USING LOCAL INFERENCE
IN MASSIVELY DISTRIBUTED SYSTEMS

Safe Zones for Tracking the Median Function

Assaf Schuster, Technion

Daniel Keren, Haifa University

Tsachi Sharfman, Technion

Minos Garofalakis, Technical University of Crete

Vasilis Samoladas, Technical University of Crete

Goal

- Construct safe zones that ***contain*** the safe zones defined by the geometric method
  - Guarantees that resulting constraints remain effective for longer periods of time
  - Can potentially reduce communications by orders of magnitude

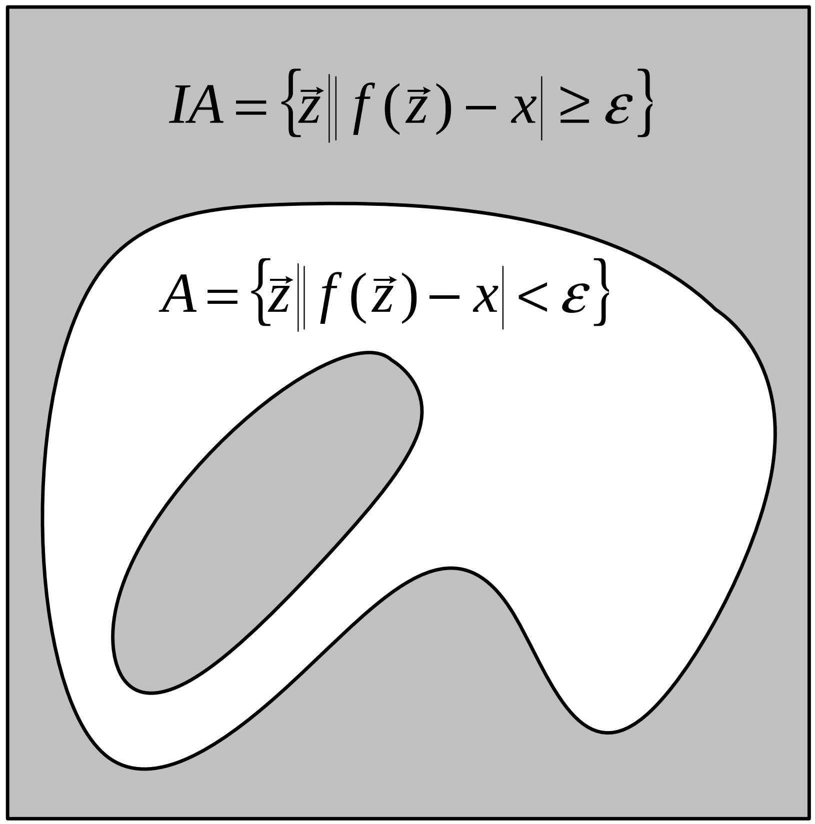
Tracking the Median Function

- We have *n* nodes, each holding a time varying *d* dimensional vector denoted by $\mathbf{v}_i$.
- We denote their average by **v.**
- We are given the median function *med*(**v**) and an approximation margin ε.
- At any time we would like to hold a number *x* , referred to as the approximation value, such that $|med(\mathbf{v}) - x| < \varepsilon$.

