Two Deletion Correcting Codes from Indicator Vectors

Jin Sima, **Netanel Raviv**, and **Jehoshua Bruck**
Department of Electrical Engineering, California Institute of Technology, Pasadena 91125, CA, USA

***Abstract*—Construction of capacity achieving deletion correcting codes has been a baffling challenge for decades. A recent breakthrough by Brakensiek *et al*., alongside novel applications in DNA storage, have reignited the interest in this longstanding open problem. In spite of recent advances, the amount of redundancy in existing codes is still orders of magnitude away from being optimal. In this paper, a novel approach for constructing binary two-deletion correcting codes is proposed. By this approach, parity symbols are computed from indicator vectors (i.e., vectors that indicate the positions of certain patterns) of the encoded message, rather than from the message itself. Most interestingly, the parity symbols and the proof of correctness are a direct generalization of their counterparts in the Varshamov-Tenengolts construction. Our techniques require $7\log(n) + o(\log(n)$ redundant bits to encode an n-bit message, which is near-optimal.**

I. INTRODUCTION

A *deletion* in a binary sequence $\mathbf{c} = (c_1, \ldots, c_n) \in \{0,1\}^n$ is the case where a symbol is removed from $\mathbf{c}$, which results in a subsequence length $n-1$. Similarly, the result of a k-deletion is a subsequence of $\mathbf{c}$ of length $n-k$. A k-deletion code $\mathcal{C}$ is a set of n-bit sequences, no two of which share a common subsequence of length $n-k$; and clearly, such a code can correct any k-deletion.

It has been proved in [1] that the largest size $L_k(n)$ of a k-deletion code satisfies

$$\frac{2^k(k!)^2 2^n}{n^{2k}} \lesssim L_k(n) \lesssim \frac{k! 2^n}{n^k}, \tag{1}$$

which implies the existence of a k-deletion code with at most $2k\log(n) + o(\log n)$ bits of redundancy for a constant k. However, to this day no explicit construction of such code is known beyond the case $k = 1$.

For $k = 1$, the well-known Varshamov-Tenengolts (VT) [2] construction

$$\left\{ \mathbf{c} : \sum_{i=1}^{n} i c_i = 0 \bmod (n+1) \right\} \tag{2}$$

can correct one deletion with at most $\log(n+1)$ bits of redundancy [1]. Several attempts to generalize the VT construction to $k > 1$ have been made. In the construction of [3], a modified Fibonacci sequence is used as weights instead of $(1, 2, \ldots, n)$ in (2). In [4], number-theoretic arguments are used to obtain k-deletion correction in run-length limited sequences. Yet, both [3] and [4] have rates that are asymptotically bounded away from 1.

The problem of finding an explicit k-deletion code of rate that approaches 1 as n grows has long been unsettled. Only recently, a code with $O(k^2 \log k \log n)$ redundancy bits and encoding/decoding complexity[1] of $O_k(n \log^4 n)$ was proposed in [5]. This code is based on a k-deletion code of length $\log n$, which is constructed using computer search. Nevertheless, the constants that are involved in the work of [5] are orders of magnitude away from the lower bound in (1) even for $k = 2$. Moreover, finding a k-deletion correcting code with an asymptotic rate 1 as an extension of the VT construction remains widely open[2].

One such potential extension is using higher order parity checks $\sum_{i=1}^{n} i^j c_i = 0 \bmod (n^j + 1)$ for $j = 1, \ldots, t$, but counterexamples are easily constructible even for $k = 2$. In this paper, we find that similar higher order parity checks work when $t = 3$, given that we restrict our attention to sequences with no consecutive ones. Consequently, applying these parity checks on certain *indicator vectors* yields the desired result. For a and b in $\{0,1\}$ and a binary sequence $\mathbf{c} = (c_i)_{i=1}^n$, the ab-indicator $\mathbb{1}_{ab}(\mathbf{c}) \in \{0,1\}^{n-1}$ of $\mathbf{c}$ is

$$\mathbb{1}_{ab}(\mathbf{c})_i = \begin{cases} 1 & \text{if } c_i = a \text{ and } c_{i+1} = b \\ 0 & \text{else} \end{cases}.$$

Since any two 10 or 01 patterns are at least two positions apart, the 10- and 01-indicators of any n-bit sequence do not contain consecutive ones, and hence higher order parity checks can be applied.

The parity checks in the proposed code rely on the following integer vectors.

$$\mathbf{m}^{(0)} \triangleq (1, 2, \ldots, n-1)$$
$$\mathbf{m}^{(1)} \triangleq \left(1, 1+2, 1+2+3, \ldots, \frac{n(n-1)}{2}\right)$$
$$\mathbf{m}^{(2)} \triangleq \left(1^2, 1^2+2^2, 1^2+2^2+3^2, \ldots, \frac{n(n-1)(2n-1)}{6}\right).$$

Further, for $\mathbf{c} \in \{0,1\}^n$ let

$$f(\mathbf{c}) \triangleq (\mathbb{1}_{10}(\mathbf{c}) \cdot \mathbf{m}^{(0)} \bmod 2n, \mathbb{1}_{10}(\mathbf{c}) \cdot \mathbf{m}^{(1)} \bmod n^2, \mathbb{1}_{10}(\mathbf{c}) \cdot \mathbf{m}^{(2)} \bmod n^3), \text{ and}$$
$$h(\mathbf{c}) \triangleq (\mathbb{1}_{01}(\mathbf{c}) \cdot \mathbb{1} \bmod 3, \mathbb{1}_{01}(\mathbf{c}) \cdot \mathbf{m}^{(1)} \bmod 2n),$$

[1]Here O_k denotes *parameterized complexity*, i.e., $O_k(n\log^4 n) = f(k)O(n\log^4 n)$ for some function f.

[2]For $k = 2$, [6] has very recently improved the redundancy up to $8\log n$ using techniques similar to [5], our techniques incur lower redundancy and complexity, and use a fundamentally different approach.

The work was supported in part by NSF grant CCF-1717884.
The work of Netanel Raviv was supported in part by the postdoctoral fellowship of the Center for the Mathematics of Information (CMI), Caltech, and in part by the Lester-Deutsch postdoctoral fellowship.

where $\cdot$ denotes inner product over the integers, and $\mathbb{1}$ denotes the all 1's vector.

