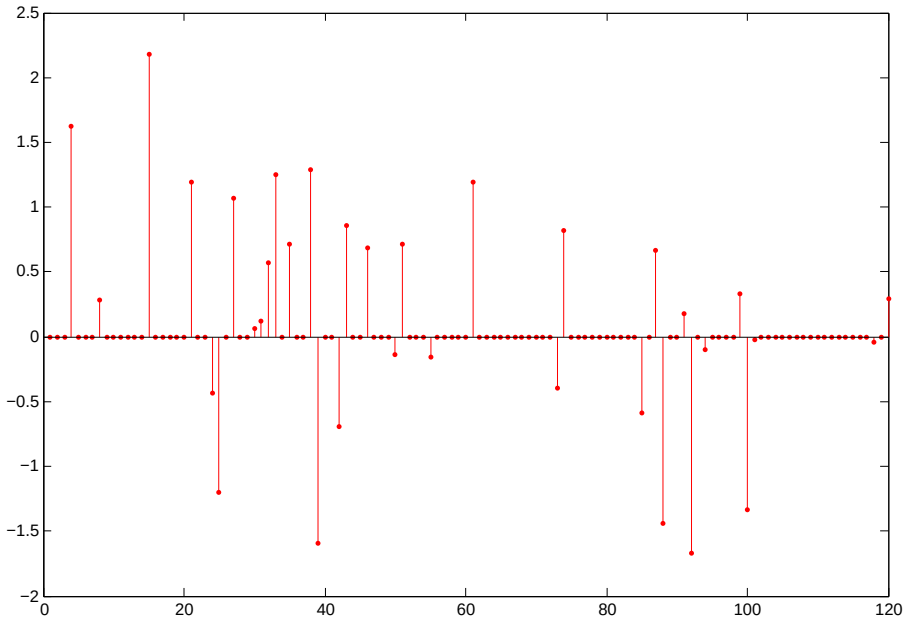
Near Theoretical Lower Bound Sparse Representation and Compressive Sampling

Vahid Tarokh

School of Engineering and Applied Sciences
Harvard University

Sparse Signals

- Let $\mathbf{x} \in \mathbb{C}^M$. Let $\text{supp}(\mathbf{x}) = \{i | x_i \neq 0\}$ and $||\mathbf{x}||_0 = |\text{supp}(\mathbf{x})|$.
- $\mathbf{x}$ is said to be L-sparse if $||\mathbf{x}||_0 = L << M$.



Motivation

- Many human-made and natural signals are sparse in transform domains.

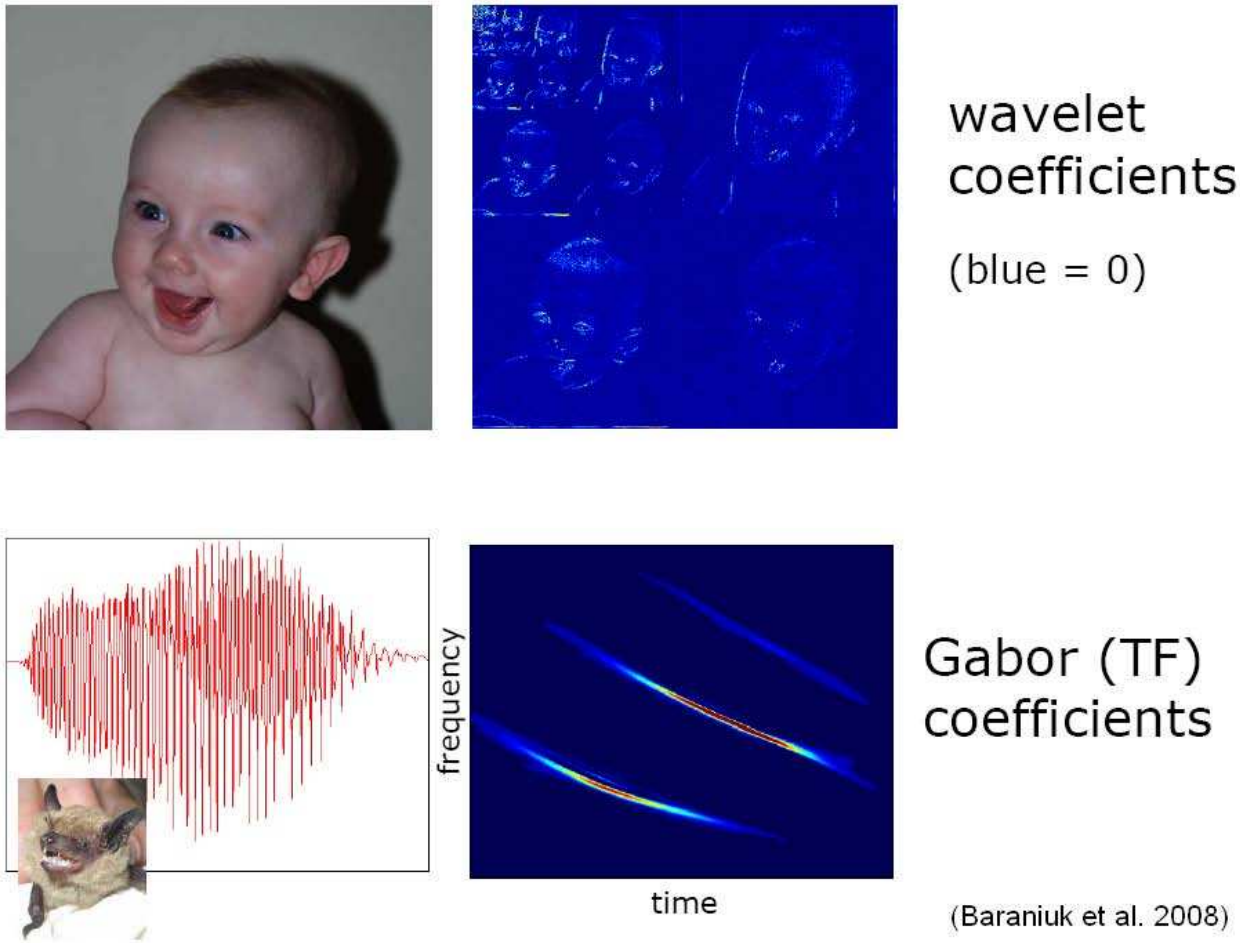


Figure: Sparsity in Transform Domain

Sparse Representations

- An interesting problem is sparse representation of a vector $\mathbf{r} \in \mathbb{C}^N$ using a given dictionary (frame) $\mathcal{F}$ of $M > N$ non-zero vectors $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_M\}$.
- What is the most compact (e.g. with least number of non-zero coefficients) way to describe $\mathbf{r}$ a linear combination of the elements of $\mathcal{F}$ with some error measure?
- This happens often in real life applications:
- Given a matrix $\mathbf{F} \in \mathbb{C}^{N \times M}$ with columns elements of $\mathcal{F}$, solve
  - Sparsity Constrained Approximation

$$\min_{\mathbf{c} \in \mathbb{C}^M} ||\mathbf{r} - \mathbf{F}\mathbf{c}||_2^2 \quad \text{s.t. } ||\mathbf{c}||_0 \leq L \quad \text{(SCA)} \tag{1}$$

- If some rich frames for which the problem of sparse representation can be quickly solved can be found (with their associated approximation algorithms), then it can essentially replace Fourier basis in many applications, e.g. compression, estimation, etc.

Sampling of Sparse Signals

- Another interesting problem is the sampling of sparse signals.
- Obviously, if a signal is sparse then it can be efficiently sampled.
- This has been done often for signals that are sparse in transform domain, etc.
- If $\mathbf{x}$ is sparse and we use b bits to describe each complex number, we need at most $L \log N + Lb$ bits to describe a sparse signal (compared to Nb).

Sampling of Sparse Signals

- Another way to sample an L-sparse signal $\mathbf{x}$ is by compressed sensing:
- Let $\mathbf{A}$ be a $N \times M$ measurement matrix. We measure

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}.$$

($\mathbf{n}$ is sampling noise).

