Appendix for Fast column generation for atomic norm regularization

A Proof of Proposition 1

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

Proof. We being by showing that the Frank-Wolfe directions computed for the regularized and the constrained problems are related via a simple relation, already discussed in Yu et al. (2014); Harchaoui et al. (2015).

First note that, the set of extreme points of the truncated cone $\{(x, \tau) \mid \gamma_{\mathcal{A}}(x) \leq \tau \leq \rho\}$ is

$$\bar{\mathcal{A}} = \{(0; 0)\} \cup \big\{(\rho a; \rho) \mid a \in \mathcal{A}\big\}.$$

so that all its non zero extreme points are in bijection with those of $\mathcal{A}$. Then, for a given point x, the Frank-Wolfe directions computed respectively by FCFW in problems (5) and (4) are

$$\begin{cases} a^* & := \arg\max_{a \in \mathcal{A}} \langle \nabla f(x), a \rangle \\ \bar{a}^* & := \arg\max_{(\rho a; \rho u) \in \bar{\mathcal{A}}} \langle \nabla f(x), a \rangle + u, \end{cases}$$

and we have

$$\bar{a}^* = \begin{cases} (0; 0) & \text{if} \quad \gamma^\circ_{\mathcal{A}}(\nabla f(x)) \leq 1 \\ (\rho a^*; \rho) & \text{otherwise,} \end{cases}$$

which shows that unless the atom $(0; 0)$ is selected in $\bar{\mathcal{A}}$, it is the image of the regular FW direction mapped via $a \mapsto (\rho a; \rho)$. Note also that the atom $(0; 0)$ is special in $\bar{\mathcal{A}}$ in that it is the only one for which the second component is different than ρ.

We now prove, by induction on t, the following statement:
$\mathcal{P}^t$: *Letting $(\bar{x}^t, \bar{A}^t, \bar{c}^t)$ denote the triple of values of x, the matrix of active atoms of $\bar{\mathcal{A}}$ and the vector of coefficients $\bar{c}$ of decomposition of x on these atoms, all generated by the FCFW algorithm, then (a) the first column of $\bar{A}^t$ is a column of zeroes corresponding to the atom $(0; 0)$, so that we can write*

$$\bar{A}^t = \begin{pmatrix} 0 & \rho A^t \\ 0 & \rho u^t \end{pmatrix} \in \mathbb{R}^{(d+1) \times (1+k_t)} \quad \text{and} \quad \bar{c}^t = \begin{pmatrix} d_0^t \\ d^t \end{pmatrix} \in \mathbb{R}^{1+k_t},$$

(b) setting $x^t := \bar{x}^t$, and $c^t = \rho d^t$ we have that (x^t, A^t, c^t) is the t-th corresponding triple produced by Algorithm (1) and (c) $\tau^t = \rho(1 - d_0^t) < \rho$ so that the truncation constraint $\{\tau \leq \rho\}$ is inactive.

To prove $\mathcal{P}^0$, note that if $(x^0; \tau^0) = (0; 0)$, then we trivially have $\bar{A}^0 = (0; 0)$ and $\bar{\mathcal{A}}^0$ has the desired form, we have $\bar{x}^0 = c_0^0 \cdot 0_d = x^0$ with $\bar{c}^0 = d_0^0 = 1$ so that $\bar{c}^0$ satisfies the simplex constraints; finally $\tau^0 < \rho$.

We now assume $\mathcal{P}^{t-1}$ is true and prove that so is $\mathcal{P}^t$. In the FCFW algorithm, the new direction chosen cannot be $(0; 0)$ since $d_0^{t-1} > 0$, which entails this atoms is already in the active set and because the algorithm is fully corrective (which prevents the forward direction to be an atom already in the active set), so that it must be of the form $(\rho a^t, \rho)$ which, given that by induction $\bar{x}^{t-1} = x^{t-1}$, entails that a^t is indeed the same direction as the one chosen by Algorithm (1) .

Letting $\bar{A}^t$ is the matrix whose columns are the atoms used in the expansion of x^t, then $\bar{x}^t = A^t c^t$ and letting $x^t = \bar{x}^t$, then the triple (x^t, A^t, c^t) is the one generated by Algorithm (1) . This entails that $\bar{A}^t$ is indeed of the announced form and that the sub-matrix A^t is indeed the one used by Algorithm (1) .

Now the optimization problem solved in the corrective step of FCFW is thus

$$\min_{x, \tau, d} \quad f(x) + \tau \quad \text{s.t.}$$
$$x = \rho A^t d, \quad \tau = \rho u^t d, \quad \bar{c} = (c_0; d) \in \Delta^{k_t+1},$$

with $u^t = 1_{k_t}^\top$ and k_t the number of currently active atoms.

