ON EXACT l_q DENOISING

Goran Marjanovic, Member, IEEE and Victor Solo, Fellow, IEEE

School of Electrical Engineering and Telecommunications
The University of New South Wales, Sydney, Australia

ABSTRACT

Recently, a lot of attention has been given to penalized least squares problem formulations for sparse signal reconstruction in the presence of noise. The penalty is responsible for inducing sparsity, where the common choice used is the convex l_1 norm. While an l_0 penalty generates maximum sparsity it has been avoided due to lack of convexity. With the hope of gaining improved sparsity but more computational tractability there has been recent interest in the l_q penalty. In this paper we provide a novel cyclic descent algorithm for optimizing the l_q penalized least squares problem when $0 < q < 1$. Optimality conditions for this problem are derived and competing ones are clarified. We illustrate with simulations comparing the reconstruction quality with three penalty functions: l_0, l_1 and l_q, $0 < q < 1$.

Index Terms— sparsity, l_q optimization, nonconvex, inverse problem.

1. INTRODUCTION

Sparse regression has become a very popular topic of interest in the last decade. It is widely used in many applications such as machine learning, denoising, inpainting, deblurring, compressed sensing, source separation and more, see [1–6]. The usual regression model is given by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y}_{n\times 1}$ are the observations, $\boldsymbol{\beta}_{m\times 1}$ is a sparse vector of interest, $\mathbf{X}_{n\times m}$ is the regression matrix and $\boldsymbol{\epsilon}_{n\times 1}$ is the noise. Following the parsimony principle, the aim is to choose the simplest model i.e. the sparsest $\boldsymbol{\beta}$ that adequately explains the data $\mathbf{y}$. The sparsity requirement improves interpretability and prevents overfitting.

To estimate a sparse $\boldsymbol{\beta}$ in (1) attention has been given to minimizing sparsity Penalized Least Squares (PLS) objective functions [7–14]. The least squares term measures the goodness-of-fit of the estimator while the penalty forces many of its components to become zero. The most common sparsity inducing penalties belong to the power family [15] and can be characterized using the l_q "norm"[1], which for $0 < q \leq 1$ is defined by:

$$\|\boldsymbol{\beta}\|_q := \left(\sum_{i=1}^{m} |\beta_i|^q\right)^{\frac{1}{q}} \tag{2}$$

The l_q penalty considered in the PLS problem is given by $\|\boldsymbol{\beta}\|_q^q$ for $0 < q \leq 1$. Since $\|\boldsymbol{\beta}\|_q^q$ approaches the total number of nonzero components in $\boldsymbol{\beta}$ as $q \to 0^+$, for $q = 0$ the l_q penalty is precisely this limit, which we denote by $\|\boldsymbol{\beta}\|_0$ and is more commonly referred to as the l_0 "norm". Indeed, $\|\boldsymbol{\beta}\|_0$ is the natural penalty for inducing sparsity. Despite the fact that it is nonconvex [10, 16] have developed a Majorization-Minimization (MM) type algorithm for optimizing the corresponding l_0 PLS problem when the singular values of $\mathbf{X}$ are strictly less than one. A cyclic descent (CD) algorithm has been developed in [17] for a related l_0 PLS problem.

A convex relaxation of $\|\boldsymbol{\beta}\|_0$ is the l_1 norm i.e. $\|\boldsymbol{\beta}\|_q^q$ with $q = 1$. Due to this favourable property, there have been numerous methods developed for solving the resulting l_1 PLS problem. We cannot list them all here but a small sample is [7–9, 11, 18–21]. These methods rely on gradient projection, fixed point, MM and Iteratively Reweighted Least Squares (IRLS) procedures.

The natural question to ask is what happens if the l_1 penalty is replaced by the l_q penalty $\|\boldsymbol{\beta}\|_q^q$ with $0 < q < 1$. It has recently been noted that using nonconvex penalties such as this can alleviate some of the shortcomings of the l_1 norm [22–24]. For example, it is expected that the l_q estimator with $0 < q < 1$ is sparser and less biased than the l_1 estimator [25]. As we will show, the l_q penalty also zeroes out estimator components when $0 < q < 1$ but is less aggressive than the l_0 penalty, and often performs very well when the underlying model is very sparse. This was evidenced in related problems, see [24, 26–28]. The resulting l_q PLS problem is nonconvex, and for some of the methods used for its optimization see [7, 11, 13, 21]. These rely on MM and IRLS or fixed point type procedures.

