Uncertainty Principles and Sparse Signal Representations Using Overcomplete Representations

Todd P. Coleman

October 2, 2002

This discussion sparse representations of signals in $\mathbb{R}^n$. The sparsity of a signal is quantified by the number of nonzero components in its representation. Such representations of signals are useful in signal processing, lossy source coding, image processing, etc. We first speak of an uncertainty principle regarding the sparsity of any two different orthonormal basis representations of a signal $\underline{S}$. Next, describing a signal as an overcomplete description involving a pair of orthonormal bases is considered. Because this is an overcomplete description, many possible representations exist. The hope is that the most sparse representation is a lot better than any representation using a single orthonormal basis. The uncertainty principle can be exploited to provide conditions upon when the most sparse overcomple description is unique. Next, performing the optimization search is considered, which in general is nonconvex and combinatorial. However, it is illustrated that if the most sparse representation is unique and is sufficiently sparse, it can be found using a linear programming formulation, which is considerably more computationally affordable. The notion of 'sufficiently sparse' depends upon the pair of orthonormal bases, and in particular their mutual incoherence. We then explore typical mutual incoherence between pairs of bases, and discuss some (idealized) applications.

We consider a signal $S \in R^N$ of unit l_2 energy. We are provided two different orthonormal bases A and B, described in terms of matrices where each column is one of the orthonormal vectors, i.e. $A = [\underline{a}_1\underline{a}_2...\underline{a}_N]$ and $B = [\underline{b}_1\underline{b}_2...\underline{b}_N]$. We note that for each of the bases individually, S is uniquely described in terms of linear combinations of the orthonormal vectors:

$$\begin{aligned} \underline{S} &= A\underline{\alpha} \\ &= \sum_{i=1}^{N} \alpha_i \underline{a}_i \\ \underline{S} &= B\underline{\beta} \\ &= \sum_{i=1}^{N} \beta_i \underline{b}_i \\ \alpha_i &= < \underline{S}, \underline{a}_i > \\ \beta_i &= < \underline{S}, \underline{b}_i > \end{aligned}$$

We are interested in the coefficient vectors $\underline{\alpha}$, $\underline{\alpha}$, and in particular the number of nonzero components. This is described quantitatively by defining the l_0 quasi-norm of a vector $x \in \mathbb{R}^N$ to be $\|x\|_0 = \sum_{i=1}^{N} 1_{\{x_i \neq 0\}}$. There is an uncertainty principle which illustrates that a signal cannot have multiple sparse representations in any two distinct bases. What is meant by sparse above actually depends upon the pairs of bases through their 'mutual incoherence' [DH01], $M = \max_{i,j} | < \underline{a}_i, \underline{b}_j > |$:

Theorem 1:
Given a signal $\underline{S} \in R^N$ and two orthonormal bases A and B , where $\underline{S} = A\underline{\alpha} = B\underline{\beta}$, the following bound is always satisfied:

$$\frac{\|\underline{\alpha}\|_0 + \|\underline{\beta}\|_0}{2} \geq \sqrt{\|\underline{\alpha}\|_0 \cdot \|\underline{\beta}\|_0} \geq \frac{1}{M}$$

Proof: Define $\|\underline{\alpha}\|_0 = l_A$ and $\|\underline{\beta}\|_0 = l_B$.

$$
\begin{aligned}
1 &= \underline{S}^T \underline{S} \\
&= [\alpha_1 ... \alpha_N] \begin{bmatrix} \underline{a}_1^T \underline{b}_1 & \underline{a}_1^T \underline{b}_2 & \cdots & \underline{a}_1^T \underline{b}_N \\ \underline{a}_2^T \underline{b}_1 & \underline{a}_2^T \underline{b}_2 & \cdots & \underline{a}_1^T \underline{b}_N \\ \vdots & \vdots & & \vdots \\ \underline{a}_N^T \underline{b}_1 & \underline{a}_N^T \underline{b}_2 & \cdots & \underline{a}_1^T \underline{b}_N \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_N \end{bmatrix} \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i < \underline{a}_i, \underline{b}_j > \beta_j \\
&= \sum_{i'=1}^{l_A} \sum_{j'=1}^{l_B} \alpha_i < \underline{a}_{i'}, \underline{b}_{j'} > \beta_{j'} \\
&\leq \sum_{i'=1}^{l_A} \sum_{j'=1}^{l_B} |\alpha_i|| < \underline{a}_{i'}, \underline{b}_{j'} > ||\beta_{j'}| \\
&\leq M \left(\sum_{i'=1}^{l_A} |\alpha_{i'}| \right) \left(\sum_{j'=1}^{l_B} \beta_{j'} \right) \\
&\leq M \sqrt{l_A l_B}
\end{aligned}
$$

