Sparsity in Information Theory and Biology

Olgica Milenkovic

ECE Department, UIUC
Joint work and work in progress with W. Dai, P. Hoa, S. Meyn, UIUC

Information Beyond Shannon, December 29, 2008

Sparsity: When only "a few" out of many options are possible...

- Sparsity in information theory:
  - Error-control codes: when only a "few errors" are possible;
  - Superimposed Euclidean and group testing codes: when only "a few" items are biased, "a few" individuals infected, "a few" users active, etc.
  - Digital fingerprinting (CS): when only "a few" colluders align.
  - Signal processing - compressed sensing (CS): when only "a few" coefficients in a linear superposition of real-valued signatures are non-zero.

- Where does sparsity arise: data storage and transmission; wireless communication; signal processing; life sciences; fault tolerant computing.

- Topics of current interest: Sparsity/sparse superpositions in information theory and life sciences.

Sparsity: When only "a few" out of many options are possible...

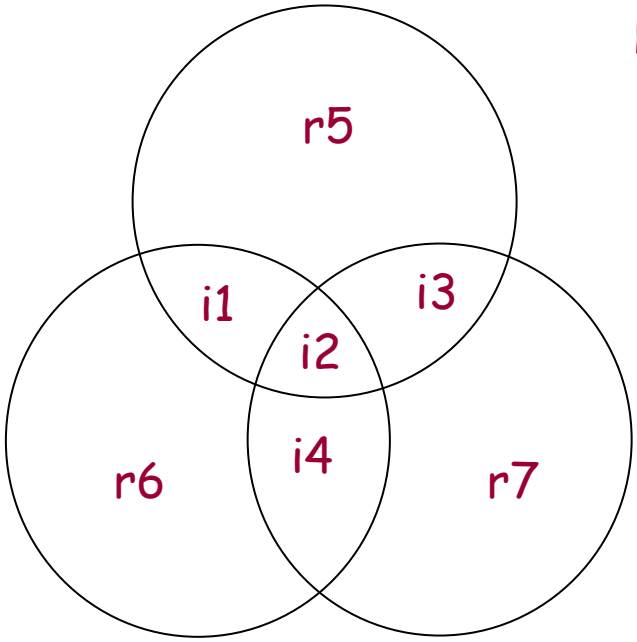
- Sparsity in biology:
  - Observation I: Biological systems evolved in complex environments with almost unlimited number of external stimula (large dimensional signal spaces!).
  - Observation II: Developing individual response mechanisms for each stimulus prohibitively costly.
  - Observation III: Fortunately, only a few signals present at the same time and/or location.
  - Observation IV: Based on group tests, have to determine which signals were present.

- Where does sparsity arise in biology: Neuroscience - group testing in sensory systems, sparse (multidimensional) neural coding, sparse network interactions.

- Where does sparsity arise in biology: Bioinformatics - group testing in immunology, sparse gene/protein network interactions, etc.

Information theory: Error-control coding

Hamming codes: [7,4,3]



r5+i1+i2+i3=0 modulo 2

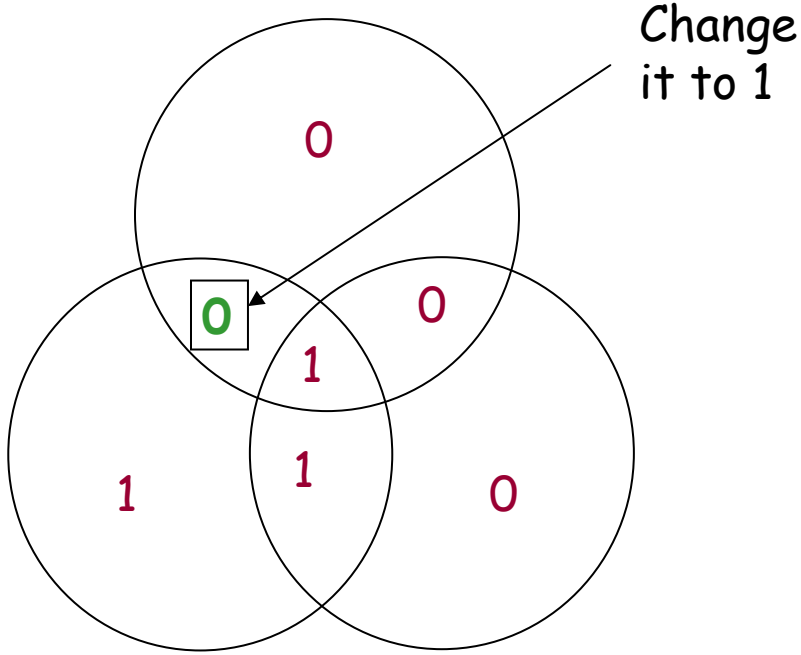
r6+i1+i2+i4=0 modulo 2

r7+i2+i3+i4=0 modulo 2

r5-=i1+i2+i3 modulo 2

r6=i1+i2+i4 modulo 2

r7=i2+i3+i4 modulo 2



$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

message = [i1 i2 i3 i4] G

$$H = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

syndrome = H [i1 i2 i3 i4 r5 r6 r7]

Linear Block Codes (LBCs) over $\mathbb{F}_q$

- **Definition:** A linear binary code $\mathcal{C}$ is a collection of codewords of length n, with k information symbols and $n-k$ parity-check symbols. The code rate is defined as $R = k/n$.

- A set of $m = n - k$ parity-check equations, arranged row-wise, form a parity-check matrix of the code, H. Clearly,

$$\mathbf{x} \in \mathcal{C} \iff \mathbf{Hx} = 0.$$

  The rows represent basis-vectors of the null-space of $\mathcal{C}$.

Error-control Coding and Sparse Superpositions

- **Error-control coding:** The support of e, supp(e), is the set of indices in $[1, \ldots, n]$ for which $e_i \neq 0$. Hence

$$\mathbf{Hy} = \sum_{i \in \mathbf{supp}\,(\mathbf{e})} e_i \, \mathbf{h_i},$$

  where $\mathbf{h}_i$ is the i-th column of H.

- **Error-control coding:** With an abuse of standard coding-theoretic language, refer to the columns of H as **code-words**. Then an r-error correcting code is a set of n codewords $\mathbf{h_i}$, $i = 1, \ldots, n$, with the property that all the $\mathbb{F}_q$-**linear combinations** of collections of not more than r codewords ("a few" $\leq r$) are distinct.

- **Robust error-control coding:** A s-robust, r'-error correcting code is a collection of n codewords $\mathbf{h_i}$, with the property that any two distinct $\mathbb{F}_q$-linear combinations of collections involving not more than r' codewords have Hamming distance at least s.

Information theory: group testing

Given 12 identical-looking coins with one defective coin (heavier, say, than others), identify the defective coin using at most three weighings of a balance.

Test coins in group:

"1" denotes defective – heavier

Test is "positive" if at least one element is defective.

Non-adaptive group testing works only if small number of coins are defective.

Codes over $\mathbb{F}_2$: OR (Group Testing) Codes

- **Generalizations:** A $\mathbb{F}_2$-sum is just the Boolean XOR function. Since we are working with the syndrome, can claim that "superposition=linear function" of columns of H is all we need for decoding. Can we use other functions (superposition strategies) instead?

