Denoising Formulations Based on Support Functions

İlker Bayram

Abstract—**We study a denoising formulation that consists of a quadratic data term and a 'support function' acting as a regularizer. We show that this abstract formulation covers a number of popular denoising formulations. Making use of the properties of support functions, we derive a characterization of the minimizer. Given the characterization, we provide algorithms for obtaining the minimizer for a number of different scenarios of potential interest.**

I. Introduction

Several denoising formulations involve the minimization of a functional like

$$J(t) = \frac{1}{2}\|y-t\|_2^2 + g(t), \tag{1}$$

where 'y' is a given noisy observation of an underlying original signal. Here, the first term, namely $\frac{1}{2}\|y-t\|_2^2$, is usually called the data term and ensures that the denoised estimate is not too far away from the observation. The second term, $g(t)$, reflects our prior information about the original signal. A few popular formulations that follow this general schema are,

(a) Analysis/Synthesis prior (for orthonormal transforms) [25], [13], [8] : $\min_t \frac{1}{2}\|y-t\|_2^2 + \lambda\|t\|_1$.
(b) Analysis prior [5], [11], [30] : $\min_t \frac{1}{2}\|y-t\|_2^2 + \lambda\|A\,t\|_1$ where A is a matrix.
(c) Mixed Norm (synthesis [23], [22] or analysis) prior : $\min_t \frac{1}{2}\|y-t\|_2^2 + \lambda\|A\,t\|_{2,1}$.
(d) Total Variation [27], [5] : $\min_t \frac{1}{2}\|y-t\|_2^2 + \lambda\,\mathrm{TV}(t)$, where $\mathrm{TV}(t)$ denotes the total variation of t, regarded as a one or two-dimensional discrete-time signal.
(e) A variant of the 'Elastic Net' [35] : $\min_t \frac{1}{2}\|y-t\|_2^2 + \lambda_1\|t\|_1 + \lambda_2\|t\|_2$.

Notice that the 'prior information' term, $g(\cdot)$, in all of these problems[1] is convex and satisfies $g(c\,t) = c\,g(t)$ for $c > 0$. The following result, which can be found in [18], leads to a new perspective to look at these problems.

Proposition 1. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a convex function that satisfies $g(c\,t) = c\,g(t)$ for $c > 0$. Then, there exists a convex set K such that $g(t) = \sup_{z\in K}\langle z, t\rangle$.*

The function '$\sup_{z\in K}\langle z,t\rangle$', which will be denoted as $\sigma_K(t)$, is called the *support function of K* [18].

İ. Bayram is with the Department of Electronics and Communication Engineering, Istanbul Technical University, Maslak, 34469, Istanbul, Turkey. E-mail: ibayram@itu.edu.tr

[1]For the different functionals in the list, it is (a)$\lambda\|t\|_1$, (b)$\lambda\|A\,t\|_1$, (c)$\lambda\|A\,t\|_{2,1}$, (d)$\lambda\,\mathrm{TV}(t)$, (e)$\lambda_1\|t\|_1 + \lambda_2\|t\|_2$.

Proposition 1 can be used to relate the problems in the list to the problem,

$$\min_t \frac{1}{2}\|y-t\|_2^2 + \sigma_K(t), \tag{2}$$

for a proper choice of K. A fundamental result for (2) is,

Proposition 2. *Let K be a closed, convex set. Then*

$$\operatorname*{argmin}_t \frac{1}{2}\|y-t\|_2^2 + \sigma_K(t) = y - P_K(y), \tag{3}$$

where $P_K(y)$ denotes the projection of y onto K.

A simple proof of this proposition is presented in Appendix A.

Proposition 2 allows us to consider as equivalent, the minimization problem (2) and the projection problem onto K. That is, if we know how to solve (2), then we can easily project onto K. Conversely, if we know how to project onto K, we can easily solve (2). It is this particular connection that we would like to stress here. Even if we do not readily have a one-step projection operator onto K, if we can devise algorithms that can achieve the desired projection, we obtain, equivalently, algorithms that solve (2). To that end, we derive and present algorithms that solve instances of the problems listed above.

In addition to deriving some known algorithms through the discussed framework, we provide an algorithm for the particular problem 'analysis prior mixed norms' (with $A \neq I$ in (c)) which has not been considered previously, as far as we are aware. We also present a result (Proposition 3), related to the elastic net variant, which has not appeared before.

Relation to Linear Inverse Problems

The minimization problem (2) is also relevant for linear inverse problem formulations that require the minimization of

$$J(t) = \frac{1}{2}\|y - A\,t\|_2^2 + g(t). \tag{4}$$

where A is a linear operator and $g(\cdot)$ is a support function. Indeed, iterative methods derived using majorization-minimization schemes [12] (also see [13] for an EM based approach) ask to minimize functionals of the form

$$J_k(t) = \frac{1}{2}\|y_k - t\|_2^2 + g(t). \tag{5}$$

at each iteration. Minimization of (5) also turns out to be the backward step in forward-backward splitting algorithms [7] for minimizing (4). The cost function in (4) appears in dictionary learning as well (see e.g. [21] where the authors employ 'hierarchical norms' – they also provide a different derivation of Algorithm 2.)

Finally, one can show that, if t^* is a minimizer of (4), then for $z = A^T\big(y - A\,t^*\big)$,

$$\operatorname*{argmin}_{t} \frac{1}{2}\,\|z - t\|_2^2 + g(t) \tag{6}$$

also minimizes (4) (it is in fact equal to t^*). In words, the minimizers of (4) are also the minimizers of a related denoising formulation. Although this point does not lead to any practical method (since we do not have access to t^* to start with), it implies that we can gain insight about the nature[2] of the minimizers of (4) by studying the related denoising problems, which are simpler.

Given a mixture signal y, which is a linear combination of two components, methods based on the 'morphological diversity' of the components (see e.g. [31], [4], [29] and the references therein) aim to separate the components by exploiting the differences in the distributions of their coefficients when possibly different bases/frames[3] are used. To that end, the formulation

$$\min_{t_1, t_2} \frac{1}{2}\,\|y - t_1 - t_2\|_2^2 + \|A_1\,t_1\|_a + \|A_2\,t_2\|_b \tag{7}$$

where A_1, A_2 are possibly different transforms and $\|\cdot\|_a$, $\|\cdot\|_b$ different norms, could also make use of the algorithms presented in this paper. In particular, we think that mixed norms with different groupings could lead to interesting results. We hope to pursue this in future research.

Upto differences in terminology, linear regression formulations (with regularization for subset selection) also pose minimization problems where the cost functions are similar to $J(\cdot)$ in (4) (see [17], Chp 3). It is interesting to trace the evolution of the 'prior terms', that is, $g(\cdot)$ in (4), in the statistics literature. In order to improve 'ridge regression', for which $g(t) = \lambda\,\|t\|_2$, Tibshirani proposes the 'Lasso', setting $g(t) = \lambda\,\|t\|_1$ and he notes that this gives a regularization where the minimizers of (4) have fewer non-zero elements [32]. The introduction of 'least angle regression' (LARS) [10], which allows computing the 'lasso' solutions for all values of λ, hence also addressing the problem of the selection of λ, sparks further research that leads to the modification of the prior term. Yuan and Lin [33], adapt LARS for a prior term, which is equivalent to what we call here 'a mixed norm with non-overlapping groups' (see Section IV-D). Zhao et al. [34], consider extensions such as 'mixed norms with overlapping groups'[4] and hierarchical norms. Jacob et al. [20] further extend this work, introducing norms which force more parsimonious representations in terms of groups, compared to mixed norms with overlapping groups. We provide a discussion about these norms using some concrete examples in Section II.

[2]That is, see whether they are sparse, group-sparse or if certain hierarchies between coefficients exist or not.

[3]It is not strictly required that the frames be different. One could make use of the different distributions of the components alone.

[4]See also the work by Fuchs [14] that addresses the selection of the tuning parameters by considering the dual problem.

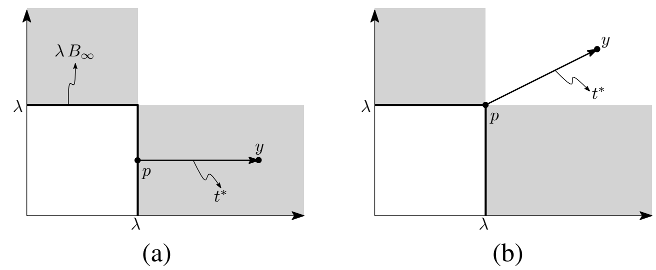


Fig. 1. Soft thresholding y with threshold λ may be regarded as the difference of y from its projection onto $\lambda\,B_\infty$ where B_∞ is the unit ℓ_∞ ball. In the figures above, the projection is denoted by p and the difference, $y - p$, by t^*. Depending on the location of y, the result may be (a) sparse or (b) not.

Contribution

Our main goal in writing this paper is to promote the formulation (2) that consists of a quadratic and a support function acting as the signal prior. We demonstrate, for different cases, that, once a minimization problem is recognized to follow this schema, the characterization in Proposition 2 can be utilized for devising algorithms that solve the problem. Besides the development of new algorithms, Proposition 2 is also useful for understanding the characteristics of the signal prior by an investigation of the support set K. This sheds new light on well-known operators, priors as well, and provides a new way to think about these objects.

Outline

In Section II, we provide some simple examples to demonstrate the norms discussed in the Introduction. Following that, we discuss some generic denoising formulations and derive algorithms for different scenarios in Section III. In Section IV, we specialize these algorithms to the problems listed in the Introduction. Experiments that demonstrate the use of the algorithms along with discussions are provided in Section V. Section VI is the conclusion.

II. Some Low Dimensional Examples

In order to gain some insight, consider the problem

$$\min_{t} \frac{1}{2}\,\|y - t\|_2^2 + \lambda\,\|t\|_1 \tag{8}$$

where $y \in \mathbb{R}^2$. We note that $\|t\|_1 = \sup_{|z_1| \leq \lambda} \langle z, t\rangle$ where $z \in \mathbb{R}^2$. Therefore, by Proposition 2, we need to find the projection of y onto $\lambda\,B_\infty = \{z \in \mathbb{R}^2 : |z_i| \leq \lambda\}$ (call it p) and subtract this projection from y. This is depicted in Figure 1 for two different cases. Notice that if y is in the shaded region, then $t^* = y - p$ is 'sparse' (i.e. has only a single non-zero component in this case). This example demonstrates how the sparsity of the solution is associated with the dual ball ($\lambda\,B_\infty$ in this case). That is, if the tangent to the boundary of the dual ball at p is parallel to an axis, the solution t^* has zero coordinate on that particular axis (as in Figure 1a).

