Lecture 2: From DNA Storage to Combinatorial Reconstruction:
Sequence, Matrix, and Hypermatrix k-Deck Problems

Lecturer: Professor Xiande Zhang

Notes written by Danni Li with assistance from ChatGPT

August 2026

Contents

1 DNA Storage and Random Trace Reconstruction

DNA storage gives a natural motivation for reconstruction problems. Digital data are encoded as strings over the alphabet $\{A, C, G, T\}$. During synthesis, storage, and sequencing, the stored strand may suffer insertion, deletion, and substitution errors. In practice, one often obtains several noisy reads and tries to recover the original strand [1].

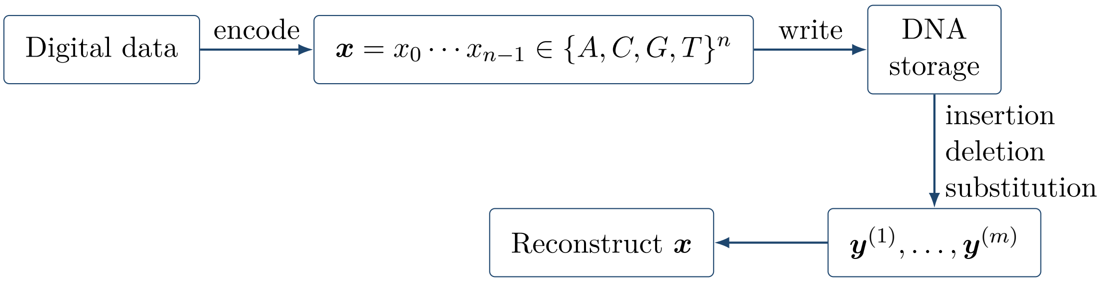

Figure 1: A schematic view of DNA storage and reconstruction.

A simple model keeps only deletion errors. Let $\boldsymbol{x} = x_0 \cdots x_{n-1} \in \Sigma^n$. A *trace* is a random subsequence obtained by retaining each position independently with a fixed probability. The main question in trace reconstruction is to determine the minimum number of traces needed to reconstruct the original sequence uniquely with high probability.

If a trace is conditioned to have length k, then every k-subset of the positions is equally likely. This leads to a deterministic object that records all length-k subsequences with their multiplicities. For $\boldsymbol{w} \in \Sigma^k$, write

$$N(\boldsymbol{w}; \boldsymbol{x}) = \#\{0 \le i_1 < \cdots < i_k \le n-1 : (x_{i_1}, \ldots, x_{i_k}) = \boldsymbol{w}\}.$$

Thus $N(\boldsymbol{w}; \boldsymbol{x})$ is the number of occurrences of $\boldsymbol{w}$ as a subsequence of $\boldsymbol{x}$.

2 Sequence Reconstruction from k-Decks

2.1 Definitions and basic properties

Definition 2.1 (k-deck)**.** For $\boldsymbol{x} = x_0 \cdots x_{n-1} \in \Sigma^n$, its *$k$-deck* is the multiset

$$\mathcal{D}_k(\boldsymbol{x}) = \{(x_{i_1}, \ldots, x_{i_k}) : 0 \le i_1 < \cdots < i_k \le n-1\},$$

where repeated subsequences are counted with multiplicity. Two words $\boldsymbol{x}$ and $\boldsymbol{y}$ are *k-equivalent*, written $\boldsymbol{x} \equiv_k \boldsymbol{y}$, if $\mathcal{D}_k(\boldsymbol{x}) = \mathcal{D}_k(\boldsymbol{y})$. A word $\boldsymbol{x}$ is *k-reconstructible* if no different word has the same k-deck.

Let $\boldsymbol{x}, \boldsymbol{y} \in \Sigma^n$. The k-deck has the following basic properties.

(1) We have $\mathcal{D}_k(\boldsymbol{x}) = \mathcal{D}_k(\boldsymbol{y})$ if and only if $N(\boldsymbol{w}; \boldsymbol{x}) = N(\boldsymbol{w}; \boldsymbol{y})$ for every $\boldsymbol{w} \in \Sigma^k$.

(2) If $\mathcal{D}_k(\boldsymbol{x}) = \mathcal{D}_k(\boldsymbol{y})$, then $\mathcal{D}_\ell(\boldsymbol{x}) = \mathcal{D}_\ell(\boldsymbol{y})$ for every $\ell < k$.

(3) Let $\boldsymbol{u} \in \Sigma^\ell$ with $\ell < k$. If $N(\boldsymbol{w}; \boldsymbol{x}) = N(\boldsymbol{w}; \boldsymbol{y})$ for every $\boldsymbol{w} \in \Sigma^k$, then $N(\boldsymbol{u}; \boldsymbol{x}) = N(\boldsymbol{u}; \boldsymbol{y})$.