- Then we need to find $\mathbf{x}$ from $\mathbf{y}$.
- This is called *noisy* or *noiseless* compressive sampling depending on the noise n being present of not.
- In noiseless case, one way to find $\mathbf{x}$ is by solving

$$\min_{\mathbf{x} \in \mathbb{C}^M} ||\mathbf{x}||_0 \quad \text{s. t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x}.$$

Historical Remarks

- Sparse representation has existed in speech literature, and been used in JPEG, etc.
- An influential paper by Stephan Mallat in 1991, proposed the matching pursuit approach and popularized this topic.
- Compressed sensing has also been around in the literature: recovery of a signal from limited number of Fourier coefficients, etc.
- Tibshirani proposed an L_1 approach to solving the compressed sensing problem in 1996.

Historical Remarks

- Recent papers by Donoho, Candes and Tao re-defined this problem, and studied the L_1 approach in details.
- In recent years, we argued that these problems are better understood in *Information Theoretic* context.
- Using this, we improved on the existing results in terms of theoretical limits, performance, and algorithmic complexity.
- We also came up with simple constructions.
- This talk summarizes some of our results.

The L_1 Approach

- To start let us consider **noiseless** complex sampling model $\mathbf{y} = \mathbf{A}\mathbf{x}$ where $\mathbf{x}$ is L-sparse.
- This of course more or less never happens in reality, but is good for developing the required intuition.
- To recover a sparse $\mathbf{x}$ from observations $\mathbf{y} = \mathbf{A}\mathbf{x}$, the ℓ_1 regularization

$$\min_{\mathbf{x} \in \mathbb{C}^M} ||\mathbf{x}||_1 \quad \text{s. t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x}$$

  can be solved via linear programming.
- What should be the minimum number of samples N (dimension of $\mathbf{y}$) to be able to reconstruct $\mathbf{x}$ from $\mathbf{y}$ uniquely?
- What properties A must hold for this to be possible?
- What is the complexity of implementation of the recovery algorithm?

The L_1 Approach

- Candes and Tao proved (using L_1 approach and Gaussian measurement matrices) a result that gives (for instance):
  - When $\frac{N}{M} = \frac{3}{4}$, then L-sparse signals with $L \leq 3.6 \times 10^{-4} M$ can be recovered.
  - When $\frac{N}{M} = \frac{2}{3}$, then L-sparse signals with $L \leq 3.2 \times 10^{-4} M$ can be recovered.
  - When $\frac{N}{M} = \frac{1}{2}$, then L-sparse signals with $L \leq 2.3 \times 10^{-4} M$ can be recovered.
- In contrast, we have shown that $L \leq 0.5N$ is necessary and sufficient and have constructed matrices $\mathbf{A}$ for which this can be done with $O(ML)$ complexity.

The First Insight

The Main Idea

- Consider the vector space

$$\mathcal{V} = \{\mathbf{d} \in \mathbb{C}^M : \mathbf{Ad} = \mathbf{0}\}.$$

- We refer to $\mathcal{V}$ as the underlying *code* of measurement matrix A.
- Note $\dim(\mathcal{V}) = M - N$ (assuming $\mathbf{A}$ being full-rank).
- If $\mathbf{c} \in \mathbb{C}^M$ such that $\mathbf{r} = \mathbf{Ac}$, then all the representations of $\mathbf{r}$ are given by $\mathbf{c} - \mathcal{V} = \{\mathbf{c} - \mathbf{d} \mid \mathbf{d} \in \mathcal{V}\}$.
- Thus the problem is equivalent to finding $\mathbf{d} \in \mathcal{V}$ which minimizes $\|\mathbf{c} - \mathbf{d}\|_0$.
- $\mathcal{V}$ a linear code over $\mathbb{C} \Rightarrow$ Find the error vector $\mathbf{e} = \mathbf{c} - \mathbf{d}$ of minimum Hamming weight ($||\mathbf{e}||_0$) over all codewords $\mathbf{d} \in \mathcal{V}$.
- Well-studied in coding theory.

MDS Bound

- Let us recall the following Theorem from coding theory.

Theorem

If a linear code has length M and dimension $M-N$, then its minimum Hamming distance is $\leq N+1$. Equality is achieved if the code is Maximum Distance Separable (MDS) code. A code with minimum distance d_{min} can correct up to $\lfloor \frac{d_{min}-1}{2} \rfloor$.

- So we uniquely recover up to iat most $L = \frac{N}{2}$, and this can be done if the measurement matrix is MDS.
- For all Gaussian i.i.d matrices, the measurement matrix is MDS with probability 1.
- Inspired by Reed-Solomon codes, one can give a measurement matrix for which unique recovery can be done in real time.

Vandermonde Measurement Matrix

- We define the Vandermonde measurement matrix as

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & \cdots & \cdots & 1 \\ z_1 & z_2 & \cdots & \cdots & z_M \\ z_1^2 & z_2^2 & \cdots & \cdots & z_M^2 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ z_1^{N-1} & z_2^{N-1} & \cdots & \cdots & z_M^{N-1} \end{pmatrix} \tag{2}$$

  where $z_i,\ i = 1, 2, \cdots, M$ are *distinct, non-zero* complex numbers.

- Any arbitrary set of N distinct columns of $\mathbf{A}$ are linearly independent.

Vandermonde Frames

- The code $\mathcal{V}$ has the following interesting property.

Lemma

For any non-zero vector $\mathbf{v} \in \mathcal{V}$, we have $\|\mathbf{v}\|_0 > N$. Moreover, there exist vectors $\mathbf{v} \in \mathcal{V}$ with $\|\mathbf{v}\|_0 = N + 1$.

- In the terminology of coding theory the code $\mathcal{V}$ is a $[M, M - N, N + 1]$ maximum distance separable (MDS) linear code.

Vandermonde Frames

- We can decode this code up to half minimum distance bound.

Lemma

Let $\hat{\mathbf{c}}$ be a solution to the ℓ_0 minimization problem

$$\min_{\mathbf{c}\in\mathbb{C}^M} ||\mathbf{c}||_0 \quad s.\ t. \quad \mathbf{r} = \mathbf{F}\mathbf{c}.$$

If $||\hat{\mathbf{c}}||_0 \leq N/2$ then there is a unique solution to the ℓ_0 minimization problem.

- We have also designed an $O(MN)$ complexity decoder that achieves this bound.
- For more details see: M. Akçakaya and V. Tarokh, “A Frame Construction and A Universal Distortion Bound for Sparse Representations,” IEEE Trans. Signal Processing, Vol. 56, Number 6, pp. 2443-2550, June 2008.

The Noisy Problem

Sparse Estimation Scenario

- Consider $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$ with $\mathbf{n}$ i.i.d. Gaussian noise with mean zero and variance σ^2 per dimension.
- Suppose that $\mathbf{x}$ has at most L non-zero elements.
- If I, the location (indices) of these non-zero elements are given by a genie, then let $\mathbf{A}_I$, $\mathbf{x}_I$ be respectively the columns and rows of $\mathbf{A}$ and $\mathbf{x}$ corresponding to indices in I.
- Then $\mathbf{y} = \mathbf{A}_I\mathbf{x}_I + \mathbf{n}$.
- It is well-known that the pseudo-inverse estimator $(\mathbf{A}_I^H\mathbf{A}_I)^{-1}\mathbf{A}_I^H\mathbf{y}$ achieves the mean square error given by Cramer-Rao bound $Tr[(\mathbf{A}_I^H\mathbf{A}_I)^{-1}]\sigma^2$.
- What happens if we do not know the indices I?

Dantzig Selector

- Candes and Tao analyze this situation in the paper "The Dantzig Selector" in Annals of Statistics 2007.
- The Dantzig Selector solves

$$\min ||\mathbf{x}||_1 \quad \text{s. t.} \quad ||\mathbf{A}^*(\mathbf{Ax} - \mathbf{y})||_\infty \leq \gamma,$$

  where γ is a function of σ and M.
- They prove that the Dantzig selector achieves an error that is larger 'only' by a factor of $\log N$ than $Tr[(\mathbf{A}_I^H \mathbf{A}_I)^{-1}]\sigma^2$. *It seems intuitively appealing: They state that: "Hence, Theorem 1.1 says that the minimum L_1 estimator achieves a loss within a logarithmic factor of the ideal mean squared error; the logarithmic factor is the price we pay for adaptivity, that is, for not knowing ahead of time where the nonzero parameter values actually are".*
- *The Dantzig Selector is additionally very complex to implement.*
- *In contrast, in a 2009 paper we proved that the Genie bound is asymptotically achievable (no logarithmic penalty exists!).*

The L_1 Approach

- Similarly for noisy compressed sensing, we provide **significantly** better trade-offs and much lower complexity of implementation.
- Also, for sparse representation, we can provide theoretical bounds on the performance and design dictionaries for which the problem can be solved in real time.
- What enables us to do better than L_1 approach is that we are inspired by theories of information.
- However, our underlying proofs are significantly more complicated than those of Information Theory.
- I will give the underlying insights next.

