Sparsity Within and Across Overlapping Groups

İlker Bayram, *Senior Member, IEEE*

***Abstract*—Recently, penalty functions promoting signals that are sparse within and across groups have been proposed. In this letter, we propose a modified penalty function that offers additional flexibility in forming groups. We study the properties of the penalty function and propose a new algorithm that can be used in energy minimization formulations that employ it. We demonstrate the effects of using the penalty function on a simple linear inverse problem.**

***Index Terms*—Structured sparsity, elitist LASSO, exclusive LASSO, sparsity within and across groups**

I. Introduction

Sparsity has played a major role in signal processing in the last two decades. However, for many natural signals, plain sparsity falls short of capturing the intrinsic characteristics of the signal of interest. In recent work [4], we addressed a specific form of sparsity, useful for signals that are composed of a few number of groups where within each group, only a few coefficients are active. We call this characteristic 'sparsity within and across groups' (SWAG) (see also [23]). In this letter, we propose a modification of this penalty to introduce further flexibility in the definition of the groups.

A. The SWAG Penalty and Threshold Function

The SWAG penalty in [4] is a group-separable function. Suppose we are given a vector $x = (x_1, x_2, \ldots, x_N)$. We partition x into x^1, x^2, …, x^k. Each x^m is a collection of distinct variables from x, referred to as a group. The SWAG penalty in [4] is defined as

$$P(x) = \|x\|_1 + \frac{\gamma}{2} \sum_m \sum_{\substack{i,j \\ i\neq j}} |x_i^m \, x_j^m|. \tag{1}$$

The associated threshold function, or proximity operator [1], [11] is defined as

$$T(z) = \arg\min_{x\in\mathbb{C}^n} \frac{1}{2}\|z-x\|_2^2 + \lambda\, P(x). \tag{2}$$

$T(z)$ is well-defined if $\lambda\gamma < 1$, in which case it can be computed with a finite terminating procedure [4].

In the SWAG penalty, the groups are required to be non-overlapping. If the groups share variables, the penalty/threshold function is no longer group-separable, and a finite terminating procedure for realizing the threshold function is not readily available. In that case, one way to realize the threshold function is to split variables and employ group-separable penalties iteratively in a splitting scheme such as the Douglas-Rachford algorithm [10], [11], or ADMM [6]. Other than the increase in the number of variables, such an approach may not be feasible because some formulations may require to compute infinite iterations within iterations – a procedure not realizable in principle.

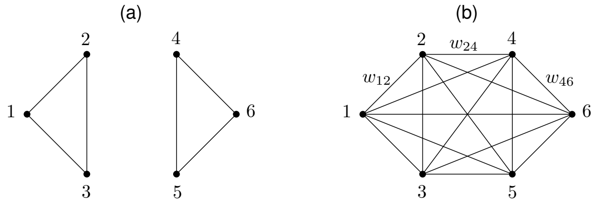

Fig. 1. A visual description of the group structure using graphs. Each variable is represented by a node. (a) This graph represents the partition $x^1 = (x_1, x_2, x_3)$, $x^2 = (x_4, x_5, x_6)$. (b) The proposed generalization employs a weighted complete graph.

B. The Proposed Penalty

The proposed generalization is easy to describe using graphs. Consider a vector $x = (x_1, x_2, \ldots, x_6)$. Suppose we partition x as $x^1 = (x_1, x_2, x_3)$, $x^2 = (x_4, x_5, x_6)$. This partition can be represented by the graph in Fig. 1a. Notice that each group is associated with a complete graph. Since the groups do not share variables, there are two disjoint complete graphs.

The generalization we propose in this letter is to use a complete weighted graph, as shown in Fig. 1b. The modified penalty on $\mathbb{C}^n$ is then defined as

$$P_W(x) = \|x\|_1 + \frac{1}{2}\sum_{i,j} w_{ij}\,|x_i\, x_j|, \tag{3}$$

For a specific choice of the weight matrix W, we can recover the penalty (1). Therefore, P_W is a generalization of P in (1).

The associated threshold function is defined as,

$$T_{\lambda,W}(z) = \arg\min_{x\in\mathbb{C}^n} \Big\{ D_{\lambda,W}(x;z) = \frac{1}{2}\|z-x\|_2^2 + \lambda\, P_W(x) \Big\}. \tag{4}$$

We show in this letter that $T_{\lambda,W}(z)$ is well-defined, if λ satisfies an upper bound determined by W.