(4) For every $\boldsymbol{u} \in \Sigma^\ell$ with $\ell < k$,

$$\sum_{\boldsymbol{w} \in \Sigma^k} N(\boldsymbol{u}; \boldsymbol{w}) N(\boldsymbol{w}; \boldsymbol{x}) = \binom{n-\ell}{k-\ell} N(\boldsymbol{u}; \boldsymbol{x}).$$

(5) If a word is k-reconstructible, then it is also $(k+1)$-reconstructible.

For binary words, define the reconstruction threshold

$$\begin{aligned} f(n) &= \min\{k : \text{every word in } \{0,1\}^n \text{ is reconstructible from its } k\text{-deck}\} \\ &= \min\{k : \mathcal{D}_k(\boldsymbol{x}) = \mathcal{D}_k(\boldsymbol{y}) \text{ implies } \boldsymbol{x} = \boldsymbol{y} \text{ for all } \boldsymbol{x}, \boldsymbol{y} \in \{0,1\}^n\}. \end{aligned}$$

The following bounds describe the main development discussed in the lecture.

Result	**Bound**
Manvel et al. [2]	$\log_2 n \lesssim f(n) \le \lfloor n/2 \rfloor$.
Krasikov–Roditty [3]	$f(n) = O(\sqrt{n})$.
Dudík–Schulman [6]	$f(n) = \exp(\Omega(\sqrt{\log n}))$.

2.2 The $\lfloor n/2 \rfloor$ Upper Bound

Theorem 2.2 (Manvel et al. [2])**.** *For $n > 7$, $f(n) \le \lfloor \frac{n}{2} \rfloor$.*

Proof. We first prove the main part of the argument. Suppose that $\boldsymbol{x} \in \{0,1\}^n$ contains exactly $k-1$ zeros. Then the zeros divide $\boldsymbol{x}$ uniquely into k blocks of ones, so

$$\boldsymbol{x} = 1^{a_0} 0\, 1^{a_1} 0 \cdots 0\, 1^{a_{k-2}} 0\, 1^{a_{k-1}},$$

where $a_0, \ldots, a_{k-1}$ are nonnegative integers. Empty blocks are allowed.

For each $j \in \{0, \ldots, k-1\}$, consider the length-k word $0^j 1\, 0^{k-1-j}$. Since $\boldsymbol{x}$ contains exactly $k-1$ zeros, every occurrence of this word as a subsequence of $\boldsymbol{x}$ must use all zeros of $\boldsymbol{x}$. Therefore its unique 1 must be chosen from the block 1^{a_j}, and hence

$$N(0^j 1\, 0^{k-1-j}; \boldsymbol{x}) = a_j.$$

Thus the k-deck gives all the numbers $a_0, \ldots, a_{k-1}$, which uniquely specify $\boldsymbol{x}$. Hence $\boldsymbol{x}$ is k-reconstructible.

If $\boldsymbol{x}$ contains fewer than $k-1$ zeros, then by the basic properties of decks, the k-deck gives the shorter deck whose length is one more than the number of zeros. The same argument applies. By symmetry, the same conclusion holds if the number of ones is at most $k-1$.

Now suppose $n < 2k$. Then at least one of the symbols 0 and 1 occurs at most $k-1$ times, so every binary word of length n is k-reconstructible. Taking $k = \lfloor \frac{n}{2} \rfloor + 1$ already gives $f(n) \le \lfloor \frac{n}{2} \rfloor + 1$.

A further treatment of the remaining boundary case improves this by one, giving $f(n) \le \lfloor \frac{n}{2} \rfloor$ for $n > 7$. □

2.3 The Logarithmic Lower Bound

Theorem 2.3 (Manvel et al. [2])**.** *For every $n \ge 2$, $f(n) \ge \lfloor \log_2 n \rfloor + 1$.*

Proof. We first construct, for every $k \ge 1$, two distinct binary words of length 2^k with the same k-deck.

Let $\boldsymbol{x}_1 = 01$ and $\boldsymbol{y}_1 = 10$. For $k \ge 1$, define recursively

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k \boldsymbol{y}_k, \qquad \boldsymbol{y}_{k+1} = \boldsymbol{y}_k \boldsymbol{x}_k,$$

where juxtaposition denotes concatenation. Clearly, $|\boldsymbol{x}_k| = |\boldsymbol{y}_k| = 2^k$ and $\boldsymbol{x}_k \ne \boldsymbol{y}_k$ for every k.

