COMPRESSED SIGNAL RECONSTRUCTION USING THE CORRENTROPY INDUCED METRIC

Sohan Seth and José C. Príncipe

Computational NeuroEngineering Laboratory, University of Florida, Gainesville, USA
sohan@cnel.ufl.edu, principe@cnel.ufl.edu

ABSTRACT

Recovering a sparse signal from insufficient number of measurements has become a popular area of research under the name of Compressed Sensing or Compressive Sampling. The reconstruction algorithm of compressed sensing tries to find the sparsest vector (minimum l_0-norm) satisfying a series of linear constraints. However, l_0-norm minimization, being a NP hard problem is replaced by l_1-norm minimization with the cost of higher number of measurements in the sampling process. In this paper we propose to minimize an approximation of l_0-norm to reduce the required number of measurements. We use the recently introduced Correntropy Induced Metric (CIM) as an approximation of l_0-norm, which is also a novel application of CIM. We show that by reducing the kernel size appropriately we can approximate the l_0-norm, theoretically, with arbitary accuracy.

Index Terms— Compressed Sensing, Correntropy Induced Metric, l_0-norm, Gradient Descent.

1. INTRODUCTION

Recovering a sparse signal from insufficient number of measurements has become a popular area of research under the name of *compressed sensing* or *compressive sampling* (CS) [1][2][3]. In Nyquist sampling scheme a signal is sampled uniformly. After sampling, the signal is often compressed to reduce storage requirement. In compressive sampling, however, the idea is to sample the signal in a way that it is already compressed. To achieve this instant compression, CS takes advantage of the sparsity or compressibility of the signal; analogous to Nyquist sampling that exploits the band-limitedness of the signal. The sampling scheme of CS involves projecting the original signal onto a set of fewer number of bases, than the actual number of bases needed to represent the signal, and the reconstruction scheme of CS involves finding the *sparsest* signal (with minimum l_0-norm) satisfying a series of linear constraints. l_0-norm minimization, however, being a NP hard problem, is not solved. Instead its closest linear counterpart, l_1-norm, is minimized. l_1-norm minimization indeed achieves a near perfect solution as that of l_0-norm. However, the cost of having simpler solution is paid by increasing the number of measurements needed for exact reconstruction. It has been shown that the required number of measurements can be reduced if any l_p-norm with $0 < p < 1$ is minimized [4]. The required number of measurements become lesser as p is decreased. Minimizing l_p-norm is a nonconvex optimization problem. In this paper, we propose to solve an approximation of l_0-norm to reduce the required number of measurements.

This work was partially supported by NSF grant ECS-0601271. Sohan thanks Il Park and Weifeng Liu.

We use the *Correntropy Induced Metric* (CIM) as an approximation of l_0-norm [5]. The idea of CIM originates from the novel idea of *correntropy* which is a generalization of *correlation* [5]. The definition of correntropy allows us to induce a *Reproducing Kernel Hilbert Space* (RKHS), called VRKHS, corresponding to the input space. The correlation function in VRKHS is referred to as correntropy in the input space and the Euclidean distance measure in VRKHS is referred to as CIM in the input space. As the mapping between the original input space and VRKHS is nonlinear, CIM is a nonlinear distance measure in the original space. Due to the inherent nonlinearity, CIM saturates if two points are far apart in the input space. This feature makes CIM insensitive to outliers and makes it an appropriate choice for regression problems involving impulsive noise [5]. In this paper we discuss l_0-norm approximation as a novel application of CIM.

