Lp Pattern Matching in a Stream
Tatiana Starikovskaya, Michal Svagerka, Przemyslaw Uznański

To cite this version:
Tatiana Starikovskaya, Michal Svagerka, Przemyslaw Uznański. Lp Pattern Matching in a Stream. Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques APPROX-RANDOM 2020, 2020, Virtual conference, France. 10.4230/LIPIcs(APPROX/RANDOM.2020.35. hal-03943021

HAL Id: hal-03943021
https://hal.science/hal-03943021
Submitted on 17 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract

We consider the problem of computing distance between a pattern of length $n$ and all $n$-length subwords of a text in the streaming model.

In the streaming setting, only the Hamming distance ($L_0$) has been studied. It is known that computing the exact Hamming distance between a pattern and a streaming text requires $\Omega(n)$ space (folklore). Therefore, to develop sublinear-space solutions, one must relax their requirements. One possibility to do so is to compute only the distances bounded by a threshold $k$, see [SODA’19, Clifford, Kociumaka, Porat] and references therein. The motivation for this variant of this problem is that we are interested in subwords of the text that are similar to the pattern, i.e. in subwords such that the distance between them and the pattern is relatively small.

On the other hand, the main application of the streaming setting is processing large-scale data, such as biological data. Recent advances in hardware technology allow generating such data at a very high speed, but unfortunately, the produced data may contain about 10% of noise [Biol. Direct.’07, Klebanov and Yakovlev]. To analyse such data, it is not sufficient to consider small distances only. A possible workaround for this issue is the $(1 \pm \varepsilon)$-approximation. This line of research was initiated in [ICALP’16, Clifford and Starikovskaya] who gave a $(1 \pm \varepsilon)$-approximation algorithm with space $\tilde{O}(\varepsilon^{-5} \sqrt{n})$.

In this work, we show a suite of new streaming algorithms for computing the Hamming, $L_1$, $L_2$ and general $L_p$ ($0 < p < 2$) distances between the pattern and the text. Our results significantly extend over the previous result in this setting. In particular, for the Hamming distance and for the $L_p$ distance when $0 < p \leq 1$ we show a streaming algorithm that uses $\tilde{O}(\varepsilon^{-2} \sqrt{n})$ space for polynomial-size alphabets.

1 Introduction

In the problem of pattern matching, we are given a pattern $P$ of length $n$ and a text $T$ and must find all occurrences of $P$ in $T$. A particularly relevant variant of this fundamental question is approximate pattern matching, where the goal is to find all subwords of the text that are similar to the pattern. This can be restated in the following way: given a
pattern \( P \), a text \( T \), and a distance function, compute the distance between \( P \) and every \( n \)-length subword of \( T \). A very natural similarity measure for words is the Hamming distance. Furthermore, if both \( P \) and \( T \) are over an integer alphabet \( \Sigma \), one can consider the Manhattan distance or the Euclidean distance.

**Definition 1** (Hamming, Manhattan and Euclidean distances). For a vector \( U = u_1u_2 \ldots u_n \), its Hamming norm is defined as \( \|U\|_H = |\{i : u_i \neq 0\}| \), Manhattan norm is defined as \( \|U\|_1 = \sum_i |u_i| \) and Euclidean norm is defined as \( \|U\|_2 = (\sum_i u_i^2)^{1/2} \). For two words \( V = v_1v_2 \ldots v_n \) and \( W = w_1w_2 \ldots w_n \), their Hamming distance is defined as \( \|V - W\|_H \), their Manhattan distance as \( \|V - W\|_1 \), and their Euclidean distance as \( \|V - W\|_2 \).

Those distance functions naturally generalize to the so called \( L_p \) distances, where \( p > 0 \) is the exponent.

**Definition 2** (\( p' \)th moment, \( p' \)th norm). For a vector \( U = u_1u_2 \ldots u_n \) and \( p \geq 0 \), its \( p' \)th moment \( F_p \) is defined as \( F_p(U) = \sum_i |u_i|^p \), and for \( p > 0 \) its \( L_p \) norm is defined as \( \|U\|_p = F_p(U)^{1/p} = (\sum_i |u_i|^p)^{1/p} \). For two words \( V = v_1v_2 \ldots v_n \) and \( W = w_1w_2 \ldots w_n \) considered as vectors, the \( p' \)th moment of their difference is \( F_p(V - W) \) and their \( L_p \) distance is defined as \( \|V - W\|_p = F_p(V - W)^{1/p} = (\sum_i |v_i - w_i|^p)^{1/p} \).

In other words, the Manhattan distance is the \( L_1 \) distance, the Euclidean distance is the \( L_2 \) distance, and the Hamming distance can be considered as the \( L_0 \) distance.

Below we assume that the length of the text is \( 2n \), as any algorithm on a text of larger length can be reduced to repeated application of an algorithm that runs on texts of length \( 2n \). This is done by splitting the text into blocks of length \( 2n \) which overlap by \( n \) characters.

**Offline setting.** For the Hamming distance, the problem has been extensively studied in the offline setting, where we assume random access to the input. The first algorithm, for a constant-size alphabet, was shown by Fischer and Paterson [21]. The algorithm uses \( \mathcal{O}(n \log n) \) time and in substance computes the Boolean convolution of two vectors a constant number of times. This was later extended to polynomial-size alphabets in [1, 33]. With a somewhat similar approach, the same complexity can be achieved for the \( L_1 \) distance in [13]. Later, in [34, 35] the authors proved that these problems must have equal (up to polylogarithmic factors) complexities by showing reductions from the Hamming to the \( L_1 \) distance and back.

To improve the complexity for large alphabets, the natural next step was to study approximation algorithms. Until very recently, the fastest \( (1 \pm \varepsilon) \)-approximation algorithm for computing the Hamming distances was by Karloff [29]. The algorithm combines random projections from an arbitrary alphabet to the binary one and Boolean convolution to solve the problem in \( \mathcal{O}(\varepsilon^{-2} n \log^3 n) \) time. In a breakthrough paper Kopelowitz and Porat [31] gave a new approximation algorithm improving the time complexity to \( \mathcal{O}(\varepsilon^{-1} n \log^4 n \log \varepsilon^{-1}) \), which was later significantly simplified [32]. Using a similar technique, Gawrychowski and Uznański [23] showed an approximation algorithm for computing the \( L_1 \) distance in \( \mathcal{O}(\varepsilon^{-4} n \log^4 n) \) (randomized) time, later made deterministic in time \( \mathcal{O}(\varepsilon^{-1} n \log^2 n) \) in [39]. Using similar techniques, the authors of [39] gave \( \mathcal{O}(\varepsilon^{-1} n) \)-time \( (1 + \varepsilon) \)-approximation algorithm for \( L_p \) distances for any constant positive \( p \).

\(^1\) Across the paper we use \( \tilde{\mathcal{O}} \) to indicate that we are suppressing poly-log(n) factors.
Streaming setting. In the streaming setting, we assume that the pattern and the text arrive as streams, one character at a time (the pattern arrives before the text). The main objective is to design algorithms that use as little space as possible, and we must account for all the space used by the algorithm, including the space required to store the input, in full or in part. It is also often the case that the text arrives at a very high speed and we must be able to process it faster than it arrives to fulfill the space guarantees, preferably, in real time. To this aim, the time complexity of streaming algorithms is defined as the worst-case amount of time spent on processing one character of the text, i.e. per arrival.

In the streaming setting, only the Hamming distance ($L_0$) has been studied. It is known that computing the Hamming distance between a pattern and a streaming text exactly requires $\Omega(n)$ space, even for the binary alphabet and with a small probability error allowed, which can be shown by a straightforward reduction to communication complexity (folklore).

Therefore, to develop sublinear-space solutions, one must relax their requirements. One possibility to do so is to compute only the distances bounded by a threshold $k$. This variant of the problem is often referred to as $k$-mismatch problem. The $k$-mismatch problem has been extensively studied in the literature [15, 16, 25, 38], with this line of work reaching $\tilde{O}(k)$ memory complexity and $\tilde{O}(\sqrt{k})$ time per input character. The motivation for this variant of this problem is that we are interested in subwords of the text that are similar to the pattern, in other words, the distance between the pattern and the text should be relatively small. On the other hand, the main application of the streaming setting is processing large-scale data, such as biological data. To decrease the cost of generating such data, recently new hardware approaches have been developed. They have become widely used due to cost efficiency, but unfortunately, the produced data may contain about 10% of noise [30]. To analyse such data, it is not sufficient to consider small distances only, and a possible workaround for this issue is $(1 \pm \varepsilon)$-approximation. This line of research was initiated by Clifford and Starikovskaya [17] who gave a $(1 \pm \varepsilon)$-approximation algorithm with space $\tilde{O}(\varepsilon^{-5} \sqrt{n})$ that uses $\tilde{O}(\varepsilon^{-4})$ time per arriving character of the text.

Independently and in parallel with this work, authors of [12] showed a $(1 \pm \varepsilon)$-approximation streaming algorithm for the $k$-mismatch problem that uses $\tilde{O}(\varepsilon^{-3} \sqrt{k})$ space. For a special case of $k = n$, they show how to reduce the space further to $\tilde{O}(\varepsilon^{-1.5} \sqrt{n})$. Compared to our solution, their algorithm has worse time complexity of $\tilde{O}(\varepsilon^{-3})$ per arrival, and more importantly, it is not obvious whether it can be generalised to other $L_p$ norms as it uses a very different set of techniques.

Sliding window. The problem of computing distance between $P$ and every $n$-length subword of $T$ in the streaming setting resembles the problem of maintaining the $L_p$ norm of a $n$-length suffix of a streaming text, also referred to as sliding window. In fact, the latter is a simplification of the former, with setting $P = [0, 0, \ldots, 0]$. There is an extensive line of work on maintaining the $L_p$ norm of a sliding window, refer to [4, 5, 6, 7, 8, 18] and references therein. The main message is that the norm of a sliding window can be maintained efficiently, e.g. for $1 \leq p \leq 2$ the $L_p$ norms can be maintained $(1 \pm \varepsilon)$-approximately in space $\tilde{O}(\varepsilon^{-1})$. However, those results do not translate to our case: in the sliding window, one can easily isolate “heavy hitters”, that is updates with a significant contribution to the output. In our case, the contribution of an update depends on its relative position to the pattern, and one can easily construct instances where a contribution of a position in the text changes drastically relative to its alignment with the pattern, which necessitates a significantly different approach.
35:4 L_p Pattern Matching in a Stream

1.1 Our results

In this work, we show a suite of new streaming algorithms for computing the Hamming, L_1, L_2 and general L_p (0 < p ≤ 2) distances between the pattern and the text. Our results significantly improve and extend the results of [17].

▶ Theorem 3. Given a pattern P of length n and a text T over an alphabet Σ = [1, 2, ..., σ], where σ = n^{O(1)}, there is a streaming algorithm that computes a (1 ± ε)-approximation of the L_p distance between P and every n-length subword of T correctly w.h.p.

