Some recent results on local testing of sparse linear codes

Swastik Kopparty (MIT)
Shubhangi Saraf (MIT)

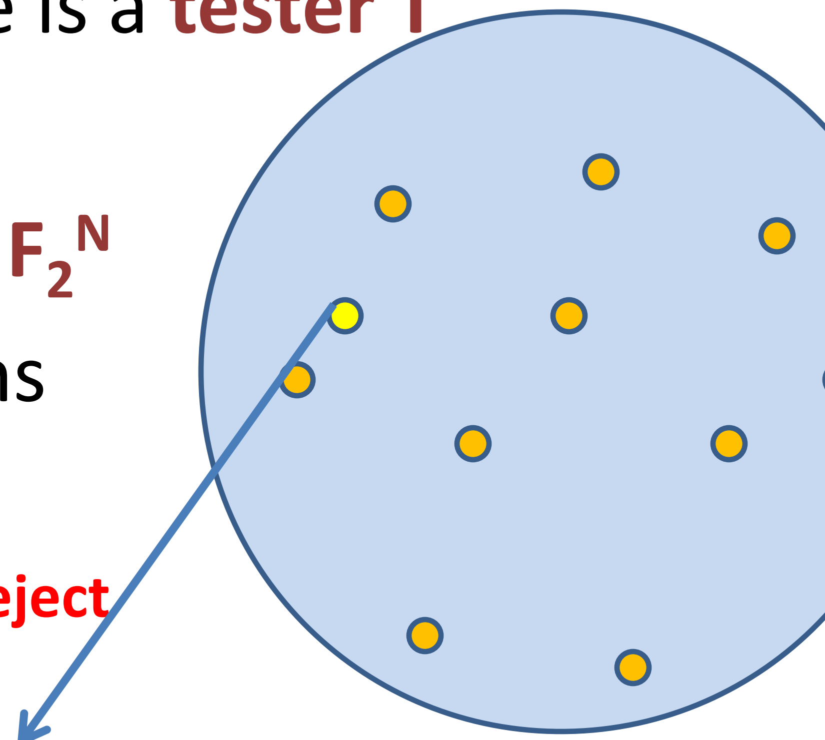
Locally Testable Codes

- Let $\mathbf{C} \subset \mathbf{F_2}^{\mathbf{N}}$ be a linear code.
- **C** is *locally testable* if there is a **tester T** such that :

  Given oracle access to $\mathbf{r} \in \mathbf{F_2}^{\mathbf{N}}$

  T queries **r** in few locations

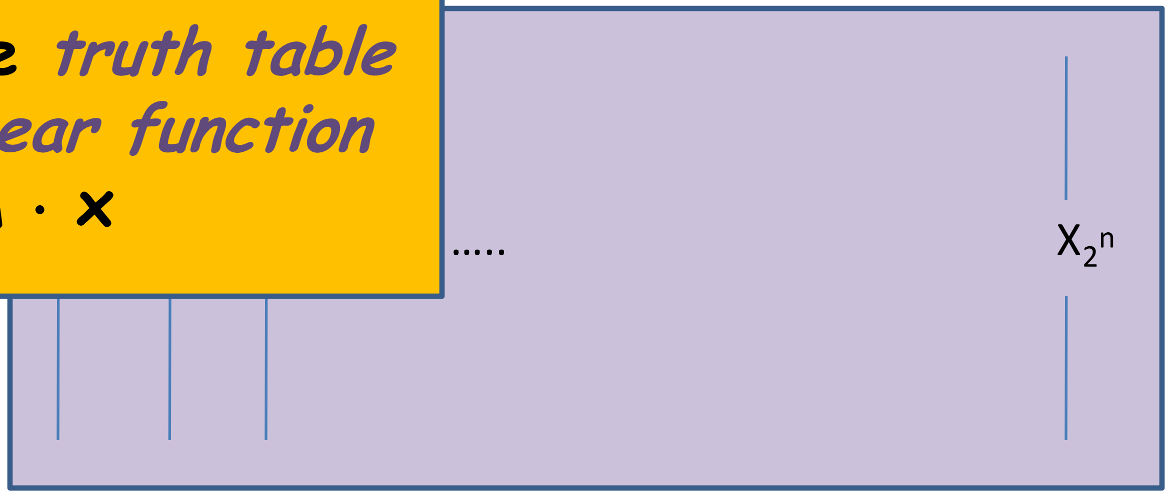
  - If $\mathbf{r} \in \mathbf{C}$, then **Accept**
  - If **r is** ϵ**-far from C**, then **Reject**



