



HAL
open science

Inexact and Stochastic Generalized Conditional Gradient with Augmented Lagrangian and Proximal Step

Antonio Silveti-Falls, Cesare Molinari, Jalal Fadili

► **To cite this version:**

Antonio Silveti-Falls, Cesare Molinari, Jalal Fadili. Inexact and Stochastic Generalized Conditional Gradient with Augmented Lagrangian and Proximal Step. *Journal of Nonsmooth Analysis and Optimization*, 2021, Volume 2 (6480), pp.41. 10.46298/jnsao-2021-6480 . hal-02569925v4

HAL Id: hal-02569925

<https://hal.science/hal-02569925v4>

Submitted on 31 Aug 2021

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

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

INEXACT AND STOCHASTIC GENERALIZED CONDITIONAL GRADIENT WITH AUGMENTED LAGRANGIAN AND PROXIMAL STEP

Antonio Silveti-Falls*

Cesare Molinari*

Jalal Fadili*

Abstract In this paper we propose and analyze inexact and stochastic versions of the CGALP algorithm developed in [34], which we denote ICGALP, that allow for errors in the computation of several important quantities. In particular this allows one to compute some gradients, proximal terms, and/or linear minimization oracles in an inexact fashion that facilitates the practical application of the algorithm to computationally intensive settings, e.g., in high (or possibly infinite) dimensional Hilbert spaces commonly found in machine learning problems. The algorithm is able to solve composite minimization problems involving the sum of three convex proper lower-semicontinuous functions subject to an affine constraint of the form $Ax = b$ for some bounded linear operator A . Only one of the functions in the objective is assumed to be differentiable, the other two are assumed to have an accessible proximal operator and a linear minimization oracle. As main results, we show convergence of the Lagrangian values (so-called convergence in the Bregman sense) and asymptotic feasibility of the affine constraint as well as strong convergence of the sequence of dual variables to a solution of the dual problem, in an almost sure sense. Almost sure convergence rates are given for the Lagrangian values and the feasibility gap for the ergodic primal variables. Rates in expectation are given for the Lagrangian values and the feasibility gap subsequentially in the pointwise sense. Numerical experiments verifying the predicted rates of convergence are shown as well.

1 INTRODUCTION

1.1 PROBLEM STATEMENT

We consider the following composite minimization problem,

$$(\mathcal{P}) \quad \min_{x \in \mathcal{H}_p} \{f(x) + g(Tx) + h(x) : Ax = b\},$$

and its associated *dual problem*,

$$(\mathcal{D}) \quad \min_{\mu \in \mathcal{H}_d} (f + g \circ T + h)^*(-A^*\mu) + \langle \mu, b \rangle,$$

where we have denoted by $*$ both the *Legendre-Fenchel conjugate* and the *adjoint operator*, to be understood from context. We consider \mathcal{H}_p , \mathcal{H}_d , and \mathcal{H}_v to be arbitrary real Hilbert spaces, possibly infinite-dimensional, whose indices correspond to a primal, dual, and auxiliary space, respectively;

ASF was supported by the ERC Consolidated grant NORIA. JF was partly supported by Institut Universitaire de France. CM was supported by Project MONOMADS funded by Conseil Régional de Normandie. ASF would like to thank Jingwei Liang for useful discussions had during his visit to Cambridge University.

*Normandie Université, ENSICAEN, UNICAEN, CNRS, GREYC, France (tonys.falls@gmail.com, cecio.molinari@gmail.com, Jalal.Fadili@ensicaen.fr)

$A : \mathcal{H}_p \rightarrow \mathcal{H}_d$ and $T : \mathcal{H}_p \rightarrow \mathcal{H}_v$ to be bounded linear operators with $b \in \text{ran}(A)$; functions f , g , and h to all be convex, closed, and proper real-valued functions. Additionally, we will assume that the function f satisfies a certain differentiability condition generalizing Lipschitz-smoothness, Hölder-smoothness, etc (see Definition 2.7), that the function g has a proximal mapping which is accessible, and that the function h admits an accessible linearly-perturbed minimization oracle with $C \stackrel{\text{def}}{=} \text{dom}(h)$ a weakly compact subset of \mathcal{H}_p .

In fact, the problem under consideration here is exactly the same as that of [34], however, in this work, we consider an inexact extension of the algorithm presented and analyzed in [34] to solve (\mathcal{P}) . The extension amounts to allowing either deterministic or stochastic errors in the computation of several quantities, including the gradient or proximal terms, e.g. ∇f , $\text{prox}_{\beta g}$, and the linear minimization oracle itself.

1.2 CONTRIBUTION AND PRIOR WORK

The primary contribution of this work is to analyze inexact and stochastic variants of the CGALP algorithm presented in [34] to address (\mathcal{P}) . We coin this algorithm **Inexact Conditional Gradient with Augmented Lagrangian and Proximal-step (ICGALP)**. Although there has been a great deal of work on developing and analyzing Frank-Wolfe or conditional gradient style algorithms, first studied in the 1950's in [17] and later in [24], in both the stochastic and deterministic case, e.g. [20, 21, 32, 18, 16, 35, 27, 19], or [26], little to no work has been done to analyze the generalized version of these algorithms for nonsmooth problems or problems involving an affine constraint, as we consider here. To the best of our knowledge, the only such work is [25], where the authors consider a stochastic conditional gradient algorithm applied to a composite problem of the form

$$\min_{x \in \mathcal{X} \subset \mathbb{R}^n} \mathbb{E} [f(x, \eta)] + g(Ax)$$

where the expectation is over the random variable η and with g possibly nonsmooth. The nonsmooth term is possibly an affine constraint but, in such cases, it is addressed through smoothing rather than through an augmented Lagrangian with a dual variable, in contrast to our work. They consider only finite-dimensional problems and their problem formulation doesn't allow for inexactness with respect to g .

We show asymptotic feasibility of the primal iterates for the affine constraint, convergence of the Lagrangian values at each iteration to an optimum value, strong convergence of the sequence of dual iterates, and provide worst-case rates of convergence for the feasibility gap and the Lagrangian values; all these results are in an almost sure sense. The rates of convergence for both the Lagrangian and the feasibility gap are given globally, i.e., for the entire sequence of iterates, in the ergodic sense where the Cesàro means are taken with respect to the primal step size, in an almost sure sense. We also show rates in expectation which hold pointwise but subsequentially. In the case where (\mathcal{P}) admits a unique solution, we furthermore have that the sequence of primal iterates converges weakly to the solution almost surely. These results are established for a family of parameters satisfying abstract open loop conditions, i.e. sequences of parameters which do not depend on the iterates themselves. We exemplify the framework on problem instances involving a smooth risk minimization where the gradient is computed inexactly either with stochastic noise or a deterministic error. In the stochastic case, we show that our conditions outlined in Section 3 for convergence are satisfied via increasing batch size or variance reduction. In the deterministic setting for minimizing an empirical risk, a sweeping approach is described.

1.3 MOTIVATION

Conditional gradient methods have seen increased interest in the last decade, due to their applicability to solving a large variety of problems, in particular in large-scale signal/image processing and statistical

learning. A chief advantage of conditional gradient is its ability to take advantage of the atomic structure of the set (or level sets of the function) that come into play, and they can do so for problems posed even in an infinite-dimensional setting. Besides this, conditional gradient methods offer an alternative to projecting onto a closed convex set in constrained optimization by instead utilizing a linear minimization oracle (lmo). The difference in complexity for these two operations depends on the set in question; a survey comparing the two for sets commonly found in practice can be found in [11]. Note that for some constraint sets, projection may not be even possible in closed form while solving a linear minimization oracle is. One of the main motivations behind this work is to be able to handle composite problems where not only one atomic constraint set is involved, but finitely many of them. There are many problems in data processing and/or machine learning where such situation is encountered, typically when one uses several regularizers or constraints in a variational formulation (e.g., recover a sparse low rank matrix). We are also motivated by composite problems where it is favourable to hybridize conditional gradient involving an lmo step on some (or several) atomic set, and a proximal/projection step on another set; see [10] where the constraint is the intersection of the semidefinite positive cone and the set of Toeplitz matrices to solve a semidefinite relaxation of the Beurling lasso [15]. Our interest in the infinite-dimensional setting is motivated by the ability of conditional gradient type algorithms to handle infinite-dimensional problems, most notably sparse inverse problems [7], the Beurling lasso [15] to recover sparse Radon measures, training neural networks with one hidden layer [4, 14], and optimal transport problems [8]. In the latter instances, the problems are posed over infinite dimensional nonreflexive Banach spaces, a nontrivial endeavor. In this sense, our work can be seen as an important intermediary step towards this goal, with proofs that provide insight to how the infinite-dimensional case might be handled.

1.4 ORGANIZATION

The remainder of the paper is divided into four sections. In Section 2 the necessary notation and prior results are recalled, consisting primarily of convex analysis, real analysis, and elementary probability. In Section 3 the assumptions on the problem structure and the parameters are noted, the ICGALP algorithm itself is presented. In Section 4, the main results, e.g. feasibility, Lagrangian convergence, and rates, are established. The analysis and results are far-reaching extensions of those in [34] to the inexact and stochastic setting, and require quite delicate new arguments. In Section 5 and Section 6, we consider different problem instances where inexact deterministic or stochastic computations are involved. Numerical results are reported in Section 7 to support our theoretical findings. Finally, in Section 8, we summarize the work and provide some closing remarks.

2 NOTATION AND PRELIMINARIES

Many of the following notations for probabilistic concepts are adopted from [13]. A sequence $(x_k)_{k \in \mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$ will be called strongly convergent to $x \in \mathcal{H}$, denoted $x_k \rightarrow x$, iff $\|x_k - x\| \rightarrow 0$; it will be called weakly convergent to $x \in \mathcal{H}$, denoted $x_k \rightharpoonup x$, iff, for any $u \in \mathcal{H}$, $\langle x_k, u \rangle \rightarrow \langle x, u \rangle$. We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space with set of events Ω , σ -algebra \mathcal{F} , and probability measure \mathbb{P} . When discussing random variables we will assume that any Hilbert space \mathcal{H} is endowed with the Borel σ -algebra, $\mathcal{B}(\mathcal{H})$. We denote a *filtration* by $\mathfrak{F} = (\mathcal{F}_k)_{k \in \mathbb{N}}$, i.e. a sequence of sub- σ -algebras which satisfies $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all $k \in \mathbb{N}$. Given a set of random variables $\{a_0, \dots, a_n\}$, we denote by $\sigma(a_0, \dots, a_n)$ the σ -algebra generated by a_0, \dots, a_n . An expression (P) is said to hold (\mathbb{P} -a.s.) if $\mathbb{P}(\{\omega \in \Omega : (P) \text{ holds}\}) = 1$. Throughout the paper, both equalities and inequalities involving random quantities should be understood as holding \mathbb{P} -almost surely, whether or not it is explicitly written.

Definition 2.1. Given a filtration \mathfrak{F} , we denote by $\ell_+(\mathfrak{F})$ the set of sequences of $[0, +\infty[$ -valued random variables $(a_k)_{k \in \mathbb{N}}$ such that, for each $k \in \mathbb{N}$, a_k is \mathcal{F}_k measurable. Then, we also define the following set,

$$\ell_+^1(\mathfrak{F}) \stackrel{\text{def}}{=} \left\{ (a_k)_{k \in \mathbb{N}} \in \ell_+(\mathfrak{F}) : \sum_{k \in \mathbb{N}} a_k < +\infty \text{ (\mathbb{P}\text{-a.s.})} \right\}$$

Lemma 2.2. Given a filtration \mathfrak{F} and the sequences of random variables $(r_k)_{k \in \mathbb{N}} \in \ell_+(\mathfrak{F})$, $(a_k)_{k \in \mathbb{N}} \in \ell_+(\mathfrak{F})$, and $(z_k)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{F})$ satisfying,

$$\mathbb{E}[r_{k+1} | \mathcal{F}_k] - r_k \leq -a_k + z_k \text{ (\mathbb{P}\text{-a.s.})}$$

then $(a_k)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{F})$ and $(r_k)_{k \in \mathbb{N}}$ converges (\mathbb{P} -a.s.) to a random variable with value in $[0, +\infty[$.

Proof. See [33, Theorem 1]. □

Lemma 2.3. Given a filtration \mathfrak{F} and a sequence of random variables $(w_k)_{k \in \mathbb{N}} \in \ell_+(\mathfrak{F})$ and a sequence of real numbers $(\gamma_k)_{k \in \mathbb{N}} \in \ell_+$ such that $(\gamma_k w_k)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{F})$ and $(\gamma_k)_{k \in \mathbb{N}} \notin \ell^1$, then:

- (i) There exists a subsequence $(w_{k_j})_{j \in \mathbb{N}}$ such that $\liminf_k w_k = 0$ (\mathbb{P} -a.s.),
- (ii) Furthermore, if there exists a constant $\alpha > 0$ such that $w_k - \mathbb{E}[w_{k+1} | \mathcal{F}_k] \leq \alpha \gamma_k$ (\mathbb{P} -a.s.) for every $k \in \mathbb{N}$, then

$$\lim_k w_k = 0 \text{ (\mathbb{P}\text{-a.s.}) .}$$

Proof. The second result is directly from [5, Lemma 2.2] and the first follows from [1] trivially extended to the stochastic setting. □

Lemma 2.4. Consider the real sequences $(r_k)_{k \in \mathbb{N}} \in \ell_+$, $(p_k)_{k \in \mathbb{N}} \in \ell_+$, $(w_k)_{k \in \mathbb{N}} \in \ell_+$, and $(z_k)_{k \in \mathbb{N}} \in \ell_+^1$. Suppose further that $(p_k)_{k \in \mathbb{N}} \notin \ell^1$ and that, for some $\alpha > 0$, the following inequalities are satisfied for every $k \in \mathbb{N}$:

$$(2.1) \quad \begin{cases} r_{k+1} \leq r_k - p_k w_k + z_k; \\ w_k - w_{k+1} \leq \alpha p_k. \end{cases}$$

Then,

- (i) $(r_k)_{k \in \mathbb{N}}$ is convergent and $(p_k w_k)_{k \in \mathbb{N}} \in \ell_+^1$.
- (ii) $\lim_k w_k = 0$.
- (iii) For every $k \in \mathbb{N}$, $\inf_{1 \leq i \leq k} w_i \leq (r_0 + E)/P_k$, where, again, $P_n = \sum_{k=1}^n p_k$ and $E = \sum_{k=1}^{+\infty} z_k$.
- (iv) There exists a subsequence $(w_{k_j})_{j \in \mathbb{N}}$ such that, for all $j \in \mathbb{N}$, $w_{k_j} \leq P_{k_j}^{-1}$.

Proof. See [1] for the proof. □

We denote by $\Gamma_0(\mathcal{H})$ the set of proper, convex, and lower semi-continuous functions $f : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$. We also consider the *domain* of a function f to be $\text{dom}(f) \stackrel{\text{def}}{=} \{x \in \mathcal{H} : f(x) < +\infty\}$ and the *Legendre-Fenchel conjugate* of f to be the function $f^* : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$ such that, $\forall y \in \mathcal{H}$,

$$f^*(y) \stackrel{\text{def}}{=} \sup_{x \in \mathcal{H}} \{\langle y, x \rangle - f(x)\}.$$

Throughout, differentiability will be intended in Fréchet sense, and we denote ∇f the (Fréchet) gradient of a differentiable function f . The *proximal mapping* (or *proximal operator*) associated to the function f with parameter β is given by

$$\text{prox}_{\beta f}(x) \stackrel{\text{def}}{=} \underset{y \in \mathcal{H}}{\text{argmin}} \left\{ f(y) + \frac{1}{2\beta} \|x - y\|^2 \right\}.$$

The following elementary result from convex analysis regarding proximal mappings will be used in the proof of optimality.

Proposition 2.5. *Let $f \in \Gamma_0(\mathcal{H})$ and denote $x^+ = \text{prox}_f(x)$. Then, for all $y \in \mathcal{H}$,*

$$2(f(x^+) - f(x)) + \|x^+ - y\|^2 - \|x - y\|^2 + \|x^+ - x\|^2 \leq 0.$$

Proof. The result is classical and the proof is readily available, e.g. in [30, Chapter 6.2.1]. □

The *subdifferential* of a function f is the set-valued operator $\partial f : \mathcal{H} \rightarrow 2^{\mathcal{H}}$ such that, for every $x \in \mathcal{H}$,

$$(2.2) \quad \partial f(x) \stackrel{\text{def}}{=} \{u \in \mathcal{H} : f(y) \geq f(x) + \langle u, y - x \rangle \quad \forall y \in \mathcal{H}\}$$

We denote $\text{dom}(\partial f) \stackrel{\text{def}}{=} \{x \in \mathcal{H} : \partial f(x) \neq \emptyset\}$ as the *domain of the subdifferential*. For $x \in \text{dom}(\partial f)$, the *minimal norm selection* of $\partial f(x)$ is denoted by $[\partial f(x)]^0 \stackrel{\text{def}}{=} \underset{y \in \partial f(x)}{\text{argmin}} \|y\|$. The *Moreau envelope* of the function f with parameter β is given by,

$$f^\beta(x) \stackrel{\text{def}}{=} \inf_{y \in \mathcal{H}} \left\{ f(y) + \frac{1}{2\beta} \|x - y\|^2 \right\}.$$

The following proposition recalls some key properties of the Moreau envelope which we will utilize in the analysis of the algorithm.

Proposition 2.6 (Moreau envelope properties). *Given a function $f \in \Gamma_0(\mathcal{H})$, the following holds:*

- (i) *The Moreau envelope, f^β , is convex, real-valued, and continuous.*
- (ii) *Lax-Hopf formula: the Moreau envelope is the viscosity solution to the following Hamilton Jacobi equation:*

$$(2.3) \quad \begin{cases} \frac{\partial}{\partial \beta} f^\beta(x) = -\frac{1}{2} \|\nabla f^\beta(x)\|^2 & (x, \beta) \in \mathcal{H} \times (0, +\infty) \\ f^0(x) = f(x) & x \in \mathcal{H}. \end{cases}$$

- (iii) *The gradient of the Moreau envelope, ∇f^β , is $\frac{1}{\beta}$ -Lipschitz continuous and is given by the expression*

$$\nabla f^\beta(x) = \frac{x - \text{prox}_{\beta f}(x)}{\beta}.$$

- (iv) $\forall x \in \text{dom}(\partial f)$, $\|\nabla f^\beta(x)\| \nearrow \|[\partial f(x)]^0\|$ as $\beta \searrow 0$.

(v) $\forall x \in \mathcal{H}, f^\beta(x) \nearrow f(x)$ as $\beta \searrow 0$. In addition, given two positive real numbers $\beta' < \beta$, for all $x \in \mathcal{H}$ we have

$$0 \leq f^{\beta'}(x) - f^\beta(x) \leq \frac{\beta - \beta'}{2} \|\nabla f^{\beta'}(x)\|^2;$$

$$0 \leq f(x) - f^\beta(x) \leq \frac{\beta}{2} \|[\partial f(x)]^0\|^2.$$

Proof. (i): see [6, Proposition 12.15]. The proof for (ii) can be found in [3, Lemma 3.27 and Remark 3.32] (see also [22] or [2, Section 3.1]). The proof for claim (iii) can be found in [6, Proposition 12.29] and the proof for claim (iv) can be found in [6, Corollary 23.46]. For the first part in (v), see [6, Proposition 12.32(i)]. To show the first inequality in (v), combine (ii) and convexity of the function $\beta \mapsto g^\beta(x)$ for every $x \in \mathcal{H}$. The second inequality follows from the first one and (iv), taking the limit as $\beta' \rightarrow 0$. \square

Given a closed, convex set C , we write $d_C \stackrel{\text{def}}{=} \sup_{x,y \in C} \|x - y\|$ to denote the *diameter* of C . We denote the *Bregman divergence* of a differentiable function F by

$$D_F(x, y) \stackrel{\text{def}}{=} F(x) - F(y) - \langle \nabla F(y), x - y \rangle.$$

Definition 2.7 ((F, ζ)-smoothness). Let $F : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$ and $\zeta :]0, 1[\rightarrow \mathbb{R}_+$. The pair (f, C) , where $f : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$ and $C \subset \text{dom}(f)$, is said to be (F, ζ) -smooth if there exists an open set C_0 such that $C \subset C_0 \subset \text{int}(\text{dom}(F))$ and

- (i) F and f are differentiable on C_0 ;
- (ii) $F - f$ is convex on C_0 ;
- (iii) it holds

$$(2.4) \quad K_{(F, \zeta, C)} \stackrel{\text{def}}{=} \sup_{\substack{x, s \in C; \gamma \in]0, 1[\\ z = x + \gamma(s - x)}} \frac{D_F(z, x)}{\zeta(\gamma)} < +\infty.$$

Remark 2.8. An important consequence of Definition 2.7(i) and Definition 2.7(ii) in (F, ζ) -smoothness is the following. Let (f, C) be (F, ζ) smooth. Then, for any $x, y \in C$, we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + D_F(y, x).$$

Moreover, by Definition 2.7(iii), if $y = x + \gamma(s - x)$ for some $s \in C$ and $\gamma \in]0, 1[$, we have,

$$(2.5) \quad D_F(y, x) \leq K_{(F, \zeta, C)} \zeta(\gamma).$$

Definition 2.9 (ω -smoothness). Consider a function $\omega : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that $\omega(0) = 0$ and $\xi(s) \stackrel{\text{def}}{=} \int_0^1 \omega(st) dt$ is nondecreasing. A differentiable function $g : \mathcal{H} \rightarrow \mathbb{R}$ is said to be ω -smooth if, for every $x, y \in \mathcal{H}$,

$$\|\nabla g(x) - \nabla g(y)\| \leq \omega(\|x - y\|)$$

Remark 2.10. A classical consequence of ω -smoothness is the following. If $g : \mathcal{H} \rightarrow \mathbb{R}$ is ω -smooth, for every $x, y \in \mathcal{H}$ we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \xi(\|y - x\|) \|y - x\|.$$

Remark 2.11. Note that being ω -smooth is a stronger condition than being (F, ζ) -smooth since every ω -smooth function f is also (F, ζ) -smooth with $F = f, \zeta(t) = d_C t \xi(d_C t)$ and $K_{(F, \zeta, C)} \leq 1$. Additionally, the assumptions on ξ being nondecreasing can be replaced by the sufficient condition that $\lim_{t \rightarrow 0^+} \omega(t) = \omega(0) = 0$.

3 ALGORITHM AND ASSUMPTIONS

For each $k \in \mathbb{N}$, we denote by λ_k and λ_k^s random variables from $(\Omega, \mathcal{F}, \mathbb{P})$ to \mathcal{H}_p and \mathbb{R}_+ respectively. In this context, λ_k will represent the error in the gradient or proximal terms and λ_k^s will represent the error in the linear minimization oracle itself.

Algorithm 1: Inexact Conditional Gradient with Augmented Lagrangian and Proximal-step (ICGALP)

Input: $x_0 \in C \stackrel{\text{def}}{=} \text{dom}(h)$; $\mu_0 \in \text{ran}(A)$; $(\gamma_k)_{k \in \mathbb{N}}, (\beta_k)_{k \in \mathbb{N}}, (\theta_k)_{k \in \mathbb{N}}, (\rho_k)_{k \in \mathbb{N}} \in \ell_+$.

$k = 0$

repeat

$$y_k = \text{prox}_{\beta_k g}(Tx_k)$$

$$z_k = \nabla f(x_k) + T^*(Tx_k - y_k) / \beta_k + A^* \mu_k + \rho_k A^*(Ax_k - b) + \lambda_k$$

$$s_k \in \text{Argmin}_{s \in \mathcal{H}_p} \{h(s) + \langle z_k, s \rangle\}$$

$$\widehat{s}_k \in \{s \in \mathcal{H}_p : h(s) + \langle z_k, s \rangle \leq h(s_k) + \langle z_k, s_k \rangle + \lambda_k^s\}$$

$$x_{k+1} = x_k - \gamma_k (x_k - \widehat{s}_k)$$

$$\mu_{k+1} = \mu_k + \theta_k (Ax_{k+1} - b)$$

$$k \leftarrow k + 1$$

until convergence;

Output: x_{k+1} .

