Invariance in Property Testing

Madhu Sudan
~~MIT~~
MSR

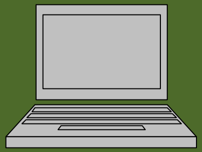
Joint work with Elena Grigorescu & Tali Kaufman.

Data Processing (Prehistoric)

Big computers

Tiny Data

Modern Data Processing



Small computers

Enormous Data

Algorithmic Challenge

- Design Algorithms to process such massive data, when there’s not enough time to read it all!

- Can such algorithms exist?
  - We seem to be using many such heuristics …
  - What guarantees do they provide?

Reasons for optimism

- Statistics:
  - Classical field aimed at studying how to ascertain properties of massive data with random samples.
    - E.g., Polling before elections ...
- Computer Science (Property Testing):
  - 1990 onwards.
  - Algorithms to check data for linearity, multilinearity, low-degree, regularity, uniformity, 3-colorability ...
- (Qualitatively ... what is CS doing that is different from Statistics?)

Property Testing

- Goal: "Efficiently" determine if some "data" "essentially" satisfies some given "property".
- Formalism:
  - Data: $f : D \to R$ given as oracle

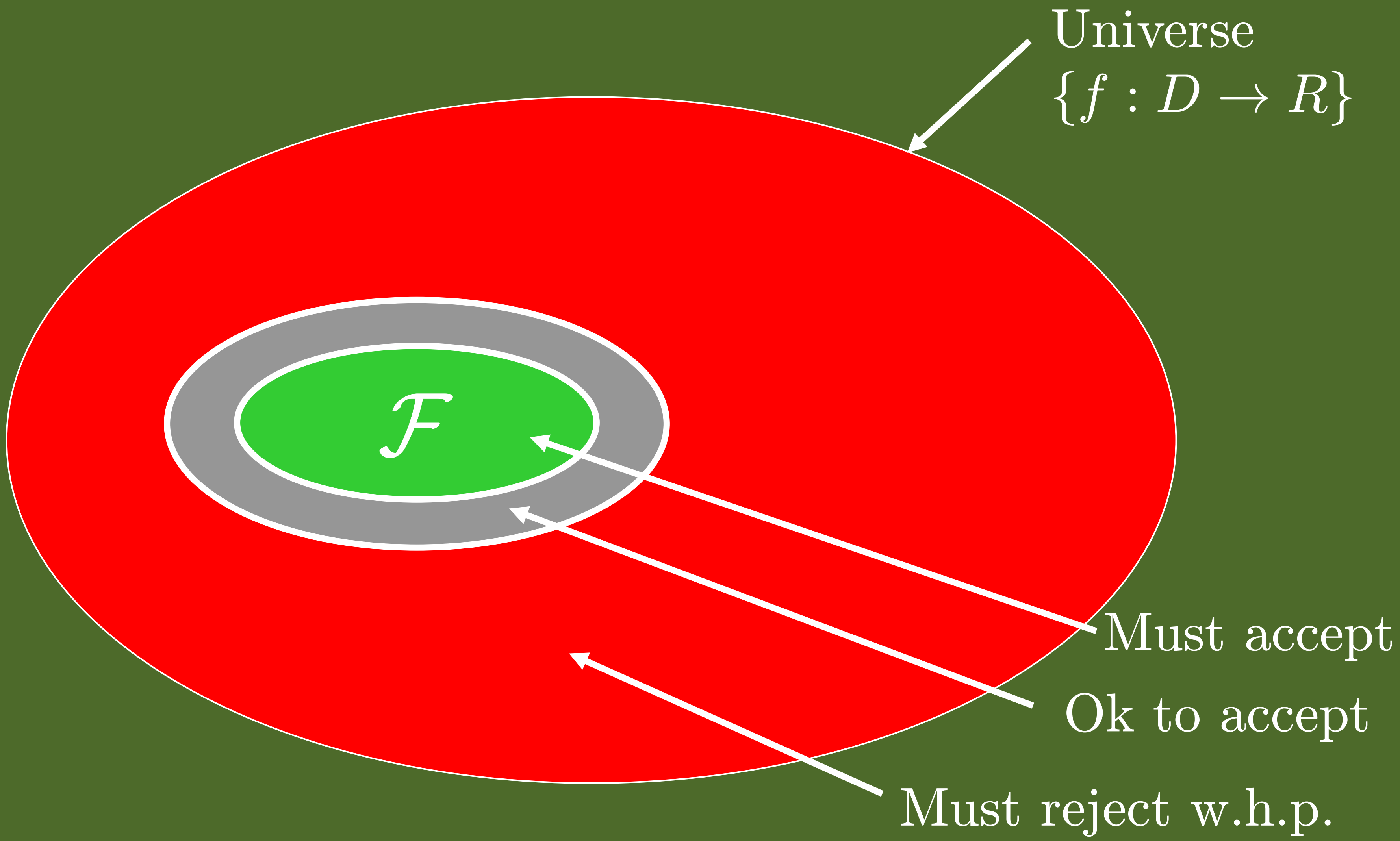
    D finite, but huge. R finite, possibly small
  - Property: Given by $\mathcal{F} \subseteq \{f : D \to R\}$
  - Efficiently: $o(D)$ queries into f. Even $O(1)$!
  - Essentially: Must accept if $f \in \mathcal{F}$

