Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2021, Vol. 57, No. 2, 627–643
https://doi.org/10.1214/20-AIHP1089

New lower bounds for trace reconstruction

Zachary Chase

Mathematical Institute, Andrew Wiles Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, UK.
E-mail: zachary.chase@maths.ox.ac.uk

Received 13 November 2019; revised 13 April 2020; accepted 16 July 2020

Abstract. We improve the lower bound on worst case trace reconstruction from $\Omega(\frac{n^{5/4}}{\sqrt{\log n}})$ to $\Omega(\frac{n^{3/2}}{\log^7 n})$. As a consequence, we improve the lower bound on average case trace reconstruction from $\Omega(\frac{\log^{9/4} n}{\sqrt{\log\log n}})$ to $\Omega(\frac{\log^{5/2} n}{(\log\log n)^7})$.

Résumé. Nous améliorons la borne inférieure, dans le cas pire, de la reconstruction par trace de $\Omega(\frac{n^{5/4}}{\sqrt{\log n}})$ à $\Omega(\frac{n^{3/2}}{\log^7 n})$. Comme conséquence, nous améliorons la borne inférieure, dans le cas moyenné, de la reconstruction par trace de $\Omega(\frac{\log^{9/4} n}{\sqrt{\log\log n}})$ à $\Omega(\frac{\log^{5/2} n}{(\log\log n)^7})$.

MSC2020 subject classifications: 62B10

Keywords: Statistics

1. Introduction

Given a string $x \in \{0,1\}^n$, a *trace* of x is obtained by deleting each bit of x with probability q, independently, and concatenating the remaining string. For example, a trace of 11001 could be 101, obtained by deleting bits 2 and 3. The goal of the trace reconstruction problem is to determine an unknown string x, with high probability, by looking at as few independently generated traces of x as possible.

More precisely, fix $\delta, q \in (0,1)$. Take n large. For each $x \in \{0,1\}^n$, let μ_x be the probability distribution on $\{0,1\}^{\leq n}$ given by $\mu_x(w) = (1-q)^{|w|} q^{n-|w|} f(w;x)$, where $f(w;x)$ is the number of times w appears as a subsequence in x, that is, the number of strictly increasing tuples $(i_1, \dots, i_{|w|})$ such that $x_{i_j} = w_j$ for $1 \leq j \leq |w|$. The problem is to determine the minimum value of $T = T(n)$ for which there exists a function $f : (\{0,1\}^{\leq n})^T \to \{0,1\}^n$ satisfying $\mathbb{P}_{\mu_x^T}[f(\widetilde{U}^1, \dots, \widetilde{U}^T) = x] \geq 1 - \delta$ for each $x \in \{0,1\}^n$ (where the $\widetilde{U}^j$ denote the T independently generated traces).

The problem of trace reconstruction was introduced by Batu, Kannan, Khanna, and McGregor [2] as "an abstraction and simplification of a fundamental problem in bioinformatics, where one desires to reconstruct a common ancestor of several organisms given genetic sequences from those organisms." [8]

Holenstein, Mitzenmacher, Panigrahy, and Wieder [10] established an upper bound, that $\exp(\widetilde{O}(n^{1/2}))$ traces suffice. Nazarov and Peres [12] and De, O'Donnell, and Servedio [6] simultaneously obtained the best upper bound known, that $\exp(O(n^{1/3}))$ traces suffice. The lower bound of $\Omega(n)$ was established in [2], by considering the strings $0^{\frac{n}{2}-1}10^{\frac{n}{2}}$ and $0^{\frac{n}{2}}10^{\frac{n}{2}-1}$. Holden and Lyons [8] obtained the (previous) best lower bound known, by presenting two strings $x'_n \neq y'_n \in \{0,1\}^n$ which require $\Omega(n^{5/4}/\sqrt{\log n})$ traces to distinguish between. Their idea was to keep a 1 as a "defect" in the middle of the string, but to "pad" with 01's instead of 0's.

In this paper, we improve the lower bound, exhibiting two strings $x_n \neq y_n \in \{0,1\}^n$ which require $\Omega(n^{3/2}/\log^7 n)$ traces to distinguish between. In fact, our methods show that $\Omega(n^{3/2}/\log^7 n)$ traces are required to distinguish between x'_n and y'_n as well (a (messier) analogue of (12) holds). We also use the idea of padding a "defect" 1 with 01's. Our strings are slightly different than those considered in [8], for computational ease.

Let $k \geq 1$, $n = 4k+3$, and $x_n = (01)^k 1 (01)^{k+1}$, $y_n = (01)^{k+1} 1 (01)^k$, i.e.

$$x_n = 0101...0101 \quad 1 \quad 01 \quad 0101...0101$$
$$y_n = 0101...0101 \quad 01 \quad 1 \quad 0101...0101.$$

Theorem 1. *Fix $q, \delta \in (0,1)$. Then there exists some constant $c = c(q,\delta) > 0$ so that at least $cn^{3/2}/\log^7 n$ traces are required to distinguish between x_n and y_n with probability at least $1-\delta$, under trace reconstruction with deletion probability q.*

The main reason we are able to obtain an improvement over $n^{5/4}$ is that we explicitly compute (something very similar to) the quantity relevant to determining the number of samples needed, rather than relying on a coupling argument to determine only the total variation distance of the measures induced on subsequences.

A variant of the trace reconstruction problem is, instead of being required to reconstruct any string x from traces of it, one must reconstruct a string x chosen uniformly at random from traces of it. For a formal statement of the problem, see Section 1.2 of [8]. The best upper bound known, due to Holden, Pemantle, and Peres, is that $\exp(O(\log^{1/3} n))$ traces suffice [9]. The (previous) best lower bound known was $\Omega(\frac{\log^{9/4} n}{\sqrt{\log\log n}})$ [8]. Proposition 4.1 of [8] together with Theorem 1 implies

Theorem 2. *For all $q \in (0,1)$, there is $c = c(q) > 0$ so that for all large n, the probability of reconstructing a random n-bit string from $c\log^{5/2}(n)/(\log\log n)^7$ traces is at most $\exp(-n^{0.15})$, under trace reconstruction with deletion probability q.*

Very recently, other variants of the trace reconstruction problem have been considered. The interested reader should refer to [1,4,5], and [11].

Here is an outline of the paper. In Section 2, we recall "the distance" (namely, the Hellinger distance) between two probability measures that is directly relevant for determining the number of samples needed to distinguish between them, and we deduce Theorem 1 assuming an appropriate estimate. In Section 3, we prove the estimate by obtaining closed form expressions for the probability distributions induced by the traces of x_n and y_n and related expressions. In Section 4, we give the proofs of some lemmas used throughout Section 3. Finally, in Section 5 we establish a result of independent interest, a nontrivial bound on the number of traces that suffice to distinguish between any pair of strings with a very large Hamming distance (in contrast to the small Hamming distance pair considered to get Theorem 1).

2. A warmup to the proof of Theorem 1

Throughout the proof, $A \lesssim B$ means $A \leq CB$ for some absolute constant C, and $A \asymp B$ means $A \lesssim B$ and $B \lesssim A$. We take $q = 1/2$ for ease; the (analogous) proof works for any $q \in (0,1)$. The variables (to be introduced later) j, t, a, b, f, m will always be integers, the variables ϵ_j, ϵ_t will always be integer multiples of $\frac{1}{3}$, and all expressions occurring in binomial coefficients will be integers (we clearly state when it appears otherwise due to slight abuse of notation). For a string w, we let $|w|$ denote the length of w, and for any positive integers a, b with $a \leq b$, we denote by $w_{a,b}$ the contiguous substring $w_a, w_{a+1}, \ldots, w_b$.

Fix $n \equiv 3 \pmod 4$ large. Let $k = \frac{n-3}{4}$. Let μ be the probability measure for the traces of x_n and ν be the probability measure for the traces of y_n. Let E be a subset of $\bigcup_{0\leq k\leq n}\{0,1\}^k$ with $\mu(E), \nu(E) \geq 1 - O(e^{-\frac{1}{2}\log^2 n})$. We define E in Section 3.2.

It is well known, though seemingly folklore, that the number of samples needed to distinguish between two probability distributions with high probability is proportional to the inverse square of the Hellinger distance between them (see, e.g., Lemma A.5 of [8]):

$$\frac{1}{H(\mu,\nu)^2} = \frac{1}{\sum_w (\sqrt{\mu(w)} - \sqrt{\nu(w)})^2}.$$

Note

$$\sum_w \big(\sqrt{\mu(w)} - \sqrt{\nu(w)}\big)^2 \leq \sum_{w\in E} \big(\sqrt{\mu(w)} - \sqrt{\nu(w)}\big)^2 + \sum_{w\notin E} \big(\mu(w) + \nu(w)\big),$$

so since

$$\mu\big(E^c\big), \nu\big(E^c\big) \le O\big(e^{-\frac{1}{2}\log^2 n}\big),$$

to show that $\Omega(\frac{n^{3/2}}{\log^7 n})$ traces are necessary to distinguish between x_n and y_n, it suffices to show that

$$\sum_{w\in E}\big(\sqrt{\mu(w)}-\sqrt{\nu(w)}\big)^2 \lesssim \frac{\log^7 n}{n^{3/2}}.$$

And since

$$\sum_{w\in E}\big(\sqrt{\mu(w)}-\sqrt{\nu(w)}\big)^2 \le \sum_{w\in E}\left[\sqrt{\frac{\mu(w)}{\nu(w)}}+1\right]^2\big(\sqrt{\mu(w)}-\sqrt{\nu(w)}\big)^2 = \sum_{w\in E}\frac{(\mu(w)-\nu(w))^2}{\nu(w)},$$

to prove Theorem 1, it suffices to show

$$\sum_{w\in E}\frac{(\mu(w)-\nu(w))^2}{\nu(w)} \lesssim \frac{\log^7 n}{n^{3/2}}. \tag{1}$$

3. Proving inequality (1)

3.1. *Obtaining closed form expressions for μ and ν*

In this subsection, we obtain closed form expressions for the probability distributions of the traces of x_n and y_n. Let $s_k = (01)^k = 0101\ldots01$ be of length $2k$. Let $f_c(w)$ denote the number of contiguous 01 appearances in w.

We will use the following simple and fortuitous combinatorial lemma. It is the main reason we are able to obtain a simple(r) closed form expression.

