Sample-Based Distance-Approximation for Subsequence-Freeness

Omer Cohen Sidon
Tel Aviv University, Israel

Dana Ron
Tel Aviv University, Israel

Abstract

In this work, we study the problem of approximating the distance to subsequence-freeness in the sample-based distribution-free model. For a given subsequence (word) $w = w_1 \dots w_k$, a sequence (text) $T = t_1 \dots t_n$ is said to contain w if there exist indices $1 \le i_1 < \dots < i_k \le n$ such that $t_{i_j} = w_j$ for every $1 \le j \le k$. Otherwise, T is w-free. Ron and Rosin (ACM TOCT 2022) showed that the number of samples both necessary and sufficient for one-sided error testing of subsequence-freeness in the sample-based distribution-free model is $\Theta(k/\epsilon)$.

Denoting by $\Delta(T, w, p)$ the distance of T to w-freeness under a distribution $p : [n] \to [0, 1]$, we are interested in obtaining an estimate $\widehat{\Delta}$, such that $|\widehat{\Delta} - \Delta(T, w, p)| \le \delta$ with probability at least $2/3$, for a given distance parameter δ. Our main result is an algorithm whose sample complexity is $\tilde{O}(k^2/\delta^2)$. We first present an algorithm that works when the underlying distribution p is uniform, and then show how it can be modified to work for any (unknown) distribution p. We also show that a quadratic dependence on $1/\delta$ is necessary.

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

Keywords and phrases Property Testing, Distance Approximation

Digital Object Identifier 10.4230/LIPIcs.ICALP.2023.44

Category Track A: Algorithms, Complexity and Games

Related Version *Full Version*: `https://arxiv.org/abs/2305.01358`

Funding *Dana Ron*: Supported by the Israel Science Foundation (grant number 1146/18) and the Kadar-family award.

1 Introduction

Distance approximation algorithms, as defined in [29], are sublinear algorithms that approximate (with constant success probability) the distance of objects from satisfying a prespecified property $\mathcal{P}$. Distance approximation (and the closely related notion of tolerant testing) is an extension of property testing [31, 20], where the goal is to distinguish between objects that satisfy a property $\mathcal{P}$ and those that are far from satisfying the property.[1] In this work we consider the property of subsequence-freeness. For a given subsequence (word) $w_1 \dots w_k$ over some alphabet Σ, a sequence (text) $T = t_1 \dots t_n$ over Σ is said to be w-free if there do not exist indices $1 \le j_1 < \dots < j_k \le n$ such that $t_{j_i} = w_i$ for every $i \in [k]$.[2]

In most previous works on property testing and distance approximation, the algorithm is allowed query access to the object, and distance to satisfying the property in question, $\mathcal{P}$, is defined as the minimum Hamming distance to an object that satisfies $\mathcal{P}$, normalized by

[1] Tolerant testing algorithms are required to distinguish between objects that are close to satisfying a property and those that are far from satisfying it.

[2] For an integer x, we use $[x]$ to denote the set of integers $\{1, \dots, x\}$

50th International Colloquium on Automata, Languages, and Programming (ICALP 2023).
Editors: Kousha Etessami, Uriel Feige, and Gabriele Puppis; Article No. 44; pp. 44:1–44:19
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

the size of the object. In this work we consider the more challenging, and sometimes more suitable, sample-based model in which the algorithm is only given a random sample from the object. In particular, when the object is a sequence $T = t_1 \dots t_n$, each element in the sample is a pair (j, t_j).

We study both the case in which the underlying distribution according to which each index j is selected (independently) is the uniform distribution over $[n]$, and the more general case in which the underlying distribution is some arbitrary unknown $p : [n] \to [0,1]$. We refer to the former as the *uniform sample-based model*, and to the latter as the *distribution-free sample-based model.* The distance (to satisfying the property) is determined by the underlying distribution. Namely, it is the minimum total weight according to p of indices j such that t_j must be modified so as to make the sequence w-free. Hence, in the uniform sample-based model, the distance measure is simply the Hamming distance normalized by n.

The related problem of testing the property of subsequence-freeness in the distribution-free sample-based model was studied by Ron and Rosin [30]. They showed that the sample-complexity of one-sided error testing of subsequence-freeness in this model is $\Theta(k/\epsilon)$ (where ϵ is the given distance parameter). A natural question is whether we can design a sublinear algorithm, with small sample complexity, that actually approximates the distance of a text T to w-freeness. It is worth noting that, in general, tolerant testing (and hence distance-approximation) for a property may be much harder than testing the property [18, 3].

1.1 Our results

In what follows, when we say that a sample is selected uniformly from T, we mean that for each sample point (j, t_j), j is selected uniformly and independently from $[n]$. This generalizes to the case in which the underlying distribution is an arbitrary distribution p.

We start by designing a distance-approximation algorithm in the uniform sample-based model. Let $\Delta(T, w)$ denote the distance under the uniform distribution of T from being w-free (which equals the fraction of symbols in T that must be modified so as to obtain a w-free text), and let $\delta \in (0,1)$ denote the error parameter given to the algorithm.

▶ **Theorem 1.** *There exists a sample-based distance-approximation algorithm for subsequence-freeness under the uniform distribution, that takes a sample of size $\Theta\left(\frac{k^2}{\delta^2} \cdot \log\left(\frac{k}{\delta}\right)\right)$ and outputs an estimate $\widehat{\Delta}$ such that $|\widehat{\Delta} - \Delta(T, w)| \leq \delta$ with probability at least* $2/3$.[3]

We then turn to extending this result to the distribution-free sample-based model. For a distribution $p : [n] \to [0,1]$, we use $\Delta(T, w, p)$ to denote the distance of T from w-freeness under the distribution p (i.e., the minimum weight, according to p, of the symbols in T that must be modified so as to obtain a w-free text).

▶ **Theorem 2.** *There exists a sample-based distribution-free distance-approximation algorithm for subsequence-freeness, that takes a sample of size $\Theta\left(\frac{k^2}{\delta^2} \cdot \log\left(\frac{k}{\delta}\right)\right)$ from T, distributed according to an unknown distribution p, and outputs an estimate $\widehat{\Delta}$ such that $|\widehat{\Delta} - \Delta(T, w, p)| \leq \delta$ with probability at least $\frac{2}{3}$.*

Finally, we address the question of how tight is our upper bound. We show (using a fairly simple argument) that the quadratic dependence on $1/\delta$ is indeed necessary, even for the uniform distribution. To be precise, denoting by k_d the number of distinct symbols in w, we

[3] As usual, we can increase the success probability to $1 - \eta$, for any $\eta > 0$ at a multiplicative cost of $O(\log(1/\eta))$ in the sample complexity.

give a lower bound of $\Omega(1/(k_d\delta^2))$ under the uniform distribution (that holds for every w with k_d distinct symbols, sufficiently large n and sufficiently small δ – for a precise statement, see Theorem 27).

1.2 A high-level discussion of our algorithms

Our starting point is a structural characterization of the distance to w-freeness under the uniform distribution, which is proved in [30, Sec. 3.1].[4] In order to state their characterization, we introduce the notion of copies of w in T, and more specifically, role-disjoint copies.

A *copy* of $w = w_1 \dots w_k$ in $T = t_1 \dots t_n$ is a sequence of indices $(j_1, \dots, j_k)$ such that $1 \leq j_1 < \dots < j_k \leq n$ and $t_{j_1} \dots t_{j_k} = w$. It will be convenient to represent a copy as an array C of size k where $C[i] = j_i$. A set of copies $\{C_\ell\}$ is said to be *role-disjoint* if for every $i \in [k]$, the indices in $\{C_\ell[i]\}$ are distinct (though it is possible that $C_\ell[i] = C_{\ell'}[i']$ for $i \neq i'$ (and $\ell \neq \ell'$)). In the special case where the symbols of w are all different from each other, a set of copies is role disjoint simply if it consists of disjoint copies. Ron and Rosin prove [30, Theorem 3.4 + Claim 3.1] that $\Delta(T, w)$ equals the maximum number of role-disjoint copies of w in T, divided by n.

Note that the analysis of the sample complexity of one-sided error sample-based testing of subsequence-freeness translates to bounding the size of the sample that is sufficient and necessary for ensuring that the sample contains evidence that T is not w-free when $\Delta(T, w) > \epsilon$. Here evidence is in the form of a copy of w in the sample, so that the testing algorithm simply checks whether such a copy exists. On the other hand, the question of distance-approximation has a more algorithmic flavor, as it is not determined by the problem what must be done by the algorithm given a sample.

Focusing first on the uniform case, Ron and Rosin used their characterization (more precisely, the direction by which if $\Delta(T, w) > \epsilon$, then T contains more than ϵn role-disjoint copies of w), to prove that a sample of size $\Theta(k/\epsilon)$ contains at least one copy of w with probability at least 2/3. In this work we go further by designing an algorithm that actually approximates the number of role-disjoint copies of w in T (and hence approximates $\Delta(T, w)$), given a uniformly selected sample from T. It is worth noting that the probability of obtaining a copy in the sample might be quite different for texts that have *exactly the same* number of role-disjoint copies of w (and hence the same distance to being w-free).[5]

In the next subsection we discuss the aforementioned algorithm (for the uniform case), and in the following one address the distribution-free case.

1.2.1 The uniform case

Let $R(T, w)$ denote the number of role-disjoint copies of w in T. In a nutshell, the algorithm works by computing estimates of the numbers of occurrences of symbols of w in a relatively small number of prefixes of T, and using them to derive an estimate of $R(T, w)$. The more precise description of the algorithm and its analysis are based on several combinatorial claims that we present and which we discuss shortly next.

Let $R_i^j(T, w)$ denote the number of role-disjoint copies of the length-i prefix of w, $w_1 \dots w_i$, in the length-j prefix of T, $t_1 \dots t_j$, and let $N_i^j(T, w)$ denote the number of occurrences of the symbol w_i in $t_1 \dots t_j$. In our first combinatorial claim, we show that for every $i \in [k]$

[4] Indeed, Ron and Rosin note that: "The characterization may be useful for proving further results regarding property testing of subsequence-freeness, as well as (sublinear) distance approximation."