C. Related Work and Contribution

The sparsity characteristic sought in this letter is different than that sought in many papers using group-based penalty functions. Specifically, [28], [20], [19], [2], [8] describe approaches for promoting signals that can be represented with a few groups, where in a non-zero group, all of the coefficients are non-zero. In contrast, [20], [21], [29], [26], [23], [4] host approaches that aim a similar characteristic as the proposed penalty. In this latter collection, the SWAG penalty [4], which we aim to generalize, separates from the rest in that it is

İ. Bayram was with the Dept. of Electronics and Communications Eng., Istanbul Technical University, Istanbul, Turkey. He is now with Analog Devices Inc. / Analog Garage, Boston, MA, USA. E-mail : ibayram@ieee.org.

a non-convex penalty. In [4], it was argued that this helps reduce the bias in the non-zero estimates produced by the threshold function (see also [7], [25] for related discussions). For a more detailed comparison between the SWAG penalty and the penalties in [20], [21], [29], we refer to [4].

An interesting difference, pointed out by one of the reviewers, between the SWAG penalty and Sparse Group Lasso (SGL) [26], [23] concerns the distribution of non-zero variables across groups. The SGL penalty consists of the sum of an ℓ_1 norm, and the sum of the ℓ_2 norms of the groups (i.e., an $\ell_{2,1}$ norm [20]). Because of the $\ell_{2,1}$ norm, SGL prefers fewer non-zero groups, where within each active group, multiple variables may be non-zero. In contrast, the SWAG penalty in (1) favors solutions where the number of active variables in each group is smaller but the number of active groups is larger. While this appears to be a fundamental difference, the effect also depends on how the groups are formed. If variables that are associated with similar responses are grouped together, then we expect this difference to be reduced.

The proposed modification to the SWAG penalty aims to introduce further flexibility in forming the groups. First, groups are allowed to overlap. Second, while the original SWAG penalty in [4] uses constant weights within each group, the proposed penalty allows the weights within a group to vary. These in turn allow to achieve a more localized and translation-invariant behavior, which is of interest for processing time-domain signals. However, these modifications come at an expense. Unlike the SWAG threshold function, the threshold function for the proposed penalty cannot be computed with a finite terminating procedure. Therefore, forward-backward splitting type algorithms that might utilize $T_{\lambda,W}$ [11], [12], [14], [3] are not readily applicable for the proposed penalty. We describe instead a descent algorithm for a generic formulation that employs the proposed penalty.

Notation: For $x \in \mathbb{C}^n$, $|x|$ denotes the magnitude vector of the same size. Therefore, $\sum_{\substack{i,j \\ i\neq j}} w_{i,j}\,|x_i\,x_j| = |x|^T\,W\,|x|$. $\mathbf{1}$ denotes a vector of ones. For non-zero $z \in \mathbb{C}^n$, $e^{j\angle z}$ denotes a unit vector in the direction of z. For two vectors x, z in $\mathbb{C}^n$, $x\,z$ denotes the vector obtained by element-wise multiplication.

$\mathbb{C}^n$ appears as a domain for some functions in the letter. For inner products and gradients, we interpret $\mathbb{C}^n$ as $\mathbb{R}^{2n}$. Thus, on $\mathbb{C}^n$, we use the inner product $\langle x, y\rangle = \sum_i \operatorname{real}(x_i\,y_i^*)$.

Outline: In Section II, we derive a condition that ensures P_W is weakly-convex, which implies that T_W is well-defined. We discuss in Section III how to construct a descent algorithm when P_W is used in a simple minimization formulation. We demonstrate the utility of the proposed penalty in Section IV. Section V contains an outlook.

II. Weak Convexity of the Proposed Penalty

In this section, we study the proposed penalty function and show that it is weakly convex [27].

Definition 1. A function g is said to be *α-weakly convex* if $\dfrac{\alpha}{2}\,\|x\|_2^2 + g(x)$ is convex.

