Distributed Query Monitoring through Convex Analysis: Towards Composable Safe Zones

Minos Garofalakis[1] **and Vasilis Samoladas**[1]

1 **Technical University of Crete**
{minos,vsam}@softnet.tuc.gr

Abstract

Continuous tracking of complex data analytics queries over high-speed distributed streams is becoming increasingly important. Query tracking can be reduced to continuous monitoring of a condition over the global stream. Communication-efficient monitoring relies on locally processing stream data at the sites where it is generated, by deriving site-local conditions which collectively guarantee the global condition. Recently proposed geometric techniques offer a generic approach for splitting an arbitrary global condition into local geometric monitoring constraints (known as "Safe Zones"); still, their application to various problem domains has so far been based on heuristics and lacking a principled, compositional methodology. In this paper, we present the first known formal results on the difficult problem of effective Safe Zone (SZ) design for complex query monitoring over distributed streams. Exploiting tools from convex analysis, our approach relies on an algebraic representation of SZs which allows us to: (1) Formally define the notion of a "good" SZ for distributed monitoring problems; and, most importantly, (2) Tackle and solve the important problem of systematically composing SZs for monitored conditions expressed as *Boolean formulas* over simpler conditions (for which SZs are known); furthermore, we prove that, under broad assumptions, the composed SZ is good if the component SZs are good. Our results are, therefore, a first step towards a principled compositional solution to SZ design for distributed query monitoring. Finally, we discuss a number of important applications for our SZ design algorithms, also demonstrating how earlier geometric techniques can be seen as special cases of our framework.

1998 ACM Subject Classification H.2.m, C.2.4

Keywords and phrases Distributed Data Streams, Geometric Method

1 Introduction

As we are moving from network-centric computing into the era of Internet of Things, large-scale event monitoring applications become ever more important. Such applications rely on *continuous* monitoring queries over the union of local, high-speed data streams. The scale of these applications, as well as power or bandwidth limitations, often impose critical communication contraints that prohibit the centralization of streaming data. Instead, the applications must rely on novel algorithmic paradigms for processing local streams of data *in situ* (i.e., locally at the sites where the data is observed). This obviously raises the problem of effectively decomposing the global monitoring query into "safe" local queries that can be tracked independently at each site while guaranteeing correctness for the global monitoring operation. Such a decomposition enables truly distributed, *push-based* monitoring, where sites track their local queries and communicate (e.g., with a "coordinator" site) only when some local query constraints are violated. Still, the problem of effectively decomposing

a complex (e.g., non-linear) query over the global distributed stream into such safe local constraints can be far from straightforward.

Problem Setup and the Geometric Method. In an abstract setting, our system architecture comprises a collection of k physically-distributed sites, where each site $p \in \{1, \ldots, k\}$ observes updates to the state of its local stream which is represented as a dynamic, high-dimensional vector $\boldsymbol{X}^{(p)} \in \mathbb{R}^m$. (Note that this is the standard model for general data streams used in the streaming algorithms literature, e.g., [26, 12].) The state of the global, distributed stream $\boldsymbol{X}$ is a convex combination of the local states; that is, $\boldsymbol{X} = \sum_p a_p \boldsymbol{X}^{(p)}$, with $a_p \geq 0, \sum_p a_p = 1$. Arbitrary linear combinations, e.g., summation of local frequency distribution vectors, can be captured by simply multiplying the local vectors by constant factors. In applications, these states often comprise of one or more frequency distributions of streams, or (linear) sketches thereof.

Let $F(\boldsymbol{X})$ denote a global query function, e.g., the norm or entropy of a dynamic global distribution, or the inner-product (i.e., equi-join size) of two underlying distributed streams $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ (note that in this case, $\boldsymbol{X} = \boldsymbol{X_1} \oplus \boldsymbol{X_2}$, the concatenation of $\boldsymbol{X_1}, \boldsymbol{X_2}$). A natural global monitoring condition is a *threshold query* $F(\boldsymbol{X}) < T$ (or, $> T$), where T is some constant. Threshold queries can naturally express more complex monitoring tasks, including *approximate function monitoring* [13].

A general approach to tracking threshold queries over distributed streams was pioneered by Sharfman et al.'s *Geometric Method* [32, 21]. For arbitrary $F()$, it is generally impossible to relate the locally-observed values of $F(\boldsymbol{X}^{(p)})$ to the global value $F(\boldsymbol{X})$; thus, their key idea is to employ geometric arguments to monitor the *domain* (rather than the range) of the monitored function $F()$. More formally, given the threshold query $F(\boldsymbol{X}) < T$, define the set $A \triangleq \{\boldsymbol{x} \in \mathbb{R}^m | F(\boldsymbol{x}) < T\} \subseteq \mathbb{R}^m$ as the query's *admissible region*. Clearly, the condition $F(\boldsymbol{X}) < T$ is equivalent to the condition $\boldsymbol{X} \in A$, and this geometric condition in $\mathbb{R}^m$ is the one being monitored — action needs to be taken only when $\boldsymbol{X}$ leaves A.

A key concept in geometric monitoring is that of a *Safe Zone (SZ)*, which is defined as a *convex subset* of the admissible region; that is, a SZ is a convex set Z such that $Z \subseteq A$ [21]. Let $\boldsymbol{X}^{(p)}(t_0)$ be the state of site p at some initial synchronization time t_0, and let $\boldsymbol{X}_0 = \sum_p a_p \boldsymbol{X}^{(p)}(t_0) = \boldsymbol{X}(t_0)$. As updates arrive at any site p, the site maintains its local *drift vector* $\boldsymbol{u}^{(p)}(t) = \boldsymbol{X}^{(p)}(t) - \boldsymbol{X}^{(p)}(t_0) + \boldsymbol{X}_0$. It is trivial to show that, at any time t, the convex combination of the local drift vectors is exactly the state of the global stream at time t; that is, $\sum_p a_p \boldsymbol{u}^{(p)}(t) = \boldsymbol{X}(t)$ [32]. Thus, as long as at every site p, we have $\boldsymbol{u}^{(p)} \in Z$, by convexity of Z we also have $\boldsymbol{X} \in Z$ and, therefore, $\boldsymbol{X} \in A$.

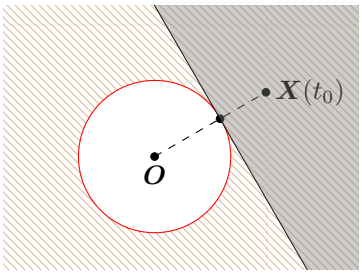

Figure 1 The admissible region (hatched) for $\|\boldsymbol{X}\| \geq T$ and a good safe zone (grayed).

The Safe Zone Design Problem. The Geometric Method has been extended and successfully applied to various monitoring problems in a number of recent papers [5, 15, 13, 20, 23, 27, 29]. A survey of this body of work reveals that a crucial, and non-trivial aspect of the technique is the issue of *Safe Zone Design:* Given a particular admissible region A, and the initial global state $\boldsymbol{X}(t_0) \in A$, define a "good" (convex) safe zone $Z \subseteq A$. As a simple example depicted in Fig. 1, if A is defined by constraint $\|\boldsymbol{X}\| \geq T$, a good SZ Z can be defined by the constraint $\boldsymbol{X} \cdot \boldsymbol{X}(t_0) \geq T\|\boldsymbol{X}(t_0)\|$.

For complex queries, safe zone design is often analytically challenging, and general methodologies are quite helpful. Simple solutions can be obtained using the original "covering spheres" method of Sharfman et al. [32], but the quality of the safe zones and the performance obtained can be far from satisfactory. A more recent work [23] introduces the "convex decomposition" method, which addresses some of the problems of "covering spheres" and provides effective solutions for several problems, but the method is lacking a systematic foundation.

The aforementioned methods, although valuable, suffer from two important drawbacks. First, they do not provide any systematic guidance for designing safe zones of provable quality, making the evaluation of a SZ design a purely experimental task. Second, and more important, they do not provide for composable designs. To clarify our (envisioned) concept of a composable SZ, we draw an analogy to the well-known chain rule for the differential operator: $D\big(f(g(x))\big) = Df(g(x))Dg(x)$. Ideally, we would wish for a "safe zone" transform, that, similar to D, would provide safe zones for complex queries by combining safe zones for simpler queries.

Our Results. In this paper, we present the first compositional method for SZ design, where the composed safe zones inherit quality features from their components. We apply our method to the important problem of inner-product queries (tracking the inner product of two vectors), both exactly and approximately, via AMS sketches.

We formally define the notion of a *good* SZ $Z \subseteq A$, for an admissible region A and a reference point $\boldsymbol{E} \in A$ ($\boldsymbol{E}$ is usually the initial system state $\boldsymbol{X}(t_0)$). A *good* SZ has two "largeness" properties; the *maximum distance* property states that the distance of $\boldsymbol{E} \in Z$ from the boundary of Z is equal to the distance of $\boldsymbol{E}$ from the boundary of A. The *maximality* property states that there does not exist a proper superset of Z which is also a safe zone for A. (Note that the SZ defined above for $\|\boldsymbol{X}\| \geq T$ clearly satisfies both properties.)

Our compositional method is primarily applicable to query functions with separable subcomponents. The global state vector $\boldsymbol{X}$ is taken to be the concatenation of n (not necessarily equidimensional) subvectors: $\boldsymbol{X} = \boldsymbol{X}_1 \oplus \boldsymbol{X}_2 \oplus \cdots \oplus \boldsymbol{X}_n$. For given safe zones on query functions $f_i(\boldsymbol{X}_i)$, we study the design of safe zones for a query function $F(f_1(\boldsymbol{X}_1), \ldots, f_n(\boldsymbol{X}_n))$. Our compositional approach relies on expressing the constraint on F as a conjunction of separable disjunctions. In particular, we present design methods for two important cases of aggregate functions F:

Boolean Functions: Here, f_i are boolean-valued functions, $F : \{0,1\}^n \to \{0,1\}$ is a boolean function, and we monitor the condition "F is true". Given *good* safe zones for subproblems $f_i(\boldsymbol{X}_i)$, our method can compose a *good* safe zone for the overall condition.

