Lecturers: Hung Q. Ngo and Atri Rudra

Strongly explicit constructions of list-disjunct matrices from randomness extractors and expanders

We have seen two methods of constructing efficiently decodable list-disjunct matrices from (not necessarily efficiently decodable) list-disjunct matrices. The first method is a recursive method. The second method is based on list-recoverable codes. The idea of constructing list-disjunct matrices from list-recoverable codes were conceived in [3] and [1] independently, though through slightly different routes. Both of these methods assume the existence of a family of list-disjunct matrices of "few" rows. If this family is (strongly) explicit, then the final efficiently decodable list-disjunct matrices is (strongly) explicit. In this lecture, we describe two methods of constructing list-disjunct matrices in a strongly explicit way. The first method is based on randomness extractors, first proposed in Cheraghchi [1]. The second method is based on expanders, first proposed in Indyk-Ngo-Rudra [3]. The second method is slightly worse in terms of the number of tests; however, it allows for a more precise control of the list size which is important in some applications.

1 List-disjunct matrices from extractors

This construction is from [1]. Randomness extractors are functions which "convert" biased and correlated random bits into almost uniform random bits. Extractors have numerous applications in (theoretical) Computer Science[1]. In this section, we use extractors to explicitly construct good list-disjunct matrices.

1.1 Min entropy and variational distance

Let $\mathcal{D}$ be a distribution on a finite sample space Ω. The *min entropy* of $\mathcal{D}$ is defined to be

$$H_\infty(\mathcal{D}) := \min_{\omega\in\Omega} \min\left\{\log_2 \frac{1}{\text{Prob}_{\mathcal{D}}[\omega]}\right\}.$$

Here $\text{Prob}_{\mathcal{D}}[\omega]$ is the probability mass which the distribution $\mathcal{D}$ assigns to ω. From the definition, if $H_\infty(\mathcal{D}) \geq k$ then $\text{Prob}_{\mathcal{D}}[\omega] \leq 1/2^k$ for every $\omega \in \Omega$.

The *total variational distance* between two distributions $\mathcal{P}$ and $\mathcal{Q}$ on Ω is defined to be

$$\|\mathcal{P} - \mathcal{Q}\|_{\text{TV}} = \max_{A\subseteq\Omega} |\mathcal{P}(A) - \mathcal{Q}(A)| = \frac{1}{2}\sum_{\omega\in\Omega} |\mathcal{P}(\omega) - \mathcal{Q}(\omega)|.$$

The first equality is the definition of total variational distance. The second equality can be derived from the definition. Two distributions are said to be ϵ-*close* if their variational distance is at most ϵ.

[1] http://people.seas.harvard.edu/~salil/pseudorandomness/extractors.pdf

1.2 Condensers and extractors

For any positive integer n, let $\mathcal{U}_n$ denote the uniform distribution on $\mathbb{F}_2^n$.

Let a, b, m be positive integers. A function $C : \mathbb{F}_2^a \times \mathbb{F}_2^b \to \mathbb{F}_2^m$ is called a *strong* $k \to_\epsilon k'$ *condenser* if it satisfies the following:

- for every distribution $\mathcal{A}$ on $\mathbb{F}_2^a$ with min entropy $H_\infty(\mathcal{A}) \geq k$
- for any random variable $A \sim \mathcal{A}$
- any "seed" variable $B \sim \mathcal{U}_b$
- the distribution of $(B, C(A, B))$ is ϵ-close to some distribution $(\mathcal{U}_b, \mathcal{Z})$ on $\mathbb{F}_2^{b+m}$ with min entropy at least $b + k'$.

Here, the quantity ϵ is called the *error*, the quantity $k - k'$ is called the *entropy loss*, and $m - k'$ is called the *overhead* of the condenser. A *lossless condenser* is a condenser with no entropy loss. A *strong* (k, ϵ)*-extractor* is a condenser with no overhead.