Our interest in showing the weak convexity of the proposed penalty stems from available schemes such as [25], [3] that make use of the weak convexity of the penalties. In addition, weak-convexity of P_W also implies that $T_{\lambda,W}$ is well-defined. To see this, observe that P_W is $1/\lambda$-weakly convex if and only if $D_{\lambda,W}(\cdot\,;z)$ is convex. In particular, if $D_{\lambda,W}(\cdot\,;z)$ is strictly convex, it has a unique minimizer, namely $T_{\lambda,W}(z)$. However we will later see in Prop. 3 that $T_{\lambda,W}$ is well-defined for an extended range of λ values than those implied by Prop. 1 of this section.

Strict convexity of $|x|^T\,(I+\lambda\,W)\,|x|$ implies strict convexity of $D_{\lambda,W}(x;z)$ with respect to x. But due to the magnitude operator $|\cdot|$, positive definity of $I+\lambda\,W$ does not automatically imply convexity of $|x|^T\,(I+\lambda\,W)\,|x|$. Nevertheless, if $I+\lambda\,W$ admits a decomposition of the form

$$I + \lambda\,W = R^T\,R, \text{ with } R_{i,j} \geq 0, \text{ for all } i, j, \tag{5}$$

then $D_{\lambda,W}$ is convex. To see this, observe that

$$|x|^T\,R^T\,R\,|x| = \sum_i \Bigl(\sum_j r_{ij}\,|x_j|\Bigr)^2. \tag{6}$$

Since $r_{ij} \geq 0$ for all i, j, the term enclosed in parentheses in (6) is convex for all i, from which convexity of $D_{\lambda,W}$ follows.

Matrices that admit a decomposition as in (5) are called completely positive [5]. Unfortunately, checking whether an arbitrary psd matrix is completely positive or not is not trivial when the size of the matrix exceeds 4×4 [16], [5]. However, it is relatively simple to find an upper bound for λ so that $I + \lambda W$ is completely positive [5].

Proposition 1. For a non-negative W, $D_{\lambda,W}(\cdot\,;z)$ is strictly convex if

$$\lambda\,\Bigl(\max_i \sum_j w_{i,j}\Bigr) < 1. \tag{7}$$

Sketch of a Proof. $D_{\lambda,W}$ can be expressed as

$$\Bigl[\frac{1}{2}\|z\|_2^2 - 2\langle z, x\rangle + \lambda\,\|x\|_1\Bigr] + \Bigl\{|x|^T\,(I+\lambda\,W)|x|\Bigr\}. \tag{8}$$

The term inside the square brackets is convex with respect to x, and $(I + \lambda\,W)$ can be shown to be completely positive, after some algebra (or, see Thm. 3 in [5]). □

This proposition implies that $T_{\lambda,W}$ is well-defined if (7) holds. However, even though $T_{\lambda,W}$ is well-defined, it is not easy to realize numerically. We discuss this issue in the sequel.

III. Descent Algorithms

We derive a descent algorithm for a problem of the form

$$\min_{x\in\mathbb{C}^n} \bigl\{C(x) = f(x) + \lambda\,P_W(x)\bigr\}, \tag{9}$$

where $f(\cdot) : \mathbb{C}^n \to \mathbb{R}$ is a convex function. Viewing $\mathbb{C}^n$ as $\mathbb{R}^{2n}$, we also assume that f is Fréchet-differentiable [1], [22], and its Fréchet-derivative, ∇f, is Lipschitz-continuous with parameter L, i.e.,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \text{ for all } x, y. \tag{10}$$

We will derive the algorithm based on the majorization-minimization scheme [18], [13].

Definition 2. A function $g : \mathbb{C}^n \to \mathbb{R}$ is said to be a *majorizer for $h : \mathbb{C}^n \to \mathbb{R}$ at x^** if

(i) $h(x^*) = g(x^*)$,
(ii) $h(x) \le g(x)$ for all $x \in \mathbb{C}^n$.

A. Majorizing '$f(\cdot)$'

By the assumptions on f, we have (see, e.g., Cor.18.14, (i)⇒(iv) in [1])

$$f(x) \le f(y) + \langle \nabla f(y), x-y \rangle + \frac{L}{2}\,\|x-y\|_2^2, \text{ for all } x, y. \tag{11}$$

Using this, we obtain a simple majorizer after some algebra.