In this paper we provide a novel CD algorithm (l_qCD) for optimizing the l_q PLS problem when $0 < q < 1$. Optimality conditions for this problem are also derived and the competing ones are clarified.

This work was partly supported by an ARC (Australian Research Council) grant.

[1] The l_q function is not a norm when $q < 1$.

The remainder of the paper is organized as follows. In Section 2 we state the relation to prior work. Section 3 contains the optimality conditions for the l_q PLS problem when $0 < q < 1$, while Section 4 gives the l_qCD algorithm. Section 5 contains the simulations comparing the reconstruction quality of three penalty functions: l_0, l_1 and l_q, $0 < q < 1$. Section 6 contains concluding remarks.

Notation: $\text{sgn}(\beta)$ is the sign of β if $\beta \neq 0$ and 0 otherwise. $\mathbf{e}_i$ is the i-th unit vector from the standard canonical basis. $\text{diag}(\boldsymbol{\beta})$ is a diagonal matrix with $\boldsymbol{\beta}$ on the main diagonal. $\|\mathbf{A}\|$ is the spectral norm of $\mathbf{A}$.

2. RELATION TO PRIOR WORK

2.1. Optimality

Even though existing literature, such as [7, 11–13, 19, 29], deals with the l_q PLS problem, no prior work except [13] presents optimality conditions for $0 < q < 1$. We show that the MM based conditions in [13] are suboptimal and reduce to ours only in a special case, see Theorem 4 and Remark 4.

When $q = 0$, the optimality conditions for the l_q PLS problem have been provided in [10, 16, 17], while for $q = 1$, they are given in [9, 11, 20]. In order to derive the corresponding optimality conditions for $0 < q < 1$ previous work cannot be applied. For example, the approach taken in [7] cannot be extended as it indirectly assumes that the optimal solution contains only nonzero components, see Remark 3. The theory in [30] cannot be applied as the objective function does not satisfy the posed convex type assumptions, while the analysis in [12] only considers differentiable penalties. Consequently, our derivation is novel and it takes advantage of features unique to the l_q PLS objective function.

2.2. Cyclic Descent (CD)

CD procedures for general objective functions are given in [31,32], while CD methods that have been explicitly designed to handle the PLS structure are given in [12, 14, 30]. None of these CD algorithms can handle the l_q PLS problem when $0 < q < 1$. For example, even though [12] analyses the l_q penalty and [30] deals with PLS objective functions, both works including the works in [14,31,32] do not provide a way to construct CD updates when $0 < q < 1$. As a result, our method is significantly different to those in [12, 30, 31, 33].

3. OPTIMALITY CONDITIONS

From now on it is assumed $0 < q < 1$ unless stated otherwise. Then, the l_q PLS problem is:

$$\min_{\boldsymbol{\beta}} \; J(\boldsymbol{\beta}) := \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_q^q \tag{3}$$

where $\lambda > 0$ is a constant penalty parameter. We denote the columns of $\mathbf{X}$ by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ and w.l.o.g. column scale i.e. $\|\mathbf{x}_i\|_2 = 1$ for all $i \in \{1, \ldots, m\}$. Further discussion requires the non-trivial result:

Theorem 1. Consider the following scalar optimization:

$$\min_{\beta} J_{z,\lambda}(\beta) := \frac{1}{2}(z - \beta)^2 + \lambda|\beta|^q \tag{4}$$

Then all its solutions are give by:

$$\tau_\lambda(z) = \begin{cases} 0 & \text{if } |z| < h_\lambda \\ \{0, \text{sgn}(z)\beta_\lambda\} & \text{if } |z| = h_\lambda \\ \text{sgn}(z)\bar{\beta} & \text{if } |z| > h_\lambda \end{cases}$$

where

$$\beta_\lambda := [2\lambda(1-q)]^{\frac{1}{2-q}} \text{ and } h_\lambda := \beta_\lambda + \lambda q \beta_\lambda^{q-1}$$

and $\bar{\beta} > 0$ satisfies $\bar{\beta} + \lambda q \bar{\beta}^{q-1} = |z|$. There are two solutions to this equation and $\bar{\beta} \in (\beta_\lambda, |z|)$ is the larger one. It can be computed from the iteration, $\beta^0 \in [\beta_\lambda, |z|]$:

$$\beta^{k+1} = \rho(\beta^k) \text{ where } \rho(\beta) := |z| - \lambda q \beta^{q-1} \tag{5}$$

Proof. See [27, Theorem 1].