An important application for this type of query is threshold monitoring for the median (or, other order statistics); a query of the form $\text{median}\{g_1(\boldsymbol{X}_1), \ldots, g_n(\boldsymbol{X}_n)\} \leq T$ is equivalent to the case where each $f_i(\boldsymbol{X}_i)$ equals the boolean value of condition $g_i(\boldsymbol{X}_i) \leq T$, and F is the majority function. Such order statistics queries arise often in monitoring robust estimators, distributed voting schemes, and so on.

Separable Sums: Here, f_i are real-valued functions and F is summation; that is, we are interested in the condition $f_1(\boldsymbol{X}_1) + \cdots + f_n(\boldsymbol{X}_n) \leq T$. It is easy to show (e.g., by negating the above condition) that the above condition can be written equivalently as a conjunction of separable disjunctions:

$$\forall(\tau_1, \ldots, \tau_n) \in \Sigma_T^n \ : \ \bigvee_{j=1}^{n} f_j(\boldsymbol{X}_j) \leq \tau_j, \tag{1}$$

where $\Sigma_T^n = \{(\tau_1, \ldots, \tau_n) \in \mathbb{R}^n \mid \tau_1 + \cdots + \tau_n = T\}$.
For this more complicated problem, our method's scope is more limited; it can compose a safe zone with the maximum distance property, provided maximum-distance safe zones for queries of the form $f_i(\boldsymbol{X}_i) \leq \tau$ are known, but maximality is harder to obtain in general.

A Motivating Example. The problem of estimating the inner product of two vectors is of fundamental importance for distributed stream processing. This problem abstracts the situation where the vectors capture the (dynamic) frequency distribution of values in two distinct data streams, and we wish to monitor the degree of correlation in the two streams. (Note that this is also equivalent to tracking the size of the equi-join of the two streams [1].)

The global state comprises of a pair of vectors $(\boldsymbol{X}_1, \boldsymbol{X}_2)$, whose inner product, $\boldsymbol{X}_1 \cdot \boldsymbol{X}_2$, we wish to track. Often, the dimension of the raw streaming vectors can be too large for exact tracking to be realistic, since we can only afford to maintain a *synopsis/summary* of the streaming data. This problem has been addressed via the use of AGMS sketch synopses [2, 1]. Succinctly, the AGMS sketch of the frequency vector $\boldsymbol{X}_i, i = 1, 2$ is a sequence of d l-dimensional vectors $\hat{\boldsymbol{X}}_i = (\hat{\boldsymbol{x}}_{i,1}, \ldots, \hat{\boldsymbol{x}}_{i,d})$, with $\hat{\boldsymbol{x}}_{i,j} \in \mathbb{R}^l$. Then, as shown by Alon et al. [1], the inner product $\boldsymbol{X}_1 \cdot \boldsymbol{X}_2$ can be approximated, with an accuracy of $\epsilon = \|\hat{\boldsymbol{X}}_1\|\|\hat{\boldsymbol{X}}_2\|/\sqrt{l}$, with probability at least $1 - O(1/2^d)$, by $F(\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}_2) = \text{median}\{\ \hat{\boldsymbol{x}}_{1,1} \cdot \hat{\boldsymbol{x}}_{2,1}, \ \ldots, \hat{\boldsymbol{x}}_{1,d} \cdot \hat{\boldsymbol{x}}_{2,d}\}$.

In a distributed stream setting, the problem has been addressed in [8, 13]. Both of the above papers rely on a safe zone approach, but these safe zones have not been proven to be either maximum-distance, or maximal. None of these, or any other known techniques, provide guarantees on communication cost. Using our techniques, we design *good* safe zones for both exact and approximate tracking (via AGMS sketches) of the inner product. For exact tracking, the inner product query can be rewritten as a separable sum and the designed safe zone is good (maximum-distance and maximal). Furthermore, the designed safe zone is composed into a design of a *good* safe zone for approximate tracking (which is just tracking the median of d inner products).

2 Safe Zone Design

In this section, we introduce some basic notation and definitions, and give the initial mathematical formulation of safe zone representation and composition. Throughout the paper, we use the notation of vector calculus; boldface letters stand for vectors. All vector spaces in this paper are Euclidean. We use the letter V to denote a vector space $\mathbb{R}^d$, equipped with the standard inner product. Vector inner product is written as $\boldsymbol{xy}$ and self-product is written as squaring, therefore $\boldsymbol{x}^2 = \|\boldsymbol{x}\|^2$. We also use $\boldsymbol{x} \oplus \boldsymbol{y}$ to denote the so-called *direct sum* of two vectors, that is, the vector resulting from the concatenation of $\boldsymbol{x}$ and $\boldsymbol{y}$.

In addition, given a Boolean predicate $\phi : V \to \{0, 1\}$, we write $\{\phi\} = \{\boldsymbol{X} \in V \mid \phi(\boldsymbol{X})\}$.

We will need some simple topological concepts. Assume some vector space $V = \mathbb{R}^d$. Let $\text{Ball}(\boldsymbol{c}, \rho)$ denote the open ball centered at c with radius ρ, i.e., $\text{Ball}(\boldsymbol{c}, \rho) = \{\boldsymbol{x} \in$

$V \mid \|\boldsymbol{x} - \boldsymbol{c}\| < \rho\}$. Given a set of points $A \subseteq V$ and $\boldsymbol{x} \in A$, we say that $\boldsymbol{x}$ is *interior* to A (denoted $\boldsymbol{x} \in \mathbf{int}\, A$) iff there exists some $\epsilon > 0$, so that $\text{Ball}(\boldsymbol{x}, \epsilon) \subseteq A$. A point $\boldsymbol{x} \in V$ is *exterior* to A (denoted $\boldsymbol{x} \in \mathbf{ext}\, A$) iff it is interior to the complement of A, $\overline{A} = V - A$. A point $\boldsymbol{x} \in V$ which is neither exterior nor interior to A is a *boundary* point for A (denoted $\boldsymbol{x} \in \mathbf{bd}\, A$). Set A is *open* iff it only contains interior points; that is, $A = \mathbf{int}\, A$. Dually, set A is closed iff its complement $\overline{A}$ is open; in this case, $A = \mathbf{int}\, A \cup \mathbf{bd}\, A$.

For completeness, we also state a few basic facts about convex sets. A set $Z \subseteq V$ is convex iff $\forall \boldsymbol{x}, \boldsymbol{y} \in Z$, $\forall t \in [0,1]$, $(1-t)\boldsymbol{x} + t\boldsymbol{y} \in Z$. Alternatively, Z is convex iff it is closed under convex combinations of its elements. An important property of convex sets is that the intersection of any collection of convex sets is convex.

A (closed) halfspace h is a *supporting halfspace* of convex set Z, iff (a) $h \supseteq Z$, and (b) the boundaries of h and Z intersect in at least one point $\boldsymbol{p}$. We say that h supports Z at $\boldsymbol{p}$. At every boundary point $\boldsymbol{p} \in \mathbf{bd}\, Z$, there is at least one halfspace h supporting Z at $\boldsymbol{p}$. Where there is exactly one, $\boldsymbol{p}$ is called *smooth* and h is called *tangent*. A well-known theorem states that any convex set Z is the intersection of its supporting halfspaces. In this paper, we shall employ a stronger, but much less known theorem:

▶ **Theorem 1** (Rockafellar [28], Thm. 18.8)**.** *A closed convex set $Z \subseteq V$ is the intersection of the closed halfspaces tangent to it.*

A function $f : V \to \mathbb{R}$ is convex iff, for every $\boldsymbol{x}, \boldsymbol{y} \in V$ and $\lambda \in [0,1]$, it is $f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$. A function f is concave iff $-f$ is convex. We will be interested primarily in concave functions. A function f is both concave and convex, iff it is *affine*, that is, $f(\boldsymbol{x}) = \boldsymbol{w}\boldsymbol{x} + a$, for some $\boldsymbol{w} \in V$ and $a \in \mathbb{R}$.

2.1 Safe Zone Specification

Consider a monitored query on V, which corresponds to an admissible region $A \subseteq V$. The problem of determining a good safe zone Z requires additional information; some *reference* must be made by the user, as to the preferred locus of the safe zone, among mutually exclusive alternatives. Motivated by previous work, we adopt the concept of a reference point $\boldsymbol{E} \in \mathbf{int}\, A$, which our safe zone must include.

Therefore, the safe zone design problem can be stated as follows: given admissible region $A \subseteq V$ and reference point $\boldsymbol{E} \in \mathbf{int}\, A$, select a safe zone $Z \subseteq A$ such that Z is convex and $\boldsymbol{E} \in \mathbf{int}\, Z$. Towards a compositional approach, we assume that our admissible region A is expressible as a set-algebraic combination (involving intersection and union) of subsets of V.

Now, consider a family of sets $A_i \subseteq V$, $i \in I$, where $A = \cap_{i \in I} A_i$, with $\boldsymbol{E} \in \mathbf{int}\, A_i$ for all $i \in I$, and let $Z_i \subseteq A_i$ be safe zones. Towards a compositional approach, it is natural to consider $Z = \cap_{i \in I} Z_i$ as a candidate safe zone for admissible region $A = \cap_{i \in I} A_i$. Indeed, Z is convex and contains $\boldsymbol{E}$.

