

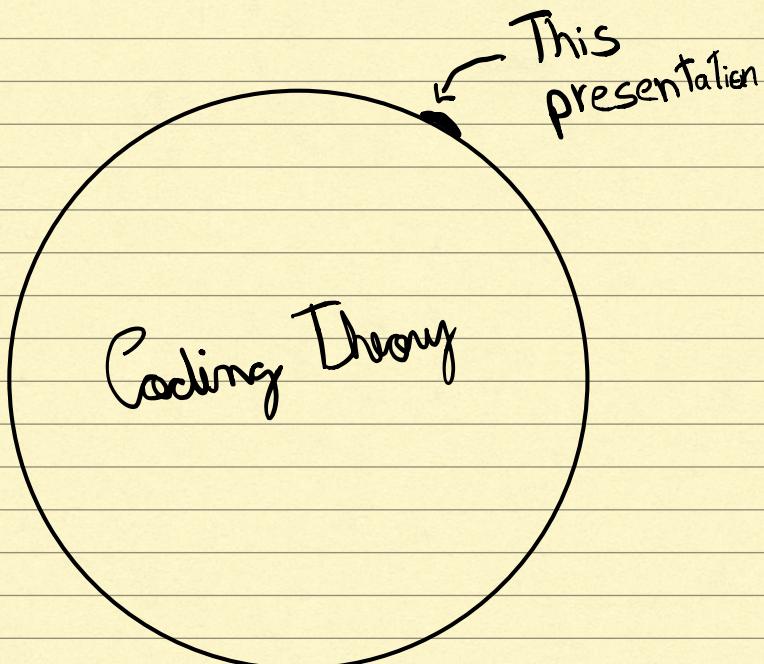
Some Open Problems in Coding Theory

Plan

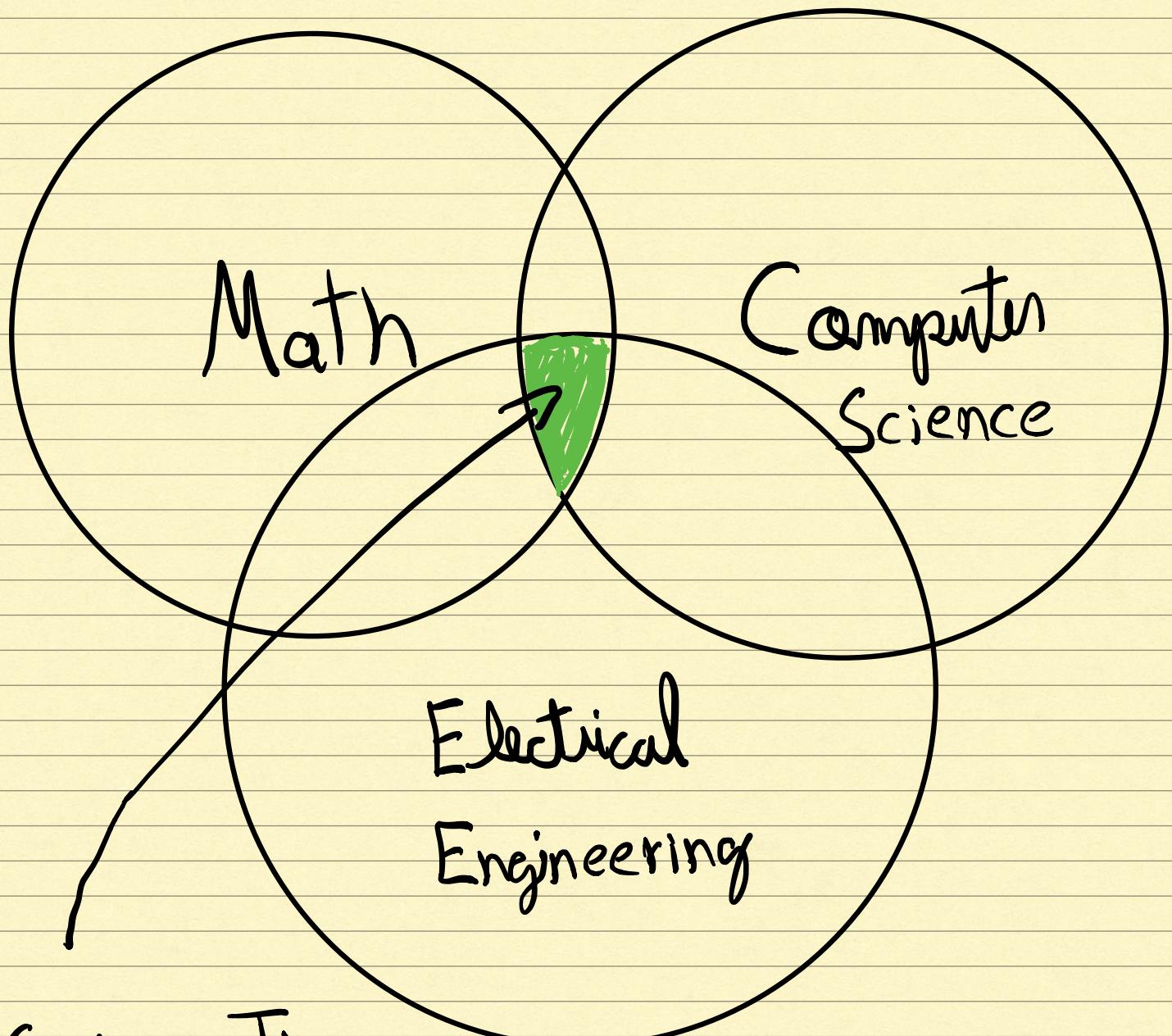
- Brief overview of some aspects
of Coding Theory
- (10 minute break)
- Open Problems

Disclaimer

- The choice of topics is biased towards my interests and knowledge of the field.
- The problems are open to the best of my knowledge.
- Some open problems are well-known longstanding problems (possibly challenging).



Coding Theory at a Macroscopic Scale



Coding Theory

Intersections

Combinatorics

Expanders/HDXs

SOS

Coding Theory

Local Algorithms
LDPCs

Hardness
PCPs

Algebraic
Methods

Highly
Parameter
Oriented

Probabilistic
Methods

What is a code?

Σ alphabet with q symbols
(e.g., $\Sigma = \{0, 1, \dots, q-1\}$)

$n \in \mathbb{N}$ the block length

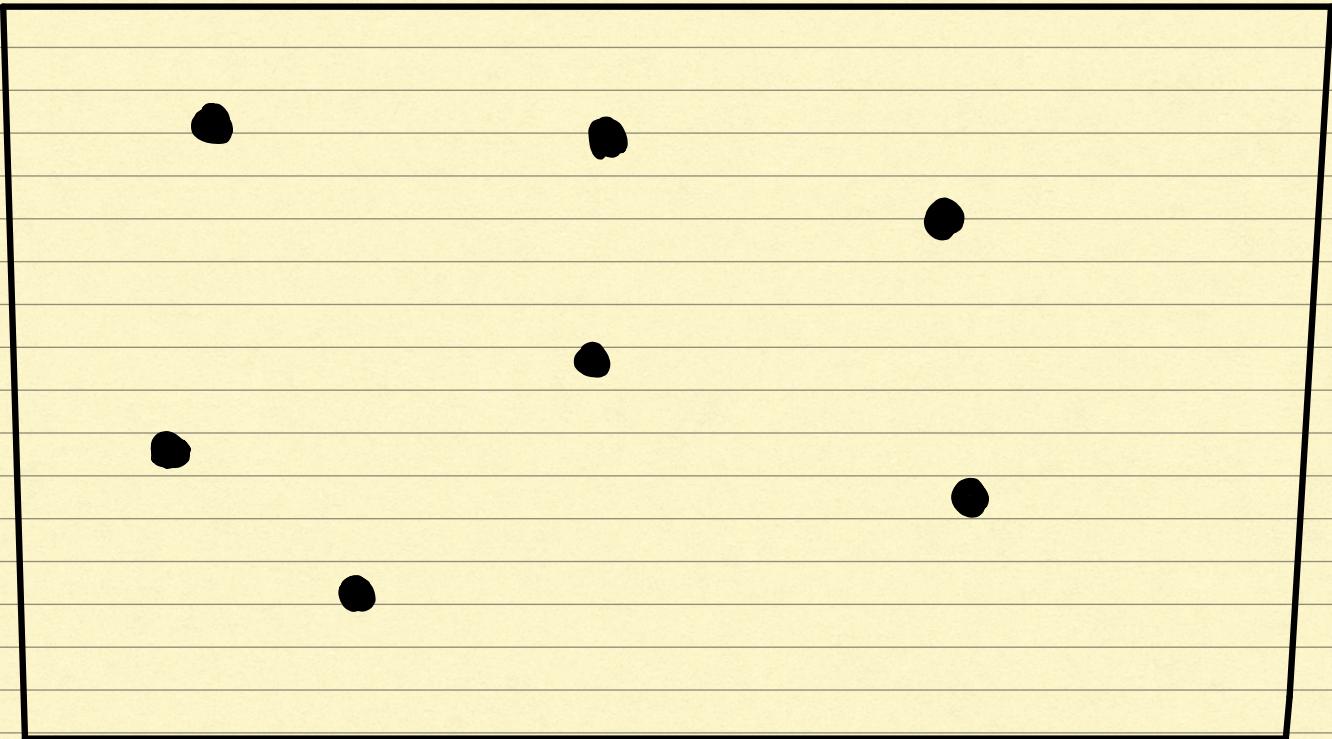
A code is a subset

$C \subseteq \Sigma^n$.