- One "neglected" example: Kautz and Singleton's (KS) superimposed codes, 1964.
  Motivation: database retrieval (signature files) (KS, 1964), quality control testing (Colbourn et.al., 1996), de-randomization of pattern-matching algorithms (Indyk, 1997).

  Definition: A superimposed design is a set of n codewords of length m, with the property that all bit-wise logical OR functions of collections of not more than r ("a few") codewords are distinct.

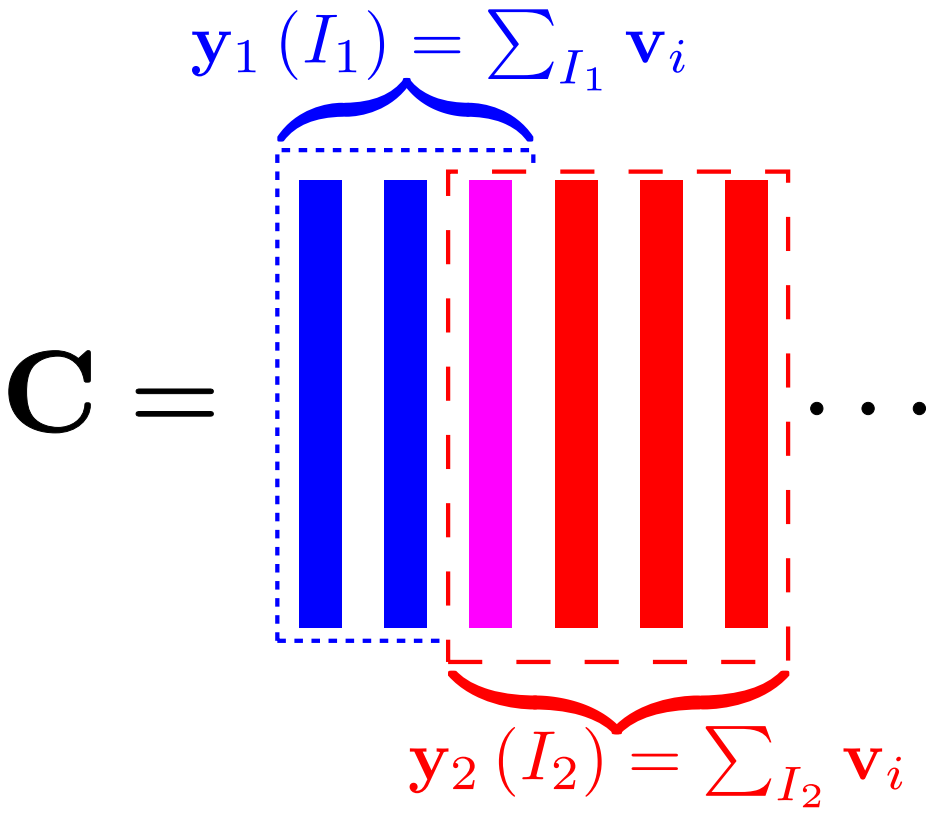
Codes over $\mathbb{F}_2$: Superimposed Coding and Beyond

- **Generalizations:** A robust superimposed code obeys the more restrictive constraint that the distinct OR functions are at Hamming distance at lest s from each other. One may also impose "joint constraints" on the codewords, such as fixed weight of the rows of the superimposed code (design) matrix (Renýi search model, Dyachkov et.al. 1990).

- **Some more recent work:** Use "thresholded" $\mathbb{F}_q$-sums, logical AND and other non-linear tests...

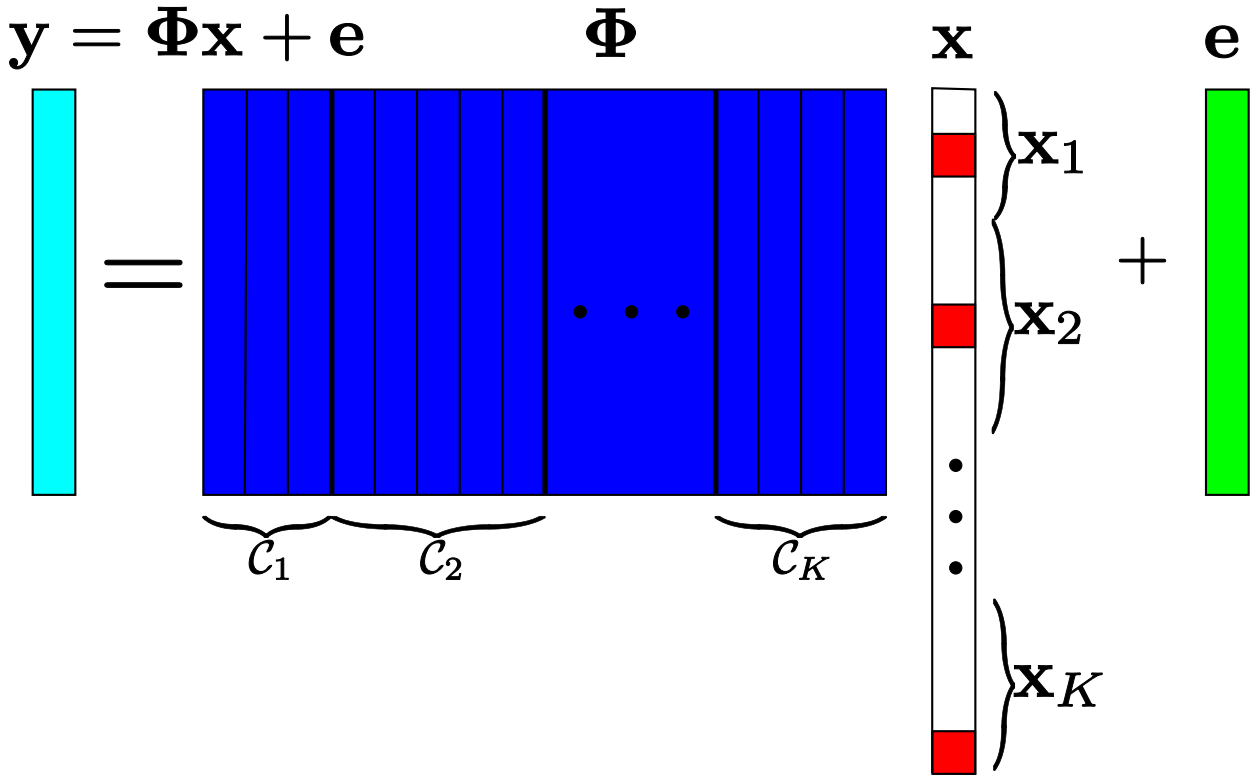
Information theory: multi-access channels

Codes over $\mathbb{R}^n$: Euclidean Superimposed Codes



User $\leftrightarrow$ signature v_i, at most K users active. Norm constraint $\leftrightarrow$ power constraint. Goal is to identify active users.

