Distance Sensitive Bloom Filters Without False Negatives*

Mayank Goswami[1], Rasmus Pagh[2], Francesco Silvestri[2], and Johan Sivertsen[2]

[1]Max-Planck Institute for Informatics, Germany, gmayank@mpi-inf.mpg.de
[2]IT University of Copenhagen, Denmark, {pagh,fras,jovt}@itu.dk

July 18, 2016

Abstract

A Bloom filter is a widely used data-structure for representing a set S and answering queries of the form "Is x in S?". By allowing some false positive answers (saying 'yes' when the answer is in fact 'no') Bloom filters use space significantly below what is required for storing S. In the *distance sensitive* setting we work with a set S of (Hamming) vectors and seek a data structure that offers of similar trade-off, but answers queries of the form "Is x *close* to an element of S?" (in Hamming distance). Previous work on distance sensitive Bloom filters have accepted false positive *and* false negative answers. Absence of false negatives is of critical importance in many applications of Bloom filters, so it is natural to ask if this can be achieved also in the distance sensitive setting. Our main contribution is upper and lower bounds (that are tight in several cases) for space usage in the distance sensitive setting where false negatives are not allowed.

*The research leading to these results has received funding from the European Research Council under the EU 7th Framework Programme, ERC grant agreement no. 614331.

1 Introduction

The Bloom filter [4] is a well-known data structure for answering *approximate membership queries* on a set S, i.e., queries of the form "Is x in S?". Bloom filters are widely used in practice because they require less space than a dictionary data structure for storing S. This is achieved by allowing a certain probability of *false positives*, i.e., 'yes' answers for queries $x \notin S$. It is critical for many applications of Bloom filters that errors are one-sided, i.e., 'no' answers are always correct. In other words, *false negatives* do not occur.

Generally the set S that we want to ask questions about is a subset from some much larger domain. In applications of Bloom filters the answer to a membership query should most often be negative, and for the vast majority of such queries the Bloom filter will give the correct answer. Whenever the filter does give a positive answer, correctness can often be checked using a slower, less space-efficient method (maybe even on a different machine). Bloom filters are often used as part of an exact two-level data structure where the first level is cheap to use and does most of the work, and the second level is more expensive but only rarely needed. Having false negatives means this setup fails, and the user would have to either accept some possibility of getting a wrong answer or perform an expensive exact query every time.

In this paper we present upper and lower bounds on the space complexity of data structures for answering *distance sensitive approximate membership queries*. These answer queries of the form "Is s close to some element S?" Specifically we address this question in the d-dimensional Hamming space where $x \in \{0,1\}^d$, $S \subset \{0,1\}^d$ and "close" means within Hamming distance r. In contrast to previous work on this problem, the data structures presented in this paper introduce no false negatives.

For some settings the problem changes a lot depending on an approximation factor $c \geq 1$ defined such that we only require a small false positive rate for points at distance cr or more from the query point,i.e, for points at distances between r and cr no guarantee is given. This kind of approximation of distances is standard in data structures for high-dimensional search.

Motivation. There are many potential applications for this kind of data structure. As a concrete example consider a journal comprising a large collection of academic papers. When accepting a new paper the journal might want to check if the new paper is very similar to any prior work already published. By using a distance-sensitive Bloom filter this can be done in a space-efficient manner. Because we do not allow false negatives any new paper passing this test (with a 'no' result) is guaranteed to be significantly different from all prior work. The test could be carried out as part of the submission process using a web system. In the rare case that a paper fails the test the submission process could be halted pending a consultation of the full archive.

Furthermore, since the filter provides very little information about the content of the papers it would not need to be subjected to the same access control as a full database of all the journals papers might be under. More interesting examples of applications for distance-sensitive Bloom filters can be found in [12] and for Bloom filters in general in [5].

Our results. We study the space required for answering distance-sensitive approximate membership queries with no false negatives, i.e., one-sided error probability ε. It turns out that, in contrast to approximate membership, we get different bounds depending on how the false positive rate is defined:

- If we desire a *point-wise* error bound (Definition 1, 2) for each query at distance $\geq cr$ from S, the space usage must be $\Omega\left(n\left(\frac{r^2}{d}+\log\frac{1}{\varepsilon}\right)\right)$ for almost all instances, and $\Omega\left(n\left(\frac{r}{c}+\frac{c}{c-1}\log\frac{1}{\varepsilon}\right)\right)$ bits if n is small.
- If it suffices to have an ε *average* false positive rate over all queries at distance $\geq cr$ from S, the space usage must be $\Omega\left(n\left(\frac{r^2}{d}+\log\frac{1}{\varepsilon}\right)\right)$ bits.

We match these lower bounds with almost tight upper bounds on space usage. We introduce the notion of vector signature, which can be seen as a succinct version of a CountSketch [6], and then show how to use them to design filters with point-wise and average errors.

Our focus is on space usage rather than query time, and indeed it would be surprising if polylogarithmic query time in n would be possible since our (point-wise) data structure could be used, say with $\varepsilon = 1/n$, to solve the c-approximate nearest neighbor problem, and the currently best data structures for this problem use $n^{\Omega(1/c)}$ time [2].

Related work. There is little prior work specifically on distance-sensitive approximate membership. The problem corresponds to querying a standard Bloom filter in a ball around the query point, but this solution is slow, time $\Omega(\binom{d}{r})$, and also not particularly space efficient since we would need to use a Bloom filter with a very small false positive rate to bound the probability that none of the queries yield a false positive. More precisely, the required space usage for this approach would be $\Omega(nr\log\frac{d}{r})$ bits.

Mitzenmacher and Kirsch [12] considered data structures that look like Bloom filters but replace standard hash functions with locality sensitive hash (LSH) functions [11] to achieve distance sensitivity. However, this approach introduces false negatives because LSH is not guaranteed to produce collisions. In order to reduce the number of false negatives the conjunction used when querying Bloom filters is replaced by a threshold function: There should just be "many" hash collisions. Unfortunately, the achieved approximation factor is large, $c = O(\log n)$. Hua et al. [10] extended the data structure of [12] with practical improvements and provided extensive experiments, confirming that false negatives also appear in practice.

There has been some recent progress on developing LSH families that can answer near neighbor queries without false negatives [15], but it seems inherent to such families that the storage cost grows exponentially with r. Thus this approach is not promising perhaps except for very small values of r.

Finally, we note that it is known that allowing a constant fraction of false negatives does not significantly affect the space usage that can be achieved by approximate membership data structures [16].

2 Problem definition and notation

The Hamming distance $D(p,q)$ between two points $p,q \in \{0,1\}^d$ is the number of positions where p and q differ. Given a set $S \subseteq \{0,1\}^d$ and a point $q \in \{0,1\}^d$, we overload the meaning of $D(\cdot)$ by defining $D(q,S)$ to be the minimum distance between q and any point in S, i.e. $D(q,S) = \min_{p \in S} D(q,p)$. We let $B_d(q,r)$ be the Hamming ball of radius r centred around q, that is $B_d(q,r) = \{p \in \{0,1\}^d, D(p,q) \leq r\}$.

In this paper we study *distance-sensitive approximate membership filters* defined as follows:

Definition 1 (Point-wise error)**.** *Let $r \geq 0$, $c \geq 1$, and $\varepsilon \in [0,1)$. Given a set $S \subset \{0,1\}^d$ define the two sets $Q_{near} = \{x \in \{0,1\}^d : D(x,S) \leq r\}$ and $Q_{far} = \{x \in \{0,1\}^d : D(x,S) \geq cr\}$. A (r,c,ε)-distance-sensitive approximate membership filter for set S is a data-structure that reports:*

- Yes *for all points in Q_{near};*
- No *with probability at least $1-\varepsilon$ for every point in Q_{far}.*

The filter has a strong per point guarantee, it requires a false positive probability at most ε for the worst case point from Q_{far}. Sometimes only a weaker guarantee is necessary, it might be acceptable that some points give false positives in every instance of the data structure, as long as only an ε total fraction of points in Q_{far} give false positives. We call this weaker, but often good enough, filter the average error version:

Definition 2 (Average error). *The definition is the same as definition 1, except the filter reports:*

- Yes *for all points in* Q_{near}*;*
- No *for at least a* $(1-\varepsilon)$ *fraction of the points in* Q_{far}.