Lemma 1. *For strings w, z, let $f(w; z)$ denote the number of times w appears as a subsequence in z, that is, the number of strictly increasing tuples $(i_1, \ldots, i_{|w|})$ such that $z_{i_j} = w_j$ for $1 \le j \le |w|$. Then, for any $k \ge 0$, $f(w; s_k) = \binom{k+f_c(w)}{m}$ if $|w| = m$.*

Proof. (Due to Russell Lyons) The idea is that every 01 occurring in w is a chance to put two consecutive indices in w in the same pair in s_k. Take any $1 \le j_1 < j_2 < \cdots < j_m \le k + f_c(w)$. Let $I_1 = j_1$ and $I_{p+1} = I_p + j_{p+1} - j_p - 1_{w_p=0=w_{p+1}-1}$ for $1 \le p \le m-1$. For each $1 \le p \le m$, let $i_p \in \{2I_p - 1, 2I_p\}$ be such that $w_p = (s_k)_{i_p}$. We thus get an occurrence of w in s_k; conversely, given any occurrence of w in s_k via $(i_p)_{1\le p\le m}$, we optain $(I_p)_{1\le p\le m}$ and then $(j_p)_{1\le p\le m}$ as above. The correspondence between $(j_p)_p$ and $(i_p)_p$ is a bijective one. □

Doing casework on whether w includes the "lone 1" (i.e. the 1 at index $2k+1$ in x, and the 1 at index $2k+3$ in y, where the convention is that the first index is 1), and if so, where it appears, Lemma 1 implies that

$$2^n\mu(w) = \binom{2k+f_c(w)}{|w|} + \sum_{\substack{1\le j\le |w|\\ w_j=1}}\binom{k+f_c(w_{1,j-1})}{j-1}\binom{k+1+f_c(w_{j+1,m})}{m-j}, \tag{2}$$

$$2^n\nu(w) = \binom{2k+f_c(w)}{|w|} + \sum_{\substack{1\le j\le |w|\\ w_j=1}}\binom{k+1+f_c(w_{1,j-1})}{j-1}\binom{k+f_c(w_{j+1,m})}{m-j}. \tag{3}$$

3.2. *The "high probability" set E*

We now define the "high probability" set used in Section 2. Let

$$E = \left\{w \in \{0,1\}^{\le n} : \big||w| - 2k\big| \le \sqrt{k}\log(k) \text{ and } \left|f_c(w) - \frac{2k}{3}\right| \le \sqrt{k}\log(k)\right\}.$$

In this subsection, we show $\mu(E), \nu(E) \geq 1 - O(e^{-\frac{1}{2}\log^2 n})$. To this end, and for the purposes of proving inequality (1), we make frequent use of the following technical lemma, used to estimate binomial coefficients. It is proven in Section 4.

Lemma 2. *For any real η bounded away from 0 and 1, any positive integers A and B such that $\eta A, \eta B \in \mathbb{Z}$, and any integers Δ and σ such that $A + \Delta$, $\eta A + \sigma$, $B - \Delta$, and $\eta B - \sigma$ are non-negative, it holds that*

$$\begin{aligned}
&\left[\frac{\binom{A+\Delta}{\eta A+\sigma}\binom{B-\Delta}{\eta B-\sigma}}{\binom{A}{\eta A}\binom{B}{\eta B}}\right]^{-1} \\
&\quad= \left(1+O\left(\frac{\sigma^3}{A^2}\right)\right)\left(1+O\left(\frac{\Delta^3}{A^2}\right)\right)\left(1+O\left(\frac{1}{A}\right)\right)\left(1+O\left(\frac{\sigma(\Delta-\sigma)^2}{A^2}\right)\right)\left(1+O\left(\frac{\Delta(\Delta-\sigma)^2}{A^2}\right)\right) \\
&\qquad\times \exp\left(\frac{1}{2}\frac{(\Delta-\sigma)^2}{(1-\eta)A}+\frac{1}{2}\frac{\sigma^2}{\eta A}-\frac{1}{2}\frac{\Delta^2}{A}\right) \\
&\qquad\times \left(1+O\left(\frac{\sigma^3}{B^2}\right)\right)\left(1+O\left(\frac{\Delta^3}{B^2}\right)\right)\left(1+O\left(\frac{1}{B}\right)\right)\left(1+O\left(\frac{\sigma(\Delta-\sigma)^2}{B^2}\right)\right)\left(1+O\left(\frac{\Delta(\Delta-\sigma)^2}{B^2}\right)\right) \\
&\qquad\times \exp\left(\frac{1}{2}\frac{(\Delta-\sigma)^2}{(1-\eta)B}+\frac{1}{2}\frac{\sigma^2}{\eta B}-\frac{1}{2}\frac{\Delta^2}{B}\right).
\end{aligned}$$

A corollary of Lemma 2 we will use frequently is that, if $A \leq B$, say, then the product $\binom{A+\Delta}{\eta A+\sigma}\binom{B-\Delta}{\eta B-\sigma}$ is, up to a $(1 + O(\frac{\log^3 A}{A}))$ multiplicative error, maximized at $\sigma = \Delta = 0$.

Formally, for any η, A, B, Δ, and σ with restrictions as in Lemma 2, we have

$$\binom{A+\Delta}{\eta A+\sigma}\binom{B-\Delta}{\eta B-\sigma} \lesssim \binom{A}{\eta A}\binom{B}{\eta B}. \tag{4}$$

For instance, (4), together with (2), implies that for any $w \in E$, if $m := |w|$ and $f := f_c(w)$,[1]

$$\begin{aligned}
2^n\mu(w) &\leq \binom{2k+f}{m} + m\max_j \binom{k+f_c(w_{1,j-1})}{j-1}\binom{k+1+f-f_c(w_{j+1,m})}{m-j} \\
&\leq \binom{2k+f}{m} + m\max_{j,a} \binom{k+a}{j-1}\binom{k+1+f-a}{m-j} \\
&\lesssim \binom{2k+f}{m} + m\binom{k+\frac{f}{2}}{\frac{m}{2}}^2 \\
&\lesssim \sqrt{k}\binom{2k+f}{m}.
\end{aligned} \tag{5}$$

The following is another simple combinatorial lemma.

Lemma 3. *For positive integers a and l, the number of $w \in \{0,1\}^l$ such that $f_c(w) = a$ is $\binom{l+1}{2a+1}$.*

Proof. The number of such strings is equal to the number of ways to place $2a+1$ indistinguishable flags in $l+1$ spots. Indeed, any such string $w = (w_1, \ldots, w_l)$ has exactly $2a+1$ indices i (a "flag"), $0 \leq i \leq l$ such that $w_i \neq w_{i+1}$, where we define $w_0 = 1$ and $w_{l+1} = 0$. And any choice of $2a+1$ flags corresponds to a w. This correspondence is a bijective one. □

Continuing from (5), Lemma 3 implies

$$\mu\left(\left\{w \in \{0,1\}^m : f_c(w) = f\right\}\right) \lesssim 2^{-n}\sqrt{k}\binom{2k+f}{m}\binom{m+1}{2f+1}. \tag{6}$$

[1] By $\frac{m}{2}$ and $\frac{f}{2}$, we mean $\lfloor\frac{m}{2}\rfloor$ and $\lfloor\frac{f}{2}\rfloor$. Similarly in the rest of the paper when $\frac{m}{2}$ and $\frac{f}{2}$ appear in binomial coefficients.

We now argue that we can restrict to m close to $2k$, allowing us to use Lemma 2 to then show that the right side of (6) is small for f far from $\frac{2k}{3}$. Since, for any m, $\sum_{w\in\{0,1\}^m}\mu(w) = 2^{-n}\binom{n}{m}$ and since $2^{-n}\binom{n}{m} = O(e^{-\log^2 n})$ for $m \notin [\frac{n}{2} - \sqrt{n}\log(n), \frac{n}{2} + \sqrt{n}\log(n)]$ (by, e.g., Lemma 2), we have

$$\mu\Bigg(\bigcup_{m\notin[2k-\sqrt{k}\log(k),2k+\sqrt{k}\log(k)]} \{0,1\}^m\Bigg) = O\big(e^{-\frac{1}{2}\log^2 n}\big). \tag{7}$$

Now assume $|m-2k| \le \sqrt{k}\log(k)$. Writing $m = 2k+\delta$ and $f = \frac{2k}{3}+\epsilon$, we see that

$$\begin{aligned}\binom{2k+f}{m}\binom{m}{2f} &= \binom{\frac{8k}{3}+\epsilon}{2k+\delta}\binom{2k+\delta}{\frac{4k}{3}+2\epsilon}\\ &= \binom{\frac{8k}{3}+\epsilon}{\frac{4k}{3}+2\epsilon}\binom{\frac{4k}{3}-\epsilon}{\frac{2k}{3}-2\epsilon+\delta}.\end{aligned}$$

Continuing from (6), using Lemma 2 with $A = \frac{8k}{3}$, $B = \frac{4k}{3}$, $\Delta = \epsilon$, $\sigma = 2\epsilon - \frac{2\delta}{3}$, and $\eta = \frac{2k+\delta}{4k} = \frac{1}{2} + O(\frac{\log k}{\sqrt{k}})$, we see that $|f - \frac{2k}{3}| > \sqrt{k}\log(k)$ implies

$$\begin{aligned}\mu\big(\{w\in\{0,1\}^m : f_c(w) = f\}\big) &\lesssim 2^{-4k}\sqrt{k}e^{-\log^2 n}\binom{8k/3}{4k/3}\binom{4k/3}{2k/3}\\ &\lesssim e^{-\log^2 n}.\end{aligned}$$

Hence, since there are at most n^2 values of (m, f), it holds that

$$\mu\Bigg(\bigcup_{\substack{m\in[2k-\sqrt{k}\log(k),2k+\sqrt{k}\log(k)]\\ f\notin[\frac{2k}{3}-\sqrt{k}\log(k),\frac{2k}{3}+\sqrt{k}\log(k)]}} \{w\in\{0,1\}^m : f_c(w) = f\}\Bigg) = O\big(e^{-\frac{1}{2}\log^2 n}\big). \tag{8}$$

Combining (7) and (8), we see

$$\mu(E) \ge 1 - O\big(e^{-\frac{1}{2}\log^2 n}\big). \tag{9}$$

The same argument shows that

$$\nu(E) \ge 1 - O\big(e^{-\frac{1}{2}\log^2 n}\big). \tag{10}$$

We take a moment to prove the following lemma, useful in the upcoming two sections, which allows us to focus on the probablistically relevant ranges of the parameters involved.

