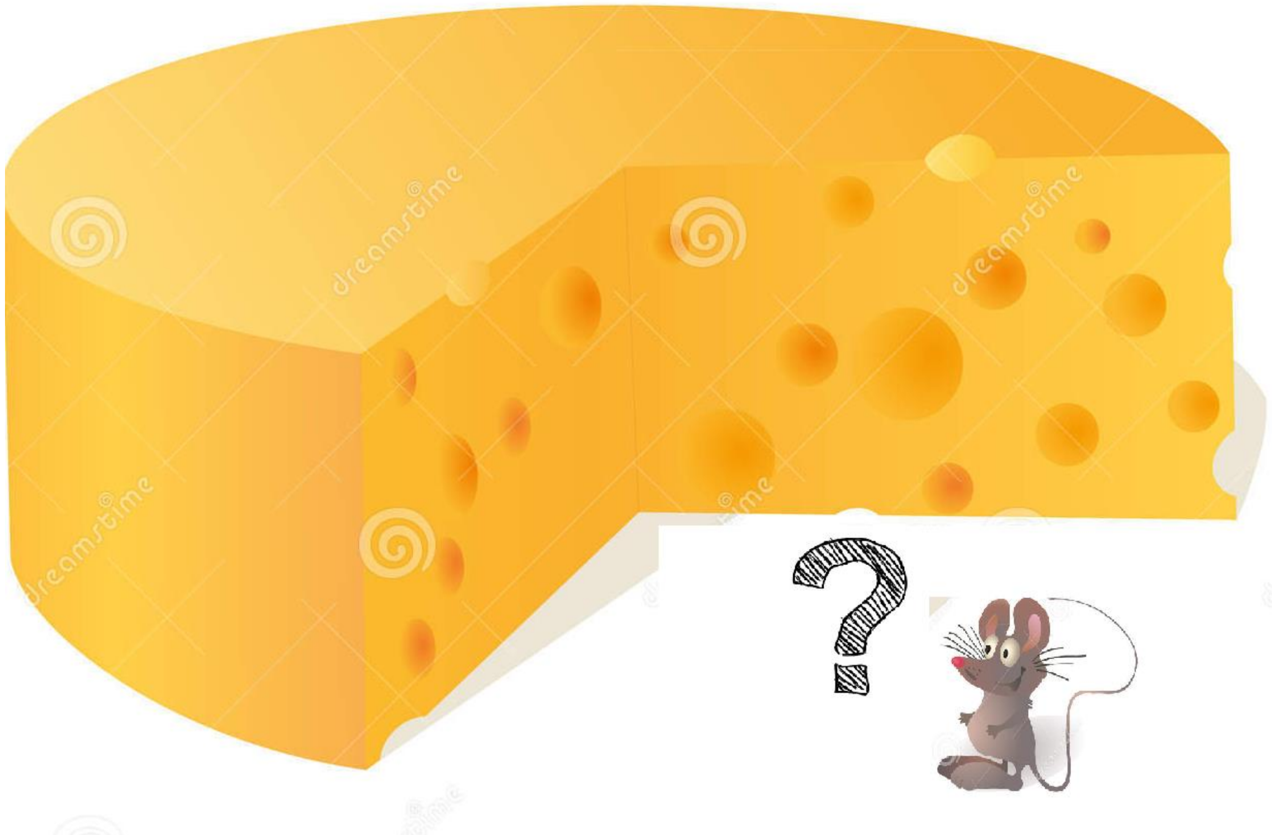
Big Data Class



LECTURER: DAN FELDMAN

TEACHING ASSISTANTS:

IBRAHIM JUBRAN

ALAA MAALOUF

Reminder, why do we need sensitivity

Query space
(P, ω, Q, f)

Coreset

Coreset
$(C, \mu), C \subseteq P, |C| \ll |P|$

Sample a subset of "important" points.

Query space
(P, ω, Q, f)

Compute
"importance"/ "sensitivity"
for each point

Sensitivity
$S: P \times \omega \to [0, \infty)$ s.t.

$$S(p) \geq \max_{q \in Q} \frac{\omega(p) \cdot f(p, q)}{\sum_{p' \in P} \omega(p') \cdot f(p', q)}$$

We will now focus on this block

Helps compute the importance for each input points by a reduction to a simpler problem

Query space
(P, ω, Q, f)

Rough approximation for the optimal solution
(α, β)-approximation

A candidate solution to the problem with provable guarantees.

Sensitivity

- Let (P, Q, w, f) be a query space, where $f: P \times Q \to [0, \infty)$.
  For every $p' \in P$. The *sensitivity* $\sigma(p')$ of p' is defined as :

$$\sigma(p') \coloneqq \sup \frac{w(p')f(p', q)}{\sum_{p \in P} w(p)(p, q)}$$

where the *sup* is over every $q \in Q$ with $\sum_{p \in P} w(p) f(p, q) > 0$.

- The *total sensitivity* of P is

$$G(P) \coloneqq \sum_{p \in P} \sigma(p)$$

Sensitivity

- Let (P, Q, w, f) be a query space, where $f: P \times Q \to [0, \infty)$. For every $p' \in P$. The *sensitivity* $\sigma(p')$ of p' is defined as :

$$\sigma(p') \coloneqq \sup \frac{w(p')f(p', q)}{\sum_{p \in P} w(p)(p, q)}$$

where the *sup* is over every $q \in Q$ with $\sum_{p \in P} w(p) f(p, q) > 0$.

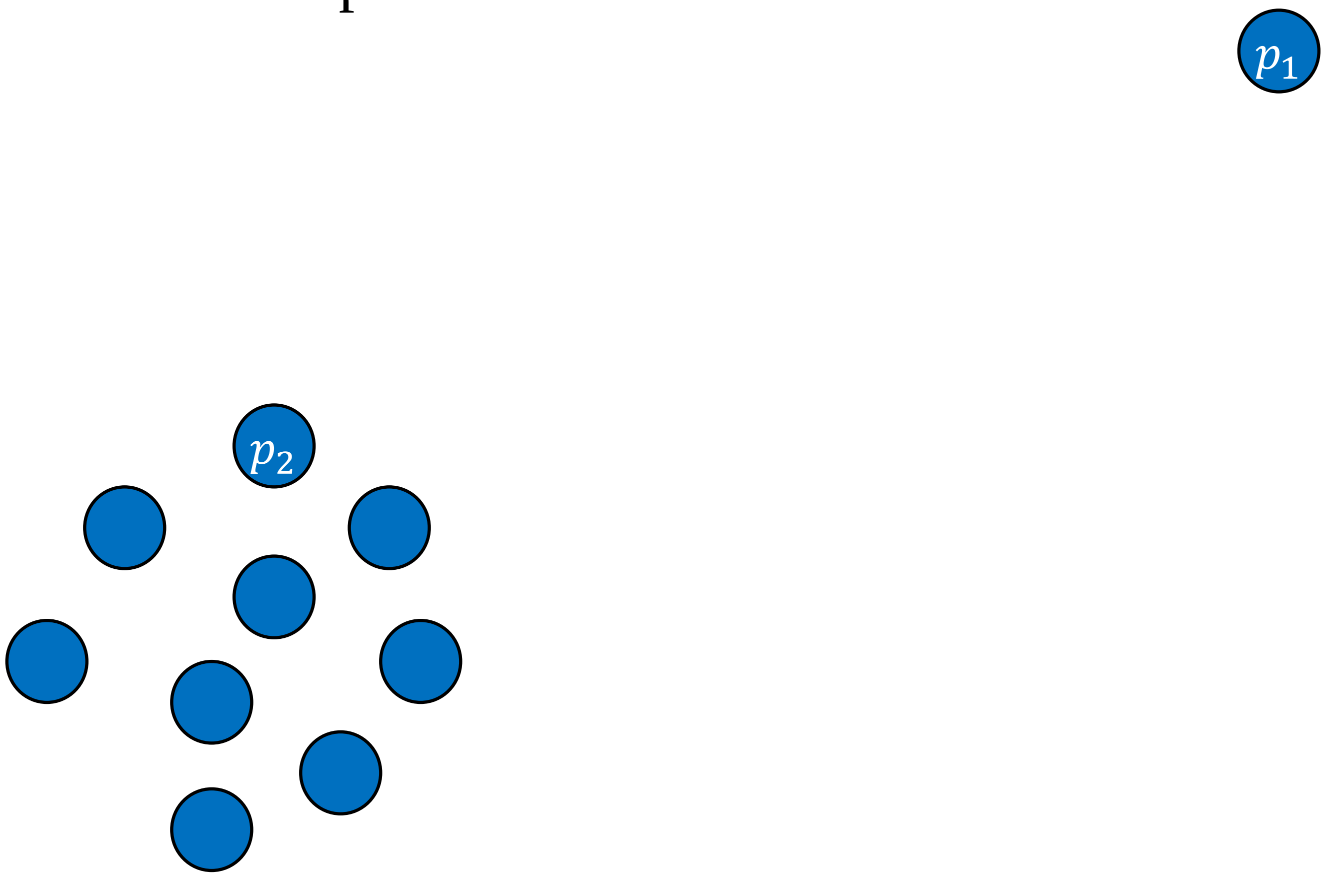
- The *total sensitivity* of P is

$$G(P) \coloneqq \sum_{p \in P} \sigma(p)$$

The sensitivity of a function (point) measures how influential that function (point) is on the optimization problem.

