List-Decodability of Structured Ensembles of Codes

Mary Wootters
Stanford University, CA, USA
marykw@stanford.edu

Abstract

What combinatorial properties are satisfied by a random subspace over a finite field? For example, is it likely that not too many points lie in any Hamming ball? What about any cube? In this talk, I will discuss the answer to these questions, along with a more general characterization of the properties that are likely to be satisfied by a random subspace. The motivation for this characterization comes from *error correcting codes.* I will discuss how to use this characterization to make progress on the questions of *list-decoding* and *list-recovery* for random linear codes, and also to establish the list-decodability of random *Low Density Parity-Check* (LDPC) codes.

This talk is based on the works [11] and [6], which are joint works with Venkatesan Guruswami, Ray Li, Jonathan Mosheiff, Nicolas Resch, Noga Ron-Zewi, and Shashwat Silas.

2012 ACM Subject Classification Theory of computation → Error-correcting codes

Keywords and phrases Error Correcting Codes, List-Decoding

Digital Object Identifier 10.4230/LIPIcs.MFCS.2020.3

Category Invited Talk

Acknowledgements This invited talk is based on the work [6], co-authored with Jonathan Mosheiff, Nicolas Resch, Noga Ron-Zewi, and Shashwat Silas; and on the work [11], co-authored with Venkatesan Guruswami, Ray Li, Jonathan Mosheiff, Nicolas Resch, and Shashwat Silas.

1 Introduction

Let $\mathbb{F}_q$ be the finite field of order q, and let $\mathcal{C} \subseteq \mathbb{F}_q^n$ be a random linear subspace of dimension k. What combinatorial properties does $\mathcal{C}$ satisfy? For example, how far apart (in Hamming distance) can we expect the closest two elements in $\mathcal{C}$ to be? What's the maximum number of elements of $\mathcal{C}$ likely to lie in any Hamming ball? What's the maximum number to lie in any cube $S_1 \times S_2 \times \cdots \times S_n$ of size ℓ^n?

These questions are interesting from a mathematical point of view, but they are also relevant to the study of *error correcting codes.* In coding-theoretic language, the set $\mathcal{C}$ is a *random linear code* of *rate* k/n. The questions above ask about the *distance, list-decodability,* and *list-recoverability* of $\mathcal{C}$, respectively. While answer to the first question – about the distance – is a classical result of Varshamov [14], the second two questions are much more challenging and aspects of them are still open.

In this talk, I will discuss a recent characterization of properties that are satisfied by a random linear code $\mathcal{C}$, from [11] (Section 3). Then I will discuss (Section 4) how to use this characterization to both make progress on the second and third questions above (list-decoding and list-recovery of random linear codes), as well as to establish the list-decodability of Gallager's ensemble of Low-Density Parity-Check (LDPC) codes. This talk is based on the works [11] and [6].

45th International Symposium on Mathematical Foundations of Computer Science (MFCS 2020).
Editors: Javier Esparza and Daniel Král'; Article No. 3; pp. 3:1–3:5
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 Background

List-decoding was introduced by Elias and Wozencraft in the 1950's [2, 16]. For $p \in [0,1]$ and integer $L \geq 1$, we say that a code $\mathcal{C} \subseteq \mathbb{F}_q^n$ is (p,L)-*list-decodable* if, for all $z \in \mathbb{F}_q^n$,

$$|\{c \in \mathcal{C} : \delta(c,z) \leq p\}| < L,$$

where $\delta(x,y) = \frac{1}{n}|\{i : x_i \neq y_i\}|$ denotes relative Hamming distance. That is, $\mathcal{C}$ is list-decodable if not too many codewords of $\mathcal{C}$ live in any small enough Hamming ball.

