Non-Convex P-norm Projection for Robust Sparsity [*]

Mithun Das Gupta
Ricoh Innovations Pvt. Ltd.
Bangalore, India
mgupta@ripl.ricoh.com

Sanjeev Kumar
Qualcomm
Bangalore, India
sanjeevk@qualcomm.com

Abstract

In this paper, we investigate the properties of L_p norm ($p \leq 1$) within a projection framework. We start with the KKT equations of the non-linear optimization problem and then use its key properties to arrive at an algorithm for L_p norm projection on the non-negative simplex. We compare with L_1 projection which needs prior knowledge of the true norm, as well as hard thresholding based sparsification proposed in recent compressed sensing literature. We show performance improvements compared to these techniques across different vision applications.

1. Introduction

Many successful algorithms for inverse problems rely on regularization for improved reconstruction by enforcing conformance to a-priori statistical properties of data. Assumptions of Laplacian, Gaussian and Generalized Gaussian priors result in L_1 (Lasso), L_2 (Ridge regression) and L_p norm regularization. In natural images, gradients, general pixel differences and transform domain coefficients can simultaneously have large peak near zero (due to smooth areas) and very heavy tail (due to presence of edges and object boundaries). The wavelet transform is tremendously popular in the signal and image processing communities, due in large part to its ability to provide parsimonious representations for signals that are smooth away from isolated discontinuities. The ability to construct low-dimensional, accurate approximations make wavelets particularly useful for image compression [23] and restoration [6]. If we compute the pdf for the empirical distribution for the wavelet coefficient magnitude and compare with the pdf's obtained by an exponential model with varying λ, we find that $\lambda < 1$ models the empirical distribution way better than Laplacian distribution ($\lambda = 1$). This affect is shown in Fig. 1, where $\lambda = 0.2$ matches the empirical distribution most closely for the Lena image.

Accurate modeling of such distributions leads to L_p norm regularization with $p < 1$. Non-convexity of L_p norm poses a major hindrance in efficient solution to resulting optimization problem. Solving an under-determined system of equations under non-convex L_p norm prior has been known to be NP hard. As a practical alternative, iterative schemes such as iteratively re-weighted least square (IRLS) [5] has been used as an approximation algorithm. However, some problems, e.g. transform domain image denoising, can be cast as projection of a vector on norm ball and does not involve solving an under-determined system of linear equations. Such problems can directly benefit from efficient projection on non-convex L_p norm ball. For other problems, iterative schemes such as gradient-projection can be devised with projection on non-convex norm ball as building block, which of course looses guarantee of global optimality, but presents a potentially attractive alternative to IRLS and other competing methods. In this work we look into least square minimization as the primary problem ($f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{v}\|_2^2$). In the literature, two complementary formulations have been used to incorporate norm-based priors:

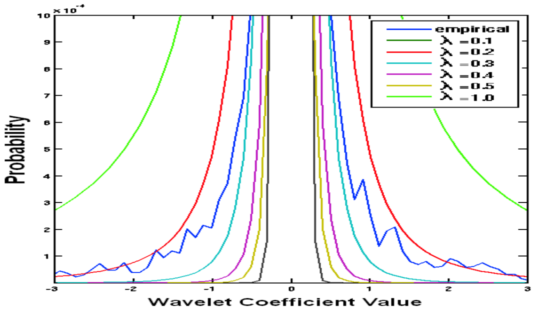

Figure 1. Empirical and model pdf.

1. projection formulation $\min f(\mathbf{x})\ s.t. \|x\| \leq z$ where norm prior term appears as constraint, and
2. penalty function formulation $\min f(\mathbf{x}) + \theta\|x\|$ where

[*] Both the authors are equal contributors in this work. This work was done before both the authors joined their current companies. Please use authors' independent email id's *mithundasgupta@gmail.com*, *sanjeev.kumar@gmail.com* for paper related queries.

norm prior term is part of objective function.

Regularization parameters z and θ serve an equivalent purpose of controlling the relative significances of data and prior terms. When L_2 norm is used for data term and L_1 norm is used for prior term in penalty function formulation, regularization path is piecewise linear [19] and every point on it corresponds to unique values of regularization parameter θ and z, thus establishing a one-to-one correspondence between θ and z. Therefore, in principle, an algorithm for any one formulation can be used to find solution of the other formulation by using bisection over the regularization parameter. Such equivalence, however, breaks down in the case of non-convex norm prior. Fig. 2 (left) shows that norm of the optimal solution exhibits vertical discontinuity at some values of regularization parameter θ for the penalty function formulation. Range of norm values corresponding to such discontinuities are never attained by optimal solution of penalty function formulation. In contrast, optimal solution of projection formulation exhibits continuous decrease of projection error with increasing z as shown in Fig. 2 (right). Consequently, for certain range of values of z, penalty function formulation can not emulate the behavior of projection formulation. Projection function formulation can be considered a more general problem with penalty function formulation as a special case.

1.1. KKT conditions and combinatorial explosion

Consider an arbitrary algorithm $\mathcal{A}$ trying to solve

$$\begin{aligned} &\min \quad f(\mathbf{x}) \\ &\texttt{s.t.} \quad g_i(\mathbf{x}) \leq 0 \; \forall i \in \mathbb{N}_n, \quad h_i(\mathbf{x}) = 0 \; \forall i \in \mathbb{N}_m \end{aligned} \tag{1}$$

where $\mathbb{N}_k = \{1, 2, \ldots, k\}$. Lagrangian for (1) is given by,

$$\mathcal{L}(\mathbf{x}, \alpha, \beta) = f(\mathbf{x}) + \sum_{i=1}^{p} \alpha_i g_i(\mathbf{x}) + \sum_{i=1}^{q} \beta_i h_i(\mathbf{x}) \tag{2}$$

