

# A Fast and Scalable Polyatomic Frank-Wolfe Algorithm for the LASSO

Adrian Jarret, Julien Fageot, Matthieu Simeoni, *Member, IEEE*

**Abstract**—We propose a fast and scalable *polyatomic Frank-Wolfe (P-FW)* algorithm for the resolution of high-dimensional LASSO regression problems. This algorithm improves upon traditional Frank-Wolfe methods by considering generalized greedy steps with *polyatomic* (i.e. linear combinations of multiple atoms) update directions, hence allowing for a more efficient exploration of the search space. To preserve sparsity of the intermediate iterates, we re-optimize the LASSO problem over the set of selected atoms at each iteration. For efficiency reasons, the accuracy of this re-optimization step is relatively low for early iterations and gradually increases with the iteration count. We provide convergence guarantees for our algorithm and validate it in simulated compressed sensing setups. Our experiments reveal that P-FW outperforms state-of-the-art methods in terms of runtime, both for FW methods and optimal first-order proximal gradient methods such as the *Fast Iterative Soft-Thresholding Algorithm (FISTA)*.

**Index Terms**—Conditional Gradient, Frank-Wolfe, LASSO, Sparse Recovery, Convex Optimisation.

## I. INTRODUCTION

LINEAR regression is at the core of many inference tasks in signal processing, statistics or machine learning. With the advent of computing power and big data, modern regression problems tend to involve a large number of features, from a few thousands to hundreds of millions depending on the applications. Such large-scale regression problems are typically encountered in computational imaging setups, as for instance interferometric imaging in radio astronomy [1]–[4] and acoustics [5], environmental monitoring [6], or medical imaging [7]–[9]. In such high dimensional settings, the LASSO regression problem [10], [11] (also known as *Basis Pursuit Denoising* in the signal processing community [12], [13]) has received particular attention from the research community [14], [15]. Thanks to an  $\ell_1$  regularization term, the LASSO promotes sparse estimates (i.e. with relatively few active degrees-of-freedom) allowing for simpler and more interpretable models. The LASSO is most often used in its penalized form:

$$\arg \min_{\mathbf{x} \in \mathbb{R}^N} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1, \quad (1)$$

First submitted on December 4th, 2021, revised on January 31st, 2022.

This work was funded by the Swiss National Science Foundation (SNSF) under grants CRSII5\_193826 AstroSignals (A. Jarret and M. Simeoni), 200 021 181 978/1, “SESAM - Sensing and Sampling: Theory and Algorithms” (M. Simeoni) and P400P2\_194364 (J. Fageot).

A. Jarret and J. Fageot are from the School of Computer and Communication Sciences and M. Simeoni is from EPFL Center for Imaging, both at École Polytechnique Fédérale de Lausanne, Ecublens, Switzerland (email: name.surname@epfl.ch).

where  $\mathbf{y} \in \mathbb{R}^L$  is a vector of observed values to be fitted,  $\mathbf{A} \in \mathbb{R}^{L \times N}$  is the so-called *design matrix*,  $\mathbf{x} \in \mathbb{R}^N$  are the unknown regression coefficients and  $\lambda > 0$  is the penalty strength. In compressed sensing setups [12], the number of measurements  $L$  is typically much smaller than the dimension  $N$  of the problem. The solution set to the LASSO problem (1) can be shown to be non-empty, and the closed convex hull of (at most)  $L$ -sparse extreme points; see for example [16, Theorem 6], [17, Theorem 6.8] and [18]–[25] for generalizations. Furthermore, if the design matrix coefficients are drawn according to a continuous probability distribution, the LASSO solution is unique with probability one, and guaranteed to be at most  $L$ -sparse [11], [26].

Numerical solvers for addressing (1) can be classified into three main categories:

- *Greedy methods*. These methods identify at each iteration the variable *most correlated* with the residuals and mark it as active. The vector of regression coefficients  $\mathbf{x}$  is then updated in a direction computed from the set of variables activated so far [15]. Examples include the *Least-Angle Regression (LARS)* algorithm [27], [28] as well as specialised *Frank-Wolfe (FW) methods*<sup>1</sup> [15], [29]–[36].
- *Coordinate Descent (CD) methods*. These methods cyclically select *one component* of  $\mathbf{x}$  at a time and update it individually via a line search [37]. A very efficient CD method for the LASSO was proposed in [38].
- *Proximal Gradient Descent (PGD) methods*. These methods update *all components* of  $\mathbf{x}$  at each iteration via successive gradient and proximal steps [39]. Examples include the well-known *Fast Iterative Soft-Thresholding Algorithm (FISTA)* and its variants [40]–[42].