Sensitivity intuition

Consider the 1-median problem.

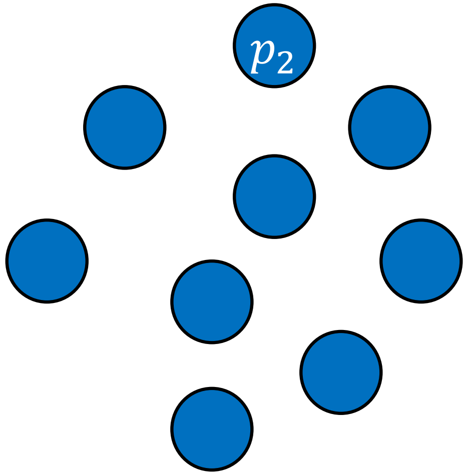


Sensitivity intuition

$\sigma(p_1)$ should be large

Consider the 1-median problem.

$\sigma(p_2)$ should be small

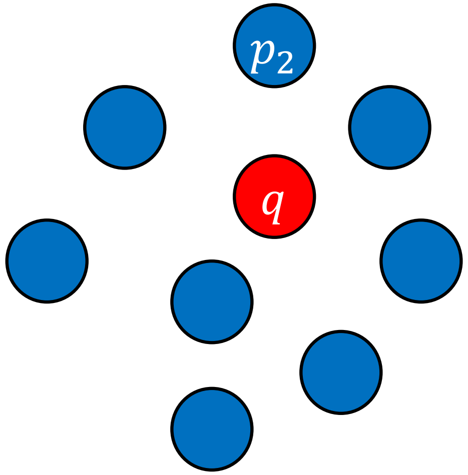


Sensitivity intuition

Consider the 1-median problem.

$\sigma(p_1)$ should be large

$\sigma(p_2)$ should be small



$$\sigma(p_1) \geq \frac{f(p_1, q)}{\sum_{p \in P} f(p, q)} \to 1$$

$$\sigma(p_2) = sup \frac{f(p_2, q)}{\sum_{p \in P} f(p, q)} \sim \frac{f(p_2, q)}{n \cdot f(p_2, q)} \to 0$$

Coreset for the Threshold Problem

Let $P \subseteq R$ be a set of n points.

Coreset for the Threshold Problem

Let $P \subseteq R$ be a set of n points.

Query: Threshold x (number).
Cost function: For every $p \in P$: $f(p, x) = \mathbf{1}(p \geq x)$
Output: $\sum_{p \in P} f(p, x)$

Coreset for the Threshold Problem

Let $P \subseteq R$ be a set of n points.

Query: Threshold x.
Cost function: For every $p \in P$: $f(p, x) = \mathbf{1}(p \geq x)$
Output: $\sum_{p \in P} f(p, x)$

$$s(p) = \max_{x \in R} \frac{f(p, x)}{\sum_{p' \in P} f(p', x)}$$

Coreset for the Threshold Problem

Let $P \subseteq R$ be a set of n points.

Query: Threshold x.
Cost function: For every $p \in P$: $f(p,x) = \mathbf{1}(p \geq x)$
Output: $\sum_{p \in P} f(p,x)$

$$s(p) = \max_{x \in R} \frac{f(p,x)}{\sum_{p' \in P} f(p',x)} = \frac{1}{\sum_{p' \in P} \mathbf{1}(p' \geq p)}$$

Coreset for the Threshold Problem

Let $P \subseteq R$ be a set of n points.

Query: Threshold x.
Cost function: For every $p \in P$: $f(p, x) = \mathbf{1}(p \geq x)$
Output: $\sum_{p \in P} f(p, x)$

$$\sum_{p \in P} s(p) = \sum_{i \in [n]} \frac{1}{i} = \ln(n) = O(\log n)$$

Coreset for the Threshold Problem (2)

Let $P \subseteq R$ be a set of n points.

Query: Threshold x right or left ($r = 1 / r = 0$).

Cost function: For every $p \in P$: $f(p, x, r) = \begin{cases} \mathbf{1}(p \geq x) & \text{if r} == 1 \\ \mathbf{1}(p \leq x) & if\ r == 0 \end{cases}$

Output: $\sum_{p \in P} f(p, x, r)$

Coreset for the Threshold Problem (2)

Let $P \subseteq R$ be a set of n points.

Query: Threshold x right or left ($r = 1/\ r = 0$).

Cost function: For every $p \in P$: $f(p, x, r) = \begin{cases} \mathbf{1}(p \geq x) & \text{if r == 1} \\ \mathbf{1}(p \leq x) & if\ r == 0 \end{cases}$

Output: $\sum_{p \in P} f(p, x, r)$

$$s(p) = \max_{x \in R, r \in \{0,1\}} \frac{f(p, x, r)}{\sum_{p' \in P} f(p', x, r)}$$

Coreset for the Threshold Problem (2)

Let $P \subseteq R$ be a set of n points.

Query: Threshold x right or left ($r = 1 / r = 0$).

Cost function: For every $p \in P$: $f(p, x, r) = \begin{cases} \mathbf{1}(p \geq x) & \text{if r == 1} \\ \mathbf{1}(p \leq x) & if\ r == 0 \end{cases}$

Output: $\sum_{p \in P} f(p, x, r)$

$$s(p) = \max_{x \in R, r \in \{0,1\}} \frac{f(p, x, r)}{\sum_{p' \in P} f(p', x, r)} \leq \max_{x \in R} \frac{f(p, x, 1)}{\sum_{p' \in P} f(p', x, 1)} + \max_{x \in R} \frac{f(p, x, 0)}{\sum_{p' \in P} f(p', x, 0)}$$

Coreset for the Threshold Problem (2)

Let $P \subseteq R$ be a set of n points.

Query: Threshold x right or left ($r = 1/\ r = 0$).

Cost function: For every $p \in P$: $f(p,x,r) = \begin{cases} \mathbf{1}(p \geq x) & \text{if r == 1} \\ \mathbf{1}(p \leq x) & if\ r == 0 \end{cases}$

