# List-Decodability of Random Linear Codes

Venkatesan Guruswami

Carnegie Mellon University

ICM 2010 Satellite Workshop
Algebraic & Probabilistic Aspects of Combinatorics & Computing
August 2010

Joint work with
Johan Håstad (KTH) and Swastik Kopparty (MIT)

## Question

Suppose we get to transmit $n$ bits over a noisy channel.
What is the best rate of information transmission if the channel flips $\approx p$ fraction of the bits?

## Question

Suppose we get to transmit $n$ bits over a noisy channel.
What is the best rate of information transmission if the channel flips $\approx p$ fraction of the bits?

- (binary) code $C \subseteq \{0,1\}^n$
  - Transmit *codewords* of $C$
  - information rate $= R(C) = \frac{\log_2 |C|}{n}$ (info per codeword bit)
- (binary) linear code: $C$ a *subspace* of $\mathbb{F}_2^n$.
- $q$-ary linear code: Subspace of $\mathbb{F}_q^n$.

# Shannon's theorem

Asymptotics: Fix $R, p$, let $n \to \infty$. Study *families* of codes.

**Capacity of binary symmetric channel**

If error $e \sim \mathrm{Binom}(n, p)$, then $\exists C$ with rate $1 - h(p) - o(1)$ and $\mathrm{Dec} : \{0,1\}^n \to C$ s.t. $\forall c \in C$

$$\Pr_e\left[\mathrm{Dec}(c+e) = c\right] \geqslant 1 - o(1) \ .$$

# Shannon's theorem

Asymptotics: Fix $R, p$, let $n \to \infty$. Study *families* of codes.

## Capacity of binary symmetric channel

If error $e \sim \mathrm{Binom}(n, p)$, then $\exists C$ with rate $1 - h(p) - o(1)$ and $\mathrm{Dec} : \{0,1\}^n \to C$ s.t. $\forall c \in C$

$$\Pr_e \left[ \mathrm{Dec}(c + e) = c \right] \geqslant 1 - o(1) \ .$$

$1 - h(p)$ is optimal (capacity):

- Given $c$, we have $\approx \binom{n}{pn} \approx 2^{h(p)n}$ likely possibilities for $y = c + e$.
- So $|\mathrm{Dec}^{-1}(c)| \approx 2^{h(p)n}$ for all codewords $c \in C$.
- So $|C| \leqslant 2^{(1-h(p)+o(1))n}$

# Worst-case errors

What if $e \in \{0,1\}^n$ is arbitrary subject to $|e| \leqslant pn$, and we want $\mathrm{Dec}(c+e) = c$ for every such $e$ (and $\forall c \in C$)?

Requires Hamming balls of radius $pn$ around the codewords to be disjoint.

- Restricts $R(C) \to 0$ for $p \geqslant 1/4$
- For $p < 1/4$, best rate $R_p$ unknown

$$1 - h(2p) \leqslant R_p \leqslant h\left(\frac{1}{2} - \sqrt{2p(1-2p)}\right) < 1 - h(p) \ .$$

# List decoding

Relaxed goal: From $c + e$, the codeword $c$ is determined up to ambiguity $L$ (a large but fixed constant, independent of $n$)

**Definition (List-decodability)**

A code $C \subset \Sigma^n$ is $(p, L)$-*list decodable* if $\forall y \in \Sigma^n$, $|B(y, pn) \cap C| \leqslant L$. Equivalently, balls of radius $pn$ around the codewords cover every point $\leqslant L$ times. ("almost-disjoint" packing)

# List decoding

Relaxed goal: From $c + e$, the codeword $c$ is determined up to ambiguity $L$ (a large but fixed constant, independent of $n$)

## Definition (List-decodability)

A code $C \subset \Sigma^n$ is $(p, L)$-*list decodable* if $\forall y \in \Sigma^n$, $|B(y, pn) \cap C| \leqslant L$. Equivalently, balls of radius $pn$ around the codewords cover every point $\leqslant L$ times. ("almost-disjoint" packing)

Above is only a combinatorial notion.

- No guarantee that we can find $B(y, pn) \cap C$ efficiently.

# Combinatorics of list decoding

- $R_L(p) =$ largest rate of binary $(p, L)$-list decodable code family.
- $R_L^{\text{lin}}(p) =$ analogous quantity for *binary linear* codes.
- $R_{L,q}(p)$ and $R_{L,q}^{\text{lin}}(p)$ analogs for $q$-ary codes.

