Accelerating ISTA with an active set strategy

Matthieu Kowalski, Pierre Weiss, Alexandre Gramfort, Sandrine Anthoine

Accelerating ISTA with an active set strategy

Matthieu Kowalski
Laboratoire des Signaux et Systèmes
Univ Paris-Sud
kowalski@lss.supelec.fr

Pierre Weiss
Institut de Mathématiques de Toulouse
INSA Toulouse
pierre.armand.weiss@gmail.com

Alexandre Gramfort
Martinos Center for Biomedical Imaging
Harvard Medical School
gramfort@nmr.mgh.harvard.edu

Sandrine Anthoine
LATP
Université Aix-Marseille I - CNRS
anthoine@cmi.univ-mrs.fr

Abstract

Starting from a practical implementation of Roth and Fisher's algorithm to solve a Lasso-type problem, we propose and study the Active Set Iterative Shrinkage/Thresholding Algorithm (AS-ISTA). The convergence is proven by observing that the algorithm can be seen as a particular case of a coordinate gradient descent algorithm with a *Gauss-Southwell-r* rule. We provide experimental evidence that the proposed method can outperform FISTA and significantly speed-up the resolution of very undetermined inverse problems when using sparse convex priors. The proposed algorithm makes brain mapping with magneto- and electroencephalography (M/EEG) significantly faster when promoting spatially sparse and temporally smooth solutions using an ℓ_1/ℓ_2 norm.

1 Introduction

The sparsity principle is now commonly used in various tasks in machine learning and signal processing. Its popularity dates back to the introduction of the Basis Pursuit by Chen, Donoho and Saunders [1] and the Lasso by Tibshirani [2] in the mid 90's. The key idea is to use the ℓ_1 norm as a regularizer, which remarkably promotes sparse solutions. Formally, the Lasso is a regression problem and it can be summarized by the following convex optimization problem:

$$\min_{x\in\mathbb{R}^N} \frac{1}{2}\|y-\Phi x\|_2^2 + \lambda\|x\|_1 \,, \tag{1}$$

where $y \in \mathbb{R}^M$, $\Phi \in \mathbb{R}^{M\times N}$, $x \in \mathbb{R}^N$ and $\lambda \in \mathbb{R}_+$. For the sake of simplicity, we stick here to the Lasso problem (1), but all the following can be extended to a more general formulation:

$$\min_{x\in\mathbb{R}^N} \mathcal{L}(y,x) + \lambda\mathcal{R}(x) \,, \tag{2}$$

where:

- $x \mapsto \mathcal{L}(y,x)$ is convex, differentiable, with Lipschitz gradient.
- $\mathcal{R}$ is a convex, separable function whose marginals are non differentiable at 0.

The functional in (1) is convex, but not differentiable. Various optimization algorithms were proposed in the literature to tackle the non-differentiability. We refer to [3] for a recent review. We focus here on two strategies: active set algorithms and proximal descent algorithms.

1.1 Contributions and organization of the paper

Motivated by a particular magneto- and electroencephalography (M/EEG) inverse problems that aims at localizing in space and in time active brain regions given some measures of the electromagnetic fields produced by neurons, we experimented with Roth's algorithm [4] in order to speed-up FISTA. It proved to be efficient in practice, but several theoretical questions remain open: the algorithm's convergence is not proved and no method is proposed in order to solve the sub-problems. In this work, we propose to mix the active-set strategy and the proximal descent and prove convergence.

Our main contribution is in Section 2 where a practical algorithm is presented, and its convergence proved. Finally, numerical experiments on a realistic M/EEG problem are presented in Section 3.

As already mentioned, the presentation is restricted to the Lasso for the sake of simplicity and clarity, but the results can be extended to more general sparsity promoting regularizers as illustrated in the experiments which use a combination of ℓ_1/ℓ_2 norms.

