Frank-Wolfe Splitting via Augmented Lagrangian Method

Gauthier Gidel
MILA, DIRO Université de Montréal

Fabian Pedregosa
UC Berkeley & ETH Zurich

Simon Lacoste-Julien
MILA, DIRO Université de Montréal

Abstract

Minimizing a function over an intersection of convex sets is an important task in optimization that is often much more challenging than minimizations over each individual constraint set. While traditional methods such as Frank-Wolfe (FW) or proximal gradient descent assume access to a linear or quadratic oracle on the intersection, splitting techniques take advantage of the structure of each sets, and only require access to the oracle on the individual constraints. In this work we develop and analyze the Frank-Wolfe Augmented Lagrangian (FW-AL) algorithm, a method for minimizing a smooth function over convex compact sets related by a linear consistency constraint that only requires access to a linear minimization oracle over the individual constraints. It is based on the Augmented Lagrangian Method (ALM), also known as Method of Multipliers, but unlike most existing splitting methods it only requires access to linear (instead of quadratic) minimization oracles. We use recent advances in the analysis of Frank-Wolfe and the alternating direction method of multipliers algorithms to prove a sublinear convergence rate over general convex compact sets and a linear convergence rate of our algorithm over for polytopes.

1 Introduction

The Frank-Wolfe (FW) or conditional gradient algorithm has seen an impressive revival in recent years, notably due to its very favorable properties for the optimization of sparse problems (Jaggi, 2013). This algorithm assumes knowledge of a linear minimization oracle (LMO) over the set of constraints. This oracle is inexpensive to compute for sets such as the ℓ_1 or trace norm ball. However, inducing complex priors often requires to consider *multiple* constraints, leading to a constraint set formed by the intersection of the original constraints. Unfortunately, evaluating the LMO over this intersection can be very challenging even if the LMOs on the individual sets are inexpensive.

The problem of minimizing over an intersection of convex constraints is pervasive in machine learning and signal processing. For example, one can seek for a matrix that is both sparse and low rank by constraining the solution to have *both* small ℓ_1 and trace norm (Richard et al., 2012) or find a set of brain maps which are both sparse and piecewise constant by constraining both the ℓ_1 and total variation pseudonorm (Gramfort et al., 2013). Furthermore, some challenging optimization problems such as multiple sequence alignment are naturally expressed over an intersection of constraints (Yen et al., 2016a) or more generally as a linear relationship between these sets (Huang et al., 2017).

The objective of this paper is to describe and analyze FW-AL, an optimization method that can solve convex optimization problems subject to multiple constraint sets, assuming we have access to a LMO on each of the set.

Previous work. The vast majority of methods proposed to solve optimization problems over an intersection of sets of constraints rely on the availability of a projection operator onto each set (see e.g. the recent reviews (Glowinski et al., 2017; Ryu and Boyd, 2016), which cover the more general proximal splitting framework). One of the most popular algorithm in this framework is the alternating direction method of multipliers (ADMM), proposed by Glowinski and Marroco (1975), studied by Gabay and Mercier (1976), and revisited many times; see for instance (Boyd et al., 2011; Yan and Yin, 2016). On some cases, such as constraints on the trace norm (Cai et al., 2010) or the latent group lasso (Obozinski et al., 2011), the projection step can be a time-consuming operation, while

Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

the Frank-Wolfe LMO is much cheaper in both cases. Moreover, for some highly structured polytopes such as those appearing in alignment constraints (Alayrac et al., 2016) or Structured SVM (Lacoste-Julien et al., 2013), there exists a fast and elegant dynamic programming algorithm to compute the LMO, while there is no known practical algorithm to compute the projection. Hence, the development of splitting methods that use the LMO instead of the proximal operator is of key practical interest.

Recently, Yen et al. (2016a) proposed a FW variant for objectives with a linear loss function over an intersection of polytopes named Greedy Direction Method of Multipliers (GDMM). A similar version of GDMM is also used in (Yen et al., 2016b; Huang et al., 2017) to optimize a function over a Cartesian product of spaces related to each other by a linear consistency constraint. The constraints are incorporated through the augmented Lagrangian method and its convergence analysis crucially uses recent progress in the analysis of ADMM by Hong and Luo (2017). Nevertheless, we argue in Sec. C.1 that there are technical issues in these analysis since some of the properties used have only been proven for ADMM and do not hold in the context of GDMM. Furthermore, even though GDMM provides good experimental results in these papers, the practical applicability of the method to other problems is dampened by overly restrictive assumptions: the loss function is required to be linear or quadratic, leaving outside loss functions such as logistic regression, and the constraint needs to be a polytope, leaving outside domains such as the trace norm ball.

Yurtsever et al. (2015) propose an algorithm (UniPDGrad) based on Lagrangian method holding splitting with LMO as a particular case. We develop the comparison with FW-AL in App. B.2.

Contributions. Our main contribution is the development of a novel variant of FW for the optimization of a function over product of spaces related to each other by a linear consistency constraint and its rigorous analysis. We name this method Frank-Wolfe via Augmented Lagrangian method (FW-AL). With respect to Yen et al. (2016a,b); Huang et al. (2017), our framework generalizes GDMM by providing a method to optimize a general class of functions over an intersection of an arbitrary number of compact sets, which are *not* restricted to be polytopes. Moreover, we argue that the previous proofs of convergence were incomplete: in this paper, we prove a new challenging technical lemma providing a growth condition on the augmented dual function which allows us to fix the missing parts.

We show that FW-AL converges for any smooth objective function. We prove that a standard gap measure converges linearly (i.e., with a geometric rate) when the constraint sets are polytopes, and sublinearly for general compact convex sets. We also show that when the function is strongly convex, the sum of this gap measure and the feasibility gives a bound on the distance to the set of optimal solutions. This is of key practical importance since the applications that we consider (e.g., minimization with trace norm constraints) verify these assumptions.

The paper is organized as follows. In Section 2, we introduce the general setting, provide some motivating applications and present the augmented Lagrangian formulation of our problem. In Section 3, we describe the algorithm FW-AL and provide its analysis in Section 4. Finally, we present in Section 5 illustrative experiments.

2 Problem Setting

We will consider the following minimization problem,

$$\begin{aligned} &\underset{\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(k)}}{\text{minimize}} \; f(\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(k)})\,, \\ &\text{s.t. } \boldsymbol{x}^{(k)} \in \mathcal{X}_k,\; k \in [K],\; \sum_{k=1}^{K} A_k \boldsymbol{x}^{(k)} = 0\,, \end{aligned} \tag{OPT}$$

where $f : \mathbb{R}^m \to \mathbb{R}$ is a convex differentiable function and for $k \in [K]$, $\mathcal{X}_k \subset \mathbb{R}^{d_k}$ are convex compact sets and A_k matrices of size $d \times d_k$. We will call the constraint $\sum_{k=1}^{K} A_k \boldsymbol{x}^{(k)} = 0$ the *linear consistency constraint*. We assume that we have access to the linear minimization oracle $\mathrm{LMO}_k(\boldsymbol{r}) \in \arg\min_{\boldsymbol{s}\in\mathcal{X}_k} \langle \boldsymbol{s}, \boldsymbol{r}\rangle$, $k \in [K]$. In Section 2.1 we detail some arising problem in machine learning and signal processing modeled by this framework. We denote by $\mathcal{X}^*$ the set of optimal points of the optimization problem (OPT) and we will assume that this problem is feasible, i.e., the set of solutions is non empty.

2.1 Motivating Applications

We now present some motivating applications of our problem setting, including examples where special case versions of FW-AL were used. This previous work provide additional evidence for the practical significance of the FW-AL algorithm.

Multiple sequence alignment and motif discovery (Yen et al., 2016a) are problems in which the domain is described as an intersection of alignment constraints and consensus constraints, two highly structured polytopes.

The linear oracle on both sets can be solved by dynamic programing whereas there is no known practical algorithm to project onto. A factorwise approach to the dual of the structured SVM objective (Yen et al., 2016b) can also be cast as constrained problem over a Cartesian product of polytopes related to each other by a linear consistency constraint. As often in structured prediction, the output domain grows exponentially, leading to extremely high dimensional polytopes. Once again, dynamic programming can be used to compute the linear oracle in structured SVMs at a lower computational cost than the potentially intractable projection. The algorithms proposed by Yen et al. (2016a) and Yen et al. (2016b) are in fact a particular instance of FW-AL, where the objective function is respectively linear and quadratic.

Finally, simultaneously sparse (ℓ_1 norm constraint) and low rank (trace norm constraint) matrices (Richard et al., 2012) is another class of problems where the constraints consists of an intersection of sets with simple linear oracle but expensive projection. This example is a novel application of FW-AL and is developed in Section 5.

2.2 Augmented Lagrangian Reformulation

It is possible to reformulate (OPT) into the problem of finding a saddle point of an augmented Lagrangian (Bertsekas, 1996), in order to split the constraints in a way in which the linear oracle is computed over a product space. We first rewrite (OPT) as follows:

$$\min_{\boldsymbol{x}^{(k)} \in \mathcal{X}_k,\, k \in [K]} f(\boldsymbol{x}) \quad \text{s.t.} \quad M\boldsymbol{x} = 0\,, \tag{1}$$

where $\boldsymbol{x} := \left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(K)}\right)$, and M is such that

$$M\boldsymbol{x} = 0 \Leftrightarrow \sum_{k=1}^{K} A_k \boldsymbol{x}^{(k)} = 0\,. \tag{2}$$

We can now consider the augmented Lagrangian formulation of (1):

$$\begin{aligned} &\underset{(\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(K)})}{\text{minimize}} \ \max_{\boldsymbol{y} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(K)}, \boldsymbol{y}) \\ &\text{s.t.} \quad \boldsymbol{x}^{(k)} \in \mathcal{X}_k, \quad k \in \{1, \dots, K\} \\ &\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) := f(\boldsymbol{x}) + \langle \boldsymbol{y}, M\boldsymbol{x} \rangle + \tfrac{\lambda}{2}\|M\boldsymbol{x}\|^2. \end{aligned} \tag{OPT2}$$

For notational simplicity, we denote $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_K \subset \mathbb{R}^{d_1 + \ldots + d_K} = \mathbb{R}^m$. This formulation is the one onto which our algorithm FW-AL is applied.

Intersection of sets. One potential application of our work is the optimization over the intersection of an arbitrary number of sets $\cap_{k=1}^{K} \mathcal{X}_k$. In that case the matrix M of the ALM formulation is such that $M\boldsymbol{x} = 0 \Leftrightarrow \boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k+1)}$, $k \in [K-1]$.

Notation and assumption. In this paper, we denote by $\|\cdot\|$ the ℓ_2 norm and $\operatorname{dist}(\boldsymbol{x}, \mathcal{C}) := \inf_{\boldsymbol{x}' \in \mathcal{C}} \|\boldsymbol{x} - \boldsymbol{x}'\|$ its associated distance to a set. Note that in finite dimensional space this infimum is achieved when $\mathcal{C}$ is a non-empty closed set. We assume that f is L-smooth on $\mathbb{R}^p$, i.e., differentiable with L-Lipschitz continuous gradient. This is characterized by the following inequality for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^p$:

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}')\| \leqslant L\|\boldsymbol{x} - \boldsymbol{x}'\|\,. \tag{3}$$

This assumption is standard in convex optimization (Nesterov, 2004). Notice that the FW algorithm does not converge if the objective function is not at least continuously differentiable (Nesterov, 2016, Example 1). In our analysis, we will also use the observation that $\frac{\lambda}{2}\|M \cdot\|^2$ is generalized strongly convex.[1] We say that a function h is *generalized strongly convex* when it takes the following general form:

$$h(\boldsymbol{x}) := g(A\boldsymbol{x}) + \langle \boldsymbol{b}, \boldsymbol{x} \rangle\,, \quad \forall \boldsymbol{x} \in \mathbb{R}^p\,, \tag{4}$$

where $A \in \mathbb{R}^{d \times p}$ and g is μ_g-*strongly convex* w.r.t. the Euclidean norm on $\mathbb{R}^d$ with $\mu_g > 0$. Recall that a μ_g-strongly (differentiable) convex function $g : \mathbb{R}^d \to \mathbb{R}$ is one such that, $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$,

$$g(\boldsymbol{x}) \geqslant g(\boldsymbol{x}') + \left\langle \boldsymbol{x} - \boldsymbol{x}', \nabla g(\boldsymbol{x}') \right\rangle + \frac{\mu_g}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2\,.$$

3 FW-AL Algorithm

Our algorithm takes inspiration from both Frank-Wolfe and the augmented Lagrangian method. The augmented Lagrangian method alternates a primal update on $\boldsymbol{x}$ (approximately) minimizing[2] the augmented Lagrangian $\mathcal{L}(\cdot, \boldsymbol{y}_t)$, with a dual update on $\boldsymbol{y}$ by taking a gradient ascent step on $\mathcal{L}(\boldsymbol{x}_{t+1}, \cdot)$. The FW-AL algorithm follows the general iteration of the augmented Lagrangian method, but with the crucial difference that Lagrangian minimization is replaced by one Frank-Wolfe step on $\mathcal{L}(\cdot, \boldsymbol{y}_t)$. The algorithm is thus composed by two loops: an outer loop presented in (5) and an inner loop noted $\mathcal{FW}$ which can be chosen to be one of the FW step variants described in Alg. 1 or 2.

[1] This notion has been studied by Wang and Lin (2014) and in the Frank-Wolfe framework by Beck and Shtern (2016) and Lacoste-Julien and Jaggi (2015).

[2] An example of approximate minimization is taking one proximal gradient step, as used for example in the Linearized ADMM algorithm (Goldfarb et al., 2013; Yang and Yuan, 2013).

FW Augmented Lagrangian method (FW-AL)

At each iteration $t \geqslant 1$, we first update the primal variable blocks $\boldsymbol{x}_t$ with a Frank-Wolfe step and then update the dual multiplier $\boldsymbol{y}_t$ using the updated primal variable:

$$\begin{cases} \boldsymbol{x}_{t+1} = \mathcal{FW}(\boldsymbol{x}_t; \mathcal{L}(\cdot, \boldsymbol{y}_t)) \,, \\ \boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \eta_t M \boldsymbol{x}_{t+1} \,, \end{cases} \tag{5}$$

where $\eta_t > 0$ is the step size for the dual update and $\mathcal{FW}$ is either Alg. 1 or Alg. 2 (see more in App. A).

FW steps. In FW-AL we need to ensure that the $\mathcal{FW}$ inner loop makes sufficient progress. For general sets, we can use one iteration of the classical Frank-Wolfe algorithm with line-search (Jaggi, 2013) as given in Algorithm 2. When working over polytopes, we can get faster (linear) convergence by taking one *non-drop* step (defined below) of the away-step variant of the FW algorithm (AFW) (Lacoste-Julien and Jaggi, 2015), as described in Algorithm 1). Other possible variants are discussed in Appendix A. We denote by $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ the iterates computed by FW-AL after t steps and by $\mathcal{A}_t$ the set of atoms previously given by the FW oracle (including the initialization point). If the constraint set is the convex hull of a set of atoms $\mathcal{A}$, the iterate $\boldsymbol{x}_t$ has a sparse representation as a convex combination of the first iterate and the atoms previously given by the FW oracle. The set of atoms which appear in this expansion with non-zero weight is called the *active set* $\mathcal{S}_t$. Similarly, since $\boldsymbol{y}_t$ is by construction in the cone generated by $\{M\boldsymbol{x}_s\}_{s \leqslant t}$, the iterate $\boldsymbol{y}_t$ is in the span of $M\mathcal{A}_t$, that is, they both have the sparse expansion:

$$\boldsymbol{x}_t = \sum_{\boldsymbol{v} \in \mathcal{S}_t} \alpha_{\boldsymbol{v}}^{(t)} \boldsymbol{v}, \quad \text{and} \quad \boldsymbol{y}_t = \sum_{\boldsymbol{v} \in \mathcal{A}_t} \xi_{\boldsymbol{v}}^{(t)} M\boldsymbol{v} \,, \tag{6}$$

When we choose to use the AFW Alg. 1 as inner loop algorithm, it can choose an *away* direction to remove mass from "bad" atoms in the active set, i.e. to reduce $\alpha_{\boldsymbol{v}}^{(t)}$ for some $\boldsymbol{v}$ (see L11 of Alg. 1), thereby avoiding the zig-zagging phenomenon that prevents FW from achieving a better convergence rate (Lacoste-Julien and Jaggi, 2015). On the other hand, the maximal step size for an *away* step can be quite small ($\gamma_{\max} = \alpha_{\boldsymbol{v}}^{(t)}/1-\alpha_{\boldsymbol{v}}^{(t)}$, where $\alpha_{\boldsymbol{v}}^{(t)}$ is the weight of the away vertex in (6)), yielding to arbitrary small suboptimality progress when the line-search is truncated to such small step-sizes. A step removing an atom from the active set is called a *drop step* (this is further discussed in Appendix A), and Alg. 1 loops until a non-drop step is obtained. In App. A.1 we prove that the number of drop steps after t iterations cannot be larger than $t+1$. This bound is crucial in order to make the convergence results theoretically significant. Indeed, if we were not able to upper bound the cumulative number of drop steps, FW-AL with Alg. 1 as inner loop could be stuck (since Alg. 1 only returns when it performs a non-drop step).