The Second Insight

Shannon's Idea

- Consider an i.i.d. sequence $X_1, X_2, X_3, \cdots$ samples of random variable X. Then

$$-\frac{\log p(X_1, X_2, \cdots, X_N)}{N} = -\frac{(\log p(X_1) + \log p(X_2) + \cdots + \log p(X_N))}{N}.$$

and

$$-\frac{(\log p(X_1) + \log p(X_2) + \cdots + \log p(X_N))}{N} \to -E \log p(X) = H(X)$$

- Thus for N large enough

$$\Pr(2^{N(H(X)-\epsilon)} < p(X_1, X_2, \cdots, X_N) < 2^{N(H(X)+\epsilon)}) > (1-\epsilon).$$

Shannon's Idea

- This means that with probability near 1 sequences $X_1, X_2, \cdots, X_N$ have $p(X_1, X_2, \cdots, X_N) \simeq 2^{-NH(X)}$.
- There can be roughly $2^{NH(X)}$ sequences of these type called **typical** sequences.
- Other sequences do not happen with very small probability.
- Shannon used this insight to answer some fundamental questions of compression and communications.
- We can also get some mileage out of his idea (by some modification).

Shannon's Other Ideas

- Instead of insisting on zero probability of error, allow for arbitrary small probability of errors.
- Let the size of problem grow large.
- Create many reasonable transmission strategies (random codes) and show that the average probability of error goes to zero over all these strategies when certain conditions are satisfied (using the typicality ideas).
- Deduce that the probability of error goes to zero asymptotically for any reasonable transmission strategy with probability 1.
- To show that the conditions are necessary, use Fano's converse.

Noisy Compressive Sampling

- Let us apply this idea to the noisy compressive sampling model

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}.$$

- In this case exact recovery of $\mathbf{x}$ is not possible.
- Fundamental questions that need to be answered are:
- What are the performance limits for a set of L, M, N and a given performance criterion?
- How do we achieve these limits in practice?

Our Performance Metrics

- Metric 1: Recovering support of $\mathbf{x}$ correctly
- Metric 2: Recovering $(1-\alpha)$ fraction of support of $\mathbf{x}$
- Metric 3: Recovering $(1-\gamma)$ fraction of the energy of $\mathbf{x}$

Problem Formulation

- A sequence of vectors, $\{\mathbf{x}^{(M)}\}$ indexed by M such that $\mathbf{x}^{(M)} \in \mathbb{C}^M$.
- Let $\mathcal{I}^{(M)} = \text{supp}(\mathbf{x}^{(M)})$, where $|\mathcal{I}^{(M)}| = L^{(M)}$.
- Let $P^{(M)} = ||\mathbf{x}^{(M)}||_2^2$.
- Consider an ensemble of $N \times M$ Gaussian measurement matrices, $\mathbf{A}^{(M)}$, where N is a function of M and $(i,j)^{\text{th}}$ term $a_{i,j} \sim \mathcal{N}_C(0,1)$.
- A decoder, $\mathcal{D}(\cdot)$ will output a set of indices, $\mathcal{D}(\mathbf{y})$.

Problem Formulation

- The average probability of error, averaged over all *Gaussian measurement matrices*:

$$p_{\text{err}}(\mathcal{D}|\mathbf{x}^{(M)}) = \mathbb{E}_{\mathbf{A}}\big(p_{\text{err}}(\mathbf{A}|\mathbf{x}^{(M)})\big), \tag{3}$$

  where $p_{\text{err}}(\mathbf{A}|\mathbf{x}^{(M)}) = \mathbb{P}(\mathcal{D}(\mathbf{y}) \neq \mathcal{I})$ for $\mathbf{y} = \mathbf{A}\mathbf{x}^{(M)} + \mathbf{n}$ and $\mathbb{P}(\cdot)$ is the probability measure.
- A decoder achieves *asymptotic reliable* sparse recovery if $p_{\text{err}}(\mathcal{D}|\mathbf{x}^{(M)}) \to 0$ as $M \to \infty$.
- Asymptotic reliable sparse recovery is not possible if $p_{\text{err}}(\mathcal{D}|\mathbf{x}^{(M)})$ stays bounded away from 0 as $M \to \infty$.

Main Results

Theorem

Let a sequence of sparse vectors, $\{\mathbf{x}^{(M)} \in \mathbb{C}^M\}_M$ *with power* $P^{(M)}$ *and* $||\mathbf{x}^{(M)}||_0 = L^{(M)} = \left\lfloor \frac{1}{\beta} M \right\rfloor$, *where* $\beta > 2$ *be given. Then asymptotic reliable recovery is possible for* $\{\mathbf{x}^{(M)}\}$ *if*

$$N \succ C_1 L$$

for some constant $C_1 > 1$ *where "* $\succ$*" denotes asymptotically greater than.*

- $N = O(L)$ holds for all performance metrics.
- Performance metric determines $P^{(M)}$ and C_1.
- For performance metrics 2 and 3, $P^{(M)}/\nu^2$ is a constant.
- For details (of constants) and also more results see: M. Akçakaya and V. Tarokh, "Shannon Theoretic Limits on Noisy Compressive Sampling, IEEE Trans. Info Theory Vol. 56, No. 1, pp. 492-504, Jan 2010. "

Converse

Theorem

Let a sequence of sparse vectors, $\{\mathbf{x}^{(M)} \in \mathbb{C}^M\}_M$ with $||\mathbf{x}^{(M)}||_0 = L = \left\lfloor \frac{1}{\beta} M \right\rfloor$, where $\beta > 2$ be given. Then asymptotic reliable recovery is not possible for $\{\mathbf{x}^{(M)}\}$ if

$$N \prec C_2 L$$

for some $C_2 \geq 0$.

- $N = O(L)$ holds for all performance metrics.
- Performance metric determines C_2.
- Thus $N = O(L)$ is necessary and sufficient when $M = \beta L$ with $\beta > 2$.

Sketch of Proof: Joint Typicality

- Let $\mathbf{A}_{\mathcal{J}}$ be the matrix formed by taking the columns of $\mathbf{A}$ specified by the set of indices $\mathcal{J}$.
- Let $\pi_{\mathbf{A}_{\mathcal{J}}}$ be the projection matrix onto the subspace spanned by the columns of $\mathbf{A}_{\mathcal{J}}$.
- *Joint Typicality*: We say an $N \times 1$ noisy observation vector, $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$ and a set of indices $\mathcal{J} \subset \{1, 2, \ldots, M\}$, with $|\mathcal{J}| = L$, are δ-jointly typical if $\text{rank}(\mathbf{A}_{\mathcal{J}}) = L$ and

$$\left| \frac{1}{N} ||\pi^{\perp}_{\mathbf{A}_{\mathcal{J}}} \mathbf{y}||^2 - \frac{N-L}{N}\sigma^2 \right| < \delta, \tag{4}$$

Sketch of Necessity Proof

- Proceed in a similar fashion to Shannon.
- Define events

$$E_{\mathcal{J}} = \{\mathbf{y} \text{ and } \mathcal{J} \text{ are } \delta\text{-jointly typical } \}.$$

- Let $K = |\mathcal{I} \cap \mathcal{J}|$ where $\mathcal{I} = \text{supp}(\mathbf{x})$.
- Using large deviations theory, we bound $\mathbb{P}(E_{\mathcal{J}})$ as a function of the overlap, K and $\min_{i \in \mathcal{I}} |x_i|$ and we bound $\mathbb{P}(E_{\mathcal{I}}^C)$.
- We count the sets of indices that overlap with $\mathcal{I}$ in K positions.
- We bound $p_{err}(\mathcal{D}|\mathbf{x})$ using the union bound, and then maximize this union bound over K.
- For achievability we choose N such that this upper bound on $p_{err}(\mathcal{D}|\mathbf{x})$ goes to 0.