1.2 Notations

As stated, the optimization problem we address reads:

$$\hat{x} = \operatorname*{arg\,min}_{x \in \mathbb{R}^N} \frac{1}{2}\|y - \Phi x\|_2^2 + \lambda \|x\|_1 = \operatorname*{arg\,min}_{x \in \mathbb{R}^N} \mathcal{F}(x) \ ,$$

where $y \in \mathbb{R}^M$, $\Phi \in \mathbb{R}^{M \times N}$ with $M \leq N$, and $x \in \mathbb{R}^N$. We denote by $\{\varphi_i\}_{i \in \{1,\dots,N\}} \in \mathbb{R}^M$ the column vectors of matrix Φ. Let us define:

- $\mathcal{A} \subseteq \{1, \dots, n\}$ an *active set*, and $\Phi_{\mathcal{A}}$ the submatrix of Φ constructed from the columns of indices $\mathcal{A}$.
- $|\mathcal{A}| = \text{card}(\mathcal{A})$ the cardinality of the set $\mathcal{A}$.
- $\mathcal{A}^c$ denotes the complementary of the set $\mathcal{A}$ in $\{1, \dots, N\}$.
- $x_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ the sub-vector of x constructed with the coordinates of indices $\mathcal{A}$.
- $\mathcal{F}_{\mathcal{A}}$ the functional restricted to the active set $\mathcal{A}$: $\mathcal{F}_{\mathcal{A}}(x) = \frac{1}{2}\|y - \Phi_{\mathcal{A}} x_{\mathcal{A}}\|_2^2 + \lambda \|x_{\mathcal{A}}\|_1$.
- $\mathcal{T}_\lambda(x) = \text{sgn}(x) \max(|x| - \lambda, 0)$, the so-called soft-thresholding operator[1].

Moreover, for the sake of simplicity and without loss of generality, we suppose that $\|\Phi\| = 1$.

2 AS-ISTA

2.1 Roth and Fisher's algorithm

In [4] , Roth and Fisher presented an active set strategy for the Group-Lasso. We present in Algorithm 1 a version adapted to the Lasso problem. One can observe directly that, by definition of x and $x_{\mathcal{A}}$ we have at each iteration $x_{\mathcal{A}^c} = \mathbf{0}_{\mathbb{R}^{|\mathcal{A}^c|}}$. The proof of convergence of this algorithm, surprisingly not given in [4], relies on the fact that at each iteration, the value of $\mathcal{F}$ decreases strictly if the minimum is not reached.

Lemma 1. *Let $\{x^{(t)}\}$ the sequence generated by Algorithm 1. If $\mathcal{A}^{(t)}$ is not a feasible active set, then*

$$\mathcal{F}(x^{(t+1)}) \leq \mathcal{F}(x^{(t)}) \ .$$

Then, one can states

Theorem 1. *Roth's algorithm converges in a finite (less than 2^n) number of iterations.*

[1]The soft-thresolding operator is actually the proximity operator of the ℓ_1 norm. However, we choose to present our algorithm without the notion of proximity operator for the sake of simplicity.

Algorithm 1: Roth and Fisher's algorithm for the Lasso.

1. Initialization: $\mathcal{A}^{(0)} = \underset{k\in\{1,\dots,N\}}{\arg\max} \{|\langle \varphi_k, y\rangle|\}$, with $i \in \{1,\dots,N\}$, $x^{(0)} = \mathbf{0}_{\mathbb{R}^N}$.
2. Find
$$x^{(t+1)}_{\mathcal{A}^{(t)}} \in \underset{x\in\mathbb{R}^{|\mathcal{A}^{(t)}|}}{\arg\min} \frac{1}{2}\|y - \Phi_{\mathcal{A}^{(t)}} x_{\mathcal{A}^{(t)}}\| + \lambda \|x_{\mathcal{A}^{(t)}}\|_1 \,. \tag{3}$$
3. Compute the dual variable $z^{(t+1)} = y - \Phi x^{(t+1)}$. Check the optimality conditions: if $\forall k \in \mathrm{supp}(x^{(t)})$, $|\langle \varphi_k, z^{(t+1)}\rangle| = \lambda$ and $\forall k \in \mathcal{A}^{(t)}$ $|\langle \varphi_k z^{(t+1)}\rangle| \leq \lambda$, then STOP the algorithm (the optimum is reached). Else, go to 4.
4. Active set update: $\mathcal{A}^{(t+1)} = \mathrm{supp}(x^{(t+1)}) \cup \left\{ k \in \underset{k\in\{1,\dots,N\}}{\arg\max} |\langle \varphi_k, z^{(t+1)}\rangle| \right\}$
5. $t \leftarrow t+1$ and go to 2.

