HIERARCHICAL SPARSITY WITHIN AND ACROSS OVERLAPPING GROUPS

İlker Bayram

Analog Devices Inc., Analog Garage, Boston, MA, USA
ibayram@ieee.org

ABSTRACT

Recently, different penalties have been proposed for signals whose non-zero coefficients reside in a small number of groups, where within each group, only few of the coefficients are active. In this paper, we extend such a penalty, and introduce an additional layer of grouping on the coefficients. Specifically, we first partition the signal into groups, and then apply the penalty on the ℓ_2 norms of the groups. We discuss how this extended penalty can be used in energy minimization formulations, and demonstrate the effects of the proposed extension on a dereverberation experiment.

Index Terms— Sparsity within and across groups, elitist Lasso, group sparsity, mixed norm, overlapping groups.

1. INTRODUCTION

Group-based sparsity inducing penalties are usually used to encourage signals that can be well-approximated using a few groups of variables. However, there exist other types of sparsity that can be observed in natural signals. In this paper, we specifically consider signals that can be decomposed into groups such that (i) only a few groups are non-zero, (ii) in a non-zero group, only a few of the coefficients are non-zero. It can be argued that the frequency representation of a nearly periodic signal exhibits such a pattern [1].

The characteristic described above has been addressed in the recent literature. The elitist-lasso [2], or exclusive lasso [3] penalty sums the squares of the ℓ_1 norms of the groups to encourage such a characteristic. An alternative is the sparse-group Lasso [4, 5], where the penalty consists of the sum of an $\ell_{2,1}$ norm and an ℓ_1 norm. These penalty functions are convex. We proposed a non-convex penalty for the mentioned characteristic in [1], where we referred to it as 'sparsity within and across groups' (SWAG). We argued in [1] that by forgoing convexity, the SWAG penalty can be used to produce estimates with reduced bias.

In the SWAG penalty, groups are required to be non-overlapping, and within a group, the variables are treated symmetrically – that is, within a group, the penalty is invariant to permutations of the variables. In order to introduce some flexibility in forming the groups, we proposed a modification in [6], and defined a penalty on the vector x as,

$$\|x\|_1 + \sum_{n,m} w_{n,m} |x_n\, x_m|, \tag{1}$$

for a given collection of weights $w_{n,m} \geq 0$. For fixed n, we can think of the collection of x_m such that $w_{n,m} > 0$ as defining a group centered around x_n. Since the weights $w_{n,m}$ are allowed to be different, this group definition is more flexible compared to that in [1].

In this paper, we introduce an additional layer of grouping to this definition. Specifically, we first partition the signal into groups, and then apply the penalty on the ℓ_2 norms of the groups. As we will demonstrate in the experiment section, this allows to capture a more realistic model for the time-frequency coefficients of audio signals such as speech or music.

Proposed Modification

Suppose we are given a partition of the input vector x as $C = \{x^1, x^2, \ldots x^K\}$, where each x^k is a vector that contains some components of x. We refer to x^k as the k^{th} group. We assume that x^k's cover x, and they do not overlap. That is, for any n, we can find a unique k such that x_n is a component of x^k. In this setting, we propose the following penalty function

$$P_W(x) = \sum_n \|x^n\|_2 + \frac{1}{2} \sum_{n,m} w_{n,m}\, \|x^n\|_2\, \|x^m\|_2, \tag{2}$$

where $w_{n,m} \geq 0$. We will also assume throughout that $w_{n,n} = 0$ for all n, and $w_{n,m} = w_{m,n}$ for all n, m.

In this paper, we consider using this penalty function in a formulation of the form

$$\min_x\, f(x) + \lambda\, P_W(x). \tag{3}$$

Provided f is convex and differentiable with a Lipschitz continuous gradient, we will derive a descent algorithm for (3).

In order to derive the promised descent algorithm, we will start with the auxiliary problem

$$\min_x \{C(x, y) = \frac{1}{2}\, \|x - y\|_2^2 + \lambda P_W(x)\}, \tag{4}$$

where y is a given vector. The problem (4) may be regarded as a denoising formulation, where y constitutes a noisy observation. The mapping that takes y to the minimizer of $C(\cdot, y)$

is also referred to as the proximity operator of P_W, when $P_W(x)$ is convex [7, 8]. We continue to use the term 'proximity operator', due to formal resemblance.

We first show in Section 2 that the problem in (4) is convex with respect to x, provided λ satisfies an upper bound. Following that, we derive a descent algorithm to obtain a minimizer of this function in Section 3. We present a dereverberation experiment in Section 4. Section 5 contains remarks pointing to a future direction to pursue.