For any integer k let $B_k(\mathbf{c})$ be the *k-deletion ball* of $\mathbf{c}$, i.e., the set of n-bit sequences that share a common $n-k$ subsequence with $\mathbf{c}$. The main result of the paper, from which a code construction is immediate, is as follows.

Theorem 1. *For $\mathbf{c},\mathbf{c}'\in\{0,1\}^n$, if $\mathbf{c}\in B_2(\mathbf{c}')$, $f(\mathbf{c})=f(\mathbf{c}')$, and $h(\mathbf{c})=h(\mathbf{c}')$, then $\mathbf{c}=\mathbf{c}'$.*

Theorem 1 readily implies that that the functions h and f can serve as the redundancy bits in a 2-deletion code, and the induced redundancy is at most $7\log(n)+o(\log n)$ (the additional term stems from protecting the redundancy bits). Furthermore, the encoding algorithm is trivial, and a linear decoding algorithm will be provided in future versions of this paper. Most interestingly, the proof of Theorem 1 can be seen as a higher dimensional variant of the proof for the VT construction, as explained in the remainder of this section.

Clearly, a length $n-1$ VT code can be seen as the set of sequences $\mathbf{c}$ for which the values of $\ell(\mathbf{c})\triangleq\mathbf{c}\cdot m_0 \bmod n$ coincide. Adopting this point of view, the proof of correctness can be done employing the following lemma.

Lemma 1. *For $\mathbf{c},\mathbf{c}'\in\{0,1\}^n$, if $\mathbf{c}\in B_1(\mathbf{c}')$ and $\ell(\mathbf{c})=\ell(\mathbf{c}')$, then $\mathbf{c}=\mathbf{c}'$.*

In turn, the proof of this lemma can be completed by defining the following function. For a vector $\mathbf{v}=(v_i)_{i=1}^{n-1}\in\mathbb{R}^{n-1}$, an integer $r\in[n-1]$, and a binary vector $\mathbf{x}=(x_i)_{i=1}^{s}$ with $r+s-2\le n-1$, let

$$\begin{aligned} g_{\mathbf{v}}(r,\mathbf{x}) &\triangleq \mathbf{x}\cdot((\mathbf{v}^{(r,r+s-2)},0)-(0,\mathbf{v}^{(r,r+s-2)})) \\ &= x_1v_r - x_s v_{r+s-2} + \sum_{t=2}^{s-1} x_t(v_{t+r-1}-v_{t+r-2}) \end{aligned} \tag{3}$$

where $\mathbf{v}^{(r,r+s-2)}\triangleq(v_r,v_{r+1},\ldots,v_{r+s-2})$. For $\mathbf{v}=\mathbf{m}^{(0)}$, one finds that $\ell(\mathbf{c})-\ell(\mathbf{c}')=g_{\mathbf{v}}(k_1,(\mathbf{c}^{(k_1,k_2)},c'_{k_2}))\bmod n$, where k_1 and k_2 ($k_1<k_2$) are the indices of the deletions after which $\mathbf{c}$ and $\mathbf{c}'$ are identical. Since $|g_{\mathbf{v}}(k_1,(\mathbf{c}^{(k_1,k_2)},c'_{k_2}))|\le n-1$, we have that $\ell(\mathbf{c})=\ell(\mathbf{c}')$ if and only if $g_{\mathbf{v}}(k_1,(\mathbf{c}^{(k_1,k_2)},c'_{k_2}))=0$. Hence, the following simple claim concludes the proof, up to a few observations that are left to the reader.

Lemma 2. *For integers r and s such that $r+s-2\le n-1$ and an s-bit binary vector $\mathbf{x}$, if $g_{\mathbf{m}^{(0)}}(r,\mathbf{x})=0$ then $\mathbf{x}$ is a constant vector.*

The crux of proving Theorem 1 boils down to the following higher dimensional variant of Lemma 2.

Lemma 3. *For integers r_1,r_2,s_1, and s_2 such that $r_2>r_1+s_1-2$ and $r_2+s_2-2\le n-1$, and binary sequences $\mathbf{x}$ and $\mathbf{y}$ of lengths s_1 and s_2, respectively, if*

$$\begin{aligned} &g_{\mathbf{m}^{(0)}}(r_1,\mathbf{x})+\lambda g_{\mathbf{m}^{(0)}}(r_2,\mathbf{y})=0, \text{ and} \\ &g_{\mathbf{m}^{(1)}}(r_1,\mathbf{x})+\lambda g_{\mathbf{m}^{(1)}}(r_2,\mathbf{y})=0, \end{aligned} \tag{4}$$

where $\lambda=\pm1$, then $\mathbf{x}$ and $\mathbf{y}$ are constant vectors.

Additional technical claims, which involve the remaining ingredients of the redundancy bits, are given in the sequel.

II. Outline

The proof of Theorem 1 is separated to the following two lemmas. In a nutshell, it is shown that for two confusable sequences, i.e., that share a common $n-2$ subsequence, if the f redundancies coincide, then so are the 10-indicators. Then, it is shown that confusable sequences with identical 10-indicators and identical h-redundancy have identical 01-indicators.

Lemma 4. *For $\mathbf{c}$ and $\mathbf{c}'$ in $\{0,1\}^n$, if $\mathbf{c}\in B_2(\mathbf{c}')$ and $f(\mathbf{c})=f(\mathbf{c}')$, then $\mathbb{1}_{10}(\mathbf{c})=\mathbb{1}_{10}(\mathbf{c}')$.*

Lemma 5. *For $\mathbf{c}$ and $\mathbf{c}'$ in $\{0,1\}^n$ such that $\mathbf{c}\in B_2(\mathbf{c}')$, if $\mathbb{1}_{10}(\mathbf{c})=\mathbb{1}_{10}(\mathbf{c}')$ and $h(\mathbf{c})=h(\mathbf{c}')$, then $\mathbb{1}_{01}(\mathbf{c})=\mathbb{1}_{01}(\mathbf{c}')$.*

From these lemmas it is clear that two n-bit sequences that share a common $n-2$ subsequence and agree on the redundancies f and h have identical 10- and 01-indicators, and hence the next simple lemma concludes the proof of Theorem 1.

