Fast column generation for atomic norm regularization
Marina Vinyes, Guillaume Obozinski

To cite this version:
Marina Vinyes, Guillaume Obozinski. Fast column generation for atomic norm regularization. The 20th International Conference on Artificial Intelligence and Statistics, Apr 2017, Fort Lauderdale, United States. hal-01502575

HAL Id: hal-01502575
https://hal-enpc.archives-ouvertes.fr/hal-01502575
Submitted on 9 Apr 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract

We consider optimization problems that consist in minimizing a quadratic function under an atomic norm\textsuperscript{1} regularization or constraint. In the line of work on conditional gradient algorithms, we show that the fully corrective Frank-Wolfe (FCFW) algorithm—which is most naturally reformulated as a column generation algorithm in the regularized case—can be made particularly efficient for difficult problems in this family by solving the simplicial or conical subproblems produced by FCFW using a special instance of a classical active set algorithm for quadratic programming (Nocedal and Wright, 2006) that generalizes the min-norm point algorithm (Wolfe 1976).

Our experiments show that the algorithm takes advantages of warm-starts and of the sparsity induced by the norm, displays fast linear convergence, and clearly outperforms the state-of-the-art, for both complex and classical norms, including the standard group Lasso.

1 INTRODUCTION

A number of problems in machine learning and structured optimization involve either structured convex constraint sets that are defined as the intersection of a number of simple convex sets or dually, norms of sets that are defined as convex hull of either extreme points or of a collection of sets. A broad class of convex regularizers that can be used to encode a priori knowledge on the structure of the objects to estimate have been described as atomic norms and atomic gauges by Chandrasekaran et al. (2012). The concept of atomic norm has found several applications to design sparsity inducing norms for vectors (Jacob et al., 2009; Obozinski et al., 2011), matrices (Richard et al., 2014; Foygel et al., 2012) and tensors (Tomioka and Suzuki, 2013; Liu et al., 2013; Wimalawarne et al., 2014).

A number of these norms remain difficult to use in practice because it is in general not possible to compute the associated proximal operator or even the norm itself at a reasonable cost. However, the dual norm which is defined as a supremum of dot products with the atoms that define the norm can often be computed efficiently because of the structure of the set of atoms. Also a number of atomic norms are actually naturally defined as infimal convolution of other norms (Jacob et al., 2009; Tomioka and Suzuki, 2013; Liu et al., 2013) and this structure has been used to design either block-coordinate descent approaches or dual ADMM optimization schemes (Tomioka and Suzuki, 2013) involving latent variables associated with the elementary norms convolved.

In this paper, we propose to solve problems regularized or constrained by atomic norms using a fully corrective Frank-Wolfe algorithm—which can be reformulated as simple column generation algorithm in the regularized case—combined with a dedicated active-set algorithm for quadratic programming. Our experiments show that we achieve state-of-the-art performance. We also include a formal proof of the correspondence between the column generation algorithm and Fully Corrective Frank-Wolfe.

After a review of the concept of atomic norms, as well some illustrations, we present a number of the main algorithmic approaches that have been proposed. We then present the scheme we propose and finally some experiments on synthetic and real datasets.

\textsuperscript{1}or more generally atomic gauge
1.1 Notations

$[p]$ denotes the set $\{1, \ldots, p\}$. If $x \in \mathbb{R}^p$, $x_G$ denotes the subvector of $x$ whose entries are indexed by a set $G \subseteq [p]$. Given a function $\psi$, $\psi^*$ denotes its Fenchel conjugate $\psi^*(s) := \max_x \{s^T x - \psi(x)\}$.

$\|M\|_1$ denotes the trace norm of the matrix $M$ defined as the $\ell_1$-norm of its singular values.

2 ATOMIC NORMS

In many machine learning applications, particularly for ill-posed problems, models are constrained structurally so they have a simple representation in terms of some fundamental elements. Examples of such elements include sparse vectors for many sparsity inducing norms, rank-one matrices for the trace norm or low-rank tensors as used in the nuclear tensor norm (Liu et al., 2013). We call atoms these elements and atomic set $\mathcal{A}$ their (possibly infinite) collection. Assuming $\mathcal{A}$ is bounded and centrally symmetric, and provided its convex hull $C_\mathcal{A}$ has non empty interior, we can define an atomic norm $\gamma_\mathcal{A}$ as the norm of unit ball $C_\mathcal{A}$. It can be shown that (in a finite dimensional space) $\gamma_\mathcal{A}(x) := \inf \{\sum_{a \in \mathcal{A}} c_a a^T a \mid \sum_{a \in \mathcal{A}} c_a = x, c_a \geq 0, a \in \mathcal{A}\}$.