where the sum up to $l_A (l_B)$ is only over the nonzero components of $\underline{\alpha}$ $(\underline{\beta})$, and the last inequality holds by noting that

$$
\begin{aligned}
\max \quad & \sum_{i=1}^{K} x_i \\
s.t. \quad & \sum_{i=1}^{K} x_i^2 = 1 \\
& x_i > 0
\end{aligned}
$$

has a solution of $\sqrt{K}$. The sum of squares to unity constraint is due to Parseval's relation and the assumption that the signal $\underline{S}$ has unit l_2 energy. Thus by using the geometric mean and arithmetic mean inequality, we arrive at

$$\frac{l_A + l_B}{2} \geq (l_A l_B) \geq \frac{1}{M}.$$

Note that in general $\frac{1}{\sqrt{N}} \leq M \leq 1$. The lower bound is so because we note that $A^T B$ is orthonormal with the sum of all of its entries N. The lower bound is met w/ equality, for instance, with the $\{\underline{a}_i = e_i\}$ and $\{\underline{b}_i = \frac{1}{\sqrt{N}} e^{\frac{j2\pi ki}{N}}\}$, i.e. the spike basis and inverse DFT basis. The upper bound is met with equality whenever two bases share a common vector. As we can see, the bound depends heavily on M.

1 Uniqueness of an Overcomplete Sparse Representation

As a consequence of the sparsity principle just stated, it can be shown that for a signal $\underline{S}$ represented as an overcomplete description, i.e., in terms of the concatenation of two bases A and B, if a very sparse representation exists, it is unique.

Theorem 2:
Suppose we represent $\underline{S}$ in an overcomplete representation, i.e.

$$\begin{aligned} \underline{S} &= A\underline{v}^A + B\underline{v}^B \\ &= [A\ B] \begin{bmatrix} \underline{v}^A \\ \underline{v}^B \end{bmatrix} \\ &= [A\ B]\,\underline{v} \end{aligned}$$

Then if we consider any two different representations, $\underline{v}_1, \underline{v}_2 \in \mathbb{R}^{2N}$, then we have

$$\|\underline{v}_1\|_0 + \|\underline{v}_2\|_0 \geq \frac{2}{M}.$$

Thus, any representation $\underline{v}$ is guaranteed to be uniquely sparse if

$$\|\underline{v}\|_0 < \frac{1}{M}.$$

Proof:
Let $\underline{v}_1$, $\underline{v}_2 \in \mathbb{R}^{2N}$ be two different coefficient representations of $\underline{S}$, i.e.

$$\underline{S} = [A\ B]\underline{v}_1 = [A\ B]\underline{v}_2$$

Note that $x = \underline{v}_1 - \underline{v}_2$ lies in the nullspace of $[A\ B]$, i.e.:

$$\begin{aligned} 0 &= [AB](\underline{v}_1 - \underline{v}_2) \\ &= [A\ B]x \\ &= Ax^A + Bx^B \\ Ax^A &= -Bx^B = W \end{aligned}$$

where $W \in \mathbb{R}^N$ is nonzero. Exploiting the uncertainty principle, we have that

$\underline{S} = A\underline{v}^A + B\underline{v}^B$, and

$$\underline{S} = [A\ B]\,\underline{v}$$

Suppose $\|x^A\|_0 = l_A$ and $\|x^B\|_0 = l_B$. Then from the uncertainty principle,

$$\frac{l_A + l_B}{2} \geq (l_A l_B) \geq \frac{1}{M}.$$

If both the original overcomplete representations were sparse, i.e., $\underline{v}_1 < \frac{1}{M}$ and $\underline{v}_2 < \frac{1}{M}$, then

$$\begin{aligned} l_A + l_B &= \|x^A\|_0 + \|x^B\|_0 \\ &= \|x\|_0 \\ &= \|\underline{v}_1 - \underline{v}_2\|_0 \\ &< \|\underline{v}_1\|_0 + \|\underline{v}_2\|_0 \\ &< \frac{2}{M} \end{aligned}$$

But that would contradict the uncertainty principle. Thus we arrive at the conclusion.