Similarly, if $A = \cup_{i \in I} A_i$, with $\boldsymbol{E} \in \mathbf{int}\, A_i$ for some $i \in I$ (note that $\boldsymbol{E}$ need not belong to all A_i), it is natural to consider $\cup_{i \in I} Z_i$ as a candidate safe zone for A. Unfortunately, this is not valid, as $\cup_i Z_i$ is *not* convex in general. Therefore, we need to restrict to a convex subset $Z \subseteq \cup_{i \in I} Z_i$, that contains $\boldsymbol{E}$; moreover, Z should inherit good qualities of zones Z_i. To overcome the difficulties of handling unions, our method focuses on the special case of decomposition into a *separable union*. In this case, $V = \mathbb{R}^d$ is the product space $V_1 \times \cdots \times V_n$ of a collection $V_i = \mathbb{R}^{d_i}, i = 1, \ldots, n$ of vector spaces (with $\sum d_i = d$), and each $\boldsymbol{x} \in V$ is the direct sum of $\boldsymbol{x}_i \in V_i$, where each $\boldsymbol{x}_i$ is the *projection* of $\boldsymbol{x}$ on V_i. For $A_i \subseteq V_i$, the separable union $\bigvee_{i=1}^n A_i$ is simply the set $\{\boldsymbol{x}_1 \oplus \cdots \oplus \boldsymbol{x}_n \in V \mid \bigvee_{i=1}^n \boldsymbol{x}_i \in A_i\}$. Equivalently, for $n = 2$,

$A_1 \vee A_2 = (A_1 \times V_2) \cup (V_1 \times A_2)$. Separable intersection $\bigwedge_{i=1}^n A_i$, defined similarly, is just the Cartesian product $A_1 \times \cdots \times A_n$.

Given safe zones $Z_i \subseteq A_i \subseteq V_i$, the separable union $\bigvee_{i=1}^n Z_i$ is not a convex set. However, it will be shown subsequently that properly selected subsets of the separable union do inherit the good qualities of the safe zones Z_i.

2.2 Safe Zone Representation and Composition

In our method, a convex set is represented as the level set of a concave function.

▶ **Definition 2** (Level set)**.** Let $f : V \to \mathbb{R}$ be any function. The level set $L(f)$ of f is the set $L(f) = \{\boldsymbol{x} \in V \mid f(\boldsymbol{x}) \geq 0\} = \{f(\boldsymbol{x}) \geq 0\}$.

When f is concave, $L(f)$ is convex. We wish to avoid the degenerate cases where $L(f)$ is empty, or the whole space, or has empty interior. This is equivalent to the following requirement:

▶ **Definition 3** (Safe Zone function)**.** A safe zone function is a concave function $f : V \to \mathbb{R}$, which attains both a positive and a negative value over V.

Safe zone functions are positive in the interior of $L(f)$, negative on the exterior and vanish on the boundary. An important property of $L(f)$ is monotonicity with respect to pointwise dominance. Given functions $f, g : V \to \mathbb{R}$, f is dominated by g (denoted by $f \leq g$) iff $\forall \boldsymbol{x} \in V, f(\boldsymbol{x}) \leq g(\boldsymbol{x})$. Clearly, $f \leq g$ directly implies $L(f) \subseteq L(g)$.

We are interested in the compositions of safe zones for intersections and unions of safe zones. We introduce two operations that construct the composite safe zone function from component safe zone functions. The first operation, used to capture the intersection of safe zones, is the pointwise-infimum of a (possibly infinite) family of safe zone functions. Given family of safe zone functions $\zeta_i : V \to \mathbb{R}$, $i \in I$, let $\zeta(\boldsymbol{x}) = \inf_{i \in I} \zeta_i(\boldsymbol{x})$. Then, $L(\zeta) = \bigcap_{i \in I} L(\zeta_i)$. The second operation is weighted sum with non-negative weights, sometimes called conical combination. It is used to construct a convex subset of the union of a finite family of safe zones.

▶ **Theorem 4.** *For safe zone functions $\zeta_i : V \to \mathbb{R}$, $i = 1, \ldots, n$, and reals $a_i \geq 0$, not all zero, let*

$$\zeta(\boldsymbol{X}) = \sum_{i=1}^{n} a_i \zeta_i(\boldsymbol{X}). \tag{2}$$

Then, $L(\zeta)$ is convex and $\bigcap_{i=1}^n L(\zeta_i) \subseteq L(\zeta) \subseteq \bigcup_{i=1}^n L(\zeta_i)$.

Proof. It is $\bigcap_{i=1}^n L(\zeta_i) \subseteq L(\inf_i a_i \zeta_i(\boldsymbol{X}))$ and $L(\sup_i a_i \zeta_i(\boldsymbol{X})) \subseteq \bigcup_{i=1}^n L(\zeta_i)$ (with equality holding when all $a_i > 0$). Since $n \inf_i a_i \zeta_i(\boldsymbol{X}) \leq \zeta(\boldsymbol{X}) \leq n \sup_i a_i \zeta_i(\boldsymbol{X})$, the theorem follows directly from these formulas and monotonicity of L. ◀

The above theorem specializes to separable union straightforwardly. If $\zeta_i : V_i \to \mathbb{R}$ are safe zone functions, then $\zeta(\boldsymbol{x}_1 \oplus \cdots \oplus \boldsymbol{x}_n) = \sum_{i=1}^n a_i \zeta_i(\boldsymbol{x}_i)$ is a safe zone function and $\times_{i=1}^n L(\zeta_i) \subseteq L(\zeta) \subseteq \bigvee_{i=1}^n L(\zeta_i)$.

2.3 Functional Analysis For Safe Zone Functions

Having defined the two fundamental operations on safe zone functions, in the following sections we proceed to study the conditions under which these operations maintain the qualities of composed safe zones. We end this section with some foundational facts from convex analysis.

All safe zone functions over V are continuous and differentiable almost everywhere in V (that is, everywhere, except for a set of measure 0). The gradient $\nabla\zeta(\boldsymbol{x}) = (\frac{\partial\zeta}{\partial x_1}, \ldots, \frac{\partial\zeta}{\partial x_d})$ is the multi-dimensional analog of the derivative. At any point $\boldsymbol{x}$, where ζ is differentiable, this derivative is a vector $\boldsymbol{v_x}$, pointing to the direction of maximum increase of ζ, and its norm $\|\boldsymbol{v_x}\|$ is proportional to the rate of change.

Let a safe zone function ζ be differentiable at point $\boldsymbol{x_0}$. The affine function $h(\boldsymbol{x}) = \nabla\zeta(\boldsymbol{x_0})(\boldsymbol{x} - \boldsymbol{x_0}) + \zeta(\boldsymbol{x_0})$ is called the *tangent* of ζ at $\boldsymbol{x_0}$. By virtue of concavity, $\zeta \leq h$.

In general, every safe zone function ζ is the pointwise infimum of some (non-unique) family $\mathcal{H}$ of affine functions. This is denoted by $\zeta = \inf \mathcal{H}$. In particular, ζ is the pointwise infimum of all its tangents.

Given a collection ζ_i, $i \in I$, of safe zone functions, and corresponding families of affine functions $\mathcal{H}_i$, such that $\zeta_i = \inf \mathcal{H}_i$, the following two properties are very important in the rest of this paper:

1. For arbitrary I, $\inf_{i\in I} \zeta_i = \inf(\bigcup_{i\in I} \mathcal{H}_i)$, and
2. For I finite, and $a_i \geq 0$, not all 0, it is $\sum_{i\in I} a_i\zeta_i = \inf(\sum_{i\in I} a_i\mathcal{H}_i)$, where $\sum_{i\in I} a_i\mathcal{H}_i$ is the set of affine functions $\{\sum_{i\in I} a_i h_i \mid h_i \in \mathcal{H}_i, i \in I\}$.

3 Maximum Distance

Given an admissible region $A \subseteq V$ and reference point $\boldsymbol{E} \in \mathbf{int}\, A$, several safe zones containing $\boldsymbol{E}$ can be constructed. Given two such safe zones, Z and Z', one can argue that one is "better" than the other, by assuming isotropy; that is, all directions *around* the reference point are equally desirable. This assumption correlates with a monitoring situation where each coordinate of the state vector behaves (or is assumed to behave) as an IID random walk. Under this assumption, we have the following criterion:

▶ Criterion 1. Safe zone Z is "better" than safe zone Z' if $\mathrm{dist}(\boldsymbol{E}, \overline{Z}) \geq \mathrm{dist}(\boldsymbol{E}, \overline{Z'})$.

With respect to criterion 1, any safe zone containing a ball that touches the admissible region's boundary is best possible.

▶ **Definition 5** (Maximum distance)**.** Let A be an admissible region and $\boldsymbol{E} \in \mathbf{int}\, A$. A safe zone $Z \subseteq A$ has maximum distance in A with respect to $\boldsymbol{E}$, iff $\mathrm{dist}(\boldsymbol{E}, \overline{Z}) = \mathrm{dist}(\boldsymbol{E}, \overline{A})$.

3.1 Preservation Of Maximum Distance Under Composition

The intersection operation always preserves the maximum distance of its operands. Given safe zones $Z_i \subseteq A_i$, and $\boldsymbol{E} \in \mathbf{int}\, A_i$, for each i, if all Z_i are maximum distance, then $\bigcap_{i\in I} Z_i$ is also maximum distance. As a matter of fact, if $D = \mathrm{dist}(\boldsymbol{E}, \overline{A}) = \inf_{i\in I} \mathrm{dist}(\boldsymbol{E}, \overline{A_i})$, it is sufficient and necessary to have $\mathrm{dist}(\boldsymbol{E}, \overline{Z_i}) \geq D$. Consequently, the pointwise-inf operation on any family of safe zone functions preserves the maximum distance property of its operands.

