

Faster Lagrangian-Based Methods in Convex Optimization

Shoham Sabach*

Marc Teboulle†

October 27, 2020

Abstract

In this paper, we aim at unifying, simplifying, and improving the convergence rate analysis of Lagrangian-based methods for convex optimization problems. We first introduce the notion of nice primal algorithmic map, which plays a central role in the unification and in the simplification of the analysis of all Lagrangian-based methods. Equipped with a nice primal algorithmic map, we then introduce a versatile generic scheme, which allows for the design and analysis of Faster LAGrangian (FLAG) methods with a new provably sublinear rate of convergence expressed in terms of functions values and feasibility violation of the original (non-ergodic) generated sequence. To demonstrate the power and versatility of our approach and results, we show that all well-known iconic Lagrangian-based schemes admit a nice primal algorithmic map, and hence share the new faster rate of convergence results within their corresponding FLAG.

2010 Mathematics Subject Classification: 90C25, 65K05.

Keywords: Augmented Lagrangian, Lagrangian multiplier methods, proximal multiplier algorithms, convex composite minimization, alternating direction method of multiplier, non-smooth optimization, fast non-ergodic global rate of convergence.

1 Introduction

Over the last decade, we are witnessing a resurgence of Lagrangian-based methods. This surge of interest is due to the flexibility that they provide in exploiting special problem structures that abound in modern and disparate applications (*e.g.*, image science, machine learning, communication systems). The Augmented Lagrangian methods (also known as multiplier methods) which are a typical prototype, were invented about half a century ago by Hestenes [16] and Powell [20], are the core of the analysis and development of many extensions of Lagrangian-based schemes capable of tackling optimization problems with various types of structures and constraints. This is reflected by a rich and very large volume of literature on various practical and theoretical aspects of many Lagrangian-based methods. Two classical references providing the central developments and ideas underlying Lagrangian-based and related decomposition methods are the books of Bertsekas [3], and Bertsekas and Tsitsiklis [4], which include a comprehensive list of relevant references for these developments until the nineties. For more modern developments, results, and applications, see *e.g.*, the more recent works [5, 23, 7, 22], and references therein.

A recent and voluminous literature on Lagrangian-based methods has focused on rate of convergence results for the so-called Alternating Direction of Multipliers Method [12, 11, 10], with an almost absolute predominance on *ergodic* type of rate of convergence results. Since the literature on this topic is very wide, let us just mention the first two central works [6, 14], which have obtained an ergodic rate of $O(1/N)$. For many more rate of convergence results of the ergodic type, and relevant references, we refer the readers to, *e.g.*, [8], and the very recent monograph [18], which also includes an up to date extensive bibliography.

*Faculty of Industrial Engineering and Management, The Technion, Haifa, 32000, Israel. E-mail: ssabach@ie.technion.ac.il.

†School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel. E-mail: teboulle@post.tau.ac.il. This research was partially supported by the Israel Science Foundation, under ISF Grants 1844-16, and 2619-20.

The focus of this paper is on faster and non-ergodic global rate of convergence analysis for Lagrangian-based methods applied to convex optimization, where the rate is measured in terms of function values and feasibility violation of the original produced sequence. The literature on papers obtaining non-ergodic rate of convergence results remains very limited, in particular, in term of the measures just mentioned. For instance, in [15], the authors obtained a non-ergodic $O(1/N)$ result for the squared distance between two successive primal iterates, while in [6], a non-ergodic $O(1/N^2)$ result, in the strongly convex setting, is proven for the squared distance between the primal iterates to an optimal primal solution. One exception is the very recent work [17], which derives a new non-ergodic $O(1/N)$ rate of convergence result in terms of function values and feasibility violation of the original produced sequence for the specific Linearized ADMM, through a rather involved and nontrivial analysis (compare with the results developed in Section 5 for this particular algorithm).

The analysis of Lagrangian-based methods is usually intricate and complicated, which results in lengthy proofs. Recently, in [22], we have shown that all well-known Lagrangian-based methods and their decomposition variants can be captured and analysed through a single scheme (that we called the Perturbed Proximal Method of Multipliers, see details in [22]). It allows us to derive an ergodic $O(1/N)$ rate of convergence in terms of function values and feasibility violation. Much like most of the existing literature, the approach we developed there was limited to Lagrangian-based methods with classical sublinear ergodic type results, and hence naturally leads to the question: *Can we develop an algorithmic framework capable of producing non-ergodic rate as well as faster Lagrangian-based schemes?* In this paper, we provide a positive answer, within a surprisingly simple and elegant approach. This is developed in four main sections which contents are now briefly outlined.

The formulation of the problem under study, some basic preliminaries on Lagrangians in the convex setting, and some important examples of convex optimization problems which fit our model, can be found in Section 2. The starting point of our study is that all Lagrangian-based methods can be simply described along the following protocol. Consider the corresponding augmented Lagrangian of the problem at hand, which consists of primal and dual (multiplier) variables, building a Lagrangian-based method requires to choose updating rules for both the primal and the dual variables. Since, in all Lagrangian-based methods, the dual variable is updated via an explicit formula, the main difference between Lagrangian-based methods is encapsulated in the choice of a *primal algorithmic map* which updates the primal variables. This choice is governed by the problem's structure and data information, which must be best exploited, and can be seen as a single step of any appropriate optimization method that applied on the augmented Lagrangian itself, or a variation of it. For instance, the classical Augmented Lagrangian method [16, 20], is a typical example whereby the primal algorithmic map is nothing else but an *exact minimization* applied on the augmented Lagrangian itself associated to problem (P).

Adopting this point of view for Lagrangian-based methods, we introduce in Section 3 the notion of *nice primal algorithmic map*, which will plays a central role in unifying the analysis of all Lagrangian-based methods into a unitary and simple framework. In turns, equipped with a nice primal algorithmic map, we propose a generic algorithmic framework, called (FLAG) for Faster LAGrangian based methods that aims at achieving faster (as well as classical) convergence rates of all Lagrangian-based methods *at once*. FLAG which is equipped with a nice primal algorithmic maps allows for the design and analysis of faster schemes with new provable sublinear *Non-ergodic* rate of $O(1/N^2)$ in the strongly convex case and a *Non-ergodic* $O(1/N)$ in the convex case (see Theorem 4.1 and Theorem 4.2, respectively). As a by product, we also derive classical and faster ergodic results in Corollaries 4.1 and 4.2; see Section 4. To illustrate the power and versatility of FLAG and our new results, in Section 5, we consider various block and other models of problem (P) and demonstrate that all well-known iconic schemes, *e.g.*, Linearized Proximal Method of Multipliers, ADMM and its linearized versions, Jacobi type decomposition, Predictor Corrector Proximal Multipliers, and many more, all admit a nice primal algorithmic map, and hence can be used in FLAG to obtain new rate of convergence results. This provides a significant list of faster schemes for fundamental well-known Lagrangian-based methods.

2 Problem Formulation

In this paper, we consider the linearly constrained convex optimization model,

$$(P) \quad \min \{ \Psi(x) : \mathcal{A}x = b, x \in \mathbb{R}^n \},$$

where $\Psi : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ is a proper, lower semi-continuous and convex function, $\mathcal{A} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a linear mapping, and $b \in \mathbb{R}^m$. Throughout this work, the feasible set of problem (P) is denoted by $\mathcal{F} = \{x \in \mathbb{R}^n : \mathcal{A}x = b\}$.

Despite its apparent simplicity, with an additional structural/data information on the objective function or/and on the linear mapping (see examples below), the optimization model (P) essentially captures a very wide range of convex optimization problems arising in many modern applications. Since we are interested in analyzing this model also in the strongly convex case, we instead assume that $\Psi(\cdot)$ is σ -strongly convex with $\sigma \geq 0$ (i.e., when $\sigma = 0$ we recover the convex setting). At this juncture, it is important to mention that in many applied problems the function Ψ is defined by the sum of two or more functions, see e.g., [7]. Therefore in such cases, to warrant the σ -strong convexity of $\Psi(\cdot)$, it is enough to assume that *one* of the functions is σ -strongly convex, see Section 5.2 for more details.

Let us illustrate few basic and important instances of model (P).

Example 2.1 (Linear composite model). The linear composite optimization model, which naturally appears in many applications, fuels very much the study of Lagrangian-based methods and is given by

$$\min_{u \in \mathbb{R}^p} \{ f(u) + g(Au) \}, \quad (2.1)$$

where $f : \mathbb{R}^p \rightarrow (-\infty, +\infty]$ and $g : \mathbb{R}^q \rightarrow (-\infty, +\infty]$ are proper, lower semi-continuous and convex functions, and $A : \mathbb{R}^p \rightarrow \mathbb{R}^q$ is a linear mapping. Rewriting problem (2.1) as an equivalent linearly constrained optimization problem (which requires the definition of an auxiliary variable) yields

$$\min_{u \in \mathbb{R}^p, v \in \mathbb{R}^q} \{ f(u) + g(v) : Au = v \}.$$

We easily see that it fits into model (P), with $x = (u^T, v^T)^T$, $\Psi(x) := f(u) + g(v)$ and $\mathcal{A}x = Au - v$ (i.e., $b = \mathbf{0}$).

Example 2.2 (Block linear constrained model). Another useful and more general model which includes (2.1) is the following

$$\min_{u \in \mathbb{R}^p, v \in \mathbb{R}^q} \{ f(u) + g(v) : Au + Bv = b \}, \quad (2.2)$$

where $f : \mathbb{R}^p \rightarrow (-\infty, +\infty]$ and $g : \mathbb{R}^q \rightarrow (-\infty, +\infty]$ are proper, lower semi-continuous and convex functions, $A : \mathbb{R}^p \rightarrow \mathbb{R}^m$ and $B : \mathbb{R}^q \rightarrow \mathbb{R}^m$ are linear mappings. Again we can easily see that it fits into model (P), with $x = (u^T, v^T)^T$, $\Psi(x) := f(u) + g(v)$ and $\mathcal{A}x = Au + Bv$.

Example 2.3 (Additive smooth/non-smooth composite model). A very classical model in optimization is of additive composite nature that include the sum of two functions one which is smooth and one which is not necessarily smooth. Here we consider this model with an additional linear constraint, that is

$$\min_{x \in \mathbb{R}^n} \{ f(x) + h(x) : \mathcal{A}x = b \}, \quad (2.3)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a proper, lower semi-continuous and convex function, while $h : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ is a continuously differentiable function with a Lipschitz continuous gradient. Again we easily see that it fits into model (P), with $\Psi(x) := f(x) + h(x)$, being the sum of a smooth function and a non-smooth function. This additional structure in the objective function that can be beneficially exploited when developing an adequate Lagrangian-based method, as we discuss in details in Section 5.3.

We conclude this section with some basic properties of Lagrangians in convex optimization. For the equality constrained problem given in model (P), the *Lagrangian* $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \rightarrow (-\infty, +\infty]$ is defined by

$$\mathcal{L}(x, y) \equiv \Psi(x) + \langle y, \mathcal{A}x - b \rangle, \quad (2.4)$$

where $y \in \mathbb{R}^m$ is the *Lagrange multiplier* associated with the linear constraint $\mathcal{A}x = b$. In this study we will also use the *augmented Lagrangian* associated to problem (P) and defined by

$$\mathcal{L}_\rho(x, y) := \mathcal{L}(x, y) + \frac{\rho}{2} \|\mathcal{A}x - b\|^2 = \Psi(x) + \langle y, \mathcal{A}x - b \rangle + \frac{\rho}{2} \|\mathcal{A}x - b\|^2, \quad (2.5)$$

where $\rho > 0$ is a *penalty parameter*. Clearly, when $\rho = 0$ we recover the Lagrangian itself, *i.e.*, $\mathcal{L} \equiv \mathcal{L}_0$.