2 Optimization techniques for finding Sparse Overcomplete Representations

Let us denote the sparsest representation of the signal $\underline{S}$ using orthonormal bases A and B as $\underline{v}$. Attempting to search for $\underline{v}$ turns out to be very difficult in general. Such an optimization formulation would be of the form P_0:

$$\begin{aligned} \min \quad & \|\underline{v}\|_0 = \sum_{i=1}^{2N} 1_{\{v_i \neq 0\}} \\ s.t. \quad & \underline{S} = [A\ B]\underline{v} \end{aligned}$$

This is a difficult combinatorial optimization problem in general, because $f(x) = \|\underline{v}\|_0$ is nonconvex. For instance, consider

$$\begin{aligned} x_1 &= (1,0,0,0), f(x_1) = 1 \\ x_2 &= (0,0,0,1), f(x_2) = 1 \\ \text{but} \quad & f(\lambda x_1 + (1-\lambda)x_2) = 2 > \lambda f(x_1) + (1-\lambda) f(x_2) = 1. \end{aligned}$$

Instead, one might consider the l_1 optimization problem (P_1), of the form

$$\begin{array}{ll} \min & f_1(\underline{v}) = \|\underline{v}\|_1 = \sum_{i=1}^{N} |v_i| \\ s.t. & \underline{S} = [A\ B]\underline{v} \end{array}$$

It turns out that, the sparsity condition

$$\|\underline{v}\|_0 < F(M) = \frac{\sqrt{2} - 0.5}{M}$$

guarantees that (P_1) returns the same optimal solution as (P_0). Note that (P_1) has a piece-wise linear convex objective function, and it can be solved using linear programming, which is very efficient.

The technique is as follows. It must be shown that for all $\tilde{\underline{v}}$ such that $[A\ B]\tilde{\underline{v}} = S$,

$$\|\underline{v}\|_0 < F(M) \Rightarrow \|\underline{v}\|_1 < \|\tilde{\underline{v}}\|_1.$$

This is shown in steps. First, by observing the difference vector $x = \underline{v} - \tilde{\underline{v}}$, we have

$$\begin{aligned} [A\ B]x &= 0 \\ Ax^A = -Bx^B & \end{aligned}$$

And the optimality condition becomes

$$\begin{aligned} \sum_{i=1}^{N} |v_k + x_k| - \sum_{i=1}^{N} |v_k| &\geq 0 \\ \text{i.e.} \sum_{\text{off}(v)} |x_k| + \sum_{\text{on}(v)} (|v_k + x_k| - |v_k|) &\geq 0 \end{aligned}$$

where $\text{off}(v) = \{i : v_i = 0\}$ and $\text{on}(v) = \{i : v_i \neq 0\}$. Next, we note that due to the relation $|v + m| \geq |v| - |m|$, we have that

$$\sum_{\text{off}(v)} |x_k| - \sum_{\text{on}(v)} |x_k| \geq 0 \Rightarrow \sum_{\text{off}(v)} |x_k| + \sum_{\text{on}(v)} (|v_k + x_k| - |v_k|) \geq 0$$

Thus, we can now focus our attention to guaranteeing that

$$\sum_{\text{off}(v)} |x_k| - \sum_{\text{on}(v)} |x_k| \geq 0 \Leftrightarrow \frac{1}{2}\sum_{i=1}^{2N} |x_k| - \sum_{\text{on}(v)} |x_k| \geq 0.$$

