# Capacity-Approaching Pseudorandom Code for Stochastic Edit Channels

Anonymous Submission

**Abstract.** Pseudorandom codes (PRCs) are error-correcting codes whose codewords are computationally indistinguishable from uniformly random strings to any efficient adversary. We study public-key multi-bit PRCs over the stochastic memoryless binary substitution-insertion-deletion channel and its $q$-ary generalization. Here $P_i, P_d$ are the insertion and deletion probabilities, and $P_s$ is the substitution probability conditional on transmission.

We first show that the majority-based construction of Christ and Gunn yields a binary PRC with vanishing rate that recovers the transmitted message with overwhelming probability whenever $P_i + P_d < 1$ and $P_s < 1/2$.

We then use this binary PRC to protect control information in a high-rate, uniquely decodable PRC over an unbounded family of constant-size alphabets. For fixed channel parameters satisfying $P_i + P_d < 1$ and $P_s < 1$, its decoder recovers the transmitted message with overwhelming probability. The alphabet size may depend on the channel parameters and desired rate gap, but not on the security parameter. As the gap tends to zero and the alphabet grows through the supported sizes, the rate approaches the capacity upper bound obtained by revealing insertion and deletion locations to the receiver. The constructions use the assumptions underlying the Christ–Gunn public-key PRC and secure pseudorandom generators.

**Keywords:** Pseudorandom codes · stochastic edit channels · synchronization errors

# 1 Introduction

## 1.1 Motivation

Error-correcting codes allow a receiver to recover a message despite transmission errors, but their codewords may have recognizable structure. We seek codes that provide reliable recovery while making their codewords computationally indistinguishable from uniformly random strings of the same length.

We study these requirements for the binary substitution-insertion-deletion (BSID) channel of Briffa et al. [BSW10] and its $q$-ary generalization. The channel samples edits according to a fixed stochastic law: inserted symbols are uniform, and each substitution replaces the transmitted symbol uniformly by a different symbol. The precise model is given in Section 2.6.

Introduced by Christ and Gunn [CG24], pseudorandom code (PRC) provides a natural cryptography primitive for this setting. PRC is a type of error-correcting code that is computationally indistinguishable from uniform randomness. Per [CG24], besides error-correcting ability (robustness) and pseudorandomness, PRC should also satisfy soundness: every fixed received string chosen independently of the keys is rejected except with negligible probability over key generation.

The majority-based construction of Christ and Gunn [CG24] is robust against deletions followed by bit flips. We aim to extend its robustness analysis to the BSID channel, which also introduces insertions. Since its majority expansion gives vanishing rate, we then seek high-rate pseudorandom codes over sufficiently large constant alphabets.

## 1.2 Our Contributions

This paper focuses on pseudorandom code for BSID and $q$-ary SID channels, a generalized version of BSID channel. The two channels are modeled in the same manner: for each current input symbol, the channel inserts a uniformly random symbol with probability $P_i$ without advancing the input, deletes the current symbol with probability $P_d$ and advances to the next input symbol, or transmits it with probability $1 - P_i - P_d$ and advances. Conditioned on transmission, with probability $P_s$ the symbol is replaced uniformly by one of the other $q - 1$ alphabet symbols. Our two contributions are summarized as follows:

- Corollary 1 shows that Construction 6 of Christ and Gunn [CG24], instantiated with their public-key multi-bit PRC, yields a binary PRC that recovers the transmitted message over the BSID channel with overwhelming probability whenever $P_i + P_d < 1$ and $P_s < 1/2$. Its rate vanishes because each bit of the inner codeword of length $n$ is expanded into a majority block of length $m$, where $m$ grows with $n$. The key step, Theorem 1, shows that majority decoding recovers more than half the coordinates of a uniformly random binary input by a constant margin with overwhelming probability, using the positive correlation between each input bit and the majority of its reconstructed block.

- Fix channel parameters with $P_i+P_d < 1$ and $P_s < 1$. Write $R_* := \frac{(1-P_i-P_d)(1-P_s)}{1-P_i}$ for the expected fraction of input symbols transmitted without deletion or substitution. Revealing insertion and deletion locations to the receiver gives an upper bound on the capacity of the original channel. By Proposition 1, this upper bound tends to $R_*$ as $q \to \infty$. For every constant target gap $0 < \zeta < R_*$, Theorem 2 gives a polynomial-time public-key multi-bit PRC of rate $R \geq R_* - \zeta$ that recovers the transmitted message over the $q$-ary SID channel with overwhelming probability. Its constant-size alphabet is chosen from the family in Theorem 2; its size may depend on the channel parameters and $\zeta$, but not on the security parameter $\lambda$. For fixed channel parameters, Proposition 1 shows that our rate can approach this upper bound as $\zeta \to 0$ and $q$ grows through the supported alphabet sizes.

The binary PRC of Corollary 1 has vanishing rate, but it can protect the seeds of the mask and self-matching string, together with the hash key and hash value. Theorem 2 uses this PRC for that control information and a separate code for the message. The resulting higher-rate PRC operates over a sufficiently large constant alphabet.

Our binary control code uses Construction 6 of Christ and Gunn [CG24], instantiated with their public-key multi-bit PRC from Theorem 3. Corollary 1 extends its robustness guarantee to the BSID channel. This instantiation relies on Assumption 1 of [CG24]: either subexponential hardness of learning parity with noise (LPN), or polynomial-time hardness of LPN and the low-density planted XOR problem. We also use standard PRGs to generate the mask and index string. Our BSID extension, the HRW list-recovery and repositioning arguments, and almost-universal hashing introduce no further computational assumptions.

## 1.3 Literature Review

Construction 6 and Theorem 4 of Christ and Gunn [CG24] give a binary PRC robust to independent deletions with any constant probability below 1, followed by independent bit flips with any constant probability below 1/2. Its majority encoding gives vanishing rate. Corollary 1 extends this construction to the BSID channel, including insertions. Theorem 2 then uses it to protect control information in a high-rate PRC over larger constant alphabets.

Huang et al. [HLMZ26] give high-rate public-key PRCs against oblivious edit channels. Their binary and constant-alphabet constructions tolerate sublinear numbers of edits,

**Table 1:** Public-key multi-bit PRCs under their respective channel models. Here $\lambda$ is the security parameter and $N$ is the transmitted codeword length. Edit budgets count insertions and deletions; a substitution counts as two operations.

| Work | Rate | Alphabet size | Channel | Edit budget | Channel probabilities |
|---|---|---|---|---|---|
| Christ–Gunn [CG24] | $o(1)$ | binary | stochastic | — | $P_d < 1$, $P_s < 1/2$ |
| Huang et al. [HLMZ26] | $1 - \varepsilon_{\mathrm{H}}$ | $\mathsf{poly}(\lambda, 1/\eta_{\mathrm{H}})$ | oblivious | $(\varepsilon_{\mathrm{H}} - \eta_{\mathrm{H}})N$ | — |
| Huang et al. [HLMZ26] | $1 - \varepsilon_{\mathrm{H}}$ | $\mathsf{poly}(1/\varepsilon_{\mathrm{H}})$ | oblivious | $O(N^{1-c})$ | — |
| Huang et al. [HLMZ26] | $1/2 - \varepsilon_{\mathrm{H}}$ | binary | oblivious | $O(N^{1-c})$ | — |
| Theorem 2 | $\geq R_* - \zeta$ | $q$ (constant) | stochastic SID | — | $P_i + P_d < 1$, $P_s < 1$ |

In the Christ–Gunn row, $P_d$ and $P_s$ are the independent deletion and subsequent bit-flip probabilities. The Huang et al. rows use sufficiently small constants $0 < \eta_{\mathrm{H}} < \varepsilon_{\mathrm{H}}$ and $0 < c < 1$ from Corollaries 5.10, 5.9, and 6.14 of [HLMZ26]. For our result, $R_* = (1 - P_i - P_d)(1 - P_s)/(1 - P_i)$ and $0 < \zeta < R_*$; $q$ is constant in $\lambda$ and depends on $P_i, P_d, P_s, \zeta$. The rows use different channel models.

with rates approaching $1/2$ and $1$, respectively. For codeword length $N$, their polynomial-alphabet construction has rate $1 - \varepsilon_{\mathrm{H}}$ and tolerates $(\varepsilon_{\mathrm{H}} - \eta_{\mathrm{H}})N$ insertions and deletions, for sufficiently small constants $0 < \eta_{\mathrm{H}} < \varepsilon_{\mathrm{H}}$; Table 1 gives the alphabet sizes and error budgets. These guarantees concern oblivious edit patterns independent of the keys and codeword. We study the stochastic law in Section 2.6, where repeated insertions can make the total edit count exceed the input length, while at most $N$ input symbols can be deleted. The Singleton tradeoff concerns a prescribed worst-case insertion/deletion budget; our rate benchmark is the stochastic-channel capacity upper bound in Proposition 1. Our constant-alphabet rate guarantee applies to the memoryless channel in Section 2.6. The alphabet is chosen according to these channel parameters and the desired rate gap.

The zero-bit edit-robust PRC constructions of Golowich and Moitra [GM24], Christ et al. [CGG+26], and Huang et al. [HLMZ26] address detection of an encoded signal. Our results concern recovery of a chosen multi-bit message and the associated rate.

Additionally, Dehmel and Li [DL26] prove an upper bound on the fraction of adversarial deletions that fixed-alphabet multi-bit public-key PRCs can tolerate with unique decoding. The adversary may choose which symbols to delete using both the public key and the transmitted codeword. The channel in Section 2.6 instead samples its edits from a fixed stochastic law independently of the keys.

Our construction also uses classical error-correcting codes to encode the payload. Theorem 2.8 of Haeupler and Shahrasbi [HS21] gives a high-rate, list-decodable error-correcting code for insertion and deletion errors. It combines a self-matching index string with the high-rate list-recoverable codes of Hemenway, Ron-Zewi, and Wootters [HRZW17]. Repositioning using this string produces sets of candidate symbols for each original position, which list recovery converts into candidate messages. We combine these tools with pseudorandom masking, the binary control PRC of Corollary 1, and hash-based message selection to obtain Theorem 2.

## 1.4 Technical Overview

### 1.4.1 Extending the CG24 majority construction to the BSID channel

To encode each bit of the inner PRC codeword, the majority encoder samples an odd-length binary block uniformly conditioned on that bit being its majority. Insertions and deletions

obscure the boundaries between these blocks, so the decoder partitions the received string into nearly equal-length blocks, takes their majorities, and applies the inner PRC decoder to correct the remaining errors. Lemmas 4 and 6 show that, with the majority-block length specified in Theorem 1, all but a vanishing fraction of the surviving bits from each original block lie in its corresponding reconstructed block with overwhelming probability.

A majority may be decided by a small difference between the numbers of zeros and ones, so we compare this difference in each original block and its reconstructed block. For uniformly random input to the majority encoder, random inserted bits and bits from neighboring blocks contribute noise independent of the intended original block. Correctly placed surviving bits preserve positive correlation between the two differences when $P_i + P_d < 1$ and $P_s < 1/2$. Lemma 7 converts this correlation into a majority-agreement probability greater than $1/2$ by a constant.

The alignment bounds restrict each reconstructed block to channel output from its intended original block and its immediate neighbors. The proof of Theorem 1 conditions on the insertion counts and deletion decisions and groups every third reconstructed block, obtaining independence of majority-agreement events within each group. Concentration within these groups gives a constant margin above one half of the input bits correctly decoded, with overwhelming probability.

We retain the majority encoder and decoder of Construction 6 of Christ and Gunn [CG24] and adapt the alignment and majority-agreement strategy of their Lemma 15 to account for fluctuations in the number of inserted bits and the noise contributed by their values.

### 1.4.2 Obtaining a high-rate construction

The binary PRC of Corollary 1 has vanishing rate because it expands every bit of the length-$n$ inner codeword into a majority block of length $m$, where $m$ grows with $n$. We reserve it for control information and encode the message using the high-rate construction of Theorem 2.8 of Haeupler and Shahrasbi [HS21], whose decoder returns a list. To obtain a PRC, we must make the codewords pseudorandom and identify the transmitted message in the decoded list.