The intuition behind the above definitions are as follows. Suppose we have a "weak" random source which gives a random variable A on $\mathbb{F}_2^a$ (i.e. a random bits). The random bits from A are not necessarily uniform, but the total entropy is at least k. We have access to a small number b of truly uniform random bits represented by B. From A and B, we would like to "extract" as many uniform random bits as possible. In the ideal case, we want to extract $b + k$ uniform random bits, because there is certainly enough entropy (i.e. randomness) to do so. However, the task is not easy and we have to settle for $b + k'$ uniform random bits with $k' \leq k$. Thus, $k - k'$ is called the entropy loss. Before extracting the uniform random bits, we may want to "condense" the randomness down to $m \geq k'$ bits, which hopefully will make uniform bit extraction easier. The condensation is done via the function C, and thus $m - k'$ is called the overhead.

1.3 Codes from condensers/extractors

From a function $C : \mathbb{F}_2^a \times \mathbb{F}_2^b \to \mathbb{F}_2^m$ we can define the corresponding *induced code* $\mathcal{I}(C)$ as follows. This code has alphabet $\Sigma = \mathbb{F}_2^m$, length $n = 2^b$, and size (i.e., number of codewords) $N = 2^a$. For any $A \in \mathbb{F}_2^a$, the Ath codeword of the code is defined to be the vector whose Bth component is $C(A, B)$, where the components of the codewords are indexed by $B \in \mathbb{F}_2^b$.

For any finite alphabet Σ and a positive integer n, a sequence $S = (S_1, \cdots, S_n)$ where $\emptyset \neq S_i \subseteq \Sigma$ is called a *mixture* on Σ^n. For any word $\mathbf{w} = (w_1, \ldots, w_n) \in \Sigma^n$, the *agreement of* $\mathbf{w}$ *with* S is defined to be

$$\text{Agr}(\mathbf{w}, S) := \frac{|\{i \in [n] \mid w_i \in S_i\}|}{n}.$$

Define $\rho(S)$ to be the expected agreement of a randomly chosen word in Σ^n with S, namely

$$\rho(S) := \frac{|S_1| + \cdots + |S_n|}{n|\Sigma|}.$$

Consider any code $C \subseteq \Sigma^n$ and $\alpha \in [0, 1]$. The list of codewords whose agreement with S is *more than* α is denoted by $\text{LIST}_C(S, \alpha)$. When $\alpha = 1$, the list consists of codewords with 100% agreements. More precisely,

$$\text{LIST}_C(S, \alpha) := \begin{cases} \{\mathbf{w} \in \Sigma^n \mid \text{Agr}(\mathbf{w}, S) > \alpha\} & \alpha < 1 \\ \{\mathbf{w} \in \Sigma^n \mid \text{Agr}(\mathbf{w}, S) = 1\} & \alpha = 1. \end{cases}$$

The following important theorem was first observed in [5] (see also [2]).

Theorem 1.1. *Let $C : \mathbb{F}_2^a \times \mathbb{F}_2^b \to \mathbb{F}_2^m$ be a $k \to_\epsilon k'$ condenser. For any mixture S of $(\mathbb{F}_2^m)^{2^b}$, if $\rho(S)2^{m-k'} + \epsilon < 1$ then*

$$\text{LIST}_{\mathcal{I}(C)}(S, \rho(S)2^{m-k'} + \epsilon) < 2^k.$$

Proof. Assume to the contrary that $\text{LIST}_{\mathcal{I}(C)}(S, \rho(S)2^{m-k'} + \epsilon) \geq 2^k$. Let $A = (A_1, \cdots, A_{2^b})$ be a random codeword uniformly chosen from $\text{LIST}_{\mathcal{I}(C)}(S, \rho(S)2^{m-k'} + \epsilon)$. Then, as a distribution on $\mathcal{F}_2^a$ the distribution of the random variable A has min-entropy at least k. Let $B \sim \mathcal{U}_b$ be a uniformly distributed random variable chosen from $\mathbb{F}_2^b$. (Think of B as a random position of a codeword.) Then, from the fact that C is a strong $k \to_\epsilon k'$ condenser, the distribution of $(B, C(A, B)) = (B, A_B)$ is supposed to be ϵ-close to *some* distribution on $\mathbb{F}_2^{b+m}$ with min entropy at least $b + k'$. We will show that such is not the case, reaching a contradiction.

Let $\mathcal{D}$ be an arbitrary distribution on $\mathbb{F}_2^{b+m}$ with min-entropy at least $b + k'$. We shall show that $\mathcal{D}$ and the distribution of (B, A_B) are not ϵ-close by specifying an event on $\mathcal{F}_2^{b+m}$ for which the two distributions differ by more than ϵ.