Eliminating x and τ we obtain

$$\min_{d \geq 0} f(\rho A^t d) + \rho\, 1_{k_t}^\top d \quad \text{s.t.} \quad 1_{k_t}^\top d \leq 1,$$

and with the change of variable $c = \rho d$, we get

$$\min_{c \geq 0} f(A^t c) + \|c\|_1 \quad \text{s.t.} \quad \|c\|_1 \leq \rho$$

But, since $\gamma_{\mathcal{A}^t}(x) = \inf\big\{\|c\|_1 \mid c \in \mathbb{R}_+^{k_t}, x = A^t c\big\}$, we can rewrite the previous problem equivalently as

$$\min_x f(x) + \gamma_{\mathcal{A}^t}(x) \quad \text{s.t.} \quad \gamma_{\mathcal{A}^t}(x) \leq \rho.$$

We first conclude the argument assuming $f \geq 0$ and $\rho > f(0)$. In that case, we have

$$\gamma_{\mathcal{A}^t}(x^t) \leq f(x^t) + \gamma_{\mathcal{A}^t}(x^t) \leq f(0) + \gamma_{\mathcal{A}^t}(0) = f(0) < \rho,$$

so that the inequality constraint is inactive for all t at the optimum in the two last problems above and can be removed. We thus showed that the optimization problem of the corrective step of the FCFW algorithm on problem (4) is equivalent to the problem solved at step 6 of Algorithm (1) , and that $\|c^t\|_1 < \rho$ which entails that $d_0^t = 1 - \|d^t\|_1 = 1 - \frac{1}{\rho}\|c\|_1 > 0$ and so that the atom $(0;0)$ remains in $\bar{\mathcal{A}}^{t+1}$. The induction step is completed which thus proves the result.

Now, if we do not assume that f is lower bounded, but we assume instead that the level sets of $f + \gamma_{\mathcal{A}}$ are bounded, then Algorithm (1) generates a sequence x^t which is bounded since the sequence $\left(f(x^t) + \gamma_{\mathcal{A}^t}(x^t)\right)_t$ is a monotonically decreasing sequence. But since for all x, $f(x) + \gamma_{\mathcal{A}^t}(x) \geq f(x) + \gamma_{\mathcal{A}}(x)$, the monotonicity also implies that the sequence $(x^t)_t$ remains in the bounded set $\{x \mid f(x) + \gamma_{\mathcal{A}}(x) \leq f(0)\}$. Since f is assumed continuous this entails that $\left(f(x^t)\right)_t$ is bounded which entails that so is $\left(\gamma_{\mathcal{A}^t}(x^t)\right)_t$ so if ρ is chosen such that $\rho > \sup_t \gamma_{\mathcal{A}^t}(x^t)$ then the FCFW algorithm applied on problem (4) will generate the same sequence as Algorithm (1) . This value of ρ is not known a priori, but is required by neither algorithms. □

Connection with cutting plane algorithms

It is well known that the Frank-Wolfe algorithm is an instance of a column generation algorithm (Forsgren et al., 2015). We explain in this section how Algorithm 1 is naturally derived as such.

Column generation algorithms correspond to cutting plane algorithms in the dual. The principle of the latter algorithms is to solve a sequence of constrained optimization problems that are relaxations of the original problem, where the constraints introduced are gradually tightening the relaxation around the optimum. The new constraint introduced at each iteration is called a *cut* since it cuts the previous relaxed constraint set in order to reduce it. A new cut is typically determined as a constraint of the original problem which is violated by a current solution s^t to the relaxed problem. Such as new constraint is called a *deep cut*. For problems of the form $\min_{s \in C_{\mathcal{A}}^\circ} f^*(s)$ and given that $C_{\mathcal{A}}^\circ = \{s \mid \langle s, a\rangle \leq 1,\ a \in \mathcal{A}\}$, a most violated constraint by a dual variable s can be computed as the inequality $\langle s, a\rangle \leq 1$ for the atom a which is a *conjugate direction* to s, that is a solution to $\max_{a \in \mathcal{A}} \langle s, a\rangle$. Indeed, this yields an atom a such that $\langle a, s\rangle$ is maximal.

After t iterations the relaxed problem to solve in the dual is of the form

$$\min_s f^*(-s) \quad \text{s.t.} \quad \langle a_i, s\rangle \leq 1,\ \forall i \in [\![t]\!], \tag{1}$$

for $\mathcal{A}^t := (a_i)_{i \in [\![t]\!]}$ a sequence of atoms of $\mathcal{A}$.