Denote by $\mathcal{G}$ the set of global minimizers of $J(\cdot)$. Then introduce the "adjusted gradient" quantity:

$$z_i = z(\boldsymbol{\beta}_{-i}) := \mathbf{x}_i^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{-i}) \tag{6}$$

where $\boldsymbol{\beta}_{-i}$ is $\boldsymbol{\beta}$ with the i-th component set to zero. We now give two theorems for deducing the optimality conditions for the l_q PLS problem:

Theorem 2. $\mathcal{G} \subseteq \mathcal{F}$ where $\mathcal{F} := \bigcap_{i=1}^m \{\boldsymbol{\beta} : \beta_i \in \tau_\lambda(z_i)\}$

Proof. See the Appendix.

Theorem 3. (Optimality Conditions) Suppose $\boldsymbol{\beta}^* \in \mathcal{F}$, and define $\mathcal{Z} := \{i : \beta_i^* = 0\}$ and $\mathcal{Z}^c := \{i : \beta_i^* \neq 0\}$. Then:

C_1: For $i \in \mathcal{Z}$, $|\mathbf{x}_i^T(\mathbf{X}\boldsymbol{\beta}^* - \mathbf{y})| \leq h_\lambda$

C_2: For $i \in \mathcal{Z}^c$, $|\beta_i^*| \geq \beta_\lambda$

C_3: For $i \in \mathcal{Z}^c$, $\mathbf{x}_i^T(\mathbf{X}\boldsymbol{\beta}^* - \mathbf{y}) + \lambda q|\beta_i^*|^{q-1}\text{sgn}(\beta_i^*) = 0$

Proof. See the Appendix.

Remark 1. By Theorem 2, conditions $\text{C}_{1,2,3}$ are satisfied by any global minimizer of $J(\cdot)$, and hence are the optimality conditions for the l_q PLS problem.

Remark 2. When $q = 0$ is directly substituted in, $\text{C}_{1,2,3}$ become the optimality conditions for the l_0 PLS problem provided in [10, 16, 17]. As $q \to 1^-$, $\text{C}_{1,2,3}$ approach the optimality conditions for the l_1 PLS problem stated in [9, 11, 20].

Remark 3. Let $\boldsymbol{\beta}_c^*$ denote the vector of nonzero components of $\boldsymbol{\beta}^* \in \mathcal{F}$ only, and let $\mathbf{X}_c$ denote the corresponding columns in $\mathbf{X}$. Using the relation $\text{sgn}(\beta_i) = \beta_i/|\beta_i|$ for $\beta_i \neq 0$, C_3 implies:

$$\mathbf{X}_c^T(\mathbf{X}_c\boldsymbol{\beta}_c^* - \mathbf{y}) + \lambda q \mathbf{D}(\boldsymbol{\beta}_c^*)\boldsymbol{\beta}_c^* = \mathbf{0} \tag{7}$$

where $\mathbf{D}(\boldsymbol{\beta}_c) := \text{diag}\left(|\boldsymbol{\beta}_c^*|^{q-2}\right)$. The necessary condition (7) is condition (10) on p.762 in [7] only when $\mathcal{Z} = \emptyset$. As a result, the claimed optimality condition in [7] is incorrect in cases when $\mathcal{Z} \neq \emptyset$, which occurs in sparse signal estimation.

Theorem 4. Let $\mathcal{C}_{CD}$ denote the set of points satisfying the optimality conditions $\mathrm{C}_{1,2,3}$. Also, let $\mathcal{C}_{MM}$ denote the set of points satisfying the MM based optimality conditions in [13]. Then $\mathcal{G} \subseteq \mathcal{C}_{CD} \subseteq \mathcal{C}_{MM}$.

Proof. See the Appendix.

It is trivial to verify that Theorem 4 holds without column scaling $\mathbf{X}$.