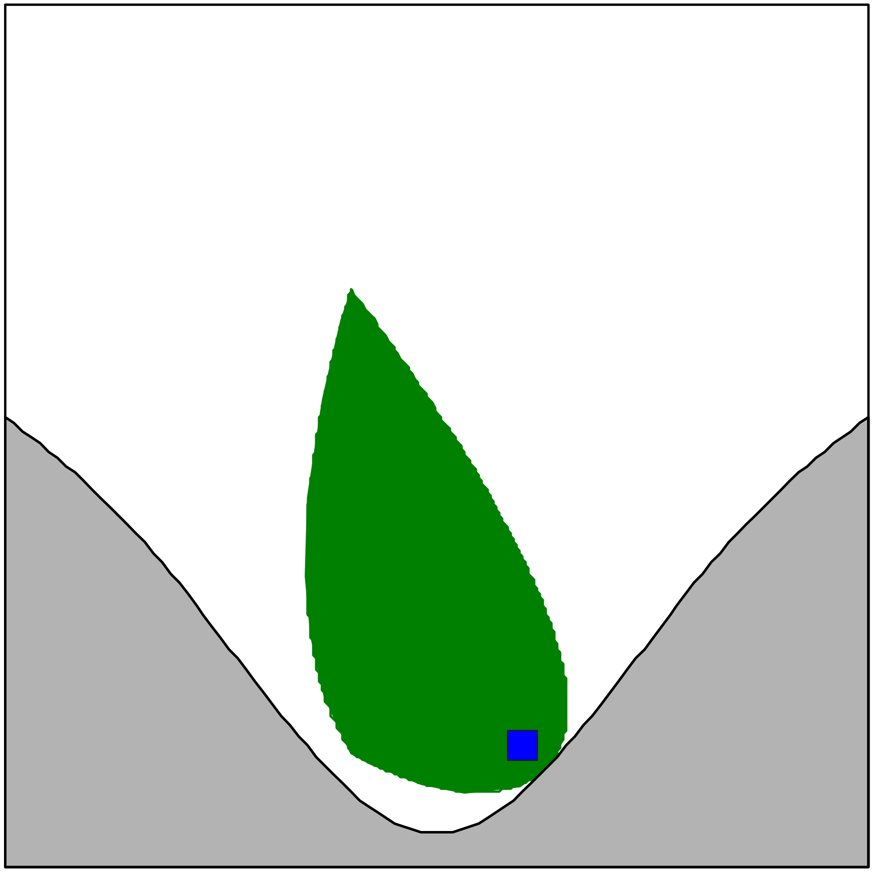
Admissible and Inadmissible Regions

- Given a function f, an estimate value x, and an approximation margin ε, we define the following:
  - Let the *admissible region*, denoted by $A(f,x,\varepsilon)$ be the set of vectors for which the value of the function is within the approximation range, i.e. $A(f,x,\varepsilon) = \{\mathbf{z} \mid |f(\mathbf{z})-x| < \varepsilon\}$
  - Similarly, let the *inadmissible region*, denoted by $IA(f,x,\varepsilon)$ be the set of vectors for which the value of the function is not within the approximation range, i.e. $A(f,x,\varepsilon) = \{\mathbf{z} \mid |f(\mathbf{z})-x| \geq \varepsilon\}$

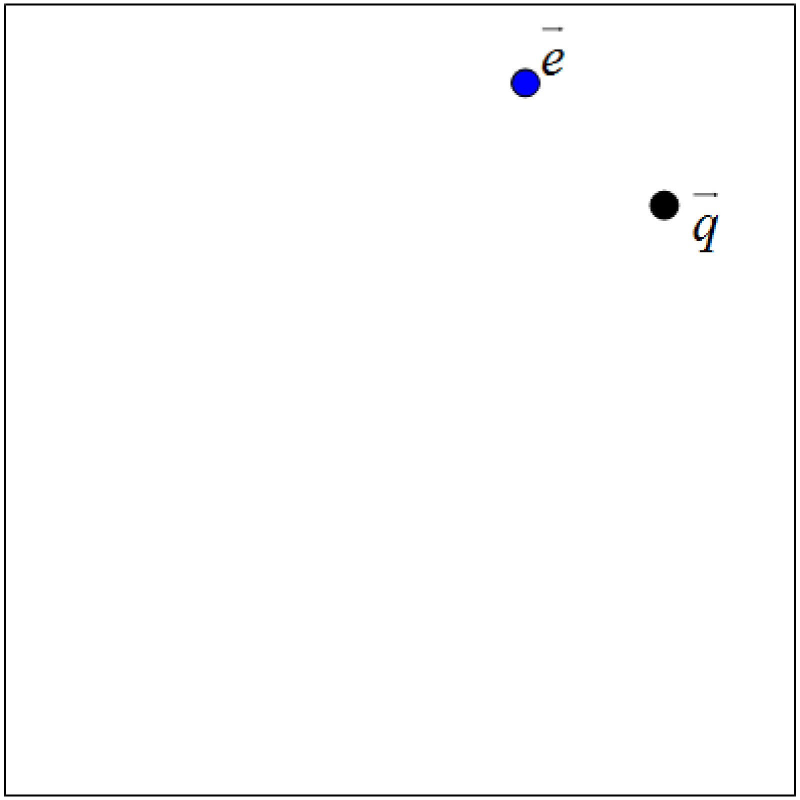
Admissible and Inadmissible Regions (Cont.)