Lemma 4. *Let f and m be positive integers such that $|f - \frac{2k}{3}|, |m-2k| \le \sqrt{k}\log(k)$. Then, for any positive integers a, j, it holds that $\binom{k+a}{j-1}\binom{k+1+f-a}{m-j} \lesssim e^{-\log^2 k}\binom{\frac{4k}{3}}{m/2^2}$ unless $|a - \frac{f}{2}| \le \sqrt{k}\log(k)$ and $|j - \frac{m}{2}| \le \sqrt{k}\log(k)$.*

Proof. Lemma 2 implies, for any λ, $\beta = O(A^{1/6})$ and η bounded away from 0 and 1,

$$\binom{A+\lambda\sqrt{A}}{\eta A+\beta\sqrt{A}}\binom{A-\lambda\sqrt{A}}{\eta A-\beta\sqrt{A}} \lesssim e^{\lambda^2-\beta^2/\eta-(\lambda-\beta)^2/(1-\eta)}\binom{A}{\eta A}\binom{A}{\eta A}.$$

We use $A = \lfloor k + \frac{f}{2}\rfloor$, $\eta = \frac{m/2}{k+\frac{f}{2}} = \frac{3}{4} + O(\frac{\log k}{\sqrt{k}})$, $\lambda = \frac{a-\frac{f}{2}}{\sqrt{k+\frac{f}{2}}}$, and $\beta = \frac{j-\frac{m}{2}}{\sqrt{k+\frac{f}{2}}}$. □

3.3. *A closed form expression*

In this subsection, we obtain a closed form expression for an upper bound of $\sum_{w\in E}\frac{(\mu(w)-\nu(w))^2}{\nu(w)}$, up to an acceptable (for the purposes of proving (1)) error. By the definition of E and an obvious lower bound on ν coming from (3), we have

$$\sum_{w\in E}\frac{(\mu(w)-\nu(w))^2}{\nu(w)}\leq \sum_{\substack{m\in[2k-\sqrt{k}\log(k),2k+\sqrt{k}\log(k)]\\ f\in[\frac{2k}{3}-\sqrt{k}\log(k),\frac{2k}{3}+\sqrt{k}\log(k)]}}\frac{1}{2^n\binom{2k+f}{m}}\sum_{\substack{|w|=m\\ f_c(w)=f}}\big(2^n\mu(w)-2^n\nu(w)\big)^2. \tag{11}$$

We fix m and f and focus on estimating

$$\begin{aligned}
&\sum_{\substack{|w|=m\\ f_c(w)=f}}\big(2^n\mu(w)-2^n\nu(w)\big)^2\\
&=\sum_{\substack{|w|=m\\ f_c(w)=f}}\Bigg(\sum_{1\leq j\leq m:w_j=1}\binom{k+f_c(w_{1,j-1})}{j-1}\binom{k+1+f_c(w_{j+1,m})}{m-j}\\
&\quad-\binom{k+1+f_c(w_{1,j-1})}{j-1}\binom{k+f_c(w_{j+1,m})}{m-j}\Bigg)^2\\
&=\sum_{1\leq j,t\leq m}\sum_{\substack{|w|=m\\ f_c(w)=f\\ w_j=1\\ w_t=1}}\left[\binom{k+f_c(w_{1,j-1})}{j-1}\binom{k+1+f_c(w_{j+1,m})}{m-j}-\binom{k+1+f_c(w_{1,j-1})}{j-1}\binom{k+f_c(w_{j+1,m})}{m-j}\right]\\
&\quad\times\left[\binom{k+f_c(w_{1,t-1})}{t-1}\binom{k+1+f_c(w_{t+1,m})}{m-t}-\binom{k+1+f_c(w_{1,t-1})}{t-1}\binom{k+f_c(w_{t+1,m})}{m-t}\right],
\end{aligned}\tag{12}$$

where (12) refers to the expression occupying the final two lines. The first equality follows from (2) and (3), and the second follows by expanding out the square and interchanging summations.

We take the following page and a half to make restrictions on the variables involved in (12), allowing us to make future estimates more effectively.

We may restrict (12) to $j,t\in[\frac{m}{2}-\sqrt{k}\log(k),\frac{m}{2}+\sqrt{k}\log(k)]$ and w with $|f_c(w_{1,j-1})-\frac{f}{2}|\leq\sqrt{k}\log(k)$ and $|f_c(w_{1,t-1})-\frac{f}{2}|\leq\sqrt{k}\log(k)$. Indeed, if at least one of those four restrictions does not hold, then by Lemma 4 and (4),

$$\binom{k+f_c(w_{1,j-1})}{j-1}\binom{k+f_c(w_{j+1,m})}{m-j}\binom{k+f_c(w_{1,t-1})}{t-1}\binom{k+f_c(w_{t+1,m})}{m-t}\lesssim e^{-\log^2 k}\binom{\frac{4k}{3}}{m/2}^2.$$[2]

A quick calculation shows that

$$\begin{aligned}
&\binom{k+f_c(w_{1,j-1})}{j-1}\binom{k+1+f_c(w_{j+1,m})}{m-j}-\binom{k+1+f_c(w_{1,j-1})}{j-1}\binom{k+f_c(w_{j+1,m})}{m-j}\\
&=\binom{k+f_c(w_{1,j-1})}{j-1}\binom{k+f_c(w_{j+1,m})}{m-j}\\
&\quad\times\left[\frac{m-j}{k+1+f_c(w_{j+1,m})-(m-j)}-\frac{j-1}{k+1+f_c(w_{1,j-1})-(j-1)}\right].
\end{aligned}$$

The restrictions just made ensure that

$$\frac{m-j}{k+1+f_c(w_{j+1,m})-(m-j)}-\frac{j-1}{k+1+f_c(w_{1,j-1})-(j-1)}=O\left(\frac{\log(k)}{\sqrt{k}}\right). \tag{13}$$

[2]The fact that this bound is (more than) sufficient to indeed make the said restrictions follows from the same argument, about to be made, yielding (14).

Indeed, since $k+1+f_c(w_{j+1,m})-(m-j) \geq \frac{k}{3} - O(\sqrt{k}\log(k))$ and $k+1+f_c(w_{1,j-1})-(j-1) \geq \frac{k}{3} - O(\sqrt{k}\log(k))$, we have

$$\begin{aligned}
&\frac{m-j}{k+1+f_c(w_{j+1,m})-(m-j)} - \frac{j-1}{k+1+f_c(w_{1,j-1})-(j-1)} \\
&= \frac{m-j}{k+f_c(w_{j+1,m})-(m-j)} - \frac{j}{k+f_c(w_{1,j-1})-j} + O\left(\frac{1}{k}\right) \\
&= \frac{mk-2jk-jf_c(w_{j+1,m})+(m-j)f_c(w_{1,j-1})}{(k+f_c(w_{j+1,m})-m+j)(k+f_c(w_{1,j-1})-j)} + O\left(\frac{1}{k}\right) \\
&= \frac{O(k\sqrt{k}\log(k))}{\Omega(k^2)} \\
&= O\left(\frac{\log(k)}{\sqrt{k}}\right).
\end{aligned}$$

Up to a multiplicative factor of 2, we may restrict (12) to $t > j$ (the argument about to be made shows the diagonal $t = j$ term is sufficiently small). Furthermore, we may in fact restrict to $t > j+5$; indeed, by (4), Lemma 3, and (13), we see that expression (12) with the first sum restricted to $j < t \leq j+5$ is upper bounded by

$$5 \sum_{j \in [k-\sqrt{k}\log(k), k+\sqrt{k}\log(k)]} \sum_{\substack{|w|=m \\ f_c(w)=f}} \binom{k+\frac{f}{2}}{\frac{m}{2}}^4 \frac{\log^2(k)}{k} \lesssim \frac{\log^3(k)}{\sqrt{k}} \binom{k+\frac{f}{2}}{\frac{m}{2}}^4 \binom{m}{2f},$$

and so summing this over $|m-2k| \leq \sqrt{k}\log(k)$ and $|f - \frac{2k}{3}| \leq \sqrt{k}\log(k)$ with weights $\frac{1}{2^n\binom{2k+f}{m}}$, we obtain an upper bound up to a multiplicative constant for

$$\begin{aligned}
&\sum_{\substack{|m-2k| \leq \sqrt{k}\log(k) \\ |f-\frac{2k}{3}| \leq \sqrt{k}\log(k)}} \frac{1}{2^n\binom{2k+f}{m}} \sum_{\substack{1 \leq j < t \leq m \\ t \leq j+5}} \sum_{\substack{|w|=m \\ f_c(w)=f \\ w_j=1, w_t=1}} \left[\binom{k+f_c(w_{1,j-1})}{j-1} \binom{k+1+f_c(w_{j+1,m})}{m-j} \right. \\
&\left. - \binom{k+1+f_c(w_{1,j-1})}{j-1} \binom{k+f_c(w_{j+1,m})}{m-j} \right] \\
&\times \left[\binom{k+f_c(w_{1,t-1})}{t-1} \binom{k+1+f_c(w_{t+1,m})}{m-t} - \binom{k+1+f_c(w_{1,t-1})}{t-1} \binom{k+f_c(w_{t+1,m})}{m-t} \right]
\end{aligned}$$

of

$$\left(\sqrt{k}\log(k)\right)^2 \sup_{\substack{|m-2k| \leq \sqrt{k}\log(k) \\ |f-\frac{2k}{3}| \leq \sqrt{k}\log(k)}} \frac{\log^3(k)}{\sqrt{k}} \frac{\binom{k+\frac{f}{2}}{\frac{m}{2}}}{\binom{2k+f}{m}} \frac{\binom{k+\frac{f}{2}}{\frac{m}{2}}\binom{m}{2f}}{2^{4k}} \lesssim \frac{\log^5(k)}{k^{3/2}} \lesssim \frac{\log^5(n)}{n^{3/2}}. \tag{14}$$

One should compare to (11), the equation involving (12), and (1).