Thanks to their limited update scope, LARS, CD and FW scale relatively well with the number of features  $N$ . Indeed, the intermediate iterates they produce are often very sparse, hence alleviating the memory and computation bottlenecks of high dimensional regression. This is especially true for FW methods, which often yield significantly sparser solutions than CD or LARS [15]. These algorithms are however relatively slow to converge with high accuracy, due to suboptimal convergence rates of  $\mathcal{O}(1/k)$  [31], [43], [44]. This is in contrast with PGD methods which, although less scalable due to their dense intermediate iterates, can achieve first-order optimal convergence rates of  $\mathcal{O}(1/k^2)$  when used in conjunction with Nesterov’s acceleration schemes [40], [42], [45].

<sup>1</sup>Frank-Wolfe methods are also referred to as *Conditional Gradient (CG) methods* in some communities.

In this work, we accelerate the FW method to make it competitive with optimal first-order proximal gradient descent methods such as FISTA. To this end, we propose generalized greedy steps, which we refer to as *polyatomic* updates. Unlike standard greedy steps, the latter can activate multiple variables at the same time, allowing for a more efficient exploration of the search space. To ensure sparse intermediate iterates and further accelerate the algorithm, we also consider a partially-corrective step, which re-optimizes the LASSO problem over the set of previously activated atoms at each iteration. To minimize the computational cost of each iteration, this re-optimization step is performed with reduced, but gradually increasing, accuracy. The resulting algorithm is called *Polyatomic Frank-Wolfe (P-FW)*. We provide convergence guarantees for the latter, and compare it to state-of-the-art FW methods and FISTA in simulated compressed sensing setups. In all investigated setups P-FW converges the fastest, sometimes by a very significant margin (*e.g.*  $\sim 4$  times faster than FISTA,  $\sim 20$  times faster than FW methods).

The rest of this article is organized as follows. In Sections II and III, we review the classical Frank-Wolfe algorithm and some of its variants, first in a generic setting and then more specifically for the LASSO problem. In Section IV, we derive the P-FW algorithm and establish its convergence. In Section V, we test the performance of P-FW on simulated data.

## II. FRANK-WOLFE AND ITS VARIANTS

### A. Vanilla Frank-Wolfe

The Frank-Wolfe algorithm, as presented in 1956 [46], aims at minimizing a general constrained convex optimization problem of the form

$$\arg \min_{\mathbf{x} \in \mathcal{D}} f(\mathbf{x}) \quad (2)$$

where  $f$  is a convex and continuously differentiable cost function and the domain  $\mathcal{D}$  is a compact convex subset of a some Banach space. The vanilla version of the procedure is provided in Algorithm 1 below.

---

#### Algorithm 1: Vanilla Frank-Wolfe Algorithm (V-FW)

---

```

Initialize  $\mathbf{x}_0 \in \mathcal{D}$ 
for  $k = 1, 2 \dots$  do
1) Find an update direction:
    $\mathbf{s}_k \in \arg \min_{\mathbf{s} \in \mathcal{D}} \langle \nabla f(\mathbf{x}_k), \mathbf{s} \rangle$ 
2.a) Step size:  $\gamma_k \leftarrow \frac{2}{k+2}$ 
2.b) Reweight:  $\mathbf{x}_{k+1} \leftarrow (1 - \gamma_k)\mathbf{x}_k + \gamma_k \mathbf{s}_k$ 
end

```

---

The update direction search in step 1) consists in minimizing a linear function over a convex set, which leads to solutions lying on the set's boundary. When the domain  $\mathcal{D}$  is a polytope, the minimum is necessarily reached for one of its extreme points or *atoms*, hence allowing to restrict step 1) to the generating atoms of  $\mathcal{D}$  only. The computation of the update direction is very cheap since it is projection-free. This is in contrast with *Projected Gradient Descent* methods for (2), which could involve very expensive projections onto the set  $\mathcal{D}$  at each iteration.

### B. Known Frank-Wolfe variants

Several variants of the FW algorithm have been proposed in the literature. Adopting the same formalism as in [43], we review some of these variants that we will later rely on.

1) *Exact Line Search*: It is possible to replace step 2.a) by an exact line search, leading to Algorithm 2.

---

#### Algorithm 2: Exact Line Search

---

```

... Same as Algorithm 1, replacing steps 2.a) by:
2.a') Solve:  $\gamma_k \leftarrow \arg \min_{\gamma \in [0,1]} f((1-\gamma)\mathbf{x}_k + \gamma \mathbf{s}_k)$ 

```

---

This step is typically performed when a closed-form solution for step 2.a') exists, *e.g.* with quadratic objective functionals.

