# Balanced Aspect Ratio Trees Revisited

Amitabh Chaudhary and Michael T. Goodrich

Department of Information & Computer Science,
University of California, Irvine CA 92697, USA,
{amic,goodrich}@ics.uci.edu

**Abstract.** Spatial databases support a variety of geometric queries on point data such as range searches, nearest neighbor searches, etc. Balanced Aspect Ratio (BAR) trees are hierarchical space decomposition structures that are general-purpose and space-efficient, and, in addition, enjoy a worst case performance poly-logarithmic in the number of points for approximate queries. They maintain limits on their depth, as well as on the aspect ratio (intuitively, how skinny the regions can be). BAR trees were initially developed for 2 dimensional spaces and a fixed set of partitioning planes, and then extended to $d$ dimensional spaces and more general partitioning planes. Here we revist 2 dimensional spaces and show that, for any given set of 3 partitioning planes, it is not only possible to construct such trees, it is also possible to derive a simple closed-form upper bound on the aspect ratio. This bound, and the resulting algorithm, are much simpler than what is known for general BAR trees. We call the resulting BAR trees Parameterized BAR trees and empirically evaluate them for different paritioning planes. Our experiments show that our theoretical bound converges to the empirically obtained values in the lower ranges, and also make a case for using evenly oriented partitioning planes.

## 1 Introduction

Spatial databases for scientific applications need efficient data structures to solve a variety of geometric queries. For example, a scientific application with which we have direct experience is the Sloan Digital Sky Survey (SDSS) [17, 18]. It stores light intensities for over a 100 million celestial objects as points on a two-dimensional sphere, and needs support for geometric queries like the nearest neighbor queries, proximity queries, and general range queries (not just axis orthogonal). Similar needs arise in geographical information systems.

There are many access methods based on the hierarchical space decomposition data structures that are useful in solving geometric queries on point data. Quad trees and $k$-d trees are widely popular examples (e.g., see Samet [20, 21]). In these data structures two properties, *depth* and *aspect ratio*, play a crucial role in determining their efficiency in solving queries. The depth of a tree characterizes the number of nodes that have to be visited to find regions of interest. The aspect ratio of a tree (intuitively, how "skinny" a region can be) characterizes the

number of wasteful nodes that can be in any particular region of interest. Most queries require both of these values to be small for them to be solved efficiently.

Unfortunately, both quad trees and $k$-d trees optimize one of these properties at the expense of the other. Quad trees produce regions with optimal aspect ratios, but they can have terrible depth. $K$-d trees, on the other hand, have optimal (logarithmic) depth, but they often produce lots of long-and-skinny regions, which slow query times.

Balanced Aspect Ratio (BAR) trees are hierarchical space decomposition data structures that have logarithmic depth and bounded aspect ratios. As a result they have a worst case performance poly-logarithmic in the number of points for approximate queries such as the approximate nearest neighbor, approximate farthest neighbor, approximate range query, etc. They were initially developed for 2 dimensional spaces [12] for a particular fixed set $(0, \pi/4, \pi/2)$ of partitioning planes. Then [10], they were extended to $d$ dimensional spaces, and the partitioning planes could be chosen flexibly as long as certain conditions were met. These conditions when applied to small dimensional spaces give bounds that are known to be very loose. For instance, in $d$ dimensional spaces, as long as $d$ of the partitioning planes are axis orthogonal, the aspect ratio (we give a precise definition later) is bounded by $50\sqrt{d}+55$. In 2 dimensional spaces, the aspect ratio is bounded by a number very close to 6. We are interested in developing simpler conditions for choosing partitioning planes flexibly in BAR trees for small dimensions. This will allow us to derive tighter bounds for the aspect ratio, and discover the best set of partitioning planes for a particular application.

### 1.1 Related Prior Work

In this subsection, we briefly review some known general-purpose hierarchical spatial decomposition trees for a set of points, $S$.

**The Binary Space Partitioning (BSP) Trees.** The BSP tree [15, 14] is a recursive subdivision of space into regions by means of half-planes. Each node $u$ in the tree represents a convex region and the points from $S$ lying in it. Initially, the root node $r$ of the tree represents a bounding region of the point set $S$. At each node $u$, an associated line partitions the region of $u$, $R_u$, into two disjoint regions $R_l$ and $R_r$. The node $u$ then has two child nodes $l$ and $r$ representing $R_l$ and $R_r$ respectively. If the number of points from $S$ in $R_u$ is less than some constant, $u$ is not partitioned and becomes a leaf. These structures satisfy our condition of being general-purpose, but without further restricting how the cutting lines are chosen, these structures are inefficient. Thus, much work has been done on methods for specializing BSP trees to be more efficient, which we review next.

**The $k$-d Tree.** This structure was introduced by Bentley [3–5] and has been extensively studied. It is a special class of the BSP tree: the partitioning line is orthogonal to one of the axes and such that it divides the set of points at the node in half by cardinality. This guarantees that the depth of the tree is $O(\log n)$. So point location queries, which take time proportional to the depth,

can be answered efficiently. But, since there are no guarantees on the aspect ratio of the regions produced, with some exceptions [9], the running times of queries can nevertheless be poor.

**The Quadtree.** The quadtree (e.g., see [20, 21]) is another special class of the BSP tree. The point set $S$ is initially bounded by a square, and the partitions are such that a square region is divided into four smaller squares of equal area. (This notion can be extended to $d$ dimensional space, giving rise to a structure called the *octree*.) The aspect ratios of regions in quadtrees is bounded by a constant, but these trees can have unbounded depth. So even basic point location queries can take unbounded time. If the point set is uniformly distributed, however, then the depth is bounded, and in those situations quadtrees perform well for some geometric queries.

**Balanced Box Decomposition Trees** In [1, 2], Arya *et al.* describe a relative of a binary-space partitioning tree called the *Balanced Box Decomposition* (BBD) tree. This structure is based on the fair-split tree of Callahan and Kosaraju [6, 7] and is defined such that its depth is $O(\log n)$ and all of the associated regions have low combinatorial complexity and bounded aspect ratio. Arya *et al.* show how BBD trees guarantee excellent performance in approximating general range queries and nearest-neighbor queries. (Approximate queries are like the regular exact versions, except they allow an error. See [10] for formal definitions.) The aspect ratio bound on each region allows them to bound the number of nodes visited during various approximate query searches by limiting the number of nodes that can be packed inside a query region. However, since these trees rely on using hole cuts during construction, they produce non-convex regions and thus are not true BSP trees. This is also a drawback with respect to several applications in computer graphics and graph drawing, where convexity of the partitioned regions is desirable (e.g., see [12, 16]).

**Balanced Aspect Ratio Trees** Duncan *et al.* [12, 11] introduced the *Balanced Aspect Ratio* (BAR) trees. These are similar to $k$-d trees in 2-dimensional space, except that instead of allowing only axis-orthogonal partitions, they also allow a third partition orthogonal to a vector at a $\pi/4$ angle to the axes. This extra cut allows them to find partitions that not only divide the point set in a region evenly, but also ensure that the child regions have good (bounded) aspect ratio. In [10], Duncan extended BAR trees to $d$ dimensions. He showed that if a certain set of conditions is satisfied, BAR trees with bounded aspect ratio can be constructed. He also proved bounds on the running time of approximate queries. The $(1+\epsilon)$-nearest neighbor query and the $(1-\epsilon)$-farthest neighbor query can be answered in $O(\log n + (1/\epsilon)\log(1/\epsilon))$ time, where $n$ is the number of points. The $\epsilon$-range query and the $\epsilon$-proximity query can be answered in $O(\log n + (1/\epsilon) + k)$ time, where $k$ is the size of the output.

### 1.2 Our Contributions

In this paper, we introduce the Parameterized Balanced Aspect Ratio (PBAR) trees in 2 dimensions, which take any three vectors as the partitioning planes. They enjoy all the advantages of BAR trees: general pupose, space efficient,

logarithmic depth, bounded aspect ratio, poly-logarithmic worst case bounds for approximate versions of spatial queries. In addition, they have a bound on the aspect ratio which is a simple closed-form function of the given partitioning planes. The proofs used are significantly different from those for earlier BAR trees: they use the advantages of 2 dimensional spaces and yet work of any given set of partitioning vectors.

Our motivation for introducing PBAR trees comes from our experience with the Sloan Digital Sky Survey (SDSS) [17, 18]. Because objects in the SDSS are indexed by their positions on the night sky, the data can be viewed at a first level of indexing as two-dimensional points on a sphere. To allow for efficient access, astronomers overlay a quasi-uniform triangular "grid" on this sphere, to reduce the curvature of each "leaf" triangle to be "almost" planar and to reduce the number of points in each such triangle to a few hundred thousand. The difficulty is that when these leaf triangles are mapped to a projection plane to allow for fast queries via a secondary data structure, there are many different side angles that must be dealt with. A data structure like PBAR trees can conveniently use the given angles of a bounding triangle as its possible partitioning directions. Without this convenience, we would get poorly-shaped regions near the boundaries of these triangles.