Algorithm 1 Away-step Frank-Wolfe (one non-drop step) : (Lacoste-Julien and Jaggi, 2015)

input: $(\boldsymbol{x}, \mathcal{S}, \mathcal{A}, \varphi)$
`drop_step` $\leftarrow$ `true` (initialization of the boolean)
while `drop_step` = `true` **do**
  $\boldsymbol{s} \leftarrow \text{LMO}\left(\nabla\varphi(\boldsymbol{x})\right)$
  $\boldsymbol{v} \in \arg\max_{\boldsymbol{v} \in \mathcal{S}} \langle \nabla\varphi(\boldsymbol{x}), \boldsymbol{v} \rangle$
  $g^{FW} \leftarrow \langle \nabla\varphi(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{s} \rangle$ *(Frank-Wolfe gap)*
  $g^{A} \leftarrow \langle \nabla\varphi(\boldsymbol{x}), \boldsymbol{v} - \boldsymbol{x} \rangle$ *(Away gap)*
  if $g^{FW} \geqslant g^{A}$ **then** *(FW direction is better)*
    $\boldsymbol{d} \leftarrow \boldsymbol{s} - \boldsymbol{x}$ and $\gamma_{\max} \leftarrow 1$
  else *(Away direction is better)*
    $\boldsymbol{d} \leftarrow \boldsymbol{x} - \boldsymbol{v}$ and $\gamma_{\max} \leftarrow \alpha_{\boldsymbol{v}}/(1 - \alpha_{\boldsymbol{v}})$
  end if
  Compute $\gamma \in \arg\min_{\gamma \in [0, \gamma_{\max}]} \varphi\left(\boldsymbol{x} + \gamma \boldsymbol{d}\right)$
  if $\gamma < \gamma_{\max}$ **then** *(first non-drop step)*
    `drop_step` $\leftarrow$ `false`
  end if
  Update $\boldsymbol{x} \leftarrow \boldsymbol{x} + \gamma \boldsymbol{d}$
  Update $\alpha_{\boldsymbol{v}}$ according to (6)
  Update $\mathcal{S} \leftarrow \{\boldsymbol{v} \in \mathcal{A} \text{ s.t. } \alpha_{\boldsymbol{v}} > 0\}$ *(active set)*
end while
return: $(\boldsymbol{x}, \mathcal{S})$

Algorithm 2 FW(one step) : (Frank and Wolfe, 1956)

input: $(\boldsymbol{x}, \varphi)$
Compute $\boldsymbol{r} \leftarrow \nabla\varphi(\boldsymbol{x})$
Compute $\boldsymbol{s} \leftarrow \arg\min_{\boldsymbol{s} \in \mathcal{X}} \langle \boldsymbol{s}, \boldsymbol{r} \rangle$
$\gamma \in \arg\min_{\gamma \in [0,1]} \varphi(\boldsymbol{x} + \gamma(\boldsymbol{s} - \boldsymbol{x}))$
Update $\boldsymbol{x} \leftarrow (1 - \gamma)\boldsymbol{x} + \gamma \boldsymbol{s}$
return: $\boldsymbol{x}$

4 Analysis of FW-AL

Solutions of (OPT2) are called saddle points, equivalently a vector $(\boldsymbol{x}^*, \boldsymbol{y}^*) \in \mathcal{X} \times \mathbb{R}^d$ is said to be a saddle point if the following is verified for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathbb{R}^d$,

$$\mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}) \leqslant \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^*) \leqslant \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*) \,. \tag{7}$$

Our assumptions (convexity of f and $\mathcal{X}$, feasibility of $M\boldsymbol{x} = 0$, and crucially boundedness of $\mathcal{X}$) are sufficient for strong duality to hold (Boyd and Vandenberghe, 2004, Exercise 5.25(e)). Hence, the set of saddle points is not empty and is equal to $\mathcal{X}^* \times \mathcal{Y}^*$, where $\mathcal{X}^*$ is the set of minimizer of (OPT) and $\mathcal{Y}^*$ the set of maximizers of the augmented dual function d:

$$d(\boldsymbol{y}) := \min_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) \,. \tag{8}$$

One of the issue of ALM is that it is a non feasible method and consequently the function suboptimality is no longer a satisfying convergence criterion (since it can be negative). In the following section we wonder what could be the quantities to look at in order to get a sufficient condition of convergence.

4.1 Convergence Measures

Variants of ALM (also known as the methods of multipliers) update at each iteration both the primal variable and the dual variable. For the purpose of analyzing the popular ADMM algorithm, Hong and Luo (2017) introduced two positive quantities which they called primal and dual gaps that we re-use in the analysis of our algorithm. Let $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ be the current primal and dual variables after t iterations of the FW-AL algorithm (5), the dual gap is defined as

$$\Delta_t^{(d)} := d^* - d(\boldsymbol{y}_t) \quad \text{where} \quad d(\boldsymbol{y}_t) := \min_{\boldsymbol{x}\in\mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_t) \quad (9)$$

and $d^* := \max_{\boldsymbol{y}\in\mathbb{R}^d} d(\boldsymbol{y})$. It represents the dual suboptimality at the t-th iteration. On the other hand, the "primal gap" at iteration t is defined as

$$\Delta_t^{(p)} := \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_t) - d(\boldsymbol{y}_t), \quad t \geqslant 0\,. \quad (10)$$

Notice that $\Delta_t^{(p)}$ is *not* the suboptimality associated with the primal function $p(\cdot) := \max_{\boldsymbol{y}\in\mathbb{R}^d} \mathcal{L}(\cdot, \boldsymbol{y})$ (which is infinite for every $\boldsymbol{x}$ non feasible). In this paper, we also define

$$\Delta_t := \Delta_t^{(p)} + \Delta_t^{(d)}\,. \quad (11)$$

It is important to realize that since ALM is a non-feasible method, the standard convergence convex minimization certificates could become meaningless. In particular, the quantity $f(\boldsymbol{x}_t) - f^*$ might be negative since $\boldsymbol{x}_t$ does not necessarily belong to the constraint set of (OPT). Without any additional assumption on f, the primal and dual gap introduced in (9) and (10) can be small whereas the current iterate can be arbitrary far from the optimal point.

Illustrative example. Previous theoretical results for GDMM (Yen et al., 2016a,b; Huang et al., 2017) only provided a rate on both gaps (9) and (10) which is not sufficient to derive guarantees on how close is iterate to the optimal point. In this paper, we are able to prove that the feasibility $\|M\boldsymbol{x}\|^2$ converges to 0 as fast as Δ_t. But even with these quantities vanishing, the suboptimality of the closest feasible point can be arbitrary bigger than the suboptimality of a point ϵ-close to the optimum. To illustrate this point, we will propose an objective function and two constraint sets following the intuition that a small value of $f(\boldsymbol{x}_t) - f^*$ can allow the point $\boldsymbol{x}_t$ to be really far from $\boldsymbol{x}^*$ where the gradient of f is larger that around the optimum, implying that feasible points close to $\boldsymbol{x}_t$ can have large suboptimality. Concretely, let $r \gg \epsilon > 0$ two fixed variables and let,

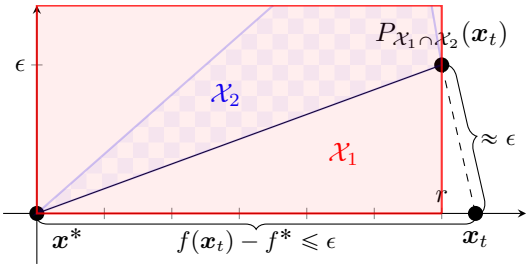


Figure 1: Even with $f(\boldsymbol{x}_t) - f^*$ small, an $\boldsymbol{x}_t$ far from $\boldsymbol{x}^*$ can lead to large value for $f(P_{\mathcal{X}_1\cap\mathcal{X}_2}(\boldsymbol{x}_t)) - f^*$

$$f(u,v) = \epsilon\tfrac{u^2 e^{r^4 v}}{(r+\epsilon)^2}, \quad \mathcal{X}_1 := [0,r]^2 \quad \text{and}$$
$$\mathcal{X}_2 := \{(u,v)\,;\, \|(u,v)\|_1 \leqslant r+\epsilon\,;\, v \geqslant \tfrac{\epsilon}{r}u,\ v \leqslant u\}\,.$$

The sets are compact and the function f is Lipschitz on these sets. The sets and the position of the non feasible iterate $\boldsymbol{x}_t$ are represented in Figure 1. We clearly have $f^* = 0$ and $\boldsymbol{x}^* = (0,0)$. We set $\boldsymbol{x}_t := (0, r')$ such that its projection $P_{\mathcal{X}_1\cap\mathcal{X}_2}(\boldsymbol{x}_t)$ is (r,ϵ), explicitly we can show that $r' = r + \epsilon^2/r$. We then have

$$f(\boldsymbol{x}_t) \leqslant \epsilon, \quad \text{and} \quad f(P_{\mathcal{X}_1\cap\mathcal{X}_2}(\boldsymbol{x}_t)) = \epsilon\tfrac{r^2 e^{r^4\epsilon}}{(r+\epsilon)^2} \geqslant \tfrac{\epsilon}{2} e^{r^4\epsilon}\,,$$

whereas $\forall \boldsymbol{x} \in B(\boldsymbol{x}^*, \epsilon)$, $f(\boldsymbol{x}) \leqslant \frac{\epsilon^3 e^{r^4\epsilon/2}}{r^2}$ which can be arbitrary smaller than $f(P_{\mathcal{X}_1\cap\mathcal{X}_2}(\boldsymbol{x}_t))$.

In conclusion, a rate on Δ_t and on the feasibility can lead to a bad approximate solution for the original problem (OPT). In order to obtain a convergence certificate leading to a "good" approximate solution we will also consider strongly convex functions. With this additional assumption, we proved a convergence rate on the distance of the iterates to the solution of (OPT).

4.2 Properties of the augmented Lagrangian dual function

The augmented dual function plays a key role in our convergence analysis. One of our main technical contribution is the proof of a new property of this function which can be understood as a growth condition. This property is due to the smoothness of the objective function and the compactness of the constraint set. We will need an additional technical assumption called *interior qualification* (a.k.a *Slater's conditions*).

Assumption 1. *The sets $(\mathcal{X}_k)$ are convex compact sets such that there exist an interior feasible point, i.e., there exist $\boldsymbol{x}^{(k)} \in \operatorname{relint}(\mathcal{X}_k)$, $k \in [K]$ such that $\sum_{k=1}^K A_k \boldsymbol{x}^{(k)} = 0$.*

Recall that $\boldsymbol{x} \in \operatorname{relint}(\mathcal{X})$ if and only if $\boldsymbol{x}$ is an interior point relative to the affine hull of $\mathcal{X}$. This assumption is pretty standard and weak in practice. It is a particular case of constraint qualifications (Holmes, 1975; Gowda and Teboulle, 1990). With this assumption we can deduce a global property on the augmented dual function that can be summarized as quadratic growth condition on a ball of size $L_\lambda D^2$ and a linear growth condition outside of this ball. Optimization literature named such properties *error bounds* (Pang, 1997).

Theorem 1 (Error bound)**.** *Let d be the augmented dual function (8), if f is a L-smooth convex function, $\mathcal{X}$ a compact convex set and if Assump. 1 holds, then there exist a constant $\alpha > 0$ such that for all $\boldsymbol{y} \in \mathbb{R}^d$,*

$$d^* - d(\boldsymbol{y}) \geqslant \frac{\alpha^2}{2} \min\left\{ \frac{\operatorname{dist}(\boldsymbol{y}, \mathcal{Y}^*)^2}{L_\lambda D^2}, \operatorname{dist}(\boldsymbol{y}, \mathcal{Y}^*) \right\}, \quad (12)$$

where $D := \max_{(\boldsymbol{x},\boldsymbol{x}') \in \mathcal{X}^2} \|\boldsymbol{x} - \boldsymbol{x}'\|$ is the diameter of $\mathcal{X}$ and $L_\lambda := L + \lambda \|M^\top M\|$.

This theorem, proven in App. C.1, is crucial in our analysis. In our *descent lemma* (21), we want to relate the gap decrease to a quantity proportional to the gap. A consequence of (12) is a lower bound of interest: (22).

Issue in previous proofs. In previous work, Yen et al. (2016a, Theorem 2) have a constant called R_Y in the upper bound of Δ_t which may be infinite and lead to the trivial bound $\Delta_t \leqslant \infty$. Actually R_Y is an upper bound on the distance of the dual iterate $\boldsymbol{y}_t$ to the optimal solution set $\mathcal{Y}^*$ of the augmented dual function, since this quantity is not proven to be bounded an element is missing in the convergence analysis. In their convergence proof, Yen et al. (2016b) and Huang et al. (2017) use Lemma 3.1 in (Hong and Luo, 2012) (which also appears as Lemma 3.1 in the published version (Hong and Luo, 2017)). This lemma states a result not holding for all $\boldsymbol{y} \in \mathbb{R}^d$ but instead for $(\boldsymbol{y}_t)_{t \in \mathbb{N}}$, which is the sequence of dual variables computed by the ADMM algorithm used in (Hong and Luo, 2017). This sequence cannot be assimilated to the sequence of dual variables computed by the GDMM algorithm since the update rule for the primal variables in each algorithm is different, the primal variable are updated with FW steps in one algorithm and with a proximal step in the other. The properties of this proximal step are intrinsically different from the FW steps computing the updates on the primal variables of FW-AL. To our knowledge, there is no easy fix to get a result as the one claimed by Yen et al. (2016b, Lem. 4) and Huang et al. (2017, Lem. 4). More details are provided in App. B.1.

4.3 Specific analysis for FW-AL

Convergence over general convex sets. When $\mathcal{X}$ is a general convex compact set and f is L-smooth, Algorithms 1 and 2 are able to perform a decrease on the objective value proportional to the square of the suboptimality (Jaggi, 2013, Lemma 5), (Lacoste-Julien and Jaggi, 2015, (31)), we will call this a *sublinear decrease* since it leads to a sublinear rate for the suboptimality: for any $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{y} \in \mathbb{R}^d$ they compute $\boldsymbol{x}^+ := \mathcal{FW}(\boldsymbol{x}; \mathcal{L}(\cdot, \boldsymbol{y}))$, such that for all $\gamma \in [0, 1]$,

$$\mathcal{L}(\boldsymbol{x}^+, \boldsymbol{y}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) \leqslant \gamma(d(\boldsymbol{y}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})) + \frac{\gamma^2 L_\lambda D^2}{2}, \quad (13)$$

where L_λ is the Lipschitz constant of $\mathcal{L}$ and D the diameter of $\mathcal{X}$. Recall that $d(\boldsymbol{y}) := \min_{\boldsymbol{x}' \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}', \boldsymbol{y})$. Note that setting $\gamma = 0$ provides us $\mathcal{L}(\boldsymbol{x}^+, \boldsymbol{y}) \leqslant \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$ and optimizing the RHS respect to γ provides us a decrease proportional to $(d(\boldsymbol{y}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}))^2$. The GDMM algorithm of (Yen et al., 2016a,b; Huang et al., 2017) relied on the assumption of $\mathcal{X}$ being polytope, hence we obtain from this sublinear decrease a completely new result on ALM with FW. This result covers the case of the simultaneously sparse and low rank matrices (29) where the trace norm ball is not a polytope.

Theorem 2 (Rate of FW-AL with Alg. 2)**.** *Under Assumption 1, if $\mathcal{X}$ is a convex compact set and f is a L-smooth convex function and M has the form described in (2), then using any algorithm with* sublinear decrease *(13) as inner loop in* FW-AL *(5) and $\eta_t := \min\left\{\frac{2}{\lambda}, \frac{\alpha^2}{2\delta}\right\} \frac{2}{t+2}$ we have that there exists a bounded $t_0 \geqslant 0$ such that $\forall t \geqslant t_0$,*

$$\Delta_t \leqslant \frac{4\delta(t_0 + 2)}{t + 2}, \quad \min_{t_0 \leqslant s-1 \leqslant t} \|M\boldsymbol{x}_s\|^2 \leqslant \frac{O(1)}{t - t_0 + 1} \quad (14)$$

where $D := \max_{\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}} \|\boldsymbol{x} - \boldsymbol{x}'\|$ is the diameter of $\mathcal{X}$, $L_\lambda := L + \lambda \|M^\top M\|$ the Lipschitz constant of $\mathcal{L}$ the AL function, $\delta := L_\lambda D^2$ and α is defined in Thm.1.

In App. D.2 we provide an analysis for different step size schemes and explicit bounds on t_0.

Convergence over Polytopes. On the other hand, if $\mathcal{X}$ is a polytope, recent advances on FW proposed global linear convergence rates for a generalized strongly convex objective using FW with away steps (Lacoste-Julien and Jaggi, 2015; Garber et al., 2016). Note that in the augmented formulation, $\lambda > 0$ and thus $\frac{1}{2}\|M \cdot\|^2$ is a generalized strongly convex function, making $\mathcal{L}(\cdot, \boldsymbol{y})$ a generalized strongly convex function for any $\boldsymbol{y} \in \mathbb{R}^d$. We can then use such linearly convergent algorithms to improve the rate of FW-AL. More precisely, we will use the fact that Algorithm 1 performs *geometric decrease* (Lacoste-Julien and Jaggi, 2015, Theorem 1): for $\boldsymbol{x}^+ := \mathcal{FW}(\boldsymbol{x}; \mathcal{L}(\cdot, \boldsymbol{y}))$, there exists $\rho_A < 1$ such that for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathbb{R}^d$,

$$\mathcal{L}(\boldsymbol{x}^+, \boldsymbol{y}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) \leqslant \rho_A \big[\min_{\boldsymbol{x}' \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}', \boldsymbol{y}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) \big]. \quad (15)$$

The constant ρ_A (Lacoste-Julien and Jaggi, 2015) depends on the smoothness, the generalized strong convexity of $\mathcal{L}(\cdot, \boldsymbol{y})$ (does not depend on $\boldsymbol{y}$, but depends on M) and the condition number of the set $\mathcal{X}$ depending on its geometry (more details in App. A.3).

Theorem 3 (Rate of FW-AL with inner loop Alg. 1)**.** *Under the same assumptions as in Thm.2 and if moreover $\mathcal{X}$ is a polytope and f a generalized strongly convex function, then using Alg 1 as inner loop and a constant step size $\eta_t = \frac{\lambda \rho_A}{4}$, the quantity Δ_t decreases by a uniform amount for finite number of steps t_0 as,*

$$\Delta_{t+1} - \Delta_t \leqslant -\frac{\lambda \alpha^2 \rho_A}{8} \,, \tag{16}$$

until $\Delta_{t_0} \leqslant L_\lambda D^2$. Then for all $t \geqslant t_0$ we have that the gap and the feasibility violation decrease linearly as,

$$\Delta_t \leqslant \frac{\Delta_{t_0}}{(1+\kappa)^{t-t_0}} \,, \quad \|M\boldsymbol{x}_{t+1}\|^2 \leqslant \frac{16}{\lambda \cdot \rho_A} \frac{\Delta_{t_0}}{(1+\kappa)^{t-t_0}} \,,$$

where $\kappa := \min\left\{\frac{\rho_A}{2}, \frac{\rho_A \lambda \alpha^2}{8 L_\lambda D^2}\right\}$ and $L_\lambda := L + \lambda \|M^\top M\|$.

Strongly convex functions. When the objective function f is strongly convex, we are able to give a convergence rate for the distance of the primal iterate to the optimum. As argued in Sec. 4.1, an iterate close to the optimal point lead to a "better" approximate solution than an iterate achieving a small gap value.

Theorem 4. *Under the same assumptions as in Thm. 2, if f is a μ-strongly convex function, then the set of optimal solutions $\mathcal{X}^*$ is reduced to $\{\boldsymbol{x}^*\}$ and,*

$$\min_{8t_0+15 \leqslant s \leqslant t} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 \leqslant \frac{4K}{\mu} \frac{O(1)}{t - 8t_0 - 14} \,. \tag{17}$$

Moreover if $\mathcal{X}$ is a compact polytope the distance of the current point to the optimal set vanishes as,

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \leqslant \frac{2K\Delta_{t_0}(\sqrt{2}+1)}{\mu(\sqrt{1+\kappa})^{t-t_0}} \,. \tag{18}$$

For an intersection of sets, the three theorems above give stronger results than (Yen et al., 2016b; Huang et al., 2017) since we prove that the distance to the optimal point as well as the feasibility condition vanish linearly.