Coding Theory is essentially the
study of properties of
(sets of) strings.
aka codewords

Pictorial View

$$e \in \Sigma^n$$



- coordinates of e

(a bit too unstructured)

Adding a Metric

— Hamming Distance

$$x, y \in \sum^n$$

$$\Delta(x, y) := \sum_{i=1}^n \frac{1}{n} [x_i \neq y_i]$$

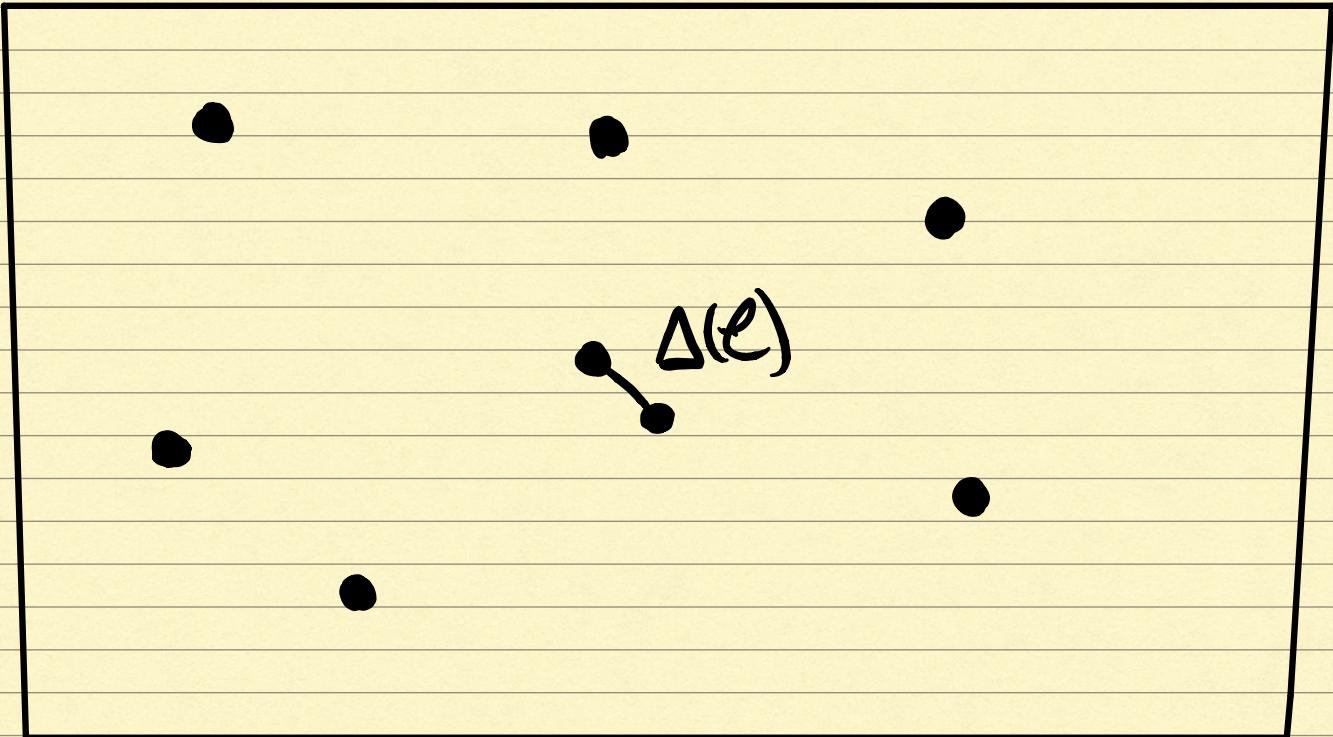
Ex: $x = 010111$

$$y = 111111$$

$$\Delta(x, y) = 2/6$$

Minimum Distance

$$e \subseteq \Sigma^n$$



- codewords of e

The (min.) distance of a code e

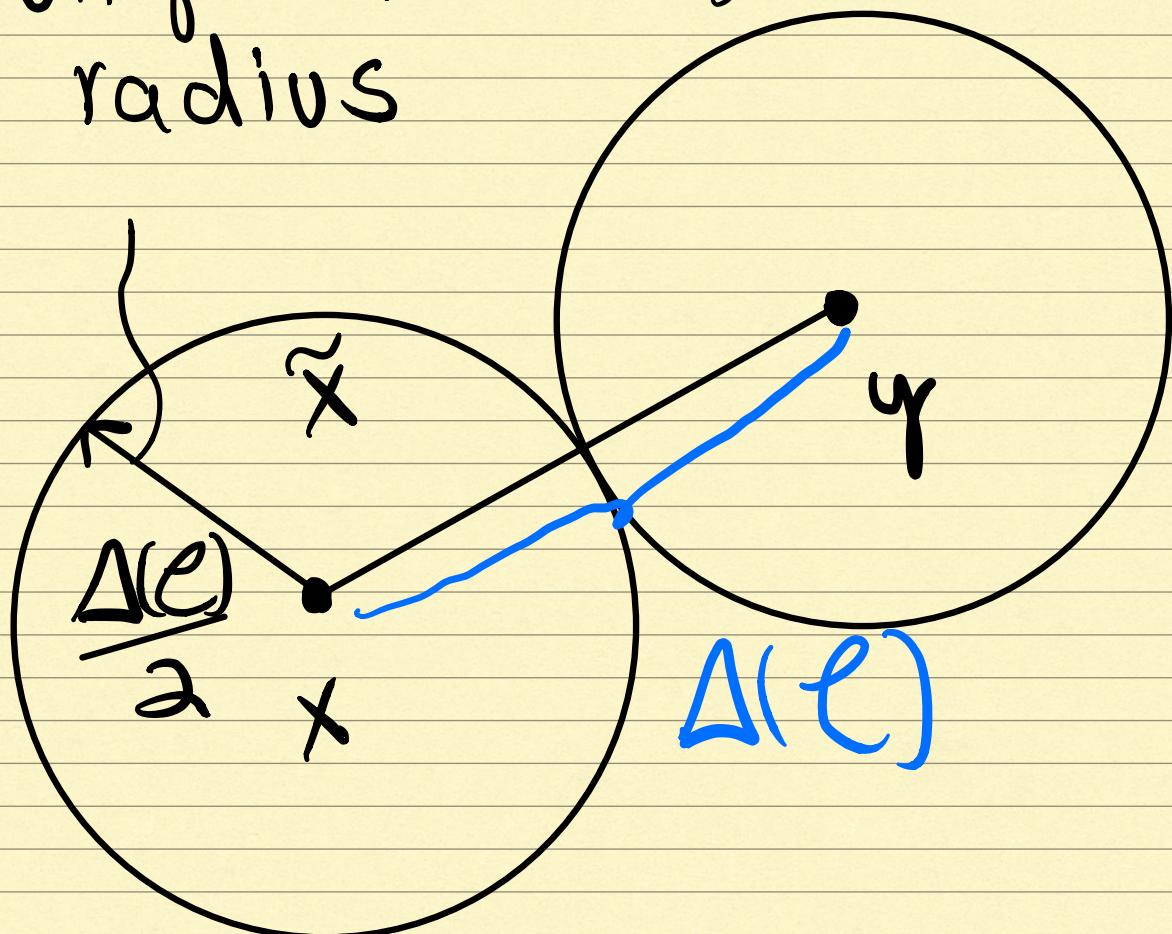
$$\Delta(e) := \min_{x \neq y} \Delta(x, y)$$

$$\begin{matrix} x \neq y \\ x, y \in e \end{matrix}$$

Why is $\Delta(e)$ important?

$x, y \in \mathcal{C}$ with $\Delta(x, y) = \Delta(\mathcal{C})$

Unique decoding
radius



x corrupted to $\tilde{x} \in B(x, \Delta(e)/2)$

Achieving Large $\Delta(e)$

Ex: $\Sigma = \{0, 1\}$

$\mathcal{C} = \underbrace{\{00\dots 0\}}_{n \text{ times}}, \underbrace{\{11\dots 1\}}_{n \text{ times}}$

$\Delta(e) = 1$ ✓ (great)

$|\mathcal{C}| = 2$ ↗ (bad)

Rate of ϵ

Want $|\epsilon|$ as large as possible.

The rate of ϵ is

$$r(\epsilon) := \log_{|\Sigma|}(|\epsilon|_n).$$

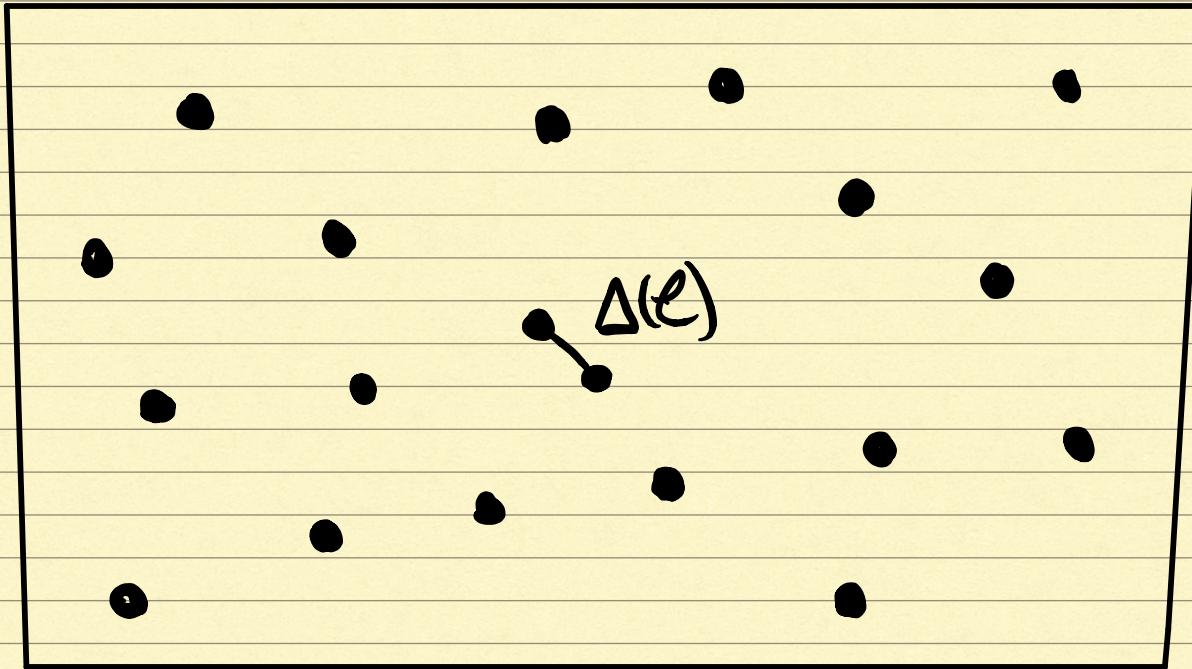
(the larger the better)

Conflicting Goals

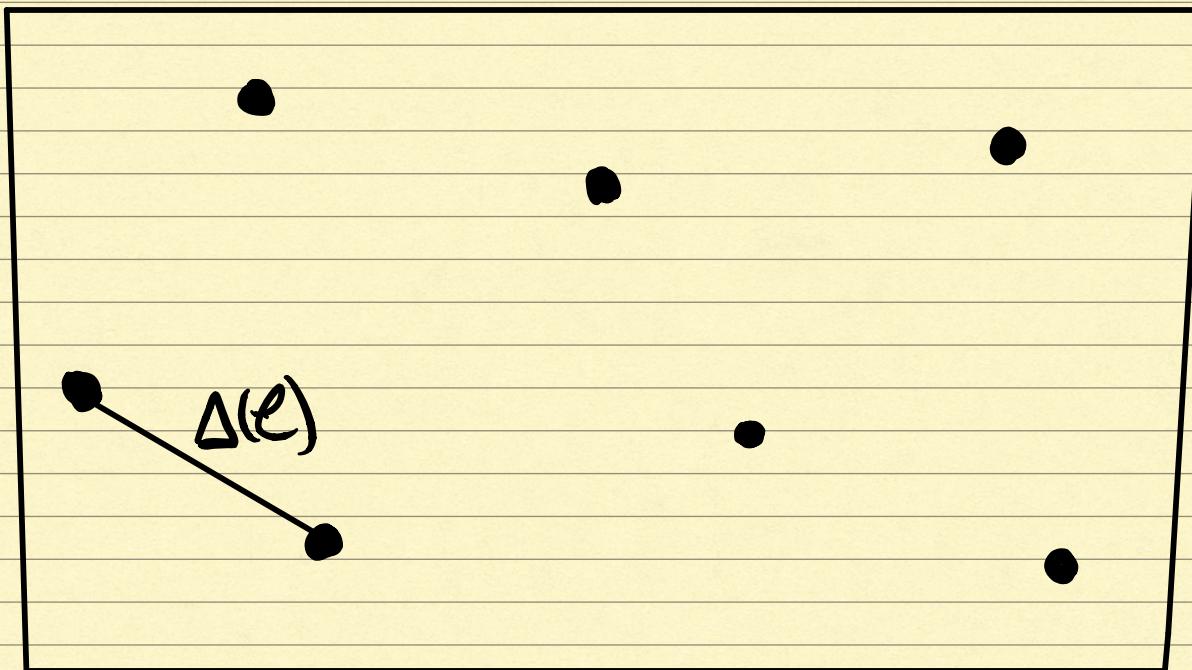
For a code \mathcal{C}

- Want $\Delta(\mathcal{C})$ large, and
- Want $r(\mathcal{C})$ large.