    Ok to accept if $f \approx g \in \mathcal{F}$.

Property Testing (in excruciating detail)

- **Distance:** $\delta(f,g) = \Pr_{x\in D}[f(x) \neq g(x)]$
  $\delta(f,\mathcal{F}) = \min_{g\in\mathcal{F}}\{\delta(f,g)\}$
  $f \approx_\epsilon g$ if $\delta(f,g) \leq \epsilon$.
- **Definition:**
  $\mathcal{F}$ is (q,α)-locally testable if
  $\exists$ a q-query tester that
  accepts $f \in \mathcal{F}$ with probability $1-\epsilon$
  rejects $f \notin \mathcal{F}$ with probability $\geq \alpha \cdot \delta(f,\mathcal{F})$.
- **Notes:** q-locally testable implies $\exists \alpha > 0$
  locally testable implies $\exists q = O(1)$

Property Testing (Pictorially)



History of Property Testing

- Statistics = Prehistory
- First “modern” Property Test: Linearity Test [Blum, Luby, Rubinfeld ’90].
- Formal Definition: [Rubinfeld & S. ’93-’96].
- Systematic study: [Goldreich, Goldwasser, Ron ’96].
- 1990-2009: Many non-trivial tests.

Modern Day Example: Testing Linearity

- Domain = Vector space $\mathbb{F}_2^n$
  Range = Field $\mathbb{F}_2$

- Property: $\mathcal{F}$ = linear functions
  i.e., $\{f(x) = \sum_{i=1}^n a_i x_i | a_i \in \mathbb{F}_2\}$

- Theorem [Blum,Luby,Rubinfeld '90]:

  Linearity is 3-query testable.

- Test: Pick $x, y \in \mathbb{F}_2^n$ uniformly.
  Accept iff $f(x) + f(y) = f(x+y)$