To improve readability, we list some notation for the functionals we will employ throughout the analysis of the algorithm,

$$(3.1) \quad \begin{aligned} \Phi(x) &\stackrel{\text{def}}{=} f(x) + g(Tx) + h(x); \\ \mathcal{L}(x, \mu) &\stackrel{\text{def}}{=} f(x) + g(Tx) + h(x) + \langle \mu, Ax - b \rangle; \\ \mathcal{L}_k(x, \mu) &\stackrel{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + h(x) + \langle \mu, Ax - b \rangle + \frac{\rho_k}{2} \|Ax - b\|^2; \\ \mathcal{E}_k(x, \mu) &\stackrel{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + \langle \mu, Ax - b \rangle + \frac{\rho_k}{2} \|Ax - b\|^2; \\ \Phi_k(x) &\stackrel{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + h(x). \end{aligned}$$

We can recognize $\mathcal{L}(x, \mu)$ as the classical Lagrangian, $\mathcal{L}_k(x, \mu)$ as the augmented Lagrangian with smoothed g , $\mathcal{E}_k(x, \mu)$ as the smooth part of $\mathcal{L}_k(x, \mu)$, and $\Phi_k(x)$ as the primal objective with smoothed g . With this notation in mind, we can see z_k as $\nabla_x \mathcal{E}_k(x_k, \mu_k)$ and λ_k as the error in the computation of $\nabla_x \mathcal{E}_k(x_k, \mu_k)$.

We define the following filtration

$$(3.2) \quad \mathfrak{S} \stackrel{\text{def}}{=} (\mathcal{S}_k)_{k \in \mathbb{N}} \quad \text{where} \quad \mathcal{S}_k \stackrel{\text{def}}{=} \sigma(x_0, \mu_0, \widehat{s}_0, \dots, \widehat{s}_k)$$

and $\sigma(x_0, \mu_0, \widehat{s}_0, \dots, \widehat{s}_k)$ is the σ -algebra generated by the random variables $x_0, \mu_0, \widehat{s}_0, \dots, \widehat{s}_k$ defined in Algorithm 1. Furthermore, due to the error terms being contained in the direction finding step, we have that x_{k+1} and μ_{k+1} are completely determined by \mathcal{S}_k . Another noteworthy consequence of the error terms being contained in the direction finding step is that the primal iterates $(x_k)_{k \in \mathbb{N}}$ remain in C , as in the classical Frank-Wolfe algorithm, while the dual iterates $(\mu_k)_{k \in \mathbb{N}}$ remain in $\text{ran}(A)$.

Finally, we define the notation for the set of solutions for (\mathcal{P}) and (\mathcal{D}) to be

$$(3.3) \quad \begin{aligned} \mathcal{S}_{\mathcal{P}} &\stackrel{\text{def}}{=} \text{Argmin}_{x \in \mathcal{H}_p} \{f(x) + g(x) + h(x) : Ax = b\} \quad \text{and} \quad \mathcal{S}_{\mathcal{D}} \stackrel{\text{def}}{=} \text{Argmin}_{\mu \in \mathcal{H}_d} \{(f + g + h)^*(-A^* \mu) + \langle \mu, b \rangle\} \end{aligned}$$

and the notation for the set of weak cluster points of a sequence $(x_k)_{k \in \mathbb{N}}$ in \mathcal{H}_p to be

$$(3.4) \quad \mathfrak{W} [(x_k)_{k \in \mathbb{N}}] \stackrel{\text{def}}{=} \left\{ x \in \mathcal{H}_p : \exists (x_{k_j})_{j \in \mathbb{N}}, x_{k_j} \rightharpoonup x \right\}.$$

3.1 ASSUMPTIONS

3.1.1 ASSUMPTIONS ON THE FUNCTIONS

We impose the following assumptions on the problem we consider; for some results, only a subset of them will be necessary:

- (A.1) The functions $f, g \circ T$, and h belong to $\Gamma_0(\mathcal{H}_p)$.
- (A.2) The pair (f, C) is (F, ζ) -smooth (see Definition 2.7), where we recall $C \stackrel{\text{def}}{=} \text{dom}(h)$.
- (A.3) The set C is weakly compact (and thus contained in a ball of radius $R > 0$).
- (A.4) It holds $TC \subset \text{dom}(\partial g)$ and $\sup_{x \in C} \|\partial g(Tx)\| = M < \infty$.
- (A.5) The function h is Lipschitz continuous relative to its domain C with constant $L_h \geq 0$, i.e., $\forall (x, z) \in C^2, |h(x) - h(z)| \leq L_h \|x - z\|$.
- (A.6) There exists a saddle-point $(x^*, \mu^*) \in \mathcal{H}_p \times \mathcal{H}_d$ for the Lagrangian \mathcal{L} .
- (A.7) The set $\text{ran}(A)$ is closed.
- (A.8) One of the following holds:
 - (a) $A^{-1}(b) \cap \text{int}(\text{dom}(g \circ T)) \cap \text{int}(C) \neq \emptyset$, where $A^{-1}(b)$ is the pre-image of b under A .
 - (b) \mathcal{H}_p and \mathcal{H}_d are finite-dimensional and

$$(3.5) \quad \begin{cases} A^{-1}(b) \cap \text{ri}(\text{dom}(g \circ T)) \cap \text{ri}(C) \neq \emptyset \\ \text{and} \\ \text{ran}(A^*) \cap \text{par}(\text{dom}(g \circ T) \cap C)^\perp = \{0\}. \end{cases}$$

- (A.9) The space \mathcal{H}_d is separable.
- (A.10) The set-valued mappings $\left(\partial \left(\Phi_k^* \circ (-A^*) \right) \right)_{k \in \mathbb{N}}$ satisfy the following property: for any sequence $((p_k, q_k))_{k \in \mathbb{N}}$ satisfying, for each $k \in \mathbb{N}$,

$$p_k \in \partial \left(\Phi_k^* \circ (-A^*) \right) (q_k),$$

with $p_k \rightarrow p$ and $q_k \rightarrow q$, the sequence $(q_k)_{k \in \mathbb{N}}$ admits a strong cluster point.

The following lemmas outline sufficient conditions ensure that assumption (A.4) holds for g and show why it's unnecessary to make a similar assumption for f in light of (A.1) and (A.2).

Lemma 3.1. *Let $T : \mathcal{H}_p \rightarrow \mathcal{H}_v$ be a bounded linear operator. Assume that one of the following holds:*

- (i) $g \in \Gamma_0(\mathcal{H}_v)$, $TC \subset \text{int}(\text{dom}(g))$ and C is a nonempty compact subset of \mathcal{H}_p .
- (ii) $g : \mathcal{H}_v \rightarrow \mathbb{R}$ is continuous, convex and bounded on bounded sets of \mathcal{H}_v , and C is a nonempty bounded subset of \mathcal{H}_p .

(iii) \mathcal{H}_v and \mathcal{H}_p are finite dimensional, and either $g \in \Gamma_0(\mathcal{H}_v)$, $TC \subset \text{int}(\text{dom}(g))$ and C is closed and bounded, or $g : \mathcal{H}_v \rightarrow \mathbb{R}$ is continuous and convex and C is a nonempty bounded subset of \mathcal{H}_p .

Then (A.4) holds.

Proof. (i) Since $g \in \Gamma_0(\mathcal{H}_p)$, it follows from [6, Proposition 16.21] that

$$TC \subset \text{int}(\text{dom}(g)) \subset \text{dom}(\partial g).$$

Moreover, by [6, Corollary 8.30(ii) and Proposition 16.14], we have that ∂g is locally bounded on $\text{int}(\text{dom}(g))$. In particular, as we assume that C is bounded, so is TC , and since $TC \subset \text{int}(\text{dom}(g))$, it means that for each $z \in TC$ there exists an open neighborhood of z , denoted by U_z , such that $\partial g(U_z)$ is bounded. Since $(U_z)_{z \in C}$ is an open cover of TC and TC is compact, there exists a finite subcover $(U_{z_k})_{k=1}^n$. Then,

$$\bigcup_{x \in C} \partial g(Tx) \subset \bigcup_{k=1}^n \partial g(U_{z_k}).$$

Since the right-hand-side is bounded (as it is a finite union of bounded sets),

$$\sup_{x \in C, u \in \partial g(Tx)} \|u\| < +\infty,$$

whence the desired conclusion trivially follows.

(ii) From the equivalence [6, Proposition 16.17(i) \iff (iii)], it follows that $\text{dom}(\partial g) = \mathcal{H}_v$ and thus $TC \subset \text{dom}(\partial g)$ trivially holds. Moreover, ∂g is bounded on every bounded set of \mathcal{H}_v , and in particular on C .

(iii) In finite dimension, the claim follows trivially from (i) for the first case by a simple compactness argument, and from (ii) in the second case since a continuous and convex is bounded on bounded sets in finite dimension; see [6, Proposition 16.17]. □

Lemma 3.2. *The assumptions (A.1) and (A.2) are sufficient to ensure that*

$$\sup_{x \in C} \|\nabla f(x)\| \leq D$$

for some $D < +\infty$.

Proof. Fix $s \in C$ and let $x \in C$. We have

$$f^*(\nabla f(x)) + f(s) - \langle \nabla f(x), s \rangle = f(s) - f(x) - \langle \nabla f(x), s - x \rangle = D_f(s, x) \leq D_F(s, x) \leq K_{(F, \zeta, C)} \zeta(1),$$

where we used the Fenchel identity ([6, Proposition 17.27]) in the first equality, Remark 2.8 in the first inequality and Definition 2.7 in the second one. By [6, Corollary 9.20], f is bounded from below on C which entails

$$f^*(\nabla f(x)) - \langle \nabla f(x), s \rangle \leq D_F(s, x) \leq K_{(F, \zeta, C)} \zeta(1) + c,$$

for some real constant c . Now, since

$$s \in C \subset \text{dom} \nabla f \subset \text{int}(\text{dom} f)$$

by Definition 2.7 and [6, Proposition 17.41], we infer from [6, Theorem 14.17 and Proposition 14.16] (recall that s is fixed), that there exists $a_1 > 0$ and $a_2 \in \mathbb{R}$ such that, for all $x \in C$,

$$a_1 \|\nabla f(x)\| + a_2 \leq K_{(F,\zeta,C)} \zeta(1) + c.$$

Taking the supremum over $x \in C$ entails the desired claim with $D = a_1^{-1} (K_{(F,\zeta,C)} \zeta(1) + c - a_2)$. \square

Remark 3.3. If the dimension of \mathcal{H}_d is finite, then (A.10) is satisfied because weakly compact sets are compact in such spaces. Alternatively, another sufficient condition is to impose that the sublevel sets of the functions $(\Phi_k^* \circ (-A^*))_{k \in \mathbb{N}}$ are compact, for instance if the functions are uniformly convex, uniformly in k .

3.1.2 ASSUMPTIONS ON THE PARAMETERS AND ERROR TERMS

We impose the following assumptions on the parameters and error terms and, as with the assumptions above, for some results only a subset will be necessary:

- (P.1) $(\gamma_k)_{k \in \mathbb{N}} \subset]0, 1]$ and the sequences $(\zeta(\gamma_k))_{k \in \mathbb{N}}$, $(\gamma_k^2 / \beta_k)_{k \in \mathbb{N}}$ and $(\gamma_k \beta_k)_{k \in \mathbb{N}}$ belong to ℓ_+^1 .
- (P.2) $(\gamma_k)_{k \in \mathbb{N}} \notin \ell^1$.
- (P.3) $(\beta_k)_{k \in \mathbb{N}} \in \ell_+$ is nonincreasing and converges to 0.
- (P.4) $(\rho_k)_{k \in \mathbb{N}} \in \ell_+$ is nondecreasing with $0 < \underline{\rho} \leq \rho_k \leq \bar{\rho} < +\infty$.
- (P.5) For some positive constants \underline{M} and \bar{M} , $\underline{M} \leq (\gamma_k / \gamma_{k+1}) \leq \bar{M}$.
- (P.6) $(\theta_k)_{k \in \mathbb{N}}$ satisfies $\theta_k = \frac{\gamma_k}{c}$ for some $c > 0$ such that $\frac{\bar{M}}{c} - \frac{\rho}{2} < 0$.
- (P.7) The sequences $(\rho_k)_{k \in \mathbb{N}}$ and $(\gamma_k)_{k \in \mathbb{N}}$ satisfy $(1 - \gamma_{k+1}) \rho_{k+1} - \rho_k + \frac{2}{c} \gamma_k - \frac{\gamma_k^2}{c} \leq 0$ with c as in (P.6).
- (P.8) $(\gamma_{k+1} \mathbb{E} [\|\lambda_{k+1}\| \mid \mathcal{S}_k])_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$ and $(\gamma_{k+1} \mathbb{E} [\lambda_{k+1}^s \mid \mathcal{S}_k])_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$.
- (P.9) $(\gamma_{k+1} \mathbb{E} [\|\lambda_{k+1}\|])_{k \in \mathbb{N}} \in \ell_+^1$ and $(\gamma_{k+1} \mathbb{E} [\lambda_{k+1}^s])_{k \in \mathbb{N}} \in \ell_+^1$.

Remark 3.4. To satisfy (P.7), it suffices to take $(\rho_k)_{k \in \mathbb{N}}$ to be a constant sequence, i.e. $\rho_k \equiv \rho$, with ρ sufficiently large to satisfy $2\frac{\bar{M}}{c} < \rho$, a similar requirement as in (P.6). The condition (P.7) would then be satisfied as follows,

$$\begin{aligned} (1 - \gamma_{k+1}) \rho - \rho + \frac{2}{c} \gamma_k - \frac{\gamma_k^2}{c} &= -\gamma_{k+1} \rho + \frac{\gamma_k}{c} (2 - \gamma_k) \\ &\leq -\gamma_{k+1} \rho + \frac{2\gamma_k}{c} \\ &\leq \gamma_{k+1} \left(2\frac{\bar{M}}{c} - \rho \right) \\ &< 0. \end{aligned}$$

Remark 3.5. We will also denote the gradient of \mathcal{E}_k with errors as

$$\widehat{\nabla_x \mathcal{E}_k}(x, \mu) \stackrel{\text{def}}{=} \nabla_x \mathcal{E}_k(x, \mu) + \lambda_k.$$

It is possible to further decompose the error term λ_k , for instance, into $\lambda_k^f - T^* \lambda_k^g / \beta_k$ where λ_k^f is the error in computing $\nabla f(x_k)$ and λ_k^g is the error in evaluating $\text{prox}_{\beta_k g}(Tx_k)$. In this case, the condition $(\gamma_{k+1} \mathbb{E} [\|\lambda_{k+1}\| \mid \mathcal{S}_k])_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$ in (P.8) is sufficiently satisfied by demanding that $(\gamma_{k+1} \mathbb{E} [\|\lambda_{k+1}^f\| \mid \mathcal{S}_k])_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$ and $(\frac{\gamma_{k+1}}{\beta_{k+1}} \mathbb{E} [\|\lambda_{k+1}^g\| \mid \mathcal{S}_k])_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$.

4 MAIN RESULTS

4.1 PREPARATORY RESULTS

Lemma 4.1. *Suppose (A.1), (A.2) and (P.1) hold. For each $k \in \mathbb{N}$, define the quantity*

$$(4.1) \quad L_k \stackrel{\text{def}}{=} \frac{\|T\|^2}{\beta_k} + \|A\|^2 \rho_k.$$

Then, for each $k \in \mathbb{N}$, we have the following inequality,

$$\begin{aligned} \mathcal{E}_k(x_{k+1}, \mu_k) &\leq \mathcal{E}_k(x_k, \mu_k) + \langle \nabla_x \mathcal{E}_k(x_k, \mu_k), x_{k+1} - x_k \rangle + D_F(x_{k+1}, x_k) \\ &\quad + \frac{L_k}{2} \|x_{k+1} - x_k\|^2. \end{aligned}$$

Proof. See [34, Lemma 4.5] □

Lemma 4.2. *Suppose (A.1) and (A.2) hold. Then, for each $k \in \mathbb{N}$ and for every $x \in \mathcal{H}_p$,*

$$\mathcal{E}_k(x, \mu_k) \geq \mathcal{E}_k(x_k, \mu_k) + \langle \nabla_x \mathcal{E}_k(x_k, \mu_k), x - x_k \rangle + \frac{\rho_k}{2} \|A(x - x_k)\|^2.$$

Proof. See [34, Lemma 4.6]. □

Lemma 4.3. *Assume that (A.3) and (P.4) hold. Let $(x_k)_{k \in \mathbb{N}}$ be the sequence of primal iterates generated by Algorithm 1 and $\mathfrak{S} = (\mathcal{S}_k)_{k \in \mathbb{N}}$ as given by (3.2). Then, for each $k \in \mathbb{N}$, we have the following estimate,*

$$\frac{\rho_k}{2} \|Ax_k - b\|^2 - \frac{\rho_{k+1}}{2} \mathbb{E} [\|Ax_{k+1} - b\|^2 \mid \mathcal{S}_{k-1}] \leq \bar{\rho} d_C \|A\| (\|A\| R + \|b\|) \gamma_k \quad (\mathbb{P}\text{-a.s.}) .$$

Proof. For each $k \in \mathbb{N}$, by convexity of the function $\frac{\rho_{k+1}}{2} \|A \cdot -b\|^2$ and the assumption (P.4) that $(\rho_k)_{k \in \mathbb{N}}$ is nondecreasing, we have,

$$\begin{aligned} \frac{\rho_k}{2} \|Ax_k - b\|^2 - \frac{\rho_{k+1}}{2} \|Ax_{k+1} - b\|^2 &\leq \frac{\rho_{k+1}}{2} \|Ax_k - b\|^2 - \frac{\rho_{k+1}}{2} \|Ax_{k+1} - b\|^2 \\ &\leq \left\langle \nabla \left(\frac{\rho_{k+1}}{2} \|A \cdot -b\|^2(x_k), x_k - x_{k+1} \right) \right\rangle \\ &= \rho_{k+1} \langle Ax_k - b, A(x_k - x_{k+1}) \rangle . \end{aligned}$$

Recall that, for each $k \in \mathbb{N}$, $x_{k+1} = x_k - \gamma_k(x_k - \widehat{s}_k)$ and take the expectation to find,

$$\begin{aligned} \frac{\rho_k}{2} \|Ax_k - b\|^2 - \mathbb{E} \left[\frac{\rho_{k+1}}{2} \|Ax_{k+1} - b\|^2 \mid \mathcal{S}_{k-1} \right] &\leq \bar{\rho} \gamma_k \mathbb{E} [\langle Ax_k - b, A(x_k - \widehat{s}_k) \rangle \mid \mathcal{S}_{k-1}] \\ &\leq \bar{\rho} \gamma_k d_C \|A\| (\|A\| R + \|b\|), \end{aligned}$$

where we have used the Cauchy-Schwarz inequality and the boundedness of C , assumed in (A.3), in the last inequality. □

Remark 4.4. The above result still holds if we replace both ρ_k and ρ_{k+1} by the constant 2 and shift the index by 1, i.e., for each $k \in \mathbb{N}$,

$$\|Ax_{k+1} - b\|^2 - \mathbb{E} [\|Ax_{k+2} - b\|^2 \mid \mathcal{S}_k] \leq 2d_C \|A\| (\|A\| R + \|b\|) \gamma_{k+1} \quad (\mathbb{P}\text{-a.s.})$$

Lemma 4.5. *Suppose that (A.1)-(A.6) hold. Let $(x_k)_{k \in \mathbb{N}}$ be the sequence of primal iterates generated by Algorithm 1 and μ^* a solution, which exists by (A.6), of the dual problem, and recall the constant D from Lemma 3.2. Then, using the filtration \mathfrak{S} given in (3.2), for each $k \in \mathbb{N}$, we have the following estimate,*

$$\mathcal{L}(x_k, \mu^*) - \mathbb{E} [\mathcal{L}(x_{k+1}, \mu^*) \mid \mathcal{S}_{k-1}] \leq \gamma_k d_C \left(M \|T\| + D + L_h + \|\mu^*\| \|A\| \right) \quad (\mathbb{P}\text{-a.s.}) .$$

Proof. We recall the proof from [34, Lemma 4.7] with a slight modification to account for the inexactness of the algorithm. Define $u_k \stackrel{\text{def}}{=} [\partial g(Tx_k)]^0$ and recall that, by (A.4) and the fact that for all $k \in \mathbb{N}$, $x_k \in C$, we have $\|u_k\| \leq M$. By (A.1), the function $\Phi(x) \stackrel{\text{def}}{=} f(x) + g(Tx) + h(x)$ is convex. Then, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathcal{L}(x_k, \mu^*) - \mathcal{L}(x_{k+1}, \mu^*) &= \Phi(x_k) - \Phi(x_{k+1}) + \langle \mu^*, A(x_k - x_{k+1}) \rangle \\ &\leq \langle u_k, T(x_k - x_{k+1}) \rangle + \langle \nabla f(x_k), x_k - x_{k+1} \rangle \\ &\quad + L_h \|x_k - x_{k+1}\| + \|\mu^*\| \|A\| \|x_k - x_{k+1}\|, \end{aligned}$$

where we used the subdifferential inequality (2.2) on g and f , the L_h -Lipschitz continuity of h relative to C (see (A.5)), and the Cauchy-Schwarz inequality on the inner product. Since, for each $k \in \mathbb{N}$, $x_{k+1} = x_k + \gamma_k (\widehat{s}_k - x_k)$, we obtain, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathcal{L}(x_k, \mu^*) - \mathcal{L}(x_{k+1}, \mu^*) &\leq \gamma_k \left(\langle u_k, T(x_k - \widehat{s}_k) \rangle + \langle \nabla f(x_k), x_k - \widehat{s}_k \rangle + L_h \|x_k - \widehat{s}_k\| \right. \\ &\quad \left. + \|\mu^*\| \|A\| \|x_k - \widehat{s}_k\| \right) \end{aligned}$$

Now take the expectation with respect to the filtration \mathcal{S}_{k-1} , such that x_k is completely determined, to get, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathcal{L}(x_k, \mu^*) - \mathbb{E} [\mathcal{L}(x_{k+1}, \mu^*) \mid \mathcal{S}_{k-1}] &\leq \gamma_k \left(\mathbb{E} [\langle u_k, T(x_k - \widehat{s}_k) \rangle \mid \mathcal{S}_{k-1}] + \mathbb{E} [\langle \nabla f(x_k), x_k - \widehat{s}_k \rangle \mid \mathcal{S}_{k-1}] \right. \\ &\quad \left. + L_h \mathbb{E} [\|x_k - \widehat{s}_k\| \mid \mathcal{S}_{k-1}] + \|\mu^*\| \|A\| \mathbb{E} [\|x_k - \widehat{s}_k\| \mid \mathcal{S}_{k-1}] \right) \\ &\leq \gamma_k d_C (M \|T\| + D + L_h + \|\mu^*\| \|A\|), \end{aligned}$$

where we have used the Cauchy-Schwarz inequality, the boundedness of the set C by (A.3), the boundedness of u_k by M by (A.4), and the boundedness of $\|\nabla f(x)\|$ by D , the constant in (A.4). \square

