The Computational Hardness of Estimating Edit Distance*

Alexandr Andoni[†]
MIT
andoni@mit.edu

Robert Krauthgamer[‡]
The Weizmann Institute of Sciences
robert.krauthgamer@weizmann.ac.il

July 3, 2009

Abstract

We prove the first non-trivial communication complexity lower bound for the problem of estimating the edit distance (aka Levenshtein distance) between two strings. To the best of our knowledge, this is the first computational setting in which the complexity of estimating the edit distance is provably larger than that of Hamming distance.

Our lower bound exhibits a trade-off between approximation and communication, asserting, for example, that protocols with $O(1)$ bits of communication can only obtain approximation $\alpha \geq \Omega(\log d / \log\log d)$, where d is the length of the input strings. This case of $O(1)$ communication is of particular importance since it captures constant-size sketches as well as embeddings into spaces like L_1 and squared-L_2, two prevailing algorithmic approaches for dealing with edit distance. Furthermore, the bound holds not only for strings over the binary alphabet, but also for strings that are permutations (aka the Ulam metric). Currently, all known non-trivial communication upper bounds are derived from embeddings into L_1.

Besides being applicable to a much richer class of algorithms than all previous results, our lower bounds are near-tight in at least one case, namely of embedding permutations into L_1. The proof uses a new technique, that relies on Fourier analysis in a rather elementary way.

1 Introduction

The *edit distance* (aka *Levenshtein distance*) between two strings is the number of insertions, deletions, and substitutions needed to transform one string into the other. This distance is of key importance in several fields, such as computational biology and text processing, and consequently computational problems involving the edit distance were studied quite extensively. The most basic problem is that of computing the edit distance between two strings of length d over alphabet Σ. The fastest algorithm known for the case of constant-size alphabet remains the algorithm of Masek and Paterson [MP80] from 1980, that runs in time $O(d^2/\log^2 d)$. Unfortunately, such near-quadratic time is prohibitive when working on large datasets, which is common in areas such as computational biology. A possible approach is to trade accuracy for speed and employ faster algorithms that compute the edit distance approximately (possibly as a preliminary filtering step). Currently, the

*A preliminary version appeared in FOCS'07.

†Part of this work was done while the author was visiting IBM Almaden. Supported in part by NSF CAREER Grant #0133849 "Approximate Algorithms for High-dimensional Geometric Problems".

‡Part of this work was done while the author was at IBM Almaden. Work supported in part by a grant from the Fusfeld Research Fund.

best near-linear time algorithm, due to Andoni and Onak [AO09], achieves approximation factor of $2^{\tilde{O}(\sqrt{\log d})}$, improving over earlier results of [BEK$^+$03, BJKK04, BES06].

Another major algorithmic challenge is to design a scheme for Nearest Neighbor Search (NNS) under the edit distance. In this problem, we wish to design a data structure that preprocesses a dataset of n strings of length d each, so that when a query string is given, the query's nearest neighbor (i.e., a dataset string with the smallest edit distance to the query string) can be reported quickly. However, no efficient solutions for this problem are known, even if one allows a small approximation. All known algorithms with fast query time (polynomial in d and $\log n$) either require large space or have large approximation error — Indyk [Ind04] achieves constant approximation using $n^{d^{\Omega(1)}}$ space, and Ostrovsky and Rabani [OR07] obtain $2^{O(\sqrt{\log d \log\log d})}$ approximation using space that is polynomial in d and n.

It is thus natural to ask: is it really "hard" to design algorithms for the edit distance? A natural benchmark is the Hamming distance, which is equal to the number of positions where the two strings differ. Hamming distance can be seen as edit distance where the only operations allowed are substitution. For Hamming distance, much better algorithms are known: (i) the distance between two strings can clearly be computed in $O(d)$ time, and (ii) NNS schemes by Indyk and Motwani [IM98] and by Kushilevitz, Ostrovsky and Rabani [KOR00] achieve $1+\epsilon$ approximation using space that is polynomial in d and in n^{1/ϵ^2}. Empirically, edit distance appears to be more difficult than Hamming distance, and the reason is quite clear — insertions and deletions cause portions of the string to move and create an alignment problem — but there is no rigorous evidence that supports this intuition. In particular, we are not aware of a computational model in which the complexity of approximating edit distance is provably larger than that of Hamming distance.[1]

We give the first rigorous evidence for the *computational* hardness of approximating the edit distance. In fact, we show a computational model in which the complexity of estimating edit distance is significantly larger than that of Hamming distance, and this is the first setting where such a separation is known. Our results hold for two important metrics:

1. *standard edit metric*, i.e. edit distance on $\{0,1\}^d$;
2. the *Ulam metric*, which is the edit distance on permutations of length d.

Here and throughout, a *permutation* is a string consisting of distinct characters coming from a large alphabet, $|\Sigma| \geq d$. This definition of permutations is non-standard but it is more convenient for us technically, and our results extend to hold under the standard one (where $|\Sigma| = d$), see Fact 2.4. Our results immediately imply lower bounds for sketching algorithms and for metric embeddings. These two algorithmic techniques lately received a lot of attention as promising approaches to many metric problems. We will discuss these implications in more detail after stating our main results.

[1]We are aware of only two results that come close. First, if the operations on the symbols of the strings are restricted to tests of equality, then computing edit distance between two strings over a large alphabet requires $\Omega(d^2)$ comparisons [WC76]. However, this lower bound holds only for exact computation (or $1+o(1)$ approximation) and for strings over a large alphabet (but not for binary strings). In fact, the lower bound breaks down even in the comparison model (when we can compare the relative order of two symbols): e.g., the algorithm of [BFC08] runs in time $O(d^2 \frac{\log^2 \log d}{\log^2 d})$ for computing edit distance between strings over arbitrarily large alphabet.

Second, if we restrict attention to *sublinear*-time, i.e., algorithms that probe only a small part of the two input strings, then there exists a simple separation in terms of query complexity. Specifically, deciding whether the edit distance is $\Omega(d)$ or $O(d^{1-\epsilon})$ requires reading at least $\Omega(d^{1/2-\epsilon/2})$ positions of the strings [BEK$^+$03], while the same decision under Hamming distance is achieved easily by sampling $O(1)$ positions. This separation has limited computational implications since it essentially shows that estimating edit distance requires reading "many" positions of the input strings.

1.1 Main Results

Our main result is stated in terms of communication complexity of the *distance threshold estimation problem (DTEP)* and holds for both edit metric over $\Sigma = \{0,1\}$, and for the Ulam metric. In DTEP [SS02], for a threshold R and an approximation $\alpha \geq 1$ fixed as parameters, we are given inputs x, y and we want to decide whether $\mathrm{ed}(x,y) > R$ or $\mathrm{ed}(x,y) \leq R/\alpha$.

In the communication protocols setting, Alice and Bob, who have access to a common source of randomness, receive strings x and y respectively as their inputs, and their goal is to solve DTEP by exchanging messages. The communication complexity of the protocol is then defined as the minimum number of bits Alice and Bob need to exchange in order to succeed with probability at least $2/3$. When x, y come from the standard edit metric, we denote the communication complexity by $\mathrm{CC}_{\alpha,R}^{\{0,1\}^d}$. Similarly, when x, y come from the Ulam metric, we denote the communication complexity by $\mathrm{CC}_{\alpha,R}^{\mathrm{Ulam}_d}$. Our main theorem provides a lower bound on the latter, exhibiting a trade-off between communication and approximation.

Theorem 1.1 (Main Theorem)**.** *There exists a constant $c > 0$, such that for every string length $d > 1$, approximation $\alpha > 1$, and R satisfying $d^{0.1} \leq R \leq d^{0.49}$,*

$$c \cdot \mathrm{CC}_{\alpha,R}^{\mathrm{Ulam}_d} + \log(\alpha \log \alpha) \geq \log \log d.$$

We extend this result from the Ulam metric to the standard edit metric by reducing the latter to the former. The key idea, which may be of independent interest, is that substituting every alphabet symbol independently with a random bit is likely to preserve the edit distance, up to a constant factor, as stated in the following theorem.

Theorem 1.2. *Let $P, Q \in \Sigma^d$ be two permutations, and let $\pi : \Sigma \mapsto \{0,1\}$ be a random function. Then*

- $\mathrm{ed}(\pi(P), \pi(Q)) \leq \mathrm{ed}(P,Q)$ *for any choice of π, and*
- $\Pr_\pi \left[\mathrm{ed}(\pi(P), \pi(Q)) \geq \Omega(1) \cdot \mathrm{ed}(P,Q) \right] \geq 1 - 2^{-\Omega(\mathrm{ed}(P,Q))}$.

Using our two theorems, we obtain the following.

Corollary 1.3. *There exists a constant $c > 0$, such that for every string length $d > 1$, approximation $\alpha > 1$, and R satisfying $d^{0.1} \leq R \leq d^{0.49}$,*

$$c \cdot \mathrm{CC}_{\alpha,R}^{\{0,1\}^d} + \log(\alpha \log \alpha) \geq \log \log d.$$

The previously known lower bounds for $\mathrm{CC}_{\alpha,R}^{\{0,1\}^d}$ and $\mathrm{CC}_{\alpha,R}^{\mathrm{Ulam}_d}$ are all obtained by a straightforward reduction from the same problem on Hamming metric. These bounds assert that the communication complexity for $\alpha = 1+\epsilon$ is $\Omega(1/\epsilon)$, and in the case of sketching (aka simultaneous)[2] protocols $\Omega(1/\epsilon^2)$ [Woo04] (see also [Woo07, Chapter 4]), and both are clearly uninformative for (say) $\alpha \geq 2$. See also [SU04] for other related results.

The only non-trivial upper bounds currently known are: (i) $\mathrm{CC}_{\alpha,R}^{\{0,1\}^d} \leq O(1)$ for suitable $\alpha = 2^{O(\sqrt{\log d \log \log d})}$; and (ii) $\mathrm{CC}_{\alpha,R}^{\mathrm{Ulam}_d} \leq O(1)$ for suitable $\alpha = O(\log d)$; and they both follow via embedding into ℓ_1. See Section 1.2 and Table 1 for more details.

[2] See the formal definition of sketching on page 6.

Comparison with Hamming distance. The next proposition, proved (implicitly) by Kushilevitz, Ostrovsky, and Rabani [KOR00], upper bounds the communication complexity of DTEP over the Hamming metric. Let $\mathrm{H}(x, y)$ be the Hamming distance between x and y.

Proposition 1.4 ([KOR00])**.** *Let $d > 1$, $R > 1$ and $\epsilon > 0$. Then there exists a communication protocol (in fact, a sketching algorithm) that given inputs $x, y \in \Sigma^d$ distinguishes whether $\mathrm{H}(x, y) > R$ or $\mathrm{H}(x, y) \leq R/(1 + \epsilon)$, using $O(1/\epsilon^2)$ bits of communication.*

Observe that for approximation factor α that is a constant (namely, independent of d and R), the complexity of the Hamming metric is $O(1)$, while that of edit metric is $\Omega(\log \log d)$. It thus follows that edit distance is indeed provably harder to compute than Hamming, in the context of communication protocols.

1.2 Implications and Related Work

Two promising approaches to designing algorithms for the edit metrics are via metric embeddings and via sketching, and our results preclude good approximation algorithms obtained via either of these approaches.

Embedding of edit distance into normed metrics. A current line of attack on edit distance is by embedding it into a computationally easier metric, for which efficient algorithms are known. An *embedding* is a mapping f from the strings into, say, ℓ_1 metric, such that for all strings x, y,

$$\mathrm{ed}(x, y) \leq \|f(x) - f(y)\|_1 \leq D \cdot \mathrm{ed}(x, y),$$

and $D \geq 1$ is called the embedding's *distortion* (approximation factor). An embedding with low distortion would have major consequences since it allows porting a host of existing algorithms for ℓ_1 metric to the case of edit distance. For example, an (efficiently computable) embedding with distortion D gives an efficient nearest neighbor data structure for approximation (say) $2D$, by applying the embedding and reverting to [IM98, KOR00].

Naturally, researchers were keen to find the least distortion for an embedding into ℓ_1 – the problem is cited in Matoušek's list of open problems [Mat07], as well as in Indyk's survey [Ind01]. Table 1 summarizes the previously known upper and lower bounds, as well as the implications of our theorems. The reader may find more background on some variants of the edit distance in [Sah08].

It is readily seen from the table that the only previous super-constant distortion lower bound is $\Omega(\log d)$ for embedding of edit distance into ℓ_1, due to Krauthgamer and Rabani [KR06], building on a technique of Khot and Naor [KN06] who gave a bound of $\Omega\left((\log d)^{1/2 - o(1)}\right)$. Although this lower bound is important, one can potentially overcome such a lower bound by, say, embedding edit distance into a richer space, such as squared-ℓ_2, a real space with squared Euclidean distance, with a possibly smaller distortion — the major implications of an embedding into squared-ℓ_2 are precisely the same as those of an embedding into ℓ_1. On this front, much weaker lower bounds were known: the previous lower bound is only 3/2 [ADG$^+$03]. To further stress how little was known, we note that one can consider even richer metrics, such as any fixed power of ℓ_2 (essentially equivalent to embedding a fixed root of edit distance into ℓ_2), which also has an efficient nearest neighbor data structure. For sufficiently high (but fixed) power of ℓ_2, even the 3/2 bound of [ADG$^+$03] gets weaker and becomes arbitrarily close to 1.

Metric	Reference	ℓ_1 embedding	squared-ℓ_2 embedding	$O(1)$-size sketch
Edit on $\{0,1\}^d$	[OR07]	$2^{O(\sqrt{\log d \log\log d})}$	$\longrightarrow$	$\longrightarrow$
	[KN06, KR06]	$\Omega(\log d)$	N/A	N/A
	[ADG+03]	$\longleftarrow$	$\geq 3/2$	N/A
	This paper	$\longleftarrow$	$\longleftarrow$	$\Omega(\frac{\log d}{\log\log d})$
Ulam (edit distance on permutations)	[CK06]	$O(\log d)$	$\longrightarrow$	$\longrightarrow$
	[Cor03]	$\longleftarrow$	$\geq 4/3$	N/A
	This paper	$\longleftarrow$	$\longleftarrow$	$\Omega(\frac{\log d}{\log\log d})$
Block edit distance	[CPSV00, MS00, Cor03]	$O(\log d \log^* d)$	$\longrightarrow$	$\longrightarrow$
Edit distance with moves	[CM07]	$O(\log d \log^* d)$	$\longrightarrow$	$\longrightarrow$

Table 1: Known bounds on distortion/approximation of embedding variants of edit distance into ℓ_1, squared-ℓ_2, and the approximation for achieving $O(1)$-size sketch. Since ℓ_1 embeds isometrically into squared-ℓ_2 and the latter has $O(1)$-size sketch for 2-approximation, the upper bounds transfer from left to right, and the lower bounds transfer from right to left (as suggested by the arrows). N/A means no result is given (even implicitly).

Our results rule out all such embeddings indirectly, by targeting a richer class of metrics — metrics for which the respective DTEP problem admits a protocol with $O(1)$ bits of communication and $O(1)$ approximation. (Proposition 1.4 shows this class of metrics is indeed richer.) It follows from our communication lower bounds that every embedding of edit distance (either on $0-1$ strings or on permutations) into a metric in that richer class must incur distortion $D \geq \Omega\left(\frac{\log d}{\log\log d}\right)$, without requiring that the embedding is efficiently computable. For completeness, we state and prove this distortion lower bound explicitly for metrics which are a fixed power of ℓ_2.

Corollary 1.5. *For every fixed $p \geq 1$, embedding the standard edit metric or the Ulam metric into $(\ell_2)^p$, the p-th power of ℓ_2, requires distortion $\Omega\left(\frac{\log d}{\log\log d}\right)$. The same is true also for embedding into ℓ_1.*

Proof. Suppose $p \geq 1$ is fixed and the edit metric ed (or similarly the Ulam metric) embeds into $(\ell_2)^p$ with distortion $D \geq 1$. In other words, the metric $\mathrm{ed}^{1/p}$ (i.e. $1/p$-power of every distance) embeds into ℓ_2 with distortion $D^{1/p}$. The DTEP problem for ℓ_2 metrics can be solved with (say) approximation $1 + \frac{1}{p}$ and communication $O(p^2)$ using Proposition 1.4 (since finite ℓ_2 metric embed isometrically into Hamming space). Together, we obtain a protocol for the DTEP problem on the metric $\mathrm{ed}^{1/p}$, which achieves approximation $D^{1/p}(1 + \frac{1}{p})$ and communication $O(p^2)$. Observe that the same protocol solves also DTEP on the edit metric ed, except that the threshold now is R^p instead of R, and the approximation is $(D^{1/p}(1+\frac{1}{p}))^p < De$. The communication is the same $O(p^2)$, and thus Corollary 1.3 (or Theorem 1.1 respectively) implies that $De \log(De) \geq 2^{-O(p^2)} \log d$, For fixed p this completes the proof. ∎

For the Ulam metric, this distortion lower bound of $\Omega\left(\frac{\log d}{\log\log d}\right)$ is near-optimal, since that metric embeds into ℓ_1 with $O(\log d)$ distortion [CK06]. The previous distortion lower bound was

4/3 [Cor03]. Other upper and lower bounds for low-distortion embeddings appear in Table 1.