2) *Approximate Linear Subproblem*: In cases where the gradient minimization step 1) is expensive (*e.g.* complex domain shape), the latter can be performed approximately. This yields Algorithm 3, where  $\delta > 0$  controls the approximation quality:

---

#### Algorithm 3: Approximate Linear Subproblem of quality $\delta > 0$

---

```

... Same as Algorithm 1 or 2, replacing step 1) by:
1') Find:  $\mathbf{s}_k \in \mathcal{D}$  such that
    $\langle \nabla f(\mathbf{x}_k), \mathbf{s}_k \rangle \geq \min_{\mathbf{s} \in \mathcal{D}} \langle \nabla f(\mathbf{x}_k), \mathbf{s} \rangle + \gamma_k \delta$ 
   with  $\gamma_k = 2/(k+2)$ 

```

---

3) *Fully-Corrective FW*: Another commonly used variant is known as *Fully-Corrective FW (FC-FW)*, which re-optimizes  $f$  over the convex-hull of all previously selected atoms at each iteration. This results in Algorithm 4 below, where  $\text{Conv}$  denotes the convex hull of a given set.

---

#### Algorithm 4: Fully-Corrective Variant (FC-FW)

---

```

... Same as Algorithm 1 or 3, replacing steps 2.a)/2.b) by:
2') Solve:  $\mathbf{x}_{k+1} \leftarrow \arg \min_{\mathbf{x} \in \text{Conv}(\mathbf{s}_1, \dots, \mathbf{s}_k)} f(\mathbf{x})$ 

```

---

The re-optimization step 2') allows for more progress to be made at each iteration and can potentially discard wrongly selected atoms. This procedure however requires solving at each iteration an optimization problem that may turn out to be as difficult as the original one, and thus computationally expensive.

### C. Convergence guarantees

Theorem 1 of [43] shows that the sequence of iterates  $\mathbf{x}_k$  produced by FW or its variants achieves a convergence rate of  $\mathcal{O}(1/k)$  in terms of the objective functional. More specifically, let  $C_f > 0$  denote the *curvature constant*<sup>2</sup> of  $f$ , and  $f^*$  its optimal value, the iterates of Algorithms 1 to 4 satisfy

$$f(\mathbf{x}_k) - f^* \leq \frac{2}{k+2}(C_f + 2\delta), \quad \forall k \geq 1, \quad (3)$$

with  $\delta = 0$  for Algorithms 1, 2, and potentially 4 if the latter does not perform the optional approximate gradient minimization step.

<sup>2</sup>See [43] for a definition.

### III. FRANK-WOLFE FOR THE LASSO

While not of the form (2) due to its non-differentiable cost functional, the penalized LASSO regression problem (1) can still be brought in the scope of the FW algorithm by means of an *epigraphical lift*, as proposed in [29], [31]. To this end, we introduce the bounded norm cone  $C = \{(t, \mathbf{x}) \in \mathbb{R}_+ \times \mathbb{R}^N : \|\mathbf{x}\|_1 \leq t \leq M\}$  with  $M = \|\mathbf{y}\|_2^2 / 2\lambda$ . With similar arguments as in [31, Lemma 4], it is then possible to show that (1) is equivalent to the constrained optimization problem:

$$\arg \min_{(t, \mathbf{x}) \in C} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda t. \quad (4)$$

Observe that (4) is indeed of the form (2), since the objective function  $f(t, \mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda t$  is continuously differentiable and  $C \subset \mathbb{R}_+ \times \mathbb{R}^N$  is compact and convex. Applied to (4), the update direction step 1) in Algorithm 1 becomes

$$(t_k, \mathbf{s}_k) \in \arg \min_{(t, \mathbf{x}) \in C} \langle \boldsymbol{\eta}_k, \mathbf{x} \rangle + \lambda t \quad (5)$$

where  $\boldsymbol{\eta}_k = \frac{1}{\lambda} \mathbf{A}^T (\mathbf{y} - \mathbf{A}\mathbf{x}_k)$  is the so-called *empirical dual certificate* at iteration  $k$  [31]. Since the extreme points of  $C$  are easily shown to be of the form  $(M, \pm M \mathbf{e}_{i_k})$  (with  $\mathbf{e}_j$  the  $j$ -th canonical basis vector), solving (5) is equivalent to finding the index  $i_k$  such that

$$i_k = \arg \max_{i \in \{1, \dots, N\}} |(\boldsymbol{\eta}_k)_i|. \quad (6)$$

Note that the lift variable  $t$  is constant across iterations and is thus omitted in what follows. To sum up, the update direction step 1) amounts to the creation of a new atom  $\pm M \mathbf{e}_{i_k}$  that will be reweighted and added to the current iterate  $\mathbf{x}_k$ . The selected index  $i_k$  is associated to the column of  $\mathbf{A}$  which is most correlated with the residuals  $\mathbf{y} - \mathbf{A}\mathbf{x}$ . Note that the exact same update step appears in the greedy *Matching Pursuit (MP)* [47] and *Orthogonal Matching Pursuit (OMP)* [48], [49] algorithms, commonly used in signal processing for sparse recovery.