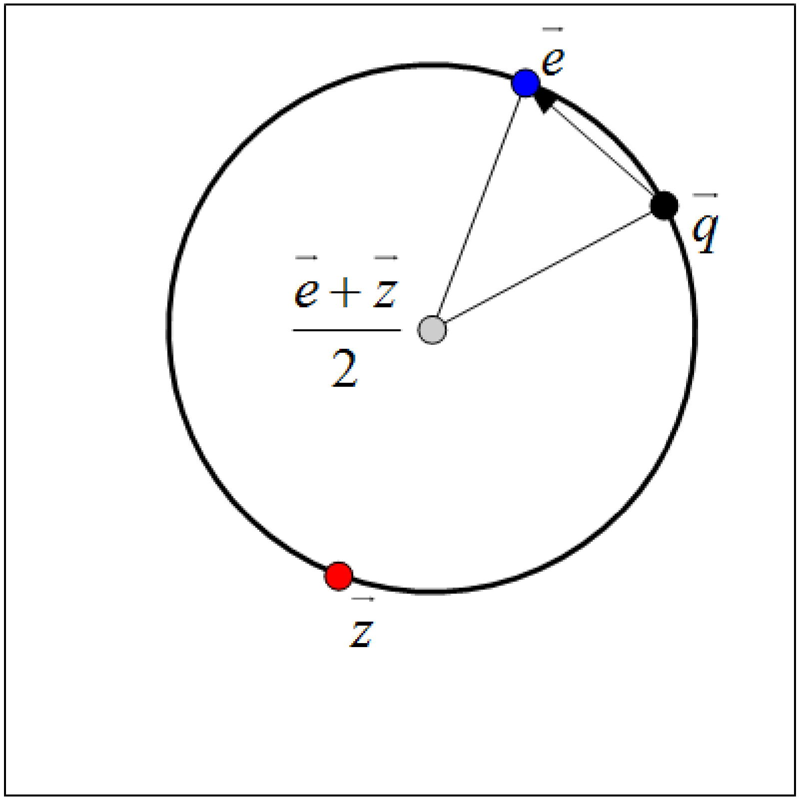
Safe Zone Induced by the Geometric Method