r 

The Hadamard Code

E(m) is the *truth table* of the *linear function* $m \cdot x$



$n \times 2^n$

E(m) = $m \cdot X_1$ $m \cdot X_2$ … … $m \cdot X_{2^n}$

R = r_1 r_2 … … r_{2^n}

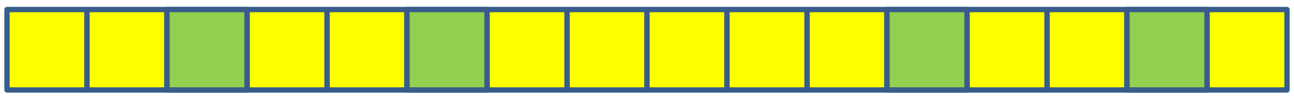
The received word R is *some function* on all on F_2^n

Linearity Testing of Boolean Functions

Given oracle access to

$f : F_2^n \rightarrow F_2$

f



F_2^n

Test using few queries if **f** is linear.

- If f **is linear**, **Accept**
- If f is **ϵ-far** from all g that are linear, **Reject**

$Pr_x[f(x) \neq g(x)] > \epsilon$

BLR Linearity Test

- Choose x, y uniformly at random
- Query f(x), f(y) and f(x+y)
  - Check if
    - If **Yes**, then ***Accept***.
    - If **No**, then ***Reject***

$$f(x) + f(y) \overset{?}{=} f(x+y)$$

Theorem [BLR '90]:

Hadamard Code is locally testable

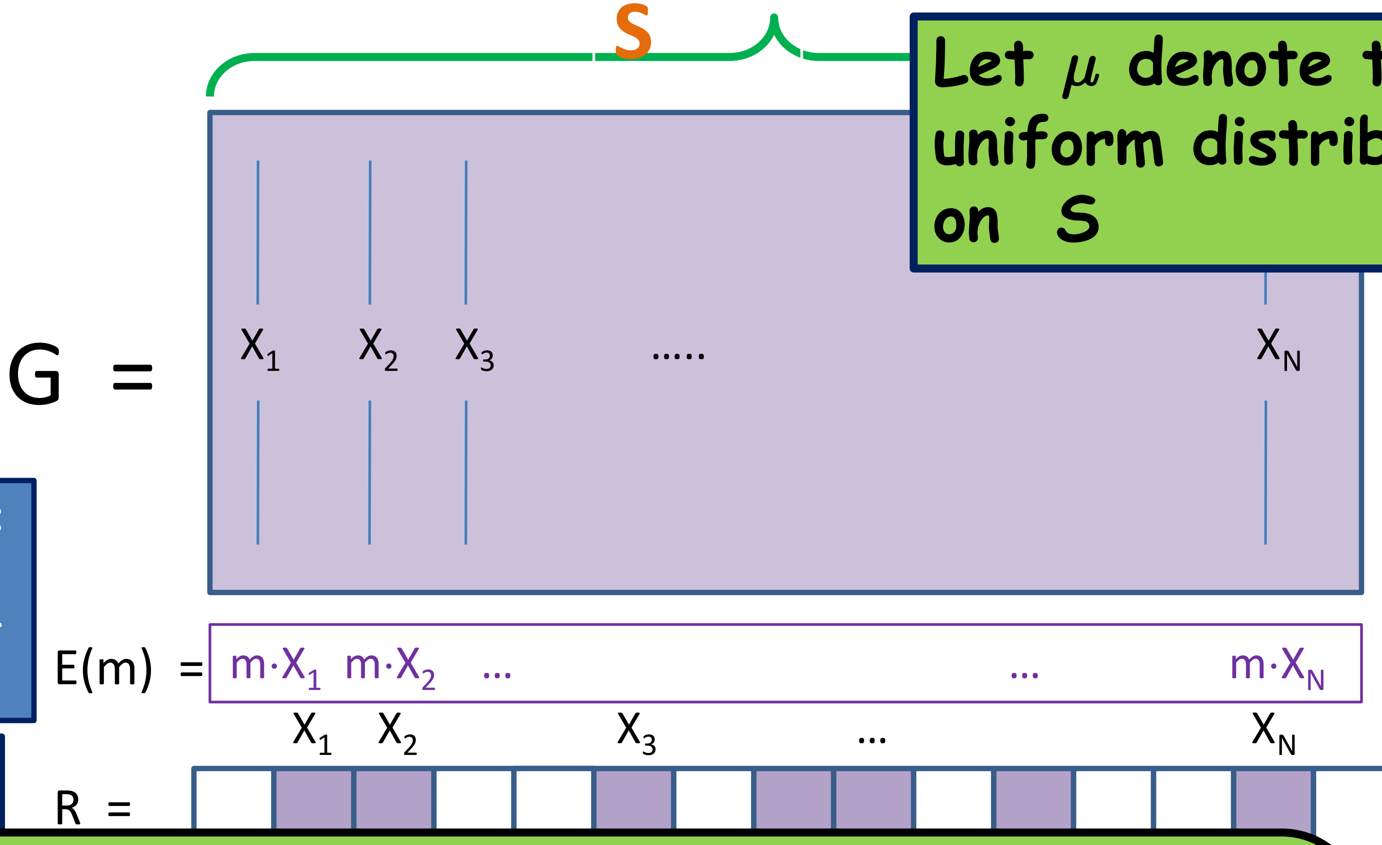
- If f **is linear**, then test ***accepts*** with probability 1.
- If *f is* ***ϵ-far*** *(under the uniform distribution)* from being linear, test ***rejects*** with probability $> \epsilon_0 > 0$

