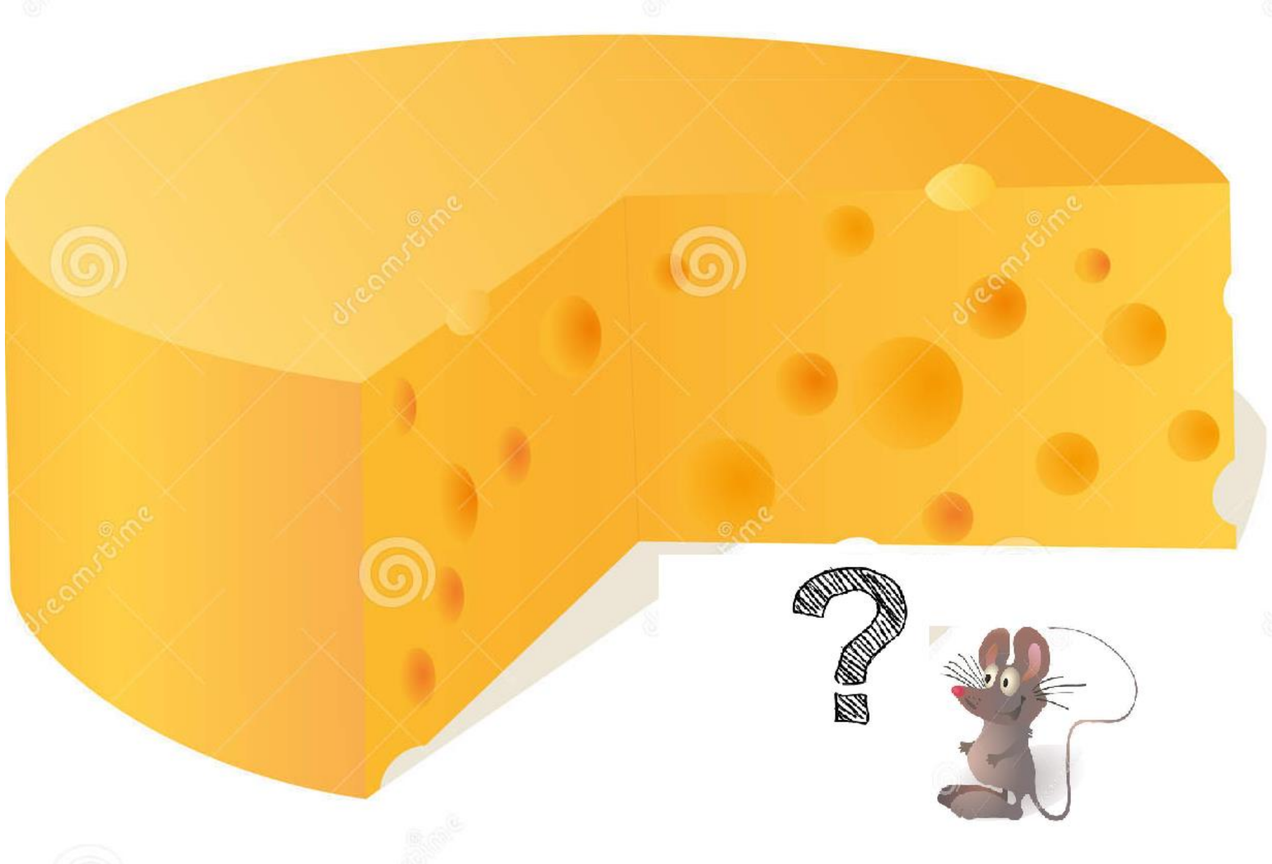
# Big Data Class



LECTURER: DAN FELDMAN

TEACHING ASSISTANTS:

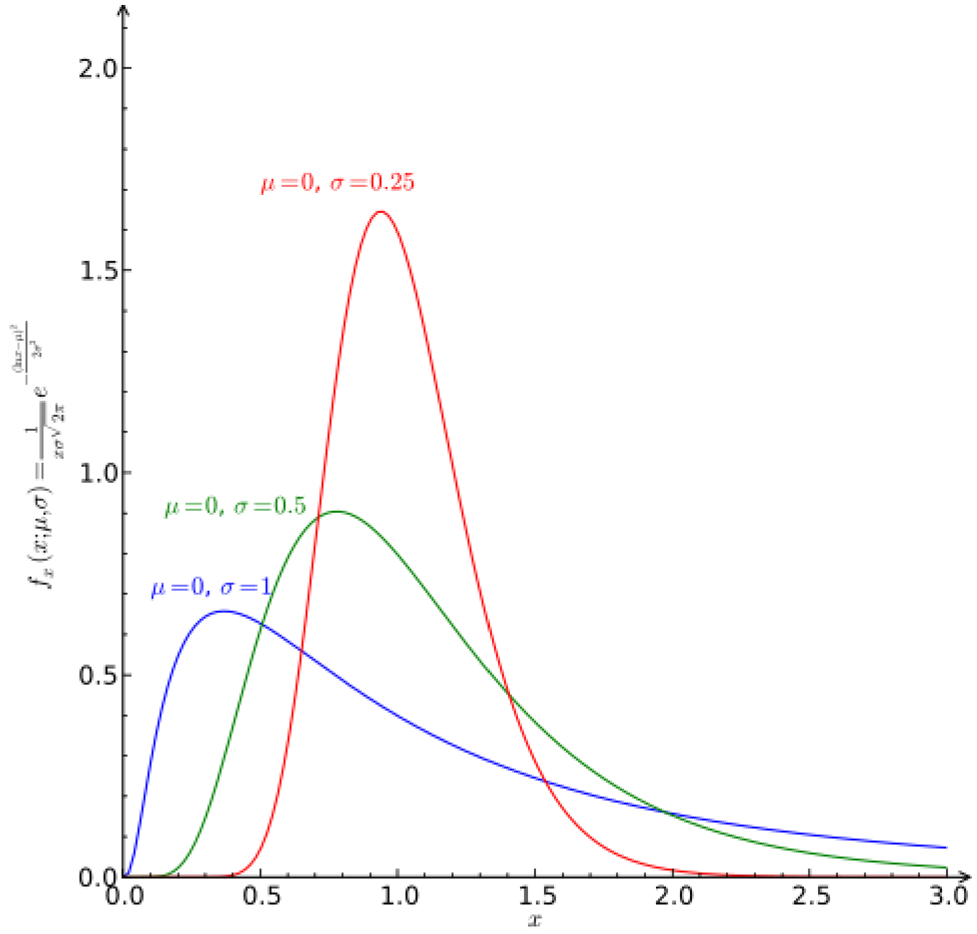
IBRAHIM JUBRAN

ALAA MAALOUF

# Example for Kernel Functions

<u>Gaussian Kernel:</u>

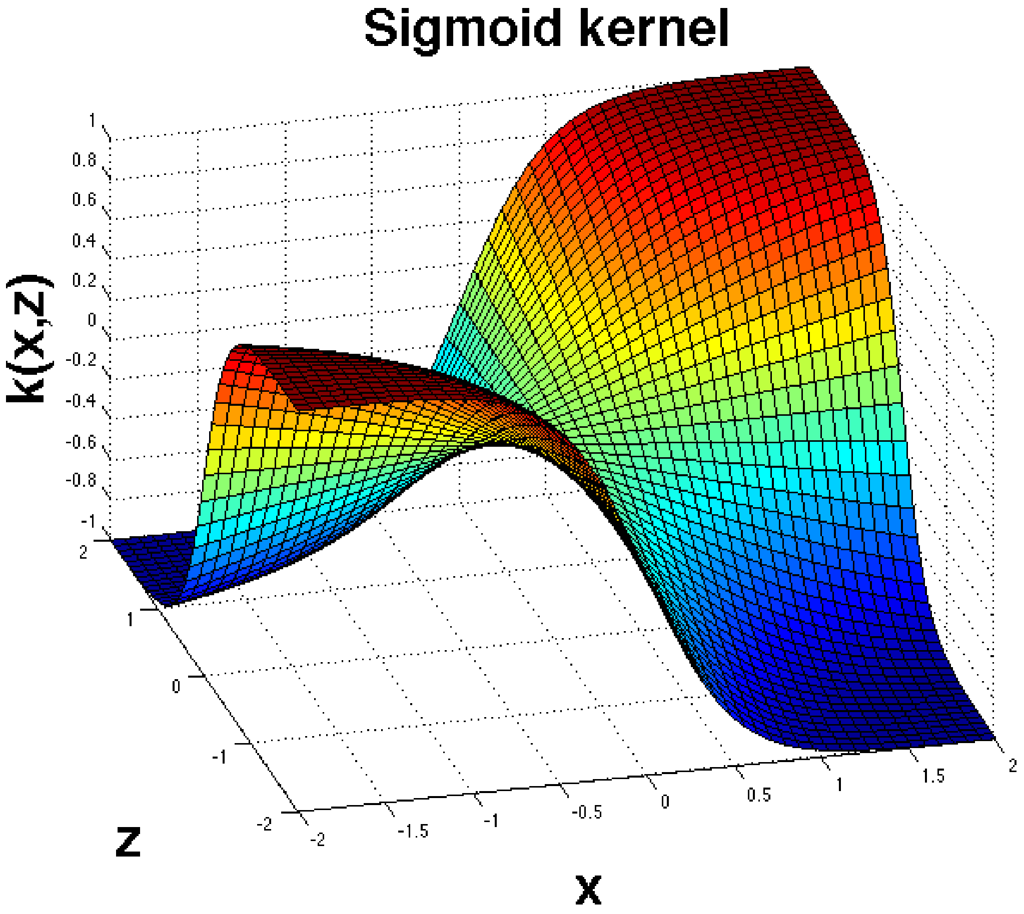
$$k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$$



# Example for Kernel Functions

Sigmoid Kernel:

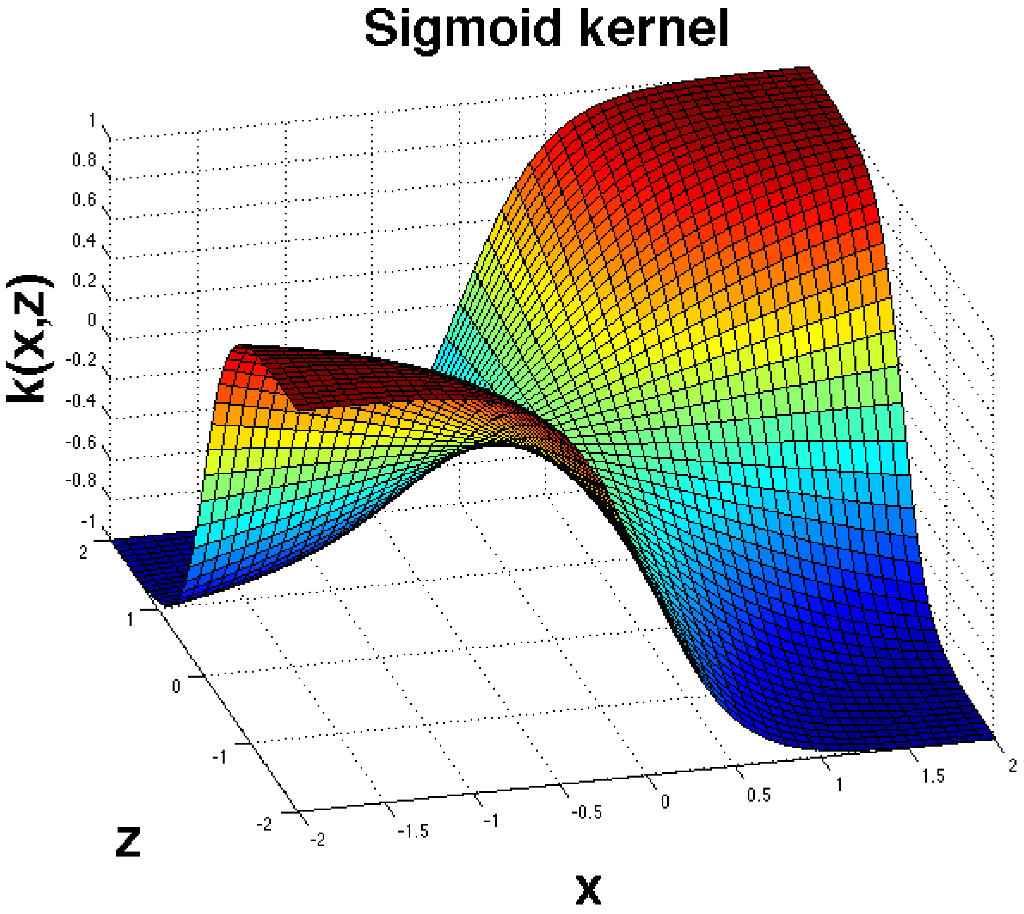
$$k(x, y) = \tanh(\alpha x^T y + c)$$



# Example for Kernel Functions

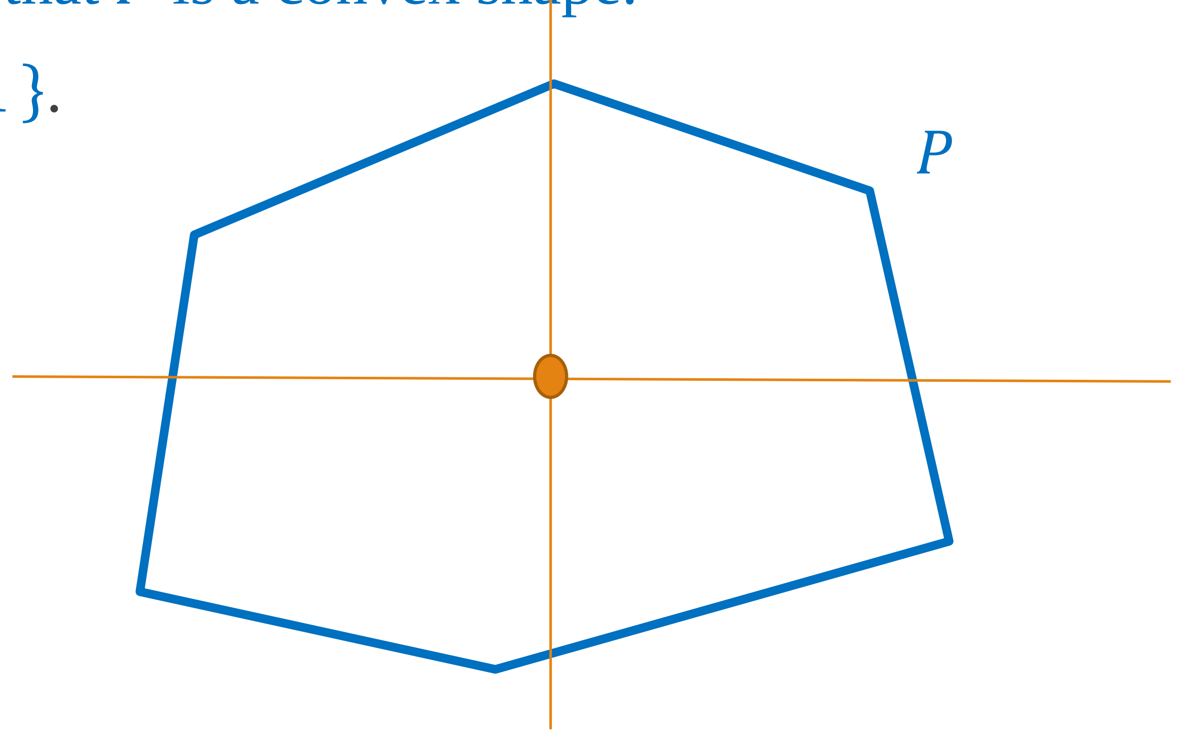
<u>Sigmoid Kernel:</u>

$$k(x, y) = \tanh(\alpha x^T y + c)$$



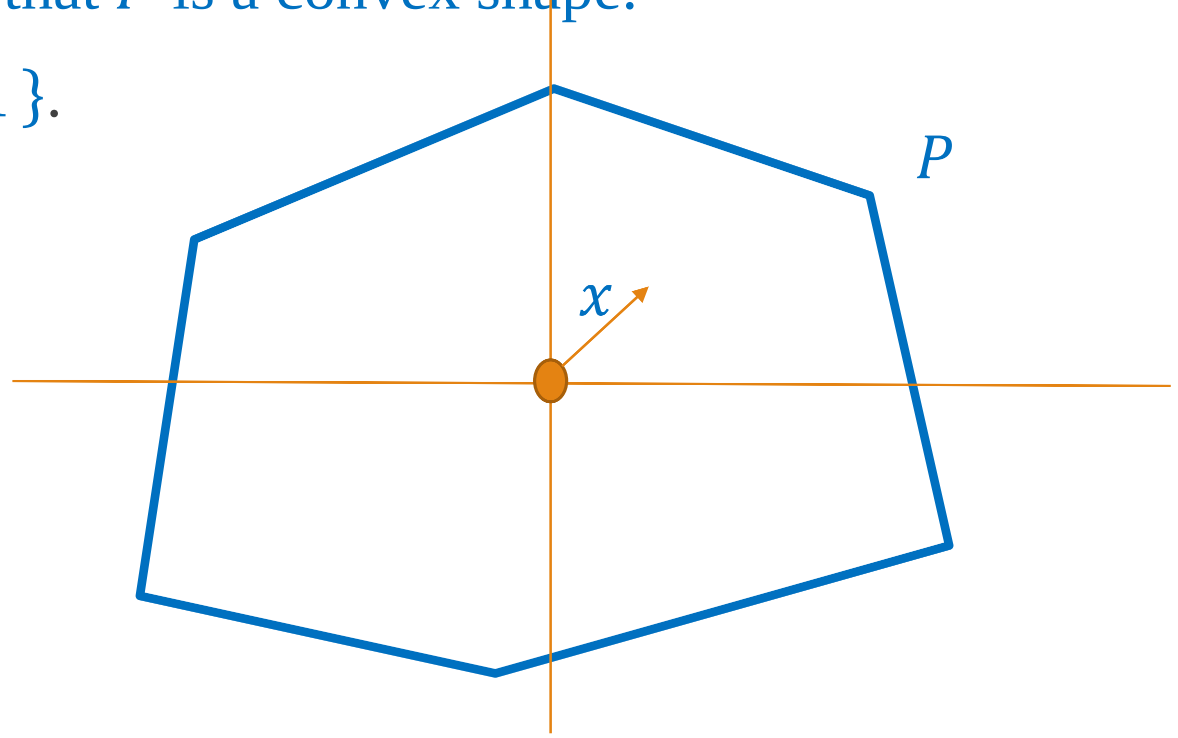
# Sensitivity for Convex Shapes

- Input: $P \subseteq R^d, |P| = n^d$ such that $P$ is a convex shape.
- Query space: $Q = \{ x \in R^d \mid ||x|| = 1 \}$.
- Cost function: $f(P, x) = \max_{p \in P} p^T x$



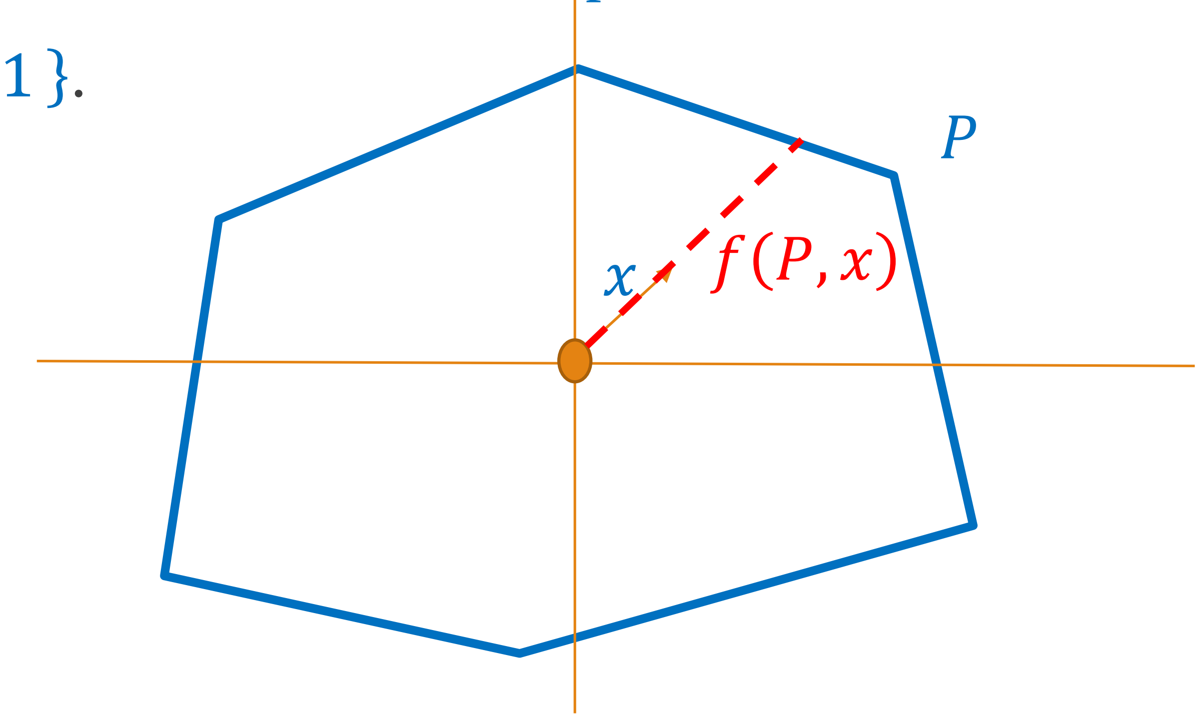
# Sensitivity for Convex Shapes

- Input: $P \subseteq R^d, |P| = n^d$ such that $P$ is a convex shape.
- Query space: $Q = \{ x \in R^d \mid ||x|| = 1 \}$.
- Cost function: $f(P, x) = \max_{p \in P} p^T x$



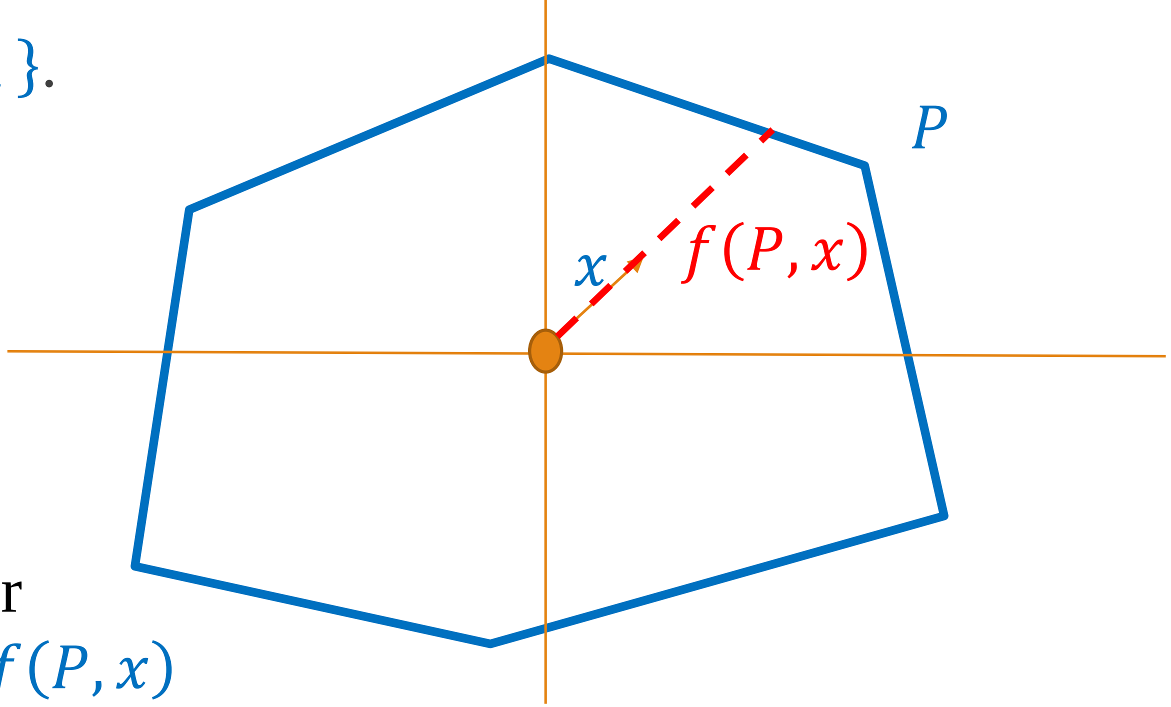
# Sensitivity for Convex Shapes

- Input: $P \subseteq R^d, |P| = n^d$ such that $P$ is a convex shape.
- Query space: $Q = \{ x \in R^d \mid \|x\| = 1 \}$.
- Cost function: $f(P, x) = \max_{p \in P} p^T x$