Throughout this work we make the following standing assumption.

Assumption A. The Lagrangian \mathcal{L} has a saddle point, that is, there exists a pair (x^*, y^*) such that

$$\mathcal{L}(x^*, y) \leq \mathcal{L}(x^*, y^*) \leq \mathcal{L}(x, y^*), \quad \forall x \in \mathbb{R}^n, \quad \forall y \in \mathbb{R}^m. \quad (2.6)$$

It is well-known (see [21]) that (x^*, y^*) is a saddle point of \mathcal{L} if and only if the following conditions hold:

- (i) x^* is a solution of problem (P), *i.e.*, $\Psi(x^*)$ is finite valued and $\mathcal{A}x^* = b$.
- (ii) The multiplier y^* is an optimal solution of the dual problem associated to problem (P), given by

$$(D) \quad d^* = \sup_{y \in \mathbb{R}^m} \{d(y) := \Psi^*(-\mathcal{A}^T y) - \langle y, b \rangle\},$$

and $p^* := \mathcal{L}(x^*, y^*) = d^*$, where $\Psi^*(\cdot)$ stands for the Fenchel conjugate function of Ψ (see [21]).

Thanks to the convexity of problem (P), recall that the existence of an optimal dual Lagrange multiplier y^* attached to the constraint $\mathcal{A}x = b$ is ensured under the standard constraint qualification for problem (P), that can be formulated as follows: there exists a $\bar{x} \in \text{ri}(\text{dom } \Psi)$ satisfying $\mathcal{A}\bar{x} = b$, and that strong duality holds, *i.e.*, $p^* = d^* \in \mathbb{R}$. Moreover, the set of optimal dual solutions is non-empty, convex and compact, see *e.g.*, [21].

We end this section with some notations that will be used throughout the paper. We denote by \mathbb{S}_+^n and \mathbb{S}_{++}^n the sets of all $n \times n$ *positive semi-definite matrices* and *positive definite matrices*, respectively. Given $P \in \mathbb{S}_+^n$, $\|u\|_P := \sqrt{\langle u, Pu \rangle}$ stands for the semi-norm of $u \in \mathbb{R}^n$. For any $u, v, w \in \mathbb{R}^n$, we define the following quantity

$$\Delta_P(u, v, w) := \frac{1}{2} \left(\|u - v\|_P^2 - \|u - w\|_P^2 \right). \quad (2.7)$$

In the case that $P = I$, where I_n denotes the $n \times n$ *identity matrix*, we simply denote $\Delta := \Delta_P$.

3 A Unified Framework for Lagrangian-based Methods

The starting point of our study is that all Lagrangian-based methods can be simply described along the following unified approach. Considering the corresponding augmented Lagrangian of the problem at hand, which consists of primal and dual (multiplier) variables, building a Lagrangian-based method requires to choose updating rules for both the primal and the dual variables. Since, in all Lagrangian-based methods, the dual variable is updated via an explicit formula (see (3.2) below), the main difference between Lagrangian-based methods is encapsulated in the choice of a *primal algorithmic map* that updates the primal variables. Therefore, any Lagrangian-based method updates a couple (z, y) by first choosing a primal algorithmic map \mathcal{P} , and will update the primal and dual (multiplier) variables via

$$z^+ \in \mathcal{P}(z, y), \quad (3.1)$$

$$y^+ = y + \mu\rho (\mathcal{A}z^+ - b), \quad (3.2)$$

where $\mu > 0$ is a scaling parameter. The update of the primal variable which depends on the algorithmic map \mathcal{P} , can be seen as a single step of any optimization method that applied on the augmented Lagrangian itself or a variation of it. For example, the classical Augmented Lagrangian method [16, 20], is a typical prototype where the mapping \mathcal{P} is nothing but an *exact minimization* applied on the augmented Lagrangian itself.

Adopting this point of view of Lagrangian-based methods, we propose a general framework that aims at achieving convergence rates (classical and fast) of all Lagrangian-based methods at once. Considering the optimization problem (P) with the data $[\Psi, \mathcal{A}, b, \sigma]$ and its associated augmented Lagrangian $\mathcal{L}_\rho(\cdot)$, which is defined in (2.5), we introduce the notion of *nice primal algorithmic map*, that plays a central role in unifying the analysis of all Lagrangian-based methods into a single and simple framework.

Definition 3.1 (Nice primal algorithmic map). Given $(z, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m$ and the parameters $\rho, t > 0$ we let $\rho_t = \rho$ and $\tau_t = t^{-1}$ (when $\sigma = 0$) or $\rho_t = \rho t$ and $\tau_t = t$ (when $\sigma > 0$). A primal algorithmic map $\text{Prim}_t : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$, which is applied on the augmented Lagrangian $\mathcal{L}_{\rho_t}(z, \lambda)$, that generates z^+ via $z^+ \in \text{Prim}_t(z, \lambda)$, is called *nice*, if there exist a parameter $\delta \in (0, 1]$ and matrices $P, Q \in \mathbb{S}_+^n$, such that for any $\xi \in \mathcal{F}$ we have

$$\mathcal{L}_{\rho_t}(z^+, \lambda) - \mathcal{L}_{\rho_t}(\xi, \lambda) \leq \tau_t \Delta_P(\xi, z, z^+) - \frac{\tau_t}{2} \|z^+ - z\|_Q^2 - \frac{\sigma}{2} \|\xi - z^+\|^2 - \frac{\delta \rho_t}{2} \|\mathcal{A}z^+ - b\|^2. \quad (3.3)$$

The essence of our ability to both unify the analysis, and to accelerate the convergence rate of all Lagrangian-based methods at once, comes from the algorithmic framework that we introduce now. This framework crucially uses as an input a nice primal algorithmic map $\text{Prim}_t(\cdot)$ with the induced parameter $\delta \in (0, 1]$ and matrices $P, Q \in \mathbb{S}_+^n$.

FLAG – Faster LAGrangian based method

1. **Input:** The problem data $[\Psi, \mathcal{A}, b, \sigma]$, and a nice primal algorithmic map $\text{Prim}_t(\cdot)$.
2. **Initialization:** Set $\mu \in (0, \delta]$ and $\rho > 0$. Start with any $(x^0, z^0, y^0) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m$, $t_0 = 1$.
3. **Iterations:** Sequences $\{(x^k, z^k, y^k)\}_{k \in \mathbb{N}}$ and $\{t_k\}_{k \in \mathbb{N}}$ are generated as follows: for each $k \geq 0$
 - 3.1. If $\sigma = 0$, let $\tau_k = t_k^{-1}$ and $\rho_k = \rho$, or, if $\sigma > 0$, let $\tau_k = t_k$ and $\rho_k = \rho t_k$. Compute

$$\lambda^k = y^k + \rho_k (t_k - 1) (\mathcal{A}x^k - b). \quad (3.4)$$

- 3.2. Update the sequence $\{(x^k, z^k, y^k)\}_{k \in \mathbb{N}}$ by

$$z^{k+1} \in \text{Prim}_k(z^k, \lambda^k), \quad (3.5)$$

$$y^{k+1} = y^k + \mu \rho_k (\mathcal{A}z^{k+1} - b), \quad (3.6)$$

$$x^{k+1} = (1 - t_k^{-1}) x^k + t_k^{-1} z^{k+1}. \quad (3.7)$$

- 3.3. Update the sequence $\{t_k\}_{k \in \mathbb{N}}$ by solving the equation $t_{k+1}^p - t_k^p = t_{k+1}^{p-1}$, i.e.,

$$t_{k+1} = \begin{cases} t_k + 1, & p = 1 \text{ (convex case),} \\ (1 + \sqrt{1 + 4t_k^2})/2, & p = 2 \text{ (strongly convex case).} \end{cases} \quad (3.8)$$

Remark 3.1 (Few comments on the algorithmic framework FLAG). (i) Borrowing ideas from well-known accelerated first order minimization methods (see, for instance, [19, 1, 2]), here the sequence $\{t_k\}_{k \in \mathbb{N}}$ plays a key role in accelerating the input map, *i.e.*, the nice primal algorithmic map $\text{Prim}_t(\cdot)$. To this end, the augmented parameter ρ_k and the prox parameter τ_k are determined and chosen through the definition of the recursion which defines the sequence $\{t_k\}_{k \in \mathbb{N}}$ (*cf.* (3.8)).

(ii) Clearly, setting $t_k \equiv 1$, for all $k \in \mathbb{N}$, in FLAG, implies $\rho_k \equiv \rho$, $\lambda^k \equiv y^k$, and $x^k \equiv z^k$, thus recovering the classical basic Lagrangian-based method discussed above.

(iii) A main new feature in our algorithmic framework FLAG is the introduction of the fundamental auxiliary sequence $\{\lambda^k\}_{k \in \mathbb{N}}$, which will enable us to derive new faster rate of convergence results in terms of the *original* sequence produced by FLAG. As we shall see, in Section 4.3, when λ^k coincides with y^k , only ergodic type rates (classical and fast) will be obtained.

In FLAG, once the nice algorithmic map $\text{Prim}_t(\cdot)$ has been chosen, *i.e.*, a parameter $\delta \in (0, 1]$ and matrices $P, Q \in \mathbb{S}_+^n$, have been determined, Definition 3.1 translates in terms of the iterative sequence $\{(x^k, z^k, y^k)\}_{k \in \mathbb{N}}$ produced by FLAG as follows: for each $k \geq 0$ and $\xi \in \mathcal{F}$,

$$\begin{aligned} \mathcal{L}_{\rho_k}(z^{k+1}, \lambda^k) - \mathcal{L}_{\rho_k}(\xi, \lambda^k) &\leq \tau_k \Delta_P(\xi, z^k, z^{k+1}) - \frac{\tau_k}{2} \|z^{k+1} - z^k\|_Q^2 - \frac{\sigma}{2} \|\xi - z^{k+1}\|^2 \\ &\quad - \frac{\delta \rho_k}{2} \|\mathcal{A}z^{k+1} - b\|^2, \end{aligned} \quad (3.9)$$

where $\tau_k = t_k^{-1}$ and $\rho_k = \rho$ (when $\sigma = 0$) or $\tau_k = t_k$ and $\rho_k = \rho t_k$ (when $\sigma > 0$).

To conclude this part, and before we prove that the property of being a nice primal algorithmic map (as defined in Definition 3.1) is *all we need* to guarantee rate of convergence results (classical and fast), we would like to note that there is no need to enter into the mechanics of the proofs to be presented next in order to derive theoretical guarantees of newly developed/analyzed Lagrangian-based methods, but just proving that the primal algorithmic map at hand is nice! This will be extensively illustrated in Section 5 whereby, after providing a *recipe* for obtaining rate of a given method, we will prove that all iconic Lagrangian-based schemes admit a nice primal map, and hence through FLAG will generate new faster schemes.

4 Convergence Analysis of FLAG

The section is divided into three parts, where we start with two key lemmas, which will serve us in deriving the main *non-ergodic* rate of convergence results in Section 4.2, followed by the ergodic type results in Section 4.3.