LTCs and Linearity Testing

G is a generator matrix for the code C

Each codeword: partial truth-table of a linear function

Received word



Let μ denote the uniform distribution on S

C is locally testable if (using queries to S)
Can distinguish between
1) R is a ***linear function on S***
2) R is ϵ-***far*** (on S) ***from all linear g***

$\Pr_{X \in \mu}[f(x) \neq g(x)] > \epsilon$

Talk Overview

- Locally testable codes ⇐ Testing linearity under some distribution μ

- Criterion for testing under μ

- Local List Decoding and Testing with high error

- Time Complexity
  - Dual BCH codes
  - connections to the noisy parity problem

Testing Linearity under General Distributions

- Given
  - a distribution μ over F_2^n
    - μ **distance** (g,h) = $\mathbf{Pr}_{x\in\mu}\,[g(x) \neq h(x)]$
  - Oracle access to $\mathbf{f : F_2^n \rightarrow F_2}$

A 3 query test actually works for all distributions! [HK07]

A Goal: If f ***is linear***, ***Accept***

If f is ***ϵ-far*** from all linear functions in ***μ distance***, ***Reject***

An odd consequence

Goal: If f ***is linear***, ***Accept***

If f is ***ϵ-far*** from all linear functions in ***<u>μ distance</u>***, ***Reject***

The moral

The tester should make queries essentially according to μ

Stronger Goal

Stronger Goal: With high probability, accept functions that are close to linear

"**Tolerant property testing**" [PRR]

Tolerant Linearity Testing

- Given
  - a distribution μ over F_2^n
  - Oracle access to $\mathbf{f : F_2^n \to F_2}$
- If f is ***close*** to linear in μ ***-distance***, then **Accept** with high probability
- If f is ***far*** from linear in μ ***-distance***, then **Reject** with noticeable probability

The BLR Linearity Tester is a Tolerant Tester for U_n

Connection to Locally Testable Codes

- For every **linear code C**, there is a distribution μ such that
  - **C** is locally testable ⇔ linearity is ***tolerantly*** testable under μ.

Uniform distribution on the columns of the generator matrix for C

- ***Tolerance crucial***

Talk Overview

- Locally testable codes ⇐ Testing linearity under some distribution μ

- Criterion for tolerant testing under μ

- Local List Decoding and Testing with high error

- Time Complexity
  - Dual BCH codes
  - connections to the noisy parity problem

Need for correlation

- Say the test queries $x_1, \ldots, x_k$

- Each query $x_i \in F_2^n$ essentially according to μ

- x_i's should satisfy some linear relation

- *A bare minimum for testing:*
  - The existence of such a correlated distribution.

Uniform Correlatability

- Definition: μ is **k-uniformly correlatable** if

there exists a joint distribution

$X_1 \; X_2 \; \ldots \; X_k$

$\boxed{\sum X_i = X}$

1. Each X_i is distributed as μ
2. $X = \sum X_i$ is distributed uniformly

Let $\mu^{(k)}$ denote this joint distribution

Theorem

If μ is k-uniformly correlatable, then linearity is tolerantly testable under μ in O(k) queries

Holds for tolerantly testing homomorphisms between any two abelian groups (under general distributions).

Tolerantly testable distributions

- Corollary: Linearity is tolerantly testable with a constant number of queries under:
  1. Product distributions
  2. Symmetric distributions supported on words of weight $\in [\gamma n, (1-\gamma) n]$
  3. Low Fourier-bias distributions
     - e.g. uniform distribution over a large random subset
       - "Sparse random linear codes are locally testable" [KS07]
       - Generalizes [KS07] to arbitrary groups

Theorem:

If $C \subseteq \{0,1\}^N$ is a linear code which is

1. Sparse: $|C| \leq N^c$
2. "unbiased": Each nonzero codeword has weight $\in (1/2 - N^{-\gamma}, \frac{1}{2} + N^{-\gamma})$

Then C is locally testable with constantly many queries.

