MIXED NORMS WITH OVERLAPPING GROUPS AS SIGNAL PRIORS

İlker Bayram

İstanbul Technical University,
Dept. of Electronics and Communication Eng.,
Maslak, 34469, İstanbul, Turkey

ABSTRACT

In a number of signal processing applications, problem formulations based on the ℓ_1 norm as a sparsity inducing signal prior lead to simple algorithms with good performance. However, ℓ_1 norm is not flexible enough to handle certain signal structures that are represented using a few *groups* of coefficients. Formulations that make use of mixed norms provide an alternative that can handle such signals by forcing sparsity on a group level and allowing non-sparse distributions within the groups. However, conventional mixed norms allow only non-overlapping groups – a restriction that leads to characteristics unlikely for natural signals. In this paper, we investigate mixed norms with overlapping groups. We consider a simple denoising formulation that gives a convex optimization problem and provide an algorithm that solves the problem. We use the algorithm to evaluate the performance of mixed norms with overlapping groups as signal priors.

Index Terms— Mixed norm, Minkowski sum, denoising, rational dilation wavelets.

1. INTRODUCTION

Signal processing based on sparsity measures has received significant attention in the last two decades. In particular, it has been observed that the ℓ_1 norm as a signal prior leads to simple formulations/algorithms with good performance. However, from a Bayesian perspective, ℓ_1 norm implicitly assumes that the signal coefficients are independent. Since it has been observed that there is some correlation between groups of coefficients, this points to a shortcoming of the ℓ_1 norm as a signal prior. 'Mixed norms' provide an alternative that addresses this issue, retaining at the same time, the simplicity of the ℓ_1 norm. For certain choices of the parameters, mixed norms encourage coefficients to form groups. Within the groups, the coefficients are allowed to follow non-sparse distributions and the sparsity requirement is in force only on a group level. In other words, it is the number of non-zero groups that matters rather than the separate coefficients.

Mixed norms are defined as follows. Given a set of coefficients $\mathbf{x} = \{x_1, \ldots, x_n\}$, we form J groups with K coefficients each[1] where the k^{th} coefficient of the j^{th} group is given by $x_{n(j,k)}$. In this setting, the $\ell_{p,q}$ norm of $\mathbf{x}$ is given by,

$$\|\mathbf{x}\|_{p,q} = \left(\sum_{j=1}^{J} \left(\sum_{k=1}^{K} |x_{n(j,k)}|^p \right)^{q/p} \right)^{1/q}. \tag{1}$$

[1]The number of coefficients in each group could be different. Here, this choice is made to keep the notation simpler.

In the definition of the groups above, we note that the index function, $n(j,k)$, might or might not be invertible. If $n(j,k)$ is invertible, the groups do not overlap, i.e. a coefficient belongs to a single group at most. This case has been investigated in a number of publications (see [1, 2, 3] and the references therein).

We remark that when $n(j,k)$ is invertible, i.e. when the groups do not have overlaps, the groups 'communicate' only over the outer sum in (1). For $q = 1$, which will be of interest here, this leads to an independence assumption among the groups. Whether a group of coefficients is kept or discarded, is decided regardless of the 'neighboring' groups. We think that it is unlikely for natural signals to conform to such predefined rigid-group structures. Moreover, inter-group independence is likely to lead to 'blocking effects'. Therefore, it is plausible to consider groups with overlaps, which can be achieved by letting $n(j,k)$ be non-invertible.

Use of overlapping or non-overlapping groups is not merely a matter of choice. When groups have overlaps, certain modifications are required in the algorithms (see [4] for an interesting discussion of an heuristic algorithm). In the sequel, we will discuss this for a denoising problem (which could also be used in a linear inverse problem setting by employing the 'Majorization-Minimization' method – see [5]). We restrict our attention to the $p = 2$, $q = 1$ case, which allows simple algorithms for the formulation we consider. Our intent is not to seek a state-of-the-art denoising method, but rather to investigate the behavior of $\ell_{2,1}$ norms with overlapping groups as signal priors and compare them to ℓ_1 and $\ell_{2,1}$ norms with non-overlapping groups.