## This talk

Understanding above quantities, specifically lower bounding $R_{L,q}^{\text{lin}}(p)$

- list-decodability of *random* linear codes

Focus on $q = 2$; our proof generalizes (with $h_q(\cdot)$ replacing $h(\cdot)$).

# Shannon capacity still a limit

$R_L(p) \leqslant 1 - h(p)$

- Pick $y$ u.a.r. from $\{0,1\}^n$.
- $\mathbb{E}_y[|B(y,pn) \cap C|] = |C|\mathrm{Vol}(n,pn)/2^n \geqslant |C|2^{(h(p)-1-o(1))n}$.

# Shannon capacity still a limit

$R_L(p) \leqslant 1 - h(p)$

- Pick $y$ u.a.r. from $\{0,1\}^n$.
- $\mathbb{E}_y[|B(y,pn) \cap C|] = |C|\mathrm{Vol}(n,pn)/2^n \geqslant |C|2^{(h(p)-1-o(1))n}$.

Surprisingly (?)

$$\limsup_{L\to\infty} R_L(p) = \limsup_{L\to\infty} R_L^{\mathrm{lin}}(p) = 1 - h(p) \ .$$

(Equals $1 - h_q(p)$ in $q$-ary case.)

Allowing for list decoding, we can (non-constructively) approach Shannon capacity even for worst-case errors.

# Existence of list-decodable codes

**Theorem (Zyablov and Pinsker'81, Elias'91)**

*For* $p \in (0, 1/2)$, $R_L(p) \geqslant 1 - h(p) - 1/L$.

# Existence of list-decodable codes

### Theorem (Zyablov and Pinsker'81, Elias'91)

*For* $p \in (0, 1/2)$, $R_L(p) \geqslant 1 - h(p) - 1/L$.

### Proof.

Random coding: Pick $M = 2^{(1-h(p)-1/L)n}$ codewords u.a.r. from $\{0,1\}^n$.
Will show that resulting code $C$ is $(p, L)$-list decodable w.h.p.

# Existence of list-decodable codes

## Theorem (Zyablov and Pinsker'81, Elias'91)

*For $p \in (0, 1/2)$, $R_L(p) \geqslant 1 - h(p) - 1/L$.*

## Proof.

Random coding: Pick $M = 2^{(1-h(p)-1/L)n}$ codewords u.a.r. from $\{0,1\}^n$. Will show that resulting code $C$ is $(p, L)$-list decodable w.h.p.

- Fix $y \in \{0,1\}^n$ and a subset $S$ of $L+1$ codewords.
- Prob. that all codewords in $S$ fall in $B(y, pn)$ equals $\left(\frac{\mathrm{Vol}(n,pn)}{2^n}\right)^{L+1} \leqslant 2^{(h(p)-1)(L+1)n}$
- Union bound over $2^n$ $y$'s and $\leqslant M^{L+1}$ subsets $S$ shows that
$$\Pr[C \text{ is not } (p, L)\text{-list decodable}] \leqslant e^{-\Omega(n)}.$$

□

# What about linear codes?

Random linear code $C$: pick a random matrix $G \in \mathbb{F}_2^{n \times k}$; $\qquad (k = Rn)$

Set $C = \{Gx \mid x \in \mathbb{F}_2^k\}$.

# What about linear codes?

Random linear code $C$: pick a random matrix $G \in \mathbb{F}_2^{n \times k}$; $\quad (k = Rn)$
Set $C = \{Gx \mid x \in \mathbb{F}_2^k\}$.

- For a subset $\{x_1, x_2, \ldots, x_{L+1}\}$, the codewords $Gx_1, \ldots, Gx_{L+1}$ are *not* in general independent.

# What about linear codes?

Random linear code $C$: pick a random matrix $G \in \mathbb{F}_2^{n \times k}$; $\qquad (k = Rn)$
Set $C = \{Gx \mid x \in \mathbb{F}_2^k\}$.

- For a subset $\{x_1, x_2, \ldots, x_{L+1}\}$, the codewords $Gx_1, \ldots, Gx_{L+1}$ are *not* in general independent.
- Any $(L+1)$-sized set has a subset of $\geqslant \log_2(L+1)$ linearly independent vectors.
  Images of these under random $G$ **are** independent.

# What about linear codes?

Random linear code $C$: pick a random matrix $G \in \mathbb{F}_2^{n \times k}$; $\quad (k = Rn)$
Set $C = \{Gx \mid x \in \mathbb{F}_2^k\}$.

