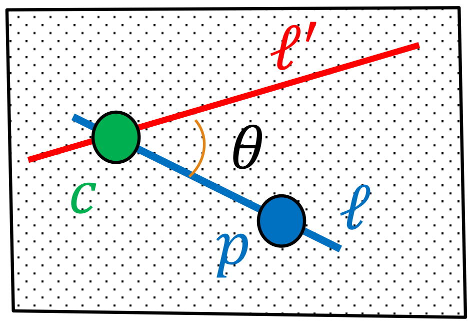
θ is the angle between the line directions

Distance between $p \in \ell$ and ℓ'

point on the line that spans the shortest distance between ℓ *and* ℓ'



Top view

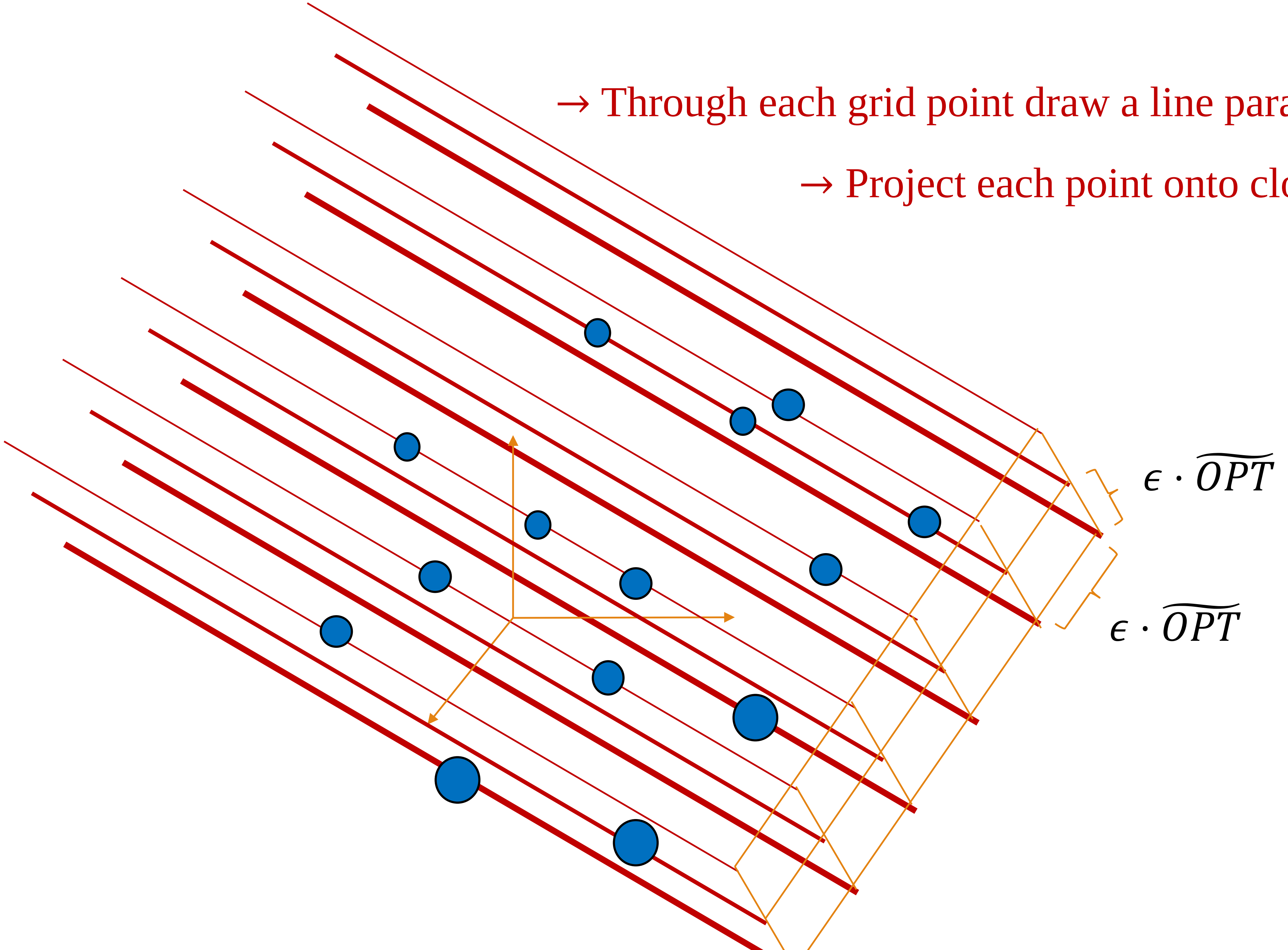


θ is the angle between the line directions

$\rightarrow \boldsymbol{dist(p, \ell') = \sin\theta \cdot dist(p, c)}$

→ Through each grid point draw a line parallel to ℓ''