The following very important paragraph, which ignores multiplicative constants, explains the motivation behind the rest of the calculations in this paper.

In the calculations just above, we used the trivial upper bound of $\binom{k+\frac{f}{2}}{\frac{m}{2}}^4 \frac{\log^2(k)}{k}$ for the summands of (12). If we did not restrict to $t \leq j+5$ in the calculation just above and used that same trivial upper bound (which is indeed valid for $j, t \in [\frac{m}{2} - \sqrt{k}\log(k), \frac{m}{2} + \sqrt{k}\log(k)]$), we would get an upper bound for the right hand side of (11) of $\frac{\log^5(n)}{n^{3/2}}\sqrt{n}\log(n) = \frac{\log^6(n)}{n}$, since there are $\sqrt{k}\log(k)$ values of t rather than just 5. Therefore, we just need a savings of $\sqrt{k}/\log(k)$ over that trivial upper bound to obtain (1). Note in that trivial upper bound, we just bounded each term individually, not using any cancellation amongst the different summands. Our goal in Section 3.4 is to analyze the left hand side of (13) very carefully, in order to exploit cancellation between different summands of (12). To make the paper significantly shorter,

we do not repeatedly make the type of calculation just made above; rather, we point out where $\widetilde{\Omega}(\sqrt{k})$ savings come from as we go along.

Fix some t and j with $t > j + 5$.[3] We will now separate the sum over w in (12) based on $f_c(w_{1,j-1})$ and $f_c(w_{1,t-1})$. To relate $f_c(w_{1,j-1})$ to $f_c(w_{j+1,m})$ and $f_c(w_{1,t-1})$ to $f_c(w_{t+1,m})$ given $f_c(w)$, we need to do casework on w_{j-1} and w_{t-1}. We first do the case of $w_{j-1} = w_{t-1} = 0$. In this case, $f_c(w_{j+1,m}) = f - f_c(w_{1,j-1}) - 1$ and $f_c(w_{t+1,m}) = f - f_c(w_{1,t-1}) - 1$. This gives the "first case" of (12):

$$\sum_{\substack{|w|=m\\ f_c(w)=f\\ w_{j-1}=0, w_j=1\\ w_{t-1}=0, w_t=1}} \left[\binom{k+f_c(w_{1,j-1})}{j-1}\binom{k+1+f_c(w_{j+1,m})}{m-j} - \binom{k+1+f_c(w_{1,j-1})}{j-1}\binom{k+f_c(w_{j+1,m})}{m-j}\right]$$

$$\times\left[\binom{k+f_c(w_{1,t-1})}{t-1}\binom{k+1+f_c(w_{t+1,m})}{m-t} - \binom{k+1+f_c(w_{1,t-1})}{t-1}\binom{k+f_c(w_{t+1,m})}{m-t}\right]$$

$$= \sum_{a,b\geq 0} \sum_{\substack{|w|=m\\ f_c(w)=f\\ w_{j-1}=0, w_j=1\\ w_{t-1}=0, w_t=1\\ f_c(w_{1,j-1})=a\\ f_c(w_{1,t-1})=b}} \left[\binom{k+a}{j-1}\binom{k+f-a}{m-j} - \binom{k+1+a}{j-1}\binom{k+f-a-1}{m-j}\right]$$

$$\times\left[\binom{k+b}{t-1}\binom{k+f-b}{m-t} - \binom{k+1+b}{t-1}\binom{k+f-b-1}{m-t}\right].$$

Removing the product (that does not depend on w) from the inner sum, we wish to count the set of w with $|w| = m$, $f_c(w) = f$, $w_{j-1} = 0$, $w_j = 1$, $w_{t-1} = 0$, $w_t = 1$, $f_c(w_{1,j-1}) = a$, and $f_c(w_{1,t-1}) = b$. Noting that $f_c(w_{1,j-1}) = f_c(w_{1,j-2})$, we use

$$f_c(w_{1,t-1}) = f_c(w_{1,j-1}) + f_c(w_{j-1,t-1}) = f_c(w_{1,j-1}) + 1 + f_c(w_{j+1,t-1})$$

and

$$f_c(w_{j+1,t-1}) = f_c(w_{j+1,t-2})$$

together with Lemma 3 to get that the number of such w is $\binom{j-1}{2a+1}\binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1}$. So, the case of $w_{j-1} = w_{t-1} = 0$ yields expression (15):

$$\sum_{\substack{t>j+5\\ a,b\geq 0}} \left[\binom{k+a}{j-1}\binom{k+f-a}{m-j} - \binom{k+1+a}{j-1}\binom{k+f-a-1}{m-j}\right]\binom{j-1}{2a+1}\binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1}$$

$$\times\left[\binom{k+b}{t-1}\binom{k+f-b}{m-t} - \binom{k+1+b}{t-1}\binom{k+f-b-1}{m-t}\right]. \tag{15}$$

The other three cases of the value of the pair (w_{j-1}, w_{t-1}) yield very similar expressions. The only difference between the expressions is that some binomial coefficients have $-1, -2, +1, +2$, or 0 in certain places. However, these minor differences will not affect our proceeding arguments. That is, our argument for a $\sqrt{k}/\log(k)$ savings for the $(w_{j-1}, w_{t-1}) = (0, 0)$ case would show a $\sqrt{k}/\log(k)$ savings for the other 3 cases. Therefore, we may restrict attention to the case $(w_{j-1}, w_{t-1}) = (0, 0)$.

3.4. *Finishing the proof of* (1)

In this final subsection, we appropriately bound (15), thereby proving (1). As explained in the last section, we may assume $a \in [\frac{f}{2} - \sqrt{k}\log(k), \frac{f}{2} + \sqrt{k}\log(k)]$, thereby, as before, yielding

$$\binom{k+a}{j-1}\binom{k+f-a}{m-j} - \binom{k+1+a}{j-1}\binom{k+f-a-1}{m-j}$$

[3] We wanted to restrict to $t > j + 5$ so that the following case analysis has no "boundary issues".

$$= \binom{k+a}{j-1}\binom{k+f-a-1}{m-j}\left[\frac{m-j}{k+f-a-(m-j)} - \frac{j}{k+a-j} + O\left(\frac{1}{k}\right)\right].$$

Let δ_j and ϵ_j be defined so that

$$j = \frac{m}{2} + \delta_j$$

and

$$a = \frac{j}{3} + \frac{f}{2} - \frac{m}{6} + \epsilon_j.$$

Observe that

$$\frac{m-j}{k+f-a-(m-j)} - \frac{j}{k+a-j} = \frac{-2k\delta_j + \frac{m\delta_j}{3} + m\epsilon_j - f\delta_j}{(k+f-a-\frac{m}{2}+\delta_j)(k+a-\frac{m}{2}-\delta_j)}.$$

Since $a \in [\frac{f}{2} - \sqrt{k}\log(k), \frac{f}{2} + \sqrt{k}\log(k)]$, we have $\epsilon_j = O(\sqrt{k}\log(k))$. Since also $m = 2k + O(\sqrt{k}\log(k))$ and $f = \frac{2k}{3} + O(\sqrt{k}\log(k))$, we see that

$$\frac{m-j}{k+f-a-(m-j)} - \frac{j}{k+a-j} = 18\frac{1}{k}[\epsilon_j - \delta_j] + O\left(\frac{\log^2(k)}{k}\right).$$

Therefore, defining δ_t[4] and ϵ_t so that

$$t = \frac{m}{2} + \delta_t$$

and

$$b = \frac{t}{3} + \frac{f}{2} - \frac{m}{6} + \epsilon_t,$$

we see that (15) takes the form

$$\frac{324}{k^2}\sum_{a,b,t,j}\binom{k+a}{j-1}\binom{k+f-a-1}{m-j}\binom{j-1}{2a+1}\binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1}$$
$$\times\binom{k+b}{t-1}\binom{k+f-b-1}{m-t}[\delta_j-\epsilon_j]\cdot[\delta_t-\epsilon_t] \tag{16}$$

up to an acceptable error (the error is acceptable since it replaces a bound of $\frac{\log(k)}{\sqrt{k}}$ for $|\delta_j - \epsilon_j|$, say, with $\frac{\log^2(k)}{k}$, giving our desired $\log(k)/\sqrt{k}$ savings). Recall that we are summing over $t, j \in [k - \sqrt{k}\log(k), k + \sqrt{k}\log(k)]$.

We now claim that, unless $b = a + \frac{t-j}{3} + O(\sqrt{t-j}\log(k))$, the magintude of the summand corresponding to a, b, j, t is sufficiently small. Note that $\binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1} \asymp \binom{\delta_t-\delta_j}{\frac{2}{3}(\delta_t-\delta_j)+2\epsilon_t-2\epsilon_j}\binom{\frac{m}{2}-\delta_t}{f-\frac{2\delta_t}{3}-2\epsilon_t}$. We may use Lemma 2 with $A = \delta_t - \delta_j$, $B = \frac{m}{2} - \delta_t$, $\eta = \frac{f-\frac{2}{3}\delta_j-2\epsilon_j}{\frac{m}{2}-\delta_j} = \frac{2}{3} + O(\frac{\log k}{\sqrt{k}})$, $\Delta = 0$, $\sigma = 2\epsilon_t - 2\epsilon_j - (\frac{f-\frac{2}{3}\delta_j-2\epsilon_j}{\frac{m}{2}-\delta_j} - \frac{2}{3})(\delta_t - \delta_j)$ to deduce that $(b - a - \frac{t-j}{3})^2 > (t-j)\log^2(k)$ implies an $e^{-\log^2(n)}$ savings, verifying the claim.

Lemma 2 also implies that[5]

$$\binom{k+b}{t-1}\binom{k+f-b-1}{m-t} = \binom{k+a+\frac{t-j}{3}}{t-1}\binom{k+f-a-\frac{t-j}{3}-1}{m-t}\left(1+O\left(\frac{\log^5(k)}{\sqrt{k}}\right)\right)$$

[4]We are abusing notation here. Formally, define a function δ by $\delta(x) = x - \frac{m}{2}$; we use δ_j as shorthand for $\delta(j)$ and δ_t as shorthand for $\delta(t)$. Analogously for ϵ_j, ϵ_t.