Proof sketch of Thm 2 and 3 Our goal is to obtain a convergence rate on the sum gaps (9) and (10). First we show that the dual gap verifies

$$\Delta^{(d)}_{t+1} - \Delta^{(d)}_t \leqslant -\eta_t \langle M\boldsymbol{x}_{t+1}, M\hat{\boldsymbol{x}}_{t+1}\rangle \tag{19}$$

where $\hat{\boldsymbol{x}}_{t+1} := \arg\min_{\boldsymbol{x}\in\mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_{t+1})$. Similarly, we prove the following inequality for the primal gap

$$\begin{aligned}\Delta^{(p)}_{t+1} - \Delta^{(p)}_t \leqslant \eta_t \|M\boldsymbol{x}_{t+1}\|^2 \\ + \mathcal{L}(\boldsymbol{x}_{t+2}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) \\ - \eta_t \langle M\boldsymbol{x}_{t+1}, M\hat{\boldsymbol{x}}_{t+1}\rangle \,.\end{aligned} \tag{20}$$

Summing (19) and (20) and using that $\|M\boldsymbol{x}_{t+1} - M\hat{\boldsymbol{x}}_{t+1}\|^2 \leqslant \frac{2}{\lambda}\left(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1})\right)$, we get the following *fundamental descent lemma*,

$$\begin{aligned}\Delta_{t+1} - \Delta_t \leqslant \mathcal{L}(\boldsymbol{x}_{t+2}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) \\ + \frac{2\eta_t}{\lambda}\left(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1})\right) \\ - \eta_t \|M\hat{\boldsymbol{x}}_{t+1}\|^2 \,.\end{aligned} \tag{21}$$

We now crucially combine (12) in Thm. 1 and the fact that $\Delta^{(d)}_t \leqslant \operatorname{dist}(\boldsymbol{y}^t, \mathcal{Y}^*)\|M\hat{\boldsymbol{x}}_{t+1}\|$ to obtain,

$$\tfrac{\alpha^2}{2L_\lambda D^2} \min\{\Delta^{(d)}_{t+1}, L_\lambda D^2\} \leqslant \|M\hat{\boldsymbol{x}}_{t+1}\|^2 \,, \tag{22}$$

and then,

$$\begin{aligned}\Delta_{t+1} - \Delta_t \leqslant \mathcal{L}(\boldsymbol{x}_{t+2}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) \\ + \frac{2\eta_t}{\lambda}\left(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1})\right) \\ - \eta_t \tfrac{\alpha^2}{2L_\lambda D^2} \min\{\Delta^{(d)}_{t+1}, L_\lambda D^2\} \,.\end{aligned} \tag{23}$$

Now the choice of the algorithm to get $\boldsymbol{x}_{t+2}$ from $\boldsymbol{x}_{t+1}$ and $\boldsymbol{y}_{t+1}$ is decisive:

If $\mathcal{X}$ is a polytope and if an algorithm with a *geometric decrease* (15) is used, setting $\eta_t = \frac{\lambda \cdot \rho_A}{4}$ we obtain

$$\begin{aligned}\Delta_{t+1} - \Delta_t \leqslant -\frac{\rho_A}{2}\left(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1})\right) \\ - \frac{\lambda \cdot \rho_A}{4}\|M\boldsymbol{x}_{t+1}\|^2 \,.\end{aligned}$$

Since $\mathcal{L}(\boldsymbol{x}_{t+2}, \boldsymbol{y}_{t+1}) \leqslant \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1})$ (L13), we have

$$\Delta^{(p)}_{t+1} \leqslant \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1}) \,, \tag{24}$$

leading us to a geometric decrease for all $t \geqslant t_0$,

$$\Delta_{t+1} \leqslant \frac{\Delta_t}{1+\kappa} \quad \text{where} \quad \kappa := \tfrac{\rho_A}{2} \min\left\{1, \tfrac{\lambda \alpha^2}{8 L_\lambda D^2}\right\}. \tag{25}$$

Additionally we can deduce from (21) that,

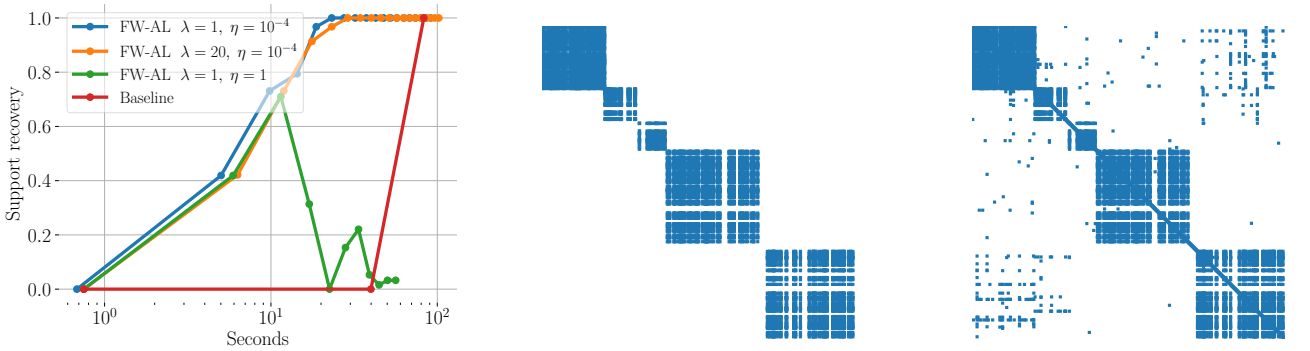
$$\eta_t \|M\hat{\boldsymbol{x}}_{t+1}\|^2 \leqslant \Delta_t \quad \text{and} \quad \eta_t \|M\boldsymbol{x}_{t+1}\|^2 \leqslant 4\Delta_t \,. \tag{26}$$

If $\mathcal{X}$ is not a polytope, we can use an algorithm with a *sublinear decrease* (13) to get from (23) that $\forall t \geqslant 0$,

$$\Delta_{t+1} - \Delta_t \leqslant -a\eta_t \min\{\Delta_{t+1}, \delta\} + (a\eta_t)^2 \tfrac{C}{2} \,, \tag{27}$$

where a, δ and C are three positive constants. Setting $\eta_t = \frac{2}{a(t+2)}$ we can prove that there exists $t_0 \geqslant \frac{C}{\delta}$ s.t.,

$$\Delta_{t+1} \leqslant \frac{4\delta(2+t_0)}{(t+2)}, \quad \forall t \geqslant t_0 \,. \tag{28}$$



(a) Fraction of the support recovered. (b) Matrix recovered with FW-AL. (c) Matrix recov. with the baseline.

Figure 2: Fig. 2a represent the fraction of the support of Σ recovered ($d^2 = 1.6 \cdot 10^7$ and the matrix computed is thresholded at 10^{-2}). The baseline is the generalized forward backward algorithm. Fig 2b and 2c represent the matrices recovered for $d^2 = 10^6$ after one minute of computation.

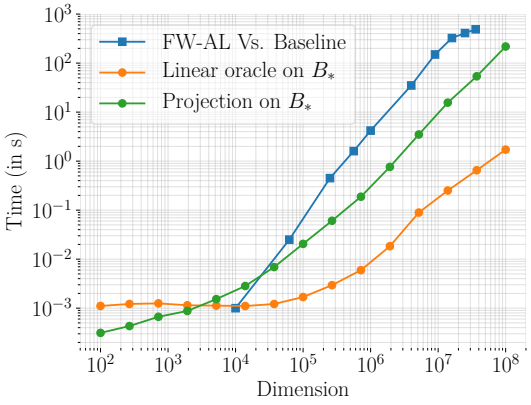


Figure 3: Time complexity of the linear oracle and the projection on the trace norm ball. The blue curve represent time spent by the generalized forward backward algorithm to reach a better point than the one computed by FW-AL.

5 Illustrative Experiments

Recovering a matrix that is simultaneously low rank and sparse has applications in problems such as covariance matrix estimation, graph denoising and link prediction (Richard et al., 2012). We compared our algorithm with proximal splitting method on a covariance matrix estimation problem. We define the $\|\cdot\|_1$ norm of a matrix S as $\|S\|_1 := \sum_{i,j} |S_{i,j}|$ and its trace norm as $\|S\|_* := \sum_{i=1}^{\operatorname{rank}(S)} \sigma_i$, where σ_i are the singular values of S. Given a symmetric positive definite matrix $\hat{\Sigma} \in \mathbb{R}^{d\times d}$ the objective function is defined as

$$\min_{S\succeq 0, \|S\|_1 \leqslant \beta_1, \|S\|_* \leqslant \beta_2} \|S - \hat{\Sigma}\|_2^2 \,. \tag{29}$$

The linear oracle for $\mathcal{X}_1 := \{S \succeq 0,\ \|S\|_1 \leqslant \beta_1\}$ is

$$\operatorname{LMO}_{\mathcal{X}_1}(D) := \beta_1 \tfrac{E_{ij}+E_{ji}}{2},\ (i,j) \in \operatorname*{arg\,min}_{(i,j)\in d\times d} D_{i,j} + D_{j,i}$$

where (E_{ij}) is the standard basis of $\mathbb{R}^{d\times d}$. The linear oracle for $\mathcal{X}_2 := \{S \succeq 0,\ \|S\|_* \leqslant \beta_2\}$ is

$$\operatorname{LMO}_{\mathcal{X}_2}(D) := \beta_2 \cdot U_1^\top U_1 \,, \tag{30}$$

where $D = [U_1, \dots, U_d] \operatorname{diag}(\sigma_1, \dots, \sigma_d) [U_1, \dots, U_d]^\top$. It can be computed efficiently by the Lanczos algorithm (Paige, 1971; Kuczyński and Woźniakowski, 1992) whereas the standard splitting method to solve (29) requires to compute projections over the trace norm ball via a complete diagonalization which is $O(d^3)$. For large d, the full diagonalization becomes untractable, while the Lanczos algorithm is more scalable and requires less storage (see Fig. 3).

The experimental setting is done following Richard et al. (2012): we generated a block diagonal covariance matrix Σ to draw n vectors $\boldsymbol{x}_i \sim \mathcal{N}(0, \Sigma)$. We use 5 blocks of the form $\boldsymbol{v}\boldsymbol{v}^\top$ where $\boldsymbol{v} \sim \mathcal{U}([-1, 1])$. In order to enforce sparsity we only kept the entries (i, j) such that $|\Sigma_{i,j}| > .9$. Finally, we add a gaussian noise $\mathcal{N}(0, \sigma)$ on each entry $\boldsymbol{x}_i$ and observe $\hat{\Sigma} = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T$. In our experiment $n = d, \sigma = 0.6$. We apply our method, as well as the baseline from (Richard et al., 2012), which is the generalized forward backward splitting of Raguet et al. (2013), to optimize (29) performing projections over the constraint sets. The results are presented in Fig. 2 and 3 . The oracle for this algorithm is also slower in the large scale, as can be seen in Fig. 3. We can say that our algorithm performs better than the baseline for high dimensional problems for two reasons: in high dimensions, only one projection on the trace norm ball B_* can take hours (green curve) whereas solving a LMO over B_* takes few seconds, additionally, the iterates computed by FW-AL are naturally sparse and low rank, so we then get a better estimation at the beginning of the optimization as illustrated in Fig. 2b and 2c.

Acknowledgements

Thanks to an anonymous reviewer for helpful comments. Work partially supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodorowska-Curie grant agreement 748900.

References

J.-B. Alayrac, P. Bojanowski, N. Agrawal, I. Laptev, J. Sivic, and S. Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *CVPR*, 2016.

A. Beck and S. Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Math. Program.*, 2016.

D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Athena Scientific, 1996.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 2011.

J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 2010.

D. J. M. Danskin. The directional derivative. In *The Theory of Max-Min and Its Application to Weapons Allocation Problems*. Springer Berlin Heidelberg, 1967.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics*, 1956.

D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 1976.

D. Garber, D. Garber, and O. Meshi. Linear-memory and decomposition-invariant linearly convergent conditional gradient algorithm for structured polytopes. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *NIPS*, 2016.

R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis*, 1975.

R. Glowinski, S. J. Osher, and W. Yin. *Splitting Methods in Communication, Imaging, Science, and Engineering*. Springer, 2017.

D. Goldfarb, S. Ma, and K. Scheinberg. Fast alternating linearization methods for minimizing the sum of two convex functions. *Mathematical Programming*, 2013.

M. S. Gowda and M. Teboulle. A comparison of constraint qualifications in infinite-dimensional convex programming. *SIAM Journal on Control and Optimization*, 1990.

A. Gramfort, B. Thirion, and G. Varoquaux. Identifying predictive regions from fMRI with TV-L1 prior. In *International Workshop on Pattern Recognition in Neuroimaging*. IEEE, 2013.

R. B. Holmes. Geometric functional analysis and its applications. 1975.

M. Hong and Z.-Q. Luo. On the linear convergence of the alternating direction method of multipliers. *arXiv:1208.3922*, 2012.

M. Hong and Z.-Q. Luo. On the linear convergence of the alternating direction method of multipliers. *Math. Program.*, 2017.

X. Huang, I. E.-H. Yen, R. Zhang, Q. Huang, P. Ravikumar, and I. Dhillon. Greedy direction method of multiplier for MAP inference of large output domain. In *AISTATS*, 2017.

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML*, 2013.

S. Kakade, S. Shalev-Shwartz, and A. Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript*, 2009.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. *ECML*, 2016.

J. Kuczyński and H. Woźniakowski. Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start. *SIAM. J. Matrix Anal. & Appl.*, 1992.

S. Lacoste-Julien and M. Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. 2015.

S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-Coordinate Frank-Wolfe Optimization for Structural SVMs. In *ICML*, 2013.

S. Łojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 1963.

Y. Nesterov. *Introductory Lectures on Convex Optimization*. Applied Optimization. Springer US, 2004.

Y. Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *CORE Discussion Paper*, 2016.

G. Obozinski, L. Jacob, and J.-P. Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv:1110.0413*, 2011.

C. C. Paige. *The computation of eigenvalues and eigenvectors of very large sparse matrices.* PhD thesis, University of London, 1971.

J.-S. Pang. A posteriori error bounds for the linearly-constrained variational inequality problem. *Mathematics of Operations Research*, 1987.

J.-S. Pang. Error bounds in mathematical programming. *Math. Program.*, 1997.

B. T. Polyak. Gradient methods for minimizing functionals. *Zh. Vychisl. Mat. Mat. Fiz.*, 1963.

H. Raguet, J. Fadili, and G. Peyré. A generalized forward-backward splitting. *SIAM Journal on Imaging Sciences*, 2013.

E. Richard, P.-A. Savalle, and N. Vayatis. Estimation of simultaneously sparse and low rank matrices. In *ICML*, 2012.

R. T. Rockafellar and R. J. Wets. Variational analysis. 1998.

E. Ryu and S. Boyd. Primer on monotone operator methods. *Appl. Comput. Math*, 2016.

S. Shalev-Shwartz and Y. Singer. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. *Mach. Learn.*, 2010.

P.-W. Wang and C.-J. Lin. Iteration complexity of feasible descent methods for convex optimization. *Journal of Machine Learning Research*, 2014.

M. Yan and W. Yin. Self equivalence of the alternating direction method of multipliers. In *Splitting Methods in Communication, Imaging, Science, and Engineering*. Springer, 2016.

J. Yang and X. Yuan. Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of computation*, 2013.

I. Yen, X. Huang, K. Zhong, R. Zhang, P. Ravikumar, and I. Dhillon. Dual decomposed learning with factorwise oracle for structural SVM with large output domain. In *NIPS*, 2016b.

I. E.-H. Yen, X. Lin, J. Zhang, P. Ravikumar, and I. Dhillon. A convex atomic-norm approach to multiple sequence alignment and motif discovery. In *ICML*, 2016a.

A. Yurtsever, Q. T. Dinh, and V. Cevher. A universal primal-dual convex optimization framework. In *NIPS*, 2015.

A Frank Wolfe inner Algorithms

A.1 Upper bound on the number of drop steps

Proposition 1 (Sparsity of the iterates and upper bound on the number of drop steps)**.** *The iterates computed by* FW-AL *have the following properties,*

1. *After t iterations, the iterates $\boldsymbol{x}_t$ (resp. $\boldsymbol{y}_t$) are a convex (resp. conic) combination of their initialization and the oracle's outputs (resp. times M) for the first t iterations.*
2. *If the algorithm $\mathcal{FW}$ is* AFW *(Alg. 1), and if we initialize our algorithm at a vertex, after t iterations of the main loop the cumulative number of drop steps performed in the inner algorithm 1 is upper bounded by $t + 1$.*

Proof. The first point comes from (6).

A *drop step* happens when $\gamma_t = \gamma_{\max}$ in the away-step update L. (11) of Alg. 1. In that case, at least one vertex is removed from the active set. The upper bound on the number of drop step can be proven with the same technique as in (Lacoste-Julien and Jaggi, 2015, Proof of Thm.8). Let us call A_t the number of FW steps (which potentially adds an atom in $\mathcal{S}_t$) and D_t the number of *drop steps*, i.e., the number of *away steps* where at least one atom from $\mathcal{S}_t$ have been removed (and thus $\gamma_t = \gamma_{\max}$ for these). Considering FW-AL with AFW after t iterations we have performed t non drop steps in the inner loop, since it is the condition to end the inner loop, then

$$A_t \leqslant t, \quad \text{and} \quad A_t - D_t + |\mathcal{S}_0| \geqslant |\mathcal{S}_t| \geqslant 0 \,. \tag{31}$$

Since by assumption $|\mathcal{S}_0| = 1$, this leads directly to $D_t \leqslant A_t + 1 \leqslant t + 1$. □

A.2 Other FW Algorithms Available

Any Frank Wolfe algorithm performing *geometric decrease* (15) or *sublinear decrease* (13) can be used an inner loop algorithm. For instance, Block coordinate Frank Wolfe (Lacoste-Julien et al., 2013) performs sublinear decrease and Fully corrective Frank Wolfe (Lacoste-Julien and Jaggi, 2015) or Garber et al. (2016)'s algorithm perform geometric decrease.

A.3 Constants for the sublinear and geometric decrease

In order to be self contained, we will introduce the definitions of the constants introduced in the definition of *sublinear decrease* (13) and *geometric decrease* (15).