Proposition 2. Suppose f is convex and ∇f is Lipschitz continuous with parameter L. If $\alpha \le 1/L$, then

$$\begin{aligned} C^k(x) = \frac{1}{2\alpha}\,\big\|x - \big(x^k - \alpha \nabla f(x^k)\big)\big\|_2^2 + \lambda\, P_w(x) \\ + \Big[f(x^k) - \frac{\alpha}{2}\,\|\nabla f(x^k)\|_2^2\Big] \end{aligned} \tag{12}$$

is a majorizer for $C(\cdot)$ at x^k.

If we set x^* to be a minimizer of $C^k(\cdot)$, then $C(x^*) \le C(x^k)$. However, to minimize $C^k(\cdot)$, we essentially need to solve (4), for which a numerical procedure is not readily available. Nevertheless, if $C^k(\hat{x}) \le C^k(x^k)$ for some $\hat{x}$, then $C(\hat{x}) \le C(x^k)$. In the following, we show that such a $\hat{x}$ can be found with a simple update rule.

B. Majorizing the Proposed Penalty

The condition $C^k(\hat{x}) \le C^k(x^k)$ is equivalent to $D_{\beta,W}(\hat{x}; z) \le D_{\beta,W}(x^k; z)$ for $z = x^k - \alpha \nabla f(x^k)$, and $\beta = \alpha\,\lambda$. Observe now that

$$D_{\beta,W}\big(|x|; |z|\big) = D_{\beta,W}\big(|x|\, e^{j\angle z}; z\big) \le D_{\beta,W}\big(x; z\big). \tag{13}$$

for all x, z in $\mathbb{C}^n$. This suggests that, instead of minimizing $D_{\beta,W}(x; z)$, we can consider

$$\min_{x \in \mathbb{R}_+^n} D_{\beta,W}(x; |z|). \tag{14}$$

On $\mathbb{R}_+^n$, $D_{\beta,W}(\cdot; |z|)$ is simply a quadratic function. This has the following consequence.

Proposition 3. If $I + \beta\, W$ is positive definite, then $D_{\beta,W}(\cdot; z)$ has a unique minimizer.

Proof. The problem in (14) can be expressed as

$$\min_{x \in \mathbb{R}_+^n} \frac{1}{2}\, x^T\, (I + \beta\, W)\, x - \big\langle |z| + \beta \mathbf{1},\, x \big\rangle. \tag{15}$$

If $I + \beta W$ is positive definite, then the cost function in (15) is strictly convex, and (15) has a unique solution, $x^* \in \mathbb{R}_+^n$. Suppose now that, for some x,

$$D_{\beta,W}(x^*\, e^{j\angle z}, z) \ge D_{\beta,W}(x, z). \tag{16}$$

We then have by (13) that $D_{\beta,W}(x^*, |z|) \ge D_{\beta,W}(|x|, |z|)$. By the uniqueness of the solution of (15), we conclude that $x^* = |x|$. Now, if $e^{j\angle x} \ne e^{j\angle z}$, then $D_{\beta,W}(|x| e^{j\angle z}, z) < D_{\beta,W}(x, z)$. Consequently,

$$D_{\beta,W}(x^* e^{j\angle z}, z) = D_{\beta,W}(|x| e^{j\angle z}, z) < D_{\beta,W}(x, z), \tag{17}$$

contradicting (16). Thus, $e^{j\angle x} = e^{j\angle z}$, and $x = |x|\, e^{j\angle z} = x^*\, e^{j\angle z}$, which implies the uniqueness claim. □

Prop. 3 extends the range implied by Prop. 1 over which $T_{\lambda,W}$ is well-defined. However, it does not imply strict convexity of $D_{\beta,W}(\cdot; z)$, as Prop. 1 does.

The problem in (14) is a constrained convex minimization problem. Thus, descent on (14) can be achieved by applying any finite number of iterations of the projected gradient algorithm (PGA) [15]. This observation leads to the following proposition.