Assuming appropriate regularization conditions are met, necessary conditions for local optimality are given by KKT first order condition $\frac{\partial \mathcal{L}}{\partial \mathbf{x}} = 0$, and

$$\underbrace{g_i(\mathbf{x}) \leq 0}_{a}, \; \underbrace{h_i(\mathbf{x}) = 0}_{b}, \; \underbrace{\alpha_i \geq 0}_{c}, \; \underbrace{\alpha_i g_i(\mathbf{x}) = 0}_{d}, \quad \forall i \in \mathbb{N}_n \; or \; \mathbb{N}_m \tag{3}$$

Let us further assume that $\mathcal{A}$ attempts to solve the problem by exhaustive search of solution space of KKT system of equations. Even for convenient functional forms of f, g_i, h_i, there is an inherent combinatorial branching involved in complementary slackness condition Eq. 3(d). For ease of future reference we refer to it as complementary slackness condition (CSC) branching stage. In worst case, CSC could lead to exponential (2^n) number of systems of equations. For $\mathcal{A}$ to be potentially polynomial time, it must avoid combinatorial explosion at CSC. Note that this necessity is only for family of algorithms which exhaustively search the solution space of KKT system of equations, and does not apply to other methods such as interior point based techniques.

Once $\mathcal{A}$ has branched over all CSC possibilities, e.g. in an outer loop, it will need to solve multiple instances of a reduced KKT system of equations (without CSC condition). In the simplest of cases, this reduced system of equations could be a system of full rank linear equations and vacuously satisfied inequalities, immediately yielding a solution. However, in more general situations e.g. quadratic constraints, corresponding elimination step involves further branching. We refer to it as Elimination in Simultaneous Equation (ESE) branching. For $\mathcal{A}$ to be polynomial time, it must avoid combinatorial explosion at ESE as well as CSC.

Once $\mathcal{A}$ has branched over all ESE branching possibilities, e.g. in a middle loop, it will need to solve multiple instances of a single equation in a small number (may be just one) of independent variables. We refer to this stage as Roots of Single Equation (RSE) stage. In general, RSE can also have exponential complexity. In present work, we show that for globally optimal non-convex Lp norm projection problem, number of CSC and ESE branches can be restricted to polynomial. Furthermore, we propose a conjecture, which guarantees polynomial time complexity for RSE branch as well, thus making complete problem polynomial time. Even if conjecture fails for very stringent norms $p \leq 0.2$, RSE stage, being a one dimensional root finding problem, is conducive for generic global optimization methods such as branch and bound [22, 3]. This makes complete problem (reducible to polynomial number of RSE stages) efficiently solvable by branch and bound.

2. Lp Norm Projection for non-negative data

L_p norm projection problem is given by,

$$\begin{aligned} &\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x} - \mathbf{v}\|_2^2, \\ &\texttt{s.t.} \; \sum_{j=1}^{n} |x_j|^p \leq z, \quad \texttt{for} \; 0 < p \leq 1 \end{aligned} \tag{4}$$

Except for the trivial case, when $\mathbf{v}$ lies inside norm ball and optimal solution is $\mathbf{x} = \mathbf{v}$, optimal solution of above problem will lie at boundary of norm ball. This can be proved by contradiction, by joining putative solution inside norm ball with $\mathbf{v}$ by a line segment and showing that intersection of this line segment with norm ball will result in better solution. Based on Lemma 3 from [7], we can work with the absolute values of the elements of $\mathbf{v}$ and add the sign to the projection $\mathbf{x}$ later such that $v_i x_i > 0$. This simplification adds an additional set of constraints for the non-negativity of the projections themselves. For now let

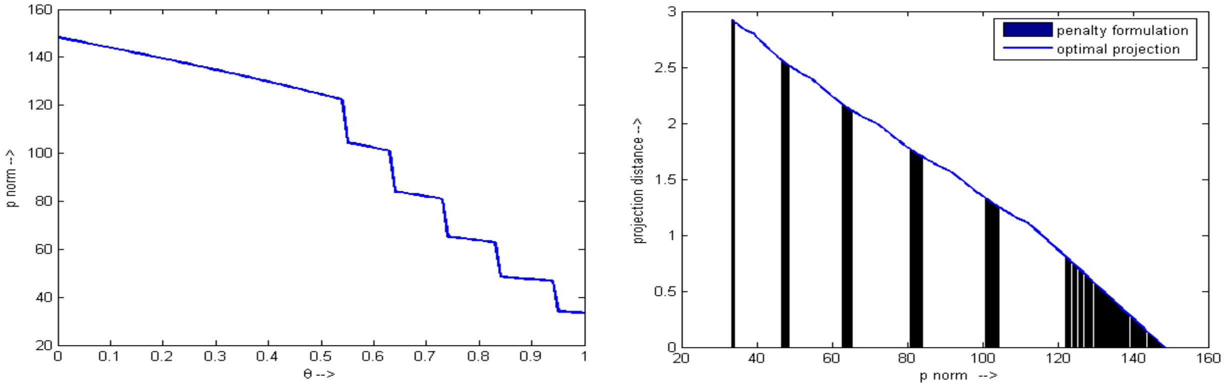

Figure 2. Left: norm of solution vs regularization parameter θ in penalty function formulation. Decrease of norm exhibits vertical discontinuity and some norm values are never attained by regularization path. Right: Projection distance $\|\mathbf{x}-\mathbf{v}\|_2^2$ vs norm of optimal solution for penalty function formulation (black bars) and projection formulation (blue curve). As regularization parameter z increases optimal solution $\mathbf{x}$ for projection formulation continuously gets closer to target $\mathbf{v}$. However corresponding decrease is discontinuous in case of penalty function formulation.