4.2 ASYMPTOTIC FEASIBILITY

Lemma 4.6 (Feasibility estimate). *Suppose that (A.1) - (A.4) and (A.6) all hold. Consider the sequence of iterates $(x_k)_{k \in \mathbb{N}}$ generated by Algorithm 1 with parameters satisfying (P.1) and (P.3)-(P.6). For each $k \in \mathbb{N}$, define the two quantities, Δ_k^p and Δ_k^d in the following way,*

$$\Delta_k^p \stackrel{\text{def}}{=} \mathcal{L}_k(x_{k+1}, \mu_k) - \tilde{\mathcal{L}}_k(\mu_k), \quad \Delta_k^d \stackrel{\text{def}}{=} \tilde{\mathcal{L}} - \tilde{\mathcal{L}}_k(\mu_k),$$

where we have denoted $\tilde{\mathcal{L}}_k(\mu_k) \stackrel{\text{def}}{=} \min_x \mathcal{L}_k(x, \mu_k)$ and $\tilde{\mathcal{L}} \stackrel{\text{def}}{=} \mathcal{L}(x^*, \mu^*)$. Furthermore, for each $k \in \mathbb{N}$, denote the sum $\Delta_k \stackrel{\text{def}}{=} \Delta_k^p + \Delta_k^d$. We then have, using the filtration \mathfrak{S} given in (3.2), for each $k \in \mathbb{N}$,

$$\begin{aligned} (4.2) \quad \mathbb{E} [\Delta_{k+1} \mid \mathcal{S}_k] - \Delta_k &\leq -\gamma_{k+1} \left(\frac{M}{c} \|A\tilde{x}_{k+1} - b\|^2 + \delta \|A(x_{k+1} - \tilde{x}_{k+1})\|^2 \right) + \gamma_{k+1}^2 \frac{L_{k+1}}{2} d_C^2 \\ &\quad + K_{(F,\zeta,C)} \zeta (\gamma_{k+1}) + \frac{\beta_k - \beta_{k+1}}{2} M^2 + (\rho_{k+1} - \rho_k) (\|A\|^2 R^2 + \|b\|^2) \\ &\quad + \gamma_{k+1} \mathbb{E} [\lambda_{k+1}^s \mid \mathcal{S}_k] + d_C \gamma_{k+1} \mathbb{E} [\|\lambda_{k+1}\| \mid \mathcal{S}_k]. \end{aligned}$$

Proof. The proof here is adapted from the analogous result found in [34, Theorem 4.1]. As before, the quantity $\Delta_k^p \geq 0$ and can be seen as a primal gap at iteration k while Δ_k^d may be negative but is uniformly bounded from below by our assumptions (see [34, Theorem 4.1]). We denote a minimizer of $\mathcal{L}_k(x, \mu_k)$ by $\tilde{x}_k \in \underset{x \in \mathcal{H}_p}{\text{Argmin}} \mathcal{L}_k(x, \mu_k)$, which exists and belongs to \mathcal{C} by (A.1)-(A.3). We have, for each $k \in \mathbb{N}$,

$$\begin{aligned} \Delta_{k+1} - \Delta_k &= \mathcal{L}_{k+1}(x_{k+2}, \mu_{k+1}) - \mathcal{L}_k(x_{k+1}, \mu_{k+1}) + \theta_k \|Ax_{k+1} - b\|^2 \\ &\quad + 2 [\mathcal{L}_k(\tilde{x}_k, \mu_k) - \mathcal{L}_{k+1}(\tilde{x}_{k+1}, \mu_{k+1})]. \end{aligned}$$

Recall that $\tilde{x}_k \in \underset{x \in \mathcal{H}_p}{\text{Argmin}} \mathcal{L}_k(x, \mu_k)$, that $g^{\beta_k} \leq g^{\beta_{k+1}}$ due to (P.3) and Proposition 2.6(v), and that $\rho_k \leq \rho_{k+1}$ by (P.4). Then, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathcal{L}_k(\tilde{x}_k, \mu_k) - \mathcal{L}_{k+1}(\tilde{x}_{k+1}, \mu_{k+1}) &\leq \mathcal{L}_k(\tilde{x}_{k+1}, \mu_k) - \mathcal{L}_{k+1}(\tilde{x}_{k+1}, \mu_{k+1}) \\ &= [g^{\beta_k} - g^{\beta_{k+1}}] (T\tilde{x}_{k+1}) + \frac{1}{2} [\rho_k - \rho_{k+1}] \|A\tilde{x}_{k+1} - b\|^2 \\ &\quad + \langle \mu_k - \mu_{k+1}, A\tilde{x}_{k+1} - b \rangle \\ &\leq -\theta_k \langle Ax_{k+1} - b, A\tilde{x}_{k+1} - b \rangle, \end{aligned}$$

where we have used the fact that $\mu_{k+1} = \mu_k + \theta_k (Ax_{k+1} - b)$ coming from Algorithm 1. So we get, for each $k \in \mathbb{N}$,

$$\begin{aligned} \Delta_{k+1} - \Delta_k &\leq \mathcal{L}_{k+1}(x_{k+2}, \mu_{k+1}) - \mathcal{L}_k(x_{k+1}, \mu_{k+1}) + \theta_k \|Ax_{k+1} - b\|^2 \\ &\quad - 2\theta_k \langle Ax_{k+1} - b, A\tilde{x}_{k+1} - b \rangle. \end{aligned}$$

Note that, for each $k \in \mathbb{N}$,

$$\mathcal{L}_k(x_{k+1}, \mu_{k+1}) = \mathcal{L}_{k+1}(x_{k+1}, \mu_{k+1}) - [g^{\beta_{k+1}} - g^{\beta_k}] (Tx_{k+1}) - \left(\frac{\rho_{k+1} - \rho_k}{2}\right) \|Ax_{k+1} - b\|^2.$$

Then, for each $k \in \mathbb{N}$,

$$\begin{aligned} \Delta_{k+1} - \Delta_k &\leq \mathcal{L}_{k+1}(x_{k+2}, \mu_{k+1}) - \mathcal{L}_{k+1}(x_{k+1}, \mu_{k+1}) + g^{\beta_{k+1}} (Tx_{k+1}) - g^{\beta_k} (Tx_{k+1}) \\ &\quad + \left(\frac{\rho_{k+1} - \rho_k}{2}\right) \|Ax_{k+1} - b\|^2 + \theta_k \|Ax_{k+1} - b\|^2 - 2\theta_k \langle Ax_{k+1} - b, A\tilde{x}_{k+1} - b \rangle. \end{aligned}$$

We denote by $\mathbf{T1} \stackrel{\text{def}}{=} \mathcal{L}_{k+1}(x_{k+2}, \mu_{k+1}) - \mathcal{L}_{k+1}(x_{k+1}, \mu_{k+1})$ and the remaining part of the right-hand side by $\mathbf{T2}$. For the moment, we focus our attention on $\mathbf{T1}$. Recall that $\mathcal{L}_k(x, \mu_k) = \mathcal{E}_k(x, \mu_k) + h(x)$ and apply Lemma 4.1 between points x_{k+2} and x_{k+1} , to get, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathbf{T1} &\leq h(x_{k+2}) - h(x_{k+1}) + \langle \nabla_x \mathcal{E}_{k+1}(x_{k+1}, \mu_{k+1}), x_{k+2} - x_{k+1} \rangle \\ &\quad + \frac{L_{k+1}}{2} \|x_{k+2} - x_{k+1}\|^2 + D_F(x_{k+2}, x_{k+1}). \end{aligned}$$

By (A.1) we have that h is convex and thus, since x_{k+2} is a convex combination of x_{k+1} and \widehat{s}_{k+1} , we get,

for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathbf{T}_1 &\leq \gamma_{k+1} \left(h(\widehat{s}_{k+1}) - h(x_{k+1}) + \langle \nabla_x \mathcal{E}_{k+1}(x_{k+1}, \mu_{k+1}), \widehat{s}_{k+1} - x_{k+1} \rangle \right) \\ &\quad + \frac{L_{k+1}}{2} \|x_{k+2} - x_{k+1}\|^2 + D_F(x_{k+2}, x_{k+1}) \\ &= \gamma_{k+1} \left(h(\widehat{s}_{k+1}) - h(x_{k+1}) + \left\langle \widehat{\nabla}_x \mathcal{E}_{k+1}(x_{k+1}, \mu_{k+1}), \widehat{s}_{k+1} - x_{k+1} \right\rangle \right) \\ &\quad + \left\langle \nabla_x \mathcal{E}_{k+1}(x_{k+1}, \mu_{k+1}) - \widehat{\nabla}_x \mathcal{E}_{k+1}(x_{k+1}, \mu_{k+1}), \widehat{s}_{k+1} - x_{k+1} \right\rangle \\ &\quad + \frac{L_{k+1}}{2} \|x_{k+2} - x_{k+1}\|^2 + D_F(x_{k+2}, x_{k+1}) \\ &= \gamma_{k+1} \left(h(\widehat{s}_{k+1}) - h(x_{k+1}) + \left\langle \widehat{\nabla}_x \mathcal{E}_{k+1}(x_{k+1}, \mu_{k+1}), \widehat{s}_{k+1} - x_{k+1} \right\rangle \right) \\ &\quad - \langle \lambda_{k+1}, \widehat{s}_{k+1} - x_{k+1} \rangle + \frac{L_{k+1}}{2} \|x_{k+2} - x_{k+1}\|^2 + D_F(x_{k+2}, x_{k+1}). \end{aligned}$$

Applying the definition of \widehat{s}_k as the approximate minimizer of the linear minimization oracle gives, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathbf{T}_1 &\leq \gamma_{k+1} \left(h(s_{k+1}) - h(x_{k+1}) + \left\langle \widehat{\nabla}_x \mathcal{E}_{k+1}(x_{k+1}, \mu_{k+1}), s_{k+1} - x_{k+1} \right\rangle + \lambda_{k+1}^s \right) \\ &\quad - \langle \lambda_{k+1}, \widehat{s}_{k+1} - x_{k+1} \rangle + \frac{L_{k+1}}{2} \|x_{k+2} - x_{k+1}\|^2 + D_F(x_{k+2}, x_{k+1}). \end{aligned}$$

We can apply the definition of s_{k+1} as the minimizer of the linear minimization oracle and Lemma 4.2 to get, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathbf{T}_1 &\leq \gamma_{k+1} \left(h(\tilde{x}_{k+1}) - h(x_{k+1}) + \left\langle \widehat{\nabla}_x \mathcal{E}_{k+1}(x_{k+1}, \mu_{k+1}), \tilde{x}_{k+1} - x_{k+1} \right\rangle + \lambda_{k+1}^s \right) \\ &\quad - \langle \lambda_{k+1}, \widehat{s}_{k+1} - x_{k+1} \rangle + \frac{L_{k+1}}{2} \|x_{k+2} - x_{k+1}\|^2 + D_F(x_{k+2}, x_{k+1}) \\ &= \gamma_{k+1} \left(h(\tilde{x}_{k+1}) - h(x_{k+1}) + \langle \nabla_x \mathcal{E}_{k+1}(x_{k+1}, \mu_{k+1}), x_{k+1} - \tilde{x}_{k+1} \rangle + \lambda_{k+1}^s \right) \\ &\quad - \langle \lambda_{k+1}, \widehat{s}_{k+1} - \tilde{x}_{k+1} \rangle + \frac{L_{k+1}}{2} \|x_{k+2} - x_{k+1}\|^2 + D_F(x_{k+2}, x_{k+1}) \\ &\leq \gamma_{k+1} \left(h(\tilde{x}_{k+1}) - h(x_{k+1}) + \mathcal{E}_{k+1}(\tilde{x}_{k+1}, \mu_{k+1}) - \mathcal{E}_{k+1}(x_{k+1}, \mu_{k+1}) - \frac{\rho_{k+1}}{2} \|A(x_{k+1} - \tilde{x}_{k+1})\|^2 \right) \\ &\quad + \lambda_{k+1}^s - \langle \lambda_{k+1}, \widehat{s}_{k+1} - \tilde{x}_{k+1} \rangle + \frac{L_{k+1}}{2} \|x_{k+2} - x_{k+1}\|^2 + D_F(x_{k+2}, x_{k+1}) \\ &= \gamma_{k+1} \left(\mathcal{L}_{k+1}(\tilde{x}_{k+1}, \mu_{k+1}) - \mathcal{L}_{k+1}(x_{k+1}, \mu_{k+1}) - \frac{\rho_{k+1}}{2} \|A(x_{k+1} - \tilde{x}_{k+1})\|^2 + \lambda_{k+1}^s \right) \\ &\quad - \langle \lambda_{k+1}, \widehat{s}_{k+1} - \tilde{x}_{k+1} \rangle + \frac{L_{k+1}}{2} \|x_{k+2} - x_{k+1}\|^2 + D_F(x_{k+2}, x_{k+1}) \\ &\leq -\frac{\gamma_{k+1} \rho_{k+1}}{2} \|A(x_{k+1} - \tilde{x}_{k+1})\|^2 + \gamma_{k+1} \left(\lambda_{k+1}^s + \langle \lambda_{k+1}, \tilde{x}_{k+1} - \widehat{s}_{k+1} \rangle \right) \\ &\quad + \frac{L_{k+1}}{2} \|x_{k+2} - x_{k+1}\|^2 + D_F(x_{k+2}, x_{k+1}), \end{aligned}$$

where we used that \tilde{x}_{k+1} is a minimizer of $\mathcal{L}_{k+1}(\cdot, \mu_{k+1})$ in the last inequality. Combining \mathbf{T}_1 and \mathbf{T}_2 and using the Pythagoras identity we have, for each $k \in \mathbb{N}$,

$$\begin{aligned} (4.3) \quad \Delta_{k+1} - \Delta_k &\leq -\theta_k \|A\tilde{x}_{k+1} - b\|^2 + \left(\theta_k - \gamma_{k+1} \frac{\rho_{k+1}}{2} \right) \|A(x_{k+1} - \tilde{x}_{k+1})\|^2 \\ &\quad + \frac{L_{k+1}}{2} \|x_{k+2} - x_{k+1}\|^2 + D_F(x_{k+2}, x_{k+1}) + \left[g^{\beta_{k+1}} - g^{\beta_k} \right] (Tx_{k+1}) \\ &\quad + \frac{\rho_{k+1} - \rho_k}{2} \|Ax_{k+1} - b\|^2 + \gamma_{k+1} \left(\lambda_{k+1}^s + \langle \lambda_{k+1}, \tilde{x}_{k+1} - \widehat{s}_{k+1} \rangle \right). \end{aligned}$$

Now take the expectation with respect to $\mathcal{S}_k = \sigma(x_0, \mu_0, \widehat{s}_0, \dots, \widehat{s}_k)$, which completely determines x_{k+1} , \tilde{x}_{k+1} , and μ_{k+1} . We are also going to perform the following estimations.

- Under (P.5) and (P.6), we have that, for each $k \in \mathbb{N}$, $\theta_k = \gamma_k/c$ with $\underline{M}\gamma_{k+1} \leq \gamma_k$ and so that

$$-\theta_k \leq -\frac{\underline{M}}{c}\gamma_{k+1}.$$

- Again by (P.6), we have, for each $k \in \mathbb{N}$, $\theta_k = \gamma_k/c$ for some $c > 0$ such that

$$\exists \delta > 0, \quad \frac{\overline{M}}{c} - \frac{\rho}{2} = -\delta < 0,$$

where \overline{M} is the constant such that, for each $k \in \mathbb{N}$, $\gamma_k \leq \overline{M}\gamma_{k+1}$ (see (P.5)). Then, using again (P.5) and the above inequality, for each $k \in \mathbb{N}$,

$$(4.4) \quad \theta_k - \gamma_{k+1} \frac{\rho_{k+1}}{2} \leq \left(\frac{\overline{M}}{c} - \frac{\rho_{k+1}}{2} \right) \gamma_{k+1} \leq \left(\frac{\overline{M}}{c} - \frac{\rho}{2} \right) \gamma_{k+1} = -\delta \gamma_{k+1}.$$

- By Algorithm 1, for each $k \in \mathbb{N}$, $x_{k+2} - x_{k+1} = \gamma_{k+1}(\widehat{s}_{k+1} - x_{k+1})$. Since \widehat{s}_{k+1} and x_{k+1} are both in C and C is bounded due to (A.3), for each $k \in \mathbb{N}$,

$$\frac{L_{k+1}}{2} \mathbb{E} [\|x_{k+2} - x_{k+1}\|^2 \mid \mathcal{S}_k] = \frac{L_{k+1}}{2} \gamma_{k+1}^2 \mathbb{E} [\|\widehat{s}_{k+1} - x_{k+1}\|^2 \mid \mathcal{S}_k] \leq \frac{L_{k+1}}{2} \gamma_{k+1}^2 d_C^2.$$

- Recall that, by (A.2), f is (F, ζ) -smooth and invoke Remark 2.8, to get

$$\mathbb{E} [D_F(x_{k+2}, x_{k+1}) \mid \mathcal{S}_k] \leq K_{(F, \zeta, C)} \zeta(\gamma_{k+1}).$$

- By Proposition 2.6(v) and assumption (A.4),

$$\mathbb{E} \left[\left[g^{\beta_{k+1}} - g^{\beta_k} \right] (Tx_{k+1}) \mid \mathcal{S}_k \right] \leq \frac{\beta_k - \beta_{k+1}}{2} \mathbb{E} \left[\|\partial g(Tx_{k+1})\|^2 \mid \mathcal{S}_k \right] \leq \frac{\beta_k - \beta_{k+1}}{2} M^2.$$

- We also have, using Jensen's inequality and (A.3), for each $k \in \mathbb{N}$,

$$\left(\frac{\rho_{k+1} - \rho_k}{2} \right) \mathbb{E} [\|Ax_{k+1} - b\|^2 \mid \mathcal{S}_k] \leq (\rho_{k+1} - \rho_k) (\|A\|^2 R^2 + \|b\|^2).$$

In total, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathbb{E} [\Delta_{k+1} \mid \mathcal{S}_k] - \Delta_k &\leq -\frac{\underline{M}}{c} \gamma_{k+1} \|A\tilde{x}_{k+1} - b\|^2 - \delta \gamma_{k+1} \|A(x_{k+1} - \tilde{x}_{k+1})\|^2 \\ &\quad + \frac{L_{k+1}}{2} \gamma_{k+1}^2 d_C^2 + K_{(F, \zeta, C)} \zeta(\gamma_{k+1}) \\ &\quad + \frac{\beta_k - \beta_{k+1}}{2} M^2 + (\rho_{k+1} - \rho_k) (\|A\|^2 R^2 + \|b\|^2) \\ &\quad + \gamma_{k+1} \left(\mathbb{E} [\lambda_{k+1}^s \mid \mathcal{S}_k] + \mathbb{E} [\langle \lambda_{k+1}, \tilde{x}_{k+1} - \widehat{s}_{k+1} \rangle \mid \mathcal{S}_k] \right). \end{aligned}$$

Using Cauchy-Schwarz together with the fact that \tilde{x}_{k+1} and \widehat{s}_{k+1} are in C , which is bounded by (A.3), we also have, for each $k \in \mathbb{N}$,

$$(4.5) \quad \gamma_{k+1} \mathbb{E} [\langle \lambda_{k+1}, \tilde{x}_{k+1} - \widehat{s}_{k+1} \rangle \mid \mathcal{S}_k] \leq \gamma_{k+1} d_C \mathbb{E} [\|\lambda_{k+1}\| \mid \mathcal{S}_k],$$

which gives, for each $k \in \mathbb{N}$,

$$(4.6) \quad \mathbb{E} [\Delta_{k+1} \mid \mathcal{S}_k] - \Delta_k \leq -\frac{M}{c} \gamma_{k+1} \|A\tilde{x}_{k+1} - b\|^2 - \delta \gamma_{k+1} \|A(x_{k+1} - \tilde{x}_{k+1})\|^2 + \gamma_{k+1}^2 \frac{L_{k+1}}{2} d_C^2 \\ + K_{(F,\zeta,C)} \zeta(\gamma_{k+1}) + \frac{\beta_k - \beta_{k+1}}{2} M^2 + (\rho_{k+1} - \rho_k) (\|A\|^2 R^2 + \|b\|^2) \\ + \gamma_{k+1} \mathbb{E} [\lambda_{k+1}^s \mid \mathcal{S}_k] + \gamma_{k+1} d_C \mathbb{E} [\|\lambda_{k+1}\| \mid \mathcal{S}_k],$$

and (4.2) follows by trivial manipulations. □

Theorem 4.7 (Feasibility). *Suppose that (A.1)-(A.4) and (A.6) all hold and recall $\Gamma_k \stackrel{\text{def}}{=} \sum_{i=0}^k \gamma_i$. For a sequence $(x_k)_{k \in \mathbb{N}}$ generated by Algorithm 1 using parameters satisfying (P.1) - (P.6) and (P.8) we have,*

(i) *Asymptotic feasibility:* $\lim_{k \rightarrow \infty} \|Ax_k - b\| = 0$ (\mathbb{P} -a.s.).

(ii) *Ergodic rate:* let $\bar{x}_k \stackrel{\text{def}}{=} \sum_{i=0}^k \gamma_i x_i / \Gamma_k$. Then

$$(4.7) \quad \|A\bar{x}_k - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right) \text{ (\mathbb{P} -a.s.) .}$$

(iii) *It holds $(\gamma_{k+1} \|A\tilde{x}_{k+1} - b\|^2)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$ and $(\gamma_{k+1} \|Ax_{k+1} - b\|^2)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$ where \mathfrak{S} is given in (3.2).*

Additionally, if (P.9) also holds then we have the following pointwise rates in expectation,

(iv) *It holds $\inf_{0 \leq i \leq k} \mathbb{E} [\|Ax_i - b\|] \in O\left(\frac{1}{\sqrt{\Gamma_k}}\right)$.*

(v) *There exists a subsequence $(x_{k_j})_{j \in \mathbb{N}}$ such that $\mathbb{E} [\|Ax_{k_j} - b\|] \leq \frac{1}{\sqrt{\Gamma_{k_j}}}$.*

(vi) *It holds $(\gamma_k \mathbb{E} [\|A\tilde{x}_k - b\|^2])_{k \in \mathbb{N}} \in \ell_+^1$ and $(\gamma_k \mathbb{E} [\|Ax_k - b\|^2])_{k \in \mathbb{N}} \in \ell_+^1$.*

Proof. Our goal is to first apply Lemma 2.2 and then apply Lemma 2.3. By Lemma 4.6, we have, for each $k \in \mathbb{N}$,

$$(4.8) \quad \mathbb{E} [\Delta_{k+1} \mid \mathcal{S}_k] - \Delta_k \leq -\gamma_{k+1} \left(\frac{M}{c} \|A\tilde{x}_{k+1} - b\|^2 + \delta \|A(x_{k+1} - \tilde{x}_{k+1})\|^2 \right) + \gamma_{k+1}^2 \frac{L_{k+1}}{2} d_C^2 \\ + K_{(F,\zeta,C)} \zeta(\gamma_{k+1}) + \frac{\beta_k - \beta_{k+1}}{2} M^2 + (\rho_{k+1} - \rho_k) (\|A\|^2 R^2 + \|b\|^2) \\ + \gamma_{k+1} \mathbb{E} [\lambda_{k+1}^s \mid \mathcal{S}_k] + d_C \gamma_{k+1} \mathbb{E} [\|\lambda_{k+1}\| \mid \mathcal{S}_k].$$

Because of (P.1) and (P.4), and in view of the definition of L_{k+1} in (4.1), we have the following,

$$\left(\frac{L_{k+1}}{2} \gamma_{k+1}^2 d_C^2 \right)_{k \in \mathbb{N}} = \left(\frac{1}{2} \left(\frac{\|T\|^2}{\beta_{k+1}} + \|A\|^2 \rho_{k+1} \right) \gamma_{k+1}^2 d_C^2 \right)_{k \in \mathbb{N}} \in \ell_+^1.$$

For the telescopic terms from the right hand side of (4.8) we have

$$\left(\frac{\beta_k - \beta_{k+1}}{2} M^2 \right)_{k \in \mathbb{N}} \in \ell_+^1 \quad \text{and} \quad ((\rho_{k+1} - \rho_k) (\|A\|^2 R^2 + \|b\|^2))_{k \in \mathbb{N}} \in \ell_+^1,$$

where R is the constant arising from (A.3). Under (P.1) we also have that

$$(K_{(F,\zeta,C)} \zeta(\gamma_{k+1}))_{k \in \mathbb{N}} \in \ell_+^1.$$

Finally, due to (P.8), we also have

$$(\gamma_{k+1} \mathbb{E} [\lambda_{k+1}^s \mid \mathcal{S}_k])_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S}) \quad \text{and} \quad (d_C \gamma_{k+1} \mathbb{E} [\|\lambda_{k+1}\| \mid \mathcal{S}_k])_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S}).$$

Using the notation of Lemma 2.2, we set, for each $k \in \mathbb{N}$,

$$\begin{aligned} r_k &= \Delta_k, \quad a_k = \gamma_{k+1} \left(\frac{M}{c} \|A\tilde{x}_{k+1} - b\|^2 + \delta \|A(x_{k+1} - \tilde{x}_{k+1})\|^2 \right), \quad \text{and} \\ z_k &= \frac{L_{k+1}}{2} \gamma_{k+1}^2 d_C^2 + K_{(F,\zeta,C)} \zeta(\gamma_{k+1}) + \frac{\beta_k - \beta_{k+1}}{2} M^2 + \left(\frac{\rho_{k+1} - \rho_k}{2} \right) \|Ax_{k+1} - b\|^2 \\ &\quad + \gamma_{k+1} \mathbb{E} [\lambda_{k+1}^s \mid \mathcal{S}_k] + d_C \gamma_{k+1} \mathbb{E} [\|\lambda_{k+1}\| \mid \mathcal{S}_k]. \end{aligned}$$

We have shown above that , for each $k \in \mathbb{N}$,

$$\mathbb{E} [r_{k+1} \mid \mathcal{S}_k] - r_k \leq -a_k + z_k,$$

where $(z_k)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$, and r_k is bounded from below. We then deduce using Lemma 2.2 that $(r_k)_{k \in \mathbb{N}}$ is convergent (\mathbb{P} -a.s.) and

$$(4.9) \quad (\gamma_{k+1} \|A\tilde{x}_{k+1} - b\|^2)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S}) \quad \text{and} \quad (\gamma_{k+1} \|A(x_{k+1} - \tilde{x}_{k+1})\|^2)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$$

satisfying (iii). Consequently,

$$(4.10) \quad (\gamma_{k+1} \|Ax_{k+1} - b\|^2)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S}),$$

since by the Cauchy-Schwarz inequality,

$$\sum_{k=1}^{\infty} \gamma_k \|Ax_k - b\|^2 \leq 2 \sum_{k=1}^{\infty} \gamma_k (\|A(x_k - \tilde{x}_k)\|^2 + \|A\tilde{x}_k - b\|^2) < +\infty.$$

To finish proving (i) we simply apply Lemma 4.3 (with Remark 4.4) and the conditions of Lemma 2.3 are satisfied. Then, (ii) follows directly from the application of Jensen’s inequality as in the results of [34, Theorem 4.1].

