Threshold rates for properties of random codes*

Venkatesan Guruswami[1], Jonathan Mosheiff[1], Nicolas Resch[2],
Shashwat Silas[3], and Mary Wootters[3]

[1]Carnegie Mellon University
[2]Centrum Wiskunde en Informatica
[3]Stanford University

November 2, 2021

Abstract

Suppose that $\mathcal{P}$ is a property that may be satisfied by a random code $C \subset \Sigma^n$. For example, for some $p \in (0,1)$, $\mathcal{P}$ might be the property that there exist three elements of C that lie in some Hamming ball of radius pn. We say that R^* is the *threshold rate* for $\mathcal{P}$ if a random code of rate $R^* + \varepsilon$ is very likely to satisfy $\mathcal{P}$, while a random code of rate $R^* - \varepsilon$ is very unlikely to satisfy $\mathcal{P}$. While random codes are well-studied in coding theory, even the threshold rates for relatively simple properties like the one above are not well understood.

We characterize threshold rates for a rich class of properties. These properties, like the example above, are defined by the inclusion of specific sets of codewords which are also suitably "symmetric." For properties in this class, we show that the threshold rate is in fact *equal* to the lower bound that a simple first-moment calculation obtains. Our techniques not only pin down the threshold rate for the property $\mathcal{P}$ above, they give sharp bounds on the threshold rate for *list-recovery* in several parameter regimes, as well as an efficient algorithm for estimating the threshold rates for list-recovery in general.

*SS and MW are partially funded by NSF-CAREER grant CCF-1844628, NSF-BSF grant CCF-1814629, and a Sloan Research Fellowship. SS is partially supported by a Google Graduate Fellowship. VG, JM and NR are partially funded by NSF grants CCF-1563742 and CCF-1814603 and a Simons Investigator Award. NR is also partially supported by ERC H2020 grant No.74079 (ALGSTRONGCRYPTO).

1 Introduction

Random codes are ubiquitous in the theory of error correcting codes: when thinking about the "right" trade-offs for a particular problem, a coding theorist's first instinct may be to try a random code. A *random code* here is simply a random set. That is, let $C \subseteq \Sigma^n$ be chosen so that each $x \in \Sigma^n$ is included in C with probability $|\Sigma|^{-n(1-R)}$ for some parameter R, which is called the (expected[1]) *rate* of the code C. Random codes are used in the proofs of the Gilbert-Varshamov bound, Shannon's channel coding theorem, and the list-decoding capacity theorem, to name just a few. This success may lead to the intuition that random codes are "easy" to analyze, and that the hard part is finding explicit constructions that match (or in rare cases, exceed) the parameters of random codes. However, there is still much we do not know about random codes, especially if we want extremely precise answers.

In particular, the question of *threshold rates*, of broader interest in probability theory, is something that we do not understand well for random codes. In more detail, suppose that $\mathcal{P}$ is a code property. For example, perhaps $\mathcal{P}$ is the property that there is some pair of codewords $c^{(1)}, c^{(2)} \in C$ that both lie in some Hamming ball of radius pn. Or perhaps $\mathcal{P}$ is the property that there are three codewords $c^{(1)}, c^{(2)}, c^{(3)} \in C$ that lie in such a Hamming ball. A value $R^* \in (0,1)$ is a *threshold rate* for $\mathcal{P}$ if a random code of rate $R^* + \varepsilon$ is very likely to satisfy $\mathcal{P}$, but a random code of rate $R^* - \varepsilon$ is very unlikely to satisfy $\mathcal{C}$. For the first example above, about pairs of codewords, the property in question is just the property of the code having *minimum distance* less than $2pn$, and this is not too hard to understand. However, already for the second example above—called *list-of-two decoding*—the threshold rate was not known.

1.1 Contributions

In this paper, we characterize threshold rates for a rich class of natural properties of random codes. We apply our characterization to obtain threshold rates for list-of-two decoding, as well as to properties like *list-decoding* and *perfect hashing codes,* and more generally to *list-recovery.* We outline our contributions below.

A characterization of the threshold rate R^* for symmetric properties. Suppose that $\mathcal{P}$ is a property defined by the inclusion of certain "bad" sets. For example, the list-of-two decoding property described above is defined by the inclusion of three codewords that lie in a radius-pn Hamming ball. For such properties that are also "symmetric enough," our main technical result, Theorem 1.1, characterizes the threshold rate R^*. Moreover, we show that this threshold rate is exactly the same as the lower bound that one obtains from a simple first-moment calculation! This is in contrast to recent work of [MRRZ+19] for random *linear* codes, which shows that the corresponding first-moment calculation is not the correct answer in that setting.

Part of our contribution is formalizing the correct notion of "symmetric enough." As we describe in the technical overview in Section 1.2, this definition turns out to be fairly subtle. Moreover, we give an example in Appendix A that shows that this definition is necessary: there are natural properties that do not meet this requirement, for which the simple first-moment calculation is *not* the correct threshold rate.

[1] Throughout, we refer to R as the rate of the code, and drop the adjective "expected."

Estimates of R^* for list-recovery. We give precise estimates of the threshold rate R^* for *list-recovery.* We say that a code $C \subseteq \Sigma^n$ is (p, ℓ, L)-list-recoverable if for all sets $K_i \subseteq \Sigma$ (for $1 \le i \le n$) with $|K_i| \le \ell$,

$$|\{c \in C : \Pr_{i \sim [n]}[c_i \notin K_i] \le p\}| < L.$$

List-recovery is a useful primitive in list-decoding, algorithm design, and pseudorandomness (see, e.g., [RW18, GUV09, Vad12]). In particular, it generalizes the list-of-two decoding example above (when $\ell = 1$ and $L = 3$), as well as other interesting properties, such as list-decoding and perfect hashing codes, discussed below.

Our characterization allows us to estimate or even exactly compute the threshold rate for (p, ℓ, L)-list-recovery in a wide variety of parameter regimes. To demonstrate this, we include several results along these lines. First, in Section 4 (Corollary 4.5), we give estimates that are quite sharp when $\frac{q \log L}{L}$ is small. In Section 5 (Lemma 5.1), we give an exact formula for the case $p = 0$, which is relevant for perfect hashing codes. In Section 6 (Theorem 6.1(I)), we give an exact formula for the case that $L = 3$ and $\ell = 1$, relevant for list-of-two decoding. Moreover, in Section 7 (Corollary 7.5) we use our characterization to develop an efficient algorithm to compute the threshold rate up to an additive error of $\varepsilon > 0$; our algorithm runs in time $O_p(L^q + \mathrm{poly}(q, L, \log(1/\varepsilon))$.

List-of-two decoding and a separation between random codes and random linear codes. We obtain new results for list-of-two decoding, the example discussed above. List-of-two decoding is a special case of *list-decoding,* which itself the special case of list-recovery where $\ell = 1$. We say that a code is (p, L)-list-decodable if there is no Hamming ball of radius pn containing L codewords; list-of-two decoding is the special case of $L = 3$.[2] We show in Section 6 (Theorem 6.1) that the threshold rate for this question, for random binary codes, is $R^* = 1 - \frac{1 - h_2(3p) + 3p \log_2 3}{3}$. That is, above this rate, a random binary code is very likely to have three codewords contained in a radius pn ball, while below this rate, the code most likely avoids all such triples.

This result is interesting for two reasons. First, it demonstrates that our techniques are refined enough to pin down the threshold rate in this parameter regime. Second, the particular value of R^* is interesting because it is *different* than the corresponding threshold rate for random *linear* codes. A *random linear code* over $\mathbb{F}_q$ of rate R is a random linear subspace of $\mathbb{F}_q^n$, of dimension Rn. The list-decodability of random linear codes has been extensively studied, and it is known (e.g., [ZP81, GHK11]) that the (p, L)-list-decoding threshold rate for both random linear codes and random codes is $1 - h_q(p)$, for sufficiently large list sizes L.[3] On the other hand, it is well-known that the distance of a random linear code (corresponding to the $L = 2$ case) is better than the distance of a completely random code. Our results show that this difference continues to hold for $L = 3$ (list-of-two decoding). We compute the threshold for random codes in Section 6, and in Appendix C, we show how to use the techniques of [MRRZ+19, GLM+20] to establish the list-of-two decoding threshold for random linear codes. The $L = 3$ case is interesting on its own, and moreover this result is a proof-of-concept to show that these techniques could perhaps be used to pin down the difference between random codes and random linear codes for larger (but still small) values of L.

[2] It is called list-of-*two* decoding, even though L is *three*, because any Hamming ball contains at most *two* codewords.

[3] Here, $h_q(x) = x \log_q(q-1) - x \log_q(x) - (1-x) \log_q(1-x)$ is the q-ary entropy.

Limitations of random codes for perfect hashing. Another special case of list-recovery is *perfect hashing codes.* Suppose that $|\Sigma| = q$. A code $C \subseteq \Sigma^n$ is said to be a q-hash code if, for any set of q distinct codewords $c^{(1)}, c^{(2)}, \ldots, c^{(q)} \in C$, there is at least one $i \in [n]$ so that $\{c_i^{(1)}, c_i^{(2)}, \ldots, c_i^{(q)}\} = \Sigma$; that is, if the set of symbols that appear in position i are all distinct. Thus, C is a q-hash code if and only if it is $(0, q-1, q)$-list-recoverable. As the name suggests, q-hash codes have applications in constructing small perfect hash families, and it is a classical question to determine the largest rate possible for a q-hash code.[4]

A simple random coding argument shows that a random code of rate $R = \frac{1}{q}\log_q \frac{1}{1-q!/q^q} - o(1)$ is a q-hash code with high probability [FK84, Kör86]. However, it is still an area of active research to do significantly better than this bound for any q. It is known that $R < \frac{q!}{q^{q-1}}$ for any q-hash code [FK84, GR19], and for large q, there is a gap of a multiplicative factor of about q^2 between these upper and lower bounds. Körner and Matron gave a construction that beats the random bound for $q = 3$ [KM88], and recently Xing and Yuan gave a construction that beats the random bound for infinitely many q's [XY19].

One might have hoped that a random code might in fact do better than the straightforward probabilistic argument (which follows from a union bound). Unfortunately, our results show that this is not the case. In Corollary 5.2, we use our characterization to pin down the threshold rate for perfect hashing codes, and show that, for random codes, the threshold rate is in fact $R^* = \frac{1}{q}\log_q \frac{1}{1-q!/q^q}$.

A broader view. Taking a broader view, threshold phenomena in other combinatorial domains, notably random graphs and Boolean functions, have been the subject of extensive study at least since Erdős and Rényi's seminal work [ER59]. Some of the deeper results in this field (e.g. [Fri99]), deal simultaneously with a wide class of properties, rather than a specific one. Other works, such as the recent [FKNP19], are general enough to cover not only multiple properties, but also multiple domains. Our work (as with the work of [MRRZ+19] on random linear codes, discussed below) is not as general as these, but we are able to get more precise results. It would be interesting to find a general framework that connects threshold phenomena in a variety of random code models, with analogues from random graphs and other natural combinatorial structures.

1.2 Technical Overview

As mentioned above, we study properties defined by the inclusion of bad subsets. We organize bad subsets of size b into matrices $B \in \Sigma^{n \times b}$, interpreting the columns of B as the elements of the set. We write "$B \subseteq C$" to mean that the columns of B are all contained in the code C.

As a running example—and also our motivating example—consider list recovery, defined above. The property $\mathcal{P}$ of *not* being (p, ℓ, L)-list-recoverable is defined by the inclusion of "bad" matrices $B \in \Sigma^{n \times L}$ so that for some sets $K_1, \ldots, K_n \subset \Sigma$ of size at most ℓ, $\Pr_{i \sim [n]}[B_{ij} \notin K_i] \leq p$ for each $j \in [L]$. Moreover we require the columns of B to be distinct.

[4] A q-hash code naturally gives rise to a perfect hash family: suppose that C is a universe of items, and define a hash function $h_i : C \to \Sigma$ given by $h_i(c) = c_i$. Then the property of being a q-hash code is equivalent to the property that, for any set of q items in the universe, there exists some hash function h_i for $1 \leq i \leq n$ that maps each item to a different value.

Analyzing a property as a union of types. Following the approach of [MRRZ+19] for random linear codes, we group the bad matrices into *types* based on their row distributions. That is, for a bad matrix $B \in \Sigma^{n \times b}$, let τ denote the row distribution

$$\tau(v) = \frac{|\{i \in [n] : B_{i,\star} = v\}|}{n},$$

where $B_{i,\star}$ denotes the i'th row of B. We say that B has *type* τ. Consider the set $\mathcal{B}$ of all of the matrices of type τ; equivalently, $\mathcal{B}$ is the set of matrices obtained by permuting the rows of B.

For example, if $B \subset \{\alpha, \beta, \gamma\}^{n \times 3}$ is given by

$$B = \begin{bmatrix} \alpha & \alpha & \beta \\ \alpha & \alpha & \beta \\ & \vdots & \\ \alpha & \alpha & \beta \\ \gamma & \beta & \beta \\ & \vdots & \\ \gamma & \beta & \beta \\ \gamma & \beta & \beta \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \} n/2 \\ \\ \\ \\ \} n/2 \end{matrix}$$

then B is bad for $(0, 2, 3)$-list-recovery, the row distribution τ of B is given by

$$\tau((\alpha, \alpha, \beta)) = \tau((\gamma, \beta, \beta)) = \frac{1}{2},$$

and the set $\mathcal{B}$ consists of all $n \times 3$ matrices with half the rows (α, α, β) and the other half (γ, β, β).

We note that possible types τ depend on n, because of divisibility constraints. For simplicity, let us ignore these restrictions for now (we will deal with them later), and suppose that a single type τ can appear for all n.

First-moment bound and main theorem. We can use a simple first-moment approach to give a lower bound on the threshold rate. In more detail, the probability that a particular B is contained in C is $q^{-nb(1-R)}$, assuming that B has b distinct columns. Using the fact that $|\mathcal{B}| \approx q^{H_q(\tau) \cdot n}$, where $H_q(\tau)$ is the base-q entropy of τ (see Section 2), and applying a union bound over all $B \in \mathcal{B}$, we see that the probability that any $B \in \mathcal{B}$ is contained in C is at most

$$q^{nb(H_q(\tau) - (1-R))}.$$

Thus, if $R \leq 1 - \frac{H_q(\tau)}{b} - \varepsilon$ for some small $\varepsilon > 0$, it is very unlikely that τ will be represented in C.

Now suppose that our collection of bad sets, which define the property $\mathcal{P}$, is closed under row permutations. This means that $\mathcal{P}$ can be represented as a collection T of types τ; note that the size of T is polynomial in n. Union bounding over all of these types, the computation above shows that a random code C of rate $R < 1 - \max_{\tau \in T} \frac{H_q(\tau)}{b} - \varepsilon$ will, with high probability, not satisfy $\mathcal{P}$.

The question is, could the rate be larger? Might it be the case that $\mathcal{P}$ still not satisfied (with high probability) by a random code of rate R significantly larger than $1 - \max_\tau H_q(\tau)/b$? In

[MRRZ+19], it was shown that the answer for random *linear* codes is "yes." If $\mathcal{P}$ exhibits certain linear structure, then it may be possible that a higher rate random linear code still does not satisfy $\mathcal{P}$ with high probability. One may conjecture that something similar holds for random codes.

Our main technical result, Theorem 3.6, is that, for random codes, for sufficiently symmetric properties, the answer to this question is "no." That is, the simple calculation above *does* give the right answer for random codes!