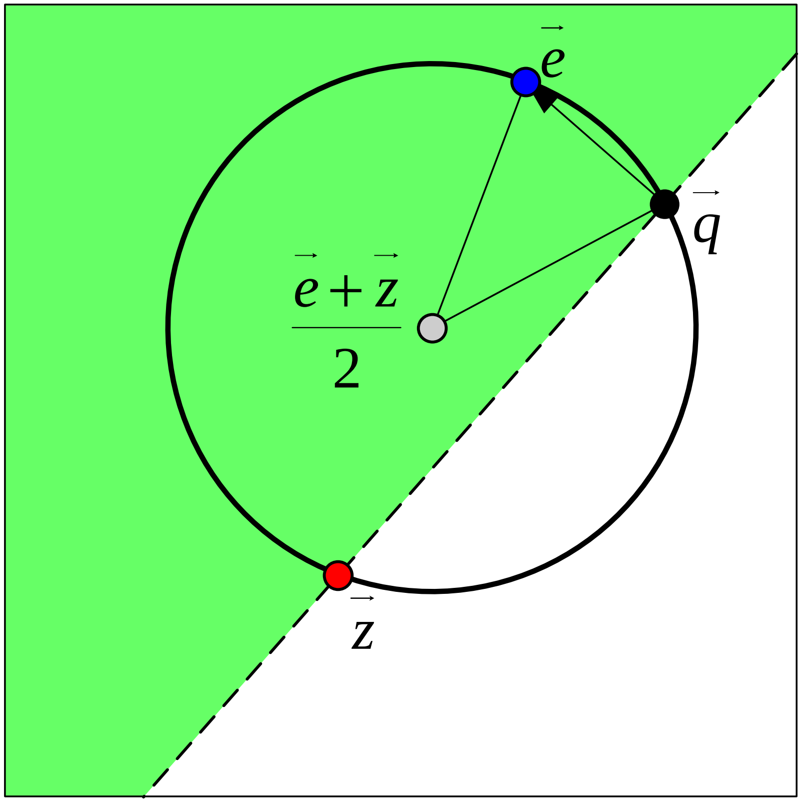
Safe Zone Induced by a Singular Vector



Safe Zone Induced by a Singular Vector



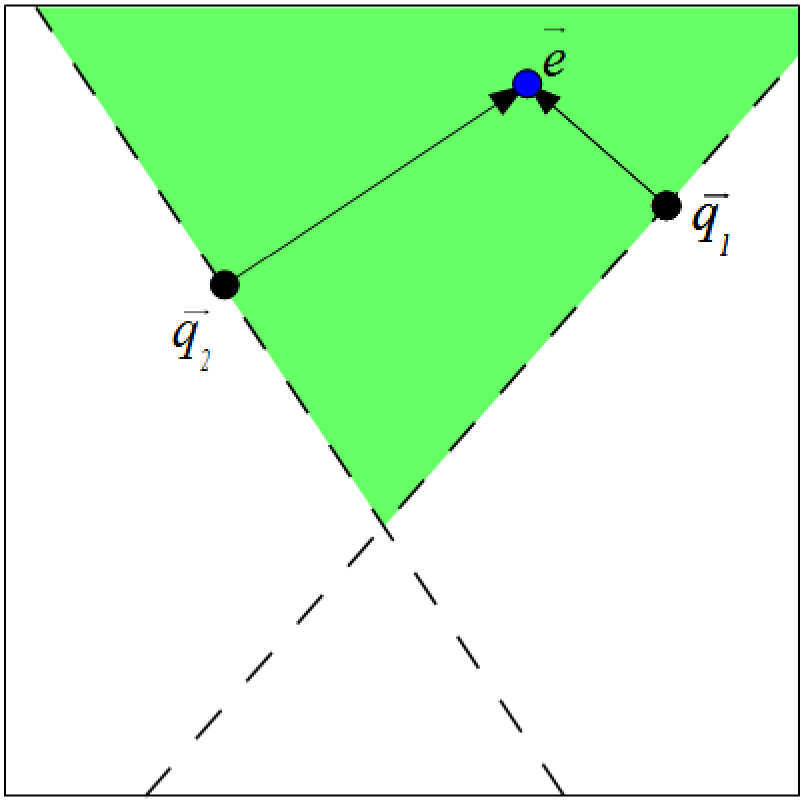
Safe Zone Induced by a Singular Vector



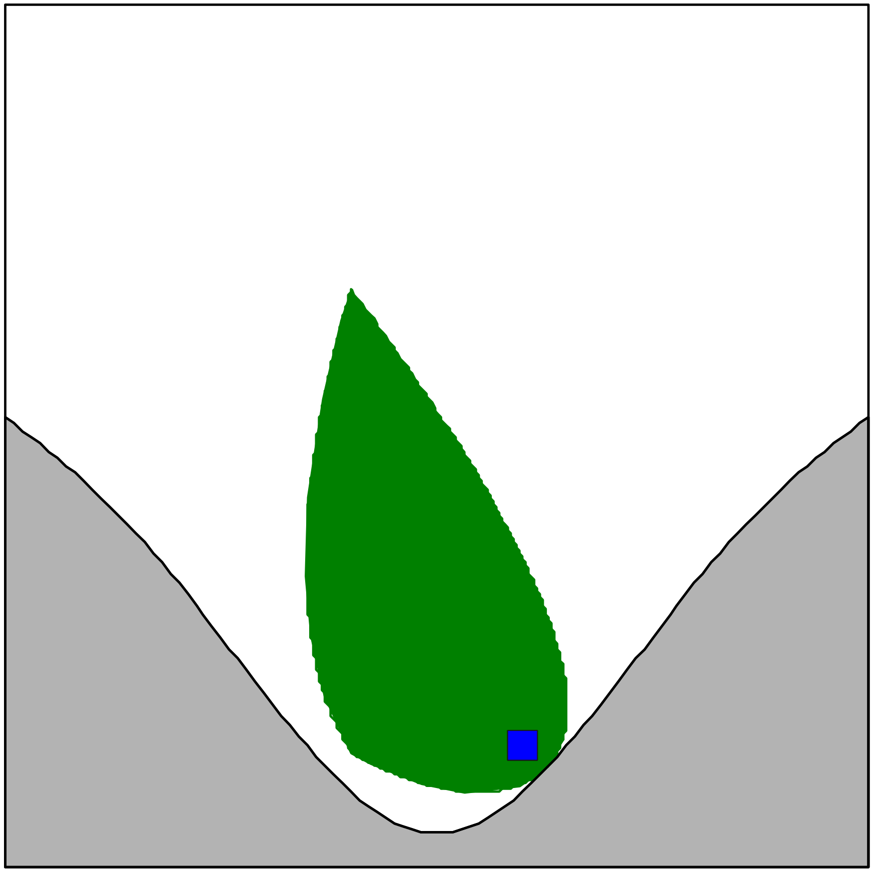
An Induced Half-Space

- Given an inadmissible vector ***q*** and an estimate vector ***e***, let *H*(***e***,***q***) be the half-space defined by the hyper-plane passing through ***q***, orthogonal to ***e-q***, and containing ***e***. We refer to *H*(***e***,***q***) as the half-space induced by ***q***.

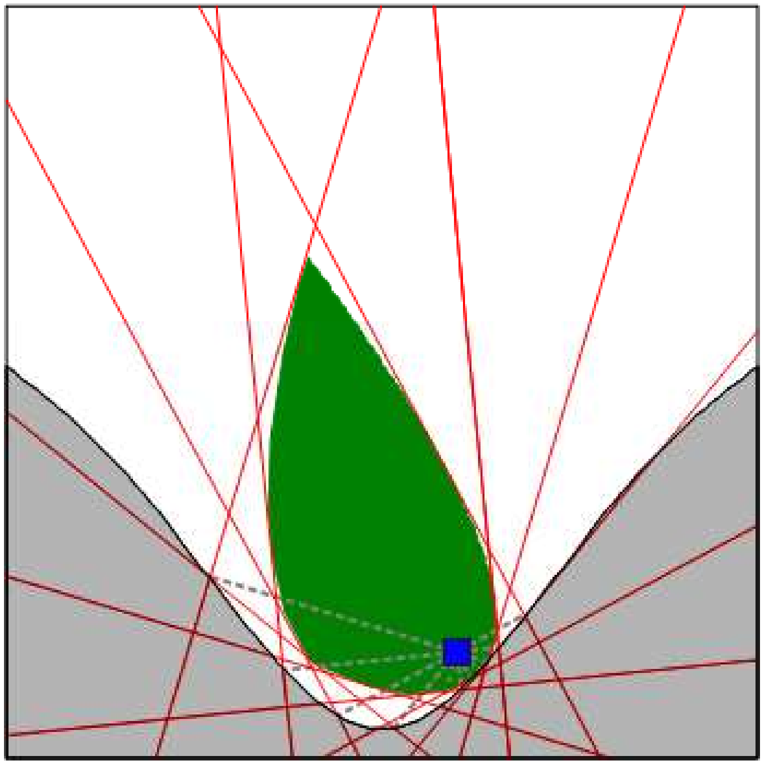
Safe Zone Induced by Two Singular Vectors