Proof that Uniform Correlatability $\Longrightarrow$ testability

Recall:

Given distribution μ that is **k-uniformly correlatable**.

There exists $\quad X_1\ X_2\ \ldots\ X_k \qquad \boxed{\sum X_i = X}$

Such that

1. Each X_i is distributed as μ
2. $X = \sum X_i$ is distributed uniformly over F_2^n

- Let $\mu^{(k)}$ denote the **joint distribution $(X_1, \ldots, X_k)$**
- Let $\mu^{(k)} \mid \sum X_i = X$ denote the joint distribution of $(X_1, \ldots, X_k)$ conditioned on $\sum X_i = X$
- Let U_n denote the **uniform distribution on F_2^n**

Rough idea

- Use $\mu^{(k)}$ to generate ***correlated queries*** satisfying ***linear relations***.

- 2 carefully designed tests: **Test 1** and **Test 2**

TEST 1

- Sample X and Y indep. from $\mathbf{U_n}$. Let $Z = X+Y$
- Sample $(X_1, \ldots, X_k)$ from $\mu^{(k)} \mid \sum X_i = X$

  $(Y_1, \ldots, Y_k)$ from $\mu^{(k)} \mid \sum Y_i = Y$

  and $(Z_1, \ldots, Z_k)$ from $\mu^{(k)} \mid \sum Z_i = Z$

Check if $\sum \mathbf{f(X_i)} + \sum \mathbf{f(Y_i)} = \sum \mathbf{f(Z_i)}$

in spirit: the BLR test!

Rewriting Test 1

Defn: Let $\mathbf{h(X) = f(X_1) + \cdots + f(X_k)}$,

where $(X_1, \ldots, X_k) \in \mu^{(k)} \mid \sum X_i = X$

- h is a *probabilistic function*.
- Test 1 rewritten: Sample X, Y from U_n. Let Z=X+Y.

Check: $h(X) + h(Y) = h(Z)$

The BLR test!

Test 1 passes whp $\Rightarrow$

A related function ***h*** is close to a ***linear function g*** under the *uniform distribution*

<u>TEST 2</u>

- Sample Z from μ. Sample X, Y from $\mathbf{U_n}$ such that X+ Y = Z

- Sample $(X_1, \ldots, X_k)$ from $\boldsymbol{\mu}^{(\mathbf{k})} \mid \sum X_i = X$
 and $(Y_1, \ldots, Y_k)$ from $\boldsymbol{\mu}^{(\mathbf{k})} \mid \sum Y_i = Y$

Check if $\sum \mathbf{f(X_i) + \sum f(Y_i) = f(Z)}$

Understanding Test 2

Assume Test 1 passes whp. So h $\approx$ linear g.

Want to show: **for Z $\in \mu$, f(Z) $\approx$ g(Z)**

If Test 2 passes, $f(Z) \approx \sum f(X_i) + \sum f(Y_i)$

But by defn of h, $\sum f(X_i) + \sum f(Y_i) = h(X) + h(Y)$

Since Test 1 passes, $h(X) + h(Y) \approx g(X) + g(Y)$

Since g is linear $g(X) + g(Y) = g(Z)$

Test 1 passes whp $\Rightarrow$
A related function ***h*** is close to ***a linear function g*** under the *uniform distribution*

If Test 2 also passes whp $\Rightarrow$
f is close to ***the linear function g*** under the μ *Distribution*

To summarize

- "Extend" f defined on μ to h defined on F_2^n
  - uniform-correlatability
- Test if h is close to a linear function g under U_n
  - the BLR test
- Test if f is close to g under μ

Some Questions

- What distributions are correlatable?
- Under what distributions is linearity testable?
- Are all* sparse linear codes are locally testable?