High $r(\ell)$ causes small $\Delta(\ell)$



High $\Delta(\ell)$ causes small $r(\ell)$



Fundamental Question

"What is the best trade-off
between rate and distance
of a code?"

Much of coding theory research
(for the Hamming model) revolves
around this question.

Wish List for a Code

Ideally, we may want a subset of:

- \mathcal{C} is explicit
- \mathcal{C} achieves the best parameter trade-off
- \mathcal{C} is efficiently encodable
- \mathcal{C} is // decodable/
list decodable
- \mathcal{C} has some local properties
- \mathcal{C} is over a small alphabet

Examples of Codes

To illustrate we will talk about :

- Reed-Solomon Codes
- Hadamard Codes
- Expanders and Codes

Some Notation

$$\mathcal{C} \subseteq \Sigma^n$$

For convenience, suppose $|\mathcal{C}| = |\Sigma|^k$ for some integer k .

Let $M = \Sigma^k$ be the message space.

We may want an encoding

function $\text{Enc} : M \rightarrow \mathcal{C}$

(Enc is simply a bijection).

Linear Codes

(possibly the most important)
class of codes

$$\Sigma = \mathbb{F} \text{ finite field}$$

Def. (Linear Code)

$\mathcal{C} \subseteq \mathbb{F}^n$ is a linear subspace.

$$n \cdot r(\mathcal{C}) = \dim(\mathcal{C}) =: k$$

\mathcal{C} can be specified either by

- a "generation matrix" $G \in \mathbb{F}^{n \times k}$

$$\text{s.t. } \mathcal{C} = \{Gz \mid z \in \mathbb{F}^k\}$$

- a "parity check matrix" $H \in \mathbb{F}^{k \times n}$

$$\text{s.t. } \mathcal{C} = \{x \in \mathbb{F}^n \mid Hx = 0\}$$

Def (Hamming weight)
 Let $\Sigma = \{0, 1, \dots, q-1\}$ and $x \in \Sigma^n$.
 The Hamming weight $\|x\|$ of x is

$$\|x\| = \sum_i \mathbb{I}[x_i \neq 0] / n.$$

Claim: If \mathcal{L} is linear, then

$$\Delta(\mathcal{L}) = \min_{0 \neq x \in \mathcal{L}} \|x\|.$$

Proof Follows from

$$\Delta(x, y) = \Delta(x-y, 0) = \|x-y\|,$$

but $x-y \in \mathcal{L}$ since \mathcal{L} is linear. □

Reed-Solomon Codes

Parameters

- { n the block length
- K message length (rate K/n)
- F a field of size $\geq n$
- $\alpha_1, \alpha_2, \dots, \alpha_n \in F$ distinct

Set of degree $K-1$ polynomials

$$\mathcal{P}_{K-1} := \left\{ p(x) = c_0 + c_1 x + \dots + c_{K-1} x^{K-1} \mid (c_0, c_1, \dots, c_{K-1}) \in F^K \right\}$$

Reed-Solomon Code

$$\mathcal{C} = \left\{ (p(\alpha_1), p(\alpha_2), \dots, p(\alpha_n)) \mid p \in \mathcal{P}_{K-1} \right\}$$

Well-known Montra

Fact: $(\forall 0 \neq p \in \mathcal{P}_{k-1}) (p \text{ has } \leq k-1 \text{ roots})$.

This fact implies amazingly large distance for Reed-Solomon.

Proof

$\forall p, q \in \mathcal{P}_{k-1}$ with $p \neq q$

$$0 \neq p - q \in \mathcal{P}_{k-1} \quad \text{abuse of notation}$$

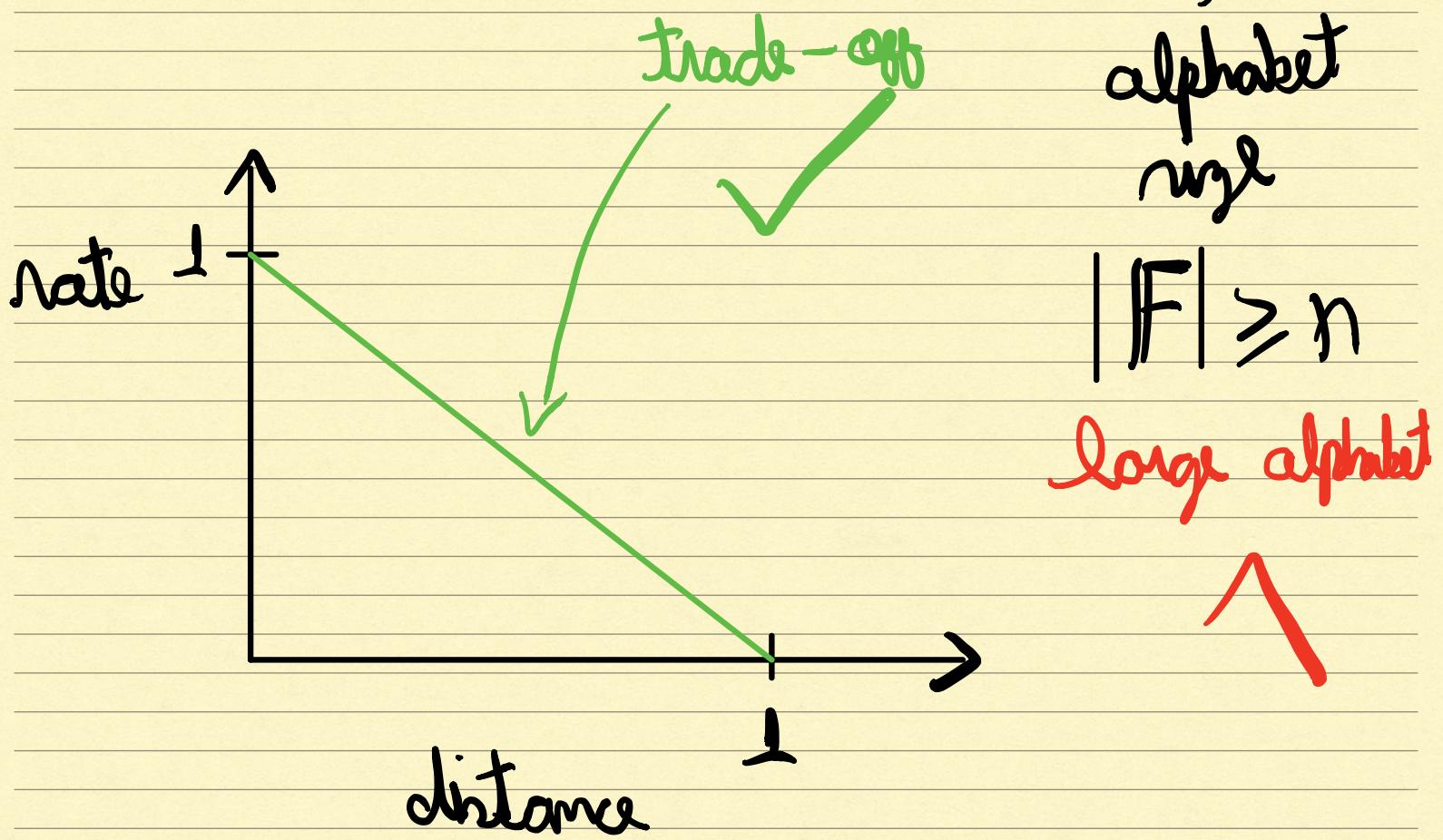
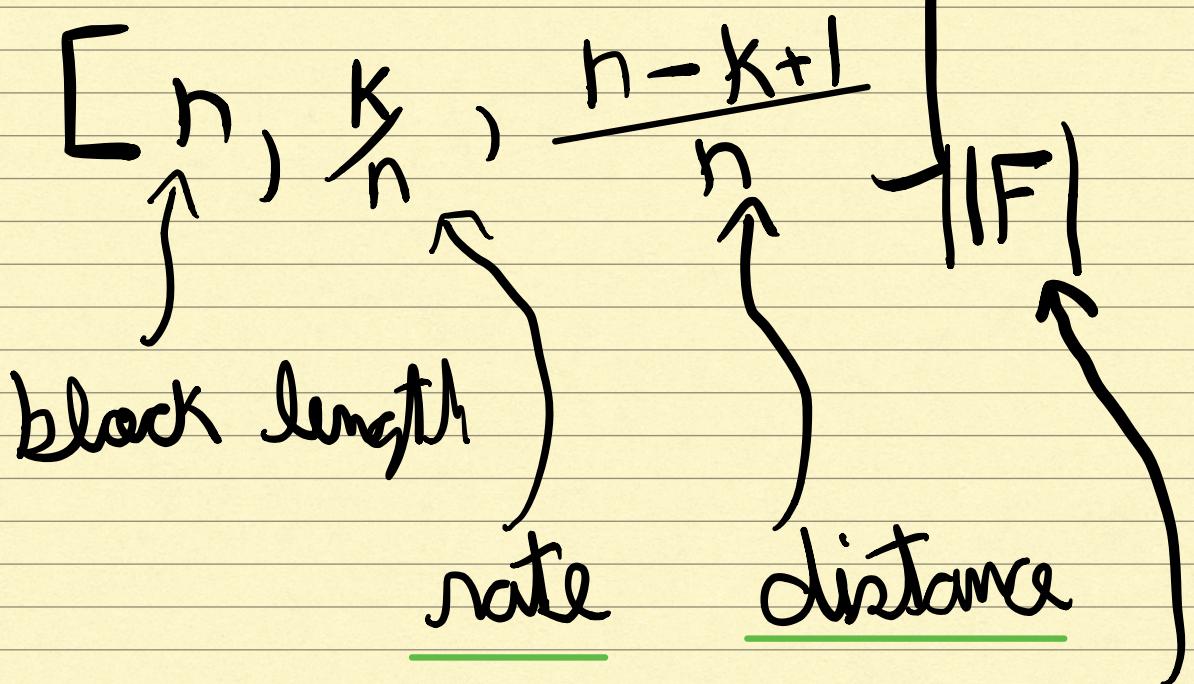
\Downarrow