Theorem 1.1 (Informal; see Theorem 3.6 for the formal version)**.** *Let $\mathcal{P}$ be a "symmetric" property defined by the inclusion of a type among the types in T. Let*

$$R^* = 1 - \frac{\max_{\tau \in T} H_q(\tau)}{b}$$

Then for all $\varepsilon > 0$, a random code of rate $R \geq R^* + \varepsilon$ satisfies $\mathcal{P}$ with probability $1 - o(1)$, while a random code of rate $R^* - \varepsilon$ satisfies $\mathcal{P}$ with probability $o(1)$.

Sketch of proof: second moment method. Below, we sketch the proof of Theorem 1.1, and explain what the assumption of "symmetry" means. As noted above, it is straightforward to show that the threshold rate R^* is at least $1 - \max_{\tau \in T} \frac{H_q(\tau)}{b}$, so the challenge is to show that it is not larger. The proof of Theorem 1.1 uses the second-moment method to show that for any *histogram type* τ (we discuss histogram types more below), a random code C of rate $1 - H_q(\tau)/b + \varepsilon$ is very likely to contain some matrix B with type τ. Thus, the threshold rate is at most $1 - \max_\tau H_q(\tau)/b$, where the maximum is over all histogram types τ that appear in T. Our eventual definition of "symmetric" will guarantee that it is legitimate to restrict our attention to histogram types.

Histogram types and the meaning of "symmetry." In order to apply the second moment method, we bound the variance of $\sum_{B \sim \tau} \mathbf{1}[B \subset C]$, where the sum is over all matrices B of type τ. This turns out to be possible when τ has the following symmetry property: for any $u \in \Sigma^b$, and for any permutation $\pi : [b] \to [b]$, it holds that $\tau(u) = \tau(\pi(u))$, where $\pi(u)$ denotes the corresponding coordinate permutation of u. We call such a type τ a *histogram-type* (Definition 3.3) because the probability of a particular vector u under τ depends only on the histogram of u.

A first attempt to formulate a definition of "symmetry" for Theorem 1.1 is thus to require $\mathcal{P}$ to be defined by histogram types. This results in a true statement, but unfortunately it is too restrictive: it is not hard to see that, for example, the property of not being list-decodable contains types τ that are not histogram types. Fortunately, for the logic above to go through, it is enough to show that T contains a type τ that is *both* a maximum entropy distribution in T, and is also a histogram type. Thus, the assumption of "symmetry" we will use is that T, the collection of types represented in the property $\mathcal{P}$, forms a convex set. Then, using the fact that $\mathcal{P}$ is defined by the inclusion of bad sets (which do not care about the order of the columns in the corresponding matrices), we can always find a maximum entropy histogram type by "symmetrizing" and taking a convex combination of column permutations of some maximum entropy type τ.

One might wonder if this symmetrization step (and the resulting assumption about convexity) is necessary. In fact, it is. In Appendix A, we give an example of a property $\mathcal{P}$ that is given by the union of bad types, but which is not closed under such "symmetrization". For this property, the threshold rate turns out to be larger than the rate predicted by Theorem 1.1.

Taking a limit as $n \to \infty$. There is one more challenge to consider, which is that in the description above, we have ignored the fact that we would like our characterization to work for a sequence of values of n. However, a type τ only works for certain values of n due to divisibility restrictions. To get around this, we work instead with a sequence of types τ_n which tend to τ. This leads us to our final definition of "symmetric" (Definition 2.19). Suppose that $\mathcal{P}$ is a property defined by the inclusion of size-b bad sets. Then for each n, there is some collection T^n of bad types τ_n, each of which is a distribution on Σ^b. We say that $\mathcal{P}$ is *symmetric* if the sets T^n approach some convex set T as n goes to infinity. The logic above then goes through to give Theorem 1.1.

Applications to list-recovery. Finally, in order to apply Theorem 1.1, we need to understand the maximum entropy distribution τ for our property $\mathcal{P}$. We do this for the property $\mathcal{P}$ of not being (p, ℓ, L)-list-recoverable in a variety of parameter regimes in Sections 4, 5 and 6, and along the way obtain our results about list-of-two decoding and perfect hashing codes. Finally, in Section 7, we use our framework to develop an algorithm to efficiently calculate the threshold rate for (p, ℓ, L)-list-recovery.

1.3 Related Work

Below, we briefly survey work in a few categories that relates to our results.

Limitations of random codes. Random codes have been studied in coding theory since its inception, as they are used to prove fundamental existential results. However, typically these constructions show that a random code is "good" (e.g., list-decodable, or a q-hash code) with high probability, not that a random code of slightly higher rate is *not* good. Our work establishes a threshold rate R^* for random codes, meaning that in particular it establishes both a positive and a negative result. There are a few works that establish limitations for random codes. For example, it is known that random codes of rate $1 - h_q(p) - \varepsilon$ require list sizes of at least $\Omega(1/\varepsilon)$ [GN14, LW18], which is a larger lower bound than that which is known for general codes.

Sharp thresholds for random linear codes. Our work takes inspiration from the techniques of [MRRZ+19], which recently developed sharp thresholds for random *linear* codes. The starting points for that work and our work are similar: in particular, that work also classified bad sets into types based on their row distribution. However, as discussed above, the situation for random linear codes is different than the situation for random codes, because for random linear codes looking at the entropy-maximizing distribution τ in a property is not enough: one has to look at suitable projections of these distributions τ. One of the contributions of this work is showing that for random codes, there are no such complications, at least for suitably symmetric properties. The entropy-maximizing distribution in a property directly determines the threshold rate.

List-of-two decoding. As mentioned above, list-of-two decoding is a special case of (p, L)-list-decoding when $L = 3$. For larger L, a classical result known as the list-decoding capacity theorem says that a random code $C \subseteq \Sigma^n$ of rate $R \leq 1 - h_q(p) - \varepsilon$, where $q = |\Sigma|$, is $(p, O(1/\varepsilon))$ list-decodable with high probability; while if $R \geq 1 - h_q(p) + \varepsilon$, then no code of rate R is (p, L) list-decodable for any $L = o(2^n)$. Thus, for large enough list sizes, the threshold rate for list-decoding

is $R^* = 1 - h_q(p)$.

While the list-size L is often viewed as a growing parameter in the list-decoding literature, it is also interesting to consider fixed values of L. A natural first step is to consider list-of-two decoding, which in our notation is $(p, 3)$-list-decodability. We focus on the case where $q = 2$. Many prior works have studied the best rate (in terms of p) at which list-of-two decoding is possible. First of all, Blinovsky [Bli86] demonstrated that in order to have positive rate $(p, 3)$-list-decodable codes for infinitely many block lengths n, it is necessary and sufficient to have $p < 1/4$. Moreover, [Bli86] gave bounds on this best rate; these bounds were further improved in [ABL00]. Results concerning list-of-2 decoding in the zero-rate regime (that is, where the rate tends to zero as n tends to infinity) have also been obtained in, e.g. [Eli91, ABP18]. The works summarized here have focused on the *best* possible rate for list-of-two-decoding. In contrast, our work focuses on random codes in particular, with the goal of being able to pin down the threshold rate very precisely. Our result shows that random codes are *not* the best for list-of-two decoding.

1.4 Future directions and open questions

We have given a characterization of the threshold rate for a rich class of properties to be satisfied by a random code, and we have focused on list-recovery as an example property. We hope that this framework, and its potential extensions, will be useful more broadly. We mention a few open questions and directions for future research.

Extending the zero-rate regime? Due to the list-recovery capacity theorem, it is known that when $p \geq 1-\ell/q$, there are no positive rate (p, ℓ, L)-list-recoverable codes over an alphabet of size q, for any (subexponential) list size L. However, for small values of L, it is natural to expect that even for values of p slightly less than $1-\ell/q$, there are still no positive-rate (p, ℓ, L)-list-recoverable codes. For example, if we want a positive rate $(p, 2, 3)$-list-recoverable code over an alphabet of size 5, must we have $p \leq 0.599$? We show that this is the case for "most" codes: more precisely, for any integers ℓ, L there exists a $p^* < 1 - \ell/q$ such that a positive rate random code is (p, ℓ, L)-list-recoverable with high probability if and only if $p < p^*$ (see Remark 7.4). It would be interesting to determine whether or not this holds for *all* codes: namely, can we prove that a (p, ℓ, L)-list-recoverable code must have zero rate if $p \geq p^*$ for the same value of p^*?

Other properties of random codes? We have focused on list-recovery and its special cases as examples of symmetric properties. What other examples are there? Could we extend our work beyond symmetric properties? As the example in Appendix A shows, extending our results beyond symmetric properties would require a more complicated expression than the one in Theorem 1.1.

Other ensembles of codes? Our framework draws inspiration from the work [MRRZ+19] that develops a similar framework for random linear codes. What other random ensembles of codes are amenable to such a framework? Is there some meta-framework that would encompass more of these? A potential starting point would be to study *pseudolinear* codes [Gur04, GI01] which in some sense interpolate between uniformly random codes (as we study in this work) and random linear codes (as addressed by [MRRZ+19, GLM+20]). It would be also be interesting to study random (and random linear) codes in the rank-metric [Din15, GR18]. Random subspaces of Euclidean space

are of significant interest in many areas including compressed sensing, dimensionality reduction, and Euclidean sections, and it would be interesting to investigate to what extent, if at all, the frameworks in [MRRZ+19, GLM+20] or this work might apply to these settings.

Sharp thresholds for list-of-$(L-1)$ decoding? We used our framework to compute the threshold for list-of-two decoding (e.g., $L = 3$), and we believe that our techniques could be used for pinning down the threshold rate, in terms of L, for any L. While this is known asymptotically for large L [GN14, LW18], the precise value remains open for constant $L > 3$.

1.5 Organization

In Section 2, we introduce notation, and also set up the definitions we need about types, thresholds, properties, and various notions of symmetry. We also introduce (non-)list-recoverability as a property, and prove in Corollary 2.23 that it is symmetric.

In Section 3, we state and prove Theorem 3.6, the formal version of the characterization theorem (Theorem 1.1 above). At the end of Section 3, we begin to apply Theorem 3.6 to list-recovery, and in particular define several notions we will need to analyze list recovery in the subsequent sections.

In the remaining sections, we specialize to list-recovery. In Section 4, we develop bounds on the threshold rate R^* for list-recovery that are tight when $(q \log L)/L$ is small. In Section 5, we compute the threshold rate R^* exactly for zero-error list-recovery (that is, when $p = 0$), and use this to compute the threshold rate for perfect hashing. In Section 6, we compute the threshold rate R^* for list-of-two decoding (e.g., list-recovery when $\ell = 1$ and $L = 3$), and use this to quantify the gap between random codes and random linear codes for list-of-two decoding. Finally, in Section 7, we give an efficient algorithm to compute the threshold rate.

2 Preliminaries

First, we fix some basic notation. Throughout, we consider codes $C \subseteq \Sigma^n$ of block length n over an alphabet Σ, where $|\Sigma| = q$. When we use $\log(x)$ without an explicit base, we mean $\log_2(x)$. We use H_q to denote the base-q entropy: for a distribution τ,

$$H_q(\tau) := -\sum_x \tau(x) \log_q(\tau(x)).$$

When q is clear from context, we will use $H(\tau)$ to denote $H_q(\tau)$. If u is a random variable distributed according to τ, then we abuse notation slightly and define $H(u) := H(\tau)$. We use $h_q(x) := x \log_q(q-1) - x \log_q(x) - (1-x) \log_q(1-x)$ to denote the q-ary entropy of $x \in (0,1)$. Again, when q is clear from context we will use $h(x)$ to denote $h_q(x)$.

For a vector $x \in \Sigma^k$ and $I \subseteq [k]$, we use x_I to refer to the vector $(x_i)_{i \in I} \in \Sigma^I$. Given a vector $u \in \Sigma^k$ and a permutation $\pi : [k] \to [k]$, we let $\pi(u) \in \Sigma^k$ denote the corresponding coordinate permutation of u.

Given distributions τ, μ on the same finite set, we define their ℓ_∞-distance by

$$d_\infty(\tau, \mu) := \max_x |\tau(x) - \mu(x)| \ .$$

Given a set of distributions T, we define the ℓ_∞ distance from μ to T by

$$d_\infty(\mu, T) := \inf_{\tau \in T} d_\infty(\mu, \tau) \ .$$

2.1 Basic notions

As mentioned in the introduction, we will organize our "bad" sets into matrices. We formalize this with the following two definitions.

Definition 2.1 (Matrices with distinct columns)**.** *Let $\Sigma^{n \times b}_{\mathrm{distinct}}$ denote the collection of all matrices $B \in \Sigma^{n \times b}$ such that each column of B is distinct.*

Definition 2.2 (Subsets as matrices)**.** *Let $C \subseteq \Sigma^n$ be a code, and let $B \in \Sigma^{n \times b}$ be a matrix. We write $B \subseteq C$ to mean that each column of B is an element of C. If $A \subseteq \Sigma^n$, let $\mathcal{B}_A \subseteq \Sigma^{n \times |A|}$ denote the collection all matrices $B \in \Sigma^{n \times |A|}_{\mathrm{distinct}}$ such that the columns of B are the elements of A.*

For completeness, we reiterate our definition of a random code from the introduction.

Definition 2.3 (Random code)**.** *Let Σ be a finite set with $q := |\Sigma| \geq 2$. For $n \in \mathbb{N}$ and $R \in [0, 1]$, let $C^n_{\mathrm{RC}}(R)$ denote an* expected-rate R random code (over the alphabet Σ) *$C \subseteq \Sigma^n$. Namely, for each $x \in \mathbb{F}^n_q$ we have $\Pr[x \in C] = q^{-n(1-R)}$, and these events are independent over all x.*

We record a useful fact about random codes, which is the probability that any particular matrix B is contained in one.

Fact 2.4 (Probability that a random code contains a matrix)**.** *Let $B \in \Sigma^{n \times b}$. Then,*

$$\Pr[B \subseteq C^n_{\mathrm{RC}}(R)] = q^{-n(1-R)t},$$

where t is the number of distinct columns in B.

We study (noisy) list-recovery, which generalizes both the list-decoding and perfect hashing examples mentioned in the introduction. We repeat the definition, so that we may formally define a "bad" matrix for list-recovery.

Definition 2.5 (Noisy list-recovery)**.** *Let $p \in [0, 1]$, $1 \leq \ell \leq q$, and $L \in N$. Say that a matrix $B \in \Sigma^{L \times n}_{\mathrm{distinct}}$ is* (p, ℓ, L)-bad *for (p, ℓ, L)-list-recovery if there exist sets $K_i \subseteq \Sigma$ $(1 \leq i \leq n)$, each of size ℓ, such that for every $1 \leq j \leq L$,*

$$\Pr_{i \sim [n]}[B_{i,j} \notin K_i] \leq p. \tag{1}$$

A code $C \subseteq \Sigma^n$ is (p, ℓ, L)-list-recoverable *if it does not contain a (p, ℓ, L)-bad matrix.*

2.2 Monotone-increasing properties and thresholds

We study the threshold rate R^* for random codes to satisfy certain properties. This was discussed informally in the introduction and the definitions below formalize what "threshold rate" means.