The situation with separable union is much more involved. It is not sufficient for the safe zones of the operands to be maximum distance; the actual safe zone functions that describe these safe zones must carry sufficient distance information, so that some conical combination

can yield a maximum-distance subset of the union. This is best illustrated by example. With reference to Fig. 2, consider the case in $\mathbb{R}^2$, where $Z_1 = L(f(x_1))$ and $Z_2 = L(f(x_2))$, with

$$f(x) = \begin{cases} x/\lambda & \text{if } x \geq 0 \\ \lambda x & \text{if } x < 0 \end{cases} \tag{3}$$

where $\lambda \geq 1$ (so that f is concave). Note that, independently of the value of λ, $Z_i = \{x_i \geq 0\}$. The union $Z_1 \vee Z_2$ is not convex, but maximal convex subsets are all halfspaces whose boundary supports the positive quadrant $(Z_1 \times Z_2)$. Yet, none of the sets $L(a_1 f(x_1) + a_2 f(x_2))$ have maximum distance for any reference point in $Z_1 \times Z_2$, unless $\lambda = 1$. In fact, as λ grows, safe zones $L(a_1 f(x_1) + a_2 f(x_2))$ shrink towards $Z_1 \cap Z_2$, which is their lower bound. The

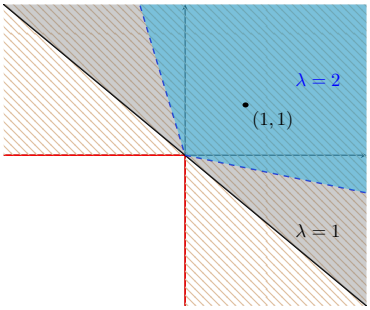

Figure 2 Example of the safe zone of union $\{x_1 \geq 0\} \vee \{x_2 \geq 0\}$, derived by with suboptimal functions. The safe zone for $\lambda = 1$, whose boundary is the solid black line, is good for reference point $(1, 1)$. The safe zone for $\lambda = 2$, whose boundary is the dashed blue line, is neither maximum-distance nor maximal.

intuitive reason of the failure in this example is that, the shape of the region generated by Eq. (2) depends crucially on the values of f outside of $L(f)$.

To ameliorate this situation, we introduce a class of safe zone functions which contain sufficient distance information.

▶ **Definition 6** (Affine Distance Function (ADF))**.** An affine function $h : V \to \mathbb{R}$ with $h(\boldsymbol{x}) = \boldsymbol{w}\boldsymbol{x} + a$ is an ADF iff $\|\nabla h\| = \|\boldsymbol{w}\| = 1$.

Every closed halfspace of V is equal to $L(h)$ for a unique ADF h. For each $\boldsymbol{x} \in V$, $h(\boldsymbol{x})$ is the signed distance of $\boldsymbol{x}$ from the boundary of $L(h)$, non-negative if $\boldsymbol{x} \in L(h)$ and negative if $\boldsymbol{x} \notin L(h)$.

▶ **Definition 7** (Eikonal function)**.** A safe zone function ζ is eikonal iff it is the pointwise-infimum of a collection of ADFs.

A useful alternative characterization of eikonal functions is the following:

▶ **Theorem 8** (Eikonal characterization)**.** *Let $\zeta : V \to \mathbb{R}$ be concave, and almost everywhere differentiable. Then, ζ is eikonal, if and only if, $\|\nabla\zeta\| = 1$ at every point where it is differentiable.*

(Due to space constraints, this and other ommitted proofs will appear in the full version of the paper.)

The equation $\|\nabla\zeta\| = 1$ is known as the (Euclidean) *eikonal* differential equation. As a consequence of this equation, it can be shown, using the mean-value theorem of analysis, that every eikonal function ζ is non-expansive: $|\zeta(\boldsymbol{x}) - \zeta(\boldsymbol{y})| \leq \|\boldsymbol{x} - \boldsymbol{y}\|$.

Signed Distance Functions. One important member in the family of eikonal functions is the Signed Distance Function (SDF) of a convex set Z.

$$\delta_Z(\boldsymbol{x}) = \begin{cases} \operatorname{dist}(\boldsymbol{x}, \overline{Z}) & \text{if } \boldsymbol{x} \in Z, \\ -\operatorname{dist}(\boldsymbol{x}, Z) & \text{if } \boldsymbol{x} \in \overline{Z}. \end{cases} \tag{4}$$

The SDF of a convex set is concave. Also, since $L(\delta_Z) = Z$, the SDF is positive in the interior of Z, negative on the exterior, and vanishes on the boundary; therefore, it is a safe zone function. Naturally, every ADF h is the SDF of $L(h)$. However, this is not the case for every eikonal function ζ.

Let $Z = L(\zeta)$ and let δ_Z be the SDF of $L(Z)$. Then, $\delta_Z \le \zeta$. In fact, for every $\boldsymbol{x} \in L(f)$, $\zeta(\boldsymbol{x}) = \delta_Z(\boldsymbol{x})$. But, ζ may strictly dominate δ_Z outside $L(\zeta)$.

The family of eikonal functions is well-behaved under our composition operators. The pointwise-infimum of any family of eikonal functions is also eikonal. For the case of separable union, we have the following:

▶ **Theorem 9.** *Let $\zeta_i : V_i \to \mathbb{R}$ be eikonal functions, $a_i \ge 0$, and $\zeta(\boldsymbol{x}_1 \oplus \cdots \oplus \boldsymbol{x}_n) = \sum_{i=1}^n a_i \zeta_i(\boldsymbol{x}_i)$. Then, ζ is eikonal iff $\sum_{i=1}^n a_i^2 = 1$.*

Proof. Since $\nabla\zeta_i \in V_i$ are orthogonal, $\|\nabla\zeta\|^2 = \left(\sum_{i=1}^k a_i \nabla\zeta_i\right)^2 = \sum_{i=1}^k a_i^2 \|\nabla\zeta_i\|^2 = \sum_{i=1}^k a_i^2$. ◀

Thus, any separable conical combination ζ of eikonal functions can always be scaled to an eikonal function, by dividing it by $\sqrt{\sum_i a_i^2}$, which of course does not affect the described safe zone $L(\zeta)$.

We now turn our attention to separable union. Let admissible region $A = \bigvee_{i=1}^n A_i$ and let $\boldsymbol{E} = \boldsymbol{E_1} \oplus \cdots \oplus \boldsymbol{E}_n$ be the reference point. Consider eikonal functions ζ_i such that $L(\zeta_i) \subseteq A_i$, and let $Z = \bigvee_{i=1}^n L(\zeta_i)$. What is the radius D of the largest ball centered at $\boldsymbol{E}$, that can be attained by a conical combination of ζ_i? Clearly, $D \le d_Z = \operatorname{dist}(\boldsymbol{E}, \overline{Z})$. In general, it is $d_Z \le d_A = \operatorname{dist}(\boldsymbol{E}, \overline{A})$. The following theorem specifies the conditions under which maximum distance can be achieved.

▶ **Theorem 10.** *Let $A = \bigvee_{i=1}^n A_i$, where $A_i \subseteq V_i$, and $\boldsymbol{E} = \boldsymbol{E_1} \oplus \cdots \oplus \boldsymbol{E}_n \in \mathbf{int}\, A$. Let $\zeta_i : V_i \to \mathbb{R}$ be eikonal functions, such that $L(\zeta_i) \subseteq A_i$, and let $Z = \bigvee_{i=1}^n L(\zeta_i)$. Then:*

1. *For any conical combination $\zeta = \sum a_i \zeta_i$, it is $\operatorname{dist}(\boldsymbol{E}, \overline{L(\zeta)}) = \dfrac{\sum_{i=1}^n a_i \zeta(\boldsymbol{E}_i)}{\sqrt{\sum_{i=1}^n a_i^2}} = D$.*
2. *It is $D \le \operatorname{dist}(\boldsymbol{E}, \overline{Z}) = d_Z$, with equality holding iff, for some $\lambda > 0$,*

$$a_i = \begin{cases} \lambda \zeta_i(\boldsymbol{E}_i) & \text{if } \zeta_i(\boldsymbol{E}_i) > 0, \\ 0 & \text{otherwise}, \end{cases}$$

3. *It is $d_Z \le d_A = \operatorname{dist}(\boldsymbol{E}, \overline{A})$, with equality holding iff, for every i such that $\boldsymbol{E}_i \in \mathbf{int}\, A_i$, $L(\zeta_i)$ has maximum distance in A_i w.r.t. $\boldsymbol{E}_i$.*

4 Maximality

During distributed monitoring using a safe zone $Z \subseteq A$, condition $\boldsymbol{X} \in Z$ may be violated, while $\boldsymbol{X} \in A$. We call such local violations, *false violations*. Typically, the performance of distributed monitoring depends crucially on minimizing false violations. Therefore,

▶ Criterion 2. Safe zone Z is "better" that safe zone Z' if $Z \supseteq Z'$

This criterion is a rather obviously desirable; providing a larger safe zone will tend, other things being equal, to reduce false violation.

With respect to criterion 2, the best possible for a safe zone is to be a $\subseteq$-wise maximal convex subset of the admissible region A.

▶ **Definition 11** (Maximality)**.** Let A be the admissible region. A safe zone Z is maximal in A (with respect to set containment), if and only if, no convex subset of A is a proper superset of Z.

We now develop a convenient characterization of maximality.

