First-order noneuclidean splitting methods for large-scale optimization : deterministic and stochastic algorithms

Antonio Silveti Falls

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité MATHEMATIQUES

Préparée au sein de l'Université de Caen Normandie

First-Order Noneuclidean Splitting Methods for Large-Scale Optimization: Deterministic and Stochastic Algorithms

Présentée et soutenue par
Antonio SILVETI FALLS

Thèse soutenue publiquement le 11/02/2021 devant le jury composé de		
M. AMIR BECK	Professeur des universités, Tel-Aviv University	Rapporteur du jury
Mme SILVIA VILLA	Professeur des universités, Université de Genova - Italie	Rapporteur du jury
M. JÉRÔME BOLTE	Professeur des universités, Université Toulouse 1 Capitole	Membre du jury
Mme EMILIE CHOUZENOUX	Chargé de recherche HDR, INRIA Paris	Membre du jury
M. ALEXANDRE D'ASPREMONT	Directeur de recherche au CNRS, ENS-PSL	Membre du jury
M. JALAL FADILI	Professeur des universités, ENSICAEN	Directeur de thèse
M. GABRIEL PEYRE	Directeur de recherche au CNRS, École Normale Supérieure de Paris	Co-directeur de thèse
M. ANTONIN CHAMBOLLE	Directeur de recherche au CNRS, Université Paris-Dauphine	Président du jury

Thèse dirigée par JALAL FADILI et GABRIEL PEYRE, Groupe de recherche en informatique, image, automatique et instrumentation

UNIVERSITÉ CAEN NORMANDIE

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

MIIS

GREYC

Abstract

In this work we develop and examine two novel first-order splitting algorithms for solving large-scale composite optimization problems in infinite-dimensional spaces. Such problems are ubiquitous in many areas of science and engineering, particularly in data science and imaging sciences. Our work is focused on relaxing the Lipschitz-smoothness assumptions generally required by first-order splitting algorithms by replacing the Euclidean energy with a Bregman divergence. These developments allow one to solve problems having more exotic geometry than that of the usual Euclidean setting. One algorithm is hybridization of the conditional gradient algorithm, making use of a linear minimization oracle at each iteration, with an augmented Lagrangian algorithm, allowing for affine constraints. The other algorithm is a primal-dual splitting algorithm incorporating Bregman divergences for computing the associated proximal operators. For both of these algorithms, our analysis shows convergence of the Lagrangian values, subsequential weak convergence of the iterates to solutions, and rates of convergence. In addition to these novel deterministic algorithms, we introduce and study also the stochastic extensions of these algorithms through a perturbation perspective. Our results in this part include almost sure convergence results for all the same quantities as in the deterministic setting, with rates as well. Finally, we tackle new problems that are only accessible through the relaxed assumptions our algorithms allow. We demonstrate numerical efficiency and verify our theoretical results on problems like low rank, sparse matrix completion, inverse problems on the simplex, and entropically regularized Wasserstein inverse problems.

Keywords: nonsmooth optimization, first-order optimization, primal-dual splitting, conditional gradient, frank-wolfe, chambolle-pock, Bregman divergence, mirror descent, Moreau envelope, relative smoothness, Kullback-Leibler divergence, Wasserstein barycenter, Wasserstein inverse problem.

Résumé

Dans ce travail, nous développons et examinons deux nouveaux algorithmes d'éclatement du premier ordre pour résoudre des problèmes d'optimisation composites à grande échelle dans des espaces à dimensions infinies. Ces problèmes sont au coeur de nombres de domaines scientifiques et d'ingénierie, en particulier la science des données et l'imagerie. Notre travail est axé sur l'assouplissement des hypothèses de régularité de Lipschitz généralement requises par les algorithmes de fractionnement du premier ordre en remplaçant l'énergie euclidienne par une divergence de Bregman. Ces développements permettent de résoudre des problèmes ayant une géométrie plus exotique que celle du cadre euclidien habituel. Un des algorithmes développés est l'hybridation de l'algorithme de gradient conditionnel, utilisant un oracle de minimisation linéaire à chaque itération, avec méthode du Lagrangien augmenté, permettant ainsi la prise en compte de contraintes affines. L'autre algorithme est un schéma d'éclatement primal-dual incorporant les divergences de Bregman pour le calcul des opérateurs proximaux associés. Pour ces deux algorithmes, nous montrons la convergence des valeurs Lagrangiennes, la convergence faible des itérés vers les solutions ainsi que les taux de convergence. En plus de ces nouveaux algorithmes déterministes, nous introduisons et étudions également leurs extensions stochastiques au travers d'un point de vue d'analyse de stablité aux perturbations. Nos résultats dans cette partie comprennent des résultats de convergence presque sûre pour les mêmes quantités que dans le cadre déterministe, avec des taux de convergence également. Enfin, nous abordons de nouveaux problèmes qui ne sont accessibles qu'à travers les hypothèses relâchées que nos algorithmes permettent. Nous démontrons l'efficacité numérique et illustrons nos résultats théoriques sur des problèmes comme la complétion de matrice parcimonieuse de rang faible, les problèmes inverses sur le simplexe, ou encore les problèmes inverses impliquant la distance de Wasserstein régularisée.

Mots-clés: optimisation non-lisse, optimisation du premier ordre, dédoublement primal-dual, gradient conditionnel, divergence de Bregman, descente en miroir, enveloppe de Moreau, lisse relative, divergence Kullback-Leibler, barycentre de Wasserstein, problème inverse de Wasserstein.

Table of contents

Chapter 1

Introduction

Contents

1.1 Context

The field of convex optimization is ubiquitous in science and applied mathematics. From signal processing to machine learning to operations research, the flexibility and utility offered through casting problems as convex optimization problems is well established. While many methods exist to solve convex optimization problems, first-order methods stand out when the problems at hand are extremely large and require only moderately precise solutions; often the case in imaging sciences like computer vision or machine learning where data is collected, and expected to be processed, at a huge scale. First-order methods generally scale well with the problem dimension, in contrast to second (or higher) order methods in which not even a single iteration can be performed because of storage constraints.

In practice, it is common to run into composite problems in which the objective function to be minimized is a sum of two or more functions, perhaps also composed with a linear operator. These problems usually admit some structure than can be exploited, e.g., differentiability, strong convexity, prox-friendliness[1], etc, but often the usefulness of this structure is impeded by the lack of separability in the objective. By this we mean that, even if f is differentiable, it's not necessarily true that $f + g$ is differentiable. Similarly, the proximal operator

[1] By prox-friendly, we mean a function whose proximal operator is computable in closed form or in another computationally accessible way.

of g, denoted prox_g and defined formally for some real Hilbert space $\mathcal{H}$,

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

may be accessible but not necessarily $\operatorname{prox}_{f+g}$ or $\operatorname{prox}_{g\circ T}$ for a bounded linear operator T. To really reap the benefits of this structure, it is essential to develop algorithms which separate or split the composite problem in a way that utilizes individually the structures present.

We describe in the following sections some prototypical first-order algorithms for solving different convex optimization problems with composite structure. We examine the ideas that lead one to these algorithms and give some history as well. Then we discuss briefly the concept of relative smoothness and stochastic algorithms. The purpose of this exposition is to set the stage for describing how the work in this thesis extends the landscape of problems solvable by first-order convex optimization methods. Throughout the rest of the section, we let $\mathcal{H}$ be a real Hilbert space and assume that f and g belong to the space $\Gamma_0(\mathcal{H})$ of convex, proper, lower semicontinuous functions from $\mathcal{H}$ to the extended real numbers $\mathbb{R}\cup\{+\infty\}$ (sometimes taking $\mathcal{H}=\mathbb{R}^n$).

1.1.1 Forward-Backward Splitting

A well-known example of a splitting algorithm is the forward-backward algorithm for solving problems of the form

$$\min_{x\in\mathcal{H}} f(x) + g(x) \tag{1.1.1}$$

where f is Lipschitz smooth,[2] and g is prox-friendly.

Algorithm 1: Forward-Backward Splitting

Input: x_0, γ

$k = 0$

repeat

$x_{k+1} = \operatorname*{argmin}_{x\in\mathcal{H}} \left\{ g(x) + \langle \nabla f(x_k), x\rangle + \frac{1}{2\gamma}\|x-x_k\|^2 \right\}$

$k \leftarrow k+1$

until *convergence*;

Output: x_{k+1}.

The updates of the algorithm allow one to split the forward (gradient) step, involving $\gamma\nabla f$, from the backward (proximal) step, involving $\operatorname{prox}_{\gamma g}$. A particular case of (1.1.1) is constrained optimization over some convex closed set $\mathcal{C}$

$$\min_{x\in\mathcal{C}} f(x).$$

This is (1.1.1) with the particular choice

$$g(x) = \iota_{\mathcal{C}}(x) \stackrel{\text{def}}{=} \begin{cases} 0 & x\in\mathcal{C}, \\ +\infty & x\notin\mathcal{C}. \end{cases}$$

In this case, the proximal operator associated to g is the projection operator onto the set $\mathcal{C}$,

$$\operatorname{prox}_g(x) = \operatorname*{argmin}_{y\in\mathcal{C}} \left\{ \frac{1}{2}\|x-y\|^2 \right\}. \tag{1.1.2}$$

The convergence properties of Algorithm 1 have been the subject of numerous research works, starting in [76] and [91], and we avoid giving a complete history here; we refer the interested reader to [38].

[2] We call a function f Lipschitz-smooth if its gradient, ∇f, is Lipschitz-continuous.

1.1.2 Primal-Dual Splitting

Many convex optimization problems admit a dual problem, to be made more precise in what follows, which can be exploited to improve the efficiency or accessibility of a problem. Some problems, in their original form (henceforth referred to as the primal problem), are not practically accessible in the sense that applying first-order methods to the primal problem alone is considered practically impossible. A common issue in this setting is when one has a nonsmooth, prox-friendly function g which has been composed with a linear operator $T : \mathbb{R}^n \to \mathbb{R}^m$, for which the composition $g \circ T$ is not prox-friendly. By Moreau's decomposition (see [10, Theorem 14.3]), g being prox-friendly means that the Legendre-Fenchel conjugate, denoted g^* and to be defined more precisely later, is also prox-friendly. By considering a primal-dual formulation of the problem we are able to untangle $g \circ T$ and handle each separately. We demonstrate what we mean by primal-dual formulation with the following example. Consider the problem

$$\min_{x \in \mathbb{R}^n} f(x) + g(Tx)$$

where f is Lipschitz-smooth and g is prox-friendly. This problem is no longer solvable by the forward-backward algorithm in general because of the linear operator T, which complicates things even if the proximal operator associated to g is accessible. We can associate a primal-dual problem (through Lagrangian duality)

$$\min_{x \in \mathbb{R}^n} \max_{\mu \in \mathbb{R}^m} f(x) - g^*(\mu) + \langle Tx, \mu \rangle .$$

In this form, with the linear operator T now separated from the function g, we can utilize the prox-friendliness of g regardless of whether $g \circ T$ has an accessible proximal operator or not. Furthermore, if strong duality holds and we exchange the $\min$ and $\max$ operators, we arrive at the Fenchel-Rockafellar dual problem

$$\max_{\mu \in \mathbb{R}^m} -\left(f^*(-T^*\mu) - g^*(\mu)\right)$$

which can be much more computationally tractable if $m \ll n$.

Such primal-dual problems are interesting in their own right and first-order algorithms have been developed to solve them. A well-known example of the aforementioned primal-dual algorithms is the Chambolle-Pock algorithm of [29], whose main steps are outlined in Algorithm 2. Other variations of this algorithm, designed to tackle even more general primal-dual problem formulations than that of [29] or its followup [31], were proposed, for instance in [112], [36], and [35].

Algorithm 2: Primal-Dual Splitting

Input: x_0, μ_0, λ, ν

$k = 0$

repeat

$$x_{k+1} = \operatorname*{argmin}_x \left\{ f(x) + \langle x, T^*\mu_k \rangle + \tfrac{1}{2\lambda} \|x - x_k\|^2 \right\}$$

$$\mu_{k+1} = \operatorname*{argmin}_\mu \left\{ g^*(\mu) - \langle T(2x_{k+1} - x_k), \mu \rangle + \tfrac{1}{2\nu} \|\mu - \mu_k\|^2 \right\}$$

$k \leftarrow k + 1$

until *convergence*;

Output: x_{k+1}, μ_{k+1}.

1.1.3 Alternating Direction Method of Multipliers

The Alternating Direction Method of Multipliers algorithm, or ADMM for short, is a splitting algorithm most commonly used for composite optimization problems over a real Hilbert space $\mathcal{H}$ of the form

$$\min_{x \in \mathcal{H}} f(x) + g(Tx)$$

where f and g are possibly nonsmooth but both are prox-friendly. We can rewrite the problem using Lagrangian duality as

$$\min_{x\in\mathcal{H},y\in\mathcal{H}} \max_{\mu\in\mathcal{H}} f(x) + g(y) + \langle \mu, Tx - y\rangle$$

Algorithm 3: Alternating Direction Method of Multipliers

Input: x_0, y_0, z_0, γ

$k = 0$

repeat

$$x_{k+1} \in \operatorname*{argmin}_x \left\{ f(x) + g(y_k) + \langle Tx - y_k, z_k\rangle + \tfrac{\gamma}{2}\|Tx - y_k\|^2 \right\}$$

$$y_{k+1} = \operatorname*{argmin}_y \left\{ f(x_{k+1}) + g(y) + \langle Tx_{k+1} - y, z_k\rangle + \tfrac{\gamma}{2}\|Tx_{k+1} - y\|^2 \right\}$$

$$z_{k+1} = z_k + \gamma(Tx_{k+1} - y_{k+1})$$

$k \leftarrow k + 1$

until *convergence*;

Output: x_{k+1}.

Originally studied in the context of partial differential equations in [57], its usefulness as a splitting algorithm in optimization was later realized and applied to solve problems in various fields. The key to the algorithm is the additional quadratic term in the updates of x_{k+1} and y_{k+1}; such additions constitute an *augmentation* to the Lagrangian

$$\mathcal{L}(x, y, z) \overset{\text{def}}{=} f(x) + g(y) + \langle Tx - y, z\rangle$$

and thus algorithms making use of this idea are often categorized as augmented Lagrangian methods.

The advantage of ADMM is the splitting it provides by minimizing the augmented Lagrangian

$$L_\gamma(x, y, z) \overset{\text{def}}{=} f(x) + g(y) + \langle Tx - y, z\rangle + \frac{\gamma}{2}\|Tx - y\|^2$$

separately over both x and y which allows one to avoid computing the, often inaccessible, proximal operator associated to $f + g$. After this alternating minimization, the dual variable z is updated by gradient ascent on the augmented Lagrangian $\mathcal{L}_\gamma$.

The convergence guarantees of ADMM have been widely studied in the literature under various assumptions, see [50], [51], [34], [90], [20], [64], [107], [105] and the references therein.

1.1.4 Conditional Gradient Algorithm

In the 1950's Frank and Wolfe developed the so-called Frank-Wolfe algorithm in [53], also commonly referred to as the conditional gradient algorithm [75, 44, 49], for solving problems of the form

$$\min_{x\in\mathcal{C}\subset\mathbb{R}^n} f(x). \tag{1.1.3}$$

where $f \in \Gamma_0(\mathbb{R}^n)$, the set of convex, proper, lower semicontinuous, extended real-valued functions on $\mathbb{R}^n$, is a Lipschitz-smooth function and $\mathcal{C}$ is a compact convex set. At that time in history, the distinction between linear and nonlinear problems was felt to be very strong, although now the community focuses more on the distinction between convex and nonconvex. Because of the perspective of the time, an algorithm was sought to solve the assumed nonlinear (1.1.3) by means of linear subproblems. Thus the main idea of the Frank-Wolfe algorithm is to replace the objective function f with a linear model at each iteration and solve the resulting linear optimization problem; the solution to the linear model is used as a step direction and the next iterate is computed as a convex combination of the current iterate and the step direction. Notice that the problem (1.1.3)

Algorithm 4: Conditional Gradient Algorithm

Input: $x_0 \in \mathcal{C}$.

$k = 0$

repeat

$s_k \in \underset{s \in \mathcal{C}}{\operatorname{argmin}} \{\langle \nabla f(x_k), s \rangle\}$

$x_{k+1} = x_k - \frac{2}{k+1}(x_k - s_k)$

$k \leftarrow k + 1$

until *convergence*;

Output: x_{k+1}.

can be solved by the forward-backward algorithm in principle, and often in practice, by choosing $g(x) = \iota_{\mathcal{C}}$ where $\iota_{\mathcal{C}}(x)$ is the indicator function, enforcing that $x \in \mathcal{C}$.

At each iteration k, the conditional gradient algorithm requires the solution to the problem

$$s_k \in \underset{s \in \mathbb{R}^n}{\operatorname{argmin}} \{\langle \nabla f(x_k), s \rangle + \iota_{\mathcal{C}}(s)\}. \tag{1.1.4}$$

While proximal methods utilize a quadratic term $\frac{\gamma}{2}\|x_k - x\|^2$ in their updates (similarly to (1.1.2)), the conditional gradient requires only access to a linear minimization oracle. This difference can be very important in practice where computing projections and computing linear minimization oracles can have very different computational complexities, e.g., when the set $\mathcal{C}$ is the nuclear norm ball for matrix problems. In other words, there are problems where the $\mathcal{C}$ is more amenable to proximal operator type oracles and other problems where the set $\mathcal{C}$ is more amenable to linear minimization type oracles.

It's also worth mentioning that the conditional gradient framework, in contrast to classical proximal methods, doesn't depend on the Hilbertian inner product structure. It is possible to define linear minimization oracles, and by extension conditional gradient algorithms, in nonreflexive Banach spaces. The key ingredient is the directional derivative, which allows one to properly define a linear minimization oracle. However, we do not explore conditional gradient algorithms beyond Hilbert spaces in this work.

When this algorithm was first studied in [53], a Lipschitz-smoothness assumption was made in the argument to prove convergence of the gap $f(x_k) - f^\star$. It was not until much later in the thesis [68] where it was made more precise what was required of the function f to guarantee convergence; a so-called curvature constant was introduced,

$$\mathcal{K}_f \overset{\text{def}}{=} \sup_{\substack{x,s \in \mathcal{C}; \\ \gamma \in [0,1]; \\ y = x + \gamma(s-x)}} \left\{ \frac{2}{\gamma^2} \left(f(y) - f(x) - \langle \nabla f(x), y - x \rangle \right) \right\}.$$

The curvature constant being bounded was deemed to be at least as important as Lipschitz-smoothness of the function f to the convergence of the conditional gradient algorithm; Lipschitz-smoothness of f implies that the curvature constant is bounded and this is sufficient to show convergence [66].

1.2 Motivation

1.2.1 Hybridizing Conditional Gradient Algorithms with Proximal Algorithms

Conditional gradient algorithms have received a lot of attention in the modern era due to their effectiveness in fields with high-dimensional problems like machine learning and signal processing (without being exhaustive, see, e.g., [66, 15, 72, 60, 118, 85, 27]). Consider the following problem over a real Hilbert space $\mathcal{H}$,

$$\min_{Ax=b} f(x) + g(Tx) + h(x) \tag{1.2.1}$$

where f is differentiable but not necessarily Lipschitz-smooth, g is prox-friendly, h admits an accessible linear minimization oracle (usually $h(x) = \iota_{\mathcal{C}}(x)$)

$$\operatorname*{lmo}_{h}(z) \stackrel{\text{def}}{=} \operatorname*{argmin}_{x \in \mathcal{H}} \{\langle z, x\rangle + h(x)\},$$

and T and A are bounded linear operators. Our aim is to solve (1.2.1) by hybrizing an ADMM style update for with a conditional gradient update, splitting and using the individual structures present as efficiently as possible.

In the past, composite constrained problems like (1.2.1) have primarily been approached using proximal splitting methods, e.g., generalized forward-backward as developed in [98] or forward-Douglas-Rachford [83]. As was touched on before, such approaches require one to compute the proximal mapping associated to the function h. Alternatively, when the objective function satisfies some regularity conditions and when the constraint set is well behaved, one can forgo computing a proximal mapping, instead computing a linear minimization oracle. The computation of the proximal step can be prohibitively expensive; for example, when h is the indicator function of the nuclear norm ball, computing the proximal operator of h requires a full singular value decomposition while the linear minimization oracle over the nuclear norm ball requires only the leading singular vector to be computed ([67], [117]). Unfortunately, the regularity assumptions required by generalized conditional gradient style algorithms are too restrictive to apply to general problems like (1.2.1) due to the lack of Lipschitz-smoothness, the affine constraint $Ax = b$ and the nonsmooth function g.

The linear minimization oracle used in the conditional gradient algorithm requires that the objective function is differentiable. Many problems in practice are convex but nonsmooth; is it possible to reconcile this with the differentiability requirement in the conditional gradient algorithm? Indeed, the answer is yes as was shown in [5] by using the Moreau envelope. While this is an appealing solution to the issue of nonsmoothness, the arguments they present require domain qualification conditions that preclude their use when the domain of the nonsmooth function is poorly behaved in the sense that it does not contain the set $\mathcal{C}$. So, one cannot use this method to include, for instance, an affine constraint of the form $Ax = b$ unless it is trivially satisfied for all $x \in \mathcal{C}$. A parallel work to ours, [116] developed a method that uses the Moreau envelope for such affine constraints in a conditional gradient framework but their work is only for finite-dimensional spaces; we will compare our work with theirs in Chapter 3.

We have seen how augmented Lagrangian methods like ADMM are able to handle these sorts of affine constraints by introducing a quadratic penalty at each iteration. However, applying such methods with a function like $h = \iota_{\mathcal{C}}$ will rely on projection onto the set $\mathcal{C}$. The question, then, is if it is possible to combine such augmented Lagrangian methods with generalized conditional gradient methods to take advantage of the efficiency each method offers for each piece of the problem. This is the motivation for the development of the Conditional Gradient with Augmeneted Lagrangian and Proximal step; a desire to unify the augmented Lagrangian method and the conditional gradient method in such a way that both the linear minimization oracle and the quadratic penalties can be utilized to their fullest.

1.2.2 Generalizing to Relatively Smooth Functions

The notion of relative smoothness is key to the analysis of differentiable but not Lipschitz-smooth optimization problems. The earliest reference to this notion can be found in an economics paper [17] where it is used to address a problem in game theory involving fisher markets. Later on it was developed in [13] and then in [80], although first coined relative smoothness in [80]. This idea allows one to apply arguments involving descent lemmas which are normally relegated to Lipschitz-smooth problems. The concept has been extended, for instance to define relative Lipschitz-continuity in [78], in [79] for the stochastic generalized conditional gradient, and to define a generalized curvature constant for the generalized conditional gradient algorithm in [108]. The analogous idea of relative strong convexity, while noted before in [31], was not explored in detail.

Given the recent advances in relative smoothness in papers like [13] and [80], it is natural to wonder if it is possible to extend the primal-dual splitting or conditional gradient algorithm and prove convergence for more

general functions than the class of Lipschitz-smooth functions. There have been small steps in this direction, for instance in the conditional gradient case in the work [116] where it was remarked that their analysis could be carried out for Hölder-smooth functions but even there the analysis was not actually carried out and presented beyond the remark.

To this end, we consider the following problem

$$\min_{x\in\mathcal{C}_p\subset\mathcal{X}_p}\max_{\mu\in\mathcal{C}_d\subset\mathcal{X}_d}\left\{\mathcal{L}(x,\mu)\overset{\text{def}}{=}f(x)+g(x)+\langle Tx,\ \mu\rangle-h^*(\mu)-l^*(\mu)\right\} \tag{1.2.2}$$

with real reflexive Banach spaces $\mathcal{X}_p$ and $\mathcal{X}_d$, $f\in\Gamma_0(\mathcal{X}_p)$ and $h^*\in\Gamma_0(\mathcal{X}_d)$ relatively smooth, $T:\mathcal{X}_p\to\mathcal{X}_d$ a bounded linear operator, and with the prox of $g\in\Gamma_0(\mathcal{X}_p)$ and $l^*\in\Gamma_0(\mathcal{X}_d)$ computable with respect to the Bregman divergences induced by the entropies[3] which f and h^* are relatively smooth with respect to. We handle the constraints $\mathcal{C}_p$ and $\mathcal{C}_d$ implicitly by choice of the entropy we consider in the updates we outline later in Chapter 5.

The relative smoothness does present challenges for the analysis of a primal-dual splitting type of algorithm; which is, roughly speaking, forward-backward on both the primal and the dual in some sense. D-prox mappings were utilized successfully in [31] which developed a primal-dual algorithm for composite minimization, but relative smoothness was not used. There are many technical obstacles that appear when one begins to replace Lipschitz-smoothness assumptions with relative smoothness and Euclidean prox mappings with D-prox mappings, for instance lack of an Opial-esque lemma for Bregman divergences. Of note is the inability to consolidate duality pairing terms with Bregman divergences into a single term. Such a consolidation is possible with a norm due to tools relating duality pairings to norms like the Cauchy-Schwarz inequality and Young's inequality.

The authors of [31] were able to evade this obstacle by assuming that the entropies were strongly convex. When the entropy ϕ is strongly convex with respect to some norm, one can estimate the Bregman divergence D_ϕ with the norm that ϕ is strongly convex with respect to. Once we are back to estimations involving only norms and duality pairings, the usual tricks can be applied. Thus we seek to develop, without assuming strong convexity of the entropies, an assumption that attempts to quantify the relationship between the Bregman divergence and the duality pairing which appears in the analysis of our algorithm.

1.2.3 Allowing for Stochasticity

Up to this point, our discussion of convex optimization has been relegated to deterministic problems and algorithms. The incorporation of probabilistic uncertainty into convex optimization is known as stochastic convex optimization. The study of stochastic convex optimization has its roots in [101].

As has been touched on, large-scale problems are very common in imaging sciences. The scale can be so large that deterministic methods are no longer practically feasible; storing a single gradient can be too computationally expensive. It is necessary, then, to develop stochastic algorithms which require only practically feasible amounts of data at each iteration. For all of the algorithms we have discussed thus far, we have also developed and analyzed stochastic extensions which allow for inexact computations of the gradients (or sometimes the proximal operators) involved at each iteration. Our study of these algorithms is through a perturbation perspective; we consider at each iteration some noise which corrupts the computation of gradients or proximal operators in an additive sense.

Through this perspective, we are able to propose several different methods for approximating the gradient inexactly, either stochastically or deterministically, which ensure convergence of the algorithm. These methods allow different practical sampling routines depending on the problem and its regularity, making them efficient in practice.

[3] We abuse the term entropy here, and throughout the thesis, to mean the function ϕ which induces the Bregman divergence D_ϕ that we are considering.

1.3 Contribution

1.3.1 Generalized Conditional Gradient with Augmented Lagrangian and Proximal Step

Results Recall the setting of (1.1.3); we generalize the conditional gradient algorithm to handle problems fitting (1.2.1). This is general enough to include various composite optimization problems involving both smooth and nonsmooth terms, intersection of multiple constraint sets, and also affine constraints.

We develop and analyze a novel algorithm, which we call **C**onditional **G**radient with **A**ugmented **L**agrangian and **P**roximal step, to solve (1.2.1) which combines penalization for the nonsmooth function g (essentially the Moreau envelope) with the augmented Lagrangian method for the affine constraint $Ax = b$. In turn, this achieves full splitting of all the parts in the composite problem (1.2.1) by using the proximal mapping of g (assumed prox-friendly) and a linear oracle for h. We can recognize in Algorithm 5 $(\gamma_k)_{k\in\mathbb{N}}$ as the sequence of step sizes, $(\beta_k)_{k\in\mathbb{N}}$ a sequence of smoothing parameters, $(\theta_k)_{k\in\mathbb{N}}$ the sequence of dual step sizes, and $(\rho_k)_{k\in\mathbb{N}}$ a sequence of augmented Lagrangian parameters.

Algorithm 5: Conditional Gradient with Augmented Lagrangian and Proximal-step (CGALP)

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

$k = 0$

repeat

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

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

$$s_k \in \operatorname{argmin}_{s\in\mathcal{H}}\{h(s) + \langle z_k, s\rangle\}$$

$$x_{k+1} = x_k - \gamma_k(x_k - s_k)$$

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

$$k \leftarrow k + 1$$

until *convergence*;

Output: x_{k+1}.

Our analysis shows:

- The sequence of iterates is asymptotically feasible for the affine constraint.
- The sequence of dual variables is strongly convergent to a solution of the dual problem.
- The associated Lagrangian converges to optimality.
- Convergence rates for a family of sequences of step sizes and sequences of smoothing/penalization parameters which satisfy so-called "open loop" rules in the sense of [96] and [49]. This means that the allowable sequences of parameters do not depend on the iterates, in contrast to a "closed loop" rule, e.g. line search or other adaptive step sizes.
- (In the case where (1.2.1) admits a unique minimizer) Weak convergence of the whole sequence of primal iterates to the solution with a rate of convergence on $\|x_k - x\|^2$.

The type of theorem one can find in Chapter 3 takes the following form:

Theorem 1.3.1. *Suppose that [mild assumptions on the functions and parameters] hold. Let $(x_k)_{k\in\mathbb{N}}$ be the sequence of primal iterates generated by CGALP and $(x^\star, \mu^\star)$ a saddle-point pair for the Lagrangian. Then, the following holds*

(i) Asymptotic feasibility:

$$\lim_{k\to\infty} \|Ax_k - b\| = 0.$$

(ii) Convergence of the Lagrangian:

$$\lim_{k\to\infty} \mathcal{L}\left(x_k, \mu^\star\right) = \mathcal{L}\left(x^\star, \mu^\star\right).$$

(iii) Every weak cluster point $\bar{x}$ of $(x_k)_{k\in\mathbb{N}}$ is a solution of the primal problem, and $(\mu_k)_{k\in\mathbb{N}}$ converges strongly to $\bar{\mu}$ a solution of the dual problem, i.e., $(\bar{x}, \bar{\mu})$ is a saddle point of the Lagrangian $\mathcal{L}$.

(iv) Ergodic rate: for each $k \in \mathbb{N}$, let $\bar{x}_k \overset{\text{def}}{=} \sum_{i=0}^{k} \gamma_i x_{i+1}/\Gamma_k$ where γ_i is the stepsize of the CGALP algorithm at iteration i. Then

$$\|A\bar{x}_k - b\|^2 = O\left(\frac{1}{\Gamma_k}\right),$$
$$\mathcal{L}\left(\bar{x}_k, \mu^\star\right) - \mathcal{L}\left(x^\star, \mu^\star\right) = O\left(\frac{1}{\Gamma_k}\right).$$

The structure of (1.2.1) generalizes (1.1.3) in several ways. First, we allow for a possibly nonsmooth term g. Second, we consider h beyond the case of an indicator function where the linear oracle of the form

$$\operatorname{lmo}_h \overset{\text{def}}{=} \operatorname*{argmin}_{s\in\mathcal{H}} h\left(s\right) + \langle x, s\rangle \tag{1.3.1}$$

can be easily solved. This oracle is reminiscent of that in the generalized conditional gradient method [21, 22, 14, 8]. Third, the regularity assumptions on f are also greatly weakened to go far beyond the standard Lipschitz gradient case. Finally, handling an affine constraint in our problem means that our framework can be applied to the splitting of a wide range of composite optimization problems, through a product space technique, including those involving finitely many functions h_i and g_i, and, in particular, intersection of finitely many nonempty bounded closed convex sets; see Section 3.5.

Practical Applications The general form of (1.2.1) allows to solve many practical problems efficiently. As will be demonstrated in Chapter 3, one can apply CGALP to matrix completion problems involving a nuclear norm constraint. This constraint is typically used as the convex relaxation of a low rank constraint. Solving such problems with nuclear norm constraints using proximal methods requires one to project onto the nuclear norm ball. Such a projection requires a full singular value decomposition of the matrix and this is computationally infeasible for very large matrices. However, in contrast, CGALP requires only a linear minimization oracle over the nuclear norm ball and this can be computed in reasonable time.

For problems involving multiple constraint sets $\mathcal{C}_i$, CGALP only requires access each linear minimization oracle $\operatorname{lmo}_{\mathcal{C}_i}$ individually with requiring $\operatorname{lmo}_{\cap_i \mathcal{C}_i}$. This is also demonstrated in the matrix completion problem, in which there is an ℓ^1 norm constraint in addition to the nuclear norm ball constraint.

Prior Work A similar algorithm to CGALP was studied, unknown to the present authors, in [5] but it does not allow for an affine constraint $Ax = b$. While finalizing this work, we became aware of the recent work of [116], who independently developed a conditional gradient-based framework which allows one to solve composite optimization problems involving a Lipschitz-smooth function f and a nonsmooth function g,

$$\min_{x\in\mathcal{C}} \left\{f(x) + g\left(Tx\right)\right\}. \tag{1.3.2}$$

The main idea is to replace g with its Moreau envelope of index β_k at each iteration k, with the index parameter β_k going to 0. This is equivalent to partial minimization with a quadratic penalization term, as in our algorithm. Like our algorithm, that of [116] is able to handle problems involving both smooth and nonsmooth terms, intersection of multiple constraint sets and affine constraints, however their algorithms employ different methods for these situations. Our algorithm uses an augmented Lagrangian to handle the affine constraint while the conditional gradient framework treats the affine constraint as a nonsmooth term g and uses penalization to smooth the indicator function corresponding to the affine constraint. In particular circumstances, outlined in more detail in Section 3.4, our algorithms agree completely.

Another recent and parallel work to ours is that of [56], where the Frank-Wolfe via Augmented Lagrangian (FW-AL) is developed to approach the problem of minimizing a Lipschitz-smooth function over a convex, compact set with a linear constraint,

$$\min_{x\in\mathcal{C}} \{f(x) : \ Ax = 0\}. \tag{1.3.3}$$

The main idea of FW-AL is to use the augmented Lagrangian to handle the linear constraint and then apply the classical augmented Lagrangian algorithm, except that the marginal minimization on the primal variable that is usually performed is replaced by an inner loop of Frank-Wolfe. It turns out that the problem they consider is a particular case of (1.2.1), discussed in Section 3.4.

We summarize the place of our work among contemporary works with the following table. By arbitrary h, we mean an arbitrary $h \in \Gamma_0(\mathcal{H})$ which has a compact domain and admits an accessible linear minimization oracle.

	Smoothness of f?	Nonsmooth $g \circ T$?	Arbitrary h?	$\mathbb{R}^n$ or $\mathcal{H}$?	$Ax = b$?
Yurtsever et. al. [116]	Lipschitz-smooth	Moreau envelope	$h \equiv \iota_{\mathcal{C}}$	$\mathbb{R}^n$	Moreau envelope
Argyriou et. al. [5]	Lipschitz-smooth	Moreau envelope	h arbitrary	$\mathcal{H}$	**X**
Gidel et. al. [56]	Lipschitz-smooth	**X**	$h \equiv \iota_{\mathcal{C}}$	$\mathbb{R}^n$	augm. Lagrangian
This work	Relatively smooth	Moreau envelope	h arbitrary	$\mathcal{H}$	augm. Lagrangian

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

Results The primary contribution of this work is to analyze inexact and stochastic variants of the CGALP algorithm presented in [108] to address (1.2.1). We coin this algorithm **I**nexact **C**onditional **G**radient with **A**ugemented **L**agrangian and **P**roximal-step (ICGALP). We now have two error terms in Algorithm 6: $(\lambda_k)_{k\in\mathbb{N}}$ representing error in the computation of the grad terms and $(\lambda_k^s)_{k\in\mathbb{N}}$ representing error in the linear minimization oracle itself.

Algorithm 6: Inexact Conditional Gradient with Augmented Lagrangian and Proximal-step (ICGALP)

Input: $x_0 \in \mathcal{C} \overset{\text{def}}{=} \operatorname{dom}(h)$; $\mu_0 \in \operatorname{ran}(A)$; $(\gamma_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}$, $(\rho_k)_{k\in\mathbb{N}} \in \ell_+$.
$k = 0$
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}.

We show:

- Asymptotic feasibility of the primal iterates for the affine constraint ($\mathbb{P}$-a.s.).
- Convergence of the Lagrangian values at each iteration to an optimal value ($\mathbb{P}$-a.s.).

- Strong convergence of the sequence of dual iterates ($\mathbb{P}$-a.s.).
- Worst-case rates of convergence for the feasibility gap and the Lagrangian values in a ($\mathbb{P}$-a.s.) sense and also in expectation.

The rates of convergence for both the Lagrangian and the feasibility gap are given globally, i.e., for the entire sequence of iterates, in the ergodic sense where the Cesáro means are taken with respect to the primal step size, in an almost sure sense. We also show rates in expectation which hold pointwise but subsequentially. In the case where (1.2.1) admits a unique solution, we furthermore have that the sequence of primal iterates converges weakly to the solution almost surely. These results are established for a family of parameters satisfying abstract open loop conditions, i.e. sequences of parameters which do not depend on the iterates themselves. We exemplify the framework on problem instances involving a smooth risk minimization where the gradient is computed inexactly either with stochastic noise or a deterministic error. In the stochastic case, we show that our conditions outlined in Section 4.1.2 for convergence are satisfied via increasing batch size or variance reduction. In the deterministic setting for minimizing an empirical risk, a sweeping approach is described.