Let $x \in \Sigma_C^N$ be the message's Hemenway–Ron-Zewi–Wootters (HRW) encoding [HRZW17]. A pseudorandom mask $R \in \Sigma_C^N$ hides its code structure. A pseudorandom index string $S \in \Sigma_S^N$ supplies position information after edits while keeping the index coordinate indistinguishable from uniform; Lemma 10 guarantees its required self-matching property with overwhelming probability. The binary PRC protects both seeds, an independent almost-universal hash key, and the message's hash value for selecting a candidate from the decoded list. The binary control codeword and the HRW codeword can have different lengths. We choose the HRW length $N$ to be at least the binary control length and extend the binary codeword to length $N$ with independent uniform bits, obtaining $b$. We transmit $c_i = (x_i + R_i, S_i, b_i)$ over an alphabet of size $q = 2|\Sigma_C||\Sigma_S|$.

The binary projection keeps insertion and deletion probabilities $P_i, P_d$, and uniform inserted symbols give uniform bits. Conditioned on transmission, its bit flips with probability $qP_s/(2(q-1)) < 1/2$ for the alphabets in Theorem 2. Insertions and deletions obscure the control codeword's endpoint in the received string. Lemma 9 shows that control information can still be recovered by decoding the first $p_{\mathrm{PRC}}$ fraction, or prefix, of the received binary string, where $p_{\mathrm{PRC}}$ is the unpadded control length divided by $N$. The repositioning algorithm uses the full received sequence of payload and index pairs. Counting each substitution as one deletion and one insertion upper-bounds the edits to the payload and index coordinates, as in Section 4.1. With the recovered seeds, the decoder regenerates $R, S$, uses $S$ to form candidate sets for the original positions, subtracts $R_i$ from each candidate at position $i$, and applies HRW list recovery to obtain a list containing the message with overwhelming probability. The hash check selects the transmitted

message with overwhelming probability: the robustness analysis in Section 4.3 establishes independence of the list computed with the encoder's true seeds from the hash key and bounds collisions by almost-universality.

The hybrids in Section 4.3 prove pseudorandomness. Soundness follows from Lemma 9, since the complete decoder rejects whenever control decoding rejects.

Every transmitted symbol carries one control bit and one index symbol. Choosing a sufficiently large constant payload alphabet makes their $\log(2|\Sigma_S|)$ bits small relative to $\log|\Sigma_C|$, as specified in (3), keeping the rate close to that of the HRW code. With $\eta + \varepsilon \le \zeta$ and the other conditions of Theorem 2, the rate is $R \ge R_* - \zeta$. For fixed channel parameters, Proposition 1 shows that this rate approaches the capacity upper bound obtained by revealing insertion and deletion locations as $\zeta \to 0$ and $q$ grows through the supported alphabet sizes.

# 2 Preliminaries

## 2.1 Notation

- *Indices.* For $n \in \mathbb{N}$, let $[n] := \{1, \dots, n\}$.
- *Sampling and concatenation.* We write $x \leftarrow \mathcal{D}$ for sampling from a distribution $\mathcal{D}$, $x \xleftarrow{\$} S$ for uniform sampling from a finite set $S$, and $x \,\|\, y$ for string concatenation.
- *Distances.* For equal-length strings, $\mathsf{dist}(x, y)$ denotes their Hamming distance; for sets $A, B$, $A \,\Delta\, B$ is their symmetric difference.
- *Asymptotic notation.* All logarithms are base two unless a base is displayed. A function $\mathsf{negl}(t)$ is negligible if it is smaller than $t^{-c}$ for every constant $c > 0$ and all sufficiently large $t$; an event holds with overwhelming probability if its complement has negligible probability. We write $\mathsf{poly}(t)$ for an unspecified polynomial in $t$.
- *Computational indistinguishability.* We write $X \approx_c Y$ for computational indistinguishability.

## 2.2 Hash Function

### 2.2.1 $\varepsilon_H$-Almost-Universal Hashing [Sti94]

A $\varepsilon_H$-almost-universal hash family $\mathcal{H} = \{H_\tau : \mathcal{X} \to \mathcal{Y}\}_{\tau \in \mathcal{K}}$ is $\varepsilon_H$-almost-universal ($\varepsilon_H$-AU) if, for every distinct $x, x' \in \mathcal{X}$,

$$\Pr_{\tau \leftarrow \mathcal{K}}[H_\tau(x) = H_\tau(x')] \le \varepsilon_H.$$

We use polynomial evaluation hashing over $\mathbb{F}_{2^\kappa}$ introduced in [Sar11]. For $x = (x_0, \dots, x_{d-1}) \in (\mathbb{F}_{2^\kappa})^d$ and $\tau \leftarrow \mathbb{F}_{2^\kappa}$, define

$$H_\tau(x) = \sum_{i=0}^{d-1} x_i \tau^i.$$

This family is $\varepsilon_H = (d-1)/2^\kappa$-almost-universal. When $d = \mathsf{poly}(\kappa), \varepsilon_H = \mathsf{negl}(\kappa)$.

## 2.3 Pseudorandom Objects

### 2.3.1 Pseudorandom Generators

Over security parameter $\lambda$, a PRG is a polynomial-time map $G : \{0,1\}^\kappa \to \Sigma^N$ for $\kappa = \Omega(\lambda)$ and $N = \mathsf{poly}(\lambda)$. The PRG's output with a uniform seed as input is computationally indistinguishable from uniform over $\Sigma^N$.

### 2.3.2 Pseudorandom Codes

We follow the definition of public-key pseudorandom codes of Christ and Gunn [CG24]. A public-key PRC over an alphabet $\Sigma$ consists of polynomial-time randomized algorithms

$$(\mathsf{Init}, \mathsf{Enc}, \mathsf{Dec}),$$

where $\mathsf{Init}(1^\lambda)$ outputs the public key $pk$ and the secret key $sk$. All probabilities in the security properties below include the internal random choices of the algorithms used.

Let $\mathcal{M}$ denote the message space, with efficiently represented messages of bit length $\mathsf{poly}(\lambda)$, and let $N = \mathsf{poly}(\lambda)$ be the codeword length. The encoding and decoding algorithms have interfaces

$$\mathsf{Enc}_{pk} : \mathcal{M} \longrightarrow \Sigma^N,$$

$$\mathsf{Dec}_{sk} : \Sigma^* \longrightarrow \mathcal{M} \cup \{\perp\}.$$

- *Pseudorandomness.* For every PPT adversary given $pk$, adaptive oracle access to $\mathsf{Enc}_{pk}(\cdot)$ is computationally indistinguishable from access to an ideal oracle that returns a fresh independent uniform element of $\Sigma^N$ on every query, including repeated queries.

- *Soundness.* For every fixed $c \in \Sigma^*$ independent of $(pk, sk)$,

$$\Pr_{(pk,sk)\leftarrow\mathsf{Init}(1^\lambda)}\big[\mathsf{Dec}_{sk}(c) = \perp\big] = 1 - \mathsf{negl}(\lambda).$$

- *Robustness.* For a channel $\mathcal{E}$, for every message $m$,

$$\Pr[m = \mathsf{Dec}_{sk}(\mathcal{E}(\mathsf{Enc}_{pk}(m)))] \geq 1 - \mathsf{negl}(\lambda),$$

  where the probability is over key generation, encoding, the channel, and decoding randomness.

- *Rate.* The rate is

$$R = \frac{\log |\mathcal{M}|}{N \log |\Sigma|}.$$

## 2.4 The Majority Wrapper

### 2.4.1 Majority Encoding

For odd $m$, $\mathsf{MajEnc}_m(x_1, \ldots, x_n)$ independently samples, for every $j$, a uniform block in $\{0,1\}^m$ conditioned on having majority $x_j$, and concatenates the blocks.

### 2.4.2 Majority Decoding

Given a received string $w$ of length $L$, $\mathsf{MajDec}_n$ sets $E_j = \lfloor jL/n \rfloor$ for $0 \leq j \leq n$. For each $j \in [n]$, its $j$-th output bit is the majority of the substring $w_{E_{j-1}+1} \cdots w_{E_j}$. It outputs 0 when the substring contains equally many zeros and ones, including when the substring is empty.

### 2.4.3 $p$-Bounded Channels

**Definition 1** ($p$-bounded channel)**.** A length-preserving binary channel $\mathcal{E} : \{0,1\}^* \to \{0,1\}^*$ is *$p$-bounded* if, for every $n \in \mathbb{N}$,

$$\Pr_{x \leftarrow \{0,1\}^n} \left[ \mathrm{dist}\big(\mathcal{E}(x), x\big) > pn \right] \le \mathrm{negl}(n),$$

where the probability is taken over the uniform choice of $x$ and the internal randomness of $\mathcal{E}$.

## 2.5 Synchronization and List Recovery

### 2.5.1 List Recovery

Let $\mathcal{C} \subseteq \Sigma_C^N$ be a code. It is $(a, l, L)$-list-recoverable if, given sets $A_1, \ldots, A_N \subseteq \Sigma_C$ of size at most $l$, its list-recovery algorithm outputs at most $L$ codewords and includes every $c \in \mathcal{C}$ satisfying $|\{i : c_i \in A_i\}| \ge aN$.

### 2.5.2 Self-Matching Strings

A monotone matching of a string $S \in \Sigma_S^N$ with itself consists of pairs $(i_r, j_r)$, $r \in [t]$, such that both index sequences are strictly increasing and $S_{i_r} = S_{j_r}$. The string is $\varepsilon_s$-self-matching if every such matching contains fewer than $\varepsilon_s N$ pairs with $i_r \neq j_r$.

### 2.5.3 LCS Repositioning

We write $\mathsf{ListRep}_K$ for the $K$-round longest-common-subsequence (LCS) repositioning procedure of Haeupler and Shahrasbi [HS21], which uses an index string to produce at most $K$ candidate symbols for each transmitted position. Their Theorem 2.8 combines this procedure with a high-rate list-recoverable code: for constants $0 \le \delta < 1$, $\gamma \ge 0$, and $\varepsilon > 0$, it gives a polynomial-time code tolerating $\delta N$ deletions and $\gamma N$ insertions, of rate at least $1 - \delta - \varepsilon$ over a constant-size alphabet, with list size $L(N) = \exp(\exp(\exp(\log^* N)))$.

**HS21 repositioning guarantee.** The analysis in the proof of Theorem 2.8 of Haeupler and Shahrasbi [HS21] gives the following interface. Let $((u_i, S_i))_{i \in [N]}$ be an indexed sequence, where $S \in \Sigma_S^N$ is $\varepsilon_s$-self-matching. If the sequence undergoes at most $\delta N$ deletions and $\gamma N$ insertions, then $\mathsf{ListRep}_K$ outputs candidate sets $A_1, \ldots, A_N$ satisfying $|A_i| \le K$ and

$$|\{i \in [N] : u_i \in A_i\}| \ge \left(1 - \delta - \frac{1+\gamma}{K} - K\varepsilon_s\right) N. \tag{1}$$

## 2.6 Channel Model

We consider a $q$-ary generalization of the Binary Substitution, Insertion, and Deletion (BSID) channel of Briffa, Schaathun, and Wesemeyer [BSW10], derived from the random insertion/deletion channel of Davey and MacKay [DM01]. Let $\Sigma$ be an alphabet of size $q \ge 2$ and let $x = (x_1, \ldots, x_N) \in \Sigma^N$. The channel has parameters $P_i, P_d, P_s \in [0,1]$ with $P_i + P_d < 1$. At the current input position $i$, independently of previous events, it performs one of three actions:

- With probability $P_i$, it outputs $u \xleftarrow{\$} \Sigma$ and remains at position $i$.
- With probability $P_d$, it deletes $x_i$ and advances to $i+1$.
- With probability $1 - P_i - P_d$, it transmits $x_i$ and advances to $i+1$.