Sublinear Decrease. Let us first recall Equation (13) describing sublinear decrease,

$$\mathcal{L}(\boldsymbol{x}^+, \boldsymbol{y}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) \leqslant -\gamma \left(\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) - \min_{\boldsymbol{x}' \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}', \boldsymbol{y}) \right) + \gamma^2 \frac{L_\lambda D^2}{2} \,.$$

Since $\mathcal{L}$ is a L_λ-smooth convex-concave function, this property follows directly from the definition of Lipschitz constant of $\mathcal{L}$. If f is L-smooth we have that the function $\mathcal{L}(\cdot, \boldsymbol{y})$ is $L_\lambda := L + \lambda \|M^\top M\|$-smooth for any $\boldsymbol{y} \in \mathbb{R}^d$, and then,

$$L_\lambda D^2 \leqslant \left(L + \lambda \|M^\top M\|\right) D_{\mathcal{X}}^2 \,. \tag{32}$$

Geometric Decrease. The same way as in the previous paragraph we can say that since $\mathcal{L}(\cdot, \boldsymbol{y})$ is a generalized strongly convex function (with a constant uniform on $\boldsymbol{y}$) and $\mathcal{X}$ a polytope, we have the geometric descent lemma from Lacoste-Julien and Jaggi (2015, Theorem 1). The constant ρ_A is the following

$$\rho_A := \frac{\mu_\lambda}{4 L_\lambda} \left(\frac{\delta_{\mathcal{X}}}{D_{\mathcal{X}}} \right)^2 \,, \tag{33}$$

where μ_λ and L_λ are respectively the generalized strong convexity constant (Lacoste-Julien and Jaggi, 2015, Lemma 9) and the smoothness constant of $\boldsymbol{x} \mapsto f(\boldsymbol{x}) + \frac{\lambda}{2}\|M\boldsymbol{x}\|^2$, and $D_{\mathcal{X}}$ and $\delta_{\mathcal{X}}$ are respectively the diameter and the pyramidal width of $\mathcal{X}$ are defined in (Lacoste-Julien and Jaggi, 2015). Note that if M is full rank the strong convexity constant μ is lower bounded by $\lambda\sigma^2_{min}$ where σ^2_{min} is the smallest singular value of M. Otherwise, if M is not full rank one can still use the lower bound on the generalized strong convexity constant given by Lacoste-Julien and Jaggi (2015, Lemma 9).

B Previous work

B.1 Discussion on previous proofs

The convergence result stated by Yen et al. (2016a, Theorem 2) is the following (with our notation)

$$\Delta_t^{(p)} + \Delta_t^{(d)} \leqslant \frac{\omega}{t} \quad \text{where} \quad \omega := \frac{4}{1-\rho_A} \max\left(\Delta_0^{(p)} + \Delta_0^{(d)}, 2R_Y^2/\lambda\right) , \tag{34}$$

and $R_Y := \sup_{t\geqslant 0} \operatorname{dist}(\boldsymbol{y}_t, \mathcal{Y}^*)$. This quantity was introduced in the last lines of the appendix without any mention to its boundedness. In our opinion, it is as challenging to prove that this quantity is bounded as to prove that Δ_t converges.

In more recent work, Yen et al. (2016b) and Huang et al. (2017) use a different proof technique in order to proove a linear convergence rate for their algorithm. In order to avoid to make appear the same problematic quantity R_Y, they use Lemma 3.1 in (Hong and Luo, 2012) (which also appears as Lemma 3.1 in the published version (Hong and Luo, 2017)). This lemma states a result not holding for all $\boldsymbol{y} \in \mathbb{R}^d$ but instead for $(\boldsymbol{y}_t)_{t\in\mathbb{N}}$, which is the sequence of dual variables computed by the algorithm introduced in (Hong and Luo, 2017). This sequence cannot be assimilated to the sequence of dual variables computed by the GDMM algorithm since the update rule for the primal variables in each algorithm is different, the primal variable are updated with FW steps in one algorithm and with a proximal step in the other. The properties of this proximal step are intrinsically different from the FW steps computing the updates on the primal variables of FW-AL. One way to adapt this Lemma for FW-AL (or GDMM) would be to use (Hong and Luo, 2017, Lemma 2.3 c)). Unfortunately, this is result a local result, only true for all $\boldsymbol{y} \in \mathcal{Y}$ such that $\|\nabla d(\boldsymbol{y})\| \leqslant \delta$ with δ fixed, whereas a global result (true for all δ) is required with the proof technique used in (Yen et al., 2016b; Huang et al., 2017). It is also mentioned in (Hong and Luo, 2017, proof of Lemma 2.3 c)) that "if in addition $\boldsymbol{y}$ also lies in some compact set $\mathcal{Y}$, then the dual error bound hold true for all $\boldsymbol{y} \in \mathcal{Y}$" then showing that R_Y is bounded would fix the issue, but as we mentioned before, we think that this is at least as challenging as showing convergence of Δ_t. To our knowledge there is no easy fix to get a result as the one claimed by Yen et al. (2016b, Lemma 4) or Huang et al. (2017, Lemma 4).

B.2 Comparison with UniPDGrad

The Universal Primal-Dual Gradient Method (UniPDGrad) Yurtsever et al. (2015) is a general method to optimize problem of the form,

$$\min_{\boldsymbol{x}\in\mathcal{X}} \{ f(\boldsymbol{x}) \ : \ A\boldsymbol{x} - \boldsymbol{b} \in \mathcal{K} \} \tag{35}$$

where f is a convex function A is a matrix, $\boldsymbol{b}$ a vector and $\mathcal{X}$ and $\mathcal{K}$ two closed convex sets.

They derive their algorithm optimizing the Lagrange dual function. Starting from (OPT) and introducing the slack variable $\boldsymbol{r}$ with the constraint $\boldsymbol{x} = \boldsymbol{r}$ we get the Lagrange function,

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{r}, \boldsymbol{y}, \boldsymbol{\lambda}) := f(\boldsymbol{r}) - \langle \boldsymbol{\lambda}, \boldsymbol{r} - \boldsymbol{x} \rangle + \langle \boldsymbol{y}, M\boldsymbol{x} \rangle \tag{36}$$

where $\boldsymbol{\lambda}$ is the dual variable associated with the constrain $\boldsymbol{r} = \boldsymbol{x}$. Then, the (negative) Lagrange dual function is,

$$g(\boldsymbol{\lambda}, \boldsymbol{y}) = -\min_{\boldsymbol{x}\in\mathcal{X}, \boldsymbol{r}\in\mathbb{R}^p} f(\boldsymbol{r}) - \langle \boldsymbol{\lambda}, \boldsymbol{r} - \boldsymbol{x} \rangle + \langle \boldsymbol{y}, M\boldsymbol{x} \rangle = -\min_{\boldsymbol{r}\in\mathbb{R}^p} f(\boldsymbol{r}) - \langle \boldsymbol{\lambda}, \boldsymbol{r} \rangle - \min_{\boldsymbol{x}\in\mathcal{X}} \left\langle M^\top \boldsymbol{y} + \boldsymbol{\lambda}, \boldsymbol{x} \right\rangle \tag{37}$$

where $p := d_1 + \ldots + d_K$. Computing the subgradients of the function g require to compute the Fenchel conjugate of f and a LMO. Note that our algorithm does not require the efficient computation of the Fenchel conjugate. In terms of rate, the accelerated version of their algorithm has an optimal $O(1/t^2)$ rate but does not cover the geometric rate when the constraint set $\mathcal{X}$ is a polytope.

C Technical results on the Augmented Lagrangian formulation

Let us recall that the Augmented Lagrangian function is defined as

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) := f(\boldsymbol{x}) + \mathbf{1}_{\mathcal{X}}(\boldsymbol{x}) + \langle \boldsymbol{y}, M\boldsymbol{x} \rangle + \tfrac{\lambda}{2}\|M\boldsymbol{x}\|^2 \,, \quad \forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^m \times \mathbb{R}^d \,, \tag{38}$$

where f is an L-smooth function, $\mathbf{1}_{\mathcal{X}}$ is the indicator function over the convex compact set $\mathcal{X} := \mathcal{X}_1 \times \ldots \times \mathcal{X}_K \subset \mathbb{R}^m$, M is the matrix defined in (1), and $m = d_1 + \ldots + d_K$. The augmented dual function d is $d(\boldsymbol{y}) := \max_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$. Strong duality ensures that $\mathcal{X}^* \times \mathcal{Y}^*$ is the set of saddle points of $\mathcal{L}$ where $\mathcal{X}^*$ is the optimal set of $p(\cdot) := \max_{\boldsymbol{y} \in \mathbb{R}^d} \mathcal{L}(\cdot, \boldsymbol{y})$ and $\mathcal{Y}^*$ is the optimal set of d. In this section we will first prove that the augmented dual function is smooth and have a property similar to strong convexity around its optimal set. It will be useful for subsequent analyses to detail the properties of the augmented Lagrangian function $\mathcal{L}$.

C.1 Properties of the dual function $d(\cdot)$

The dual function $d(\cdot)$ can be written as the composition of a linear transformation and the Fenchel conjugate of $f_\lambda(\boldsymbol{x}) := f(\boldsymbol{x}) + \frac{\lambda}{2}\|M\boldsymbol{x}\|^2 + \mathbf{1}_{\mathcal{X}}(\boldsymbol{x})$ where $\mathbf{1}_{\mathcal{X}}$ is the indicator function of $\mathcal{X}$. More precisely, if we denote by $^\star : f \mapsto f^*$ the Fenchel conjugate operator, then we have,

$$d(\boldsymbol{y}) := \min_{\boldsymbol{x} \in \mathbb{R}^m} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = - \max_{\boldsymbol{x} \in \mathbb{R}^m} \left\langle -M^\top \boldsymbol{y}, \boldsymbol{x} \right\rangle - f_\lambda(\boldsymbol{x}) = -f_\lambda^\star(-M^\top \boldsymbol{y}) \,. \tag{39}$$

Smoothness of the augmented dual function. The smoothness of the augmented dual function is due to the duality between strong convexity and strong smoothness (Kakade et al., 2009). In order to be self contended we provide the proof of this property given by Hong and Luo (2017).

Proposition 2 (Lemma 2.2 (Hong and Luo, 2017))**.** *If f is convex, the dual function d is $1/\lambda$-smooth, i.e.,*

$$\nabla d(\boldsymbol{y}) = M\hat{\boldsymbol{x}}(\boldsymbol{y}), \quad \textit{where} \quad \hat{\boldsymbol{x}}(\boldsymbol{y}) \in \arg\min_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) \,, \ \forall \boldsymbol{y} \in \mathbb{R}^d \,, \tag{40}$$

and

$$\|\nabla d(\boldsymbol{y}) - \nabla d(\boldsymbol{y}')\| \leqslant \frac{1}{\lambda}\|\boldsymbol{y} - \boldsymbol{y}'\| \quad \forall \, \boldsymbol{y}, \boldsymbol{y}' \in \mathbb{R}^d \,. \tag{41}$$

Proof. We will start by showing that the quantity $M\hat{\boldsymbol{x}}(\boldsymbol{y})$ has the same value for all $\hat{\boldsymbol{x}}(\boldsymbol{y}) \in \arg\min_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$. We reason by contradiction and assume there exists $\boldsymbol{x}, \boldsymbol{x}' \in \arg\min_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$ such that $M\boldsymbol{x} \neq M\boldsymbol{x}'$. Then by convexity of f and strong convexity of $\|\cdot\|^2$ we have that

$$d(\boldsymbol{y}) = \frac{1}{2}\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) + \frac{1}{2}\mathcal{L}(\boldsymbol{x}', \boldsymbol{y}) > f(\bar{\boldsymbol{x}}) + \langle \boldsymbol{y}, M\bar{\boldsymbol{x}} \rangle + \frac{\lambda}{2}\|M\bar{\boldsymbol{x}}\|^2 = \mathcal{L}(\bar{\boldsymbol{x}}, \boldsymbol{y}) \,, \tag{42}$$

where $\bar{\boldsymbol{x}} := \frac{\boldsymbol{x} + \boldsymbol{x}'}{2}$ and the inequality is strict because we assumed $M\boldsymbol{x} \neq M\boldsymbol{x}'$. This contradict the assumption that $\boldsymbol{x}, \boldsymbol{x}' \in \arg\min_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$. To conclude, Danskin (1967)'s Theorem claims that $\partial d(\boldsymbol{y}) = \{M\hat{\boldsymbol{x}}(\boldsymbol{y}), \, | \, \hat{\boldsymbol{x}}(\boldsymbol{y}) \in \arg\min_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})\}$ which is a singleton in that case. The function d is then differentiable.

For the second part of the proof, let $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$ and let $\boldsymbol{x}, \boldsymbol{x}'$ be two respective minimizers of $\mathcal{L}(\cdot, \boldsymbol{y})$ and $\mathcal{L}(\cdot, \boldsymbol{y}')$. Then by the first order optimality conditions we have

$$\langle \nabla f(\boldsymbol{x}) + M^\top \boldsymbol{y} + \lambda M^\top M\boldsymbol{x}, \boldsymbol{x}' - \boldsymbol{x} \rangle \geqslant 0, \ \ \langle \nabla f(\boldsymbol{x}') + M^\top \boldsymbol{y}' + \lambda M^\top M\boldsymbol{x}', \boldsymbol{x} - \boldsymbol{x}' \rangle \geqslant 0 \,. \tag{43}$$

Adding these two equation gives,

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}') + M^\top(\boldsymbol{y} - \boldsymbol{y}') + \lambda M^\top M(\boldsymbol{x} - \boldsymbol{x}'), \boldsymbol{x}' - \boldsymbol{x} \rangle \geqslant 0 \,, \tag{44}$$

but since f is convex, $\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}' \rangle \geqslant 0$, and so

$$\left\langle \boldsymbol{y} - \boldsymbol{y}', M(\boldsymbol{x}' - \boldsymbol{x}) \right\rangle \geqslant -\lambda \left\langle M(\boldsymbol{x} - \boldsymbol{x}'), M(\boldsymbol{x}' - \boldsymbol{x}) \right\rangle \,. \tag{45}$$

Finally, by the Cauchy-Schwarz inequality we have

$$\|\boldsymbol{y} - \boldsymbol{y}'\| \geqslant \lambda \|M\boldsymbol{x} - M\boldsymbol{x}'\| = \lambda \|\nabla d(\boldsymbol{y}) - \nabla d(\boldsymbol{y}')\| \,. \tag{46}$$

□

Error bound on the augmented dual function. After having proved that the dual function is smooth we will prove that the augmented dual function has a property similar to the Polyak-Łojasiewic (PL) condition first introduced by Polyak (1963) and the same year in a more general setting by Łojasiewicz (1963). Recently, convergence under this condition has been studied with a machine learning perspective by Karimi et al. (2016). The PL condition is particular case of what is called *error bounds* that have been widely studied in optimization literature see for instance (Pang, 1997, 1987).

We start our proof with some dual computation. These are similar to the dualization of strong convexity (see for instance (Rockafellar and Wets, 1998)). Let us consider $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^p$, $\boldsymbol{n} \in N_c^{\mathcal{X}}(\boldsymbol{x})$ and the function

$$g_{\boldsymbol{x}}(\boldsymbol{u}) = f_\lambda(\boldsymbol{u}+\boldsymbol{x}) - f_\lambda(\boldsymbol{x}) - \langle \boldsymbol{u}, \nabla f(\boldsymbol{x}) + \boldsymbol{n}\rangle \ , \forall \boldsymbol{u} \in \mathbb{R}^p \ . \tag{47}$$

Note that $\{\nabla f(\boldsymbol{x}) + \boldsymbol{n} \ ; \ \boldsymbol{n} \in N_c^{\mathcal{X}}(\boldsymbol{x})\}$ is equal to $\partial f_\lambda(\boldsymbol{x})$, the set of subgradients of f_λ at $\boldsymbol{x}$. This function is similar to the function g introduced by Kakade et al. (2009, App. A.1). Since $f + \frac{\lambda}{2}\|M\cdot\|^2$ is L_λ-Lipschitz, we have that $g_{\boldsymbol{x}}(\boldsymbol{u}) \leqslant \frac{L_\lambda}{2}\|\boldsymbol{u}\|^2 + \mathbf{1}_{\mathcal{X}}(\boldsymbol{u}+\boldsymbol{x}) =: h_{\boldsymbol{x}}(\boldsymbol{u}),\ \forall \boldsymbol{u} \in \mathbb{R}^m$. By (Shalev-Shwartz and Singer, 2010, Lemma 19) we know that

$$g_{\boldsymbol{x}}(\boldsymbol{u}) \leqslant h_{\boldsymbol{x}}(\boldsymbol{u}),\ \forall \boldsymbol{u} \in \mathbb{R}^m \Rightarrow g^\star_{\boldsymbol{x}}(\boldsymbol{v}) \geqslant h^\star_{\boldsymbol{x}}(\boldsymbol{v}) <,\ \forall \boldsymbol{v} \in \mathbb{R}^m \, . \tag{48}$$

Dual computations give us for all $\boldsymbol{v}$,

$$\begin{aligned} g^\star_{\boldsymbol{x}}(\boldsymbol{v}) &= \max_{\boldsymbol{u}\in\mathbb{R}^m} \left[\langle \boldsymbol{u}, \boldsymbol{v}\rangle - f_\lambda(\boldsymbol{u}+\boldsymbol{x}) - \langle \boldsymbol{u}, \nabla f(\boldsymbol{x}) + \boldsymbol{n}\rangle\right] + f_\lambda(\boldsymbol{x}) \\ &= \max_{\boldsymbol{u}\in\mathbb{R}^m} \left[\langle \boldsymbol{u}, \boldsymbol{v} - \nabla f(\boldsymbol{x}) - \boldsymbol{n}\rangle - f_\lambda(\boldsymbol{u}+\boldsymbol{x})\right] + f_\lambda(\boldsymbol{x}) \\ &= f^\star_\lambda(\boldsymbol{v} + \nabla f(\boldsymbol{x}) + \boldsymbol{n}) + f_\lambda(\boldsymbol{x}) + \langle \boldsymbol{x}, \boldsymbol{v} + \nabla f_\lambda(\boldsymbol{x}) + \boldsymbol{n}\rangle \\ &= f^\star_\lambda(\boldsymbol{v} + \nabla f(\boldsymbol{x}) + \boldsymbol{n}) - f^*_\lambda(\nabla f(\boldsymbol{x}) + \boldsymbol{n}) - \langle \boldsymbol{x}, \boldsymbol{v}\rangle \ , \end{aligned} \tag{49}$$

where in he last line we used that $\nabla f(\boldsymbol{x}) + \boldsymbol{n} \in \partial f_\lambda(\boldsymbol{x})$ and that $\forall \boldsymbol{d} \in \partial f_\lambda(\boldsymbol{x})$, $\langle \boldsymbol{x}, \boldsymbol{d}\rangle = f_\lambda(\boldsymbol{x}) + f^*_\lambda(\boldsymbol{d})$ (Shalev-Shwartz and Singer, 2010, Lemma 17).