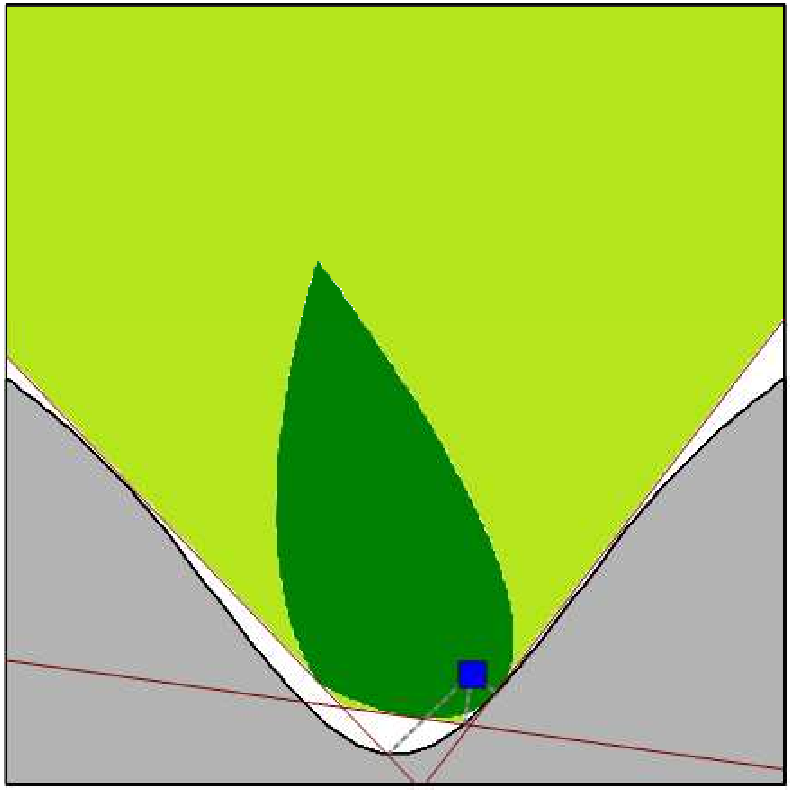
Safe Zone Induced by the Geometric Method



Safe Zone as an Intersection of Half-Spaces



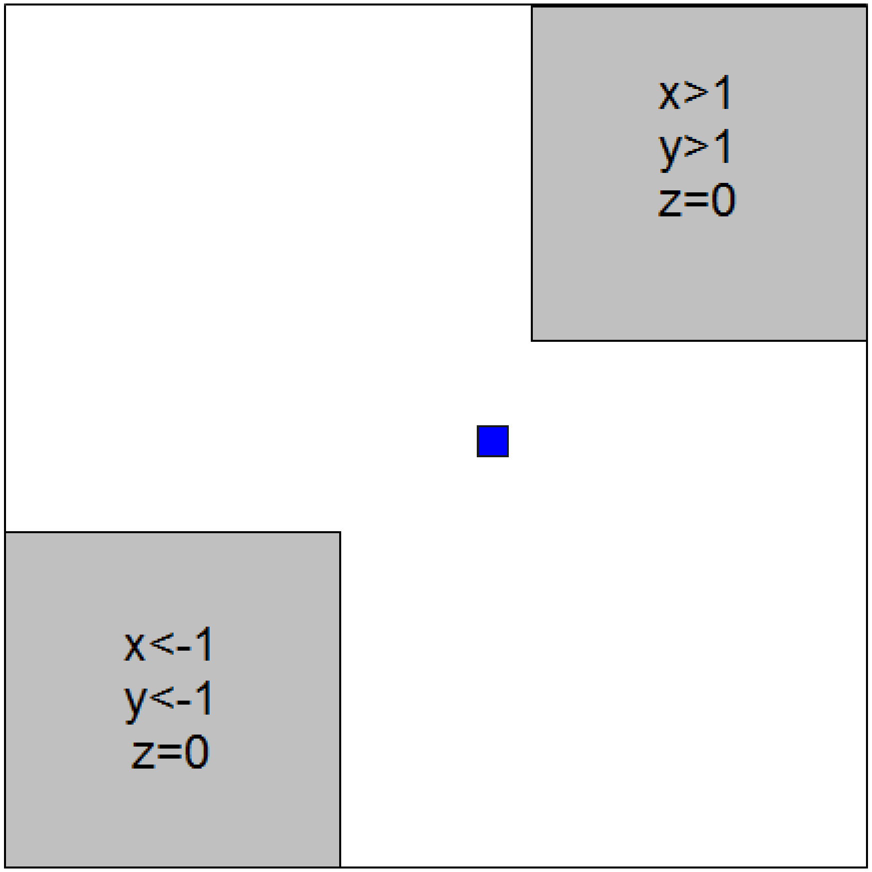
Safe Zone as an Intersection of Half-Spaces