Output: $\sum_{p \in P} f(p,x,r)$

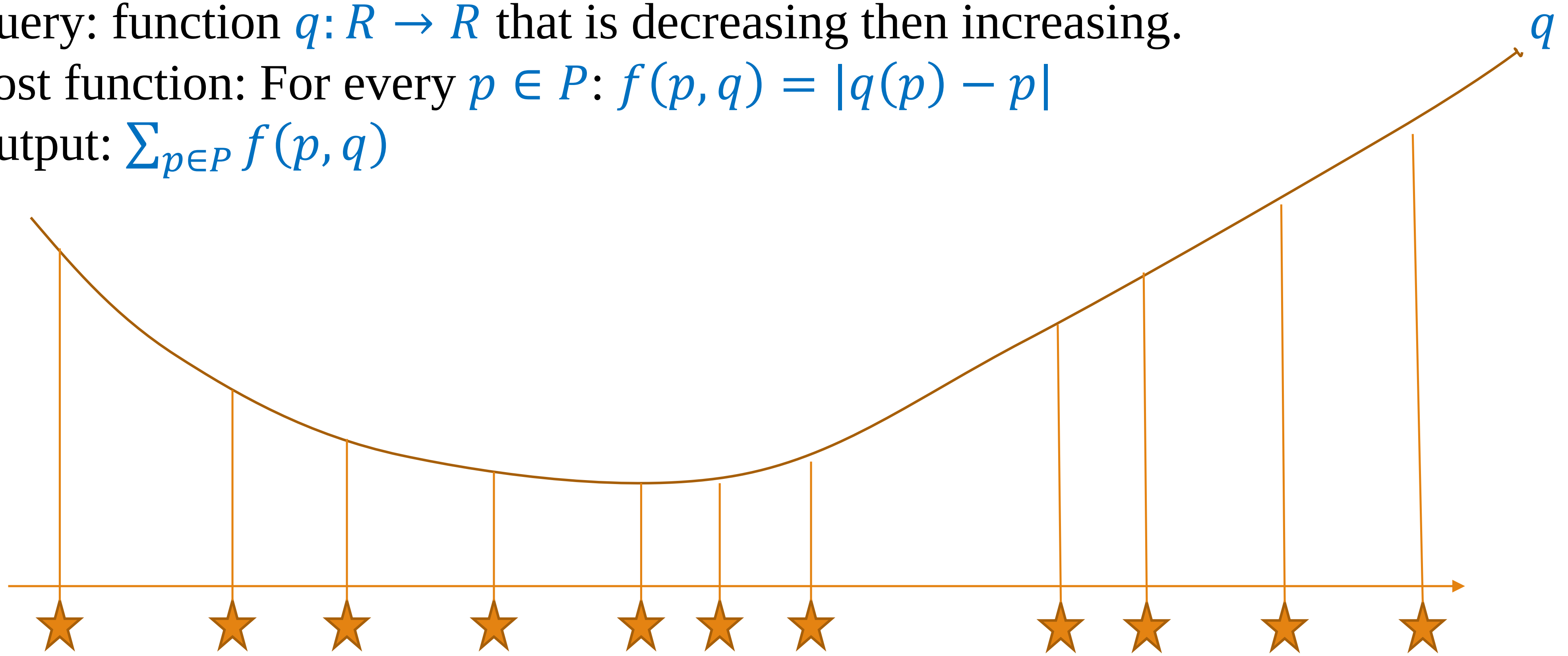
$$s(p) = \max_{x \in R, r \in \{0,1\}} \frac{f(p,x,r)}{\sum_{p' \in P} f(p',x,r)} \leq \max_{x \in R} \frac{f(p,x,1)}{\sum_{p' \in P} f(p',x,1)} + \max_{x \in R} \frac{f(p,x,0)}{\sum_{p' \in P} f(p',x,0)}$$

$$\leq \log n + \log n = O(\log n)$$

Coreset for the Convex Function

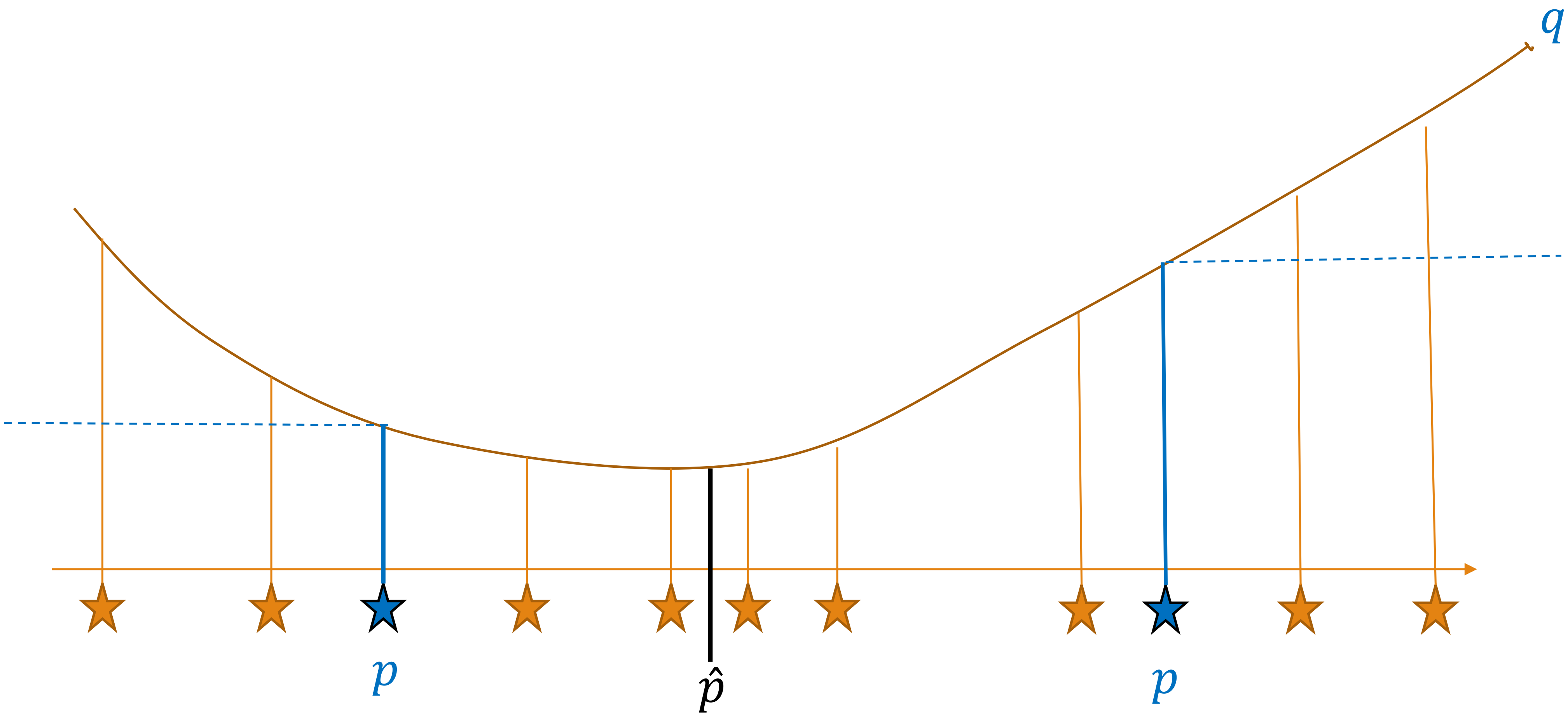
Let $P \subseteq R$ be a set of n points.

Query: function $q: R \to R$ that is decreasing then increasing.
Cost function: For every $p \in P$: $f(p, q) = |q(p) - p|$
Output: $\sum_{p \in P} f(p, q)$