1) in $\tilde{O}(\varepsilon^{-2}\sqrt{n} + \log \sigma)$ space, and $\tilde{O}(\varepsilon^{-2})$ time per arrival when $p = 0$ (Hamming distance);
2) in $\tilde{O}(\varepsilon^{-2}\sqrt{n} + \log^2 \sigma)$ space and $\tilde{O}(\sqrt{n} \log \sigma)$ time per arrival when $p = 1$ (Manhattan distance);
3) in $\tilde{O}(\varepsilon^{-2}\sqrt{n} + \log^2 \sigma)$ space and $\tilde{O}(\varepsilon^{-2}\sqrt{n})$ time per arrival when $0 < p < 1/2$;
4) in $\tilde{O}(\varepsilon^{-2}\sqrt{n} + \log^2 \sigma)$ space and $\tilde{O}(\varepsilon^{-3}\sqrt{n})$ time per arrival when $p = 1/2$;
5) in $\tilde{O}(\varepsilon^{-2}\sqrt{n} + \log^2 \sigma)$ space and $\tilde{O}(\sigma^{\frac{1}{2p-1}} \sqrt{n}/\varepsilon^{2+\frac{2}{p}})$ time per arrival when $1/2 < p < 1$;
6) in $\tilde{O}(\varepsilon^{-2p/2}\sqrt{n} \log^2 \sigma)$ space and $\tilde{O}(\varepsilon^{-p/2}\sqrt{n} + \varepsilon^{-2} \log \sigma)$ time per arrival for $1 < p ≤ 2$.

We also improve and extend the space lower bound of [17], who showed that any streaming algorithm that computes a (1 ± ε)-approximation of the Hamming distance between a pattern and a streaming text must use $\Omega(\varepsilon^{-2} \log^2 n)$ bits for all ε such that $1/\varepsilon < n^{1/2-\gamma}$ for some constant γ (condition inherited from [27]). We show the following result:

▶ Lemma 4. Let $2 ≤ 1/\varepsilon < n$ and $0 ≤ p ≤ 2$. Any (1 ± ε)-approximation algorithm that computes the L_p distance between a pattern and a streaming text for each alignment, must use $\Omega(\min(1/\varepsilon^2, n))$ bits of space.

1.2 Techniques

At a very high level, the structure of all algorithms presented in this paper is similar to that of [17]. We process the text by blocks of length $b \approx \sqrt{n}$. To compute an approximation of the distance / the p'th moment at a particular alignment, we divide the pattern into two parts: a prefix of length $\leq b$ aligned with a suffix of some block of the text, and the remaining suffix (see Fig. 1). We compute an approximation of the distance / the p'th moment for both of the parts and sum them up to obtain the final answer. Our main contribution is a set of new tools that allows computing the approximations efficiently.

To be able to compute the approximation of the distance / the p'th moment between the prefix and the corresponding block of the text, we compute, while reading each block of the text, its compact lossy description that we refer to as prefix encoding. The prefix encoding captures the relation between the read block and the prefix of the pattern of length b. To compute the distance / the p'th moment between the suffix and the text, we will use suffix sketches. For each position i of the text, the suffix sketch describes the subword $T[b \cdot k + 1, i]$ of the text where $k$ is the smallest integer such that $i - b \cdot k \leq n$ (see Fig. 1).

For the Hamming distance, we define the prefix encodings in Section 2.1 and the suffix sketches in Section 3.1. Our Hamming prefix encoding introduces a novel use of a known technique called subsampling. The prefix encodings are used to approximate the distance between any suffix of one word and the prefix of another word of the same length. In brief, the idea is to replace each character of the two words by the don’t care character “?”; a special character that matches any other character of the alphabet. We repeat the process a logarithmic number of times to create a logarithmic number of pairs of “subsamples”. For each pair, we find the longest suffix of one subsample that matches the prefix of the
second subsample up to at most $\Theta(1/\varepsilon^2)$ mismatches. We then show that this information can be used to approximate the Hamming distance between any suffix-prefix pair. Similar techniques were used in [3, 19, 22, 24, 28, 37] for estimating the Hamming norm in streams. The crucial difference with our approach is that we must be able to compute the Hamming norm of any suffix-prefix pair of the two words, and we must be able to do it efficiently. As for the suffix sketches, for the binary alphabet we use the sketches introduced in [17]. We then show a reduction from arbitrary alphabets to the binary alphabet, which improves the space consumption of Hamming suffix sketches by a factor of $1/\varepsilon^2$.

We can solve the problem of $L_1$ (Manhattan distance) pattern matching by replacing each character of the pattern and of the stream with its unary encoding and running the solution for the Hamming distance. However, this would introduce a multiplicative factor of $\sigma$ (the size of the alphabet) to the time complexity. We show efficient randomised reductions from the Manhattan to Hamming distance that allow simulating the solution for the Hamming distance without a significant overhead. In particular, to design the prefix encodings we use random shifting and rounding, while for the suffix sketches we use range-summable hash functions [9]. We show the Manhattan prefix encodings in Section 2.2 and the Manhattan suffix sketches in Section 3.2.

For generic $L_p$ distances, $0 < p \leq 2$, we discuss the prefix encodings in Section 2.4 and the suffix sketches in Section 3.3. Our approach to $L_p$ prefix encodings is rather involved. In the case of $0 < p < 1$, we construct a novel embedding from $L_p$ space into the Hamming space, which might be of independent interest. While the target dimension of the Hamming space is large, we construct the embedding in such a way that each value is mapped into a compressible sequence of form $c_1^{d_1} \ldots c_t^{d_t}$ for some small value of $t$, and where values of $d_1, \ldots, d_t$ are constant across all input values. Such compressed representation allows us to efficiently apply the subsampling framework and reduce the problem to the Hamming distance case. For $1 < p \leq 2$, we identify a logarithmic number of anchor suffixes, and partition each of them into $\varepsilon^{-p}$ words of roughly even contribution to the distance. We then use the partition to decode prefix-suffix distance queries for arbitrary length queries. Such construction is a generalization and improvement of the approach presented in [17]. For suffix sketches, we simply use the $p$-stable distributions [26].

Finally, we combine the prefix encodings and the suffix sketches to prove Theorem 3 in Section 4. To simplify the notation, we use $x \leq y$ to denote $(1 - \varepsilon)y \leq x \leq (1 + \varepsilon)y$ from now on. We will also use the fact that for $p > 0$ we can speak of approximating the $p$’th moment of differences between the pattern and the $n$-length substrings of the text and the $L_p$ distances between the pattern and the $n$-length substrings of the text interchangeably, it changes the complexities up to a constant factor only:
Observation 5. For any constant $p > 0$ and $\varepsilon < 1/2$, there is a constant $C_p$ such that finding a $(1 \pm C_p \cdot p\varepsilon)$ approximation of the $p$’th moment of a vector suffices for $(1 \pm \varepsilon)$-approximating its $p$’th norm, and finding a $(1 \pm C_p \cdot \varepsilon/p)$ approximation of its $p$’th norm suffices for $(1 \pm \varepsilon)$-approximating its $p$’th moment.

2 Prefix encodings

In this section we present a solution to the following problem. Imagine we have a block of text $T'[1, b] = T[i + 1, i + b]$ and a prefix of the pattern $P' = P[1, b]$. We want to find a compressed representation (encoding) of $T'$ so that the following is possible: given any $1 \leq d \leq b$, the compressed representation of $T'$, and $P'$ (explicitly), we can $1 \pm \varepsilon$ approximate $\|T'' - P''\|_p$, where $T'' = T'[b - d + 1, b]$ is a suffix of $T'$ and $P'' = P'[1, d]$ is a prefix of $P'$.

We start by presenting a solution to the Hamming distance case, which is a basis to our solution for all other $L_p$ norms for $0 < p \leq 2$.

2.1 Hamming ($L_0$) distance

Recall that “?” is the don’t care character, a special character that matches any other character of the alphabet.

Definition 6 (Hamming subsampling). Consider a word $U$ of length $n$. Let $q = [3 \log n]$ and let $h(i) : [n] \rightarrow \{0,1\}^q$ be a function drawn at random from a pairwise independent family. For $r = 0, \ldots, q$, we define the $r$-th level Hamming subsample of $U$, $\text{hSub}_r(U)$, as follows:

$$\text{hSub}_r(U)[i] = \begin{cases} U[i], & \text{if the } r \text{ lowest bits of } h(i) \text{ are all } 0; \\ \text{?}, & \text{otherwise.} \end{cases}$$

In particular, $\text{hSub}_0(U) = U$.

Fix an integer $k = \Theta(1/\varepsilon^2)$ large enough. For two words $U, V$, consider the following estimation procedure:

Algorithm 7.

1. Denote $X_r$ to be the Hamming distance between $\text{hSub}_r(U)$ and $V$ and let $f = \min\{i : X_i \leq k\}$.\footnote{We emphasize that $\text{hSub}_r(U)$ contains don’t care characters, so the Hamming distance is defined as the number of pairs of characters of $\text{hSub}_r(U)$ and $V$ that do not match.}

2. Output $Z_f = 2^f \cdot X_f$ as an estimate of $\|U - V\|_H$.

The following lemma is a rephrasing of a known result regarding subsampling in estimation of the Hamming norm (cf. [3, Theorem 3], or [24, Theorem 2]).

Lemma 8. For $Z_f$ as in Algorithm 7 there is $Z_f \leq \|U - V\|_H$ with probability at least $3/4$. Since the subsampling is performed independently for each position, one can use subsampling to approximate the Hamming distance between any suffix of $B$ and any prefix of $P$ of equal lengths in a similar fashion.

We are now ready to define the Hamming prefix encoding of a block. For brevity, let $B'_j = \text{hSub}_r(B)[b - j + 1, b]$ and $P'_j = P[1, j]$ (the same for all $r$). Furthermore, given two words $U, V$ of equal length, define the mismatch information $\text{MI}(U, V) = \{(i, U[i], V[i]) : U[i] \text{ does not match } V[i]\}$.\footnote{We emphasize that $\text{hSub}_r(U)$ contains don’t care characters, so the Hamming distance is defined as the number of pairs of characters of $\text{hSub}_r(U)$ and $V$ that do not match.}
Definition 9. Consider a $b$-length block $B$ of the text $T$. For each $0 \leq r \leq \lceil 3 \log n \rceil$, let $j^*(r)$ be the maximal integer such that the Hamming distance between $B^r_{j^*(r)}$ and $P^r_{j^*(r)}$ is at most $k = \Theta(\varepsilon^{-2})$. We define the Hamming prefix encoding of $B$ to be a tuple of pairs $j^*(r), M(B^r_{j^*(r)}, P^r_{j^*(r)})$.

Note that the prefix encoding of $B$ uses $O(k \log n) = O(\varepsilon^{-2} \log n)$ space. We can compute it efficiently:

Lemma 10. Assume constant-time random access to $P[1, b]$. Given a $b$-length block $B$ of the text $T$, its Hamming prefix encoding can be computed in $\tilde{O}(kb) = \tilde{O}(b \varepsilon^{-2})$ time.