The polar norm or dual norm is defined as: $\gamma_\mathcal{A}^*(s) := \sup_{a \in \mathcal{A}} (s^T a)$. If $\mathcal{A}$ is not symmetric, or if $C_\mathcal{A}$ is empty, as long as $\mathcal{A}$ contains the origin and is closed, $\gamma_\mathcal{A}$ can still be defined as a gauge instead of a norm and the theory and algorithms presented in this paper still apply. We restrict the discussion to norms for simplicity. For a reference on gauges, see Rockafellar (1997).

We consider in this paper formulations in which an atomic norm is used as a regularizer, and which lead to an optimization problem of the form

$$\min_{x \in \mathbb{R}^p} f(x) + \gamma_\mathcal{A}(x),$$

where $f$ is a quadratic function. The case where $f$ is more generally twice differentiable is obviously of interest, but beyond the scope of this work.

2.1 Examples of atomic norms

Lasso. The Lasso is a natural example of atomic norm, whose atoms are the $(x e_i)_{i \in [p]}$, where $(e_i)_{i \in [p]}$ is the canonical basis of $\mathbb{R}^p$. The Lasso polar norm is defined as $\Omega_{\text{Lasso}}^*(s) = \max_{i \in [p]} |s_i|$.

Latent group lasso (LGL). The norms introduced in Jacob et al. (2009) are a strong motivating example. For instance, Obozinski and Bach (2016) show that a broad family of tight relaxations for structured sparsity can be written in LGL form. Given a collection of sets $\mathcal{B}$ covering $[p]$ and which can overlap, and fixed positive weights $\delta_B$ for each set $B \in \mathcal{B}$, the atoms of LGL norm are the vectors of norm $\delta_B$ and support in $\mathcal{B}$. The polar LGL norm is defined as $\Omega_{\text{LGL}}^*(s) = \max_{B \in \mathcal{B}} \delta_B \| s_B \|_2$.

In the particular case where $\mathcal{B}$ form a partition of $[p]$ we recover the group Lasso norm. Maurer and Pontil (2012) consider a generalization to a broader family of atomic norms with dual norms of the form $\sup_{M \in \mathcal{M}} \|Ms\|_2$, where $\mathcal{M}$ is a collection of operators. Matrix counterparts of the latent group Lasso norms are the latent group trace norms (Tomisaka and Suzuki 2013; Wimalawarne et al. 2014).

Additive decompositions. There has been interest in the literature for additive matrix decompositions (Agarwal et al. 2012), the most classical example being “sparse+low rank decompositions” which have been proposed for robust PCA and multitask learning (Candès et al. 2011; Chandrasekaran et al. 2011). This formulation leads to a problem of the form $\min_{M \in \mathcal{M}} f(L+S) + \mu \|M\|_1$, which under the form $\min_{M \in \mathcal{M}} f(M) + \gamma_\mathcal{A}(M)$ with $\gamma_\mathcal{A}$ the atomic norm where $\mathcal{A} \in \mathbb{R}^{p_1 \times p_2}$ is defined as

$$\mathcal{A} := \lambda A_1 \cup \mu A_2, \quad \text{where}$$

$$A_1 := \{\pm e_i e_j^T, (i, j) \in [p_1] \times [p_2]\},$$

$$A_2 := \{uv^T, \|u\|_2 = \|v\|_2 = 1\}.$$

As a consequence, $C_{\mathcal{A}}^* = \frac{1}{\lambda} C_{\mathcal{A}_1}^* \cap \frac{1}{\mu} C_{\mathcal{A}_2}^*$, with $C_{\mathcal{A}}^*$ a unit $\ell_\infty$ ball and $C_{\mathcal{A}}^*$ a unit spectral norm ball.

Convex sparse SVD and PCA. A third example are the norms introduced in Richard et al. (2014), including the $(k, q)$-trace norm for which

$$\mathcal{A} := \bigcup_{I \subseteq [p_1]} \{A_{I, J} \mid |I| = k, |J| = q\},$$

with

$$A_{I, J} := \{uv^T \in \mathcal{A} \mid \|u\|_0 \leq k, \|v\|_0 \leq q\},$$

and the sparse-PCA norm for which

$$\mathcal{A} := \bigcup_{I \subseteq [p_1]} \{A_{I, I} \mid |I| = k\},$$

with

$$A_{I, I} := \{uu^T \mid u \in \mathcal{A}_I\},$$

and $A_I$ defined like $A_B$ for LGL.