2. PROBLEM FORMULATION

Given noisy observations of a signal, $\mathbf{y}$, the $\ell_{p,q}$-regularized denoising formulation we consider is

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \ \frac{1}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_{p,q}. \tag{2}$$

Here $\hat{\mathbf{x}}$ denotes the estimate of the underlying signal. With non-overlapping groups, for $p > 1$, $q = 1$, the denoising formulation in (2) does indeed achieve the desired effect: sparsity only on a group level (see [1] for a more detailed discussion). In particular, for $p = 2$, $q = 1$, and invertible index function $n(\cdot,\cdot)$, the minimizer of the above functional is given by,

Algorithm 1. *For $j = 1, \ldots, J$, set*

$$\hat{\mathbf{x}}_j = \text{soft}\{\|\mathbf{y}_j\|_2, \lambda\} \frac{\mathbf{y}_j}{\|\mathbf{y}_j\|_2}, \tag{3}$$

where

$$\text{soft}(z, \lambda) = \text{sgn}(z) \ \max\{|z| - \lambda, 0\} \tag{4}$$

and $\mathbf{y}_j$ is the j^{th} group, i.e. $\mathbf{y}_j = (y_{n(j,1)}, \ldots, y_{n(j,K)})$.

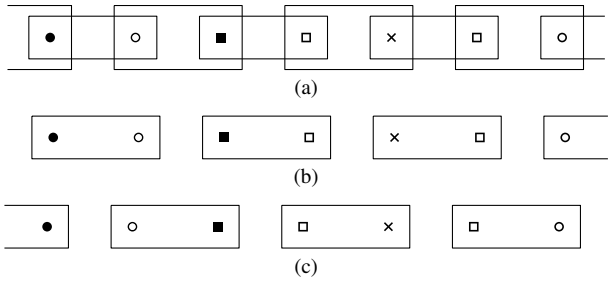


Fig. 1: (a) Groups with overlaps. This collection of groups can be decomposed into two non-overlapping collection of groups as depicted in (b) and (c).

Let us now turn to mixed norms with overlapping groups. Consider a grouping system as shown in Fig. 1(a). We can decompose this system into two subsystems with non-overlapping groups as depicted in Fig. 1(b,c). If we denote the $\ell_{p,q}$ norm according to the group systems in Fig 1(a,b,c) as $\|\cdot\|_a$, $\|\cdot\|_b$, $\|\cdot\|_c$ respectively, thanks to $q = 1$, we have $\|\cdot\|_a = \|\cdot\|_b + \|\cdot\|_c$. Thus we can write the functional

$$J(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_a, \tag{5}$$

as

$$J(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_b + \lambda\|\mathbf{x}\|_c. \tag{6}$$

We remark that if either of $\|\mathbf{x}\|_c$ or $\|\mathbf{x}\|_b$ was missing in (6), we could minimize $J(\mathbf{x})$ using Algorithm 1. Thus we can assume that we have two mappings $M_b(\mathbf{y})$, $M_c(y)$ defined as,

$$M_b(\mathbf{y}) = \operatorname*{argmin}_x \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_b, \tag{7}$$

$$M_c(\mathbf{y}) = \operatorname*{argmin}_x \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_c. \tag{8}$$

Given these, an algorithm that minimizes (6) is,

Algorithm 2. *Initialize $\mathbf{z}_1$, $\mathbf{z}_2$ by setting them to zero.*

(I) Repeat until some convergence criterion is met,

$$\mathbf{z}_1 = \mathbf{y} - \mathbf{z}_2 - M_b(\mathbf{y} - \mathbf{z}_2), \tag{9}$$

$$\mathbf{z}_2 = \mathbf{y} - \mathbf{z}_1 - M_c(\mathbf{y} - \mathbf{z}_1), \tag{10}$$

(II) Set $\hat{\mathbf{x}} = \mathbf{y} - (\mathbf{z}_1 + \mathbf{z}_2)$.

In general, given a more complicated group system than that of Fig. 1(a), we can always decompose it into non-overlapping subsystems, possibly using more than two such subsystems. In that case, Algorithm 2 will utilize more than two $M(\cdot)$ functions. We will consider the general problem in the next section.