Proof. To compute the encoding, we use the algorithm of [14]. Formally, for each $r$, we create a word $T'$ by appending $b$ don’t care characters to the subsample $\text{hSub}_r$. The algorithm of [14] can be used to find all $b$-length subwords of $T'$ that match $P[1, b]$ up to $k$ mismatches, moreover for each of these subwords the algorithm outputs the mismatch information. We take the leftmost subword only, which corresponds to $j^*(r)$ because of the don’t care characters. In total, our algorithm uses $\tilde{O}(kb) = \tilde{O}(\varepsilon^{-2}b)$ time.

We now show how to compute the Hamming distance between any $j$-length suffix of $B$ and any $j$-length prefix of $P$ given $P[1, b]$ and the Hamming prefix encoding of a block $B$.

Lemma 11. Given the prefix encoding of a $b$-length block $B$ of the text $T$, there is an algorithm that computes, for any $j = 1, \ldots, b$, a $(1 + \varepsilon)$-approximation of the Hamming distance between the $j$-length suffix of $B$ and the $j$-length prefix of $P$ in $\tilde{O}(kb) = \tilde{O}(b \varepsilon^{-2})$ time.

Proof. Denote $X_r$ to be the Hamming distance between $P^j_r$ and $B^r_j$. We compute the smallest $f$ such that $X_f \leq k$ in the following way. For each $r$, we use $M(B^r_{j^*(r)}, P^r_{j^*(r)})$ to restore $B^r_{j^*(r)}$. We then append $P^r_{j^*(r)}$ with $b$ don’t care characters and run the algorithm of [14] for the resulting text and the pattern. This allows to compute $X_r$ for all $j \leq j^*(r)$, and if $j > j^*(r)$, then $X_f > k$ by definition. In total, the algorithm takes $\tilde{O}(kb) = \tilde{O}(\varepsilon^{-2}b)$ time.

2.2 Manhattan ($L_1$) distance

Recall a word morphism $\nu : \Sigma \to \{0, 1\}^\sigma$, $\nu(a) = 1^a0^\sigma-a^a$. Our goal in this section is to simulate implicitly procedures from Lemma 10 and Lemma 11 on words $\nu(B)$ and $\nu(T)$ without introducing any significant overhead.

Definition 12 (Manhattan scaling). Consider a word $U$ of length $n$. Let $q = \lceil 3 \log n \sigma \rceil$ and let $h : \mathbb{N} \to 2^q$ be a function drawn at random from a 4-wise independent family. For $r = 0, \ldots, q$, we define the $r$-th level Manhattan subsample of $U$, $\text{mSub}_r(U)$, as a word of length $n$ such that $\text{mSub}_r(U)[i] = \lfloor \frac{U[i] + h(i) \mod 2^q}{2} \rfloor$. In particular, $\text{mSub}_0(U) = U$.

Fix an integer $k = \Theta(1/\varepsilon^2)$ large enough. For words $U, V$, consider $\text{mSub}_r(U), \text{mSub}_r(V)$ for all $r = 0, \ldots, q$, and the following estimation procedure:

Algorithm 13.
1. Denote $X_r = \|\text{mSub}_r(U) - \text{mSub}_r(V)\|_1$ and let $f = \min\{i : X_i \leq k\}$.
2. Output $Z_f = 2^f \cdot X_f$ as an estimate of $\|U - V\|_1$.

Lemma 14. For $Z_f$ as in Algorithm 13 there is $Z_f \tilde{=} \|U - V\|_1$ with probability $\geq 3/4$. 

APPROX/RANDOM 2020
To approximate the Manhattan distance between any suffix of $B$ and any prefix of $P$ of equal lengths, we define the encoding similar to the Hamming distance case. Specifically, we still use the mismatch information, building on the fact that for any two words $\| U - V \|_1 \leq \| U - V \|_1$ and from the mismatch information the exact value of $\| U - V \|_1$ can be found. We define $B_j^i = \text{Sub}_j(B)[b - j + 1, b]$ as before, but change the definition of $P_j^i$ slightly. Intuitively, we define $P_j^i$ to be the $j$-length prefix of $P$ subsampled in a synchronized way with $B_j^i$. Formally, $P_j^i[i] = \left\lceil \frac{P_j^i[i]+(b(b-j+i) \mod 2^j)}{b} \right\rceil$.

▶ **Definition 15.** Consider a $b$-length block $B$ of the text $T$. For each $0 \leq r \leq \lfloor 3 \log n \sigma \rfloor$, let $j^*(r)$ be the maximal integer such that the Manhattan distance between $B_{j^*(r)}^i$ and $P_{j^*(r)}^i$ is at most $k = \Theta(\varepsilon^{-2})$. We define the Manhattan prefix encoding of $B$ to be a tuple of pairs $j^*(r), \text{MI}(B_{j^*(r)}^i, P_{j^*(r)}^i)$.

Note that the prefix encoding of $B$ uses $O(k \log n \sigma) = O(\varepsilon^{-2} \log n)$ space.

▶ **Lemma 16.** Assume constant-time random access to $P[1, b]$. Given a $b$-length block $B$ of the text $T$, its Manhattan prefix encoding can be computed in $\tilde{O}(b^2)$ time and $O(b)$ space.

**Proof.** Let $q = \lfloor 3 \log n \sigma \rfloor$. For each $r = 0, \ldots, q$ and $j = 1, \ldots, b$ we compare $B_j^r$ and $P_j^r$ character by character in $O(b)$ time to find $j^*(r)$ and the corresponding mismatch information. The claim follows.

▶ **Lemma 17.** Given the prefix encoding of a $b$-length block $B$ of the text $T$, there is an algorithm that computes, for all $j = 1, \ldots, b$, a $(1 + \varepsilon)$-approximation of the Manhattan distance between the $j$-length suffix of $B$ and the $j$-length prefix of $P$ in $\tilde{O}(b^2)$ time.

**Proof.** Denote $X_r = \| P_j^i - B_j^i \|_H$. We compute the smallest $f$ such that $X_f \leq k$ in the following way. For each $r$, we use $\text{MI}(B_{j^*(r)}^i, P_{j^*(r)}^i)$ to restore $B_{j^*(r)}^i$. If $j > j^*(r)$, the Manhattan distance between $P_j^i$ and $B_j^i$ is at least $k$. Otherwise, we compare $P_j^i$ and $B_j^i$ character by character to compute the Manhattan distance in $O(b)$ time. The claim follows.

### 2.3 Generic $\langle L_p \rangle$ distance for $0 < p < 1$

Our goal is to construct a morphism (parametrised by $p$) acting as a randomized embedding of $\langle L_p \rangle$ into the Hamming distance. The intuition behind our approach is as follows. Let $r_0, r_1, \ldots \in [0, 1]$ be a sequence of real numbers picked independently and u.a.r. Define a sequence of values

$$d_i = \begin{cases} 
\varepsilon^{-1} \cdot (1 + \varepsilon)^p, & \text{when } i > 0 \\
\varepsilon^{-1} \cdot \frac{(1+\varepsilon)^p}{(1+\varepsilon)^{p-1}}, & \text{when } i = 0 
\end{cases}$$

and for a character $c \in \Sigma$ consider sequence of characters $c_0, c_1, \ldots$ where $c_i = \lfloor \frac{c_i}{(1+\varepsilon)^\ell} + r_i \rfloor$ (similarly, a character $c'$ defines a sequence $c'_0, c'_1, \ldots$). Now consider two characters $c, c' \in \Sigma$ such that $|c - c'| = (1 + \varepsilon)^\ell$ for some integer $\ell$ and a random variable $x = \sum_{i=0}^{\infty} d_i \cdot \|c_i - c'_i\|_H$. 
There is
\[
\mathbb{E}[x] = \sum_{i=0}^{\infty} d_i \cdot \Pr[c_i \neq c'_i] = \sum_{i=0}^{t} d_i \cdot 1 + \sum_{i=t+1}^{\infty} d_i \cdot \frac{|c - c'|}{(1 + \varepsilon)^i} =
\]
\[
eq \varepsilon^{-1} \left( \frac{(1 + \varepsilon)^p}{(1 + \varepsilon)^p - 1} + \varepsilon^{-1} \sum_{i=1}^{\ell} \frac{(1 + \varepsilon)^p}{1 - (1 + \varepsilon)^p} \right) =
\]
\[
eq \varepsilon^{-1} \left( \frac{(1 + \varepsilon)^{t+1} p}{(1 + \varepsilon)^p - 1} + (c - c') \frac{(1 + \varepsilon)^{t+1}}{1 - (1 + \varepsilon)^p} \right) =
\]
\[
\approx \varepsilon^{-2} |c - c'| \frac{1}{p(1 - p)}. \tag{1}
\]

We thus see that an idealized morphism of the form \( \varphi : c \to \epsilon_0^{d_0} \epsilon_1^{d_1} \cdots \) would have the property that \( \|U - V\|_p \sim \|\varphi(U) - \varphi(V)\|_H \) on words of length \( n \). But there are the following issues: (i) characters are mapped into infinite length words, (ii) number of repetitions of characters \( (d_i) \) is fractional, (iii) we cannot guarantee that character distance is always of form \( (1 + \varepsilon)^t \) and (iv) the distance is preserved only in expectation. We show how to overcome these issues to achieve the following result:

- **Theorem 18.** Given \( 0 < p < 1 \) and \( \varepsilon > 0 \) there is a word morphism \( \varphi : c \to \epsilon_0^{d_0} \epsilon_1^{d_1} \cdots \epsilon_{t-1}^{d_{t-1}} \) such that:
  1. \( t = \bar{O}(\varepsilon^{-2}) \) when \( 0 < p < 1/2 \), \( t = \bar{O}(\varepsilon^{-3}) \) when \( p = 1/2 \) and \( t = \bar{O}(\sigma^{2p-1} / \varepsilon^{2+3} 2^{p-1}) \) when \( 1/2 < p < 1 \).
  2. values of \( t \) and \( d_0, \ldots, d_{t-1} \) do not depend on \( c \),
  3. there exists a constant \( \alpha = \alpha(p, \varepsilon) \) such that for any two words \( U, V \) of length at most \( n \), we have \( \|U - V\|_p \leq \alpha \cdot \|\varphi(U) - \varphi(V)\|_H \) with probability at least \( 9/10 \),
  4. it is enough for the randomness to be realized by a hash function \( r : [t] \to [D] \) from a \( 4 \)-wise independent hash function family for some \( D = \text{poly}(n \sigma \varepsilon^{-1}) \), which can be generated from a \( \bar{O}(\log \sigma) \) bits size seed.

We now describe how to use the morphism \( \varphi \) to approximate the \( L_p \) distances in a small space. To design an efficient algorithm, we take advantage of the fact that \( \varphi(U) \) has a compressed representation of size comparable with the length of \( U \) (at least when \( p \leq 1/2 \)).

