Optimized Projections for Compressed Sensing

Michael Elad

The Department of Computer Science

The Technion – Israel Institute of Technology

Haifa 32000 Israel

Email: elad@cs.technion.ac.il.

October 19, 2006

Abstract

Compressed-Sensing (CS) offers a joint compression- and sensing-processes, based on the existence of a sparse representation of the treated signal and a set of projected measurements. Work on CS thus far typically assumes that the projections are drawn at random. In this paper we consider the optimization of these projections. As a direct such optimization is prohibitive, we target an average measure of the *mutual-coherence* of the effective dictionary, and demonstrate that this leads to better CS reconstruction performance. Both the Basis-Pursuit and the Orthogonal-Matching-Pursuit are shown to benefit from the newly designed projections, with a reduction of the error-rate by a factor of 10 and beyond.

keywords: Compressed-sensing, Sparse and redundant representations, optimized projections, Basis-Pursuit, Orthogonal Matching-Pursuit, Mutual-Coherence.

1 Introduction

Consider a family of signals $\{\mathbf{x}_j\}_j \in \mathcal{R}^n$, known to have sparse representations over a fixed dictionary $\mathbf{D} \in \mathcal{R}^{n\times k}$. Thus we have that these signals can be described by

$$\forall j, \quad \mathbf{x}_j = \mathbf{D}\alpha_j, \tag{1}$$

with $\|\alpha_j\|_0 \leq T \ll n$ for all j. The ℓ^0-norm used here simply counts the number of non-zeros in α_j.

Compressed-Sensing (CS) offers a joint sensing- and compression-processes for such signals [1, 2, 3, 4, 5, 6, 7]. Using a projection matrix $\mathbf{P} \in \mathcal{R}^{p \times n}$ with $T < p \ll n$, CS suggests to represent $\mathbf{x}_j$ by p scalars given by

$$\mathbf{y}_j = \mathbf{P}\mathbf{x}_j. \tag{2}$$

The original signal $\mathbf{x}_j$ can be reconstructed from $\mathbf{y}_j$ by exploiting the sparsity of it's representation – i.e., among all possible α satisfying $\mathbf{y}_j = \mathbf{PD}\alpha$ we seek the sparsest. If this representation coincides with α_j, we get a perfect reconstruction of the signal using Equation (1). This reconstruction thus requires the solution of

$$\min_{\alpha} \|\alpha\|_0 \quad s.t. \quad \mathbf{y}_j = \mathbf{PD}\alpha, \tag{3}$$

which is known to be NP-hard even for moderate-sizes of the linear system in the constraint [8, 9]. Approximation techniques, known as pursuit algorithms are deployed, and are proven to lead to the true result for very sparse solutions [11, 12, 10].

Work on CS thus far assumes that $\mathbf{P}$ is drawn at random, which simplifies its theoretical analysis, and also facilitates a simple implementation [1, 2, 3, 4, 5, 6, 7]. In this paper we show that by optimizing the choice of $\mathbf{P}$ such that it leads to better coherence of the effective dictionary, a substantially better CS reconstruction performance is obtained, with both the Basis-Pursuit (BP) [10] and the Orthogonal Matching-Pursuit (OMP) algorithms [11, 12].

In the next section we provide the intuition behind CS, along with a statement of the

main results in the literature regarding its expected performance, which are related to this work. Section 3 concentrates on a proposed iterative method for improving the projections based on the *mutual-coherence* (as will be defined shortly) of the overall new dictionary. We demonstrate experimental results Section 4 and show the performance gain obtained with the optimized projections. As this work is the first to consider the design of the projections, and as it approaches this problem indirectly by improving the *mutual coherence*, there is clearly a room for future work and improvements. Ideas on how to further extend this work are brought in Section 5.

2 Compressed Sensing – The Basics

We have described above the core idea behind Compressed-Sensing. The first question one must ask is – why will it work at all? In order to answer this question, we need to recall the definition of the *mutual-coherence* of a dictionary [13, 14].

Definition 1: For a dictionary $\mathbf{D}$, its *mutual-coherence* is defined as the largest absolute and normalized inner product between different columns in $\mathbf{D}$. Put formally, this reads

$$\mu\{\mathbf{D}\} = \max_{1\leq i,j\leq k \text{ and } i\neq j} \frac{\left|\mathbf{d}_i^T \mathbf{d}_j\right|}{\|\mathbf{d}_i\| \cdot \|\mathbf{d}_j\|}. \tag{4}$$

The *mutual-coherence* provides a measure of the worst similarity between the dictionary columns, a value that exposes the dictionary's vulnerability, as such two closely related columns may confuse any pursuit technique.