We claim that $\mathcal{D}_k(\boldsymbol{x}_k) = \mathcal{D}_k(\boldsymbol{y}_k)$ for every $k \ge 1$. We prove this by induction on k. The case $k = 1$ is immediate. Suppose that $\mathcal{D}_k(\boldsymbol{x}_k) = \mathcal{D}_k(\boldsymbol{y}_k)$. By the basic properties of decks, $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$

have the same ℓ-deck for every $\ell \leq k$. Fix a word $\boldsymbol{w} \in \{0,1\}^{k+1}$. An occurrence of $\boldsymbol{w}$ in $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k \boldsymbol{y}_k$ is obtained by choosing a prefix $\boldsymbol{u}$ of $\boldsymbol{w}$ from $\boldsymbol{x}_k$ and the remaining suffix $\boldsymbol{v}$ from $\boldsymbol{y}_k$. Therefore

$$N(\boldsymbol{w}; \boldsymbol{x}_{k+1}) = \sum_{\boldsymbol{w}=\boldsymbol{u}\boldsymbol{v}} N(\boldsymbol{u}; \boldsymbol{x}_k) N(\boldsymbol{v}; \boldsymbol{y}_k),$$

where the sum is over all decompositions of $\boldsymbol{w}$ into a prefix $\boldsymbol{u}$ and a suffix $\boldsymbol{v}$, allowing either of them to be empty. Similarly,

$$N(\boldsymbol{w}; \boldsymbol{y}_{k+1}) = \sum_{\boldsymbol{w}=\boldsymbol{u}\boldsymbol{v}} N(\boldsymbol{u}; \boldsymbol{y}_k) N(\boldsymbol{v}; \boldsymbol{x}_k).$$

For every nontrivial decomposition $\boldsymbol{w} = \boldsymbol{u}\boldsymbol{v}$, both $|\boldsymbol{u}|$ and $|\boldsymbol{v}|$ are at most k. Hence

$$N(\boldsymbol{u}; \boldsymbol{x}_k) = N(\boldsymbol{u}; \boldsymbol{y}_k) \quad \text{and} \quad N(\boldsymbol{v}; \boldsymbol{x}_k) = N(\boldsymbol{v}; \boldsymbol{y}_k).$$

The two trivial decompositions contribute $N(\boldsymbol{w}; \boldsymbol{x}_k) + N(\boldsymbol{w}; \boldsymbol{y}_k)$ to both sums. Thus $N(\boldsymbol{w}; \boldsymbol{x}_{k+1}) = N(\boldsymbol{w}; \boldsymbol{y}_{k+1})$ for every $\boldsymbol{w} \in \{0,1\}^{k+1}$. Hence

$$\mathcal{D}_{k+1}(\boldsymbol{x}_{k+1}) = \mathcal{D}_{k+1}(\boldsymbol{y}_{k+1}),$$

which proves the claim.

Now let $k = \lfloor \log_2 n \rfloor$, so $2^k \leq n$. If $n = 2^k$, the distinct words $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ already have the same k-deck. If $n > 2^k$, append the same binary word $\boldsymbol{z}$ of length $n - 2^k$ to both words. Since $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ have the same ℓ-deck for every $\ell \leq k$, the same decomposition argument shows that $\boldsymbol{x}_k \boldsymbol{z}$ and $\boldsymbol{y}_k \boldsymbol{z}$ have the same k-deck.

Thus there are two distinct words in $\{0,1\}^n$ with the same k-deck. Therefore not every word of length n is k-reconstructible, and hence $f(n) > k = \lfloor \log_2 n \rfloor$. This gives the desired bound. $\square$

3 The Krasikov–Roditty Upper Bound

We now improve the linear upper bound to an $O(\sqrt{n})$ upper bound. The main idea is to convert equality of k-decks into algebraic identities.

Theorem 3.1 (Krasikov–Roditty [3])**.** *Every binary word of length n is reconstructible from its k-deck whenever $k \geq \left\lfloor \sqrt{\frac{16n}{7}} \right\rfloor + 5$. In particular, $f(n) = O(\sqrt{n})$.*

We prove the theorem by contradiction. Suppose that two distinct words $\boldsymbol{x}, \boldsymbol{y} \in \{0,1\}^n$ have the same k-deck, and write

$$\boldsymbol{x} = (x_0, \dots, x_{n-1}), \qquad \boldsymbol{y} = (y_0, \dots, y_{n-1}),$$

and $\delta_i = x_i - y_i$ for $i \in \{0, \dots, n-1\}$.

The proof consists of two steps.

(1) We use simple statistics of the k-deck to show that

$$\sum_{i=0}^{n-1} p(i)\delta_i = 0$$

for every real polynomial p of degree less than k.

(2) We construct a polynomial p of degree less than k that detects the first position at which $\boldsymbol{x}$ and $\boldsymbol{y}$ differ. For this polynomial,

$$\sum_{i=0}^{n-1} p(i)\delta_i \neq 0,$$

which gives a contradiction.

The first step follows directly from the k-deck. The second step uses a polynomial construction of Borwein, Erdélyi, and Kós.

3.1 From Deck Equality to Vanishing Moments

Let $\boldsymbol{x} = (x_0, \ldots, x_{n-1}) \in \{0,1\}^n$. For each $j \in \{0, \ldots, k-1\}$, let $S_j(\boldsymbol{x})$ be the total number of occurrences of the symbol 1 in the jth coordinate among all members of $\mathcal{D}_k(\boldsymbol{x})$, counted with multiplicity.