By strong duality we have that $\mathcal{X}^* \times \mathcal{Y}^*$ is the set of saddle points, where $\mathcal{X}^*$ and $\mathcal{Y}^*$ are respectively the optimal sets of $f(\cdot)$ and $d(\cdot)$, respectively introduced in (38) and (39). In the following we will fix a pair $(\boldsymbol{x}^*, \boldsymbol{y}^*) \in \mathcal{X}^* \times \mathcal{Y}^*$. Then by the stationary conditions we have

$$\nabla f(\boldsymbol{x}^*) + M^\top \boldsymbol{y}^* \in -N_c(\boldsymbol{x}^*), \quad \text{and} \quad M\boldsymbol{x}^* = 0 \, . \tag{50}$$

Equivalently, there exist $\boldsymbol{u} \in N_c^{\mathcal{X}}(\boldsymbol{x}^*)$ such that

$$\nabla f(\boldsymbol{x}^*) + \boldsymbol{u} = -M^\top \boldsymbol{y}^* \, . \tag{51}$$

For all $\boldsymbol{y}^* \in \mathcal{Y}^*$ we can set $\boldsymbol{x} = \boldsymbol{x}^*$ and $\boldsymbol{n} = \boldsymbol{u} \in N_c^{\mathcal{X}}(\boldsymbol{x}^*)$ in (47) to get the following inequality,

$$\begin{aligned} d^* - d(\boldsymbol{v} + \boldsymbol{y}^*) &= f^\star_\lambda(-M^\top \boldsymbol{v} - M^\top \boldsymbol{y}^*) - f^\star_\lambda(-M^\top \boldsymbol{y}^*) \\ &\overset{(51)}{=} f^\star_\lambda(-M^\top \boldsymbol{v} + \nabla f(\boldsymbol{x}^*) + \boldsymbol{n}) - f^\star_\lambda(\nabla f(\boldsymbol{x}^*) + \boldsymbol{n}) \\ &\overset{(50)}{=} f^\star_\lambda(-M^\top \boldsymbol{v} + \nabla f(\boldsymbol{x}^*) + \boldsymbol{n}) - f^\star_\lambda(\nabla f(\boldsymbol{x}^*) + \boldsymbol{n}) - \left\langle \boldsymbol{x}^*, -M^\top \boldsymbol{v}\right\rangle \\ &\overset{(49)}{=} g^\star(-M^\top \boldsymbol{v}) \\ &\overset{(48)}{\geqslant} h^\star_{\boldsymbol{x}^*}(-M^\top \boldsymbol{v}) \ , \quad \forall \boldsymbol{v} \in \mathbb{R}^d \, , \end{aligned} \tag{52}$$

where for all $\boldsymbol{v} \in \mathbb{R}^d$,

$$h^*_{\boldsymbol{x}^*}(-M^\top \boldsymbol{v}) := \max_{\boldsymbol{x}\in\mathbb{R}^m} \left[\left\langle \boldsymbol{x}, -M^\top \boldsymbol{v}\right\rangle - h_{\boldsymbol{x}^*}(\boldsymbol{x})\right] \tag{53}$$

$$= \max_{\boldsymbol{x}\in\mathbb{R}^m} \left[\left\langle \boldsymbol{x}, -M^\top \boldsymbol{v}\right\rangle - \frac{L_\lambda}{2}\|\boldsymbol{x}\|^2 - \mathbf{1}_{\mathcal{X}}(\boldsymbol{x} + \boldsymbol{x}^*)\right] \tag{54}$$

$$= \max_{\boldsymbol{x}+\boldsymbol{x}^*\in\mathcal{X}} \left[\left\langle \boldsymbol{x}, -M^\top \boldsymbol{v}\right\rangle - \frac{L_\lambda}{2}\|\boldsymbol{x}\|^2\right] \tag{55}$$

$$= \frac{1}{2L_\lambda}\left[\|M^\top \boldsymbol{v}\|^2 - \|M^\top \boldsymbol{v} + L_\lambda(P_{\mathcal{X}}(\boldsymbol{x}^* - M^\top \boldsymbol{v}/L_\lambda) - \boldsymbol{x}^*)\|^2\right] . \tag{56}$$

We noted $P_{\mathcal{X}}$ the projection onto the convex set $\mathcal{X}$. Let us choose $\boldsymbol{y} \in \mathbb{R}^d$ and set $\boldsymbol{v} = \boldsymbol{y} - \boldsymbol{y}^*$, where $\boldsymbol{y}^* = P_{\mathcal{Y}^*}(\boldsymbol{y})$. Then combining (52) and (55) we get for all $\boldsymbol{x} \in \mathcal{X}$, and $\gamma \in [0,1]$ that $\gamma \boldsymbol{x} + (1-\gamma)\boldsymbol{x}^* \in \mathcal{X}$ and then,

$$d^* - d(\boldsymbol{y}) \geqslant \left[-\gamma \left\langle M^\top (\boldsymbol{y} - \boldsymbol{y}^*), \boldsymbol{x} - \boldsymbol{x}^* \right\rangle - \frac{L_\lambda}{2} \gamma^2 \|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \right] \tag{57}$$

$$\geqslant \frac{1}{2} \left[2\gamma \langle \boldsymbol{y} - \boldsymbol{y}^*, -M\boldsymbol{x} \rangle - \gamma^2 L_\lambda D^2 \right] , \tag{58}$$

where $D := \max_{(\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{X}^2} \|\boldsymbol{x} - \boldsymbol{x}'\|$ is the diameter of $\mathcal{X}$. Since $d^* \geqslant d(\boldsymbol{y})$ the last equation can give a non trivial lower bound when $\max_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{y} - \boldsymbol{y}^*, -M\boldsymbol{x} \rangle > 0$, we will now prove that is it always the case when $\boldsymbol{y} \notin \mathcal{Y}^*$.

If $\boldsymbol{y} \notin \mathcal{Y}^*$, then the necessary and sufficient stationary conditions lead to

$$\nabla f(\boldsymbol{x}^*) + M^\top \boldsymbol{y} \notin -N_c(\boldsymbol{x}^*) , \tag{59}$$

that is, there exist $\boldsymbol{x} \in \mathcal{X}$ such that $\left\langle \nabla f(\boldsymbol{x}^*) + M^\top \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{x}^* \right\rangle < 0$. Using (51) gives

$$\begin{aligned} 0 &> \left\langle \nabla f(\boldsymbol{x}^*) + M^\top \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{x}^* \right\rangle \\ &\overset{(51)}{=} \left\langle -M^\top \boldsymbol{y}^* - \boldsymbol{u} + M^\top \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{x}^* \right\rangle \\ &\geqslant \langle \boldsymbol{y} - \boldsymbol{y}^*, M\boldsymbol{x} \rangle , \end{aligned} \tag{60}$$

where for the last inequality we use the fact that $\boldsymbol{u} \in N_c(\boldsymbol{x}^*)$ and $M\boldsymbol{x}^* = 0$. Then we have

$$\max_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{y} - \boldsymbol{y}^*, -M\boldsymbol{x} \rangle > 0, \quad \forall \boldsymbol{y} \notin \mathcal{Y}^* . \tag{61}$$

Optimizing Eq. (58) with respect to $\gamma \in [0,1]$ we get the following:

- If $0 < \max_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{y}^* - \boldsymbol{y}, M\boldsymbol{x} \rangle \leqslant L_\lambda D^2$, the optimum of (58) is achieved for $\gamma = \frac{\max_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{y}^* - \boldsymbol{y}, M\boldsymbol{x} \rangle}{L_\lambda D^2} \leqslant 1$ and we have,
$$d^* - d(\boldsymbol{y}) \geqslant \frac{1}{2L_\lambda D^2} \max_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{y}^* - \boldsymbol{y}, M\boldsymbol{x} \rangle^2 , \tag{62}$$
- Otherwise, if $\max_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{y}^* - \boldsymbol{y}, M\boldsymbol{x} \rangle > L_\lambda D^2$, the optimum of (58) is achieved for $\gamma = 1$, giving
$$d^* - d(\boldsymbol{y}) \geqslant \frac{1}{2} \max_{\boldsymbol{x} \in \mathcal{X}} \left[2 \langle \boldsymbol{y}^* - \boldsymbol{y}, M\boldsymbol{x} \rangle - L_\lambda D^2 \right] \geqslant \frac{1}{2} \max_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{y}^* - \boldsymbol{y}, M\boldsymbol{x} \rangle . \tag{63}$$

Combining both cases leads to

$$d^* - d(\boldsymbol{y}) \geqslant \frac{1}{2L_\lambda D^2} \min \left(\max_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{y}^* - \boldsymbol{y}, M\boldsymbol{x} \rangle^2 , L_\lambda D^2 \max_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{y}^* - \boldsymbol{y}, M\boldsymbol{x} \rangle \right) \tag{64}$$

Since our goal is to get an error bound on the dual function d we divide and multiply by $\|\boldsymbol{y} - \boldsymbol{y}^*\|$ making appear the desired norm and a constant α defined as

$$\alpha := \inf_{\substack{\boldsymbol{y} \in \mathbb{R}^d \setminus \mathcal{Y}^* \\ \boldsymbol{y}^* = P_{\mathcal{Y}^*}(\boldsymbol{y})}} \sup_{\boldsymbol{x} \in \mathcal{X}} \left\langle \frac{\boldsymbol{y}^* - \boldsymbol{y}}{\|\boldsymbol{y}^* - \boldsymbol{y}\|}, M\boldsymbol{x} \right\rangle . \tag{65}$$

Recall that $\boldsymbol{y}^* := P_{\mathcal{Y}^*}(\boldsymbol{y})$ and consequently $\|\boldsymbol{y} - \boldsymbol{y}^*\| = \operatorname{dist}(\boldsymbol{y}, \mathcal{Y}^*)$. Our goal is now to show that $\alpha > 0$.

Proof that α is positive. In order to prove that α is positive we need to get results on the structure of $\mathcal{Y}^*$. Let us recall the full assumptions needed for our lemma,

Assumption' 1. *There exists $\boldsymbol{x} \in \operatorname{relint}(\mathcal{X}^{feas})$, i.e., $\exists\, \bar{\boldsymbol{x}}^{(k)} \in \operatorname{relint}(\mathcal{X}_k)$, $k \in \{1, \dots, K\}$, s.t., $\sum_{k=0}^K A_k \boldsymbol{x}^{(k)} = 0$.*

Lemma 1. *Under Assumption 1, the optimal set $\mathcal{Y}^*$ of the augmented dual function $d(\cdot)$ (39) can be written as*

$$\mathcal{Y}^* = \{ \boldsymbol{\kappa} + \boldsymbol{v} \ : \ \boldsymbol{\kappa} \in \mathcal{K} , \ \boldsymbol{v} \in V \} , \tag{66}$$

where $V := \cap_{k=1}^K \left(A_k (Span(\mathcal{X}_k - \bar{\boldsymbol{x}}^{(k)})) \right)^\perp$ and $\mathcal{K} \subset V^\perp$ is a compact set.

We define $\mathrm{Span}(\mathcal{X}_k - \bar{\boldsymbol{x}}^{(k)})$ is the linear span of the feasible direction from $\bar{\boldsymbol{x}}^{(k)}$, since $\bar{\boldsymbol{x}}^{(k)}$ is an interior point we have $\mathrm{Span}(\mathcal{X}_k - \bar{\boldsymbol{x}}^{(k)}) = \{\lambda(\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}^{(k)}) \ : \ \boldsymbol{x}^{(k)} \in \mathcal{X}_k,\ \lambda > 0\}$.

Proof. For any $\boldsymbol{x}^* \in \mathcal{X}^*$, a necessary and sufficient condition for any $\boldsymbol{y}^*$ to be in $\mathcal{Y}^*$ is

$$\nabla f(\boldsymbol{x}^*) + M^\top \boldsymbol{y}^* \in -N_c(\boldsymbol{x}^*)\,, \tag{67}$$

meaning that

$$-A_k^\top \boldsymbol{y}^* \in N_c^{\mathcal{X}_k}(\boldsymbol{x}^*) + \nabla_{\boldsymbol{x}^{(k)}} f(\boldsymbol{x}^*)\,, \quad k \in \{1,\dots,K\}\,. \tag{68}$$

Then noting $g_k := \nabla_{\boldsymbol{x}^{(k)}} f(\boldsymbol{x}^*)$ we have the following equivalences,

$$\begin{aligned}
\boldsymbol{y}^* \in \mathcal{Y}^* &\Leftrightarrow -A_k^\top \boldsymbol{y}^* \in N_c^{\mathcal{X}_k}(\boldsymbol{x}^*) + g_k\,, \quad k \in \{1,\dots,K\} \\
&\Leftrightarrow A_k^\top \boldsymbol{y}^* + g_k \in -N_c^{\mathcal{X}_k}(\boldsymbol{x}^*)\,, \quad k \in \{1,\dots,K\} \\
&\Leftrightarrow \left\langle -A_k^\top \boldsymbol{y}^* - g_k, \boldsymbol{x}^{(k)} - (\boldsymbol{x}^*)^{(k)} \right\rangle \leqslant 0\ ;\ \forall \boldsymbol{x}^{(k)} \in \mathcal{X}_k\,, \quad k \in \{1,\dots,K\} \\
&\Leftrightarrow \left\langle -\boldsymbol{y}^*, A_k(\boldsymbol{x}^{(k)} - (\boldsymbol{x}^*)^{(k)}) \right\rangle \leqslant \left\langle g_k, \boldsymbol{x}^{(k)} - (\boldsymbol{x}^*)^{(k)} \right\rangle\ ;\ \forall \boldsymbol{x}^{(k)} \in \mathcal{X}_k\,, \quad k \in \{1,\dots,K\}
\end{aligned}$$

Then we can notice that if we write $\boldsymbol{y}^* = \boldsymbol{y}_1^* + \boldsymbol{y}_2^*$ with $\boldsymbol{y}_1^* \in V := \cap_{k=1}^K \big(A_k(\mathrm{Span}(\mathcal{X}_k - \bar{\boldsymbol{x}}^{(k)})\big)^\perp$ and $\boldsymbol{y}_2^* \in V^\perp$ we get,

$$\boldsymbol{y}^* \in \mathcal{Y}^* \Leftrightarrow \left\langle -\boldsymbol{y}_2^*, A_k(\boldsymbol{x}^{(k)} - (\boldsymbol{x}^*)^{(k)}) \right\rangle \leqslant \left\langle g_k, \boldsymbol{x}^{(k)} - (\boldsymbol{x}^*)^{(k)} \right\rangle\ ;\ \forall \boldsymbol{x}^{(k)} \in \mathcal{X}_k\,, \quad k \in [K]\,. \tag{69}$$

There is then no conditions on $\boldsymbol{y}_1^*$.

Let us get a necessary condition on $\boldsymbol{y}_2^*$. (69) implies,

$$\begin{aligned}
\boldsymbol{y}^* \in \mathcal{Y}^* &\Rightarrow \left\langle -\boldsymbol{y}_2^*, \sum_{k=1}^K A_k(\boldsymbol{x}^{(k)} - (\boldsymbol{x}^*)^{(k)}) \right\rangle \leqslant \sum_{k=1}^K \left\langle g_k, \boldsymbol{x}^{(k)} - (\boldsymbol{x}^*)^{(k)} \right\rangle\ ;\ \forall \boldsymbol{x}^{(k)} \in \mathcal{X}_k\,, \quad k \in [K] \\
&\Rightarrow \left\langle -\boldsymbol{y}_2^*, \sum_{k=1}^K A_k(\boldsymbol{x}^{(k)} - (\boldsymbol{x}^*)^{(k)}) \right\rangle \leqslant \sum_{k=1}^K \|g_k\| \|\boldsymbol{x}^{(k)} - (\boldsymbol{x}^*)^{(k)}\|\ ;\ \forall \boldsymbol{x}^{(k)} \in \mathcal{X}_k\,, \quad k \in [K] \\
&\Rightarrow \left\langle -\boldsymbol{y}_2^*, \sum_{k=1}^K A_k(\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}^{(k)}) \right\rangle \leqslant \sum_{k=1}^K \|g_k\| \operatorname{diam}(X_k) \quad ; \quad \forall \boldsymbol{x}^{(k)} \in \mathcal{X}_k\,, \quad k \in [K]\,, \\
&\qquad (\boldsymbol{x}^* \text{ and } \bar{\boldsymbol{x}} \text{ are feasible, i.e., } \sum_{k=1}^K A_k \bar{\boldsymbol{x}}^{(k)} = \sum_{k=1}^K A_k (\boldsymbol{x}^*)^{(k)} = 0)
\end{aligned}$$

where $\bar{\boldsymbol{x}} \in \mathrm{relint}(\mathcal{X}^{feas})$ (hypothesis of the theorem). Moreover, since $V := \cap_{k=1}^K \big(A_k(\mathrm{Span}(\mathcal{X}_k - \bar{\boldsymbol{x}}^{(k)}))\big)^\perp$ we have that $V^\perp = \sum_{k=1}^K A_k(\mathrm{Span}(\mathcal{X}_k - \bar{\boldsymbol{x}}^{(k)}))$. We can consequently write $\boldsymbol{y}_2^* = \sum_{k=1}^K \boldsymbol{y}_{2,k}^*$ where $\boldsymbol{y}_{2,k}^* \in A_k(\mathrm{Span}(\mathcal{X}_k - \bar{\boldsymbol{x}}^{(k)}))$. Then, since $\bar{\boldsymbol{x}} \in \mathrm{relint}(\mathcal{X}^{feas})$ there exists $\delta > 0$ such that for all $\boldsymbol{y}_{2,k}^*$ we can set $\boldsymbol{x}^{(k)} \in \mathcal{X}_k$ such that $A_k(\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}^{(k)}) = -\delta \boldsymbol{y}_{2,k}^*$. Finally, we get that,

$$\boldsymbol{y}^* \in \mathcal{Y}^* \Rightarrow \delta \left\langle \boldsymbol{y}_2^*, \sum_{k=1}^K \boldsymbol{y}_{2,k}^* \right\rangle \leqslant \sum_{k=1}^K \|g_k\| \operatorname{diam}(X_k) \Rightarrow \|\boldsymbol{y}_2^*\|_2^2 \leqslant \sum_{k=1}^K \frac{\|g_k\| \operatorname{diam}(X_k)}{\delta}\,. \tag{70}$$

Proposition 3. *If Assumption 1 holds, then the set of normal directions to $\mathcal{Y}^*$,*

$$\mathcal{D} := \left\{ \boldsymbol{d}\ ;\ \boldsymbol{d} \in N_c^{\mathcal{Y}^*}(\boldsymbol{y}^*) \text{ for } \boldsymbol{y}^* \in \mathcal{Y}^*\,,\ \|\boldsymbol{d}\| = 1 \right\}\,, \tag{71}$$

is closed and consequently compact.

Proof. Let us first show that $\mathcal{D} = \{\boldsymbol{y} - P_{\mathcal{Y}^*}(\boldsymbol{y})\ ;\ \boldsymbol{y} \in \mathbb{R}^d \backslash \mathcal{Y}^*, ; \|\boldsymbol{y} - P_{\mathcal{Y}^*}(\boldsymbol{y})\| = 1\}$. Let $\boldsymbol{y} \in \mathbb{R}^d \backslash \mathcal{Y}^*$ by definition of the normal cone and the projection onto a convex set we have that $\boldsymbol{y} - P_{\mathcal{Y}^*}(\boldsymbol{y}) \in N_c^{\mathcal{Y}^*}(P_{\mathcal{Y}^*}(\boldsymbol{y}))$. Reciprocally, for any $\boldsymbol{y}^* \in \mathcal{Y}^*$ and $\boldsymbol{d} \in N_c^{\mathcal{Y}^*}(\boldsymbol{y}^*)$ such that $\|\boldsymbol{d}\| = 1$ we have that $\boldsymbol{y}^* = P_{\mathcal{Y}^*}(\boldsymbol{y}^* + \boldsymbol{d})$ and $\boldsymbol{y}^* + \boldsymbol{d} \notin \mathcal{Y}^*$.