Major classes of problems

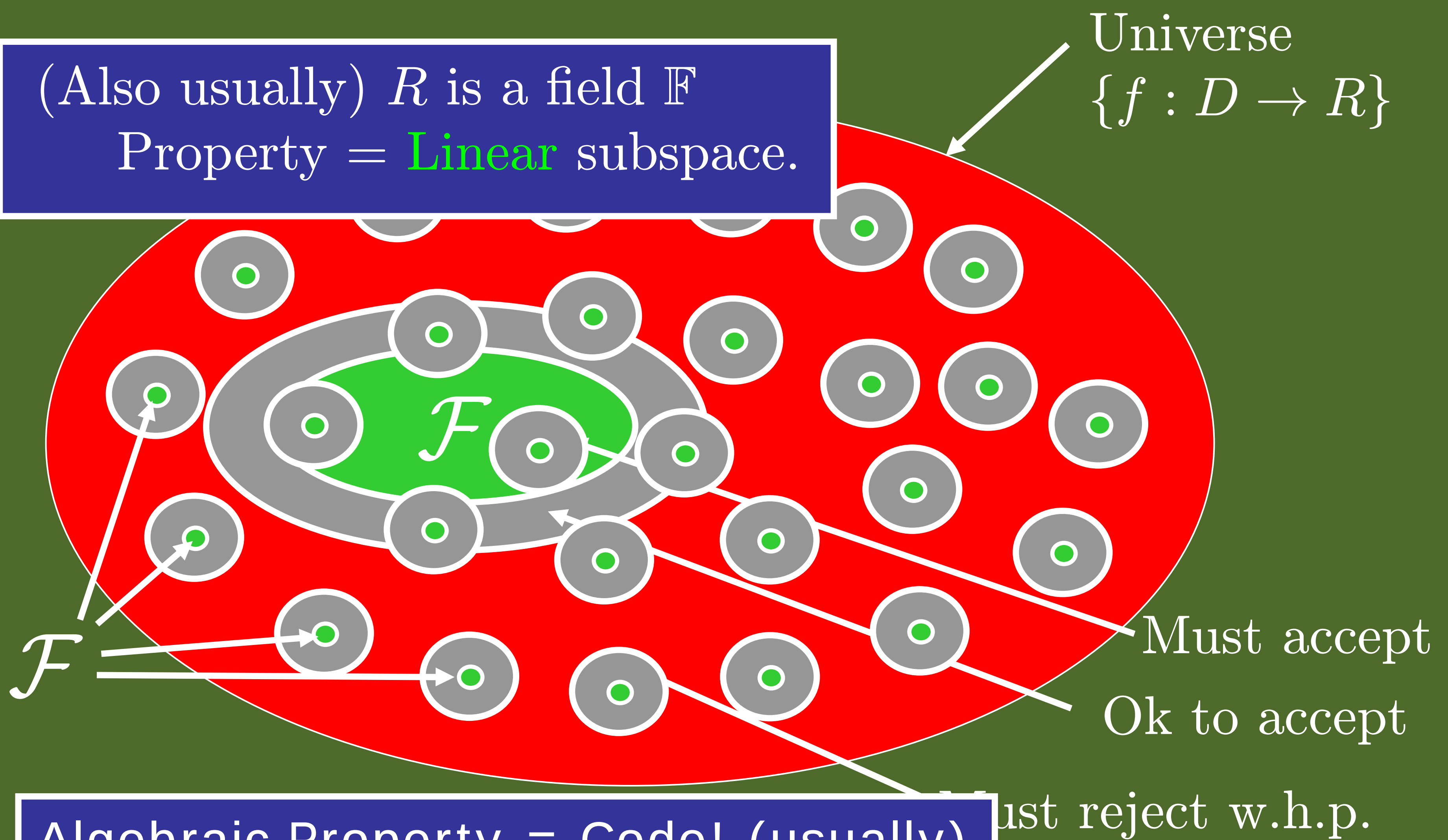
- Graph Property Testing:
  - The web-surfer's problem
    - Does the web graph have small diameter?
    - Is it expanding?
    - Is it bipartite (essentially)?
- Statistical Property Testing:
  - The gambler's problem
    - Are the dice unbiased?
    - Is there a difference between two slot machines?
- Algebraic Property Testing:
  - Kepler's problem
    - Is all this data I am seeing fitting some polynomial?

Main Results

- … [Alon, Shapira], [Alon, Fisher, Newman, Shapira], [Borgs, Chayes, Lovasz, Sos, Szegedy,Vesztergombi]:
  - Monotone graph properties are testable.
  - “Regular” graph properties $\Leftrightarrow$ testable.
- [P. Valiant]
  - Symmetric Statistical Properties $\Leftrightarrow$ testable.
- [BLR,BFL,BFLS,GLRSW,RS,AKKLR,KR,JPRZ]:
  - Laundry list of algebraic properties testable.

Algebraic vs. Combinatorial / Statistical P.T.

(Also usually) R is a field $\mathbb{F}$
Property = Linear subspace.