Sketching of edit distance. The sketch of a string x is a (randomized) mapping of x into a short "fingerprint" $\mathbf{sk}(x)$, such that sketches of two strings, $\mathbf{sk}(x)$ and $\mathbf{sk}(y)$, are sufficient to distinguish between the case where edit distance is $\mathrm{ed}(x,y) \le R/\alpha$, and the case where $\mathrm{ed}(x,y) > R$, for fixed approximation factor $\alpha > 1$ and parameter $R > 1$. The main parameter of a sketching algorithm is its *sketch size*, the length of $\mathbf{sk}(x)$.

The sketching model can also be described as a (randomized) simultaneous communication protocol, as follows. Alice receives x and computes $\mathbf{sk}(x)$, Bob receives y and computes $\mathbf{sk}(y)$, and then they send their computed values to a "referee", who needs to decide whether x, y are close or far based only on the sketches. By letting either Alice or Bob play the role of the referee in this simultaneous protocol, one easily sees that the sketch size required by a sketching algorithm is always no smaller than the number of communication bits required by a (general) protocol. The following corollary thus follows immediately from our preceding communication lower bounds.

Corollary 1.6. *For every $d > 1$ and $d^{0.1} \le R \le d^{0.49}$, every $O(1)$-size sketching algorithm of the standard edit metric or of the Ulam metric can achieve approximation of only $\Omega\left(\frac{\log d}{\log\log d}\right)$.*

Sketching with constant sketch size can be viewed as a generalization of the "embeddings approach" presented above, by using Proposition 1.4, albeit with an arbitrarily small constant factor loss in the approximation factor. An important observation is that this more general approach suffices for the purpose of designing an NNS scheme with efficient query time (assuming that computing the sketch can be done efficiently) and with polynomial storage.[3] Indeed, the nearest neighbor data structure for Hamming metric of [KOR00] could be viewed as an instantiation of the last step. In addition, sketching can be useful for the original goal of quickly estimating the distance (e.g., as a filtering step).

The sketching model is also important as a basic computational notion for massive data sets, and in recent years, an intensive research effort has led to several sketching algorithms for DTEP over different metrics. Prior to our work, there were essentially three metrics for which a sketch size lower bounds were known: ℓ_1 [Woo04] (equivalently, for ℓ_p, $p \in (1,2]$), ℓ_∞ [SS02, BJKS04] (implying lower bounds for ℓ_p, $p > 2$), and the Earth-mover distance over $\{0,1\}^d$ [AIK08].

Sketching of edit distance was studied in [BEK$^+$03, BJKK04, OR07, CK06], but the only lower bound known for sketching of edit distance is trivial in the sense that it follows immediately from Hamming distance (by a straightforward reduction). This lower bound on the sketch size is $\Omega(1/\epsilon^2)$ for approximation $\alpha = 1+\epsilon$ [Woo04], which becomes uninformative for even a 2-approximation. In fact, Bar-Yossef et al. [BJKK04] write that "The state of affairs indicates that proving sketching lower bounds for edit distance may be quite hard."

1.3 Our Techniques

Our proof of Theorem 1.1 consists of three steps. Generally speaking, we design two input distributions: $\tilde{\mu}_0$ over "far" pairs (x,y) (i.e. $\mathrm{ed}(x,y) > R$), and $\tilde{\mu}_1$ over "close" pairs (i.e. $\mathrm{ed}(x,y) \le R/\alpha$).

[3] In particular, one can first amplify the sketching's probability of success to $1 - n^{-\Omega(1)}$, where n is the number of points in the dataset, using sketch size $O(\log n)$. Then, the data structure pre-indexes all possible sketches in the amplified protocol, using only $2^{O(\log n)} = n^{O(1)}$ space. For each possible value of the amplified sketch, the data structure stores the answer that the sketching referee would conclude from the sketch of the query and that of each dataset point. Note that, in fact, s-size sketches imply $n^{O(s)}$-size NN data structure.

The goal then becomes to show that these distributions are indistinguishable by protocols with low communication complexity. By Yao's minimax principle, it suffices to consider deterministic protocols.

The first step reduces the problem to proving that the two distributions $\tilde{\mu}_0, \tilde{\mu}_1$ are indistinguishable by boolean functions over $\mathbb{Z}_p^d$. Roughly speaking, we show that if there is a protocol using at most l bits of communication, then there exists a (deterministic) sketching protocol that uses sketch size of 1 bit and achieves an advantage of at least $\Omega(2^{-l})$ in distinguishing between the two distributions. Let $\mathcal{H}^A, \mathcal{H}^B : \mathbb{Z}_p^d \to \{-1, +1\}$ be the boolean functions that Alice and Bob, respectively, use as their sketch functions. We can then further restrict the sketching protocol so that the referee decides by checking whether $\mathcal{H}^A(x) = \mathcal{H}^B(y)$ or not. This step follows the approach employed earlier in [AIK08], with some minor technical differences.

The second step's main goal is to further characterize the advantage achieved by $\mathcal{H}^A, \mathcal{H}^B$ in terms of a carefully crafted measure of statistical distance between the two input distributions $\tilde{\mu}_0, \tilde{\mu}_1$. For this approach to be effective, it is important that the functions $\mathcal{H}^A, \mathcal{H}^B$ depend only on a few coordinates of their inputs, and in order to guarantee this (indirectly), we include in $\tilde{\mu}_0, \tilde{\mu}_1$ a noise component, which effectively destroys any dependence of $\mathcal{H}^A, \mathcal{H}^B$ on many coordinates. Specifically, this step assumes that, each distribution $\tilde{\mu}_t$, $t \in \{0, 1\}$, has the following structure: choose $x \in \mathbb{Z}_p^d$ uniformly at random, and then generate y from x via a sequence of two randomized operations. The first of the two is a noise operator with rate $\rho \in (0, 1)$, i.e., each coordinate is modified independently with probability $1 - \rho$ into a randomly chosen value. The second operation permutes the coordinates according to a permutation drawn from a distribution $\mathcal{D}_t$. Given this $\mathcal{D}_t$, consider the following derived distribution: take a vector $u \in \mathbb{Z}_p^d$ with λ non-zero positions (called a λ-*test*) and apply a random permutation $\pi \in \mathcal{D}_t$ to it; let $A_u^{(t,\lambda)}$ be the resulting distribution of vectors. (Note that the support of $A_u^{(t,\lambda)}$ contains only vectors with precisely λ non-zero entries.) Our measure Δ_λ, *called λ-test distinguishability*, is the maximum, over all such λ-tests u, of the total variation distance between $A_u^{(0,\lambda)}$ and $A_u^{(1,\lambda)}$. It pretty much captures the statistical advantage in distinguishing $\mathcal{D}_0$ from $\mathcal{D}_1$ (and thus $\tilde{\mu}_0$ from $\tilde{\mu}_1$) achievable by inspecting only λ positions of, say, y (e.g., by tracing them back to x). Altogether, our upper bound on the advantage achieved by $\mathcal{H}^A, \mathcal{H}^B$ takes roots in the following dichotomy. If $\mathcal{H}^B$ essentially depends on many coordinates of y (e.g., a linear function with many terms), then the advantage is bounded by ρ^λ (i.e., the noise destroys almost all the information), and if $\mathcal{H}^B$ essentially depends on a few, say λ, coordinates, then the advantage is bounded by the aforementioned Δ_λ. To prove this dichotomy, we rely on Fourier analysis which expands $\mathcal{H}^A, \mathcal{H}^B$ into linear functions at different levels λ.

In the third step, we complete the description of $\tilde{\mu}_0, \tilde{\mu}_1$ by detailing the construction of $\mathcal{D}_0, \mathcal{D}_1$, and give an upper bound on the λ-test distinguishability Δ_λ for these distributions. In a simplified view, each distribution $\mathcal{D}_t$ is generated by a block rotation operation, namely, choosing a random block of length L and applying to it $\epsilon_t L$ cyclic shifts. The difference between the two distributions is in the magnitude of the rotation (namely, ϵ_t).

Our use of Fourier analysis is elementary, and does not involve the KKL theorem [KKL88] or Bourgain's noise sensitivity theorem [Bou02], which were used in the previous non-embeddability results for edit distance [KN06, KR06]. We also note that our hard distribution is notably different from the distributions of [KR06] or [KN06], which do admit efficient communication protocols.

To prove Theorem 1.2, we give a new characterization of the Ulam distance between two strings. In particular, building on the work of [SU04, GJKK07], we prove that if two strings (permutations) P, Q are at distance $k = \text{ed}(P, Q)$, then there exist $\Theta(k)$ pairs of characters in P, all characters at distinct positions, such that for each pair (a, b), their order in P is opposite to that in Q (if they appear in Q at all). We then exploit this characterization by a careful counting of the number of the possible low-cost alignments between P and Q, tailored to the aforementioned $\Theta(k)$ positions.

2 Preliminaries

We use the notation $[d] = \{1, 2 \ldots d\}$ and $\mathbb{Z}_p = \{0, 1 \ldots p-1\}$. For a vector $u \in \mathbb{Z}_p^d$, define the *weight* of u, denoted $\mathrm{wt}(u)$, to be the number of coordinates in u that are non-zero.

Definition 2.1. *For matrix $A \in M_{n,n}(\mathbb{R})$ and $p \in [1, \infty]$, the* p-norm *of A is defined by* $\|A\|_p = \max\{\|Av\|_p : v \in \mathbb{C}^n, \|v\|_p = 1\}$.

2.1 Fourier Analysis over $\mathbb{Z}_p^d$

We review basic Fourier Analysis over $\mathbb{Z}_p^d$ for a prime $p \geq 2$.

The collection of functions $f : \mathbb{Z}_p^d \to \mathbb{C}$ is a vector space of dimension p^d, equipped with an inner product given by $\langle f, g \rangle = \mathbb{E}_{x \in \mathbb{Z}_p^d}\left[f(x) \cdot \overline{g(x)}\right]$. For $u \in \mathbb{Z}_p^d$, define a character $\chi_u(x) = e^{\frac{2\pi i}{p}(x \cdot u)}$, where $x \cdot u$ is the scalar product of $x, u \in \mathbb{Z}_p^d$. The set of characters $\{\chi_u \mid u \in \mathbb{Z}_p^d\}$ forms an orthonormal basis, called the Fourier basis. Thus every function $f : \mathbb{Z}_p^d \to \mathbb{C}$ admits a Fourier expansion $f = \sum_{u \in \mathbb{Z}_p^d} \hat{f}_u \chi_u$, where $\hat{f}_u = \langle f, \chi_u \rangle$ is called the Fourier coefficient of f corresponding to u. Parseval's identity states that $\mathbb{E}_{x \in \mathbb{Z}_p^d}\left[f(x)\overline{g(x)}\right] = \sum_{u \in \mathbb{Z}_p^d} \hat{f}_u \overline{\hat{g}_u}$.

We let N_ρ stand for a *noise* vector over $\mathbb{Z}_p^d$, namely, a vector where each coordinate is set independently at random as follows: with probability ρ it is set to zero, and with probability $1 - \rho$ it is set to a random value from $\mathbb{Z}_p$. We refer to ρ as the *rate* of the noise.

The *noise operator* T_ρ (also called Bonami-Beckner operator) operates on functions $f : \mathbb{Z}_p^d \to \mathbb{R}$, and is defined by $(T_\rho f)(x) = \mathbb{E}_{N_\rho}[f(x + N_\rho)]$. The following standard fact relates the Fourier coefficients of f with those of $T_\rho f$.

Fact 2.2. *For every vector $u \in \mathbb{Z}_p^d$, $\widehat{(T_\rho f)}_u = \hat{f}_u \cdot \rho^{\mathrm{wt}(u)}$.*

Proof. We can write $(T_\rho f)(x) = \mathbb{E}_{N_\rho}[f(x + N_\rho)]$ as

$$\mathbb{E}_{N_\rho}\left[\sum_{u \in \mathbb{Z}_p^d} \hat{f}_u e^{\frac{2\pi i}{p} u \cdot (x + N_\rho)}\right] = \sum_{u \in \mathbb{Z}_p^d} \hat{f}_u e^{\frac{2\pi i}{p} u \cdot x} \mathbb{E}_{N_\rho}\left[e^{\frac{2\pi i}{p} u \cdot N_\rho}\right] = \sum_{u \in \mathbb{Z}_p^d} \hat{f}_u \rho^{\mathrm{wt}(u)} \chi_u,$$

where we used the fact that for every $w \in \mathbb{Z}_p \setminus \{0\}$ we have $\mathbb{E}_{v \in \mathbb{Z}_p}\left[e^{\frac{2\pi i}{p} wv}\right] = 0$. □

Note that, for $p = 2$, i.e. Fourier expansion over $\{0, 1\}^d$, this is equivalent to having $\widehat{(T_\rho f)}_S = \hat{f}_S \rho^{|S|}$ for every $S \subseteq [d]$.

2.2 Edit metric and Ulam metric

Let Σ be the alphabet; we mostly consider $\Sigma = \{0,1\}$ or $\Sigma = \mathbb{Z}_p = \{0,1,\dots p-1\}$ for $p \in \mathbb{N}$ (we will use $p = \Theta(d^3)$).

For $x \in \Sigma^d$, we let x_i denote the i^{th} position in x whenever $i \in [d]$, and extend the notation to $i \notin [d]$ by defining $x_i = x_j$ where $i \equiv j \pmod d$ and $j \in [d]$.

Definition 2.3 (Edit metrics)**.** *Let d be a positive integer. The* edit metric over Σ *is the space Σ^d endowed with distance function* $\mathrm{ed}(x,y)$*, which is defined as the minimum number of character substitutions/insertions/deletions to transform x into y.*

When $|\Sigma| \geq d$, let the Ulam metric *be the space of permutations $x \in \Sigma^d$, where x is called a* permutation *if no symbol $c \in \Sigma$ appears more than once in x. This space is endowed with the same distance function* $\mathrm{ed}(x,y)$.

We note that allowing alphabets Σ bigger than $[d]$ does not make Ulam metric harder (at least in our communication complexity setting) and thus our main theorem carries over to Ulam metric over permutations with alphabet $\Sigma = [d]$ (i.e. the standard notion of permutations). In particular, one can perform the following reduction from the former problem to the latter problem.

Fact 2.4. *For any string length d, and alphabet Σ, $|\Sigma| \geq d$, there is a function $f : \Sigma^d \to \Sigma^{|\Sigma|}$ such that for every pair of permutations $x, y \in \Sigma^d$, we have that $f(x), f(y)$ are permutations over Σ and*

$$\mathrm{ed}(x,y) \leq \mathrm{ed}(f(x), f(y)) \leq 3\,\mathrm{ed}(x,y).$$

Proof. For given $x \in \Sigma^d$, construct $f(x) \in \Sigma^{|\Sigma|}$ by appending all the alphabet symbols that are missing from x in an increasing order. Then, clearly $\mathrm{ed}(f(x), f(y)) \geq \mathrm{ed}(x,y)$. Furthermore, we claim that $\mathrm{ed}(f(x), f(y)) \leq 3\,\mathrm{ed}(x,y)$. Indeed, edit distance between the starting block of length d of $f(x)$ and of $f(y)$ is $\mathrm{ed}(x,y)$. Also, if $z \leq \mathrm{ed}(x,y)$ is the number of symbols that appear in x but not in y and vice-versa, then the edit distance between the ending block of length $|\Sigma| - d$ of $f(x)$ and $f(y)$ is $2z$. Total edit distance between $f(x)$ and $f(y)$ is at most $3\,\mathrm{ed}(x,y)$. ☐

Note that when $|\Sigma| = p = \Theta(d^3)$ as is the case in our main theorem, then $\log|\Sigma| = \Theta(\log d)$, and thus the logarithmic lower bound carries over.

We will also use the following operation on strings which is illustrated in Fig. 1.

Definition 2.5 (Rotation operations)**.** *Fix a positive integer d and an alphabet Σ. For $s, L \in [d]$, define the* right rotation operation $\overrightarrow{R}_{s,L} : \Sigma^d \to \Sigma^d$ *as follows. When applied to a string x, it takes the substring of x of length L starting at position s (with wrap-around), and performs on it one cyclic shift to the right (by 1 position); the rest of x remains unchanged. A* left rotation $\overleftarrow{R}_{s,L}$ *is defined similarly. We call L the* length *of the rotation operation.*

Note that $\overrightarrow{R}_{s,L}$ works as a permutation (and thus is a bijection on the space of strings). Also, for $i \in [L]$, $\left(\overrightarrow{R}_{s,L}\right)^i$ is a rotation of the same block by i positions to the right. Note that a rotation operation $\overrightarrow{R}_{s,L}$ can be simulated by at most two deletions and two insertions (and only one of each when the rotation block does not wrap-around at the string's boundary). Thus, $\mathrm{ed}\left(x, \left(\overrightarrow{R}_{s,L}\right)^i(x)\right) = O(i)$ for every x and i.