Remark 4. From the proof of Theorem 4, $\mathcal{C}_{MM}$ depends on $L := \|\mathbf{X}^T\mathbf{X}\|$ and $\mathcal{C}_{CD} = \mathcal{C}_{MM}$ for $L = 1$. But, since $\|\mathbf{x}_i\|_2 = 1$ for all i, $L = 1$ if and only if $\mathbf{X}^T\mathbf{X} = \mathbf{I}$. In general, $\mathbf{X}^T\mathbf{X} \neq \mathbf{I}$, and so, $\mathcal{C}_{CD} \subset \mathcal{C}_{MM}$ which makes $\mathcal{C}_{MM}$ a strictly larger set. Identical reasoning with rescaling can be applied when $\mathbf{X}$ is not column scaled.

Remark 5. The optimality conditions from [13] are derived by considering the MM algorithm for the lq PLS problem. By Theorem 4 and Remark 4, they in general form a strictly larger set around the global minimizers than conditions $\mathrm{C}_{1,2,3}$. Hence our conditions are "tighter".

4. THE ALGORITHM

The l_qCD algorithm is summarized in the table below:

The l_qCD Algorithm

For $i = 1, 2, \ldots, m, 1, 2, \ldots, m, \ldots$ repeat (1)-(3):

(1) Let $\boldsymbol{\beta}$ be the current iterate

(2) Calculate $z_i = z(\boldsymbol{\beta}_{-i})$ and choose $\beta_i^+ \in \tau_\lambda(z_i)$

(3) The next iterate becomes $\boldsymbol{\beta}^+ = \boldsymbol{\beta}_{-i} + \beta_i^+ \mathbf{e}_i$.

Theorem 5. If $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^+$ are the current and the next iterate of the l_qCD algorithm respectively then $J(\boldsymbol{\beta}^+) \leq J(\boldsymbol{\beta})$.

Proof. See the Appendix.

5. SIMULATIONS

Here we compare the quality of estimators $\widehat{\boldsymbol{\beta}}$ obtained by optimizing the l_q PLS problem with $0 \leq q \leq 1$. For $q = 0$ and $q = 1$ the CD algorithms from [17] and [14] were used respectively. The aim is to reconstruct a signal $\boldsymbol{\beta}$ of length $m = 256$ from $n = 100$ observations. This signal has 24 randomly placed ± 1 spikes. The observations $\mathbf{y}$ is generated according to (1) where the noise is i.i.d. Gaussian with mean zero and variance σ^2. Matrix $\mathbf{X}$ is filled with independent samples of the Gaussian distribution and its columns are then orthonormalized.

Letting MSE $:= \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_2/m$, the values of (q, λ) used to generate Fig.1 correspond to the minimum MSE for the given σ^2. Next, Fig.2 shows contour plots of the average MSE over 20 trials for each (q, λ) considered. In Fig.'s 2 and 3, it can clearly be seen that the (optimal) values of q that minimize the MSE strongly depend on the noise level (σ^2).

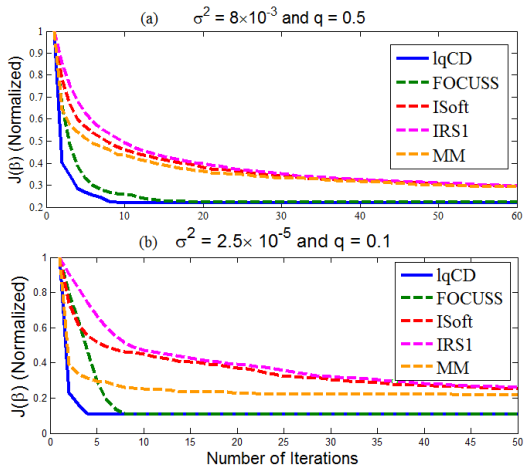

Fig. 1. Examples of $J(\boldsymbol{\beta})$ vs. no. of iterations. All algorithms are initialized with $\boldsymbol{\beta} = \|\mathbf{X}^T\mathbf{X}\|^{-1}\mathbf{X}^T\mathbf{y}$. FOCUSS is from [7], ISoft and IRS1 are from [11], and MM is from [13]. MM is very similar to IST from [11]. To be consistent, updating all components in $\boldsymbol{\beta}$ denotes a single iteration for all algorithms. The stepsize used in ISoft, IRS1 and MM corresponds to the Lipschitz constant $\|\mathbf{X}^T\mathbf{X}\|$. In (a) $q = 0.5$ (b) $q = 0.1$, it is interesting to see that l_qCD and MM do not converge to the same value of $J(\cdot)$, which we confirm by comparing $J(\cdot)$ after 200 iterations. This happens due to severe nonconvexity, but l_qCD achieves a lower value of $J(\cdot)$.