We now assume that (P.9) holds. By Lemma 4.6, we can take the total expectation and use the law of total expectation to have, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathbb{E} [\Delta_{k+1}] - \mathbb{E} [\Delta_k] &\leq -\gamma_{k+1} \left(\frac{M}{c} \mathbb{E} [\|A\tilde{x}_{k+1} - b\|^2] + \delta \mathbb{E} [\|A(x_{k+1} - \tilde{x}_{k+1})\|^2] \right) + \gamma_{k+1}^2 \frac{L_{k+1}}{2} d_C^2 \\ &\quad + K_{(F,\zeta,C)} \zeta(\gamma_{k+1}) + \frac{\beta_k - \beta_{k+1}}{2} M^2 + (\rho_{k+1} - \rho_k) (\|A\|^2 R^2 + \|b\|^2) \\ &\quad + \gamma_{k+1} \mathbb{E} [\lambda_{k+1}^s] + d_C \gamma_{k+1} \mathbb{E} [\|\lambda_{k+1}\|]. \end{aligned}$$

Define the following, for each $k \in \mathbb{N}$,

$$\begin{aligned} \tilde{r}_k &= \mathbb{E} [\Delta_k], \quad \tilde{p}_k = \gamma_{k+1}, \quad \tilde{w}_k = \left(\frac{M}{c} \mathbb{E} [\|A\tilde{x}_{k+1} - b\|^2] + \delta \mathbb{E} [\|A(x_{k+1} - \tilde{x}_{k+1})\|^2] \right), \quad \text{and} \\ \tilde{z}_k &= \frac{L_{k+1}}{2} \gamma_{k+1}^2 d_C^2 + K_{(F,\zeta,C)} \zeta(\gamma_{k+1}) + \frac{\beta_k - \beta_{k+1}}{2} M^2 + \left(\frac{\rho_{k+1} - \rho_k}{2} \right) \mathbb{E} [\|Ax_{k+1} - b\|^2] \\ &\quad + \gamma_{k+1} \mathbb{E} [\lambda_{k+1}^s] + d_C \gamma_{k+1} \mathbb{E} [\|\lambda_{k+1}\|]. \end{aligned}$$

By the argument of the previous paragraph, in conjunction with (P.9), we have that $(\tilde{z}_k)_{k \in \mathbb{N}} \in \ell_+^1$. We can apply the total expectation to the results of both Lemma 4.3 and Lemma 4.5 and then the claims of interest follow from Lemma 2.4 applied with $(\tilde{r}_k)_{k \in \mathbb{N}}$, $(\tilde{p}_k)_{k \in \mathbb{N}}$, $(\tilde{w}_k)_{k \in \mathbb{N}}$, and $(\tilde{z}_k)_{k \in \mathbb{N}}$ defined as above. □

4.3 OPTIMALITY

The following lemmas regard the boundedness of the sequence of dual iterates $(\mu_k)_{k \in \mathbb{N}}$ and the uniform boundedness of the Lagrangian. They were shown in the deterministic setting in [34] and easily extend to the stochastic case in light of Theorem 4.7.

Lemma 4.8. *Suppose that (A.1)-(A.3), (A.6)-(A.8), and (P.1)-(P.6) all hold and define, for each $k \in \mathbb{N}$,*

$$(4.11) \quad \varphi_k(\mu) \stackrel{\text{def}}{=} - \inf_{x \in \mathcal{H}_p} \mathcal{L}_k(x, \mu) \quad \text{and} \quad \bar{\varphi} \stackrel{\text{def}}{=} f(x) + g(Tx) + h(x) + \frac{\bar{\rho}}{2} \|Ax - b\|^2.$$

Then the sequence of dual iterates $(\mu_k)_{k \in \mathbb{N}}$ generated by Algorithm 1 is bounded (\mathbb{P} -a.s.), for each $k \in \mathbb{N}$ the function $\varphi_k(\mu)$ is convex and differentiable with gradient

$$(4.12) \quad \nabla \varphi_k(\mu) = \rho_k^{-1} \left(\mu - \text{prox}_{\rho_k \Phi_k^* \circ (-A^*)}(\mu - \rho_k b) \right),$$

and it holds, for each $k \in \mathbb{N}$,

$$(4.13) \quad \nabla \varphi_k(\mu_k) = A\tilde{x}_k - b.$$

Proof. Note that here we have denoted, for each $k \in \mathbb{N}$, $\Phi_k(x) = f(x) + g^{\beta_k}(x) + h(x)$ while in [34] it is defined differently with $\phi_k(x) = f(x) + g^{\beta_k}(x) + h(x)$ and $\Phi_k(x)$ having different meaning. For brevity, we defer to the proof in [34, Lemma 4.11], noting that since $(\gamma_{k+1} \|\tilde{x}_{k+1} - b\|^2)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$ and $(\gamma_{k+1} \|x_{k+1} - b\|^2)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$, there exists $\tilde{\Omega} \subset \mathcal{F}$ with $\mathbb{P}(\tilde{\Omega}) = 1$ such that $(\varphi_k(\mu_k(\omega)))_{k \in \mathbb{N}}$ is convergent and thus bounded, and the uniform coercivity of $(\varphi_k)_{k \in \mathbb{N}}$ is unaffected by the inexactness, i.e., $(\mu_k(\omega))_{k \in \mathbb{N}}$ is bounded. \square

Lemma 4.9. *Under (A.1)-(A.8) and (P.1)-(P.6), the composite function $f + g \circ T + h$ is uniformly bounded on C and we have*

$$(4.14) \quad \tilde{M} \stackrel{\text{def}}{=} \sup_{x \in C} |f(x) + g(Tx) + h(x)| + \sup_{k \in \mathbb{N}} \|\mu_k\| (\|A\|R + b) < +\infty \quad (\mathbb{P}\text{-a.s.}),$$

where R is the radius from (A.3).

Proof. The proof follows in a (\mathbb{P} -a.s.) sense from [34, Lemma 4.12] with the addition of Theorem 4.7. \square

We now begin with the main energy estimate needed to show the convergence of the Lagrangian values to optimality.

Lemma 4.10 (Optimality estimate). *Recall the constants c, L_k, M, D , and L_h from (P.6), Lemma 4.1, (A.4), Lemma 4.3, and (A.5), respectively. Define, for each $k \in \mathbb{N}$,*

$$r_k \stackrel{\text{def}}{=} (1 - \gamma_k) \mathcal{L}_k(x_k, \mu_k) + \frac{c}{2} \|\mu_k - \mu^*\|^2$$

and

$$C_k \stackrel{\text{def}}{=} \frac{L_k}{2} d_C^2 + d_C (M\|T\| + D + L_h + \|\mu^*\| \|A\|).$$

Then, under (A.1)-(A.8) and (P.1)-(P.7) with $\underline{M} \geq 1$, for the sequences $(x_k)_{k \in \mathbb{N}}$ and $(\mu_k)_{k \in \mathbb{N}}$ generated by Algorithm 1, using the filtration $\mathfrak{S}' = (\mathcal{S}_{k-1})_{k \in \mathbb{N}}$ with $\mathcal{S}_k \stackrel{\text{def}}{=} \sigma(x_0, \mu_0, \widehat{s}_0, \dots, \widehat{s}_k)$ as before, the following inequality holds, for each $k \in \mathbb{N}$ with $k > 0$,

$$(4.15) \quad \mathbb{E}[r_{k+1} \mid \mathcal{S}_{k-1}] - r_k \leq -\gamma_k \left(\mathcal{L}(x_k, \mu^*) - \mathcal{L}(x^*, \mu^*) + \frac{\rho_k}{2} \|Ax_k - b\|^2 \right) + (\beta_k - \beta_{k+1}) \frac{M^2}{2} \\ + (\gamma_k - \gamma_{k+1}) \tilde{M} + \gamma_k \beta_k \frac{M^2}{2} + K_{(F, \zeta, C)} \zeta(\gamma_k) + \gamma_k^2 C_k + d_C \gamma_k \mathbb{E}[\|\lambda_k\| \mid \mathcal{S}_{k-1}] \\ + \gamma_k \mathbb{E}[\lambda_k^s \mid \mathcal{S}_{k-1}] \quad (\mathbb{P}\text{-a.s.}).$$

Proof. Applying Lemma 4.2 to the points x^* and x_k we have, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathcal{E}_k(x^*, \mu_k) &\geq \mathcal{E}_k(x_k, \mu_k) + \langle \nabla_x \mathcal{E}_k(x_k, \mu_k), x^* - x_k \rangle + \frac{\rho_k}{2} \|A(x^* - x_k)\|^2 \\ &= \mathcal{E}_k(x_k, \mu_k) + \left\langle \widehat{\nabla}_x \mathcal{E}_k(x_k, \mu_k), x^* - x_k \right\rangle + \langle \lambda_k, x_k - x^* \rangle + \frac{\rho_k}{2} \|A(x^* - x_k)\|^2 \\ &= \mathcal{E}_k(x_k, \mu_k) + \left\langle \widehat{\nabla}_x \mathcal{E}_k(x_k, \mu_k), x^* - x_k \right\rangle + h(x^*) - h(x_k) + \langle \lambda_k, x_k - x^* \rangle \\ &\quad + \frac{\rho_k}{2} \|A(x^* - x_k)\|^2. \end{aligned}$$

By the definition of s_k as a minimizer and the definition of \widehat{s}_k we further have, for each $k \in \mathbb{N}$,

$$\begin{aligned} (4.16) \quad \mathcal{E}_k(x^*, \mu_k) &\geq \mathcal{E}_k(x_k, \mu_k) + \left\langle \widehat{\nabla}_x \mathcal{E}_k(x_k, \mu_k), s_k - x_k \right\rangle + h(s_k) - h(x^*) + \langle \lambda_k, x_k - x^* \rangle \\ &\quad + \frac{\rho_k}{2} \|A(x^* - x_k)\|^2 \\ &\geq \mathcal{E}_k(x_k, \mu_k) + \left\langle \widehat{\nabla}_x \mathcal{E}_k(x_k, \mu_k), \widehat{s}_k - x_k \right\rangle + h(\widehat{s}_k) - \lambda_k^s - h(x^*) + \langle \lambda_k, x_k - x^* \rangle \\ &\quad + \frac{\rho_k}{2} \|A(x^* - x_k)\|^2. \end{aligned}$$

From Lemma 4.1 applied to the points x_{k+1} and x_k and by definition of $x_{k+1} \stackrel{\text{def}}{=} x_k + \gamma_k (\widehat{s}_k - x_k)$ in Algorithm 1, we also have, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathcal{E}_k(x_{k+1}, \mu_k) &\leq \mathcal{E}_k(x_k, \mu_k) + \langle \nabla_x \mathcal{E}_k(x_k, \mu_k), x_{k+1} - x_k \rangle + D_F(x_{k+1}, x_k) + \frac{L_k}{2} \|x_{k+1} - x_k\|^2 \\ &= \mathcal{E}_k(x_k, \mu_k) + \gamma_k \langle \nabla_x \mathcal{E}_k(x_k, \mu_k), \widehat{s}_k - x_k \rangle + D_F(x_{k+1}, x_k) + \gamma_k^2 \frac{L_k}{2} \|\widehat{s}_k - x_k\|^2 \\ &= \mathcal{E}_k(x_k, \mu_k) + \gamma_k \langle \widehat{\nabla}_x \mathcal{E}_k(x_k, \mu_k), \widehat{s}_k - x_k \rangle + \gamma_k \langle \lambda_k, x_k - \widehat{s}_k \rangle + D_F(x_{k+1}, x_k) \\ &\quad + \gamma_k^2 \frac{L_k}{2} \|\widehat{s}_k - x_k\|^2. \end{aligned}$$

We combine the latter with (4.16), to get, for each $k \in \mathbb{N}$,

$$\begin{aligned} (4.17) \quad \mathcal{E}_k(x_{k+1}, \mu_k) &\leq \mathcal{E}_k(x_k, \mu_k) + \gamma_k \langle \lambda_k, x^* - \widehat{s}_k \rangle + D_F(x_{k+1}, x_k) + \gamma_k^2 \frac{L_k}{2} \|\widehat{s}_k - x_k\|^2 \\ &\quad + \gamma_k \left(\mathcal{E}_k(x^*, \mu_k) + h(x^*) - \mathcal{E}_k(x_k, \mu_k) - h(\widehat{s}_k) - \frac{\rho_k}{2} \|Ax_k - b\|^2 + \lambda_k^s \right). \end{aligned}$$

By convexity of h from (A.1) and the definition of x_{k+1} , we have, for each $k \in \mathbb{N}$,

$$\begin{aligned} (4.18) \quad \mathcal{L}_k(x_{k+1}, \mu_k) - \mathcal{L}_k(x_k, \mu_k) &= \mathcal{E}_k(x_{k+1}, \mu_k) - \mathcal{E}_k(x_k, \mu_k) + h(x_{k+1}) - h(x_k) \\ &\leq \mathcal{E}_k(x_{k+1}, \mu_k) - \mathcal{E}_k(x_k, \mu_k) + \gamma_k (h(\widehat{s}_k) - h(x_k)) \end{aligned}$$

Combining (4.17) and (4.18), we obtain, for each $k \in \mathbb{N}$,

$$\begin{aligned} (4.19) \quad \mathcal{L}_k(x_{k+1}, \mu_k) - \mathcal{L}_k(x_k, \mu_k) &\leq \gamma_k (\mathcal{E}_k(x^*, \mu_k) + h(x^*) - \mathcal{E}_k(x_k, \mu_k) - h(x_k)) + D_F(x_{k+1}, x_k) + \\ &\quad \gamma_k^2 \frac{L_k}{2} \|\widehat{s}_k - x_k\|^2 + \gamma_k \left(\langle \lambda_k, x^* - \widehat{s}_k \rangle - \frac{\rho_k}{2} \|Ax_k - b\|^2 + \lambda_k^s \right) \\ &= \gamma_k (\mathcal{L}_k(x^*, \mu_k) - \mathcal{L}_k(x_k, \mu_k)) + D_F(x_{k+1}, x_k) + \gamma_k^2 \frac{L_k}{2} \|\widehat{s}_k - x_k\|^2 \\ &\quad + \gamma_k \left(\langle \lambda_k, x^* - \widehat{s}_k \rangle - \frac{\rho_k}{2} \|Ax_k - b\|^2 + \lambda_k^s \right) \end{aligned}$$

Recalling the definition of $\mu_{k+1} \stackrel{\text{def}}{=} \mu_k + A(x_{k+1} - b)$ in Algorithm 1, we have, for each $k \in \mathbb{N}$,

$$\mathcal{L}_k(x_{k+1}, \mu_{k+1}) - \mathcal{L}_k(x_{k+1}, \mu_k) = \langle \mu_{k+1} - \mu_k, Ax_{k+1} \rangle = \theta_k \|Ax_{k+1} - b\|^2.$$

We combine the above and (4.19) to get, for each $k \in \mathbb{N}$,

$$(4.20) \quad \begin{aligned} \mathcal{L}_k(x_{k+1}, \mu_{k+1}) - \mathcal{L}_k(x_k, \mu_k) &\leq \theta_k \|Ax_{k+1} - b\|^2 + \gamma_k (\mathcal{L}_k(x^*, \mu_k) - \mathcal{L}_k(x_k, \mu_k)) + D_F(x_{k+1}, x_k) \\ &\quad + \gamma_k^2 \frac{L_k}{2} \|\widehat{s}_k - x_k\|^2 + \gamma_k \left(\langle \lambda_k, x^* - \widehat{s}_k \rangle - \frac{\rho_k}{2} \|Ax_k - b\|^2 + \lambda_k^s \right). \end{aligned}$$

Notice that the update of the dual variable μ can be interpreted as a proximal operator in the following way,

$$\mu_{k+1} = \operatorname{argmin}_{\mu \in \mathcal{H}_d} \left\{ -\mathcal{L}_k(x_{k+1}, \mu) + \frac{1}{2\theta_k} \|\mu - \mu_k\|^2 \right\}.$$

Then, using Lemma 2.5, we get, for each $k \in \mathbb{N}$,

$$(4.21) \quad \begin{aligned} 0 &\geq \theta_k (\mathcal{L}_k(x_{k+1}, \mu^*) - \mathcal{L}_k(x_{k+1}, \mu_{k+1})) + \frac{1}{2} \left(\|\mu_{k+1} - \mu^*\|^2 - \|\mu_k - \mu^*\|^2 + \|\mu_{k+1} - \mu_k\|^2 \right) \\ &= \theta_k (\mathcal{L}_k(x_{k+1}, \mu^*) - \mathcal{L}_k(x_{k+1}, \mu_{k+1})) + \frac{1}{2} \left(\|\mu_{k+1} - \mu^*\|^2 - \|\mu_k - \mu^*\|^2 + \theta_k^2 \|Ax_{k+1} - b\|^2 \right). \end{aligned}$$

Recall that, by (P.6), $\theta_k = \gamma_k/c$. Multiply (4.21) by c and sum with (4.20), to obtain, for each $k \in \mathbb{N}$,

$$\begin{aligned} &(1 - c\theta_k) \mathcal{L}_k(x_{k+1}, \mu_{k+1}) - (1 - c\theta_k) \mathcal{L}_k(x_k, \mu_k) + \frac{c}{2} \left(\|\mu_{k+1} - \mu^*\|^2 - \|\mu_k - \mu^*\|^2 \right) \\ &\leq \left(\theta_k - \frac{c\theta_k^2}{2} \right) \|Ax_{k+1} - b\|^2 + \gamma_k (\mathcal{L}_k(x^*, \mu_k) - \mathcal{L}_k(x_k, \mu_k)) - c\theta_k (\mathcal{L}_k(x_{k+1}, \mu) - \mathcal{L}_k(x_k, \mu_k)) \\ &\quad - \frac{\rho_k \gamma_k}{2} \|Ax_k - b\|^2 + D_F(x_{k+1}, x_k) + \gamma_k^2 \frac{L_k}{2} \|\widehat{s}_k - x_k\|^2 + \gamma_k \left(\langle \lambda_k, x^* - \widehat{s}_k \rangle + \lambda_k^s \right). \end{aligned}$$

The previous inequality can be re-written, by trivial manipulations, as, for each $k \in \mathbb{N}$,

$$(4.22) \quad \begin{aligned} &(1 - c\theta_{k+1}) \mathcal{L}_{k+1}(x_{k+1}, \mu_{k+1}) - (1 - c\theta_k) \mathcal{L}_k(x_k, \mu_k) + \frac{c}{2} \left(\|\mu_{k+1} - \mu^*\|^2 - \|\mu_k - \mu^*\|^2 \right) \\ &\leq (1 - c\theta_{k+1}) \mathcal{L}_{k+1}(x_{k+1}, \mu_{k+1}) - (1 - c\theta_k) \mathcal{L}_k(x_{k+1}, \mu_{k+1}) + \left(\theta_k - \frac{c\theta_k^2}{2} \right) \|Ax_{k+1} - b\|^2 \\ &\quad + \gamma_k (\mathcal{L}_k(x^*, \mu_k) - \mathcal{L}_k(x_k, \mu_k)) - c\theta_k (\mathcal{L}_k(x_{k+1}, \mu^*) - \mathcal{L}_k(x_k, \mu_k)) - \frac{\rho_k \gamma_k}{2} \|Ax_k - b\|^2 \\ &\quad + D_F(x_{k+1}, x_k) + \gamma_k^2 \frac{L_k}{2} \|\widehat{s}_k - x_k\|^2 + \gamma_k \left(\langle \lambda_k, x^* - \widehat{s}_k \rangle + \lambda_k^s \right) \\ &= c(\theta_k - \theta_{k+1})(f + h + \langle \mu_{k+1}, A \cdot -b \rangle)(x_{k+1}) + \left((1 - c\theta_{k+1})g^{\beta_{k+1}} - (1 - c\theta_k)g^{\beta_k} \right)(Tx_{k+1}) \\ &\quad + \frac{1}{2} \left((1 - c\theta_{k+1})\rho_{k+1} - (1 - c\theta_k)\rho_k + 2\theta_k - c\theta_k^2 \right) \|Ax_{k+1} - b\|^2 + \gamma_k (\mathcal{L}_k(x^*, \mu_k) - \mathcal{L}_k(x_k, \mu_k)) \\ &\quad - c\theta_k (\mathcal{L}_k(x_{k+1}, \mu^*) - \mathcal{L}_k(x_k, \mu_k)) - \frac{\rho_k \gamma_k}{2} \|Ax_k - b\|^2 + D_F(x_{k+1}, x_k) + \gamma_k^2 \frac{L_k}{2} \|\widehat{s}_k - x_k\|^2 \\ &\quad + \gamma_k \left(\langle \lambda_k, x^* - \widehat{s}_k \rangle + \lambda_k^s \right). \end{aligned}$$

By (P.5), (P.6) and the assumption that $\underline{M} \geq 1$, we have $\theta_{k+1} \leq \underline{M}^{-1}\theta_k \leq \theta_k$. In view of (P.3), we also have $\beta_{k+1} \leq \beta_k$. In particular, $g^{\beta_k} \leq g^{\beta_{k+1}} \leq g$ pointwise. By Proposition 2.6(iv) and assumption (A.4),

we are able to, for each $k \in \mathbb{N}$, estimate the quantity

$$\begin{aligned} & \left((1 - c\theta_{k+1}) g^{\beta_{k+1}} - (1 - c\theta_k) g^{\beta_k} \right) (Tx_{k+1}) \\ &= \left(g^{\beta_{k+1}} - g^{\beta_k} \right) (Tx_{k+1}) + c \left(\theta_k g^{\beta_k} - \theta_{k+1} g^{\beta_{k+1}} \right) (Tx_{k+1}) \\ &\leq \frac{1}{2} (\beta_k - \beta_{k+1}) \left\| (\partial g(Tx_{k+1}))^0 \right\|^2 + c \left(\theta_k g^{\beta_k} - \theta_{k+1} g^{\beta_{k+1}} \right) (Tx_{k+1}) \\ &\leq \frac{1}{2} (\beta_k - \beta_{k+1}) \left\| (\partial g(Tx_{k+1}))^0 \right\|^2 + c (\theta_k - \theta_{k+1}) g(Tx_{k+1}). \end{aligned}$$