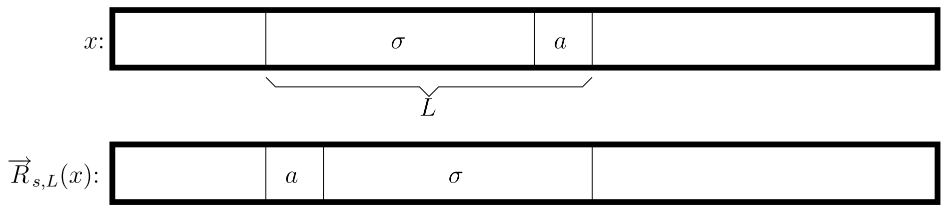

Figure 1: The rotation operation $\overrightarrow{R}_{s,L}(\cdot)$. Here, σ is the substring of length $L-1$ starting at position s in x, and a is the character at position $s+L-1$ in x.

3 Proof of Main Theorem

In this section we prove Theorem 1.1. Fix the values of d and R, and let us use the alphabet $\Sigma = \mathbb{Z}_p$ for p sufficiently large so that a random string from Σ^d is a permutation with high probability (e.g., it suffices to set p to be the smallest prime greater than d^3). For the rest of this section, we denote our hard distribution by $\tilde{\mu} = \frac{\tilde{\mu}_0+\tilde{\mu}_1}{2}$, where $\tilde{\mu}_0$ will be a distribution over *far* pairs of strings (x,y) and $\tilde{\mu}_1$ will be a distribution over *close* pairs (x,y), i.e., $\mathrm{ed}(x,y) > R$ and $\mathrm{ed}(x,y) \le R/\alpha$, respectively.

We will follow the steps outlined in Section 1.3 and eventually put all the pieces together in Section 3.3. Our general approach to proving the theorem uses just a few simple properties of the hard distribution, which we will specify along the way. To differentiate the underlying technique from the specifics of our hard distribution, we describe the hard distribution and prove its required properties separately in Section 4.

3.1 Reduction to Boolean Functions

Our first lemma says that if there is an efficient communication protocol, then there are boolean functions with a non-negligible advantage in distinguishing the distribution $\tilde{\mu}_0$ from $\tilde{\mu}_1$. This lemma is based on the ideas from [AIK08], although the presented proof is simpler than in [AIK08].

Lemma 3.1. *Let $\tilde{\mu}_0$ and $\tilde{\mu}_1$ be distributions over far and close pairs, respectively. If $\mathrm{CC}^{\mathrm{Ulam}_d}_{\alpha,R} \le l$ for some $l \ge 1$, then there exist boolean functions $\mathcal{H}^A, \mathcal{H}^B : \mathbb{Z}_p^d \to \{-1,+1\}$, such that*

$$\Pr_{\tilde{\mu}_0}[\mathcal{H}^A(x) \neq \mathcal{H}^B(y)] - \Pr_{\tilde{\mu}_1}[\mathcal{H}^A(x) \neq \mathcal{H}^B(y)] \ge \tfrac{1}{3} \cdot 2^{-l}.$$

Proof. The idea is to reduce the general communication protocol to a simultaneous (i.e. sketching) protocol where Alice and Bob each send a sketch of one bit only, and the referee performs an equality test on these two bits. Then, using Yao's minimax principle, we easily obtain two deterministic boolean functions $\mathcal{H}^A$ and $\mathcal{H}^B$ that complete the proof.

To accomplish the reduction, consider an actual l-bit (randomized) protocol Π. We construct a one-bit sketching protocol as follows: Alice and Bob make a random guess of the entire transcript of an l-bit protocol using the public coins, uniform over the space of all 2^l protocols (the guess is independent of the actual inputs). Each of them then checks whether the guessed transcript describes the messages they would send in the actual protocol Π, using the guessed transcript to simulate the other party's messages. For example, Alice starts the protocol Π (that depends on her input), but instead of sending the messages to Bob, she verifies that her messages are exactly the same as the ones appearing in the guessed protocol. Alice also uses the messages from the guessed protocol to simulate Bob's answers.

If at any moment Alice (or Bob) spots an inconsistency, she (or he) sends a bit chosen independently at random. Otherwise, Alice outputs 1, and Bob outputs the outcome of the guessed transcript. Observe that if the guessed transcript is not equal to the actual protocol they would have run, then at least one of the two players notices an inconsistency, and one of the bits output by Alice or Bob is random.

Thus, if x and y are such that $\mathrm{ed}(x, y) \leq R/\alpha$ (close pair), then Alice and Bob's bits are equal with probability at least $\frac{2}{3} \cdot 2^{-l} + (1 - 2^{-l})\frac{1}{2} = \frac{1}{2} + \frac{1}{6}2^{-l}$ (where $\frac{2}{3}$ is the probability that the original protocol Π succeeds on (x, y)). Similarly, if x and y are such that $\mathrm{ed}(x, y) > R$ (far pair), then Alice and Bob's bits are equal with probability at most $\frac{1}{3} \cdot 2^{-l} + (1 - 2^{-l}) \cdot \frac{1}{2} = \frac{1}{2} - \frac{1}{6}2^{-l}$. Using Yao's minimax principle, we conclude that, for given distributions $\tilde{\mu}_0$ and $\tilde{\mu}_1$ over far and close pairs respectively, there exist some fixed boolean functions $\mathcal{H}^A, \mathcal{H}^B$ that achieve a success probability at least $\frac{1}{2} + \frac{1}{6}2^{-l}$ on the distribution $\tilde{\mu} = \frac{\tilde{\mu}_0 + \tilde{\mu}_1}{2}$, or, formally,

$$\frac{1}{2} \Pr_{\tilde{\mu}_0}[\mathcal{H}^A(x) \neq \mathcal{H}^B(y)] + \frac{1}{2} \Pr_{\tilde{\mu}_1}[\mathcal{H}^A(x) = \mathcal{H}^B(y)] \geq \frac{1}{2} + \frac{1}{6} \cdot 2^{-l}.$$

We conclude that $\Pr_{\tilde{\mu}_0}[\mathcal{H}^A(x) \neq \mathcal{H}^B(y)] - \Pr_{\tilde{\mu}_1}[\mathcal{H}^A(x) \neq \mathcal{H}^B(y)] \geq \frac{1}{3} \cdot 2^{-l}$. $\square$

The rest of the proof of Theorem 1.1 uses these boolean functions $\mathcal{H}^A, \mathcal{H}^B$.

3.2 From Boolean Functions to λ-Tests

Next we provide a method to lower bound the advantage achieved by the boolean functions $\mathcal{H}^A, \mathcal{H}^B$, by relating it to a certain statistical property of the hard distribution $\tilde{\mu}$. Our hard distribution $\tilde{\mu} = \frac{\tilde{\mu}_0 + \tilde{\mu}_1}{2}$ will have a specific generic construction that we describe next. For each $t \in \{0, 1\}$, the distribution $\tilde{\mu}_t$ is formed via a small modification of another distribution μ_t, which is easier to analyze (due to certain independencies), but might (rarely) produce invalid inputs. Specifically, each $\tilde{\mu}_t$ is the distribution μ_t conditioned on the fact that the pair $(x, y) \in \mu_t$ is valid in the sense that x and y are both permutations and the pair (x, y) is respectively a far (when $t = 0$) or a close (when $t = 1$) pair. We analyze below the distributions μ_0 and μ_1 (specifically, in Lemma 3.4). For completeness, we mention that, in the next section, we show that this analysis extends to distributions $\tilde{\mu}_0$ and $\tilde{\mu}_1$ using the fact that $\tilde{\mu}_0$ and $\tilde{\mu}_1$ are statistically very close to distributions μ_0 and μ_1 respectively.

The distribution μ_t consists of pairs (x, y) chosen as follows: $x \in \mathbb{Z}_p^d$ is chosen uniformly at random, and y is constructed from x in two steps. In the first step, let $z \triangleq x + N_\rho$, where N_ρ, defined in the preliminaries, is noise of rate $\rho \in (0, 1)$, independent of t. In the second step, y is obtained from z by permuting the coordinates of z according to a distribution $\mathcal{D}_t$. Formally, $\mathcal{D}_t$ is a distribution over permutation operations, where a *permutation operation* is a function $\pi : \mathbb{Z}_p^d \to \mathbb{Z}_p^d$ for which there exists a permutation $\hat{\pi} : [d] \to [d]$ such that $\pi(x) \equiv (x_{\hat{\pi}(1)}, \ldots x_{\hat{\pi}(d)})$. We will require that $\mathcal{D}_t$ is *symmetric* in the sense that, for every π, the permutation operations π and π^{-1} are equi-probable (in it). Notice that y has the same marginal distribution as x, i.e. uniform over $\mathbb{Z}_p^d$.

We now quantify the "difference" between the distributions $\mathcal{D}_0, \mathcal{D}_1$ from the perspective of what we call λ-*tests*. For $\lambda \in [d]$, we define a λ-*test* to be a vector $u \in \mathbb{Z}_p^d$ with precisely λ non-zero entries, i.e., $\mathrm{wt}(u) = \lambda$. For a distribution $\mathcal{D}_t$ and $\lambda \in [d]$, let the matrix $A^{(t,\lambda)}$ be the transition matrix of a Markov chain whose states are all the λ-tests, and whose transitions are according to $\mathcal{D}_t$, i.e., at a λ-test u, the process picks $\pi \in \mathcal{D}_t$ and moves to state $\pi(u)$ (which is also a λ-test). In other words, a row corresponding to u in $A^{(t,\lambda)}$ is a vector, that has, for every λ-test w, a coordinate of value $\Pr_{\pi \in \mathcal{D}_t}[\pi(u) = w]$. We denote this row by $A_u^{(t,\lambda)}$. Note that the matrix $A^{(t,\lambda)}$ is symmetric (since $\mathcal{D}_t$ is symmetric) and thus it is doubly-stochastic.

Definition 3.2. *The* λ-test distinguishability *of* $\mathcal{D}_0, \mathcal{D}_1$*, denoted* Δ_λ*, is the maximum, over all* λ*-tests* u*, of the total variation distance between the distributions* $A_u^{(0,\lambda)}$ *and* $A_u^{(1,\lambda)}$*.*

We can also write Δ_λ using matrix norms (as per Definition 2.1), and the easy fact that $\|B\|_\infty = \max_{i \in [n]} \sum_{j \in [n]} |B_{ij}|$ for all $B \in M_{n,n}(\mathbb{R})$. Later (in Fact 3.8) we shall use known inequalities between different matrix norms (in particular ℓ_∞ and ℓ_2).

Fact 3.3. $\Delta_\lambda = \|A^{(0,\lambda)} - A^{(1,\lambda)}\|_\infty / 2$.

The following lemma bounds the advantage achieved by $\mathcal{H}^A, \mathcal{H}^B$ in terms of the λ-test distinguishability Δ_λ of distributions $\mathcal{D}_0$ and $\mathcal{D}_1$ for any pair of distributions $\mathcal{D}_0, \mathcal{D}_1$. Note that we have not yet specified the distributions $\mathcal{D}_0$ and $\mathcal{D}_1$ themselves. We will specify the distributions $\mathcal{D}_0$ and $\mathcal{D}_1$ in Section 4, thus completing the definition of the hard distribution $\tilde{\mu}$.

Lemma 3.4. *Consider* $\mathcal{H}^A, \mathcal{H}^B : \mathbb{Z}_p^d \to \{-1, +1\}$ *and* $\rho \in (0,1)$*. If each* μ_t*, for* $t \in \{0,1\}$*, is defined as above from a symmetric distributions* $\mathcal{D}_t$ *over permutation operations, then*

$$\Pr_{\mu_0}[\mathcal{H}^A(x) \neq \mathcal{H}^B(y)] - \Pr_{\mu_1}[\mathcal{H}^A(x) \neq \mathcal{H}^B(y)] \leq \max_{\lambda \in [d]} \Delta_\lambda \rho^\lambda.$$

Proof. For $t \in \{0,1\}$, define $C^{(t)} \triangleq \mathbb{E}_{\mu_t}\left[\mathcal{H}^A(x)\mathcal{H}^B(y)\right]$ to be the *correlation* between the two boolean functions. Note that, $\Pr_{\mu_t}[\mathcal{H}^A(x) \neq \mathcal{H}^B(y)] = \frac{1}{4}\mathbb{E}_{\mu_t}\left[\mathcal{H}^A(x) - \mathcal{H}^B(y)\right]^2 = 1/2 - C^{(t)}/2$. Thus,

$$\Pr_{\mu_0}[\mathcal{H}^A(x) \neq \mathcal{H}^B(y)] - \Pr_{\mu_1}[\mathcal{H}^A(x) \neq \mathcal{H}^B(y)] = \frac{C^{(1)} - C^{(0)}}{2}.$$

We will show that $C^{(1)} - C^{(0)} \leq 2\max_{\lambda \in [d]} \Delta_\lambda \rho^\lambda$. For this purpose, it is more convenient to express each $C^{(t)}$ in terms of the Fourier coefficients of $\mathcal{H}^A$ and $\mathcal{H}^B$. Recall that μ_t is generated by picking a random x, and constructing y from x by adding to it the noise N_ρ and then applying a random permutation drawn from $\mathcal{D}_t$, namely, $y = \pi(x + N_\rho)$, where $\pi \in \mathcal{D}_t$. Let $\mu_t|x$ denote the distribution μ_t conditioned on the value of x. Thus,

$$\mathbb{E}_{\mu_t}\left[\mathcal{H}^A(x)\mathcal{H}^B(y)\right] = \mathbb{E}_{x \in \mathbb{Z}_p^d}\left[\mathcal{H}^A(x) \cdot \mathbb{E}_{\mu_t|x}\left[\mathcal{H}^B(y)\right]\right]$$

Define $f^{(t)}(x) \triangleq \mathbb{E}_{\mu_t|x}\left[\mathcal{H}^B(y)\right]$. Then

$$f^{(t)}(x) = \mathbb{E}_{N_\rho}\left[\mathbb{E}_{\pi \in \mathcal{D}_t}\left[\mathcal{H}^B(\pi(x + N_\rho))\right]\right].$$

Since $C^{(t)} = \mathbb{E}_x\left[\mathcal{H}^A(x)f^{(t)}(x)\right]$, we can switch to the Fourier basis by applying Parseval's identity, and get

$$C^{(t)} = \sum_{u \in \mathbb{Z}_p^d} \widehat{(\mathcal{H}^A)}_u \overline{\widehat{(f^{(t)})}_u}, \tag{1}$$

where $\widehat{(\mathcal{H}^A)}_u$ and $\widehat{(f^{(t)})}_u$ are the Fourier coefficients of $\mathcal{H}^A$ and $f^{(t)}$ respectively.

The next proposition, which we shall prove shortly, expresses the level λ Fourier coefficients of $f^{(t)}$ in terms of those of $\mathcal{H}^B$. Let $\left(\widehat{(f^{(t)})}_u\right)_{u:\mathrm{wt}(u)=\lambda}$ be the vector of the Fourier coefficients of $f^{(t)}$ indexed by u's of weight $\mathrm{wt}(u) = \lambda$. Define $\left(\widehat{(\mathcal{H}^B)}_u\right)_{u:\mathrm{wt}(u)=\lambda}$ similarly.

Proposition 3.5. *For all $\lambda \in [d]$ and $\mathcal{H}^B : \mathbb{Z}_p^d \to \mathbb{C}$,*

$$\left(\widehat{(f^{(t)})}_u\right)_{u:\mathrm{wt}(u)=\lambda} = \rho^\lambda A^{(t,\lambda)} \cdot \left(\widehat{(\mathcal{H}^B)}_u\right)_{u:\mathrm{wt}(u)=\lambda} \tag{2}$$

This proposition naturally leads us to break each $C^{(t)}$ into the terms corresponding to each Fourier level λ. Define the λ^{th}-*correlation* to be

$$C_\lambda^{(t)} \triangleq \sum_{u \in \mathbb{Z}_p^d : \mathrm{wt}(u)=\lambda} \widehat{(\mathcal{H}^A)}_u \overline{\widehat{(f^{(t)})}_u}. \tag{3}$$

Then, $C^{(1)} - C^{(0)} = \sum_{\lambda=0}^d \left(C_\lambda^{(1)} - C_\lambda^{(0)}\right)$. We can now bound each $C_\lambda^{(1)} - C_\lambda^{(0)}$ in terms of Δ_λ and ρ.

Let $\omega_\lambda^A = \left\|\left(\widehat{(\mathcal{H}^A)}_u\right)_{u:\mathrm{wt}(u)=\lambda}\right\|_2$ be the ℓ_2-weight of the level λ Fourier coefficients of $\mathcal{H}^A$, and define similarly ω_λ^B. By Parseval's identity, $\sum_{\lambda=0}^d \left(\omega_\lambda^A\right)^2 = \mathbb{E}_x\left[\mathcal{H}^A(x) \cdot \overline{\mathcal{H}^A(x)}\right] = 1$, and similarly $\sum_{\lambda=0}^d \left(\omega_\lambda^B\right)^2 = 1$.