▶ **Definition 12** (Flats of an affine family)**.** Given a family $\mathcal{H}$ of affine functions on V, let $\Phi_{\mathcal{H}} : \mathcal{H} \to 2^V$ be

$$\Phi_{\mathcal{H}}(h) = \{\boldsymbol{p} \in V \mid h(\boldsymbol{p}) = 0 \text{ and } \forall h' \in \mathcal{H}, h' \neq h \implies h'(\boldsymbol{p}) > 0\} \tag{5}$$

Set $\Phi_{\mathcal{H}}(h)$ is the flat of h in $\mathcal{H}$.

Fix some affine family $\mathcal{H}$ and let $\zeta = \inf \mathcal{H}$ and $Z = L(\zeta)$. Let $h \in \mathcal{H}$ be an affine function with non-empty flat $\Phi(h)$. Each $\boldsymbol{p} \in \Phi(h)$ is a boundary point of Z, since $\zeta(\boldsymbol{p}) = 0$. Also, $\boldsymbol{p}$ is smooth, since ζ is differentiable at $\boldsymbol{p}$ (with $\nabla\zeta(\boldsymbol{p}) = \nabla h$). Therefore, h is a tangent halfspace to Z.

Some examples of flats; a ball has a flat for each point on its boundary. A planar triangle has three flats, each corresponding to a side minus the corners (which are not smooth). A cylinder in 3-d has two 2-d flats, its top and bottom (minus the edges) and infinite 1-d flats which are segments from top to bottom (minus the endpoints). Finally, the positive quadrant in $\mathbb{R}^2$ has two flats, the rays $(0, +\infty) \times \{0\}$ and $\{0\} \times (0, +\infty)$.

▶ **Definition 13** (Non-redundant affine family)**.** A family $\mathcal{H}$ of affine functions is non-redundant iff, for every $h \in \mathcal{H}$, $\Phi_{\mathcal{H}}(h) \neq \emptyset$.

According to the above, a non-redundant affine family $\mathcal{H}$ contains only tangent halfspaces (represented as affine functions) of $Z = L(\inf \mathcal{H})$. However, not all tangents of Z need be contained in $\mathcal{H}$. As an example, consider the planar unit disk $Z = \{x_1^2 + x_2^2 \leq 1\}$; although there is a tangent at every point of the unit circle, only a countable family of tangent halfspaces (say, those whose slope is rational) is enough to define it!

▶ **Definition 14** (Witness)**.** For admissible region $A \subseteq V$ and an affine family $\mathcal{H}$, so that $L(\inf \mathcal{H}) \subseteq A$, a point $\boldsymbol{p} \in \mathbf{bd}\, A$ is a witness iff, for some $h \in \mathcal{H}$, $\boldsymbol{p} \in \Phi_{\mathcal{H}}(h)$.

▶ **Theorem 15** (Witnessed maximality criterion)**.** *For admissible region $A \subseteq V$ and affine family $\mathcal{H}$, so that $Z = L(\inf \mathcal{H}) \subseteq A$, if every flat of $\mathcal{H}$ contains a witness, then Z is maximal in A.*

Witnessed maximality is sufficient for maximality, but not necessary. As an example, consider the 2-d case where $Z = \{y \leq 0\}$ and $A = \{y \leq 1/|x| \vee x = 0\}$. Then, Z is maximal in A, but there is no witness to the unique tangent halfspace that is the whole of Z.

The witnessed maximality criterion is handy, because it relates maximality of a set Z to a requirement on each flat of a description of Z. It is possible, but cumbersome, to extend the concept of a witness, so that a sufficient *and* necessary condition for maximality can be obtained. Because of space constraints, this extension will be presented in the full paper.

4.1 Preservation Of Maximality Under Composition

In contrast to maximum distance, intersection does not preserve maximality in general. This is quite well-known; in fact, loss of maximality under intersection is the reason for the unsatisfactory behavior of previous safe zone design approaches, such as the Covering Spheres method.

Fortunately, under suitable conditions to be explored below, separable union does preserve maximality; given maximal safe zones $L(\zeta_i) \subseteq A_i$, it is possible to select a maximal convex subset of $\bigvee L(\zeta_i)$ which is also maximal in $\bigvee A_i$. However, as in the case for maximum distance, the safe zone functions ζ_i must a special requirement, *non-redundancy*.

▶ **Definition 16** (Non-redundant safe zone function)**.** A safe zone function ζ is non-redundant iff ζ is the pointwise infimum of a non-redundant affine family; else it is redundant.

Intuitively, a non-redundant safe zone function ζ for $Z = L(\zeta)$ is one which is (pointwise) maximal, among all safe zone functions g with $L(g) = Z$; that is, if $L(f) = L(g)$ and $f < g$ (that is, $f \leq g$ and at some $\boldsymbol{x}_0$, $f(\boldsymbol{x}_0) < g(\boldsymbol{x}_0)$), then f is redundant.

An important observation pertains to eikonal non-redundant functions. Given any convex set Z, the family $\mathcal{F}$ of eikonal functions f with $L(f) = Z$ is known to contain a $\leq$-wise least element, the SDF of Z. It can be shown that it also contains a unique $\leq$-wise greatest element η_Z, which is the (unique) non-redundant function in $\mathcal{F}$. Importantly, when Z is smooth (all points of its boundary are smooth), then $\eta_Z = \delta_Z$.

We now proceed to the main result of this section, which determines maximality of safe zones for an admissible region composed as an intersection of unions. To describe the intersection of unions, we use an *antichain* on $[n] = \{1, \ldots, n\}$, i.e., a non-empty collection C of subsets of $[n]$, such that no element of C is a subset of another. A single separable union is specified $C = \{[n]\}$. A (separable) intersection on the other hand is specified as $C = \big\{\{1\}, \ldots, \{n\}\big\}$.

▶ **Theorem 17.** *Let $\zeta_i : V_i \to \mathbb{R}$, $i = 1, \ldots, n$ be non-redundant safe zone functions, and C an antichain on $[n]$. Let $\zeta = \inf_{\Gamma \in C} \sum_{i \in \Gamma} a_i(\Gamma)\zeta_i$, where $a_i(\Gamma) > 0$. Then,*

1. *ζ is non-redundant, and*
2. *if, for $A_i \subseteq V_i$, each $L(\zeta_i)$ is witnessed maximal in A_i, then $L(\zeta)$ is witnessed maximal in $A = \bigcap_{\Gamma \in C} \bigvee_{i \in \Gamma} A_i$.*

5 Safe Zones For Boolean Functions

We now apply our method to the class of (separable) boolean query functions. Let $F : \{0,1\}^n \to \{0,1\}$ be any boolean function and let $f_i : V_i \to \{0,1\}$ be predicates on V_i respectively. We are interested in a safe zone for admissible region $A = \{\boldsymbol{X} \in V \mid F(f_1(\boldsymbol{X}_1), \ldots, f_n(\boldsymbol{X}_n))\}$, containing the reference point $\boldsymbol{E} = \boldsymbol{E}_1 \oplus \cdots \oplus \boldsymbol{E}_n$, with $\boldsymbol{E} \in \mathbf{int}\, A$.

We assume that F is given to us as a conjunction of clauses, where each clause is a disjunction of literals $b_i, \overline{b_i}$, for $i = 1, \ldots, n$. A clause Γ can be represented as a subset of $\{1, \ldots, n\} \times \{+, -\}$, where pair $(j, s) \in \Gamma$ means that the clause contains b_j if $s = +$, or $\overline{b_j}$ if $s = -$.

Letting, $A_i^{(+)} = \{f_i(\boldsymbol{X}_i)\}$ and $A_i^{(-)} = V_i - A_i^{(+)}$, the admissible region A can be decomposed as $A = \bigcap_{\Gamma \in F} \bigvee_{(i,s) \in \Gamma} A_i^{(s)}$. By the observations for maximum distance, we are allowed to *reduce* F to a (stronger) boolean function $\tilde{F}$. In particular, for each $i = 1, \ldots, n$, we eliminate at least one of the literals $b_i, \overline{b_i}$ from all clauses: if $\boldsymbol{E}_i \notin \mathbf{int}\, A_i^{(-)}$, we eliminate

literal $\overline{b_i}$, and if $\boldsymbol{E}_i \notin \mathbf{int}\, A_i^{(+)}$ we eliminate b_i. This justified is because, by Thm. 10, a maximum-distance safe zone for each clause would eliminate the corresponding components. Once $\tilde{F}$ is obtained, it can be further reduced by expressing it as a conjunction of its prime implicates, so that no clause is weaker than another. This last step is needed in order to apply Thm. 17.

Then, given safe zone functions ζ_i for the remaining admissible regions on V_i, the safe zone function

$$\zeta(\boldsymbol{X}_1 \oplus \ldots \oplus \boldsymbol{X}_n) = \inf_{\Gamma \in \tilde{F}} \frac{\sum_{(i,s)\in\Gamma} \zeta_i(\boldsymbol{E}_i)\zeta_i(\boldsymbol{X}_i)}{\sqrt{\sum_{(i,s)\in\Gamma} \zeta_i^2(\boldsymbol{E}_i)}} \tag{6}$$

defines a safe zone for A. Furthermore, if, for all i, ζ_i are eikonal and $L(\zeta_i)$ are maximum distance, then $L(\zeta)$ is maximum distance and ζ is eikonal. With respect to maximality, by virtue of Thm. 17, if all ζ_i are non-redundant, and $L(\zeta_i)$ are (witnessed) maximal in A_i, then $L(\zeta)$ is also (witnessed) maximal in A.

5.1 Monitoring Quantiles

Consider the query function $Q_k(g_1(\boldsymbol{X}_1), \ldots, g_n(\boldsymbol{X}_n))$, where $Q_k : \mathbb{R}^n \to \mathbb{R}$ returns the k-th least value among its arguments. These queries arise in monitoring robust statistics of (functions of) the state, such as the median or the inter-quartile range.