$$n \Delta(p, q) = n \|p - q\| = n - \#\{\text{roots of } p - q\}$$

$$n \Delta(e) \geq n - k + 1$$



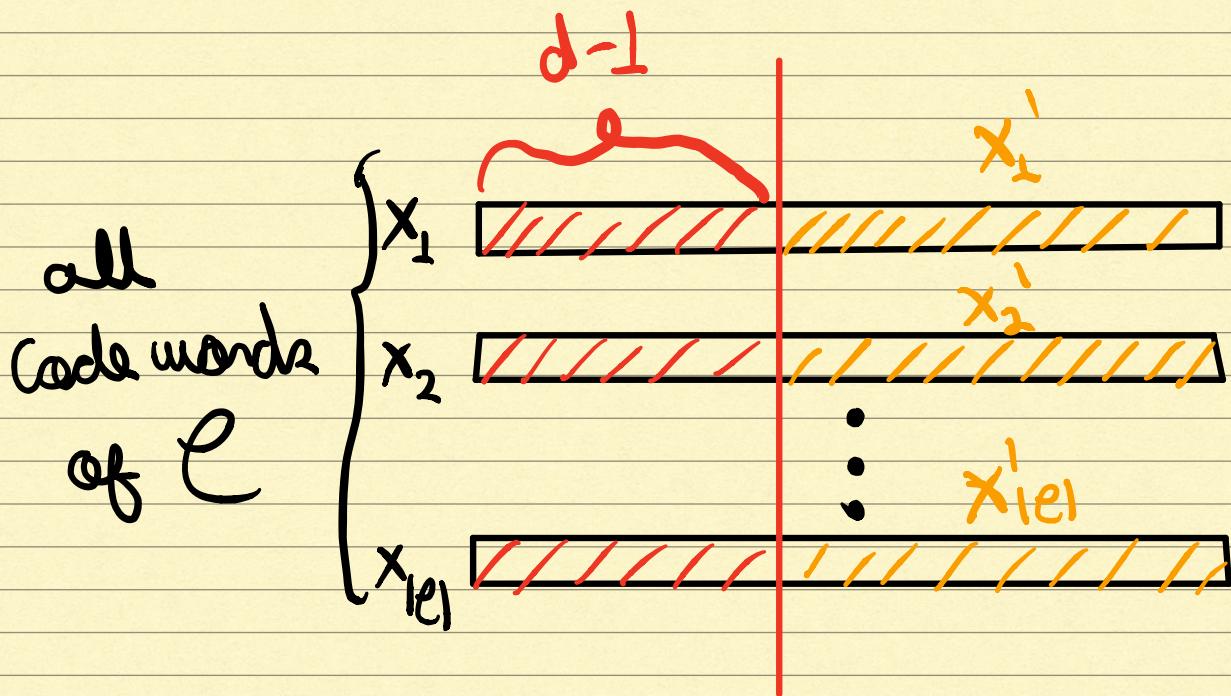
Parameters of Reed-Solomon



We will see that this trade-off is optimum

Singleton Bound

Let \mathcal{C} be $[n, k, d]_q$



delete the first
 $d-1$ symbols

$x'_1, \dots, x'_{i_{el}}$ are distinct
since $\Delta(\mathcal{C}) \geq d$

$$q^k = |\mathcal{C}| \leq q^{n-(d-1)} \implies k \leq n-d+1$$

Hadamard Codes

For $x \in \mathbb{F}_2^K$, define the linear function

$$f_x(y) = \langle x, y \rangle \\ = \sum_i x_i y_i \pmod{2}.$$

Define the evaluation table of f_x as

$$\text{Eval}(f_x) = (f_x(y))_{y \in \mathbb{F}_2^K}$$

The Hadamard code is

$$\text{Had}_K := \{\text{Eval}(f_x) \mid x \in \mathbb{F}_2^K\}.$$

2th block length is $n = 2^K$ and

$$|\text{Had}_K| = 2^K.$$

$$\text{The rate } r(\text{Hack}_k) = \frac{k}{2^k}$$

$$= \frac{\log n}{n}.$$

To analyse $\Delta(\text{Hack}_k)$ we will use a bit of Fourier analysis.

f_x, f_y with $x \neq y$

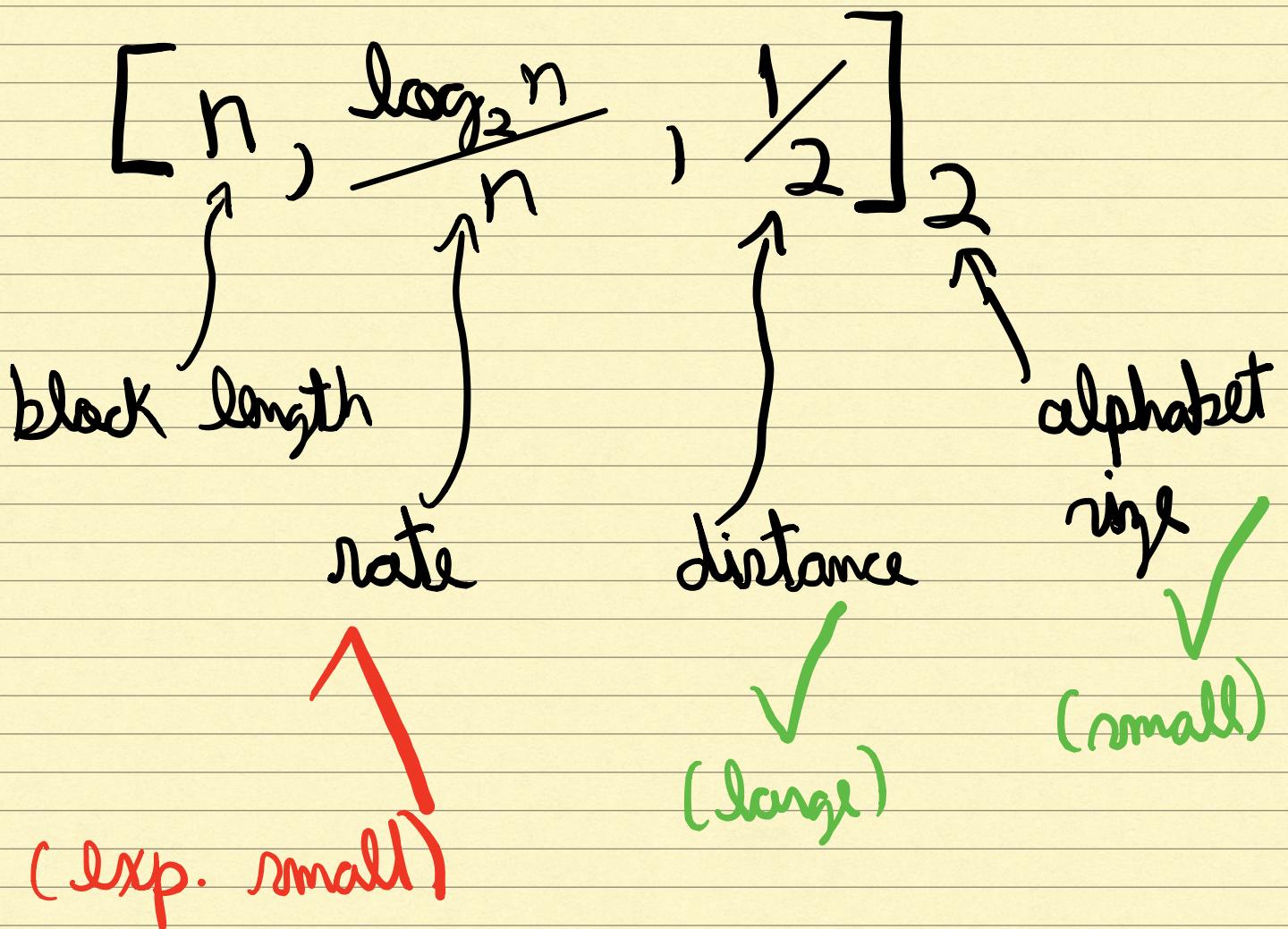
$$f_x + f_y = f_{x+y} \quad (\text{Property of linear function})$$

$$\Delta(f_x, f_y) = \|f_{x+y}\| \quad (\text{since Hack}_k \text{ is a linear code})$$

$$\sum_{y \in \mathbb{F}_2^k} (-1)^{\langle x+y, y \rangle} = 0 \quad \text{for } x \neq y$$

$$\Rightarrow \Delta(f_x, f_y) = \frac{1}{2}.$$

Hach_K parameters ($k = \log_2 n$)



Local Properties

Hadamard codes have amazingly local properties.

"informal"
Def. (Locally Testable Codes)

We say that a code $\mathcal{C} \subseteq \Sigma^n$ is an $\mathsf{l-LTC}$ if \exists a tester \mathcal{T}

s.t. $\forall x \in \Sigma^n$, \mathcal{T} queries

$\leq l$ positions of x and

- (Completeness) If $x \in \mathcal{C}$, then \mathcal{T} accepts w.p. $\geq 2/3$

- (Soundness) If $\Delta(x, \mathcal{C}) = \Omega(1)$, then \mathcal{T} accepts w.p. $\leq 1/3$.

Consider the following tester \mathcal{T}
for Had_K :

Input: oracle to $f: \mathbb{F}_2^K \rightarrow \mathbb{F}_2$

\mathcal{T} samples $\gamma_1, \gamma_2 \in \mathbb{F}_2^K$

queries $f(\gamma_1), f(\gamma_2), f(\gamma_1 + \gamma_2)$

accepts iff $f(\gamma_1) + f(\gamma_2) = f(\gamma_1 + \gamma_2)$.

Theorem (BLR linearity test)

Had_K is a 3-LTC with
tester \mathcal{T} (above).

Proof Exercise. Hint: Fourier analysis.

Expanders and Codes

Roughly, expanders have two main uses in Coding Theory:

- Distance amplification
- Define the parity check matrix (Tanner codes)

What is an Expander?

Expander: "well connected sparse graph"
"derandomization of a complete graph"
"random like explicit sparse graph"

Roughly, we will have:

- Being sparse will be helpful to achieve good rates.
- Random like will imply good distance.
- Being explicit will make codes explicit.

Distance Amplification

We illustrate with a (high-level) example.

(i) Start with a base code \mathcal{C}_0

$[n, r_0, d_0]_2$ with $r_0, d_0 = \mathcal{Q}(1)$.