[5] For example, consider $w = 1 \dots k$, $T_1 = (1 \dots k)^{n/k}$ and $T_2 = 1^{n/k} \dots k^{n/k}$.

and $j \in [n]$, the value of $R_i^j(T,w)$ can be expressed in terms of the values of $N_i^{j'}(T,w)$ for $j' \in [j]$ (in particular, $N_i^j(T,w)$) and the values of $R_{i-1}^{j'-1}(T,w)$ for $j' \in [j]$. In other words, we establish a recursive expression which implies that if we know what are $R_{i-1}^{j'-1}(T,w)$ and $N_i^{j'}(T,w)$ for every $j' \in [j]$, then we can compute $R_i^j(T,w)$ (and as an end result, compute $R(T,w) = R_k^n(T,w)$).

In our second combinatorial claim we show that if we only want an approximation of $R(T,w)$, then it suffices to define (also in a recursive manner) a measure that depends on the values of $N_i^j(T,w)$ for every $i \in [k]$ but only for a relatively small number of choices of j, which are evenly spaced. To be precise, each such j belongs to the set $J = \{r \cdot \gamma n\}_{r=1}^{1/\gamma}$ for $\gamma = \Theta(\delta/k)$. We prove that since each interval $[(r-1)\gamma n + 1, r\gamma n]$ is of size γn for this choice of γ, we can ensure that the aforementioned measure (which uses only $j \in J$) approximates $R(T,w)$ to within $O(\delta n)$.

We then prove that if we replace each $N_i^j(T,w)$ for these choices of j (and for every $i \in [k]$) by a sufficiently good estimate, then we incur a bounded error in the approximation of $R(T,w)$. Finally, such estimates are obtained using (uniform) sampling, with a sample of size $\tilde{O}(k^2/\delta^2)$.

1.2.2 The distribution-free case

In [30, Sec. 4] it is shown that, given a word w, a text T and a distribution p, it is possible to define a word $\widetilde{w}$ and a text $\widetilde{T}$ for which the following holds. First, $\Delta(T,w,p)$ is closely related to $\Delta(\widetilde{T},\widetilde{w})$. Second, the probability of observing a copy of w in a sample selected from T according to p is closely related to the probability of observing a copy of $\widetilde{w}$ in a sample selected uniformly from $\widetilde{T}$.

We use the first relation stated above (i.e., between $\Delta(T,w,p)$ and $\Delta(\widetilde{T},\widetilde{w})$). However, since we are interested in distance-approximation rather than one-sided error testing, the second relation stated above (between the probability of observing a copy of w in T and that of observing a copy of $\widetilde{w}$ in $\widetilde{T}$) is not sufficient for our needs, and we need to take a different (once again, more algorithmic) path, as we explain shortly next.

Ideally, we would have liked to sample uniformly from $\widetilde{T}$, and then run the algorithm discussed in the previous subsection using this sample (and $\widetilde{w}$). However, we only have sampling access to T according to the underlying distribution p, and we do not have direct sampling access to uniform samples from $\widetilde{T}$. Furthermore, since $\widetilde{T}$ is defined based on (the unknown) p, it is not clear how to determine the aforementioned subset of (evenly spaced) indices J.

For the sake of clarity, we continue the current exposition while making two assumptions. The first is that the distribution p is such that there exists a value β, such that p_j/β is an integer for every $j \in [n]$ (the value of β need not be known). The second is that in w there are no two consecutive symbols that are the same. Under these assumptions, $\widetilde{T} = t_1^{p_1/\beta} \dots t_n^{p_n/\beta}$, $\widetilde{w} = w$, and $\Delta(\widetilde{T},\widetilde{w}) = \Delta(T,w,p)$ (where t_j^x for an integer x is the subsequence that consists of x repetitions of t_j).

Our algorithm for the distribution-free case (working under the aforementioned assumptions), starts by taking a sample distributed according to p and using it to select a (relatively small) subset of indices in $[n]$. Denoting these indices by $b_0, b_1, \dots, b_\ell$, where $b_0 = 0 < b_1 < \dots < b_{\ell-1} < b_\ell = n$, we would have liked to ensure that the weight according to p of each interval $[b_{u-1}+1, b_u]$ is approximately the same (as is the case when considering the intervals defined by the subset J in the uniform case). To be precise, we would have liked each interval to have relatively small weight, while the total number of intervals is not too large. However, since it is possible that for some single indices $j \in [n]$, the probability p_j is large, we also allow intervals with large weight, where these intervals consist of a single index (and there are few of them).

The algorithm next takes an additional sample, to approximate, for each $i \in [k]$ and $u \in [\ell]$, the weight, according to p, of the occurrences of the symbol w_i in the length-b_u prefix of T. Observe that prefixes of T correspond to prefixes of $\widetilde{T}$. Furthermore, the weight according to p of occurrences of symbols in such prefixes, translates to numbers of occurrences of symbols in the corresponding prefixes in $\widetilde{T}$, normalized by the length of $\widetilde{T}$. The algorithm then uses these approximations to obtain an estimate of $\Delta(\widetilde{T}, \widetilde{w})$.

We note that some pairs of consecutive prefixes in $\widetilde{T}$ might be far apart, as opposed to what we had in the algorithm for the uniform case described in Section 1.2.1. However, this is always due to single-index intervals in T (for j such that p_j is large). Each such interval corresponds to a consecutive subsequence in $\widetilde{T}$ with repetitions of the same symbol, and we show that no additional error is incurred because of such intervals.

1.3 Related results

As we have previously mentioned, the work most closely related to ours is that of Ron and Rosin on distribution-free sample-based testing of subsequence-freeness [30]. For other related results on property testing (e.g., testing other properties of sequences, sample-based testing of other types of properties and distribution-free testing (possibly with queries)), see the introduction of [30], and in particular Section 1.4. For another line of work, on sublinear approximation of the longest increasing subsequence, see [27] and references within. Here we shortly discuss related results on distance approximation / tolerant testing.

As already noted, distance approximation and tolerant testing were first formally defined in [29], and were shown to be significantly harder for some properties in [18, 3]. Almost all previous results are query-based, and where the distance measure is with respect to the uniform distribution. These include [21, 19, 1, 26, 16, 11, 23, 7, 25, 17, 28]. Kopparty and Saraf [24] present results for query-based tolerant testing of linearity under several families of distributions. Berman, Raskhodnikova and Yaroslavtsev [5] give tolerant (query based) L_p-testing algorithms for monotonicity. Berman, Murzbulatov and Raskhodnikova [4] give a sample-based distance-approximation algorithms for image properties that works under the uniform distribution.

Canonne et al. [12] study the property of k-monotonicity of Boolean functions over various posets. A Boolean function over a finite poset domain D is k-monotone if it alternates between the values 0 and 1 at most k times on any ascending chain in D. For the special case of $D = [n]$, the property of k-monotonicity is equivalent to being free of w of length $k + 2$ where $w_1 \in \{0, 1\}$ and $w_i = 1 - w_{i-1}$ for every $i \in [2, k + 2]$. One of their results implies an upper bound of $\widetilde{O}\left(\frac{k}{\delta^3}\right)$ on the sample complexity of distance-approximation for k-monotonicity of functions $f : [n] \to \{0, 1\}$ under the uniform distribution (and hence for w-freeness when w is a binary subsequence of a specific form). This result generalizes to k-monotonicity in higher dimensions (at an exponential cost in the dimension d).

Blum and Hu [9] study distance-approximation for k-interval (Boolean) functions over the line in the distribution-free active setting. In this setting, an algorithm gets an unlabeled sample and asks queries on a subset of sample points. Focusing on the sample complexity, they show that for any underlying distribution p on the line, a sample of size $\widetilde{O}\left(\frac{k}{\delta^2}\right)$ is sufficient for approximating the distance to being a k-interval function up to an additive error of δ. This implies a sample-based distribution-free distance-approximation algorithm with the same sample complexity for the special case of being free of the same pair of w's described in the previous paragraph, replacing $k + 2$ by $k + 1$.

Blais, Ferreira Pinto Jr. and Harms [8] introduce a variant of the VC-dimension and use it to prove lower and upper bounds on the sample complexity of distribution-free testing for a variety of properties. In particular, one of their results implies that the linear dependence on k in the result of [9] is essentially optimal.

Finally we mention that our procedure in the distribution-free case for constructing "almost-equal-weight" intervals by sampling is somewhat reminiscent of techniques used in other contexts of testing when dealing with non-uniform distributions [6, 22, 10].

1.4 Further research

The main open problem left by this work is closing the gap between the upper and lower bounds that we give, and in particular understanding the precise dependence on k, or possibly other parameters determined by w (such as k_d). One step in this direction can be found in the Master Thesis of the first author [13].

1.5 Organization

In Section 2 we present our algorithm for distance-approximation under the uniform distribution. Some of the main details of the distribution-free case appears in Section 3, and in Section 4 we prove our lower bound. All missing details and proofs can be found in the full version of this paper [14].

2 Distance approximation under the uniform distribution

In this section, we address the problem of distance approximation when the underlying distribution is the uniform distribution. As mentioned in the introduction, Ron and Rosin showed [30, Thm. 3.4] that $\Delta(T,w)$ (the distance of T from w-freeness under the uniform distribution), equals the number of role-disjoint copies of w in T, divided by $n = |T|$ (where role-disjoint copies are as defined in the introduction – see Section 1.2). We may use $T[j]$ to denote the j^{th} symbol of T (so that $T[j] = t_j$).

We start by introducing the following notations.

