Local Algorithms & Error-correction

Madhu Sudan
Microsoft Research
Prelude

- Algorithmic Problems in Coding Theory
- New Paradigm in Algorithms
- The Marriage: Local Error-Detection & Correction
Algorithmic Problems in Coding Theory

- Code: $\Sigma = \text{finite alphabet (e.g., \{0,1\}, \{A \ldots Z\})}$
  - $E: \Sigma^k \rightarrow \Sigma^n; \text{Image}(E) = C \subseteq \Sigma^n$
  - $R(C) = k/n; \delta(C) = \text{normalized Hamming distance}$

- Encoding:
  - Fix code $C$ and associated $E$.
  - Given $m \in \Sigma^k$, compute $E(m)$.

- Error-detection ($\epsilon$-Testing):
  - Given $x \in \Sigma^n$, decide if $\exists m \text{ s.t. } x = E(m)$.
  - Given $x$, decide if $\exists m \text{ s.t. } \delta(x,E(m)) \leq \epsilon$.

- Error-correction (Decoding):
  - Given $x \in \Sigma^n$, compute (all) $m \text{ s.t. } \delta(x,E(m)) \leq \epsilon$ (if any exist).
Sublinear time algorithmics

- Given $f : \{0,1\}^k \rightarrow \{0,1\}^n$ can $f$ be “computed” in $o(k,n)$ time?

- **Answer 1:** Clearly NO, since that is the time it takes to even read the input/write the output.
  1. Present input implicitly (by an oracle).
  2. Represent output implicitly
  3. Compute function on approximation to input.

Extends to computing relations as well.
Sub-linear time algorithms

- Initiated in late eighties in context of
  - Program checking [BlumKannan, BlumLubyRubinfeld]
  - Interactive Proofs/PCPs [BabaiFortnowLund]
- Now successful in many more contexts
  - Property testing/Graph-theoretic algorithms
  - Sorting/Searching
  - Statistics/Entropy computations
  - (High-dim.) Computational geometry
- Many initial results are coding-theoretic!
Sub-linear time algorithms & Coding

- Encoding: Not reasonable to expect in sub-linear time.

  - In fact many initial results do so!

- Codes that admit efficient ...
  - ... testing: Locally Testable Codes (LTCs)
  - ... decoding: Locally Decodable Codes (LDCs).
Rest of this talk

- Definitions of LDCs and LTCs
- Quick description of known results
- The first result: Hadamard codes
- Some basic constructions
- Recent constructions of LDCs.
  - [Kopparty-Saraf-Yekhanin '11]
  - [Yekhanin '07, Raghavendra '08, Efremenko '09]
Definitions
Locally Decodable Code

$C: \Sigma^k \rightarrow \Sigma^n$ is $(q, \epsilon)$-Locally Decodable if there exists a decoder $D$ such that given $i \in [k]$, and oracle $w : [n] \rightarrow \Sigma$ such that there exists $m$ such that $\delta(w, C(m)) \leq \epsilon \leq \delta(C)/2$, $D(i)$ outputs $m_i$. $D(i)$ reads $q(n)$ random positions of $w$ and outputs $m_i$ w.p. $\geq 2/3$.

What if $\epsilon > \delta(C)/2$? Might need to report a list of codewords.
Locally List-Decodable Code

C is \((\varepsilon,L)\)-list-decodable if \(\forall w \in \Sigma^n\) # codewords \(c \in C\) s.t. \(\delta(w,c) \leq \varepsilon\) is at most \(L\).

C is \((q,\varepsilon,L)\)-locally-list-decodable if \(\exists\) decoder \(D\) s.t. given oracle \(w: [n] \to \Sigma\), \(\forall m \in \Sigma^k\), s.t. \(\delta(w,C(m)) \leq \varepsilon\), \(\exists j \in [L]\) s.t., \(\forall i \in [k]\), \(D^w(i,j)\) output \(m_i\) w.p. \(2/3\).

\(D(i,j)\) reads \(q(n)\) random positions of \(w\) and outputs \(m_i\) w.p. \(\geq 2/3\).
History of definitions

- Constructions predate formal definitions
  - [Goldreich-Levin ’89].
  - [Beaver-Feigenbaum ’90, Lipton ’91].
  - [Blum-Luby-Rubinfeld ’90].
- Hints at definition (in particular, interpretation in the context of error-correcting codes): [Babai-Fortnow-Levin-Szegedy ’91].
- Formal definitions
  - [S.-Trevisan-Vadhan ’99] (local list-decoding).
  - [Katz-Trevisan ’00]
Locally Testable Codes

C is \((q, \epsilon)\)-Locally Testable if \(\exists\) tester \(T\) s.t.

\[ T \text{ reads } q(n) \text{ positions (probabilistically):} \]

If \(w \in C\), \(T\) accepts w.p. 1.
If \(\delta(w, C) > \epsilon\), \(T\) rejects w.p. \(\geq \frac{1}{2}\).

“Weak” definition: hinted at in [BFLS], explicit in [RS’96, Arora’94, Spielman’94, FS’95].
C is \((q, \epsilon)-(\text{strongly})\) Locally Testable if \(\exists\) tester \(T\) s.t.

