Title

Targeted Sampling to Overcome Data Availability Attacks in Blockchain Systems: Joint Code-Sampler Design for 2D Reed-Solomon Codes

ISBN

9798190936667

Author

Madisetti, Sonali

Publication Date

2026-09-11

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Targeted Sampling to Overcome Data Availability Attacks in Blockchain Systems:

Joint Code-Sampler Design for 2D Reed-Solomon Codes

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Electrical and Computer Engineering

by

Sonali Madisetti

2026

ABSTRACT OF THE THESIS

Targeted Sampling to Overcome Data Availability Attacks in Blockchain Systems:
Joint Code-Sampler Design for 2D Reed-Solomon Codes

by

Sonali Madisetti
Master of Science in Electrical and Computer Engineering
University of California, Los Angeles, 2026
Professor Lara Dolecek, Chair

In blockchain systems, light nodes are vulnerable to *data availability* (DA) attacks, in which a malicious block producer withholds block data from the network. To protect against DA attacks, block data is encoded with an erasure code, and each light node performs *data availability sampling* (DAS), in which a node randomly samples a small set of coded symbols to test whether data is available to the network. Two-dimensional Reed-Solomon (2D-RS) codes are a leading choice for use in DAS systems.

Standard DAS analysis treats the erasure code and sampling strategy separately. The undecodable erasure patterns of a 2D-RS code have structured geometry that can be exploited using joint code-sampler design. We measure each sampler design by a light node's *probability of failure* to detect a DA attack. Holding the code fixed, we design a *biregular sampler* which spreads sample queries evenly across rows and columns. In the regime where light nodes take few samples, we prove that the biregular sampler minimizes probability of failure against minimum-distance erasure patterns. We then derive upper bounds for the worst-case probability of failure over all undecodable erasure patterns. We then study our

samplers asymptotically, where we characterize the exponential decay rate of probability of failure and identify an explicit code parameter regime in which minimum-distance erasure patterns are provably worst-case. Holding the sampler fixed, we design a sampler-aware 2D-RS subcode whose added parity checks improve worst-case probability of failure behavior for our sampler. Our joint sampler-subcode design shows that worst-case probability of failure can be improved by targeting a code's structural properties beyond the minimum distance.

The thesis of Sonali Madisetti is approved.

Danijela Cabric

Ian Roberts

Lara Dolecek, Committee Chair

University of California, Los Angeles

2026

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

I thank my advisor, Professor Lara Dolecek, for many helpful discussions and for feedback on the manuscript. I am especially grateful to my labmate Lev Tauz, whose input and careful reading improved this work throughout. His insights and guidance made this work much stronger.

I thank my family for their support and encouragement.

Chapter 3 extends and generalizes results first presented in S. Madisetti, L. Tauz, and L. Dolecek, "Data Availability Attacks: Targeted Sampling for 2-D Reed-Solomon Codes in Blockchain Systems," in Proc. 2025 IEEE Information Theory Workshop (ITW), Sep. 2025, pp. 758–763, doi:10.1109/ITW62417.2025.11240529. A journal version is in preparation for publication.

This work was supported in part by NSF-BSF:CIF grant no. 2330309 and NSF:CIF grant no. 2312872.

CHAPTER 1

Introduction

1.1 Blockchain Scalability

Blockchain systems are decentralized, immutable, and tamper-proof, leading to large impacts in decentralized finance [Nak], healthcare systems [Abd23], voting systems [JAS21], and supply chain management systems [KKA25]. However, blockchain systems struggle with scalability in both storage overhead and throughput [RCS24]. Ethereum's main layer can process a substantially lower number of transactions than VISA or other conventional payment methods [Kil]. Any full node joining Ethereum must verify and custody up to 1 TB of transactions and maintain the current blockchain state, leading to prohibitive hardware requirements on full nodes. Improving blockchain scalability is critical for a wider adoption of blockchain systems.

Using lightweight nodes in blockchain systems is key for blockchain scalability. A *light node* only stores a block's header and does not store the entire ledger of transactions. Light nodes lower both storage and computational costs, which allow for more participants in the network to preserve system decentralization. However, light nodes compromise blockchain security. Without access to full block data, light nodes cannot directly verify transaction validity or whether a transaction is available to the network. Full nodes must generate and disseminate *fraud proofs* to light nodes in the network if any transaction is found to be incorrect. However, light nodes must individually determine whether data is available to the honest full nodes in the network.

1.2 Data Availability

The *data availability* problem for light nodes, or for any client with partial block data (e.g. sharded blockchain systems), is to verify that all data is fully available to the network. Without any assurances of data availability, light nodes may be susceptible to adversarial attacks. A *Data Availability (DA)* attack occurs when a malicious block producer intentionally withholds data from the network [ASB21]. A light node may then be tricked into accepting a block with missing (and potentially invalid) transactions. For example, at the 2016 MIT Expo, Peter Todd demonstrated that a light node could be tricked into thinking that he owns over 21 million bitcoin [Tod16]. Each light node must be able to individually ascertain whether data is available to the network without downloading the entire proposed block.

A simple and popular approach to provide data availability assurances for light nodes is *Data Availability Sampling* (DAS), in which each light node requests small random portions of the proposed block data [ASB21, HSW23]. Each returned symbol provides probabilistic assurances about a block's availability. A light node's *probability of failure* is the probability that the node accepts unavailable block data as available.

An adversarial node may selectively withhold only a tiny fraction of the block to avoid detection by light nodes with high probability. To prevent adversarial nodes from only hiding a small fraction of the block, DAS systems encode the block data with an *erasure code*, which forces an adversarial node to withhold a larger fraction of the block data to prevent block recovery [ASB21, BNN25, YSL20, HSW23].

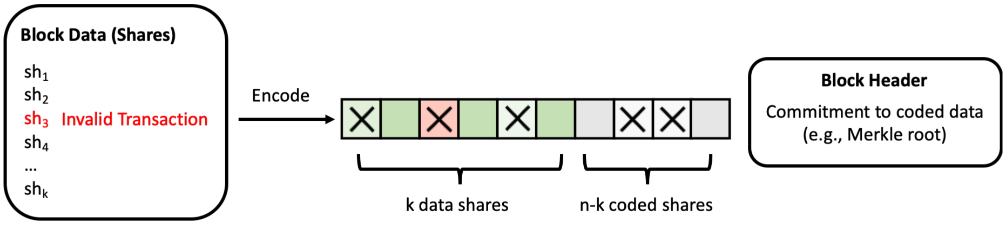
In Figure 1.1, we visualize a DA attack. In Figure 1.1(a), a malicious block producer encodes k uncoded transaction shares into n coded shares using an erasure code with minimum distance $d_{\min}$. The adversary then selects $d_{\min}$ codeword symbols (indicated by an X) to withhold from the network. Then, the block producer generates a commitment (e.g., a Merkle root) to the coded block data. In Figure 1.1(b), the malicious block producer sends the proposed block (without the erased shares) to the full nodes on the network. The light nodes receive only the block header. Each honest full node is unable to generate fraud-proofs for the missing shares. In Figure 1.1(c), a light node queries the block producer with index positions for requested coded symbols. If all symbols are returned, the light node incorrectly accepts the block as available.

1.3 The Coding-Sampling Problem for DAS Systems

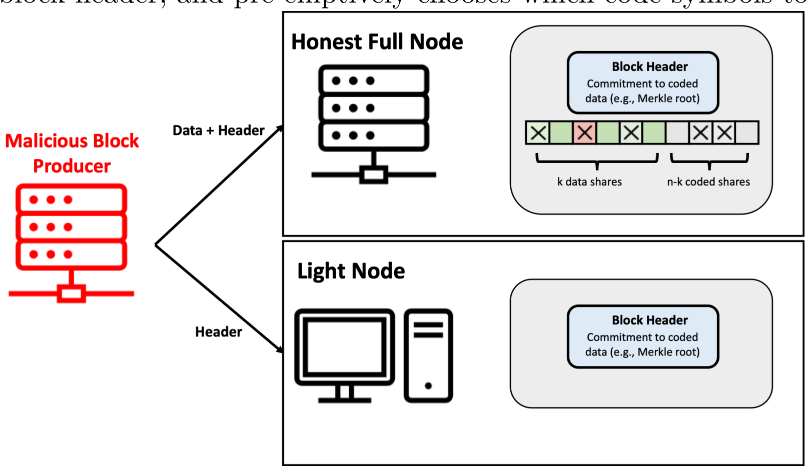
Codes for DAS systems aim for high code distance, a low encoding-decoding complexity, efficient cryptogrpahic commitment compatibility, and cheap symbol repair. Reed-Solomon codes were first proposed as MDS codes with high code distance, and are currently being implemented in Ethereum PeerDAS [ASB21, Eth26]. However, the cost of incorrect coding proofs or single-symbol repair was too prohibitive. A wide variety of codes have been considered for DAS systems operating among the set of trade-offs [ASB21, MTD23, SVH25, BNN25].

2D Reed-Solomon codes, which are product codes with Reed-Solomon row and column codes, are planned for implementation in Ethereum [KB25, KBR24a], and have been widely studied in the literature [SRB22, BST26, HSW23]. 2D Reed-Solomon codes are prized for simple encoding and decoding, local symbol repairability using rows or columns, and compatibility with cryptographic primitives. However, 2D Reed-Solomon codes compromise on code distance compared to other code constructions.

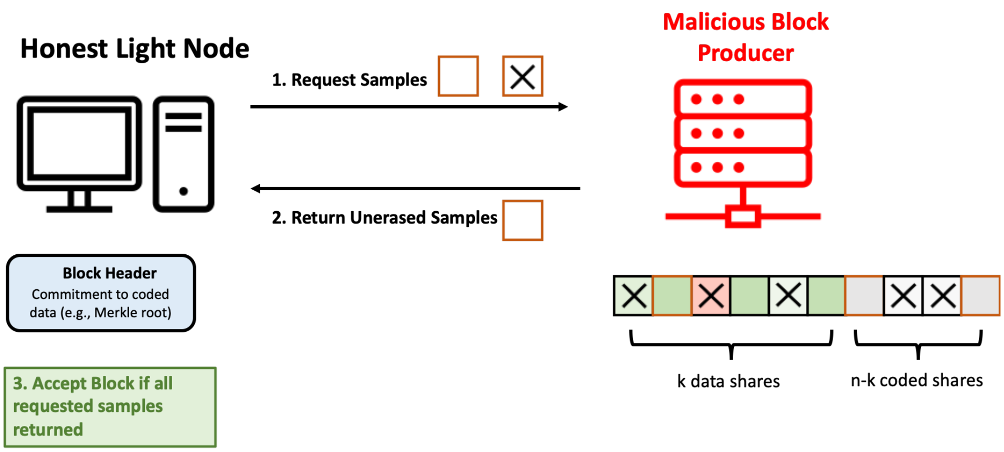
Improving the weakened probability of failure guarantees for 2D Reed-Solomon codes is the core objective of the thesis. Code distance measures only the cardinality of the undecodable erasure patterns. Standard DAS analysis, which uses uniform random sampling



(a) DA Attack: A Malicious block producer encodes the block, generates a commitment to place in the block header, and pre-emptively chooses which code symbols to hide.



(b) DA Attack: The malicious block producer distributes block headers to the light nodes, and sends block data with missing transactions to honest full nodes.



(c) DA Attack: A light node requests samples from the block producer. The light node accepts the block if all requested samples are returned, and rejects the block otherwise.

Figure 1.1: DAS Model for a DA Attack

without replacement, relies solely on the code distance for probability of failure. For a product code, however, the set of undecodable erasure patterns is not an arbitrary set. Each undecodable erasure pattern for the product code has a rectangular shape and combinatorial properties. A light node which samples to exploit structured erasure pattern geometry may improve probability of failure far beyond what code distance alone implies.

In this thesis, we build a joint code-sampler design for 2D Reed-Solomon codes. First, holding the code fixed, we design samplers which improve probability of failure against the rectangular geometry of undecodable erasure patterns. Then, holding our sampler fixed, we design a subcode which favorably reshapes the surviving erasure pattern geometry to be more sampler-compatible. Minimum distance should not be the sole design objective for DAS over the set of product codes. Our best subcode construction has a lower guaranteed distance than our alternative construction but a strictly better worst-case probability of failure.

1.4 Contributions

We study the joint design of a sampling strategy and subcode for 2D Reed-Solomon encoded DAS systems. Our performance comparison is a light node's worst case *probability of failure* for detecting a DA attack. Our contributions are as follows.

1. **Geometry-Aware Sampler Design.** We design a *biregular sampler* for 2D-RS codes, which enforces uniform sample coverage over rows and columns. For $Q \leq \min(n_1, n_2)$ samples, our sample set coincides with those of DiDAS [KBR24b] and the Improved Sampling Protocol [MTD25]. We prove that it is optimal over the class of minimum distance erasure patterns in this regime. For $Q > \min(n_1, n_2)$, the biregular sampler shows continued probability of failure improvements.

2. **Asymptotic Sampler Performance.** We evaluate the performance of our sampler for $\frac{Q}{n_1 n_2} \to \rho$, where ρ is a fixed fraction of the block. We find the exponential decay

rates in probability of failure for the biregular sampler against minimum distance erasure patterns. Then, we prove a sufficient region in which the analyzed minimum distance erasure patterns are the worst-case erasure pattern for probability of failure analysis.

3. **Sampler-Aware Subcode Design** We design a 2D-RS subcode to improve probability of failure using a biregular sampler with $Q = n$ queries. We use two families of parities: one enlarges the rectangular support of each surviving erasure pattern, and one suppresses all failure-increasing partial matching counts for each surviving erasure pattern. We obtain a subcode with a guaranteed and explicit worst-case probability of failure guarantee.

1.5 Thesis Organization

Chapter 2 establishes the notation and coding theory preliminaries used throughout the paper, defines the DAS system model and trust assumptions used, and characterizes the undecodable erasure patterns of 2D Reed-Solomon Codes.

In Chapter 3, we develop the *biregular sampler*. We motivate sample set regularity from the structure of minimum distance erasure patterns, and analyze row-regular, column-regular, and biregular samplers for a low number of samples. Then, we compare the asymptotics of each sampler, and prove that minimum distance erasure patterns are the worst-case erasure pattern for the biregular sampler when $\delta < \frac{1}{3}$. We provide a sufficient condition to prove whether the code's distance and percentage of the block sampled have worst-case probability of failure for minimum distance erasure patterns. The chapter closes with numerical comparisons in probability of failure to the *Improved Sampling Protocol* [MTD25] and uniform random sampling, as well as a discussion on sampler implementation complexity.

In Chapter 4, we fix our sampler design as a biregular sampler and design a 2D RS subcode to improve probability of failure. We construct a sampler-aware 2D-RS subcode using two families of heavy parities and bound the worst-case probability of failure for the joint subcode-sampler design.

In Chapter 5, we summarize our contributions and list some open problems.

In Appendix A.1, we provide the asymptotic proof for distance trade-offs between the minimum distance and second-weight erasure patterns in the sparse sample regime. Then, in Appendix B, we provide the proof for our asymptotic sampling regime to determine which code distances allow have minimum distance erasure patterns as the worst-case probability of failure event.

CHAPTER 2

Background

2.1 Notation and General Preliminaries

Let $[n] = \{1, \ldots, n\}$, and let $\binom{[n]}{k}$ denote the collection of all k-subsets of $[n]$. For a matrix $X \in [0,1]^{m\times n}$, let $H(X) = \sum_{ij} h(x_{ij})$, where $h(x_{ij}) = -x_{ij}\ln(x_{ij}) - (1-x_{ij})\ln(1-x_{ij})$ is the natural entropy function. Let $\mathbb{F}_q$ denote a finite field of size q. Let $(x)_+$ denote the expression $\max(x, 0)$. Let $(x)_m = x(x-1)\ldots(x-m+1)$ denote a falling factorial.

2.1.1 Graph Theory Preliminaries

Let K_{n_1,n_2} denote a complete bipartite graph with n_1 left-vertices and n_2 right vertices. We label the left-vertices $\{1, \ldots, n_1\}$ and the right-vertices $\{1, \ldots, n_2\}$. Then, K_{n_1,n_2} has an edge set $[n_1] \times [n_2]$. For a set $\mathcal{F} \subseteq [n_1] \times [n_2]$, let graph $G_{\mathcal{F}} \subseteq K_{n_1,n_2}$ denote a bipartite graph with edges specified by the positions in $\mathcal{F}$. For graph $G_{\mathcal{F}}$, we define $r_i(\mathcal{F})$ as the set of right-vertices adjacent to left-vertex i, and $c_j(\mathcal{F})$ as the set of left-vertices adjacent to right-vertex j. For a set $\mathcal{A} \subseteq K_{n_1,n_2}$, let $m_j(\mathcal{A})$ denote the number of j-partial matchings between j distinct left-vertices and j distinct right-vertices.

2.2 Coding Theory Preliminaries

We introduce basic coding theory preliminaries used throughout the thesis. An $[n, k]$ linear code $\mathcal{C} \subseteq \mathbb{F}_q^n$ is a linear subspace of $\mathbb{F}_q^n$ of dimension k. Code $\mathcal{C}$ encodes k transaction symbols into n coded symbols. The *distance* of a code, $d(\mathcal{C})$, is the minimum Hamming weight, wt() over all non-zero codewords. Up to $d(\mathcal{C}) - 1$ erased symbols may be corrected using an ML decoder.

Suppose a coordinate set $\mathcal{E} \subseteq [n]$ of symbols is erased. Then, we call $\mathcal{E}$ an *erasure pattern.* For a fixed decoder, $\mathcal{E}$ is *undecodable* if the decoder cannot fully recover the original codeword with the remaining symbols. The set of undecodable erasure patterns is denoted $\mathcal{E}_u$, where $\mathcal{E}_u \subseteq 2^{[n]}$.

Definition 2.1 (Inclusion-Minimal Erasure Pattern)**.** For a code $\mathcal{C}$ with support $[n]$, an erasure pattern $\mathcal{E} \subseteq [n]$ is *inclusion-minimal* if it is undecodable, and there does not exist any other undecodable erasure pattern $\mathcal{E}'$ such that $\mathcal{E}' \subsetneq \mathcal{E}$. We let $\mathcal{E}_{\text{inc}}$ denote the set of all inclusion-minimal erasure patterns for code $\mathcal{C}$ for a specified decoder.

Under ML decoding, $\mathcal{E}$ is undecodable if and only if it contains the support of a nonzero codeword. Hence, every undecodable erasure pattern has a cardinality of at least $d(\mathcal{C})$, and $\mathcal{E}_{\text{inc}}$ is exactly the set of minimal codeword supports. For an iterative or peeling decoder, there may exist undecodable erasure patterns below the minimum distance. The set of undecodable erasure patterns for the ML-decoder is a subset of the undecodable erasure patterns for an iterative decoder.

2.3 Blockchain and DAS Model

2.3.1 Blockchain Block Model

Blockchain is a distributed ledger, in which all fully participating nodes in the network fully download, validate, and custody block transactions. Each blockchain is composed of a sequence of *blocks*. A *block* consists of a set of transaction data, which is batched into k units called *shares*. Each block also contains a *block header*, which holds the block's consensus information (e.g., Proof-of-Work or Proof-of-Stake), a hash linking to the previous block header, and a cryptographic *commitment* to the block data, which is a cryptographic proof which allows any node in the network to verify a given transactional share is included in a block.

In a *coded* blockchain system, k shares of block data are encoded into n coded shares using linear erasure code $\mathcal{C}$. A commitment to the encoded block data is stored in the block header [ASB21]. For coded blockchains, each node must have a mechanism to prove that the data was encoded correctly. An *incorrect-coding proof* is a mathematical proof that a codeword was not correctly encoded, allowing honest nodes to reject an incorrectly encoded block. The incorrect coding proof size is dependent on both the code and the commitment. For 2D Reed-Solomon codes and KZG commitments, for example, the commitment also functions as a correct-encoding proof, and an extra incorrect coding proof mechanism is not needed [HSW23, 0xb24]

2.3.2 Block Distribution

A coded blockchain system consists of three agents: the *block producer*, *light (or verifier) nodes*, and *full nodes*. A block producer encodes a set of k *shares* of transactions using code $\mathcal{C}$, and then generates a commitment over the encoded data. An honest block producer disseminates the block header to all nodes in the network, and sends the encoded block data to full nodes in the network. Full nodes download, store, and verify proposed block data, and broadcast fraud proofs if any transaction is incorrect or incorrectly encoded. Light nodes only store the block header, and may verify transaction inclusion but not correctness. The block distribution model is illustrated in Figure 1.1.

2.3.3 Trust and Network Assumptions

We follow the network protocols in [ASB21]. First, we assume a synchronous network, in which all nodes will successfully receive a sent transaction within a fixed delay. If one honest node receives a message, then all other honest nodes may also receive the message if requested. We now outline three main assumptions required for our sampler analysis.

Assumption 2.2 (Light Node Connectivity)**.** *Each light node is connected to at least one honest full node in the network. The network may consist of a dishonest majority of full nodes.*

Assumption 2.3 (Query Unlinkability)**.** *No node can link a query or sequence of queries to a single light node.*

Assumption 2.4 (Non-Adaptivity of Adversary)**.** *An adversarial node is non-adaptive. The set of withheld symbols is chosen before any sample requests are made.*

Under Assumption 2.4, the malicious block producer selects the withheld symbol set before any light node requests samples. Additionally, we assume that a sufficient number of light nodes sample the block to allow for block recovery.

The security assumptions are also discussed in [joa22] and are used in other coded DAS papers [GRK26, YSL20, MTD25]. In Ethereum PeerDAS, every honest full node is able to download any code symbol received by another honest full node in the network [Eth26].

2.3.4 Data Availability Attack and DAS Protocol

In a DA attack, an adversarial block producer encodes a set of k shares using erasure code $\mathcal{C}$, and generates a commitment over the coded data. The adversary pre-emptively selects a set $\mathcal{E} \subseteq [n]$ of code symbols to withhold from the network such that the remaining symbols $[n] \setminus \mathcal{E}$ cannot decode the block. Then, the adversary broadcasts the remaining symbols to the network of honest full nodes.

Each light node draws a sample set $\mathcal{S} \subseteq [n]$ for a code of length n with cardinality $|\mathcal{S}| \leq Q$ for query budget Q. The light node then queries the block producer for coded symbols at all index positions in $\mathcal{S}$. Each symbol is returned if and only if it is available to at least one honest full node (Assumption 2.2). The light node accepts the block as available if every requested coordinate is returned, i.e., $\mathcal{S} \cap \mathcal{E} = \emptyset$. If a DA attack is successful, light nodes accept an invalid block, which may cause funds to be stolen or a fork in the blockchain. We define a light node's sampling strategy as follows.

Definition 2.5 (Sampling Strategy)**.** A *sampling strategy* is a probability distribution π over $\{\mathcal{S} \subseteq [n] : |\mathcal{S}| \leq Q\}$. A light node samples using strategy π by drawing set $\mathcal{S} \sim \pi$ and querying the coordinate positions in $\mathcal{S}$. The per-node *probability of failure* to detect a DA attack for an adversarial node withholding a fixed erasure pattern $\mathcal{E} \in \mathcal{E}_u$ is

$$p_\pi(\mathcal{E}) = \Pr_{S \sim \pi}[\mathcal{S} \cap \mathcal{E} = \emptyset]. \tag{2.1}$$

For any set of undecodable erasure patterns $\mathcal{E}' \subseteq \mathcal{E}_u$, the worst-case probability of failure is

$$p_\pi(\mathcal{E}') = \max_{\mathcal{E} \in \mathcal{E}'} p_\pi(\mathcal{E}). \tag{2.2}$$

Then, the optimal sampling strategy for a fixed code and decoder is

$$\pi^* \in \arg\min_{\pi} p_\pi(\mathcal{E}_{\text{inc}}). \tag{2.3}$$

In Equation 2.3, we optimize over the set of inclusion-minimal erasure patterns $\mathcal{E}_{\text{inc}}$ instead of $\mathcal{E}_u$ without loss of generality. Any $\mathcal{E} \in \mathcal{E}_u \setminus \mathcal{E}_{\text{inc}}$ contains an inclusion-minimal erasure pattern as a subset and therefore has strictly lower probability of failure. Therefore, it suffices to analyze the inclusion-minimal erasure patterns for worst-case probability of failure performance.

2.4 2D Reed-Solomon Code Background

In this section, we define 2D Reed-Solomon codes and analyze the properties of the code's undecodable erasure patterns. For two linear codes $\mathcal{C}_1 \subseteq \mathbb{F}_q^{n_1}$ and $\mathcal{C}_2 \subseteq \mathbb{F}_q^{n_2}$ for positive integers n_1 and n_2, a product code $\mathcal{C} = \mathcal{C}_1 \otimes \mathcal{C}_2$ is a subspace of $\mathbb{F}_q^{n_1 \times n_2}$ which is generated by the tensors $c_1 \otimes c_2$ for codewords $c_1 \in \mathcal{C}_1$ and $c_2 \in \mathcal{C}_2$. If d_1 is the minimum distance of $\mathcal{C}_1$ and d_2 is the minimum distance of $\mathcal{C}_2$, then the minimum distance of $\mathcal{C}$ is $d_1 d_2$ [Eli54]. For a finite field $\mathbb{F}_q$ of size $q \geq n$, let set $T = \{\alpha_1, \dots, \alpha_n\} \subseteq \mathbb{F}_q$ be a finite evaluation set of $\mathbb{F}_q$.

Definition 2.6 (Reed-Solomon Code [RS60])**.** For evaluation set $T \subseteq \mathbb{F}_q$, a Reed-Solomon code is defined as

$$RS_{\mathbb{F}_q, T}[n, k] = \{(p(\alpha))_{\alpha \in T} \mid p \in \mathbb{F}_q[X] \text{ is a polynomial of degree } < k\}. \tag{2.4}$$

A Reed-Solomon code is MDS, and therefore has code distance $d = n - k + 1$. A Reed-Solomon code can correct from up to $d - 1$ erasures.

A 2D Reed-Solomon code is a product code with Reed-Solomon constituent codes. The code may also be viewed as the evaluation code of a bivariate polynomial, which we use in our subcode design.

Definition 2.7 (2D Reed-Solomon Code)**.** A 2D Reed-Solomon code $\mathcal{C} \subseteq \mathbb{F}_q^{n_1 \times n_2}$ is an $[n_1 n_2, k_1 k_2, d_1 d_2]$ product code

$$\mathcal{C} = \mathcal{C}_1 \otimes \mathcal{C}_2,$$

for constituent Reed-Solomon codes

$$\mathcal{C}_1 = RS_{\mathbb{F}_q,S}[n_1, k_1, d_1], \quad \mathcal{C}_2 = RS_{\mathbb{F}_q,T}[n_2, k_2, d_2],$$

where $S \subseteq \mathbb{F}_q$ of size n_1 and $T \subseteq \mathbb{F}_q$ of size n_2. The code has dimension $\dim(C) = \dim(C_1)\dim(C_2) = k_1k_2$, minimum distance d_1d_2, and coordinate set $[n_1] \times [n_2]$.

