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 [16], which we denote ICGALP , that allows 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 prox operator and a linear minimization oracle. As main results, we show convergence of the Lagrangian to an optimum and asymptotic feasibility of the affine constraint as well as weak convergence of the dual variable to a solution of the dual problem, all in an almost sure sense. Almost sure convergence rates, both pointwise and ergodic, are given for the Lagrangian values and the feasibility gap. Numerical experiments verifying the predicted rates of convergence are shown as well.

Key words. Conditional gradient; Augmented Lagrangian; Composite minimization; Proximal mapping; Moreau envelope.

AMS subject classifications. 49J52, 65K05, 65K10.

1 Introduction

1.1 Problem Statement

We consider the following composite minimization problem,

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

and its associated *dual problem*,

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

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,

*Normandie Université, ENSICAEN, UNICAEN, CNRS, GREYC, France. E-mail: tonys.falls@gmail.com, cecio.molinari@gmail.com, Jalal.Fadili@ensicaen.fr.

whose indices correspond to a primal, dual, and auxilliary space, respectively; $A : \mathcal{H}_p \to \mathcal{H}_d$ and $T : \mathcal{H}_p \to \mathcal{H}_v$ to be bounded linear operators with $b \in \mathrm{ran}(A)$; 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 2.6), that the function g has a proximal mapping which is accessible, and that the function h admits an accessible linearly-perturbed minimization oracle with $C \stackrel{\text{def}}{=} \mathrm{dom}\,(h)$ a weakly compact subset of $\mathcal{H}_p$.

In fact, the problem under consideration here is exactly the same as that of [16], however, in this work, we consider an inexact extension of the algorithm presented and analyzed in [16] to solve ($\mathscr{P}$). The extension amounts to allowing either deterministic or stochastic errors in the computation of several quantities, including the gradient or prox terms, e.g. ∇f, $\mathrm{prox}_{\beta g}$, 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 [16] to address ($\mathscr{P}$). We coin this algorithm **I**nexact **C**onditional **G**radient with **A**ugemented **L**agrangian and **P**roximal-step (ICGALP). Although there has been a great deal of work on developing and analyzing Frank-Wolfe or conditional gradient style algorithms in both the stochastic and deterministic case, e.g. [6, 7, 14, 4, 3, 17, 10, 5], or [9], 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 [8], where the authors consider a stochastic conditional gradient algorithm applied to a composite problem allowing nonsmooth terms. The nonsmooth term is possibly an affine constraint but it is addressed through smoothing rather than through an augmented Lagrangian with a dual variable, in contrast to our work.

We show asymptotic feasibility of the primal iterates for the affine constraint, convergence of the Lagrangian values at each iteration to an optimum value, weak convergence of the sequence of dual iterates to a solution of the dual problem, and provide worst-case rates of convergence for the feasibility gap and the Lagranian values. The rates of convergence are given both subsequentially in the pointwise sense and globally, i.e. for the entire sequence of iterates, in the ergodic sense where the Cesáro means are taken with respect to the primal step size. In the case where ($\mathscr{P}$) admits a unique solution, we furthermore have that the sequence of primal iterates converges weakly to the solution. These results are shown to hold almost surely and 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 determinstc setting for minimizing an empirical risk, a sweeping approach is described.

1.3 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 extend those of [16] to the inexact and stochastic setting. 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 [2]. 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 $\{a_0, \ldots, a_n\}$, we denote by $\sigma(a_0, \ldots, a_n)$ the σ-algebra generated by $a_0, \ldots, a_n$. An expression (P) is said to hold $(\mathbb{P}\text{-a.s.})$ 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,

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

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,*

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

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

Proof. See [15, 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 $\left(w_{k_j}\right)_{j\in\mathbb{N}}$ such that

$$w_{k_j} \leq \Gamma_{k_j}^{-1} \ (\mathbb{P}\text{-a.s.})$$

where $\Gamma_{k_j} = \sum_{n=1}^{k_j} \gamma_n$. In particular, $\liminf_k w_k = 0$ $(\mathbb{P}\text{-a.s.})$.

(ii) Furthermore, if there exists a constant $\alpha > 0$ such that $w_k - \mathbb{E}[w_{k+1} \mid \mathcal{F}_k] \leq \alpha\gamma_k$ $(\mathbb{P}\text{-a.s.})$ for every $k \in \mathbb{N}$, then

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

Proof. The main result is directly from [1, Lemma 2.2] and the rates follow from [18] trivially extended to the stochastic setting. □

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

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

The *proximal mapping* (or *proximal operator*) associated to the function f with parameter β is given by,

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

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

Proposition 2.4. *Let $f \in \Gamma_0(\mathcal{H})$ and denote $x^+ = \mathrm{prox}_f(x)$. Then, for all $y \in \mathcal{H}$,*

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

Proof. The result is classical and the proof is readily available, e.g. in [12, Chapter 6.2.1]. □

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

$$\partial f(x) \stackrel{\text{def}}{=} \{u \in \mathcal{H} : f(y) \geq f(x) + \langle u, y - x \rangle \quad \forall y \in \mathcal{H}\} \tag{2.1}$$

We denote $\mathrm{dom}(\partial f) \stackrel{\text{def}}{=} \{x \in \mathcal{H} : \partial f(x) \neq \emptyset\}$ as the *domain of the subdifferential*. For $x \in \mathrm{dom}(\partial f)$, the *minimal norm selection* of $\partial f(x)$ is denoted by $[\partial f(x)]^0 \stackrel{\text{def}}{=} \operatorname*{argmin}_{y \in \partial f(x)} \|y\|$. 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 2.5 (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:

$$\begin{cases} \frac{\partial}{\partial \beta} f^\beta(x) = -\frac{1}{2} \left\|\nabla_x f^\beta(x)\right\|^2 & (x, \beta) \in \mathcal{H} \times (0, +\infty) \\ f^0(x) = f(x) & x \in \mathcal{H}. \end{cases} \tag{2.2}$$

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

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

(iv) $\forall x \in \mathrm{dom}(\partial f)$, $\left\|\nabla f^\beta(x)\right\| \nearrow \left\|[\partial f(x)]^0\right\|$ 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*

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

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

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

Definition 2.6 ((F, ζ)-smoothness). Let $F : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ and $\zeta :]0, 1] \to \mathbb{R}_+$. The pair $(f, \mathcal{C})$, where $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ and $\mathcal{C} \subset \operatorname{dom}(f)$, is said to be (F, ζ)-smooth if there exists an open set $\mathcal{C}_0$ such that $\mathcal{C} \subset \mathcal{C}_0 \subset \operatorname{int}(\operatorname{dom}(F))$ and,

(i) F and f are differentiable on $\mathcal{C}_0$;
(ii) $F - f$ is convex on $\mathcal{C}_0$;
(iii) it holds

$$K_{(F,\zeta,\mathcal{C})} \overset{\text{def}}{=} \sup_{\substack{x,s \in \mathcal{C}; \gamma \in]0,1] \\ z = x + \gamma(s - x)}} \frac{D_F(z, x)}{\zeta(\gamma)} < +\infty. \tag{2.3}$$

Remark 2.7. An important consequence of Definition 2.6(i) and Definition 2.6(ii) in (F, ζ)-smoothness is the following. Let $(f, \mathcal{C})$ be (F, ζ) smooth. Then, for any $x, y \in \mathcal{C}$, we have,

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

Moreover, by Definition 2.6(iii), if $y = x + \gamma(s - x)$ for some $s \in \mathcal{C}$ and $\gamma \in]0, 1]$, we have,

$$D_F(y, x) \leq K_{(F,\zeta,\mathcal{C})} \zeta(\gamma). \tag{2.4}$$

Definition 2.8 (ω-smoothness). Consider a function $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\omega(0) = 0$ and $\xi(s) \overset{\text{def}}{=} \int_0^1 \omega(st)\, dt$ is nondecreasing. A differentiable function $g : \mathcal{H} \to \mathbb{R}$ 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 2.9. A classical consequence of ω-smoothness is the following. If $g : \mathcal{H} \to \mathbb{R}$ 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 2.10. 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 $F = f$, $\zeta(t) = d_{\mathcal{C}} t \xi(d_{\mathcal{C}} t)$ and $K_{(F,\zeta,\mathcal{C})} \leq 1$. Additionally, the assumptions on ξ being nondecreasing can be replaced by the sufficient condition that $\lim_{t \to 0^+} \omega(t) = \omega(0) = 0$.