- For a subset $\{x_1, x_2, \ldots, x_{L+1}\}$, the codewords $Gx_1, \ldots, Gx_{L+1}$ are *not* in general independent.
- Any $(L+1)$-sized set has a subset of $\geqslant \log_2(L+1)$ linearly independent vectors.
  Images of these under random $G$ **are** independent.
- Union bound over centers $y$ and $\log_2(L+1)$-sized sets of linearly independent elements in $\mathbb{F}_2^k$.
- Similar calculation, with $\log_2(L+1)$ replacing $L$

**Theorem (Zyablov and Pinsker'81)**

*For* $p \in (0, 1/2)$, $R_L^{\text{lin}}(p) \geqslant 1 - h(p) - \frac{1}{\log_2(L+1)}$.

# Linear vs. non-linear

Stated in different notation:

1. Random $q$-ary code of rate $1 - h_q(p) - \varepsilon$ is $(p, O(1/\varepsilon))$-list decodable w.h.p.
2. Random $q$-ary *linear* code of rate $1 - h_q(p) - \varepsilon$ is $(p, q^{O(1/\varepsilon)})$-list decodable w.h.p.

**Motivation of this work**

Is this exponential discrepancy in list size inherent,
or an artifact of the proof technique?

Conjectured to be the latter [Elias'91]

# Main Result

**Theorem**

*For every prime power $q$, $p \in (0, 1 - 1/q)$, and $\varepsilon > 0$,*
*a random $q$-ary linear code of rate $1 - h_q(p) - \varepsilon$*
*is $(p, a_{p,q}/\varepsilon)$-list decodable with $1 - \exp(-\Omega(n))$ probability.*

# A previous result

**Theorem (G., Håstad, Sudan, Zuckerman'02)**

*For every $p \in (0, 1/2)$ and $\varepsilon > 0$, there exists a binary linear code family of rate $1 - h(p) - \varepsilon$ that is $(p, 1/\varepsilon)$-list decodable.*

# A previous result

**Theorem (G., Håstad, Sudan, Zuckerman'02)**

*For every $p \in (0, 1/2)$ and $\varepsilon > 0$, there exists a binary linear code family of rate $1 - h(p) - \varepsilon$ that is $(p, 1/\varepsilon)$-list decodable.*

**Comments**

- *Not* a high probability result. Existence proof via semi-random method.
- Applies only to *binary* linear codes.
- Conjectured that both restrictions can be removed.

# Digression: Lower bound on list size

[Blinovsky'86] $R_L(p) < 1 - h(p)$ for every fixed $L$.

- Unbounded list size needed to approach capacity $1 - h(p)$.
- Existence of $(p, L)$-list decodable code of rate $1 - h(p) - \varepsilon$ implies $L \geqslant \Omega(\log(1/\varepsilon))$.

# Digression: Lower bound on list size

[Blinovsky'86] $R_L(p) < 1 - h(p)$ for every fixed $L$.

- Unbounded list size needed to approach capacity $1 - h(p)$.
- Existence of $(p, L)$-list decodable code of rate $1 - h(p) - \varepsilon$ implies $L \geqslant \Omega(\log(1/\varepsilon))$.

**Open question**

Close (or shrink) the exponential gap between $\Omega(\log(1/\varepsilon))$ lower bound and $O(1/\varepsilon)$ upper bound.

# Digression: Lower bound on list size

[Blinovsky'86] $R_L(p) < 1 - h(p)$ for every fixed $L$.

- Unbounded list size needed to approach capacity $1 - h(p)$.
- Existence of $(p, L)$-list decodable code of rate $1 - h(p) - \varepsilon$ implies $L \geqslant \Omega(\log(1/\varepsilon))$.

**Open question**

Close (or shrink) the exponential gap between $\Omega(\log(1/\varepsilon))$ lower bound and $O(1/\varepsilon)$ upper bound.

- My guess is $\Theta(1/\varepsilon)$ is closer to the truth.
- For *random* codes, $O(1/\varepsilon)$ list size bound is tight.
  - [Rudra'09] W.h.p. a *random* rate $(1 - h(p) - \varepsilon)$ code is *not* $(p, c_p/\varepsilon)$-list decodable
  - [G.-Narayanan'10] Same holds for *random linear* codes

# Rest of the talk

Proof of main theorem (for binary codes)

**Theorem**

For every $p \in (0, 1/2)$, and $\varepsilon > 0$, a random linear code $C \subseteq \mathbb{F}_2^n$ of rate $1 - h(p) - \varepsilon$ is $(p, a_p/\varepsilon)$-list decodable with $1 - \exp(-\Omega(n))$ probability.