With the same notation as Lemma 1 we can write $\boldsymbol{y} \in \mathbb{R}^d$ a unique way as $\boldsymbol{y} = \boldsymbol{y}_1 + \boldsymbol{y}_2$ where $\boldsymbol{y}_1 \in V$ and $\boldsymbol{y}_2 \in V^\perp$. Then since $\mathcal{Y}^* = V \oplus \mathcal{K}$ we get that $P_{\mathcal{Y}^*}(\boldsymbol{y}) = \boldsymbol{y}_1 + \boldsymbol{\kappa}$ where $\boldsymbol{\kappa} \in \mathcal{K}$. Then $\boldsymbol{y} - P_{\mathcal{Y}^*}(\boldsymbol{y}) = \boldsymbol{y}_2 - \boldsymbol{\kappa}$ where $P_{\mathcal{K}}(\boldsymbol{y}_2) = \boldsymbol{\kappa}$. Reciprocally, for any couple $(\boldsymbol{y}_2, \kappa) \in V^\perp \times \mathcal{K}$ such that $P_{\mathcal{K}}(\boldsymbol{y}_2) = \boldsymbol{\kappa}$ we have that $\boldsymbol{y}_2 - \boldsymbol{\kappa} \in N_c^{\mathcal{K}}(\boldsymbol{\kappa})$.

If we call $\phi : \boldsymbol{y} \mapsto \boldsymbol{y} - P_{\mathcal{K}}(\boldsymbol{y})$, then $\mathcal{D} = \phi(A)$ where $A = \{\boldsymbol{y}_2 \in V^\perp \; ; \; \operatorname{dist}(\boldsymbol{y}_2, \mathcal{K}) = 1\}$ is a compact (because $\mathcal{K}$ is compact). Then since ϕ is continuous, $\mathcal{D}$ is a compact. □

Now we can apply this result to bound the α constant introduced in Eq. (65). We notice that we can write that definition as

$$\alpha = \inf_{\substack{\boldsymbol{y} \in \mathbb{R}^d \backslash \mathcal{Y}^* \\ \boldsymbol{y}^* = P_{\mathcal{Y}^*}(\boldsymbol{y}) \\ \boldsymbol{d} = \boldsymbol{y}^* - \boldsymbol{y}, \, \|\boldsymbol{d}\| = 1}} \sup_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{d}, M\boldsymbol{x} \rangle \; . \tag{72}$$

The function $\boldsymbol{d} \mapsto \sup_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{d}, M\boldsymbol{x} \rangle$ is convex (as a supremum of convex function) and then is continuous on the interior of its domain which is $\mathbb{R}^d$ because the supremum is achieved for all $\boldsymbol{d} \in \mathbb{R}^d$. With a similar argument of continuity and compactness as in previous lemma there exist $(\boldsymbol{y}, \boldsymbol{y}^*) \in (\mathbb{R}^d \backslash \mathcal{Y}^*) \times \mathcal{Y}^*$ such that $\|\boldsymbol{y}^* - \boldsymbol{y}\| = 1$ and $\alpha = \max_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{y}^* - \boldsymbol{y}, M\boldsymbol{x} \rangle$. Equation (61) claiming that for all non optimal point $\boldsymbol{y} \in \mathbb{R}^d \backslash \mathcal{Y}^*$, we have

$$\exists \, (\boldsymbol{y}, \boldsymbol{y}^*) \in (\mathbb{R}^d \backslash \mathcal{Y}^*) \times \mathcal{Y}^* \text{ s.t. } \alpha = \max_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{y}^* - \boldsymbol{y}, M\boldsymbol{x} \rangle > 0. \tag{73}$$

Proof of Thm.1 and a Corollary.

Theorem' 1. *Let d be the augmented dual function (39), if f is a smooth convex function and $\mathcal{X}$ a compact convex set and if Assumption 1 holds, then for all $\boldsymbol{y} \in \mathbb{R}^d$ there exist a constant $\alpha > 0$ such that,*

$$d^* - d(\boldsymbol{y}) \geqslant \frac{1}{2L_\lambda D^2} \min \left\{ \alpha^2 \operatorname{dist}(\boldsymbol{y}, \mathcal{Y}^*)^2, \alpha L_\lambda D^2 \operatorname{dist}(\boldsymbol{y}, \mathcal{Y}^*) \right\} \; , \tag{74}$$

where $D := \max_{(\boldsymbol{x}, \boldsymbol{x}') = \in \mathcal{X}^2}$ is the diameter of $\mathcal{X}$.

Proof. Recall that we proved

$$d^* - d(\boldsymbol{y}) \geqslant \frac{1}{2L_\lambda D^2} \min \left(\max_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{y}^* - \boldsymbol{y}, M\boldsymbol{x} \rangle^2 , L_\lambda D^2 \max_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{y}^* - \boldsymbol{y}, M\boldsymbol{x} \rangle \right) \; , \tag{75}$$

and that α defined in (65) was positive (73). Then for all $\boldsymbol{y} \notin \mathcal{Y}^*$,

$$d^* - d(\boldsymbol{y}) \geqslant \frac{1}{2L_\lambda D^2} \min \left(\alpha^2 \operatorname{dist}(\boldsymbol{y}, \mathcal{Y}^*)^2, L_\lambda D^2 \alpha \operatorname{dist}(\boldsymbol{y}, \mathcal{Y}^*) \right) \; . \tag{76}$$

The same result is trivially true for $\boldsymbol{y} \in \mathcal{Y}^*$ (since in that case we have $d(\boldsymbol{y}) = d^*$). □

This Theorem leads to an immediate corollary on the norm of the gradient of d.

Corollary 1. *Under the same assumption as Theorem 1, for all $\boldsymbol{y} \in \mathbb{R}^d$ there exist a constant α such that,*

$$\|\nabla d(\boldsymbol{y})\| \geqslant \frac{1}{2L_\lambda D^2} \min\{\alpha^2 \operatorname{dist}(\boldsymbol{y}, \mathcal{Y}^*), \alpha L_\lambda D^2\} \geqslant \frac{\alpha}{\sqrt{2L_\lambda D^2}} \min\{\sqrt{d^* - d(\boldsymbol{y})}, \sqrt{L_\lambda D^2}\} \; . \tag{77}$$

Proof. We just need to notice that by convexity for all $\boldsymbol{y}^* \in \mathcal{Y}^*$ the suboptimality is upper bounded by the linearization of the function: $d^* - d(\boldsymbol{y}) \leqslant \langle \boldsymbol{y}^* - \boldsymbol{y}, \nabla d(\boldsymbol{y}) \rangle \leqslant \operatorname{dist}(\boldsymbol{y}, \mathcal{Y}^*) \|\nabla d(\boldsymbol{y})\|$. □

The second lower bound on the norm of the gradient (77) does not contradict the fact that there exist optimal points since it holds only for point $\boldsymbol{y}$ far from $\mathcal{Y}^*$. Actually, $0 \leqslant \max_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{y}^* - \boldsymbol{y}, M\boldsymbol{x} \rangle \leqslant \operatorname{dist}(\boldsymbol{y}, \mathcal{Y}^*) \max_{\boldsymbol{x} \in \mathcal{X}} \|M\boldsymbol{x}\|$ goes to 0 when $\boldsymbol{y} \to \mathcal{Y}^*$ and then eventually becomes smaller than D^2.

C.2 Properties of the function $\mathcal{L}$

We will first prove that for any $\boldsymbol{y} \in \mathbb{R}^d$ the function $\mathcal{L}(\cdot, \boldsymbol{y})$ has a property similar to strong convexity respect to the variable $M\boldsymbol{x}$, if $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$ is close to its minimum with respect to $\boldsymbol{x}$ then $M\boldsymbol{x}$ is close to the image by M of the minimizer of $\mathcal{L}(\cdot, \boldsymbol{y})$. More precisely,

Proposition 4. *for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathbb{R}^d$, if f is convex,*

$$\|M\boldsymbol{x} - M\hat{\boldsymbol{x}}(\boldsymbol{y})\|^2 \leqslant \frac{2}{\lambda}\left(\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) - \mathcal{L}(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y})\right) \quad \textit{where} \quad \hat{\boldsymbol{x}}(\boldsymbol{y}) \in \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) \;, \tag{78}$$

and $(\boldsymbol{x}, \boldsymbol{y}) := f(\boldsymbol{x}) + \mathbf{1}_{\mathcal{X}}(\boldsymbol{x}) + \langle \boldsymbol{y}, M\boldsymbol{x} \rangle + \frac{\lambda}{2}\|M\boldsymbol{x}\|^2$

Proof. By convexity of f we have that,

$$f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}}(\boldsymbol{y})) \geqslant \langle \nabla f(\hat{\boldsymbol{x}}(\boldsymbol{y})), \boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{y}) \rangle \;, \tag{79}$$

then by simple algebra (noting $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}(\boldsymbol{y})$),

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) - \mathcal{L}(\hat{\boldsymbol{x}}, \boldsymbol{y}) \geqslant \left\langle \nabla f(\hat{\boldsymbol{x}}) + M^\top \boldsymbol{y} + \lambda M^\top M\hat{\boldsymbol{x}}, \boldsymbol{x} - \hat{\boldsymbol{x}} \right\rangle + \frac{\lambda}{2}\|M\boldsymbol{x} - M\hat{\boldsymbol{x}}\|^2 \tag{80}$$

$$\geqslant \frac{\lambda}{2}\|M\boldsymbol{x} - M\hat{\boldsymbol{x}}\|^2 \;. \tag{81}$$

The last inequality come from the first order optimality condition on $\mathcal{L}(\cdot, \boldsymbol{y})$. □

Now let us introduce the key property allowing us to insure that $\boldsymbol{x}_t$ actually converge to $\boldsymbol{x}^*$. This proposition states that the primal gap $\Delta_t^{(p)}$ upper-bounds the squared distance to the optimum.

Proposition 5. *If f is a μ-strongly convex function then, $\mathcal{X}^* = \{\boldsymbol{x}^*\}$ and we have for all $t \geqslant 0$,*

$$\frac{\mu}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \leqslant \Delta_t^{(p)} - \Delta_t^{(d)} + \max\left(2\Delta_t^{(d)}, \sqrt{2L_\lambda D^2 \Delta_t^{(d)}}\right) \;. \tag{82}$$

and also

$$\frac{\mu}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \leqslant \Delta_t^{(p)} + \frac{2L_\lambda D^2}{\alpha^2}\|M\boldsymbol{x}_{t+1}\|\|M\hat{\boldsymbol{x}}_t\| \;, \; \forall t \in \mathbb{N} \;;\; \operatorname{dist}(\boldsymbol{y}_t, \mathcal{Y}^*) \leqslant \frac{L_\lambda D^2}{\alpha} \;. \tag{83}$$

Proof. We start from the identity

$$f(\boldsymbol{x}_{t+1}) - f^* = \Delta_t^{(p)} - \Delta_t^{(d)} - \langle \boldsymbol{y}_t, M\boldsymbol{x}_{t+1} \rangle - \frac{\lambda}{2}\|M\boldsymbol{x}_{t+1}\|^2 \;. \tag{84}$$

From first order optimality conditions we get for any $\boldsymbol{y}^* \in \mathcal{Y}^*$ and any $\boldsymbol{x} \in \mathcal{X}$,

$$\left\langle \boldsymbol{x} - \boldsymbol{x}^*, \nabla f(\boldsymbol{x}^*) + M^\top \boldsymbol{y}^* + \lambda M^\top M\boldsymbol{x}^* \right\rangle \geqslant 0 \quad \text{and} \quad M\boldsymbol{x}^* = 0 \;, \tag{85}$$

then for $\boldsymbol{x} = \boldsymbol{x}_{t+1}$,

$$\langle \boldsymbol{y}^*, M\boldsymbol{x}_{t+1} \rangle \geqslant -\langle \boldsymbol{x}_{t+1} - \boldsymbol{x}^*, \nabla f(\boldsymbol{x}^*) \rangle \;, \tag{86}$$

if f is μ-strongly convex then,

$$-\langle \boldsymbol{x}_{t+1} - \boldsymbol{x}^*, \nabla f(\boldsymbol{x}^*) \rangle \geqslant -f(\boldsymbol{x}_{t+1}) + f^* + \frac{\mu}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \;, \tag{87}$$

then combining (84), (86) and (87) we get for any $\boldsymbol{y}^* \in \mathcal{Y}^*$

$$\frac{\mu}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \leqslant \Delta_t^{(p)} - \Delta_t^{(d)} + \langle \boldsymbol{y}^* - \boldsymbol{y}_t, M\boldsymbol{x}_{t+1} \rangle \leqslant \Delta_t^{(p)} - \Delta_t^{(d)} + \operatorname{dist}(\boldsymbol{y}_t, \mathcal{Y}^*)\|M\boldsymbol{x}_{t+1}\| \;, \tag{88}$$

we then get using the fact that in (77),

$$\text{either} \quad \operatorname{dist}(\boldsymbol{y}_t, \mathcal{Y}^*) \geqslant \frac{L_\lambda D^2}{\alpha} \quad \text{or} \quad \operatorname{dist}(\boldsymbol{y}_t, \mathcal{Y}^*) \leqslant \frac{2L_\lambda D^2}{\alpha^2}\|\nabla d(\boldsymbol{y}_t)\| = \frac{2L_\lambda D^2}{\alpha^2}\|M\hat{\boldsymbol{x}}_t\| \;, \tag{89}$$

leading to

$$\frac{\mu}{2}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}^*\|^2 \leqslant \Delta_t^{(p)} - \Delta_t^{(d)} + \frac{2L_\lambda D^2}{\alpha^2}\|M\boldsymbol{x}_{t+1}\|\|M\hat{\boldsymbol{x}}_t\|\,,\ \forall t\in\mathbb{N}\,;\ \operatorname{dist}(\boldsymbol{y}_t,\mathcal{Y}^*)\leqslant\frac{L_\lambda D^2}{\alpha}\,. \tag{90}$$

Similarly, combining (88) and (64) gives us,

$$\frac{\mu}{2}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}^*\|^2 \leqslant \Delta_t^{(p)} - \Delta_t^{(d)} + \max\left(2\Delta_t^{(d)}, \sqrt{2L_\lambda D^2\Delta_t^{(d)}}\right)\,. \tag{91}$$

□

D Proof of Theorem 2 and Theorem 3

This section is decomposed into 3 subsections. First, we prove some intermediate results on the sequence computed by our algorithm to get the fundamental equation (99) that we will use to prove the convergence of $(\Delta_t)_{t\in\mathbb{N}}$. Then in subsection D.2 (respectively Subsection D.3) we prove Thm.2 (resp. Thm. 3). Let us recall that the Augmented Lagrangian function is defined as

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) := f(\boldsymbol{x}) + \mathbf{1}_{\mathcal{X}}(\boldsymbol{x}) + \langle \boldsymbol{y}, M\boldsymbol{x}\rangle + \tfrac{\lambda}{2}\|M\boldsymbol{x}\|^2\,,\quad \forall(\boldsymbol{x},\boldsymbol{y})\in\mathbb{R}^m\times\mathbb{R}^d\,, \tag{92}$$

where f is a smooth function, $\mathbf{1}_{\mathcal{X}}$ is the indicator function of a convex compact set $\mathcal{X}\subset\mathbb{R}^m$. The augmented dual function d is $d(\boldsymbol{y}) := \max_{\boldsymbol{x}\in\mathcal{X}}\mathcal{L}(\boldsymbol{x},\boldsymbol{y})$. The FW-AL algorithm computes

$$\begin{cases} \boldsymbol{x}_{t+1} = \mathcal{FW}(\boldsymbol{x}_t; \mathcal{L}(\cdot,\boldsymbol{y}_t))\,, \\ \boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \eta_t M\boldsymbol{x}_{t+1}\,, \end{cases} \tag{93}$$

where $\mathcal{FW}(\boldsymbol{x}_t;\mathcal{L}(\cdot,\boldsymbol{y}_t))$ is roughly a FW step. (More details in App. A).

D.1 Lemmas on the sequences computed by FW-AL

The two following lemmas do not require any assumption on the sets or the functions, they only rely on the definition of the algorithm. They provide upper bounds on the decrease of the primal and the dual gaps. They are true for all functions f and constraint set $\mathcal{X}$. Recall that we respectively defined the primal and the dual gap as,

$$\Delta_t^{(d)} := d^* - d(\boldsymbol{y}_t) \quad\text{and}\quad \Delta_t^{(p)} := \mathcal{L}(\boldsymbol{x}_{t+1};\boldsymbol{y}_t) - d(\boldsymbol{y}_t)\,. \tag{94}$$

The first lemma upper bounds the decrease of the dual suboptimality, note that Hong and Luo (2017) are probably not the firsts to provide such lemma. We are citing them because we provide the proof proposed in their paper.

Lemma 2 (Lemma 3.2 (Hong and Luo, 2017))**.** *For any $t\geqslant 1$, there holds*

$$\Delta_{t+1}^{(d)} - \Delta_t^{(d)} \leqslant -\eta_t\langle M\boldsymbol{x}_{t+1}, M\hat{\boldsymbol{x}}_{t+1}\rangle. \tag{95}$$

Proof.

$$\begin{aligned} \Delta_{t+1}^{(d)} - \Delta_t^{(d)} &= d(\boldsymbol{y}_t) - d(\boldsymbol{y}_{t+1}) \\ &= \mathcal{L}(\hat{\boldsymbol{x}}_t,\boldsymbol{y}_t) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1},\boldsymbol{y}_{t+1}) \\ &\overset{(\star)}{\leqslant} \mathcal{L}(\hat{\boldsymbol{x}}_{t+1},\boldsymbol{y}_t) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1},\boldsymbol{y}_{t+1}) \\ &= \langle \boldsymbol{y}_t - \boldsymbol{y}_{t+1}, M\hat{\boldsymbol{x}}_{t+1}\rangle \\ &= -\eta_t\langle M\boldsymbol{x}_{t+1}, M\hat{\boldsymbol{x}}_{t+1}\rangle\,, \end{aligned} \tag{96}$$

where $(\star)$ is because $\hat{\boldsymbol{x}}_t$ is the minimizer of $\mathcal{L}(\cdot,\boldsymbol{y}_t)$. □

Next we proceed to bound the decrease of the primal gap $\Delta_{t+1}^{(p)}$.

