Streaming Diameter of High-Dimensional Points

Magnús M. Halldórsson
Department of Computer Science, Reykjavik University, Iceland

Nicolaos Matsakis
Department of Computer Science, Reykjavik University, Iceland

Pavel Veselý
Computer Science Institute of Charles University, Prague, Czech Republic

Abstract

We improve the space bound for streaming approximation of Diameter but also of Farthest Neighbor queries, Minimum Enclosing Ball and its Coreset, in high-dimensional Euclidean spaces. In particular, our deterministic streaming algorithms store $\mathcal{O}(\varepsilon^{-2}\log(\frac{1}{\varepsilon}))$ points. This improves by a factor of ε^{-1} the previous space bound of Agarwal and Sharathkumar (SODA 2010), while retaining the state-of-the-art approximation guarantees, such as $\sqrt{2}+\varepsilon$ for Diameter or Farthest Neighbor queries, and also offering a simpler and more complete argument. Moreover, we show that storing $\Omega(\varepsilon^{-1})$ points is necessary for a streaming $(\sqrt{2}+\varepsilon)$-approximation of Farthest Pair and Farthest Neighbor queries.

2012 ACM Subject Classification Theory of computation → Streaming, sublinear and near linear time algorithms

Keywords and phrases streaming algorithm, farthest pair, diameter, minimum enclosing ball, coreset

Digital Object Identifier 10.4230/LIPIcs.ESA.2025.58

Related Version *Full Version*: https://arxiv.org/abs/2505.16720

Funding *Magnús M. Halldórsson*: Partially supported by Icelandic Research Fund grant 2511609.
Nicolaos Matsakis: Partially supported by Icelandic Research Fund grant 2511609 and by Czech Science Foundation project 22-22997S.
Pavel Veselý: Partially supported by Czech Science Foundation project 22-22997S, by ERC-CZ project LL2406 of the Ministry of Education of Czech Republic, and by Center for Foundations of Modern Computer Science (Charles Univ. project UNCE 24/SCI/008).

Acknowledgements We would like to thank Martin Tancer for helpful discussions on high-dimensional geometry.

1 Introduction

In the streaming model, the input data is assumed to be vast and must be processed using limited memory in one or a few passes. Therefore, streaming algorithms "sketch" the input, yielding a small data structure that still accurately preserves desired properties. The research on streaming algorithms has been remarkably fruitful and we now have optimal or near-optimal algorithms for counting distinct elements, frequency moments, quantiles and a plethora of other problems (see [14] for a comprehensive exposition).

We focus on high-dimensional geometric streams, where the input S consists of points in $\mathbb{R}^d$ for d large, a topic of recent interest, e.g., [28, 17, 26, 15, 25, 11, 10, 12, 23]. Extent measures, such as the Diameter or the Minimum Enclosing Ball, are fundamental statistics of a set of points, having a body of work in both streaming and non-streaming settings [2, 22, 19, 3, 8, 7].

In an influential work, Agarwal and Sharathkumar [5] gave a streaming algorithmic framework for several high-dimensional extent problems. Their Blurred Ball Cover data structure maintains a collection of $\mathcal{O}(\frac{1}{\varepsilon^2}\log(\frac{1}{\varepsilon}))$ balls, whose union approximately covers the

33rd Annual European Symposium on Algorithms (ESA 2025).
Editors: Anne Benoit, Haim Kaplan, Sebastian Wild, and Grzegorz Herman; Article No. 58; pp. 58:1–58:10
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

input S. It was then used to approximate a number of high-dimensional extent problems. The claim is that each ball is represented by a coreset of $\mathcal{O}(\frac{1}{\varepsilon})$ points of S, for a total space of $\mathcal{O}(\frac{1}{\varepsilon^3}\log(\frac{1}{\varepsilon}))$ points. It appears though that a somewhat higher space bound is needed for the claimed approximations; see Section A.

We give a modified data structure, Guarded Ball Cover, building on [5]. It allows for both a simpler and more complete treatment, and also results in the smaller space bound of $\mathcal{O}(\frac{1}{\varepsilon^2}\log(\frac{1}{\varepsilon}))$ points. The improved space bound extends to all four applications: approximate Farthest Neighbor queries and for maintaining approximate Farthest Pair (providing an estimate for Diameter), Minimum Enclosing Ball, and Coreset for Minimum Enclosing Ball. This is feasible by storing only a single point per ball, along with its center and radius. Correctness arguments are simplified by also storing the first point of S as a proxy for all points deleted later from memory. We retain the approximation guarantees of all four applications, as in [5]: $1.22+\varepsilon$ for Minimum Enclosing Ball and $\sqrt{2}+\varepsilon$ for each of Diameter, Farthest Neighbor Queries, and Coreset for Minimum Enclosing Ball.

We also show that $\Omega(\varepsilon^{-1})$ points need to be stored for a comparable $(\sqrt{2}+\varepsilon)$-approximation of Farthest Neighbor queries and also for maintaining $(\sqrt{2}+\varepsilon)$-approximate Farthest Pair. This applies to a computational model where the algorithm must return an input point upon a query and the space is determined by the number of points stored; crucially, once a point is deleted from memory, it cannot be retrieved.