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: $x_0 \in \mathcal{C} \stackrel{\text{def}}{=} \operatorname{dom}(h)$; $\mu_0 \in \operatorname{ran}(A)$; $(\gamma_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}$, $(\rho_k)_{k\in\mathbb{N}} \in \ell_+$.
$k = 0$
repeat
$y_k = \operatorname{prox}_{\beta_k g}(Tx_k)$
$z_k = \nabla f(x_k) + T^*(Tx_k - y_k)/\beta_k + A^*\mu_k + \rho_k A^*(Ax_k - b) + \lambda_k$
$s_k \in \operatorname{Argmin}_{s\in\mathcal{H}_p}\{h(s) + \langle z_k, s\rangle\}$
$\widehat{s}_k \in \{s \in \mathcal{H}_p : h(s) + \langle z_k, s\rangle \leq h(s_k) + \langle z_k, s_k\rangle + \lambda_k^s\}$
$x_{k+1} = x_k - \gamma_k(x_k - \widehat{s}_k)$
$\mu_{k+1} = \mu_k + \theta_k(Ax_{k+1} - b)$
$k \leftarrow k + 1$
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,

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

We can recognize $\mathcal{L}(x,\mu)$ as the classical Lagrangian, $\mathcal{L}_k(x,\mu)$ as the augmented Lagrangian with smoothed g, and $\mathcal{E}_k(x,\mu)$ as the smooth part of $\mathcal{L}_k(x,\mu)$. 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 filtration $\mathfrak{S} \stackrel{\text{def}}{=} (\mathcal{S}_k)_{k\in\mathbb{N}}$ where $\mathcal{S}_k \stackrel{\text{def}}{=} \sigma(x_0, \mu_0, \widehat{s}_0, \ldots, \widehat{s}_k)$ is the σ-algebra generated by the random variables $x_0, \mu_0, \widehat{s}_0, \ldots, \widehat{s}_k$. 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 $\mathcal{C}$, as in the classical Frank-Wolfe algorithm, while the dual iterates $(\mu_k)_{k\in\mathbb{N}}$ remain in $\operatorname{ran}(A)$.

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) $f, g \circ T$, and h belong to $\Gamma_0(\mathcal{H}_p)$
(A.2) The pair $(f, \mathcal{C})$ is (F, ζ)-smooth (see Definition 2.6), where we recall $\mathcal{C} \stackrel{\text{def}}{=} \operatorname{dom}(h)$
(A.3) $\mathcal{C}$ is weakly compact (and thus contained in a ball of radius $R > 0$)
(A.4) $T\mathcal{C} \subset \operatorname{dom}(\partial g)$ and $\sup_{x \in \mathcal{C}} \left\| [\partial g(Tx)]^0 \right\| = M < \infty$
(A.5) h is Lipschitz continuous relative to its domain $\mathcal{C}$ with constant $L_h \geq 0$, i.e., $\forall (x, z) \in \mathcal{C}^2$, $|h(x) - h(z)| \leq L_h \|x - z\|$.
(A.6) There exists a saddle-point $(x^\star, \mu^\star) \in \mathcal{H}_p \times \mathcal{H}_d$ for the Lagrangian $\mathcal{L}$
(A.7) $\operatorname{ran}(A)$ is closed
(A.8) One of the following holds:
  (a) $A^{-1}(b) \cap \operatorname{int}(\operatorname{dom}(g \circ T)) \cap \operatorname{int}(\mathcal{C}) \neq \emptyset$, where $A^{-1}(b)$ is the pre-image of b under A
  (b) $\mathcal{H}_p$ and $\mathcal{H}_d$ are finite-dimensional and

$$\begin{cases} A^{-1}(b) \cap \operatorname{ri}(\operatorname{dom}(g \circ T)) \cap \operatorname{ri}(\mathcal{C}) \neq \emptyset \\ \text{and} \\ \operatorname{ran}(A^*) \cap \operatorname{par}(\operatorname{dom}(g \circ T) \cap \mathcal{C})^\perp = \{0\}. \end{cases} \tag{3.2}$$

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 $0 < \underline{\rho} \leq \rho_k \leq \overline{\rho} < +\infty$
(P.5) For some positive constants $\underline{M}$ and $\overline{M}$, $\underline{M} \leq (\gamma_k/\gamma_{k+1}) \leq \overline{M}$
(P.6) $(\theta_k)_{k \in \mathbb{N}}$ satisfies $\theta_k = \frac{\gamma_k}{c}$ for some $c > 0$ such that $\frac{\overline{M}}{c} - \frac{\rho}{2} < 0$
(P.7) $(\gamma_k)_{k \in \mathbb{N}}$ and $(\rho_k)_{k \in \mathbb{N}}$ satisfy $\rho_{k+1} - \rho_k - \gamma_{k+1}\rho_{k+1} + \frac{2}{c}\gamma_k - \frac{\gamma_k^2}{c} \leq \gamma_{k+1}$ for c 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 $\left(\gamma_{k+1} \mathbb{E}\left[\lambda_{k+1}^s \mid \mathcal{S}_k\right]\right)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$.

Remark 3.1. 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 $\operatorname{prox}_{\beta_k g}(Tx_k)$. 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 $\left(\gamma_{k+1} \mathbb{E}\left[\left\|\lambda_{k+1}^f\right\| \mid \mathcal{S}_k\right]\right)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$ and $\left(\frac{\gamma_{k+1}}{\beta_{k+1}} \mathbb{E}\left[\left\|\lambda_{k+1}^g\right\| \mid \mathcal{S}_k\right]\right)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$.

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*

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

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

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

Proof. See [16, Lemma 4.5] □

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

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

Proof. See [16, 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 before. Then, for each $k \in \mathbb{N}$, we have the following estimate,*

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

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,

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

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

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

where we have used the Cauchy-Schwartz inequality and the boundedness of $\mathcal{C}$, 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 $k \in \mathbb{N}$,

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

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 $\mu^\star$ a solution, which exists by (A.6), of the dual problem. Then, for each $k \in \mathbb{N}$, we have the following estimate,*

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

Proof. We recall the proof from [16, Lemma 4.7] with a slight modification to account for the inexactness of the algorithm. Define $u_k \overset{\text{def}}{=} [\partial g(Tx_k)]^0$ and recall that, by (A.4) and the fact that for all $k \in \mathbb{N}$, $x_k \in \mathcal{C}$, we have $\|u_k\| \leq M$. By (A.1), the function $\Phi(x) \overset{\text{def}}{=} f(x) + g(Tx) + h(x)$ is convex. Then, for each $k \in \mathbb{N}$, ,

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

where we used the subdifferential inequality (2.1) on g and f, the L_h-Lipschitz continuity of h relative to $\mathcal{C}$ (see (A.5)), and the Cauchy-Schwartz 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 $k \in \mathbb{N}$,

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

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

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

where we have used the Cauchy-Schwartz inequality, the boundedness of the set $\mathcal{C}$ by (A.3), the boundedness of u_k by M by (A.4), and denoted by D the constant $D \overset{\text{def}}{=} \sup_{x\in\mathcal{C}} \|\nabla f(x)\| < +\infty$. Note that D exists and is bounded since f is differentiable on an open set $\mathcal{C}_0$ containing $\mathcal{C}$ by (A.2) and Definition 2.6. □

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,*

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

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

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

Proof. The proof here is adapted from the analogous result found in [16, 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 [16, Theorem 4.1]). We denote a minimizer of $\mathcal{L}_k(x, \mu_k)$ by $\tilde{x}_k \in \underset{x \in \mathcal{H}_p}{\operatorname{Argmin}} \mathcal{L}_k(x, \mu_k)$, which exists and belongs to $\mathcal{C}$ by (A.1)-(A.3). We have, for each $k \in \mathbb{N}$,

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

Recall that $\tilde{x}_k \in \underset{x \in \mathcal{H}_p}{\operatorname{Argmin}} \mathcal{L}_k(x, \mu_k)$, that $g^{\beta_k} \leq g^{\beta_{k+1}}$ due to (P.3) and Proposition 2.5(v), and that $\rho_k \leq \rho_{k+1}$ by (P.4). Then, for each $k \in \mathbb{N}$,

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

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 $k \in \mathbb{N}$,

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

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

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

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

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

We denote by $\mathbf{T1} \overset{\text{def}}{=} \mathcal{L}_{k+1}(x_{k+2}, \mu_{k+1}) - \mathcal{L}_{k+1}(x_{k+1}, \mu_{k+1})$ 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 $k \in \mathbb{N}$,