Then, for each $k \in \mathbb{N}$,

$$\begin{aligned} (4.23) \quad & c (\theta_k - \theta_{k+1}) (f + h + \langle \mu_{k+1}, A \cdot - b \rangle) (x_{k+1}) + \left((1 - c\theta_{k+1}) g^{\beta_{k+1}} - (1 - c\theta_k) g^{\beta_k} \right) (Tx_{k+1}) \\ &\leq c (\theta_k - \theta_{k+1}) \mathcal{L} (x_{k+1}, \mu_{k+1}) + \frac{1}{2} (\beta_k - \beta_{k+1}) \left\| (\partial g(Tx_{k+1}))^0 \right\|^2. \end{aligned}$$

Recall the definition of r_k in (4.10). Coming back to (4.22) and using (4.23), we obtain, for each $k \in \mathbb{N}$, (4.24)

$$\begin{aligned} r_{k+1} - r_k &\leq \frac{1}{2} \left((1 - \gamma_{k+1}) \rho_{k+1} - (1 - \gamma_k) \rho_k + \frac{2}{c} \gamma_k - \frac{\gamma_k^2}{c} \right) \|Ax_{k+1} - b\|^2 \\ &\quad + \gamma_k (\mathcal{L}_k (x^*, \mu_k) - \mathcal{L}_k (x_{k+1}, \mu^*)) - \frac{\rho_k \gamma_k}{2} \|Ax_k - b\|^2 + \frac{\beta_k - \beta_{k+1}}{2} \left\| (\partial g(Tx_{k+1}))^0 \right\|^2 \\ &\quad + (\gamma_k - \gamma_{k+1}) \mathcal{L} (x_{k+1}, \mu_{k+1}) + D_F (x_{k+1}, x_k) + \gamma_k^2 \frac{L_k}{2} \|\widehat{s}_k - x_k\|^2 + \gamma_k \left(\langle \lambda_k, x^* - \widehat{s}_k \rangle + \lambda_k^s \right). \end{aligned}$$

Recall that, by feasibility of x^* for the affine constraint, $\mathcal{L} (x^*, \mu_k) = \mathcal{L} (x^*, \mu^*)$ and thus, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathcal{L}_k (x^*, \mu_k) - \mathcal{L}_k (x_{k+1}, \mu^*) &= \mathcal{L} (x^*, \mu^*) - \mathcal{L} (x_{k+1}, \mu^*) + \left(g^{\beta_k} - g \right) (Tx^*) + \left(g - g^{\beta_k} \right) (Tx_{k+1}) \\ &\quad - \frac{\rho_k}{2} \|Ax_{k+1} - b\|^2 \\ &= \mathcal{L} (x^*, \mu^*) - \mathcal{L} (x_k, \mu^*) + \mathcal{L} (x_k, \mu^*) - \mathcal{L} (x_{k+1}, \mu^*) \\ &\quad + \left(g^{\beta_k} - g \right) (Tx^*) + \left(g - g^{\beta_k} \right) (Tx_{k+1}) - \frac{\rho_k}{2} \|Ax_{k+1} - b\|^2 \\ &\leq \mathcal{L} (x^*, \mu^*) - \mathcal{L} (x_k, \mu^*) + \mathcal{L} (x_k, \mu^*) - \mathcal{L} (x_{k+1}, \mu^*) + \frac{\beta_k}{2} \left\| (\partial g(Tx_{k+1}))^0 \right\|^2 \\ &\quad - \frac{\rho_k}{2} \|Ax_{k+1} - b\|^2, \end{aligned}$$

where in the inequality we have used the fact that $g^{\beta_k} \leq g$ pointwise and that, by Proposition 2.6(v), for each $k \in \mathbb{N}$,

$$\left(g - g^{\beta_k} \right) (Tx_{k+1}) \leq \frac{\beta_k}{2} \left\| (\partial g(Tx_{k+1}))^0 \right\|^2.$$

Substituting the above into (4.24) we have, for each $k \in \mathbb{N}$,

$$\begin{aligned} (4.25) \quad r_{k+1} - r_k &\leq \frac{1}{2} \left((1 - \gamma_{k+1}) \rho_{k+1} - \rho_k + \frac{2}{c} \gamma_k - \frac{\gamma_k^2}{c} \right) \|Ax_{k+1} - b\|^2 \\ &\quad + \gamma_k (\mathcal{L} (x^*, \mu^*) - \mathcal{L} (x_k, \mu^*)) + \gamma_k (\mathcal{L} (x_k, \mu^*) - \mathcal{L} (x_{k+1}, \mu^*)) \\ &\quad - \frac{\rho_k \gamma_k}{2} \|Ax_k - b\|^2 + \frac{\beta_k - \beta_{k+1}}{2} \left\| (\partial g(Tx_{k+1}))^0 \right\|^2 + (\gamma_k - \gamma_{k+1}) \mathcal{L} (x_{k+1}, \mu_{k+1}) \\ &\quad + \gamma_k \frac{\beta_k}{2} \left\| (\partial g(Tx_{k+1}))^0 \right\|^2 + D_F (x_{k+1}, x_k) + \gamma_k^2 \frac{L_k}{2} \|\widehat{s}_k - x_k\|^2 \\ &\quad + \gamma_k \left(\langle \lambda_k, x^* - \widehat{s}_k \rangle + \lambda_k^s \right). \end{aligned}$$

Take the expectation with respect to $\mathcal{S}_{k-1} \stackrel{\text{def}}{=} \sigma(x_0, \mu_0, \widehat{s}_0, \dots, \widehat{s}_{k-1})$, which will completely determine x_k and μ_k , and perform the following estimations.

- From (P.7), we have, for each $k \in \mathbb{N}$,

$$(1 - \gamma_{k+1}) \rho_{k+1} - \rho_k + \frac{2}{c} \gamma_k - \frac{\gamma_k^2}{c} \leq 0.$$

- By assumption (A.4), for each $k \in \mathbb{N}$,

$$\mathbb{E} \left[\left\| (\partial g(Tx_{k+1}))^0 \right\|^2 \mid \mathcal{S}_{k-1} \right] \leq M^2.$$

- By Lemma 4.9, for each $k \in \mathbb{N}$,

$$\mathbb{E} [\mathcal{L}(x_{k+1}, \mu_{k+1}) \mid \mathcal{S}_{k-1}] \leq \tilde{M}.$$

- Recall that, by (A.2), f is (F, ζ) -smooth and invoke Remark 2.8, to get, for each $k \in \mathbb{N}$,

$$\mathbb{E} [D_F(x_{k+1}, x_k) \mid \mathcal{S}_{k-1}] \leq K_{(F, \zeta, C)} \zeta(\gamma_k).$$

- Since, for each $k \in \mathbb{N}$, \widehat{s}_k and x_k are both in C , we have

$$\mathbb{E} [\|\widehat{s}_k - x_k\| \mid \mathcal{S}_{k-1}] \leq d_C.$$

We have, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathbb{E} [r_{k+1} \mid \mathcal{S}_{k-1}] - r_k &\leq \gamma_k (\mathcal{L}(x^*, \mu^*) - \mathcal{L}(x_k, \mu^*)) + \gamma_k (\mathcal{L}(x_k, \mu^*) - \mathbb{E} [\mathcal{L}(x_{k+1}, \mu^*) \mid \mathcal{S}_{k-1}]) \\ &\quad - \frac{\rho_k \gamma_k}{2} \|Ax_k - b\|^2 + \frac{\beta_k - \beta_{k+1}}{2} M^2 + (\gamma_k - \gamma_{k+1}) \tilde{M} + \gamma_k \frac{\beta_k}{2} M^2 \\ &\quad + K_{(F, \zeta, C)} \zeta(\gamma_k) + \gamma_k^2 \frac{L_k}{2} d_C^2 + \gamma_k \mathbb{E} [\langle \lambda_k, x^* - \widehat{s}_k \rangle + \lambda_k^s \mid \mathcal{S}_{k-1}]. \end{aligned}$$

We can bound the inner product involving the error terms using the Cauchy-Schwarz inequality and the boundedness of C . Applying Lemma 4.5 and regrouping terms with γ_k^2 we get, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathbb{E} [r_{k+1} \mid \mathcal{S}_{k-1}] - r_k &\leq \gamma_k (\mathcal{L}(x^*, \mu^*) - \mathcal{L}(x_k, \mu^*)) - \frac{\rho_k \gamma_k}{2} \|Ax_k - b\|^2 + (\beta_k - \beta_{k+1}) \frac{M^2}{2} + \gamma_k^2 C_k \\ &\quad + (\gamma_k - \gamma_{k+1}) \tilde{M} + \gamma_k \beta_k \frac{M^2}{2} + K_{(F, \zeta, C)} \zeta(\gamma_k) + \gamma_k \mathbb{E} [d_C \|\lambda_k\| + \lambda_k^s \mid \mathcal{S}_{k-1}]. \end{aligned}$$

We conclude by trivial manipulations. □

We now proceed to prove the main theorem regarding optimality, recalling the notation of (3.3) for the terms $\mathcal{S}_{\mathcal{D}}$ and $\mathcal{S}_{\mathcal{D}}$ and (3.4) for $\mathfrak{B}[(x_k)_{k \in \mathbb{N}}]$. The convergence and rates of the Lagrangian values will be shown in terms $\mathcal{L}(x_k, \mu^*) - \mathcal{L}(x^*, \mu^*)$, which is non-negative since (x^*, μ^*) is a saddle point. This is however not a primal-dual gap. Nevertheless, observe that in view of [6, Proposition 19.21(v)], we have

$$\mathcal{L}(x_k, \mu^*) - \mathcal{L}(x^*, \mu^*) = \Phi(x) - \Phi(x^*) + \langle A^* \mu^*, x_k - x^* \rangle,$$

which is nothing but the Bregman divergence of Φ with the subgradient $-A^* \mu$ between x_k and x^* . This Bregman divergence appears then as a good candidate to quantify the convergence rate of Algorithm 1 given that it captures both the discrepancy of the primal objective to the optimal value and violation of the affine constraint.

Theorem 4.11 (Optimality). *Suppose that (A.1)-(A.10) and (P.1)-(P.8) hold, with $\underline{M} \geq 1$. Let $(x_k)_{k \in \mathbb{N}}$ be the sequence of primal iterates generated by Algorithm 1 and (x^*, μ^*) a saddle-point pair for the Lagrangian. Then, in addition to the results of Theorem 4.7, the following holds*

(i) *Convergence of the Lagrangian:*

$$(4.26) \quad \lim_{k \rightarrow \infty} \mathcal{L}(x_k, \mu^*) = \mathcal{L}(x^*, \mu^*) \quad (\mathbb{P}\text{-a.s.}).$$

(ii) *The sequence $(x_k)_{k \in \mathbb{N}}$ satisfies $\mathfrak{B}[(x_k)_{k \in \mathbb{N}}] \subset \mathcal{S}_{\mathcal{D}}$ (\mathbb{P} -a.s.) and there exists $\bar{\mu}$, an $\mathcal{S}_{\mathcal{D}}$ -valued random variable, such that $\mu_k \rightarrow \bar{\mu}$ (\mathbb{P} -a.s.).*

(iii) *Ergodic rate: for each $k \in \mathbb{N}$, let $\bar{x}_k \stackrel{\text{def}}{=} \sum_{i=0}^k \gamma_i x_{i+1} / \Gamma_k$. Then, for each $k \in \mathbb{N}$,*

$$(4.27) \quad \mathcal{L}(\bar{x}_k, \mu^*) - \mathcal{L}(x^*, \mu^*) \in O\left(\frac{1}{\Gamma_k}\right) \quad (\mathbb{P}\text{-a.s.}).$$

(iv) *If the problem (\mathcal{P}) admits a unique solution x^* , then $(x_k)_{k \in \mathbb{N}}$ converges weakly (\mathbb{P} -a.s.) to x^* a solution of (\mathcal{P}) . Moreover, if Φ is uniformly convex on C with modulus of convexity $\psi : \mathbb{R}_+ \rightarrow [0, \infty]$, then $(x_k)_{k \in \mathbb{N}}$ converges strongly (\mathbb{P} -a.s.) to x^* at the ergodic rate, for each $k \in \mathbb{N}$,*

$$\psi\left(\|\bar{x}_k - x^*\|\right) \in O\left(\frac{1}{\Gamma_k}\right) \quad (\mathbb{P}\text{-a.s.}).$$

Furthermore, if (P.9) holds then we have the following pointwise rates in expectation, for any $(x^*, \mu^*) \in \mathcal{S}_{\mathcal{D}} \times \mathcal{S}_{\mathcal{D}}$,

(v) *It holds $\inf_{0 \leq i \leq k} \mathbb{E}[\mathcal{L}(x_k, \mu^*)] - \mathcal{L}(x^*, \mu^*) \in O\left(\frac{1}{\Gamma_k}\right)$.*

(vi) *There exists a subsequence $(x_{k_j})_{j \in \mathbb{N}}$ such that $\mathbb{E}[\mathcal{L}(x_{k_j}, \mu^*)] - \mathcal{L}(x^*, \mu^*) \leq \frac{1}{\Gamma_{k_j}}$.*

Proof. As in the proof of Theorem 4.7, our goal is to first apply Lemma 2.2 and then apply Lemma 2.3. By (4.15) in Lemma 4.10 we have, using the same notation, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathbb{E}[r_{k+1} \mid \mathcal{D}_{k-1}] - r_k &\leq -\gamma_k \left(\mathcal{L}(x_k, \mu^*) - \mathcal{L}(x^*, \mu^*) + \frac{\rho_k}{2} \|Ax_k - b\|^2 \right) \\ &\quad + (\beta_k - \beta_{k+1}) \frac{M^2}{2} + (\gamma_k - \gamma_{k+1}) \tilde{M} + \gamma_k \beta_k \frac{M^2}{2} + K_{(F, \zeta, C)} \zeta(\gamma_k) + \gamma_k^2 C_k \\ &\quad + d_C \gamma_k \mathbb{E}[\|\lambda_k\| \mid \mathcal{D}_{k-1}] + \gamma_k \mathbb{E}[\lambda_k^s \mid \mathcal{D}_{k-1}]. \end{aligned}$$

Let, for each $k \in \mathbb{N}$, $a_k = \gamma_k \left(\mathcal{L}(x_k, \mu^*) - \mathcal{L}(x^*, \mu^*) + \frac{\rho_k}{2} \|Ax_k - b\|^2 \right)$ and denote what remains on the r.h.s. by z_k . Then, to apply Lemma 2.2, we must show $(z_k)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S}')$. The terms $(\beta_k - \beta_{k+1}) \frac{M^2}{2}$ and $(\gamma_k - \gamma_{k+1}) \tilde{M}$ are bounded and telescopic, hence in ℓ_+^1 . The terms $\gamma_k \beta_k \frac{M^2}{2}$ and $K_{(F, \zeta, C)} \zeta(\gamma_k)$ are in ℓ_+^1 by (P.1). Recalling the definition of C_k , we have, for each $k \in \mathbb{N}$,

$$\begin{aligned} \gamma_k^2 C_k &= \gamma_k^2 \left(\frac{L_k}{2} d_C^2 + d_C (M\|T\| + D + L_h + \|\mu^*\| \|A\|) \right) \\ &= \left(\frac{d_C^2 \|T\|^2}{2} \right) \frac{\gamma_k^2}{\beta_k} + \left(\frac{d_C^2 \|A\|^2 \rho_k}{2} + d_C (M\|T\| + D + L_h + \|\mu^*\| \|A\|) \right) \gamma_k^2 \\ &\leq \left(\frac{d_C^2 \|T\|^2}{2} \right) \frac{\gamma_k^2}{\beta_k} + \left(\frac{d_C^2 \|A\|^2 \bar{\rho}}{2} + d_C (M\|T\| + D + L_h + \|\mu^*\| \|A\|) \right) \gamma_k^2 \end{aligned}$$

which is in ℓ_+^1 by (P.1) and (P.3). The remaining terms,

$$d_C \gamma_k \mathbb{E} [\|\lambda_k\| \mid \mathcal{S}_{k-1}] + \gamma_k \mathbb{E} [\lambda_k^s \mid \mathcal{S}_{k-1}],$$

coming from the inexactness of the algorithm, are in $\ell_+^1(\mathfrak{S}')$ by (P.8). Thus, the r.h.s. belongs to $\ell_+^1(\mathfrak{S}')$ and so by Lemma 2.2 we have,

$$a_k = \gamma_k \left(\mathcal{L}(x_k, \mu^*) - \mathcal{L}(x^*, \mu^*) + \frac{\rho k}{2} \|Ax_k - b\|^2 \right) \in \ell_+^1(\mathfrak{S}') \quad (\mathbb{P}\text{-a.s.}),$$

and also that $(r_k)_{k \in \mathbb{N}}$ converges (\mathbb{P} -a.s.).

The first claim (i) follows by applying Lemma 2.3, the conditions of which are satisfied directly from Lemma 4.3 and Lemma 4.5.

The second claim, (ii), follows from the same arguments as in [34, Theorem 4.2(ii)] but adapted to the stochastic case. For the claims about $(x_k)_{k \in \mathbb{N}}$, the proof is trivially extended to the stochastic setting (\mathbb{P} -a.s.). However, the claims about $(\mu_k)_{k \in \mathbb{N}}$ are more delicate to adapt so we write explicitly the arguments below.

By Theorem 4.7(iii) we have $(\gamma_k \|A\tilde{x}_k - b\|^2)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$ which, by Lemma 2.4 implies that there exists a subsequence $(\tilde{x}_{k_j})_{j \in \mathbb{N}}$ with $\|A\tilde{x}_{k_j} - b\| \rightarrow 0$ (\mathbb{P} -a.s.). Since the sequence $(\mu_k)_{k \in \mathbb{N}}$ is bounded (\mathbb{P} -a.s.) by Lemma 4.8, the subsequence $(\mu_{k_j})_{j \in \mathbb{N}}$ is bounded (\mathbb{P} -a.s.) as well and thus admits a weakly (\mathbb{P} -a.s.) convergent subsequence $(\mu_{k_{j_i}})_{i \in \mathbb{N}}$ with $\mu_{k_{j_i}} \rightarrow \bar{\mu}$ for some \mathcal{H}_d -valued random variable $\bar{\mu}$. By Fermat's rule ([6, Theorem 16.2]), the weak (\mathbb{P} -a.s.) sequential cluster point $\bar{\mu}$ is a solution to (\mathcal{D}) iff

$$0 \in \partial(\Phi^* \circ (-A^*))(\bar{\mu}) + b \quad (\mathbb{P}\text{-a.s.}).$$

The proximal operator is the resolvent of the subdifferential and so it follows that (4.12) is equivalent, for each $i \in \mathbb{N}$, to

$$(4.28) \quad \nabla \varphi_{k_{j_i}}(\mu_{k_{j_i}}) - b \in \partial(\Phi_{k_{j_i}}^* \circ (-A^*))(\mu_{k_{j_i}} - \rho_{k_{j_i}} \nabla \varphi_{k_{j_i}}(\mu_{k_{j_i}})) \quad (\mathbb{P}\text{-a.s.}).$$

Since $(A\tilde{x}_{k_j})_{j \in \mathbb{N}}$ converges strongly to b (\mathbb{P} -a.s.), and in view of (4.13), it holds that $\nabla \varphi_{k_j}(\mu_{k_j})$ converges strongly to 0 (\mathbb{P} -a.s.). However, $\mu_{k_{j_i}} - \rho_{k_{j_i}} \nabla \varphi_{k_{j_i}}(\mu_{k_{j_i}})$ converges weakly to $\bar{\mu}$ (\mathbb{P} -a.s.). We henceforth argue that we can pass to the limit in (4.28) by sequential closedness using a Mosco convergence (weak-strong epigraphical convergence) argument (see [9] and [3, Definition 3.7]). Indeed, it was shown in the proof of [34, Theorem 4.2], which defer to for brevity, that $\Phi_{k_{j_i}}^* \circ (-A^*)$ Mosco-converges to $(\Phi)^* \circ (-A^*)$. This implies, via [3, Theorem 3.66], that $\partial \Phi_{k_{j_i}}^* \circ (-A^*)$ graph-converges to $\partial \Phi^* \circ (-A^*)$, and [3, Proposition 3.59] shows that $(\partial \Phi_{k_{j_i}}^* \circ (-A^*))_{i \in \mathbb{N}}$ is sequentially closed for graph-convergence in the weak-strong topology on \mathcal{H}_d , i.e., for any sequence $((v_{k_{j_i}}, \eta_{k_{j_i}}))_{i \in \mathbb{N}}$ in the graph of $(\partial \Phi_{k_{j_i}}^* \circ (-A^*))_{i \in \mathbb{N}}$ such that $v_{k_{j_i}}$ converges weakly to \bar{v} and $\eta_{k_{j_i}}$ converges strongly to $\bar{\eta}$, we have $\bar{\eta} \in \partial \Phi^* \circ (-A^*)(\bar{v})$. Let, for each $i \in \mathbb{N}$, $v_{k_{j_i}} = \nabla \varphi_{k_{j_i}}(\mu_{k_{j_i}}) - b$ and $\eta_{k_{j_i}} = \mu_{k_{j_i}} - \rho_{k_{j_i}} \nabla \varphi_{k_{j_i}}(\mu_{k_{j_i}})$, which converge strongly (\mathbb{P} -a.s.) and weakly (\mathbb{P} -a.s.) respectively, and let $\tilde{\Omega} \subset \mathcal{F}$ such that $\mathbb{P}(\tilde{\Omega}) = 1$ and, for all $\omega \in \tilde{\Omega}$, $v_{k_{j_i}}(\omega) \rightarrow b$ and $\eta_{k_{j_i}}(\omega) \rightarrow \bar{\mu}(\omega)$. We conclude that, for each $\omega \in \tilde{\Omega}$,

$$0 \in \partial(\Phi^* \circ (-A^*))(\bar{\mu}(\omega)) + b \quad (\mathbb{P}\text{-a.s.}),$$

i.e., $\bar{\mu}$ is a solution of the dual problem (\mathcal{D}) (\mathbb{P} -a.s.).

We now prove the existence of a set $\tilde{\Omega} \subset \mathcal{F}$ such that $\mathbb{P}(\tilde{\Omega}) = 1$ and, for all $\omega \in \tilde{\Omega}$, for any $\mu^* \in S_{\mathcal{D}}$,

$$\Theta(\mu^*, \omega) \stackrel{\text{def}}{=} \lim_k \|\mu_k(\omega) - \mu^*\|^2$$

exists. This does indeed hold (\mathbb{P} -a.s.) for each fixed $\mu^* \in S_{\mathcal{D}}$ by the argument in the proof of [34, Theorem 4.2] but $S_{\mathcal{D}}$ may be uncountable and so the entire statement for any $\mu^* \in S_{\mathcal{D}}$ may not necessarily hold (\mathbb{P} -a.s.). To rectify this situation, we make the assumption (A.9) and argue as in [12, Proposition 2.3(iii)].

