Tight Bounds for Monotone Minimal Perfect Hashing*

Sepehr Assadi
Rutgers University

Martín Farach-Colton
Rutgers University

William Kuszmaul
MIT

Abstract

The monotone minimal perfect hash function (MMPHF) problem is the following indexing problem. Given a set $S = \{s_1, \ldots, s_n\}$ of n distinct keys from a universe U of size u, create a data structure D that answers the following query:

$$\mathrm{Rank}(q) = \begin{cases} \text{rank of } q \text{ in } S & q \in S \\ \text{arbitrary answer} & \text{otherwise.} \end{cases}$$

Solutions to the MMPHF problem are in widespread use in both theory and practice.

The best upper bound known for the problem encodes D in $O(n \log\log\log u)$ bits and performs queries in $O(\log u)$ time. It has been an open problem to either improve the space upper bound or to show that this somewhat odd looking bound is tight.

In this paper, we show the latter: any data structure (deterministic or randomized) for monotone minimal perfect hashing of any collection of n elements from a universe of size u requires $\Omega(n \log\log\log u)$ expected bits to answer every query correctly.

We achieve our lower bound by dening a graph G where the nodes are the possible $\binom{u}{n}$ inputs and where two nodes are adjacent if they cannot share the same D. The size of D is then lower bounded by the log of the chromatic number of G. Finally, we show that the fractional chromatic number (and hence the chromatic number) of G is lower bounded by $2^{\Omega(n \log\log\log u)}$.

1 Introduction

The monotone minimal perfect hash function (MMPHF) problem is the following indexing problem. Given a set $S = \{s_1, \ldots, s_n\}$ of n distinct keys from a universe U of size u, create a data structure D that answers the following query:

$$\mathrm{Rank}(q) = \begin{cases} \text{rank of } q \text{ in } S & q \in S \\ \text{arbitrary answer} & \text{otherwise.} \end{cases}$$

The name MMPHF comes from interpreting the data structure D as a hash function: given a sorted array $A = [a_1, \ldots, a_n]$, D is hashing each a_i to its position i. The hash function is minimal, meaning it maps n items to n distinct positions, and monotone, meaning $a_i < a_j$ iff $D(a_i) < D(a_j)$.

It may seem strange at rst glance that D is permitted to return arbitrary answers on negative queries. A key insight, however, is that this relaxation allows for asymptotic improvements in space eciency: whereas the set S would require $\Omega(n \log(u/n))$ bits to encode, Belazzougui, Boldi, Pagh and Vigna [BBPV09] show that it is possible to construct an MMPHF D using as few as $O(n \log\log\log u)$ bits, while supporting $O(\log u)$-time queries.

The remarkable space eciency of MMPHF makes it useful for a variety of practical applications (e.g., in security [BCO11], key-value stores [LFAK11] and information retrieval [Nav14]). A high-performance implementation can be found in the Sux4J library [BV08, BBPV11]. MMPHF has also been widely used in

*A full version of the paper is available at: https://arxiv.org/abs/2207.10556.

the theory community for the design of space-ecient combinatorial pattern-matching algorithms (see, e.g., [BN14, GNP20, Bel14, BN15, CFP+15, BCKM20, BGMP16, GOR10]).

Despite the widespread use of MMPHF, it remains an open question [BBPV09, Bol15, D+18] to determine the optimal bounds for solving this problem. The best lower bound achieved so far [BBPV11, D+18] is (n) bits (which follows immediately from the same lower bound for minimal perfect hashing [Meh82]). Even disregarding applications (and the running time to answer queries), the information-theoretic question as to how many bits a MMPHF requires has been posed as a problem of independent combinatorial interest [D+18].

Our result. We fully settle this question by establishing the following result:

> **Theorem 1** (Formalized in Theorem 2). Any data structure (deterministic or randomized) for monotone minimal perfect hashing of any collection of n elements from a universe of size u requires $(n \log\log\log u)$ expected bits to answer every query correctly. The lower bound holds whenever u is at least $n^{1+1/\sqrt{\log n}}$ and at most $\exp(\exp(\mathrm{poly}(n)))$.

Thus, surprisingly, the $O(n \log\log\log u)$ bound achieved by [BBPV09] is asymptotically optimal. The boundary conditions on u in Theorem 1 are also natural in the following sense. There are two trivial solutions for MMPHF. One encodes the input set S in $O(u)$ bits and the other builds a perfect hash table from elements of S to their rank in $O(n \log n)$ space. When u is very small, say, $u = O(n)$, the rst solution uses $O(u) = o(n \log\log\log u)$ bits. And when u is very large, that is when u is even beyond $\exp(\exp(\mathrm{poly}(n)))$, then the $O(n \log n)$-bit solution uses $o(n \log\log\log u)$ bits. (See also the variable-size bucketing reduction of [BBPV11] which reduces the universe size from u to u/n). Our lower bound in Theorem 1 covers almost the entire range in between.

The lower bound achieved by Theorem 1 is remarkably general: it applies independently of the running time of the data structure; and it applies even to randomized data structures that are permitted to store their random bits for free.

Our techniques. The most intuitive approach toward proving a lower bound of d bits on the size of an MMPHF is to encode a d-bit string into the state of the data structure. This approach is already hindered by the fact that MMPHFs only support positive queries, however. If the user already knows which elements are in the input, then the MMPHF encodes no interesting information — but if the user only has partial information about the input, then the user can only get useful information from a small portion of possible MMPHF queries. The previous (n) lower bound of [Meh82, BBPV11, D+18] addresses this as follows: consider any bit-string $x \in \{0,1\}^d$ and dene:

$$S(x) := \{3, 6, \ldots, 3d\} \cup \{3i+1 \mid i \in [d], x_i = 1\} \cup \{3i - 1 \mid i \in [d], x_i = 0\}.$$

For every $i \in [d]$, rstly, $3i$ belongs to $S(x)$ and thus is a positive query, and secondly, $\mathrm{Rank}(3i) = 2(i-1) + x_i$. This allows us to recover x from any MMPHF for $S(x)$, proving a lower bound of $d = (n)$ bits for MMPHF on size-n subsets of universe $[3n+1]$. This approach, however, seems to be stuck at proving any $\omega(n)$ lower bound as these "direct encodings" ignore the delicate interaction between dierent elements in the input set[1].

To get around these obstacles, we take a dierent approach to proving Theorem 1. We construct a "conict graph" G whose vertices are the possible inputs to an MMPHF problem for a xed n and u. Two vertices are adjacent in G if they cannot have the same MMPHF index, that is, if the vertices share an element but with a dierent rank. Any MMPHF induces a proper coloring of this graph, where the color of a vertex corresponds to its MMPHF representation. As a result, the chromatic number of the conict graph is a lower bound on how many dierent MMPHF representations we must have, implying that some input must have a representation of size at least $\log \chi(G)$ bits. This reduces our task to combinatorial problem of lower bounding $\chi(G)$.[2]

The problem of bounding chromatic number of graphs dened over these types of set-systems has a rich history in the discrete math literature; see, e.g. [EH66, FHRT92, DLR95, ST11]. For instance, Erdős and Hajnal [EH66]

[1] Any lower bound of d bits for a data structure immediately implies an encoding of d-bit strings in the state of the data structure by just assigning one bit-string to each state. This means that there is never a formal proof that one cannot encode a bit-string in a data structure and still prove a lower bound.

[2] Slightly more care must be taken when bounding the *expected* size of a MMPHF that is permitted to take dierent sizes on dierent inputs.

study shift-graphs that have vertices corresponding to n-element subsets of $[u]$ and edges between vertices $(a_1; a_2; \dots; a_n)$ and $(a_2; \dots; a_n; a_{n+1})$ for all $a_1 < a_2 < \dots < a_{n+1}$. They prove that the chromatic number of the shift-graph is $(1 + o(1)) \cdot \log^{(n-1)}(u)$, namely, the $(n-1)$-th iterated logarithm of u. The shift-graph is a subgraph of our conflict graph. Thus, by taking $u = 2 \uparrow\uparrow \Theta(n)$, i.e., the tower of twos of height $\Theta(n)$, we can have $\chi(G) = 2^{\omega(n)}$, and thus prove an $\omega(n)$ lower bound for MMPHF on n-subsets of (extremely large) universes of size $u = 2 \uparrow\uparrow \Theta(n)$. This is the starting point of our approach. We now need to dramatically decrease the size of the universe, while also dramatically increasing the bound on the chromatic number by considering the conflict graph itself, and not only its shift-subgraph.

To lower bound the chromatic number of the conflict graph, we consider the relaxation of this problem via fractional colorings (see Section 2.2). Given that this latter problem can be formulated as a linear program (LP), a natural way for proving a lower bound on its value is to exhibit a feasible dual solution instead[3]. This corresponds to the following problem: exhibit a distribution on vertices of the graph so that for any independent set, the probability that a vertex sampled from the distribution belongs to the independent set is bounded by p; this then implies that the fractional chromatic number (and in turn the chromatic number) are lower bounded by $1/p$. The main technical novelty of our work lies in the introduction of a highly non-trivial such distribution and the analysis of this probability bound for each independent set (we postpone the overview of this part to Section 4.1 after we setup the required background). This allows us to lower bound the (fractional) chromatic number of the conflict-graph by $\Omega(n \log n)$ when the universe is of size $u = 2^{2^{\mathrm{poly}(n)}}$ which gives an $\Omega(n \log\log\log u)$ lower bound for MMPHF on such universes.

Working with fractional colorings, beside being an immensely helpful analytical tool, has several additional benefits for us. Firstly, unlike standard (integral) colorings, fractional colorings admit a natural direct product property over a certain union of graphs; this allows us to extend the lower bound for MMPHF from universes of size doubly exponential in n (which are admittedly not the most interesting setting of parameters), all the way down to universes of size $n^{1+o(1)}$. Secondly, unlike the (integral) chromatic number, which yields a lower bound only on the space of deterministic MMPHFs, we show that lower bounding the fractional chromatic number allows us to prove a lower bound even for randomized MMPHFs that have access to their randomness for free. We believe this technique, namely, defining a proper conflict graph and bounding its fractional coloring by exhibiting a feasible dual solution, may be applicable to many other data structure problems and is therefore interesting in its own right.

2 Preliminaries

Notation. For any integer $t \geq s \geq 1$, we let $[t] := \{1; \dots; t\}$ and let $[s : t] = \{s; \dots; t\}$. For a tuple $(X_1; \dots; X_t)$, we further define $X_{<i} := (X_1; \dots; X_{i-1})$ and $X_{-i} := (X_1; \dots; X_{i-1}; X_{i+1}; \dots; X_t)$.