2. COMPRESSED SENSING

Let $\mathbf{s} \in \mathbb{R}^D$ be a real valued, finite length, one dimensional, discrete signal of length D. We represent the signal by a column vector $[\mathbf{s}]_{D\times 1}$. Suppose that this signal is sparse in a particular domain $\mathbf{\Psi} \in \mathbb{R}^{D\times D}$ i.e. $\mathbf{s}$ can be completely represented by only M ($M \ll D$) nonzero projections on a set of orthonormal bases $\mathbf{\Psi}$. If the representation of $\mathbf{s}$ in that sparse domain is $\boldsymbol{\theta} \in \mathbb{R}^D$ then $[\boldsymbol{\theta}]_{D\times 1}$ is a column vector with only M nonzero entries and $\mathbf{s} = \mathbf{\Psi}^T\boldsymbol{\theta}$. $[\mathbf{\Psi}]_{D\times D}$ is the sparsifying basis whose columns are the bases that captures the sparsity of the signal. $(.)^T$ is the transpose operation. If the signal is sparse in time domain then $\mathbf{\Psi} = \mathbf{I}$.

The sampling scheme of CS projects the signal $\mathbf{s}$ on a set of new bases $\mathbf{\Phi} \in \mathbb{R}^{D\times N}$, where $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are *incoherent* i.e. any column vector $\{\mathbf{\Psi}_{j,j=1,2,\ldots,D}\}$of $\mathbf{\Psi}$ can not be sparsely represented by the column vectors $\{\mathbf{\Phi}_{j,j=1,2,\ldots,N}\}$ of $\mathbf{\Phi}$ and vice versa [3]. The number of such bases (N) are much smaller compared to the actual number of bases (D). We express the sampling scheme as $\mathbb{R}^N \ni \mathbf{y} = \mathbf{\Phi}^T\mathbf{s}$ where $[\mathbf{y}]_{N\times 1}$ is a column vector of measurements and $[\mathbf{\Phi}]_{D\times N}$ is the measurement matrix whose columns are the measurement bases. The incoherence between the sparse bases and the measurement bases ensures that reconstruction is possible even from less number of measurements. A simple way to assure this property is to choose the elements of $\mathbf{\Phi}$ from Gaussian distribution i.e. $\mathbf{\Phi} \sim \mathbf{N}\left(0, \sigma^2\right)$. For instance, when the sparsity basis is $\mathbf{\Psi} = \mathbf{I}$ then it is easily visible that $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are incoherent. For such combination of measurement matrix and basis matrix it can be shown that $N \geq cM\log(D) \ll D$ number of measurements is sufficient to assure exact reconstruction with very high probability [1]. c is a constant. Another way of constructing $\mathbf{\Phi}$ is to choose the elements from Rademacher distribution. Rademacher distribution is a discrete probability distribution which has equal chance for either

1 or -1.

The sampling part assures that the reconstruction is possible but the inverse problem is ill posed as we have fewer number of measurements (N) than unknowns (D). This problem is tackled by using the sparsity of the signal [3]. The reconstruction algorithm try to find the sparse signal $\boldsymbol{\theta}$ instead of $\mathbf{s}$. The classical approach of solving such ill posed inverse problem is to find the minimum energy solution i.e. to minimize the least square cost function. Let $\boldsymbol{\Theta} = \boldsymbol{\Psi}\boldsymbol{\Phi}$ then the solution can be expressed as

$$\boldsymbol{\theta} = \arg\min \left\|\boldsymbol{\theta}'\right\|_2 \ such\ that\ \mathbf{y} = \boldsymbol{\Theta}^T\boldsymbol{\theta}' \tag{1}$$

(See Section 4 for the definition of $\|.\|_p$).Though this quadratic programming problem has a closed form solution, the solution unfortunately is not the sparsest one. The sparsest solution is guaranteed, however, if we minimize the cardinality of the solution. The solution in this case can be expressed as

$$\boldsymbol{\theta} = \arg\min \left\|\boldsymbol{\theta}'\right\|_0 \ such\ that\ \mathbf{y} = \boldsymbol{\Theta}^T\boldsymbol{\theta}' \tag{2}$$