The bounds on the running times for approximate queries, in [10], depend only on the fact that both BBD and BAR trees have $O(\log n)$ depth and ensure a constant bound on the aspect ratio of all their regions. These two conditions are satisfied by PBAR trees as well. So the same bounds hold for PBAR trees as well. We present empirical results for the $(1+\epsilon)$-nearest neighbor query using PBAR trees with various different partitioning planes using artificial data as well as real datasets from the SDSS. Our experiments also indicate that our bound is tight in some respects.

In the next section we give the foundations and definition for PBAR trees. In Section 3 we describe the algorithm for constructing PBAR trees and prove its correctness. In the last section we present our empirical results. We include details for the pseudo-code and proofs of correctness in an optional appendix in this extended abstract.

## 2 Parameterizing BAR Trees

PBAR trees are for point data in 2-dimensional space. The *distance* $\delta(p, q)$ between two points $p = (p_1, p_2)$ and $q = (q_1, q_2)$ is $\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$. Extending this notion, the distance between two sets of points $P$ and $Q$ is

$$\delta(P, Q) = \min_{p \in P, q \in Q} \delta(p, q).$$

$S$ is the set of $n$ points given as input. The size $|R|$ of region $R$ is the number of points from $S$ in $R$.

*Partitioning Vectors.* We use vectors from $\mathbb{R}^2$ to specify partitioning directions. Note that partitioning vectors $l$ and $-l$ are equivalent. The *angle* $\theta_{lm}$ between

two partitioning vectors $l$ and $m$ is the angle from $l$ to $m$ in the counterclockwise direction, except that we take into account that $m$ and $-m$ are equivalent. In the context of trigonometric functions, we shall prefer to use, for example, the short $\sin(lm)$ instead of $\sin(\theta_{lm})$.

To construct a PBAR tree, we use the 3 partitioning vectors from the given set $V = \{\lambda, \mu, \nu\}$. We make all partitions by taking a region $R$ and dividing it into two subregions, $R_1$ and $R_2$, with a line $c'_l$, called a *cut*, orthogonal to some $l \in V$. A cut orthogonal to $l$ is also called an *$l$-cut*. Note that if $R$ is convex, both $R_1$ and $R_2$ are convex too. We divide the set of points in $R$, call it $S$, between $R_1$ and $R_2$ in the natural fashion. For points in $S$ that are on $c_l$, we assign each of them to either $R_1$ or $R_2$ as per convenience.

Let the sequence $(\lambda, \mu, \nu)$ be in the counterclockwise order. All the 3 sequences in the set $\mathcal{P}(V) = \{(\lambda, \mu, \nu), (\mu, \nu, \lambda), (\nu, \lambda, \mu)\}$ are equivalent for our purpose. So, often, we shall speak in terms of the general $(l, m, n)$, where $(l, m, n) \in \mathcal{P}(V)$.

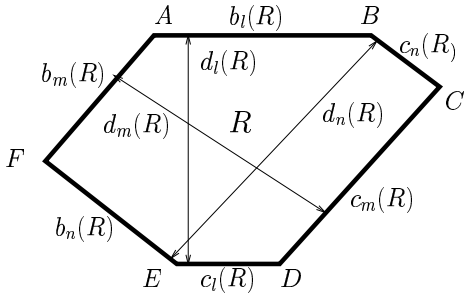


**Fig. 1.** The names used for the sides and the diameters.

*Canonical Regions and Canonical Aspect Ratios.* We assume that the given set of points $S$ has an initial convex bounding region with sides that are orthogonal to $\lambda, \mu$, or $\nu$. Since in constructing PBAR trees we make all partitions with lines orthogonal to these 3 partitioning vectors, the regions we construct are always hexagons with sides orthogonal to $\lambda, \mu$, or $\nu$. Some sides may be degenerate, that is, of length 0. We call these hexagonal regions *canonical regions*. See Figure 1. In a canonical region $R$, $b_l(R)$ and $c_l(R)$ are the two unique opposing sides orthogonal to the partitioning vector $l$. The *diameter* $d_l(R)$ of $R$ with respect to the partitioning vector $l$ is the distance $\delta(b_l(R), c_l(R))$. The *maximum diameter* of $R$ is $d_{\max}(R) = \max_{l \in V} d_l(R)$, and the *minimum diameter* of $R$ is $d_{\min}(R) = \min_{l \in V} d_l(R)$. When the region is understood from the context we drop the argument in the above notations and use, for example, $b_l$ instead of $b_l(R)$. A *canonical trapezoidal region* is of special interest, and is a canonical region that is quadilateral and has exactly one pair of parallel sides.

The *canonical aspect ratio* $\mathsf{casp}(R)$ of canonical region $R$ is the ratio of $d_{\max}(R)$ to $d_{\min}(R)$.

*PBAR Trees, One-Cuts, and Two-Cuts.* Given a *balancing factor* $\alpha$, $R$ is $\alpha$-*balanced* or has a *balanced aspect ratio* if $\mathsf{casp}(R) \leq \alpha$. $R$ is *critically balanced* if $\mathsf{casp}(R) = \alpha$. A cut $c_l$ orthogonal to $l \in V$ that divides an $\alpha$-balanced region $R$ into $R_1$ and $R_2$ is *feasible* if both $R_1$ and $R_2$ are $\alpha$-balanced.

Given a set $S$ of $n$ points in 2-dimensional space, a set of 3 partitioning vectors $V = \{\lambda, \mu, \nu\}$, a balancing factor $\alpha$, $\alpha \geq 1$, and a reduction factor $\beta$, $0.5 \leq \beta < 1$, a *Parameterized Balanced Aspect Ratio* tree $T$ is a BSP tree on $S$ such that

1. All partitions are made with cuts orthogonal to the vectors in $V$;
2. The canonical aspect ratio of each region is at most $\alpha$;
3. The number of points in each leaf cell of $T$ is a constant with respect to $n$.
4. The depth of $T$ is $O(\log_{1/\beta} n)$.

Given a balancing factor $\alpha$ and reduction factor $\beta$, an $\alpha$-balanced region $R$ is *one-cuttable* if there is a cut $c$, called a *one-cut*, orthogonal to a vector in $V$ that divides $R$ into two canonical subregions $R_1$ and $R_2$ such that

1. $c$ is feasible;
2. $|R_1| \leq \beta|R|$ and $|R_2| \leq \beta|R|$.

(Note that if there is a continuum of feasible cuts that cover the entire region $R$, then $R$ is one-cuttable, as at least one of these cuts will satisfy 2 above.) A region $R$ is *two-cuttable* if there is a cut $c$, called a *two-cut*, orthogonal to a vector in $V$ that divides $R$ into two canonical subregions $R_1$ and $R_2$ such that

1. $c$ is feasible;
2. $|R_1| \leq \beta|R|$;
3. $|R_2| \leq \beta|R|$ or $R_2$ is one-cuttable.

*Shield Regions.* Let $R$ be an $\alpha$-balanced canonical region and let $x_l$ be a side of $R$, $x \in \{b, c\}$, $l \in V$. Now sweep a cut $x'_l$ starting from the side opposite to $x_l$ toward $x_l$. Let $P$ be the subregion formed between $x_l$ and $x'_l$. In the beginning, $\mathsf{casp}(P) \leq \alpha$. Sweep $x'_l$ toward $x_l$ and stop when $P$ is critically balanced. $P$ is called the *shield region* $\mathsf{shield}_{x_l}(R)$ of $R$ with respect to $x_l$. $x'_l$ is the *cut for* $\mathsf{shield}_{x_l}(R)$. $R$ has two shield regions for each $l \in V$, $\mathsf{shield}_{b_l}(R)$ and $\mathsf{shield}_{c_l}(R)$. Note that $R$ has a feasible $l$-cut if and only if $\mathsf{shield}_{b_l}(R) \cap \mathsf{shield}_{c_l}(R) = \emptyset$.

For a given $l \in V$, the *maximal shield region* $\mathsf{maxshield}_l(R)$ of $R$ with respect to $l$ is one among $\mathsf{shield}_{b_l}(R)$ and $\mathsf{shield}_{c_l}(R)$ that has the maximum size. Note that $R$ has a one-cut orthogonal to $l$ only when $|\mathsf{maxshield}_l(R)| \leq \beta|R|$.

## 3 The PBAR Tree Algorithm

In this section we present the PBAR tree algorithm that, given a set of partitioning vectors $V = \{\lambda, \mu, \nu\}$, a reduction factor $\beta$, and a balancing factor $\alpha$,

constructs a PBAR tree on any set $S$ of $n$ points in 2-dimensional space; as long as $0.5 \leq \beta < 1$, and $\alpha$ is at least

$$f(V) = \frac{4.38}{\sin(\theta_{\min}) \sin(\lambda\mu) \sin(\mu\nu) \sin(\nu\lambda)},$$

where $\theta_{\min}$ is the minimum among the angles $(\lambda\mu), (\mu\nu), (\nu\lambda)$.

The PBAR tree algorithm takes an initial $\alpha$-balanced canonical region $R$ that bounds $S$ and recursively subdivides it by first searching for a one-cut, and if no such cut exists, by searching for a two-cut. For details see Figure C-1.