2 Preliminaries

Let S be a multiset of points in $\mathbb{R}^d$ that arrive sequentially in a stream. Upon arrival, each point $p \in S$ is either stored in memory (and, possibly, deleted later) or irrevocably discarded. We assume one-pass streaming algorithms in the insertion-only setting. By $\varepsilon \in (0,1]$ we denote an error parameter and by $\alpha > 1$ an approximation guarantee.

An extent measure of a set of points computes certain statistics of either this set or a geometric shape enclosing it [2]. Let $\|pq\|$ denote the Euclidean distance between points $p \in \mathbb{R}^d$ and $q \in \mathbb{R}^d$. The *Diameter* is the maximum Euclidean distance between any pair of points in S and the *Farthest Pair* $\mathsf{FP}(S)$ is a pair of points of S having Euclidean distance equal to the Diameter. The *Farthest Neighbor* of a point q is a point p of the largest Euclidean distance from q. An α-farthest-neighbor α-$\mathsf{FN}(q)$ of a query point $q \in \mathbb{R}^d$ is a point $x \in S$ such that for every $p \in S$ it is $\|qp\| \leq \alpha \cdot \|xq\|$.

By $B(c(B), r(B))$ we denote a ball centered at point $c(B)$ with radius $r(B)$. The $(1+\varepsilon)$-expansion of $B(c,r)$ is defined as $B(c,(1+\varepsilon)r)$. The *Minimum Enclosing Ball* $\mathsf{MEB}(S)$ is the ball of minimum radius containing all points of S. A ball B is α-$\mathsf{MEB}(S)$ if $r(B) \leq \alpha r(\mathsf{MEB}(S))$ and each point of S is within Euclidean distance $r(B)$ from $c(B)$.

For a set of points S, a *coreset* is a set $S' \subseteq S$ preserving a geometric property of S [21]. A set $S' \subseteq S$ is α-$\mathsf{coreset}(S)$ for MEB if each point of S is contained in the α-expansion of $\mathsf{MEB}(S')$.

We focus on computing $(\sqrt{2}+\varepsilon)$-$\mathsf{FN}(q)$ for any query $q \in \mathbb{R}^d$ and maintaining $(1.22+\varepsilon)$-$\mathsf{MEB}(S)$, $(\sqrt{2}+\varepsilon)$-$\mathsf{coreset}(S)$ and $(\sqrt{2}+\varepsilon)$-$\mathsf{FP}(S)$; these ratios are the same as in [5].

2.1 Related Work

The streaming algorithm of Gonzalez [20] computes a 2-MEB by storing the first point p_1 of S and its farthest neighbor q; the enclosing ball is simply $B(p_1, \|p_1 q\|)$. Zarrabi-Zadeh and Chan [29] improved the guarantee of α-MEB to $\alpha = 1.5$ by giving a one-pass streaming algorithm that stores one ball. They also gave a lower bound of $\frac{\sqrt{2}+1}{2} \approx 1.207$ for the guarantee of any deterministic algorithm for α-MEB that stores only one ball.

Badŏiu et al. [8] showed that the number of coreset points approximating $(1+\varepsilon)$-$\mathsf{MEB}(S')$ for a set S' in $\mathbb{R}^d$ does not depend on d. Improved algorithms were given in [24, 6, 7]; however, these algorithms do not work in a streaming fashion.

Preceding the work of Agarwal and Sharathkumar [4], a simple $(1/\sqrt{3})$-approximate two-pass algorithm for the Diameter was given by Egecioglu and Kalantari [16], working in space $\mathcal{O}(d)$.

Following the conference result of [4] that maintains a $(\frac{1+\sqrt{3}}{2}+\varepsilon)$-$\mathsf{MEB}(S)$, Chan and Pathak [9] improved this guarantee to $\alpha = 1.22+\varepsilon$ by employing a detailed analysis to this algorithm. Subsequently, Agarwal and Sharathkumar in their journal paper [5] observed that the guarantee of their algorithm is slightly greater than $\frac{\sqrt{2}+1}{2} \approx 1.207$ by presenting an input for $d = 3$.

On the negative side, any randomized streaming algorithm that maintains α-$\mathsf{FP}(S)$, α-$\mathsf{MEB}(S)$ or α-$\mathsf{coreset}(S)$ with probability at least $2/3$ requires $\Omega(\min\{n, \exp(d^{1/3})\})$ space for certain values of α. These values are $\alpha < \sqrt{2}(1-2/d^{1/3})$ for α-$\mathsf{FP}(S)$ and α-$\mathsf{coreset}(S)$ and $a < (1+\sqrt{2})(1/2-1/d^{1/3})$ for α-$\mathsf{MEB}(S)$, as shown by Agarwal and Sharathkumar [5].

For low d, such as $d = O(1)$ or $d = O(\log\log n)$, the lower bounds do not apply as it is possible to maintain $(1+\varepsilon)$-$\mathsf{FP}(S)$ or answer $(1+\varepsilon)$-$\mathsf{FN}(x)$ queries in a poly-logarithmic space, using an optimized version of the sampling approach of [18]; this applies also to dynamic streams where points may be deleted. For high-dimensional dynamic streams, the best streaming algorithm follows from asymmetric embedding techniques of Indyk [22] but provides $O(1)$-approximation only at the cost of using space polynomial in n.