The event we want is defined by a function $f : \mathbb{F}_2^{b+m} \to \{0, 1\}$, where for $i \in \mathbb{F}_2^b$ and $X \in \mathbb{F}_2^m$, we define $f(i, X) = 1$ iff $X \in S_i$. We next estimate the probabilities that the distribution $\mathcal{D}$ and the distribution of (B, A_B) assign to the event f.

First, consider a random point (B, A_B). (Intuitively, we picked a uniformly random coordinate B of a random codeword A chosen as above.)

$$\Pr_{A,B}[f(B, A_B) = 1] = \Pr_{A,B}[A_B \in S_B] = \text{Agr}A, S > \rho(S)2^{m-k'} + \epsilon.$$

Second, consider the distribution $\mathcal{D}$ which has min-entropy at least $b + k'$.

$$\begin{aligned}
\Pr_{\mathcal{D}}[f(i, X) = 1] &= \sum_{\substack{(i,X) \in \mathbb{F}_2^b \times \mathbb{F}_2^m \\ X \in S_i}} \Pr_{\mathcal{D}}[(i, X)] \\
&\leq \frac{1}{2^{b+k'}} \sum_{i \in \mathbb{F}_2^b} |S_i| \\
&= \rho(S)2^{m-k'}.
\end{aligned}$$

□

From the above theorem, Cheraghchi [1] observed the following, which was the main result in that paper. The language that Cheraghchi used was not code concatenation and he did not use the term *list-separable/disjunct*, but we can easily see the analogy. It is also not hard to see that the basic idea is viewing the induced code of a condenser as as list-recoverable code with list size 2^k.

Corollary 1.2. *Let $C : \mathbb{F}_2^a \times \mathbb{F}_2^b \to \mathbb{F}_2^m$ be a strong $k \to_\epsilon k'$ condenser. Then, the concatenation of $\mathcal{I}(C) \circ \text{ID}_{2^m}$ is a $(d, 2^k - d)$-list-separable matrix for any d satisfying the following constraints: $d \leq 2^m$, $d \leq (1 - \epsilon)2^{k'}$. The matrix has $t = 2^{b+m}$ rows and $N = 2^a$ columns. Furthermore, the total decoding time is $O(2^{a+b+m})$.*

Proof. We simply specify a decoding algorithm. We decode a set S_i for each position $i \in \mathbb{F}_2^b$. Note that $|S_i| \leq d$ for each i because there are at most d positives. Thus, $\rho(S) = \sum_{i \in \mathbb{F}_2^b} |S_i|/2^{b+m} \leq d/2^m$. From Theorem 1.1 we know

$$\textsc{list}_{\mathcal{C}}(S, (d/2^m)2^{m-k'} + \epsilon) = \textsc{list}_{\mathcal{C}}(S, d/2^{k'} + \epsilon) < 2^k.$$

Furthermore, all positive items correspond to codewords with 100% agreement with S. Because $1 \geq d/2^{k'} + \epsilon$, the codewords corresponding to positive items all belong to $\textsc{list}_{\mathcal{C}}(S, d/2^{k'} + \epsilon)$ which means we output less than 2^k codewords (including the false positives). The algorithm is simply to output all such codewords. The running time is $O(2^{a+b+m})$, and the number of codewords outputted is less than 2^k. □

Now that we know of a way to convert a condenser into a list-disjunct matrix, we look for known explicit constructions of condensers with favorable parameters. One such construction was given in [2].

Theorem 1.3 (Explicit extractor from [2])**.** *For integers $a \geq k$, $\epsilon > 0$, there exists an explicit strong (k, ϵ)-extractor $Ext : \mathbb{F}_2^a \times \mathbb{F}_2^b \to \mathbb{F}_2^m$ with $m = k - 2\log(1/\epsilon) - O(1)$ and $b = \log a + O(\log k \cdot \log(k/\epsilon))$.*

Recall that a strong (k, ϵ)-extractor is a strong $k \to_\epsilon k'$ condenser with $m = k'$. From the above theorem and Corollary 1.2, we obtain the following construction.