2.1 Problem Definition and Model of Computation For any integer $n; u \geq 1$, we let $\mathcal{D}(n; u)$ be an MMPHF indexing algorithm for size-n subsets of $[u]$. That is, if $\mathcal{S}_{n;u} = \{S \subseteq [u] \text{ s.t. } |S| = n\}$ then for all $S \in \mathcal{S}_{m;u}$, $\mathcal{D}(S)$ is the MMPHF index for S.

For any fixed choice of random bits r, we use $\mathcal{D}^r$ to denote the resulting MMPHF with random bits r. Note that for any fixed choice of r, $\mathcal{D}^r$ is deterministic. For any $S \in \mathcal{S}_{n;u}$ and randomness r, define $d^r(S)$ as the size in bits of the MMPHF index $\mathcal{D}^r(S)$. Define:

$$d(n; u) := \max_{S \in \mathcal{S}_{n;u}} \mathop{\mathbb{E}}_{r}[d^r(S)] :$$

When n and u are clear, we drop them and refer simply to $\mathcal{D}$ and d.

In this definition of size, we are not charging the algorithm for storing its randomness. In other words, the algorithm has access to a tape of random bits chosen independent of the input that it can use for both creating the index as well as answering the queries. Furthermore, we also allow the algorithm unbounded computation time[4]. Thus, the only measure of interest for us is the size of the index. Finally, any deterministic MMPHF in this

[3] This is an inherently different technique than the one used in [EH66] for the shift-graph, as it is known that the fractional chromatic number of the shift-graph is $O(1)$ (see, e.g. [ST11]).

[4] In this (non-uniform) information-theoretical setting, one can remove random bits entirely by increasing the space with $O(\log n + \log\log u)$ bits (see, e.g., Newman's Theorem in communication complexity [New91]), but this extra $O(\log\log u)$ is in

model is simply a randomized MMPHF that ignores its random bits and thus we will only focus on randomized MMPHFs from now on.

2.2 Fractional Colorings A key tool that we use in establishing our lower bound is the notion of a *fractional coloring* of a graph. We now review the basics of fractional colorings, which we need in our proofs. The results mentioned in this subsection are all standard; see, e.g. [SU11] (we present self-contained proofs of these results in Appendix A for completeness).

Let $G = (V, E)$ be any undirected graph. A proper coloring of G is any assignment of colors to vertices of G so that no edge is monochromatic. The chromatic number $\chi(G)$ is the minimum number of colors in any proper coloring of G.

Let $\mathcal{I}(G) \subseteq 2^V$ denote the set of all independent sets in G, and for any vertex $v \in V$, dene $\mathcal{I}(G, v)$ as the set of all independent sets that contain the vertex v. A fractional coloring of G is any assignment of $x \in [0, 1]^{\mathcal{I}(G)}$ to the independent sets of G satisfying the following constraint:

$$\text{for every vertex } v \in V: \quad \sum_{I \in \mathcal{I}(G, v)} x_I \geqslant 1.$$

The value $|x|$ of a fractional coloring x is given by $\sum_{I \in \mathcal{I}(G,v)} x_I$. The *fractional chromatic number* $\chi_f(G)$ is the minimum value of any fractional coloring of G. This quantity can be formalized as a linear program (LP):

$$\chi_f(G) := \min_{x \in [0,1]^{\mathcal{I}(G)}} \sum_{I \in \mathcal{I}(G)} x_I \quad \text{subject to} \quad \sum_{v \in \mathcal{I}(G, v)} x_I \geqslant 1 \quad \forall v \in V. \tag{2.1}$$

Any proper coloring of G with k colors induces a solution x of value k to this LP, where x_I is set to 1 for the independent sets I that correspond to (whole) color classes in the coloring. Thus the LP given by Eq (2.1) is indeed a relaxation of the original coloring problem.

Fact 2.1. *For any graph G, $\chi_f(G) \leqslant \chi(G)$.*

It is worth mentioning that at the same time $\chi(G) = O(\log |V(G)|) \cdot \chi_f(G)$ using the standard randomized rounding argument (we do not use this direction explicitly in our paper).

A primal-dual analysis of the fractional-chromatic-number LP implies the following results. These results are standard but we provide proofs in Appendix A for completeness.

Proposition 2.2. *Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be arbitrary graphs. Dene $G_1 \vee G_2$ as a graph on vertices $V_1 \times V_2$ and dene an edge between vertices (v_1, v_2) and (w_1, w_2) whenever (v_1, w_1) is an edge in G_1 or (v_2, w_2) is an edge in G_2. Then, $\chi_f(G_1 \vee G_2) = \chi_f(G_1) \cdot \chi_f(G_2)$.*

Proposition 2.2 allows us to determine χ_f of a product of several graphs by focusing on each individual graph separately.

Proposition 2.3. *For any graph $G = (V, E)$,*

$$\chi_f(G) = \max_{\text{distribution } \mu \text{ on } V} \left(\min_{I \in \mathcal{I}(G)} \Pr_{v \sim \mu}(v \in I) \right)^{-1}.$$

Proposition 2.3 provides us with a tool to lower bound χ_f by nding a suitable distribution on the vertices so that no independent set has a signicant probability of containing a vertex sampled from this distribution.

3 A Lower Bound for MMPHF via Fractional Colorings

We can now formally state the main theorem of this paper.

general unavoidable (see, e.g. [HT01] and references therein), and can be prohibitive for us when u is suciently large. Hence, we still explicitly account for randomized data structures in our lower bound.

Theorem 2 (Formalization of Theorem 1). *For any $n, u \in \mathbb{N}^+$ such that $n \cdot 2^{\sqrt{\log n}} \leq u \leq 2^{n^{n^2+n}}$, and for any MMPHF algorithm $D(n; u)$,*

$$d(n; u) = \Omega(n \log \log \log u).$$

The rest of the paper presents the proof of Theorem 2. We spend the rest of the section reframing the theorem in terms of the fractional chromatic number of a certain graph associated with the MMPHF problem. We will then show how to lower bound the fractional chromatic number in the next section.

3.1 Conflict Graph and its Fractional Chromatic Number

Let $m \geq 1$ be an integer and define $M := 2^{m^{m^2+m}}$. Define the graph $G(m) := (V(m), E(m))$ as:

- The vertex set is $V(m) = \mathcal{S}_{m,M}$, that is, the size-m subsets of $[M]$. We denote each vertex $v \in V(M)$ by the m-tuple $v := (v_1, \ldots, v_m)$ where $0 < v_1 < v_2 < \cdots < v_m \leq M$.
- The edge set $E(m)$ is defined as follows. Let $v = (v_1, \ldots, v_m)$ and $w = (w_1, \ldots, w_m)$ be any two vertices in $V(M)$. Then, there is an edge $(v, w) \in G(m)$ iff there exists some pair of indexes $i \neq j \in [m]$ such that $v_i = w_j$.

We refer to $G(m)$ as the **conflict graph** of m. The following lemma clarifies our interest in this graph by showing that fractional chromatic number of $G(m)$ can be used to lower bound size of any MMPHF (for certain parameters of input).

Lemma 3.1. *Let $m \geq 1$ be an integer and let $M = 2^{m^{m^2+m}}$. For any MMPHF $D(m; M)$,*

$$d(m; M) \geq (\log \chi_f(G(m)) - 2)/2.$$

Proof. Consider any two vertices $v, w \in G(m)$. If there is an edge between v and w, then there exists an element $z = v_i = w_j$; $i \neq j$. Therefore for every choice of randomness r, $D^r(v) \neq D^r(w)$, because query z must return i on $D^r(v)$ and j on $D^r(w)$. This implies that for every r, the set of vertices v with the same $D^r(v)$ form an independent set in $G(m)$ (and the collection of these sets is a coloring of $G(m)$). We use $\mathcal{I}^r$ to denote these independent sets in $G(m)$ for this choice of r.

On the other hand, by Proposition 2.3, there exists a distribution μ on $V(m)$ such that

$$\chi_f(G(m)) = \left(\min_{I \in \mathcal{I}(G(m))} \Pr_{v \sim \mu}(v \in I) \right)^{-1}. \tag{3.2}$$

Let us fix that distribution. Under this distribution, by the definition of d,

$$d = d(m; M) = \max_{v \in V(m)} \mathbb{E}_r[d^r(v)] \geq \mathbb{E}_{v \sim \mu} \mathbb{E}_r[d^r(v)] = \mathbb{E}_r \mathbb{E}_{v \sim \mu}[d^r(v)].$$

An averaging argument now implies that there exists a choice r of random bits such that

$$\mathbb{E}_{v \sim \mu}\left[d^r(v) \right] \leq d.$$

By Markov's inequality, with probability at least $1/2$, for $v \sim \mu$, we have that $d^r(v) \leq 2d$.

Recall that $D^r(v)$ corresponds to an independent set in $\mathcal{I}_r$. Moreover, there can be at most $2^{2d+1} - 2$ independent sets I in $\mathcal{I}_r$ such that for all $v \in I$, $d^r(v) \leq 2d$; this is because there are at most $2^{2d+1} - 2$ choices for $D^r(v)$ across all $v \in V(m)$ that can use up to $2d$ bits in their index (as the number of non-empty binary strings of length at most $2d$ is $2^{2d+1} - 2$). Since a random $v \sim \mu$ belongs to one of these $2^{2d+1} - 2$ independent sets with probability at least half, we necessarily have some independent set $I \in \mathcal{I}_r$ where

$$\Pr_{v \sim \mu}(v \in I) \geq \frac{1}{2 \cdot (2^{2d+1} - 2)} \geq \frac{1}{2^{2d+2}}.$$

Plugging in this bound in Eq (3.2), we have,

$$\chi_f(G(m)) \leq 2^{2d+2},$$

which implies that $d \geq (\log \chi_f(G(m)) - 2)/2$, concluding the proof. ∎

Lemma 3.1 reduces our task of proving Theorem 2 to establishing a lower bound on $(G(m))$. This will be accomplished by the following lemma, which we prove in Section 4.

Lemma 3.2. There is an absolute constant > 0 such that for every suciently large $m > 1$,

$$_f(G(m)) > m^m:$$

By plugging in the lower bound of $_f(G(m))$ from Lemma 3.2 inside Lemma 3.1, we get that for any suciently large $n > 1$ and universe size $u = 2^{n^{n^2+n}}$, the lower bound on the MMPHF problem is $(n \log$ $n)$ $=$ $(n \log\log\log u)$ as $\log n = (\log\log\log u)$ here.

Thus Lemmas 3.1 and 3.2 can be combined to prove Theorem 2 modulo a serious caveat: the lower bound only holds for instances of the problem wherein the universe size is larger than doubly exponential in n, which is admittedly not the most interesting setting of the parameters. In the next subsection, we use a simple graph product argument (plus Proposition 2.2) to extend this lower bound to the whole range of parameters u considered by Theorem 2.