The type of theorem one can find in Chapter 4 takes the following form:

Theorem 1.3.2. *Suppose that [mild assumptions on parameters and functions] all hold and recall* $\Gamma_k \overset{\text{def}}{=} \sum_{i=0}^{k} \gamma_i$. *For a sequence* $(x_k)_{k\in\mathbb{N}}$ *generated by ICGALP we have:*

(i) Asymptotic feasibility ($\mathbb{P}$-a.s.):

$$\lim_{k\to\infty} \|Ax_k - b\| = 0 \ \ (\mathbb{P}\text{-a.s.}) .$$

(ii) Convergence of the Lagrangian ($\mathbb{P}$-a.s.):

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

(iii) The set of ($\mathbb{P}$-a.s.) weak cluster points of $(x_k)_{k\in\mathbb{N}}$ *is contained in the set of solutions to the primal problem ($\mathbb{P}$-a.s.) and the sequence* $(\mu_k)_{k\in\mathbb{N}}$ *converges ($\mathbb{P}$-a.s.) strongly to a solution of the dual problem.*

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

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

Practical Applications We introduce three different methods for computing the gradient inexactly which are compatible with the assumptions made in our convergence analysis for ICGALP. The first is for empirical risk minimization, in which we take an increasing batch size of samples at each iteration. The second is a variance reduction method which allows for as little as a single sample at each iteration. Last is a deterministic sweeping method which takes a single sample at each iteration in a predetermined way. The variance reduction and sweeping methods are demonstrated on a model problem with an affine constraint in addition to an ℓ^1 ball constraint. This problem, while simple, is not solvable by stochastic Frank-Wolfe methods due to the affine constraint. Besides the work [77], which uses the Moreau envelope instead of an augmented Lagrangian to handle the affine constraint, our algorithm is the only stochastic Frank-Wolfe algorithm which can solve the model problem presented. In contrast to [77], we are able to inexactly compute the terms related to our penalization of the affine constraint.

Prior Work 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. [62, 63, 100, 58, 46, 115, 82, 61], or [80], 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 [77], where the authors consider a stochastic conditional gradient algorithm applied to a composite problem of the form

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

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

1.3.3 Stochastic Bregman Primal-Dual Splitting

Results We introduce and analyze the **S**tochastic **B**regman **P**rimal-**D**ual (SBPD) algorithm to solve (1.2.2). To our knowledge, our work is the first to analyze solving (1.2.2) under a relative smoothness condition with D-prox mappings, even in the deterministic setting. Additionally, we are the first to include stochastic error, denoted δ_k^p and δ_k^d in Algorithm 7, in the computation of the gradient terms for (1.2.2) under these assumptions.

Algorithm 7: Stochastic Bregman Primal-Dual Splitting (SBPD)

for $k = 0, 1, \ldots$ **do**

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{C}_p} \left\{ g(x) + \langle \nabla f(x_k) + \delta_k^p, x \rangle + \langle Tx, \bar{\mu}_k \rangle + \frac{1}{\lambda_k} D_p\left(x, x_k\right) \right\}$$

$$\mu_{k+1} = \operatorname*{argmin}_{\mu \in \mathcal{C}_d} \left\{ l^*(\mu) + \langle \nabla h^*(\mu_k) + \delta_k^d, \mu \rangle - \langle T\bar{x}_k, \mu \rangle + \frac{1}{\nu_k} D_d\left(\mu, \mu_k\right) \right\}$$

where $\bar{\mu}_k = \mu_k$ and $\bar{x}_k = 2x_{k+1} - x_k$.

We are able to show:

- Convergence of the Lagrangian gap $\mathbb{E}\left[\mathcal{L}\left(\bar{x}_k, \mu^\star\right) - \mathcal{L}\left(x^\star, \bar{\mu}_k\right)\right]$ for the ergodic sequence of iterates with a $O\left(1/k\right)$ rate of convergence.
- Every ($\mathbb{P}$-a.s.) weak sequential cluster point of the ergodic sequence is optimal in expectation.
- ($\mathbb{P}$-a.s.) weak convergence of the pointwise sequence of iterates to a solution.
- ($\mathbb{P}$-a.s.) strong convergence of the pointwise sequence of iterates to a solution if the entropies are totally convex and the objective is relatively strongly convex (to be made more precise in Chapter 5).

The type of theorem one can find in Chapter 5 takes the following form:

Theorem 1.3.3. *Let [mild assumptions on the parameters and functions] hold. Then we have the following convergence rate: for each $k \in \mathbb{N}$, for every $(x, \mu) \in \mathcal{C}_p \times \mathcal{C}_d$,*

$$\mathbb{E}\left[\mathcal{L}\left(\bar{x}_k, \mu\right) - \mathcal{L}\left(x, \bar{\mu}_k\right)\right] = O\left(1/k\right). \tag{1.3.4}$$

Furthermore, under additional mild assumptions on the entropies, we have ($\mathbb{P}$-a.s.) weak convergence of the pointwise sequence of iterates $\left(\left(x_k, \mu_k\right)\right)_{k \in \mathbb{N}}$ to a saddle point $(x^\star, \mu^\star)$.

Practical Applications The generality afforded by relative smoothness allows us to solve problems involving the Kullback-Liebler divergence on the simplex with total variation regularization. Such problems were previously inaccessible due to the lack of Lipschitz-smoothness and the lack of prox-friendliness of the total variation regularizer. We solve two of these problems in Chapter 5, giving explicit step size calculations and showing the assumptions are satisfied. In addition to these problems, we demonstrate the effectiveness of the algorithm on entropically regularized Wasserstein inverse problems. Although it was technically possible to solve this problem using the algorithm of [31], it was not done until now. Our reformulation as a saddle-point problem allows us to dramatically reduce the dimensionality.

Prior Work As has been discussed, the idea of using primal-dual methods to solve convex-concave saddle-point problems has been around since the 1960s, e.g., [84], [102], [74], and [81]. For an introduction into the use of primal-dual methods in convex optimization, we refer the reader to [71]. More recently, without being exhaustive, there were the notbale works [36], [29], [39], [112], and [31] which examined problems quite similar to the one posed here using first order primal-dual methods.

In particular, [31] studied (1.2.2) using D-prox mappings, i.e., proximal mappings where the euclidean energy has been replaced by a suitable Bregman divergence, under the assumption that f and h^* are Lipschitz-smooth Γ_0 functions. They show ergodic convergence of the Lagrangian optimality gap with a rate of $O\left(1/k\right)$ under mild assumptions and also faster rates, e.g., $O\left(1/k^2\right)$ and linear convergence, under stricter assumptions involving strong convexity. We generalize their results by relaxing the Lipschitz-smoothness assumption to a relative smoothness assumption, by analyzing the totally convex and relatively strongly convex case, by introducing stochastic error to the algorithm, and by showing almost sure weak convergence of the pointwise iterates themselves.

Generalizations of [31] involving inexactness already exist in the form of [99] and [30], however, [99] only considers determinstic inexactness and proximal operators computed in the euclidean sense, i.e., with entropy equal to the euclidean energy, and requires Lipschitz-smoothness, although it's worth noting that inexactness in their paper is extended to the computation of the proximal operators in contrast to our work which allows for inexactness, in the form of stochastic error, only in the computation of gradient terms. The paper [30] allows for a very particular kind of stochastic error in which one randomly samples a set of indices at each iteration in an arbitrary but fixed way, i.e., according to some fixed distribution. However, the stochastic error we consider here is more general while encompassing the previous cases.

Another related work is that of [59] which generalizes the problem considered [31] by allowing for a non-linear coupling $\Phi\left(x,\mu\right)$ in (1.2.2) instead of $\langle Tx,\mu\rangle$, although they maintain essentially the same Lipschitz-smoothness assumptions as in [31] translated to $\Phi\left(x,\mu\right)$. They are able to show a $O\left(1/k\right)$ convergence rate for the ergodic Lagrangian optimality gap under mild assumptions and an accelerated rate $O\left(1/k^2\right)$ when g in (1.2.2) is strongly convex with another assumption on the coupling $\Phi\left(x,\mu\right)$.

1.4 Outline

The remainder of the thesis, with the exception of the mathematical background chapter, Chapter 2, and the final conclusion chapter, Chapter 6, is divided into three chapters, each corresponding to an individual research work. Chapter 2 gathers the basic notation and terminology that will be used throughout. It is divided into two sections; one dealing primarily with convex analysis and one dealing with probability theory. Chapter 6 sums up the main ideas and contributions of the thesis. In general, we structure each of the three main chapters, Chapter 3, Chapter 4, and Chapter 5, in the same way, although some chapters have extra sections in addition to these core sections.

Introduction: states the problem under consideration, the proposed algorithm to solve it, the main assumptions on the problem and the parameters involved in the algorithm, and the organization of the chapter.

Estimations: prepares the basic inequalities and estimations from convex analysis and probability theory that will be used in the arguments of the convergence analysis section.

Convergence: develops the main theorems and their proofs; typically convergence of some measure of optimality with rates of convergence as well.

Applications: details how to apply the algorithm and problem formulation to different contexts, practical implementations of the algorithm are developed and the proposed theoretical claims of the previous section are verified numerically.

Chapter 2

Background

Contents

We assemble in this chapter the relevant background material and some notations that will be used throughout the following chapters. There are four lemmas in this chapter which are new: Lemma 2.1.16, Lemma 2.1.17, Lemma 2.1.18, Lemma 2.2.5. We roughly divide the chapter into results from convex analysis, results from real analysis, and results from probability theory, first stating some common notation below.

Given a reflexive Banach space $\mathcal{X}$ with norm $\|\cdot\|_{\mathcal{X}}$, we denote by $\mathcal{X}^*$ is topological dual, the space of all continuous linear functionals, and by $\langle u, x\rangle_{\mathcal{X}}$ the duality pairing for $x \in \mathcal{X}$ and $u \in \mathcal{X}^*$. In general, we will not write the subscript $\mathcal{X}$ for the norm or duality pairing of $\mathcal{X}$, relying on context to convey which space the norm comes from. We will also let $\mathcal{H}$ denote an arbitrary real Hilbert space with norm $\|\cdot\|_{\mathcal{H}}$ and inner product $\langle x, y\rangle_{\mathcal{H}}$ for $x, y \in \mathcal{H}$. When referring to the differentiability or the gradient of a function $f : \mathcal{X} \to \mathbb{R}$ we mean in the sense of the Gâteaux derivative. We say a sequence $(x_k)_{k\in\mathbb{N}}$ with $x_k \in \mathcal{X}$ for each $k \in \mathbb{N}$ converges strongly to some $x \in \mathcal{X}$, denoted $x_k \to x$, iff

$$\|x_k - x\|_{\mathcal{X}} \to 0.$$

On the other hand, we say a sequence $(x_k)_{k\in\mathbb{N}}$ with $x_k \in \mathcal{X}$ for each $k \in \mathbb{N}$ converges weakly to some $x \in \mathcal{X}$, denoted $x_n \rightharpoonup x$, iff, for every $u \in \mathcal{X}^*$,

$$\langle u, x_k\rangle_{\mathcal{X}} \to \langle u, x\rangle_{\mathcal{X}}\,.$$

Finally, when referring the interior of a set U, denoted $\mathrm{int}U$, the boundary, denoted $\mathrm{bd}\,(U)$, or the closure, denoted $\overline{U}$, we mean with respect to the norm topology (also referred to as the strong topology) on $\mathcal{X}$.

2.1 Convex Analysis

We recall some important definitions and results from convex analysis. For a more comprehensive coverage we refer the interested reader to [10, 92] and [103] in the finite dimensional case. Throughout, we let g be an arbitrary function from $\mathcal{X}$ to the real extended line, namely $g : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$. The function g is said to belong to $\Gamma_0\,(\mathcal{X})$ if it is proper, convex, and lower semicontinuous. The *domain* of g is defined to be $\mathrm{dom}\,(g) \stackrel{\text{def}}{=} \{x \in \mathcal{X} : g\,(x) < +\infty\}$. The *Legendre-Fenchel conjugate* of g is the function $g^* : \mathcal{X}^* \to \mathbb{R}\cup\{+\infty\}$ such that, for every $u \in \mathcal{X}^*$,

$$g^*\,(u) \stackrel{\text{def}}{=} \sup_{x\in\mathcal{X}} \{\langle u, x\rangle - g\,(x)\}\,.$$

Notice that

$$g_1 \leq g_2 \quad \Longrightarrow \quad g_2^* \leq g_1^*. \tag{2.1.1}$$

Moreau proximal mapping and envelope We discuss a construction for a function $g : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$. The *proximal operator* for the function g is defined to be

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

and its *Moreau envelope* with parameter β as

$$g^\beta(x) \stackrel{\text{def}}{=} \inf_{y \in \mathcal{H}} \left\{ g(y) + \frac{1}{2\beta}\|x - y\|^2 \right\}. \tag{2.1.2}$$

Proposition 2.1.1. *Let $g \in \Gamma_0(\mathcal{H})$ and denote $x^+ = \operatorname{prox}_g(x)$. Then, for all $y \in \mathcal{H}$,*

$$2\left(g\left(x^+\right) - g(x)\right) + \left\|x^+ - y\right\|^2 - \|x - y\|^2 + \left\|x^+ - x\right\|^2 \leq 0. \tag{2.1.3}$$

Proof. The result is well-known and the proof is readily available, e.g. in [92, Chapter 6.2.1]. Indeed, to show this result we simply apply strong convexity to the function $g(\cdot) + \frac{1}{2}\|\cdot - x\|^2$ and note that the $\operatorname{prox}_g(x)$ is the minimizer of this function. □

We recall that the *subdifferential* of the function g is defined as the set-valued operator $\partial g : \mathcal{H} \to 2^{\mathcal{H}}$ such that, for every x in $\mathcal{H}$,

$$\partial g(x) \stackrel{\text{def}}{=} \left\{ u \in \mathcal{H} : \; g(y) \geq g(x) + \langle u, y - x \rangle \quad \forall y \in \mathcal{H} \right\}. \tag{2.1.4}$$

We denote $\operatorname{dom}(\partial g) \stackrel{\text{def}}{=} \left\{ x \in \mathcal{H} : \; \partial g(x) \neq \emptyset \right\}$. When g belongs to $\Gamma_0(\mathcal{H})$, it is well-known that the subdifferential is a maximal monotone operator. If, moreover, the function is Gâteaux differentiable at $x \in \mathcal{H}$, then $\partial g(x) = \{\nabla g(x)\}$. For $x \in \operatorname{dom}(\partial g)$, the *minimal norm selection* of $\partial g(x)$ is defined to be the unique element $\left\{ [\partial g(x)]^0 \right\} \stackrel{\text{def}}{=} \operatorname*{Argmin}_{y \in \partial g(x)} \|y\|$. Then we have the following fundamental result about Moreau envelopes.

Proposition 2.1.2. *Given a function $g \in \Gamma_0(\mathcal{H})$, we have the following:*

(i) The Moreau envelope is convex, real-valued, and continuous.

(ii) Lax-Hopf formula: the Moreau envelope is the viscosity solution to the following Hamilton Jacobi equation:

$$\begin{cases} \frac{\partial}{\partial \beta} g^\beta(x) = -\frac{1}{2}\left\|\nabla_x g^\beta(x)\right\|^2 & (x, \beta) \in \mathcal{H} \times (0, +\infty) \\ g^0(x) = g(x) & x \in \mathcal{H}. \end{cases} \tag{2.1.5}$$

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

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

(iv) $\forall x \in \operatorname{dom}(\partial g)$, $\left\|\nabla g^\beta(x)\right\| \nearrow \left\|[\partial g(x)]^0\right\|$ as $\beta \searrow 0$.

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

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

Proof. (i): see [10, Proposition 12.15]. The proof for (ii) can be found in [6, Lemma 3.27 and Remark 3.32] (see also [65] or [4, Section 3.1]). The proof for claim (iii) can be found in [10, Proposition 12.29] and the proof for claim (iv) can be found in [10, Corollary 23.46]. For the first part in (v), see [10, Proposition 12.32(i)]. To show the first inequality in (v), combine (ii) and convexity of the function $\beta \mapsto g^\beta(x)$ for every $x \in \mathcal{H}$. The second inequality follows from the first one and (iv), taking the limit as $\beta' \to 0$. □

Remark 2.1.3.

(i) While the regularity claim in Proposition 2.1.2(iii) of the Moreau envelope $g^{\beta}(x)$ with respect to x is well-known, a less known result is the C^1-regularity with respect to β for any $x \in \mathcal{H}$ (Proposition 2.1.2(ii)). To our knowledge, the proof goes back, at least, to the book of [6]. Though it has been rediscovered in the recent literature in less general settings.

(ii) For given functions $H : \mathcal{H} \to \mathbb{R}$ and $g_0 : \mathcal{H} \to \mathbb{R}$, a natural generalization of the Hamilton-Jacobi equation in (2.1.5) is

$$\begin{cases} \frac{\partial}{\partial \beta} g(x,\beta) + H(\nabla_x g(x,\beta)) = 0 & (x,\beta) \in \mathcal{H} \times (0,+\infty), \\ g(x,0) = g_0(x) & x \in \mathcal{H}. \end{cases}$$

Supposing that H is convex and that $\lim_{\|p\| \to +\infty} H(p)/\|p\| = +\infty$, the solution of the above system is given by the Lax-Hopf formula (see [52, Theorem 5, Section 3.3.2][1]):

$$g(x,t) \stackrel{\text{def}}{=} \inf_{y \in \mathcal{H}} \left\{ g_0(y) + tH^* \left(\frac{y-x}{t} \right) \right\}.$$

If $H(p) = \frac{1}{2}\|p\|^2$, then $H^*(p) = \frac{1}{2}\|p\|^2$ and we recover the result in Proposition 2.1.2.

Regularity of differentiable functions In what follows, we introduce some definitions related with regularity of differentiable functions. They will provide useful upper-bounds and descent properties. Notice that, besides Lemma 2.1.17 and Lemma 2.1.18, the notions and results of this part are independent from convexity. The lemmas in this section are new results, although quite similar to the results cited in the proofs.

Definition 2.1.4 (ω**-smoothness).** Consider a function $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\omega(0) = 0$ and

$$\xi(s) \stackrel{\text{def}}{=} \int_0^1 \omega(st)\, dt \tag{2.1.6}$$

is nondecreasing. A differentiable function $g : \mathcal{H} \to \mathbb{R}$ is said to belong to $C^{1,\omega}(\mathcal{H})$ or to be ω-smooth if the following inequality is satisfied for every $x, y \in \mathcal{H}$:

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

Lemma 2.1.5 (ω**-smooth Descent Lemma).** *Given a function $g \in C^{1,\omega}(\mathcal{H})$ we have the following inequality: for every x and y in $\mathcal{H}$,*

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

where ξ is defined in (2.1.6).

Proof. We recall here the simple proof for completeness:

$$\begin{aligned} g(y) - g(x) &= \int_0^1 \frac{d}{dt} g(x + t(y-x))\, dt \\ &= \int_0^1 \langle \nabla g(x), y-x \rangle\, dt + \int_0^1 \langle \nabla g(x + t(y-x)) - \nabla g(x), y-x \rangle\, dt \\ &\leq \langle \nabla g(x), y-x \rangle + \|y-x\| \int_0^1 \|\nabla g(x + t(y-x)) - \nabla g(x)\|\, dt \\ &\leq \langle \nabla g(x), y-x \rangle + \|y-x\| \int_0^1 \omega(t\|y-x\|)\, dt, \end{aligned}$$

where, in the first inequality, we used Cauchy-Schwarz and, in the second, Definition 2.1.4. We conclude using the definition of ξ (see (2.1.6)). □

[1] The proof in [52] is given in the finite-dimensional case but it extends readily to any real Hilbert space.

Remark 2.1.6. For $L > 0$ and $\omega(t) = Lt^{\nu}$, $\nu \in]0,1]$, $C^{1,\omega}(\mathcal{H})$ is the space of differentiable functions with Hölder continuous gradients, in which case $\xi(s) = Ls^{\nu}/(1+\nu)$ and the Descent Lemma reads

$$g(y) - g(x) \leq \langle \nabla g(x), y - x \rangle + \frac{L}{1+\nu} \|y - x\|^{1+\nu}, \tag{2.1.7}$$

see e.g., [86, 87]. When $\nu = 1$, we have that $C^{1,\omega}(\mathcal{H})$ is the class of differentiable functions with L-Lipschitz continuous gradient, and one recovers the classical Descent Lemma.

Now, following [13], we introduce some notions that allow one to further generalize (2.1.7). We state a subset of these results only for Banach spaces.

Definition 2.1.7 (Bregman divergence). Given a Banach space $\mathcal{X}$ with duality pairing $\langle \cdot, \cdot \rangle$ and a function, often referred to as the entropy, $\phi : \mathcal{X} \to \mathbb{R}$ differentiable on its domain, its Bregman divergence is defined by

$$D_{\phi}(x,y) \overset{\text{def}}{=} \begin{cases} \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle & \text{if } x \in \operatorname{dom}(\phi) \text{ and } y \in \operatorname{intdom}(\phi), \\ +\infty & \text{else.} \end{cases} \tag{2.1.8}$$

Notice that, if ϕ belongs to $\Gamma_0(\mathcal{X})$, then the Bregman divergence associated to ϕ is always nonnegative by the subdifferential inequality. A function F is said to be *essentially smooth* if it is differentiable on the interior of its domain and if it satisfies, for each sequence $(x_k)_{k\in\mathbb{N}}$ in $\operatorname{intdom}(F)$ such that $x_k \to x \in \operatorname{bd}(\operatorname{dom}(F))$,

$$\|\nabla f(x_k)\| \to +\infty$$

Definition 2.1.8 (Legendre function). The function ϕ is called a Legendre function if $\partial\phi$ is both locally bounded and single-valued on its domain, $(\partial\phi)^{-1}$ is locally bounded on its domain, and ϕ is strictly convex on every convex subset of $\operatorname{dom}\partial\phi$.

Definition 2.1.9 (Relative smoothness). Given a differentiable function $\phi : \mathcal{X} \to \mathbb{R}$, we say that the function $f : \mathcal{X} \to \mathbb{R}$ is L-smooth with respect to ϕ if it is differentiable and $L\phi - f$ is convex; namely, if for every $x, y \in \mathcal{X}$

$$D_f(x,y) \leq L D_{\phi}(x,y).$$

Remark 2.1.10. The relative smoothness property, used notably in [13], implies the following fact which can be interpreted as a "generalized descent lemma"; for every $x, y \in \mathcal{X}$,

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + L D_{\phi}(x,y). \tag{2.1.9}$$

When ϕ is the Euclidean square norm, or energy, relative smoothness is equivalent to Lipschitz-smoothness, i.e., Lipschitz-continuity of the gradient of f.

Definition 2.1.11 (Relative strong convexity). Given a differentiable function $\phi : \mathcal{X} \to \mathbb{R}$, we say that f is m-strongly convex with respect to ϕ if $f - m\phi$ is convex.

Note that the idea of relative strong convexity can be found in a footnote of [31] but it was not explored as it was not clear if there are any interesting examples for which it is true. For our purposes, we slightly changed the definition in [13, Definition 1] in a weaker sense, and we have the following result.

Lemma 2.1.12 (Generalized Descent Lemma, [13]). *Let F and f be differentiable on $\mathcal{C}_0$, where $\mathcal{C}_0$ is an open subset of* $\operatorname{int}(\operatorname{dom}(F))$. *Assume that $F - f$ is convex on $\mathcal{C}_0$. Then, for every x and y in $\mathcal{C}_0$,*

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

Proof. For our purpose, we intentionally weakened the hypothesis needed in the original result of [13, Lemma 1]. We repeat their argument but show the result is still valid under our weaker assumption. Let x and y be in $\mathcal{C}_0$, where, by hypothesis, $\mathcal{C}_0$ is open and contained in $\operatorname{int}(\operatorname{dom}(F))$. As $F - f$ is convex and differentiable on $\mathcal{C}_0$, from the gradient inequality (2.1.4) we have, for all $y \in \mathcal{C}_0$,

$$(F - f)(y) \geq (F - f)(x) + \langle \nabla(F - f)(x), y - x \rangle.$$

Rearranging the terms and using the definition of D_F in (2.1.8), we obtain the claim. □

The previous lemma suggests the introduction of the following definition, which extends Definition 2.1.4 and 2.1.9 by incorporating the idea of a curvature constant of F with respect to the set $\mathcal{C}$.

Definition 2.1.13 ((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

$$K_{(F,\zeta,\mathcal{C})} \overset{\text{def}}{=} \sup_{\substack{x,s\in\mathcal{C};\ \gamma\in]0,1] \\ z=x+\gamma(s-x)}} \frac{D_F(z,x)}{\zeta(\gamma)} \quad < \quad +\infty. \tag{2.1.10}$$

$K_{(F,\zeta,\mathcal{C})}$ is a generalization of the standard curvature constant widely used in the literature of conditional gradient. The curvature constant was first studied in the context of conditional gradient algorithms in [66].

Remark 2.1.14. Assume that $(f, \mathcal{C})$ is (F,ζ)-smooth. Using first Lemma 2.1.12 and then the definition in (2.1.10), we have the following descent property: for every $x, s \in \mathcal{C}$ and for every $\gamma \in]0,1]$,

$$\begin{aligned} f(x+\gamma(s-x)) &\leq f(x) + \gamma\langle \nabla f(x), s-x\rangle + D_F(x+\gamma(s-x), x) \\ &\leq f(x) + \gamma\langle \nabla f(x), s-x\rangle + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma). \end{aligned}$$

Notice that, as in the previous definition, we do not require $\mathcal{C}$ to be convex. So, in general, the point $z = x + \gamma(s-x)$ may not lie in $\mathcal{C}$.

Remark 2.1.15. 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

$$\lim_{t\to 0^+} \omega(t) = \omega(0) = 0.$$

We will denot the *diameter* of a set $\mathcal{C}$ by the following,

$$d_{\mathcal{C}} \overset{\text{def}}{=} \sup_{x,y\in\mathcal{C}} \|x-y\|$$

Lemma 2.1.16. *Suppose that the set $\mathcal{C}$ is bounded with diameter $d_{\mathcal{C}}$. Moreover, assume that the function f is ω-smooth on some open and convex subset $\mathcal{C}_0$ containing $\mathcal{C}$. Set $\zeta(\gamma) \overset{\text{def}}{=} \xi(d_{\mathcal{C}}\gamma)$, where ξ is given in (2.1.6). Then the pair $(f, \mathcal{C})$ is (f,ζ)-smooth with $K_{(f,\zeta,\mathcal{C})} \leq d_{\mathcal{C}}$.*

Proof. With $F = f$ and f being ω-smooth on $\mathcal{C}_0$, both F and f are differentiable on $\mathcal{C}_0$ and $F - f \equiv 0$ is convex on $\mathcal{C}_0$. Thus, all conditions required in Definition 2.1.13 hold true. It then remains to show (2.1.10) with the bound $K_{(f,\zeta,\mathcal{C})} \leq d_{\mathcal{C}}$. First notice that, for every $x, s \in \mathcal{C}$ and for every $\gamma \in]0,1]$, the point $z = x + \gamma(s-x)$ belongs to $\mathcal{C}_0$. Indeed, $\mathcal{C} \subset \mathcal{C}_0$ and $\mathcal{C}_0$ is convex by hypothesis. In particular, as f is ω-smooth on $\mathcal{C}_0$, the Descent

Lemma, Lemma 2.1.5, holds between the points x and z. Then

$$\begin{aligned} K_{(f,\zeta,\mathcal{C})} &= \sup_{\substack{x,s\in\mathcal{C};\ \gamma\in]0,1] \\ z=x+\gamma(s-x)}} \frac{D_f(z,x)}{\zeta(\gamma)} \\ &= \sup_{\substack{x,s\in\mathcal{C};\ \gamma\in]0,1] \\ z=x+\gamma(s-x)}} \frac{f(z)-f(x)-\langle\nabla f(x), z-x\rangle}{\xi(d_{\mathcal{C}}\gamma)} \\ &\leq \sup_{\substack{x,s\in\mathcal{C};\ \gamma\in]0,1] \\ z=x+\gamma(s-x)}} \frac{\|z-x\|\,\xi(\|z-x\|)}{\xi(d_{\mathcal{C}}\gamma)} \\ &= \sup_{x,s\in\mathcal{C};\ \gamma\in]0,1]} \frac{\gamma\|s-x\|\,\xi(\gamma\|s-x\|)}{\xi(d_{\mathcal{C}}\gamma)} \\ &\leq \sup_{\gamma\in]0,1]} \frac{\gamma d_{\mathcal{C}}\xi(d_{\mathcal{C}}\gamma)}{\xi(d_{\mathcal{C}}\gamma)} = d_{\mathcal{C}}. \end{aligned}$$

In the first inequality we used Lemma 2.1.5, while in the second we used that $\|s-x\| \leq d_{\mathcal{C}}$ (both x and s belong to $\mathcal{C}$, that is bounded by hypothesis) and the monotonicity of the function ξ (see Definition 2.1.4). □

The following lemma gives several sufficient conditions to ensure that the minimal norm selection of the subdifferential of a convex function is bounded over some set $\mathcal{C}$.

Lemma 2.1.17. *Let $T : \mathcal{H}_p \to \mathcal{H}_v$ be a bounded linear operator. Assume that one of the following holds:*

(i) $g \in \Gamma_0(\mathcal{H}_v)$, $T\mathcal{C} \subset \operatorname{int}(\operatorname{dom}(g))$ and $\mathcal{C}$ is a nonempty compact subset of $\mathcal{H}_p$.

(ii) $g : \mathcal{H}_v \to \mathbb{R}$ is continuous, convex and bounded on bounded sets of $\mathcal{H}_v$, and $\mathcal{C}$ is a nonempty bounded subset of $\mathcal{H}_p$.

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

Then $T\mathcal{C} \subset \operatorname{dom}(\partial g)$ and $\sup_{x\in\mathcal{C}} \left\| [\partial g(Tx)]^0 \right\| \leq M < \infty$, where M is a positive constant.

Proof. (i) Since $g \in \Gamma_0(\mathcal{H}_p)$, it follows from [10, Proposition 16.21] that

$$T\mathcal{C} \subset \operatorname{int}(\operatorname{dom}(g)) \subset \operatorname{dom}(\partial g).$$

Moreover, by [10, Corollary 8.30(ii) and Proposition 16.14], we have that ∂g is locally bounded on $\operatorname{int}(\operatorname{dom}(g))$. In particular, as we assume that $\mathcal{C}$ is bounded, so is $T\mathcal{C}$, and since $T\mathcal{C} \subset \operatorname{int}(\operatorname{dom}(g))$, it means that for each $z \in T\mathcal{C}$ there exists an open neighborhood of z, denoted by U_z, such that $\partial g(U_z)$ is bounded. Since $(U_z)_{z\in\mathcal{C}}$ is an open cover of $T\mathcal{C}$ and $T\mathcal{C}$ is compact, there exists a finite subcover $(U_{z_k})_{k=1}^n$. Then,

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

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

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

whence the desired conclusion trivially follows.

(ii) From the equivalence [10, Proposition 16.17(i) $\iff$ (iii)], it follows that $\operatorname{dom}(\partial g) = \mathcal{H}_v$ and thus $T\mathcal{C} \subset \operatorname{dom}(\partial g)$ trivially holds. Moreover, ∂g is bounded on every bounded set of $\mathcal{H}_v$, and in particular on $\mathcal{C}$.

(iii) In finite dimension, the claim follows trivially from (i) for the first case by a simple compactness argument, and from (ii) in the second case since a continuous and convex is bounded on bounded sets in finite dimension; see [10, Proposition 16.17].

□

Similarly, we show in the following lemma that (F,ζ)-smoothness is sufficient to ensure that a function $f \in \Gamma_0(\mathcal{H})$ has uniformly bounded gradient.

Lemma 2.1.18. *Assume that $f \in \Gamma_0(\mathcal{H})$ and is (F,ζ)-smooth for some F and ζ over the set $\mathcal{C}$. Then*

$$\sup_{x\in\mathcal{C}} \|\nabla f(x)\| \leq D < +\infty$$

for some positive constant D.

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

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

where we used the Fenchel identity ([10, Proposition 17.27]) in the first equality, Lemma 2.1.12 in the first inequality and Definition 2.1.13 in the second one. By [10, Corollary 9.20], f is bounded from below on $\mathcal{C}$ which entails

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

for some real constant c. Now, since

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

by Definition 2.1.13 and [10, Proposition 17.41], we infer from [10, Theorem 14.17 and Proposition 14.16] (recall that s is fixed), that there exists $a_1 > 0$ and $a_2 \in \mathbb{R}$ such that, for all $x \in \mathcal{C}$,

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

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

Indicator and support functions Given a subset $\mathcal{C} \subset \mathcal{X}$, we define its *indicator function* as

$$\iota_{\mathcal{C}}(x) \overset{\text{def}}{=} \begin{cases} 0 & x \in \mathcal{C} \\ +\infty & x \notin \mathcal{C}. \end{cases}$$

Recall that, if $\mathcal{C}$ is nonempty, closed, and convex, then $\iota_{\mathcal{C}}$ belongs to $\Gamma_0(\mathcal{X})$. Remember also the definition of the *support function* of $\mathcal{C}$, $\sigma_{\mathcal{C}} \overset{\text{def}}{=} \iota_{\mathcal{C}}^*$. Equivalently, $\sigma_{\mathcal{C}}(x) \overset{\text{def}}{=} \sup\{\langle z, x\rangle : z \in \mathcal{C}\}$. We denote by $\mathrm{ri}(\mathcal{C})$ the *relative interior* of the set $\mathcal{C}$ (in finite dimension, it is the interior for the topology relative to its affine full). We denote $\mathrm{par}(C)$ as the subspace parallel to $\mathcal{C}$ which, in finite dimension, takes the form $\mathbb{R}(C - C)$.

We have the following characterization of the support function from the relative interior in finite dimension.

Proposition 2.1.19 ([111]). *Let $\mathcal{H}$ be finite-dimensional and $\mathcal{C} \subset \mathcal{H}$ a nonempty, closed bounded and convex subset. If $0 \in \mathrm{ri}(\mathcal{C})$, then $\sigma_{\mathcal{C}} \in \Gamma_0(\mathbb{R}^n)$ is sublinear, nonnegative and finite-valued, and*

$$\sigma_{\mathcal{C}}(x) = 0 \iff x \in (\mathrm{par}(\mathcal{C}))^{\perp}.$$

Proof. The proof can be found in [111, Lemma 1]. □

Coercivity We recall that a function g is *coercive* if $\lim_{\|x\|\to+\infty} g(x) = +\infty$ and that coercivity is equivalent to the boundedness of the sublevel-sets [10, Proposition 11.11]. We have the following result, that relates coercivity to properties of the Fenchel conjugate.

Proposition 2.1.20 ([10]). *Given g in $\Gamma_0(\mathcal{H})$, g^* is coercive if and only if $0 \in \mathrm{int}(\mathrm{dom}(g))$.*

Proof. The proof can be found in [10, Theorem 14.17]. □

The *recession function* (sometimes referred to as the *horizon function*) of g at a given point $d \in \mathbb{R}^n$ is defined to be the function $g^{d,\infty} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ such that, for every $x \in \mathbb{R}^n$,

$$g^{d,\infty}(x) \stackrel{\text{def}}{=} \lim_{\alpha \to \infty} \frac{g(d + \alpha x) - g(d)}{\alpha}.$$

Recall that, if g is convex, the recession function is independent from the selection of the point $d \in \mathbb{R}^n$ and can be then simply denoted as g^∞. In finite dimension, the following result relates coercivity to properties of the recession function.

Proposition 2.1.21. *Let $g \in \Gamma_0(\mathbb{R}^n)$ and $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear operator. Then,*

(i) *g coercive $\iff$ $g^\infty(x) > 0 \quad \forall x \neq 0$.*

(ii) *$g^\infty \equiv \sigma_{\mathrm{dom}(g^*)}$.*

(iii) *$(g \circ A)^\infty \equiv g^\infty \circ A$.*

In particular, we deduce that $g \circ A$ is coercive if and only if $\sigma_{\mathrm{dom}(g^*)}(Ax) > 0$ for every $x \neq 0$.