2D RS codes may be represented as evaluation codes over bivariate polynomials with constrained monomial degrees. For $S, T \subseteq \mathbb{F}_q$, we define the space of bivariate, degree-constrained polynomials as

$$V_{a,b} = \{f \in \mathbb{F}_q[x, y] : \deg_x(f) < a, \deg_y(f) < b\}. \tag{2.5}$$

A polynomial $f \in V_{a,b}$ has maximum monomial degrees $\deg_x(f) \leq a-1$ and $\deg_y(f) \leq b-1$. Then, a 2D-RS codeword may be represented as an evaluation mapping

$$\begin{aligned} &\mathrm{ev}_{S,T} : V_{k_1,k_2} \to \mathbb{F}_q^{n_1 \times n_2} \\ &\mathrm{ev}_{S,T}(f) := \{f(\alpha, \beta) : (\alpha, \beta) \in S \times T\} \end{aligned}$$

over all bivariate polynomials of constrained monomial degree. We use the bivariate polynomial interpretation of 2D-RS codes in our subcode design.

2.4.1 2D-RS Erasure Pattern Analysis

For our DAS sampler design in Chapter 3, target the set $\mathcal{E}_{\text{inc}}$ in our sampler design. In this section, we represent erasure pattern supports using a rectangular support profile.

We use similar notational conventions for our erasure pattern analysis as [JB16, GHK17]. Using an iterative row-column decoder, we categorize each undecodable erasure patterns by its *rectangular support* of rows and columns with erased symbol positions.

Definition 2.8 (Erasure Pattern Rectangular Support [JB16])**.** Let $\mathcal{E} \subseteq [n_1] \times [n_2]$ denote a set of symbol positions in an $[n_1n_2, k_1k_2]$ 2D-RS code. The set of row positions associated to $\mathcal{E}$ is $\mathcal{U}(\mathcal{E}) = \{i_1, \dots, i_{d_1+u}\}$, and the set of column positions associated to $\mathcal{E}$ is $\mathcal{V}(\mathcal{E}) = \{j_1, \dots, j_{d_2+v}\}$ for non-negative integers u and v. The rectangular support of $\mathcal{E}$ is

$$\mathcal{X}(\mathcal{E}) = \mathcal{U}(\mathcal{E}) \times \mathcal{V}(\mathcal{E}).$$

Any erasure pattern $\mathcal{E} \in \mathcal{E}_u$ may be represented as

$$\mathcal{E} = \mathcal{X}(\mathcal{E}) \backslash \mathcal{A}(\mathcal{E}) = (\mathcal{U}(\mathcal{E}) \times \mathcal{V}(\mathcal{E})) \backslash \mathcal{A}(\mathcal{E}),$$

where $\mathcal{A}(\mathcal{E}) = \mathcal{X} \setminus \mathcal{E}$ is the set of unerased positions in the rectangular support.

When the erasure pattern $\mathcal{E}$ is clear from context, we label the rectangular support as $\mathcal{X} = \mathcal{U} \times \mathcal{V}$ with a slight abuse of notation. For the set of inclusion-minimal erasure patterns, $\mathcal{E}_{\text{inc}}$, from Definition 2.1, each erasure pattern has important structural properties.

Lemma 2.9 (Inclusion-Minimal Erasure Pattern Properties)**.** *For the set of inclusion-minimal erasure patterns $\mathcal{E}_{inc}$, any erasure pattern $\mathcal{E} \in \mathcal{E}_{inc}$ has the following properties:*

1. *Each row in $\mathcal{U}$ has a maximum of v unerased symbols in $\mathcal{A}$*
2. *Each column in $\mathcal{V}$ has a maximum of u unerased symbols in $\mathcal{A}$*
3. *For any $(i, j) \in \mathcal{E}$, either row i has exactly v unerased symbols in $\mathcal{A}$, row j has exactly u unerased symbols in $\mathcal{A}$, or both conditions hold.*
4. *The cardinality of $\mathcal{A}$ is upper bounded as*

$$|\mathcal{A}| \leq \min\left((d_1 + u)v, (d_2 + v)u\right). \tag{2.6}$$

Proof. We split the proof into three sections: a proof of claims 1 and 2, a proof of claim 3, and a proof of claim 4.

Proof of claims 1 *and* 2*.* Suppose there exists an $\mathcal{E} \in \mathcal{E}_{\text{inc}}$ such that one row i has more than v erased symbols. Then, there are greater than $(n-d-v)+v = k-1$ code symbols available in row i, and row i may be successfully decoded. Then, either $\mathcal{E}$ is not undecodable, or else there exists an erasure pattern $\mathcal{E}' \subseteq \mathcal{E}$ which is also undecodable. This is a contradiction, as $\mathcal{E}$ is inclusion-minimal. The proof of claim 2 follows similarly.

Proof of claim 3. For $\mathcal{E} \in \mathcal{E}_{\text{inc}}$ with support $\mathcal{U} \times \mathcal{V} \setminus \mathcal{A}$, suppose there exists an $(i, j) \in \mathcal{E}$ where neither row i nor column j have v or u unerased symbols. Then, for erasure pattern $\mathcal{E}' = \mathcal{E} \setminus \{(i, j)\}$, row i has at least d_2 erased symbols and column j has at least d_1 erased symbols. Thus, neither row i nor column j can be decoded in $\mathcal{E}'$, and $\mathcal{E}'$ is still undecodable. This is a contradiction of inclusion-minimality.

Proof of claim 4. The rectangular support occupies $(d_1 + u)(d_2 + v)$ total positions. Each row must have a minimum of d_2 erasures, and each column must have a minimum of d_1 erasures. Then, there are a minimum of $\max((d_1 + u)d_2, (d_2 + v)d_1)$ erasures in $\mathcal{X}$. Therefore, there can be at most $(d_1 + u)(d_2 + v) - \max((d_1 + u)d_2, (d_2 + v)d_1)$ unerased positions in $\mathcal{X}$. To simplify the expression, we expand terms to find

$$\begin{aligned}(d_1 + u)(d_2 + v) - \max((d_1 + u)d_2, (d_2 + v)d_1) &= \min(ud_2, vd_1) + uv \\ &= \min((d_1 + u)v, (d_2 + v)u).\end{aligned}$$

□

The cardinality of the rectangular support for inclusion-minimal erasure patterns may be viewed as follows. Each row with erased symbols must have less than k_2 symbols, or else the row may be recovered. Each column must have less than k_1 symbols, or else the column may be recovered. For any one column to be erased, at least d_1 row positions must be erased. The same argument applies for rows. Then, each undecodable erasure pattern will occupy a rectangular subset of the block of size at least $d_1 \times d_2$.

For a fixed erasure pattern $\mathcal{E}$, let $r_i(\mathcal{E})$ denote the set of erased positions in row i, and let $c_j(\mathcal{E})$ denote the set of erased positions in column j. Formally, we define

$$r_i(\mathcal{E}) = \{j \in [n_2] : (i, j) \in \mathcal{E}\}, \qquad c_j(\mathcal{E}) = \{[i] \in [n_1] : (i, j) \in \mathcal{E}\}.$$

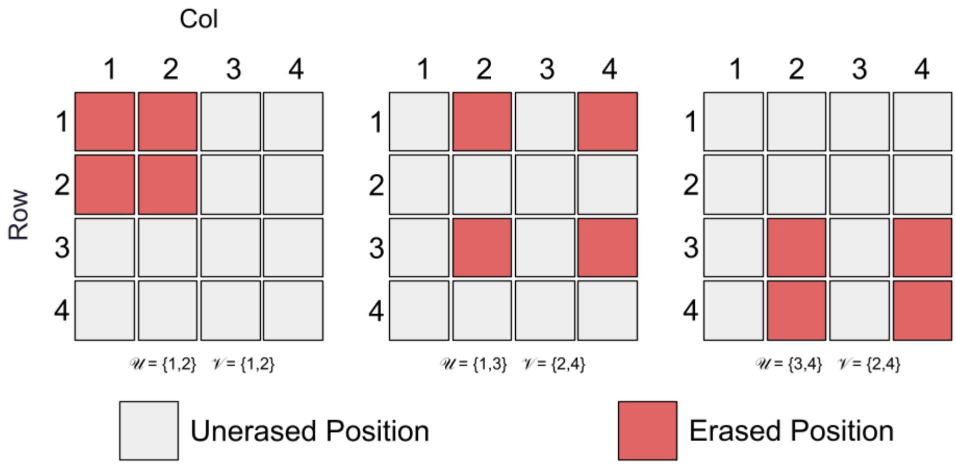


Figure 2.1: Three minimum distance erasure patterns for a $[5^2, 4^2]$ 2D-RS code.

Now, we state the properties of several classes of low-cardinality erasure patterns, which we consider in our sampler probability of failure analysis.

Lemma 2.10 (Minimum Distance Erasure Patterns [JB16]). *For an $[n_1 n_2, k_1 k_2]$ 2D-RS code, the set of minimum distance erasure patterns $\mathcal{E}_{min}$ is*

$$\mathcal{E}_{min} = \left\{ \mathcal{U} \times \mathcal{V} \subseteq [n_1] \times [n_2] : \mathcal{U} \in \binom{[n_1]}{d_1}, \mathcal{V} \in \binom{[n_2]}{d_2} \right\}. \tag{2.7}$$

An example of three minimum distance erasure patterns for a $[5^2, 4^2]$ 2D-RS code is illustrated in Figure 2.1. In the example, $d_1 = d_2 = 2$, and the minimum distance erasure patterns have a rectangular support of size $d_1 \times d_2$.

Next, we examine the structural properties of the second lowest cardinality erasure patterns for 2D-RS codes with $d_1 = d_2 = d$ [JB16]. Any second cardinality erasure pattern which contains a minimum distance erasure pattern as a subset has strictly worse probability of failure. Hence, we examine the inclusion-minimal erasure patterns for probability of failure analysis.

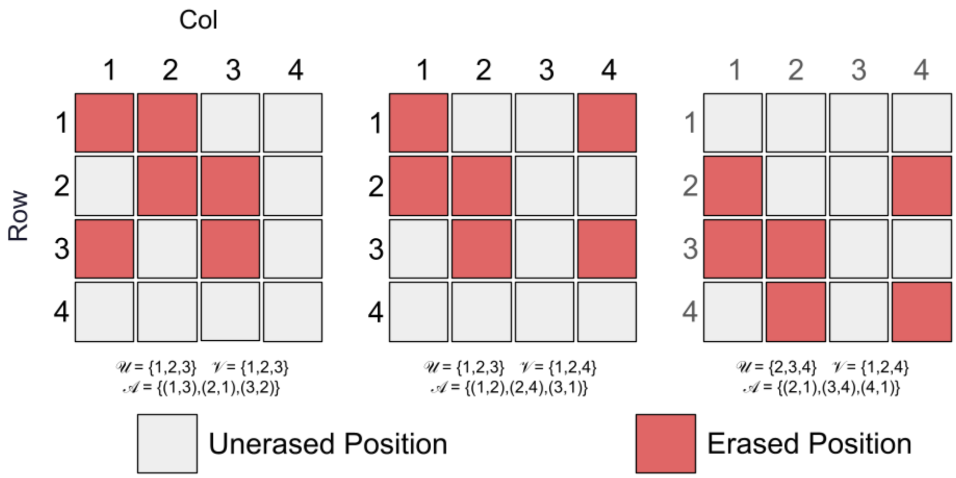


Figure 2.2: Three second-weight, inclusion-minimal erasure patterns for a $[5^2, 4^2]$ 2D-RS code.

Lemma 2.11 (Second-Cardinality Inclusion-Minimal Erasure Patterns [JB16]). *For an $[n_1 n_2, k^2]$ 2D-RS code with $d_1 = d_2 = d$, the set of second-weight inclusion-minimal erasure patterns is*

$$\mathcal{E}_2 = \{\mathcal{U} \times \mathcal{V} \setminus \mathcal{A} : \mathcal{U} \in \binom{[n_1]}{d+1}, \mathcal{V} \in \binom{[n_2]}{d+1}, \mathcal{A} = (i, \sigma(i)) \text{ for } i \in \mathcal{U}\}, \tag{2.8}$$

where $\sigma : \mathcal{U} \to \mathcal{V}$ is a bijection.

An example of three inclusion-minimal erasure patterns in $\mathcal{E}_2$ for a $[5^2, 4^2]$ 2D-RS code is shown in Figure 2.2. The rectangular support occupies a 3×3 subgrid of code, and each row and column in the support has maximum one unerased position, forming a matching on the rectangular support. Any 2×3 erasure pattern which occupies the entire rectangular support is not inclusion-minimal.

CHAPTER 3

2D-RS Sampler Design

We design targeted sampling strategies for DAS systems using a $[n_1 n_2, k_1 k_2]$ 2D-RS code. The code-sampler co-design problem has been explored for both LDPC and Polar codes [MTD23, MTD22]. For coded DAS systems, uniform random sampling is the assumed sampling model, and relies only on code distance properties. However, many codes which are favorable for other code properties, such as local symbol repair, compromise on code distance, and thus probability of failure, guarantees [SVH25, ASB21]. 2D Reed-Solomon codes are heavily being considered for implementation due to their structured row and column local codes and commitment compatibility [KB25, HSW23]. For this reason, we design samplers for 2D-RS codes to improve probability of failure given the code constraints.

Our sampling algorithms exploit the structured, rectangular-like erasure patterns inherent to product codes. We compare our samplers to our previous work, in which we developed the *Improved Sampling Protocol* (ISP), DiDAS, and with uniform random sampling without replacement [MTD25, KBR24b]. For this section, we assume an iterative row-column decoder is used on the code. Probability of failure performance analysis under an ML-decoder can only lower worst-case probability of failure for 2D-RS codes, as the set of ML-undecodable erasure patterns is a subset of those undecodable using an iterative row-column decoder.

First, we motivate our sampler design by examining the probability of failure against minimum distance erasure patterns for a sample set $\mathcal{S}$ randomized using a distinct row and column permutation. We introduce a low-complexity sampler which has provable improvement in probability of failure.

Then, we evaluate our sampler in the *sparse-sample* regime, in which we sample $Q \leq \min(n_1, n_2)$ queries. This region is desirable for light nodes, as light nodes incur a sample download cost proportional to the number of samples taken. We prove that our sampler is optimal against *minimum distance* erasure patterns. Then, we contribute a further analysis of our sampler against higher-order erasure patterns. We find the crossover point for probability of failure performance against second-order erasure patterns, and provide an upper bound on the maximum probability of failure of our sampler against any inclusion-minimal erasure pattern.

For the asymptotic sample regime, we find the exponential decay rate of the probability of failure for our samplers, and compare the asymptotic probability of failure improvements to a uniform random sampler without replacement. We prove that minimum distance erasure patterns are the worst case erasure patterns for our sampler for $\frac{d}{n} = \delta < \frac{1}{3}$ and $2\rho + 3\delta \leq 1$.

3.1 Sampler Comparison Baseline: Uniform Random Sampling

Probability of failure analysis under uniform random sampling only relies on the code's distance, and does not account for any structural properties of codeword supports or undecodable erasure patterns. Before we design our sampler, we state the probability of failure for uniform random sampling without replacement, which is the baseline against which we evaluate our results.

Lemma 3.1 (Uniform Sampler). *For a query budget $Q \in \mathbb{Z}^+$, a uniform sampler $\pi_{U,Q}$ draws a sample set $\mathcal{S}$ uniformly among all of the Q-subsets of $[n_1] \times [n_2]$. The worst case probability of failure for the uniform sampler is attained by the set of minimum distance* *erasure patterns, $\mathcal{E}_{min}$, and is equal to*

$$p_{\pi_{U,Q}}(\mathcal{E}_{min}) = \frac{\binom{n_1 n_2 - d_1 d_2}{Q}}{\binom{n_1 n_2}{Q}}. \tag{3.1}$$

Proof. Uniform random sampling without replacement draws Q distinct placements out of $n_1 n_2$ total placements. There are $\binom{n_1 n_2}{Q}$ total ways to place Q distinct samples in $n_1 n_2$ positions. An adversary pre-emptively chooses an erasure pattern $\mathcal{E} \in \mathcal{E}_{\text{inc}}$ to withhold from the network. A light node samples one of the withheld positions if any Q samples land in the $|\mathcal{E}|$ withheld positions. There are $\binom{n_1 n_2 - |\mathcal{E}|}{Q}$ ways to place Q distinct positions on the remaining $n_1 n_2 - d_1 d_2$ positions not withheld by the block producer. Then, as $\binom{n_1 n_2 - |\mathcal{E}|}{Q}$ is monotonically decreasing in $|\mathcal{E}|$, the optimal adversary selects a minimum cardinality erasure pattern. Therefore, an adversary chooses an erasure pattern from set $\mathcal{E}_{\text{min}}$ with cardinality $d_1 d_2$. □

3.2 Motivation and Row-Regular Sampling

For 2D-RS codes, the baseline uniform sampler has a worst case probability of failure against $\mathcal{E}_{\text{min}}$. Thus, we first analyze the sampler design problem against $\mathcal{E}_{\text{min}}$ for improvements in probability of failure.

Suppose we design an informed sampler which applies a random row and column permutation to a fixed sample set $\mathcal{S} \subseteq [n_1] \times [n_2]$. We denote this strategy as $\pi_{\mathcal{S}}$. We want to evaluate the worst-case probability of failure of strategy $\pi_{\mathcal{S}}$ against an adversarial block producer constrained to withhold a *minimum distance erasure pattern* $\mathcal{E} \in \mathcal{E}_{\text{min}}$.

Theorem 3.2 (Probability of Failure for Minimum Distance Erasure Patterns)**.** *For a $[n_1 n_2, k_1 k_2]$ 2D-RS code, let $\pi_{\mathcal{S}}$ denote a sampling strategy which applies a random row and column permutation to sample set $\mathcal{S}$. Then, the probability of failure of $\pi_{\mathcal{S}}$ against a* *minimum distance erasure pattern in $\mathcal{E}_{min}$ is*

$$\begin{aligned} p_{\pi_{\mathcal{S}}}(\mathcal{E}_{min}) &= \frac{1}{\binom{n_2}{d_2}\binom{n_1}{d_1}} \sum_{\mathcal{R}\in\binom{[n_1]}{d_1}} \binom{n_2 - |\cup_{i\in\mathcal{R}} r_i(\mathcal{S})|}{d_2} \\ &= \frac{1}{\binom{n_2}{d_2}} \mathbb{E}_{\mathcal{R}}\left[\binom{n_2 - |\cup_{i\in\mathcal{R}} r_i(\mathcal{S})|}{d_2}\right], \end{aligned} \tag{3.2}$$

where $\mathbb{E}_{\mathcal{R}}[]$ is the expected value over $\mathcal{R}$ uniformly distributed on $\binom{[n_1]}{d_1}$.

Proof. For a fixed sample set $\mathcal{S}$, the sampler requests a random row and column permutation of $\mathcal{S}$. The adversarial block producer knows strategy $\pi_{\mathcal{S}}$, and therefore knows set $\mathcal{S}$. By Assumption 2.4, the adversarial node is non-adaptive. Therefore, the adversarial node pre-emptively selects an $\mathcal{E} \in \mathcal{E}_{\min}$ to withhold from the system of full nodes.

First, we prove that there does not exist some $\mathcal{E}, \mathcal{E}' \in \mathcal{E}_{\min}$ such that $p_{\pi_{\mathcal{S}}}(\mathcal{E}) \neq p_{\pi_{\mathcal{S}}}(\mathcal{E}')$. The set of erasure patterns $\mathcal{E}_{\min}$ is invariant to any row and column permutation. For any $\mathcal{E} \in \mathcal{E}_{\min}$, applying any row and column permutation will result in some $\mathcal{E}' \in \mathcal{E}_{\min}$. By design, $\pi_{\mathcal{S}}$ is also invariant to any row and/or column permutation. Then, for a random row index permutation σ and column index permutation τ,

$$p_{\pi_{\mathcal{S}}}(\mathcal{E}_{\min}) = \max_{\mathcal{E}\in\mathcal{E}_{\min}} \Pr[(\sigma \times \tau(\mathcal{S})) \cap \mathcal{E} = \emptyset].$$

Any $\mathcal{E} \in \mathcal{E}_{\min}$ may be represented as $\mathcal{E} = \sigma' \times \tau'([d_1] \times [d_2])$ for row permutation σ' and column permutation τ'. Then,

$$\begin{aligned} p_{\pi_{\mathcal{S}}}(\mathcal{E}_{\min}) &= \max_{\sigma',\tau'} \Pr[(\sigma \times \tau(\mathcal{S})) \cap \sigma' \times \tau'([d_1] \times [d_2]) = \emptyset] \\ &= \max_{\sigma',\tau'} \Pr[(\sigma')^{-1}\sigma \times (\tau')^{-1}\tau(\mathcal{S}) \cap ([d_1] \times [d_2]) = \emptyset], \end{aligned}$$

where $(\sigma')^{-1}$ and $(\tau')^{-1}$ represent the inverse permutation operation. Then, $(\sigma')^{-1}\sigma$ and $(\tau')^{-1}\tau$ are a random row and column permutation regardless of choice of σ' and τ', as σ and τ are random. Therefore, there does not exist some σ' or τ' which changes the probability of failure.

The proof then reduces to a counting argument on the number of minimum distance erasure patterns $\mathcal{E} \in \mathcal{E}_{\min}$ which do not intersect with $\mathcal{S}$. The probability of failure is then

$$p_{\pi_{\mathcal{S}}}(\mathcal{E}_{\min}) = \frac{|\{G_{\mathcal{E}} \subseteq K_{n_1,n_2} : \mathcal{E} \in \mathcal{E}_{\min}, \mathcal{E} \cap \mathcal{S} = \emptyset\}|}{|\mathcal{E}_{\min}|}$$

The cardinality of $\mathcal{E}_{\min}$ is $|\mathcal{E}_{\min}| = \binom{n_1}{d_1}\binom{n_2}{d_2}$. This count follows directly from Lemma 2.10. We evaluate the set of erasure patterns with a fixed row-set $\mathcal{R}$ which avoid sample set $\mathcal{S}$. Let

$$f_{\mathcal{S}}(\mathcal{R}) = \bigcup_{i \in \mathcal{R}} r_i(\mathcal{S})$$

denote the set of columns which are sampled by at least one row position in $\mathcal{R}$. Then, if $f_{\mathcal{S}}(\mathcal{R}) \cap \mathcal{Y} \neq \emptyset$ for any column set $\mathcal{Y} \in \binom{[n_2]}{d_2}$, erasure pattern $\mathcal{R} \times \mathcal{Y}$ is sampled by $\pi_{\mathcal{S}}$. We count how many $\mathcal{Y}$ avoid the restricted sample set. As $\mathcal{Y} \in \binom{[n_2]}{d_2}$ selects d_2 column indices, there are $\binom{n_2 - |f_{\mathcal{S}}(\mathcal{R})|}{d_2} = \binom{n_2 - |\bigcup_{i \in \mathcal{R}} r_i(\mathcal{S})|}{d_2}$ ways to select $\mathcal{Y}$ so that no position intersects with $f_{\mathcal{S}}(\mathcal{R})$. Then, the counting argument is finished by summing over all possible $\mathcal{R} \in \binom{n_1}{d_1}$.

We may model each $\mathcal{R} \in \binom{[n_1]}{d_1}$ as being drawn uniformly at random with probability $\frac{1}{\binom{n_1}{d_1}}$, as each $\mathcal{R}$ selection has equal probability of failure. Then, we may express the probability of failure using the expectation of $\mathcal{R}$ over $\binom{n - |f_{\mathcal{S}}(\mathcal{R})|}{d_2}$. $\square$

There are several consequences to the minimum distance probability of failure formula in Theorem 3.2. First, for a deterministic sample set $\mathcal{S}$, calculating the probability of failure in Equation 3.2 requires summing over an exponential number of terms. Therefore, using a greedy algorithm to sequentially select samples is computationally infeasible. In our previous work, the ISP greedy algorithm selected samples sequentially based on Bonferroni upper and lower bounds on $p_{\pi_{\mathcal{S}}}(\mathcal{E}_{\min})$ [MTD25]. However, these bounds become trivial as Q grows, and the ISP terminates and defaults to random sampling without replacement at approximately $Q = 2n$ samples. The probability of failure of all three samplers in the sparse-sample regime is plotted in Figure 3.6. Then, for an extended number of samples, the ISP bounds trivialize in Figure 3.5.

The properties of $p_{\pi_S}(\mathcal{E}_{\min})$ allow for sampler design with continued gains in probability of failure. Probability of failure $p_{\pi_S}(\mathcal{E}_{\min})$ is the expected value of a discretely convex function in $\mathcal{R}$, as $f(x) = \binom{n_2-x}{d_2}$ is discretely convex and decreasing in x. Thus, an application of Jensen's inequality lower bounds the probability of failure. The question of optimal sampler design against minimum distance erasure patterns has an equivalent problem in the supersaturation literature [Nag17]. Nagy proved that a biregular sample set was optimal when $d_1 = d_2 = 2$, and examined the optimal strategies for $d_1 = 2$, and $d_2 \gg d_1$ [Nag17]. Additionally, Nagy proved that the minimized lower bound for Equation 3.2 results from row-regular or column-regular graphs, as graphs with this property maximize the first moment $\mathbb{E}_{\mathcal{R}}\left[|\cup_{i\in\mathcal{R}} r_i(\mathcal{S})|\right]$ or the column-equivalent expression [Nag17]. However, exact optimal sampler design is an open problem in the equivalent graph theory literature, even for minimum distance erasure patterns alone [Nag17, Con22]. We therefore base our sampler design on these graph-regularity properties.

We first introduce a *row-regular* sampler, which draws a uniform number of samples from each row without replacement. Because a row-regular sampler samples each row independently, its worst-case probability of failure has a closed-form solution for 2D-RS codes. Additionally, the worst-case erasure patterns for probability of failure are exactly the set of *minimum distance* erasure patterns for the row-regular sampler.

Definition 3.3 (Row-regular Sampler)**.** For an $[n_1n_2, k_1k_2]$ 2D-RS code and sample set $\mathcal{S} \subseteq [n_1] \times [n_2]$ of size Q, a row-regular sampler, which we denote $\pi_{\mathcal{R},Q}$, draws uniformly from set

$$\mathcal{R}_Q = \left\{ G_{\mathcal{S}} \subseteq K_{n_1,n_2} : |r_i(\mathcal{S})| \in \{d_R, d_R+1\} \ \forall i \ \in [n_1] \right\}, \tag{3.3}$$

where $d_R = \lfloor \frac{Q}{n_1} \rfloor$.

A row-regular sampler may be instantiated as follows. If $n_1 \mid Q$, then $\frac{Q}{n_1}$ samples are drawn without replacement in each row independently. If $n_1 \nmid Q$, then a row-regular sampler randomly selects $Q \mod n_1$ rows where $d_R + 1$ samples are independently drawn without replacement. The remaining rows independently sample d_R samples without replacement. A column-regular sampler may be defined similarly, which we denote as strategy $\pi_{C,Q}$, with $d_C = \lfloor \frac{Q}{n_2} \rfloor$. We analyze the probability of failure for the row-regular sampler against minimum distance erasure patterns, and prove that this is the worst-case probability of failure for the sampler over all $\mathcal{E} \in \mathcal{E}_{\text{inc}}$.