Lemma 6. *For $\mathbf{c}$ and $\mathbf{c}'$ in $\{0,1\}^n$ such that $\mathbf{c}\in B_2(\mathbf{c}')$, if $\mathbb{1}_{10}(\mathbf{c})=\mathbb{1}_{10}(\mathbf{c}')$ and $\mathbb{1}_{01}(\mathbf{c})=\mathbb{1}_{01}(\mathbf{c}')$ then $\mathbf{c}=\mathbf{c}'$.*

The proofs of Lemma 4 and Lemma 5 make extensive use of the following two technical claims, that are easy to prove.

Lemma 7. *For $\mathbf{c}$ and $\mathbf{c}'$ in $\{0,1\}^n$, if $\mathbf{c}\in B_2(\mathbf{c}')$ then $\mathbb{1}_{10}(\mathbf{c})\in B_2(\mathbb{1}_{10}(\mathbf{c}'))$ and $\mathbb{1}_{01}(\mathbf{c})\in B_2(\mathbb{1}_{01}(\mathbf{c}'))$.*

Lemma 8. *For $\mathbf{c},\mathbf{c}'\in\{0,1\}^n$, if $\mathbf{c}\in B_2(\mathbf{c}')$ and $\mathbb{1}_{01}(\mathbf{c})\cdot\mathbb{1}=\mathbb{1}_{01}(\mathbf{c}')\cdot\mathbb{1}\bmod 3$, then $\mathbb{1}_{01}(c)\cdot\mathbb{1}=\mathbb{1}_{01}(c')\cdot\mathbb{1}$.*

In addition, one of the cases of the proof of Lemma 4 requires a specialized variant of Lemma 3.

Lemma 9. *Let r_1,r_2,s_1,s_2 and s_3 be positive integers that satisfy $r_2=r_1+s_1$ and $r_2+s_2+s_3\le n-1$, and let $\mathbf{x}\in\{0,1\}^{s_1+s_2+1}$ and $\mathbf{y}\in\{0,1\}^{s_2+s_3+1}$ be such that*

$$(x_{s_1+1},x_{s_1+2},\ldots,x_{s_1+s_2})=(y_2,y_3,\ldots,y_{s_2+1}),$$

and such that $(x_{s_1+1},x_{s_1+2},\ldots,x_{s_1+s_2})$ has no adjacent 1's. If

$$\begin{aligned} &g_{\mathbf{m}^{(0)}}(r_1,\mathbf{x})+g_{\mathbf{m}^{(0)}}(r_2,\mathbf{y})=0, \\ &g_{\mathbf{m}^{(1)}}(r_1,\mathbf{x})+g_{\mathbf{m}^{(1)}}(r_2,\mathbf{y})=0, \text{ and} \\ &g_{\mathbf{m}^{(2)}}(r_1,\mathbf{x})+g_{\mathbf{m}^{(2)}}(r_2,\mathbf{y})=0, \end{aligned} \tag{5}$$

then either $x_1=\ldots=x_{s_1+s_2+1}=y_1=\ldots=y_{s_2+s_3+1}$ or

$$\begin{aligned} &x_1=x_2=\ldots=x_{s_1+1}=1-y_1, \\ &x_t+x_{t+1}=1, \text{ for } t\in\{s_1+1,\ldots,s_1+s_2-1\}, \\ &x_{s_1+s_2+1}+y_{s_2+1}=1, \text{ and} \\ &y_{s_2+1}=\ldots=y_{s_2+s_3+1}. \end{aligned} \tag{6}$$

The proofs of Lemma 6, Lemma 7, and Lemma 8 are simple, and omitted due to space constraints. Lemma 5 is proved in Section III, and its more involved counterpart Lemma 4 is proved in Section IV. Finally, Lemma 3 and Lemma 9 are proved in Section V.

III. Proof of Lemma 5

Since $\mathbf{c}$ and $\mathbf{c}'$ have an identical 10-indicator, they can be written as

$$\begin{aligned} \mathbf{c} &= 0^{\pi_0}1^{\pi_1}0^{\pi_2}1^{\pi_3}\cdots0^{\pi_{2\ell}}1^{\pi_{2\ell+1}}, \\ \mathbf{c}' &= 0^{\tau_0}1^{\tau_1}0^{\tau_2}1^{\tau_3}\cdots0^{\tau_{2\ell}}1^{\tau_{2\ell+1}}, \end{aligned} \tag{7}$$

where $\{\pi_i\}_{i=0}^{2\ell+1}$ and $\{\tau_i\}_{i=0}^{2\ell+1}$ are nonnegative integers such that π_i and τ_i are strictly positive for every $i \notin \{0, 2\ell+1\}$, and such that $\pi_{2i}+\pi_{2i+1} = \tau_{2i}+\tau_{2i+1}$ for all $i \in \{0,1,\ldots,\ell\}$. In addition, since $h(\mathbf{c})_1 = h(\mathbf{c}')_1$ it follows from Lemma 8 that the number of 1's in the 01-indicators is equal. Hence, if π_0 and $\pi_{2\ell+1}$ (resp. τ_0 and $\tau_{2\ell+1}$) are both positive then this number is $\ell+1$, if precisely one of them is positive then it is ℓ, and if they are both zero it is $\ell-1$.

Let $\mathbf{d} = 0^{\gamma_0}1^{\gamma_1}0^{\gamma_2}1^{\gamma_3}\cdots 0^{\gamma_{2\ell}}1^{\gamma_{2\ell+1}} \in \{0,1\}^{n-2}$ be a common subsequence of $\mathbf{c}$ and $\mathbf{c}'$ which is obtained by deleting two bits from either $\mathbf{c}$ or $\mathbf{c}'$, where $\gamma_i \geq 0$ for all i. Then, it is readily verified that

$$\sum_{i=0}^{2\ell+1}(\pi_i - \gamma_i) = 2, \quad \sum_{i=0}^{2\ell+1}(\tau_i - \gamma_i) = 2, \text{, and hence}$$
$$\sum_{i=1}^{2\ell+1}|\pi_i - \tau_i| \leq \sum_{i=1}^{2\ell+1}|\pi_i - \gamma_i| + \sum_{i=1}^{2\ell+1}|\tau_i - \gamma_i| = 4.$$

Moreover, since $\pi_{2i} + \pi_{2i+1} = \tau_{2i} + \tau_{2i+1}$ for all $i \in \{0,1,\ldots,\ell\}$, it follows that $|\pi_{2i} - \tau_{2i}| = |\pi_{2i+1} - \tau_{2i+1}|$. Assuming for contradiction that the 01-indicators do not coincide leaves us with either of the following cases.