Lemma 3 (weaker version of Lemma 3.3 (Hong and Luo, 2017)). *Then for any $t \geqslant 1$, we have*

$$\Delta_{t+1}^{(p)} - \Delta_t^{(p)} \leqslant \eta_t \|M\boldsymbol{x}_{t+1}\|^2 + (\mathcal{L}(\boldsymbol{x}_{t+2}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1})) - \eta_t \langle M\boldsymbol{x}_{t+1}, M\hat{\boldsymbol{x}}_{t+1}\rangle . \tag{97}$$

Proof. We start using the definition of $\Delta_{t+1}^{(p)}$,

$$\begin{aligned}
\Delta_{t+1}^{(p)} - \Delta_t^{(p)} &= \mathcal{L}(\boldsymbol{x}_{t+2}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1}) - (\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_t) - \mathcal{L}(\hat{\boldsymbol{x}}_t, \boldsymbol{y}_t)) \\
&= \mathcal{L}(\boldsymbol{x}_{t+2}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_t) + (\mathcal{L}(\hat{\boldsymbol{x}}_t, \boldsymbol{y}_t) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1})) \\
&\overset{(96)}{\leqslant} \mathcal{L}(\boldsymbol{x}_{t+2}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) + \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_t) - \eta_t \langle M\boldsymbol{x}_{t+1}, M\hat{\boldsymbol{x}}_{t+1}\rangle \\
&\overset{(\star)}{=} (\mathcal{L}(\boldsymbol{x}_{t+2}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1})) + \eta_t \|M\boldsymbol{x}_{t+1}\|^2 - \eta_t \langle M\boldsymbol{x}_{t+1}, M\hat{\boldsymbol{x}}_{t+1}\rangle ,
\end{aligned}$$

where the last inequality $(\star)$ is by definition of $\mathcal{L}$ and because $\boldsymbol{y}_{t+1} - \boldsymbol{y}_t = \eta_t M\boldsymbol{x}_{t+1}$. □

We can now combine Lemma 2 and Lemma 3 with our technical result Cor. 1 on the dual suboptimality to get our fundamental descent lemma only valid under Assumption 1.

Lemma 4 (Fundamental descent Lemma). *Under Assumption 1 we have that for all $t \geqslant 0$,*

$$\Delta_{t+1} - \Delta_t \leqslant \frac{2\eta_t}{\lambda} (\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1})) + \mathcal{L}(\boldsymbol{x}_{t+2}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \eta_t \frac{\alpha^2}{2L_\lambda D^2} \min\{\Delta_{t+1}^{(d)}, L_\lambda D^2\} ,$$

Proof. Combining (95) and (97) gives us,

$$\begin{aligned}
\Delta_{t+1} - \Delta_t &= [\Delta_{t+1}^{(p)} - \Delta_t^{(p)}] + [\Delta_{t+1}^{(d)} - \Delta_t^{(d)}] \\
&\leqslant \eta_t \|M\boldsymbol{x}_{t+1}\|^2 + \mathcal{L}(\boldsymbol{x}_{t+2}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - 2\eta_t \langle M\boldsymbol{x}_{t+1}, M\hat{\boldsymbol{x}}_{t+1}\rangle \\
&= \eta_t \|M\boldsymbol{x}_{t+1} - M\hat{\boldsymbol{x}}_{t+1}\|^2 - \eta_t \|M\hat{\boldsymbol{x}}_{t+1}\|^2 + \mathcal{L}(\boldsymbol{x}_{t+2}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) .
\end{aligned} \tag{98}$$

Finally, from the "strong convexity" of $\mathcal{L}(\cdot, \boldsymbol{y}_t)$ respect to $M\boldsymbol{x}$ (78) we obtain,

$$\Delta_{t+1} - \Delta_t \leqslant \frac{2\eta_t}{\lambda} (\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1})) + \mathcal{L}(\boldsymbol{x}_{t+2}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \eta_t \|M\hat{\boldsymbol{x}}_{t+1}\|^2 , \tag{99}$$

where $\Delta_{t+1} := \Delta_{t+1}^{(p)} + \Delta_{t+1}^{(d)}$. Then we can use our fundamental technical result (Corollary (1)) relating the dual suboptimality and the norm of its gradient,

$$\|M\hat{\boldsymbol{x}}_{t+1}\|^2 = \|\nabla d(\boldsymbol{y}_{t+1})\|^2 \geqslant \frac{\alpha^2}{2L_\lambda D^2} \min\{\Delta_{t+1}^{(d)}, L_\lambda D^2\} , \tag{100}$$

to get the desired lemma. □

The two following sections respectively deal with the proof of Theorem 3 and Theorem 3 they both start from our fundamental descent lemma (Lemma 4).

D.2 Proof of Theorem 2

Let us first recall the setting and propose a detailed version of the first part of Thm.2. The second part of Thm.2 is proposed in Corollary 2.

Theorem' 2. *If $\mathcal{X}$ is a compact convex set and f is L-smooth, using any algorithm with* sublinear decrease (13) *as inner loop in* FW-AL (5) *and $\eta_t := \min\left\{\frac{2}{\lambda}, \frac{\alpha^2}{2\delta}\right\}\frac{2}{t+2}$ then there exists a bounded $t_0 \geqslant 0$ such that,*

$$\Delta_t \leqslant \min\left\{\frac{4\delta(t_0+2)}{t+2}, \delta\right\} \quad \forall t \geqslant t_0 \quad \textit{and} \quad t_0 \leqslant \left(\tfrac{C}{\delta} + 2\right) \exp\left(\frac{\Delta_0 - \delta + 2C}{2\delta}\right) . \tag{101}$$

where $C := 8\delta \max\left(1, \frac{64\delta}{\lambda^2\alpha^2}\right)$ *and* $\delta := L_\lambda D^2$.

If we set $\eta_t = \min\left\{\frac{2}{\lambda}, \frac{\alpha^2}{4\delta}\right\}\frac{C}{\delta}$ *for at least* t_0 *iterations and then* $\eta_t := \min\left\{\frac{2}{\lambda}, \frac{\alpha^2}{2\delta}\right\}\frac{2}{t+2}$ *we get*

$$\Delta_t \leqslant \min\left\{\frac{4\delta(t_0+2)}{t+2}, \delta\right\} \quad \forall t \geqslant t_0 \quad \textit{where} \quad t_0 = \max\left\{1 + \frac{2(\Delta_0 - \delta)C}{\delta^2}, \frac{C}{\delta}\right\}. \tag{102}$$

Proof. This proof will start from Lemma 4 and use the fact that if $\mathcal{X}$ is a general convex compact set, a usual Frank Wolfe step (Alg. 2) produces a sublinear decrease (13). It leads to the following equation holding for any $\gamma \in [0, 1]$,

$$\Delta_{t+1} - \Delta_t \leqslant \left(\frac{2\eta_t}{\lambda} - \gamma\right)\left(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1})\right) + \gamma^2\frac{L_\lambda D^2}{2} - \eta_t\frac{\alpha^2}{2L_\lambda D^2}\min\{\Delta_{t+1}^{(d)}, L_\lambda D^2\}, \tag{103}$$

Then for $\gamma = \frac{4\eta_t}{\lambda}$ we get,

$$\Delta_{t+1} - \Delta_t \leqslant -\frac{2\eta_t}{\lambda}\left(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1})\right) + \left(\frac{4\eta_t}{\lambda}\right)^2\frac{L_\lambda D^2}{2} - \eta_t\frac{\alpha^2}{2L_\lambda D^2}\min\{\Delta_{t+1}^{(d)}, L_\lambda D^2\}. \tag{104}$$

Since we are doing line-search we know that $\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) \geqslant \mathcal{L}(\boldsymbol{x}_{t+2}, \boldsymbol{y}_{t+1})$ implying that

$$\Delta_{t+1} - \Delta_t \leqslant -\frac{2\eta_t}{\lambda}\Delta_{t+1}^{(p)} + \left(\frac{4\eta_t}{\lambda}\right)^2\frac{L_\lambda D^2}{2} - \eta_t\frac{\alpha^2}{2L_\lambda D^2}\min\{\Delta_{t+1}^{(d)}, L_\lambda D^2\}, \tag{105}$$

In order to make appear Δ_{t+1} in the RHS we will introduce $a = \min\left\{\frac{2}{\lambda}, \frac{\alpha^2}{2L_\lambda D^2}\right\}$, this constant depends on λ which is a hyperparameter. It seems that λ helps to scale the decrease of the primal with to the one of the dual.

$$\Delta_{t+1} - \Delta_t \leqslant -a\eta_t\min\{\Delta_{t+1}, L_\lambda D^2\} + \left(\frac{4\eta_t}{\lambda}\right)^2\frac{L_\lambda D^2}{2}. \tag{106}$$

Then we have either that,

$$\Delta_{t+1} - \Delta_t \leqslant -aL_\lambda D^2\eta_t + \left(\frac{4\eta_t}{\lambda}\right)^2\frac{L_\lambda D^2}{2}, \tag{107}$$

giving a uniform (in time) decrease with a small enough constant step size η_t or we have,

$$\Delta_{t+1} - \Delta_t \leqslant -a\eta_t\Delta_{t+1} + \left(\frac{4\eta_t}{\lambda}\right)^2\frac{L_\lambda D^2}{2}, \tag{108}$$

giving a usual Frank-Wolfe recurrence scheme leading to a sublinear decrease with a decreasing step size $\eta_t \sim 1/t$. The problem here is that since we don't have access to Δ_{t+1} we do not know in which regime we are. The same way it seems hard to get an adaptive step size. In order to tackle this problem we will consider an upper bound looser than (106) leading to a separation of the two regimes. Let us introduce $\bar{\eta}_t := a\eta_t$, $\delta := L_\lambda D^2$ and $C := 8\delta\max\left(1, \frac{64\delta}{\lambda^2\alpha^2}\right)$, we have that (106) implies

$$\Delta_{t+1} - \Delta_t \leqslant -\bar{\eta}_t\min\{\Delta_{t+1}, \delta\} + \bar{\eta}_t^2\frac{C}{2}. \tag{109}$$

Lemma 5. *If there exists* $t_0 > \frac{C}{\delta} - 2$ *such that* $\Delta_{t_0} \leqslant \delta$ *and if we set* $\bar{\eta}_t = \frac{2}{2+t}$ *then,*

$$\Delta_t \leqslant \min\left\{\frac{4\delta(t_0+2)}{t+2}, \delta\right\} \quad \forall t \geqslant t_0. \tag{110}$$

Proof. For $t = t_0$ the result comes from the fact that we assumed that $\Delta_{t_0} \leqslant \delta$, now let us assume that for a $t \geqslant t_0$, $\Delta_t \leqslant \frac{4\delta(t_0+2)}{t+2}$ then if Δ_{t+1} was greater than δ we would have get,

$$\delta \leqslant \Delta_{t+1} \leqslant \Delta_t - \frac{2}{2+t}\delta + \left(\frac{2}{2+t}\right)^2\frac{C}{2} \leqslant \delta - \frac{2}{2+t}\delta + \left(\frac{2}{2+t}\right)^2\frac{C}{2}, \tag{111}$$

implying that,

$$\delta \leqslant \frac{C}{2+t} \qquad \text{and then} \qquad t \leqslant \frac{C}{\delta} - 2 \tag{112}$$

which contradicts the assumption $t > \frac{C}{\delta} - 2$. Leading to $\Delta_t \leqslant \delta\,,\ \forall t \geqslant t_0$.

Moreover, we have for all $t \geqslant t_0$,

$$\Delta_{t+1} \leqslant \Delta_t - \frac{2}{2+t}\Delta_{t+1} + \left(\frac{2}{2+t}\right)^2 \frac{C}{2} \tag{113}$$

$$\frac{t+4}{t+2}\Delta_{t+1} \leqslant \Delta_t + \left(\frac{2}{2+t}\right)^2 \frac{C}{2} \tag{114}$$

$$\Delta_{t+1} \leqslant \frac{t+2}{t+4}\Delta_t + \frac{2C}{(2+t)(t+4)} \tag{115}$$

$$\overset{(\star)}{\leqslant} \frac{t+2}{t+4}\frac{4\delta(t_0+2)}{t+2} + \frac{2C}{(t+2)(t+4)} \tag{116}$$

$$\leqslant \frac{4\delta(t_0+2)}{t+3}\left[\frac{t+3}{t+4}\left(1 + \frac{1}{2(t+2)}\right)\right]\,, \tag{117}$$

where $(\star)$ is due to the induction hypothesis and the last inequality is due to the fact that $\delta(t_0+2) \geqslant C$. Then, we just need to show that

$$\left[\frac{t+3}{t+4}\left(1 + \frac{1}{2(t+2)}\right)\right] \leqslant 1\,, \quad \forall t \geqslant 1\,. \tag{118}$$

That is true because

$$\left[\frac{t+3}{t+4}\left(1 + \frac{1}{2(t+2)}\right)\right] \leqslant 1 \tag{119}$$

$$\Leftrightarrow (t+3)(t+\tfrac{5}{2}) \leqslant (2+t)(t+4) \tag{120}$$

$$\Leftrightarrow\ -\frac{1}{2}t + \frac{15}{2} \leqslant 8 \tag{121}$$

$$\Leftrightarrow\ t \geqslant 1\,. \tag{122}$$

□

Now we have to show that in a finite number of iterations t_0 we can reach a point such that $\Delta_{t_0} \leqslant \delta$.

Let us assume that $\Delta_0 \geqslant \delta$, then we cannot initialize the recurrence (110). Instead we will show the following

Lemma 6. *Let $(\Delta_t)_{t\in\mathbb{N}}$ a sequence such that $\Delta_{t+1} - \Delta_t \leqslant -\bar{\eta}_t \min\{\Delta_{t+1}, \delta\} + \bar{\eta}_t^2 \frac{C}{2}\,,\ \forall t \in \mathbb{N}$. We have that,*

- *If $\bar{\eta}_t = \frac{\delta}{C}$ then there exists $t_0 \in \mathbb{N}$ such that,*

$$\Delta_{t_0} \leqslant \delta\,, \quad \Delta_t \leqslant \delta \quad ; \quad \forall\, t \geqslant t_0\,, \quad \text{and} \quad t_0 \leqslant 1 + \frac{2(\Delta_0 - \delta)C}{\delta^2}\,. \tag{123}$$

- *If $\bar{\eta}_t = \frac{2}{2+t}$ then there exists $t_0 \geqslant \frac{C}{\delta} - 2$ such that,*

$$\Delta_{t_0} \leqslant \delta \quad \text{and} \quad t_0 \leqslant \left(\tfrac{C}{\delta} + 2\right)\exp\left(\frac{\Delta_0 - \delta + 2C}{2\delta}\right)\,. \tag{124}$$

Proof. If t_0 did not exist would (109) gives us for all $t \in \mathbb{N}$,

$$\Delta_{t+1} - \Delta_t \leqslant -\bar{\eta}_t\delta + \bar{\eta}_t^2\frac{C}{2}\,. \tag{125}$$

Then we would have either for $\bar{\eta}_t = \frac{\delta}{C}$ or $\bar{\eta}_t = \frac{2}{2+t}$,

$$\infty = \delta\sum_{t=0}^{\infty}\bar{\eta}_t \leqslant \Delta_0 + \frac{C}{2}\sum_{t=0}^{\infty}\bar{\eta}_t^2 < \infty\,. \tag{126}$$

Then let us consider the smallest time t_0 such that $\Delta_{t_0} \leqslant \delta$.

- If we set $\bar{\eta}_t = \frac{\delta}{C}$, we get for all $t < t_0$

$$\Delta_{t+1} - \Delta_t \leqslant -\frac{\delta^2}{2C} \tag{127}$$

and then summing for $0 \leqslant t \leqslant t_0 - 2$

$$\delta - \Delta_0 \leqslant \Delta_{t_0-1} - \Delta_0 \leqslant -\frac{(t_0-1)\delta^2}{2C} \,, \tag{128}$$

implying that

$$t_0 \leqslant 1 + \frac{2(\Delta_0 - \delta)C}{\delta^2} \tag{129}$$

then, let us show by recurrence that $\forall t \geqslant t_0,\ \Delta_t \leqslant \delta$. The result for $t = t_0$ is true by definition of t_0. Let us assume that it is true for a $t \geqslant t_0$, then either $\Delta_{t+1} \geqslant \delta/2$ and in that case (109) gives us

$$\Delta_{t+1} \leqslant \Delta_t - \frac{\Delta_t \delta}{2C} + \frac{\delta^2}{2C} \leqslant \Delta_t \leqslant \delta \,, \tag{130}$$

otherwise $\Delta_{t+1} \leqslant \delta/2 \leqslant \delta$.

- If $\bar{\eta}_t = \frac{2}{2+t}$, we want a $t_0 \geqslant \frac{C}{\delta} - 2$ so if $\Delta_{\lfloor \frac{C}{\delta} - 1 \rfloor} \leqslant \delta$ we are done, otherwise

$$\delta \sum_{t=\lfloor \frac{C}{\delta}-1 \rfloor}^{t_0-2} \frac{2}{2+t} \leqslant \Delta_0 - \delta + \frac{C}{2} \sum_{t=0}^{\infty} \frac{4}{(2+t)^2} \leqslant \Delta_0 - \delta + 2C\Big(\frac{\pi^2}{6} - 1\Big) \leqslant \Delta_0 - \delta + 2C \,. \tag{131}$$

Since $\sum_{t=\lfloor \frac{C}{\delta}+1 \rfloor}^{t_0} \frac{1}{t} \geqslant \ln(t_0) - \ln(\frac{C}{\delta} + 2)$ we get that

$$t_0 \leqslant \big(\tfrac{C}{\delta} + 2\big) \exp\left(\frac{\Delta_0 - \delta + 2C}{2\delta}\right) \,. \tag{132}$$

□

To sum up, we can either set $\bar{\eta}_t = \frac{\delta}{C}$ for a fixed number of iterations or we can use a decreasing step size leading to a very bad upper bound on t_0. Nevertheless this bound for the decreasing step size is very conservative and even if the best theoretical rates are given by a constant step size $\bar{\eta}_t$ for a number of iterations proportional to $\frac{C}{\delta}$ and then a sublinear step size $\bar{\eta}_t = \frac{2}{2+t}$, in practice, we can directly start with a decreasing step size.