Definition 2.6 (Monotone-increasing property)**.** *A code property $\mathcal{P}$ is* monotone-increasing *if given a code C satisfying $\mathcal{P}$, it holds that every code C' such that $C \subseteq C'$ also satisfies $\mathcal{P}$.*

For example, the property of being *not* (p, ℓ, L)-list-recoverable (that is, the property of containing a (p, ℓ, L)-bad matrix) is a monotone-increasing property.

Definition 2.7 (Minimal-set)**.** *Let P_n be a monotone-increasing property of length-n codes. A set $A \subseteq \Sigma^n$ is a* minimal element *of P_n if A satisfies P_n but no strict subset of A satisfies P_n. The* minimal set *for P_n is the collection of matrices*

$$\bigcup_{A \text{ is a minimal element of } P_n} \mathcal{B}_A.$$

For example, the minimal set for the property P_n of being *not* (p, ℓ, L)-list-recoverable is the set of (p, ℓ, L)-bad matrices.

Note that a code satisfies P_n if and only if it contains some matrix belonging to the minimal set of P_n. If $\mathcal{P}$ is a monotone-increasing property of codes, we denote its associated *threshold rate* by $R^n_{\mathrm{RC}}(\mathcal{P})$. This is defined as

$$\sup\left\{R \in [0,1] : \Pr\left[C^n_{\mathrm{RC}}(R) \text{ satisfies } \mathcal{P}\right] \leq \frac{1}{2}\right\}$$

if there is such an R, and 0 otherwise.

Remark 2.8. *If $\mathcal{P}$ is monotone-increasing then the function $\Pr\left[C^n_{\mathrm{RC}}(R) \text{satisfies } \mathcal{P}\right]$ is monotone-increasing in R. This can be proved by a standard coupling argument, akin to [Bol01, Thm. 2.1].*

Definition 2.9 (Sharpness for random codes)**.** *A monotone-increasing property $\mathcal{P}$ is* sharp for random codes *if*

$$\lim_{n\to\infty} \Pr\left[C^n_{\mathrm{RC}}\left(R^n_{\mathrm{RC}}(\mathcal{P}) - \varepsilon\right) \text{ satisfies } \mathcal{P}\right] = 0$$

and

$$\lim_{n\to\infty} \Pr\left[C^n_{\mathrm{RC}}\left(R^n_{\mathrm{RC}}(\mathcal{P}) + \varepsilon\right) \text{ satisfies } \mathcal{P}\right] = 1$$

for every $\varepsilon > 0$.

2.3 Local and row-symmetric properties

As discussed in the introduction, we study properties that can be written as a union of "types," where each type corresponds to a row distribution τ of a matrix M. The following definitions make this notion precise.

Definition 2.10 (Row-permutation invariant collection of matrices)**.** *A collection of matrices $\mathcal{B} \subseteq \Sigma^{n\times b}$ is* row-permutation invariant *if, given a matrix $B \in \mathcal{B}$, every row permutation of B (that is, a matrix resulting from applying the same coordinate permutation to each column of B) also belongs to $\mathcal{B}$.*

Definition 2.11 (Local and row-symmetric properties)**.** *Let $\mathcal{P} = \{P_n\}_{n\in\mathbb{N}}$ be a monotone-increasing property, and let M_n denote the minimal set of P_n.*

- *If there exists some $b \in \mathbb{N}$ such that $M_n \subseteq \Sigma^{n\times b}$ for every n, we say that $\mathcal{P}$ is* b-local.
- *If every M_n is row-permutation invariant, we say that $\mathcal{P}$ is* row-symmetric.

Remark 2.12. *Every monotone-increasing property is trivially* column-symmetric*, in the sense that permuting the columns of a matrix in M_n results in another matrix in M_n. This naturally reflects the fact that containment of a matrix does not depend on the ordering of the columns, and follows immediately from the definition of a minimal set.*

Let $B \in \Sigma^{n\times b}$, and consider the collection $\mathcal{B}$ of all row-permutations of B. Let τ_B denote the *row-distribution* of B. That is, τ is the probability distribution, over Σ^b, of the row $B_{i,\star}$, where i is sampled uniformly from $[n]$. Observe that every matrix in $\mathcal{B}$ has the same row-distribution as B. Moreover, $\mathcal{B}$ can be characterized as the set of all matrices with the row distribution τ_B. These observations motivate the following definitions.

Definition 2.13 (Type of a matrix)**.** *Let $B \in \Sigma^{n\times b}$. We define its* type *τ_B as the distribution of a uniformly random row of B. That is, τ_B is the distribution over Σ^b, such that*

$$\tau_B(x) = \frac{|\{i \in [n] \mid B_i = x\}|}{n}$$

for every $x \in \Sigma^b$. Let

$$\mathcal{T}_b^n = \{\tau_B \mid B \in \Sigma^{n\times b}_{\text{distinct}}\}$$

denote the set of all possible types of $n \times b$ matrices with distinct columns. Given $\tau \in \mathcal{T}_b^n$, we denote

$$M_\tau = \{B \in \Sigma^{n\times b} \mid \tau_B = \tau\}.$$

Remark 2.14. *The type of a matrix $B \in \Sigma^{n\times b}$ determines whether $B \in \Sigma^{n\times b}_{\text{distinct}}$. Therefore, for $\tau \in \mathcal{T}_b^n$,*

$$M_\tau = \{B \in \Sigma^{n\times b}_{\text{distinct}} \mid \tau_B = \tau\}.$$

The following fact now follows from the above discussion.

Fact 2.15 (Decomposition of a row-permutation invariant collection)**.** *Let $\mathcal{B} \subseteq \Sigma^{n\times b}$ be a row-permutation invariant collection. Then, there exists a set of types $T \subseteq \mathcal{T}_{n,b}$ such that*

$$\mathcal{B} = \bigcup_{\tau\in T} M_\tau.$$

Note that a type in $\mathcal{T}_b^n$ is defined by the number of occurrences of each of $|\Sigma^b|$ possible rows, in a matrix consisting of n rows. In particular, each row occurs between 0 and n times. Thus,

$$|\mathcal{T}_b^n| \leq (n+1)^{|\Sigma^b|} = (n+1)^{q^b}. \tag{2}$$

Crucially for our purposes, this upper bound is polynomial in n.

2.4 Symmetric properties and convex approximations

Definition 2.16. *Let $\mathcal{T}_b$ denote the simplex of all probability distributions over Σ^b.*

It is generally more convenient to work in $\mathcal{T}_b$ rather than $\mathcal{T}_b^n$, since the former is continuous, while the latter is discrete and involves certain divisibility conditions. This motivates the following definition.

Definition 2.17 (Permutation-closed type sets)**.** *A set $T \subseteq \mathcal{T}_b$ is called* permutation-closed *if for every $\tau \in T$ and every permutation $\pi : [b] \to [b]$, the distribution of $\pi(u)$ (where $u \sim \tau$) also belongs to T.*

Definition 2.18 (Approximating sets of types)**.** *Fix $b \in \mathbb{N}$. Let $\{T^n\}_{n\in\mathbb{N}}$ be a sequence of sets of types, such that $T^n \subseteq \mathcal{T}_b^n$. A (topologically) closed and permutation-closed set $T \subseteq \mathcal{T}_b$ is an* approximation *for $\{T^n\}_{n\in\mathbb{N}}$ if $T^n \subseteq T$ for every n, and*

$$\lim_{n\to\infty} \max_{\tau\in T} d_\infty(\tau, T^n) = 0. \tag{3}$$

Definition 2.19 (Symmetric property and convex approximation)**.** *Let $\mathcal{P} = \{P_n\}_{n\in\mathbb{N}}$ be a b-local, row-symmetric, monotone-increasing property. Due to Fact 2.15, for every n there exists a set $T_n \subseteq \mathcal{T}_{n,b}$ such that the minimal set of P_n is $\bigcup_{\tau\in T_n} M_\tau$. If the sequence $\{T_n\}_{n\in N}$ has a convex approximation T, we say that T is a* convex approximation *for $\mathcal{P}$. In this case, we say that $\mathcal{P}$ is* symmetric*.*

2.5 Non-list-recoverability as a property

Our motivating property is that of being *not* list-recoverable. In this section, we show that non-(p, ℓ, L)-list-recoverability is a symmetric property, and we define the convex set $T_{p,\ell,L}$ that is a convex approximation for it.

Fix $p \in [0, 1]$, $1 \le \ell \le q$ and $L \in \mathbb{N}$. Let $\mathcal{P} = (P_n)_{n\in\mathbb{N}}$ denote the property of being *not* (p, ℓ, L)-list-recoverable. That is, a code $C \subseteq \Sigma^n$ satisfies P_n if it contains a (p, ℓ, L)-bad matrix. We now show that $\mathcal{P}$ is a symmetric property.

Clearly, $\mathcal{P}$ is monotone-increasing, and its minimal set is exactly the set of (p, ℓ, L)-bad matrices, which we denote $M_n \subseteq \Sigma^{n\times L}_{\text{distinct}}$. It follows immediately that $\mathcal{P}$ is L-local. Furthermore since the left-hand side of (1) is invariant to row-permutations of B, the collection M_n is row-permutation invariant, and so $\mathcal{P}$ is row-symmetric.

Fact 2.15 says that we can write $M_n = \bigcup_{\tau\in T^n_{p,\ell,L}} M_\tau$ for some $T^n_{p,\ell,L} \subseteq \mathcal{T}^n_L$. Indeed, (1) yields the following description of $\mathcal{T}^n_{p,\ell,L}$: A type $\tau \in \mathcal{T}^n_L$ belongs to $T^n_{p,\ell,L}$ if and only if there exists a distribution ρ over $\Sigma^L \times \binom{\Sigma}{\ell}$ such that, given $(u, K) \sim \rho$, the following holds:

1. The distribution of u is τ.
2. For every $1 \le j \le L$, it holds that $\Pr[u_j \notin K] \le p$.
3. $n \cdot \rho((u, K)) \in \mathbb{N}$ for every u and K.

To see this, let ρ be the joint distribution (B_i, K_i) for i uniformly sampled from $[n]$, where B and K are as in (1). Note that ρ must satisfy the three conditions above. In the other direction, it is not hard to see that any such distribution ρ as above gives rise to a matrix of type τ, satisfying (1).

We next construct a convex approximation for $\mathcal{P}$. Let $T_{p,\ell,L}$ denote the set of all types $\tau \in \mathcal{T}_L$ for which there exists a distribution ρ satisfying Conditions 1 and 2, but not necessarily Condition 3:

Definition 2.20. *Let $1 \le \ell \le 1$, $L \in \mathbb{N}$ and $0 \le p \le 1$. Let τ be a distribution over Σ^L. We say that τ belongs to the set $T_{p,\ell,L}$ if there exists a distribution ρ over $\Sigma^L \times \binom{\Sigma}{L}$ such that:*

1. *If $(u, K) \sim \rho$ then the vector u is τ-distributed.*
2. *For every $1 \le j \le L$ it holds that*

$$\Pr_{(u,K)\sim\rho}[u_j \notin K] \le p. \tag{4}$$

Clearly, $T^n_{p,\ell,L} \subseteq T_{p,\ell,L}$ for all $n \in \mathbb{N}$. It is also immediate to verify that $T_{p,\ell,L}$ is permutation-closed.

Lemma 2.21. *The set $T_{p,\ell,L}$ is convex.*

Proof. Let $\tau_0, \tau_1 \in T_{p,\ell,L}$. Let $t \in [0,1]$ and let τ_t denote the mixture distribution $(1-t)\tau_0 + t\tau_1$. Let ρ_0 and ρ_1 be distributions over $\Sigma^L \times \binom{\Sigma}{\ell}$, satisfying Conditions 1 and 2 for τ_0 and τ_1, respectively. Let ρ_t be the mixture distribution $(1-t)\rho_0 + t\rho_1$. It is straightforward to verify that ρ_t satisfies Conditions 1 and 2 with respect to τ_t. Hence, $\tau_t \in T_{p,\ell,L}$. □

The following lemma, proven in Appendix B shows that $T_{p,\ell,L}$ satisfies (3). Namely, every type in $T_{p,\ell,L}$ can be realized with low error as a type from $T^n_{p,\ell,L}$, for large enough n.

Lemma 2.22.

$$\lim_{n\to\infty} \sup_{\tau \in T_{p,\ell,L}} d_\infty(\tau, T^n_{p,\ell,L}) = 0.$$

We record the results of this section in the following corollary.

Corollary 2.23. *Being* not *(p,ℓ,L)-list-recoverable is a symmetric property. Furthermore, $T_{p,\ell,L}$ is a convex approximation for this property.*

3 Characterization theorem

In this section, we prove our main characterization theorem, Theorem 1.1, which is formally stated below as Theorem 3.6. Before stating and proving the theorem, we record a few useful lemmas.

Lemma 3.1 ([CS04, Lemma 2.2])**.** *Let $\tau \in \mathcal{T}^n_b$. Then,*

$$q^{H(\tau)n} \cdot n^{-O_{q,b}(1)} \le |M_\tau| \le q^{H(\tau)n}.$$

Lemma 3.2. *Let $M \subseteq \Sigma^{n\times b}$. Then,*

$$|M| \le (n+1)^{q^b} \cdot q^{n \cdot \max_{B\in M} H(\tau_B)}.$$

Proof. Let $T = \{\tau_B \mid B \in M\}$. Note that

$$M \subseteq \bigcup_{\tau\in T} M_\tau.$$

Thus,

$$\begin{aligned} |M| &\le \sum_{\tau\in T} |M_\tau| \\ &\le |T| \cdot \max_{\tau\in T} |M_\tau| \\ &\le |\mathcal{T}_{n,b}| \cdot \max_{\tau\in T} |M_\tau|. \end{aligned}$$

The claim follows from (2) and Lemma 3.1. □

We say that a type is a *histogram type* if it is indifferent to the ordering of a given vector's entries, and thus, only cares about the histogram of the vector. Formally, we make the following definition.

Definition 3.3 (Histogram type)**.** *A type $\tau \in T_b$ is called a* histogram-type *if $\tau(u) = \tau(\pi(u))$ for every $u \in \Sigma^b$ and every permutation $\pi : [b] \to [b]$.*

Lemma 3.4. *Let $T \subseteq \mathcal{T}_b$ be a closed, permutation-closed, convex, set of types. Then there exists a histogram type $\tau \in T$ such that $H(\tau) = \max_{\tau' \in T} H(\tau')$.*

Proof. Since T is closed and bounded, it is compact. Thus, there is some $\tau' \in T$ such that $H(\tau')$ is maximal. Given a permutation $\pi : [b] \to [b]$, let $\pi(\tau')$ denote the distribution of the vector $\pi(u)$, where $u \sim \tau$. Let

$$\tau = \frac{\sum_{\pi \in \mathrm{Sym}_b} \pi(\tau')}{b!}.$$

Since T is permutation-closed and convex, $\tau \in T$. By concavity of entropy,

$$\begin{aligned} H(\tau) &\geq \frac{\sum_{\pi \in \mathrm{Sym}_b} H(\pi(\tau'))}{b!} \\ &= \frac{\sum_{\pi \in \mathrm{Sym}_k} H(\tau')}{b!} \\ &= H(\tau'). \end{aligned}$$

Thus, τ has maximum entropy in T, and is clearly a histogram-type. □

The following technical lemma facilitates our use of an approximation for a set of types.

Lemma 3.5. *Let $\tau, \tau' \in T_b$ such that $d_\infty(\tau, \tau') \leq \varepsilon$. Then,*

$$|H_{u \sim \tau}(u \mid u_I) - H_{u \sim \tau'}(u \mid u_I)| \leq O_{b,q}\left(\varepsilon \cdot \log \frac{1}{\varepsilon}\right)$$

for any $I \subseteq [b]$.