When considering the approximated subproblem variant, step 1') of Algorithm 3 becomes

$$\begin{aligned} &\text{Find } i_k \in \{1, \dots, N\}, \text{ such that} \\ &|(\boldsymbol{\eta}_k)_{i_k}| \geq \max_{i \in \{1, \dots, N\}} |(\boldsymbol{\eta}_k)_i| - \gamma_k \delta = \|\boldsymbol{\eta}_k\|_\infty - \gamma_k \delta. \end{aligned} \quad (7)$$

Moreover, the fully-corrective step 2') of Algorithm 4 becomes:

$$\begin{aligned} \tilde{\mathbf{x}}_{k+1} &\leftarrow \arg \min_{\tilde{\mathbf{x}} \in \mathbb{R}^{\text{Card}(\mathcal{S}_k)}} \|\mathbf{y} - \mathbf{A}_{\mathcal{S}_k} \tilde{\mathbf{x}}\|_2^2 + \lambda \|\tilde{\mathbf{x}}\|_1 \quad (8) \\ \mathbf{x}_{k+1}[\mathcal{S}_k] &\leftarrow \tilde{\mathbf{x}}_{k+1}, \\ \mathbf{x}_{k+1}[\overline{\mathcal{S}}_k] &\leftarrow \mathbf{0}, \end{aligned}$$

where  $\mathcal{S}_k := \{i_1, \dots, i_k\} \subset \{1, \dots, N\}$  denotes the set of the active indices up to iteration  $k$  and  $\mathbf{A}_{\mathcal{S}_k} \in \mathbb{R}^{L \times \text{Card}(\mathcal{S}_k)}$  denotes the restriction of  $\mathbf{A}$  to its columns with indices in  $\mathcal{S}_k$ . Note that in comparison with the initial problem (1), the dimension of (8) is dramatically reduced, from  $N$  to  $\text{Card}(\mathcal{S}_k)$  at iteration  $k$ .

### IV. POLYATOMIC FRANK-WOLFE

Observe that there might exist multiple indices that simultaneously satisfy the approximate gradient minimization step (7), each yielding different but equally valid atomic update directions. Unlike state-of-the-art approaches which limit themselves to selecting one of these directions arbitrarily, we propose taking *polyatomic* update directions (*i.e.* expanding the set of active indices with multiple indices at once). This is indeed possible since any convex combination of the potential atomic update directions is also a solution to (7) –although in general not on the boundary of the constraint convex set  $C$  anymore. The idea behind such polyatomic update directions is to accelerate the search space exploration, and consequently the convergence of the algorithm. Polyatomic update steps have already been used in other contexts (see for example [50] for OMP), but to the best of our knowledge, its combination with a FW-type algorithm has not yet been explored.

The main difficulty in working with polyatomic update directions consists in correctly estimating the weights associated to each atom in the convex combination. When a fully-corrective step is additionally performed, choosing the convex optimization weights optimally is no longer a necessity since (8) will anyway re-optimize these weights. We therefore propose to use suboptimal polyatomic update steps with arbitrary weights (*e.g.* the mean of all atoms) followed by fully-corrective steps. In practice, to further reduce the computational cost of each iteration, we also propose to prematurely stop the fully-corrective steps, resulting in *partially-corrective* steps. Thanks to an adaptive stopping criterion, these steps are designed to be more and more accurate as the number of iterations increases. Indeed, early partially-corrective steps need not be very accurate since the iterates evolve a lot from one iteration to the other.

Putting all these ingredients together results in Algorithm 5, which we call the *Polyatomic Frank-Wolfe (P-FW)* algorithm. The partial-correction procedure is described in Algorithm 6. It consists in a warm-started iterative soft-thresholding algorithm with a stopping criterion based on the relative improvement of the iterates.

Theorem 1 proves the convergence of P-FW, relying on [43, Theorem 1].