Talk Overview

- Locally testable codes ⇐ Testing linearity under some distribution μ
- Criterion for tolerant testing under μ
- Local List Decoding and Testing with high error
- Time Complexity
  - Dual BCH codes
  - connections to the noisy parity problem

The *high error* regime

Recall: for **Local testability:**

If **$r \in C$,** then **Accept (with prob 1),**

If **r is ϵ-far from C,** then **Reject (with noticable probability)**

In the **high error regime:**

If **$\Delta(r, C) <$ ½ - ϵ,** then **Accept**

If **$\Delta(r, C) \approx$ ½ ,** then **Reject**

Distance estimation:

For $0 < \epsilon_2 < \epsilon_1 < ½$,

If **$\Delta(r, C) <$ ½ - ϵ_1 ,** then **Accept**

If **$\Delta(r, C) >$ ½ - ϵ_2 ,** then **Reject**

Theorem:

If $C \subseteq \{0,1\}^N$ is a linear code which is

1. Sparse: $|C| \leq N^c$
2. "unbiased": Each nonzero codeword has weight $\in (1/2 - N^{-\gamma}, \frac{1}{2} + N^{-\gamma})$

Then C is locally testable and locally list decodable from $\frac{1}{2} - \epsilon$ fraction errors using only poly$(1/\epsilon)$ queries.

Corollary:

Random sparse linear codes are locally testable and locally list decodable with $\frac{1}{2}-\epsilon$ fraction errors using only poly($1/\epsilon$) queries.

Dual BCH codes are locally testable and locally list decodable with $\frac{1}{2}-\epsilon$ fraction errors using only poly($1/\epsilon$) queries.

Proof

Reduce to the Hadamard Code!

The Hadamard code

- **[BCHKS'96]:** Fourier analysis proof of BLR Test
  - Hadamard Code is testable in the high error regime

- **[GL'89]:** Hadamard Code is locally list decodable up to **$1/2-\epsilon$** fraction errors with **poly($1/\epsilon$)** queries.

- **Distance estimation:** For **$0 < \epsilon_2 < \epsilon_1 < ½$,** In **poly($1/\epsilon_1 - \epsilon_2$)** queries, can distinguish between
  1. **($1/2 - \epsilon_1$)** close to a codeword
  2. **($1/2 - \epsilon_2$)** far from every codeword

Recall: Low error testing

- "Extend" f defined on μ to h defined on F_2^n
  - uniform-correlatability
- Test if h is close to a linear function g under U_n
  - the BLR test
- Test if f is close to g under μ

Recall: ***from f to h*** - Uniform Correlatability

<u>The problem in the high error regime:</u>

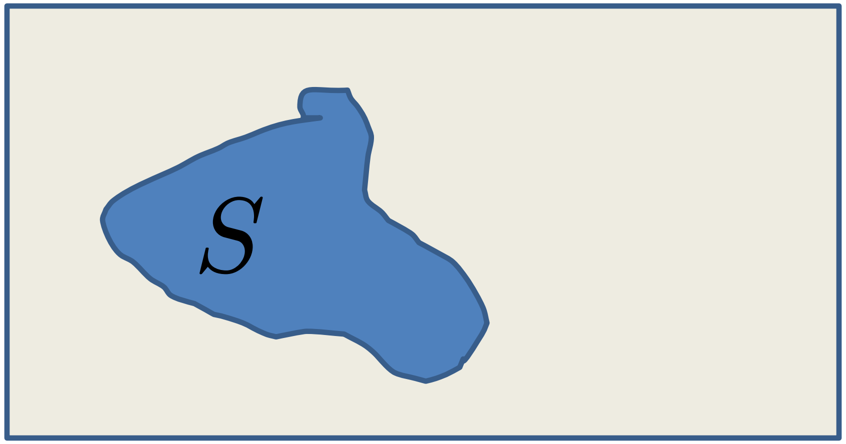
f could be *somewhat close* to linear, but h could be very far from linear.

So can't deduce anything about closeness of f from closeness of h ☹

Independent uniform correlatability

- C: Sparse, unbiased code
- S: Set of columns of generator matrix
  - S is a large set **($|S| \approx 2^{n/k}$)** with small Fourier bias **($\approx 2^{-n/10k}$).**