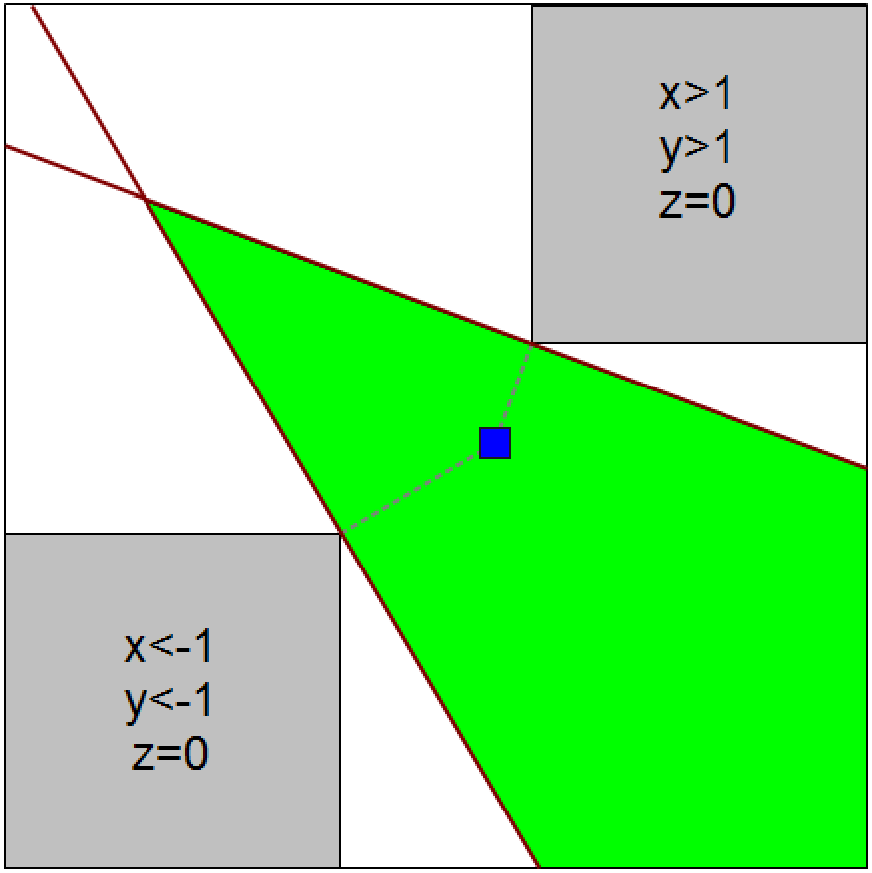
Support Vectors

- Given an inadmissible region and an estimate vector, we would like to select a small set of vectors from the set such that the intersection of the half spaces they induce is contained in the admissible region.
- We refer to these vectors as support vectors

The Median Function



The Median Function

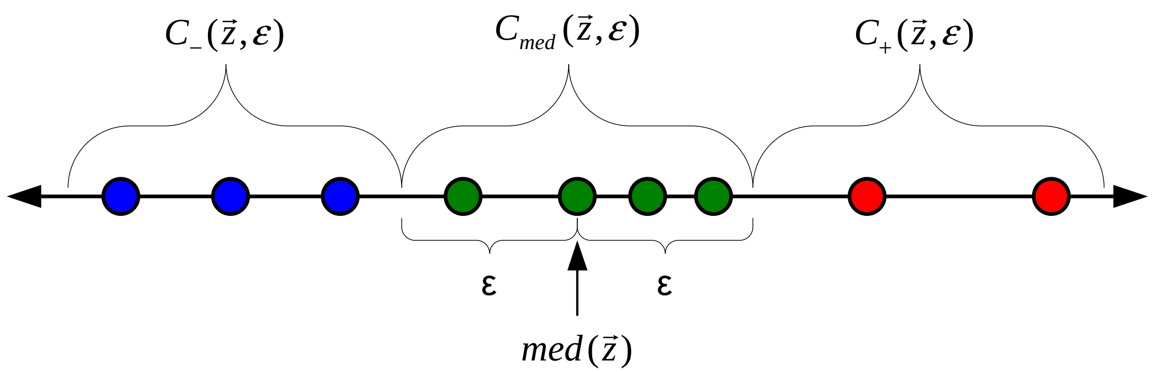


Notations

- Let I denote a subset of the natural numbers in the range $\{1,\ldots,d\}$.
- Let r be a real value, and let $\mathbf{z}$ be a d-dimensional vector. Let $\mathbf{z}^{(I,r)}$ denote the vector $\mathbf{z}$ with the value of the components whose index is in I set to r.
- Let $\begin{bmatrix} I \\ m \end{bmatrix}$ be the set of all m sized subsets of I.