---

#### Algorithm 5: Polyatomic FW (P-FW) of quality $\delta > 0$

---

```

Initialize:  $\mathbf{x}_0 \leftarrow \mathbf{0}, \mathcal{S}_0 \leftarrow \emptyset$ 
for  $k = 1, 2 \dots$  do
   $\gamma_k \leftarrow 2/(k+2)$ 
  1a.a) Polyatomic exploration:
     $\mathcal{I}_k = \{1 \leq j \leq N : |(\boldsymbol{\eta}_k)_j| \geq \|\boldsymbol{\eta}_k\|_\infty - \delta \gamma_k\}$ 
     $\mathbf{s}_k \leftarrow (\sum_{i \in \mathcal{I}_k} \mathbf{e}_i) / \text{Card}(\mathcal{I}_k)$ 
  1a.b) Update active indices:  $\mathcal{S}_k \leftarrow \mathcal{S}_{k-1} \cup \mathcal{I}_k$ 
  2a.a) Set accuracy threshold:  $\varepsilon_k = \varepsilon_0 \gamma_k$ 
  2a.b) Update active weights:
     $\mathbf{x}_{k+1/2} \leftarrow (1 - \gamma_k) \mathbf{x}_k + \gamma_k \mathbf{s}_k$ 
     $\mathbf{x}_{k+1} \leftarrow \text{partial\_correction}(\mathbf{x}_{k+1/2}, \mathcal{S}_k, \varepsilon_k)$ 
     $k \leftarrow k + 1$ 
end

```

---

**Algorithm 6:** `partial_correction`( $\mathbf{x}, \mathcal{S}, \varepsilon$ )

---

Initialize:  $\tilde{\mathbf{u}}_0 \leftarrow \mathbf{x}[\mathcal{S}], k \leftarrow 1, \tau \leftarrow 1/\|\mathbf{A}_{\mathcal{S}}\|_2$   
**while**  $\|\tilde{\mathbf{u}}_k - \tilde{\mathbf{u}}_{k-1}\|_2 > \varepsilon \|\tilde{\mathbf{u}}_{k-1}\|_2$  **do**  
 $\tilde{\mathbf{v}}_{k+1} = \tilde{\mathbf{u}}_k - \tau \mathbf{A}_{\mathcal{S}}^T (\mathbf{A}_{\mathcal{S}} \tilde{\mathbf{u}}_k - \mathbf{y})$   
 $\tilde{\mathbf{u}}_{k+1} = \text{sgn}(\tilde{\mathbf{v}}_{k+1}) \max(\mathbf{0}, |\tilde{\mathbf{v}}_{k+1}| - \tau \lambda)$   
 $k \leftarrow k + 1$   
**end**  
Return:  $\mathbf{u} \in \mathbb{R}^N$  with  $\mathbf{u}[\mathcal{S}] = \tilde{\mathbf{u}}_k, \mathbf{u}[\bar{\mathcal{S}}] = \mathbf{0}$

---

**Theorem 1** (Convergence of Polyatomic FW). *Consider Algorithm 5 with parameter  $\delta > 0$ . Denote by  $\mathcal{L}(\mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1$  the LASSO cost function and  $\mathcal{L}^*$  its minimal value. For each  $k \geq 1$ , the iterates  $\mathbf{x}_k$  satisfy*

$$\mathcal{L}(\mathbf{x}_k) - \mathcal{L}^* \leq \frac{2}{k+2} (C_f + 2\delta)$$

where  $C_f$  is the curvature constant of (4).

*Proof.* First, we remark that Lemma 5 of [43] holds for  $\mathbf{s}_k$  being any convex combination of solutions of (7). Then we only need  $\mathcal{L}(\mathbf{x}_{k+1}) \leq \mathcal{L}(\mathbf{x}_{k+1/2})$  to apply the proof of [43]. This is necessarily the case since the partial correction solver ISTA is monotonic and initialized with  $\mathbf{x}_{k+1/2}$ .  $\square$

Note that Theorem 1 only provides an upper bound on the convergence rate of Algorithm 5. In practice, the algorithm tends to converge much faster, as illustrated in the next section.

## V. NUMERICAL SIMULATIONS

For our simulations, we consider the following data model:

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{w}, \quad (9)$$

where  $\mathbf{x}_0 \in \mathbb{R}^N$  is a  $k$ -sparse vector of dimension  $N = 16384$ ,  $\mathbf{A} \in \mathbb{R}^{L \times N}$  is a Gaussian random matrix and  $\mathbf{w}$  an additive Gaussian white noise with a PSNR of 20 dB. We set the number of measurements to  $L = \alpha K \ll N$  for some oversampling factor  $\alpha > 1$ . We then solve a LASSO problem (1) with input data  $\mathbf{y}$  with four algorithms<sup>3</sup>: *FISTA*, *Vanilla FW* (V-FW) (with optimal line search, Algorithm 2), *Fully-Corrective FW* (FC-FW) Algorithm 4, and our proposed *Polyatomic FW* (P-FW). We let each algorithm run for a maximum of 4 seconds and monitor the evolution of the LASSO objective functional value across reconstruction time. We perform this experiment for various values of  $K$  and  $\alpha$ , each time re-running the algorithms for 15 independent noise realizations. Fig. 1 shows the median LASSO objective functional value vs time, as well as the interquartile distance (shaded).