us assume that v_i's are all positive so that the notations are simpler to follow. For data with mixed sign we can remove all the signs, solve the optimization problem and plug back the signs again later. The simplified problem we investigate is defined as: given $\mathbf{v} \in \mathbb{R}^{n+}$, solve

$$\min_{\mathbf{x}} \tfrac{1}{2}\|\mathbf{x}-\mathbf{v}\|_2^2, \ \texttt{s.t.} \ \sum_{j=1}^{n} |x_j|^p \le z, \qquad (5)$$
$$\text{for } 0 < p \le 1, \ 0 \le x_i, \ \forall i = \{1, 2, \dots, n\}$$

Duchi et al. [7] present a solution for the case where $p = 1$. This is a specific case of the method investigated in this paper. At $p = 1$ the problem is convex and hence well studied [13, 21]. We argue that a generic p-norm constraint provides greater control over the amount of sparsity obtained in the projected domain albeit with additional non-linearity in its formulation. [14] introduces a similar system, but does not provide a generic framework. Only $p = 1/2, 2/3$ are explored since analytic solutions can be obtained for these systems. Our next proposition asserts that ordering of components of $\mathbf{v}$ determines ordering of components of optimal solution $\mathbf{x}$.

Proposition 2.1. *If $v_i > v_j$, then in optimal solution $x_i \ge x_j$.*

Proof: If $x_i < x_j$, then swapping x_i and x_j can be shown to result in a better solution.□

We note that as a consequence of Prop. 2.1, Lemma 1. from [7] restated below for ease of reference, still holds for this problem.

Proposition 2.2. *If $v_i > v_j$, and in optimal solution $x_i = 0$, then $x_j = 0$.*

An important consequence of Prop. 2.2 is that, combinatorial explosion for CSC can be avoided. More specifically, number of CSC branches will be upper bounded by n, the dimensionality of $\mathbf{v}$. The Lagrangian for the above optimization problem can be written as

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\zeta}, \theta) = \frac{1}{2}\|\mathbf{x}-\mathbf{v}\|_2^2 + \theta\Big(\sum_{j=1}^{n} x_i^p - z\Big) - \boldsymbol{\zeta}.\mathbf{x} \qquad (6)$$

Eq.(6) is not differentiable whenever, at least one of x_i's is zero. This precludes unconditional use of first order optimality as a KKT necessary condition. When at least one of x_i's *is* zero, consideration of local geometry of constraint surfaces reveals that surface normal vector for $x_i \ge 0$, and that for p-norm constraint become linearly dependent, thus violating the linearly independent constraint qualification (LICQ) criterion [18]. Hence, we adopt the more fundamental necessary condition that for the feasibility of local optimality at a point, it must be impossible to decrease objective function further by any local movement remaining feasible with the constraints. Whenever any x_i is zero and norm constraint is satisfied, any local movement with nonzero component along x_i will result in constraint violation. This leads us to following modified optimality condition:

$\mathbf{x}$ is optimal only if, for all the non-zero components $\{x_i | i \in \mathbb{N}_{non\ zero} \subset \mathbb{N}_n\}$, $\sum x_i^p = z$ and

$$x_i - v_i + p\theta x_i^{p-1} = 0 \qquad (7)$$

The above condition is the starting point of our analysis. We plot the representative curves for different values of v_i but fixed p and θ (Fig. 3 left). We observe that the whole plot shifts down with increasing v_i. For some value of v_i, the curve touches the horizontal line at 0 at *one* point (tangent), indicating that (7) has one unique zero. We denote this as $v_{i_t}(p, \theta)$, since this value depends on p and θ. For all values of $v_i > v_{i_t}(p, \theta)$, there are two roots for (7) denoted as x_i^l and x_i^r.

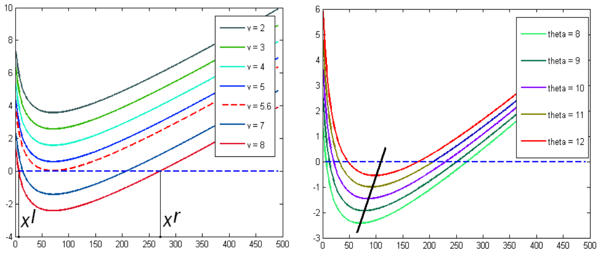

Figure 3. Plot for $x_i - v_i + p\theta x_i^{p-1}$. Left: Changing v_i for $p = 0.6$, $\theta = 8$. The curves move vertically down for increasing v_i. For the dotted red curve ($v_i = 5.6$) the horizontal line at 0 is a tangent. The two possible solutions for $v_i = 8$ are denoted in the figure as x_i^l and x_i^r. Right: Changing θ for $p = 0.6$, $v_i = 8$. The solid black line joins the minimum points for the different curves.

Proposition 2.3. *The left root x_k^l corresponding to any v_k is NOT part of the optimal solution for (5), except possibly for the smallest v_i among those with non-zero optimal projection x_i.*

Proof: We prove the proposition in two stages. In the first stage, we note that since x_i^l moves towards left with increasing v_i, in optimal solution, two distinct entries v_i and v_j can not have the same root $x_i^l = x_j^l$ as it will violate Prop. 2.1. This means at most one v_i can have corresponding left solution at x_i^l. In second stage of proof, we note that since, irrespective of ordering of v_i's, x_i^r for any v_i is larger than x_j^l for all other v_j's, $j \neq i$. Hence, any v_i with corresponding solution at x_i^l must be the smallest v_i among those with non-zero x_i, otherwise Prop. 2.1 would again be violated. □

An important consequence of Prop. 2.3 is that, combinatorial explosion of ESE can be avoided, since the number of ESE branches are upper bounded by 2. One branch corresponds to solution where *smallest* non-zero x_i corresponds to the left root and all other x_j's correspond to the right roots. The other branch corresponds to solution where all non-zero x_i's correspond to right root. We denote the two corresponding solutions as $\mathbf{x}_L$ and $\mathbf{x}_R$ respectively.