- **Definition 19** \( (L_p \text{ scaling}) \). Consider a word \( S = s_1^{e_1} s_2^{e_2} \ldots s_m^{e_m} \) of length \( m' = \sum_i e_i \). Let \( h : [m] \to 2^q \) be a function drawn at random from a \( 4 \)-wise independent family, where \( q = \lfloor 3 \log m' \rfloor \). For \( r = 0, \ldots, q \), we define the \( r \)-th level subsample of \( S \),

\[
\text{Sub}_r(S) = (s_1)^{\lfloor (r + (h(1) \mod 2^q)) / 2 \rfloor} (s_2)^{\lfloor (r + (h(2) \mod 2^q)) / 2 \rfloor} \ldots (s_m)^{\lfloor (r + (h(m) \mod 2^q)) / 2 \rfloor}
\]

In particular, \( \text{Sub}_0(U) = U \).

Consider two words \( S, Q \) of form \( S = s_1^{e_1} \ldots s_m^{e_m} \) and \( Q = q_1^{e_1} \ldots q_m^{e_m} \). Fix an integer \( k = \Theta(1/\varepsilon^2) \) large enough and consider \( \text{Sub}_r(S), \text{Sub}_r(Q) \) for all \( r = 0, 1, \ldots, \lfloor 3 \log m' \rfloor \), where \( m' = \sum_i e_i \).

- **Algorithm 20.**
  1. Denote \( X_r = \|\text{Sub}_r(S) - \text{Sub}_r(Q)\|_H \) and let \( f = \min\{i : X_i \leq k\} \).
  2. Output \( Z_f = 2^f \cdot X_f \) as an estimate of \( \|S - Q\|_H \).

- **Lemma 21.** For \( Z_f \) as in Algorithm 20 there is \( Z_f \approx \|S - Q\|_H \) with probability \( \geq 3/4 \).
We are now ready to define $L_p$ prefix encodings. Consider a $b$-length block $B$ of the text and define $B_i^j = \text{Sub}_{[b]}(\varphi(B))[(b - j)t + 1, bt]$ ($t$ is defined as in Theorem 18). Also, define $P_i^j$ to be the $(tj)$-length prefix of $\varphi(P)$ subsampled in a synchronized way with $B_i^j$.

**Definition 22.** Consider a $b$-length block $B$ of the text $T$. For each $r = 0, \ldots, \lceil 3 \log n' \rceil$, where $n' = |\varphi(B)|$, let $j^*(r)$ be the maximal integer such that the Hamming distance between $B_i^j$ and $P_i^j$ is at most $k = \Theta(\varepsilon^{-2})$. We define the $L_p$ prefix encoding of $B$ to be a tuple of pairs $j^*(r), \text{MI}(B_i^j, P_i^j)$.

The $L_p$ prefix encoding of $B$ uses $O(k \log n') = O(\varepsilon^{-2} \log(n \sigma \varepsilon^{-1}))$ space.

**Lemma 23.** Assume constant-time random access to $P[1, b]$. Given a $b$-length block $B$ of the text $T$, its $L_p$ prefix encoding can be computed in $O(b^2 \cdot t \log n \sigma \varepsilon^{-1})$ time and $O(b + \varepsilon^{-2} \log n \sigma \varepsilon^{-1})$ space.

**Proof.** For each $r = 0, \ldots, \lceil 3 \log n' \rceil$ and $j = 1, \ldots, b$, we compute the Hamming distance between $B_i^j$ and $P_i^j$ in $O(bt)$ time using the compressed representation to find $j^*(r)$ and the corresponding mismatch information. The claim follows.

**Lemma 24.** Given the $L_p$ prefix encoding of a $b$-length block $B$ of the text $T$, there is an algorithm that computes, for all $j = 1, \ldots, b$, a $(1 \pm \varepsilon)$-approximation of the $L_p$ distance between the $j$-length suffix of $B$ and the $j$-length prefix of $P$ in $O(b^2 \cdot t \log n \sigma \varepsilon^{-1})$ time and $O(b + \varepsilon^{-2} \log n \sigma \varepsilon^{-1})$ space.

**Proof.** Denote $X_r = \|P_i^j - B_i^j\|_H$. We compute the smallest $f$ such that $X_f \leq k$ in the following way. For each $r$, we use $\text{MI}(B_i^j, P_i^j)$ to restore $B_i^j$. If $j > j^*(r)$, the Hamming distance between $P_i^j$ and $B_i^j$ is at least $k$. Otherwise, we compare $P_i^j$ and $B_i^j$ to compute the Hamming distance in $O(bt)$ time. The claim follows.

### 2.4 Generic ($L_p$) distance for $1 < p \leq 2$

For $1 < p \leq 2$, we use a scheme similar to the one developed in [17] for the Hamming distance, but adapt it to generic $L_p$ distances. Particularly, we plug in a standard tool used in this situation, the $p$-stable distribution. We additionally have to adapt the scheme a bit, taking into account that $L_p$ norm is sub-additive under concatenation when $p > 1$.

**Definition 25 ($p$-stable distribution [40]).** For a parameter $p > 0$, we say that a distribution $D$ is $p$-stable if for all $a, b \in \mathbb{R}$ and random variables $X, Y$ drawn independently from $D$, the variable $aX + bY$ is distributed as $(|a|^p + |b|^p)^{1/p} Z$, where $Z$ is a random variable with distribution $D$.

Consider a word $X = x_1x_2 \ldots x_n$, and let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be independent random variables drawn from a $p$-stable distribution $D$ with expected value $\mu_D$. By Definition 25, we have $E[\sum_i \alpha_i x_i] / \mu_D = ||X||_p$. The $p$-stable distributions exist for all $0 < p \leq 2$, and a random variable $X$ from a $p$-stable distribution can be generated using the formula

$$X = \frac{\sin(p\Theta)}{\cos^{1/p}(\Theta)} \left( \frac{\cos(\Theta(1-p))}{\ln(1/r)} \right)^{(1-p)/p}$$

[11, 40], where $\Theta$ is uniform on $[-\pi/2, \pi/2]$ and $r$ is uniform on $[0, 1]$.

However, to be able to design an efficient sketching scheme that allows to approximate the $L_p$ norm with high probability, there are three technicalities to be overcome: First, one must show that $\sum \alpha_i x_i$ concentrates well, second, the formula above assumes infinite precision of computation, and finally, one cannot use fully independent random variables $\alpha_i$ as above.
Given the sketches choose \( q \) or the alphabet of size \( k, j \) and between \( B_2 \) and \( P_2 \) we use the sketches.

as this would require much space. To overcome these issues, Indyk [26] combined \( p \)-stable distributions and pseudorandom generators for bounded space computation [36]. We restate the final result of Indyk below, in the form that will be convenient for us later.

\[ \text{Lemma 29.} \]

\[ \text{O}(\varepsilon^{-2} \log \log(n/\varepsilon)) \]

\[ \text{Lemma 28.} \]

\[ \text{O}(\varepsilon^{-2} \log(\varepsilon n/\varepsilon) \log(n/\varepsilon)) \]

\[ \text{Theorem 26 (cf. Theorem 2, Theorem 4 [26]).} \]

For any \( 0 < p \leq 2 \), there is a non-uniform streaming algorithm that maintains a sketch \( \text{Sketch}_p(S) \) of a word \( S \) of length \( n \) over an alphabet of size \( \sigma \) such that:

1) when a new character of \( S \) arrives, the sketch can be updated in \( O(\varepsilon^{-2} \log(n/\varepsilon)) \) time;
2) the algorithm and the sketch use \( O(\varepsilon^{-2} \log(\sigma n/\varepsilon) \log(n/\varepsilon)) \) bits of space.

Given the sketches \( \text{Sketch}_p(X), \text{Sketch}_p(Y) \) of two words \( X, Y \) of length \( n \), one can estimate \( \|X - Y\|_p \) up to a factor \( 1 \pm \varepsilon \) with probability at least \( 9/10 \) in time \( \tilde{O}(1/\varepsilon^2) \).

We now proceed to building the \( L_p \) prefix encoding by using \( \text{Sketch}_p \) and the landmarking technique.

\[ \text{Definition 27 (L_p prefix encoding).} \]

Let \( 1 < p \leq 2 \). Consider a word \( S \) of length \( b \) on the alphabet of size \( \sigma \). Define \( q_0 = b \). For \( k = 0, \ldots, \lfloor \log b \sigma^p \rfloor \), let \( q_k \leq q_{k-1} \) be the leftmost position such that the \( p \)th moment of the difference between \( S[q_k, b] \) and \( P[1, b - q_k + 1] \), i.e.

\[ \|S[q_k, b] - P[1, b - q_k + 1]\|_p^p \] is at most \( 2^k \).

Further, divide \( S[q_k, b] \) into \( \Theta(1/\varepsilon^p) \) blocks such that each block is either a single character, or the \( p \)th moment of the difference between each block and the corresponding subword of \( P[1, b - q_k + 1] \) is at most \( \varepsilon^p \cdot 2^k \). Let \( q_k = q_k^0 \leq q_k^1 \leq \ldots \leq q_k^i = b \) be the block borders. We choose \( q_k^1, q_k^2, \ldots, q_k^i \) from left to right, and each position \( q_k^i \) is chosen to be the rightmost possible.

The \( L_p \) prefix encoding of \( S \) is defined to contain sorted lists of the positions \( q_k \) and \( q_k^i \), characters \( S[q_k^i] \), and sketches for \( (1 \pm C_p : \varepsilon/p) \)-approximating the \( p \)th norm of \( S[q_k^i, b] \). for all \( k, j \) and \( C_p \) as in Observation 5, see also Theorem 26.

The encoding takes \( \tilde{O}(\varepsilon^{-2} \log \sigma \log(\sigma n/\varepsilon) \log(n/\varepsilon)) \) bits of space. We now show that given the \( L_p \) prefix encoding of a block \( B \) of the text of length \( b \), one can compute a \( (1 \pm \varepsilon) \)-approximation of the \( L_p \) distance between any prefix \( P[1, b - j + 1] \) of the pattern \( P \) and the corresponding suffix \( B[j, b] \) of \( B \).

\[ \text{Lemma 28.} \]

Let \( 1 < p \leq 2 \). For any two vectors \( X, Y \) of equal length, \( \|X + Y\|_p - \|X\|_p = O(\|Y\|_p + \|Y\|_p \cdot \|X\|_p^{-1}) \).

\[ \text{Lemma 29.} \]

Let \( 1 < p \leq 2 \). Given the \( L_p \) prefix encoding of a block \( B \) of the text \( T \) of length \( b \), one can find \( (1 \pm \varepsilon) \)-approximation of the \( p \)th moment of the difference between any prefix \( P[1, b - j + 1] \) of the pattern \( P \) and the corresponding suffix \( B[j, b] \) of \( B \) in time \( \tilde{O}(\varepsilon^{-2} + \log \sigma) \).
**Proof.** Let $q_k$ be the position that is closest to $i$ from the left, and $q_k^+ < j < q_k^{i+1}$ (see Fig. 2). We can find $q_k, q_k^+, q_k^{i+1}$ in time $O(\log(b\sigma) + 1/\varepsilon_p)$ by iterating over the sorted lists.