Proof. Given $x \in \Sigma^I$ and a distribution τ over Σ^b, write $\tau_I(x) = \Pr_{u \sim \tau}[u_I = x]$. By our assumption,

$$\begin{aligned} |\tau_I(x) - \tau'_I(x)| &= \left| \sum_{\substack{y \in \Sigma^b \\ y_I = x}} (\tau_I(y) - \tau'_I(y)) \right| \\ &\leq \sum_{\substack{y \in \Sigma^b \\ y_I = x}} \left| \tau_I(y) - \tau'_I(y) \right| \\ &\leq |\Sigma^{[b] \setminus I}| \cdot \varepsilon = q^{b - |I|} \cdot \varepsilon \end{aligned} \tag{5}$$

for all $I \subseteq [b]$, $x \in \Sigma^I$.

We also need the following fact [Zha07, Eq. (4)]: Let θ and θ' be two probability distributions on a set of N elements, such that their ℓ_∞ distance is at most δ. Then,

$$H(\theta) - H(\theta') \leq \log_q 2 \cdot h_2(2N\delta) + 2N\delta \log_q N. \tag{6}$$

Now,

$$\begin{aligned}
&|H_{u\sim\tau}(u \mid u_I) - H_{u\sim\tau'}(u \mid u_I)| \\
&= \Bigg| \sum_{x\in\Sigma^I} \tau_I(x) H_{u\sim\tau}(u \mid u_I = x) \\
&\quad - \tau'_I(x) H_{u\sim\tau'}(u \mid u_I = x) \Bigg| \\
&\leq \sum_{x\in\Sigma^I} |\tau_I(x) H_{u\sim\tau}(u \mid u_I = x) \\
&\quad - \tau'_I(x) H_{u\sim\tau'}(u \mid u_I = x)| \qquad (7)
\end{aligned}$$

We turn to bounding each term of this sum. Let $x \in \Sigma^I$ and assume without loss of generality that $\tau_I(x) \leq \tau'_I(x)$. By the triangle inequality,

$$\begin{aligned}
&|\tau_I(x) H_{u\sim\tau}(u \mid u_I = x) - \tau'_I(x) H_{u\sim\tau'}(u \mid u_I = x)| \\
&\leq |\tau_I(x) - \tau'_I(x)| \cdot H_{u\sim\tau'}(u \mid u_I = x) \\
&\quad + \tau_I(x) \cdot |H_{u\sim\tau}(u \mid u_I = x) \\
&\quad - H_{u\sim\tau'}(u \mid u_I = x)|. \qquad (8)
\end{aligned}$$

Due to the fact that entropy can only decrease on conditioning, we can say that $H_{u\sim\tau}(u \mid u_I = x) \leq H_{u\sim\tau}(u) \leq b$. So,

$$\begin{aligned}
&\left|\left(\tau_I(x) - \tau'_I(x)\right) H_{u\sim\tau'}(u \mid u_I = x)\right| \qquad (9) \\
&\leq b \cdot q^{b-|I|} \cdot \varepsilon
\end{aligned}$$

due to (5).

Let θ (resp. θ') denote the distribution of $u \sim \tau$ (resp. $u \sim \tau'$) conditioned on $u_I = x$. We bound the L_∞ distance of θ and θ'. For $u \in \Sigma^b$ such that $u_I = x$, we have

$$\begin{aligned}
&\left|\theta(u) - \theta'(u)\right| = \left|\frac{\tau(u)}{\tau_I(x)} - \frac{\tau'(u)}{\tau'_I(x)}\right| \\
&\leq \left|\frac{\tau(u)}{\tau_I(x)} - \frac{\tau(u)}{\tau'_I(x)}\right| + \left|\frac{\tau(u)}{\tau'_I(x)} - \frac{\tau'(u)}{\tau'_I(x)}\right| \\
&= \tau(u) \cdot \left|\frac{\tau'_I(x) - \tau_I(x)}{\tau_I(x) \cdot \tau'_I(x)}\right| + \frac{1}{\tau'_I(x)} \left|\tau(u) - \tau'(u)\right| \\
&\leq \tau(u) \cdot \frac{q^{b-|I|} \cdot \varepsilon}{\tau_I(x) \cdot \tau'_I(x)} + \frac{\varepsilon}{\tau'_I(x)} \\
&\leq \frac{q^{b-|I|} + 1}{\tau'_I(x)} \cdot \varepsilon,
\end{aligned}$$

where the inequalities follow respectively from (5), and from the fact that $\tau(u) \leq \tau_I(x)$. Hence,

$$\begin{aligned}
&\tau_I(x) \left| H_{u\sim\tau}(u \mid u_I = x) - H_{u\sim\tau'}(u \mid u_I = x) \right| \\
&= \tau_I(x) \left| H(\theta) - H(\theta') \right| \\
&\leq \tau_I(x) \left(\log_q 2 \cdot h\left(2\frac{q^b + q^{|I|}}{\tau'_I(x)} \cdot \varepsilon \right) + 2q^{|I|}|I|\varepsilon \right) \\
&\leq \tau'_I(x) \left(\log_q 2 \cdot h\left(2\frac{q^b + q^{|I|}}{\tau'_I(x)} \cdot \varepsilon \right) + 2q^{|I|}|I|\varepsilon \right) \\
&\leq \tau'_I(x) \left(\log_q 2 \cdot h\left(2\frac{2q^b \cdot \varepsilon}{\tau'_I(x)} \right) + 2q^b b\varepsilon \right) \\
&\leq O_{b,q}\left(\varepsilon \log_q \frac{1}{\varepsilon} \right) \qquad (10)
\end{aligned}$$

due to (6). The lemma follows from (7), (8), (9) and (10). □

We now prove that every monotone-increasing, local and row-symmetric property with a convex approximation is sharp for random codes. Furthermore, we identify the threshold rate as the maximal entropy in the approximating set.

Theorem 3.6. *Fix $b \in \mathbb{N}$. Let $\mathcal{P} = \{P_n\}_{n\in\mathbb{N}}$ be a symmetric property with locality parameter b, and let T be a convex approximation for $\mathcal{P}$. Denote $R^* = 1 - \frac{\max_{\tau\in T} H(\tau)}{b}$. Fix $\varepsilon > 0$ and let $R \in [0,1]$. The following now holds.*

1. *If $R \leq R^* - \varepsilon$ then*
$$\lim_{n\to\infty} \Pr\left[C^n_{\mathrm{RC}}(R) \text{ satisfies } \mathcal{P}\right] = 0.$$
2. *If $R \geq R^* + \varepsilon$ then*
$$\lim_{n\to\infty} \Pr\left[C^n_{\mathrm{RC}}(R) \text{ satisfies } \mathcal{P}\right] = 1.$$

Proof. For $b \in \mathbb{N}$ and a matrix $B \in \Sigma^{b\times n}_{\mathrm{distinct}}$, let X_B be an indicator variable for the event that $B \in C^n_{\mathrm{RC}}(R)$. For a set $M \subseteq \Sigma^{b\times n}_{\mathrm{distinct}}$, let $X_M = \sum_{B\in M} X_B$. By Fact 2.4,

$$\mathbb{E}\left[X_M\right] = |M| \cdot q^{-n(1-R)b}. \qquad (11)$$

Let M_n denote the minimal set for P_n and let $T_n = \{\tau_B \mid B \in M_n\}$.

The first statement now follows from Markov's inequality, (11), and Lemma 3.2:

$$\begin{aligned}
&\Pr\left[C \text{ satisfies } \mathcal{P}\right] \\
&= \Pr\left[\exists B \in M_n \;\; B \subseteq C^n_{\mathrm{RC}}(R)\right] \\
&\leq \Pr\left[X_M \geq 1\right] \\
&\leq \mathbb{E}\left[X_M\right] \\
&= |M| \cdot q^{-n(1-R)b} \\
&\leq (n+1)^{q^b} \cdot q^{n\cdot\max_{\tau\in T_n} H(\tau)} \cdot q^{-n(1-R)b} \\
&\leq (n+1)^{q^b} \cdot q^{n\cdot\max_{\tau\in T} H(\tau)} \cdot q^{-n(1-R)b} \\
&\leq (n+1)^{q^b} \cdot q^{-nb\varepsilon} \leq e^{-\Omega(n)}.
\end{aligned}$$

Above, we used the fact that $T_n \subseteq T$.

For the second statement, let $\tau \in T$ have maximum entropy. By definition 2.18, T is closed and permutation-closed, in addition to being convex. Consequently, due to Lemma 3.4, we may assume that τ is a histogram-type. Let $\tau_n \in T_n$ such that $d_\infty(\tau, \tau_n) = o_{n\to\infty}(1)$. Our plan is to use a second-moment argument to show that $C^n_{\mathrm{RC}}(R)$ likely contains a matrix of type τ_n.

By (11) and Lemma 3.1,

$$\begin{aligned}\mathbb{E}\left[X_{M_{\tau_n}}\right] &= |M_{\tau_n}| q^{-n(1-R)b} \\ &\geq q^{(H(\tau_n)-(1-R)b)n+o(n)} \\ &\geq q^{(H(\tau)-(1-R)b)n+o(n)}\end{aligned}$$

We turn to bounding the variance of $X_{M_{\tau_n}}$. Fact 2.4 yields

$$\begin{aligned}&\mathrm{Var}\left[X_{M_{\tau_n}}\right] \\ &= \sum_{B,B' \in M_{\tau_n}} \left(\Pr\left[X_B = X_{B'} = 1\right]\right. \\ &\quad \left. - \Pr\left[X_B = 1\right]\Pr\left[X_{B'} = 1\right]\right) \\ &= \sum_{B,B' \in M_{\tau_n}} \left(q^{-n(1-R)(2b-\alpha(B,B'))} - q^{-2n(1-R)b}\right) \\ &\leq \sum_{\substack{B,B' \in M_{\tau_n} \\ \alpha(B,B') \geq 1}} q^{-n(1-R)(2b-\alpha(B,B'))}\end{aligned}$$

where $\alpha(B, B')$ is the number of columns in B' that also appear in B.

In order to bound this sum, we need an estimate on the number of pairs B, B' with a given $\alpha(B, B')$. For $0 \leq r \leq b$, let

$$W_r = \{(B, B') \mid B, B' \in M_{\tau_n} \text{ and } \alpha(B, B') = r\}$$

and denote $S_r = \{\tau_{B\|B'} \mid (B, B') \in W_r\}$. Here, $B\|B'$ is the $n \times 2b$ matrix whose first (resp. last) b columns are B (resp. B'). By Lemma 3.2,

$$|W_r| \leq (n+1)^{2q^b} \cdot q^{n \max_{\nu \in S_r} H(\nu)}.$$

Let $(B, B') \in W_r$ and let $\nu = \tau_{B\|B'}$. Assume without loss of generality that the first r columns of B are identical to the first r columns of B'. Let $u \sim \nu$. Note that, since $B, B' \in M_{\tau_n}$, the random variables $u_{[b]}$ and $u_{[2b]\setminus[b]}$ are both τ_n-distributed. Hence,

$$\begin{aligned}H(\nu) = H(u) &= H(u_{[2b]\setminus[b]}) + H(u_{[b]} \mid u_{[2b]\setminus[b]}) \\ &= H(\tau_n) + H(u_{[b]} \mid u_{[2b]\setminus[b]}) \\ &\leq H(\tau_n) + H(u_{[b]} \mid u_{[r]}) \\ &= H(\tau_n) + H(u_{[b]\setminus[r]} \mid u_{[r]}).\end{aligned}$$

Note that the inequality follows from the observation that $H(u_{[b]} \mid u_{[2b]\setminus[b]}) \leq H(u_{[b]} \mid u_{[r]})$ since entropy can only decrease on conditioning.

Lemma 3.5 yields

$$\begin{aligned} H(u_{[b]\setminus[r]} \mid u_{[r]}) &\le H_{v\sim\tau}(v_{[b]\setminus[r]} \mid v_{[r]}) + o(1) \\ &= \sum_{i=r+1}^{b} H_{v\sim\tau}(v_i \mid v_{[i-1]}) + o(1) \\ &= \sum_{i=r+1}^{b} H_{v\sim\tau}(v_b \mid v_{[i-1]}) + o(1), \end{aligned}$$

where the last equality is due to τ being a histogram-type. Writing

$$f(r) = \sum_{i=r+1}^{b} H_{v\sim\tau}(v_b \mid v_{[i-1]}),$$

we conclude that

$$H(\nu) \le f(r) + H(\tau) + o(1),$$

so that

$$|W_r| \le q^{(f(r)+H(\tau))n+o(n)},$$

and

$$\begin{aligned} &\mathrm{Var}\left[X_{M_{\tau_n}}\right] \\ &\le \sum_{r=1}^{b} |W_r| \cdot q^{-n(1-R)(2b-r)} \\ &\le \sum_{r=1}^{b} q^{(f(r)+H(\tau)-(1-R)(2b-r))n+o(n)} \\ &\le \max_{1\le r\le b} q^{(f(r)+H(\tau)-(1-R)(2b-r))n+o(n)} \end{aligned}$$

By Chebyshev's inequality,

$$\begin{aligned} &\Pr\left[X_{M_{\tau_n}} = 0\right] \\ &\le \frac{\mathrm{Var}\left[X_{M_{\tau_n}}\right]}{\mathbb{E}\left[X_{M_{\tau_n}}\right]^2} \\ &\le \max_{1\le r\le b} q^{(f(r)-H(\tau)+r(1-R))n+o_{b,q}(n)}. \end{aligned} \tag{12}$$

We claim that $(f(r))_{r=0}^{b}$ is a convex sequence. Indeed,

$$\begin{aligned} &f(r-1) + f(r+1) - 2f(r) \\ &= H_{v\sim\tau}(v_b \mid v_{[r-1]}) - H_{v\sim\tau}(v_b \mid v_{[r]}) \\ &\ge 0. \end{aligned}$$

Therefore, the maximum in the right-hand side of (12) is achieved either by $r = 1$ or $r = b$. In the former case, note that

$$\begin{aligned} f(1) &= \sum_{i=2}^{b} H_{v\sim\tau}(v_b \mid v_{[i-1]}) \\ &= \sum_{i=2}^{b} H_{v\sim\tau}(v_i \mid v_{[i-1]}) = H_{v\sim\tau}(v \mid v_1) \\ &\le H(\tau) - H_{v\sim\tau}(v_1) \le H(\tau) \cdot \frac{b-1}{b}. \end{aligned}$$

In the last inequality above, we used the fact that $H_{v\sim\tau}v_1 = H_{v\sim\tau}v_i$ for all $i \in [b]$, due to τ being a histogram-type. Thus, for $r = 1$, the corresponding exponent in (12) is

$$\begin{aligned} &(f(1) - H(\tau) + (1-R))n \\ &\le \left((1-R) - \frac{H(\tau)}{b}\right) n \\ &\le -\varepsilon n. \end{aligned}$$

In the latter case, since $f(b) = 0$, the exponent is

$$(-H(\tau) + (1-R)b)n \le -\varepsilon bn.$$

We conclude that

$$\begin{aligned} \Pr\left[C^n_{\mathrm{RC}}(R) \text{ does not satisfy } \mathcal{P}\right] &\le \Pr(X_{M_\rho} = 0) \\ &\le q^{-\varepsilon n + o(n)}. \end{aligned}$$

□

Applying the framework to list-recovery. In the rest of the paper, we use Theorem 3.6 to compute the threshold rate for (p, ℓ, L) list-recovery in several different settings. In order to do that, we set up a few useful definitions.