3.2 Extending the MMPHF Lower Bound to Small Universes

For every integers $m; ` > 1$, dene $G(m; `) = (V(m; `); E(m; `))$ as the `-oset conict graph where the vertex set $V(m; `)$ is the set of all size-m subsets of $[` + 1 : M + `]$ for $M := 2^{m^{m^2+m}}$ dened earlier, and the edge set $E(m; `)$ is dened as in normal conict graphs. (Thus $G(m; 0) = G(m)$.)

Furthermore, for every integer $m; k > 1$, we dene the k-fold conict graph, denoted by $G^k(m)$, as the graph:

$$G^k(m) = (V^k(m); E^k(m)) := G(m; 0) \_ G(m; M) \_ G(m; 2M) \_ \_ G(m; (k \quad 1)M);$$

where '_' denotes the graph product in Proposition 2.2. The direct interpretation of the nodes of $V^k(m)$ is a product of tuples from disjoint ranges, but we can also interpret it as a single tuple of length $k \; m$. This way, $G^k(m)$ is a subset of the conict graph on km-size subsets of $[kM]$ and it makes sense to compute $D(v)$ for any $v \; 2 \; V^k(m)$.

Therefore, by Lemma 3.1, we again have a lower bound of $(\log{}_f(G^k(m)))$ for MMPHF on tuples of length $n = km$ from a universe of size $u = kM$.

By Proposition 2.2, combined with Lemma 3.2, we have,

$$\log{}_f(G^k(m)) = \sum_{i=1}^{k} \log{}_f(G(m; i \quad 1)) = k \; \log{}_f(G(m)) > (k \; m \; \log m) = (n \log m);$$

where the second equality is because $_f(G(m; i \quad 1)) = {}_f(G(m))$ for all $i \; 2 \; [k]$, as these graphs are all isomorphic to each other. Consider a choice of

$$m = (\log\log n)^{1=6} \quad \text{and} \quad k = n=(\log\log n)^{1=6};$$

which in turn gives us

$$u = k \; 2^{m^{m^2+m}} \quad k \; 2^{2^{m^3}} = \frac{n}{(\log\log n)^{1=6}} \; 2^{2^{p \log\log n}} \quad n \; 2^{p \log n}:$$

By the above equation, we have a lower bound of $(n \log\log\log u)$ for MMPHF given that in this case, $\log m = (\log\log\log u)$. Thus, so far, we have proven Theorem 2 on both its boundary cases, namely, when $u = n \; 2^{p \log n}$ and when $u = 2^{n^{n^2+n}}$. The proof can now be extended to the full range of the parameters in the middle by re-parameterizing k appropriately; see Appendix B for the complete argument.

We conclude that in order to nish the proof of Theorem 2, we need only establish Lemma 3.2.

4 Fractional Chromatic Number of Conict Graphs

In this section, we establish a lower bound on the fractional chromatic number of the conict graph $G(m)$ for any (large enough) $m > 1$, and thereby prove Lemma 3.2.

Proposition 2.3 gives us a clear path for proving the lower bound on $\chi_f(G(m))$ in Lemma 3.2: we can design a distribution μ on vertices of $V(m)$ and then, for every independent set $I \in \mathcal{I}(G(m))$, we can upper bound the probability that v sampled from μ belongs to I. As χ_f in Proposition 2.3 is maximum over all possible distributions, our distribution provides a lower bound for $\chi_f(G(m))$.

To continue, we need the following interpretation of the (maximal) independent sets in $G(m)$.

Observation 4.1. *Any maximal independent set I in $G(m)$ can be uniquely identified by a function $f_I : [M] \to [m]$ such that for every vertex $v = (v_1, \ldots, v_m) \in V(m)$, we have $f_I(v_i) = i$.*

Proof. Consider any two vertices $v, w \in I$. Since there is no edge between $v = (v_1, \ldots, v_m)$ and $w = (w_1, \ldots, w_m)$ in $G(m)$, whenever $v_i = w_j$, we necessarily have that $i = j$. Thus, any element of $e \in [M]$ can only appear in a single index $i_e \in [m]$ throughout all vertices $v \in I$ (or does not appear at all in v). We can thus define $f_I(e)$ to be i_e, giving us a functin f_I with the desired property.

We now show that f_I uniquely identies I. Define I^0 as set of vertices $v = (v_1, \ldots, v_m) \in V(m)$ satisfying $f_I(v_i) = i$ for all $i \in [m]$. I^0 is an independent set satisfying $I \subseteq I^0$. Since I is assumed to be maximal, it follows that $I = I^0$, meaning that we recover I from f_I. ∎

Observation 4.1 allows us to reduce Lemma 3.2 to the following lemma about m-tuples of increasing integers. Proving Lemma 4.2 is the main technical contribution of our work.

Lemma 4.2. *There is an absolute constant $\delta > 0$ such that for any suciently large $m \geq 1$ and $M = 2^{m^{m^2+m}}$, the following is true. There exists a distribution on m-tuples of increasing numbers $X_1 < \cdots < X_m$ from $[M]$ such that for any function $f : [M] \to [m]$,*

$$\Pr_{(X_1, \ldots, X_m)} (\forall i \in [m] : f(X_i) = i) \leq m^{-\delta m}.$$

Before proving Lemma 4.2, we show how it implies Lemma 3.2.

Proof of Lemma 3.2 (assuming Lemma 4.2). Any choice of $(X_1, \ldots, X_m)$ in Lemma 4.2 can be mapped to a unique vertex $v \in G(m)$ and vice versa. Thus, $(X_1, \ldots, X_m)$ induces a distribution μ on vertices $V(m)$: sample $(X_1, \ldots, X_m)$ and return the vertex $v = (v_1, \ldots, v_m)$ where $v_i = X_i$ for all $i \in [m]$. Moreover, for any maximal independent set $I \in \mathcal{I}(G)$, by Observation 4.1, the vertex corresponding to $(X_1, \ldots, X_m)$ belongs to I iff $f_I(X_i) = i$ for all $i \in [m]$. Thus,

$$\Pr_{v \sim \mu} (v \in I) = \Pr_{(X_1, \ldots, X_m)} (\forall i \in [m] : f(X_i) = i) \leq m^{-\delta m}.$$

As every independent set of $G(m)$ is a subset of some maximal independent set, the upper bound continues to hold for every independent set in $G(m)$.

By Proposition 2.3,

$$\chi_f(G(m)) \geq \left(\min_{I \in \mathcal{I}(G(m))} \Pr_{v \sim \mu} (v \in I) \right)^{-1} \geq m^{\delta m};$$

concluding the proof. ∎

The rest of the section proves Lemma 4.2. We start with a high-level overview in Section 4.1. We then dene the distribution that we will use for the proof of Lemma 4.2 (Section 4) and analyze it to establish Lemma 4.2 (Section 4.3). The probability distribution that we construct in these sections should be viewed intuitively as a "hard" input distribution on inputs to the MMPHF problem (in the spirit of Yao's minimax principle).

4.1 A High-Level Overview of the Proof The proof of Lemma 4.2 is quite dense and requires both a highly delicate probability distribution and several intricate technical arguments. Thus, before getting into the details of this proof, we provide a (very) high-level overview of the logic behind it. In order to convey the intuition, we omit many details from this subsection, instead limiting ourselves to an informal discussion.

The distribution in Lemma 4.2 is roughly as follows: we start with a "window" Win_1 which is the interval $[1 : M]$, and then sample X_1 uniformly at random from Win_1. We then pick window Win_2 to be $[X_1 + 1 : X_1 + w_2]$

for an integer $w_2 > 1$ chosen randomly from a carefully designed distribution. Similarly to before, X_2 will be chosen uniformly from Win_2. We continue like this by picking a new window $Win_i = [X_{i-1}+1 : X_{i-1}+w_i]$ for each $i \in [m]$ by sampling each w_i from a distribution that is constructed based on $(w_1, \ldots, w_{i-1})$, and then sampling X_i from Win_i. Note that, by design, we will satisfy $X_1 < X_2 < \cdots < X_m$.

The key property that this distribution achieves can be explained informally as follows. For any index $i \in [m]$, there is a recursive partitioning of the window Win_i into "dense" and "sparse" intervals, where an interval $I \subseteq Win_i$ is dense (with respect to the function f and the index i) if at least an $(1/m)$ fraction of entries $j \in I$ satisfy $f(j) = i$, and otherwise I is sparse. The central property that our distribution ensures is that, if the random choice of X_i places it in a dense interval, then (with very high probability) the final window Win_m will itself end up being dense (i.e., for at least a $2/m$ fraction of $j \in Win_m$, $f(j) = i$).

Establishing this property is quite challenging and involves defining the distribution of w_i's in a highly non-uniform manner (in terms of their values); this is also the source of the doubly exponential dependence of range M on the number of indices m. We postpone the details on how this property can be achieved to the actual proof and focus on why it is a useful property for us.

The analysis of the distribution now uses the property in a potential-function style argument. For each X_i, it is either sampled from a sparse interval or a dense one. If X_i is sampled from a sparse interval I, then no matter the past iterations, the probability that $f(X_i) = i$ is at most $(2/m)$, since at most $(2/m)$ fraction of I can have value $f(j) = i$ by the definition of it being sparse. On the other hand, if X_i is chosen from a dense interval, then at least a $(2/m)$ fraction of entries of Win_m should be mapped to i by f as well (by our property). Seeing Win_m as a potential function now, we have that this latter step can only happen for $(m/2)$ iterations $i \in [m]$—indeed, each time that this happens for some i, we commit some $(2/m)$ fraction of indices $j \in Win_m$ to having $f(j) = i$, and these sets indices must be disjoint. As a result, we have that at least $(m/2)$ iterations $i \in [m]$ sample X_i from a sparse interval. Thus,

$$\Pr(f(X_1) = 1, \ldots, f(X_m) = m) \leqslant \prod_{\substack{i:\, X_i \text{ chosen from} \\ \text{a sparse interval}}} \Pr\left(f(X_i) = i \mid f(X_1) = 1, \ldots, f(X_{i-1}) = (i-1)\right)$$

$$\leqslant O\left(\frac{1}{m}\right)^{m/2} = m^{-\Omega(m)},$$

as desired for the proof of Lemma 4.2.

The main challenge in formalizing the above argument is the design and analysis of the distribution so that the property discussed above holds. Note also that the property cannot hold deterministically—another challenge is to show that it holds with such high probability that the risk of the property ever failing (across the entire construction) can be ignored.

4.2 The Hard Input Distribution in Lemma 4.2

The distribution is defined as follows.

(i) Let $k = m^m$, $S_0 = k^{m+1}$, and $X_0 = 0$.

(ii) For $i = 1$ to m:

  (a) Sample two random numbers Y_i from $[2^{S_{i-1}}]$ and Z_i from $[k-1]$ uniformly at random.