But unfortunately this problem is NP hard and takes a long time to solve. Therefore instead of minimizing l_0-norm, its closest linear counterpart l_1-norm minimized. l_1-norm minimization is a convex optimization problem which can be solved using a linear programing method known as *basis pursuit* whose complexity is $O\left(D^3\right)$ [3]. The solution, therefore, is given by

$$\boldsymbol{\theta} = \arg\min \left\|\boldsymbol{\theta}'\right\|_1 \ such\ that\ \mathbf{y} = \boldsymbol{\Theta}^T\boldsymbol{\theta}' \tag{3}$$

Surprisingly the signal can be reconstructed with very high accuracy by solving the l_1 minimization problem provided we have sufficient number of measurements $N > cM\log D$ [1]. In practical cases, however, $N > 4M$ number of measurements gives sufficiently accurate result [2]. Recently Chartrand has shown that we can also achieve a near perfect solution by minimizing the l_p-norm ($0 < p < 1$). The sufficient number of measurements becomes lesser as we decrease p [4]. Minimizing l_p-norm with $0 < p < 1$ is, however, a nonconvex optimization problem and it becomes difficult to control as we decrease p. Chartrand has shown results for $p = 0.5$, after which the parameters for the gradient descent method become difficult to tune [4].

In this paper we reconstruct the signal by solving the following optimization problem

$$\boldsymbol{\theta} = \arg\min CIM\left(\boldsymbol{\theta}', 0\right) \ such\ that\ \mathbf{y} = \boldsymbol{\Theta}^T\boldsymbol{\theta}' \tag{4}$$

(See Section 3 for the definition of CIM). Since we use CIM as an approximation of l_0-norm, we expect to reduce the number of measurements further without degrading the reconstruction accuracy.

3. CORRENTROPY AND CIM

The correntropy of two vectors (or two one-dimensional discrete signals)

$$\begin{aligned} X &= [x_1, x_2, ..., x_N] \\ Y &= [y_1, y_2, ..., y_N] \end{aligned}$$

is defined as

$$V(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\kappa(x_i, y_i) \tag{5}$$

where $\kappa(.)$ is a reproducing kernel. We use the Gaussian kernel,

$$\kappa(x,y) = \kappa(x-y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{6}$$

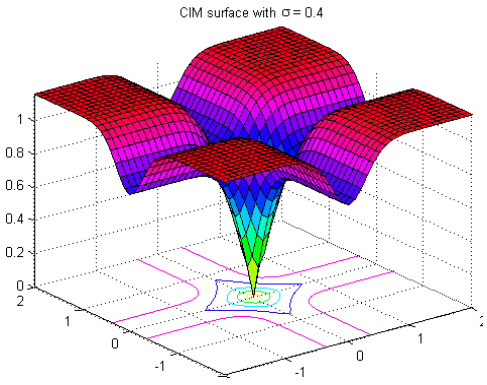

Fig. 1. CIM surface plot showing to distince regions; Euclidean (quadratic shaped region) and rectification (flat surface). The width of the convex Euclidean region is proportional to the kernel size.

where σ is the kernel size. $\kappa(.)$ satisfies the Mercer's theorem and therefore there exists a nonlinear transformation $\Phi : \mathbb{R} \to \mathbb{R}^{\mathbb{R}}$ from the input space to a RKHS $\mathbb{F}$ where the inner product produced equals the evaluation of the kernel i.e.

$$\langle \Phi(x), \Phi(y) \rangle_{\mathbb{F}} = \kappa(x,y) \tag{7}$$

We use this nonlinear transformation to map X and Y to

$$\begin{aligned} \widetilde{X} &= [\Phi(x_1), \Phi(x_2), ..., \Phi(x_N)] \\ \widetilde{Y} &= [\Phi(y_1), \Phi(y_2), ..., \Phi(y_N)] \end{aligned} \tag{8}$$

Then the Euclidean distance between $\widetilde{X}$ and $\widetilde{Y}$ in $\mathbb{F}$ is given by [5]

$$\left\{\left(\widetilde{X}-\widetilde{Y}\right)^T\left(\widetilde{X}-\widetilde{Y}\right)\right\}^{\frac{1}{2}} = \sqrt{2N}\left\{\kappa(0) - V(X,Y)\right\}^{\frac{1}{2}}$$