Since an insertion does not advance the input position, arbitrarily many inserted symbols may precede any deletion or transmission.

Conditioned on transmission, the output equals $x_i$ with probability $1 - P_s$; for every $a \in \Sigma \setminus \{x_i\}$, it equals $a$ with probability $P_s/(q-1)$. Thus a substitution replaces $x_i$ uniformly by one of the other $q-1$ symbols. The channel maps $x \in \Sigma^N$ to a random output $y \in \Sigma^{N'}$ of random length $N'$. For $q = 2$, every substitution flips the transmitted bit and the model reduces to the binary BSID channel; we call its non-binary extension the *q-ary SID channel*.

# 3 The CG24 Code over the BSID Channel

Christ and Gunn construct a pseudorandom code for the deletion channel by composing a PRC robust to bounded Hamming errors with the majority encoder and decoder $(\mathsf{MajEnc}, \mathsf{MajDec})$ [CG24]. Their original analysis proves this construction robust over the composed channel $\mathrm{BSC}_q \circ \mathrm{BDC}_p$, which independently deletes bits and then flips surviving bits, but does not account for insertions. We instead consider the BSID channel, a natural memoryless channel in which insertions, deletions, and substitutions occur within a single process. Since inserted bits change the received length and shift the block boundaries used by $\mathsf{MajDec}$, the CG24 robustness result does not apply directly. We show that, when $1 - P_i - P_d > 0$, $P_s < \frac{1}{2}$, and $m = \Omega(n \log^6 n)$ is odd, the majority code still recovers more than half of the original coordinates by a constant margin with overwhelming probability. This is the key step that extends the CG24 construction to the BSID channel and is stated formally in the following theorem.

**Theorem 1** (Majority-Voted Code Survives BSID)**.** *Let $\mathcal{E}$ be a BSID channel with parameters $\{P_i, P_d, P_s\}$, and let $x \leftarrow \{0,1\}^n$. If $1 - P_i - P_d > 0$, $P_s < \frac{1}{2}$, and $m = \Omega(n \log^6 n)$ is odd, then there exists a constant $\varepsilon > 0$ such that*

$$\Pr\big[\, \mathsf{dist}(x, \mathsf{MajDec}_n(\mathcal{E}(\mathsf{MajEnc}_m(x)))) \le (\tfrac{1}{2} - \varepsilon)n \,\big] \ge 1 - \mathsf{negl}(n).$$

## 3.1 Channel Geometry and Block Alignment

We begin the proof by sampling $X \xleftarrow{\$} \{0,1\}^n$ and then $Y \leftarrow \mathsf{MajEnc}_m(X)$. We consider what happens when $Y$ goes through the BSID channel $\mathcal{E}$. By Lemma 1 below, $Y$ is also a uniform string. The $i$-th bit of $Y$ is labeled as $Y_i$; when the memoryless channel $\mathcal{E}$ takes $Y_i$ in, the output would be some inserted uniform bits in the front, concatenated with the original $Y_i$ with probability $\frac{(1-P_i-P_d)(1-P_s)}{1-P_i}$, or a flipped $Y_i$ with probability $\frac{(1-P_i-P_d)P_s}{1-P_i}$, or nothing if $Y_i$ is deleted with probability $\frac{P_d}{1-P_i}$. For every $Y_i$ we name the number of inserted uniform bits as $K_i$, where $\Pr[K_i = k] = (1-P_i)P_i^k$, and name the concatenated string channel $\mathcal{E}$ outputs on input $Y_i$ as $G_i$, so $\mathcal{E}(Y) = G_1 \parallel G_2 \parallel \cdots \parallel G_{mn} = G$. For convenience, we name $G_{(j-1)m+1} \parallel G_{(j-1)m+2} \parallel \cdots \parallel G_{jm}$ as $H_j$, to represent the section of $G$ originated from $X_j$.

**Lemma 1** (Uniformity of Majority Encoding)**.** *For uniformly random $X \in \{0,1\}^n$ and odd $m$, $\mathsf{MajEnc}_m(X)$ is uniformly distributed over $\{0,1\}^{mn}$. Hence its $mn$ coordinates may be regarded as i.i.d. uniform bits.*

*Proof.* For odd $m$, exactly $2^{m-1}$ strings in $\{0,1\}^m$ have majority 0, and exactly $2^{m-1}$ have majority 1. The encoder samples each block uniformly conditioned on its majority being the corresponding bit of $X$. Averaging over a uniform input bit therefore makes each block uniform over $\{0,1\}^m$; independence across the input bits and encoder samples makes the concatenation uniform over $\{0,1\}^{mn}$. This is also the pseudorandomness observation in the proof of Theorem 4 of Christ and Gunn [CG24]. □

$\mathsf{MajDec}_n(G)$ will first partition $G$ into $n$ equal length blocks, which we call $S_1, S_2, \ldots, S_n$, with length deviation of 1 at most, then it will solve the majority symbol within each $S_j$ and output them in sequence. Conditioning on any insertion/deletion pattern satisfying (2) for every $j \in [n]$, we will show that each event $\mathsf{MajDec}_1(S_j) = X_j$ has conditional probability at least $\frac{1}{2} + \gamma$ for a fixed constant $\gamma > 0$, and that the agreement events are mutually independent within each residue class of $j$ modulo 3. Chernoff bounds within the three residue classes then prove Theorem 1.

We analyze two properties to prove Theorem 1. First, how much do $S_j$ and $H_j$ overlap and misalign? Second, under such a misalignment, what is the probability that $\mathsf{MajDec}_1(S_j) = X_j$? Besides directly influencing this probability, the misalignment also determines the correlation between agreement events, because correlation arises when a misaligned $S_j$ overlaps with $H_k$ for $k \neq j$. In other words, if $S_j = H_j$ for every $j$, then the events $\mathsf{MajDec}_1(S_j) = X_j$ are independent because the $X_j$ are mutually independent and each $H_j$ depends only on its corresponding input block.

To reveal the misalignment between $S_j$ and $H_j$, we check $H_j$'s property first. We define the expected length of every $G_i$ as $\mu := \mathbb{E}[|G_i|] = \mathbb{E}[K_i] + \mathbb{E}[\mathrm{Ber}(\frac{1-P_i-P_d}{1-P_i})] = \frac{1-P_d}{1-P_i}$, and the cumulative length from $G_1$ to $G_t$ as $C_t := \sum_{i=1}^{t} |G_i|$. The only factor causing the misalignment between $S_j$ and $H_j$ is the accumulated deviation between the actual $|G_i|$ and $\mu$: if every $|G_i|$ equaled $\mu$, then every $H_j$ would have the fixed length $\mu m$ and would be parsed precisely as $S_j$ by $\mathsf{MajDec}$. Here we measure the deviation of $|G_i|$ from $\mu$ first.

**Lemma 2** (Channel Output Length per Input Bit)**.** *There exists constant $C$ and $c$, such that for every integer $x \geq 1$ $\Pr[|G_i| - \mu \geq x] \leq Ce^{-cx}$.*

*Proof.* The expression of $|G_i| - \mu$ is as follows: $|G_i| - \mu = |G_i| - \mathbb{E}[|G_i|] = K_i + \mathrm{Ber}\left(\frac{1-P_i-P_d}{1-P_i}\right) - \mathbb{E}[|G_i|]$. Within $G_i$, $K_i$ follows a geometric distribution, where $\Pr[K_i \geq k] = P_i^k$, thus for $0 < P_i < 1$, $\Pr[|G_i| - \mu \geq x] < \Pr[K_i \geq x - 1] = P_i^{x-1} = \frac{1}{P_i} e^{-\ln(1/P_i)x}$; for $P_i = 0$, the probability is outright zero. Here ln denotes the natural logarithm. □

Lemma 2 demonstrated that $|G_i| - \mu$ is subexponentially bounded in positive $|G_i| - \mu > 0$ direction, while in negative direction $|G_i| - \mu < 0$, the tail cuts off when $|G_i| = 0$ at $|G_i| - \mu = -\mu$. With the zero mean and subexponential nature of $|G_i| - \mu$, we can now bound the accumulated deviation $C_t - t\mu$ by applying Bernstein's Inequality.

**Lemma 3** (Channel Output Length per Input String)**.** *There exists a constant $c > 0$ such that, for $u := c\sqrt{mn}\log n$ and all $j$,*

$$\Pr[|C_{jm} - jm\mu| \geq u] = \mathsf{negl}(n).$$

*Proof.* Because $|G_i| - \mu$ is subexponential, its sub-exponential norm, denoted with $\|\cdot\|_{\psi_1}$, is constant. Set $k := \max\{1, \||G_i| - \mu\|_{\psi_1}\}$, a positive constant upper bound on this norm. We have $C_t - t\mu = \sum_{i=1}^{t}(|G_i| - \mu)$. Per Bernstein's Inequality, there exist a constant $c_1$ such that

$$\Pr\left[\left|\sum_{i=1}^{jm}(|G_i| - \mu)\right| \geq c_2\sqrt{mn}\log n\right] \leq 2\exp\left(-c_1 \min\left\{\frac{c_2^2 n \log^2 n}{jk^2}, \frac{c_2\sqrt{mn}\log n}{k}\right\}\right) = \mathsf{negl}(n). \quad \square$$

Next, we locate each $S_j$ and compare it with $H_j$, thereby quantifying their misalignment.

We distinguish bit values from position labels as follows. Every original bit $Y_i$ has label $i$, and every inserted bit receives a new label distinct from all other bit labels. Let $B'_j := \{(j-1)m+1, \ldots, jm\}$, and let $H'_j$ and $S'_j$ be the sets of labels of the output symbols appearing in $H_j$ and $S_j$, respectively. A deleted original label belongs to $B'_j$ but to neither $H'_j$ nor $S'_j$, while a substituted bit retains its original label. The size $|S'_j \,\Delta\, H'_j|$ counts the misaligned output positions of $S_j$ and $H_j$.

**Lemma 4** ($S_j, H_j$ Misalignment)**.** *For all $j$, the length $|S'_j \,\Delta\, H'_j|$ of the misaligned portion of $S_j$ and $H_j$ satisfies*

$$\Pr\big[|S'_j \,\Delta\, H'_j| \le 4u + 2\big] \ge 1 - \mathsf{negl}(n).$$

*Proof.* By the definition of $\mathsf{MajDec}_n$, the $j$-th boundary of the partition of $G$ is $E_j = \lfloor j|G|/n \rfloor$, so $\left|E_j - \frac{j|G|}{n}\right| < 1$. By Lemma 3, a union bound gives $|C_t - \mu t| \le u$ for $t \in \{(j-1)m, jm, nm\}$ except with probability $\mathsf{negl}(n)$. On this event, we can estimate the distance between the right boundaries of $S_j$ and $H_j$, which are $E_j$ and $C_{jm}$:

$$|E_j - C_{jm}| \le \left|E_j - \frac{j|G|}{n}\right| + \frac{j}{n}\big||G| - \mu mn\big| + |\mu jm - C_{jm}| \le 2u + 1.$$

Therefore,

$$|S'_j \,\Delta\, H'_j| \le |E_{j-1} - C_{(j-1)m}| + |E_j - C_{jm}| \le 4u + 2 = O(u).$$

□

With the length of misalignment between $S_j$ and $H_j$ measured, to evaluate the probability of $\mathsf{MajDec}_1(S_j) = X_j$, we need to determine the number of shared bits between $S_j$ and $\mathsf{MajEnc}_m(X_j) = Y_{(j-1)m+1} \,\|\, Y_{(j-1)m+2} \,\|\, \cdots \,\|\, Y_{jm}$. For convenience, we define $B_j := Y_{(j-1)m+1} \,\|\, Y_{(j-1)m+2} \,\|\, \cdots \,\|\, Y_{jm}$.

Since $\mathcal{E}(B_j) = H_j$, we first determine the number of surviving $B'_j$ bit positions in $H'_j$.