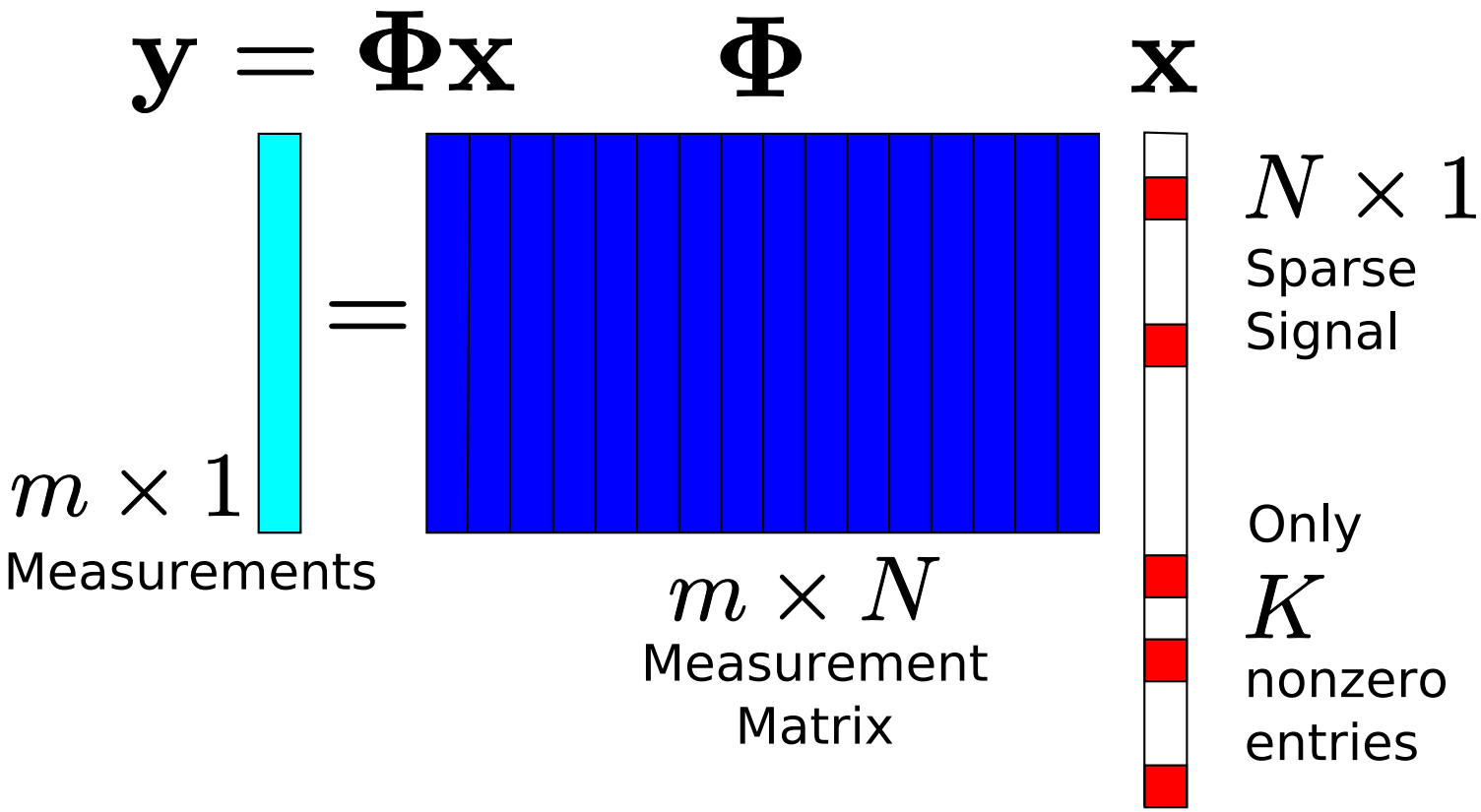
Codes over $\mathbb{R}^n$: Partitioned Euclidean Superimposed Codes



Each user has a codebook of signatures, and at most K users active.

Information theory (?): compressed sensing

Compressed sensing: Codewords over $\mathbb{R}^m$, **weights from** $\mathbb{R}$, **$\mathbb{R}$-linear combinations.** As for superimposed codes, it is assumed that there is a bound on the number of active users/components: $||\mathbf{x}||_0 \leq \mathbf{K}$.



Sparsity as side information: Knowledge about signal being sparse allows for simple, information-preserving dimensionality reductions! In addition, reconstruction algorithms are polynomial time.

CS, Group testing, and sparse superpositions in Biology

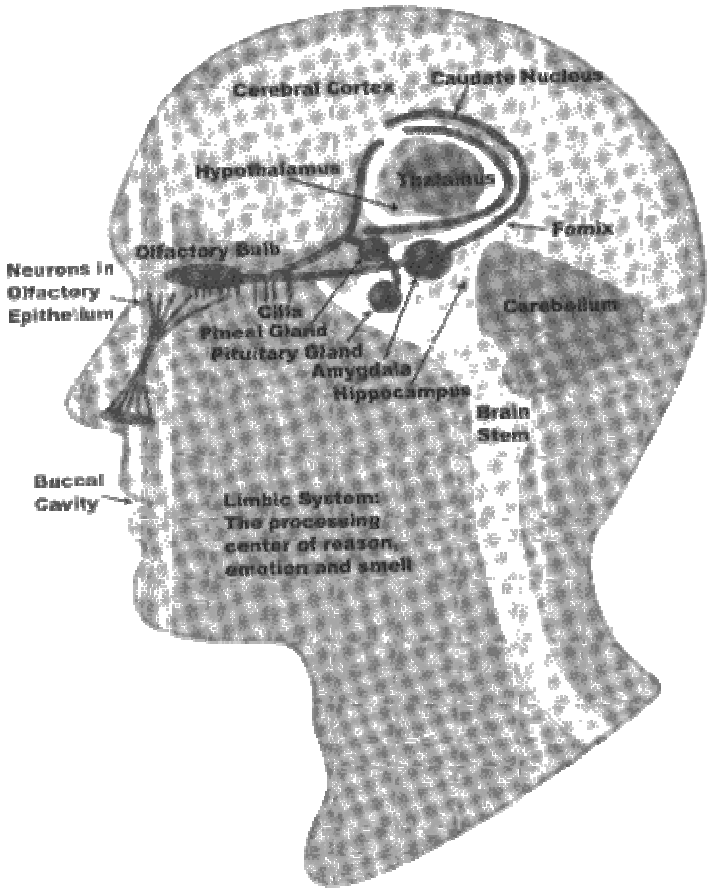
Group testing and CS - Neuroscience (with D. Wilson, Oklahoma University)

The Olfactory System

"Smell system" in mammals responsible for:
1) Creating representation of smells.
2) Determining the concentration of smells.
3) Distinguishing new smells from background.

Buck and Malnic (Nobel Prize, 2004)
Sense of smell uses "combinatorial approach to recognizing and processing odors".

Instead of assigning to each (10 000) smell a unique receptor, the olfactory system uses a small collection (100) of "combinatorial group testing sensors" to create a specific smell response within the neurons of the brain.



http://www.infinitevitalheart.com/Olfactory_System.html

Group testing in the epithelium: Shape-based, one receptor protein (methaloprotein) locks onto several "basic odorants".