# Sensitivity for Convex Shapes

- Input: $P \subseteq R^d, |P| = n^d$ such that $P$ is a convex shape.
- Query space: $Q = \{ x \in R^d \mid \|x\| = 1 \}$.
- Cost function: $f(P, x) = \max_{p \in P} p^T x$

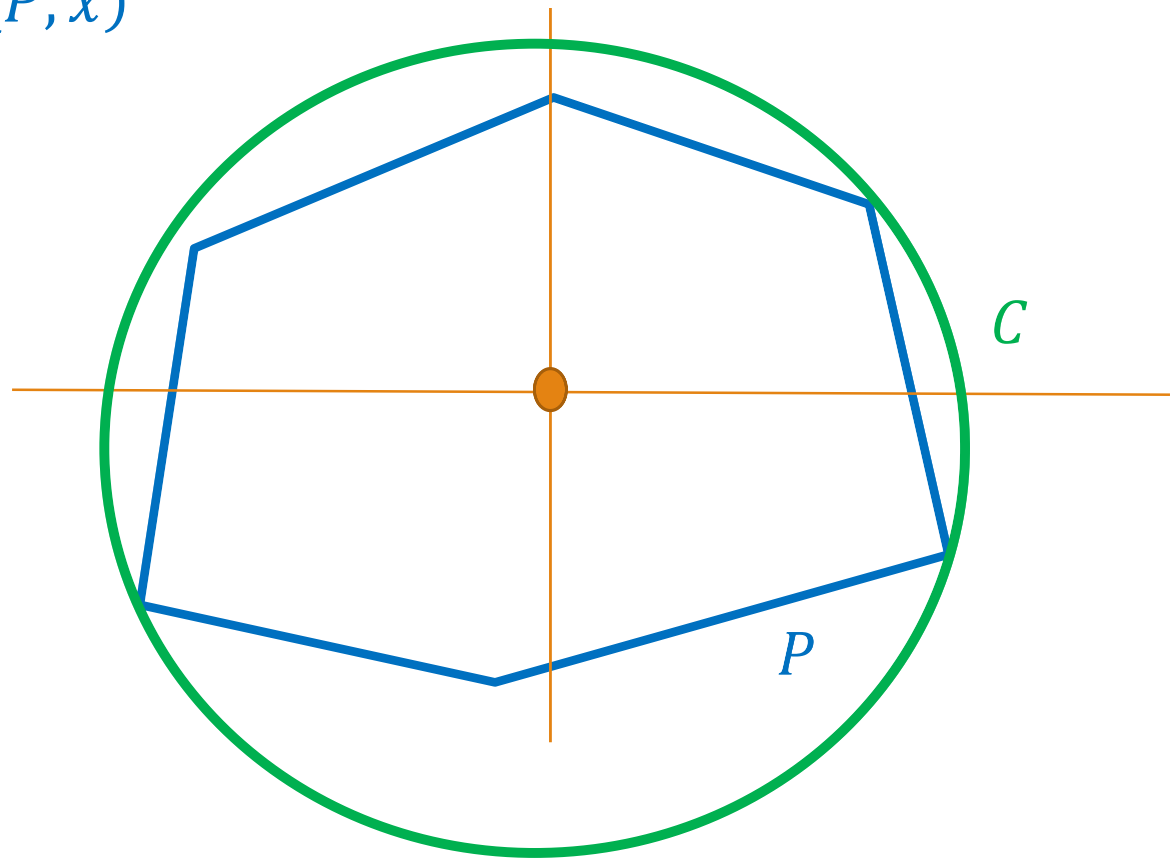



- Goal: Find another shape $C$ that can be represented by $O(d)$ vectors such that for every $x \in Q$: $f(P, x) \le f(C, x) \le \sqrt{d} \cdot f(P, x)$

# Sensitivity for Convex Shapes

- Goal: Find another shape $C$ that can be represented by $O(d)$ vectors such that for every $x \in Q$: $f(P,x) \leq f(C,x) \leq \alpha \cdot f(P,x)$

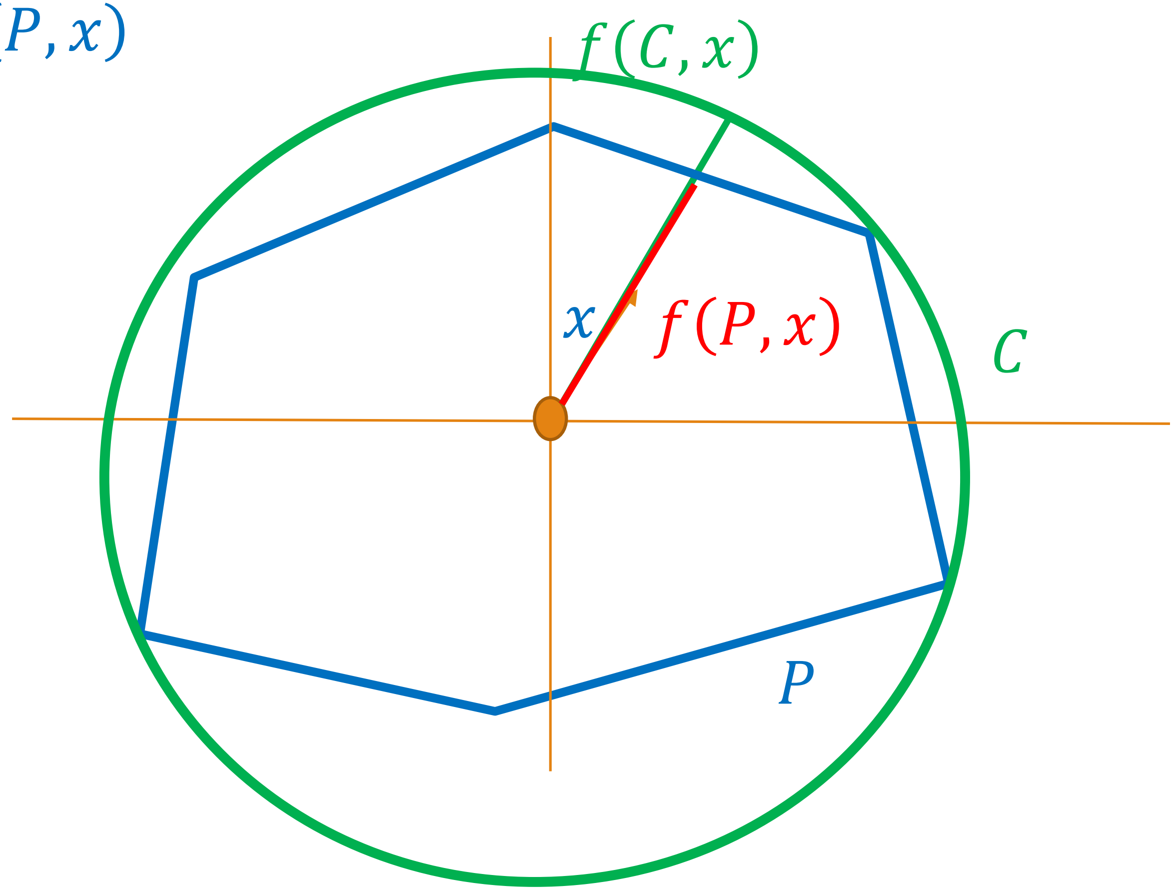
- Suggestion 1: Set $C$ to be the minimum enclosing circle of $P$.



# Sensitivity for Convex Shapes

- Goal: Find another shape $C$ that can be represented by $O(d)$ vectors such that for every $x \in Q$: $f(P, x) \leq f(C, x) \leq \alpha \cdot f(P, x)$

- Suggestion 1: Set $C$ to be the minimum enclosing circle of $P$.

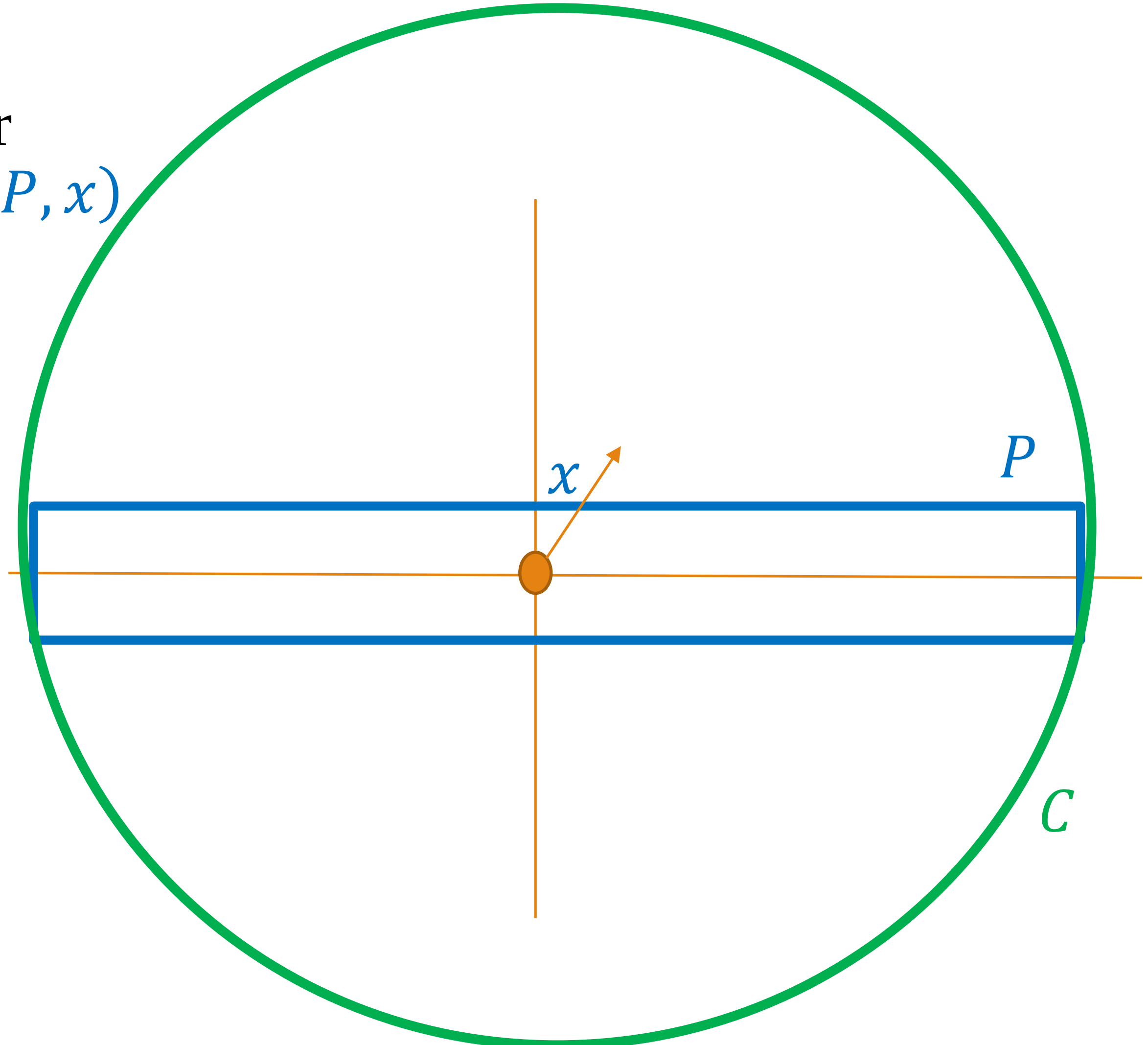


# Sensitivity for Convex Shapes

- Goal: Find another shape $C$ that can be represented by $O(d)$ vectors such that for every $x \in Q$: $f(P,x) \leq f(C,x) \leq \alpha \cdot f(P,x)$

- Suggestion 1: Set $C$ to be the minimum enclosing circle of $P$.