A different way to understand the *mutual-coherence* is by considering the Gram matrix $\mathbf{G} = \tilde{\mathbf{D}}^T\tilde{\mathbf{D}}$, computed using the dictionary after normalizing each of its columns. The

off-diagonal entries in $\mathbf{G}$ are the inner products that appear in Equation (4). The *mutual-coherence* is the off-diagonal entry g_{ij} with the largest magnitude.

Suppose that the signal $\mathbf{x}_0$ has been constructed by $\mathbf{x}_0 = \mathbf{D}\alpha_0$ with a sparse representation. Suppose further that the following inequality is satisfied:

$$\|\alpha_0\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu\{\mathbf{D}\}}\right). \tag{5}$$

A fundamental set of results state that [13, 14, 15]:

1. The vector α_0 is necessarily the sparsest one to describe $\mathbf{x}_0$, i.e. it is the solution of

$$\min_{\alpha} \|\alpha\|_0 \quad s.t. \quad \mathbf{x}_0 = \mathbf{D}\alpha. \tag{6}$$

2. The BP algorithm for approximating α_0, which is solving the linear program

$$\min_{\alpha} \|\alpha\|_1 \quad s.t. \quad \mathbf{x}_0 = \mathbf{D}\alpha, \tag{7}$$

   is guaranteed to find α_0 exactly. And

3. The OMP for approximating α_0 is also guaranteed to succeed. The OMP is a greedy and sequential method that accumulates the non-zeros in α_0 one at a time, while attempting to obtain the fastest decrease of the residual error $\|\mathbf{x}_0 - \mathbf{D}\alpha\|$.

Based on the above, suppose that the projection matrix $\mathbf{P}$ has been chosen and we are to solve

$$\min_{\alpha} \|\alpha\|_0 \quad s.t. \quad \mathbf{y}_0 = \mathbf{P}\mathbf{x}_0 = \mathbf{P}\mathbf{D}\alpha. \tag{8}$$

If the original representation is satisfying the more strict requirement

$$\|\alpha_0\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu\{\mathbf{PD}\}}\right) \leq \frac{1}{2}\left(1 + \frac{1}{\mu\{\mathbf{D}\}}\right), \tag{9}$$

then necessarily, the original α_0 is the solution of the problem posed in (8), both pursuit methods will manage to recover it perfectly, and thus reconstruct $\mathbf{x}_0$ well.

The above implies that if $\mathbf{P}$ is designed such that $\mu\{\mathbf{PD}\}$ is as small as possible, this allows a wider set of candidate signals to reside under the umbrella of successful CS behavior. While this conclusion is true from a worst-case stand-point, it turns out that the *mutual-coherence* as defined above does not do justice to the actual behavior of sparse representations and pursuit algorithms' performance. Thus, if we relax our expectations and allow a small fraction of signals with the same representation's cardinality to fail, than values of $\|\alpha_0\|_0$ substantially beyond the above bound are still leading to successful CS. Considering the average performance of CS as a function of this cardinality, an "average" measure of coherence is more likely to describe its true behavior.

Another fundamental question in CS is the following: How many measurements are required for successful reconstruction? Assuming that the cardinality of the representation, $\|\alpha_0\|_0 = T$, is known, one needs at least $2T$ measurements to form a non-linear set of $2T$ equations with $2T$ unknowns (the indices of the non-zeros and their coefficients). Recent work has established that indeed, for a high success-rate of CS, it is enough to use $\mathcal{O}\{T\}$ measurements, with an appropriate coefficient (e.g. $\text{Const} \cdot \log(n) \cdot T$, as found in [4]). These results are typically accompanied by an assumption about the specific dictionary structure, the use of random projections, and considering an asymptotic case where the relative sizes grow to infinity.

If we address this very question of the required number of projections from the point of view of the value of $\mu\{\mathbf{PD}\}$, we are likely to find that $\mathcal{O}\{n\}$ measurements are needed,

loosing all the compressibility potential in CS. Again we find that replacing the measure $\mu\{\mathbf{PD}\}$ with a parallel one that considers average absolute inner-products may do more justice to the conclusion about the required number of measurements.

3 Optimizing the Projection Matrix

In this section we shall consider a different *mutual-coherence*, which reflects average behavior. We define it as follows:

Definition 2: For a dictionary $\mathbf{D}$, its t-averaged *mutual-coherence* is defined as the average of all absolute and normalized inner products between different columns in $\mathbf{D}$ (denoted as g_{ij}) that are above t. Put formally,

$$\mu_t\{\mathbf{D}\} = \frac{\sum_{1\leq i,j\leq k \text{ and } i\neq j} (|g_{ij}| \geq t) \cdot |g_{ij}|}{\sum_{1\leq i,j\leq k \text{ and } i\neq j} (|g_{ij}| \geq t)}. \tag{10}$$

As the value of t grows, we obtain that $\mu_t\{\mathbf{D}\}$ grows and approaches $\mu\{\mathbf{D}\}$ from below. Also, it is obvious from the definition that $\mu_t\{\mathbf{D}\} \geq t$. In the optimization procedure we are about to describe we will target this value and minimize it iteratively.