4.1 Main Pillars of the Analysis

Before proceeding, we recall the following fundamental *Pythagoras three-points identity* that is frequently used below. Given any matrix $P \in \mathbb{S}_+^n$, we have

$$2 \langle u - w, P(w - v) \rangle = \|u - v\|_P^2 - \|u - w\|_P^2 - \|v - w\|_P^2, \quad \forall u, v, w \in \mathbb{R}^n. \quad (4.1)$$

We also recall the main parameter sequences introduced in FLAG and which, for convenience, will be used often and systematically as follows:

$$\rho_k = \rho t_k^{p-1}, \quad (\rho > 0) \quad \text{and} \quad t_k^p - t_{k-1}^p = t_k^{p-1}, \quad (t_{-1} \equiv 0 \text{ and } t_0 = 1),$$

where $p = 1$ (if $\sigma = 0$) or $p = 2$ (if $\sigma > 0$), and $\tau_k = t_k^{-1}$ if $\sigma = 0$, or $\tau_k = t_k$ if $\sigma > 0$. We will also often use the following expression. For any $k \geq 0$ and any $\xi \in \mathcal{F}$ we define,

$$s_k := \mathcal{L}_{\rho t_{k-1}^p}(x^k, \eta) - \mathcal{L}_{\rho t_{k-1}^p}(\xi, \eta) \equiv \mathcal{L}_{\rho t_{k-1}^p}(x^k, \eta) - \Psi(\xi). \quad (4.2)$$

We start with the first useful technical lemma.

Lemma 4.1. *Let $\{(x^k, z^k, y^k)\}_{k \in \mathbb{N}}$ be a sequence generated by FLAG. Then, for any $\xi \in \mathcal{F}$, $\eta \in \mathbb{R}^m$ and $k \geq 0$, we have*

$$\begin{aligned} \mathcal{L}_{\rho_k}(z^{k+1}, \eta) - \mathcal{L}_{\rho_k}(\xi, \eta) &\leq \tau_k \Delta_P(\xi, z^k, z^{k+1}) - \frac{\sigma}{2} \|\xi - z^{k+1}\|^2 + \frac{1}{\mu \rho_k} \Delta(\eta, y^k, y^{k+1}) \\ &\quad - \rho t_{k-1}^p \langle \mathcal{A}x^k - b, \mathcal{A}z^{k+1} - b \rangle. \end{aligned}$$

Proof. Since $\{(x^k, z^k, y^k)\}_{k \in \mathbb{N}}$ is generated by FLAG, we obtain from (3.9) after omitting the non-negative term $(\tau_k/2) \|z^{k+1} - z^k\|_Q^2$, that

$$\mathcal{L}_{\rho_k}(z^{k+1}, \lambda^k) - \mathcal{L}_{\rho_k}(\xi, \lambda^k) \leq \tau_k \Delta_P(\xi, z^k, z^{k+1}) - \frac{\sigma}{2} \|\xi - z^{k+1}\|^2 - \frac{\delta \rho_k}{2} \|\mathcal{A}z^{k+1} - b\|^2. \quad (4.3)$$

From the multiplier update (3.6) and the three-points identity (4.1), we obtain, for all $\eta \in \mathbb{R}^m$,

$$\begin{aligned} \langle \eta - y^k, \mathcal{A}z^{k+1} - b \rangle &= \frac{1}{\mu \rho_k} \langle \eta - y^k, y^{k+1} - y^k \rangle = \frac{1}{\mu \rho_k} \Delta(\eta, y^k, y^{k+1}) + \frac{1}{2\mu \rho_k} \|y^{k+1} - y^k\|^2 \\ &= \frac{1}{\mu \rho_k} \Delta(\eta, y^k, y^{k+1}) + \frac{\mu \rho_k}{2} \|\mathcal{A}z^{k+1} - b\|^2. \end{aligned} \quad (4.4)$$

Using the update rule of the sequence $\{t_k\}_{k \in \mathbb{N}}$ (cf. (3.8)) we have that $\rho_k(t_k - 1) = \rho t_{k-1}^p$ and thus

$$\lambda^k = y^k + \rho_k(t_k - 1)(\mathcal{A}x^k - b) = y^k + \rho t_{k-1}^p(\mathcal{A}x^k - b).$$

Using this together with (4.4) yields, for all $\eta \in \mathbb{R}^m$, that

$$\begin{aligned} \langle \eta - \lambda^k, \mathcal{A}z^{k+1} - b \rangle &= \langle \eta - y^k, \mathcal{A}z^{k+1} - b \rangle - \rho t_{k-1}^p \langle \mathcal{A}x^k - b, \mathcal{A}z^{k+1} - b \rangle \\ &= \frac{1}{\mu \rho_k} \Delta(\eta, y^k, y^{k+1}) + \frac{\mu \rho_k}{2} \|\mathcal{A}z^{k+1} - b\|^2 - \rho t_{k-1}^p \langle \mathcal{A}x^k - b, \mathcal{A}z^{k+1} - b \rangle. \end{aligned} \quad (4.5)$$

Adding (4.3) to (4.5) yields (recall that $\mu \leq \delta$) the desired result. \square

From this result, we will get the (classical and fast) ergodic rate of convergence results to be presented in Section 4.3 (see Corollaries 4.1 and 4.2). However, in order to obtain the *non-ergodic* results in Section 4.2, the following result is the main step towards the derivation of the promised (classical and fast) rate of convergence results.

Lemma 4.2. *Let $\{(x^k, z^k, y^k)\}_{k \in \mathbb{N}}$ be a sequence generated by FLAG. Then, for any $\xi \in \mathcal{F}$, $\eta \in \mathbb{R}^m$ and $k \geq 0$, we have*

$$t_k^p s_{k+1} - t_{k-1}^p s_k \leq \frac{\tau_k \rho_k}{\rho} \Delta_P(\xi, z^k, z^{k+1}) - \frac{\rho_k \sigma}{2\rho} \|\xi - z^{k+1}\|^2 + \frac{1}{\mu \rho} \Delta(\eta, y^k, y^{k+1}),$$

where $s_k = \mathcal{L}_{\rho t_{k-1}^p}(x^k, \eta) - \mathcal{L}_{\rho t_{k-1}^p}(\xi, \eta)$ (cf. 4.2).

Proof. Since $x^{k+1} = (1 - t_k^{-1})x^k + t_k^{-1}z^{k+1}$ (cf. (3.7)) we obtain from the convexity of $\Psi(\cdot)$ that

$$\Psi(x^{k+1}) \leq (1 - t_k^{-1})\Psi(x^k) + t_k^{-1}\Psi(z^{k+1}).$$

Therefore, multiplying both sides by t_k^p (recalling that $t_k^p - t_k^{p-1} = t_{k-1}^p$), we obtain

$$t_k^p (\Psi(x^{k+1}) - \Psi(\xi)) - t_{k-1}^p (\Psi(x^k) - \Psi(\xi)) \leq t_k^{p-1} (\Psi(z^{k+1}) - \Psi(\xi)).$$

In addition, using again (3.7), yields that

$$t_k^p \langle \eta, \mathcal{A}x^{k+1} - b \rangle - t_{k-1}^p \langle \eta, \mathcal{A}x^k - b \rangle = t_k^{p-1} \langle \eta, \mathcal{A}z^{k+1} - b \rangle.$$

Combining these two facts yields (using the definition of the Lagrangian \mathcal{L})

$$t_k^p \left(\mathcal{L}(x^{k+1}, \eta) - \mathcal{L}(\xi, \eta) \right) - t_{k-1}^p \left(\mathcal{L}(x^k, \eta) - \mathcal{L}(\xi, \eta) \right) \leq t_k^{p-1} \left(\mathcal{L}(z^{k+1}, \eta) - \mathcal{L}(\xi, \eta) \right). \quad (4.6)$$

Now, we will again use that $x^{k+1} = (1 - t_k^{-1})x^k + t_k^{-1}z^{k+1}$ to obtain

$$\|\mathcal{A}x^{k+1} - b\|^2 = (1 - t_k^{-1})^2 \|\mathcal{A}x^k - b\|^2 + t_k^{-2} \|\mathcal{A}z^{k+1} - b\|^2 + 2(1 - t_k^{-1})t_k^{-1} \langle \mathcal{A}x^k - b, \mathcal{A}z^{k+1} - b \rangle.$$

Multiplying both sides of the above equality by $\rho t_k^{2p}/2$ (recalling $\rho_k = \rho t_k^{p-1}$ and $t_k^p - t_k^{p-1} = t_{k-1}^p$) yields

$$\frac{\rho t_k^{2p}}{2} \|\mathcal{A}x^{k+1} - b\|^2 - \frac{\rho t_{k-1}^{2p}}{2} \|\mathcal{A}x^k - b\|^2 = \frac{\rho_k t_k^{p-1}}{2} \|\mathcal{A}z^{k+1} - b\|^2 + \rho_k t_{k-1}^p \langle \mathcal{A}x^k - b, \mathcal{A}z^{k+1} - b \rangle. \quad (4.7)$$

Therefore, adding (4.6) to (4.7) yields (recall that $\|\mathcal{A}\xi - b\|^2 = 0$) with $s_k = \mathcal{L}_{\rho t_k^p}(x^k, \eta) - \mathcal{L}_{\rho t_{k-1}^p}(\xi, \eta)$,

$$t_k^p s_{k+1} - t_{k-1}^p s_k \leq t_k^{p-1} \left(\mathcal{L}_{\rho_k}(z^{k+1}, \eta) - \mathcal{L}_{\rho_k}(\xi, \eta) \right) + \rho_k t_{k-1}^p \langle \mathcal{A}x^k - b, \mathcal{A}z^{k+1} - b \rangle.$$

From Lemma 4.1, after we multiplied both sides by t_k^{p-1} (recall that $\rho_k = \rho t_k^{p-1}$), we obtain that

$$\begin{aligned} t_k^{p-1} \left(\mathcal{L}_{\rho_k}(z^{k+1}, \eta) - \mathcal{L}_{\rho_k}(\xi, \eta) \right) + \rho_k t_{k-1}^p \langle \mathcal{A}x^k - b, \mathcal{A}z^{k+1} - b \rangle &\leq \frac{\tau_k \rho_k}{\rho} \Delta_P(\xi, z^k, z^{k+1}) - \frac{\rho_k \sigma}{2\rho} \|\xi - z^{k+1}\|^2 \\ &\quad + \frac{1}{\mu\rho} \Delta(\eta, y^k, y^{k+1}). \end{aligned}$$

By combining the last two inequalities, we thus get

$$t_k^p s_{k+1} - t_{k-1}^p s_k \leq \frac{\tau_k \rho_k}{\rho} \Delta_P(\xi, z^k, z^{k+1}) - \frac{\rho_k \sigma}{2\rho} \|\xi - z^{k+1}\|^2 + \frac{1}{\mu\rho} \Delta(\eta, y^k, y^{k+1}),$$

which proves the desired result. \square

We conclude this section with the following simple result which summarizes some useful properties of the sequence $\{t_k\}_{k \in \mathbb{N}}$ (cf. (3.8)) that will be used below.

Lemma 4.3. *Let $\{t_k\}_{k \in \mathbb{N}}$ be the sequence of positive numbers generated from $t_0 = 1$ via the equation $t_{k+1}^p - t_k^p = t_{k+1}^{p-1}$, with $p = 1$ or $p = 2$. Then, for any $k \geq 0$, we have*

(i) For $p = 1$, we have $t_k = k + 1$.

(ii) For $p = 2$, we have $t_k \geq (k + 1)/2$. Moreover, one has $t_{k+1}^2 \leq t_k^2 + 2t_k$.