Definition 3.7 ($\beta(p,\ell,L)$ and $\bar{T}_{p,\ell,L}$)**.** *Given $L \in \mathbb{N}$, $\ell \le L$ and $p \in [0,1)$, let $\bar{T}_{p,\ell,L}$ denote the set of all histogram-types in $T_{p,\ell,L}$. Let*

$$\beta(p,\ell,L) = \max_{\tau \in \bar{T}_{p,\ell,L}} H(\tau).$$

Theorem 3.6 allows us to characterize the threshold rate for (p,ℓ,L)-list recovery in terms of $\beta(p,\ell,L)$:

Corollary 3.8. *Fix $L \in \mathbb{N}$, $\ell \le L$ and $p \in [0,1)$. The threshold rate for (p,ℓ,L) list-recovery is*

$$R^* = 1 - \frac{\beta(p,\ell,L)}{L}.$$

Proof. By Corollary 2.23 and Lemma 3.4,

$$\beta(p,\ell,L) = \max_{\tau \in T_{p,\ell,L}} H(\tau).$$

The claim now follows from Corollary 2.23 and Theorem 3.6. □

Finally, we introduce the following notation, which will be used for the rest of the paper.

Definition 3.9 ($P_\ell(\cdot)$ and $D_{d,\ell,L}$)**.** *Fix $\ell \le L$. Given a vector $v \in \Sigma^L$ let*

$$P_\ell(v) = \min_{A \in \binom{\Sigma}{\ell}} |\{i \in [L] \mid v_i \notin A\}|$$

We use the notation $D_{d,\ell,L} = \{v \in \Sigma^L \mid P_\ell(v) = d\}$.

4 Bounds on the threshold rate for noisy list-recovery

The main result in this section is an estimate of $\beta(p, \ell, L)$ (Proposition 4.4 below), which leads to an estimate on the threshold rate for list-recovery (Corollary 4.5). This estimate is very sharp when $\frac{q \log L}{L}$ is small; in subsequent sections we will derive estimates which are more precise for certain parameter regimes.

Before coming to these bounds, we begin with a few useful lemmas that bound $|D_{d,\ell,L}|$ and characterize $\bar{T}_{p,\ell,L}$.

Lemma 4.1. *Let $r = 1 - \frac{\ell}{q}$ and $s = \frac{d}{L}$. Suppose that $s < r$. Then,*

$$\begin{aligned}
&\binom{q}{rq}\binom{L}{sL}\Bigg(\underbrace{\frac{(1-s)L}{(1-r)q}, \dots \frac{(1-s)L}{(1-r)q}}_{\ell}\Bigg) \\
&\cdot \Bigg(\underbrace{\frac{sL}{rq}, \dots \frac{sL}{rq}}_{q-\ell}\Bigg) \le |D_{d,\ell,L}| \\
&\le \binom{q}{rq}\left(\sum_{i=0}^{sL}\binom{L}{i}((1-r)q)^{L-i}(rq)^i\right).
\end{aligned}$$

Proof. For the upper bound, note that a vector $v \in D_{d,\ell,L}$ is uniquely determined by the combination of the following:

- The set $A \in \binom{\Sigma}{\ell}$, consisting of the ℓ most common entries in v (with some arbitrary tie-breaking rule). There are $\binom{q}{rq}$ ways to choose this set.
- An assignment of elements from Σ to v_i ($i \in [L]$) so that at most sL of the entries do not belong to A. There are $\sum_{i=0}^{sL}\binom{L}{i} \cdot ((1-r)q)^{L-i} \cdot (rq)^i$ such assignments.

The upper bound is obtained by multiplying these two counts.

For the lower bound, let M denote the collection of all vectors $v \in \Sigma^L$ for which there exists a set $A_v \in \binom{\Sigma}{\ell}$ such that:

- Each element of A_v appears in v exactly $\frac{L-d}{\ell}$ times.

- Each element of $\Sigma \setminus A_v$ appears in v exactly $\frac{d}{q-\ell}$ times.

Since we know that $r > s$, we can say that $L > \frac{dq}{q-\ell} = \frac{dl}{q-\ell} + \frac{d(q-\ell)}{q-\ell}$ and therefore $\frac{L-d}{\ell} > \frac{d}{q-\ell}$. From this we may conclude that $M \subseteq D_{d,\ell,L}$. So,

$$\begin{aligned}
|D_{d,\ell,L}| &\geq |M| \\
&\geq \binom{q}{rq}\binom{L}{sL}\Bigg(\underbrace{\frac{(1-s)L}{(1-r)q}, \ldots, \frac{(1-s)L}{(1-r)q}}_{\ell}\Bigg) \\
&\cdot \Bigg(\underbrace{\frac{sL}{rq}, \ldots, \frac{sL}{rq}}_{q-\ell}\Bigg).
\end{aligned}$$

□

Using Stirling's approximation, Lemma 4.1 immediately yields the following.

Corollary 4.2. *In the setting of Lemma 4.1, suppose that $s < r$. Then,*

$$\log_q |D_{d,\ell,L}| = L(1 - D_{\mathrm{KL}_q}(s \parallel r)) \pm O(q \log L),$$

where the underlying constant is universal.

In order to compute $\beta(p, \ell, L)$, we will make use of the following characterization of $\bar{T}_{p,\ell,L}$ (Definition 3.7). Intuitively, this lemma says that a histogram-type τ is bad for (p, ℓ, L)-list-recovery if and only if it has many symbols inside the most frequent ℓ symbols in expectation.

Lemma 4.3. *Let $1 \leq \ell \leq q$, $L \in \mathbb{N}$ and $0 \leq p \leq 1$. Let τ be a distribution over Σ^L and suppose that τ is a histogram-type. Then, $\tau \in \bar{T}_{p,\ell,L}$ if and only if*

$$\mathop{\mathbb{E}}_{u \sim \tau}[P_\ell(u)] \leq pL. \tag{13}$$

Proof. Suppose that $\tau \in \bar{T}_{p,\ell,L}$. A fortiori, $\tau \in T_{p,\ell,L}$. Let ρ be as in Definition 2.20. Then,

$$\mathop{\mathbb{E}}_{v \sim \tau}[P_\ell(u)] \leq \mathop{\mathbb{E}}_{(u,K) \sim \rho}[|\{i \in [L] \mid u_i \notin K\}|] \leq pL,$$

where the first inequality is due to Condition 1 of Definition 2.20, and the second inequality follows from (4).

Conversely, suppose that (13) holds. Let ρ be the distribution of the pair (u, K), where u is first sampled from τ, and let $K = K(u) \in \binom{\Sigma}{\ell}$ be the lexicographically minimal set for which $|\{i \in [L] \mid u_i \notin K\}| = P_\ell(v)$. Clearly, ρ satisfies Condition 1 of Definition 2.20.

We next show that it also satisfies Condition 2. Observe that $K(u) = K(\pi(u))$ for every permutation π over $[L]$. Hence, since τ is a histogram property,

$$\Pr_{(u,K) \sim \rho}[u_i \notin K] = \Pr_{(u,K) \sim \rho}[u_1 \notin K]$$

for every $i \in [L]$. Consequently, for any $i \in [L]$,

$$\begin{aligned}\Pr_{(u,K)\sim\rho}[u_i \notin K] &= \frac{1}{L}\mathop{\mathbb{E}}_{(u,K)\sim\rho}[|\{i \in [L] \mid u_i \notin K\}|] \\ &= \frac{1}{L}\mathop{\mathbb{E}}_{v\sim\tau}[P_\ell(v)] \le p.\end{aligned}$$

It follows that $\tau \in T_{p,\ell,L}$. Since τ is a histogram-type, it also belongs to $\bar{T}_{p,\ell,L}$. □

Now, we come to our estimate on the threshold rate for (p, ℓ, L) list-recovery in the regime where $L \to \infty$ and $q \le o(\frac{\log L}{L})$. We begin with the following proposition, which bounds the quantity $\beta(p, \ell, L)$.

Proposition 4.4. *Let $r = 1 - \frac{\ell}{q}$ and suppose that $p \le r$. Then,*

$$\beta(p, \ell, L) = L(1 - D_{\mathrm{KL}_q}(p \parallel r)) \pm O(q \log L).$$

Proof. We first bound $\beta(p, \ell, L)$ from below. Let τ denote the uniform distribution on $D_{pL,\ell,L}$. This distribution clearly belongs to $\bar{T}_{p,\ell,L}$, so Corollary 4.2 yields

$$\begin{aligned}\beta(p, \ell, L) &\ge H(\tau) \\ &\ge L(1 - D_{\mathrm{KL}_q}(p \parallel r)) \pm O(q \log L).\end{aligned}$$

We turn to proving a matching upper bound. Let $\tau \in \bar{T}_{p,\ell,L}$ such that $\beta(p, \ell, L) = H(\tau)$. Let $v \sim \tau$ and denote the distribution of $P_\ell(v)$ by μ. Note that

$$\begin{aligned}H(\tau) &= H(\mu) + H(v \mid P_\ell(v)) \\ &\le H(v \mid P_\ell(v)) + \log_q L,\end{aligned}$$

since $P_\ell(v)$ has at most L possible values. Now,

$$\begin{aligned}H(v \mid P_\ell(v)) &= \mathbb{E}_{d\sim\mu}[H(v \mid P(v) = d)] \\ &\le \mathbb{E}_{d\sim\mu}\left[\log_q |D_{d,\ell,L}|\right].\end{aligned} \tag{14}$$

Let

$$f(d) = \begin{cases} L\left(1 - D_{\mathrm{KL}_q}\left(\frac{d}{L} \parallel r\right)\right) & \text{if } \frac{d}{L} < r \\ L & \text{if } \frac{d}{L} \ge r. \end{cases}$$

By (14), Corollary 4.2, and the concavity of f,

$$\begin{aligned}H(v \mid P_\ell(v)) &\le \mathbb{E}_{d\sim\mu}[f(d)] + O(q \log L) \\ &\le f(\mathbb{E}_{d\sim\mu}[d]) + O(q \log L).\end{aligned}$$

By (4), we have

$$\mathop{\mathbb{E}}_{d\sim\mu}[d] \le p.$$

Since f is non-decreasing, it follows that

$$\begin{aligned}H(v \mid P_\ell(v)) &\le f(p) + O(q \log L) \\ &= L(1 - D_{\mathrm{KL}_q}(p \parallel r)) + O(q \log L),\end{aligned}$$

establishing the upper bound. □

Corollary 4.5. *The threshold rate for (p, ℓ, L) list-recovery of a random code is*

$$R^* = \begin{cases} D_{\mathrm{KL}q}\left(p \parallel r\right) \pm O\left(\frac{q \log L}{L}\right) & \text{if } p < r \\ 0 & \text{if } p \geq r, \end{cases}$$

where $r = 1 - \frac{\ell}{q}$.

Remark 4.6. *In order to better illustrate the threshold rate computed in Corollary 4.5, one can verify the identity*

$$\begin{aligned} &D_{\mathrm{KL}q}\left(p \parallel 1 - \ell/q\right) \\ &= 1 - p \log_q \left(\frac{q-\ell}{p}\right) - (1-p) \log_q \left(\frac{\ell}{1-p}\right). \end{aligned}$$

Substituting $\ell = 1$, we find $D_{\mathrm{KL}q}\left(p \parallel 1 - 1/q\right) = 1 - h_q(p)$, agreeing with the list decoding capacity theorem. For larger ℓ, this expression agrees with the list-recovery capacity theorem*, as stated in e.g. [Res20].*

Proof. The case $p < r$ is immediate from Corollary 3.8 and Proposition 4.4.

For $p \geq r$, let τ denote the uniform distribution on Σ^L. We claim that $\tau \in \bar{T}_{p,\ell,L}$. Since τ is clearly a histogram-type, it suffices to show that it satisfies (13). Fix some arbitrary set $A \in \binom{\Sigma}{\ell}$. Then,

$$\begin{aligned} \mathop{\mathbb{E}}_{u \sim \tau}\left[P_\ell(u)\right] &\leq \mathop{\mathbb{E}}_{u \sim \tau}\left[\left|\{i \in [L] \mid u_i \notin A\}\right|\right] \\ &\leq L \cdot \left(1 - \frac{\ell}{q}\right) = Lr \leq Lp, \end{aligned}$$

and the claim follows.

Corollary 3.8 now yields

$$R^* = 1 - \frac{\beta(p, \ell, L)}{L} \leq 1 - \frac{H(\tau)}{L} = 0. \quad \square$$

5 Zero-error list-recovery and perfect hashing codes

In this section we analyze the threshold rate for zero-error list-recovery (that is, when $p = 0$), and give a more precise version of Corollary 4.5 in this setting. We use this to compute the threshold rate for a random code be to a perfect hash code, which is the same as being $(0, q-1, q)$ list-recoverable.

Lemma 5.1. *Let $p^* = |D_{0,\ell,L}|/q^L$. The threshold rate for $(0, \ell, L)$ list-recovery of a random code is*

$$R^* = \frac{-\log_q(p^*)}{L}$$

Proof. Let μ be the uniform distribution on $D_{0,\ell,L}$. Note that $\mu \in T_{0,\ell,L}$, so

$$\beta(0, \ell, L) \geq H(\mu) = \log_q(|D_{0,\ell,L}|) = \log_q(p^*) + L. \tag{15}$$

On the other hand, due to (4), every distribution in $T_{0,\ell,L}$ is supported in $D_{0,\ell,L}$. Consequently, (15) is in fact an equality (since $\log_q(|D_{0,\ell,L}|)$ is simply the entropy of the uniform distribution, which is the maximal entropy distribution on $D_{0,\ell,L}$). It now follows from Corollary 3.8 that

$$R^* = 1 - \frac{\beta(0,\ell,L)}{L}. \quad \square$$

Corollary 5.2. *The threshold rate for $(0, q-1, q)$ list-recovery of a random code is*

$$R^* = \frac{1}{q}\log_q\left(\frac{1}{1-q!/q^q}\right)$$

Proof. Due to Lemma 5.1 it suffices to show that

$$|D_{0,q-1,q}| = q^q - q!.$$

Indeed, $|D_{0,q-1,q}|$ counts the number of vectors of length q, over an alphabet of size q such that not all q distinct letters show up in the vector. This number is exactly all possible vectors after removing the permutations of $(1, 2, \ldots, q)$, namely, $q^q - q!$. $\square$

6 List of two decoding of random and random linear codes

In this section, we study the list-of-2 decodability of two random ensembles of codes. In detail, we precisely compute the threshold rate for $(p, 3)$-list-decoding for random codes and for random *linear* codes. Denote by $\mathcal{P}$ the monotone increasing property of *not* being $(p, 3)$-list-decodable. Note that we cannot immediately apply Corollary 4.5, as the error term of $O\left(\frac{q \log L}{L}\right)$ is not negligible in this regime. We specialize to the case of $q = 2$, and recall our convention that log denotes the base-2 logarithm. Recall from the introduction that whenever $p < 1/4$ there exist $(p, 3)$-list-decodable codes with positive rate, but whenever $p > 1/4$ the only $(p, 3)$-list-decodable codes are of bounded size, independent of n.

Our main result of this section is a demonstration that the list-of-2 decoding threshold rate for random *linear* codes is in fact greater than the corresponding threshold rate for random codes. This result demonstrates that our techniques are precise enough to allow us to sharply delineate between different natural ensembles of codes.

