Self Basis Selection in a Finite Set*

Ishai Ilani
Actelis Networks, Israel
iilani@actelis.com

Ram Zamir
Tel Aviv University, Israel
zamir@eng.tau.ac.il

Abstract

Given a set of n vectors in $\mathcal{R}^m$, $m < n$, we wish to find a subset of m vectors that are good "predictors" for the complementary set. We consider two criteria of goodness, one leads to requiring that the least-squares expansion coefficients of the complementary set be bounded by one, the other leads to maximizing the determinant of the selected subset. Exhaustive search requires checking all n choose m possible subsets. We present a low-complexity iterative selection algorithm, and examine its worst loss with respect to the optimum solution under both goodness criteria. We show that with linear complexity in n, the proposed algorithm achieves expansion coefficients which are uniformly bounded by $1+\epsilon$, while the determinant of the selected subset is at most $m^{m/2}$ below the true maximum determinant.

1 Introduction

Finding a good basis for a vector space is a classical question in harmonic analysis and frame theory [1]. A more restrictive question is to find a good basis from a given *finite* set of vectors, a topic treated in the signal processing area under names like "matching pursuit" [2] and "basis pursuit" [3].

Our work on basis selection is motivated by a question arising in multiple source coding or in multiple channel coding. In the former, we look for "context sources" for encoding a vector of correlated sources. In the latter, we look for "sensor channels" for interference cancellation over a vector of adjacent channels. Both scenarios may be either in a point-to-point or in a multipoint-to-point configuration. These two problems lead to the following question: Given a set $\mathbf{a}_1, \ldots, \mathbf{a}_n$ of n vectors in $\mathcal{R}^m$, where $m < n$, find a subset S of size m' which can serve as a good basis for the remaining $n - m'$ vectors. Hence, in this problem the goodness of the basis is evaluated relative to the *same set* from which it is selected. We shall mainly be interested in the $m' = m$ case.

To make the notion of a "good basis" concrete, consider the following model of n linearly distorted noisy measurements of a vector $\mathbf{u} = (u_1, \ldots, u_m)^T$:

$$x_i =< \mathbf{a}_i, \mathbf{u} > +z_i, \;\; i = 1, \ldots, n \tag{1}$$

where x_i is the i-th measurement, $\mathbf{a}_i = (a_{i1}, \ldots, a_{im})^T$ is the corresponding vector of linear distortion coefficients, z_i is the corresponding noise, and $< .. >$ denotes inner product. We wish to find a subset $S \subset \{1, \ldots, n\}$ of size $|S| = m'$ such that the measurements

*This research is supported in part by the Office of the Chief Scientist.

$\{x_k, k \in S\}$ are a "good context" or "good sensors" for the remaining measurements $\{x_i, i \notin S\}$.

We define subset goodness in two ways:

- Low noise amplification;
- Small residual entropy.

The former notion leads to a criterion of *small expansion coefficients* of $\{a_i, i \notin S\}$ in terms of $\{a_k, k \in S\}$. The latter notion leads, for $m' = m$, to a criterion of *maximum absolute determinant* of the matrix A_S composed of the basis vectors $\{a_k, k \in S\}$. It's worth noting that in general, the best basis of size $l+1$ is *not* an extension of the best basis of size l. See an example in the next section.

Noise Amplification

We arrive at the two criteria above from probabilistic arguments. Assume that $(u_1, \ldots, u_m)$ in (1) are i.i.d. random variables $\sim \mathcal{N}(0, \sigma_u^2)$, mutually independent of $(z_1, \ldots, z_m)$ which are i.i.d. $\sim \mathcal{N}(0, \sigma_z^2)$. Since all the random variables are jointly Gaussian, the minimum Mean Squared Error (MSE) estimate of x_i from $\{x_k, k \in S\}$ takes a linear form:

$$\hat{x}_i =< \boldsymbol{g}_i, \mathbf{x}_S > \tag{2}$$

where $\mathbf{x}_S$ is the vector of the measurements $\{x_k, k \in S\}$, and $\boldsymbol{g}_i = (g_{i1}, \ldots, g_{im'})^T$ is a vector of linear estimation coefficients. Moreover, as the "signal-to-noise" ratio (SNR)