  (b) Define the random variables of iteration i as:

$$X_i = X_{i-1} + Y_i \quad \text{and} \quad S_i = S_{i-1} - k^{m-i+1} \cdot Z_i.$$

(iii) Return $(X_1, \ldots, X_m)$ as the resulting random variables.

To avoid ambiguity, we use lower case letters (s_i, x_i, y_i, z_i) to denote realizations of random variables (S_i, X_i, Y_i, Z_i) for $i \in [m]$.

We have the following basic observation on the range of numbers created in this distribution.

Observation 4.3. Every choice of $(X_1, \dots, X_m)$ and $(S_1, \dots, S_m)$ satisfy the following properties:

(i) "Monotonicity": for all $i \in [m]$, $X_i \geq X_{i-1}$ and $S_i \leq S_{i-1} - m^m$ (and S_i, X_i are integers).

(ii) "Boundedness": for every $i \in [m]$, $X_m \leq X_i + (m-i) \cdot 2^{S_i}$ and $S_m \geq S_i - k^{m-i+1} > 0$.

Proof. Monotonicity of X_i's holds as Y_i's are positive. Monotonicity for S_i's holds because Z_i's are positive and $k^{m-i+1} \geq k^{m-m+1} \geq k = m^m$, meaning that we always have $S_i \leq S_{i-1} - m^m$.

For part (ii), we have,

$$X_m = X_i + \sum_{j=i+1}^{m} Y_j \leq X_i + \sum_{j=i+1}^{m} 2^{S_{j-1}} \leq X_i + (m-i) \cdot 2^{S_i};$$

which proves the boundedness of X_i's. For S_i's,

(as $\sum_{j=i}^{m-1} k^{-j} \leq \sum_{j=i}^{\infty} k^{-j} = k^{-i+1} \cdot (k-1)^{-1}$)

$$S_m = S_i - \sum_{j=i+1}^{m} k^{m-j+1} \cdot Z_j \geq S_i - k^m \cdot (k-1) \cdot \sum_{j=i}^{m-1} k^{-j} \geq S_i - k^{m-i+1}.$$

Finally, by this bound, we have $S_m \geq S_0 - k^{m+1} > 0$ as $S_0 = k^{m+1}$. ∎

When discussing $(X_1, \dots, X_m)$, we will also need some further definitions:

- For any realization $(s_{<i}, x_{<i})$, we define the **window** of iteration $i \in [m]$, $\mathrm{Win}_i := \mathrm{Win}_i(s_{<i}, x_{<i})$, as the support of the random variable X_i conditioned on $(s_{<i}, x_{<i})$, i.e.,

$$\mathrm{Win}_i := \mathrm{Win}_i(s_{<i}, x_{<i}) = [x_{i-1} + 1 : x_{i-1} + 2^{S_{i-1}}].$$

  Notice that $|\mathrm{Win}_i(s_{<i}, x_{<i})| = 2^{S_{i-1}}$ and Win_i is determined by $(s_{<i}, x_{<i})$.

- Similarly, for any fixed choice of $(s_{<i}, x_{<i})$, consider the following numbers:

$$w_{i,j} := 2^{S_{i-1} - j \cdot k^{(m-i+1)}} \quad \text{for all } j \in \{0, \dots, k\}. \tag{4.3}$$

  This way, $|\mathrm{Win}_{i+1}(s_{<i}, x_{<i})|$ is chosen uniformly at random from $\{w_{i,1}, \dots, w_{i,k-1}\}$ (depending solely on the choice of $Z_i \in [k-1]$ which also determines S_i). Moreover, the ratio of $w_{i,j}$ and $w_{i,j+1}$ is fixed for any $j \in \{0, \dots, k-1\}$ and we define this quantity as

$$r_i := 2^{k^{m-i+1}} = \frac{w_{i,j}}{w_{i,j+1}} \quad \text{for any } j \in \{0, \dots, k-1\}. \tag{4.4}$$

Observation 4.4. For any fixed $(s_{<i}, x_{<i})$, the supports of random variables $|\mathrm{Win}_{i+1}|, \dots, |\mathrm{Win}_m|$ are subsets of the interval $[2^{m^m} \cdot w_{i,Z_i+1} : w_{i,Z_i}]$.

Proof. By definition,

$$|\mathrm{Win}_{i+1}| = 2^{S_i} = 2^{S_{i-1} - k^{m-i+1} Z_i} = w_{i,Z_i}.$$

Moreover, by Observation 4.3, for any $j \in \{i+1, \dots, m\}$, we have $|\mathrm{Win}_j| \leq |\mathrm{Win}_{i+1}|$. Thus each of these windows can have length at most w_{i,Z_i}, proving the upper bound side.

For the lower bound, for any $j \in \{i+1, \dots, m\}$, we have,

$$\text{(by parts (i), (ii) of Observation 4.3)} \qquad |\mathrm{Win}_j| \geq |\mathrm{Win}_m| = 2^{S_{m-1}} \geq 2^{S_i - k^{m-i+1} + m^m}$$
$$= 2^{m^m} \cdot 2^{S_i} \cdot 2^{-k^{m-i+1}} = 2^{m^m} \cdot w_{i,Z_i} \cdot r_i^{-1} = 2^{m^m} \cdot w_{i,Z_i+1}.$$

This concludes the proof. ∎

We need one nal denition for now:

- For the function $f : [M] \to [m]$, we dene the **density** of index $i \in [m]$ in f over a window Win, denoted by $\text{density}_f(\mathrm{Win}; i)$, as

$$\text{density}_f(\mathrm{Win}; i) := \frac{|\{j \in \mathrm{Win} : f(j) = i\}|}{|\mathrm{Win}|};$$

  namely, the fraction of entries of the window that are equal to i.

Observation 4.5. For any choice of $(s_{<i}, x_{<i})$, we have,

$$\Pr(f(X_i) = i \mid s_{<i}, x_{<i}) = \text{density}_f(\mathrm{Win}_i(s_{<i}, x_{<i}); i).$$

Proof. Conditioned on $(s_{<i}, x_{<i})$, X_i is chosen uniformly at random from $\mathrm{Win}_i(s_{<i}, x_{<i})$. The observation therefore follows from the denition of $\text{density}_f(\mathrm{Win}_i(s_{<i}, x_{<i}); i)$. ∎

4.3 Analysis of the Hard Distribution – Proof of Lemma 4.2

We prove Lemma 4.2 by individually considering each iteration in the distribution.

Lemma 4.6. For any iteration $i \in [m-1]$ and conditioned on any choice of $(s_{<i}, x_{<i})$, <u>at least one</u> of the following two conditions is true:

$$\text{(i)}\ \Pr(f(X_i) = i \mid s_{<i}, x_{<i}) \leq \frac{101}{m} \quad \text{or} \quad \text{(ii)}\ \Pr\left(\text{density}_f(\mathrm{Win}_m; i) < \frac{2}{m} \mid s_{<i}, x_{<i}\right) < \frac{.1}{k^{1/3}}$$

The guarantee in Lemma 4.6 does not apply to the last iteration (omitted for technical reasons).

The main bulk of this section is to prove Lemma 4.6. We then show at the end of the section that this lemma easily implies Lemma 4.2. To continue, we need some denitions.

Denition 4.7. The **window-tree** of iteration $i \in [m]$ for $(s_{<i}, x_{<i})$, denoted by $T_i := T(s_{<i}, x_{<i})$, is the following rooted tree with $k+1$ levels (the root is at level 0):

(i) Every non-leaf node of the tree has r_i many child-nodes.

(ii) Every node at a level $\ell \in \{0, \ldots, k\}$ is associated with a window Win() of length $w_{i,\ell}$.

(iii) The root $_r$ is associated with the window $\mathrm{Win}(_r) := \mathrm{Win}_i(s_{<i}, x_{<i})$. The windows associated with child-nodes of a node at level ℓ partition Win() of length $w_{i,\ell}$ into equal-size windows of length $w_{i,\ell+1}$ (recall that has $r_i = w_{i,\ell}/w_{i,\ell+1}$ child-nodes). Moreover, the left most child-node receives the window in the partition with the smallest starting point, the next child-node on the right receives the next window with smallest part, and so on.

(iv) The **density** of a node with respect to any function $f : [M] \to [m]$ is dened as

$$\text{density}_f() := \text{density}_f(\mathrm{Win}(); i).$$

One way we use the window-tree in our analysis is to consider the process of sampling X_i (which is uniform over $\mathrm{Win}_i(s_{<i}, x_{<i})$ at this stage) as traversing the window-tree via a root-to-leaf path. This is formalized in the following observation.

Observation 4.8. The distribution of X_i conditioned on $(s_{<i}, x_{<i})$ can be alternatively seen as: (i) Sample a root-to-leaf path $_0, _1, \ldots, _k$ where $_0$ is the root of T_i and where each $_{\ell+1}$ is a child-node of $_\ell$ chosen uniformly at random; then, (ii) sample X_i uniformly at random from $\mathrm{Win}(_k)$. We refer to $_0, \ldots, _k$ as the **sampling path** of X_i.

Proof. X_i is distributed uniformly over Win_i and leaf-nodes of T_i form an equipartition of Win_i. ∎

In addition, we dene a pruning procedure for any window-tree T as follows.

Denition 4.9. Fix a function $f : [M] \to [m]$ and a window-tree T_i for some $i \in [m]$. We say that a node $\in T_i$ is sparse i

$$\mathsf{density}_f() \leq \frac{100}{m}.$$

Consider the following procedure for pruning T_i: Start from the root down to the leaf-nodes and prune any sparse node, as well as the whole subtree rooted at that node. We refer to a sparse node that was pruned on its own (i.e., any node that is sparse and has no sparse ancestors) as a *directly pruned* node and to other pruned nodes as *indirectly pruned*.

Finally, for $\ell \in \{0, \ldots, k\}$, dene p_ℓ as the fraction of directly pruned nodes at level ℓ of the tree over all level-ℓ nodes that are not indirectly pruned.

It is worth noting that pruning is deterministic conditioned on $(s_{<i}, x_{<i})$.

With these denitions, we can now start proving Lemma 4.6. This will be done by considering some dierent cases handled by the following claims. The rst (and easiest) case is when most nodes of the window-tree are pruned, in which case we achieve property (i) of Lemma 4.6.

Claim 4.10 (Case I: "Many Directly Pruned Nodes"). Suppose

$$\prod_{\ell=0}^{k}(1 - p_\ell) \leq \frac{1}{m}.$$

Then, for any choice of $(s_{<i}, x_{<i})$,

$$\Pr_{X_i}(f(X_i) = i \mid s_{<i}, x_{<i}) \leq \frac{101}{m}.$$

Proof. Let W_{rem} denote the subset of Win_i that remains after removing windows of all pruned leaf-nodes from Win_i. We have that

$$|W_{rem}| = \frac{\#\ \text{leaf-nodes of } T_i \text{ that are not pruned}}{\#\ \text{leaf-nodes of } T_i} \cdot |Win_i| = \prod_{\ell=0}^{k}(1 - p_\ell) \cdot |Win_i| \leq \frac{|Win_i|}{m},$$

where the second equality is because at each level ℓ of the tree, the number of not pruned nodes drops by a factor of $(1 - p_\ell)$ by the denition of p_ℓ.