Proof. The proofs can be found in [104, Theorem 3.26], [104, Theorem 11.5] and [73, Corollary 3.2] respectively. □

2.2 Real Analysis

We list in this section some definitions and lemmas for real sequences that will be used to prove the convergence properties of the algorithms in later chapters. We denote ℓ_+ as the set of all sequences taking values in $[0, +\infty[$. Given $p \in [1, +\infty[$, ℓ^p is the space of real sequences $(r_k)_{k\in\mathbb{N}}$ such that

$$\left(\sum_{k=1}^{\infty} |r_k|^p\right)^{1/p} < +\infty.$$

For $p = +\infty$, we denote by ℓ^∞ the space of bounded sequences. Furthermore, we will use the notation $\ell^p_+ \stackrel{\text{def}}{=} \ell^p \cap \ell_+$. In the next, we recall some key results about real sequences.

Lemma 2.2.1 ([33, 97]). *Consider three sequences $(r_k)_{k\in\mathbb{N}} \in \ell_+$, $(a_k)_{k\in\mathbb{N}} \in \ell_+$, and $(z_k)_{k\in\mathbb{N}} \in \ell^1_+$, such that*

$$r_{k+1} \leq r_k - a_k + z_k, \quad \forall k \in \mathbb{N}.$$

Then $(r_k)_{k\in\mathbb{N}}$ is convergent and $(a_k)_{k\in\mathbb{N}} \in \ell^1_+$.

Proof. This result is found more recently in [33, Lemma 3.1] or [97, Lemma 2, page 44]. □

Lemma 2.2.2 ([2]). *Consider two sequences $(p_k)_{k\in\mathbb{N}} \in \ell_+$ and $(w_k)_{k\in\mathbb{N}} \in \ell_+$ such that $(p_k w_k)_{k\in\mathbb{N}} \in \ell^1_+$ and $(p_k)_{k\in\mathbb{N}} \notin \ell^1$. Then the following holds:*

(i) There exists a subsequence $\left(w_{k_j}\right)_{j\in\mathbb{N}}$ such that, for all $j \in \mathbb{N}$,

$$w_{k_j} \leq P_{k_j}^{-1},$$

where $P_n \stackrel{\text{def}}{=} \sum_{k=1}^n p_k$. In particular, $\liminf_k w_k = 0$.

(ii) If moreover there exists a constant $\alpha > 0$ such that

$$w_k - w_{k+1} \leq \alpha p_k$$

for every $k \in \mathbb{N}$, then

$$\lim_k w_k = 0.$$

Proof. The proofs can be found in [2, Theorem 2] and [2, Proposition 2(ii)]. □

Lemma 2.2.3. *Consider the sequences* $(r_k)_{k\in\mathbb{N}} \in \ell_+$, $(p_k)_{k\in\mathbb{N}} \in \ell_+$, $(w_k)_{k\in\mathbb{N}} \in \ell_+$, *and* $(z_k)_{k\in\mathbb{N}} \in \ell_+$. *Suppose that* $(z_k)_{k\in\mathbb{N}} \in \ell_+^1$, $(p_k)_{k\in\mathbb{N}} \notin \ell^1$, *and that, for some* $\alpha > 0$, *the following inequalities are satisfied for every* $k \in \mathbb{N}$*:*

$$\begin{aligned} r_{k+1} &\leq r_k - p_k w_k + z_k; \\ w_k - w_{k+1} &\leq \alpha p_k. \end{aligned} \tag{2.2.1}$$

Then,

(i) $(r_k)_{k\in\mathbb{N}}$ *is convergent and* $(p_k w_k)_{k\in\mathbb{N}} \in \ell_+^1$.

(ii) $\lim_k w_k = 0$.

(iii) For every $k \in \mathbb{N}$, $\inf_{1\leq i\leq k} w_i \leq (r_0 + Z_\infty)/P_k$, *where, again,* $P_n \overset{\text{def}}{=} \sum_{k=1}^n p_k$ *and* $Z_\infty \overset{\text{def}}{=} \sum_{k=1}^{+\infty} z_k$.

(iv) There exists a subsequence $\left(w_{k_j}\right)_{j\in\mathbb{N}}$ *such that, for all* $j \in \mathbb{N}$, $w_{k_j} \leq P_{k_j}^{-1}$.

Proof. (i) See Lemma 2.2.1.

(ii) Claim (ii) follows by combining (i) and Lemma 2.2.2(ii).

(iii) Sum (2.2.1) using a telescoping property and summability of $(z_k)_{k\in\mathbb{N}}$.

(iv) Claim (iv) follows by combining (i) and Lemma 2.2.2(i).

□

Remark 2.2.4. Notice that the conclusions of Lemma 2.2.3 remain true if nonnegativity of the sequence $(r_k)_{k\in\mathbb{N}}$ is replaced with the assumption that it is bounded from below by a trivial translation argument. Observe also that Lemma 2.2.3 guarantees the convergence of the whole sequence to zero, but it gives a convergence rate only on a subsequence.

Lemma 2.2.5. *Consider two positive sequences* $(u_k)_{k\in\mathbb{N}}$ *and* $(\gamma_k)_{k\in\mathbb{N}}$ *which satisfy, for some* $c, d > 0$, *for each* $k \in \mathbb{N}$,

$$u_{k+1} \leq (1 - c\gamma_k^s)\, u_k + d\gamma_k^t, \tag{2.2.2}$$

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

$$\frac{\gamma_k}{\gamma_{k+1}} \leq 1 + o\left(\gamma_k^s\right), \tag{2.2.3}$$

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 $\nu_k \overset{\text{def}}{=} \gamma_k^{s-t} u_k - \frac{d}{c}$ such that $u_k = \gamma_k^{t-s}\left(\nu_k + \frac{d}{c}\right)$. Then, by (2.2.2),

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

By (2.2.3), we then have, for each $k \in \mathbb{N}$,

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

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

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

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,

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

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 [97, Ch.2, Lemma 3], that $\limsup_k \nu_k\leq 0$. Thus, by definition of ν_k,

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

□

Finally, we state a well-known result, Pinsker's inequality, which lower bounds the Kullback-Leibler divergence on the simplex by the total variation norm.

Lemma 2.2.6 (Pinsker's Inequality [95]). *Let $x,y\in\Sigma^n\overset{\text{def}}{=}\left\{u\in\mathbb{R}^n:u\geq 0,u^T\mathbb{1}=1\right\}$ and let K be the Shannon-Boltzmann entropy; $K(x)=\sum_{i=1}^n x_i\log(x_i)$. Then it holds*

$$\frac{1}{2}\|x-y\|_1^2\leq D_K(x,y).$$

2.3 Probability and Random Variables

Many of the following notations for probabilistic concepts are adopted directly from [37]. We denote by $(\Omega,\mathcal{F},\mathbb{P})$ a *probability space* with a *set of events* Ω, a *σ-algebra* $\mathcal{F}$, and a *probability measure* $\mathbb{P}$. When discussing random variables we will assume that any Hilbert space $\mathcal{H}$ or Banach space $\mathcal{X}$ is endowed with the Borel σ-algebra, $\mathcal{B}(\mathcal{H})$, induced by the strong topology. We denote a *filtration* by $\mathfrak{F}=(\mathcal{F}_k)_{k\in\mathbb{N}}$, i.e. a sequence of sub-σ-algebras which satisfies $\mathcal{F}_k\subset\mathcal{F}_{k+1}\in\mathcal{F}$ 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 ($\mathbb{P}$-almost surely) (denoted ($\mathbb{P}$-a.s.)) if

$$\mathbb{P}(\{\omega\in\Omega:(P)\text{ holds}\})=1.$$

Throughout the manuscript, both equalities and inequalities involving random quantities should be understood as holding ($\mathbb{P}$-almost surely), whether or not it is explicitly written.

Definition 2.3.1. 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 2.3.2 ([101]). *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 ($\mathbb{P}$-a.s.) to a random variable with value in $[0,+\infty[$.

Proof. See [101, Theorem 1]. □

Lemma 2.3.3. *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 $\liminf_k w_k = 0$ ($\mathbb{P}$-a.s.),

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

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

Proof. The second result is directly from [9, Lemma 2.2] and the first follows from [2] trivially extended to the stochastic setting. □

Lemma 2.3.4. *If $(x_n)_{k\in\mathbb{N}}$ is a sequence of random variables such that $(\mathbb{E}(\|x_k\|^q))_{k\in\mathbb{N}} \in \ell_+^1$ for some $q \in]0, +\infty[$, then $x_k \to 0$ ($\mathbb{P}$-a.s.).*

Proof. For every $\varepsilon > 0$, by Markov inequality,

$$\sum_{n=0}^{N} \mathbb{P}(\|x_n\|^q \geq \varepsilon) \leq \frac{1}{\varepsilon} \sum_{n=0}^{N} \mathbb{E}(\|x_n\|^q). \tag{2.3.1}$$

Taking the limit for $N \to +\infty$ and using the assumption $(\mathbb{E}(\|x_k\|^q))_{k\in\mathbb{N}} \in \ell_+^1$, we get that, for every $\varepsilon > 0$, also $\mathbb{P}(\|x_n\|^q \geq \varepsilon)$ belongs to ℓ_+^1. As a consequence of the Borel-Cantelli Lemma, $\|x_n\|^q \to 0$ ($\mathbb{P}$-a.s.) and thus $x_n \to 0$ almost surely. □

Chapter 3

Generalized Conditional Gradient with Augmented Lagrangian and Proximal Step

In this chapter we propose a novel splitting scheme which hybridizes generalized conditional gradient with the augmented Lagrangian method, which we call CGALP algorithm, for minimizing

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

where $f \in \Gamma_0(\mathcal{H})$ satisfies a relaxed differentiability condition, $g \circ T \in \Gamma_0(\mathcal{H})$ with g prox-friendly, $h \in \Gamma_0(\mathcal{H})$ has a weakly compact domain and admits an accessible linear minimization oracle, and T and A are bounded linear operators. While classical conditional gradient methods require Lipschitz-continuity of the gradient of the differentiable part of the objective, CGALP needs only differentiability (on an appropriate subset), hence circumventing the intricate question of Lipschitz continuity of gradients. For the functions f and g in the objective, we do not require any additional regularity assumption. g possibly nonsmooth, is assumed simple, i.e., the associated proximal mapping is easily computable. The affine constraint is addressed by the augmented Lagrangian approach. Our problem formulation is novel in several ways, starting with the posing of the problem over a general Hilbert space $\mathcal{H}$. We are also the first to consider a relative smoothness condition for f with a generalized curvature constant, which allows for a wider class of functions than contemporary works. Our generalized curvature constant also gives insight into how certain parameters, e.g., step sizes, should be chosen according to the regularity of f to maintain convergence guarantees. Furthermore, with such wide choice of algorithm parameters satisfying so called "open loop" rules, we are the first to rigorously prove convergence guarantees for the Lagrangian values, the feasibility gap, and for the sequence of dual variables for a conditional gradient algorithm with an affine constraint and augmented Lagrangian. Our main contributions and findings can be summarized as follows:

Main contributions of this chapter

- Convergence guarantees and rates for both the feasibility $(\|Ax_k - b\|)_{k\in\mathbb{N}}$ and the optimality $(\mathcal{L}(x_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star))_{k\in\mathbb{N}}$.
- Strong convergence of the sequence of dual variables $(\mu_k)_{k\in\mathbb{N}}$ to a solution $\mu^\star$ of the dual problem.
- A detailed outline of how to use our algorithm to solve problems as general as

$$\min_{x\in\bigcap_{i=1}^{n}\mathcal{C}_i} \{f(x) + g(Tx)\} \quad \text{and} \quad \min_{x\in\mathcal{H}} \left\{ f(x) + \sum_{i=1}^{n} g_i(T_i x) + h(x) \right\}$$

  in a separable way, i.e., utilizing individually the linear minimization oracle over each compact convex set $\mathcal{C}_i$ or utilizing individually the proximal operators $\operatorname{prox}_{g_i}$. This outline makes important use of a product space technique which induces an affine constraint $Ax = b$, highlighting the importance

of its inclusion in our problem. Our algorithm is the only provably convergent conditional gradient algorithm which handles the affine constraint *without* smoothing it using the Moreau envelope.

- A practical demonstration of the aforementioned outline implemented to solve a matrix completion problem involving both nuclear norm and ℓ^1 norm ball constraints with a nonsmooth data fidelity term. Such nuclear norm ball constraints are challenging for ordinary proximal algorithms, which we implement and compare to our algorithm, because projecting on the nuclear norm ball is much more computationally intense than the corresponding linear minimization oracle. Our practical findings match the rates of convergence predicted by the theorems.

The content of this chapter appeared in [108].

Contents

3.1 Introduction

3.1.1 Problem Statement

In this chapter, we consider the composite optimization problem,

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

where $\mathcal{H}_p, \mathcal{H}_d, \mathcal{H}_v$ are real Hilbert spaces (the subscripts p, d, and v denoting the "primal", the "dual" and an auxiliary space, respectively), endowed with the associated scalar products, $\langle\cdot,\cdot\rangle$, and norms, $\|\cdot\|$ (to be understood from the context), $A : \mathcal{H}_p \to \mathcal{H}_d$ and $T : \mathcal{H}_p \to \mathcal{H}_v$ are bounded linear operators, $b \in \mathcal{H}_d$ and f, g, h are proper, convex, and lower semi-continuous functions with $\mathcal{C} \overset{\text{def}}{=} \operatorname{dom}(h)$ being a weakly compact subset of $\mathcal{H}_p$. We allow for some *asymmetry* in regularity between the functions involved in the objective. While g is assumed to be prox-friendly, for h we assume that it is easy to compute a linearly-perturbed oracle (see (1.3.1)). On the other hand, f is assumed to be differentiable and satisfies a condition that generalizes Lipschitz-continuity of the gradient (see Definition 2.1.13).

Problem ($\mathscr{P}$) can be seen as a generalization of the classical Frank-Wolfe problem in [53] of minimizing a Lipschitz-smooth function f on a convex closed bounded subset $\mathcal{C} \subset \mathcal{H}_p$,

$$\min_{x\in\mathcal{H}_p} \{f(x) : \ x \in \mathcal{C}\} \tag{3.1.1}$$

In fact, if $A \equiv 0$, $b \equiv 0$, $g \equiv 0$, and $h \equiv \iota_{\mathcal{C}}$ is the indicator function of $\mathcal{C}$ then we recover exactly (3.1.1) from ($\mathscr{P}$).

3.1.2 Algorithm

As described in the previous section, we combine penalization with the augmented Lagrangian approach to form the following functional

$$\mathcal{J}_k(x,y,\mu) = f(x) + g(y) + h(x) + \langle \mu, Ax-b\rangle + \frac{\rho_k}{2}\|Ax-b\|^2 + \frac{1}{2\beta_k}\|y-Tx\|^2, \tag{3.1.2}$$

where μ is the dual multiplier, and ρ_k and β_k are non-negative parameters. The steps of our scheme, then, are summarized in Algorithm 8.

Algorithm 8: Conditional Gradient with Augmented Lagrangian and Proximal-step (CGALP)

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

$k = 0$

repeat

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

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

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

$x_{k+1} = x_k - \gamma_k(x_k - s_k)$

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

$k \leftarrow k+1$

until *convergence*;

Output: x_{k+1}.

For the interpretation of the algorithm, notice that the first step is equivalent to

$$\{y_k\} = \operatorname*{Argmin}_{y\in\mathcal{H}_v} \mathcal{J}_k(x_k, y, \mu_k).$$

Now define the functional $\mathcal{E}_k(x,\mu) \overset{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + \langle \mu, Ax-b\rangle + \frac{\rho_k}{2}\|Ax-b\|^2$. By convexity of the set $\mathcal{C}$ and the definition of x_{k+1} as a convex combination of x_k and s_k, the sequence $(x_k)_{k\in\mathbb{N}}$ remains in $\mathcal{C}$ for all k, although the affine constraint $Ax_k = b$ might only be satisfied asymptotically. It is an augmented Lagrangian, where we do not consider the non-differentiable function h and we replace g by its Moreau envelope. Notice that

$$\begin{aligned}\nabla_x \mathcal{E}_k(x,\mu_k) &= \nabla f(x) + T^*[\nabla g^{\beta_k}](Tx) + A^*\mu_k + \rho_k A^*(Ax-b)\\ &= \nabla f(x) + \frac{1}{\beta_k}T^*\left(Tx - \operatorname{prox}_{\beta_k g}(Tx)\right) + A^*\mu_k + \rho_k A^*(Ax-b)\end{aligned} \tag{3.1.3}$$

where in the second equality we used 2.1.2(iii). Then z_k is just $\nabla_x \mathcal{E}_k(x_k,\mu_k)$ and the first three steps of the algorithm can be condensed in

$$s_k \in \operatorname*{Argmin}_{s\in\mathcal{H}_p}\{h(s) + \langle \nabla_x \mathcal{E}_k(x_k,\mu_k), s\rangle\}. \tag{3.1.4}$$

Thus the primal variable update of each step of our algorithm boils down to conditional gradient applied to the function $\mathcal{E}_k(\cdot,\mu_k)$, where the next iterate is a convex combination between the previous one and the new direction s_k. A standard update of the Lagrange multiplier μ_k follows.

3.1.3 Assumptions

3.1.3.1 Assumptions on the functions

In order to help the reading, we recall in a compact form the following notation that we will use to refer to various functionals throughout the chapter:

$$\begin{aligned}
\Phi(x) &\overset{\text{def}}{=} f(x) + g(Tx) + h(x);\\
\phi_k(x) &\overset{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + h(x);\\
\Phi_k(x) &\overset{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + h(x) + \frac{\rho_k}{2}\|Ax-b\|^2;\\
\bar{\Phi}(x) &\overset{\text{def}}{=} \Phi(x) + (\overline{\rho}/2)\|Ax-b\|^2;\\
\bar{\varphi}(\mu) &\overset{\text{def}}{=} \bar{\Phi}^*(-A^*\mu) + \langle b,\ \mu\rangle;\\
\mathcal{L}(x,\mu) &\overset{\text{def}}{=} f(x) + g(Tx) + h(x) + \langle \mu, Ax-b\rangle;\\
\mathcal{L}_k(x,\mu) &\overset{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + h(x) + \langle \mu, Ax-b\rangle + \frac{\rho_k}{2}\|Ax-b\|^2;\\
\mathcal{E}_k(x,\mu) &\overset{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + \langle \mu, Ax-b\rangle + \frac{\rho_k}{2}\|Ax-b\|^2,
\end{aligned} \tag{3.1.5}$$

where $\overline{\rho}$ is defined in Assumption ($\mathbf{P}_4$) to be $\overline{\rho} = \sup_k \rho_k$.

In the list (3.1.5), we can recognize Φ as the objective, ϕ_k as the objective with smoothed g, Φ_k as the smoothed objective augmented with a quadratic penalization of the constraint, and $\mathcal{L}_k$ as a smoothed augmented Lagrangian. $\mathcal{L}$ denotes the classical Lagrangian. Recall that $(x^\star,\mu^\star) \in \mathcal{H}_p \times \mathcal{H}_d$ is a saddle-point for the Lagrangian $\mathcal{L}$ if for every $(x,\mu) \in \mathcal{H}_p \times \mathcal{H}_d$,

$$\mathcal{L}(x^\star,\mu) \leq \mathcal{L}(x^\star,\mu^\star) \leq \mathcal{L}(x,\mu^\star). \tag{3.1.6}$$

It is well-known from standard Lagrange duality, see e.g. [10, Proposition 19.19] or [92, Theorem 3.68], that the existence of a saddle point $(x^\star,\mu^\star)$ ensures strong duality, that $x^\star$ solves ($\mathscr{P}$) and $\mu^\star$ solves the dual problem,

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

The following assumptions on the problem will be used throughout the convergence analysis (for some results only a subset of these assumptions will be needed):

- ($\mathbf{A}_1$) $f, g\circ T$, and h belong to $\Gamma_0(\mathcal{H}_p)$.
- ($\mathbf{A}_2$) The pair $(f,\mathcal{C})$ is (F,ζ)-smooth (see Definition 2.1.13), where we recall $\mathcal{C} \overset{\text{def}}{=} \operatorname{dom}(h)$.
- ($\mathbf{A}_3$) $\mathcal{C}$ is weakly compact (and thus contained in a ball of radius $R>0$).
- ($\mathbf{A}_4$) $T\mathcal{C} \subset \operatorname{dom}(\partial g)$ and $\sup_{x\in\mathcal{C}} \left\|[\partial g(Tx)]^0\right\| \leq M < \infty$, where M is a positive constant.
- ($\mathbf{A}_5$) h is Lipschitz continuous relative to its domain $\mathcal{C}$ with constant $L_h \geq 0$, i.e., $\forall (x,z)\in\mathcal{C}^2$, $|h(x)-h(z)| \leq L_h\|x-z\|$.
- ($\mathbf{A}_6$) There exists a saddle-point $(x^\star,\mu^\star) \in \mathcal{H}_p \times \mathcal{H}_d$ for the Lagrangian $\mathcal{L}$.
- ($\mathbf{A}_7$) $\operatorname{ran}(A)$ is closed.
- ($\mathbf{A}_8$) One of the following holds:
  - **(I)** $A^{-1}(b) \cap \operatorname{int}(\operatorname{dom}(g\circ T)) \cap \operatorname{int}(\mathcal{C}) \neq \emptyset$, where $A^{-1}(b)$ is the pre-image of b under A.
  - **(II)** $\mathcal{H}_p$ and $\mathcal{H}_d$ are finite dimensional and

$$\begin{cases} A^{-1}(b) \cap \operatorname{ri}(\operatorname{dom}(g\circ T)) \cap \operatorname{ri}(\mathcal{C}) \neq \emptyset \\ \text{and} \\ \operatorname{ran}(A^*) \cap \operatorname{par}(\operatorname{dom}(g\circ T)\cap\mathcal{C})^\perp = \{0\}. \end{cases} \tag{3.1.7}$$

($\mathbf{A}_9$) The set-valued mappings $\left(\partial\left(\phi_k^* \circ (-A^*)\right)\right)_{k\in\mathbb{N}}$ satisfy the following property: for any sequence $\left((p_k, q_k)\right)_{k\in\mathbb{N}}$ satisfying, for each $k \in \mathbb{N}$,

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

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

At this stage, a few remarks are in order.

Remark 3.1.1.

(i) By Assumption ($\mathbf{A}_1$), $\mathcal{C}$ is also closed and convex. This together with Assumption ($\mathbf{A}_3$) entail, upon using [10, Lemma 3.29 and Theorem 3.32], that $\mathcal{C}$ is weakly compact.

(ii) Since the sequence of iterates $(x_k)_{k\in\mathbb{N}}$ generated by Algorithm 8 is guaranteed to belong to $\mathcal{C}$ under ($\mathbf{P}_1$), we have from ($\mathbf{A}_4$)

$$\sup_k \left\| [\partial g(Tx_k)]^0 \right\| \leq M \tag{3.1.8}$$

where M is a positive constant.

(iii) Assumption ($\mathbf{A}_5$) will only be needed in the proof of convergence to optimality (Theorem 3.3.3). It is not needed to show asymptotic feasibility (Theorem 3.3.2).

(iv) Assume that $A^{-1}(b) \cap \operatorname{dom}(g \circ T) \cap \mathcal{C} \neq \emptyset$, which entails that the set of minimizers of ($\mathscr{P}$) is a non-empty convex closed bounded set under ($\mathbf{A}_1$)-($\mathbf{A}_3$). Then there are various domain qualification conditions, e.g., one of the conditions in [10, Proposition 15.24 and Fact 15.25], that ensure the existence of a saddle-point for the Lagrangian $\mathcal{L}$ (see [10, Theorem 19.1 and Proposition 9.19(v)]).

(v) Observe that under the inclusion assumption of Lemma 2.1.17, ($\mathbf{A}_8$)(**I**) is equivalent to $A^{-1}(b) \cap \operatorname{int}(\mathcal{C}) \neq \emptyset$.

(vi) Assumption ($\mathbf{A}_8$) will be crucial to show that $\bar{\varphi}$ is coercive on $\ker(A^*)^\perp = \operatorname{ran}(A)$ (the last equality follows from ($\mathbf{A}_7$)), and hence boundedness of the dual multiplier sequence $(\mu_k)_{k\in\mathbb{N}}$ provided by Algorithm 8 (see Lemma 3.2.6 and Lemma 3.2.7).

(vii) If the dimension of $\mathcal{H}_d$ is finite, then ($\mathbf{A}_9$) is satisfied because weakly compact sets are compact in such spaces. Alternatively, another sufficient condition is to impose that the sublevel sets of the functions $\left(\Phi_k^* \circ (-A^*)\right)_{k\in\mathbb{N}}$ are compact, for instance if the functions are uniformly convex, unfiormly in k.

The uniform boundedness of the minimal norm selection of ∂g on $\mathcal{C}$, as required in Assumption ($\mathbf{A}_4$), is important when we will invoke Proposition 2.1.2(v) in our proofs to get meaningful estimates. We recall that Lemma 2.1.17 gives some sufficient conditions under which ($\mathbf{A}_4$) holds (in fact even stronger claims) for g. An immediate consequence of assuming ($\mathbf{A}_1$) and ($\mathbf{A}_2$) due to Lemma 2.1.18 is that

$$\sup_k \|\nabla f(x_k)\| \leq D < +\infty \tag{3.1.9}$$

for $(x_k)_{k\in\mathbb{N}}$ generated by Algorithm 8.

3.1.3.2 Assumptions on the parameters

We also use the following assumptions on the parameters of Algorithm 8 (recall the function ζ in Definition 2.1.13):

($\mathbf{P}_1$) $(\gamma_k)_{k\in\mathbb{N}} \subset]0, 1]$ and the sequences $\left(\zeta(\gamma_k)\right)_{k\in\mathbb{N}}$, $\left(\gamma_k^2/\beta_k\right)_{k\in\mathbb{N}}$ and $(\gamma_k\beta_k)_{k\in\mathbb{N}}$ belong to ℓ_+^1.

($\mathbf{P}_2$) $(\gamma_k)_{k\in\mathbb{N}} \notin \ell^1$.

($\mathbf{P}_3$) $(\beta_k)_{k\in\mathbb{N}} \in \ell_+$ is non-increasing and converges to 0.

($\mathbf{P}_4$) $(\rho_k)_{k\in\mathbb{N}} \in \ell_+$ is non-decreasing with $0 < \underline{\rho} = \inf_k \rho_k \leq \sup_k \rho_k = \overline{\rho} < +\infty$.

($\mathbf{P}_5$) For some positive constants $\underline{M}$ and $\overline{M}$, $\underline{M} \leq \inf_k (\gamma_k/\gamma_{k+1}) \leq \sup_k (\gamma_k/\gamma_{k+1}) \leq \overline{M}$.

($\mathbf{P}_6$) $(\theta_k)_{k\in\mathbb{N}}$ satisfies $\theta_k = \frac{\gamma_k}{c}$ for all $k \in \mathbb{N}$ for some $c > 0$ such that $\frac{\overline{M}}{c} - \frac{\rho}{2} < 0$.

($\mathbf{P}_7$) $(\gamma_k)_{k\in\mathbb{N}}$ 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 all $k \in \mathbb{N}$ and for c in ($\mathbf{P}_6$).

Remark 3.1.2.

(i) One can recognize that the update of the dual multiplier μ_k in Algorithm 8 has a flavour of gradient ascent applied to the augmented dual with step-size θ_k. However, unlike the standard method of multipliers with the augmented Lagrangian, Assumption ($\mathbf{P}_6$) requires θ_k to vanish in our setting. The underlying reason is that our update can be seen as an inexact dual ascent (i.e., exactness stems from the conditional gradient-based update on x_k which is not a minimization of over x of the augmented Lagrangian $\mathcal{L}_k$). Thus θ_k must annihilate this error asymptotically.

(ii) A sufficient condition for ($\mathbf{P}_7$) to hold consists of taking $\rho_k \equiv \rho > 0$ and $\gamma_{k+1} \geq \frac{2}{c(1+\rho)}\gamma_k$. In particular, if $(\gamma_k)_{k\in\mathbb{N}}$ satisfies ($\mathbf{P}_5$), then, for ($\mathbf{P}_7$) to hold, it is sufficient to take $\rho_k \equiv \rho > 2\overline{M}/c$ as supposed in ($\mathbf{P}_6$).

(iii) The relevance of having ρ_k vary is that it allows for more general and less stringent choice of the step-size γ_k. It is, however, possible (and easier in practice), to simply pick $\rho_k \equiv \rho$ for all $k \in \mathbb{N}$ as described above.

There is a large class of sequences that fulfill the requirements ($\mathbf{P}_1$)-($\mathbf{P}_7$). A typical one is as follows.

Example 3.1.3. Take[1], for $k \in \mathbb{N}$,

$$\rho_k \equiv \rho > 0, \gamma_k = \frac{(\log(k+2))^a}{(k+1)^{1-b}}, \beta_k = \frac{1}{(k+1)^{1-\delta}}, \quad \text{with}$$
$$a \geq 0, 0 \leq 2b < \delta < 1, \delta < 1-b, \rho > 2^{2-b}/c, c > 0.$$

In this case, one can take the crude bounds $\underline{M} = (\log(2)/\log(3))^a$ and $\overline{M} = 2^{1-b}$, and choose $\rho > 2\overline{M}/c$ as devised in Remark 3.1.2(ii). In turn, ($\mathbf{P}_4$)-($\mathbf{P}_7$) hold. In addition, suppose that f has a ν-Hölder continuous gradient (see (2.1.7)). Thus for ($\mathbf{P}_1$)-($\mathbf{P}_2$) to hold, simple algebra shows that the allowable choice of b is in $\left[0, \min\left(1/3, \frac{\nu}{1+\nu}\right)\right[$.

3.1.4 Organization of the Chapter

In Section 3.2 we present the preliminary estimations that will be used to prove the main convergence results that are given in Section 3.3. Section 3.2 is divided in three main parts, the feasibility, the boundedness of $(\mu_k)_{k\in\mathbb{N}}$, and then finally the optimality. In Section 3.3 we show the asymptotic feasibility, i.e., the limit $\lim_k \|Ax_k - b\| = 0$, and finally the optimality guarantees, i.e., strong convergence of the sequence $(\mu_k)_{k\in\mathbb{N}}$ to a solution of the dual problem, weak subsequential convergence of the sequence $(x_k)_{k\in\mathbb{N}}$ to a solution of the primal problem, and convergence of the Lagrangian values, and with convergence rates. In Section 3.4 we provide a more detailed discussion comparing CGALP to contemporary work on similar algorithms. Meanwhile, in Section 3.5 we describe how our algorithm can be instantiated to solve a variety of composite optimization problems, demonstrating how the inclusion of an affine constraint $Ax = b$ in the problem formulation allows one to lift otherwise unwieldy problems to a tractable form. Finally, in Section 3.6, numerical results are reported on two different problem, utilizing the aforementioned lifting scheme in the matrix completion problem.

[1] Of course, one can add a scaling factor in the choice of the parameters which would allow for more practical flexibility. But this does not change anything to our discussion nor to the bahviour of the CGALP algorithm for k large enough.

3.2 Preliminary Estimations

3.2.1 Preparatory Results

The next result is a direct application of the Descent Lemma 2.1.7 and the generalized one in Lemma 2.1.12 to the specific case of Algorithm 8. It allows to obtain a descent property for the function $\mathcal{E}_k(\cdot, \mu_k)$ between the previous iterate x_k and next one x_{k+1}.

Lemma 3.2.1. *Suppose Assumptions* $(\mathbf{A}_1)$, $(\mathbf{A}_2)$ *and* $(\mathbf{P}_1)$ *hold. For each* $k \in \mathbb{N}$, *define the quantity*

$$L_k \stackrel{\text{def}}{=} \frac{\|T\|^2}{\beta_k} + \|A\|^2 \rho_k. \tag{3.2.1}$$

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

$$\begin{aligned}\mathcal{E}_k(x_{k+1}, \mu_k) \leq \mathcal{E}_k(x_k, \mu_k) + \langle \nabla_x \mathcal{E}_k(x_k, \mu_k),\ x_{k+1} - x_k \rangle + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_k) \\ + \frac{L_k}{2}\|x_{k+1} - x_k\|^2.\end{aligned}$$

Proof. Define for each $k \in \mathbb{N}$,

$$\tilde{\mathcal{E}}_k(x, \mu) \stackrel{\text{def}}{=} g^{\beta_k}(Tx) + \langle \mu, Ax - b \rangle + \frac{\rho_k}{2}\|Ax - b\|^2,$$

so that $\mathcal{E}_k(x, \mu) = f(x) + \tilde{\mathcal{E}}_k(x, \mu)$. Compute

$$\nabla_x \tilde{\mathcal{E}}_k(x, \mu) = T^* \nabla g^{\beta_k}(Tx) + A^* \mu + \rho_k A^*(Ax - b),$$

which is Lipschitz-continuous with constant $L_k = \frac{\|T\|^2}{\beta_k} + \|A\|^2 \rho_k$ by virtue of $(\mathbf{A}_1)$ and Proposition 2.1.2(iii). Then we can use the Descent Lemma (2.1.7) with $\nu = 1$ on $\tilde{\mathcal{E}}_k(\cdot, \mu_k)$ between the points x_k and x_{k+1}, to obtain, for each $k \in \mathbb{N}$,

$$\tilde{\mathcal{E}}_k(x_{k+1}, \mu_k) \leq \tilde{\mathcal{E}}_k(x_k, \mu_k) + \langle \nabla \tilde{\mathcal{E}}_k(x_k, \mu_k), x_{k+1} - x_k \rangle + \frac{L_k}{2}\|x_{k+1} - x_k\|^2. \tag{3.2.2}$$

From Assumption $(\mathbf{A}_2)$, Lemma 2.1.12 and Remark 2.1.14, we have, for each $k \in \mathbb{N}$,

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + D_F(x_{k+1}, x_k) \\ &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_k), \end{aligned}$$

where we used that both x_k and s_k lie in $\mathcal{C}$, that γ_k belongs to $]0, 1]$ by $(\mathbf{P}_1)$ and thus $x_{k+1} = x_k + \gamma_k(s_k - x_k) \in \mathcal{C}$. Summing (3.2.2) with the latter and recalling that $\mathcal{E}_k(x, \mu_k) = f(x) + \tilde{\mathcal{E}}_k(x, \mu_k)$, we obtain the claim. □

Again for the function $\mathcal{E}_k(\cdot, \mu_k)$, we also have a lower-bound, presented in the next lemma.

Lemma 3.2.2. *Suppose Assumptions* $(\mathbf{A}_1)$ *and* $(\mathbf{A}_2)$ *hold. Then, for all* $k \in \mathbb{N}$, *for all* $x, x' \in \mathcal{H}_p$ *and for all* $\mu \in \mathcal{H}_d$,

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

Proof. First, split the function $\mathcal{E}_k(\cdot, \mu)$ as $\mathcal{E}_k(x, \mu) = \mathcal{E}_k^0(x, \mu) + \frac{\rho_k}{2}\|Ax - b\|^2$ for an opportune definition of $\mathcal{E}_k^0(\cdot, \mu)$. For the first term, simply by convexity, we have

$$\mathcal{E}_k^0(x, \mu) \geq \mathcal{E}_k^0(x', \mu) + \langle \nabla_x \mathcal{E}_k^0(x', \mu),\ x - x' \rangle. \tag{3.2.3}$$

Now use the strong convexity of the term $(\rho_k/2)\|\cdot - b\|^2$ between points Ax and Ax', to affirm that

$$\frac{\rho_k}{2}\|Ax - b\|^2 \geq \frac{\rho_k}{2}\|Ax' - b\|^2 + \left\langle \nabla\left(\frac{\rho_k}{2}\|\cdot - b\|^2\right)(Ax'), Ax - Ax' \right\rangle + \frac{\rho_k}{2}\|A(x - x')\|^2. \tag{3.2.4}$$

Compute

$$\begin{aligned}\left\langle \nabla\left(\frac{\rho_k}{2}\|\cdot - b\|^2\right)(Ax'), Ax - Ax' \right\rangle &= \rho_k \langle A^*(Ax' - b), x - x' \rangle \\ &= \left\langle \nabla\left(\frac{\rho_k}{2}\|A\cdot - b\|^2\right)(x'), x - x' \right\rangle.\end{aligned}$$

Summing (3.2.3) and (3.2.4) and invoking the gradient computation above, we obtain the claim. □

Lemma 3.2.3. *Suppose that assumptions* $(\mathbf{A}_1)$-$(\mathbf{A}_8)$ *and* $(\mathbf{P}_1)$-$(\mathbf{P}_7)$ *hold, with* $\underline{M} \geq 1$. *Let* $(x_k)_{k\in\mathbb{N}}$ *be the sequence of primal iterates generated by* Algorithm 8 *and* $\mu^\star$ *a solution of the dual problem* $(\mathscr{D})$. *Then we have the following estimate,*

$$\mathcal{L}(x_k, \mu^\star) - \mathcal{L}(x_{k+1}, \mu^\star) \leq \gamma_k d_{\mathcal{C}} \left(M \|T\| + D + L_h + \|A\| \, \|\mu^\star\|\right).$$

Proof. First define $u_k \overset{\text{def}}{=} [\partial g(Tx_k)]^0$ and recall that, by $(\mathbf{A}_2)$ and $(\mathbf{A}_4)$ and there consequences in (3.1.8) and (3.1.9), $\|u_k\| \leq M$ and $\|\nabla f(x_k)\| \leq D$ for every $k \in \mathbb{N}$. Then,

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

where we used the subdifferential inequality (2.1.4) on g, the gradient inequality on f, the L_h-Lipschitz continuity of h relative to $\mathcal{C}$ (see $(\mathbf{A}_5)$), and the Cauchy-Schwartz inequality on the scalar product. Since $x_{k+1} = x_k + \gamma_k(x_k - s_k)$, we obtain

$$\begin{aligned}\mathcal{L}(x_k, \mu^\star) - \mathcal{L}(x_{k+1}, \mu^\star) &\leq \gamma_k \Big(\langle u_k,\ T(x_k - s_k)\rangle + \langle \nabla f(x_k),\ x_k - s_k\rangle + L_h \|x_k - s_k\| \\ &\quad + \|\mu^\star\| \, \|A\| \, \|x_k - s_k\| \Big) \\ &\leq \gamma_k d_{\mathcal{C}} \left(M\|T\| + D + L_h + \|\mu^\star\| \, \|A\|\right).\end{aligned}$$

□

Lemma 3.2.4. *Suppose that assumptions* $(\mathbf{A}_3)$ *and* $(\mathbf{P}_4)$ *hold. Let* $(x_k)_{k\in\mathbb{N}}$ *be the sequence of primal iterates generated by* Algorithm 8. *Then we have the following estimate,*

$$\frac{\rho_k}{2} \|Ax_k - b\|^2 - \frac{\rho_{k+1}}{2} \|Ax_{k+1} - b\|^2 \leq \overline{\rho} d_{\mathcal{C}} \|A\| \left(\|A\|R + \|b\|\right) \gamma_k,$$

where R *is the radius of the ball containing* $\mathcal{C}$ *and* $\overline{\rho} = \sup_k \rho_k$.