Though the difference between these two definitions seems small, their properties and analysis differs substantially. In the rest of the paper we will give space bounds for both versions, referring to them respectively as the *point-wise error* or *average error* setting of the problem.

3 Lower bounds

We first investigate what can be done when no errors are allowed.

Theorem 1 (The $\varepsilon=0$ case). *Assume that* $\varepsilon=0$*; then any data structure for the average error version of the problem must use* $n\log(2^d/en|B_d(cr)|)$ *bits in the worst case. If* $\log n=o(d)$ *and* $cr=o(d/\log d)$ *then any data structure must use* $\Omega(nd)$ *bits. Thus up to constant factors, the optimal data structure is no better than one that stores the exact set explicitly.*

Proof. The proof is an encoding argument. Assume that a set $S\subseteq\{0,1\}^d$ of size n is to be encoded. Assume the optimal data structure uses s bits in the worst case. The encoder inserts the given set S into the data structure, and runs the query algorithm on each point in the universe. The data structure says yes to at most $n|B_d(cr)|$ many points (including S). The encoder encodes S as a subset of these positives, thereby using at most $\log\binom{n|B_d(cr)|}{n}$ bits. The encoder then send these bits and the at most s bits to the decoder.

The decoding procedure is fairly straightforward. The decoder runs the query on all points in the universe on the representation of s bits. This returns the set of positives. Using the second set of bits written by the encoder, the decoder retrieves the set S, concluding the procedure.

Since every set S can be encoded, we get that

$$\begin{aligned} s+\log\binom{n|B_d(cr)|}{n} &\geq \log\binom{2^d}{n} \\ \Rightarrow s &\geq \log\left(\left(\frac{2^d}{n}\right)^n \Big/ \left(\frac{en|B_d(cr)|}{n}\right)^n\right) \\ \Rightarrow s &\geq n\log\left(\frac{2^d}{en|B_d(cr)|}|\right) \end{aligned}$$

This simplifies to $s\geq nd-n\log(en)-n\log|B_d(cr)|$. If $\log n=o(d)$, we get $s=\Omega(nd-n\log|B_d(cr)|)$. Further, using that $|B_d(cr)|=\sum_{i=0}^{cr}\binom{d}{i}<d^{cr}$ for $cr<d/2$, we get that $s=\Omega(nd-ncr\log d)$, which is $\Omega(nd)$ when $cr=o(d/\log d)$, establishing the theorem.

□

3.1 Average error

Next we investigate the average error distance-sensitive membership problem with $\varepsilon>0$. We let $TP(n,r,c,d)$ (True Positives) denote the quantity that is n times the volume of a ball of radius cr in a d-dimensional space, i.e., $TP(n,r,c,d)=n|B_d(cr)|$. TP is an upper bound on the number of positives not considered *false*. We fix n,r,c and d, and abbreviate $TP(n,r,c,d)$ as TP.

Theorem 2. *Assume that* $TP/2^d<\varepsilon<1/4$. *Then any distance-sensitive membership data structure for the average error version must use* $\Omega(n(r^2/d+\log(1/\varepsilon)))$ *bits in the worst case. Note that as long as* $TP/2^d<\varepsilon$ *we can have* $c=1$ *in this setting.*

Remarks:

1. The above theorem holds as long as $TP < 2^{d-2}$ (the union of cr balls around the input point set is less than a quarter of the full Hamming space) and $TP/2^d < \varepsilon < 1/4$. This is the most interesting range of parameters. As we will see later, the $\Omega(nr^2/d)$ lower bound holds as long as $TP < 2^{d-1}$, and it starts to deteriorate when TP approaches 2^d. This makes sense; if the union of cr balls is of size $2^d - O(n/d)$, then storing the complement exactly in $O(n)$ bits suffices. Also note that at the lower limit of ε, this lower bound matches the lower bound of the $\varepsilon = 0$ case in Theorem 1.

2. The conditions on TP are essentially a condition on $|B_d(cr)|$ and therefore a condition on c and r. The reason why we cannot put this in a closed form solely as a condition on c and r is because there is no closed form for $|B_d(cr)|$ for all values of c and r. We refer the reader to Lemma 11 in the Appendix for bounds on volumes of Hamming balls.

The rest of this section is devoted to the proof of Theorem 2.

We first prove a $n\log(1/\varepsilon)$ lower bound. The proof is an encoding argument that extends the scheme presented in the proof of Theorem 1. The encoder receives a set S of size n from the universe to encode. Assume the optimal distance sensitive data structure with ε average error uses s bits in the worst case. The encoder inserts S into the data structure, and runs the query algorithm on all points in the universe. We first claim that the number of points the data structure answers yes to is at most $2^{d+1}\varepsilon$. First, the number of true/legitimate positives TP is less than $2^d\varepsilon$. Also the number of false positives is always less than $2^d\varepsilon$. Adding these, we find that the total number of positives is at most $2^{d+1}\varepsilon$. The encoder then encodes the set S as a subset of these positives, using at most $\log\binom{2^{d+1}\varepsilon}{n}$ bits. He then sends these bits along with the at most s bits representing the set S to the decoder.

The decoder runs the query algorithm on the s bits, retrieves the set of positives, and uses the bits written by the encoder to figure out the set S. We have that:

$$
\begin{aligned}
s + \log\binom{2^{d+1}\varepsilon}{n} &\geq \log\binom{2^d}{n} \\
\Rightarrow s &\geq n\log(1/2e\varepsilon) \in \Omega(n\log(1/\varepsilon))
\end{aligned}
$$

To prove the nr^2/d lower bound, we first develop some notation. Consider the usual graph on the d-dimensional Hamming cube where two points p and q have an edge between them if they are related by a bit flip. Given a set $A \subset \{0,1\}^d$, let A^c denote its complement, and define ∂A to be the set of points in A that have an edge to a point in A^c (when either A^c or A is empty, ∂A is defined to be the empty set). Also, given an integer $r > 0$, define $A^{-r} = A \setminus \cup_{x\in\partial A} B(x, r-1)$. A^{-r} is the maximal subset of A that has the property that for any $x \in A^{-r}$, the ball $B(x,r)$ is contained inside A.

A data structure that uses s bits can be viewed as a function $\mathcal{D} : [\binom{2^d}{n}] \to 2^s$; given a set $S \subseteq \{0,1\}^d$ of size n, $\mathcal{D}(S)$ returns a memory representation using at most s bits. Let $V(S) = |\cup_{x\in S} B(x,r)| + \varepsilon(2^d - |\cup_{x\in S} B(x,r)|)$

Running the query algorithm on all points in the Hamming cube for the representation $\mathcal{D}(S)$ returns a set A of positives (A^c of negatives) such that $|A| \leq V(S)$. Varying over all $S \in [\binom{2^d}{n}]$, we get a family $\mathcal{F}$ of sets A such that:

1. $\forall S \in [\binom{2^d}{n}]$, $\exists A \in \mathcal{F}$ such that $B(x,r) \subset A$ for all $x \in S$.

2. $\forall S$ such that $\mathcal{D}(S) = A$, $|A| \leq V(S)$.

The query algorithm is therefore a function from 2^s to $\mathcal{F}$, the image of which is all of $\mathcal{F}$. This implies that $s \geq \log|\mathcal{F}|$. So in order to get a lower bound on s we need a lower bound on the size of the smallest family $\mathcal{F}$ with the above properties. Let D denote the composition of $\mathcal{D}$ with the query algorithm; i.e., $D:[\binom{2^d}{n}]\to\mathcal{F}$, where $D(S)$ equals the set A of positives that the query algorithm answers yes to on the representation $\mathcal{D}(S)$.