Let DP denote the set of all nodes in the tree T_i that were directly pruned. Note that the windows $Win()$ for $\in DP$ partition $Win_i \setminus W_{rem}$. This implies that

(by the denition of $\mathsf{density}_f()$ function)

$$\mathsf{density}_f(Win_i, i) = \frac{1}{|Win_i|} \cdot \left(|W_{rem}| \cdot \mathsf{density}_f(W_{rem}, i) + \sum_{\in DP} \mathsf{density}_f() \cdot |Win()| \right)$$

(as $\mathsf{density}_f() \leq 100/m$ by the denition of sparsity, and $\mathsf{density}_f(W_{rem}, i) \leq 1$)

$$\leq \frac{1}{|Win_i|} \cdot \left(|W_{rem}| + \sum_{\in DP} \frac{100}{m}|Win()| \right)$$

(as $|W_{rem}|/|Win_i| \leq 1/m$ as established above, and $\sum_{\in DP} |Win()| \leq |Win_i|$)

$$\leq \frac{1}{m} + \frac{100}{m} = \frac{101}{m}.$$

By Observation 4.5, we have,

$$\Pr_{X_i}(f(X_i) = i \mid s_{<i}, x_{<i}) = \mathsf{density}_f(Win_i, i) \leq \frac{101}{m},$$

concluding the proof. ∎

We now consider the complementary case, while also taking the randomness of Z_i into account. Recall that Z_i is uniform over $[k \quad 1]$ and that $|Win_{i+1}| = w_{i;z_i}$. For any xed realization z_i of Z_i, recall the sampling-path-based process of sampling X_i outlined in Observation 4.8. Consider the rst z_i vertices in this path, namely, $_0; \ldots; _{z_i \; 1}$ that start from the root and end at a level $z_i \quad 1$ node of T_i.

Dene $E(s_{<i}; x_{<i}; z_i; X_i)$ to be the event that none of the nodes in $_0; \ldots; _{z_i \; 1}$ are pruned. Event $E(s_{<i}; x_{<i}; z_i; X_i)$ depends only on the choice of X_i (to traverse the root-to-leaf path), and is conditioned on $s_{<i}; x_{<i}$ (which determine the window-tree T_i) and z_i (which determines the level of the tree that we focus on). To avoid clutter, when it is clear from the context, we refer to this event simply by E_i.

We partition the remaining cases based on whether or not the event E_i happens.

Claim 4.11 (Case II: "A Pruned Node on the Sampling Path"). Fix any choice of z_i and $(s_{<i}; x_{<i})$. In the case that the event E_i does <u>not</u> happen, we have,

$$\Pr_{X_i}(f(X_i) = i \mid s_{<i}; x_{<i}; z_i; \overline{E(s_{<i}; x_{<i}; z_i; X_i)}) 6 \frac{100}{m}:$$

Proof. After conditioning on $(s_{<i}; x_{<i}; z_i)$, the event E_i is only a function of the sampling process of X_i outlined in Observation 4.8. Assuming E_i does not happen, we know that there exists a unique node $_j$ on the path $_0; \ldots; _{z_i \; 1}$ such that $_j$ is sparse and is directly pruned. By additionally conditioning on the subpath $_0; \ldots; _j$, we have that X_i is chosen uniformly at random from $Win(_j)$ at this point. Thus,

$$\Pr_{X_i}(f(X_i) = i \mid s_{<i}; x_{<i}; z_i; \overline{E_i})$$

(as these subpaths partition all possible choices for E_i to not happen)

$$= \sum_{j=0}^{z_i \; 1} \sum_{\substack{(_0;\ldots;_j):\\ _j \text{ is directly pruned}}} \Pr_{X_i}(f(X_i) = i \wedge (_0; \ldots; _j) \text{ is on the sampling path} \mid s_{<i}; x_{<i}; z_i; \overline{E_i})$$

(as X_i is chosen uniformly from $Win(_j)$ under these conditions)

$$= \sum_{j=0}^{z_i \; 1} \sum_{\substack{(_0;\ldots;_j):\\ _j \text{ is directly pruned}}} \Pr_{X_i}((_0; \ldots; _j) \text{ is on the sampling path} \mid s_{<i}; x_{<i}; z_i; \overline{E_i}) \cdot \frac{|\{t \in Win(_j) : f(t) = i\}|}{|Win(_j)|}$$

(by the denition of density_f)

$$= \sum_{j=0}^{z_i \; 1} \sum_{\substack{(_0;\ldots;_j):\\ _j \text{ is directly pruned}}} \Pr_{X_i}((_0; \ldots; _j) \text{ is on the sampling path} \mid s_{<i}; x_{<i}; z_i; \overline{E_i}) \cdot \text{density}_f(Win(_j); i)$$

(as $_j$ needs to be sparse to be directly pruned)

$$6 \sum_{j=0}^{z_i \; 1} \sum_{\substack{(_0;\ldots;_j):\\ _j \text{ is directly pruned}}} \Pr_{X_i}((_0; \ldots; _j) \text{ is on the sampling path} \mid s_{<i}; x_{<i}; z_i; \overline{E_i}) \cdot \frac{100}{m}:$$

This can now be further upper bounded by $100/m$ as the probability terms are summing over all disjoint events that can lead to $\overline{E_i}$ (conditioned on this event) and thus add up to one. ∎

Finally, we have the following case which handles the situation when E_i happens. The following claim is the heart of the proof.

Claim 4.12 (Case III: "No Pruned Nodes on the Sampling Path"). Fix any choice of z_i and $(s_{<i}; x_{<i})$. In the case that the event E_i happens, we have,

$$\Pr_{X_i}\left(\text{density}_f(Win_m; i) < \frac{2}{m} \mid s_{<i}; x_{<i}; z_i; E_i\right) < 4 \left(p_{z_i} + p_{z_i+1} + \frac{m}{r_i}\right):$$

Proof. Throughout this proof, we always condition on $s_{<i}; x_{<i}; z_i;$ and $E_i = E(s_{<i}; x_{<i}; z_i; X_i)$ and so may not mention this explicitly in the probability terms. This is the information we have so far:

- None of the nodes $_0; \ldots ; _{z_i - 1}$ on the sampling path are pruned as we conditioned on the event E_i (although $_{z_i - 1}$ is still a random variable and is not xed yet just by these conditions).
- Window Win_m is going to have size at least $2^{m^m} \, w_{i;z_i+1}$ and at most $w_{i;z_i}$ by Observation 4.4.
- By Observation 4.3,

  (by the denition of $w_{i;z_i} = 2^{S_i}$) $\quad X_m \leqslant X_i + (m - i) \cdot 2^{S_i} = X_i + (m - i) \cdot w_{i;z_i}:$

- Win_m starts at X_{m-1} and ends at $X_{m-1} + |Win_m|$. We can think of the process of sampling Win_m as rst sampling $|Win_m|$, then sampling the oset $O_{i;m} := X_{m-1} - X_i = \sum_{j=i+1}^{m-1} Y_j$ conditioned on $|Win_m|$, and then sampling X_i conditioned on $O_{i;m}$, and $|Win_m|$.
- We further have that X_i conditioned on $O_{i;m}$ and $|Win_m|$ is still uniform over $Win(_{z_i - 1})$. This is because $|Win_m|$ is only a function of $Z_{i+1}; \ldots ; Z_m$, and $X_{m-1} - X_i$ is only a function of $Y_{i+1}; \ldots ; Y_{m-1}$, while X_i is only a function of Y_i; nally, Y_i is independent of $Y_{i+1}; \ldots ; Y_m$ and $Z_{i+1}; \ldots ; Z_m$ and is chosen uniformly from $[2^{S_i} - 1]$ (recall that $i < m$ in this lemma).

In the following, we condition on any xed choice of oset $o_{i;m}$ for $O_{i;m}$ and on $|Win_m|$. We have already established that

$$2^{m^m} \, w_{i;z_i+1} \leqslant |Win_m| \leqslant w_{i;z_i} \quad \text{and} \quad o_{i;m} \leqslant (m - i) \cdot w_{i;z_i}: \tag{4.5}$$

Moreover, the distribution of Win_m conditioned on $o_{i;m}; |Win_m|$ (and $s_{<i}; x_{<i}; z_i; E_i$ that we always condition on in this proof), is $X_i + o_{i;m}$ for X_i chosen randomly from $Win(_{z_i - 1})$. Also, given that $o_{i;m} \leqslant (m - i) w_{i;z_i}$ while $|Win(_{z_i - 1})| = w_{i;z_i - 1} = r_i \cdot w_{i;z_i}$ and $r_i = 2^{k^{m-i+1}} > 2^k$ as $i \leqslant m$, the distribution of X_i and $X_i + o_{i;m}$ are quite close to each other modulo a negligible factor. Thus, for intuition, we can think of X_i itself as the distribution of starting point for Win_m in this context (although we will of course take this dierence into account explicitly in the proof). We now use this information to prove the claim. To simplify the exposition, we are going to separate the analysis based on level z_i and level z_{i+1} of the window-tree.

Analysis on level z_i of the window-tree. Firstly, since $|Win_m| \leqslant w_{i;z_i}$, and each node at level z_i of the window-tree T_i has a window of length $w_{i;z_i}$, we get that Win_m intersects with windows of at most two consecutive nodes at level z_i of T_i, which are solely determined by the choice of X_i. We use $_1(X_i)$ and $_2(X_i)$ to denote these two nodes with $_1$ being the one where the starting point of Win_m, namely, $X_i + o_{i;m}$, lies in, and $_2(X_i)$ being the one containing the endpoint $X_i + o_{i;m} + |Win_m|$ (note that it is possible that $_2 = _1$).