We could possibly consider focusing on the problem:

$$\begin{aligned} \min \quad & f_2(x) = \frac{1}{2}\sum_{i=1}^{2N} |x_k| - \sum_{\text{on}(v)} |x_k| \\ s.t. \quad & Ax^A = -Bx^B \end{aligned}$$

and consider when the result is greater than 0. Note that as $\|\underline{v}\|_0$ increases, the objective function decreases in value, and eventually becomes negative. We are interested in finding the largest value of $\|\underline{v}\|_0$ such that the result of the above optimization is still positive. But we note that the zero vector is feasible. However, since we are only interested in the sign of the result, we could impose the extra linear constraint $\sum_{i=1}^{2N} |x_k| = 1$ to avoid the zero vector being a feasible solution. Doing so replaces the first sum in the above objective function with 1. This formulation still depends crucially on the two matrices and on the positions of the nonzero components, however. So we consider minimizing the same objective function over a set of weaker constraints:

$$\begin{aligned} \min \quad & f_2(x) = \frac{1}{2} - \underline{1}_{v1}^T \underline{X}_1 - \underline{1}_{v2}^T \underline{X}_2 \\ s.t. \quad & \underline{X}_1 \leq M 1_{N\times N} \underline{X}_2 \\ & \underline{X}_2 \leq M 1_{N\times N} \underline{X}_1 \\ & \underline{1}^T(\underline{X}_1 + \underline{X}_2) = 1 \\ & \underline{X}_1 \geq 0, \underline{X}_2 \geq 0 \end{aligned}$$

Where $\underline{X}_1$ ($\underline{X}_2$) is the vector of absolute values of x^A (x^B). The first two inequality constraints arise by noting that they are satisfied by any feasible solution of the previous constraint set, because $x^A = -A^T B x^B$ and the definition of M. Next, we can exploit the symmetric nature of the above problem, and note that the locations of the nonzero components of v_1,v_2 are not required, only how many of them is. This is a linear programming problem of the form (P):

$$\min \ \underline{C}^T \underline{X} \ \text{ s.t. } A\underline{X} \geq \underline{B}, \ \underline{X} \geq 0.$$

Since we are interested in ensuring that the optimal solution to the above problem is positive, we may exploit weak duality in linear programming [BT97], which states that for the LP dual problem (D),

$$\max\ \underline{P}^T \underline{B} \text{ s.t. } A^T \underline{P} \leq \underline{C},\ \underline{P} \geq 0,$$

for any feasible solution $\underline{X}$ to (P) and any feasible solution $\underline{P}$ to (D), we have that

$$\underline{P}^T \underline{B} \leq \underline{C}^T \underline{X}.$$

Thus, if we impose conditions on $\underline{v}$ so that some feasible solution of the dual problem is greater than 0, then we imply that the optimal solution to the primal problem is as well. This then implies, through our chain of implications, that the original l_1 linear optimization will return $\underline{v}$ as its solution. It is shown by selecting a feasible solution of (D) parametrically of the form

$$\underline{U}^T = \left[\underline{1}_{v1}^T\ \alpha\ \underline{1}_{v2}^T\ \beta\ \delta\right]$$

with equation manipulation that if $K_1(K_2)$ is the number of nonzero components in v_1 (v_2), then the inequality

$$K_1 + K_2 = \|\underline{v}\|_0 \leq \frac{\sqrt{2} - 0.5}{M}$$

guarantees a feasible solution of the parametric form described above and furthermore that its cost is greater than 0. This then implies through the chain of implications that the original l_1 LP will find $\underline{v}$. Thus, the results concluded in [EB02] are that if the sparsest overcomplete representation representation of the signal s, given by $\underline{v}$, satisfies $\|\underline{v}\|_0 < \frac{1}{M}$, then it is unique. And furthermore, if $\|\underline{v}\|_0 < \frac{\sqrt{2}-0.5}{M} = \frac{0.9142}{M}$, then it can be found using linear programming methods.

3 Typical Mutual Incoherence

We might ask the question about mutual incoherence between pairs of orthonormal matrices in general. We can consider this by picking pairs of random orthonormal matrices, and observing what M is typically. Suppose we generate two matrices independently from the uniform distribution on the group of all orthnormal matrices $O(N)$. For large N, it turns out that they in fact are typically quite incoherent: M is typically never greater than $2\sqrt{\frac{\log N}{N}}$. Finding what M is typically is the same as finding what largest magnitude entry in a random orthogonal matrix typically (orient yourself in the coordinate basis direction of one of the two). It turns out that

$$P\left(\max_{i,j}|U_{i,j}| > 2(1+\epsilon)\sqrt{\frac{\log N}{N}}\right)$$

tends to 0 as N tends to ∞. Any entry $U_{i,j}$ is the projection of a randomly chosen point on the N-sphere, to one coordinate. Such a random point is typically near the equator, and $U_{i,j}$ behaves like a $\mathcal{N}(0,\frac{1}{N})$ random variable.