Beyond these examples a number of structured convex optimization problems encountered in machine learning and operations research that involve combinatorial or structured tasks such as finding permutations or alignments, convex relaxation of structured matrix factorization problems (Bach et al. 2008; Ding et al. 2010), Procrustes analysis, etc., involve difficult convex constraint sets such as ellipitope, the Birkhoff polytope, the set of doubly nonnegative matrices that are naturally written (themselves or their polar) as intersections of simpler sets such as the p.s.d. cone, the positive orthant, simplices, hypercubes, etc, and which lead to optimization problems whose duals are regularized by associated atomic norms.

In fact this is not a norm but only a gauge.
2.2 Existing algorithmic approaches

2.2.1 Conditional gradient algorithms

For many of these norms, it is assumed that an efficient algorithm is available to compute $\arg\max_{a \in A} (a, s)$. For the case of the constrained problem

$$\min_x f(x) \quad \text{s.t.} \quad x \in C_A, \quad (2)$$

this has motivated a number of authors to suggest variants of the conditional gradient algorithm, also known as the Frank-Wolfe algorithm when the objective is quadratic, as a tool of choice to solve problems with atomic norm constraints. Indeed, the principle of conditional gradient algorithms is to build a sequence of approximations to the solution of the problem as convex combinations of extreme points of the constraint sets, which here correspond to atoms, so that the expansion take the form $x = \sum_{i=1}^{k} c_i a_i$, with $\sum_{i=1}^{k} c_i = 1$. This procedure guarantees a feasible sequence. At each iteration a new atom, also called Frank-Wolfe direction or forward direction, is added in the expansion. This atom is the extreme point of the constraint set defined by $a^{t+1} := \arg\max_{a \in A} (a, -\nabla f(x^t))$. The Frank-Wolfe (FW) algorithm writes

$$x^{t+1}_F = (1 - \eta^t) x^t + \eta^t a^{t+1},$$

where $\eta^t \in [0, 1]$ is a scalar stepsize and $x^0 = 0$. It can be set to $\frac{1}{1+t}$ or found by line search.

Other variants of FW algorithms have been proposed, notably, FW with away steps (which we do not describe here), pairwise FW (PWFW) and fully corrective Frank-Wolfe (FCFW). We refer the reader to Lacoste-Julien and Jaggi (2015) for a detailed presentation and summarize hereafter the form of the different updates for PWFW and FCFW. The active set of atoms $A^t$ at time $t$ is recursively defined by $A^{t+1} = \tilde{A}^t \cup \{a^{t+1}\}$ with $\tilde{A}$ the set of active atoms of $A^t$ at the end of iteration $t$, i.e. the ones that contributed with a non-zero coefficient in the expansion of $x^t$.

PWFW makes use of a backward direction also called away atom, and defined as $a^{t+1}_B = \arg\max_{a \in A} (a, -\nabla f(x^t))$, i.e. it is the active atom of largest projection on the gradient direction. The idea in PWFW is to move by transferring weight from the away atom $a^{t+1}_B$ to the FW atom $a^{t+1}$:

$$x^{t+1}_{PWFW} = x^t + \eta^t_B (a^{t+1}_B - a^{t+1})$$

where $\eta^t_B \in [0, c^t_B]$, with $c^t_B \geq 0$ the weight attributed to atom $a^{t+1}_B$ at iteration $t$, and $\eta^t_B$ is found by line search.

The optimal step sizes $\eta^t \in \mathbb{R}$ for FW and PWFW are easily obtained in closed form when $f$ is quadratic.

In FCFW, all weights are reoptimized at each iteration:

$$x^{t+1}_{FCFW} = \arg\min_x f(x) \quad \text{s.t.} \quad x \in \text{Conv} \text{hull}(A^{t+1}).$$

If $A^t = \{a_1, \ldots, a_k\}$, where $k \leq t$ is the number of atoms in $A^t$, the subproblem that has to be solved at each iteration $t$ of FCFW rewrites

$$\min_{c \geq 0} f(\sum_{i=1}^{k} c_i a_i) \quad \text{s.t.} \quad \sum_{i=1}^{k} c_i = 1. \quad (3)$$

Lacoste-Julien and Jaggi (2015) show that PWFW and FCFW converge linearly for strongly convex objectives when $A$ is finite.

Rao et al. (2015) propose a variant of FCFW to solve (2) for $f$ smooth and specifically for atomic norm constraints, with an enhancing “backward step” which applies hard-thresholding to the coefficients $c_i$. To solve (3) they use a projected gradient algorithm.