One remark about Algorithm 2 is in order. Apparently, the algorithm is known (see [6] for a derivation and an application to dictionary learning). However, we are not aware of a previous description in this context and therefore we provide a derivation in the following for convenience. The derivation makes use of the dual problem which is easier to solve.

3. THE DUAL-PROBLEM AND ITS SOLUTION

Notation

In the rest of the paper, all of the variables are assumed to be vectors in $\mathbb{R}^n$. We will not need to refer to particular entries of these vectors so we use regular small case letters for vectors. The subscripts will be used to differentiate between different vectors.

We first note that we can rewrite (see [7]) any norm $\|\cdot\|$ as the support function $\sigma_K(\cdot)$ of some closed convex set K, which is defined by,

$$\sigma_K(\cdot) := \sup_{z\in K}\langle z, \cdot\rangle. \tag{11}$$

For example, $\|\cdot\|_p = \sigma_{B_q}(\cdot)$ where B_q denotes the unit ball according to the ℓ_q norm with $p^{-1} + q^{-1} = 1$.

Therefore, given $y \in \mathbb{R}^n$, we can rewrite the original problem as the minimization of

$$J(x) = \frac{1}{2}\|y - x\|_2^2 + \lambda_1\,\sigma_{K_1}(x) + \ldots + \lambda_n\,\sigma_{K_n}(x) \tag{12}$$

where K_i's are convex sets and $\sigma_{K_i}(\cdot)$ is the support function of K_i. Here we take

$$\|\cdot\|_{2,1} = \lambda_1\,\sigma_{K_1}(\cdot) + \ldots + \lambda_n\,\sigma_{K_n}(\cdot), \tag{13}$$

where the mixed norm (which appears in the original problem (2)) is defined according to some grouping system. Now, assuming that we know how to minimize

$$J_i(x) = \frac{1}{2}\|y - x\|_2^2 + \lambda_i\,\sigma_{K_i}(x) \tag{14}$$

or, equivalently (see Prop. 1), that we know how to find the projection of any point z onto $\lambda_i K_i$, for $i \in \{1, \ldots, n\}$, the following algorithm can be used to obtain the minimizer of (12). In the following, we denote the projection operator onto a set C as $\mathrm{Proj}_C(\cdot)$.

Algorithm 3. *Initialize $z_k \in (\lambda_k K_k)$ for $k = 1, 2, \ldots, n$, and the iteration count $i = 1$.*

(I) For $k = 1$ to n,

(i) Set
$$v = y - \sum_{\substack{m=1,\ldots,n\\ m\neq k}} z_m. \tag{15}$$

(ii) Update $z_k = \mathrm{Proj}_{\lambda_k K_k}(v)$.

(II) Set $x_i = y - \sum_{m=1}^n z_m$, update $i = i + 1$, go to (II).

The algorithm produces a sequence $\{x_i\}_{i=1}^{\infty}$ which converges to the minimizer of $J(x)$ in (12).

Algorithm 3 implies Algorithm 2 when only two support functions σ_{K_i} appear in (13). In general, since the dual formulation is simpler to describe, and we lack further space, we will not translate Algorithm 3 to the primal formulation, which could be done straightforwardly.

4. DERIVATION AND CONVERGENCE OF THE ALGORITHM

We first note that $\lambda\,\sigma_C(x) = \sigma_{(\lambda\,C)}(x)$. It also follows from the definition of a support function that $\sigma_{C_1}(x) + \sigma_{C_2}(x) = \sigma_{C_1+C_2}(x)$. Thus, $J(x)$ in (12) can be written as,

$$J(x) = \frac{1}{2}\|y - x\|_2^2 + \sigma_K(x) \tag{16}$$

where $K = \lambda_1 K_1 + \lambda_2 K_2 + \ldots + \lambda_n K_n$. This problem can be solved easily if we know how to project any given vector to the set K. Let the projection operator onto the set K be denoted as $\mathrm{Proj}_K(\cdot)$. The following proposition has been used by a number of authors [8, 9]. We omit the proof.