Algebraic Property = Code! (usually)

Main Results

- … [Alon, Shapira], [Alon, Fisher, Newman, Shapira], [Borgs, Chayes, Lovasz, Sos, Szegedy,Vesztergombi]:
  - Monotone graph properties are testable.
  - "Regular" graph properties $\Leftrightarrow$ testable.
- [P. Valiant]
  - Symmetric Statistical Properties $\Leftrightarrow$ testable.
- [BLR,BFL,BFLS,GLRSW,RS,AKKLR,KR,JPRZ]:
  - Laundry list of algebraic properties testable.

Why are properties locally testable?

- Answer 1: They are not global properties!
  - What does Warren Buffett think?

- Answer 2: They are not (very) sensitive to individual names
  - What does Joe the plumber think?
    - Even if he's not Joe, or plumber,

- To formalize Answer 2: Study “Invariances” of properties.

Invariance & Property testing

- Recall: Property $= \mathcal{F} \subseteq \{D \to R\}$

- Invariances (Automorphism groups):

  For permutation $\pi : D \to D$, $\mathcal{F}$ is π-invariant if
  $f \in \mathcal{F}$ implies $f \circ \pi \in \mathcal{F}$.

  $\mathrm{Aut}(\mathcal{F}) = \{\pi \mid \mathcal{F} \text{ is } \pi\text{-invariant}\}$
  Forms group under composition.

- Hope: If Automorphism group is "large" (or "nice"), then property is testable at least iff some well-studied parameter is small.

Examples

- Majority: (Pre-election polling)
  - Aut group $= S_D$ (full group).
  - Easy Fact: If $\mathrm{Aut}(\mathcal{F}) = S_D$ then $\mathcal{F}$ is $\mathrm{poly}(R, 1/\epsilon)$-locally testable.
- Graph Properties:
  - Aut. group given by renaming of vertices
  - [AFNS, Borgs et al.] implies *regular* properties with this Aut group are testable.
- Statistical Properties: Closed under every permutation of domain and range.
- Algebraic Properties: What symmetries do they have?

Motivating example

- Multivariate polynomials over finite fields:
  - Kepler … (mod p)

$\mathbb{F} = \mathbb{F}_p =$ finite field with p elements.

$\mathcal{F} = \mathcal{F}_{n,d,p} = \{n\text{-variate poly of (total) degree} \leq d\}$

- Example:

$$f(x,y,z) = 3xyz + 2x^2 - 5xz^2$$

Polynomial of degree 3

Theorem [RS 96]: Deg. d poly $\Rightarrow$ $d+2$-query testable.

if $d \ll p$

Invariances of low-deg. polynomials

- Invariant under affine transformations:
- Example:

$f(x, y, z)$ is a deg. d poly
$\Rightarrow f(3x + 2y + z, 2z + 1, 3x - y + 2)$ is also a deg d poly

- So we consider affine-invariant families

$A : \mathbb{F}^n \to \mathbb{F}^n$ affine if $A(\vec{x}) = M \cdot \vec{x} + \vec{b}$

$\mathcal{F}$ affine-invariant if $\forall f \in \mathcal{F}, A$ affine, $f \circ A \in \mathcal{F}$

Our class

- Affine-invariant Property $\mathcal{F}$

- Additionally, Linear:

$$f, g \in \mathcal{F};\ \alpha \in \mathbb{F} \Rightarrow \alpha f, f + g \in \mathcal{F}$$

Why? Because there's light there ...

- Additionally, Locally Constrained:

$$\exists x_1, \ldots, x_k \in \mathbb{F}^n; V \subsetneq \mathbb{F}^k \text{ s.t.}$$
$$\forall f \in \mathcal{F}\ \ f(x_1) \cdots f(x_k) \in V$$

Why? Because its necessary ...

Examples:

- Affine functions:

$$\mathcal{F} = \{a_0 + \sum_{i=1}^{n} a_i x_i | a_0, \ldots, a_n \in \mathbb{F}\}$$

  - Affine-invariant!
  - Linear!
  - Local Constraint:

$$x_1 = a, x_2 = b, x_3 = c; x_4 = a + b + c$$

$$V = \{(\alpha, \beta, \gamma, \alpha + \beta + \gamma) | \alpha, \beta, \gamma \in \mathbb{F}\}$$

Our Results

- Theorem: $\mathcal{F} \subseteq \{\mathbb{F}^n \to \mathbb{F}\}$ linear, affine-invariant,
  with k-local constraint
  implies $\mathcal{F}$ is $f(\mathbb{F}, k)$-query testable.