Sum of few *independent* samples from S
<u>nearly uniform</u>



Extending f to all of F_2^n

Defn: Let $h(X) = f(X_1) + \cdots + f(X_k)$,

where X_i are sampled independently from $\mu \mid \sum X_i = X$

Defn: Let $Corr_\mu(f,g) = 1 - 2\,\Delta_\mu(f,g)$

$$Corr_U(h,g) \approx Corr_{x \in \mu(k)}(h(X),g(X))$$

$$= Corr_{x1,..,xk \in \mu}(f(X_1)+ ..+f(X_k)\,,\, g(X_1) + ..+ g(X_k))$$

$$= [Corr_\mu(f,g)]^k$$

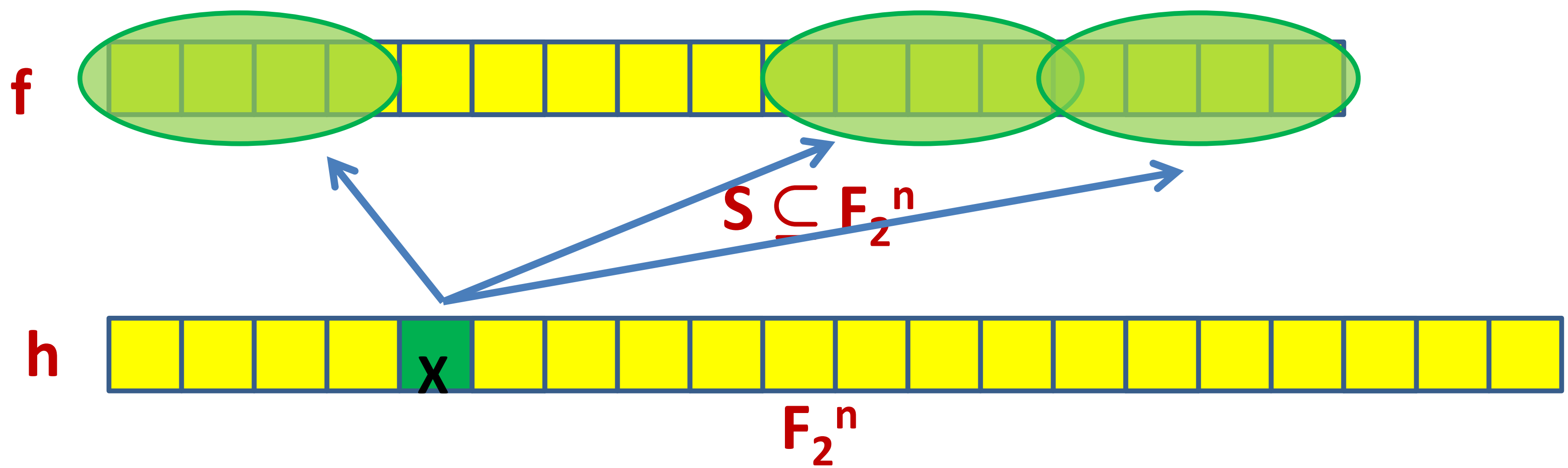
- If $\Delta_\mu(f,g) = (1 - \alpha)/2$, then $\Delta_{Un}(h,g) \approx (1 - \alpha^k)/2$

Getting oracle access to h

Recall: $h(X) = f(X_1) + \cdots + f(X_k)$,

where X_i are sampled independently from $\mu \mid \sum X_i = X$

Given oracle access to f, can simulate oracle access to h.



From f to h

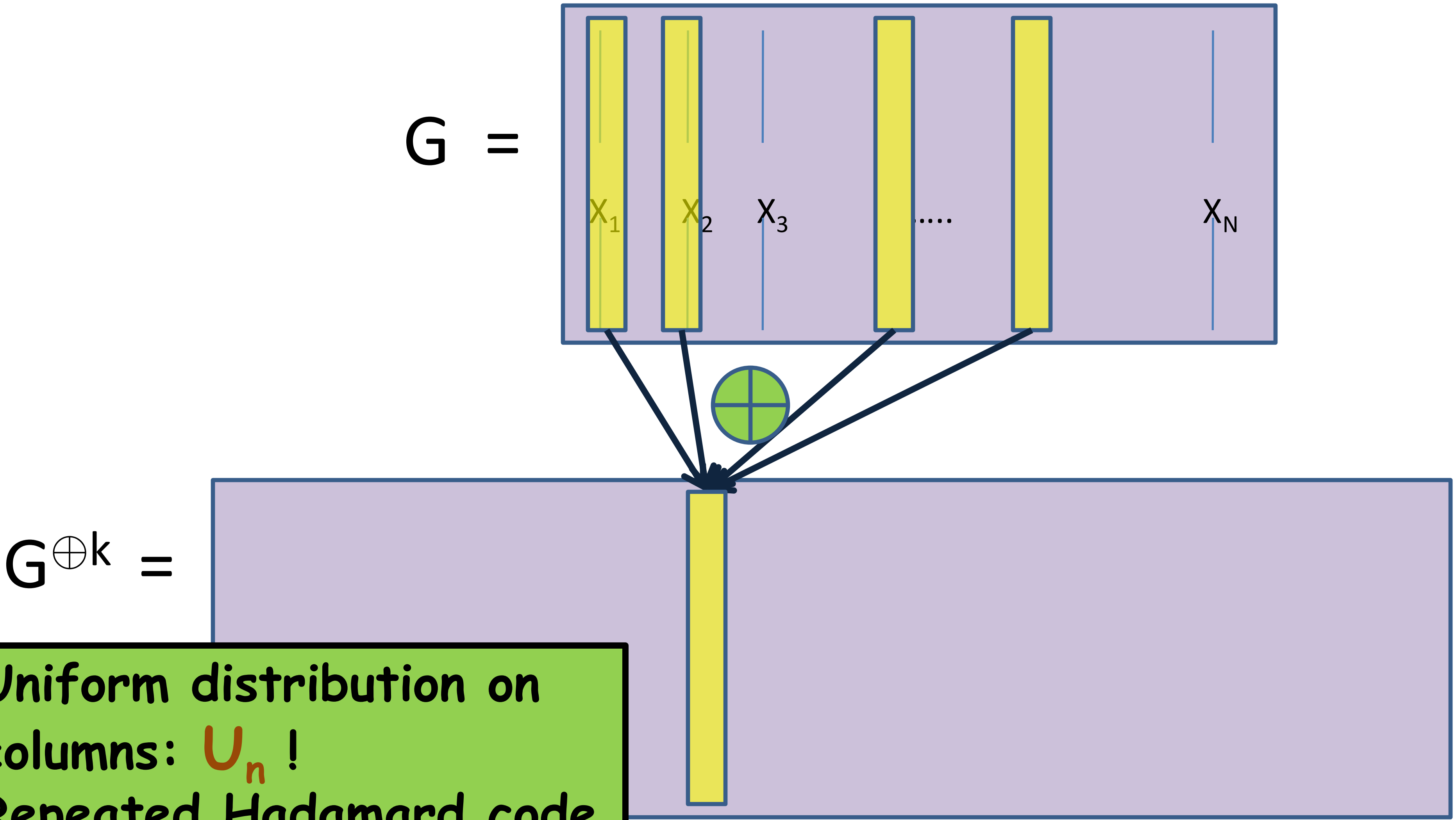
- Given oracle access to **f**, can simulate oracle access to the *extended function* **h.**

