Coding Theory

July 17, 2020

1 Introduction

Roughly speaking, coding theory studies properties and algorithmic questions of subsets of strings $\mathcal{C} \subseteq \Sigma^n$ over a finite alphabet Σ, where a subset $\mathcal{C}$ is appropriately denoted a *code* and its elements are denoted as *codewords*. Despite this seemingly simplistic description, to really "unlock" interesting properties one usually brings to bear way more structure such as algebraic properties of low degree polynomials (in which case Σ is taken to be a finite field $\mathbb{F}$) or combinatorial properties of expander graphs. Codes are usually designed with a given error model in mind. Are symbols erased, changed, inserted and/or deleted? For instance, are corruptions caused adversarially with access to the entire codeword or done probabilistically and independently one symbol at time?

Probabilistic code constructions typically enjoy great success with several examples achieving optimal/nearly optimal parameters, and for this reason probabilistic methods have flourished within coding theory. However, it can be notoriously hard to find explicit [1] constructions. Why should one bother? One difficulty may be that with high probability a sampled random code could have all the desired properties, but certifying that a given one indeed does might be computationally hard. Figuratively, this line of coding theory research is trying to find hay in a haystack and it can be surprisingly difficult to do so.

Besides the quest for explicitness in code constructions, one might also ask for efficient algorithms that can convert a message into a codeword of $\mathcal{C}$ and/or conversely given a word in Σ^n (possibly satisfying some closeness assumption to $\mathcal{C}$) efficiently finds the closest (or a list of closest) codewords in $\mathcal{C}$. This search for algorithmic efficiency goes under the umbrella of algorithmic coding theory. In contrast, combinatorial coding theory seeks to understand bounds on what is achievable or impossible regardless of algorithmic efficiency or explicitness.

The development of Probabilistically Checkable Proofs (PCPs) has placed a strong emphasis on local properties of codes. An example is the need to test whether a function (given as an evaluation table) is a low-degree multivariate polynomial by querying it only at a few entries. In other words, this is asking whether Reed–Muller codes can be locally tested. The search for codes admitting local properties has motivated many developments in the past three decades and it is still a very active field within coding theory with longstanding open problems.

Another central notion in modern coding theory is *list decoding*. It is a relaxed decoding model, where given an arbitrary word in Σ^n we seek all codewords of a code, say $\mathcal{C}$, within a given radius ρ of it. The difference is that ρ can be taken to be much larger than the unique decoding radius of the code $\mathcal{C}$, in which case there may be a list of codewords rather than a single one. This concept is not new dating back to Elias [Eli57]. However, it was made algorithmic much more recently

[1] Explicit means that the construction can be done in polynomial time in n.

with a breakthrough result of Sudan [Sud97] for list decoding Reed–Solomon codes. Subsequently, the algorithmic list decoding radius was improved by Guruswami–Sudan [GS98] all the way to the Johnson bound and finally to the best possible by Guruswami–Rudra [GR08] (achieving *capacity* for the so-called folded Reed–Solomon codes). This concept is now widely used in coding theory used even to the standard unique decoding regime. Moreover, list decoding has found applications outside such as worst-case to average-case hardness among others.

2 Expander and Codes

The pseudorandom properties of expander graphs can be very useful in coding theory. Expanders are usually employed to either perform distance amplification or to define the parity check matrix of a code. In some cases, decoding algorithms of expander based codes are very efficient (specially compared to some algebraic decoding methods). In this section, we mention only two older but very representative results (see some other sections in this document and the HDX document) for more references.

2.1 Distance Amplification

Expander graphs can be used to boost the distance of codes.

- In [ABN$^+$92], Alon et al. show how the pseudorandom properties of expander graphs can be used for distance amplification.

2.2 Parity Check Matrix

The adjacency relation of an expander graph can be interpreted as a parity check matrix.

- In [SS96], Sipser and Spielman use the adjacency relation of bipartite (lossless) exapander graphs to define the parity check matrix of good binary codes. They also provide a very efficient decoding algorithm for their codes.

Open Problem 2.1. *Can some HDX (possibly algebraic) yield good expander codes? Note that ensuring constant rate seems quite challenging.*

2.3 Explicit Constructions

2.4 Large Alphabet

Thanks to a host of algebraic objects and techniques suitable for large fields, codes overs large alphabets are reasonably well-understood (specially compared to the binary case for which several mysteries remain).