Bad example:

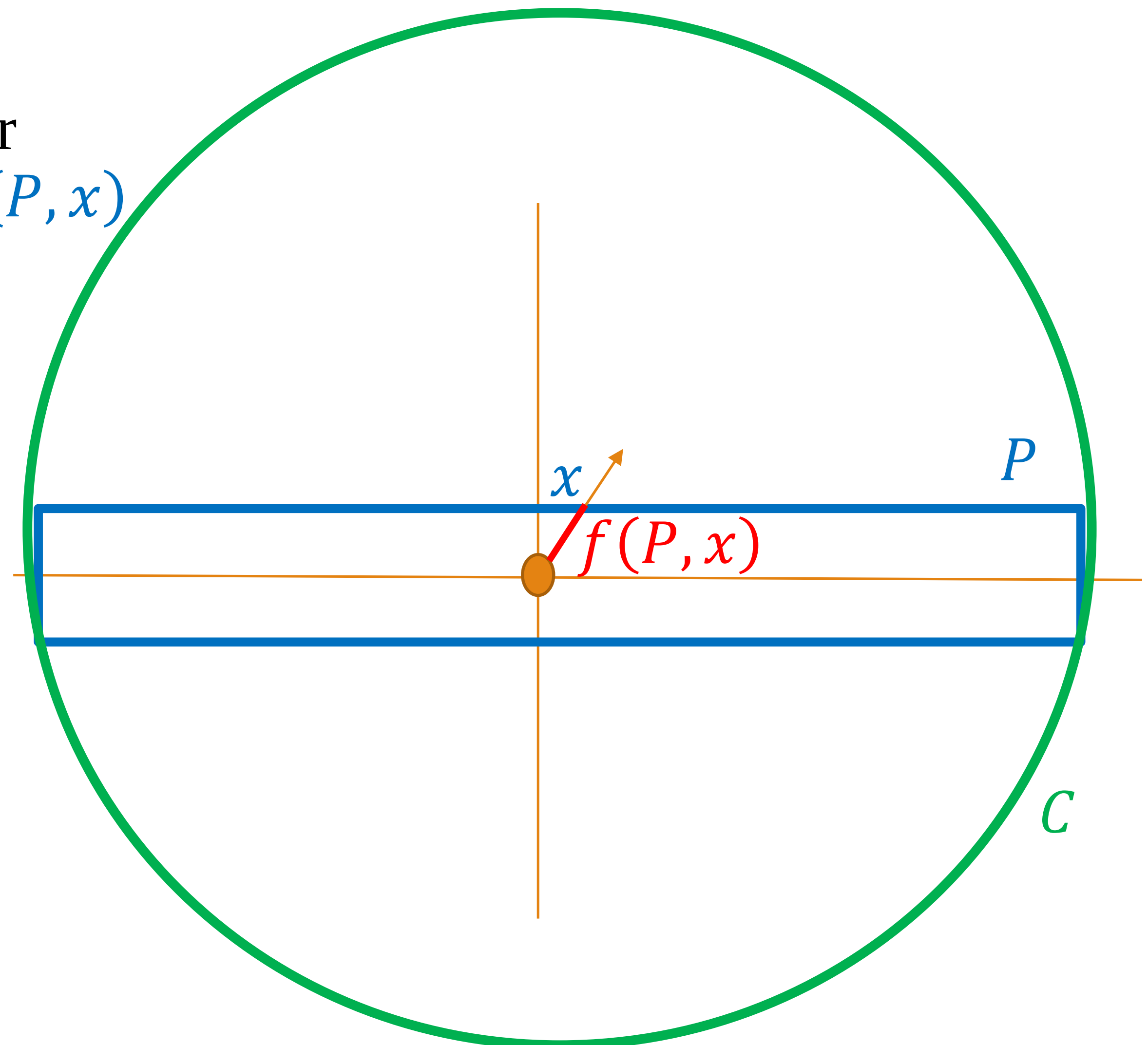


# Sensitivity for Convex Shapes

Bad example:

- Goal: Find another shape $C$ that can be represented by $O(d)$ vectors such that for every $x \in Q$: $f(P,x) \leq f(C,x) \leq \alpha \cdot f(P,x)$

- Suggestion 1: Set $C$ to be the minimum enclosing circle of $P$.



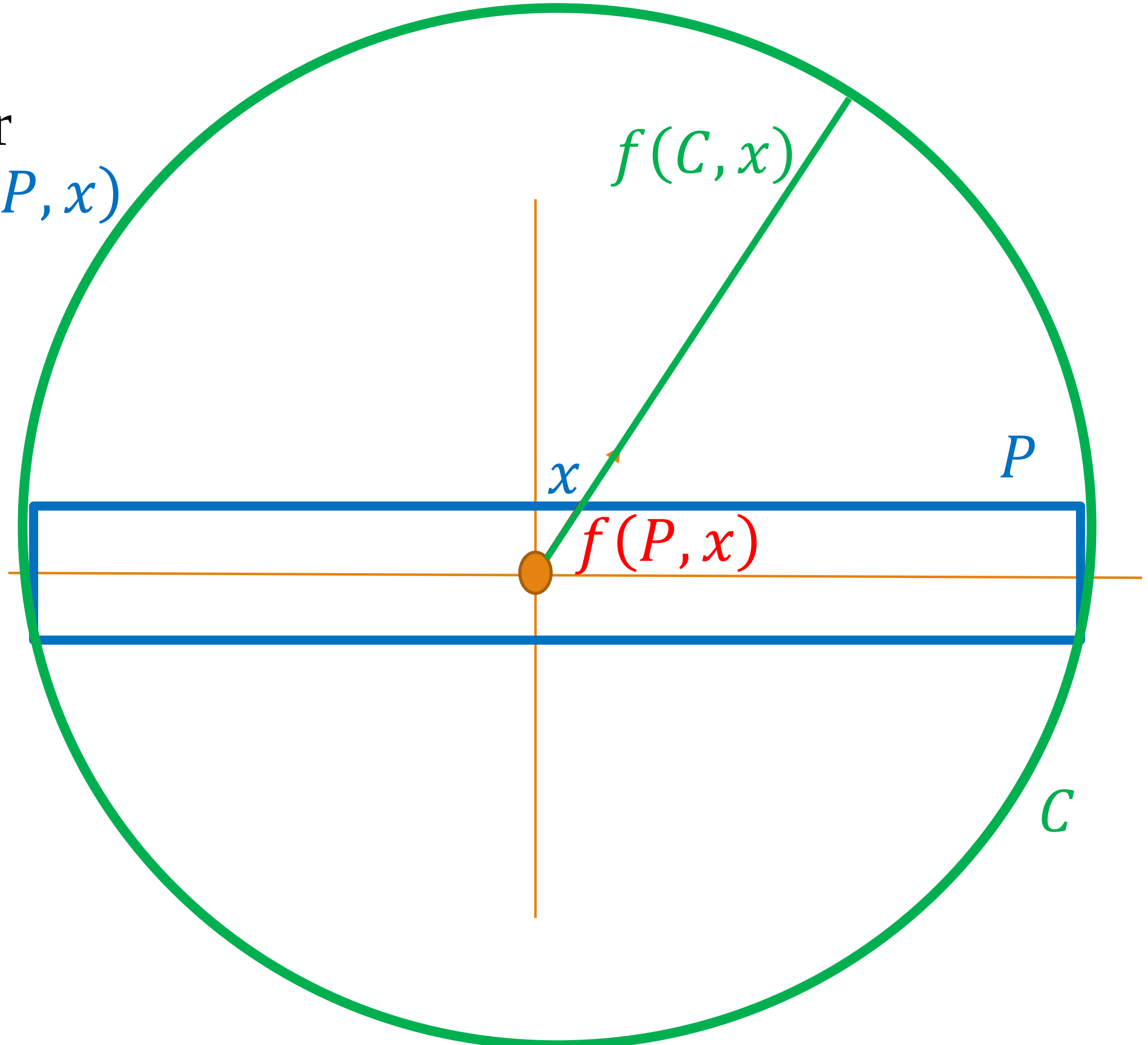
# Sensitivity for Convex Shapes

- Goal: Find another shape $C$ that can be represented by $O(d)$ vectors such that for every x $\in Q$: $f(P,x) \leq f(C,x) \leq \alpha \cdot f(P,x)$

- Suggestion 1: Set $C$ to be the minimum enclosing circle of $P$.

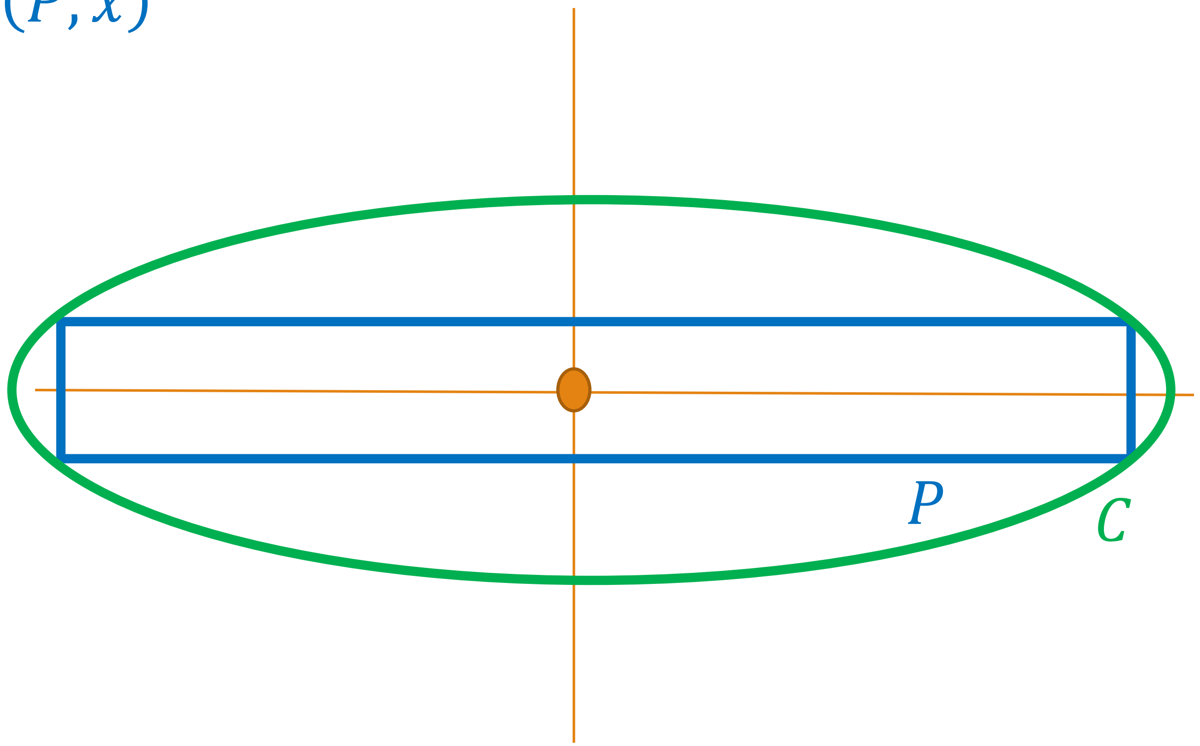
$\rightarrow$ Circle is not a good approximation since $\frac{f(C,x)}{f(P,x)} \rightarrow \infty$

Bad example:



# Sensitivity for Convex Shapes

- Goal: Find another shape $C$ that can be represented by $O(d)$ vectors such that for every $x \in Q$: $f(P,x) \leq f(C,x) \leq \alpha \cdot f(P,x)$
- Suggestion 2: Set $C$ to be the minimum enclosing ellipsoid of $P$.