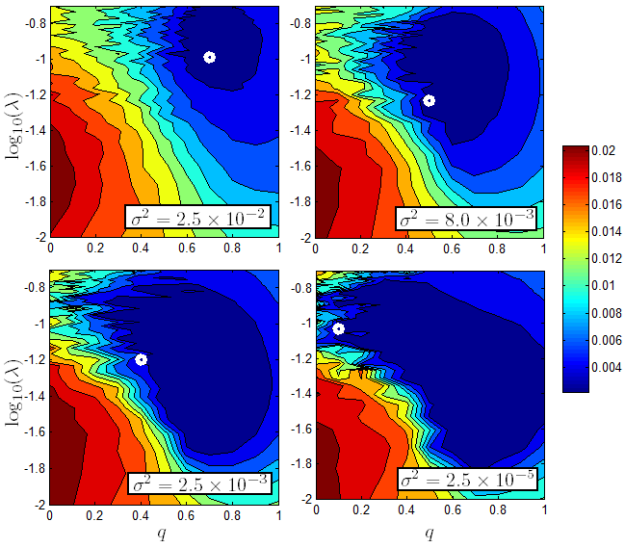

Fig. 2. Av. MSE contour plots for different σ^2 as a function of $(q, \log_{10}(\lambda))$. The white circle is the location of the min. av. MSE. We see that as σ^2 varies so does the (optimal) value of q. As σ^2 rises so does the optimal q.

6. CONCLUSION

We developed a novel CD algorithm for optimizing the l_q PLS problem $(0 < q < 1)$, and derived the corresponding optimality conditions. The set of these contains all the global minimizers but is a subset of the MM based conditions, and so, is

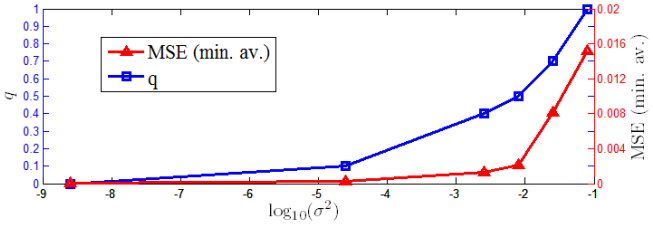

Fig. 3. Min. av. MSE and the corresponding (optimal) value of q as functions of $\log_{10}(\sigma^2)$. These values of $0 < q < 1$ correspond to those in Fig.2 indicated by the white circle.

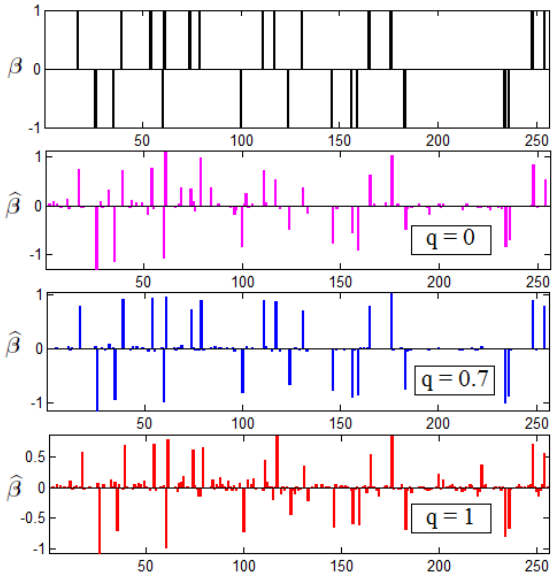

Fig. 4. Example comparing $\boldsymbol{\beta}$ (original signal) and av. reconstructions $\widehat{\boldsymbol{\beta}}$ (over 20 runs) by optimizing the l_q PLS problem when $\sigma^2 = 2.5 \times 10^{-2}$. Min. av. MSE is achieved with $q = 0.7$, and we see that the corresponding $\widehat{\boldsymbol{\beta}}$ best approximates $\boldsymbol{\beta}$ compared to $\widehat{\boldsymbol{\beta}}$ with $q = 0$ and $q = 1$.

tighter. Simulations at varying noise levels show that q with $0 < q < 1$ generally improves on $q = 0$ or $q = 1$.