2.1. Algorithm

Based on aforementioned propositions, we present an outline of the proposed method in Algorithm 1. RSE$_L$ and RSE$_R$ represent RSE stage operations corresponding to choice of which root (left or right) is used for different equations in KKT system.

As can be seen from Fig. 3 (left), for all v_i and θ, the curve has a unique minimum point, which can be shown to be

$$x_{i_{min}} = [p(1-p)\theta]^{\frac{1}{2-p}} \tag{8}$$

Next we evaluate (7), but for different values of θ, as shown in Fig. 3 (right). The entire curve goes up and right

Algorithm 1 Algorithm outline for projection on L^p norm ball from positive orthant

```
REQUIRE $\mathbf{v} \in \mathbb{R}^{n+}$, $z > 0$, $p \in (0,1)$
$\mathbf{v} \leftarrow$ sort($\mathbf{v}$) /* decreasing order */
$\mathbf{x}_{opt} \leftarrow 0$
$OBJ_{opt} \leftarrow \infty$
for $\rho$ = 1 to n /* CSC branching */ do
    for j = 1 to 2 /* ESE branching */ do
        if j == 1 /* first ESE branch: smallest non-zero $x_i$ comes from left root */ then
            $X_L \leftarrow$ RSE$_L(v(1:\rho), z, p)$ (Eq. 7 left root)
        else if j == 2 /* second ESE branch: all $x_i$s come from right root */ then
            $X_R \leftarrow$ RSE$_R(v(1:\rho), z, p)$ (Eq. 7 right root)
        end if
    end for
    $\mathbf{x} \leftarrow \text{argmin}_{\mathbf{x}(1:\rho) \in X_L \cup X_R, \mathbf{x}(\rho+1:n)=0} \|(\mathbf{x}-\mathbf{v})\|_2$
    if $\|(\mathbf{x}-\mathbf{v})\|_2 < OBJ_{opt}$ then
        $OBJ_{opt} \leftarrow \|(\mathbf{x}-\mathbf{v})\|_2$
        $\mathbf{x}_{opt} \leftarrow \mathbf{x}$
    end if
end for
reorder $\mathbf{x}_{opt}$ according to initial sorting of $\mathbf{v}$
return $\mathbf{x}_{opt}$
```

with increasing θ, for constant v_i and p. The above observations can be combined to draw the following conclusion.

Proposition 2.4. *For every v_i there exists a θ (and corresponding x_i), given p, such that Eq. 7 has a unique 0 (tangent). The values are given by*

$$\theta_{tan}(v_i, p) = \frac{v_i^{2-p}}{p(1-p)[\frac{1}{1-p}+1]^{2-p}}, \tag{9}$$

$$x_{i_{tan}}(v,p) = \frac{1-p}{2-p} v_i \tag{10}$$

Proof: Substituting (8) in (7) and equating to 0 gives the expression for θ_{tan}. Substituting it back into (8) gives expression for $x_{i_{tan}}$. □

To explore the behavior of the two ESE branches we chose a random $\mathbf{v}$, and set a suitable norm such that none of the elements of its projection $\mathbf{x}$ can be zero. This allows us to study the behavior of the two solutions $\mathbf{x}_L$ and $\mathbf{x}_R$. We find $\theta_{tangent}$ corresponding to v_s, the smallest element of $\mathbf{v}$ and then go on reducing it until zero. This leads to the curves shown in Fig. 4. For $p \geq 0.5$ the nature of the curves obtained remains same. But for more stringent sparsity requirements $p < 0.5$ and higher dimensional problems the solution norm for $\mathbf{x}_L$ can intersect the norm constraint line multiple times as shown in Fig. 4 (right). This observation leads to the following conjecture:

Conjecture: For some integer r, the r^{th} derivative of the norm curve for $\mathbf{x}_L$ against θ (Fig. 4 right, blue curve) has only one zero crossing.

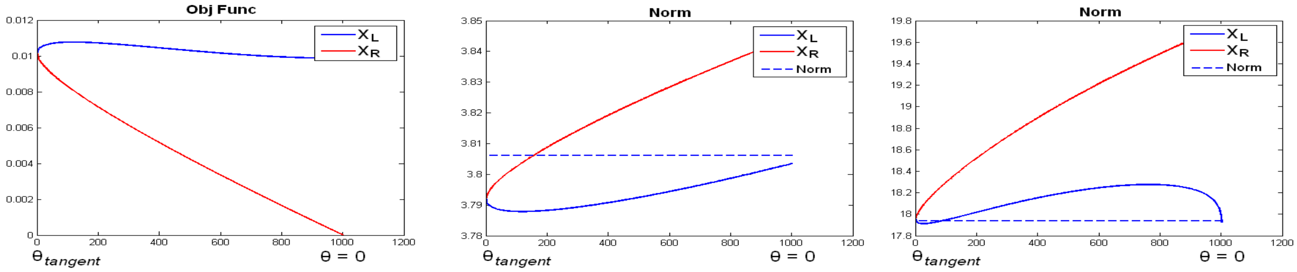

Figure 4. Left: Plot of objective value for $\mathbf{x}_L$ (blue) and $\mathbf{x}_R$ (red) with changing θ, with $p = 0.6$. Center: similar curves for the norms of the solutions. The horizontal axis denotes decreasing θ from $\theta_{tangent}$ to 0 in 1000 steps. Right: for higher dimensions and stringent $p = 0.3$ norm value wrt $\mathbf{x}_L$ touches the norm constraint line at multiple points.