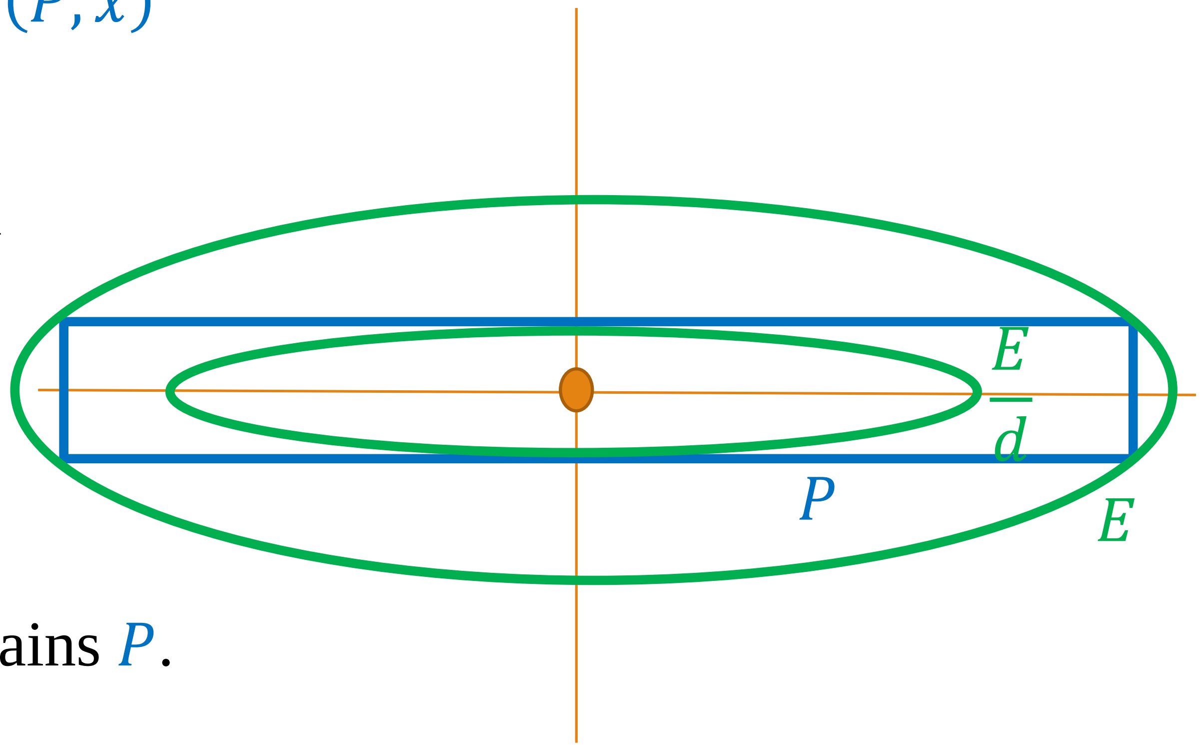


# Sensitivity for Convex Shapes

- Goal: Find another shape $C$ that can be represented by $O(d)$ vectors such that for every $x \in Q$: $f(P,x) \leq f(C,x) \leq \alpha \cdot f(P,x)$

- Suggestion 2: Set $C$ to be the minimum enclosing ellipsoid of $P$.

**Theorem:** **(John's Ellipsoid)**
Every convex shape $P$ contains an ellipsoid $\frac{E}{d}$ such that the ellipsoid $E$ contains $P$.

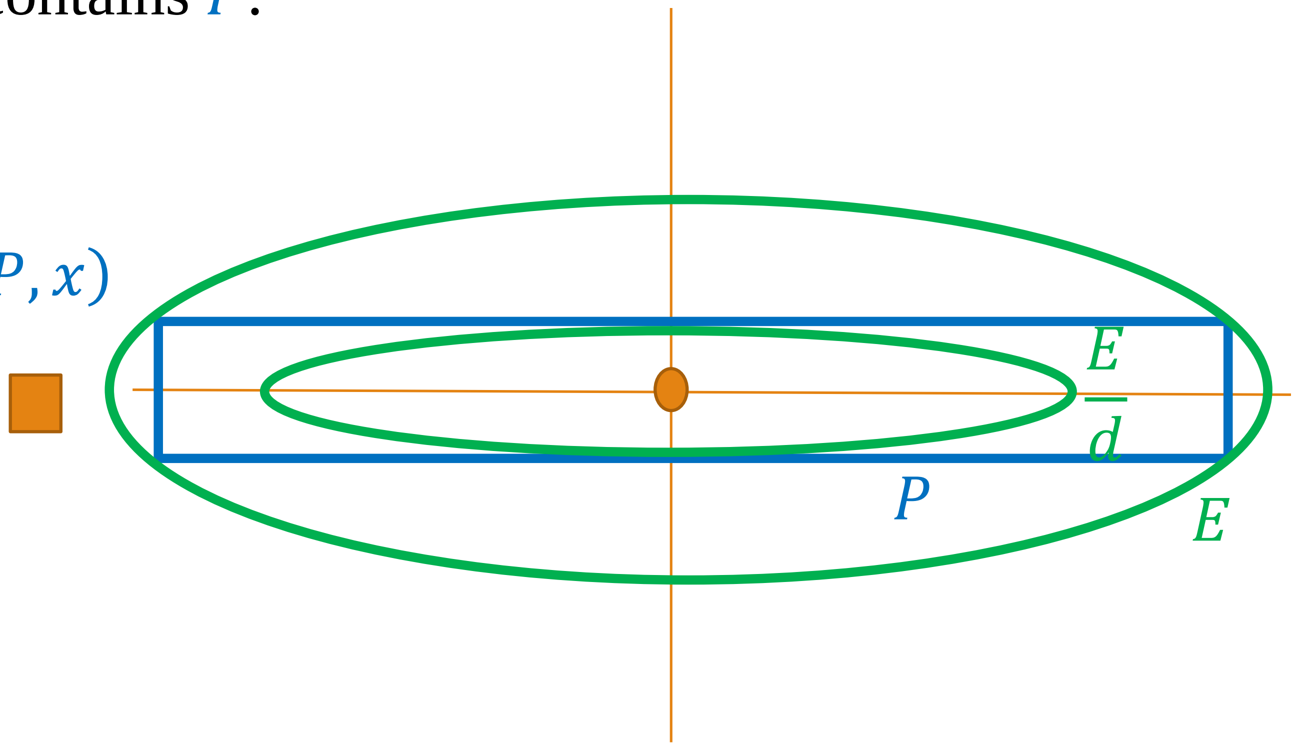


# Sensitivity for Convex Shapes

**Theorem:** **(John's Ellipsoid)**
Every convex shape $P$ contains an ellipsoid $\frac{E}{d}$ such that the ellipsoid $E$ contains $P$.

→ For every $x \in Q$:

$$f(P,x) \leq f(E,x) \leq d \cdot f\left(\frac{E}{d},x\right) \leq d \cdot f(P,x)$$



# Sensitivity for Convex Optimization

- Input: $P \in R^{n \times d}$ such that $P$ is a convex shape.
- Query space: $Q = \{\, x \in R^d \mid \|x\| = 1 \,\}$.
- Cost function: $k(p, x) = |px|, f(P, x) = \sum_{p \in P} k(p, x) = \|Px\|_1$

  $k(p, x) \sim g(|px|)$

# Sensitivity for Convex Optimization

- Input: $P \in R^{n \times d}$ such that $P$ is a convex shape.
- Query space: $Q = \{ x \in R^d \mid \|x\| = 1 \}$.
- Cost function: $k(p,x) = |px|, f(P,x) = \sum_{p \in P} k(p,x) = \|Px\|_1$

  $k(p,x) \sim g(|px|)$

- Notice that: $f(x) \sim \|Ex\| = \|DV^T x\|$

**Lemma:**

The sensitivity of a point $p \in P$ is at most

$$e_i = (0, \cdots, 0, \overset{i}{1}, 0, \cdots, 0)$$

$$\max_{x \in Q} \frac{k(p,x)}{f(P,x)} \leq \sum_{i=1}^{d} k(p, E^{-1} e_i)$$

# Sensitivity for Convex Optimization

**Lemma:**

The sensitivity of a point $p \in P$ is at most

$$\max_{x \in Q} \frac{k(p,x)}{f(P,x)} \leq \sum_{i=1}^{d} k(p, E^{-1} e_i)$$

**Proof:**

$$\frac{k(p,x)}{f(P,x)} \sim \frac{k(p,x)}{\|Ex\|} \sim k\left(p, \frac{x}{\|Ex\|}\right) = k(uE, E^{-1}, y) \sim g(|uy|) \leq g(\|u\|_2)$$

$$\leq g(\|u\|_1) = g\left(\sum_{i=1}^{d} |ue_i|\right) \sim \sum_{i=1}^{d} g(|ue_i|) \sim \sum_{i=1}^{d} k(uE, E^{-1} e_i)$$

$$= \sum_{i=1}^{d} k(p, E^{-1} e_i)$$

# Sensitivity for Convex Optimization

**Lemma:**

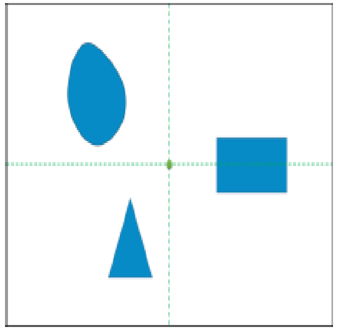
The sensitivity of a point $p \in P$ is at most

$$\max_{x \in Q} \frac{k(p,x)}{f(P,x)} \leq \sum_{i=1}^{d} k(p, E^{-1} e_i)$$

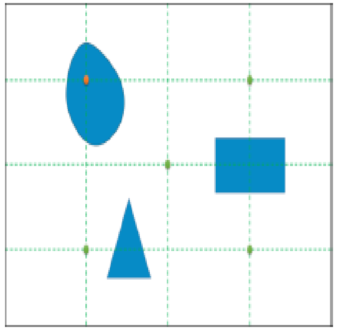
Hence, the total sensitivity is:

$$\sum_{p \in P} \sum_{i=1}^{d} k(p, E^{-1} e_i) = \sum_{i=1}^{d} \sum_{p \in P} k(p, E^{-1} e_i)$$

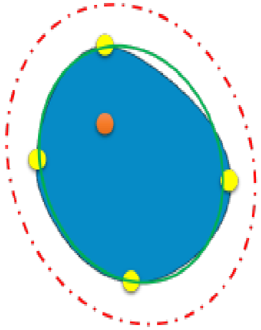
$$= \sum_{i=1}^{d} f(E^{-1}, e_i) \sim \sum_{i=1}^{d} \|E \cdot E^{-1}\, e_i\| \sim \sum_{i=1}^{d} \|\, e_i\| = d$$



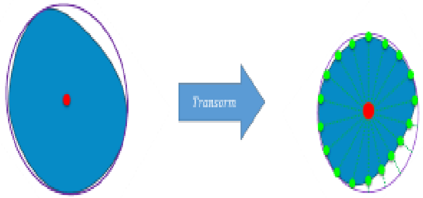
(a) Epsilon grid sampling; First iteration



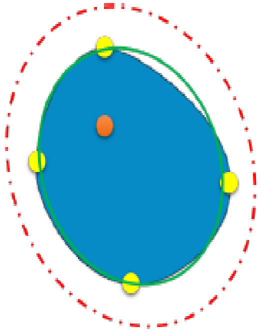
(b) Epsilon grid sampling; Second iteration