(ii) Take a bipartite expander

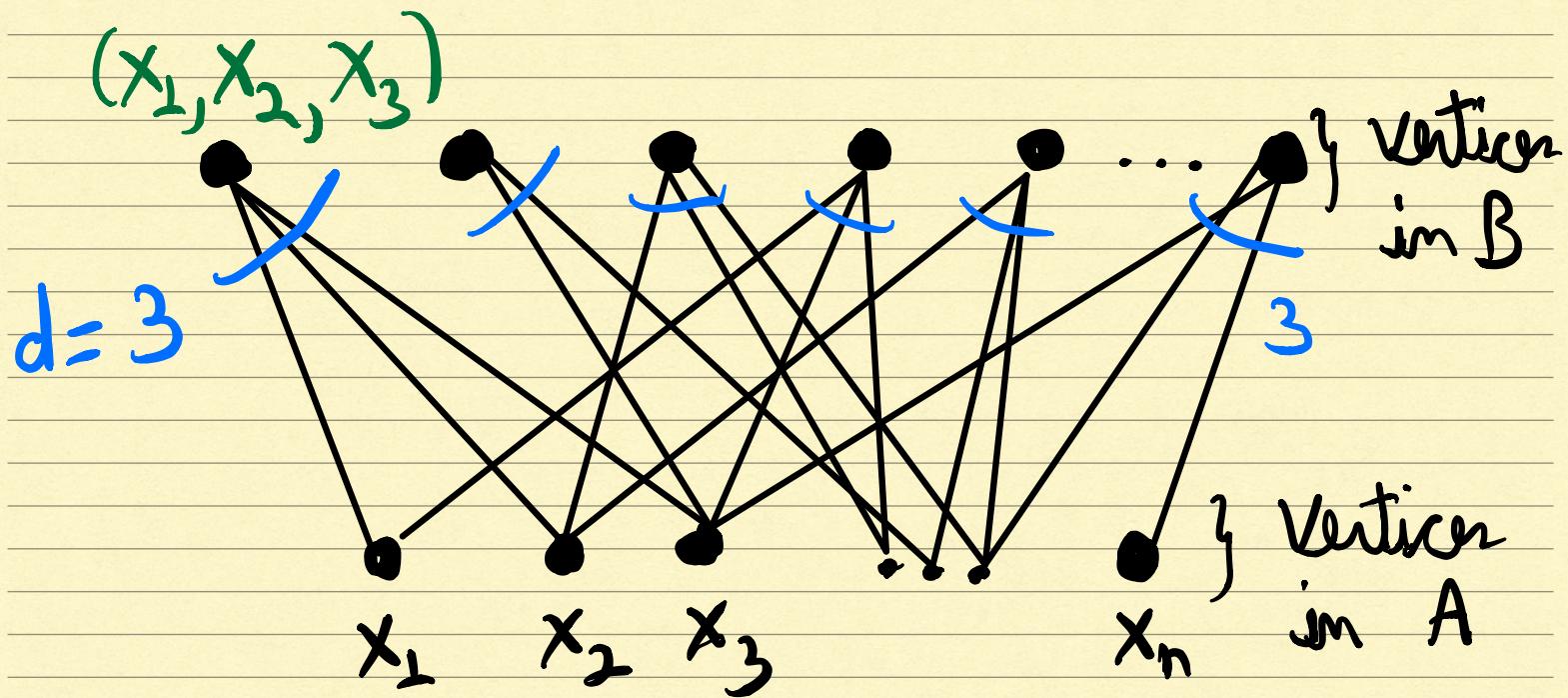
$G = (A \cup B, E)$ with

$$|A| = n$$

$$\deg(v) = d \quad \forall v \in B.$$

(The expansion of G needs to be strong enough)

$$x \in F_2^n, G = (A \cup B, E)$$



Assume $\forall v \in B, N(v) = \{v_1, v_2, v_3\}$.
(i.e., neighbors of v are labeled)

Form a new codeword from x
 denoted $\text{lift}(x) \in (F_2^d)^{|B|}$ as:

$$\text{lift}(x)_v := (x_{v_1}, x_{v_2}, x_{v_3})$$

for every $v \in B$.

Why is the distance amplified?

We use the "sampling" property of an expander:

$$\forall S \subseteq A \text{ with } |S| \geq \lambda_0 \cdot |A|$$

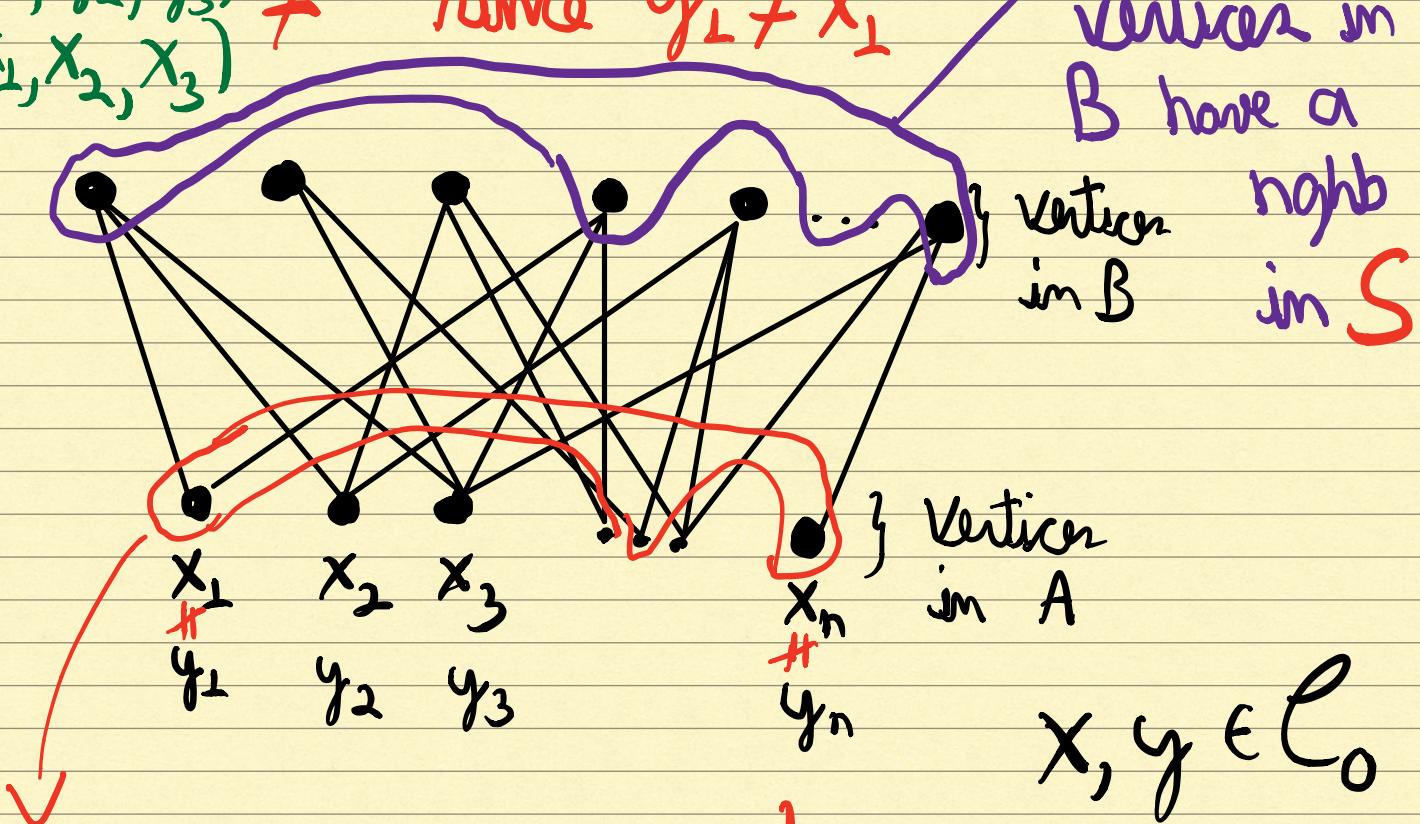
most vertices (depending on λ_0 and the expansion of G) of B

have neighbors in S .

Ex:

$$(y_1, y_2, y_3) \neq \text{ since } y_1 \neq x_1$$

$$(x_1, x_2, x_3)$$



$$S = \{i \in A \mid x_i \neq y_i\}$$

$$|S| = \Delta(x, y) > \Delta(\ell_0)$$

This implies that

$$\Delta(\text{lift}(x), \text{lift}(y)) = 1 - \varepsilon,$$

where $\varepsilon = \varepsilon(\lambda, \delta_0)$.

(spectral gap)

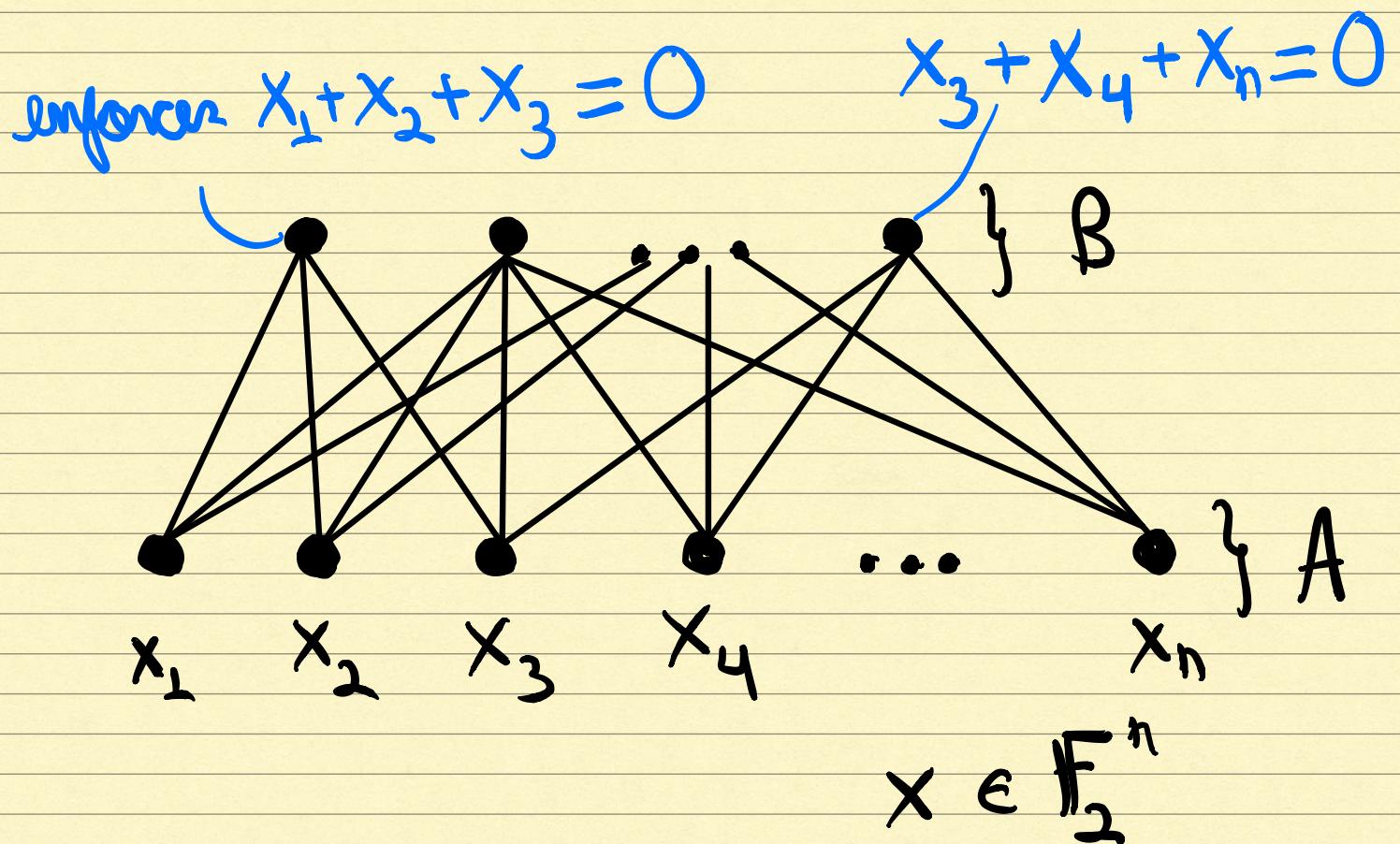
Exercise: Compute bounds on ε .