▶ **Definition 3.** *For every $i \in [k]$ and $j \in [n]$, let $N_i^j(T,w)$ denote the number of occurrences of the symbol w_i in the length j prefix of T, $T[1,j] = T[1]\dots T[j]$.*[6] *Let $R_i^j(T,w)$ denote the number of role-disjoint copies of the subsequence $w_1 \dots w_i$ in $T[1,j]$. When $i = k$ and $j = n$, we use the shorthand $R(T,w)$ for $R_k^n(T,w)$ (the total number of role-disjoint copies of w in T).*

Observe that $R_1^j(T,w)$ equals $N_1^j(T,w)$ for every $j \in [n]$.

Since, as noted above, $\Delta(T,w) = R(T,w)/n$, we would like to estimate $R(T,w)$. More precisely, given $\delta > 0$ we would like to obtain an estimate $\widehat{R}$, such that: $\left|\widehat{R} - R(T,w)\right| \le \delta n$. To this end, we first establish two combinatorial claims. The first claim shows that the value of each $R_i^j(T,w)$ can be expressed in terms of the values of $N_i^{j'}(T,w)$ for $j' \in [j]$ (in particular, $N_i^j(T,w)$) and the values of $R_{i-1}^{j'-1}(T,w)$ for $j' \in [j]$. In other words, if we know what are $R_{i-1}^{j'-1}(T,w)$ and $N_i^{j'}(T,w)$ for every $j' \in [j]$, then we can compute $R_i^j(T,w)$.

▷ Claim 4. For every $i \in \{2,\dots,k\}$ and $j \in [n]$,

$$R_i^j(T,w) = N_i^j(T,w) - \max_{j' \in [j]} \left\{ N_i^{j'}(T,w) - R_{i-1}^{j'-1}(T,w) \right\} .$$

[6] Indeed, if $w_i = w_{i'}$ for $i \neq i'$, then $N_i^j(T,w) = N_{i'}^j(T,w)$ for every j.

Clearly, $R_i^j(T,w) \le N_i^j(T,w)$ (for every $i \in \{2,\dots,k\}$ and $j \in [n]$), since each role-disjoint copy of $w_1 \dots w_i$ in $T[1,j]$ must end with a distinct occurrence of w_i in $T[1,j]$. Claim 4 states by exactly how much is $R_i^j(T,w)$ smaller than $N_i^j(T,w)$. Roughly speaking, the expression $\max_{j' \in [j]} \left\{ N_i^{j'}(T,w) - R_{i-1}^{j'-1}(T,w) \right\}$ accounts for the number of occurrences of w_i in $T[1,j]$ that cannot be used in role-disjoint copies of $w_1 \dots w_i$ in $T[1,j]$.

Proof. For simplicity (in terms of notation), we prove the claim for the case that $i = k$ and $j = n$. The proof for general $i \in \{2,\dots,k\}$ and $j \in [n]$ is essentially the same up to renaming of indices. Since T and w are fixed throughout the proof, we shall use the shorthand N_i^j for $N_i^j(T,w)$ and R_i^j for $R_i^j(T,w)$.

For the sake of the analysis, we start by describing a simple greedy procedure, that constructs $R = R_k^n$ role-disjoint copies of w in T. The correctness of this procedure follows from [30, Claim 3.5] and a simple inductive argument (details are provided in the full version of the paper [14]). Every copy C_m, for $m \in [R]$ is an array of size k whose values are monotonically increasing, where for every $i \in [k]$ we have that $C_m[i] \in [n]$, and $T[C_m[i]] = w_i$. Furthermore, for every $i \in [k]$ the indices $C_1[i], \dots, C_R[i]$ are distinct. For every $m = 1, \dots, R$ and $i = 1, \dots, k$, the procedure scans T, starting from $T[C_m[i-1]+1]$ (where we define $C_m[0]$ to be 0) and ending at $T[n]$ until it finds the first index j such that $T[j] = w_i$ and $j \notin \{C_1[i], \dots, C_{m-1}[i]\}$. It then sets $C_m[i] = j$. For $i > 1$ we say in such a case that the procedure *matches* j to the partial copy $C_m[1], \dots, C_m[i-1]$.

For $i \in [k]$, define: $G_i = \{j \in [n] : T[j] = w_i\}$. Also define: $G_i^+ = \{j \in G_i : \exists m, C_m[i] = j\}$ and $G_i^- = \{j \in G_i : \nexists m, C_m[i] = j\}$ (recall that $C_m[i]$ is the i-th index in the m-th greedy copy).

It is easy to verify that $|G_i| = N_i^n$, $\left|G_i^+\right| = R_i^n$ and $|G_i| = \left|G_i^+\right| + \left|G_i^-\right|$. To complete the proof, we will show that $\left|G_i^-\right| = \max_{j \in [n]} \left\{ N_i^j - R_{i-1}^{j-1} \right\}$.

Let j^* be an index j that maximizes $\left\{ N_i^j - R_{i-1}^{j-1} \right\}$. In the interval $[j^*]$ we have $N_i^{j^*}$ occurrences of w_i, and in the interval $[j^*-1]$ we only have $R_{i-1}^{j^*-1}$ role-disjoint copies of $w_1 \dots w_{i-1}$. This implies that in the interval $[j^*]$ there are at least $N_i^{j^*} - R_{i-1}^{j^*-1}$ occurrences of w_i that cannot be the i-th index of any greedy copy, and so we have

$$\left|G_i^-\right| \ge N_i^{j^*} - R_{i-1}^{j^*-1} = \max_{j \in [n]} \left\{ N_i^j - R_{i-1}^{j-1} \right\} . \tag{1}$$

On the other hand, denote by j^{**} the largest index in G_i^-. Since each index $j \in [j^{**}]$ such that $T[j] = w_i$ is either the i-th element of some copy or is not the i-th element of any copy, $N_i^{j^{**}} = R_i^{j^{**}-1} + \left|G_i^-\right|$. We claim that $R_i^{j^{**}-1} = R_{i-1}^{j^{**}-1}$. Otherwise, $R_i^{j^{**}-1} < R_{i-1}^{j^{**}-1}$, in which case the index j^{**} would have to be the the i-th element of a greedy copy. Hence,

$$\left|G_i^-\right| = N_i^{j^{**}} - R_{i-1}^{j^{**}-1} \le \max_{j \in [n]} \left\{ N_i^j - R_{i-1}^{j-1} \right\} . \tag{2}$$

In conclusion,

$$\left|G_i^-\right| = \max_{j \in [n]} \left\{ N_i^j - R_{i-1}^{j-1} \right\} , \tag{3}$$

and the claim follows. ◀

In order to state our next combinatorial claim, we first introduce one more definition, which will play a central role in obtaining an estimate for $R(T,w)$.

▶ **Definition 5.** *For $\ell \leq n$, let $\mathcal{N}$ be a $k \times \ell$ matrix of non-negative numbers, where we shall use $\mathcal{N}_i^r$ to denote $\mathcal{N}[i][r]$. For every $r \in [\ell]$ let $M_1^r(\mathcal{N}) = \mathcal{N}_1^r$, and for every $i \in \{2, \dots, k\}$, let*

$$M_i^r(\mathcal{N}) \stackrel{\text{def}}{=} \mathcal{N}_i^r - \max_{r' \leq r} \left\{ \mathcal{N}_i^{r'} - M_{i-1}^{r'}(\mathcal{N}) \right\} \ .$$

When $i = k$ and $r = \ell$ we use the shorthand $M(\mathcal{N})$ for $M_k^\ell(\mathcal{N})$.

In our second combinatorial claim we show that for an appropriate choice of a matrix $\mathcal{N}$, whose entries are a subset of all values in $\left\{ N_i^j(T, w) \right\}_{i \in [k]}^{j \in [n]}$, we can bound the difference between $M(\mathcal{N})$ and $R(T, w)$. We later use sampling to obtain an estimated version of $\mathcal{N}$.

▷ Claim 6. Let $J = \{j_0, j_1, \dots, j_\ell\}$ be a set of indices satisfying $j_0 = 0 < j_1 < j_2 < \dots < j_\ell = n$. Let $\mathcal{N} = \mathcal{N}(J, T, w)$ be the matrix whose entries are $\mathcal{N}_i^r = N_i^{j_r}(T, w)$, for every $i \in [k]$ and $r \in [\ell]$. Then we have

$$|M(\mathcal{N}) - R(T, w)| \leq (k - 1) \cdot \max_{\tau \in [\ell]} \{j_\tau - j_{\tau-1}\} \ .$$

Proof. Recall that $M(\mathcal{N}) = M_k^\ell(\mathcal{N})$ and $R(T, w) = R_k^{j_\ell}(T, w)$. We shall prove that for every $i \in [k]$ and for every $r \in [\ell]$, $\left| M_i^r(\mathcal{N}) - R_i^{j_r}(T, w) \right| \leq (i - 1) \cdot \max_{\tau \in [r]} \{j_\tau - j_{\tau-1}\}$. We prove this by induction on i.
For $i = 1$ and every $r \in [\ell]$,

$$\left| M_1^r(\mathcal{N}) - R_1^{j_r}(T, w) \right| = \left| N_1^{j_r}(T, w) - N_1^{j_r}(T, w) \right| = 0 \leq (1 - 1) \cdot \max_{\tau \in [1]} \{j_\tau - j_{\tau-1}\} \ , \qquad (4)$$

where the first equality follows from the setting of $\mathcal{N}$ and the definitions of $M_1^r(\mathcal{N})$ and $R_1^{j_r}(T, w)$.
For the induction step, we assume the claim holds for $i - 1 \geq 1$ (and every $r \in [\ell]$) and prove it for i. We have,

$$\begin{aligned} & M_i^r(\mathcal{N}) - R_i^{j_r}(T, w) \\ &= N_i^{j_r}(T, w) - \max_{b \in [r]} \left\{ N_i^{j_b}(T, w) - M_{i-1}^b(\mathcal{N}) \right\} - R_i^{j_r}(T, w) && (5) \\ &= \max_{j \in [j_r]} \left\{ N_i^j(T, w) - R_{i-1}^{j-1}(T, w) \right\} - \max_{b \in [r]} \left\{ N_i^{j_b}(T, w) - M_{i-1}^b(\mathcal{N}) \right\} \ , && (6) \end{aligned}$$