The algorithm for searching for a one-cut, OneCut, considers each partitioning vector in turn. For a partitioning vector $l$, a one-cut orthogonal to it exists if and only if the shield regions with respect to $l$ don't overlap and the maximal shield region contains at most $\beta|R|$ points. Details are in Figure C-2.

The algorithm for searching for a two-cut, TwoCut, considers very few cuts as potential two-cuts. Only cuts for the maximal shield regions for the 3 partitioning vectors are considered as potential two-cuts. This is sufficient as long as $\alpha \geq f(V)$ — that this is true is our main result and we prove it in the rest of the section. Details for TwoCut are in Figure C-3.

**Theorem 1 (Main Result).** *Given a set $S$ of $n$ points in 2-dimensional space, a set of 3 partitioning vectors $V = \{\lambda, \mu, \nu\}$, a balancing factor $\alpha$, $\alpha \geq f(V)$, and a reduction factor $\beta$, $0.5 \leq \beta < 1$, the PBAR tree algorithm constructs a PBAR tree on $S$ in $O(n \log n)$ time.*

We first prove some preliminary lemmas. For all of these we shall assume that $0.5 \leq \beta < 1$ and $\alpha \geq f(V)$.

**Lemma 1.** *Given a set of partitioning vectors $V$, a balancing factor $\alpha$, and a reduction factor $\beta$, if every $\alpha$-balanced region $R$ is two-cuttable, then a PBAR tree can be constructed for every set $S$ of $n$ points.*

*Proof.* Start with any initial $\alpha$-balanced canonical region that bounds $S$. Since this region is two-cuttable, divide it into a maximum of 3 $\alpha$-balanced subregions such that each contains less than $\beta n$ points. Repeat this process for each of the resulting subregions until each of the final leaf regions has at most a constant number of points. The process, along any path of subregions, cannot be repeated more than $O(\log_{1/\beta} n)$ times.

**Lemma 2.** *A canonical region $R$ that is a triangle is always $\alpha$-balanced.*

Proofs for the above and all the following lemmas are in Appendix A.

**Lemma 3.** *Let $(l, m, n) \in \mathcal{P}(V)$. Let $R$ be an $\alpha$-balanced canonical region that is not critically balanced. If $P$ is a critically balanced subregion created by partitioning $R$ with an $l$-cut, then the minimum diameter of $P$ is $d_l(P)$.*

**Corollary 1.** *For a critically balanced canonical region $R$ that is a trapezoid, if $b_l$ and $c_l$ are the two parallel sides, then the minimum diameter of $R$ is $d_l$.*

**Lemma 4.** *Let $(l, m, n) \in \mathcal{P}(V)$, and let $R$ be an $\alpha$-balanced region that has no feasible $l$-cut. Let $S$ be the region formed by extending $R$ such that $b_n(S)$ is of length 0. If $P$ is $\mathsf{shield}_{c_m}(S)$, then $c_l(P) \leq c_l(R)$.*

**Lemma 5.** *Let $(l, m, n) \in \mathcal{P}(V)$. If an $\alpha$-balanced region $R$ has no feasible $l$-cut, then it has a feasible $m$-cut and a feasible $n$-cut.*

**Corollary 2.** *Let $(l, m, n) \in \mathcal{P}(V)$. If an $\alpha$-balanced region $R$ has no one-cut and no feasible $l$-cut, then the maximal shield regions with respect to $m$ and $n$ intersect.*

**Lemma 6.** *A critically balanced region $R$ that is a trapezoidal is one-cuttable.*

**Lemma 7.** *Let $(l, m, n) \in \mathcal{P}(V)$. If an $\alpha$-balanced region $R$ has no one-cut and no feasible $l$-cut, then $R$ has a two-cut.*

**Lemma 8.** *Let $(l, m, n) \in \mathcal{P}(V)$. If an $\alpha$-balanced region $R$ does not have a one-cut and yet there are feasible cuts along all 3 partitioning vectors $l$, $m$, and $n$, then $R$ has a two-cut.*

Proofs in Appendix A.

**Proof of Theorem 1.** The PBAR tree algorithm recursively subdivides the initial bounding region by first searching for a one-cut, and if no such cut exists, by searching for a two-cut. By Lemma 1, if it always succeeds in finding a one-cut or a two-cut, it constructs a PBAR tree. It is easy to see that when the algorithm does not find a one-cut, no such cut exists. In such a situation, the algorithm searches for a two-cut by checking if any of the 3 maximal shield regions are one-cuttable. The proofs for Lemmas 7 and 8 show that at least one of these shield regions is one-cuttable, and so the algorithm always succeeds in finding either a one-cut or a two-cut. For the time analysis see proof in Appendix A. □

## 4 Empirical Tests

In this section we present the preliminary empirical results we have obtained and analyze them. We look at measures like number of nodes created, the depth of the tree, and the number of leaves visited instead of the actual time or space requirements. This is because the time and space measures are dependent on the efficiency of the implementation, the load on the machine during testing, etc. The other measures are not as dependent on the kind of testing carried out. In addition, the number of nodes visited is the dominant term if the data structure is stored in external memory (as is the case in SDSS).

We took 2 data sets and varied the partitioning planes in small increments and for each we found the best aspect ratio that can be obtained. First, we present plots that summarize the results of this experiment. Later, we present detailed results for 6 different data sets in which we compare BAR trees with $(0, \pi/4, \pi/2)$ partitioning angles with 2 instances of PBAR trees.

### 4.1 Varying the partitioning planes

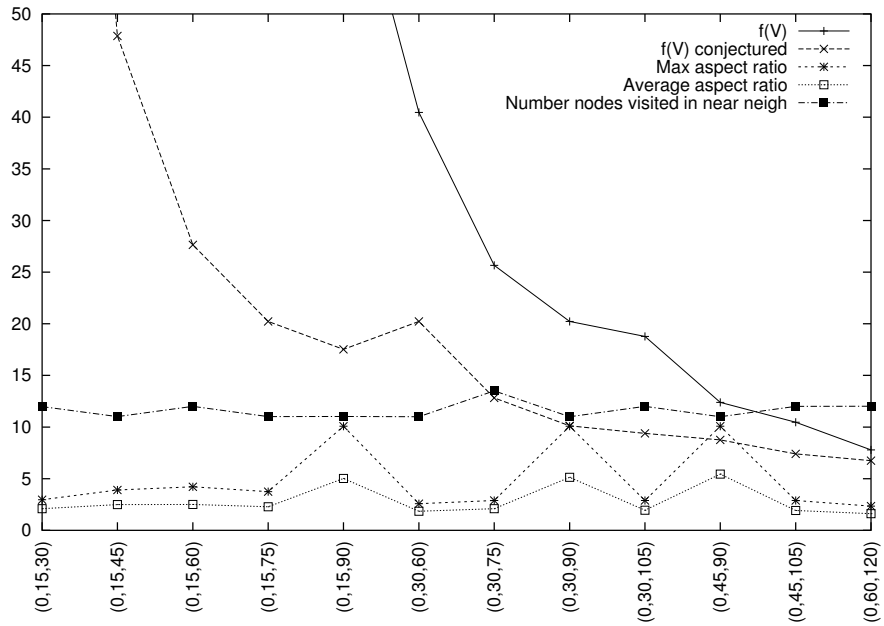


**Fig. 2.** Effects of varying the partitioning planes on SDSS data.

We varied the set of partitioning planes for a given dataset and found the best possible aspect ratio that can be obtained for that dataset. We plot this best empirical aspect ratio in Figures 2 and 3 (it is called the Maximum aspect ratio in the figures). We also plot, alongside, the bound $f(V)$ on the aspect ratio that we have proved. We had conjectured that $f(V)$ can be tightened by removing the $\sin(\theta_{\min})$ term to

$$f(V) = \frac{4.38}{\sin(\lambda\mu)\sin(\mu\nu)\sin(\nu\lambda)}.$$

We plot the conjectured bound as well. We also plot the average aspect ratio of the nodes in tree.