Theorem 1.4. *Let $1 \leq d \leq N$ be integers. Then, there exists a strongly-explicit $t \times N$ matrix $\mathbf{M}$ that is $(d, O(d))$-list-disjunct with $t = O(d^{1+o(1)} \log N)$ rows.*

Proof. Fix small $\epsilon > 0$. Let $k' = m$ be the least positive integer suchthat $d \leq (1 - \epsilon)2^m$. Let C be the (k, ϵ)-extractor from Theorem 1.3, where we choose $a \approx \log N$ and $k = \log d + 2\log(1/\epsilon) + O(1)$ (so that $d \leq 2^m$). Then, the concatenated code $\mathcal{I}(C) \circ \mathrm{ID}_{2^m}$ is certainly strongly explicit. And, the corresponding matrix by Corollary 1.2 is $(d, O(d))$-list-separable with N columns and $t = 2^{b+m}$ rows. Note that

$$\begin{aligned} t &= 2^{b+m} \\ &= 2^{\log a + O(\log k \cdot \log(k/\epsilon)) + k - 2\log(1/\epsilon) - O(1)} \\ &= O(\epsilon^2) \cdot (d \log N) \cdot k^{O(\log(k/\epsilon))} \\ &= (d \log N) \cdot (\log d)^{O(\log \log d))} \\ &= O(d^{1+o(1)} \log N). \end{aligned}$$

□

Finally, combining the above theorem with the two construction methods we have discussed, we obtain the following results. (**Note to students:** whoever present the following corollaries should work out the details. They are all mechanical, but instructive! This forces you to read and understand previous lectures.)

Corollary 1.5 (Combination with PV^s-based method, Section 3 of Lecture 10)**.** *Let $\epsilon > 0$ be a real number and let $1 \leq d \leq N$ be integers. Then there exists a strongly-explicit $t \times N$ matrix that is $(d, (1/\epsilon)^{O(1/\epsilon)} \cdot d^{1+\epsilon})$-list-disjunct with $t = (1/\epsilon)^{O(1/\epsilon)} \cdot d^{1+\epsilon} \cdot \log N$ rows that can be decoded in time $t^{O(1/\epsilon)}$.*

Corollary 1.6 (Combination with recursive construction method, Section 2 of Lecture 9)**.** *Let $1 \leq d \leq N$ be integers. For any constant $\alpha \in (0, 1)$ there exists a strongly-explicit $t \times N$ matrix that is $(d, O(d))$-list disjunct with $t = O(d^{1+o(1)} \log N \log \log N)$ rows and can be decoded in* $\mathrm{poly}(t)$ *time.*

2 List-disjunct matrices from expanders

A W-left regular bipartite graph $[N] \times [W] \to [T]$ is called an (N, W, T, D, α)-*expander* if every subset $S \subset [N]$ of size at most D has a neighborhood, denoted by $\Gamma_G(S)$, of size at least $\alpha \cdot |S|$. The *neighborhood* of S is the set of all vertices which are adjacent to at least one vertex in S. The quantity α is called the *expansion rate* of this expander. Given such a bipartite expander G, consider the $T \times N$ incidence matrix $\mathbf{M}_G$ of G, which is the binary matrix whose rows are indexed by $[T]$ and whose columns are indexed by $[N]$, and there is a 1 in the (i, j) entry of the matrix if and only if (i, j) is an edge of G.

Proposition 2.1. *Let G be an $(N, w, t, d + \ell, \alpha)$-expander. If $\alpha > \frac{wd}{d+\ell}$ then $\mathbf{M}_G$ is a (d, ℓ)-list disjunct matrix with t rows and N columns.*

Proof. Recall that by definition a matrix $\mathbf{M}$ is (d, ℓ)-list disjunct if the following is true: for any two disjoint subsets S_1 and S_2 of columns of size d and ℓ respectively, there must be a row in $\mathbf{M}$ in which S_2 has a 1 but S_1 has all 0s.

When $\mathbf{M} = \mathbf{M}_G$, this property is equivalent to the following property on G. For every two subsets S_1 and S_2 of vertices in $[N]$ of size d and ℓ respectively, there is some vertex in $[T]$ which is adjacent to S_2 but not to S_1. In other words, $\Gamma_G(S_2) \setminus \Gamma_G(S_1) \neq \emptyset$.