Ignoring the constant term $\sqrt{2N}$ in the expression, CIM is defined as

$$CIM(X,Y) = \left\{\kappa(0) - V(X,Y)\right\}^{\frac{1}{2}} \tag{9}$$

Due to its relation with correntropy $V(X,Y)$, this nonlinear metric is called the *correntropy* induced metric. CIM is a nonlinear metric in the input space. This metric divide the space in three regions named Euclidean region, Transition region and Rectification region [5]. In the Euclidean region CIM behaves as l_2-norm (convex function), in transition region CIM behaves as l_1-norm and in rectification region CIM behaves like l_0-norm (nonconvex function) (Figure 1). The width of the convex region is proportional to the kernel size. We will use this feature of CIM later.

4. CIM AS l_0-NORM APPROXIMATOR

The l_p-norm of a N dimensional vector $X = [x_1, x_2, ..., x_N]$ for any $0 < p < \infty$ is defined as

$$\|X\|_p = \left(\sum_{i=1}^{N}|x_i|^p\right)^{\frac{1}{p}} \tag{10}$$

l_0-norm and l_∞-norm are defined separately. In the limiting case $p \to 0$, l_0-norm (Referring l_0 as "norm" is a slight abuse of terminology), is defined as the number of nonzero entries in the vector i.e.

$$\|X\|_0 = card\{x_i : x_i \neq 0\} \tag{11}$$

where card is set cardinality [6].

Minimizing l_0-norm is a NP hard problem [6]. Therefore l_0-norm is often approximated by continuous functions. A popular approximation used by many authors is

$$\|X\|_0 \sim \sum_{i=1}^{N} \{1 - \exp(-\alpha |x_i|)\} \tag{12}$$

where the parameter α has to be chosen by the user. For practical purposes α is either set to some finite value like 5 or is increased slowly throughout the optimization process for better approximation [6]. Another approximation suggested by Weston et al. is given by

$$\|X\|_0 \sim \sum_{i=1}^{N} \log(|x_i| + \epsilon) \tag{13}$$

where ϵ is a small positive number. If X_0 is the solution achieved by minimizing (11) and X_l is the solution achieved by minimizing (13) then Weston et al. has shown that

$$\|X_l\|_0 < \|X_0\|_0 + O\left(\frac{1}{\ln \epsilon}\right) \tag{14}$$

provided, the absolute value of the nonzero entries of X are bounded below by a small positive number δ, i.e. by making $\epsilon \rightarrow 0$ we can get arbitrarily close to the l_0-norm solution. In this paper we propose to use CIM as an approximation of l_0-norm. Therefore, we approximate l_0-norm by

$$\begin{aligned} \|X\|_0 &\sim CIM(X,0) = \sqrt{\kappa(0) - \frac{1}{N}\sum_{i=1}^{N}\kappa(x_i,0)} \\ &= \sqrt{\frac{\kappa(0)}{N}\sum_{i=1}^{N}\left\{1 - \exp\left(-\frac{x_i^2}{2\sigma^2}\right)\right\}} \end{aligned} \tag{15}$$

We simplify the expression by removing the square root operator. Thus the approximation is given by

$$\|X\|_0 \sim CIM^2(X,0) = \frac{\kappa(0)}{N}\sum_{i=1}^{N}\left\{1 - \exp\left(-\frac{x_i^2}{2\sigma^2}\right)\right\} \tag{16}$$

It can be shown that if $|x_i| > \delta \; \forall i : x_i \neq 0$ then by making $\sigma \rightarrow 0$, we can get arbitrarily close to the l_0-norm. The proof is very similar to the one described by Weston et al. See the appendix for detailed derivation.