Sketch of Proofs - Converse

- To prove the converse result, we consider a related genie-aided decoding problem. The error probability of this genie-aided decoding problem gives a lower bound on the error probability of the actual recovery problem.
- Then we derive conditions under which this lower bound stays bounded away from 0.
- Let $\text{supp}(\mathbf{x}) = \{i_1, i_2, \ldots, i_L\}$ with $i_1 < i_2 < \cdots < i_L$.
- Assume a genie provides $\mathbf{x}_{\mathcal{I}} = (x_{i_1}, x_{i_2}, \ldots, x_{i_L})^T$.
- We treat $\mathbf{x}_{\mathcal{I}}$ as channel coefficients of a vector channel.

Sketch of Proofs - Converse

- The transmission scenario is as follows:
  - The encoder takes the index set $\mathcal{J}$ as input and outputs the k^{th} row of $\mathbf{A}_{\mathcal{J}}$ at the k^{th} transmission.
  - This vector is transmitted over the vector channel with an additive Gaussian noise, $\mathcal{N}_C(0, \nu^2)$.
  - In the case that the input to the encoder was the index set $\mathcal{I}$, after N transmissions, the received vector is $\mathbf{y} = \mathbf{A}_{\mathcal{I}}\mathbf{x}_{\mathcal{I}} + \mathbf{n}$.
- Thus the information transmitted over the channel is contained in the index set.
- We next define the codebook.

Sketch of Proofs - Converse

- Let $\mathcal{J} = \{j_1, j_2, \ldots, j_L\}$ with $j_1 < j_2 < \cdots < j_L$.
- Let $\mathbf{z}_k^{\mathcal{J}} = (a_{k,j_1}, a_{k,j_2}, \ldots, a_{k,j_L})^T$, where $a_{m,n}$ is the $(m,n)^{\text{th}}$ term of $\mathbf{A}$. i.e. $\mathbf{z}_k^{\mathcal{J}}$ is the transpose of the k^{th} row of $\mathbf{A}_{\mathcal{J}}$.
- The codebook is specified by

$$\mathcal{C} = \left\{ \begin{pmatrix} \mathbf{z}_1^{\mathcal{J}} & \mathbf{z}_2^{\mathcal{J}} & \ldots & \mathbf{z}_N^{\mathcal{J}} \end{pmatrix} \middle| \mathcal{J} \subset \{1, 2, \ldots, M\}, |\mathcal{J}| = L \right\},$$

  and has size $\binom{M}{L}$.
- We define the rate of the code to be $R = \log |\mathcal{C}|$.
- Asymptotically R grows as $MH(L/M)$. We denote this by $R \doteq MH(L/M)$.

Sketch of Proofs - Converse

- The output of the channel, $\mathbf{y}$ is

$$y_k = \mathbf{H}\mathbf{z}_k^{\mathcal{I}} + n_k \quad \text{for} \quad k = 1, 2, \ldots, N,$$

  where $\mathbf{H} = (\mathbf{x}_{\mathcal{I}})^T$; y_k and n_k are the k^{th} coordinates of $\mathbf{y}$ and $\mathbf{n}$.
- By Fano's converse to the Shannon Channel Coding Theorem, if $R \doteq MH(L/M) > C$ then error probability is bounded away from 0, where

$$C = N \log \left(1 + \frac{1}{L} \frac{\mathbb{E}(||\mathbf{z}_k^{\mathcal{J}}||^2)}{\mathbb{E}n_k^2} \mathbf{H}\mathbf{H}^\dagger \right) = N \log \left(1 + \frac{P}{\nu^2} \right).$$

- For performance metrics 2 and 3, we also allow distortion.
- Rate is dictated by the rate distortion function of the source.
- If $R(D) > C$ then error probability is bounded away from 0.

Back To The Estimation Problem

Sparse Estimation Scenario

- Consider $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$ with $\mathbf{n}$ i.i.d. Gaussian noise with mean zero and variance σ^2 per dimension.
- Suppose that $\mathbf{x}$ has at most L non-zero elements.
- If I, the location (indices) of these non-zero elements are given by a genie, then let $\mathbf{A}_I$, $\mathbf{x}_I$ be respectively the columns and rows of $\mathbf{A}$ and $\mathbf{x}$ corresponding to indices in I.
- Then $\mathbf{y} = \mathbf{A}_I\mathbf{x}_I + \mathbf{n}$.
- It is well-known that the pseudo-inverse estimator $(\mathbf{A}_I^H\mathbf{A}_I)^{-1}\mathbf{A}_I^H\mathbf{y}$ achieves the mean square error given by Cramer-Rao bound $e_G(N) = Tr[(\mathbf{A}_I^H\mathbf{A}_I)^{-1}]\sigma^2$.
- What happens if we do not know the indices I?

Main Result

- A Joint Typicality Estimator finds a jointly typical subspace with $\mathbf{y}$ and produces $\hat{\mathbf{x}}$ by projecting $\mathbf{y}$ onto this subspace.

Theorem

Without any knowledge of $\mathcal{I}$, Joint Typicality Estimator has mean squared error $e_\delta(N)$. Under very mild conditions, we proved that

$$\lim_{N\to\infty} |e_\delta(N) - e_G(N)| = 0$$

- For more details, please see: B. Babadi, N. Kalouptsidis, and V. Tarokh, “Asymptotic Achievability Of The Cramér-Rao Bound For Noisy Compressive Sampling,” IEEE Transactions on Signal Processing Vol. 57, No. 3, pp. 1233-1236, March 2009.

Shannon and Sparse Representation

Sparse Representations

- Closely related to compressive sampling is the problem of finding a sparse representation for a vector in terms of a given dictionary.
- Given a vector $\mathbf{r} \in \mathbb{C}^N$ and a matrix $\mathbf{F} \in \mathbb{C}^{N \times M}$ whose non-zero columns from a **frame** $\mathcal{F}$. We have $|\mathcal{F}| = M$, and the elements of $\mathcal{F}$ span $\mathbb{C}^N$.
- We are interested in solving one of the following
  - Error Constrained Sparse Approximation

$$\min_{\mathbf{c} \in \mathbb{C}^M} ||\mathbf{c}||_0 \quad \text{s.t.} \ ||\mathbf{r} - \mathbf{F}\mathbf{c}||_2^2 \leq \delta ||\mathbf{r}||^2 \quad \text{(ECSA)} \tag{5}$$

  - Sparsity Constrained Approximation

$$\min_{\mathbf{c} \in \mathbb{C}^M} ||\mathbf{r} - \mathbf{F}\mathbf{c}||_2^2 \quad \text{s.t.} \ ||\mathbf{c}||_0 \leq L \quad \text{(SCA)} \tag{6}$$

Sparse Representations

- If the SCA problem can be solved in polynomial time, then so can the ECSA problem. Thus we focus on SCA, which is a minimum distance decoding problem.
- The solution to SCA is invariant to scaling. Thus we can normalize any given vector $\mathbf{r}$ to have length $\sqrt{N}$.
- We use the average distortion as the measure of quality of a given frame

$$D(\mathcal{F}) = \frac{1}{N}\mathbb{E}_{\mathbf{r}} \min ||\mathbf{r} - \mathbf{F}\mathbf{c}||^2,$$

  where $\mathbf{r}$ is uniformly distributed on the N-dimensional hypersphere of radius $\sqrt{N}$.
- This is the worst case characterization of distortion since it assumes no knowledge of the distribution of $\mathbf{r}$.