We prove that with high probability, neither of these nodes are pruned. Let us focus on $_1(X_i)$ rst (the analysis is almost identical for $_2(X_i)$ and we can then apply a union bound). For any $\ell \in \{0; \ldots ; k - 1\}$, let $P(\ell)$ (resp. $DP(\ell)$) denote the set of pruned (resp. directly pruned) nodes at level ℓ of T_i; similarly, for a node $\in T_i$, let $P()$ (resp. $DP()$) denote the set of child-nodes of that are pruned (resp. directly pruned). For any xed choice of $_{z_i - 1}$ on the sampling path of X_i,

(as $_1$ is in level z_i and $P(z_i)$ is the set of all pruned nodes of this level)

$$\Pr_{X_i}(_1(X_i) \text{ is pruned} \mid _{z_i - 1}) = \sum_{\in P(z_i)} \Pr_{X_i}(_1(X_i) = \mid _{z_i - 1})$$

(by partitioning the nodes in level z_i between child-nodes of $_{z_i - 1}$ and remaining ones)

$$= \sum_{\in P(_{z_i - 1})} \Pr_{X_i}(_1(X_i) = \mid _{z_i - 1}) + \sum_{\in P(z_i) \setminus P(_{z_i - 1})} \Pr_{X_i}(_1(X_i) = \mid _{z_i - 1})$$

$$\leqslant |P(_{z_i - 1})| \cdot \frac{1}{r_i} + (m - i) \cdot \frac{1}{r_i}; \tag{4.6}$$

where the last inequality holds because of the following reasoning. Firstly, the probability that $_1(X_i)$ is equal to any xed node at level z_i is at most $1=r_i$. This is because

$$\Pr\ _1(X_i) = \ j\ _{z_i}\ _1 = \Pr(X_i + o_{i;m}\ 2\ \mathrm{Win}()\ j\ _{z_i}\ 1)\ 6\ \frac{jWin()j}{jWin(\ _{z_i}\ _1)j} = \frac{1}{r_i};$$

because X_i is chosen uniformly from $\mathrm{Win}(_{z_i}\ 1)$, and $j\mathrm{Win}()j = j\mathrm{Win}(_{z_i}\ 1)j = r_i$ as is at level z_i. This immediately implies the rst term in the RHS of Eq (4.6). For the second term, for $_1(X)$ to intersect with a node not in the subtree of $_{z_i}\ 1$, we need to have $X_i + o_{i;m}\ 2\ \mathrm{Win}(_{z_i}\ 1)$, while we know $X_i\ 2\ \mathrm{Win}(_{z_i}\ 1)$. As $o_{i;m}\ 6\ (m\ i)\ w_{i;z_i}$ by Eq (4.5), and any node at level z_i has a window of length $w_{i;z_i}$, we get that there are most $(m\ i)$ choices of outside child-nodes of $_{z_i}\ 1$ that can also become $_1(X_i)$. The second part of RHS in Eq (4.6) now follows from this and the upper bound of $1=r_i$ on the probability of each node.

Finally, by taking the expectation over the choice of $_{z_i}\ 1$,

(by the law of total probability, over the choice of $_{z_i}\ 1$ in the sampling path)

$$\Pr_{X_i}(_1(X_i) \text{ is pruned}) = \mathop{E}_{z_i\ 1}\ \Pr_{X_i}(_1(X_i) \text{ is pruned}\ j\ _{z_i}\ 1)$$

(by Eq (4.6))

$$6\ \mathop{E}_{z_i\ 1}\ \frac{jP(_{z_i}\ 1)j}{r_i} + \frac{(m\ i)}{r_i}$$

$$= p_{z_i} + \frac{(m\ i)}{r_i};$$

where in the nal equality, we used the fact that $_{z_i}\ 1$ is chosen from non-pruned nodes (by conditioning on E_i), and thus $jP(_{z_i}\ 1)j = r_i$ is the fraction of pruned nodes over all not indirectly pruned nodes at level z_i, which by denition is p_{z_i}.

Doing the same exact analysis, we can bound the probability that $_2(X_i)$ is pruned also as

$$\Pr_X(_2(X_i) \text{ is pruned})\ 6\ p_{z_i} + \frac{(m\ i) + 1}{r_i};$$

where the $+1$ term in the RHS compared to the one for $_1$ comes from the fact that $_2(X_i)$ can have $(m\ i + 1)$ choices outside subtree of $_{z_i}\ 1$ (because we are now considering $X_i + o_{i;m} + j\mathrm{Win}_m j\ 6\ X_i + (m\ i + 1)\ w_{i;z_i}$ instead). By the union bound on the probabilities for $_1(X_i)$ and $_2(X_i)$,

(4.7) $$\Pr_X(\text{either of } _1(X_i) \text{ or } _2(X_i) \text{ is pruned})\ 6\ 2\ p_{z_i} + 2\ \frac{m\ i}{r_i}:$$

Analysis on level $z_i + 1$ of the window-tree. For the next step, let $_1(X_i); : : : ; _t(X_i)$ denote the child-nodes of $_1(X_i)$ and $_2(X_i)$ such that $\mathrm{Win}(_j(X_i))$ is entirely contained in Win_m. Again, the choice of $_1; : : : ; _t$ is only a function of X_i. Moreover, since $j\mathrm{Win}_m j > 2^m\ w_{i;z_i+1}$ by Eq (4.5), while the window of each node at level $z_i + 1$ is of size $w_{i;z_i+1}$, we have that $t > 2^m\ 2$ always. We now bound the probability that each $_j$ is (directly) pruned, for $j\ 2\ [t]$. This part of the analysis is quite similar to that of level z_i with only minor changes.

For any choice of $_1(X_i)$ and $_2(X_i)$,

(because $\mathrm{Win}_m\ \mathrm{Win}(_1)\ [\ \mathrm{Win}(_2)$ and thus $_j$ has no choice outside child-nodes of $_1$ or $_2$)

$$\Pr_{X_i}(_j(X_i) \text{ is directly pruned}\ j\ _1;_2) = \sum_{2\ DP(_1)[DP(_2)} \Pr_{X_i}(_j(X_i) = \ j\ _1;_2)$$

$$6\ \left(jDP(_1)j + jDP(_2)j\right)\ \frac{1}{r_i}; \qquad (4.8)$$

where we are again going to argue that the probability that $_j(X_i)$ is equal to any xed node is at most $1=r_i$ conditioned on the choice of $_1$ and $_2$. For $_j(X_i)$ to be equal to a node we need to have that $X_i + o_{i;m} + (j\ 1)\ w_{i;z_i+1}\ 2\ \mathrm{Win}()$; this is because $_j(X_i)$ appears after $(j\ 1)$ nodes of level $z_i + 1$ that are fully inside Win_m and each such window has length $w_{i;z_i+1}$ (note that this is a necessary but not a sucient condition). Thus,

$$\Pr_X(_j(X_i) = \quad j\ {}_1;{}_2) 6 \Pr_{X_{i;z_i}}(X_i + o_{i;m} + (j \quad 1)\ w_{i;z_i+1}\ 2\ \mathrm{Win}()\ j\ {}_1;{}_2) 6 \ \frac{j\mathrm{Win}()j}{} = \frac{1}{w_i}; \quad \frac{}{r}$$

where the last inequality is because conditioned on Win_m intersecting with ${}_1;{}_2$, X_i is chosen uniformly at random from a window of length $w_{i;z_i}$ (equal to length of $\mathrm{Win}(_1)$ and $\mathrm{Win}(_2)$); the nal equality also uses that $j\mathrm{Win}()j = w_{i;z_i+1} = w_{i;z_i} = r_i$. Hence, we get Eq (4.8).

We can now deduce that

$$\mathop{E}_X[\#\text{ of }{}_1(X_i);:::;{}_t(X_i)\text{ that are directly pruned}]$$

(by the law of total probability over the choices of ${}_1;{}_2$)

$$= \mathop{E}_{1;2}\ \mathop{E}_{X_i}[\#\text{ of }{}_1(X_i);:::;{}_t(X_i)\text{ that are directly pruned } j\ {}_1;{}_2]$$

(4.9)
$$= \mathop{E}_{1;2}[jDP(_1)j + jDP(_2)j] \quad r_i; \quad \frac{t}{}$$

where the last inequality is by Eq (4.8).

Let $\overline{P}(z_i)$ denote the set of not pruned nodes in level z_i and let $\hat{P}(z_i)$ denote the set of nodes in level z_i whose parents are not pruned. Since we are conditioning on E_i, we know that X_i is uniformly random from the interval $[2_{\hat{P}(z_i)}\mathrm{Win}()$. It follows that $X_{m\quad 1} = X_i + o_{i;m}$ is uniformly random in a range whose size is also $` = \sum_{2\hat{P}(z_i)} j\mathrm{Win}()j$. Thus, for any level-$z_i$ node , we have that

$$\Pr[_1 = \ ] = \Pr[X_{m\quad 1}\ 2\ \mathrm{Win}()]\ 6\ \frac{j\mathrm{Win}()j}{`} = \frac{w_{i;z_i}}{`} = \frac{1}{jP^\wedge(z_i)j};$$

Summing over the level-$(z_i + 1)$ nodes that are directly pruned, we have that

$$E\ jDP(_1)j = \sum_{i\quad i} \Pr[_1\text{ is the parent of }]\ 6 \sum_D \frac{j\ P(z_i + 1)j}{j\hat{P}(z_i^D)j} 6 \frac{j\ P(z_i+1)j}{j\hat{P}(z_i)j}; \frac{2DP(z_i+1)}{}$$

using the upper bound established above on the probability that $_1$ is any xed node. Note that

$$p_{z_i+1} = \frac{jDP(z_i + 1)j}{r_i\ j\overline{P}(z_i)j};$$

i.e., the number of directly pruned nodes in level $z_i + 1$ divided by the number of nodes with not pruned parents. Therefore, $E\ jDP(_1)j\ 6\ r_i\ p_{z_i+1}$. By the same reasoning (but applied to $_2$, which contains the endpoint of X_m), we have that $E\ jDP(_2)j\ 6\ r_i\ p_{z_i+1}$.

Thus, we can use Eq (4.9) to conclude that

$$\mathop{E}_X[\#\text{ of }{}_1(X_i);:::;{}_t(X_i)\text{ that are directly pruned}]\ 6\ 2p_{z_i+1}\ t:$$

By Markov's inequality,

(4.10)
$$\Pr_X(\text{more than }t{=}2\text{ of }{}_1(X_i);:::;{}_t(X_i)\text{ are directly pruned})\ 6\ 4\ p_{z_i+1}:$$

Finally, by considering the possibility that at least one of $_1$ or $_2$ could be pruned also we have,

$$\Pr_X(\text{more than }t{=}2\text{ of }{}_1;:::;{}_t\text{ are pruned})$$
$$6\ \Pr(\text{more than }t{=}2\text{ of }{}_1(X_i);:::;{}_t(X_i)\text{ are directly pruned})$$
$$+ \Pr(\text{either of }{}_1\text{ or }{}_2\text{ are pruned})\ 6$$

(4.11)
$$4p_{z_i+1} + 2p_{z_i} + \frac{2m}{r_i};$$

by Eq (4.7) and Eq (4.10).