7. APPENDIX

Proof of Theorem 2: By simple linear algebra, separability of $\|\cdot\|_q^q$ and using z_i in (6), we have:

$$\begin{aligned} J(\boldsymbol{\beta}) &= J(\boldsymbol{\beta}_{-i} + \beta_i \mathbf{e}_i) \\ &\overset{(3)}{=} \frac{1}{2}\|\beta_i \mathbf{x}_i - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{-i})\|_2^2 + \lambda\|\boldsymbol{\beta}_{-i}\|_q^q + \lambda|\beta_i|^q \\ &= J(\boldsymbol{\beta}_{-i}) + \frac{1}{2}\beta_i^2 - z_i\beta_i + \lambda|\beta_i|^q \\ &\overset{(4)}{=} J(\boldsymbol{\beta}_{-i}) - \frac{1}{2}z_i + J_{z_i,\lambda}(\beta_i) \\ &\geq J(\boldsymbol{\beta}_{-i}) - \frac{1}{2}z_i + \min_{\beta} J_{z_i,\lambda}(\beta) \qquad (8) \\ &= J(\boldsymbol{\beta}_{-i} + \tau_\lambda(z_i)\mathbf{e}_i) \end{aligned}$$

The last equality in the above holds by Theorem 1. For $\boldsymbol{\beta} = \boldsymbol{\beta}^0 \in \mathcal{G}$ we have $z_i = z_i^0$. Also, there must be an equality in (8) implying $\beta_i^0 \in \tau_\lambda(z_i^0)$. This holds for any $i \in \{1, \ldots, m\}$.

Proof of Theorem 3: For a particular $i \in \{1, \ldots, m\}$ we examine cases, noting that by (6) we have $z_i^* = \mathbf{x}_i^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{-i}^*)$.

If $i \in \mathcal{Z}$, then by the fact that $\beta_i^* \in \tau_\lambda(z_i^*)$ we have $|z_i^*| \leq h_\lambda$. Hence, by z_i^* and the fact that $\boldsymbol{\beta}_{-i}^* = \boldsymbol{\beta}^*$ we obtain C_1.

If $i \in \mathcal{Z}^c$, then again by the fact that $\beta_i^* \in \tau_\lambda(z_i^*)$ we have $|z_i| \geq h_\lambda$. If $|z_i^*| = h_\lambda$ then $|\beta_i^*| = \beta_\lambda$. If $|z_i^*| > h_\lambda$ then by Theorem 1, $|\beta_i^*|$ satisfies:

$$|\beta_i^*| + \lambda q|\beta_i^*|^{q-1} = |z_i^*| \qquad (9)$$

Define $f(x) := x + \lambda q x^{q-1}$ for $x > 0$, and notice $f'(x) = 1 - \lambda q(1-q)x^{q-2}$. Hence, $f'(x) > 0$ iff $x > \beta'_\lambda := [\lambda q(1-q)]^{\frac{1}{2-q}}$, and so, $f(x)$ is strictly increasing for $x > \beta'_\lambda$. So, since $\beta_\lambda > \beta'_\lambda$, for $f(|\beta_i^*|) = |z_i^*| > h_\lambda$ we have $|\beta_i^*| > \beta_\lambda$, establishing C_2.

Lastly, by Theorem 1 for $i \in \mathcal{Z}^c$ note that $\mathrm{sgn}(z_i^*) = \mathrm{sgn}(\beta_i^*) \neq 0$. Using this as well as z_i^*, (9) becomes:

$$\begin{aligned} & (z_i^* - \beta_i^*)\mathrm{sgn}(\beta_i^*) - \lambda q|\beta_i^*|^{q-1} = 0 \\ \Leftrightarrow \quad & \{\mathbf{x}_i^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{-i}^*) - \beta_i^*\} - \lambda q|\beta_i^*|^{q-1}\mathrm{sgn}(\beta_i^*) = 0 \\ \Leftrightarrow \quad & \mathbf{x}_i^T(\mathbf{X}\boldsymbol{\beta}^* - \mathbf{y}) + \lambda q|\beta_i^*|^{q-1}\mathrm{sgn}(\beta_i^*) = 0 \end{aligned}$$

establishing C_3.