- Other stuff:
  - Extension to Linear-invariant properties (*)
  - Extension when Domain-field extends range.
  - Study Linear-invariant Properties.
  - Counterexample to AKKLR conjecture.

Implications

- Unifies most previous results on Algebraic Property Testing.
- Simpler, combined proof (than recent papers).
- Many new properties: E.g.,
  - Homogenous polynomials
  - Polynomials supported on degree {2,3,5} …
  - Some v. high-degree polynomials
- Counterexample to
  - Conjecture [Alon, Kaufman, Krivelevich, Litsyn, Ron] : Linear code with k-local constraint and 2-transitive group of symmetries must be testable.

Local Testing

Key Notion: Single Orbit Property

– $\mathcal{F}$ has single orbit property if
  $\exists$ a *single* constraint $C = (\langle x_1, \ldots, x_k \rangle, V)$ such that $\{C \circ \pi\}_{\pi \in \mathrm{Aut}(\mathcal{F})}$ characterize $\mathcal{F}$.

– Single orbit property applies to all known algebraic properties, possibly with the exception of BCH codes.

Theorem: Every linear invariant $\mathcal{F}$ with a k-local characterization, has the single orbit property under some $f(k, \mathbb{K})$-local constraint

Theorem: If $\mathcal{F}$ has single orbit property with a k-local constraint (with some restrictions) then it is k-locally testable.

BLR (and our) analysis

The tests

- BLR: Pick $x, y \in_R \mathbb{F}^n$ and check
$$f(x) + f(y) = f(x+y)$$

  Need to show:
  $$\exists g \text{ s.t. } \delta(f,g) \leq C \cdot \Pr_{x,y}[f(x) + f(y) \neq f(x+y)]$$

- Ours: $\mathcal{F}$ given by $x_1, \ldots, x_k; V$

  Pick linear/affine $L : \mathbb{K}^n \to \mathbb{K}^n$ at random
  Verify $\langle f(L(x_1)), \ldots, f(L(x_k)) \rangle \in V$

  Need to show $\exists g \in \mathcal{F}$ s.t.
  $$\delta(f,g) \leq C \cdot \Pr_L[\langle f(L(x_1)), \ldots, f(L(x_k)) \rangle \notin V$$

BLR Analysis: Outline

- Have f s.t. $\Pr_{x,y}[f(x) + f(y) \neq f(x+y)] = \delta < 1/20$.
  Want to show f close to some $g \in \mathcal{F}$.
- Define $g(x) = \text{most likely}_y\{f(x+y) - f(y)\}$.
- If f close to $\mathcal{F}$ then g will be in $\mathcal{F}$ and close to f.
- But if f not close? g may not even be uniquely defined!
- Steps:
  - Step 0: Prove f close to g
  - Step 1: Prove *most likely* is overwhelming majority.
  - Step 2: Prove that g is in $\mathcal{F}$.

BLR Analysis: Step 0

- Define $g(x) = \text{most likely }_y \{f(x+y) - f(y)\}$.

Claim: $\Pr_x[f(x) \neq g(x)] \leq 2\delta$

- Let $B = \{x | \Pr_y[f(x) \neq f(x+y) - f(y)] \geq \frac{1}{2}\}$
- $\Pr_{x,y}[\text{linearity test rejects } | x \in B] \geq \frac{1}{2}$

$$\Rightarrow \Pr_x[x \in B] \leq 2\delta$$

- If $x \notin B$ then $f(x) = g(x)$

$\text{Vote}_x(y)$

BLR Analysis: Step 1

- Define $g(x) =$ most likely $_y\{f(x+y) - f(y)\}$.
- Suppose for some x, $\exists$ two equally likely values. Presumably, only one leads to linear x, so which one?
- If we wish to show g linear, then need to rule out this case.

Lemma: $\forall\, x$, $\Pr_{y,z}[\text{Vote}_x(y) \neq \text{Vote}_x(z))] \leq 4\delta$

$\text{Vote}_x(y)$

BLR Analysis: Step 1

- Define $g(x) =$ most likely $_y\{f(x+y) - f(y)\}$.
- Suppose for some x, $\exists$ two equally likely values.
  Presumably, only one leads to linear x, so which one?
- If we wish to show g linear,
  then need to rule out this case.