Theorem 3.4 (Row-Regular Probability of Failure)**.** *For an $[n_1 n_2, k_1 k_2]$ 2D-RS code and a row-regular sampler $\pi_{\mathcal{R},Q}$ which requests Q sample positions, let $d_R = \lfloor \frac{Q}{n_1} \rfloor$ and $b = Q$* mod n_1*. The probability of failure for minimum distance erasure patterns is*

$$p_{\pi_{\mathcal{R},Q}}(\mathcal{E}_{min}) = \sum_{i=0}^{\min(b,d_1)} \frac{\binom{b}{i}\binom{n_1-b}{d_1-i}}{\binom{n_1}{d_1}} \left(\frac{\binom{n_2-d_2}{d_R+1}}{\binom{n_2}{d_R+1}} \right)^i \left(\frac{\binom{n_2-d_2}{d_R}}{\binom{n_2}{d_R}} \right)^{d_1-i}. \tag{3.4}$$

When $n_1 \mid Q$, then the probability of failure may be simplified as

$$p_{\pi_{\mathcal{R},Q}}(\mathcal{E}_{min}) = \left(\frac{\binom{n_2-d_2}{d_R}}{\binom{n_2}{d_R}} \right)^{d_1}. \tag{3.5}$$

The worst-case probability of failure for the row-regular sampler is achieved by the set of minimum distance erasure patterns.

Proof. The probability of failure in Eq 3.4 is found as follows. Any erasure pattern $\mathcal{E} \in \mathcal{E}_{\text{min}}$ may be entirely represented using rectangular support $\mathcal{R} \times \mathcal{Y}$, for $\mathcal{R} \in \binom{[n_1]}{d_1}$ and $\mathcal{Y} \in \binom{[n_2]}{d_2}$. We know that exactly b rows have exactly $d_R + 1$ samples drawn without replacement, and $n_1 - b$ rows have exactly d_R samples drawn without replacement. We sum over the number of ways i rows in $\mathcal{R}$ are sampled $d_R + 1$ times, which is

$$\frac{\binom{b}{i}\binom{n_1-b}{d_1-i}}{\binom{n_1}{d_1}}.$$

There are $\binom{n_1}{d_1}$ ways to select d_1 rows, and exactly $\binom{b}{i}\binom{n_1-b}{d_1-i}$ ways for i rows to land in the b rows with d_R+1 samples and the rest to land in the n_1-b rows with only d_R samples. Then, each row is sampled independently. For the $d_1 - i$ rows which are sampled d_R times, there are $\binom{n_2-d_2}{d_R}$ out of $\binom{n_2}{d_R}$ ways for the sample set to avoid the pre-emptively chosen column set. The same product applies for the i rows with d_R+1 samples in the summation.

For simplicity of analysis, we prove that minimum distance erasure patterns are the worst-case erasure patterns for probability of failure when $n_1 \mid Q$. For any $\mathcal{E} \in \mathcal{E}_{\text{inc}}$ the probability of failure is

$$p_{\pi_{R,Q}}(\mathcal{E}) = \prod_{i=1}^{n_1} \frac{\binom{n_2-|r_i(\mathcal{E})|}{d_R}}{\binom{n_2}{d_R}}.$$

A minimum of d_2 symbols must be erased per row for $\mathcal{E}$ to be undecodable and inclusion-minimal. Each term of form $f(x) = \binom{n_2-x}{d_R}$ is discretely convex and is decreasing in x. Then, each term of the product is maximized by minimizing $|r_i(\mathcal{E}_{\min})|$ subject to the erasure pattern undecodability constraints. For a row $i \in \mathcal{U}$ with erased symbols, $|r_i(\mathcal{E})| \geq d_2$, and there must be at least d_1 erased rows. Thus, the minimum distance erasure patterns achieve both constraints with equality.

When $n_1 \nmid Q$, a similar argument may be applied. For both $\binom{n_2-d_2}{d_R+1}$ and $\binom{n_2-d_2}{d_R}$, each term may be individually minimized by having d_2 erasures per selected row. There are a minimum of d_1 rows which must be in the rectangular support. $\square$

Next, we examine a *biregular* sampler which enforces both row and column regularity. For certain parameter regimes, the biregular sampler offers further improvement in worst-case probability of failure compared to the row-regular sampler at a larger sampler cost, which we define in Section 3.4.

3.3 Biregular Sampler

In this section, we design and analyze a *biregular sampler* for 2D-RS codes. We examine our sampler in two parameter regimes:

- **Sparse-Sample Regime:** When $Q \leq \min(n_1, n_2)$, we call this regime the sparse-sample regime. Here, a light node samples a small, fixed number of samples per DAS round.

- **Linear (Asymptotic) Regime:** When $\frac{Q}{n_1 n_2} \to \rho$ for a fixed $\rho \in (0, 1-\delta)$, each light node samples a fixed fraction ρ of the block per DAS round.

We study the sparse-sample regime, where $Q \leq \min(n_1, n_2)$ samples for tractable analysis of our sampler. The sparse-sample regime is favorable for light nodes, as each light node minimizes the total download cost of a DAS sampling round. We prove that our sampler is optimal against minimum distance erasure patterns. Then, we analyze the probability of failure behavior of second-order erasure patterns and identify a code distance cross-over point in which minimum distance erasure patterns are not the worst-case erasure pattern for probability of failure. Finally, we state worst-case probability of failure bounds for our sampler.

In the linear (asymptotic) sampling regime, a light node samples a fixed fraction of the block. Some blockchain systems increase the number of transactions processed on the block or through external roll-ups by increasing the block size [KB25]. The linear sampling regime allows for analysis of probability of failure for such cases. Against minimum distance erasure patterns, we prove probability of failure improvement in the first order of the exponent compared to random sampling without replacement. Then, we prove that minimum distance erasure patterns are the asymptotic worst case erasure pattern for square, symmetric 2D-RS codes with distance approximately $\delta = \frac{d}{n} \leq \frac{1}{3}$.

Definition 3.5 (Biregular Sampler)**.** For an $[n_1 n_2, k_1 k_2]$ 2D-RS code and sample set $\mathcal{S} \subseteq [n_1] \times [n_2]$ of size Q, let $G_{\mathcal{S}} \subseteq K_{n_1,n_2}$ denote a bipartite graph with sampled positions (edges) denoted by set $\mathcal{S}$. Let $d_R = \lfloor \frac{Q}{n_1} \rfloor$ and $d_C = \lfloor \frac{Q}{n_2} \rfloor$. Then, the biregular sampler $\pi_{\mathcal{B},Q}$ draws uniformly from

$$\mathcal{B}_Q = \Big\{ G_{\mathcal{S}} \subseteq K_{n_1,n_2} : |r_i(\mathcal{S})| \in \{d_R, d_R + 1\} \ \forall i \in [n_1], |c_j(\mathcal{S})| \in \{d_C, d_C + 1\} \ \forall j \in [n_2] \Big\}. \tag{3.6}$$

When $d_R n_1 = d_C n_2$, then the biregular sampler draws a (d_R, d_C)-biregular graph uniformly at random.

The biregular sampler draws an approximately (d_R, d_C)-regular graph uniformly at random, allowing for both uniform row and column sample coverage.

Remark 3.6. When $n_1 = n_2 = n$ and $Q = n$, then $d_R = d_C = 1$. Then $\mathcal{B}_n$ is exactly the set of perfect matchings of $K_{n,n}$. Then, the biregular sampler is a uniform random permutation of $[n]$, and a failure event occurs when the permutation avoids $\mathcal{E}$.

3.3.1 Sparse-Sample Regime

The sparse-sample regime, where $Q \leq \min(n_1, n_2)$, is desirable for light nodes. Light nodes wish to minimize the number of samples and the corresponding download size required for data availability guarantees. While the biregular sampler improves probability of failure beyond the row-regular sampler for the class of minimum distance erasure patterns, higher-cardinality erasure patterns may yield a higher probability of failure than minimum distance erasure patterns under the biregular sampler. Therefore, we must evaluate the probability of failure for both the class of minimum distance erasure patterns and higher-cardinality erasure patterns.

We first give an exact probability of failure formula for an arbitrary erasure pattern $\mathcal{E} \subseteq [n_1] \times [n_2]$. We then provide the probability of failure of minimum and second-weight erasure patterns as direct corollaries.

Theorem 3.7 (Probability of Failure for Biregular Sampler)**.** *For an $[n_1 n_2, k_1 k_2]$ 2D-RS code and a biregular sampler $\pi_{\mathcal{B},Q}$ with $Q \leq \min(n_1, n_2)$ samples, fix an erasure pattern support $\mathcal{E} = \mathcal{U} \times \mathcal{V} \setminus \mathcal{A} \subseteq [n_1] \times [n_2]$ such that $|\mathcal{U}| = d_1 + u$ and $|\mathcal{V}| = d_2 + v$ for non-negative integers u and v. Here, we require $d_1 + u \leq n_1$ and $d_2 + v \leq n_2$. The probability of failure* *of $\mathcal{E}$ is*

$$p_{\pi_{\mathcal{B},Q}}(\mathcal{E}) = \frac{\sum_{i=0}^{\min(Q,d_1+u,d_2+v)} m_i(\mathcal{A}) M_{Q,i}(n_1,n_2,d_1+u,d_2+v)}{\binom{n_1}{Q}\binom{n_2}{Q}Q!},$$

where $m_i(\mathcal{A})$ denotes the number of ways to place i symbols in distinct row and column positions on set $\mathcal{A} \subseteq [n_1] \times [n_2]$, and

$$M_{Q,i}(n_1,n_2,s_1,s_2) = \sum_{x+y+z=Q-i,x,y,z\geq 0} \binom{s_1-i}{x}\binom{n_2-s_2}{x}x!\binom{n_1-s_1}{y}\binom{s_2-i}{y}y! \times \binom{n_1-s_1-y}{z}\binom{n_2-s_2-x}{z}z!.$$

Proof. For a fixed erasure pattern $\mathcal{E}$, the probability of failure of the biregular sampler is counting problem

$$\frac{|\{G_{\mathcal{S}} \in \mathcal{B}_Q : \mathcal{S} \cap \mathcal{E} = \emptyset\}|}{|\{G_{\mathcal{S}} \in \mathcal{B}_Q\}|}.$$

We count the number of placements of Q samples in distinct rows and columns which avoid $\mathcal{E}$. First, for an $m \times n$ grid for non-negative integers m and n, there are $\binom{m}{Q}\binom{n}{Q}Q!$ possible Q-placements on the grid for non-negative integer $Q \leq \min(m,n)$.

Now, we visualize the regions of the board as follows:

$$\begin{bmatrix} \mathcal{U}\times\mathcal{V} & \mathcal{U}\times\mathcal{V}^c \\ \mathcal{U}^c\times\mathcal{V} & \mathcal{U}^c\times\mathcal{V}^c, \end{bmatrix} \tag{3.7}$$

where $\mathcal{U}\times\mathcal{V}$ denotes an erasure pattern's rectangular support. First, assume that i of the Q samples are placed in $\mathcal{U}\times\mathcal{V}$.

Let $m_i(\mathcal{A})$ denote the number of ways to place i samples in distinct rows and columns in set $\mathcal{A}$. The remaining $Q-i$ samples must be placed in distinct rows and columns outside of $\mathcal{U}\times\mathcal{V}$. Let x denote the number of samples in $\mathcal{U}\times\mathcal{V}^c$, y denote the number of samples in $\mathcal{U}^c\times\mathcal{V}$, and z denote the number of samples in $\mathcal{U}^c\times\mathcal{V}^c$.

In $\mathcal{U}\times\mathcal{V}^c$, there are d_1+u-i available rows to place samples in, as i of the rows already have a sample in $\mathcal{U}\times\mathcal{V}$. Then, there are n_2-d_2-v allowed column placements. Then, there are $\binom{d_1+u-i}{x}\binom{n_2-d_2-v}{x}x!$ valid placements of x samples with distinct rows and columns in $\mathcal{U} \times \mathcal{V}^c$.

In $\mathcal{U}^c \times \mathcal{V}$, there are $n_1 - d_1 - u$ available rows to place samples in. Then, there are $d_2 + v - i$ available columns to place samples in. We then count $\binom{n_1-d_1-u}{y}\binom{d_2+v-i}{y}(y)!$ ways to place valid configurations.

For $\mathcal{U}^c \times \mathcal{V}^c$, suppose we take z samples. There are $n_1 - d_1 - u - y$ available rows, and $n_2 - d_2 - v - x$ available columns.

The formula applies generally for a forbidden square of size $s_1 \times s_2$. There must be a total of $x + y + z = Q - i$ samples taken in these three regions. Therefore, we sum over each (x, y, z) of sampler placements. For each (x, y, z), we multiply the number of sampling configurations permitted per region. $\square$

Then, we arrive at several corollaries for the minimum and low cardinality erasure patterns for a 2D-RS code.

Corollary 3.8 (Probability of Failure for Minimum Distance Undecodable Erasure Patterns)**.** *The probability of failure for a minimum distance erasure pattern $E \in \mathcal{E}_{min}$ using a biregular sampler $\pi_{\mathcal{B},Q}$ for $Q \leq \min(n_1, n_2)$ is*

$$p_{\pi_{\mathcal{B},Q}}(\mathcal{E}_{min}) = \frac{M_{Q,0}(n_1, n_2, d_1, d_2)}{\binom{n_1}{Q}\binom{n_2}{Q}Q!}. \tag{3.8}$$

For the second weight erasure patterns, we consider the case where $d_1 = d_2 = d$.

Corollary 3.9 (Probability of Failure for Second Weight Inclusion-Minimal Erasure Patterns)**.** *For an $[n_1n_2, k^2]$ 2D-RS code with $Q \leq \min(n_1, n_2)$ samples, the probability of failure for $\mathcal{E}_2$ is*

$$p_{\pi_{\mathcal{B},Q}}(\mathcal{E}_2) = \frac{\sum_{j=0}^{d+1}\binom{d+1}{j}M_{Q,j}(n_1, n_2, d+1, d+1)}{\binom{n_1}{Q}\binom{n_2}{Q}Q!} \tag{3.9}$$

For the sparse-sample regime, we prove that $\pi_{\mathcal{B},Q}$ is optimal against the set of minimum distance erasure patterns $\mathcal{E}_{\min}$. The biregular sampler is a sufficient condition for probability of failure optimality against $\mathcal{E}_{\min}$.

Theorem 3.10 (Sparse-Sample Regime Optimality)**.** *For an $[n_1 n_2, k_1 k_2]$ 2D-RS code and any sampling strategy π with $Q \leq \min(n_1, n_2)$ samples,*

$$p_\pi(\mathcal{E}_{min}) \geq \frac{M_{Q,0}(n_1, n_2, d_1, d_2)}{\binom{n_1}{Q}\binom{n_2}{Q}Q!}. \tag{3.10}$$

The bound is achieved with equality for the biregular sampler $\pi_{\mathcal{B},Q}$.

Proof. We first consider permutation-invariant sampling strategies. Let σ and τ denote a random row and column permutation. For sampling strategy π, let π' denote a sampling strategy which draws $\mathcal{S} \sim \pi$ and sample $\sigma \times \tau(\mathcal{S})$. We know that $\mathcal{E}_{\min}$ is invariant under a row and column permutation. Then, for every $\mathcal{E} \in \mathcal{E}_{\min}$

$$p_{\pi'}(\mathcal{E}) = \mathbb{E}_{\sigma,\tau}\left[p_\pi((\sigma \times \tau)^{-1}(\mathcal{E}))\right] \leq \max_{\mathcal{E}' \in \mathcal{E}_{\min}} p_\pi(\mathcal{E}).$$

Therefore, it suffices to only consider permutation invariant strategies.

For $Q \leq \min(n_1, n_2)$, the set of biregular sample sets is invariant under a random row and column permutation, and is equivalent to strategy $\pi_{\mathcal{S}}$ for $\mathcal{S} = \{(\sigma(i), \tau(i)) : i \in [Q]\}$ for row permutation σ and column permutation τ. Therefore, we apply Theorem 3.2 to express the probability of failure as

$$\frac{1}{\binom{n_2}{d_2}}\mathbb{E}_{\mathcal{R}}\left[\binom{n_2 - |f_{\mathcal{S}}(\mathcal{R})|}{d_2}\right],$$

where $f(\mathcal{R}) = \bigcup_{i\in\mathcal{R}} r_i(\mathcal{S})$ is the column coverage of samples in a row in $\mathcal{R}$. The function $g(x) = \binom{n_2-x}{d_2}$ is discretely convex and decreasing in x. Therefore,

$$g(x+2) + g(x) - 2g(x+1) \geq 0 \implies g(x+2) - g(x+1) \geq g(x+1) - g(x).$$

We now apply a switching argument to prove that placing samples in distinct rows and columns does not decrease probability of failure for $Q \geq 2$. Suppose we have sample set

$\mathcal{S} \notin \mathcal{B}_Q$ Select an edge $e = (i, j)$ such that either $|r_i(\mathcal{S})| > 1$ or $|c_j(\mathcal{S})| > 1$. Without loss of generality, we prove our argument for the case where $|r_i(\mathcal{S})| > 1$. As $\mathcal{S} \notin \mathcal{B}_Q$, there must exist at least one row or column with more than one sample. As $Q \leq \min(n_1, n_2)$, there must exist at least one $i' \in [n_1]$ such that $|r_{i'}(\mathcal{S})| = 0$.

We propose a forward switching $\{(i, j) \mapsto (i', j)\}$, which maps $\mathcal{S}$ to some $\mathcal{S}'$. We prove that $p_{\pi_{\mathcal{S}}}(\mathcal{E}_{\min}) - p_{\pi_{\mathcal{S}'}}(\mathcal{E}_{\min}) \geq 0$. When evaluating $\mathbb{E}_{\mathcal{R}}\left[\binom{n_2 - |f_{\mathcal{S}}(\mathcal{R})|}{d_2}\right]$, the only $\mathcal{R} \in \binom{[n_1]}{d_1}$ which will cause a change in the summation involve subsets $\mathcal{R} \in \binom{[n_1]\setminus\{i,i'\}}{d_1-1} \cup \{i'\}$ and $\mathcal{R} \in \binom{[n_1]\setminus\{i,i'\}}{d_1-1} \cup \{i\}$, where either i or i' is in $\mathcal{R}$, but not both.

Let $\mathcal{R}_{i,i'} = \binom{[n_1]\setminus\{i,i'\}}{d_1-1}$.

$$\binom{n_1}{d_1}\binom{n_2}{d_2}(p_{\mathcal{S}}(\mathcal{E}_{\min}) - p_{\mathcal{S}'}(\mathcal{E}_{\min})) =$$
$$\sum_{\mathcal{R} \in \mathcal{R}_{i,i'}} \binom{n_2 - |f_{\mathcal{S}}(\mathcal{R} \cup \{i\})|}{d_2} + \binom{n_2 - |f_{\mathcal{S}}(\mathcal{R} \cup \{i'\})|}{d_2}$$
$$- \binom{n_2 - |f_{\mathcal{S}'}(\mathcal{R} \cup \{i\})|}{d_2} - \binom{n_2 - |f_{\mathcal{S}'}(\mathcal{R} \cup \{i'\})|}{d_2}.$$

For a fixed $\mathcal{R} \in \mathcal{R}_{i,i'}$, we know that $|f_{\mathcal{S}}(\mathcal{R} \cup \{i\})| = |f_{\mathcal{S}'}(\mathcal{R} \cup \{i\})| + 1$, and that $|f_{\mathcal{S}}(\mathcal{R} \cup \{i'\})| = |f_{\mathcal{S}}(\mathcal{R})| = |f_{\mathcal{S}'}(\mathcal{R} \cup \{i'\})| - 1$ from our proposed switch. Let $a = |f_{\mathcal{S}}(\mathcal{R} \cup \{i\})|$, and let $b = |f_{\mathcal{S}}(\mathcal{R} \cup \{i'\})|$. Then,

$$\binom{n_1}{d_1}\binom{n_2}{d_2}(p_{\mathcal{S}}(\mathcal{E}_{\min}) - p_{\mathcal{S}'}(\mathcal{E}_{\min})) =$$
$$\sum_{\mathcal{R} \in \mathcal{R}_{i,i'}} \binom{n_2 - a}{d_2} + \binom{n_2 - b}{d_2} - \binom{n_2 - a + 1}{d_2} - \binom{n_2 - b - 1}{d_2}.$$

By the definition of a and b, we know that $a \geq b$. Suppose that $a = b + k$ for some integer $k \geq 0$. Then, we apply the discrete convexity inequality k times to lower bound the difference. The difference must be ≥ 0 for each term in the summation. Therefore, the probability of $\mathcal{S}'$ is equal to or lower after applying a forward switch.

For a sample set $\mathcal{S} \notin \mathcal{B}_Q$, suppose we apply repeated row and column switchings. Then, the probability of failure is non-increasing for each switch. Therefore, a sample set $\mathcal{S}'$ with samples in distinct rows and columns must have equal or lower probability of failure to any sample set $S \notin \mathcal{B}_Q$. For $Q \leq \min(n_1, n_2)$, each $\mathcal{S} \in \mathcal{B}_Q$ is equivalent under a row and column permutation. Therefore, the entire set $\mathcal{B}_Q$ has an equal probability of failure which is optimal.

$\square$

For the sparse-sample regime, our biregular sampler is optimal against the set of minimum distance erasure patterns. However, we must also consider higher-order erasure patterns in our analysis, as each may have a higher probability of failure than $\mathcal{E}_{\min}$ under certain parameter regimes. We demonstrate that the second-order inclusion-minimal erasure patterns can have worse probability of failure than minimum distance erasure patterns for the biregular sampler. We compare the probability of failure of the biregular sampler for the first versus second weight erasure patterns. We focus on a symmetric, square 2D-RS code, such that $n_1 = n_2 = n$ and $d_1 = d_2 = d$. Additionally, we impose the constraint that $Q = n$ samples. We examine this case for simplicity of analysis.

Theorem 3.11 (Second-Weight Erasure Comparison). *Let $n_1 = n_2 = Q = n$, $d_1 = d_2 = d$, and $\frac{d}{n} \to \delta \in (0, \frac{1}{2})$. Let $\mathcal{E} \in \mathcal{E}_{min}$ denote a fixed minimum distance erasure pattern, and let $\mathcal{E}_2$ denote an inclusion-minimal second-weight erasure pattern. Then,*

$$\frac{p_{\pi_{\mathcal{B},n}}(\mathcal{E}_2)}{p_{\pi_{\mathcal{B},n}}(\mathcal{E}_{min})} = \frac{1}{(n-d)^2} \sum_{k=0}^{d+1} \binom{d+1}{k} \frac{(n-2d)!}{(n-2d-2+k)!} \tag{3.11}$$

As $\frac{d}{n} \to \delta$,

$$\frac{p_{\pi_{\mathcal{B},n}}(\mathcal{E}_2)}{p_{\pi_{\mathcal{B},n}}(\mathcal{E}_{min})} \to \left(\frac{1-2\delta}{1-\delta}\right)^2 \exp(\frac{\delta}{1-2\delta}). \tag{3.12}$$

The second support-class has higher probability of failure than the minimum distance erasure pattern for $\delta > \delta_2^*$, where

$$\delta_2^* \approx 0.417.$$

Proof. As $Q = n$, we may count the number of permutations in $[n] \times [n]$ which avoid a fixed $d \times d$ rectangle. Here $m_j(\mathcal{A}) = 0$ for every $j > 0$. Using Corollary 3.8 and Corollary 3.9,

$$p_{\pi_{\mathcal{B},n}}(\mathcal{E}_{\min}) = \frac{\binom{n-d}{d}}{\binom{n}{d}} = \frac{(n-d)!^2}{n!(n-2d)!} \tag{3.13}$$

$$p_{\pi\mathcal{B},n}(\mathcal{E}_2) = \frac{1}{n!} \sum_{k=0}^{d+1} \binom{d+1}{k} \frac{(n-d-1)!^2}{(n-2d-2+k)!}. \tag{3.14}$$

The simplified version of $p_{\pi\mathcal{B},n}(\mathcal{E}_2)$ is derived using rook-theory equations for perfect matchings [HR01]. There are $\binom{d+1}{k}$ partial k-matchings in $\mathcal{U} \times \mathcal{V}$. The number of permutations which avoid $\mathcal{E}_2$ is

$$(n-d-1)!^2 \sum_{k=0}^{d+1} \binom{d+1}{k} \frac{1}{(n-2d-2+k)!}.$$

Therefore, by taking the ratio of the two permutation counts,

$$\frac{p_{\pi\mathcal{B},n}(\mathcal{E}_2)}{p_{\pi\mathcal{B},n}(\mathcal{E}_{\min})} = \frac{1}{(n-d)^2} \sum_{k=0}^{d+1} \binom{d+1}{k} \frac{(n-2d)!}{(n-2d-2+k)!}.$$

The asymptotic analysis is in Appendix A.1. □

Theorem 3.11 demonstrates that there exists a code distance crossover point in which minimum distance erasure patterns do not define worst-case probability of failure. Asymptotically, we find that for $\delta < \delta_2^*$, minimum distance erasure patterns have worse probability of failure than second-weight erasure patterns.

So far, we have only considered the two lowest weight erasure pattern structures. For our upper bound analysis, we once again use a square, symmetric $[n^2, d^2]$ 2D-RS code. Then, we may use tools from perfect matching literature to upper bound our sampler's worst case probability of failure.

To evaluate the performance of our sampler, we find upper bounds on the worst-case probability of failure against higher order erasure patterns.

Lemma 3.12 (Upper Bounds for Worst-Case Erasure Patterns)**.** *For an* $[n^2, k^2]$ *2D-RS code and inclusion-minimal erasure pattern* $\mathcal{E}$*, a worst case partial-matching upper bound for* $m_j(\mathcal{A})$ *and a given erasure support size is*

$$m_k(\mathcal{A}) \leq \min\left(\binom{d+u}{k}(\frac{h_m}{d+u})^k, \binom{d+v}{k}(\frac{h_m}{d+v})^k\right) \tag{3.15}$$

$$m_k(\mathcal{A}) \leq \min\left(\binom{d+u}{k}\frac{(d+v)!^{\frac{d+v-k}{d+v}}}{(d+v-k)!}(v!)^{\frac{k}{v}}, \binom{d+v}{k}\frac{(d+u)!^{\frac{d+u-k}{d+u}}}{(d+u-k)!}(u!)^{\frac{k}{u}}\right) \tag{3.16}$$

where $h_m = \min((d+u)v, (d+v)u)$*.*

Proof. For set $\mathcal{A}$, let $(a_1, \ldots, a_r)$ denote the row degrees for $r = \max(u, v)$. For any fixed set $\mathcal{R}$ of i rows, there are at most $\prod_{y\in\mathcal{R}} a_y$ matchings which fully saturate those rows. Therefore,

$$m_j(\mathcal{A}) \leq \sum_{\substack{\mathcal{R}\subseteq\mathcal{U} \\ |\mathcal{R}|=j}} \prod_{y\in\mathcal{R}} a_y.$$

Then, by applying MacLaurin's inequality,

$$\frac{1}{\binom{d+u}{j}} \sum_{\substack{\mathcal{R}\subseteq\mathcal{U} \\ |\mathcal{R}|=j}} \prod_{y\in\mathcal{S}} a_y \leq \binom{d+u}{j}\left(\frac{a_1 + \ldots a_r}{d+u}\right)^j.$$

Then, there are a maximum of $(d+u)v$ erasures per row or $(d+v)u$ erasures per column.