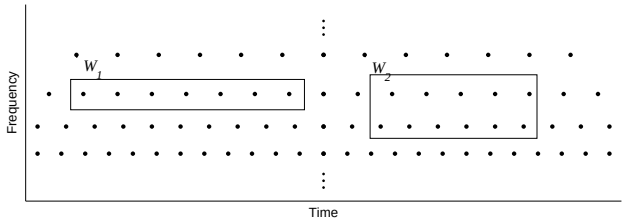


Fig. 2: Time-Frequency sampling pattern of the wavelet basis with dilation factor $d = 6/5$. The rectangular windows W_1 and W_2 are used to define group systems for the mixed $\ell_{2,1}$ norms.

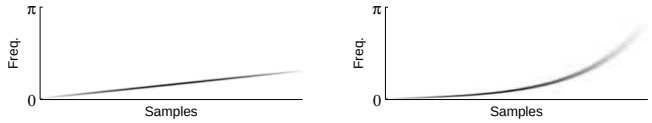


Fig. 3: Spectrograms of the two chirp signals used in Experiment 1. Left : 'Linear chirp'. Right : 'Logarithmic Chirp'.

Proposition 1. *Let C be a closed, convex set. Suppose we are given some $y \in \mathbb{R}^n$. The convex functional*

$$J(x) = \frac{1}{2}\|y - x\|_2^2 + \sigma_C(x) \tag{17}$$

achieves its unique minimum at $x^* = y - \mathrm{Proj}_C(y)$.

This proposition is useful bothways. We may know how to project onto C *or* how to minimize $J(\cdot)$. The proposition can be invoked to pass from the minimizer of $J(\cdot)$ to the projection and vice versa. In the following, we will have this in mind when we talk about projections – given a set and a point y, we will assume that we either know how to minimize the associated cost function $J(\cdot)$ or project y onto the set.

For our task of minimizing (12), under assumption (14), the proposition allows us to transform the original problem into :

Problem 1. *Let $K = \lambda_1 K_1 + \lambda_2 K_2 + \ldots + \lambda_n K_n$. Given $\mathrm{Proj}_{\lambda_i K_i}(\cdot)$, how can we construct $\mathrm{Proj}_K(\cdot)$?*

In order to see the convergence of the algorithm, we first write the projection as a minimization algorithm.

$$\mathrm{Proj}_K(y) = \underset{z \in K}{\mathrm{argmin}} \|y - z\|_2^2. \tag{18}$$

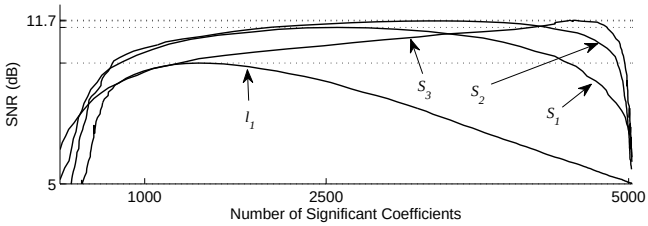
We remark that $z \in K$ if and only if it can be written as,

$$z = \sum_{i=1}^{n} z_i \qquad \text{for some } z_i \in K_i, \quad i = 1, \ldots, n. \tag{19}$$

Therefore, if we define the convex function

$$F_y(z_1, z_2, \ldots, z_n) = \begin{cases} \left\|y - \sum_{i=1}^n z_i\right\|_2^2 & \text{if } z_i \in K_i \quad \forall i \in \{1, \ldots, n\}, \\ \infty & \text{if } z_i \notin K_i \text{ for some } i \in \{1, \ldots, n\}. \end{cases}$$

we can write, $\mathrm{Proj}_K(y) = \mathrm{argmin}_{z_1, z_2, \ldots, z_n} F_y(z_1, z_2, \ldots, z_n)$. With this notation, Algorithm 3 can be seen to be a coordinate descent type algorithm [10] for the cost function F_y.