Lemma: $\forall\, x$, $\Pr_{y,z}[\text{Vote}_x(y) \neq \text{Vote}_x(z))] \leq 4\delta$

BLR Analysis: Step 1

$\text{Vote}_x(y)$

- Define $g(x) = \text{most likely}_y\{f(x+y) - f(y)\}$.

Lemma: $\forall\, x,\ \Pr_{y,z}[\text{Vote}_x(y) \neq \text{Vote}_x(z))] \leq 2\delta$

?	$-f(y)$	$f(x+y)$
$-f(z)$	0	$f(z)$
$f(x+z)$	$f(y)$	$-f(x+y+z)$

Prob. Row/column sum non-zero $\leq \delta$.

BLR Analysis: Step 1

$\text{Vote}_x(y)$

- Define $g(x) = \text{most likely}_y\{f(x+y) - f(y)\}$.

Lemma: $\forall\, x,\ \Pr_{y,z}[\text{Vote}_x(y) \neq \text{Vote}_x(z))] \leq 2\delta$

?	$-f(y)$	$f(x+y)$
$-f(z)$	0	$f(z)$
$f(x+z)$	$f(y)$	$-f(x+y+z)$

Prob. Row/column sum non-zero $\leq \delta$.

BLR Analysis: Step 2 (Similar)

Lemma: If $\delta < \frac{1}{20}$, then $\forall\, x, y,\ g(x) + g(y) = g(x+y)$

$g(x)$	$g(y)$	$-g(x+y)$
$f(z)$	$f(y+z)$	$-f(y+2z)$
$-f(x+z)$	$-f(2y+z)$	$f(x+2y+2z)$

Prob. Row/column sum non-zero $\leq 4\delta$.

Our Analysis: Outline

- f s.t. $\Pr_L[\langle f(L(x_1), \ldots, f(L(x_k))\rangle \notin V] = \delta \ll 1.$
- Define $g(x) = \alpha$ that maximizes $\Pr_{\{L|L(x_1)=x\}}[\langle \alpha, f(L(x_2)), \ldots, f(L(x_k))\rangle \in V]$

- Steps:
  - Step 0: Prove f close to g
  - Step 1: Prove "most likely" is overwhelming majority.
  - Step 2: Prove that g is in $\mathcal{F}$.

Our Analysis: Outline

- f s.t. $\Pr_L[\langle f(L(x_1),\ldots,f(L(x_k))\rangle \notin V] = \delta \ll 1.$
- Define $g(x) = \alpha$ that maximizes
  $\Pr_{\{L|L(x_1)=x\}}[\langle \alpha, f(L(x_2)),\ldots,f(L(x_k))\rangle \in V]$

Same as before

- Steps:
  - Step 0: Prove f close to g
  - Step 1: Prove "most likely" is overwhelming majority.
  - Step 2: Prove that g is in $\mathcal{F}$.

$\text{Vote}_x(L)$

Matrix Magic?

- Define $g(x) = \alpha$ that maximizes
  $$\Pr_{\{L \mid L(x_1)=x\}}[\langle \alpha, f(L(x_2)), \ldots, f(L(x_k))\rangle \in V]$$

Lemma: $\forall\, x,\ \Pr_{L,K}[\text{Vote}_x(L) \neq \text{Vote}_x(K))] \leq 2(k-1)\delta$