$$\gamma^2 = \frac{\sigma_u^2}{\sigma_z^2}$$

becomes large, the optimal estimation coefficients vector $\boldsymbol{g}_i$ approaches the Least Squares (LS) solution, i.e., the expansion of the projection of $\mathbf{a}_i$ on the linear sub-space spanned by the basis vectors $\{a_k, k \in S\}$. If $m' \geq m$, and assuming the basis spans the whole space, we have

$$\mathbf{a}_i = \sum_{k \in S} g_{ik} \cdot \mathbf{a}_k = A_S \cdot \boldsymbol{g}_i \tag{3}$$

where

$$A_S = [\mathbf{a}_{i_1}; \ldots; \mathbf{a}_{i_{m'}}]$$

is the $m \times m'$ matrix whose columns are the basis vectors $\{a_k, k \in S\}$, where here $S = \{i_1, \ldots, i_{m'}\}$. Combining (1), (2) and (3), it follows that for high SNR and if $m' \geq m$, the error vector contains only noise elements, and it is given by

$$\hat{x}_i - x_i =< \boldsymbol{g}_i, \mathbf{z}_S > -z_i \tag{4}$$

where $\mathbf{z}_S$ is the vector of noises $\{z_k, k \in S\}$. Hence the resulting MSE is

$$E(\hat{x}_i - x_i)^2 = \sigma_z^2(\|\boldsymbol{g}_i\|^2 + 1). \tag{5}$$

(Note that this is the MSE of the LS solution for any signal-to-noise ratio γ, not only high, and in the general case it is an upper bound on the optimum (Baysian) MMSE.)

We see that the noise is amplified by the expansion coefficients of $\mathbf{a}_i$ relative to the basis S. We say that a basis S is an *α-amplifier* if the expansion coefficients of all the vectors outside S are absolutely bounded by α, i.e., for $i \notin S$

$$|g_{ik}| \leq \alpha \quad \forall k \in S. \tag{6}$$

Finally, we say that a basis is good in the sense of noise amplification if it is a *1-amplifier*, i.e., if $|g_{ik}| \leq 1$ for all i, and all k in S.

Residual Entropy and Basis Determinant

We turn to motivate the second criterion of basis goodness. The residual entropy of the measurements relative to a basis S is defined as the conditional differential entropy of the measurements outside S given the measurements in S,

$$h(\mathbf{x}_{S^c}|\mathbf{x}_S)$$

(see [4] for the definition of $h(.)$). This quantity determines the Shannon capacity of an n-lines vector channel, with additive noises $x_1, \ldots, x_n$, assuming lines $k \in S$ act as "sensors" (provide channel side information) for the rest of the lines. The smaller the residual entropy is, the higher is the capacity of lines $i \notin S$. This is also the minimum rate in compressing an n-component vector source $x_1, \ldots, x_n$, if sources $k \in S$ serve as "context" for encoding the rest of the sources. Now, by the chain rule for joint entropy we have, (cf. [4]):

$$h(x_1, \ldots, x_n) = h(\mathbf{x}_S) + h(\mathbf{x}_{S^c}|\mathbf{x}_S) \tag{7}$$

so minimizing $h(\mathbf{x}_{S^c}|\mathbf{x}_S)$ over S is equivalent to maximizing $h(\mathbf{x}_S)$. Furthermore, since $x_1, \ldots, x_n$ are zero-mean jointly Gaussian, $\mathbf{x}_S = A_S^T\mathbf{u} + \mathbf{z}_S$ is a Gaussian vector with covariance $\sigma_u^2 A_S^T A_S + \sigma_z^2 I$, and we have

$$e^{2h(\mathbf{x}_S)} = K \cdot \det(A_S^T A_S + \gamma^{-2} I) \tag{8}$$

where $K = (\sigma_u^2\sqrt{2\pi e})^{m'}$ is a constant. Assuming $m' = m$ and $\gamma \to \infty$, it follows that minimizing the residual entropy over the choice of S amounts to maximizing the absolute determinant of the $m \times m$ matrix A_S:

$$S^* = \arg \max_{S:\ |S|=m} |\det(A_S)|. \tag{9}$$

Since the set $\mathbf{a}_1, \ldots, \mathbf{a}_n$ is finite, this maximum is always achieved. In the sequel we assume that the set $\mathbf{a}_1, \ldots, \mathbf{a}_n$ is not degenerate (i.e., it is not contained in a strict sub-space of $\mathcal{R}^m$), so the maximum in (9) is always strictly positive.