# Shortcoming of earlier proof

An $(L+1)$-element set $\{x_1, x_2, \ldots, x_{L+1}\}$ has $\ell \geqslant \log_2(L+1)$ linearly independent elements (say $x_1, \ldots, x_\ell$).
We used

$$\Pr[Gx_1, Gx_2, \ldots, Gx_{L+1} \text{ all lie in } B(y, pn)]$$
$$\leqslant \Pr[Gx_1, Gx_2, \ldots, Gx_\ell \text{ all lie in } B(y, pn)] = 2^{(h(p)-1)\ell n} \ .$$

Wasteful; ignores all remaining events $Gx_i \in B(y, pn)$ for $i > \ell$.

# Shortcoming of earlier proof

An $(L+1)$-element set $\{x_1, x_2, \ldots, x_{L+1}\}$ has $\ell \geqslant \log_2(L+1)$ linearly independent elements (say $x_1, \ldots, x_\ell$).
We used

$$\begin{aligned}&\Pr[Gx_1, Gx_2, \ldots, Gx_{L+1} \text{ all lie in } B(y, pn)]\\ &\leqslant \Pr[Gx_1, Gx_2, \ldots, Gx_\ell \text{ all lie in } B(y, pn)] = 2^{(h(p)-1)\ell n} \ .\end{aligned}$$

Wasteful; ignores all remaining events $Gx_i \in B(y, pn)$ for $i > \ell$.

**Key issue: Correlation of linear spaces and Hamming balls**

If we pick $\ell$ random vectors from $B(0, pn) \subset \mathbb{F}_2^n$, what is the probability that $\geqslant L$ vectors from their $\mathbb{F}_2$-span lie in $B(0, pn)$? (Here $\ell \leqslant L \leqslant 2^\ell$.)

# Moving center to origin

Let $R = 1 - h(p) - \varepsilon$ and $L = c_p/\varepsilon$.
It suffices to prove for random $C$ of dimension $Rn$:

$$\begin{aligned} \Pr_C\left[\exists y,\ |B(y,pn)\cap C| \geqslant L\right] &\leqslant 2^{-n} \\ \Longleftarrow \Pr_{C,y}\left[|B(y,pn)\cap C| \geqslant L\right] &\leqslant 2^{-2n} \end{aligned}$$

# Moving center to origin

Let $R = 1 - h(p) - \varepsilon$ and $L = c_p/\varepsilon$.
It suffices to prove for random $C$ of dimension $Rn$:

$$\begin{aligned} \Pr_C\left[\exists y,\ |B(y,pn) \cap C| \geqslant L\right] &\leqslant 2^{-n} \\ \Longleftarrow \Pr_{C,y}\left[|B(y,pn) \cap C| \geqslant L\right] &\leqslant 2^{-2n} \end{aligned}$$

$$\Longleftrightarrow \Pr_{C,y}\left[|B(0,pn) \cap (C+y)| \geqslant L\right] \leqslant 2^{-2n}$$

# Moving center to origin

Let $R = 1 - h(p) - \varepsilon$ and $L = c_p/\varepsilon$.
It suffices to prove for random $C$ of dimension $Rn$:

$$\begin{aligned} \Pr_C\big[\exists y,\ |B(y,pn)\cap C| \geqslant L\big] &\leqslant 2^{-n} \\ \Longleftarrow \Pr_{C,y}\big[|B(y,pn)\cap C| \geqslant L\big] &\leqslant 2^{-2n} \end{aligned}$$

$$\begin{aligned} \Longleftrightarrow \ \Pr_{C,y}\big[|B(0,pn)\cap (C+y)| \geqslant L\big] &\leqslant 2^{-2n} \\ \Longleftarrow \ \Pr_{C,y}\big[|B(0,pn)\cap \mathrm{span}(C,y)| \geqslant L\big] &\leqslant 2^{-2n} \\ \Longleftarrow \ \Pr_{C^*}\big[|B(0,pn)\cap C^*| \geqslant L\big] &\leqslant 2^{-2n} \end{aligned}$$

where $C^*$ is a random linear code of dimension $Rn+1$. Call it $C$.