- \(T\) reads \(q(n)\) positions (probabilistically):
  - If \(w \in C\), \(T\) accepts w.p. 1.
  - \(\forall w \in \Sigma^n, T\) rejects w.p. \(\geq \Omega(\delta(w,C))\).

“Strong” Definition: [Goldreich-S. ‘02]
Motivations
Local Decoding: Worst-case vs. Average-case

- Suppose $C \subseteq \Sigma^N$ is locally-decodable for $N = 2^n$. (Furthermore assume can locally decode all bits of the codeword, and not just message bits.)
- $c \in C$ can be viewed as $c: \{0,1\}^n \rightarrow \Sigma$.
- Local decoding $\sim \Rightarrow$ can compute $c(x)$, $\forall x$, if can compute $c(x')$ for most $x'$.
- Relates average case complexity to worst-case complexity. [Lipton, STV].
- Alternate interpretation:
  - Can compute $c(x)$ without revealing $x$.
  - Leads to Instance Hiding Schemes [BF], Private Information Retrieval [CGKS].
Motivation for Local-testing

- No generic applications known.
- However,
  - Interesting phenomenon on its own.
  - Intangible connection to Probabilistically Checkable Proofs (PCPs).
  - Potentially good approach to understanding limitations of PCPs (though all resulting work has led to improvements).
Contrast between decoding and testing

- **Decoding**: Property of words near codewords.
- **Testing**: Property of words far from code.

**Decoding:**
- Motivations happy with \( n = \text{quasi-poly}(k) \), and \( q = \text{poly log } n \).
- Lower bounds show \( q = O(1) \) and \( n = \text{nearly-linear}(k) \) impossible.

**Testing:** Better tradeoffs possible! Likely more useful in practice.
- Even conceivable: \( n = O(k) \) with \( q = O(1) \)?
Some LDCs and LTCs
Hadamard (1st Order RM) Codes

- **Messages:**
  - (Coefficients of) Linear functions \( \{ L : F_2^k \rightarrow F_2 \} \).

- **Encoding:**
  - Evaluations of \( L \) on all of \( F_2^k \).

- **Parameters:**
  - \( k \) bit messages \( \rightarrow 2^k \) bit codewords.

- **Locality:**
  - 2-Locally Decodable [Folklore/Exercise]
  - 3-Locally Testable [BlumLubyRubinfeld]
Hadamard (1\textsuperscript{st} Order RM) Codes

- Summary:
  - There exist infinite families of codes
  - With constant locality (for testing and correcting).
Codes via Multivariate Polynomials

- **Message**: Coefficients of degree $t$, $m$-variate polynomial over (finite field) $F$

  - Encoding: Evaluations of $P$ over all of $F^m$
  - Parameters: $k \approx (t/m)^m$; $n = F^m$; $\delta(C) \approx 1 - t/F.$
Basic insight to locality

- m-variate polynomial of degree t, restricted to $m' < m$ dim. affine subspace is poly of deg. t.

**Local Decoding:**
- Given oracle for $w \approx P$, and $x \in F^m$
- Pick subspace $A$ through $x$.
- Query $w$ on $A$ and decode for $P|_A$
- Query complexity: $q = F^{m'}$; Time = poly(q);
  $m' = o(m) \Rightarrow$ sublinear!

**Local Testing:**
- Verify $w$ restricted to subspace is of degree t.
- Same complexity; Analysis much harder.
Polynomial Codes

- Many parameters: \( m, t, F \)

- Many tradeoffs possible:
  - Locality \((\log k)^2\) with \( n = k^4 \);
  - Locality \( \epsilon k \) with \( n = O(k) \);
  - Locality (constant) \( q \), with \( n = \exp(k^{(1/q-1)}) \)
Are Polynomial Codes (Roughly) Best?

- No! [Ambainis97] [GoldreichS.00] ...

- No!! [Beimel,Ishai,Kushilevitz,Raymond]

- Really ... Seriously ... No!!!!
  [Yekhanin07,Raghavendra08,Efremenko09]
  [Kopparty-Saraf-Yekhanin ‘10]
Recent LDCs - I
[Kopparty-Saraf-Yekhanin '10] s
The Concern

- Poor rate of polynomial codes:
  - Best rate (for any non-trivial locality): $\frac{1}{2}$
    (bivariate polynomials, $\sqrt{n}$ locality).
  - Locality $n^\epsilon$: Rate $\epsilon^{(1/\epsilon)}$
    (use $1/\epsilon$ variables).

- Practical codes use high rates (say 80%)
Bivariate Polynomials

- Use \( t = (1 - \rho).F \); \( \rho \to 0 \)
- Yields \( \delta(C) \approx \rho \).
- \# coefficients: \( k < \frac{1}{2}(1 - \rho)^2.F^2 \)
- Encoding length: \( n = F^2 \).
- Rate \( \approx \frac{1}{2}(1 - \rho)^2 \)

- Can’t use degree > F; Hence Rate < \( \frac{1}{2} \)!
Mutiplicity Codes

- Idea:
  - Encode polynomial $P(x,y)$ by its evaluations, and evaluations of its (partial) derivatives!