Although FC-FW is slightly faster than V-FW, these two algorithms are significantly slower to converge than P-FW or FISTA. Remarkably, P-FW is able to outperform the optimal first-order method FISTA in most setups. The most striking example comes from configuration (d) of Fig. 1, in which FISTA takes on average more than 2.0s to reach an objective functional value reached by P-FW in less than 0.5s ( $\sim 4\times$  speedup). Furthermore, we remark that the performances of V-FW and FC-FW degrade with larger values of  $K$  (ground truth  $\mathbf{x}_0$  less sparse), whereas this is not the case for P-FW.

<sup>3</sup>Our implementation is based on the `PyCso` optimization package [51].

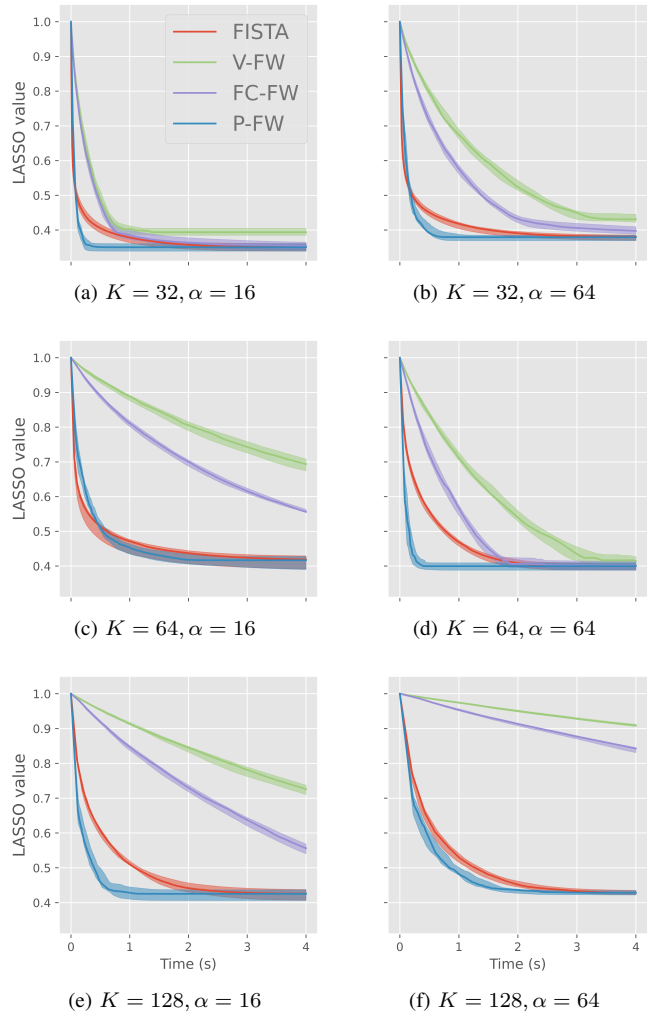


Fig. 1: Value of the LASSO objective functional across reconstruction time for the four algorithms benchmarked (FISTA, V-FW, FC-FW, P-FW) and various values of  $K$  and  $\alpha$ . The timings are for a ThinkPad T14, Intel Core i7 (4C/8T) @ 1.8GHz with 32GB RAM. These results can be reproduced using the Python scripts provided in our GitHub repository [52].

It is able to remain competitive with FISTA, even in the less sparse setups ( $K = 128$ ) of plots (e) and (f).

## VI. CONCLUSION

In this paper, we introduce a novel polyatomic FW algorithm, tailored to large-scale LASSO problems. Our procedure leverages polyatomic update directions and partially-corrective re-optimization steps for fast convergence. By preserving sparse intermediate iterates, the algorithm is well-suited for high dimensional regression problems. As revealed by our numerical simulations, the algorithm converges faster than standard state-of-the-art LASSO solvers (from 1.5 to 20 times faster depending on the setup and algorithm), notably including the optimal first-order method FISTA. Future work will include the application of this promising algorithm to the challenging problems of sparse radio-interferometric imaging [1].

**Acknowledgement.** The authors are grateful to Martin Vetterli for his expert advice as well as Quentin Denoyelle and Eric Bezzam for their useful feedback and/or proofreading on earlier versions of this work.