As we shall see next, the two notions of goodness (unit noise amplification and maximum basis determinant) are closely related via Cramer's law.

Proposition 1 (Existence of a 1-Amplifier Basis) *Every* locally *optimal solution for (9) (i.e., a subset such that replacing one vector does not increase its determinant) is a 1-amplifier basis. In particular, any determinant maximizing set S^* is a 1-amplifier basis.*

Proof: This is a simple consequence of Cramer's law,

$$g_{ik} = \frac{\det(A_{S,k})}{\det(A_S)}, \tag{10}$$

where $A_{S,k}$ is obtained by replacing the k-th column of A_S by $\mathbf{a}_i$.

Note that the opposite, however, is not true; not every 1-amplifier basis achieves the global maximum in (9).

Geometrically, the determinant of A_S amounts to the product of the lengths of $\{a_k, k \in S\}$ and the sines of the angles between each vector and the linear subspace spanned by the previous vectors (in some order). Hence, a large determinant corresponds to long and close to orthogonal vectors. This partially resembles a search for a short basis for a

lattice. All bases of a given lattice have the same determinant (it is the volume of the lattice basic cell), so minimizing the vectors' lengths is equivalent to making the angles as close to 90^o as possible. See the LLL algorithm, [5], for an efficient search for a short basis for a lattice.

As mentioned above, a greedy search is in general not optimal. Thus, solving (9) requires, in principle, searching all $\binom{n}{m}$ subsets and calculating their determinants. This implies $\sim n^m$ determinant calculations (i.e. polynomial in n, exponential in m). On the other hand, a greedy solution (as in matching pursuit [2]) sequentially selects the longest residual vector in a Gram-Schmidt-like process, implying *linear* complexity in n. However, this solution only guarantees a 2^{m-1}-amplifier basis (see Section 3), and a far from optimum basis determinant.

In this work we investigate the gap in performance between the optimum solution and low complexity variations on the greedy solution above. We consider both the noise amplification and the maximum determinant basis selection criteria. The next section introduces the algorithmic approach to basis selection. Section 3 upper bounds the noise amplification achieved by greedy selection. Section 4 presents an iterative enhancement algorithm which approaches a 1-amplifier basis. Section 5 gives some pessimistic results about the complexity of determinant maximization. The last section discusses alternative approaches and goodness criteria for basis selection.

2 Basis Selection Algorithms

Before discussing the algorithmic approach we are going to take, we illustrate why "greedy" basis selection is in general sub-optimal. Consider first the following simple example. Let the vectors $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ be orthogonal, and suppose we wish to expand the vector $\mathbf{a}$ in terms of a subset of l vectors from the set

$$\{\mathbf{a}+\mathbf{b}, \mathbf{a}+\mathbf{b}+\mathbf{c}, \mathbf{b}+\mathbf{c}\}$$

with minimum squared error. For $l = 1$, the best subset is clearly the first vector (resulting in squared error $\|\mathbf{b}\|^2$), while for $l = 2$ the best choice is clearly the second and third vectors (resulting in zero error). Hence, from the point of view of "estimating" the vector $\mathbf{a}$, the best subset of size two is not an extension of the best subset of size one!

Consider next the set

$$\{\mathbf{a},\ (1-2\epsilon)(\mathbf{a}+\mathbf{b}),\ (1-2\epsilon)(\mathbf{a}-\mathbf{b})\} \tag{11}$$

where the vectors $\mathbf{a}$ and $\mathbf{b}$ are orthogonal as above, and $\|\mathbf{b}\| = \epsilon\|\mathbf{a}\|$ for some $\epsilon \ll 1$. Note that the length of the second and the third vectors is roughly $(1-\epsilon)$ times the length of the first vector. The only 1-amplifier basis of size one is the first vector, $\mathbf{a}$. On the other hand, the only 1-amplifier basis of size two is composed of the second and third vectors. ($\mathbf{a}$ is roughly 0.5 (second vector) + 0.5 (third vector), while the third vector is roughly $2\mathbf{a}$ - (second vector).) Also the determinant of the latter two is almost twice the determinant of any combination of $\mathbf{a}$ with another vector. Hence, we see that also with respect to our general criteria for basis selection, the best subset of size two is not an extension of the best subset of size one.