First repeat the argument made in the proof of [34, Theorem 4.2] to show that, for each fixed $\mu^* \in S_{\mathcal{D}}$, there exists $\Omega_{\mu^*} \subset \mathcal{F}$ with $\mathbb{P}(\Omega_{\mu^*}) = 1$ such that, for any $\omega \in \Omega_{\mu^*}$, $\Theta(\mu^*, \omega)$ exists. Let $\mu^* \in S_{\mathcal{D}}$ and recall $(r_k)_{k \in \mathbb{N}}$ in (4.10), for each $k \in \mathbb{N}$,

$$r_k \stackrel{\text{def}}{=} (1 - \gamma_k) \mathcal{L}_k(x_k, \mu_k) + \frac{c}{2} \|\mu_k - \mu^*\|^2.$$

We have already shown that $(r_k)_{k \in \mathbb{N}}$ is convergent (\mathbb{P} -a.s.). We also have, for each $k \in \mathbb{N}$,

$$\begin{aligned} -\mathcal{L}_k(x_k, \mu_k) &= (\mathcal{L}(x_k, \mu^*) - \mathcal{L}_k(x_k, \mu_k)) - \mathcal{L}(x_k, \mu^*) \\ &= g(Tx_k) - g^{\beta_k}(Tx_k) + \langle \mu^* - \mu_k, Ax_k - b \rangle - \frac{\rho_k}{2} \|Ax_k - b\|^2 \\ &\quad - \mathcal{L}(x_k, \mu^*). \end{aligned}$$

We have from Theorem 4.7 that $\frac{\rho_k}{2} \|Ax_k - b\|^2 \rightarrow 0$ (\mathbb{P} -a.s.). Therefore,

$$\langle \mu^* - \mu_k, Ax_k - b \rangle \rightarrow 0$$

since $(\mu_k)_{k \in \mathbb{N}}$ is bounded (\mathbb{P} -a.s.). We also have, by claim (i) of this theorem, that $\mathcal{L}(x_k, \mu^*) \rightarrow \mathcal{L}(x^*, \mu^*)$ (\mathbb{P} -a.s.). By Lemma 2.6 and (A.4), we get

$$0 \leq (g(Tx_k) - g^{\beta_k}(Tx_k)) \leq \frac{\beta_k}{2} M^2 \quad (\mathbb{P}\text{-a.s.})$$

which implies, in light of (P.3), that $g(Tx_k) - g^{\beta_k}(Tx_k) \rightarrow 0$ (\mathbb{P} -a.s.). Altogether, it holds that $\mathcal{L}_k(x_k, \mu_k) \rightarrow \mathcal{L}(x^*, \mu^*)$ (\mathbb{P} -a.s.) and thus the limit

$$\lim_k \|\mu_k - \mu^*\|^2 = 2/c \left(\lim_k r_k - \mathcal{L}(x^*, \mu^*) \right)$$

exists (\mathbb{P} -a.s.) for each $\mu^* \in S_{\mathcal{D}}$.

Since \mathcal{H}_d is separable by (A.9), there exists a countable set S such that $\bar{S} = S_{\mathcal{D}}$. The previous paragraph has shown that, for every $\mu^* \in S_{\mathcal{D}}$, there exists $\Omega_{\mu^*} \subset \mathcal{F}$ such that $\mathbb{P}(\Omega_{\mu^*}) = 1$ and, for any $\omega \in \Omega_{\mu^*}$, $\Theta(\mu^*, \omega)$ exists. Set $\tilde{\Omega} = \bigcap_{\mu^* \in S} \Omega_{\mu^*}$ and let $\tilde{\Omega}^c$ be its set-theoretic complement. By the countability of S ,

$$\mathbb{P}(\tilde{\Omega}) = 1 - \mathbb{P}(\tilde{\Omega}^c) = 1 - \mathbb{P}\left(\bigcup_{\mu^* \in S} \Omega_{\mu^*}^c\right) \geq 1 - \sum_{\mu^* \in S} \mathbb{P}(\Omega_{\mu^*}^c) = 1,$$

i.e., $\mathbb{P}(\tilde{\Omega}) = 1$. Fix $\mu^* \in S_{\mathcal{D}}$; since $\bar{S} = S_{\mathcal{D}}$, there exists a sequence $(\mu_n^*)_{n \in \mathbb{N}}$ such that, for each $n \in \mathbb{N}$, $\mu_n^* \in S$ and $\mu_n^* \rightarrow \mu^*$. As was already shown, for each $n \in \mathbb{N}$, for any $\omega \in \Omega_{\mu_n^*}$, $\Theta(\mu_n^*, \omega)$ exists. Let $\omega \in \tilde{\Omega}$, then we have, for each $n \in \mathbb{N}$, for each $k \in \mathbb{N}$,

$$-\|\mu_n^* - \mu^*\| \leq \|\mu_k(\omega) - \mu^*\| - \|\mu_k(\omega) - \mu_n^*\| \leq \|\mu_n^* - \mu^*\|$$

and thus, for each $n \in \mathbb{N}$,

$$\begin{aligned} -\|\mu_n^* - \mu^*\| &\leq \liminf_k \|\mu_k(\omega) - \mu^*\| - \lim_k \|\mu_k(\omega) - \mu_n^*\| \\ &= \liminf_k \|\mu_k(\omega) - \mu^*\| - \Theta(\mu_n^*, \omega) \\ &\leq \limsup_k \|\mu_k(\omega) - \mu^*\| - \Theta(\mu_n^*, \omega) \\ &= \limsup_k \|\mu_k(\omega) - \mu^*\| - \lim_k \|\mu_k(\omega) - \mu_n^*\| \\ &\leq \|\mu_n^* - \mu^*\|. \end{aligned}$$

Taking the limit as $n \rightarrow \infty$ then gives that the sequence $(\Theta(\mu_n^*, \omega))_{n \in \mathbb{N}}$ converges to $\Theta(\mu^*, \omega)$ for any $\omega \in \tilde{\Omega}$ where $\tilde{\Omega}$ does not depend on μ^* .

We now aim to use (A.10), for which we denote $(p_i)_{i \in \mathbb{N}} = (\nabla \varphi_{k_{j_i}}(\mu_{k_{j_i}}) - b)_{i \in \mathbb{N}}$ and $(q_i)_{i \in \mathbb{N}} = (\mu_{k_{j_i}} - \rho_{k_{j_i}} \nabla \varphi_{k_{j_i}}(\mu_{k_{j_i}}))_{i \in \mathbb{N}}$. We've shown that $(p_i)_{i \in \mathbb{N}}$ converges strongly to 0 (\mathbb{P} -a.s.) and that $(q_i)_{i \in \mathbb{N}}$ converges weakly to $\bar{\mu}$ (\mathbb{P} -a.s.) and so there exists $\tilde{\Omega} \subset \mathcal{F}$ with $\mathbb{P}(\tilde{\Omega}) = 1$ such that, for any $\omega \in \tilde{\Omega}$, $p_i(\omega) \rightarrow p(\omega)$ and $q_i(\omega) \rightarrow q(\omega)$. Due to (4.28), we furthermore have, for each $\omega \in \tilde{\Omega}$, for each $i \in \mathbb{N}$,

$$p_i(\omega) \in \partial(\Phi_{k_{j_i}}^* \circ (-A^*)) (q_i(\omega)) \quad (\mathbb{P}\text{-a.s.}),$$

and thus by (A.10), for each $\omega \in \tilde{\Omega}$, $(q_i(\omega))_{i \in \mathbb{N}}$ admits a subsequence $(q_{i_l}(\omega))_{l \in \mathbb{N}}$ such that $q_{i_l}(\omega) \rightarrow \bar{q}(\omega)$, i.e., the sequence $(\mu_{k_{j_{i_l}}} - \rho_{k_{j_{i_l}}} \nabla \varphi_{k_{j_{i_l}}}(\mu_{k_{j_{i_l}}}))_{l \in \mathbb{N}}$ is strongly convergent (\mathbb{P} -a.s.). Thus, the subsequence $(\mu_{k_{j_{i_l}}})_{l \in \mathbb{N}}$ is strongly convergent to $\bar{\mu}$ (\mathbb{P} -a.s.). Since $\bar{\mu}$ is a solution to (\mathcal{D}) , it holds that $\lim_k \|\mu_k - \bar{\mu}\|$ exists (\mathbb{P} -a.s.). At the same time, we have shown that $\lim_l \|\mu_{k_{j_{i_l}}} - \bar{\mu}\| = 0$ (\mathbb{P} -a.s.) and so the whole sequence $(\mu_k)_{k \in \mathbb{N}}$ converges strongly to $\bar{\mu} \in S_{\mathcal{D}}$ (\mathbb{P} -a.s.).

Meanwhile the third claim, (iii), follows from the argument of [34, Theorem 4.2(iv)] directly applied to the (\mathbb{P} -a.s.) setting and similarly for (iv) following from the argument of [34, Corollary 4.3].

Finally, assume that (P.9) holds. By taking the total expectation of (4.15) in Lemma 4.10 and using the law of total expectation we have, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathbb{E}[r_{k+1}] - \mathbb{E}[r_k] &\leq -\gamma_k \left(\mathbb{E}[\mathcal{L}(x_k, \mu^*)] - \mathcal{L}(x^*, \mu^*) + \frac{\rho_k}{2} \mathbb{E}[\|Ax_k - b\|^2] \right) + (\beta_k - \beta_{k+1}) \frac{M^2}{2} \\ &\quad + (\gamma_k - \gamma_{k+1}) \tilde{M} + \gamma_k \beta_k \frac{M^2}{2} + K_{(F, \zeta, C)} \zeta(\gamma_k) + \gamma_k^2 C_k + d_C \gamma_k \mathbb{E}[\|\lambda_k\|] \\ &\quad + \gamma_k \mathbb{E}[\lambda_k^s] \quad (\mathbb{P}\text{-a.s.}). \end{aligned}$$

Define the following, for each $k \in \mathbb{N}$,

$$\tilde{r}_k = \mathbb{E}[r_k] \quad \text{and} \quad \tilde{p}_k = \gamma_k \quad \text{and} \quad \tilde{w}_k = \mathbb{E}[\mathcal{L}(x_k, \mu^*)] - \mathcal{L}(x^*, \mu^*) + \frac{\rho_k}{2} \mathbb{E}[\|Ax_k - b\|^2]$$

and denote what remains, for each $k \in \mathbb{N}$,

$$\tilde{z}_k = (\beta_k - \beta_{k+1}) \frac{M^2}{2} + (\gamma_k - \gamma_{k+1}) \tilde{M} + \gamma_k \beta_k \frac{M^2}{2} + K_{(F, \zeta, C)} \zeta(\gamma_k) + \gamma_k^2 C_k + d_C \gamma_k \mathbb{E}[\|\lambda_k\|] + \gamma_k \mathbb{E}[\lambda_k^s]$$

By repeating the arguments of the previous paragraph, we have that $(\tilde{z}_k)_{k \in \mathbb{N}} \in \ell_+^1$. Invoking Lemma 2.4, again noting Lemma 4.3 and Lemma 4.5 hold with the total expectation as well, with $(\tilde{r}_k)_{k \in \mathbb{N}}$, $(\tilde{p}_k)_{k \in \mathbb{N}}$, $(\tilde{w}_k)_{k \in \mathbb{N}}$, and $(\tilde{z}_k)_{k \in \mathbb{N}}$ defined as above, we obtain the remaining claims. \square

Remark 4.12. The assumption (A.9) is only necessary for showing that the sequence of dual variables $(\mu_k)_{k \in \mathbb{N}}$ admits an optimal weak cluster point. The other results, e.g., convergence of the Lagrangian values, the containment $\mathfrak{B}[(x_k)_{k \in \mathbb{N}}] \subset \mathcal{S}_{\mathcal{D}}$ (\mathbb{P} -a.s.), etc, do not require the separability imposed by (A.9). Likewise, something similar can be said for (A.10), which is only necessary for ensuring the strong convergence of the sequence of dual variables $(\mu_k)_{k \in \mathbb{N}}$ and can otherwise be omitted.

Remark 4.13. The combination of the Lagrangian convergence rate (4.27) with convexity and the feasibility rate (4.7) leads directly to a rate for the objective itself evaluated at the ergodic iterate,

$$|\Phi(\bar{x}_k) - \Phi(x^*)| = O\left(1/\sqrt{\Gamma_k}\right) \text{ (}\mathbb{P}\text{-a.s.) .}$$

5 STOCHASTIC EXAMPLES

We examine the problem of risk minimization using two different ways to inexactly calculate the gradient with stochastic noise to demonstrate that the assumptions on the error can be satisfied in order to apply ICGALP.

Consider the following,

$$(\mathcal{P}_1) \quad \min_{\substack{x \in \mathcal{C} \subset \mathcal{H} \\ Ax=b}} \left\{ f(x) \stackrel{\text{def}}{=} \mathbb{E}[L(x, \eta)] \right\},$$

where $L(\cdot, \eta)$ is differentiable for every η , and η is a random variable.

We will impose the following assumptions, or a subset of them depending on the context. Indeed, only (E.1) and (E.2) will be used for risk minimization with increasing batch size while (E.3) and (E.4) will be needed for the results on risk minimization with variance reduction.

(E.1) It holds, for all $x \in \mathcal{H}_p$, $\nabla f(x) = \mathbb{E}[\nabla_x L(x, \eta)]$.

(E.2) For all η , the function $L(\cdot, \eta)$ is ω -smooth (see Definition 2.9) with ω nondecreasing.

(E.3) The function f is ω -smooth with ω nondecreasing.

(E.4) The function f is Hölder-smooth with constant C_f and exponent τ .

Remark 5.1. In practical contexts, it's unrealistic that one will have access to the function f or knowledge of its regularity. To this end, we note that the assumptions (E.1) and (E.2), which depend only on the function $L(x, \eta)$, are sufficient to ensure that (E.3) holds and similarly for (E.4) if one adjusts (E.2) for Hölder-smoothness. Moreover, since Hölder-smoothness is a special case of ω -smoothness, (E.4) \implies (E.3).

Remark 5.2. With the above choice for λ_k , the terms in $\nabla_x \mathcal{E}_k(x_k, \mu_k)$ coming from the augmented Lagrangian are computed exactly, however our analysis extends to the case where $\nabla_x \left(\frac{\rho_k}{2} \|Ax_k - b\|^2\right) = \rho_k A^*(Ax_k - b)$ is computed inexactly as well, as this function is always Lipschitz-continuous. We demonstrate this alternative choice in Section 7 by sampling the components $\rho_k A^*(Ax_k - b)^{(i)}$ in the numerical experiments.

For the sake of clarity, we demonstrate only the case where, for each $k \in \mathbb{N}$, $\lambda_k \equiv \lambda_k^f$ with $\lambda_k^f = \widehat{\nabla} f_k - \nabla f(x_k)$ and $\widehat{\nabla} f_k$ is our inexact computation of $\nabla f(x_k)$, to be defined in the following subsections.

5.1 RISK MINIMIZATION WITH INCREASING BATCH SIZE

Consider (\mathcal{P}_1) and define, for each $k \in \mathbb{N}$,

$$\widehat{\nabla} f_k \stackrel{\text{def}}{=} \frac{1}{n(k)} \sum_{i=1}^{n(k)} \nabla_x L(x_k, \eta_i)$$

where $n(k)$ is the number of samples to be taken at iteration k . We assume that each η_i is i.i.d., according to some fixed distribution, and that n is a function of k , i.e., the number of samples taken to estimate the expectation is dependent on the iteration number itself.

Lemma 5.3. *Under assumptions (E.1) and (E.2), denote*

$$C = 2 \left(\omega(d_C)^2 + \mathbb{E} \left[\|\nabla L(x^*, \eta)\|^2 \right] \right)$$

where x^* is a solution to (\mathcal{P}_1) and, for each $k \in \mathbb{N}$, $\mathcal{S}_k = \sigma(x_0, \mu_0, \widehat{s}_0, \dots, \widehat{s}_k)$ as in (3.2). Then, for each $k \in \mathbb{N}$, the following holds,

$$\mathbb{E} \left[\left\| \lambda_{k+1}^f \right\| \mid \mathcal{S}_k \right] \leq \sqrt{\frac{C}{n(k+1)}}.$$

Proof. By Jensen's inequality, for each $k \in \mathbb{N}$,

$$\mathbb{E} \left[\left\| \lambda_{k+1}^f \right\| \mid \mathcal{S}_k \right]^2 \leq \mathbb{E} \left[\left\| \lambda_{k+1}^f \right\|^2 \mid \mathcal{S}_k \right] = \mathbb{E} \left[\left\| \nabla f(x_{k+1}) - \widehat{\nabla} f_{k+1} \right\|^2 \mid \mathcal{S}_k \right].$$

Then, since $\widehat{\nabla} f_{k+1}$ is an unbiased estimator for $\nabla f(x_{k+1})$, we have, for each $k \in \mathbb{N}$,

$$\begin{aligned} \mathbb{E} \left[\left\| \nabla f(x_{k+1}) - \widehat{\nabla} f_{k+1} \right\|^2 \mid \mathcal{S}_k \right] &= \mathbb{E} \left[\left\| \mathbb{E} \left[\widehat{\nabla} f_{k+1} \right] - \widehat{\nabla} f_{k+1} \right\|^2 \mid \mathcal{S}_k \right] \\ &= \text{Var} \left[\widehat{\nabla} f_{k+1} \mid \mathcal{S}_k \right] \\ &= \text{Var} \left[\frac{1}{n(k+1)} \sum_{i=1}^{n(k+1)} \nabla L(x_{k+1}, \eta_i) \mid \mathcal{S}_k \right] \\ &= \frac{1}{n(k+1)} \text{Var} \left[\nabla L(x_{k+1}, \eta) \mid \mathcal{S}_k \right], \end{aligned}$$

where the last equality follows from the independence and identical distribution of η_i . Applying the definition of conditional variance yields, for each $k \in \mathbb{N}$,

$$\begin{aligned} \frac{1}{n(k+1)} \text{Var} \left[\nabla L(x_{k+1}, \eta) \mid \mathcal{S}_k \right] &= \frac{1}{n(k+1)} \left(\mathbb{E} \left[\|\nabla L(x_{k+1}, \eta)\|^2 \mid \mathcal{S}_k \right] - \|\mathbb{E} \left[\nabla L(x_{k+1}, \eta) \mid \mathcal{S}_k \right]\|^2 \right) \\ &\leq \frac{1}{n(k+1)} \mathbb{E} \left[\|\nabla L(x_{k+1}, \eta)\|^2 \mid \mathcal{S}_k \right]. \end{aligned}$$

We again use Jensen's inequality, then ω -smoothness, and finally the fact that ω is nondecreasing together with the fact that x_{k+1} and x^* are both in C to find, for each $k \in \mathbb{N}$,

$$\begin{aligned} \frac{1}{n(k+1)} \mathbb{E} \left[\|\nabla L(x_{k+1}, \eta)\|^2 \mid \mathcal{S}_k \right] &\leq \frac{2}{n(k+1)} \left(\mathbb{E} \left[\|\nabla L(x_{k+1}, \eta) - \nabla L(x^*, \eta)\|^2 \mid \mathcal{S}_k \right] \right. \\ &\quad \left. + \mathbb{E} \left[\|\nabla L(x^*, \eta)\|^2 \right] \right) \\ &\leq \frac{2}{n(k+1)} \left(\mathbb{E} \left[\omega \left(\|x_{k+1} - x^*\|^2 \right) \mid \mathcal{S}_k \right] + \mathbb{E} \left[\|\nabla L(x^*, \eta)\|^2 \right] \right) \\ &\leq \frac{2}{n(k+1)} \left(\omega(d_C)^2 + \mathbb{E} \left[\|\nabla L(x^*, \eta)\|^2 \right] \right) \\ &= \frac{C}{n(k+1)}. \end{aligned}$$

The above shows that, for each $k \in \mathbb{N}$, $\mathbb{E} \left[\left\| \lambda_{k+1}^f \right\|^2 \mid \mathcal{S}_k \right] \leq \frac{C}{n(k+1)}$ and so $\mathbb{E} \left[\left\| \lambda_{k+1}^f \right\| \mid \mathcal{S}_k \right] \leq \sqrt{\frac{C}{n(k+1)}}$ as desired. \square

Proposition 5.4. *Under (E.1) and (E.2), assume that the number of samples $n(k)$ at iteration k is lower bounded by $\left(\frac{\gamma_k}{\zeta(\gamma_k)}\right)^2$, i.e. for some $\alpha > 0$, $n(k) \geq \alpha \left(\frac{\gamma_k}{\zeta(\gamma_k)}\right)^2$. Then, the summability of the error in (P.8) is satisfied; namely,*

$$\gamma_{k+1} \mathbb{E} \left[\left\| \lambda_{k+1}^f \right\|^2 \mid \mathcal{S}_k \right] \in \ell^1(\mathfrak{S}).$$

Proof. By Lemma 5.3 we have, for each $k \in \mathbb{N}$,

$$\gamma_{k+1} \mathbb{E} \left[\left\| \lambda_{k+1}^f \right\|^2 \mid \mathcal{S}_k \right] \leq \gamma_{k+1} \sqrt{\frac{C}{n(k+1)}} \leq \sqrt{\frac{C}{\alpha}} \zeta(\gamma_{k+1}).$$

The summability of $\zeta(\gamma_{k+1})$ is given by (P.1) and thus $\gamma_{k+1} \mathbb{E} \left[\left\| \lambda_{k+1}^f \right\|^2 \mid \mathcal{S}_k \right] \in \ell^1(\mathfrak{S})$ □

Remark 5.5. The lower bound $n(k) \geq \alpha \left(\frac{\gamma_k}{\zeta(\gamma_k)}\right)^2$ is sufficient but not necessary; one can alternatively choose $n(k)$ to be lower bounded by $\alpha \left(\frac{\beta_k}{\gamma_k}\right)^2$ or $\alpha \left(\frac{1}{\beta_k}\right)^2$ and, due to (P.1), the result will still hold.

5.2 RISK MINIMIZATION WITH VARIANCE REDUCTION

We reconsider (\mathcal{P}_1) as before but now with a different $\widehat{\nabla}f$. We define a stochastic-averaged gradient, which will serve as a form of variance reduction, such that the number of samples at each iteration need not increase as in the previous subsection. For each $k \in \mathbb{N}$, let $\nu_k \in [0, 1]$ and define

$$(5.1) \quad \widehat{\nabla}f_k \stackrel{\text{def}}{=} (1 - \nu_k) \widehat{\nabla}f_{k-1} + \nu_k \nabla_x L(x_k, \eta_k)$$

with $\widehat{\nabla}f_{-1} = 0$ and with each η_i i.i.d.. We call $\widehat{\nabla}f_k$ the stochastic average of sampled gradients with weight ν_k . This method of approximating $\nabla f(x_k)$ is reminiscent of both the SARAH estimator [28] and the SVRG estimator [23], although distinct from both since the SARAH estimator is biased and the SVRG estimator requires an inner and outer loops.

In the previous section, we have used the number of batches $n(k)$ to ensure the error summability condition. This in turn means that the number of gradient evaluations increases with k (in particular, for finite-sum objectives, one has to evaluate all gradients after finitely many iterations). This is in stark contrast with variance reduction proposed in this section where we are able to take a single (or a larger but fixed batch size) gradient sample at each iteration, while taking full advantage of the flexibility offered by the choice of ν_k to reduce the stochastic error variance as we now show.

Lemma 5.6. *Under (E.1) and (E.3), denote, for each $k \in \mathbb{N}$,*

$$(5.2) \quad \sigma_k^2 \stackrel{\text{def}}{=} \mathbb{E} \left[\left\| \nabla_x L(x_k, \eta_k) - \nabla f(x_k) \right\|^2 \mid \mathcal{S}_{k-1} \right]$$

where \mathcal{S}_k is defined in (3.2) and assume that $\sigma^2 = \sup_k \sigma_k^2 < \infty$. Then, for each $k \in \mathbb{N}$, the following inequality holds,

$$\mathbb{E} \left[\left\| \lambda_{k+1}^f \right\|^2 \mid \mathcal{S}_k \right] \leq \left(1 - \frac{\nu_{k+1}}{2}\right) \left\| \lambda_k^f \right\|^2 + \nu_{k+1}^2 \sigma^2 + 2 \frac{\omega(d_C \gamma_k)^2}{\nu_{k+1}}.$$

Remark 5.7. Since (E.2) implies (E.3), see Remark 5.1, it is clear that Lemma 5.6 holds under (E.1)-(E.2). Moreover, by arguing similarly to the end of the proof of Lemma 5.3, it can be easily shown that the uniform boundedness assumption on σ_k^2 is in force under (E.2).