Spatial grouping of receptors: Responses of receptors for the same group of odorants (i.e., same type of receptors) converge to the same glomeruli region in the olfactory bulb.

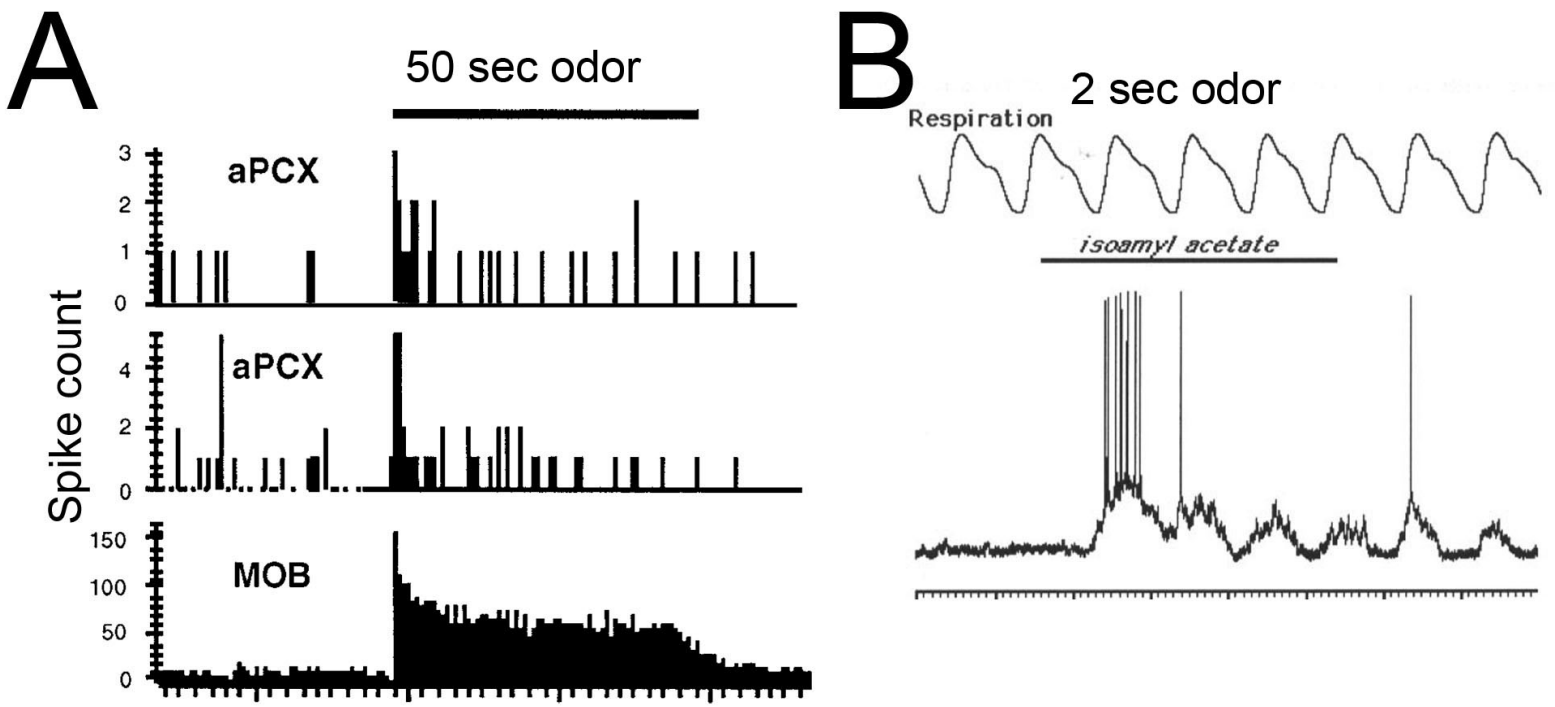
Detection, estimation, and classification is performed in the reduced dimensional space: CS theory - see work by Baraniuk et. al., although sometimes "fanning in - fanning out" effects are possible.

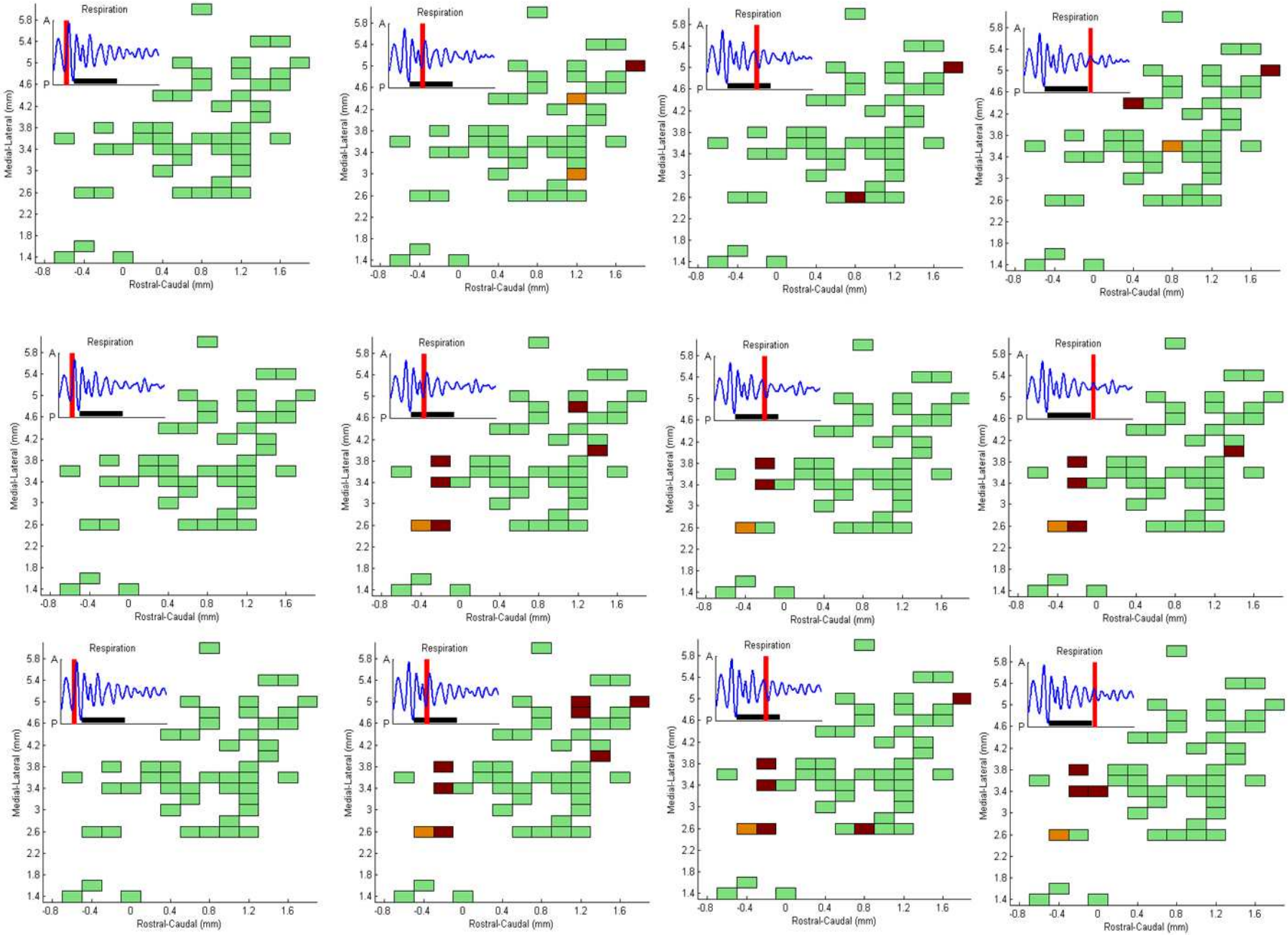
Sparse Spatio-Temporal Coding

- **Sparse spatial coding:** At each point of time, only certain groups of neurons are active ("a few" groups).