Fix $A\in\mathcal{F}$. Define $D^{-1}(A)=\{S:D(S)=A\}$. Any ball of radius r around a point $p\in S$ such that $S\in D^{-1}(A)$ must be completely contained inside A. The maximum number of such points p is $|A^{-r}|$. Thus we get that $|\cup_{S\in D^{-1}(A)}S|\le|A^{-r}|$. This implies that $|D^{-1}(A)|\le\binom{|A^{-r}|}{n}$.

Since the entire space $[\binom{2^d}{n}]$ needs to be covered, we get that $|\mathcal{F}|\ge\binom{2^d}{n}/\binom{|A^{-r}|}{n}$. Since $TP\le 2^{d-2}$, and the number of allowed false positives is less than $2^d\varepsilon$, which because $\varepsilon<1/4$ is also less than 2^{d-2}, we get that the total number of positives is less than 2^{d-1}. The next lemma derives an upper bound on $|A^{-r}|$

Lemma 1. *Fix a $1\le V\le 2^d$, and let $r>0$.*

(i) Of all sets A such that $|A|\le V$, the largest Hamming ball of size V maximizes $|A^{-r}|$.

(ii) If $2^d\varepsilon\le V\le 2^{d-1}$ and A is any set of size V, then $|A^{-r}|\le 2^d\exp(-2r^2/d)$.

Proof. We prove (i) by induction. The case $r=1$ is just the edge isoperimetric inequality, which states that of all sets of a given size, Hamming balls have the smallest vertex boundary. Assume the statement is true for $r=k$, and fails at $r=k+1$ for some set A, i.e., $|A^{-(k+1)}|>|H(V)^{-(k+1)}|$, were $H(V)$ is a Hamming ball of volume V. However, the isoperimetric inequality can also be stated as: if a set A (that is not a ball) has size greater then or equal to that of the Hamming ball of radius R, then $|A\cup\Gamma(A)|$ is larger than the volume of Hamming ball of radius $R+1$, where $\Gamma(A)$ is the set of neighbors of A. Thus we get that $|A^{-(k+1)}|\cup\Gamma(A^{-(k+1)})|>|H(V)^{-k}|$, implying that $|A^{-k}|>|H(V)^{-k}|$, which contradicts the induction hypothesis.

The proof of (ii) follows from (i) and the additive Chernoff bound for binomial random variables. If X_i denotes the outcome of the ith coin toss with an unbiased coin, and $X=\sum_{i=1}^d X_i$, then $P(X\le\mu-a)\le\exp(-2a^2/d)$, for all $0<a<\mu$, where $\mu=\mathbb{E}[X]=d/2$.

To prove (ii), let $r(V)$ be the radius of a Hamming ball of volume V. First note that $|A^{-r}|$ is at most the size of the Hamming ball of radius $r(V)-r$ by (i) above. Let $X\sim\text{Bin}(d,0.5)$. Now we have that

$$\begin{aligned}|A^{-r}|&\le 2^dP[X\le r(V)-r]\\&=2^dP[X\le d/2-(r+(d/2-r(V)))]\\&\le 2^d\exp\left(-2\frac{(r+(d/2-r(V)))^2}{d}\right)\le 2^d\exp\left(-2r^2/d\right)\end{aligned}$$

□

We now use the lemma to derive the lower bound. We get

$$|\mathcal{F}|\ge\binom{2^d}{n}/\binom{|A^{-r}|}{n}\ge\left(e2^d/|A^{-r}|\right)^n\ge\left(\exp\left(2r^2/d\right)+1\right)^n,$$

which implies that $s\ge\log\mathcal{F}=\Omega\left(nr^2/d\right)$ Combining our bounds, we get that when n,r and c satisfy the condition that $TP\le 2^{d-2}$, any data structure must use $\Omega(n(r^2/d+\log(1/\varepsilon)))$ bits in the worst case.

3.2 Point-wise error

The lower bound for the average case as stated in Theorem 2 also applies to a filter with point-wise error guarantees. A (r,c,ε)-filter with point-wise error is also a (r,c,ε)-filter with average error ε. However, a stronger lower bound holds if the number of points n is not too large.

Theorem 3. *Consider an (r,c,ε)-distance sensitive approximate membership filter with point-wise error guarantees, $\mathcal{F}$, on a set of n points in $\{0,1\}^d$. Define $TP(n,r,c,d) = n|B_d(cr)|$ as before. Then the filter must use:*

- *(Almost all n)* $\Omega\big(n\big(r^2/d + \log(1/\varepsilon)\big)\big)$ *bits if* $TP(n,r,c,d)/2^d < \varepsilon < 1/4$.
- *(Small n)* $\Omega(n(r/\delta c + \log(1/\varepsilon)))$ *bits if* $n\exp(-\delta cr D(\frac{1}{\delta}||0.5)) < \varepsilon < 1/4$*, where* $D(p||0.5)$ *is the Kullback-Leibler divergence function (see Appendix Lemma 11).*

Proof. We claim that the $H = \Omega\big(n\big(r^2/d + \log(1/\varepsilon)\big)\big)$ lower bound in Theorem 2 for the average error case applies to $\mathcal{F}$ as well. Suppose that $\mathcal{F}$ requires H' bits with $H' < H$. Let Q_{far} be the set of points at distance at least cr from all points in S. Since the point-wise filter says *no* to a point in Q_{far} with probability at least $1-\varepsilon$, the expected number of points in Q_{far} with a correct answer is $(1-\varepsilon)|Q_{\text{far}}|$. There must exist random values for which the filter provides the correct solution for at least $(1-\varepsilon)|Q_{\text{far}}|$ points: by using these values, we obtain a deterministic average error filter with space complexity $H' < H$, which is a contradiction. Therefore the lower bound H holds for point-wise filters as well.

We now observe that the lower bound H can be further improved. Indeed, the lower bound H is *decreasing* in d and a higher value is reach when $d = \Theta(rc)$. However, a filter for d-dimensional points with point-wise guarantees is also a filter for d'-dimensional points with the same guarantees for any $d > d'$. The claimed result then follows, as long as the conditions of Theorem 2 are met. This is where we require n to be small; let $d = \delta rc$, then the condition states that $n|B_{\delta cr}(cr)/|2^{\delta cr} < \varepsilon < 1/4$. BY using the Hoeffding-Chernoff bound for the volume of a Hamming ball (see Appendix Lemma 11), this simplifies to $n\exp(-\delta cr D(\delta^{-1}||0.5)) < \varepsilon < 1/4$.

We observe that this trick does not hold for the average case since the average error rate relatively to a subspace (e.g., $\{0,1\}^{d'}$) can be much larger than the one in the complete space (i.e., $\{0,1\}^d$). ☐

As we will see later this lower bound is asymptotically tight if $c \geq 2$. In the upper bound we show there is a $1/(c-1)$ overhead if $1 < c < 2$ and ε is sufficiently small. The next proof shows that this overhead is unavoidable when $1 < c < 2$. To help in assessing the hypothesis, we notice that, when $c = 1 + \sqrt{r}$, the theorem holds for $n \leq 2^{\Theta(r)}$, $\varepsilon \leq 2^{-\Theta(r)}$, $d = 2^{\Omega(\sqrt{r})}$ and it gives a $\Omega\big(r^{3/2}\big)$ bound, whereas the previous theorem only gave $\Omega(r)$. We note that the next theorem can be integrated with the previous Theorem 3 to get an additive r/c or r^2/d more (according to the parameters).

Theorem 4. *Let $c \leq 2$, $\varepsilon \leq (c-1)/n$ and $d(c-1) \geq ((c-1)/\varepsilon)^{6/(r(c-1))} + (r(c-1))^3$. Consider an (r,c,ε)-distance sensitive approximate membership filter with point-wise error guarantees, $\mathcal{F}$, on a set of n points in $\{0,1\}^d$. Then, $\mathcal{F}$ requires $\Omega\Big(\frac{n}{c-1}\log(1/\varepsilon)\Big)$ bits.*

Proof. The main idea of the proof is to use filter $\mathcal{F}$ in a one-way randomized protocol between two players (Alice and Bob) to send an arbitrary element x of a given set $\mathcal{S}$ from Alice to Bob: it is known (indexing problem) [13] that such a protocol requires $\Omega(\log|S|)$ bits if the protocol succeeds with probability at least $2/3$ and the two players share random bits. We introduce two families of error correcting codes, $\mathcal{C}$ and $\mathcal{M}$, that are explained below and we assume, without loss of generality, that are both known to Alice and Bob (the code families can be constructed with a simple deterministic brute-force algorithm).