Fix $i \in \{0, \ldots, n-1\}$ with $x_i = 1$. If x_i appears in the jth coordinate of a length-k subsequence, then exactly j positions must be chosen from $\{0, \ldots, i-1\}$, and exactly $k-j-1$ positions must be chosen from $\{i+1, \ldots, n-1\}$. Hence the position i contributes

$$\binom{i}{j}\binom{n-i-1}{k-j-1}$$

to $S_j(\boldsymbol{x})$. Summing over all positions, we obtain

$$S_j(\boldsymbol{x}) = \sum_{i=0}^{n-1} \binom{i}{j}\binom{n-i-1}{k-j-1} x_i.$$

Now suppose that $\mathcal{D}_k(\boldsymbol{x}) = \mathcal{D}_k(\boldsymbol{y})$. Then $S_j(\boldsymbol{x}) = S_j(\boldsymbol{y})$ for every $j \in \{0, \ldots, k-1\}$. Therefore

$$\sum_{i=0}^{n-1} \binom{i}{j}\binom{n-i-1}{k-j-1} \delta_i = 0 \qquad \text{for every } j \in \{0, \ldots, k-1\}.$$

For $j \in \{0, \ldots, k-1\}$, define

$$F_j(t) = \binom{t}{j}\binom{n-t-1}{k-j-1},$$

where the binomial coefficients are viewed as polynomials in t. Since

$$\deg \binom{t}{j} = j \qquad \text{and} \qquad \deg \binom{n-t-1}{k-j-1} = k-j-1,$$

we have $\deg F_j \le k-1$.

Lemma 3.2. *The polynomials $F_0, \ldots, F_{k-1}$ form a basis of the real polynomials of degree at most $k-1$.*

Proof. The vector space of real polynomials of degree at most $k-1$ has dimension k, so it is enough to prove that $F_0, \ldots, F_{k-1}$ are linearly independent.

Suppose

$$\sum_{j=0}^{k-1} a_j F_j(t) = 0$$

and not all a_j are zero. Let j_0 be the smallest index such that $a_{j_0} \neq 0$. Evaluating at $t = j_0$, all terms with $j < j_0$ vanish because their coefficients are zero, while all terms with $j > j_0$ vanish because $\binom{j_0}{j} = 0$. Thus $a_{j_0} F_{j_0}(j_0) = 0$. However,

$$F_{j_0}(j_0) = \binom{n-j_0-1}{k-j_0-1} > 0,$$

a contradiction. Hence the polynomials are linearly independent. □

The preceding identities therefore hold not only for the polynomials F_j, but for every polynomial of degree less than k.

Proposition 3.3 (Moment-vanishing principle [3])**.** *If $\boldsymbol{x}, \boldsymbol{y} \in \{0,1\}^n$ satisfy $\mathcal{D}_k(\boldsymbol{x}) = \mathcal{D}_k(\boldsymbol{y})$, then*

$$\sum_{i=0}^{n-1} p(i)\delta_i = 0$$

for every real polynomial p with $\deg p < k$.

Proof. For each $j \in \{0, \ldots, k-1\}$, equality of the k-decks gives

$$\sum_{i=0}^{n-1} F_j(i)\delta_i = 0.$$

By Lemma 3.2, every polynomial p of degree less than k can be written as $p = \sum_{j=0}^{k-1} c_j F_j$ for some real numbers $c_0, \ldots, c_{k-1}$. Therefore

$$\sum_{i=0}^{n-1} p(i)\delta_i = \sum_{j=0}^{k-1} c_j \sum_{i=0}^{n-1} F_j(i)\delta_i = 0.$$

□

3.2 A Polynomial Detecting the First Difference

The moment identities alone do not yet give a contradiction. We need a polynomial of degree less than k whose value at the first position where $\boldsymbol{x}$ and $\boldsymbol{y}$ differ dominates its values at all later positions.

The required polynomial is given by the following result.

Theorem 3.4 (Borwein–Erdélyi–Kós [4])**.** *For every integer $N \geq 1$, there is a real polynomial q such that $|q(0)| > \sum_{j=1}^{N} |q(j)|$ and $\deg q \leq \left\lfloor \sqrt{\frac{16N}{7}} \right\rfloor + 4$.*

Proof of the Krasikov–Roditty theorem. Suppose, for a contradiction, that $\boldsymbol{x} \neq \boldsymbol{y}$ and $\mathcal{D}_k(\boldsymbol{x}) = \mathcal{D}_k(\boldsymbol{y})$. Let $a = \min\{i : x_i \neq y_i\}$. Thus $\delta_i = 0$ for every $i < a$, while $|\delta_a| = 1$. If $a = n-1$, then taking the constant polynomial $p(t) = 1$ in Proposition 3.3 gives $0 = \sum_{i=0}^{n-1} \delta_i = \delta_a$, a contradiction. We may therefore assume that $a < n-1$ and set $N = n - 1 - a \geq 1$.