A Geometrcal Insight

- Let $L = \epsilon N$ be the sparsity constraint.
- Let $M = rN$, $r \geq 1$.
- There are at most $T = \binom{M}{L}$ L-dimensional subspaces, $\{\mathcal{P}_k\}$ of $\mathbb{C}^N$ that are spanned by $\{\phi_j\}_{j \in \mathcal{I}_k}$ as $\mathcal{I}_k$ ranges over all possible L combinations of the indices frame elements.
- Given a vector $\mathbf{r}$, we want to find the closest $\mathcal{P}_k$, i.e. minimize $||\mathbf{r} - \pi_{\mathcal{P}_k}\mathbf{r}||^2$, where $\pi_{\mathcal{P}_k}$ is the projection operator onto $\mathcal{P}_k$.
- Thus

$$D(\mathcal{F}) = N\mathbb{E}(\min_k d^2(\mathbf{x}, \mathcal{P}_k)) = N \int_0^\infty \mathbb{P}(\min_k d^2(\mathbf{x}, \mathcal{P}_k) \geq \eta) d\eta \qquad (7)$$

Generalized spherical Caps

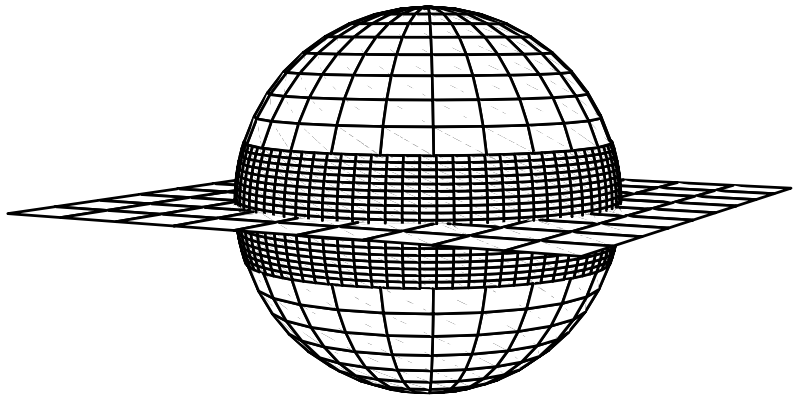
- We next define an L-dimensional complex generalized cap of radius $\sqrt{\rho}$ around an L-dimensional plane $\mathcal{P}_k$ as

$$GC_L(\rho, \mathcal{P}_k) = \{\mathbf{x} \in \mathbb{S}^N : ||\mathbf{x} - \pi_{\mathcal{P}_k}\mathbf{x}||^2 \leq \rho\} \tag{8}$$

  where $\mathbb{S}^N$ is the N dimensional complex unit hypersphere, $\{\mathbf{x} \in \mathbb{C}^N : ||\mathbf{x}||^2 = 1\}$.

Generalized spherical Caps

- The following figure depicts a generalized cap (in $\mathbb{R}^N$) for $N = 3$ and $L = 2$



Key Insight

- Let $d^2(\mathbf{x}, \mathcal{P}_k) = \frac{1}{N}||\mathbf{x} - \pi_{\mathcal{P}_k}\mathbf{x}||^2$ for $||\mathbf{x}||^2 = N$.
- The key observation about the distribution of $\min_k d^2(\mathbf{x}, \mathcal{P}_k)$ is the following

$$\begin{aligned}
&\mathbb{P}(\min_k d^2(\mathbf{x}, \mathcal{P}_k) \leq \eta \mid \mathbf{x}) \\
&\quad = \mathbb{P}(\text{There exists a plane within distance } \sqrt{\eta} \text{ of } \mathbf{x} \mid \mathbf{x}) \\
&\quad = \mathbb{P}(\mathbf{x} \text{ is in the area covered by the generalized} \\
&\qquad \text{caps of radius } \sqrt{\eta} \mid \mathbf{x}) \\
&\quad = \mathbb{P}\left(\mathbf{x} \in \bigcup_{k=1}^{T} GC_L(\mathcal{P}_k, \eta) \;\middle|\; \mathbf{x}\right)
\end{aligned}$$

Universal Performance Limits

- By bounding the right side from the above, we derived a universal lower bound on the average distortion as a function of L, N, M.

Theorem

For any frame $\mathcal{F}$ and $\mathbf{r}$ uniformly distributed on the N-dimensional hypersphere of radius $\sqrt{N}$, we have the following asymptotic distortion lower bound

$$\lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{\mathbf{r}}\left(\min_{1\leq k\leq T} ||\mathbf{r} - \Pi_{\mathcal{P}_k}\mathbf{r}||^2 \right) \geq \frac{\kappa_0(1-\epsilon)}{1-\epsilon\kappa_0} \tag{9}$$

where

$$\kappa_0 = 2^{-\frac{r}{1-\epsilon}H\left(\frac{\epsilon}{r}\right)} \epsilon^{\frac{\epsilon}{1-\epsilon}} \tag{10}$$

- For more details see: M. Akçakaya and V. Tarokh, “A Frame Construction and A Universal Distortion Bound for Sparse Representations,” IEEE Trans. Signal Processing, Vol. 56, Number 6, pp. 2443-2550, June 2008.

Constructions

Historical Remark

- Shannon's theoretical limits are derived with random codes.
- It took about 50 years for constructive methods to achieve performance near Shannon limit.
- Our proofs work for random measurement matrices and frames. Thus,these random structures achieve our limits.
- We have tried to achieve our theoretical limits constructively and with low complexity.
- Our goal is to construct $\mathbf{A}$ with structure that can be used to design high-performance and low-complexity recovery algorithms. But performance also needs to be near theoretical limits.
- In constructing our frames and compressed sensing metrices, we use the connections with coding theory that we discussed above.
- In particular, inspired by low density codes, we next create low density frames.

Low Density Frames

Low Density Frames

- Let

$$\mathbf{F} = \begin{pmatrix} \phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,M} \\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,M} \\ \vdots & \ddots & \ddots & \vdots \\ \phi_{N,1} & \phi_{N,2} & \cdots & \phi_{N,M} \end{pmatrix}. \tag{11}$$

- We define a (d_v, d_c)-*regular* LDF as a matrix $\mathbf{F}$ that has d_c non-zero elements in each row and d_v non-zero elements in each column. Clearly $Md_v = Nd_c$.
- We restrict ourselves to *binary* regular LDFs, where the non-zero elements of $\mathbf{F}$ are ones.
- The density ρ of a frame $\mathbf{F}$ is the ratio of the number of non-zero entries of $\mathbf{F}$ to the dimension of $\mathbf{F}$. We consider regular LDFs for which $\rho = (Md_v)/(MN) = d_v/N$ is small.

Low Density Frames

- The **first component** of our construction is to use **low density frames** (LDF).
- For the rest of the talk, we will consider the noisy compressive sampling problem

$$\mathbf{y} = \mathbf{F}\mathbf{x} + \mathbf{n},$$

  where $\mathbf{F}$ is an LDF and $\mathbf{n} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_N)$.
- Our decoding algorithms combine various ideas from coding theory, statistical learning theory and theory of estimation.
- I will discuss only the gist of the main underlying ideas (in the interest of time).

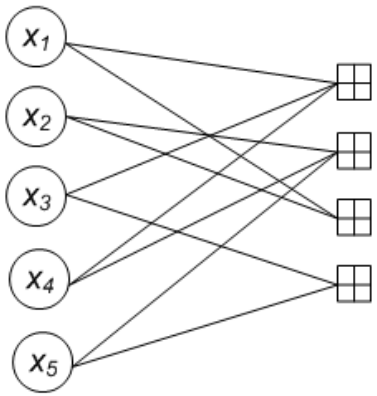
Graphical Representation

- The **second component** is that of graphical representation. It is natural to represent LDFs using bipartite graphs.
- We create a bipartite graph with two set of nodes V (variables nodes) and C (check nodes). Thus we have $|V| = M$ and $|C| = N$.
- Node j in V will be connected to node i in C if and only if the $(i,j)^{\text{th}}$ element of $\mathbf{F}$ is non-zero. For an LDF, this leads to a sparse bipartite graph.
- We denote check nodes by $\boxplus$.
- Nodes of V correspond to the coordinates of $\mathbf{x}$.
- The j^{th} parity check node has the property that the variable nodes connected to it sum to y_j.

Graphical Representation of LDFs

- A simple example is depicted in the following

$$\mathbf{F} = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix},$$



Inference on LDF graphs

- The standard decoding algorithm for graphical models (e.g. based on codes) is the message passing sum-product algorithm (SPA).
- Given a set of observations, this algorithm can be used to approximate the posterior marginal distributions.
- The main difficulty in using the SPA in compressive sensing setting is that the variables of interest are continuous.
- The **third component** of our construction is a way to circumvent these difficulties, by using **Gaussian Scale Mixtures priors with Jeffreys' non-informative hyperprior**.
- We use these priors to devise the Sum Product with Expectation Maximization (SuPrEM) algorithm.

Gaussian SPA

- Any Gaussian pdf $\mathcal{N}(x|a,A)$ can be determined by its mean a and variance A, these constitute the messages in this setting (we pass Gaussians).
- At the check nodes, we have

$$\mathcal{N}(x|a_1,A_1) * \mathcal{N}(x|a_2,A_2) \propto \mathcal{N}(x|a_1+a_2,A_1+A_2),$$

and at the variable nodes we have

$$\mathcal{N}(x|a_1,A_1) \cdot \mathcal{N}(x|a_2,A_2) \propto \mathcal{N}(x|b,B),$$

where $\propto$ denotes normalization up to a constant, and

$$B = (A_1^{-1} + A_2^{-1})^{-1},$$

$$b = B(A_1^{-1}a_1 + A_2^{-1}a_2).$$

- All the underlying operations for SPA preserve the Gaussian structure.