Consider now the following mixed-norms defined on $\mathbb{R}^3$:

$$\big\|(x,y,z)\big\|_{m_1} = \sqrt{x^2+y^2} + |z|, \tag{9}$$

$$\big\|(x,y,z)\big\|_{m_2} = \sqrt{x^2/4+y^2} + \sqrt{x^2/4+z^2}, \tag{10}$$

$$\big\|(x,y,z)\big\|_{m_3} = \frac{1}{2}\left(\sqrt{x^2+y^2} + \sqrt{x^2+z^2} + \sqrt{y^2+z^2}\right). \tag{11}$$

The unit balls of these norms are shown in Figure 2(b,c,d). Using these norms to regularize denoising problems lead to solutions with different characteristics. Noting the sharp corners or edges of the unit balls, we can see that the solutions will be sparse or group-sparse. As in the 2D example above, the characteristics of the norms can also be deduced from the dual-balls. Considering Figure 2(f) for instance, the existence of a flat region on the boundary parallel to the xy plane implies that a number of solutions will be sparse with only a z-component – this will be the case if the x and y coordinates of the observation vector fall within the flat infinite cylinder, parallel to the xy plane, obtained by extending the cylinder along the z axis. Again in Figure 2(f), the existence of the smooth region on the boundary parallel to the z axis implies that a number of solutions will have only x and y components. Thus the mixed norm (9) achieves group sparsity in a sense, where the groups are defined as $\{x,y\}$ and $\{z\}$. In this regard, if we want the groups to overlap, say, to allow the denoised estimates to have either $\{x,y\}$ or $\{x,z\}$ components only, we could utilize the norm in (10). This was the idea in [1]. The unit ball for this norm is shown in Figure 2(c); its dual ball is shown in Figure 2(g). This norm does achieve the desired effect to some extent. We see that a number of solutions will be sparse with only a y-component; a number of solutions will be sparse with only a z-component; a number of solutions will be *almost* group-sparse with either $\{x,y\}$ or $\{x,z\}$ components.

An interesting construction of norms that achieve the desired effect above (*exact* group sparsity with either $\{x,z\}$ or $\{y,z\}$ groups, for instance) was provided by Jacob et al. in [20]. Although the authors define the norms directly in [20], we think that the dual-ball is easier to describe. Indeed, the dual-ball in Figure 3(e) is the intersection of the cylinders $S_y := \{(x,y,z) : \sqrt{x^2+z^2} \leq 1\}$ and $S_z := \{(x,y,z) : \sqrt{x^2+y^2} \leq 1\}$. Since S_y has boundary parallel to the y-axis, a number of solutions will not have a y-component. Similarly, since S_z has boundary parallel to the z-axis, a number of solutions will not have a z-component. The norm associated with this dual ball, that is $\sigma_{S_y \cap S_z}(t)$, is the infimal convolution of $\sigma_{S_y}(t)$ and $\sigma_{S_z}(t)$ (see [18] – also see [20] for the equivalent expression in this context). For the particular case we are discussing, the infimal convolution is given by,

$$\sigma_{S_y\cap S_z}(x,y,z) = \inf_{\alpha\in[0,1]} \sqrt{(\alpha\, x)^2 + y^2} + \sqrt{\big((1-\alpha)\,x\big)^2 + z^2} \tag{12}$$

Although we will not discuss these norms in greater depth, we note that Proposition 2 proves useful for this example too. Indeed, even though we do not have a closed form expression for $\sigma_{S_y\cap S_z}$, we can minimize

$$J(t) = \frac{1}{2}\|u-t\|_2^2 + \sigma_{S_y\cap S_z}(t) \tag{13}$$

if we can project u onto $S_y \cap S_z$. Dykstra's algorithm (see [9], [16]), which computes the projection onto the intersection of a number of convex sets, can be utilized for this purpose[5]. We also invite the reader to compare Figure 2(d,h) with Figure 3(b,f).

Finally, we show in Figure 3(c,d,g,h) the so-called hierarchical norms that favor a certain subgroup within a given group. Specifically, the norm defined by $\sqrt{x^2+y^2+z^2} + |z|$ (see Figure 3(c,g)) ensures that the z-coordinate assumes a non-zero value only if the (x,y) coordinates are non-zero as well. We refer to [34], [21] for more detailed discussions. We note that these norms are instances of mixed norms with overlapping groups. Therefore, one can make use of Algorithms 2, 3 below, when they are utilized as prior terms.

III. Related Problems

In this section, we consider several variations of the problem (2). Starting from 'base operation(s)', like projection(s) onto some simple set(s), we derive algorithms that solve each particular case. The discussions and formulations are generic; specialization to the problems listed in the Introduction are provided in Section IV.

A. Support Functions Under Linear Mappings

We first investigate the case for which the term $\sigma_K(t)$ in (2) is replaced by $\sigma_K(A\,t)$, where A is a matrix. That is, we are interested in minimizing,

$$J(t) = \frac{1}{2}\,\|y-t\|_2^2 + \sigma_K(A\,t). \tag{14}$$

We will assume that we know how to solve the problem when $A = I$, i.e. we know how to project onto K. We first note that,

Lemma 1. *Let K be a closed, convex set, A be a matrix, and $A^T K$ be the set $\{A^T z : z \in K\}$. Then, $\sigma_K(A\,t) = \sigma_{A^T K}(t)$.*

The lemma follows directly from the definition of a support function. We can now write $J(t)$ as,

$$J(t) = \frac{1}{2}\,\|y-t\|_2^2 + \sigma_{A^T K}(t). \tag{15}$$

By Prop. 2, to obtain the minimizer, we need to compute the projection of y onto $A^T K$. Let us denote the projection by p. Notice that $p = A^T z^*$ where z^* is a solution[6] of the minimization problem,

$$z^* \in \operatorname*{argmin}_{z\in K} \underbrace{\|y - A^T z\|_2^2}_{C(z)}. \tag{16}$$

[5] This is not the only option. Thanks to the appearance of an 'inf' term in the definition of the norm, the authors propose another method in [20]. In fact, Dykstra's algorithm can be derived similarly, by considering the dual problem [15].

[6] If A has a non-trivial null-space, the minimizer of (16) may not be unique.

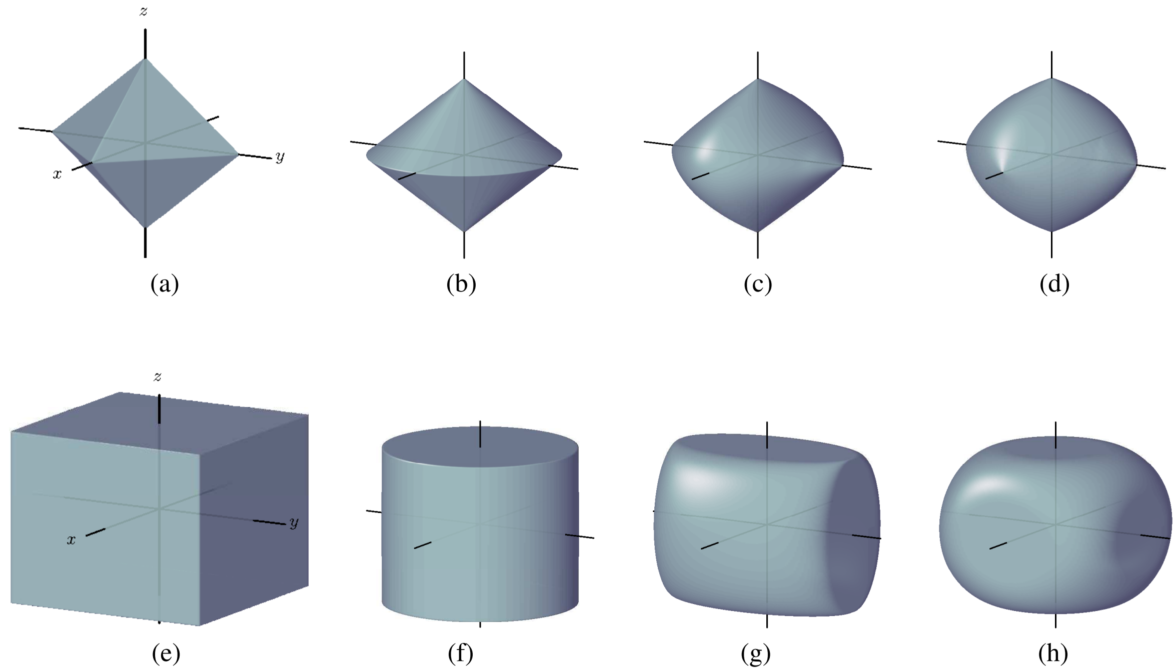


Fig. 2. Unit balls of (a) the ℓ_1 norm, (b) $\sqrt{x^2+y^2}+|z|$, (c) $\sqrt{x^2/4+y^2}+\sqrt{x^2/4+z^2}$, (d) $\frac{1}{2}\left(\sqrt{x^2+y^2}+\sqrt{x^2+z^2}+\sqrt{y^2+z^2}\right)$. (e,f,g,h) The dual balls of the norms in (a,b,c,d).

To come up with an algorithm, we make use of the majorization-minimization procedure [12], [19]. We will construct a sequence $\{z^{(m)}\}_{m\in\mathbb{Z}}$ that converges to a minimizer z^*.

For this, let γ be a positive constant such that $\gamma\, I - A\, A^T$ is a positive-definite matrix. Given $z^{(m)}$, our goal is to find $z^{(m+1)}$ with

$$C(z^{(m+1)}) \leq C(z^{(m)}). \tag{17}$$

Now set

$$z^{(m+1)} = \operatorname*{argmin}_{z\in K} \underbrace{C(z) + \left(z - z^{(m)}\right)^T \left(\gamma\, I - A\, A^T\right) \left(z - z^{(m)}\right)}_{C_m(z)} \tag{18}$$

Notice that,

(i) $C_m(z) \geq C(z)$, because of the positive definiteness of $\gamma\, I - A\, A^T$.
(ii) $C_m\big(z^{(m)}\big) = C\big(z^{(m)}\big)$.

These two conditions ensure that $C(z^{(m+1)}) \leq C(z^{(m)})$, as desired.