[5]Technically, we are adding and substracting $\lfloor a + \frac{t-j}{3} \rfloor$ rather than $a + \frac{t-j}{3}$.

for $b = a + \frac{t-j}{3} + O(k^{1/4}\log^{3/2}(k))$. Therefore, we see that (16) is, up to a multiplicative factor of $1 + O(\frac{\log^5(k)}{\sqrt{k}})$, equal to

$$\frac{324}{k^2}\sum_{a,b,j,t}\binom{k+a}{j-1}\binom{k+f-a-1}{m-j}\binom{j-1}{2a+1}\binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1}$$
$$\times\binom{k+a+\frac{t-j}{3}}{t-1}\binom{k+f-a-\frac{t-j}{3}-1}{m-t}[\delta_j-\epsilon_j]\cdot[\delta_t-\epsilon_t], \tag{17}$$

where the sum is restricted to $|b-a-\frac{t-j}{3}| \le \sqrt{t-j}\log(k)$.

Our strategy now to exploit cancellation occurring between different summands is as follows. We split the term $\delta_t - \epsilon_t$ into three terms and deal with each separately, each by fixing j, t, and a, and summing over b. We get cancellation from the second term by pairing the summand corresponding to b to the summand corresponding to the reflection of b about a natural symmetry (explained below). The third term has magnitude a factor of $\sqrt{k}$ less than $\delta_t - \epsilon_t$ (i.e. it is $O(1)$), so it can be ignored. The first term requires the most work and is dealt with after the second and third are handled.

Specifically, we split up

$$[\delta_t-\epsilon_t] = [\delta_t-\epsilon_j] + \left[\epsilon_j + \left(\frac{f-\frac{2}{3}\delta_j-2\epsilon_j}{m-2\delta_j}-\frac{1}{3}\right)(\delta_t-\delta_j)-\epsilon_t\right]$$
$$-\left[\left(\frac{f-\frac{2}{3}\delta_j-2\epsilon_j}{m-2\delta_j}-\frac{1}{3}\right)(\delta_t-\delta_j)\right].$$

For any fixed a, j, and t, by Lemma 2 with $A = \delta_t - \delta_j$, $B = \frac{m}{2} - \delta_t$, $\eta = \frac{f-\frac{2}{3}\delta_j-2\epsilon_j}{\frac{m}{2}-\delta_j} = \frac{2}{3} + O(\frac{\log k}{\sqrt{k}})$, $\Delta = 0, \sigma = 2\epsilon_t - 2\epsilon_j - (\frac{f-\frac{2}{3}\delta_j-2\epsilon_j}{\frac{m}{2}-\delta_j} - \frac{2}{3})(\delta_t-\delta_j)$, we have that

$$\binom{\delta_t-\delta_j}{\frac{2}{3}(\delta_t-\delta_j)+2\epsilon_t-2\epsilon_j}\binom{\frac{m}{2}-\delta_t}{f-\frac{2}{3}\delta_t-2\epsilon_t}$$
$$=\left(1+O\left(\frac{\log^3(k)}{\sqrt{\delta_t-\delta_j}}\right)\right)\binom{\delta_t-\delta_j}{\frac{2}{3}(\delta_t-\delta_j)+2\epsilon_t^*-2\epsilon_j}\binom{\frac{m}{2}-\delta_j}{f-\frac{2}{3}\delta_t-2\epsilon_t^*},$$

where ϵ_t^* is the reflection[6] of ϵ_t about $\epsilon_j + \frac{1}{2}(\frac{f-\frac{2}{3}\delta_j-2\epsilon_j}{\frac{m}{2}-\delta_j} - \frac{2}{3})(\delta_t-\delta_j)$.[7] And therefore, since

$$\epsilon_j + \frac{1}{2}\left(\frac{f-\frac{2}{3}\delta_j-2\epsilon_j}{\frac{m}{2}-\delta_j}-\frac{2}{3}\right)(\delta_t-\delta_j)-\epsilon_t = O\left(\sqrt{\delta_t-\delta_j}\log(k)+\log^2(k)\right),$$

letting b^* denote the b corresponding to ϵ_t^*, we deduce that

$$\left|\sum_b\binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1}\left[\epsilon_j+\left(\frac{f-\frac{2}{3}\delta_j-2\epsilon_j}{m-2\delta_j}-\frac{1}{3}\right)(\delta_t-\delta_j)-\epsilon_t\right]\right|$$
$$=\frac{1}{2}\left|\sum_b\binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1}\left[\epsilon_j+\left(\frac{f-\frac{2}{3}\delta_j-2\epsilon_j}{m-2\delta_j}-\frac{1}{3}\right)(\delta_t-\delta_j)-\epsilon_t\right]\right.$$
$$\left.+\binom{t-j-1}{2b^*-2a-1}\binom{m-t+1}{2f-2b^*-1}\left[\epsilon_j+\left(\frac{f-\frac{2}{3}\delta_j-2\epsilon_j}{m-2\delta_j}-\frac{1}{3}\right)(\delta_t-\delta_j)-\epsilon_t^*\right]\right|$$
$$=\frac{1}{2}\left|\sum_b\binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1}\left[\epsilon_j+\left(\frac{f-\frac{2}{3}\delta_j-2\epsilon_j}{m-2\delta_j}-\frac{1}{3}\right)(\delta_t-\delta_j)-\epsilon_t\right]\right.$$

[6] To be precise, the reflection of x about y is defined to be $2y - x$.

[7] We might have to round ϵ_t^* a bit (so that $\frac{t}{3} + \frac{f}{2} - \frac{m}{6} + \epsilon_t^*$ is an integer), but the induced error in this rounding is negligible, by Lemma 2.

$$
\begin{aligned}
&\quad - \left(1 + O\left(\frac{\log^3(k)}{\sqrt{\delta_t - \delta_j}}\right)\right)\binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1}\left[\epsilon_j + \left(\frac{f - \frac{2}{3}\delta_j - 2\epsilon_j}{m - 2\delta_j} - \frac{1}{3}\right)(\delta_t - \delta_j) - \epsilon_t\right]\Bigg| \\
&\lesssim \sum_b \binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1} O\left(\frac{\log^3(k)}{\sqrt{\delta_t - \delta_j}}\right) O\left(\sqrt{\delta_t - \delta_j}\log(k) + \log^2(k)\right) \\
&\lesssim \sum_b \binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1}\log^5(k)
\end{aligned}
$$

is small enough, since we rid of a factor of $\widetilde{\Omega}(\sqrt{k})$ potentially coming from $\delta_t - \epsilon_t$. And since $(\frac{f-\frac{2}{3}\delta_j-2\epsilon_j}{m-2\delta_j} - \frac{1}{3})(\delta_t - \delta_j) = O(1)$ rather than $\Omega(\sqrt{k})$, the expression corresponding to the second term, namely

$$
\sum_b \binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1}\left[\left(\frac{f - \frac{2}{3}\delta_j - 2\epsilon_j}{m - 2\delta_j} - \frac{1}{3}\right)(\delta_t - \delta_j)\right],
$$

is small enough. Therefore, for any fixed a, j, t, the part of the sum in (17) with terms containing b is, up to negligible error, the expression corresponding to the remaining term:

$$
\sum_b \binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1}[\delta_t - \epsilon_j]. \tag{18}
$$

If $t > j + 5$, Lemma 7, proven in Section 4, states that

$$
\sum_b \binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1} = \left(\frac{1}{2} + O\left(\frac{\log^2(k)}{t-j}\right)\right)\sum_b \binom{t-j-1}{b-2a-1}\binom{m-t+1}{2f-b-1}.
$$

And using the general combinatorial identity

$$
\sum_C \binom{D}{C}\binom{E}{F-C} = \binom{D+E}{F}
$$

(we may extend the range of b and restrict it freely, since the b outside $a + \frac{t-j}{3} \pm \sqrt{t-j}\log(k)$ yield exponentially (in $\log^2 n$) small terms), we see that

$$
\sum_b \binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1} = \left(\frac{1}{2} + O\left(\frac{\log^2 k}{t-j}\right)\right)\binom{m-j}{2f-2a-2}. \tag{19}
$$

Therefore, noting that $\delta_t - \epsilon_j$ does not depend on b and then plugging (19) into (18), we see that (17) is, up to a negligible error, equal to

$$
\begin{aligned}
&\frac{162}{k^2}\sum_{a,j,t}\left(1 + O\left(\frac{\log^2(k)}{t-j}\right)\right)\binom{k+a}{j-1}\binom{k+f-a-1}{m-j}\binom{j-1}{2a+1}\binom{m-j}{2f-2a-2} \\
&\quad \times \binom{k+a+\frac{t-j}{3}}{t-1}\binom{k+f-a-\frac{t-j}{3}-1}{m-t}[\delta_j - \epsilon_j]\cdot[\delta_t - \epsilon_j],
\end{aligned} \tag{20}
$$

where, to reiterate, the sum is restricted to $t > j + 5$.

We can rid of the $O(\frac{\log^2(k)}{t-j})$ term trivially. Indeed, using (4), we can upper bound

$$
\begin{aligned}
&\binom{k+a}{j-1}\binom{k+f-a-1}{m-j} \lesssim \binom{k+\frac{f}{2}}{\frac{m}{2}}^2, \\
&\binom{k+a+\frac{t-j}{3}}{t-1}\binom{k+f-a-\frac{t-j}{3}-1}{m-t} \lesssim \binom{k+\frac{f}{2}}{\frac{m}{2}}^2,
\end{aligned}
$$

and

$$\binom{j-1}{2a+1}\binom{m-j}{2f-2a-2} \lesssim \binom{\frac{m}{2}}{f}^2;$$

noting that for each $\Delta \geq 5$, the number of pairs $(t, j) \in [\frac{m}{2} - \sqrt{k}\log(k), \frac{m}{2} + \sqrt{k}\log(k)]$ with $t - j = \Delta$ is at most $\sqrt{k}\log(k)$, we thus obtain an upper bound of

$$\frac{162}{k^2}\sqrt{k}\log(k)\sqrt{k}\log(k)\binom{k+\frac{f}{2}}{\frac{m}{2}}^4\binom{\frac{m}{2}}{f}^2\sum_{\Delta=5}^{\sqrt{k}\log(k)}\frac{\log^2(k)}{\Delta},$$

which is small enough; i.e., $t - j$ is on average $\sqrt{k}$, which gives us the required savings (note we get the $\log^7(n)$ from here, since summing over m and f picks up two extra $\log(k)$ factors). Note that we needed the error in Lemma 7 to be $O(\frac{\log^2(k)}{t-j})$ rather than the trivial $O(\frac{\log^2(k)}{\sqrt{t-j}})$, since the latter would have led to the sum $\sum_{\Delta=5}^{\sqrt{k}\log(k)}\frac{\log^2(k)}{\sqrt{\Delta}}$, which would have yielded a $k^{1/4}$ factor rather than a $\log(k)$ factor.