Sparsity Enhancing Priors

- However the Gaussian pdf is not “sparsity-enhancing”.
- Some authors propose the use of the Laplacian prior

$$p(\mathbf{x}) = \prod_i p_{x_i}(x_i) = \prod_i \frac{\lambda}{2} \exp(-\lambda |x_i|).$$

- Clearly with this prior and for Gaussian noise $\mathbf{n}$

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}) \propto \exp\left(- ||\mathbf{y} - \mathbf{A}\mathbf{x}||_2^2 - \lambda' ||\mathbf{x}||_1 \right),$$

and maximization of $p(\mathbf{x}|\mathbf{y})$ is equivalent to minimizing

$$||\mathbf{y} - \mathbf{A}\mathbf{x}||_2^2 + \lambda' ||\mathbf{x}||_1,$$

which is the objective function for the L_1 based algorithms.

Gaussian Scale Mixtures

- The **third component** of our construction is to consider the family of Gaussian Scale Mixtures (GSM) densities given by

$$x = \sqrt{\beta} u,$$

where u is a zero-mean Gaussian and $\sqrt{\beta}$ is a positive scalar random variable.

- Hence

$$p_{x|\beta}(x|\beta) \sim \mathcal{N}(x|0, \beta),$$

and

$$p_x(x) = \int_0^\infty p_{x|\beta}(x|\beta) p_\beta(\beta) d\beta.$$

- This family of densities are symmetric, zero-mean and have heavier tails than a Gaussian.
- Successfully employed in image processing and learning theory.

Jeffreys' Prior

- We now specify a pdf for $p_\beta(\beta)$, given by

$$p_\beta(\beta) \propto \sqrt{\det(I(\beta))}, \quad I(\beta) = \mathbb{E}\left(- \frac{\partial^2 \log p_{x|\beta}(x|\beta)}{\partial \beta^2} \middle| \beta \right)$$

  where $I(\beta)$ is the Fisher information matrix.
- This is referred to as the Jeffreys' prior, which can be shown to be a scalar invariant prior suitable for sparse estimation.
- In our model, the prior is given by

$$p_{\beta_i}(\beta_i) = \frac{1}{\beta_i},$$

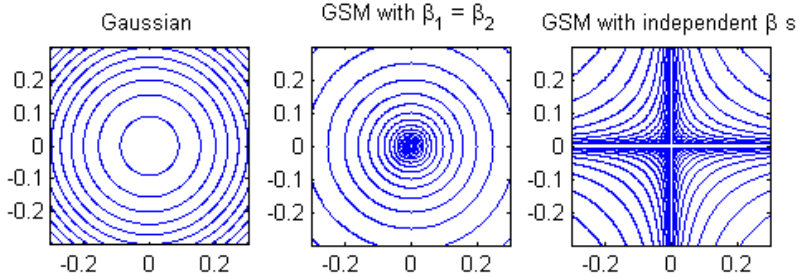
  which has no parameters to optimize.

Jeffreys' Prior

- This is an improper density, i.e. it cannot be normalized. However, in Bayesian statistics, only the relative weight of the prior determines the a-posteriori density.
- Has a singularity at the origin. This fact is usually ignored as long as it does not create computational problems. As an alternative one might set the prior to 0 in a small interval $\beta \in [0, \beta_{min})$.
- With this choice for $p_{\beta_i}(\beta_i)$, $p_{x_i}(x_i) \propto 1/|x_i|$, which is a very heavy-tailed density.

Jeffreys' Prior

- To enhance sparsity in each coordinate, it is important to have independent β_i for all i.



- In our model, we will assume that

$$p(\mathbf{x}, \boldsymbol{\beta}) = \prod_{i=1}^{M} p(x_i|\beta_i) \prod_{i=1}^{M} p(\beta_i)$$

in order to enhance sparsity in all coordinates.

Our Algorithm (SuPrEM)

- Start with an initial guess for β and zero means for the variable nodes.
- At each check node operate as in the sum-product algorithm.
- Start with an initial guess for β.
- At each variable node, first update the estimate for β (using one update of EM algorithm) and then estimate $\mathbf{x}$.
- Pass the updated estimates and continue iteration until convergence is achieved.

Simulation Setup

- In our simulations we used LDFs with parameters $(3, 6)$, $(3, 12)$ and $(3, 24)$ for $M/N = 2, 4, 8$ and $M = 10000$.
- Simulations will be presented for SNR$= 12, 24, 36$ dB, as well as the noiseless case.
- For various choices of L and SNR, we ran 1000 Monte-Carlo simulations for each value, where $\mathbf{x}$ is generated as a signal with L non-zero elements that are picked from a Gaussian distribution.
- The support of $\mathbf{x}$ is picked uniformly at random.
- Once $\mathbf{x}$ is generated, it is normalized such that $||\mathbf{Fx}||_2 = \sqrt{N}$. Thus SNR$= 10 \log_{10} \frac{1}{\sigma^2}$.

Performance Criteria

- Let $\mathcal{G}$ be the genie decoder that has full information about $\text{supp}(\mathbf{x}) = \{i : x_i \neq 0\}$. Let the output of this decoder be $\hat{\mathbf{x}}_{genie} = \mathcal{G}(\mathbf{r}) = \mathbf{F}^{\dagger}_{\text{supp}(\mathbf{x})}\mathbf{r}$.
- We define the following genie distortion measure:

$$\bar{d}_g(\mathbf{x}, \hat{\mathbf{x}}_{genie}) = \frac{||\mathbf{x} - \hat{\mathbf{x}}_{genie}||_2^2}{||\mathbf{x}||_2^2}.$$

- For any other recovery algorithm that outputs an estimate $\hat{\mathbf{x}}$, we let

$$\bar{d}_e(\mathbf{x}, \hat{\mathbf{x}}_e) = \frac{||\mathbf{x} - \hat{\mathbf{x}}_e||_2^2}{||\mathbf{x}||_2^2},$$

where the subscript e denotes the estimation procedure.

Performance Criteria

- We define

$$\mathcal{D}_{e/g}(\mathbf{x}, \hat{\mathbf{x}}_e, \hat{\mathbf{x}}_{genie}) = \frac{||\mathbf{x} - \hat{\mathbf{x}}_e||_2^2}{||\mathbf{x} - \hat{\mathbf{x}}_{genie}||_2^2} = \frac{\bar{d}_e(\mathbf{x}, \hat{\mathbf{x}}_e)}{\bar{d}_g(\mathbf{x}, \hat{\mathbf{x}}_{genie})}.$$

- We will be interested in this quantity averaged over K Monte-Carlo simulations, and converted to dB. The closer this quantity is to 0 dB means the closer the performance of the estimation procedure is to the performance of the genie decoder.

Performance Criteria

- For the noiseless case, we will be interested in the empirical probability of recovery. For K Monte-Carlo simulations, this is given by

$$P_{rec} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}(\mathbf{x} \sim \hat{\mathbf{x}}_e),$$