Expanding and arranging the terms in (18), we obtain,

$$\begin{aligned} z^{(m+1)} &= \operatorname*{argmin}_{z\in K} \left\| \left(\gamma^{-1} A\big(y - A^T z^{(m)}\big) + z^{(m)}\right) - z \right\|_2^2 \\ &= P_K\left(\gamma^{-1} A\big(y - A^T z^{(m)}\big) + z^{(m)}\right). \end{aligned}$$

In summary, an algorithm that computes the minimizer of (14) is,

Algorithm 1. *Let z be a point in K. Let γ be greater than the largest eigenvalue of $A\, A^T/2$. Let $P_K(\cdot)$ denote the projection operator onto K.*

(I) Repeat until convergence :
Update $z = P_K\left(\gamma^{-1} A\big(y - A^T z\big) + z\right)$
(II) Set $t^* = y - A^T z$.

Remark 1. *Convergence of this algorithm is shown in Appendix B.*

Remark 2. *Although in the derivation of the algorithm, we required that $\gamma\, I - A\, A^T$ be positive-definite, actually it suffices that $2\gamma\, I - A\, A^T$ be positive definite. The difference is that, if $\gamma\, I - A\, A^T$ is positive-definite, then the algorithm monotonely decreases the function $C(\cdot)$ in* (16). *When only $2\gamma\, I - A\, A^T$ is positive-definite, the algorithm produces a sequence z_k that approaches, with each iteration, to the set of minimizers Z^* of* (16). *These claims follow from the derivation of the algorithm above and the proof in Appendix B.*

The second problem of our list in Section I is a special case of the discussed problem with $K = \lambda\, B_\infty$ where B_∞ is the unit ball of the ℓ_∞ norm (see (37)).

B. Multiple Support Functions

In this section, we consider another variation of (2). For this, let K_1, K_2, $\ldots K_k$ be closed, convex sets. We investigate the minimization of

$$J(t) = \frac{1}{2}\, \|y - t\|_2^2 + \sigma_{K_1}(t) + \sigma_{K_2}(t) + \ldots + \sigma_{K_k}(t). \tag{19}$$

In the following, we assume that we know how to project onto K_j for $j = 1, 2, \ldots, k$.

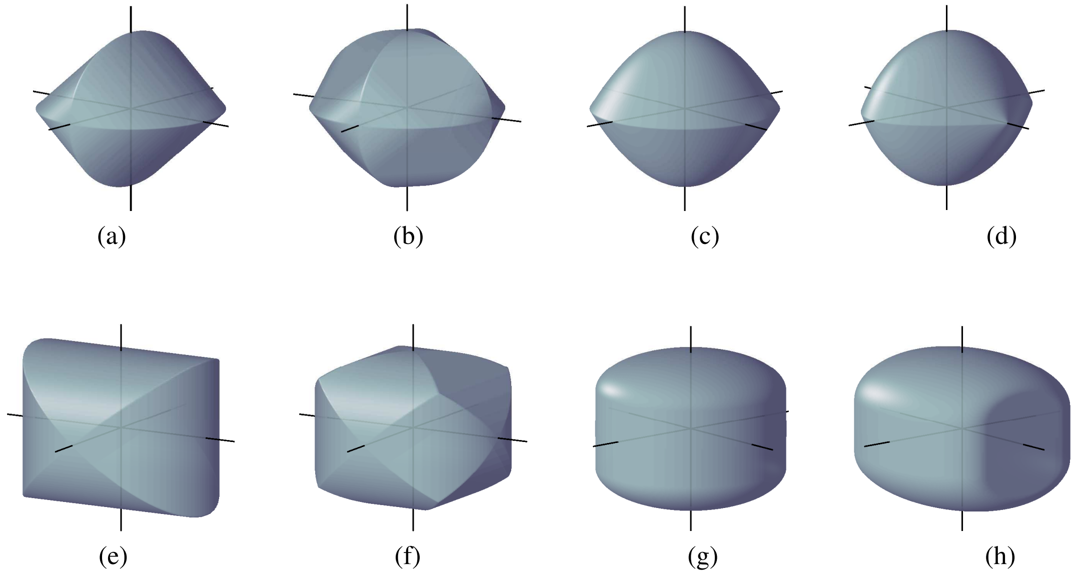


Fig. 3. The dual ball in (e) is defined as the intersection of the cylinders $S_y := \{(x,y,z) : \sqrt{x^2+z^2} \leq 1\}$, $S_z := \{(x,y,z) : \sqrt{x^2+y^2} \leq 1\}$. The dual ball in (f) is defined as the intersection of the cylinders S_y, S_x above and $S_z := \{(x,y,z) : \sqrt{x^2+y^2} \leq 1\}$. (a,b) are unit balls of the support function of the sets in (e,f) respectively. (c) Unit ball of the norm $\sqrt{x^2+y^2+z^2}+|z|$, (d) unit ball of the norm $\sqrt{x^2+y^2+z^2}+\sqrt{x^2+z^2}+|z|$. (g,h) dual balls of the norms in (c,d).

We start by noting that,

Lemma 2. $\sigma_{K_1}(t) + \sigma_{K_2}(t) = \sigma_{K_1+K_2}(t)$.

The lemma follows from the definition of a support function. As a corollary we obtain,

Corollary 1. $\sigma_{K_1}(t) + \sigma_{K_2}(t) + \ldots + \sigma_{K_k}(t) = \sigma_{K_1+K_2+\ldots+K_k}(t)$.

Using this corollary, we write (19) as,

$$t^* = \operatorname*{argmin}_t \frac{1}{2}\,\|y-t\|_2^2 + \sigma_{K_1+K_2+\ldots+K_k}(t). \tag{20}$$

By Prop. 2, we can obtain t^* if we can compute $P_{K_1+K_2+\ldots+K_k}(y)$. Recall that,

$$P_{K_1+K_2+\ldots+K_k}(y) = \operatorname*{argmin}_{z\in K_1+K_2+\ldots+K_k} \|y-z\|_2^2. \tag{21}$$

But any $z \in K_1+K_2+\ldots+K_k$ is expressed as $z = z_1+z_2+\ldots+z_k$ where $z_j \in K_j$. Therefore the minimization problem (21) can be written as,

$$\min_{z_1\in K_1, z_2\in K_2,\ldots,z_k\in K_k} \big\|y-(z_1+z_2+\ldots+z_k)\big\|_2^2. \tag{22}$$

If $(z_1^*, z_2^*, \ldots, z_k^*)$ is a collection of points that minimize the function in (22), $z^* = z_1^*+z_2^*+\ldots+z_k^*$ minimizes (21). Notice that a natural 'coordinate-descent' type algorithm (see e.g. [24] for a general description of coordinate-descent algorithms) for the constrained minimization problem can be constructed for (22). This allows us to solve (19).

Algorithm 2. *Select some $z_1 \in K_1$, $z_2 \in K_2$, ..., $z_k \in K_k$.*

(I) Repeat until some convergence criterion is met,
For $j = 1$ to k,

Let $\tilde{z}_j = \sum_{i\neq j} z_i$

Update $z_j := P_{K_j}(y - \tilde{z}_j)$

(II) Set $t^* = y - \sum_{i=1}^k z_i$.

Remark 3. *In the above algorithm, $\sum z_i$ converges to $P_{K_1+K_2+\ldots+K_k}(y)$.*

Remark 4. *Notice that the update $z_j := P_{K_j}(y-\tilde{z}_j)$ is equivalent to*

$$z_j := y - \tilde{z}_j - \operatorname*{argmin}_t \left(\frac{1}{2}\,\big\|(y-\tilde{z}_j) - t\big\|_2^2 + \sigma_{K_j}(t)\right).$$

C. Multiple Support Functions Under Linear Mappings

Consider the minimization of

$$J(t) = \frac{1}{2}\,\|y-t\|_2^2 + \sigma_{K_1}(A\,t) + \sigma_{K_2}(A\,t) + \ldots + \sigma_{K_k}(A\,t). \tag{23}$$

We transform $J(t)$ as,

$$\begin{aligned} J(t) &= \frac{1}{2}\,\|y-t\|_2^2 + \sigma_{(A^T K_1)}(t) + \ldots + \sigma_{(A^T K_k)}(t) \\ &= \frac{1}{2}\,\|y-t\|_2^2 + \sigma_{A^T(K_1+K_2+\ldots+K_k)}(t). \end{aligned}$$

Therefore, the minimizer of $J(t)$ is equal to $y - P_{A^T(K_1+K_2+\ldots+K_k)}(y)$. We remark that $P_{A^T(K_1+K_2+\ldots+K_k)}(y)$ is given by, $u^* = A^T z^*$, where

$$z^* \in \operatorname*{argmin}_{z\in K_1+K_2+\ldots+K_k} \|y - A^T z\|_2^2. \tag{24}$$

Or, $z^* = z_1^* + z_2^* + \ldots + z_k^*$ where $z_i^* \in K_i$ and

$$(z_1^*, z_2^*, \ldots z_k^*) \in \operatorname*{argmin}_{z_1\in K_1, z_2\in K_2,\ldots,z_k\in K_k} \big\|y - A^T\left(z_1+z_2+\ldots+z_k\right)\big\|_2^2.$$

To realize an algorithm that computes z^*, we once again resort to the majorization-minimization procedure and construct a sequence $\{z^{(m)}\}_{m\in\mathbb{Z}}$ that converges to such a z^*.

Let now $z_j^{(m)} \in K_j$ be given and set $z^{(m)} = \sum_{j=1}^k z_j^{(m)}$. Also let $\gamma\, I - A\, A^T$ be positive-definite. Define,

$$\begin{aligned} C(z) &= \|y - A^T z\|_2^2, \\ C_m(z) &= C(z) + \left(z - z^{(k)}\right)^T \left(\gamma\, I - A\, A^T\right) \left(z - z^{(k)}\right) \\ &= \gamma \left\| \left(\gamma^{-1} A^T \big(y - A\, z^{(m)}\big) + z^{(m)}\right) - z \right\|_2^2 + c \end{aligned}$$

where c is independent of z. Since $C(z^{(m)}) = C_m(z^{(m)})$ and $C_m(z) \geq C(z)$, all we need to do is to find a point $z^{(m+1)} \in \sum K_i$ with $C_m(z^{(m+1)}) \leq C_m(z^{(m+1)})$. But this is exactly the same problem discussed in Section III-B. For instance, for $d = \gamma^{-1} A^T \big(y - A\, z^{(m)}\big) + z^{(m)}$, if we set,

- $z_1^{(m+1)} = P_{K_1}\left(d - \sum_{j>1} z_j^{(m)}\right)$,
- $z_2^{(m+1)} = P_{K_2}\left(d - \sum_{j<2} z_j^{(m+1)} - \sum_{j>2} z_j^{(m)}\right)$,
- $z_3^{(m+1)} = P_{K_3}\left(d - \sum_{j<3} z_j^{(m+1)} - \sum_{j>3} z_j^{(m)}\right)$,

  $\vdots$
- $z_k^{(m+1)} = P_{K_k}\left(d - \sum_{j<k} z_j^{(m+1)}\right)$,

then we obtain a set of vectors $z_j^{(m+1)} \in K_j$ such that, for $z^{(m+1)} = \sum z_j^{(m+1)}$, we have $C_m(z^{(m+1)}) \leq C_m(z^{(m+1)})$.