Case (a). There exists an integer j such that $|\pi_{2j} - \tau_{2j}|$ is either 1 or 2 and $\pi_{2i} = \tau_{2i}$ for $i \neq j$.

Case (b). There exist two integers m and r (where $m < r$) such that $|\pi_{2m} - \tau_{2m}| = |\pi_{2r} - \tau_{2r}| = 1$, and $\pi_{2i} = \tau_{2i}$ for $i \notin \{m, r\}$.

In Case (a), since $\pi_{2i} + \pi_{2i+1} = \tau_{2i} + \tau_{2i+1}$ for every i and $\pi_{2i} = \tau_{2i}$ for every $i \neq j$, it follows that $\mathbb{1}_{01}(\mathbf{c})$ and $\mathbb{1}_{01}(\mathbf{c}')$ differ in precisely two positions s and t such that $1 \leq s - t \leq 2$. Hence, since the number of 1's in the 01-indicators is equal, it follows that $\mathbb{1}_{01}(\mathbf{c})_s = \mathbb{1}_{01}(\mathbf{c}')_t$, $\mathbb{1}_{01}(\mathbf{c})_t = \mathbb{1}_{01}(\mathbf{c}')_s$, and $\mathbb{1}_{01}(\mathbf{c})_s \neq \mathbb{1}_{01}(\mathbf{c})_t$, and therefore

$$\begin{aligned} h(\mathbf{c})_2 - h(\mathbf{c}')_2 &= (\mathbb{1}_{01}(\mathbf{c})_s - \mathbb{1}_{01}(\mathbf{c}')_s)\binom{s+1}{2} + \\ &\quad (\mathbb{1}_{01}(\mathbf{c})_t - \mathbb{1}_{01}(\mathbf{c}')_t)\binom{t+1}{2} \\ &= \pm\left(\binom{s+1}{2} - \binom{t+1}{2}\right). \end{aligned} \tag{8}$$

Since $1 \leq s-t \leq 2$, it follows that (8) equals either $\pm(t+1)$ or $\pm(2t+3)$, and a contradiction follows since neither of which is 0 modulo $2n$, .

Similarly, in Case (b), if non of $\pi_{2m}, \tau_{2m}, \pi_{2m+1}, \tau_{2m+1}, \pi_{2r}, \tau_{2r}, \pi_{2r+1}, \tau_{2r+1}$ is zero, then $\mathbb{1}_{01}(\mathbf{c})$ and $\mathbb{1}_{01}(\mathbf{c}')$ differ in four positions $s, s+1, t$, and $t+1$, and hence

$$\begin{aligned} h(\mathbf{c})_2 - h(\mathbf{c}')_2 &= (\mathbb{1}_{01}(\mathbf{c})_s - \mathbb{1}_{01}(\mathbf{c}')_s)\binom{s+1}{2} + \\ &\quad (\mathbb{1}_{01}(\mathbf{c})_{s+1} - \mathbb{1}_{01}(\mathbf{c}')_{s+1})\binom{s+2}{2} + \\ &\quad (\mathbb{1}_{01}(\mathbf{c})_t - \mathbb{1}_{01}(\mathbf{c}')_t)\binom{t+1}{2} + \\ &\quad (\mathbb{1}_{01}(\mathbf{c})_{t+1} - \mathbb{1}_{01}(\mathbf{c}')_{t+1})\binom{t+2}{2}. \end{aligned} \tag{9}$$

Once again, since $\mathbb{1}_{01}(\mathbf{c})$ and $\mathbb{1}_{01}(\mathbf{c}')$ have an identical number of 1's, we have that

$$\begin{aligned} \mathbb{1}_{01}(\mathbf{c})_s &= \mathbb{1}_{01}(\mathbf{c}')_{s+1} & \mathbb{1}_{01}(\mathbf{c})_{s+1} &= \mathbb{1}_{01}(\mathbf{c}')_s \\ \mathbb{1}_{01}(\mathbf{c})_t &= \mathbb{1}_{01}(\mathbf{c}')_{t+1} & \mathbb{1}_{01}(\mathbf{c})_{t+1} &= \mathbb{1}_{01}(\mathbf{c}')_t \\ \mathbb{1}_{01}(\mathbf{c})_s &\neq \mathbb{1}_{01}(\mathbf{c}')_s & \mathbb{1}_{01}(\mathbf{c})_t &\neq \mathbb{1}_{01}(\mathbf{c}')_t. \end{aligned}$$

This readily implies that (9) equals either $\pm(s-t)$ or $\pm(s+t+2)$, and since non of which is 0 modulo $2n$, another contradiction is obtained. If $\pi_{2m} = 0$ (resp. $\tau_{2m} = 0$), by the discussion after Eq. (7) it follows that $\tau_{2r+1} = 0$ (resp. $\pi_{2r+1} = 0$), and hence $\mathbb{1}_{01}(\mathbf{c})$ and $\mathbb{1}_{01}(\mathbf{c}')$ differ in the first and last positions. Hence, (9) becomes $\pm(1 - \frac{n(n-1)}{2})$, which is nonzero modulo $2n$, and the claim follows.

IV. Proof of Lemma 4

Since $c \in B_2(c')$ it follows that there exist integers i_1, i_2, j_1, and j_2 such that

$$c \xrightarrow{\text{del' } i_1} d \xrightarrow{\text{del' } j_1} e$$
$$c' \xrightarrow{\text{del' } i_2} d' \xrightarrow{\text{del' } j_2} e$$

and by Lemma 7 it follows that there exist integers ℓ_1, ℓ_2, k_1, and k_2 such that

$$\mathbb{1}_{10}(c) \xrightarrow{\text{del' } \ell_1} \mathbb{1}_{10}(d) \xrightarrow{\text{del' } k_1} \mathbb{1}_{10}(e)$$
$$\mathbb{1}_{10}(c') \xrightarrow{\text{del' } \ell_2} \mathbb{1}_{10}(d') \xrightarrow{\text{del' } k_2} \mathbb{1}_{10}(e).$$

Due to symmetry between $\mathbf{c}$ and $\mathbf{c}'$, we distinguish between the following three cases. In each case, the difference between the f values of $\mathbf{c}$ and $\mathbf{c}'$ are given in terms of the function g (Eq. (3)). Further, the computation of these three differences, which is tedious but straightforward, is deferred to the full version of this paper. Notice that the equalities below are modular, and yet ordinary equality holds due to trivial bounds on g.