In the following, $C^n_{\mathrm{RLC}}(R)$ denotes a random linear code of block length n and rate R. We define the threshold rate for random linear codes, denoted $R^n_{\mathrm{RLC}}(\mathcal{P})$, in a manner analogous to the definition for random codes. It is,

$$\sup\left\{R \in [0,1] \,:\, \Pr\left[C^n_{\mathrm{RLC}}(R) \text{ satisfies } \mathcal{P}\right] \leq \tfrac{1}{2}\right\}$$

if such an R exists, and 0 otherwise.

Theorem 6.1. *Let $p \in (0, 1/4)$.*

1. *The threshold rate for $(p, 3)$-list-decoding for random codes satisfies*

$$\lim_{n\to\infty} R^n_{\mathrm{RC}}(\mathcal{P}) = 1 - \frac{1 + h(3p) + 3p\log 3}{3}.$$

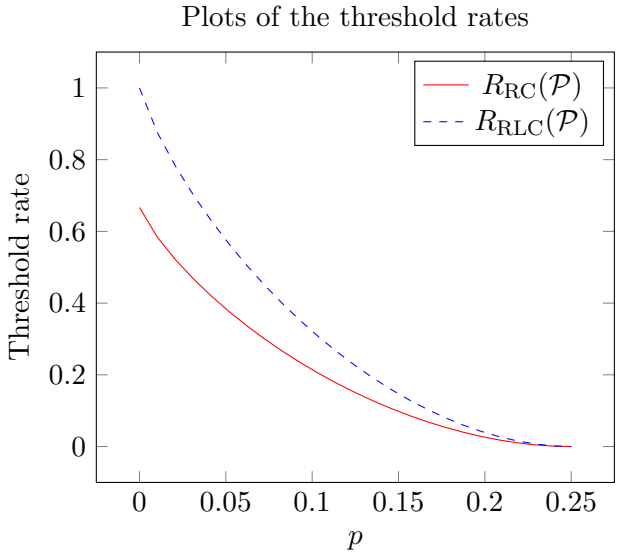

Figure 1: The threshold rate R_{RC} (red) for $(p, 3)$-list-decodability of random codes, and the threshold rate R_{RLC} (blue, dashed) for $(p, 3)$-list-decodability of random *linear* codes. Note that, uniformly over p, random linear codes have the greater threshold rate.

2. *The threshold rate for* $(p, 3)$*-list-decoding for random* linear *codes satisfies*

$$\lim_{n\to\infty} R^n_{\mathrm{RLC}}(\mathcal{P}) = 1 - \frac{h(3p) + 3p\log 3}{2}.$$

Note that the threshold rate for random linear codes is greater than the threshold rate for random codes, uniformly over $p \in (0, 1/4)$. See Figure 1.

The proof of Theorem 6.1 is split into two parts. Part 1, which computes the threshold rate for random codes, is proved in the remainder of this subsection. The proof of Part 2 concerning the threshold rate for random linear codes is deferred to Appendix C, as it requires certain tools and techniques developed in other works ([MRRZ+19, Res20]) that we have not yet introduced in this paper.

Proof of Theorem 6.1, Part 1. By Corollary 3.8, to compute the threshold rate, it suffices to show

$$\beta(p, 1, 3) = \max_{\tau\in\bar{T}_{p,1,3}} H(\mu) = 1 + h_2(3p) + 3p\log 3.$$

Once we have done this, Corollary 3.8 will tell us that

$$1 - \frac{\beta(p, 1, 3)}{3} = 1 - \frac{1 + h(3p) + 3p\log 3}{3},$$

as required. We will also use the characterization of $\bar{T}_{p,1,3}$ provided by Lemma 4.3: namely, $\tau \in \bar{T}_{p,1,3}$ if and only if $\mathbb{E}_{u\sim\tau}[P_1(u)] \le 3p$. We first prove an upper bound on $\beta(p, 1, 3)$, followed by a matching lower bound.

Upper bound on $\beta(p,1,3)$ Let $\tau \in \bar{T}_{p,1,3}$. Let $A = \{(0,0,0),(1,1,1)\}$ and let $B = \mathbb{F}_2^3 \setminus A$. Define $x := \mu(B) = \Pr_{v\sim\tau}[v \in B]$, and note that $\mathbb{E}_{v\sim\tau}[P_1(v)] = x$. Indeed, $P_1(v) = 1$ if $v \in B$ and $P_1(v) = 0$ if $v \in A$. Thus, we deduce $x \le 3p$. Since $|A| = 2$ and $|B| = 6$, we may upper bound

$$\begin{aligned} H(\tau) &\le x \log \frac{6}{x} + (1-x)\log\frac{2}{1-x} \\ &= 1 + h(x) + x\log 3. \end{aligned}$$

As $x \le 3p < \frac{3}{4}$, one can see that the quantity $1 + h(x) + x\log 3$ is increasing in x: indeed, defining $g : [0, 1/4] \to \mathbb{R}$ by $x \mapsto 1 + h(x) + x\log 3$, note that $g'(x) = \log\left(\frac{3(1-x)}{x}\right)$ is positive if and only if $3(1-x) < x$, which is satisfied precisely when $x \in (0, \frac{3}{4})$. Therefore, $H(\tau) \le 1 + h(3p) + 3p\log 3$. Thus,

$$\begin{aligned} \beta(p,1,3) &= \max_{\tau\in\bar{T}_{p,1,3}} H(\tau) \\ &\le 1 + h(3p) + 3p\log 3. \end{aligned}$$

Lower bound on $\beta(p,1,3)$ Now, define the distribution τ^* by $\tau^*(v) = \frac{1-3p}{2}$ for $v \in A$ and $\tau^*(v) = \frac{3p}{6} = \frac{p}{2}$ for $v \in B$. Reasoning as above, $\mathbb{E}_{v\sim\tau^*}[P_1(v)] = |B| \cdot \frac{3p}{6} = 3p$, and moreover it is clear that $\tau(\pi(v)) = \tau^*(v)$ for all permutations $\pi : [3] \to [3]$ and $v \in \mathbb{F}_2^3$. Thus, $\tau^* \in \bar{T}_{p,1,3}$. Furthermore

$$\begin{aligned} H(\tau^*) &= |B| \cdot \frac{p}{2}\log\frac{2}{p} + |A| \cdot \frac{1-3p}{2}\log\frac{2}{1-3p} \\ &= 1 + h(3p) + 3p\log 3, \end{aligned}$$

so

$$\begin{aligned} \beta(p,1,3) &= \max_{\tau\in\bar{T}_{p,1,3}} H(\tau) \ge H(\tau^*) \\ &= 1 + h(3p) + 3p\log 3. \end{aligned}$$

□

7 Computing the threshold rate for list-recovery efficiently

In the previous sections, we gave precise analytical expressions for the threshold rate for list-recovery in certain parameter regimes. However, there are some regimes where these bounds aren't precise. In this section, we consider the question of computing the threshold rate R^* algorithmically, given p, ℓ and L. We use tools from the study of entropy-maximizing distributions to develop a simple binary-search-based procedure to pinpoint R^* up to arbitrarily small additive error.

We begin with a lemma that shows that we can compute the cardinality $|D_{d,\ell,L}|$ efficiently; we will use this as a subroutine in our final algorithm.

Lemma 7.1. *Given $0 \le d \le L$ and $1 \le \ell \le q$, the cardinality $|D_{d,\ell,L}|$ can be computed in time*

$$O\left((L+1)^q + \mathrm{poly}(q, L)\right).$$

Proof. Given a vector $v \in \Sigma^L$, define its *histogram* $\eta_v : \Sigma \to \mathbb{Z}_{\geq 0}$ by

$$\eta_v(\sigma) = |\{i \in [L] \mid v_i = \sigma\}|.$$

Note that the number of vectors in Σ^L with a given histogram η is given by the multinomial

$$\binom{L}{(\eta(\sigma))_{\sigma\in\Sigma}}.$$

Also, a vector v belongs to $D_{d,\ell,L}$ if and only if

$$L - \max_{A\in\binom{\Sigma}{\ell}} \sum_{\sigma\in A} \eta(\sigma) = d. \tag{16}$$

Thus,

$$|D_{d,\ell,L}| = \sum_{\eta} \binom{L}{(\eta(\sigma))_{\sigma\in\Sigma}},$$

where the sum is over all distributions η satisfying (16). In particular, $|D_{d,\ell,L}|$ can be computed by going over all $(L+1)^q$ functions $\eta : \Sigma \to \{0, \ldots, L\}$ and summing the terms corresponding to those functions that satisfy (16). □

We recall the following standard facts from the theory of entropy-maximizing distributions.

Lemma 7.2. *[WJ08, Sec. 3] Let Ω be a finite nonempty set, $f : \Omega \to \mathbb{R}$ and $t \in \mathbb{R}$. Let S_t denote the set of all distributions τ over Ω such that $\mathbb{E}_{\omega\sim\tau}[f(\omega)] = t$. Let*

$$F(t) = \max_{\tau\in S_t} H(\tau).$$

Then

$$F(t) = \inf_{\alpha\in\mathbb{R}} \left[\log_q \left(\sum_{\omega\in\Omega} q^{\alpha\cdot f(\omega)}\right) - \alpha t\right].$$

Furthermore:

1. *If τ is the entropy maximizing distribution, then $\tau(\omega) = \tau(\omega')$ for every $\omega, \omega' \in \Omega$ such that $f(\omega) = f(\omega')$.*
2. *Let $t^* = \mathbb{E}_{\omega\sim\mathrm{Uniform}(\Omega)}[f(\omega)]$. Then, $F(t^*) = \log|\Omega|$, and $F(t)$ is nondecreasing (resp. nonincreasing) in the range $t < t^*$ (resp. $t > t^*$).*
3. *The function*

$$\log_q \left(\sum_{\omega\in\Omega} q^{\alpha\cdot f(\omega)}\right) - \alpha t$$

is convex in α.

Lemma 7.3. *Let $\ell \le q$, $L \in \mathbb{N}$ and $0 < p \le 1$, and let $t^* = q^{-L} \cdot \sum_{d=0}^{L} d \cdot |D_{d,\ell,L}|$. Then, in the case when $p < t^*$,*

$$\beta(p,\ell,L) = \inf_{\alpha \in \mathbb{R}} \left[\log_q \left(\sum_{d=0}^{L} |D_{d,\ell,L}| \cdot q^{\alpha d} \right) - \alpha p L \right]$$

and when $p \ge t^*$,

$$\beta(p,\ell,L) = L.$$

Remark 7.4. *In general, $\frac{t^*}{L}$ is slightly smaller than $1 - \frac{\ell}{q}$. Thus, Lemma 7.3 lemma extends the range in which the threshold is 0 from $\left[1 - \frac{\ell}{q}, 1\right]$ (Corollary 4.5) to $[t^*, 1]$.*

Proof. Suppose that $p \ge t^*$. Let τ be the uniform distribution over Σ^L. Recalling the Definition of $P_\ell(u)$ from Definition 3.9, we have $\mathbb{E}_{u\sim\tau}[P_\ell(u)] = t^* L$. Consequently, Lemma 4.3 implies that $\tau \in \bar{T}_{p,\ell,L}$. Hence, $\beta(p,\ell,L) \ge H(\tau) = L$, and this bound is clearly tight since no distribution over Σ^L has entropy larger than L.

We proceed, assuming that $p < t^*$. Let $S_{t,\ell,L}$ be the set of all distributions τ over Σ^L such that $\mathbb{E}_{u\sim\tau}[P_\ell(u)] = tL$. By Lemma 7.2(1), the distribution τ that has maximal entropy in $S_{t,\ell,L}$, satisfies $\tau(u) = \tau(v)$ whenever $P_\ell(u) = P_\ell(v)$. In particular, τ is invariant to coordinate permutations, so it is a histogram-type. Therefore, using Lemma 4.3 and the definition of $\beta(p,\ell,L)$,

$$\beta(p,\ell,L) = \max_{t \le p} \max_{\tau \in S_t} H(\tau).$$

By Lemma 7.2(2), this expression is maximized by $t = p$, namely

$$\beta(p,\ell,L) = \max_{\tau \in S_p} H(\tau).$$

Consequently, by the main part of Lemma 7.2,

$$\begin{aligned}
&\beta(p,\ell,L) \\
&= \min_{\alpha \in \mathbb{R}} \left[\log_q \left(\sum_{u \in \Sigma^L} q^{\alpha P_\ell(u)} \right) - \alpha p L \right] \\
&= \min_{\alpha \in \mathbb{R}} \left[\log_q \left(\sum_{d=0}^{L} |D_{d,\ell,L}| \cdot q^{\alpha d} \right) - \alpha p L \right].
\end{aligned}$$

□

Theorem 7.5. *There is an algorithm, that, given p, ℓ, L and $\varepsilon > 0$, computes the threshold-rate for (p,ℓ,L)-list-recovery, within an additive error of ε, in time $O\left((L+1)^q + \mathrm{poly}(q, L, \log\frac{1}{\varepsilon}, \beta(p))\right)$, where*

$$\beta(p) = \begin{cases} \log\frac{1}{p} & \text{if } p > 0 \\ 1 & \text{if } p = 0. \end{cases}$$

Proof. Consider the following algorithm:

1. Compute the coefficients $|D_{d,\ell,L}|$ for $0 \le d \le L$.

2. If $p = 0$, return the expression for R^* given in Lemma 5.1.

3. If $p \geq t^*$ (where t^* is as in Lemma 7.3.), return 0.

4. If $p < t^*$, use the bisection method (i.e., binary search) to approximate the minimum M of the function
$$g(\alpha) := \log_q \left(\sum_{d=0}^{L} |D_{d,\ell,L}| \cdot q^{\alpha d} \right) - \alpha p L$$
to within an additive error of εL. Return $1 - \frac{M}{L}$.

The correctness of the algorithm follows from Lemma 7.3. The first step can be completed in time $O\left((L+1)^q + \text{poly}(q, L)\right)$ due to Lemma 7.1.

To analyze the last step, we first note that $g(\alpha)$ is convex, due to Lemma 7.2(3), and thus has a unique minimum. In addition, it is L-Lipshitz. Finally, as we show in Claim 7.6 below, the minimizing α lies in the range $\left[-\left(L + \log_q \frac{1}{p} \right), 0 \right]$. It follows that this step requires at most $O(\log L + \log\log \frac{1}{p} + \log \frac{1}{\varepsilon})$ bisection iterations.

It remains to prove the following claim.

Claim 7.6.
$$\text{argmin}_\alpha g(\alpha) \in \left[-\left(L + \log_q \frac{1}{p} \right), 0 \right].$$

Proof. We first compute
$$\frac{dg}{d\alpha} = \frac{\left(\sum_{d=0}^{L} |D_{d,\ell,L}| \cdot d \cdot q^{\alpha d} \right)}{\left(\sum_{d=0}^{L} |D_{d,\ell,L}| \cdot q^{\alpha d} \right)} - pL.$$

The derivative at 0 is positive due to our assumption that $p < t^*$, which, along with the fact that g is convex, implies that the minimizer is less than zero. It is left to prove that the derivative at $\alpha_0 := -(L + \log_q \frac{1}{p})$ is negative, which will imply that the minimizer is at least $-(L + \log_q \frac{1}{p})$.