For the second set of bounds, we first select i rows in $\mathcal{A}$ as a set $\mathcal{R}_i$. Then, we evaluate the corresponding $\mathcal{R}_i \times \mathcal{V}$ submatrix of $\mathcal{A}$, which we denote $A_{\mathcal{R}_i}$. To make the submatrix square, we append $|\mathcal{V}| - i$ extra rows which are all ones (allowed positions). Then,

$$B_i = \begin{bmatrix} A_i \\ \mathbf{1}_{|\mathcal{V}|-i,|\mathcal{V}|} \end{bmatrix}$$

is the square completion of A_i. For any fixed partial matching in A_i, there exists $(C-i)!$ perfect matchings in B_i. Therefore, $\mathrm{Per}(B_i) = (C-i)!\,\mathrm{Per}(A_i)$. Then, using the Bregman-Minc inequality [Bre73] for square matrix permanents,

$$\mathrm{Per}(B_i) \leq (v!)^{\frac{i}{v}}(d+v)!^{\frac{d+v-i}{d+v}}.$$

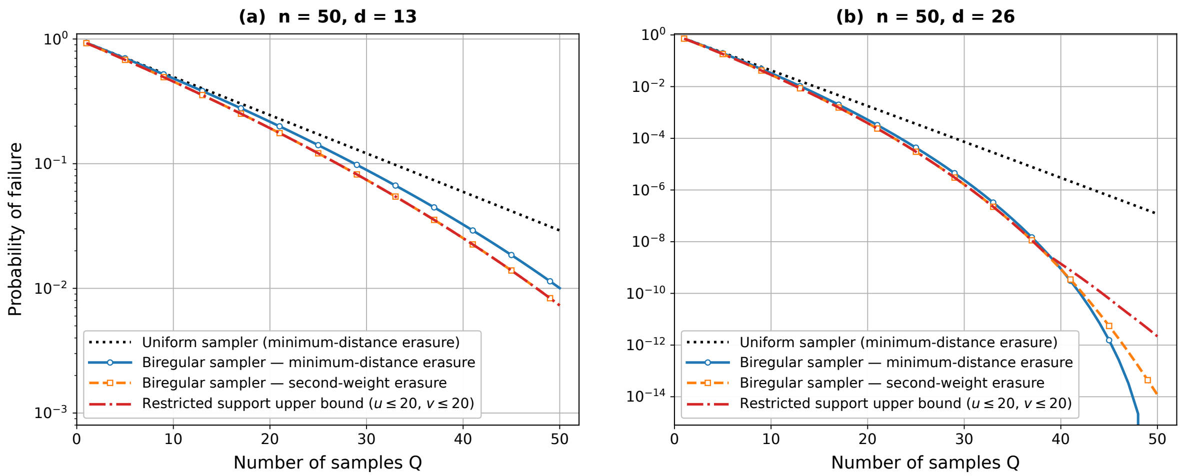


Figure 3.1: Probability of Failure and Upper Bound Comparisons for a $[50^2, 38^2]$ 2D-RS code and $[50^2, 25^2]$ 2D-RS code for $u, v \leq 20$

Finally, $\mathrm{Per}(A_i) \leq \frac{1}{(d+v-i)!}(v!)^{\frac{i}{v}}(d+v)!^{\frac{d+v-i}{d+v}}$. Then, there are $\binom{d+u}{i}$ ways to select i rows from the rectangular support. $\square$

Lemma 3.12 is a worst-case bound over all inclusion-minimal erasure patterns which may be found for any rectangular support size. We supply two upper bounds. The Maclaurin upper bound is favorable for small u, v. The Bregman-Minc inequality for matrix permanents is more accurate for larger u, v. The distance crossover is visible for finite blocklengths. In Figure 3.1, the worst-case probability of failure for a biregular sampler using a $[50^2, 38^2]$ 2D-RS code is found to be the set of minimum distance erasure patterns. For a $[50^2, 25^2]$ 2D-RS code, however, both the second-cardinality and worst-case probability of failure bounds are larger than those of the minimum distance erasure patterns.

3.3.2 Asymptotic-Sample Regime

We study the asymptotic properties of the biregular sampler. Evaluating the exact finite performance of the biregular sampler requires a matrix permanent, which is $\#P$-Complete to determine [Val79]. First, we derive an asymptotic bound for the probability of failure of a biregular sampler. We use the bound as a point of comparison for the exponential rate of decay of the probability of failure compared to the row-regular and uniform random sampler. Then, we prove that for code distance $\frac{d}{n} < \frac{1}{3}$, minimum distance erasure patterns are the worst-case erasure patterns for probability of failure performance.

We consider the regime where we sample a fixed fraction, ρ, of the total blocklength. We do not asymptotically analyze the sparse-sample regime, as the sampled fraction of the blocklength $\frac{Q}{n_1 n_2} \to 0$ as the blocklength increases.

We specify the asymptotic regime where $\frac{d_1}{n_1} \to \alpha, \frac{d_2}{n_2} \to \beta$, and $\frac{Q}{n_1 n_2} \to \rho$. For our analysis, we require $0 \leq \alpha, \beta, \rho \leq 1$, where $\alpha + \beta \leq 1$. We examine probability of failure in the asymptotic regime in which $d_1 = \Theta(n_1)$, $d_2 = \Theta(n_2)$, and $d_R = \Theta(n)$, where the number of samples is linear in the blocklength.

In this section, we assume the biregular sampler $\pi_{\mathcal{B},Q}$ is exactly biregular (i.e., $d_R n_1 = d_C n_2$ for $d_R = \frac{Q}{n_1}$ and $d_C = \frac{Q}{n_2}$). We study strictly biregular samplers for cleaner asymptotic analysis.

Theorem 3.13 (Asymptotic Minimum Distance Probability of Failure)**.** *For $\alpha, \beta > 0$, $\alpha + \beta < 1$, and $0 < \rho < \min(1-\alpha, 1-\beta)$, the asymptotic probability of failure for any minimum distance erasure pattern $E \in \mathcal{E}_{min}$ goes to*

$$-\frac{1}{n_1 n_2}\log(p_{\pi_{\mathcal{B},Q}}(\mathcal{E}_{min})) = I_B(\alpha, \beta, \rho) + o(1), \tag{3.17}$$

for

$$\begin{aligned} I_B(\alpha, \beta, \rho) = h(\rho) - \alpha(1-\beta)h(\frac{\rho}{1-\beta}) - \beta(1-\alpha)h(\frac{\rho}{1-\alpha}) \\ - (1-\alpha)(1-\beta)h(\frac{\rho(1-\alpha-\beta)}{(1-\alpha)(1-\beta)}). \end{aligned}$$

Proof. To find the asymptotic probability of failure, we consider the adjacency matrix of graph $G_{\mathcal{S}} \in \mathcal{B}_Q$. The adjacency matrix $M_{\mathcal{S}} \in \{0,1\}^{n_1 \times n_2}$ is defined as

$$(M_{\mathcal{S}})_{ij} = \begin{cases} 1 & (i,j) \in \mathcal{S} \\ 0 & \text{otherwise.} \end{cases} \tag{3.18}$$

The probability of failure is a counting argument for the ratio of biregular graphs which avoid $\mathcal{E} \in \mathcal{E}_{\min}$ with the total number of biregular matrices. First, any $\mathcal{E} \in \mathcal{E}_{\min}$ is equivalent to a binary matrix with forbidden positions in the first $d_1 \times d_2$ indexed square of the grid. Then, a biregular sample set which avoids a minimum distance erasure pattern can be represented as a block matrix

$$\begin{array}{cc} & \begin{array}{cc} d_2 & n_2 - d_2 \end{array} \\ \begin{array}{c} d_1 \\ n_1 - d_1 \end{array} & \begin{bmatrix} 0 & B \\ C & D \end{bmatrix} \end{array}$$

Our proof relies on the biregularity of the sample set, and the fact that no samples can be in the $d_1 \times d_2$ erased region for a probability of failure event to occur. First, we know that there must be d_R samples per row, and d_C samples per column. We condition on the fact that each row must have d_R samples, and each column must have d_C samples. In block matrix B, there must be exactly $d_1 d_R$ samples. In block matrix C, there must be exactly $d_2 d_C$ samples, as no samples can be in the $d_1 \times d_2$ region. In block matrix D, we must have $Q - d_1 d_R - d_2 d_C$ remaining samples. We define the density of samples in each block as

$$\rho_B = \frac{d_R d_1}{d_1(n_2 - d_2)} \to \frac{\rho}{1-\beta} \tag{3.19}$$

$$\rho_C = \frac{d_C d_2}{d_2(n_1 - d_1)} \to \frac{\rho}{1-\alpha} \tag{3.20}$$

$$\rho_D = \frac{Q - d_R d_1 - d_C d_2}{(n_1 - d_1)(n_2 - d_2)} \to \frac{\rho(1-\alpha-\beta)}{(1-\alpha)(1-\beta)}. \tag{3.21}$$

Next, we use a maximum entropy approach, derived from [Bar10], to count the number of biregular sample sets which avoid $\mathcal{E}$. Let $W_{\mathcal{E}}$ denote the allowed positions, and

$\Sigma(d_R, d_C; W_{\mathcal{E}}) \subseteq \mathcal{B}_Q$ denote the set of (d_R, d_C)-biregular matrices which avoid $\mathcal{E}$. Then, Barvinok [Bar10] states that

$$\log|\Sigma(d_R, d_C; W_{\mathcal{E}})| \leq \max_{Z \in \mathcal{P}(d_R, d_C; W)} \sum_{i,j} h(z_{ij}) + o(n_1 n_2),$$

for maximum entropy matrix Z subject to polytope constraints $\mathcal{P}(d_R, d_C; W)$. Here, $Z \in [0,1]^{n_1 \times n_2}$, and $H(Z) = \sum_{ij} h(z_{ij})$ denotes the sum of the natural entropy of each matrix element. The maximum entropy is constrained by the row and column sum and forbidden (erased) position constraints, which are characterized using polytope

$$\mathcal{P}(d_R, d_C; \mathcal{E}) = \Big\{ X \in [0,1]^{n_1 \times n_2} : \sum_j x_{ij} = d_R, \sum_i x_{ij} = d_C, \\ x_{ij} = 0 \text{ whenever } (w_{\mathcal{E}})_{ij} = 0 \Big\}$$

The natural entropy function is strictly concave. Therefore, for all positions (i, j) in block matrix B, we know that $z_{ij} = \rho_B$ will maximize the total entropy sum. Here, $z_{ij} = \rho_B$ is strictly feasible in polytope $\mathcal{P}(d_R, d_C; \mathcal{E})$, as $(n_2 - d_2)\rho_B = d_R$ per row. The column constraints for block matrix C are strictly feasible and the entropy sum is maximized for $z_{ij} = \rho_C$ for all $(i, j) \in C$. Then, ρ_D enforces both row and column regularity in block matrix D. Therefore, the maximum entropy matrix is

$$H(Z_{\mathcal{E}}) = n_1 n_2 \left[\alpha(1-\beta)h(\rho_B) + \beta(1-\alpha)h(\rho_C) + (1-\alpha)(1-\beta)h(\rho_D)\right] \tag{3.22}$$

For each block matrix, the maximum entropy z_{ij} within the block is the same. Therefore, we may multiply each $h(\rho_F)$ for block matrix $F \in \{B, C, D\}$ by the cardinality of F.

The total number of (d_R, d_C)-biregular sampler sets has a maximum entropy value of $z_{ij} = \rho$ for all $(i, j) \in [n_1] \times [n_2]$ within the feasible polytope, which follows from the strict concavity of entropy and required row and column regularity constraints. The cardinality of the number of (d_R, d_C)-biregular sampler sets may be lower bounded by [Bar10] as

$$|\Sigma(d_R, d_C)| \geq (n_1 n_2)^{\gamma(n_1+n_2)} e^{H(Z)}. \tag{3.23}$$

Then,

$$\log|\Sigma(d_R, d_C)| \geq H(Z) + \gamma(n_1+n_2)\log(n_1 n_2),$$

where $\gamma(n_1+n_2)\log(n_1n_2) = O(n_1 \log n_1)$. For the biregular matrices which avoid $\mathcal{E}$,

$$\log|\Sigma(d_R, d_C)| \leq H(Z_{\mathcal{E}}).$$

Then, by dividing both terms, we find that

$$-\frac{1}{n_1n_2}\log p_{\pi_{\mathcal{B},Q}}(\mathcal{E}_{\min}) = h(\rho) - \alpha(1-\beta)h(\rho_B) - \beta(1-\alpha)h(\rho_c) \tag{3.24}$$

$$-(1-\alpha)(1-\beta)h(\rho_D) + o(1). \tag{3.25}$$

□

Our asymptotic analysis for the rate of exponential decay of $p_{\pi_{\mathcal{B},Q}}(\mathcal{E}_{\min})$ is accurate to the first order of the exponent. Our asymptotic bound is primarily useful for comparison of the rate of exponential decay against uniform random sampling without replacement and the row-regular sampler. We use the results of Theorem 3.13 and maximum entropy methods [Bar10] to prove that no higher-order erasure pattern is worse when $\delta < \delta_2^*$.

For a square, symmetric 2D-RS code, we prove that the minimum distance biregular probability of failure is the worst-case probability of failure for $\delta < \frac{1}{3}$.

Theorem 3.14. *Let $n_1 = n_2 = n$, $d_1 = d_2 = d$, and $\delta = \frac{d}{n}$. We assume the biregular sampler is uniform over simple a-regular bipartite graphs, such that $Q = an$ and $\rho = \frac{a}{n}$.*

Let (δ, ρ) lie in a compact subset of

$$\Omega = \left\{0 < \delta < \frac{1}{2}, 0 < \frac{\rho}{1-\delta} < \frac{1}{2}, \max_{c,l}(F(c,l), G(c,l)) < 0\right\},$$

for every $0 \le c \le \frac{\delta}{1-\delta}$ *and* $0 \le l \le \frac{\rho}{1-\delta}$*. We define* $F(c,l)$ *and* $G(c,l)$ *as*

$$\begin{aligned} F(c,l) = &-\frac{c}{1+c}\log(1-p_{c,l}) - \frac{2(1-c)}{1+c}\log(1-l) \\ &+ \frac{2}{1+c}\log(1-t_{c,l}) - \frac{l}{1+c}\log\left(\frac{p_{c,l}(1-t_{c,l})}{t_{c,l}(1-p_{c,l})}\right), \end{aligned} \tag{3.26}$$

$$\begin{aligned} G(c,l) = &-\left(1+\frac{(1+c)l}{p_{c,l}}\right)\log(1-p_{c,l}) \\ &+ 4\log(1-l) - 2\log(1-t_{c,l}) - cl\log\left(\frac{p_{c,l}(1-t_{c,l})}{t_{c,l}(1-p_{c,l})}\right). \end{aligned} \tag{3.27}$$

Then, there exists a constant $m > 0$ *and* $n \ge n_0$ *such that the following holds. If an inclusion-minimal erasure pattern* E *has* $\max(|\mathcal{U}(E)|, |\mathcal{V}(E)|) = d + r$*, then*

$$\log \frac{p_{\pi_{\mathcal{B},Q}}(\mathcal{E})}{p_{\pi_{\mathcal{B},Q}}(\mathcal{E}_{min})} \le -mnr.$$

Proof. The technical details of this proof are in Appendix B We provide a brief proof sketch. We use the maximum entropy counting methods of [Bar10] to upper and lower bound the number of failed sample sets for a given erasure pattern $\mathcal{E}$.

First, for any arbitrary inclusion-minimal $\mathcal{E}$, we upper bound the number of failed sample sets by upper bounding the maximum entropy optimization problem. We define and upper bound the convex dual, in which row and column slack variables θ_1 and θ_2 define probabilities (p, l, t) for symbol entropy in the primal program based on the symbol entropies in $\mathcal{U} \times \mathcal{V}$, $\mathcal{U}^c \times \mathcal{V}$ and $\mathcal{U} \times \mathcal{V}^c$, and $\mathcal{U}^c \times \mathcal{V}^c$.

Then, we apply transformations to the dual bound to evaluate when its derivative is negative, and higher order erasure patterns are not the worst-case probability of failure. □

Using Theorem 3.14, we provide sufficiency conditions for when minimum distance erasure patterns are the worst-case probability of failure for a code. This region is desirable, as the minimum distance erasure pattern behavior is tractable.

Importantly, we may use Theorem 3.14 to determine fully analytic regions in which minimum distance erasure patterns have the worst-case probability of failure. By using simple upper bound on $F(c,l)$ and $G(c,l)$, we prove that minimum distance erasure patterns are the worst-case probability of failure case for 2D-RS codes with $\delta < \frac{1}{3}$ and $\rho < \frac{1-3\delta}{2}$.

Theorem 3.15 (Analytic Boundary Region)**.** *For $\delta > 0, \rho > 0,$ and $3c + 2\rho \leq 1,$ both $F(c,l) < 0$ and $G(c,l) < 0$ for all $0 < c \leq \frac{\delta}{1-\delta}$ and $0 < l \leq \frac{\rho}{1-\delta}$.*

Then, minimum distance erasure patterns $\mathcal{E}_{min}$ have the largest probability of failure value compared to any other $\mathcal{E} \in \mathcal{E}_{min}$ in the parameter region $\delta < \frac{1}{3}$ and $\rho < \frac{1-3\delta}{2}$.

Proof. As $c \leq \frac{\delta}{1-\delta}$ and $\rho \leq \frac{\rho}{1-\delta}$, we know that $c + l \leq \frac{\delta+\rho}{1-\delta}$. Let us assume that $c < \frac{1}{2}$.

From Definition B.9, we are given that $p_{c,l} = \frac{l(1-(1-c)l)}{(1-c)+(2c-1)l}$ and that $t_{c,l} = (1-c)l$ for entropy optimizers $p_{c,l}$ and $t_{c,l}$. First, we substitute the values for $p_{c,l}$ and $t_{c,l}$ into $F(c,l)$ to find

$$(1+c)F(c,l) = -c\log(1-p_{c,l}) - 2(1-c)\log(1-l) + 2\log(1-t_{c,l}) - l\log(\frac{p_{c,l}(1-t_{c,l}))}{t_{c,l}(1-p_{c,l})}) \tag{3.28}$$

$$= c\log\left(1 + \frac{(2c-1)l}{1-c}\right) + 2(1-l)\log\left(\frac{1-(1-c)l}{1-l}\right) + 2l\log(1-c). \tag{3.29}$$

Then, we apply the following logarithm bounds. For any constant $y > -1$, $\log(1+y) \leq y$. Then, for any constant $y' < 1$, $\log(1-y') \leq -y'$. We apply the following inequalities to find

$$(1+c)F(c,l) < \frac{c(2c-1)l}{(1-c)} + 2cl - 2cl = \frac{c(2c-1)l}{(1-c)}. \tag{3.30}$$

Using the assumption that $c < \frac{1}{2}$, term $(2c-1) < 0$, and $F(c,l)$ is negative. We choose $c < \frac{1}{2}$ as a constraint to enforce $F(c,l) < 0$. Next, we use the fact that

$$\frac{(1+c)l}{p_{c,l}} = 1 - \frac{c^2(1-2l)}{1-(1-c)l}.$$

Then, since $\log(1-y') \leq -y'$, we find that

$$-\left(\frac{(1+c)l}{p_{c,l}}-1\right)\log(1-p_{c,l}) =$$
$$\frac{c^2(1-2l)}{1-(1-c)l}\log(1-p_{c,l}) \leq -\frac{c^2(2c-1)l}{(1-c)l}p_{c,l} = \frac{c^2(2l-1)}{(1-c)+(2c-1)l}.$$

Then, the next set of terms we consider are

$$\begin{aligned}2(-\log(1-p_{c,l})+2\log(1-l)-\log(1-t_{c,l})) &= 2\log\left(\frac{(1-l)^2}{(1-p_{c,l})(1-t_{c,l})}\right)\\ &\leq 2\frac{c^2 l}{(1-c)(1-(1-c)l)}.\end{aligned}$$

Finally, we upper bound $\log(\frac{p_{c,l}(1-t_{c,l}))}{t_{c,l}(1-p_{c,l})})$ as

$$\begin{aligned}\log\left(\frac{p_{c,l}(1-t_{c,l}))}{t_{c,l}(1-p_{c,l})}\right) &= 2\log(1+\frac{cl}{1-l})-2\log(1-c)\\ &\geq \frac{2cl}{1-(1-c)l}+2c.\end{aligned}$$

Then,

$$\begin{aligned}G(c,l) = &-\left(1+\frac{(1+c)l}{p_{c,l}}\right)\log(1-p_{c,l})+4\log(1-l)\\ &-2\log(1-t_{c,l})-cl\log(\frac{p_{c,l}(1-t_{c,l}))}{t_{c,l}(1-p_{c,l})}) &(3.31)\\ < &\frac{c^2 l(1-2l)}{(1-c)+(2c-1)l}+2\frac{c^2 l}{(1-c)(1-(1-c)l)}-cl\left(\frac{2cl}{1-(1-c)l}+2c\right). &(3.32)\end{aligned}$$

We simplify the remaining expression to

$$\begin{aligned}G(c,l) &< c^2 l\left(-\frac{1-2l}{(1-c)+(2c-1)l}+\frac{2}{1-c}-2\right)\\ &= c^2 l\left(\frac{2c}{1-c}-\frac{1-2l}{(1-c)+(2c-1)l}\right).\end{aligned}$$

Then, when $c \leq \frac{1}{2}$, $0 < (1-c)+(2c-1)l \leq 1-c$. If we restrict $c+l \leq \frac{1}{2}$, then we also find that $1-2l \geq 2c$. Then, the boundary point for $G(c,l)$ to change signs is when

$$\frac{1-2l}{(1-c)+(2c-1)l} = \frac{2c}{1-c}.$$

The equivalent condition is $3\delta+2\rho \leq 1$. □

Using Theorem 3.15, we have proven that the region in which $\delta < \frac{1}{3}$ and $\rho < \frac{1-3\delta}{2}$ has a worst-case probability of failure defined by $\mathcal{E}_{\text{min}}$. We provide Theorem 3.15 as a concrete, analytical proof found only through manipulation of $F(c,l)$ and $G(c,l)$ with loose bounds on the logarithms. Our proven sampler performance region in Theorem 3.15 is a conservative region, as we require that both an upper bound on $F(c,l)$ and $G(c,l)$ must be less than zero. We conjecture that the admissible δ values in which minimum distance erasure patterns dominate worst-case probability of failure performance may be further sharpened by more careful analysis of Theorem 3.14.

As a point of comparison, we examine the asymptotic bounds for uniform random sampling without replacement, row-regular, and column-regular samplers.

Theorem 3.16 (Comparison Asymptotics)**.** *For an $[n^2, k^2]$ 2D-RS code, the row-regular and uniform random sampling asymptotics are*

$$-\frac{1}{n_1 n_2}\log p_{\pi_{U,Q}}(\mathcal{E}_{min}) = I_{unif} + O(\frac{1}{n_1 n_2}) \tag{3.33}$$

$$-\frac{1}{n_1 n_2}\log(p_{\pi_{R,Q}}(\mathcal{E}_{min}) = \alpha I_{row} + O(\frac{1}{\sqrt{n_1 n_2}}) \tag{3.34}$$

$$-\frac{1}{n_1 n_2}\log(p_{\pi_{C,Q}}(\mathcal{E}_{min}) = \beta I_{col} + O(\frac{1}{\sqrt{n_1 n_2}}) \tag{3.35}$$

for I_{unif} and I_{row} values

$$I_{unif} = h(\rho) - (1-\alpha\beta)h(\frac{\rho}{1-\alpha\beta}) \tag{3.36}$$

$$I_{row} = h(\rho) - (1-\beta)h(\frac{\rho}{1-\beta}) \tag{3.37}$$

$$I_{col} = h(\rho) - (1-\alpha)h(\frac{\rho}{1-\alpha}) \tag{3.38}$$

Proof. Both uniform random sampling without replacement and row-regular sampling have an exact finite probability of failure evaluation. Thus, we may directly apply Stirling asymptotics to each formula. We use the Stirling approximation

$$\log\binom{n}{k} = \log(1 + O(\frac{1}{n})) + \log(\sqrt{\frac{n}{2\pi k(n-k)}}) + k\log(\frac{n}{k}) + (n-k)\log(\frac{n}{n-k}), \tag{3.39}$$

when $k = \Omega(n)$ [Das16].

For a uniform sampler,

$$p_{\pi_{U,Q}} = \frac{\binom{n_1 n_2 - d_1 d_2}{Q}}{\binom{n_1 n_2}{Q}} = \frac{\binom{n_1 n_2 (1-\alpha\beta)}{\rho n_1 n_2}}{\binom{n_1 n_2}{\rho n_1 n_2}}.$$

Then, by applying Equation 3.39,

$$\begin{aligned} \log \frac{\binom{n_1 n_2(1-\alpha\beta)}{\rho n_1 n_2}}{\binom{n_1 n_2}{\rho n_1 n_2}} &= -n_1 n_2 h(\rho) + n_1 n_2 (1-\alpha\beta) h(\frac{\rho}{1-\alpha\beta}) \\ &+ \log(1 + O(\frac{1}{n_1 n_2})) + \log\left(\sqrt{\frac{(1-\alpha\beta)(1-\rho)}{1-\alpha\beta-\rho}}\right). \end{aligned} \tag{3.40}$$

Then, $\log(1+O(\frac{1}{n_1 n_2})) + \log\left(\sqrt{\frac{(1-\alpha\beta)(1-\rho)}{1-\alpha\beta-\rho}}\right)$ is a constant. The $O(\frac{1}{n_1 n_2})$ term contributes $O(1)$ after dividing by $n_1 n_2$. The *row-regular* probability of failure is

$$p_{\pi_{R,Q}}(\mathcal{E}_{\min}) = \left(\frac{\binom{n_2(1-\beta)}{\rho n_2}}{\binom{n_2}{\rho n_2}}\right)^{\alpha n_1}.$$

Then, we apply Stirling asymptotics such that

$$\alpha n_1 \log(\frac{\binom{n_2(1-\beta)}{\rho n_2}}{\binom{n_2}{\rho n_2}}) = -\alpha n_1 n_2 (h(\rho) - (1-\beta) h(\frac{\rho}{1-\beta})) + O(\sqrt{n_1 n_2}). \tag{3.41}$$

The α results from raising the combinatorial quantity to exponent $d_1 = \alpha n_1$. The term $\alpha n_1 O(1) = O(n_1)$. As $n_1 n_2$ increase at the same rate, this term is $O(\sqrt{n_1 n_2})$. Then, multiplying each side by $-\frac{1}{n_1 n_2}$,

$$-\frac{1}{n_1 n_2} \alpha n_1 \log\left(\frac{\binom{n_2(1-\beta)}{\rho n_2}}{\binom{n_2}{\rho n_2}}\right) = \alpha\left(h(\rho) - (1-\beta) h(\frac{\rho}{1-\beta})\right) + O\left(\frac{1}{\sqrt{n_1 n_2}}\right). \tag{3.42}$$

$\square$

The asymptotics for the column-regular sampler are found similarly.

We compare the performance of row-regular and column-regular samplers. The probability of failure improvement is determined by the row and column code distance to code length ratio.

Lemma 3.17 (Regime Comparison for Row vs Column Sampling)**.** *For* $\rho > 0$ *and* $0 < \alpha, \beta < 1$ *such that* $\rho < \min(1-\alpha, 1-\beta)$,

$$\alpha I_{row} > \beta I_{col} \iff \alpha < \beta \tag{3.43}$$

$$\tag{3.44}$$

Proof. Function $f(x) = H(\rho) - (1-x)H(\frac{\rho}{1-x})$ is strictly convex for $x \in [0, 1-\rho]$.

Let $g(x,t) = -tH(\frac{x}{t})$. As binary entropy is concave, $g(x,t)$ is jointly convex in both x and t (here, negative entropy is convex). So, $f(x)$ is convex in x, and $f(0) = 0$.