Let $k = 1/(c-1)$. The error correcting binary code $\mathcal{C}$ has $m = 1/(n\varepsilon k)$ code words, each one with length $d_{\mathcal{C}} = d/k$ bits, weight $w = r/k$ and minimum Hamming distance between two code words $\delta = r/k$. Such an error correcting code exists for [9, Theorem 6], which shows that there exists a constant-weight code of size at least

$$\frac{d_{\mathcal{C}}^{w-\delta/2+1}}{\delta!} \geq \frac{(d(c-1))^{r(c-1)/2}}{(r(c-1))^{r(c-1)}} \geq (d(c-1))^{r(c-1)/6} \geq \frac{c-1}{\varepsilon}$$

where, in the second from the end and in the last steps, we exploit the fact that $d(c-1) \geq (r(c-1))^3$ and $d(c-1) \geq ((c-1)/\varepsilon)^{6/(r(c-1))}$, respectively.

The error correcting binary code $\mathcal{M}$ has n codewords and minimum Hamming distance rc (there is no requirement on codewords weights); we let $\mathcal{M} = \{m_1,...,m_n\}$. By the Gilbert-Varshamov bound such a code $\mathcal{M}$ exists with length $d_{\mathcal{M}} = rc + \log n$.

Alice arbitrary selects n codes $x_i = (x_{i,1},...,x_{i,k-1})$ from the set $\mathcal{C}^k$. Then, she encodes each x_i into $\hat{x}_i = x_{i,1} \cdot ... \cdot x_{i,k} \cdot z_0 \cdot m_i$, where $\cdot$ denotes the concatenation of binary sequences, z_0 is a sequence of $r/k = r(c-1)$ zeros, and $m_i \in \mathcal{M}$. The length of each $\hat{x}_i$ is $d_x = kd_{\mathcal{C}} + d_{\mathcal{M}} + r/k = d + \log n + r(2c-1)$. Finally, Alice inserts $\hat{x}_0,...,\hat{x}_{n-1}$ into the filter $\mathcal{F}$ and sends $\mathcal{F}$ to Bob using $S_{\mathcal{F}}(n,d_X,c,r)$ bits.

We now show that Bob can reconstruct each codeword x_i by querying filter $\mathcal{F}$ at most $1/\varepsilon$ times. Codeword $x_{i,1}$ is obtained by performing a query with $q = q' \cdot z_2 \cdot z_3 \cdot m_i$ for every possible codeword $q' \in \mathcal{C}$, where z_2 is a sequence of $(k-1)\delta = (k-1)r(c-1)$ zeros, z_3 is a sequence of r/k ones, and $m_i \in \mathcal{M}$. The distance between q' and any $\hat{x}_j$ in $\mathcal{F}$ is $D(\hat{x}_j,q) = D(x_{j,1},q') + D(x_{j,2} \cdot ... \cdot x_{i,k},z_2) + D(z_0,z_3) + D(m_j,m_i)$. It holds that:

1. $D(x_{i,1},q') \geq r(c-1)$ if $q \neq x_{i,1}$ and 0 otherwise;
2. $D(x_{j,2} \cdot ... x_{i,k},z_1) = (k-1)r(c-1) = r - r(c-1)$ since each codeword in $\mathcal{C}$ has weight $r(c-1)$;
3. $D(z_0,z_3) = r(c-1)$;
4. $D(m_j,m_i) \geq rc$ if $m_j \neq m_i$ and 0 otherwise.

Therefore, $D(\hat{x}_j,q) = r$ if $x_{i,1} = q'$ and $m_i = m_j$, and $D(\hat{x}_j,q) \geq rc$ otherwise. A similar procedure holds for computing $x_{i,j}$ for each i and j.

Bob performs mk queries per x_i and $nkm = 1/\varepsilon$ queries in total. The expected number of wrong queries is then 1 and, if the protocol is repeated independently, there is a constant probability that all queries succeed. Since Bob is able to reconstruct an entry from the set $\mathcal{S} = \mathcal{C}^{nk}$, by the aforementioned result in [13], we have

$$S_{\mathcal{F}}(n,d_x,c,r,\varepsilon) \geq \Omega(\log \mathcal{S}) \Rightarrow S_{\mathcal{F}}(n,d_x,c,r,\varepsilon) \geq \Omega\left(\log |\mathcal{C}|^{nk}\right) \geq \frac{n}{c-1} \log(1/\varepsilon).$$

□

4 Upper bounds

4.1 The vector signature method

In this section we present a filter data structure with point-wise error and prove its correctness. We will later show how a filter with average error follows as a special case.

Our filter data structure can be seen as a succinct version of a CountSketch [6]. It is easy to see that the CountSketch of a vector, x, will never have ℓ_1 norm larger than $\|x\|_1$. By linearity we can use the difference between the CountSketches of two vectors to derive upper bounds on the ℓ_1 distance between them. The number of entries (and hence the space usage) in each CountSketch can be used to control the approximation error.

It turns out that we can throw away much of the information in a standard CountSketch without sacrificing asymptotic error bounds. Specifically we introduce the notion of a vector signature as a function mapping a vector from $\{0,1\}^d$ into $O(r/(c-1) + (c/(c-1))\log(1/\varepsilon))$ bits.

The key feature of the vector signature is that a suitable function of the signatures of two vectors x and y is smaller than or equal to a certain threshold Ψ if $D(x,y) \leq r$, while it is larger than Ψ with probability $1-\varepsilon$ if $D(x,y) \geq cr$ as formalized in theorem 5.

Signature construction The construction of the signature uses four parameters m,c_{mod},c_{div} and δ that all depend on r, c and ε. Their values will be provided in the two following proofs.

Let M be an $m\times d$ random matrix. For any $i\in\{1,\ldots,m\}$,$j\in\{1,\ldots,d\}$, $M_{i,j}$ denotes the element in the ith row and jth column of M, and m_i denotes the ith row. For $m=O\big(r/(c-1)+(c/(c-1))^2\log(1/\varepsilon)\big)$. Every entry of M is initially set to 0. For each column, j', $\delta=O\left(\lceil(c/r)\log(1/\varepsilon)\rceil\right)$ updates are performed as follows:

(1) Select a value, s, i.d.d. from $\{-1,1\}$. (2) Select a row, i', uniformly at random from $\{1,\ldots,m\}$.(3) Update the entry at $M_{i',j'}$ by adding s.

We let u_i denote the number of updates performed on all entries of row m_i; it holds that $\|m_i\|_1\le u_i$ since two updates can affect the same entry and erase each other.

Let c_{div},c_{mod} be suitable values with asymptotic value $O(c)$. The *signature* of a vector $x\in\{0,1\}^d$ is then the m-dimensional vector $\sigma(x)$ defined by

$$\sigma(x)_i=\left\lfloor\frac{(Mx)_i \bmod_* c_{\mathrm{mod}}}{c_{\mathrm{div}}}\right\rfloor$$

For notational simplicity, we define the $\bmod_*$ operator that is similar to the standard modulo operator but maps into a range symmetric around zero, i.e,

$$\alpha \bmod_* c_{\mathrm{mod}}=((\alpha+\lceil c_{\mathrm{mod}}/2\rceil) \quad \mathrm{mod} c_{\mathrm{mod}})-\lceil c_{\mathrm{mod}}/2\rceil$$

where $\bmod$ denotes the standard modulo operation into $[0,c_{\mathrm{mod}})$. We refer to $\Gamma(x,y)$, defined by $\Gamma(x,y)_i=c_{\mathrm{div}}(\sigma(x)_i-\sigma(y)_i) \bmod_* c_{\mathrm{mod}}$, as the *gap vector* between signatures of vectors x and y, and to $\gamma(x,y)=\|\Gamma(x,y)\|_1$ as their *gap*.