List-recovery is a more recent notion, which arose in the context of designing list-decodable codes, but which has since found other applications and become interesting in its own right. For integers $\ell < L$, we say that a code $\mathcal{C} \subseteq \mathbb{F}_q^n$ is (ℓ, L)-*list-recoverable* if, for all $S_1, \ldots, S_n \subseteq \mathbb{F}_q$ so that $|S_i| \leq \ell$,

$$|\{c \in \mathcal{C} : c_i \in S_i \forall i\}| < L.$$

That is, $\mathcal{C}$ is list-recoverable if not too many codewords of $\mathcal{C}$ live in any $\ell \times \ell \times \cdots \times \ell$ cube.[1]

For both list-decoding and list-recovery, we are interested in the trade-offs between the parameters (p, L) or (ℓ, L) and the rate of the code. Codes which acheive the largest rate possible, in terms of p or ℓ, while keeping the list size L small (constant), are said to *achieve capacity* for list-decoding or list-recovery respectively.

The list-decodability and list-recoverability of random linear codes has been well-studied [17, 4, 1, 15, 12, 13, 10], and by now it is known in many (but not all) parameter regimes that random linear codes achieve list-decoding capacity with high probability. Understanding the list-decodability/recoverability of random linear codes is motivated both as a fundamental question and because it can lead to improvements in other constructions that use random linear codes as a building block [5, 7, 9, 8]. Moreover, as was shown in [11] and as we will see later in this extended abstract, understanding the list-decodability/recoverability of random linear codes can lead to an understanding of another structured ensembles of codes: *Gallager's ensemble* of LDPC codes [3].

An LDPC code is based on a sparse (constant-degree) bipartite graph, $G = (V, W, E)$ with $|V| = n, |W| = m$, and $m \leq n$. The n symbols of a codeword $c \in \mathcal{C}$ are identified with n vertices in V. The vertices in W are *parity checks:* a vector $c \in \mathbb{F}_q^n$ is in $\mathcal{C}$ if, for every vertex $j \in W$, we have $\sum_{i \in \Gamma(j)} \alpha_{i,j} c_i = 0$, where $\Gamma(i)$ denotes the neighbors of i in G and $\alpha_{i,j} \in \mathbb{F}_q$ are some fixed coefficients. LDPC codes are notable for their extremely efficient algorithms for unique decoding (that is, the case that $L = 1$), and are ubiquitous in both theory and applications. Gallager's ensemble is given by a particular distribution on random graphs. Gallager showed that the codes arising from such graphs have good distance with high probability, but until the work [11], nothing was known about their list-decodability.

3 Characterization of sets in a random linear code

The problems of list-decoding and list-recovery of random linear codes are special cases of the following question:

> Let $R \in (0,1)$. Let $\mathcal{B}$ be a collection of subsets $B \subseteq \mathbb{F}_q^n$, so that $|B| \leq L$ for all $B \in \mathcal{B}$. Is it likely that $\mathcal{B}$ is represented in a random subspace $\mathcal{C} \subseteq \mathbb{F}_q^n$ of dimension $k = Rn$?

[1] This definition of "list-recoverability" is often referred to as "zero-error list-recoverability." There is a more general notion of (p, ℓ, L)-list-recoverability, where the requirement is that $c_i \in S_i$ for at least a $(1-p)$ fraction of the coordinates $i \in [n]$. However, since we will not discuss this more general notion here, we use "list-recovery" to refer to the zero-error setting.

Above, by "$\mathcal{B}$ is represented in $\mathcal{C}$," we mean that there is some $B \in \mathcal{B}$ so that $B \subseteq \mathcal{C}$. For list-decoding, $\mathcal{B}$ is the collection of all sets of L vectors that lie in a Hamming ball; for list-recovery, $\mathcal{B}$ is the collection of all sets of L vectors that lie in an $\ell \times \ldots \times \ell$ cube.

One of the contributions of [11] is a characterization of the collections $\mathcal{B}$ that are represented in a random linear code of a given rate. More precisely, suppose that $\mathcal{B}$ is permutation-invariant (in the sense that $\pi(B) \in \mathcal{B}$ for all $B \in \mathcal{B}$ and for all $\pi \in S_n$, where π acts by permuting the coordinates of a vector $c \in \mathbb{F}_q^n$). Then we have the following theorem.

▶ **Theorem 1** ([11]). *Let $\varepsilon > 0$. For any permutation-invariant collection $\mathcal{B}$ of sets of size at most L, there is a threshold rate R^* so that the following holds. Let R be the rate of a random linear code $\mathcal{C}$. If $R \geq R^* + \varepsilon$, then w.h.p. $\mathcal{B}$ is represented in $\mathcal{C}$; while if $R < R^* - \varepsilon$, the w.h.p. $\mathcal{B}$ is not represented in $\mathcal{C}$.*

Moreover, [11] gives a characterization of the threshold rate R^*. We describe the intuition behind this characterization below, and refer the reader to [11] for the precise statements.