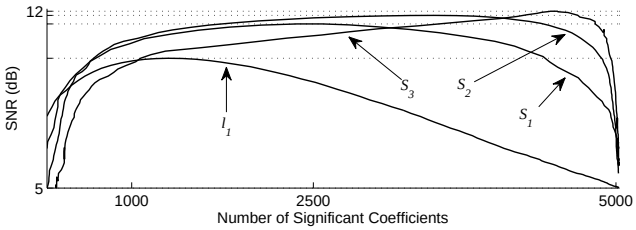


Fig. 4: Output SNR's for Experiment 1. A coefficient is counted as 'significant' if its magnitude exceeds 0.001. Top panel : Linear chirp. Bottom panel : Logarithmic Chirp. For the linear chirp, the the tick marks, which show the best SNR values obtained with S_3, S_2, S_1 and ℓ_1 norms are 11.72, 11.64, 11.28 and 9.93 dB respectively. For the quadratic chirp, the best SNR values are (in the same order) 12.03, 11.87, 11.59 and 10.12 dB.

5. EXPERIMENTS

For our experiments, we used an orthonormal wavelet basis with a dilation factor $d = 6/5$ (see [11] for a more detailed discussion). This is *roughly* a wavelet basis of the form $\{d^{n/2}\,\psi(dx - k)\}_{n,k \in \mathbb{Z}}$. This basis yields a time-frequency (T-F) sampling pattern as depicted in Fig. 2. To define the groups, we used two different windows defined on this T-F lattice. The first window, W_1, hosts six coeficients in a particular subband of the wavelet transform. The second window, W_2, hosts coefficients from two consecutive subbands (six from the coarser, five from the finer). We anticipated that W_1 would be effective for capturing ridges that change subbands slowly whereas W_2 would be suitable for more complicated structures involving neighboring subband interactions.

Using W_1, we constructed two grouping systems. The first one, denoted by S_1, employs non-overlapping groups obtained by shifting W_1 along the frequency and time axes (along the time axis, each shift of the window is by six coefficients)[2]. The second one, denoted by S_2, employs all shifts of the window W_1 on the T-F lattice and therefore contains overlapping groups (overlaps along the frequency axis only). Using the second window, W_2, we constructed a single grouping system, denoted by S_3, which consists of the groups obtained by considering all the shifts of W_3 on the T-F plane (thus overlaps are along the frequency *and* time axes). In the following, we will denote the $\ell_{2,1}$ norms associated with these grouping systems as $\|\cdot\|_{S_1}$, $\|\cdot\|_{S_2}$, $\|\cdot\|_{S_3}$.

Experiment 1. Our first experiment involves two artificial chirp signals and aims to highlight the difference between the overlapping group systems S_2 and S_3 above. The first chirp signal is linear and sweeps the frequencies between $[\pi/80, 5\pi/16]$ rad/samples in 5000 samples. The second chirp signal is 'logarithmic' and sweeps the frequencies between $[\pi/80, 7\pi/8]$ rad/samples in the same number

[2] As the windows move along the frequency axis, their widths and heights are scaled using the dilation factor so as to keep the number of included coefficients the same.

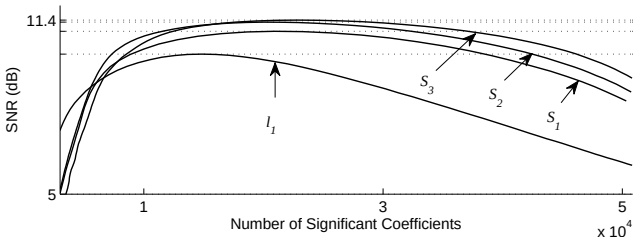


Fig. 5: SNR outputs for Experiment 2. A coefficient is counted as 'significant' if its magnitude exceeds 0.01. The input is a noisy speech signal (SNR = 5dB). The best SNR values obtained by using S_3, S_2, S_1 and ℓ_1 norms are 11.35, 11.29, 10.99 and 10.11 dB respectively.

of samples. The spectrogram of the linear and logarithmic chirp signals are depicted in the left and right panels of Fig. 3.