The following theorem describes the main property of signature vectors.

Theorem 5. *There exists a value $\Psi=O(\delta r)$, such that for each pair of vectors $x,y\in\{0,1\}^d$:*

- *if $D(x,y)\le r$ then $\gamma(x,y)\le\Psi$, and*
- *if $D(x,y)>cr$ then $\gamma(x,y)\ge\Psi$ with probability at least $1-\varepsilon$.*

We give two different proofs of Theorem 5 depending on the value of the approximation factor c. First for $c=O(1)$ we obtain the claimed result when $m=O\big(r/(c-1)+(1/(c-1))^2\log(1/\varepsilon)\big)$. Then we prove the theorem for larger approximation factors with $m=O(r/c+\log(1/\varepsilon))$.

For two given vectors x and y, we will assume for notational convenience that they differ on the first $D(x,y)$ positions. We let x' and y' denote the prefix of length $D(x,y)$ of x and y (i.e., the positions where they differ), M' denote the first $D(x,y)$ columns of M, m_i' the ith row of M', and u_i' the number of updates affecting m_i'.

***Proof of Theorem 5 with* $\mathbf{c=O(1)}$.** For the case $c=O(1)$, we set $m=24\frac{c^2}{c-1}\max\left\{r;\frac{2}{c-1}\log(1/\varepsilon)\right\}$, $c_{\mathrm{div}}=1$, $c_{\mathrm{mod}}=2$ and, $\delta=1$. With these values, the signature becomes $\sigma(x)_i=(Mx)_i \bmod_* 2$, where each column of M is a random vector with exactly one entry in $\{-1,1\}$ and the remaining $m-1$ entries set to zero. Further, the threshold is $\Psi=r$ and the gap vector is defined by

$$\Gamma(x,y)_i=M(x-y)_i \bmod_* 2=M'(x'-y')_i \bmod_* 2.$$

The second equality is true because $c_{\mathrm{div}}=1$ so there is no rounding and σ is a linear function of x and y. We observe that the bit positions where x and y are equal do not affect the gap vector.

When $D(x,y)\le r$, M' contains at most r entries in $\{-1,1\}$ and hence $\gamma(x,y)=\|M'(x'-y')\|_1\le r$, proving the first part of Theorem 5.

Consider now the case $D(x,y)\ge cr$. The second part of Theorem 5 follows by the following two claims:

Claim 1: With probability $1-\varepsilon$, there are more than r rows of M' affected by an odd number of updates; we refer to these rows as *odd rows.*

Claim 2: If m'_i is an odd row, then $|\Gamma(x,y)_i|=1$.

Lemma 2 and 3 below show that the claims hold. We then have that $\gamma(x,y)=\sum_{i=1}^{m}|\Gamma_i(x,y)|>r=\Psi$ and Theorem 5 follows. □

Lemma 2 (Claim 1)**.** *Let x,y be two input vectors in $\{0,1\}^d$, and let M' be the sub-matrix of M associated to the positions where x and y differ. Then, if x and y have distance at least cr we have that with probability $1-\varepsilon$, there are more than r odd rows in M'.*

Proof. Consider the $D(x,y)$ updates used in the construction of M'. If after the first $D(x,y)-cr$ updates there are at least $(c+1)r$ rows with an odd number of updates, then the theorem follows: The remaining cr updates can decrease the number of odd rows by at most cr.

Suppose now that there are $Y_o<(c+1)r$ odd rows after the first $D(x,y)-cr$ updates, and consider the last cr updates. Let Y_j, with $j\in\{1,...cr\}$ be a random variable set to 1 if the jth update affects an odd row, which then becomes an even row; Y_i is set to 0 otherwise. The probability that $Y_j=1$ is $p=(Y_o+j-1)/m\leq 3cr/m$ since there can be at most Y_o+j-1 odd rows before the jth update: the initial Y_o odd rows and the rows affected by the previous $j-1$ updates. Let $Y=\sum_{j=1}^{cr}Y_j$. The expected value of Y is $\mu=pcr\leq 3(cr)^2/m$. Let $\eta=(c-1)r/(2\mu)-1$ (note that $\eta\geq 0$). By a Chernoff bound, the probability that $Y\geq(c-1)r$ is $\Pr[Y\geq(c-1)r/2]=\Pr[Y\geq\mu(1+\eta)]\leq e^{-\eta^2\mu/2}$. So we have:

$$\Pr[Y\geq(c-1)r/2]\leq e^{-\left(\left(\frac{c-1}{c}\right)^2\frac{m}{6}+\frac{3(cr)^2}{2m}-(c-1)r\right)}\leq e^{-\left(\left(\frac{c-1}{c}\right)^2\frac{m}{6}-(c-1)r\right)}\leq\varepsilon$$

where the last step follows since $m=24\frac{c^2}{c-1}\max\left\{r;\frac{2}{c-1}\log(1/\varepsilon)\right\}$. Therefore, with probability $1-\varepsilon$, at most $Y\leq(c-1)r/2$ updates affect odd rows, making them even. It follows that the number of odd rows after all updates is: $Y_0+(cr-Y)-Y\geq cr-2Y\geq r$. □

Lemma 3 (Claim 2)**.** *If row m'_i is odd, then $|\Gamma_i(x,y)|=1$.*

Proof. Let $h_1,...,h_{u_i}$ be the non zero entries in m'_i. We have that $m_i(x'-y')=\sum_{j=1}^{u_i}M'_{i,h_j}(x'_{h_j}-y'_{h_j})$. Since $(x'_{h_j}-y'_{h_j})$ and $M'_{i,j}$ are in $\{-1,1\}$ and since u_i is odd, then the sum must be odd and $|\Gamma_i(x,y)|=|m'_i(x'-y') \bmod_* 2|=1$. □

Proof of Theorem 5 for $\mathbf{c=\Omega(1)}$**.** Let $\beta=4/(p_1p_2)^2$ where p_1 and p_2 are suitable constants ($p_1\approx 0.6$, $p_2\approx 0.06$) that will be derived in the proof. The proof presented here holds for $c\geq\sqrt{\beta}/p_2\approx 925$. We conjecture that a smaller approximation factor c can be obtained with a more careful analysis of the constants. The parameters used in the signature construction are set as follows: $m=\beta\max\{r/c,\log(2/\varepsilon)\}$, $\delta=\lceil c\log(2/\varepsilon)/r\rceil$, $c_{\text{div}}=c/\beta$, $c_{\text{mod}}=8c$, and the threshold is $\Psi=\delta r+\max\{r,c\log(1/\varepsilon)\}$.

In contrast to the $c=O(1)$ case, the gap vector and the gap cannot be expressed as a function of only the positions where x and y differ (i.e., x' and y'). In fact, due to the division by c_{div} and the floor operation, the gap vector may depend on the positions where x and y coincide. However, we can still provide upper and lower bounds on the gap that depend only on x' and y'. It holds that:

$$|m'_i(x'-y') \bmod_* c_{\text{div}}|-c_{\text{div}}\leq|\Gamma_i(x,y)|\leq|m'_i(x'-y') \bmod_* c_{\text{mod}}|+c_{\text{div}}. \tag{1}$$

Suppose $D(x,y) \le r$, then by (1) the gap can be upper bounded as follows:

$$\gamma(x,y) = \sum_{i=1}^{m} |\Gamma_i(x,y)| \le \sum_{i=1}^{m} \big(|m_i'(x'-y') \bmod_* c_{\text{mod}}| + c_{\text{div}} \big)$$
$$\le c_{\text{div}} m + \sum_{i=1}^{m} |m_i'|_1 \le \delta r + \max\{r, c\log(1/\varepsilon)\} = \Psi.$$

The last two steps follow since entries in $x'-y'$ are in $\{-1,1\}$ and M' contains at most δr updates. We observe that it is crucial to use $\bmod_*$ instead of $\bmod$ since it guarantees that $|\alpha \bmod_* c_{\text{mod}}| \le |\alpha|$. The first part of the theorem follows.