Note that a different and more direct approach towards the design of the projection matrix would be its learning based on signal examples and tests involving the pursuit algorithm deployed. We believe that such a method is likely to lead to better performance compared to the method described here. Nevertheless, such a direct scheme is also expected to be far more complex and involved, and thus its replacement with the optimization of $\mu_t\{\mathbf{PD}\}$ is an appealing alternative.

Put very simply, our goal is to minimize $\mu_t\{\mathbf{PD}\}$ with respect to $\mathbf{P}$, assuming that the

dictionary $\mathbf{D}$ and the parameter t are known and fixed. Since $\mu_t\{\mathbf{PD}\}$ is defined via the entries of the Gram-matrix, we propose an iterative scheme that includes transformations from- and to- the Gram matrix in every iteration. This algorithm is inspired by a similar approach adopted in [16] for the design of Grassmanian frames that minimize the *mutual-coherence* of a desired dictionary. While the work in [16] considers $\mathbf{D}$ as an unknown to be built, we target $\mathbf{P}$ in the expression $\mu_t\{\mathbf{PD}\}$, which adds more complications.

A slightly different mode of operation of the above algorithm can be proposed, where t varies from one iteration to another, by addressing at all times a constant fraction of the entries in the Gram matrix. For example, the value t can be updated at each iteration such that it targets the top 20% of the inner-products. We shall denote the average *mutual-coherence* of the top $t\%$ by $\mu_{t\%}\{\mathbf{PD}\}$, and, as we shall see in the next section, it is this measure that we will work with. The algorithm for optimizing $\mathbf{P}$ with the above two options is described in Figure 1.

In this algorithm we start with a random set of p projections stored in the matrix $\mathbf{P}$. As our main objective is the reduction of the inner-products that are above t in absolute value (assuming the first mode of operation), the Gram matrix of the normalized effective dictionary is computed, and these values are "shrinked" multiplying their values by $0 < \gamma < 1$. As such decrease in the magnitude should lead to monotonic non-increasing function, entries in $\mathbf{G}$ with magnitude below t but above γt are "shrinked" by a smaller amount,

Objective: Minimize $\mu_t\{\mathbf{PD}\}$ with respect to $\mathbf{P}$.

Input: Use the following parameters:

- t or $t\%$ - coherence (fixed or relative) threshold,
- $\mathbf{D}$ - the dictionary,
- p - number of measurements,
- γ - down-scaling factor, and
- *Iter* -number of iterations.

Initialization: Set $\mathbf{P}_0 \in \mathcal{R}^{p\times n}$ to be an arbitrary random matrix.

Loop: Set $k = 0$ and repeat *Iter* times:

1. *Normalize:* Normalize the columns in the matrix $\mathbf{P}_k\mathbf{D}$ and obtain the effective dictionary $\hat{\mathbf{D}}_k$.
2. *Compute Gram Matrix:* $\mathbf{G}_k = \hat{\mathbf{D}}_k^T\hat{\mathbf{D}}_k$
3. *Set Threshold:* If mode of operation is fixed, use t as threshold. Otherwise, choose t such that $t\%$ of the off-diagonal entries in $\mathbf{G}_k$ are above it.
4. *Shrink:* Update the Gram matrix and obtain $\hat{\mathbf{G}}_k$ by

$$g_{ij} = \begin{cases} \gamma g_{ij} & |g_{ij}| \geq t \\ \gamma t \cdot sign(g_{ij}) & t > |g_{ij}| \geq \gamma t \\ g_{ij} & \gamma t > |g_{ij}| \end{cases} .$$

5. *Reduce Rank:* Apply SVD and force the rank of $\hat{\mathbf{G}}_k$ to be equal to p.
6. *Squared-Root:* Build the squared-root of $\hat{\mathbf{G}}_k$, $\mathbf{S}_k^T\mathbf{S}_k = \hat{\mathbf{G}}_k$, where $\mathbf{S}_k$ is of size $p \times n$.
7. *Update* $\mathbf{P}$*:* Find $\mathbf{P}_{k+1}$ that minimizes the error $\|\mathbf{S}_k - \mathbf{PD}\|_F^2$.
8. *Advance:* Set $k = k + 1$.

Result: The output of the above algorithm is $\mathbf{P}_{Iter}$.