We now argue that the property holds if G has an expansion rate of $(wd + 1)/(d + \ell)$. Because each vertex of G on the left has degree at most w, $\Gamma_G(S_1) \leq w|S_1| = wd$. On the other hand, because G is an $(N, w, t, d + \ell, (wd + 1)/(d + \ell))$-expander, and because $|S_1 \cup S_2| = d + \ell$, we know

$$|\Gamma(S_1 \cup S_2)| \geq \alpha(d + \ell) > \left(\frac{wd}{d + \ell}\right) \cdot (d + \ell) = wd \geq |\Gamma_G(S_1)|.$$

Hence, there must be at least one neighbor in $\Gamma(S_2)$ which does not belong to $\Gamma(S_1)$. □

Thus, from suitable expanders we can construct list-disjunct matrices. However known (explicit) constructions of expanders in the literature do not have the parameter range we prefer. This is mainly due to the fact that we prefer to minimize the number of rows T of the constructed matrix, while existing expanders were designed with different objectives (e.g., minimizing both W and T). For example, the following results are known.

Theorem 2.2 ([2])**.** *Let $\epsilon > 0$. There exists an* explicit *$(N_1, W_1, T_1, D_1, W_1(1 - \epsilon))$ expander with $T_1 \leq (4D_1)^{\log W_1}$ and $W_1 \leq 2(\log N_1) \cdot (\log D_1)/\epsilon$.*

Theorem 2.3 ([4])**.** *Let $\epsilon > 0$ be a* constant*. Then there exists an* explicit *$(N_2, W_2, T_2, D_2, W_2(1 - \epsilon))$-expander with $T_2 = O(D_2 W_2)$ and $W_2 = 2^{O(\log\log N_2 + (\log\log D_2)^3)}$.*

Fortunately, we can combine them to form families of expanders which have our range of parameters. The above two expanders can be combined using the following well known technique.

Proposition 2.4. *Let G_1 be an $(N, W_1, T_1, D, W_1(1-\epsilon))$-expander and G_2 be a $(T_1, W_2, T_2, DW_1, W_2(1-\epsilon))$-expander. Then from G_1 and G_2 we can construct an $(N, W_1W_2, T_2, D, W_1W_2(1 - 2\epsilon))$-expander G. Furthermore, if G_1 and G_2 are both explicit then so is G.*

Proof. The graph G is constructed by "concatenating" G_1 and G_2. In particular, construct the following intermediate tripartite graph G' (on the vertex sets $[N]$, $[T_1]$ and $[T_2]$ respectively), where one identifies $[T_1]$ once as the right vertex set for G_1 and once as the left vertex set of G_2. The final graph G is a bipartite graph on the "left vertices" $[N]$ and "right vertices $[T_2]$) where there is an edge from $i \in [N]$ to $j \in [T_2]$ if and only if there is a corresponding path of length 2 in the tripartite graph G'.

We verity that G is indeed an $(N, W_1W_2, T_2, D, W_1W_2(1 - 2\epsilon))$-expander:

- Consider a vertex $i \in [N]$: i can reach W_1 vertices in $[T_1]$, from each of them we can reach W_2 vertices in $[T_2]$. Hence, overall from i there are W_1W_2 paths of length 2 to $[T_2]$. Each of these paths correspond to an edge from i in the G. Hence, the left-degree of G is W_1W_2. (Note that there can be multi-edges in the sense that there might be more than one path in G' of length 2 from i to the same vertex j in $[T_2]$. But that's ok.)

- Now, consider a subset $S \subseteq [N]$ where $|S| \leq D$. Then, $|\Gamma_{G_1}(S)| \geq W_1(1-\epsilon)|S|$. Note that $\Gamma_{G_1}(S) \subseteq [T_1]$ and $|\Gamma_{G_1}(S)| \leq W_1 D$. Hence, $\Gamma_{G_1}(S)$ will "expand" at most by a factor of $W_2(1-\epsilon)$ in G_2. Overall, the number of vertices in $[T_2]$ which S can reach is at least

$$W_2(1-\epsilon)|\Gamma_{G_1}(S)| \geq W_2(1-\epsilon)W_1(1-\epsilon)|S| \geq W_1W_2(1-2\epsilon)|S|.$$

  Thus, the expansion rate of G is at least $W_1W_2(1-2\epsilon)$ as desired.