$$\begin{aligned}
P\left(|U_{i,j}| > \frac{x}{\sqrt{N-2}}\right) &\leq 2e^{-\frac{x^2}{2}} \\
P\left(\text{any } |U_{i,j}| > \frac{x}{\sqrt{N}}\right) &\leq \sum_{i,j} P\left(|U_{i,j}| > \frac{x}{\sqrt{N}}\right) \\
&\leq 2N^2\exp\left(-\frac{N-2}{N}\frac{x^2}{2}\right).
\end{aligned}$$

By letting $x = (1+\epsilon)\sqrt{\log N}$, the resulting probability tends to 0 with N. Thus, typically, pairs of orthonormal bases of $\mathbb{R}^N$ have mutual incoherence on the order of $\sqrt{\frac{\log N}{N}}$, which decays in N.

4 Interesting Idealized Applications

Due to the fact that linear programming can be executed in polynomial time, along with the guarantee that for certain ranges of M, the LP optimization is guaranteed to give the sparsest representation, interesting idealized applications have been mentioned [DH01].

4.1 Error-Correcting Encryption

Wyner introduced some methods of encrypting signals in $\mathbb{R}^N$ by using random orthogonal matrices. Take the signal of interest $\underline{S} \in \mathbb{R}^N$, along with an $N \times N$ matrix U, and perform the encryption $\underline{E} = U\underline{S}$. It is considered an encryption scheme because an observer that does not know U only knows the marginal distribution of $\underline{E}$ is uniform on the sphere of radius $\|S\|$. So $\underline{E}$ may be transmitted and the receiver, who knows U, decrypts it by $\underline{S} = U^T \underline{E}$.

From before we know that the LP minimization formulation with an over-complete dictionary is guaranteed to find the sparsest representation provided it has less than $\frac{\sqrt{2}-0.5}{M}$ nonzero entries. Now suppose we form our signal $\underline{S}$ and it has $K < \frac{\sqrt{2}-0.5}{2M}$ nonzero entries. Suppose the received vector is $\underline{\tilde{E}} = \underline{E} + \underline{Z}$, where Z has K or less nonzero entries. Then note that $\underline{\tilde{E}} = U\underline{S} + I\underline{Z}$, i.e., it is a superposition of K terms from the U dictionary and K or less terms from the $\{\underline{e}_i\}$ dictionary. Since $2K < \frac{\sqrt{2}-0.5}{M}$, the LP formulation will recover perfectly the columns of U where the transmitted data occurs, along with the locations of of the nonzero entries of Z. It may recover precisely the original signal S, regardless of the size of z in the K components that are nonzero.

4.2 Idealized (Noiseless) Uncoordinated Source Separation

Suppose we have two uncoordinated users that would like to send messages to an ideal receiver (with no noise). The receiver sees

$$\underline{R} = \underline{E}_1 + \underline{E}_2.$$

Suppose both users are provided random orthogonal matrices, U_1, U_2, unknown to each other. The only thing they know is their mutual incoherence, M. Each user encodes a signal $\underline{S}_i \in \mathbb{R}^N$ that has at most $K < \frac{\sqrt{2}-0.5}{2M}$ nonzero entries by $\underline{E}_i = U_i \underline{S}_i$. Using the LP formulation, the users' two messages may be decoded perfectly. Interesting key features include:

- users do not know each others random matrices U_i, the only thing they know is their mutual incoherence, M.
- transmission is encrypted. Any receiver that observes $\underline{R}$ without U_1 and U_2 will not be able to access the components.
- the scheme works perfectly regardless of the ratio of the two users signals. All that is needed is $K < \frac{\sqrt{2}-0.5}{2M}$.

References

[BT97] D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization.* Athena Scientific, 1997.

[Coh89] L. Cohen. Time-frequency distributions - a review. *Proceedings of the IEEE*, 77(7), 1989.

[DH01] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7), November 2001.

[EB02] M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 48(9), September 2002.