# Breaking down by rank

$$\Pr_C\left[\ |B(0,pn)\cap C| \geqslant L\right] \quad \leqslant \quad \sum_{W \in \binom{B(0,pn)}{L}} \Pr_C[W \subseteq C]$$

# Breaking down by rank

$$\begin{aligned}\Pr_C\left[\, |B(0,pn)\cap C| \geqslant L\right] &\leqslant \sum_{W \in \binom{B(0,pn)}{L}} \Pr_C[W \subseteq C] \\ &\leqslant \sum_{\ell=\log L}^{L} |\mathcal{F}_\ell| \left(\frac{2^{Rn}}{2^n}\right)^{\ell}\end{aligned}$$

where

$$\mathcal{F}_\ell = \left\{ U \in \binom{B(0,pn)}{\ell} \;\middle|\; U \text{ is linearly indep. } \& \; |\mathrm{span}(U) \cap B(0,pn)| \geqslant L \right\}$$

# Breaking down by rank

$$\Pr_C\left[\ |B(0,pn)\cap C| \geqslant L\right] \ \leqslant \sum_{W\in\binom{B(0,pn)}{L}} \Pr_C[W\subseteq C]$$

$$\leqslant \sum_{\ell=\log L}^{L} |\mathcal{F}_\ell|\left(\frac{2^{Rn}}{2^n}\right)^{\ell} = \sum_{\ell=\log L}^{L} \frac{|\mathcal{F}_\ell|}{2^{h(p)n\ell}}\ 2^{-\varepsilon n\ell}$$

where

$\mathcal{F}_\ell = \left\{U \in \binom{B(0,pn)}{\ell} \ \middle|\ U \text{ is linearly indep. \& } |\mathrm{span}(U)\cap B(0,pn)| \geqslant L\right\}$

# Breaking down by rank

$$\Pr_C\left[\ |B(0,pn)\cap C| \geqslant L\right] \ \leqslant \sum_{W\in\binom{B(0,pn)}{L}} \Pr_C[W\subseteq C]$$

$$\leqslant \sum_{\ell=\log L}^{L} |\mathcal{F}_\ell|\left(\frac{2^{Rn}}{2^n}\right)^{\ell} = \sum_{\ell=\log L}^{L} \frac{|\mathcal{F}_\ell|}{2^{h(p)n\ell}}\ 2^{-\varepsilon n\ell}$$

where

$\mathcal{F}_\ell = \left\{U \in \binom{B(0,pn)}{\ell} \ \middle|\ U \text{ is linearly indep. \& } |\mathrm{span}(U)\cap B(0,pn)| \geqslant L\right\}$

- For large $\ell \geqslant 10/\varepsilon$, the trivial bound $|\mathcal{F}_\ell| \leqslant 2^{h(p)n\ell}$ suffices.
- For $\ell < 10/\varepsilon$, we have $L > A_p \cdot \ell$, and we prove $\frac{|\mathcal{F}_\ell|}{2^{h(p)n\ell}} \leqslant 2^{-5n}$.

# Balls vs. Subspaces

## Main technical theorem

For every $p \in (0, 1/2)$, there exists $A' = A_p < \infty$ such that for all $\ell$, and sufficiently large $n$, if $n$-bit strings $x_1, x_2, \ldots, x_\ell$ are picked u.a.r and independently from $B(0, pn)$,

$$\Pr\left[\ |\mathrm{span}(x_1, \ldots, x_\ell) \cap B(0, pn)| > A' \cdot \ell\right] \leqslant 2^{-5n} \ .$$

# Balls vs. Subspaces

## Main technical theorem

For every $p \in (0, 1/2)$, there exists $A' = A_p < \infty$ such that for all $\ell$, and sufficiently large $n$, if $n$-bit strings $x_1, x_2, \ldots, x_\ell$ are picked u.a.r and independently from $B(0, pn)$,

$$\Pr\left[ \, |\mathrm{span}(x_1, \ldots, x_\ell) \cap B(0, pn)| > A' \cdot \ell \right] \leqslant 2^{-5n} \; .$$

Implies $|\mathcal{F}_\ell| \leqslant 2^{h(p)n\ell} \cdot 2^{-5n}$ for $L > A' \cdot \ell$.

# Balls vs. Subspaces

## Main technical theorem

For every $p \in (0, 1/2)$, there exists $A' = A_p < \infty$ such that for all $\ell$, and sufficiently large $n$, if $n$-bit strings $x_1, x_2, \ldots, x_\ell$ are picked u.a.r and independently from $B(0, pn)$,

$$\Pr\left[\ |\mathrm{span}(x_1, \ldots, x_\ell) \cap B(0, pn)| > A' \cdot \ell\right] \leqslant 2^{-5n} \ .$$

Implies $|\mathcal{F}_\ell| \leqslant 2^{h(p)n\ell} \cdot 2^{-5n}$ for $L > A' \cdot \ell$.