Proof of Theorem 4: $\mathcal{G} \subseteq \mathcal{C}_{CD}$ holds by Remark 1. Next, using our notation, let $\boldsymbol{\beta}^* \in \mathcal{G}$. We have $L := \|\mathbf{X}^T\mathbf{X}\|$, and by using Proposition 3.14 (b) and (c), together with Lemma 4.1 both in [13], the optimality conditions from [13] are:

C_1': For $i \in \mathcal{Z}$, $|\mathbf{x}_i^T(\mathbf{X}\boldsymbol{\beta}^* - \mathbf{y})| \leq L^{\frac{1-q}{2-q}} h_\lambda$

C_2': For $i \in \mathcal{Z}^c$, $|\beta_i^*| \geq L^{-\frac{1}{2-q}} \beta_\lambda$

C_3': For $i \in \mathcal{Z}^c$, $\mathbf{x}_i^T(\mathbf{X}\boldsymbol{\beta}^* - \mathbf{y}) + \lambda q|\beta_i^*|^{q-1}\mathrm{sgn}(\beta_i^*) = 0$

Having $\|\mathbf{x}_i\|_2^2 \leq \|\mathbf{X}^T\mathbf{X}\|$ and $\|\mathbf{x}_i\|_2 = 1$ implies $L \geq 1$. So, suppose $\boldsymbol{\beta} \in \mathcal{C}_{CD}$ and consider any $i \in \{1, \ldots, m\}$. If C_1 holds, since $h_\lambda \leq L^{\frac{1-q}{2-q}} h_\lambda$ implies C_1' holds. Also, if C_2 holds, since $\beta_\lambda \geq L^{-\frac{1}{2-q}} \beta_\lambda$ implies C_2' holds. Lastly, C_3 and C_3' are equal, so $\boldsymbol{\beta} \in \mathcal{C}_{MM}$ and the result follows.

Proof of Theorem 5: Expression (8) becomes: $J(\boldsymbol{\beta}_{-i} + \beta_i^+ \mathbf{e}_i)$ where $\beta_i^+ \in \tau_\lambda(z_i)$ by Theorem 1. The result follows since $\boldsymbol{\beta}^+ = \boldsymbol{\beta}_{-i} + \beta_i^+ \mathbf{e}_i$.

8. REFERENCES

[1] J.Yang, A.Bouzerdoum, and S.Phung, "A training algorithm for sparse ls-svm using compressive sampling," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2054–2057, 2010.

[2] M.Zibulevsky and M.Elad, "L1-l2 optimization in signal and image processing," *IEEE Signal Proc. Mag.*, vol. 27, no. 3, pp. 76–88, 2010.

[3] J.L.Starck, E.J.Candes, and D.L.Donoho, "The curvelet transform for image denoising," *IEEE T. Image Process.*, vol. 11, no. 6, pp. 670–684, 2002.

[4] M.Elad, J.L.Starck, P.Querre, and D.L.Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis (mca)," *J. Appl. Comput. Harmon. Anal.*, vol. 19, no. 3, pp. 340–358, 2005.

[5] E.J.Candes and T.Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?" *IEEE T. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[6] D.L.Donoho, "Compressed sensing," *IEEE T. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[7] B.D.Rao, K.Engan, S.F.Cotter, J.Palmer, and K.K.Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE T. Signal Process.*, vol. 51, no. 3, pp. 760–770, 2003.

[8] M.A.T.Figueiredo, R.D.Nowak, and S.J.Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Top. Signa.*, vol. 1, no. 4, pp. 586–597, 2007.

[9] J.J.Fuchs, "Convergence of a sparse representations algorithm applicable to real or complex data," *IEEE J. Sel. Top. Signa.*, vol. 1, no. 4, pp. 598–605, 2007.