In summary, the following algorithm produces a minimizer of (23).

Algorithm 3. *Select some $z_1 \in K_1$, $z_2 \in K_2$, $\ldots$, $z_k \in K_k$. Set $z = \sum_j z_j$. Initialize, d, $\tilde{z}_1$, $\ldots$, $\tilde{z}_k$.*

(I) Repeat until some convergence criterion is met,

(i) Update $d = \gamma^{-1} A\big(y - A^T z\big) + z$.

(ii) For $j = 1$ to k,

$$\text{Update } \tilde{z}_j = \sum_{i\neq j} z_i \tag{25}$$

$$\text{Update } z_j = P_{K_j}(d - \tilde{z}_j) \tag{26}$$

(iii) Update $z = \sum_j z_j$.

(II) Set $t^* = y - A^T z$.

D. Convex Combinations with an ℓ_2 term

Finally, we consider a special case of the problem discussed in Section III-B. This time, the cost function is,

$$J(t) = \frac{1}{2}\,\|y - t\|_2^2 + \sigma_K(t) + \lambda \|t\|_2, \tag{27}$$

where K is a closed, convex set.

We remark that Algorithm 2, discussed in Section III-B can be adapted for the minimization of (27). But Algorithm 2 typically requires infinitely many iterations. The following proposition shows that, if we know how to project onto K (which we assume in Section III-B), we can actually reach the minimizer in two steps.

Proposition 3. *Let K be a convex set and B_2 be the unit ℓ_2 ball, both in $\mathbb{R}^n$. Let $C = K + \lambda B_2$. Also, let the projection of a given y onto K be p. Then, the projection of y onto C is $d = p + z$ where z is the projection of $y - p$ onto λB_2, which is given by,*

$$z = \lambda\, \frac{y - p}{\max(\lambda, \|y - p\|_2)}. \tag{28}$$

Proof: We note that the claim is trivially true if $y \in K + \lambda\, B_2$. In the following, we assume that $y \notin K + \lambda\, B_2$. Also, we will take $\lambda = 1$ for simplicity. In this case, we have,

$$z = \frac{y - p}{\|y - p\|_2} \text{ and } d = p + \frac{y - p}{\|y - p\|_2}. \tag{29}$$

We will show that $\langle y - d, x - d\rangle \leq 0$ for all $x \in K + B_2$.

We first note that

$$\langle y - p, u - p\rangle \leq 0 \quad \forall u \in K. \tag{30}$$

Now pick an arbitrary point $x \in K + B_2$. By definition, there exists a pair, $u \in K$, $v \in B_2$ such that $x = u + v$ (this pair need not be unique though). Set $\alpha = 1 - 1/\|y - p\|_2$. We have,

$$\begin{aligned} &\langle y - d, x - d\rangle \\ &= \left\langle y - p - \frac{y - p}{\|y - p\|_2}, u - p + v - \frac{y - p}{\|y - p\|_2} \right\rangle \\ &= \alpha\langle y - p, u - p\rangle + \alpha\langle y - p, v\rangle - \alpha \left\langle y - p, \frac{y - p}{\|y - p\|_2} \right\rangle \\ &\leq \alpha\langle y - p, u - p\rangle + \alpha\|y - p\|_2\, \|v\|_2 - \alpha\|y - p\|_2 \qquad (31) \\ &\leq \alpha\langle y - p, u - p\rangle \qquad (32) \\ &\leq 0. \qquad (33) \end{aligned}$$

Here, (31) *follows by applying the Cauchy-Schwarz inequality to the second term,* (32) *by noting that $\|v\|_2 \leq 1$ and* (33) *is just* (30)*. Since x was an arbitrary point of $K + B_2$, the claim follows.* ■

IV. Application to Different Formulations

We now specialize the algorithms discussed so far to the problems listed in the Introduction.

Let us first recall some facts about ℓ_p norms. For $t \in \mathbb{R}^n$, the ℓ_p norm is defined as,

$$\|t\|_p = \left(\sum_{k=1}^n |t_i|^p\right)^{1/p} \quad \text{for } 1 \leq p < \infty, \tag{34}$$

$$\|t\|_\infty = \max(t_1, t_2, \ldots, t_n). \tag{35}$$

If we denote the unit ball of the ℓ_q norm by B_q, that is,

$$B_q = \{z : \|z\|_q \leq 1\}, \tag{36}$$

then it can be shown, by Hölder's inequality, that, for $1/q + 1/p = 1$,

$$\|t\|_p = \sup_{z\in B_q} \langle z, t\rangle. \tag{37}$$

The 'mixed norm' (defined in Section IV-D) and the convex combination of ℓ_1 and ℓ_2 norms used in the elastic net can also be written as support functions of a set K. For mixed norms, we will not specifically make use of the set K. For the convex combination of ℓ_1 and ℓ_2 norms, we note that K is a sum of the unit balls of ℓ_∞ and ℓ_2 norms. This can be used to gain some geometric insight on the parameters of the elastic net.

A. Analysis/Synthesis Prior Denoising For Orthonormal Transforms

We would like to compute the minimizer of

$$J(t) = \frac{1}{2}\,\|y - t\|_2^2 + \lambda\,\|t\|_1. \tag{38}$$

Noting that $\lambda\,\sigma_K(t) = \sigma_{\lambda\,K}(t)$ (follows from the definition), we can rewrite this as,

$$J(t) = \frac{1}{2}\,\|y - t\|_2^2 + \sigma_{\lambda\,B_\infty}(t). \tag{39}$$

Invoking Prop. 2 the minimizer, t^*, is,

$$t^* = y - P_{\lambda\,B_\infty}(y). \tag{40}$$

If we denote the projection of y to $\lambda\,B_\infty$, as z, we have,

$$z_i = \operatorname{sign}(y_i)\ \max(|y_i|, \lambda) \text{ for } i = 1, 2, \ldots, n. \tag{41}$$

Therefore,

$$t_i^* = y_i - z_i = \operatorname{sign}(y_i)\ \max(|y_i| - \lambda, 0) \text{ for } i = 1, 2, \ldots, n.$$

This is the well-known soft-thresholding function.

For later reference, we define the 'clip' function as,

$$\operatorname{clip}(y, \lambda) := \operatorname{sign}(y)\ \max(|y|, \lambda). \tag{42}$$

Thus, soft thresholding y with λ can be expressed as $y - \operatorname{clip}(y, \lambda)$.

B. Analysis Prior Denoising

We would like to compute the minimizer of

$$J(t) = \frac{1}{2}\,\|y - t\|_2^2 + \lambda\,\|A\,t\|_1, \tag{43}$$

where A is some matrix. This is a particular case of the minimization problem considered in (14), where $K = \lambda\,B_\infty$.

Recalling that $P_{\lambda\,B_\infty} = \operatorname{clip}(y, \lambda)$. Algorithm 1 can be adapted for minimizing $J(t)$ in (43) as,

Algorithm 4. *Let z be a point in $\lambda\,B_\infty$. Let γ be greater than the largest eigenvalue of $A\,A^T/2$.*

(I) Repeat until some convergence criterion is met :
Update $z = \operatorname{clip}\Big(\gamma^{-1}\,A^T\big(y - A\,z\big) + z, \lambda\Big)$

(II) Set $t^* = y - A^T\,z$.

C. Orthonormal Ridge Regression

Before delving into mixed norms, we study the problem

$$t^* = \operatorname*{argmin}_t \frac{1}{2}\|y - t\|_2^2 + \lambda\,\|t\|_2 \tag{44}$$

The solution of this problem will be useful when we consider mixed norms. This problem is also referred to as ridge regression (for an orthogonal design) [17].

Since $\|t\|_2 = \sup_{z \in B_2}\langle z, t\rangle$ (recall (37)), by Prop. 2, we see that,

$$t^* = y - P_{\lambda\,B_2}(y). \tag{45}$$

The projection of y onto $\lambda\,B_2$ is simply $\lambda\,y/\max\big(\|y\|_2, \lambda\big)$. Therefore,

$$t^* = y - \frac{y}{\max\big(\|y\|_2/\lambda, 1\big)}. \tag{46}$$

In the following, we will make use of this result.

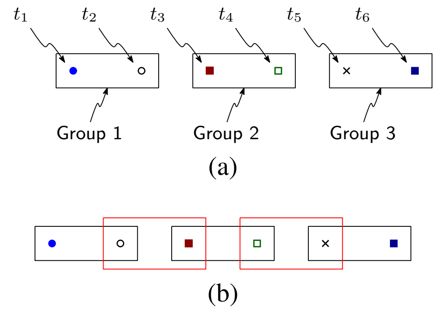


Fig. 4. Mixed ($\ell_{2,1}$) norms are defined as the sum of the ℓ_2 norms of the groups. The mixed norm associated with the grouping system in (a) is given in (47), wheras the mixed norm associated with the grouping system in (b) is given in (48).

D. Mixed Norms

Briefly, the mixed $\ell_{2,1}$ norm of a vector is defined to be the sum of the ℓ_1 norms of groups of coefficients. More concretely, for the vector $t = (t_1, t_2, \ldots, t_6)$ suppose we first form three groups, $g_1 = (t_1, t_2)$, $g_2 = (t_3, t_4)$, $g_3 = (t_5, t_6)$ (see Figure 4a). Given this grouping, the mixed $\ell_{2,1}$ norm is defined as,

$$\|t\|_{2,1} = \sum_{k=1}^{3} \|g_k\|_2 = \sqrt{t_1^2 + t_2^2} + \sqrt{t_3^2 + t_4^2} + \sqrt{t_5^2 + t_6^2}. \tag{47}$$

Notice that in this case, the groups do not overlap. That is, a particular coefficient t_j appears only in a single group. There is no such restriction in general. A grouping system with overlapping groups is shown in Figure 4b. For the grouping system in Figure 4b, the mixed norm becomes,

$$\begin{aligned}\|t\|_{2,1} = {} & \sqrt{t_1^2 + t_2^2} + \sqrt{t_2^2 + t_3^2} \\ & + \sqrt{t_3^2 + t_4^2} + \sqrt{t_4^2 + t_5^2} + \sqrt{t_5^2 + t_6^2}.\end{aligned} \tag{48}$$

In general, given a vector $t = (t_1, t_2, \ldots, t_n)$, we first form a number of groups,

$$\begin{aligned} g_1 &= (g_{1,1}, g_{1,2}, \ldots, g_{1,j_1}) \\ g_2 &= (g_{2,1}, g_{2,2}, \ldots, g_{2,j_2}) \\ &\vdots \\ g_k &= (g_{k,1}, g_{k,2}, \ldots, g_{2,j_k}) \end{aligned}$$

where $g_{i,j} = t_k$ for some k. Based on these groups, the $\ell_{p,q}$ norm of t is,

$$\|t\|_{p,q} = \left(\sum_{i=1}^{k} \|g_i\|_p^q\right)^{1/q}. \tag{49}$$

In this paper, we will take $p = 2$, $q = 1$. For a discussion of the more general case, we refer to [23]. For $p = 2$, $q = 1$, the expression becomes

$$\|t\|_{2,1} = \sum_{i=1}^{k} \|g_i\|_2. \tag{50}$$

We note that the groups are allowed to share coefficients. When this is the case, we refer to the mixed norm as a 'mixed norm with overlapping groups'. Otherwise, we will call it a 'mixed norm with non-overlapping groups'. The latter, as we will see, is easier to handle for our particular problem formulation. However, allowing the groups to overlap leads to better performance [1].