Suppose $\alpha I_{\text{row}} > \beta I_{\text{col}}$. Then, we know that

$$\frac{f(\beta)}{\beta} > \frac{f(\alpha)}{\alpha}. \tag{3.45}$$

As f is strictly convex and $f(0) = 0$, $\frac{f(x)}{x}$ is strictly increasing in x. Therefore, for $\alpha I_{\text{row}} > \beta I_{\text{col}}$, it holds that $\beta > \alpha$. The converse follows similarly. □

3.3.3 Results and Comparison

For a biregular sampler in the asymptotic regime, we compare the gains in the first order of the exponent between $I_{\text{bireg}}, I_{\text{row}}$, and I_{unif}. In Figure 3.4, for $\alpha = 0.3$ and $\beta = 0.2$, the first order exponent increases by a factor of 1.25 compared to uniform random sampling without replacement.

The contributions of row- versus column- regularity to the performance evaluation of the biregular sampler is dependent on the (α, β) regime of the code parameters. As seen in Figure 3.2, column-regular sampling has lower probability of failure than row-regular sampling when $\beta < \alpha$. In Figure 3.3, we find that $\alpha < \beta$, and row-regular sampling has a lower probability of failure.

Next, we empirically compare the probability of failure of the biregular sampler, row-regular sampler, and ISP sampler for a $[25^2, 20^2]$ 2D-RS code. As seen in Figure 3.5, probability of failure performance for the biregular sampler improves probability of failure for a

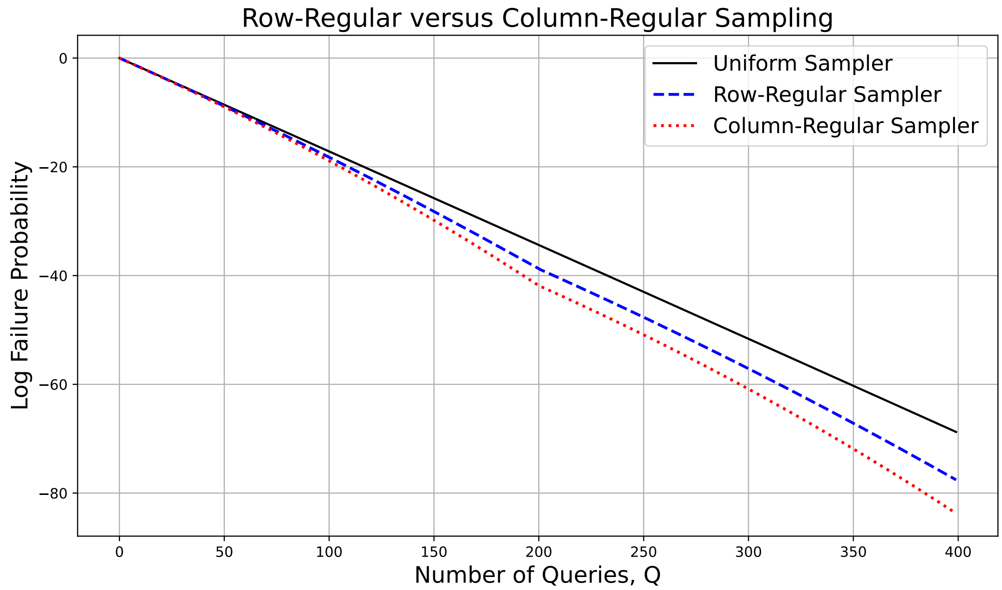


Figure 3.2: Probability of Failure Comparison for Row-Regular vs Column-Regular Samplers for $\alpha = \frac{d_1}{n_1} = \frac{90}{200}$ and $\beta = \frac{d_2}{n_2} = \frac{70}{200}$.

much larger number of samples than ISP. For ISP, the worst-case probability of failure upper bound degrades around $Q = 50$ samples, and the protocol terminates to random sampling without replacement. However, the biregular and row-regular sampler have continued improvement in probability of failure. In Figure 3.6, we visualize the lower-sample regime with $Q \leq 2n$. The biregular and ISP sampler have similar probability of failure in this regime. However, the biregular sampler has lower implementation complexity for $Q \geq \min(n_1, n_2)$, as it is not a greedy algorithm.

While the biregular sampler has an improved probability of failure, as proven in the square, symmetric 2D RS case with $\delta < 0.417$, the implementation complexity is higher than row-regular and column-regular samplers. Implementing a biregular sampler when sampling a fraction ρ of the blocklength can be implemented using Markov-Chain switches on a base graph, which has a mixing time of $O((an)^2 \log an)$ for regular bipartite graphs [Gre21].

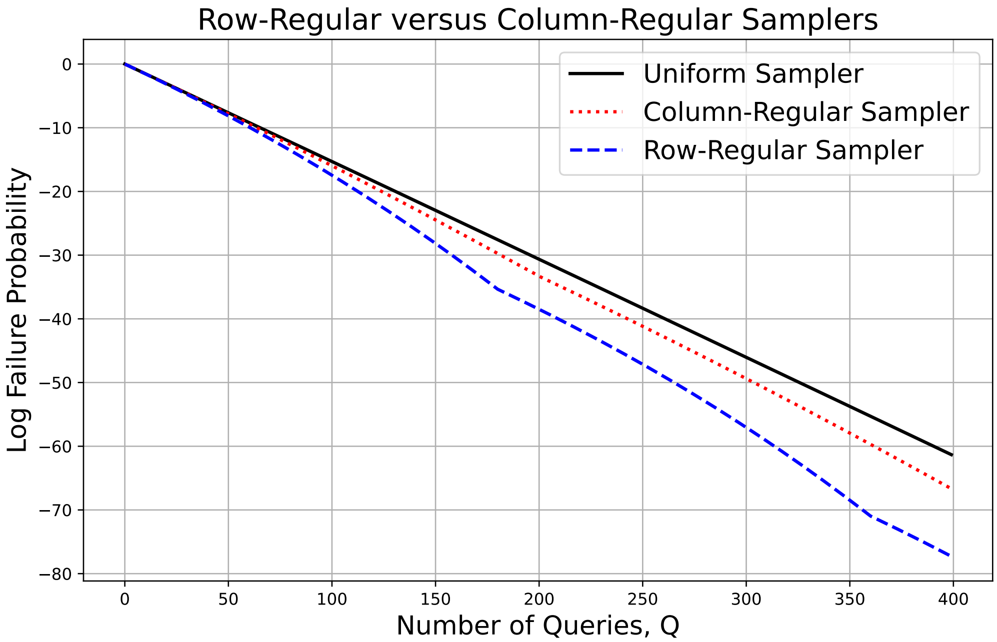


Figure 3.3: Probability of Failure Comparison for Row-Regular vs Column-Regular Samplers for $\alpha = \frac{d_1}{n_1} = \frac{51}{180}$ and $\beta = \frac{d_2}{n_2} = \frac{100}{200}$.

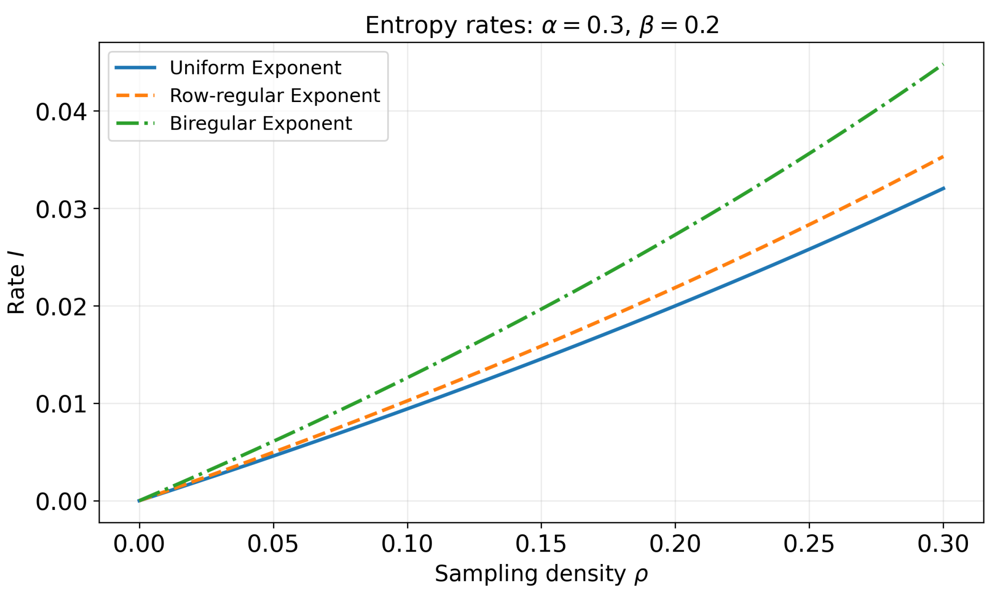


Figure 3.4: First Order Exponent Comparison for $I_{\text{unif}}, \alpha I_{\text{row}}, I_{\text{bireg}}$ for a 2D RS code with $\alpha = 0.3, \beta = 0.2$ for increasing ρ.

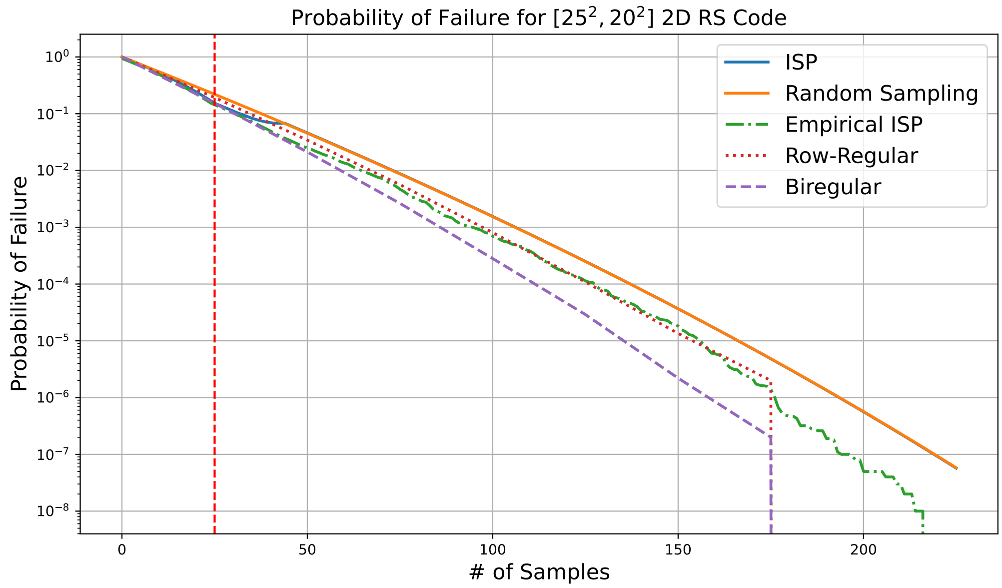


Figure 3.5: Empirical Probability of Failure Analysis for Minimum Distance Erasure Patterns

Finally, we compare our sampling algorithm for a continued number of samples to the *Improved Sampling Protocol*, as proposed in [MTD25].

3.4 Sampler Implementation Complexity

Each of the three samplers differs in implementation cost. Using random sampling without replacement algorithms in [Tin21], both uniform random sampling without replacement and row-regular sampling are implemented in $\Theta(Q)$ time and $\Theta(Q)$ space.

A biregular sampler uniformly draws from $\mathcal{B}_Q$. There does not exist a direct construction to draw a uniform graph from $\mathcal{B}_Q$. When sampling a fixed fraction ρ of the blocklength, the biregular sampler may be implemented using Markov-chain switches on a base biregular graph, with a mixing time of $O((an)^2 \log(an))$ for regular bipartite graphs [Gre21]. The time complexity is much larger than that of the row-regular sampler, $\Theta(Q)$. Compared to the ISP, however, the biregular sampler has lower sampling complexity, as ISP is a greedy algorithm which must calculate upper and lower probability of failure bounds for every code

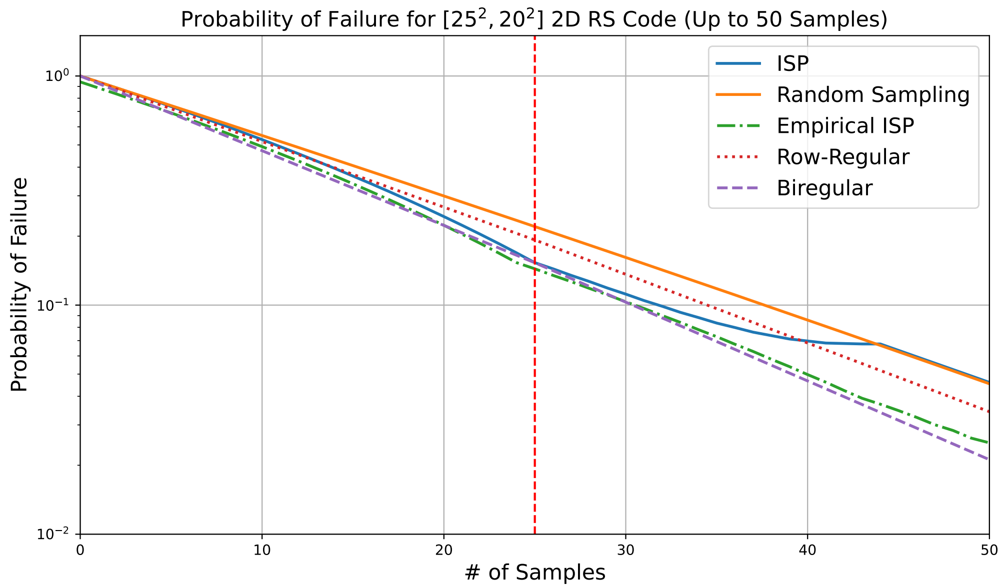


Figure 3.6: Empirical Probability of Failure Analysis for Minimum Distance Erasure Patterns for $Q \leq 50$ Samples

symbol at every sample step.

3.5 Scope and Open Problems

There are several open problems for the biregular sampler design problem. While the biregular sampler is optimal over $\mathcal{E}_{\min}$ for $Q \leq \min(n_1, n_2)$ samples, we do not claim the optimality of the biregular sampler outside of this regime. Optimal sampler design remains an open problem in the equivalent supersaturation literature [Nag17]. Optimal sampler design for $Q = an$ samples involves controlling the total column coverage, $f_{\mathcal{S}}(\mathcal{R})$, for all possible d_1-subsets $\mathcal{R}$, and corresponds to a t-design problem (as also found in [Nag17]). The asymptotic performance and performance against higher cardinality erasure patterns for refined biregular samplers remains an open problem.

Theorem 3.11 demonstrates that $\mathcal{E}_{\min}$ is not the worst-case class of erasures for all possible code rates in the sparse-sample regime. For our results, we particularly focus on the higher code rate regime, in which $\delta < \delta_2^*$, for our sampler analysis. The exact worst-case probability of failure for $\delta > \delta_2^*$ remains open.

For the linear, asymptotic sample regime, we prove a sufficient condition for minimum distance erasure patterns to be the worst-case probability of failure in Theorem 3.14. For a sufficiently large code with $\frac{d}{n} = \delta$ and $\frac{Q}{n^2} = \rho$ in the specified region of Theorem 4.3, we are able to asymptotically evaluate the worst-case probability of failure for the biregular sampler in Theorem 3.13 and demonstrate performance improvements over uniform random sampling.

An open problem in the asymptotic region is to fully evaluate the worst-case probability of failure for lower rate codes under a biregular sampler. We have currently only proven when minimum distance erasure patterns are the worst case for our sampler. Our Theorem 3.14 provides a sufficient condition for our sampler and the specified 2D-RS code to have our stated performance guarantees. However, there may exist a sharper analysis of the provided region. Our fully analytic proof in Theorem 4.3 uses a conservative bound $\log(1+y) \leq y$ to find upper bounds on both $F(c,l)$ and $G(c,l)$. The analytical region may be further tightened to prove performance guarantees for a larger δ or ρ.

CHAPTER 4

Joint Code-Sampler Design

In the previous chapter, we developed a sampler for the standard 2D-RS code. In this section, we design a sampler-aware 2D-RS subcode to further improve the probability of failure for the biregular sampler. For simplicity, we assume the 2D-RS code is square and symmetric.

Let $T \subseteq \mathbb{F}_q$ denote a set of n evaluation points in field $\mathbb{F}_q$. Let integer $2 \leq k < n$ and $d = n - k + 1$, denote the dimension and distance of a constituent Reed-Solomon code. The base 2D-RS code $\mathcal{C}_0$ is the evaluation on $T \times T$ of $V_{k,k}$.

All coefficients, check nodes, and stored symbols lie in $\mathbb{F}_q$. The necessary field size for our construction is given in Lemma 4.6.

An iterative row-column decoder, as considered in Chapter 3, does not consider heavy parity checks when decoding. We therefore analyze our subcode design under the ML-decoder.The set of ML-undecodable erasure patterns is a subset of the undecodable erasure patterns studied in Chapter 3 for 2D-RS codes.

For a $\pi_{B,n}$ biregular sampler with $Q = n$ distinct samples, the probability of failure for an erasure pattern $\mathcal{E} = \mathcal{U} \times \mathcal{V} \setminus \mathcal{A}$ is

$$p_{\pi_{B,n}}(\mathcal{E}) = \frac{1}{n!} \sum_{i=0}^{\min(n,d+u,d+v)} m_i(\mathcal{A}) M_{n,i}(n, n, d+u, d+v), \tag{4.1}$$

where $m_i(\mathcal{A})$ denotes the number of ways to place i samples in distinct rows and columns of $\mathcal{A}$.

A subcode $\mathcal{C} \subseteq \mathcal{C}_0$, which is defined by t independent parity checks, has dimension $k^2 - t$. Our design aim is to eliminate any surviving codeword support structures which have a large probability of failure. From Equation 4.1, the worst-case probability of failure upper bound may be lowered by restricting $m_j(\mathcal{A})$ or increasing the minimum required rectangular support size.

4.1 Codeword Factorization and Rectangular Support

An erasure pattern is ML-undecodable if and only if it contains the support of a nonzero codeword. For any two undecodable erasure patterns $\mathcal{E}$ and $\mathcal{E}'$ such that $\mathcal{E}' \subseteq \mathcal{E}$, probability of failure is non-increasing for $\mathcal{E}'$. Therefore, it suffices to study the property of codeword supports under ML-decoding. However, the structural properties of codeword supports for even a plain product code is an open problem in the maximal-recoverability literature [GHK17]. Thus, our subcode construction focuses on bounding each codeword support algebraically to be more favorable to the biregular sampler.

Any erasure pattern $\mathcal{E}$ may be characterized by its rectangular support, $\mathcal{U} \times \mathcal{V}$. In this section, which is used for both parity check designs, we connect the rectangular support interpretation of 2D-RS erasure patterns to the algebraic properties of the code. For every 2D-RS codeword whose support lies entirely within $\mathcal{U} \times \mathcal{V}$, the remaining rows and columns are all zero polynomials. Therefore, we may factor any codeword support.

Lemma 4.1 (Codeword Factorization)**.** *For a 2D-RS code C_0, let $f \in V_{k,k}$ satisfy $f(a,b) = 0$ for every $(a,b) \in (T \times T) \setminus (U \times V)$, and let $|U| = d + u$ and $|V| = d + v$ for $0 \leq u, v \leq k-1$. Then, there exists a polynomial $H \in \mathbb{F}_q[x,y]$ such that*

$$f(x,y) = V_U(x) V_V(y) H(x,y), \tag{4.2}$$

where $V_U(x)$ and $V_V(y)$ are vanishing polynomials on set $T \setminus U$ and $T \setminus V$

$$V_U(x) = \prod_{\tilde{u} \in T \setminus U} (x - \tilde{u}), \quad V_V(y) = \prod_{\tilde{v} \in T \setminus V} (y - \tilde{v}).$$

Additionally, $\deg_x(H) \leq u$ *and* $\deg_y(H) \leq v$*.*

Proof. For a codeword support f which is zero for rows evaluated at $T \setminus U$, take a field element $\tilde{u} \in T \setminus U$. Then, $f(\tilde{u}, y) = 0 \ \forall y \in T$. Here, $f(\tilde{u}, y) = 0$ has n distinct roots, and $\deg_y f < k$. Then, $f(\tilde{u}, y)$ is the zero polynomial, and $(x - \tilde{u})$ divides f. For every $\tilde{v} \in T \setminus V$, the proof follows similarly such that $(y - \tilde{v})$ divides f. Hence,

$$f(x, y) = \left(\prod_{\tilde{u} \in T \setminus U} (x - \tilde{u}) \prod_{\tilde{v} \in T \setminus V} (y - \tilde{v}) \right) H(x, y),$$

where $H(x, y)$ is a residual polynomial. Since $\deg_x(V_U) \leq n - d - u = n - (n - k + 1) - u = k - 1 - u$, we find $\deg_x(H) \leq u$. Similarly, $\deg_y(H) \leq v$. □

For any codeword with support $\mathcal{E} = \mathcal{U} \times \mathcal{V} \backslash \mathcal{A}$, the surviving polynomial $H(x, y)$ determines the zero placements $\mathcal{A}$ in the rectangular support. If we design parity checks to lower the x or y degree of H, then the surviving codewords must have a larger rectangular support. We design h parity checks to increase the required rectangular support size for all surviving codewords, and to tractably bound the maximum number of zeros per row or column in $H(x, y)$.

4.2 Rectangular Parities

Our subcode design uses two types of heavy parities. The first heavy parity type, which we develop in this section, enlarges the size of the rectangular support for every surviving codeword. Our parity construction allows for tractable worst-case probability of failure improvements based on each surviving erasure pattern's structure.

To design h parity checks to increase the rectangular support, we represent $H(x, y)$ in a bivariate Newton basis, which is a lower-triangular basis of the code [Nei19]. Suppose we have two sets of field elements $\eta = (\eta_0, \ldots, \eta_{M-1})$ and $\upsilon = (\upsilon_0, \ldots, \upsilon_{M-1})$, where each $\upsilon_i \neq \upsilon_j$ and $\eta_i \neq \eta_j$ for $i \neq j$. Then, each Newton basis term $N_{k,l}$ is

$$N_{k,l} = \prod_{i=0}^{k-1}(x - \upsilon_i) \prod_{j=0}^{l-1}(y - \eta_j).$$

Any bivariate polynomial $H(x, y)$ with $\deg_x(H) \leq M - 1$ and $\deg_y(H) \leq M - 1$ may be expressed using the Newton basis as

$$H(x, y) = \sum_{k=0}^{M-1} \sum_{l=0}^{M-1} c_{k,l} N_{k,l}$$

for coefficients $c_{k,l}$. A Newton basis is lower-triangular, as each term $c_{k,l} N_{k,l}$ vanishes at any υ_i for $i < k$ and at any η_j for $j < l$.

We impose constraints $f(\alpha, \beta) = 0$ for selected (α, β) pairs with $\alpha, \beta \in \mathbb{F}_q \setminus T$. For any codeword support factorization $f = V_{\mathcal{U}}(x)V_{\mathcal{V}}(y)H(x, y)$, both $V_{\mathcal{U}}(\alpha)$ and $V_{\mathcal{V}}(\beta)$ are non-zero, as $(\alpha, \beta) \notin T \times T$. Then, the constraint implies that $H(\alpha, \beta) = 0$. We use lower-triangular parity selection on off-grid points in order to enlarge the surviving rectangular support size and reduce the number of unerased symbols allowed within a rectangular support.

Theorem 4.2 (Rectangular Support Parity Construction)**.** *For base code $\mathcal{C}_0$ evaluated on $T \times T$, choose $0 \leq M \leq k - 1$ such that $h = \frac{M(M+1)}{2}$ parity checks are implemented. Then, select two sets of M off-grid elements*

$$(\upsilon_0, \ldots, \upsilon_{M-1}), \quad (\eta_0, \ldots, \eta_{M-1}),$$

where $\upsilon_i \notin T$ for $i \in \{0, \ldots, M - 1\}$ and $\eta_j \notin T$ for $j \in \{0, \ldots, M - 1\}$. Each set of elements is pairwise distinct. However, both υ and η may have common elements. Define a lower-triangular set of points

$$\Lambda = \{(\upsilon_i, \eta_j) : 0 \leq i \leq M - 1, 0 \leq j \leq M - 1, i + j \leq M - 1\}.$$

For $h = \frac{M(M+1)}{2}$ parity equations, impose parity constraint

$$f(a, b) = 0 \quad \text{for all } (a, b) \in \Lambda.$$

Then, for each surviving codeword support $\mathcal{E}$ with $|\mathcal{U}| = d + u$ and $|\mathcal{V}| = d + v$,

$$\prod_{i=0}^{(M-v)_+-1} (x - \upsilon_i) \mid H(x, y), \qquad \prod_{j=0}^{(M-u)_+-1} (y - \eta_j) \mid H(x, y).$$

Each surviving polynomial codeword may then be factored as

$$H(x, y) = V_\upsilon(x) V_\eta(y) \tilde{H}(x, y)$$

with degrees $\deg_x(\tilde{H}) \leq u - (M - v)_+$ *and* $\deg_y(\tilde{H}) \leq v - (M - u)_+$.

Proof. The proof follows from the Newton polynomial basis. First, each basis term

$$N_{k,l} = \prod_{i=0}^{k-1} (x - \upsilon_i) \prod_{j=0}^{l-1} (y - \eta_j)$$

has degree $\deg_x(N_{k,l}) = k$ and $\deg_y(N_{k,l}) = l$. Therefore, if $c_{k,l} \neq 0$, the degree of any surviving polynomial must be greater than or equal to (k, l).

For a fixed codeword support $\mathcal{E}$ when $i \leq M - v - 1$, the parity construction gives that $H(\upsilon_i, \eta_0) = H(\upsilon_i, \eta_1) = \dots H(\upsilon_i, \eta_{M-i-1}) = 0$. Then, the restricted univariate polynomial $H(\upsilon_i, y)$ is the zero polynomial, as the number of roots $M - i \geq v + 1$. Then, $(x - \upsilon_i) \mid H(x, y)$. A similar proof follows for all η variables. $\square$

As an example, for constraint $H(\upsilon_0, \eta_0) = 0$, we know that $c_{0,0} = 0$, as it is the only term which does not have a zero at υ_0 or η_0. Suppose there exists a codeword with rectangular support polynomial $H(x, y)$ with $\deg_x(H) = \deg_y(H) = 0$. As $H(\upsilon_0, \eta_0) = 0$, there is a contradiction.

Next, we set $H(\upsilon_1, \eta_0) = 0$ and $H(\upsilon_0, \eta_1) = 0$. Then,

$$H(\upsilon_1, \eta_0) = 0 = c_{0,0} + c_{1,0}(\upsilon_1 - \upsilon_0), \quad H(\upsilon_0, \eta_1) = 0 = c_{0,0} + c_{0,1}(\eta_1 - \eta_0).$$

As each υ_i is distinct, $(\upsilon_1 - \upsilon_0) \neq 0$, and $c_{0,0} = 0$ must hold from constraint $H(\upsilon_0, \eta_0) = 0$. Therefore, coefficients $c_{1,0}$ and $c_{0,1}$ are also forced to zero. Thus, any surviving polynomial has $\deg_x(\tilde{H}) + \deg_y(\tilde{H}) \geq 2$. The process follows similarly for the remaining parities.