where Equation (5) follows from the setting of $\mathcal{N}$ and the definition of $M_i^r(\mathcal{N})$, and Equation (6) is implied by Claim 4. Denote by j^* an index $j \in [j_r]$ that maximizes the first max term and let b^* be the largest index such that $j_{b^*} \leq j^*$. We have:

$$\begin{aligned} & \max_{j \in [j_r]} \left\{ N_i^j(T, w) - R_{i-1}^{j-1}(T, w) \right\} - \max_{b \in [r]} \left\{ N_i^{j_b}(T, w) - M_{i-1}^b(\mathcal{N}) \right\} \\ &\leq N_i^{j^*}(T, w) - R_{i-1}^{j^*-1}(T, w) - N_i^{j_{b^*}}(T, w) + M_{i-1}^{b^*}(\mathcal{N}) \\ &= N_i^{j^*}(T, w) + R_{i-1}^{j_{b^*}}(T, w) - R_{i-1}^{j_{b^*}}(T, w) - R_{i-1}^{j^*-1}(T, w) - N_i^{j_{b^*}}(T.w) + M_{i-1}^{b^*}(\mathcal{N}) \\ &\leq \left(M_{i-1}^{b^*}(\mathcal{N}) - R_{i-1}^{j_{b^*}}(T, w) \right) + \left(N_i^{j^*}(T, w) - N_i^{j_{b^*}}(T, w) \right) + \left(R_{i-1}^{j_{b^*}}(T, w) - R_{i-1}^{j^*-1}(T, w) \right) \\ &\leq (i - 2) \max_{\tau \in [r]} \{j_\tau - j_{\tau-1}\} + \left(j^* - j^{b^*} \right) + \left(j^{b^*} - (j^* - 1) \right) && (7) \\ &= (i - 2) \max_{\tau \in [r]} \{j_\tau - j_{\tau-1}\} + 1 \\ &\leq (i - 2) \max_{\tau \in [r]} \{j_\tau - j_{\tau-1}\} + \max_{\tau \in [r]} \{j_\tau - j_{\tau-1}\} \\ &= (i - 1) \max_{\tau \in [r]} \{j_\tau - j_{\tau-1}\} \ , && (8) \end{aligned}$$

where in Equation (7) we used the induction hypothesis. By combining Equations (6) and (8), we get that

$$M_i^r(\mathcal{N}) - R_i^{j_r}(T,w) \le (i-1)\max_{\tau\in[r]}\{j_\tau - j_{\tau-1}\} \ . \tag{9}$$

Similarly to Equation (6),

$$R_i^{j_r}(T,w) - M_i^r(\mathcal{N}) = \max_{b\in[r]}\left\{N_i^{j_b}(T,w) - M_{i-1}^b(\mathcal{N})\right\} - \max_{j\in[j_r]}\left\{N_i^j(T,w) - R_{i-1}^{j-1}(T,w)\right\} \ . \tag{10}$$

Let b^{**} be the index $b \in [r]$ that maximizes the first max term. We have

$$\begin{aligned}
&\max_{b\in[r]}\left\{N_i^{j_b}(T,w) - M_{i-1}^b(\mathcal{N})\right\} - \max_{j\in[j_r]}\left\{N_i^j(T,w) - R_{i-1}^{j-1}(T,w)\right\}\\
&\quad\le\ N_i^{j_{b^{**}}}(T,w) - M_{i-1}^{b^{**}}(\mathcal{N}) - N_i^{j_{b^{**}}}(T,w) + R_{i-1}^{j_{b^{**}}-1}(T,w)\\
&\quad\le\ R_{i-1}^{j_{b^{**}}}(T,w) - M_{i-1}^{b^{**}}(\mathcal{N}) \ \le\ \left|R_{i-1}^{j_{b^{**}}}(T,w) - M_{i-1}^{b^{**}}(\mathcal{N})\right|\\
&\quad\le\ (i-2)\max_{\tau\in[r]}\{j_\tau - j_{\tau-1}\} \ \le\ (i-1)\max_{\tau\in[r]}\{j_\tau - j_{\tau-1}\} \ .
\end{aligned} \tag{11}$$

Hence (combining Equations (10) and (11)),[7]

$$R_i^{j_r}(T,w) - M_i^r(\mathcal{N}) \le (i-1)\max_{\tau\in[r]}\{j_\tau - j_{\tau-1}\} \ . \tag{12}$$

Together, Equations (9) and (12) give us that

$$\left|M_i^r(\mathcal{N}) - R_i^{j_r}(T,w)\right| \le (i-1)\max_{\tau\in[r]}\{j_\tau - j_{\tau-1}\} \ , \tag{13}$$

and the proof is completed. ◁

In our next claim we bound the difference between $M(\widehat{\mathcal{N}}) - M(\widetilde{\mathcal{N}})$ for any two matrices (with dimensions $k \times \ell$), given a bound on the L_∞ distance between them. We later apply this claim with $\widetilde{\mathcal{N}} = \mathcal{N}$ for $\mathcal{N}$ as defined in Claim 6, and $\widehat{\mathcal{N}}$ being a matrix that contains estimates $\widehat{N}_i^r$ of $N_i^{j_r}(T,w)$ (respectively). We discuss how to obtain $\widehat{\mathcal{N}}$ in Claim 8.

▷ Claim 7. Let $\gamma \in (0,1)$, and let $\widehat{\mathcal{N}}$ and $\widetilde{\mathcal{N}}$ be two $k \times \ell$ matrices. If for every $i \in [t]$ and $r \in [\ell]$, $\left|\widehat{\mathcal{N}}_i^r - \widetilde{\mathcal{N}}_i^r\right| \le \gamma n$, then $\left|M(\widehat{\mathcal{N}}) - M(\widetilde{\mathcal{N}})\right| \le (2k-1)\gamma n$.

Proof. We shall prove that for every $t \in [k]$ and for every $r \in [\ell]$, $\left|M_t^r(\widehat{\mathcal{N}}) - M_t^r(\widetilde{\mathcal{N}})\right| \le (2t-1)\gamma n$. We prove this by induction on t.
For $t = 1$ and every $r \in [\ell]$, we have

$$\left|M_1^r(\widehat{\mathcal{N}}) - M_1^r(\widetilde{\mathcal{N}})\right| \ = \ \left|\widehat{\mathcal{N}}_1^r - \widetilde{\mathcal{N}}_1^r\right| \ \le \ \gamma n \, . \tag{14}$$

Now assume the claim is true for $t-1 \ge 1$ and for every $r \in [\ell]$, and we prove it for t. For any $r \in [\ell]$, by the definition of $M_t^r(\cdot)$,

$$\begin{aligned}
&\left|M_t^r(\widehat{\mathcal{N}}) - M_t^r(\widetilde{\mathcal{N}})\right|\\
&\quad=\ \left|\widehat{\mathcal{N}}_t^r - \max_{r''\in[r]}\left\{\widehat{\mathcal{N}}_t^{r''} - M_{t-1}^{r''}(\widehat{\mathcal{N}})\right\} - \widetilde{\mathcal{N}}_t^r + \max_{r'\in[r]}\left\{\widetilde{\mathcal{N}}_t^{r'} - M_{t-1}^{r'}(\widetilde{\mathcal{N}})\right\}\right|\\
&\quad\le\ \gamma n + \left|\max_{r'\in[r]}\left\{\widetilde{\mathcal{N}}_t^{r'} - M_{t-1}^{r'}(\widetilde{\mathcal{N}})\right\} - \max_{r''\in[r]}\left\{\widehat{\mathcal{N}}_t^{r''} - M_{t-1}^{r''}(\widehat{\mathcal{N}})\right\}\right| \ ,
\end{aligned} \tag{15}$$

[7] It actually holds that $M_i^r(\mathcal{N}) \ge R_i^{j_r}(T,w)$, so that $R_i^{j_r}(T,w) - M_i^r(\mathcal{N}) \le 0$, but for the sake of simplicity of the inductive argument, we prove the same upper bound on $R_i^{j_r}(T,w) - M_i^r(\mathcal{N})$ as on $M_i^r(\mathcal{N}) - R_i^{j_r}(T,w)$.

where in the last inequality we used the premise of the claim. Assume that the first max term in Equation (15) is at least as large as the second (the case that the second term is larger than the first is dealt with analogously), and let r^* be the index that maximizes the first max term. Then,

$$\begin{aligned}
&\left|\max_{r'\in[r]}\left\{\widetilde{\mathcal{N}}_t^{r'} - M_{t-1}^{r'}(\widetilde{\mathcal{N}})\right\} - \max_{r''\in[r]}\left\{\widehat{\mathcal{N}}_t^{r''} - M_{t-1}^{r''}(\widehat{\mathcal{N}})\right\}\right| \\
&\quad\le \left|\left(\widetilde{\mathcal{N}}_t^{r^*} - \widehat{\mathcal{N}}_t^{r^*}\right) + \left(M_{t-1}^{r^*}(\widehat{\mathcal{N}}) - M_{t-1}^{r^*}(\widetilde{\mathcal{N}})\right)\right| \\
&\quad\le \left|\widetilde{\mathcal{N}}_t^{r^*} - \widehat{\mathcal{N}}_t^{r^*}\right| + \left|M_{t-1}^{r^*}(\widehat{\mathcal{N}}) - M_{t-1}^{r^*}(\widetilde{\mathcal{N}})\right| \\
&\quad\le \gamma n + (2t-3)\gamma n = (2t-2)\gamma n \ , \qquad (16)
\end{aligned}$$

where we used the premise of the claim once again, and the induction hypothesis. The claim follows by combining Equation (15) with Equation (16). ◁

The next claim states that we can obtain good estimates for all values in $\left\{N_i^{j_r}(T,w)\right\}_{i\in[k]}^{r\in[\ell]}$ (with a sufficiently large sample). Its (standard) proof is deferred to the full version of this paper [14].