Figure 1: The numerical algorithm for optimizing the projection matrix $\mathbf{P}$.

using the function

$$y = \begin{cases} \gamma x & |x| \geq t \\ \gamma t \cdot sign(x) & t > |x| \geq \gamma t \\ x & \gamma t > |x| \end{cases} . \tag{11}$$

This function is described graphically for $t = 0.5$ and $\gamma = 0.6$ in Figure 2. For convenience, the functions $y = x$ and $y = \gamma x$ are also shown. As can be seen, the proposed function is embraced between these two, switching to the slower one at t.

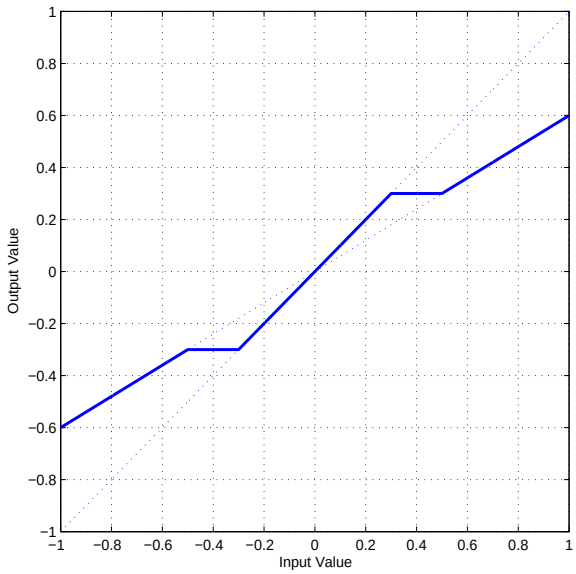

Figure 2: The shrink operation employed in the algorithm for $t = 0.5$ and $\gamma = 0.6$.

The above shrinking operation causes the resulting Gram matrix to become full-rank in the general case. Thus, the next steps are mending this by forcing a rank p and finding the matrix $\mathbf{P}$ that could best describe the squared-root of the obtained Gram matrix. Thus, steps 1-4 in the algorithm are addressing the objective of the process – the reduction of $\mu_t\{\mathbf{PD}\}$, and steps 5-7 are responsible for the feasibility of the proposed new Gram matrix

and the identity of the emerged projection matrix.

Regarding convergence properties, not much can be said in general. The overall problem is far from being convex, and convergence is guaranteed only of γ is chosen very close to 1. However, as we show next, in practice one can choose $\gamma = 0.5$ and still get convergence, and in fact an accelerated one, compared to the use of higher values. Since the objective function can be evaluated after every iteration with almost no additional cost, this could be used for an automatic stopping of the algorithm (in case an assent is started), and even for tuning the parameters dynamically form one iteration to another.

To illustrate the behavior of the above algorithm, we provide a demonstration of its results in Figure 3. Considering a random dictionary (every entry is drawn from iid zero mean and unit variance Gaussian distribution) of size 200×400, we seek the best projection matrix containing 30 projections, such that $\mu_t\{\mathbf{PD}\}$ is minimized for $t = 0.2$. The initialization is a random matrix $\mathbf{P}_0$ of size 30×200, built the same way as the dictionary.

We use several values of γ, from 0.55 to 0.95. In all cases we obtain convergence, and it is faster as γ is smaller. The value of $\mu_t\{\mathbf{PD}\}$ is by definition above t, but as can be seen, it gets smaller quite effectively.

Figure 4 presents the histogram of the absolute off-diagonal entries of $\mathbf{G} = \mathbf{D}^T\mathbf{P}^T\mathbf{PD}$ before the optimization and after 50 iterations (using $\gamma = 0.5$ and $t = 0.2$). As can be seen, there is a marked shift towards the origin of the histogram after optimization, with an emphasis on the right tail which represents the higher values. A similar effect is seen also in Figure 5, which presents similar histograms, this time working with $t\% = 40\%$. Thus, in this run we target at every iteration the minimization of the average of the top 40% of