We consider the formulation,

$$\hat{x} = \frac{1}{2} \operatorname*{argmin}_x \|y - x\|_2^2 + \lambda \|x\| \tag{20}$$

where $\|\cdot\|$ is one of $\|\cdot\|_1$ (the regular ℓ_1 norm), $\|\cdot\|_{S_1}$, $\|\cdot\|_{S_2}$, $\|\cdot\|_{S_3}$. Taking as input the chirp signals contaminated with white noise (input SNR = 5dB), we varied λ to obtain reconstructions with different number of non-zero coefficients. The output SNR, with respect to the number of coefficients whose magnitude exceed 0.001 (the greatest coefficient magnitude is ≈ 3.8) are depicted in Fig. 4 for the linear and logarithmic chirp in the lower and upper panel respectively. For both signals, we see that the best SNR with the mixed norms is clearly higher than that of the ℓ_1 norm. Again in both cases, S_2 and S_3 (overlapping groups) perform better than S_1 (non-overlapping groups). We repetitively observed this behavior in our experiments. For the linear chirp, where the ridge moves slowly along the frequency axis, we see that S_3 performs only slightly better (by 0.08 dB) than S_2. However, for the logarithmic chirp, where the ridge movement along the frequency axis is faster, the performance gap between S_3 (with its groups hosting coefficients from neighboring subbands) and S_2 is wider (0.16 dB). Even though the improvement in the performance difference between S_2 and S_3 for the quadratic chirp is less than we expected, we believe that the results could be further improved by modifying the structure of the groups.

One observation common to both cases is, as we progress from S_1 to S_3, the best reconstruction becomes less sparse. This is more or less expected as one switches from ℓ_1 to S_1, because mixed norms put less emphasis on sparsity (this can also be observed in Fig. 3 of [1]). What demands an explanation is the observation that the best reconstruction with S_3 uses almost all of the coefficients and with fewer coefficients, the performance achieved with this norm degrades quickly. We think that this unexpected behavior stems from the rigidity in the 'sparsity measure' – here the number of coefficients that exceed a certain threshold. Indeed, when we look at the histogram of the coefficients, those obtained via S_3 have a slightly higher kurtosis (than those obtained using the other norms) indicating a sparse distribution. In the following experiment, we increase the threshold for a coefficient to be counted 'significant' and we observe that the number of coefficients for the best reconstructions with the ℓ_1 and S_3 norms get even closer.

Experiment 2. In our second experiment, we took as input a speech signal (4 sec., 16000 samples/sec.) contaminated with white noise (input SNR = 5dB) using the same formulation and the norms in Experiment 1. The outputs SNR's are shown in Fig. 5. ℓ_1 norm is best if one restrics herself to few coefficients. However, mixed norms are able to extract more out of the signal, when more coefficients are allowed. Of course, the results are dependent on the particular basis chosen (in particular the dilation factor) but we can nevertheless conclude that overlapping groups enhance performance and the difference between S_2 and S_3, even though slight, encourages inter-subband group systems.

6. CONCLUSION

In this paper, we evaluated the performance of mixed norms with overlapping groups as signal priors. We considered a particular denoising formulation, that leads to a convex minimization problem, in order to evaluate different grouping systems and described an algorithm that solves that minimization problem. Our simulation results on synthetic as well as real signals demonstrate the potential of such priors and provide a motivation to optimize the grouping system. Inverse problem formulations based on wavelet domain regularization could also benefit from such optimization.

7. REFERENCES

[1] M. Kowalski, "Sparse regression using mixed norms," *J. of Appl. and Comp. Harm. Analysis*, vol. 27, no. 3, pp. 303–324, Nov. 2009.

[2] M. Kowalski and B. Torrésani, "Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients," *Signal, Image and Video Processing*, vol. 3, no. 3, pp. 251–264, Sept. 2009.

[3] M. Szafranski, Y. Grandvalet, and P. Morizet-Mahoudeaux, "Hierarchical penalization," in *Advances in Neural Information Processing Systems, vol. 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. MIT Press, 2008.

[4] M. Kowalski and B. Torrésani, "Structured sparsity : From mixed norms to structured shrinkage," in *Proc. Signal Proc. with Adaptive Sparse Structured Representations (SPARS)*, 2009.

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

[6] 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.

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

[8] A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical Imaging and Vision*, vol. 20, no. 1-2, pp. 89–97, January-March 2004.

[9] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Trans. Image Processing*, vol. 18, no. 11, pp. 2419–2434, 2009.

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

[11] T. Blu, "Iterated filter banks with rational rate changes–connection with discrete wavelet transforms," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3232–3244, Dec. 1993.