▷ Claim 8. For any $\gamma \in (0,1)$ and $J = \{j_1,\dots,j_\ell\}$ (such that $1 \le j_1 < \dots < j_\ell = n$), by taking a sample of size $\Theta\left(\frac{\log(k\cdot\ell)}{\gamma^2}\cdot\right)$ from T, we can obtain with probability at least $2/3$ estimates $\left\{\widehat{\mathcal{N}}_i^r\right\}_{i\in[k]}^{r\in[\ell]}$, such that

$$\left|\widehat{\mathcal{N}}_i^r - N_i^{j_r}(T,w)\right| \le \gamma n \ , \qquad (17)$$

for every $i \in [k]$ and $r \in [\ell]$.

We can now restate and prove our main theorem for distance approximation under the uniform distribution.

▶ **Theorem 1.** *There exists a sample-based distance-approximation algorithm for subsequence-freeness under the uniform distribution, that takes a sample of size $\Theta\left(\frac{k^2}{\delta^2}\cdot\log\left(\frac{k}{\delta}\right)\right)$ and outputs an estimate $\widehat{\Delta}$ such that $|\widehat{\Delta} - \Delta(T,w)| \le \delta$ with probability at least $2/3$.*[8]

While our focus is on the sample complexity of the algorithm, we note that its running time is linear in the size of the sample.

Proof. The algorithm sets $\gamma = \delta/(3k)$ and $J = \{\gamma n, 2\gamma n, \dots, n\}$. It first applies Claim 8 with the above setting of γ to obtain the estimates $\left\{\widehat{\mathcal{N}}_i^r\right\}$ for every $i \in [k]$ and $r \in [\ell]$, which with probability at least $2/3$ are as stated in Equation (17). If we take $\widetilde{\mathcal{N}} = \mathcal{N}$ for $\mathcal{N}$ as defined in Claim 6, then the premise of Claim 7 holds. We can hence apply Claim 7, and combining with Claim 6 and the definition of J, we get that with probability at least $2/3$, for the matrix $\widehat{\mathcal{N}}$,

$$\left|M(\widehat{\mathcal{N}}) - R(T,w)\right| \le (2k-1)\gamma n + (k-1)\gamma n = (3k-2)\gamma n \le \delta n \ . \qquad (18)$$

The algorithm hence computes $M(\widehat{\mathcal{N}}) = M_k^\ell(\widehat{\mathcal{N}})$ in an iterative manner, based on Definition 5, and outputs $\widehat{\Delta} = M(\widehat{\mathcal{N}})/n$. Since $R(T,w)/n = \Delta(T,w)$, the theorem follows. ◀

[8] As usual, we can increase the success probability to $1-\eta$, for any $\eta > 0$ at a multiplicative cost of $O(\log(1/\eta))$ in the sample complexity.

3 Distribution-free distance approximation

As noted in the introduction, our algorithm for approximating the distance from subsequence-freeness under a general distribution p works by reducing the problem to approximating the distance from subsequence-freeness under the uniform distribution. However, we won't be able to use the algorithm presented in Section 2 as is. There are two main obstacles, explained shortly next. In the reduction, given a word w and access to samples from a text T, distributed according to p, we define a word $\widetilde{w}$ and a text $\widetilde{T}$ such that if we can obtain a good approximation of $\Delta(\widetilde{T}, \widetilde{w})$ then we get a good approximation of $\Delta(T, w, p)$. (Recall that $\Delta(T, w, p)$ denotes the distance of T from being w-free under the distribution p.) However, first, we don't actually have direct access to uniformly distributed samples from $\widetilde{T}$, and second, we cannot work with a set J of indices that induce equally sized intervals (of a bounded size), as we did in Section 2.

We address these challenges (as well as precisely define $\widetilde{T}$ and $\widetilde{w}$) in several stages. We start, in Sections 3.1 and 3.2, by using sampling according to p, in order to construct intervals in T that have certain properties (with sufficiently high probability). The role of these intervals will become clear as we proceed. Due to space constraints, several proofs are deferred to the full version of this paper [14].

3.1 Interval construction and classification

We begin this subsection by defining intervals in $[n]$ that are determined by p (which is unknown to the algorithm). We then construct intervals by sampling from p, where the latter intervals are in a sense approximations of the former (this will be formalized subsequently). Each constructed interval will be classified as either "heavy" or "light", depending on its (approximated) weight according to p. Ideally, we would have liked all intervals to be light, but not too light, so that their number won't be too large (as was the case when we worked under the uniform distribution and simply defined intervals of equal size). However, for a general distribution p we might have single indices $j \in [n]$ for which p_j is large, and hence we also need to allow heavy intervals (each consisting of a single index). We shall make use of the following two definitions.

▶ **Definition 9.** *For any two integers $j_1 \leq j_2$, let $[j_1, j_2]$ denote the interval $\{j_1, \ldots, j_2\}$. For every $j_1, j_2 \in [n]$, define $\mathrm{wt}_p([j_1, j_2]) \stackrel{\text{def}}{=} \sum_{j=j_1}^{j_2} p_j$ to be the weight of the interval $[j_1, j_2]$ according to p. We shall use the shorthand $\mathrm{wt}_p(j)$ for $\mathrm{wt}_p([j, j])$.*

▶ **Definition 10.** *Let S be a multiset of size s, with elements from $[n]$. For every $j \in [n]$, let $N_S(j)$ be the number of elements in S that equal j. For every $j_1, j_2 \in [n]$, define $\mathrm{wt}_S([j_1, j_2]) \stackrel{\text{def}}{=} \frac{1}{s} \sum_{j=j_1}^{j_2} N_S(j)$ to be the estimated weight of the interval $[j_1, j_2]$ according to S. We shall use the shorthand $\mathrm{wt}_S(j)$ for $\mathrm{wt}_S([j, j])$.*

In the next definition, and the remainder of this section, we shall use

$$z = c_z \frac{k}{\delta} , \tag{19}$$

where let $c_z = 100$.

We next define the aforementioned set of intervals, based on p. Roughly speaking, we try to make the intervals as equally weighted as possible, keeping in mind that some indices might have a large weight, so we assign each to an interval of its own.

▶ **Definition 11.** *Define a sequence of indices in the following iterative manner. Let $h_0 = 0$ and for $\ell = 1, 2, \ldots,$ as long as $h_{\ell-1} < n$, let h_ℓ be defined as follows. If $\mathrm{wt}_p(h_{\ell-1}+1) > \frac{1}{8z}$, then $h_\ell = h_{\ell-1}+1$. Otherwise, let h_ℓ be the maximum index $h'_\ell \in [h_{\ell-1}+1, n]$ such that $\mathrm{wt}_p([h_{\ell-1}+1, h'_\ell]) \le \frac{1}{4z}$ and for every $h''_\ell \in [h_{\ell-1}+1, h'_\ell]$, $\mathrm{wt}_p(h''_\ell) \le \frac{1}{8z}$. Let L be such that $h_L = n$.*

Based on the indices $\{h_\ell\}_{\ell=0}^{L}$ defined above, for every $\ell \in [L]$, let $H_\ell = [h_{\ell-1}+1, h_\ell]$ and let $\mathcal{H} = \{H_\ell\}_{\ell=1}^{L}$. We partition $\mathcal{H}$ into three subsets as follows. Let $\mathcal{H}_{sin}$ be the subset of all $H \in \mathcal{H}$ such that $|H| = 1$ and $\mathrm{wt}_p(H) > \frac{1}{8z}$. Let $\mathcal{H}_{med}$ be the set of all $H \in \mathcal{H}$ such that $|H| \ne 1$ and $\frac{1}{8z} \le \mathrm{wt}_p(H) \le \frac{1}{4z}$. Let $\mathcal{H}_{sml}$ be the set of all $H \in \mathcal{H}$ such that $\mathrm{wt}_p(H) < \frac{1}{8z}$.

Observe that since $\mathrm{wt}_p(T) = 1$, then $|\mathcal{H}_{sin} \cup \mathcal{H}_{med}| \le 8z$. In addition, since between each $H', H'' \in \mathcal{H}_{sml}$ there has to be at least one $H \in \mathcal{H}_{sin}$, then we also have $|\mathcal{H}_{sml}| \le 8z + 1$.

By its definition, $\mathcal{H}$ is determined by p. We next construct a set of intervals $\mathcal{B}$ based on sampling according to p (in a similar, but not identical, fashion to Definition 11). Consider a sample S_1 of size s_1 selected according to p (with repetitions), where s_1 will be set subsequently.

▶ **Definition 12.** *Given a sample S_1 (multiset of elements in $[n]$) of size s_1, determine a sequence of indices in the following iterative manner. Let $b_0 = 0$ and for $u = 1, 2, \ldots,$ as long as $b_{u-1} < n$, let b_u be defined as follows. If $\mathrm{wt}_{S_1}(b_{u-1}+1) > 1/z$, then $b_u = b_{u-1}+1$. Otherwise, let b_u be the maximum index $b'_u \in [b_{u-1}+1, n]$ such that $\mathrm{wt}_{S_1}([b_{u-1}+1, b'_u]) \le \frac{1}{z}$. Let U be such that $b_U = n$.*

Based on the indices $\{b_u\}_{u=0}^{U}$ defined above, for every $u \in [U]$, let $B_u = [b_{u-1}+1, b_u]$, and let $\mathcal{B} = \{B_u\}_{u=1}^{U}$. For every $u \in [U]$, if $\mathrm{wt}_{S_1}(B_u) > \frac{1}{z}$, then we say that B_u is heavy, otherwise it is light.

Observe that each heavy interval consists of a single element.

In order to relate between $\mathcal{H}$ and $\mathcal{B}$, we introduce the following event, based on the sample S_1.