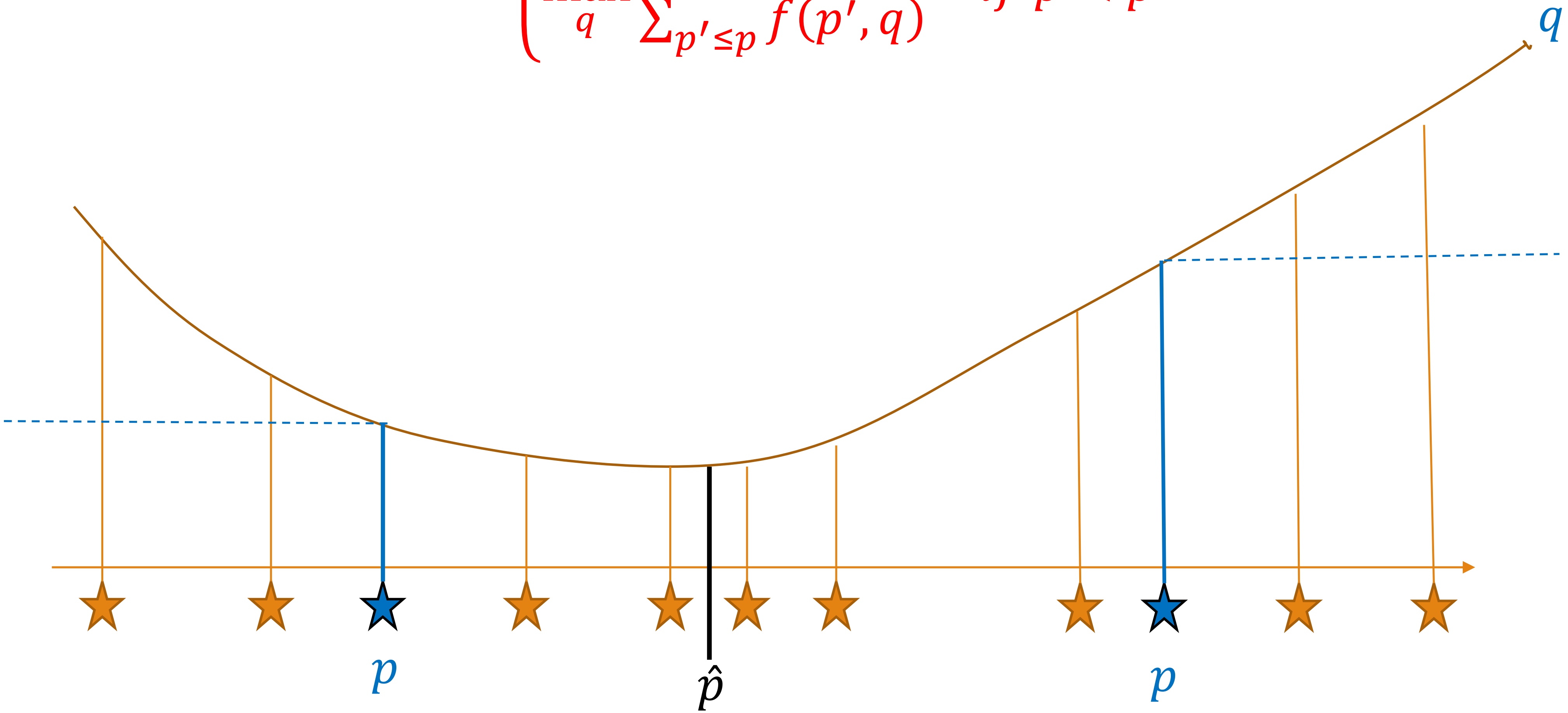
Coreset for the Convex Function

$$s(p) = \max_q \frac{f(p,q)}{\sum_{p' \in P} f(p',q)}$$



Coreset for the Convex Function

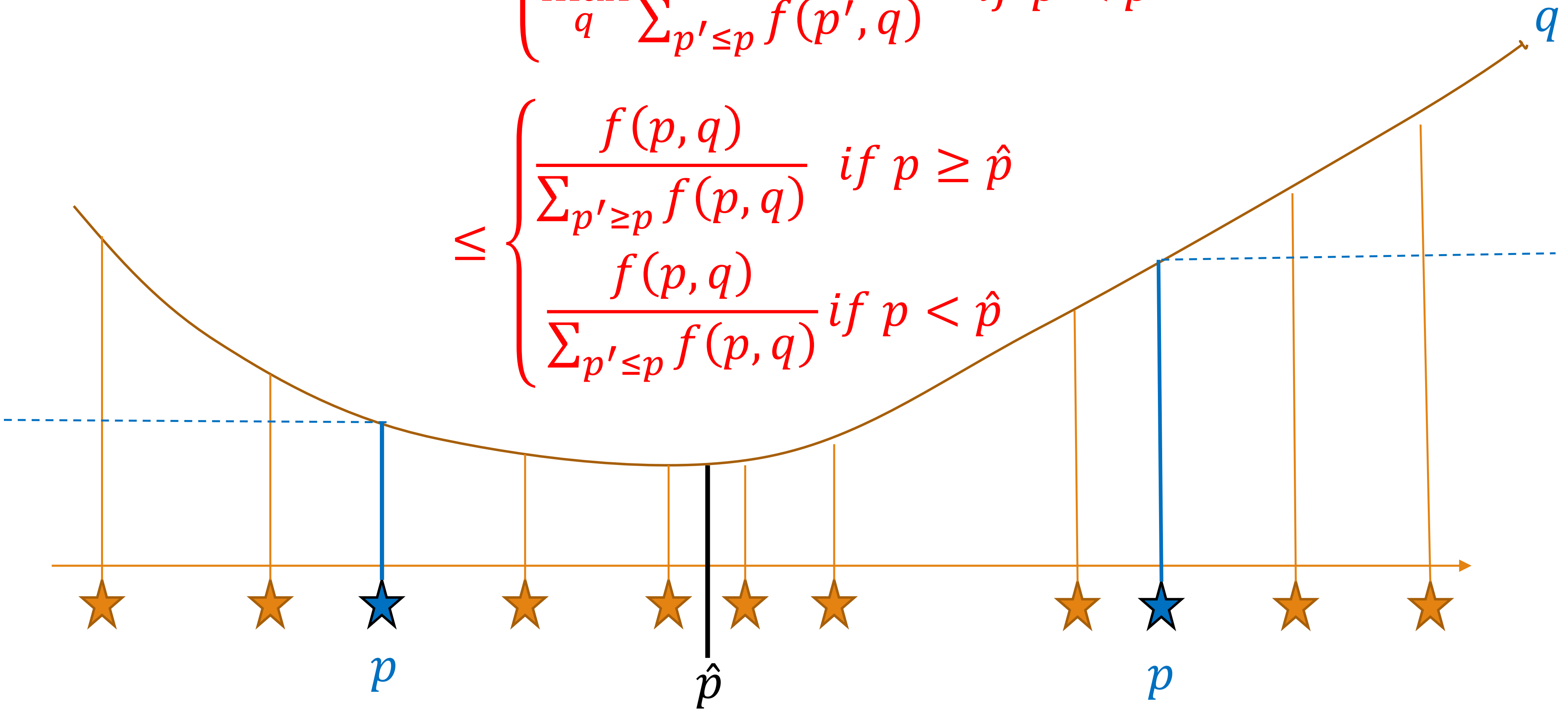
$$s(p) = \max_q \frac{f(p,q)}{\sum_{p' \in P} f(p',q)} \leq \begin{cases} \max_q \frac{f(p,q)}{\sum_{p' \geq p} f(p',q)} & if\ p \geq \hat{p} \\ \max_q \frac{f(p,q)}{\sum_{p' \leq p} f(p',q)} & if\ p < \hat{p} \end{cases}$$



Coreset for the Convex Function

$$s(p) = \max_q \frac{f(p,q)}{\sum_{p'\in P} f(p',q)} \le \begin{cases} \max_q \dfrac{f(p,q)}{\sum_{p'\ge p} f(p',q)} & if\ p \ge \hat{p} \\ \max_q \dfrac{f(p,q)}{\sum_{p'\le p} f(p',q)} & if\ p < \hat{p} \end{cases}$$

$$\le \begin{cases} \dfrac{f(p,q)}{\sum_{p'\ge p} f(p,q)} & if\ p \ge \hat{p} \\ \dfrac{f(p,q)}{\sum_{p'\le p} f(p,q)} & if\ p < \hat{p} \end{cases}$$

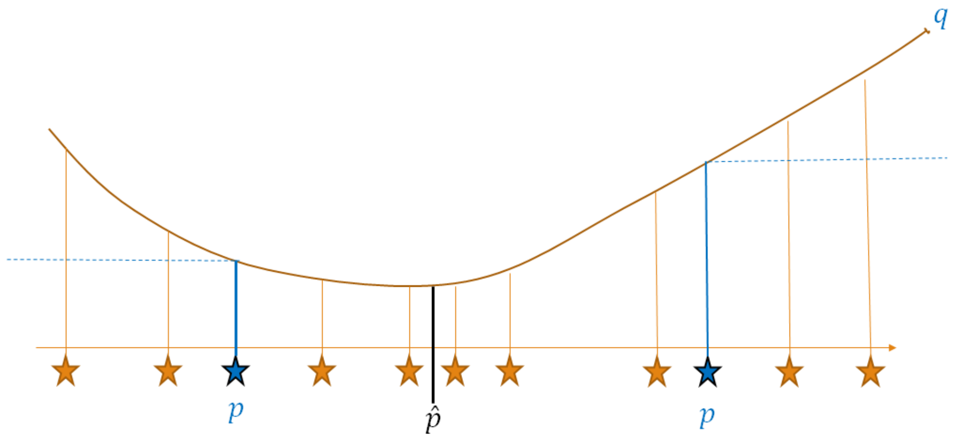


Coreset for the Convex Function

$$s(p) = \max_q \frac{f(p,q)}{\sum_{p' \in P} f(p',q)} \leq \begin{cases} \max_q \dfrac{f(p,q)}{\sum_{p' \geq p} f(p',q)} & if\ p \geq \hat{p} \\ \max_q \dfrac{f(p,q)}{\sum_{p' \leq p} f(p',q)} & if\ p < \hat{p} \end{cases}$$

$$\leq \begin{cases} \dfrac{f(p,q)}{\sum_{p' \geq p} f(p,q)} & if\ p \geq \hat{p} \\ \dfrac{f(p,q)}{\sum_{p' \leq p} f(p,q)} & if\ p < \hat{p} \end{cases}$$