Note : The specific "lifting" function used in our example is the well-known direct product encoding.

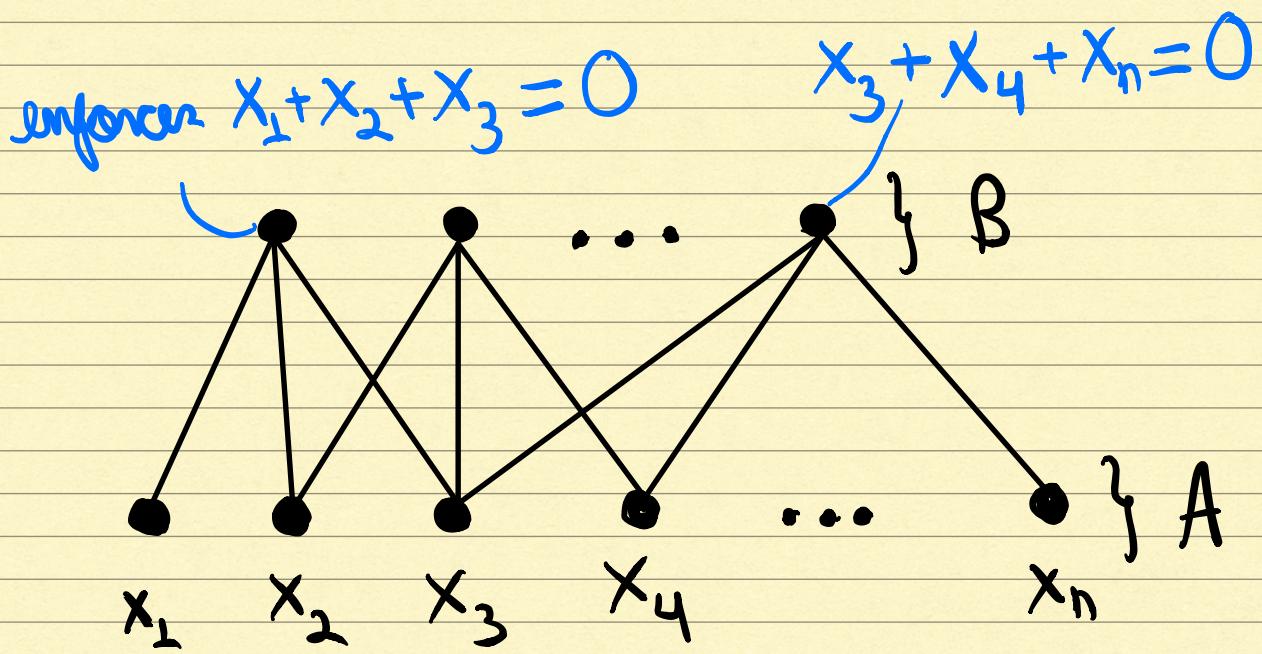
Expanders and Parity Check

The adjacency relation of a bipartite (vertex) expander is used to define the parity check matrix of a code.

$$G = (A \cup B, E)$$



Require: $|B| \leq p |A|$ $p \in (0, 1)$



Naturally define $H \in \mathbb{F}_2^{B \times A}$ as

$$H_{v,j} := \mathbb{1}_{[v \sim_G j]}$$

to be the parity check matrix
of \mathcal{C} .

$$\dim(\mathcal{C}) \geq |A| - |B| \geq (1-\rho)|A|$$

$$= (1-\rho) \cdot n.$$

Minimum distance

Exercise: use a sufficiently strong vertex expansion assumption to obtain minimum distance.

Decoding can be done using a simple local algorithm.

The above construction is due to Sipser and Spielman.

Their code belongs to the more general class of so-called

Tammer codes (which allow

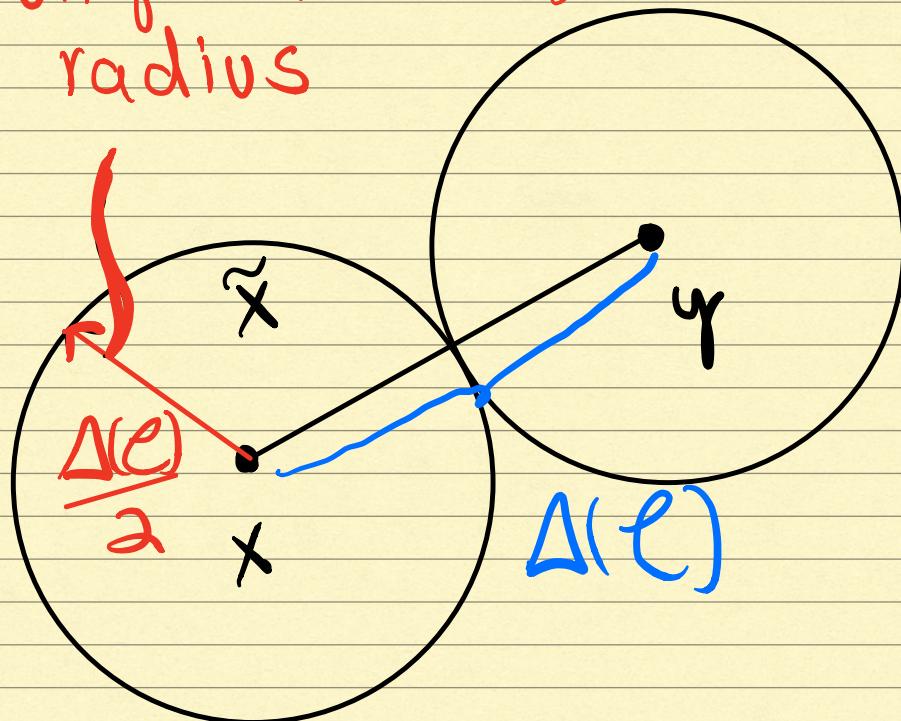
more general parity constraints).

[See also Guruswami's LDPC survey
and Montanari's book.]

List Decoding

List decoding is a relaxed decoding scheme that allows one to roughly double the unique decoding radius (for well behaved codes) with the caveat of allowing a small list of codewords.

Unique decoding radius



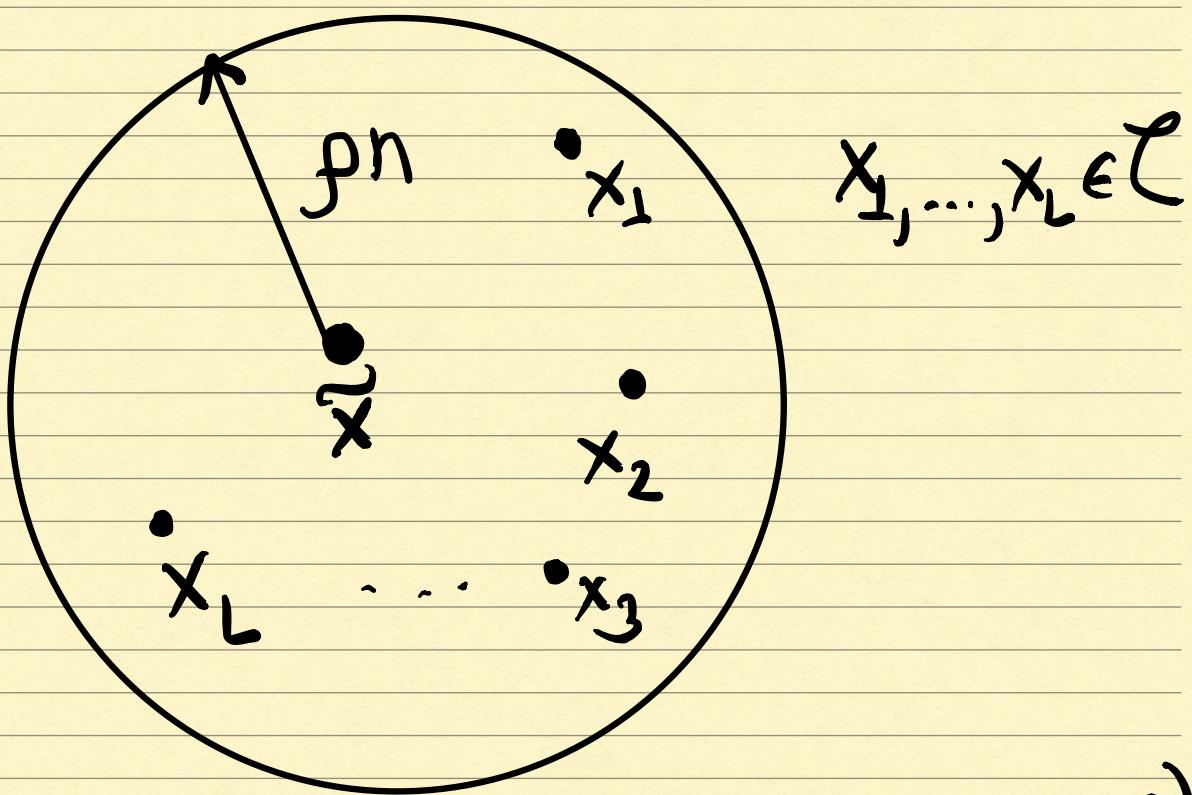
Setting $\mathcal{C} \subseteq \Sigma^n$, $p \in (0,1)$

and arbitrary $\tilde{x} \in \Sigma^n$.

Define the list

$$L(\tilde{x}, p \cdot n) := \{x \in \mathcal{C} \mid \Delta(x, \tilde{x}) \leq pn\}.$$

Ex:



Possibly $p \approx \Delta(\mathcal{C})$ (rather than $\frac{\Delta(\mathcal{C})}{2}$)

Def: $\mathcal{C} \subseteq \Sigma^n$ is (p, L) -list decodable

if $|L(\tilde{x}, pn)| \leq L \quad \forall \tilde{x} \in \Sigma^n$.

If $\mathcal{L} \subseteq \Sigma^n$ is (p, L) -list

decodable with $L \leq \text{poly}(n)$,

then \mathcal{L} is said to be

combinatorially list decodable.

List decoding is an important topic in modern Coding Theory.

Gilbert-Varshamov Bound

Non-constructive existential bound
for codes.

Theorem : (GV Bound)

$$(\forall q \geq 2)(\forall p \in (0, 1 - \frac{1}{q}))(\forall \varepsilon > 0)$$

(\exists) a family of q -ary \mathcal{C}_n codes s.t.

(i) $\Delta(\mathcal{C}_n) \geq p$, and

(ii) $r(\mathcal{C}_n) \geq 1 - H_q(p) - \varepsilon$.

n is the block length

The GV bound follows from
the trivial bound:

Fact 1: $\alpha(G) \geq \frac{|V(G)|}{\deg(G)+1}$,

where G is a regular graph of
degree $\deg(G)$.