Condition $Q_k() \leq T$ can be written as a boolean function $F_k(f_1(\boldsymbol{X}_1), \ldots, f_n(\boldsymbol{X}_n))$, where $f_i(\boldsymbol{X}_i)$ is the boolean value of "$g_i(\boldsymbol{X}_i) \leq T$" and F_k is true iff k or more of its inputs are true. With respect to F_k, the safe zone design of this section yields a safe zone function ζ, given safe zone functions ζ_i for each constraint $g_i(\boldsymbol{X}_i) \leq T$.

A practical point is related to the computational cost of checking membership in $L(\zeta)$, which, if done straightforwardly, requires time $O(2^n)$. For counting queries, it is possible to test condition $\zeta(\boldsymbol{X}) \geq 0$ in time $O(n)$. To see this, note that the set of clauses of $F_k(b_1, \ldots, b_n)$ is the set of all $n-k+1$-subsets of $\{b_1, \ldots, b_n\}$ (since, if every $n-k+1$-subset of literals contains a true literal, there are at most $n-k$ false literals overall, and therefore at least k true literals). Therefore, to check $\zeta(\boldsymbol{X}) \geq 0$, it is sufficient to compute the sum S of the least $n-k+1$ elements of $\{\zeta_i(\boldsymbol{E}_i)\zeta_i(\boldsymbol{X}_i) \mid i = 1, \ldots, n\}$, as it is $\zeta(\boldsymbol{X}) \geq 0$ iff $S \geq 0$.

6 Safe Zones For Separable Sums

Separable sum queries refer to conditions of the form $\sum_{i=1}^n f_i(\boldsymbol{X}_i) \leq T$, where f_i are arbitrary real functions. As shown in Eq. 1, this condition can be written as a universal quantification (which relates to conjunction) of a family of finite disjunctions. Therefore, at least in principle, our composition operators can be used to derive a safe zone formula. In this section, we demonstrate that this approach can go well beyond the principle, into an analytic method of safe zone design.

The approach is straightforward; as shown in Eq. 1, the separable sum threshold condition is rewritten as $\forall \boldsymbol{\tau} \in \Sigma_T^n$, $\bigvee_{i=1}^n f_i(\boldsymbol{X}_i) \leq \tau_i$. Let the reference point be $\boldsymbol{E} = \boldsymbol{E}_1 \oplus \cdots \oplus \boldsymbol{E}_n$. The booleanized condition decomposes as the intersection of a family $A_{\boldsymbol{\tau}} \subseteq V$, for $\boldsymbol{\tau} \in \Sigma_T^n$, of admissible regions, each corresponding to a disjunctive clause. In order to write a safe zone function for each clause, we need to have a parametrized family of safe zone functions for conditions $f_i(\boldsymbol{X}_i) \leq \tau_i$. Let $\zeta_i(\boldsymbol{X}_i; \tau_i)$, $i = 1, \ldots, n$ denote such a parameterized family. Then, a safe zone for $A_{\boldsymbol{\tau}}$ can be given by function $\zeta_{\boldsymbol{\tau}}(\boldsymbol{X}) = \sum_{i=1}^n a_i(\boldsymbol{\tau})\zeta_i(\boldsymbol{X}_i; \tau_i)$, for suitably selected $a_i(\boldsymbol{\tau})$. Finally, the overall safe zone function is $\zeta(\boldsymbol{X}) = \inf_{\tau \in \Sigma_T^n} \zeta_\tau(\boldsymbol{X})$.

As a first example, we solve a trivial problem; the linear constraint $\sum_{i=1}^{n} w_i x_i = \boldsymbol{w}\boldsymbol{x} \leq T$. Booleanization gives $\forall \boldsymbol{\tau} \in \Sigma_T^n : \bigvee_{i=1}^{n} w_i x_i \leq \tau_i$. Each constraint $w_i x_i \leq \tau_i$ has a good safe zone described by the eikonal and non-redundant affine function $\zeta_i(x_i; \tau_i) = (\tau_i - w_i x_i)/|w_i|$. Therefore, we have an overall expression of

$$\zeta(\boldsymbol{x}) = \inf_{\boldsymbol{\tau} \in \Sigma_T^n} \zeta_{\boldsymbol{\tau}}(\boldsymbol{x}) = \inf_{\boldsymbol{\tau} \in \Sigma_T^n} \sum_{i=1}^{n} a_i(\boldsymbol{\tau}) \frac{\tau_i - w_i x_i}{|w_i|}.$$

Although any choice for $a_i(\boldsymbol{\tau})$ will yield a legal safe zone, to obtain a good solution, we need to select $a_i(\boldsymbol{\tau})$ more carefully. But, notice that if we select $a_i(\boldsymbol{\tau}) = |w_i|/\|\boldsymbol{w}\|$, we obtain $\zeta_{\boldsymbol{\tau}}(\boldsymbol{x}) = \frac{T - \boldsymbol{w}\boldsymbol{x}}{\|\boldsymbol{w}\|}$, which is independent of $\boldsymbol{\tau}$. Therefore, the inf operator becomes redundant. It is easy to see that the solution obtained is best possible; it is less clear whether there is a systematic strategy for selecting $a_i(\boldsymbol{\tau})$, for more complex problems. Below we introduce two such systematic strategies.

6.1 Dominating Index

It turns out that the previous example exhibits a *dominating index*. Assume that every $\zeta_i(\boldsymbol{x}_i; \tau_i)$ is eikonal and maximum distance. Define function

$$\Delta(\boldsymbol{\tau}) = \sqrt{\sum_{i=1}^{n} \max\big(\zeta_i(\boldsymbol{E}_i; \tau_i), 0\big)^2} = \operatorname{dist}(\boldsymbol{E}, \overline{A_{\boldsymbol{\tau}}})$$

where the second equality is a consequence of Thm. 10. Assume that Δ minimizes at a unique index $\boldsymbol{\tau}^*$, where, naturally, $\Delta(\boldsymbol{\tau}^*)$ is $\operatorname{dist}(\boldsymbol{E}, \overline{A})$. We can select $a_i(\boldsymbol{\tau}^*)$ as per Thm. 10, that is, $a_i(\boldsymbol{\tau}^*) = \zeta_i(\boldsymbol{E}_i; \tau_i^*)/\Delta(\boldsymbol{\tau}^*)$, and obtain $\zeta_{\boldsymbol{\tau}^*}$. Now, if it so happens that $L(\zeta_{\boldsymbol{\tau}^*}) \subseteq A$, then, we say that $\boldsymbol{\tau}^*$ is a *dominating index*: we can select $\zeta = \zeta_{\boldsymbol{\tau}^*}$, eliminating the inf operation. The overall safe zone $L(\zeta)$ is obviously maximum distance. Also, if $L(\zeta_{\boldsymbol{\tau}^*})$ is maximal in $A_{\boldsymbol{\tau}^*}$, $L(\zeta)$ (i.e., $L(\zeta_{\boldsymbol{\tau}^*})$) is maximal in $A \subseteq A_{\boldsymbol{\tau}}$.

6.2 Alignment

Independently of the existence of a dominating index, alignment is a simple strategy for selecting weights $a_i(\boldsymbol{\tau})$, when the query $F(\boldsymbol{X}) = \sum_{i=1}^{n} f_i(\boldsymbol{X}_i)$ is differentiable.

Fix some clause $\boldsymbol{\tau}$. Assume that each $\zeta_i(\boldsymbol{X}_i; \boldsymbol{\tau})$ is witnessed maximal; for simplicity, assume that it is also eikonal. Now, consider a point $\breve{\boldsymbol{X}} = \breve{\boldsymbol{X}}_1 \oplus \cdots \oplus \breve{\boldsymbol{X}}_n$, such that, for every i, $f_i(\breve{\boldsymbol{X}}_i) = \tau_i$ and $\zeta_i(\breve{\boldsymbol{X}}_i; \tau_i) = 0$. That is, each $\breve{\boldsymbol{X}}_i$ is a maximality witness for $\zeta_i(\boldsymbol{X}_i; \tau_i)$. By smoothness of f_i, its gradient $\boldsymbol{g}_i = \nabla f_i(\breve{\boldsymbol{X}}_i)$ will be parallel to the gradient $\nabla \zeta_i(\breve{\boldsymbol{X}}_i; \tau_i)$ at this witness point. Then, we can choose $a_i(\boldsymbol{\tau})$ so as to *align* (make parallel) the gradients of F and $\zeta_{\boldsymbol{\tau}}$ at $\breve{\boldsymbol{X}}$, which is achieved by $a_i(\boldsymbol{\tau}) = \|\boldsymbol{g}_i\|$. Note also that $\sum_{i=1}^{n} \boldsymbol{g}_i^2 = \|\nabla F(\breve{\boldsymbol{X}})\|^2$, thus, a choice of $a_i = \|\boldsymbol{g}_i\|/\|\nabla F(\breve{\boldsymbol{X}})\|$ yields normalized weights.

This justifies the suitability of the choice $a_i = |w_i|/\|\boldsymbol{w}\|$ for every clause $\boldsymbol{\tau}$ in the previous example. Also, with respect to the previous section, it is easy to see that at the minimizer $\boldsymbol{\tau}^*$ of $\Delta(\boldsymbol{\tau})$, since $\breve{\boldsymbol{X}}$ is a nearest neighbor of $\boldsymbol{E}$ on the boundary of A, by smoothness of F, $\|\boldsymbol{g}_i\|$ will be proportional to $\zeta_i(\boldsymbol{E}_i; \tau_i^*)$.