Corollary 2. *Under the same assumption as Thm. 2. Let the $t_0 \in \mathbb{N}$ stated in Thm. 2, then for all $T \geqslant t_0$,*

$$\min_{t_0 \leqslant t \leqslant T} \|M\hat{\boldsymbol{x}}_{t+1}\|^2 \leqslant \frac{2\Theta}{T - t_0 + 1} \quad \textit{and} \quad \min_{t_0 \leqslant t \leqslant T} \|M\boldsymbol{x}_{t+1}\|^2 \leqslant \frac{8\Theta}{T - t_0 + 1} \,. \tag{133}$$

where $\Theta := \frac{\lambda}{2} + \frac{2^8\delta^2}{\alpha^4\lambda}$. Moreover, if f is μ-strongly convex we have that,

$$\min_{8t_0+15 \leqslant t \leqslant T} \|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \leqslant \frac{4K}{\mu}\left[\frac{\lambda}{2} + \frac{8\sqrt{2}\delta}{\alpha^2}\right] \frac{\Theta}{T - 8t_0 - 14} \,. \tag{134}$$

Proof. This proof follows the same idea as the proof of (Lacoste-Julien et al., 2013, Thm C.3). Since we are working with different quantities and that the rates are slightly different from the ones provided in (Lacoste-Julien et al., 2013) we will provide a complete proof of this result. We start from the fundamental descent lemma (99). We use the fact that a usual Frank-Wolfe step produces a sublinear decrease (A.3) that we specify for $\gamma = \frac{4\eta_t}{\lambda}$ to get a similar equation as (105),

$$\Delta_{t+1} - \Delta_t \leqslant -\frac{2\eta_t}{\lambda}\left(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1})\right) + \left(\frac{4\eta_t}{\lambda}\right)^2 \frac{\delta}{2} - \eta_t \|M\hat{\boldsymbol{x}}_{t+1}\|^2 \,, \tag{135}$$

noting $\delta := L_\lambda D^2$. Then introducing $h_{t+1} := (\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1}))$, (note that because of the line search $\Delta_{t+1}^{(p)} \geqslant h_{t+1} \geqslant 0$) it leads to

$$\left(\frac{2}{\lambda} h_{t+1} + \|M\hat{\boldsymbol{x}}_{t+1}\|^2\right) \leqslant \frac{\Delta_t - \Delta_{t+1}}{\eta_t} + \eta_t \left(\frac{4}{\lambda}\right)^2 \frac{\delta}{2} \,. \tag{136}$$

which is similar equation as (Lacoste-Julien et al., 2013, Eq.(22)). We will then use the same proof technique. Let $\{w_t\}_{t_0}^T$ a sequence of positive weights and let $\rho_t := {}^{w_t}/_{\sum_{t=t_0}^T w_t}$ the associated normalized weights. The convex combination of (136) give us,

$$\begin{aligned}
\sum_{t=t_0}^{T} \rho_t \left(\frac{2}{\lambda} h_{t+1} + \|M\hat{\boldsymbol{x}}_{t+1}\|^2\right) &\leqslant \sum_{t=t_0}^{T} \rho_t \frac{\Delta_t - \Delta_{t+1}}{\eta_t} + \sum_{t=t_0}^{T} \rho_t \eta_t \left(\frac{4}{\lambda}\right)^2 \frac{\delta}{2} \\
&= \frac{\rho_{t_0}}{\eta_{t_0}} \Delta_{t_0} - \frac{\rho_T}{\eta_T} \Delta_{T+1} + \sum_{t=t_0}^{T-1} \Delta_t \left(\frac{\rho_{t+1}}{\eta_{t+1}} - \frac{\rho_t}{\eta_t}\right) + \sum_{t=t_0}^{T} \rho_t \eta_t \left(\frac{4}{\lambda}\right)^2 \frac{\delta}{2} \\
&\leqslant \frac{\rho_{t_0}}{\eta_{t_0}} \Delta_{t_0} + \sum_{t=t_0}^{T-1} \Delta_t \left(\frac{\rho_{t+1}}{\eta_{t+1}} - \frac{\rho_t}{\eta_t}\right) + \sum_{t=t_0}^{T} \rho_t \eta_t \left(\frac{4}{\lambda}\right)^2 \frac{\delta}{2} \,.
\end{aligned} \tag{137}$$

We can now use a *weighted average* such as $w_t = t - t_0$. This kind of average leads to

$$\frac{\rho_{t+1}}{\eta_{t+1}} - \frac{\rho_t}{\eta_t} = \frac{(t - t_0 + 1)(t+3) - (t - t_0)(t+2)}{a(T - t_0)(T - t_0 + 1)} = \frac{2t - t_0 + 3}{a(T - t_0)(T - t_0 + 1)} \,. \tag{138}$$

where $\eta_t := \min\left\{\frac{2}{\lambda}, \frac{\alpha^2}{4\delta}\right\} \frac{2}{t+2} = a\frac{2}{t+2}$. Then we can plug that $\Delta_t \leqslant \min\left\{\frac{4\delta(t_0+2)}{t+2}, \delta\right\}$, $\forall t \geqslant t_0$ to get,

$$\sum_{t=t_0}^{T} \rho_t \left(\frac{2}{\lambda} h_{t+1} + \|M\hat{\boldsymbol{x}}_{t+1}\|^2\right) \leqslant \frac{\delta \rho_{t_0}}{\eta_{t_0}} + \sum_{t=t_0}^{T-1} \frac{4\delta(t_0+2)}{t+2} \left(\frac{\rho_{t+1}}{\eta_{t+1}} - \frac{\rho_t}{\eta_t}\right) + \sum_{t=t_0}^{T} \rho_t \eta_t \left(\frac{4}{\lambda}\right)^2 \frac{\delta}{2} \tag{139}$$

$$\leqslant \frac{2}{(T - t_0)(T - t_0 + 1)} \left[\sum_{t=t_0}^{T-1} \frac{4\delta(t_0+2)(2t - t_0 + 3)}{a(t+2)} + \sum_{t=t_0}^{T} \frac{2a(t - t_0)}{2+t} \left(\tfrac{4}{\lambda}\right)^2 \tfrac{\delta}{2}\right] \tag{140}$$

$$\leqslant \frac{2}{T - t_0 + 1} \left[\frac{8\delta(t_0+2)}{a} + \frac{16a\delta}{\lambda^2}\right] \tag{141}$$

Then,

$$\min_{t_0 \leqslant t \leqslant T} \left[\frac{2}{\lambda} h_{t+1} + \|M\hat{\boldsymbol{x}}_{t+1}\|^2\right] \leqslant \frac{2}{T - t_0 + 1} \left[\frac{8\delta(t_0+2)}{a} + \frac{16a\delta}{\lambda^2}\right] \,. \tag{142}$$

To upper bound $\|M\boldsymbol{x}_{t+1}\|^2$ the idea is to combine the previous equation with $\|M\boldsymbol{x}_{t+1} - M\hat{\boldsymbol{x}}_t\|^2 \leqslant \frac{2}{\lambda}\left(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_t) - \mathcal{L}(\hat{\boldsymbol{x}}_t, \boldsymbol{y}_t)\right) =: \frac{2}{\lambda}\Delta_t^{(p)} \leqslant \frac{2}{\lambda} h_t$ (Prop. 4 + the fact that we perform line search) giving,

$$\|M\boldsymbol{x}_{t+1}\|^2 \leqslant \frac{8}{\lambda} h_t + 4\|M\hat{\boldsymbol{x}}_t\|^2 \tag{143}$$

$$\min_{t_0+1 \leqslant t \leqslant T+1} \|M\boldsymbol{x}_{t+1}\|^2 \leqslant 4 \min_{t_0+1 \leqslant t \leqslant T+1} \left(\|M\hat{\boldsymbol{x}}_t\|^2 + \frac{2}{\lambda} h_t\right) \leqslant \frac{8}{T - t_0 + 1} \left[\frac{8\delta(t_0+2)}{a} + \frac{16a\delta}{\lambda^2}\right] \tag{144}$$

If f is μ-strongly convex we can use Prop. 5 to get,

$$\frac{\mu}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \leqslant \Delta_t^{(p)} + \frac{2\delta}{\alpha^2}\|M\boldsymbol{x}_{t+1}\|\|M\hat{\boldsymbol{x}}_{t+1}\| \,, \ \forall t \in \mathbb{N} \,; \ \operatorname{dist}(\boldsymbol{y}_t, \mathcal{Y}^*) \leqslant \frac{\delta}{\alpha} \,. \tag{145}$$

In order to show that at some point we have $\operatorname{dist}(\boldsymbol{y}_t, \mathcal{Y}^*) \leqslant \frac{\delta}{\alpha}$ we will use Thm. 1 and (110) to get,

$$\frac{4\delta(t_0+2)}{t+2} \geqslant \Delta_t^p \geqslant \frac{1}{2\delta} \min\left\{\alpha^2 \operatorname{dist}(\boldsymbol{y}_t, \mathcal{Y}^*)^2, \alpha\delta \operatorname{dist}(\boldsymbol{y}_t, \mathcal{Y}^*)\right\} \,, \ \forall t \geqslant t_0 \,, \tag{146}$$

Then for all $t \geqslant t_0$ such that $\operatorname{dist}(\boldsymbol{y}_t, \mathcal{Y}^*) > \frac{\delta}{\alpha}$ we have that,

$$\frac{8\delta(t_0+2)}{t+2} \geqslant \alpha \operatorname{dist}(\boldsymbol{y}_t, \mathcal{Y}^*) \tag{147}$$

implying that for $t \geqslant 8(t_0+2) - 2 = 8t_0 + 14$ we have that $\alpha \operatorname{dist}(\boldsymbol{y}_t, \mathcal{Y}^*) \leqslant \delta$ and then,

$$\begin{aligned}
\frac{\mu}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 &\leqslant \Delta_t^{(p)} + \frac{2\delta}{\alpha^2}\|M\boldsymbol{x}_{t+1}\|\|M\hat{\boldsymbol{x}}_t\| && (148)\\
&\leqslant h_t + \frac{2\delta}{\alpha^2} 4\sqrt{2}\|M\hat{\boldsymbol{x}}_t\|\sqrt{\frac{2}{\lambda}h_t + \|M\hat{\boldsymbol{x}}_t\|^2} && (149)\\
&\leqslant \frac{\lambda}{2}\|M\hat{\boldsymbol{x}}_t\|^2 + h_t + \frac{2^8\delta^2}{\alpha^4\lambda}\left(\frac{2}{\lambda}h_t + \|M\hat{\boldsymbol{x}}_t\|^2\right) && (150)\\
&\leqslant \left(\frac{\lambda}{2} + \frac{2^8\delta^2}{\alpha^4\lambda}\right)\left(\frac{2}{\lambda}h_t + \|M\hat{\boldsymbol{x}}_t\|^2\right) . && (151)
\end{aligned}$$

It then implies that,

$$\begin{aligned}
\min_{8t_0+15 \leqslant t \leqslant T+1} \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 &\leqslant \frac{2K}{\mu}\left(\frac{\lambda}{2} + \frac{2^8\delta^2}{\alpha^4\lambda}\right) \min_{8t_0+15 \leqslant t \leqslant T}\left(\frac{2}{\lambda}h_t + \|M\hat{\boldsymbol{x}}_t\|^2\right) && (152)\\
&\leqslant \frac{2K}{\mu}\left(\frac{\lambda}{2} + \frac{2^8\delta^2}{\alpha^4\lambda}\right)\frac{2}{T - 8t_0 - 14}\left[\frac{8\delta(8t_0+17)}{a} + \frac{16a\delta}{\lambda^2}\right] . && (153)
\end{aligned}$$

□

D.3 Proof of Theorem 3

This proof starts with the fundamental descent lemma (Lemma 4). It uses the fact that if $\mathcal{X}$ is a polytope and if we use an algorithm with a geometric decrease (15) such as Alg. 1 then with a small enough constant step size η_t we can upper bound the decrease of $\Delta_{t+1} - \Delta_t$.

Lemma 7. *for all $t \geqslant 1$,*

$$\Delta_{t+1} - \Delta_t \leqslant -\frac{\rho_A}{2}\Delta_{t+1}^{(p)} - \frac{\lambda\alpha^2\rho_A}{8L_\lambda D^2}\min\{\Delta_{t+1}^{(d)}, L_\lambda D^2\} . \tag{154}$$

Proof. To prove Lemma 7, we start from Lemma 4 to obtain

$$\begin{aligned}
\Delta_{t+1} - \Delta_t &\leqslant \mathcal{L}(\boldsymbol{x}_{t+2}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) + \frac{2\eta_t}{\lambda}\left(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1})\right) - \eta_t\frac{\alpha^2}{2L_\lambda D^2}\min\{\Delta_{t+1}^{(d)}, L_\lambda D^2\} ,\\
&\overset{(15)}{\leqslant} \left(\frac{2\eta_t}{\lambda} - \rho_A\right)\left(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1})\right) - \eta_t\frac{\alpha^2}{2L_\lambda D^2}\min\{\Delta_{t+1}^{(d)}, L_\lambda D^2\} . && (155)
\end{aligned}$$

Now we can choose $\eta_t = \frac{\lambda\cdot\rho_A}{4}$ giving us Lemma 7. □

Note that, the optimal λ in Lemma 7 is $\lambda = \frac{4L_\lambda D^2}{\alpha^2}$. From this lemma we can deduce a constant decrease for a finite number of step and eventually a geometric decrease.

Lemma 8. *For all $\lambda > 0$ if we set $\eta_t = \frac{\lambda\rho_A}{4}$ for finite number of steps t_0 the quantity Δ_t decreases by a uniform amount as,*

$$\Delta_{t+1} - \Delta_t \leqslant -\frac{\lambda\alpha^2\rho_A}{8} \quad \textit{where} \quad t_0(\Delta_0) \leqslant 1 + \frac{8(\Delta_0 - L_\lambda D^2)}{\lambda\alpha^2} , \tag{156}$$

otherwise Δ_t decrease geometrically as,

$$\Delta_{t+1} \leqslant \frac{1}{1+\kappa}\Delta_t \quad \textit{where} \quad \kappa := \frac{\rho_A}{2}\min\left\{1, \frac{\lambda\alpha^2}{4L_\lambda D^2}\right\} , \tag{157}$$

Particularly, if we set $\lambda = \frac{4L_\lambda D^2}{\alpha^2}$ *and* $\eta_t = \frac{2L_\lambda D^2 \rho_A}{\alpha^2}$ *for* t_0 *step the sum of the gaps decreases by an uniform amount as,*

$$\Delta_{t+1} \leqslant \Delta_t - \frac{\rho_A L D^2}{2} \quad \textit{with} \quad t_0(\Delta_0) \leqslant 1 + \frac{2(\Delta_0 - L_\lambda D^2)}{\rho_A L_\lambda D^2} \,, \tag{158}$$

and otherwise geometrically, as

$$\Delta_{t+1} \leqslant \frac{\Delta_t}{1+\kappa} \quad \textit{where} \quad \kappa := \frac{\rho_A}{2} \,. \tag{159}$$

Proof. We start from Lemma 7, if $\Delta^{(d)}_{t+1} \leqslant L_\lambda D^2$,

$$\Delta_{t+1} - \Delta_t \leqslant -\frac{\lambda \alpha^2 \rho_A}{16} \,, \tag{160}$$

and otherwise,

$$\Delta_{t+1} - \Delta_t \leqslant -\frac{\rho_A}{2}\Delta^{(p)}_{t+1} - \frac{\lambda \alpha^2 \rho_A}{8 L_\lambda D^2}\Delta^{(d)}_{t+1} \leqslant -\kappa \Delta_{t+1} \,. \tag{161}$$

Our goal is then just to show the upper bound on t_0. First let us notice that (Δ_t) is decreasing then this non negative sequence cannot decrease by a uniform amount an infinite number of time, then we can sum for $t = 1, \dots, t_0 - 1$ such that (160) holds to get,

$$L_\lambda D^2 - \Delta_0 \leqslant \Delta^{(d)}_{t_0 - 1} - \Delta_0 \leqslant \sum_{t=0}^{t_0 - 2} \Delta_{t+1} - \Delta_t \leqslant -(t_0 - 1)\frac{\lambda \alpha^2 \rho_A}{16} \tag{162}$$

□

One can deduce several convergence properties from this lemma which are compiled in Theorem 3.

Corollary 3. *There exist* $t_0 \leqslant 1 + \frac{16(\Delta_0 - L_\lambda D^2)}{\lambda \alpha^2}$ *such that for all* $t \geqslant t_0$ *we have the following properties,*

1. *The gap decreases linearly,*

$$\Delta_t \leqslant \frac{L_\lambda D^2}{(1+\kappa)^{t - t_0}} \,. \tag{163}$$

2. *The sequences of feasibility violations at points* $\boldsymbol{x}_{t+1}$ *and* $\hat{\boldsymbol{x}}_{t+1}$ *decrease linearly,*

$$\|M\hat{\boldsymbol{x}}_{t+1}\|^2 \leqslant \frac{4}{\lambda \cdot \rho_A}\frac{L_\lambda D^2}{(1+\kappa)^{t-t_0}} \quad \textit{and} \quad \|M\boldsymbol{x}_{t+1}\|^2 \leqslant \frac{16}{\lambda \cdot \rho_A}\frac{L_\lambda D^2}{(1+\kappa)^{t-t_0}}, \tag{164}$$

where $\kappa := \frac{\rho_A}{2}\min\left\{1, \frac{\lambda \alpha^2}{8 L_\lambda D^2}\right\}$.

Finally, if f *is* μ_f*-strongly convex, the distance of the current point to the optimal set vanishes as,*

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \leqslant \frac{2 L_\lambda D^2 (\sqrt{2}+1)}{\mu_f (\sqrt{1+\kappa})^{t-t_0}} \,. \tag{165}$$

Proof. To prove the first statement let us start from Lemma 7, for all $t \geqslant 0$,

$$[\Delta^{(p)}_{t+1} + \Delta^{(d)}_{t+1}] - [\Delta^{(p)}_t + \Delta^{(d)}_t] \leqslant -\frac{\rho_A}{2}\left(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1})\right) - \frac{\lambda \cdot \rho_A}{4}\|M\hat{\boldsymbol{x}}_{t+1}\|^2 \,, \tag{166}$$

leading us directly to

$$\frac{\lambda \cdot \rho_A}{4}\|M\hat{\boldsymbol{x}}_{t+1}\|^2 \leqslant \Delta_t \quad \text{and} \quad \frac{\rho_A}{2}\left(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1})\right) \leqslant \Delta_t \,. \tag{167}$$

To upper bound $\|M\boldsymbol{x}_{t+1}\|^2$ the idea is to combine the two previous equations with $\|M\boldsymbol{x}_{t+1} - M\hat{\boldsymbol{x}}_{t+1}\|^2 \leqslant \frac{2}{\lambda}\left(\mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{x}}_{t+1}, \boldsymbol{y}_{t+1})\right)$ (Prop. 4) giving,

$$\|M\boldsymbol{x}_{t+1}\|^2 \leqslant 2\|M\hat{\boldsymbol{x}}_{t+1}\|^2 + 2\|M\boldsymbol{x}_{t+1} - M\hat{\boldsymbol{x}}_{t+1}\|^2 \leqslant \frac{16}{\lambda \rho_A}\Delta_t \,. \tag{168}$$

The last statement directly follows from the fact (Δ_{t+1}) decreases linearly (Lemma 8) and the fact that one can upper bound $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2$ with the primal and dual suboptimalities (Proposition 5),

$$\frac{\mu}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \leqslant \Delta_t^{(p)} - \Delta_t^{(d)} + \max\left(2\Delta_t^{(d)}, \sqrt{2L_\lambda D^2 \Delta_t^{(d)}}\right) . \tag{169}$$

Then, it easily follows that

$$\frac{\mu}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \leqslant \frac{(\sqrt{2}+1)L_\lambda D^2}{(\sqrt{1+\kappa})^{t-t_0}} . \tag{170}$$

□