Let

$$f(\delta_j, \epsilon_j) = \frac{162}{k^2}\binom{k+a}{j}\binom{k+f-a-1}{m-j}\binom{j-1}{2a+1}\binom{m-j}{2f-2a-2}$$

and

$$g(\delta_t, \epsilon_j) = \binom{k+a+\frac{t-j}{3}}{t-1}\binom{k+f-a-\frac{t-j}{3}-1}{m-t}.$$

We break up the remaining expression, i.e., expression (20) without the $O(\frac{\log^2(k)}{t-j})$ term, as follows:

$$\begin{aligned}
&\sum_{\substack{\epsilon_j,\delta_j,\delta_t \\ \delta_t > \delta_j + 5}} f(\epsilon_j, \delta_j)[\delta_j - \epsilon_j]g(\delta_t, \epsilon_j)[\delta_t - \epsilon_j] \\
&= \sum_{\epsilon_j}\sum_{\delta_j > \epsilon_j}\Bigg[f(\epsilon_j, \delta_j)(\delta_j - \epsilon_j)\sum_{\delta_t > \delta_j + 5} g(\delta_t, \epsilon_j)(\delta_t - \epsilon_j) \\
&\quad + f(\epsilon_j, 2\epsilon_j - \delta_j)(\epsilon_j - \delta_j)\sum_{\delta_t > 2\epsilon_j - \delta_j + 5} g(\delta_t, \epsilon_j)(\delta_t - \epsilon_j)\Bigg]. \qquad (21)
\end{aligned}$$

We claim that g has symmetry[8] in δ_t about ϵ_j and f has symmetry in δ_j about ϵ_j: $g(\delta_t, \epsilon_j) \approx g(2\epsilon_j - \delta_t, \epsilon_j)$ and $f(\epsilon_j, \delta_j) \approx f(\epsilon_j, 2\epsilon_j - \delta_j)$. This is the content of the quite fortuitous Lemmas 5 and 7, respectively.[9]

Lemma 5. *For any positive integers f and m with $|f - \frac{2k}{3}|, |m - 2k| \leq \sqrt{k}\log k$ and for any integers δ_t, ϵ_j with $|\delta_t|, |\epsilon_j| \leq \sqrt{k}\log k$ and $\frac{\delta_t}{3} + \epsilon_j \in \mathbb{Z}$, it holds that*

$$\binom{k+\frac{f}{2}+\frac{\delta_t}{3}+\epsilon_j}{\frac{m}{2}+\delta_t}\binom{k+\frac{f}{2}-\frac{\delta_t}{3}-\epsilon_j}{\frac{m}{2}-\delta_t} = \left(1+O\left(\frac{\log^3(k)}{\sqrt{k}}\right)\right)\binom{k+\frac{f}{2}+\frac{\delta_t}{3}-\frac{5\epsilon_j}{3}}{\frac{m}{2}+\delta_t-2\epsilon_j}\binom{k+\frac{f}{2}-\frac{\delta_t}{3}+\frac{5\epsilon_j}{3}}{\frac{m}{2}-\delta_t+2\epsilon_j}.$$

Proof. Lemma 2, with $A = k + \frac{f}{2}$, $B = k + \frac{f}{2}$, $\eta = \frac{m/2}{k+\frac{f}{2}} = \frac{3}{4} + O(\frac{\log k}{\sqrt{k}})$, $\Delta = \frac{\delta_t}{3} + \epsilon_j$, $\sigma = \delta_t$ and $\Delta = \frac{\delta_t}{3} - \frac{5\epsilon_j}{3}$, $\sigma = \delta_t - 2\epsilon_j$ shows that both products of binomial coefficients are $(1 + O(\frac{\log^3(k)}{\sqrt{k}}))\exp(-\frac{3(\delta_t-\epsilon_j)^2}{k+\frac{f}{2}})\binom{k+f/2}{m/2}^2$. □

[8] See footnote 2 on page 632.

[9] The additive factors of $-1, +1$, and -2 have been omitted for ease. The proofs are the same with them present.

Lemma 6. *For any positive integers f and m with $|f - \frac{2k}{3}|, |m - 2k| \le \sqrt{k}\log k$ and for any integers δ_j, ϵ_j with $|\delta_j|, |\epsilon_j| \le \sqrt{k}\log k$ and $\frac{\delta_j}{3} + \epsilon_j \in \mathbb{Z}$, it holds that*

$$\binom{k + \frac{f}{2} + \frac{\delta_j}{3} + \epsilon_j}{\frac{m}{2} + \delta_j}\binom{k + \frac{f}{2} - \frac{\delta_j}{3} - \epsilon_j}{\frac{m}{2} - \delta_j} = \left(1 + O\left(\frac{\log^3(k)}{\sqrt{k}}\right)\right)\binom{k + \frac{f}{2} + \frac{\delta_j}{3} - \frac{5\epsilon_j}{3}}{\frac{m}{2} + \delta_j - 2\epsilon_j}\binom{k + \frac{f}{2} - \frac{\delta_j}{3} + \frac{5\epsilon_j}{3}}{\frac{m}{2} - \delta_j + 2\epsilon_j}$$

and

$$\binom{\frac{m}{2} + \delta_j}{f + \frac{2\delta_j}{3} + 2\epsilon_j}\binom{\frac{m}{2} - \delta_j}{f - \frac{2\delta_j}{3} - 2\epsilon_j} = \left(1 + O\left(\frac{\log^3(k)}{\sqrt{k}}\right)\right)\binom{\frac{m}{2} + 2\epsilon_j - \delta_j}{f + \frac{10\epsilon_j}{3} - \frac{2\delta_j}{3}}\binom{\frac{m}{2} - 2\epsilon_j + \delta_j}{f - \frac{10\epsilon_j}{3} + \frac{2\delta_j}{3}}.$$

Proof. The first approximation is the content of Lemma 5. For the second, use Lemma 2 with $A = \frac{m}{2}$, $B = \frac{m}{2}$, $\eta = \frac{2f}{m} = \frac{2}{3} + O(\frac{\log k}{\sqrt{k}})$, $\Delta = \delta_j, \sigma = \frac{2\delta_j}{3} + 2\epsilon_j$ and $\Delta = 2\epsilon_j - \delta_j, \sigma = \frac{10\epsilon_j}{3} - \frac{2\delta_j}{3}$ to see that both products of binomial coefficients are $(1 + O(\frac{\log^3(k)}{\sqrt{k}}))\exp(-\frac{18\epsilon_j^2}{m/2})\binom{\frac{m}{2}}{f}^2$. □

Lemma 5 implies that, for each fixed δ_j and ϵ_j, we have

$$\sum_{\delta_t > 2\epsilon_j - \delta_j + 5} g(\delta_t, \epsilon_j)(\delta_t - \epsilon_j) = \sum_{\delta_t > \delta_j + 5} g(\delta_t, \epsilon_j)(\delta_t - \epsilon_j) + O\left(\frac{\log^3(k)}{\sqrt{k}}\right)\sum_{\delta_t} g(\delta_t, \epsilon_j)|\delta_t - \epsilon_j|.$$

Indeed, for example, if $\delta_j < \epsilon_j$, then

$$\begin{aligned}
&\sum_{\delta_t > \delta_j + 5} g(\delta_t, \epsilon_j)(\delta_t - \epsilon_j) - \sum_{\delta_t > 2\epsilon_j - \delta_j + 5} g(\delta_t, \epsilon_j)(\delta_t - \epsilon_j) \\
&= \sum_{\epsilon_j - (\epsilon_j - \delta_j) + 5 < \delta_t \le \epsilon_j + (\epsilon_j - \delta_j) + 5} g(\delta_t, \epsilon_j)(\delta_t - \epsilon_j) \\
&= \sum_{\epsilon_j < \delta_t \le \epsilon_j + (\epsilon_j - \delta_j) + 5} \big[g(\delta_t, \epsilon_j) - g(2\epsilon_j - \delta_t, \epsilon_j)\big](\delta_t - \epsilon_j) \\
&= \sum_{\epsilon_j < \delta_t \le \epsilon_j + (\epsilon_j - \delta_j) + 5} O\left(\frac{\log^3(k)}{\sqrt{k}}\right) g(\delta_t, \epsilon_j)(\delta_t - \epsilon_j).
\end{aligned}$$

Therefore, (21) is, up to negligible error, equal to

$$\sum_{\epsilon_j}\sum_{\delta_j > \epsilon_j}\left(\big[f(\epsilon_j, \delta_j)(\delta_j - \epsilon_j) + f(\epsilon_j, 2\epsilon_j - \delta_j)(\epsilon_j - \delta_j)\big]\sum_{\delta_t > \delta_j + 5} g(\delta_t, \epsilon_j)(\delta_t - \epsilon_j)\right). \tag{22}$$

Lemma 6 then allows us to write (22) as

$$\sum_{\epsilon_j}\sum_{\delta_j > \epsilon_j}\left(\big[(\delta_j - \epsilon_j) + (\epsilon_j - \delta_j)\big] f(\epsilon_j, \delta_j)\sum_{\delta_t > \delta_j + 5} g(\delta_t, \epsilon_j)(\delta_t - \epsilon_j)\right)$$

up to a negligible error. But this is just 0, and so we've established (1).

4. Remaining proofs of lemmas

In this section, we prove Lemmas 7 and 2, restated here for the reader's convenience.