It is immediate to check that the corresponding primal algorithm is a version of Algorithm 1 in which all atoms are stored. The classical constrained version of Frank-Wolfe correspond a cutting plane algorithm in the dual problem regularized by the dual norm, where this regularization is reformulated as a conic constraint like in formulation (4) in the main paper.

B Rank one updates of the Hessian and its inverse in active-set

Let H^t be the Hessian of the quadratic problem in active-set algorithm and B^t its inverse. Let Q be the Hessian of the quadratic function f. We have $H^t = A^{t\top} Q A^t$. We use the Sherman–Morrison–Woodbury matrix inversion formula in the following equations.

When we add an atom a_{t+1}, we have updates

$$H^{t+1} = \begin{bmatrix} H^t & v \\ v^\top & a_{t+1}^\top Q a_{t+1} \end{bmatrix}$$

and

$$B^{t+1} = \begin{bmatrix} B^t + \alpha B^t v v^\top B^t & -\alpha B^t v \\ -\alpha (B^t v)^\top & \alpha \end{bmatrix}$$

where $v = A^{t\top} Q a_{t+1}$ and $\alpha = (a_{t+1}^\top Q a_{t+1} - v^\top B v)^{-1}$.

When removing an atom, H^{t+1} is obtained removing the corresponding column and row. For clarity, let us assume that we want to remove the last atom. We have

$$H^t = \begin{bmatrix} \tilde{H}^t & v \\ v^\top & \nu \end{bmatrix}$$

and

$$B^{t+1} = \begin{bmatrix} \tilde{B}^t & w \\ w^\top & \beta \end{bmatrix}.$$

Then,

$$H^{t+1} = \tilde{H}^t,$$

$$B^{t+1} = \tilde{B}^t + \frac{\beta \tilde{B}^t v v^\top \tilde{B}^t - (w^\top v - 1)(w v^\top \tilde{B}^t + \tilde{B}^t v w^\top) + v^\top \tilde{B} v w w^\top}{(w^\top v - 1)^2 - \beta v^\top \tilde{B}^t v}.$$

C Additional experiments

C.1 Hierarchical sparsity

Additional plot for experiment of section Hierarchical sparsity on simulated dataset to give a better idea of the improvement brought over interior points methods

Figure 1 shows the number of matrix inversions per size of the matrix. The interior point solver requires 6-7 times more matrix inversions that the active-set algorithm for most of the iterations of the algorithm (in particular, the ones involving larger Hessian), and for the active-set algorithm the inverse Hessian updates could be done in time $\mathcal{O}(k^2)$ instead of $\mathcal{O}(k^3)$.

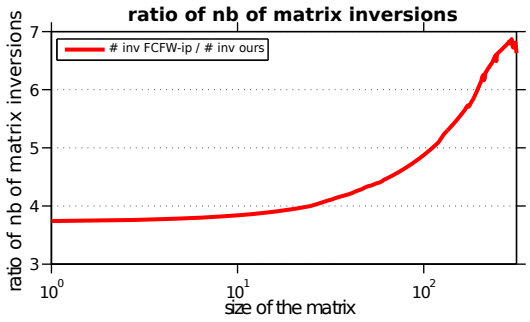

Figure 1: Number of matrix inversions for FCFW-ip divided by the number of matrix inversions in our method per size of the matrix (on WH simulated data).

C.2 Sparse PCA

See Figure 2 for a representation of the ground truth matrix and the noisy covariance used in experiments.

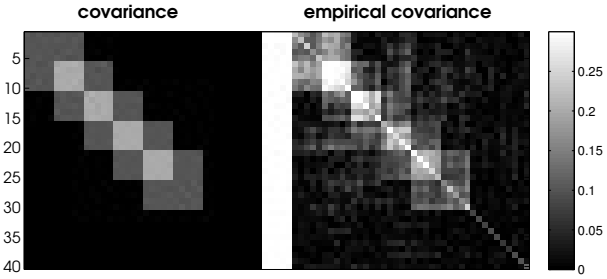

Figure 2: Zoom on 40 first variables of true covariance(left) and empirical covariance(right) for a noise level $\sigma = 0.3$.

References

Forsgren, A., Gill, P. E., and Wong, E. (2015). Primal and dual active-set methods for convex quadratic programming. *Mathematical Programming*, pages 1–40.

Harchaoui, Z., Juditsky, A., and Nemirovski, A. (2015). Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1–2):75–112.

Yu, Y., Zhang, X., and Schuurmans, D. (2014). Generalized conditional gradient for sparse estimation. *arXiv preprint arXiv:1410.4828*.