3 The Guarded Ball Cover

The Guarded Ball Cover is a collection $\mathcal{B}$ of balls that approximately cover all points of S. We represent each ball of $\mathcal{B}$ by a triplet $B = (c, r, q)$, where c is the center of B (possibly $c \notin S$), r is its radius, and q is a point of S inside B. The point q is referred to as the *guard* of B. Our algorithm maintains a coreset Q that consists of the guard points. We treat the first point $p_1 \in S$ specially by always having $p_1 \in Q$.

Let $(1+\varepsilon)\mathcal{B} = \{(c, (1+\varepsilon)r, q) : (c, r, q) \in \mathcal{B}\}$ be the collection of the $(1+\varepsilon)$-expansions of the balls in $\mathcal{B}$. If the arriving point $p \in S$ belongs to $(1+\varepsilon)\mathcal{B}$, then it is discarded. Otherwise, a new ball is added to $\mathcal{B}$. Finally, all balls of too small radius are removed from $\mathcal{B}$.

As the space bound is our primary measure, we assume an exact MEB algorithm, but a good approximation also suffices, specifically within a factor of $1+\varepsilon^2/16$ (using the algorithm of [6] as a subroutine).

Algorithm 1 Algorithm for processing a new point $p \in S$ (excluding the first point p_1).

```
1: if $p \notin (1+\varepsilon)\mathcal{B}$ then                    ▷ If $p$ is outside of the expansions of all guarded balls
2:     $(c, r) \leftarrow \mathsf{MEB}(Q \cup \{p\})$                                  ▷ Compute new MEB
3:     $\mathcal{B} \leftarrow \{(c, r, p)\} \cup \{(c', r', p') \in \mathcal{B} : r' \geq \varepsilon^2 r/80\}$    ▷ Add new ball, delete small balls
4:     $Q \leftarrow \{p_1\} \cup \bigcup_{(c'', r'', q) \in \mathcal{B}} \{q\}$                              ▷ Update coreset
```

The following lemma holds for every MEB computed in line 2 of Algorithm 1:

▶ **Lemma 1** (Lemma 2.2 in [8]). *If $B = B(c, r)$ is the MEB of a set X of points, then any closed half-space containing c also contains a point of X on the boundary of B.*

To analyze the algorithm, we first observe that deleted balls are "guarded" by p_1.

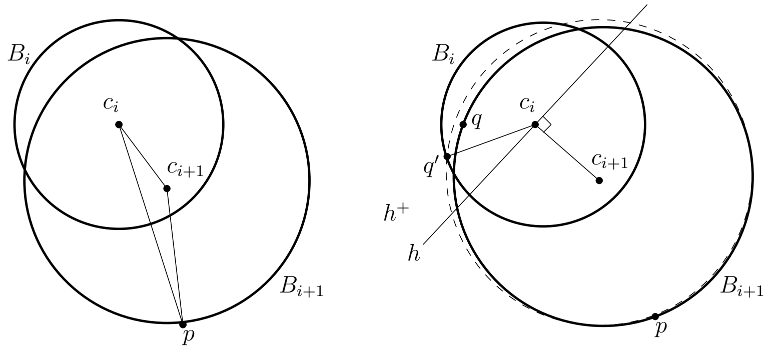

Figure 1 Left: Case $\|c_i c_{i+1}\| \le 5\varepsilon r_i/6$, Right: Case $\|c_i c_{i+1}\| > 5\varepsilon r_i/6$ of Lemma 3. The ball with dashed boundaries is the MEB of all points, assuming that q' is deleted before B_{i+1} is created.

▶ **Lemma 2.** *Suppose ball B is deleted when a ball of radius r is added to $\mathcal{B}$. Then, B is contained in a ball of radius $\varepsilon^2 r/40$ centered at p_1. Consequently, for any guard point q' that has been deleted up to the current time, it holds that $\|q'p_1\| \le \varepsilon^2 r/40$.*

Proof. B contains p_1 and when deleted in line 3, it has radius at most $\varepsilon^2 r/80$. Since the guard of B (which is also evicted from memory) is inside B, it is within distance at most $\varepsilon^2 r/40$ from p_1. ◀

The main technical part is to show that the radii of new balls increase exponentially:

▶ **Lemma 3.** *If $B_{i+1} = (c_{i+1}, r_{i+1}, p)$ is added to $\mathcal{B}$ following $B_i = (c_i, r_i, p_i)$, then it holds that $r_{i+1} \ge (1+\varepsilon^2/8) \cdot r_i$.*

Proof. The proof is similar to that of Lemma 2 in [5] and Claim 2.4 in [8]. We first assume that the exact MEB is computed in line 2 of Algorithm 1. Let Q, $\hat{Q}_D$ be the point sets such that $B_{i+1} = \mathsf{MEB}(Q \cup \{p\})$ and $B_i = \mathsf{MEB}(Q \cup \hat{Q}_D)$, i.e., Q is the coreset just before computing B_{i+1}. Consider two cases:

If $\|c_i c_{i+1}\| \le 5\varepsilon r_i/6$ then $r_{i+1} \ge \|c_{i+1}p\| \ge \|c_i p\| - \|c_i c_{i+1}\| \ge (1+\varepsilon)r_i - 5\varepsilon r_i/6 \ge (1+\varepsilon^2/6)r_i$ (Figure 1, left), using the triangle inequality and that $p \notin (1+\varepsilon)B_i$ (by line 1).