x	$L(x_2)$	$\cdots$	$L(x_k)$
$K(x_2)$			
$\vdots$		?	
$K(x_k)$			

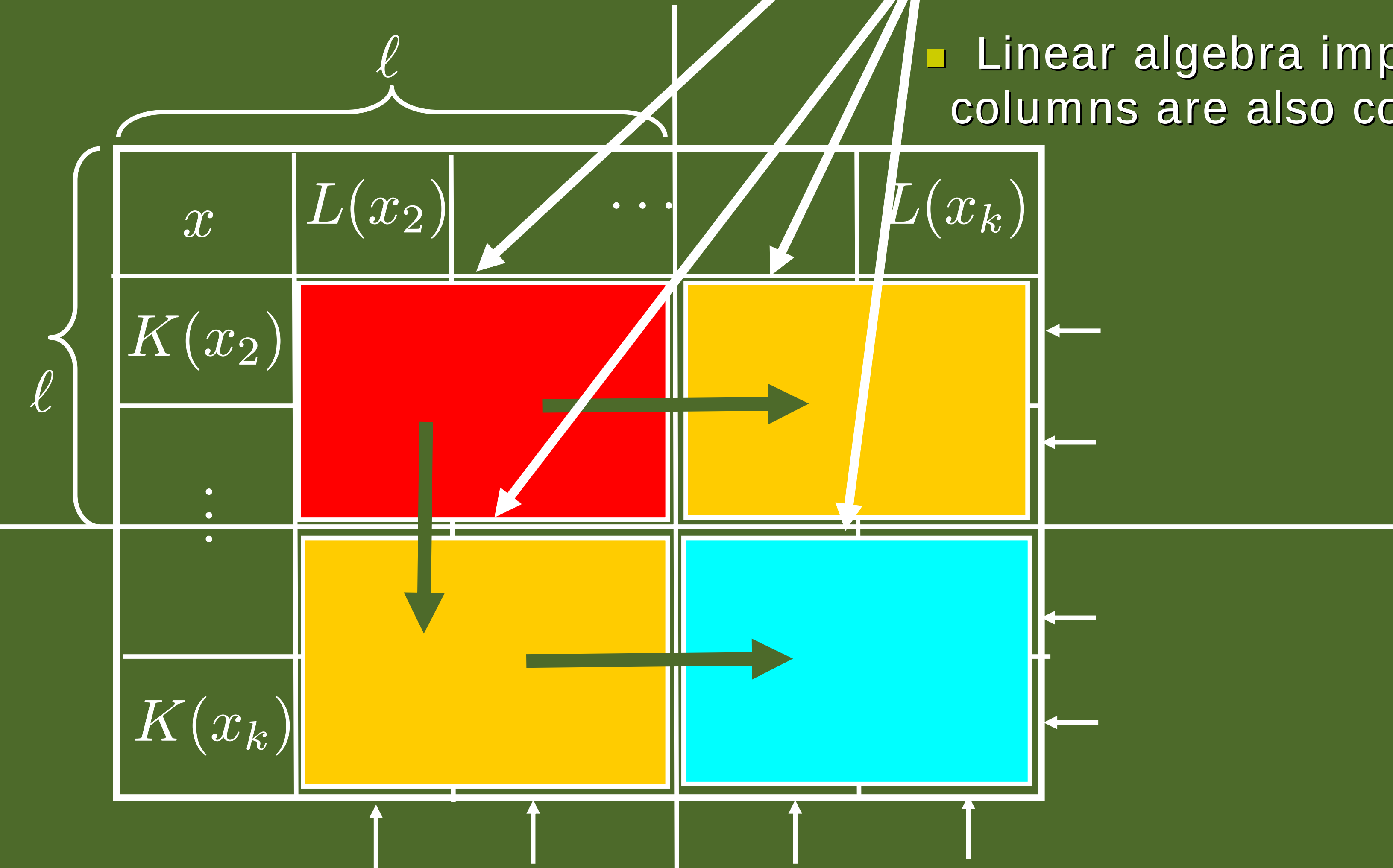
Matrix Magic?

x	$L(x_2)$	$\cdots$	$L(x_k)$
$K(x_2)$			
$\vdots$		?	
$K(x_k)$			

- Want marked rows to be random constraints.
- Suppose $x_1, \ldots, x_\ell$ linearly independent; and rest dependent on them.

Matrix Magic?