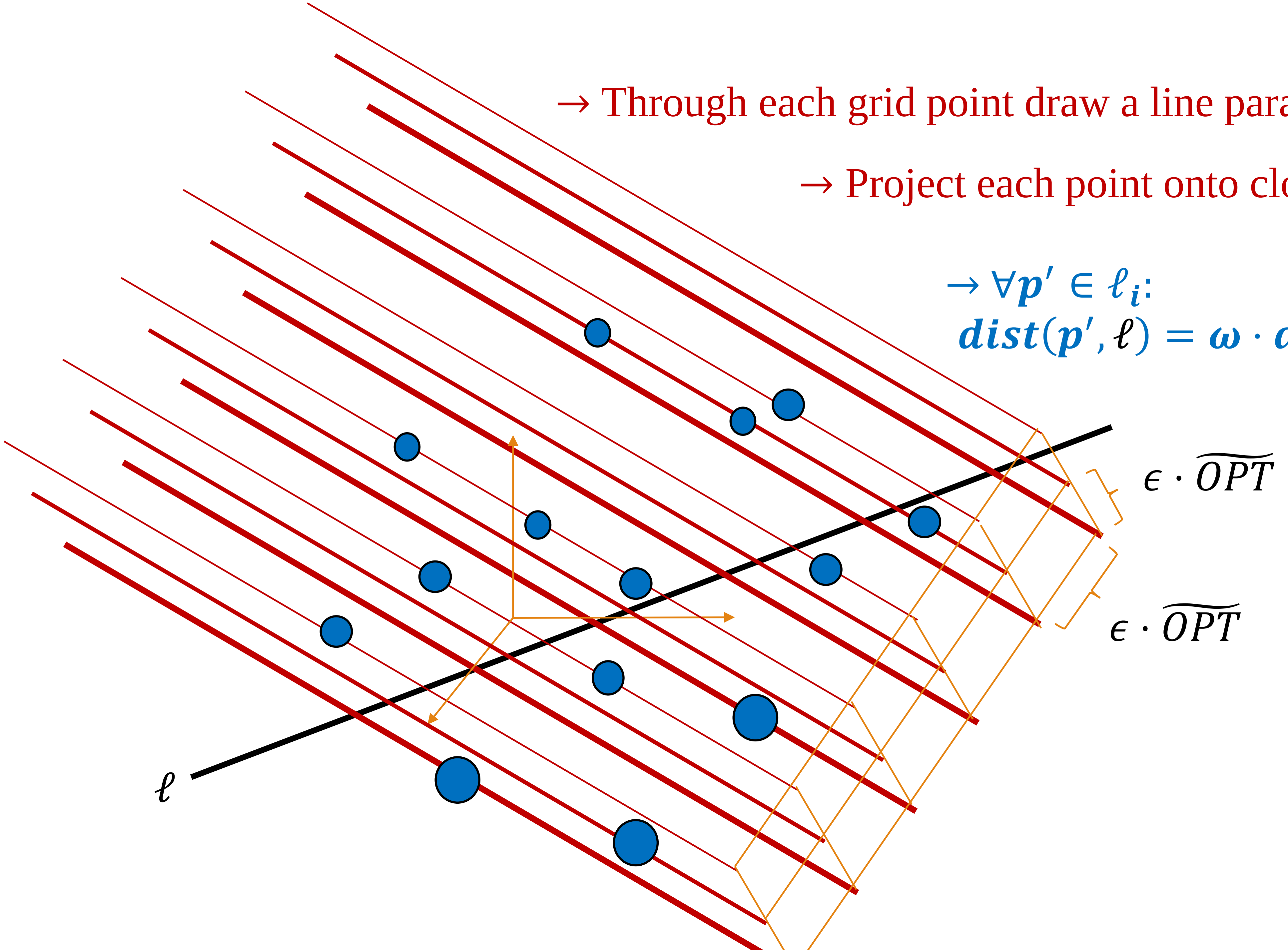
→ Project each point onto closest line



→ Through each grid point draw a line parallel to ℓ''

→ Project each point onto closest line

→ $\forall \boldsymbol{p'} \in \ell_{\boldsymbol{i}}$:
$\boldsymbol{dist}(\boldsymbol{p'}, \ell) = \boldsymbol{\omega} \cdot \boldsymbol{dist}(\boldsymbol{p'}, \boldsymbol{c_i})$