$$\leq \min\left\{\frac{1}{i}, \frac{1}{n-i}\right\}$$



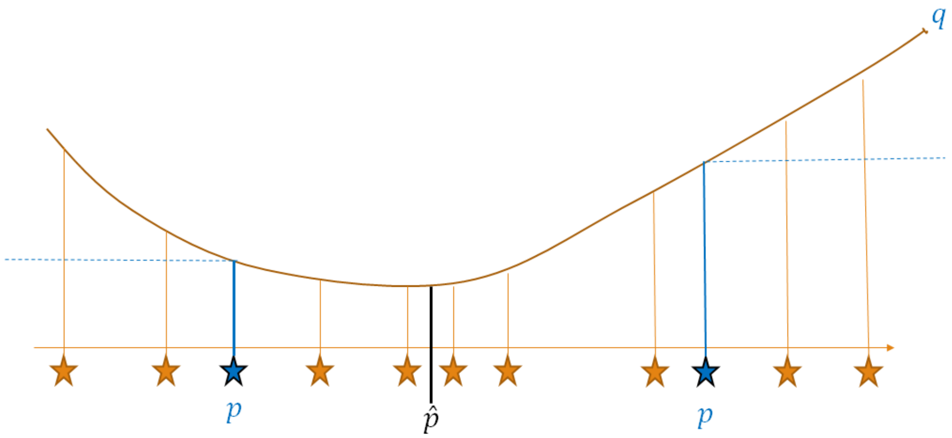
Coreset for the Convex Function

$$s(p) = \max_q \frac{f(p,q)}{\sum_{p' \in P} f(p',q)} \leq \begin{cases} \max_q \frac{f(p,q)}{\sum_{p' \geq p} f(p',q)} & if\ p \geq \hat{p} \\ \max_q \frac{f(p,q)}{\sum_{p' \leq p} f(p',q)} & if\ p < \hat{p} \end{cases}$$

$$\leq \begin{cases} \frac{f(p,q)}{\sum_{p' \geq p} f(p,q)} & if\ p \geq \hat{p} \\ \frac{f(p,q)}{\sum_{p' \leq p} f(p,q)} & if\ p < \hat{p} \end{cases}$$

$$\leq \min\left\{\frac{1}{i}, \frac{1}{n-i}\right\}$$

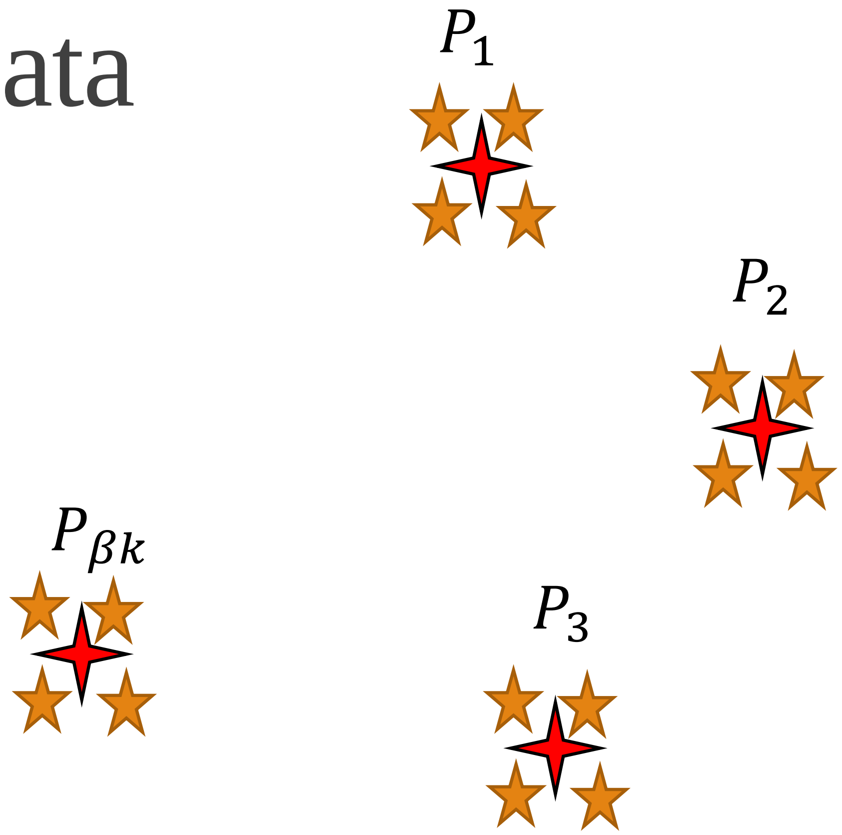
$$\rightarrow \sum_{p \in P} s(p) = O(\log n)$$



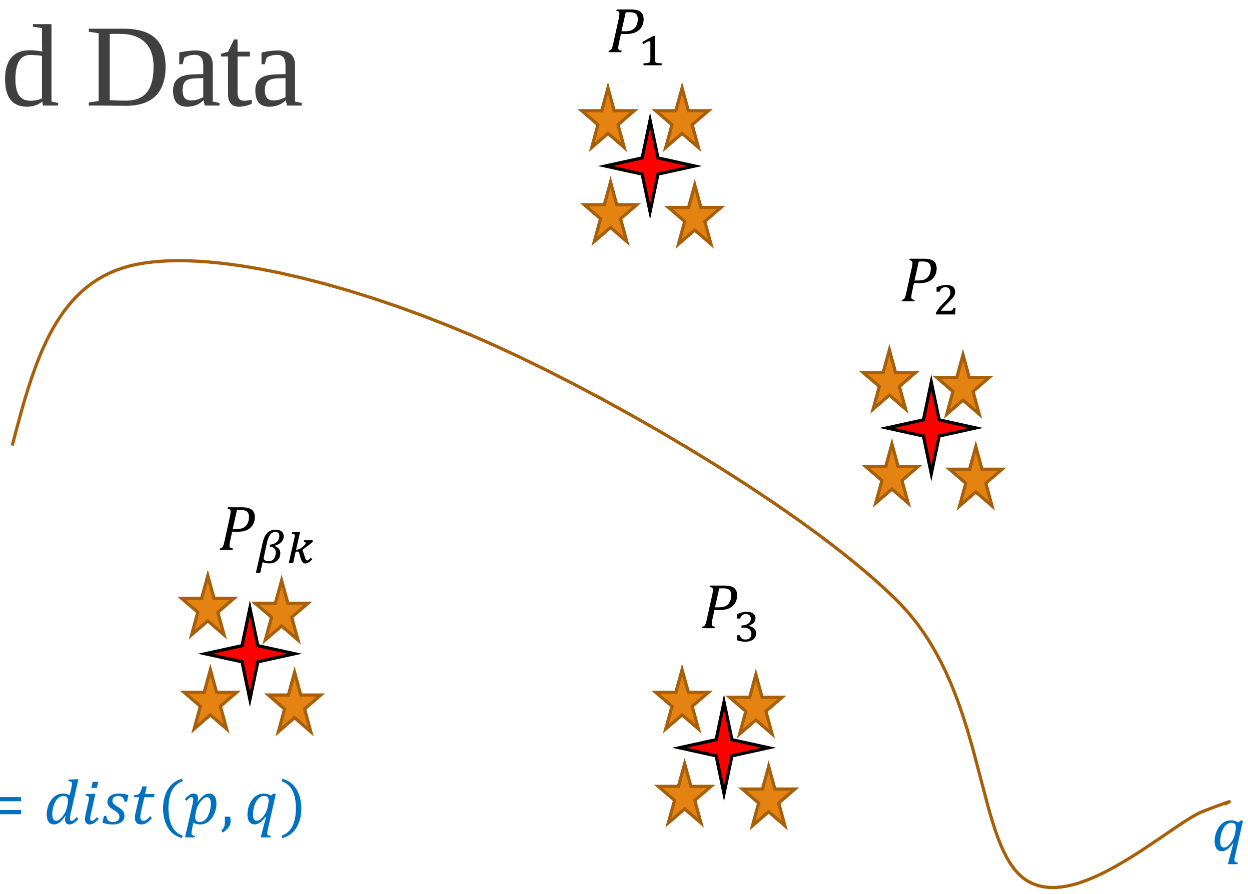
Sensitivity for Clustered Data

Let:

- $p_1, \cdots p_{\beta \cdot k} \in R^d$ be $\beta \cdot k$ centers.
- $P_i = \{p_i, p_i, \ldots, p_i\}, \; |P_i| = \frac{n}{\beta \cdot k}$.
- $P = P_1 \cup P_2 \cup \cdots \cup P_{\beta \cdot k}$



Sensitivity for Clustered Data



Let:

- $p_1, \cdots p_{\beta \cdot k} \in R^d$ be $\beta \cdot k$ centers.
- $P_i = \{p_i, p_i, \dots, p_i\}, \ |P_i| = \frac{n}{\beta \cdot k}$.
- $P = P_1 \cup P_2 \cup \cdots \cup P_{\beta \cdot k}$