The bound and the conjectured bound both decrease dramatically as the planes become evenly oriented. But the empirically obtained values do not follow their lead, though there is a reasonable amount of variation in the best (maximum) aspect ratio possible. The value of $f(V)$ converges towards the empirical value as it reduces, which indicates that it is possibly tight for the evenly oriented planes. The same is true for the conjectured bound; but, if you look closely at Figure 3 for the orientation (0, 45, 90) the conjectured value is actually lower

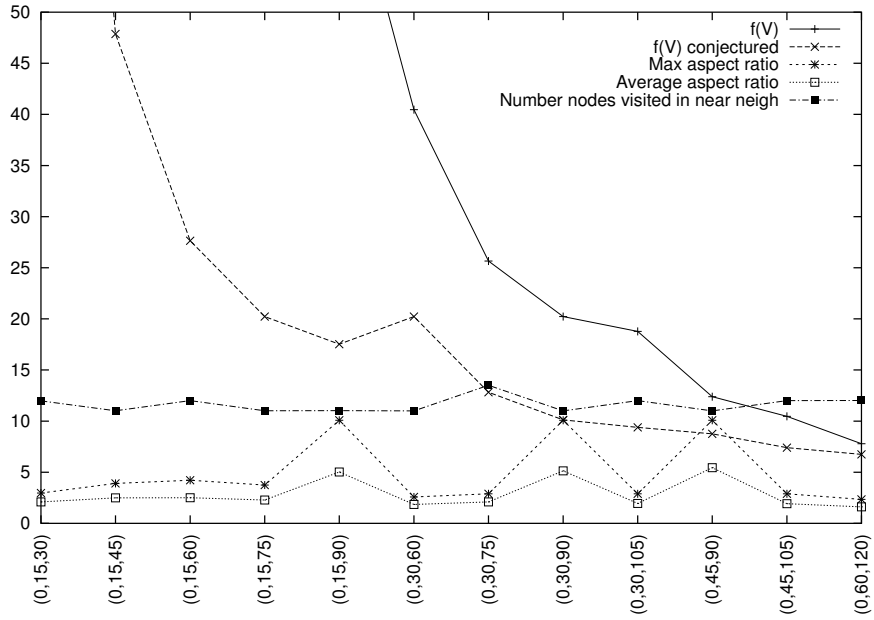


**Fig. 3.** Effects of varying the partitioning planes on data created uniformly at random along a circle.

than the best aspect ratio obtained through experiments. This indicates that the conjecture is wrong. We also plot the number of nodes visited during the nearest neighbor searches (the details for these searches are decribed in the next section) for the various planes. There is very slight variation in this, or in the average aspect ratio with the change in plane orientations.

### 4.2 Comparing BAR trees and PBAR trees

To compare the performance of BAR trees (with mostly axis orthogonal planes: $0, \pi/4, \pi/2$) and PBAR trees we constructed PBAR trees with three different sets of partitioning vectors $V$. In the first set, $V$ is such that $\theta_{\lambda\mu} = \pi/4$ and $\theta_{\lambda\nu} = \pi/2$. In the second set, $V$ is such that $\theta_{\lambda\mu} = \pi/3$ and $\theta_{\lambda\nu} = 2\pi/3$, and in the third set, $V$ is such that $\theta_{\lambda\mu} = \pi/6$ and $\theta_{\lambda\nu} = \pi/2$. Note that the partitioning vectors in the first set are that used by BAR trees. The PBAR trees constructed in this case closely mimic BAR trees and the performance is representative of the performance of BAR trees. In the second case, the partitioning vectors are more evenly oriented, while in the third case they are less evenly oriented, than the first case. We refer to the former case as the BAR tree results and the latter two cases as the PBAR-even tree and PBAR-uneven tree results respectively. In all cases $\alpha$ is 20, $\beta$ is 0.6, and the maximum points in any leaf $k$ is 5. A 100 $(1 + \epsilon)$-nearest neighbor queries were solved using each tree. For each data

set, a point $q$ is first chosen uniformly at random from among the data points. This is the first query point. Then a random increment is chosen and repeatedly added to $q$ to get 99 other query points. $\epsilon$ for the queries is always 0.001. The results for BAR tree are in Figure 4, for PBAR-even tree are in Figure 5, and for PBAR-uneven tree are in Figure 6. Of the 6 data sets data sets, 4 were chosen rather arbitrarily, and the last two are real data from the Sloan Digital Sky Survey(SDSS). The data sets are described in Appendix B.

| Data Set | Nodes in tree | Depth of tree | Avg. $\mathsf{casp}(\cdot)$ of regions | Nodes visited during query | Leaves visited during query |
|---|---|---|---|---|---|
| Set 1 | 4615 | 12 | 9.70 | 57.2 | 11.74 |
| Set 2 | 4719 | 13 | 8.78 | 73.64 | 13.97 |
| Set 3 | 4709 | 12 | 9.08 | 148.51 | 30.97 |
| Set 4 | 4631 | 12 | 7.84 | 16.47 | 2.49 |
| Set 5 | 4749 | 12 | 8.98 | 13.49 | 1.60 |
| Set 6 | 2079 | 11 | 9.13 | 21.84 | 3.68 |

**Fig. 4.** Results for $(1+\epsilon)$-nearest neighbor queries on BAR trees. (Number of nodes and leaves visited during query are averaged over a 100 queries.)

| Data Set | Nodes in tree | Depth of tree | Avg. $\mathsf{casp}(\cdot)$ of regions | Nodes visited during query | Leaves visited during query |
|---|---|---|---|---|---|
| Set 1 | 4647 | 12 | 9.93 | 21.5 | 3.66 |
| Set 2 | 4791 | 12 | 8.78 | 22.61 | 4.38 |
| Set 3 | 4741 | 12 | 9.36 | 13.7 | 1.48 |
| Set 4 | 4641 | 12 | 8.14 | 15.59 | 2.23 |
| Set 5 | 4673 | 12 | 8.41 | 32.44 | 5.91 |
| Set 6 | 2067 | 11 | 8.77 | 21.19 | 3.49 |

**Fig. 5.** Results for $(1+\epsilon)$-nearest neighbor queries on PBAR-even trees; $V$ such that $\theta_{\lambda\mu} = \pi/3$ and $\theta_{\lambda\nu} = 2\pi/3$. (Number of nodes and leaves visited during query are averaged over a 100 queries.)

| Data Set | Nodes in tree | Depth of tree | Avg. $\mathsf{casp}(\cdot)$ of regions | Nodes visited during query | Leaves visited during query |
|---|---|---|---|---|---|
| Set 1 | 4635 | 12 | 9.64 | 58.04 | 12.03 |
| Set 2 | 4697 | 13 | 8.58 | 76.32 | 14.91 |
| Set 3 | 4655 | 12 | 8.86 | 152.81 | 31.62 |
| Set 4 | 4629 | 12 | 7.44 | 15.96 | 2.37 |
| Set 5 | 4683 | 12 | 8.57 | 13.48 | 1.6 |
| Set 6 | 2077 | 11 | 8.82 | 21.21 | 3.51 |

**Fig. 6.** Results for $(1+\epsilon)$-nearest neighbor queries on PBAR-uneven trees; $V$ such that $\theta_{\lambda\mu} = \pi/6$ and $\theta_{\lambda\nu} = \pi/2$. (Number of nodes and leaves visited during query are averaged over a 100 queries.)

The number of nodes in the tree are about the same, for the first 5 data sets, irrespective of $V$. This is expected as the number of data points and the

maximum size of a leaf are the same in all cases. The depth of the trees are about the same too, irrespective of the data set. This, again, is expected as the $\beta$ values are the same, and we don't except too many regions that require two-cuts. Set 6 has far fewer points and so has much fewer nodes.

Surprisingly, the average values of the canonical aspect ratio are not very different in the three trees for the different data sets. Neither is one of the trees always better than the other. Such is not the case for number of nodes and number of leaves visited during query processing. PBAR-even trees almost always visit fewer nodes and fewer leaves and in one particular case the difference with both BAR and PBAR-uneven trees is a factor of 10. That PBAR-even trees perform better at approximate nearest neighbor searches may be expected given the theoretical results, it is surprising that this should be the case when the canonical aspect ratio values are about the same. For Set 5, which is real data set from the SDSS, however, PBAR-even trees are not the best. For this set, both BAR trees and PBAR-uneven trees perform better, with PBAR-uneven slightly ahead of BAR. For the other real data, Set 6, again PBAR-even is the best.

In conclusion, our experiments show that the flexibility of PBAR trees can help in increasing the efficiency of approximate nearest neighbor searches.

## 5 Conclusion and Future Work

In this paper we revisited BAR trees in 2 dimensional spaces and developed the Parameterized Balanced Aspect Ratio (PBAR) trees. These allow any given set of 3 partitioning planes and yet retain all the advantages of BAR trees — general purpose data stuctures, space efficient, logarithmic depth, bounded aspect ratio, and poly-logarithmic approximate query processing. These are the first known "BAR-type" trees in which the aspect ratio can be bounded by a simple closed-form function (it depends on the orientation of the partitioning planes). We conducted empirical tests that show that in many instances the evenly oriented partitioning planes are better than the mostly axis orthogonal planes that have been mostly studied prior to this. In addition, our experiments indicate that our bound is tight in some respects: it converges to empirical values for evenly oriented planes, and that a natural modification to tighten it (our conjecture that $\sin(\theta_{\min})$ factor can be removed) is wrong.

Having bounds on the aspect ratio can be useful in ways other than solving queries faster. For example, PBAR trees can be used to efficiently compute the density of a region around a given point. This can be useful in detecting density based outliers. We want to explore this and other possible applications of PBAR trees in spatial data mining.

#### Acknowledgment

The authors will like to thank Breno de Medeiros for many helpful suggestions, including helping tighten the bounds in Lemma 7; Christian Duncan for discussions; and Tanu Malik for discussions, proof reading, and suggestions.