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

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 $k \in \mathbb{N}$,

$$\begin{aligned}
\mathbf{T1} &\leq \gamma_{k+1}\left(h\left(\widehat{s}_{k+1}\right) - h\left(x_{k+1}\right) + \left\langle \nabla_x \mathcal{E}_{k+1}\left(x_{k+1}, \mu_{k+1}\right), \widehat{s}_{k+1} - x_{k+1}\right\rangle\right) \\
&\quad + \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + D_F\left(x_{k+2}, x_{k+1}\right) \\
&= \gamma_{k+1}\Big(h\left(\widehat{s}_{k+1}\right) - h\left(x_{k+1}\right) + \left\langle \widehat{\nabla_x \mathcal{E}}_{k+1}\left(x_{k+1}, \mu_{k+1}\right), \widehat{s}_{k+1} - x_{k+1}\right\rangle \\
&\quad + \left\langle \nabla_x \mathcal{E}_{k+1}\left(x_{k+1}, \mu_{k+1}\right) - \widehat{\nabla_x \mathcal{E}}_{k+1}\left(x_{k+1}, \mu_{k+1}\right), \widehat{s}_{k+1} - x_{k+1}\right\rangle\Big) \\
&\quad + \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + D_F\left(x_{k+2}, x_{k+1}\right) \\
&= \gamma_{k+1}\Big(h\left(\widehat{s}_{k+1}\right) - h\left(x_{k+1}\right) + \left\langle \widehat{\nabla_x \mathcal{E}}_{k+1}\left(x_{k+1}, \mu_{k+1}\right), \widehat{s}_{k+1} - x_{k+1}\right\rangle \\
&\quad - \left\langle \lambda_{k+1}, \widehat{s}_{k+1} - x_{k+1}\right\rangle\Big) + \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + D_F\left(x_{k+2}, x_{k+1}\right)
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{T1} &\leq \gamma_{k+1}\Big(h\left(s_{k+1}\right) - h\left(x_{k+1}\right) + \left\langle \widehat{\nabla_x \mathcal{E}}_{k+1}\left(x_{k+1}, \mu_{k+1}\right), s_{k+1} - x_{k+1}\right\rangle + \lambda^s_{k+1} \\
&\quad - \left\langle \lambda_{k+1}, \widehat{s}_{k+1} - x_{k+1}\right\rangle\Big) + \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + D_F\left(x_{k+2}, x_{k+1}\right).
\end{aligned}$$

Now 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{T1} &\leq \gamma_{k+1}\Big(h\left(\tilde{x}_{k+1}\right) - h\left(x_{k+1}\right) + \left\langle \widehat{\nabla_x \mathcal{E}}_{k+1}\left(x_{k+1}, \mu_{k+1}\right), \tilde{x}_{k+1} - x_{k+1}\right\rangle + \lambda^s_{k+1} \\
&\quad - \left\langle \lambda_{k+1}, \widehat{s}_{k+1} - x_{k+1}\right\rangle\Big) + \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + D_F\left(x_{k+2}, x_{k+1}\right) \\
&= \gamma_{k+1}\Big(h\left(\tilde{x}_{k+1}\right) - h\left(x_{k+1}\right) + \left\langle \nabla_x \mathcal{E}_{k+1}\left(x_{k+1}, \mu_{k+1}\right), x_{k+1} - \tilde{x}_{k+1}\right\rangle + \lambda^s_{k+1} \\
&\quad - \left\langle \lambda_{k+1}, \widehat{s}_{k+1} - \tilde{x}_{k+1}\right\rangle\Big) + \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + D_F\left(x_{k+2}, x_{k+1}\right) \\
&\leq \gamma_{k+1}\Big(h\left(\tilde{x}_{k+1}\right) - h\left(x_{k+1}\right) + \mathcal{E}_{k+1}\left(\tilde{x}_{k+1}, \mu_{k+1}\right) - \mathcal{E}_{k+1}\left(x_{k+1}, \mu_{k+1}\right) - \frac{\rho_{k+1}}{2}\left\|A\left(x_{k+1} - \tilde{x}_{k+1}\right)\right\|^2 \\
&\quad + \lambda^s_{k+1} - \left\langle \lambda_{k+1}, \widehat{s}_{k+1} - \tilde{x}_{k+1}\right\rangle\Big) + \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + D_F\left(x_{k+2}, x_{k+1}\right) \\
&= \gamma_{k+1}\Big(\mathcal{L}_{k+1}\left(\tilde{x}_{k+1}, \mu_{k+1}\right) - \mathcal{L}_{k+1}\left(x_{k+1}, \mu_{k+1}\right) - \frac{\rho_{k+1}}{2}\left\|A\left(x_{k+1} - \tilde{x}_{k+1}\right)\right\|^2 + \lambda^s_{k+1} \\
&\quad - \left\langle \lambda_{k+1}, \widehat{s}_{k+1} - \tilde{x}_{k+1}\right\rangle\Big) + \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + D_F\left(x_{k+2}, x_{k+1}\right) \\
&\leq -\frac{\gamma_{k+1}\rho_{k+1}}{2}\left\|A\left(x_{k+1} - \tilde{x}_{k+1}\right)\right\|^2 + \gamma_{k+1}\Big(\lambda^s_{k+1} + \left\langle \lambda_{k+1}, \tilde{x}_{k+1} - \widehat{s}_{k+1}\right\rangle\Big) \\
&\quad + \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + D_F\left(x_{k+2}, x_{k+1}\right),
\end{aligned}$$

where we used that $\tilde{x}_{k+1}$ is a minimizer of $\mathcal{L}_{k+1}\left(\cdot, \mu_{k+1}\right)$ in the last inequality. Now combining **T1** and **T2**

and using the Pythagoras identity we have, for each $k \in \mathbb{N}$,

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

Now take the expectation with respect to $\mathcal{F}_k = \mathcal{S}_k = \sigma\left(x_0, \mu_0, \widehat{s}_0, \ldots, \widehat{s}_k\right)$, 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
$$-\theta_k \leq -\tfrac{\underline{M}}{c}\gamma_{k+1}.$$
- Again by (P.6), we have, for each $k \in \mathbb{N}$, $\theta_k = \gamma_k/c$ for some $c > 0$ such that
$$\exists \delta > 0, \quad \frac{\overline{M}}{c} - \frac{\underline{\rho}}{2} = -\delta < 0,$$
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 $k \in \mathbb{N}$,
$$\theta_k - \gamma_{k+1}\frac{\rho_{k+1}}{2} \leq \left(\frac{\overline{M}}{c} - \frac{\rho_{k+1}}{2}\right)\gamma_{k+1} \leq \left(\frac{\overline{M}}{c} - \frac{\underline{\rho}}{2}\right)\gamma_{k+1} = -\delta\gamma_{k+1}. \tag{4.2}$$
- By Algorithm 1, for each $k \in \mathbb{N}$, $x_{k+2} - x_{k+1} = \gamma_{k+1}\left(\widehat{s}_{k+1} - x_{k+1}\right)$. Since $\widehat{s}_{k+1}$ and x_{k+1} are both in $\mathcal{C}$ and $\mathcal{C}$ is bounded due to (A.3), for each $k \in \mathbb{N}$,
$$\frac{L_{k+1}}{2}\mathbb{E}\left[\left\|x_{k+2} - x_{k+1}\right\|^2 \mid \mathcal{F}_k\right] = \frac{L_{k+1}}{2}\gamma_{k+1}^2\mathbb{E}\left[\left\|\widehat{s}_{k+1} - x_{k+1}\right\|^2 \mid \mathcal{F}_k\right] \leq \frac{L_{k+1}}{2}\gamma_{k+1}^2 d_{\mathcal{C}}^2.$$
- Recall that, by (A.2), f is (F, ζ)-smooth and invoke Remark 2.7, to get
$$\mathbb{E}\left[D_F\left(x_{k+2}, x_{k+1}\right) \mid \mathcal{F}_k\right] \leq K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_{k+1}\right).$$
- By Proposition 2.5(v) and assumption (A.4),
$$\mathbb{E}\left[\left[g^{\beta_{k+1}} - g^{\beta_k}\right](Tx_{k+1}) \mid \mathcal{F}_k\right] \leq \frac{\beta_k - \beta_{k+1}}{2}\mathbb{E}\left[\left\|\left[\partial g\left(Tx_{k+1}\right)\right]^0\right\|^2 \mid \mathcal{F}_k\right] \leq \frac{\beta_k - \beta_{k+1}}{2}M^2.$$
- We also have, using Jensen's inequality and (A.3), for each $k \in \mathbb{N}$,
$$\left(\frac{\rho_{k+1} - \rho_k}{2}\right)\mathbb{E}\left[\left\|Ax_{k+1} - b\right\|^2 \mid \mathcal{F}_k\right] \leq \left(\rho_{k+1} - \rho_k\right)\left(\|A\|^2 R^2 + \|b\|^2\right).$$

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

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

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

$$\gamma_{k+1}\mathbb{E}\left[\langle \lambda_{k+1}, \tilde{x}_{k+1} - \widehat{s}_{k+1}\rangle \mid \mathcal{F}_k\right] \leq \gamma_{k+1} d_{\mathcal{C}} \mathbb{E}\left[\|\lambda_{k+1}\| \mid \mathcal{F}_k\right], \tag{4.3}$$