These examples show that simple "greedy" selection algorithms may fail to find the best basis. In the following sections, we shall consider two algorithms for selecting a self-basis of size m from the set $\mathbf{a}_1, \ldots, \mathbf{a}_n$ of vectors in $\mathcal{R}^m$:

- Longest Residual Vector (LRV) Selection.
- One by One Replacement Algorithm.

The former is an m-stage recursive selection algorithm, which starts with a single vector set, S_1, and at each stage adds a new vector to the existing set. That is, for $l = 0, \ldots, m-1$, we have $S_{l+1} = S_l \cup i$ for some $i \notin S_l$, where S_0 is the empty set. The latter is an iterative algorithm, which starts with some set S of size m, and at each iteration tries to improve its goodness by replacing one vector from S by a vector outside S. Eventually, we shall combine the two algorithms to get an algorithm with linear complexity in n, for choosing a basis close to a 1-amplifier basis.

In the last section we discuss an alternative selection approach using SDP relaxation.

3 Longest Residual Vector Selection

Under any "reasonable" criterion, the best subset of size *one* is the longest vector in the set $\mathbf{a}_1, \ldots, \mathbf{a}_n$. Choosing a vector achieving

$$A_{S_1} = \arg\max_{\mathbf{a}_i} \|\mathbf{a}_i\|$$

clearly maximizes the general determinant criterion in (8). And from the noise amplification point of view, this choice guarantees $g_i \leq 1$ for every $i \notin S_1$. Thus, the longest vector is also a 1-amplifier basis.

If we are restricted to *recursive* selection, then the most natural next choice is the longest *residual* vector (LRV). That is, for $1 \leq l < m$, suppose S_l is given, and we must create S_{l+1} by augmenting S_l by a single vector. Let $\tilde{\mathbf{a}}_i$ denote the projection error of $\mathbf{a}_i$ on the vectors in S_l. Then, LRV selects

$$S_{l+1} = S_l \cup i_{l+1},$$

where

$$i_{l+1} = \arg\max_{i \notin S_l} \|\tilde{\mathbf{a}}_i\|. \tag{12}$$

More specifically, let A_{S_l} denote the $m \times l$ matrix whose columns are the vectors $\mathbf{a}_k$, with k in S_l. Let $A_{S_l} = Q_l R_l$ be QR decomposition of A_{S_l}, where Q_l is an $m \times l$ matrix with orthonormal columns which span the column space of A_{S_l}, and R_l is an $l \times l$ upper triangular matrix. Then, the projection error $\tilde{\mathbf{a}}_i$ in (12) is given by

$$\tilde{\mathbf{a}}_i = (I - Q_l Q_l^T)\mathbf{a}_i.$$

Note that in this recursion, the next orthonormal matrix Q_{l+1} results in by augmenting Q_l with a unit vector in the direction of the projection error of the newly selected vector $\mathbf{a}_{i_{l+1}}$.

After m recursive steps of LRV, we get the final selected basis $S = S_m$. The total complexity is linear in n. It includes m steps of n linear projections in $\mathcal{R}^m$, that is $\sim m^3 n$ operations.

3.1 Performance Bounds for LRV Selection

It follows from the example regarding the set in (11), that the LRV recursive selection is not optimal for $l > 1$. In what sense, then, LRV is a good choice? As a greedy algorithm, LRV guarantees that at step l the coefficient g_{il} is bounded by one for all i (while possibly increasing g_{ij} for $j < l$). It also makes the largest increase in the determinant of $A_S^T A_S$ (given that we cannot change previously selected vectors).

The principle of this algorithm is similar to that of Matching Pursuit [2]. It is a classical result in Matching Pursuit [2] that for high SNR, the LRV is the MSE minimizing choice.

From our point of view, the most appealing feature of LRV for self-basis selection is that it provides some *uniform* bounds on the noise amplification and on the maximal determinant, although these bounds are quite high.