The position $q_k^{i+1}$ divides $P[1, b - j + 1]$ into two parts, $P_1$ and $P_2$. Denote $B_1$ and $B_2$ the respective subwords of $B$ they are aligned with (see Fig. 2). Let $m_1 = F_p(P_1 - B_1)$ and $m_2 = F_p(P_2 - B_2)$. Then $m = F_p(P[1, b - j + 1] - B[j, b])$, being the value we need to approximate, is equal to $m_1 + m_2$.

We can find $m_2^{\prime} = m_2$ using the sketches for $B_2 = B[q_k^{i+1}, b]$ and $P_2$ in time $O(1/\varepsilon^2)$. Furthermore, if $q_k^+ = q_k^{i+1} - 1$, then we can compute $m_1$ exactly as we store $B[q_k^+]$. Otherwise, we consider the subword $\tilde{P} = P[j - q_k + 1, q_k^{i+1} - q_k + 1]$ of the pattern $P$. Denote $m_1^{\prime} = F_p(P_1 - \tilde{P})$ and use it as our estimation of $m_1$.

Since $1 \leq p \leq 2$, by definition, $F_p(B_1 - \tilde{P}) \leq \varepsilon^p \cdot 2^{k-1}$, and $F_p(P_1 - B_1) \leq 2^k$. By Lemma 28 with $X = P_1 - B_1$ and $Y = B_1 - \tilde{P}$,

$$|m_1^{\prime} - m_1| = O(||B_1 - \tilde{P}||^p_p + ||B_1 - \tilde{P}||_p||P_1 - B_1||^{p-1}_p) = O(\varepsilon^p 2^k + \varepsilon (2^k)^{2\sigma^2} b) = O(\varepsilon 2^k)$$

and finally $|m_1^{\prime} - m_1| \leq O(\varepsilon m) + \varepsilon m_2 = O(\varepsilon m)$. ▶

**Lemma 30.** Let $1 < p \leq 2$. The $L_p$ prefix encoding of a $b$-length block $B$ of the text can be computed in time $O(b^2 + \varepsilon^{-2} b \log \sigma)$ and space $O(b + \varepsilon^{-2} b \log^2 \sigma)$.

**Proof.** For $j = 1, \ldots, b$, we naively compute the $L_p$ distance between the suffix of $B$ and the prefix of $P$ in $O(b)$ time. We then find the positions $q_k$. For each $k = 0, \ldots, \lceil \log b\sigma \rceil$, we can find the positions $q_k^+$ in $O(b)$ time and compute the sketches in $O(\varepsilon^{-2} b)$ time by Theorem 26. ▶

## 3 Suffix sketches

In this section, we give the definitions and explain how we maintain the suffix sketches for each of the distances.

### 3.1 Hamming distance

We first recall Euclidean suffix sketches as presented in [17]. In fact, we will not use them for the Euclidean distance as for it we can use the generic solution of Section 3.3, but they will serve as a foundation of Hamming suffix sketches.

All sketches presented in this section are correct with constant probability, which can be amplified to $1 - \delta$ for arbitrarily small $\delta$ by a standard method of repeating sketching independently $\Theta(\log \delta^{-1})$ times and taking the median of the estimates.

**Lemma 31 (Euclidean sketches [2]).** Let $M$ be a random matrix of size $d \times n$ filled with 4-wise independent random $\pm 1$ variables, for $d = \Theta(\varepsilon^{-2})$ chosen big enough. For a vector $X \in \mathbb{R}^n$ there is $\frac{1}{\sqrt{d}} ||MX||_2 \leq ||X||_2$ with constant probability $9/10$, taken over all possible choices of $M$. We say that a vector $MX$ of dimension $d$ is a Euclidean sketch of $X$.

**Definition 32 (Euclidean suffix sketches [17]).** Consider a word $X$ of length $n$. We define its Euclidean suffix sketch as follows.

Let $b$ be the block length. Let $\mathcal{R}$ be a random matrix of size $d \times b$ filled with 4-wise independent random $\pm 1$ variables and let $\alpha_1, \ldots, \alpha_{n/b}$ be 4-wise independent random coefficients with values $\pm 1$ as well. We define a matrix $M$ of size $d \times n$ such that $M_{i,jb+k} = \alpha_j \cdot \mathcal{R}_{i,k}$.

Let $X'$ be a word of length $\lceil n/b \rceil \cdot b$ obtained from $X$ by appending an appropriate number of zeros. The Euclidean suffix sketch of $X$ is defined as $\text{eSketch}(X) = MX'$, where $X'$ is considered as a vector.
Observe that the matrix $M$ does not need to be accessed explicitly. Indeed, from $MX' = \sum_{i} \alpha_i \cdot R_i \cdot [X'[b], \ldots, X'[b+i-1]]^T$, it follows that the Euclidean suffix sketch can be computed by first sketching each block of $X'$ using the matrix $R_i$, and then taking a linear combination of the sketches of the blocks (using the random ±1 coefficients $\alpha_i$).

Lemma 33 ([17]). Selecting $d = \Theta(\varepsilon^{-2})$ gives $\frac{1}{\sqrt{d}}\|eSketch(X)\|_2 \leq \|X\|_2$ with probability at least 9/10 (taken over all possible choices of $R_i, \alpha_i$).

By linearity of sketches, we obtain $\|X - Y\|_2 \leq \frac{1}{\sqrt{d}}\|eSketch(X) - eSketch(Y)\|_2$ with probability at least 9/10 as well.

We now define Hamming suffix sketches. First note that for binary words $X, Y$ there is $Ham(X, Y) = \|X - Y\|_2$, and therefore in the case of the binary alphabet we can use the Euclidean suffix sketches. We will now show how to reduce the case of arbitrary polynomial-size alphabets to the case of the binary alphabet.

To this end, [17] used a random mapping of Karloff [29] as a black-box reduction, which led to sketches of size $\sim \varepsilon^{-4}$. We now show a more careful reduction to avoid this overhead and to achieve dependency $\varepsilon^{-2}$ in total. Consider a word morphism defined on alphabet as $\mu : \Sigma \to \{0, 1\}^\ast$, $\mu(a) = 0^a1^{\sigma_a-1}$ (and acting on words by concatenating the images of each character of the input word). Note that $\|\mu(X) - \mu(Y)\|^2_2 = 2^{-\|X - Y\|_H}$, thus using the Euclidean suffix sketches on top of $\mu(X)$ and $\mu(Y)$ allows computation of the respective Hamming distance. Formally,

Definition 34 (Hamming suffix sketches [17]). Consider a word $X$ of length $n$ on the alphabet of size $\sigma$. We define its Hamming suffix sketch as follows.

Let $b$ be the block length, $R$ be a random matrix of size $d \times \sigma b$ filled with 4-wise independent random variables, and $\alpha_1, \ldots, \alpha_{\lceil n/b \rceil}$ be 4-wise independent random coefficients with values ±1 as well. We define a matrix $M$ of size $d \times \sigma n$ such that $M_{i, \sigma b+k} = \alpha_i \cdot R_{i,k}$.

Let $X'$ be a word of length $\lceil n/b \rceil \cdot b$ obtained from $X$ by appending an appropriate number of zeroes. The Hamming suffix sketch of $X$ is defined as $hSketch(X) = M\mu(X')$, where $\mu(X')$ is considered as a vector.

Lemma 35. Selecting $d = \Theta(\varepsilon^{-2})$ gives $\frac{1}{2d}\|hSketch(X)\|_2 \leq \|X\|_H$ with probability at least 9/10 (taken over all possible choices of $R_i, \alpha_i$).

Proof. Follows immediately as a corollary of Lemma 33 and the properties of the embedding $\mu$. In more detail, the following holds with probability at least 9/10:

$$\frac{1}{2d} \cdot \|hSketch(X)\|_2^2 = \frac{1}{2d} \cdot \|M\mu(X')\|_2^2 = \frac{1}{2d} \cdot \|M\mu(X)\|_2^2 = \frac{1}{2d} \cdot \|eSketch(\mu(X))\|_2^2 \leq \frac{1}{2} \cdot \|\mu(X)\|_2\| \leq \|X\|_H.$$ 

As $\mu(X), \mu(Y)$ are sparse, there is an efficient streaming algorithm for maintaining the Hamming suffix sketches of a text:

Lemma 36. Given a text $T$, there is a streaming algorithm that for every position $i$ outputs the Hamming suffix sketch of a word $T[b \cdot k+1, i]$, where $k$ is the largest integer such that $i - b \cdot k \leq n$. The algorithm takes $O(dn/b + \log d\sigma n)$ space and $O(d + n/b^2)$ time per character.
3.2 Manhattan ($L_1$) distance

To show efficient suffix sketches for the Manhattan distance, we consider a word morphism $\nu : \Sigma \rightarrow \{0, 1\}^a$, $\nu(a) = 1^{a0^{a-1}}$. Note that $\|\nu(X) - \nu(Y)\|_2^2 = \|\nu(X) - \nu(Y)\|_H = \|X - Y\|_1$, thus using the Hamming suffix sketches on top of $\nu(X)$ and $\nu(Y)$ allows computation of the respective Manhattan distance.

However, if we apply the morphism straightforwardly, we will have to pay an extra $\sigma$ factor per character to compute the Manhattan suffix sketches. To improve the running time, we will use range-summable hash functions. Range-summable hash functions were introduced by Feigenbaum et al. [20], and later their construction was improved by Calderbank et al. [9].

Definition 37 (cf. [9]). A family $\mathcal{H}$ of hash functions $h(x; \xi) : [t] \times \{0, 1\}^a \rightarrow \{-1, 1\}$ (here $x$ is the argument and $\xi$ is the seed) is called $k$-independent, range-summable if it satisfies the following properties for any $h \in \mathcal{H}$:
1. (k-independent) for all distinct $0 \leq x_1, \ldots, x_k < t$ and all $b_1, \ldots, b_k \in \{-1, +1\}$,

   $$\Pr_{\xi \in \{0, 1\}^k}[h(x_1; \xi) = b_1 \land \cdots \land h(x_k; \xi) = b_k] = 2^{-k}$$

2. (range-summable) there exists a function $g$ such that given a pair of integers $0 \leq \alpha, \beta \leq \sigma$, and a seed $\xi$, the value $g(\alpha, \beta; \xi) = \sum_{0 \leq z < \beta} h(x; \xi)$ can be computed in time polynomial in $\log t$.

Corollary 38 (cf. Theorem 3.1 [9]). There is a $4$-independent, range-summable family of hash functions $h(x; \xi) : [t] \times \{0, 1\}^a \rightarrow \{-1, +1\}$ with a random seed $\xi$ of length $s = O(\log^2 t)$ such that any range-sum $g(\alpha, \beta; \xi)$ can be computed in $O(\log^3 t)$ time.