- Fix $T \subseteq \mathbb{F}_2^\ell \setminus \{0, e_1, \ldots, e_\ell\}$ of size $(A' - 1)\ell = A \cdot \ell$.
- Upper bound probability that all vectors $(X_v)_{v \in T}$ lie in $B(0, pn)$ (where $X_v = \sum_{i=1}^{\ell} v_i x_i$)
- Union bound over all choices of $T$ (at most $2^{O(\ell^2)}$)

# An idealized case

Suppose $T$ has many ($d = d_p$, think $10$) vectors with *disjoint* support.

Concretely, say $(X_v)_{v \in T}$ contains the linear combinations

$$x_1 + x_2, \qquad x_3 + x_4, \quad \cdots \quad x_{2d-1} + x_{2d} \ .$$

# An idealized case

Suppose $T$ has many ($d = d_p$, think $10$) vectors with *disjoint* support.

Concretely, say $(X_v)_{v \in T}$ contains the linear combinations

$$x_1 + x_2, \qquad x_3 + x_4, \quad \cdots \quad x_{2d-1} + x_{2d} \ .$$

The events that these belong to $B(0, pn)$ are independent, and each occurs with probability $\leqslant 2^{-\delta_p n}$

- Each is essentially a random point in $B(0, 2p(1-p)n)$

Prob. that all of them lie in $B(0, pn)$ is $\leqslant \left(2^{-\delta_p n}\right)^d \leqslant 2^{-6n}$.

# An idealized case

Suppose $T$ has many ($d = d_p$, think $10$) vectors with *disjoint* support.

Concretely, say $(X_v)_{v \in T}$ contains the linear combinations

$$x_1 + x_2, \qquad x_3 + x_4, \quad \cdots \quad x_{2d-1} + x_{2d} \ .$$

The events that these belong to $B(0, pn)$ are independent, and each occurs with probability $\leqslant 2^{-\delta_p n}$

- Each is essentially a random point in $B(0, 2p(1-p)n)$

Prob. that all of them lie in $B(0, pn)$ is $\leqslant \left(2^{-\delta_p n}\right)^d \leqslant 2^{-6n}$.

Can we always find many such disjoint vectors?

# Hunting for independence

Can we always find many such disjoint vectors?

# Hunting for independence

Can we always find many such disjoint vectors?

- Of course not! A family might not even have *two* disjoint sets

# Hunting for independence

Can we always find many such disjoint vectors?

- Of course not! A family might not even have *two* disjoint sets

Disjointness is too strong and unnecessary.

"Ordered" disjointness or increasing chain is enough.

- Eg., $x_1 + x_2$, $x_1 + x_3 + x_4$, $x_2 + x_3 + x_4 + x_5 + x_6$, $x_1 + x_3 + x_5 + x_7 + x_8$, $\cdots$

# Hunting for independence

Can we always find many such disjoint vectors?

- Of course not! A family might not even have *two* disjoint sets

Disjointness is too strong and unnecessary.

"Ordered" disjointness or increasing chain is enough.

- Eg., $x_1 + x_2$, $x_1 + x_3 + x_4$, $x_2 + x_3 + x_4 + x_5 + x_6$, $x_1 + x_3 + x_5 + x_7 + x_8$, $\cdots$
- Prob. that each linear combination is in $B(0, pn)$ *conditioned on choice of $x_i$'s that occur in previous combinations* is also small. Why?

# Hunting for independence

Can we always find many such disjoint vectors?

- Of course not! A family might not even have *two* disjoint sets

Disjointness is too strong and unnecessary.

"Ordered" disjointness or increasing chain is enough.

- Eg., $x_1 + x_2$, $x_1 + x_3 + x_4$, $x_2 + x_3 + x_4 + x_5 + x_6$, $x_1 + x_3 + x_5 + x_7 + x_8$, $\cdots$
- Prob. that each linear combination is in $B(0, pn)$ *conditioned on choice of $x_i$'s that occur in previous combinations* is also small. Why?

## Relaxed goal

In any family of $A \cdot \ell$ subsets of $\{1, 2, \ldots, \ell\}$, can we always find a 2-increasing chain of size $10$, i.e., a sequence of $10$ sets each of which has $\geqslant 2$ fresh elements (that don't belong to previous sets in the sequence)?