To see this, let $Z = \sum_{d=0}^{L} |D_{d,\ell,L}| \cdot q^{\alpha_0 d}$. Then,
$$\begin{aligned} \frac{dg}{d\alpha}_{|\alpha=\alpha_0} &= \frac{\left(\sum_{d=0}^{L} |D_{d,\ell,L}| \cdot d \cdot q^{\alpha_0 d} \right)}{Z} - pL \\ &\leq \frac{L \cdot \left(\sum_{d=1}^{L} |D_{d,\ell,L}| \cdot \cdot q^{\alpha_0 d} \right)}{Z} - pL \end{aligned}$$
so it suffices to show that
$$\left(\sum_{d=1}^{L} |D_{d,\ell,L}| \cdot \cdot q^{\alpha_0 d} \right) \leq pZ.$$
The left-hand side is $Z - |D_{0,\ell,L}|$, so the above is equivalent to
$$\frac{|D_{0,\ell,L}|}{Z} \geq 1 - p.$$

Now,

$$Z \leq |D_{0,\ell,L}| + q^L \cdot q^{\alpha_0} = |D_{0,\ell,L}| + p$$

so indeed,

$$\frac{|D_{0,\ell,L}|}{Z} \geq \frac{|D_{0,\ell,L}|}{|D_{0,\ell,L}| + p}$$

which is at least $1 - p$ since $|D_{0,\ell,L}| \geq 1$. □

This completes the proof of the claim, and thus of Corollary 7.5. □

8 Acknowledgements

We would like to thank Ray Li for helpful conversations.

References

[ABL00] Alexei Ashikhmin, Alexander Barg, and Simon Litsyn. A new upper bound on codes decodable into size-2 lists. In *Numbers, Information and Complexity*, pages 239–244. Springer, 2000.

[ABP18] Noga Alon, Boris Bukh, and Yury Polyanskiy. List-decodable zero-rate codes. *IEEE Transactions on Information Theory*, 65(3):1657–1667, 2018.

[Bli86] Volodia M Blinovsky. Bounds for codes in the case of list decoding of finite volume. *Problems of Information Transmission*, 22(1):7–19, 1986.

[Bol01] Béla Bollobás. *Random Graphs, Second Edition*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2001.

[CS04] Imre Csiszár and Paul C Shields. *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.

[Din15] Y. Ding. On list-decodability of random rank metric codes and subspace codes. *IEEE Transactions on Information Theory*, 61(1):51–59, 2015.

[Eli91] Peter Elias. Error-correcting codes for list decoding. *IEEE Transactions on Information Theory*, 37(1):5–12, 1991.

[ER59] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[FK84] Michael L Fredman and János Komlós. On the size of separating systems and families of perfect hash functions. *SIAM Journal on Algebraic Discrete Methods*, 5(1):61–68, 1984.

[FKNP19] Keith Frankston, Jeff Kahn, Bhargav Narayanan, and Jinyoung Park. Thresholds versus fractional expectation-thresholds. *arXiv preprint arXiv:1910.13433*, 2019.

[Fri99] Ehud Friedgut. Sharp thresholds of graph properties, and the k-sat problem. *J. Amer. Math. Soc.*, 12(4):1017–1054, 1999. With an appendix by Jean Bourgain.

[GHK11] Venkatesan Guruswami, Johan Håstad, and Swastik Kopparty. On the list-decodability of random linear codes. *IEEE Trans. Information Theory*, 57(2):718–725, 2011.

[GI01] Venkatesan Guruswami and Piotr Indyk. Linear-time codes to correct a maximum possible fraction of errors. In *Proceedings of the Annual Allerton Conference on Communication Control and Computing*, volume 39, pages 857–866. The University; 1998, 2001.

[GLM+20] Venkatesan Guruswami, Ray Li, Jonathan Mosheiff, Nicolas Resch, Shashwat Silas, and Mary Wootters. Bounds for list-decoding and list-recovery of random linear codes. *arXiv preprint arXiv:2004.13247*, 2020.

[GN14] Venkatesan Guruswami and Srivatsan Narayanan. Combinatorial limitations of average-radius list-decoding. *IEEE Trans. Information Theory*, 60(10):5827–5842, 2014.

[GR18] Venkatesan Guruswami and Nicolas Resch. On the list-decodability of random linear rank-metric codes. In *2018 IEEE International Symposium on Information Theory, ISIT 2018, Vail, CO, USA, June 17-22, 2018*, pages 1505–1509. IEEE, 2018.

[GR19] Venkatesan Guruswami and Andrii Riazanov. Beating fredman-komlós for perfect k-hashing. In *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[Gur04] Venkatesan Guruswami. *List decoding of error-correcting codes: winning thesis of the 2002 ACM doctoral dissertation competition*, volume 3282. Springer Science & Business Media, 2004.

[GUV09] Venkatesan Guruswami, Christopher Umans, and Salil Vadhan. Unbalanced expanders and randomness extractors from parvaresh–vardy codes. *Journal of the ACM (JACM)*, 56(4):1–34, 2009.

[KM88] J Korner and Katalin Marton. New bounds for perfect hashing via information theory. *European Journal of Combinatorics*, 9(6):523–530, 1988.

[Kör86] Jénos Körner. Fredman–komlós bounds and information theory. *SIAM Journal on Algebraic Discrete Methods*, 7(4):560–570, 1986.

[LW18] Ray Li and Mary Wootters. Improved list-decodability of random linear binary codes. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[MRRZ+19] Jonathan Mosheiff, Nicolas Resch, Noga Ron-Zewi, Shashwat Silas, and Mary Wootters. Ldpc codes achieve list decoding capacity. *arXiv preprint arXiv:1909.06430*, 2019.

[Res20] Nicolas Resch. *List-Decodable Codes:(Randomized) Constructions and Applications*. PhD thesis, Carnegie Mellon University, 2020.

[RW18] Atri Rudra and Mary Wootters. Average-radius list-recovery of random linear codes. In *Proceedings of the 2018 ACM-SIAM Symposium on Discrete Algorithms, SODA*, 2018.

[Vad12] Salil P. Vadhan. Pseudorandomness. *Foundations and Trends in Theoretical Computer Science*, 7(1-3):1–336, 2012.

[WJ08] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[XY19] Chaoping Xing and Chen Yuan. Beating the probabilistic lower bound on perfect hashing. *arXiv preprint arXiv:1908.08792*, 2019.

[Zha07] Zhengmin Zhang. Estimating mutual information via kolmogorov distance. *IEEE Trans. Inf. Theory*, 53(9):3280–3282, 2007.

[ZP81] Victor Vasilievich Zyablov and Mark Semenovich Pinsker. List concatenated decoding. *Problemy Peredachi Informatsii*, 17(4):29–33, 1981.

A Example of a non-symmetric property

In this section, we give an example of a property that is row-symmetric (aka, is defined by a collection of bad types), but is not symmetric. Then we will see that Theorem 3.6 is *not* satisfied for this property.

Here is one such property $\mathcal{P} = \{P_n\}_{n\in\mathbb{N}}$. Let $p \in (0, 1/2)$. Say that $C \subset \{0,1\}^n$ satisfies $\mathcal{P}$ if there are three codewords $c^{(1)}, c^{(2)}, c^{(3)} \in C$ so that for some point $z \in \{0,1\}^n$, $d(c^{(1)}, z), d(c^{(2)}, z) \leq \frac{pn}{2}$, and $d(c^{(3)}, z) \leq pn$.

That is, this property is similar to list-of-two decoding, except that two of the codewords must be closer to the central point than the third must be. This property is not symmetric, and Theorem 3.6 does not hold for it. However, in order to see an example that is as clean as possible, let's instead consider a "toy" version of this property.

Define the property $\mathcal{P}^{\text{toy}}$ as follows. Let $T = \{\tau_1, \tau_2, \tau_3\}$, where

$$\begin{aligned}
\tau_1((1,0,0)) &= \tau_1((0,1,0)) = \frac{p}{2}\\
\tau_1((0,0,1)) &= p\\
\tau_1((0,0,0)) &= 1-2p
\end{aligned}$$

$$\begin{aligned}
\tau_2((1,0,0)) &= \tau_2((0,0,1)) = \frac{p}{2}\\
\tau_2((0,1,0)) &= p\\
\tau_2((0,0,0)) &= 1-2p
\end{aligned}$$

$$\tau_3((0,0,1)) = \tau_3((0,1,0)) = \frac{p}{2}$$
$$\tau_3((1,0,0)) = p$$
$$\tau_3((0,0,0)) = 1 - 2p$$

(and $\tau_i((1,1,1)) = 0$ for all i). That is, τ_1, τ_2, and τ_3 correspond to the matrices B so that two columns have exactly a $p/2$ fraction of 1's, one column has exactly a p fraction of 1's, and the support of the columns are disjoint. (For simplicity, assume that $pn/2$ is an integer.) Then $\mathcal{P}^{\text{toy}}$ is defined by the inclusion of at least one type in T.

The toy version $\mathcal{P}^{\text{toy}}$ is in spirit the same as the property $\mathcal{P}$ above. The differences are that (a) we are fixing the vector z to be the all-zero vector, (b) we demand that the distance be exactly p or $p/2$, rather than at most p or $p/2$, and (c) we are not considering configurations where the three vectors have overlapping ones. The computations are similar for both $\mathcal{P}^{\text{toy}}$ and $\mathcal{P}$. We mention both, but analyze $\mathcal{P}^{\text{toy}}$, since $\mathcal{P}$ is perhaps more natural, while $\mathcal{P}^{\text{toy}}$ is simpler to analyze.

First, observe that $\mathcal{P}^{\text{toy}}$ is not symmetric. In particular, the distribution $\frac{1}{3}(\tau_1 + \tau_2 + \tau_3)$, which corresponds to a matrix with $2p/3$ fraction of 1's in each column, is not contained in T.

Theorem 3.6 would predict that the threshold rate for $\mathcal{P}^{toy}$ would be

$$\begin{aligned} R^{\text{theorem}} &= 1 - \max_{\tau \in T} H(\tau)/3 \\ &= 1 - \frac{1}{3}\left(p \log\left(\frac{2}{p^2}\right)\right. \\ &\qquad \left. + (1-2p)\log\left(\frac{1}{1-2p}\right)\right). \end{aligned}$$

However, the actual threshold rate is larger than this. To see this, observe that the probability that C satisfies $\mathcal{P}^{\text{toy}}$ is at most the probability that C contains the type $\tau^\dagger$ on $\{0,1\}^2$ given by

$$\tau^\dagger((1,0)) = \tau^\dagger((0,1)) = \frac{p}{2}$$
$$\tau^\dagger((0,0)) = 1 - p$$
$$\tau^\dagger((1,1)) = 0.$$

Indeed, suppose that C contains τ_i for some $i = 1, 2, 3$. Then in particular it contains two codewords of weight $p/2$, meaning that it contains $\tau^\dagger$. Therefore,

$$R^* \geq R^\dagger,$$

where $R^\dagger$ is the rate threshold for the property corresponding to $\{\tau^\dagger\}$. But the rate threshold for $\tau^\dagger$ is

$$\begin{aligned} R^\dagger &= 1 - \frac{H(\tau^\dagger)}{2} \\ &= 1 - \frac{1}{2}\left(p\log\left(\frac{2}{p}\right) + (1-p)\log\left(\frac{1}{1-p}\right)\right). \end{aligned}$$

Figure 2 plots these two values, and it is clear that $R^\dagger$ (and hence R^*) is larger than R^{theorem} for $p \leq 0.3$.

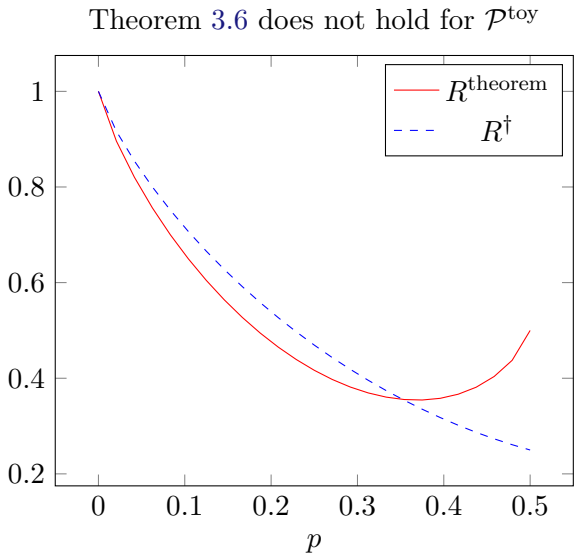

Figure 2: The value R^{theorem} that Theorem 3.6 would predict for $\mathcal{P}^{\text{toy}}$, and the value $R^{\dagger}$ that is a lower bound on the actual value of R^*, for $p \in (0, 0.5)$.

B Proofs of claims from Sections 2 and 3

We restate and prove the deferred lemma.

Lemma 2.22.

$$\lim_{n\to\infty} \sup_{\tau \in T_{p,\ell,L}} d_\infty(\tau, T^n_{p,\ell,L}) = 0.$$

Proof. Let $\tau_0 \in T_{p,\ell,L}$, and let ρ_0 be a distribution over $\Sigma^L \times \binom{\Sigma}{\ell}$ that satisfies Conditions 1 and 2 with regard to τ_0. Let τ_1 denote the atomic distribution over Σ^L that always returns the all-σ vector, for some arbitrary $\sigma \in \Sigma$. Let ρ_1 be the atomic distribution over $\Sigma^L \times \binom{\Sigma}{\ell}$ that always returns (σ, W), where $W \in \binom{\Sigma}{\ell}$ is some arbitrary set containing σ.

Fix some $\varepsilon > 0$ and let τ_ε and ρ_ε denote the mixture distributions $(1-\varepsilon)\tau_0 + \varepsilon\tau_1$ and $(1-\varepsilon)\rho_0 + \varepsilon\rho_1$, respectively. Let $\{(u^i, K^i)\}_{i=1}^n$ denote a sequence of n independent samples from ρ_ε.

Sample i uniformly from $[n]$. Let $\hat{\tau}$ denote the distribution of (u^i) and let $\hat{\rho}$ be the distribution of (u^i, K^i). It is straightforward to verify that $\hat{\rho}$ satisfies Conditions 1 and 3 with regard to $\hat{\tau}$. If $p = 0$ then Condition 2 also clearly holds.

Suppose that $p > 0$. We claim that, with some positive probability, Condition 2 is satisfied as well, and consequently, $\hat{\tau} \in T^n_{p,\ell,L}$.

Let X_j denote the number of coordinates i for which $u^i_j \notin K^i$. Note that the CDF of X_j is bounded from below by that of a Binomial$(n, (1-\varepsilon)p)$ variable. Consequently, by Hoeffding's

bound,

$$\begin{aligned}
&\Pr_{\hat{\rho}}\left[\exists j \in L \text{ s.t. } \Pr_{(u,K)\sim\hat{\rho}}[u_j \in K] > p\right] \\
&\leq L \cdot \Pr_{\hat{\rho}}\left[\Pr_{(u,K)\sim\hat{\rho}}[u_j \in K] > p\right] \\
&= L \cdot \Pr\left[X_j > pn\right] \\
&\leq L \cdot e^{-2(p\varepsilon)^2 n}.
\end{aligned}$$

Hence, taking $\varepsilon > \sqrt{\frac{\log_q L}{2p^2 n}}$, Condition 2 is satisfied with positive probability. Therefore, there exists $\hat{\tau} \in T_{p,\ell,L}$ with $d_\infty(\tau, \hat{\tau}) \leq \varepsilon = O\left(\sqrt{\frac{\log_q L}{p^2 n}}\right)$. □

C Threshold rate for list-of-two decoding of random linear codes

In this section, we prove Part 2 of Theorem 6.1. For convenience, the statement is provided below.