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

$$\begin{aligned}\mathbb{E}\left[\Delta_{k+1} \mid \mathcal{F}_k\right] - \Delta_k \leq & -\frac{M}{c}\gamma_{k+1}\left\|A\tilde{x}_{k+1} - b\right\|^2 - \delta\gamma_{k+1}\left\|A\left(x_{k+1} - \tilde{x}_{k+1}\right)\right\|^2 + \gamma_{k+1}^2 \frac{L_{k+1}}{2} d_{\mathcal{C}}^2 \\ & + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_{k+1}\right) + \frac{\beta_k - \beta_{k+1}}{2}M^2 + \left(\rho_{k+1} - \rho_k\right)\left(\|A\|^2 R^2 + \|b\|^2\right) \\ & + \gamma_{k+1}\mathbb{E}\left[\lambda_{k+1}^s \mid \mathcal{F}_k\right] + \gamma_{k+1} d_{\mathcal{C}} \mathbb{E}\left[\|\lambda_{k+1}\| \mid \mathcal{F}_k\right],\end{aligned} \tag{4.4}$$

and we conclude by trivial manipulations. □

Theorem 4.7 (Feasibility). *Suppose that (A.1)-(A.4) and (A.6) all hold. 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 feasbility: $\lim_{k\to\infty} \|Ax_k - b\| = 0$ *($\mathbb{P}$-a.s.)*

(ii) Pointwise rate:

$$\begin{aligned}\inf_{0\leq i\leq k} \|Ax_i - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right) \ (\mathbb{P}\text{-a.s.}) \quad \text{and} \\ \exists \text{ a subsequence } \left(x_{k_j}\right)_{j\in\mathbb{N}} \text{ such that } \|Ax_{k_j} - b\| \leq \frac{1}{\sqrt{\Gamma_{k_j}}} \ (\mathbb{P}\text{-a.s.}) \ .\end{aligned} \tag{4.5}$$

where $\Gamma_k \overset{\text{def}}{=} \sum_{i=0}^k \gamma_i$.

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

$$\|A\bar{x}_k - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right) \ (\mathbb{P}\text{-a.s.}) \ . \tag{4.6}$$

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

$$\begin{aligned}\mathbb{E}\left[\Delta_{k+1} \mid \mathcal{F}_k\right] - \Delta_k \leq & -\gamma_{k+1}\left(\frac{M}{c}\left\|A\tilde{x}_{k+1} - b\right\|^2 + \delta\left\|A\left(x_{k+1} - \tilde{x}_{k+1}\right)\right\|^2\right) + \gamma_{k+1}^2 \frac{L_{k+1}}{2} d_{\mathcal{C}}^2 \\ & + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_{k+1}\right) + \frac{\beta_k - \beta_{k+1}}{2}M^2 + \left(\rho_{k+1} - \rho_k\right)\left(\|A\|^2 R^2 + \|b\|^2\right) \\ & + \gamma_{k+1}\mathbb{E}\left[\lambda_{k+1}^s \mid \mathcal{F}_k\right] + d_{\mathcal{C}}\gamma_{k+1}\mathbb{E}\left[\|\lambda_{k+1}\| \mid \mathcal{F}_k\right].\end{aligned} \tag{4.7}$$

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

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

For the telescopic terms from the right hand side of (4.7) we have

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

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

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

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

$$\left(\gamma_{k+1}\mathbb{E}\left[\lambda_{k+1}^s \mid \mathcal{F}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1\left(\mathfrak{F}\right), \quad \left(d_{\mathcal{C}}\gamma_{k+1}\mathbb{E}\left[\|\lambda_{k+1}\| \mid \mathcal{F}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1\left(\mathfrak{F}\right).$$

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

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

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

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

where $(z_k)_{k\in\mathbb{N}} \in \ell_+^1\left(\mathfrak{F}\right)$, 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

$$\left(\gamma_k\left\|A\tilde{x}_k-b\right\|^2\right)_{k\in\mathbb{N}} \in \ell_+^1\left(\mathfrak{F}\right), \quad \left(\gamma_k\left\|A\left(x_k-\tilde{x}_k\right)\right\|^2\right)_{k\in\mathbb{N}} \in \ell_+^1\left(\mathfrak{F}\right). \tag{4.8}$$

Consequently,

$$\left(\gamma_k\left\|Ax_k-b\right\|^2\right)_{k\in\mathbb{N}} \in \ell_+^1\left(\mathfrak{F}\right), \tag{4.9}$$

since by the Cauchy-Schwarz inequality,

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

To finish proving (i) we simply apply Lemma 4.3 (with the remark which follows) and the conditions of Lemma 2.3 are satisfied. Then, (ii) and (iii) follow directly from the results of [16, Theorem 4.1]. □

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 [16] and trivially 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. Then the sequence of dual iterates $(\mu_k)_{k\in\mathbb{N}}$ generated by* Algorithm 1 *is bounded.*

Proof. See [16, Lemma 4.9]. □

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 $\mathcal{C}$ and we have*

$$\tilde{M} \overset{\text{def}}{=} \sup_{x \in \mathcal{C}} |f(x) + g(Tx) + h(x)| + \sup_{k \in \mathbb{N}} \|\mu_k\| (\|A\| R + b) < +\infty, \tag{4.10}$$

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

Proof. The proof follows directly from [16, Lemma 4.10] with the addition of Theorem 4.7. □

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 $k \in \mathbb{N}$,*

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

and

$$C_k \overset{\text{def}}{=} \frac{L_k}{2} d_{\mathcal{C}}^2 + d_{\mathcal{C}} (M\|T\| + D + L_h + \|\mu^\star\| \, \|A\|).$$

Then, under (A.1)-(A.8) and (P.1)-(P.7), for the sequences $(x_k)_{k \in \mathbb{N}}$ and $(\mu_k)_{k \in \mathbb{N}}$ generated by Algorithm 1, using the filtration $\mathfrak{F} = (\mathcal{F}_k)_{k \in \mathbb{N}}$ with $\mathcal{F}_k = \mathcal{S}_{k-1}$, the following inequality holds, for each $k \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}[r_{k+1} \mid \mathcal{F}_k] - r_k \leq & -\gamma_k \left(\mathcal{L}(x_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) + \frac{\rho_k}{2} \|Ax_k - b\|^2 \right) + \frac{\gamma_{k+1}}{2} \mathbb{E}\left[\|Ax_{k+1} - b\|^2 \mid \mathcal{F}_k \right] \\
& + (\beta_k - \beta_{k+1}) \frac{M^2}{2} + (\gamma_k - \gamma_{k+1}) \tilde{M} + \gamma_k \beta_k \frac{M^2}{2} + K_{(F,\zeta,\mathcal{C})} \zeta(\gamma_k) + \gamma_k^2 C_k \\
& + d_{\mathcal{C}} \gamma_k \mathbb{E}[\|\lambda_k\| \mid \mathcal{F}_k] + \gamma_k \mathbb{E}[\lambda_k^s \mid \mathcal{F}_k] \ (\mathbb{P}\text{-a.s.}) \ .
\end{aligned} \tag{4.11}$$

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

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

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

$$\begin{aligned}
\mathcal{E}_k(x^\star, \mu_k) &\geq \mathcal{E}_k(x_k, \mu_k) + \left\langle \widehat{\nabla_x \mathcal{E}_k}(x_k, \mu_k), s_k - x_k \right\rangle + h(s_k) - h(x^\star) + \langle \lambda_k, x_k - x^\star \rangle \\
&\quad + \frac{\rho_k}{2} \|A(x^\star - x_k)\|^2 \\
&\geq \mathcal{E}_k(x_k, \mu_k) + \left\langle \widehat{\nabla_x \mathcal{E}_k}(x_k, \mu_k), \widehat{s}_k - x_k \right\rangle + h(\widehat{s}_k) - \lambda_k^s - h(x^\star) + \langle \lambda_k, x_k - x^\star \rangle \\
&\quad + \frac{\rho_k}{2} \|A(x^\star - x_k)\|^2 .
\end{aligned} \tag{4.12}$$

From Lemma 4.1 applied to the points x_{k+1} and x_k and by definition of $x_{k+1} \overset{\text{def}}{=} x_k + \gamma_k(\widehat{s}_k - x_k)$ 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 (4.12), to get, for each $k \in \mathbb{N}$,

$$\begin{aligned}
\mathcal{E}_k(x_{k+1}, \mu_k) \leq \mathcal{E}_k(x_k, \mu_k) &+ \gamma_k \langle \lambda_k, x^\star - \widehat{s}_k \rangle + D_F(x_{k+1}, x_k) + \gamma_k^2 \frac{L_k}{2}\|\widehat{s}_k - x_k\|^2 \\
&+ \gamma_k \left(\mathcal{E}_k(x^\star, \mu_k) + h(x^\star) - \mathcal{E}_k(x_k, \mu_k) - h(\widehat{s}_k) - \frac{\rho_k}{2}\|Ax_k - b\|^2 + \lambda_k^s \right).
\end{aligned} \tag{4.13}$$