Observation 39. For a word $X = x_1x_2\ldots x_n$, let $Y = \nu(X) = y_1y_2\ldots y_{ns}$. Let $h, g$ be as in Corollary 38 with $t = ns$. Then $\sum_{i=1}^n g(\sigma_i, \sigma_i + x_i; \xi) = \sum_{i=1}^n y_i h(i; \xi)$.

Thus, we see that range-summable hash functions can be used to efficiently simulate $\nu$.

Definition 40 (Manhattan suffix sketches). Consider $X$ be a word of length $n$. We define its Manhattan suffix sketch as follows.

Let $b$ be the block length. Let $h, g$ be as in Corollary 38 with $t = bd\sigma$. Let $R$ be a random matrix of size $d \times (d\sigma)^t$ filled with $d$-wise independent random $\pm 1$ variables, such that $R_{i,k} = h(\sigma_i + k; \xi)$ and let $\alpha_1, \ldots, \alpha_{(n/b)}$ be $d$-wise independent random coefficients with values $\pm 1$ as well. We define a matrix $M$ of size $d \times (d\sigma)^t$ such that $M_{\alpha, \sigma_i + k} = \alpha_{i,k} R_{i,k} = \alpha_{i,k} h(\sigma_i + k; \xi)$.

Let $X'$ be a word of length $\lceil n/b \rceil b$ obtained from $X$ by appending an appropriate number of zeroes. The Manhattan suffix sketch of $X$ is defined as $\text{mSketch}(X) = M \nu(X')$, where $\nu(X')$ is considered as a vector.

Lemma 41. Selecting $d = \Theta(1/\varepsilon^2)$ gives $\frac{1}{d} \|\text{mSketch}(X)\|_2^2 \leq 2 \|X\|_1$ with probability at least $9/10$ (taken over all possible choices of $\alpha_i$ and $\xi$).

Proof. Follows immediately as a corollary of Lemma 33 and the properties of the embedding $\nu$. In more detail, the following holds with probability at least $9/10$:

$$\frac{1}{d} \|\text{mSketch}(X)\|_2^2 = \frac{1}{d} \|M \nu(X')\|_2^2 = \frac{1}{d} \|M \nu(X)\|_2^2 = \frac{1}{d} \|\text{eSketch}(\nu(X))\|_2^2 \leq \|\nu(X)\|_2^2 = \|X\|_1.$$
Lemma 42. Given a text $T$, there is a streaming algorithm that for every position $i$ outputs the Manhattan suffix sketch of a word $T[b \cdot k + 1, i]$, where $k$ is the smallest integer such that $i - b \cdot k \leq n$. The algorithm takes $O(d \cdot (n/b) + \log^2 \sigma)$ space, and $O(d(1 + n/b^2) \cdot \log^3(bd\sigma))$ time per character.

3.3 Generic ($L_p$) distance for $0 < p \leq 2$

For generic $L_p$ distances, we use the approach of [26] based on $p$-stable distributions.

Corollary 43. Given a text $T$, there is a streaming algorithm that for every position $i$ outputs the $L_p$ suffix sketch of a word $T[b \cdot k + 1, i]$, where $k$ is the smallest integer such that $i - b \cdot k \leq n$. The algorithm takes $O(\varepsilon^{-p}(n/b) \cdot \log(\sigma n/\varepsilon) \log(n/\varepsilon))$ bits of space and $O(\varepsilon^{-p}(n/b) \log(n))$ time per character.

Proof. We start a new instance of the sketching algorithm of Theorem 26 at every block border and continue running it for the next $[n/b]$ blocks. At each moment, there are $O(n/b)$ active instances of the algorithm. The bounds follow.

4 Proof of Theorem 3

Recall the structure of the algorithms. During the preprocessing, we compute the suffix sketches of suffixes $P[1, n], P[2, n], \ldots, P[b, n]$ of $P$. During the main stage, the text is processed by blocks of length $b$. To compute an approximation of the distance / the $p$'th moment at a particular alignment, we divide the pattern into two parts: a prefix of length at most $b$, and the remaining suffix. We compute an approximation of the distance / the $p$'th moment for both of the parts and sum them up to obtain the final answer. To compute an approximation of the distance / the $p$'th moment between the prefix and the corresponding block of the text, we compute, while reading each block of the text, its prefix encoding, and to compute an approximation of the distance / the $p$'th moment between the suffix and the text, we use the suffix sketches.

1) Hamming ($L_0$) distance. When we receive a new block of the text, we compute its Hamming prefix encoding using the algorithm of Lemma 10 in $O(b)$ space. We de-amortize the computation over the subsequent block and spend $\tilde{O}(\varepsilon^{-2})$ time per character. We store the resulting encoding for the next $O(n/b)$ blocks. In total, the encodings require $\tilde{O}(\varepsilon^{-2}n/b)$ space. The Hamming suffix sketches of $P[1, n], P[2, n], \ldots, P[b, n]$ occupy $O(\varepsilon^{-2}b)$ space. The algorithm of Lemma 36 that computes the suffix sketches takes $O(\varepsilon^{-2}b + \log(\varepsilon^{-2}n))$ space and $O(\varepsilon^{-2}(1 + n/b^2))$ time per character. Consider a block starting with position $p$. To compute the Hamming distances between $n$-length subwords that end in this block and the pattern, we apply the following approach. First, while reading the block preceding the current one, we decode the Hamming prefix encoding of the block that starts at position $p - n$ using Lemma 11. We de-amortize the algorithm to spend $\tilde{O}(\varepsilon^{-2})$ time per character. Hence, at the position $i$, we know the $(1 \pm \varepsilon)$-approximation between the prefixes of the pattern and the corresponding subwords of the text. At each position, we can compute the Hamming distance between the corresponding suffix of the pattern and the text in $\tilde{O}(\varepsilon^{-2})$ time using the Hamming suffix sketch. By taking $b = \sqrt{n}$, we obtain the claim.

2) Manhattan ($L_1$) distance. We proceed analogously to the Hamming distance case. The Manhattan prefix encoding of each block is computed using Lemma 16, in $\tilde{O}(b)$ time per character. We store the resulting encoding for the next $O(n/b)$ blocks, giving
in total $\tilde{O}(\varepsilon^{-2}n/b)$ space. The Manhattan suffix sketches of $P[1, n], P[2, n], \ldots, P[b, n]$ occupy $O(\varepsilon^{-2}b)$ space. Algorithm of Lemma 42 takes $\tilde{O}(\varepsilon^{-2}(b + \sqrt{n}b + \log^2 \sigma)$ space and $\tilde{O}(\varepsilon^{-2}(1 + \sqrt{n}b^2))$ time per character. For decoding the prefix encoding we use Lemma 17, spending $\tilde{O}(b)$ time per character. Once again we take $b = \sqrt{n}$, and assume w.l.o.g. $\varepsilon^{-1} \leq \sqrt{n}$ (as otherwise we can use a naive algorithm with $O(n)$ space and $O(n)$ time per character).

3) **Generic ($L_p$) distance for $0 < p < 1$.** The $L_p$ prefix encodings of the blocks are computed using Lemma 23, using $\tilde{O}(t \cdot b)$ time per character. We store the resulting encodings for the next $O(n/b)$ blocks, giving in total $\tilde{O}(\varepsilon^{-2}n/b)$ space. The $L_p$ suffix sketches of $P[1, n], P[2, n], \ldots, P[b, n]$ occupy $\tilde{O}(\varepsilon^{-2}b \log \sigma)$ space. Algorithm of Corollary 43 computes the $L_p$ suffix sketches for the text in $\tilde{O}(\varepsilon^{-2}(n/b) \log \sigma)$ space and $\tilde{O}(\varepsilon^{-2}n/b)$ time per character. For decoding the prefix encoding we use Lemma 24, spending $\tilde{O}(t \cdot b)$ time per character. We take $b = \sqrt{n}$, and substitute $t$ accordingly to Theorem 18.

4) **Generic ($L_p$) distance for $1 < p < 2$.** Note that for $\varepsilon < 1/n$ we can use a naive algorithm, that is to store $S$ itself in $O(n)$ space. The update takes constant time, and computing the $L_p$ norm takes $O(n)$ time which is better than the guarantees of the theorem for such values of $\varepsilon$. For $\varepsilon \geq 1/n$, the algorithm of Lemma 30 computes the $L_p$ prefix encodings of the blocks in $\tilde{O}(b + \varepsilon^{-2}p \log^2 \sigma)$ space and $\tilde{O}(b + \varepsilon^{-2} \log \sigma)$ time per character. The encodings occupy $\tilde{O}(\varepsilon^{-2}p(n/b) \log^2 \sigma)$ space. The $L_p$ suffix sketches of $P[1, n], P[2, n], \ldots, P[b, n]$ occupy $\tilde{O}(\varepsilon^{-2}b \log \sigma)$ space. Algorithm of Corollary 43 computes the $L_p$ suffix sketches for the text in $\tilde{O}(\varepsilon^{-2}(n/b) \log \sigma)$ space and $\tilde{O}(\varepsilon^{-2}n/b)$ time per character. Taking $b = \varepsilon^{-p/2}\sqrt{n}$ and assuming w.l.o.g. $\varepsilon^{-1} < \sqrt{n}$, we obtain the claim.

5) **Conclusion**

We pose several open questions. First is whether the time-complexity for $1/2 < p < 1$ can be improved to not involve any dependency on $\sigma$. For this we need a better technique than bounding variance of the embedding into Hamming distance: in our technique, the tail gets "too heavy". Another pressing question is whether for all values of $p > 0$ we could improve upon $\sqrt{n}$ time per character. We also remark that it seems unlikely that an embedding to Hamming space could be used to reduce space complexity for $p > 1$: $L_p^p$ does not admit the triangle inequality while the Hamming distance does, and the $L_p$ distance is not additive with respect to concatenation, while the Hamming distance is.

References

1. Karl R. Abrahamson. Generalized string matching. *SIAM J. Comput.*, 16(6):1039–1051, 1987.
2. Dimitris Achlioptas. Database-friendly random projections: Johnson–Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003. doi:10.1016/S0022-0000(03)00025-4.
3. Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, D. Sivakumar, and Luca Trevisan. Counting distinct elements in a data stream. In *RANDOM 2002*, pages 1–10, 2002. doi:10.1007/3-540-45726-7_1.
4. Vladimir Braverman, Ran Gelles, and Rafail Ostrovsky. How to catch $L_2$-heavy-hitters on sliding windows. *Theor. Comput. Sci.*, 554:82–94, 2014.
5. Vladimir Braverman and Rafail Ostrovsky. Smooth histograms for sliding windows. In *FOCS 2007*, pages 283–293, 2007.
6. Vladimir Braverman and Rafail Ostrovsky. Effective computations on sliding windows. *SIAM J. Comput.*, 39(6):2113–2131, 2010.
**Lemma 4.** Let $2 \leq 1/\varepsilon < n$ and $0 \leq p \leq 2$. Any $(1 \pm \varepsilon)$-approximation algorithm that computes the $L_p$ distance between a pattern and a streaming text for each alignment, must use $\Omega(\min(1/\varepsilon^2, n))$ bits of space.