Proof. The proof of this theorem is inspired by a similar construction found in [29, Lemma 2]. By definition of λ_{k+1}^f and $\widehat{\nabla}f_{k+1}$, we have, for all $k \in \mathbb{N}$,

$$\left\| \lambda_{k+1}^f \right\|^2 = \left\| \widehat{\nabla}f_{k+1} - \nabla f(x_{k+1}) \right\|^2 = \left\| (1 - \nu_{k+1}) \widehat{\nabla}f_k + \nu_{k+1} \nabla_x L(x_{k+1}, \eta_{k+1}) - \nabla f(x_{k+1}) \right\|^2.$$

We add and subtract $(1 - \nu_{k+1}) \nabla f(x_k)$ to get,

$$\left\| \lambda_{k+1}^f \right\|^2 = \left\| (1 - \nu_{k+1}) \lambda_k^f + \nu_{k+1} (\nabla_x L(x_{k+1}, \eta_{k+1}) - \nabla f(x_{k+1})) + (1 - \nu_{k+1}) (\nabla f(x_k) - \nabla f(x_{k+1})) \right\|^2.$$

Applying the pythagorean identity then gives,

$$\begin{aligned} \left\| \lambda_{k+1}^f \right\|^2 &= (1 - \nu_{k+1})^2 \left\| \lambda_k^f \right\|^2 + \nu_{k+1}^2 \left\| \nabla_x L(x_{k+1}, \eta_{k+1}) - \nabla f(x_{k+1}) \right\|^2 \\ &\quad + (1 - \nu_{k+1})^2 \left\| \nabla f(x_k) - \nabla f(x_{k+1}) \right\|^2 \\ &\quad + 2 \left\langle (1 - \nu_{k+1}) \left(\lambda_k^f + \nabla f(x_k) - \nabla f(x_{k+1}) \right), \nu_{k+1} (\nabla_x L(x_{k+1}, \eta_{k+1}) - \nabla f(x_{k+1})) \right\rangle \\ &\quad + 2 \left\langle (1 - \nu_{k+1}) \lambda_k^f, (1 - \nu_{k+1}) (\nabla f(x_k) - \nabla f(x_{k+1})) \right\rangle. \end{aligned}$$

Using Young’s inequality on the last inner product, we find,

$$\begin{aligned} \left\| \lambda_{k+1}^f \right\|^2 &\leq (1 - \nu_{k+1})^2 \left\| \lambda_k^f \right\|^2 + \nu_{k+1}^2 \left\| \nabla_x L(x_{k+1}, \eta_{k+1}) - \nabla f(x_{k+1}) \right\|^2 \\ &\quad + (1 - \nu_{k+1})^2 \left\| \nabla f(x_k) - \nabla f(x_{k+1}) \right\|^2 \\ &\quad + 2 \left\langle (1 - \nu_{k+1}) \left(\lambda_k^f + \nabla f(x_k) - \nabla f(x_{k+1}) \right), \nu_{k+1} (\nabla_x L(x_{k+1}, \eta_{k+1}) - \nabla f(x_{k+1})) \right\rangle \\ &\quad + \frac{\nu_{k+1}}{2} \left\| \lambda_k^f \right\|^2 + \frac{2}{\nu_{k+1}} \left\| (1 - \nu_{k+1}) (\nabla f(x_k) - \nabla f(x_{k+1})) \right\|^2. \end{aligned}$$

Notice that $1 - \nu_{k+1} \leq 1$ and thus $(1 - \nu_{k+1})^4 \leq 1 - \nu_{k+1}$ for all $k \in \mathbb{N}$. This leads to

$$\begin{aligned} \left\| \lambda_{k+1}^f \right\|^2 &\leq \left(1 - \frac{\nu_{k+1}}{2} \right) \left\| \lambda_k^f \right\|^2 + \nu_{k+1}^2 \left\| \nabla_x L(x_{k+1}, \eta_{k+1}) - \nabla f(x_{k+1}) \right\|^2 + \left\| \nabla f(x_k) - \nabla f(x_{k+1}) \right\|^2 \\ &\quad + 2 \left\langle (1 - \nu_{k+1}) \left(\lambda_k^f + \nabla f(x_k) - \nabla f(x_{k+1}) \right), \nu_{k+1} (\nabla_x L(x_{k+1}, \eta_{k+1}) - \nabla f(x_{k+1})) \right\rangle \\ &\quad + \frac{2(1 - \nu_{k+1})}{\nu_{k+1}} \left\| (\nabla f(x_k) - \nabla f(x_{k+1})) \right\|^2 \\ &\leq \left(1 - \frac{\nu_{k+1}}{2} \right) \left\| \lambda_k^f \right\|^2 + \nu_{k+1}^2 \left\| \nabla_x L(x_{k+1}, \eta_{k+1}) - \nabla f(x_{k+1}) \right\|^2 + \left(\frac{2}{\nu_{k+1}} \right) \left\| \nabla f(x_k) - \nabla f(x_{k+1}) \right\|^2 \\ &\quad + 2 \left\langle (1 - \nu_{k+1}) \left(\lambda_k^f + \nabla f(x_k) - \nabla f(x_{k+1}) \right), \nu_{k+1} (\nabla_x L(x_{k+1}, \eta_{k+1}) - \nabla f(x_{k+1})) \right\rangle. \end{aligned}$$

Recall that, by (E.3), f is ω -smooth with ω is nondecreasing. Furthermore, using the fact that $x_{k+1} = x_k - \gamma_k(x_k - \widehat{s}_k)$, we find

$$\begin{aligned} \left\| \lambda_{k+1}^f \right\|^2 &\leq \left(1 - \frac{\nu_{k+1}}{2} \right) \left\| \lambda_k^f \right\|^2 + \nu_{k+1}^2 \left\| \nabla_x L(x_{k+1}, \eta_{k+1}) - \nabla f(x_{k+1}) \right\|^2 + \left(\frac{2}{\nu_{k+1}} \right) \omega(\|x_k - x_{k+1}\|)^2 \\ &\quad + 2 \left\langle (1 - \nu_{k+1}) \left(\lambda_k^f + \nabla f(x_k) - \nabla f(x_{k+1}) \right), \nu_{k+1} (\nabla_x L(x_{k+1}, \eta_{k+1}) - \nabla f(x_{k+1})) \right\rangle \\ &\leq \left(1 - \frac{\nu_{k+1}}{2} \right) \left\| \lambda_k^f \right\|^2 + \nu_{k+1}^2 \left\| \nabla_x L(x_{k+1}, \eta_{k+1}) - \nabla f(x_{k+1}) \right\|^2 + \left(\frac{2}{\nu_{k+1}} \right) \omega(d_C \gamma_k)^2 \\ &\quad + 2 \left\langle (1 - \nu_{k+1}) \left(\lambda_k^f + \nabla f(x_k) - \nabla f(x_{k+1}) \right), \nu_{k+1} (\nabla_x L(x_{k+1}, \eta_{k+1}) - \nabla f(x_{k+1})) \right\rangle \end{aligned}$$

We take the expectation on both sides, recalling the definition of σ_k (see (5.2)), σ , and that

$$\mathbb{E} [\nabla_x L(x_k, \eta_k) \mid \mathcal{S}_{k-1}] = \nabla f(x_k),$$

to find

$$\mathbb{E} \left[\left\| \lambda_{k+1}^f \right\|^2 \mid \mathcal{S}_k \right] \leq \left(1 - \frac{\nu_{k+1}}{2} \right) \left\| \lambda_k^f \right\|^2 + \nu_{k+1}^2 \sigma^2 + \left(\frac{2}{\nu_{k+1}} \right) \omega(d_C \gamma_k)^2. \quad \square$$

In the following proposition, we analyze a particular case of parameter choices under the assumption (E.4) of Hölder smoothness of f , i.e. $\exists C_f, \tau > 0$ such that $\omega : t \rightarrow C_f t^\tau$.

Proposition 5.8. *Under (E.1) and (E.4), for each $k \in \mathbb{N}$, let $\widehat{\nabla} f_k$ be defined as in (5.1) with weight $\nu_k = \gamma_k^\alpha$ for some $\alpha \in]0, \tau[$. If the following conditions on the sequence $(\gamma_k)_{k \in \mathbb{N}}$ hold,*

$$(5.3) \quad \left(\gamma_k^{1+\min\{\frac{\alpha}{2}, \tau-\alpha\}} \right)_{k \in \mathbb{N}} \in \ell^1,$$

and, for k sufficiently large,

$$(5.4) \quad \frac{\gamma_k}{\gamma_{k+1}} \leq 1 + o(\gamma_k^\alpha),$$

then the summability condition in (P.8) is satisfied; namely,

$$\gamma_{k+1} \mathbb{E} \left[\left\| \lambda_{k+1}^f \right\|^2 \mid \mathcal{S}_k \right] \in \ell^1(\mathfrak{S}).$$

Proof. Since (E.4) \implies (E.3), the assumptions (E.1) and (E.3) are satisfied and Lemma 5.6 gives, for all $k \in \mathbb{N}$,

$$\mathbb{E} \left[\left\| \lambda_{k+1}^f \right\|^2 \mid \mathcal{S}_k \right] \leq \left(1 - \frac{\gamma_{k+1}^\alpha}{2} \right) \left\| \lambda_k^f \right\|^2 + \sigma^2 \gamma_{k+1}^{2\alpha} + \frac{2C_f^2 d_C^{2\tau} \gamma_k^{2\tau}}{\gamma_{k+1}^\alpha}.$$

By (P.5) we have, for all $k \in \mathbb{N}$, $\gamma_k \leq \overline{M} \gamma_{k+1}$. It follows that, for each $k \in \mathbb{N}$,

$$\mathbb{E} \left[\left\| \lambda_{k+1}^f \right\|^2 \mid \mathcal{S}_k \right] \leq \left(1 - \frac{\gamma_{k+1}^\alpha}{2} \right) \left\| \lambda_k^f \right\|^2 + \sigma^2 \gamma_{k+1}^{2\alpha} + 2\overline{M}^{2\tau} C_f^2 d_C^{2\tau} \gamma_{k+1}^{2\tau-\alpha}.$$

Consolidating higher order terms gives, for each $k \in \mathbb{N}$,

$$\mathbb{E} \left[\left\| \lambda_{k+1}^f \right\|^2 \mid \mathcal{S}_k \right] \leq \left(1 - \frac{\gamma_{k+1}^\alpha}{2} \right) \left\| \lambda_k^f \right\|^2 + \left(\sigma^2 + 2\overline{M}^{2\tau} C_f^2 d_C^{2\tau} \right) \gamma_{k+1}^{\min\{2\alpha, 2\tau-\alpha\}}.$$

Since $\alpha < \tau \leq 1$ by 5.3, it holds that $\alpha < \min\{1, 2\tau - \alpha\}$, and the first condition of Lemma A.1 is satisfied. Additionally, by (5.4), we have that the second condition, (A.2), of Lemma A.1 is satisfied as well and we can apply Lemma A.1 with

$$u_k = \left\| \lambda_k^f \right\|^2, \quad c = \frac{1}{2}, \quad s = \alpha, \quad d = \left(\sigma^2 + 2\overline{M}^{2\tau} C_f^2 d_C^{2\tau} \right), \quad \text{and } t = \min\{2\alpha, 2\tau - \alpha\},$$

to find, for k sufficiently large,

$$\mathbb{E} \left[\left\| \lambda_{k+1}^f \right\|^2 \mid \mathcal{S}_k \right] \leq 2\tilde{C} \gamma_{k+1}^{\min\{\alpha, 2(\tau-\alpha)\}} + o\left(\gamma_{k+1}^{\min\{\alpha, 2(\tau-\alpha)\}}\right)$$

and, by extension, for k sufficiently large,

$$\mathbb{E} \left[\left\| \lambda_{k+1}^f \right\| \mid \mathcal{S}_k \right] \leq \sqrt{2\tilde{C}} \gamma_{k+1}^{\min\{\frac{\alpha}{2}, \tau - \alpha\}} + o \left(\gamma_{k+1}^{\min\{\frac{\alpha}{2}, \tau - \alpha\}} \right).$$

Then, for k sufficiently large,

$$\begin{aligned} \gamma_{k+1} \mathbb{E} \left[\left\| \lambda_{k+1}^f \right\| \mid \mathcal{S}_k \right] &\leq \gamma_{k+1} \left(\sqrt{2\tilde{C}} \gamma_{k+1}^{\min\{\frac{\alpha}{2}, \tau - \alpha\}} + o \left(\gamma_{k+1}^{\min\{\frac{\alpha}{2}, \tau - \alpha\}} \right) \right) \\ &\leq \sqrt{2\tilde{C}} \gamma_{k+1}^{1+\min\{\frac{\alpha}{2}, \tau - \alpha\}} + o \left(\gamma_{k+1}^{1+\min\{\frac{\alpha}{2}, \tau - \alpha\}} \right). \end{aligned}$$

Under the assumptions 5.3 we have $\gamma_k^{1+\min\{\frac{\alpha}{2}, \tau - \alpha\}} \in \ell^1$ and thus the summability condition of (P.8) is satisfied. □

Example 5.9. The condition (5.3) in Proposition 5.8 can be satisfied, for example, by taking $\gamma_k = \frac{1}{(k+1)^{1-b}}$, $b > 0$. In this case, the condition (5.3) reduces to picking b such that the following holds,

$$(1 - b) \left(1 + \min \left\{ \frac{\alpha}{2}, \tau - \alpha \right\} \right) > 1.$$

Rearranging, we find that this is equivalent to,

$$(5.5) \quad b < 1 - \left(1 + \min \left\{ \frac{\alpha}{2}, \tau - \alpha \right\} \right)^{-1}.$$

The condition (5.4) in Proposition 5.8 can be satisfied under this choice of γ_k as well. We have,

$$\frac{\gamma_k}{\gamma_{k+1}} = \left(\frac{k+2}{k+1} \right)^{1-b} = \left(1 + \frac{1}{k+1} \right)^{1-b} \approx 1 + \frac{1-b}{k+1} = 1 + o \left(\gamma_k^\epsilon \right)$$

for any $0 < \epsilon < 1$, for k sufficiently large.

Using [34, Example 19], the predicted convergence rates for the ergodic iterates \bar{x}_k given by Theorem 4.7 and Theorem 4.11 under the above choice of γ_k read,

$$\|A\bar{x}_k - b\| = O \left(\frac{1}{(k+2)^b} \right) \quad (\mathbb{P}\text{-a.s.}) \quad \text{and} \quad \mathcal{L}(\bar{x}_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) = O \left(\frac{1}{(k+2)^b} \right) \quad (\mathbb{P}\text{-a.s.}) .$$

Thus, choosing b to be as large as possible is desired. For a given value of τ corresponding to the Hölder exponent of the gradient, the best choice for α is $\frac{2}{3}\tau$. In turn, the largest possible choice of b is $\tau/(3 + \tau)$. If the function f is Lipschitz-smooth, then $\tau = 1$, and we get $\alpha = \frac{2}{3}$ and $b = 1/4$ (to be compared with the strict upper-bound $1/3$ in the exact case, see [34, Example 15]).

Notice that the choice of α does not directly affect the predicted rates of convergence, which now depends only on the constant b . However, the choice of α dictates the largest possible choice for b satisfying the assumptions and thus, indirectly, the rates of convergence as well.

6 SWEEPING

We now consider an example in which the errors in the computation of ∇f are deterministic; a finite sum minimization problem,

$$(P_2) \quad \min_{\substack{x \in C \subset \mathcal{H} \\ Ax=b}} \frac{1}{n} \sum_{i=1}^n f_i(x)$$

where $n > 1$ is fixed. We assume that:

(F.1) f_i is ω -smooth (see Definition 2.9) for $1 \leq i \leq n$ with ω nondecreasing.

(F.2) $(\gamma_k)_{k \in \mathbb{N}}$ a nonincreasing sequence.

As in the previous section, Section 5, we examine only the case where, for each $k \in \mathbb{N}$, $\lambda_k \equiv \lambda_k^f = \nabla f(x_k) - \widehat{\nabla} f_k$, with $\widehat{\nabla} f_k$ to be defined below, although our analysis is straightforward to adapt to the more general case where one computes $\rho_k A^*(Ax_k - b)$ inexactly as well, at the expense of brevity (see Remark 5.2). We will sweep, or cycle, through the functions f_i , taking the gradient of a single one at each iteration and recursively averaging with the past gradients. For notation, fixed n , we take $\text{mod}(k) \stackrel{\text{def}}{=} (k \bmod n)$ with the convention that $\text{mod}(n) \stackrel{\text{def}}{=} n$. We define the inexact gradient in the following way,

$$\widehat{\nabla} f_k \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^k \nabla f_i(x_i) \quad (\forall k \leq n)$$

and

$$\widehat{\nabla} f_k \stackrel{\text{def}}{=} \widehat{\nabla} f_{k-1} + \frac{1}{n} (\nabla f_{\text{mod}(k)}(x_k) - \nabla f_{\text{mod}(k)}(x_{k-n})) \quad (\forall k \geq n + 1).$$

For $k \geq n + 1$ it can also be written in closed form as,

$$\widehat{\nabla} f_k = \frac{1}{n} \left(\sum_{i=1}^{\text{mod}(k)} \nabla f_i(x_{i+k-\text{mod}(k)}) + \sum_{i=\text{mod}(k)+1}^n \nabla f_i(x_{i+k-n-\text{mod}(k)}) \right).$$

Lemma 6.1. *Let $B = \frac{1}{n} (n(n-1) + (n-1)(2n-1))$. Under (F.1) and (F.2), we then have, for all $k \geq 2n - 1$, the following,*

$$\|\lambda_{k+1}^f\| \leq B\omega(\gamma_{k+2-2n}d_C).$$

Proof. Using the definition of λ_{k+1}^f for $k \geq 2n - 1 \geq n + 1$, we have

$$\begin{aligned} \|\lambda_{k+1}^f\| &= \|\nabla f(x_{k+1}) - \widehat{\nabla} f_{k+1}\| \\ &= \frac{1}{n} \left\| \left(\sum_{i=1}^{\text{mod}(k+1)} \nabla f_i(x_{k+1}) - \nabla f_i(x_{i+k+1-\text{mod}(k+1)}) \right) \right. \\ &\quad \left. + \left(\sum_{i=\text{mod}(k+1)+1}^n \nabla f_i(x_{k+1}) - \nabla f_i(x_{i+k+1-n-\text{mod}(k+1)}) \right) \right\|. \end{aligned}$$

Then, we apply the triangle inequality and ω -smoothness of f_i assumed in (F.1),

$$\begin{aligned} \|\lambda_{k+1}^f\| &\leq \frac{1}{n} \left(\sum_{i=1}^{\text{mod}(k+1)} \|\nabla f_i(x_{k+1}) - \nabla f_i(x_{i+k+1-\text{mod}(k+1)})\| \right. \\ &\quad \left. + \sum_{i=\text{mod}(k+1)+1}^n \|\nabla f_i(x_{k+1}) - \nabla f_i(x_{i+k+1-n-\text{mod}(k+1)})\| \right) \\ &\leq \frac{1}{n} \left(\sum_{i=1}^{\text{mod}(k+1)} \omega(\|x_{k+1} - x_{i+k+1-\text{mod}(k+1)}\|) \right. \\ &\quad \left. + \sum_{i=\text{mod}(k+1)+1}^n \omega(\|x_{k+1} - x_{i+k+1-n-\text{mod}(k+1)}\|) \right). \end{aligned}$$

Now we add and subtract the iterates in between x_{k+1} and $x_{i+k+1-\text{mod}(k+1)}$ then use the definition $x_{k+1} = x_k + \gamma_k (\widehat{s}_k - x_k)$ and the fact that, for all $k \in \mathbb{N}$, \widehat{s}_k and x_k are in C ,

$$\begin{aligned} \|\lambda_{k+1}^f\| &\leq \frac{1}{n} \left(\sum_{i=1}^{\text{mod}(k+1)} \sum_{j=1}^{\text{mod}(k+1)-i} \omega \left(\|x_{k+2-j} - x_{k+1-j}\| \right) \right. \\ &\quad \left. + \sum_{i=\text{mod}(k+1)+1}^n \sum_{j=1}^{\text{mod}(k+1)-i+n} \omega \left(\|x_{k+2-j} - x_{k+1-j}\| \right) \right) \\ &\leq \frac{1}{n} \left(\sum_{i=1}^{\text{mod}(k+1)} \sum_{j=1}^{\text{mod}(k+1)-i} \omega (\gamma_{k+1-j} d_C) \right. \\ &\quad \left. + \sum_{i=\text{mod}(k+1)+1}^n \sum_{j=1}^{\text{mod}(k+1)-i+n} \omega (\gamma_{k+1-j} d_C) \right). \end{aligned}$$

Recall that, by (F.2), $(\gamma_k)_{k \in \mathbb{N}}$ is nonincreasing, by (F.1), ω is a nondecreasing function, and, for each $k \in \mathbb{N}$, $\text{mod}(k) \leq n$. Then,

$$\begin{aligned} \|\lambda_{k+1}^f\| &\leq \frac{1}{n} \left(\sum_{i=1}^{\text{mod}(k+1)} (-i + \text{mod}(k+1)) \omega (\gamma_{k+1+i-\text{mod}(k+1)} d_C) \right. \\ &\quad \left. + \sum_{i=\text{mod}(k+1)+1}^n (-i + n + \text{mod}(k+1)) \omega (\gamma_{k+1+i-n-\text{mod}(k+1)} d_C) \right) \\ &\leq \frac{1}{n} (\text{mod}(k+1) (-1 + \text{mod}(k+1)) \omega (\gamma_{k+2-\text{mod}(k+1)} d_C) \\ &\quad + (n - \text{mod}(k+1)) (-1 + n + \text{mod}(k+1)) \omega (\gamma_{k+2-n-\text{mod}(k+1)} d_C)) \\ &\leq \frac{1}{n} (n(n-1) \omega (\gamma_{k+2-n} d_C) + (n-1)(2n-1) \omega (\gamma_{k+2-2n} d_C)) \\ &\leq \frac{1}{n} (n(n-1) + (n-1)(2n-1)) \omega (\gamma_{k+2-2n} d_C). \end{aligned}$$

□

Proposition 6.2. Under (F.1) and (F.2), and assuming that $(\gamma_k \omega(d_C \gamma_k))_{k \in \mathbb{N}} \in \ell^1$, the summability condition of (P.8) holds; namely,

$$\gamma_{k+1} \|\lambda_{k+1}^f\| \in \ell^1.$$

Proof. By Lemma 6.1, we have, for all $k \geq 2n - 1$,

$$\gamma_{k+1} \|\lambda_{k+1}^f\| \leq B \gamma_{k+1} \omega(d_C \gamma_{k+2-2n}) \leq B \gamma_{k+2-2n} \omega(d_C \gamma_{k+2-2n})$$

where we have used the fact that $(\gamma_k)_{k \in \mathbb{N}}$ is a nonincreasing sequence by (F.2). Since $(\gamma_k \omega(d_C \gamma_k))_{k \in \mathbb{N}} \in \ell^1$, the desired claim follows. □

6.1 COMPARISON OF COMPLEXITIES

We summarize here the complexity of the three methods we have proposed. To make things clear and concrete, we consider a Lipschitz-smooth function $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$ with all f_i Lipschitz-smooth as

well, such that all of our relevant assumptions hold. We take $\gamma_k = 1/(k + 1)^{1-b}$ with the rest of the parameters chosen to satisfy the assumptions laid out in the algorithm (see Section 7 for more details), and all fixed batch sizes to be 1.