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

$$\begin{aligned}
\mathcal{L}_k(x_{k+1}, \mu_k) - \mathcal{L}_k(x_k, \mu_k) &= \mathcal{E}_k(x_{k+1}, \mu_k) - \mathcal{E}_k(x_k, \mu_k) + h(x_{k+1}) - h(x_k) \\
&\leq \mathcal{E}_k(x_{k+1}, \mu_k) - \mathcal{E}_k(x_k, \mu_k) + \gamma_k (h(\widehat{s}_k) - h(x_k))
\end{aligned} \tag{4.14}$$

Combining (4.13) and (4.14), we obtain, for each $k \in \mathbb{N}$,

$$\begin{aligned}
\mathcal{L}_k(x_{k+1}, \mu_k) - \mathcal{L}_k(x_k, \mu_k) &\leq \gamma_k (\mathcal{E}_k(x^\star, \mu_k) + h(x^\star) - \mathcal{E}_k(x_k, \mu_k) - h(x_k)) + D_F(x_{k+1}, x_k) + \\
&\quad \gamma_k^2 \frac{L_k}{2}\|\widehat{s}_k - x_k\|^2 + \gamma_k \left(\langle \lambda_k, x^\star - \widehat{s}_k \rangle - \frac{\rho_k}{2}\|Ax_k - b\|^2 + \lambda_k^s \right) \\
&= \gamma_k (\mathcal{L}_k(x^\star, \mu_k) - \mathcal{L}_k(x_k, \mu_k)) + D_F(x_{k+1}, x_k) + \gamma_k^2 \frac{L_k}{2}\|\widehat{s}_k - x_k\|^2 \\
&\quad + \gamma_k \left(\langle \lambda_k, x^\star - \widehat{s}_k \rangle - \frac{\rho_k}{2}\|Ax_k - b\|^2 + \lambda_k^s \right)
\end{aligned} \tag{4.15}$$

Recalling the definition of $\mu_{k+1} \overset{\text{def}}{=} \mu_k + A(x_{k+1} - b)$ 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 (4.15) to get, for each $k \in \mathbb{N}$,

$$\begin{aligned}
\mathcal{L}_k(x_{k+1}, \mu_{k+1}) - \mathcal{L}_k(x_k, \mu_k) &\leq \theta_k \|Ax_{k+1} - b\|^2 + \gamma_k (\mathcal{L}_k(x^\star, \mu_k) - \mathcal{L}_k(x_k, \mu_k)) + D_F(x_{k+1}, x_k) \\
&\quad + \gamma_k^2 \frac{L_k}{2}\|\widehat{s}_k - x_k\|^2 + \gamma_k \left(\langle \lambda_k, x^\star - \widehat{s}_k \rangle - \frac{\rho_k}{2}\|Ax_k - b\|^2 + \lambda_k^s \right).
\end{aligned} \tag{4.16}$$

Notice that the update of the dual variable μ can be interpreted as a prox 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 2.4, we get, for each $k \in \mathbb{N}$,

$$\begin{aligned} 0 &\geq \theta_k \left(\mathcal{L}_k\left(x_{k+1}, \mu^\star\right) - \mathcal{L}_k\left(x_{k+1}, \mu_{k+1}\right)\right) + \frac{1}{2}\left(\left\|\mu_{k+1} - \mu^\star\right\|^2 - \left\|\mu_k - \mu^\star\right\|^2 + \left\|\mu_{k+1} - \mu_k\right\|^2\right) \\ &= \theta_k \left(\mathcal{L}_k\left(x_{k+1}, \mu^\star\right) - \mathcal{L}_k\left(x_{k+1}, \mu_{k+1}\right)\right) + \frac{1}{2}\left(\left\|\mu_{k+1} - \mu^\star\right\|^2 - \left\|\mu_k - \mu^\star\right\|^2 + \theta_k^2 \left\|Ax_{k+1} - b\right\|^2\right). \end{aligned} \tag{4.17}$$

Recall that, by (P.6), $\theta_k = \gamma_k / c$. Multiply (4.17) by c and sum with (4.16), to obtain, for each $k \in \mathbb{N}$,

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

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

$$\begin{aligned} &(1 - c\theta_{k+1})\mathcal{L}_{k+1}\left(x_{k+1}, \mu_{k+1}\right) - (1 - c\theta_k)\mathcal{L}_k\left(x_k, \mu_k\right) + \frac{c}{2}\left(\left\|\mu_{k+1} - \mu^\star\right\|^2 - \left\|\mu_k - \mu^\star\right\|^2\right) \\ \leq\ & (1 - c\theta_{k+1})\mathcal{L}_{k+1}\left(x_{k+1}, \mu_{k+1}\right) - (1 - c\theta_k)\mathcal{L}_k\left(x_{k+1}, \mu_{k+1}\right) + \left(\theta_k - \frac{c\theta_k^2}{2}\right)\left\|Ax_{k+1} - b\right\|^2 \\ & + \gamma_k\left(\mathcal{L}_k\left(x^\star, \mu_k\right) - \mathcal{L}_k\left(x_k, \mu_k\right)\right) - c\theta_k\left(\mathcal{L}_k\left(x_{k+1}, \mu^\star\right) - \mathcal{L}_k\left(x_k, \mu_k\right)\right) - \frac{\rho_k\gamma_k}{2}\left\|Ax_k - b\right\|^2 \\ & + D_F\left(x_{k+1}, x_k\right) + \gamma_k^2\frac{L_k}{2}\left\|\widehat{s}_k - x_k\right\|^2 + \gamma_k\left(\left\langle\lambda_k, x^\star - \widehat{s}_k\right\rangle + \lambda_k^s\right) \\ =\ & c\left(\theta_k - \theta_{k+1}\right)\left(f + h + \left\langle\mu_{k+1},\ A\cdot - b\right\rangle\right)\left(x_{k+1}\right) + \left(\left(1 - c\theta_{k+1}\right)g^{\beta_{k+1}} - \left(1 - c\theta_k\right)g^{\beta_k}\right)\left(Tx_{k+1}\right) \\ & + \frac{1}{2}\left(\left(1 - c\theta_{k+1}\right)\rho_{k+1} - \left(1 - c\theta_k\right)\rho_k + 2\theta_k - c\theta_k^2\right)\left\|Ax_{k+1} - b\right\|^2 + \gamma_k\left(\mathcal{L}_k\left(x^\star, \mu_k\right) - \mathcal{L}_k\left(x_k, \mu_k\right)\right) \\ & - c\theta_k\left(\mathcal{L}_k\left(x_{k+1}, \mu^\star\right) - \mathcal{L}_k\left(x_k, \mu_k\right)\right) - \frac{\rho_k\gamma_k}{2}\left\|Ax_k - b\right\|^2 + D_F\left(x_{k+1}, x_k\right) + \gamma_k^2\frac{L_k}{2}\left\|\widehat{s}_k - x_k\right\|^2 \\ & + \gamma_k\left(\left\langle\lambda_k, x^\star - \widehat{s}_k\right\rangle + \lambda_k^s\right). \end{aligned} \tag{4.18}$$

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 2.5(iv) and assumption (A.4), we are able to, for each $k \in \mathbb{N}$, estimate the quantity

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

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

$$\begin{aligned} & c\left(\theta_k-\theta_{k+1}\right)\left(f+h+\left\langle\mu_{k+1},\ A\cdot-b\right\rangle\right)\left(x_{k+1}\right)+\left(\left(1-c\theta_{k+1}\right)g^{\beta_{k+1}}-\left(1-c\theta_k\right)g^{\beta_k}\right)\left(Tx_{k+1}\right) \\ \leq\ & c\left(\theta_k-\theta_{k+1}\right)\mathcal{L}\left(x_{k+1},\mu_{k+1}\right)+\frac{1}{2}\left(\beta_k-\beta_{k+1}\right)\left\|\left(\partial g(Tx_{k+1})\right)^0\right\|^2. \end{aligned} \tag{4.19}$$

Recall the definition of r_k in (4.10). Coming back to (4.18) and using (4.19), we obtain, for each $k \in \mathbb{N}$,

$$\begin{aligned} r_{k+1}-r_k\ \leq\ & \frac{1}{2}\left(\left(1-\gamma_{k+1}\right)\rho_{k+1}-\left(1-\gamma_k\right)\rho_k+\frac{2}{c}\gamma_k-\frac{\gamma_k^2}{c}\right)\left\|Ax_{k+1}-b\right\|^2+\gamma_k\left(\mathcal{L}_k\left(x^\star,\mu_k\right)-\mathcal{L}_k\left(x_{k+1},\mu^\star\right)\right) \\ & -\frac{\rho_k\gamma_k}{2}\left\|Ax_k-b\right\|^2+\frac{\beta_k-\beta_{k+1}}{2}\left\|\left(\partial g(Tx_{k+1})\right)^0\right\|^2+\left(\gamma_k-\gamma_{k+1}\right)\mathcal{L}\left(x_{k+1},\mu_{k+1}\right) \\ & +D_F\left(x_{k+1},x_k\right)+\gamma_k^2\frac{L_k}{2}\left\|\widehat{s}_k-x_k\right\|^2+\gamma_k\left(\left\langle\lambda_k,x^\star-\widehat{s}_k\right\rangle+\lambda_k^s\right). \end{aligned} \tag{4.20}$$