Proof. The first item is immediate as well as the first part of the second item (see, for instance, [2, Lemma 4.3]). Moreover, note that by the given recursion $t_{k+1}^2 - t_k^2 = t_{k+1}$, one also has $t_{k+1} = t_k + t_{k+1}/(t_k + t_{k+1})$, and hence it follows that $t_{k+1}^2 = t_k^2 + t_{k+1} \leq t_k^2 + 2t_k$, since $t_k \geq 1$. \square

In the rest of this section we aim at proving rate of convergence results of FLAG. We will focus, in this paper, on iteration complexity of Lagrangian-based methods, using the following two classical measures at a given iterate $k \in \mathbb{N}$:

(i) *Function values gap* in terms of $\Psi(x^k) - \Psi(x^*)$.

(ii) *Feasibility violation* of the constraints of problem (P) in terms of $\|\mathcal{A}x^k - b\|$.

When discussing the measures mentioned above for the iterates produced by FLAG, we will also distinguish our results between rates expressed in terms of the original produced sequence or of the ergodic sequence, as defined below in subsection 4.3.

4.2 Non-Ergodic Rate of Convergence

Throughout the rest of this section, recalling the fact that the set of optimal dual solutions is compact (see Section 2), we take $c > 0$ to be a constant for which $c \geq 2\|y^*\|$, where y^* is an optimal solution of the dual problem.

We are now ready to prove our main result in the strongly convex setting (*i.e.*, $p = 2$) on the sequence itself.

Theorem 4.1 (A fast non-ergodic function values and feasibility violation rates). *Let $\{(x^k, z^k, y^k)\}_{k \in \mathbb{N}}$ be a sequence generated by FLAG. Suppose that $\sigma > 0$ and $0 \preceq P \preceq (\sigma/2)I_n$. Then, for any optimal solution x^* of problem (P), we have*

$$\Psi(x^N) - \Psi(x^*) \leq \frac{B_{\rho,c}(x^*)}{2N^2}, \quad (4.8)$$

$$\|\mathcal{A}x^N - b\| \leq \frac{B_{\rho,c}(x^*)}{cN^2}, \quad (4.9)$$

where $B_{\rho,c}(x^*) := 4\left(\|x^* - z^0\|_P^2 + \frac{1}{\mu\rho}(\|y^0\| + c)^2\right)$.

Proof. From Lemma 4.2, after we substitute $\xi = x^*$ and $p = 2$ (recall that $\rho_k = \rho t_k$ and $\tau_k = t_k$), we obtain

$$t_k^2 s_{k+1} - t_{k-1}^2 s_k \leq t_k^2 \Delta_P(x^*, z^k, z^{k+1}) - \frac{t_k \sigma}{2} \|x^* - z^{k+1}\|^2 + \frac{1}{\mu\rho} \Delta(\eta, y^k, y^{k+1}).$$

Using the definition of Δ_P and Δ (see (2.7)) and the fact that $P \preceq (\sigma/2)I_n$ we have that

$$\begin{aligned} t_k^2 s_{k+1} - t_{k-1}^2 s_k &\leq \frac{t_k^2}{2} \left(\|x^* - z^k\|_P^2 - \|x^* - z^{k+1}\|_P^2 \right) - t_k \|x^* - z^{k+1}\|_P^2 + \frac{1}{2\mu\rho} \left(\|\eta - y^k\|^2 - \|\eta - y^{k+1}\|^2 \right) \\ &\leq \frac{1}{2} \left(t_k^2 \|x^* - z^k\|_P^2 - t_{k+1}^2 \|x^* - z^{k+1}\|_P^2 \right) + \frac{1}{2\mu\rho} \left(\|\eta - y^k\|^2 - \|\eta - y^{k+1}\|^2 \right), \end{aligned}$$

where the last inequality follows from Lemma 4.3(ii). Summing this inequality for all $k = 0, 1, \dots, N-1$ (recall that $t_{-1} = 0$ and $t_0 = 1$), it follows (where we use the definition of s_N , see (4.2), for the first inequality below)

$$t_{N-1}^2 (\Psi(x^N) - \Psi(x^*) + \langle \eta, \mathcal{A}x^N - b \rangle) \leq t_{N-1}^2 s_N - t_{-1}^2 s_0 \leq \frac{1}{2} \|x^* - z^0\|_P^2 + \frac{1}{2\mu\rho} \|\eta - y^0\|^2.$$

Therefore, by taking the maximum of both sides over $\|\eta\| \leq c$, proves that (recall that $t_{N-1}^2 \geq N^2/4$ using Lemma 4.3(ii))

$$\Psi(x^N) - \Psi(x^*) + c \|\mathcal{A}x^N - b\| \leq \frac{B_{\rho,c}(x^*)}{2N^2} := \beta,$$

from which the estimate (4.8) immediately follows. Moreover, since (x^*, y^*) is a saddle point of problem (P), with $c \geq 2\|y^*\|$, using the above inequality it follows that

$$c \|\mathcal{A}x^N - b\| \leq \Psi(x^*) - \Psi(x^N) + \beta \leq \langle y^*, \mathcal{A}x^N - b \rangle + \beta \leq \frac{c}{2} \|\mathcal{A}x^N - b\| + \beta,$$

and the desired estimate (4.9) is proved. \square

Our main result in the convex setting (*i.e.*, $\sigma = 0$ and $p = 1$) on the sequence itself is recorded next.

Theorem 4.2 (A non-ergodic function values and feasibility violation rates). *Let $\{(x^k, z^k, y^k)\}_{k \in \mathbb{N}}$ be a sequence generated by FLAG and suppose that $\sigma = 0$. Then, for any optimal solution x^* of problem (P), we have*

$$\Psi(x^N) - \Psi(x^*) \leq \frac{B_{\rho,c}(x^*)}{2N}, \quad (4.10)$$

$$\|\mathcal{A}x^N - b\| \leq \frac{B_{\rho,c}(x^*)}{cN}, \quad (4.11)$$

where $B_{\rho,c}(x^*) := 2\left(\|x^* - z^0\|_P^2 + \frac{1}{\mu\rho}(\|y^0\| + c)^2\right)$.

Proof. From Lemma 4.2, after we substitute $\xi = x^*$, $\sigma = 0$ and $p = 1$ (recall that $\rho_k = \rho$ and $\tau_k = t_k^{-1}$), we obtain

$$t_k s_{k+1} - t_{k-1} s_k \leq t_k^{-1} \Delta_P \left(x^*, z^k, z^{k+1} \right) + \frac{1}{\mu \rho} \Delta \left(\eta, y^k, y^{k+1} \right).$$

Using the definition of Δ_P and Δ (see (2.7)) we have that

$$\begin{aligned} t_k s_{k+1} - t_{k-1} s_k &\leq \frac{1}{2t_k} \left(\|x^* - z^k\|_P^2 - \|x^* - z^{k+1}\|_P^2 \right) + \frac{1}{2\mu\rho} \left(\|\eta - y^k\|^2 - \|\eta - y^{k+1}\|^2 \right) \\ &\leq \frac{1}{2} \left(\frac{1}{t_k} \|x^* - z^k\|_P^2 - \frac{1}{t_{k+1}} \|x^* - z^{k+1}\|_P^2 \right) + \frac{1}{2\mu\rho} \left(\|\eta - y^k\|^2 - \|\eta - y^{k+1}\|^2 \right), \end{aligned}$$

where the second inequality follows from the fact that $-t_k^{-1} \leq -t_{k+1}^{-1}$. Summing this inequality for all $k = 0, 1, \dots, N-1$ (recall that $t_{-1} = 0$ and $t_0 = 1$) it follows

$$t_{N-1} \left(\Psi(x^N) - \Psi(x^*) + \langle \eta, \mathcal{A}x^N - b \rangle \right) \leq t_{N-1} s_N - t_{-1} s_0 = t_{N-1} s_N \leq \frac{1}{2} \|x^* - z^0\|_P^2 + \frac{1}{2\mu\rho} \|\eta - y^0\|^2.$$

Therefore, by taking the maximum of both sides over $\|\eta\| \leq c$, proves that (recall that $t_{N-1} = N$ using Lemma 4.3(i))

$$\Psi(x^N) - \Psi(x^*) + c \|\mathcal{A}x^N - b\| \leq \frac{B_{\rho,c}(x^*)}{2N},$$

and the desired results follow exactly as done at the end of the proof of Theorem 4.1. \square

4.3 Classical Ergodic Rate of Convergence

Although deriving ergodic type results was not our primary goal, to conclude this section we further illustrate the versatility of our framework by providing the fast and classical *ergodic* rate of convergence results. Before doing so, it should be noted that in the setting of producing such weaker results of ergodic type, two main features of our main framework FLAG are simplified as follows:

- (i) The auxiliary sequence $\{\lambda^k\}_{k \in \mathbb{N}}$ is simply replaced by the multiplier sequence $\{y^k\}_{k \in \mathbb{N}}$, i.e., the sequence $\{x^k\}_{k \in \mathbb{N}}$ plays no role. Thus, $\{z^k\}_{k \in \mathbb{N}}$ is the only primal sequence, so that the main iteration (3.5) reads now $z^{k+1} \in \text{Prim}_k(z^k, y^k)$.
- (ii) Only the parameter ρ_k and τ_k are still used to include both the fast and classical results.

According to the above, since we replace the auxiliary multiplier λ^k with y^k , it is easy to verify from the proof of Lemma 4.1 (by combining (4.3) and (4.4) there) that in this case we have

$$\mathcal{L}_{\gamma_k} \left(z^{k+1}, \eta \right) - \mathcal{L}_{\gamma_k} \left(\xi, \eta \right) \leq \tau_k \Delta_P \left(\xi, z^k, z^{k+1} \right) - \frac{\sigma}{2} \left\| \xi - z^{k+1} \right\|^2 + \frac{1}{\mu \rho_k} \Delta \left(\eta, y^k, y^{k+1} \right), \quad (4.12)$$

where $\gamma_k := (1 + \delta - \mu) \rho_k \geq 0$. Therefore, in this case we can consider scaling parameters μ in the larger interval $(0, 1 + \delta]$, instead of $(0, \delta]$ as stated in the original version of FLAG above.

We begin with the result in the strongly convex setting (*i.e.*, $p = 2$).

Corollary 4.1 (A fast ergodic function values and feasibility violation rates). *Let $\{(z^k, y^k)\}_{k \in \mathbb{N}}$ be a sequence generated by FLAG. Suppose that $\sigma > 0$ and $0 \preceq P \preceq (\sigma/2) I_n$. Then, for any optimal solution z^* of problem (P), the following holds for the ergodic sequence $\bar{z}^N = t_{N-1}^{-2} \sum_{k=0}^{N-1} t_k z^{k+1}$*

$$\Psi(\bar{z}^N) - \Psi(z^*) \leq \frac{B_{\rho,c}(z^*)}{2N^2}, \quad (4.13)$$

$$\|\mathcal{A}\bar{z}^N - b\| \leq \frac{B_{\rho,c}(z^*)}{cN^2}, \quad (4.14)$$

where $B_{\rho,c}(z^*) := 4 \left(\|z^* - z^0\|_P^2 + \frac{1}{\mu\rho} (\|y^0\| + c)^2 \right)$.