Proof. To prove the convergence of the algorithm, we just have to prove that we cannot have $\mathcal{F}(x^{(t+1)}) = \mathcal{F}(x^{(t)})$, except at the optimum. Let us do that by contradiction. Assume that both $x^{(t+1)}_{\mathcal{A}^{(t)}}$ and $x^{(t)}_{\mathcal{A}^{(t)}}$ are minimizers of $\mathcal{F}_{\mathcal{A}^{(t)}}$ and that $x^{(t)}_{\mathcal{A}^{(t)}}$ is not an optimum of $\mathcal{F}$. The KKT conditions corresponding to $x^{(t)}_{\mathcal{A}^{(t)}} \in \arg\min F_{\mathcal{A}^{(t)}}$ imply: $\forall k \in \mathcal{A}^{(t)}, |\langle \varphi_k, z^{(t)}\rangle| < \lambda \Rightarrow x^{(t)}_k = 0$.

Let $k_t = \arg\max |\langle \varphi_k, z^{(t)}\rangle|$. By definition, we have $k_t \in \mathcal{A}^{(t)}$ and $k_t \notin \mathcal{A}^{(t-1)}$. Moreover, as we are not at the optimum of $\mathcal{F}$, we have $|\langle \varphi_{k_t}, z^{(t)}\rangle| > \lambda$ which is in contradiction with the KKT condition. Finally, as we stand in finite dimension, the number of possible active set is finite and the conclusion follows. ∎

This algorithm is generic, in the sense that one can choose any algorithm to find the minimizer of Problem (3) at step 2. Here, we propose to study the algorithm when a proximal descent algorithm is used to solve Problem (3). However, in this case, only a approximate solution is found at step 2. We call our algorithm AS-ISTA (for Active Set Iterative Shrinkage/Thresholding algorithm). Theorem 1 cannot be applied to prove convergence as the minimizer of (3) is not computed exactly at step 2. In the following, we prove convergence of the algorithm in the case where the minimizer of (3) is computed inexactly using ISTA-like algorithms step and when several coordinates are added at each iteration (and many can be removed).

2.2 AS-ISTA algorithm

Algorithm 2, summarizes our AS-ISTA which consists in a modified Roth and Fisher's algorithm where several steps of proximal descent are done at step 2, and where a subset of the coordinates which violate the constraint $\|\Phi^* z\|_\infty \leq \lambda$ are added to the active set simultaneously.

Remark 1. *Step 2 in Algorithm 2 consists in $J^{(t)}$ iterations of ISTA. One can use a faster version, FISTA, in these inner iterations instead. Moreover, during the update of the active set, $\{ \underset{k\in\{1,\dots,N\}}{\arg\max} |\langle \varphi_k, z\rangle|\}$ may already be a subset of $\mathcal{A}^{(t)}$, and $a^{(t)}$ may be the empty set.*

Actually, Algorithm 2 can be seen as a coordinate gradient descent method. This class of methods has been widely studied by Paul Tseng, in particular in [5]. This observation allows us to state the following theorem, which is a consequence of Tseng's results.

Theorem 2. *Let $\{x^{(t)}\}$ be the sequence generated by Algorithm 2. Then, denoting by X^* the set of the minimizers of $\mathcal{F}$, $\lim_{t\to\infty}$ dist$(x^{(t)}, X^*) = 0$.*

Proof. The differences between Algorithm 2 and a coordinate gradient descent are: i) the presence of the inner loop ISTA, where several iterations can be made and ii) the choice of the active set.

Algorithm 2: AS-ISTA

1. Initialization: $\mathcal{A}^{(0)} \subseteq \{k \text{ such that } |\langle \varphi_k, y \rangle| > \lambda\}$, $x^{(0)} = \mathbf{0}_{\mathbb{R}^N}$.
2. Let $\tilde{x}^{(0)}_{\mathcal{A}^{(t)}} = x^{(t)}_{\mathcal{A}^{(t)}}$. Iterate $J^{(t)}$ times
$$\tilde{x}^{(j+1)}_{\mathcal{A}^{(t)}} = \mathcal{T}_\lambda \left(\tilde{x}^{(j)}_{(\mathcal{A}^{(t)}} + \Phi^*_{\mathcal{A}^{(t)}} \left(y - \Phi_{\mathcal{A}^{(t)}} \tilde{x}^{(j)}_{\mathcal{A}^{(t)}} \right) \right)$$
and define $x^{(t+1)}$ by $x^{(t+1)}_{\mathcal{A}^{(t)}} = \tilde{x}^{J^{(t)}}_{\mathcal{A}^{(t)}}$.
3. Compute the dual variable $z^{(t+1)} = y - \Phi x^{(t+1)}$.
4. Let $a^{(t+1)} \subseteq \{k, \text{ such that } |\langle \varphi_k, z^{(t+1)} \rangle| > \lambda\}$. Update the active set:
$$\mathcal{A}^{(t+1)} = \text{supp}(x^{(t+1)}) \cup \left\{ \underset{k \in \{1,\dots,N\}}{\arg\max} \ |\langle \varphi_k, z^{(t+1)} \rangle| \right\} \cup a^{(t+1)}.$$
5. $t \leftarrow t+1$ and go to 2.