The lower-triangular parity selection from Theorem 4.2 eliminates all codewords which have rectangular supports with x-degree $u < (M-v)_+$ or $v < (M-u)_+$. For $h = 1$ (i.e., $M = 1$), no codeword support with $u = v = 0$ survives. The code distance is at least $d(d+1)$.

Theorem 4.3 (Subcode Distance)**.** *For a 2D-RS subcode $\mathcal{C} \subseteq \mathcal{C}_0$ with $h = \frac{M(M+1)}{2}$ heavy parities selected using Theorem* 4.2*, the code distance is*

$$d(\mathcal{C}) \geq d(d+M). \tag{4.3}$$

Proof. We consider non-zero codeword support $E = \mathcal{U} \times \mathcal{V} \setminus A$, where $|\mathcal{U}| = d+u, |\mathcal{V}| = d+v$. If $v \geq M$, then the codeword must have at least $d+M$ active columns with support at least d. Then, $\mathrm{wt}(f) \geq d(d+M)$, where wt() is the Hamming weight.

Suppose $v < M$. Then, by Theorem 4.2, $\tilde{u} := u - (M-v)$. Then, each column in the rectangular support contains at most $\tilde{u}$ unerased symbols. The support size is at least $(d+u) - \tilde{u} = d + M - v$. Therefore, $\mathrm{wt}(f) \geq (d+v)(d+M-v) \geq d(d+M)$. □

Our parity construction provides several advantages for improving probability of failure. First, we know a lower bound on the minimum distance of $\mathcal{C}$ from Theorem 4.3. Second, for every surviving codeword, we additionally know that $\deg_x(\tilde{H}) \leq \tilde{u}$ and $\deg_y(\tilde{H}) \leq \tilde{v}$. Therefore, we have a maximum cap on the number of zero positions per row or per column of a surviving rectangular support, and can upper bound the worst-case $m_k(\mathcal{A})$ of every surviving rectangular support shape. The smallest surviving support shape, where $u + v = M$, is rectangular, since both $\tilde{u} = 0$ and $\tilde{v} := v - (M-u)_+ = 0$.

4.2.1 Matching-Based Parities

The rectangular parity design imposes restrictions on both the minimum support size of each codeword and the residual degree bounds $\tilde{u}, \tilde{v}$ of all surviving codewords. Next, we design *rook parities* to complement our rectangular parities. The rook parities are designed to constrain the leading coefficient term of $\tilde{H}$.

Our rook parity construction targets the leading coefficient layer in x and y, and aims to restrict codewords with many internal zero positions in the rectangular support. An undecodable erasure pattern for the iterative row-column decoder which is worst-case given our parity structural restrictions is not necessarily a codeword support. Thus, our construction improves the worst-case probability of failure.

Let $\Upsilon = (\upsilon_0, \dots, \upsilon_{M-1})$ and $H = (\eta_0, \dots, \eta_{M-1})$ denote the field elements used for the rectangular parity checks from Theorem 4.2.

Choose r_x distinct elements $\zeta_i \in \mathbb{F}_q \setminus (T \cup H)$ and r_y distinct elements $\varsigma_j \setminus \mathbb{F}_q \setminus (T \cup \Upsilon)$. Then, we define *rook*-parities

$$[x^{k-1}]f(\zeta) = 0, \qquad 1 \leq i \leq r_x, \tag{4.4}$$

$$[y^{k-1}]f(\zeta_i) = 0, \qquad 1 \leq j \leq r_y. \tag{4.5}$$

Here, $[x^{k-1}]f$ is a polynomial in y, and $[y^{k-1}]f$ is similarly a polynomial in x. Then, $r_x + r_y$ additional parity checks are added.

Lemma 4.4. *Let $f = V_U(x)V_V(y)V_\upsilon(x)V_\eta(y)\tilde{H}(x,y)$, and let $c_{\tilde{u}}(y) = [x^{\tilde{u}}]\tilde{H}(x,y)$. Then,*

$$[x^{k-1}]f(x,y) = V_V(y)V_\eta(y)c_{\tilde{u}}(y) \tag{4.6}$$

for the vanishing polynomials described above. If $[x^{k-1}]f(x,\zeta_i) = 0$, then $c_{\tilde{u}}(\zeta_i) = 0$.

Proof. First, $\deg_x V_U(x)V_\upsilon(x) = (k-1-u)+(M-v)_+ = k-1-\tilde{u}$. As $\tilde{H}(x,y)$ has a degree at most $\tilde{u}$, the only product of x terms which contributes to $[x^{k-1}]f$ are the leading terms of $\deg_x V_U(x)V_\upsilon(x)$ and $\tilde{H}(x,y)$. Then, the coefficient of x^{k-1} in f includes $c_{\tilde{u}}(y)$. As ζ is not on the product grid and not a part of the h parities, $V_V(\zeta)V_\eta(\zeta) \neq 0$, so $c_{\tilde{u}}(\zeta) = 0$. $\square$

The parity design in Lemma 4.4 enforce a zero condition on $c_{\tilde{u}}(y)$ for the $\tilde{u}$ specified for any codeword. Then, the worst-case $m_j(\mathcal{A})$ upper bound is tightened for all surviving rectangular supports.

Theorem 4.5 (Rook-Parity Degree Reduction)**.** *Let $(\tilde{u}, \tilde{v})$ be the residual degree bounds on $\tilde{H}(x, y)$ after h rectangular-support parities, and let r_x and r_y be the number of rook-parities.*

If $r_x \geq \tilde{v} + 1$, then $\deg_x \tilde{H} \leq \tilde{u} - 1$. If $r_y \geq \tilde{u} + 1$, then $\deg_y \tilde{H} \leq \tilde{v} - 1$.

Proof. The r_x rook-parities give r_x distinct roots of $c_{\tilde{u}}(y)$, which satisfies $\deg c_{\tilde{u}}(y) \leq \tilde{v}$. Then, for $r_x \geq \tilde{v} + 1$ enforced zeros, $c_{\tilde{u}}(y)$ must be the zero polynomial. The proof for y-parities follows similarly. □

Lemma 4.6 (Field Size and Rate)**.** *Suppose that $0 \leq M, r_x, r_y \leq k - 1$. Then, the specified off-grid parity checks in Theorem 4.2 and Lemma 4.4 may be chosen if and only if*

$$q - n \geq M + \max(r_x, r_y). \tag{4.7}$$

When condition (4.7) is met, all $t = \frac{M(M+1)}{2} + r_x + r_y$ parity checks are independent, and the subcode has parameters $[n^2, k^2 - t]$ and a rate $R = \frac{k^2 - t}{n^2}$.

Proof. The x coordinate parity checks require at least $M + r_y$ distinct off-grid field elements. The y coordinate parity checks require at least $M + r_x$ distinct off-grid field elements. Then, (4.7) is both a sufficient and necessary condition for our construction.

The rectangular parity checks are linearly independent because the associated Newton basis evaluation matrix is a triangular matrix with a non-zero diagonal. Since each Newton polynomial has both x and y degrees at most $M - 1 \leq k - 2$, they satisfy the rook checks.

For the x-rook parities, suppose we have a monic polynomial $G(x)$ such that $\deg_x G = k - 1$ and G vanishes at $v_0, \dots, v_{M-1}$. Then, for any polynomial $L(y)$ with $\deg_y L \leq k - 2$, the polynomial $f = G(x)L(y)$ satisfies both the rectangular checks and y-rook checks. Then, $L(\zeta_i)$ for $i = 1, \dots, r_x$ is a polynomial satisfying $\deg L \leq r_x - 1 \leq k - 2$, and the x-rook parities can be chosen independently. The same argument applies for the y-rook checks. □

4.3 Worst-Case Bounds and Parity Allocation

Suppose we build a subcode with $t = h + r_x + r_y$ heavy parities, where $h = \frac{M(M+1)}{2}$. Let

$$\tilde{u} = u - (M - v)_+, \quad \tilde{v} = v - (M - u)_+$$

denote the residual degree bounds after applying the rectangular parities. Then, for the rook parities, we reduce the x degree bound by one if $r_x \geq \tilde{v} + 1$, and reduce the y degree bound by one if $r_y \geq \tilde{u} + 1$. If one bound decreases, it will impact the other bound. Let $\tilde{u}'$ and $\tilde{v}'$ denote maximum x and y degrees of the surviving codewords.

Theorem 4.7 (Worst Case Probability of Failure of Subcode)**.** *Let $\mathcal{E} = \mathcal{U} \times \mathcal{V} \setminus \mathcal{A}$ be a surviving codeword support with $|\mathcal{U}| = d + u, |\mathcal{V}| = d + v$, and $m_k(\mathcal{A})$ partial k-matchings on $\mathcal{A}$. Here, we assume that $u + v \geq M$. For a biregular sampler $\pi_{\mathcal{B},n}$ with $Q = n$ samples,*

$$p_{\pi_{\mathcal{B},n}}(\mathcal{E}) \leq \sum_k m_k(\mathcal{A}) \frac{(n-d-v)_{d+u-k}}{(n)_{d+u}} \tag{4.8}$$

for falling factorial $(x)_m$. Let $h_{max} = \min((d+u)\tilde{v}', (d+v)\tilde{u}')$. Then, for $1 \leq i \leq \min(d+u, d+v)$, we upper bound $m_k(\mathcal{A})$ as

$$m_k(\mathcal{A}) \leq \min\left(\binom{d+u}{k}(\frac{h_{max}}{d+u})^k, \binom{d+v}{k}(\frac{h_{max}}{d+v})^k\right) \tag{4.9}$$

$$m_k(\mathcal{A}) \leq \min\left(\binom{d+u}{k}\frac{(d+v)!^{\frac{d+v-k}{d+v}}}{(d+v-k)!}(\tilde{v}!')^{\frac{k}{\tilde{v}'}}, \binom{d+v}{k}\frac{(d+u)!^{\frac{d+u-k}{d+u}}}{(d+u-k)!}(\tilde{u}!')^{\frac{k}{\tilde{u}'}}\right), \tag{4.10}$$

where $\tilde{u}'\tilde{v}' > 0$

Proof. For row degrees $r_1, \dots, r_{d+u}$, Maclaurin's inequality gives us

$$m_k(\mathcal{A}) \leq \binom{d+u}{k}(\frac{|\mathcal{A}|}{d+u})^k \leq \binom{d+u}{k}(\frac{h_{\max}}{d+u})^k.$$

Then, the second bound follows similarly from Lemma 3.12, in which each row has a maximum degree $\tilde{v}'$ and each column has maximum degree $\tilde{u}'$. $\square$

Then, adding $r_x \geq \tilde{v}+1$ row rook-parities enforces a maximum of $\tilde{u}' = \tilde{u}-1$ roots in x for each rectangular support. Finally, we bound the probability of failure for our construction for $t = h + r_x + r_y$ heavy parities as follows.

Let $m_{k,\max}(u,v) = \min(\binom{d+v}{k}(\tilde{u}!)^{\frac{k}{\tilde{u}}}, \binom{d+u}{k}(\tilde{v}!)^{\frac{k}{\tilde{v}}})$ denote the worst-case partial rook number. Then, the worst-case probability of failure is

$$p_{\pi_{B,n}}(\mathcal{E}) \leq \min\left(1, \max_{u,v}\left(\sum_k m_k(u,v)\frac{(n-d-v)_{d+u-k}}{(n)_{d+u}}\right)\right)$$

over the surviving erasure support sizes and internal polynomial degrees.

For a budget of $t = h + r_x + r_y$ heavy parities, we design a sampler-aware subcode by optimizing over selection of (h, r_x, r_y).

4.3.1 Example: Parity Budget Allocation

First, suppose we start with a $[50^2, 36^2]$ 2D-RS code $\mathcal{C}_0$. Then, suppose we have a budget of t heavy parities. For $h = \frac{M(M+1)}{2}$ rectangular support parities, the smallest surviving rectangular supports have $u + v = M$, and have no internal unerased points. The code has distance $d(\mathcal{C}) \geq d(d+M)$. Suppose we allocate $t = 36$ heavy parities. By Lemma 4.6, each subcode has parameters $[2500, 1260]$. We compare our subcode design to a 2D-RS code with a $[50, 35]$ and $[50, 36]$ row and column RS code. Any finite field with $q \geq 68$ is sufficient for independent parity checks. For example, we use $\mathbb{F}_{128}$.

For each subcode, we sample $Q = 50$ samples using the biregular sampler. Then, the probability of failure of multiple parity allocations is shown in Table 4.1, where the bounds were evaluated for $0 \leq u, v \leq 35$.

The two parity families both provide complementary control over surviving codewords. In Table 4.1, our parity designs improve worst-case probability of failure against a plain 2D RS code with the same rate. The probability of failure for the base code is exact (minimum distance erasure pattern), while our subcode probability of failure assignments are a worst-

Table 4.1: Worst-Case probability of failure comparison for $[2500, 1260]$ subcode designs against a $[50^2, 36(35)]$ 2D RS code.

Design	(h, r_x, r_y)	Worst-Case Bound
$[50^2, 35(36)]$ 2D RS Code	(0,0,0)	8.25×10^{-4}
Rectangular Only	(36,0,0)	2.17×10^{-4}
Matching Only	(0,18,18)	4.48×10^{-4}
Mixed	(15,10,11)	1.38×10^{-4}

case upper bound on probability of failure.

Both rectangular and rook parities place restrictions codeword supports and the internal zero patterns of a codeword's support profile. Rectangular parity impose constraints on the support size by requiring $u + v \geq M$. Rook checks reduce the residual degree of all surviving erasure patterns, which lowers $m_j(\mathcal{A})$ of all surviving patterns. Neither parity check type alone proves more effective. A mixed parity check design provides the smallest worst-case upper bound on probability of failure.

Our analysis does not directly establish that the extremal patterns permitted by the bounds occur as a surviving codeword support. Rather, we constrain the probability of failure of a subcode by constraining the surviving erasure patterns to have a more sampler-favorable structure. Fully determining which of the extremal erasure patterns remain ML-undecodable is a natural direction for sharpening our subcode design guarantees.

CHAPTER 5

Conclusion

This thesis reframes both code and sampler design for DAS systems using 2D Reed-Solomon codes. The rectangular-like structure of the code's undecodable erasure patterns create opportunities for targeted sampler design. We design a *biregular sampler*, where we analyze probability of failure performance in both the *sparse-sample* and *asymptotic* regime. Then, fixing our sampler, we build a *sampler-aware* subcode to minimize worst-case probability of failure performance guarantees. Our biregular sampler is optimal against minimum distance erasure patterns in the sparse-sample regime. For a $[25^2, 20^2]$ code, the worst-case probability of failure for $Q = 25$ samples is 1.4 times lower than uniform random sampling. We prove an asymptotic crossover point in which minimum distance erasure patterns do not dominate probability of failure performance, and provide worst-case upper bounds on probability of failure for all $\mathcal{E} \in \mathcal{E}_{\text{inc}}$.

In the asymptotic sampling regime, in which each light node samples a fixed fraction of the block, we analyze the exponential decay rate of probability of failure compared to a uniform random sampler and a *row-regular* sampler. We found an improvement in the first-order exponent of probability of failure by a factor of roughly $1.3 - 1.5$ which raises as the row and column code distances increase. We also provide an explicit region, $3\delta + 2\rho \leq 1$, in which minimum distance erasure patterns are the worst-case erasure patterns for probability of failure. Our numerical simulations show that there is an asymptotic crossover near $\delta^* \approx 0.417$,matching the finite-length crossover of the sparse-sample regime.

Then, fixing our biregular sampler with $Q = n$ queries, we designed a sampler-aware subcode whose parities reshape all surviving erasure patterns to be favorable for the biregular sampler. For a $[50^2, 36^2]$ base code with $t = 36$ parities, our subcode lowers the worst-case probability of failure bound from an exact 8.25×10^{-4} for a base code of equivalent rate to 1.38×10^{-4}. Our mixed parity allocation, using 15 rectangular parities and 21 rook parities, has the smallest worst-case probability of failure bound by a factor of approximately 6. The existence of a construction with lower code distance but more favorable probability of performance shows opportunities for joint code-sampler design for blockchain systems.

Several problems remain open. First, our finite-regime probability of failure performance for both the biregular sampler and subcode design relies on an upper bound of $m_j(\mathcal{A})$, where $m_j(\mathcal{A})$ is the number of ways to place j points in distinct rows and columns of $\mathcal{A}$. We also do not have an exact characterization of which codeword support structures are ML-decodable.

For our asymptotic sampling regime, our region Ω is a sufficient condition for minimum distance erasure patterns having the largest probability of failure compared to any other $\mathcal{E} \in \mathcal{E}_{\text{inc}}$. In Theorem 3.15, we prove a smaller, conservative region in which $3\delta + 2\rho \leq 1$ for our conditions to hold. However, numerical simulations imply the true distance threshold δ^* approaches the second-weight crossover $\delta_2^* \approx 0.417$ of Theorem 3.11. We conjecture that the true δ crossover where $\rho \to 0$ is around 0.417. An open problem is whether the proven region $3\delta + 2\rho \leq 1$ can be further improved towards the boundary by more careful treatment of Ω.

Finally, the cost of exact biregular sampling grows with the blocklength and fraction of the block sampled. There is a required Markov-chain switch mixing time of $O((an)^2 \log(an))$ per sampling round [Gre21]. Whether an approximately biregular sampler retains most of the first-order exponent gain of the exact sampler, and at what blocklengths the switching ceases to be viable for a light node, is open.

APPENDIX A

Proof for Sparse Sampling Regime

A.1 Sparse-Sample Asymptotics

This appendix proves the asymptotic statement of Theorem 3.11.

Proof. First, for $n_1 = n_2 = n$, $d_1 = d_2 = d$, and exactly $Q = n$ queries using the biregular sampler, the ratio in probability of failure between the second-weight and minimum distance erasure patterns is

$$\frac{p_{\pi_{B,n}}(E_1)}{p_{\pi_{B,n}}(E)} = \frac{1}{(n-d)^2} \sum_{k=0}^{d+1} \binom{d+1}{k} \frac{(n-2d)!}{(n-2d-2+k)!}. \tag{A.1}$$

The minimum distance erasure patterns have a larger probability of failure when $\frac{p_{\pi_{B,n}}(\mathcal{E}_1)}{p_{\pi_{B,n}}(\mathcal{E})}$ is less than or equal to 1. We find the asymptotic cross-over point where the second weight erasure patterns have higher probability of failure than the minimum distance erasure patterns using the biregular sampler. We upper bound each term in the summation. When $k > 2$, we may exactly express the k^{th} term in the sum as

$$\frac{1}{(n-d)^2} \binom{d+1}{k} \frac{1}{\prod_{j=1}^{k-2}(n-2d+j)}.$$

Then, we may upper bound $\binom{d+1}{k} \leq \frac{(d+1)^k}{k!}$. Next, we may upper bound the product term using

$$\frac{1}{\prod_{j=1}^{k-2}(n-2d+j)} \leq \frac{1}{\prod_{j=1}^{k-2}(n-2d)} = (n-2d)^{2-k}, \tag{A.2}$$

by lower bounding each term in the product. We upper bound each summation term for $k > 2$ as

$$\frac{(n-2d)^2}{(n-d)^2}\left(\frac{d+1}{n-2d}\right)^k \frac{1}{k!}. \tag{A.3}$$

When $k = 0$, the summation is equal to $\frac{(n-2d)(n-2d+1)}{(n-d)^2} \leq \frac{(n-2d)^2}{(n-d)^2}$. When $k = 1$, the summation is equal to $\frac{(d+1)(n-2d)}{(n-d)^2} = \frac{(n-2d)^2}{(n-d)^2}\frac{(d+1)}{(n-2d)}$. When $k = 2$, the summation is equal to $\frac{(d+1)d}{2(n-d)^2} = \frac{(n-2d)^2}{(n-d)^2}\frac{(d+1)d}{2(n-2d)^2)}$.

Then, we find that the ratio is upper bounded as

$$\frac{p_{\pi_{B,n}}(E_1)}{p_{\pi_{B,n}}(E)} \leq \left(\frac{(n-2d)}{(n-d)}\right)^2 \sum_{k=0}^{\infty}\left(\frac{d+1}{n-2d}\right)^k \frac{1}{k!}, \tag{A.4}$$

where we further upper bound the summation using ∞ in the summation. The above is an exponential Taylor series multiplied by a fixed constant. Now, in the asymptotic regime, suppose $\frac{d}{n} \to \delta$, for some $\delta < \frac{1}{2}$. Here, we specify $\delta < \frac{1}{2}$ as $n - 2d > 0$ must hold in the summation. This can be viewed conceptually as follows: $Q = n$ samples will detect all minimum distance erasure patterns, and so the ration will be undefined if $\frac{d}{n} > \frac{1}{2}$. Then, each term in the summation converges to

$$\frac{(n-2d)^2}{(n-d)^2}\left(\frac{d+1}{n-2d}\right)^k \frac{1}{k!} \to (\frac{1-2\delta}{1-\delta})^2(\frac{\delta}{1-2\delta})^k\frac{1}{k!}. \tag{A.5}$$

Then, we find that the ratio has an asymptotic upper bound

$$\frac{p_{\pi_{B,n}}(E_1)}{p_{\pi_{B,n}}(E)} \to \left(\frac{(1-2\delta)}{(1-\delta)}\right)^2 \exp\left(\frac{\delta}{1-2\delta}\right). \tag{A.6}$$

To find the crossover point, we see what value of δ has the ratio equal to 1. We find that

$$1 = \left(\frac{(1-2\delta)}{(1-\delta)}\right)^2 \exp\left(\frac{\delta}{1-2\delta}\right) \implies 2\ln(\frac{1-\delta}{1-2\delta}) = \frac{\delta}{1-2\delta}.$$

Then, the equality holds at $\delta^* \approx 0.417$.

$\square$

APPENDIX B

Proofs for Asymptotic Linear Sampling Regime

B.1 Problem Setup

For this proof, we assume a square, symmetric $[n^2, k^2]$ 2D Reed-Solomon code and an a-regular sampler, $\pi_{\mathcal{B},Q}$ with $Q = an = \Theta(n)$. Our proof is in asymptotic region

$$\frac{d}{n} \to \delta, \qquad \frac{a}{n} \to \rho \tag{B.1}$$

for initial assumptions $\delta \in (0, \frac{1}{2})$ and $\rho \in (0, \delta)$. For an inclusion-minimal erasure pattern $\mathcal{E} \in \mathcal{E}_{\text{inc}}$, let $\mathcal{X} = \mathcal{U} \times \mathcal{V}$ denote the rectangular support, with $|\mathcal{U}| = d + u$ and $|\mathcal{V}| = d + v$. Both u and v are non-negative integers such that $u, v \leq n - d$. Let $r = \max(u, v)$, and $\mathcal{A} = \mathcal{X} \setminus \mathcal{E}$. The set of allowed positions for the sampler to sample during a failure event is $\mathcal{W} = K_{n,n} \setminus \mathcal{E}$.

Let $\mathcal{E}_0 \in \mathcal{E}_{\text{min}}$ denote a minimum distance erasure pattern, with rectangular support $\mathcal{X}_0 = \mathcal{U}_0 \times \mathcal{V}_0$ and allowed positions $\mathcal{W}_0 = K_{n,n} \setminus \mathcal{E}_0$.

Let $x = \delta + \frac{r}{n}$. Then, for $\delta \le x < 1$, we define function

$$\begin{aligned}\Phi_{\delta,\rho}(x) = \inf_{\alpha,\beta\in\mathbb{R}} \{&x(x-\delta)\log(1+e^{\alpha}) + 2x(1-x)\log(1+e^{\alpha+\beta}) \\ &+ (1-x)^2\log(1+e^{2\beta}) - 2\rho(x\alpha + (1-x)\beta)\}.\end{aligned} \tag{B.2}$$

We now prove the following component of the theorem.

Theorem B.1 (Entropy Gap)**.** *Define parameter region*

$$\begin{aligned}\Omega = \Bigg\{ (\delta,\rho) : 0 < \delta < \frac{1}{2}, \quad 0 < \rho < \frac{(1-\delta)}{2} \\ F(c,l) < 0, G(c,l) < 0 \ \forall\ 0 < c \le \frac{1-\delta}{\delta}, 0 < l \le \frac{\rho}{1-\delta} \Bigg\}.\end{aligned} \tag{B.3}$$

For a compact subset $K \subset \Omega$, suppose there exists a positive constant β_0 such that

$$\Phi_{\delta,\rho}(x) \le \Phi_{\delta,\rho}(\delta) - \beta_0(x-\delta) \tag{B.4}$$

for all $(\delta,\rho) \in K$ and all $x \in (\delta, 1)$. Then, there exists a positive constant β_1 and n_0 such that

$$\log\left(\frac{p_{\pi_{B,Q}}(\mathcal{E})}{p_{\pi_{B,Q}}(\mathcal{E}_0)}\right) \le -\beta_1 nr \tag{B.5}$$

for all $n \ge n_0$ and every $r \ge 1$.

As $\delta \to \delta_2^* \approx 0.417$, the region in which minimum distance erasure patterns define the probability of failure performance begins to close. We note that this numerically evaluated threshold is very similar to the one found for the sparse-sample regime, δ_2^*. Both the sparse-sample regime and the asymptotic regime approach the same distance threshold. However, our proof does not fully extend to $\delta_2^* \approx 0.417$, and we only see this approximation numerically.

B.1.1 Maximum Entropy Preliminaries

In this section, we establish the necessary notation and theorems used from [Bar10]. For an inclusion-minimal erasure pattern $\mathcal{E}$, let $\Sigma(R, C; \mathcal{W})$ for $R = (r_1, \ldots, r_m)$ and $C = (c_1, \ldots, c_n)$

denote the number of $m \times n$ bipartite graphs with row sums R, column sums C, and allowed positions $\mathcal{W}$. Here, we require $\sum_i r_i = \sum_j c_j$. Our quantity of interest is $\Sigma(\mathbf{a}, \mathbf{a}; \mathcal{W})$, where $\mathbf{a} = (a, \dots, a)$ is of length n and imposes that all row and column sums must equal a.

The *feasible polytope* $\mathcal{P}(R, C; \mathcal{W})$ [Bar10] is the set of fractional edges $X = (x)_{ij}$ such that

$$\mathcal{P}(R, C; W) = \Big\{ X = (x)_{ij} \in [n] \times [n] : z_{ij} = 0 \;\; \forall (i,j) \notin \mathcal{W}, \quad 0 \le z_{ij} \le 1, \\ \sum_j z_{ij} = r_i \;\forall i \in [n], \quad \sum_i z_{ij} = c_j \;\forall j \in [n] \Big\}. \tag{B.6}$$

Then, for natural entropy function $h(z) = -z \ln(z) - (1-z)\ln(1-z)$ for $z \in [0,1]$, the maximum entropy matrix [Bar10] is defined as

$$H(Z_{\mathcal{W}}) = \max_{Z \in \mathcal{P}(R,C;\mathcal{W})} \sum_{(i,j) \in \mathcal{W}} h(z_{ij}^{\mathcal{W}}). \tag{B.7}$$

There are three important quantities of interest, which we state here. The count of graphs in $\Sigma(R, C; \mathcal{W})$ may be bounded as [Bar10, Theorem 5]

$$e^{H(Z_{\mathcal{W}})} \ge |\Sigma(R, C; \mathcal{W})| \ge \frac{(mn)!}{(mn)^{mn}} \left(\prod_{i=1}^{m} \frac{(n - r_i)^{n - r_i}}{(n - r_i)!} \right) \left(\prod_{j=1}^{n} \frac{c_j^{c_j}}{c_j!} \right) e^{H(Z_{\mathcal{W}})}. \tag{B.8}$$

Thus, $|\Sigma(R, C; W)|$ is both upper and lower bounded by the maximum entropy. Second, the maximum entropy matrix $Z_{\mathcal{W}}$ has a solution of the form [Bar10, Lemma 6]

$$z_{ij}^{\mathcal{W}} = \frac{e^{\alpha_i + \beta_j}}{1 + e^{\alpha_i + \beta_j}},$$

with found α_i and β_j for each row and column.