- **Sparse and dense temporal coding:** Neuronal spikes are infrequent/frequent in time.

Example: Sparse/dense temporal coding




Respiration
A
P
Medial-Lateral (mm)
Rostral-Caudal (mm)

Sparse Spatio-Temporal Coding

- **Question 1:** What is the exact nature of non-linear superposition mechanisms?

- **Question 2:** How does the type of coding method relate to the function of group of neurons?

- **Question 3:** What kind of processing algorithms (estimation, detection, classification) does the neural system use for non-linear CS data?

Group testing and CS - Bioinformatics (with J. Dingel, A. MacNeil, J. Shisler)

Group testing and CS as part of the immune system response: "Shape-based", one T cell type recognizes many viral epitopes. Competition of immune system cells is regulated in such a way that only a few of the most efficient T cells are produced during equilibrium response. Good from the perspective of energy preservation, big drawback when fighting HIV viruses (original antigenic sin). Also of importance when studying oncolytic viral treatments.

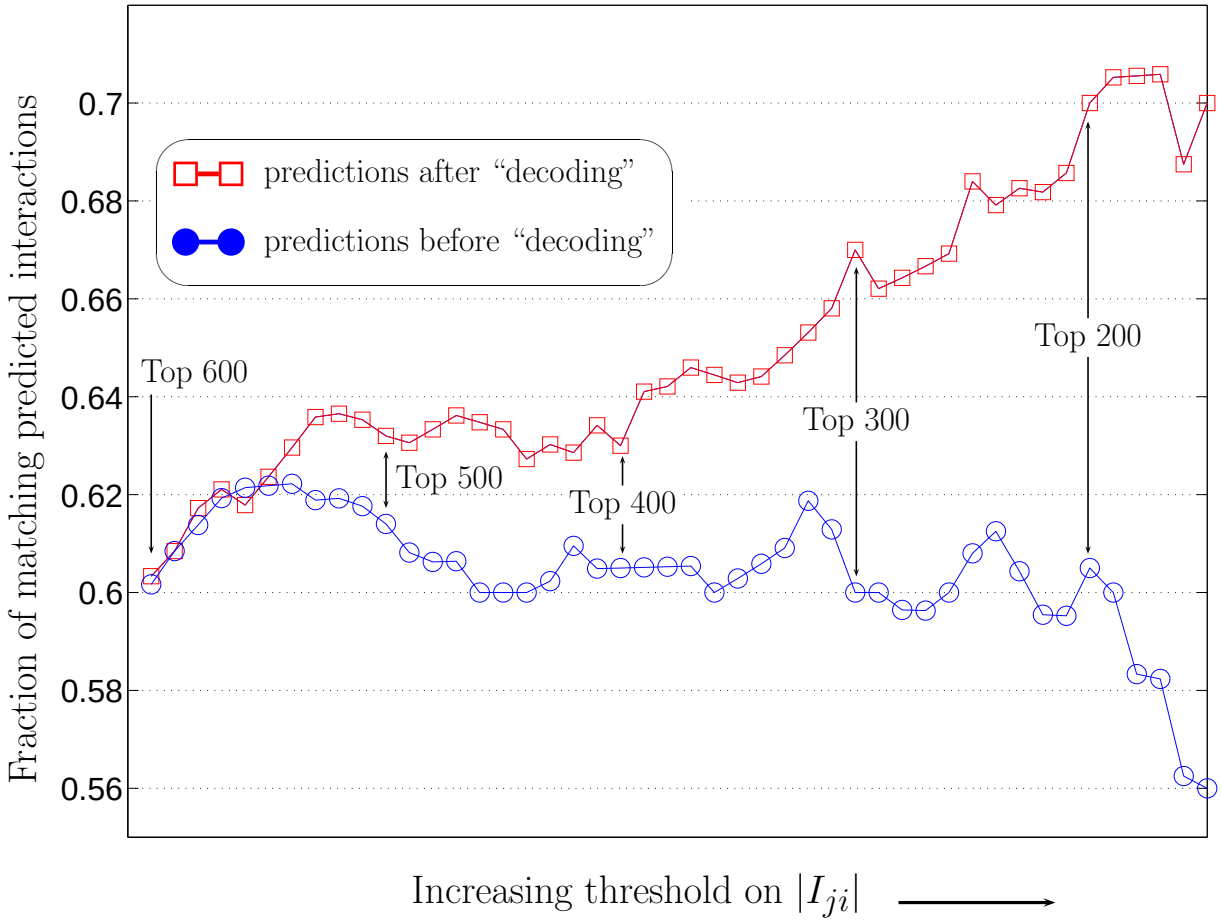
In coding theoretic language, only keep the projections with length exceeding a certain threshold (some form of quantization). How do these projections "preserve information" when the input signal changes?.

Inferring topology/dynamics of sparse gene regulatory networks: E. coli SOS network

- **Except for a few exceptions, most genes are regulated by only a few other genes:** can assume that gene response is a (linear?) superposition of input responses of a few regulatory genes.

- **How do we do this inference efficiently:** coding-theoretic inspired reconstruction algorithms for CS and group testing.

Linear superposition model - improvement in interaction prediction

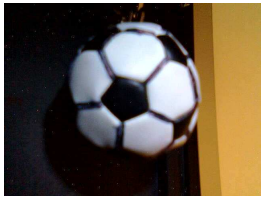


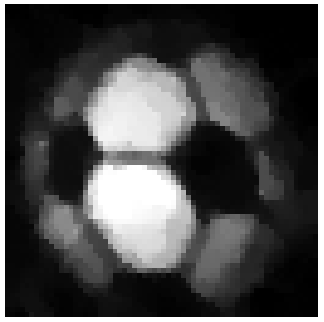
Biologically inspired sensing systems:

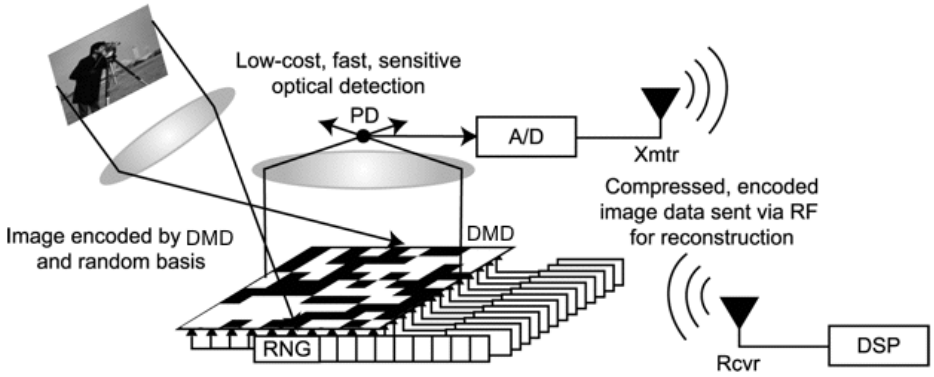
Artificial nose technology by Ken Suslick (UIUC).