Concluding the proof. Let us now condition on the event that at least $t/2$ of nodes ${}_1, \ldots, {}_t$ are not pruned, namely, the complement of the event in Eq (4.11). Given that Win_m can have non-empty intersection with at most two other level-(z_i+1) nodes beside ${}_1, \ldots, {}_t$, and that non-pruned nodes are all dense, conditioned on the above event, we have,

$$\mathrm{density}_f(\mathrm{Win}_m; i) \geq \frac{(t/2) \cdot 100/m}{t+2} \geq \frac{100}{3m} \geq \frac{2}{m};$$

as $t \geq 2^{m^m} - 2 \geq 1$. Thus, by Eq (4.11), we have,

$$\Pr_X\left(\mathrm{density}_f(\mathrm{Win}_m; i) \leq \frac{2}{m}\right) \leq 2p_{z_i} + 4 \cdot \left(p_{z_i+1} + \frac{2m}{r_i}\right) < 4\left(p_{z_i} + p_{z_i+1} + \frac{m}{r_i}\right);$$

concluding the proof. ∎

Claims 4.10 to 4.12 now cover all possible cases and allow us to prove Lemma 4.6.

Proof of Lemma 4.6. Fix the tree T_i and consider its pruning process. If $\prod_{\ell=0}^{k}(1-p_\ell) \leq 1/m$, we achieve the first condition of the lemma by Claim 4.10 and are thus done. Now consider the complement case. In this case, we have,

$$\frac{1}{m} < \prod_{\ell=0}^{k}(1-p_\ell) \leq \exp\left(-\sum_{\ell=0}^{k} p_\ell\right);$$

which implies that $\sum_{\ell=0}^{k} p_\ell \leq \ln m$. Recall that the choice of Z_i in the distribution is uniform over $[k-1]$ regardless of conditioning on $(s_{<i}, x_{<i})$. Since Z_i is chosen uniformly from $[k-1]$, we have,

$$\mathbb{E}_{Z_i}\left[p_{Z_i} + p_{Z_i+1}\right] \leq \frac{1}{k-1}\sum_{\ell=1}^{k-1} p_\ell + \frac{1}{k-1}\sum_{\ell=2}^{k} p_\ell \leq \frac{2}{k-1}\sum_{\ell=0}^{k} p_\ell \leq \frac{2\ln m}{(k-1)}.$$

By Markov bound, we have,

$$\Pr_{Z_i}\left(p_{Z_i} + p_{Z_i+1} \geq \frac{4\ln m}{k^{1/2}}\right) \leq \frac{1}{k^{1/3}}.$$

We can now condition on any choice z_i of Z_i such that $p_{z_i} + p_{z_i+1} \leq (4\ln m)/k^{1/2}$. At this point, either event E_i does not happen, in which case, by Claim 4.11, we again obtain condition (i) of the lemma; or the event E_i happens, which by Claim 4.12 and the choice of r_i in Eq (4.4) implies

$$\Pr_{X_i}\left(\mathrm{density}_f(\mathrm{Win}_m; i) \leq \frac{2}{m} \mid s_{<i}, x_{<i}\right) \leq 4\left(\frac{4\ln m}{k^{1/2}} + \frac{m}{2^{k^{m-i}}}\right) \leq \frac{1}{k^{1/3}};$$

as $i \leq m-1$ and thus $m/2^{k^{m-i}} \leq m/2^k \leq 1/k^{1/3}$, as $k = m^m$. Taking the union bound over the above two events, we also obtain condition (ii) of the lemma. ∎

Finally, we use this lemma to conclude the proof of Lemma 4.2.

Proof of Lemma 4.2. Let $T_1, T_2 \subseteq [m]$ denote, respectively, the iterations in which condition (i) or condition (ii) of Lemma 4.6 happens. Note that T_1 and T_2 are random variables over the randomness of S_i's and X_i's. We first claim that with high probability $|T_2| < m/2$. This is because for any iteration $i \in T_2$ and any choice of $(s_{<i}, x_{<i})$ of prior iterations, by Lemma 4.6,

$$\Pr_{X_i}\left(\mathrm{density}_f(\mathrm{Win}_m; i) \leq \frac{2}{m} \mid s_{<i}, x_{<i}\right) \leq \frac{1}{k^{1/3}}.$$

A union bound on at most m choices for indices on T_2 then implies that with probability at least $1 - m/k^{1/3}$, we have $\mathrm{density}_f(\mathrm{Win}_m; i) > \frac{2}{m}$ for all $i \in T_2$. But then conditioned on this event, the size of T_2 cannot be $m/2$ or larger as otherwise Win_m contains $m/2$ disjoint sets each of which contains than a $2/m$ fraction of the window, which is a contradiction. Thus,

$$\text{(as } k = m^{m}\text{)} \qquad \Pr(|T_2| > m/2) \leqslant \frac{m}{k^{1/3}} \leqslant \frac{1}{k^{1/4}}.$$

We condition on the complement of this event in the following, namely, that $|T_2| < m/2$. Let $i_1, \ldots, i_{m/2}$ denote the rst $m/2$ indices of T_1 which by the conditioning on the size of T_2 is well dened. We have,

$$\Pr(\text{for all } j \in [m/2]\colon f(X_{i_j}) = i_j) = \prod_{1 \leqslant j \in [m/2]} \Pr\left(f(X_{i_j}) = i_j \mid f(X_{i_1}) = i_1, \ldots, f(X_{i_{j-1}}) = i_j\right)$$

(since these are type (i) iterations and we can apply condition (i) of Lemma 4.6)

$$\leqslant \left(\frac{101}{m}\right)^{m/2}.$$

Putting these two together, combined with the value of $k = m^{m}$, implies that,

$$\Pr_{(X_1,\ldots,X_m)}(\forall i \in [m] : f(X_i) = i) \leqslant \frac{1}{k^{1/4}} + \left(\frac{101}{m}\right)^{m/2} \leqslant m^{-m};$$

for some constant > 0 (taking $= 1/100$ certainly suces). This concludes the proof. ∎

Acknowledgements We are thankful to the anonymous reviewers of SODA 2023 for many helpful comments on the presentation of the paper.

William Kuszmaul was partially sponsored by the United States Air Force Research Laboratory and the United States Air Force Articial Intelligence Accelerator under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the ocial policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Additionally, Sepehr Assadi was supported in part by a NSF CAREER grant CCF-2047061, a Google Research gift, and a Fulcrum award from Rutgers Research Council. Martin Farach-Colton was supported by NSF grants CCF-2118620 and CCF-2106999. William Kuszmaul was supported by an NSF GRFP fellowship and a Hertz fellowship.

References

[BBPV09] Djamal Belazzougui, Paolo Boldi, Rasmus Pagh, and Sebastiano Vigna. Monotone minimal perfect hashing: searching a sorted table with O(1) accesses. In SODA, pages 785{794. SIAM, 2009. 1, 2

[BBPV11] Djamal Belazzougui, Paolo Boldi, Rasmus Pagh, and Sebastiano Vigna. Theory and practice of monotone minimal perfect hashing. Journal of Experimental Algorithmics (JEA), 16:3{1, 2011. 1, 2

[BCKM20] Djamal Belazzougui, Fabio Cunial, Juha Kärkkäinen, and Veli Mäkinen. Linear-time string indexing and analysis in small space. ACM Transactions on Algorithms (TALG), 16(2):1{54, 2020. 2

[BCO11] Alexandra Boldyreva, Nathan Chenette, and Adam O'Neill. Order-preserving encryption revisited: Improved security analysis and alternative solutions. In Annual Cryptology Conference, pages 578{595. Springer, 2011. 1

[Bel14] Djamal Belazzougui. Linear time construction of compressed text indices in compact space. In Proceedings of the forty-sixth Annual ACM Symposium on Theory of Computing, pages 148{193, 2014. 2

[BGMP16] Djamal Belazzougui, Travis Gagie, Veli Mäkinen, and Marco Previtali. Fully dynamic de bruijn graphs. In International symposium on string processing and information retrieval, pages 145{152. Springer, 2016. 2

[BN14] Djamal Belazzougui and Gonzalo Navarro. Alphabet-independent compressed text indexing. ACM Transactions on Algorithms (TALG), 10(4):1{19, 2014. 2

[BN15] Djamal Belazzougui and Gonzalo Navarro. Optimal lower and upper bounds for representing sequences. ACM Transactions on Algorithms (TALG), 11(4):1{21, 2015. 2

[Bol15] Paolo Boldi. Minimal and monotone minimal perfect hash functions. In Giuseppe F. Italiano, Giovanni Pighizzini, and Donald Sannella, editors, Mathematical Foundations of Computer Science 2015 - 40th International Symposium, MFCS 2015, Milan, Italy, August 24-28, 2015, Proceedings, Part I, volume 9234 of Lecture Notes in Computer Science, pages 3{17. Springer, 2015. 2

[BV08] P Boldi and S Vigna. Sux4j 1.0. 2008. 1

[CFP+15] Raphael Cliord, Allyx Fontaine, Ely Porat, Benjamin Sach, and Tatiana Starikovskaya. Dictionary matching in a stream. In Algorithms-ESA 2015, pages 361{372. Springer, 2015. 2

[D+18] Martin Dietzfelbinger et al. 4.3 space complexity of monotone minimal perfect hashing. Dagstuhl Reports, Vol. 7, Issue 5 ISSN 2192-5283, page 19, 2018. 2

[DLR95] Dwight Duus, Hannon Lefmann, and Vojtech Rodl. Shift graphs and lower bounds on ramsey numbers rk (l; r). Discrete Mathematics, 137(1-3):177{187, 1995. 2

[EH66] Paul Erdos and Andras Hajnal. On chromatic number of graphs and set-systems. Acta Math. Acad. Sci. Hungar, 17(61-99):1, 1966. 2, 3

[FHRT92] Zoltan Furedi, Peter Hajnal, Vojtech Rodl, and William T Trotter. Interval orders and shift graphs. 1992. 2

[GNP20] Travis Gagie, Gonzalo Navarro, and Nicola Prezza. Fully functional sux trees and optimal text searching in bwt-runs bounded space. Journal of the ACM (JACM), 67(1):1{54, 2020. 2

[GOR10] Roberto Grossi, Alessio Orlandi, and Rajeev Raman. Optimal trade-os for succinct string indexes. In International Colloquium on Automata, Languages, and Programming, pages 678{689. Springer, 2010. 2

[HT01] Torben Hagerup and Torsten Tholey. Ecient minimal perfect hashing in nearly minimal space. In Afonso Ferreira and Horst Reichel, editors, STACS 2001, 18th Annual Symposium on Theoretical Aspects of Computer Science, Dresden, Germany, February 15-17, 2001, Proceedings, volume 2010 of Lecture Notes in Computer Science, pages 317{326. Springer, 2001. 4

[LFAK11] Hyeontaek Lim, Bin Fan, David G Andersen, and Michael Kaminsky. Silt: A memory-ecient, high-performance key-value store. In Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles, pages 1{13, 2011. 1

[Meh82] Kurt Mehlhorn. On the program size of perfect and universal hash functions. In 23rd Annual Symposium on Foundations of Computer Science (sfcs 1982), pages 170{175. IEEE, 1982. 2

[Nav14] Gonzalo Navarro. Spaces, trees, and colors: The algorithmic landscape of document retrieval on sequences. ACM Computing Surveys (CSUR), 46(4):1{47, 2014. 1

[New91] Ilan Newman. Private vs. common random bits in communication complexity. Inf. Process. Lett., 39(2):67{71, 1991. 3

[ST11] Gabor Simonyi and Gabor Tardos. On directed local chromatic number, shift graphs, and borsuk-like graphs. Journal of Graph Theory, 66(1):65{82, 2011. 2, 3

[SU11] Edward R Scheinerman and Daniel H Ullman. Fractional graph theory: a rational approach to the theory of graphs. Courier Corporation, 2011. 4, 19

Appendix

A Proofs of Standard Results in Fractional Coloring

We prove Propositions 2.2 and 2.3 here for completeness. These proofs are standard; see, e.g. [SU11]. We start by presenting the dual view of fractional colorings that is the key to these proofs.