	Maximum b	Iterations k	# ∇f_i calls per iter.	Total # ∇f_i calls
Increasing Batch Size	$\approx \frac{1}{3}$	$\Omega(\epsilon^{-3})$	$(k + 1)^{\frac{4}{3}}$	$O\left(\sum_{i=1}^{\epsilon^{-3}} (i + 1)^{\frac{4}{3}}\right)$
Variance Reduction	$\approx \frac{1}{4}$	$\Omega(\epsilon^{-4})$	1	$O(\epsilon^{-4})$
Sweeping	$\approx \frac{1}{3}$	$\Omega(\epsilon^{-3})$	1	$O(\epsilon^{-3})$
Deterministic (CGALP)	$\approx \frac{1}{3}$	$\Omega(\epsilon^{-3})$	n	$O((n\epsilon)^{-3})$

7 NUMERICAL EXPERIMENTS

We apply the sweeping method and the variance reduction method to solve the following projection problem,

$$(7.1) \quad \min_{\substack{\|x\|_1 \leq 1 \\ Ax=0}} \frac{1}{2n} \|x - y\|^2,$$

where x and y are in \mathbb{R}^n . Notice that this problem fits both the risk minimization and the sweeping problem structures. By choosing $f_i(x) = \frac{1}{2}(x_i - y_i)^2$ we can rewrite the problem to apply the sweeping method of Section 6. Alternatively, we can let η be a random variable taking values in the set $\{1, \dots, n\}$ and write $L(x, \eta) = \frac{1}{2}(x_\eta - y_\eta)^2$ to cast the problem as risk minimization as in Section 5. In both of these cases, it is possible by our analysis to consider also sampling components of the components of the gradient term $\nabla_x \frac{\rho_k}{2} \|Ax_k\|^2 = \rho_k A^* Ax_k$.

The assumptions (E.1) - (E.4) and (F.1) all hold as the function f is Lipschitz-smooth and the functions $L(\cdot, \eta)$ are all Lipschitz-smooth for every η as well. The assumptions ((A.1)) to ((a)) all hold as f is Lipschitz-smooth and has full domain.

For parameters, we take $\gamma_k = 1/(k + 1)^{1-b}$, $\rho_k \equiv \rho = 2^{2-b} + 1$, $\theta_k = \gamma_k$. If we take $b < \frac{1}{2}$ then all the assumptions (P.1) to (P.7) are satisfied, as well as (F.2). In particular, to satisfy (P.8) in the variance reduction case, we will take $b \in \{\frac{1}{4} - 0.15, \frac{1}{4} - 0.01\}$. With this choice of b , we have $O\left(\frac{1}{\Gamma_k}\right) = O(k^{-0.1})$ and $O\left(\frac{1}{\Gamma_k}\right) = O(k^{-0.24})$, respectively. The weight v_k in the variance reduction is chosen to be $v_k = \gamma_k^\alpha$ with $\alpha = 2/3$ since the problem is Lipschitz-smooth, i.e. the Hölder exponent is $\tau = 1$. With this choice, the condition (5.3) in Proposition 5.8 is satisfied as was discussed in Example 5.9.

Since the problem (7.1) is strongly convex, we show $\|\bar{x}_k - x^*\|^2$ in addition to the feasibility gap, $\|A\bar{x}_k\|^2$ where \bar{x}_k is the ergodic variable; for each $k \in \mathbb{N}$,

$$\bar{x}_k \stackrel{\text{def}}{=} \sum_{i=0}^k \gamma_i x_{i+1} / \Gamma_k.$$

The result in Theorem 4.11 ensures convergence of $\|\bar{x}_k - x^*\|^2$ with a $O\left(\frac{1}{\Gamma_k}\right)$ rate since strong convexity is a special case of uniform convexity with $\psi(t) = \frac{1}{2}t^2$.

We initialize $y \in \mathbb{R}^n$ and $A \in \mathbb{R}^{2 \times n}$ randomly. To find the solution x^* to high precision, we use generalized forward-backward before running the experiments. As a baseline, we run CGALP, the exact counterpart to ICGALP, and display the results. We run the sweeping method on $\nabla f(x_k)$ for two different step size choices, displayed in Figures 1 and 2. For the variance reduction, we examine both the case where $\nabla L(x_k, \eta_k)$ is sampled and the case where the gradient of the quadratic term is sampled (see Remark 5.2), for two different step size and weight choices as well as different batch sizes (1, 64, or 256), displayed in Figures 1 and 2.

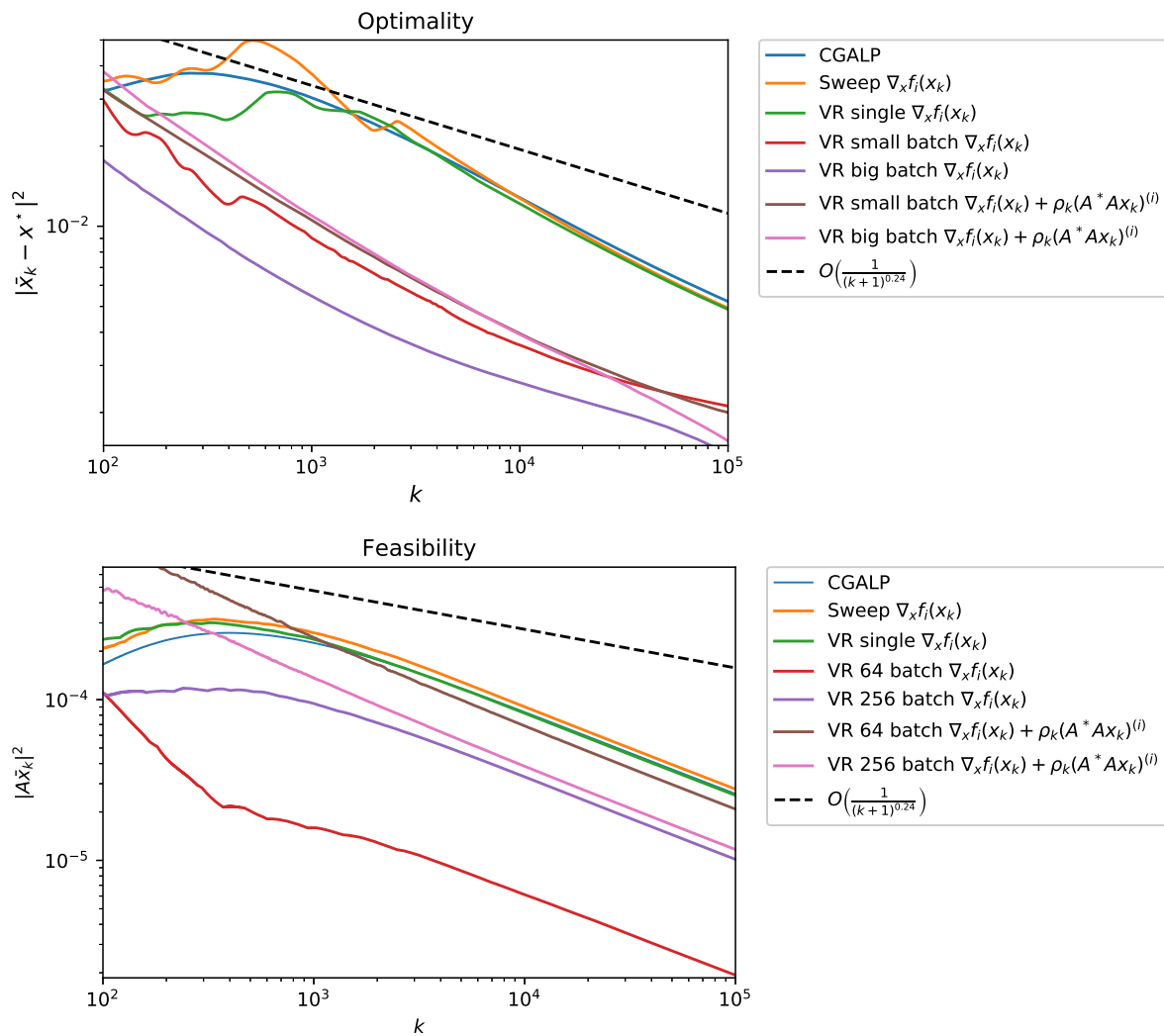


Figure 1: Ergodic convergence profiles for ICGALP applied to the projection problem (7.1) with $n = 1024$. The step size is, for each $k \in \mathbb{N}$, $\gamma_k = (k + 1)^{-(1-\frac{1}{4}+0.01)}$ and the weight for variance reduction is, for each $k \in \mathbb{N}$, $\nu_k = \gamma_k^{2/3}$.

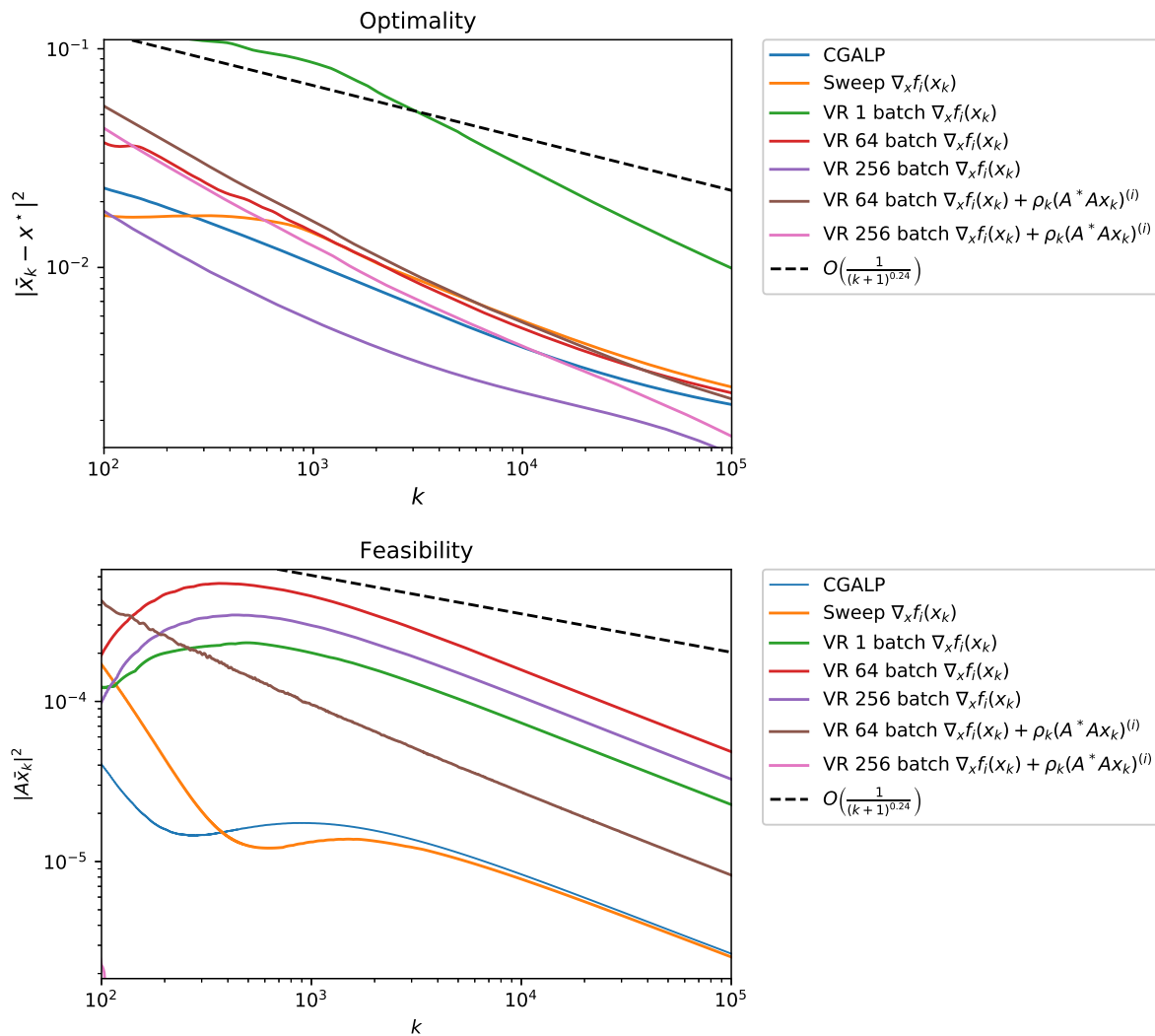


Figure 2: Ergodic convergence profiles for ICGALP applied to the projection problem (7.1) with $n = 1024$. The step size is, for each $k \in \mathbb{N}$, $\gamma_k = (k + 1)^{-(1-\frac{1}{4}+0.15)}$ and the weight for variance reduction is, for each $k \in \mathbb{N}$, $v_k = \gamma_k^{2/3}$.

8 CONCLUSION

We introduced an inexact extension of the CGALP algorithm, given in [34], which allows for either stochastic or deterministic errors in the computation of several important quantities. The main benefit of this extension will be in the high-dimensional setting, where computing the terms ∇f , $\text{prox}_{\beta g}$, or the linear minimization oracle can be impractical. Several different methods were considered which demonstrated how the gradient ∇f could be computed in such a way that the summability conditions of ICGALP would be satisfied. The main drawbacks of using the inexact variant of the algorithm emerge from the restrictions on the parameters one is free to choose. Indeed, here the choices of step sizes are more strict than in the CGALP setting. However, the predicted convergence rates for both the optimality and feasibility maintain the same dependence on parameters as was observed for CGALP in an almost sure sense.

APPENDIX A

Lemma A.1. Consider a positive sequence $(u_k)_{k \in \mathbb{N}}$ which satisfies, for each $k \in \mathbb{N}$,

$$(A.1) \quad u_{k+1} \leq (1 - c\gamma_k^s) u_k + d\gamma_k^t,$$

for some real numbers s and t satisfying $0 < s < \min\{1, t\}$. If, in addition, the sequence $(\gamma_k)_{k \in \mathbb{N}}$ satisfies, for each $k \in \mathbb{N}$,

$$(A.2) \quad \frac{\gamma_k}{\gamma_{k+1}} \leq 1 + o(\gamma_k^s),$$

then, for k sufficiently large, it holds,

$$u_k \leq \frac{d}{c} \gamma_k^{t-s} + o(\gamma_k^{t-s})$$

Proof. For each $k \in \mathbb{N}$, we denote $v_k \stackrel{\text{def}}{=} \gamma_k^{s-t} u_k - \frac{d}{c}$ such that $u_k = \gamma_k^{t-s} \left(v_k + \frac{d}{c} \right)$. Then, by (A.1),

$$v_{k+1} = \gamma_{k+1}^{s-t} u_{k+1} - \frac{d}{c} \leq \gamma_{k+1}^{s-t} \left((1 - c\gamma_k^s) u_k + d\gamma_k^t \right) - \frac{d}{c} = \gamma_k^{s-t} \left(\frac{\gamma_k}{\gamma_{k+1}} \right)^{t-s} \left((1 - c\gamma_k^s) u_k + d\gamma_k^t \right) - \frac{d}{c}.$$

By (A.2), we then have, for each $k \in \mathbb{N}$,

$$v_{k+1} \leq \gamma_k^{s-t} (1 + o(\gamma_k^s))^{t-s} \left((1 - c\gamma_k^s) u_k + d\gamma_k^t \right) - \frac{d}{c}.$$

Substituting for u_k using the definition of v_k we find, for each $k \in \mathbb{N}$,

$$v_{k+1} \leq \gamma_k^{s-t} (1 + o(\gamma_k^s))^{t-s} \left((1 - c\gamma_k^s) \left(v_k + \frac{d}{c} \right) \gamma_k^{t-s} + d\gamma_k^t \right) - \frac{d}{c}.$$

Now, we take a Taylor expansion for the term $(1 + o(\gamma_k^s))^{t-s} \approx (1 + o(\gamma_k^s))$ to get, for k sufficiently large,

$$v_{k+1} \leq \gamma_k^{s-t} (1 + o(\gamma_k^s)) \left((1 - c\gamma_k^s) \left(v_k + \frac{d}{c} \right) \gamma_k^{t-s} + d\gamma_k^t \right) - \frac{d}{c}.$$

We distribute the γ_k^{s-t} and then expand parentheses,

$$\begin{aligned} v_{k+1} &\leq (1 + o(\gamma_k^s)) \left((1 - c\gamma_k^s) \left(v_k + \frac{d}{c} \right) + d\gamma_k^s \right) - \frac{d}{c} \\ &= (1 - c\gamma_k^s) v_k + (1 - c\gamma_k^s) \frac{d}{c} + d\gamma_k^s + o(\gamma_k^s) \left((1 - c\gamma_k^s) \left(v_k + \frac{d}{c} \right) + d\gamma_k^s \right) - \frac{d}{c} \\ &= (1 - c\gamma_k^s) v_k + (1 - c\gamma_k^s) \frac{d}{c} + d\gamma_k^s + o(\gamma_k^s) (1 - c\gamma_k^s) v_k + o(\gamma_k^s) (1 - c\gamma_k^s) \frac{d}{c} + o(\gamma_k^s) d\gamma_k^s - \frac{d}{c} \\ &= (1 - c\gamma_k^s + o(\gamma_k^s)) v_k + o(\gamma_k^s). \end{aligned}$$

Fix $0 < \tilde{c} < c$. Then, by definition of $o(\gamma_k^s)$, $\exists k_0 \in \mathbb{N}$ such that, $\forall k > k_0$, $o(\gamma_k^s) \leq (c - \tilde{c})\gamma_k^s$. Then,

$$(1 - c\gamma_k^s + o(\gamma_k^s)) v_k \leq (1 - \tilde{c}\gamma_k^s) v_k.$$

From this we conclude, by [31, Ch.2, Lemma 3], that $\limsup_k v_k \leq 0$. Thus, by definition of v_k ,

$$u_{k+1} \leq \frac{d}{c} \gamma_k^{t-s} + o(\gamma_k^{t-s}). \quad \square$$

REFERENCES

- [1] Y. I. Alber, A. N. Iusem, and M. V. Solodov, On the projected subgradient method for nonsmooth convex optimization in a Hilbert space, *Mathematical Programming* 81 (1998), 23–35.
- [2] L. Ambrosio, N. Gigli, and G. Savare, *Gradient Flows*, Lectures in Mathematics. ETH Zürich, Birkhäuser Basel, 2008.
- [3] H. Attouch, *Variational Convergence for Functions and Operators*, Applicable mathematics series, Pitman Advanced Publishing Program, 1984.
- [4] F. Bach, Breaking the curse of dimensionality with convex neural networks, *The Journal of Machine Learning Research* 18 (2017), 629–681.
- [5] K. Barty, J. S. Roy, and C. Strugarek, Hilbert-valued perturbed subgradient algorithms, *Mathematics of Operations Research* 32 (2007), 551–562, <http://www.jstor.org/stable/25151809>.
- [6] H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2011.
- [7] N. Boyd, G. Schiebinger, and B. Recht, The alternating descent conditional gradient method for sparse inverse problems, *SIAM Journal on Optimization* 27 (2017), 616–639.
- [8] K. Bredies, M. Carioni, S. Fanzon, and F. Romero, A generalized conditional gradient method for dynamic inverse problems with optimal transport regularization (2020), [arXiv:2012.11706](https://arxiv.org/abs/2012.11706).
- [9] H. Brezis and A. Pazy, Convergence and approximation of semigroups of nonlinear operators in Banach spaces, *J. Functional Analysis* 9 (1972), 63–74.
- [10] P. Catala, *Relaxations Semi-Définies Positives pour l’Imagerie*, PhD thesis, PSL University, 2020.
- [11] C. W. Combettes and S. Pokutta, Complexity of linear minimization and projection on some sets (2021), [arXiv:2101.10040](https://arxiv.org/abs/2101.10040).

- [12] P. L. Combettes and J. C. Pesquet, Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping, *SIAM Journal on Optimization* 25 (2015), 1221–1248.
- [13] P. L. Combettes and J. C. Pesquet, Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping II: mean-square and linear convergence, *Mathematical Programming* 174 (2019), 433–451, [doi:10.1007/s10107-018-1296-y](https://doi.org/10.1007/s10107-018-1296-y).
- [14] A. d’Aspremont and M. Pilanci, Global Convergence of Frank Wolfe on One Hidden Layer Networks, *arXiv preprint arXiv:2002.02208* (2020).
- [15] Q. Denoyelle, V. Duval, G. Peyré, and E. Soubies, The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy, *Inverse Problems* 36 (2019), 014001.
- [16] L. Ding and M. Udell, Frank-Wolfe style algorithms for large scale optimization, in *Large-Scale and Distributed Optimization*, P. Giselsson and A. Rantzer (eds.), Springer International Publishing, Cham, 2018, 215–245, [doi:10.1007/978-3-319-97478-1_9](https://doi.org/10.1007/978-3-319-97478-1_9).
- [17] M. Franke and P. Wolfe, An algorithm for quadratic programming, *Naval Research Logistics Quarterly* 3 (1956), 95–110.
- [18] D. Goldfarb, G. Iyengar, and C. Zhou, Linear convergence of stochastic Frank Wolfe variants (2017), [arXiv:1703.07269](https://arxiv.org/abs/1703.07269).
- [19] H. Hassani, A. Karbasi, A. Mokhtari, and Z. Shen, Stochastic Conditional Gradient++ (2019), [arXiv:1902.06992](https://arxiv.org/abs/1902.06992).
- [20] E. Hazan and S. Kale, Projection-free online learning, in *ICML*, 2012.
- [21] E. Hazan and H. Luo, Variance-reduced and projection-free stochastic optimization, in *ICML*, 2016.
- [22] C. Imbert, Convex Analysis techniques for Hopf-Lax formulae in Hamilton-Jacobi equations, *Journal of Nonlinear and Convex Analysis* 2 (2001), 333–343.
- [23] R. Johnson and T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction, *Advances in neural information processing systems* 26 (2013), 315–323.
- [24] E. Levitin and B. Polyak, Constrained minimization methods, *USSR Computational Mathematics and Mathematical Physics* 6 (1966), 1 – 50, [doi:10.1016/0041-5553\(66\)90114-5](https://doi.org/10.1016/0041-5553(66)90114-5).
- [25] F. Locatello, A. Yurtsever, O. Fercoq, and V. Cevher, Stochastic conditional gradient method for composite convex minimization (2019), [arXiv:1901.10348](https://arxiv.org/abs/1901.10348), [arXiv:1901.10348](https://arxiv.org/abs/1901.10348).
- [26] H. Lu and R. M. Freund, Generalized stochastic Frank-Wolfe algorithm with stochastic substitute gradient for structured convex optimization, *Mathematical Programming* (2020), [doi:10.1007/s10107-020-01480-7](https://doi.org/10.1007/s10107-020-01480-7).
- [27] A. Mokhtari, H. Hassani, and A. Karbasi, Stochastic conditional gradient methods: from convex minimization to submodular maximization, *arXiv e-prints* (2018), [arXiv:1804.09554](https://arxiv.org/abs/1804.09554), [arXiv:1804.09554](https://arxiv.org/abs/1804.09554).
- [28] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč, SARAH: A novel method for machine learning problems using stochastic recursive gradient, in *International Conference on Machine Learning*, 2017, 2613–2621.

- [29] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takac, SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient, in *ICML*, 2017.
- [30] J. Peypouquet, *Convex Optimization in Normed Spaces: Theory, Methods and Examples*, Springer, 2015.
- [31] B. T. Polyak, *Introduction to Optimization*, Optimization Software, 1987.
- [32] S. J. Reddi, S. Sra, B. Póczos, and A. Smola, Stochastic Frank-Wolfe methods for nonconvex optimization, *arXiv e-prints* (2016), arXiv:1607.08254.
- [33] H. Robbins and D. Siegmund, A convergence theorem for non negative almost supermartingales and some applications, in *Herbert Robbins Selected Papers*, Springer New York, 1985, 111–135, doi: [10.1007/978-1-4612-5110-1_10](https://doi.org/10.1007/978-1-4612-5110-1_10).
- [34] A. Silveti-Falls, C. Molinari, and J. Fadili, Generalized conditional gradient with augmented Lagrangian for composite minimization, *SIAM Journal on Optimization* 30 (2020), 2687–2725, doi: [10.1137/19m1240460](https://doi.org/10.1137/19m1240460).
- [35] X. Wei and M. J. Neely, Primal-dual Frank-Wolfe for constrained stochastic programs with convex and non-convex objectives (2018), [arXiv:1806.00709](https://arxiv.org/abs/1806.00709).