(c) $d^{2d}$ approximation to John Ellipsoid



(d) Applying "Epsilon Star" on the transform space



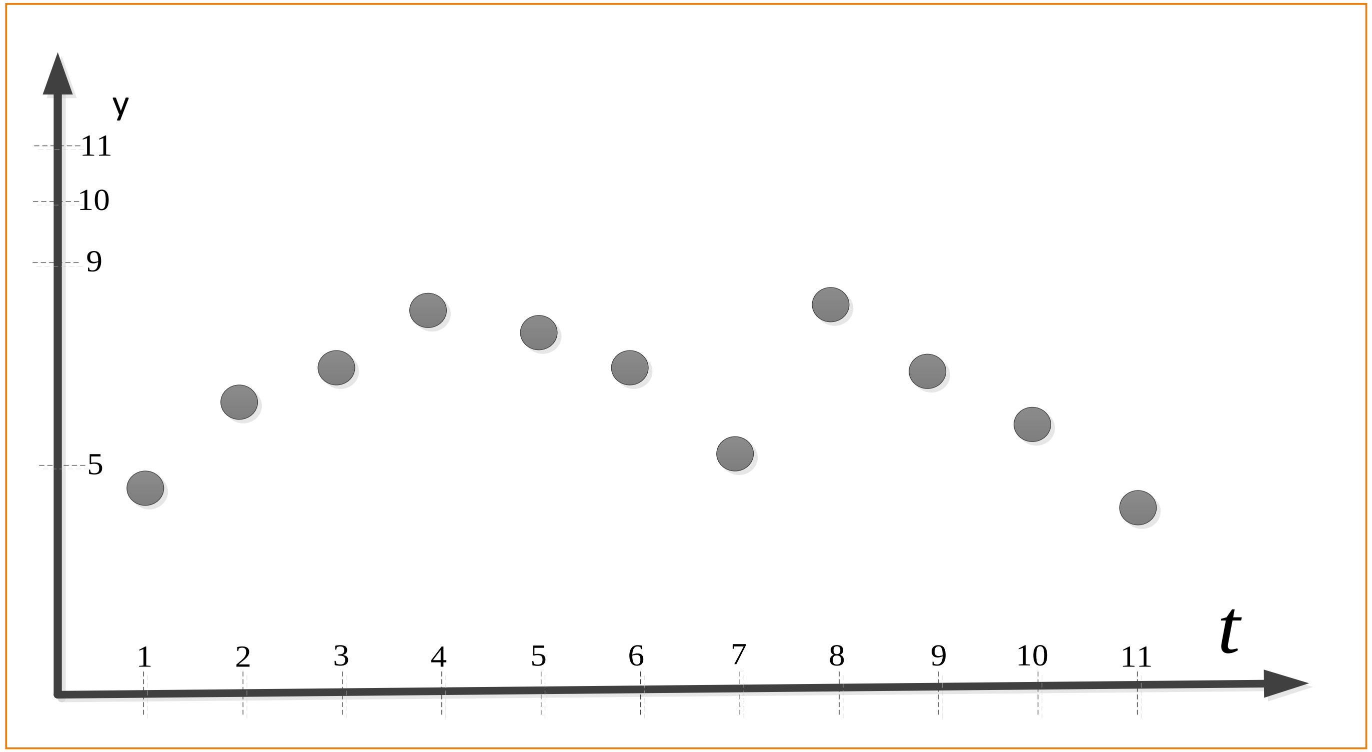
(e) $1 + \epsilon$ approximation to the real convex bodies

# $k$-Segment mean

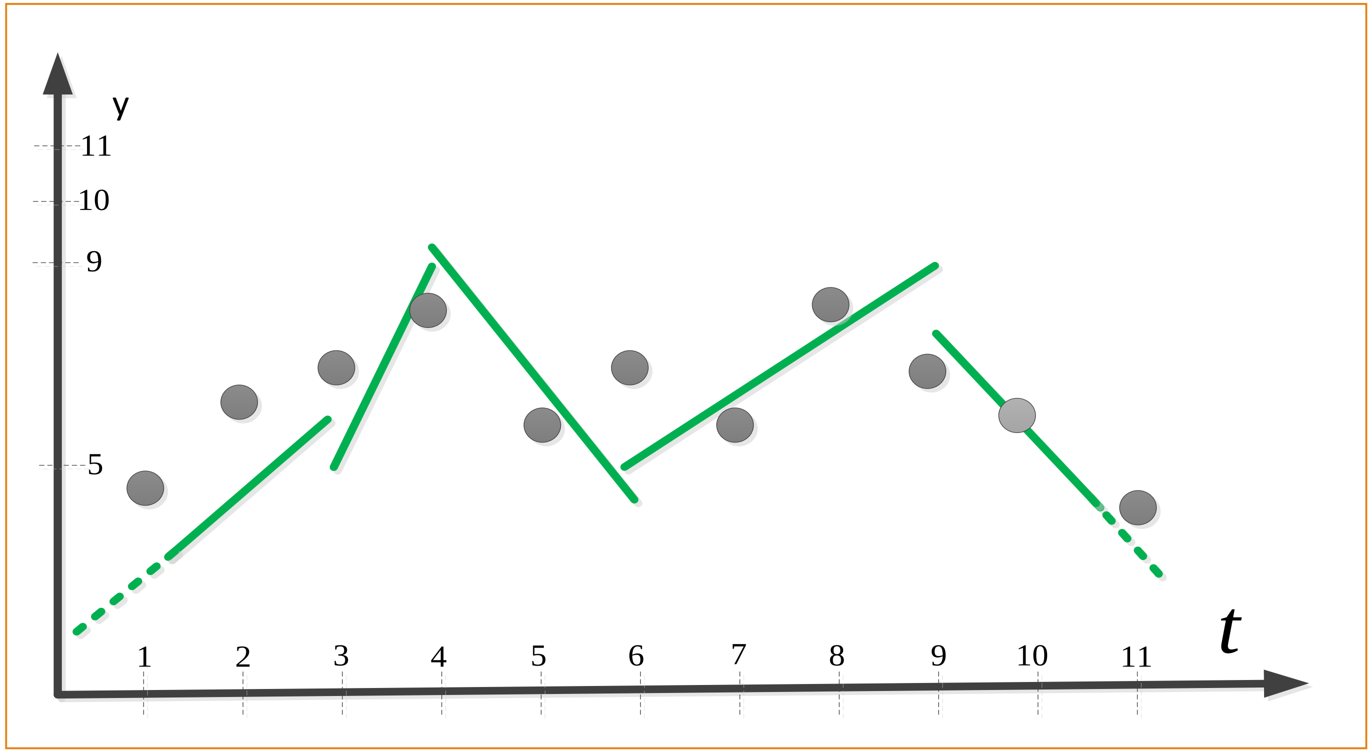
- Input: $P = \{(1, p_1), \ldots, (1, p_n) \mid p_i \in R^d\} \subseteq R^{d+1}$
- $k$-segment: $f: R \to R^d$
- Query space: $Q = \{f \mid f \; is \; a \; k - segment\}\}$
- Cost function: $cost(P, f) = \sum_{i=1}^{n} \|p_i - f(i)\|_2^2$

- $\mathrm{OPT} = \min_{f \in Q} cost(P, f)$
- k-segment mean $f^* = argmin_f \; cost(P, f)$

# $k$-Segment mean

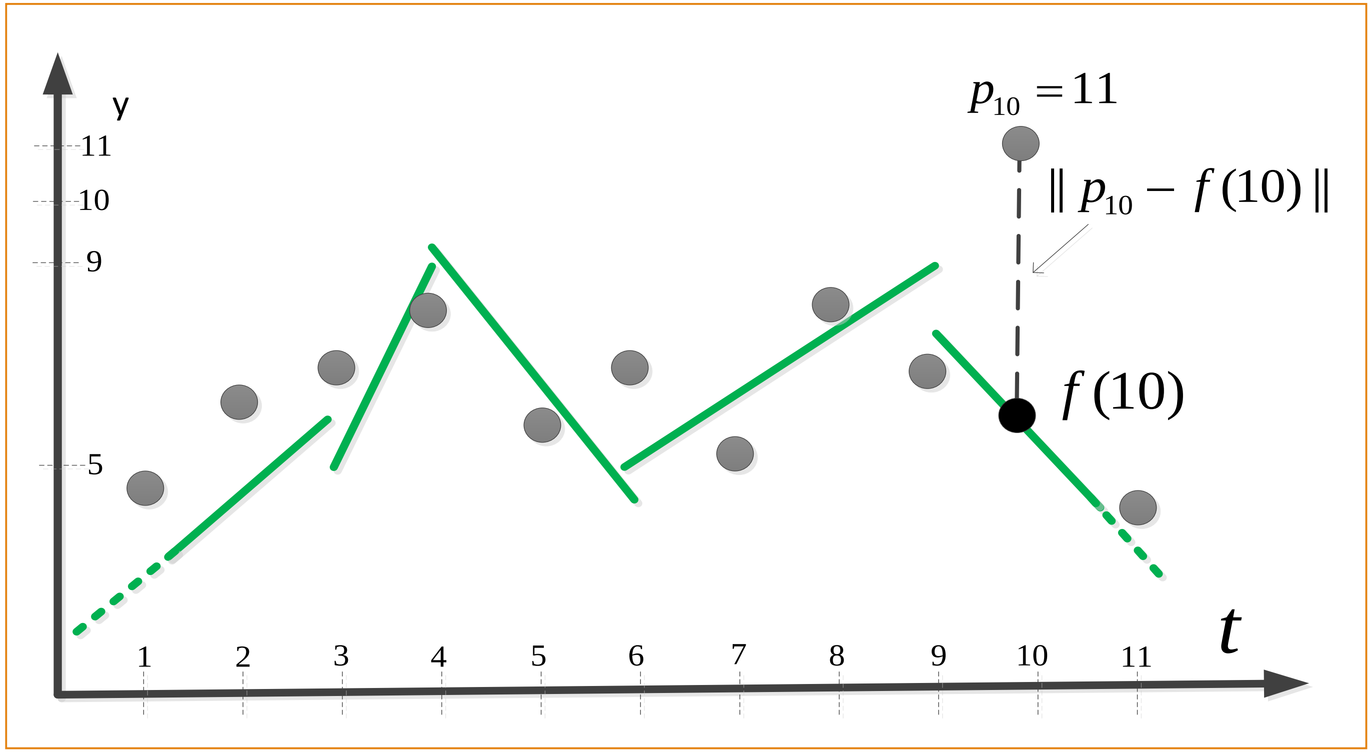


# $k$-Segment mean



# $k$-Segment mean

$$cost(P, f) = \sum_{i=1}^{n} \|p_i - f(i)\|_2^2$$



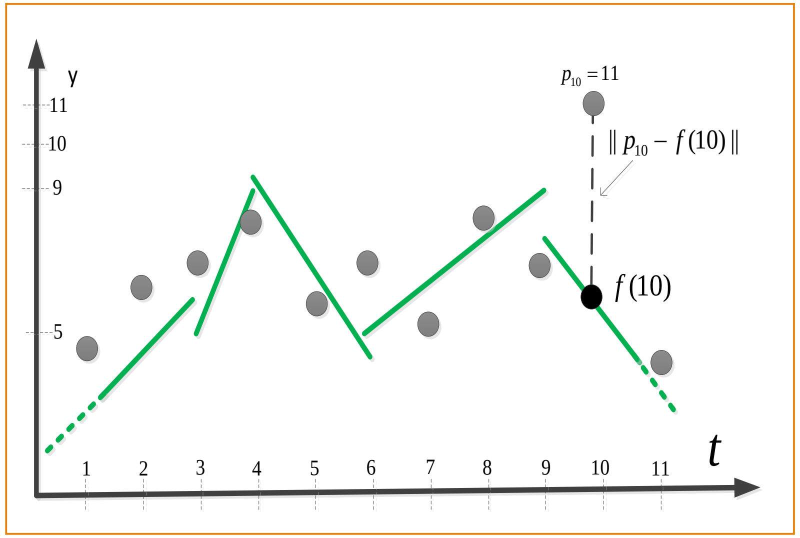
# Coreset for $k$-Segment mean

- Input: $P = \{(1, p_1), \ldots, (1, p_n) \mid p_i \in R^d\} \subseteq R^{d+1}$
- $k$-segment: $f: R \to R^d$
- Query space: $Q = \{f \mid f \text{ is a } k - \text{segment}\}$
- Cost function: $cost(P, f) = \sum_{i=1}^{n} \|p_i - f(i)\|_2^2$
- Output: $(\boldsymbol{C}, \boldsymbol{\omega})$ where $C \subseteq P,\ \omega: C \to R\ \ s.t.\ \forall f \in Q$:

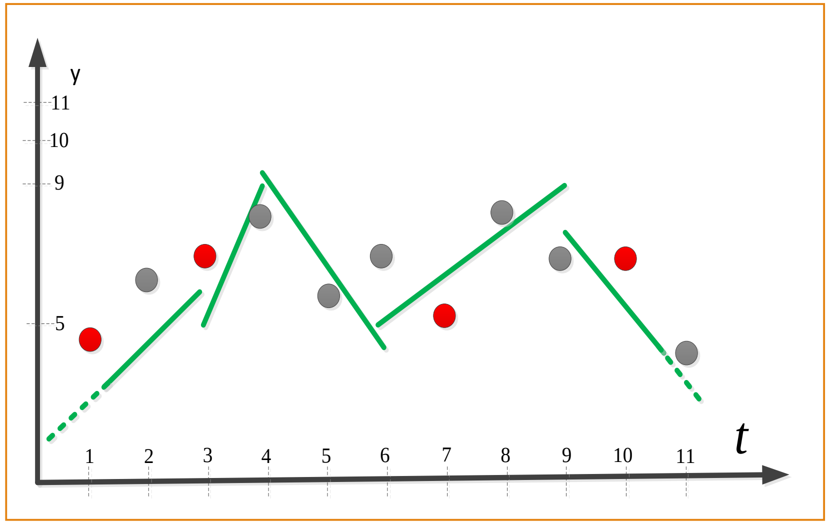
$$\left| \sum_{p_i \in P} \|p_i - f(i)\|_2^2 - \sum_{p_i \in C} \boldsymbol{\omega}(p_i) \cdot \|p_i - f(i)\|_2^2 \right| \leq \epsilon \cdot \sum_{p_i \in P} \|p_i - f(i)\|_2^2$$

# Coreset for $k$-Segment mean

$$\left|\sum_{p_i \in P} \|p_i - f(i)\|_2^2 - \sum_{p_i \in C} \omega(p_i) \cdot \|p_i - f(i)\|_2^2\right| \le \epsilon \cdot \sum_{p_i \in P} \|p_i - f(i)\|_2^2$$




$\sim$




$(1 \pm \epsilon)$

# Theorem [Feldman, Langberg, STOC'11]

Let $P, Q$ and $dist: PxQ \rightarrow R^+$.

A sample $C \subseteq P$, from the distribution

$$sensitivity(p) = \max_{q \in Q} \frac{dist(p,q)}{\sum_{p \in P} dist(p,q)},$$

is a coreset if

$$|C| \geq \frac{dimension\ of\ Q}{\epsilon^2} \cdot \sum_{p \in P} sensitivity(p)$$

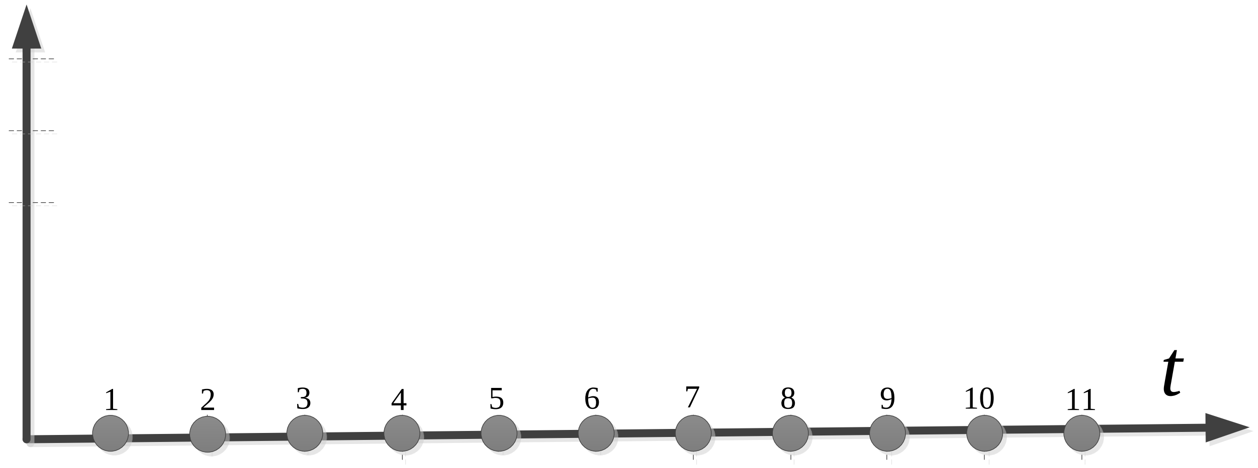
# Coreset for $k$-Segment mean

**Observation:**

No small coreset $C \subset P$ exists for $k$-segment queries

# Illustration for observation

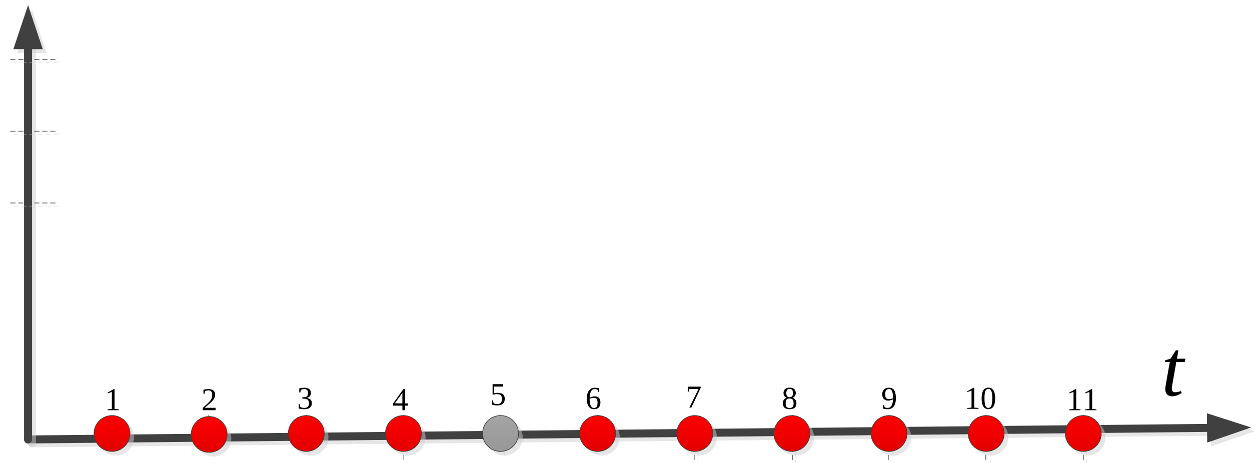
Input P: $n$ points on the $x$-axis



# Illustration for observation

Input P: $n$ points on the $x$-axis

Coreset $C$: all points except one

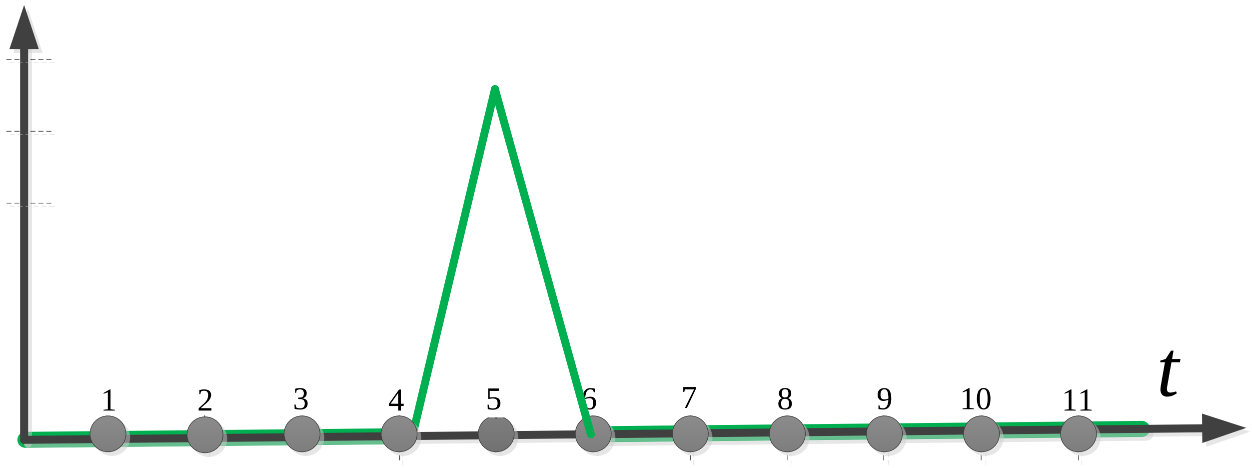


# Illustration for observation

Input P: $n$ points on the $x$-axis

Coreset $C$: all points except one

Query $f$: covers all except this one

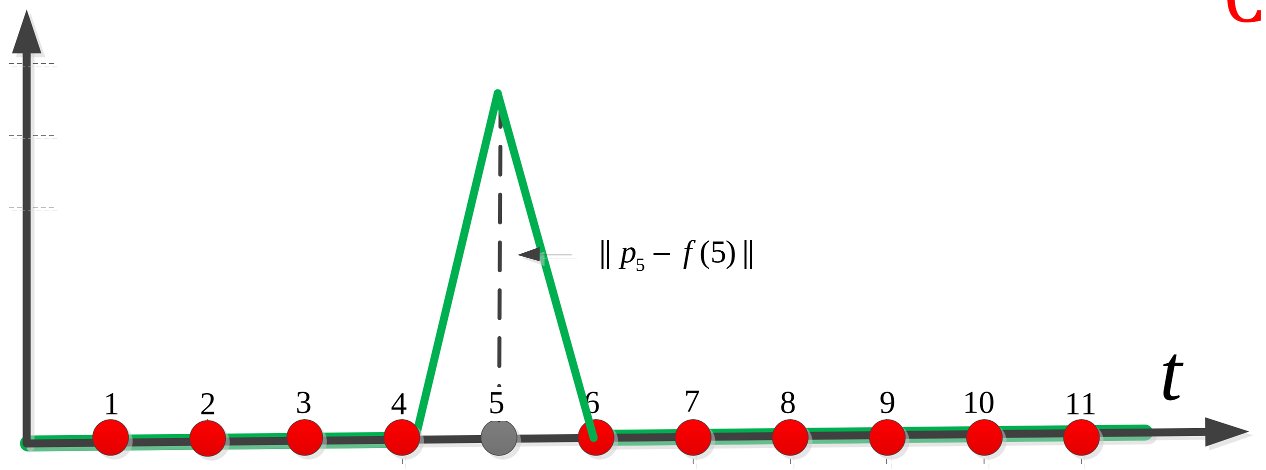


# Illustration for observation

Input P: $n$ points on the $x$-axis

Coreset $C$: all points except one

Query $f$: covers all except this one




$$\frac{\mathrm{Cost}(P, f) > 0}{\mathrm{Cost}(C, f) = 0}$$

Unbounded factor approximation

# Coreset for $k$-Segment mean

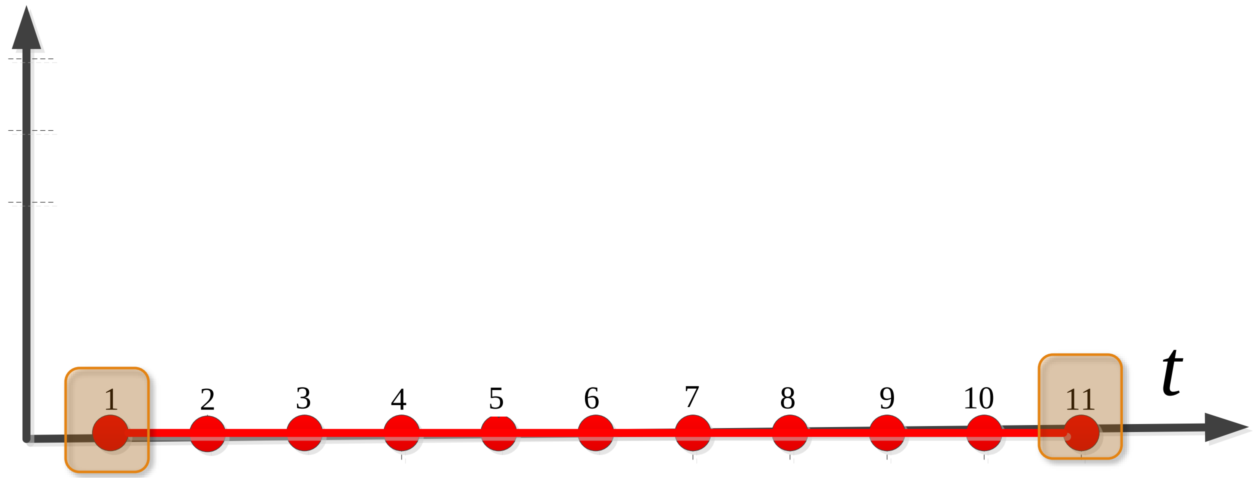
$$\forall p \in P: \quad sensitivity(p) = \max_{q \in Q} \frac{dist(p,q)}{\sum_{p \in P} dist(p,q)} = 1$$