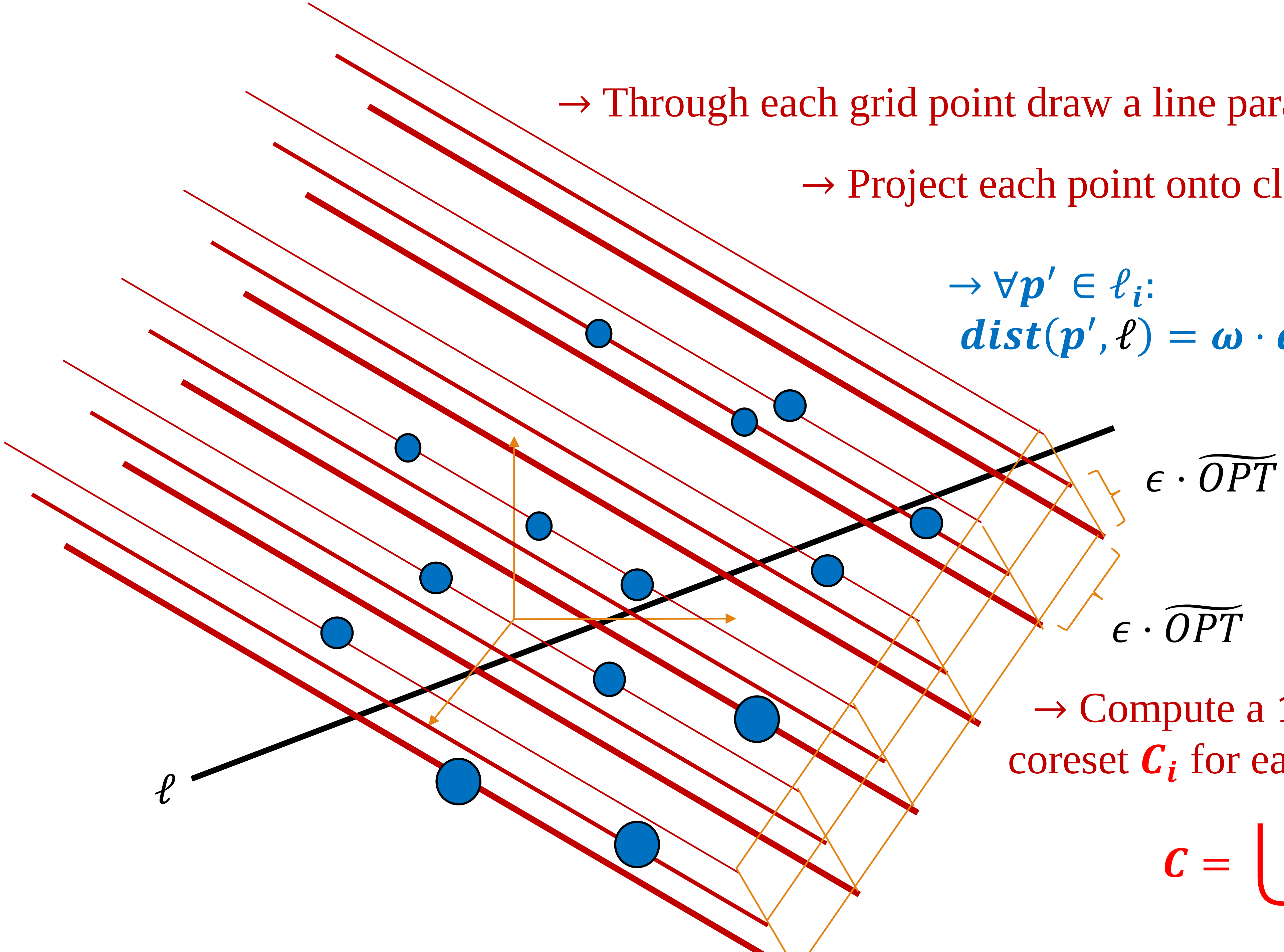


→ Through each grid point draw a line parallel to ℓ''

→ Project each point onto closest line

→ $\forall \boldsymbol{p'} \in \ell_{\boldsymbol{i}}$:

$$\boldsymbol{dist}(\boldsymbol{p'}, \ell) = \boldsymbol{\omega} \cdot \boldsymbol{dist}(\boldsymbol{p'}, \boldsymbol{c_i})$$



→ Compute a **1-Center** coreset $\boldsymbol{C_i}$ for each **line** $\ell_{\boldsymbol{i}}$!

$$\boldsymbol{C} = \bigcup \boldsymbol{C_i}$$

Coreset for j-subspace in R^d

Claim 1:

Let S be an r-dimensional subspace of R^d and let L be an $(r+j)$-dimensional subspace of R^d that contains S. Let V be a j-dimensional subspace of R^d. Then there is an orthogonal matrix U such that $Ux = x$ for every $x \in S$, and $Uc \in L$ for every $c \in V$.