Proof. Substituting in (4.12) $\xi = z^*$ and multiplying both sides by t_k yields (recall that $\rho_k = \rho t_k$ and $\tau_k = t_k$)

$$t_k \left(\Psi \left(z^{k+1} \right) - \Psi \left(z^* \right) + \left\langle \eta, \mathcal{A}z^{k+1} - b \right\rangle \right) \leq t_k^2 \Delta_P \left(z^*, z^k, z^{k+1} \right) - \frac{t_k \sigma}{2} \left\| z^* - z^{k+1} \right\|^2 + \frac{1}{\mu \rho} \Delta \left(\eta, y^k, y^{k+1} \right).$$

where we have omitted the non-negative augmented term, which is present only in $\mathcal{L}_{\gamma_k} \left(z^{k+1}, \eta \right)$. Using the definition of Δ_P and Δ (see (2.7)) and the fact that $P \preceq (\sigma/2) I_n$ we thus obtain

$$\begin{aligned} t_k \left(\Psi \left(z^{k+1} \right) - \Psi \left(z^* \right) + \left\langle \eta, \mathcal{A}z^{k+1} - b \right\rangle \right) &\leq \frac{t_k^2}{2} \left(\left\| z^* - z^k \right\|_P^2 - \left\| z^* - z^{k+1} \right\|_P^2 \right) - t_k \left\| z^* - z^{k+1} \right\|_P^2 \\ &\quad + \frac{1}{\mu \rho} \Delta \left(\eta, y^k, y^{k+1} \right) \\ &\leq \frac{1}{2} \left(t_k^2 \left\| z^* - z^k \right\|_P^2 - t_{k+1}^2 \left\| z^* - z^{k+1} \right\|_P^2 \right) \\ &\quad + \frac{1}{2\mu\rho} \left(\left\| \eta - y^k \right\|^2 - \left\| \eta - y^{k+1} \right\|^2 \right), \end{aligned}$$

where the last inequality follows from Lemma 4.3(ii). Summing this inequality for all $k = 0, 1, \dots, N-1$ it follows that (recall that $t_0 = 1$)

$$\sum_{k=0}^{N-1} t_k \left(\Psi \left(z^{k+1} \right) - \Psi \left(z^* \right) + \left\langle \eta, \mathcal{A}z^{k+1} - b \right\rangle \right) \leq \frac{1}{2} \left\| z^* - z^0 \right\|_P^2 + \frac{1}{2\mu\rho} \left\| \eta - y^0 \right\|^2.$$

Using the relation $t_k = t_k^2 - t_{k-1}^2$ and the definition of \bar{z}^N , it easily follows (recall that $t_{-1} = 0$)

$$\sum_{k=0}^{N-1} t_k = t_{N-1}^2 \quad \text{and} \quad \sum_{k=0}^{N-1} t_k \left\langle \eta, \mathcal{A}z^{k+1} - b \right\rangle = t_{N-1}^2 \left\langle \eta, \mathcal{A}\bar{z}^N - b \right\rangle.$$

Therefore, thanks to Jensen's inequality for the convex function Ψ , and these relations, it follows from the inequality above that

$$t_{N-1}^2 \left(\Psi \left(\bar{z}^N \right) - \Psi \left(z^* \right) + \left\langle \eta, \mathcal{A}\bar{z}^N - b \right\rangle \right) \leq \frac{1}{2} \left\| z^* - z^0 \right\|_P^2 + \frac{1}{2\mu\rho} \left\| \eta - y^0 \right\|^2.$$

Taking the maximum of both sides over $\|\eta\| \leq c$, proves that (recall that $t_{N-1}^2 \geq N^2/4$ using Lemma 4.3(ii))

$$\Psi \left(\bar{z}^N \right) - \Psi \left(z^* \right) + c \left\| \mathcal{A}\bar{z}^N - b \right\| \leq \frac{B_{\rho,c} \left(z^* \right)}{N^2}.$$

The required results now follow as done at the end of the proof of Theorem 4.1. \square

The classical ergodic rate of convergence result in the convex setting (*i.e.*, $\sigma = 0$ and $p = 1$) is recorded next.

Corollary 4.2 (An ergodic function values and feasibility violation rates). *Let $\{(z^k, y^k)\}_{k \in \mathbb{N}}$ be a sequence generated by FLAG with $\sigma = 0$ and $t_k = 1$ for all $k \in \mathbb{N}$. Then, for any optimal solution z^* of problem (P), the following holds for the ergodic sequence $\bar{z}^N = N^{-1} \sum_{k=0}^{N-1} z^{k+1}$*

$$\Psi \left(\bar{z}^N \right) - \Psi \left(z^* \right) \leq \frac{B_{\rho,c} \left(z^* \right)}{2N}, \tag{4.15}$$

$$\left\| \mathcal{A}\bar{z}^N - b \right\| \leq \frac{B_{\rho,c} \left(z^* \right)}{cN}, \tag{4.16}$$

where $B_{\rho,c} \left(z^* \right) := 2 \left(\left\| z^* - x^0 \right\|_P^2 + \frac{1}{\mu\rho} \left(\left\| y^0 \right\| + c \right)^2 \right)$.

Proof. The result trivially follows from Corollary 4.1 since $t_k = 1$ for all $k \in \mathbb{N}$. \square

As discussed in the introduction, similar ergodic $O(1/N)$ results have been obtained in many works in the literature and on many variants of Lagrangian-based methods (see, for instance, [6, 14, 23, 22, 17, 18] and references therein).

5 Applications: Nice Primal Algorithmic Maps and Their FLAG

In this section we will show that several well-known iconic Lagrangian-based methods *all admit* nice primal algorithmic maps. Recall that the notion of nice primal algorithmic map deals only with the update of the primal variable(s), while the multiplier update is the same for all Lagrangian-based methods, which in FLAG, is simply given by:

$$y^+ = y + \mu \rho_t (Az^+ - b), \quad (5.1)$$

where $\rho_t = \rho$ in the convex case or $\rho_t = \rho t$ in the strongly convex case. Thus, given any Lagrangian-based method, we only need to consider the corresponding primal algorithmic map $\text{Prim}_t(\cdot)$ and use it in FLAG to produce a new variant of this algorithm with faster *non-ergodic* rate of convergence: $O(1/N^2)$ when the objective is strongly convex and $O(1/N)$ when the objective is convex, as proved in Theorems 4.1 and 4.2, respectively.

It is well-known that a major drawback in Lagrangian-based methods is the presence of the quadratic augmented penalty term, which often renders a primal minimization step as a difficult one. This difficulty can be overcome by *linearizing the quadratic* expression around a given current iterate. An important point to stress in this respect, is that in all of the instances below, we will show that we can always consider linearizing the quadratic penalty term, and prove that it generates a corresponding *nice* primal algorithmic map $\text{Prim}_t(\cdot)$.

Before presenting these results, for a potential user who is interested in developing and/or analyzing a Lagrangian-based method for solving a certain convex optimization problem, we offer the following informal recipe for obtaining faster rates of the designed method.

A recipe for finding the rate of convergence of Lagrangian-based methods.

- (i) Formulate the problem at hand via model (P), *i.e.*, identify the relevant problem data $[\Psi, \mathcal{A}, b, \sigma]$, where the value of the modulus of convexity σ will infer the value of p (1 or 2) to be used in FLAG, and which determine the type of rate that can be achieved (classical or fast).
- (ii) Define the desired iterative step(s) of the primal algorithmic map $\text{Prim}_t(\cdot)$ applied on the augmented Lagrangian $\mathcal{L}_{\rho_t}(\cdot)$ of model (P).
- (iii) Show that the defined primal algorithmic map is nice according to Definition 3.1 (*i.e.*, determine the parameter δ and the matrices P and Q) using information on the involved objective function $\Psi(\cdot)$, the linear mapping \mathcal{A} and the iterative step(s) defining $\text{Prim}_t(\cdot)$.
- (iv) Apply Theorem 4.1 (if $p = 2$) or Theorem 4.2 (if $p = 1$) to obtain a faster non-ergodic rate of convergence for the designed method.

To proceed, we recall the following well-known result that will be useful below. For completeness we include its simple proof.

Lemma 5.1 (Composite proximal inequality). *Let $\varphi : \mathbb{R}^d \rightarrow (-\infty, +\infty]$ be a proper, lower-semi continuous and σ -strongly convex function ($\sigma \geq 0$) and let $c : \mathbb{R}^d \rightarrow \mathbb{R}$ be a differentiable function. Given a matrix $W \in \mathbb{S}_+^d$, we define:*

$$w^+ \in \operatorname{argmin}_{\xi \in \mathbb{R}^d} \left\{ \varphi(\xi) + c(\xi) + \frac{1}{2} \|\xi - w\|_W^2 \right\}.$$

Then, for any $\xi \in \mathbb{R}^d$, we have

$$\varphi(w^+) - \varphi(\xi) + \langle \nabla c(w^+), w^+ - \xi \rangle \leq \frac{1}{2} \left(\|\xi - w\|_W^2 - \|\xi - w^+\|_W^2 - \|w^+ - w\|_W^2 \right) - \frac{\sigma}{2} \|\xi - w^+\|^2.$$

Proof. Writing the first-order optimality condition of the optimization problem yields, for any $\xi \in \mathbb{R}^d$

$$\mathbf{0} \in \partial\varphi(w^+) + \nabla c(w^+) + W(w^+ - w),$$

which means that

$$\nabla c(w^+) + W(w^+ - w) \equiv -\xi^+ \in -\partial\varphi(w^+).$$

On the other hand, using the σ -strong convexity of φ , for all $\xi \in \mathbb{R}^d$, we have

$$\varphi(w^+) - \varphi(\xi) \leq \langle -\xi^+, \xi - w^+ \rangle - \frac{\sigma}{2} \|\xi - w^+\|^2 = \langle \nabla c(w^+) + W(w^+ - w), \xi - w^+ \rangle - \frac{\sigma}{2} \|\xi - w^+\|^2.$$

By rearranging terms and using the Pythagoras identity (4.1) we obtain the desired result. \square

5.1 Methods for the Basic Model (P)

As we discussed in Section 3, the most basic example of a primal algorithmic map is the mapping $\text{Prim}_t(\cdot)$ that performs an *exact minimization* on the augmented Lagrangian itself, or a proximal variant of it.

5.1.1 Example I: Augmented Lagrangian

For the classical (Proximal) Augmented Lagrangian method, the primal step of FLAG is given, in this case, by

$$z^+ = \text{Prim}_t(z, \lambda) \equiv \operatorname{argmin}_{\xi} \left\{ \Psi(\xi) + \langle \lambda, \mathcal{A}\xi - b \rangle + \frac{\rho_t}{2} \|\mathcal{A}\xi - b\|^2 + \frac{\tau_t}{2} \|\xi - z\|_M^2 \right\}, \quad (5.2)$$

where $M \succeq 0$ is a weight matrix defining the additional proximal term. Therefore, if $M \neq 0$ then the algorithmic map $\text{Prim}_t(\cdot)$ can be seen as a *proximal minimization* applied on the augmented Lagrangian $\mathcal{L}_{\rho_t}(\cdot, y)$, for a fixed $y \in \mathbb{R}^m$. In this case, we can show the following result, which simple proof is omitted; instead see the proof of Lemma 5.3 below (which essentially includes this lemma with an adequate choice of the involved matrices).

Lemma 5.2 (Proximal AL is nice). *Let $M \succeq 0$, the algorithmic map $\text{Prim}_t(\cdot)$ defined in (5.2) is nice with $\delta = 1$ and $P = Q = M \in \mathbb{S}_+^n$.*

5.1.2 Example II: Proximal Linearized Augmented Lagrangian

It should be noted that the Proximal AL method mentioned above requires to solve an optimization problem, which is usually not easy (even when $\Psi(\cdot)$ is prox-friendly) mainly because of the *quadratic augmented* term. One classical way to overcome this difficulty is to *linearize* the augmented term around the current iterate. Therefore, by choosing the algorithmic map $\text{Prim}_t(\cdot)$ to be a *proximal gradient* step applied on the augmented Lagrangian \mathcal{L}_{ρ_t} yields the *Proximal Linearized AL Method*, see, e.g., [14, 23]. In this case, we prove the following result.