# After increasing chains

**Relaxed goal**

In any family of $A \cdot \ell$ subsets of $\{1, 2, \ldots, \ell\}$, can we always find a sequence of $10$ sets each of which has $\geqslant 2$ elements that don't belong to previous sets in the sequence?

# After increasing chains

## Relaxed goal

In any family of $A \cdot \ell$ subsets of $\{1, 2, \ldots, \ell\}$, can we always find a sequence of $10$ sets each of which has $\geqslant 2$ elements that don't belong to previous sets in the sequence?

Unfortunately no!

# After increasing chains

**Relaxed goal**

In any family of $A \cdot \ell$ subsets of $\{1, 2, \ldots, \ell\}$, can we always find a sequence of $10$ sets each of which has $\geqslant 2$ elements that don't belong to previous sets in the sequence?

Unfortunately no! Take the family to be all $\ell - 2$ element subsets.

# $2$-increasing chains in hiding

So are these linear combinations $(X_v)_{v \in \mathbb{F}_2^\ell, |v| = \ell - 2}$ in fact bad?

# $2$-increasing chains in hiding

So are these linear combinations $(X_v)_{v \in \mathbb{F}_2^\ell, |v| = \ell - 2}$ in fact bad?

- If all these lie in $B(0, pn)$, then $(X_v)_{v \in \mathbb{F}_2^\ell, |v| = 2}$ must all lie in $B(w, pn)$ where $w = x_1 + x_2 + \cdots + x_\ell$.

# 2-increasing chains in hiding

So are these linear combinations $(X_v)_{v \in \mathbb{F}_2^\ell, |v|=\ell-2}$ in fact bad?

- If all these lie in $B(0, pn)$, then $(X_v)_{v \in \mathbb{F}_2^\ell, |v|=2}$ must all lie in $B(w, pn)$ where $w = x_1 + x_2 + \cdots + x_\ell$.
- $\{v \in \mathbb{F}_2^\ell \mid |v| = 2\}$ has a long 2-increasing chain (in fact $\approx \ell/2$ disjoint vectors), but now the center $w$ is not $0$ but depends on $x_i$'s.

# 2-increasing chains in hiding

So are these linear combinations $(X_v)_{v \in \mathbb{F}_2^\ell, |v|=\ell-2}$ in fact bad?

- If all these lie in $B(0, pn)$, then $(X_v)_{v \in \mathbb{F}_2^\ell, |v|=2}$ must all lie in $B(w, pn)$ where $w = x_1 + x_2 + \cdots + x_\ell$.
- $\{v \in \mathbb{F}_2^\ell \mid |v| = 2\}$ has a long 2-increasing chain (in fact $\approx \ell/2$ disjoint vectors), but now the center $w$ is not $0$ but depends on $x_i$'s.
- Turns out this is okay.

**Lemma (Increasing chains are good for *every* center)**

*Let $\mathcal{C} \subseteq \mathbb{F}_2^\ell$ be a 2-increasing chain of size $d$.*
*Then the probability (over choice of $x_1, \ldots, x_\ell$ from $B(0, pn)$) that* **there exists** *$y \in \mathbb{F}_2^n$ such that all $(X_v)_{v \in \mathcal{C}}$ belong to $B(y, pn)$ is at most $2^n \cdot 2^{-\delta_p dn}$ (and thus $\leqslant 2^{-6n}$ if $d \geqslant d_p$).*

# Translating to find $2$-increasing chain

Can *always* find a translate that has a long $2$-increasing chain.

## Theorem

*For every subset $T \subseteq \mathbb{F}_2^\ell$ there exists a $z \in \mathbb{F}_2^\ell$ such that $T + z$ contains a $2$-increasing chain $\mathcal{C}$ of size $\Omega\left(\log \frac{|T|}{\ell}\right)$.*

## Corollary

*We can get a $2$-increasing chain in a translate of $T$ of size $d_p$ if $|T| \geqslant A_p \ell$.*

# Translating to find $2$-increasing chain

Can *always* find a translate that has a long $2$-increasing chain.