Case (a). If $\ell_1 \leq \ell_2 < k_2 \leq k_1$ then

$$\begin{aligned} \mathbb{1}_{10}(\mathbf{c})_{t+1} &= \mathbb{1}_{10}(\mathbf{c}')_t && \text{if } \ell_1 \leq t \leq \ell_2 - 1, \\ \mathbb{1}_{10}(\mathbf{c})_t &= \mathbb{1}_{10}(\mathbf{c}')_{t+1} && \text{if } k_2 \leq t \leq k_1 - 1, \end{aligned}$$

and $\mathbb{1}_{10}(\mathbf{c})_t = \mathbb{1}_{10}(\mathbf{c}')_t$ for any other $t \notin \{l_2, k_1\}$. Thus, for $e \in \{0,1,2\}$,

$$\begin{aligned} (f(\mathbf{c}) - f(\mathbf{c}'))_e = g_{\mathbf{m}^{(e)}}(\ell_1, (\mathbb{1}_{10}(\mathbf{c})^{(\ell_1,\ell_2)}, \mathbb{1}_{10}(\mathbf{c}')_{\ell_2})) - \\ g_{\mathbf{m}^{(e)}}(k_2, (\mathbb{1}_{10}(\mathbf{c}')^{(k_2,k_1)}, \mathbb{1}_{10}(\mathbf{c})_{k_1})). \end{aligned}$$

Case (b). If $\ell_1 \leq \ell_2 < k_1 \leq k_2$ then

$$\begin{aligned} \mathbb{1}_{10}(\mathbf{c})_{t+1} = \mathbb{1}_{10}(\mathbf{c}')_t \quad &\text{if } \ell_1 \leq t \leq \ell_2 - 1 \\ &\text{or } k_1 \leq t \leq k_2 - 1. \end{aligned}$$

and $\mathbb{1}_{10}(\mathbf{c})_t = \mathbb{1}_{10}(\mathbf{c}')_t$ for any other $t \notin \{l_2, k_2\}$. Thus, for $e \in \{0,1,2\}$,

$$\begin{aligned} (f(\mathbf{c}) - f(\mathbf{c}'))_e = g_{\mathbf{m}^{(e)}}(\ell_1, (\mathbb{1}_{10}(\mathbf{c})^{(\ell_1,\ell_2)}, \mathbb{1}_{10}(\mathbf{c}')_{\ell_2})) + \\ g_{\mathbf{m}^{(e)}}(k_1, (\mathbb{1}_{10}(\mathbf{c})^{(k_1,k_2)}, \mathbb{1}_{10}(\mathbf{c}')_{k_2})). \end{aligned}$$

Case (c). If $\ell_1 < k_1 \leq \ell_2 < k_2$ then

$$\begin{aligned} \mathbb{1}_{10}(\mathbf{c})_{t+1} = \mathbb{1}_{10}(\mathbf{c}')_t \quad &\text{if } \ell_1 \leq t \leq k_1 - 2 \\ &\text{or } \ell_2 + 1 \leq t \leq k_2 - 1, \\ \mathbb{1}_{10}(\mathbf{c})_{t+2} = \mathbb{1}_{10}(\mathbf{c}')_t \quad &\text{if } k_1 - 1 \leq t \leq \ell_2 - 1. \end{aligned}$$

and $\mathbb{1}_{10}(\mathbf{c})_t = \mathbb{1}_{10}(\mathbf{c}')_t$ for any other $t \notin \{l_2, k_2\}$. Thus, for $e \in \{0,1,2\}$,

$$(f(\mathbf{c}) - f(\mathbf{c}'))_e = g_{\mathbf{m}^{(e)}}(\ell_1, (\mathbb{1}_{10}(\mathbf{c})^{(\ell_1,k_1-1)}, \mathbb{1}_{10}(\mathbf{c})^{(k_1+1,\ell_2+1)}, \mathbb{1}_{10}(\mathbf{c}')_{\ell_2})) + g_{\mathbf{m}^{(e)}}(k_1, (\mathbb{1}_{10}(\mathbf{c})^{(k_1,k_2)}, \mathbb{1}_{10}(\mathbf{c}')_{k_2})).$$

For Case (a), Lemma 3 implies that

$$\mathbb{1}_{10}(\mathbf{c})_{\ell_1} = \ldots = \mathbb{1}_{10}(\mathbf{c})_{\ell_2} = \mathbb{1}_{10}(\mathbf{c}')_{\ell_2}$$
$$\mathbb{1}_{10}(\mathbf{c}')_{k_2} = \ldots = \mathbb{1}_{10}(\mathbf{c}')_{k_1} = \mathbb{1}_{10}(\mathbf{c})_{k_1},$$

which readily implies that $\mathbb{1}_{10}(\mathbf{c}) = \mathbb{1}_{10}(\mathbf{c}')$. In addition, Case (b) is similar, switching between k_1 and k_2.

For Case (c), Lemma 9 implies that either

$$\mathbb{1}_{10}(\mathbf{c})_{\ell_1} = \ldots = \mathbb{1}_{10}(\mathbf{c})_{k_2} = \mathbb{1}_{10}(\mathbf{c}')_{\ell_2} = \mathbb{1}_{10}(\mathbf{c}')_{k_2} \quad (10)$$

or

$$\begin{aligned}
&\mathbb{1}_{10}(\mathbf{c})_{\ell_1} = \ldots = \mathbb{1}_{10}(\mathbf{c})_{k_1-1} = \mathbb{1}_{10}(\mathbf{c})_{k_1+1},\\
&\mathbb{1}_{10}(\mathbf{c})_i + \mathbb{1}_{10}(\mathbf{c})_{i+1} = 1 \text{ for } i \in \{k_1, \ldots, \ell_2\},\\
&\mathbb{1}_{10}(\mathbf{c}')_{\ell_2} + \mathbb{1}_{10}(\mathbf{c}')_{k_2} = 1, \text{ and}\\
&\mathbb{1}_{10}(\mathbf{c})_{\ell_2+1} = \ldots = \mathbb{1}_{10}(\mathbf{c})_{k_2} = \mathbb{1}_{10}(\mathbf{c}')_{k_2}. \quad (11)
\end{aligned}$$

In either (10) or (11), one can verify that $\mathbb{1}_{10}(\mathbf{c})_t = \mathbb{1}_{10}(\mathbf{c}')_t$ for all t by an incremental argument that follows the above. For example, for $t < \ell_1$ the claim is obvious, for $\ell_1 \le t \le k_1 - 2$ we have that $\mathbb{1}_{10}(\mathbf{c})_t = \mathbb{1}_{10}(\mathbf{c})_{t+1} = \mathbb{1}_{10}(\mathbf{c}')_t$, etc.