Like dominating index, this method is limited to query functions that exhibit symmetry; if there are several witness points $\breve{\boldsymbol{X}}$, each giving a different value for a_i, then it is not clear what is the best choice.

6.3 Safe Zones For Inner Product

We turn to the non-trivial problem of designing a good safe zone for the inner product of two vectors. Previous solutions [14, 23, 22] strove for the same qualities, but were suboptimal in terms of maximum distance [23] or maximality [14, 22].

The problem is defined by condition $\boldsymbol{XY} \geq t$. Instead of decomposing this problem on a per-dimension basis, we apply the polarization identity, by the change of variables $\boldsymbol{x} = (\boldsymbol{X}+\boldsymbol{Y})/\sqrt{2}$, and $\boldsymbol{y} = (\boldsymbol{X}-\boldsymbol{Y})/\sqrt{2}$. It is easy to see that $\boldsymbol{x}^2 - \boldsymbol{y}^2 = 2\boldsymbol{XY}$ and that the change of variables is a rotation, that is, it preserves all distances. Therefore, we focus on the (slightly more general) problem of monitoring $\boldsymbol{x}^2 - \boldsymbol{y}^2 \geq T$, with $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^m$. We shall design a safe zone for this problem, for reference point $\boldsymbol{E} = \boldsymbol{\xi} \oplus \boldsymbol{\psi}$, where $\boldsymbol{\xi}^2 - \boldsymbol{\psi}^2 > T$.

First note that, when $\boldsymbol{\xi}$ is 0 (which implies that $T < 0$), the constraint $\|\boldsymbol{y}\| \leq -T$ defines a good (maximum-distance and maximal) safe zone for the original problem. Therefore, we assume $\|\boldsymbol{\xi}\| \neq 0$ and we denote $\hat{\boldsymbol{\xi}} = \boldsymbol{\xi}/\|\boldsymbol{\xi}\|$.

We treat this problem as a separable sum of two functions, $f_1(\boldsymbol{x}) = \boldsymbol{x}^2$ and $f_2(\boldsymbol{y}) = -\boldsymbol{y}^2$, of which f_1 is convex and f_2 is concave (thus, $-f2$ is convex). The condition $f_1 + f_2 \geq T$ can be booleanized as

$$\forall u, v \geq 0 : u^2 - v^2 = T : \quad \boldsymbol{x}^2 \geq u^2 \vee \boldsymbol{y}^2 \leq v^2,$$

where we have applied a convenient change of variables to the index tuple $(\tau_1, \tau_2) = (u^2, -v^2)$.

The problem $\boldsymbol{x}^2 \geq u^2$ is solved optimally by the affine eikonal function $\zeta_1(\boldsymbol{x}; u) = \boldsymbol{x}\hat{\boldsymbol{\xi}} - u$. Similarly, the problem $\boldsymbol{y}^2 \leq v^2$ is solved optimally by the eikonal, non-redundant solution $\zeta_2(\boldsymbol{y}; v) = v - \|\boldsymbol{y}\|$. Putting everything together, we obtain formula

$$\zeta(\boldsymbol{x} \oplus \boldsymbol{y}) = \inf_{u^2 - v^2 = T} \zeta_{u,v}(\boldsymbol{x} \oplus \boldsymbol{y}) \qquad \text{where} \quad \zeta_{u,v}(\boldsymbol{x} \oplus \boldsymbol{y}) = \alpha(u,v)(\boldsymbol{x}\hat{\boldsymbol{\xi}} - u) + \beta(u,v)(v - \|\boldsymbol{y}\|).$$

By alignment, we select (unnormalized) $\alpha(u, v) = u$ and $\beta(u, v) = v$, and get

$$\zeta_{u,v}(\boldsymbol{x} \oplus \boldsymbol{y}) = u\boldsymbol{x}\hat{\boldsymbol{\xi}} - v\|\boldsymbol{y}\| - T.$$

which is tangent to the boundary of the admissible region at every point $u\hat{\boldsymbol{\xi}} \oplus v\hat{\boldsymbol{y}}$, where $\|\hat{\boldsymbol{y}}\| = 1$.

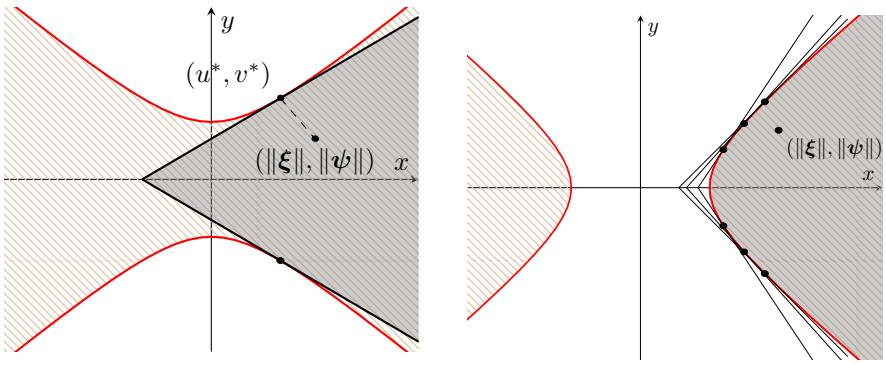

Figure 3 Safe zones (grayed) for $x^2 - y^2 \geq T$, when $T \leq 0$ (left) and $T > 0$ (right). The admissible regions are hatched. On the left, the safe zone is a single cone $u^*x - v^*|y| \geq T$, where (u^*, v^*) is the dominating index. On the right, the safe zone is defined by the intersection of all cones $ux - v|y| \geq T$ (three such cones are shown).

Instead of continuing the treatment of the problem directly, we will examine the case $n = m = 1$; that is, the equivalent planar problem with reference point $\boldsymbol{e} = (\|\boldsymbol{\xi}\|, \|\boldsymbol{\psi}\|)$. This

problem (and its solution) is depicted in Fig. 3. In fact, treating the 2-d problem is equivalent to treating the problem in any dimension, by the change of variables $x = \boldsymbol{x}\hat{\boldsymbol{\xi}}$ and $y = \|\boldsymbol{y}\|$.

Case $T \leq 0$. When $T \leq 0$, the problem admits a dominating index, (u^*, v^*), which is found by minimizing $\Delta(u, v)$ over the index set.

Case $T > 0$. In this case there is no dominating index. The safe zone is the right lobe of the hyperbola, that is $Z = \{x \geq \sqrt{y^2 + T}\}$. This set is smooth everywhere on its boundary, therefore its SDF $\delta_Z(x, y)$ is non-redundant.

Overall, we have the following:

▶ **Theorem 18.** *Given constraint $\boldsymbol{x}^2 - \boldsymbol{y}^2 \geq T$ and reference point $\boldsymbol{\xi} \oplus \boldsymbol{\psi}$ in any dimension, a good safe zone is given by $\zeta(\boldsymbol{x} \oplus \boldsymbol{y}) = \zeta_2(\boldsymbol{x}\hat{\boldsymbol{\xi}}, \|\boldsymbol{y}\|)$, where ζ_2 is the safe zone for the planar constraint $x^2 - y^2 \geq T$ with reference point $(\|\boldsymbol{\xi}\|, \|\boldsymbol{\psi}\|)$. More precisely,*
1. *if $T \leq 0$, $\zeta_2(x, y) = (u^*x - v^*|y| - T)/\sqrt{u^{*2} + v^{*2}}$, and*
2. *if $T > 0$, $\zeta_2(x, y)$ is the SDF of convex set $\{x \geq \sqrt{y^2 + T}\}$.*
Both ζ and ζ_2 are eikonal and non-redundant, and define maximum-distance and witnessed maximal safe zones.

Computation cost for monitoring the inner product. Computing the safe zone function for l-dimensional vectors, takes time $O(l)$, in order to compute $\boldsymbol{x}\hat{\boldsymbol{\xi}}$ and $\|y\|$. Then, computing the actual value can be done in time $O(1)$. When the vectors between successive computations of the safe zone function change in only $O(1)$ coordinates (which is quite standard in stream monitoring, when a stream update changes $O(1)$ locations of the state vector), it is possible simply update cached previous values of $\boldsymbol{x}\hat{\boldsymbol{\xi}}$ and $\|\boldsymbol{y}\|$ and reduce the cost to $O(1)$.

Safe zones for AGMS sketches. As discussed, monitoring the inner product of two streams by AGMS sketches, involves monitoring the median of d inner products, of dimension l each. The result's accuracy is within $O(1/\sqrt{l})$, with probability at least $1 - O(1/2^d)$. For practical purposes, d will be of the order of 10, but l of the order of 1000.

Combining our safe zones for inner product with the safe zone for the median, we obtain the first provably good safe zone for AGMS sketches. The computational cost of testing membership in the safe zone is important, as it is likely to be performed for every stream update. Using the FastAGMS sketch of [10], each stream update changes only 1 counter in each of the d vectors of a sketch. Therefore, the change to each monitored inner product can be computed in $O(1)$ time. The median's safe zone requires $O(d)$ time for testing membership. In conclusion, our safe zones can be used for membership testing with only $O(d)$ cost per update.