**Proof.** Let us first show the lower bound for $p = 0$, i.e., for Hamming distance. We show the lower bound by reduction to a two-party communication complexity problem called GAP-Hamming-distance. In this problem, the two parties, Alice and Bob are given two binary words of length $n$ and a parameter $g = \varepsilon n$, $1 \leq g \leq n/2$. Alice sends Bob a message, and Bob’s task is to output 1 if the Hamming distance between his and Alice’s word is larger than $n/2 + g$, and zero if it is at most $n/2 - g$. Otherwise, he can output “don’t know”. By Proposition 4.4 [10], the communication complexity of this problem is $\Omega(\min(1/\varepsilon^2, n))$.

We can now show a space lower bound for any $(1 \pm \varepsilon)$-approximate algorithm for computing the Hamming distance between the pattern and the text by a standard reduction. Suppose that $2 \leq 1/\varepsilon \leq n$ there is an algorithm that uses $o(\min(1/\varepsilon^2, n))$ bits of space. Let $P$ be Alice’s word, $T$ Bob’s word. After reading $P$, the algorithm stores all the information about it in $o(\min(1/\varepsilon^2, n))$ bits of space. We construct the communication protocol as follows: Alice sends the information about $P$ to Bob. Using it, Bob can continue running the algorithm and compute the approximation of the Hamming distance between $P$ and $T$. We have thus developed a communication protocol with complexity $o(\min(1/\varepsilon^2, n))$, a contradiction.

We can now show the lower bound for $0 < p \leq 2$. We immediately obtain a space lower bound for any $(1 \pm \varepsilon)$-approximate algorithm for computing the $p$’th moment between the pattern and the text at every alignment. Indeed, on binary words the $p$’th moment is equal to the Hamming distance for all $0 < p \leq 2$. The lower bound for the $L_p$ distance follows by Observation 5.
Lemma 8. For $Z_f$ as in Algorithm 7 there is $Z_f \overset{d}{=} \|U - V\|_H$ with probability at least $3/4$.

Proof. Denote $m = \|U - V\|_H$. Consider a fixed value $r$. Let $I_1, I_2, \ldots, I_n$ be binary variables indicating existence of a mismatch between $h_{\text{Sub}}(U)$ and $V$ at positions $1, \ldots, n$, so that $X_r = \sum_j f_j$. We observe that $\mathbb{E}[X_r] = m/2^r$ and therefore $\mathbb{E}[Z_r] = m$, because each of the $m$ positions with mismatch between $U$ and $V$ generates a mismatch between $h_{\text{Sub}}(U)$ and $V$ with probability $1/2^r$.

Furthermore, as the function $h$ in Definition 6 is drawn from a pairwise independent family, there is $\text{Var}[X_r] = \sum_j \text{Var}[I_j] \leq \sum_j \mathbb{E}[(I_j)^2] = \sum_j \mathbb{E}[I_j] = \mathbb{E}[X_r] = m/2^r$. Let $c = \min\{i : \mathbb{E}[X_i] \leq k\} = \lceil \log_2 \left( \frac{k}{m} \right) \rceil$. By Chebyshev's inequality, we have

$$\Pr[|Z_r - m| \geq 4\sqrt{m^2(e + 1)}] \geq \frac{2^{2+(r+1-c)/2}}{m^2/2^r} \leq 1/2^{4+(c+1-r)} \quad (2)$$

We estimate $\Pr[f > c + 1] = \Pr[X_{c+1} > k]$. Assume w.l.o.g. that $k \geq 32$. Observe that $m/2^e \leq k$, which implies, for $k \geq 32$, $m/2^{e+1} + 4\sqrt{m/2^{e+1}} \leq k/2 + 4\sqrt{k/2} \leq k$. By Equation 2, there is

$$\Pr[X_{c+1} > k] \leq \Pr[X_{c+1} \geq m/2^{e+1} + 4\sqrt{m/2^{e+1}}] \leq 1/16.$$

It follows that $\Pr[f > c + 1] = \Pr[X_{c+1} > k] \leq 1/16$. Hence, we obtain

$$\Pr[|Z_f - m| \geq 4\sqrt{2/k \cdot m}] \leq \Pr[|Z_f - m| \geq 4\sqrt{m^{2^{e+1}}}] \leq \Pr[f > c + 1] + \sum_{r=0}^{c+1} \Pr[|Z_f - m| \geq 4\sqrt{m^{2^{e+1}}} \text{ and } f = r] \leq \Pr[f > c + 1] + \sum_{r=0}^{c+1} \Pr[|Z_r - m| \geq 4\sqrt{m^{2^{e+1}}}] \leq 1/16 + \sum_{r=1}^{c+1} 1/2^{4+(c+1-r)} < 1/4.$$

It follows that we can choose $k = \Theta(1/e^2)$ large enough so that $Z_f \overset{d}{=} \|U - V\|_H$ with probability $\geq 3/4$.

Lemma 14. For $Z_f$ as in Algorithm 13 there is $Z_f \overset{d}{=} \|U - V\|_1$ with probability $\geq 3/4$.

Proof. Take some position $i$ and denote for short $a = m_{\text{Sub}}(U)[i]$ and $b = m_{\text{Sub}}(V)[i]$ and $c = \frac{U[i] - V[i]}{2}$. There is $|a - b| \in \{ |c|, |\lfloor c \rfloor| \}$ and $\mathbb{E}[|a - b|] = |c|$. Since $|a - b| - |\lfloor c \rfloor|$ is a 0/1 variable, there is $\text{Var}[|a - b|] = \text{Var}[|(a - b) - |\lfloor c \rfloor||] \leq \mathbb{E}[|(a - b) - |\lfloor c \rfloor||] = \mathbb{E}[|a - b|]$. Summing for all values of $i$, we reach that

$$\text{Var}[X_r] = \text{Var}[|m_{\text{Sub}}(U) - m_{\text{Sub}}(V)|] \leq \mathbb{E}[|m_{\text{Sub}}(U) - m_{\text{Sub}}(V)|_1] = \mathbb{E}[X_r].$$

Since we have reached an identical variance bound, the proof follows step-by-step the proof of Lemma 8.

Theorem 18. Given $0 < p < 1$ and $\varepsilon > 0$ there is a word morphism $\varphi : c \in \Sigma \rightarrow c_0^{d_0} c_1^{d_1} \ldots c_{d-1}^{d_{d-1}}$ such that:

1) $t = \tilde{O}(\varepsilon^{-2})$ when $0 < p < 1/2$, $t = \tilde{O}(\varepsilon^{-3})$ when $p = 1/2$ and $t = \tilde{O}(\sigma \varepsilon^{-1}/\varepsilon^{2+3\varepsilon^{-1}})$ when $1/2 < p < 1$.

2) values of $t$ and $d_0, \ldots, d_{d-1}$ do not depend on $c$. 

\text{APPROX/RANDOM 2020}
3) there exists a constant \( \alpha = \alpha(p, \varepsilon) \) such that for any two words \( U, V \) of length at most \( n \), we have \( \|U - V\|_p^p \leq \alpha \cdot \|\varphi(U) - \varphi(V)\|_H \) with probability at least \( 9/10 \),

4) it is enough for the randomness to be realized by a hash function \( r : [t] \to [D] \) from a

\( 4 \)-independent hash function family for some \( D = \text{poly}(n\sigma^{-1}) \), which can be generated from a \( \tilde{O}(\log \sigma) \) bits size seed.

**Proof.** We will consider three cases: \( 0 < p < 1/2 \), \( p = 1/2 \), and \( 1/2 < p < 1 \).

**Case \( 0 < p < 1/2 \).** Our plan is to build upon the scheme highlighted earlier in this section. Specifically, we preserve the values of \( c_i \).

Consider a pair of characters \( c, c' \). First, note that \( \mathbb{E}[x] \) is an increasing function of \( |c - c'| \).
From this and Equation 1 we obtain that

\[
\mathbb{E}[x] \leq |c - c'|^{p-1} (1 - (1 + \varepsilon)^{p-1}) \leq \frac{\varepsilon^{-1} \sigma}{(1 - (1 + \varepsilon)^{p-1}) \sigma^e^{-1}} = \Theta(\varepsilon).
\]

We also round \( d_i \) down to the nearest integer, which introduces an additional factor \( 1 \pm \Theta(\varepsilon) \) relative error, since \( \forall d_i \geq \varepsilon^{-1} \) Finally, we set \( \varphi(c) = c_0 \ldots c_{q-1} \). We then have

\[
\mathbb{E}[\|\varphi(c) - \varphi(c')\|_H] = \Theta(\varepsilon^2 |c - c'| |p - 1| \sigma^e^{-1}).
\]

To guarantee that the equality holds with probability at least \( 9/10 \) and not just in expectation, we repeat the scheme several times, with independent random seeds. That is, consider morphisms \( \varphi_1(c), \varphi_2(c), \ldots, \varphi_s(c) \) and define a morphism \( \varphi(c) = \varphi_1(c) \varphi_2(c) \ldots \varphi_s(c) \) with property:

\[
\mathbb{E}[\|\varphi(c) - \varphi(c')\|_H] = \Theta(\varepsilon^{-2} |c - c'| |p - 1| \sigma^e^{-1}).
\]

Assume w.l.o.g. that \( (1 + \varepsilon)^t < |c - c'| \leq (1 + \varepsilon)^t \). We proceed to bound

\[
\text{Var}[\|\varphi(c) - \varphi(c')\|_H] \leq \sum_{i=\ell+1}^q (d_i)^2 \cdot \text{Pr}[c_i \neq c'_i] \leq \sum_{i=\ell+1}^q \varepsilon^{-2} |c - c'| (1 + \varepsilon)^t \leq \varepsilon^{-2} |c - c'| \sum_{i=\ell+1}^\infty (1 + \varepsilon)^{2p-1} \leq \varepsilon^{-2} |c - c'| \frac{(1 + \varepsilon)^{2p-1}}{1 - (1 + \varepsilon)^{2p-1}} = \Theta(\varepsilon^{-2} |c - c'|^2 \sigma^e^{-3}).
\]

We set \( s = \Theta\left(\frac{|c - c'|^{2p-3}(p-1)^2}{\varepsilon^{-1} \sigma^e^{-3}}\right) = \Theta(\varepsilon^{-1} \frac{1}{1 - 2p}) \) for the claim to hold via Chebyshev’s inequality. The error probability coming from Chebyshev’s inequality can be made arbitrarily small constant by fixing the constant factor in \( s \) to be large enough. We finally set \( t = sq \).
Case $p = 1/2$. Note that for $p, p'$ such that $|p - p'| \leq \log_2(1 + \varepsilon)$ we have $|x|^p \overset{\Delta}{=} |x|^{p'}$ for all $-\sigma \leq x \leq \sigma$. We can therefore reduce this case to $p = 1/2 - \log_2(1 + \varepsilon)$. However, we have to take into account that the asymptotic growth of $t$ hides $1/(1 - 2p)$ dependency on $p$ for $0 < p < 1/2$, hence $t = O(\varepsilon^{-3})$ for $p = 1/2$.