## References


1. S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Wu. An optimal algorithm for approximate nearest neighbor searching. In *Proc. 5th ACM-SIAM Sympos. Discrete Algorithms*, pages 573–582, 1994.
2. Sunil Arya and David M. Mount. Approximate range searching. In *Proc. 11th Annu. ACM Sympos. Comput. Geom.*, pages 172–181, 1995.
3. J. L. Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, September 1975.
4. J. L. Bentley. Multidimensional binary search trees in database applications. *IEEE Trans. Softw. Eng.*, SE-5:333–340, 1979.
5. J. L. Bentley. $K$-d trees for semidynamic point sets. In *Proc. 6th Annu. ACM Sympos. Comput. Geom.*, pages 187–197, 1990.
6. P. B. Callahan and S. R. Kosaraju. A decomposition of multidimensional point sets with applications to $k$-nearest-neighbors and $n$-body potential fields. *J. ACM*, 42:67–90, 1995.
7. Paul B. Callahan and S. Rao Kosaraju. Algorithms for dynamic closest-pair and $n$-body potential fields. In *Proc. 6th ACM-SIAM Sympos. Discrete Algorithms*, pages 263–272, 1995.
8. Mark de Berg, Marc van Kreveld, Mark Overmars, and Otfried Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, Berlin, 1997.
9. M. Dickerson, C. A. Duncan, and M. T. Goodrich. K-D trees are better when cut on the longest side. In *Proc. 8th European Symp. on Algorithms*, volume 1879 of *Lecture Notes Comput. Sci.*, pages 179–190. Springer-Verlag, 2000.
10. C. Duncan. *Balanced Aspect Ratio Trees*. PhD thesis, The Johns Hopkins University, Baltimore, Maryland, Sep 1999.
11. C. A. Duncan, M. T. Goodrich, and S. Kobourov. Balanced aspect ratio trees: combining the advantages of k-d trees and octrees. In *Proc. 10th Annu. ACM-SIAM Sympos. Discrete Alg.*, pages 300–309, 1999.
12. C. A. Duncan, M. T. Goodrich, and S. G. Kobourov. Balanced aspect ratio trees and their use for drawing very large graphs. In *Graph Drawing*, Lecture Notes in Computer Science, pages 111–124. Springer-Verlag, 1998.
13. H. Edelsbrunner. *Algorithms in Combinatorial Geometry*, volume 10 of *EATCS Monographs on Theoretical Computer Science*. Springer-Verlag, Heidelberg, West Germany, 1987.
14. H. Fuchs, G. D. Abrams, and E. D. Grant. Near real-time shaded display of rigid objects. *Comput. Graph.*, 17(3):65–72, 1983. Proc. SIGGRAPH '83.
15. H. Fuchs, Z. M. Kedem, and B. Naylor. On visible surface generation by a priori tree structures. *Comput. Graph.*, 14(3):124–133, 1980. Proc. SIGGRAPH '80.
16. David Luebke and Carl Erikson. View-dependent simplification of arbitrary polygonal environments. In Turner Whitted, editor, *SIGGRAPH 97 Conference Proceedings*, Annual Conference Series, pages 199–208. ACM SIGGRAPH, Addison Wesley, August 1997. ISBN 0-89791-896-7.
17. Robert Lupton, F. Miller Maley, and Neal Young. Sloan digital sky survey. http://www.sdss.org/sdss.html.
18. Robert Lupton, F. Miller Maley, and Neal Young. Data collection for the Sloan Digital Sky Survey—A network-flow heuristic. *Journal of Algorithms*, 27(2):339–356, 1998.
19. F. P. Preparata and M. I. Shamos. *Computational Geometry: An Introduction*. Springer-Verlag, New York, NY, 1985.

20. H. Samet. *Spatial Data Structures: Quadtrees, Octrees, and Other Hierarchical Methods*. Addison-Wesley, Reading, MA, 1989.
21. H. Samet. *Applications of Spatial Data Structures: Computer Graphics, Image Processing, and GIS*. Addison-Wesley, Reading, MA, 1990.

# A Proofs

**Observation 2** For a canonical region $R$ (see Figure 1) the following are true:

$$d_l(R) = c_n(R)\sin(ln) + c_m(R)\sin(lm) = b_n(R)\sin(ln) + b_m(R)\sin(lm), \quad \text{(A-1)}$$
$$d_m(R) = c_l(R)\sin(lm) + b_n(R)\sin(mn) = c_n(R)\sin(mn) + b_l(R)\sin(lm), \quad \text{(A-2)}$$
$$d_n(R) = c_m(R)\sin(mn) + c_l(R)\sin(ln) = b_l(R)\sin(ln) + b_m(R)\sin(mn). \quad \text{(A-3)}$$

Note that $b_n(R) \leq d_l(R)/\sin(ln)$. Combining it with Equation A-2 gives

$$d_m(R) \leq c_l(R)\sin(lm) + \frac{d_l(R)}{\sin(ln)}\sin(mn), \quad \text{(A-4)}$$

which can be re-written as

$$c_l(R) \geq \frac{d_m(R)}{\sin(lm)} - \frac{d_l(R)\sin(mn)}{\sin(lm)\sin(ln)}. \quad \text{(A-5)}$$

Similarly, we get the following from Equation A-3 and the inequality $c_m(R) \leq d_l(R)/\sin(lm)$[1]:

$$d_n(R) \leq c_l(R)\sin(ln) + \tfrac{d_l(R)}{\sin(lm)}\sin(mn), \quad \text{(A-6)}$$
$$c_l(R) \geq \tfrac{d_n(R)}{\sin(ln)} - \tfrac{d_l(R)\sin(mn)}{\sin(lm)\sin(ln)}. \quad \text{(A-7)}$$

Combining Inequalities A-5 and A-7 we get

$$c_l(R) \geq \max\{\frac{d_m(R)}{\sin(lm)}, \frac{d_n(R)}{\sin(ln)}\} - \frac{d_l(R)\sin(mn)}{\sin(lm)\sin(ln)},$$

which, when $R$ is critically balanced (or even unbalanced) and $d_{\min}(R)$ is $d_l(R)$, gives

$$c_l(R) \geq d_l(R)\left(\alpha - \frac{\sin(mn)}{\sin(lm)\sin(ln)}\right). \quad \text{(A-8)}$$

This can be re-written as

$$d_l(R) \leq \frac{c_l(R)\sin(lm)\sin(ln)}{\alpha\sin(lm)\sin(ln) - \sin(mn)}. \quad \text{(A-9)}$$

**Lemma 9 (Lemma 2).** *A canonical region $R$ that is a triangle is always $\alpha$-balanced.*

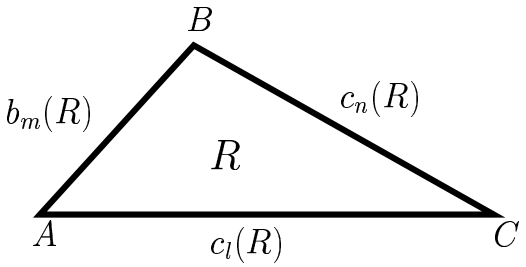


**Fig. A-1.** For Lemma 2

*Proof.* Consider Figure A-1. Let $(l, m, n) \in \mathcal{P}(V)$. Without loss of generality let $d_{\max}$ be $d_n$ and $d_{\min}$ be $d_l$. Note that $d_n = b_m \sin(mn)$ and $d_l = b_m \sin(lm)$. This implies that

$$\frac{d_n}{d_l} = \frac{\sin(mn)}{\sin(lm)} \leq \alpha.$$

**Lemma 10 (Lemma 3).** *Let $(l, m, n) \in \mathcal{P}(V)$. Let $R$ be an $\alpha$-balanced canonical region that is not critically balanced. If $P$ is a critically balanced subregion created by partitioning $R$ with an $l$-cut, then the minimum diameter of $P$ is $d_l(P)$.*

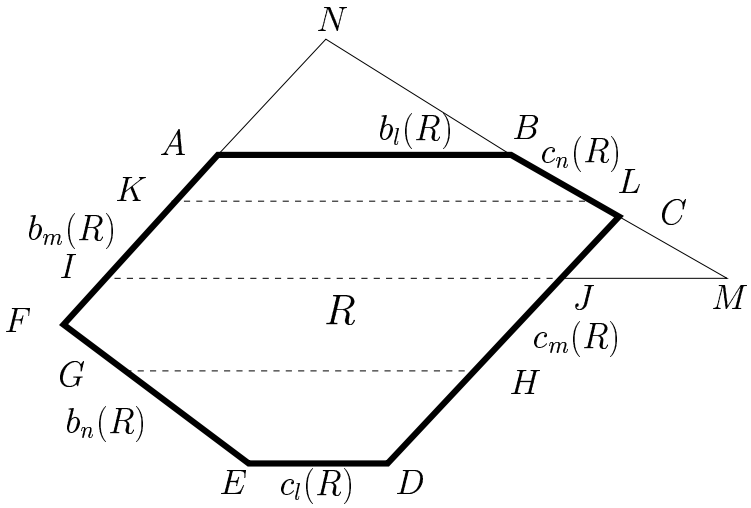


**Fig. A-2.** For Lemma 3

[1] Note that here, and elsewhere, we do not necessarily present all permutations of the partitioning vectors that result in relations of a certain type. We present just the ones that we use in proofs to come. The rest can easily be inferred.

*Proof.* See Figure A-2. Let ABCDEF be the given balanced region $R$. There are three cases for the relative position of the $l$-cut, at GH, at IJ, and at KL.