1) Non-Overlapping Groups: As a concrete example, consider the norm (47). The functional of interest in this case is,

$$C(t) = \frac{1}{2}\,\|y-t\|_2^2+\lambda\left(\sqrt{t_1^2+t_2^2}+\sqrt{t_3^2+t_4^2}+\sqrt{t_5^2+t_6^2}\right).$$

$C(t)$ is separable with respect to the groups. That is, for

$$C_1(t_1,t_2) = \frac{1}{2}\sum_{i=1}^{2}\|y_i-t_i\|_2^2+\lambda\sqrt{t_1^2+t_2^2},$$

$$C_2(t_3,t_4) = \frac{1}{2}\sum_{i=3}^{4}\|y_i-t_i\|_2^2+\lambda\sqrt{t_3^2+t_4^2},$$

$$C_3(t_5,t_6) = \frac{1}{2}\sum_{i=5}^{6}\|y_i-t_i\|_2^2+\lambda\sqrt{t_5^2+t_6^2},$$

we see that $C(t) = C_1(t_1,t_2) + C_2(t_3,t_4) + C_3(t_5,t_6)$. Minimization of C_i can be performed separately to obtain the minimizer of $C(t)$. We remark that Section IV-C provides the solution for minimizing $C_i(\cdot,\cdot)$.

2) Overlapping Groups: For mixed norms with overlapping groups, the algorithm outlined above cannot be applied because in this case the cost function is not separable with respect to the groups. Once again, we outline an algorithm for a concrete example. Consider the mixed norm $\|t\|_{2,1}$ in (48). Notice that we can write this norm as the sum of two components with non-overlapping groups,

$$\|t\|_{2,1} = \underbrace{\sqrt{t_1^2+t_2^2}+\sqrt{t_3^2+t_4^2}+\sqrt{t_5^2+t_6^2}}_{\|t\|_a} + \underbrace{\sqrt{t_2^2+t_3^2}+\sqrt{t_4^2+t_5^2}}_{\|t\|_b}. \tag{51}$$

Notice that $\|t\|_a$ and $\|t\|_b$ come from the groups formed by the black and red blocks respectively in Figure 4b. Our minimization problem is,

$$\min_t \frac{1}{2}\,\|y-t\|_2^2+\lambda\,\|t\|_a+\lambda\,\|t\|_b \tag{52}$$

We are in the application domain of Algorithm 2 because we know how to solve each of,

$$\min_t \frac{1}{2}\,\|y-t\|_2^2+\lambda\,\|t\|_a, \tag{53}$$

$$\min_t \frac{1}{2}\,\|y-t\|_2^2+\lambda\,\|t\|_b, \tag{54}$$

from our analysis in the previous section. We remark that solving (53) and (54) is essentially equivalent to projecting y onto the sets A and B, where A and B are such that $\|\cdot\|_a = \sigma_A(\cdot)$, $\|\cdot\|_b = \sigma_B(\cdot)$.

Translating Algorithm 2 for (52), we therefore see that the following routine yields a minimizer of (52) :

(I) Repeat until some convergence criterion is met :
Update

$$t_a = y-t_b-\left\{\operatorname*{argmin}_t\left(\frac{1}{2}\big\|(y-t_b)-t\big\|_2^2+\lambda\,\|t\|_a\right)\right\}$$

Update

$$t_b = y-t_a-\left\{\operatorname*{argmin}_t\left(\frac{1}{2}\big\|(y-t_a)-t\big\|_2^2+\lambda\,\|t\|_b\right)\right\}$$

(II) Set $t^* = y-(t_a+t_b)$.

We remark that we can always express a mixed norm with overlapping groups, $\|t\|_{2,1}$, as the sum of mixed norms with non-overlapping groups, possibly using more than two grouping systems. In particular, suppose we have that $\|t\|_{2,1} = \|t\|_{a_1}+\|t\|_{a_2}+\ldots+\|t\|_{a_k}$ and we know how to minimize each of

$$J_i(k) = \frac{1}{2}\|y-t\|_2^2+\lambda\,\|t\|_{a_i}. \tag{55}$$

Then, to compute

$$\operatorname*{argmin}_t \frac{1}{2}\|y-t\|_2^2+\lambda\,\|t\|_{2,1}. \tag{56}$$

the following algorithm can be used.

Algorithm 5. *Initialize z_1, z_2, $\ldots z_k$ such that $z_i \in K_i$.*

(I) Repeat until some convergence criterion is met,
For $j=1$ to k,

$$\textit{Update } \tilde{z}_j = \left(\sum_{i\neq j} z_i\right)$$

$$\begin{aligned}\textit{Update } z_j &= P_{K_j}(y-\tilde{z}_j)\\ &= y-\tilde{z}_j-\left\{\operatorname*{argmin}_t\left(\frac{1}{2}\,\big\|(y-\tilde{z}_j)-t\big\|_2^2+\sigma_{K_j}(t)\right)\right\}\end{aligned}$$

(II) Set $t^* = y-\sum_{i=1}^k z_i$.

Remark 5. *This is an adaptation of Algorithm 2.*

Mixed Norms Under Linear Mappings: Suppose $\|\cdot\|$ is a mixed norm with overlapping groups and that we can write it as,

$$\|\cdot\| = \|\cdot\|_{a_1}+\|\cdot\|_{a_2}+\ldots+\|\cdot\|_{a_n}, \tag{57}$$

where for each $\|\cdot\|_{a_i}$, we know how to minimize,

$$J_i(t) = \frac{1}{2}\,\|y-t\|_2^2+\lambda\,\|t\|_{a_i}. \tag{58}$$

To minimize

$$J(t) = \frac{1}{2}\,\|y-t\|_2^2+\lambda\,\|A\,t\|, \tag{59}$$

the following algorithm, which is an adaptation of Algorithm 4 can be used.

Algorithm 6. *Initialize z_1, z_2, $\ldots$, z_n such that z_i is an element of the support set for $\|\cdot\|_{a_i}$. Set $z = z_1+\ldots+z_n$. Choose γ, greater than the largest eigenvalue of $A\,A^T/2$.*

(I) Repeat until some convergence criterion is met,

(i) Update $d = \gamma^{-1} A\,(y - A^T\, z) + z$

(ii) For $j = 1$ *to* n,

$$\text{Update } \tilde{z}_j = \sum_{i \neq j} z_i$$

$$\text{Update } u = \operatorname*{argmin}_t \; \frac{1}{2}\,\|d - \tilde{z}_j - t\|_2^2 + \lambda\,\|t\|_{a_j}$$

$$\text{Update } z_j = d - \tilde{z}_j - u$$

(iii) Update $z = \sum_j z_j$.

(II) Set $t^* = y - A^T\, z$.

See experiment 1 in Section V for a comparison of different uses of this algorithm.

E. Total Variation Denoising

Let t be an $N \times N$ image whose $(i,j)^{\text{th}}$ pixel is denoted by $t_{i,j}$. In this setting, the total variation (TV) of t is,

$$\mathrm{TV}(t) = \sum_{i,j=1}^{N} \left\| \Big(\big(D_x\, t\big)_{i,j}, \big(D_y\, t\big)_{i,j} \Big) \right\|_2 \tag{60}$$

where D_x and D_y are finite difference operators, which can be realized by LTI filtering t with $[1\;-1]$ in the horizontal and vertical directions respectively.

We recognize $\mathrm{TV}(t)$ as a mixed norm, with non-overlapping groups under the linear mapping
$A = \begin{bmatrix} D_x & D_y \end{bmatrix}^T$.

Remark 6. *The spectral radius of* $A\,A^T$ *satisfies*
$\rho(A\,A^T) \leq 8$.

In this setting, an algorithm for minimizing

$$J(t) = \frac{1}{2}\|y - t\|_2^2 + \lambda\,\mathrm{TV}(t) \tag{61}$$

can be given as,

Algorithm 7. *Set* $\gamma > 4$ *(assuming that* D_x *and* D_y *are realized as described above). Initialize* z_x, z_y *by setting to zero.*

(I) Repeat until some convergence criterion is met,

(i) Update $d = y - D_x^T\, z_x - D_y^T\, z_y$

(ii) Update $d_x = \gamma^{-1}\, D_x\, d + z_x$

(iii) Update $d_y = \gamma^{-1}\, D_y\, d + z_y$

(iv) For $i, j = 1$ *to* N, *update*

$$c = \max\left(\sqrt{(d_x)_{i,j}^2 + (d_y)_{i,j}^2}, \lambda \right)$$

$$(z_x)_{i,j} = \lambda\,(d_x)_{i,j}/c$$

$$(z_y)_{i,j} = \lambda\,(d_y)_{i,j}/c$$

(II) Set $t^* = y - D_x^T\, z_x - D_y^T\, z_y$.

This algorithm can also be found in [6], [3]. An interesting observation pertaining to this algorithm is that, although the iterates work on the dual formulation, the associated primal variables seem to monotonely decrease the cost function in (61). This is discussed in Experiment 2.