Definitions

- Given a *d*-dimensional vector **z**, an estimate value *x* and an approximation margin ε, we group the vector's components as follows:
  - C_+(**z**,ε): A set of the indices of the components of **z** whose value is greater than or equal to *median*(**z**)+ε.
  - C_{med}(**z**,ε): A set of the indices of the components of **z** whose value is in the range *median*(**z**)± ε.
  - C_-(**z**,ε): A set of the indices of the components of **z** whose value is smaller than or equal to *median*(**z**)-ε



Support Vectors

- The safe zone consists of the half-spaces defined by the estimate vector and a set of vectors referred to as the support vectors.
- The set of support vectors is the union of the following two sets:

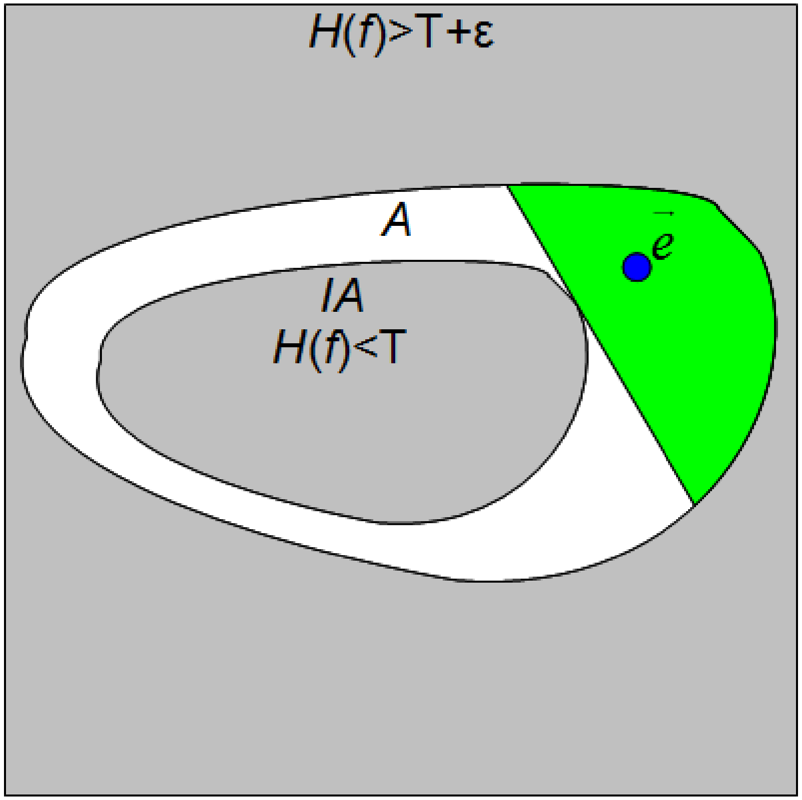
$$Support_{+}(\vec{e},\varepsilon) = \left\{ \vec{e}^{(I, med(\vec{e})-\varepsilon)} \middle| I \in \begin{bmatrix} C_{med}(\vec{e},\varepsilon) \bigcup C_{+}(\vec{e},\varepsilon) \\ \frac{d+1}{2} - \left\| C_{-}(\vec{e},\varepsilon) \right\| \end{bmatrix} \right\}$$

$$Support_{-}(\vec{e},\varepsilon) = \left\{ \vec{e}^{(I, med(\vec{e})+\varepsilon)} \middle| I \in \begin{bmatrix} C_{med}(\vec{e},\varepsilon) \bigcup C_{-}(\vec{e},\varepsilon) \\ \frac{d+1}{2} - \left\| C_{+}(\vec{e},\varepsilon) \right\| \end{bmatrix} \right\}$$

Monitoring Entropy

- We are given a distributed set of time-varying frequency vectors $\boldsymbol{f}_i$. let their $\boldsymbol{f}$ denote their sum.
- We would like to monitor the entropy of the sum, $H(\boldsymbol{f})$, with an approximation error of ε.
- It is easy to show that the inadmissible region *IA* defined by $H(\boldsymbol{f})>T$ is convex.
- Consequently, the admissible region *A* defined by $H(\boldsymbol{f})<T+\varepsilon$ is convex.
- Given an estimate vector $\boldsymbol{e}$ (that belongs to $A \backslash IA$) let $\boldsymbol{e}^*$ be the vector in *IA* that is closest to $\boldsymbol{e}$.
- We can construct a safe zone by taking the intersection of $H(\boldsymbol{e},\boldsymbol{e}^*)$ and *A*.