Empirically we found that the 2nd derivative has one zero crossing, leading to maximum 3 zero crossings for the solution norm, as shown in Fig. 4 (right). In such cases, plot consists of one (initial) convex segment, and one (later) concave segment concatenated together. The location of the zero crossing can be found using piecewise binary search. We found that the conjecture fails in a few trials, less than 1%, depending on value of p (very small), and distribution of the elements of $\mathbf{v}$. We could still argue that our algorithm based on the above conjecture, will give us globally optimal solution with very high probability. A similar argument is proposed by Bredies et al. [1] (Sec. 4) where they claim that lp penalties converge to a solution whenever one of two theorems are applicable.

3. General sparse coding problem

The above developments were all concentrated on solving a projection problem over separable variables. Now we look into the more generic problem,

$$\tilde{\mathbf{w}} \quad = \quad \arg\min_{\mathbf{w}} ||\mathbf{Y} - \mathbf{A}\mathbf{w}||^2 + \beta|\boldsymbol{\phi}^{\mathbf{T}}\mathbf{w}|_p \qquad (11)$$

Using the half-quadratic penalty method [11, 12, 24], we now introduce the auxiliary vector $\mathbf{x}$, such that

$$\tilde{\mathbf{w}} = \arg\min_{\mathbf{w}} ||\mathbf{Y} - \mathbf{A}\mathbf{w}||^2 + \frac{\zeta}{2}\|\mathbf{x} - \boldsymbol{\phi}^{\mathbf{T}}\mathbf{w}\|_2^2 + \beta|\mathbf{x}|_p \ (12)$$

This equivalent formulation decouples the non-linear norm term from the principal variable with the additional property that it is exactly same as the previous formulation when $\zeta \rightarrow \infty$. This can now be subdivided into two parts

$$\tilde{\mathbf{w}}_s = \arg\min_{\mathbf{w}} ||\mathbf{Y} - \mathbf{A}\mathbf{w}||^2 + \tfrac{\zeta}{2}\|\mathbf{x} - \boldsymbol{\phi}^{\mathbf{T}}\mathbf{w}\|_2^2 \ (13)$$

$$\mathbf{x}_s = \arg\min_{\mathbf{x}} \tfrac{\zeta}{2}\|\mathbf{x} - \boldsymbol{\phi}^{\mathbf{T}}\mathbf{w}\|_2^2 + \beta|\mathbf{x}|_p \qquad (14)$$

Alternating solution for the two subproblems leads to the solution of the the original system (Eq. 11) as $\zeta \rightarrow \infty$.

4. Experiments

4.1. Comparison against L_1 norm minimization techniques

In this section we describe comparative experiments with other algorithms, primarily with algorithms which produce sparsity in otherwise more established methods, specifically, projection onto convex sets (POCS) [2]. We compare against alternate projection algorithms proposed by Candes and Romburg [4](L_1POCS) and Lu Gan [8](KPOCS) where the projection alternates between a POCS projection and a sparsifying projections which typically bounds the L_1 norm of the projected vector. The problem is the reconstruction of $\mathbf{x}$ from limited measurements denoted as $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$, where $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^N$, and $m < N$. Let the original signal be $\hat{\mathbf{x}}$. The reconstruction is guaranteed if $\hat{\mathbf{x}}$ is K sparse and $\mathbf{\Phi}$ follows the restrictive isometry property (RIP). The solution of sparse POCS lies in the intersection between the hyperplane $P = \{\mathbf{x} : \mathbf{\Phi}\mathbf{x} = \mathbf{y}\}$ and the L_1 ball B of radius $||\mathbf{x}||_1$. If the exact norm of the original signal is known, i.e. $||\hat{\mathbf{x}}||_1$, then L_1POCS iteratively projects onto the convex sets P and B to find the common intersection point (i.e. recover $\mathbf{x}$ from $\mathbf{y}$). In the absence of the knowledge of the true norm of the unknown vector, L_1POCS generates a sparse approximation of the original vector.

Jacques[1] proposed a simplification of the technique proposed by Lu Gan [8]. A hard thresholding step, keeping the K largest components of a vector is used as the second projection step. The principle difference in the two schemes is that in [8], the extra information is the sparsity of x (as in CoSaMP [17]), whereas in [4] the expected L_1 norm of the original signal is required. A similar method to KPOCS called gradient descent with sparsification (GraDes) has been independently proposed by Garg and Khandelwal [9]. We replaced the norm projection scheme in these two techniques by our method (L_pPOCS). The measurement matrix $\mathbf{\Phi}$ is a Gaussian random matrix. The parameter $p = 0.6$ if not stated otherwise and we keep the norm constraint to be

[1] http://nuit-blanche.blogspot.com/2008/04/monday-morning-algorithm-16-s-pocs.html

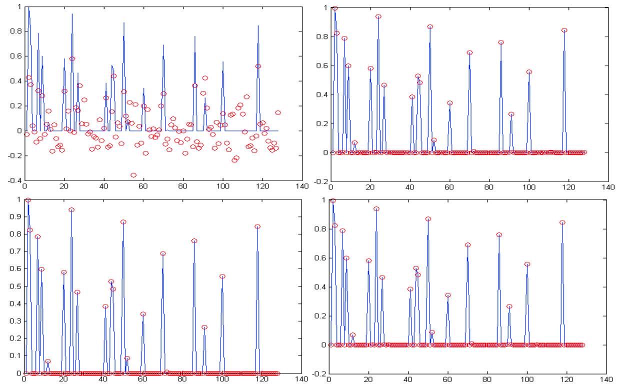

Figure 5. Plot of reconstruction from sparse measurements. Blue curve is the true data. Red dots are the reconstruction points. Left to right: L_1POCS with $\|\mathbf{y}\|_1$ estimate, L_1POCS with true $\|\hat{\mathbf{x}}\|_1$ estimate, KPOCS with $K = 3.5 * m$, L_pPOCS with $p = 0.6$ and norm constraint equal to $\|\mathbf{y}\|_1$. The experiment reports result when there was no additional noise introduced on $\mathbf{y}$. Problem specifics are: $N = 128$, $m = 20$, $K = 3.5 * m$.