Theorem C.1 (Theorem 6.1, Part 2, Restated)**.** *Let $p \in (0, 1/4)$. The threshold rate for $(p, 3)$-list-decoding for random linear codes satisfies*

$$\lim_{n\to\infty} R^n_{\mathrm{RLC}}(\mathcal{P}) = 1 - \frac{h(3p) + 3p\log 3}{2}\ .$$

Following [MRRZ+19], we can again characterize the threshold rate for random linear codes in terms of types.[5] However, the characterization is now more complicated; in particular, we need to account for so-called *implied* distributions. We now introduce the necessary machinery, specialized to the $q = 2$ case.

For a distribution $\tau \sim \mathbb{F}_2^b$, consider the span of the support of τ. The dimension of this subspace is denoted by $\dim(\tau)$.

Definition C.2 (Implied Distribution)**.** *Let $b \in \mathbb{N}$, let τ be a distribution over $\mathbb{F}_2^b$ and let $A \in \mathbb{F}_2^{m\times\ell}$ be a rank m matrix for some $m \leq \ell$. The distribution of $Au \in \mathbb{F}_2^m$, where u is sampled according to τ, is said to be* τ-implied*. We denote the set of τ-implied types by $\mathcal{I}_\tau$.*

We briefly motivate the above definition; for further details, the reader is referred to [MRRZ+19], specifically Section 2. Suppose a linear code C contains a matrix $M \in \mathbb{F}_2^{n\times b}$ of type τ. Then, it also contains any matrix $M' \in \mathbb{F}_2^{n\times m}$ whose columns lie in the column span of M. That is, for any full-rank matrix $A \in \mathbb{F}_2^{m\times b}$, it contains the matrix MA^T. Moreover, note that if M has type τ, then the type of $M' = MA^T$ is τ' as in Definition C.2.

We now quote the result from [MRRZ+19] that we require.

Theorem C.3 (Theorem 2.8 of [MRRZ+19])**.** *Fix $b \in \mathbb{N}$. Let $(P_n)_{n\in\mathbb{N}}$ be a sequence of monotone-increasing properties such that $T_n \subseteq \mathcal{T}_b^n$ is a minimal set for P_n for all $n \in \mathbb{N}$. Then*

$$R^n_{\mathrm{RLC}}(P_n) = 1 - \max_{\tau\in T_n} \min_{\tau'\in\mathcal{I}_\tau} \frac{H(\tau')}{\dim(\tau')} \pm o_{n\to\infty}(1).$$

[5] Again, we warn the reader that in that work monotone-*decreasing* properties were considered, whereas here we consider monotone-increasing properties.

With these tools, we may prove Theorem C.1.

Proof of Theorem C.1. Specializing to the specific property $\mathcal{P} = (P_n)_{n\in\mathbb{N}}$ of $(p,3)$-list-decodability, recall that T_n should consist of all types $\tau \in \mathcal{T}_3^n$ for which there exists a distribution ρ over $\mathbb{F}_2^3 \times \mathbb{F}_2$[6] such that, given $(u,z) \sim \rho$, the following holds:

1. The distribution of u is τ.
2. $\Pr[u_j \neq z] \leq p$ for every $1 \leq j \leq 3$.
3. $n\rho((u,z)) \in \mathbb{N}$ for every $u \in \mathbb{F}_2^3$ and $z \in \mathbb{F}_2$.
4. Any matrix $M \in M_\tau$ has distinct columns, i.e., $M \in (\mathbb{F}_2)^{n\times 3}_{\text{distinct}}$. Stated differently, $\Pr[u_i \neq u_j] > 0$ for any $1 \leq i < j \leq 3$.

Define

$$\bar{\gamma}(p,1,3) := \lim_{n\to\infty} \max_{\tau\in T_n} \min_{\tau'\in\mathcal{I}_\tau} \frac{H(\tau')}{\dim(\tau')};$$

Theorem C.3 says that $\lim_{n\to\infty} R^n_{\text{RLC}}(P_n) = 1 - \bar{\gamma}(p,1,3)$. We will show $\bar{\gamma}(p,1,3) = \frac{h(3p)+3p\log 3}{2}$. We begin by proving an upper bound on $\bar{\gamma}(p,1,3)$.

Upper bound on $\bar{\gamma}(p,1,3)$. Let $\tau \in T_n$ and let $\rho \sim \mathbb{F}_2^3 \times \mathbb{F}_2$ be the promised distribution satisfying Conditions 1–4. Condition 2 implies that

$$\frac{1}{3}\sum_{j=1}^{3} \Pr_{(u,z)\sim\rho}[u_j \neq z] \leq p. \tag{17}$$

Let $\text{MAJ}(u)$ denote the majority element of the vector u,[7] and note that if $z = \text{MAJ}(u)$ the left-hand side of (17) can only decrease. Hence,

$$\frac{1}{3}\sum_{i=1}^{3} \Pr_{u\sim\tau}[u_i \neq \text{MAJ}(u)] \leq p. \tag{18}$$

Now, as in Section 6 let[8] $A = \{000, 111\}$ and $B = \mathbb{F}_2^3 \setminus A$. Then, defining $x = \tau(B)$, Condition (18) becomes

$$x \leq 3p.$$

Now, consider the implied type $\tau^* \in \mathcal{I}_\tau$ defined by the linear map $(a,b,c) \mapsto (a+b, a+c)$. Note that the kernel of this map is $\{000, 111\}$, and so $\tau^*(00) = 1 - x$. Hence, $\tau^*(10) + \tau^*(01) + \tau^*(11) = x$.

[6] Technically, to be completely consistent with Section 2, we should consider distributions over $\mathbb{F}_2^3 \times \binom{\mathbb{F}_2}{1}$, but we just identify elements with the corresponding singleton sets in the natural way.

[7] In symbols, $\text{MAJ}(u) = \text{argmax}\{b \in \mathbb{F}_2 : |\{j \in [3] : u_j = b\}|\}$.

[8] In this proof, we denote a vector by the corresponding string for readability.

We may therefore bound the entropy from above as follows:

$$\begin{aligned} H(\tau^*) &= \tau^*(00) \cdot \log \tfrac{1}{\tau^*(00)} + \tau^*(01) \cdot \log \tfrac{1}{\tau^*(01)} \\ &\quad + \tau^*(10) \cdot \log \tfrac{1}{\tau^*(10)} + \tau^*(11) \cdot \log \tfrac{1}{\tau^*(11)} \\ &\leq (1-x) \log \tfrac{1}{1-x} + x \log \tfrac{3}{x} \\ &= h(x) + x \log 3 \ . \end{aligned}$$

The above inequality uses the concavity of the function $y \mapsto y \log \frac{1}{y}$. Note that in the range $[0, 3/4)$, the function $x \mapsto h(x) + x \log 3$ is increasing: clearly $h_2(0) + 0 \cdot \log_2 3 = 0$, and moreover the derivative of $h(x) + x \log_2 3$ with respect to x is $\log\left(\frac{3(1-x)}{x}\right)$, which is positive assuming $\frac{3(1-x)}{x} > 1$, which rearranges to $x < 3/4$. Hence, as $x \leq 3p$ and $p < 1/4$, we conclude

$$H(\tau^*) \leq h(3p) + 3p \log 3 \ .$$

Now, we claim that $\dim(\tau^*) = 2$. Let $U = \mathrm{span}(\mathrm{supp}(\tau))$. If $\dim(\tau^*) \leq 1$, then $\dim(U) \leq 2$ and $111 \in U$ (recall that the kernel of $(a, b, c) \mapsto (a + b, b + c)$ is $\{000, 111\}$). This implies

$$\begin{aligned} U \in \{&\{000\}, \{000, 111\}, \{000, 111, 001, 110\}, \\ &\{000, 111, 010, 101\}, \{000, 111, 100, 011\}\} \ . \end{aligned}$$

In any of the above cases, we find that τ contradicts Condition 4. For example, if $U = \{000, 111, 001, 110\}$, then $\Pr[u_i \neq u_j] = 0$.

As $\tau \in T_n$ was arbitrary, we conclude that

$$\begin{aligned} \bar{\gamma}(p, 1, 3) &= \lim_{n \to \infty} \max_{\tau \in T_n} \min_{\tau' \in \mathcal{I}_\tau} \frac{H(\tau')}{\dim(\tau')} \\ &\leq \frac{h(3p) + 3p \log 3}{2}, \end{aligned}$$

as desired.

Lower bound on $\bar{\gamma}(p, 1, 3)$. For any $n \in \mathbb{N}$, we define a type $\tau_{n,0} \in T_n$ and show that

$$\min_{\tau' \in \mathcal{I}_{\tau_{n,0}}} \frac{H(\tau')}{\dim(\tau')} \geq \frac{h(3p) + 3p \log 3}{2} - o_{n \to \infty}(1);$$

taking limits will then show $\bar{\gamma}(p, 1, 3) \geq \frac{h(3p) + 3p \log 3}{2}$, as desired.

Define $\rho_n \sim \mathbb{F}_2^{3 \times 1}$ by assigning probability mass $\frac{1}{n} \left\lfloor \frac{p}{2} \cdot n \right\rfloor$ to each of the elements $(001, 0), (010, 0), (100, 0), (011, 1), (101, 1), (110, 1) \in \mathbb{F}_2^3 \times \mathbb{F}_2$, and splitting the remaining probability mass as evenly as possible between the $(000, 0)$ and $(111, 1)$, being sure to obey the condition that $n \cdot \rho_n(000, 0)$ and $n \cdot \rho_n(111, 1)$ are integers. Intuitively, ρ_n is tending to the "limit" distribution ρ defined by

$$\begin{aligned} \rho(000, 0) = \rho(111, 1) &= \frac{1 - 3p}{2} \quad \text{and} \\ \rho(001, 0) = \rho(010, 0) &= \rho(100, 0) = \rho(011, 1) \\ = \rho(101, 1) &= \rho(110, 1) = \frac{p}{2}; \end{aligned}$$

however, as we require a type lying in T_n, we cannot directly work with ρ. Let $\tau_{n,0}$ denote the distribution of u for $(u,z) \sim \rho_n$. It is immediate that Conditions 2, 3 and 4 are satisfied for ρ_n and $\tau_{n,0}$. Also, observe that $\dim(\tau_{n,0}) = 3$, as $\tau_{n,0}$ has full-support for sufficiently large n (this uses the assumption $p \in (0, 1/4)$).

Our plan now is to consider all of the implied types of $\tau_{n,0}$ one-by-one. We will show that each such implied type $\tau'_{n,0}$ satisfies

$$\frac{H(\tau'_{n,0})}{\dim(\tau'_{n,0})} \geq \frac{h(3p) + 3p\log 3}{2} - o_{n\to\infty}(1),$$

and therefore, we will deduce

$$\min_{\tau'_{n,0} \in \mathcal{I}_{\tau_{n,0}}} \frac{H(\tau'_{n,0})}{\dim(\tau'_{n,0})} \geq \frac{h(3p) + 3p\log 3}{2} - o_{n\to\infty}(1).$$

First, one can compute that

$$\frac{H(\tau_{n,0})}{\dim(\tau_{n,0})} \geq \frac{h_2(3p) + 1 + 3p\log_2 3}{3} - o_{n\to\infty}(1).$$

Next, consider the type $\tau_{n,1}$ implied by the map $(a,b,c) \mapsto (a+b, a+c)$. One can calculate

$$\frac{H(\tau_{n,1})}{\dim(\tau_{n,1})} \geq \frac{h_2(3p) + 3p\log 3}{2} - o_{n\to\infty}(1).$$

Next, consider any type $\tau_{n,2}$ implied by a full-rank map $\mathbb{F}_2^3 \to \mathbb{F}_2^2$ with a vector from B in the kernel. We compute

$$\frac{H(\tau_{n,2})}{\dim(\tau_2)} \geq \frac{h(2p) + 1}{2} - o_{n\to\infty}(1).$$

Next, consider any type $\tau_{n,3}$ implied by a full-rank map $\mathbb{F}_2^3 \to \mathbb{F}_2$ with 111 in the kernel. In this case,

$$\frac{H(\tau_{n,3})}{\dim(\tau_{n,3})} \geq h(2p) - o_{n\to\infty}(1).$$

Finally, consider any type $\tau_{n,4}$ implied by a full-rank map $\mathbb{F}_2^3 \to \mathbb{F}_2$ without 111 in the kernel. In this case,

$$\frac{H(\tau_{n,4})}{\dim(\tau_{n,4})} \geq 1 - o_{n\to\infty}(1).$$

This completes the computation of $\frac{H(\tau'_n)}{\dim(\tau'_n)}$ for each $\tau'_n \in \mathcal{I}_{\tau_{n,0}}$.

It is now just a finite check to verify that

$$\frac{H(\tau'_n)}{\dim(\tau'_n)} \geq \frac{h(3p) + \log 3}{2} - o_{n\to\infty}(1).$$

For example, to show that

$$\begin{aligned}\frac{H(\tau_{n,0})}{\dim(\tau_{n,0})} &= \frac{h(3p) + 1 + 3p\log 3}{3} - o_{n\to\infty}(1) \\ &\geq \frac{h(3p) + \log 3}{2} - o_{n\to\infty}(1),\end{aligned}$$

one can reason as follows. First, it clearly suffices to show

$$\frac{h(3p) + 1 + 3p \log 3}{3} \geq \frac{h(3p) + \log 3}{2}.$$

Now, note that both the left-hand side and the right-hand side are 0 at $p = 1/4$, and moreover the derivative of the right-hand side minus the left-hand side is $\frac{1}{6}\left(\log\left(\frac{3p}{1-3p}\right) - \log 3\right)$, which is negative if $p < 1/4$. We omit the remaining computations, which are completely routine; for a pictorial proof, see Figure 3.

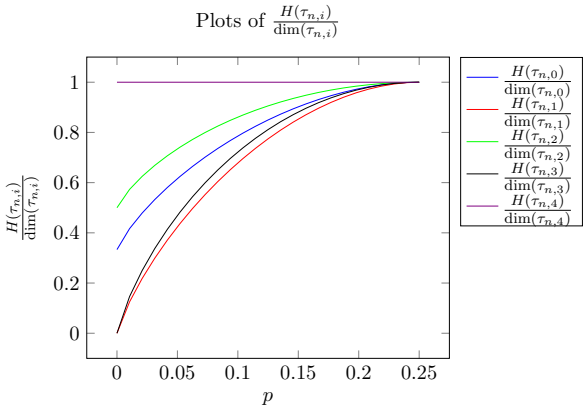

Figure 3: Plots of $\frac{H(\tau_{n,i})}{\dim(\tau_{n,i})}$ for each $i \in \{0, 1, 2, 3, 4\}$, ignoring $o_{n\to\infty}(1)$ terms. One can see that, uniformly over $p \in (0, 1/4)$, the minimum is obtained by $\frac{H(\tau_{n,1})}{\dim(\tau_{n,1})}$.

Thus, we conclude that

$$\bar{\gamma}(p, 1, 3) = \lim_{n\to\infty} \max_{\tau \in T_n} \min_{\tau' \in \mathcal{I}_\tau} \frac{H(\tau')}{\dim(\tau')} \geq \frac{h(3p) + 3p \log 3}{2},$$

completing the proof. □