Proposition 4. Suppose $I + \beta\, W$ is positive semi-definite with spectral norm σ; f is convex, ∇f is Lipschitz continuous with parameter L; and $\alpha \le 1/L$. Let $z = x^k - \alpha\, \nabla f(x^k)$, and $P_+(\cdot)$ denote the projection operator onto $\mathbb{R}_+^n$. Let $S : \mathbb{R}^n \to \mathbb{R}^n$ denote the operator that maps x to $\hat{x}$ where,

$$\hat{x} = P_+\left(x - \eta\left[\Big(I + \beta\, W\Big)x + \beta \mathbf{1} - |z|\right]\right). \tag{18}$$

Finally, let S^m denote S iterated m times. If $\eta \le 2/\sigma$, then for any $m \ge 1$, and C as in (9), we have

$$C\big(S^m(|x^k|)\, e^{j\angle z}\big) \le C\big(x^k\big). \tag{19}$$

Further, if equality holds in (19), then,

(i) x^k is a stationary point of $C(\cdot)$, i.e., 0 is in the proximal subdifferential [9] of $C(\cdot)$,
(ii) $x^k = S^m(|x^k|)\, e^{j\angle z}$, i.e., x^k is a fixed point of the iterations.

Proof. Applying S corresponds to one iteration of PGA on (14). Specifically, the correction term in (18) coincides with $\nabla D_{\beta,W}(x, |z|)$. It follows by the properties of P_+ that

$$\langle x - S(x), x - \eta \nabla D_{\beta,W}(x, |z|) - S(x) \rangle \le 0. \tag{20}$$

After rearranging this inequality, and invoking (11) with $L = \sigma$, one obtains that if $x \in \mathbb{R}_+^n$, then

$$\begin{aligned} D_{\beta,W}\big(x, |z|\big) - D_{\beta,W}\big(S(x), |z|\big) \\ \ge \Big(\frac{1}{\eta} - \frac{\sigma}{2}\Big)\, \big\|x - S(x)\big\|_2^2. \end{aligned} \tag{21}$$

The assumption $\eta\sigma < 2$, implies that the rhs is non-negative. Repeatedly invoking this inequality, we obtain $D_{\beta,W}\big(S^m(|x^k|), |z|\big) \le D_{\beta,W}\big(|x^k|, |z|\big)$. By (13), we then have $D_{\beta,W}(S^m(|x^k|)\, e^{j\angle z}; z) \le D_{\beta,W}(x^k; z)$, which implies (19).

Suppose now equality holds in (19). This implies that $D_{\beta,W}(S^m(|x^k|); |z|) = D_{\beta,W}(|x^k|; |z|)$. But by (21), this is possible only if $S(|x^k|) = |x^k|$. This in turn implies

$$|x^k| - |z| + \beta \mathbf{1} + W|x^k| \in -N_+(|x^k|), \tag{22}$$

where $N_+(|x^k|)$ is the normal cone [17] of $\mathbb{R}_+^n$ at $|x^k|$. (22) coincides with the optimality condition for $|x^k|$ for (14). It then follows from the train of inequalities in (13) that $x^* = S^m(|x^k|)\, e^{j\angle z}$ minimizes $D_{\beta,W}(\cdot; z)$. But $C(x^*) = C(x^k)$ implies $C^k(x^k) \le C^k(x^*)$, which is equivalent to $D_{\beta,W}(x^*; z) \le D_{\beta,W}(x^k; z)$. Therefore, x^k also minimizes $D_{\beta,W}(\cdot; z)$. Thus

$$0 \in x^k - z + \beta\, \partial P_W(x^k), \tag{23}$$

where $\partial P_W(x^k)$ is the proximal subdifferential of P_W [9]. Plugging in $z = x^k - \alpha\, \nabla f(x^k)$ and $\beta = \alpha\, \lambda$, we obtain,

$$0 \in \nabla f(x^k) + \lambda\, \partial P_W(x^k). \tag{24}$$

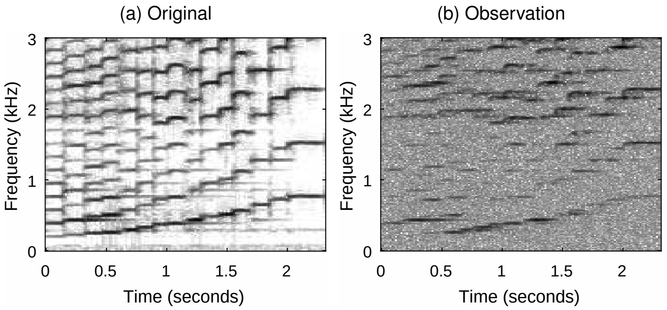

Fig. 2. Spectrograms of (a) the original signal, and (b) the reverberant and noisy observation (SNR = 5 dB) used in the experiment.