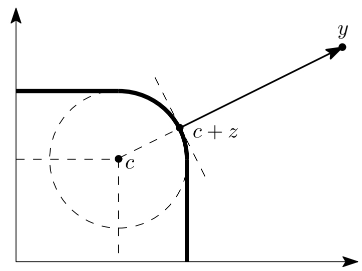
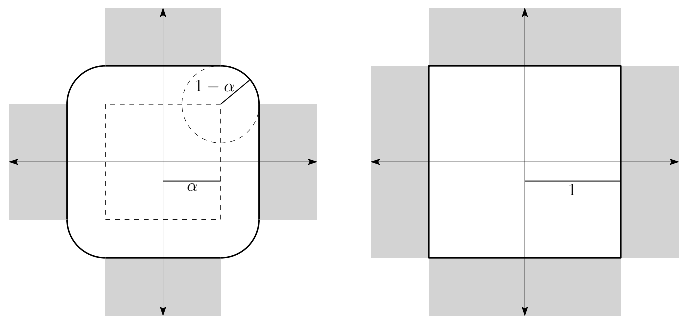


Fig. 5. Projecting y onto $B = \alpha\, B_\infty + (1 - \alpha)\, B_2$ can be performed in two steps. First, project y onto $\alpha\, B_\infty$ to obtain c. Then, project $y - c$ onto $(1 - \alpha)\, B_2$ to obtain z. $c + z$ gives the desired projection.



(a) (b)

Fig. 6. The dual balls of (a) the elastic net, (b) the ℓ_1 norm. If y is in the gray regions, the solution will be 'sparse'.

F. Elastic Net [7]

Consider the minimization problem

$$\min_t \frac{1}{2}\,\|y - t\|_2^2 + \lambda\,\big(\alpha\,\|t\|_1 + (1 - \alpha)\,\|t\|_2\big). \tag{62}$$

To find the solution, we need to project onto $B = \lambda\,\big(\alpha\, B_\infty + (1 - \alpha)\, B_2\big)$. Thanks to Proposition 3, this can be achieved with two projections :

(I) Set $c := P_{\lambda\,\alpha\, B_1}(y) = \mathrm{clip}(y, \lambda\,\alpha)$.

(II) Set $z := P_{\lambda\,(1-\alpha)\, B_2}(y - c) = \lambda\,(1 - \alpha)\frac{y - c}{\max(\|y - c\|_2, \lambda\,(1-\alpha))}$.

(III) Set $t^* = y - c - z$.

The procedure is demonstrated in Figure 5.

To understand the difference with ℓ_1-regularization, we can compare the dual-ball B with the dual ball of the ℓ_1 norm, namely B_∞ (see Figure 6). Like B_∞, B leads to sparse solutions provided the observations fall in the shaded region. However, this 'sparse region' is smaller than that of B_∞. Consequently, elastic net solutions are typically not as sparse as ℓ_1 regularized solutions. On the other hand, when the observations fall in the remaining white regions, the elastic net solutions preserve the direction of the observation vector better. This in turn means that, compared to ℓ_1 regularization, the elastic net provides solutions that are more faithful to the correlations among the coefficients. In short, we can say that the elastic net allows a trade-off between sparsity and the preservation of the correlations. We refer to Experiment 3 for a further discussion.

[7]The discussion included in this subsection is partially adapted from [2].

V. Experiments and Discussion

In this section, we consider some experiments to evaluate and discuss the presented algorithms[8].

Experiment 1. We consider a signal that consists of three clicks, whose Short-Time Fourier Transform coefficients are shown in Figure 7(a). We add Gaussian noise to this signal to obtain the 'noisy observation', y (rms = 2.86) whose STFT coefficients are shown in Figure 7(b). We try three different denoising methods. First, we simply apply a soft threshold to the STFT coefficients where the threshold is selected so as to yield the best output rms (Figure 7(c). The other two minimize a functional of the form

$$J(t) = \frac{1}{2}\,\|y - t\|_2^2 + \lambda \|A\,t\|_{2,1}, \tag{63}$$

using Algorithm 6. These two methods differ in the way the groups are formed. In Figure 7(d) the output when the groups are formed along the time-axis (G_1 in Figure 8) are shown. Figure 7(e) shows the output when the groups are formed along the frequency-axis (G_2 in Figure 8). We note that both visually, in terms of rms and perceptually, this last scheme outperforms the others. An interesting thing to note is that, despite the higher rms value, mixed norms of Figure 7(d) do not suffer from musical noise which is clearly heard in soft-thresholding.

Experiment 2. We apply Algorithm 7 (TV denoising) to the noisy image in Figure 9a. In 100 iterations, we reach the 'approximately denoised' image in Figure 9b. Logarithm of the primal cost function $J(t)$ evaluated at the iterates are shown in Figure 9c for $\gamma = 4$ and $\gamma = 8$. For $\gamma = 8$, we see that the primal cost function is decreasing with iterations. We note that the descent in the primal cost function is not an obvious consequence of the derivation of the algorithm because the development of the algorithm is based on the dual problem. We do not have a proof for this interesting behavior, but we would like to note that if it is true in general, one could modify the readily used algorithms for linear inverse problems with TV regularizers. Especially, 'iterated soft-thresholding algorithms' employ iterations that consist of a Landweber step (which can be performed easily) followed by a TV denoising step (see e.g. [3]). Even though the exact TV denoising step requires infinitely many iterations, one can stop as soon as a descent in the cost function is achieved, but the number of iterations required to reach a descent in the cost function are usually not known beforehand. Therefore it would be of interest to show that, for certain choiced of γ, the algorithm monotonely decreases the primal cost function.

As a final note on the choice of γ, we show in Figure 9d the distance from the limit as the as a function of iterations, for $\gamma = 4$ and $\gamma = 8$. In order to compute this limit numerically, we first obtained an approximate limit image by running the algorithm with $\gamma = 8$ for 5000 iterations. We see that the choice $\gamma = 4$, although not monotonely decreasing the primal cost function, leads to faster convergence to the limit image.

[8]Matlab Codes of these experiments are available at `http://web.itu.edu.tr/ibayram/Temp/DenoiseSupport.zip`

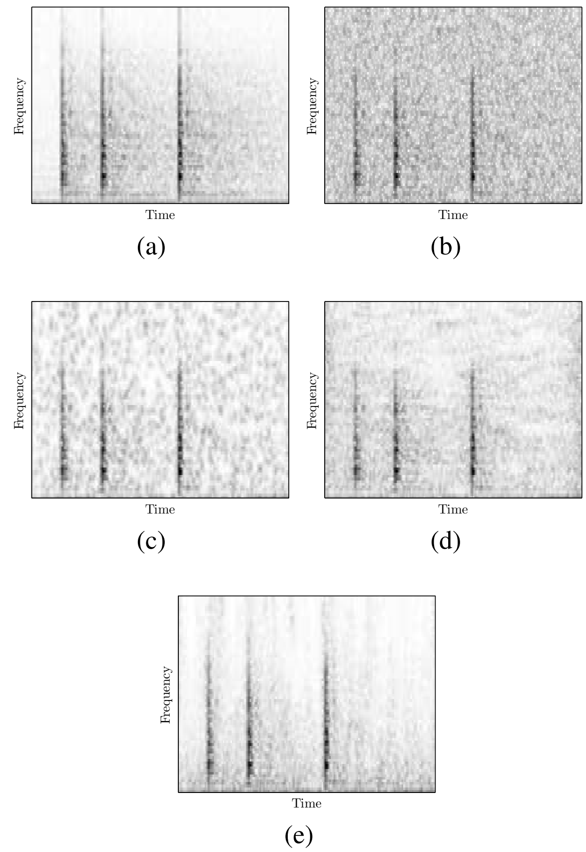


Fig. 7. Time-frequency maps (spectrogram magnitudes) for a denoising experiment. The same nonlinear function is applied to the spectrograms above to enhance visibility. (a) Clean signal, (b) Noisy observation (rms = 2.86), (c) Soft threshold output (rms = 0.82), (d) Mixed norm denoising, groups formed along ridges parallel to the time-axis (rms = 1.00), (e) Mixed norm denoising, groups formed along ridges parallel to the frequency-axis (rms = 0.75). Despite the higher 'rms' value, the output in (d) sounds perceptually better than (c), which suffers from musical noise. The output in (e) performs better than the other two methods, both perceptually and in terms of 'rms'.

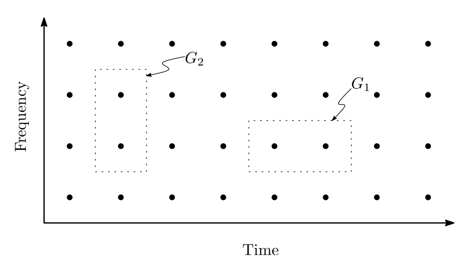


Fig. 8. To define the mixed norms used in Example 1, two different grouping systems on the STFT coefficients are used. The first one, G_1, groups coefficients along the time-axis (groups consist of all the shifts of G_1). The second one, G_2, groups coefficients along the frequency-axis (groups consist of all the shifts of G_1). Although in this figure the groups host two coefficients, in practice we used groups hosting 15 coefficients.

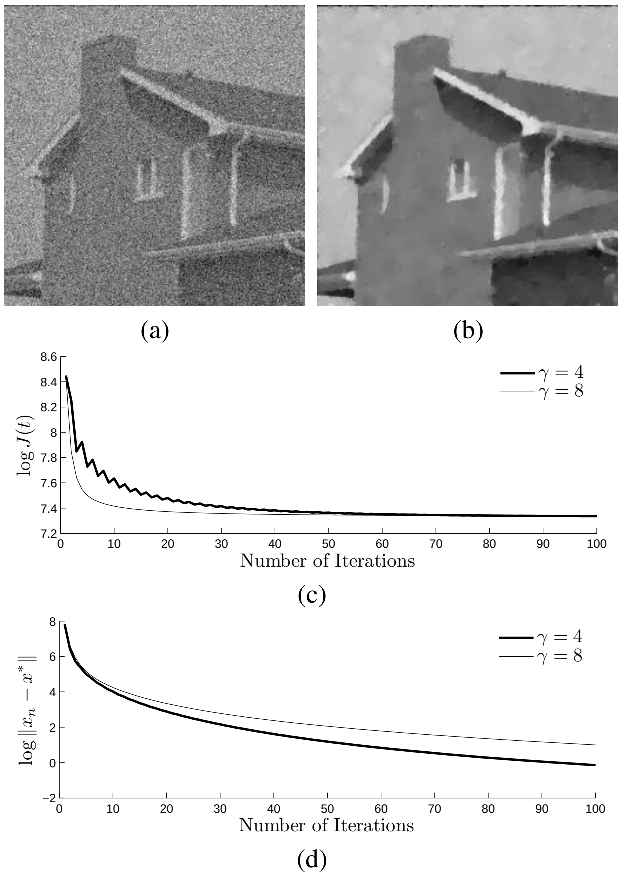


Fig. 9. (a) Noisy observation, (b) Result of applying Algorithm 7 to the noisy observation with 500 iterations, (c) Evolution of the primal cost function in (61) for $\gamma = 4$ and $\gamma = 8$, (d) log-Distance from the limit TV denoised image for $\gamma = 4$ and $\gamma = 8$.