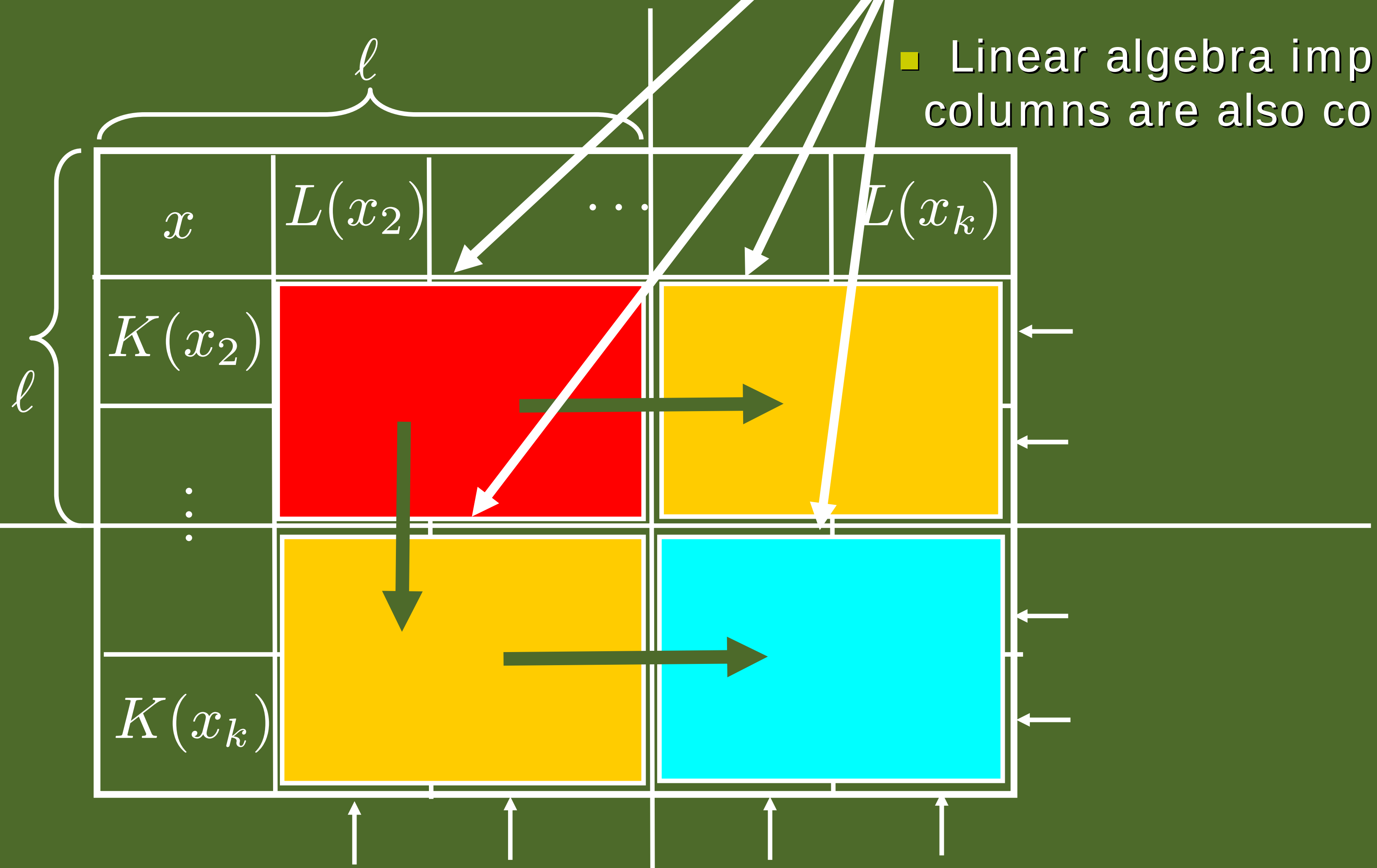
- Fill with random entries
- Fill so as to form constraints
- Linear algebra implies final columns are also constraints.



- Suppose $x_1, \ldots, x_\ell$ linearly independent; and rest dependent on them.

Matrix Magic?

- Fill with random entries
- Fill so as to form constraints
- Linear algebra implies final columns are also constraints.



- Suppose $x_1, \ldots, x_\ell$ linearly independent; and rest dependent on them.

Conclusions

- Invariance is important in property testing.
- Linear-invariance suffices to explain many algebraic tests (and shows some new ones).
- Future work: What are other invariances that lead to testability (from characterizations)?

Thanks!