Thus, x^k is a stationary point of $C(\cdot)$, as claimed in (i).

Observe now that if x^k minimizes $D_{\beta,W}(\cdot;z)$, then we must have $e^{j\angle z} = e^{j\angle x^k}$. Therefore, $x^* = S(|x^k|)\,|\,e^{j\angle z} = |x^k|\,e^{j\angle x^k} = x^k$, as claimed in (ii). $\square$

In view of Prop. 4, Algorithm 1 achieves descent for (9).

Algorithm 1 A Descent Algorithm for (9) – See Prop. 4

```
Set $\alpha < 1/L$, $\beta \leftarrow \alpha\,\lambda$, $\eta < 2/\sigma\big(I + \beta\,W\big)$, initialize $x$
repeat
    $z \leftarrow x - \alpha \nabla f(x)$
    $x \leftarrow |x|$
    for $K$ iterations do
        $x \leftarrow P_+\Big(x - \eta\big[(\,I + \beta\,W)x + \beta\mathbf{1} - |z|\big]\Big)$
    end for
    $x \leftarrow x\,e^{j\,\angle z}$
until convergence
```

IV. Demonstration of the Proposed Penalty/Algorithm

We demonstrate the utility and behavior of the proposed penalty and the algorithm on a dereverberation experiment.

The clean signal consists of a violin playing a chromatic scale (see Fig. 2a). We observe a reverberant and noisy version of this signal. The (known) reverberation is represented in the short-time Fourier transform (STFT) domain by a linear operator H [24]. Denoting the STFT coefficients of the observations as y, we consider a reconstruction formulation for the STFT coefficients of the clean signal as

$$\min_x \underbrace{\frac{1}{2}\|y - H\,x\|_2^2}_{f(x)} + P(x). \tag{25}$$

Here $P(x)$ is the penalty term, which is either the ℓ_1 norm, the SWAG penalty, or the proposed penalty.

We select the weight of the ℓ_1 norm with a sweep search so as to maximize the output SNR. For both SWAG and the proposed penalty, we set λ to be a quarter of the weight used for the ℓ_1 penalty (which is sub-optimal). For these penalties, we form groups over each slice along the frequency axis in the STFT domain. For SWAG, we partition this slice with 960 coefficients into groups of size 15 and set γ to 40. For the proposed penalty, we set the weight matrix W to be a symmetric Toeplitz matrix of size 960×960, where the first column is as shown in Fig. 3d. The sequence is non-negative and sums to 900, so that $\sigma(W) \leq 900$. A Toeplitz W with non-zeros close to the main diagonal allows to obtain a localized effect in along the frequency axis.

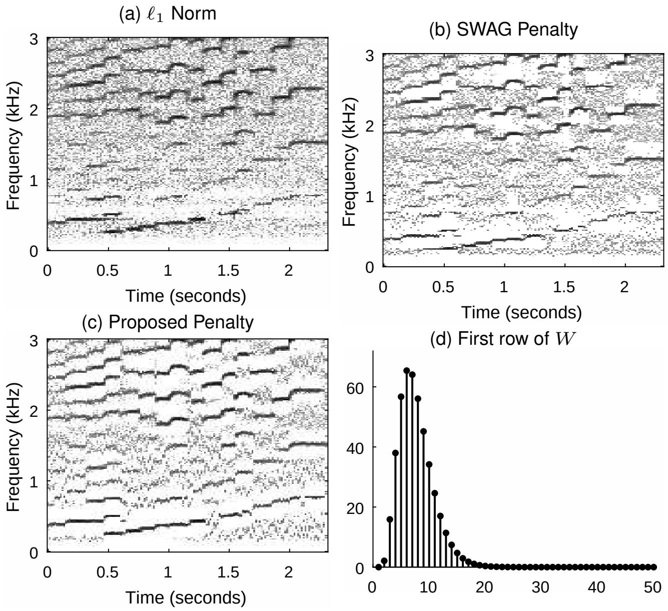

Fig. 3. Spectrograms of the dereverbed signals using (a) the ℓ_1 norm (7.75 dB), (b) SWAG penalty from [4] (5.76 dB), (c) proposed penalty (7.80 dB). (d) The first 50 coefficients from the first row of the Toeplitz weight matrix W, used for defining $P_W(\cdot)$.