Notation

For a vector z, we write $e^{j\angle z}$ to denote the unit vector in the direction of z. For a vector z of the same size as x above, we assume that we have a partition of z similar to that of x, and write z^k to denote the k^{th} group of z. That is, x_n is in x^k if and only if z_n is in z^k.

For a complex number z, $\text{Re}(z)$ denotes its real part.

2. WELL-POSEDNESS OF THE DENOISING PROBLEM

$P_W(x)$, defined in (2), is a non-convex function of x. However, we show below that the cost function $C(\cdot, y)$ in (4) is strictly convex, provided λ is small enough. In that case, $C(\cdot, y)$ has a unique minimizer.

Observe now that

$$C(x,y) = \left[\frac{1}{2}\|y\|_2^2 - \text{Re}\langle x, y\rangle + \lambda \sum_n \|x^n\|_2\right] + \frac{1}{2}\left\{\sum_n \|x^n\|_2^2 + \lambda \sum_{n,m} w_{n,m}\, \|x^n\|_2\, \|x^m\|_2\right\}. \quad (5)$$

The term inside the square brackets is convex for any choice of λ. Therefore, if we can find a condition to ensure that the term inside the curly brackets is convex, we are done. Notice that we can rewrite that term as

$$\sum_n \|x^n\|_2^2 - \lambda \sum_{n,m} w_{n,m}\, \frac{1}{2}\big(\|x^n\|_2^2 + \|x^m\|_2^2\big) + \frac{\lambda}{2} \sum_{n,m} w_{n,m} \Big(\|x^n\|_2^2 + \|x^m\|_2^2 + 2\|x^n\|_2\, \|x^m\|_2\Big). \quad (6)$$

Rearranging, this is equal to

$$\sum_n \Big(1 - \lambda \sum_m w_{n,m}\Big) \|x^n\|_2^2 + \frac{\lambda}{2} \sum_{n,m} w_{n,m} \Big(\|x^n\|_2 + \|x^m\|_2\Big)^2. \quad (7)$$

The second term in this expression is convex. The first term is strictly convex if

$$\lambda < \Big(\sum_m w_{n,m}\Big)^{-1}, \text{ for all } n. \quad (8)$$

To summarize, we obtained the following result.

Proposition 1. Suppose $w_{m,n} = w_{n,m} \geq 0$ for all n, m, and $w_{n,n} = 0$ for all n. If (8) holds, then for fixed y, $C(\cdot, y)$ in (4) is strictly convex, and it has a unique minimizer.

This condition implies that $I + \lambda\, W$ can be decomposed as $D^T\, D$, where D has non-negative entries. In this case, $I + \lambda\, W$ is said to be completely positive [9], although this fact is not explicitly used for obtaining Prop 1. We refer to [6] for a related discussion.

3. A DESCENT ALGORITHM

In this section, we derive a descent algorithm for (3). But before we tackle this general problem, we consider (4), where f is replaced with a simple energy term.

3.1. Denoising with the Proposed Penalty

In order not to complicate the notation, let us start by defining a length-K vector u, such that $u_k = \|x^k\|_2$ for $k = 1, \ldots, K$. Using u, we can write,

$$P_W(x) = \|u\|_1 + \frac{1}{2}\, u^T\, W\, u. \quad (9)$$

The set $\{(u, x) : u_i = \|x^i\|_2, \text{ for all } i\}$ is not convex. However, the change of variables from u to x will be useful for deriving a simple update step.

Suppose we partition y similarly as x to obtain the non-overlapping groups y^k, and let $v_k = \|y^k\|_2$. We have the following lemma, which is key to our development.

Lemma 1. Let $\tilde{x}$ be the vector whose k^{th} group is defined as $\tilde{x}^k = \|x^k\|_2\, e^{j\angle y^k}$. We have,

$$C(x; y) \geq C(\tilde{x}; y) \quad (10)$$
$$= \Big\{D(u; v) = \frac{1}{2}\|u - v\|_2^2 + \lambda\, \|u\|_1 + \frac{\lambda}{2}\, u^T\, W\, u\Big\}.$$

Proof. We first note that

$$\text{Re}(\langle x^k, y^k\rangle) \leq \|x^k\|_2\, \|y^k\|_2 = \text{Re}(\langle \tilde{x}^k, y^k\rangle). \quad (11)$$

Therefore,

$$\|x - y\|_2^2 = \sum_k \|x^k - y^k\|_2^2 \quad (12)$$
$$= \sum_k \|x^k\|_2^2 + \|y^k\|_2^2 - 2\text{Re}(\langle x^k, y^k\rangle) \quad (13)$$
$$\geq \sum_k \|\tilde{x}^k\|_2^2 + \|y^k\|_2^2 - 2\text{Re}(\langle \tilde{x}^k, y^k\rangle) \quad (14)$$
$$= \|\tilde{x} - y\|_2^2 \quad (15)$$

Combining this observation with $P_W(x) = P_W(\tilde{x})$, the inequality $C(x; y) \geq C(\tilde{x}; y)$ follows.