2.5 Capacity Achieving Codes

Intuitively, codes of distance $1 - R$ can have rate at most R. Codes of rate R that can be combinatorially list decoded within radius $1 - R - \epsilon$ (for arbitrarily small $\epsilon > 0$) are known as capacity achieving codes. Over larger alphabets, there are known explicit constructions of such capacity achieving codes which also admit efficient list decoding algorithms.

- In a breakthrough result [GR08], Guruswami and Rudra show that starting from the well-known Reed–Solomon (RS) codes and using a folding operation [2] yields codes achieving capacity. These codes were dubbed folded-RS codes.
- Kopparty [Kop15] found another explicit construction achieving capacity which is also algebraic. It uses the evaluations of a polynomial as well as its derivatives in a construction known as multiplicity codes.
- In [GW13], Guruswami and Wang greatly simplify and speedup the folded-RS decoding. This simplification is achieved by using linear algebraic methods instead of more complex root-finding procedures over extension fields.
- In [KRSW18], Kopparty et al. refine the list decoding parameters of folded-RS codes.

2.6 Binary Alphabet

Contrary to large alphabet codes, binary codes as not as well understood.

- In [TS17], Ta-Shma gives a breakthrough explicit construction of binary codes achieving nearly optimal distance versus rate trade-off the so-called Gilbert–Varshamov bound. This construction is based on a "higher-order" version of the celebrated zig-zag product [RVW00].
- Many binary code results are obtained by code concatenation starting from powerful code constructions over large alphabets. Unfortunately, no such explicit construction is known yielding optimal or near optimal distance versus rate trade-off.

Open Problem 2.2 (☕☕☕☕...☕)**.** *Find explicit binary codes of rate $\Omega(\epsilon^2)$ efficiently list decodable within radius $1/2 - \epsilon$.*

3 Locally Testable Codes

Locally Testable Codes (LTCs) are codes $\mathcal{C}$ whose membership can be probabilistically tested by reading only a constant [3] number of symbols of a purported codeword x (non-codewords are rejected with probability proportional to their distance from $\mathcal{C}$).

Open Problem 3.1 (☕☕...☕)**.** *Construct* good *LTCs, i.e., LTCs of constant relative distance and constant rate.*

- In [BSGH$^+$04], Ben-Sasson et al. show, in particular, an interesting connection between PCPs of Proximity (PCPP) and LTCs.
- In [Din06], Dinur besides giving a combinatorial proof of the PCP theorem also gives a LTC of block length $O(t \cdot \text{polylog}(t))$ using a PCPP strengthening of her result.
- In [Mei09], Meir gives a combinatorial proof of LTCs of block length $O(t \cdot \text{polylog}(t))$ (recall that Dinur's proof achieving this near linear block length started from a suitable algebraic construction of PCPPs).

[2] Actually, it is an interpretation rather than an operation.

[3] The non constant query regime is also interesting.

- In [DDHRZ20], Dikstein et al. propose an approach towards good LTC using HDXs. They also use HDX as a unifying language for LTCs.

4 Locally Decodable Codes

Locally Decodable Codes (LDCs) are codes whose symbols of the original message can be probabilistically decoded from a corrupted codeword (under some distance assumption) by reading only a constant [4] number of symbols.

Open Problem 4.1. *Reduce the gap between upper and lower bounds on the rate of LDCs.*

4.1 Constructions

- In [CY20], Cohen and Yankovitz use expanders to perform query-efficient distance amplification of LDCs.
- In [Efr12], Efremenko gives a framework for constructing locally decodable codes from irreducible representations.

4.2 Lower Bounds

- In [BCG19], Bhattacharyya et al. give a direct combinatorial proof of a known lower boud on 3-query LDC.
- In [BDSS11], Bhattacharyya et al. give a lower bound for 2-query LCCs. A LCC are closely related to LDC, but its definition only asks the local decoding of a codeword symbol rather than a message symbol.
- Katz–Trevisan [KT00] lower bounds on LDCs ruling out the existence of good LDCs.