Proof. By $(\mathbf{P}_4)$ and convexity of the function $\frac{\rho_{k+1}}{2}\|A \cdot - b\|^2$, we have

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

Now compute the gradient and use the definition of x_{k+1}, to obtain

$$\begin{aligned}\frac{\rho_k}{2}\|Ax_k - b\|^2 - \frac{\rho_{k+1}}{2}\|Ax_{k+1} - b\|^2 &\leq \rho_{k+1}\gamma_k \langle Ax_k - b,\ A(x_k - s_k)\rangle \\ &\leq \overline{\rho} d_{\mathcal{C}} \|A\| \left(\|A\|R + \|b\|\right) \gamma_k.\end{aligned}$$

In the last inequality, we used Cauchy-Schwartz inequality, triangle inequality, the fact that $\|x_k - s_k\| \leq d_{\mathcal{C}}$, and assumptions $(\mathbf{A}_3)$ and $(\mathbf{P}_4)$ (respectively, $\sup_{x\in\mathcal{C}} \|x\| \leq R$ and $\rho_{k+1} \leq \overline{\rho}$). □

3.2.2 Feasibility Esimation

We proceed with an intermediary lemma establishing the main feasibility estimation and some summability results that will also be used in the main energy estimation used in the proof of optimality.

Lemma 3.2.5. *Suppose that Assumptions* $(\mathbf{A}_1)$-$(\mathbf{A}_4)$ *and* $(\mathbf{A}_6)$ *hold. Consider the sequence of iterates* $(x_k)_{k\in\mathbb{N}}$ *from Algorithm 8 with parameters satisfying Assumptions* $(\mathbf{P}_1)$-$(\mathbf{P}_6)$. *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* $\tilde{\mathcal{L}} \overset{\text{def}}{=} \mathcal{L}(x^\star,\mu^\star)$. *Denote the sum* $\Delta_k \overset{\text{def}}{=} \Delta_k^p + \Delta_k^d$. *Then we have the following estimation,*

$$\Delta_{k+1} \leq \Delta_k - \gamma_{k+1}\left(\frac{M}{c}\left\|A\tilde{x}_{k+1}-b\right\|^2 + \delta\left\|A\left(x_{k+1}-\tilde{x}_{k+1}\right)\right\|^2\right) + \frac{L_{k+1}}{2}\gamma_{k+1}^2 d_{\mathcal{C}}^2$$
$$+K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_{k+1}\right) + \frac{\beta_k-\beta_{k+1}}{2}M + \left(\frac{\rho_{k+1}-\rho_k}{2}\right)\left\|Ax_{k+1}-b\right\|^2,$$

and, moreover,

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

Proof. First notice that 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 bounded from below by our assumptions. Indeed, in view of $(\mathbf{A}_1)$, $(\mathbf{A}_6)$ and Remark 3.1.1(iv), $\tilde{\mathcal{L}}_k(\mu_k)$ is bounded from above since

$$\begin{aligned}\tilde{\mathcal{L}}_k(\mu_k) &\leq \mathcal{L}_k(x^\star,\mu_k)\\ &= f(x^\star) + g^{\beta_k}(Tx^\star) + h(x^\star) + \langle\mu_k, Ax^\star-b\rangle + \frac{\rho_k}{2}\left\|Ax^\star-b\right\|^2\\ &= f(x^\star) + g^{\beta_k}(Tx^\star) + h(x^\star)\\ &\leq f(x^\star) + g(Tx^\star) + h(x^\star) < +\infty,\end{aligned}$$

where we used Proposition 2.1.2(v) in the last inequality.

We denote a minimizer of $\mathcal{L}_k(x,\mu_k)$ by $\tilde{x}_k \in \operatorname{Argmin}_{x\in\mathcal{H}_p} \mathcal{L}_k(x,\mu_k)$, which exists and belongs to $\mathcal{C}$ by $(\mathbf{A}_1)$-$(\mathbf{A}_3)$. Then, we have, for each $k\in\mathbb{N}$,

$$\Delta_{k+1}^d - \Delta_k^d = \mathcal{L}_k(\tilde{x}_k,\mu_k) - \mathcal{L}_{k+1}(\tilde{x}_{k+1},\mu_{k+1}). \tag{3.2.5}$$

Since $\tilde{x}_k$ is a minimizer of $\mathcal{L}_k(x,\mu_k)$ we have that $\mathcal{L}_k(\tilde{x}_k,\mu_k) \leq \mathcal{L}_k(\tilde{x}_{k+1},\mu_k)$ which leads to

$$\begin{aligned}\mathcal{L}_k(\tilde{x}_{k+1},\mu_k) &= \mathcal{L}_{k+1}(\tilde{x}_{k+1},\mu_k) + g^{\beta_k}(T\tilde{x}_{k+1}) - g^{\beta_{k+1}}(T\tilde{x}_{k+1}) + \tfrac{\rho_k-\rho_{k+1}}{2}\left\|A\tilde{x}_{k+1}-b\right\|^2\\ &\leq \mathcal{L}_{k+1}(\tilde{x}_{k+1},\mu_k),\end{aligned}$$

where the last inequality comes from Proposition 2.1.2(v) and the assumptions $(\mathbf{P}_3)$ and $(\mathbf{P}_4)$. Combining this with (3.2.5), for each $k\in\mathbb{N}$,

$$\begin{aligned}\Delta_{k+1}^d - \Delta_k^d &\leq \mathcal{L}_{k+1}(\tilde{x}_{k+1},\mu_k) - \mathcal{L}_{k+1}(\tilde{x}_{k+1},\mu_{k+1})\\ &= \langle\mu_k-\mu_{k+1}, A\tilde{x}_{k+1}-b\rangle\\ &= -\theta_k\langle Ax_{k+1}-b, A\tilde{x}_{k+1}-b\rangle,\end{aligned} \tag{3.2.6}$$

where in the last equality we used the definition of μ_{k+1}. Meanwhile, for the primal gap we have, for each $k\in\mathbb{N}$,

$$\Delta_{k+1}^p - \Delta_k^p = \left(\mathcal{L}_{k+1}(x_{k+2},\mu_{k+1}) - \mathcal{L}_k(x_{k+1},\mu_k)\right) + \left(\mathcal{L}_k(\tilde{x}_k,\mu_k) - \mathcal{L}_{k+1}(\tilde{x}_{k+1},\mu_{k+1})\right).$$

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

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

and estimate $\mathcal{L}_k(\tilde{x}_k,\mu_k) - \mathcal{L}_{k+1}(\tilde{x}_{k+1},\mu_{k+1})$ as in (3.2.6), to get

$$\Delta_{k+1}^p - \Delta_k^p \leq \mathcal{L}_{k+1}(x_{k+2},\mu_{k+1}) - \mathcal{L}_k(x_{k+1},\mu_{k+1}) + \theta_k\left\|Ax_{k+1}-b\right\|^2 - \theta_k\langle Ax_{k+1}-b, A\tilde{x}_{k+1}-b\rangle. \tag{3.2.7}$$

Using (3.2.6) and (3.2.7), we then have, for each $k\in\mathbb{N}$,

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

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

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

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

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

We denote by $\mathbf{T1} = \mathcal{L}_{k+1}\left(x_{k+2}, \mu_{k+1}\right) - \mathcal{L}_{k+1}\left(x_{k+1}, \mu_{k+1}\right)$ and the remaining part of the right-hand side by **T2**. For the moment, we focus our attention on **T1**. Recall that $\mathcal{L}_k\left(x, \mu_k\right) = \mathcal{E}_k\left(x, \mu_k\right) + h\left(x\right)$ and apply Lemma 3.2.1 between points x_{k+2} and x_{k+1}, to get

$$\begin{aligned}\mathbf{T1} \leq{}& h\left(x_{k+2}\right) - h\left(x_{k+1}\right) + \left\langle\nabla_x \mathcal{E}_{k+1}\left(x_{k+1}, \mu_{k+1}\right), x_{k+2} - x_{k+1}\right\rangle\\ &+ K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_{k+1}\right) + \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2.\end{aligned}$$

By $(\mathbf{A}_1)$ we have that h is convex and thus, since x_{k+2} is a convex combination of x_{k+1} and s_{k+1}, we get

$$\begin{aligned}\mathbf{T1} \leq{}& -\gamma_{k+1}\left(h\left(x_{k+1}\right) - h\left(s_{k+1}\right) + \left\langle\nabla_x \mathcal{E}_{k+1}\left(x_{k+1}, \mu_{k+1}\right), x_{k+1} - s_{k+1}\right\rangle\right)\\ &+ \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_{k+1}\right).\end{aligned}$$

Applying the definition of s_k as the minimizer of the linear minimization oracle and Lemma 3.2.2 at the points $\tilde{x}_{k+1}$, x_{k+1}, and μ_{k+1} gives,

$$\begin{aligned}\mathbf{T1} &\leq -\gamma_{k+1}\left(h\left(x_{k+1}\right) - h\left(\tilde{x}_{k+1}\right) + \left\langle\nabla_x \mathcal{E}_{k+1}\left(x_{k+1}, \mu_{k+1}\right), x_{k+1} - \tilde{x}_{k+1}\right\rangle\right)\\ &\quad + \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_{k+1}\right)\\ &\leq -\gamma_{k+1}\Big(h\left(x_{k+1}\right) - h\left(\tilde{x}_{k+1}\right) + \mathcal{E}_{k+1}\left(x_{k+1}, \mu_{k+1}\right) - \mathcal{E}_{k+1}\left(\tilde{x}_{k+1}, \mu_{k+1}\right)\\ &\quad + \frac{\rho_{k+1}}{2}\left\|A\left(x_{k+1} - \tilde{x}_{k+1}\right)\right\|^2\Big) + \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_{k+1}\right)\\ &= -\gamma_{k+1}\left(\mathcal{L}_{k+1}\left(x_{k+1}, \mu_{k+1}\right) - \mathcal{L}_{k+1}\left(\tilde{x}_{k+1}, \mu_{k+1}\right) + \frac{\rho_{k+1}}{2}\left\|A\left(x_{k+1} - \tilde{x}_{k+1}\right)\right\|^2\right)\\ &\quad + \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_{k+1}\right)\\ &\leq -\frac{\gamma_{k+1}\rho_{k+1}}{2}\left\|A\left(x_{k+1} - \tilde{x}_{k+1}\right)\right\|^2 + \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_{k+1}\right),\end{aligned}$$

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 $k \in \mathbb{N}$,

$$\begin{aligned}\Delta_{k+1} - \Delta_k \leq{}& -\theta_k\left\|A\tilde{x}_{k+1} - b\right\|^2 + \left(\theta_k - \gamma_{k+1}\frac{\rho_{k+1}}{2}\right)\left\|A\left(x_{k+1} - \tilde{x}_{k+1}\right)\right\|^2\\ &+ \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_{k+1}\right) + \left[g^{\beta_{k+1}} - g^{\beta_k}\right]\left(Tx_{k+1}\right)\\ &+ \frac{\rho_{k+1} - \rho_k}{2}\left\|Ax_{k+1} - b\right\|^2. \quad (3.2.8)\end{aligned}$$

Under $(\mathbf{P}_6)$ we have $\theta_k = \frac{\gamma_k}{c}$ 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 $\gamma_k \leq \overline{M}\gamma_{k+1}$ (see Assumption $(\mathbf{P}_5)$). Then, using $(\mathbf{P}_5)$ and the above inequality, for each $k \in \mathbb{N}$,

$$\theta_k - \gamma_{k+1}\frac{\rho_{k+1}}{2} \leq \left(\frac{\overline{M}}{c} - \frac{\rho_{k+1}}{2}\right)\gamma_{k+1} \leq \left(\frac{\overline{M}}{c} - \frac{\underline{\rho}}{2}\right)\gamma_{k+1} = -\delta\gamma_{k+1} \text{ and } \theta_k \geq \tfrac{\underline{M}\gamma_{k+1}}{c}. \quad (3.2.9)$$

Now use the fact that, for each $k \in \mathbb{N}$, $x_{k+2} = x_{k+1} + \gamma_{k+1}(s_{k+1} - x_{k+1})$ to estimate

$$\|x_{k+2} - x_{k+1}\|^2 \leq \gamma_{k+1}^2 d_{\mathcal{C}}^2. \tag{3.2.10}$$

Moreover, by the two assumptions $(\mathbf{P}_3)$, $(\mathbf{A}_4)$ and Proposition 2.1.2(v), (3.1.8) holds with a constant $M > 0$, and thus with Proposition 2.1.2(iv) we obtain, for each $k \in \mathbb{N}$,

$$\left[g^{\beta_{k+1}} - g^{\beta_k}\right](Tx_{k+1}) \leq \frac{\beta_k - \beta_{k+1}}{2} \left\| \left[\partial g\left(Tx_{k+1}\right)\right]^0 \right\|^2 \leq \frac{\beta_k - \beta_{k+1}}{2} M. \tag{3.2.11}$$

Plugging (3.2.9), (3.2.10) and (3.2.11) into (3.2.8), we get, for each $k \in \mathbb{N}$,

$$\begin{aligned} \Delta_{k+1} - \Delta_k \leq & -\frac{M}{c}\gamma_{k+1}\|A\tilde{x}_{k+1} - b\|^2 - \delta\gamma_{k+1}\|A(x_{k+1} - \tilde{x}_{k+1})\|^2 + \frac{L_{k+1}}{2}\gamma_{k+1}^2 d_{\mathcal{C}}^2 \\ & + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_{k+1}) + \frac{\beta_k - \beta_{k+1}}{2} M + \left(\frac{\rho_{k+1} - \rho_k}{2}\right)\|Ax_{k+1} - b\|^2. \end{aligned} \tag{3.2.12}$$

Because of the assumptions $(\mathbf{P}_1)$ and $(\mathbf{P}_4)$, and in view of the definition of L_k in (3.2.1), we have the following,

$$\frac{L_k}{2}\gamma_k^2 d_{\mathcal{C}}^2 = \frac{1}{2}\left(\frac{\|T\|^2}{\beta_k} + \|A\|^2 \rho_k\right)\gamma_k^2 d_{\mathcal{C}}^2 \in \ell_+^1.$$

For the telescopic terms from the right hand side of (3.2.12) we have

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

where R is the constant arising from $(\mathbf{A}_3)$. Under $(\mathbf{P}_1)$ we also have that

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

Using the notation of Lemma 2.2.3, we set, for each $k \in \mathbb{N}$,

$$\begin{aligned} r_k &= \Delta_k, \quad p_k = \gamma_{k+1}, \quad w_k = \left(\frac{M}{c}\|A\tilde{x}_{k+1} - b\|^2 + \delta\|A(x_{k+1} - \tilde{x}_{k+1})\|^2\right), \\ z_k &= \frac{L_{k+1}}{2}\gamma_{k+1}^2 d_{\mathcal{C}}^2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_{k+1}) + \frac{\beta_k - \beta_{k+1}}{2} M + \left(\frac{\rho_{k+1} - \rho_k}{2}\right)\|Ax_{k+1} - b\|^2. \end{aligned}$$

We have shown above that

$$r_{k+1} \leq r_k - p_k w_k + z_k,$$

where $(z_k)_{k\in\mathbb{N}} \in \ell_+^1$, and r_k is bounded from below. We then deduce using Lemma 2.2.3(i) that $(r_k)_{k\in\mathbb{N}}$ is convergent and

$$\left(\gamma_k\|A\tilde{x}_k - b\|^2\right)_{k\in\mathbb{N}} \in \ell_+^1, \quad \left(\gamma_k\|A(x_k - \tilde{x}_k)\|^2\right)_{k\in\mathbb{N}} \in \ell_+^1. \tag{3.2.13}$$

Consequently,

$$\left(\gamma_k\|Ax_k - b\|^2\right)_{k\in\mathbb{N}} \in \ell_+^1, \tag{3.2.14}$$

since, by Jensen's inequality,

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

□

3.2.3 Boundedness of $(\mu_k)_{k\in\mathbb{N}}$

In the following two lemmas, we provide an argument that shows the sequence of dual variables $(\mu_k)_{k\in\mathbb{N}}$ generated by Algorithm 8 is bounded. We start by studying coercivity of $\bar{\varphi}$.

Lemma 3.2.6. *Suppose that Assumptions $(\mathbf{A}_1)$-$(\mathbf{A}_3)$ and $(\mathbf{A}_6)$-$(\mathbf{A}_8)$ hold. Then $\bar{\varphi}$ is coercive on* $\operatorname{ran}(A)$.

Proof. From (3.1.5), we have, for any $c \in A^{-1}(b)$, that

$$\bar{\varphi}(\mu) = \left(\bar{\Phi}^* + \langle -c, \cdot \rangle\right)(-A^*\mu).$$

Moreover, Assumptions $(\mathbf{A}_1)$ and $(\mathbf{A}_7)$ entail that $\bar{\Phi} \in \Gamma_0(\mathcal{H}_p)$. We now consider separately the two assumptions.

(a) Case of $(\mathbf{A}_8)$**(I)**: If follows from the Fenchel-Moreau theorem ([10, Theorem 13.32]) that

$$\left(\bar{\Phi}^* - \langle c, \cdot \rangle\right)^* = \bar{\Phi}^{**}(\cdot + c) = \bar{\Phi}(\cdot + c).$$

Using this, together with Proposition 2.1.20 and $(\mathbf{A}_2)$, we can assert that $\bar{\Phi}^* - \langle c, \cdot \rangle$ is coercive if and only if

$$\begin{aligned} 0 \in \operatorname{int}\left(\operatorname{dom}\left(\bar{\Phi}(\cdot + c)\right)\right) = \operatorname{int}(\operatorname{dom}(\Phi)) - c &= \operatorname{int}(\operatorname{dom}(g \circ T) \cap \mathcal{C}) - c \\ &= \operatorname{int}(\operatorname{dom}(g \circ T)) \cap \operatorname{int}(\mathcal{C}) - c. \end{aligned}$$

But this is precisely what $(\mathbf{A}_8)$**(I)** guarantees. In turn, using [10, Proposition 14.15], $(\mathbf{A}_8)$**(I)** is equivalent to

$$\exists (a > 0, \beta \in \mathbb{R}), \quad \bar{\Phi}^* - \langle c, \cdot \rangle \geq a \|\cdot\| + \beta.$$

Using standard results on linear operators in Hilbert spaces [10, Facts 2.18 and 2.19], we have

$$(\mathbf{A}_7) \iff (\exists \alpha > 0), (\forall \mu \in \operatorname{ran}(A)), \quad \|A^*\mu\| \geq \alpha \|\mu\|.$$

Combining the last two inequalities, we deduce that under $(\mathbf{A}_8)$**(I)**,

$$\exists (a > 0, \alpha > 0, \beta \in \mathbb{R}), (\forall \mu \in \operatorname{ran}(A)), \quad \bar{\varphi}(\mu) \geq a \|A^*\mu\| + \beta \geq a\alpha \|\mu\| + \beta,$$

which in turn is equivalent to coercivity of $\bar{\varphi}$ on $\operatorname{ran}(A)$ by [10, Proposition 14.15].

(b) Case of $(\mathbf{A}_8)$**(II)**: Since $\mathcal{H}_d$ is finite dimensional, We have, $\forall u \in \mathcal{H}_d$,

$$\begin{aligned} \bar{\varphi}^\infty(u) &= \left(\left(\bar{\Phi}^* + \langle -c, \cdot \rangle\right) \circ (-A^*)\right)^\infty(u) \\ \text{(Proposition 2.1.21(iii))} \quad &= \left(\bar{\Phi}^* + \langle -c, \cdot \rangle\right)^\infty(-A^*u) \\ \text{(Proposition 2.1.21(ii))} \quad &= \sigma_{\operatorname{dom}\left(\bar{\Phi}^* + \langle -c, \cdot \rangle\right)^*}(-A^*u) \\ &= \sigma_{\operatorname{dom}\left(\bar{\Phi}(\cdot + c)\right)}(-A^*u) \\ &= \sigma_{\operatorname{dom}(\bar{\Phi}) - c}(-A^*u) \\ \text{(by } (\mathbf{A}_2)) \quad &= \sigma_{\operatorname{dom}(g \circ T) \cap \mathcal{C} - c}(-A^*u). \end{aligned}$$

Notice that, by Assumption $(\mathbf{A}_4)$, we have $\operatorname{dom}(g \circ T) \cap \mathcal{C} = \mathcal{C}$. Thus, using Proposition 2.1.21(i), we have the following chain of equivalences

$$\begin{aligned} \bar{\varphi} \text{ is coercive on } \operatorname{ran}(A) &\iff \bar{\varphi}^\infty(u) > 0, \quad \forall u \in \operatorname{ran}(A) \setminus \{0\} \\ &\iff \sigma_{\mathcal{C} - c}(-A^*u) > 0, \quad \forall u \in \operatorname{ran}(A) \setminus \{0\}. \end{aligned}$$

For this to hold, and since $\operatorname{ran}(A) = \ker(A^*)^\perp$, a sufficient condition is that

$$\sigma_{\mathcal{C} - c}(x) > 0, \quad \forall x \in \operatorname{ran}(A^*) \setminus \{0\}. \tag{3.2.15}$$

It remains to check that the latter condition holds under $(\mathbf{A}_8)$**(II)**. First, observe that $\mathcal{C}$ is a nonempty bounded convex set thanks to $(\mathbf{A}_1)$ and $(\mathbf{A}_3)$. The first condition in $(\mathbf{A}_8)$**(II)** is equivalent to $0 \in \operatorname{ri}(\mathcal{C} - c)$ for some $c \in A^{-1}(b)$. It then follows from Proposition 2.1.19 that

$$\sigma_{\mathcal{C} - c}(x) > 0, \forall x \notin \operatorname{par}(\mathcal{C} - c)^\perp = \operatorname{par}(\mathcal{C})^\perp,$$

which then implies (3.2.15) thanks to the second condition in $(\mathbf{A}_8)$**(II)**.

□

Lemma 3.2.7. *Suppose that assumptions* $(\mathbf{A}_1)$-$(\mathbf{A}_3)$ *and* $(\mathbf{A}_6)$-$(\mathbf{A}_8)$ *and* $(\mathbf{P}_1)$-$(\mathbf{P}_6)$ *hold. Then the sequence of dual iterates* $(\mu_k)_{k \in \mathbb{N}}$ *generated by* Algorithm 8 *is bounded.*

Proof. Using the notation in (3.1.5), the primal problem:

$$\min_{x\in\mathcal{H}_p}\{\Phi(x):\ \ Ax=b\}=\min_{x\in\mathcal{H}_p}\sup_{\mu\in\mathcal{H}_d}\ \mathcal{L}(x,\mu),$$

is obviously equivalent to

$$\min_{x\in\mathcal{H}_p}\left\{\Phi(x)+\frac{\rho_k}{2}\|Ax-b\|^2:\ \ Ax=b\right\}=\min_{x\in\mathcal{H}_p}\sup_{\mu\in\mathcal{H}_d}\left\{\mathcal{L}(x,\mu)+\frac{\rho_k}{2}\|Ax-b\|^2\right\}.$$

We associate to the previous the following regularized primal problem:

$$\min_{x\in\mathcal{H}_p}\{\Phi_k(x):\ \ Ax=b\}=\min_{x\in\mathcal{H}_p}\sup_{\mu\in\mathcal{H}_d}\ \mathcal{L}_k(x,\mu)$$

and its Lagrangian dual, namely:

$$\sup_{\mu\in\mathcal{H}_d}\inf_{x\in\mathcal{H}_p}\ \mathcal{L}_k(x,\mu)=-\inf_{\mu\in\mathcal{H}_d}\sup_{x\in\mathcal{H}_p}\ -\mathcal{L}_k(x,\mu).$$

Now consider the dual function in the latter, namely $\varphi_k(\mu)\overset{\text{def}}{=}-\inf_{x\in\mathcal{H}_p}\ \mathcal{L}_k(x,\mu)$. Observe that the minimum is actually attained owing to $(\mathbf{A}_1)$ and $(\mathbf{A}_3)$. Now we claim that φ_k is continuously differentiable with $L_{\nabla\varphi_k}$-Lipschitz gradient, and $1/\underline{\rho}$ (see $(\mathbf{P}_4)$) is an upper-bound for $(L_{\nabla\varphi_k})_{k\in\mathbb{N}}$. In order to show it, introduce the notation

$$\begin{aligned}\phi_k(x)&\overset{\text{def}}{=}f(x)+g^{\beta_k}(Tx)+h(x);\\ \psi_k(v)&\overset{\text{def}}{=}\frac{\rho_k}{2}\|v-b\|^2.\end{aligned}$$

By definition, we have

$$\begin{aligned}\varphi_k(\mu)&=-\min_{x\in\mathcal{H}_p}\left\{f(x)+g^{\beta_k}(Tx)+h(x)+\langle\mu,\ Ax-b\rangle+\frac{\rho_k}{2}\|Ax-b\|^2\right\}\\&=-\min_{x\in\mathcal{H}_p}\{\phi_k(x)+\langle A^*\mu,\ x\rangle+\psi_k(Ax)\}+\langle\mu,\ b\rangle.\end{aligned}\tag{3.2.16}$$

Using Fenchel-Rockafellar duality and strong duality, which holds by $(\mathbf{P}_4)$ and continuity of ψ_k (see, for instance, [92, Theorem 3.51]), we have the following equality,

$$\begin{aligned}\min_{x\in\mathcal{H}_p}\{\phi_k(x)+\langle A^*\mu,x\rangle+\psi_k(Ax)\}&=-\min_{v\in\mathcal{H}_d}\{(\phi_k(\cdot)+\langle A^*\mu,\cdot\rangle)^*(-A^*v)+\psi_k^*(v)\}\\&=-\min_{v\in\mathcal{H}_d}\{\phi_k^*(-A^*v-A^*\mu)+\psi_k^*(v)\}\end{aligned}$$

where we have used the fact that the conjugate of a linear perturbation is the translation of the conjugate in the last line. Substituting the above into (3.2.16) we find

$$\begin{aligned}\varphi_k(\mu)&=\min_{v\in\mathcal{H}_d}\left\{\phi_k^*(-A^*(v+\mu))+\frac{1}{2\rho_k}\|v\|^2+\langle v,\ b\rangle\right\}+\langle\mu,\ b\rangle\\&=\min_{v\in\mathcal{H}_d}\left\{\phi_k^*(-A^*(v+\mu))+\frac{1}{2\rho_k}\|v+\rho_kb\|^2\right\}+\langle\mu,\ b\rangle-\frac{\rho_k}{2}\|b\|^2\end{aligned}$$

Moreover, from the primal-dual extremality relationships [92, Theorem 3.51(i)], we have

$$-\tilde{v}=\nabla\psi_k(A\tilde{x})=\rho_k(A\tilde{x}-b),\tag{3.2.17}$$

where $\tilde{x}$ is a minimizer (which exists and belongs to $\mathcal{C}$) of the primal objective $\mathcal{L}_k(\cdot,\mu)$ and $\tilde{v}$ is the unique minimizer to the associated dual objective. Now, using the change of variable $u=v+\mu$, we get

$$\begin{aligned}\varphi_k(\mu)&=\inf_{u\in\mathcal{H}_d}\left\{\phi_k^*(-A^*u)+\frac{1}{2\rho_k}\|u-\mu+\rho_kb\|^2\right\}+\langle\mu,\ b\rangle-\frac{\rho_k}{2}\|b\|^2\\&=[\phi_k^*\circ(-A^*)]^{\rho_k}(\mu-\rho_kb)+\langle\mu,\ b\rangle-\frac{\rho_k}{2}\|b\|^2,\end{aligned}$$

where the notation $[\cdot]^{\rho_k}$ denotes the Moreau envelope with parameter ρ_k as defined in (2.1.2). It follows from Proposition 2.1.2(i) and (iii), that φ_k is convex, real-valued and its gradient, given by

$$\nabla\varphi_k(\mu)=\rho_k^{-1}(\mu-\rho_kb-\tilde{u})+b=\rho_k^{-1}(\mu-\tilde{u}),\quad\text{where}\quad\tilde{u}=\operatorname{prox}_{\rho_k\phi_k^*\circ(-A^*)}(\mu-\rho_kb),\tag{3.2.18}$$

is $1/\rho_k$-Lipschitz continuous since the gradient of a Moreau envelope with parameter ρ_k is $1/\rho_k$-Lipschitz continuous (see Proposition 2.1.2(iii)). As ρ_k is non-decreasing, $1/\rho_k \leq 1/\underline{\rho}$ and the sequence of functions $(\nabla\varphi_k)_{k\in\mathbb{N}}$ is uniformly Lipschitz-continuous with constant $1/\underline{\rho}$. In addition, combining (3.2.17) and (3.2.18), and recalling the change of variable $\tilde{u} = \tilde{v} + \mu$, we get that

$$\nabla\varphi_k(\mu) = \rho_k^{-1}(\mu - \tilde{u}) = -\rho_k^{-1}\tilde{v} = A\tilde{x} - b. \tag{3.2.19}$$

As in Lemma 3.2.5, we are going to denote $\tilde{x}_k$ a minimizer of $\mathcal{L}_k(x, \mu_k)$. Then, from the Descent Lemma (see Proposition 2.1.5 and inequality (2.1.7)), we have

$$\varphi_k(\mu_{k+1}) \leq \varphi_k(\mu_k) + \langle \nabla\varphi_k(\mu_k),\ \mu_{k+1} - \mu_k\rangle + \frac{1}{2\underline{\rho}}\|\mu_{k+1} - \mu_k\|^2.$$

Now substitute in the right-hand-side the expression $\nabla\varphi_k(\mu_k) = A\tilde{x}_k - b$ in (3.2.19) and the update $\mu_{k+1} = \mu_k + \theta_k(Ax_{k+1} - b)$ from the algorithm, to obtain

$$\begin{aligned}\varphi_k(\mu_{k+1}) &\leq \varphi_k(\mu_k) + \theta_k\langle A\tilde{x}_k - b,\ Ax_{k+1} - b\rangle + \frac{\theta_k^2}{2\underline{\rho}}\|Ax_{k+1} - b\|^2 \\ &\leq \varphi_k(\mu_k) + \frac{\theta_k}{2}\|A\tilde{x}_k - b\|^2 + \frac{\theta_k}{2}\left(\frac{\theta_k}{\underline{\rho}} + 1\right)\|Ax_{k+1} - b\|^2,\end{aligned} \tag{3.2.20}$$

where we estimated the scalar product by Cauchy-Schwartz and Young inequality. Moreover, by definition,

$$\begin{aligned}\varphi_{k+1}(\mu_{k+1}) &= -\inf_{x\in\mathcal{H}_p}\left\{f(x) + g^{\beta_{k+1}}(Tx) + h(x) + \langle\mu_{k+1},\ Ax - b\rangle + \frac{\rho_{k+1}}{2}\|Ax - b\|^2\right\} \\ &= \sup_{x\in\mathcal{H}_p}\left\{-\mathcal{L}_k(x, \mu_{k+1}) + \left[g^{\beta_k} - g^{\beta_{k+1}}\right](Tx) + \frac{1}{2}(\rho_k - \rho_{k+1})\|Ax - b\|^2\right\}.\end{aligned} \tag{3.2.21}$$

Now recall assumptions $(\mathbf{P}_3)$ and $(\mathbf{P}_4)$: for β_k non-increasing, $\left[g^{\beta_k} - g^{\beta_{k+1}}\right](Tx) \leq 0$ for every $x \in \mathcal{H}_p$ by Proposition 2.1.2(v) and, for ρ_k non-decreasing, $\rho_k - \rho_{k+1} \leq 0$. Then we can estimate the right-hand-side of (3.2.21) to obtain

$$\varphi_{k+1}(\mu_{k+1}) \leq \sup_{x\in\mathcal{H}_p} -\mathcal{L}_k(x, \mu_{k+1}) = \varphi_k(\mu_{k+1}).$$

Sum (3.2.20) with the latter, to obtain

$$\varphi_{k+1}(\mu_{k+1}) - \varphi_k(\mu_k) \leq \frac{\theta_k}{2}\|A\tilde{x}_k - b\|^2 + \frac{\theta_k}{2}\left(\frac{\theta_k}{\underline{\rho}} + 1\right)\|Ax_{k+1} - b\|^2.$$

By Assumption $(\mathbf{P}_6)$, $\theta_k = \gamma_k/c$ where $\gamma_k \leq 1$. Moreover, by assumption $(\mathbf{P}_5)$, $\gamma_k \leq \overline{M}\gamma_{k+1}$. Then,

$$\varphi_{k+1}(\mu_{k+1}) - \varphi_k(\mu_k) \leq \frac{\gamma_k}{2c}\|A\tilde{x}_k - b\|^2 + \frac{\overline{M}}{2c}\left(\frac{1}{\underline{\rho c}} + 1\right)\gamma_{k+1}\|Ax_{k+1} - b\|^2. \tag{3.2.22}$$

Notice that the right-hand-side is in ℓ^1_+, because both $\left(\gamma_k\|Ax_k - b\|^2\right)_{k\in\mathbb{N}}$ and $\left(\gamma_k\|A\tilde{x}_k - b\|^2\right)_{k\in\mathbb{N}}$ are in ℓ^1_+ by Lemma 3.2.5. Additionally, $(\varphi_k(\mu_k))_{k\in\mathbb{N}}$ is bounded from below. Indeed, by virtue of $(\mathbf{A}_6)$ and Remark 3.1.1(iv), we have

$$\begin{aligned}\varphi_k(\mu_k) &\geq -\mathcal{L}_k(x^\star, \mu_k) \\ &\geq -\left[f(x^\star) + g(Tx^\star) + h(x^\star)\right] > -\infty.\end{aligned}$$

Then we can use Lemma 2.2.3(i) on inequality (3.2.22) to conclude that $(\varphi_k(\mu_k))_{k\in\mathbb{N}}$ is convergent and, in particular, bounded. Now recall Φ_k, $\bar{\Phi}$ and $\bar{\varphi}$ from (3.1.5). Notice that

$$\begin{aligned}\varphi_k(\mu) &= \sup_{x\in\mathcal{H}_p}\{\langle\mu,\ b - Ax\rangle - \Phi_k(x)\} \\ &= \sup_{x\in\mathcal{H}_p}\{\langle -A^*\mu,\ x\rangle - \Phi_k(x)\} + \langle b,\ \mu\rangle \\ &= \Phi_k^*(-A^*\mu) + \langle b,\ \mu\rangle.\end{aligned}$$

It then follows that

$$g^{\beta_k} \leq g \quad \Longrightarrow \quad \Phi_k \leq \bar{\Phi} \quad \Longleftrightarrow \quad \bar{\Phi}^* \leq \Phi_k^* \quad \Longrightarrow \quad \bar{\varphi} \leq \varphi_k, \tag{3.2.23}$$

where we used Proposition 2.1.2(v) and the fact in (2.1.1). We are now in position to invoke Lemma 3.2.6 which shows that $\bar{\varphi}$ is coercive on $\operatorname{ran}(A)$, and thus, by (3.2.23), $(\varphi_k)_{k\in\mathbb{N}}$ is equi-coercive on $\operatorname{ran}(A)$. In turn, since $\operatorname{ran}(A)$ is closed and $(\mu_k)_{k\in\mathbb{N}} \subset \operatorname{ran}(A) = \ker(A^*)^\perp$, we have from (3.2.23) and the proof of Lemma 3.2.6 that

$$\exists (a>0, \alpha>0, \beta\in\mathbb{R}), (\forall k\in\mathbb{N}), \quad \varphi_k(\mu_k) \geq \bar{\varphi}(\mu_k) \geq a\left\|A^*\mu_k\right\| + \beta \geq a\alpha\left\|\mu_k\right\| + \beta,$$

which shows that $(\mu_k)_{k\in\mathbb{N}}$ is indeed bounded by boundedness of $\left(\varphi_k\left(\mu_k\right)\right)_{k\in\mathbb{N}}$. □

Lemma 3.2.8. *Under assumptions* $(\mathbf{A}_1)$-$(\mathbf{A}_8)$ *and* $(\mathbf{P}_1)$-$(\mathbf{P}_6)$*, the objective* Φ *is bounded on* $\mathcal{C}$*, and thus*

$$\tilde{M} \stackrel{\text{def}}{=} \sup_{x\in\mathcal{C}} |\Phi(x)| + \sup_{k\in\mathbb{N}} \|\mu_k\| \left(\|A\| \, R + \|b\|\right) < +\infty, \tag{3.2.24}$$

where we recall the radius R *from assumption* $(\mathbf{A}_3)$*.*

Proof. By assumption $(\mathbf{A}_4)$, g is subdifferentiable at Tx for any $x \in \mathcal{C}$. Thus convexity of g implies that for any $x \in \mathcal{C}$

$$\begin{aligned} g(Tx) &\leq g(Tx^\star) + \left\langle \left[\partial g\left(Tx\right)\right]^0, Tx - Tx^\star \right\rangle \leq g(Tx^\star) + \left\| \left[\partial g\left(Tx\right)\right]^0 \right\| \|T\| d_{\mathcal{C}} \\ g(Tx) &\geq g(Tx^\star) + \left\langle \left[\partial g\left(Tx^\star\right)\right]^0, Tx - Tx^\star \right\rangle \geq g(Tx^\star) - \left\| \left[\partial g\left(Tx^\star\right)\right]^0 \right\| \|T\| d_{\mathcal{C}}. \end{aligned} \tag{3.2.25}$$

By assumption $(\mathbf{A}_4)$, $\|\nabla f\|$ is uniformly bounded on $\mathcal{C}$ and we have

$$\sup_{x\in\mathcal{C}} \|\nabla f\left(x\right)\| < +\infty. \tag{3.2.26}$$

In turn, convexity entails that for any $x \in \mathcal{C}$

$$\begin{aligned} f(x) &\leq f(x^\star) + \left\langle \nabla f\left(x\right), x - x^\star \right\rangle \leq f(x^\star) + \|\nabla f\left(x\right)\| \, d_{\mathcal{C}}, \\ f(x) &\geq f(x^\star) + \left\langle \nabla f\left(x^\star\right), x - x^\star \right\rangle \geq f(x^\star) - \|\nabla f\left(x^\star\right)\| \, d_{\mathcal{C}}. \end{aligned} \tag{3.2.27}$$

From assumption $(\mathbf{A}_5)$, we also have for any $x \in \mathcal{C}$

$$h(x^\star) - L_h d_{\mathcal{C}} \leq h(x) \leq h(x^\star) + L_h d_{\mathcal{C}}. \tag{3.2.28}$$

Summing (3.2.25), (3.2.27) and (3.2.28), using (3.2.26) and assumption $(\mathbf{A}_4)$, we get

$$|\Phi(x)| \leq |\Phi(x^\star)| + \left(L_h + \|T\| \sup_{x\in\mathcal{C}} \left\| \left[\partial g\left(Tx\right)\right]^0 \right\| + \sup_{x\in\mathcal{C}} \|\nabla f\left(x\right)\| \right).$$

From Lemma 3.2.7, we know that the sequence of dual variables $(\mu_k)_{k\in\mathbb{N}}$ is bounded which concludes the proof. □

3.2.4 Optimality Estimation

With the boundedness results of Section 3.2.3, we can prove the main energy estimation; a key inequality used to show that the Lagrangian values converge to the optimum. Define $C_k \stackrel{\text{def}}{=} \frac{L_k}{2} d_{\mathcal{C}}^2 + d_{\mathcal{C}} \left(D + M\|T\| + L_h + \|A\| \, \|\mu^\star\|\right)$, where L_k is given in (3.2.1) and the constants D, M, and L_h are as in Lemma 3.2.3. We then have the following lemma, in which we state the main energy estimation.