We can break up a set $\mathcal{B}$ into permutation-invariant classes: if we view a list $B \in \mathcal{B}$ as a matrix $B \in \mathbb{F}_q^{n \times L}$ with the elements of B as columns, then a class corresponds to a distribution τ on $\mathbb{F}_q^L$ given by the rows of B. For simplicity, suppose that $\mathcal{B}$ consists of only one such class, given by the distribution τ. Below, we will conflate the list $B \subset \mathbb{F}_q^n$ with the matrix $B \in \mathbb{F}_q^{n \times L}$ as above.

Intuitively, if the entropy $H(\tau)$ of this distribution is small, then there are not many sets $B \in \mathcal{B}$; therefore, from the union bound, it is not likely that a random linear code will contain them. Quantitatively, suppose that the support of τ has dimension[2] L, and that $H(\tau) < \gamma L \log(q)$ for some $\gamma \in (0,1)$. It is not hard to see that there are at most $q^{Ln(\gamma - o(1))}$ sets $B \in \mathcal{B}$. Since the elements of B are linearly independent, the probability that they are all contained in a random linear code of rate R is at most q^{-RnL}. By a union bound, if $R < 1 - \gamma - \varepsilon$ for some small $\varepsilon > 0$, then no $B \in \mathcal{B}$ is contained in $\mathcal{C}$ with high probability.

This allows us to partially answer the question above: if $\mathcal{B}$ corresponds to a class τ so that $\dim(\mathrm{Supp}(\tau)) = L$ and so that $H(\tau) < (1-R)L \log q$, then $\mathcal{B}$ is not likely to be represented in $\mathcal{C}$. However, it is not hard to see that this picture is incomplete. In particular, there are examples of distributions τ of full rank L where $H(\tau)$ is significantly larger than $(1-R)L \log q$, but so that $\mathcal{B}$ is not likely to be represented in a random linear code of rate R. One example of when this can happen is when there is some linear map $A : \mathbb{F}_q^L \to \mathbb{F}_q^{L'}$ for $L' < L$ so that $H(A(\tau)) < (1-R)L' \log q$. If this happens, then consider the collection $\mathcal{B}'$ given by $\tau' = A(\tau)$. (So, elements of $\mathcal{B}'$ are sets $B' \subseteq \mathbb{F}_q^n$ of size L'). The logic above shows that $\mathcal{B}'$ is not likely to be represented in $\mathcal{C}$. But for any $B' \in \mathcal{B}'$, we have $B' = BA^T$ for some $B \in \mathcal{B}$, and hence $B \subseteq \mathcal{C}$ implies that $B' \subseteq \mathcal{C}$ as $\mathcal{C}$ is linear. Thus, if $\mathcal{B}'$ is not likely to be represented in $\mathcal{C}$, then neither is $\mathcal{B}$.

The characterization of [11] shows that in fact this is the *only* way that the simple computation above is incomplete. That is, we can characterize the threshold rate R^* for a set $\mathcal{B}$ in terms of the entropy of linear maps of the corresponding row distributions τ:

$$R^* = \min_{\tau} \max_{\tau' = A(\tau)} 1 - \frac{H(\tau')}{\dim(\mathrm{Supp}(\tau')) \log q},$$

where the minimum is over all distributions τ that appear as permutation-invariant classes in $\mathcal{B}$, and the maximum is over all linear maps A.

[2] A similar argument will hold if τ has dimension smaller than L, after a suitable projection.

4 Applications

We conclude by briefly mentioning some applications of the characterization described above.

List-decodability of Gallager's ensemble of LDPC codes. The work [11] uses the characterization above to show that if a collection $\mathcal{B}$ is likely to be represented in a random LDPC code from Gallager's ensemble, then it is likely to be represented in a random linear code. Since recent works have shown that random linear codes achieve list-decoding capacity, this implies that Gallager's ensemble does as well.