Using (11), we have

$$\|\tilde{x}^k - y^k\|_2^2 = (\|\tilde{x}^k\|_2 - \|y^k\|_2)^2 = (u_k - v_k)^2. \quad (16)$$

Thus $\|\tilde{x} - y\|_2^2 = \|u - v\|_2^2$. Since $P_W(x) = P_W(\tilde{x})$, noting (9), we obtain the equality $C(\tilde{x}; y) = D(u; v)$. □

Observe that if all of the enries of u are non-negative, then for $x^k = u_k\, e^{j\angle y^k}$, we have $\|x^k\|_2 = u_k$, and $D(u;v) = C(x,y)$. This observation, along with Lemma 1 leads to the following corollary.

Corollary 1. For given u, v, x, y, defined as above, suppose $\hat{u} \geq 0$ satisfies

$$D(\hat{u};v) \leq D(u;v) - d, \tag{17}$$

for some $d \geq 0$. Let $\hat{x}$ be defined so that $\hat{x}^k = \hat{u}_k e^{j\angle y^k}$. Then,

$$C(\hat{x};y) \leq C(x;y) - d. \tag{18}$$

Further, if $\hat{u}$ minimizes $D(\cdot;v)$ over the positive orthant, then $\hat{x}$ minimizes $C(\cdot;y)$.

This corollary suggests that, instead of the problem in (4), we can consider the constrained problem

$$\min_{u\in\mathbb{R}^K_+} D(u;v) \tag{19}$$

For $u \in \mathbb{R}^K_+$, $D(u;v)$ is actually a quadratic function of u. Therefore, the minimization problem in (19) is equivalent to,

$$\min_{u\in\mathbb{R}^K_+} \frac{1}{2} u^T\,(I + \lambda\, W)\,u + u^T\,(\lambda\mathbf{1} - v), \tag{20}$$

where $\mathbf{1}$ denotes a vector of all ones. Notice that, if $(I + \lambda\, W)$ positive definite, this problem is strictly convex, and it has a unique minimizer. This is a less stringent condition than (8), which ensures strict convexity of $C(x,y)$. In the following, we will assume that $I + \lambda W$ is positive definite, and (20) is a convex problem.

The cost function in (20) is differentiable with a Lipschitz continuous gradient. Further, the constraint set, namely $\mathbb{R}^K_+$, has a simple projection operator. These two features make the projected gradient algorithm (PGA) [10] a feasible choice to obtain a descent algorithm for (20). At each iteration, PGA consists of iterations of the form

$$u \leftarrow P_+\Big(u - \eta\,\big[(I + \lambda\, W)u + (\lambda\mathbf{1} - v)\big]\Big), \tag{21}$$

where η is a step-size. Suppose now that σ is the spectral norm of $I + \lambda W$. It can be shown similarly as in [6] (see the proof of Prop.4), that if $\eta < 2/\sigma$, then the iterations in (21) achieve descent on (20). This, along with the observation in Cor. 1 allows us to obtain a descent algorithm for $C(\cdot, y)$, as described in the following. Another option, pointed out by one of our reviewers would be to employ an active set method, which is guaranteed to terminate in a finite number of steps for this problem (see e.g. [11], Sec. 16.4).

3.2. An Algorithm for (3)

We now consider the main problem (3), where the cost function is of the form

$$Q(x) = f(x) + \lambda\, P_W(x). \tag{22}$$

We assume that $f : \mathbb{C}^n \to \mathbb{R}^n$ is a convex function. Further, interpreting $\mathbb{C}^n$ as $\mathbb{R}^{2n}$, we assume that f is Fréchet-differentiable, and that the Fréchet derivative is Lipschitz continuous with parameter L. That is,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\,\|x - y\|_2, \text{ for all } x, y. \tag{23}$$

Suppose that we set up an iterative algorithm for minimizing $Q(\cdot)$. Also, let $\tilde{x}$ be our current iterate. Consider the following function, defined using $\tilde{x}$:

$$\hat{Q}(x;\tilde{x}) = \frac{1}{2\alpha}\left\|x - \big(\tilde{x} - \alpha\nabla f(\tilde{x})\big)\right\|_2^2 + \lambda\, P_W(x). \tag{24}$$

In this setting, following the majorization-minimization framework [12], it can be shown (see for instance [6], Prop.2) that, for $\alpha < 1/L$, the inequality $\hat{Q}(x;\tilde{x}) \leq \hat{Q}(\tilde{x};\tilde{x})$ implies $Q(x) \leq Q(\tilde{x})$. In words, it is sufficient to perform descent on $\hat{Q}$, in order to achieve descent on Q. But upto a factor of α, $\hat{Q}$ is in the form of a denoising formulation studied in Section 3.1, and we already derived how to achieve descent on denoising formulations. Thus, we can perform descent for (3). The resulting algorithm is summarized in Algorithm 1, for convenience.