Otherwise $\|c_i c_{i+1}\| > 5\varepsilon r_i/6$. Then let h be the hyperplane passing through c_i with direction $c_i c_{i+1}$ as its normal and let h^+ be the halfspace bounded by h that does not contain c_{i+1}. There is a point $q' \in (Q \cup \hat{Q}_D) \bigcap h^+$ at Euclidean distance r_i from c_i, by Lemma 1. Then, $\|q'c_{i+1}\| \ge (\|c_i c_{i+1}\|^2 + \|q'c_i\|^2)^{1/2} \ge ((5\varepsilon r_i/6)^2 + r_i^2)^{1/2} \ge (1+\varepsilon^2/4)r_i$ (Figure 1, right), where the first inequality follows from the cosine law. By Lemma 2, there is a point $q \in Q$ such that $\|qq'\| \le (\varepsilon^2/40)r_i$ (if $q' \in Q$ then $q = q'$, and $q = p_1$ otherwise). Hence, $r_{i+1} \ge \|qc_{i+1}\| \ge \|q'c_{i+1}\| - \|qq'\| \ge (1+\varepsilon^2/5)r_i$.

Finally, if we use a $(1+\varepsilon^2/16)$-approximate MEB, then we still have that $r_{i+1} \ge (1+\varepsilon^2/5)r_i/(1+\varepsilon^2/16) \ge (1+\varepsilon^2/8)r_i$. ◀

Finally, we show that at any time, the $(1+\varepsilon)$-expansion of any deleted ball is contained in the $(1+\varepsilon)$-expansion of each ball created after the deletion of the former ball.

▶ **Lemma 4.** *If ball $\hat{B}$ was deleted then $(1+\varepsilon)\hat{B} \subset (1+\varepsilon)B_i$, for each $B_i \in \mathcal{B}$ created after the deletion of $\hat{B}$.*

Proof. Let $B_i \in \mathcal{B}$. By Lemmas 2 and 3, it is $\hat{B} \subset B(p_1, \varepsilon^2 r(B_i)/40)$; therefore, we have $(1+\varepsilon)\hat{B} \subset B(p_1, \varepsilon r(B_i)/20)$ as $\varepsilon \le 1$. Since p_1 is in B, it follows that $(1+\varepsilon)\hat{B} \subset (1+\varepsilon/20)B_i$. ◀

Our main result follows from the preceding lemmas:

▶ **Theorem 5.** *$\mathcal{B}$ contains $\mathcal{O}((1/\varepsilon^2)\log(1/\varepsilon))$ balls and $S \subset (1+\varepsilon)\mathcal{B}$.*

Proof. The first claim follows from Lemma 3 and line 3 of Algorithm 1. For the second claim, let $p \in S$. By construction, p is in the $(1+\varepsilon)$-expansion of a ball B that entered $\mathcal{B}$. By Lemma 4, $(1+\varepsilon)B \subset \cup_{B' \in \mathcal{B}}(1+\varepsilon)B'$, whether B was deleted or not. ◀

Differences to the Blurred Ball Cover

The Blurred Ball Cover [5], similarly to the Guarded Ball Cover, initiates the computation of a new ball when an arriving point is not contained in any $(1+\epsilon)$-expansion of a stored ball. The key difference is that in the Blurred Ball Cover, the coreset on the boundary of the new ball, which is returned by the MEB computation and comprises up to $\lceil 1/\epsilon \rceil$ points, is explicitly stored in memory for each ball. (In fact, $\Omega(1/\epsilon^2)$ points may need to be stored for each ball in the Blurred Ball Cover as we discuss in Appendix A.) Coresets belonging to sufficiently small balls are also deleted from the Blurred Ball Cover, though the radius threshold for this deletion is $\approx \varepsilon r$ instead of $\approx \varepsilon^2 r$ in our algorithm. Finally, the first point $p_1 \in S$ is never deleted by the Guarded Ball Cover, contrary to the Blurred Ball Cover which treats p_1 as any other point of S.

Update time

The worst-case update time of our algorithm is $\mathcal{O}(d \cdot \text{poly}(1/\varepsilon))$ when an approximate MEB computation is used [7], which is comparable to that of [5]; the precise update times primarily depend on the MEB computation. (Note that the update times of [5] are also affected by the possible issue mentioned in Appendix A.) We remark that the Blurred Ball Cover of [5] allows for better *amortized* update time as it is possible to perform batched updates, i.e., storing incoming points in a buffer and only running the update procedure when the buffer gets full (see also the second paragraph in Appendix A). This is not possible in our case as we require storing one guard per ball; that is, if we run an update on our sketch with a batch of size C, we may need to store up to C guard points for B_{i+1} if they all end up in the coreset of B_{i+1} (and thus on its boundary).

4 Applications

Farthest Neighbor Queries and Diameter

We largely follow the analysis of [5]. For a query point $x \in \mathbb{R}^d$, the algorithm computes and returns the farthest point in Q: $q' = \arg\max_{q \in Q} \|qx\|$. We show that $\|p'x\| \le (\sqrt{2}+2\varepsilon)\|q'x\|$, where $p' = \arg\max_{p \in S} \|px\|$ is one of the (optimal) farthest points from x.