▶ **Definition 13.** *Denote by E_1 the event where*

$$\forall H \in \mathcal{H}_{sin} \cup \mathcal{H}_{med}, \;\; \frac{1}{2}\mathrm{wt}_p(H) \le \mathrm{wt}_{S_1}(H) \le \frac{3}{2}\mathrm{wt}_p(H) \;, \tag{20}$$

$$\forall H \in \mathcal{H}_{sml}, \;\; \mathrm{wt}_{S_1}(H) \le \frac{1}{2z} \;. \tag{21}$$

▷ Claim 14. If the size of the sample S_1 is $s_1 = 120z\log(240z)$, then $\Pr[E_1] \ge \frac{8}{10}$, where the probability is over the choice of S_1.

▷ Claim 15. Conditioned on the event E_1, for every $u \in [U]$ such that B_u is light, $\mathrm{wt}_p(B_u) < \frac{6}{z}$.

3.2 Estimation of symbol density and weight of intervals

In this subsection we estimate the weight, according to p, of every interval $[b_u]$ for $u \in U$, as well as its symbol density, focusing on symbols that occur in w. Note that $[b_u]$ is the union of the intervals $B_1, \ldots, B_u$. We first introduce some notations.

For any word w^*, text T^*, $i \in [|w^*|]$ and $j \in [|T^*|]$, let $I_i^j(T^*, w^*) = 1$ if $T^*[j] = w^*_i$ and 0 otherwise. We next set

$$\xi_i^u = \sum_{j \in [b_u]} I_i^j(T, w) p_j \;. \tag{22}$$

Consider a sample S_2 of size s_2 selected according to p (with repetitions), where s_2 will be set subsequently. For every $u \in [U]$ and $i \in [k]$, set

$$\breve{\xi}_i^u = \frac{1}{s_2} \sum_{j \in [b_u]} I_i^j(T, w) N_{S_2}(j) \; . \tag{23}$$

▶ **Definition 16.** *The event E_2 (based on S_2) is defined as follows. For every $i \in [k]$ and $u \in [U]$,*

$$\left| \breve{\xi}_i^u - \xi_i^u \right| \le \frac{1}{z} \; , \tag{24}$$

and for every $u \in [U]$

$$|\mathrm{wt}_{S_2}([b_u]) - \mathrm{wt}_p([b_u])| \le \frac{1}{z} \; . \tag{25}$$

▷ Claim 17. If the size of the sample S_2 is $s_2 = z^2 \log(40kU)$, then $\Pr[E_2] \ge \frac{9}{10}$, where the probability is over the choice of S_2.

3.3 Reducing from distribution-free to uniform

In this subsection we give the aforementioned reduction from the distribution-free case to the uniform case, using the intervals and estimators that were defined in the previous subsections. We start by providing three definitions, taken from [30], which will be used in the reduction. The first two definitions are for the notion of *splitting* (variants of this notion were also used in previous works, e.g., [15]).

▶ **Definition 18.** *For a text $T = t_1 \dots t_n$, a text $\widetilde{T}$ is said to be a splitting of T if $\widetilde{T} = t_1^{\alpha_1} \dots t_n^{\alpha_n}$ for some $\alpha_1 \dots \alpha_n \in \mathbb{N}^+$. We denote by ϕ the splitting map, which maps each (index of a) symbol of $\widetilde{T}$ to its origin in T. Formally, $\phi : [|\widetilde{T}|] \to [n]$ is defined as follows. For every $\ell \in [|\widetilde{T}|] = [\sum_{i=1}^n \alpha_i]$, let $\phi(\ell)$ be the unique $i \in [n]$ that satisfies $\sum_{r=1}^{i-1} \alpha_r < \ell < \sum_{r=1}^{i} \alpha_r$.*

Note that by this definition, ϕ is a non-decreasing surjective map, satisfying $\widetilde{T}[\ell] = T[\phi(\ell)]$ for every $\ell \in [|\widetilde{T}|]$. For a set $S \subseteq [|\widetilde{T}|]$ we let $\phi(S) = \{\phi(\ell) : \ell \in S\}$. With a slight abuse of notation, for any $i \in [n]$ we use $\phi^{-1}(i)$ to denote the set $\left\{\ell \in [|\widetilde{T}|] : \phi(\ell) = i\right\}$, and for a set $S \subseteq [n]$ we let $\phi^{-1}(S) = \left\{\ell \in [|\widetilde{T}|] : \phi(\ell) \in S\right\}$

▶ **Definition 19.** *Given text $T = t_1 \dots t_n$ and a corresponding probability distribution $p = (p_1, \dots, p_n)$, a splitting of (T, p) is a text $\widetilde{T}$ along with a corresponding probability distribution $\hat{p} = (\hat{p}_1, \dots, \hat{p}_{|\widetilde{T}|})$, such that $\widetilde{T}$ is a splitting of T and $\sum_{\ell \in \phi^{-1}(i)} \hat{p}_\ell = p_i$ for every $i \in [n]$.*

The third definition is of a set of words, where no two consecutive symbols are the same.

▶ **Definition 20.** *Let $\mathcal{W}_c = \{w : w_{j+1} \ne w_j, \forall j \in [k-1]\}$.*

3.3.1 A basis for reducing from distribution-free to uniform

Let $\widetilde{w}$ be a word of length $\widetilde{k}$ and $\widetilde{T}$ a text of length $\widetilde{n}$. In this subsection we establish a claim, which gives sufficient conditions on a (normalized version) of an estimation matrix $\widehat{\mathcal{N}}$, under which it can be used to obtain an estimate of $\Delta(\widetilde{T}, \widetilde{w})$ with a small additive error.

We first state a claim that is similar to Claim 6, with a small, but important difference, that takes into account intervals in $\widetilde{T}$ (determined by a set of indices J) that consist of repetitions of a single symbol. Recall that $M(\cdot)$ was defined in Definition 5, and that $R(\widetilde{T}, \widetilde{w})$ denotes the number of role-disjoint copies of $\widetilde{w}$ in $\widetilde{T}$.

▷ Claim 21. Let $J = \{j_0, j_1, \dots, j_\ell\}$ be a set of indices satisfying $j_0 = 0 < j_1 < j_2 < \cdots < j_\ell = \widetilde{n}$. Let $\mathcal{N}$ be the matrix whose entries are $\mathcal{N}_i^r = N_i^{j_r}(\widetilde{T}, \widetilde{w})$ for every $i \in [\widetilde{k}]$ and $r \in [\ell]$. Let $J' = \{r \in [\ell] \,:\, \widetilde{T}[j_{r-1}+1] = \cdots = \widetilde{T}[j_r]\}$. Then

$$\left| M(\mathcal{N}) - R(\widetilde{T}, \widetilde{w}) \right| \le (\widetilde{k} - 1) \cdot \max_{r \in [\ell] \setminus J'} \{(j_r - j_{r-1})\} \ .$$

The following observation can be easily proved by induction.

▶ **Observation 22.** *Let $\widehat{\mathcal{N}}$ be a matrix of size $\widetilde{k} \times \ell$. Then*

$$\frac{1}{\widetilde{n}} M(\widehat{\mathcal{N}}) = M\left(\frac{\widehat{\mathcal{N}}}{\widetilde{n}}\right) \ . \tag{26}$$

The next claim will serve as the basis for our reduction from the general, distribution-free case, to the uniform case.

▷ Claim 23. Let $\widehat{\mathcal{N}}$ be a $\widetilde{k} \times \ell$ matrix, $J = \{j_0, j_1, j_2, \dots, j_\ell\}$ be a set of indices satisfying $j_0 = 0 < j_1 < j_2 < \cdots < j_\ell = \widetilde{n}$ and let c_1 and c_2 be constants. Suppose that the following conditions are satisfied.

1. For every $r \in [\ell]$, if $j_r - j_{r-1} > c_1 \cdot \frac{\delta \widetilde{n}}{\widetilde{k}}$, then $\widetilde{T}[j_{r-1}+1] = \cdots = \widetilde{T}[j_r]$.
2. For every $i \in [\widetilde{k}]$ and $r \in [\ell]$, $\left| \widehat{\mathcal{N}}_i^r - N_i^{j_r}(\widetilde{T}, \widetilde{w}) \right| \le c_2 \cdot \frac{\delta \widetilde{n}}{\widetilde{k}}$.

Then,

$$\left| M\left(\frac{\widehat{\mathcal{N}}}{\widetilde{n}}\right) - \Delta(\widetilde{T}, \widetilde{w}) \right| \le (c_1 + 2c_2)\delta \ .$$

3.3.2 Establishing the reduction for $w \in \mathcal{W}_c$ and quantized p

For the ease of readability, in this subsection we address the special case in which $w \in \mathcal{W}_c$ (recall Definition 20), and in the full version of this paper [14] we show how to deal with the general case.

For the case considered in this subsection, let $\widetilde{T} = t_1^{\alpha_1} \dots t_n^{\alpha_n}$ where $\alpha_j = \frac{p_j}{\beta}$ for every $j \in [n]$, so that $|\widetilde{T}| = \frac{1}{\beta}$. Define $\widetilde{p}$ by $\widetilde{p}_j = \beta$ for every $j \in [|\widetilde{T}|]$, so that $\widetilde{p}$ is the uniform distribution. Since $p_j = \beta \cdot \alpha_j$, for every $j \in [n]$, we get that $(\widetilde{T}, \widetilde{p})$ is a splitting of (T, p) (recall Definition 19), and hence by [30, Clm. 4.4] (using the assumption that $w \in \mathcal{W}_c$),

$$\Delta(\widetilde{T}, w, \widetilde{p}) = \Delta(T, w, p) \ . \tag{27}$$

Denote $\widetilde{n} = |\widetilde{T}|$. We begin by defining a set of intervals of $[\widetilde{n}]$, where $\{b_0, \dots, b_U\}$ and $\mathcal{B} = \{B_1, \dots, B_U\}$ are as defined in Section 3.1, and ϕ is as in Definition 19.