5 Randomized Analysis

The use of randomized analysis to derive bounds on code parameters has greatly evolved with the use of more sophisticated methods such as the chaining argument [Tal05]. Currently, we know very sharp bounds on the list sizes of random codes.

- In [MRRZ$^+$19], Mosheiff et al. show that random ensembles of LDPCs achieve list decoding capacity.
- In [GLM$^+$20], Guruswami et al. obtain very sharp bounds for the list-decoding and list-recovery of random linear codes.
- In [RW14], Rudra and Wootters show that randomly puncturing a code of sufficiently large distance yields w.h.p. combinatorially list decodable codes from large list decoding radius. They use the chaining method (which is a sophisticated form of union bound).
- In [Woo13], Wootters give list size bound for random linear codes of large distance.

[4] Similarly to LTCs, the non constant query regime is also interesting.

6 Limits on Binary Codes

As surprising as it may sound we still do not know the precise rate upper bound for binary codes given their distance.

Open Problem 6.1 (☕☕☕☕. . . ☕)**.** *What is the maximum rate of binary codes of distance d?*

- In [MRRW77], McEliece et al. derive rate upper bounds for binary codes using linear programming and identities involving orthogonal polynomials naturally arising in the Hamming scheme [Del75].
- In [FT05], Friedman and Tillich use spectral graph theory and notions of algebraic topology as an approach to obtain rate upper bounds.
- In [NS05], Navon and Samorodnitsky show that for large distances the bounds of MRRW based on linear programming are almost tight (of course this does not rule out tighter relaxations). They also derive the first linear programming bound of MRRW using a simpler Fourier analytic proof.
- In [Alo09], Alon gives a rate upper bound of $O(\epsilon^2 \log(1/\epsilon))$ for ϵ-balanced codes using a rank lower bound for diagonally dominant matrices.

7 Shannon Error Model

In the Shannon error model, errors are caused probabilistically and independently. Think of a communication channel that flips every transmitted bit with a fixed probability p. An extremely elegant mathematical theory of information was developed to understand this kind of model and some probabilistic code constructions of the original theory took decades to see explicit counterparts.

- In the landmark paper [Sha48], Shannon lays the mathematical foundation of information theory and communication.
- After 60 years after Shannon result [Sha48], Arikan [Ari09] introduces the so-called polar codes which are explicit codes asymptotically achieving capacity.
- In [GRY20], Guruswami et al. explore larger *kernels* to achieve a near optimal convergence to capacity.

Open Problem 7.1. *Can the encoding time of [GRY20] be improved?*

8 Beyond the Johnson Bound

The Johnson bound is a generic result about the combinatorial list decodability of any code. It establishes list decoding radii and list size bounds based only on the code distance. In essence, it says that codes are combinatorially list decodable within arbitrarily large radius provided their distance is sufficiently large.

- In [ST20], Shangguan and Tamo show that explicit Reed–Solomon codes with exponentially large field sizes can beat the Johnson bound.
- In [BKR10], Ben-Sasson et al. show limitations on the list decoding of Reed-Solomon codes which uses all field elements in their evaluations.
- In [DGKS08], Dinur et al. show decodability of group homomorphism codes beyond the Johnson bound.

Open Problem 8.1. *Reduce the field size to explicit Reed-Solomon codes beating the Johnson bound and improving on [ST20].*

9 Beating the Gilbert–Varshamov Bound

The Gilbert–Varshmov (GV) bound [Gil52, Var57] is a trade-off between distance and rate of code achieved by random codes. For every $q \geq 49$, algebraic geometry codes are known beating the GV for some distance interval. This is established by Tsfasman-Vladut-Zink bound [Sti08, Chapter 8].

10 Handling Insertion/Deletions

Naturally, the error model crucially determines the nature of the code. To handle insertions and deletions completely different constructions are needed (with some built-in notions of varying symbol "frequency" to implicitly give a spatial anchoring of symbols).

- In [GHS20], Guruswami et al. design optimally resilient codes for list decoding from insertion and deletions.

11 Hardness of Approximation and Coding Theory

A big consumer (and also producer) of coding theory results is the field of hardness of approximation. This is specially true for codes admitting some local property.

- In [BGH$^+$12], Barak et al. use coding theoretic results to prove that the short-code graph is a small set expander with high threshold rank.
- In [ALM$^+$98], Arora et al. heavily use error correcting codes to prove the PCP theorem.