CS (group testing) DNA microarrays and aptamer arrays (UIUC).

Single pixel camera (Rice university).



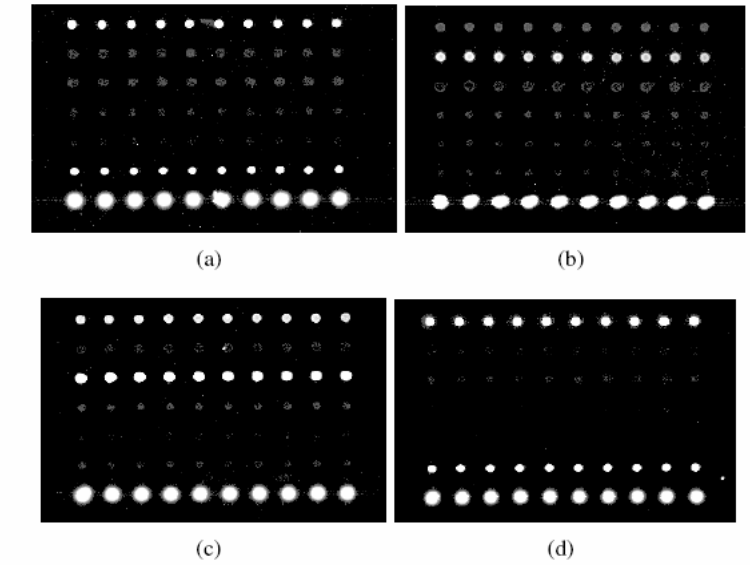




Original Object

4096 Pixels
800 Measurements (20%)

Single Pixel Camera: Baraniuk et.al., Rice University



Compressed Sensing DNA Microarrays: Milenkovic et.al., UIUC

INTERESTING MATHEMATICS? ERROR-CONTROL AND SOURCE CODING, ALGORITHMS,...

CS and Superimposed Coding

Hybrids Between ESC and CS: Constrained and Nonlinear CS

- **Settings allow for handling three important drawbacks of CS strategies:** a) noise intolerance; b) lack of deterministic design strategies for Φ; c) uncertainties in sensing matrix; d) additional constraints imposed on the structure of sensing matrices (non-negativity, ℓ_1, ℓ_2 norm constraints, etc.); f) non-linearities ("higher harmonics", "polynomial CS").

- **Amenable for low-complexity decoding:** Combination of algorithmic decoding/reconstruction techniques from CS and CT theory, such as list decoding, belief-propagation decoding, and orthogonal matching pursuit algorithms (OMP, ROMP, CSOMP).

WESC and Non-linear SC: Extensions

Let $B_t = \{-t, -t+1, \cdots, -1, 1, \cdots, t\} = [-t, t]$, $t \in \mathbb{Z}^+$, **be a symmetric, bounded set of integers. For a given set** $I \in [1, N]$ **and a coefficient vector** $\mathrm{b} \in B_t^{|I|}$, **let**

$$f(I, \mathrm{b}) = \sum_{i \in I} b_i \, \mathbf{v}_i,$$

where b_i **is the** i^{th} **element of** b **and** $\mathbf{v}_i$ **is the** i^{th} **column of** C. **Define, as before,**

$$d_E(\mathcal{C}, K) = \min_{((I_1, \mathbf{b}_1), (I_2, \mathbf{b}_2))} \| f(I_1, \mathbf{b}_1) - f(I_2, \mathbf{b}_2) \|_2 ,$$

where $I_{1,2} \in \mathcal{I}_K$, $(I_1, \mathbf{b}_1) \neq (I_2, \mathbf{b}_2)$.

Definition: A code $\mathcal{C}$ **is said to be a weighted ESC (WESC) with parameters** (N, m, K, d, B_t) **if** $d_E(\mathcal{C}, K) \geq d$, **for some** $0 \leq d \leq 1$.

WESC and Non-Linear SC: Extensions

Let $B_t = \{-t, -t+1, \cdots, -1, 1, \cdots, t\} = [-t, t]$, $t \in \mathbb{Z}^+$, **be a symmetric, bounded set of integers. For a given set** $I \in [1, N]$ **and a coefficient vector** $\mathrm{b} \in B_t^{|I|}$, **let**

$$f(I, \mathrm{b}) = \sum_{i \in I} b_i \, \mathbf{v}_i,$$

where b_i **is the** i^{th} **element of** b **and** $\mathbf{v}_i$ **is the** i^{th} **column of** C. **Define, as before,**

$$d_E(\mathcal{C}, K) = \min_{((I_1, \mathbf{b}_1), (I_2, \mathbf{b}_2))} \| f(I_1, \mathbf{b}_1) - f(I_2, \mathbf{b}_2) \|_2 ,$$

where $I_{1,2} \in \mathcal{I}_K$, $(I_1, \mathbf{b}_1) \neq (I_2, \mathbf{b}_2)$.

Definition: A code $\mathcal{C}$ **is said to be a weighted ESC (WESC) with parameters** (N, m, K, d, B_t) **if** $d_E(\mathcal{C}, K) \geq d$, **for some** $0 \leq d \leq 1$.

WESC and Non-linear SC: Extensions

Definition: Let $\mathcal{C}$ be a set of N codewords (vectors) $\sum_{d=1}^{D_i} a_{i,d} v_i^d$, where $v_i \in \mathbb{R}^{m \times 1}$, $i = 1, 2, \cdots, N$ and D_i is the degree of the polynomial associated with a vector v_i. A code $\mathcal{C}$ is said to be a polynomial wESC (WESC) with parameters (N, m, K, d, B_t) if it is a WESC over the extended set of polynomial codewords.

Less formally, it is a family of codes in which each codeword can have several "harmonics". For the example of a 2-harmonic code, one can take K_2 to be the number of selected columns having exactly two harmonics, so that $K_2 + \|\mathbf{b}\|_0 \leq K$.

Theoretical Results: Fundamental Reconstruction Limits for WESCs

- **Definition:** **Let**

$$N\left(m, K, d, B_t\right) := \max\left\{N : \ \mathcal{C}\left(N, m, K, d, B_t\right) \neq \phi\right\}.$$

The asymptotic code exponent is defined as

$$R\left(K, d, B_t\right) := \lim_{m\to\infty} \sup \frac{\log N\left(m, K, d, B_t\right)}{m}.$$

- **Theorem:** **For constant t, the asymptotic code exponent of WESCs can be bounded as**

$$\frac{\log K}{4K}\left(1 + o_{t,d}\left(1\right)\right) \leq R\left(K, d, B_t\right) \leq \frac{\log K}{2K}\left(1 + o_{t,d}\left(1\right)\right)$$

where $o_{t,d}\left(1\right)$ is a function of t and d, and $o_{t,d}\left(1\right) \to 0$ as $K \to \infty$.