Lemma 7. *For any fixed positive integers a, j, t, m, f with $|m - 2k|, |j - \frac{m}{2}|, |t - \frac{m}{2}|, |f - \frac{2k}{3}|, |a - \frac{f}{2}| \le \sqrt{k}\log(k)$ and $t > j$, the following holds*:

$$\sum_b \binom{t - j - 1}{2b - 2a - 1}\binom{m - t + 1}{2f - 2b - 1} = \left(\frac{1}{2} + O\left(\frac{\log^2(k)}{t - j}\right)\right)\sum_b \binom{t - j - 1}{b - 2a - 1}\binom{m - t + 1}{2f - b - 1},$$

where the first sum is restricted to b with $|b-a-\frac{t-j}{3}| \le \sqrt{t-j}\log(k)$, and the second sum is restricted to b with $|b-2a-2\frac{t-j}{3}| \le 2\sqrt{t-j}\log(k)$.

Proof. The sum on the right contains all b in the range $[2a+2\frac{t-j}{3}-2\sqrt{t-j}\log(k), 2a+2\frac{t-j}{3}+2\sqrt{t-j}\log(k)]$, while the sum on the left contains only even b in that range. Therefore, due to the factor of $\frac{1}{2}$, we wish to show (23):

$$\sum_{b \text{ even}} \binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1} - \sum_{b \text{ odd}} \binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b-1}$$
$$= O\left(\frac{\log^2(k)}{t-j}\right)\sum_{b} \binom{t-j-1}{b-2a-1}\binom{m-t+1}{2f-2b-1}, \tag{23}$$

where the range of b is restricted to $|b-2a-2\frac{t-j}{3}| \le 2\sqrt{t-j}\log(k)$.

The idea of the proof is to pair every even-b term with $\frac{2}{3}$ times the (odd) term before it and $\frac{1}{3}$ times the (odd) term after it. Specifically, to establish (23), it suffices to show (24):

$$\frac{2}{3}\binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b+1} + \frac{1}{3}\binom{t-j-1}{2b-2a+1}\binom{m-t+1}{2f-2b-1}$$
$$= \left(1+O\left(\frac{\log^2(k)}{t-j}\right)\right)\binom{t-j-1}{2b-2a}\binom{m-t+1}{2f-2b}. \tag{24}$$

As mentioned on pages 638–639, the error $O(\frac{\log^2(k)}{\sqrt{t-j}})$ is trivial (it follows from pairing even-b terms with odd-b terms); our $\frac{2}{3}$-$\frac{1}{3}$ weighting gives the (necessary) improvement to $O(\frac{\log^2(k)}{t-j})$. Observe that

$$\frac{2}{3}\binom{t-j-1}{2b-2a-1}\binom{m-t+1}{2f-2b+1} + \frac{1}{3}\binom{t-j-1}{2b-2a+1}\binom{m-t+1}{2f-2b-1}$$

is, by using the equations $\binom{c}{d-1} = \frac{d}{c-d+1}\binom{c}{d}$ and $\binom{c}{d+1} = \frac{c-d}{d+1}\binom{c}{d}$, equal to

$$\binom{t-j-1}{2b-2a}\binom{m-t+1}{2f-2b}$$
$$\times\left[\frac{2}{3}\frac{2b-2a}{t-j-1-(2b-2a)+1}\frac{m-t+1-(2f-2b)}{2f-2b+2}\right.$$
$$\left.+\frac{1}{3}\frac{t-j-1-(2b-2a)}{2b-2a+1}\frac{2f-2b}{m-t+1-(2f-2b)+1}\right].$$

Since

$$\frac{m-t+1-(2f-2b)}{2f-2b+2} = \frac{1}{2}+O\left(\frac{\log(k)}{\sqrt{k}}\right)$$

and

$$\frac{2f-2b}{m-t+1-(2f-2b)+1} = 2+O\left(\frac{\log(k)}{\sqrt{k}}\right),$$

we may replace the expression above in brackets with, up to an acceptable error,

$$\frac{2}{3}\frac{b-a}{t-j-1-(2b-2a)}+\frac{1}{3}\frac{t-j-1-(2b-2a)}{b-a}. \tag{25}$$

Writing $b = a+\frac{t-j-1}{3}+\Delta$ transforms (25) into

$$\frac{(\frac{t-j-1}{3})^2+2\Delta^2}{(\frac{t-j-1}{3})^2-\frac{t-j-1}{3}\Delta-2\Delta^2},$$

which is $1+O(\frac{\log^2(k)}{t-j})$, the critical point being the lack of a $\frac{t-j-1}{3}\Delta$ term (which is why we chose the factors $\frac{2}{3}$ and $\frac{1}{3}$). This finishes the proof of (24) and thus (23). □

Lemma 2. *For any real η bounded away from 0 and 1, any positive integers A and B such that $\eta A, \eta B \in \mathbb{Z}$, and any integers Δ and σ such that $A+\Delta$, $\eta A+\sigma$, $B-\Delta$, and $\eta B-\sigma$ are non-negative, it holds that*

$$\begin{aligned}
&\left[\frac{\binom{A+\Delta}{\eta A+\sigma}\binom{B-\Delta}{\eta B-\sigma}}{\binom{A}{\eta A}\binom{B}{\eta B}}\right]^{-1}\\
&\quad=\left(1+O\left(\frac{\sigma^3}{A^2}\right)\right)\left(1+O\left(\frac{\Delta^3}{A^2}\right)\right)\left(1+O\left(\frac{1}{A}\right)\right)\left(1+O\left(\frac{\sigma(\Delta-\sigma)^2}{A^2}\right)\right)\left(1+O\left(\frac{\Delta(\Delta-\sigma)^2}{A^2}\right)\right)\\
&\qquad\times\exp\left(\frac{1}{2}\frac{(\Delta-\sigma)^2}{(1-\eta)A}+\frac{1}{2}\frac{\sigma^2}{\eta A}-\frac{1}{2}\frac{\Delta^2}{A}\right)\\
&\qquad\times\left(1+O\left(\frac{\sigma^3}{B^2}\right)\right)\left(1+O\left(\frac{\Delta^3}{B^2}\right)\right)\left(1+O\left(\frac{1}{B}\right)\right)\left(1+O\left(\frac{\sigma(\Delta-\sigma)^2}{B^2}\right)\right)\left(1+O\left(\frac{\Delta(\Delta-\sigma)^2}{B^2}\right)\right)\\
&\qquad\times\exp\left(\frac{1}{2}\frac{(\Delta-\sigma)^2}{(1-\eta)B}+\frac{1}{2}\frac{\sigma^2}{\eta B}-\frac{1}{2}\frac{\Delta^2}{B}\right).
\end{aligned}$$

Proof. Using Stirling's approximation

$$n!=\left(1+O\left(\frac{1}{n}\right)\right)\frac{n^n}{e^n}\sqrt{2\pi n},$$

we obtain

$$\begin{aligned}
&\left[\frac{\binom{A+\Delta}{\eta A+\sigma}\binom{B-\Delta}{\eta B-\sigma}}{\binom{A}{\eta A}\binom{B}{\eta b}}\right]^{-1}\\
&\quad=\left(1+O\left(\frac{1}{A}\right)\right)\left(1+O\left(\frac{1}{B}\right)\right)\\
&\qquad\times\frac{(\eta A+\sigma)^{\eta A+\sigma}((1-\eta)A+\Delta-\sigma)^{(1-\eta)A+\Delta-\sigma}(\eta B-\sigma)^{\eta B-\sigma}((1-\eta)B-(\Delta-\sigma))^{(1-\eta)B-(\Delta-\sigma)}A^A B^B}{(\eta A)^{\eta A}((1-\eta)A)^{(1-\eta)A}(\eta B)^{\eta B}((1-\eta)B)^{(1-\eta)B}(A+\Delta)^{A+\Delta}(B-\Delta)^{B-\Delta}}\\
&\quad=\left(1+O\left(\frac{1}{A}\right)\right)\left(1+O\left(\frac{1}{B}\right)\right)\left[\frac{\eta A+\sigma}{(1-\eta)A+(\Delta-\sigma)}\frac{(1-\eta)B-(\Delta-\sigma)}{\eta B-\sigma}\right]^{\sigma}\\
&\qquad\times\left[\frac{(1-\eta)A+(\Delta-\sigma)}{A+\Delta}\frac{B-\Delta}{(1-\eta)B-(\Delta-\sigma)}\right]^{\Delta}\left(1+\frac{\sigma}{\eta A}\right)^{\eta A}\left(1+\frac{\Delta-\sigma}{(1-\eta)A}\right)^{(1-\eta)A}\\
&\qquad\times\left(1-\frac{\sigma}{\eta B}\right)^{\eta B}\left(1-\frac{\Delta}{A+\Delta}\right)^{A}\left(1-\frac{\Delta-\sigma}{(1-\eta)B}\right)^{(1-\eta)B}\left(1+\frac{\Delta}{B-\Delta}\right)^{B}.
\end{aligned}$$

Now, using that $\log(1+x)=x-\frac{x^2}{2}+O(x^3)$ for small x,

$$\begin{aligned}
&\left(1+\frac{\sigma}{\eta A}\right)^{\eta A}\left(1+\frac{\Delta-\sigma}{(1-\eta)A}\right)^{(1-\eta)A}\left(1-\frac{\sigma}{\eta B}\right)^{\eta B}\left(1-\frac{\Delta}{A+\Delta}\right)^{A}\left(1-\frac{\Delta-\sigma}{(1-\eta)B}\right)^{(1-\eta)B}\left(1+\frac{\Delta}{B-\Delta}\right)^{B}\\
&\quad=\exp\left(\eta A\left(\frac{\sigma}{\eta A}-\frac{1}{2}\frac{\sigma^2}{\eta^2A^2}+O\left(\frac{\sigma^3}{A^3}\right)\right)\right)\exp\left((1-\eta)A\left(\frac{\Delta-\sigma}{(1-\eta)A}-\frac{1}{2}\frac{(\Delta-\sigma)^2}{(1-\eta)^2A^2}+O\left(\frac{(\Delta-\sigma)^3}{A^3}\right)\right)\right)\\
&\qquad\times\exp\left(-A\left(\frac{\Delta}{A+\Delta}+\frac{1}{2}\frac{\Delta^2}{(A+\Delta)^2}+O\left(\frac{\Delta^3}{(A+\Delta)^3}\right)\right)\right)\exp\left(-\eta B\left(\frac{\sigma}{\eta B}+\frac{1}{2}\frac{\sigma^2}{\eta^2B^2}+O\left(\frac{\sigma^3}{B^3}\right)\right)\right)\\
&\qquad\times\exp\left(-(1-\eta)B\left(\frac{\Delta-\sigma}{(1-\eta)B}+\frac{1}{2}\frac{(\Delta-\sigma)^2}{(1-\eta)^2B^2}+O\left(\frac{(\Delta-\sigma)^3}{B^3}\right)\right)\right)
\end{aligned}$$