Proposition 3.6. *For all $\lambda \in [d]$,*

$$C_\lambda^{(1)} - C_\lambda^{(0)} \le 2\Delta_\lambda \rho^\lambda \cdot \omega_\lambda^A \omega_\lambda^B.$$

We will prove the proposition shortly by a straightforward calculation. In addition, $C_0^{(1)} = C_0^{(0)}$ because the 0-th level Fourier coefficient of $f^{(t)}$ equals $\mathbb{E}_{x\in\mathbb{Z}_p^d}\left[f^{(t)}(x)\right] = \mathbb{E}_{y\in\mathbb{Z}_p^d}\left[\mathcal{H}^B(y)\right]$, which does not depend on $t \in \{0,1\}$. Given the above proposition, we thus have

$$\begin{aligned} C^{(1)} - C^{(0)} &= \sum_{\lambda=0}^d \left(C_\lambda^{(1)} - C_\lambda^{(0)}\right) \le \sum_{\lambda=1}^d 2\Delta_\lambda \rho^\lambda \cdot \omega_\lambda^A \omega_\lambda^B \\ &\le \sum_{\lambda=1}^d 2\Delta_\lambda \rho^\lambda \cdot \frac{\left(\omega_\lambda^A\right)^2 + \left(\omega_\lambda^B\right)^2}{2} \le 2 \max_{\lambda \in [d]} \Delta_\lambda \rho^\lambda, \end{aligned}$$

where we used the geometric–arithmetic mean inequality. This finishes the proof of Lemma 3.4. $\square$

It remains to prove Propositions 3.5 and 3.6.

Proof of Proposition 3.5. Define a new function $g^{(t)} : \mathbb{Z}_p^d \to \mathbb{R}$ as

$$g^{(t)}(z) \triangleq \mathbb{E}_{\pi \in \mathcal{D}_t}\left[\mathcal{H}^B(\pi(z))\right].$$

Then $f^{(t)} = T_\rho g^{(t)}$, and thus $\widehat{(f^{(t)})}_u = \widehat{(g^{(t)})}_u \cdot \rho^{\mathrm{wt}(u)}$ for all $u \in \mathbb{Z}_p^d$ (by Fact 2.2). It remains to prove that

$$\left(\widehat{(g^{(t)})}_u\right)_{u:\mathrm{wt}(u)=\lambda} = A^{(t,\lambda)} \cdot \left(\widehat{(\mathcal{H}^B)}_u\right)_{u:\mathrm{wt}(u)=\lambda} \tag{4}$$

Similarly to the operator T_ρ, we define the operator $\mathcal{O}_t$ as $(\mathcal{O}_t\mathcal{H}^B)(x) \triangleq \mathbb{E}_{\pi\in\mathcal{D}_t}\left[\mathcal{H}^B(\pi(x))\right]$. Since $g^{(t)} = \mathcal{O}_t\mathcal{H}^B$, we proceed to analyze how the operator $\mathcal{O}_t$ works on the Fourier coefficients of a function $\mathcal{H}^B$.

Fact 3.7. *For a permutation operation π, define $\mathcal{P}_\pi$ to be an operator on functions $\psi : \mathbb{Z}_p^d \to \mathbb{R}$, given by $(\mathcal{P}_\pi\psi)(x) \triangleq \psi(\pi(x))$. Then, $\widehat{(\mathcal{P}_\pi\psi)}_u = \hat{\psi}_{\pi(u)}$.*

Now, the operator $\mathcal{O}_t$ defined earlier is simply a convex combination of several $\mathcal{P}_\pi$, where π is drawn from $\mathcal{D}_t$. Thus, with the above fact, for every $u \in \mathbb{Z}_p^d$,

$$\widehat{(g^{(t)})}_u = \widehat{(\mathcal{O}_t\mathcal{H}^B)}_u = \mathbb{E}_{\pi\in\mathcal{D}_t}\left[\widehat{(\mathcal{H}^B)}_{\pi(u)}\right]. \tag{5}$$

Consequently, the vector of level λ Fourier coefficients of $g^{(t)}$ can be written as a product of the matrix $A^{(t,\lambda)}$ and the vector of the (same) level λ Fourier coefficients of $\mathcal{H}^B$, which proves Proposition 3.5. ☐

We will need the following fact for the proof of Proposition 3.6. Recall that $\|A\|_p$ denotes the p-norm of such a matrix A, as per Definition 2.1.

Fact 3.8. *Let $B \in M_{n,n}(\mathbb{R})$ be a symmetric matrix. Then, $\|B\|_2 \le \|B\|_\infty$.*

Proof. It is known that $\|B\|_1 = \max_{j\in[n]}\sum_{i\in n}|B_{ij}|$ and $\|B\|_\infty = \max_{i\in[n]}\sum_{j\in[n]}|B_{ij}|$, and since B is symmetric, these two norms are equal. By Riesz-Thorin interpolation theorem, $\|B\|_2 \le \max\{\|B\|_1, \|B\|_\infty\} = \|B\|_\infty$. (The Riesz-Thorin interpolation theorem states that for every $1 \le p < q < r \le \infty$ and a real matrix A, we have $\|A\|_q \le \max\{\|A\|_p, \|A\|_r\}$.) ☐

Proof of Proposition 3.6. For every λ, the matrix $A^{(t,\lambda)}$ is symmetric, and so is $A^{(1,\lambda)} - A^{(0,\lambda)}$. Thus,

$$\begin{aligned}
C_\lambda^{(1)} - C_\lambda^{(0)} &= \sum_{u\in\mathbb{Z}_p^d:\mathrm{wt}(u)=\lambda} \widehat{(\mathcal{H}^A)}_u \cdot \left(\overline{\widehat{(f^{(1)})}_u} - \overline{\widehat{(f^{(0)})}_u}\right) \\
&\le \left\|\left(\widehat{(\mathcal{H}^A)}_u\right)_{u:\mathrm{wt}(u)=\lambda}\right\|_2 \cdot \left\|\left(\overline{\widehat{(f^{(1)})}_u} - \overline{\widehat{(f^{(0)})}_u}\right)_{u:\mathrm{wt}(u)=\lambda}\right\|_2 \\
&= \omega_\lambda^A \cdot \left\|\rho^\lambda\left(A^{(1,\lambda)} - A^{(0,\lambda)}\right)\left(\overline{\widehat{(\mathcal{H}^B)}_u}\right)_{u:\mathrm{wt}(u)=\lambda}\right\|_2 \\
&\le \rho^\lambda \cdot \omega_\lambda^A \cdot \left\|A^{(1,\lambda)} - A^{(0,\lambda)}\right\|_2 \left\|\left(\widehat{(\mathcal{H}^B)}_u\right)_{u:\mathrm{wt}(u)=\lambda}\right\|_2 \\
&\le \rho^\lambda \cdot \omega_\lambda^A\omega_\lambda^B \cdot \left\|A^{(1,\lambda)} - A^{(0,\lambda)}\right\|_\infty \\
&= 2\Delta_\lambda \cdot \rho^\lambda \cdot \omega_\lambda^A\omega_\lambda^B;
\end{aligned}$$

where we used Eqn. (3), Cauchy-Schwarz, Proposition 3.5, Definition 2.1, Fact 3.8, and Definition 3.3, respectively. ☐

3.3 Putting It All Together

We proceed to proving Theorem 1.1, using the machinery just developed in Sections 3.1 and 3.2. Recall that we still need to exhibit a suitable hard distribution. We outlined the construction of our hard distribution in Section 3.2 ; the construction relies on two distributions $\mathcal{D}_0$ and $\mathcal{D}_1$ which were not specified. The next lemma asserts that the desired hard distribution exists. More precisely, it asserts that it can be constructed to satisfy the required properties, such as a small λ-test distinguishability.

Lemma 3.9 (Hard Distribution)**.** *There exist constants $\theta, c_1, c_2, d_0 > 0$, such that for all $d > d_0$, $p > d^3$, $1 < \alpha \le O(\frac{\log d}{\log\log d})$, $d^{0.1} \le R \le d^{0.49}$, there exist symmetric distributions $\mathcal{D}_0^*$ and $\mathcal{D}_1^*$ over permutation operations on $\mathbb{Z}_p^d$ (as defined in Section 3.2) with the following guarantees.*

(a). For all $\lambda \ge 1$, the λ-test distinguishability of $\mathcal{D}_0^*$ and $\mathcal{D}_1^*$ is $\Delta_\lambda \le c_1 \cdot \lambda \frac{\log \alpha}{\log d} \cdot \frac{R}{d}$.

(b). Define each distribution μ_t from $\mathcal{D}_t^*$ as described in Section 3.2, setting $\rho = 1 - \theta \frac{R/\alpha}{d}$. Define the distribution $\tilde{\mu}_t$ to be the restriction (i.e. conditioning) of μ_t to the event that the sampled pair $(x, y) \in \mu_t$ is legal, in the sense that $x, y \in \mathbb{Z}_p^d$ are permutations and are respectively a far pair (for $t = 0$) or a close pair (for $t = 1$). Then for each $t \in \{0, 1\}$, the total variation distance between $\tilde{\mu}_t$ and μ_t is at most d^{-c_2}.

We prove this lemma separately in Section 4, where we include a full description of $\mathcal{D}_0^*$ and $\mathcal{D}_1^*$. Here, we use the lemma to complete the proof of the main theorem.

Proof of Theorem 1.1. First, consider the hard distribution given by Lemma 3.9. Next, by Lemma 3.1, there must exist functions $\mathcal{H}^A, \mathcal{H}^B$ such that

$$\Pr_{\tilde{\mu}_0}[\mathcal{H}^A(x) \ne \mathcal{H}^B(y)] - \Pr_{\tilde{\mu}_1}[\mathcal{H}^A(x) \ne \mathcal{H}^B(y)] \ge \tfrac{1}{3} \cdot 2^{-\mathrm{CC}_{\alpha,R}^{\mathrm{Ulam}_d}}.$$

Applying Lemma 3.4 to the distributions μ_0, μ_1, and using the fact that $\tilde{\mu}_0$ and $\tilde{\mu}_1$, respectively are statistically close to μ_0 and μ_1 (Lemma 3.9(b)), we deduce that

$$\Pr_{\tilde{\mu}_0}[\mathcal{H}^A(x) \ne \mathcal{H}^B(y)] - \Pr_{\tilde{\mu}_1}[\mathcal{H}^A(x) \ne \mathcal{H}^B(y)] \le \max_{\lambda \in [d]} \Delta_\lambda \rho^\lambda + d^{-c_2}.$$

Combining the two inequalities above and plugging in the upper bound on Δ_λ and the value of ρ from Lemma 3.9(a), we have

$$\begin{aligned}
\tfrac{1}{3} \cdot 2^{-\mathrm{CC}_{\alpha,R}^{\mathrm{Ulam}_d}} &\le \max_{\lambda \in [d]} \left[c_1 \cdot \lambda \tfrac{\log \alpha}{\log d} \cdot \tfrac{R}{d} \cdot \left(1 - \theta \tfrac{R/\alpha}{d}\right)^\lambda \right] + d^{-c_2} \\
&\le \tfrac{c_1}{\theta} \cdot \alpha \cdot \tfrac{\log \alpha}{\log d} \cdot \max_{x \ge 0} x \cdot e^{-x} + d^{-c_2} \\
&= O\left(\tfrac{\alpha \log \alpha}{\log d}\right),
\end{aligned}$$

which concludes the proof of Theorem 1.1. $\square$

4 Construction of The Hard Distribution

In this section we prove Lemma 3.9. We start by giving the detailed construction of our hard distribution $\tilde{\mu} = \frac{\tilde{\mu}_0+\tilde{\mu}_1}{2}$. Then, in Sections 4.1 and 4.2 we prove respectively λ-test indistinguishability (part (a)) and statistical closeness (part (b)) properties of the hard distribution.

The hard distribution construction follows the outline given in Section 3.2. We first specify the distributions $\mathcal{D}_0^*, \mathcal{D}_1^*$ over permutation operators, which forms the bulk of the construction. Once these distributions are specified, we obtain the intermediary distributions μ_0 and μ_1 as already described in Section 3.2. We finalize the description by constructing $\tilde{\mu}_t$ from μ_t, for each $t \in \{0,1\}$, by conditioning on the pair $(x,y) \in \mu_t$ being a legal pair, namely that $x, y \in \mathbb{Z}_p^d$ are permutations and are respectively a far pair (for $t=0$) or a close pair (for $t=1$).

Fix $\epsilon_0 \triangleq 1/2$ and select $\epsilon_1 = \Theta(\frac{1}{\alpha})$ as follows. Let $\beta \triangleq \frac{1-\epsilon_1}{1-\epsilon_0} = 2(1-\epsilon_1)$, and $\xi_1 \triangleq \lceil \log_2(C_1\alpha) \rceil$, for a sufficiently large constant $C_1 > 0$ (in particular $C_1 = 805$ will suffice). Let ϵ_1 be the solution to the equation $(1-\epsilon_1) = \epsilon_1 \beta^{\xi_1}$ satisfying $\epsilon_1 \le \frac{2}{C_1\alpha}$. The existence of ϵ_1 follows from the following claim, whose proof is deferred to the end of the construction.

Claim 4.1. *Let $\alpha > 1$ and $C_1 > 1$ be sufficiently large. Then there exists ϵ_1 with $\frac{1}{3C_1\alpha} < \epsilon_1 \le \frac{2}{C_1\alpha}$ such that $(1-\epsilon_1) = \epsilon_1(2(1-\epsilon_1))^{\xi_1}$, where $\xi_1 = \lceil \log_2(C_1\alpha) \rceil$.*

We thus have, by construction,

$$\epsilon_0 = (1-\epsilon_0) = (1-\epsilon_1)\beta^{-1} = \epsilon_1 \beta^{\xi_1 - 1}. \tag{6}$$

For each $t \in \{0,1\}$, we define the distribution μ_t over (x,y) such that $\operatorname{ed}(x,y)$ is almost surely $\Theta(\epsilon_t R)$. Choose $x \in \Sigma^d = \mathbb{Z}_p^d$ uniformly at random. Then set $z \triangleq x + N_\rho$ where $N_\rho \in \mathbb{Z}_p^d$ is a noise of rate $\rho \triangleq 1 - \epsilon_1 R/d$ (i.e., each position is randomized with probability $1-\rho = \epsilon_1 R/d$). We shall obtain y from z, by applying a number of random rotation operations, each picked independently from a specific distribution. We use the following notation:

- $m \triangleq 0.01 \cdot \log_\beta d = \Theta(\log d)$ is the number of possible lengths of a rotation operation;
- $L_{\min} \triangleq d^{0.01}$ determines the minimum length of a rotation operation (modulo a factor of β);
- $w \triangleq C_2 \cdot \frac{R}{m \cdot L_{\min}}$ is the number of rotation operations that we apply, for a sufficiently large constant $C_2 > 0$ to be determined later (in Section 4.2).

Generate a sequence $(r_1, r_2, \ldots, r_w)$ of w rotations by picking each r_i i.i.d. according to the following distribution $\mathcal{D}_t^{\text{rot}}$:

1. Pick $l_i \in [m]$ randomly so that $\Pr[l_i = l] = \frac{\beta^{-l}}{\zeta}$ for each $l \in [m]$, where $\zeta = \sum_{l=1}^m \beta^{-l}$ is the normalization constant.
2. Pick a starting position $s_i \in [d]$ uniformly at random, and rotate the block that starts at position s_i and has length (with wrap-around) $L_i = \beta^{l_i} L_{\min}$ by $\epsilon_t L_i$ positions, either to the right or to the left, at random. We choose r_i at random from the set $\left\{ (\tilde{R}_{s,L_i})^{\epsilon_t L_i} \mid s \in [d], \tilde{R} \in \{\overrightarrow{R}, \overleftarrow{R}\} \right\}$.

   We note that $(\tilde{R}_{s,L_i})^{\epsilon_t L_i}$ is not well defined when $\epsilon_t L_i$ or L_i are not integers. Overloading the notation, we define $(\overrightarrow{R}_{s,L_i})^{\epsilon_t L_i}$ for non-integer $\epsilon_t L_i, L_i$ as follows. Let B_1 be the block that starts at position s_i and has length $\lfloor(1-\epsilon_t)L_i\rfloor$, and let B_2 be the block immediately following B_1 of length $\lfloor\epsilon_t L_i\rfloor$, i.e.

$$B_1 = [s : s + \lfloor(1-\epsilon_t)L_i\rfloor - 1], \quad B_2 = [s + \lfloor(1-\epsilon_t)L_i\rfloor : s + \lfloor(1-\epsilon_t)L_i\rfloor + \lfloor\epsilon_t L_i\rfloor - 1].$$

Then, $(\overrightarrow{R}_{s,L_i})^{\epsilon_t L_i}$ swaps blocks B_1 and B_2. We define $(\overleftarrow{R}_{s,L_i})^{\epsilon_t L_i}$ similarly.