V. PROOFS OF g-LEMMAS

Proof: (of Lemma 3) For $\lambda = -1$, we distinguish between four cases according to the value of (y_1, y_{s_2}).

If $(y_1, y_{s_2}) = (0,1)$ then

$$\begin{aligned}
g_{\mathbf{m}^{(0)}}(r_1,\mathbf{x}) - g_{\mathbf{m}^{(0)}}(r_2,\mathbf{y}) &= \mathbf{m}^{(0)}_{r_1}x_1 + \sum_{t=2}^{s_1-1}(\mathbf{m}^{(0)}_{t+r_1-1} - \mathbf{m}^{(0)}_{t+r_1-2})x_t - \\
&\quad \mathbf{m}^{(0)}_{r_1+s_1-2}x_{s_1} - \mathbf{m}^{(0)}_{r_2}y_1 - \\
&\quad \sum_{t=2}^{s_2-1}(\mathbf{m}^{(0)}_{t+r_2-1} - \mathbf{m}^{(0)}_{t+r_2-2})y_t + \mathbf{m}^{(0)}_{r_2+s_2-2}y_{s_2}\\
&\ge -\mathbf{m}^{(0)}_{r_1+s_1-2} - \sum_{t=2}^{s_2-1}(\mathbf{m}^{(0)}_{t+r_2-1} - \mathbf{m}^{(0)}_{t+r_2-2}) + \\
&\quad \mathbf{m}^{(0)}_{r_2+s_2-2}\\
&= \mathbf{m}^{(0)}_{r_2} - \mathbf{m}^{(0)}_{r_1+s_1-2} > 0,
\end{aligned}$$

a contradiction. If $(y_1, y_{s_2}) = (1,0)$, the proof is similar.

If $(y_1, y_{s_2}) = (1,1)$ let

$$S_1 \triangleq \{j : y_{j-r_2+1} = 1, r_2+1 \le j \le r_2+s_2-2\}, \text{ and}$$
$$S_1^c \triangleq \{j : y_{j-r_2+1} = 0, r_2+1 \le j \le r_2+s_2-2\},$$

and notice that

$$\begin{aligned}
g_{\mathbf{m}^{(0)}}(r_2,\mathbf{y}) &= \mathbf{m}^{(0)}_{r_2} - \mathbf{m}^{(0)}_{r_2+s_2-2} + \sum_{j\in S_1}(\mathbf{m}^{(0)}_j - \mathbf{m}^{(0)}_{j-1})\\
&= -\sum_{j=r_2+1}^{r_2+s_2-2}(\mathbf{m}^{(0)}_j - \mathbf{m}^{(0)}_{j-1}) + \sum_{j\in S_1}(\mathbf{m}^{(0)}_j - \mathbf{m}^{(0)}_{j-1})\\
&= -\sum_{j\in S_1^c}(\mathbf{m}^{(0)}_j - \mathbf{m}^{(0)}_{j-1}) = -\sum_{j\in S_1^c} 1, \text{ and}\\
g_{\mathbf{m}^{(1)}}(r_2,\mathbf{y}) &= -\sum_{j\in S_1^c}(\mathbf{m}^{(1)}_j - \mathbf{m}^{(1)}_{j-1}) = -\sum_{j\in S_1^c} j. \quad (12)
\end{aligned}$$

Now, on the one hand if $x_{s_1} = 0$ we have

$$g_{\mathbf{m}^{(0)}}(r_1,\mathbf{x}) = \mathbf{m}^{(0)}_{r_1}x_1 + \sum_{t=2}^{s_1-1}(\mathbf{m}^{(0)}_{t+r_1-1} - \mathbf{m}^{(0)}_{t+r_1-2})x_t \ge 0, \quad (13)$$

and hence, (12) and (13) imply that $g_{\mathbf{m}^{(0)}}(r_1,\mathbf{x}) - g_{\mathbf{m}^{(0)}}(r_2,\mathbf{y}) \ge 0$, and equality holds only when $g_{\mathbf{m}^{(0)}}(r_1,\mathbf{x})$ and $g_{\mathbf{m}^{(0)}}(r_2,\mathbf{y})$ are both 0, which by Lemma 2 implies that $\mathbf{x}$ and $\mathbf{y}$ are constant vectors. On the other hand, if $x_{s_1} = 1$ let $S_2 = \{j : x_{\max\{j-r_1+1,1\}} = 0, 1 \le j \le r_1+s_1-2\}$, and notice that

$$\begin{aligned}
g_{\mathbf{m}^{(0)}}(r_1,\mathbf{x}) &= \mathbf{m}^{(0)}_{r_1}(x_1 - 1) + \\
&\quad \sum_{t=2}^{s_1-1}(\mathbf{m}^{(0)}_{t+r_1-1} - \mathbf{m}^{(0)}_{t+r_1-2})(x_t - 1)\\
&= -\sum_{t\in S_2} 1, \text{ and}\\
g_{\mathbf{m}^{(1)}}(r_1,\mathbf{x}) &= -\sum_{t\in S_2} t. \quad (14)
\end{aligned}$$

Plugging (12) and (14) into (4), we have

$$-\sum_{t\in S_2} 1 + \sum_{j\in S_1^c} 1 = 0,$$
$$-\sum_{t\in S_2} t + \sum_{j\in S_1^c} j = 0.$$

This implies that the sets S_1^c and S_2 have the same cardinality and the same sum of elements. However, the maximum element in S_2 is smaller than the minimum element in S_1^c. Therefore S_1^c and S_2 are empty, which implies that $\mathbf{x}$ is the 0 vector and $\mathbf{y}$ is the all 1's vector.