Apply Theorem 3.4 with this value of N, and let q be the resulting polynomial. Define $p(t) = q(t-a)$. Since $N \leq n$, we have

$$\deg p = \deg q \leq \left\lfloor \sqrt{\frac{16N}{7}} \right\rfloor + 4 \leq \left\lfloor \sqrt{\frac{16n}{7}} \right\rfloor + 4 < k.$$

Hence Proposition 3.3 gives $\sum_{i=0}^{n-1} p(i)\delta_i = 0$.

On the other hand, since $\delta_i = 0$ for $i < a$ and $|\delta_i| \leq 1$ for every i, we have

$$\left|\sum_{i=0}^{n-1} p(i)\delta_i\right| = \left|\delta_a q(0) + \sum_{i=a+1}^{n-1} \delta_i q(i-a)\right| \geq |q(0)| - \sum_{j=1}^{N} |q(j)| > 0,$$

contradicting the moment-vanishing principle.

Therefore two distinct binary words cannot have the same k-deck when $k \geq \left\lfloor \sqrt{\frac{16n}{7}} \right\rfloor + 5$. This proves the theorem. □

Remark 3.5 (q-ary sequences)**.** The reconstruction threshold does not depend on the size of the alphabet. More precisely, for every $q \geq 2$, every word in $[q]^n$ is k-reconstructible if and only if every binary word in $\{0,1\}^n$ is k-reconstructible.

Remark 3.6 (Matrices and hypermatrices)**.** The same reconstruction problem can be considered for matrices and higher-dimensional arrays. For an $n \times n$ matrix, the k-deck is the multiset of all $k \times k$ submatrices obtained by choosing k rows and k columns while preserving their orders.

Kós, Ligeti, and Sziklai [5] proved that, for all sufficiently large n, every $n \times n$ matrix can be reconstructed from its k-deck when $k \geq 38n^{2/3}$. More generally, Zhong and Zhang [7] obtained an $O(n^{d/(d+1)})$ upper bound for fixed d-dimensional hypermatrices. For $d \geq 3$, the upper bound has been improved to $O\left(n^{2/3}\right)$ recently by the group of Zhang.

4 The Dudík–Schulman Lower Bound

We now turn to the lower bound. The logarithmic construction in Section 2 produces explicit pairs of k-equivalent words, but their length grows only exponentially in k. Dudík and Schulman gave a much more efficient recursive construction.

Theorem 4.1 (Dudík–Schulman [6])**.** *For binary sequence reconstruction,*

$$f(n) \geq 3^{(\sqrt{2/3}-o(1))\sqrt{\log_3 n}} = \exp(\Omega(\sqrt{\log n})).$$

The main idea is to construct pairs of distinct words with the same k-deck recursively. Define

$$S(k) := \min\left\{m : \text{there exist distinct } \boldsymbol{x}, \boldsymbol{y} \in \{0,1\}^m \text{ with } \mathcal{D}_k(\boldsymbol{x}) = \mathcal{D}_k(\boldsymbol{y})\right\}.$$

If $S(k) \leq n$, then such a pair can be extended to length n by appending the same symbols to both words. Hence $f(n) > k$. Therefore, to obtain a lower bound on $f(n)$, it is enough to find a good upper bound on $S(k)$.

Dudík and Schulman achieve this through the recurrence

$$S(3k+R) \leq S(k)k^{3+o(1)} \qquad (R \in \{0,1,2\}).$$

We first explain the construction behind this recurrence.

4.1 The Recursive Construction

Take distinct $\boldsymbol{x}, \boldsymbol{y} \in \{0,1\}^{S(k)}$ with $\mathcal{D}_k(\boldsymbol{x}) = \mathcal{D}_k(\boldsymbol{y})$. We regard $\boldsymbol{x}$ and $\boldsymbol{y}$ as two large symbols X and Y. For a template $\boldsymbol{p} = p_1 \cdots p_m \in \{X, Y\}^m$, define $h(X) := \boldsymbol{x}$, $h(Y) := \boldsymbol{y}$, and

$$h(\boldsymbol{p}) := h(p_1) \cdots h(p_m).$$

Thus $h(\boldsymbol{p})$ is a binary word of length $mS(k)$.

Our goal is to find two short distinct templates $\boldsymbol{p}, \boldsymbol{q}$ such that $\mathcal{D}_{3k+R}(h(\boldsymbol{p})) = \mathcal{D}_{3k+R}(h(\boldsymbol{q}))$. This would give

$$S(3k+R) \leq mS(k).$$

We therefore need to understand what information about a template can be detected by a subsequence of length $3k+R$.