## REFERENCES

- [1] Y. Wiaux, L. Jacques, G. Puy, A. M. M. Scaife, and P. Vanderghenst, "Compressed sensing imaging techniques for radio interferometry," *Monthly Notices of the Royal Astronomical Society*, vol. 395, pp. 1733–1742, May 2009.
- [2] H. Pan, M. Simeoni, P. Hurley, T. Blu, and M. Vetterli, "LEAP: Looking beyond pixels with continuous-space Estimation of Point sources," *Astronomy & Astrophysics*, vol. 608, p. A136, Dec. 2017.
- [3] M. Simeoni, "Towards more accurate and efficient beamformed radio interferometry imaging," tech. rep., EPFL, 2015.
- [4] M. Simeoni and P. Hurley, "SIML: Sieved Maximum Likelihood for array signal processing," in *IEEE ICASSP 2021*, pp. 4535–4539, IEEE, 2021.
- [5] M. Simeoni, S. Kashani, P. Hurley, and M. Vetterli, "Deepwave: a recurrent neural-network for real-time acoustic imaging," *Advances In Neural Information Processing Systems 32 (Nips 2019)*, vol. 32, no. CONF, 2019.
- [6] M. Simeoni, "Functional penalised basis pursuit on spheres," *Applied and Computational Harmonic Analysis*, vol. 53, pp. 1–53, 2021.
- [7] S. Ravishanker, J. C. Ye, and J. A. Fessler, "Image Reconstruction: From Sparsity to Data-adaptive Methods and Machine Learning," *Proceedings of the IEEE*, vol. 108, pp. 86–109, Jan. 2020.
- [8] A. Besson, L. Roquette, D. Perdios, M. Simeoni, M. Arditi, P. Hurley, Y. Wiaux, and J.-P. Thiran, "A physical model of nonstationary blur in ultrasound imaging," *IEEE Transactions on Computational Imaging*, vol. 5, no. 3, pp. 381–394, 2019.
- [9] L. Roquette, M. Simeoni, and P. Hurley, "A functional framework for ultrasound imaging," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 1837–1841, IEEE, 2018.
- [10] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
- [11] R. J. Tibshirani, "The lasso problem and uniqueness," *Electronic Journal of Statistics*, vol. 7, Jan. 2013.
- [12] S. Foucart and H. Rauhut, "An invitation to compressive sensing," in *A mathematical introduction to compressive sensing*, Springer, 2013.
- [13] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.
- [14] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 3, pp. 273–282, 2011.
- [15] E. Frandi, R. Nanculef, S. Lodi, C. Sartori, and J. A. K. Suykens, "Fast and Scalable Lasso via Stochastic Frank-Wolfe Methods with a Convergence Guarantee," *arXiv:1510.07169 [cs, math, stat]*, Oct. 2015.
- [16] M. Unser, J. Fageot, and H. Gupta, "Representer Theorems for Sparsity-Promoting  $\ell_1$  Regularization," *IEEE Transactions on Information Theory*, vol. 62, pp. 5167–5180, Sept. 2016.
- [17] M. Simeoni, "Functional inverse problems on spheres: theory, algorithms and applications," tech. rep., EPFL, 2020.
- [18] M. Unser, J. Fageot, and J. P. Ward, "Splines are universal solutions of linear inverse problems with generalized TV regularization," *SIAM Review*, vol. 59, no. 4, pp. 769–793, 2017.
- [19] H. Gupta, J. Fageot, and M. Unser, "Continuous-domain solutions of linear inverse problems with tikhonov versus generalized tv regularization," *IEEE Transactions on Signal Processing*, vol. 66, no. 17, pp. 4670–4684, 2018.
- [20] J. Fageot and M. Simeoni, "TV-based reconstruction of periodic functions," *Inverse Problems*, vol. 36, no. 11, p. 115015, 2020.
- [21] M. Unser, "A unifying representer theorem for inverse problems and machine learning," *Foundations of Computational Mathematics*, vol. 21, no. 4, pp. 941–960, 2021.
- [22] C. Boyer, A. Chambolle, Y. D. Castro, V. Duval, F. De Gournay, and P. Weiss, "On representer theorems and convex regularization," *SIAM Journal on Optimization*, vol. 29, no. 2, pp. 1260–1281, 2019.
- [23] K. Bredies and M. Carioni, "Sparsity of solutions for variational inverse problems with finite-dimensional data," *Calculus of Variations and Partial Differential Equations*, vol. 59, no. 1, pp. 1–26, 2020.
- [24] A. Badoual, J. Fageot, and M. Unser, "Periodic splines and gaussian processes for the resolution of linear inverse problems," *IEEE Transactions on Signal Processing*, vol. 66, no. 22, pp. 6047–6061, 2018.
- [25] T. Debarre, Q. Denoyelle, M. Unser, and J. Fageot, "Sparsest continuous piecewise-linear representation of data," *arXiv preprint arXiv:2003.10112*, 2020.
- [26] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis, New York, NY: Springer New York, 2013.