Let $B_i = (c_i, r_i)$ be the ball in $\mathcal{B}$ of greatest radius that contains p' in its $(1+\varepsilon)$-expansion, which exists by Theorem 5. Applying the triangle inequality, followed by the inequality $x + y \le \sqrt{2(x^2+y^2)}$, we have that

$$\|xp'\| \le \|xc_i\| + \|c_ip'\| \le \|xc_i\| + (1+\varepsilon)r_i \le \sqrt{2}(\|xc_i\|^2 + r_i^2)^{1/2} + \varepsilon r_i \ . \tag{1}$$

By Lemma 1 (for half-space with direction c_ix as normal, c_i on the boundary, and not containing x), when B_i was created, there was a guard z such that: i) $\|zc_i\| = r_i$, ii) $\|xz\| \ge r_i$, and iii) $\angle zc_ix \ge 90°$. By i), iii), and the cosine law, we have

$$\|xz\| \ge (\|xc_i\|^2 + r_i^2)^{1/2} \ . \tag{2}$$

Combining (1) and (2) and using ii), we get that

$$\|xp'\| \leq \sqrt{2}\|xz\| + \varepsilon r_i \leq (\sqrt{2}+\varepsilon)\|xz\| \ . \tag{3}$$

Note that $\|q'x\| \geq r_m/2$, where r_m is the radius of the largest ball in $\mathcal{B}$, as otherwise there is a ball containing Q of radius less than r_m. Also, by Lemma 2, there is a point $w \in Q$ (z or p_1) of distance at most $\varepsilon^2 r_m/80 \leq \varepsilon^2\|q'x\|/40$ from z. By definition of q', $\|wx\| \leq \|q'x\|$. Thus, $\|xz\| \leq \|wx\| + \|wz\| \leq (1+\varepsilon^2/40)\|q'x\|$. Combining this with (3) gives that q' is a $(\sqrt{2}+\varepsilon')$-$\mathsf{FN}(x)$, for $\varepsilon' = 2\varepsilon$.

Finally, for the closely related problem of Diameter (Farthest Pair), we return $\mathsf{FN}(p)$ for each point $p \in S$. If $\bar{p} = \mathsf{FN}(p)$ and $\|p\bar{p}\|$ exceeds the stored value for Diameter, then we replace the old Farthest Pair with the new pair $(p, \bar{p})$.

▶ **Corollary 6.** *For a stream S of points in $\mathbb{R}^d$, the Guarded Ball Cover of $\mathcal{O}((1/\varepsilon)^2\log(1/\varepsilon))$ stored points answers $(\sqrt{2}+\varepsilon)$-FN(x) for any query $x \in \mathbb{R}^d$, and maintains $(\sqrt{2}+\varepsilon)$-FP(S).*

Minimum Enclosing Ball

The following theorem was shown by Chan and Pathak [9] and improved the guarantee of the approximate MEB algorithm of Agarwal and Sharathkumar to $1.22+\varepsilon$ (see [5], p. 91):

▶ **Theorem 7** (Theorem 1 in [9], Theorem 1 in [27]). *Let $K_1, ..., K_u$ be subsets of a point set S in $\mathbb{R}^d$, with $B_i = \mathit{MEB}(K_i)$ such that: i) $r(B_i)$ is increasing over i, and ii) $K_i \subset (1+\varepsilon)B_j$, for each $i < j$. Then, $r(B) \leq (1.22+\varepsilon)\cdot r(\mathit{MEB}(S))$, where $B = \mathit{MEB}(\bigcup_{i=1}^{u} B_i)$.*

In our case, K_i is the coreset on the boundary of the MEB computed in line 2 of Algorithm 1. Therefore, the first requirement of Theorem 7 holds by Lemma 3. The second requirement follows immediately for points in Q and by Lemma 4 for points deleted from Q.

▶ **Corollary 8.** *For a stream S of points in $\mathbb{R}^d$, the Guarded Ball Cover of $\mathcal{O}((1/\varepsilon^2)\log(1/\varepsilon))$ stored points maintains $(1.22+\varepsilon)$-MEB(S).*

Coreset for Minimum Enclosing Ball

We mostly follow the analysis of [5]. Let $B = (c, r_m)$ be the most recently created ball added to $\mathcal{B}$ and let Q be the coreset at the time of computing B, thus $B = \mathsf{MEB}(Q)$. Note that $\mathsf{MEB}(S)$ has radius at least r_m, since B is an MEB of a subset of S. We claim that $(\sqrt{2}+\varepsilon)B$ contains all points in S, which implies that Q forms a $(\sqrt{2}+\varepsilon)$-coreset(S). Namely, we show that each point $y \in S$ has $\|yc\| \leq (\sqrt{2}+\varepsilon')r_m$, for $\varepsilon' = 2\varepsilon$.