Theorem 1 (Noise Amplification Bound) *If the basis S is selected by the LRV algorithm, then for every i not in S, the k-th expansion coefficient (2) satisfies*

$$|g_{ik}| \leq 2^{m-k}, \quad k = 1, \ldots, m \tag{13}$$

where g_{i1} is the coefficient of the first selected vector, g_{i2} is the coefficient of the second selected vector, etc. Thus, S is a 2^{m-1} -amplifier. Furthermore, the bound is arbitrarily tight in the sense that there exists examples of a vector set for which LRV selection achieves the bound with almost equality.

Proof: 1. Upper bound. For ease of notation, assume that the set selected by LRV is $S = \{1, \ldots, m\}$ in that order, i.e, the first selected vector is $\mathbf{a}_1$, the second is $\mathbf{a}_2$, and so on. Let $\tilde{\mathbf{a}}_1, \ldots, \tilde{\mathbf{a}}_m$ denote the corresponding residual vectors as defined above, where $\tilde{\mathbf{a}}_1 \equiv \mathbf{a}_1$. By construction, the $\tilde{\mathbf{a}}_i$'s are orthogonal. Furthermore, each of the basis vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m$ can be written as

$$\mathbf{a}_i = \beta_{i,1}\tilde{\mathbf{a}}_1 + \ldots + \beta_{i,i-1}\tilde{\mathbf{a}}_{i-1} + \tilde{\mathbf{a}}_i \tag{14}$$

while each vector outside S can be written as

$$\mathbf{a}_i = \beta_{i,1}\tilde{\mathbf{a}}_1 + \ldots + \beta_{i,m}\tilde{\mathbf{a}}_m \tag{15}$$

for $m + 1 \leq i \leq n$. The key to the proof is the observation that the LRV selection rule implies that

$$|\beta_{ij}| \leq 1 \quad \forall i, j,$$

otherwise the residual of $\mathbf{a}_i$ at step j was longer than that of the vector actually selected. This implies that $\mathbf{a}_2$ contains $\mathbf{a}_1$ at most once; $\mathbf{a}_3$ contains $\mathbf{a}_1$ at most twice, one directly and one through $\tilde{\mathbf{a}}_2$; $\mathbf{a}_4$ contains $\mathbf{a}_1$ at most four times, one directly, one through $\tilde{\mathbf{a}}_2$, and two through $\tilde{\mathbf{a}}_3$; and so on. To make this argument formal, we write the residual vectors as

$$\tilde{\mathbf{a}}_i = \mathbf{a}_i - [g_{i,1}\mathbf{a}_1 + \ldots + g_{i,i-1}\mathbf{a}_{i-1}]. \tag{16}$$

Combining (14) and (16), we see that $g_{i,j}$ (the total contribution of $\mathbf{a}_j$ in $\mathbf{a}_j$), satisfies the following recursion in $i = 1, \ldots, m$

$$g_{ij} = \beta_{ij} - [\beta_{i,j+1}g_{j+1,j} + \ldots + \beta_{i,i-1}g_{i-1,j}], \quad 1 \leq j \leq i-1.$$

Since $|\beta_{ij}| \leq 1$, it follows that

$$|g_{ij}| \leq 1 + |g_{j+1,j}| + \ldots + |g_{i-1,j}|, \quad 1 \leq j \leq i-1.$$

Using the initial condition $|g_{21}| \leq 1$, a simple induction in i implies

$$|g_{ij}| \leq 2^{i-j-1}, \tag{17}$$

$1 \leq i \leq m$, $1 \leq j \leq i-1$. Extending the derivation to the vectors outside S, we have for $m+1 \leq i \leq n$, $|g_{ij}| \leq 1 + \sum_{l=j+1}^{m} |g_{l,j}|$, which together with (17) implies the desired result (13).