Beyond constrained optimization problems, the basic conditional gradient algorithm (corresponding to plain FW when $f$ is quadratic) has been generalized to solve problems of the form $\min_x f(x) + \psi(x)$ where the set constraint $C_A$ is replaced by a proper convex function $\psi$ for which the subgradient of $\psi^*$ can be computed efficiently (Bredies et al., 2009; Yu et al., 2014; Bach, 2015) shows that the obtained algorithm can be interpreted as a dual mirror descent algorithm. Yu et al. (2014); Bach (2015) and Nesterov et al. (2015) prove sublinear convergence rates for these algorithms. Corresponding generalizations of PWFW and FCFW are however not obvious. As exploited in (Yu et al. 2014; Harchaoui et al. 2015), if $\psi = h + \gamma_A$, with $h$ a nondecreasing convex function and $\gamma_A$ an atomic norm, and if an upper bound $\rho$ can be specified a priori on $\gamma_A(x^*)$ for $x^*$ a solution of the problem, it can reformulated as

$$\min_{x, \tau} f(x) + h(\tau) \quad \text{s.t.} \quad \gamma_A(x) \leq \tau, \quad \tau \leq \rho, \quad (4)$$

and it is natural to apply the different variant of Frank-Wolfe on the variable $(x, \tau)$, because the FW direction is easy to compute (see Section 3.1).

2.2.2 Proximal block-coordinate descent

In the context where they are applicable, proximal gradient methods provide an appealing alternative to Frank-Wolfe algorithms. However, the former require to be able to compute efficiently the proximal operator of the norm $\gamma_A$ appearing in the objective, which is typically more difficult to compute than the Frank-Wolfe direction.
For a number of atomic norms, we have $A = \bigcup_{j=1}^{J} C_j$ where $C_j$ are convex sets. As a consequence the polar norm takes the form $\gamma_A^*(s) = \max_{C_j} \gamma_{C_j}^*(s)$, with $\gamma_{C_j}$ the atomic norm (or gauge) associated with the set $C_j$, and it is a standard result that

$$\gamma_A(x) = \inf \{ \gamma_{C_1}(z_1) + \ldots + \gamma_{C_J}(z_J) \mid z_1 + \ldots + z_J = x \}.$$ 

Technically, $\gamma_A$ is called the infimal convolution of the norms $(\gamma_{C_j})_j$ (see Rockafellar 1997). In fact most of the norms that we presented in section 2.1 are of this form, including LGL norms, latent group trace norms, norms arising from additive decomposition (obviously by construction), and the norms for sparse SVD and sparse PCA.

For all these norms, problem (1) can be reformulated as

$$\min_{z_1, \ldots, z_J} f(z_1 + \ldots + z_J) + \gamma_{C_1}(z_1) + \ldots + \gamma_{C_J}(z_J).$$

Since the objective is then a sum of a smooth and of a separable function, randomized proximal block-coordinate descent algorithm are typical candidates. These algorithms have attracted a lot of attention in the recent literature (see ? and reference therein) and have been applied successfully to a number of formulations involving convex sparsity inducing regularizers (Shalev-Shwartz and Tewari 2011, Friedman et al. 2010, Gu et al. 2016), where they achieve state-of-the-art performance. Such BCD algorithms where the ones proposed for the norms proposed in Jacob et al. (2009) and Richard et al. (2014).

Unfortunately these algorithms are slow in general even if $f$ is strongly convex because of the composition with the linear mapping $(z_1, \ldots, z_J) \mapsto z_1 + \ldots + z_J$. Intuitively if the atoms of the different norms are similar, then the formulation is badly conditioned. If they are different or essentially decorrelated, BCD remains one of the most efficient algorithms (Shalev-Shwartz and Tewari 2011, Gu et al. 2016).

3  PIVOTING FRANK-WOLFE

After reviewing the form of the corrective step of FCFW and reformulating FCFW in the regularized case as a column generation algorithm, we introduce active-set algorithms to solve efficiently sequences of corrective steps.

3.1 Simplicial and conical subproblems

We focus on the sequence of subproblems that need to be solved at the corrective step of FCFW. Let $k_t := |A^t|$ be the number of selected atoms at iteration $t$, and $A^t \in \mathbb{R}^{p \times k_t}$, the matrix whose columns are the atoms $A^t$,

then, for the constrained problem (2), the subproblem is the simplicial problem:

$$\min_c f(A^t c) \quad \text{s.t.} \quad c \in \Delta^k,$$ (5)

with $\Delta^k := \{ c \in \mathbb{R}^k : \sum_{i=1}^{k} c_i = 1 \}$ the canonical simplex. The regularized problem (4) can be reformulated as the constrained optimization problem (6) on a truncated cone, provided the truncation level $\rho$ is an upper bound of the value of $\gamma_A$ at the optimum. Actually, if $\rho$ is sufficiently large, several Frank-Wolfe algorithms do not depend any longer on the value of $\rho$ and can be interpreted as algorithms in which whole extreme rays of the cone $\{(x, \tau) \mid \gamma_A(x) \leq \tau \}$ enter the active set via the linear minimization oracle, and where the original cone is locally approximated from inside by the simplicial cone obtained as their conical hull. In particular in the case of FCFW, the subproblem considered at the $t$-th iteration takes the form of the conical problem