Suppose now that $D(x,y) \ge cr$. We say that row m_i' is dense if the number of updates u_i is at least $\delta D(x,y)/(2m)$. The proof that the gap is larger than Ψ with probability at least $1-\varepsilon$ relies on the following claims:

Claim 1: With probability $1-\varepsilon/2$, the number of dense rows is at least $p_1 m$, where p_1 is a suitable constant;

Claim 2: For each dense row, the probability that $|\Gamma_i(x,y)| \ge c/\sqrt{\beta}$ is a constant p_2;

Claim 3: With probability $1-\varepsilon/2$, there are at least $p_1 p_2 m$ entries in the gap vector such that $|\Gamma_i(x,y)| \ge c/\sqrt{\beta}$;

Then, we have that $\gamma(x,y) = \sum_{i=1}^{m} |\Gamma_i(x,y)| \ge c p_1 p_2 m/(2\sqrt{\beta}) \ge \Psi$ since $m = \beta \max\{r/c, \log(1/\varepsilon)\}$ and $\beta = 4/(p_1 p_2)^2$. Thus, the second part of Theorem 5 follows. □

Before proving the claims , we will need three technical lemmas. Lemma 4 gives a load bound on a balls and bins problem by using the bounded differences method to manage dependent random variables. Lemma 5 bounds the probability of a sum of $\{-1,1\}$ random variables to be in a specified interval after a modular operation. Finally, lemma 6 gives a lower bound on the tail distribution of the sum of $\{-1,1\}$ random variables by leveraging the Berry-Esseen theorem.

Lemma 4. *Consider p balls thrown uniformly and independently at random in q bins, with $p \ge q$, and let $\alpha > 0$ be any arbitrary value. Then, with probability at least $1-\varepsilon$, there are more than $q(1-e^{-\alpha}) - \sqrt{(q/2)\log(1/\varepsilon)}$ bins with at least $p/q - \sqrt{2\alpha p/q}$ balls.*

Proof. For every $i \in \{1,...,p\}$ and $j \in \{1,...,q\}$, define the following random variable:

$$X_{i,j} = \begin{cases} 1 \text{ if ball } i \text{ landed in bin } j \\ 0 \text{ otherwise} \end{cases}$$

Let also $X_j = \sum_{i\in[p]} X_{i,j}$ be the number of balls in the jth bin; the expected value of X_j is $\mu = p/q$ for each j. Since the balls are throw independently a Chernoff bound gives:

$$\Pr\Big[X_j \le \mu - \sqrt{2\alpha\mu}\Big] \le e^{-\alpha}$$

Consider now the random variable Y_j:

$$Y_j = \begin{cases} 1 \text{ if } X_j \ge \mu - \sqrt{2\alpha\mu} \\ 0 \text{ otherwise} \end{cases}$$

Let $Y = \sum_{j=1}^{q} Y_j$; we use $Y_{Y_1,..,Y_q}$ to denote the actual value of Y with the specified values. Since there is dependency among the Y_j terms, we use the method of bounded differences [8] to bound the tail distribution, instead of a Chernoff bound. The random variable Y satisfies the Lipschitz property with constant 1, that is:

$$|Y_{Y_1,\ldots,Y_i,\ldots,Y_q} - Y_{Y_1,\ldots,Y_i',\ldots,Y_q}| = |Y_i - Y_i'| \leq 1$$

whenever $Y_i \neq Y_i'$ for every $i \in \{1,\ldots,q\}$. By the method of bounded differences [8, Corollary 5.2], we get $\Pr[Y \leq \mathrm{E}[Y] - t] \leq e^{-2t^2/q}$. Then, $\Pr[Y > \mathrm{E}[Y] - t] \geq 1 - \varepsilon$ if $t = \sqrt{(q/2)\log(1/\varepsilon)}$. Being $\mathrm{E}[Y] \geq q(1 - \Pr[X_j \leq \mu - \sqrt{2\alpha\mu}]) \geq q(1 - e^{-\alpha})$, the claim follows. □

Lemma 5. *Consider a sequence $s_1,\ldots,s_k$ of independent and evenly distributed random variables in $\{1,-1\}$, and an arbitrary value $q \in \mathbb{N}$. Let $S = \sum_{i=1}^{k} s_i$ and $S_q = S \bmod_* q$. Then for all values a,b such that $0 \leq a \leq b < \lceil q/2 \rceil$ and $b - a \geq q/4$, we have:*

$$\frac{\Pr[|S| \geq a]}{2} < \Pr[a \leq |S_q| < b] < \Pr[|S| \geq a]. \tag{2}$$

Proof. Let $k' = \lceil 2k/q \rceil - 1$, and define the following four quantities:

$$H_1 = \sum_{\ell=0}^{\lfloor k'/2 \rfloor} \Pr[a + \ell q \leq |S| < b + \ell q]$$

$$H_2 = \sum_{\ell=0}^{\lfloor k'/2 \rfloor} \Pr[\ell q + b \leq S < (\ell + 1/2)q] + \Pr[(\ell+1)q \leq S < (\ell+1)q + a]$$

$$H_3 = \sum_{\ell=0}^{\lfloor k'/2 \rfloor} \Pr[(\ell+1)q - b < |S| \leq (\ell+1)q - a]$$

$$H_4 = \sum_{\ell=0}^{\lfloor k'/2 \rfloor} \Pr[(\ell+1/2)q \leq S < (\ell+1)q - b] + \Pr[(\ell+1)q - a \leq S < (\ell+1)q].$$

Standard computations show that: $\Pr[a \leq |S_q| < b] = H_1 + H_3$ and that $\Pr[|S| \geq a] = H_1 + H_2 + H_3 + H_4$. We then have that $\Pr[a \leq |S_q| < b] < \Pr[|S| \geq a]$, and the right side of the inequality in (2) follows.

We now focus on the other side of the inequality. We first prove that $H_1 \geq H_2$. The random variable S has value i, with $i \in [-k,k]$ if there are $(k+i)/2$ terms set to $+1$ and $(k-i)/2$ terms set to -1. Since the s_i terms are independent and evenly distributed, we have $\Pr[S = i] = \binom{k}{(k+i)/2}\frac{1}{2^k}$ (note that the probability is decreasing in i). We then have, for any integer $\ell \geq 0$, that :

$$\Pr[a + \ell q \leq |S| < b + \ell q] = 2 \sum_{j=a+\ell q}^{b+\ell q - 1} \binom{k}{(k+j)/2} \frac{1}{2^k}.$$

Since $(b-a) \geq q/4$ and $\Pr[S = i]$ is decreasing in i, it then follows that

$$\begin{aligned}\Pr[a + \ell q \leq |S| < b + \ell q] &\geq 2 \sum_{j=\ell q + b}^{(\ell+1/2)q+b} \binom{k}{(k+j)/2} \frac{1}{2^k} + 2 \sum_{j=(\ell+1)q}^{(\ell+1)q+a} \binom{k}{(k+j)/2} \frac{1}{2^k} \\ &\geq \Pr[\ell q + b \leq S < (\ell+1/2)q] + \Pr[(\ell+1)q \leq S < (\ell+1)q + a].\end{aligned}$$

It then follows that $H_1 \geq H_2$. Similarly, it can be shown that $H_3 \geq H_4$. Therefore, $\Pr[|S| \geq a] = H_1 + H_2 + H_3 + H_4 \leq 2(H_1 + H_3) \leq 2\Pr[a \leq |S_q| < b]$. The left side of the inequality in (2) follows. □