To obtain y, we apply to $z = x + N_\rho$ the sequence of rotations $r_1, \ldots, r_w$, i.e.,

$$y \triangleq r_w(r_{w-1}(\ldots r_1(z)\ldots)) = (r_w \circ \ldots \circ r_2 \circ r_1)(x + N_\rho).$$

In the language of Section 3.2, the distribution $\mathcal{D}_t^*$ of permutation operations is simply the distribution of $\pi = r_w \circ r_{w-1} \circ \ldots \circ r_1$, where $r_1, \ldots r_w$ are drawn independently from $\mathcal{D}_t^{\mathrm{rot}}$.

Intuitively, each rotation operation r_i, or more precisely its distribution $\mathcal{D}_t^{\mathrm{rot}}$, is designed to achieve the following goal. Consider a position $j \in [d]$ and assume for simplicity $j \in [0.1d, 0.9d]$. Let the random variable $Z_{t,j} \in \mathbb{Z}$ be the displacement (change in position) of position j under a (random) rotation operation $r_i \in \mathcal{D}_t^{\mathrm{rot}}$, i.e. $Z_{t,j} \in \mathbb{Z}$ is the unique value such that $r_i(e_j) = e_{j+Z_{t,j}}$, where e_k denotes the k-th standard basis vector. By construction, $Z_{t,j}$ is symmetric around 0, i.e. $\Pr[Z_{t,j} = k] = \Pr[Z_{t,j} = -k]$, and its distribution does not depend on j, i.e. $Z_{t,j}$ and $Z_{t,j'}$ have the same distribution (but they are correlated). Moreover, its support, i.e. values $k > 0$ with probability $\Pr[Z_{t,j} = k] > 0$, forms a geometric sequence (because the block length L has a geometric distribution). Let us now condition on the event that position j is included in the rotation block, i.e. $Z_{t,j} \neq 0$. Then the distribution of $Z_{t,j}$ is almost uniform over the support — this follows from the distribution of L and of s, and by Eqn. (6). Furthermore, the distributions of $Z_{0,j}$ and $Z_{1,j}$ (when we condition on them being nonzero) are almost identical, because their supports differ only at the boundaries, i.e. at the smallest and largest displacements, again due to Eqn. 6, and they are both almost uniform. We repeat the rotation operation many times in order to obtain a high concentration in the distance between y and z.

To finalize the construction, it remains to define $\tilde{\mu}_t$ for $t \in \{0,1\}$. We note that we cannot set $\tilde{\mu}_t$ to be exactly μ_t because the latter may sometimes generate pairs (x,y) that are not far or close respectively, or are not even permutations altogether. (x and y are not always permutations since each of the two strings is uniformly at random and may have a multiple occurrence of the same symbol.) We thus define $\tilde{\mu}_0$ to be the distribution μ_0 restricted to (i.e. conditioned on) pairs of permutations (x,y) with $\mathrm{ed}(x,y) > R$, and similarly $\tilde{\mu}_1$ is the distribution μ_1 restricted to pairs of permutations with $\mathrm{ed}(x,y) \le R/\alpha$.

It only remains to prove Claim 4.1, namely that the desired ϵ_1 exists.

Proof of Claim 4.1. Define function $f(x) : [0,1] \to \mathbb{R}$ as $f(x) = x \cdot (1-x)^{\xi_1 - 1} 2^{\xi_1} - 1$. Note that ϵ_1 is the solution to the equation $f(x) = 0$. For $x = 1/(3C_1\alpha)$, $f(x) \le \frac{1}{3C_1\alpha}(1 - \frac{1}{3C_1\alpha})^{\xi_1 - 1} \cdot 2^{\log_2(C_1\alpha)+1} - 1 < 0$. Similarly, for $x = \frac{2}{C_1\alpha}$, $f(x) \ge \frac{2}{C_1\alpha}(1 - \frac{2(\xi_1-1)}{C_1\alpha}) \cdot 2^{\log_2(C_1\alpha)} - 1 \ge 2(1 - \frac{2(\log_2(C_1\alpha)-1)}{C_1\alpha}) - 1 > 0$ provided C_1 is a sufficiently large constant. By the continuity of $f(x)$, there exists some $x \in [\frac{1}{3C_1\alpha}, \frac{2}{C_1\alpha}]$ satisfying $f(x) = 0$. $\square$

In the rest of this section we prove the two properties required from our hard distribution, stated in Lemma 3.9: that $\mathcal{D}_0^*$ and $\mathcal{D}_1^*$ have small λ-test distinguishability (Lemma 3.9 (a)), and that each $\tilde{\mu}_t$ is very close to μ_t, for both $t \in \{0,1\}$ (Lemma 3.9 (b)).

Here and throughout the big $O(\cdot)$ notation may hide dependence on constants used in the construction of the hard distribution, namely C_1 and C_2. Furthermore, although the parameters

β and ζ are not constants (they depend on α), we can bound $1.5 < \beta < 2$, which guarantees that $\frac{1}{1-\beta^{-1}} \le O(1)$ and $\frac{1}{\zeta} \le \beta \le O(1)$.

4.1 λ-test Indistinguishability

We prove Lemma 3.9 (a) via the following lemma.

Lemma 4.2. *Let Δ_λ be the λ-test distinguishability of $\mathcal{D}_0^*$ and $\mathcal{D}_1^*$. Then for all $\lambda \ge 1$, we have $\Delta_\lambda \le O\left(\lambda \frac{\log \alpha}{\log d} \cdot \frac{R}{d}\right)$.*

Proof. Fix a λ-test $u \in \mathbb{Z}_p^d$ and let $\delta_\lambda(u) = \max_{T \subseteq \mathbb{Z}_p^d} \left| \Pr[r^{(0)}(u) \in T] - \Pr[r^{(1)}(u) \in T] \right|$ be the total variation distance between the distributions $r^{(0)}(u)$ and $r^{(1)}(u)$, where $r^{(t)} \in \mathcal{D}_t^{\mathrm{rot}}$ for $t \in \{0,1\}$. The heart of this lemma is the following bound, which we shall prove below:

$$\delta_\lambda(u) \le O\left(\lambda \log \alpha \cdot \frac{L_{\min}}{d}\right). \tag{7}$$

We shall also prove shortly the claim that $\Delta_\lambda \le w \cdot \max_u \delta_\lambda(u)$. The lemma then follows immediately from Eqn. (7) and this claim, by plugging the former into the latter and recalling $w = C_2 \cdot \frac{R}{m \cdot L_{\min}}$ is the number of rotation operations. Since $\lambda \frac{\log \alpha}{\log d} \cdot \frac{R}{d} > 1$ for $\lambda \ge d^{0.95}$, it actually suffices to prove (7) only for $\lambda < d^{0.95}$.

We now prove the above claim, that $\Delta_\lambda \le w \cdot \max_u \delta_\lambda(u)$, by induction. Let $v_i^t = r_i^{(t)}(r_{i-1}^{(t)}(\ldots r_1^{(t)}(u) \ldots))$ for $t \in \{0,1\}$ and $i \in [w]$. We prove that, for any $T \subseteq \mathbb{Z}_p^d$, we have $|\Pr[v_i^0 \in T] - \Pr[v_i^1 \in T]| \le i \cdot \max_v \delta_\lambda(v)$. The base case $i = 1$ holds by the definition of δ_λ, and so we turn to the inductive step:

$$\begin{aligned}
\Pr[v_i^0 \in T] &= \sum_v \Pr[v_{i-1}^0 = v] \Pr[r_i^{(0)}(v) \in T] \\
&\le \sum_v \Pr[v_{i-1}^0 = v] \left(\Pr[r_i^{(1)}(v) \in T] + \delta_\lambda(v)\right) \\
&\le \max_v \delta_\lambda(v) + \sum_r \Pr[r_i^{(1)} = r] \Pr[r(v_{i-1}^0) \in T] \\
&\le \max_v \delta_\lambda(v) + \sum_r \Pr[r_i^{(1)} = r] \left(\Pr[r(v_{i-1}^1) \in T] + (i-1) \cdot \max_v \delta_\lambda(v)\right) \\
&= i \cdot \max_v \delta_\lambda(v) + \Pr[v_i^1 \in T].
\end{aligned}$$

Proving the same inequality with the roles of $t = 0$ and $t = 1$ reversed, we obtain that $\Delta_\lambda = \max_{T \subseteq \mathbb{Z}_p^d} |\Pr[v_w^0 \in T] - \Pr[v_w^1 \in T]| \le w \cdot \max_u \delta_\lambda(u)$.

In the rest of the proof of Lemma 4.2, we prove the bound (7). The proof consists of two parts. The first part proves the bound for $\lambda = 1$, and contains the main intuition why our distribution is hard. The second part builds on the first one to show the bound for general λ.

Part 1: $\lambda = 1$. We prove that $\delta_1(u) \le O(\log \alpha \cdot \frac{L_{\min}}{d})$ next. In this part, we shall assume that L and $\epsilon_t L$ are integers, deferring the full treatment of this technicality to the second part.

Since $\lambda = 1$, we have only one non-zero entry in u, say at position j. For $t \in \{0,1\}$, let j_t be the random variable denoting the position of the symbol u_j in the vector $r^{(t)}(u)$ obtained by applying the random rotation $r^{(t)} \in \mathcal{D}_t^{\mathrm{rot}}$ on u. Also, let Z_t be the displacement of j_t with respect to j on the cycle $\mathbb{Z}_p$, and namely $Z_t = (j_t - j + d/2)(\mathrm{mod}\ d) - d/2$ (where the addition/subtraction of $d/2$ is for the purpose of accounting for string boundaries). It is not hard to see that the distribution of Z_t does not depend on the value of j.

The total variation distance between the distributions of $r^{(0)}(u)$ and of $r^{(1)}(u)$ equals to the total variation distance between Z_0 and Z_1. We compute the latter via its complement, i.e. the probability mass that is "common" to the two distributions, which is, formally, $\sum_{z\in[-d,d]} \min_{t\in\{0,1\}} \Pr_{r^{(t)}}[Z_t = z]$.

First, we can compute the probability that $Z_t = 0$, i.e., the symbol u_j remains at position j, as follows:

$$\Pr[Z_t = 0] = \frac{d - \mathbb{E}[L]}{d} = 1 - m \cdot \frac{L_{\min}}{\zeta d},$$

irrespective of the value of $t \in \{0, 1\}$.

Next, consider the case when $Z_t \neq 0$ and note that $\Pr_{r^{(0)}}[Z_0 \neq 0] = \Pr_{r^{(1)}}[Z_1 \neq 0] = m \cdot \frac{L_{\min}}{\zeta d}$. We show that, conditioned on $Z_t \neq 0$, the variable Z_t is uniform over most of its support, denoted S_t. Moreover S_0 and S_1 have almost the same size and almost completely overlap. Formally, we prove the following claim.

Claim 4.3. *There exists a set $S \subset \mathbb{Z} \setminus \{0\}$ satisfying:*

- *There is $\nu > 0$ such that for each $t \in \{0,1\}$ and $z \in S$ we have $\Pr_{r^{(t)}}[Z_t = z] = \nu$; and*
- *For each $t \in \{0,1\}$ we have $\Pr_{r^{(t)}}[Z_t \in S] \geq \frac{m-\xi_1}{m} \cdot \Pr_{r^{(t)}}[Z_t \neq 0]$.*

We first show how Claim 4.3 lets us prove that $\delta_1(u) \leq O(\log\alpha \cdot \frac{L_{\min}}{d})$. Indeed, one can observe that $\delta_1(u)$ is bounded by the probability that $\Pr_{r^{(0)}}[Z_0 \notin S \cup \{0\}] = \Pr_{r^{(1)}}[Z_1 \notin S \cup \{0\}]$, which we can bound as

$$\delta_1(u) \leq 1 - \Pr_{r^{(t)}}[Z_t = 0] - \Pr_{r^{(t)}}[Z_t \in S] \leq \frac{\xi_1}{m} \cdot \Pr_{r^{(t)}}[Z_t \neq 0] = O(\log\alpha) \cdot \frac{L_{\min}}{\zeta d}.$$

Proof of Claim 4.3. We show the claim for $S = \{\pm(1-\epsilon_1)\beta^l L_{\min} \mid l = 1, \ldots, m - \xi_1\}$ and $\nu = \frac{1}{2} \cdot \frac{L_{\min}}{\zeta d}$.

Let us consider the case that $Z_t \neq 0$. Then, the magnitude of the displacement, $|Z_t|$, must be either $\epsilon_t L$ or $(1-\epsilon_t)L$ where $L = \beta^l L_{\min}$ for some $l \in [m]$. In particular, $Z_t \neq 0$ iff the position j falls inside the rotation block of the operation $r^{(t)}$, and either: (i) j falls into the bigger part of size $(1-\epsilon_t)L$ (that does not wrap-around), hence $|Z_t| = \epsilon_t L$; or (ii) j falls into the smaller part of size $\epsilon_t L$ (that does wrap-around), hence $|Z_t| = L - \epsilon_t L = (1-\epsilon_t)L$. Moreover, conditioned on the magnitude of Z_t, the sign of Z_t is equi-probable to be either positive or negative (depending on whether the rotation block rotates to the right or left).

When $t = 0$, we can compute the probability that $|Z_0| = \frac{1}{2}L = \frac{1}{2}\beta^l L_{\min}$ for some $l \in [m]$ as follows. We have $Z_0 = L/2$ when we choose block length $L = \beta^l L_{\min}$, which happens with probability β^{-l}/ζ, and additionally either (i) position j is inside the "bigger" part of the block, of size $(1-\epsilon_0)L = L/2$, and the block moves to right, or (ii) position j is inside the "smaller" part of the block, of size $\epsilon_0 L = L/2$, and the block moves to left. Formally,

$$\Pr_{r^{(0)}}[Z_0 = L/2] = \Pr_{r^{(0)}}[Z_0 = -L/2] = \tfrac{\beta^{-l}}{\zeta} \cdot \tfrac{(1-\epsilon_0)\beta^l L_{\min}}{d} \cdot \tfrac{1}{2} + \tfrac{\beta^{-l}}{\zeta} \cdot \tfrac{\epsilon_0 \beta^l L_{\min}}{d} \cdot \tfrac{1}{2} = \tfrac{L_{\min}}{\zeta d} \cdot \tfrac{1}{2} = \nu.$$

Note that $z = \frac{1}{2}\beta^l L_{\min}$ may be written as $z = (1-\epsilon_1)\beta^{l-1} L_{\min}$ (using Eqn. (6)) and thus $z \in S$ whenever $l \in \{2, \dots m - \xi_1 + 1\}$.

Now let $t = 1$. When $|Z_1| = \epsilon_1 \beta^{l+\xi_1} L_{\min} = (1-\epsilon_1) \cdot \beta^l L_{\min} \in S$ for $l \in \{1, \dots, m - \xi_1\}$ (the equality here is by Eqn. (6)), we again have that

$$\Pr_{r^{(1)}}[Z_1 = \epsilon_1 \beta^{l+\xi_1}] = \tfrac{\beta^{-l-\xi_1}}{\zeta} \cdot \tfrac{(1-\epsilon_1)\beta^{l+\xi_1} L_{\min}}{d} \cdot \tfrac{1}{2} + \tfrac{\beta^{-l}}{\zeta} \cdot \tfrac{\epsilon_1 \beta^l L_{\min}}{d} \cdot \tfrac{1}{2} = \tfrac{L_{\min}}{\zeta d} \cdot \tfrac{1}{2} = \nu.$$

Finally, note that $\Pr_{r^{(t)}}[Z_t \in S] = \sum_{z \in S} \Pr_{r^{(t)}}[Z_t = z] = 2(m - \xi_1) \cdot \nu = \frac{m-\xi_1}{m} \cdot \Pr_{r^{(t)}}[Z_t \neq 0]$. This concludes the proof of Claim 4.3. $\square$

Part 2: $\lambda \geq 2$. When we have $\lambda \geq 2$ non-zero entries in u, the intuition is to group these non-zero entries into one or more "atomic intervals" and then reduce to the case $\lambda = 1$ with the role of "symbol u_j" being replaced by an atomic interval. For example, when there are $\lambda = 2$ non-zero entries in u, most of the block lengths L fall into two categories:

- L is much larger than the distance between the positions of the two non-zero entries — in this case, the two non-zero symbols from u move jointly (atomically) most of the time, and thus the interval connecting the two symbols behaves roughly as the "symbol u_j" in the $\lambda = 1$ scenario;
- L is much smaller than the distance between the two positions — in this case, each of the two non-zero entries can be treated independently as in $\lambda = 1$ case, and we lose only a factor of λ (by "union bound").