First, one can check that we have

$$x^{(t+1)} = \underset{x \in \mathbb{R}^N}{\arg\min} \left\{ \frac{1}{2} \|x - x^{(t)} - \Phi^*(y - \Phi x^{(t)})\|_2^2 + \lambda \|x\|_1 \ s.t. \ x_k = x^{(t)}_k, k \notin \mathcal{A}^{(t)} \right\}$$

which corresponds to an update of the coordinate gradient descent with a constant step of size 1.

One can check that $\mathcal{A}^{(t)}$ is a particular case of the *Gauss-Southwell-r* rules [5]. To prove that, one must show there exists $\nu \in (0,1]$ such that: $\|d^{(t)}\|_\infty \leq \nu \|d^{(t)}_{\mathcal{A}^{(t)}}\|_\infty$ where $d^{(t)} = \mathcal{T}_\lambda(x^{(t)} + \Phi^*(y - \Phi x^{(t)})) - x^{(t)}$ and $d^{(t)}_{\mathcal{A}^{(t)}} = \mathcal{T}_\lambda(x^{(t)} + \Phi^*(y - \Phi_{\mathcal{A}^{(t)}} x^{(t)}_{\mathcal{A}^{(t)}})) - x^{(t)}$.

We have $\|d^{(t)}_{\mathcal{A}^{(t)}}\|_\infty = \max\limits_{i \in \mathcal{A}^{(t)}} |d^{(t)}_{\mathcal{A}^{(t)}}[i]|$. where $d[i]$ stands for the i^{th} coordinate of d. Moreover, we have $\|d^{(t)}\|_\infty = \max\left(\max_{i \in \text{supp}(x^{(t)})} |d(t)[i]|, \max_{i \in \text{supp}(x^{(t)})^c} |d(t)[i]| \right)$.

Let $i^{(t)} = \max\limits_{i \in \text{supp}(x^{(t)})^c} (|\langle \varphi_i, y - \Phi x^{(t)} \rangle|)$. We have $\forall i \in \text{supp}(x^{(t)})^c, d^{(t)}[i] = \mathcal{T}_\lambda(\langle \varphi_i, y - \Phi x^{(t)} \rangle)$, then $\underset{i \in \text{supp}(x^{(t)})^c}{\arg\max} |d^{(t)}[i]| = i^{(t)}$, so that

$$\|d^{(t)}\|_\infty = \max \left(\max_{i \in \text{supp}(x^{(t)})} |d^{(t)}[i]|, |d^{(t)}[i^{(t)}]| \right) .$$

Then, we have $\|d^{(t)}\|_\infty = \|d^{(t)}_{\mathcal{A}^{(t)}}\|_\infty$ in the following cases:

- if $i^{(t)} = \underset{i \in \{1,\dots,N\}}{\arg\max} (|\langle \varphi_i, y - \Phi x^{(t)} \rangle|)$, by construction of $\mathcal{A}^{(t)}$;
- **else** if $|d^{(t)}[i^{(t)}]| > \lambda$, then $|d^{(t)}[i^{(t)}]| \leq \max\limits_{i \in \mathcal{A}^{(t)}} (|\langle \varphi_i, y - \Phi x^{(t)} \rangle|)$;
- **else** if $|d^{(t)}[i^{(t)}]| \leq \lambda$ and $\max\limits_{i \in \mathcal{A}^{(t)}} (|\langle \varphi_i, y - \Phi x^{(t)} \rangle|) > \lambda$.

The only remaining case is when $|d^{(t)}[i^{(t)}]| \leq \lambda$ and $\max\limits_{i \in \mathcal{A}^{(t)}} |\langle \varphi_i, y - \Phi x^{(t)} \rangle| \leq \lambda$, which implies that $\mathcal{A}^{(t)}$ is feasible. Then, the algorithm reduces to the classical ISTA. So the AS-ISTA algorithm iterates until one of the previous case is checked, and the Gauss-Southwell-r rules is verified, or until convergence.