7 Related Work

The problem of tracking distributed streams through *in situ* constraints has attracted significant attention in recent years. Still, most existing work has focused on purpose-built solutions for specific query classes; these include, for instance, the simpler cases of thresholding *linear* functions [19, 18, 24]. top-k monitoring [4, 25], ratio threshold queries [16], and tracking polynomials of simple scalar variables [30]. All these techniques typically rely on some form of locally-installed "adaptive filters" – that is, bounds around the value of distributed variables that can grow or shrink over time (e.g., based on variability), while guaranteeing a global bound on the overall uncertainty. Similar local filtering ideas are also employed by Huang et al. [17] to monitor the eigenvalues of a network traffic matrix through

perturbation analysis, and by Wolf et al. [33] to threshold the norm of the average vector in a distributed system. Cormode and Garofalakis [7, 8] introduce the use of *sketch synopses* [9] for effectively summarizing local data streams, and propose sketch-based schemes for the communication-efficient, approximate monitoring of join aggregates over distributed streams. Finally, Cormode et al. [11] give a theoretical study of the distributed function monitoring problem focusing, in particular, on providing *communication lower bounds* for the case of various L_p-norm functions, assuming a "cash-register" (i.e., insert only) streaming model. Lower bounds for distributed norm monitoring are also given by Arackaparambil et al. [3], who demonstrate that, for general "turnstile" streams (i.e., allowing both inserts and deletes), the worst-case communication lower bounds are *linear* in the size of the stream. [1]

The Geometric Method of Sharfman et al. [31, 32] introduced the first generic approach for efficiently thresholding a *general function/query* over distributed data. Their solution relies on a function-agnostic, geometric "covering spheres" technique for breaking the global condition into safe local constraints. Extensions of the basic method as well as the more general notion of convex Safe Zones (SZs) are discussed in a later paper [21], and a broad range of applications have been explored, including distributed outlier detection [6], prediction-based distributed stream monitoring [15], sketch-based monitoring of norms and range aggregates [13], and distributed skyline tracking [27].

As demonstrated in our recent work [23], the safe zones (implicitly) defined by the "covering spheres" method are often far from optimal, and geometric convexity arguments (based on decomposing the problem into convex pieces) can give provably better results in certain important cases. Still, the methodology and results in [23] are heuristic, refer to specific classes of monitoring functions, and do not offer any hard quality guarantees for the resulting safe zones; furthermore, they do not consider the important problem of safe zone composition. Instead, our work is based on a novel functional representation of safe zones which allows us to effectively deal with general Boolean safe zone composition with provable quality guarantees. Our Boolean formalism is, in fact, much more powerful, and can easily express the methodology of [23] as a special case. The very recent work of Lazerson et al. [22] proposes another broad method based on defining "convex bounds" for the monitored function $F()$ using functional approximation techniques (assuming $F()$ is differentiable). However, no optimality properties are formally shown for the resulting SZs, and the problem of effective SZ composition is not addressed. The worst-case communication complexity of geometric techniques for distributed monitoring has not been studied, but empirical studies demonstrate significant communication gains in real problems and query workloads.

8 Conclusions

In this paper, we have presented the first formal framework for the compositional design of convex Safe Zones (SZs), for problems with separable constraints. To this end, we have introduced a functional safe zone representation that conserves, under Boolean composition, the quality guarantees of their component constraints. We have also applied our new framework to general function monitoring scenarios of practical interest.

Important problems remain open for future research, mainly relating the quality of safe zones to actual guarantees on the communication cost of monitoring, and extending our compositional approach beyond boolean, to other types of composite queries.

[1] The problem of communication lower bounds for general functions under the cash-register model remains open.

References

1 Noga Alon, Phillip B. Gibbons, Yossi Matias, and Mario Szegedy. "Tracking Join and Self-Join Sizes in Limited Storage". In *Proc. of the 18th ACM Symposium on Principles of Database Systems*, Philadeplphia, Pennsylvania, May 1999.

2 Noga Alon, Yossi Matias, and Mario Szegedy. "The Space Complexity of Approximating the Frequency Moments". In *Proc. of the 28th Annual ACM Symposium on the Theory of Computing*, pages 20–29, Philadelphia, Pennsylvania, May 1996.

3 Chrisil Arackaparambil, Joshua Brody, and Amit Chakrabarti. Functional monitoring without monotonicity. In *ICALP (1)*, 2009.

4 B. Babcock and C. Olston. Distributed top-k monitoring. In *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, New York, NY, USA, 2003. ACM.

5 Sabbas Burdakis and Antonios Deligiannakis. "Detecting Outliers in Sensor Networks Using the Geometric Approach". In *Proc. of the 28th Intl. Conference on Data Engineering*, April 2012.

6 Sabbas Burdakis and Antonios Deligiannakis. Detecting outliers in sensor networks using the geometric approach. In *ICDE*, 2012.

7 Graham Cormode and Minos Garofalakis. "Sketching Streams Through the Net: Distributed Approximate Query Tracking". In *Proc. of the 31st Intl. Conference on Very Large Data Bases*, Trondheim, Norway, September 2005.

8 Graham Cormode and Minos Garofalakis. "Approximate Continuous Querying over Distributed Streams". *ACM Transactions on Database Systems*, 33(2), June 2008.

9 Graham Cormode, Minos Garofalakis, Peter J. Haas, and Chris Jermaine. "Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches". *Foundations and Trends in Databases*, 4(1-3), 2012.

10 Graham Cormode and Minos N. Garofalakis. Sketching streams through the net: Distributed approximate query tracking. In *VLDB*, 2005.

11 Graham Cormode, S. Muthukrishnan, and Ke Yi. Algorithms for distributed functional monitoring. In *SODA*, 2008.

12 Minos Garofalakis, Johannes Gehrke, and Rajeev Rastogi. *"Data-Stream Management – Processing High-Speed Data Streams"*. Springer-Verlag New York (Data-Centric Systems and Applications Series), 2016.

13 Minos Garofalakis, Daniel Keren, and Vasilis Samoladas. "Sketch-based Geometric Monitoring of Distributed Stream Queries". In *Proc. of the 39th Intl. Conference on Very Large Data Bases*, Trento, Italy, August 2013.

14 Minos N. Garofalakis, Daniel Keren, and Vasilis Samoladas. Sketch-based geometric monitoring of distributed stream queries. *PVLDB*, 2013.

15 Nikos Giatrakos, Antonios Deligiannakis, Minos Garofalakis, Izchak Sharfman, and Assaf Schuster. "Prediction-based Geometric Monitoring of Distributed Data Streams". In *Proc. of the 2012 ACM SIGMOD Intl. Conference on Management of Data*, Scottsdale, Arizona, May 2012.

16 Rajeev Gupta, Krithi Ramamritham, and Mukesh K. Mohania. "Ratio threshold queries over distributed data sources". In *Proc. of the 39th Intl. Conference on Very Large Data Bases*, Trento, Italy, August 2013.

17 Ling Huang, XuanLong Nguyen, Minos N. Garofalakis, Joseph M. Hellerstein, Michael I. Jordan, Anthony D. Joseph, and Nina Taft. Communication-efficient online detection of network-wide anomalies. In *INFOCOM*, 2007.

18 Srinivas R. Kashyap, Jeyashankher Ramamirtham, Rajeev Rastogi, and Pushpraj Shukla. Efficient constraint monitoring using adaptive thresholds. In *ICDE*, pages 526–535, 2008.

19 Ram Keralapura, Graham Cormode, and Jeyashankher Ramamirtham. Communication-efficient distributed monitoring of thresholded counts. In *SIGMOD*, 2006.

20 Daniel Keren, Guy Sagy, Amir Abboud, David Ben-David, Assaf Schuster, Izchak Sharfman, and Antonios Deligiannakis. "Geometric Monitoring of Heterogeneous Streams". *IEEE Transactions on Knowledge and Data Engineering*, 26(8), August 2014.

21 Daniel Keren, Izchak Sharfman, Assaf Schuster, and Avishay Livne. Shape sensitive geometric monitoring. *IEEE Trans. Knowl. Data Eng.*, 24(8), 2012.

22 Arnon Lazerson, Daniel Keren, and Assaf Schuster. Lightweight monitoring of distributed streams. In *Proc. of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1685–1694, New York, NY, USA, 2016. ACM.

23 Arnon Lazerson, Izchak Sharfman, Daniel Keren, Assaf Schuster, Minos Garofalakis, and Vasilis Samoladas. "Monitoring Distributed Streams using Convex Decompositions". In *Proc. of the 41st Intl. Conference on Very Large Data Bases*, August 2015.

24 Shicong Meng, Ting Wang, and Ling Liu. Monitoring continuous state violation in datacenters: Exploring the time dimension. In *ICDE*, pages 968–979, 2010.

25 Sebastian Michel, Peter Triantafillou, and Gerhard Weikum. Klee: a framework for distributed top-k query algorithms. In *VLDB '05*. VLDB Endowment, 2005.

26 S. Muthukrishnan. "Data Streams: Algorithms and Applications". *Foundations and Trends in Theoretical Computer Science*, 1(2), 2005.

27 Odysseas Papapetrou and Minos Garofalakis. "Continuous Fragmented Skylines over Distributed Streams". In *Proc. of the 30th Intl. Conference on Data Engineering*, Chicago, Illinois, April 2014.

28 R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

29 G. Sagy, D. Keren, I. Sharfman, and A. Schuster. "Distributed Threshold Querying of General Functions by a Difference of Monotonic Representation". In *Proc. of the 36th Intl. Conference on Very Large Data Bases*, August 2010.

30 Shetal Shah and Krithi Ramamritham. Handling non-linear polynomial queries over dynamic data. In *ICDE*, 2008.

31 Izchak Sharfman, Assaf Schuster, and Daniel Keren. "A geometric approach to monitoring threshold functions over distributed data streams". In *SIGMOD*, 2006.

32 Izchak Sharfman, Assaf Schuster, and Daniel Keren. "A geometric approach to monitoring threshold functions over distributed data streams". *ACM Trans. Database Syst.*, 32(4), 2007.

33 Ran Wolff, Kanishka Bhaduri, and Hillol Kargupta. A generic local algorithm for mining data streams in large distributed systems. *IEEE Trans. on Knowl. and Data Eng.*, 21(4), 2009.