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

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

where in the inequality we have used the fact that $g^{\beta_k} \leq g$ pointwise and that, by Proposition 2.5(v), for each $k \in \mathbb{N}$,

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

Substituting the above into (4.20) we have, for each $k \in \mathbb{N}$,

$$\begin{aligned} r_{k+1}-r_k\ \leq\ & \frac{1}{2}\left(\left(1-\gamma_{k+1}\right)\rho_{k+1}-\rho_k+\frac{2}{c}\gamma_k-\frac{\gamma_k^2}{c}\right)\left\|Ax_{k+1}-b\right\|^2 \\ & +\gamma_k\left(\mathcal{L}\left(x^\star,\mu^\star\right)-\mathcal{L}\left(x_k,\mu^\star\right)\right)+\gamma_k\left(\mathcal{L}\left(x_k,\mu^\star\right)-\mathcal{L}\left(x_{k+1},\mu^\star\right)\right) \\ & -\frac{\rho_k\gamma_k}{2}\left\|Ax_k-b\right\|^2+\frac{\beta_k-\beta_{k+1}}{2}\left\|\left(\partial g(Tx_{k+1})\right)^0\right\|^2+\left(\gamma_k-\gamma_{k+1}\right)\mathcal{L}\left(x_{k+1},\mu_{k+1}\right) \\ & +\gamma_k\frac{\beta_k}{2}\left\|\left(\partial g(Tx_{k+1})\right)^0\right\|^2+D_F\left(x_{k+1},x_k\right)+\gamma_k^2\frac{L_k}{2}\left\|\widehat{s}_k-x_k\right\|^2 \\ & +\gamma_k\left(\left\langle\lambda_k,x^\star-\widehat{s}_k\right\rangle+\lambda_k^s\right). \end{aligned} \tag{4.21}$$

Now we take the expectation with respect to $\mathcal{F}_k=\mathcal{S}_{k-1}=\sigma\left(x_0,\mu_0,\widehat{s}_0,\ldots,\widehat{s}_{k-1}\right)$, which will completely determine x_k and μ_k, and we are perform the following estimations.

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

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

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

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

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

$$\mathbb{E}\left[\mathcal{L}\left(x_{k+1}, \mu_{k+1}\right) \mid \mathcal{F}_k\right] \leq \tilde{M}.$$

- Recall that, by (A.2), f is (F, ζ)-smooth and invoke Remark 2.7, to get, for each $k \in \mathbb{N}$,

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

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

$$\mathbb{E}\left[\|\widehat{s}_k - x_k\| \mid \mathcal{F}_k\right] \leq d_{\mathcal{C}}.$$

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

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

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

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

We conclude by trivial manipulations. □

We now proceed to prove the main theorem regarding optimality.

Theorem 4.11 (Optimality). *Suppose that (A.1)-(A.8) 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 $(x^\star, \mu^\star)$ 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:

$$\lim_{k\to\infty}\mathcal{L}\left(x_k, \mu^\star\right) = \mathcal{L}\left(x^\star, \mu^\star\right) \ (\mathbb{P}\text{-a.s.}) \ . \tag{4.22}$$

(ii) Every weak cluster point $\bar{x}$ of $(x_k)_{k\in\mathbb{N}}$ is a solution of the primal problem ($\mathscr{P}$), and $(\mu_k)_{k\in\mathbb{N}}$ converges weakly to $\bar{\mu}$ a solution of the dual problem ($\mathscr{D}$), i.e., $(\bar{x},\bar{\mu})$ is a saddle point of $\mathcal{L}$ ($\mathbb{P}$-a.s.).

(iii) Pointwise rate:

$$\begin{aligned}&\forall k\in\mathbb{N},\ \inf_{0\leq i\leq k}\mathcal{L}(x_i,\mu^\star)-\mathcal{L}(x^\star,\mu^\star)=O\left(\frac{1}{\Gamma_k}\right)\ (\mathbb{P}\text{-a.s.})\quad \text{and}\\ &\exists\ \text{a subsequence } \left(x_{k_j}\right)_{j\in\mathbb{N}}\ \text{s.t. } \forall j\in\mathbb{N},\mathcal{L}\left(x_{k_j+1},\mu^\star\right)-\mathcal{L}(x^\star,\mu^\star)\leq\frac{1}{\Gamma_{k_j}}\ (\mathbb{P}\text{-a.s.}).\end{aligned}\tag{4.23}$$

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

$$\mathcal{L}(\bar{x}_k,\mu^\star)-\mathcal{L}(x^\star,\mu^\star)=O\left(\frac{1}{\Gamma_k}\right)\ (\mathbb{P}\text{-a.s.}).\tag{4.24}$$

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

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

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 Lemma 4.10 we have, using the same notation, for each $k\in\mathbb{N}$,

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

Let, for each $k\in\mathbb{N}$, $a_k=\gamma_k\left(\mathcal{L}(x_k,\mu^\star)-\mathcal{L}(x^\star,\mu^\star)+\frac{\rho_k}{2}\|Ax_k-b\|^2\right)$ and denote what remains on the r.h.s. by z_k. Then, to apply Lemma 2.2, we must show $(z_k)_{k\in\mathbb{N}}\in\ell^1_+(\mathfrak{F})$. The first term, $\gamma_{k+1}\mathbb{E}\left[\|Ax_{k+1}-b\|^2\mid\mathfrak{F}_k\right]$, is in $\ell^1_+(\mathfrak{F})$ by 4.7. 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 $K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_k)$ are in ℓ^1_+ by (P.1). Recalling the definition of C_k, we have, for each $k\in\mathbb{N}$,

$$\begin{aligned}\gamma_k^2C_k&=\gamma_k^2\left(\frac{L_k}{2}d_{\mathcal{C}}^2+d_{\mathcal{C}}\left(M\|T\|+D+L_h+\|\mu^\star\|\,\|A\|\right)\right)\\&=\left(\frac{d_{\mathcal{C}}^2\|T\|^2}{2}\right)\frac{\gamma_k^2}{\beta_k}+\left(\frac{d_{\mathcal{C}}^2\|A\|^2\rho_k}{2}+d_{\mathcal{C}}\left(M\|T\|+D+L_h+\|\mu^\star\|\,\|A\|\right)\right)\gamma_k^2\\&\leq\left(\frac{d_{\mathcal{C}}^2\|T\|^2}{2}\right)\frac{\gamma_k^2}{\beta_k}+\left(\frac{d_{\mathcal{C}}^2\|A\|^2\overline{\rho}}{2}+d_{\mathcal{C}}\left(M\|T\|+D+L_h+\|\mu^\star\|\,\|A\|\right)\right)\gamma_k^2\end{aligned}$$

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

$$d_{\mathcal{C}}\gamma_k\mathbb{E}\left[\|\lambda_k\|\mid\mathcal{F}_k\right]+\gamma_k\mathbb{E}\left[\lambda_k^s\mid\mathcal{F}_k\right],$$

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

$$a_k = \gamma_k \left(\mathcal{L}(x_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) + \frac{\rho_k}{2} \|Ax_k - b\|^2 \right) \in \ell_+^1(\mathfrak{F}) \ (\mathbb{P}\text{-a.s.}) \ .$$

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 following three claims, (ii), (iii), and (iv), all follow from [16, Theorem 4.2]. The final claim, (v), follows from [16, Corollary 19]. □

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,

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

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:

(E.1) It holds $\nabla_x f(x) = \mathbb{E}\left[\nabla_x L(x, \eta) \right]$ ($\mathbb{P}$-a.e.)

(E.2) For all η, the function $L(\cdot, \eta)$ is ω-smooth (see Definition 2.8) 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.

Notice that (E.4) $\implies$ (E.3). For the sake of clarity, we analyze 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.

Remark 5.1. 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.

5.1 Risk minimization with increasing batch size

Consider ($\mathscr{P}_1$) 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 5.2. *Under assumptions (E.1) and (E.2), denote*

$$C = 2\left(\omega\left(d_{\mathcal{C}}\right)^2 + \mathbb{E}\left[\left\|\nabla L\left(x^\star, \eta\right)\right\|^2 \mid \mathcal{S}_k\right]\right)$$

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

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

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

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

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

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

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

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

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 $x^\star$ are both in $\mathcal{C}$ to find, for each $k \in \mathbb{N}$,

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

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

Proposition 5.3. *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,*

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

Proof. By Lemma 5.2 we have, for each $k \in \mathbb{N}$,

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

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

Remark 5.4. 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 ($\mathscr{P}_1$) 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

$$\widehat{\nabla f}_k \overset{\text{def}}{=} (1-\nu_k)\widehat{\nabla f}_{k-1} + \nu_k \nabla_x L(x_k, \eta_k) \tag{5.1}$$

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. In this way, we are able to take a single gradient sample (or a larger fixed batch size) at each iteration, in contrast to the previous subsection.