$\|\mathbf{y}\|_1$. The norm parameter p works analogous to the forced sparsity K. Qualitative results for the noiseless case are shown in Fig. 5. Reducing the value of p generates sparser results. As we introduce more noise all the methods start to deteriorate. This effect is most pronounced in the KPOCS technique with fixed K.

4.2. Comparison with Iterative Least Squares (IRLS) and similar techniques

We compare against iterative least squares based algorithms to establish the general acceptability of our technique. Note that several authors have used the f-measure criterion to identify the merits of sparse solutions defined as

$$f-measure = 2\frac{\text{supp}(\mathbf{x}^{true}) \cap \text{supp}(\mathbf{x}^{solve})}{\text{supp}(\mathbf{x}^{true}) + \text{supp}(\mathbf{x}^{solve})} \qquad (15)$$

where supp$((x))$ is the support of the vector $\mathbf{x}$. f-measure closure to 1 means better identification of the true support of the unknown vector. Gasso et al. [10] introduced a difference of convex functions (DC) based method to tackle a similar problem. They show improvements over the performance of the IRLS method proposed by Saab et al. [20]. We perform similar experiments, to the ones reported by Gasso et al. [10], with the measurement matrix being a Gaussian random matrix with $n = 128$ rows and $d = 256$ columns. Note that due to step 5 (branching over the entire support of $\mathbf{v}$) in Algo. 1 our method almost always finds the true support of the solution vector $\mathbf{x}$, except for very stringent norm requirements for $p \leq 0.05$ or for very noisy data. This can be easily observed from Fig. 6. Also note that the sorting step in our method enables us to solve the original non-linear problem, rather than the ϵ padded systems which are normally used in IRLS based techniques.

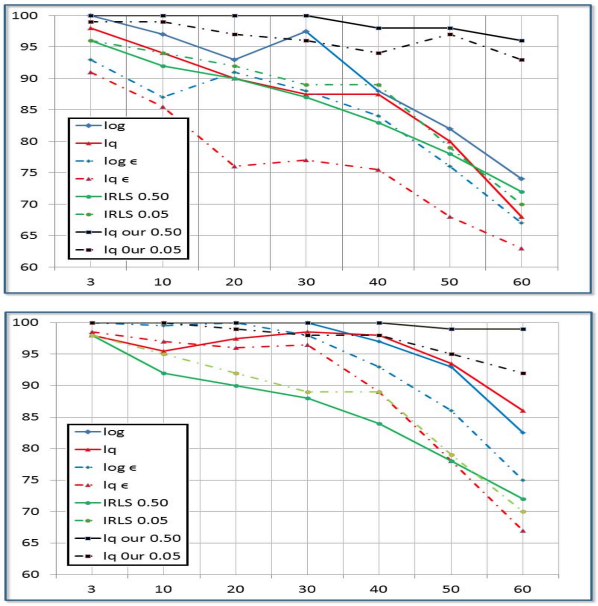

Figure 6. Comparing F-measure performance of different penalty functions with respect to the number of active elements. F-measure of the compared algorithms. Log, lq, Log ϵ and lq ϵ are same as defined in Gasso et al. [10]. IRLS is [20] the number next to it denotes the value of the norm constraint q, Lq our denotes the proposed method. Left: $SNR = 10db$, right: $SNR = 30db$.

4.3. Image denoising with L_p norm projection

In this section we present an alternative formulation to p-norm projection problem which can be motivated by considering MAP estimation of a signal $\mathbf{x} \in \mathbb{R}^n$ with component wise Generalized Gaussian prior distribution $p(x_i) = \frac{1}{2\alpha\Gamma(1+1/\beta)}exp(-(|x_i|/\alpha)^\beta)$, from its noisy version $\mathbf{v}$, corrupted by an additive noise $\mathbf{e} \in \mathbb{R}^n | e_i \sim \mathcal{N}(0, \sigma^2)$. Applying Bayes rule, $p(\mathbf{x}|\mathbf{v}) \propto p(\mathbf{x})p(\mathbf{v}|\mathbf{x}) = p(\mathbf{x})p(\mathbf{e})$, corresponding negative log likelihood function in given by,

$$\mathfrak{L}(\mathbf{x}) \quad \propto \quad \frac{\sum_{i=1}^n |x_i|^\beta}{\alpha^\beta} + \frac{\sum_{i=1}^n (x_i - v_i)^2}{2\sigma^2} \qquad (16)$$

$$\propto \quad \sum_{i=1}^n \theta |x_i|^\beta + \frac{1}{2}(x_i - v_i)^2 \qquad (17)$$

where $\theta = \sigma^2/\alpha^\beta$. For any exponent parameter β and noise variance σ^2, scale parameter α can be estimated from noisy signal itself, using the following relation.

$$var(\mathbf{v}) \quad = \quad var(\mathbf{x}) + var(\mathbf{e}) = \alpha^2 \frac{\Gamma(3/\beta)}{\Gamma(1/\beta)} + \sigma^2 (18)$$

$$\Rightarrow \quad \alpha = \sqrt{(var(\mathbf{v}) - \sigma^2)\frac{\Gamma(1/\beta)}{\Gamma(3/\beta)}} \qquad (19)$$

When components of $\mathbf{x}$ can be logically grouped as multiple disjoint sets with different scale parameter α, e.g. subbands of wavelet decomposition, (19) should be interpreted