If $(y_1, y_{s_2}) = (0,0)$, note that (4) implies that $g_{\mathbf{m}^{(e)}}(r_1,\overline{\mathbf{x}}) - g_{\mathbf{m}^{(e)}}(r_2,\overline{\mathbf{y}}) = 0$ for $e \in \{1,2\}$, where $\overline{\mathbf{x}} \triangleq \mathbb{1} - \mathbf{x}$ and $\overline{\mathbf{y}} \triangleq \mathbb{1} - \mathbf{y}$. Since $(1-y_1, 1-y_{s_2}) = (1,1)$, from the previous case we have that $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ are constant vectors, and thus so are $\mathbf{x}$ and $\mathbf{y}$. Finally, for $\lambda = 1$, it can be verified that $g_{\mathbf{m}^{(e)}}(r_2,\mathbf{y}) = -g_{\mathbf{m}^{(e)}}(r_2,\overline{\mathbf{y}})$ which reduces to the case $\lambda = -1$. ■

Proof: (of Lemma 9) We distinguish between four cases according to the value of $(x_{s_1+s_2+1}, y_{s_2+s_3+1})$. If $(x_{s_1+s_2+1}, y_{s_2+s_3+1}) = (0,0)$, then similar to (13), we have that $g_{\mathbf{m}^{(0)}}(r_1,\mathbf{x}) + g_{\mathbf{m}^{(0)}}(r_2,\mathbf{y}) \ge 0$, where equality holds only if $\mathbf{x}$ and $\mathbf{y}$ are constant 0 vectors. Similarly, if $(x_{s_1+s_2+1}, y_{s_2+s_3+1}) = (1,1)$ we have $(\overline{x}_{s_1+s_2+1}, \overline{y}_{s_2+s_3+1}) = (0,0)$ and hence $g_{\mathbf{m}^{(0)}}(r_1,\overline{\mathbf{x}}) + g_{\mathbf{m}^{(0)}}(r_2,\overline{\mathbf{y}}) \ge 0$ where equality holds when $\mathbf{x}$ and $\mathbf{y}$ are constant 1 vectors.

If $(x_{s_1+s_2+1}, y_{s_2+s_3+1}) = (0,1)$, then for $y_1 = 0$ we have

$$\begin{aligned}
&g_{\mathbf{m}^{(0)}}(r_1,\mathbf{x}) + g_{\mathbf{m}^{(0)}}(r_2,\mathbf{y}) =\\
&= \mathbf{m}^{(0)}_{r_1}x_1 + \sum_{t=2}^{s_1+1}(\mathbf{m}^{(0)}_{t+r_1-1} - \mathbf{m}^{(0)}_{t+r_1-2})x_t
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{t=s_1+1}^{s_1+s_2-1}(\mathbf{m}^{(0)}_{t+r_1}-\mathbf{m}^{(0)}_{t+r_1-1})(x_t+x_{t+1})\\
&+\sum_{t=s_2+1}^{s_2+s_3}(\mathbf{m}^{(0)}_{t+r_2-1}-\mathbf{m}^{(0)}_{t+r_2-2})y_t-\mathbf{m}^{(0)}_{r_2+s_2+s_3-1}\\
&\le 0,
\end{aligned}
$$

where equality equality holds when

$$
\begin{aligned}
&x_t = 1 \text{ for } t\in\{1,\ldots,s_1+1\},\\
&x_t+x_{t+1} = 1 \text{ for } t\in\{s_1+1,\ldots,s_1+s_2-1\}, \text{ and}\\
&y_t = 1 \text{ for } t\in\{s_2+1,\ldots,s_2+s_3\},
\end{aligned}
$$

and hence (6) holds. On the other hand, when $y_1 = 1$, let

$$
\begin{aligned}
S_1 &= \{t : x_{\max\{t-r_1+1,1\}} = 1, 1\le t\le s_1+r_1\},\\
S_2 &= \{t : x_{t-r_1}+x_{t-r_1+1} = 0, r_2+1\le t\le r_2+s_2-1\},\\
S_3 &= \{t : y_{t-r_2+1} = 0, r_2+s_2\le t\le r_2+s_2+s_3-1\},
\end{aligned}
$$

and notice that

$$
\begin{aligned}
&g_{\mathbf{m}^{(0)}}(r_1, \mathbf{x}) + g_{\mathbf{m}^{(0)},r_2}(\mathbf{y})\\
&\quad=\mathbf{m}^{(0)}_{r_1}x_1+\sum_{t=2}^{s_1+1}(\mathbf{m}^{(0)}_{t+r_1-1}-\mathbf{m}^{(0)}_{t+r_1-2})x_t+\\
&\qquad\mathbf{m}^{(0)}_{s_1+r_1}+\sum_{t=s_1+1}^{s_1+s_2-1}(\mathbf{m}^{(0)}_{t+r_1}-\mathbf{m}^{(0)}_{t+r_1-1})(x_t+x_{t+1})+\\
&\qquad\sum_{t=s_2+1}^{s_2+s_3}(\mathbf{m}^{(0)}_{t+r_2-1}-\mathbf{m}^{(0)}_{t+r_1-2})y_t-\mathbf{m}^{(0)}_{r_2+s_2+s_3-1}\\
&\quad=\sum_{t\in S_1}(\mathbf{m}^{(0)}_{t}-\mathbf{m}^{(0)}_{t-1})-\sum_{t\in S_2}(\mathbf{m}^{(0)}_{t}-\mathbf{m}^{(0)}_{t-1})-\\
&\qquad\sum_{t\in S_3}(\mathbf{m}^{(0)}_{t}-\mathbf{m}^{(0)}_{t-1})\\
&\quad=\sum_{t\in S_1}1-\sum_{t\in S_2}1-\sum_{t\in S_3}1 \qquad (15)
\end{aligned}
$$

Similarly, we have

$$
g_{\mathbf{m}^{(1)}}(r_1,\mathbf{x})+g_{\mathbf{m}^{(1)}}(r_2,\mathbf{y})=\sum_{t\in S_1}t-\sum_{t\in S_2}t-\sum_{t\in S_3}t. \quad (16)
$$

Equations (5), (15), and (16) imply that the cardinality of S_1 equals the sum of cardinalities of S_2 and S_3, and in addition, the sum of elements of S_1 equals the sum of elements of S_2 and S_3. Note that the minimum element of $S_2\cup S_3$ is larger than the maximum element of S_1. This is impossible, unless S_1, S_2, and S_3 are empty, which implies that $x_t = 0$ for $t\in\{1,\ldots,s_1+1\}$, $x_t+x_{t+1} = 1$ for $t\in\{s_1+1,\ldots,s_1+s_2-1\}$, and $y_t = 1$ for $t\in\{s_2+1,\ldots,s_2+s_3\}$, and hence (6) holds.