Case $1/2 < p < 1$. The proof follows the steps of the case $0 < p < 1/2$. We first bound the variance:

$$\text{Var}[\|\varphi(x) - \varphi(x')\|_2] \leq s \cdot \sum_{i=1}^{q} (d_i)^2 \cdot \Pr[e_i \neq e'_i] =$$

$$= s \cdot \varepsilon^{-2}|c - c'| \sum_{i=1}^{q} ((1 + \varepsilon)^{2p-1})^i =$$

$$= s \cdot O(\varepsilon^{-3}|c - c'|((1 + \varepsilon)^{2p-1}) =$$

$$= s \cdot O(\varepsilon^{-3}|c - c'|2p-1 / \varepsilon^3 2p+1 / (2p+1)).$$

We set $s = \Theta(\varepsilon^{-1}|c - c'|2p+1 / \varepsilon^3 2p+1 / (2p+1)) = O(\varepsilon^{-1}|c - c'|2p+1 / \varepsilon^3 2p+1 / (2p+1))$, so that by Chebyshev’s inequality, the probability of obtaining $\|U - V\|_p^2 \overset{\Delta}{=} \alpha \cdot \|\varphi(U) - \varphi(V)\|_H$ is an arbitrarily small constant (by setting $s$ to be large enough).

Randomness. The only source of randomness in the description are the values $r_i \in [0, 1]$ picked u.a.r. and independently. We note that the values $r_i$ can be picked instead as a finite precision floating-point numbers. Since all the values we are working with are bounded by $\text{poly}(\varepsilon, n, \sigma, \sigma)$, it is enough to set precision accordingly. We also observe that our concentration argument involves only Chebyshev’s inequality and thus only the variance and the expected value, so it suffices to require that $r_i$ be 4-wise independent.

Lemma 21. For $Z_f$ as in Algorithm 20 there is $Z_f \overset{\Delta}{=} \|S - Q\|_H$ with probability $\geq 3/4$.

Proof. Consider a fixed subsampling level $r$. For simplicity, let $\text{Sub}_r(S) = s_1 r_1^2 \ldots s_m r_m^2$ and $\text{Sub}_r(Q) = q_1 r_1^2 \ldots q_m r_m^2$. Define a random variable $x_i$ to be the contribution of $s_i r_i^2, q_i r_i^2$ to the Hamming distance $X_r$, i.e.

$$x_i = \|s_i r_i^2 - q_i r_i^2\|_H = e_i \cdot \|s_i - q_i\|_H.$$ 

Since $e_i \in \{[c_i/2^r], [c_i/2^r]\}$ and $\mathbb{E}[e_i] = e_i/2^r$, we have $\mathbb{E}[x_i] = e_i \cdot \|s_i - q_i\|_H$ and

$$\text{Var}[x_i] = \text{Var}[x_i - [e_i/2^r]] \leq \mathbb{E}[x_i - [e_i/2^r]] \leq \mathbb{E}[x_i].$$

Summing over all values of $i$, we reach $\mathbb{E}[X_r] = \|S - Q\|_H$ and $\text{Var}[X_r] \leq \mathbb{E}[X_r]$. These bounds are identical to that of Lemma 8 and we can proceed in a similar fashion to obtain the claim.

Lemma 28. Let $1 < p \leq 2$. For any two vectors $X, Y$ of equal length, $\|X + Y\|_p^p - \|X\|_p^p = \mathcal{O}(\|Y\|_p^p + \|Y\|_p \cdot \|X\|_p^{p-1})$.

Proof. Consider $x, y \in \mathbb{R}$. If $|x| \geq |y|$, then by Taylor expansion, $|x + y|^p = |x|^p (1 + O(|y|/|x|)) = |x|^p (1 + \mathcal{O}(|y|/|x|^p)) = |x|^p + \mathcal{O}(|y|/|x|^p - 1)$. If $|x| < |y|$, then $|x + y|^p = \mathcal{O}(|y|^p)$. Thus for any real values, we have

$$|x + y|^p = |x|^p + \mathcal{O}(|y|^p + |y| \cdot |x|^p - 1).$$
Denote \( X = [x_1, x_2, \ldots, x_n]^T \) and \( Y = [y_1, y_2, \ldots, y_n]^T \). There is

\[
\|X + Y\|_p^p = \sum_i |x_i + y_i|^p = \sum_i |x_i|^p + O \left( \sum_i |y_i|^p + \sum_i |y_i||x_i|^{p-1} \right).
\]

Pick \( q = p/(p - 1) \) so that \( 1/p + 1/q = 1 \). By Hölder’s inequality:

\[
\sum_i |y_i||x_i|^{p-1} \leq \left( \sum_i |y_i|^p \right)^{1/p} \left( \sum_i |x_i|^{(p-1)q} \right)^{1/q} = \|Y\|_p \|X\|_p^{1-1/p}.
\]

\[\triangleright\text{Lemma 36.} \text{ Given a text } T, \text{ there is a streaming algorithm that for every position } i \text{ outputs the Hamming suffix sketch of a word } T[b \cdot k + 1, i], \text{ where } k \text{ is the largest integer such that } i - b \cdot k \leq n. \text{ The algorithm takes } O(dn/b + \log d\sigma n) \text{ space and } O(d(1 + n/b^2)) \text{ time per character.} \]

\textbf{Proof.} We fix the matrix \( R \) and the random coefficients \( \alpha_1, \ldots, \alpha_{n/b} \) from Definition 34. We do not store \( R \) and \( \alpha_i \) explicitly, but generate them using two hash functions drawn at random from a 4-wise independent family. For example, to generate \( R \) we can consider a family of polynomials \( 2((a x^3 + b x^2 + c x + d \mod p) \mod 2) - 1 \), with parameters \( a, b, c, d \) chosen i.i.d. from the prime field \( \mathbb{F}_p \) for \( p \geq db \), and \( \alpha_i \) can be generated in a similar fashion. This way, we need to store only \( O(\log(d\sigma b) + \log(n/b)) = O(\log d\sigma n) \) random bits that define the coefficients of two polynomials to generate \( R \) and \( \alpha_i \).

We process the text \( T \) by blocks \( B_1, B_2, \ldots \) of length \( b \). For each block \( B_k \) we compute its sketch using the matrix \( R \). That is, at the beginning of each block we initialize its sketch as a zero vector of length \( d \). When a new character \( T[i] \) of a block \( B_k \) arrives, we compute and add \([M[1, i \cdot b + T[i]], \ldots, M[|d, i \cdot b + T[i]]]^T\) to the sketch, which takes \( O(d) \) time. We store the sketch of \( B_k \) until the block \( B_{k+[n/b]} \) and use it to compute the suffix sketches for the positions in this block.

Consider now a block \( B_{k+[n/b]} \). We first compute the suffix sketch for the position \( b \cdot (k + [n/b]) \), which is the position preceding the block \( B_{k+[n/b]} \). The suffix sketch for it is simply a linear combination of the sketches of the blocks \( B_{k+[n/b]-1}, B_{k+[n/b]-2}, \ldots, B_k \) with coefficients \( \alpha_1, \ldots, \alpha_{n/b}-1 \). Since each sketch is a vector of length \( d \), we can compute the linear combination in \( O(dn/b) \) time. To make this computation time-efficient, we start it \( b \) positions before position \( b \cdot (k + [n/b]) \) arrives, and de-amortise the computation over these \( b \) positions. This way, we use only \( O(dn/b^2) \) time per character.

Now, using the suffix sketch for the position \( b \cdot (k + [n/b]) \), we can compute the suffix sketches for all positions in the block \( B_{k+[n/b]} \) one-by-one, using only \( O(d) \) time per character: When a new character \( T[i] \) arrives, we add \([\alpha_{n/b} M[1, i \cdot b + T[i]], \alpha_{n/b} M[2, i \cdot b + T[i]], \ldots, \alpha_{n/b} M[|d, i \cdot b + T[i]]]^T\) to the suffix sketch to update it.

Note that at any time we store \( O(n/b) \) sketches of the blocks, so the algorithm uses \( O(dn/b + \log d\sigma n) \) space in total.

\[\triangleright\text{Lemma 42.} \text{ Given a text } T, \text{ there is a streaming algorithm that for every position } i \text{ outputs the Manhattan suffix sketch of a word } T[b \cdot k + 1, i], \text{ where } k \text{ is the smallest integer such that } i - b \cdot k \leq n. \text{ The algorithm takes } O(d \cdot (n/b) + \log^2 \sigma) \text{ space, and } O(d(1 + n/b^2) \cdot \log^3 (bd\sigma)) \text{ time per character.} \]

\textbf{Proof.} The proof mirrors the proof of Lemma 36, and we describe the key elements. We fix the random coefficients \( \alpha_1, \ldots, \alpha_{n/b} \) and the hash function \( h \) from Definition 40. As previously, we do not store the coefficients \( \alpha_i \) explicitly, but generate them using a hash
function drawn at random from a 4-wise independent family. The matrix $R$ is already defined by $h$, with the following parameters: it requires $O(\log^2(bd\sigma))$ bits of seed, and range-sum queries are answered in time $O(\log^3(bd\sigma))$.

In the sketching of blocks, we proceed in the same manner, except that when a new character $T[i]$ of a block $B_k$ arrives, we compute and add $\sum_{0 \leq j < T[i]} [M[1, i \cdot b\sigma + j], \ldots, M[d, i \cdot b\sigma + j]]^T = \alpha_{i, \lfloor n/b \rfloor} \cdot [g(b\sigma, b\sigma + T[i]; \xi), g(2b\sigma, 2b\sigma + T[i]; \xi), \ldots, g(d \cdot b\sigma, d \cdot b\sigma + T[i]; \xi)]^T$ to the sketch, which takes $O(d \cdot \log^3(bd\sigma))$ time (log$^3(bd\sigma)$ times slower as the corresponding step in Lemma 36).

Consider now a block $B_{k + \lfloor n/b \rfloor}$. When a new character $T[i]$ arrives, we update the suffix sketch by adding $\alpha_{\lfloor n/b \rfloor} \cdot [g(b\sigma, b\sigma + T[i]; \xi), g(2b\sigma, 2b\sigma + T[i]; \xi), \ldots, g(d \cdot b\sigma, d \cdot b\sigma + T[i]; \xi)]^T$ to it.

All of the operations are $O(\log^3(bd\sigma))$ time slower than the corresponding steps in Lemma 36, and the memory complexity is increased by the seed size $O(\log^2(bd\sigma))$ term ($\log^2 b$ and $\log^2 d$ terms get absorbed).