Finally, applying the results of convergence of the Gauss-Southwell-r rule and ISTA, we have that $\lim\limits_{t \to \infty} \mathcal{F}(x^{(t)}) = \min\limits_{x \in \mathbb{R}^N} \mathcal{F}(x)$. Then, applying the theorem in [6, p. 135] the conclusion follows.

■

The AS-ISTA algorithm has mainly two parameters: 1) The number $J^{(t)}$ of inner iterations of ISTA. This number can be fixed ($J^{(t)} \geq 1$), or be chosen according to a convergence criterion as the duality gap for solving Problem 3. 2) The number of coordinates added in the active set $\mathcal{A}^{(t)}$. The only constraint is that the active set must contain $\underset{k\in\{1,\dots,N\}}{\arg\max} |\langle \varphi_k, z^{(t)} \rangle|$.

3 Experiments on M/EEG data

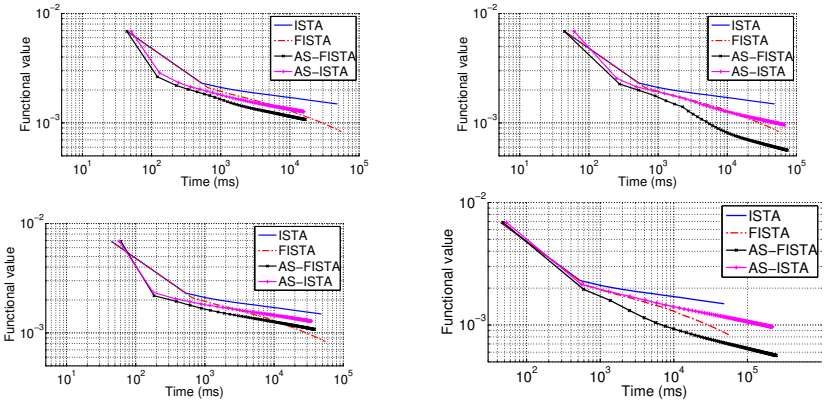

Figure 1: λ small. *Top left*: 30 inner iterations, $|\mathcal{A}^{(t)}| = 30 + |\text{supp}\{x^{(t)}\}|$. *Top right*: 300 inner iterations, $|\mathcal{A}^{(t)}| = 30 + |\text{supp}\{x^{(t)}\}|$. *Bottom left*: 30 inner iterations, $|\mathcal{A}^{(t)}| = 300 + |\text{supp}\{x^{(t)}\}|$. *Bottom right*: 300 inner iterations, $|\mathcal{A}^{(t)}| = 300 + |\text{supp}\{x^{(t)}\}|$.

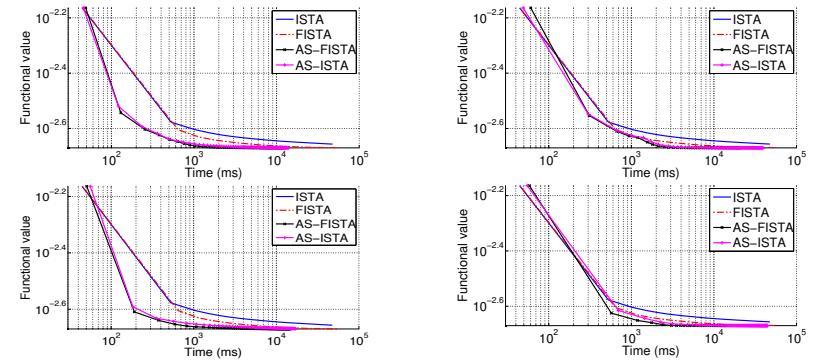

Figure 2: λ big. *Top left*: 30 inner iterations, $|\mathcal{A}^{(t)}| = 30 + |\text{supp}\{x^{(t)}\}|$. *Top right*: 300 inner iterations, $|\mathcal{A}^{(t)}| = 30 + |\text{supp}\{x^{(t)}\}|$. *Bottom left*: 30 inner iterations, $|\mathcal{A}^{(t)}| = 300 + |\text{supp}\{x^{(t)}\}|$. *Bottom right*: 300 inner iterations, $|\mathcal{A}^{(t)}| = 300 + |\text{supp}\{x^{(t)}\}|$.