- [27] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, pp. 407–499, Apr. 2004. Publisher: Institute of Mathematical Statistics.
- [28] B. A. Turlach, "On algorithms for solving least squares problems under an  $\ell_1$  penalty or an  $\ell_1$  constraint," in *2004 Proceedings of the American Statistical Association, Statistical Computing Section [CD-ROM]*, pp. 2572–2577, Citeseer, 2005.
- [29] Z. Harchaoui, A. Juditsky, and A. Nemirovski, "Conditional gradient algorithms for machine learning," in *NIPS Workshop on Optimization for ML*, vol. 3, pp. 3–2, Citeseer, 2012.
- [30] S. Shalev-Shwartz, N. Srebro, and T. Zhang, "Trading accuracy for sparsity in optimization problems with sparsity constraints," *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 2807–2832, 2010.
- [31] Q. Denoyelle, V. Duval, G. Peyré, and E. Soubies, "The sliding Frank-Wolfe algorithm and its application to super-resolution microscopy," *Inverse Problems*, vol. 36, p. 014001, Jan. 2020.
- [32] K. Bredies and H. K. Pikkarainen, "Inverse problems in spaces of measures," *ESAIM: Control, Optimisation and Calculus of Variations*, vol. 19, pp. 190–218, Jan. 2013.
- [33] K. Bredies, M. Carioni, S. Fanzon, and D. Walter, "Linear convergence of accelerated generalized conditional gradient methods," *arXiv:2110.06756 [math]*, Oct. 2021.
- [34] D. Gong, M. Tan, Y. Zhang, A. V. Hengel, and Q. Shi, "MPGL: An Efficient Matching Pursuit Method for Generalized LASSO," in *AAAI*, 2017.
- [35] N. Boyd, G. Schiebinger, and B. Recht, "The Alternating Descent Conditional Gradient Method for Sparse Inverse Problems," *arXiv:1507.01562 [math]*, July 2015.
- [36] G. Braun, S. Pokutta, D. Tu, and S. Wright, "Blended conditional gradients," in *Proceedings of the 36th International Conference on Machine Learning*, p. 735–743, PMLR, May 2019.
- [37] Y. Li, S. Osher, and UCLA Mathematics Department, Box 951555, Los Angeles, CA 90095-1555, "Coordinate descent optimization for  $\ell^1$  minimization with application to compressed sensing; a greedy algorithm," *Inverse Problems & Imaging*, vol. 3, no. 3, pp. 487–503, 2009.
- [38] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, Dec. 2007.
- [39] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in optimization*, vol. 1, no. 3, pp. 127–239, 2014.
- [40] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM Journal on Imaging Sciences*, vol. 2, pp. 183–202, Jan. 2009.
- [41] K. Scheinberg, D. Goldfarb, and X. Bai, "Fast first-order methods for composite convex optimization with backtracking," *Foundations of Computational Mathematics*, vol. 14, no. 3, pp. 389–417, 2014.
- [42] J. Liang, T. Luo, and C.-B. Schönlieb, "Improving "fast iterative shrinkage-thresholding algorithm": Faster, smarter and greedier," 2021.
- [43] M. Jaggi, "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization," in *Proceedings of the 30th International Conference on Machine Learning*, pp. 427–435, PMLR, Feb. 2013.
- [44] Y. Zhao and X. Huo, "A homotopic method to solve the lasso problems with an improved upper bound of convergence rate," *arXiv preprint arXiv:2010.13934*, 2020.
- [45] A. Chambolle and C. Dossal, "On the Convergence of the Iterates of the "Fast Iterative Shrinkage/Thresholding Algorithm"," *Journal of Optimization Theory and Applications*, vol. 166, pp. 968–982, Sept. 2015.
- [46] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval Research Logistics Quarterly*, vol. 3, no. 1-2, pp. 95–110, 1956.
- [47] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.
- [48] Y. Pati, R. Rezaifar, and P. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, (Pacific Grove, CA, USA), pp. 40–44, IEEE Comput. Soc. Press, 1993.
- [49] S. Mallat, "Adaptive time-frequency decompositions," *Optical Engineering*, vol. 33, p. 2183, July 1994.
- [50] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, "Sparse Solution of Underdetermined Systems of Linear Equations by Stagewise Orthogonal Matching Pursuit," *IEEE Transactions on Information Theory*, vol. 58, pp. 1094–1121, Feb. 2012.
- [51] M. Simeoni, *mathhieu/meo/pycsou: Pycsou 1.0.6*. Zenodo, Apr 2021.
- [52] [https://github.com/AdriaJ/PolyatomicFW\\_SPL](https://github.com/AdriaJ/PolyatomicFW_SPL). Accessed: Jan 2022.