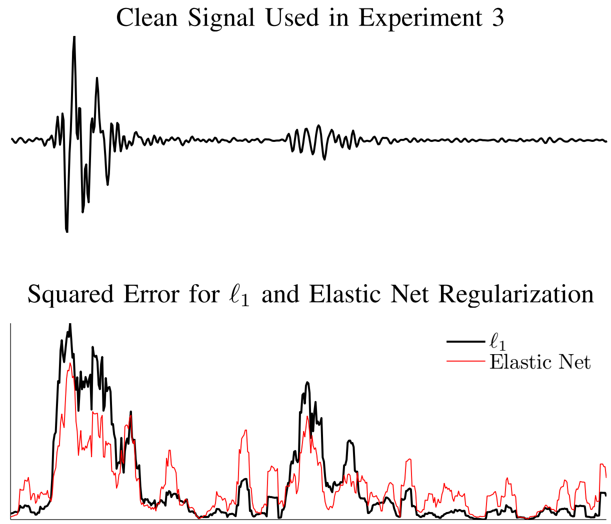


Fig. 10. A denoising experiment comparing the elastic net and ℓ_1 regularization. On the top panel, the underlying clean signal is shown. This signal is 'group-sparse'. The bottom panel shows the squared error signals (smoothed for better visualization) for ℓ_1 regularization (the thick black line) and the elastic net (the thin red line). Noting that the total error for both are the same (assured by selecting λ accordingly), we observe that the error is more evenly distributed for the elastic net.

Experiment 3. In this experiment, we aim to demonstrate the freedom that the elastic net brings by providing a trade-off between sparsity and the correlation structure of the coefficients. We take the clean signal shown on the top panel of Figure 10. Observe that some coefficients form non-zero clusters, but otherwise the signal can be considered sparse. We add noise to this signal and then denoise using ℓ_1 regularization and the elastic net. For denoising, we chose a single λ that gives the same output SNR for both methods. The squared error signal for the two cases are shown on the bottom panel of Figure 10. For ℓ_1 (the thick black line), the error signal is unevenly distributed, taking higher values where the original signal is locally non-sparse. The elastic net solution distributes the error more evenly (the thin red line).

A point we find interesting, and perhaps also a bit disappointing, is that, on this example (and a number of other examples), the elastic net did not lead to an overall increase in performance. We think that in order to achieve better performance, one needs to form more complicated structures like those of mixed norms with appropriate group structures.

VI. Conclusion

In this paper, we studied the problem

$$\min_t \frac{1}{2} \|y - t\|_2^2 + \sup_{z \in K} (t). \tag{64}$$

For proper choices of the set K, we have seen that one can recover a number of popular formulations as listed in the Introduction. We aimed to develop further insight about the problem by exploiting the close relation between the solution of (64) and the projection of y onto K. We think that recognition of a certain denoising problem as (64), however exotic the associated K may be, could lead to new/useful algorithms since projections in different scenarios are well-studied in convex analysis.

Regarding the solution of the problem (64) as essentially equivalent to a projection to the dual-ball (that is, K), one is naturally led to the design of norms by explicit constructions of K. This is demonstrated in Section II for a simple case (following [20]). This approach to the design of norms, could also help improve signal models, without sacrificing the simplicity of the resulting model.

Appendix A
Proof of Proposition 2

Compact proofs for Proposition 2 making use of convex analysis tools can be found in [5], [3]. Here we provide a simple proof from scratch, following the idea in [3].

We will see that there is a very natural dual problem [26] associated with (2). In the following, we essentially derive the solution for the dual problem and show that it leads to the solution of the primal problem as well.

Define,

$$J(t, z) = \frac{1}{2} \|y - t\|_2^2 + \langle z, t \rangle. \tag{65}$$

The problem (2) can now be expressed as,

$$\min_t \max_{z \in K} J(t, z). \tag{66}$$

Let us denote $J^* := \min_t \max_{z\in K} J(t,z)$. If we can find a pair (t^*, z^*) such that

(i) $\mathrm{argmax}_{z\in K}\langle t^*, z\rangle = z^*$,
(ii) $J^* \geq J(t^*, z^*)$,

then t^* is the minimizer we seek in (2).

Since, $J(t,z) \leq \max_{u\in K} J(t,u)$ for all $z \in K$ and t, we have,

$$\min_t J(t,z) \leq \min_t \max_{u\in K} J(t,u) = J^* \qquad \text{for all } z \in K.$$

This yields,

$$J^* \geq \max_{z\in K} \min_t J(t,z). \tag{67}$$

Now let us look at the right hand side of (67). Notice that the unique solution of $\min_t J(t,z)$ is,

$$\operatorname*{argmin}_t J(t,z) = \operatorname*{argmin}_t \frac{1}{2}\|y-t\|_2^2 + \langle z, t\rangle = y - z.$$

Inserting this into the rhs of (67), we have,

$$\max_{z\in K} \min_t J(t,z) = \max_{z\in K} \frac{1}{2}\|z\|_2^2 + \langle z, y-z\rangle \tag{68}$$
$$= \max_{z\in K} -\frac{1}{2}\|z\|_2^2 + \langle z, y\rangle \tag{69}$$
$$= \max_{z\in K} -\frac{1}{2}\|y-z\|_2^2 + \frac{1}{2}\|y\|_2^2. \tag{70}$$

We finally note that,

$$\operatorname*{argmax}_{z\in K} -\frac{1}{2}\|y-z\|_2^2 + \frac{1}{2}\|y\|_2^2 = \operatorname*{argmin}_{z\in K} \|y-z\|_2^2 \tag{71}$$
$$= P_K(y), \tag{72}$$

where $P_K(y)$ denotes the projection (the closest point) of the set K to y. Because K is closed, such a point exists in K. Thanks to convexity, it is unique.

Let us check whether the pair $(t^*, z^*) = \big(y - P_K(y), P_K(y)\big)$ satisfies the conditions (i), (ii) above. For (i), we first recall that the projection of y onto K is the unique point p that satisfies (see e.g. [18]),

$$\langle y-p, z-p\rangle \leq 0 \quad \text{ for all } z \in K. \tag{73}$$

Using this, (i) is seen to be true since

$$\operatorname*{argmax}_{z\in K} \langle t^*, z\rangle = \operatorname*{argmax}_{z\in K} \langle y - P_K(y), z\rangle \tag{74}$$
$$= \operatorname*{argmax}_{z\in K} \langle y - P_K(y), z - P_K(y)\rangle \tag{75}$$
$$= P_K(y). \tag{76}$$

Also, (ii) follows automatically from (67). Thus follows the claim.

Appendix B
Convergence of Algorithm 1

The algorithm is actually an instance of a forward-backward splitting algorithm. Therefore, its convergence is a consequence of the results of Combettes and Wajs [7]. We note that the treatment in [7] is valid for a more general setting. Here, we provide a convergence proof from scratch using less sophisticated machinery.

For convenience of notation, let us denote the two steps of Algorithm 1 by a function $\mathcal{M}_c(\cdot)$:

$$\mathcal{M}_\gamma(z) = P_K\left\{\gamma^{-1} A\big(y - A^T z\big) + z\right\}. \tag{77}$$

With this definition, we can express the algorithm as $z^{m+1} = \mathcal{M}_\gamma(z^m)$. Since the cost function $C(z)$ (see (16)) is convex and bounded from below, there exists at least one minimizer (but it need not be unique). We denote the (non-empty) set of minimizers as Z^*. Let us start by showing that the fixed points of this operator coincide with the set of minimizers.

Lemma 3. *Suppose that $\gamma I - A A^T > 0$. Then, $\mathcal{M}_c(z) = z$ if and only if $z \in Z^*$.*

Proof: Let $z^* \in Z^*$. This means that $C(z^*) \leq C(z)$ for all feasible z (i.e. for all $z \in K$). Consider $D(z) = C(z) + \big(z - z^*\big)^T \big(\gamma I - A A^T\big)\big(z - z^*\big)$ (compare with (18)). We want to show that

$$\mathcal{M}_\gamma(z^*) := \operatorname*{argmin}_{z\in K} D(z) = z^*. \tag{78}$$

Suppose this is not true. This is only possible if $D(z) \leq D(z^*)$ for some feasible z^*. This in turn means that

$$\underbrace{C(z) - C(z^*)}_{t_1} + \underbrace{(z-z^*)^T\,(\gamma I - A A^T)\,\{(z-z^*)\}}_{t_2} \leq 0.$$

Since t_1 is non-negative and t_2 is positive, this cannot be true. Therefore $\mathcal{M}_\gamma(z^*) = z^*$.

Now suppose $z^* \notin Z^*$ for a feasible z^*. We can find a feasible z s.t. $C(z) < C(z^*)$. Let $z_u = u\,z + (1-u)\,z^*$ and consider the (feasible) segment $\{z_u : u \in [0,1]\}$. On this segment, we will show that there exists $u > 0$ s.t. $D(z_u) < D(z^*)$ and therefore that $\mathcal{M}_\gamma(z^*) \neq z^*$. Consider the affine function $f(u) = u\,C(z^*) + (1-u)\,C(z)$. By convexity of $C(\cdot)$, we have $C(z_u) \leq f(u)$ with $f(0) = C(z^*)$. Also let, $g(u) = \big(z_u - z^*\big)^T\big(\gamma I - A A^T\big)\big(z_u - z^*\big)$. Notice that $g(u) = d\,u^2$ for some $d > 0$. Thus, for a fixed arbitrary $a > 0$, there exists $u \in (0,1]$ s.t. $g(u) < a\,u$. This implies that $f(u) + g(u) < f(0) + g(0)$ for some $u \in (0,1]$. Hence

$$D(z_u) \leq f(u) + g(u) < f(0) + g(0) = D(z^*), \tag{79}$$

which is the contradiction we set out to reach. ■

This is a characterization of the fixed points of $\mathcal{M}_\gamma(\cdot)$ for a restricted set of γ values. In fact, we can remove this restriction. We first show that the fixed point set of $\mathcal{M}_\gamma(\cdot)$ is not dependent on γ.