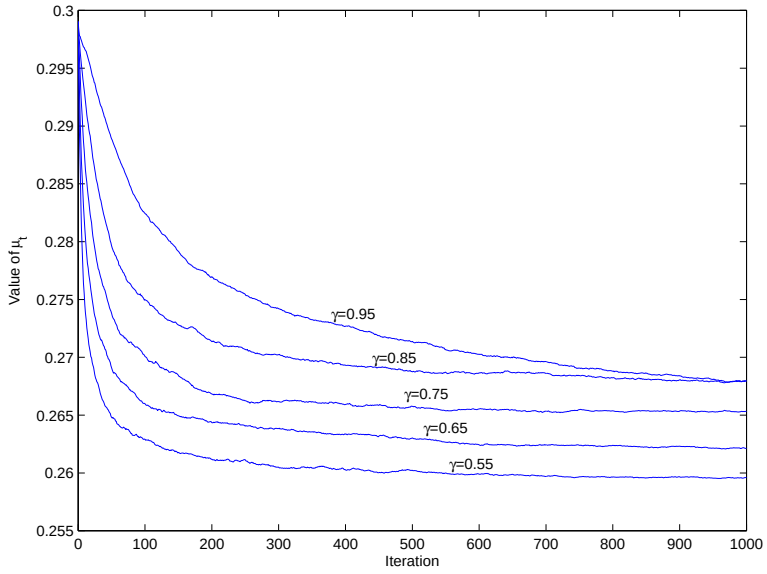

Figure 3: The value of $\mu_t\{\mathbf{P}_k\mathbf{D}\}$ as a function of the iteration for $t = 0.2$ and various values of γ.

the off-diagonal entries in the Gram matrix.

4 Compressed-Sensing: Experimental Results

It is now time to assess how the optimized projections perform in the compressed-sensing setting. We should remind the reader that in this work we assume that by optimizing $\mu_t\{\mathbf{PD}\}$ w.r.t. $\mathbf{P}$, one leads to more informative projections, which in turn leads to better CS performance. This link between μ_t and CS is yet to be theoretically analyzed and proven, and here we limit our study to an empirical one. The proposed test includes the following steps:

Stage 1 - Generate Data: Choose a dictionary $\mathbf{D} \in \mathcal{R}^{n\times k}$, and synthesize N test signals $\{\mathbf{x}_j\}_{j=1}^N$ by generating N sparse vectors $\{\alpha_{j=1}^N\}_j$ of length k each, and computing

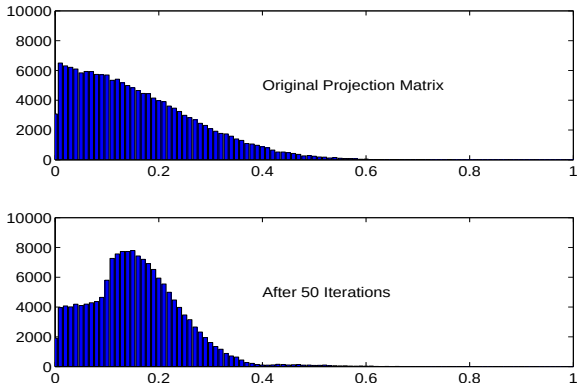

Figure 4: The histogram of the absolute off-diagonal entries of $\mathbf{G}$ before the optimization and afterwards, using a fixed threshold $t = 0.2$.

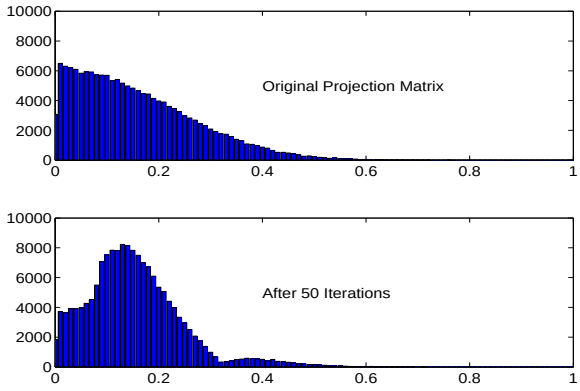

Figure 5: The histogram of the absolute off-diagonal entries of $\mathbf{G}$ before the optimization and afterwards, operating on the top 40% of the inner-products.

$\forall j,\ \mathbf{x}_j = \mathbf{D}\alpha_j$. All representations are to be built using the same low cardinality $\|\alpha\|_0 = T$.

Stage 2 - Initial Projection: For a chosen number of measurements m, create a random projection matrix $\mathbf{P}$, and apply it to the signals, obtaining $\forall j,\ \mathbf{y}_j = \mathbf{P}\mathbf{x}_j$. Compute the effective dictionary $\hat{\mathbf{D}} = \mathbf{P}\mathbf{D}$.

Stage 3 - Performance Tests: Apply the BP and the OMP to reconstruct the signals by approximating the solution of

$$\hat{\alpha}_j = arg\min_{\alpha} \|\alpha\|_0 \quad s.t. \quad \mathbf{y}_j = \hat{\mathbf{D}}\alpha,$$

and testing the error $\|\mathbf{x}_j - \mathbf{D}\hat{\alpha}_j\|_2$. Measure the average error-rate – A reconstruction with a mean-squared-error above some threshold is considered as a reconstruction failure.

Stage 4 - Optimize Projections: Use the algorithm as described in Section 3 to optimize the projection matrix $\mathbf{P}$.