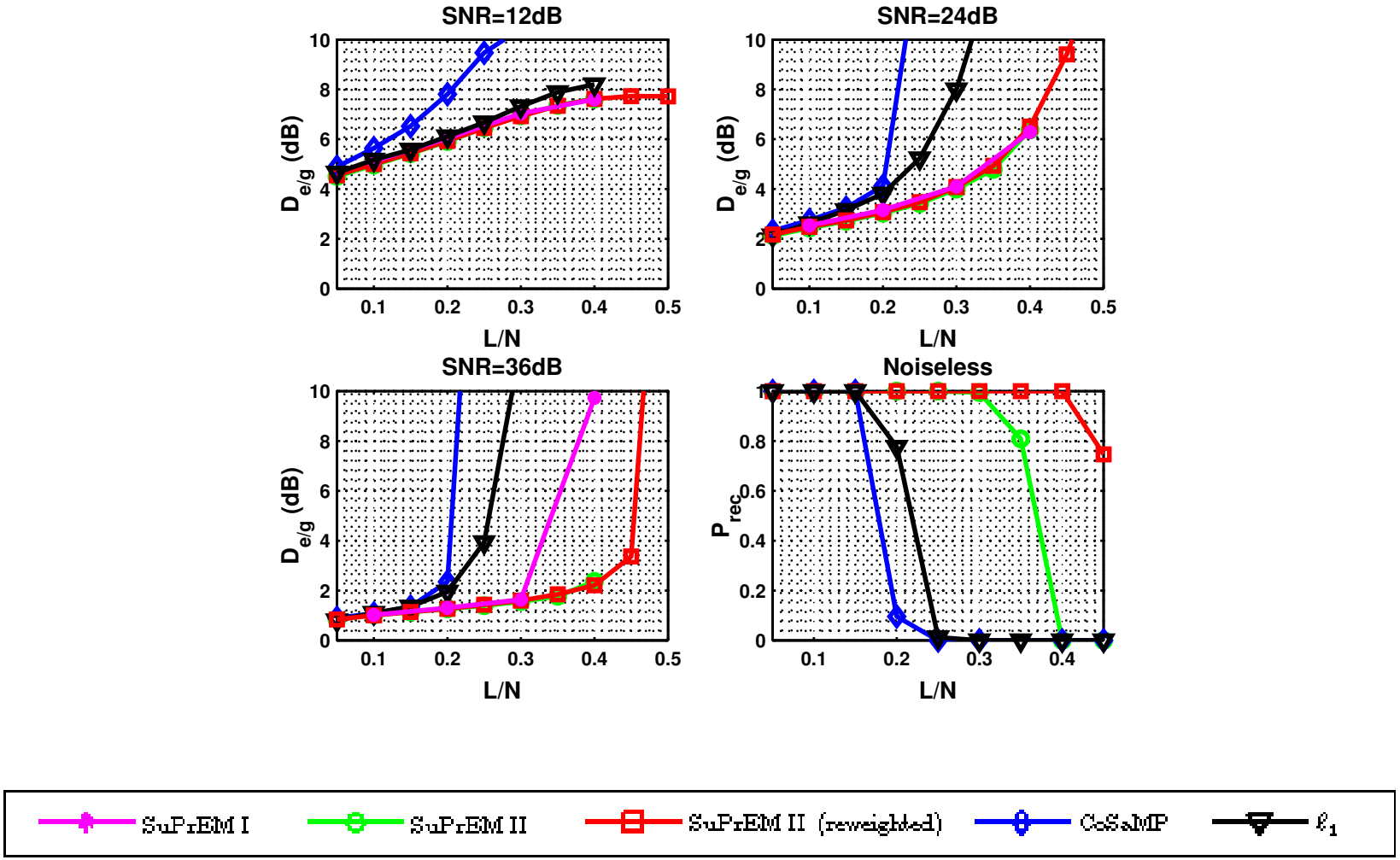
  where $\mathbb{I}(\cdot)$ is the indicator function for $(\cdot)$ (1 if $(\cdot)$ is true, 0 otherwise).
- For SuPrEM algorithms, we define the relation $\mathbf{x} \sim \hat{\mathbf{x}}_e$ to be true only if $\text{supp}(\mathbf{x}) = \text{supp}(\hat{\mathbf{x}}_e)$.
- For CoSaMP and ℓ_1, we use a milder condition, where we define $\mathbf{x} \sim \hat{\mathbf{x}}_e$ to be true if $\bar{d}_e(\mathbf{x}, \hat{\mathbf{x}}_e) < 10^{-6}$. We use this condition, since these algorithms tend to miss a small portion of $\text{supp}(\mathbf{x})$ containing elements of small magnitude.

Simulation Results

- We include results for CoSaMP and ℓ_1 based methods. For these algorithms we used partial Fourier matrices as measurement matrices.
- For ℓ_1 based methods, we used the `L1MAGIC` package in the noiseless case. In the noisy case, we used the `GPSR` package (with continuation and debiasing).
- In the implementation of GPSR we fine-tune the value of τ and observe that $\tau = 0.001||\mathbf{F}^T\mathbf{r}||_\infty$ gives the best performance.
- Since the outputs of ℓ_1 based methods and SuPrEM I are not sparse, we threshold $\mathbf{x}$ to its L largest coefficients and postulate these are the locations of the sparse coefficients.
- For all algorithms we compute $\hat{\mathbf{x}}_{\text{final}} = \mathbf{F}^{\dagger}_{\text{supp}(\hat{\mathbf{x}})}\mathbf{r}$, where $\hat{\mathbf{x}}$ is either sparse with L coefficients or has been thresholded to the largest L coefficients.