G for us will have $V = [q]^n$ and
 $x, y \in V$ are connected $\Leftrightarrow \Delta(x, y) \leq p n$

Rmk 1: Any independent set in G is a
code of distance $\geq p$.

Rmk 2: $\deg(x) = |B(x, pn)| \approx q^{H_q(p)n}$

Rmk 1 + Rmk 2 + Fact 1



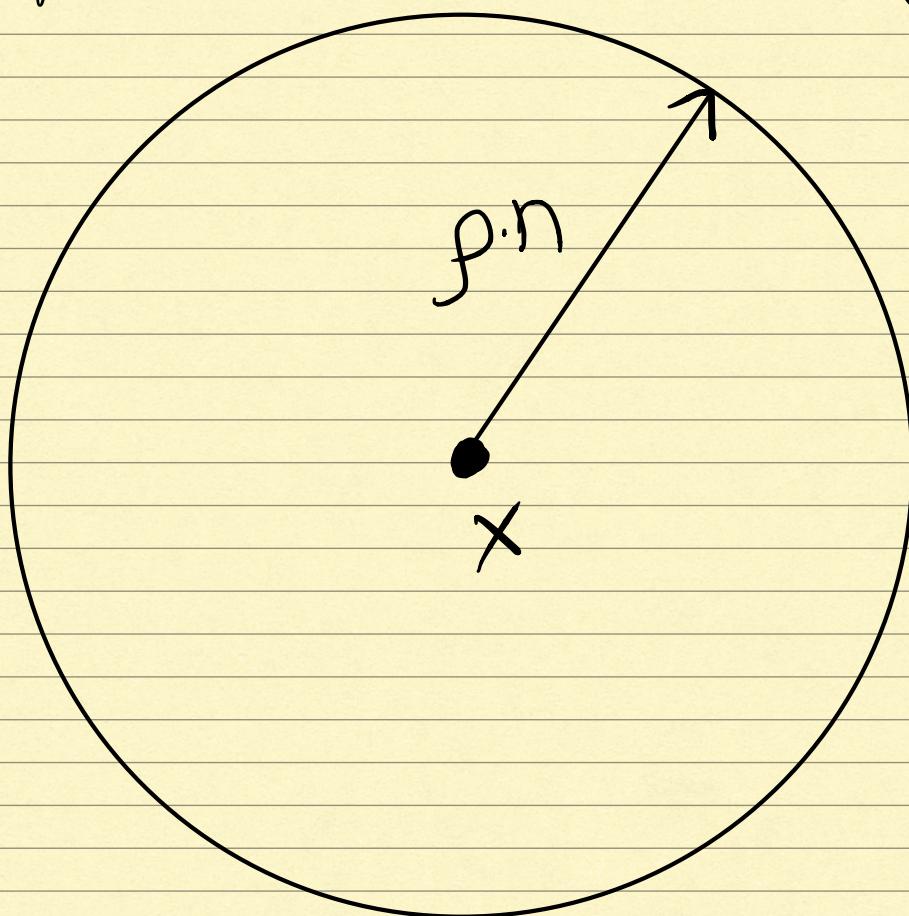
$$\begin{cases} \Delta(\mathcal{C}) \geq p, \text{ and} \\ |\mathcal{C}| \geq q^{H_q(p)n} \end{cases}$$

$$\exists \mathcal{C} \subseteq V \text{ s.t. } |\mathcal{C}| \geq q^{H_q(p)n}$$

i.e., $\Delta(\ell) \geq p$ and

$$r(\ell) \geq 1 - H_g(p) - \varepsilon.$$

In a picture, we are covering the g -ary n -cube with balls of radius p^n .

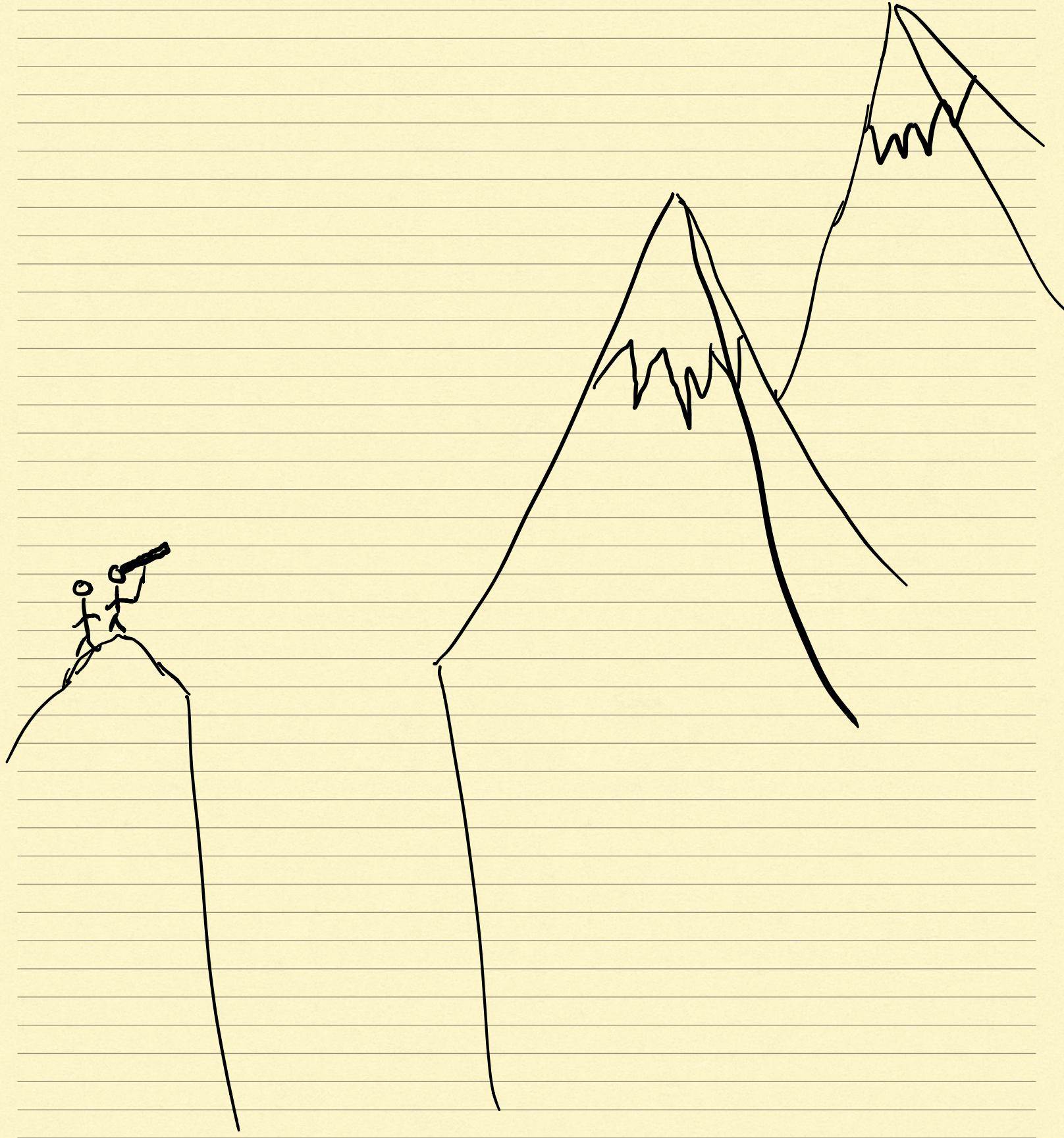


$B(x, p^n)$
Hamming ball of
radius p^n .

Despite its simple form
the GV bound is the
best known existential bound
for binary codes of constant
rate and constant relative
distance.

(The GV Bound also holds for)
linear codes.

Open Problems



List Decoding Capacity for Binary Codes

An ambitious problem in this direction.

Open Problem: (, , , ...,)

Find an explicit family of
binary codes of rate $\Omega(\varepsilon^2)$
efficiently list decodable from

radius $\frac{1}{2} - \varepsilon$.

The following relaxations are also interesting.

Open Problem: (⊤, ⊤, ⊤, ., X)

Find an explicit family of binary codes of rate $\Omega(\varepsilon^2)$ efficiently list decodable from radius $\frac{1}{2} - \varepsilon$.

Open Problem: (⊤, ⊤, ⊤, ., X)

Find an explicit family of binary codes of rate $\Omega(\varepsilon^2)$ efficiently list decodable from radius $\frac{1}{2} - \varepsilon$.

Combinatorially

What is Known?

Theorem (Guruswami-Rudra)

~~Open Problem:~~

Find an explicit family of binary codes of rate $\Omega(\varepsilon^3)$ efficiently list decodable from radius $\frac{1}{2} - \varepsilon$.

+ ...

Theorem (Guruswami-Rudra)

~~Open Problem:~~

Find a explicit family of binary codes of rate $\Omega(\varepsilon^2)$ efficiently list decodable from radius $\frac{1}{2} - \varepsilon$.

(Monteiro et al.)

+ ...

Combinatorially

Good LTCs

Open Problem: ($\text{LTC}_1, \text{LTC}_2, \dots, \text{LTC}_n$)

Find an explicit family of
 λ -LTCs with

- Constant number of queries,
i.e., $\lambda = O(1)$,
- constant rate, and
- constant relative distance.

(This problem is related to
the open problem of linear ring
PCPs)

The following relaxations are also interesting.

Open Problem: (L_P , L_P , ..., L_P)

Find an ~~explicit~~ family of
 λ -LTCs with

- Constant number of queries,
i.e., $\lambda = O(1)$,
- constant rate, and
- constant relative distance.

Open Problem: (L_P , L_P , \times , L_P)

Find an ~~explicit~~ family of
 λ -LTCs with

- Constant number of queries,
i.e., $\lambda = O(1)$,

improve the
dependence
 $\lambda(n)$

- constant rate, and
- constant relative distance.

What is Known?

Theorem (Bam-Sasson and Sudan)

Open Problem:

Find an explicit family of ℓ -LTCs with

- Constant number of queries,
i.e., $\ell = O(1)$,
- ~~constant rate~~, and
- constant relative distance.

+
(Dimin)
+
(Mein)

not entirely
explicit

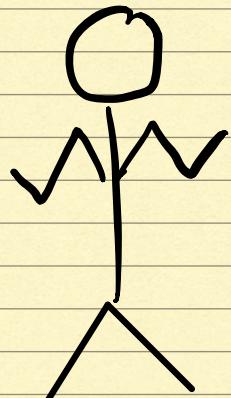
rate $1/\text{polylog}(n)$, where n is the block length

LTCs and HDX

(a diversion)

HDXs constructions can yield
linear size $\ell = O(1)$ -uniform
hypergraphs with nice properties.
Total # hyperedges = $O(\# \text{ vertices})$.

HDX machinery might be
helpful towards building
good LTCs (according to
some specialists).