1. Consider the sub-region $P$ specified by ABCHGF. Note that $d_m(P) = d_m(R)$, $d_n(P) = d_n(R)$, and $d_l(P) < d_l(R)$. Since $R$ is balanced $d_{\min}(P)$ can only be $d_l(P)$.
2. Consider the sub-region $Q$ specified by ABCJI. In this case $d_m(Q) = d_m(R)$, $d_n(Q) < d_n(R)$, and $d_l(Q) < d_l(R)$. So $d_m(Q)$ cannot be $d_{\min}(Q)$. Now consider the triangle $T$ specified by NMI. By Lemma 2 $T$ is balanced. Now $d_n(Q) = d_n(T)$, $d_m(Q) < d_m(T)$, and $d_l(Q) < d_{lT}$. So $d_n(Q)$ cannot be $d_{\min}(Q)$. Thus $d_{\min}(Q)$ can only be $d_l(Q)$.
3. Finally consider the sub-region $V$ specified by ABLK. Also look at the triangle $W$ specified by NCK. By an argument similar to above it can be shown that $d_{\min V}$ can only be $d_{lV}$.

**Corollary 3.** *For a critically balanced canonical region $R$ that is a trapezoid, if $b_l$ and $c_l$ are the two parallel sides, then the minimum diameter of $R$ is $d_l$.*

**Proof Sketch.** Use an argument similar to Case 3 in proof for Lemma 3. □

**Lemma 11 (Lemma 4).** *Let $(l, m, n) \in \mathcal{P}(V)$, and let $R$ be an $\alpha$-balanced region that has no feasible $l$-cut. Let $S$ be the region formed by extending $R$ such that $b_n(S)$ is of length 0. If $P$ is $\mathsf{shield}_{c_m}(S)$, then $c_l(P) \le c_l(R)$.*

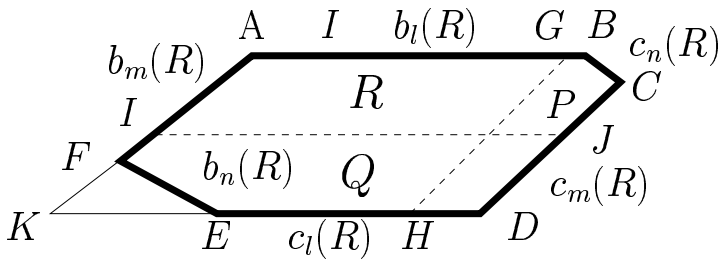


**Fig. A-3.** For Lemma 4

*Proof.* Let ABCDEF be region $R$ (See Figure A-3), and let ABCDEKF be the extended region $S$. Let GBCDH be $P$.

In $P$, from Inequality A-9,

$$d_m(P) \le \frac{c_m(P)\sin(lm)\sin(mn)}{\alpha\sin(lm)\sin(mn) - \sin(ln)}.$$

Further, using inequality $c_l(P) \le d_m(P)/\sin(lm)$, we get

$$c_l(P) \le \frac{c_m(P)\sin(mn)}{\alpha\sin(lm)\sin(mn) - \sin(ln)}. \tag{A-10}$$

Let $Q$ be the region IJDEF, a subregion of $R$ constructed by $l$-cut IJ such that $d_l(Q) = d_l(R)/2$. Since $R$ does not have a feasible $l$-cut, $Q$ is unbalanced. Therefore, from Inequality A-8, we get

$$c_l(R) = c_l(Q) \geq d_l(Q)\left(\alpha - \frac{\sin(mn)}{\sin(lm)\sin(ln)}\right) = \frac{d_l(R)}{2}\left(\alpha - \frac{\sin(mn)}{\sin(lm)\sin(ln)}\right). \tag{A-11}$$

From A-10 and A-11, $c_l(P) \leq c_l(R)$ is implied by

$$\frac{c_m(P)\sin(mn)}{\alpha\sin(lm)\sin(mn) - \sin(ln)} \leq \frac{d_l(R)}{2}\left(\alpha - \frac{\sin(mn)}{\sin(lm)\sin(ln)}\right).$$

Since $c_m(P) \leq d_l(R)/\sin(lm)$, the above is implied by

$$\alpha^2\sin^2(lm)\sin(ln)\sin(mn) + \alpha(\sin(lm)\cos^2(ln) + \sin(lm)(4\cos^2(mn) - 5)) + 2\sin(ln)\sin(mn) \geq 0.$$

which is true as long as $\alpha \geq 4.26/(\sin^2(lm)\sin(ln)\sin(mn))$.

**Lemma 12 (Lemma 5).** *Let $(l, m, n) \in \mathcal{P}(V)$. If an $\alpha$-balanced region $R$ has no feasible $l$-cut, then it has a feasible $m$-cut and a feasible $n$-cut.*

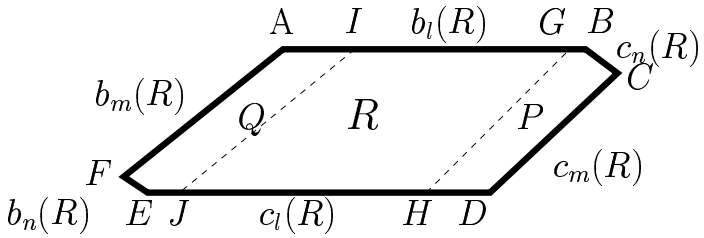


**Fig. A-4.** For Lemma 5.

*Proof.* Let ABCDEF be the region $R$ (See Figure A-4). Let GBCDH be a critically balanced region $P$ constructed by an $m$-cut. Similarly let AIJEF be a critically balanced region $Q$ constructed by an $m$-cut. From Lemma 4, the $m$-cut GH that constructs $P$ intersects $c_l(R)$. We prove the lemma by showing that both $P$ and $Q$ are disjoint — in particular, $c_l(R) \geq c_l(P) + c_l(Q)$.

In $P$, by Lemma 3, $d_{\min}(P)$ is $d_m(P)$. Further, from Equations A-2 and A-3 we get:

$$d_m(P) \geq c_l(P)\sin(lm), \tag{A-12}$$
$$d_n(P) = c_m(R)\sin(mn) + c_l(P)\sin(ln). \tag{A-13}$$

So, if $d_{\max}(P)$ is $d_n(P)$, from A-12 and A-13

$$\alpha c_l(P)\sin(lm) \leq c_m(R)\sin(mn) + c_l(P)\sin(ln),$$

which implies,

$$c_l(P) \leq \frac{c_m(R)\sin(mn)}{\alpha\sin(lm) - \sin(ln)}. \tag{A-14}$$

If, instead, $d_{\max}(P)$ is $d_l(P)$, from Inequality A-12, we get

$$c_l(P) \leq \frac{d_l(P)}{\alpha\sin(lm)} \leq \frac{d_l(R)}{\alpha\sin(lm)}. \tag{A-15}$$

In general, combining A-14 and A-15 gives us

$$c_l(P) \leq \frac{d_l(R) + c_m(R)\sin(mn)}{\alpha\sin(lm) - \sin(ln)}. \tag{A-16}$$

Now in $Q$, from Equation A-2 and Inequality A-6, respectively, we get

$$d_m(Q) = b_n(Q)\sin(mn) + c_l(Q)\sin(lm), \tag{A-17}$$

$$d_n(Q) \leq c_l(Q)\sin(ln) + \frac{d_l(R)}{\sin(lm)}\sin(mn). \tag{A-18}$$

If $d_{\max}(Q)$ is $d_n(Q)$, from A-17 and A-18,

$$\alpha(b_n(R)\sin(mn) + c_l(Q)\sin(lm)) \leq d_l(R)\sin(mn)/\sin(lm) + c_l(Q)\sin(ln),$$

which implies,

$$c_l(Q) \leq \frac{d_l(R)\sin(mn)/\sin(lm) - \alpha b_n(R)\sin(mn)}{\alpha\sin(lm) - \sin(ln)}. \tag{A-19}$$

If, instead, $d_{\max}(Q)$ is $d_l(Q)$, Equation A-17 implies

$$\alpha(b_n(R)\sin(mn) + c_l(Q)\sin(lm)) \leq d_l(Q) \leq d_l(R),$$

which, in turn, is equivalent to,

$$c_l(Q) \leq \frac{d_l(R) - \alpha b_n(R)\sin(mn)}{\alpha\sin(lm)}. \tag{A-20}$$

In general, combining Inequalities A-19 and A-20 gives

$$c_l(Q) \leq \frac{d_l(R)(1 + \sin(mn)/\sin(lm)) - \alpha b_n(R)\sin(mn)}{\alpha\sin(lm) - \sin(ln)}. \tag{A-21}$$

Now consider $R$. We can use an inequality similar to A-11, but instead, we use tighter relations by considering two separate cases.