Notice that the definition of CIM does not restrict the kernel $\kappa(.)$ to Gaussian kernel. Another possible kernel that can be used is a Laplacian kernel given by

$$\kappa(x,y) = \kappa(x-y) = \frac{\alpha}{2}\exp(-\alpha|x-y|) \tag{17}$$

If we use this kernel then the approximation of l_0-norm takes the form of equation (12).

5. NUMERICAL RESULTS

We test the performance of l_1-norm minimization, l_p-norm minimization ($0 < p < 1$) and CIM minimization on a real valued sparse sequence $\mathbf{s}$. We use $D = 512$ and $M = 16$. The positions of the 16 nonzero entries were selected randomly from the available 512 positions with equal probability and the values at those points were generated from a zero mean, unity variance Gaussian distribution. The number of measurements N were varied from 32 $(2M)$ to 96 $(6M)$ with stepsize 4 $(M/4)$ i.e. $N \in \{32, 36, ..., 96\}$. For each values of N, 25 trials were performed. The reconstruction was considered successful if the l_2-norm of the error between the original sequence and the reconstructed sequence (say $\mathbf{s}_r$) is less than 10^{-3} i.e. successful reconstruction implies $\|\mathbf{s} - \mathbf{s}_r\|_2 < 10^{-3}$. For each N the probability of success was determined numerically by dividing the number of successful reconstructions by the total number of trials. We use the l_1-*magic* package available online for l_1-norm minimization [7]. For l_1-norm minimization, the elements of the measurement matrix were generated from a zero mean, unity variance Gaussian distribution and then the matrix was orthonormalized as shown in the l_1-magic package. For l_p-norm ($0 < p < 1$) and CIM minimization the elements of the measurement matrix were chosen from the Rademacher distribution as we observed better reconstruction with this. The parameters for l_1-norm minimization were set to the default values used in the examples shown in the package.

For CIM minimization, we use the constrained gradient projection method as described in [8]. We first compute the gradient vector

$$[\mathbf{g}]_{D\times 1} = \left[\frac{\partial CIM(\boldsymbol{\theta},0)}{\partial\theta_1}, \frac{\partial CIM(\boldsymbol{\theta},0)}{\partial\theta_2}, ..., \frac{\partial CIM(\boldsymbol{\theta},0)}{\partial\theta_D}\right] \tag{18}$$

where

$$\frac{\partial CIM(\boldsymbol{\theta},0)}{\partial\theta_i} = \frac{\kappa(0)\,\theta_i}{D\sigma^2}\exp\left(-\frac{\theta_i^2}{2\sigma^2}\right) \tag{19}$$

and then project the gradient onto the null space of $\boldsymbol{\Theta}^T$

$$\widetilde{\mathbf{g}} = \left[I - \boldsymbol{\Theta}\left(\boldsymbol{\Theta}^T\boldsymbol{\Theta}\right)^{-1}\boldsymbol{\Theta}^T\right]\mathbf{g} \tag{20}$$

where $[I]_{D\times D}$ is the identity matrix. We update $\boldsymbol{\theta}$ using

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\frac{\widetilde{\mathbf{g}}}{\|\widetilde{\mathbf{g}}\|_2} \tag{21}$$

where η is the learning rate parameter and t is the iteration number. $\boldsymbol{\theta}_0 = \left(\boldsymbol{\Theta}^T\right)^{-1}\mathbf{y}$ where the inverse is a pseudoinverse; i.e. the gradient descent process starts from the minimum energy solution of the linear constraint equations.