$$\min_c f(A^t c) + \sum_i c_i \quad \text{s.t.} \quad c \geq 0,$$ (6)

which is simply a Lasso problem with positivity constraints when $f$ is quadratic. The fact that problem (6) can be solved by as sequence of problems of the form (5) is shown in Harchaoui et al. (2015, Sec. 5), who argue that this leads to an algorithm no worse and possibly better. We formally show that the simple column generating scheme presented as Algorithm 1 is in fact exactly equivalent to FCFW applied to the truncated cone formulation as soon as $\rho$ is large enough:

**Proposition 1.** If $f$ is assumed lower bounded by 0 and if $\rho > f(0)$, or more generally if the level sets of $x \mapsto f(x) + \gamma_A(x)$ are bounded and $\rho$ is sufficiently large, then the sequence $(\hat{x}_t)_{t}$ produced by the FCFW algorithm applied to the truncated cone constrained problem (4) and initialized at $(\hat{x}^0, \hat{\tau}^0) = (0, 0)$ is the same as the sequence $(x_t)_{t}$ produced by Algorithm 1 initialized with $\hat{x}^0 = 0$, with equivalent sequences of subproblems, active sets and decomposition coefficients.

See the appendix for a proof. As discussed as well in the appendix, a variant of Algorithm 1 without pruning of the atoms with zero coefficients (at step 7) is derived very naturally as the dual of a cutting plane algorithm.

3.2 Leveraging active-set algorithms for quadratic programming

Problems (5) and (6) can efficiently be solved by a number of algorithms. In particular, an appropriate variant LARS algorithm solves both problem in a finite number of iterations and it is fast if the solution in sparse, in spite of the fact that it solves exactly a sequence of linear systems. Interior point algorithms can always be used, and are often considered to be a
natural choice to solve this step in the literature. For larger scale problems, and if \( f \) has Lipschitz gradients (which is obviously the case for a quadratic function), the forward-backward proximal algorithm can be used as well, since the projection on the simplex for \([5]\) and the asymmetric soft-thresholding for \([6]\) can be computed efficiently. For the constrained case, this is the algorithm used by \( \text{[Rao et al. (2015)]} \).

In our case, we need to solve a sequence of problems of the form \([5]\) or \([6]\), that differ each from the previous one by the addition of a single atom. So being able to use warm-starts is key! If the simplicial problems remains of small size, and if the corresponding Hessians can be computed efficiently, using second order algorithms is likely to outperform first order methods. But the LARS and interior point methods cannot take advantage of warm-starts. Thus, when \( f \) is quadratic, we propose to use active set algorithms for convex quadratic programming (\( \text{[Nocedal and Wright (2006)]} \)). In particular, following \( \text{[Bach (2013)]} \), we propose to apply the active-set algorithm of \( \text{[Nocedal and Wright (2006)]} \) to iteratively solve \([5]\) and \([6]\). This algorithm takes the very simple form of Algorithm \([2]\). In fact, as noted in \( \text{[Bach (2013)]} \), this algorithm is a generalization of the famous min-norm point algorithm (\( \text{[Wolfe (1976)]} \)), the latter being recovered when the Hessian is the identity.

Algorithm \([2]\) is illustrated in Figure \([1]\). The obtained iterates always remain in the positive orthant (i.e. primal feasible). Each update of \( c \) in Algorithm \([2]\) is called a pivot, which is either full-step or drop-step. Given a collection of active atoms indexed by a set \( J \), the solution \( d \) of the non-constrained quadratic program restricted to this set of atoms and obtained by removing the positivity constraints is computed (line 4). If \( d \) lies in the positive orthant, we set \( c = d \), and we say that we perform a full-step. In that case, the index of an atom that must become active (if any), based on gradients, is added to \( J \). If \( d \notin \mathbb{R}^n_+ \), a drop-step is performed: \( c \) is updated as the intersection between segment \([c_{\text{old}}, d]\) and the positive orthant, and the index \( i \) such that \( c_i = 0 \) is dropped from \( J \) (line 13). The algorithm stops if after a full-step, no new index is added in \( J \).