**Case 1.** $d_{\max}(R')$ is $d_m(R')$. From Equation A-2 we get

$$\begin{aligned}\frac{d_l(R)}{2} = d_l(R') &\leq \frac{d_m(R')}{\alpha} \\ &= \frac{1}{\alpha}(c_l(R')\sin(lm) + b_n(R')\sin(mn)) \\ &\leq \frac{1}{\alpha}(c_l(R)\sin(lm) + b_n(R)\sin(mn)).\end{aligned}$$

The above implies

$$c_l(R) \geq \frac{\alpha d_l(R)}{2\sin(lm)} - \frac{b_n(R)\sin(mn)}{\sin(lm)}$$

Combining the above with Inequalities A-16 and A-21 implies that $c_l(R) \geq c_l(P) + c_l(Q)$ if

$$\frac{\alpha d_l(R)}{2\sin(lm)} - \frac{b_n(R)\sin(mn)}{\sin(lm)} \geq \frac{d_l(R) + c_m(R)\sin(mn)}{\alpha\sin(lm) - \sin(ln)} + \frac{d_l(R)(1 + \sin(mn)/\sin(lm)) - \alpha b_n(R)\sin(mn)}{\alpha\sin(lm) - \sin(ln)}$$

is true. The latter reduces to

$$\alpha^2 \sin(lm) - \alpha\sin(ln) - 4(\sin(lm) + \sin(mn)) \geq 0,$$

a relation that is true as long as $\alpha \geq 3.37/\sin(lm)$.

**Case 2.** $d_{\max}(R')$ is $d_n(R')$. In a manner similar to the previous case, from Equation A-3 we get

$$c_l(R) \geq \frac{\alpha d_l(R)}{2\sin(ln)} - \frac{c_m(R)\sin(mn)}{\sin(ln)}.$$

Combining with A-16 and A-21 implies that $c_l(R) \geq c_l(P) + c_l(Q)$ is true as long as

$$\frac{\alpha d_l(R)}{2\sin(ln)} - \frac{c_m(R)\sin(mn)}{\sin(ln)} \geq \frac{d_l(R) + c_m(R)\sin(mn)}{\alpha\sin(lm) - \sin(ln)} + \frac{d_l(R)(1 + \sin(mn)/\sin(lm)) - \alpha b_n(R)\sin(mn)}{\alpha\sin(lm) - \sin(ln)}$$

is true. The latter reduces to

$$\alpha^2 \sin^2(lm) - \alpha\sin(lm)(2\sin(mn) + \sin(ln)) - \sin(ln)(2\sin(mn) + 4\sin(lm)) \geq 0,$$

which is true as long as $\alpha \geq 4.37/\sin(lm)$.

**Lemma 13 (Lemma 6).** *A critically balanced region $R$ that is a trapezoidal is one-cuttable.*

*Proof.* See Figure A-5. Let $(l, m, n) \in \mathcal{P}(V)$, and ABCD be the critically balanced trapezoidal region $R$. Let AEFD be a critically balanced region $P$ constructed by an $m$-cut. Similarly, let GBCH be a critically balanced region $Q$ constructed by an $n$-cut. From Lemma 4 both GH and EF intersect $b_l(R)$. We prove the lemma by showing that that $P$ and $Q$ are disjoint; in particular, we show that $b_l(R) \geq b_l(P) + b_l(Q)$. This implies that there is a continuum of cuts covering the entire region $R$, and at least one of these cuts is a one-cut.

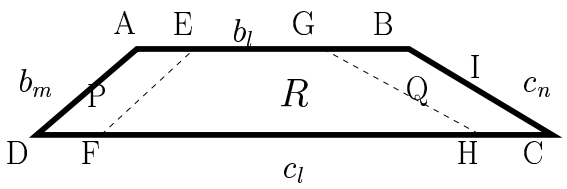


**Fig. A-5.** For Lemma 6

In $R$, Inequality A-8 gives

$$b_l(R) \geq d_l(R)(\alpha - \frac{\sin(mn)}{\sin(lm)\sin(ln)}). \tag{A-22}$$

In $P$, it is easy to see that $d_{\max}(P)$ is $d_n(P)$ and $d_{\min}(P)$ is $d_m(P)$. Further, from basic trigonometry,

$$d_{\max}(P) = \sqrt{(b_l(P))^2 + (c_m(P))^2 - 2b_l(P)c_m(P)\cos(ml)},$$
$$d_{\min}(P) = b_l(P)\sin(lm).$$

The above, along with $c_m(P) = d_l(R)/\sin(lm)$ imply

$$b_l(P) = d_l(R)\frac{\cos(lm) + \sin(lm)\sqrt{\alpha^2 - 1}}{(\alpha^2 \sin^2(lm) - 1)\sin(lm)}. \tag{A-23}$$

Using similar arguments, in $Q$,

$$b_l(Q) = d_l(R)\frac{-\cos(ln) + \sin(ln)\sqrt{\alpha^2 - 1}}{(\alpha^2 \sin^2(ln) - 1)\sin(ln)}. \tag{A-24}$$

From A-22, A-23, and A-24, $b_l(R) \geq b_l(P) + b_l(Q)$ is true if

$$d_l(R)(\alpha - \frac{\sin(mn)}{\sin(lm)\sin(ln)}) \geq$$
$$d_l(R)\frac{\cos(lm) + \sin(lm)\sqrt{\alpha^2 - 1}}{(\alpha^2 \sin^2(lm) - 1)\sin(lm)} + d_l(R)\frac{-\cos(ln) + \sin(ln)\sqrt{\alpha^2 - 1}}{(\alpha^2 \sin^2(ln) - 1)\sin(ln)},$$

which, in turn, is implied by

$$\alpha^2(\sin^2(lm)\sin^2(ln)) - \alpha(\sin(lm)\sin(ln)\sin(mn)) - 4(\sin^2(lm) + \sin^2(ln)) \geq 0.$$

This relation is satisfied as long as $\alpha \geq 3.37/(\sin(lm)\sin(ln))$.

**Lemma 14 (Lemma 7).** *Let $(l, m, n) \in \mathcal{P}(V)$. If an $\alpha$-balanced region $R$ has no one-cut and no feasible $l$-cut, then $R$ has a two-cut.*

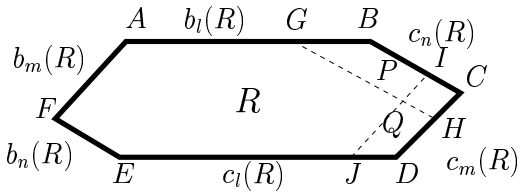


**Fig. A-6.** For Lemma 7

*Proof.* See Figure A-6. Let ABCDEF be the region $R$. Let GBCH be $\mathsf{maxshield}_n(P)$ and let ICDJ be $\mathsf{maxshield}_m(Q)$. We shall prove that at least one of $P$ or $Q$ is a critically balanced trapezoidal region. Note that from Lemma 4, $b_l(P) \leq b_l(R)$ and $c_l(Q) \leq c_l(R)$. So to prove the lemma we just show that if $c_m(P) > c_m(Q)$ then $c_n(Q) \leq c_n(P)$.

In $P$, from Inequality A-9,

$$d_n(P) \leq \frac{c_n(P)\sin(ln)\sin(mn)}{\alpha\sin(ln)\sin(mn) - \sin(lm)}.$$

Combining with inequality $c_m(P) \leq d_n(P)/\sin(mn)$ gives

$$c_m(P) \leq \frac{c_n(P)\sin(ln)}{\alpha\sin(ln)\sin(mn) - \sin(lm)}. \tag{A-25}$$

Similarly, in $Q$,

$$c_n(Q) \leq \frac{c_m(Q)\sin(lm)}{\alpha\sin(ln)\sin(mn) - \sin(lm)}. \tag{A-26}$$

From A-25 and A-26, it is sufficient to show that if

$$\frac{c_n(P)\sin(ln)}{\alpha\sin(ln)\sin(mn) - \sin(lm)} > c_m(Q)$$

then

$$\frac{c_m(Q)\sin(lm)}{\alpha\sin(ln)\sin(mn) - \sin(lm)} \leq c_n(P).$$

The above is true as long as

$$\alpha\sin(ln)\sin(mn) - \sin(lm) \geq \frac{\sin(lm)\sin(ln)}{\alpha\sin(lm)\sin(mn) - \sin(ln)},$$

which is true as long as $\alpha \geq 2/(\sin(lm)\sin(ln)\sin(mn))$.

**Lemma 15 (Lemma 8).** *Let $(l, m, n) \in \mathcal{P}(V)$. If an $\alpha$-balanced region $R$ does not have a one-cut and yet there are feasible cuts along all 3 partitioning vectors $l$, $m$, and $n$, then $R$ has a two-cut.*

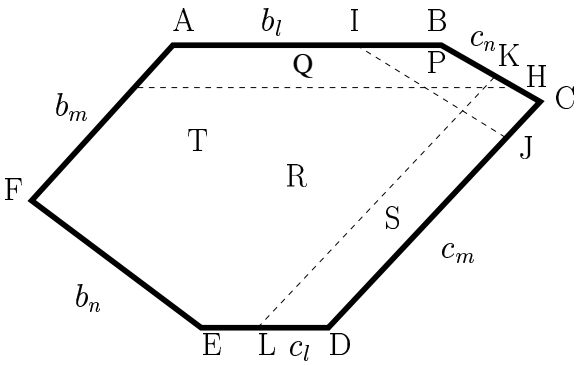


**Fig. A-7.** For Lemma 8

*Proof.* See Figure A-7. Let ABCDEF be the region $R$. Let the maximal shield region with respect to $n$ be $P$ (IBCJ), with respect to $m$ be $S$ (KCDL), and with respect to $l$ be $Q$ (ABHG). We shall show that the "middle" shield region, $P$, is a critically balanced trapezoidal region. In particular we shall show that $b_l(P) \leq b_l(R)$. Let ABKLEF be the region $T$.