Since $\boldsymbol{x}$ and $\boldsymbol{y}$ have the same decks of every length at most k, a substituted block cannot be distinguished as X or Y if the subsequence uses at most k symbols from that block. We represent such an undistinguished block by a wildcard symbol J, which may match either X or Y.

For $r \in \{1,2\}$ and $t \geq r$, let $U_r(t)$ be the family of patterns over $\{X, Y, J\}$ of length at most t with exactly r symbols from $\{X, Y\}$. For a pattern $\boldsymbol{u}$ and a template $\boldsymbol{p}$, let $N_U(\boldsymbol{u}; \boldsymbol{p})$ be the number of order-preserving embeddings of $\boldsymbol{u}$ into $\boldsymbol{p}$, where J may match either X or Y.

For example, if $\boldsymbol{p} = XYY$, then $N_U(XJ; \boldsymbol{p}) = 2$ and $N_U(JY; \boldsymbol{p}) = 3$.

For $k_1 \geq k_2$, two templates $\boldsymbol{p}$ and $\boldsymbol{q}$ are called $U(k_1, k_2)$-equivalent if

$$N_U(\boldsymbol{u}; \boldsymbol{p}) = N_U(\boldsymbol{u}; \boldsymbol{q}) \qquad \text{for every } \boldsymbol{u} \in U_1(k_1) \cup U_2(k_2).$$

Let $S_U(k_1, k_2)$ be the minimum common length of two distinct $U(k_1, k_2)$-equivalent templates.

The next lemma shows why the parameters $2k+R$ and $k+R$ are the relevant ones.

Lemma 4.2 (Dudík–Schulman [6]). *For every $R \in \{0,1,2\}$,*

$$S(3k+R) \le S(k)\, S_U(2k+R, k+R).$$

Proof. Choose distinct $U(2k+R,k+R)$-equivalent templates $\boldsymbol{p},\boldsymbol{q} \in \{X,Y\}^m$, where $m = S_U(2k+R, k+R)$. We show that $h(\boldsymbol{p})$ and $h(\boldsymbol{q})$ have the same $(3k+R)$-deck.

Fix $\boldsymbol{w} \in \{0,1\}^{3k+R}$. Consider an occurrence of $\boldsymbol{w}$ in $h(\boldsymbol{p})$ that visits ℓ substituted blocks. It uniquely splits $\boldsymbol{w}$ into nonempty consecutive pieces

$$\boldsymbol{w} = \boldsymbol{w}_1 \cdots \boldsymbol{w}_\ell,$$

where $\boldsymbol{w}_s$ is chosen from the corresponding substituted block.

Call $\boldsymbol{w}_s$ *long* if $|\boldsymbol{w}_s| > k$. If $|\boldsymbol{w}_s| \le k$, then

$$N(\boldsymbol{w}_s;\boldsymbol{x}) = N(\boldsymbol{w}_s;\boldsymbol{y}),$$

so this piece cannot distinguish whether the corresponding template symbol is X or Y. Such a position therefore behaves as a wildcard J.

There can be at most two long pieces, since three of them would use at least $3(k+1) > 3k+R$ symbols.

If there is no long piece, then the contribution of this decomposition to $N(\boldsymbol{w};h(\boldsymbol{p}))$ is independent of the template.

If there is exactly one long piece, then the template matters at only one visited position. Moreover, $3k+R \ge (k+1)+(\ell-1)$, so $\ell \le 2k+R$. Hence its contribution is determined by a statistic in $U_1(2k+R)$.

If there are exactly two long pieces, then the template matters at only two visited positions. In this case, $3k+R \ge 2(k+1)+(\ell-2)$, so $\ell \le k+R$. Hence its contribution is determined by a statistic in $U_2(k+R)$.

Therefore every possible contribution to $N(\boldsymbol{w};h(\boldsymbol{p}))$ is determined by the $U(2k+R,k+R)$-statistics of $\boldsymbol{p}$. Since $\boldsymbol{p}$ and $\boldsymbol{q}$ are $U(2k+R,k+R)$-equivalent, the same contributions occur for $h(\boldsymbol{q})$. Thus

$$N(\boldsymbol{w};h(\boldsymbol{p})) = N(\boldsymbol{w};h(\boldsymbol{q}))$$

for every $\boldsymbol{w} \in \{0,1\}^{3k+R}$.

Hence $h(\boldsymbol{p}) \equiv_{3k+R} h(\boldsymbol{q})$. Since $\boldsymbol{p} \ne \boldsymbol{q}$ and $\boldsymbol{x} \ne \boldsymbol{y}$, the two substituted words are distinct, and their common length is $mS(k)$. Therefore

$$S(3k+R) \le S(k) S_U(2k+R, k+R).$$

□

4.2 Finding Indistinguishable Templates

It remains to bound $S_U(k_1,k_2)$. The idea is to record only the one- and two-position statistics that are visible to the substitution argument.