Lemma 5.3 (Proximal Linearized AL is nice). *Let $M \succ 0$, then the algorithmic map defined by*

$$z^+ = \text{Prim}_t(z, \lambda) \equiv \operatorname{argmin}_{\xi} \left\{ \Psi(\xi) + \langle \lambda, \mathcal{A}\xi - b \rangle + \rho_t \langle \mathcal{A}z - b, \mathcal{A}\xi \rangle + \frac{\tau_t}{2} \|\xi - z\|_M^2 \right\},$$

is nice with $\delta = 1$ and $P = Q = M - \rho\mathcal{A}^T\mathcal{A} \succeq 0$.

Proof. Applying Lemma 5.1 with the function $\varphi(\cdot) = \Psi(\cdot)$, $c(\xi) = \langle \lambda + \rho_t(\mathcal{A}z - b), \mathcal{A}\xi \rangle$ and $W = \tau_t M$ yields for any $\xi \in \mathbb{R}^n$ that

$$\Psi(z^+) - \Psi(\xi) + \langle \lambda + \rho_t(\mathcal{A}z - b), \mathcal{A}z^+ - \mathcal{A}\xi \rangle \leq \tau_t \Delta_M(\xi, z, z^+) - \frac{\tau_t}{2} \|z^+ - z\|_M^2 - \frac{\sigma}{2} \|\xi - z^+\|^2. \quad (5.3)$$

Taking $\xi \in \mathcal{F}$ yields that $\mathcal{A}\xi = b$, and thus we can estimate the following term as:

$$\begin{aligned} \langle \mathcal{A}z - b, \mathcal{A}z^+ - \mathcal{A}\xi \rangle &= \langle \mathcal{A}z^+ - b, \mathcal{A}z^+ - b \rangle + \langle \mathcal{A}z - \mathcal{A}z^+, \mathcal{A}z^+ - \mathcal{A}\xi \rangle \\ &= \|\mathcal{A}z^+ - b\|^2 + \Delta_{\mathcal{A}^T \mathcal{A}}(\xi, z, z^+) - \frac{1}{2} \|z^+ - z\|_{\mathcal{A}^T \mathcal{A}}^2. \end{aligned} \quad (5.4)$$

Using (5.4) in (5.3) we obtain,

$$\begin{aligned} \Psi(z^+) - \Psi(\xi) + \langle \lambda, \mathcal{A}z^+ - b \rangle + \frac{\rho_t}{2} \|\mathcal{A}z^+ - b\|^2 &\leq \tau_t \Delta_M(\xi, z, z^+) - \frac{\tau_t}{2} \|z^+ - z\|_M^2 - \frac{\sigma}{2} \|\xi - z^+\|^2 \\ &\quad - \rho_t \Delta_{\mathcal{A}^T \mathcal{A}}(\xi, z, z^+) + \frac{\rho_t}{2} \|z^+ - z\|_{\mathcal{A}^T \mathcal{A}}^2 - \frac{\rho_t}{2} \|\mathcal{A}z^+ - b\|^2. \end{aligned}$$

Hence, the nice property of $\text{Prim}_t(\cdot)$ follows from Definition 3.1, where we also used the fact that $\rho_t \geq \rho\tau_t$ (if $\sigma > 0$ it is obvious and if $\sigma = 0$ it follows since $t \geq 1$). \square

5.2 Methods for Block Models

The Lagrangian-based algorithms to be presented below are designed to tackle the block model as discussed in Example 2.2:

$$\min_{(u,v) \in \mathbb{R}^n} \{f(u) + g(v) : Au + Bv = b\},$$

which fits the general model (P) where the linear mapping \mathcal{A} is defined by $\mathcal{A}z = Au + Bv$ and $\Psi(z) = f(u) + g(v)$ where $x = (u^T, v^T)^T \in \mathbb{R}^p \times \mathbb{R}^q$. To warrant the σ -strong convexity of Ψ with $\sigma \geq 0$, it is sufficient to require σ -strong convexity of only one function f or g , *i.e.*, only in one of the 2 blocks. Thus, without loss of generality, throughout the rest of this section, the σ -strong convexity is assumed in the function g .

The notion of nice primal algorithmic map is flexible and easily adapt to the block setting as stated below.

Definition 5.1 (Nice primal algorithmic map - Block version). Given $(z, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m$ and the parameters $\rho, t > 0$. A primal algorithmic map $\text{Prim}_t : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$, which is applied on the augmented Lagrangian $\mathcal{L}_{\rho t}(z, \lambda)$, that generates $z^+ = (u^+, v^+)$ via $z^+ \in \text{Prim}_t(z, \lambda)$, is called *nice*, if there exist a parameter $\delta \in (0, 1]$ and matrices $P_1, Q_1 \in \mathbb{S}_+^p$ and $P_2, Q_2 \in \mathbb{S}_+^q$ with $P = (P_1, P_2)$ and $Q = (Q_1, Q_2)$, such that for any $(\xi_1, \xi_2) \in \mathcal{F}$ we have

$$\begin{aligned} \mathcal{L}_{\rho t}(z^+, \lambda) - \mathcal{L}_{\rho t}(\xi, \lambda) &\leq \frac{1}{t} \Delta_{P_1}(\xi_1, u, u^+) - \frac{1}{2t} \|u^+ - u\|_{Q_1}^2 + \tau_t \Delta_{P_2}(\xi_2, v, v^+) - \frac{\tau_t}{2} \|v^+ - v\|_{Q_2}^2 \\ &\quad - \frac{\sigma}{2} \|\xi_2 - v^+\|^2 - \frac{\delta \rho t}{2} \|\mathcal{A}z^+ - b\|^2, \end{aligned}$$

where $\tau_t = t^{-1}$ and $\rho_t = \rho$ (when $\sigma = 0$) or $\tau_t = t$ and $\rho_t = \rho t$ (when $\sigma > 0$).

The adaptation made in this variant is very simple and appears only in the terms that concerns with the parameter τ_t . Indeed, recall that the choice of τ_t depends on the convexity or σ -strong convexity of the given objective function $\Psi(\cdot)$ through σ . Therefore, with respect to the block v (which correspond to the possibly strongly convex part) we use τ_t (to include both the convex and strongly convex cases). On the other hand with respect to the other block, where only convexity is assumed, we fixed $\tau_t = t^{-1}$ (which corresponds to the case where $\sigma = 0$).

Furthermore, equipped with Definition 5.1, it is then straightforward to adapt the two main pillars of the analysis given in Lemma 4.1 and 4.2, to the block model, and hence to warrant the corresponding rate of convergence results established in Sections 4.2 and 4.3; we omit the details and proofs which are essentially identical to the original ones.

5.2.1 Example III: Alternating Direction Method of Multipliers (ADMM)

Thanks to the split nature of the block model it will be worth to exploit this structure by using an Alternating Minimization-based algorithm. The ADMM [12, 11, 10] can be obtained from the Augmented Lagrangian method (see Section 5.1.1) by applying the Alternating Minimization idea on the primal updating step, which consists of minimizing the augmented Lagrangian jointly with respect to the primal variables (u, v) . In order to use FLAG in this case, all we need is to verify that the corresponding algorithmic map $(u^+, v^+) \equiv z^+ \in \text{Prim}_t(z, \lambda)$ is nice according to Definition 5.1, as recorded next without a proof; instead see the proof of the Linearized ADMM below.

Lemma 5.4 (Proximal ADMM is nice). *Let $M_1, M_2 \succeq 0$, the primal algorithmic map $\text{Prim}_t(\cdot)$ defined by*

$$\begin{aligned} u^+ &= \operatorname{argmin}_{\xi_1} \left\{ f(\xi_1) + \langle \lambda, A\xi_1 + Bv - b \rangle + \frac{\rho_t}{2} \|A\xi_1 + Bv - b\|^2 + \frac{1}{2t} \|\xi_1 - u\|_{M_1}^2 \right\}, \\ v^+ &= \operatorname{argmin}_{\xi_2} \left\{ g(\xi_2) + \langle \lambda, Au^+ + B\xi_2 - b \rangle + \frac{\rho_t}{2} \|Au^+ + B\xi_2 - b\|^2 + \frac{\tau_t}{2} \|\xi_2 - v\|_{M_2}^2 \right\}, \end{aligned}$$

is nice with $\delta = 1 - \rho\lambda_{\max}(B^T B) / (\rho\lambda_{\max}(B^T B) + \lambda_{\min}(M_2))$, $P_1 = M_1$, $P_2 = M_2 + \rho B^T B$, $Q_1 = M_1$ and $Q_2 = 0$.

5.2.2 Example IV: Proximal Linearized Alternating Direction Method of Multipliers

Now, we would like to show that also the Proximal Linearized ADMM, which consists of a linearization of the augmented term around the current iteration, also admits a nice primal algorithmic map as recorded next. It should be noted that here we consider only partial linearization (with respect to the second block v), since only this block corresponds to a strongly convex function. Clearly, if we additionally assume that also f is strongly convex then the primal algorithmic map corresponds to fully Proximal Linearized ADMM can be easily proven to be nice in a similar way, hence we omit the details.

Lemma 5.5 (Proximal Linearized ADMM is nice). *Let $M_1, M_2 \succeq 0$ with $\lambda_{\min}(M_2) > \rho\lambda_{\max}(B^T B)$, the primal algorithmic map $\text{Prim}_t(\cdot)$ defined by*

$$u^+ = \operatorname{argmin}_{\xi_1} \left\{ f(\xi_1) + \langle \lambda, A\xi_1 + Bv - b \rangle + \frac{\rho_t}{2} \|A\xi_1 + Bv - b\|^2 + \frac{1}{2t} \|\xi_1 - u\|_{M_1}^2 \right\}, \quad (5.5)$$

$$v^+ = \operatorname{argmin}_{\xi_2} \left\{ g(\xi_2) + \langle \lambda, Au^+ + B\xi_2 - b \rangle + \rho_t \langle Au^+ + Bv - b, B\xi_2 \rangle + \frac{\tau_t}{2} \|\xi_2 - v\|_{M_2}^2 \right\}, \quad (5.6)$$

is nice with $\delta = 1 - \rho\lambda_{\max}(B^T B) / \lambda_{\min}(M_2)$, $P_1 = Q_1 = M_1$, $P_2 = M_2$ and $Q_2 = 0$.