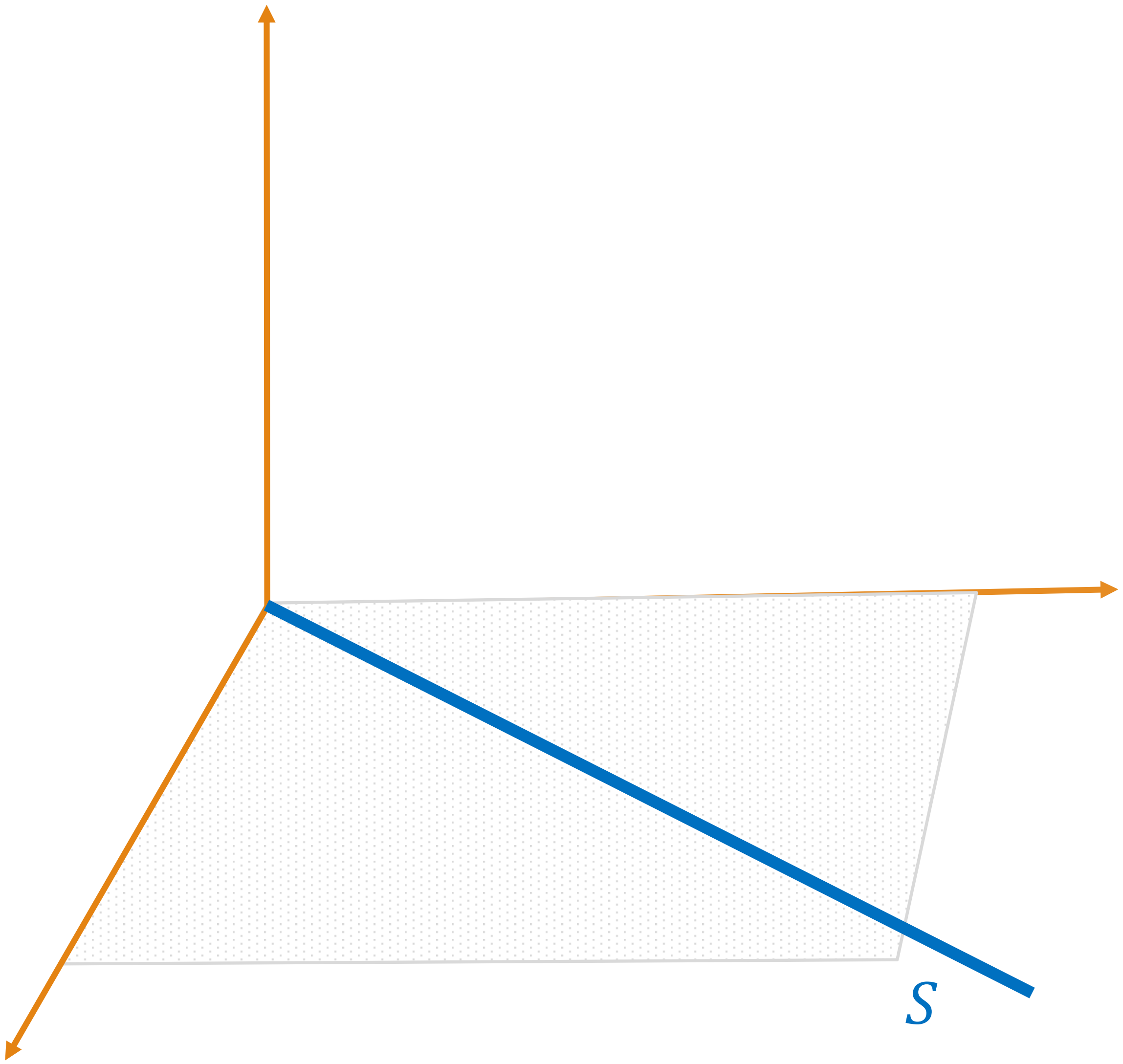
Claim 2:

Let $A \in R^{n \times d}$ be a matrix of rank r and let L be an $(r+j+1)$-dimensional subspace of R^d that contains the row vectors (A_{i*}) for every $1 \leq i \leq n$. Then for evert affine j-dimensional subspace V of R^d there is a corresponding affine j-dimensional subspace $V' \subseteq L$ such that for every $i \in [n]$ we have