Algorithm 1 A Descent Algorithm for (3)

```
Require: $L$ : Lipschitz const. of $\nabla f$; $T$ : number of inner
    iterations; $W$ : weight matrix used in $P_W$; $\lambda$ : weight of
    the penalty function
Set $\alpha < 1/L$, $\beta \leftarrow \alpha\,\lambda$, $\eta < 2/\sigma\big(I + \beta\, W\big)$, initialize $x$
repeat
    $z \leftarrow x - \alpha\nabla f(x)$
    $u_k \leftarrow \|x^k\|_2$ for $k = 1, \ldots, K$
    $v_k \leftarrow \|z^k\|_2$ for $k = 1, \ldots, K$
    for $T$ iterations do
        $u \leftarrow P_+\Big(u - \eta\,\big[(I + \beta\, W)u + (\beta\mathbf{1} - v)\big]\Big)$,
    end for
    $x^k \leftarrow u_k\, e^{j\,\angle z^k}$ for $k = 1, \ldots, K$
until convergence
```

4. NUMERICAL EXPERIMENT

To demonstrate the improvement obtained by the proposed additional layer on the penalty function, we revisit the experiment in [6]. This also gives us a chance to compare the algorithm with other penalties such as the ℓ_1 norm or the SWAG penalty from [1]. The signal of interest consists of a violin playing a chromatic scale. The observation y consists of a noisy and reverberant version of this signal. Given the room impulse response, the effect of reverberation is modelled in the STFT domain as a linear operator H [13]. The observed signal is thus obtained by applying H to the STFT coefficients and adding circular (complex-valued) white Gaussian noise, so that the SNR is 5 dB. The spectrograms of the clean and reverberant signals are shown in Fig. 1.

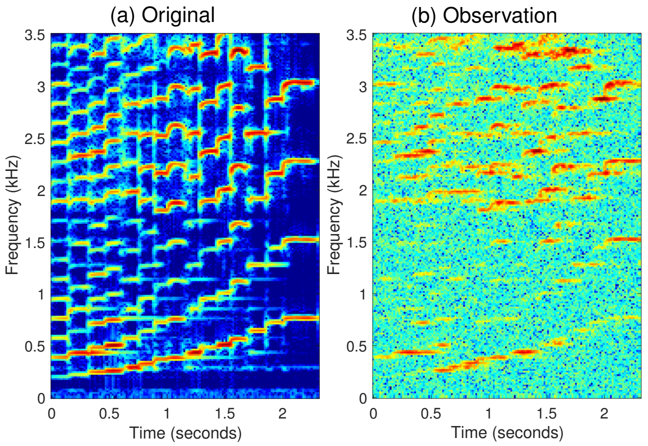

Fig. 1. (a) Original and (b) noisy reverberant observation signals in the STFT domain, used in the experiment in Section 4.

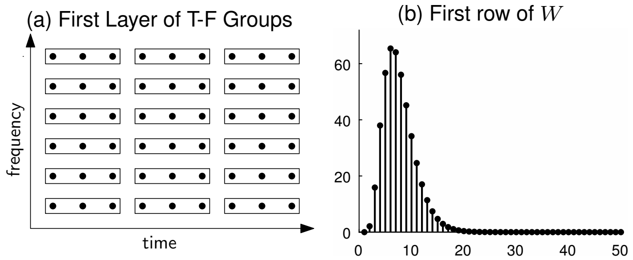

Fig. 2. Groups are formed in the time-frequency plane as shown in (a), to define the proposed penalty. Given the grouping, the sequence in (b) is used to define a Toeplitz weight matrix W.

We consider the minimization formulation

$$\min_x \frac{1}{2}\|y - Hx\|_2^2 + P(x), \tag{25}$$

where P is the penalty function.

In order to define the penalty from [6], we take a slice of the time-frequency STFT coefficients parallel to the frequency axis. For each time instance, there are 960 frequency coefficients. Regarding such a slice as a 1D signal, we employ a Toeplitz W to define the penalty on the frequency slice. The first row of this Toeplitz matrix is shown in Fig. 2b. We remark that this is the same penalty function used in the experiment in [6].