**Lemma 5** (Codeword Bit Positions Surviving the Channel)**.** *Let $a := \frac{1-P_i-P_d}{1-P_i}$. For all $j$,*

$$\Pr\big[\big||B'_j \cap H'_j| - am\big| \le u\big] \ge 1 - \mathsf{negl}(n).$$

*Proof.* Since every original label in $B'_j$ independently has probability $a$ of surviving the channel and becoming part of $H'_j$, $|B'_j \cap H'_j| \sim \mathrm{Bin}(m, a)$. An additive Chernoff bound gives

$$\Pr\big[\big||B'_j \cap H'_j| - am\big| > u\big] \le 2e^{-2u^2/m} = \mathsf{negl}(n).$$

A union bound gives the stated bound simultaneously for all $j$ with overwhelming probability. □

Using Lemmas 4 and 5, we can determine the number of shared bit positions between $B'_j$ and $S'_j$. The choice $m = \Omega(n \log^6 n)$ from Christ and Gunn [CG24] ensures $u = o(m)$.

**Lemma 6** (Codeword Bit Positions in Reconstructed Blocks)**.** *Let $m = \Omega(n \log^6 n)$. For all $j$,*

$$\Pr\big[\big||B'_j \cap S'_j| - am\big| \le 5u + 2\big] \ge 1 - \mathsf{negl}(n).$$

*Proof.* $B'_j \cap S'_j \subseteq B'_j \cap H'_j$, and $(B'_j \cap H'_j) \setminus (B'_j \cap S'_j) \subseteq S'_j \,\Delta\, H'_j$, so

$$|B'_j \cap H'_j| - |S'_j \,\Delta\, H'_j| \le |B'_j \cap S'_j| \le |B'_j \cap H'_j|.$$

By Lemmas 4 and 5, with overwhelming probability,

$$am - 5u - 2 \le |B'_j \cap S'_j| \le am + u.$$

Taking $m = \Omega(n \log^6 n)$ makes $u$ asymptotically smaller than $m$, since $\frac{u}{m} = c\sqrt{\frac{n}{m}} \log n = O\left(\frac{1}{\log^2 n}\right) = o(1)$. The displayed interval proves the claimed deviation bound. Since $u = o(m)$, it also gives $|B'_j \cap S'_j| = am + o(m)$ with overwhelming probability. □

### 3.2 Majority-Vote Agreement

We now use the overlap bound from Lemma 6 to compute the correlation between the majority of $S_j$ and the majority of $B_j$ (representing $X_j$). Let $\mathcal{F}$ denote the channel geometry generated by all insertion counts and all deletion/survival decisions. Conditioning on $\mathcal{F}$ fixes every labeled set $B'_j, H'_j, S'_j$ and every decoder boundary, but leaves the original bit values, inserted bit values, and substitution indicators mutually independent. For a fixed $j$, call a realization $\mathcal{F} = \omega$ good for block $j$ if, with $u$ from Lemma 3,

$$\begin{aligned} |C_t - \mu t| \le u \qquad (t \in \{(j-1)m, jm, nm\}), \\ \big||B'_j \cap H'_j| - am\big| \le u. \end{aligned} \tag{2}$$

These are the output-length conditions of Lemma 3 and the survivor condition of Lemma 5. Fix $j$ and such a geometry $\mathcal{F} = \omega$, and define

$$T_j := B'_j \cap S'_j, \qquad N_j := |T_j|.$$

The proof of Lemma 6 and the equal partition in Lemma 4 give

$$\begin{aligned} |N_j - am| &\le 5u + 2, \\ \big||S'_j| - \mu m\big| &\le \frac{|C_{nm} - \mu nm|}{n} + 2 \le \frac{u}{n} + 2. \end{aligned}$$

Since $u/m = o(1)$ as shown in Lemma 6, we have $N_j = am + o(m)$ and $|S'_j| = \mu m + o(m)$ uniformly over $j$ and geometries satisfying (2). For every $r \in B'_j$, let $W_r \in \{-1, 1\}$ be the sign representation of the original bit at position $r$. For every $r \in T_j$, independently let

$$F_r = \begin{cases} +1, & \text{with probability } 1 - P_s, \\ -1, & \text{with probability } P_s, \end{cases}$$

so the corresponding received bit has sign $F_r W_r$. For every output label $q \in S'_j \setminus T_j$, let $Q_q \in \{-1, 1\}$ denote its received sign. Conditional on $\mathcal{F} = \omega$, the variables $W_r$, $F_r$, and $Q_q$ used below are independent, the $W_r$ and $Q_q$ are uniform Rademacher variables, and $\mathbb{E}[F_r] = 1 - 2P_s$. Define

$$W^{(j)} := \sum_{r \in T_j} W_r, \qquad \widetilde{W}^{(j)} := \sum_{r \in T_j} F_r W_r,$$

$$V^{(j)} := \sum_{r \in B'_j \setminus T_j} W_r, \qquad Z^{(j)} := \sum_{q \in S'_j \setminus T_j} Q_q.$$

Then

$$B_j^{\mathsf{Maj}} = W^{(j)} + V^{(j)}, \qquad S_j^{\mathsf{Maj}} = \widetilde{W}^{(j)} + Z^{(j)}.$$

Therefore, conditional on $\mathcal{F} = \omega$,

$$\mathrm{Var}(B_j^{\mathsf{Maj}}) = m, \qquad \mathrm{Var}(S_j^{\mathsf{Maj}}) = |S'_j| = \mu m + o(m),$$

and only the paired contributions $(W_r, F_r W_r)$ for $r \in T_j$ contribute covariance. Hence

$$\begin{aligned} \mathrm{Cov}(B_j^{\mathsf{Maj}}, S_j^{\mathsf{Maj}} \mid \mathcal{F} = \omega) &= \sum_{r \in T_j} \mathbb{E}[F_r W_r^2] \\ &= (1 - 2P_s) N_j \\ &= (1 - 2P_s) am + o(m). \end{aligned}$$

If we define $a' := (1-2P_s)a$, the conditional correlation coefficient is

$$\begin{aligned}\rho_{j,\omega} &= \frac{a'm + o(m)}{\sqrt{m(\mu m + o(m))}} \\ &= \frac{a'}{\sqrt{\mu}} + o(1) \\ &= \frac{(1-P_i-P_d)(1-2P_s)}{\sqrt{(1-P_i)(1-P_d)}} + o(1).\end{aligned}$$

Finally, with conditional correlation coefficient $\rho_{j,\omega}$ derived as above uniformly over $j$ and geometries $\omega$ good for block $j$, we can compute the probability that $\mathsf{MajDec}_1(S_j)$ equals $X_j$.

**Lemma 7** (Majority-Vote Matching Probability)**.** *For $1-P_i-P_d > 0$, $P_s < \frac{1}{2}$, sufficiently large odd $m$, every $j$, and every geometry $\mathcal{F} = \omega$ satisfying (2) for this $j$, there exists a constant $\gamma > 0$ independent of $j$ and $\omega$ such that*

$$\Pr[\mathsf{MajDec}_1(S_j) = X_j \mid \mathcal{F} = \omega] \geq \frac{1}{2} + \gamma.$$

*Proof.* Throughout this proof, we define sgn(0) to be the sign corresponding to the bit selected by the decoder's fixed tie-breaking rule. Since $m$ is odd, the sign corresponding to $X_j$ equals the sign of $B_j^{\mathsf{Maj}}$. Thus, conditional on $\mathcal{F} = \omega$,

$$\Pr[\mathsf{MajDec}_1(S_j) = X_j \mid \mathcal{F} = \omega] = \Pr[\operatorname{sgn}(S_j^{\mathsf{Maj}}) = \operatorname{sgn}(B_j^{\mathsf{Maj}}) \mid \mathcal{F} = \omega].$$

To find the probability that $S_j^{\mathsf{Maj}}$ and $B_j^{\mathsf{Maj}}$ share the same sign, we apply the central limit theorem to $R_j$, defined as follows:

$$\begin{aligned}R_{j,\omega} &:= \left(\frac{S_j^{\mathsf{Maj}}}{\sqrt{|S'_j|}}, \frac{B_j^{\mathsf{Maj}}}{\sqrt{m}}\right) \\ &= \sum_{r\in T_j}\left(\frac{F_r W_r}{\sqrt{|S'_j|}}, \frac{W_r}{\sqrt{m}}\right) + \sum_{q\in S'_j\setminus T_j}\left(\frac{Q_q}{\sqrt{|S'_j|}}, 0\right) + \sum_{r\in B'_j\setminus T_j}\left(0, \frac{W_r}{\sqrt{m}}\right).\end{aligned}$$

These are independent, centered, bounded vector summands conditional on $\mathcal{F} = \omega$. Since $|S'_j| = \Theta(m)$, their largest Euclidean norm is $O(m^{-1/2})$, meaning all summands uniformly approaches 0, so the Lindeberg condition holds. Define $\rho := \frac{a'}{\sqrt{\mu}}$. The covariance matrices $\begin{pmatrix} 1 & \rho_{j,\omega} \\ \rho_{j,\omega} & 1 \end{pmatrix}$ converge uniformly over good geometries to $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ because $\rho_{j,\omega}$ uniformly converges to $\rho$. Applying the multivariate Lindeberg–Feller central limit theorem, for good $\omega$,

$$R_{j,\omega} \xrightarrow{d} R \sim \mathrm{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Since the convergence holds for every sequence of pairs of $j$ and good $\omega$, the convergence is uniform in $j$ and $\omega$. When $\rho = 1$, then necessarily $P_i = P_d = P_s = 0$. In this case the channel is noiseless, so $\mathsf{MajDec}_1(S_j) = X_j$ with probability one for every $j$. The claimed bound therefore holds. When $0 < \rho < 1$, for such an $R = (R_1, R_2)$, represent its coordinates using two independent standard normal variables $X, Y$ as $R_1 = X$ and $R_2 = \rho X + \sqrt{1-\rho^2}Y$. The event $R_1 > 0, R_2 > 0$ is equivalent to $X > 0$ and

$Y > -\frac{\rho}{\sqrt{1-\rho^2}}X$. These inequalities define a sector in the $(X,Y)$-plane with angular size $\frac{\pi}{2} + \arcsin(\rho)$. By rotational symmetry, its probability is $\frac{1}{4} + \frac{\arcsin\rho}{2\pi}$. Symmetry also gives $\Pr[R_1 < 0, R_2 < 0] = \Pr[R_1 > 0, R_2 > 0]$, so

$$\Pr[\mathrm{sgn}(R_1) = \mathrm{sgn}(R_2)] = \frac{1}{2} + \frac{\arcsin\rho}{\pi}.$$

Since this convergence is uniform over good geometries and the limiting normal distribution assigns zero probability to either coordinate being zero, majority ties contribute only $o(1)$. Therefore, for every good $\omega$,

$$\begin{aligned}\Pr\left[\mathrm{sgn}\left(\frac{S_j^{\mathsf{Maj}}}{\sqrt{|S'_j|}}\right) = \mathrm{sgn}\left(\frac{B_j^{\mathsf{Maj}}}{\sqrt{m}}\right) \;\middle|\; \mathcal{F} = \omega\right] \\ = \Pr[\mathrm{sgn}(S_j^{\mathsf{Maj}}) = \mathrm{sgn}(B_j^{\mathsf{Maj}}) \mid \mathcal{F} = \omega] \\ = \frac{1}{2} + \frac{\arcsin\rho}{\pi} + o(1).\end{aligned}$$

Let $0 < \gamma < \frac{\arcsin\rho}{\pi}$. Uniform convergence then implies that for sufficiently large odd $m$, every $j$, and every $\omega$ good for block $j$,

$$\Pr[\mathsf{MajDec}_1(S_j) = X_j \mid \mathcal{F} = \omega] \geq \frac{1}{2} + \gamma. \qquad \square$$

### 3.3 Global Concentration and the CG24 Corollary

We now prove Theorem 1 and derive its consequence for the CG24 construction.