Lemma 3.2.9. *Suppose that assumptions* $(\mathbf{A}_1)$-$(\mathbf{A}_8)$ *and* $(\mathbf{P}_1)$-$(\mathbf{P}_6)$ *hold, with* $\underline{M} \geq 1$*. Consider the sequence of primal-dual iterates* $\left((x_k, \mu_k)\right)_{k\in\mathbb{N}}$ *generated by* Algorithm 8 *and* $(x^\star, \mu^\star)$ *a saddle-point point of the Lagrangian as in* (3.1.6)*. Let*

$$r_k \stackrel{\text{def}}{=} (1-\gamma_k)\mathcal{L}_k\left(x_k, \mu_k\right) + \frac{c}{2}\|\mu_k - \mu^\star\|^2 + \frac{\beta_k}{2} M^2 + \gamma_k \tilde{M}. \tag{3.2.29}$$

Then, we have the following energy estimate

$$\begin{aligned} &r_{k+1} - r_k + \gamma_k \left[\mathcal{L}\left(x_k, \mu^\star\right) - \mathcal{L}\left(x^\star, \mu^\star\right) + \frac{\rho_k}{2}\|Ax_k - b\|^2\right] \leq \\ &\frac{1}{2}\left[\rho_{k+1} - \rho_k - \gamma_{k+1}\rho_{k+1} + \frac{2}{c}\gamma_k - \frac{\gamma_k^2}{c}\right]\|Ax_{k+1} - b\|^2 + \frac{\gamma_k\beta_k}{2}M^2 + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right) + C_k\gamma_k^2. \end{aligned} \tag{3.2.30}$$

Proof. Notice that the dual update $\mu_{k+1} = \mu_k + \theta_k\left(Ax_{k+1} - b\right)$ can be re-written as

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

Then, from firm nonexpansiveness of the proximal mapping (see (2.1.3)),

$$\begin{aligned} 0 &\geq \theta_k\left[\mathcal{L}_k\left(x_{k+1}, \mu^\star\right) - \mathcal{L}_k\left(x_{k+1}, \mu_{k+1}\right)\right] + \frac{1}{2}\left[\|\mu_{k+1} - \mu^\star\|^2 - \|\mu_k - \mu^\star\|^2\right. \\ &\qquad \left. + \|\mu_{k+1} - \mu_k\|^2\right] \\ &= \theta_k\left[\mathcal{L}_k\left(x_{k+1}, \mu^\star\right) - \mathcal{L}_k\left(x_{k+1}, \mu_{k+1}\right)\right] + \frac{1}{2}\left[\|\mu_{k+1} - \mu^\star\|^2 - \|\mu_k - \mu^\star\|^2\right] \\ &\qquad + \frac{\theta_k^2}{2}\|Ax_{k+1} - b\|^2. \end{aligned} \tag{3.2.31}$$

Notice that

$$\mathcal{L}_k\left(x_{k+1}, \mu_k\right) - \mathcal{L}_k\left(x_k, \mu_k\right) = \left[\mathcal{E}_k\left(x_{k+1}, \mu_k\right) + h(x_{k+1})\right] - \left[\mathcal{E}_k\left(x_k, \mu_k\right) + h(x_k)\right]$$

and that, by the definition of x_{k+1} in the algorithm and by convexity of function h,

$$\begin{aligned} h(x_{k+1}) - h(x_k) &= h((1-\gamma_k)x_k + \gamma_k s_k) - h(x_k) \\ &\leq \gamma_k\left(h(s_k) - h(x_k)\right). \end{aligned}$$

Then,

$$\mathcal{L}_k\left(x_{k+1}, \mu_k\right) - \mathcal{L}_k\left(x_k, \mu_k\right) \leq \mathcal{E}_k\left(x_{k+1}, \mu_k\right) - \mathcal{E}_k\left(x_k, \mu_k\right) + \gamma_k\left(h(s_k) - h(x_k)\right). \tag{3.2.32}$$

Now apply Lemma 3.2.2 at the points $x^\star$, x_k, and μ_k to affirm that

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

From the latter, by the alternative definition of s_k in the algorithm (see (3.1.4)), we obtain

$$\mathcal{E}_k\left(x^\star, \mu_k\right) \geq \mathcal{E}_k\left(x_k, \mu_k\right) - h(x^\star) + h(s_k) + \left\langle\nabla_x\mathcal{E}_k\left(x_k, \mu_k\right),\ s_k - x_k\right\rangle + \frac{\rho_k}{2}\|Ax_k - b\|^2. \tag{3.2.33}$$

From Lemma 3.2.1, we have also that

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

Recall that, from the algorithm, $x_{k+1} = x_k + \gamma_k\left(s_k - x_k\right)$. Then,

$$\begin{aligned} \mathcal{E}_k\left(x_{k+1}, \mu_k\right) &\leq \mathcal{E}_k\left(x_k, \mu_k\right) + \gamma_k\left\langle\nabla_x\mathcal{E}_k\left(x_k, \mu_k\right),\ s_k - x_k\right\rangle + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right) + \frac{L_k\gamma_k^2}{2}\|s_k - x_k\|^2 \\ &\leq \mathcal{E}_k\left(x_k, \mu_k\right) + \gamma_k\left[\mathcal{E}_k\left(x^\star, \mu_k\right) + h(x^\star) - \mathcal{E}_k\left(x_k, \mu_k\right) - h(s_k) - \frac{\rho_k}{2}\|Ax_k - b\|^2\right] \\ &\quad + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right) + \frac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2, \end{aligned}$$

where in the last inequality we used (3.2.33). Using the latter in (3.2.32), we obtain

$$\begin{aligned} \mathcal{L}_k\left(x_{k+1}, \mu_k\right) - \mathcal{L}_k\left(x_k, \mu_k\right) \leq &\gamma_k\left[\mathcal{L}_k\left(x^\star, \mu_k\right) - \mathcal{L}_k\left(x_k, \mu_k\right) - \frac{\rho_k}{2}\|Ax_k - b\|^2\right] \\ &+ K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right) + \frac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2. \end{aligned} \tag{3.2.34}$$

Notice also that, from the definitions of $\mathcal{L}_k\left(x_{k+1}, \cdot\right)$ and μ_{k+1} as $\mu_{k+1} = \mu_k + \theta_k\left(Ax_{k+1} - b\right)$,

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

So, from the latter and (3.2.34),

$$\mathcal{L}_k\left(x_{k+1},\mu_{k+1}\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\leq\theta_k\|Ax_{k+1}-b\|^2+\gamma_k\left[\mathcal{L}_k\left(x^\star,\mu_k\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\right]\\-\frac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2+K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right)+\frac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2.$$

Now recall that, by assumption $(\mathbf{P}_6)$, $\theta_k=\gamma_k/c$. Multiply (3.2.31) by c and sum with the latter, to obtain

$$\begin{aligned}&(1-c\theta_k)\mathcal{L}_k\left(x_{k+1},\mu_{k+1}\right)-(1-c\theta_k)\mathcal{L}_k\left(x_k,\mu_k\right)+\tfrac{c}{2}\left[\|\mu_{k+1}-\mu^\star\|^2-\|\mu_k-\mu^\star\|^2\right]\\\leq\;&\left(\theta_k-\tfrac{c\theta_k^2}{2}\right)\|Ax_{k+1}-b\|^2+\gamma_k\left[\mathcal{L}_k\left(x^\star,\mu_k\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\right]-c\theta_k\left[\mathcal{L}_k\left(x_{k+1},\mu^\star\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\right]\\&-\tfrac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2+K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right)+\tfrac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2.\end{aligned}$$

The previous inequality can be re-written, by trivial manipulations, as

$$\begin{aligned}&(1-c\theta_{k+1})\mathcal{L}_{k+1}\left(x_{k+1},\mu_{k+1}\right)-(1-c\theta_k)\mathcal{L}_k\left(x_k,\mu_k\right)+\frac{c}{2}\left[\|\mu_{k+1}-\mu^\star\|^2-\|\mu_k-\mu^\star\|^2\right]\\\leq\;&(1-c\theta_{k+1})\mathcal{L}_{k+1}\left(x_{k+1},\mu_{k+1}\right)-(1-c\theta_k)\mathcal{L}_k\left(x_{k+1},\mu_{k+1}\right)+\left(\theta_k-\frac{c\theta_k^2}{2}\right)\|Ax_{k+1}-b\|^2\\&+\gamma_k\left[\mathcal{L}_k\left(x^\star,\mu_k\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\right]-c\theta_k\left[\mathcal{L}_k\left(x_{k+1},\mu^\star\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\right]-\frac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2\\&+K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right)+\frac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2\\=\;&c\left(\theta_k-\theta_{k+1}\right)\left[f+h+\langle\mu_{k+1},\;A\cdot-b\rangle\right]\left(x_{k+1}\right)+\left[\left(1-c\theta_{k+1}\right)g^{\beta_{k+1}}-\left(1-c\theta_k\right)g^{\beta_k}\right]\left(Tx_{k+1}\right)\\&+\frac{1}{2}\left[\left(1-c\theta_{k+1}\right)\rho_{k+1}-\left(1-c\theta_k\right)\rho_k+2\theta_k-c\theta_k^2\right]\|Ax_{k+1}-b\|^2\\&+\gamma_k\left[\mathcal{L}_k\left(x^\star,\mu_k\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\right]-c\theta_k\left[\mathcal{L}_k\left(x_{k+1},\mu^\star\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\right]-\frac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2\\&+K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right)+\frac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2.\end{aligned}\tag{3.2.35}$$

By $(\mathbf{P}_5)$ and $(\mathbf{P}_6)$, 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 $(\mathbf{P}_3)$, we also have $\beta_{k+1}\leq\beta_k$ by $(\mathbf{P}_3)$. In particular, $g^{\beta_k}\leq g^{\beta_{k+1}}\leq g$. Now, by Proposition 2.1.2(iv) and the definition of the constant M in (3.1.8), we are able to estimate the quantity

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

Then,

$$\begin{aligned}&c\left(\theta_k-\theta_{k+1}\right)\left[f+h+\langle\mu_{k+1},\;A\cdot-b\rangle\right]\left(x_{k+1}\right)+\left[\left(1-c\theta_{k+1}\right)g^{\beta_{k+1}}-\left(1-c\theta_k\right)g^{\beta_k}\right]\left(Tx_{k+1}\right)\\\leq\;&c\left(\theta_k-\theta_{k+1}\right)\mathcal{L}\left(x_{k+1},\mu_{k+1}\right)+\frac{1}{2}\left(\beta_k-\beta_{k+1}\right)M^2.\end{aligned}\tag{3.2.36}$$

Recall that, by assumption $(\mathbf{A}_3)$, $\mathcal{C}$ is convex and bounded and that, by the update $x_{k+1}=x_k+\gamma_k\left(s_k-x_k\right)$ with $s_k\in\mathcal{C}$ and $\gamma_k\in]0,1]$ by $(\mathbf{P}_1)$, x_k always belongs to $\mathcal{C}$. From the assumptions, the functions f, h and $g\circ T$ are bounded on $\mathcal{C}$ and, from the algorithm and convexity, $(x_k)_{k\in\mathbb{N}}\subset\mathcal{C}$. By Lemma 3.2.7, also the sequence $(\mu_k)_{k\in\mathbb{N}}$ is bounded. Then, recalling $\tilde{M}$ from Lemma 3.2.8, we can use the Cauchy-Schwartz and the triangular inequality to affirm that

$$\mathcal{L}\left(x_k,\mu_k\right)=\Phi(x_k)+\langle\mu_k,\;Ax_k-b\rangle\leq\tilde{M}.\tag{3.2.37}$$

Recall the definition of r_k in (3.2.29). Coming back to (3.2.35) and using both (3.2.36) and (3.2.37), we obtain

$$\begin{aligned} r_{k+1} - r_k \leq & \frac{1}{2}\left[(1-\gamma_{k+1})\rho_{k+1} - (1-\gamma_k)\rho_k + \frac{2}{c}\gamma_k - \frac{\gamma_k^2}{c}\right]\|Ax_{k+1}-b\|^2 \\ & + \gamma_k\left[\mathcal{L}_k(x^\star,\mu_k) - \mathcal{L}_k(x_{k+1},\mu^\star)\right] - \frac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_k) + \frac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2. \end{aligned} \tag{3.2.38}$$

Recall that, by feasibility of $x^\star$, $\mathcal{L}(x^\star,\mu_k) = \mathcal{L}(x^\star,\mu^\star)$. Now compute

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

where in the inequality we used the facts that $g^{\beta_k} \leq g$ and that, by Proposition 2.1.2(v) and (3.1.8),

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

Then, using the latter in (3.2.38), we obtain

$$\begin{aligned} r_{k+1} - r_k \leq & \frac{1}{2}\left[\rho_{k+1} - \rho_k - \gamma_{k+1}\rho_{k+1} + \frac{2}{c}\gamma_k - \frac{\gamma_k^2}{c}\right]\|Ax_{k+1}-b\|^2 + \gamma_k\left[\mathcal{L}(x^\star,\mu^\star) - \mathcal{L}(x_{k+1},\mu^\star)\right] \\ & + \frac{\gamma_k\beta_k}{2}M^2 - \frac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_k) + \frac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2. \end{aligned}$$

We replace the term $[\mathcal{L}(x^\star,\mu^\star) - \mathcal{L}(x_{k+1},\mu^\star)]$ with $[\mathcal{L}(x^\star,\mu^\star) - \mathcal{L}(x_k,\mu^\star)] + [\mathcal{L}(x_k,\mu^\star) - \mathcal{L}(x_{k+1},\mu^\star)]$ and estimate using Lemma 3.2.3 to get the following,

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

We conclude by trivial manipulations. □

3.3 Convergence Analysis

Throughout this section, when rates of convergence are given they will be given in terms of the quantity $\Gamma_k \overset{\text{def}}{=} \sum_{i=0}^{k}\gamma_i$. Because our analysis is carried out for open loop step sizes, the rates must necessarily be stated in terms of this quantity. To give a clearer picture of what rates one can expect, e.g. in practice, we provide the following example.

Example 3.3.1. Suppose that the sequences of parameters are chosen according to Example 3.1.3. Let the function $\sigma : t \in \mathbb{R}^+ \mapsto (\log(t+2))^a/(t+1)^{1-b}$. We obviously have $\sigma(k) = \gamma_k$ for $k \in \mathbb{N}$. Moreover, it is easy to see that $\exists k' \geq 0$ (depending on a and b), such that σ is decreasing for $t \geq k'$. Thus, $\forall k \geq k'$, we have

$$\Gamma_k \geq \sum_{i=k'}^{k}\gamma_i \geq \int_{k'}^{k+1}\sigma(t)dt \geq \int_{k'+1}^{k+2}(\log(t))^a t^{b-1}dt = \int_{\log(k'+1)}^{\log(k+2)} t^a e^{bt}dt.$$

It is easy to show, using integration by parts for the first case, that

$$\Gamma_k^{-1} = \begin{cases} o\left(\frac{1}{(k+2)^b}\right) & a = 1, b > 0, \\ O\left(\frac{1}{(k+2)^b}\right) & a = 0, b > 0, \\ O\left(\frac{1}{\log(k+2)}\right) & a = 0, b = 0. \end{cases}$$

This result reveals that picking a and b as large as possible results in a faster convergence rate, with the proviso that b satisfy some conditions for $(\mathbf{P}_1)$-$(\mathbf{P}_7)$ to hold, see the discussion in Example 3.1.3 for the largest possible choice of b. In the case of Lipschitz-smooth functions, the largest possible choice of b is $1/3 - \epsilon$ for arbitrary $\epsilon > 0$ which gives $\Gamma_k^{-1} = O\left(\frac{1}{(k+2)^{1/3-\epsilon}}\right)$.

3.3.1 Asymptotic Feasibility

We are now prove Theorem 3.3.2, i.e., we show that the sequence of iterates $(x_k)_{k\in\mathbb{N}}$ is asymptotically feasible.

Theorem 3.3.2 (Asymptotic feasibility). *Suppose that Assumptions $(\mathbf{A}_1)$-$(\mathbf{A}_4)$ and $(\mathbf{A}_6)$ hold. Consider the sequence of iterates $(x_k)_{k\in\mathbb{N}}$ from Algorithm 8 with parameters satisfying Assumptions $(\mathbf{P}_1)$-$(\mathbf{P}_6)$. Then,*

(i) Ax_k converges strongly to b as $k \to \infty$, i.e., the sequence $(x_k)_{k\in\mathbb{N}}$ is asymptotically feasible for $(\mathscr{P})$ in the strong topology.

(ii) Pointwise rate:

$$\inf_{0\leq i\leq k} \|Ax_i - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right) \text{ and } \exists \text{ a subsequence } \left(x_{k_j}\right)_{j\in\mathbb{N}} \text{ s.t. for all } j \in \mathbb{N}, \|Ax_{k_j} - b\| \leq \frac{1}{\sqrt{\Gamma_{k_j}}}, \tag{3.3.1}$$

where, for all $k \in \mathbb{N}$, $\Gamma_k \overset{\text{def}}{=} \sum_{i=0}^k \gamma_i$.

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

$$\|A\bar{x}_k - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right). \tag{3.3.2}$$

Proof. (i) By Lemma 3.2.4 with $\rho_k \equiv \rho_{k+1} \equiv 2$, we have

$$\|Ax_k - b\|^2 - \|Ax_{k+1} - b\|^2 \leq 2\gamma_k d_{\mathcal{C}} \|A\| \left(\|A\| R + \|b\|\right).$$

Using this together with Lemma 3.2.5 and Assumption $(\mathbf{P}_2)$, we are in position to apply Lemma 2.2.3(ii) to conclude that $\lim_{k\to\infty} \|Ax_k - b\| = 0$.

(ii) The rates in (3.3.1) follow respectively from Lemma 2.2.3(iii) and Lemma 2.2.3(iv).

(iii) We have, by Jensen's inequality and Lemma 3.2.5, that

$$\|A\bar{x}_k - b\|^2 \leq \frac{1}{\Gamma_k}\sum_{i=0}^k \gamma_i \|Ax_i - b\|^2 \leq \frac{1}{\Gamma_k}\sum_{i=0}^{+\infty} \gamma_i \|Ax_i - b\|^2 = O\left(\frac{1}{\Gamma_k}\right).$$

□

3.3.2 Optimality

In this section we prove Theorem 3.3.3 by establishing convergence of the Lagrangian values to the optimum (i.e., the value at the saddle-point). The convergence and rates of the Lagrangian values will be shown in terms $\mathcal{L}(x_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star)$, which is non-negative since $(x^\star, \mu^\star)$ is a saddle point. This is however not a primal-dual gap, in the strictest interpretation. Nevertheless, observe that in view of [10, Proposition 19.21(v)], we have, for each $k \in \mathbb{N}$,

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

which is nothing but the Bregman divergence of Φ with the subgradient $-A^*\mu$ between x_k and $x^\star$. This Bregman divergence appears then as a good candidate to quantify the convergence rate of Algorithm 8 given that it captures both the discrepancy of the primal objective to the optimal value and violation of the affine constraint.

Theorem 3.3.3 (Convergence to optimality). *Suppose that assumptions $(\mathbf{A}_1)$-$(\mathbf{A}_8)$ and $(\mathbf{P}_1)$-$(\mathbf{P}_7)$ hold, with $\underline{M} \geq 1$. Let $(x_k)_{k\in\mathbb{N}}$ be the sequence of primal iterates generated by* Algorithm 8 *and $(x^\star, \mu^\star)$ a saddle-point pair for the Lagrangian. Then, in addition to the results of Theorem 3.3.2, the following holds*

(i) Convergence of the Lagrangian:

$$\lim_{k\to\infty} \mathcal{L}(x_k, \mu^\star) = \mathcal{L}(x^\star, \mu^\star). \tag{3.3.3}$$

(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 strongly to $\bar{\mu}$ a solution of the dual problem, ($\mathscr{D}$), i.e., $(\bar{x}, \bar{\mu})$ is a saddle point of $\mathcal{L}$.

(iii) Pointwise rate:

$$\begin{aligned} \inf_{0\le i\le k} \mathcal{L}(x_i, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) = O\left(\frac{1}{\Gamma_k}\right) \text{ and} \\ \exists \text{ a subsequence } \left(x_{k_j}\right)_{j\in\mathbb{N}} \text{ s.t. for each } j\in\mathbb{N}, \mathcal{L}\left(x_{k_j+1}, \mu^\star\right) - \mathcal{L}(x^\star, \mu^\star) \le \frac{1}{\Gamma_{k_j}}. \end{aligned} \tag{3.3.4}$$

(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

$$\mathcal{L}(\bar{x}_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) = O\left(\frac{1}{\Gamma_k}\right). \tag{3.3.5}$$

Proof. Our starting point is the main energy estimate (3.2.30). Let us focus on its right-hand-side. Under assumption ($\mathbf{P}_7$),

$$\frac{1}{2}\left[\rho_{k+1} - \rho_k - \gamma_{k+1}\rho_{k+1} + \frac{2}{c}\gamma_k - \frac{\gamma_k^2}{c}\right]\|Ax_{k+1} - b\|^2 \le \gamma_{k+1}\|Ax_{k+1} - b\|^2,$$

where the right hand side is in ℓ^1_+ by Lemma 3.2.5. Now remember that $C_k = \frac{L_k}{2}d_{\mathcal{C}}^2 + d_{\mathcal{C}}\left(D + M\|T\| + L_h + \|A\|\,\|\mu^\star\|\right)$, where $L_k = \|T\|^2/\beta_k + \|A\|^2\rho_k$. Then we have

$$\begin{aligned} \gamma_k\beta_k M^2/2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_k) + C_k\gamma_k^2 = \gamma_k\beta_k M^2/2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_k) + \|T\|^2\gamma_k^2 d_{\mathcal{C}}/(2\beta_k) + \|A\|^2\rho_k\gamma_k^2 d_{\mathcal{C}}/2 \\ + d_{\mathcal{C}}\left(D + M\|T\| + L_h + \|A\|\,\|\mu^\star\|\right)\gamma_k^2 \in \ell^1_+. \end{aligned}$$

Indeed, under assumption ($\mathbf{P}_1$), the sequences $(\gamma_k\beta_k)_{k\in\mathbb{N}}$, $(\zeta(\gamma_k))_{k\in\mathbb{N}}$, and $\left(\gamma_k^2/\beta_k\right)_{k\in\mathbb{N}}$ belong to ℓ^1_+. Moreover, we have by assumptions ($\mathbf{P}_3$) and ($\mathbf{P}_4$) that $\underline{\rho}\gamma_k^2 \le \rho_k\gamma_k^2 \le \beta_0\overline{\rho}\gamma_k^2/\beta_k$, whence we get that $\left(\rho_k\gamma_k^2\right)_{k\in\mathbb{N}} \in \ell^1_+$ and $\left(\gamma_k^2\right)_{k\in\mathbb{N}} \in \ell^1_+$ after invoking assumption ($\mathbf{P}_1$). Thus all terms on the right hand side are summable. Let

$$\begin{aligned} w_k &\overset{\text{def}}{=} \left[\mathcal{L}(x_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star)\right] + \frac{\rho_k}{2}\|Ax_k - b\|^2 \\ z_k &\overset{\text{def}}{=} \gamma_{k+1}\|Ax_{k+1} - b\|^2 + \gamma_k\beta_k M^2/2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_k) + C_k\gamma_k^2. \end{aligned}$$

So far, we have shown that

$$r_{k+1} \le r_k - \gamma_k w_k + z_k, \tag{3.3.6}$$

where r_k is bounded from below, and $(z_k)_{k\in\mathbb{N}} \in \ell^1_+$. The rest of the proof consists of invoking properly Lemma 2.2.3.

(i) In order to use Lemma 2.2.3(ii), we need to show that for some positive constant α,

$$w_k - w_{k+1} \le \alpha\gamma_k.$$

Notice that the term $\mathcal{L}(x_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star)$ is proportional to γ_k by Lemma 3.2.3. For the second term of w_k, we have by Lemma 3.2.4 that $\frac{\rho_k}{2}\|Ax_k - b\|^2 - \frac{\rho_{k+1}}{2}\|Ax_{k+1} - b\|^2$ is proportional to γ_k. The desired claim then follows from Lemma 2.2.3(ii).

(ii) By [10, Lemma 2.37], we can assert that $(x_k)_{k\in\mathbb{N}}$ possesses a weakly convergent subsequence, say $\left(x_{k_j}\right)_{j\in\mathbb{N}}$, with cluster point $\bar{x}\in\mathcal{C}$. Since $\|A\cdot - b\| \in \Gamma_0(\mathcal{H}_p)$ and in view of [10, Theorem 9.1], we have

$$\|A\bar{x} - b\| \le \liminf_j \left\|Ax_{k_j} - b\right\| = \lim_k \|Ax_k - b\| = 0,$$

where we used lower semicontinuity of the norm and Theorem 3.3.3. Thus $A\bar{x} = 0$, meaning that $\bar{x}$ is a feasible point of ($\mathscr{P}$). In turn, $\mathcal{L}(\bar{x}, \mu^\star) = \Phi(\bar{x})$. The function $\mathcal{L}(\cdot, \mu^\star)$ is lower semicontinuous by ($\mathbf{A}_1$)

and ($\mathbf{A}_6$). Thus, using [10, Theorem 9.1] and by virtue of claim (i), we have

$$\Phi(\bar{x}) = \mathcal{L}(\bar{x}, \mu^\star) \leq \liminf_j \mathcal{L}\left(x_{k_j}, \mu^\star\right) = \lim_k \mathcal{L}(x_k, \mu^\star) = \mathcal{L}(x^\star, \mu^\star) \leq \mathcal{L}(x, \mu^\star)$$

for all $x \in \mathcal{H}_p$, and in particular for all $x \in A^{-1}(b)$. Thus, for every $x \in A^{-1}(b)$, we deduce that

$$\Phi(\bar{x}) \leq \mathcal{L}(x, \mu^\star) = \Phi(x),$$

meaning that $\bar{x}$ is a solution for problem ($\mathscr{P}$).

Recall r_k from (3.2.29) which verifies (3.3.6). From Lemma 2.2.3(i), $(r_k)_{k\in\mathbb{N}}$ is convergent. By ($\mathbf{P}_1$) and ($\mathbf{P}_3$), $(\gamma_k)_{k\in\mathbb{N}}$ and $(\beta_k)_{k\in\mathbb{N}}$ both converge to 0. We also have that, for each solution to ($\mathscr{D}$), $\mu^\star$, for each $k \in \mathbb{N}$,

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

We have from Theorem 3.3.2(i) that $\frac{\rho_k}{2}\|Ax_k - b\|^2 \to 0$. In turn, for each dual solution $\mu^\star$, $\langle \mu^\star - \mu_k, Ax_k - b\rangle \to 0$ since $(\mu_k)_{k\in\mathbb{N}}$ is bounded (Lemma 3.2.7). We also have, for each dual solution $\mu^\star$, $\mathcal{L}(x_k, \mu^\star) \to \mathcal{L}(x^\star, \mu^\star)$ by claim (i) above. By Proposition 2.1.2(v) and (3.1.8), we get that

$$0 \leq \left(g(Tx_k) - g^{\beta_k}(Tx_k)\right) \leq \frac{\beta_k}{2}M^2.$$

Passing to the limit and in view of ($\mathbf{P}_3$), we conclude that $g(Tx_k) - g^{\beta_k}(Tx_k) \to 0$. Altogether, this shows that $\mathcal{L}_k(x_k, \mu_k) \to \mathcal{L}(x^\star, \mu^\star)$. In turn, we conclude that the limit

$$\lim_{k\to\infty}\|\mu_k - \mu^\star\|^2 = 2/c\left(\lim_{k\to\infty} r_k - \mathcal{L}(x^\star, \mu^\star)\right)$$

exists for each solution to the dual problem ($\mathscr{D}$), $\mu^\star$.
By Lemma 3.2.5 we have $\left(\gamma_k\|A\tilde{x}_k - b\|^2\right)_{k\in\mathbb{N}} \in \ell^1_+$ which, by Lemma 2.2.2, implies that there exists a subsequence $\left(A\tilde{x}_{k_j}\right)_{j\in\mathbb{N}}$ with $\left\|A\tilde{x}_{k_j} - b\right\| \to 0$. Since the sequence $(\mu_k)_{k\in\mathbb{N}}$ is bounded by Lemma 3.2.7, the subsequence $\left(\mu_{k_j}\right)_{j\in\mathbb{N}}$ induced by the above is also bounded and thus admits a weakly convergent subsequence $\left(\mu_{k_{j_i}}\right)_{i\in\mathbb{N}}$ with $\mu_{k_{j_i}} \rightharpoonup \bar{\mu}$ for some $\bar{\mu} \in \mathcal{H}_d$. Then, by Fermat's rule ([10, Theorem 16.2]), the weak sequential cluster point $\bar{\mu}$ is a solution to ($\mathscr{D}$) if and only if

$$0 \in \partial\left(\Phi^* \circ (-A^*)\right)(\bar{\mu}) + b.$$

Since the proximal operator is the resolvent of the subdifferential, it follows that (3.2.18) is equivalent, for each $i \in \mathbb{N}$, to

$$\nabla\varphi_{k_{j_i}}\left(\mu_{k_{j_i}}\right) - b \in \partial\left(\phi^*_{k_{j_i}} \circ (-A^*)\right)\left(\mu_{k_{j_i}} - \rho_{k_{j_i}}\nabla\varphi_{k_{j_i}}\left(\mu_{k_{j_i}}\right)\right). \tag{3.3.7}$$

Since $\left(A\tilde{x}_{k_j}\right)_{j\in\mathbb{N}}$ converges strongly to b, and combined with (3.2.19), it holds that $\nabla\varphi_{k_j}\left(\mu_{k_j}\right)$ converges strongly to 0. On the other hand, $\mu_{k_{j_i}} - \rho_{k_{j_i}}\nabla\varphi_{k_{j_i}}\left(\mu_{k_{j_i}}\right)$ converges weakly to $\bar{\mu}$. We now argue that we can pass to the limit in (3.3.7) by showing sequential closedness.
When $g \equiv 0$, we have, for all $i \in \mathbb{N}$, $\phi_{k_{j_i}} \equiv f + h$ and the rest of the argument relies on sequential closedness of the graph of the subdifferential of $\Phi^* \circ (-A^*) \in \Gamma_0(\mathcal{H}_d)$ in the weak-strong topology. For the general case, our argument will rely on the fundamental concept of Mosco convergence of functions, which is epigraphical convergence for both the weak and strong topology (see [24] and [6, Definition 3.7]). By Proposition 2.1.2(v) and assumptions ($\mathbf{A}_1$)-($\mathbf{A}_2$), $\left(\phi_{k_{j_i}}\right)_{j\in\mathbb{N}}$ is an increasing sequence of functions in $\Gamma_0(\mathcal{H}_d)$. It follows from [6, Theorem 3.20(i)] that $\phi_{k_{j_i}}$ Mosco-converges to $\sup_{i\in\mathbb{N}}\phi_{k_{j_i}} = \sup_{i\in\mathbb{N}} f +$

$g^{\beta_{k_{j_i}}} \circ T + h = f + g \circ T + h = \Phi$ since $\beta_{k_{j_i}} \to 0$ by ($\mathbf{P}_3$). Bicontinuity of the Legendre-Fenchel conjugation for the Mosco convergence (see [6, Theorem 3.18]) entails that $\phi^*_{k_{j_i}} \circ -(A^*)$ Mosco-converges to $(f + g \circ T + h)^* \circ (-A^*) = \phi^* \circ (-A^*)$. This implies, via [6, Theorem 3.66], that $\partial\phi^*_{k_{j_i}} \circ (-A^*)$ graph-converges to $\partial\Phi^* \circ (-A^*)$, and [6, Proposition 3.59] shows that $\left(\partial\phi_{k_{j_i}} \circ (-A^*)\right)_{i\in\mathbb{N}}$ is sequentially closed for graph-convergence in the weak-strong topology on $\mathcal{H}_d$, i.e., for any sequence $\left(v_{k_{j_i}}, \eta_{k_{j_i}}\right)_{i\in\mathbb{N}}$ in the graph of $\partial\phi^*_{k_{j_i}} \circ (-A^*)$ such that $v_{k_{j_i}}$ converges weakly to $\bar{v}$ and $\eta_{k_{j_i}}$ converges strongly to $\bar{\eta}$, we have $\bar{\eta} \in \partial\Phi^* \circ (-A^*)(\bar{v})$. Taking $v_{k_{j_i}} = \nabla\varphi_{k_{j_i}}\left(\mu_{k_{j_i}}\right) - b$ and $\eta_{k_{j_i}} = \mu_{k_{j_i}} - \rho_{k_{j_i}}\nabla\varphi_{k_{j_i}}\left(\mu_{k_{j_i}}\right)$, we conclude that

$$0 \in \partial\left(\Phi^* \circ (-A^*)\right)(\bar{\mu}) + b,$$

i.e., $\bar{\mu}$ is a solution of the dual problem ($\mathscr{D}$).
We now invoke ($\mathbf{A}_9$), for which we denote

$$(p_i)_{i\in\mathbb{N}} = \left(\nabla\varphi_{k_{j_i}}\left(\mu_{k_{j_i}}\right) - b\right)_{i\in\mathbb{N}} \quad \text{and} \quad (q_i)_{i\in\mathbb{N}} = \left(\mu_{k_{j_i}} - \rho_{k_{j_i}}\nabla\varphi_{k_{j_i}}\left(\mu_{k_{j_i}}\right)\right)_{i\in\mathbb{N}}.$$

We've shown that $(p_i)_{i\in\mathbb{N}}$ converges strongly to 0 and that $(q_i)_{i\in\mathbb{N}}$ converges weakly to $\bar{\mu}$. Due to (3.3.7), we furthermore have, for each $\omega \in \tilde{\Omega}$, for each $i \in \mathbb{N}$,

$$p_i \in \partial\left(\Phi^*_{k_{j_i}} \circ (-A^*)\right)(q_i),$$

and thus by ($\mathbf{A}_9$), $(q_i)_{i\in\mathbb{N}}$ admits a subsequence $(q_{i_l})_{l\in\mathbb{N}}$ such that $q_{i_l} \to \bar{q}$, i.e., the sequence

$$\left(\mu_{k_{j_{i_l}}} - \rho_{k_{j_{i_l}}}\nabla\varphi_{k_{j_{i_l}}}\left(\mu_{k_{j_{i_l}}}\right)\right)_{l\in\mathbb{N}}$$

is strongly convergent. Thus, the subsequence $\left(\mu_{k_{j_{i_l}}}\right)_{l\in\mathbb{N}}$ is strongly convergent to $\bar{\mu}$. Since $\bar{\mu}$ is a solution to ($\mathscr{D}$), it holds that $\lim_k \|\mu_k - \bar{\mu}\|$ exists. At the same time, we have shown that $\lim_l \left\|\mu_{k_{j_{i_l}}} - \bar{\mu}\right\| = 0$ and so the whole sequence $(\mu_k)_{k\in\mathbb{N}}$ converges strongly to $\bar{\mu}$.