For a template $\boldsymbol{p} \in \{X,Y\}^m$, first record, for each $j \in [k_1]$, the number of length-k_1 subsequences whose jth symbol is X. These k_1 numbers determine the corresponding one-symbol statistics with Y as well.

We first explain why these length-k_1 statistics determine all shorter one-symbol statistics. Let $\boldsymbol{u} \in U_1(k_1)$ have length $\ell < k_1$, and suppose that its unique non-wildcard symbol $\sigma \in \{X,Y\}$ occurs in position a. Let $A_j^\sigma(\boldsymbol{p})$ denote the number of length-k_1 subsequences of $\boldsymbol{p}$ whose jth symbol is σ. Double counting pairs consisting of an occurrence of $\boldsymbol{u}$ and a length-k_1 set of template positions containing it gives

$$\binom{m-\ell}{k_1-\ell} N_U(\boldsymbol{u};\boldsymbol{p}) = \sum_{j=1}^{k_1} \binom{j-1}{a-1}\binom{k_1-j}{\ell-a} A_j^\sigma(\boldsymbol{p}).$$

Hence the length-k_1 one-position counts determine all $U_1(k_1)$-statistics. In particular, they determine the one-position statistics of length k_2.

Next, for every $1 \le r < s \le k_2$, record the number of length-k_2 subsequences whose rth and sth symbols are both X. For two fixed positions there are four possibilities XX, XY, YX, YY. Their one-position marginals at length k_2 are already known from the preceding paragraph, so the XX count determines the other three counts as well. Thus all two-position statistics of length k_2 are determined.

These length-k_2 statistics also determine all shorter two-position statistics. Indeed, let $\boldsymbol{u} \in U_2(k_2)$ have length $\ell < k_2$, with its two non-wildcard symbols $\sigma, \tau \in \{X, Y\}$ in positions $a < b$. Let $B^{\sigma,\tau}_{r,s}(\boldsymbol{p})$ denote the number of length-k_2 subsequences whose rth and sth symbols are σ and τ, respectively. Double counting as above gives

$$\binom{m-\ell}{k_2-\ell} N_U(\boldsymbol{u};\boldsymbol{p}) = \sum_{1\le r<s\le k_2} \binom{r-1}{a-1}\binom{s-r-1}{b-a-1} \cdot \binom{k_2-s}{\ell-b} B^{\sigma,\tau}_{r,s}(\boldsymbol{p}).$$

Hence the recorded one- and two-position data determine all $U(k_1, k_2)$-statistics. We call these data the *signature* of $\boldsymbol{p}$.

Each one-position count has at most $\binom{m}{k_1} + 1$ possible values, and each two-position count has at most $\binom{m}{k_2} + 1$ possible values. Hence the number of possible signatures is at most

$$\left(\binom{m}{k_1} + 1\right)^{k_1} \left(\binom{m}{k_2} + 1\right)^{\binom{k_2}{2}} \le m^K, \qquad K = k_1^2 + k_2\binom{k_2}{2}.$$

On the other hand, there are 2^m templates of length m. If $2^m > m^K$, then by the pigeonhole principle two distinct templates have the same signature, and hence are $U(k_1, k_2)$-equivalent. It follows that

$$S_U(k_1, k_2) \le (1 + o(1))K \log_2 K.$$

We now take $k_1 = 2k + R$ and $k_2 = k + R$. Then

$$K = (2k + R)^2 + (k + R)\binom{k+R}{2} = \frac{1}{2}k^3 + O(k^2),$$

and hence $S_U(2k + R, k + R) \le k^{3+o(1)}$. Together with Lemma 4.2, this gives

$$S(3k + R) \le S(k)k^{3+o(1)}.$$

4.3 Iterating the Recurrence

We have proved that, for $R \in \{0, 1, 2\}$,

$$S(3r + R) \le S(r)r^{3+o(1)}.$$

We now iterate this recurrence.

Fix K, and define

$$K_0 = K, \qquad K_{j+1} = \left\lfloor \frac{K_j}{3} \right\rfloor.$$

We stop when K_t is bounded. For every $j < t$, there is some $R_j \in \{0, 1, 2\}$ such that

$$K_j = 3K_{j+1} + R_j.$$

Applying the recurrence with $r = K_{j+1}$ gives

$$\log_3 S(K_j) \le \log_3 S(K_{j+1}) + (3 + o(1)) \log_3 K_{j+1}.$$

Iterating from $K_0 = K$ down to K_t yields

$$\log_3 S(K) \le O(1) + (3 + o(1)) \sum_{j=0}^{t-1} \log_3 K_{j+1}.$$

Since at each step the parameter is reduced by a factor of approximately 3,

$$K_j = \frac{K}{3^j} + O(1), \qquad t = \log_3 K + O(1).$$

Hence

$$\begin{aligned}\sum_{j=0}^{t-1} \log_3 K_{j+1} &= \sum_{j=0}^{t-1} (\log_3 K - (j+1) + O(1)) \\ &= \frac{1}{2}(\log_3 K)^2 + O(\log K).\end{aligned}$$