2. Achievability: To see that this bound can be arbitrarily approached, consider the set below of five (column) vectors in $\mathcal{R}^4$ (i.e., $m = 4, n = 5$). To better illustrate their special structure, we write $\mathbf{a}_1, \ldots, \mathbf{a}_4$ in a 4×4 matrix and put $\mathbf{a}_5$ separately:

$$\begin{pmatrix} 1 & -1 & -1 & -1 \\ 0 & +\epsilon & -\epsilon & -\epsilon \\ 0 & 0 & +\epsilon^2 & -\epsilon^2 \\ 0 & 0 & 0 & +\epsilon^3 \end{pmatrix} \cdot D, \qquad (1-\epsilon^2)^3 \begin{pmatrix} 1 \\ \epsilon \\ \epsilon^2 \\ \epsilon^3 \end{pmatrix} \tag{18}$$

where $\epsilon \ll 1$ and D is the diagonal matrix $\mathrm{diag}[1, (1-\epsilon^2), (1-\epsilon^2)^2, (1-\epsilon^2)^3]$. The multiplication by D from the right scales the second column by $(1-\epsilon^2)$, the third column by $(1-\epsilon^2)^2$, and so on. This guarantees that the longest vector in the set is $\mathbf{a}_1$. After projecting $\mathbf{a}_2, \ldots, \mathbf{a}_5$ on $\mathbf{a}_1$, the residual vectors have exactly the same structure, so $\tilde{\mathbf{a}}_2$ is the next longest. Continuing this process, we see that the LRV selected set S is $\mathbf{a}_1, \ldots, \mathbf{a}_4$. Furthermore, it follows that

$$\begin{aligned}
\mathbf{a}_2 &= -(1-\epsilon^2)\mathbf{a}_1 + \tilde{\mathbf{a}}_2 && (19)\\
\mathbf{a}_3 &= -2(1-\epsilon^2)^2\mathbf{a}_1 - (1-\epsilon^2)\mathbf{a}_2 + \tilde{\mathbf{a}}_3 && (20)\\
\mathbf{a}_4 &= -4(1-\epsilon^2)^3\mathbf{a}_1 - 2(1-\epsilon^2)^2\mathbf{a}_2 - (1-\epsilon^2)\mathbf{a}_3 + \tilde{\mathbf{a}}_4 && (21)\\
\mathbf{a}_5 &= +8(1-\epsilon^2)^4\mathbf{a}_1 + 4(1-\epsilon^2)^3\mathbf{a}_2 + 2(1-\epsilon^2)^2\mathbf{a}_3 + (1-\epsilon^2)\mathbf{a}_4. && (22)
\end{aligned}$$

Thus, in the limit as $\epsilon \to 0$, the bounds (17) and (13) become tight. The extension of this example to any m should be clear.

Theorem 2 (Determinant Loss Bound) *Let $det^* = \det(A_{S^*})$ denote the maximum determinant of an m-subset (see (9)). Then, the LRV basis A_S satisfies*

$$|\det(A_S)| \geq \frac{det^*}{m^{m/2}}. \tag{23}$$

Furthermore, this bound may be arbitrarily tight for certain vector sets.

Proof: Let

$$\tilde{A}_S = [\tilde{\mathbf{a}}_1; \ldots; \tilde{\mathbf{a}}_m]$$

denote the matrix whose columns are the residual vectors, where again for ease of notation we assume that $S = \{1, \ldots, m\}$, in this order. Since A_S and $\tilde{A}_S$ are related via a triangular transformation with ones on the diagonal (see (14)), we have $\det(\tilde{A}_S) = \det(A_S)$. On the other hand, (14) and (15) imply that for any set of vectors S', the corresponding matrix $A_{S'}$ is related to $\tilde{A}_S$ via a transformation whose elements are all absolutely bounded by one (the β_{ij}'s). The norm of the rows of the such a transformation is bounded by

$\sqrt{m}$. Since the absolute determinant of any transformation is bounded by the product of the norms of the rows, we have

$$|\det(A_{S'})| \leq m^{m/2} |\det(\tilde{A}_S)|. \tag{24}$$

This holds in particular for $S' = S^*$, and the bound follows.