Furthermore, we can bound the number of values of L that do not satisfy one of the above properties. A relatively straight-forward bound is $O(\lambda^2)$ (all pair-wise distances between the non-zero entries), times $O(\xi_1)$ (the same extra factor as in the $\lambda = 1$ case). This analysis would give a bound of $\delta_\lambda(u) \leq O(\lambda^3 \log \alpha \cdot \frac{L_{\min}}{d})$. In the sequel we obtain a stronger bound, with only a linear dependence on λ, using a more careful analysis. (For the impact of a weaker bound see the calculation in Section 3.3.)

More generally, we partition the non-zero entries of u such that each part consists of "nearby" entries, while the parts are "far" amongst themselves. We then view each part as a contiguous *A-interval* (stands for atomic interval). Once we manage such an approximation, we have several A-intervals (at most λ), and we expect each one to move atomically: all non-zero entries from the same A-interval will move the same direction by the same displacement most of the time. The main challenge lies in the fact that the notion of nearby entries depends on the length L of the rotation block, and we say two non-zero entries are nearby if their positions differ by at most L. Thus, for each possible block length L, we have a possibly different partition of entries into A-intervals (partitions are progressively coarser with bigger L). The main technical work is to analyze the structure of these A-intervals over all lengths L.

We proceed with a complete proof below. For a block length $L = \beta^l L_{\min}$, we define the graph G_L as follows. G_L is an undirected graph on λ vertices, where each vertex corresponds to a non-zero entry in u. For convenience, we use the term "entry" when we refer to the position of a non-zero entry of u, and equivalently a vertex of G_L (in contrast, we will use the term node for another graph structure defined later). We connect two entries $i, j \in [d]$ if $|i-j|^* \leq L$, where $|i-j|^* = \min\{|i-j|, d - |i-j|\}$ computes distance on the d-cycle. For a graph G_L, we focus on its connected components, which may be viewed as intervals in $\mathbb{Z}_d$. Specifically, to each connected component $C \subset V$ we assign the interval $I(C)$, an interval defined as the minimal interval (with wrap-around) on $\mathbb{Z}_d$ that contains all entries in C. Overloading the notation, we write an interval $I(C) = [i,j]$ to mean that $I(C) = \{i, i+1, \ldots, j\}$ if $i \leq j$ and $I(C) = \{i, i+1, \ldots, d, 1, 2, \ldots j\}$ if $j < i$. The length of interval $I = [i,j]$ is $\text{len}(I) = |I| = (j - i + 1) (\text{mod } d)$. Note that, for every connected component C, every two *consecutive* entries in $I(C)$ are at distance at most L; thus, the length of any interval $I(C)$ can be at most $L \cdot \lambda < d^{0.99}$; also if $I(C) = [i,j]$ then both i and j are non-zero entries of u.

An *A-interval* is then an interval $I(C)$ that corresponds to some connected component C. Each block length L induces potentially different graph G_L, that in turn induces different set of A-intervals. The following observation relates A-intervals induced by different G_L's.

Observation 4.4. *If two entries are in the same A-interval (equivalently, connected component) in G_L for some L, then there are also in the same A-interval in $G_{L'}$ for any $L' \geq L$.*

We use this observation to define a forest on all the A-intervals, as follows. The forest consists of m levels, where nodes at level $l \in [m]$ correspond to the A-intervals for $L = \beta^l L_{\min}$ (i.e. the connected components in G_L). For a forest node v at level l we write $I(v)$ for the corresponding A-interval. The edges in the forest are defined as follows: for two forest nodes v_1, v_2 on two consecutive levels, l and $l+1$ respectively, we connect v_1 to v_2 iff $I(v_1) \subseteq I(v_2)$. This construction is well-defined due to Observation 4.4. Nodes at level 1 will be called *leaves*. Notice that every forest node at level $l > 1$ indeed has at least one edge to a node at level $l - 1$, i.e. non-leaf nodes have at least one child. Let $n_l \in [\lambda]$ be the number of nodes at level l.

We now wish to bound the error incurred by considering an A-interval to be an atomic object. Specifically, a too long A-interval is likely to move not atomically, in the sense that the interval is "cut" by the rotation block. We bound the error of our "approximation" using the probability that a random position $s \in [d]$ (one of the two block boundaries) falls inside these A-intervals at a random level l. The latter probability is proportional to expected sum of lengths of the A-intervals of G_L, when we choose the block length L randomly according to the distribution $\mathcal{D}_t^{\text{rot}}$.

Claim 4.5. *Let $s \in [d]$ be chosen uniformly at random and let $l \in [m]$ be chosen randomly with probability β^{-l}/ζ. Then,*

$$\Pr_{s,l}\Big[s \textit{ is inside one of the A-intervals at level } l\Big] \leq O\Big(\lambda \frac{L_{\min}}{d}\Big).$$

Proof. Consider any two *consecutive* non-zero entries of u and let J be the interval between them (with wrap-around), including one of the endpoints, say the left one. We compute next the probability that s is contained in this interval J, and interval J is contained in an A-interval $I(v)$ for a forest node v at level l. Note that summing this probability over all λ intervals J gives the final quantity we want.

By definition, an interval J is inside an A-interval at level l iff $|J| \leq \beta^l L_{\min}$. Thus, for a fixed J, the probability that both $s \in J$ and J is contained in an A-interval at level l is at most

$$\frac{|J|}{d} \cdot \sum_{l \in [m]:\ |J| \leq \beta^l L_{\min}} \frac{\beta^{-l}}{\zeta} \leq \frac{|J|}{d} \cdot \frac{L_{\min}}{\zeta \cdot |J|} \cdot \frac{1}{1 - \beta^{-1}} \leq O\left(\frac{L_{\min}}{d}\right).$$

We have exactly λ such intervals J, and thus the total contribution is $O(\lambda \frac{L_{\min}}{d})$. ∎

We now continue with computing the total variation distance $\delta_\lambda(u)$ between $r^{(0)}(u)$ and $r^{(1)}(u)$ where $r^{(0)} \in \mathcal{D}_0^{\text{rot}}$ and $r^{(1)} \in \mathcal{D}_1^{\text{rot}}$. As in part one ($\lambda = 1$), we will bound the total variation distance between them by estimating the probability mass "common" to the two distributions.

First we compute the probability that all non-zero entries of u stay put (as in part one).

Claim 4.6. *For each $t \in \{0,1\}$, we have that*

$$\Pr_{r^{(t)}}[r^{(t)}(u) = u] \geq 1 - O\left(\lambda \frac{L_{\min}}{d}\right) - \sum_{l=1}^{m} n_l \cdot \frac{L_{\min}}{\zeta d}.$$

Proof. The complement event is that at least one non-zero entry of u is displaced. Whenever it occurs, at least one of the following holds:

- Left or right endpoint of the rotation block belongs to an A-interval induced by G_L; or else
- The rotation block contains inside it an entire A-interval induced by G_L.

The probability of the first event is bounded by, using Claim 4.5:

$$2 \Pr_{s,L}[s \text{ is inside one of the A-intervals at level } l] \leq O\left(\lambda \frac{L_{\min}}{d}\right).$$

The probability of the second event can be bounded by the probability that the rotation block includes the leftmost endpoint of some A-interval at level l:

$$\sum_{l=1}^{m} \frac{\beta^{-l}}{\zeta} \sum_{v \text{ at level } l} \Pr_s\Big[\text{left endpoint of } I(v) \text{ is inside } [s, s+L-1]\Big] \leq$$
$$\leq \sum_{l=1}^{m} \frac{\beta^{-l}}{\zeta} \cdot \frac{n_l \cdot L}{d} = \sum_{l=1}^{m} n_l \frac{L_{\min}}{\zeta d}$$

The claim follows from the last two inequalities by applying a union and then considering the complement event. ∎

We now prove a claim that should be seen as the analogue of Claim 4.3 from part one, which characterizes the common weight of the distributions of $r^{(0)}(u)$ and $r^{(0)}(u)$ when some entries (more precisely, A-intervals) move. In contrast to part one, here we have to also consider the case when an A-interval does not behave atomically, i.e., when the rotation block intersects the A-interval of some node v at a level $l \in [m]$. This will contribute some additional error term that depends on the length of the interval $I(v)$, and which we will bound using Claim 4.5.

Let us define the random variable $Z_t(I)$, for an interval $I = I(v)$ corresponding to a forest node v, and $t \in \{0,1\}$. $Z_t(I)$ denotes the (position) displacement of the entries from the interval I under rotation $r^{(t)} \in \mathcal{D}_1^{\text{rot}}$ when the interval I moves atomically and no entry outside I moves. We set $Z_t(I) = \perp$ if the interval I does not move atomically and/or some other entry outside I moves as well under $r^{(t)}$.

Claim 4.7. *There exists a set $S \subset \mathbb{Z} \setminus \{0\}$ satisfying:*

- *For each interval $I = I(v)$ corresponding to a forest node v at level $l^* \in \{\xi_1+1, \dots m\}$, and for each $t \in \{0,1\}$ and $z \in S$,*

$$\Pr_{r^{(t)}}\left[Z_t(I) = z\right] \geq \frac{1}{2} \cdot \frac{L_{\min}}{\zeta d} - \beta^{-(l^*-\xi_1)} \cdot \frac{2\,\mathrm{len}(I)}{\zeta d};$$

- *Call two intervals $I(v)$ and $I(v')$ distinct if they have at least one distinct endpoint; then*

$$\sum_{z\in S} \sum_{\text{distinct } I=I(v)} \min_{t\in\{0,1\}} \left\{\Pr_{r^{(t)}}\left[Z_t(I) = z\right]\right\} \geq \sum_{l=\xi_1+1}^{m} n_l \frac{L_{\min}}{\zeta d} - O\left(\lambda \frac{L_{\min}}{d}\right).$$

Proof. We show the claim for $S = \left\{\pm \left\lfloor (1-\epsilon_1)\beta^l L_{\min} \right\rfloor \mid l = 1, \dots, m-\xi_1\right\}$.

Fix an interval $I = I(v)$ for a node at level $l^* \geq \xi_1 + 1$. Consider the displacement $z = \left\lfloor \epsilon_1 \beta^{l^*} L_{\min} \right\rfloor = \left\lfloor (1-\epsilon_1)\beta^{l^*-\xi_1} L_{\min} \right\rfloor \in S$ (the equality is again by Eqn. (6)). We now bound $\Pr[Z_1(I) = z]$, namely the probability that all the entries in $I(v)$ are moved (atomically) z positions to the right (and all the other entries stay put), under the distribution $\mathcal{D}_1^{\mathrm{rot}}$. We have $\Pr[Z_1(I) = z]$ when either: i) $l = l^*$, interval I is completely inside the "bigger" part of the block, of size $\left\lfloor (1-\epsilon_1)\beta^l L_{\min} \right\rfloor$, and the block moves to right, or ii) $l = l^* - \xi_1$, interval I is completely inside the "smaller" part of the block, of size $\left\lfloor \epsilon_1 \beta^{l-\xi_1} L_{\min} \right\rfloor$, and the block moves to left. Note that in both cases all entries outside I stay put as they are at (position) distance at least $\beta^{l^*} L_{\min} + 1$ from I and thus cannot be inside the rotation block. Formally,

$$\begin{aligned}
\Pr_{r^{(1)}}\left[Z_1(I) = z\right] &= \Pr_{r^{(1)}=(\tilde{R}_{s,L})^{\epsilon_1 L},\ L=\beta^l L_{\min}} \left[Z_1(I) = z,\ l = l^*,\ \tilde{R} = \overrightarrow{R}\right] \\
&\quad + \Pr_{r^{(1)}=(\tilde{R}_{s,L})^{\epsilon_1 L},\ L=\beta^l L_{\min}} \left[Z_1(I) = z,\ l = l^* - \xi_1,\ \tilde{R} = \overleftarrow{R}\right] \\
&\geq \frac{\beta^{-l^*}}{\zeta} \cdot \frac{1}{2} \cdot \frac{(1-\epsilon_1)\beta^{l^*} L_{\min} - 1 - \mathrm{len}(I)}{d} \\
&\quad + \frac{\beta^{-(l^*-\xi_1)}}{\zeta} \cdot \frac{1}{2} \cdot \frac{\epsilon_1 \beta^{l^*-\xi_1} L_{\min} - 1 - \mathrm{len}(I)}{d} \\
&\geq \frac{1}{2} \frac{L_{\min}}{\zeta d} - \frac{\beta^{-(l^*-\xi_1)}}{\zeta} \cdot \frac{2\,\mathrm{len}(I)}{d}
\end{aligned} \tag{8}$$

Similarly, we can give the exact same lower bound for each of the following four events: $Z_1(I) = \pm z$ and $Z_0(I) = \pm z$.

We can now bound the probability mass that is common to the two distributions $r^{(0)}(u)$ and $r^{(1)}(u)$ for the events that there is a distinct interval I such that $Z_t(I) = z$ for some $z \in S$:

$$\begin{aligned}
&\sum_{z\in S} \sum_{\text{distinct } I=I(v)} \min_{t\in\{0,1\}} \left\{\Pr_{r^{(t)}}\left[Z_t(I) = z\right]\right\} \\
&\qquad \geq \sum_{l=\xi_1+1}^{m} \sum_{v \text{ at level } l} \Pr\left[Z_t(I(v)) \in \left\{\pm \left\lfloor \epsilon_1 \beta^l L_{\min} \right\rfloor\right\}\right]
\end{aligned} \tag{9}$$

because, for each node v at level $l^* \geq \xi_1 + 1$, we can consider the interval $I = I(v)$ and the displacement of $z = z(v) = \left\lfloor \epsilon_1 \beta^{l^*} L_{\min} \right\rfloor \in S$. Then all the events $Z_t(I(v)) = \pm z(v)$ are mutually exclusive (over the choice of such v), and hence we obtain the sum from Eqn. (9). Furthermore, using Eqn. (8), we obtain:

$$\begin{aligned}
\sum_{z\in S}\sum_{\text{distinct } I=I(v)} \min_{t\in\{0,1\}}\left\{\Pr_{r^{(t)}}\left[Z_t(I)=z\right]\right\} \\
&\geq \sum_{l^*=\xi_1+1}^{m} 2 \sum_{v \text{ at level } l^*}\left(\frac{1}{2}\cdot\frac{L_{\min}}{\zeta d} - \frac{\beta^{-(l^*-\xi_1)}}{\zeta}\cdot\frac{2\,\mathrm{len}(I(v))}{d}\right) \\
&\geq \sum_{l^*=\xi_1+1}^{m} n_{l^*}\frac{L_{\min}}{\zeta d} - 4\sum_{l^*=1}^{m}\sum_{v \text{ at level } l^*}\frac{\beta^{-l^*}}{\zeta}\cdot\frac{\mathrm{len}(I(v))}{d}, \qquad (10)
\end{aligned}$$

where we remind that n_l is the number of nodes at level l^*. The last inequality follows from the fact that, for each interval $I=I(v)$ of a node v at level $l^*\geq\xi_1+1$, we can charge $\mathrm{len}(I(v))$ to lengths of the intervals of the descendants of v at level $l^*-\xi_1$.

Finally, observe that the last term in Eqn. (10), namely $\sum_{l^*}\sum_v \frac{\beta^{-l^*}}{\zeta}\cdot\frac{\mathrm{len}(I(v))}{d}$, is equal precisely to the probability that a random position s falls into an A-interval at level l^*, where l^* is chosen at random according to the distribution $l^*=l$ with probability β^{-l}/ζ. Thus we can use Claim 4.5 to bound it from above,

$$\sum_{l^*=1}^{m}\sum_{v \text{ at level } l^*}\frac{\beta^{-l^*}}{\zeta}\cdot\frac{\mathrm{len}(I(v))}{d} \leq O(\lambda)\cdot\frac{L_{\min}}{\zeta d},$$

which together with Eqn. (10) completes the proof of Claim 4.7. □

To summarize, the total probability mass that we accounted to be common for $t=0$ and $t=1$ is the sum of (our lower bounds on) the probability that all entries stay put, plus the probability that exactly one distinct interval $I=I(v)$ is displaced by precisely $z\in S$ positions. Combining Claims 4.6 and 4.7, and using the trivial bound of $n_l\leq\lambda$ for all $l\in[m]$, we obtain:

$$1-\delta_\lambda(u)\geq 1-O\left(\lambda\cdot\frac{L_{\min}}{d}\right)-\sum_{l=1}^{m} n_l\cdot\frac{L_{\min}}{\zeta d}+\sum_{l=\xi_1+1}^{m} n_l\frac{L_{\min}}{\zeta d}-O\left(\lambda\frac{L_{\min}}{d}\right)\geq 1-O(\lambda\xi_1)\cdot\frac{L_{\min}}{\zeta d}.$$

Finally, using the fact that $\xi_1=O(\log\alpha)$, we conclude Eqn. (7), which completes the proof of Lemma 4.2. □

4.2 Statistical Closeness of the Distributions $\tilde{\mu}_t$ and μ_t

We prove Lemma 3.9 (b) via the following lemma.