In order to illustrate the convergence rate of the algorithms detailed above, simulations have been performed using a real MEG lead field matrix $\Phi \in \mathbb{R}^{151\times 5000}$. The matrix x is jointly-sparse along the columns This corresponds to the natural cognitive assumption that only a few brain regions are active during a cognitive task. Finally, a white Gaussian noise n is added on the data: $y = \Phi x + n$, in order to obtain a signal to noise ratio of 12 dB. The cost function under consideration is then a Group-Lasso problem where groups are formed by the rows of x. The lead field matrix is highly correlated, and then the problem can be seen as a "high-scale/high-correlated" problem [3].

We compare the four algorithms ISTA, FISTA, AS-ISTA and AS-FISTA, with various choices for the number of inner iterations and the update of the active set Two values of λ were chosen in order to compare the performance when the solution has different numbers of non-zero components. The results are provided in Figures 1 and 2 respectively. These figures show the evolution of the cost function as a function of the computation time in a log/log scale. We experimented with different

values for the number of inner iterations and with the size of the active set. For the update of the active set, there are two obvious strategies: 1) $\mathcal{A}^{(t)} = \{k, x_k^{(t)} \neq 0\} \cup \{\arg\max_k |\langle \varphi_k, z^{(t)} \rangle|\}$, which we call the "minimum" strategy. 2) $\mathcal{A}^{(t)} = \{k, x_k^{(t)} \neq 0\} \cup \{k, |\langle \varphi_k, z^{(t)} \rangle| > \lambda\}$, which we call the "full" strategy. These two strategies do not give satisfactory results. A lot of outer iterations must be made with the "minimum" strategy in order to add all the good coordinates, and a lot of inner iterations must be made with the "full" strategy in order to delete all the extra coordinates. We chose here to add in the active set the K coordinates which violate the most the constraint. It appears that the number of inner iterations should be proportional to the size of the active set, while the number of outer iterations is inversely-proportional to it.

Experiments show that AS-FISTA reaches the minimum faster than FISTA when the number of inner iterations is high enough. When the parameter λ is large, the solution is sparse, and AS-ISTA is as good as AS-FISTA (see Figure 2). Furthermore, AS-(F)ISTA appears to be efficient when the number of added variables at each step is not too small. This may be explained the same way as for the "minimum" and the "full" strategy. In the context of M/EEG, one expects tens to hundreds of brain sources to be active. Our experimental results show that source estimation can be made significantly faster if one adds more than one variable at each active set update and if we perform a few hundreds of inner iterations without requiring full convergence.

4 Conclusion

Depending on the point of view, the algorithm presented in this contribution can be seen as an active set algorithm, an iterative shrinkage/thresholding algorithm, or a coordinate gradient descent. On a realistic setup, we have shown that such a strategy can speed-up the state-of-the art FISTA algorithm when applied to the M/EEG inverse problem for functional brain imaging. To go beyond our contribution, we think that several open questions should be addressed:

1. We have shown that whatever the number of coordinates the algorithm converges. However both parameters influence its speed of convergence. Although we gave some insight on how these parameters should be chosen for M/EEG, a theoretical answer still needs to be discovered. As the efficiency of (F)ISTA depends on the correlation degree of the matrix Φ, efficient strategies should be defined based on it. Also, the bigger the regularization parameter λ, the smaller the number of inner iterations.

2. One of the great advantages of active set methods such as LARS is that the solutions are computed for all the regularization path. A good (in the sense of optimal) parameter can then be chosen *a posteriori*. However, for many signal processing problems where the good active set is very large, the LARS algorithm becomes inefficient. Signal processing would therefore certainly benefit from our algorithm. A practical problem might however be that for signal processing problems, applying Φ to a subset of coordinates may not be straightforward. Assuming Φ is a dictionary of time-frequency atoms, efficient multiplication on a limited number of atoms is likely to require more algorithmic work.

References

[1] S. Chen, D. David L. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Scient. Comp.*, 20, 1998.

[2] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Serie B*, 58(1):267–288, 1996.

[3] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. In S. J. Wright S. Sra, S. Nowozin, editor, *Optimization for Machine Learning*. MIT press, 2011.

[4] V. Roth and B. Fischer. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of ICML '08*, pages 848–855, 2008.

[5] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming B*, 117:387–423, 2009.

[6] B.T. Polyak. *Introduction to Optimization*. Translation Series in Mathematics and Engineering, Optimization Software, 1987.