The reconstructions obtained with these regularizers are shown in Fig. 3. We remark that our primary purpose is not to compare output SNRs, but to demonstrate the different characteristics of the reconstructions.

For the ℓ_1 norm, a higher threshold leads to the suppression of the weaker harmonics, along with noise. The proposed penalty and SWAG avoid this dilemma by suppressing noise around the strong harmonics, and preserving the weaker harmonics, thanks to a lower λ. Overall, this still leads to an improvement in terms of SNR for the proposed penalty. This aside, the spectrograms obtained with SWAG and the proposed penalty show some differences. SWAG uses non-overlapping groups. Also, since the boundaries of the groups are not selected with respect to the positions of the harmonics, we observe that the suppressed regions surrounding the harmonics are not centered around the harmonics. In contrast, thanks to the Toeplitz nature of W, the proposed penalty essentially employs maximally overlapping groups. This leads to a reconstruction where the harmonics lie at the center of an otherwise suppressed region.

V. Outlook

One aspect of interest, that is not addresed in this letter, is the selection of the weight matrix W. While the proposed penalty offers flexibility in the choice of the groups via the introduction of W, it is not obvious what the ‘optimal’ weights for a specific application should be. An alternative to an expert selection is to *learn* W from data. We hope to investigate this issue in future work.

References

[1] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

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

[3] İ. Bayram. On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty. *IEEE Transactions on Signal Processing*, 64(6):1597–1608, March 2016.

[4] İ. Bayram and S. Bulek. A penalty function promoting sparsity within and across groups. *IEEE Transactions on Signal Processing*, 65(16):4238 – 4251, June 2017.

[5] A. Berman. Complete positivity. *Linear Algebra and its Applications*, 107:57 – 63, 1988.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[7] P.-Y. Chen and I. W. Selesnick. Group-sparse signal denoising: Non-convex regularization, convex optimization. *IEEE Transactions on Signal Processing*, 62(13):3464–3478, July 2014.

[8] P.-Y. Chen and I. W. Selesnick. Translation-invariant shrinkage/thresholding of group sparse signals. *Signal Processing*, 94:476–489, January 2014.

[9] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski. *Nonsmooth Analysis and Control Theory*. Springer, 1998.

[10] P. L. Combettes and J.-C. Pesquet. A proximal decomposition method for solving convex variational inverse problems. *Inverse Problems*, 24(6):065014, 2008.

[11] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer, New York, 2011.

[12] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM Journal on Multiscale Modelling and Simulation*, 4(4):1168–1200, November 2005.

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

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

[15] A. A. Goldstein. Convex programming in Hilbert space. *Bull. Amer. Math. Soc.*, 70(5):709–710, 1964.

[16] L.J. Gray and D.G. Wilson. Nonnegative factorization of positive semidefinite nonnegative matrices. *Linear Algebra and its Applications*, 31:119 – 127, 1980.

[17] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2004.

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

[19] L. Jacob, G. Obozinsky, and J. P. Vert. Group lasso with overlap and graph lasso. In *Proc. Int. Conf. Machine Learning (ICML)*, 2009.

[20] M. Kowalski. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27(3):303–324, November 2009.

[21] 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, 2009.

[22] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, 1970.

[23] N. Rao, R. Nowak, C. Cox, and T. Rogers. Classification with the sparse group lasso. *IEEE Transactions on Signal Processing*, 64(2):448–463, January 2016.

[24] J. P. Reilly, M. Wilbur, M. Seibert, and N. Ahmadvand. The complex subband decomposition and its application to the decimation of large adaptive filtering problems. *IEEE Transactions on Signal Processing*, 50(11):2730–2743, Nov 2002.

[25] I. W. Selesnick and İ. Bayram. Sparse signal estimation by maximally sparse convex optimization. *IEEE Transactions on Signal Processing*, 62(5):1078–1092, March 2014.

[26] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[27] J.-P. Vial. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8:231–259, May 1983.

[28] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 68(1):49–67, 2006.

[29] Y. Zhou, R. Jin, and S. Hoi. Exclusive lasso for multi-task feature selection. In *Proc. Int. Conf. Artificial Intelligence and Statististics*, 2010.