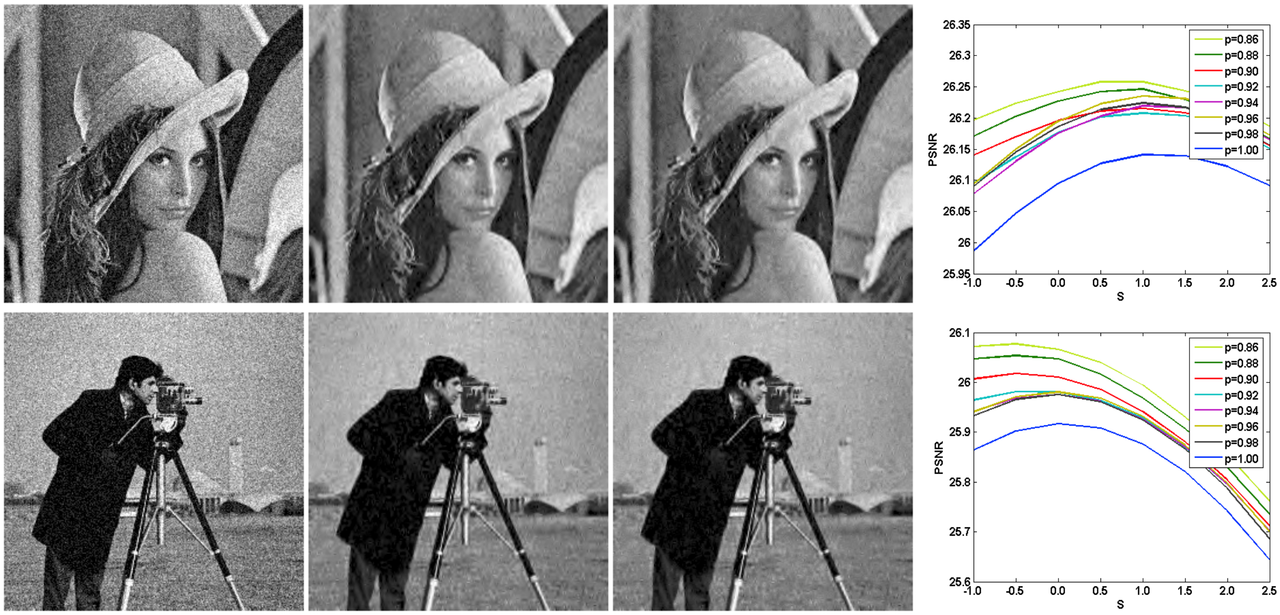

Figure 7. Denoising Results. First to third column: noisy image, best L_1 reconstruction (PSNR = 25.9dB) and best L_P reconstruction (PSNR = 26.7dB) respectively. Last column: PSNR values for different p against the scale correction factor S.

as an independent relation for each set. Due to separability of the likelihood we need to solve a set of one dimensional optimization problems of the form

$$f(x) = \theta x^p + \frac{1}{2}(x - v)^2 \text{ s.t. } x \in [0, v] \qquad (20)$$

Krishnan et al. [14] proposed an LUT based approach for general p and analytical approach for $p = 1/2$ and $p = 2/3$ to find optimal solution of (20). We propose an efficient solution for general p which can either be used without LUT or can be used to reduce the size of the LUT and number of lookups . Optimal solution will be located either at boundary of interval $[0, v]$ or where derivative is zero i.e. at some root of $f'(x) = p\theta x^{p-1} + x - v = 0$ which is same as Eq. 7.

Our algorithm does the norm minimization for all the bands of the image together and does not need the sub-band norms to be separately mentioned as done in [4]. The necessity of maintaining sub-band norms in [4] defeats the purpose of random measurements and limits the applicability of their method for unknown norms. For the generalized Gaussian prior, the α estimate in (19), usually underestimates the true parameter α and hence we introduce a correction factor of the form $(1 + \epsilon)^S$, where we empirically determine $\epsilon = 0.15 \pm 2\%$. Ideally the model should perform best for $S = 0$. This effect can be observed in Fig. 7 (left most column), where the lower p values usually peak at $s = 0$. As we approach L_1 norm ($p = 1.0$) the peak shifts to a different position (s=0.5), corresponding to a correction of about $+7\%$ in the estimate for α. The images show the noisy image, the best L_1 reconstruction, and the best L_p norm reconstruction respectively.

4.4. Sparse l_p PCA

Informative feature selection by sparse PCA has been proposed recently by Naikal et al. [15]. The key intuition in this work is as follows. Let us assume a data matrix $\mathbf{A} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m] \in \mathbb{R}^{n \times m}$, where the m vectors $\mathbf{y}_i$ are assumed to be centered. The empirical covariance matrix is given by $\Sigma_A = \frac{1}{m}\mathbf{A}\mathbf{A}^T$. Sparse PCA computes the first sparse eigenvector of Σ_A by optimizing the following objective [25]

$$\mathbf{x}_s = \arg\max\ \mathbf{x}^T \Sigma_A \mathbf{x} \quad s.t. \|\mathbf{x}\|_2 = 1, \ \ \|\mathbf{x}\|_1 < k \quad (21)$$

Once the first eigenvector is identified, the second can be estimated by repeating the same procedure for the deflated covariance matrix $\Sigma'_A = \Sigma_A - (\mathbf{x}^T\Sigma_A\mathbf{x})\mathbf{x}\mathbf{x}^T$. We replace the final l_1 norm constraint by the more generic l_p norm constraint. We report comparative results to Naikal et al. [15] in Fig. 8. Note that with the l_p norm formulation, lesser features are selected. Most of the repeated features are further dropped as compared to l_1 sparse PCA. For the four class of images shown, Naikal et al. [15] report recognition performance of $86.8 \pm 4.164\%$ with average 24.5 features selected. The performance numbers for our method are $85.6 \pm 2.14\%$ with average 20 features selected per class.