Consider a point $y \in S$ that is farthest from c. Let $B_i = (c_i, r_i)$ be a guarded ball that has not been deleted and whose $(1+\varepsilon)$-expansion contains y (B_i is well-defined by Theorem 5). By the triangle inequality and the definition of B_i, $\|yc\| \leq \|cc_i\| + \|yc_i\| \leq \|cc_i\| + (1+\varepsilon)r_i$. Let h be the hyperplane passing through c_i with direction cc_i as normal and let h^+ be the halfspace bounded by h that does not contain c. By Lemma 1 there is a guard g in h^+ with $\|gc_i\| = r_i$, and by the cosine law it is $\|gc\| \geq \sqrt{\|cc_i\|^2 + r_i^2}$. Then (using the inequality $a + b \leq \sqrt{2(a^2+b^2)}$),

$$\frac{\|yc\|}{\|gc\|} \leq \frac{\|cc_i\| + (1+\varepsilon)r_i}{\sqrt{\|cc_i\|^2 + r_i^2}} \leq \frac{\sqrt{2(\|cc_i\|^2 + r_i^2)} + \varepsilon \cdot r_i}{\sqrt{\|cc_i\|^2 + r_i^2}} \leq \sqrt{2} + \varepsilon \ .$$

By Lemma 2, there is a guard $q \in Q$ with $\|qg\| \leq \varepsilon^2 r_m/40$, so by the triangle inequality, $\|gc\| \leq \|qc\| + \|gq\| \leq (1+\varepsilon^2/40)r_m$. Hence, $\|yc\| \leq (\sqrt{2}+\varepsilon)(1+\varepsilon^2/40)r_m \leq (\sqrt{2}+2\varepsilon)r_m$.

▶ **Corollary 9.** *For a stream S of points in $\mathbb{R}^d$, the Guarded Ball Cover of $\mathcal{O}((1/\varepsilon)^2\log(1/\varepsilon))$ stored points maintains $(\sqrt{2}+\varepsilon)$-coreset(S).*

5 Lower Bound for Farthest Pair and Farthest Neighbor Queries

We show that computing a $\sqrt{2}$-approximation (with $\varepsilon = 0$) of Farthest Neighbor queries or Farthest Pair is impossible in the streaming model without returning points outside of S.

This applies to the computational model of "coreset-based algorithms" in which the space bound is counted in the number of input points stored and the algorithm must return an input point upon a query (or two input points for the Farthest Pair); crucially, once a point is deleted from memory, it cannot be retrieved. This model is akin to comparison-based model in sorting or selection, as used in [13] for streaming lower bounds for quantile estimation.

▶ **Theorem 10.** *For any $\varepsilon > 0$ and $d = \Omega(1/\varepsilon)$, any coreset-based randomized streaming algorithm answering approximate Farthest Neighbor queries or maintaining the approximate Farthest Pair in $\mathbb{R}^d$ with multiplicative error $\leq \sqrt{2} + \varepsilon$, has to store $\Omega(1/\varepsilon)$ points.*

Proof. We use the easy direction of Yao's minimax principle and design a distribution over instances (points and Farthest Neighbor queries) so that any deterministic streaming algorithm using space $o(1/\varepsilon)$ will, with high constant probability, answer a $\mathsf{FN}(q)$ query incorrectly, i.e., the point returned on the query q will be more than $(\sqrt{2} + \varepsilon)$-factor closer to q than the farthest point. The same argument applies to Farthest Pair.

Suppose without loss of generality that $2/\varepsilon \in \mathbb{Z}$. We insert $k := 1/(2\varepsilon) + 1$ points of the standard basis, i.e., $\mathbf{e}_i = (0, 0, \dots, 1, 0, \dots, 0)$, where the i-th coordinate is 1, for $i = 0, \dots, k-1$ (the order of insertions does not matter). The random part of the construction is to choose $j \in \{0, \dots, k-1\}$ uniformly at random and make a Farthest Neighbor query for point $q_j = (2\varepsilon, 2\varepsilon, \dots, -1, 2\varepsilon, \dots, 2\varepsilon, 0, \dots, 0)$, where the coordinate j is -1 and only the first k coordinates are not 0. Clearly, the farthest point from q_j is e_j and their Euclidean distance is $\sqrt{4 + (k-1) \cdot 4\varepsilon^2} = \sqrt{4 + 2\varepsilon}$, using the choice of k.

However, with some constant probability, point e_j is not stored as the algorithm stores $o(1/\varepsilon) = o(d)$ points. Conditioning on the event that e_j is not stored, the algorithm needs to answer the query with a point e_i for $i \neq j$. However, the distance between q_j and e_i with $i \neq j$ is $\sqrt{(1-2\varepsilon)^2 + 1 + (k-2) \cdot 4\varepsilon^2} = \sqrt{2 - 4\varepsilon + (k-1) \cdot 4\varepsilon^2} = \sqrt{2 - 2\varepsilon}$, using the choice of k. Thus, the approximation ratio of the algorithm is at least $\sqrt{(4 + 2\varepsilon)/(2 - 2\varepsilon)} > \sqrt{2} + \varepsilon$. ◀

6 Conclusions and Open Problems

We have designed streaming algorithms storing $\mathcal{O}((1/\varepsilon^2) \log(1/\varepsilon))$ points from $\mathbb{R}^d$ that can estimate several extent statistics of the input. All error guarantees are almost optimal, with the only exception of the Minimum Enclosing Ball application, where there exists a small gap between the approximation guarantee of $1.22 + \varepsilon$ and the lower bound converging to $(\sqrt{2} + 1)/2 \approx 1.207$ for $d \to \infty$. This is achieved by simplifying (and also fixing) the Blurred Ball Cover from [5] into a "Guarded Ball Cover", where we store $\Theta(1)$ points per ball, compared to $\text{poly}(1/\varepsilon)$ points per ball for the Blurred Ball Cover.