The accumulated error from the $o(1)$ terms is $o((\log K)^2)$. Therefore

$$\log_3 S(K) \le \left(\frac{3}{2} + o(1)\right) (\log_3 K)^2,$$

or equivalently,

$$S(K) \le 3^{(3/2+o(1))(\log_3 K)^2}.$$

Replacing K again by k, we obtain

$$S(k) \le 3^{(3/2+o(1))(\log_3 k)^2}.$$

Finally, taking

$$k = 3^{(\sqrt{2/3}-o(1))\sqrt{\log_3 n}}$$

gives $S(k) \le n$, and hence $f(n) > k$. Therefore

$$f(n) \ge 3^{(\sqrt{2/3}-o(1))\sqrt{\log_3 n}} = \exp(\Omega(\sqrt{\log n})).$$

4.4 Iterating the Recurrence

We now iterate the recurrence. For powers of 3, let $k = 3^t$. Taking $R = 0$ repeatedly gives

$$\log_3 S(3^{t+1}) \le \log_3 S(3^t) + (3 + o(1))t.$$

Summing over the previous steps,

$$\log_3 S(3^t) \le (3 + o(1)) \sum_{j=1}^{t-1} j = \left(\frac{3}{2} + o(1)\right) t^2.$$

For a general k, set $k_0 = k$ and $k_{j+1} = \lfloor k_j/3 \rfloor$, and stop when k_t is bounded. For each $j < t$, write $k_j = 3k_{j+1} + R_j$ with $R_j \in \{0, 1, 2\}$. Applying the recurrence at each step gives

$$\log_3 S(k) \le O(1) + \sum_{j=0}^{t-1} (3 + o(1)) \log_3 k_{j+1}.$$

Since $k_j = k/3^j + O(1)$ and $t = \log_3 k + O(1)$,

$$\sum_{j=0}^{t-1} \log_3 k_{j+1} = \frac{1}{2}(\log_3 k)^2 + O(\log k).$$

The accumulated error from the $o(1)$ terms is $o((\log k)^2)$. Therefore, for general k,

$$S(k) \le 3^{(3/2+o(1))(\log_3 k)^2}.$$

Finally, choose

$$k = 3^{(\sqrt{2/3}-o(1))\sqrt{\log_3 n}}.$$

Then the preceding bound gives $S(k) \le n$, and therefore $f(n) > k$. We obtain the Dudík–Schulman result $f(n) \ge 3^{(\sqrt{2/3}-o(1))\sqrt{\log_3 n}} = \exp(\Omega(\sqrt{\log n}))$. This finishes the proof of Lemma 4.2.

We remark that for $d = 1$, the best known bounds remain $\exp(\Omega(\sqrt{\log n})) \le f(n) \le O(\sqrt{n})$, where the upper bound is due to Krasikov–Roditty [3] and the lower bound is due to Dudík–Schulman [6].

Remark 4.3 (Graph reconstruction)**.** Another classical reconstruction problem arises in graph theory. For a graph G on n vertices, its *deck* is the multiset $\{G - v : v \in V(G)\}$, where each card $G - v$ is considered up to isomorphism. The famous *Graph Reconstruction Conjecture* of Kelly and Ulam asserts that every graph on at least three vertices is uniquely determined, up to isomorphism, by its deck.

This problem is similar in spirit to sequence reconstruction: in both cases one tries to recover a large object from the multiset of its smaller substructures. The main difference is that a sequence carries a fixed linear order, while the cards in a graph deck are unlabeled graphs and do not retain the identities of the deleted vertices.

References

[1] S. R. Srinivasavaradhan, S. Gopi, H. D. Pfister, and S. Yekhanin, Trellis BMA: coded trace reconstruction on IDS channels for DNA storage, in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021.

[2] B. Manvel, A. Meyerowitz, A. Schwenk, K. Smith, and P. Stockmeyer, Reconstruction of sequences, *Discrete Math.* **94** (1991), 209–219.

[3] I. Krasikov and Y. Roditty, On a reconstruction problem for sequences, *J. Combin. Theory Ser. A* **77** (1997), 344–348.

[4] P. Borwein, T. Erdélyi, and G. Kós, Littlewood-type problems on $[0, 1]$, *Proc. London Math. Soc.* **79** (1999), 22–46.

[5] G. Kós, P. Ligeti, and P. Sziklai, Reconstruction of matrices from submatrices, *Math. Comp.* **78** (2009), 1733–1747.

[6] M. Dudík and L. J. Schulman, Reconstruction from subsequences, *J. Combin. Theory Ser. A* **103** (2003), 337–348.

[7] W. Zhong and X. Zhang, Reconstruction of hypermatrices from subhypermatrices, *J. Combin. Theory Ser. A* **209** (2025), 105966.