Finally, the maximum entropy approach may be viewed using typical-set like arguments. When polytope $\mathcal{P}(R, C; \mathcal{W})$ has a non-empty interior and $Z_{\mathcal{W}}$ is the maximum entropy matrix, suppose we draw random matrix $X \in \{0,1\}^{n \times n}$ such that each x_{ij} is a Bernoulli random variable drawn i.i.d. with probability $z_{ij}^{\mathcal{W}}$. Then, [Bar10, Theorem 8]

$$\mathbb{E}[X] = Z_{\mathcal{W}},$$

and

$$\Pr(X = D \text{ for fixed } D \in \Sigma(R, C; \mathcal{W})) = e^{-H(Z_{\mathcal{W}})} \ \forall \ D \in \Sigma(R, C; \mathcal{W}).$$

Importantly, we may use this to estimate $\Sigma(R, C; \mathcal{W})$ as

$$\Pr(X \in \Sigma(R, C; \mathcal{W})) = |\Sigma(R, C; \mathcal{W})| e^{-H(Z_{\mathcal{W}})}.$$

We use Isaev et. al. [IIM20] to refine our asymptotic estimate for $\Pr(X \in \Sigma(R, C; W))$ when necessary for the proof.

B.1.2 Entropy Upper Bound

Lemma B.2 (Dual Convex Bound)**.** *For an inclusion-minimal erasure pattern $\mathcal{E} \in \mathcal{E}_{inc}$ and a biregular sampler which takes $Q = an$ samples, assume that $|\Sigma(\boldsymbol{a}; \boldsymbol{a}; \mathcal{W})| > 0$. Then,*

$$\log |\Sigma(\boldsymbol{a}, \boldsymbol{a}, \mathcal{W})| \leq H(Z_{\mathcal{W}}) \leq n^2 \Phi_{\delta,\rho}(\frac{d+r}{n}), \tag{B.9}$$

for $|\mathcal{U}| = d + u$, $|\mathcal{V}| = d + v$, and $r = \max(u, v)$. Here, $H(Z_{\mathcal{W}})$ denotes the maximum entropy matrix for a-regular graphs with allowed edges $\mathcal{W}$, and

$$\begin{aligned} \Phi_{\delta,\rho}(x) = \inf_{\alpha,\beta \in \mathbb{R}} \{ & x(x-\delta) \log(1 + e^{\alpha}) + 2x(1-x) \log(1 + e^{\alpha+\beta}) \\ & + (1-x)^2 \log(1 + e^{2\beta}) - 2\rho(x\alpha + (1-x)\beta)\} \end{aligned} \tag{B.10}$$

is defined for $x \in [\delta, 1)$.

Proof. From Equation B.8 [Bar10, Theorem 5], we know that $\log(|\Sigma(\mathbf{a}, \mathbf{a}; \mathcal{W})|) \leq H(Z_{\mathcal{W}})$, for $H(Z_{\mathcal{W}})$ found using Equation B.7.

We design an upper bound for $H(Z_{\mathcal{W}})$ by finding the dual program. For objective function $H(Z_{\mathcal{W}})$ and feasibility constraints $\mathcal{P}(\mathbf{a}, \mathbf{a}; \mathcal{W})$, the Lagrange multiplier is

$$\begin{aligned} \mathcal{L}(Z, \theta_1, \theta_2) &= \sum_{(i,j) \in \mathcal{W}} h(z_{ij}) + \sum_{i=1}^{n} \theta_{1i} \left(\sum_{j=1}^{n} z_{ij} - a \right) + \sum_{j=1}^{n} \theta_{2j} \left(\sum_{i=1}^{n} z_{ij} - a \right) \\ &= \sum_{(i,j) \in \mathcal{W}} \big(h(z_{ij}) + (\theta_{1i} + \theta_{2j}) z_{ij} \big) - a \sum_{i=1}^{n} \theta_{1i} - a \sum_{j=1}^{n} \theta_{2j}, \end{aligned} \tag{B.11}$$

with dual variables θ_{1i} for row-slackness and θ_{2j} for column-slackness. We may maximize separately for each z_{ij}. The conjugate of the entropy function [Bar10] is

$$h^*(y) = \sup_{0 \leq z \leq 1} \{h(z) + yz\} = \log(1 + e^y) \tag{B.12}$$

for $z^* = \frac{e^y}{1+e^y}$. As in [Bar10], we set $y = \theta_{1i} + \theta_{2j}$, such that each $z^*_{ij} = \frac{e^{\theta_{1i}+\theta_{2j}}}{1+e^{\theta_{1i}+\theta_{2j}}}$. Then,

$$g(\theta_1, \theta_2) = \inf_{\theta_1, \theta_2} \left(\sum_{(i,j) \in \mathcal{W}} \log\left(1 + e^{\theta_{1i}+\theta_{2j}}\right) - a \sum_{i=1}^{n} \theta_{1i} - a \sum_{j=1}^{n} \theta_{2j} \right). \tag{B.13}$$

We upper bound $H(Z_{\mathcal{W}})$ as follows. Let $\mathcal{X}' = \mathcal{U}' \times \mathcal{V}'$ denote the *enlarged rectangular support* where both $|\mathcal{U}| = |\mathcal{V}| = d + r$. $\mathcal{X}'$ encapsulates the $(d + r) \times (d + r)$ square grid of points in which $\mathcal{E}$ is fully contained.

We define one feasible point in the dual. Let $\theta_{1i} = \alpha$ if $i \in \mathcal{U}'$ and $\theta_{1i} = \beta$ if $i \notin \mathcal{U}'$ for constants α and β. Similarly, let $\theta_{2j} = \alpha$ if $j \in \mathcal{V}'$ and $\theta_{2j} = \beta$ if $j \notin \mathcal{V}'$. We do not find the optimal solution to the dual; the upper bound assigns each row or column as "in" the support with value α or "out" of the support with value β.

For region $\mathcal{U}' \times \mathcal{V}'^c$, there are $(d + r)(n - d - r) = n^2(x)(1 - x)$ total positions. Then, $\theta_{1i} = \alpha$ and $\theta_{2j} = \beta$ for all $(i, j) \in \mathcal{U}' \times \mathcal{V}'^c$, and $h(z^{\mathcal{W}}_{ij}) = \log(1+e^{\alpha+\beta})$ for all $(i, j) \in \mathcal{U}' \times \mathcal{V}'^c$ using Equation B.12. For region $\mathcal{U}'^c \times \mathcal{V}'$, there are $(d + r)(n - d - r) = n^2(x)(1 - x)$ total positions, and $h(z^{\mathcal{W}}_{ij}) = \log(1 + e^{\alpha+\beta})$ for all $(i, j) \in \mathcal{U}'^c \times \mathcal{V}'$. For region $\mathcal{U}' \times \mathcal{V}'$, there are a *maximum* of $(d + r)^2 - (d + r)d = (d + r)r = n^2(x)(x - \delta)$ available positions, each with entropy $h(z^{\mathcal{W}}_{ij}) = \log(1+e^{2\alpha})$. For region $\mathcal{U}'^c \times \mathcal{V}'^c$, there are exactly $(n-d-r)^2 = n^2(1-x)^2$ positions, each with entropy $h(z^{\mathcal{W}}_{ij}) = \log(1 + e^{2\beta})$.

The term $a \sum_i \theta_1 = a((d + r)\theta_1 + (n - d - r)\theta_2) = n^2\rho(x\theta_1 + (1 - x)\theta_2)$. Additionally, we find that $a \sum_j \theta_{2j} = a((d + r)\theta_1 + (n - d - r)\theta_2) = n^2\rho(x\alpha + (1 - x)\beta)$.

Our upper bound, which results from forcing each θ_{1i} or θ_{2j} to be equal to α or β depending on its position in the rectangular support, directly follows by minimizing over constants α and β after substituting in each term. $\square$

Our convex upper bound as defined in Lemma B.2 is the optimal solution when $x = \delta$, which occurs for the set of minimum distance erasure patterns $\mathcal{E}_{\min}$.

Lemma B.3 (Minimum Distance Upper Bound)**.** *For $\Phi_{\delta,\rho}(x)$ defined in Lemma B.2,*

$$n^2\Phi_{\delta,\rho}(\delta) = H(Z_{\mathcal{W}_0}) \tag{B.14}$$

for $\mathcal{E}_0 \in \mathcal{E}_{min}$. The maximum entropy upper bound $\Phi_{\delta,\rho}(\delta)$ is met with equality for minimum distance erasure patterns.

Proof. Using Theorem 3.13, the symmetry properties of $\mathcal{E}_{\min}$ force block entropy values

$$\begin{aligned}\rho_A &= 0\\ \rho_B &= \rho_C = \frac{\rho}{1-\delta}\\ \rho_D &= \frac{\rho(1-2\delta)}{(1-\delta)^2},\end{aligned}$$

for ρ_A in region $\mathcal{U}_0 \times \mathcal{V}_0$, D in region $\mathcal{U}_0^c \times \mathcal{V}_0^c$, and ρ_B in the remaining region of the matrix. Then, for any $(\rho, \delta) \in \Omega$, as defined in Theorem B.1, $0 < \rho_B, \rho_C, \rho_D < 1$ are strictly feasible. Feasibility is proven using the arguments in [Bar10, Proof of Lemma 6]. At $x = \delta$, the upper bound in Lemma B.2 reduces to

$$\begin{aligned}\Phi_{\delta,\rho}(\delta) = \inf_{\theta_1,\theta_2\in\mathbb{R}} &\{2\delta(1-\delta)\log(1+e^{\theta_1+\theta_2})\\ &+ (1-\delta)^2\log(1+e^{2\theta_2}) - 2\rho(\delta\theta_1 + (1-\delta)\theta_2)\}.\end{aligned} \tag{B.15}$$

Then, the convex dual $\Phi_{\delta,\rho}(\delta)$ follows the exact block structure of Theorem 3.13 for $n_1 = n_2 = n$ and $d_1 = d_2 = d$, and $\Phi_{\delta,\rho}(\delta)$ is the exact dual program. The solution to the dual $\Phi_{\delta,\rho}(\delta)$ will therefore give the optimal solution to the primal. □

Lemma B.4 (Counting Arguments)**.** *There exists a constant $C_0 < \infty$ such that*

$$\log|\Sigma(\boldsymbol{a}, \boldsymbol{a}; \mathcal{W}_0)| \geq H(Z_{\mathcal{W}_0}) - C_0 n \log n. \tag{B.16}$$

Proof. From [Bar12], we may lower bound non-zero $|\Sigma(\mathbf{a}, \mathbf{a}, \mathcal{W})|$ as

$$|\Sigma(\mathbf{a}, \mathbf{a}; \mathcal{W}_0)| \geq (n^2)^{\gamma_0(2n)} e^{H(Z_{\mathcal{W}_0})}$$

for a fixed positive constant γ_0. Then, taking the logarithm on both sides, $\log((n^2)^{\gamma_0 2n}) = C_0 n \log n$ for positive constant $C_0 = 4\gamma_0$. □

B.1.3 Entropy Upper Bound: Feasibility and Derivative

Lemma B.5 (Entropy Feasibility)**.** *For $\mathcal{E} \in \mathcal{E}_{inc}$ with $\delta < x < 1$, the equivalent primal program of $\Phi_{\delta,\rho}(x)$ in Lemma B.2 has a unique interior maximizer $Z^* = (p, l, t)$. The dual infimum occurs for finite multipliers θ_1 and θ_2 using*

$$p = \frac{e^{2\theta_1}}{1 + e^{2\theta_1}}, \quad l = \frac{e^{\theta_1+\theta_2}}{1 + e^{\theta_1+\theta_2}}, \quad t = \frac{e^{2\theta_2}}{1 + e^{2\theta_2}}. \tag{B.17}$$

The KKT conditions require

$$(x - \delta)p + (1 - x)l = \rho \tag{B.18}$$

$$x(l) + (1 - x)t = \rho \tag{B.19}$$

$$pt(1 - l)^2 = l^2(1 - p)(1 - t). \tag{B.20}$$

Proof. We consider a primal, three-variable entropy maximization problem. We specify three entropy values: p for edges in $\mathcal{U}' \times \mathcal{V}'$, l for edges in $\mathcal{U}' \times \mathcal{V}'^c$ or $\mathcal{U}'^c \times \mathcal{V}$, and t for edges in $\mathcal{U}'^c \times \mathcal{V}'^c$. The normalized entropy maximization problem is then

$$\begin{aligned} \text{maximize } J(p, l, t) :=\ & x(x - \delta)h(p) + 2x(1 - x)h(l) + (1 - x)^2 h(t) \\ \text{subject to } & (x - \delta)p + (1 - x)l = \rho \\ & xl + (1 - x)t = \rho \\ & 0 \leq p, l, t \leq 1, \end{aligned} \tag{B.21}$$

where the constraints are the row and column degree constraints. In region $\delta < x < 1$, each entropy term in $J(p, l, t)$ is multiplied by a positive constant. Suppose the polytope formed by the constraints in B.21 has a non-empty interior. Then, strict concavity of the objective function implies that there exists a unique, interior maximizer (p^*, l^*, t^*) (Lemma 6 in [Bar10]).

For primal solution (p, l, t) and a fixed constant $x \in (\delta, 1)$, the Lagrange multiplier of B.21 is

$$\begin{aligned}\mathcal{L}(p, l, t, \mu_1, \mu_2) = J(p, l, t) + \mu_1((x-\delta)p + (1-x)l - \rho) \\ + \mu_2(xl + (1-x)t - \rho)\end{aligned} \tag{B.22}$$

for dual variables μ_1 and μ_2. As the primal has a strictly interior point, Slater's condition implies that strong duality holds [BV23], and the dual minimum is achieved using finite multipliers. Let $\mu_1 = 2x\theta_1$ and $\mu_2 = 2(1-x)\theta_2$. The dual program is then

$$\begin{aligned}\inf_{\theta_1, \theta_2 \in \mathbb{R}} \{x(x-\delta)\log(1+e^{2\theta_1}) + 2x(1-x)\log(1+e^{\theta_1+\theta_2}) \\ + (1-x)^2 \log(1+e^{2\theta_2}) - 2\rho(x\theta_1 + (1-x)\theta_2)\},\end{aligned} \tag{B.23}$$

which is equivalent to the dual program in Lemma B.2. The optimal primal solutions are then those specified in Equation B.17 [Bar10]. Then, from the primal constraints, the KKT conditions [BV23] enforce

$$\begin{aligned}(x-\delta)p + (1-x)l = \rho \\ xl + (1-x)t = \rho.\end{aligned}$$

Conditions on the Lagrangian gradient [BV23] enforce

$$\frac{\partial L}{\partial p} = 0 \implies h'(p) = -2\theta_1 \tag{B.24}$$

$$\frac{\partial L}{\partial l} = 0 \implies h'(l) = -\theta_1 - \theta_2 \tag{B.25}$$

$$\frac{\partial L}{\partial t} = 0 \implies h'(t) = -2\theta_2. \tag{B.26}$$

As $h'(z) = \log(\frac{1-z}{z})$, we find the entropy maximizers in Equation B.17 by solving for p, l, t. The gradient conditions imply that

$$h'(p) + h'(t) = 2h'(l).$$

Then, using $h'(z) = \log \frac{1-z}{z}$, we find that

$$\log(\frac{1-p}{p}) + \log(\frac{1-t}{t}) = 2\log(\frac{1-l}{l}) \implies pt(1-l)^2 = l^2(1-p)(1-t). \tag{B.27}$$

$\square$

Lemma B.6 (Strict Feasibility). *For $\delta < x < 1$, let*

$$L_x = \max\left(\frac{\rho-(x-\delta)}{1-x}, \frac{\rho-(1-x)}{x}\right), \tag{B.28}$$

and

$$U_x = \min(1, \frac{\rho}{1-x}, \frac{\rho}{x}). \tag{B.29}$$

Then, the three-block entropy maximization problem in Equation B.21 admits a strictly feasible, interior solution if and only if $L_x < U_x$.

Proof. First, for solution (p, l, t) to the optimization problem Equation B.21, we treat t as a free variable. Then, Equations B.18 and B.19 give us that

$$p = \frac{\rho-(1-x)l}{x-\delta}$$
$$t = \frac{\rho-xl}{1-x}.$$

There exists a strictly feasible interior solution if there exists a feasible (p, l, t) such that $0 < p, l, t < 1$. If we impose $0 < p < 1$, then

$$0 < \frac{\rho-(1-x)l}{x-\delta} < 1 \implies \frac{\rho-(x-\delta)}{1-x} < l < \frac{\rho}{1-x}.$$

For $0 < t < 1$, the constraints impose

$$0 < \frac{\rho-xl}{1-x} < 1 \implies \frac{\rho-(1-x)}{x} < l < \frac{\rho}{x}.$$

Finally, we require $0 < l < 1$, such that $0 < \min(\frac{\rho-1-x}{x}, \frac{\rho-\delta-x}{1-x})$ and $1 > \max(\frac{\rho}{1-x}, \frac{\rho}{x})$. Let $L_x = \max(\frac{\rho-1-x}{x}, \frac{\rho-\delta-x}{1-x})$ and $U_x = \min(1, \frac{\rho}{1-x}, \frac{\rho}{x})$. Then, for $1 > x > \delta$, $L_x > 0$. If $L_x < U_x$, then there exists an interior feasible (p, l, t) pair. $\square$

Corollary B.7. *For the primal program in Lemma B.5, there exists a strictly feasible, unique optimizer p, l, t such that*

$$0 < t < l < p < 1.$$

The corollary follows directly from Lemma B.6 and the constraints in Lemma B.5. We now find the derivative of $\Phi_{\delta,\rho}(x)$ for fixed δ and ρ.

Lemma B.8 (Derivative of Upper Bound)**.** *For a fixed triple (x, δ, ρ), let (p, l, t) denote the associated unique primal solution to dual optimizer (θ_1^*, θ_2^*). Then, define*

$$\begin{aligned}\Psi_{\delta,\rho}(x, \theta_1, \theta_2) = -x(x-\delta)\log(1-p) - 2x(1-x)\log(1-l) - (1-x)^2\log(1-t) \\ -2\rho(x\theta_1 + (1-x)\theta_2).\end{aligned} \tag{B.30}$$

Then, at the associated optimum (p, l, t),

$$\Phi'_{\delta,\rho}(x) = \frac{\partial \Psi_{\delta,\rho}(x, \theta_1^*, \theta_2^*)}{\partial x}. \tag{B.31}$$

The derivative is then

$$\begin{aligned}\Phi'_{\delta,\rho}(x) = -(2x-\delta)\log(1-p) - 2(1-2\delta)\log(1-l) + 2(1-x)\log(1-t) \\ -\rho\log\left(\frac{p(1-t)}{t(1-p)}\right),\end{aligned} \tag{B.32}$$

Proof. From Lemma B.5, we know that

$$\log(1+e^{2\theta_1}) = -\log(1-p)$$

for $p = \frac{e^{2\theta_1}}{1+e^{2\theta_1}}$. The substituted values for $\log(1+e^{\theta_1+\theta_2})$ and $\log(1+e^{2\theta_2})$ may be derived.

The proof then follows from an application of the envelope theorem to (B.30) [MS02]. The partial derivatives with respect to θ_1 and θ_2 vanish at θ_1^* and θ_2^*.

Then, we may express the derivative as

$$\begin{aligned}\Phi'_{\delta,\rho}(x) = -(2x-\delta)\log(1-p) - 2(1-2\delta)\log(1-l) + 2(1-x)\log(1-t) \\ -\rho\log\left(\frac{p(1-t)}{t(1-p)}\right),\end{aligned}$$

$\square$

where $\rho\log\left(\frac{p(1-t)}{t(1-p)}\right)$ is derived from relation $e^{2(\theta_1-\theta_2)} = \frac{p(1-t)}{t(1-p)}$.

B.1.4 Derivative Boundary Conditions

For (δ, ρ), where $x = \delta + \frac{r}{n}$, a sufficient condition to prove that larger erasure patterns lower probability is to determine whether $\Phi'_{\delta,\rho}(x) < 0$ for all $\delta < x < 1$.

We may view the change in maximum entropy from $\mathcal{E}$ to $\mathcal{E}_0$ as an entropy *increase* from non-zero p and $h(p)$ inside the rectangular support versus an entropy *decrease* from a smaller region outside of the enlarged rectangular support.

In order to determine regions of (δ, ρ) in which $H(Z_{\mathcal{W}_0})$ has larger entropy than any other $\mathcal{E} \in \mathcal{E}_{\min}$, we apply a transformation of variables to measure the relative contrast induced by changes in p, l, t. For a fixed l, which is the entropy probability of regions $\mathcal{U}' \times \mathcal{V}'^c$ and $\mathcal{U}'^c \times \mathcal{V}$, let

$$t = (1 - c)l \tag{B.33}$$

denote the ratio of the probability assignments between $\mathcal{U}'^c \times \mathcal{V}'^c$ and the regions adjacent to the rectangular support. Then, for any fixed pair (c, l), the value of t and p are fully determined.

Definition B.9 (Optimizers after Change of Variables)**.** For $0 < c < 1$ and $0 < l < \frac{1}{2}$, we define

$$p_{c,l} = \frac{l(1 - (1 - c)l)}{(1 - c) + (2c - 1)l} \tag{B.34}$$

$$t_{c,l} = (1 - c)l \tag{B.35}$$

as the $(p_{c,l}, l, t_{c,l})$ solution of the KKT conditions in Lemma B.5. Then, we define two functions

$$
\begin{aligned}
F(c,l) &= -\frac{c}{1+c}\log(1-p_{c,l}) - \frac{2(1-c)}{1+c}\log(1-l) \\
&\quad + \frac{2}{1+c}\log(1-t_{c,l}) - \frac{l}{1+c}\log\left(\frac{p_{c,l}(1-t_{c,l})}{t_{c,l}(1-p_{c,l})}\right), \qquad \text{(B.36)}\\
G(c,l) &= -\left(1+\frac{(1+c)l}{p_{c,l}}\right)\log(1-p_{c,l}) \\
&\quad + 4\log(1-l) - 2\log(1-t_{c,l}) - cl\log\left(\frac{p_{c,l}(1-t_{c,l})}{t_{c,l}(1-p_{c,l})}\right). \qquad \text{(B.37)}
\end{aligned}
$$

Lemma B.10. *Let $\Omega_{c,l} \subset \Omega$ denote the region where $F(c,l) < 0$ and $G(c,l) < 0$ for every $0 < c \leq \frac{\delta}{1-\delta}$ and $0 < l < \frac{\rho}{1-\delta}$.*

Then, any compact subset $K \subset \Omega_{c,l}$ achieves the entropy gap in Theorem B.1.

Proof. From Lemma B.6, $\rho < l < \frac{\rho}{1-\delta}$. Then, by rearranging the conditions in Lemma B.5,

$$\delta = x\left(1 - \frac{(1+c)l}{p_{c,l}}\right) + \frac{cl}{p_{c,l}}.$$

Additionally, $\rho = (1-c)l + xcl$. Then, we prove that $\left(1 - \frac{(1+c)l}{p_{c,l}}\right)$ is a positive quantity. By manipulating the definition of $p_{c,l}$

$$
\begin{aligned}
1 - \frac{(1+c)l}{p_{c,l}} &= 1 - \frac{(1+c)(1-c) + (2c-1)(1+c)l}{1-(1-c)l} \\
&= \frac{c^2(1-2l)}{1-(1-c)l} \\
&> 0,
\end{aligned}
$$

where the expression is positive when $l < \frac{1}{2}$ and $0 < c < 1$. As our region Ω specifies $\frac{\rho}{1-\delta} < \frac{1}{2}$, the coefficient in front of x is always positive for a compact subset K in $\Omega_{c,l}$.

Next, using the fact that $x < 1$, we know that $\delta < \frac{p_{c,l}-l}{p_{c,l}} = \frac{c(1-l)}{1-(1-c)l} < c$.

Then, we substitute δ and ρ in terms of c into B.32 to find

$$\Phi'_{\delta,\rho}(x) = F(c,l) + (x - \frac{c}{1+c})G(c,l).$$

Then, any compact region K of (δ, ρ, c, l) in which both $F(c,l) < 0$ and $G(c,l) < 0$ has $\Phi'_{\delta,\rho}(x) < 0$ for all $x \in (\delta, 1)$, as $\Phi'_{\delta,\rho}(x)$ is negative, and Φ is decreasing in x. Then, there exists some positive constant β_0 such that $\Phi'(x)$ is maximized at $-\beta_0$ for constant $\beta_0 > 0$ in the compact region K. We then find a linear upper bound

$$\Phi_{\delta,\rho}(x) - \Phi_{\delta,\rho}(\delta) \leq -\beta_0(x - \delta)$$

in the specified region K. □

Then, we apply the following counting lemma to complete the proof.

Lemma B.11 (Counting Arguments)**.** *For any fixed constant K_0 and $\mathcal{E} \in \mathcal{E}_{inc}$, suppose $1 \leq r \leq K_0 \log n$ for $r = \max(u, v)$. Then,*

$$\log \frac{|\Sigma(\boldsymbol{a}, \boldsymbol{a}; \mathcal{W}_{\mathcal{E}})|}{|\Sigma(\boldsymbol{a}, \boldsymbol{a}; \mathcal{W}_0)|} = H(Z_{\mathcal{W}_{\mathcal{E}}}) - H(Z_{\mathcal{W}_{\mathcal{E}_0}}) + o(n).$$

Proof. The proof is in Section B.2. □

Then, the proof concludes as follows. First, using Lemma B.4, we may upper bound

$$\frac{|\Sigma(\mathbf{a}, \mathbf{a}; \mathcal{W})|}{|\Sigma(\mathbf{a}, \mathbf{a}; \mathcal{W}_0)|} \leq H(Z_{\mathcal{W}}) - H(Z_{\mathcal{W}_0}) + C_0 \log n.$$

Then, using Lemma B.2,

$$\frac{|\Sigma(\mathbf{a}, \mathbf{a}; \mathcal{W})|}{|\Sigma(\mathbf{a}, \mathbf{a}; \mathcal{W}_0)|} \leq n^2 \Phi_{\delta,\rho}(\delta + \frac{r}{n}) - n^2 \Phi_{\delta,\rho}(\delta) + C_0 n \log n.$$

When $F(c,l) < 0$ and $G(c,l) < 0$, we apply Lemma B.10 to find that

$$\log \left(\frac{|\Sigma(\mathbf{a}, \mathbf{a}; \mathcal{W})|}{|\Sigma(\mathbf{a}, \mathbf{a}; \mathcal{W}_0)|} \right) \leq -\beta_0 n r + C_0 n \log n.$$

When $r > K_0 \log n$ for some positive constant K_0, then the counting estimate in Lemma B.4 is sufficient to prove that minimum distance erasure patterns have larger probability of failure. For the $r \leq K_0 \log n$ case, we apply the sharper estimate from Lemma B.11

$$\log \left(\frac{|\Sigma(\mathbf{a}, \mathbf{a}; \mathcal{W})|}{|\Sigma(\mathbf{a}, \mathbf{a}; \mathcal{W}_0)|} \right) = H(Z_{\mathcal{W}}) - H(Z_{\mathcal{W}_0}) + O(r) + O(1).$$

Then,

$$\frac{|\Sigma(\mathbf{a},\mathbf{a};\mathcal{W})|}{|\Sigma(\mathbf{a},\mathbf{a};\mathcal{W}_0)|} = -\beta_0 nr + O(r) + O(1)$$
$$\leq -\beta_1 nr.$$

B.2 Proof for small r

B.2.1 Graph Orientation Preliminaries

Isaev et. al. view the same counting problem using *oriented graphs* [IIM20]. A vertex v has an *imbalance* $b_v := \mathrm{outdeg}(v) - \mathrm{indeg}(v)$, and the imbalance sequence of a graph G is $\mathbf{b} := (b_1, \dots, b_n)$ for n vertices. Then, probability

$$\lambda_{jk} := \frac{r_j}{r_j + r_k}$$

corresponds to $r_j := e^{\alpha_j}$ and $r_k := e^{-\beta_k}$ in the maximum entropy literature, where $\lambda_{ij} = z_{ij}$ in [Bar10]. Of import to us is Theorem 1 in [IIM20], which provides a much sharper asymptotic estimate of $|\Sigma(\mathbf{a},\mathbf{a};\mathcal{W})|$ when the theorem conditions apply.