Lemma 5.5. *Under (E.1) and (E.3), denote, for each $k \in \mathbb{N}$,*

$$\sigma_k^2 \overset{\text{def}}{=} \mathbb{E}\left[\|\nabla_x L(x_k, \eta_k) - \nabla f(x_k)\|^2 \mid \mathcal{S}_k\right] \tag{5.2}$$

and assume that $\exists \sigma > 0$ such that $\sup_k \sigma_k^2 = \sigma^2 < \infty$. Then, for each $k \in \mathbb{N}$, the following inequality holds,

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

Proof. The proof of this theorem is inspired by a similar construction found in [11, 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 pythagoreas identity then gives,

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

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

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

Notice that $1-\nu_{k+1} \le 1$ and thus $(1-\nu_{k+1})^2 \le 1-\nu_{k+1}$ for all $k \in \mathbb{N}$. This leads to

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

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

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

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

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

to find,

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

□

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 $\omega : t \to C_f t^\tau$.

Proposition 5.6. *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,*

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

and, for k sufficiently large,

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

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

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

Proof. Since (E.4) $\implies$ (E.3), the assumptions (E.1) and (E.3) are satisfied and Lemma 5.5 gives, for all $k \in \mathbb{N}$,

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

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

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

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

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

Since $\alpha < \tau \leq 1$ by 5.3, it holds that $\alpha < \min\{1, 2\tau-\alpha\}$, and the first condition of Lemma 9.1 is satisfied. Additionally, by (5.4), we have that the second condition, (9.2), of Lemma 9.1 is satisfied as well and we can apply Lemma 9.1 with

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

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,

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

Then, for k sufficiently large,

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

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

Example 5.7. The condition (5.3) in Proposition 5.6 can be satisfied, for example, by taking $\gamma_k = \frac{1}{(k+1)^{1-b}}$. 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,

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

The condition (5.4) in Proposition 5.6 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.

Recall that the predicted convergence rates for the ergodic iterates $\bar{x}_k$ given by Theorem 4.7 and Theorem 4.11 under this choice of step size are,

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

where $\Gamma_k = \sum_{i=0}^{k}\gamma_i = \sum_{i=0}^{k}\frac{1}{(i+1)^{1-b}}$. 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 $\frac{2}{3}\tau$. If the problem is Lipschitz-smooth, then $\tau = 1$ and we get $\alpha = \frac{2}{3}$.

Notice that the choice of α does not directly affect the predicted rates of convergence, which now depend only on the constant b. However, the choice of α dictates the possible choices for b which satisfy the assumptions and thus, indirectly, the rates of convergence as well. In the Lipschitz-smooth case, choosing $\alpha = \frac{2}{3}$ leads one to pick $b < 1 - (4/3)^{-1} = \frac{1}{4}$

6 Sweeping

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

$$\min_{\substack{x\in\mathcal{C}\subset\mathcal{H}\\ Ax=b}} \frac{1}{n}\sum_{i=1}^{n} f_i(x) \tag{$\mathscr{P}_2$}$$

where $n > 1$ is fixed. We assume that:
(F.1) f_i is ω-smooth (see Definition 2.8) 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 5.1). 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 $mod(k) \overset{\text{def}}{=} (k \mod n)$ with the convention that $mod(n) \overset{\text{def}}{=} n$. We define the inexact gradient in the following way,

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

and

$$\widehat{\nabla f}_k \overset{\text{def}}{=} \widehat{\nabla f}_{k-1} + \frac{1}{n}\left(\nabla f_{mod(k)}(x_k) - \nabla f_{mod(k)}(x_{k-n})\right) \quad (\forall k \geq n+1).$$

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

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

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

$$\left\|\lambda_{k+1}^f\right\| \leq C\omega(\gamma_{k+2-2n} d_{\mathcal{C}}).$$

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

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

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

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

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

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

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

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

□

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

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

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

$$\gamma_{k+1}\left\|\lambda_{k+1}^f\right\| \leq C\gamma_{k+1}\omega(d_{\mathcal{C}}\gamma_{k+2-2n}) \leq C\gamma_{k+2-2n}\omega(d_{\mathcal{C}}\gamma_{k+2-2n})$$

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

7 Numerical Experiments

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

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

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 $\{1,\ldots,n\}$ 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 $b \in \left\{\frac{1}{4} - 0.15, \frac{1}{3} - 0.01\right\}$. The weight ν_k in the variance reduction is chosen to be $\nu_k = \gamma_k^\alpha$ 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 5.6 is satisfied as was discussed in Example 5.7.

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

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

We initialize $y \in \mathbb{R}^n$ and $A \in \mathbb{R}^{2\times n}$ randomly. To find the solution $x^\star$ 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 including the gradient of the quadratic term is sampled (see Remark 5.1), for two different step size and weight choices as well as different batch sizes (1, 64, or 256), displayed in Figures 1 and 2.

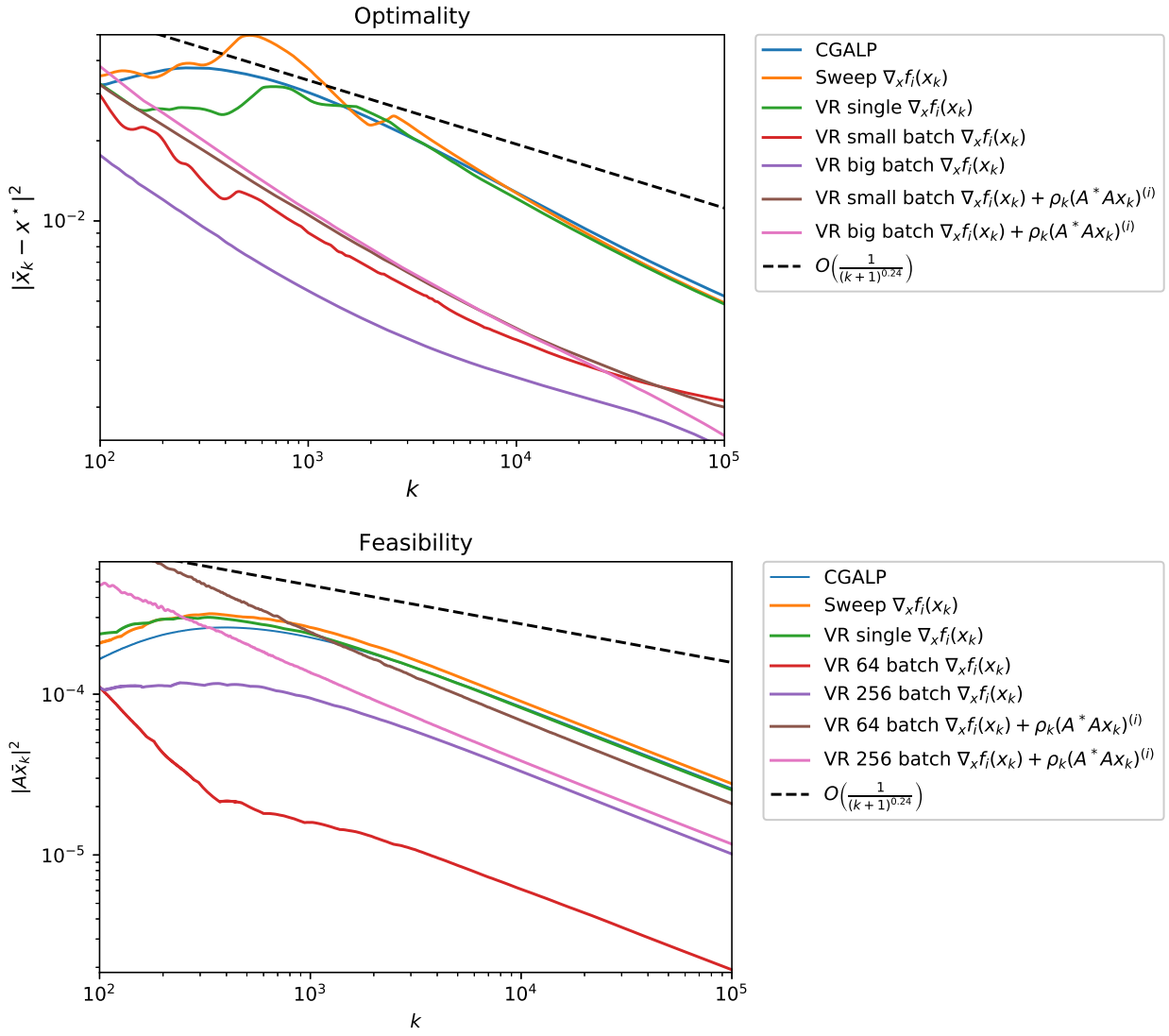

Figure 1: Ergodic convergence profiles for ICGALP applied to the projection problem (7.1) with $n = 1024$. The step size is, for each $k \in \mathbb{N}$, $\gamma_k = (k+1)^{-\left(1-\frac{1}{4}+0.01\right)}$ and the weight for variance reduction is, for each $k \in \mathbb{N}$, $\nu_k = \gamma_k^{2/3}$.