Stage 5 - Re-evaluation of CS performance: Re-apply the performance tests of the BP and the OMP as described above, and see how the newly designed projections influence the CS behavior.

We have followed the above stages in the following two experiments. The first experiment studies the performance of CS before and after the optimization of the projections, with BP and OMP, and for varying amounts of measurements. The second one studies the effect of the cardinality of the representations.

In the first experiment we used a random dictionary of size 80×120 (other options such as a redundant DCT dictionary, where tested too, and found to lead to qualitatively the same results, and thus omitted). This size was chosen as it enables the CS performance evaluation in reasonable time. We generated $N = 100,000$ sparse vectors of length $k = 120$ with $T = 4$ non-zeros in each. The non-zeros locations were chosen at random and populated with iid zero-mean and unit variance Gaussian values. These sparse vectors were used to create the example-signals with which to evaluate the CS performance. CS performance was tested with varying values of m in the range $16 \div 40$. The relative error rate was evaluated as a function of m for both the BP and the OMP, before and after the projection optimization. The projection optimization (per every value of m) was done using up to $1,000$ iterations[1] using $\gamma = 0.95$ and varying $\%t = 20\%$. The results are shown in Figure 6.

Each point in the shown graph represents an average performance, accumulated over a possibly varying number of experiments. While every point is supposed to present an average performance over $N = 100,000$ examples, in cases where more than 300 errors were accumulated, the test was stopped and the average so far was used instead. This was done in order to reduce the overall test run-time. Another substantial speed-up was obtained by replacing the BP direct test (which requires a linear programming solver) with a much much faster alternative, as described in Appendix A.

As can be seen and as expected, the results of both pursuit techniques improve as m increases. In this test the BP performs much better than the OMP. The optimized projections are indeed leading to improved performance for both algorithms. For some values

[1]The algorithm is stopped in case of an increase in the value $\mu_{t\%}$.

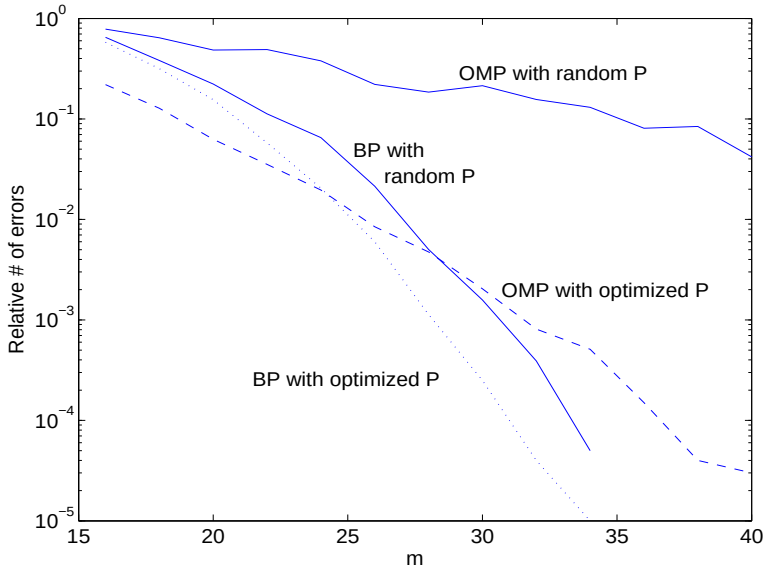

Figure 6: Compressed-Sensing relative errors as a function of the number of measurements m, with random projections and optimized projections. Note: A vanishing graph implies a zero error rate.

of m there is nearly a $10:1$ improvement factor for the BP and more than $100:1$ improvement for the OMP. Indeed, the OMP with the optimized projections lead to better performance compared to the original BP, for low and mid-values of m.

The second experiment is similar to the first one, this time fixing $m = 25$ and varying T in the range $1 \div 7$. The results are shown in Figure 7. As expected, as T grows, performance deteriorates. However, the optimized projections are consistent in their improved performance.

We should emphasize that the presented results do not include a thorough optimization of the parameters γ and t, and the relation between μ_t and the CS performance remains still obscure at this stage. Also, our experiments concentrated on one specific choice of dictionary size that enables reasonable run-time simulation, and this has an impact on