Lemma 6. *Let $S = \sum_{i=1}^{k} s_i$, where the s_i terms are independent and evenly distributed random variables in $\{-1,+1\}$ and let $\alpha > 0$ be any arbitrary value. Then,*

$$\Pr[|S| \geq \alpha\sqrt{k}] \geq \frac{2\alpha}{\sqrt{2\pi}(\alpha^2+1)e^{\alpha^2/2}} - \frac{1}{\sqrt{k}}.$$

Proof. We observe that $E[s_i]=0$, $\sigma^2=E[s_i^2]=1$ and $\rho=E[|s_i|^3]=1$. By the Berry-Esseen theorem [3], we have that the random variable $Q=S/(\sqrt{k}\sigma)=S/\sqrt{k}$ can be approximate by a standard normal distribution $\mathcal{N}(0,1)$ with error

$$|\Pr[Q \leq x] - \Psi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{k}}.$$

where $\Psi(x)$ is the cumulative distribution function of $\mathcal{N}(0,1)$, C is a suitable constant smaller than 1/2 [17]. The above inequality can be rewritten as

$$|\Pr[Q > x] - \Psi^c(x)| \leq \frac{1}{\sqrt{k}}.$$

with $\Psi^c(t) = 1-\Psi(x)$. Since $\Psi^c(x) \geq x/(\sqrt{2\pi}(x^2+1)e^{x^2/2})$ [7, 1], we then get

$$\Pr[|S| \geq \alpha\sqrt{k}] = 2\Pr[Q \geq \alpha]] \geq 2\Psi^c(\alpha) - \frac{1}{\sqrt{k}},$$

The lemma follows by inserting the bound for $\Psi^c(x)$. □

We are now ready to prove the three claims used in the proof of Theorem 5 for $c=\Omega(1)$.

Lemma 7 (Claim 1)**.** *With probability $1-\varepsilon/2$, the number of dense rows in M' is at least $p_1 m$, where $p_1 \approx 0.6$.*

Proof. Matrix M' is obtained by performing δ random updates per column independently and uniformly distributed. The number of updates u_i affecting row m'_i is distributed as the number of balls in a bin after randomly throwing $\delta D(x,y)$ balls into m bins. By applying Lemma 4 with $\alpha=1$, it follows that, with probability at least $1-\varepsilon$, there are more than $m' \geq (1-1/e-\sqrt{\log(2/\varepsilon)/m})m \geq p_1 m$ dense rows where $u_i \geq (\delta D(x,y)/m)(1-\sqrt{m/(\delta D(x,y))}) \geq \delta D(x,y)/(2m)$ as soon as $c \geq 2\sqrt{\beta}$. □

Lemma 8 (Claim 2)**.** *If m'_i is dense, then $|\Gamma_i(x,y)| \geq c\sqrt{\beta}$ with constant probability $p_2 \approx 0.06$.*

Proof. The statement follows by showing that $|m'_i(x'-y') \bmod_* c_{\text{mod}}| \geq c/\sqrt{\beta}(1+1/\sqrt{\beta}) \triangleq K$: indeed, by Equation 1, we have that $|\Gamma_i(x,y)| \geq |m'_i(x'-y') \bmod_* c_{\text{mod}}| - c_{\text{div}} \geq c/\beta + c/\sqrt{\beta} - c_{\text{div}} = c/\sqrt{\beta}$.

We observe that $m'_i(x'-y')$ can be rewritten as $\sum_{j=1}^{u_i} \sigma_j (x'-y')_{f(j)}$, where $f(j)$ is the position in m'_i affected by the jth update. Since $(x'-y')$ has entries in $\{-1,1\}$ and the σ_j's are independent, $m'_i(x'-y')$ has the same density function of $S=\sum_{j=1}^{u_i}\sigma_j$. The probability that $|\Gamma_i(x,y)| \geq K$, is then

$$\Pr[|M'_i(x'-y') \bmod_* c_{\text{mod}}| \geq K] = \Pr[|S \bmod_* c_{\text{mod}}| \geq K] \geq \frac{\Pr[|S| \geq K]}{2},$$

where in the last step we applied Lemma 5 with $a=K$, $b=c_{\text{mod}}/2$ and $q=c_{\text{mod}}$ since $b-a \geq c_{\text{mod}}/4$. To lower bound $\Pr[|S| \geq K]$ we apply Lemma 6 with $\alpha = 1+1/\sqrt{\beta}$ since $K \leq \sqrt{u_i}(1+1/\sqrt{\beta})$. Hence,

$$\frac{\Pr[|S| \geq K]}{2} \geq \frac{\Pr[|S| \geq (1+1/\sqrt{\beta})\sqrt{u_i}]}{2} \geq \frac{1+1/\sqrt{\beta}}{\sqrt{2\pi}((1+1/\sqrt{\beta})^2+1)e^{(1+1/\sqrt{\beta})^2/2}} - \frac{1}{\sqrt{u_i}} \approx 0.06 \triangleq p_2$$

where the last step follows since $\sqrt{u_i} \geq c/\sqrt{\beta} \geq 1/p_2$. □

Lemma 9 (Claim 3)**.** *With probability* $1-\varepsilon/2$*, there are at least* $p_1p_2m/2$ *entries in the gap vector such that* $|\Gamma_i(x,y)| \geq c$*;*

Proof. By the first claim there are at least $m' = p_1 m$ dense rows with probability $1-\varepsilon/2$. For each dense row, let Y_i be a random variable sets to 1 if $|\Gamma_i(x,y)| \geq c$, and 0 otherwise. By the previous lemma, we have that $\Pr[Y_i = 1] \geq p_2$. Let $Y = \sum_{i=1}^{m'} Y_i$. Since the Y_i are independent, we get with a Chernoff bound::

$$\Pr[Y \geq p_2 m'(1-\sqrt{\log(2/\varepsilon)/(p_2 m')} \geq p_2 m'/2] \geq 1-\varepsilon/2$$

using $m = \beta \max\{r/c, \log(2/\varepsilon)\}$ and $\beta \geq 2/p_2$. ☐

4.2 A filter with point-wise error

A filter with point-wise error follows by just storing the n signatures of the points S, using space $O\big(n\big(r/(c-1)+(c/(c-1))^2\log(n/\varepsilon)\big)\big)$. We must also store enough information to recover M, but this cost is dominated by the cost of storing the signatures. However, a better encoding allows to us remove the $\log n$ dependency, as shown in the following theorem.

Theorem 6. *There exists a* (r,c,ε)*-distance sensitive approximate membership filter with point-wise error which requires* $O\big(n\big(r/(c-1)+(c/(c-1))^2\log(1/\varepsilon)\big)\big)$ *bits for any* $c>1$ *on a set* S *of* n *points. When* $c \geq 2$ *and* $|\cup_{x\in S} B(x,cr)| \leq 2^{d-2}$*, the filter requires optimal* $O\big(n\big(r/c+\log(1/\varepsilon)\big)\big)$ *bits.*

Proof. Consider the n signatures of points in S constructed with error $\varepsilon' = \varepsilon/n$. Each requires $O\big(r/(c-1)+(c/(c-1))^2\log(1/\varepsilon')\big) = O\big(r/(c-1)+(c/(c-1))^2\log(n/\varepsilon)\big)$ bits. Group all signatures by the value of the $\log n$ most significant bits, and store all signatures in the same group without the common $\log n$ bits. This gives the claimed result. The probability of a false positive is $n\varepsilon' = \varepsilon$ due to an union bound on the n signatures. Finally, the optimality of the filter follows from Theorem 3. ☐

4.3 A filter with average error

The point-wise error filters are of course also valid average error filters, but in this setting we can also construct space efficient filters with a $c=1$ approximation factor. Define $Q_{r\text{-far}} = \{x \in \{0,1\}^d | D(x,S) \geq r\}$ and similarly $Q_{(r;cr)\text{-far}} = \{x \in \{0,1\}^d | r \leq D(x,S) \leq cr\}$.