*Proof of Theorem 1.* Let $\mathsf{Good}$ be the event, measurable with respect to $\mathcal{F}$, on which (2) holds for every $j \in [n]$. Lemmas 3 and 5 give a negligible probability that any required bound fails for a fixed block $j$, uniformly over $j \in [n]$. A union bound therefore gives

$$\begin{aligned}\Pr[\neg\mathsf{Good}] &\leq \sum_{j=1}^{n} \Pr[\text{a required bound fails for block } j] \\ &\leq n \cdot \mathsf{negl}(n) = \mathsf{negl}(n).\end{aligned}$$

Hence $\Pr[\mathsf{Good}] \geq 1 - \mathsf{negl}(n)$. Fix any geometry $\mathcal{F} = \omega \in \mathsf{Good}$. On this geometry,

$$|E_j - C_{jm}| = o(m), \qquad |H'_j| = C_{jm} - C_{(j-1)m} = \mu m + o(m) = \Theta(m)$$

simultaneously for all $j$. Hence, for sufficiently large $m$, neither endpoint of $S_j$ can cross an entire adjacent $H$-block, and therefore $S'_j \subseteq H'_{j-1} \cup H'_j \cup H'_{j+1}$, with the nonexistent boundary blocks omitted. Thus the event $I_j := \mathbf{1}[\mathsf{MajDec}_1(S_j) = X_j]$ depends only on the original bits, inserted bits, and substitution indicators associated with input blocks $j-1, j, j+1$. Conditional on the fixed geometry $\omega$, these underlying random variables are independent across disjoint input-block neighborhoods. In particular, the events $I_j$ and $I_{j'}$ are independent whenever $|j - j'| \geq 3$.

We therefore separate the events into three groups $A^{(1)}, A^{(2)}, A^{(3)}$ according to $j \equiv 0, 1, 2 \pmod 3$. Within each group the events are mutually independent conditional on $\mathcal{F} = \omega$, and the preceding lemma gives $\Pr[I_j = 1 \mid \mathcal{F} = \omega] \geq \frac{1}{2} + \gamma$. Thus, for $l \in \{1, 2, 3\}$, a Chernoff bound gives

$$\Pr\left[\sum_{j \in A^{(l)}} \mathbf{1}[\mathsf{MajDec}_1(S_j) = X_j] < |A^{(l)}| \left(\frac{1}{2} + \frac{\gamma}{2}\right) \;\middle|\; \mathcal{F} = \omega\right] \leq \exp(-\Omega(\gamma^2 n)).$$

Taking a union bound over the three groups, for every fixed good geometry $\omega$ we have

$$\Pr\left[\sum_{j=1}^{n} \mathbf{1}[\mathsf{MajDec}_1(S_j) = X_j] < n\left(\frac{1}{2} + \frac{\gamma}{2}\right) \;\middle|\; \mathcal{F} = \omega\right] \le 3\exp(-\Omega(\gamma^2 n)) = \mathsf{negl}(n).$$

Since this conditional bound is uniform over $\omega \in \mathsf{Good}$,

$$\begin{aligned}\Pr[\text{failure}] &\le \Pr[\neg\mathsf{Good}] + \sup_{\omega\in\mathsf{Good}} \Pr[\text{failure} \mid \mathcal{F} = \omega] \\ &= \mathsf{negl}(n).\end{aligned}$$

Let $\varepsilon = \gamma/2$. We obtain the main theorem when $1 - P_i - P_d > 0$, $P_s < \frac{1}{2}$, and odd $m = \Omega(n \log^6 n)$:

$$\Pr\big[\,\mathsf{dist}(x, \mathsf{MajDec}_n(\mathcal{E}(\mathsf{MajEnc}_m(x)))) \le (\tfrac{1}{2} - \varepsilon)n\,\big] \ge 1 - \mathsf{negl}(n). \qquad \square$$

The Majority-Voted Code Survives BSID Theorem, which parallels Lemma 15 of Christ and Gunn [CG24], immediately implies that their deletion-channel PRC works for the BSID channel under the parameter constraints above.

**Corollary 1** (CG24 Code for BSID)**.** *Let $\mathcal{E}$ be a BSID channel with parameters $\{P_i, P_d, P_s\}$, and security parameter be $\lambda$, and assume that $n$ and $m$ are polynomially bounded in $\lambda$. Let a binary multi-bit pseudorandom code, such as the code of Theorem 3 of Christ and Gunn [CG24], be robust to every key-independent $p$-bounded Hamming channel under rate $\rho_{\mathsf{in}}$ with block length $n \ge \lambda$, under the assumptions required by the inner PRC. If $\frac{1}{2} - \varepsilon < p < \frac{1}{2}$, $1 - P_i - P_d > 0$, $P_s < \frac{1}{2}$, and $m = \Omega(n \log^6 n)$ is sufficiently large and odd, then Construction 6 of Christ and Gunn [CG24] yields a PRC of rate $\rho_{\mathsf{in}}/m$ that is robust to $\mathcal{E}$, computationally pseudorandom, and sound.*

*Proof.* By Theorem 1, $\mathsf{dist}(x, \mathsf{MajDec}_n(\mathcal{E}(\mathsf{MajEnc}_m(x))))$ is $\frac{1}{2} - \varepsilon$ bounded for uniform $x$. The same bound holds when $x$ is a pseudorandom codeword: otherwise, sampling the channel, applying $\mathsf{MajEnc}_m$ and $\mathsf{MajDec}_n$, and testing the resulting distance would distinguish a PRC codeword from a uniform string. Choosing the underlying PRC to tolerate any $p > \frac{1}{2} - \varepsilon$ bounded channel, the robustness of the inner PRC therefore applies to $\mathsf{MajDec}_n(\mathcal{E}(\mathsf{MajEnc}_m(\cdot)))$, proving robustness of the wrapped construction over $\mathcal{E}$ exactly as the proof of Theorem 4 of Christ and Gunn [CG24]. The soundness and pseudorandomness arguments of their Construction 6 are unchanged. $\square$

# 4 A Uniquely Decodable Pseudorandom Code for the $q$-ary SID Channel

We construct a uniquely decodable pseudorandom code under the $q$-ary SID channel $\mathcal{E}'$ by combining the Corollary 1 PRC and the error-correcting code against synchronization errors from Theorem 2.8 of Haeupler and Shahrasbi [HS21]. Its rate is at least $1 - \delta - \varepsilon$.

**Theorem 2.** *Assume the existence of the binary public-key PRC of Corollary 1 and secure pseudorandom generators. Fix constants $P_i, P_d, P_s \in [0,1]$ with $P_i + P_d < 1$ and $P_s < 1$, and $\eta, \varepsilon > 0$ satisfying $\eta + \varepsilon < \frac{(1-P_i-P_d)(1-P_s)}{1-P_i}$. Let $\gamma = \frac{P_i + (1-P_i-P_d)P_s}{1-P_i} + \eta$ and require $8(1+\gamma)/\varepsilon \in \mathbb{Z}$. There exist a finite index alphabet $\Sigma_S$ and an even power $q_0 \ge 16/\varepsilon^2$ of a prime, depending only on these constants, such that for every sufficiently large integer $h_C$, fixed independently of the security parameter $\lambda$, with $\Sigma_C = \mathbb{F}_{q_0^{h_C}}$ and $q = 2|\Sigma_S| q_0^{h_C} > \frac{1}{1-P_s}$, there is a polynomial-time public-key pseudorandom code over $\Sigma_C \times \Sigma_S \times \{0,1\}$ with rate $R \ge 1 - \delta - \varepsilon$, where $\delta = \frac{P_d + (1-P_i-P_d)P_s}{1-P_i} + \eta$. The code is uniquely decodable over*

*the q-ary SID channel with parameters $(P_i, P_d, P_s)$ except with negligible probability in $\lambda$; its codewords are computationally indistinguishable from uniform strings over the code alphabet, and it is sound with overwhelming probability.*

In this construction, the code of Christ and Gunn [CG24] carries the control information including the seeds for a random mask and a self-matching string, a hash key and hash value of the message, through $\mathcal{E}'$ and is decoded first. If the control information is recovered correctly and the generated string is a valid self-matching string, the remaining decoding procedure is the original list decoder from Theorem 2.8 of Haeupler and Shahrasbi [HS21] operating on a $q$-ary SID channel followed by a hash check to select the transmitted message from the list. Section 4.3 proves that the complete decoder succeeds with overwhelming probability.

## 4.1 Parameters

Per input symbol through $\mathcal{E}'$, the expected insertion, deletion, and substitution ratios are respectively $\frac{P_i}{1-P_i}$, $\frac{P_d}{1-P_i}$, and $\frac{(1-P_i-P_d)P_s}{1-P_i}$. Over a length-$N$ codeword, these ratios have $o(1)$ fluctuations with overwhelming probability as proved in the follwing lemma:

**Lemma 8** (Concentration of SID error ratios)**.** *Let $I_N, D_N, S_N$ be the numbers of insertions, deletions, and substitutions produced by feeding $N$ symbols through $\mathcal{E}'$, then with $1-\mathsf{negl}(N)$ probability,*

$$\frac{I_N}{N} = \frac{P_i}{1-P_i} + o(1), \qquad \frac{D_N}{N} = \frac{P_d}{1-P_i} + o(1), \qquad \frac{S_N}{N} = \frac{(1-P_i-P_d)P_s}{1-P_i} + o(1).$$

*Proof.* For each input coordinate, let $K_i$ denote the number of preceding insertions. As observed in the proof of Lemma 2, the $K_i$'s are independent geometric random variables with $\mathbb{E}K_i = \frac{P_i}{1-P_i}$ and constant subexponential norm. The deletion and substitution indicators are independent across input coordinates, bounded, and have respective means $\frac{P_d}{1-P_i}$ and $\frac{(1-P_i-P_d)P_s}{1-P_i}$. Applying Bernstein's inequality as in Lemma 3 to each of the three length-$N$ sums gives deviations $O(\sqrt{N}\log N) = o(N)$ from their expectations, except with probability $e^{-\Omega(\log^2 N)} = \mathsf{negl}(N)$. A union bound over the three counts and division by $N$ demonstrate fluctuation of $o(1)$ and prove then claim.

□

With $N = \mathsf{poly}(\lambda)$ chosen as in "Choosing lengths with random padding" below, we have $N \geq n \geq \lambda$. Hence the $\mathsf{negl}(N)$ failure probability in Lemma 8 is also negligible in $\lambda$.

Since Theorem 2.8 of Haeupler and Shahrasbi [HS21] is formulated in edit distance, we interpret each substitution as one deletion followed by one insertion. Fix a small constant concentration slack $\tau > 0$; with overwhelming probability, the full codeword therefore has insertion and deletion ratios at most $\gamma_0 = \frac{P_i+(1-P_i-P_d)P_s}{1-P_i} + \tau$ and $\delta_0 = \frac{P_d+(1-P_i-P_d)P_s}{1-P_i} + \tau$, respectively.