12 Hardness of Coding Theory Tasks

Several coding theory tasks can be computationally very hard in the worst-case such as decoding and certifying the minimum distance.

- In [SV19], Stephens-Davidowitz and Vaikuntanathan obtain very strong hardness results for coding theory tasks under the SETH.

Open Problem 12.1. *How much can the SETH assumption of [SV19] be weakened?*

13 Codes for Distributed Storage

Coding schemes for distributed storage requires new insights and lead to nice combinatorics.

- In [KLR17], Kane–Lovett–Rao give alphabet lower bounds for a family of distributed storage codes (maximally recoverable codes on a grid-like topolgy). Their proof uses a beautiful represention theoretic argument.

14 Books

- Essential Coding Theory by Guruswami, Rudra and Sudan [GRS19].
- Algebraic Function Fields and Codes by Stichtenoth [Sti08].
- Introduction to Coding Theory by van Lint [vL99].

References

[ABN+92] N. Alon, J. Bruck, J. Naor, M. Naor, and R. Roth. Construction of asymptotically good, low-rate error-correcting codes through pseudo-random graphs. *IEEE Transactions on Information Theory*, 28:509–516, 1992.

[ALM+98] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *J. ACM*, 45(3):501555, May 1998.

[Alo09] Noga Alon. Perturbed identity matrices have high rank: Proof and applications. *Comb. Probab. Comput.*, 18(12):315, 2009.

[Ari09] E. Arikan. Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Transactions on Information Theory*, 55(7):3051–3073, July 2009.

[BCG19] Arnab Bhattacharyya, L. Sunil Chandran, and Suprovat Ghoshal. Combinatorial lower bounds for 3-query ldcs, 2019.

[BDSS11] A. Bhattacharyya, Z. Dvir, A. Shpilka, and S. Saraf. Tight lower bounds for 2-query lccs over finite fields. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 638–647, 2011.

[BGH+12] Boaz Barak, Parikshit Gopalan, Johan Håstad, Raghu Meka, Prasad Raghavendra, and David Steurer. Making the long code shorter. In *Proceedings of the 53rd IEEE Symposium on Foundations of Computer Science*, pages 370–379, 2012.

[BKR10] E. Ben-Sasson, S. Kopparty, and J. Radhakrishnan. Subspace polynomials and limits to list decoding of reedsolomon codes. *IEEE Transactions on Information Theory*, 56(1):113–120, 2010.

[BSGH$^+$04] Eli Ben-Sasson, Oded Goldreich, Prahladh Harsha, Madhu Sudan, and Salil Vadhan. Robust pcps of proximity, shorter pcps and applications to coding. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing*, STOC 04, page 110, 2004.

[CY20] Gil Cohen and Tal Yankovitz. Rate amplification and query-efficient distance amplification for locally decodable codes, 2020.

[DDHRZ20] Yotam Dikstein, Irit Dinur, Prahladh Harsha, and Noga Ron-Zewi. Locally testable codes via high-dimensional expanders, 2020.

[Del75] P. Delsarte. The association schemes of coding theory. In *Combinatorics*, pages 143–161. Springer Netherlands, 1975.

[DGKS08] Irit Dinur, Elena Grigorescu, Swastik Kopparty, and Madhu Sudan. Decodability of group homomorphisms beyond the johnson bound. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC 08, page 275284, 2008.

[Din06] Irit Dinur. The PCP theorem by gap amplification. In *Proc. 38th ACM Symp. on Theory of Computing*, pages 241–250, 2006.

[Efr12] Klim Efremenko. From irreducible representations to locally decodable codes. In *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing*, STOC 12, page 327338, 2012.

[Eli57] Peter Elias. List decoding for noisy channels, 1957.

[FT05] Joel Friedman and Jean-Pierre Tillich. Generalized alon–boppana theorems and error-correcting codes. *SIAM J. Discret. Math.*, page 700718, July 2005.

[GHS20] Venkatesan Guruswami, Bernhard Haeupler, and Amirbehshad Shahrasbi. Optimally resilient codes for list-decoding from insertions and deletions. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 524537, 2020.