To see that the bound can be approached, note that to get equality in (24) all the $\beta_{i,j}$'s must be either +1 or -1, and the rows of the transformation from $\tilde{A}_S$ to $A_{S'}$ should be orthogonal. These two conditions are met if $A_{S'} = \tilde{A}_S H$, where H is a *Hadamard* matrix, in which case $\det(H) = m^{m/2}$. In dimension 4 a Hadamard matrix exists; it consists of the four orthogonal vectors $[1,1,1,1]$, $[1,-1,1,-1]$, $[1,1,-1,-1]$ and $[1,-1,-1,1]$, the norm of each is equal to $\sqrt{4} = 2$, so $\det(H) = 16$. We use this transformation to extend the example given in the proof of achievability in Theorem 1 above. See the four vectors basis A_S in (18), whose corresponding matrix of residual vectors is $\tilde{A}_S = \text{diag}[1, \epsilon, \epsilon^2, \epsilon^3]$. We augment the set A_S by the the following four (column) vectors, which form the matrix $A_{S'}$:

$$(1-\epsilon^2)^3 \begin{pmatrix} +1 & +1 & +1 & +1 \\ +\epsilon & -\epsilon & +\epsilon & -\epsilon \\ +\epsilon^2 & +\epsilon^2 & -\epsilon^2 & -\epsilon^2 \\ +\epsilon^3 & -\epsilon^3 & -\epsilon^3 & +\epsilon^3 \end{pmatrix}.$$

It's not hard to verify that, again, the scaling by $(1-\epsilon^2)^3$ guarantees that the set selected by LRV will remain the vectors $A_S = [\mathbf{a}_1, \ldots, \mathbf{a}_4]$ of example (18). Their determinant is roughly $1 \cdot \epsilon \cdot \epsilon^2 \cdot \epsilon^3 = \epsilon^6$. On the other hand, due to the Hadamard transformation, the determinant of the new basis $A_{S'}$ is almost $m^{m/2} = 4^2 = 16$ larger. In the limit as $\epsilon \to 0$, these approximations become exact, and the bound (23) is arbitrarily approached. This example can easily be extended to any dimension m for which a Hadamard matrix exists.

4 One by One Replacement Algorithm

We next propose a simple iterative replacement algorithm. Given a first choice for an m subset S and the corresponding matrix A_S, the algorithm proceeds as follows:

a. Define a threshold $\alpha > 1$ and a maximal number of iterations N.

b. Compute the expansion coefficients of all the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$, relative to the given basis (i.e. for each $\mathbf{a}_i$ compute $\boldsymbol{g}_i = A_S^{-1}\mathbf{a}_i$).

c. Find $G = \max_{i,j}(|g_{ij}|)$ and the coordinates i, j for which the maximum is achieved.

d. Compare G with α. If it is greater than α, then replace the j-th vector of A_S by $\mathbf{a}_i$, and update the set S and the matrix A_S. If the maximal number of iterations N has not been reached, repeat the algorithm from step b.

e. If G is not greater than α, then the algorithm stops, and by Cramer's law the current basis is an α-amplifier or better.

By the bound on the determinant loss of LRV selection in Theorem 2, it follows that if this algorithm starts with the LRV set (or better) and completes N replacements, then the resulting determinant will be at least

$$|\det(A_S)| \geq \frac{det^*}{m^{m/2}} \, \alpha^N.$$

In this case, Cramer's Law (10) implies that the basis is a $(\sqrt{m})^m/\alpha^N$-amplifier. On the other hand, the stopping rule above implies that if we stop in less than N iterations then we get an α-amplifier. It follows that combining the LRV selection algorithm for initialization with the one-by-one replacement algorithm above guarantees noise amplification which is the maximum between α and $m^{m/2}/\alpha^N$. Optimizing over α, we get the following corollary.

Corollary 1 (Close to 1-amplifier) *If we choose $\alpha = m^{m/(2(N+1))}$, then after at most N iterations we get an $m^{m/(2(N+1))}$-amplifier.*

Substituting, for example, $N = m^2/2$, we get an $m^{1/m}$-amplifier. Note that $m^{1/m}$ goes to 1 as m goes to infinity, and it is always less than 1.5.

The complexity of the one-by-one replacement algorithm is $O(n \cdot m^2 \cdot N)$, which for $N = O(m^2)$ becomes $O(n \cdot m^4)$ (i.e. linear in n, polynomial in m), and it results in a basis which is quite close to a 1-amplifier. As noted above, the complexity of full search is $O(\binom{n}{m})$ which can be considerably higher than $O(n \cdot m^4)$ for $n \gg m \gg 1$.

5 Maximum Determinant Results

Small residual entropy is associated with finding a set with maximal determinant. The results we have in this case are more pessimistic, i.e. our results suggest that finding the maximal determinant may be a problem of high complexity. The proofs are omitted for lack of space.