Theoretical Results: Fundamental Reconstruction Limits for WESCs

- **Theorem:** **The polynomial code superposition rate is upper bounded by**

$$\frac{\log K}{2K}\left(1+F(t,d)\right),$$

where

$$F(t,d)=\frac{2}{\log K}\log\left(\frac{2\sqrt{A_m}\,(t+1)}{d}+\frac{1}{\sqrt{K}}\right), \qquad (1)$$

with $A=\max\{\sum_{d=1}^{D_{i_j}}\left|a_{i_j,d}\right|\}$.

Interpretation

- **The compression parameters m and N satisfy**

$$2\,K\frac{\log\,N}{\log\,K} \leq m \leq 4\,K\frac{\log\,N}{\log\,K}. \qquad (2)$$

- **Order of asymptotic code exponent does not depend on minimum Euclidean distance - can make the distance arbitrarily close to one.**

WESC: More Extensions

- **Features of WSEC I:** The parameter t can be a **constant**, or it can **grow** with K or m.

- **Features of WSEC II:** Can impose additional restrictions on the weighting set/alphabet B_t - and include **rational values**. Can try to bridge the "gap to real numbers" using the fact that for every real number ψ and an integer Q, there exists an irreducible rational number a/q such that

$$0 < q \leq n, \quad \left| \psi - \frac{a}{q} \right| \leq \frac{1}{q\,(Q+1)}.$$

  By restricting the alphabet of the weights to integers/rationals, can enforce minimum distance constraints - i.e., make the **schemes robust to errors/noise**.

- **Features of WSEC III:** Can enforce **"norm distribution"** on the codewords in order to improve code rate.

- **Features of WSEC IV:** Can work with **different normed spaces** - both with respect to distance measure and codewords. Interesting connection to Milman's theorem on **Almost Euclidean Quotient Spaces/Volumes of Convex Bodies.**

WESC with Code Uncertainty

Rather than having one signature sequence, each user can have a signature code with W codewords. This can also be seen as an instant of CS with sensing matrix uncertainty.

Laczay (2005) showed that the optimal asymptotic code rate $\log N/n$ **satisfies**

$$\frac{\log K}{4K} - \frac{\log W}{n} \leq \frac{\log N}{n} \leq \frac{\log K}{2K} - \frac{\log W}{n}.$$

Decoding/Reconstruction: Dense and Sparse WESCs

Why Dense: Most sensing matrices are dense, and need general reconstruction algorithms (redundant WESC decoders, Subspace Pursuit (SP), etc).

Why Sparse: Sparse problems can be solved more efficiently - Matching Pursuit, LP, Belief Propagation.

For the latter case, deal with sparse WESC: A WSEC code $\mathcal{C}_s$ is said to be a **regular, sparse code**, with sparsity s (where $s|m$), if every codeword $\mathbf{v} \in \mathcal{C}_s$ has support size m/s.

Discouraging fact: Loose a lot with sparsity requirement with correlation decoder! Will briefly discuss a new method that combines sparse/dense reconstruction!

Redundant WESCs Decoding

The WESC pursuit decoder: Given the measurement $\mathbf{y}$, find the i^{th} element of the input signal x via

$$x_i = -\arg\min_{a \in B_t \bigcup \{0\}} \|a\mathbf{v}_i + \mathbf{y}\|^2 .$$

Iterate process with adequate changes in $\mathbf{v}$. Use "redundant codewords" in the WESCs matrix.

Computational complexity: Smaller than that of OMP, since we essentially **only need to compute the inner product of $\mathbf{v}_i$ and $\mathbf{y}$ once**. In OMP, similar type of inner product has to be evaluated K times for each $\mathbf{v}_i$.

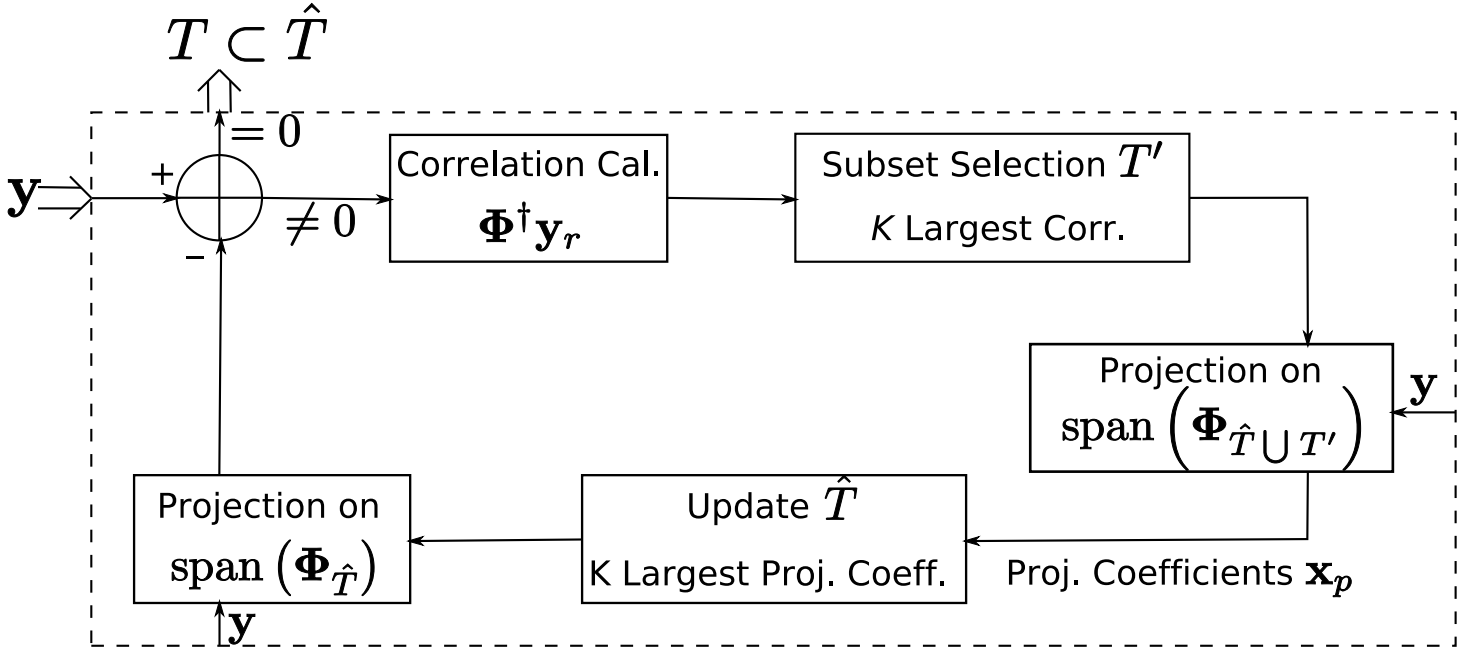
Theorem: Consider a measurement matrix $\mathbf{V} \in \mathbb{R}^{m \times N}$ with unit norm columns. For given K and t, m and N sufficiently large, if

$$\frac{\log N}{m} > \frac{1}{8K^2t^2}\left(1 + o_K(1)\right),$$

then there exists a $\mathbf{V}$ such that the WESC pursuit decoding algorithm can reconstruct every K-sparse signal.