▶ **Definition 24.** *Let $\widetilde{b}_0 = 0$, and for every $u \in [U]$, let $\widetilde{b}_u = \max\{h \in [\widetilde{n}] : \phi(h) = b_u\}$. For every $u \in [U]$ let $\widetilde{B}_u = [\widetilde{b}_{u-1}+1, \widetilde{b}_u]$, and define $\widetilde{\mathcal{B}} = \left\{\widetilde{B}_u\right\}_{u=1}^{U}$.*

We next introduce a notation for the weights, according to $\widetilde{p}$, of unions of these intervals. For every $i \in [k]$ and $u \in [U]$,

$$\widetilde{\xi}_i^u = \sum_{j \in [\widetilde{b}_u]} I_i^j(\widetilde{T}, w) \widetilde{p}_j \ . \tag{28}$$

Note that

$$\widetilde{\xi}_i^u = \frac{1}{\widetilde{n}} N_i^{b_u}(\widetilde{T}, w) \ . \tag{29}$$

▷ Claim 25. For every $i \in [k]$ and $u \in [U]$ $\widetilde{\xi}_i^u = \xi_i^u$, where ξ_i^u is as defined in Equation (22).

We can now state and prove the following lemma.

▶ **Lemma 26.** *Let w be a word of length k in $\mathcal{W}_c$, T a text of length n, and p a distribution over $[n]$ for which there exists $\beta \in (0,1)$ such that p_j/β is an integer for every $j \in [n]$. There exists an algorithm that, given a parameter $\delta \in (0,1)$, takes a sample of size $\Theta\left(\frac{k^2}{\delta^2} \cdot \log\left(\frac{k}{\delta}\right)\right)$ from T, distributed according to p, and outputs an estimate $\widehat{\Delta}$ such that $|\widehat{\Delta} - \Delta(T, w, p)| \leq \delta$ with probability at least $2/3$.*

As in the uniform case, the running time of the algorithm is linear in the size of the sample.

Proof. The algorithm first takes a sample S_1 of size $s_1 = 120z\log(240z)$ and constructs a set of intervals $\mathcal{B}$ as defined in Definition 12. Next the algorithm takes another sample, S_2, of size $s_2 = z^2\log(40kU)$ according to which it defines an estimation matrix $\widehat{\xi}$ of size $k \times U$ as follows. For every $i \in [k]$ and $u \in [U]$, it sets $\widehat{\xi}[i][u] = \breve{\xi}_i^u$, where $\breve{\xi}_i^u$ is as defined in Equation (23). Lastly the algorithm outputs $\widehat{\Delta} = M(\widehat{\xi})$, where M is as defined in Definition 5.

We would like to apply Claim 23 in order to show that $|\widehat{\Delta} - \Delta(\widetilde{T}, w)| \leq \delta$ with probability of at least $\frac{2}{3}$. By the setting of s_1, applying Claim 14 gives us that with probability at least $\frac{8}{10}$, the event E_1, as defined in Definition 13, holds. By the setting of s_2, applying Claim 17 gives us that with probability at least $\frac{9}{10}$, the event E_2, as defined in Definition 16, holds. We henceforth condition on both events (where they hold together with probability at least 7/10).

In order to apply Claim 23, we set $\widetilde{w} = w$, $J = \left\{\widetilde{b}_0, \widetilde{b}_1, \ldots, \widetilde{b}_U\right\}$ (recall Definition 24) and $\widehat{\mathcal{N}} = \widetilde{n}\widehat{\xi}$, for $\widehat{\xi}$ as defined above. Also, we set $c_1 = \frac{1}{2}$ and $c_2 = \frac{1}{4}$. We next show that both items in the premise of the claim are satisfied.

To show that Item 1 is satisfied, we first note that since $\widetilde{p}$ is uniform, then for every $u \in U$, $\mathrm{wt}_{\widetilde{p}}(b_u) = \frac{\widetilde{b}_u - \widetilde{b}_{u-1}}{\widetilde{n}}$. We use the consequence of Claim 15 (recall that we condition on E_1) by which for every u such that $\frac{\widetilde{b}_u - \widetilde{b}_{u-1}}{\widetilde{n}} \geq \frac{6}{z}$, B_u is heavy (since for every $u \in U$, $\mathrm{wt}_{\widetilde{p}}(\widetilde{B}_u) = \mathrm{wt}_p(B_u)$). By Definition 12 this implies that B_u contains only one index, and so $\widetilde{T}[\widetilde{b}_{u-1} + 1] = \cdots = \widetilde{T}[\widetilde{b}_u]$. By the definition of z (Equation (19)) and the setting of c_1, the item is satisfied.

To show that Item 2 is satisfied, we use the definition of E_2 (Definition 16, Equation (24)) together with Claim 25, which give us $|\widehat{\xi}_i^u - \widetilde{\xi}_i^u| \leq \frac{1}{z}$ for every $i \in [k]$ and $u \in [U]$. By Equation (29), the definition of z and the setting of c_2, we get that the item is satisfied.

After applying Claim 23 we get that $|\widehat{\Delta} - \Delta(\widetilde{T}, w)| \leq (c_1 + 2c_2)\delta$, which by the setting of c_1 and c_2 is at most δ. Since $\widetilde{p}$ is the uniform distribution, $\Delta(\widetilde{T}, w) = \Delta(\widetilde{T}, w, \widetilde{p})$ and since $\Delta(\widetilde{T}, w, \widetilde{p}) = \Delta(T, w, p)$ (by Equation (27)), the lemma follows. ◀

In the full version of this paper [14] we address the general case where we do not necessarily have that $w \in \mathcal{W}_c$ or that there exists a value β such that for every $j \in [n]$, p_j/β is an integer.

4 A lower bound for distance approximation

In this section we give a lower bound for the number of samples required to perform distance-approximation from w-freeness of a text T. The lower bound holds when the underlying distribution is the uniform distribution.

▶ **Theorem 27.** *Let k_d be the number of distinct symbols in w. Any distance-approximation algorithm for w-freeness under the uniform distribution must take a sample of size $\Omega(\frac{1}{k_d\delta^2})$, conditioned on $\delta \leq \frac{1}{300k_d}$ and $n > \max\left\{\frac{8k}{\delta}, \frac{200}{k_d\delta^2}\right\}$.*

Note that if $\delta \geq 1/k_d$, then the algorithm can simply output 0. This is true since the number of role disjoint copies of w in T is at most the number of occurrences of the symbol in w that is least frequent in T. This number is upper bounded by $\frac{n}{k_d}$, and so the distance from w-freeness is at most $\frac{1}{k_d}$. In this case no sampling is needed, so only the trivial lower bound holds. The proof will deal with the case of $\delta \in (0, \frac{1}{300k_d}]$.

Proof. The proof is based on the difficulty of distinguishing between an unbiased coin and a coin with a small bias. Precise details follow.

Let $V = \{v_1, \ldots, v_{k_d}\}$ be the set of distinct symbols in w, and let 0 be a symbol that does not belong to V. We define two distributions over texts, $\mathcal{T}_1$ and $\mathcal{T}_2$ as follows. For each $\tau \in [\frac{n}{k_d}]$ and $\rho \in [0,1]$, let λ_ρ^τ be a random variable that equals 0 with probability ρ and equals v_1 with probability $1-\rho$. Let $\delta' = 3k_d\delta$ and consider the following two distributions over texts

$$\mathcal{T}_1 = \left[\lambda_{\frac{1}{2}}^1, v_2, v_3, \ldots, v_{k_d}, \lambda_{\frac{1}{2}}^2, v_2, v_3, \ldots, v_{k_d}, \ldots\ldots, \lambda_{\frac{1}{2}}^{n/k_d}, v_2, v_3, \ldots, v_{k_d}\right], \tag{30}$$

$$\mathcal{T}_2 = \left[\lambda_{\frac{1}{2}+\delta'}^1, v_2, v_3, \ldots, v_{k_d}, \lambda_{\frac{1}{2}+\delta'}^2, v_2, v_3, \ldots, v_{k_d}, \ldots\ldots, \lambda_{\frac{1}{2}+\delta'}^{n/k_d}, v_2, v_3, \ldots, v_{k_d}\right]. \tag{31}$$

Namely, the supports of both distributions contain texts that consist of n/k_d blocks of size k_d each. For $i \in \{2, \ldots, k_d\}$, the i-th symbol in each block is v_i. The distributions differ only in the way the first symbol in each block is selected. In $\mathcal{T}_1$ it is 0 with probability 1/2 and v_1 with probability 1/2, while in $\mathcal{T}_2$ it is 0 with probability $1/2 + \delta' = 1/2 + 3\delta k_d$, and v_1 with probability $1/2 - \delta'$.

For $b \in \{1,2\}$, consider selecting a text T_b according to $\mathcal{T}_b$ (denoted by $T_b \sim \mathcal{T}_b$), and let O_b be the number of occurrences of v_1 in the text (so that O_b is a random variable). Observe that $\mathbb{E}[O_1] = \frac{n}{2k_d}$ and $\mathbb{E}[O_2] = \frac{n}{2k_d} - 3\delta n$. By applying the additive Chernoff bound (Theorem 28) and using the premise of the theorem regarding n,

$$\Pr_{T_1\sim\mathcal{T}_1}[O_1 < \mathbb{E}[O_1] - \delta n/8] \leq \exp(-2(k_d\delta/8)^2 \cdot n/k_d) \leq \frac{1}{100}, \tag{32}$$

and

$$\Pr_{T_2\sim\mathcal{T}_2}[O_2 < \mathbb{E}[O_2] + \delta n/8] \leq \exp(-2(k_d\delta/8)^2 \cdot n/k_d) \leq \frac{1}{100}. \tag{33}$$

For $b \in \{1,2\}$ let $R_b = R(T_b, w)$ (recall that $R(T_b, w)$ denotes the number of disjoint copies of w in T_b, and note that R_b is a random variable). Observe that $R_1 \geq O_1 - k + 1$, and $R_2 \leq O_2$.