- Δ_{U_n}**(h, L)** essentially captures Δ_{μ}**(f, L)**

- We understand testing over $\mathbf{U_n}$ very well.

- **We can *transfer* questions of list decoding, testing, distance estimation over μ to those over $\mathbf{U_{n}}$.**

In the language of codes: XOR



Uniform distribution on columns: U_n !
Repeated Hadamard code

$x_{ijkh} = x_i + x_j + x_k + x_h$

Talk Overview

- Locally testable codes ⇐ Testing linearity under some distribution μ

- Criterion for tolerant testing under μ

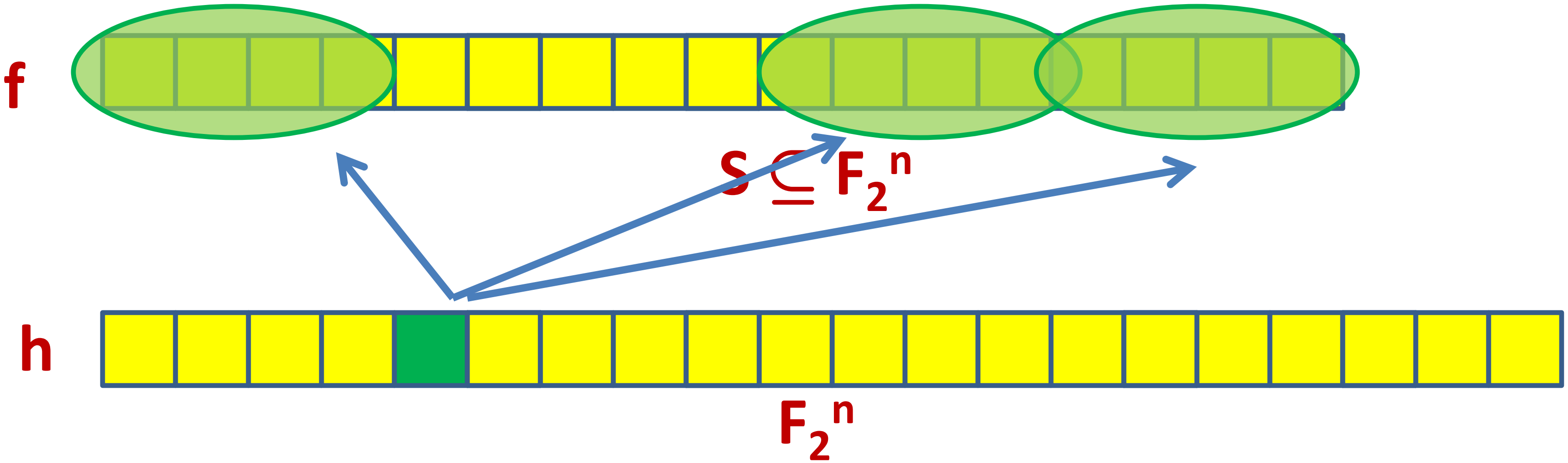
- Local List Decoding and Testing with high error