### Algorithm 1: Column generation

1. **Require:** \( f \) convex differentiable, tolerance \( \epsilon \)
2. **Initialization:** \( x^0 = 0 \), \( A^0 = \emptyset \), \( k_0 = 0 \), \( t = 1 \)
3. **repeat**
   4. \( a_t \leftarrow \text{arg max}_{a \in A} (-\nabla f(x^{t-1}), a) \)
   5. \( A_t \leftarrow \{ A_t^{t-1}, a_t \} \)
   6. \( c_t \leftarrow \text{arg min}_{c \geq 0} f(A_t c) + \|c\|_1 \)
   7. \( I_t \leftarrow \{ i \mid c_t^i > 0 \} \)
   8. \( c_t \leftarrow c_t^{-I_t} \)
   9. \( A_t \leftarrow A_t^{I_t} \)
   10. \( x_t \leftarrow A_t c_t \)
   11. \( t \leftarrow t + 1 \)
4. **until** \( \text{max}_{a \in A} (-\nabla f(x^{t-1}), a) \leq \epsilon \)

![Figure 1: Illustration of Algorithm 2. Here, it converges after a drop-step (variable \( c_2 \) is dropped) leading to \( c^1 \) followed by a full-step (along \( c_1 \)) leading to \( c^2 \).](image)

### Algorithm 2 \([c,J]=\text{Active-set}(H,b,c_0,J_0)\)

1. **Solves:** \( P := \text{min}_{c} \{ \text{H}c + b^T c \}. \text{ s.t. } c \geq 0 \)
2. **Initialization:** \( c = c_0 \), \( J = J_0 \)
3. **repeat**
   4. \( d \leftarrow H_{J,J}^{-1} b_J \)
   5. **if** \( d \geq 0 \), **then**
       6. \( c \leftarrow d \) \( \triangleright \) full-step
   7. \( g \leftarrow Hc + b \)
   8. \( k \leftarrow \text{arg min}_{k \notin J, \emptyset, J_0 \cup J} g_k \)
   9. **if** \( g_k \geq 0 \), **then** break, else \( J \leftarrow J \cup \{k\} \) **end**
   10. **else**
       11. \( i^* \leftarrow \text{arg min}_{i} \frac{c_i}{c_i - d_i} \text{ s.t. } c_i - d_i > 0, d_i < 0 \)
       12. \( \tau \leftarrow e_i^T c_{\text{old}} - d_i c_{\text{old}} \quad \triangleright \) drop-step
       13. \( J \leftarrow J \setminus \{i^*\} \)
       14. \( c \leftarrow c + \tau (d - c) \)
       15. **end**
   16. **until** \( g_{J_0 \cup J} \geq 0 \)
17. **return** \( c, J \)

### 3.3 Convergence and computational cost

In this section, we discuss first the convergence of the algorithm and the number of pivots needed for convergence, and the the cost of each pivot.

Algorithm \([2]\) is an instance of min-norm point (MNP) with a general quadratic instead of Euclidean distance,
but the algorithm is affine invariant, so the convergence is the same. MNP is known to be finitely convergent. The positive orthonormal in dimension \( k \) has at most \( 2^k \) faces which is a naive bound on the number of pivots in the active-set at iteration \( t \) of FCFW. But, Lacoste-Julien and Jaggi (2015) prove that MNP is linearly convergent. In practice, the solution is most of the time either strictly inside the orthonormal or in one of the \( k-1 \) dimensional faces in which case it is in fact found in just 1 or respectively 2 iterations! The number of pivots per call is illustrated in Figure 3 upper left.

Let \( s = \max_{a \in A} |a|_0 \) be the sparsity of the atoms, \( k \) the number of active atoms at iteration \( t \) and \( H^t = A^t Q A^t \) the Hessian of the quadratic problem in the active set, where \( Q \) is the Hessian of the quadratic function \( f \).

The cost of one pivot is the cost of computing the Hessian \( H^t \) and its inverse, which is \( O(\min(k^2 s^2, kps + k^3 s)) \) for building the Hessian and an extra \( O(k^3) \) for the inversion. In the active-set with warm starts we only add or remove one atom at a time. We can take advantage of this to efficiently update the Hessian \( H^t \) and its inverse with rank one updates. The computational cost for updating the Hessian is \( O(\min(k s^2, ps + ks)) \) when an atom is added and \( O(k) \) when removing an atom. The additional cost to update \( (H^t)^{-1} \) is then just \( O(k^2) \) in both cases. See the appendix for more details on the rank one updates.

4 EXPERIMENTS

In this section, we report experiments that illustrate the computational efficiency of the proposed algorithm. We consider linear regression problems of the form of (1)

\[
\min_{x} \frac{1}{2} \|Xw - y\|^2 \quad \text{subject to} \quad \gamma Aw \leq \rho,
\]

where \( X \) is a design matrix and \( \gamma A \) the LGL or the sparse-PCA norms described in Section 2. We also considered the constrained version for LGL, \( \min_x f(x) \) s.t. \( \Omega_{\text{LGL}}(w) \leq \rho \), in section 4.2.