5. Conclusion

In this paper we have looked into the projection onto the L_p norm ball and provided some insights into constructing an exhaustive search algorithm. The results for simulated as well as real experiments encourage us to believe that controllable sparsity afforded by $p < 1$ models can be used

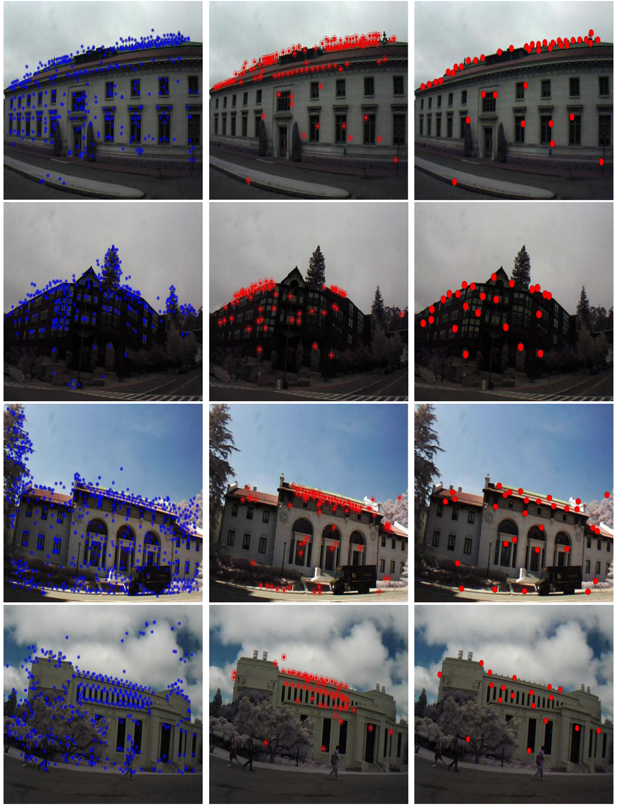

Figure 8. Top: Images of 4 objects in the BMW database [16] with superimposed SURF features; Middle: Informative features detected by the l_1 sparse PCA approach from [15]; Bottom: Informative features detected by l_p sparse PCA.

instead of the prevalent $p = 1$ model, albeit at the cost of convexity of formulation.

References

[1] K. Bredies and D. A. Lorenz. Minimization of non-smooth, non-convex functionals by iterative thresholding, 2009. 5

[2] L. M. Bregman. The method of successive projection for finding a common point of convex sets. In *Soviet Math. Dokl.*, volume 6, pages 688–692, 1965. 5

[3] P. Brucker, B. Jurisch, and B. Sievers. A branch and bound algorithm for the job-shop scheduling problem. *Discrete Applied Mathematics*, 49(1-3):107–127, 1994. 2

[4] E. J. Candes and J. Romberg. Practical signal recovery from random projections. In *Proc. SPIE Computational Imaging*, volume 5674, pages 76–86, 2005. 5, 7

[5] I. Daubechies, R. Devore, M. Fornasier, and C. S. Gntrk. Iteratively reweighted least squares minimization for sparse recovery. *Comm. Pure Appl. Math*, 2008. 1

[6] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994. 1

[7] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l1-ball for learning in high dimensions. In *ICML '08*, 2008. 2, 3

[8] L. Gan. Block compressed sensing of natural images. In *Proc. of Intl. Conf. Digital Signal Processing*, pages 403–406, 2007. 5

[9] R. Garg and R. Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML '09*, pages 337–344, 2009. 5

[10] G. Gasso, A. Rakotomamonjy, and S. Canu. Recovering sparse signals with a certain family of nonconvex penalties and dc programming. *IEEE Trans. Sig. Proc.*, 57(12):4686–4698, Dec. 2009. 6

[11] D. Geman and G. Reynolds. Constrained restoration and recovery of discontinuities. *PAMI*, 14(3):367–383, 1992. 5

[12] D. Geman and C. Yang. Nonlinear image recovery with half-quadratic regularization. *PAMI*, 4(3):932–946, 1995. 5

[13] M. D. Gupta, S. Kumar, and J. Xiao. L1 projections with box constraints. *CoRR*, abs/1010.0141, 2010. 3

[14] D. Krishnan and R. Fergus. Analytic hyper-laplacian priors for fast image deconvolution. In *NIPS*, 2009. 3, 7

[15] N. Naikal, A. Yang, and S. S. Sastry. Informative feature selection for object recognition via sparse pca. Technical Report UCB/EECS-2011-27, EECS Dept., Univ. California, Berkeley, Apr 2011. 7, 8

[16] N. Naikal, A. Y. Yang, and S. S. Sastry. Towards an efficient distributed object recognition system in wireless. *Information Fusion*, 2010. 8

[17] D. Needell and J. A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, Apr 2008. 5

[18] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2000. 3

[19] S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *Annals of Statistics*, 35:1012, 2007. 2

[20] R. Saab, R. Chartrand, and O. Yilmaz. Stable sparse approximations via nonconvex optimization. In *Proc. 33rd Int. Conf. Accoustic Speech and Signal Processing, ICASSP*, 2008. 6

[21] H. Su, A. W. Yu, and L. Fei-Fei. Efficient euclidean projections onto the intersection of norm balls. In *International Conference on Machine Learning (ICML)*, Edinburgh, UK, June 2012. 3

[22] H. Tuy. Concave programming under linear constraints. *Soviet Math*, (5):1437–1440, 1964. 2

[23] B. E. Usevich. A tutorial on modern lossy image compression: Foundations of JPEG 2000. *IEEE Signal Processing Mag.*, pages 22–35, September 2001. 1

[24] Y. Wang, J. Yang, W. Yin, , and Y. Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM J. Imaging Sciences*, 1(3):248–272, 2008. 5

[25] H. Zou, T. Hastie, and R. Tibshirani. Sparse Principal Component Analysis. *J. Computational and Graphical Statistics*, 15, 2004. 7