□

Next, we prove the following result by combining all the ingredients above. We genearlly want to construct $(d, \delta d)$-list-disjunct matrices for some fixed constant $1 \geq \delta > 0$. From Proposition 2.1, to do so we will need an expander where each subset of $(1+\delta)d$ vertices on the left expands by at least a factor of $\alpha > \frac{wd}{d(1+\delta)} = \frac{w}{1+\delta}$. Setting $\alpha = (1-\delta/3)w$ is sufficient because

$$(1-\delta/3)(1+\delta) = 1 + 2\delta/3 - \delta^2/3 = 1 + \delta(2/3 - \delta/3) \geq 1 + \delta/3 > 1.$$

So, if from what we know above we can construct explicitly an expander where each subset of at most $(1+\delta)d$ vertices on the left expands by a factor of at least $(1-\delta/3)w$ then we would get a $(d, \delta d)$-list-disjunct matrix explicitly.

Theorem 2.5. *For every $1 \leq d \leq N$ and $0 < \delta \leq 1$, there exists an explicit $(N, W, T, D = (1+\delta)d, (1-\delta/3)W)$-expander with $T = O(D \log N \cdot f(D, N))$, where*

$$f(D, N) = \frac{12}{\delta} \log D \cdot 2^{O((\log\log D + \log\log(12/\delta) + \log\log\log N)^3)}.$$

Note that $f(D, N) = (D \log N)^{o(1)}$ for a fixed δ.

We will set $D = (1+\delta)d$. By Theorem 2.2, there exists an explicit $(N, W_1, T_1, D, (1-\delta/6)W_1)$-expander, where

$$\begin{aligned} W_1 &\leq \frac{12}{\delta} \log N \log D, \\ T_1 &\leq (4D)^{\log W_1}. \end{aligned}$$

By Theorem 2.3, there exists an explicit $(T_1, W_2, T_2, D_2, (1-\delta/6)W_2)$-expander, where

$$\begin{aligned} D_2 &= DW_1 \\ W_2 &= 2^{O(\log\log T_1 + (\log\log D_2)^3)} \\ T_2 &= O(D_2 W_2). \end{aligned}$$

We bound W_2 first:

$$\begin{aligned} W_2 &= 2^{O(\log\log T_1 + (\log\log D_2)^3)} \\ &= 2^{O(\log\log D + \log\log W_1 + (\log\log D + \log\log W_1)^3)} \\ &= 2^{O((\log\log D + \log\log(12/\delta) + \log\log\log N)^3)}. \end{aligned}$$

From that we can bound T_2, which is big-O of

$$\begin{aligned} W_2 D_2 &\leq \frac{12D}{\delta} \log N \log D \cdot 2^{O((\log\log D + \log\log(12/\delta) + \log\log\log N)^3)} \\ &\leq D \log N \cdot f(D, N), \end{aligned}$$

as desired.

Corollary 2.6. *Let $1 \leq d \leq N$ be integers and $\delta > 0$ be any given constant. Then, there is an explicit $t \times N$ matrix that is $(d, \delta d)$-list disjunct with $t = (d \log N)^{1+o(1)}$ rows.*

References

[1] M. CHERAGHCHI, *Noise-resilient group testing: Limitations and constructions*, in FCT, 2009, pp. 62–73.

[2] V. GURUSWAMI, C. UMANS, AND S. P. VADHAN, *Unbalanced expanders and randomness extractors from parvaresh-vardy codes*, in Proceedings of the 22nd Annual IEEE Conference on Computational Complexity, 2007, pp. 96–108.

[3] P. INDYK, H. Q. NGO, AND A. RUDRA, *Efficiently decodable non-adaptive group testing*, in Proceedings of the Twenty First Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'2010), New York, 2010, ACM, pp. 1126–1142.

[4] A. TA-SHMA, C. UMANS, AND D. ZUCKERMAN, *Lossless condensers, unbalanced expanders, and extractors*, Combinatorica, 27 (2007), pp. 213–240.

[5] A. TA-SHMA AND D. ZUCKERMAN, *Extractor codes*, IEEE Transactions on Information Theory, 50 (2004), pp. 3015–3025.