$$dist(A_{i*}, V) = dist(A_{i*}, V').$$

Coreset for j-subspace in R^d

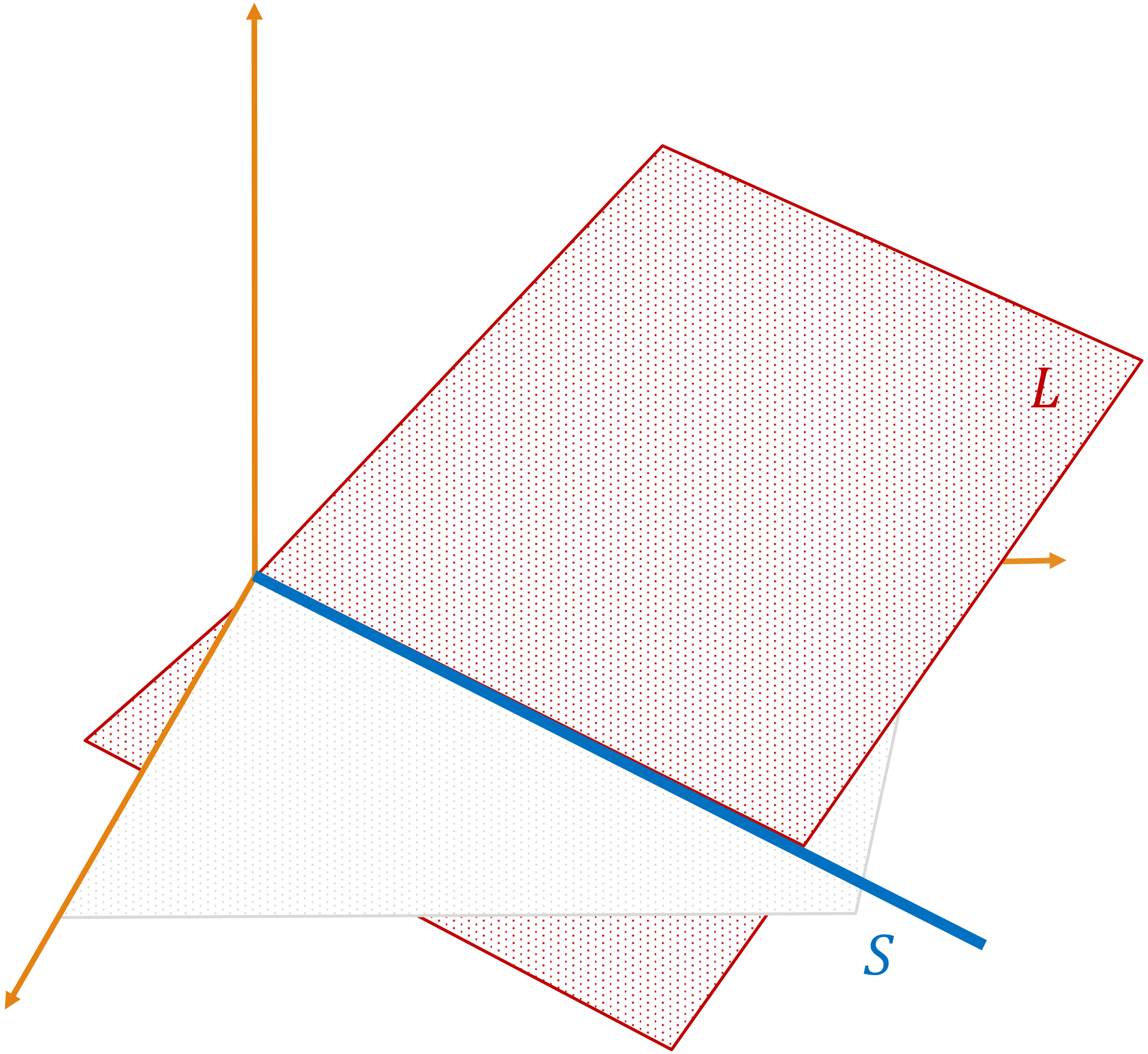
$r = j = 1$



Coreset for j-subspace in R^d

$r = j = 1$

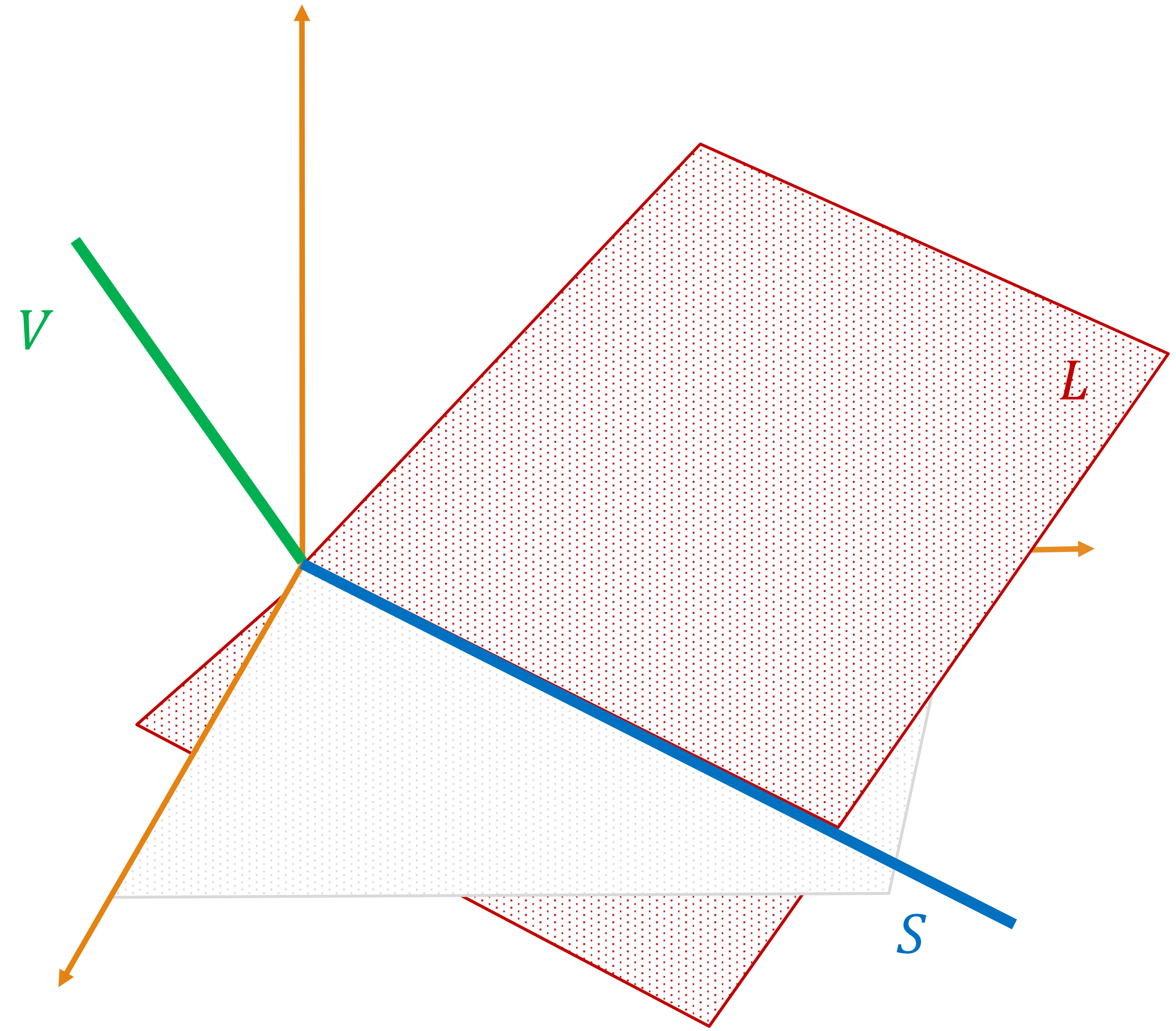
L contains S



Coreset for j-subspace in R^d