Monitoring Entropy (Cont')



Summary and Future Work

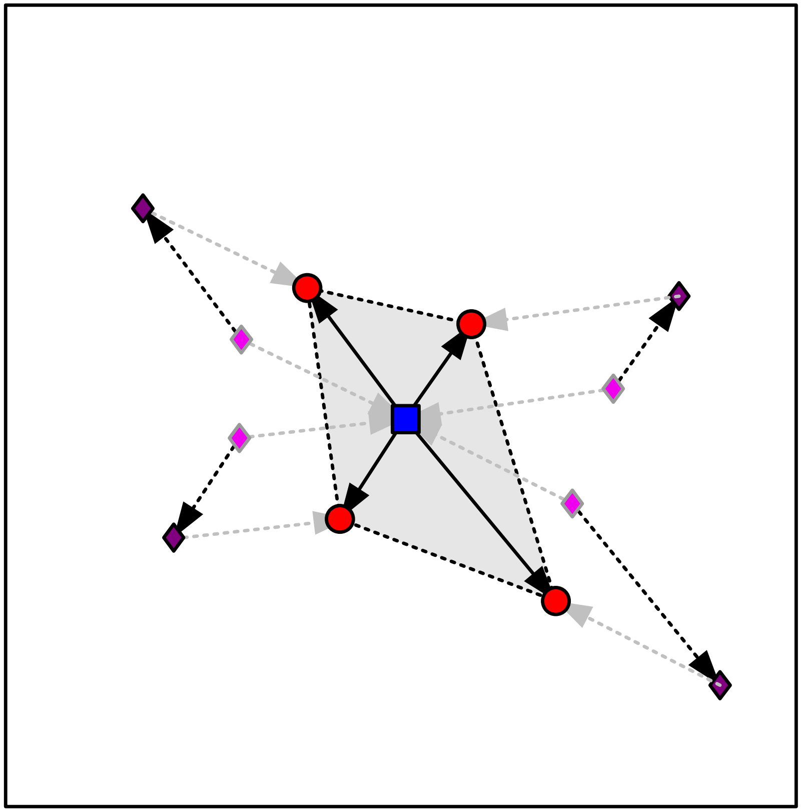
- Extend technique to general AMS sketches
- Extend technique to generic families of functions
- Refine and extend the notion of optimality

Backup Slides

Algorithmic Framework

- From time to time, as specified by the protocol, we will perform a synchronization procedure, which consists of:
  - Collecting local vectors from nodes. Value sent by the i^{th} vector is denoted by $\mathbf{v'_i}$.
  - Determining their average, denoted by $\mathbf{e}$.
  - Sending the average vector back to the nodes.
- After synchronization:
  - We set $x=f(\mathbf{e})$
  - At each node determine a vector $\boldsymbol{\delta}_\mathbf{i}= \mathbf{e}-\mathbf{v'_i}$. This value remains fixed until the next synchronization event.
  - Each node holds an additional vector $\mathbf{u_i}$ referred to as the drift vector, $\mathbf{u_i}=\mathbf{v_i}+\boldsymbol{\delta}_\mathbf{i}$
- At any time, the average of the drift vectors is equal to the average of data vectors. After synchronization drift vectors are equal to the estimate vector.

Algorithmic Framework (cont.)



AMS Sketches

- Given two (very) large *d* vectors, $\mathbf{v_a}$ and $\mathbf{v_b}$, their inner product $<\mathbf{v_a},\mathbf{v_b}>$ can be approximated as follows:
  - Given (ε,δ), let $n = O(1/\varepsilon^2)$ and $m = O(\ln(1/\delta))$
  - Each vector **v** is represented by a $m \times n$ matrix **S(v)** referred as the vector's sketch:
    - We use $m \times n$ four-wise independent *d* dimensional random vectors $\mathbf{r}_{i,j}$ where each component receives either 1 or -1 with equal probability
    - Given a vector **v** let $\mathbf{S(v)}_{i,j} = <\mathbf{v},\mathbf{r}_{i,j}>$
  - Let $\mathbf{S(v_a)} \bullet \mathbf{S(v_b)}$ denote the component-wise product of $\mathbf{S(v_a)}$ and $\mathbf{S(v_b)}$
  - The inner product of $\mathbf{v_a}$ and $\mathbf{v_b}$ is approximated by the median of the average of the rows of $\mathbf{S(v_a)} \bullet \mathbf{S(v_b)}$
- The AMS estimate has a error of $\varepsilon\|\mathbf{v_a}\|\cdot\|\mathbf{v_b}\|$ with probability of are least $1- \delta$