- Time Complexity
  - Dual BCH codes
  - connections to the noisy parity problem

Time Complexity

Recall: $h(X) = f(X_1) + \cdots + f(X_k)$,

where X_i are sampled independently from $\mu \mid \sum X_i = X$

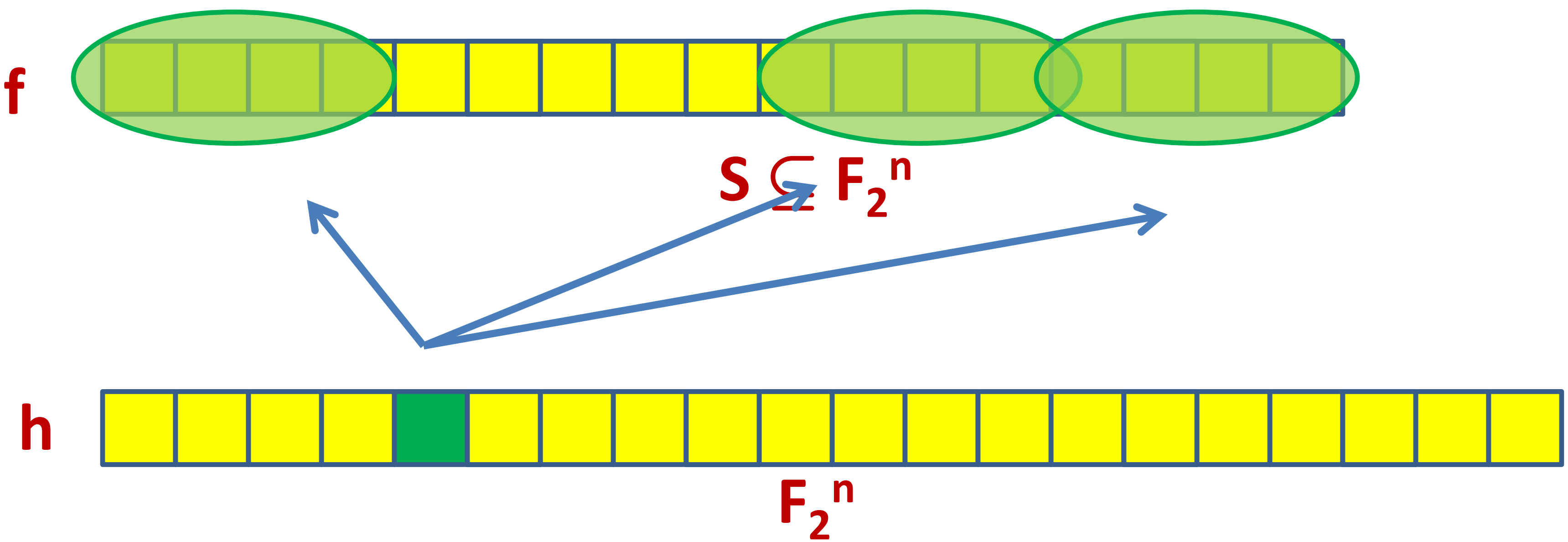


- **Need to *"Back Sample"*.**

  In general (for a random set S) could take time poly(|S|).

Dual-BCH Codes

- The set $\mathbf{S} \subseteq \mathbf{F_2}^n$ is structured.
  - [KL05]: For $X \in F_2^n$, can efficiently compute (in time **polylog (|S|)**) which k-subsets of S sum to X.



Time complexity: decoding a random linear code

Theorem:

If $C \subseteq \{0,1\}^N$ is a linear code of bias = $N^{-\gamma}$ then C is list decodable with $\frac{1}{2}-\epsilon$ fraction errors in time $\exp(n/\log\log n)$

Proof: Reduce to the Hadamard code!

[BKW03, Ly05]: Learning noisy parities:

The Hadamard code can be decoded from random samples from a received word (**a code word corrupted with *random errors***) in time exp(n/log n)

[FGKP06]: Agnostically learning parities:

The Hadamard code can be list decoded from random samples from a received word in time exp(n/log n)

Main features

- Need to take super-constantly many sums of S to get to Hadamard
  - Noise rate gets very high
- **Getting random samples from h is easy** given access to random samples from f.
  - Back sampling not needed.

Thank you!