Lemma 4.8. *For every $t\in\{0,1\}$, the total variation distance between $\tilde{\mu}_t$ and μ_t is at most $d^{-\Omega(1)}$.*

Proof. First we recall that $\tilde{\mu}_t$ is equal to the distribution μ conditioned on the fact that the generated pair $(x,y)\in\mu$ is legal, i.e., both x and y are permutations and (x,y) are far or close for $t=0$ or $t=1$ respectively. Since both x and y are random from $\mathbb{Z}_p^d$, and $p>d^3$, then x and y are both permutations with probability at least $1-O(1/d)$.

Thus, the total variation distance between $\tilde{\mu}_0$ and μ_0 is at most $\Pr_{\mu_0}[\mathrm{ed}(x,y)\leq R]+O(1/d)$. Similarly, total variation distance between $\tilde{\mu}_1$ and μ_1 is at most $\Pr_{\mu_1}[\mathrm{ed}(x,y)>R/\alpha]+O(1/d)$.

Thus, it suffices to prove that $\Pr_{\mu_0}[\mathrm{ed}(x,y) \le R] \le d^{-\Omega(1)}$ and $\Pr_{\mu_1}[\mathrm{ed}(x,y) > R/\alpha] \le d^{-\Omega(1)}$. Remember that x is chosen at random, then $z = x + N_\rho$, and y is obtained from z via a sequence of rotation operations.

We choose the constant $C_2 = 20\zeta/\epsilon_0 = 40\zeta$, and condition on the event that x, y, and z are all permutations, which happens with probability $\ge 1 - O(1/d)$. We can describe the distribution μ_t also as follows. Start with a permutation z, and let x be the permutation obtained by modifying every coordinate in z to a new symbol independently with probability $1-\rho$. We may assume, without loss of generality (by renaming symbols), that z is the identity permutation of length d, i.e. for all $i \in [d]$ we have $z(i) = i$, and furthermore with probability ρ we have $x(i) = z(i)$ and $x(i) = i + d$ otherwise. Next, let y be the permutation obtained from z by applying w random rotation operations chosen from $\mathcal{D}_t^{\mathrm{rot}}$.

It will then suffice to prove the following two claims.

Claim 4.9. *For both $t \in \{0,1\}$,*

$$\Pr_{\mu_t}\Big[\mathrm{ed}(x,z) \le 2\epsilon_1 R\Big] \ge 1 - e^{-d^{\Omega(1)}}.$$

Claim 4.10. *For both $t \in \{0,1\}$,*

$$\Pr_{\mu_t}\left[0.1 \le \frac{\mathrm{ed}(z,y)}{R \cdot C_2\epsilon_t/\zeta} \le 10\right] \ge 1 - d^{-\Omega(1)}.$$

We can now obtain the lemma statement from the above two claims, using a union bound, and applying the triangle inequality $|\,\mathrm{ed}(x,y) - \mathrm{ed}(z,y)| \le \mathrm{ed}(x,z)$ (see also Figure 2). Indeed, we obtain (i) that for the distribution μ_0, with high probability, $\mathrm{ed}(x,y) \ge (0.1C_2\epsilon_0/\zeta - 2\epsilon_1)R = (2-2\epsilon_1)R > R$; and (ii) that for the distribution μ_1, with high probability, $\mathrm{ed}(x,y) \le (10C_2\epsilon_1/\zeta + 2\epsilon_1)R = 402\epsilon_1 R \le \frac{804}{C_1\alpha} \cdot R < R/\alpha$.

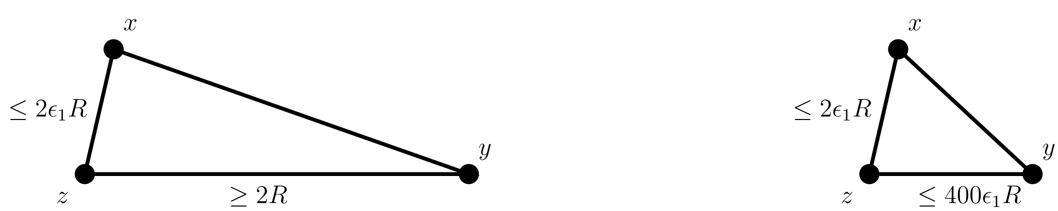

Figure 2: The relative positions of x, y, z under the distributions μ_0 and μ_1 respectively.

It remains to prove Claims 4.9 and 4.10.

Proof of Claim 4.9. One can verify that $\mathrm{ed}(x,z)$ is upper bounded by the number of substitutions performed when constructing x from z. This number of substitutions may be bounded using a straightforward Chernoff bound:

Theorem 4.11 (Chernoff Bound, cf. [MR95])**.** *Let X_i, $i = 1 \dots d$, be i.i.d. random Poisson trials with $\mathbb{E}[X_i] = q$ for some $q \in (0,1)$. Then* $\Pr[|\sum X_i - qd| > \frac{1}{2}qd] \le 2e^{-qd/10}$.

In our case probability of substitution is $q = (1-\rho)(1-1/p)$, where the second factor is the probability that the substituted symbol is different from the original symbol. Since $\rho = 1 - \epsilon_1 R/d$, we get

$$\Pr_{\mu_t}\Big[\mathrm{ed}(x,z) \le 2\epsilon_1 R\Big] \ge 1 - e^{-\Omega(\epsilon_1 R)} \ge 1 - e^{-d^{\Omega(1)}}.$$

$\square$

Proof of Claim 4.10. We first show an upper bound on $\mathrm{ed}(z, y)$ by analyzing the sum of magnitudes of all the rotation operations. Recall that there are w rotation operations; a single rotation operation works on a block of (random) length $L = \beta^l L_{\min}$, and incurs edit distance at most (in fact, exactly) $2\lfloor \epsilon_t L \rfloor$. For $l \in [m]$, let the random variable Z_l denote the number of rotation operations in which the block length equals $\beta^l L_{\min}$. Observe that Z_l has Binomial distribution $B(w, \frac{\beta^{-l}}{\zeta})$ and its expectation is $\mathbb{E}[Z_l] = w \cdot \frac{\beta^{-l}}{\zeta} \geq \frac{C_2 R}{m L_{\min}} \cdot \frac{d^{-0.01}}{\zeta} \geq d^{\Omega(1)}$. By a straightforward Chernoff bound (Theorem 4.11),

$$\Pr\Big[Z_l \geq 2\,\mathbb{E}[Z_l]\Big] \leq e^{-\Omega(\mathbb{E}[Z_l])} \leq e^{-d^{\Omega(1)}}.$$

Taking a union bound over these events for $l = 1, \dots, m$, we conclude that with high probability

$$\mathrm{ed}(z, y) \leq \sum_{l=1}^{m} (2w \frac{\beta^{-l}}{\zeta} \cdot 2\epsilon_t \beta^l L_{\min}) = \frac{4 C_2 \epsilon_t R}{\zeta}.$$

We proceed to show a lower bound on $\mathrm{ed}(z, y)$, by counting inversions, i.e. pairs of symbols $(a_1, b_1), \dots, (a_k, b_k)$ such that each a_j appears before b_j in z, but a_j appears after b_j in y. It is easy to verify that if the inversions are disjoint, in the sense that the symbols $a_1, b_1, \dots, a_k, b_k$ are all distinct, then $\mathrm{ed}(z, y) \geq k$ (because in every alignment of z with y, for each $j = 1, \dots, k$, at least one of a_j, b_j must incur an edit operation). For each of the w rotation operations we take $\lfloor \epsilon_t L \rfloor$ pairs — simply take the $\lfloor \epsilon_t L \rfloor$ symbols that were at the beginning of the block and match them to the $\lfloor \epsilon_t L \rfloor$ symbols that were at the end of the block. It follows, using Chernoff bounds as above, that with probability at least $1 - e^{-d^{\Omega(1)}}$ this process picks at least $\frac{1}{2} \cdot \frac{C_2 \epsilon_t R}{\zeta}$ pairs of symbols, but this count might include repetitions. Furthermore, a pair "inverted" in one rotation operation may be inverted back by another rotation. To mitigate this concern, fix a pair (a, b) taken at some j-th rotation operation. The probability that symbol a was inside a rotated block in at least one other rotation is at most (using the independence between rotations and a union bound)

$$(w-1) \sum_{l=1}^{m} \left(\frac{\beta^{-l}}{\zeta} \cdot \frac{\beta^l L_{\min}}{d} \right) < \frac{w m L_{\min}}{\zeta d} = \frac{C_2 R}{\zeta d}.$$

A similar argument applies to symbol b, and clearly if both a and b were not inside a rotated block of any of the other $w-1$ rotations, then either (a, b) or (b, a) is an inversion between z and y. It remains to apply a union bound over the $\frac{C_2 \epsilon_t R}{2\zeta}$ pairs of symbols the above process produces, and indeed the probability that at least one of them fails is at most

$$2 \cdot \frac{C_2 \epsilon_t R}{2\zeta} \cdot \frac{C_2 R}{\zeta d} \leq O\left(\frac{R^2}{d}\right) \leq d^{-\Omega(1)}.$$

We conclude that with probability at least $1 - d^{-\Omega(1)}$, the above process produces $\frac{C_2 \epsilon_t R}{2\zeta}$ disjoint inversions, and thus $\mathrm{ed}(y, z) \geq \frac{C_2 \epsilon_t R}{2\zeta}$. This completes the proof of Claim 4.10. $\square$

We thus finalized the proof of Lemma 4.8. $\square$

5 Reducing Ulam to Edit Distance on 0-1 strings

In this section, we prove Theorem 1.2. We make no attempt to optimize the constants.

The basic intuition behind this proof is quite simple. The first part (the upper bound on $\mathrm{ed}(\pi(P), \pi(Q)))$ is immediate, and the main challenge is to prove the lower bound on $\mathrm{ed}(\pi(P), \pi(Q))$. To prove the lower bound, we proceed by ruling out all "potential certificates" that $\mathrm{ed}(\pi(P), \pi(Q))$ is small. Specifically, a "potential certificate" is a potential fixed alignment between $\pi(P)$ and $\pi(Q)$ of low cost, i.e. a fixed monotone mapping that matches monotonically all but at most $\frac{1}{100}\,\mathrm{ed}(P, Q)$ of the positions in $\pi(P)$ and $\pi(Q)$. We then analyze the probability that such an alignment is "successful", in the sense that every pair of positions that is matched under the potential alignment has equal symbols. Indeed, we show this probability is exponentially small because many of the pairs matched are independent coin tosses. We then apply a union bound over all potential alignment of small cost. Although a direct union bound is not sufficient (there are too many potential alignments to consider), we reduce the number of potential low-cost alignments by partitioning the set of all such alignments into a smaller number of groups of "equivalent alignments".

We proceed to set up some basic terminology and notation and to provide two lemmas that will be used in the proof of the theorem.

For two permutations P, Q, we say that an index (position) $i \in [d]$ in P is *missing* (from Q) if the symbol $P(i)$ does not appear inside Q.[4] We say that a pair of indices $\{i, j\} \subseteq [d]$ is an *inversion* (in P with respect to Q) if the two characters $P(i), P(j)$ appear in Q but in the opposite relative order than in P, formally given by $(i - j)(Q^{-1}(P(i)) - Q^{-1}(P(j))) < 0$. We also say that index j is *inverted* with respect to i.

An *alignment* of two strings $x, y \in \Sigma^d$ is a mapping $A : [d] \mapsto [d] \cup \{\perp\}$ that is monotonically increasing on $A^{-1}([d]) = \{i \in [d] \mid A(i) \in [d]\}$. Intuitively, A models a *candidate* longest common subsequence between x and y, and thus it maps indices in x to their respective indices in y and takes the value $\perp$ when there is no respective index in y (i.e., the respective position of x is not in the candidate subsequence). A *disagreement* in the alignment A is an index $i \in [d]$ for which $A(i) \neq \perp$ and $x(i) \neq y(A(i))$. The alignment is called *successful* if it has no disagreements. The *cost* of an alignment is the number of positions in x (equivalently, in y) that are not mapped to a respective index in the other string, namely $|A^{-1}(\perp)| = d - |A^{-1}([d])| = d - |A([d])|$, where $A([d]) = \{A(i) \mid i \in [d]\}$. It is easy to verify that for all x, y,

$$\tfrac{1}{2}\,\mathrm{ed}(x, y) \leq \min_A \mathrm{cost}(A) \leq \mathrm{ed}(x, y), \tag{11}$$

where the minimum is taken over all successful alignments A.

In the following lemma, we present a property of strings P and Q that will let us prove that, for a fixed potential alignment between $\pi(P)$ and $\pi(Q)$, the probability of the alignment being successful is very small.

Lemma 5.1. *Let P, Q be two permutations of length d that contain the same symbols, i.e. $P([d]) = Q([d])$. Then there exists a collection of $m \geq \mathrm{ed}(P, Q)/4$ inversions $\{i_1, j_1\}, \ldots, \{i_m, j_m\}$ such that $i_1, j_1, \ldots, i_m, j_m$ are all distinct.*

Proof. Fix P, Q. Define an (undirected) graph G with vertex set $[d]$ and an edge $\{i, j\}$ whenever $\{i, j\}$ is an inversion. Let $E^* \subseteq E(G)$ be a matching in G (i.e. no two edges in E^* share an endpoint)

[4] Remember that we have defined a permutation P as a string with a large alphabet where every symbol appears at most once.

that is maximal with respect to containment. Observe that E^* is a collection of inversions whose indices are all distinct (as desired), and it only remains to bound $m = |E^*|$ from below. Following [SU04], we achieve the latter using the well-known relation between maximal matching and vertex-cover.[5]

Let V^* be the set of vertices incident to any edge in E^*, thus $|V^*| = 2|E^*|$. Clearly, V^* is a vertex-cover of G, namely every edge (inversion) must have at least one endpoint in V^*. It follows that $V \setminus V^*$ contains no edges (inversions), and thus immediately yields a successful alignment A between P and Q. Formally, the subsequence of P obtained by removing the positions V^* is also a subsequence of Q, and A is the monotone map matching them. Thus, $A(i) = \perp$ if and only if $i \in V^*$ and $\operatorname{cost}(A) = 2|E^*|$. Finally, using (11) we get that $m = |E^*| = \frac{1}{2}\operatorname{cost}(A) \geq \frac{1}{4}\operatorname{ed}(P, Q)$. □

We now give a lemma that essentially lets us partition all potential alignments into a small number of groups of equivalent alignments.

Lemma 5.2. *Let P, Q be two permutations of length d. Fix $0 < \gamma < 1/2$ and a subset $S \subseteq [d]$. For an alignment A of P and Q (not necessarily successful), let $A_{|S} : S \to [d] \cup \{\perp\}$ be a function that is equal to the function A on the domain S. Define*

$$F = \{A_{|S} \mid A \text{ is an alignment of } P \text{ and } Q \text{ with } \operatorname{cost}(A) \leq \gamma|S|\}.$$

Then $|F| \leq (3e/\gamma)^{2\gamma|S|}$.

Proof. Let us denote $s = |S|$. An alignment of P with Q of cost at most γs can be described as deleting exactly γs symbols from P and exactly γs symbols from Q. (We assume here for simplicity that γs is an integer; otherwise, we round it up and change constants accordingly.) Clearly, we can bound $|F|$ by the number of such alignments between P and Q, namely $|F| \leq \binom{d}{\gamma s}\binom{d}{\gamma s}$, but we aim to get a bound that depends on $s = |S|$ and not on d, by more carefully counting restrictions $A_{|S}$.

An alignment A of P with Q of cost at most γs can be described as first deleting exactly γs characters from P and then inserting into the resulting string exactly γs characters. Observe that $A_{|S}$ is completely determined from the following information: (a) which positions in S are deleted; (b) how many characters are deleted between every two successive indices in S; and (c) how many characters are inserted between every two successive indices in S. (When we say two successive indices in S, it should be interpreted to include also 0 and $d+1$ as indices in S, and in particular (b) describes also how many characters before the first index in S are deleted from P.) Indeed, for each $i \in S$, data (a) determines whether $A(i) = \perp$. If $A(i) \neq \perp$, then $A(i) = i - d_i + a_i$ where d_i is the total number of deletions among indices $1, \ldots, i-1$, which can be determined from data (a) and (b), and a_i is the total number of insertions before position i, which can be determined from data (c).

It remains to upper bound the number of possible outcomes to data (a)–(c). Clearly, the outcomes for (a) and (b) together can be upper bounded by the number of outcomes of throwing γs indistinguishable balls into $2s+2$ bins (a bin per element in S which may get at most one ball, a bin per each interval between elements in S and one extra bin to account for case when the cost is strictly less than γs). This upper bound is equal to $\binom{2s+2+\gamma s}{\gamma s}$ possible outcomes. The outcomes of data (c) can be similarly upper bounded by $\binom{s+1+\gamma s}{\gamma s}$. Together, we obtain that

$$|F| \leq \binom{2s+2+\gamma s}{\gamma s}\binom{s+1+\gamma s}{\gamma s} \leq \Big(\frac{e(2+2\gamma)}{\gamma}\Big)^{\gamma s}\Big(\frac{e(1+2\gamma)}{\gamma}\Big)^{\gamma s} \leq \Big(\frac{3e}{\gamma}\Big)^{2\gamma s},$$

[5] Another proof may be obtained using the $O(1)$-approximation in [GJKK07, Theorem 3.3].

which proves the lemma. $\square$

Having established the two lemmas, we proceed to prove the theorem, which states that with high probability, $\Omega(\mathrm{ed}(P,Q)) \le \mathrm{ed}(\pi(P),\pi(Q)) \le \mathrm{ed}(P,Q)$.