A very low kernel size creates local minima and makes the gradient descent process unstable. To tackle this problem we use the kernel annealing method as proposed in [9]. We vary the kernel size exponentially throughout the experiment starting from a large value ($\sigma_{\max} = 5$) to a small value ($\sigma_{\min} = 0.001$) in every 100 iterations using the function

$$\sigma = \sigma_{\max}\exp\left(-\frac{\beta t}{T}\right) + \sigma_{\min} \tag{22}$$

where T is the total number of iterations and t is the current iteration. The parameter β is the exponential decay rate. We choose $T = 10000$ and $\beta = 10$. Though l_0-norm is a nonconvex optimization problem, CIM minimization can be solved as a convex optimization problem by kernel annealing. We start with a large kernel size to ensure that the initial guess of the solution lies in the convex Euclidean region. As we go closer to the sparse solution we keep reducing the kernel size to ensure a l_0-norm solution. Though we decrease the kernel size we still expect to be in the convex region as the solution is sparse.

After changing the kernel size we adapt the best step size through a line search. We use the steepest descent method for line search. We select the stepsize η that gives the minimum $CIM(\boldsymbol{\theta}_{t+1}, 0)$.

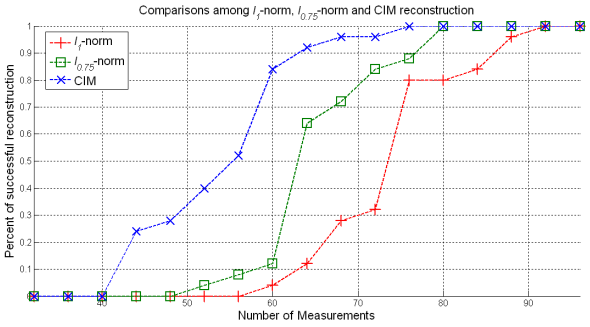

Fig. 2. Performance of l_1-norm minimization, $l_{0.75}$-norm minimization and CIM minimization in reconstructing compressed signal. The original signal length is $D = 512$ and has sparsity $M = 16$.

For l_p-norm minimization we select $p = 0.75$. This value is chosen as the gradient descent process in easier to control with $p > 0.5$ and the number of measurements needed for $p = 0.75$ is as low as that for $p = 0.5$ [4]. We use the same gradient descent method with same parameter values. We start the experiment with stepsize 1 and select the best stepsize η, using same line search method, that gives minimum $\|\boldsymbol{\theta}_{t+1}\|_{0.75}$ in every 100 iterations.

Figure 2 shows the performance of all three reconstruction schemes. The figure shows improvement in performance over both l_1-norm minimization and $l_{0.75}$-norm minimization when we use CIM minimization. We observe that under the described experimental settings CIM minimization requires around $4M$ number of measurements to give a 100% reconstruction probability whereas l_1-norm minimization and $l_{0.75}$-norm minimization require around $6M$ and $5M$ number of measurement respectively. Moreover we notice that exact reconstruction is possible even with as low as around $2M$ number of measurements when we use CIM minimization.

CIM as an approximate l_0-norm minimization performs better than both l_1-norm and $l_{0.75}$-norm minimization in terms of required number of measurements. But the computation involved in minimizing $l_{0.75}$-norm and CIM are nearly the same and are much higher that that of l_1-norm minimization. We use a simple gradient descent method and kernel annealing technique to minimize CIM. As mentioned earlier, CIM is a nonlinear function and it has many local minima. Therefore there still lies possibility for the gradient descent to converge in the local minima and to give improper result. Also the CIM approximation is valid when the values of the nonzero entries in the signal are greater than a threshold (See appendix). Therefore if the signal contains very low values then CIM minimization fails to reconstruct properly.

6. SUMMARY

In this paper we show a novel way to reconstruct in compressive sampling using also a novel metric the CIM to reduce the number of measurements. The future works include using other sophisticated optimization tools for CIM minimization and comparing CIM with other l_0-norm approximators.