The Subspace-Pursuit (SP) Algorithm

Similar to order-statistics Dykstra's algorithm, known in coding theory as A* (Han and Hartmann, 1992). Extensions: produce list of candidate data vectors.



The Subspace-Pursuit (SP) Algorithm: Theoretical Guarantees

Definitions: **A matrix $\Phi \in \mathbb{R}^{m \times N}$ satisfies the Restricted Isometry Property (RIP) with parameters (K, δ) for $K \leq m$, if for all index sets $I \subset \{1, \cdots, N\}$ such that $|I| \leq K$ and for all $\mathbf{q} \in \mathbb{R}^{|K|}$, it holds**

$$(1 - \delta) \|\mathbf{q}\|_2^2 \leq \|\Phi_I \mathbf{q}\|_2^2 \leq (1 + \delta) \|\mathbf{q}\|_2^2 .$$

For an RIP matrix, define δ_K as

$$\delta_K := \inf \Big\{ \delta : \ (1 - \delta) \|\mathbf{q}\|_2^2 \leq \|\Phi_I \mathbf{q}\|_2^2 \leq (1 + \delta) \|\mathbf{q}\|_2^2 , \\ \forall \mathbf{q} \in \mathbb{R}^K, \ \forall |I| \leq K \Big\} .$$

Theorem: **Assume that $\mathbf{x} \in \mathbb{R}^N$ is an arbitrary K-sparse signal, and let the weighted sum of the codewords be $\mathbf{y} = \Phi \mathbf{x} \in \mathbb{R}^m$. If the measurement matrix Φ satisfies the RIP with parameter**

$$\delta_{3K} < 6 - \sqrt{35} \approx 0.084, \tag{3}$$

then the SP algorithm can exactly recover $\mathbf{x}$ from $\mathbf{y}$, using a finite number of iterations.

The Subspace-Pursuit (SP) Algorithm: Complexity

- **LP-Decoding:** **Condition for exact recovery**

$$\delta_{3K} < \mathbf{0.33};$$

Complexity: $O(N^3)$.

- **ROMP:** **Condition for exact recovery**

$$\delta_{3K} < \frac{\mathbf{0.03}}{\log K};$$

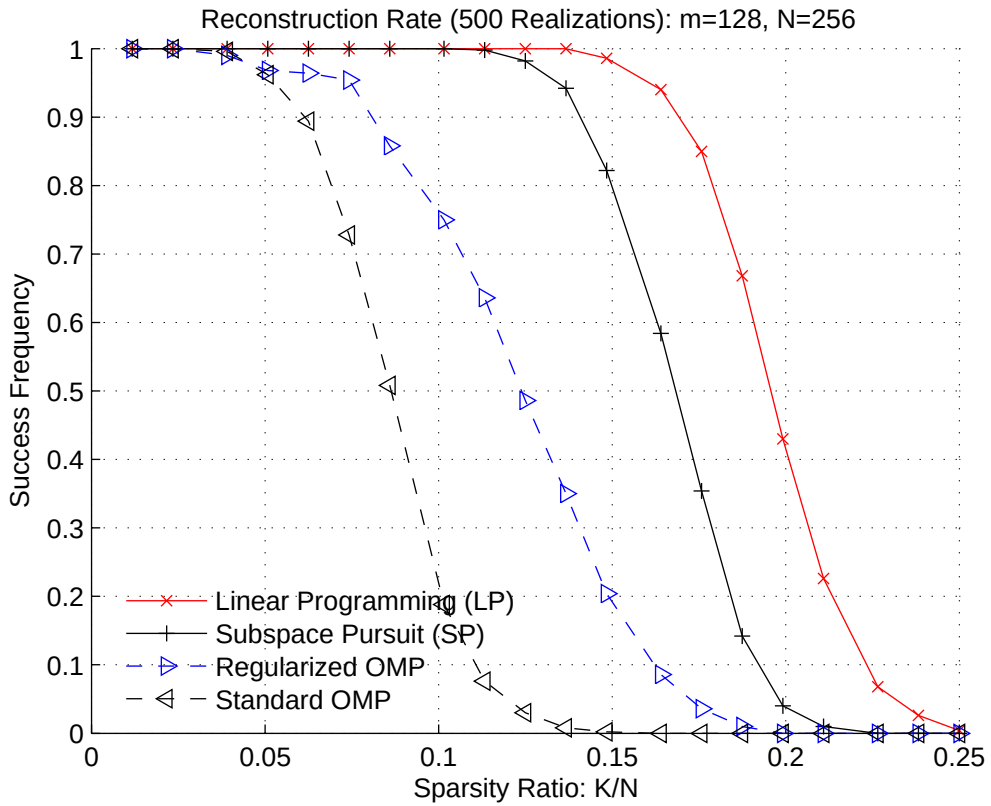
Complexity: $O(m\,N\,K)$.

- **SP-Decoding:** **Condition for exact recovery**

$$\delta_{3K} < \mathbf{0.12};$$

Complexity: $O(m\,N\,K)$ **or** $O(m\,N\,\log\,K)$**, depending on the compressibility of the signal (and given** p**).**

The SP Algorithm: $0-1$ signals



The Sparse-Dense Codes and Reconstruction Algorithms

Special Structure of Φ: the sensing matrix consists of all non-zero codewords of a low-density parity-check (LDPC) code. Computation of correlation between y and the codewords boils down to LDPC decoding. For the latter, use LP or BP decoding, complexity only of the order of m^3 or m, respectively.

Deterministic Code/Matrix Constructions

The Design Approaches

- **Spherical Codes:** Ericsson and Zinoviev, 2001. Sophisticated constructions involving specialized binary trees and nested codes assigned to interior nodes of the tree.

- **Superimposed Codes:** Can be generated from spherical codes (Danev, 2004) or based on real-valued mappings from q-ary error-control codes, or primitive polynomials.

- **Example:** Let C be a binary linear $[N, K, D]$ block code that contains the all-ones codeword. Delete all codewords starting with "1" and puncture the remaining ones in the first position. Apply the mapping

$$a \rightarrow \frac{a}{\sqrt{N-1}},\ a \in \{0, 1\}.$$

  Provided that D, K, N and d satisfy certain conditions, the code can be shown to be ESC. Similar mappings (but slightly more involved) can be devised for WSEC and other CS categories).

The Design Approaches

- **Definition: B_h Sequences: A sequence of distinct non-negative integers $n_1, n_2, n_3, \ldots, n_M$, $n_i \in [1, N]$, is a B_h sequence if the sums of not more than h elements are all distinct. More details in Halberstam and Roth, Sequences, 1983.**

- **Expansions of elements of B_h sequences lead to columns of WSEC/CS matrices.**

THANK YOU!