(iii) Recalling that $(\gamma_k)_{k\in\mathbb{N}} \notin \ell^1_+$ (see assumption ($\mathbf{P}_2$)), the rates in (3.3.4) follow by applying Lemma 2.2.3(iii)-(iv) to (3.3.6). Notice that both terms in w_k are positive and that $\rho_k \geq \underline{\rho} > 0$ (see again assumption ($\mathbf{P}_4$)). Therefore we have that, for the same subsequence $\left(x_{k_j}\right)_{j\in\mathbb{N}}$, (3.3.8) holds.

(iv) The ergodic rate (3.3.2) follows by applying the Jensen's inequality to the convex function $\mathcal{L}(\cdot, \mu^\star)$.

□

An important observation is that Theorem 3.3.3, which will be proved in Section 3.3.2, actually shows that

$$\lim_{k\to\infty}\left[\mathcal{L}(x_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) + \frac{\rho_k}{2}\|Ax_k - b\|^2\right] = 0,$$

and subsequentially, for each $j \in \mathbb{N}$,

$$\mathcal{L}\left(x_{k_j}, \mu^\star\right) - \mathcal{L}(x^\star, \mu^\star) + \frac{\rho_{k_j}}{2}\|Ax_{k_j} - b\|^2 \leq \frac{1}{\Gamma_{k_j}}. \tag{3.3.8}$$

This means, in particular, that the pointwise rate for feasibility and optimality hold simulatenously for the same subsequence.

The following corollary is immediate.

Corollary 3.3.4. *Under the assumptions of* Theorem 3.3.3, *if the problem* ($\mathscr{P}$) *admits a unique solution* $x^\star$, *then the primal sequence* $(x_k)_{k\in\mathbb{N}}$ *converges weakly to a solution of the primal problem,* ($\mathscr{P}$). *Moreover, if* Φ *is uniformly convex on* $\mathcal{C}$ *with modulus* $\psi : \mathbb{R}_+ \to [0, +\infty]$, *then* $(x_k)_{k\in\mathbb{N}}$ *converges strongly to* $x^\star$ *at the ergodic rate*

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

Proof. By uniqueness, it follows from Theorem 3.3.3(ii) that $(x_k)_{k\in\mathbb{N}}$ has exactly one weak sequential cluster point which is the solution to $(\mathscr{P})$. Weak convergence of the sequence $(x_k)_{k\in\mathbb{N}}$ then follows from [10, Lemma 2.38].

From [10, Proposition 19.21(v)], we know that $-A^*\mu^\star \in \partial\Phi(x^\star)$. This together with ψ-uniform convexity of Φ imply that

$$\Phi(x) \geq \Phi(x^\star) + \langle -A^*\mu^\star,\ x - x^\star\rangle + \psi\left(\|x - x^\star\|\right), \quad \forall x \in \mathcal{C},$$

where ψ is a an increasing non-negative funtion that vanishes only at 0. This is equivalent to

$$\psi\left(\|x - x^\star\|\right) \leq \mathcal{L}\left(x, \mu^\star\right) - \mathcal{L}\left(x^\star, \mu^\star\right), \quad \forall x \in \mathcal{C}.$$

Applying this inequality to $x = x_k$, passing to the limit and using (3.3.3), we get $\psi\left(\|x_k - x^\star\|\right) \to 0$ which forces strong convergence of x_k by assumption on ψ. The ergodic rate follows from the same above inequality applied to $x = \bar{x}_k$. □

3.4 Comparison

In this section we compare and contrast CGALP with some other contemporary works which studied algorithms sharing elements with CGALP . We have deferred this comparison to now to make use of the previous section to concretely highlight the differences in the algorithms and approaches.

3.4.1 Conditional Gradient Framework

In [116] the following problem was analyzed in the finite-dimensional setting,

$$\min_{x\in\mathcal{C}} \{f(x) + g(Tx)\} \tag{3.4.1}$$

where $f \in \Gamma_0(\mathbb{R}^n)$ is Lipschitz-smooth, $T \in \mathbb{R}^{d\times n}$ is a linear operator, $g \circ T \in \Gamma_0(\mathbb{R}^n)$, and $\mathcal{C}$ is a compact, convex subset of $\mathbb{R}^n$. They develop an algorithm which avoids projecting onto the set $\mathcal{C}$, instead utilizing a linear minimization oracle $\mathrm{lmo}_{\mathcal{C}}(v) = \operatorname{Argmin}_{x\in\mathcal{C}} \langle x, v\rangle$, and replaces the function $g\circ T$ with the smooth function $g_k^\beta \circ T$. They consider only functions f which are Lipschitz-smooth and finite dimensional spaces, i.e. $\mathbb{R}^n$, compared to CGALP which weakens the assumptions on f to be differentiable and (F,ζ)-smooth (see Definition 2.1.13) with an arbitrary real Hilbert space $\mathcal{H}_p$ (possibly infinite dimensional). Furthermore, the analysis in [116] is restricted to the parameter choices $\gamma_k = \frac{2}{k+1}$ and $\beta_k = \frac{\beta_0}{\sqrt{k+1}}$ exclusively, although they do include a section in which they consider two variants of an inexact linear minimization oracle: one with additive noise and one with multiplicative noise. In contrast, the results we present in Section 3.3 show optimality and feasibility for a wider choice for both the sequence of stepsizes $(\gamma_k)_{k\in\mathbb{N}}$ and the sequence of smoothing parameters $(\beta_k)_{k\in\mathbb{N}}$, although we only consider exact linear perturbation oracles of the form $\operatorname{Argmin}_{s\in\mathcal{H}_p} \{h(s) + \langle x, s\rangle\}$. Finally, for solving (3.4.1) with an exact linear minimization oracle, our algorithm encompasses the algorithm in [116] by choosing $h(x) = \iota_{\mathcal{C}}(x)$, $A \equiv 0$, and restricting f to be in $C^{1,1}(\mathcal{H})$ with $\mathcal{H} = \mathbb{R}^n$.

In [116, Section 5] there is a discussion on splitting and affine constraints using the conditonal gradient framework presented. In this setting, i.e. assuming exact oracles, the primary difference between CGALP and the conditional gradient framework is the approach each algorithm takes to handle affine constraints. In CGALP , the augmented Lagrangian formulation is used to account for the affine constraints, introducing a dual variable μ and both a linear and quadratic term for the constraint $Ax - b = 0$. In contrast, in [116] the affine constraint is treated the same as the nonsmooth term $g \circ T$ and thus handled by quadratic penalization/smoothing alone. The consequence of smoothing for the affine constraint $Ax = b$ comes from calculating the gradient of the squared-distance to the constraint. This will involve solving a least squares problem at each iteration which can be computationally expensive. Our algorithm does not need to solve such a linear system.

The difference in the approaches is highlighted when both methods are applied to problem presented in Section 3.5.3 with $n = 2$ since this problem necessitates an affine constraint $\Pi_{\mathcal{V}^\perp} \boldsymbol{x} = 0$ for splitting. According to [116, Section 5], we reformulate the problem to be

$$\min_{\substack{x^{(1)} \in \mathcal{C}_1 \\ x^{(2)} \in C_2}} \left\{ \frac{1}{2} \left(f\left(x^{(1)}\right) + f\left(x^{(2)}\right) \right) + \iota_{\left\{x^{(1)}\right\}} \left(x^{(2)}\right) \right\}.$$

Note that the inclusion of the function $\iota_{\left\{x^{(1)}\right\}} \left(x^{(2)}\right)$ in the objective is equivalent to the affine constraint $\Pi_{\mathcal{V}^\perp} \boldsymbol{x} = 0$ in the $n = 2$ case. Apply the conditional gradient framework on the variable $\left(x^{(1)}, x^{(2)}\right)$ to get

$$\boldsymbol{s}_k \in \underset{\substack{s^{(1)} \in \mathcal{C}_1 \\ s^{(2)} \in \mathcal{C}_2}}{\operatorname{Argmin}} \left\{ \left\langle \begin{pmatrix} s^{(1)} \\ s^{(2)} \end{pmatrix}, \begin{pmatrix} \nabla_{x^{(1)}} \left[\frac{1}{2} f\left(x_k^{(1)}\right) + \iota^{\beta_k}_{x_k^{(2)}} \left(x_k^{(1)}\right) \right] \\ \nabla_{x^{(2)}} \left[\frac{1}{2} f\left(x_k^{(2)}\right) + \iota^{\beta_k}_{x_k^{(1)}} \left(x_k^{(2)}\right) \right] \end{pmatrix} \right\rangle \right\},$$

which leads to a separable scheme that can be computed component-wise,

$$\begin{aligned} s_k^{(1)} &\in \underset{s \in \mathcal{C}_1}{\operatorname{Argmin}} \left\langle s, \frac{1}{2} \nabla f\left(x_k^{(1)}\right) + \frac{x_k^{(1)} - x_k^{(2)}}{\beta_k} \right\rangle \\ s_k^{(2)} &\in \underset{s \in \mathcal{C}_2}{\operatorname{Argmin}} \left\langle s, \frac{1}{2} \nabla f\left(x_k^{(2)}\right) + \frac{x_k^{(2)} - x_k^{(1)}}{\beta_k} \right\rangle. \end{aligned} \tag{3.4.2}$$

Compare the direction obtained in (3.4.2) to the one obtained in (3.5.4), the components of which we rewrite below for $n = 2$,

$$\begin{aligned} s_k^{(1)} &\in \underset{s \in \mathcal{C}_1}{\operatorname{Argmin}} \left\langle s, \frac{1}{2} \nabla f\left(x_k^{(1)}\right) + \frac{1}{2}\left(\mu_k^{(1)} - \mu_k^{(2)}\right) + \frac{\rho_k}{2}\left(x_k^{(1)} - x_k^{(2)}\right) \right\rangle \\ s_k^{(2)} &\in \underset{s \in \mathcal{C}_2}{\operatorname{Argmin}} \left\langle s, \frac{1}{2} \nabla f\left(x_k^{(2)}\right) + \frac{1}{2}\left(\mu_k^{(2)} - \mu_k^{(1)}\right) + \frac{\rho_k}{2}\left(x_k^{(2)} - x_k^{(1)}\right) \right\rangle. \end{aligned} \tag{3.4.3}$$

Due to affine constraint, the computation of the direction in (3.4.2) necessitates smoothing and, as a consequence, the parameter β_k, which is necessarily going to 0. In CGALP , the introduction of the dual variable $\boldsymbol{\mu}_k$ in place of smoothing the affine constraint avoids the parameter β_k. Instead, we have the parameter ρ_k but ρ_k can be picked to be constant without issue.

3.4.2 FW-AL Algorithm

In [56] the following problem was analyzed,

$$\min_{\substack{x \in \bigcap_{i=1}^n \mathcal{C}_i \\ Ax = 0}} f(x)$$

using a combination of the Frank-Wolfe algorithm with the augmented Lagrangian to account for the constraint $Ax = 0$. The function f is assumed to be Lipschitz-smooth, in contrast to our approach. The perspective used in their paper is to modify the classic ADMM algorithm, replacing the marginal minimization with respect to the primal variable by a Frank-Wolfe step instead, although their analysis is not restricted only to Frank-Wolfe steps. Indeed, in all the scenarios where one can apply FW-AL using a Frank-Wolfe step our algorithm encompasses FW-AL as a special case, discussed in Section 3.5.3. The primary differences between CGALP and FW-AL are in the convergence results and the generality of CGALP . The results in [56] prove convergence of the objective in the case where the sets $\mathcal{C}_i$ are polytopes and convergence of the iterates in the case where the sets $\mathcal{C}_i$ are polytopes and f is a generalized strongly convex function,[2] but they do not prove convergence of the objective, convergence

[2] A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be generalized strongly convex if it can be written $f(x) = g(Ax) + \langle b, x \rangle$ for all $x \in \mathbb{R}^n$ where $b \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times m}$ is a linear operator, and $g : \mathbb{R}^m \to \mathbb{R}$ is strongly convex with respect to the euclidean norm.

(or even boundedness) of the dual variable, or asymptotic feasibility of the iterates in the general case where each $\mathcal{C}_i$ is a compact, convex set. Instead, they prove two theorems which imply subsequential convergence of the objective and subsequential asymptotic feasibility in the general case and subsequential convergence of the iterates to the optimum in the generalized strongly convex case in [56, Theorem 2] and [56, Corollary 2] respectively. Unfortunately, each of these results is obtained separately and so the subsequences that produce each result are not guaranteed to coincide with one another.

Interestingly, the results they obtain are not unique to Frank-Wolfe style algorithms as their analysis is from the perspective of a modified ADMM algorithm; they only require that the algorithm used to replace the marginal minimization on the primal variable in ADMM produces sublinear decrease in the objective. Finally, they do not provide conditions for the dual multiplier sequence, μ_k in our notation, to be bounded as they discuss in their analysis of issues with similar proofs, e.g. in GDMM. This is a crucial issue as the constants in their bounds depend on the norm of these dual multipliers.

3.5 Applications

3.5.1 Sum of Several Nonsmooth Functions

In this section we explore the applications of Algorithm 8 to splitting in composite optimization problems, where we allow the presence of more than one nonsmooth function g or h in the objective:

$$\min_{x\in\mathcal{H}_p}\left\{f(x)+\sum_{i=1}^{n}g_i(T_ix)+\sum_{i=1}^{n}h_i(x)\right\}. \tag{3.5.1}$$

First, we denote the product space by $\boldsymbol{\mathcal{H}}_p \overset{\text{def}}{=} \mathcal{H}_p^n$ endowed with the scalar product $\langle \boldsymbol{x},\boldsymbol{y}\rangle = \frac{1}{n}\sum_{i=1}^{n}\langle x^{(i)},y^{(i)}\rangle$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors in $\boldsymbol{\mathcal{H}}_p$ with $\boldsymbol{x} \overset{\text{def}}{=} \left(x^{(1)},\ldots,x^{(n)}\right)^\top$. We define also $\mathcal{V}$ as the diagonal subspace of $\boldsymbol{\mathcal{H}}_p$, i.e. $\mathcal{V} \overset{\text{def}}{=} \{\boldsymbol{x}\in\boldsymbol{\mathcal{H}}_p : x^{(1)}=\ldots=x^{(n)}\}$, $\mathcal{V}^\perp$ the orthogonal subspace to $\mathcal{V}$, and $\Pi_{\mathcal{V}},\Pi_{\mathcal{V}^\perp}$ the orthogonal projections onto $\mathcal{V}$, $\mathcal{V}^\perp$ - respectively. We finally introduce the (diagonal) linear operator $\boldsymbol{T}:\boldsymbol{\mathcal{H}}_p\to\boldsymbol{\mathcal{H}}_p$ defined by

$$[\boldsymbol{T}(\boldsymbol{x})]^{(i)} = T_ix^{(i)}$$

and the functions

$$F(\boldsymbol{x}) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}f\left(x^{(i)}\right);\quad G(\boldsymbol{T}\boldsymbol{x}) \overset{\text{def}}{=} \sum_{i=1}^{n}g_i\left(T_ix^{(i)}\right);\quad H(\boldsymbol{x}) \overset{\text{def}}{=} \sum_{i=1}^{n}h_i\left(x^{(i)}\right).$$

Then problem (3.5.1) is obviously equivalent to

$$\min_{\boldsymbol{x}\in\boldsymbol{\mathcal{H}}_p}\left\{F(\boldsymbol{x})+G(\boldsymbol{T}\boldsymbol{x})+H(\boldsymbol{x}):\ \Pi_{\mathcal{V}^\perp}\boldsymbol{x}=0\right\}, \tag{3.5.2}$$

which fits in the setting of our main problem ($\mathscr{P}$). In order to the presentation clearer, we separate the two cases of multiple g and multiple h, that can be trivially combined. Moreover, we focus on the main case involving indicator functions, e.g., $h_i=\iota_{\mathcal{C}_i}$.

3.5.2 Sum of Several Simple Functions Over a Compact Set

Consider the following composite minimization problem,

$$\min_{x\in\mathcal{C}}\left\{f(x)+\sum_{i=1}^{n}g_i(T_ix)\right\}. \tag{3.5.3}$$

We can reformulate the problem in the product space $\boldsymbol{\mathcal{H}}_p$ using the above notation to get,

$$\min_{\boldsymbol{x}\in\mathcal{C}^n\cap\mathcal{V}}\left\{F(\boldsymbol{x})+G(\boldsymbol{T}\boldsymbol{x})\right\}.$$

Applying Algorithm 8 to this problem gives a completely separable scheme; we first compute the direction,

$$\boldsymbol{s}_k \in \underset{\boldsymbol{s}\in\mathcal{C}^n\cap\mathcal{V}}{\operatorname{Argmin}} \langle \nabla \left(F\left(\boldsymbol{x}_k\right) + G^{\beta_k}\left(\boldsymbol{T}\boldsymbol{x}_k\right)\right), \boldsymbol{s}\rangle,$$

which reduces to the following computation since $\boldsymbol{s}_k = \begin{pmatrix} s_k \\ \vdots \\ s_k \end{pmatrix}$ has identical components,

$$s_k \in \underset{s\in\mathcal{C}}{\operatorname{Argmin}} \left\langle \sum_{i=1}^{n} \left(\frac{1}{n}\nabla f\left(x_k^{(i)}\right) + \nabla g_i^{\beta_k}\left(T_i x_k^{(i)}\right)\right), s \right\rangle.$$

The term $\nabla g_i^{\beta_k}$ has a closed form given in Proposition 2.1.2 which can be used to get the following formula for the direction,

$$s_k \in \underset{s\in\mathcal{C}}{\operatorname{Argmin}} \left\langle \sum_{i=1}^{n} \left(\frac{1}{n}\nabla f\left(x_k^{(i)}\right) + \frac{1}{\beta_k} T_i^*\left(T_i x_k^{(i)} - \operatorname{prox}_{\beta g}\left(T_i x_k^{(i)}\right)\right)\right), s \right\rangle.$$

3.5.3 Minimizing Over Intersection of Compact Sets

A classical problem found in machine learning is to minimize a Lipschitz-smooth function f over the intersection of convex, compact sets $\mathcal{C}_i$ in some real Hilbert space $\mathcal{H}$,

$$\min_{x\in\bigcap_{i=1}^{n}\mathcal{C}_i} f(x) = \min_{x\in\mathcal{H}} \left\{ f(x) + \sum_{i=1}^{n} h_i(x) \right\},$$

where $h_i \equiv \iota_{\mathcal{C}_i}$. Reformulating the problem in the product space $\boldsymbol{\mathcal{H}}_p$ gives,

$$\min_{\substack{\boldsymbol{x}\in\boldsymbol{\mathcal{H}}_p \\ \Pi_{\mathcal{V}^\perp}\boldsymbol{x}=0}} \left\{ F(\boldsymbol{x}) + H(\boldsymbol{x}) \right\}.$$

Then, we can apply Algorithm 8 and compute the step direction

$$\boldsymbol{s}_k \in \underset{\boldsymbol{s}\in\mathcal{C}_1\times\ldots\times\mathcal{C}_n}{\operatorname{Argmin}} \langle \boldsymbol{s}, \nabla \left[F(\boldsymbol{x}) + \langle \boldsymbol{\mu}_k, \Pi_{\mathcal{V}^\perp}\boldsymbol{x}_k\rangle_+ \frac{\rho_k}{2} \left\|\Pi_{\mathcal{V}^\perp}\boldsymbol{x}_k\right\|^2 \right] \rangle$$

which gives a separable scheme for each component of $\boldsymbol{s}_k = \begin{pmatrix} s_k^{(1)} \\ \vdots \\ s_k^{(n)} \end{pmatrix}$,

$$\begin{aligned} s_k^{(i)} &\in \underset{s\in\mathcal{C}_i}{\operatorname{Argmin}} \left\langle s, \frac{1}{n}\nabla f\left(x_k^{(i)}\right) + \left(\Pi_{\mathcal{V}^\perp}\boldsymbol{\mu}_k\right)^{(i)} + \rho_k \left(\Pi_{\mathcal{V}^\perp}\boldsymbol{x}_k\right)^{(i)} \right\rangle \\ &= \underset{s\in\mathcal{C}_i}{\operatorname{Argmin}} \left\langle s, \frac{1}{n}\nabla f\left(x_k^{(i)}\right) + \mu_k^{(i)} - \frac{1}{n}\sum_{j=1}^{n}\mu_k^{(j)} + \rho_k \left(x_k^{(i)} - \frac{1}{n}\sum_{j=1}^{n} x_k^{(j)} \right) \right\rangle. \end{aligned} \tag{3.5.4}$$

3.6 Numerical Experiments

In this section we present some numerical experiments comparing the performance of Algorithm 8 and a proximal algorithm applied to splitting in composite optimization problems.

3.6.1 Projection Problem

First, we consider a simple projection problem,

$$\min_{x\in\mathbb{R}^2} \left\{ \frac{1}{2}\|x-y\|_2^2 : \ \|x\|_1 \le 1, Ax = 0 \right\}, \tag{3.6.1}$$

where $y \in \mathbb{R}^2$ is the vector to be projected and $A : \mathbb{R}^2 \to \mathbb{R}^2$ is a rank-one matrix. To exclude trivial projections, we choose randomly $y \notin \mathbb{B}_1^1 \cap \ker(A)$, where $\mathbb{B}_1^1$ is the unit ℓ^1 ball centered at the origin. Then Problem (3.6.1) is nothing but Problem ($\mathscr{P}$) with $f(x) = \frac{1}{2}\|x - y\|_2^2$, $g \equiv 0$, $h \equiv \iota_{\mathbb{B}_1^1}$ and $\mathcal{C} = \mathbb{B}_1^1$.

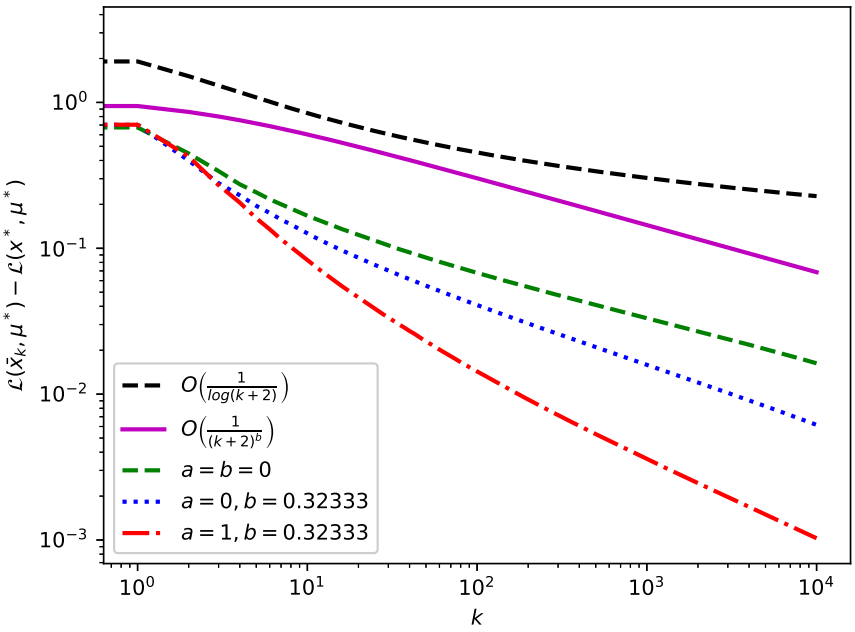

Figure 3.1: Ergodic convergence profiles for CGALP applied to the simple projection problem.

The assumptions mentioned previously, i.e. $(\mathbf{A}_1)$-$(\mathbf{A}_8)$, all hold in this finite-dimensional case as f, g, and h are all in $\Gamma_0\left(\mathbb{R}^2\right)$, f is Lipschitz-smooth, h is the indicator function for a compact convex set, g has full domain and $0 \in \ker(A) \cap \operatorname{int}(\mathcal{C})$. Regarding the parameters and the associated assumptions, we choose γ_k according to Example 3.1.3 with $(a, b) \in \{(0, 0), (0, 1/3 - 0.01), (1, 1/3 - 0.01)\}$, $\theta_k = \gamma_k$, and $\rho = 2^{2-b} + 1$. The ergodic convergence profiles of the Lagrangian are displayed in Figure 3.1 along with the theoretical rates (see Theorem 3.3.3 and Example 3.3.1). The observed rates agree with the predicted ones of $O\left(\frac{1}{\log(k+2)}\right)$, $O\left(\frac{1}{(k+2)^b}\right)$ and $o\left(\frac{1}{(k+2)^b}\right)$ for the respective choices of (a, b).

3.6.2 Matrix Completion Problem

We also consider the following, more complicated matrix completion problem,

$$\min_{X \in \mathbb{R}^{N \times N}} \left\{ \|\Omega X - y\|_1 : \ \|X\|_* \leq \delta_1, \|X\|_1 \leq \delta_2 \right\}, \tag{3.6.2}$$

where δ_1 and δ_2 are positive constants, $\Omega : \mathbb{R}^{N \times N} \to \mathbb{R}^p$ is a masking operator, $y \in \mathbb{R}^p$ is a vector of observations, and $\|\cdot\|_*$ and $\|\cdot\|_1$ are respectively the nuclear and ℓ^1 norms. The mask operator Ω is generated randomly by specifying a sampling density, in our case 0.8. We generate the vector y randomly in the following way. We first generate a sparse vector $\tilde{y} \in \mathbb{R}^N$ with $N/5$ non-zero entries independently uniformly distributed in $[-1, 1]$. We take the exterior product $\tilde{y}\tilde{y}^\top = X_0$ to get a rank-1 sparse matrix which we then mask to get ΩX_0. The radii of the contraints in (3.6.2) are chosen according to the nuclear norm and ℓ^1 norm of X_0, $\delta_1 = \frac{\|X_0\|_*}{2}$ and $\delta_2 = \frac{\|X_0\|_1}{2}$.

3.6.2.1 CGALP

Problem (3.6.2) is a special instance of (3.5.1) with $n = 2$, $f \equiv 0$, $g_i = \|\cdot - y\|_1/2$, $T_i = \Omega$, $h_1 = \iota_{\mathbb{B}_*^{\delta_1}}$, $h_2 = \iota_{\mathbb{B}_1^{\delta_2}}$, where $\mathbb{B}_*^{\delta_1}$ and $\mathbb{B}_1^{\delta_2}$ are the nuclear and ℓ^1 balls of radii δ_1 and δ_2. We then follow the same steps as in Section 3.5.1. Let $\mathcal{H}_p = \mathbb{R}^{N\times N}$, $\boldsymbol{\mathcal{H}}_p = \mathcal{H}_p^2$, $\boldsymbol{X} = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \in \boldsymbol{\mathcal{H}}_p$. We then have $G(\Omega \boldsymbol{X}) = \frac{1}{2}\left(\left\|\Omega X^{(1)} - y\right\|_1 + \left\|\Omega X^{(2)} - y\right\|_1\right)$, and $H(\boldsymbol{X}) = \iota_{\mathbb{B}_*^{\delta_1}}(X^{(1)}) + \iota_{\mathbb{B}_1^{\delta_2}}(X^{(2)})$. Then problem (3.6.2) is obviously equivalent to

$$\min_{\boldsymbol{X}\in\boldsymbol{\mathcal{H}}_p} \left\{G(\Omega\boldsymbol{X}) + H(\boldsymbol{X}) : \Pi_{\mathcal{V}^\perp}\boldsymbol{X} = 0\right\}, \tag{3.6.3}$$

which is a special case of (3.5.2) with $F \equiv 0$. It is immediate to check that our assumptions $(\mathbf{A}_1)$-$(\mathbf{A}_8)$ hold. Indeed, all functions are in $\Gamma_0(\mathcal{H}_p)$ and $F \equiv 0$, and thus $(\mathbf{A}_1)$ and $(\mathbf{A}_2)$ are verified. $\mathcal{C} = \mathbb{B}_*^{\delta_1} \times \mathbb{B}_1^{\delta_2}$ which is a non-mepty convex compact set. We also have $\Omega\mathcal{C} \subset \mathrm{dom}(\partial G) = \mathbb{R}^p \times \mathbb{R}^p$, and for any $\boldsymbol{z} \in \mathbb{R}^p \times \mathbb{R}^p$, $\partial G(\boldsymbol{z}) \subset \mathbb{B}_\infty^{1/2} \times \mathbb{B}_\infty^{1/2}$ and thus $(\mathbf{A}_4)$ is verified. $(\mathbf{A}_5)$ also holds with $L_h = 0$. $\mathcal{V}$ is closed as we are in finite dimension, and thus $(\mathbf{A}_7)$ is fulfilled. We also have, since $\mathrm{dom}(G\circ\Omega) = \boldsymbol{\mathcal{H}}_p$,

$$\mathbf{0} \in \mathcal{V} \cap \mathrm{int}\left(\mathrm{dom}(G\circ\Omega)\right) \cap \mathrm{int}\left(\mathcal{C}\right) = \mathcal{V} \cap \mathrm{int}(\mathbb{B}_*^{\delta_1}) \times \mathrm{int}(\mathbb{B}_1^{\delta_2}),$$

which shows that $(\mathbf{A}_8)$ is verified. The latter is nothing but the condition in [10, Fact 15.25(i)]. It then follows from the discussion in Remark 3.1.1(iv) that $(\mathbf{A}_6)$ holds true.

We use Algorithm 8 by choosing the sequence of parameters $\gamma_k = \frac{1}{k+1}$, $\beta_k = \frac{1}{\sqrt{k+1}}$, $\theta_k = \gamma_k$, and $\rho_k \equiv 15$, which verify all our assumptions $(\mathbf{P}_1)$-$(\mathbf{P}_7)$ in view of Example 3.1.3. Our choice of γ_k is the most common in the literature, and it can be improved according to our discussion in the previous section.

Finding the direction $\boldsymbol{S}_k$ by solving the linear minimization oracle is a separable problem, and thus each component is given by,

$$\begin{aligned}
S_k^{(1)} \in \underset{S^{(1)}\in\mathbb{B}^{\delta_1}_{\|\cdot\|_*}}{\mathrm{Argmin}} \Bigg\langle & \frac{\Omega^*\left(\Omega X_k^{(1)} - y - \mathrm{prox}_{\frac{\beta_k}{2}\|\cdot\|_1}\left(\Omega X_k^{(1)} - y\right)\right)}{\beta_k} \\
& + \frac{1}{2}\left(\mu_k^{(1)} - \mu_k^{(2)} + \rho_k\left(X_k^{(1)} - X_k^{(2)}\right)\right), S^{(1)}\Bigg\rangle, \\
S_k^{(2)} \in \underset{S^{(2)}\in\mathbb{B}^{\delta_2}_{\|\cdot\|_1}}{\mathrm{Argmin}} \Bigg\langle & \frac{\Omega^*\left(\Omega X_k^{(2)} - y - \mathrm{prox}_{\frac{\beta_k}{2}\|\cdot\|_1}\left(\Omega X_k^{(2)} - y\right)\right)}{\beta_k} \\
& + \frac{1}{2}\left(\mu_k^{(2)} - \mu_k^{(1)} + \rho_k\left(X_k^{(2)} - X_k^{(1)}\right)\right), S^{(2)}\Bigg\rangle.
\end{aligned} \tag{3.6.4}$$

Because of the structure of the sets $\mathbb{B}^{\delta_1}_{\|\cdot\|_*}$ and $\mathbb{B}^{\delta_2}_{\|\cdot\|_1}$, finding the first component of $\boldsymbol{S}_k$ reduces to computing the leading right and left singular vectors of

$$\frac{\Omega^*\left(\Omega X_k^{(1)} - y - \mathrm{prox}_{\frac{\beta_k}{2}\|\cdot\|_1}\left(\Omega X_k^{(1)} - y\right)\right)}{\beta_k} + \frac{1}{2}\left(\mu_k^{(1)} - \mu_k^{(2)} + \rho_k\left(X_k^{(1)} - X_k^{(2)}\right)\right)$$

while finding the second component reduces to computing the largest entry of

$$\left|\left(\frac{\Omega^*\left(\Omega X_k^{(2)} - y - \mathrm{prox}_{\frac{\beta_k}{2}\|\cdot\|_1}\left(\Omega X_k^{(2)} - y\right)\right)}{\beta_k} + \frac{1}{2}\left(\mu_k^{(2)} - \mu_k^{(1)} + \rho_k\left(X_k^{(2)} - X_k^{(1)}\right)\right)\right)_{(i,j)}\right|$$

over all the entries (i,j). The dual variable update is given by,

$$\boldsymbol{\mu}_{k+1} \overset{\text{def}}{=} \begin{pmatrix} \mu_{k+1}^{(1)} \\ \mu_{k+1}^{(2)} \end{pmatrix} = \begin{pmatrix} \mu_k^{(1)} \\ \mu_k^{(2)} \end{pmatrix} + \frac{\gamma_k}{2}\begin{pmatrix} X_{k+1}^{(1)} - X_{k+1}^{(2)} \\ X_{k+1}^{(2)} - X_{k+1}^{(1)} \end{pmatrix}.$$

3.6.2.2 GFB

Let $\mathcal{H}_p = \mathbb{R}^{N\times N}$, $\boldsymbol{\mathcal{H}}_p = \mathcal{H}_p^3$, $\boldsymbol{W} = \begin{pmatrix} W^{(1)} \\ W^{(2)} \\ W^{(3)} \end{pmatrix} \in \boldsymbol{\mathcal{H}}_p$, $Q(\boldsymbol{W}) = \left\|\Omega W^{(1)} - y\right\|_1 + \iota_{\mathbb{B}^{\delta_1}_{\|\cdot\|_*}}\left(W^{(2)}\right) + \iota_{\mathbb{B}^{\delta_2}_{\|\cdot\|_1}}\left(W^{(3)}\right)$. Then we reformulate problem (3.6.2) as

$$\min_{\boldsymbol{W}\in\boldsymbol{\mathcal{H}}_p} \left\{Q(\boldsymbol{W}) : \boldsymbol{W} \in \mathcal{V}\right\}, \tag{3.6.5}$$

which fits the framework to apply the GFB algorithm proposed in [98] (in fact Douglas-Rachford since the smooth part vanishes).