$$\Rightarrow total\ sensitivity: \quad n$$

# Coreset for $k$-Segment mean

**Observation:**

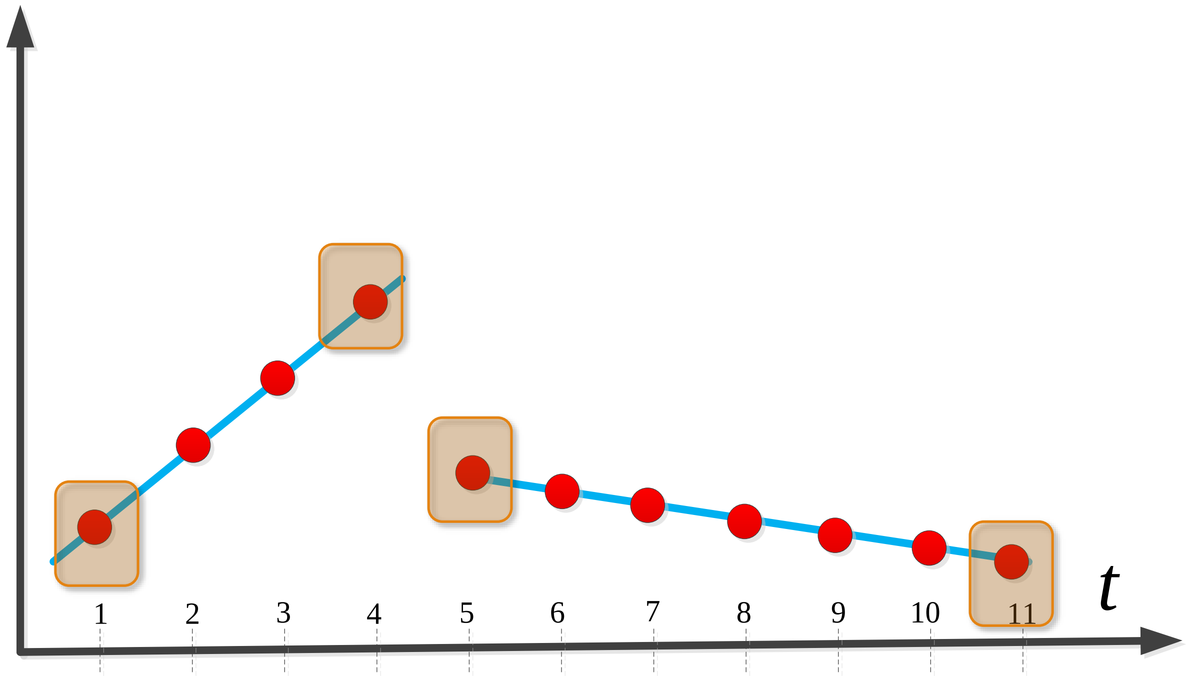
Points on a segment can be stored by the two indexes of their end-points



# Coreset for $k$-Segment mean

**Observation:**

Points on a segment can be stored by the two indexes of their end-points and the slope of the segment.



# Coreset for $k$-Segment mean (new definition)

- Input: $P = \{(1, p_1), \dots, (1, p_n) \mid p_i \in R^d\} \subseteq R^{d+1}$
- $k$-segment: $f: R \to R^d$
- Query space: $Q = \{\{f_1, \cdots, f_k\} \mid f_i \; is \; a \; segment\}\}$
- Cost function: $cost(P, f) = \sum_{i=1}^{n} \|p_i - f(i)\|_2^2$
- Output: $(\boldsymbol{C}, \boldsymbol{\omega})$ $where \; C \cancel{\subseteq} P, \; \omega: C \to R \;\; s.t. \; \forall f \in Q$:

$$\left| \sum_{p_i \in P} \|p_i - f(i)\|_2^2 - \sum_{p_i \in C} \boldsymbol{\omega}(p_i) \cdot \|p_i - f(i)\|_2^2 \right| \le \epsilon \cdot \sum_{p_i \in P} \|p_i - f(i)\|_2^2$$

# Coreset for $k$-Segment mean

**Input:** $(P, Q)$ and an $(\alpha, \beta)$-approximation $B$. Let $p' = \mathrm{proj}(\mathrm{p}, \mathrm{B})$.

**Goal:** To compute a set $(\boldsymbol{C}, \boldsymbol{\omega})$ such that for every $\forall f \in Q$:

$$\left| \sum_{p_i \in P} \|p_i - f(i)\|_2^2 - \sum_{p_i \in C} \boldsymbol{\omega}(p_i) \cdot \|p_i - f(i)\|_2^2 \right| \leq \epsilon \cdot \sum_{p_i \in P} \|p_i - f(i)\|_2^2$$

# Coreset for $k$-Segment mean

**Input:** $(P, Q)$ and an $(\alpha, \beta)$-approximation $B$. Let $p' = \mathrm{proj(p, B)}$.

**Goal:** To compute a set $(\boldsymbol{C}, \boldsymbol{\omega})$ such that for every $\forall f \in Q$:

$$\left|\sum_{p_i \in P} \|p_i - f(i)\|_2^2 - \sum_{p_i \in C} \boldsymbol{\omega}(p_i) \cdot \|p_i - f(i)\|_2^2\right| \leq \epsilon \cdot \sum_{p_i \in P} \|p_i - f(i)\|_2^2$$

$$\rightarrow \left|\sum_{p_i \in P} \|p_i - f(i)\|_2^2 - \sum_{p_i \in P} \|p_i' - f(i)\|_2^2 + \sum_{p_i \in P} \|p_i' - f(i)\|_2^2 - \sum_{p_i \in C} \boldsymbol{\omega}(p_i) \cdot \|p_i - f(i)\|_2^2\right|$$

$$\leq \epsilon \cdot \sum_{p_i \in P} \|p_i - f(i)\|_2^2$$

# Coreset for $k$-Segment mean

**Input:** $(P, Q)$ and an $(\alpha, \beta)$-approximation $B$. Let $p' = \mathrm{proj}(\mathrm{p}, \mathrm{B})$.

**Goal:** To compute a set $(\boldsymbol{C}, \boldsymbol{\omega})$ such that for every $\forall f \in Q$:

$$\left|\sum_{p_i \in P} \|p_i - f(i)\|_2^2 - \sum_{p_i \in C} \boldsymbol{\omega}(p_i) \cdot \|p_i - f(i)\|_2^2\right| \leq \epsilon \cdot \sum_{p_i \in P} \|p_i - f(i)\|_2^2$$

$$\to \left|\sum_{p_i \in P} \|p_i - f(i)\|_2^2 - \sum_{p_i \in P} \|p_i' - f(i)\|_2^2 + \sum_{p_i \in P} \|p_i' - f(i)\|_2^2 - \sum_{p_i \in C} \boldsymbol{\omega}(p_i) \cdot \|p_i - f(i)\|_2^2\right|$$

$$\leq \epsilon \cdot \sum_{p_i \in P} \|p_i - f(i)\|_2^2$$

$$\to \left|\frac{\sum_{p_i \in P} \|p_i - f(i)\|_2^2 - \sum_{p_i \in P} \|p_i' - f(i)\|_2^2 + \sum_{p_i \in P} \|p_i' - f(i)\|_2^2 - \sum_{p_i \in C} \boldsymbol{\omega}(p_i) \cdot \|p_i - f(i)\|_2^2}{\sum_{p_i \in P} \|p_i - f(i)\|_2^2}\right| \leq \epsilon$$

# Coreset for $k$-Segment mean

**Input:** $(P, Q)$ and an $(\alpha, \beta)$-approximation $B$. Let $p' = \mathrm{proj}(\mathrm{p}, \mathrm{B})$.

**Goal:** To compute a set $(\boldsymbol{C}, \boldsymbol{\omega})$ such that for every $\forall f \in Q$:

$$\left| \sum_{p_i \in P} \|p_i - f(i)\|_2^2 - \sum_{p_i \in C} \boldsymbol{\omega}(p_i) \cdot \|p_i - f(i)\|_2^2 \right| \leq \epsilon \cdot \sum_{p_i \in P} \|p_i - f(i)\|_2^2$$

$$\rightarrow \left| \sum_{p_i \in P} \|p_i - f(i)\|_2^2 - \sum_{p_i \in P} \|p_i' - f(i)\|_2^2 + \sum_{p_i \in P} \|p_i' - f(i)\|_2^2 - \sum_{p_i \in C} \boldsymbol{\omega}(p_i) \cdot \|p_i - f(i)\|_2^2 \right|$$

$$\leq \epsilon \cdot \sum_{p_i \in P} \|p_i - f(i)\|_2^2$$

✓

$$\rightarrow \left| \frac{\sum_{p_i \in P} \|p_i - f(i)\|_2^2 - \sum_{p_i \in P} \|p_i' - f(i)\|_2^2 + \sum_{p_i \in P} \|p_i' - f(i)\|_2^2 - \sum_{p_i \in C} \boldsymbol{\omega}(p_i) \cdot \|p_i - f(i)\|_2^2}{\sum_{p_i \in P} \|p_i - f(i)\|_2^2} \right| \leq \epsilon$$

Add the projections to the coreset $C$

# Coreset for $k$-Segment mean

**Input:** $(P, Q)$ and an $(\alpha, \beta)$-approximation $B$. Let $p' = \text{proj}(\text{p}, \text{B})$.

**Goal:** To compute a set $(\boldsymbol{C}, \boldsymbol{\omega})$ such that for every $\forall f \in Q$:

$$\left|\sum_{p_i \in P} \|p_i - f(i)\|_2^2 - \sum_{p_i \in C} \boldsymbol{\omega}(p_i) \cdot \|p_i - f(i)\|_2^2\right| \leq \epsilon \cdot \sum_{p_i \in P} \|p_i - f(i)\|_2^2$$

$$\rightarrow \left|\sum_{p_i \in P} \|p_i - f(i)\|_2^2 - \sum_{p_i \in P} \|p_i' - f(i)\|_2^2 + \sum_{p_i \in P} \|p_i' - f(i)\|_2^2 - \sum_{p_i \in C} \boldsymbol{\omega}(p_i) \cdot \|p_i - f(i)\|_2^2\right|$$

$$\leq \epsilon \cdot \sum_{p_i \in P} \|p_i - f(i)\|_2^2 \quad ✓$$

$$\rightarrow \left|\frac{\sum_{p_i \in P} \|p_i - f(i)\|_2^2 - \sum_{p_i \in P} \|p_i' - f(i)\|_2^2 + \sum_{p_i \in P} \|p_i' - f(i)\|_2^2 - \sum_{p_i \in C} \boldsymbol{\omega}(p_i) \cdot \|p_i - f(i)\|_2^2}{\sum_{p_i \in P} \|p_i - f(i)\|_2^2}\right| \leq \epsilon$$

Bound this term's sensitivity similar to $k$-means

# Theorem [Feldman, Sung, Rus, GIS'12]

For every discrete signal of $n$ points in $R^d$, there is a coreset of space $O\left(\frac{k}{\epsilon^2}\right)$ that can be computed in the big data model.