$$\times \exp\left(B\left(\frac{\Delta}{B-\Delta}+\frac{1}{2}\frac{\Delta^2}{(B-\Delta)^2}+O\left(\frac{\Delta^3}{(B-\Delta)^3}\right)\right)\right)$$
$$=\left(1+O\left(\frac{\sigma^3}{A^2}\right)\right)\left(1+O\left(\frac{\Delta^3}{A^2}\right)\right)\left(1+O\left(\frac{(\Delta-\sigma)^3}{A^2}\right)\right)\exp\left(-\frac{1}{2}\frac{\sigma^2}{\eta A}-\frac{1}{2}\frac{(\Delta-\sigma)^2}{(1-\eta)A}-\frac{1}{2}\frac{\Delta^2}{A}+\frac{\Delta^2}{A}\right)$$
$$\times\left(1+O\left(\frac{\sigma^3}{B^2}\right)\right)\left(1+O\left(\frac{\Delta^3}{B^2}\right)\right)\left(1+O\left(\frac{(\Delta-\sigma)^3}{B^2}\right)\right)\exp\left(-\frac{1}{2}\frac{\sigma^2}{\eta B}-\frac{1}{2}\frac{(\Delta-\sigma)^2}{(1-\eta)B}-\frac{1}{2}\frac{\Delta^2}{B}+\frac{\Delta^2}{B}\right).$$

And using the simpler $\log(1+x)=x+O(x^2)$ for small x,

$$\left[\frac{\eta A+\sigma}{(1-\eta)A+(\Delta-\sigma)}\frac{(1-\eta)B-(\Delta-\sigma)}{\eta B-\sigma}\right]^{\sigma}$$
$$=\left[1+\frac{(1-\eta)\sigma B-(\Delta-\sigma)\eta B+\sigma(1-\eta)A-\eta(\Delta-\sigma)A}{(1-\eta)\eta AB+(\Delta-\sigma)\eta B-\sigma(1-\eta)A-\sigma(\Delta-\sigma)}\right]^{\sigma}$$
$$=\exp\left(\sigma\left(\frac{\sigma}{\eta A}-\frac{\Delta-\sigma}{(1-\eta)A}+\frac{\sigma}{\eta B}-\frac{\Delta-\sigma}{(1-\eta)B}+O\left(\frac{\sigma^2}{A^2}\right)+O\left(\frac{(\Delta-\sigma)^2}{A^2}\right)\right.\right.$$
$$\left.\left.+O\left(\frac{\sigma^2}{B^2}\right)+O\left(\frac{(\Delta-\sigma)^2}{B^2}\right)\right)\right)$$

and

$$\left[\frac{(1-\eta)A+(\Delta-\sigma)}{A+\Delta}\frac{B-\Delta}{(1-\eta)B-(\Delta-\sigma)}\right]^{\Delta}$$
$$=\left[1+\frac{(\Delta-\sigma)B-(1-\eta)\Delta B+(\Delta-\sigma)A-(1-\eta)\Delta A}{(1-\eta)AB+(1-\eta)\Delta B-(\Delta-\sigma)A-\Delta(\Delta-\sigma)}\right]^{\Delta}$$
$$=\exp\left(\Delta\left(\frac{\Delta-\sigma}{(1-\eta)A}-\frac{\Delta}{A}+\frac{\Delta-\sigma}{(1-\eta)B}-\frac{\Delta}{B}+O\left(\frac{(\Delta-\sigma)^2}{A^2}\right)+O\left(\frac{\Delta^2}{A^2}\right)\right.\right.$$
$$\left.\left.+O\left(\frac{(\Delta-\sigma)^2}{B^2}\right)+O\left(\frac{\Delta^2}{B^2}\right)\right)\right).$$

Combining everything yields the lemma. □

5. Large Hamming distances

The lower bounds established for trace reconstruction thus far have come from pairs of strings with small Hamming distance. A natural question is what can be said about strings with very large Hamming distance. Of course, a pair of strings that differ in all but $O(1)$ indices can be distinguished very easily (in $O(1)$ traces). However, what if we insist on "padding" two strings that always differ, at the beginning and end by some arbitrary strings?

We say that a pair of strings $x, y \in \{0,1\}^n$ *essentially always differ* if there are indices $k_1, k_2 \le n$ such that x and y agree at all indices at most k_1 and at least k_2, and disagree at all indices between k_1 and k_2.

Proposition 8. *Let $x, y \in \{0,1\}^n$ be a pair of strings that essentially always differ. Then x and y can be distinguished in $\exp(C\frac{\log^3 n}{\log\log n})$ samples. Here, $C>0$ is an absolute constant.*

We use the following lemma, found as E7 on page 64 of [3].

Lemma 9. *Let $p(z)=a_nz^n+\cdots+a_1z+a_0$ be a polynomial of degree n with $a_i\in\{\pm1\}$ for each i. Then, $p(z)$ has at most $\frac{C\log^2 n}{\log\log n}$ zeros at 1, i.e., $(z-1)^m$ does not divide $p(z)$ for $m=\lfloor\frac{C\log^2 n}{\log\log n}\rfloor+1$. Here, $C>0$ is an absolute constant.*

With this lemma, we deduce Proposition 8 as follows. We first claim that there is some 0–1 string w of length at most $k:=\lfloor\frac{C\log^2 n}{\log\log n}\rfloor+1$ such that $f(w;x)\neq f(w;y)$ (see Lemma 1 for notation). Indeed, if $f(w;x)=f(w;y)$ for all w of

length at most k, that is, if the so-called "k-decks" of x and y are the same, then by Section 5 of [7], it must be that $\sum_{i=1}^{n} x_i i^m = \sum_{i=1}^{n} y_i i^m$ for all $0 \le m \le k-1$. If we let $p(z) = \sum_{i=1}^{n} [x_i - y_i] z^i$, then it's easy to see that the equalities imply $p(1), p'(1), \ldots, p^{(k-1)}(1) = 0$, which imply $(z-1)^k \mid p(z)$. Now, since x and y essentially always differ, $p(z)$ takes the form $p(z) = \epsilon_{k_1} z^{k_1} + \epsilon_{k_1+1} z^{k_1+1} \cdots + \epsilon_{k_2-1} z^{k_2-1} + \epsilon_{k_2} z^{k_2}$ for some $\epsilon_{k_1}, \ldots, \epsilon_{k_2} \in \{\pm 1\}$. Therefore, by factoring out z^{k_1}, since $k_2 - k_1 \le n$, Lemma 9 implies that $k \le \frac{C \log^2 n}{\log \log n}$, a contradiction. The claim is established.

With this claim, we can distinguish between x and y by simply looking at $f(w; \widetilde{U})$ for traces $\widetilde{U}$; indeed, $\mathbb{E}_x[f(w; \widetilde{U})] = f(w; x)(1-q)^{-|w|}$. Since $|w| \le C \frac{\log^2 n}{\log \log n}$, it holds that $\exp(C' \frac{C \log^2 n}{\log \log n} \log n)$ traces suffice to distinguish between x and y. For details, see the proof of Theorem 14 of [11].

Acknowledgements

I would like to thank Omer Tamuz for introducing me to the wonderful trace reconstruction problem, and for helpful discussions. I would also like to thank Russell Lyons for much helpful feedback on the paper, and for a bijective proof of Lemma 1. Finally, I would like to greatly thank an anonymous referee for several helpful comments, substantially improving the paper's readability and understandability.

References

[1] F. Ban, X. Chen, A. Freilich, R. Servedio and S. Sinha. Beyond trace reconstruction: Population recovery from the deletion channel. *ArXiv e-prints* (2019). Available at 1904.05532. MR3981280 https://doi.org/10.1016/j.acha.2017.09.003

[2] T. Batu, S. Kannan, S. Khanna and A. McGregor. Reconstructing strings from random traces. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 910–918. ACM, New York, 2004. MR2290981

[3] P. Borwein. *Computational Excursions in Analysis and Number Theory*. *CMS Books in Mathematics*. Springer-Verlag, Berlin, 2002. MR1912495 https://doi.org/10.1007/978-0-387-21652-2

[4] M. Cheraghchi, R. Gabrys, O. Milenkovic and J. Ribeiro. Coded trace reconstruction. *ArXiv e-prints* (2019). Available at 1903.09992.

[5] S. Davies, M. Racz and C. Rashtchian. Reconstructing trees from traces. *ArXiv e-prints* (2019). Available at 1902.05101.

[6] A. De, R. O'Donnell and R. A. Servedio. Optimal mean-based algorithms for trace reconstruction. In *STOC'17 – Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing* 1047–1056. ACM, New York, 2017. MR3678250 https://doi.org/10.1145/3055399.3055450

[7] M. Dudik and L. J. Schulman. Reconstruction from subsequences. *J. Combin. Theory Ser. A* **103** (2002) 337–348. MR1996071 https://doi.org/10.1016/S0097-3165(03)00103-1

[8] N. Holden and R. Lyons. Lower bounds for trace reconstruction. *Ann. Appl. Probab.* (2020). MR4108114 https://doi.org/10.1214/19-AAP1506

[9] N. Holden, R. Pemantle and Y. Peres. Subpolynomial trace reconstruction for random strings and arbitrary deletion probability. In *Proceedings of the 31st Conference on Learning Theory* 1799–1840. *PMLR* **75**, 2018.

[10] T. Holenstein, M. Mitzenmacher, R. Panigrahy and U. Wieder. Trace reconstruction with constant deletion probability and related results. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 389–398. ACM, New York, 2008. MR2487606

[11] A. Krishnamurthy, A. Mazumdar, A. McGregor and S. Pal, 2019. Trace reconstruction: Generalized and parameterized. *ArXiv e-prints*. Available at 1904.09618. MR4007907

[12] F. Nazarov and Y. Peres. Trace reconstruction with $\exp(O(n^{1/3}))$ samples. In *STOC'17 – Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing* 1042–1046. ACM, New York, 2017. MR3678249