Theorem 7. *Let* $|\cup_{x\in S} B(x,cr)| \leq 2^{d-1}$ *and* $|\cup_{x\in S} B(x,cr) - B(x,r)| \leq \varepsilon 2^{d-2}$*. Then an* $(r,c,\varepsilon/4)$*-point-wise error filter,* $\mathcal{F}$*, is also an* $(r,1,\varepsilon)$*-average error filter.*

Proof. Let P denote the number of false positives accepted by $\mathcal{F}$. Since there is no guarantee on points in $Q_{(r;cr)\text{-far}}$ we get $P \leq \frac{\varepsilon}{4}|Q_{cr\text{-far}}| + |Q_{(r;cr)\text{-far}}|$. We can trivially bound $|Q_{cr\text{-far}}| \leq 2^d$. From the second constraint we get $|Q_{(r;cr)\text{-far}}| \leq \varepsilon 2^{d-2}$. We see that $P \leq \frac{\varepsilon}{4}2^d + \varepsilon 2^{d-2} = \varepsilon 2^{d-1}$ By the first constrain we also have $|Q_{r\text{-far}}| \geq |Q_{cr\text{-far}}| \geq 2^{d-1}$ so we see $P \leq \varepsilon 2^{d-1} \leq \varepsilon |Q_{r\text{-far}}|$. Hence $\mathcal{F}$ fulfills all the requirements for an $(r,1,\varepsilon)$-average error filter. ☐

The two constrains used in Theorem 7 allow many different settings of r,c and ε depending on S and d. Lets consider a specific, practical setting of the parameters. By setting $c=r$ in the point-wise filter, we obtain an average error filter which matches the $\Omega(n\log(1/\varepsilon))$ lower bound of Theorem 2 for small r. Interestingly, this space bound shows that it is possible to support distance sensitive membership queries in the average error setting with the asymptotic space bound of a Bloom filter when $r \leq \sqrt{d}$.

Theorem 8. *Let* $r \leq \sqrt{d}$*,* $n \leq 2^{d/3}$ *and* $\varepsilon \geq 1/2^{d-2}$*. Then, there exists an optimal* $(r,1,\varepsilon)$*-distance sensitive approximate membership filter with average error which requires* $O(n\log(1/\varepsilon))$ *bits on a set* S *of* n *points.*

Proof. Let us consider a $(r,r,\varepsilon/4)$-filter $\mathcal{F}$ with point-wise guarantees. The amount of false positives accepted by $\mathcal{F}$ is $P \leq (\varepsilon/4)|Q_{r^2\text{-far}}| + |Q_{(r;r^2)\text{-far}}|$. We have $|Q_{(r;r^2)\text{-far}}| \leq nr^2\binom{d}{r^2} \leq (\varepsilon/4)2^d$ since $d \geq r^2$, $n \leq 2^{d/3}$ and $\varepsilon \geq 4/2^{d/2}$. Trivially, we also have that $|Q_{r^2\text{-far}}| \leq 2^d$. We see that $P \leq \varepsilon 2^{d-1}$

Now note that $|Q_{r\text{-far}}| \geq 2^d - nr\binom{d}{r} \geq 2^{d-1}$ by $d \geq r^2$ and $n \leq 2^{d/3}$.

We combine the two bounds to see $P \leq \varepsilon 2^{d-1} \leq \varepsilon |Q_{r\text{-far}}|$. The optimality of $\mathcal{F}$ follows from Theorem 2 since $r^2/d < 1$ and $n\log(1/\varepsilon)$ is a lower bound. □

5 Conclusion

To the best of our knowledge, this paper is the first that presents and gives upper and lower space bounds for the problem of distance-sensitive filters without false negatives. However, several open questions still remain: It would be interesting to improve the analysis of signatures to get a unique analysis of the two cases (i.e., $c \in O(1)$ and $c \in \Omega(1)$) with better constants. There are also settings where the bounds are not tight. It would be interesting to close the gap in the point-wise setting between the lower and upper bound when n is larger than the upper limit in Theorem 3, and likewise in the average error setting when $r > \sqrt{d}$. The signature method provides a valuable tool for distance sensitive membership queries without false negatives, it would be interesting to extend it further. We have made some preliminary experimental analysis that shows that the signatures work as intended and are easy to implement, but it would be interesting to perform more through experiments.

References

[1] Milton Abramowitz. *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables,*. Dover Publications, Incorporated, 1974.

[2] Alexandr Andoni and Ilya Razenshteyn. Optimal data-dependent hashing for approximate near neighbors. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 793–801. ACM, 2015.

[3] Andrew C. Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941.

[4] Burton H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, jul 1970.

[5] Andrei Broder and Michael Mitzenmacher. Network applications of Bloom filters: A survey. *Internet mathematics*, 1(4):485–509, 2004.

[6] Moses Charikar, Kevin C. Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3–15, 2004.

[7] John Cook. Upper and lower bounds for the normal distribution. Unpublished manuscript, http://www.johndcook.com/normalbounds.pdf, 2009.

[8] Devdatt Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.

[9] R. L. Graham and N. J. A. Sloane. Lowr bounds for constant weight codes. *IEEE Transaction on Information Theory*, 1980.

[10] Bin Hua, Yu abd Xiao, Bharadwaj Veeravalli, and Dan Feng. Locality-sensitive Bloom filter for approximate membership query. *IEEE Transactions on Computers*, 61(6):817–830, 2012.

[11] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. *Proceedings of the thirtieth annual ACM ...*, 8:321–350, 1998.

[12] Adam Kirsch and Michael Mitzenmacher. Distance-sensitive Bloom filters. *Proceedings of the 8th Workshop on Algorithm Engineering and Experiments (ALENEX)*, 2006:41–50, 2006.

[13] Eyal Kushilevitz and Noam Nisan. *Communication complexity*. Cambridge University Press, 1997.

[14] Michael Mitzenmacher, Rasmus Pagh, and Ninh Pham. Efficient estimation for high similarities using odd sketches. In *Proceedings of the 23rd international conference on World wide web*, pages 109–118. ACM, 2014.

[15] Rasmus Pagh. Locality-sensitive hashing without false negatives. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1–9. SIAM, 2016.

[16] Rasmus Pagh and Flemming Friche Rodler. Lossy dictionaries. In *European Symposium on Algorithms*, pages 300–311. Springer, 2001.

[17] I.S. Tyurin. An improvement of upper estimates of the constants in the lyapunov theorem. *Russian Mathematical Surveys*, 65:201–202, 2010.

A Binomial and Hamming ball volume bounds

Lemma 10 (Binomial bounds). *Let $1 \le k \le n = \omega(1)$. Then:*

1. *If $k = O(1)$, $\binom{n}{k} = (1+o(1))n^k/k!$.*
2. *If $k = \omega(1)$ and $k = o(\sqrt{n})$,*
$$\binom{n}{k} = (1+o(1))\frac{1}{\sqrt{2\pi k}}\left(\frac{ne}{k}\right)^k.$$
3. *If $k = o(n)$, $\log\binom{n}{k} = (1+o(1))k\log(n/k)$.*
4. *If $k = \Omega(n)$, $\log\binom{n}{k} = (1+o(1))H(k/n)n$, where $H(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function defined on* $[0,1]$.

Next, recall that $|B_d(R)|$ denotes the volume of a Hamming ball of radius R in d-dimensions.

Lemma 11. *[Hamming ball bounds] Assume that $R \le d/2$. The following are upper bounds on $|B_d(R)|$:*

1. *(Trivial) $\forall R \le d/2$, $|B_d(R)| \le R\binom{d}{R}$.*
2. *(Small R) When $R = o(\sqrt{d})$, $|B_d(R)| \le 2\binom{d}{R}$.*
3. *(Large R) Let $D(x||y) = x\log(x/y) + (1-x)\log((1-x)/(1-y))$ denote the "relative entropy" or Kullback-Leibler Divergence[1], defined for $0 \le x,y \le 1$. Then*

$$\frac{2^d}{\sqrt{2\pi d}}exp(-dD(R/d||1/2)) \le |B_d(R)| \le 2^d exp(-dD(R/d||1/2)).$$

[1] Note that $D(p||1/2) = 1 - H(p)$.