$r = j = 1$

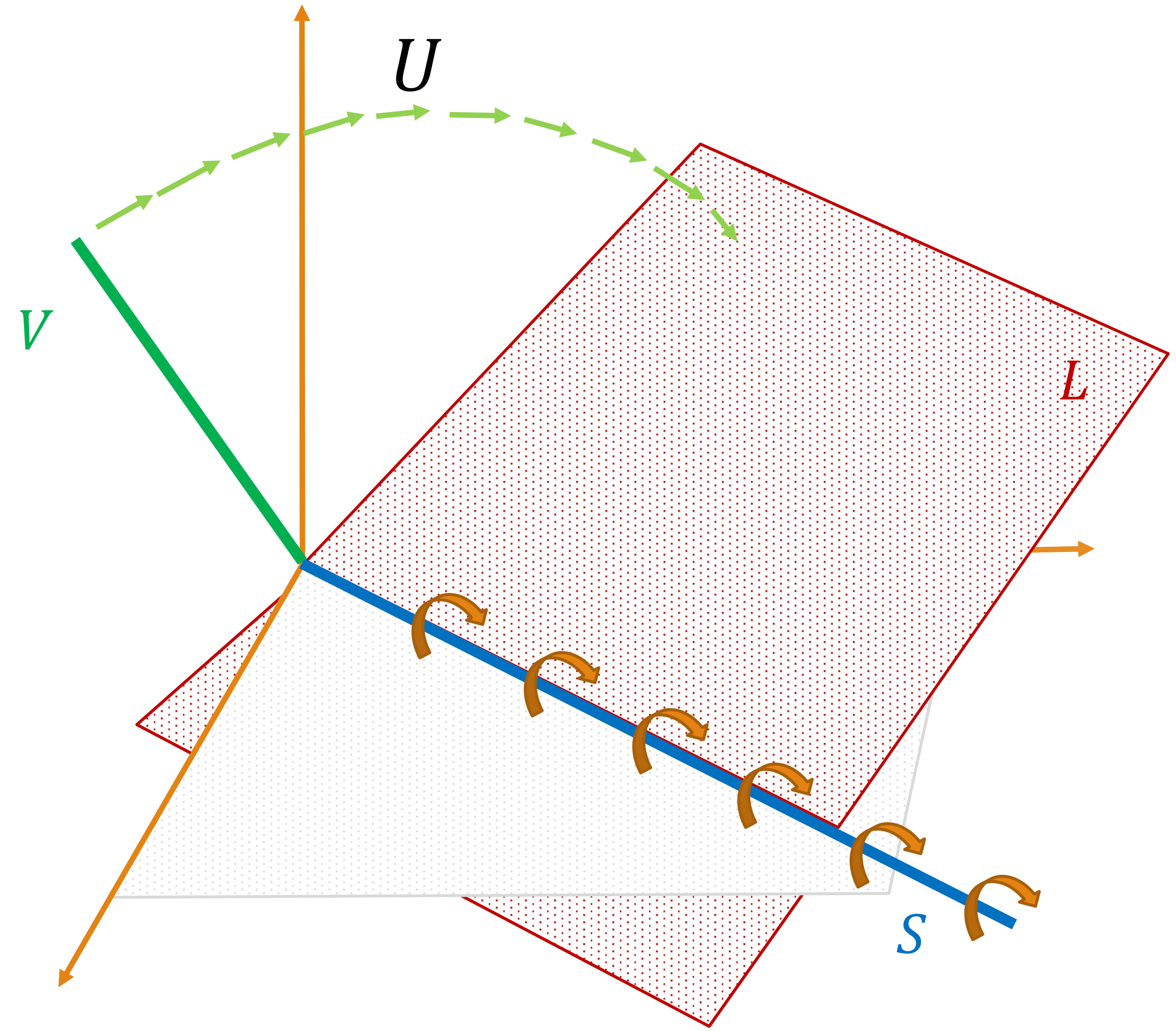
L contains S



Coreset for j-subspace in R^d

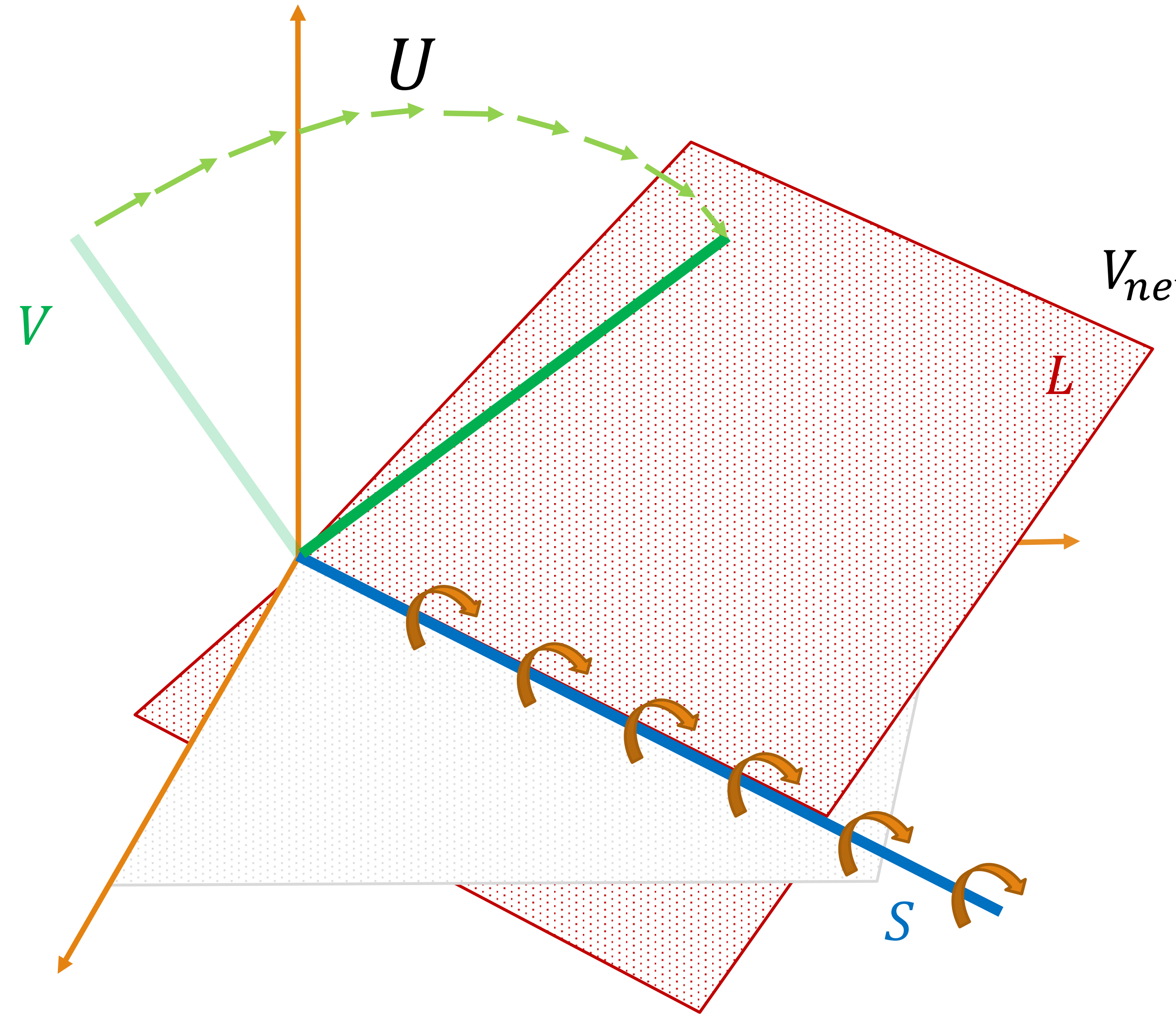
$r = j = 1$

L contains S



Coreset for j-subspace in R^d

$r = j = 1$



L contains S

L contains
$V_{new} = \{Uc \mid c \in V\}$

Coreset for j-subspace in R^d

JSubspaceCoreset(P, j):

Coreset for j-subspace in R^d

<u>JSubspaceCoreset($\boldsymbol{P}, \boldsymbol{j}$):</u>

- $h' \leftarrow$ an α-approximation for the affine j-subspace center of P.