Finally, if $(x_{s_1+s_2+1}, y_{s_2+s_3+1}) = (1,0)$, then for $y_1 = 0$, arguments similar to the above yield

$$
g_{\mathbf{m}^{(0)}}(r_1,\mathbf{x})+g_{\mathbf{m}^{(0)}}(r_2,\mathbf{y})=-\sum_{t\in S_1}1-\sum_{t\in S_2}1+\sum_{t\in S_3}1=0
$$

for

$$
\begin{aligned}
S_1 &= \{t : x_{\max\{t-r_1+1,1\}} = 0, 1\le t\le s_1+r_1\},\\
S_2 &= \{t : x_{t-r_1}+x_{t-r_1+1} = 0, r_2+1\le t\le r_2+s_2-1\},\\
S_3 &= \{t : y_{t-r_2+1} = 1, r_2+s_2\le t\le r_2+s_2+s_3-1\}.
\end{aligned}
$$

Then, we obtain sets with identical cardinalities and sum of elements, and yet the smallest element in one is greater the the largest element in the others. Therefore, it follows that $x_t = 1$ for $t\in\{1,\ldots,s_1+1\}$, $x_t+x_{t+1} = 1$ for $t\in\{s_1+1,\ldots,s_1+s_2-1\}$, and $y_t = 0$ for $t\in\{s_2+1,\ldots,s_2+s_3\}$, and hence (6) holds.

On the other hand, for $y_1 = 1$ we get

$$
g_{\mathbf{m}^{(0)}}(r_1,\mathbf{x})+g_{\mathbf{m}^{(0)}}(r_2,\mathbf{y})=\sum_{t\in S_1}1-\sum_{t\in S_2}1+\sum_{t\in S_3}1, \text{ for}
$$

$$
\begin{aligned}
S_1 &= \{t : x_{\max\{t-r_1+1,1\}} = 1, 1\le t\le s_1+r_1\},\\
S_2 &= \{t : x_{t-r_1}+x_{t-r_1+1} = 0, r_2+1\le t\le r_2+s_2-1\},\\
S_3 &= \{t : y_{t-s_2+1} = 1, r_2+s_2\le t\le r_2+s_2+s_3-1\}.
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
g_{\mathbf{m}^{(1)}}(r_1,\mathbf{x})+g_{\mathbf{m}^{(1)}}(r_2,\mathbf{y})&=\sum_{t\in S_1}t-\sum_{t\in S_2}t+\sum_{t\in S_3}t\\
g_{\mathbf{m}^{(2)}}(r_1,\mathbf{x})+g_{\mathbf{m}^{(2)}}(r_2,\mathbf{y})&=\sum_{t\in S_1}t^2-\sum_{t\in S_2}t^2+\sum_{t\in S_3}t^2
\end{aligned}
$$

which implies that the linear equation

$$
A\boldsymbol{x}=\begin{bmatrix}\sum_{t\in S_1}1 & \sum_{t\in S_2}1 & \sum_{t\in S_3}1\\ \sum_{t\in S_1}t & \sum_{t\in S_2}t & \sum_{t\in S_3}t\\ \sum_{t\in S_1}t^2 & \sum_{t\in S_2}t^2 & \sum_{t\in S_3}t^2\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=0 \quad (17)
$$

has a nonzero solution $(1,-1,1)^\top$. However, the determinant

$$
\begin{aligned}
\det(A) &= \sum_{i\in S_1, j\in S_2, k\in S_3}\det\begin{pmatrix}1&1&1\\ i&j&k\\ i^2&j^2&k^2\end{pmatrix}\\
&= \sum_{i\in S_1, j\in S_2, k\in S_3}(j-i)(k-i)(k-j) \qquad (18)
\end{aligned}
$$

is strictly positive since $\max_{i\in S_1} i < \min_{j\in S_2} j < \min_{k\in S_3} k$, where the first equality follows from the linearity of the determinant in each column. Thus, Eq. (17) has no nonzero solution unless $A = 0$, which implies that S_1, S_2, and S_3 are empty. Therefore, $x_t = 0$ for $t\in\{1,\ldots,s_1+1\}$, $x_t+x_{t+1} = 1$ for $t\in\{s_1+1,\ldots,s_1+s_2-1\}$, and $y_t = 0$ for $t\in\{s_2+1,\ldots,s_2+s_3\}$, which implies (6). ■

REFERENCES

[1] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” in *Soviet physics doklady*, vol. 10, no. 8, 1966, pp. 707–710.

[2] R. R. Varshamov and G. M. Tenengolts, “Codes which correct single asymmetric errors,” in *Autom. Remote Control*, vol. 26, no. 2, 1965, pp. 286–290.

[3] A. S. Helberg and H. C. Ferreira, “On multiple insertion/deletion correcting codes,” *IEEE Trans. on Inf. Th.*, vol. 48, no. 1, pp. 305–308, 2002.

[4] F. Paluncic, K. A. Abdel-Ghaffar, H. C. Ferreira, and W. A. Clarke, “A multiple insertion/deletion correcting code for run-length limited sequences,” *IEEE Trans. on Inf. Th.*, vol. 58, no. 3, pp. 1809–1824, 2012.

[5] J. Brakensiek, V. Guruswami, and S. Zbarsky, “Efficient low-redundancy codes for correcting multiple deletions,” in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms* (SODA), pp. 1884–1892. 2016

[6] R. Gabrys and F. Sala, “Codes correcting two deletions.” *arXiv:1712.07222* [cs.IT], 2017.