Each transmitted symbol has a masked payload coordinate, an index coordinate, and a binary coordinate. A full-symbol substitution changes the masked payload and index pair with probability $r = (q-2)/(q-1)$. Thus, with overwhelming probability, the insertion and deletion ratios seen by the HS construction are at most

$$\gamma_{HS} = \frac{P_i + (1-P_i-P_d)rP_s}{1-P_i} + \tau, \qquad \delta_{HS} = \frac{P_d + (1-P_i-P_d)rP_s}{1-P_i} + \tau.$$

For parameter setup, choose another small constant $\eta \geq \tau$ and define the conservative bounds

$$\gamma = \frac{P_i + (1-P_i-P_d)P_s}{1-P_i} + \eta \geq \gamma_{HS}, \qquad \delta = \frac{P_d + (1-P_i-P_d)P_s}{1-P_i} + \eta \geq \delta_{HS}.$$

Finally, choose $\varepsilon > 0$ such that $\eta + \varepsilon < \frac{(1-P_i-P_d)(1-P_s)}{1-P_i}$. With $\gamma$, $\delta$, and $\varepsilon$ fixed and $8(1+\gamma)/\varepsilon \in \mathbb{Z}$ as required in Theorem 2, all parameters involved in the self-matching-string list-recovery construction of Theorem 2.8 in Haeupler and Shahrasbi [HS21] can be determined from the original text, except we re-instantiate the HRW list-recoverable code with a larger maximum input size of $l = \frac{8(1+\gamma)}{\varepsilon}$, rather than using the original $l = \frac{2(1+\gamma)}{\varepsilon}$ appearing in the proof of HS21 Theorem 2.8 [HS21]. This is a direct re-instantiation of HRW Theorem A.7 [HRZW17] (see also HS21 Theorem 2.9 [HS21]), with rate $\rho = 1 - \delta - \varepsilon/2$, input-list size $l$, and HRW error parameter $\varepsilon/4$. HRW measures disagreement: its radius $1 - \rho - \varepsilon/4$ corresponds to the agreement threshold $a = \rho + \varepsilon/4 = 1 - \delta - \varepsilon/4$. Choose an even power of a prime $q_0 \geq 16/\varepsilon^2$. Here $q_0$ is the base-field size and $\Sigma_C = \mathbb{F}_{q_0^{h_C}}$ is the payload alphabet; $h_C$ denotes the extension degree, to distinguish it from the message hash $h$. HRW Theorem A.7 provides an integer threshold

$$h_{\min} = O\left(\frac{l}{\varepsilon^2} \log_{q_0}\left(\frac{4l}{\varepsilon}\right)\right)$$

and applies for every integer $h_C \geq h_{\min}$. For each fixed choice of these parameters, the output-list bound remains $L(N) = \exp(\exp(\exp(\log^* N)))$. In the repositioning algorithm ListRep, the LCS repositioning executes $K = \frac{8(1+\gamma)}{\varepsilon}$ times; and the reference string $S$ is proved later by Lemma 10 to be $\varepsilon_s = \frac{\varepsilon^2}{64(1+\gamma)}$ self-matching with overwhelming probability when $|\Sigma_S| = \Omega\left(\frac{(1+\gamma)^2}{\varepsilon^4}\right) = \Omega\left(\frac{l^2}{\varepsilon^2}\right)$. Applying equation (1) with $u_i = x_i + R_i$ and the chosen parameters shows that the candidate sets contain the correctly positioned masked symbol in at least

$$\left(1 - \delta - \frac{1+\gamma}{K} - K\varepsilon_s\right) N = \left(1 - \delta - \frac{\varepsilon}{4}\right) N = aN$$

positions, as required by the HS21 list-recovery construction.

Choose $|\Sigma_S| = \Theta\left(l^2/\varepsilon^2\right)$ large enough for Lemma 10. For the rate calculation, choose an integer extension degree satisfying

$$h_C \geq \max\left\{h_{\min}, \left\lceil \frac{4\log(2|\Sigma_S|)}{\varepsilon \log q_0} \right\rceil\right\}.$$

The first lower bound ensures applicability of HRW Theorem A.7 [HRZW17]; the second gives

$$\frac{\log(2|\Sigma_S|)}{\log|\Sigma_C|} = \frac{\log(2|\Sigma_S|)}{h_C \log q_0} \leq \frac{\varepsilon}{4}. \tag{3}$$

If $q_0 = \Theta(\varepsilon^{-2})$ and $h_C$ is the smallest integer satisfying these two lower bounds, the estimate for $h_{\min}$ gives

$$\log|\Sigma_C| = h_C \log q_0 = O\left(\frac{l}{\varepsilon^2} \log \frac{4l}{\varepsilon}\right).$$

Using $l = 8(1+\gamma)/\varepsilon$ and the chosen $|\Sigma_S|$, this yields

$$q = 2|\Sigma_C||\Sigma_S| = \left(\frac{1+\gamma}{\varepsilon^2}\right)^{O((1+\gamma)/\varepsilon^3)}.$$

Larger integer $h_C$ are also allowed by HRW Theorem A.7 and satisfy (3); this alphabet-size upper bound applies only to the smallest-degree choice with $q_0 = \Theta(\varepsilon^{-2})$. The formal discussion of information rate is reserved for Section 4.3.

**Choosing lengths with random padding.** After fixing the channel parameters and $\eta, \varepsilon$, choose the index alphabet $\Sigma_S$ and payload alphabet $\Sigma_C = \mathbb{F}_{q_0^{h_C}}$ independently of the security parameter $\lambda$. For each $\lambda$, set the seed length $\kappa = \lambda$ and instantiate the inner multi-bit PRC used in Corollary 1 with message length exactly $4\kappa$ and the required Hamming robustness. Let $n \geq \lambda$ be its actual block length, with $n = \mathsf{poly}(\lambda)$. Independently of the HRW length, choose a sufficiently large odd $m = \Theta(n \log^6 n)$ and set $N_{\mathrm{ctrl}} = nm$. Choose any supported HRW length $N = q_0^{e/2}(\sqrt{q_0} - 1)$ from Theorem A.7 of [HRZW17] such that $N \geq N_{\mathrm{ctrl}}$ and $N$ is polynomially bounded in $\lambda$. Set the public, key-independent parameters

$$d_{\mathrm{pad}} = N - N_{\mathrm{ctrl}}, \qquad p_{\mathrm{PRC}} = \frac{N_{\mathrm{ctrl}}}{N}.$$

Append $d_{\mathrm{pad}}$ independent uniform bits to the control codeword.

**Lemma 9** (Uniform padding and proportional-prefix decoding)**.** *Take the length-$nm$ binary PRC from Corollary 1. Append $N - nm$ independent uniform bits to each codeword, where $N = \mathsf{poly}(\lambda)$ and $N \geq nm$, and set $p_{\mathrm{PRC}} = nm/N$. Given a received binary string of length $N'$, apply the original control decoder to its first $T = \lfloor p_{\mathrm{PRC}} N' \rfloor$ bits: apply $\mathsf{MajDec}_n$ to this prefix, then decode with the inner PRC. This padded code retains robustness over the stated BSID channel, pseudorandomness, and soundness.*

*Proof.* We first analyze the induced channel on a uniformly random input $X \xleftarrow{\$} \{0,1\}^n$. By Lemma 1, $\mathsf{MajEnc}_m(X)$ is uniform; appending independent uniform padding therefore gives a uniform length-$N$ string. Write $C_t$ for the number of bits output by the BSID channel while processing the first $t$ bits of this padded string, and set $\mu = (1 - P_d)/(1 - P_i)$. With $u = C\sqrt{nm}\log n$ for a sufficiently large constant $C$, Lemma 3 and a union bound over $j$ give the first bound in (4) below. For the second, set $B = \max\{1, k\}$, where $k$ is the sub-exponential norm bound in the proof of Lemma 3. The Bernstein inequality used there, applied to all $N$ input bits with the same constant $c_1 > 0$, gives

$$\begin{aligned}\Pr[p_{\mathrm{PRC}}|C_N - \mu N| > u] &\leq 2\exp\left(-c_1 \min\left\{\frac{u^2}{p_{\mathrm{PRC}}^2 N B^2}, \frac{u}{p_{\mathrm{PRC}} B}\right\}\right) \\ &= \mathsf{negl}(n),\end{aligned}$$

since $p_{\mathrm{PRC}}^2 N = (nm)^2/N \leq nm$ and $p_{\mathrm{PRC}} \leq 1$. Thus both bounds hold simultaneously except with probability $\mathsf{negl}(n)$:

$$\max_{0 \leq j \leq n} |C_{jm} - \mu jm| \leq u, \qquad p_{\mathrm{PRC}}|C_N - \mu N| \leq u. \tag{4}$$

On the decoding side, $\mathsf{MajDec}_n$ first partitions the received prefix of length $T = \lfloor p_{\mathrm{PRC}} C_N \rfloor$ into $n$ equal length blocks, which we call $S_1, S_2, \ldots, S_n$, with length deviation of 1 at most. We adapt the boundary calculation in Lemma 4 to the boundaries $E_j = \lfloor jT/n \rfloor$ of these blocks for $0 \leq j \leq n$. Since $|T - p_{\mathrm{PRC}} C_N| < 1$ and $p_{\mathrm{PRC}} N = nm$,

$$\begin{aligned}|E_j - C_{jm}| &\leq \left|E_j - \frac{jT}{n}\right| + \frac{j}{n}|T - \mu nm| + |C_{jm} - \mu jm| \\ &\leq 1 + \frac{j}{n}\left(1 + p_{\mathrm{PRC}}|C_N - \mu N|\right) + |C_{jm} - \mu jm| \\ &\leq 2 + \frac{j}{n} p_{\mathrm{PRC}}|C_N - \mu N| + |C_{jm} - \mu jm| \\ &\leq 2u + 2 = o(m).\end{aligned}$$

The rounding of $E_j$ contributes at most 1, and that of $T$ contributes at most $j/n \leq 1$. Thus the boundary bound $|E_j - C_{jm}|$ changes from the lemma's original $2u + 1$ to $2u + 2$, while

the lemma's $O(u)$ misalignment conclusion is unchanged. For a fixed $j$, Lemma 5 gives $am + o(m)$ undeleted bits from the $j$-th original majority block except with probability $\mathsf{negl}(n)$, where $a = (1 - P_i - P_d)/(1 - P_i)$. Together with the output-length bounds obtained above from Lemma 3 and the boundary estimate adapted from Lemma 4, this gives $|S_j| = \mu m + o(m)$ and, by the argument of Lemma 6, $am + o(m)$ surviving bits from its intended block.

As in the majority-correlation calculation preceding Lemma 7, condition on a realization of the insertion counts and deletion decisions for all $N$ input bits such that, for this fixed $j$,

$$|C_{(j-1)m} - \mu(j-1)m| \le u, \qquad |C_{jm} - \mu jm| \le u,$$
$$|C_{nm} - \mu nm| \le u, \qquad p_{\mathrm{PRC}}|C_N - \mu N| \le u,$$
$$\big||B'_j \cap H'_j| - am\big| \le u.$$

The first four bounds are the estimates derived above using Lemma 3. The last is the survivor-count bound of Lemma 5: $|B'_j \cap H'_j|$ counts the original bits of block $j$ that survive deletion. Since $|T - C_{nm}| \le 2u + 2 = o(m)$ and $|S_n| = \mu m + o(m)$, we have $E_{n-1} = T - |S_n| < C_{nm}$ for sufficiently large $n$. Thus uniform bit padded section appended in the rear can enter only $S_n$, contributing at most $|T - C_{nm}| = o(m)$ bits. Note that surviving padding bits remain independent and uniform after substitution, as do bits inserted while processing the padding. For $j < n$, no padding output enters $S_j$, so the majority-correlation calculation and the majority-agreement bound of Lemma 7 apply unchanged. For $j = n$, the padding signs enter only the noise sum $Z^{(n)}$, independently of the original block's majority statistic $B_n^{\mathsf{Maj}}$. They add only $o(m)$ variance and zero covariance with $B_n^{\mathsf{Maj}}$, preserving the conditional variance $\mu m + o(m)$ and covariance $(1 - 2P_s)am + o(m)$ used in that lemma. The limiting correlation and hence the majority-agreement bound $1/2 + \gamma$ remain unchanged for $j = n$ as well, uniformly over these geometries.