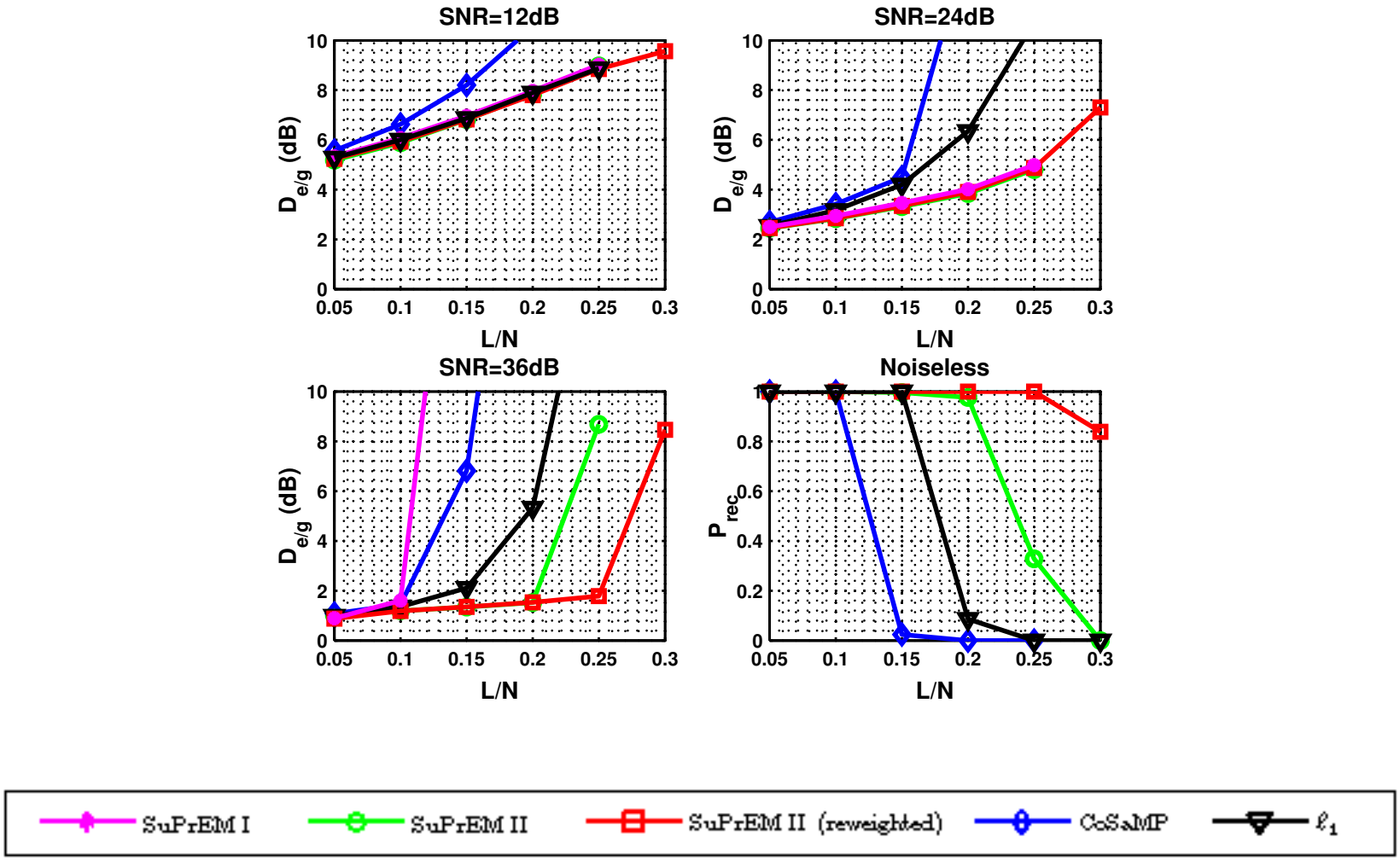
Simulation Results

- M = 10000, N = 5000



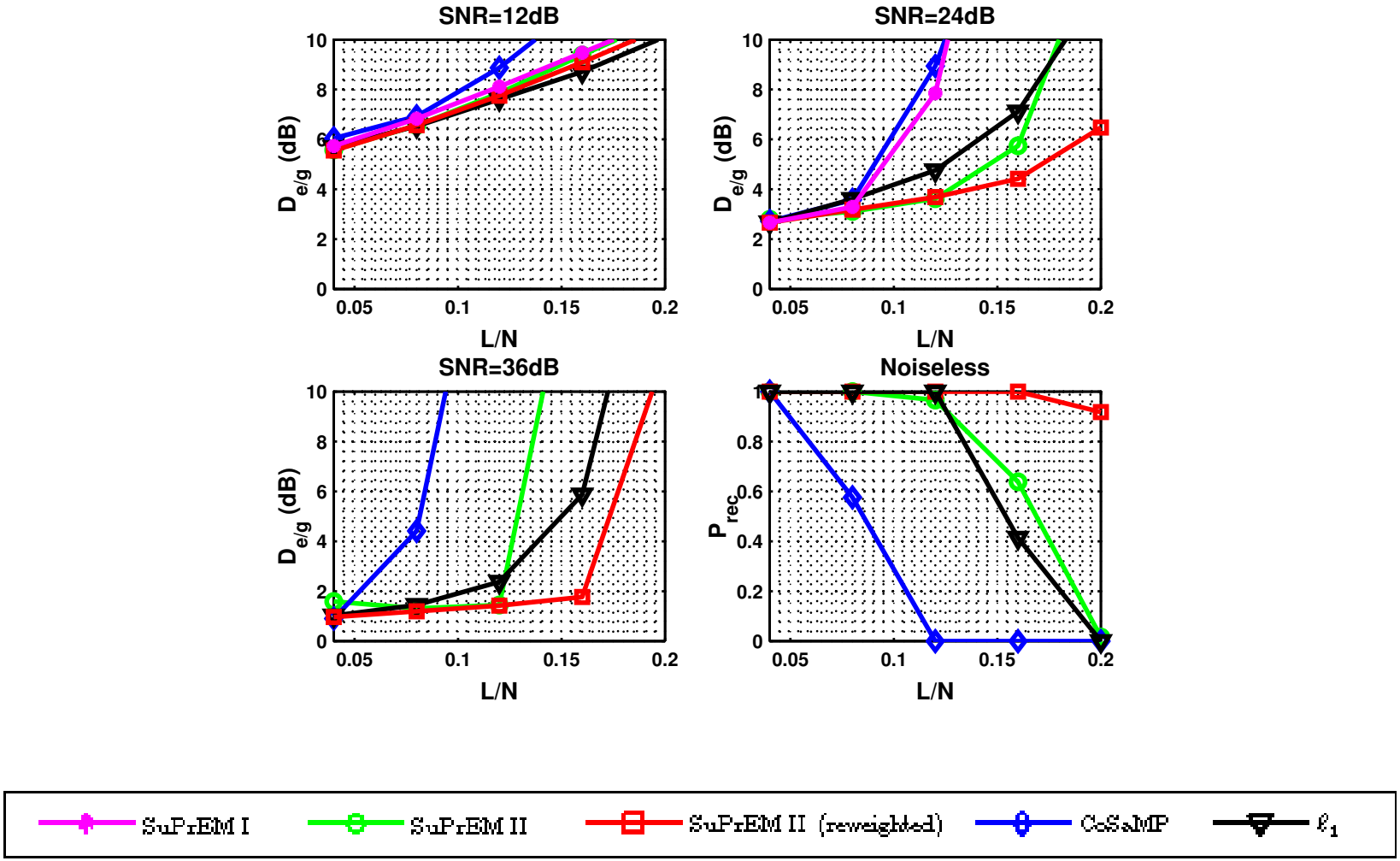
Simulation Results

- $M = 10000$, $N = 2500$



Simulation Results

- M = 10000, N = 1250



Discussion of Results

- Results indicate that the SuPrEM algorithms outperform the other state-of-the-art algorithms.
- In the low SNR regime (SNR = 12 dB), SuPrEM algorithms and the ℓ_1 methods have similar performance.
- In moderate and high SNR regimes, SuPrEM algorithms significantly outperform the other algorithms both in terms of distortion and the maximum sparsity they can work at.
- The reweighted SuPrEM II algorithm outperforms the regular SuPrEM II algorithm, even though the maximum number of iterations are the same.

Discussion of Results

- For the noiseless problem, the SuPrEM algorithms can recover signals that have a higher number of non-zero elements (even though the success condition is much milder for the other methods).
- In this case, the reweighted algorithm performs the best, and converges faster.
- We also note that for both partial Fourier matrices and LDFs, the quantity $\bar{d}_g(\mathbf{x}, \hat{\mathbf{x}}_{genie})$ is almost the same for a fixed L and SNR. i.e. $\mathcal{D}_{e/g}(\mathbf{x}, \hat{\mathbf{x}}_e, \hat{\mathbf{x}}_{genie})$ provides an objective performance criterion.

Natural Images

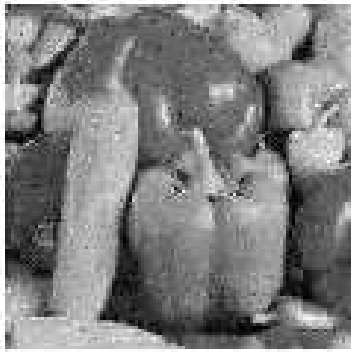
- For the testing of compressible signals, instead of using artificially generated signals, we used real-world compressible signals.
- We compressively sampled the db2 wavelet coefficients of the 256×256 (raw) peppers image using $N = 17000$ measurements.
- The PSNR values for various recovery methods are as follows: 23.41 dB for SuPrEM II, 24.79 dB for SuPrEM II (reweighted), 20.18 dB for CoSaMP, 21.62 dB for ℓ_1, 23.61 dB for LASSO.

Natural Images

Original

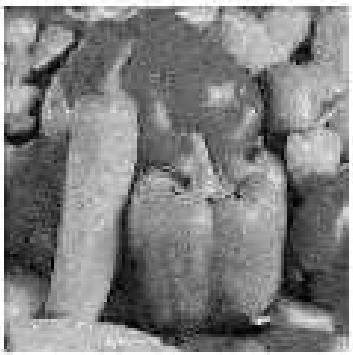
SuPrEM II (reweighted)

SuPrEM II

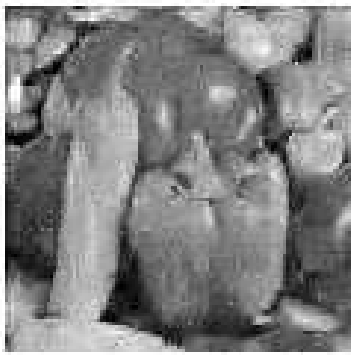


CoSaMP

LASSO (GPSR)



ℓ_1 with Equality Constraints



Conclusions

- We considered the number of measurements required for successfully recovery for compressed sensing in the presence of noise.
- For the linear regime we found $N = O(L)$ is necessary and sufficient. This is an improvement over ℓ_1 regularization methods which require $N = O(L \log(M - L))$.
- We showed a Joint Typicality Estimator asymptotically achieves the Cramér-Rao bound on the mean squared error of the Genie-Aided Estimator without any knowledge of the locations of nonzero elements of $\mathbf{x}$, as $N \to \infty$ for $\alpha = L/N$ a fixed number when $||\mathbf{x}||_2$ grows at a specified rate.

Conclusions

- We considered the sparsity constrained approximation (SCA) problem and derived a lower bound on the average distortion for a given N, as well as an asymptotic lower bound.
- We constructed Vandermonde frames and a simple algorithm that outputs the most compact representation of a given vector, as long as $\epsilon \leq 0.5$ (the sparsity factor). This is irrespective of r (the redundancy of the frame).
- We discussed the recovery algorithms from current literature, as well as deterministic constructions, for noisy compressive sampling.

Conclusions

- We constructed an ensemble of measurement matrices with small storage requirements, which we denoted low density frames.
- For these frames, we provided sparse reconstruction algorithms that have $O(M)$ complexity and that are Bayesian in nature.
- We observed that in various cases of interest, SuPrEM algorithms with LDFs outperformed the other state-of-the-art recovery algorithms.
- For Gaussian sparse signals and Gaussian noise, we are within 2 dB range of the theoretical lower bound in most cases.