For the penalty function proposed in this paper, we first form groups over the time-frequency plane as shown in Fig. 2a. Specifically, we combine M time-adjacent coefficients to form a group. Replacing each group with its ℓ_2 norm, we obtain a new time-frequency signal. Then we apply the penalty described in the previous paragraph to this new signal to obtain the proposed penalty function.

Given these two penalty functions and the observation, we need to set λ. We also replace W by γW for $\gamma > 0$, in order to introduce a tuning parameter. For the penalty from [6] we perform a sweep search over λ and γ in order to obtain the highest possible performance achieved with that penalty. The resulting best reconstruction achieves an SNR of 8.17 dB, and is shown in Fig. 3a.

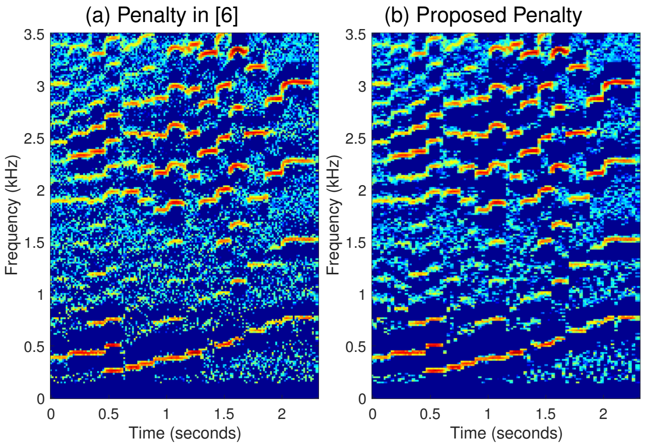

Fig. 3. Reconstructed signals using the formulation with (a) the penalty in [6] (SNR = 8.17 dB), (b) the proposed penalty (SNR = 8.43 dB).

For the proposed penalty we use the same λ found for the penalty from [6], but we replace γ with $\gamma\sqrt{M}$. We set the neighborhood size M to 2. Notice that since we do not perform a sweep search over γ, these parameters are suboptimal (such a sweep search is not possible in practice anyway). For these parameters, the reconstructed signal is shown in Fig. 3b. The obtained SNR is 8.43 dB.

Aside from the improvement in SNR, the reconstructed signal with the proposed modification contains less noise, in between the harmonics. We think this is due to the regularizing effect of the employed grouping that is performed before applying P_W. Recall that for the current choice of W, P_W encourages high valued coefficients to be isolated. However, due to noise, deciding whether a time-frequecy coefficient has a high magnitude or not cannot be performed reliably. However, thanks to the harmonic structures in audio, high magnitude coefficients appear adjacent along the time axis. Grouping makes use of this fact, and leads to a more reliable estimate of the magnitudes. This in turn enhances the effectivity of the penalty function P_W.

5. CONCLUSION

This paper extends a recently proposed group based penalty. Specifically, we first partition the input vector into groups and then apply the penalty to the ℓ_2 norms of the groups. The overall penalty then has two layers of groups. The first layer partitions the signal into non-overlapping groups. The second layer works with the ℓ_2 norms of the groups from the first layer, and allows the groups to overlap. A natural extension of the current work is to allow the first layer to contain non-overlapping groups as well. We hope to pursue this in future work.

6. REFERENCES

[1] İ. Bayram and S. Bulek, "A penalty function promoting sparsity within and across groups," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4238–4251, Aug 2017.

[2] M. Kowalski, "Sparse regression using mixed norms," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 303–324, Nov. 2009.

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

[4] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.

[5] N. Rao, R. Nowak, C. Cox, and T. Rogers, "Classification with the sparse group lasso," *IEEE Transactions on Signal Processing*, vol. 64, no. 2, pp. 448–463, Jan 2016.

[6] İ. Bayram, "Sparsity within and across overlapping groups," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 288–292, Feb 2018.

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

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

[9] A. Berman, "Complete positivity," *Linear Algebra and its Applications*, vol. 107, pp. 57 – 63, 1988.

[10] A. A. Goldstein, "Convex programming in Hilbert space," *Bull. Amer. Math. Soc.*, vol. 70, no. 5, pp. 709–710, 1964.

[11] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, 2$^{\text{nd}}$ edition, 2006.

[12] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statist.*, vol. 58, no. 1, pp. 30–37, Feb. 2004.

[13] 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*, vol. 50, no. 11, pp. 2730–2743, Nov 2002.