For global concentration, let $\mathsf{Good}_{\mathrm{pad}}$ be the event on which (4) and $\big||B'_j \cap H'_j| - am\big| \le u$ hold for every $j \in [n]$, as in the definition of $\mathsf{Good}$ in the proof of Theorem 1. A union bound over the survivor-count estimates of Lemma 5, together with (4), gives $\Pr[\mathsf{Good}_{\mathrm{pad}}] \ge 1 - \mathsf{negl}(n)$. Fix any realization of all insertion counts and deletion decisions satisfying $\mathsf{Good}_{\mathrm{pad}}$. Padding contributes independent randomness only to $S_n$, preserving conditional independence within each index class modulo 3. The global concentration argument in the proof of Theorem 1 therefore applies unchanged, giving at most $(1/2 - \varepsilon)n$ errors except with probability $\mathsf{negl}(n)$. The composition of $\mathsf{MajEnc}_m$, appending independent uniform padding, the BSID channel, taking the first $T$ received bits, and $\mathsf{MajDec}_n$ maps $n$ bits to $n$ bits without using the PRC key, defining a key-independent Hamming channel. As in Corollary 1, if this channel exceeded the error bound with nonnegligible probability on inner-PRC codewords, simulating it and testing the Hamming distance between its input and output would distinguish codewords from uniform strings. Thus the bound also holds for inner-PRC codewords, and inner-PRC robustness applies.

Appending independent uniform bits preserves pseudorandomness by the same hybrid replacing the unpadded codeword by a uniform string. For soundness, the public prefix crop followed by $\mathsf{MajDec}_n$ maps any fixed key-independent received string to a fixed key-independent input to the inner PRC decoder, whose soundness then applies. □

## 4.2 Construction

### 4.2.1 Building Blocks

- The binary control PRC from Corollary 1, with algorithms PRC.Init, PRC.Enc, and PRC.Dec. Its inner code has actual length $n$ and message length exactly $4\kappa$, and its majority encoding has length $N_{\mathrm{ctrl}} = nm$. The parameters $N, m, d_{\mathrm{pad}}, p_{\mathrm{PRC}}$ are

chosen above, with $N = N_{\mathrm{ctrl}} + d_{\mathrm{pad}}$ and $p_{\mathrm{PRC}} = N_{\mathrm{ctrl}}/N$. Lemma 9 justifies applying this control decoder to the proportional prefix of the received binary coordinate.

- The HRW list recoverable code used in Theorem 2.8 of Haeupler and Shahrasbi [HS21] and re-initialized in Section 4.1: $\mathcal{C}_{\mathrm{HRW}} \subseteq \Sigma_C^N : \{\mathsf{Enc}_{\mathsf{HRW}}, \mathsf{ListRec}_{\mathsf{HRW}}\}$, with rate at least $\rho$, block length $N$, minimum list recovery agreement ratio $a$, maximum input list size $l$ and maximum output list size $L(N)$. Let $\mathcal{M}$ be the message space of $\mathsf{Enc}_{\mathrm{HRW}}$. The algorithm $\mathsf{ListRec}_{\mathrm{HRW}}$ outputs a list of at most $L(N)$ messages in $\mathcal{M}$.
- PRG $G_0() : \{0,1\}^\kappa \to \Sigma_C^N$.
- PRG $G_1() : \{0,1\}^\kappa \to \Sigma_S^N$.
- The polynomial hash family from Section 2.2, $\{H_\tau : \mathcal{M} \to \{0,1\}^\kappa\}_{\tau \in \{0,1\}^\kappa}$, with $\varepsilon_H = \mathsf{negl}(\lambda)$.
- The payload alphabet $\Sigma_C$ is a finite additive group, so masking and unmasking are well-defined.
- The HS21 repositioning procedure $\mathsf{ListRep}_K$ returns candidate sets $A_i \subseteq \Sigma_C$ with $|A_i| \le K$.
- The alphabet size of the codeword is $q = 2|\Sigma_C||\Sigma_S| = 2q_0^{h_C}|\Sigma_S|$, with the choices and alphabet-size bound given in Section 4.1.

#### 4.2.2 Initialization $\mathsf{Init}(1^\lambda)$

- $(pk, sk) \leftarrow \mathsf{PRC.Init}(1^\lambda)$.

#### 4.2.3 Encoding $\mathsf{Enc}_{pk}(m)$

- Sample three independent uniform strings for random-mask, self-matching-string generation, and hash key: $\sigma_0, \sigma_1, \tau \xleftarrow{\$} \{0,1\}^\kappa$, where $\kappa = \lambda$.
- Compute hash value of the message $m$: $h = H_\tau(m) \in \{0,1\}^\kappa$.
- Encode the message with the HRW code: $x = \mathsf{Enc}_{\mathrm{HRW}}(m) = (x_1, \ldots, x_N) \in \Sigma_C^N$.
- Generate the random mask and self-matching string: $R = G_0(\sigma_0) = (R_1, \ldots, R_N) \in \Sigma_C^N, S = G_1(\sigma_1) = (S_1, \ldots, S_N) \in \Sigma_S^N$.
- Encode the control information with the binary PRC:

$$\widetilde{b} = \mathsf{PRC.Enc}_{pk}(\sigma_0 \| \sigma_1 \| \tau \| h) \in \{0,1\}^{N_{\mathrm{ctrl}}}.$$

  Independently sample $Z \xleftarrow{\$} \{0,1\}^{d_{\mathrm{pad}}}$ and set $b = \widetilde{b} \| Z \in \{0,1\}^N$.
- Output the codeword $c = (c_1, \ldots, c_N)$ defined coordinate-wise by $c_i = (x_i + R_i, S_i, b_i) \in \Sigma_C \times \Sigma_S \times \{0,1\}$.

#### 4.2.4 Decoding $\mathsf{Dec}_{sk}(c')$

- Let the corrupted codeword be $c' = \big((y'_1, z'_1, b'_1), \ldots, (y'_{N'}, z'_{N'}, b'_{N'})\big)$.
- Set $T = \lfloor p_{\mathrm{PRC}} N' \rfloor$ and decode the proportional prefix of the binary coordinate:

$$u' \leftarrow \mathsf{PRC.Dec}_{sk}(b'_1 \cdots b'_T).$$

  This applies $\mathsf{MajDec}_n$ to the prefix and then the inner decoder, as justified by Lemma 9. Only the binary control input is cropped; payload/index recovery below uses all $N'$ received pairs $(y'_j, z'_j)$. If the result is $\perp$, output $\perp$ and abort. Otherwise, parse $u'$ as $(\widehat{\sigma}_0, \widehat{\sigma}_1, \widehat{\tau}, \widehat{h})$.
- Regenerate the random mask and self-matching string: $\widehat{R} = G_0(\widehat{\sigma}_0), \widehat{S} = G_1(\widehat{\sigma}_1)$.
- Run the modified repositioning algorithm from Theorem 2.8 of Haeupler and Shahrasbi [HS21]:

$$(A_1, \ldots, A_N) = \mathsf{ListRep}_K\left(\widehat{S}, \big((y'_j, z'_j)\big)_{j \in [N']}\right).$$

- Unmask each positional candidate: $B_i := A_i - \widehat{R}_i = \{a - \widehat{R}_i : a \in A_i\}$.
- Compute the list of HRW messages: $\mathcal{L} \leftarrow \mathsf{ListRec}_{\mathrm{HRW}}(B_1, \ldots, B_N)$, where $\mathcal{L} \subseteq \mathcal{M}$.
- Finally, for every entry in $\mathcal{L}$, output the first $\widehat{m}_k$ such that $H_{\widehat{\tau}}(\widehat{m}_k) = \widehat{h}$; if no matching $\widehat{m}_k$ is detected, output $\perp$.

### 4.3 Analysis

In this section we discuss the robustness on $q$-ary SID channel $\mathcal{E}'$, pseudorandomness to PPT adversary, soundness, and rate of the construction above.

#### 4.3.1 Robustness

The $q$-ary SID channel $\mathcal{E}'$ satisfies $1 - P_i - P_d > 0$ and $P_s < 1$. Projecting each received symbol onto its binary coordinate preserves all insertion and deletion decisions, and each inserted symbol projects to an independent uniform bit. For every transmitted symbol, exactly $q/2$ of the $q-1$ possible substitution targets have the opposite binary coordinate, regardless of its payload and index coordinates. Together with the independence of the channel randomness across input positions, this shows that the binary projection is a BSID channel with parameters $\{P_i, P_d, qP_s/(2(q-1))\}$. Since $P_s < 1$ and $q > \frac{1}{1-P_s}$ implies $\frac{(q/2)P_s}{q-1} < \frac{1}{2}$, Lemma 9 proves robustness of the padded binary control code with proportional-prefix decoding, and hence $\Pr[(\widehat{\sigma}_0, \widehat{\sigma}_1, \widehat{\tau}, \widehat{h}) = (\sigma_0, \sigma_1, \tau, h)] \geq 1 - \mathsf{negl}(\lambda)$. With the correct seeds recovered, Theorem 2.8 of Haeupler and Shahrasbi [HS21] applies because the actual projected edit ratios satisfy $\gamma_{HS} \leq \gamma$ and $\delta_{HS} \leq \delta$, the parameters obey $1 - \delta - \varepsilon > 0$, and $S = G_1(\sigma_1)$ is $\varepsilon_s$ self-matching with overwhelming probability. Indeed, by the choice of $\eta$ and $\varepsilon$ above, $1 - \delta - \varepsilon = \frac{(1-P_i-P_d)(1-P_s)}{1-P_i} - \eta - \varepsilon > 0$.

Next, we examine whether the pseudorandom string $S$ is $\varepsilon_s$ self-matching. By Chapter 2.1.1 of HS21, a uniform random string of length $N$ has probability at most $\left(\frac{e^2}{|\Sigma|\varepsilon^2}\right)^{N\varepsilon}$ of failing to be $\varepsilon$ self-matching. Thus, for $|\Sigma| = \Omega(\varepsilon^{-2})$ with a sufficiently large hidden constant, the failure probability is $\mathsf{negl}(N)$. With this fact we can prove the following lemma.

**Lemma 10** ($S$ Is $\varepsilon_s$ Self-Matching)**.** *When* $|\Sigma_S| = \Omega\left(\frac{(1+\gamma)^2}{\varepsilon^4}\right)$*,* $S$ *is* $\varepsilon_s$ *self-matching with probability at least* $1 - \mathsf{negl}(\lambda)$.

*Proof.* For a length-$N$ uniform random string $U^N$ over $\Sigma_S$, the conclusion above implies that it is $\varepsilon_s$ self-matching with probability at least $1 - \mathsf{negl}(N)$, because $\varepsilon_s^{-2} = \frac{4096(1+\gamma)^2}{\varepsilon^4}$ and $|\Sigma_S| = \Omega(\varepsilon_s^{-2})$.

For $S = G_1(\sigma_1)$, suppose instead that the probability of failing the $\varepsilon_s$ self-matching property exceeded the uniform-string failure probability by a non-negligible function of $\lambda$. Since $N = \mathsf{poly}(\lambda)$ and the self-matching property can be checked in time $\mathsf{poly}(N)$ [HS21], this checker would distinguish $G_1(U^\kappa)$ from $U^N$ with non-negligible advantage, contradicting the security of $G_1$. Therefore $S$ is $\varepsilon_s$ self-matching with probability at least $1 - \mathsf{negl}(\lambda)$. □

Next, we discuss that $m$ can be recovered from $\mathcal{L}$ with overwhelming probability. Fix a message $m$. Let $\mathcal{L}^*$ be the list obtained by keeping only $(y'_1, z'_1), \ldots,$ $(y'_{N'}, z'_{N'})$ from $c'$ and running repositioning, unmasking, and HRW list recovery over it with the encoder's true seeds $\sigma_0, \sigma_1$, even when control recovery fails.

Under the channel model in Section 2.6, a substitution gives the original payload and index pair with probability $1/(q-1)$ and each other pair with probability $2/(q-1)$, regardless of the binary coordinate. Since insertions and deletions are input-independent and the transmitted pairs depend only on $m, \sigma_0, \sigma_1$, the ideal list $\mathcal{L}^*$ is independent of $\tau$ for fixed true seeds. Since this independence holds for every fixed seed pair, and the encoder samples the seeds independently of $\tau$, it also holds when the seeds are sampled as in the encoder.