Proof. Applying Lemma 5.1 with $\varphi(\xi) = f(\xi_1)$, $c(\xi_1) = \langle \lambda, A\xi_1 \rangle + (\rho_t/2) \|A\xi_1 + Bv - b\|^2$ and $W = t^{-1}M_1$ yields for any $\xi_1 \in \mathbb{R}^p$ that

$$f(u^+) - f(\xi_1) + \langle \lambda + \rho_t(Au^+ + Bv - b), Au^+ - A\xi_1 \rangle \leq \text{RHS}_1,$$

where

$$\text{RHS}_1 = \frac{1}{t} \Delta_{M_1}(\xi_1, u, u^+) - \frac{1}{2t} \|u^+ - u\|_{M_1}^2.$$

Similarly, by applying Lemma 5.1 with $\varphi(\xi_2) = g(\xi_2)$, $c(\xi_2) = \langle \lambda, B\xi_2 \rangle + \rho_t \langle Au^+ + Bv - b, B\xi_2 \rangle$ and $W = \tau_t M_2$ yields for any $\xi \in \mathbb{R}^q$ that

$$g(v^+) - g(\xi_2) + \langle \lambda + \rho_t(Au^+ + Bv - b), Bv^+ - B\xi_2 \rangle \leq \text{RHS}_2,$$

where

$$\text{RHS}_2 = \tau_t \Delta_{M_2}(\xi_2, v, v^+) - \frac{\tau_t}{2} \|v^+ - v\|_{M_2}^2 - \frac{\sigma}{2} \|\xi_2 - v^+\|^2.$$

By adding these two inequalities and using the fact that $\xi := (\xi_1, \xi_2) \in \mathcal{F}$, that is $A\xi_1 + B\xi_2 = b$, we get

$$\Psi(z^+) - \Psi(\xi) + \langle \lambda, \mathcal{A}z^+ - b \rangle + \rho_t \langle Au^+ + Bv - b, \mathcal{A}z^+ - b \rangle \leq \text{RHS}_1 + \text{RHS}_2, \quad (5.7)$$

where we used that $Au^+ + Bv^+ = \mathcal{A}z^+$ and $z = (u, v)$, $\xi = (\xi_1, \xi_2)$, $z^+ = (u^+, v^+)$. In addition, we have

$$\begin{aligned} \langle Au^+ + Bv - b, \mathcal{A}z^+ - b \rangle &= \langle Au^+ + Bv^+ - b, \mathcal{A}z^+ - b \rangle + \langle Bv - Bv^+, \mathcal{A}z^+ - b \rangle \\ &= \|\mathcal{A}z^+ - b\|^2 + \langle Bv - Bv^+, \mathcal{A}z^+ - b \rangle \\ &\geq \|\mathcal{A}z^+ - b\|^2 - \frac{1}{2\alpha} \|Bv^+ - Bv\|^2 - \frac{\alpha}{2} \|\mathcal{A}z^+ - b\|^2, \end{aligned} \quad (5.8)$$

where we used again that $Au^+ + Bv^+ = \mathcal{A}z^+$, and the simple fact:

$$\langle c, d \rangle \geq -\frac{1}{2\alpha} \|c\|^2 - \frac{\alpha}{2} \|d\|^2, \quad \forall c, d \in \mathbb{R}^n, \forall \alpha \in (0, 1). \quad (5.9)$$

By combining (5.7) with ρ_t times (5.8), we obtain, for all $\alpha < 1$, that

$$\Psi(z^+) - \Psi(\xi) + \langle \lambda, \mathcal{A}z^+ - b \rangle + \frac{\rho_t}{2} \|\mathcal{A}z^+ - b\|^2 \leq \text{RHS}_1 + \text{RHS}_2 + \frac{\rho_t}{2\alpha} \|Bv^+ - Bv\|^2 - \frac{\delta\rho_t}{2} \|\mathcal{A}z^+ - b\|^2,$$

where $\delta = 1 - \alpha$. In order to determine the best parameter $\alpha < 1$ we consider the following terms of the previous inequality (using the definition of RHS_1 and omitting ρ_t):

$$\frac{1}{\alpha} \|v^+ - v\|_{B^T B}^2 - \frac{1}{\rho} \|v^+ - v\|_{M_2}^2 \leq \left(\frac{1}{\alpha} \lambda_{\max}(B^T B) - \frac{1}{\rho} \lambda_{\min}(M_2) \right) \|v^+ - v\|^2.$$

Therefore, by taking $\alpha = \rho \lambda_{\max}(B^T B) / \lambda_{\min}(M_2)$, the desired result is proved. \square

Remark 5.1. As discussed in the introduction, we would like to mention again [17], where a non-ergodic $O(1/N)$ rate of convergence result for the Linearized ADMM is proven. It should be noted that their proposed algorithm appears very different from our scheme. In the convex case (*i.e.*, when $p = 1$), FLAG with the primal algorithmic map $\text{Prim}_t(\cdot)$ chosen be the Linearized ADMM, produce a scheme sharing the non-ergodic $O(1/N)$ rate of convergence (see Theorem 4.2).

5.2.3 Example V: Chambolle-Pock Method

The Chambolle-Pock (CP) method [6] is an example of a well-known Lagrangian-based method that was designed to tackle the linear composite model (*cf.* Example 2.1), which can be seen as a particular case of the block model discuss above with $A = I_p$ (where I_p is the $p \times p$ identity matrix), *i.e.*,

$$\min_{(u,v) \in \mathbb{R}^n} \{f(u) + g(v) : u + Bv = b\}.$$

Following [23, Section 3.3.], the CP Method can be seen as special case of the Proximal Linearized ADMM. More precisely, the primal algorithmic map Prim_t that defines the CP method is given by

$$u^+ = \text{argmin}_{\xi_1} \left\{ f(\xi_1) + \langle \lambda, \xi_1 + Bv - b \rangle + \frac{\rho_t}{2} \|\xi_1 + Bv - b\|^2 \right\}, \quad (5.10)$$

$$v^+ = \text{argmin}_{\xi_2} \left\{ g(\xi_2) + \langle \lambda, u^+ + B\xi_2 - b \rangle + \rho_t \langle u^+ + Bv - b, B\xi_2 \rangle + \frac{\tau_t}{2\alpha} \|\xi_2 - v\|^2 \right\}, \quad (5.11)$$

where $\alpha > 0$ is a proximal parameter (*i.e.*, in this case $M_1 \equiv 0$ and $M_2 = (1/\alpha)I_q$). Therefore, we immediately obtain from Lemma 5.5 that CP also admits a nice primal algorithmic map as recorded below.

Lemma 5.6 (CP is nice). *Let $\alpha^{-1} > \rho\lambda_{\max}(B^T B)$, the primal algorithmic map $\text{Prim}_t(\cdot)$ defined by (5.10) and (5.11) is nice with $\delta = 1 - \rho\alpha\lambda_{\max}(B^T B)$, $P_1 = Q_1 = 0$, $P_2 = (1/\alpha)I_q$ and $Q_2 = 0$.*

Remark 5.2. (a) Since the primal algorithmic map defined by the CP method is nice, all our results can be applied both in the convex and the strongly convex settings. In the original paper [6] an ergodic $O(1/N)$ rate was proved, and this can be recovered by our Corollary 4.2, under the same condition that $\alpha^{-1} > \rho\lambda_{\max}(B^T B)$.

(b) In [8], a faster ergodic $O(1/N^2)$ rate was proved in the strongly convex setting. Apparently, it does not seem possible to compare the method proposed there, which also includes other parameters that we do not use here. In any case, here we can use FLAG with the primal iterates as described in Lemma 5.6, and thanks to Corollary 4.1 with $\sigma\alpha \geq 2$, we obtain a variant of the CP method with the faster ergodic rate of $O(1/N^2)$.

5.2.4 Example VI: Jacobi (Parallel) Direction Method of Multipliers

In continuation of the ADMM presented above, another approach to exploit the block structure is to split them into two parallel steps, which is also known Jacobi Direction Method of Multipliers (JDMM) [13]. In this case we have the following result. Note that this algorithm is nice only when both functions f and g are treated the same, meaning, if they are both strongly convex functions or both are convex functions.

Lemma 5.7 (Proximal Jacobi is nice). *Let $M_1, M_2 \succeq 0$, the primal algorithmic map $\text{Prim}_t(\cdot)$ defined by*

$$\begin{aligned} u^+ &= \operatorname{argmin}_{\xi_1} \left\{ f(\xi_1) + \langle \lambda, A\xi_1 + Bv - b \rangle + \frac{\rho t}{2} \|A\xi_1 + Bv - b\|^2 + \frac{\tau t}{2} \|\xi_1 - u\|_{M_1}^2 \right\}, \\ v^+ &= \operatorname{argmin}_{\xi_2} \left\{ g(\xi_2) + \langle \lambda, Au + B\xi_2 - b \rangle + \frac{\rho t}{2} \|Au + B\xi_2 - b\|^2 + \frac{\tau t}{2} \|\xi_2 - v\|_{M_2}^2 \right\}, \end{aligned}$$

is nice with $\delta = 1 - 2 \max\{\alpha_1, \alpha_2\}$ where $\alpha_1 = \rho\lambda_{\max}(A^T A) / (\rho\lambda_{\max}(A^T A) + \lambda_{\min}(M_1))$ and $\alpha_2 = \rho\lambda_{\max}(B^T B) / (\rho\lambda_{\max}(B^T B) + \lambda_{\min}(M_2))$, $P_1 = M_1 + \rho A^T A$, $P_2 = M_2 + \rho B^T B$ and $Q_1 = Q_2 = 0$.

We omit the proof since it is similar to the more practical scheme described next.

5.2.5 Example VII: Predictor Corrector Proximal Multiplier Method

In continuation of the JDMM presented above, we would like also to consider the version which consists of linearizing the augmented term. This method, proposed by Chen and Teboulle in [9], is known in the literature as Predictor Corrector Proximal Multipliers (PCPM). For the sake of completeness, since we did not present the proof of the Proximal JDMM, we prove now that PCPM is nice. As mentioned above, here we should consider that both functions f and g are treated the same. Thus, in this part we assume that f is σ_f -strongly convex function with $\sigma_f \geq 0$ and g is σ_g -strongly convex function with $\sigma_g \geq 0$, *i.e.*, either $\sigma_f, \sigma_g > 0$ or $\sigma_f = \sigma_g = 0$.

Lemma 5.8 (PCPM is nice). *Let $M_1, M_2 \succeq 0$ with $\lambda_{\min}(M_1) > \rho\lambda_{\max}(A^T A)$ and $\lambda_{\min}(M_2) > \rho\lambda_{\max}(B^T B)$, the primal algorithmic map $\text{Prim}_t(\cdot)$ defined by*

$$u^+ = \operatorname{argmin}_{\xi_1} \left\{ f(\xi_1) + \langle \lambda, A\xi_1 + Bv - b \rangle + \rho_t \langle Au + Bv - b, A\xi_1 \rangle + \frac{\tau t}{2} \|\xi_1 - u\|_{M_1}^2 \right\}, \quad (5.12)$$

$$v^+ = \operatorname{argmin}_{\xi_2} \left\{ g(\xi_2) + \langle \lambda, Au + B\xi_2 - b \rangle + \rho_t \langle Au + B\xi_2 - b, B\xi_2 \rangle + \frac{\tau t}{2} \|\xi_2 - v\|_{M_2}^2 \right\}, \quad (5.13)$$

is nice with $\delta = 1 - 2 \max\{\alpha_1, \alpha_2\}$ where $\alpha_1 = \rho\lambda_{\max}(A^T A) / \lambda_{\min}(M_1)$ and $\alpha_2 = \rho\lambda_{\max}(B^T B) / \lambda_{\min}(M_2)$, $P_1 = M_1$, $P_2 = M_2$ and $Q_1 = Q_2 = 0$.