In our problem, we have a graph with $2n$ vertices, each of a maximum degree a. For each row vertex i, we define $\mathrm{outdeg}(i) = a$ and $\mathrm{indeg}(i) = d_{\mathcal{W}}(i) - a$, where $d_{\mathcal{W}}(i)$ is the number of available, non-forbidden column positions in row i. The in-degree denotes how many unsampled, unselected, and allowed columns are in row i. For each column vertex j, $\mathrm{indeg}(j) = a$ and $\mathrm{outdeg}(j) = d_{\mathcal{W}}(j) - a$. Then,

$$b_i = a - (d_{\mathcal{W}}(i) - a) = 2a - d_{\mathcal{W}}(i)$$
$$b_j = (d_{\mathcal{W}}(j) - a) - a = d_{\mathcal{W}}(j) - 2a.$$

To use the refined estimate of [IIM20] on our graph with $2n$ vertices, a total host graph degree Δ and an imbalance sequence $\mathbf{b}$ defined as stated above, we must prove that

1. $(2n)^{\frac{1}{3}+\varepsilon} \leq \Delta \leq 2n-1$ for some constant $\varepsilon > 0$.

2. $h(G) \geq \mu\Delta$ for some constant $\mu > 0$.

3. "Balance" equations $\sum_{k:(j,k)\in G} \frac{r_j - r_k}{r_j + r_k} = b_j$ for $1 \leq j \leq n$ such that $\frac{r_j}{r_k} \leq 1 + R$ for $(j,k) \in G$, where $R = R(n)$ satisfies $0 \leq R = O(1)$ and $R^2 \frac{2n}{\Delta} \log(\frac{4n}{\Delta}) = o(\log n)$.

We label the vertices $v = 1, \ldots, 2n$.

B.2.2 Full Proof

Lemma B.12 (Change in Entropy)**.** *For $\Sigma(R, C; \mathcal{W})$ and $\mathcal{W} = K_{n_1,n_2} \setminus \mathcal{E}$ for $\mathcal{E} \in \mathcal{E}_{inc}$, let Z denote the maximum entropy matrix. There exists some constant $\tau \in (0,1)$ such that*

$$\tau \leq z_{ij} \leq 1 - \tau \quad \forall (i,j) \in \mathcal{W}. \tag{B.38}$$

Proof. First, we define an erasure pattern $\mathcal{E}' = \mathcal{U} \times \mathcal{V}$, which occupies the entire rectangular support of $\mathcal{E}$. Therefore, $\mathcal{E} \subseteq \mathcal{E}'$. Let $\mathcal{W}'$ denote the allowed positions of $\mathcal{E}'$ during a failure event. Let (ρ_B, ρ_C, ρ_D) denote the strictly interior probability assignments for each z_{ij} region from Theorem 3.13 for a rectangular erasure pattern $\mathcal{E}'$.

For $\mathcal{E}'$, suppose we assign densities l_i for any $i \in \mathcal{U}$ for region $\mathcal{U} \times \mathcal{V}^c$. Let v_j for $j \in \mathcal{V}$ denote the density in block $\mathcal{U}^c \times \mathcal{V}$. Then, let t denote the remaining probability assignment in block $\mathcal{U}^c \times \mathcal{V}^c$. From row and column degree constraints,

$$a = (n-d-v)l_i + \sum_{j:(i,j)\in\mathcal{A}} z_{ij}. \tag{B.39}$$

Then, $l_i = \frac{a - \sum_{j:(i,j)\in\mathcal{A}} z_{ij}}{n-d-v}$. There are a maximum of $r = o(n)$ possible summation terms each between 0 and 1. Then,

$$\frac{a-r}{n-d-v} \leq l_i \leq \frac{a}{n-d-v} \implies \rho_B - \frac{r}{n-d-v} \leq l_i \leq \rho_B.$$

Next, $\frac{r}{n-d-v} = o(1)$ as $r = o(n)$ and $(n-d-v) = \Theta(n)$. The same argument may be applied for each v_j. For t, we find

$$\rho_D \leq \frac{a-(d+u)-(d+v)+r}{(n-d-u)(n-d-v)} \leq \rho_D + o(1).$$

Then, there exists some constant c_1 such that $c_1 \leq l_i, v_j, t \leq 1-c_1$. Additionally, maximum entropy interior solutions have form [Bar10, IIM20]

$$\frac{z_{ij}}{1-z_{ij}} = e^{x_i - y_j}.$$

suppose we take a position $(i,j) \in \mathcal{A}$. Then, the symbol probability may be described using

$$\frac{z_{ij}}{1-z_{ij}} = \frac{l_i}{1-l_i}\frac{v_j}{1-v_j}\frac{1-t}{t},$$

which is derived from Equation B.20. Then, we know each variable may be upper and lower bounded using c_1. For example,

$$\frac{c_1}{1-c_1} \leq \frac{l_i}{1-l_i} \leq \frac{(1-c_1)}{c_1}.$$

Then, there exists some constant τ, as a function of c_1, such that

$$\tau \leq z_{ij} \leq 1-\tau.$$

$\square$

Next, bound the Cheeger-constant of any subset of selected vertices. The Cheeger-constant allows us to lower bound the non-zero eigenvalues of the Laplacian.

Lemma B.13 (Expansion)**.** *For $\mathcal{W}_1$ and $\mathcal{W}_0$, and any vertex set T, there is a constant $\gamma_0 > 0$ such that*

$$|\partial_W T| \geq \mu n |T| \tag{B.40}$$

for $0 \leq |T| \leq n$.

Proof. Both $\mathcal{W}_0$ and $\mathcal{W}_1$ lie entirely inside of $\mathcal{U} \times \mathcal{V}$.

Let vertex set T contain x row vertices and y column vertices such that $x+y \leq n$. Then, we count all edges from rows in T to columns outside of T, and edges from all columns in T to rows outside of T.

For any row $i \in T$, there are a minimum of $(n - |\mathcal{V}| - y)$ available edges. Here, we lower bound by assuming $\mathcal{V}$ is all forbidden, and that the y columns in T are all in $\mathcal{V}^c$. For any row $j \in T$, there are a minimum of $(n - |\mathcal{U}| - x)$ available edges. Then,

$$
\begin{aligned}
|\partial_W T| &\geq x(n - |\mathcal{V}| - y) + y(n - |\mathcal{U}| - x) \\
&\geq n(1 - \delta - \frac{r}{n})(x+y) - 2xy \\
&\geq \frac{1}{2} n(1 - \delta - o(1))n(x+y).
\end{aligned}
\tag{B.41}
$$

Here, as $r = K_0 \log n = o(n)$, it contributes an $o(1)$ term. Then, $2xy \leq \frac{(x+y)^2}{2} \leq \frac{n(x+y)}{2}$. Then, as $|T| = x + y$, the proof is complete. $\square$

Next, for edge weights $w_{ij} = 2(z_{ij})(1 - z_{ij})$, we need to bound the first norm of the differences between w_{ij} values for $\mathcal{W}$ and $\mathcal{W}_0$.

Lemma B.14 (Bound on Laplacian Edge Weights). *For $w_{ij}^{(1)}$ as defined above,*

$$
\sum_{(i,j) \in \mathcal{W}_1 \cup \mathcal{W}_0} |w_{ij}^{(1)} - w_{ij}^{(0)}| = O(nr). \tag{B.42}
$$

Proof. First, for $w_{ij} = 2z_{ij}(1 - z_{ij})$, we use the following identity. As $0 \leq z_{ij} \leq 1$, we know that $w_{ij} = 2(\frac{1}{4} - (z_{ij} - \frac{1}{2})^2) \leq \frac{1}{2}$.

Next, we assume without loss of generality that both the rectangular support of $\mathcal{X} = \mathcal{U} \times \mathcal{V}$ and $\mathcal{X}_0 = \mathcal{U}_0 \times \mathcal{V}_0$ both occupy the first $d \times d$ vertices.

Positions in $\mathcal{X}_0 \setminus \mathcal{A}$ are forbidden in both $\mathcal{W}$ and $\mathcal{W}_0$. The positions outside of $\mathcal{X}$ are allowed in both $\mathcal{W}_0$ and $\mathcal{W}$, and each has a maximum entropy perturbation of $O(\frac{r}{n})$ by

Lemma B.12. There are a maximum of n^2 unchanged positions. Then,

$$\begin{aligned}|w_{ij}^{(1)} - w_{ij}^{(0)}| &= 2|z_{ij}^{(1)}(1 - z_{ij}^{(1)}) - z_{ij}^{(0)}(1 - z_{ij}^{(0)})| \\ &= 2|(z_{ij}^{(1)} - z_{ij}^{(0)})(1 - z_{ij}^{(1)} - z_{ij}^{(0)})| \\ &\leq 2|z_{ij}^{(1)} - z_{ij}^{(0)}|.\end{aligned}$$

The last inequality is derived from the fact that $|1 - z_{ij}^{(1)} - z_{ij}^{(0)}| \leq 1$. Then, as $|z_{ij}^{(1)} - z_{ij}^{(0)}| = O(\frac{r}{n})$ in this region, we find that the maximum weight change is $O(rn)$ for the $\mathcal{X}' \setminus \mathcal{X}_0$ region.

Next, we consider the set of allowed vertices $|\mathcal{A}|$ within the rectangular support of $\mathcal{E}$, and we consider the set of rows or columns which are incident to either $\mathcal{U}' \setminus \mathcal{U}_0$ or $\mathcal{V}' \setminus \mathcal{V}_0$, but not both. We know that $|\mathcal{A}| \leq nr$, such that $|\mathcal{A}| = O(nr)$. Then, we know there are $n(u+v)$ symbols which are incident in either row or column to $\mathcal{X}'$ but not to $\mathcal{X}_0$. There are a maximum of $|\mathcal{A}| + n(u+v) \leq 3nr$ such positions. For each specified position, we know that $|w_{ij}^{(1)} - w_{ij}^{(0)}| \leq \frac{1}{2}$, as each $w_{ij} < \frac{1}{2}$. Then, the total sum of $|w_{ij}^{(1)} - w_{ij}^{(0)}|$ for the specified vertices is $\frac{3}{2}nr = O(nr)$. Therefore,

$$\sum_{(i,j)\in\mathcal{W}_1\cup\mathcal{W}_0} |w_{ij}^{(1)} - w_{ij}^{(0)}| = O(nr).$$

□

Finally, we may make a statement on the determinant of the Laplacian, which allows us to apply corrections beyond the maximum entropy value to our asymptotic counting estimate.

Lemma B.15 (Determinant Calculation)**.** *For two allowed positions $\mathcal{W}_1$ and $\mathcal{W}_0$, let z_{ij}^1 and z_{ij}^0 denote the maximum entropy value at position $(i,j) \in \mathcal{W}_0 \cup \mathcal{W}_1$. For*

$$L_0 = \sum_{(i,j)\in\mathcal{W}_0\cup\mathcal{W}_1} 2z_{ij}^{(0)}(1 - z_{ij}^{(0)})\boldsymbol{c}_{ij}\boldsymbol{c}_{ij}^T \tag{B.43}$$

$$L_1 = \sum_{(i,j)\in\mathcal{W}_0\cup\mathcal{W}_1} 2z_{ij}^{(1)}(1 - z_{ij}^{(1)})\boldsymbol{c}_{ij}\boldsymbol{c}_{ij}^T, \tag{B.44}$$

let $w_{ij}^{(0)} = 2z_{ij}^{(0)}(1 - z_{ij}^{(0)})$ *and* $w_{ij}^{(1)} = 2z_{ij}^{(1)}(1 - z_{ij}^{(1)})$*. Suppose that*

$$\boldsymbol{x}^T L_0 \boldsymbol{x} \geq \mu n ||\boldsymbol{x}||_2^2 \text{ whenever } \sum_{v=1}^{2n} x_v = 0 \tag{B.45}$$

$$\boldsymbol{x}^T L_1 \boldsymbol{x} \geq \mu n ||\boldsymbol{x}||_2^2 \text{ whenever } \sum_{v=1}^{2n} x_v = 0 \tag{B.46}$$

for some constant μ*. Then,*

$$|\log(\frac{\kappa_1}{\kappa_0})| \leq \frac{2}{\mu n} \sum_{(i,j)} |w_{ij}^{(1)} - w_{ij}^{(0)}| = O(r). \tag{B.47}$$

Proof. First, for L_0 and L_t as defined, let

$$L_t = (1-t)L_0 + tL_1, \quad 0 \leq t \leq 1.$$

L_t is the linear combination of the two graph Laplacians. Then, similar to [IIM20], we let $M_t = L_t + \frac{1}{2n}J$, where J is the all ones matrix. Then, $\det M_t = 2n\kappa_t$ [IIM20].

Using the linearity of L_t in L_0 and L_1,

$$\mathbf{x}^T L_t \mathbf{x} = \mathbf{x}^T((1-t)L_0 + tL_1)\mathbf{x} \geq \mu n ||\mathbf{x}||_2^2.$$

For zero-sum vectors where $\sum_{ij} x_{ij} = 0$, the term $\frac{1}{m}J\mathbf{x} = \mathbf{0}$. For this case, the eigenvalues of M_t correspond to those of L_t, and are $\geq \mu$. For each zero-sum vector $\mathbf{c}_{ij}$, we the know that

$$0 \leq \mathbf{c}_{ij}^T M_t^{-1} \mathbf{c}_{ij} \leq \frac{1}{\mu n} ||\mathbf{c}_{ij}||_2^2 = \frac{2}{\mu n}$$

Here, we have used the fact that the eigenvalues associated with zero-sum eigenvectors are greater than or equal to μ, which means that those of M_t^{-1} are less than $\frac{1}{\mu n}$. Then, as $\mathbf{c}_{ij}$ has one 1 and one -1 entry, $||\mathbf{c}_{ij}||_2^2 = 2$.

Finally, we want to find quantity $\log \det M_t$. Let $f(t) = \log \det M_t$, and $\kappa_t = \frac{1}{2n} \det M_t$. Then, using the Mean Value Theorem,

$$\log(\frac{\kappa_1}{\kappa_0}) = f(1) - f(0) = f'(t^*)$$

for some $t^* \in (0,1)$. Then, using Jacobi's formula,

$$\begin{aligned} f'(t^*) &= \mathrm{Tr}(M_t^{-1}(L_1 - L_0)) \\ &= \sum_{ij} (w_{ij}^{(1)} - w_{ij}^{(0)}) \mathbf{c}_{ij}^T M_t^{-1} \mathbf{c}_{ij} \end{aligned}$$

where $\frac{dL_t}{dt} = L_1 - L_0$. Then, we may use the fact that $\mathbf{c}_{ij}^T M_t^{-1} \mathbf{c}_{ij} \leq \frac{2}{\mu}$ to bound

$$|\log(\frac{\kappa_1}{\kappa_0})| \leq \frac{2}{\mu n} \sum_{ij} |w_{ij}^{(1)} - w_{ij}^{(0)}|.$$

Finally, we know that $\sum_{ij} |w_{ij}^{(1)} - w_{ij}^{(0)}| = O(nr)$ from Lemma B.14. □

Lemma B.16 (Small-r Necessary Conditions)**.** *For an inclusion-minimal erasure pattern $\mathcal{E} \in \mathcal{E}_{inc}$ with allowed positions W and with $r \leq K_0 \log n$ for a finite, positive constant K,*

1. $(2n)^{\frac{1}{3}+\varepsilon} \leq \Delta \leq 2n - 1$ *for some constant* $\varepsilon > 0$.
2. $h(\mathcal{W}) \geq \gamma_1 \Delta$ *for some constant* $\gamma_1 > 0$.
3. *Equations* $\sum_{k:(j,k)\in G} \frac{r_j - r_k}{r_j + r_k} = b_j$ *for* $1 \leq j \leq n$ *such that* $\frac{r_j}{r_k} \leq 1 + R$ *for* $(j,k) \in G$, *where* $R = R(n)$ *satisfies* $0 \leq R = O(1)$ *and* $R^2 \frac{2n}{\Delta} \log(\frac{4n}{\Delta}) = o(\log n)$.

Proof. The maximum degree of any vertex is $\Delta = n$ for vertices in $\mathcal{U}^c \times \mathcal{V}^c$. The graph is bipartite with n row and n column vertices.

Using Lemma B.13, $h(\mathcal{W}) \geq \mu n$, and $\Delta = n$.

We use the tameness of $\tau \leq z_{ij}^{\mathcal{W}} \leq 1 - \tau$. Using our tameness result, $R = O(1)$, and $R^2 \frac{2n}{n} \log(\frac{4n}{n}) = O(1) = o(\log n)$. □

Finally, we may apply [IIM20, Theorem 1]. Then,

$$\log\left(\frac{|\Sigma(\mathbf{a}, \mathbf{a}; \mathcal{W})}{|\Sigma(\mathbf{a}, \mathbf{a}; \mathcal{W}_0)|}\right) = H(Z_{\mathcal{W}}) - H(Z_{\mathcal{W}_0}) + O(r) + O(1),$$

where the $O(r)$ is fully bounded from Lemma B.15 and the $O(1)$ is bounded by the remainder terms in [IIM20, Theorem 1] with our found values..

References

[0xb24] 0xbarchitect. "KZG polynomial commitments and their applications in blockchain (EIP-4844).", October 2024.

[Abd23] Fatma M. AbdelSalam. "Blockchain Revolutionizing Healthcare Industry: A Systematic Review of Blockchain Technology Benefits and Threats." *Perspectives in Health Information Management*, **20**(3):1b, September 2023.

[ASB21] Mustafa Al-Bassam, Alberto Sonnino, Vitalik Buterin, and Ismail Khoffi. "Fraud and Data Availability Proofs: Detecting Invalid Blocks in Light Clients." In *Financial Cryptography and Data Security: 25th International Conference, FC 2021, Virtual Event, March 1–5, 2021, Revised Selected Papers, Part II*, pp. 279–298, Berlin, Heidelberg, 2021. Springer-Verlag.

[Bar10] Alexander Barvinok. "On the number of matrices and a random matrix with prescribed row and column sums and 0–1 entries." *Advances in Mathematics*, **224**(1):316–339, May 2010.

[Bar12] Alexander Barvinok. "Matrices with prescribed row and column sums." *Linear Algebra and its Applications*, **436**(4):820–844, February 2012.

[BNN25] Dan Boneh, Joachim Neu, Valeria Nikolaenko, and Aditi Partap. "Data Availability Sampling with Repair.", 2025. Publication info: Preprint.

[Bre73] L. M. Brègman. "Some properties of nonnegative matrices and their permanents." *Doklady Akademii Nauk SSSR*, **211**(1):27–30, 1973.

[BST26] Amit Berman, Yaron Shany, and Itzhak Tamo. "Codes with Large Minimum Distance in Product Codes: Explicit Constructions and Bounds.", April 2026. arXiv:2604.15080 [cs.IT].

[BV23] Stephen P. Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge New York Melbourne New Delhi Singapore, version 29 edition, 2023.

[Con22] David Conlon. "Some remarks on the Zarankiewicz problem." *Mathematical Proceedings of the Cambridge Philosophical Society*, **173**(1):155–161, July 2022.

[Das16] Shagnik Das. "A brief note on estimates of binomial coefficients." 2016.

[Eli54] Peter Elias. *Error-free coding.* Research Laboratory of Electronics, Massachusetts Institute of Technology, 1954.

[Eth26] Ethereum.org. "PeerDAS.", June 2026.

[GHK17] Parikshit Gopalan, Guangda Hu, Swastik Kopparty, Shubhangi Saraf, Carol Wang, and Sergey Yekhanin. "Maximally Recoverable Codes for Grid-like Topologies." In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2092–2108. Society for Industrial and Applied Mathematics, January 2017.

[Gre21] Catherine Greenhill. "Generating graphs randomly." In Konrad K. Dabrowski, Maximilien Gadouleau, Nicholas Georgiou, Matthew Johnson, George B. Mertzios, and Daniël Paulusma, editors, *Surveys in Combinatorics 2021*, pp. 133–186. Cambridge University Press, 1 edition, June 2021.

[GRK26] Moritz Grundei, Aayush Rajasekaran, Kishori Konwar, and Médard Muriel. "From Indexing to Coding: A New Paradigm for Data Availability Sampling.", April 2026.

[HR01] J. Haglund and J.B. Remmel. "Rook Theory for Perfect Matchings." *Advances in Applied Mathematics*, **27**(2-3):438–481, August 2001.

[HSW23] Mathias Hall-Andersen, Mark Simkin, and Benedikt Wagner. "Foundations of Data Availability Sampling.", 2023. Publication info: Published by the IACR in CIC 2024.

[IIM20] Mikhail Isaev, Tejas Iyer, and Brendan D. McKay. "Asymptotic Enumeration of Orientations of a Graph as a Function of the Out-Degree Sequence." *The Electronic Journal of Combinatorics*, pp. P1.26–P1.26, January 2020.

[JAS21] Uzma Jafar, Mohd Juzaiddin Ab Aziz, and Zarina Shukur. "Blockchain for Electronic Voting System—Review and Open Research Challenges." *Sensors (Basel, Switzerland)*, **21**(17):5874, August 2021.

[JB16] Fanny Jardel and Joseph J. Boutros. "Edge Coloring and Stopping Sets Analysis in Product Codes with MDS components.", March 2016. arXiv:1603.01468 [cs.IT].

[joa22] joachimneu. "Data Availability Sampling: From Basics to Open Problems.", August 2022.

[KB25] Csaba Kiraly and Leonardo Bautista-Gomez. “Accelerating blob scaling with FullDASv2 (with getBlobs, mempool encoding, and possibly RLC) - Networking.”, May 2025. Section: Networking.

[KBR24a] Csaba Kiraly, Leonardo Bautista-Gomez, and Dmitriy Ryajov. “FullDAS: towards massive scalability with 32MB blocks and beyond - Sharding.”, May 2024. Section: Sharding.

[KBR24b] Csaba Kiraly, Leonardo Bautista-Gomez, and Dmitriy Ryajov. “LossyDAS: Lossy, Incremental, and Diagonal Sampling for Data Availability - Sharding.”, March 2024.

[Kil] Anzél Killian. “A Deep Dive Into Blockchain Scalability - Crypto.com International.”.

[KKA25] Narendra Kumar, Krishna Kumar, Anurag Aeron, and Filippo Verre. “Blockchain technology in supply chain management: Innovations, applications, and challenges.” *Telematics and Informatics Reports*, **18**:100204, June 2025.

[MS02] Paul Milgrom and Ilya Segal. “Envelope Theorems for Arbitrary Choice Sets.” *Econometrica*, **70**(2):583–601, March 2002.

[MTD22] Debarnab Mitra, Lev Tauz, and Lara Dolecek. “Overcoming Data Availability Attacks in Blockchain Systems: Short Code-Length LDPC Code Design for Coded Merkle Tree.” *IEEE Transactions on Communications*, **70**(9):5742–5759, September 2022. arXiv:2108.13332 [cs.IT].

[MTD23] Debarnab Mitra, Lev Tauz, and Lara Dolecek. “Graph Coded Merkle Tree: Mitigating Data Availability Attacks in Blockchain Systems Using Informed Design of Polar Factor Graphs.” *IEEE Journal on Selected Areas in Information Theory*, **4**:434–452, 2023.

[MTD25] Sonali Madisetti, Lev Tauz, and Lara Dolecek. “Data Availability Attacks: Targeted Sampling for 2-D Reed-Solomon Codes in Blockchain Systems.” In *2025 IEEE Information Theory Workshop (ITW)*, pp. 758–763, September 2025. ISSN: 2475-4218.

[Nag17] Zoltán Lóránt Nagy. “Supersaturation of C_4: from Zarankiewicz towards Erdős-Simonovits-Sidorenko.”, November 2017. arXiv:1711.09282 [math.CO].

[Nak] Satoshi Nakamoto. “Bitcoin: A Peer-to-Peer Electronic Cash System.”.

[Nei19] Richard D. Neidinger. “Multivariate Polynomial Interpolation in Newton Forms.” *SIAM Review*, **61**(2):361–381, January 2019.

[RCS24] Gabriel Antonio F. Rebello, Gustavo F. Camilo, Lucas Airam C. de Souza, Maria Potop-Butucaru, Marcelo Dias de Amorim, Miguel Elias M. Campista, and Luís Henrique M. K. Costa. "A Survey on Blockchain Scalability: From Hardware to Layer-Two Protocols." *IEEE Communications Surveys & Tutorials*, **26**(4):2411–2458, October 2024.

[RS60] I. S. Reed and G. Solomon. "Polynomial Codes Over Certain Finite Fields." *Journal of the Society for Industrial and Applied Mathematics*, **8**(2):300–304, June 1960.

[SRB22] Paolo Santini, Giulia Rafaiani, Massimo Battaglioni, Franco Chiaraluce, and Marco Baldi. "Optimization of a Reed-Solomon code-based protocol against blockchain data availability attacks.", January 2022. arXiv:2201.08261 [cs.IT].

[SVH25] Birenjith Sasidharan, Emanuele Viterbo, and Son Hoang Dau. "Block Circulant Codes With Application to Decentralized Systems." *IEEE Transactions on Communications*, **73**(10):8542–8557, October 2025.

[Tin21] Daniel Ting. "Simple, Optimal Algorithms for Random Sampling Without Replacement.", April 2021. arXiv:2104.05091 [cs.DS].

[Tod16] Peter Todd. "Fraud proofs.", 2016.

[Val79] L. G. Valiant. "The complexity of computing the permanent." *Theoretical Computer Science*, **8**(2):189–201, January 1979.

[YSL20] Mingchao Yu, Saeid Sahraei, Songze Li, Salman Avestimehr, Sreeram Kannan, and Pramod Viswanath. "Coded Merkle Tree: Solving Data Availability Attacks in Blockchains." In Joseph Bonneau and Nadia Heninger, editors, *Financial Cryptography and Data Security*, volume 12059, pp. 114–134. Springer International Publishing, Cham, 2020. Series Title: Lecture Notes in Computer Science.