Lemma 4. *$\mathcal{M}_\gamma(z^*) = z^*$ if and only if $\mathcal{M}_{\gamma'}(z^*) = z^*$ for any $\gamma' > 0$.*

Proof: Suppose that $\mathcal{M}_\gamma(z^*) = z^*$. That is,

$$z^* = P_K\left(\gamma^{-1} A\left(y - A^T z^*\right) + z^*\right). \tag{80}$$

This is equivalent to,

$$\langle \gamma^{-1} A\left(y - A^T z^*\right), z - z^*\rangle \leq 0 \quad \forall z \in K. \tag{81}$$

Notice that we can replace γ with any positive γ' without changing the equality. Therefore $\mathcal{M}_{\gamma'}(z^*) = z^*$. Repeating the same argument by changing the roles of γ and γ', the claim of the lemma follows. ■

Combining the two lemmas, we have

Corollary 2. *$\mathcal{M}_\gamma(z) = z$ if and only if $z \in Z^*$, for any $\gamma > 0$.*

Let us now turn to the convergence issue. In the following, $\rho(A)$ denotes the spectral radius of the matrix A. We first recall a well-known result from convex analysis (see e.g. [18]).

Lemma 5. *Let K be a closed convex set and $P_K(\cdot)$ be the projection operator to K. Then $P_K(\cdot)$ is a non-expansive mapping, i.e.,*

$$\|P_K(x) - P_K(y)\|_2 \leq \|x - y\|_2. \tag{82}$$

Thanks to this lemma, we have,

$$\|\mathcal{M}_\gamma(z) - \mathcal{M}_\gamma(\tilde{z})\|_2 \leq \|(I - \gamma^{-1}\, A\, A^T)(z - \tilde{z})\|_2. \tag{83}$$

for any z, $\tilde{z}$. Therefore if $\rho(I - \gamma^{-1}\, A\, A^T) \leq 1$, $\mathcal{M}_\gamma(\cdot)$ is a nonexpansive mapping. Now for a given z^0, suppose we define $z^{i+1} = \mathcal{M}_\gamma(z^i)$. In particular, since we know from Corollary 2 that $z^* \in Z^*$ is a fixed point of this operator, we get

$$\|\mathcal{M}_\gamma(z^i) - z^*\|_2 = \|z^{i+1} - z^*\|_2 \leq \|z^i - z^*\|_2. \tag{84}$$

In words, we can only get closer to a minimizer as the iterations progress. Another consequence of (84) is that for given z^0, the set consisting of iterates of the algorithm $\{z^i\}_{i\in\mathbb{N}}$ is bounded. Since $z^i \in \mathbb{R}^n$, we can therefore extract a convergent subsequence by the Bolzano-Weierstrass theorem (see [28], Thm. 2.42). Suppose $\{z^{i_k}\}_{k\in\mathbb{N}}$ is one such subsequence and that $z^{i_k} \to z^\infty$. If we can show that this z^∞ is a fixed point of the operator, then, by (84), we would have that $\|z^i - z^*\|$ is monotone decreasing and convergent to 0, i.e. the algorithm in fact converges to a minimizer. The critical point is then to show that $z^\infty \in Z^*$.

Lemma 6. *Suppose that $2\gamma I - A\, A^T > 0$. For an arbitrary fixed z^0, define $z^{i+1} = \mathcal{M}_\gamma(z^i)$. Then, the cluster points of the sequence $\{z^i\}_{i\in\mathbb{N}}$ are in Z^*.*

Proof: We first remark that if $2\gamma I - A\, A^T > 0$, then $\sigma := \rho(I - \gamma^{-1}\, A\, A^T) < 1$ on the range space of A. We also note that if $\tilde{z}$ and $\hat{z}$ are in Z^*, then $A^T\, \tilde{z} = A^T\, \hat{z}$.

Now let z^∞ be a cluster point of $\{z^i\}_{i\in\mathbb{N}}$. From this sequence, extract a convergent subsequence $z^{i_k} \to z^\infty$. Suppose $z^\infty \notin Z^*$. This will lead to a contradiction.

Since $z^\infty \notin Z^*$, we can find $z^* \in Z^*$ with $\|z^\infty - z^*\| = d > 0$. Let us denote the projector to the range space of A as P and denote $P^\perp := I - P$. Then, since $A^T z^* \neq A^T z^\infty$, (otherwise z^∞ would be a minimizer), we have

$$\|P(z^* - z^\infty)\| := d_1 > 0, \tag{85}$$
$$\|P^\perp(z^* - z^\infty)\| := d_2 \geq 0 \tag{86}$$

and $d^2 = d_1^2 + d_2^2$. Now for any given $\epsilon > 0$ we can find an integer $K(\epsilon)$ such that if $k \geq K(\epsilon)$, we have $\|P(z^{i_k} - z^\infty)\| < \epsilon$, $\|P^\perp(z^{i_k} - z^\infty)\| < \epsilon$. Notice that in this case,

$$\|P(z^{i_k} - z^*)\| < d_1 + \epsilon, \tag{87}$$
$$\|P^\perp(z^{i_k} - z^*)\| < d_2 + \epsilon. \tag{88}$$

Since $\mathcal{M}_\gamma(z^*) = z^*$ and $P_K(\cdot)$ is nonexpansive, using the definition of $\mathcal{M}_\gamma(\cdot)$, we have

$$\begin{aligned}
\|\mathcal{M}_\gamma(z^{i_k}) - z^*\|^2 &= \|\mathcal{M}_\gamma(z^{i_k}) - \mathcal{M}_\gamma(z^*)\|_2^2 \\
&\leq \|(I - \gamma^{-1}\, A\, A^T)\left\{(z^{i_k} - z^*)\right\}\|_2^2 \\
&= \|P(I - \gamma^{-1}\, A\, A^T)\left\{(z^{i_k} - z^*)\right\}\|^2 \\
&\quad + \|P^\perp(I - \gamma^{-1}\, A\, A^T)\left\{(z^{i_k} - z^*)\right\}\|^2 \\
&= \|(I - \gamma^{-1}\, A\, A^T)P\, \left\{(z^{i_k} - z^*)\right\}\|^2 \\
&\quad + \|(I - \gamma^{-1}\, A\, A^T)P^\perp \left\{(z^{i_k} - z^*)\right\}\|^2 \\
&\leq \sigma^2\, (d_1 + \epsilon)^2 + (d_2 + \epsilon)^2.
\end{aligned}$$

Now if,

$$\sigma^2\, (d_1 + \epsilon)^2 + (d_2 + \epsilon)^2 \leq d_1^2 + d_2^2, \tag{89}$$

then we would have $\|z^{i_k+1} - z^*\| \leq \|z^\infty - z^*\|$ and since $\|z^{i_{k+r}} - z^*\| \leq \|z^{i_k+1} - z^*\|$ for any positive integer r, it would follow that $z^{i_k} \nrightarrow z^\infty$, a contradiction that would conclude the proof.

Let us see if we can make (89) true by selecting ϵ small enough. Notice that (89) is equivalent to,

$$\sigma^2 \leq \frac{d_1^2 + d_2^2 - (d_2 + \epsilon)^2}{(d_1 + \epsilon)^2} = \frac{1 - \epsilon(2d_2 + \epsilon)/d_1^2}{1 + \epsilon(2 + \epsilon)/d_1^2}. \tag{90}$$

But the term on the rhs goes to 1 as ϵ goes to 0 and therefore we can find some ϵ such that (89) holds, and the proof follows. ■

By the discussion preceding Lemma 6, we have

Proposition 4. *If $2\gamma I - A\, A^T > 0$, then Algorithm 1 converges to a minimizer.*

References

[1] İ. Bayram. Mixed-norms with overlapping groups as signal priors. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2011.

[2] İ. Bayram. On convex combinations of norms for group sparsity. In *Proc. IEEE Conference on Signal Processing and Communications Applications (SİU)*, 2011.

[3] A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Trans. Image Processing*, 18(11):2419–2434, 2009.

[4] J. Bobin, J.-L. Starck, M.J. Fadili, and Y. Moudden. Sparsity, morphological diversity and blind source separation. *IEEE Trans. Image Processing*, 16(11):2662–2674, November 2007.

[5] A. Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20(1-2):89–97, January-March 2004.

[6] A. Chambolle. Total variation minimization and a class of binary MRF models. *Lecture notes in computer science*, 3757:136–152, 2005.

[7] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM J. Multiscale Model. Simul.*, 4(4):1168–1200, November 2005.

[8] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. Pure Appl. Math.*, 57(11):1413–1457, August 2004.

[9] R. L. Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78:837–842, 1983.

[10] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[11] M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse Problems*, 23(3):947–968, June 2007.

[12] M. A. T. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak. Majorization-minimization algorithms for wavelet-based image restoration. *IEEE Trans. Image Processing*, 16(12):2980–2991, December 2007.

[13] M. A. T. Figueiredo and R. D. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Trans. Image Processing*, 12(8):906–916, August 2003.

[14] J.-J. Fuchs. Extension of the global matched filter to structured groups of atoms: Application to harmonic signals. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2011.

[15] N. Gaffke and R. Mathar. A cyclic projection algorithm via duality. *Metrika*, 36:29–54, 1989.

[16] S.-P. Han. A successive projection method. *Mathematical Programming*, 40:1–14, 1988.

[17] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[18] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundemantals of Convex Analysis*. Springer, 2001.

[19] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *Amer. Statist.*, 58(1):30–37, February 2004.

[20] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proc.* 26^{th} *Int. Conf. on Machine Learning*, 2009.

[21] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2010.

[22] M. Kowalski. Sparse regression using mixed norms. *J. of Appl. and Comp. Harm. Analysis*, 27(3):303–324, November 2009.

[23] M. Kowalski and B. Torrésani. Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients. *Signal, Image and Video Processing*, 3(3):251–264, September 2009.

[24] D. G. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley, 1984.

[25] C. De Mol and M. Defrise. A note on wavelet-based inversion methods. In M. Z. Nashed and O. Scherzer, editors, *Inverse Problems, Image Analysis and Medical Imaging*. American Mathematical Society, 2002.

[26] R. T. Rockafellar. *Conjugate Duality and Optimization*. SIAM, 1974.

[27] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60(1-4):259–268, November 1992.

[28] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 3^{rd} edition, 1976.

[29] I. W. Selesnick. Resonance-based signal decomposition: A new sparsity-enabled signal analysis method. *Signal Processing*, 2010.

[30] I.W. Selesnick and M. A. T. Figueiredo. Signal restoration with overcomplete wavelet transforms: comparison of analysis and synthesis priors. In *Proc. SPIE, vol 7446 (Wavelets XII)*, 2009.

[31] J.-L. Starck, M. Elad, and D. Donoho. Image decomposition via the combination of sparse representation and a variational approach. *IEEE Trans. Image Processing*, 14(10):1570–1582, October 2005.

[32] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B*, 58(1):267–288, 1996.

[33] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68:49–67, 2006.

[34] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.

[35] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67(2):301–320, April 2005.