The algorithm has three steps, each of which is separable in the components. We choose the step sizes $\lambda_k = \gamma = 1$ in the GFB to get,

$$\begin{cases} \boldsymbol{U}_{k+1} = \begin{pmatrix} 2W_k^{(1)} - Z_k^{(1)} + \Omega^*\left(y - \Omega\left(2W_k^{(1)} - Z_k^{(1)}\right) + \operatorname{prox}_{\|\cdot\|_1}\left(\Omega\left(2W_k^{(1)} - Z_k^{(1)}\right) - y\right)\right) \\ \Pi_{\mathbb{B}^{\delta_1}_{\|\cdot\|_*}}\left(2W_k^{(2)} - Z_k^{(2)}\right) \\ \Pi_{\mathbb{B}^{\delta_2}_{\|\cdot\|_1}}\left(2W_k^{(3)} - Z_k^{(3)}\right) \end{pmatrix} \\ \boldsymbol{Z}_{k+1} = \boldsymbol{Z}_k + \boldsymbol{U}_{k+1} - \boldsymbol{W}_k \\ \boldsymbol{W}_{k+1} = \begin{pmatrix} \sum_{i=1}^3 Z_{k+1}^{(i)}/3 \\ \sum_{i=1}^3 Z_{k+1}^{(i)}/3 \\ \sum_{i=1}^3 Z_{k+1}^{(i)}/3 \end{pmatrix} \end{cases} \tag{3.6.6}$$

We know from [98] that $\boldsymbol{Z}_k$ converges to $\boldsymbol{Z}^\star$, and $\boldsymbol{W}_k$ and $\boldsymbol{U}_k$ both converge to $\boldsymbol{W}^\star = \Pi_{\mathcal{V}}(\boldsymbol{Z}^\star) = (X^\star, X^\star, X^\star)$, where $X^\star$ is a minimizer of (3.6.2).

3.6.2.3 Results

We compare the performance of CGALP with GFB for varying dimension, N, using their respective ergodic convergence criteria. For CGALP this is the quantity $\mathcal{L}\left(\bar{\boldsymbol{X}}_k, \mu^*\right) - \mathcal{L}\left(\boldsymbol{X}^\star, \boldsymbol{\mu}^\star\right)$ where $\bar{\boldsymbol{X}}_k = \sum_{i=0}^{k} \gamma_i \boldsymbol{X}_i/\Gamma_k$. Meanwhile, for GFB, we know from [83] that the Bregman divergence $D_Q^{\boldsymbol{v}^\star}\left(\bar{\boldsymbol{U}}_k\right) = Q(\bar{\boldsymbol{U}}_k) - Q(\boldsymbol{W}^\star) - \left\langle \boldsymbol{v}^\star, \bar{\boldsymbol{U}}_k - \boldsymbol{W}^\star\right\rangle$, with $\bar{\boldsymbol{U}}_k = \sum_{i=0}^{k} \boldsymbol{U}_i/(k+1)$ and $\boldsymbol{v}^\star = (\boldsymbol{W}^\star - \boldsymbol{Z}^\star)/\gamma$, converges at the rate $O(1/(k+1))$. To compute the convergence criteria, we first run each algorithm for 10^5 iterations to approximate the optimal variables ($\boldsymbol{X}^\star$ and $\boldsymbol{\mu}^\star$ for CGALP, and $\boldsymbol{Z}^\star$ and $\boldsymbol{W}^\star$ for GFB). Then, we run each algorithm again for 10^5 iterations, this time recording the convergence criteria at each iteration. The results are displayed in Figure 3.2.

It can be observed that our theoretically predicted rate (which is $O\left(1/\log(k+2)\right)$ for CGALP according to Theorem 3.3.3 and Example 3.3.1) is in close agreement with the observed one. On the other hand, as is very well-known, employing a proximal step for the nuclear ball constraint will necessitate to compute an SVD which is much more time consuming than computing the linear minimization oracle for large N. For this reason, even though the rates of convergence guaranteed for CGALP are slower than for GFB, one can expect CGALP to be a more time computationally efficient algorithm for large N.

Visualizing the resulting matrices, we can see that both have correctly identified the support. Similarly, both have identified the correct sign; for visualization purpose we have thus taken the absolute values of the entries of the solutions to assist the visualization.

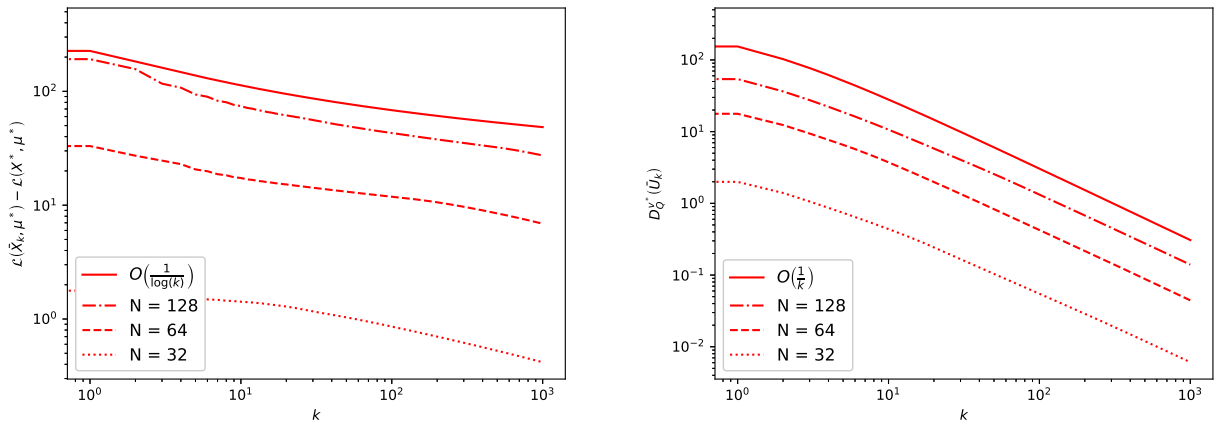

Figure 3.2: Convergence profiles for CGALP (left) and GFB (right) for $N = 32$, $N = 64$, and $N = 128$.

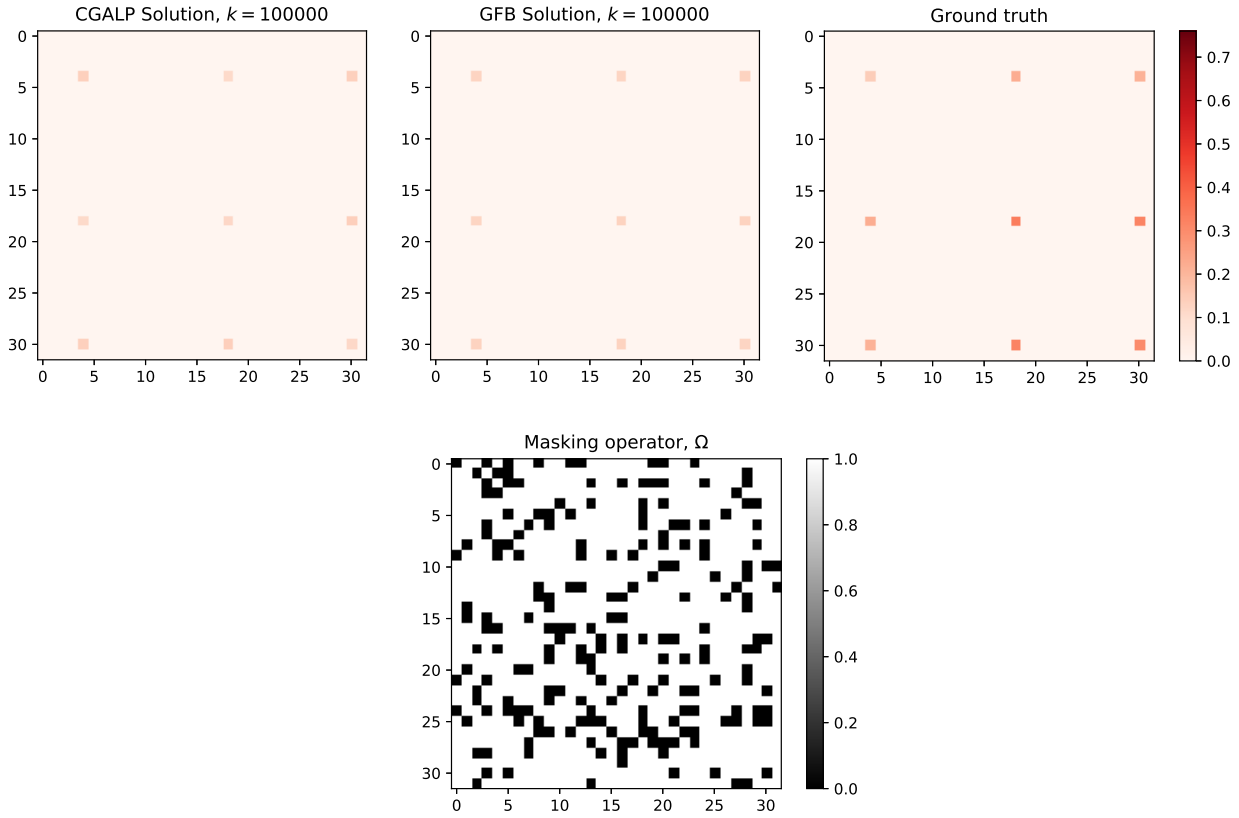

Figure 3.3: Completed matrices for the $n = 32$ matrix completion problem, with ground truth and masking matrix shown as well.

Chapter 4

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

In this chapter, we propose and analyze inexact and stochastic versions of the CGALP algorithm developed in Chapter 3, which we denote ICGALP , that allow for solving problems of the form

$$\min_{x\in\mathcal{H}} \{f(x) + g(Tx) + h(x) : Ax = b\} \tag{4.0.1}$$

where $f \in \Gamma_0(\mathcal{H})$ satisfies a relative smoothness condition, $g \circ T \in \Gamma_0(\mathcal{H})$ with g prox-friendly, $h \in \Gamma_0(\mathcal{H})$ has compact domain and admits an accessible linear minimization oracle, and T and A are bounded linear operators. Inexactness pertains to errors in the computation of several important quantities. In particular, we propose three different inexact and stochastic methods of computing the gradient ∇f in a practically implementable way. We detail how to show that the practical sampling routines induced by these methods are compatible with ICGALP by analyzing the summability of the induced errors. The first is a stochastic method of approximating the gradient ∇f using a necessarily increasing batch size for risk minimization problems. The second is a variance reduction method which allows us to take as little as a single sample at each iteration, greatly reducing the computational load and storage requirements. Finally, we propose a deterministic sweeping method which samples in a predetermined way. All of these methods allow to approximate not only ∇f but also the gradient of $\|A \cdot -b\|^2$ in the augmented Lagrangian coming from the affine constraint, or when computing the proximal mapping of g. Our main contributions and findings can be summarized as follows:

Main contributions of this chapter

- Convergence guarantees and rates for both the feasibility $(\|Ax_k - b\|)_{k\in\mathbb{N}}$ and the optimality $(\mathcal{L}(x_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star))_{k\in\mathbb{N}}$ in a $\mathbb{P}$-almost sure sense.
- Convergence rates for the feasibility and optimality in terms of their expectation.
- $\mathbb{P}$-almost sure strong convergence of the sequence of dual variables $(\mu_k)_{k\in\mathbb{N}}$.
- A numerical verification of the theorems proposed for the variance reduction and sweeping methods. Our algorithm is the only conditional gradient algorithm, i.e., using the linear minimization oracle over $\mathcal{C}$, which can solve this problem with stochastic approximations of both ∇f and the augmented Lagrangian penalty term, in contrast to contemporary work like [77]. We compare the variance reduction and sweeping methods for various different batch sizes and step sizes, with plots validating our claimed convergence results.

The content of this chapter appeared in [109].

Contents

4.1 Introduction

4.1.1 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, 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 $b \in \operatorname{ran}(A)$; 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 2.1.13), that the function g has a proximal mapping which is accessible, and that the function h admits an accessible linearly-perturbed minimization oracle with $C \overset{\text{def}}{=} \operatorname{dom}(h)$ a weakly compact subset of $\mathcal{H}_p$.

In fact, the problem under consideration here is exactly the same as that of Chapter 3 (and by extension, [108]), however, in this chapter, we consider an inexact extension of the algorithm presented and analyzed in Chapter 3 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, $\operatorname{prox}_{\beta g}$, and the linear minimization oracle itself.

4.1.2 Algorithm

For each $k \in \mathbb{N}$, we denote by λ_k and λ_k^s random variables, i.e., measurable mappings 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 9: Inexact Conditional Gradient with Augmented Lagrangian and Proximal-step (IC-GALP)

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 (similarly to Chapter 3),

$$\begin{aligned}
\Phi(x) &\stackrel{\text{def}}{=} f(x) + g(Tx) + h(x);\\
\mathcal{L}(x,\mu) &\stackrel{\text{def}}{=} f(x) + g(Tx) + h(x) + \langle \mu, Ax - b\rangle;\\
\mathcal{L}_k(x,\mu) &\stackrel{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + h(x) + \langle \mu, Ax - b\rangle + \frac{\rho_k}{2}\|Ax - b\|^2;\\
\mathcal{E}_k(x,\mu) &\stackrel{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + \langle \mu, Ax - b\rangle + \frac{\rho_k}{2}\|Ax - b\|^2;\\
\phi_k(x) &\stackrel{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + h(x).
\end{aligned} \tag{4.1.1}$$

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

We define the 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 $x_0, \mu_0, \widehat{s}_0, \ldots, \widehat{s}_k$ (see Section 2.3). 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 $\operatorname{ran}(A)$.

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

$$\boldsymbol{S}_{\mathscr{P}} \stackrel{\text{def}}{=} \operatorname*{Argmin}_{x\in\mathcal{H}_p}\{f(x) + g(x) + h(x) : Ax = b\} \quad \text{and} \quad \boldsymbol{S}_{\mathscr{D}} \stackrel{\text{def}}{=} \operatorname*{Argmin}_{\mu\in\mathcal{H}_d}\{(f + g + h)^*(-A^*\mu) + \langle \mu, b\rangle\} \tag{4.1.2}$$

and the notation for $\mathbb{P}$-a.s. weak cluster points of a sequence of $\mathcal{H}_p$-valued random variables $(x_k)_{k\in\mathbb{N}}$ to be

$$\mathfrak{W}\left[(x_k)_{k\in\mathbb{N}}\right] \stackrel{\text{def}}{=} \left\{x \in \mathcal{H}_p : \exists\, (x_{k_j})_{j\in\mathbb{N}}, x_{k_j} \rightharpoonup x\right\}. \tag{4.1.3}$$

4.1.3 Assumptions

4.1.3.1 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:

($\mathbf{A}_1$) The functions $f, g \circ T$, and h belong to $\Gamma_0(\mathcal{H}_p)$.

($\mathbf{A}_2$) The pair $(f, \mathcal{C})$ is (F, ζ)-smooth (see Definition 2.1.13), where we recall $\mathcal{C} \overset{\text{def}}{=} \operatorname{dom}(h)$.

($\mathbf{A}_3$) The set $\mathcal{C}$ is weakly compact (and thus contained in a ball of radius $R > 0$).

($\mathbf{A}_4$) It holds $T\mathcal{C} \subset \operatorname{dom}(\partial g)$ and $\sup_{x \in \mathcal{C}} \left\| [\partial g(Tx)]^0 \right\| = M < +\infty$.

($\mathbf{A}_5$) The function h is Lipschitz continuous relative to its domain $\mathcal{C}$ with constant $L_h \geq 0$, i.e., $\forall (x, z) \in \mathcal{C}^2$, $|h(x) - h(z)| \leq L_h \|x - z\|$.

($\mathbf{A}_6$) There exists a saddle-point $(x^\star, \mu^\star) \in \mathcal{H}_p \times \mathcal{H}_d$ for the Lagrangian $\mathcal{L}$.

($\mathbf{A}_7$) The set $\operatorname{ran}(A)$ is closed.

($\mathbf{A}_8$) One of the following holds:

(I) $A^{-1}(b) \cap \operatorname{int}(\operatorname{dom}(g \circ T)) \cap \operatorname{int}(\mathcal{C}) \neq \emptyset$, where $A^{-1}(b)$ is the pre-image of b under A.

(II) $\mathcal{H}_p$ and $\mathcal{H}_d$ are finite-dimensional and

$$\begin{cases} A^{-1}(b) \cap \operatorname{ri}(\operatorname{dom}(g \circ T)) \cap \operatorname{ri}(\mathcal{C}) \neq \emptyset \\ \text{and} \\ \operatorname{ran}(A^*) \cap \operatorname{par}(\operatorname{dom}(g \circ T) \cap \mathcal{C})^\perp = \{0\}. \end{cases} \tag{4.1.4}$$

($\mathbf{A}_9$) The space $\mathcal{H}_d$ is separable.

($\mathbf{A}_{10}$) The set-valued mappings $(\partial(\phi_k^* \circ (-A^*)))_{k \in \mathbb{N}}$ satisfy the following property: for any sequence $((p_k, q_k))_{k \in \mathbb{N}}$ satisfying, for each $k \in \mathbb{N}$,

$$p_k \in \partial(\phi_k^* \circ (-A^*))(q_k),$$

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

We recall that the lemmas Lemma 2.1.17 and Lemma 2.1.18 outline sufficient conditions which ensure that assumption **($\mathbf{A}_4$)** holds for g and show why it's unnecessary to make a similar assumption for f in light of **($\mathbf{A}_1$)** and **($\mathbf{A}_2$)**.

Remark 4.1.1. If the dimension of $\mathcal{H}_d$ is finite, then **($\mathbf{A}_{10}$)** is satisfied because weakly compact sets are compact in such spaces. Alternatively, another sufficient condition is to impose that the sublevel sets of the functions $(\phi_k^* \circ (-A^*))_{k \in \mathbb{N}}$ are compact, for instance if the functions are uniformly convex, unfiormly in k.

4.1.3.2 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:

($\mathbf{P}_1$) $(\gamma_k)_{k \in \mathbb{N}} \subset]0, 1]$ 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.

($\mathbf{P}_2$) $(\gamma_k)_{k \in \mathbb{N}} \notin \ell^1$.

($\mathbf{P}_3$) $(\beta_k)_{k \in \mathbb{N}} \in \ell_+$ is nonincreasing and converges to 0.

($\mathbf{P}_4$) $(\rho_k)_{k \in \mathbb{N}} \in \ell_+$ is nondecreasing with $0 < \underline{\rho} \leq \rho_k \leq \overline{\rho} < +\infty$.

($\mathbf{P}_5$) For some positive constants $\underline{M}$ and $\overline{M}$, $\underline{M} \leq (\gamma_k / \gamma_{k+1}) \leq \overline{M}$.

($\mathbf{P}_6$) $(\theta_k)_{k \in \mathbb{N}}$ satisfies $\theta_k = \frac{\gamma_k}{c}$ for some $c > 0$ such that $\frac{\overline{M}}{c} - \frac{\rho}{2} < 0$.

($\mathbf{P}_7$) $(\gamma_k)_{k\in\mathbb{N}}$ and $(\rho_k)_{k\in\mathbb{N}}$ satisfy $\rho_{k+1} - \rho_k - \gamma_{k+1}\rho_{k+1} + \frac{2}{c}\gamma_k - \frac{\gamma_k^2}{c} \leq 0$ for c in ($\mathbf{P}_6$).

($\mathbf{P}_8$) $(\gamma_{k+1}\mathbb{E}[\|\lambda_{k+1}\| \mid \mathcal{S}_k])_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S})$ and $\left(\gamma_{k+1}\mathbb{E}\left[\lambda_{k+1}^s \mid \mathcal{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S})$.

($\mathbf{P}_9$) $(\gamma_{k+1}\mathbb{E}[\|\lambda_{k+1}\|])_{k\in\mathbb{N}} \in \ell_+^1$ and $\left(\gamma_{k+1}\mathbb{E}\left[\lambda_{k+1}^s\right]\right)_{k\in\mathbb{N}} \in \ell_+^1$.

Remark 4.1.2. To satisfy ($\mathbf{P}_7$), it suffices to take $(\rho_k)_{k\in\mathbb{N}}$ to be a constant sequence, i.e. $\rho_k \equiv \rho$, with ρ sufficiently large to satisfy $2\frac{\overline{M}}{c} < \rho$, a similar requirement as in ($\mathbf{P}_6$). The condition ($\mathbf{P}_7$) would then be satisfied as follows,

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

Remark 4.1.3. We will also denote the gradient of $\mathcal{E}_k$ with errors as

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

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 ($\mathbf{P}_8$) is sufficiently satisfied by demanding that

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

4.1.4 Organization of the Chapter

The remainder of the chapter is divided into four sections. In Section 4.2, the main estimations, e.g. for the feasibility and Lagrangian convergence, are established and then used in the proceeding Section 4.3. The analysis and results extend those of Chapter 3 to the inexact and stochastic setting. In Section 4.4, we consider different problem instances where inexact deterministic or stochastic computations are involved. Finally, numerical results are reported in Section 4.5 to support our theoretical findings.

4.2 Preliminary Estimations

4.2.1 Preparatory Results

The following two lemmas come directly from Chapter 3, we include their statement for convenience.

Lemma 4.2.1. *Suppose ($\mathbf{A}_1$), ($\mathbf{A}_2$) and ($\mathbf{P}_1$) hold. For each $k \in \mathbb{N}$, define the quantity*

$$L_k \overset{\text{def}}{=} \frac{\|T\|^2}{\beta_k} + \|A\|^2\rho_k. \tag{4.2.1}$$

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

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

Proof. See Lemma 3.2.1 □

Lemma 4.2.2. *Suppose* ($\mathbf{A}_1$) *and* ($\mathbf{A}_2$) *hold. Then, for each* $k \in \mathbb{N}$ *and for every* $x \in \mathcal{H}_p$,

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

Proof. See Lemma 3.2.2. □

Lemma 4.2.3. *Assume that* ($\mathbf{A}_3$) *and* ($\mathbf{P}_4$) *hold. Let* $(x_k)_{k\in\mathbb{N}}$ *be the sequence of primal iterates generated by* Algorithm 9 *with* $\mathcal{S}_k \overset{\text{def}}{=} \sigma(x_0, \mu_0, \widehat{s}_0, \ldots, \widehat{s}_k)$ *as before. Then, for each* $k \in \mathbb{N}$, *we have the following estimate,*

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

Proof. For each $k \in \mathbb{N}$, by convexity of the function $\frac{\rho_{k+1}}{2} \|A \cdot -b\|^2$ and the assumption ($\mathbf{P}_4$) that $(\rho_k)_{k\in\mathbb{N}}$ is nondecreasing, we have,

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

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

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

where we have used the Cauchy-Schwarz inequality and the boundedness of $\mathcal{C}$, assumed in ($\mathbf{A}_3$), in the last inequality. □

Remark 4.2.4. 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 $k \in \mathbb{N}$,

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

Lemma 4.2.5. *Suppose that* ($\mathbf{A}_1$)-($\mathbf{A}_6$) *hold. Let* $(x_k)_{k\in\mathbb{N}}$ *be the sequence of primal iterates generated by* Algorithm 9 *and* $\mu^\star$ *a solution, which exists by* ($\mathbf{A}_6$), *of the dual problem, and recall the constant* D *from* Lemma 2.1.18. *Then, for each* $k \in \mathbb{N}$, *we have the following estimate,*

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

Proof. We recall the proof from Lemma 3.2.3 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 ($\mathbf{A}_4$) and the fact that for all $k \in \mathbb{N}$, $x_k \in \mathcal{C}$, we have $\|u_k\| \leq M$. By ($\mathbf{A}_1$), 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 (2.1.4) on g and f, the L_h-Lipschitz continuity of h relative to $\mathcal{C}$ (see ($\mathbf{A}_5$)), and the Cauchy-Schwarz inequality on the inner product. Since, for each $k \in \mathbb{N}$, $x_{k+1} = x_k + \gamma_k (\widehat{s}_k - x_k)$, we obtain, for each $k \in \mathbb{N}$,

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

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

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

where we have used the Cauchy-Schwarz inequality, the boundedness of the set $\mathcal{C}$ by $(\mathbf{A}_3)$, the boundedness of u_k by M by $(\mathbf{A}_4)$, and the boundedness of $\|\nabla f(x)\|$ by D, the constant in Lemma 2.1.18. □

4.2.2 Feasibility Estimation

Lemma 4.2.6 (Feasibility estimate). *Suppose that $(\mathbf{A}_1)$ - $(\mathbf{A}_4)$ and $(\mathbf{A}_6)$ all hold. Consider the sequence of iterates $(x_k)_{k\in\mathbb{N}}$ generated by Algorithm 9 with parameters satisfying $(\mathbf{P}_1)$ and $(\mathbf{P}_3)$-$(\mathbf{P}_6)$. 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 $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\Delta_{k+1} \mid \mathcal{S}_k\right] - \Delta_k \leq -\gamma_{k+1}\left(\frac{M}{c}\|A\tilde{x}_{k+1} - b\|^2 + \delta\|A(x_{k+1} - \tilde{x}_{k+1})\|^2\right) + \gamma_{k+1}^2 \frac{L_{k+1}}{2} d_{\mathcal{C}}^2 \\ + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_{k+1}) + \frac{\beta_k - \beta_{k+1}}{2} M^2 + (\rho_{k+1} - \rho_k)\left(\|A\|^2 R^2 + \|b\|^2\right) \\ + \gamma_{k+1}\mathbb{E}\left[\lambda_{k+1}^s \mid \mathcal{S}_k\right] + d_{\mathcal{C}}\gamma_{k+1}\mathbb{E}\left[\|\lambda_{k+1}\| \mid \mathcal{S}_k\right].\end{aligned} \tag{4.2.2}$$

Additionally, if the assumptions are strengthened to include $(\mathbf{P}_1)$-$(\mathbf{P}_6)$ and $(\mathbf{P}_8)$, then it holds

$$\left(\gamma_{k+1}\|A\tilde{x}_{k+1} - b\|^2\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S}) \quad \text{and} \quad \left(\gamma_{k+1}\|Ax_{k+1} - b\|^2\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S}), \tag{4.2.3}$$

where $\mathfrak{S} \overset{\text{def}}{=} (\mathcal{S}_k)_{k\in\mathbb{N}}$.

Proof. The proof here is adapted from the analogous result found in Lemma 3.2.5. 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 Lemma 3.2.5). We denote a minimizer of $\mathcal{L}_k(x, \mu_k)$ by $\tilde{x}_k \in \operatorname{Argmin}_{x\in\mathcal{H}_p} \mathcal{L}_k(x, \mu_k)$, which exists and belongs to $\mathcal{C}$ by $(\mathbf{A}_1)$-$(\mathbf{A}_3)$. We have, for each $k \in \mathbb{N}$,

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

Recall that $\tilde{x}_k \in \operatorname{Argmin}_{x\in\mathcal{H}_p} \mathcal{L}_k(x, \mu_k)$, that $g^{\beta_k} \leq g^{\beta_{k+1}}$ due to $(\mathbf{P}_3)$ and Proposition 2.1.2(v), and that $\rho_k \leq \rho_{k+1}$ by $(\mathbf{P}_4)$. 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 9. 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 $k \in \mathbb{N}$,

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

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

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

We denote by $\mathbf{T1} \overset{\text{def}}{=} \mathcal{L}_{k+1}\left(x_{k+2}, \mu_{k+1}\right) - \mathcal{L}_{k+1}\left(x_{k+1}, \mu_{k+1}\right)$ and the remaining part of the right-hand side by **T2**. For the moment, we focus our attention on **T1**. Recall that $\mathcal{L}_k\left(x, \mu_k\right) = \mathcal{E}_k\left(x, \mu_k\right) + h\left(x\right)$ and apply Lemma 4.2.1 between points x_{k+2} and x_{k+1}, to get, for each $k \in \mathbb{N}$,

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

By $(\mathbf{A}_1)$ 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 $k \in \mathbb{N}$,

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

We can apply the definition of s_{k+1} as the minimizer of the linear minimization oracle and Lemma 4.2.2 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. Combining **T1** and **T2** and using the Pythagoras identity we have, for each $k \in \mathbb{N}$,

$$
\begin{aligned}
\Delta_{k+1} - \Delta_k \leq\ & -\theta_k\left\|A\tilde{x}_{k+1} - b\right\|^2 + \left(\theta_k - \gamma_{k+1}\frac{\rho_{k+1}}{2}\right)\left\|A\left(x_{k+1} - \tilde{x}_{k+1}\right)\right\|^2 \\
& + \frac{L_{k+1}}{2}\left\|x_{k+2} - x_{k+1}\right\|^2 + D_F\left(x_{k+2}, x_{k+1}\right) + \left[g^{\beta_{k+1}} - g^{\beta_k}\right]\left(Tx_{k+1}\right) \\
& + \frac{\rho_{k+1} - \rho_k}{2}\left\|Ax_{k+1} - b\right\|^2 + \gamma_{k+1}\Big(\lambda^s_{k+1} + \left\langle \lambda_{k+1}, \tilde{x}_{k+1} - \widehat{s}_{k+1}\right\rangle\Big).
\end{aligned} \tag{4.2.4}
$$

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

- Under $(\mathbf{P}_5)$ and $(\mathbf{P}_6)$, 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 $(\mathbf{P}_6)$, 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 $(\mathbf{P}_5)$). Then, using again $(\mathbf{P}_5)$ and the above inequality, for each $k \in \mathbb{N}$,
$$\theta_k - \gamma_{k+1}\frac{\rho_{k+1}}{2} \leq \left(\frac{\overline{M}}{c} - \frac{\rho_{k+1}}{2}\right)\gamma_{k+1} \leq \left(\frac{\overline{M}}{c} - \frac{\underline{\rho}}{2}\right)\gamma_{k+1} = -\delta\gamma_{k+1}. \tag{4.2.5}$$
- By Algorithm 9, 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 $(\mathbf{A}_3)$, for each $k \in \mathbb{N}$,
$$\frac{L_{k+1}}{2}\mathbb{E}\left[\left\|x_{k+2} - x_{k+1}\right\|^2 \mid \mathcal{S}_k\right] = \frac{L_{k+1}}{2}\gamma^2_{k+1}\mathbb{E}\left[\left\|\widehat{s}_{k+1} - x_{k+1}\right\|^2 \mid \mathcal{S}_k\right] \leq \frac{L_{k+1}}{2}\gamma^2_{k+1}d^2_{\mathcal{C}}.$$
- Recall that, by $(\mathbf{A}_2)$, f is (F, ζ)-smooth and invoke Remark 2.1.14, to get
$$\mathbb{E}\left[D_F\left(x_{k+2}, x_{k+1}\right) \mid \mathcal{S}_k\right] \leq K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_{k+1}\right).$$
- By Proposition 2.1.2(v) and assumption $(\mathbf{A}_4)$,
$$\mathbb{E}\left[\left[g^{\beta_{k+1}} - g^{\beta_k}\right]\left(Tx_{k+1}\right) \mid \mathcal{S}_k\right] \leq \frac{\beta_k - \beta_{k+1}}{2}\mathbb{E}\left[\left\|\left[\partial g\left(Tx_{k+1}\right)\right]^0\right\|^2 \mid \mathcal{S}_k\right] \leq \frac{\beta_k - \beta_{k+1}}{2}M^2.$$

- We also have, using Jensen's inequality and $(\mathbf{A}_3)$, for each $k \in \mathbb{N}$,

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

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

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

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 $(\mathbf{A}_3)$, we also have, for each $k \in \mathbb{N}$,

$$\gamma_{k+1}\mathbb{E}\left[\langle \lambda_{k+1}, \tilde{x}_{k+1}-\widehat{s}_{k+1}\rangle \mid \mathcal{S}_k\right] \leq \gamma_{k+1} d_{\mathcal{C}}\mathbb{E}\left[\|\lambda_{k+1}\| \mid \mathcal{S}_k\right], \tag{4.2.6}$$

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

$$\begin{aligned}\mathbb{E}\left[\Delta_{k+1} \mid \mathcal{S}_k\right] - \Delta_k \leq & -\frac{M}{c}\gamma_{k+1}\|A\tilde{x}_{k+1}-b\|^2 - \delta\gamma_{k+1}\|A\left(x_{k+1}-\tilde{x}_{k+1}\right)\|^2 + \gamma_{k+1}^2\frac{L_{k+1}}{2}d_{\mathcal{C}}^2 \\ & + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_{k+1}\right) + \frac{\beta_k-\beta_{k+1}}{2}M^2 + (\rho_{k+1}-\rho_k)\left(\|A\|^2 R^2 + \|b\|^2\right) \\ & + \gamma_{k+1}\mathbb{E}\left[\lambda_{k+1}^s \mid \mathcal{S}_k\right] + \gamma_{k+1} d_{\mathcal{C}}\mathbb{E}\left[\|\lambda_{k+1}\| \mid \mathcal{S}_k\right],\end{aligned} \tag{4.2.7}$$

and (4.2.2) follows by rearranging terms, giving, for each $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\Delta_{k+1} \mid \mathcal{S}_k\right] - \Delta_k \leq & -\gamma_{k+1}\left(\frac{M}{c}\|A\tilde{x}_{k+1}-b\|^2 + \delta\|A\left(x_{k+1}-\tilde{x}_{k+1}\right)\|^2\right) + \gamma_{k+1}^2\frac{L_{k+1}}{2}d_{\mathcal{C}}^2 \\ & + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_{k+1}\right) + \frac{\beta_k-\beta_{k+1}}{2}M^2 + (\rho_{k+1}-\rho_k)\left(\|A\|^2 R^2 + \|b\|^2\right) \\ & + \gamma_{k+1}\mathbb{E}\left[\lambda_{k+1}^s \mid \mathcal{S}_k\right] + d_{\mathcal{C}}\gamma_{k+1}\mathbb{E}\left[\|\lambda_{k+1}\| \mid \mathcal{S}_k\right].\end{aligned} \tag{4.2.8}$$

Because of $(\mathbf{P}_1)$ and $(\mathbf{P}_4)$, and in view of the definition of L_{k+1} in (4.2.1), 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 (4.2.8) we have

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

where R is the constant arising from $(\mathbf{A}_3)$. Under $(\mathbf{P}_1)$ 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 $(\mathbf{P}_8)$, we also have

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

Using the notation of Lemma 2.3.2, we set, for each $k \in \mathbb{N}$,

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

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

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

where $(z_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathfrak{S})$, and r_k is bounded from below. We then deduce using Lemma 2.3.2 that $(r_k)_{k\in\mathbb{N}}$ is convergent ($\mathbb{P}$-a.s.) and

$$\left(\gamma_{k+1}\left\|A\tilde{x}_{k+1}-b\right\|^2\right)_{k\in\mathbb{N}} \in \ell^1_+(\mathfrak{S}) \quad \text{and} \quad \left(\gamma_{k+1}\left\|A\left(x_{k+1}-\tilde{x}_{k+1}\right)\right\|^2\right)_{k\in\mathbb{N}} \in \ell^1_+(\mathfrak{S}) \tag{4.2.9}$$

satisfying 4.2.3. Consequently,

$$\left(\gamma_k\left\|Ax_k-b\right\|^2\right)_{k\in\mathbb{N}} \in \ell^1_+(\mathfrak{S}), \tag{4.2.10}$$

since by the Cauchy-Schwarz inequality,

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

□

4.2.3 ($\mathbb{P}$-a.s.) Boundedness of $(\mu_k)_{k\in\mathbb{N}}$

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 Section 3.2.3 and trivially extend to the stochastic case in light of the previous section.

Lemma 4.2.7. *Suppose that* ($\mathbf{A}_1$)-($\mathbf{A}_3$), ($\mathbf{A}_6$)-($\mathbf{A}_8$), *and* ($\mathbf{P}_1$)-($\mathbf{P}_6$) *all hold and define, for each* $k \in \mathbb{N}$,

$$\varphi_k(\mu) \overset{\text{def}}{=} -\inf_{x\in\mathcal{H}_p}\mathcal{L}_k(x,\mu) \quad \text{and} \quad \bar{\varphi} \overset{\text{def}}{=} f(x)+g(Tx)+h(x)+\frac{\overline{\rho}}{2}\left\|Ax-b\right\|^2. \tag{4.2.11}$$

Then the sequence of dual iterates $(\mu_k)_{k\in\mathbb{N}}$ *generated by* Algorithm 9 *is bounded* ($\mathbb{P}$-a.s.), *for each* $k \in \mathbb{N}$ *the function* $\varphi_k(\mu)$ *is convex and differentiable with gradient*

$$\nabla\varphi_k(\mu) = \rho_k^{-1}\left(\mu - \operatorname{prox}_{\rho_k\Phi_k^*\circ(-A^*)}(\mu-\rho_k b)\right), \tag{4.2.12}$$

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

$$\nabla\varphi_k(\mu_k) = A\tilde{x}_k - b. \tag{4.2.13}$$

Proof. Note that here we have denoted, for each $k \in \mathbb{N}$, $\phi_k(x) = f(x) + g^{\beta_k}(x) + h(x)$ as in Chapter 3. For brevity, we defer to the proof of Lemma 3.2.7 in Section 3.2.3, noting that since $\left(\gamma_{k+1}\left\|\tilde{x}_{k+1}-b\right\|^2\right)_{k\in\mathbb{N}} \in \ell^1_+(\mathfrak{S})$ and $\left(\gamma_{k+1}\left\|x_{k+1}-b\right\|^2\right)_{k\in\mathbb{N}} \in \ell^1_+(\mathfrak{S})$, there exists $\tilde{\Omega} \subset \mathcal{F}$ with $\mathbb{P}\left(\tilde{\Omega}\right) = 1$ such that $(\varphi_k(\mu_k(\omega)))_{k\in\mathbb{N}}$ is convergent and thus bounded, and the uniform coercivity of $(\varphi_k)_{k\in\mathbb{N}}$ is unaffected by the inexactness, i.e., $(\mu_k(\omega))_{k\in\mathbb{N}}$ is bounded. □

Lemma 4.2.8. *Under* ($\mathbf{A}_1$)-($\mathbf{A}_8$) *and* ($\mathbf{P}_1$)-($\mathbf{P}_6$), *the composite function* $f + g\circ T + h$ *is uniformly bounded on* $\mathcal{C}$ *and we have*

$$\tilde{M} \overset{\text{def}}{=} \sup_{x\in\mathcal{C}}\left|f(x)+g(Tx)+h(x)\right| + \sup_{k\in\mathbb{N}}\left\|\mu_k\right\|\left(\left\|A\right\|R+b\right) < +\infty \ (\mathbb{P}\text{-a.s.}), \tag{4.2.14}$$

where R *is the radius from* ($\mathbf{A}_3$).