New lower bounds for list-recovery random linear codes. The work [6] uses this characterization to obtain lower bounds on the list size for list-recovery of random linear codes. Because of this characterization, it is sufficient to exhibit a distribution π so that the class $\mathcal{B}$ associated with π is bad for list-recovery, and so that $H(A(\pi))$ is large for every linear transformation A. This can show that a random linear code of rate $1 - \log_q(\ell) - \varepsilon$ requires list size $L \geq \ell^{\Omega(1/\varepsilon)}$. This may be surprising, because for completely random codes a list size of $L = O(\ell/\varepsilon)$ suffices.

Three-point concentration of the list size for list-decoding random linear codes. The work [6] uses the characterization in a similar way to prove an extremely tight lower bound on the list size for list-decoding random linear codes. Combined with an upper bound of [10], this establishes that the list size of a binary random linear code is concentrated on at most[3] three values: $\lfloor h_q(p)/\varepsilon \rfloor + 2$, $\lfloor h_q(p)/\varepsilon \rfloor + 1$, and $\lfloor h_q(p)/\varepsilon + 0.99 \rfloor$, where $h_q(x)$ denotes the q-ary entropy.

5 Conclusion

In this extended abstract we have discussed a recent characterization of [11], which gives a precise threshold for the rate at which a random linear code satisfies a combinatorial property defined by the exclusion of a collection of small sets $\mathcal{B}$. As we have discussed, this characterization has proved useful for analyzing random linear codes themselves, as well as for analyzing other structured random codes (Gallager's ensemble). We hope that this characterization can be useful even more broadly: the main open question posed by this talk is to find more applications!

References

1 Mahdi Cheraghchi, Venkatesan Guruswami, and Ameya Velingker. Restricted isometry of fourier matrices and list decodability of random linear codes. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 432–442, 2013. doi:10.1137/1.9781611973105.31.

2 Peter Elias. List decoding for noisy channels. *Wescon Convention Record, Part 2*, pages 94–104, 1957.

3 Robert G. Gallager. Low-density parity-check codes. *IRE Trans. Information Theory*, 8(1):21–28, 1962. doi:10.1109/TIT.1962.1057683.

[3] Depending on the value of $h_q(p)/\varepsilon$, there may be only two values in this set.

4 Venkatesan Guruswami, Johan Håstad, and Swastik Kopparty. On the list-decodability of random linear codes. *IEEE Trans. Information Theory*, 57(2):718–725, 2011. doi:10.1109/TIT.2010.2095170.

5 Venkatesan Guruswami and Piotr Indyk. Efficiently decodable codes meeting Gilbert-Varshamov bound for low rates. In *SODA*, volume 4, pages 756–757, 2004.

6 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. To appear, RANDOM 2020.

7 Venkatesan Guruswami and Atri Rudra. Concatenated codes can achieve list-decoding capacity. *Electronic Colloquium on Computational Complexity (ECCC)*, 15(054), 2008. URL: http://eccc.hpi-web.de/eccc-reports/2008/TR08-054/index.html.

8 Brett Hemenway, Noga Ron-Zewi, and Mary Wootters. Local list recovery of high-rate tensor codes & applications. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 204–215. IEEE, 2017.

9 Brett Hemenway and Mary Wootters. Linear-time list recovery of high-rate expander codes. *Information and Computation*, 261:202–218, 2018.

10 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.

11 Jonathan Mosheiff, Nicolas Resch, Noga Ron-Zewi, Shashwat Silas, and Mary Wootters. LDPC codes achieve list decoding capacity. *arXiv preprint arXiv:1909.06430*, 2019.

12 Atri Rudra and Mary Wootters. Every list-decodable code for high noise has abundant near-optimal rate puncturings. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 764–773. ACM, 2014.

13 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.

14 Rom Rubenovich Varshamov. Estimate of the number of signals in error correcting codes. *Docklady Akad. Nauk, SSSR*, 117:739–741, 1957.

15 Mary Wootters. On the list decodability of random linear codes with large error rates. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 853–860, 2013. doi:10.1145/2488608.2488716.

16 Jack Wozencraft. List decoding. *Quarter Progress Report*, 48:90–95, 1958.

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