As stated in Theorem 2 above, the Longest Residual Vector Selection algorithm guarantees determinant loss of at most $\sqrt{m}^m$, and we can construct examples where this bound is tight. In particular, for each $\epsilon > 0$, there exist examples for which the Residual Longest Vector Selection results in a 1-amplifier basis (i.e. any replacement of only 1 vector will not increase the determinant of A_S), however the maximal determinant det^* satisfies

$$det^* \geq (\sqrt{m/(1+\epsilon)})^m |\det(A_S)|.$$

Moreover, for every $m > k > 1$, we can construct examples of a set of vectors where in addition to the above, replacing any subset of k vectors from A_S by any k vectors does not increase the determinant of A_S, yet the maximal determinant det^* satisfies

$$det^* \geq (\sqrt{m/k})^m |\det(A_S)|.$$

For $k = m - 1$ we have a tighter bound. If replacing any $m - 1$ vectors of A_S by any $m - 1$ vectors will not increase the determinant of A_S, then the maximal determinant det^* is bounded by

$$det^* \leq (\sqrt{m})^{m/(m-1)} |\det(A_S)|,$$

and there exist examples for which the bound is achieved.

6 Alternative Approaches and Extensions

As suggested by Y. Eldar, [7], a relaxed version of the setting discussed in this paper can be formulated as a semi-definite programming (SDP) problem. In particular, the residual entropy optimization is related to the *experiment design* variant of the *max-det* problem

considered by Vandenberghe, Boyd and Wu in [6]. They present an interior-point method to maximize the determinant of a matrix subject to linear matrix inequalities (LMIs). The relaxation of [6] is that the set of different vectors chosen may be larger than m. This can be thought of as "soft selection" of each of the n vectors by a parameter between zero and one, rather than "hard selection" of exactly m vectors. This does not necessarily lead to un-ambiguous selection of the best m-subset, and some post processing mechanism should be used to make the final selection.

The complexity of the interior-point method seems larger than the complexity of the iterative algorithm presented here. Another benefit of the current algorithm is that it only uses computations of the same type as the ones used anyway by the system after the basis is selected (i.e., prediction coefficients computations). On the other hand, SDP related methods have the advantage that they can handle a much wider variety of problems, for example, direct minimization of the total MSE in (5). We tend to believe that beyond a complexity gain, the algorithm presented here may in some cases give better results. This issue is currently under study.

Future extensions of this work include: (i) extending the discussion to signal space, i.e., selecting a good m-subset of n signals derived from m "hidden" signals; (ii) extending the discussion to *non* additive-Gaussian settings, i.e., minimization of $H(\mathbf{X}_{S^c}|\mathbf{X}_S)$ or $\sum_{I \notin S} H(X_i|\mathbf{X}_S)$ over S for more general joint distributions.

Acknowledgement

We thank Yonina Eldar for pointing out the connection of our work to Matching Pursuit and to SDP, and for many fruitful discussions. We thank Arie Yeredor for the first example in Section 2.

References

[1] I. Daubechies. Time-frequency localization operators: a geometric phase space approach. *IEEE Trans. On Information Theory*, IT-34, pp. 605-612, 1988.

[2] S. Mallat and Z. Zhang. Matching Pursuit in a time-frequency dictionary. *IEEE Trans. On Signal Processing*, vol. 41, pp. 3397-3415, 1993.

[3] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by Basis Pursuit, SIAM J. Sci Comp., vol. 20,1 pp. 33-61, 1999.

[4] T. Cover and J. Thomas, Elements of Information Theory, Wiley & Sons, New York, 1991.

[5] K. Lenstra, H. W. Lenstra, and L. Lovasz, Factoring Polynomials with Rational Coefficients. Math. Ann., vol. 261, pp. 515-534, 1982.

[6] L. Vandenberghe, S. Boyd, and S.-P. Wu, Determinant maximization with linear matrix inequality constraints. SIAM Journal on Matrix Analysis and Applications, vol. 19(2), pp. 499-533, 1998.

[7] Y. Eldar, Private communication.

[8] R. A. Horn and C.R. Johnson, Matrix Analysis. Cambridge University Press, 1985.