Hence, by Equation (32), if we select T_1 according to $\mathcal{T}_1$ and use the premise that $n > \frac{8k}{\delta}$, then $R(T_1, w) \geq \frac{n}{2k_d} - \frac{1}{8}\delta n - k + 1 \geq \frac{n}{2k_d} - \frac{2}{8}\delta n$ with probability at least 99/100, and by Equation (33), if we select T_2 according to $\mathcal{T}_2$, then $R(T_2, w) \leq \frac{n}{2k_d} - 3\delta n + \frac{1}{8}\delta n = \frac{n}{2k_d} - \frac{23}{8}\delta n$ with probability at least 99/100.

Assume, contrary to the claim, that we have a sample-based distance-approximation algorithm for subsequence-freeness that takes a sample of size $Q(k_d, \delta) = 1/(ck_d\delta^2)$, for some sufficiently large constant c, and outputs an estimate of the distance to w-freeness that has additive error at most δ, with probability at least 2/3. Consider running the algorithm on either $T_1 \sim \mathcal{T}_1$ or $T_2 \sim \mathcal{T}_2$. Let L denote the number of times that the sample landed on an index of the form $j = \ell \cdot k_d + 1$ for an integer ℓ. By Markov's inequality, the probability that $L > 10 \cdot Q(k_d, \delta)/k_d = 10/(ck_d^2\delta^2)$ is at most 1/10.

By the above, if we run the algorithm on $T_1 \sim \mathcal{T}_1$, then with probability at least $2/3 - 1/100 - 1/10$ the algorithm outputs an estimate $\widehat{\Delta} \geq \frac{n}{2k_d} - \frac{10}{8}$ while $L \leq 10/(ck_d^2\delta^2)$. Similarly, if we run it on $T_2 \sim \mathcal{T}_2$, then with probability at least $2/3 - 1/100 - 1/10$ the algorithm outputs an estimate $\widehat{\Delta} \leq \frac{n}{2k_d} - \frac{15}{8}$ while $L \leq 10/(ck_d^2\delta^2)$. (In both cases the probability is taken over the selection of $T_b \sim \mathcal{T}_b$, the sample that the algorithm gets, and possibly additional internal randomness of the algorithm.) Based on the definitions of $\mathcal{T}_1$ and $\mathcal{T}_2$, this implies that it is possible to distinguish between an unbiased coin and a coin with bias $3k_d\delta$ with probability at least $2/3 - 1/100 - 1/10 > \frac{8}{15}$, using a sample of size $\frac{1}{c'k_d^2\delta^2}$ in contradiction to the result of Bar-Yosef [2, Thm. 8] (applied with $m = 2$, $\epsilon = 3k_d\delta$. Since we have $\delta < \frac{1}{300k_d}$, then $\epsilon < \frac{1}{96}$, as the cited theorem requires). ◀

References

1 Nir Ailon, Bernard Chazelle, Seshadhri Comandur, and Ding Liu. Estimating the distance to a monotone function. *Random Structures and Algorithms*, 31(3):371–383, 2007.

2 Ziv Bar-Yossef. Sampling lower bounds via information theory. In *Proceedings of the 35th Annual ACM Symposium on the Theory of Computing*, pages 335–344, 2003.

3 Omri Ben-Eliezer, Eldar Fischer, Amit Levi, and Ron D. Rothblum. Hard properties with (very) short PCPPs and their applications. In *Proceedings of the 11th Innovations in Theoretical Computer Science conference (ITCS)*, pages 9:1–9:27, 2020.

4 Piotr Berman, Meiram Murzabulatov, and Sofya Raskhodnikova. Tolerant testers of image properties. *ACM Transactions on Algorithms*, 18(4):1–39, 2022. Article number 37.

5 Piotr Berman, Sofya Raskhodnikova, and Grigory Yaroslavtsev. Lp-testing. In *Proceedings of the 46th Annual ACM Symposium on the Theory of Computing*, pages 164–173, 2014.

6 Hadley Black, Deeparnab Chakrabarty, and C. Seshadhri. Domain reduction for monotonicity testing: A $o(d)$ tester for boolean functions in d-dimensions. In *Proceedings of the 31st Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1975–1994, 2020.

7 Eric Blais, Clément L Canonne, Talya Eden, Amit Levi, and Dana Ron. Tolerant junta testing and the connection to submodular optimization and function isomorphism. *ACM Transactions on Computation Theory*, 11(4):1–33, 2019.

8 Eric Blais, Renato Ferreira Pinto Jr., and Nathaniel Harms. VC dimension and distribution-free sample-based testing. In *Proceedings of the 53rd Annual ACM Symposium on the Theory of Computing*, pages 504–517, 2021.

9 Avrim Blum and Lunjia Hu. Active tolerant testing. In *Proceedings of the 31st Conference on Computational Learning Theory (COLT)*, pages 474–497, 2018.

10 Mark Braverman, Subhash Khot, Guy Kindler, and Dor Minzer. Improved monotonicity testers via hypercube embeddings. In *Proceedings of the 13th Innovations in Theoretical Computer Science conference (ITCS)*, pages 25:1–25:24, 2024.

11 Andrea Campagna, Alan Guo, and Ronitt Rubinfeld. Local reconstructors and tolerant testers for connectivity and diameter. In *Proceedings of the 17th International Workshop on Randomization and Computation*, pages 411–424, 2013.

12 Clément L Canonne, Elena Grigorescu, Siyao Guo, Akash Kumar, and Karl Wimmer. Testing k-monotonicity: The rise and fall of boolean functions. *Theory of Computing*, 15(1):1–55, 2019. This paper appeared in the proceedings of ITCS 2017.

13 Omer Cohen Sidon. Sample-based distance-approximation for subsequence-freeness. MSc thesis, Tel Aviv University, 2023.

14 Omer Cohen Sidon and Dana Ron. Sample-based distance-approximation for subsequence-freeness. *arXiv preprint*, 2023. `arXiv:2305.01358`.

15 Ilias Diakonikolas and Daniel Kane. A new approach for testing properties of discrete distributions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, pages 685–694, 2016.

16 Shahar Fattal and Dana Ron. Approximating the distance to monotonicity in high dimensions. *ACM Transactions on Algorithms*, 6(3):1–37, 2010.

17 Nimrod Fiat and Dana Ron. On efficient distance approximation for graph properties. In *Proceedings of the 32nd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1618–1637, 2021.

18 Eldar Fischer and Lance Fortnow. Tolerant versus intolerant testing for boolean properties. *Theory of Computing*, 2:173–183, 2006.

19 Eldar Fischer and Ilan Newman. Testing versus estimation of graph properties. *SIAM Journal on Computing*, 37(2):482–501, 2007.

20 Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connections to learning and approximation. *Journal of the ACM*, 45:653–750, 1998.

21 Venkat Guruswami and Atri Rudra. Tolerant locally testable codes. In *Proceedings of the 9th International Workshop on Randomization and Computation*, pages 306–317, 2005.

22 Nathaniel Harms and Yuichi Yoshida. Downsampling for testing and learning in product distributions, 2022.

23 Carlos Hoppen, Yoshiharu Kohayakawa, Richard Lang, Hanno Lefmann, and Henrique Stagni. Estimating the distance to a hereditary graph property. *Electronic Notes in Discrete Mathematics*, 61:607–613, 2017.

24 Swastik Kopparty and Shubhangi Saraf. Tolerant linearity testing and locally testable codes. In *Proceedings of the 13th International Workshop on Randomization and Computation*, pages 601–614, 2009.

25 Amit Levi and Erik Waingarten. Lower bounds for tolerant junta and unateness testing via rejection sampling of graphs. In *Proceedings of the 10th Innovations in Theoretical Computer Science conference (ITCS)*, pages 52:1–52:20, 2019.

26 Sharon Marko and Dana Ron. Distance approximation in bounded-degree and general sparse graphs. *Transactions on Algorithms*, 5(2), 2009. Article number 22.

27 Ilan Newman and Nithin Varma. New sublinear algorithms and lower bounds for LIS estimation. In *Automata, Languages and Programming: 48th International Colloquium*, pages 100:1–100:20, 2021.

28 Ramesh Krishnan S Pallavoor, Sofya Raskhodnikova, and Erik Waingarten. Approximating the distance to monotonicity of boolean functions. *Random Structures & Algorithms*, 60(2):233–260, 2022.

29 Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *Journal of Computer and System Sciences*, 72(6):1012–1042, 2006.

30 Dana Ron and Asaf Rosin. Optimal distribution-free sample-based testing of subsequence-freeness with one-sided error. *ACM Transactions on Computation Theory*, 14(4):1–31, 2022. An extended abstract of this work appeared in the proceedings of SODA 2021.

31 Ronitt Rubinfeld and Madhu Sudan. Robust characterization of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, 1996.

A Chernoff bounds

▶ **Theorem 28.** *Let $\chi_1, \ldots, \chi_m$ be m independent random variables where $\chi_i \in [0,1]$ for every $1 \le i \le m$. Let $p \overset{\text{def}}{=} \frac{1}{m}\sum_i \mathbb{E}[\chi_i]$. Then, for every $\gamma \in (0,1]$, the following bounds hold:*

- *(Additive Form)*

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m}\chi_i > p + \gamma\right] < \exp\left(-2\gamma^2 m\right) \tag{34}$$

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m}\chi_i < p - \gamma\right] < \exp\left(-2\gamma^2 m\right) \tag{35}$$

- *(Multiplicative Form)*

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m}\chi_i > (1+\gamma)p\right] < \exp\left(-\gamma^2 pm/3\right) \tag{36}$$

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m}\chi_i < (1-\gamma)p\right] < \exp\left(-\gamma^2 pm/2\right) \tag{37}$$