We believe that the space bound can be improved, at least by shaving off the $\log(1/\varepsilon)$ factor. One possible direction for improvement is the use of randomization as both our algorithms and those of [5] are deterministic, while the simple lower bound of $\Omega(1/\varepsilon)$ for "coreset-based" algorithms holds even when randomization is used.

While our algorithms are more space-efficient than that of [5], the amortized update times are somewhat worse, as discussed in Section 3. Thus, we ask if it is possible to optimize the $\text{poly}(d \cdot \varepsilon^{-1})$ amortized update time while retaining the near-quadratic space bound. More importantly, it would be interesting to develop a "fast" streaming algorithm for extent problems, that is, with amortized update time $\mathcal{O}(d \cdot \text{poly}(\log \varepsilon^{-1}))$.

Beyond the streaming setting, an important property of data sketches is mergeability [1], which enables to summarize the input in a parallel or distributed way and then merge the resulting sketches into one summary of the whole dataset. It is open how to design a merge operation for Guarded (or Blurred) Ball covers while retaining the space and approximation guarantees.

References

1 Pankaj K. Agarwal, Graham Cormode, Zengfeng Huang, Jeff M. Phillips, Zhewei Wei, and Ke Yi. Mergeable summaries. *ACM Trans. Database Syst.*, 38(4):26, 2013. doi:10.1145/2500128.

2 Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Approximating extent measures of points. *J. ACM*, 51(4):606–635, 2004. doi:10.1145/1008731.1008736.

3 Pankaj K. Agarwal, Jiří Matoušek, and Subhash Suri. Farthest neighbors, maximum spanning trees and related problems in higher dimensions. *Comput. Geom.*, 1:189–201, 1991. doi:10.1016/0925-7721(92)90001-9.

4 Pankaj K. Agarwal and R. Sharathkumar. Streaming algorithms for extent problems in high dimensions. In *Proceedings of the 21st ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1481–1489. SIAM, 2010. doi:10.1137/1.9781611973075.120.

5 Pankaj K. Agarwal and R. Sharathkumar. Streaming algorithms for extent problems in high dimensions. *Algorithmica*, 72(1):83–98, 2015. doi:10.1007/S00453-013-9846-4.

6 Mihai Badoiu and Kenneth L. Clarkson. Smaller core-sets for balls. In *Proceedings of the 14th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 801–802. ACM/SIAM, 2003. URL: http://dl.acm.org/citation.cfm?id=644108.644240.

7 Mihai Badoiu and Kenneth L. Clarkson. Optimal core-sets for balls. *Comput. Geom.*, 40(1):14–22, 2008. doi:10.1016/J.COMGEO.2007.04.002.

8 Mihai Badoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proceedings of the 34th ACM Symposium on Theory of Computing (STOC)*, pages 250–257. ACM, 2002. doi:10.1145/509907.509947.

9 Timothy M. Chan and Vinayak Pathak. Streaming and dynamic algorithms for minimum enclosing balls in high dimensions. In *Proceedings of the 12th Workshop on Algorithms and Data Structures (WADS)*, pages 195–206. Springer, 2011. doi:10.1007/978-3-642-22300-6_17.

10 Xi Chen, Vincent Cohen-Addad, Rajesh Jayaram, Amit Levi, and Erik Waingarten. Streaming Euclidean MST to a constant factor. In *Proceedings of the 55th ACM Symposium on Theory of Computing (STOC)*, pages 156–169. ACM, 2023. doi:10.1145/3564246.3585168.

11 Xi Chen, Rajesh Jayaram, Amit Levi, and Erik Waingarten. New streaming algorithms for high dimensional EMD and MST. In *Proceedings of the 54th ACM Symposium on Theory of Computing (STOC)*, pages 222–233. ACM, 2022. doi:10.1145/3519935.3519979.

12 Xiaoyu Chen, Shaofeng H.-C. Jiang, and Robert Krauthgamer. Streaming Euclidean max-cut: Dimension vs data reduction. In *Proceedings of the 55th ACM Symposium on Theory of Computing (STOC)*, pages 170–182. ACM, 2023. doi:10.1145/3564246.3585170.

13 Graham Cormode and Pavel Veselý. A tight lower bound for comparison-based quantile summaries. In *Proceedings of the 39th ACM Symposium on Principles of Database Systems (PODS)*, pages 81–93. ACM, 2020. doi:10.1145/3375395.3387650.

14 Graham Cormode and Ke Yi. *Small Summaries for Big Data.* Cambridge University Press, 2020.

15 Artur Czumaj, Shaofeng H.-C. Jiang, Robert Krauthgamer, Pavel Veselý, and Mingwei Yang. Streaming facility location in high dimension via geometric hashing. In *Proceedings of the 63rd IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 450–461. IEEE, 2022. doi:10.1109/FOCS54457.2022.00050.