Observe that

$$\begin{aligned} d_l(Q) &\leq \frac{b_l(R)\sin(ln)\sin(lm)}{\alpha\sin(lm)\sin(ln) - \sin(mn)}, \\ c_n(Q)\sin(ln) &\leq d_l(Q), \\ c_n(R) - c_n(S) &\leq c_n(Q), \end{aligned}$$

which imply

$$c_n(R) - c_n(S) \leq \frac{b_l(R)\sin(lm)}{\alpha\sin(lm)\sin(ln) - \sin(mn)}.$$

Again, observe that

$$\begin{aligned} c_n(S) &\leq d_m(S)/\sin(mn), \\ d_m(S) &\leq \frac{c_m(S)\sin(mn)\sin(lm)}{\alpha\sin(mn)\sin(lm) - \sin(ln)}, \\ c_m(S)\sin(mn) &\leq d_n(T), \\ d_n(T) &\leq \alpha d_m(T), \\ d_m(T) &= b_l(R)\sin(lm) + (c_n(R) - c_n(S))\sin(mn), \end{aligned}$$

which imply

$$c_n(S) \leq \frac{b_l(R)\sin(lm) + c_n(R)\sin(mn)}{\sin(mn)(\sin(mn) - \sin(ln)/(\alpha\sin(lm)) + 1)}.$$

The above inequalities can be used to prove that $b_l(P) \leq b_l(R)$ as long as

$$\alpha \geq \frac{3.65}{\sin^2(mn)\sin(lm)\sin(ln)},$$

which is true.

**Proof of Time Complexity in Theorem 1.** The PBAR tree algorithm has $O(\log n)$ levels of recursion, where $n$ is the total number of given points. At each level calls are placed to the algorithms OneCut and TwoCut. In these algorithms all the computations involving geometry take $O(1)$ steps as all the geometric objects have a "constant amount of complexity." For example, computing the $\mathsf{maxshield}_l(R)$ takes a constant number of steps as $R$ can have at most 6 sides.

The most time intensive computation in OneCut is the sorting of the points in $S$ with respect to the 3 partitioning vectors. Sorting usually takes $O(|S| \log |S|)$ steps. Note, however, that here the set $S$ to be sorted at a node is already available sorted as part of a bigger sorted set, call it $S'$, at its parent node. We can obtain the sorted order for $S$ with respect to some $l \in V$, from the corresponding sorted order of $S'$ with respect to $l$, in $O(|S'|)$ steps. Thus, at every level the calls to OneCut take $O(n)$ steps. The most time intensive computations in TwoCut are the 3 calls made to OneCut. Thus, at every level the calls to TwoCut also take $O(n)$ steps. The total steps taken by the PBAR algorithm is, therefore, $O(n \log n)$. □

## B Description of Data Sets

The original objective in creating the first 4 synthetic data sets was to come up with data distributions which show a significant difference between BAR and PBAR trees, particularly in the canonical aspect ratios of the nodes. We did not succeed in that. But, later, when we ran approximate nearest neighbor queries we did observe interesting differences in the performance of the two data structures. Set 5 and Set 6 are from the Sloan Digital Sky Survey (SDSS).

**Set 1** An equilateral triangle is taken as a bounding box. Let the vertices be denoted by vectors $a$, $b$, and $c$. Each point $p$ inside the triangle is expressed as a weighted average, $p = w_a \cdot a + w_b \cdot b + w_c \cdot c$. 10,000 points are chosen by choosing $w_a$ uniformly at random from $[0, 1]$, and then choosing $w_b$ uniformly at random from $[0, 1 - w_a]$, and then setting $w_c$ to $1 - w_a - w_b$.

**Set 2** 5,000 points are chosen as in Set 1. Another 5,000 are chosen by choosing $w_a$ uniformly at random from $[0, 0.1]$, and then setting $w_b$ and $w_c$ to $1 - w_a/2$.

**Set 3** 10,000 points are chosen as in Set 1, except that if $w_a$ happens to lie in $[0.25, 0.75]$, 0.25 is subtracted (added) to its value if it is less than (more than equal to) 0.50.

**Set 4** As in Set 1, except that $p$ is set to $w_i \cdot i + (w_i/10.0) \cdot 1, i \in \{a, b, c\}$. A total of 10,000 points are chosen, with $i$ cycling through its 3 values.

**Set 5** From the 5 dimensional color space of the SDSS, the $u$ and $g$ color intensities for 10,000 celestial objects.

**Set 6** The redshift value and the difference between the $u$ and $g$ color intensities for 4487 quasars, obtained from the SDSS.

Set 5 (SDSS) Set 6 (SDSS)

**Fig. B-1.** Two of the data distributions used to evaluate PBAR trees.

## C Pseudo Code for the PBAR Algorithm

PBART($S, R, V, \alpha, \beta$)
*Pre-conditions:*
$V$ is the set of 3 partitioning vectors.
$\alpha$ is the balancing factor, $\alpha \geq f(V)$.
$\beta$ is the reduction factor, $0.5 \leq \beta < 1$.
$S$ is a set of points bounded by $\alpha$-balanced canonical region $R$.
*Post-conditions:*
A PBAR tree $T$ on $S$ is returned.
*Algorithm:*
If($|S| \leq k$) /**Comment: $k$ is a constant with respect to the size of the original input set.*/
Return a root node representing $R$.
Else If(OneCut($S, R, V, \alpha, \beta$) returns a cut)
Let $R_1$ and $R_2$ be the subregions formed by the one-cut.
Let $S_1$ and $S_2$ be the points in $R_1$ and $R_2$ respectively.
$T_1 \leftarrow$ PBART($S_1, R_1, V, \alpha, \beta$).
$T_2 \leftarrow$ PBART($S_2, R_2, V, \alpha, \beta$).
Return a root node representing $R$ and with the roots of $T_1$ and $T_2$ as its child nodes.
Else If(TwoCut($S, R, V, \alpha, \beta$) returns a cut)
Let $R_1$ and $R_2$ be the subregions formed by the two-cut.
Let $S_1$ and $S_2$ be the points in $R_1$ and $R_2$ respectively.
$T_1 \leftarrow$ PBART($S_1, R_1, V, \alpha, \beta$).
$T_2 \leftarrow$ PBART($S_2, R_2, V, \alpha, \beta$).
Return a root node representing $R$ and with the roots of $T_1$ and $T_2$ as its child nodes.
Else Return Error./**Comment: Control never reaches here.*/

**Fig. C-1.** The PBAR tree algorithm.

OneCut($S, R, V, \alpha, \beta$)
*Pre-conditions:*
$V$ is the set of 3 partitioning vectors.
$\alpha$ is the balancing factor.
$\beta$ is the reduction factor, $0.5 \leq \beta < 1$.
$S$ is a set of points bounded by $\alpha$-balanced canonical region $R$.
*Post-condition:*
Returns a one-cut for $R$, if one exists.
*Algorithm:*
For(all $l \in V$)
If($\mathsf{shield}_{b_l}(R) \cap \mathsf{shield}_{c_l}(R) = \emptyset$ and $|\mathsf{maxshield}_l(R)| \leq \beta|R|$ )
/*Comment: one-cut orthogonal to $l$ exists.*/
Sort points in $S$ in order of increasing distance from the (infinite) line at the origin orthogonal to $l$.
Let $p$ be the $\lfloor \beta|R| \rfloor$th point in the sorted order.
Let $c_p$ be the line through $p$ and orthogonal to $l$.
If($c_p \cap \mathsf{shield}_{b_l}(R) = \emptyset$ and $c_p \cap \mathsf{shield}_{c_l}(R) = \emptyset$)
Return $c_p$.
Else If($c_p \cap \mathsf{shield}_{b_l}(R) \neq \emptyset$)
Return cut for $\mathsf{shield}_{b_l}(R)$.
Else
Return cut for $\mathsf{shield}_{c_l}(R)$.
Return. /*Comment: No one-cut exists.*/

**Fig. C-2.** The OneCut Algorithm.

TwoCut($S, R, V, \alpha, \beta$)
*Pre-conditions:*
$V$ is the set of 3 partitioning vectors.
$\alpha$ is the balancing factor.
$\beta$ is the reduction factor, $0.5 \leq \beta < 1$.
$S$ is a set of points bounded by $\alpha$-balanced canonical region $R$.
*Post-condition:*
Returns a two-cut for $R$, if it can find one.
*Algorithm:*
For(all $l \in V$)
$R_1 \leftarrow \mathsf{maxshield}_l(R)$.
$S_1 \leftarrow$ points from $S$ in $R_1$.
If(OneCut($S_1, R_1, V, \alpha, \beta$) returns a cut)
Return the cut for $\mathsf{maxshield}_l(R)$.
Return. /*Comment: Could not find a two-cut.*/

**Fig. C-3.** The TwoCut algorithm.