[Gil52] E.N. Gilbert. A comparison of signalling alphabets. *Bell System Technical Journal*, 31:504–522, 1952.

[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, 2020.

[GR08] V. Guruswami and A. Rudra. Explicit codes achieving list decoding capacity: Error-correction with optimal redundancy. *IEEE Transactions on Information Theory*, 54(1):135–150, 2008.

[GRS19] Venkatesan Guruswami, Atri Rudra, and Madhu Sudan. Essential coding theory. Available at `https://cse.buffalo.edu/faculty/atri/courses/coding-theory/book/index.html`, 2019.

[GRY20] Venkatesan Guruswami, Andrii Riazanov, and Min Ye. Arikan meets shannon: Polar codes with near-optimal convergence to channel capacity. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 552564, 2020.

[GS98] V. Guruswami and M. Sudan. Improved decoding of reed-solomon and algebraic-geometric codes. In *Proceedings 39th Annual Symposium on Foundations of Computer Science*, pages 28–37, 1998.

[GW13] V. Guruswami and C. Wang. Linear-algebraic list decoding for variants of reedsolomon codes. *IEEE Transactions on Information Theory*, 59(6):3257–3268, 2013.

[KLR17] D. Kane, S. Lovett, and S. Rao. The independence number of the birkhoff polytope graph, and applications to maximally recoverable codes. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 252–259, 2017.

[Kop15] Swastik Kopparty. List-decoding multiplicity codes. *Theory of Computing*, 11(5):149–182, 2015.

[KRSW18] S. Kopparty, N. Ron-Zewi, S. Saraf, and M. Wootters. Improved decoding of folded reed-solomon and multiplicity codes. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 212–223, 2018.

[KT00] J. Katz and L. Trevisan. On the efficiency of local decoding procedures for error correcting codes. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.

[Mei09] Or Meir. Combinatorial construction of locally testable codes. *SIAM J. Comput.*, page 491544, July 2009.

[MRRW77] R. McEliece, E. Rodemich, H. Rumsey, and L. Welch. New upper bounds on the rate of a code via the Delsarte-MacWilliams inequalities. *IEEE Transactions on Information Theory*, 23(2):157–166, 1977.

[MRRZ$^+$19] Jonathan Mosheiff, Nicolas Resch, Noga Ron-Zewi, Shashwat Silas, and Mary Wootters. Ldpc codes achieve list decoding capacity, 2019.

[NS05] M. Navon and A. Samorodnitsky. On delsarte's linear programming bounds for binary codes. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pages 327–336, 2005.

[RVW00] O. Reingold, S. Vadhan, and A. Wigderson. Entropy waves, the zig-zag graph product, and new constant-degree expanders and extractors. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 3–13, Nov 2000.

[RW14] 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*, STOC 14, page 764773, 2014.

[Sha48] Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948.

[SS96] M. Sipser and D. A. Spielman. Expander codes. *IEEE Transactions on Information Theory*, 42(6):1710–1722, 1996.

[ST20] Chong Shangguan and Itzhak Tamo. Combinatorial list-decoding of reed-solomon codes beyond the johnson radius. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 538551, 2020.

[Sti08] Henning Stichtenoth. *Algebraic Function Fields and Codes*. Springer Publishing Company, Incorporated, 2nd edition, 2008.

[Sud97] Madhu Sudan. Decoding of reed solomon codes beyond the error-correction bound. *J. Complex.*, 13(1):180193, March 1997.

[SV19] N. Stephens-Davidowitz and V. Vaikuntanathan. Seth-hardness of coding problems. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 287–301, 2019.

[Tal05] Michel Talagrand. *The Generic Chaining*. Springer-Verlag Berlin Heidelberg, 2005.

[TS17] Amnon Ta-Shma. Explicit, almost optimal, epsilon-balanced codes. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, Proceedings of the 49th ACM Symposium on Theory of Computing, pages 238–251. ACM, 2017.

[Var57] R.R. Varshamov. Estimate of the number of signals in error correcting codes. *Doklady Akademii Nauk SSSR*, 117:739–741, 1957.

[vL99] Jacobus H. van Lint. *Introduction to Coding Theory*. Springer-Verlag, 1999.

[Woo13] Mary Wootters. On the list decodability of random linear codes with large error rates. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC 13, page 853860, 2013.