Section 4.1 compares the performance of our proposed algorithm with state-of-the-art algorithms for the group Lasso. Section 4.2 presents comparisons with the variants of Frank-Wolfe and with COGEnT on problem involving the latent group Lasso. Section 4.3 provides a comparison with a version of FCFW relying on interior-point solver on larger scale problems. Sections 4.3 and 4.4 provide comparisons with randomized block proximal coordinate descent algorithms. Most experiments are on simulated data to control characteristics of the experiments, except in section 4.5.

4.1 Classical group Lasso

We consider an example with group Lasso regularization with groups of size 10, \( \mathcal{B} = \{ \{1, \ldots, 10\}, \{11, \ldots, 20\}, \ldots \} \). We choose the support of the parameter \( w_0 \in \mathbb{R}^{1000} \) of the model to be \( \{1, \ldots, 50\} \) and all non zero coefficients are set to 2. We generate \( n = 200 \) examples \( (y_i)_{i=1, \ldots, n} \) from \( y = x^T w + \varepsilon \). Block Coordinate Descent (BCD) algorithms are the standard method for this problems but they suffer slow convergence when the design matrix is highly correlated. In this experiment we choose a highly correlated design matrix (with singular values in \( \{1, 0.9^2, \ldots, 0.92^{(p-2)}, 0.92^{(p-1)}\} \) ) to highlight the advantages of our algorithm for the harder instances. We compared our algorithm to our own implementation of BCD and an enhanced BCD from Qin et al. (2013) (hyb-BCD). Figure 2 shows that we outperform both methods.

4.2 k-chain latent group Lasso

We consider a toy example involving latent group Lasso regularization where the groups are chains of continuous indices of length \( k = 8 \), that is where the collection of group is \( \mathcal{B} = \{ \{1, \ldots, k\}, \{2, \ldots, k+1\}, \ldots, \{p-k+1, \ldots, p\} \} \). We choose the support of the parameter \( w_0 \) of the model to be \( \{1, \ldots, 10\} \). Hence, three overlapping chains are needed to retrieve the support of \( w_0 \). We generate \( n = 300 \) examples \( (y_i)_{i=1, \ldots, n} \) from \( y = x^T w + \varepsilon \) where \( x \) is a standard Gaussian vector and \( \varepsilon \sim \mathcal{N}(0, \sigma^2 I_p) \). The noise level is chosen to be \( \sigma = 0.1 \). In upper Figure 3 we show a time comparison of our algorithm on the regularized problem. We implemented Algorithm 1 and three Frank-Wolfe versions: simple FW, FW with line search (FW-ls) and pairwise FW (FW-pw). We compare also with a regularized version of the forward-backward greedy algorithm from Rao et al. (2015) (CoGEnT). In the bottom plot of Figure 3 we show a comparison on the constrained problem. All codes are in MATLAB and we used Rao et al.’s code for the forward-backward greedy algorithm.

Figure 4 illustrates complexity and memory usage of our algorithm for the same experiment. Top plots show
that each call to the active-set algorithm has low cost. Indeed less than two pivots in average, i.e., drop or full steps, are needed to converge. This is clearly due to the use of warm starts. Bottom plot shows the number of active atoms during iterations.

4.3 Hierarchical sparsity

In high-dimensional linear models that involve interaction terms, statisticians usually favor variable selection obeying certain logical hierarchical constraints. In this section we consider a quadratic model (linear + interaction terms) of the form

$$ y = \sum_{i=1}^{p} \beta_i x_i + \sum_{i,j} \beta_{ij} x_i x_j. $$

*Strong and weak hierarchical sparsity* are usually distinguished (see [Bien et al. 2013](#) and reference therein). The Weak Hierarchical (WH) sparsity constraints are that if an interaction is selected, then at least one of its associated main effects is selected, i.e., \( \beta_{ij} \neq 0 \Rightarrow \beta_i \neq 0 \) or \( \beta_j \neq 0 \). We use the latent overlapping group Lasso formulation proposed in Yan and Bien (2015) to formulate our problem. The corresponding collection of groups \( \mathcal{B} \) thus contains the singletons \{i\} and contains for all pairs \{i, j\} the sets \{i, {i, j}\} and \{{j, {i, j}}\} (coupling respectively the selection of \( \beta_i \) with that of \( \beta_j \) or that of \( \beta_i \)). We focussed on WH sparsity which is more challenging here because of the group overlaps, but the approach applies also to the counterpart for strong hierarchical constraints.