- Sample parameters:
  - $n = 3F^2$ ($F^2$ evaluations of $\{P + P_x + P_y\}$).
  - However, degree can now be larger than $F$.
  - $t = 2(1 - \rho).F \Rightarrow \delta(C) = \rho$.
  - $k = 2 \cdot (1 - \rho)^2 \cdot F^2$; Rate $\approx 2/3$.
  - Locality = $O(F) = O(\sqrt{k})$

- Getting better:
  - With more multiplicity, rate goes up.
  - With more variables, locality goes down.
Multiplicity Codes: The Theorem

- **Theorem:**
  \[
  \forall \alpha, \beta > 0, \quad \\
  \exists \delta > 0 \text{ and LDC } C: \{0,1\}^k \rightarrow \{0,1\}^n \text{ with } \\
  \text{Rate} \geq 1 - \alpha, \\
  \text{Distance} \geq \delta, \\
  \text{Locality} \leq k^\beta \text{ (decodable with } k^\beta \text{ queries).}
  \]
Recent LDCs - II
[Yekhanin ‘07, Raghavendra ‘08, Efremenko ‘09]
Other end of spectrum

- Minimum locality possible?
  - $q = 2$: Hadamard codes achieve $n = 2^k$;
    - [Kerenidis, deWolf]: $n \geq \exp(k)$.
  - $q = 3$: Best possible $= ?$.
    - Till 2006: Widely held belief: $n \geq \exp(k^{0.1})$
    - [Yekhanin '07]: $n \leq \exp(k^{0.0000001})$
    - [Raghavendra '08]: Clarified above.
    - [Efremenko '09]: $n \leq \exp(\exp(\sqrt{\log k}))$ ...
Essence of the idea:

- Build “good” combinatorial matrix over $\mathbb{Z}_m$ (integers modulo $m$).
- Embed $\mathbb{Z}_m$ in multiplicative subgroup of $F$.
- Get locally decodable code over $F$. 
"Good" Combinatorial matrix

\[ A = \begin{bmatrix}
0 & \ldots & 0 \\
\ldots & 0 & \ldots \\
\ldots & \ldots & 0 \\
\ldots & \ldots & \ldots & 0 \\
\end{bmatrix} \]

- \( k \times n \) matrix over \( \mathbb{Z}_m \)
- Zeroes on diagonal
- Non-zero off-diagonal
- Columns closed under addition
Embedding into a field

- Let $A = [a_{ij}]$ be good over $\mathbb{Z}_m$.

- Let $\omega \in F$ be primitive $m^{th}$ root of unity.

- Let $G = [\omega^{a_{ij}}]$.

- Thm $[Y, R, E]$: $G$ generates an $m$ query LDC over $F$!!!

Highly non-intuitive!
Improvements

- Let $A = [a_{ij}]$ be good; Let $G = [\omega^{a_{ij}}]$.
- Off-diagonal entries of $A$ from $S$ code is $(|S|+1)$-locally decodable. (suffices for [Efremenko]).
- $\omega^S$ roots of $t$-sparse polynomial code is $t$-locally decodable. (critical for [Yekhanin]).
“Good” Matrices?

- [Yekhanin]:
  - Picked m prime.
  - Hand-constructed matrix.
  - Achieved \( n = \exp(k^{1/|S|}) \)
  - Optimal if m prime!
  - Managed to make S large \((10^6)\) with \(t=3\).

- [Efremenko]
  - m composite!
  - Achieves \(|S| = 3\) and \(n = \exp(\exp(\sqrt{\log k}))\)
    (\([\text{Beigel}, \text{Barrington}, \text{Rudich}]; [\text{Grolmusz}]\))
  - Optimal?
Limits to Local Decodability: Katz-Trevisan

- $q$ queries $\Rightarrow n = k^1 + \Omega(1/q)$

- Technique:
  - Recall $D(x)$ computes $C(x)$ whp for all $x$.
  - Can assume (with some modifications) that query pattern uniform for any fixed $x$.
  - Can find many random strings such that their query sets are disjoint.
  - In such case, random subset of $n^{1-1/q}$ coordinates of codeword contain at least one query set, for most $x$.
  - Yields desired bound.
Some general results

- **Sparse, High-Distance Codes:**
  - Are Locally Decodable and Testable
    - [KaufmanLitsyn, KaufmanS]

- **2-transitive codes of small dual-distance:**
  - Are Locally Decodable
    - [Alon, Kaufman, Krivelevich, Litsyn, Ron]

- **Linear-invariant codes of small dual-distance:**
  - Are also Locally Testable
    - [KaufmanS]
Summary

- Local algorithms in error-detection/correction lead to interesting new questions.
- Non-trivial progress so far.
- Limits largely unknown
  - $O(1)$-query LDCs must have $\text{Rate}(C) = 0$
    - [Katz-Trevisan]
Questions

- Can LTC replace RS (on your hard disks)?
  - Lower bounds?
  - Better error models?

- Simple/General near optimal constructions?
- Other applications to mathematics/computation? (PCPs necessary/sufficient)?
- Lower bounds for LDCs?/Better constructions?
Thank You!