A. PROOF OF CIM AS l_0-NORM APPROXIMATOR

Let the X_0 be the solution we get by minimizing the l_0-norm and X_l be the the solution we achieve by minimizing CIM. Then

$$CIM\left(X_l, 0\right) \leq CIM\left(X_0, 0\right)$$

$$\Rightarrow \sum_{j=1}^{N} \exp\left(-\frac{(X_l)_j^2}{2\sigma^2}\right) \geq \sum_{j=1}^{N} \exp\left(-\frac{(X_0)_j^2}{2\sigma^2}\right)$$

$$\Rightarrow \sum_{j=1,\,(X_l)_j=0}^{N} \exp\left(-\frac{(X_l)_j^2}{2\sigma^2}\right) + \sum_{j=1,\,(X_l)_j\neq 0}^{N} \exp\left(-\frac{(X_l)_j^2}{2\sigma^2}\right)$$

$$\geq \sum_{j=1,\,(X_0)_j=0}^{N} \exp\left(-\frac{(X_0)_j^2}{2\sigma^2}\right) + \sum_{j=1,\,(X_0)_j\neq 0}^{N} \exp\left(-\frac{(X_0)_j^2}{2\sigma^2}\right)$$

$$\Rightarrow \left(N - \|X_l\|_0\right) + \sum_{j=1,\,(X_l)_j\neq 0}^{N} \exp\left(-\frac{(X_l)_j^2}{2\sigma^2}\right)$$

$$\geq \left(N - \|X_0\|_0\right) + \sum_{j=1,\,(X_0)_j\neq 0}^{N} \exp\left(-\frac{(X_0)_j^2}{2\sigma^2}\right)$$

$$\Rightarrow \|X_l\|_0 - \sum_{j=1,\,(X_l)_j\neq 0}^{N} \exp\left(-\frac{(X_l)_j^2}{2\sigma^2}\right)$$

$$\leq \|X_0\|_0 - \sum_{j=1,\,(X_0)_j\neq 0}^{N} \exp\left(-\frac{(X_0)_j^2}{2\sigma^2}\right)$$

$$\Rightarrow \|X_l\|_0 - \|X_0\|_0 \leq \sum_{j=1,\,(X_l)_j\neq 0}^{N} \exp\left(-\frac{(X_l)_j^2}{2\sigma^2}\right) - \sum_{j=1,\,(X_0)_j\neq 0}^{N} \exp\left(-\frac{(X_0)_j^2}{2\sigma^2}\right)$$

Now if we assume that $\left|(X_0)_{j,(X_0)_j\neq 0}\right|, \left|(X_l)_{j,(X_l)_j\neq 0}\right| > \delta$ where δ is a small positive number then we can choose a σ to make the right hand side of the equation arbitrarily close to zero. Therefore we arrive at the condition given by $\|X_0\|_0 \leq \|X_l\|_0 \leq \|X_0\|_0 + \nu$; $\nu > 0$ where ν is small positive number.

B. REFERENCES

[1] E. J. Candès, "Compressive sampling," in *Proc. International Congress of Mathematics*, Madrid, Spain, 2006, pp. 1433–1452.

[2] D. Donoho and Y. Tsaig, "Extensions of compressed sensing," *Signal Processing*, no. 3, pp. 533–548, March 2006.

[3] R. Baraniuk, "A lecture on compressive sensing," *IEEE Signal Processing Magazine*, July 2007.

[4] R. Chartrand, "Exact reconstructions of sparse signals via non-convex minimization," *IEEE Signal Proc. Lett.*, 2007.

[5] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-gaussian signal processing," *IEEE Tran. on Signal Processing*, 2007.

[6] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of zero-norm with linear models and kernel methods," *JMLR special Issue on Variable and Feature Selection*, pp. 1439–1461, 2002.

[7] [Online]. Available: http://www.acm.caltech.edu/l1magic/

[8] R. Horie and E. Aiyoshi, "Variable metric gradient projection method and replicator equation," in *IEEE SMC '99 Conference Proceedings*, Tokyo, Japan, 1999, pp. 515–520.

[9] D. Erdogmus, "Information theoretic learning: Renyi's entropy and its applications to adaptive system training," Ph.D. dissertation, Univ. of Florida, Gainesville, 2002. [Online]. Available: www.cnel.ufl.edu