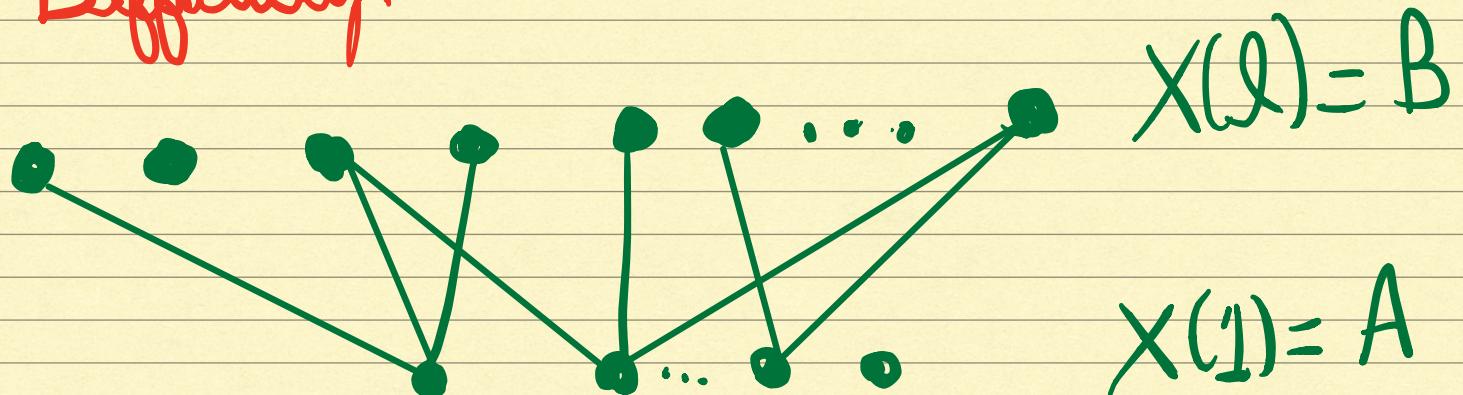
(borrowing from Mary Wootters)

In this direction, we have :

Open Problem:

Are there constant distance,
constant rate Tanner codes
on HDXs?

Difficulty:



$$|B| > |A|$$

$$H \in \mathbb{F}^{B \times A}$$

parity check

naive counting cannot ensure
non-trivial rate

Counting is against us

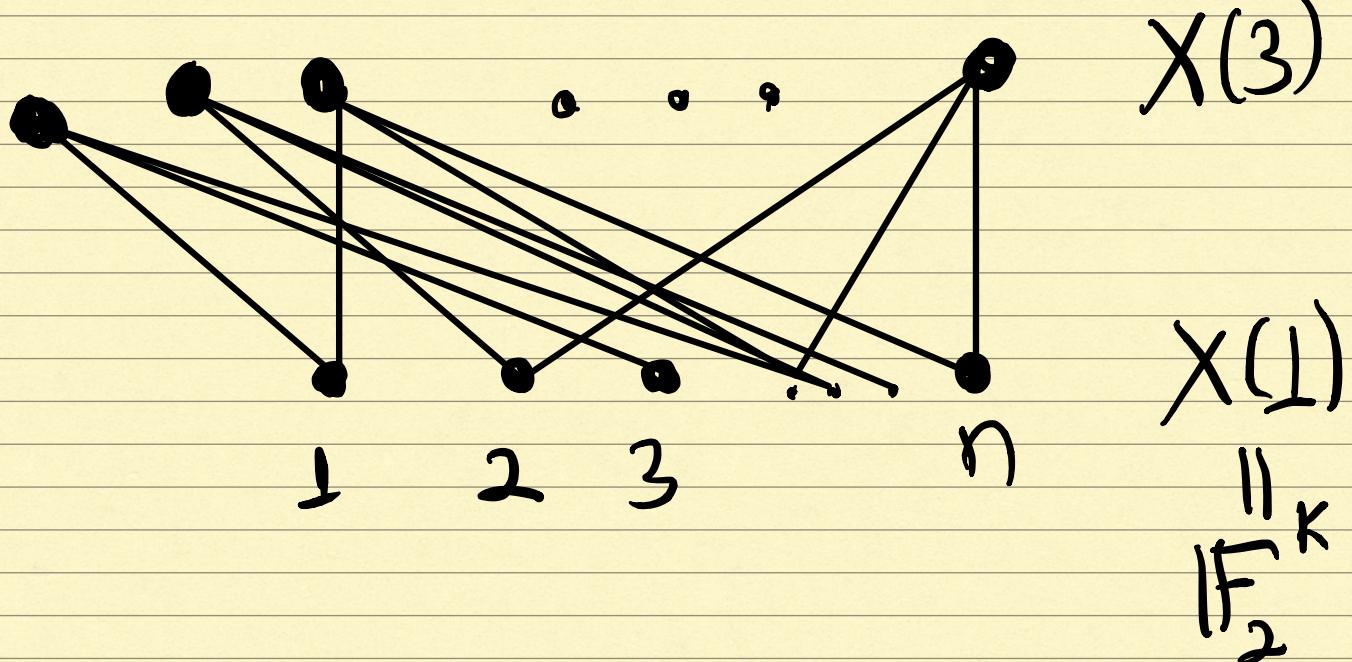
but some algebraic structure

can guarantee a non-trivial
Kernel

Ex: Had_K $f_x \quad x \in \mathbb{F}_2^K$

$$n = 2^K$$

$$\{(x, y, x+y) \mid x, y \in \mathbb{F}_2^K\}$$

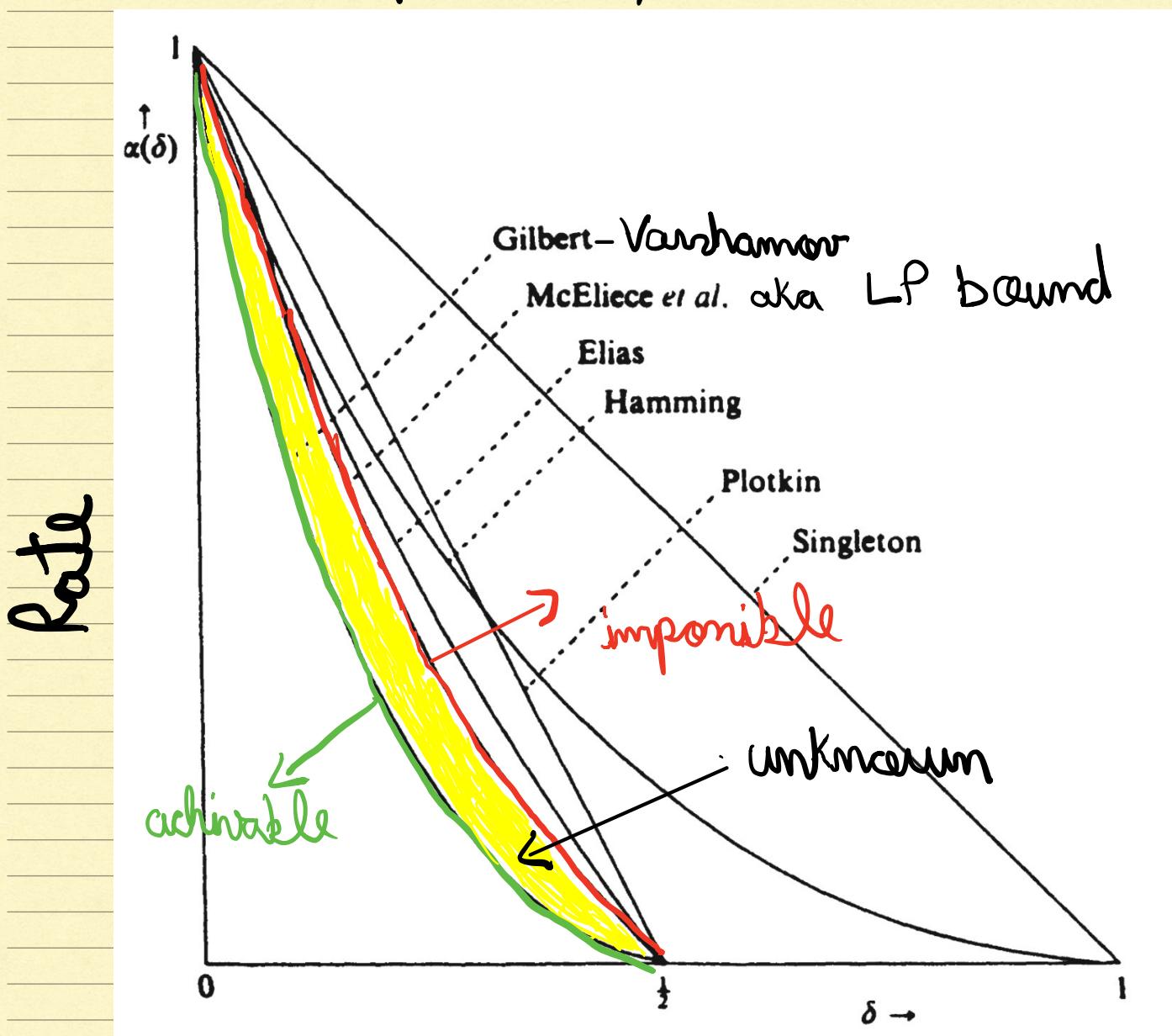


$$n^2 = |X(3)| \gg |X(1)| = n$$

$$\text{yet } r(\text{Had}_K) = \log_2 n / n \gg 1/n.$$

Distance vs Rate trade-off for Binary Codes

Noticeable gap between existential and impossibility bounds.



Distance

Ref: adapted from van Lint's book.

Open Problem: (, , , ...,)

Find the best asymptotic trade-off

between rate and distance for
binary codes.

(This is known only for convex case)

What is Known?

Asymptotically for good binary codes:

- Best lower bound:
(positive / existential)

Gilbert-Varshamov Bound
1950's

- Best upper bound:
(negative / impossibility)

LP bounds McEliece et al.

1970's

Tighten the Gap between Upper and Lower Bounds for LDCs

The rate lower bounds for LDCs
and the rate of known constructions
are very far apart.

Open Problem:

Tighten the huge rate gap of
constant query LDCs.

To give an idea of the gap
(this may be a bit outdated)

$\frac{k}{n}$ is the rate with n being
the block length

Theorem (Efremenko)

\exists 3-LDCs with

$$n = \exp(\exp(\sqrt{\log k \log \log k})).$$

Theorem (Woodruff)

\forall 3-LDCs $n = \Omega(k^2)$.

(linear)

(see Yekhanin's survey)

"Zig-Zag product" of High-dimensional Expanders

The zig-zag product of Reingold, Vadhan and Wigderson is an operation that combines two expander graphs G and H to produce a larger one $G \otimes H$. It has several applications including to coding theory.

Open Problem:

Given two MDXs (spectral link expanders) X and Y , can we produce a larger one $X^{\text{a}} \otimes^{\text{b}} Y$?

Improve the Running Time

of a Unique Decoder of
explicit ϵ -balanced near
The GV Bound

(We now have polynomial time decoding
for these codes, but nowhere near
quadratic or near-linear.

Open Problem:

Improve the decoder running
time to near-linear time.

Lower Bounds for Average

Cone Coding Tasks against the Sum-of-Squares Hierarchy

A sample question in this direction (probably requires more thinking)

Question :

Can we fool SOS into thinking that a random linear code (given by a random generator matrix) has a codeword of small Hamming weight?

I would not classify it as open at this point.

That's all.

Thank you!