[10] T.Blumensath, M.Yaghoobi, and M.E.Davies, "Iterative hard thresholding and l_0 regularization," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, pp. III–877–III–880, 2007.

[11] M.A.T.Figueiredo, J.M.Bioucas-Dias, and R.D.Nowak, "Majorization-minimization algorithms for wavelet-based image restoration," *IEEE T. Image Process.*, vol. 16, no. 12, pp. 2980–2991, 2007.

[12] R.Mazumder, J.H.Friedman, and T.Hastie, "Sparsenet: Coordinate descent with nonconvex penalties," *J. Am. Stat. Assoc.*, vol. 106, pp. 1125–1138, 2007.

[13] K.Bredies and D.A.Lorenz, "Minimization of non-smooth, non-convex functionals by iterative thresholding," *Tech. Rept.*, 2009, http://www.dfg-spp1324.de/download/preprints/preprint010.pdf.

[14] J.Friedman, T.Hastie, H.Hofling, and R.Tibshirani, "Pathwise coordinate optimization," *Ann. Appl. Stat.*, vol. 1, no. 2, pp. 302–332, 2007.

[15] J.Friedman, "Fast sparse regression and classification," *Tech. Report*, 2008, dept. Stat. Stanford University.

[16] T.Blumensath and M.E.Davies, "Iterative thresholding for sparse approximations," *J. Fourier Anal. Appl.*, vol. 14, pp. 629–654, 2008.

[17] A.J.Seneviratne and V.Solo, "On vector l_0 penalized mulitvariate regression," *ICASSP, Kyoto, Japan*, pp. 3613–3616, 2012.

[18] J.Barzilai and J.Borwein, "Two point step size gradient methods," *IMA J. Numer. Anal.*, vol. 8, pp. 141–148, 1988.

[19] W.J.Fu, "Penalized regressions: The Bridge versus the LASSO," *J. Comput. Graph. Stat.*, vol. 7, no. 3, 1998.

[20] A.Alliney and S.Ruzinsky, "An algorithm for the minimization of mixed l_1 and l_2 norms with application to bayesian estimation," *IEEE T. Signal Process.*, vol. 42, no. 3, pp. 618–627, 1994.

[21] S.J.Wright, R.D.Nowak, and M.A.T.Figueiredo, "Sparse reconstruction by separable approximation," *IEEE T. Signal Process.*, vol. 57, no. 7, pp. 3373–3376, 2009.

[22] E.J.Candes, M.B.Wakin, and S.P.Boyd, "Enhancing sparsity by reweighted l_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, pp. 877–905, 2008.

[23] R.Chartrand and W.Yin, "Iteratively reweighted algorithms for compressive sensing," *International conference on acoustics, speech and signal processing (ICASSP)*, pp. 3869–3872, 2008.

[24] R.Chartrand, "Exact reconstructions of sparse signals via nonconvex minimization," *IEEE Signal Process. Lett.*, vol. 14, pp. 707–710, 2007.

[25] R.Tibshirani, T.Hastie, and J.Friedman, *The elements of statistical learning, second edition: Data mining, inference and prediction*. Springer, New York, 2009.

[26] R.Chartrand and V.Staneva, "Restricted isometry properties and nonconvex compressive sensing," *Inverse Probl.*, vol. 24, pp. 1–14, 2008.

[27] G.Marjanovic and V.Solo, "On l_q optimization and matrix completion," *IEEE T. Signal Process.*, vol. 60, no. 11, pp. 5714–5724, 2012.

[28] ——, "l_q matrix completion," *ICASSP, Kyoto, Japan*, pp. 3885–3888, 2012.

[29] K.Knight and W.J.Fu, "Asymptotics for LASSO type estimators," *Ann. Stat.*, vol. 28, no. 5, 2000.

[30] P.Tseng, "Convergence of block coordinate descent methods for nondifferentiable minimization," *J. Optimiz. Theory App.*, vol. 109, pp. 475–494, 2001.

[31] D.Luenberger and Y.Ye, *Linear and Nonlinear Programming*. Springer, New York, 2008.

[32] P.Huard, *Mathematical Programming Study 10*. North-Holland Publishing Company, 1979.

[33] D.P.Bertsekas, *Nonlinear Programming 2nd ed.* Athena Scientific, Boston, 1999.