Since $|\mathcal{L}^*| \le L(N)$, almost-universality of the hash and a union bound give

$$\Pr[\exists m' \in \mathcal{L}^* \setminus \{m\} : H_\tau(m') = H_\tau(m)] \le L(N)\varepsilon_H.$$

Let $E$ denote the event that all four control fields are recovered correctly, and let $E^c$ denote its complement. On $E$, the actual decoder uses the correct $\sigma_0, \sigma_1$, so $\mathcal{L} = \mathcal{L}^*$, using the same randomness in both list computations if needed. Its hash key and tag also equal $\tau$ and $H_\tau(m)$. Thus the probability that $E$ occurs and an incorrect candidate passes the actual hash check is at most $L(N)\varepsilon_H$. This uses the unconditional collision bound above and does not require independence after conditioning on $E$.

When the index string is $\varepsilon_s$-self-matching and the edit counts stay within the chosen bounds, list recovery with the true seeds places $m$ in $\mathcal{L}^*$. The preceding bounds on these events and on control recovery give $\Pr[m \notin \mathcal{L}^*] = \mathsf{negl}(\lambda)$ and $\Pr[E^c] = \mathsf{negl}(\lambda)$. Moreover, $L(N) \le \mathsf{poly}(N)$, $N = \mathsf{poly}(\lambda)$, and $\varepsilon_H = \mathsf{negl}(\lambda)$, so $L(N)\varepsilon_H = \mathsf{negl}(\lambda)$. The actual decoder succeeds whenever $E$ occurs, $m \in \mathcal{L}^*$, and no incorrect candidate has the same hash value. Hence

$$\begin{aligned}&\Pr[\mathsf{Dec}_{sk}(\mathcal{E}'(\mathsf{Enc}_{pk}(m))) \neq m] \\ &\quad \le \Pr[E^c] + \Pr[m \notin \mathcal{L}^*] + L(N)\varepsilon_H = \mathsf{negl}(\lambda).\end{aligned}$$

### 4.3.2 Pseudorandomness

Fix a PPT adversary $A$ given $pk$ and making at most $Q = \mathsf{poly}(\lambda)$ adaptive message queries. Every game samples $(pk, sk) \leftarrow \mathsf{PRC.Init}(\lambda)$ once. The real oracle returns $\mathsf{Enc}_{pk}(m)$, whereas the ideal oracle returns a fresh independent uniform string in $\Sigma^N$, where $\Sigma = \Sigma_C \times \Sigma_S \times \{0,1\}$ and $|\Sigma| = q$.

*Game* $H_0$. Answer every query $m$ with $\mathsf{Enc}_{pk}(m)$, with fresh independent seeds $\sigma_0, \sigma_1$, fresh hash key $\tau$, and fresh encoding randomness. Write $x = \mathsf{Enc}_{\mathrm{HRW}}(m)$.

*Game* $H_1$. Replace every unpadded binary PRC codeword by an independent uniform string of length $N_{\mathrm{ctrl}}$ while retaining the independent $d_{\mathrm{pad}}$ padding bits. Their

concatenation is $U_B \xleftarrow{\$} \{0,1\}^N$. A distinguisher for the underlying PRC oracle simulates $A$ by passing it $pk$, sampling the two seeds and hash key on each query, submitting $\sigma_0\|\sigma_1\|\tau\|H_\tau(m)$ to its oracle, appending fresh independent uniform padding to the oracle reply, and computing the other two components locally. Thus the difference in $A$'s acceptance probabilities between $H_0$ and $H_1$ is at most $\varepsilon_{\mathsf{PRC}}(\lambda) = \mathsf{negl}(\lambda)$ by the Corollary 1 binary PRC's pseudorandomness.

*Game* $H_2$. Starting from $H_1$, replace every mask $G_0(\sigma_0)$ by $U_C \xleftarrow{\$} \Sigma_C^N$. Use intermediate games replacing the first $t$ mask outputs, for $t = 0, \ldots, Q$. A $G_0$ distinguisher embeds its challenge at a uniformly chosen query and simulates all other replies, continuing to answer adaptive queries. This is possible because $\sigma_0$ is no longer used in any other component. If $\varepsilon_0(\lambda)$ is this distinguisher's negligible advantage, the transition costs at most $Q\varepsilon_0(\lambda)$.

*Game* $H_3$. Starting from $H_2$, replace every index string $G_1(\sigma_1)$ by $U_S \xleftarrow{\$} \Sigma_S^N$. The same query-by-query argument applies to $G_1$, since $\sigma_1$ is no longer used in any other component. If $\varepsilon_1(\lambda)$ is the corresponding distinguisher's negligible advantage, this transition costs at most $Q\varepsilon_1(\lambda)$.

In $H_3$, conditional on $pk$, the prior transcript, and the current query, the reply is $(x + U_C, U_S, U_B)$. Translation by $x$ is a bijection of $\Sigma_C^N$, so this reply is uniform over $\Sigma^N$, independently of all previous replies. Hence $H_3$ is exactly the ideal oracle, whose $Q$ replies are jointly independent uniform $q$-ary strings. The total distinguishing advantage is at most $\varepsilon_{\mathsf{PRC}}(\lambda) + Q(\varepsilon_0(\lambda) + \varepsilon_1(\lambda)) = \mathsf{negl}(\lambda)$, proving the required oracle pseudorandomness.

### 4.3.3 Soundness

Fix a received $q$-ary string independent of the keys. Its binary projection, its public length-$\lfloor p_{\mathrm{PRC}} N' \rfloor$ prefix, and the result of applying $\mathsf{MajDec}_n$ to that prefix are all fixed independently of the keys. By inner-PRC soundness, the control decoder rejects except with probability $\mathsf{negl}(\lambda)$, as in Lemma 9. The complete decoder aborts whenever control decoding rejects, so it has the same soundness bound.

### 4.3.4 Rate

For the HRW payload code over $\Sigma_C$, we have $\rho \geq 1 - \delta - \varepsilon/2$. For the message space $\mathcal{M}$ of $\mathsf{Enc}_{\mathrm{HRW}}$, $\log|\mathcal{M}| \geq \rho N \log|\Sigma_C|$. Each transmitted symbol belongs to $\Sigma_C \times \Sigma_S \times \{0,1\}$, whose size is $q = 2|\Sigma_C||\Sigma_S|$. Thus the rate of the complete construction is

$$R = \frac{\log|\mathcal{M}|}{N \log q} \geq \frac{\rho N \log|\Sigma_C|}{N \log(2|\Sigma_C||\Sigma_S|)} = \frac{\rho}{1 + \dfrac{\log(2|\Sigma_S|)}{\log|\Sigma_C|}}.$$

By (3) and $0 < \rho \leq 1$,

$$R \geq \frac{\rho}{1+\varepsilon/4} \geq \rho - \frac{\varepsilon}{4} \geq 1 - \delta - \frac{3\varepsilon}{4} \geq 1 - \delta - \varepsilon = \frac{(1-P_i-P_d)(1-P_s)}{1-P_i} - \eta - \varepsilon.$$

*Proof of Theorem 2.* Polynomial running time follows from the corresponding guarantees of the underlying components. Robustness, pseudorandomness, and soundness are established in preceding discussions, and the rate calculation above gives $R \geq 1 - \delta - \varepsilon$. Together these prove the theorem. □

Finally, we compare the achieved rate $R$ with a capacity upper bound obtained by revealing insertion and deletion locations.

**Proposition 1** (Genie-aided capacity upper bound)**.** *Let $C_{\mathrm{SID}}^{(q)}$ denote the capacity of the $q$-ary SID channel and $C_{\mathrm{revealed}}^{(q)}$ its capacity when the receiver is told all insertion and deletion locations, both measured in $q$-ary symbols per input symbol. Then*

$$C_{\mathrm{SID}}^{(q)} \le C_{\mathrm{revealed}}^{(q)} = \frac{1-P_i-P_d}{1-P_i}\big(1-h_q(P_s)\big),$$

*where*

$$h_q(p) = p\log_q(q-1) - p\log_q p - (1-p)\log_q(1-p).$$

*For every constant $\zeta > 0$, one can choose $\eta, \varepsilon > 0$ satisfying the conditions in Section 4.1 and $\eta + \varepsilon < \zeta/2$ so that, for all sufficiently large admissible alphabet sizes $q$, the code of Theorem 2 has rate $R \ge C_{\mathrm{revealed}}^{(q)} - \zeta$.*

*Proof.* Under the independent channel model in Section 2.6, revealing insertion and deletion locations lets the receiver discard inserted symbols and identify deleted input positions. Each input symbol survives with probability $(1-P_i-P_d)/(1-P_i)$ and, if it survives, passes through a $q$-ary symmetric channel with substitution probability $P_s$. The inserted symbols and revealed locations are independent of the input. Uniform input achieves capacity $1-h_q(P_s)$ on the symmetric channel; multiplying by the survival probability gives the displayed value of $C_{\mathrm{revealed}}^{(q)}$ per input symbol. Revealing these locations cannot decrease capacity, so $C_{\mathrm{SID}}^{(q)} \le C_{\mathrm{revealed}}^{(q)}$.

The rate bound in Theorem 2 gives

$$C_{\mathrm{revealed}}^{(q)} - R \le \eta + \varepsilon + \frac{1-P_i-P_d}{1-P_i}\big(P_s - h_q(P_s)\big).$$

Choose $\eta, \varepsilon$ as in the proposition. Since $h_q(P_s) \to P_s$ as $q \to \infty$, the last term has absolute value less than $\zeta/2$ for all sufficiently large admissible $q$. Together with $\eta + \varepsilon < \zeta/2$, this gives the claimed rate gap. □

# References


[BSW10] J. A. Briffa, H. G. Schaathun, and S. Wesemeyer. An improved decoding algorithm for the Davey–MacKay construction. In *IEEE International Conference on Communications*, pages 1–5. IEEE, 2010. doi:10.1109/ICC.2010.5502293.

[CG24] M. Christ and S. Gunn. Pseudorandom error-correcting codes. In *Advances in Cryptology—CRYPTO 2024, Part VI*, volume 14925 of *LNCS*, pages 325–347. Springer, 2024. doi:10.1007/978-3-031-68391-6_10.

[CGG+26] M. Christ, N. Golowich, S. Gunn, A. Moitra, and D. Wichs. Improved pseudorandom codes from permuted puzzles. In *Proceedings of the 58th Annual ACM Symposium on Theory of Computing (STOC)*, pages 2130–2139. ACM, 2026. doi:10.1145/3798129.3800916.

[DL26] F. Dehmel and S. Li. A public-key-dependent adversarial-deletion ceiling for fixed-alphabet multi-bit pseudorandom codes. *CoRR*, abs/2609.02943, 2026. doi:10.48550/arXiv.2609.02943.

[DM01] M. C. Davey and D. J. C. MacKay. Reliable communication over channels with insertions, deletions, and substitutions. *IEEE Transactions on Information Theory*, 47(2):687–698, 2001. doi:10.1109/18.910582.

[GM24] N. Golowich and A. Moitra. Edit distance robust watermarks via indexing pseudorandom codes. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pages 20645–20693, 2024. doi:10.52202/079017-0652.

[HLMZ26] S. Huang, X. Li, S. Mao, and Z. Zhou. High-rate public-key pseudorandom codes for edit errors. *CoRR*, abs/2605.19402, 2026. doi:10.48550/arXiv.2605.19402.

[HRZW17] B. Hemenway, N. Ron-Zewi, and M. Wootters. Local list recovery of high-rate tensor codes and applications. In *58th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 204–215. IEEE, 2017. doi:10.1109/FOCS.2017.27.

[HS21] B. Haeupler and A. Shahrasbi. Synchronization strings and codes for insertions and deletions—a survey. *IEEE Transactions on Information Theory*, 67(6):3190–3206, 2021. doi:10.1109/TIT.2021.3056317.

[Sar11] P. Sarkar. A trade-off between collision probability and key size in universal hashing using polynomials. *Designs, Codes and Cryptography*, 58:271–278, 2011. doi:10.1007/s10623-010-9408-6.

[Sti94] D. R. Stinson. Universal hashing and authentication codes. *Designs, Codes and Cryptography*, 4(4):369–380, 1994. doi:10.1007/BF01388651.