Proof of Theorem 1.2. Fix two permutations P and Q of length d. The inequality $\mathrm{ed}(\pi(P),\pi(Q)) \le \mathrm{ed}(P,Q)$ follows immediately from the observation that every sequence of edit operations to transform P into Q can be applied also to transform $\pi(P)$ and $\pi(Q)$. It thus remains to prove the other direction. Assume for now that P and Q use the same symbols, i.e. $P([d]) = Q([d])$. We will later explain how the general case follows using a similar argument.

Apply Lemma 5.1 to P,Q, and extract $m \ge \mathrm{ed}(P,Q)/4$ inversions $\{i_1,j_1\},\ldots,\{i_m,j_m\}$ such that $i_1,j_1,\ldots,i_m,j_m$ are all distinct. Define $S = \{i_1,j_1,\ldots,i_m,j_m\}$, hence $|S| = 2m$. Fix $\gamma = 1/100$ and let F be defined as in Lemma 5.2 (with respect to our P,Q,S and γ). By that lemma, $|F| \le (3e/\gamma)^{2\gamma|S|} = (3e/\gamma)^{4\gamma m}$. Note that F does not depend on π.

For every $f \in F$, let $\mathcal{E}_f$ be the event that all $i \in S$ with $f(i) \ne \perp$ satisfy $\pi(P(i)) = \pi(Q(f(i)))$. That is

$$\mathcal{E}_f = \bigwedge_{i \in S \setminus f^{-1}(\perp)} \{\pi(P(i)) = \pi(Q(f(i))\}.$$

We claim that

$$\Pr\Big[\mathrm{ed}(\pi(P),\pi(Q)) < \tfrac{1}{2}\gamma \cdot \mathrm{ed}(P,Q)\Big] \le \Pr\Big[\bigcup_{f \in F} \mathcal{E}_f\Big]. \tag{12}$$

To prove the claim we show that $\mathrm{ed}(\pi(P),\pi(Q)) < \frac{1}{2}\gamma \cdot \mathrm{ed}(P,Q)$ implies that at least one of the events $\mathcal{E}_f$ happens. Indeed, suppose there is a successful alignment A between $\pi(P)$ and $\pi(Q)$ that has cost $\frac{1}{2}\gamma \cdot \mathrm{ed}(P,Q) \le 2\gamma m = \gamma|S|$. Since A is successful, for all $i \in S \setminus A^{-1}(\perp)$, we must have $\pi(P(i)) = \pi(Q(A(i))$. Furthermore, we can think of A as an alignment between P and Q, and then by definition, its restriction $A_{|S}$ must be in F.

We now bound $\Pr[\mathcal{E}_f]$ for any fixed $f \in F$, i.e. $f = A_{|S}$ for some alignment A of cost at most $\gamma|S| = 2\gamma m$. Since S is the union of m inversions $\{i_t,j_t\}$ with distinct indices, for at least $(1-2\gamma)m$ of these inversions, we have that $f(i_t), f(j_t) \ne \perp$. For every such inversion $\{i_t,j_t\}$, it cannot be that both $P(i_t) = Q(f(i_t))$ and $P(j_t) = Q(f(j_t))$ (as that would contradict the fact that the alignment A is increasing). Let $a_t \ne b_t$ denote these two differing symbols (i.e. either $a_t = P(i_t)$, $b_t = Q(f(i_t))$ or $a_t = P(j_t)$, $b_t = Q(f(j_t))$), the event $\mathcal{E}_f$ can only occur if $\pi(a_t) = \pi(b_t)$. We thus obtain $(1-2\gamma)m$ requirements of the form $\pi(a_t) = \pi(b_t)$. These requirements have distinct symbols a_t in their left-hand sides (since they come from distinct positions in P), and similarly, the right-hand sides contain distinct symbols b_t. Altogether, every symbol in Σ may appear in at most two requirements, and thus we can extract (say greedily) a subcollection containing at least one half of these requirements, namely, at least $(1-2\gamma)m/2 \ge m/4$ requirements, such that every symbol appears in at most one requirement. Since π is a random function, the probability that all these requirements are satisfied is at most $2^{-m/4}$, and we conclude that $\Pr[\mathcal{E}_f] \le 2^{-m/4}$.

To complete the proof of the theorem, we plug the last bound into (12) and use a union bound and Lemma 5.2, which altogether gives

$$\Pr\Big[\mathrm{ed}(\pi(P),\pi(Q)) < \tfrac{1}{2}\gamma \cdot \mathrm{ed}(P,Q)\Big] \le (3e/\gamma)^{4\gamma m} \cdot 2^{-m/4} \le 2^{-m/8}.$$

Finally, we extend the proof to the case where P and Q differ on some symbols, i.e., there is at least a symbol in P that is not in Q (and vice-versa). Define $\Sigma' = P([d]) \cap Q([d])$ to be the set of symbols that appear in both P and Q. Let P' be the string obtained by deleting from P the symbols not in Σ', and let Q' be obtained similarly from Q. Clearly, P' and Q' are permutations, they have the same length $d' = |\Sigma'|$, and they use exactly the same symbols. Furthermore, $\mathrm{ed}(P, Q) = \mathrm{ed}(P', Q') + \Theta(d - d')$. Applying Lemma 5.1 to P', Q', we get $m \geq \mathrm{ed}(P', Q')/4$ inversions $\{i'_1, j'_1\}, \ldots, \{i'_m, j'_m\}$ such that $i'_1, j'_1, \ldots, i'_m, j'_m$ are all distinct. Translating these positions to P yields m inversions $\{i_1, j_1\}, \ldots, \{i_m, j_m\}$ between P and Q, such that $i_1, j_1, \ldots, i_m, j_m$ are all distinct. We then let S contain the indices in these inversions and also the $d - d'$ indices in P containing the symbols not in Σ'. It is not difficult to see that we will still get $|S| \geq \Omega(\mathrm{ed}(P, Q))$. Inversions will give rise to requirements of the form $\pi(a) = \pi(b)$ as before, and each index i where $P(i) \notin \Sigma'$ gives rise to a requirement $\pi(P(i)) = \pi(Q(f(i)))$. Altogether, after removing indices i such that $f(i) = \perp$, we still get at least $|S|/8$ requirements whose variables $\pi(a), \pi(b)$ are all distinct. $\square$

6 Concluding Remarks

Poincaré inequality. Our communication complexity lower bounds imply that embedding the edit and Ulam metrics into ℓ_1, and into powers thereof, requires distortion $\Omega(\frac{\log d}{\log \log d})$. But our proof also yields a Poincaré-type inequality, as follows. Indeed, using (i) a variant of Lemma 3.4, where $\Pr[\mathcal{H}^A(x) \neq \mathcal{H}^B(x)]$ is replaced with $\mathbb{E}[\mathcal{H}^A(x) - \mathcal{H}^B(x)]^2$ and $\mathcal{H}^A, \mathcal{H}^B$ are real (rather than boolean) functions with $\mathbb{E}_x\left[\mathcal{H}^A(x)\right]^2 = \mathbb{E}_x\left[\mathcal{H}^B(x)\right]^2 = 1$, together with (ii) Lemma 4.2 for suitable parameters $R = d^{1/4}$ and $\alpha = \Theta(\frac{\log d}{\log \log d})$, we get that for all $f : \mathbb{Z}_p^d \to \mathbb{R}$ (and thus all $f : \mathbb{Z}_p^d \to \ell_2$)

$$\mathbb{E}_{(x,y)\in\mu_0}\left[f(x) - f(y)\right]^2 - \mathbb{E}_{(x,y)\in\mu_1}\left[f(x) - f(y)\right]^2 \leq \tfrac{1}{10}\mathbb{E}_{x,y\in\mathbb{Z}_p^d}\left[f(x) - f(y)\right]^2. \tag{13}$$

In fact, the aforementioned nonembeddability into ℓ_1 (actually into the bigger space squared-ℓ_2) can be proved *directly* from the Poincaré inequality (13), as follows. Consider a D-distortion embedding into squared-ℓ_2, namely, let $\phi : \mathbb{Z}_p^d \to \ell_2$ be such that for all permutations $x, y \in \mathbb{Z}_p^d$,

$$\mathrm{ed}(x, y)/D \leq \|\phi(x) - \phi(y)\|_2^2 \leq \mathrm{ed}(x, y)$$

Schoenberg [Sch38] proved (see e.g., [DL97, Theorem 9.1.1]) that for every $\lambda > 0$, applying the transform $x \mapsto 1 - e^{-\lambda x}$ on the distances of a squared-ℓ_2 metric always results with a squared-ℓ_2 metric. Thus, there exists a mapping $\psi : \mathbb{Z}_p^d \to \ell_2$ satisfying

$$\|\psi(x) - \psi(y)\|_2^2 = 1 - e^{-\|\phi(x)-\phi(y)\|_2^2 \cdot \alpha/R}.$$

We thus get, using Lemma 4.8, that

$$\mathbb{E}_{\mu_0}\|\psi(x) - \psi(y)\|_2^2 - \mathbb{E}_{\mu_1}\|\psi(x) - \psi(y)\|_2^2 - \tfrac{1}{10}\cdot\mathbb{E}_{x,y\in\mathbb{Z}_p^d}\|\psi(x) - \psi(y)\|_2^2 \geq \tfrac{1}{e} - \tfrac{1}{e^{\alpha/D}} - \tfrac{1}{10} - d^{-\Omega(1)}.$$

Combining this inequality with Eqn. (13) implies that $D \geq \Omega(\alpha) = \Omega(\frac{\log d}{\log \log d})$.

Overcoming non-embeddability into ℓ_1. One moral of our non-embeddability lower bound is that some of the usual approaches for designing algorithms for Ulam and edit metrics cannot give approximation better than $\tilde{\Omega}(\log d)$. In particular, to design for these metrics a nearest neighbor algorithm that achieves a sublogarithmic (in d) approximation factor, we must depart from embedding into ℓ_1 or into other spaces that admit $O(1)$-size sketches.

This challenge was accomplished for the Ulam metric in a recent paper [AIK09], which designs a nearest neighbor algorithm achieving $O(\log\log d)$ approximation using polynomial (in $n+d$) space and sublinear (in n) query time. These guarantees bypass the ℓ_1 non-embeddability barrier proved in the current paper, by relying on a constant-distortion embedding of the Ulam metric into an alternative, richer host space (namely, iterated product of simple spaces such as ℓ_1) which, despite the richer structure, turns out to have reasonably good algorithms.

Acknowledgments

We thank Parikshit Gopalan, Piotr Indyk, T.S. Jayram, Ravi Kumar, Ilan Newman, and Yuri Rabinovich for numerous early discussions on non-embeddability of the Ulam and edit metrics, which undoubtedly were a precursor of the current work. We also thank James Lee and Assaf Naor for enlightening discussions about the Poincaré inequality. Finally, we thank the reviewers for careful reading and suggestions that helped improve the clarity and exposition of the article.

References

[ADG+03] Alexandr Andoni, Michel Deza, Anupam Gupta, Piotr Indyk, and Sofya Raskhodnikova. Lower bounds for embedding edit distance into normed spaces. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 523–526, 2003.

[AIK08] Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. Earth mover distance over high-dimensional spaces. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 343–352, 2008.

[AIK09] Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. Overcoming the ℓ_1 non-embeddability barrier: Algorithms for product metrics. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 865–874, 2009.

[AO09] Alexandr Andoni and Krzysztof Onak. Approximating edit distance in near-linear time. In *Proceedings of the Symposium on Theory of Computing (STOC)*, pages 199–204, 2009.

[BEK+03] Tuğkan Batu, Funda Ergün, Joe Kilian, Avner Magen, Sofya Raskhodnikova, Ronitt Rubinfeld, and Rahul Sami. A sublinear algorithm for weakly approximating edit distance. In *Proceedings of the Symposium on Theory of Computing (STOC)*, pages 316–324, 2003.

[BES06] Tuğkan Batu, Funda Ergün, and Cenk Sahinalp. Oblivious string embeddings and edit distance approximations. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 792–801, 2006.

[BFC08] Philip Bille and Martin Farach-Colton. Fast and compact regular expression matching. *Theoretical Computer Science*, 409(28):486–496, 2008.

[BJKK04] Ziv Bar-Yossef, T. S. Jayram, Robert Krauthgamer, and Ravi Kumar. Approximating edit distance efficiently. In *Proceedings of the Symposium on Foundations of Computer Science (FOCS)*, pages 550–559, 2004.

[BJKS04] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.*, 68(4):702–732, 2004.

[Bou02] J. Bourgain. On the distributions of the Fourier spectrum of Boolean functions. *Israel J. Math.*, 131:269–276, 2002.

[CK06] Moses Charikar and Robert Krauthgamer. Embedding the ulam metric into ℓ_1. *Theory of Computing*, 2(11):207–224, 2006.

[CM07] Graham Cormode and S. Muthukrishnan. The string edit distance matching problem with moves. *ACM Trans. Algorithms*, 3(1), 2007. Special issue on SODA'02.

[Cor03] Graham Cormode. *Sequence Distance Embeddings*. Ph.D. Thesis. University of Warwick, 2003.

[CPSV00] Graham Cormode, Mike Paterson, Suleyman Cenk Sahinalp, and Uzi Vishkin. Communication complexity of document exchange. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 197–206, 2000.

[DL97] M. M. Deza and M. Laurent. *Geometry of cuts and metrics*. Springer-Verlag, Berlin, 1997.

[GJKK07] Parikshit Gopalan, T. S. Jayram, Robert Krauthgamer, and Ravi Kumar. Estimating the sortedness of a data stream. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 318–327, 2007.

[IM98] P. Indyk and R. Motwani. Approximate nearest neighbor: towards removing the curse of dimensionality. *Proceedings of the Symposium on Theory of Computing (STOC)*, pages 604–613, 1998.

[Ind01] P. Indyk. Tutorial: Algorithmic applications of low-distortion geometric embeddings. *Proceedings of the Symposium on Foundations of Computer Science (FOCS)*, pages 10–33, 2001.

[Ind04] Piotr Indyk. Approximate nearest neighbor under edit distance via product metrics. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 646–650, 2004.

[KKL88] J. Kahn, G. Kalai, and N. Linial. The influence of variables on boolean functions. In *Proceedings of the Symposium on Foundations of Computer Science (FOCS)*, pages 68–80, 1988.

[KN06] Subhash Khot and Assaf Naor. Nonembeddability theorems via fourier analysis. *Math. Ann.*, 334(4):821–852, 2006. Preliminary version appeared in FOCS'05.

[KOR00] E. Kushilevitz, R. Ostrovsky, and Y. Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM J. Comput.*, 30(2):457–474, 2000. Preliminary version appeared in STOC'98.

[KR06] Robert Krauthgamer and Yuval Rabani. Improved lower bounds for embeddings into l_1. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1010–1017, 2006.

[Mat07] J. Matoušek. Collection of open problems on low-distortion embeddings of finite metric spaces. March 2007. Available online. Last access in August, 2007.

[MP80] William J. Masek and Mike Paterson. A faster algorithm computing string edit distances. *J. Comput. Syst. Sci.*, 20(1):18–31, 1980.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[MS00] S. Muthukrishnan and C. Sahinalp. Approximate nearest neighbors and sequence comparison with block operations. *Proceedings of the Symposium on Theory of Computing (STOC)*, pages 416–424, 2000.

[OR07] Rafail Ostrovsky and Yuval Rabani. Low distortion embedding for edit distance. *J. ACM*, 54(5), 2007. Preliminary version appeared in STOC'05.

[Sah08] Süleyman Cenk Sahinalp. Edit distance under block operations. In Ming-Yang Kao, editor, *Encyclopedia of Algorithms*. Springer, 2008.

[Sch38] I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, November 1938.

[SS02] Michael Saks and Xiaodong Sun. Space lower bounds for distance approximation in the data stream model. In *Proceedings of the Symposium on Theory of Computing (STOC)*, pages 360–369, 2002.

[SU04] Cenk Sahinalp and Andrey Utis. Hardness of string similarity search and other indexing problems. In *Proceedings of International Colloquium on Automata, Languages and Programming (ICALP)*, pages 1080 – 1098, 2004.

[WC76] C. K. Wong and Ashok K. Chandra. Bounds for the string editing problem. *J. ACM*, 23(1):13–16, 1976.

[Woo04] D. Woodruff. Optimal space lower bounds for all frequency moments. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 167–175, 2004.

[Woo07] D. P. Woodruff. *Efficient and Private Distance Approximation in the Communication and Streaming Models*. PhD thesis, MIT, 2007.