### Simulated data

We consider a quadratic problem with \( p = 50 \) main features, which entails that we have \( p \times (p-1)/2 = 1225 \) potential interaction terms and simulate \( n = 1000 \) samples. We choose the parameter \( \beta \) to have 10% of the interaction terms \( \beta_{ij} \) equal to 1 and the rest equal to zero. In order to respect the WH structure, the minimal number of necessary unary terms \( \beta_i \) possible given the WH constrains are included in the model with \( \beta_i = 0.5 \). We compare our algorithm with FCFW combined with an interior point solver (FCFW-ip) instead of the active-set subroutine, and with a degraded version of our algorithm not using warm starts. Figure 3 shows that FCFW-ip becomes slower than our algorithm only beyond 200 seconds. A plausible explanation is that at the begining the sub-problems being solved are small and time is dominated by the search of the new direction; when the size of the problem grows, the active-set with warm start is faster, meaning that the active-set exploits the structure of positivity constraints better than IP, which has to invert bigger matrices. Full corrections of FCFW-ip call the *quadprog* function of MATLAB, which is an optimized C++ routine, whereas our implementation is done in MATLAB. An optimized C implementation of our active-set algorithm, in particular leveraging the
Fast column generation for atomic norm regularization

rank one updates on the inverse Hessian described in sections 3.3 should provide an additional significant speedup.

![Figure 5: Experiments on simulated data for WH sparsity. log-log plot of progress of the duality gap as a function time in seconds.](image)

California housing data set We apply the previous hierarchical mode to the California housing data (Pace and Barry, 1997). The data contains 8 variables, so with interaction terms the initial model contains 36 variables. To make the selection problem more challenging, following She and Jiang (2014), we add 20 main nuisance variables, generated as standard Gaussian random variables corresponding to 370 additional noisy interaction terms. We compare our algorithm to the greedy Forward-Backward algorithm with a truncation parameter $\eta = 0.5$ and with Block Coordinate Descent (BCD). Table 1 shows running time for different levels of regularization $\lambda$. $\lambda = 10^{-3}$ is the value selected by 10-fold cross validation on the validation risk. Figure 6 shows the running time for the different algorithms.

Table 1: Computation time in seconds needed to reach a duality gap of $10^{-3}$ on California housing data set. Time is not reported when larger than 700 seconds.

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$10^{-5}$</th>
<th>$10^{-4}$</th>
<th>$10^{-3}$</th>
<th>$10^{-2}$</th>
<th>$10^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCD</td>
<td>-</td>
<td>585</td>
<td>73</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>CoGEnT</td>
<td>-</td>
<td>1300</td>
<td>14</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>ours</td>
<td>27</td>
<td>1.4</td>
<td>0.4</td>
<td>0.06</td>
<td>0.02</td>
</tr>
</tbody>
</table>

4.4 Sparse PCA

We compare our method to the block proximal gradient descent (BCD) described in Richard et al. (2014). We generate a sparse covariance matrix $\Sigma^*$ of size $150 \times 150$ obtained as the sum of five overlapping rank one blocks $11^2$ of size $k \times k$ with $k = 10$. We generate a noisy covariance with a noise level $\sigma = 0.3$. We consider an $\ell_2$ loss and a regularization by the gauge $\gamma_{A_{\ell_2}}$, described in Section 2 with $k = 10$. The regularization parameter is $\lambda$. Figure 7 shows a time comparison with BCD.

![Figure 6: Experiments on California House data set. log-log plot of progress of the duality gap during computation time.](image)

![Figure 7: Experiments on sparse PCA. log-log plot of progress of the duality gap.](image)

5 CONCLUSION

In this paper, we have shown that to minimize a quadratic function with an atomic norm regularization or constraint, the fully corrective Frank-Wolfe algorithm, which in the regularized case corresponds exactly to a very simple column generating algorithm that is not well known, is particularly efficient given that sparsity make the computation of the reduced Hessian relatively cheap. In particular, the corrective step is solved very efficiently with a simple active-set methods for quadratic programming. The proposed algorithm takes advantage of warm-starts, and empirically outperforms other Frank-Wolfe schemes, block-coordinate descent (when applicable) and the algorithm of Rao et al. (2015). Its performance could be enhanced by low-rank updates of the inverse Hessian. In future work we intend to generalize the algorithm to smooth loss functions using sequential quadratic programming.

Acknowledgements

Marina Vinyes is funded by ANR CHORUS research grant 13-MONU-0005-10.
References


Tomioka, R. and Suzuki, T. (2013). Convex tensor decomposition via structured Schatten norm regulariza-

Wimalawarne, K., Sugiyama, M., and Tomioka, R. (2014). Multitask learning meets tensor factoriza-