16 Ömer Egecioglu and Bahman Kalantari. Approximating the diameter of a set of points in the Euclidean space. *Inf. Process. Lett.*, 32(4):205–211, 1989. doi:10.1016/0020-0190(89)90045-8.

17 Hossein Esfandiari, Praneeth Kacham, Vahab Mirrokni, David P. Woodruff, and Peilin Zhong. High-dimensional geometric streaming for nearly low rank data. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL: https://openreview.net/forum?id=yQfA0etfB7.

18 Gereon Frahling, Piotr Indyk, and Christian Sohler. Sampling in dynamic data streams and applications. *Int. J. Comput. Geom. Appl.*, 18(1/2):3–28, 2008. doi:10.1142/S0218195908002520.

19 Ashish Goel, Piotr Indyk, and Kasturi R. Varadarajan. Reductions among high dimensional proximity problems. In *Proceedings of the 12th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 769–778. ACM/SIAM, 2001. URL: http://dl.acm.org/citation.cfm?id=365411.365776.

20 Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theor. Comput. Sci.*, 38:293–306, 1985. doi:10.1016/0304-3975(85)90224-5.

21 Sariel Har-Peled. *Geometric Approximation Algorithms.* American Mathematical Society, 2011.

22 Piotr Indyk. Better algorithms for high-dimensional proximity problems via asymmetric embeddings. In *Proceedings of the 14th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 539–545. ACM/SIAM, 2003. URL: http://dl.acm.org/citation.cfm?id=644108.644200.

23 Shaofeng H.-C. Jiang, Robert Krauthgamer, and Shay Sapir. Moderate dimension reduction for k-center clustering. In *Proceedings of the 40th International Symposium on Computational Geometry (SoCG)*, volume 293, pages 64:1–64:16, 2024. doi:10.4230/LIPICS.SOCG.2024.64.

24 Piyush Kumar, Joseph S. B. Mitchell, and E. Alper Yildirim. Approximate minimum enclosing balls in high dimensions using core-sets. *ACM J. Exp. Algorithmics*, 8, 2003. doi:10.1145/996546.996548.

25 Sepideh Mahabadi, Ilya P. Razenshteyn, David P. Woodruff, and Samson Zhou. Non-adaptive adaptive sampling on turnstile streams. In *Proceedings of the 52nd ACM Symposium on Theory of Computing (STOC)*, pages 1251–1264. ACM, 2020. doi:10.1145/3357713.3384331.

26 Yury Makarychev, Naren Sarayu Manoj, and Max Ovsiankin. Near-optimal streaming ellipsoidal rounding for general convex polytopes. In *Proceedings of the 56th ACM Symposium on Theory of Computing (STOC)*, pages 1526–1537. ACM, 2024. doi:10.1145/3618260.3649692.

27 Vinayak Pathak. Streaming and dynamic algorithms for minimum enclosing balls in high dimensions. Master's thesis, University of Waterloo, 2011.

28 David P. Woodruff and Taisuke Yasuda. High-dimensional geometric streaming in polynomial space. In *Proceedings of the 63rd IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 732–743. IEEE, 2022. doi:10.1109/FOCS54457.2022.00075.

29 Hamid Zarrabi-Zadeh and Timothy M. Chan. A simple streaming algorithm for minimum enclosing balls. In *Proceedings of the 18th Annual Canadian Conference on Computational Geometry (CCCG)*, 2006.

A A Note on Lemma 2 in [5]

We report on a possible issue in the proof of Lemma 2 in [5] (a similar issue appears in the conference version [4]). Lemma 2 in [5] states that for any i, $r(B_{i+1}) \geq (1 + \varepsilon^2/8) \cdot r(B_i)$, where B_i and B_{i+1} are two consecutive balls of the Blurred Ball Cover. The property that the radii of balls increase geometrically is crucially needed to bound the space requirements.

The proof of Lemma 2 goes as follows: Ball $B_{i+1} = \textsf{APPROX-MEB}(\bigcup_{j\leq i} K_j \cup A), \varepsilon/3)$ is created (K_j is the coreset of B_j and APPROX-MEB is the subroutine of [7] that computes it) because one of the points in a buffer A is not contained in any $(1 + \varepsilon)$-expansion of a ball B_j for $j \leq i$. Define ball B' as the MEB of $\bigcup_{j \leq i} K_j \cup A$. It is subsequently claimed that $r(B_{i+1}) \geq r(B')$, without a proof. However, B' is the MEB of $\bigcup_{j \leq i} K_j \cup A$, while B_{i+1} is an *approximate* MEB for $\bigcup_{j \leq i} K_j \cup A$, namely the $(1 + \varepsilon/3)$-expansion of B_{i+1} contains all these points but a smaller expansion of B_{i+1} may not contain them all.

We observe that this is fixable by computing a tighter MEB approximation $B_{i+1} = \mathsf{APPROX\text{-}MEB}(\bigcup_{j \leq i} K_j \cup A), \varepsilon^2/16)$. That, however, may result in a coreset K_{i+1} of size $\Theta(\varepsilon^{-2})$, which (unlike our approach) increases the space bound in [5] from $\mathcal{O}(\varepsilon^{-3} \log(1/\varepsilon))$ to $\mathcal{O}(\varepsilon^{-4} \log(1/\varepsilon))$. We are unaware of a fix that does not affect the space bound or error guarantees.