Query: a function q.
Cost function: For every $p \in P$: $f(p,q) = dist(p,q)$
Output: $\sum_{p \in P} f(p,q)$

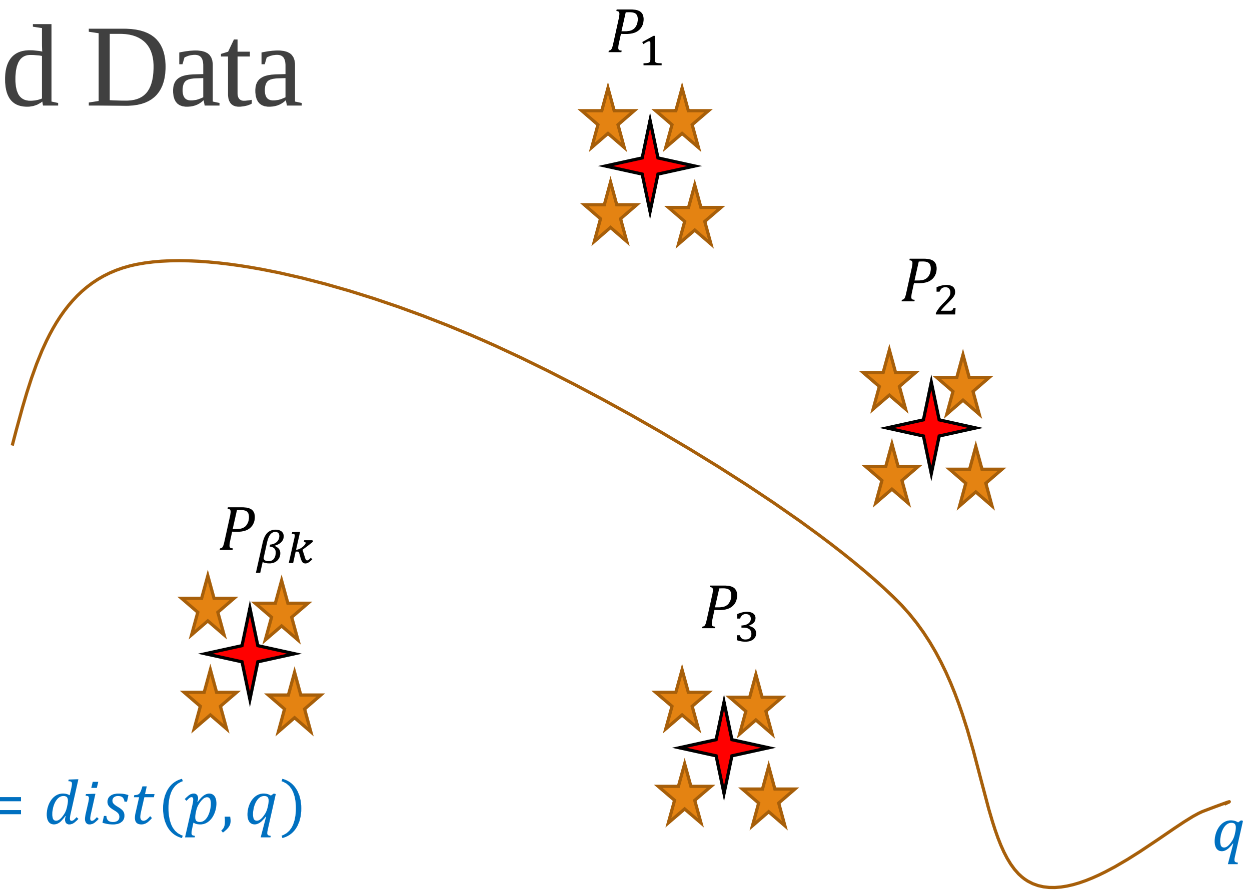
Sensitivity for Clustered Data

Let:

- $p_1, \cdots p_{\beta \cdot k} \in R^d$ be $\beta \cdot k$ centers.
- $P_i = \{p_i, p_i, \ldots, p_i\}, \ |P_i| = \frac{n}{\beta \cdot k}$.
- $P = P_1 \cup P_2 \cup \cdots \cup P_{\beta \cdot k}$

Query: a function q.
Cost function: For every $p \in P$: $f(p, q) = dist(p, q)$
Output: $\sum_{p \in P} f(p, q)$

$$s(p_i) = \max_q \frac{f(p_i, q)}{\sum_{p' \in P} f(p', q)} \le \max_q \frac{f(p_i, q)}{\sum_{p_i' \in P_i} f(p_i', q)}$$

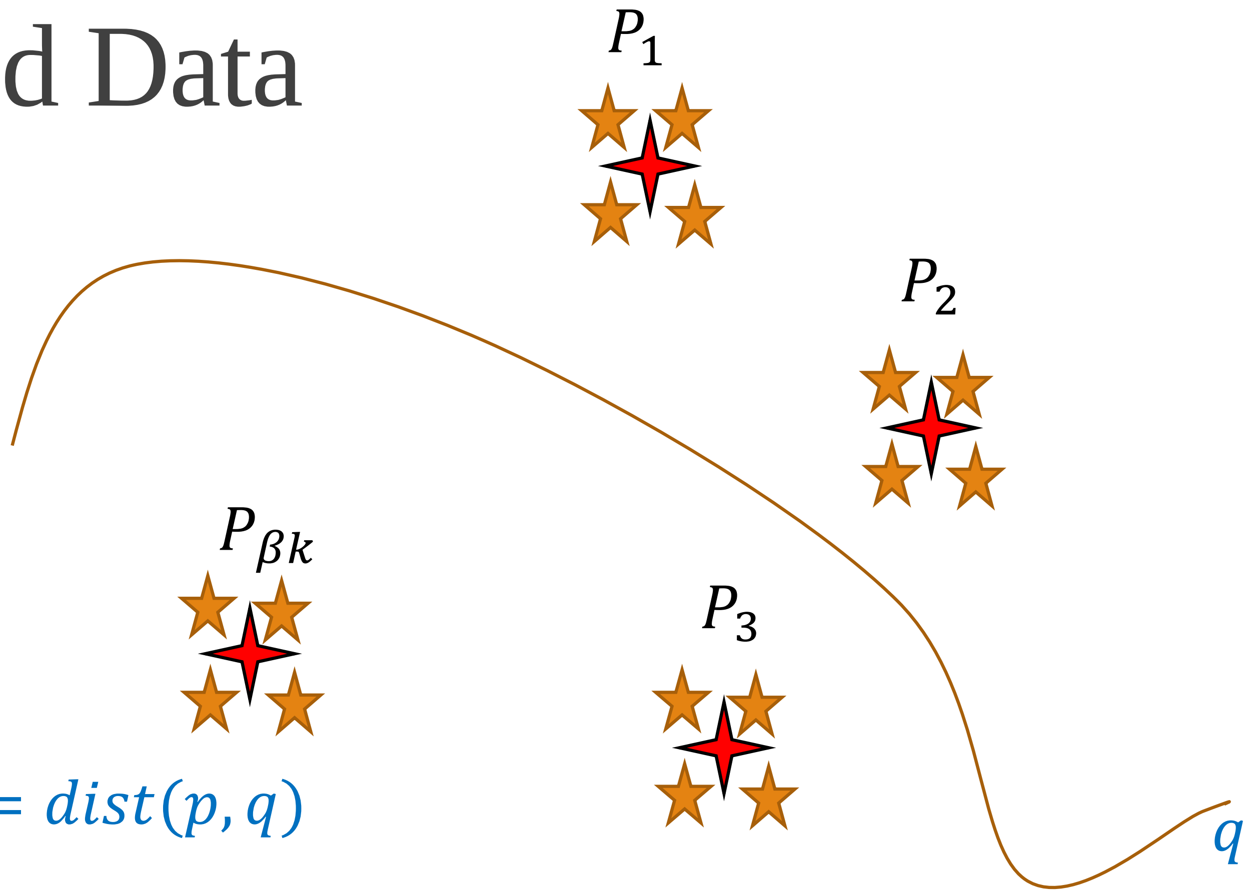


Sensitivity for Clustered Data

Let:

- $p_1, \cdots p_{\beta \cdot k} \in R^d$ be $\beta \cdot k$ centers.
- $P_i = \{p_i, p_i, \ldots, p_i\}, \; |P_i| = \frac{n}{\beta \cdot k}$.
- $P = P_1 \cup P_2 \cup \cdots \cup P_{\beta \cdot k}$