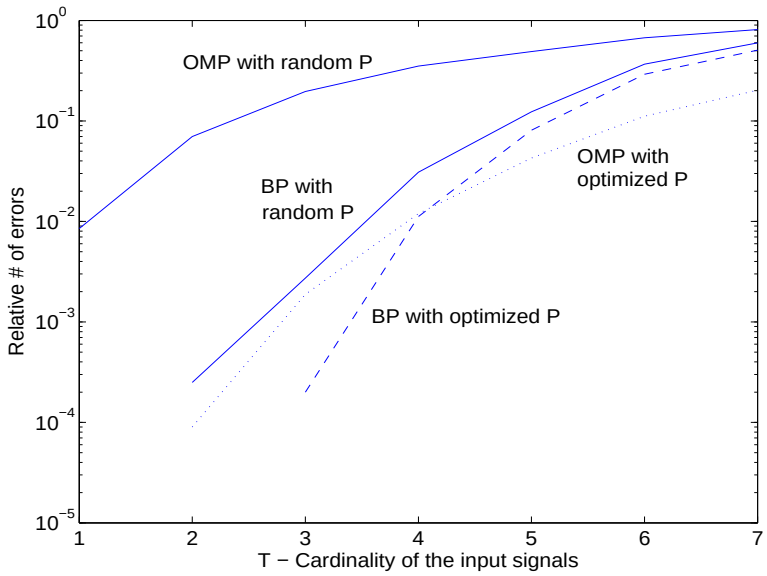

Figure 7: Compressed-Sensing relative errors as a function of the signals' cardinality T, with random projections and optimized projections.

the relatively weak performance CS shows. Other experiments we have done with much larger dictionaries show the same improvement as above, but require too long run-time for gathering fair statistics, and thus avoided. Still, the point this paper is making about the potential to get better projections and thereby improving CS performance, is clearly demonstrated.

5 Conclusions

Compressed-Sensing (CS) is an emerging field of activity with beautiful theoretical results that state that signals can be compressed and sensed at the same time. This is based on the structural assumption such signals are satisfying – having a sparse and redundant representation over a specific dictionary. A crucial ingredient in the deployment of the

CS idea is the use of linear projections that mix the signal. This operation has been traditionally chosen as a random matrix. This work aims to show that better choices of such mixtures are within reach. The projections can be designed such that the average *mutual-coherence* of the effective dictionary becomes favorable. We have defined this property, shown how to design a projection operator based on it, and demonstrated how it is indeed leading to better CS performance.

The idea of optimizing the projections is appealing and should be further studied. Here are several intriguing questions that future work could consider:

- How can the proposed optimization algorithm be performed or approximated for very high dimensions? This is important in cases where the CS is deployed on images or other signals of high-dimensions.

- Optimizing the projections can be done alternatively using a direct method that considers CS performance, rather than addressing an indirect measure as done in this work. Further work is required to explore this option, and show how effective it is compared to the one discussed in this work.

- We should develop a theoretical link between the average *mutual-coherence* as presented here, to the CS performance, so as to give better justification for the proposed work. Perhaps there is yet another simple measure of the effective dictionary $\mathbf{PD}$ that could replace $\mu_t\{\mathbf{PD}\}$ and lead to better results.

Acknowledgement

The author would like to thank Dr. Michael Zibulevsky for helpful discussions and his fruitful ideas on how to speed-up the tests carried out in this work.

Appendix A - Evaluating BP's Performance

The problem we face is the following: We generate a sparse vector α_0 and compute from it the measurement vector $\mathbf{y}_0 = \mathbf{PD}\alpha_0$. In order to determine whether BP succeeds in the recovery of the signal $\mathbf{x} = \mathbf{D}\alpha_0$, we should solve

$$\hat{\alpha} = arg\min_{\alpha} \|\alpha\|_1 \quad s.t. \quad \mathbf{y}_0 = \mathbf{PD}\alpha, \tag{A-1}$$

and check whether $\hat{\alpha} = \alpha_0$. The problem in such a direct approach is the need to deploy a linear programming solver per each test, and as we are interested in many thousands of such tests, this approach becomes prohibitive.

Since we are dealing here with a synthetic test, where the desired solution is a-priori known, we can replace the direct solution of (A-1) with a much more moderate test of considering α_0 and checking wether it is indeed it's global minimizer. In order to do so, we consider the necessary first-order KKT conditions, as emerging from the Lagrangian of (A-1). The Lagrangian is given by

$$\mathcal{L}(\alpha, \lambda) = \|\alpha\|_1 + \lambda^T(\mathbf{y}_0 - \mathbf{PD}\alpha), \tag{A-2}$$

with λ serving as the Lagrange ,multipliers. Taking its derivative with respect to α, and using the fact that the derivative of the absolute value at zero leads to the feasible interval

$[-1, 1]$ (considering the sub-gradients), we obtain

$$\mathbf{D}^T\mathbf{P}^T\lambda = \begin{cases} +1 & \alpha_0(j) > 0 \\ -1 & \alpha_0(j) < 0 \\ u_j & \alpha_0(j) = 0 \end{cases}, \tag{A-3}$$

where one must require $\forall j, \ -1 \leq u_j \leq 1$.