The dual view of fractional colorings. Given that $\chi_f(G)$ is dened as a solution to an LP, we can use duality to also express $\chi_f(G)$ via the following LP:

(A.1) $$\chi_f(G) := \max_{y \in [0,1]^{V(G)}} \sum_{v \in G} y_v \quad \text{subject to} \quad \sum_{v \in I} y_v \leq 1 \quad \forall I \in \mathcal{I}(G).$$

This LP is a fractional relaxation of the clique number of G, namely, the size of the largest clique in G (since, in any integral solution to this LP, the y-values that are 1 must be on the vertices of a clique). Interestingly, although the chromatic number and clique size are not duals, their relaxations are.

Proposition (Restatement of Proposition 2.2). Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be arbitrary graphs. Dene $G_1 \vee G_2$ as a graph on vertices $V_1 \times V_2$ and dene an edge between vertices (v_1, v_2) and (w_1, w_2) whenever (v_1, w_1) is an edge in G_1 <u>or</u> (v_2, w_2) is an edge in G_2. Then, $\chi_f(G_1 \vee G_2) = \chi_f(G_1) \cdot \chi_f(G_2)$.

Proof of Proposition 2.2. We rst prove that

(A.2) $$\chi_f(G_1 \vee G_2) \geq \chi_f(G_1) \cdot \chi_f(G_2).$$

Let $y^1 \in \mathbb{R}^{V_1}$ and $y^2 \in \mathbb{R}^{V_2}$ be optimal solutions to the dual LP given by Eq (A.1) for G_1 and G_2, respectively. Consider the assignment $y \in \mathbb{R}^{V_1 \times V_2}$ where $y_{u_1,u_2} = y^1_{u_1} \cdot y^2_{u_2}$. We clearly have that

$$\sum_{(u_1,u_2) \in V_1 \times V_2} y_{u_1,u_2} = \left(\sum_{u_1 \in V_1} y^1_{u_1}\right) \cdot \left(\sum_{u_2 \in V_2} y^2_{u_2}\right) = \chi_f(G_1) \cdot \chi_f(G_2).$$

We now argue that y is also a valid solution to the dual LP given by Eq (A.1) for $G_1 \vee G_2$. Fix any independent set $I \in \mathcal{I}(G_1 \vee G_2)$. By the denition of the product, we know that I can be written as a product set, namely, $I = I_1 \times I_2$ for $I_1 \in \mathcal{I}(G_1)$ and $I_2 \in \mathcal{I}(G_2)$. Thus,

$$\sum_{(u_1,u_2) \in I} y_{u_1,u_2} = \left(\sum_{u_1 \in I_1} y^1_{u_1}\right) \cdot \left(\sum_{u_2 \in I_2} y^2_{u_2}\right) \leq 1 \cdot 1 = 1,$$

where the inequality is by the constraint of dual LP for y^1 and y^2 each. Thus, y is a solution to the dual LP for $G_1 \vee G_2$, proving Eq (A.2).

We now prove that

(A.3) $$\chi_f(G_1 \vee G_2) \leq \chi_f(G_1) \cdot \chi_f(G_2),$$

using the primal LP instead. Let $x^1 \in \mathbb{R}^{\mathcal{I}(G_1)}$ and $x^2 \in \mathbb{R}^{\mathcal{I}(G_2)}$ be optimal solutions to primal LP from Eq (2.1) for G_1 and G_2, respectively. Consider the assignment $x \in \mathbb{R}^{\mathcal{I}(G_1 \vee G_2)}$ where $x_I = x^1_{I_1} \cdot x^2_{I_2}$, using the fact from the previous part that $I = I_1 \times I_2$ for $I_1 \in \mathcal{I}(G_1)$ and $I_2 \in \mathcal{I}(G_2)$.

We again clearly have that

$$\sum_{(u_1,u_2) \in \mathcal{I}(G_1 \vee G_2)} x_I = \left(\sum_{I_1 \in \mathcal{I}(G_1)} x^1_{I_1}\right) \cdot \left(\sum_{I_2 \in \mathcal{I}(G_2)} x^2_{I_2}\right) = \chi_f(G_1) \cdot \chi_f(G_2);$$

so it remains to prove that x is a valid solution to the primal LP from Eq (2.1) for $G_1 \vee G_2$. Fix any vertex $(u_1, u_2) \in V_1 \times V_2$ and consider all independent sets $I_1 \subseteq V_1$ that contain u_1 and $I_2 \subseteq V_2$ that contain u_2. Then, $I_1 \times I_2$ is also an independent set in $G_1 \vee G_2$ that contains (u_1, u_2). Thus,

$$\sum_{I \in \mathcal{I}(G_1 \vee G_2) : (u_1,u_2) \in I} x_I \geq \left(\sum_{I_1 \in \mathcal{I}(G_1, u_1)} x^1_{I_1} \right) \cdot \left(\sum_{I_2 \in \mathcal{I}(G_2, u_2)} x^2_{I_2} \right) \geq 1 \cdot 1 = 1,$$

where the inequality is by the constraint of primal LP from Eq (2.1) for x^1 and x^2 each. Thus, x is a solution to the primal LP from Eq (2.1) for $G_1 \vee G_2$, proving Eq (A.3). ∎

Proposition (Restatement of Proposition 2.3). For any graph $G = (V, E)$,

$$\chi_f(G) = \max_{\text{distribution } \mu \text{ on } V} \left(\min_{I \in \mathcal{I}(G)} \Pr_{v \sim \mu}(v \in I) \right)^{-1}.$$

Proof of Proposition 2.3. Let μ be any distribution on $V(G)$ and define $b := \max_{I \in \mathcal{I}(G)} \Pr(v \in I)^{-1}$. Create $y \in \mathbb{R}^{V(m)}$ such that $y_v = b \cdot \mu(v)$ for every vertex $v \in V(m)$ where $\mu(v)$ is the probability of vertex v under the distribution μ. We claim that y is a feasible dual solution in Eq (A.1).

For every independent set $I \in \mathcal{I}(G)$,

$$\sum_{v \in I} y_v = b \cdot \sum_{v \in I} \mu(v) = b \cdot \Pr_{v \sim \mu}(v \in I) \leq 1,$$

by the definition of b. Thus y is a feasible dual solution. Moreover,

$$\sum_{v \in V(G)} y_v = b \cdot \sum_{v \in V(G)} \mu(v) = b.$$

As the dual LP in Eq (A.1) is a maximization LP, we have that $\chi_f(G) \geq b = \max_{I \in \mathcal{I}(G)} \Pr(v \in I)^{-1}$, for any distribution μ on the vertices.

Conversely, let y be any optimal solution to the dual LP and let $c := \sum_{v \in V} y_v$. Define a distribution μ on the vertices V by setting $\mu(v) = y_v / c$. For any independent set $I \in \mathcal{I}(G)$, we have,

$$\Pr_{v \sim \mu}(v \in I) = \sum_{v \in I} \mu(v) = \sum_{v \in I} y_v / c \leq 1/c,$$

where the final inequality is because y is a feasible dual solution. Thus, there exists a distribution μ such that $\chi_f(G) = c \leq \max_{I \in \mathcal{I}(G)} \Pr(v \in I)^{-1}$.

Combining these two parts concludes the proof. ∎

B Covering The Full Range of the Universe Size

We now generalize the proof of Theorem 2 to the full parameter range specified in the theorem. Consider u and n satisfying

$$n 2^{2^{\sqrt{\log \log n}}} \leq u \leq 2^{n^{n^2+n}}.$$

Notice that, on the lower-bound side, we are actually covering a slightly larger range (and therefore proving a slightly stronger result) than required to establish Theorem 2.

Set

$$m = (\log \log u)^{1/6} \quad \text{and} \quad k = n/m = n/(\log \log u)^{1/6}.$$

Note that the setting of k implicitly requires that $(\log \log u)^{1/6} \leq n$, which follows from the fact that $(\log \log u)^{1/6} \leq (n^2 + n)^{1/6} \leq \sqrt{n}$.

The k-fold conict graph $G^k(m)$ has $\log_f(G^k(m)) =$
$(n \log m) =$
$(n \log\log\log u)$ as already argued in Section 3.2. To complete the proof, we must establish that the graph $G^k(m)$ has vertices that are subsets of a universe whose size u_0 satises $u_0 \; u$. Solving for u^0, we have that

$$u^0 = kM = \frac{n}{(\log\log u)^{1=6}} \; 2^{m^{m^2+m}} \quad n\, 2^{2^{m^3=2}} \quad n\, 2^{2^{\sqrt{\log\log u=2}}}:$$

On the other hand, $u \; n\, 2^{2^{\sqrt{\log\log u}}}$. It follows that

$$\frac{u_0}{u} \quad \frac{2^{2^{\sqrt{\log\log u}}}}{2^{2^{\sqrt{\log\log u=2}}}} \quad 1;$$

which completes the proof of Theorem 2 for any choice of u between $n\, 2^{2^{\sqrt{\log\log n}}}$ and $2^{n^{n^2+n}}$.

Finally, we remark that the term $2^{n^{n^2+n}}$ in the upper bound is not tight and can be replaced by any other $2^{2^{poly(n)}}$ term; this is simply because for any $u = 2^{2^{poly(n)}}$, $\log\log\log u = (\log n)$ and thus for any larger universe size u also, we can simply focus on the smallest $2^{n^{n^2+n}}$ numbers in the universe and still obtain the same asymptotic lower bound. The lower bound term is also not tight and can be replaced with $n\, 2^{2^{(\log\log n)^{"}}}$ for any constant $" \; 2 \; (0; 1=2)$ by the same argument.