**Theorem**

*For every subset $T \subseteq \mathbb{F}_2^\ell$ there exists a $z \in \mathbb{F}_2^\ell$ such that $T + z$ contains a $2$-increasing chain $\mathcal{C}$ of size $\Omega\left(\log \frac{|T|}{\ell}\right)$.*

**Corollary**

*We can get a $2$-increasing chain in a translate of $T$ of size $d_p$ if $|T| \geqslant A_p \ell$.*

$$(X_v)_{v \in T} \subset B(0, pn) \Rightarrow (X_v)_{v \in T+z} \subset B(X_z, pn) \Rightarrow (X_v)_{v \in \mathcal{C}} \subset B(X_z, pn)$$

and last event occurs with $\leqslant 2^{-\Omega(n)}$ probability.

So it remains to prove the above theorem.

# Proof by induction

We'll find a translate with $2$-increasing chain of size $\log_4 \frac{|T|}{\ell+1}$.

## Lemma (Sauer-Shelah (-Perles-Vapnik-Chervonenkis))

*If $T \subseteq \mathbb{F}_2^\ell$ has size $> \ell + 1$, then there exist $1 \leqslant i_1 < i_2 \leqslant \ell$ such that $\{(u_{i_1}, u_{i_2}) \mid u \in T\} = \{0,1\}^2$.*

If $|T| \leqslant \ell + 1$, there is nothing to prove. Otherwise, apply above lemma and let $\{i_1, i_2\} = \{1, 2\}$.

- All $4$ possibilities occur in first two positions of strings in $T$. Let $(0,0)$ be most frequent.
- Let $T' = \{v \in \mathbb{F}_2^{\ell-2} \mid (0,0,v) \in T\}$. Note $|T'| \geqslant |T|/4$.

# Proof by induction

We'll find a translate with $2$-increasing chain of size $\log_4 \frac{|T|}{\ell+1}$.

**Lemma (Sauer-Shelah (-Perles-Vapnik-Chervonenkis))**

*If $T \subseteq \mathbb{F}_2^\ell$ has size $> \ell + 1$, then there exist $1 \leqslant i_1 < i_2 \leqslant \ell$ such that $\{(u_{i_1}, u_{i_2}) \mid u \in T\} = \{0,1\}^2$.*

If $|T| \leqslant \ell + 1$, there is nothing to prove. Otherwise, apply above lemma and let $\{i_1, i_2\} = \{1, 2\}$.

- All $4$ possibilities occur in first two positions of strings in $T$. Let $(0,0)$ be most frequent.
- Let $T' = \{v \in \mathbb{F}_2^{\ell-2} \mid (0,0,v) \in T\}$. Note $|T'| \geqslant |T|/4$.
- Get $2$-increasing chain $\mathcal{C}'$ in $T' + z'$ by induction.
- Let $\mathcal{C} = \{(0,0,u) \mid u \in \mathcal{C}'\}$ and $z = (0,0,z')$.

# Proof by induction

We'll find a translate with 2-increasing chain of size $\log_4 \frac{|T|}{\ell+1}$.

**Lemma (Sauer-Shelah (-Perles-Vapnik-Chervonenkis))**

*If $T \subseteq \mathbb{F}_2^\ell$ has size $> \ell + 1$, then there exist $1 \leqslant i_1 < i_2 \leqslant \ell$ such that $\{(u_{i_1}, u_{i_2}) \mid u \in T\} = \{0,1\}^2$.*

If $|T| \leqslant \ell + 1$, there is nothing to prove. Otherwise, apply above lemma and let $\{i_1, i_2\} = \{1, 2\}$.

- All $4$ possibilities occur in first two positions of strings in $T$. Let $(0,0)$ be most frequent.
- Let $T' = \{v \in \mathbb{F}_2^{\ell-2} \mid (0,0,v) \in T\}$. Note $|T'| \geqslant |T|/4$.
- Get $2$-increasing chain $\mathcal{C}'$ in $T' + z'$ by induction.
- Let $\mathcal{C} = \{(0,0,u) \mid u \in \mathcal{C}'\}$ and $z = (0,0,z')$.
- Let $w \in T$ be such that $(w_1, w_2) = (1,1)$.
- $\mathcal{C}$ followed by $w + z$ is a $2$-increasing chain in $T + z$. $\square$

# Concluding remarks

- $q$-ary case similar, with a slightly non-standard generalization of Sauer-Shelah lemma.
- Random linear codes are nearly as good as random codes w.r.t convergence to "capacity" as function of list size.
- Technical core of the proof: A strong upper bound on probability that $\ell$ random vectors have many elements from their span lie in a Hamming ball.
- Best possible list-size for rate $1 - h(p) - \varepsilon$? Big gap between $\log(1/\varepsilon)$ lower bound and $1/\varepsilon$ upper bound.