Thus, if we find a feasible solution λ to this system, we can guarantee that α_0 is the solution of (A-1) and thus the BP is expected to succeed. If we cannot find a solution, we suspect that BP fails. Declaration of failure in such a case is definitely possible, but leads to an upper-bound on the true number of errors, as our numerical scheme for solving Equation (A-3) may fail in-spite of the BP success. Assuming that the expected number of such suspected failures is substantially smaller compared to N (as is indeed the case in our simulations), we can directly try to solve (A-1) for this few cases, and see whether failure takes place.

As for the solution of (A-3), this can be achieved in various ways. We separate the equation-set to two parts – the equality- and the inequality-constraints, denoted by $\mathbf{A}_1\lambda = \mathbf{b}$ and $-\mathbf{1} \leq \mathbf{A}_2\lambda \leq \mathbf{1}$, respectively. We use the penalty method, minimizing the function

$$f(\lambda) = \|\mathbf{A}_1\lambda - \mathbf{b}\|_2^2 + \beta\|\mathbf{W}\mathbf{A}_2\lambda\|_2^2, \tag{A-4}$$

with respect to λ. The matrix $\mathbf{W}$ is a diagonal weight matrix, initialized as $\mathbf{W} = \mathbf{I}$. Starting with a very small β, the first constraint is satisfied while the second might be violated. Iterating and increasing the value of β, the first term remains zero while the second one gets closer to the satisfaction of $-\mathbf{1} \leq \mathbf{A}_2\lambda \leq \mathbf{1}$. A more delicate update step can be proposed, where the extreme entries in the vector $|\mathbf{A}_2\lambda|$ that are above 1 are

treated by increasing their weight in $\mathbf{W}$. A finite number of such an iterative algorithm (50 iterations) was used, and shown to be 1-2 orders of magnitude faster than the full LP solver.

References

[1] Candès, E.J., Romberg, J. and Tao, T. (2006) Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, *IEEE Trans. on Inf. Theory*, Vol. 52, pp. 489–509, February.

[2] Candès, E.J. and Romberg, J. (2005) Quantitative robust uncertainty principles and optimally sparse decompositions, to appear in *Foundations of Computational Mathematics*.

[3] Candès, E.J. and Tao, T. (2006) Near optimal signal recovery from random projections: universal encoding strategies, to appear in *IEEE Trans. on Inf. Theory*.

[4] Donoho, D.L. (2006) Compressed sensing, *IEEE Trans. on Inf. Theory*, Vol. 52, pp. 1289–1306, April.

[5] Tsaig, Y. and Donoho, D.L. (2006) Extensions of compressed sensing, *Signal Processing*, Vol. 86, pp. 549–571, March.

[6] Tropp, J.A. and Gilbert A.C. (2006) Signal recovery from partial information via Orthogonal Matching Pursuit, submitted to *IEEE Trans. on Inf. Theory*.

[7] Tropp, J.A., Wakin, M.B., Duarte, M.F., Baron, D., and Baraniuk, R.G., (2006) Random filters for compressive sampling and reconstruction, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP*, Toulouse, France, May.

[8] Natarajan, B. K. (1995) Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24:227–234.

[9] Davis, G., Mallat, S., and Avellaneda, M. (1997) Greedy adaptive approximation. *J. Constr. Approx.*, 13:57–98.

[10] Chen, S.S., Donoho, D.L. and Saunders, M.A. (2001) Atomic decomposition by basis pursuit, *SIAM Review*, Volume 43, number 1, pages 129–59.

[11] Mallat, S. and Zhang, Z. (1993) Matching pursuit in a time-frequency dictionary, *IEEE Trans, on Signal Proc.*, Vol. 41, pp. 3397–3415.

[12] Pati, Y.C., Rezaiifar, R., and Krishnaprasad, P.S. (1993) Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition, *Proceedings of the 27 th Annual Asilomar Conference on Signals, Systems, and Computers.*

[13] Donoho, D.L. & Elad, M. (2002) Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ^1 minimization, *Proc. Nat. Aca. Sci.*, 100:2197–2202.

[14] Gribonval, R. & Nielsen, M. (2004) Sparse representations in unions of bases, *IEEE Trans. on Inf. Theory*, 49(12):3320–3325.

[15] Tropp, J.A. (2004) Greed is Good: Algorithmic results for sparse approximation. *IEEE Trans. on Inf. Theory*, Vol. 50(10):2231–2242

[16] Dhillon, I.S., Heath R.W.Jr. and Strohmer, T. (2005) Designing structured tight frames via alternating projection, *IEEE Trans. on Inf. Theory*, Vol. 51, pp. 188–209, January.