Proof. Applying Lemma 5.1 with $\varphi(\xi) = f(\xi_1)$, $c(\xi_1) = \langle \lambda + \rho_t(Au + Bv - b), A\xi_1 \rangle$ and $W = \tau_t M_1$ yields for any $\xi_1 \in \mathbb{R}^p$ that

$$f(u^+) - f(\xi_1) + \langle \lambda + \rho_t(Au + Bv - b), Au^+ - A\xi_1 \rangle \leq \text{RHS}_1,$$

where

$$\text{RHS}_1 = \tau_t \Delta_{M_1}(\xi_1, u, u^+) - \frac{\tau_t}{2} \|u^+ - u\|_{M_1}^2 - \frac{\sigma_f}{2} \|\xi_1 - u^+\|^2.$$

Similarly, by applying Lemma 5.1 with $\varphi(\xi_2) = g(\xi_2)$, $c(\xi_2) = \langle \lambda + \rho_t(Au + Bv - b), B\xi_2 \rangle$ and $W = \tau_t M_2$ yields for any $\xi \in \mathbb{R}^q$ that

$$g(v^+) - g(\xi_2) + \langle \lambda + \rho_t(Au + Bv - b), Bv^+ - B\xi_2 \rangle \leq \text{RHS}_2,$$

where

$$\text{RHS}_2 = \tau_t \Delta_{M_2}(\xi_2, v, v^+) - \frac{\tau_t}{2} \|v^+ - v\|_{M_2}^2 - \frac{\sigma_g}{2} \|\xi_2 - v^+\|^2.$$

By adding these two inequalities and using the fact that $\xi := (\xi_1, \xi_2) \in \mathcal{F}$, that is $A\xi_1 + B\xi_2 = b$, we get

$$\Psi(z^+) - \Psi(\xi) + \langle \lambda, \mathcal{A}z^+ - b \rangle + \rho_t \langle Au + Bv - b, \mathcal{A}z^+ - b \rangle \leq \text{RHS}_1 + \text{RHS}_2, \quad (5.14)$$

where we used that $Au^+ + Bv^+ = \mathcal{A}z^+$ and $z = (u, v)$, $\xi := (\xi_1, \xi_2)$, $z^+ = (u^+, v^+)$. Now, we have

$$\begin{aligned} \langle Au + Bv - b, \mathcal{A}z^+ - b \rangle &= \langle Au^+ + Bv^+ - b, \mathcal{A}z^+ - b \rangle + \langle Au + Bv - Au^+ - Bv^+, \mathcal{A}z^+ - b \rangle \\ &= \|\mathcal{A}z^+ - b\|^2 + \langle Au - Au^+, \mathcal{A}z^+ - b \rangle + \langle Bv - Bv^+, \mathcal{A}z^+ - b \rangle \\ &\geq \|\mathcal{A}z^+ - b\|^2 - \frac{1}{2\alpha_1} \|Au^+ - Au\|^2 - \frac{1}{2\alpha_2} \|Bv^+ - Bv\|^2 \\ &\quad - \frac{\alpha_1 + \alpha_2}{2} \|\mathcal{A}z^+ - b\|^2, \end{aligned} \quad (5.15)$$

where we used again that $Au^+ + Bv^+ = \mathcal{A}z^+$ and the inequality follows by applying (5.9). Combining (5.14) with ρ_t times (5.15), we obtain, for all $\alpha < 1$,

$$\begin{aligned} \Psi(z^+) - \Psi(\xi) + \langle \lambda, \mathcal{A}z^+ - b \rangle + \frac{\rho_t}{2} \|\mathcal{A}z^+ - b\|^2 &\leq \text{RHS}_1 + \frac{\rho_t}{2\alpha_1} \|Au^+ - Au\|^2 \\ &\quad + \text{RHS}_2 + \frac{\rho_t}{2\alpha_2} \|Bv^+ - Bv\|^2 - \frac{\delta\rho_t}{2} \|\mathcal{A}z^+ - b\|^2, \end{aligned}$$

where $\delta = 1 - 2\max\{\alpha_1, \alpha_2\}$ (due to $2\max\{\alpha_1, \alpha_2\} \geq \alpha_1 + \alpha_2$). In order to determine the best parameter $\alpha_1 < 1$ we consider the first term of the previous inequality (using the definition of RHS_1):

$$\frac{\rho_t}{2} \left\{ \frac{1}{\alpha_1} \|u^+ - u\|_{A^T A}^2 - \frac{1}{\rho} \|u^+ - u\|_{M_1}^2 \right\} \leq \frac{\rho_t}{2} \left\{ \left(\frac{1}{\alpha_1} \lambda_{\max}(A^T A) - \frac{1}{\rho} \lambda_{\min}(M_1) \right) \|u^+ - u\|^2 \right\}.$$

Therefore, we will take $\alpha_1 = \rho \lambda_{\max}(A^T A) / \lambda_{\min}(M_1)$. Similarly, we determine α_2 , and the desired result follows. \square

5.3 Model (P) with Additional Smooth Function

Another situation that might occur in applications, is when the objective function Ψ is given as the sum of two functions $f : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ and $h : \mathbb{R}^n \rightarrow \mathbb{R}$, where f is proper, lower semi-continuous and σ -strongly convex (with $\sigma \geq 0$) while h is convex and continuously differentiable with L -Lipschitz continuous gradient, as discussed in Example 2.3. That is, problems of the following form

$$\min_{x \in \mathbb{R}^n} \{f(x) + h(x) : \mathcal{A}x = b\}.$$

5.3.1 Example VIII: Proximal Augmented Lagrangian for Composite Objective

It is well-known that the best way to tackle L -smooth functions in continuous optimization algorithms is via a linearization. Therefore, independently of the Lagrangian-based method to be applied on this model, the additional smooth function h can be easily adapted to the FLAG framework. We first record the result for the Proximal Augmented Lagrangian method without a proof; instead see the next result.

Lemma 5.9 (Proximal AL is nice). *Let $M \succeq LI_n$, the algorithmic map defined by*

$$z^+ = \text{Prim}_t(z, \lambda) \equiv \text{argmin}_\xi \left\{ f(\xi) + \langle \nabla h(z), \xi \rangle + \langle \lambda, \mathcal{A}\xi - b \rangle + \frac{\rho_t}{2} \|\mathcal{A}\xi - b\|^2 + \frac{\tau_t}{2} \|\xi - z\|_M^2 \right\},$$

is nice with $\delta = 1$ and $P = M$ and $Q = M - LI_n$.

5.3.2 Example IX: Proximal Linearized Augmented Lagrangian for Composite Objective

Handling a smooth function in the given objective function is independent of the Lagrangian-based method that is used, and therefore will affect in a similar way (cf. Lemma 5.9) any Lagrangian-based method such as ADMM, Proximal Linearized ADMM, CP, Proximal JDMM and PCPM (the value of the corresponding δ might changed too). For example, if we apply the Linearized Proximal AL method we will also have a nice primal algorithmic map as shown below.

Lemma 5.10 (Linearized Proximal AL is nice). *Let $M \succeq \rho \mathcal{A}^T \mathcal{A} + LI_n$, the algorithmic map defined by*

$$z^+ = \text{Prim}_t(z, \lambda) \equiv \text{argmin}_\xi \left\{ f(\xi) + \langle \nabla h(z), \xi \rangle + \langle \lambda, \mathcal{A}\xi - b \rangle + \rho_t \langle \mathcal{A}z - b, \mathcal{A}\xi \rangle + \frac{\tau_t}{2} \|\xi - z\|_M^2 \right\},$$

is nice with $\delta = 1$ and $P = M - \rho \mathcal{A}^T \mathcal{A}$ and $Q = M - \rho \mathcal{A}^T \mathcal{A} - LI_n$.

Proof. We apply Lemma 5.1 on the functions $\varphi(\cdot) = f(\cdot)$, $c(\xi) = \langle \nabla h(z), \xi \rangle + \langle \lambda + \rho_t(\mathcal{A}z - b), \mathcal{A}\xi \rangle$ and $W = \tau_t M$ to obtain, for any $\xi \in \mathbb{R}^n$, that

$$\begin{aligned} f(z^+) - f(\xi) + \langle \lambda + \rho_t(\mathcal{A}z - b), \mathcal{A}z^+ - \mathcal{A}\xi \rangle + \langle \nabla h(z), z^+ - \xi \rangle &\leq \tau_t \Delta_M(\xi, z, z^+) - \frac{\tau_t}{2} \|z^+ - z\|_M^2 \\ &\quad - \frac{\sigma}{2} \|\xi - z^+\|^2. \end{aligned}$$

Using the Three-Points Descent Lemma applied on the smooth convex function h , we get that

$$h(z^+) - h(\xi) - \langle \nabla h(z), z^+ - \xi \rangle \leq \frac{L}{2} \|z^+ - z\|^2 \leq \frac{L\rho_t}{2\rho} \|z^+ - z\|^2,$$

where the last inequality follows from the fact that $\rho_t/\rho \geq 1$. Combining these inequalities shows that

$$\Psi(z^+) - \Psi(\xi) + \langle \lambda, \mathcal{A}z^+ - b \rangle + \rho_t \langle \mathcal{A}z - b, \mathcal{A}z^+ - \mathcal{A}\xi \rangle \leq \tau_t \Delta_M(\xi, z, z^+) - \frac{\tau_t}{2} \|z^+ - z\|_{M-LI_n}^2 - \frac{\sigma}{2} \|\xi - z^+\|^2.$$

The result now follows using the same arguments as we used in the proof of Lemma 5.3. \square

References

- [1] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim.*, 16(3):697–725, 2006.
- [2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

- [3] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Computer Science and Applied Mathematics. Academic Press Inc., New York-London, 1982.
- [4] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation*. Prentice-Hall International Editions, Englewood Cliffs, NJ, 1989.
- [5] 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.*, 3(1):1–122, 2011.
- [6] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145, 2011.
- [7] A. Chambolle and T. Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.
- [8] A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primaldual algorithm. *Math. Programming*, 159(1):253–287, 2016.
- [9] G. Chen and M. Teboulle. A proximal-based decomposition method for convex minimization problems. *Math. Programming*, 64(1):81–101, 1994.
- [10] M. Fortin and R. Glowinski. *Augmented Lagrangian Methods*, volume 15 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1983.
- [11] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.*, 2(1):17–40, 1976.
- [12] R. Glowinski and A. Marrocco. Sur l’approximation, par éléments finis d’ordre un, et la résolution, par pénalisation-dualité, d’une classe de problèmes de Dirichlet non linéaires. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Anal. Numér.*, 9, 1975.
- [13] B. He, L. Hou, and X. Yuan. On full Jacobian decomposition of the augmented Lagrangian method for separable convex programming. *SIAM J. Optim.*, 25(4):2274–2312, 2015.
- [14] B. He and X. Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM J. Numer. Anal.*, 50(2):700–709, 2012.
- [15] B. He and X. Yuan. On non-ergodic convergence rate of Douglas–Rachford alternating direction method of multipliers. *Numer. Math.*, 130(3):567–577, 2015.
- [16] M. Hestenes. Multiplier and gradient methods. *J. Optim. Theory Appl.*, 4(5):303–320, 1969.
- [17] H. Li and Z. Lin. Accelerated alternating direction method of multipliers: an optimal $O(1/k)$ nonergodic analysis. *J. Sci. Comput.*, 79(2):671–699, 2019.
- [18] Z. Lin, H. Li, and C. Fang. *Accelerated Optimization for Machine Learning*. Springer, 2020.
- [19] Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.
- [20] M. J. D. Powell. A method for nonlinear constraints in minimization problems. In R. Fletcher, editor, *Optimization*, pages 283–298. Academic Press, New York, 1969.
- [21] R. T. Rockafellar. *Convex Analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

- [22] S. Sabach and M. Teboulle. Lagrangian methods for composite optimization. In X.-C. Tai R. Kimmel, editor, *Processing, Analyzing and Learning of Images, Shapes, and Forms: Part 2*, volume 20, pages 402–436. Elsevier, 2019.
- [23] R. Shefi and M. Teboulle. Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. *SIAM J. Optim.*, 24(1):269–297, 2014.