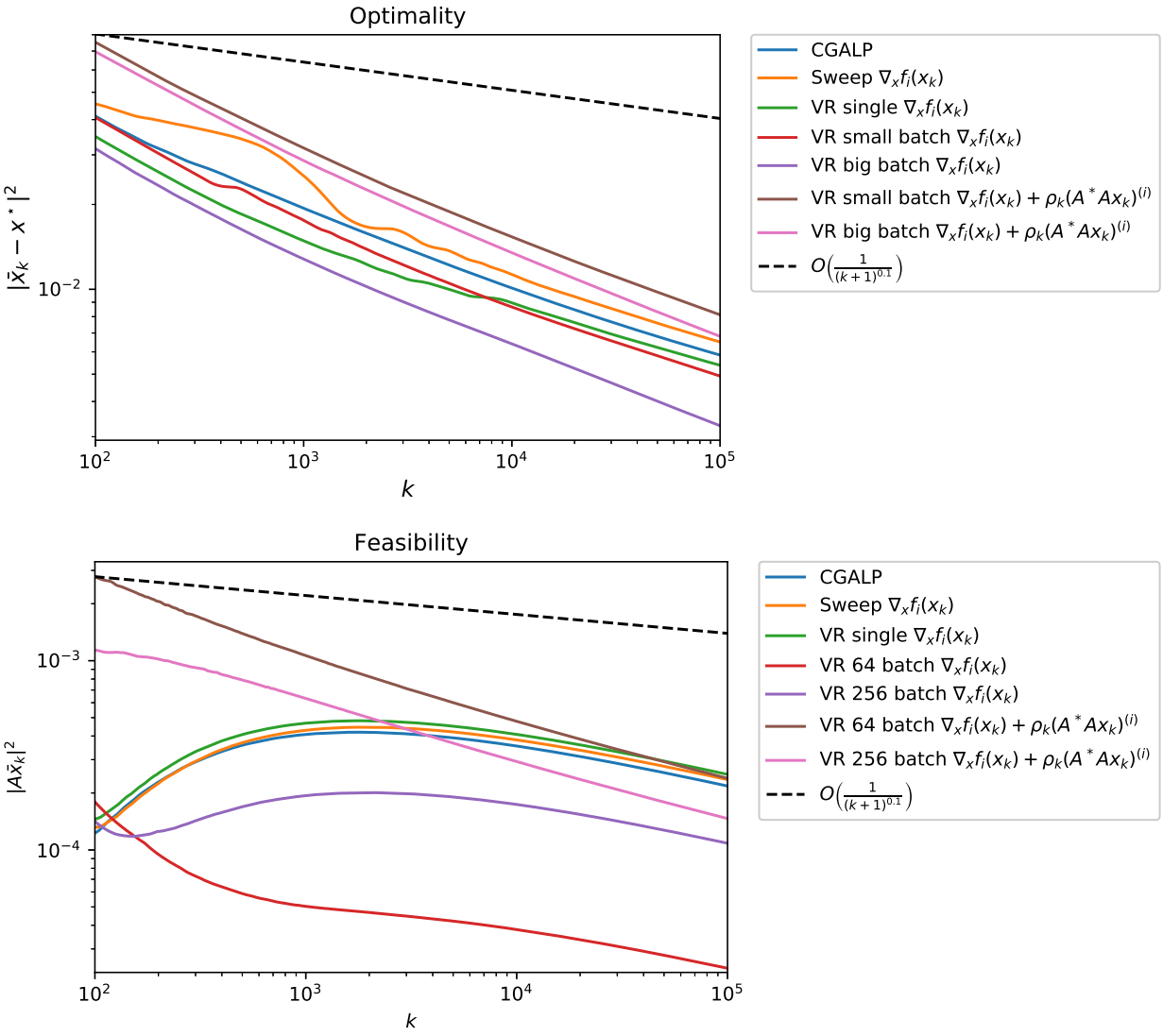

Figure 2: Ergodic convergence profiles for ICGALP applied to the projection problem (7.1) with $n = 1024$. The step size is, for each $k \in \mathbb{N}$, $\gamma_k = (k+1)^{-\left(1-\frac{1}{4}+0.15\right)}$ and the weight for variance reduction is, for each $k \in \mathbb{N}$, $\nu_k = \gamma_k^{2/3}$.

8 Conclusion

We introduced an inexact extension of the CGALP algorithm, given in [16], 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, $\operatorname{prox}_{\beta g}$, 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.

9 Appendix

Lemma 9.1. *Consider a positive sequence $(u_k)_{k\in\mathbb{N}}$ which satisfies, for each $k \in \mathbb{N}$,*

$$u_{k+1} \leq (1 - c\gamma_k^s)\, u_k + d\gamma_k^t, \tag{9.1}$$

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 $k \in \mathbb{N}$,

$$\frac{\gamma_k}{\gamma_{k+1}} \leq 1 + o\left(\gamma_k^s\right), \tag{9.2}$$

then, for k sufficiently large, it holds,

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

Proof. For each $k \in \mathbb{N}$, we denote $\nu_k \stackrel{\text{def}}{=} \gamma_k^{s-t}u_k - \frac{d}{c}$ such that $u_k = \gamma_k^{t-s}\left(\nu_k + \frac{d}{c}\right)$. Then, by (9.1),

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

By (9.2), we then have, for each $k \in \mathbb{N}$,

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

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

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

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

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

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

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

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,

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

From this we conclude, by [13, Ch.2, Lemma 3], that $\limsup_k \nu_k \leq 0$. Thus, by definition of ν_k,

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

□

Acknowledgements

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.

References

[1] Kengy Barty, Jean-Sébastien Roy, and Cyrille Strugarek. Hilbert-valued perturbed subgradient algorithms. *Mathematics of Operations Research*, 32(3):551–562, 2007.

[2] Patrick L. Combettes and Jean-Christophe Pesquet. Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping ii: mean-square and linear convergence. *Mathematical Programming*, 174(1):433–451, Mar 2019.

[3] Lijun Ding and Madeleine Udell. Frank-wolfe style algorithms for large scale optimization. In Pontus Giselsson and Anders Rantzer, editors, *Large-Scale and Distributed Optimization*, pages 215–245. Springer International Publishing, Cham, 2018.

[4] Donald Goldfarb, Garud Iyengar, and Chaoxu Zhou. Linear Convergence of Stochastic Frank Wolfe Variants. *arXiv e-prints*, page arXiv:1703.07269, Mar 2017.

[5] Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Zebang Shen. Stochastic Conditional Gradient++. *arXiv e-prints*, page arXiv:1902.06992, Feb 2019.

[6] Elad Hazan and Satyen Kale. Projection–free online learning. In *ICML*, 2012.

[7] Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *ICML*, 2016.

[8] Francesco Locatello, Alp Yurtsever, Olivier Fercoq, and Volkan Cevher. Stochastic Conditional Gradient Method for Composite Convex Minimization. *arXiv e-prints*, page arXiv:1901.10348, Jan 2019.

[9] Haihao Lu and Robert M. Freund. Generalized stochastic frank-wolfe algorithm with stochastic substitute gradient for structured convex optimization. *Mathematical Programming*, 2020.

[10] Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic Conditional Gradient Methods: From Convex Minimization to Submodular Maximization. *arXiv e-prints*, page arXiv:1804.09554, Apr 2018.

[11] Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takac. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017.

[12] J. Peypouquet. *Convex optimization in normed spaces: theory, methods and examples*. Springer, 2015.

[13] B. T. Polyak. *Introduction to optimization*. Optimization Software, 1987.

[14] Sashank J. Reddi, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic Frank-Wolfe Methods for Nonconvex Optimization. *arXiv e-prints*, page arXiv:1607.08254, Jul 2016.

[15] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Herbert Robbins Selected Papers*, pages 111–135. Springer New York, 1985.

[16] Antonio Silveti-Falls, Cesare Molinari, and Jalal Fadili. Generalized conditional gradient with augmented lagrangian for composite minimization. *SIAM Journal on Optimization*, 2020. in press.

[17] Xiaohan Wei and Michael J. Neely. Primal-Dual Frank-Wolfe for Constrained Stochastic Programs with Convex and Non-convex Objectives. *arXiv e-prints*, page arXiv:1806.00709, Jun 2018.

[18] A. N. Iusem Ya. I. Alber and M. V. Solodov. On the projected subgradient method for nonsmooth convex optimization in a hilbert space. *Mathematical Programming*, 81(1):23–35, 1998.