Query: a function q.
Cost function: For every $p \in P$: $f(p,q) = dist(p,q)$
Output: $\sum_{p \in P} f(p,q)$

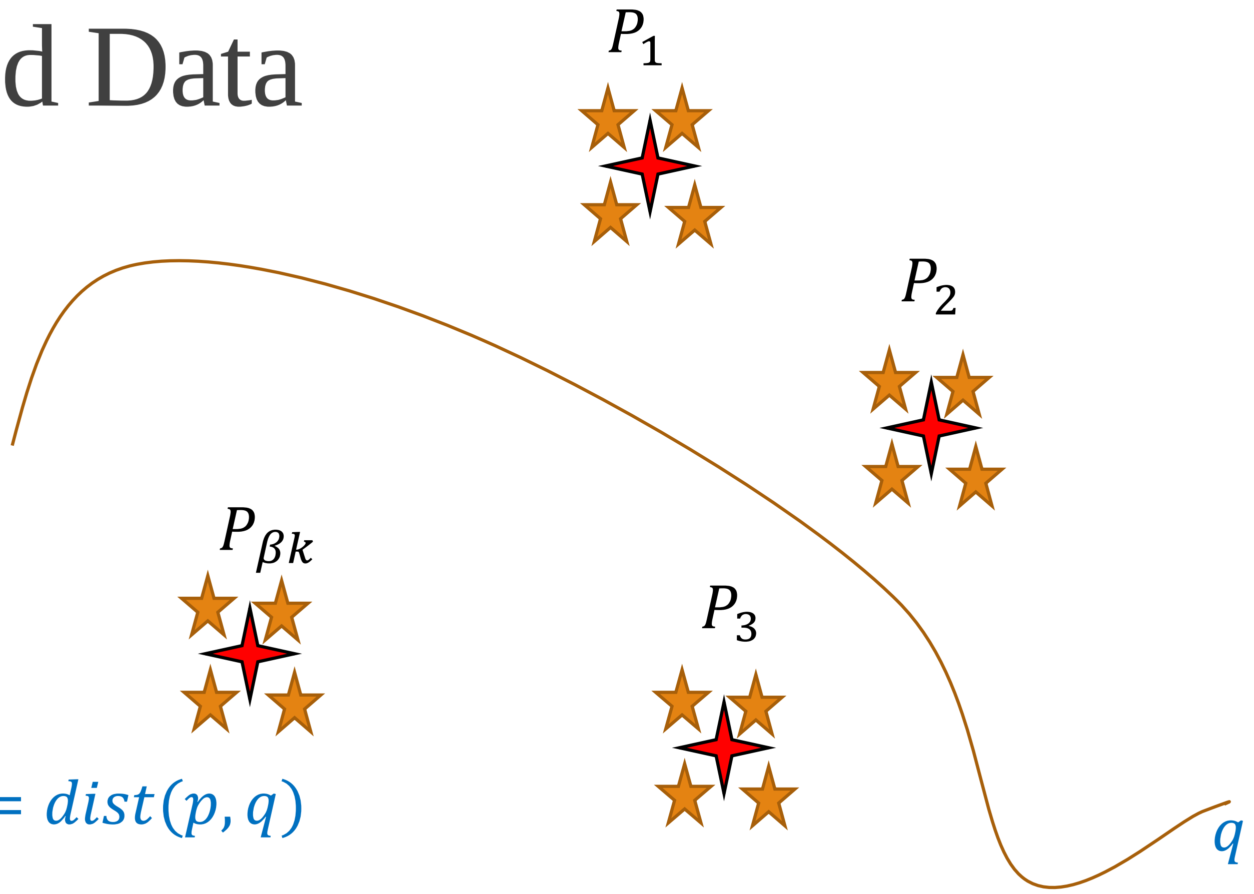


$$s(p_i) = \max_q \frac{f(p_i,q)}{\sum_{p' \in P} f(p',q)} \leq \max_q \frac{f(p_i,q)}{\sum_{p_i' \in P_i} f(p_i',q)} = \frac{1}{|P_i|}$$

Sensitivity for Clustered Data

Let:

- $p_1, \cdots p_{\beta \cdot k} \in R^d$ be $\beta \cdot k$ centers.
- $P_i = \{p_i, p_i, \dots, p_i\}, \; |P_i| = \frac{n}{\beta \cdot k}$.
- $P = P_1 \cup P_2 \cup \cdots \cup P_{\beta \cdot k}$



Query: a function q.
Cost function: For every $p \in P$: $f(p,q) = dist(p,q)$
Output: $\sum_{p \in P} f(p,q)$

$$s(p_i) = \max_q \frac{f(p_i,q)}{\sum_{p' \in P} f(p',q)} \leq \max_q \frac{f(p_i,q)}{\sum_{p_i' \in P_i} f(p_i',q)} = \frac{1}{|P_i|}$$

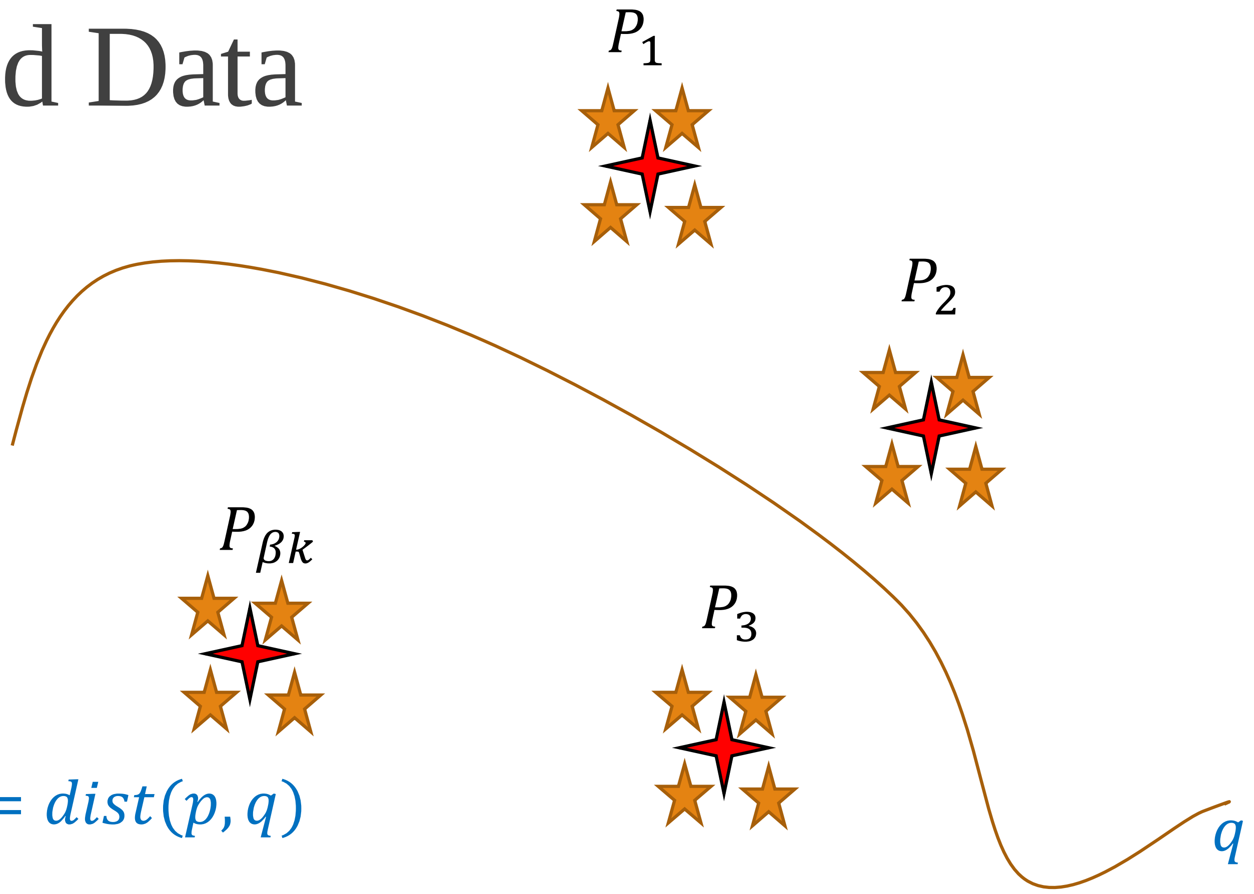
$$\sum_{p_i \in P_i} s(p_i) = \sum_{p_i \in P_i} \frac{1}{|P_i|} = 1$$

Sensitivity for Clustered Data

Let:

- $p_1, \cdots p_{\beta \cdot k} \in R^d$ be $\beta \cdot k$ centers.
- $P_i = \{p_i, p_i, \ldots, p_i\}$, $|P_i| = \frac{n}{\beta \cdot k}$.
- $P = P_1 \cup P_2 \cup \cdots \cup P_{\beta \cdot k}$

Query: a function q.
Cost function: For every $p \in P$: $f(p, q) = dist(p, q)$
Output: $\sum_{p \in P} f(p, q)$



$$s(p_i) = \max_q \frac{f(p_i, q)}{\sum_{p' \in P} f(p', q)} \leq \max_q \frac{f(p_i, q)}{\sum_{p_i' \in P_i} f(p_i', q)} = \frac{1}{|P_i|}$$

$$\sum_{p_i \in P_i} s(p_i) = \sum_{p_i \in P_i} \frac{1}{|P_i|} = 1$$

$$\sum_{p \in P} s(p_i) = \beta \cdot k$$

Bounding Sensitivity using Bicreteria

Lemma: Let $(X, dist)$ be a metric space such that the weak triangle inequality holds: for every $p, q, x \in X$: $dist(p,x) \leq \rho\big(dist(p,q) + dist(q,x)\big)$.
Let $A \subseteq X$ and Q be (possibly infinite) subsets in X.

Let $A' \subseteq X$ and suppose that there is a mapping from every $p \in A$ to a point $p' \in A'$.

If $dist(A, A') \leq \alpha \cdot \mathrm{OPT}$ where $\mathrm{OPT} = \min_{T^* \in Q} \sum_{p \in A} dist(p, T^*)$ for some $\alpha > 0$

(i.e., if A' is an (α, β)-approximation) then:

$$\sum_{p \in A} s(p) = \sum_{p \in A} \max_{T \in Q} \frac{dist(p,T)}{dist(A,T)} \leq \rho\alpha + \rho^2(1+\alpha) \sum_{p' \in A'} \max_{T \in Q} \frac{dist(p',T)}{dist(A',T)}$$

Bounding Sensitivity using Bicreteria

Proof:

Let $T \in Q$ and $p \in A$. By the weak triangle inequality:

$$\underbrace{\frac{dist(p,T)}{dist(A,T)}}_{s(p)} \leq \frac{\rho \cdot dist(p,p')}{dist(A,T)} + \frac{\rho \cdot dist(p',T)}{dist(A,T)}$$

Bounding Sensitivity using Bicreteria

Proof:

Let $T \in Q$ and $p \in A$. By the weak triangle inequality:

$$\frac{dist(p,T)}{dist(A,T)} \leq \frac{\rho \cdot dist(p,p')}{dist(A,T)} + \frac{\rho \cdot dist(p',T)}{dist(A,T)}$$

$$\leq \frac{\rho \cdot dist(p,p')}{OPT} + \frac{\rho \cdot dist(p',T)}{dist(A,T)}$$

$OPT \leq dist(A,T)$

Bounding Sensitivity using Bicreteria

Proof:

Let $T \in Q$ and $p \in A$. By the weak triangle inequality:

$$\frac{dist(p,T)}{dist(A,T)} \leq \frac{\rho \cdot dist(p,p')}{dist(A,T)} + \frac{\rho \cdot dist(p',T)}{dist(A,T)}$$

$$\leq \frac{\rho \cdot dist(p,p')}{OPT} + \frac{\rho \cdot dist(p',T)}{dist(A,T)}$$

$$\leq \frac{\rho \cdot dist(A,A')}{OPT} \cdot \frac{dist(p,p')}{dist(A,A')} + \frac{\rho \cdot dist(p',T)}{dist(A,T)}$$

Bounding Sensitivity using Bicreteria

Proof:

Let $T \in Q$ and $p \in A$. By the weak triangle inequality:

$$\frac{dist(p,T)}{dist(A,T)} \le \frac{\rho \cdot dist(p,p')}{dist(A,T)} + \frac{\rho \cdot dist(p',T)}{dist(A,T)}$$

$$\le \frac{\rho \cdot dist(p,p')}{OPT} + \frac{\rho \cdot dist(p',T)}{dist(A,T)}$$

$$\le \frac{\rho \cdot dist(A,A')}{OPT} \cdot \frac{dist(p,p')}{dist(A,A')} + \frac{\rho \cdot dist(p',T)}{dist(A,T)}$$

$$\le \rho\alpha \frac{dist(p,p')}{dist(A,A')} + \frac{\rho \cdot dist(p',T)}{dist(A,T)}$$

$$\frac{dist(A,A')}{OPT} \le \alpha$$

Bounding Sensitivity us

$$dist(A',T) \le \rho\big(dist(A',A) + dist(A,T)\big)$$
$$\le \rho\alpha \cdot OPT + \rho \cdot dist(A,T)$$
$$\le \rho(\alpha+1) \cdot dist(A,T)$$
$$\rightarrow dist(A,T) \ge \frac{dist(A',T)}{\rho(\alpha+1)}$$

Proof:

Let $T \in Q$ and $p \in A$. By the weak triang

$$\frac{dist(p,T)}{dist(A,T)} \le \frac{\rho \cdot dist(p,p')}{dist(A,T)} + \frac{\rho \cdot dist(p',T)}{dist(A,T)}$$
$$\le \frac{\rho \cdot dist(p,p')}{OPT} + \frac{\rho \cdot dist(p',T)}{dist(A,T)}$$
$$\le \frac{\rho \cdot dist(A,A')}{OPT} \cdot \frac{dist(p,p')}{dist(A,A')} + \frac{\rho \cdot dist(p',T)}{dist(A,T)}$$
$$\le \rho\alpha \frac{dist(p,p')}{dist(A,A')} + \frac{\rho \cdot dist(p',T)}{dist(A,T)}$$
$$\le \rho\alpha \frac{dist(p,p')}{dist(A,A')} + \frac{\rho^2(\alpha+1) \cdot dist(p',T)}{dist(A',T)}$$

Bounding Sensitivity using Bicreteria

Proof:

$$\rightarrow s(p) = \frac{dist(p,T)}{dist(A,T)} \leq \rho\alpha \frac{dist(p,p')}{dist(A,A')} + \frac{\rho^2(\alpha+1)\cdot dist(p',T)}{dist(A',T)}$$

Bounding Sensitivity using Bicreteria

Proof:

$$\rightarrow s(p) = \frac{dist(p,T)}{dist(A,T)} \leq \rho\alpha \frac{dist(p,p')}{dist(A,A')} + \frac{\rho^2(\alpha+1)\cdot dist(p',T)}{dist(A',T)}$$

$$\rightarrow \sum_{p\in A} s(p) \leq \sum_{p\in A}\left(\rho\alpha \frac{dist(p,p')}{dist(A,A')} + \frac{\rho^2(\alpha+1)\cdot dist(p',T)}{dist(A',T)}\right)$$

Bounding Sensitivity using Bicreteria

Proof:

$$\rightarrow s(p) = \frac{dist(p,T)}{dist(A,T)} \leq \rho\alpha \frac{dist(p,p')}{dist(A,A')} + \frac{\rho^2(\alpha+1)\cdot dist(p',T)}{dist(A',T)}$$

$$\rightarrow \sum_{p\in A} s(p) \leq \sum_{p\in A}\left(\rho\alpha \frac{dist(p,p')}{dist(A,A')} + \frac{\rho^2(\alpha+1)\cdot dist(p',T)}{dist(A',T)}\right)$$

$$\rightarrow \sum_{p\in A} s(p) \leq \rho\alpha + \rho^2(\alpha+1)\cdot \sum_{p\in A} \max_{T\in Q} \frac{dist(p',T)}{dist(A',T)}$$

Bounding Sensitivity using Bicreteria

Proof:

$$\rightarrow s(p) = \frac{dist(p,T)}{dist(A,T)} \leq \rho\alpha \frac{dist(p,p')}{dist(A,A')} + \frac{\rho^2(\alpha+1)\cdot dist(p',T)}{dist(A',T)}$$

$$\rightarrow \sum_{p\in A} s(p) \leq \sum_{p\in A} \left(\rho\alpha \frac{dist(p,p')}{dist(A,A')} + \frac{\rho^2(\alpha+1)\cdot dist(p',T)}{dist(A',T)} \right)$$

$$\rightarrow \sum_{p\in A} s(p) \leq \rho\alpha + \rho^2(\alpha+1)\cdot \sum_{p\in A} \max_{T\in Q} \frac{dist(p',T)}{dist(A',T)}$$