Coreset for j-subspace in R^d

<u>JSubspaceCoreset(P, j):</u>

- $h' \leftarrow$ an α-approximation for the affine j-subspace center of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.

Coreset for j-subspace in R^d

<u>JSubspaceCoreset(P, j):</u>

- $h' \leftarrow$ an α-approximation for the affine j-subspace center of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.
- $h^{\perp} \leftarrow$the affine $d - j$-subspace that is orthogonal to h'.

Coreset for j-subspace in R^d

<u>**JSubspaceCoreset($\boldsymbol{P, j}$):**</u>

- $h' \leftarrow$ an α-approximation for the affine j-subspace center of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.
- $h^{\perp} \leftarrow$the affine $d - j$-subspace that is orthogonal to h'.
- Construct a grid on $h^{\perp}$ whose cell length is $\epsilon \cdot \widetilde{OPT}$.

Coreset for j-subspace in R^d

<u>JSubspaceCoreset(P, j):</u>

- $h' \leftarrow$ an α-approximation for the affine j-subspace center of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.
- $h^{\perp} \leftarrow$the affine $d - j$-subspace that is orthogonal to h'.
- Construct a grid on $h^{\perp}$ whose cell length is $\epsilon \cdot \widetilde{OPT}$.
- Through each grid point construct an affine j-subspace parallel to h'.

$$\#JSubspaces = O\left(\left(\frac{2}{\epsilon}\right)^{d-j}\right).$$

Coreset for j-subspace in R^d

JSubspaceCoreset(P, j):

- $h' \leftarrow$ an α-approximation for the affine j-subspace center of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.
- $h^{\perp} \leftarrow$the affine $d - j$-subspace that is orthogonal to h'.
- Construct a grid on $h^{\perp}$ whose cell length is $\epsilon \cdot \widetilde{OPT}$.
- Through each grid point construct an affine j-subspace parallel to h'.

  #JSubspaces = $O\left(\left(\frac{2}{\epsilon}\right)^{d-j}\right)$.
- Compute the projection p' of each point $p \in P$ onto it's closest affine j-subspace h_p.

Coreset for j-subspace in R^d

<u>JSubspaceCoreset(P, j):</u>

- $h' \leftarrow$ an α-approximation for the affine j-subspace center of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.
- $h^{\perp} \leftarrow$the affine $d - j$-subspace that is orthogonal to h'.
- Construct a grid on $h^{\perp}$ whose cell length is $\epsilon \cdot \widetilde{OPT}$.
- Through each grid point construct an affine j-subspace parallel to h'.

  #JSubspaces = $O\left(\left(\frac{2}{\epsilon}\right)^{d-j}\right)$.

- Compute the projection p' of each point $p \in P$ onto it's closest affine j-subspace h_p.
- H_p ←an $R^{d \times j}$ matrix whose columns span h_p and $H_p^T H_p = I$.

Coreset for j-subspace in R^d

<u>JSubspaceCoreset(P, j):</u>

- $h' \leftarrow$ an α-approximation for the affine j-subspace center of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.
- $h^{\perp} \leftarrow$the affine $d - j$-subspace that is orthogonal to h'.
- Construct a grid on $h^{\perp}$ whose cell length is $\epsilon \cdot \widetilde{OPT}$.
- Through each grid point construct an affine j-subspace parallel to h'.

  #JSubspaces = $O\left(\left(\frac{2}{\epsilon}\right)^{d-j}\right)$.
- Compute the projection p' of each point $p \in P$ onto it's closest affine j-subspace h_p.
- $H_p \leftarrow$an $R^{d \times j}$ matrix whose columns span h_p and $H_p^T H_p = I$.
- $P' = \{H_p p' | p \in P\}$.

Coreset for j-subspace in R^d

<u>**JSubspaceCoreset($\boldsymbol{P, j}$):**</u>

- $h' \leftarrow$ an α-approximation for the affine j-subspace center of P.
- $\widetilde{OPT} = \max_{p \in P} dist(p, h')$.
- $h^{\perp} \leftarrow$the affine $d - j$-subspace that is orthogonal to h'.
- Construct a grid on $h^{\perp}$ whose cell length is $\epsilon \cdot \widetilde{OPT}$.
- Through each grid point construct an affine j-subspace parallel to h'.

  #JSubspaces = $O\left(\left(\frac{2}{\epsilon}\right)^{d-j}\right)$.
- Compute the projection p' of each point $p \in P$ onto it's closest affine j-subspace h_p.
- $H_p \leftarrow$an $R^{d \times j}$ matrix whose columns span h_p and $H_p^T H_p = I$.
- $P' = \{H_p p' | p \in P\}$.
- Call $\boldsymbol{HyperplaneCoreset}(\boldsymbol{P'}, \boldsymbol{j})$.