Proof. The proof follows in a ($\mathbb{P}$-a.s.) sense from Lemma 3.2.8 with the addition of the previous section. □

4.2.4 Optimality Estimation

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

Lemma 4.2.9 (Optimality estimate). *Recall the constants c, L_k, M, D, and L_h from* $(\mathbf{P}_6)$, Lemma 4.2.1, $(\mathbf{A}_4)$, Lemma 4.2.3, *and* $(\mathbf{A}_5)$, *respectively. Define, for each $k \in \mathbb{N}$,*

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

and

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

Then, under $(\mathbf{A}_1)$-$(\mathbf{A}_8)$ *and* $(\mathbf{P}_1)$-$(\mathbf{P}_8)$ *with* $\underline{M} \geq 1$, *for the sequences $(x_k)_{k\in\mathbb{N}}$ and $(\mu_k)_{k\in\mathbb{N}}$ generated by Algorithm 9, using the filtration $\mathfrak{S}' = (\mathcal{S}_{k-1})_{k\in\mathbb{N}}$, the following inequality holds, for each $k \in \mathbb{N}$ with $k > 0$,*

$$\begin{aligned}
\mathbb{E}\left[r_{k+1} \mid \mathcal{S}_{k-1}\right] - r_k \leq & -\gamma_k\left(\mathcal{L}\left(x_k, \mu^\star\right) - \mathcal{L}\left(x^\star, \mu^\star\right) + \frac{\rho_k}{2}\left\|Ax_k - b\right\|^2\right) + \frac{\gamma_{k+1}}{2}\mathbb{E}\left[\left\|Ax_{k+1} - b\right\|^2 \mid \mathcal{S}_{k-1}\right] \\
& + \left(\beta_k - \beta_{k+1}\right)\frac{M^2}{2} + \left(\gamma_k - \gamma_{k+1}\right)\tilde{M} + \gamma_k\beta_k\frac{M^2}{2} + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right) + \gamma_k^2 C_k \\
& + d_{\mathcal{C}}\gamma_k\mathbb{E}\left[\|\lambda_k\| \mid \mathcal{S}_{k-1}\right] + \gamma_k\mathbb{E}\left[\lambda_k^s \mid \mathcal{S}_{k-1}\right] \ (\mathbb{P}\text{-}a.s.).
\end{aligned} \tag{4.2.15}$$

Proof. Applying Lemma 4.2.2 to the points $x^\star$ and x_k we have, for each $k \in \mathbb{N}$,

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

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

$$\begin{aligned}
\mathcal{E}_k\left(x^\star, \mu_k\right) &\geq \mathcal{E}_k\left(x_k, \mu_k\right) + \left\langle \widehat{\nabla_x \mathcal{E}_k}\left(x_k, \mu_k\right), s_k - x_k\right\rangle + h\left(s_k\right) - h\left(x^\star\right) + \left\langle \lambda_k, x_k - x^\star\right\rangle \\
&\quad + \frac{\rho_k}{2}\left\|A(x^\star - x_k)\right\|^2 \\
&\geq \mathcal{E}_k\left(x_k, \mu_k\right) + \left\langle \widehat{\nabla_x \mathcal{E}_k}\left(x_k, \mu_k\right), \widehat{s}_k - x_k\right\rangle + h\left(\widehat{s}_k\right) - \lambda_k^s - h\left(x^\star\right) + \left\langle \lambda_k, x_k - x^\star\right\rangle \\
&\quad + \frac{\rho_k}{2}\left\|A(x^\star - x_k)\right\|^2.
\end{aligned} \tag{4.2.16}$$

From Lemma 4.2.1 applied to the points x_{k+1} and x_k and by definition of $x_{k+1} \overset{\text{def}}{=} x_k + \gamma_k\left(\widehat{s}_k - x_k\right)$ in Algorithm 9, we also have, for each $k \in \mathbb{N}$,

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

We combine the latter with (4.2.16), to get, for each $k \in \mathbb{N}$,

$$\begin{aligned}
\mathcal{E}_k\left(x_{k+1}, \mu_k\right) \leq & \ \mathcal{E}_k\left(x_k, \mu_k\right) + \gamma_k\left\langle \lambda_k, x^\star - \widehat{s}_k\right\rangle + D_F\left(x_{k+1}, x_k\right) + \gamma_k^2\frac{L_k}{2}\left\|\widehat{s}_k - x_k\right\|^2 \\
& + \gamma_k\left(\mathcal{E}_k\left(x^\star, \mu_k\right) + h(x^\star) - \mathcal{E}_k\left(x_k, \mu_k\right) - h(\widehat{s}_k) - \frac{\rho_k}{2}\left\|Ax_k - b\right\|^2 + \lambda_k^s\right).
\end{aligned} \tag{4.2.17}$$

By convexity of h from ($\mathbf{A}_1$) and the definition of x_{k+1}, we have, for each $k \in \mathbb{N}$,

$$\begin{aligned}\mathcal{L}_k\left(x_{k+1}, \mu_k\right) - \mathcal{L}_k\left(x_k, \mu_k\right) &= \mathcal{E}_k\left(x_{k+1}, \mu_k\right) - \mathcal{E}_k\left(x_k, \mu_k\right) + h\left(x_{k+1}\right) - h\left(x_k\right) \\ &\leq \mathcal{E}_k\left(x_{k+1}, \mu_k\right) - \mathcal{E}_k\left(x_k, \mu_k\right) + \gamma_k\left(h\left(\widehat{s}_k\right) - h\left(x_k\right)\right)\end{aligned} \tag{4.2.18}$$

Combining (4.2.17) and (4.2.18), we obtain, for each $k \in \mathbb{N}$,

$$\begin{aligned}\mathcal{L}_k\left(x_{k+1}, \mu_k\right) - \mathcal{L}_k\left(x_k, \mu_k\right) &\leq \gamma_k\left(\mathcal{E}_k\left(x^\star, \mu_k\right) + h(x^\star) - \mathcal{E}_k\left(x_k, \mu_k\right) - h\left(x_k\right)\right) + D_F\left(x_{k+1}, x_k\right) + \\ &\quad \gamma_k^2 \frac{L_k}{2}\left\|\widehat{s}_k - x_k\right\|^2 + \gamma_k\left(\left\langle\lambda_k, x^\star - \widehat{s}_k\right\rangle - \frac{\rho_k}{2}\left\|Ax_k - b\right\|^2 + \lambda_k^s\right) \\ &= \gamma_k\left(\mathcal{L}_k\left(x^\star, \mu_k\right) - \mathcal{L}_k\left(x_k, \mu_k\right)\right) + D_F\left(x_{k+1}, x_k\right) + \gamma_k^2 \frac{L_k}{2}\left\|\widehat{s}_k - x_k\right\|^2 \\ &\quad + \gamma_k\left(\left\langle\lambda_k, x^\star - \widehat{s}_k\right\rangle - \frac{\rho_k}{2}\left\|Ax_k - b\right\|^2 + \lambda_k^s\right)\end{aligned} \tag{4.2.19}$$

Recalling the definition of $\mu_{k+1} \overset{\text{def}}{=} \mu_k + A\left(x_{k+1} - b\right)$ in Algorithm 9, we have, for each $k \in \mathbb{N}$,

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

We combine the above and (4.2.19) to get, for each $k \in \mathbb{N}$,

$$\begin{aligned}\mathcal{L}_k\left(x_{k+1}, \mu_{k+1}\right) - \mathcal{L}_k\left(x_k, \mu_k\right) &\leq \theta_k\left\|Ax_{k+1} - b\right\|^2 + \gamma_k\left(\mathcal{L}_k\left(x^\star, \mu_k\right) - \mathcal{L}_k\left(x_k, \mu_k\right)\right) + D_F\left(x_{k+1}, x_k\right) \\ &\quad + \gamma_k^2 \frac{L_k}{2}\left\|\widehat{s}_k - x_k\right\|^2 + \gamma_k\left(\left\langle\lambda_k, x^\star - \widehat{s}_k\right\rangle - \frac{\rho_k}{2}\left\|Ax_k - b\right\|^2 + \lambda_k^s\right).\end{aligned} \tag{4.2.20}$$

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

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

Then, using Lemma 2.1.1, we get, for each $k \in \mathbb{N}$,

$$\begin{aligned}0 &\geq \theta_k\left(\mathcal{L}_k\left(x_{k+1}, \mu^\star\right) - \mathcal{L}_k\left(x_{k+1}, \mu_{k+1}\right)\right) + \frac{1}{2}\left(\left\|\mu_{k+1} - \mu^\star\right\|^2 - \left\|\mu_k - \mu^\star\right\|^2 + \left\|\mu_{k+1} - \mu_k\right\|^2\right) \\ &= \theta_k\left(\mathcal{L}_k\left(x_{k+1}, \mu^\star\right) - \mathcal{L}_k\left(x_{k+1}, \mu_{k+1}\right)\right) + \frac{1}{2}\left(\left\|\mu_{k+1} - \mu^\star\right\|^2 - \left\|\mu_k - \mu^\star\right\|^2 + \theta_k^2\left\|Ax_{k+1} - b\right\|^2\right).\end{aligned} \tag{4.2.21}$$

Recall that, by ($\mathbf{P}_6$), $\theta_k = \gamma_k / c$. Multiply (4.2.21) by c and sum with (4.2.20), 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 $k \in \mathbb{N}$,

$$\begin{aligned}
&(1-c\theta_{k+1})\mathcal{L}_{k+1}\left(x_{k+1},\mu_{k+1}\right)-(1-c\theta_k)\mathcal{L}_k\left(x_k,\mu_k\right)+\frac{c}{2}\left(\|\mu_{k+1}-\mu^\star\|^2-\|\mu_k-\mu^\star\|^2\right)\\
\leq\ &(1-c\theta_{k+1})\mathcal{L}_{k+1}\left(x_{k+1},\mu_{k+1}\right)-(1-c\theta_k)\mathcal{L}_k\left(x_{k+1},\mu_{k+1}\right)+\left(\theta_k-\frac{c\theta_k^2}{2}\right)\|Ax_{k+1}-b\|^2\\
&+\gamma_k\left(\mathcal{L}_k\left(x^\star,\mu_k\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\right)-c\theta_k\left(\mathcal{L}_k\left(x_{k+1},\mu^\star\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\right)-\frac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2\\
&+D_F\left(x_{k+1},x_k\right)+\gamma_k^2\frac{L_k}{2}\|\widehat{s}_k-x_k\|^2+\gamma_k\left(\langle\lambda_k,x^\star-\widehat{s}_k\rangle+\lambda_k^s\right)\\
=\ &c\left(\theta_k-\theta_{k+1}\right)\left(f+h+\langle\mu_{k+1},\ A\cdot-b\rangle\right)\left(x_{k+1}\right)+\left(\left(1-c\theta_{k+1}\right)g^{\beta_{k+1}}-\left(1-c\theta_k\right)g^{\beta_k}\right)\left(Tx_{k+1}\right)\\
&+\frac{1}{2}\left(\left(1-c\theta_{k+1}\right)\rho_{k+1}-\left(1-c\theta_k\right)\rho_k+2\theta_k-c\theta_k^2\right)\|Ax_{k+1}-b\|^2+\gamma_k\left(\mathcal{L}_k\left(x^\star,\mu_k\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\right)\\
&-c\theta_k\left(\mathcal{L}_k\left(x_{k+1},\mu^\star\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\right)-\frac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2+D_F\left(x_{k+1},x_k\right)+\gamma_k^2\frac{L_k}{2}\|\widehat{s}_k-x_k\|^2\\
&+\gamma_k\left(\langle\lambda_k,x^\star-\widehat{s}_k\rangle+\lambda_k^s\right).
\end{aligned}\tag{4.2.22}$$

By $(\mathbf{P}_5)$, $(\mathbf{P}_6)$ 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 $(\mathbf{P}_3)$, we also have $\beta_{k+1}\leq\beta_k$. In particular, $g^{\beta_k}\leq g^{\beta_{k+1}}\leq g$ pointwise. By Proposition 2.1.2(iv) and assumption $(\mathbf{A}_4)$, 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 $k\in\mathbb{N}$,

$$\begin{aligned}
&c\left(\theta_k-\theta_{k+1}\right)\left(f+h+\langle\mu_{k+1},\ A\cdot-b\rangle\right)\left(x_{k+1}\right)+\left(\left(1-c\theta_{k+1}\right)g^{\beta_{k+1}}-\left(1-c\theta_k\right)g^{\beta_k}\right)\left(Tx_{k+1}\right)\\
\leq\ &c\left(\theta_k-\theta_{k+1}\right)\mathcal{L}\left(x_{k+1},\mu_{k+1}\right)+\frac{1}{2}\left(\beta_k-\beta_{k+1}\right)\left\|\left(\partial g(Tx_{k+1})\right)^0\right\|^2.
\end{aligned}\tag{4.2.23}$$

Recall the definition of r_k in (4.2.9). Coming back to (4.2.22) and using (4.2.23), we obtain, for each $k\in\mathbb{N}$,

$$\begin{aligned}
r_{k+1}-r_k\ \leq\ &\frac{1}{2}\left(\left(1-\gamma_{k+1}\right)\rho_{k+1}-\left(1-\gamma_k\right)\rho_k+\frac{2}{c}\gamma_k-\frac{\gamma_k^2}{c}\right)\|Ax_{k+1}-b\|^2+\gamma_k\left(\mathcal{L}_k\left(x^\star,\mu_k\right)-\mathcal{L}_k\left(x_{k+1},\mu^\star\right)\right)\\
&-\frac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2+\frac{\beta_k-\beta_{k+1}}{2}\left\|\left(\partial g(Tx_{k+1})\right)^0\right\|^2+\left(\gamma_k-\gamma_{k+1}\right)\mathcal{L}\left(x_{k+1},\mu_{k+1}\right)\\
&+D_F\left(x_{k+1},x_k\right)+\gamma_k^2\frac{L_k}{2}\|\widehat{s}_k-x_k\|^2+\gamma_k\left(\langle\lambda_k,x^\star-\widehat{s}_k\rangle+\lambda_k^s\right).
\end{aligned}\tag{4.2.24}$$

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

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

where in the inequality we have used the fact that $g^{\beta_k} \leq g$ pointwise and that, by Proposition 2.1.2(v), for each $k \in \mathbb{N}$,

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

Substituting the above into (4.2.24) we have, for each $k \in \mathbb{N}$,

$$\begin{aligned}r_{k+1} - r_k \;\leq\; &\frac{1}{2}\left((1-\gamma_{k+1})\rho_{k+1} - \rho_k + \frac{2}{c}\gamma_k - \frac{\gamma_k^2}{c}\right)\|Ax_{k+1} - b\|^2 \\ &+ \gamma_k\left(\mathcal{L}(x^\star, \mu^\star) - \mathcal{L}(x_k, \mu^\star)\right) + \gamma_k\left(\mathcal{L}(x_k, \mu^\star) - \mathcal{L}(x_{k+1}, \mu^\star)\right) \\ &- \frac{\rho_k\gamma_k}{2}\|Ax_k - b\|^2 + \frac{\beta_k - \beta_{k+1}}{2}\left\|(\partial g(Tx_{k+1}))^0\right\|^2 + (\gamma_k - \gamma_{k+1})\mathcal{L}(x_{k+1}, \mu_{k+1}) \\ &+ \gamma_k\frac{\beta_k}{2}\left\|(\partial g(Tx_{k+1}))^0\right\|^2 + D_F(x_{k+1}, x_k) + \gamma_k^2\frac{L_k}{2}\|\widehat{s}_k - x_k\|^2 \\ &+ \gamma_k\left(\langle\lambda_k, x^\star - \widehat{s}_k\rangle + \lambda_k^s\right).\end{aligned} \tag{4.2.25}$$

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

- From $(\mathbf{P}_7)$, we have, for each $k \in \mathbb{N}$,
$$(1-\gamma_{k+1})\rho_{k+1} - \rho_k + \frac{2}{c}\gamma_k - \frac{\gamma_k^2}{c} \leq 0.$$
- By assumption $(\mathbf{A}_4)$, for each $k \in \mathbb{N}$,
$$\mathbb{E}\left[\left\|(\partial g(Tx_{k+1}))^0\right\|^2 \mid \mathcal{S}_{k-1}\right] \leq M^2.$$
- By Lemma 4.2.8, for each $k \in \mathbb{N}$,
$$\mathbb{E}\left[\mathcal{L}(x_{k+1}, \mu_{k+1}) \mid \mathcal{S}_{k-1}\right] \leq \tilde{M}.$$
- Recall that, by $(\mathbf{A}_2)$, f is (F, ζ)-smooth and invoke Remark 2.1.14, to get, for each $k \in \mathbb{N}$,
$$\mathbb{E}\left[D_F(x_{k+1}, x_k) \mid \mathcal{S}_{k-1}\right] \leq K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_k).$$
- Since, for each $k \in \mathbb{N}$, $\widehat{s}_k$ and x_k are both in $\mathcal{C}$, we have
$$\mathbb{E}\left[\|\widehat{s}_k - x_k\| \mid \mathcal{S}_{k-1}\right] \leq d_{\mathcal{C}}.$$

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

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

We can bound the inner product involving the error terms using the Cauchy-Schwarz inequality and the boundedness of $\mathcal{C}$. Applying Lemma 4.2.5 and regrouping terms with γ_k^2 we get, for each $k \in \mathbb{N}$,

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

We conclude by trivial manipulations. □

4.3 Convergence Analysis

As in Section 3.3 in the previous chapter, when rates of convergence are given they will be given in terms of the quantity $\Gamma_k \overset{\text{def}}{=} \sum_{i=0}^{k} \gamma_i$. The same example, Example 3.3.1, is relevant for its bounds on Γ_k, although the largest choice of b in this chapter will be more restrictive to satisfy the error summability conditions.

4.3.1 Asymptotic Feasibility

Theorem 4.3.1 (Feasibility). *Suppose that* $(\mathbf{A}_1)$*-*$(\mathbf{A}_4)$ *and* $(\mathbf{A}_6)$ *all hold and recall* $\Gamma_k \overset{\text{def}}{=} \sum_{i=0}^{k} \gamma_i$. *For a sequence* $(x_k)_{k\in\mathbb{N}}$ *generated by Algorithm 9 using parameters satisfying* $(\mathbf{P}_1)$ *-* $(\mathbf{P}_6)$ *and* $(\mathbf{P}_8)$ *we have,*

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

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

$$\|A\bar{x}_k - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right) \ (\mathbb{P}\text{-a.s.}). \tag{4.3.1}$$

Additionally, if $(\mathbf{P}_9)$ *also holds then we have the following pointwise rates in expectation,*

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

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

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

Proof. Our goal is to first apply Lemma 2.3.2 and then apply Lemma 2.3.3. To finish proving (i) we simply apply Lemma 4.2.3 (with Remark 4.2.4) and the conditions of Lemma 2.3.3 are satisfied. Then, (ii) follows directly from the application of Jensen's inequality as in the results of Theorem 3.3.2.

We now assume that $(\mathbf{P}_9)$ holds. By Lemma 4.2.6, we can take the total expectation and use the law of total expectation to have, for each $k \in \mathbb{N}$,

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

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

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

By the argument of the analogous claim for the conditional expectations in 4.2.6, in conjunction with $(\mathbf{P}_9)$ in place of $(\mathbf{P}_8)$, we have that $(\tilde{z}_k)_{k\in\mathbb{N}} \in \ell^1_+$. We can apply the total expectation to the results of both Lemma 4.2.3 and Lemma 4.2.5 and then the claims of interest follow from Lemma 2.2.3 applied with $(\tilde{r}_k)_{k\in\mathbb{N}}$, $(\tilde{p}_k)_{k\in\mathbb{N}}$, $(\tilde{w}_k)_{k\in\mathbb{N}}$, and $(\tilde{z}_k)_{k\in\mathbb{N}}$ defined as above. □

4.3.2 Optimality

We now proceed to prove the main theorem regarding optimality, recalling the notation of (4.1.2) for the terms $\boldsymbol{S}_{\mathscr{P}}$ and $\boldsymbol{S}_{\mathscr{D}}$ and (4.1.3) for $\mathfrak{W}\left[(x_k)_{k\in\mathbb{N}}\right]$.

Theorem 4.3.2 (Optimality). *Suppose that $(\mathbf{A}_1)$-$(\mathbf{A}_{10})$ and $(\mathbf{P}_1)$-$(\mathbf{P}_8)$ hold, with $\underline{M} \geq 1$. Let $(x_k)_{k\in\mathbb{N}}$ be the sequence of primal iterates generated by Algorithm 9 and $(x^\star, \mu^\star)$ a saddle-point pair for the Lagrangian. Then, in addition to the results of Theorem 4.3.1, the following holds*

(i) Convergence of the Lagrangian:

$$\lim_{k\to\infty} \mathcal{L}(x_k, \mu^\star) = \mathcal{L}(x^\star, \mu^\star) \ (\mathbb{P}\text{-a.s.}). \tag{4.3.2}$$

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

(iii) 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 $k \in \mathbb{N}$,

$$\mathcal{L}(\bar{x}_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) \in O\left(\frac{1}{\Gamma_k}\right) \ (\mathbb{P}\text{-a.s.}). \tag{4.3.3}$$

(iv) If the problem $(\mathscr{P})$ admits a unique solution $x^\star$, then $(x_k)_{k\in\mathbb{N}}$ converges weakly ($\mathbb{P}$-a.s.) to $x^\star$ a solution of $(\mathscr{P})$. 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 $k \in \mathbb{N}$,

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

Furthermore, if $(\mathbf{P}_9)$ holds then we have the following pointwise rates in expectation, for any $(x^\star, \mu^\star) \in \boldsymbol{S}_{\mathscr{P}} \times \boldsymbol{S}_{\mathscr{D}}$,

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

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

Proof. As in the proof of Theorem 4.3.1, our goal is to first apply Lemma 2.3.2 and then apply Lemma 2.3.3. By (4.2.15) in Lemma 4.2.9 we have, using the same notation, for each $k \in \mathbb{N}$,

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

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 2.3.2, we must show $(z_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathfrak{S}')$ where $\mathfrak{S}' \overset{\text{def}}{=} (\mathcal{S}_{k-1})_{k\in\mathbb{N}}$ as before. 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 $(\mathbf{P}_1)$. Recalling the definition of C_k, we have, for each $k \in \mathbb{N}$,

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

which is in ℓ^1_+ by $(\mathbf{P}_1)$ and $(\mathbf{P}_3)$. The remaining terms,

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

coming from the inexactness of the algorithm, are in $\ell^1_+(\mathfrak{S}')$ by $(\mathbf{P}_8)$. Thus, the r.h.s. belongs to $\ell^1_+(\mathfrak{S}')$ and so by Lemma 2.3.2 we have,

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

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

The first claim (i) follows by applying Lemma 2.3.3, the conditions of which are satisfied directly from Lemma 4.2.3 and Lemma 4.2.5.

The second claim, (ii), follows from the same arguments as in Theorem 3.3.3 but adapted to the stochastic case. For the claims about $(x_k)_{k\in\mathbb{N}}$, the proof is trivially extended to the stochastic setting ($\mathbb{P}$-a.s.). However, the claims about $(\mu_k)_{k\in\mathbb{N}}$ are more delicate to adapt so we write explicitly the arguments below.

By Theorem 4.3.14.2.3 we have $\left(\gamma_k\|A\tilde{x}_k - b\|^2\right)_{k\in\mathbb{N}} \in \ell^1_+(\mathfrak{S}')$ which, by Lemma 2.2.3 implies that there exists a subsequence $\left(A\tilde{x}_{k_j}\right)_{j\in\mathbb{N}}$ with $\left\|A\tilde{x}_{k_j} - b\right\| \to 0$ ($\mathbb{P}$-a.s.). Since the sequence $(\mu_k)_{k\in\mathbb{N}}$ is bounded ($\mathbb{P}$-a.s.) by Lemma 4.2.7, the subsequence $\left(\mu_{k_j}\right)_{j\in\mathbb{N}}$ is bounded ($\mathbb{P}$-a.s.) as well and thus admits a weakly ($\mathbb{P}$-a.s.) convergent subsequence $\left(\mu_{k_{j_i}}\right)_{i\in\mathbb{N}}$ with $\mu_{k_{j_i}} \rightharpoonup \bar{\mu}$ for some $\mathcal{H}_d$-valued random variable $\bar{\mu}$. By Fermat's rule ([10, Theorem 16.2]), the weak ($\mathbb{P}$-a.s.) sequential cluster point $\bar{\mu}$ is a solution to $(\mathscr{D})$ iff

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

The proximal operator is the resolvent of the subdifferential and so it follows that (4.2.12) is equivalent, for each $i \in \mathbb{N}$, to

$$\nabla\varphi_{k_{j_i}}\left(\mu_{k_{j_i}}\right) - b \in \partial\left(\phi^*_{k_{j_i}} \circ (-A^*)\right)\left(\mu_{k_{j_i}} - \rho_{k_{j_i}}\nabla\varphi_{k_{j_i}}\left(\mu_{k_{j_i}}\right)\right) \ (\mathbb{P}\text{-a.s.}). \tag{4.3.4}$$

Since $\left(A\tilde{x}_{k_j}\right)_{j\in\mathbb{N}}$ converges strongly to b ($\mathbb{P}$-a.s.), and in view of (4.2.13), it holds that $\nabla\varphi_{k_j}\left(\mu_{k_j}\right)$ converges strongly to 0 ($\mathbb{P}$-a.s.). However, $\mu_{k_{j_i}} - \rho_{k_{j_i}}\nabla\varphi_{k_{j_i}}\left(\mu_{k_{j_i}}\right)$ converges weakly to $\bar{\mu}$ ($\mathbb{P}$-a.s.). We henceforth argue that we can pass to the limit in (4.3.4) by sequential closedness using a Mosco convergence (weak-strong epigraphical convergence) argument (see [24] and [6, Definition 3.7]). Indeed, it was shown in the proof of Theorem 3.3.3 that $\Phi^*_{k_{j_i}} \circ -(A^*)$ Mosco-converges to $(\Phi)^* \circ (-A^*)$. This implies, via [6, Theorem 3.66], that $\partial\Phi^*_{k_{j_i}} \circ (-A^*)$ graph-converges to $\partial\Phi^* \circ (-A^*)$, and [6, Proposition 3.59] shows that $\left(\partial\Phi_{k_{j_i}} \circ (-A^*)\right)_{i\in\mathbb{N}}$ is sequentially closed for graph-convergence in the weak-strong topology on $\mathcal{H}_d$, i.e., for any sequence $\left(\left(v_{k_{j_i}}, \eta_{k_{j_i}}\right)\right)_{i\in\mathbb{N}}$ in the graph of $\partial\left(\Phi^*_{k_{j_i}} \circ (-A^*)\right)_{i\in\mathbb{N}}$ such that $v_{k_{j_i}}$ converges weakly to $\bar{v}$ and $\eta_{k_{j_i}}$ converges strongly to $\bar{\eta}$, we have $\bar{\eta} \in \partial\Phi^* \circ (-A^*)(\bar{v})$. Let, for each $i \in \mathbb{N}$, $v_{k_{j_i}} = \nabla\varphi_{k_{j_i}}\left(\mu_{k_{j_i}}\right) - b$ and $\eta_{k_{j_i}} = \mu_{k_{j_i}} - \rho_{k_{j_i}}\nabla\varphi_{k_{j_i}}\left(\mu_{k_{j_i}}\right)$, which converge strongly ($\mathbb{P}$-a.s.) and weakly ($\mathbb{P}$-a.s.) respectively, and let $\tilde{\Omega} \subset \mathcal{F}$ such that $\mathbb{P}\left(\tilde{\Omega}\right) = 1$ and, for all $\omega \in \tilde{\Omega}$, $v_{k_{j_i}}(\omega) \to b$ and $\eta_{k_{j_i}}(\omega) \rightharpoonup \bar{\mu}(\omega)$. We conclude that, for each $\omega \in \tilde{\Omega}$,

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

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

We now prove the existence of a set $\tilde{\Omega} \subset \mathcal{F}$ such that $\mathbb{P}\left(\tilde{\Omega}\right) = 1$ and, for all $\omega \in \tilde{\Omega}$, for any $\mu^\star \in \boldsymbol{S}_{\mathscr{D}}$,

$$\Theta\left(\mu^\star, \omega\right) \overset{\text{def}}{=} \lim_k \left\|\mu_k\left(\omega\right) - \mu^\star\right\|^2$$

exists. This does indeed hold ($\mathbb{P}$-a.s.) for each fixed $\mu^\star \in \boldsymbol{S}_{\mathscr{D}}$ by the argument in the proof of Theorem 3.3.3 but $\boldsymbol{S}_{\mathscr{D}}$ may be uncountable and so the entire statement for any $\mu^\star \in \boldsymbol{S}_{\mathscr{D}}$ may not necessarily hold ($\mathbb{P}$-a.s.) . To rectify this situation, we make the assumption $(\mathbf{A}_9)$ and argue as in [37, Proposition 2.3(iii)].

First repeat the argument made in the proof of Theorem 3.3.3 to show that, for each fixed $\mu^\star \in \boldsymbol{S}_{\mathscr{D}}$, there exists $\Omega_{\mu^\star} \subset \mathcal{F}$ with $\mathbb{P}\left(\Omega_{\mu^\star}\right) = 1$ such that, for any $\omega \in \Omega_{\mu^\star}$, $\Theta\left(\mu^\star, \omega\right)$ exists. Let $\mu^\star \in \boldsymbol{S}_{\mathscr{D}}$ and recall $\left(r_k\right)_{k\in\mathbb{N}}$ in (4.2.9), for each $k \in \mathbb{N}$,

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

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

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

We have from Theorem 4.3.1 that $\frac{\rho_k}{2}\left\|Ax_k - b\right\|^2 \to 0$ ($\mathbb{P}$-a.s.). Therefore,

$$\langle \mu^\star - \mu_k, Ax_k - b\rangle \to 0$$

since $\left(\mu_k\right)_{k\in\mathbb{N}}$ is bounded ($\mathbb{P}$-a.s.). We also have, by claim (i) of this theorem, that $\mathcal{L}\left(x_k \mu^\star\right) \to \mathcal{L}\left(x^\star, \mu^\star\right)$ ($\mathbb{P}$-a.s.). By Lemma 2.1.2 and $(\mathbf{A}_4)$, we get

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

which implies, in light of $(\mathbf{P}_3)$, that $g\left(Tx_k\right) - g^{\beta_k}\left(Tx_k\right) \to 0$ ($\mathbb{P}$-a.s.). Altogether, it holds that $\mathcal{L}_k\left(x_k, \mu_k\right) \to \mathcal{L}\left(x^\star, \mu^\star\right)$ ($\mathbb{P}$-a.s.) and thus the limit

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

exists ($\mathbb{P}$-a.s.) for each $\mu^\star \in \boldsymbol{S}_{\mathscr{D}}$.

Since $\mathcal{H}_d$ is separable by $(\mathbf{A}_9)$, there exists a countable set S such that $\bar{S} = \boldsymbol{S}_{\mathscr{D}}$. The previous paragraph has shown that, for every $\mu^\star \in \boldsymbol{S}_{\mathscr{D}}$, there exists $\Omega_{\mu^\star} \subset \mathcal{F}$ such that $\mathbb{P}\left(\Omega_{\mu^\star}\right) = 1$ and, for any $\omega \in \Omega_{\mu^\star}$, $\Theta\left(\mu^\star, \omega\right)$ exists. Set $\tilde{\Omega} = \bigcap_{\mu^\star \in S} \Omega_{\mu^\star}$ and let $\tilde{\Omega}^c$ be its set-theoretic complement. By the countability of S,

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

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

$$-\left\|\mu_n^\star - \mu^\star\right\| \leq \left\|\mu_k\left(\omega\right) - \mu^\star\right\| - \left\|\mu_k\left(\omega\right) - \mu_n^\star\right\| \leq \left\|\mu_n^\star - \mu^\star\right\|$$

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

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

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

We now aim to use $(\mathbf{A}_{10})$, for which we denote

$$\left(p_i\right)_{i\in\mathbb{N}} = \left(\nabla\varphi_{k_{j_i}}\left(\mu_{k_{j_i}}\right) - b\right)_{i\in\mathbb{N}} \quad \text{and} \quad \left(q_i\right)_{i\in\mathbb{N}} = \left(\mu_{k_{j_i}} - \rho_{k_{j_i}}\nabla\varphi_{k_{j_i}}\left(\mu_{k_{j_i}}\right)\right)_{i\in\mathbb{N}}.$$

We've shown that $\left(p_i\right)_{i\in\mathbb{N}}$ converges strongly to 0 ($\mathbb{P}$-a.s.) and that $\left(q_i\right)_{i\in\mathbb{N}}$ converges weakly to $\bar{\mu}$ ($\mathbb{P}$-a.s.) and so there exists $\tilde{\Omega} \subset \mathcal{F}$ with $\mathbb{P}\left(\tilde{\Omega}\right) = 1$ such that, for any $\omega \in \tilde{\Omega}$, $p_i\left(\omega\right) \to p\left(\omega\right)$ and $q_i\left(\omega\right) \rightharpoonup q\left(\omega\right)$. Due to (4.3.4), we furthermore have, for each $\omega \in \tilde{\Omega}$, for each $i \in \mathbb{N}$,

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

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

Meanwhile the third claim, (iii), follows from the argument of Theorem 3.3.3(3.3.5) directly applied to the ($\mathbb{P}$-a.s.) setting and similarly for (iv) following from the argument of the proof of Corollary 3.3.4.

Finally, assume that $(\mathbf{P}_9)$ holds. By taking the total expectation of (4.2.15) in Lemma 4.2.9 and using the law of total expectation we have, for each $k \in \mathbb{N}$,

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

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

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

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

$$\begin{aligned}
\tilde{z}_k = \frac{\gamma_{k+1}}{2}\mathbb{E}\left[\left\|Ax_{k+1} - b\right\|^2\right] + \left(\beta_k - \beta_{k+1}\right)\frac{M^2}{2} + \left(\gamma_k - \gamma_{k+1}\right)\tilde{M} \\
+ \gamma_k\beta_k\frac{M^2}{2} + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right) + \gamma_k^2 C_k + d_{\mathcal{C}}\gamma_k\mathbb{E}\left[\left\|\lambda_k\right\|\right] + \gamma_k\mathbb{E}\left[\lambda_k^s\right].
\end{aligned}$$

By repeating the arguments of the previous paragraph, we have that $\left(\tilde{z}_k\right)_{k\in\mathbb{N}} \in \ell_+^1$ (recall that $\left(\gamma_k\mathbb{E}\left[\left\|Ax_k - b\right\|^2\right]\right)_{k\in\mathbb{N}} \in \ell_+^1$ by Theorem 4.3.1). Invoking Lemma 2.2.3, again noting Lemma 4.2.3 and Lemma 4.2.5 hold with the total expectation as well, with $\left(\tilde{r}_k\right)_{k\in\mathbb{N}}$, $\left(\tilde{p}_k\right)_{k\in\mathbb{N}}$, $\left(\tilde{w}_k\right)_{k\in\mathbb{N}}$, and $\left(\tilde{z}_k\right)_{k\in\mathbb{N}}$ defined as above, we obtain the remaining claims. □

Remark 4.3.3. The assumption (A$_9$) is only necessary for showing that the sequence of dual variables $(\mu_k)_{k\in\mathbb{N}}$ admits an optimal weak cluster point. The other results, e.g., convergence of the Lagrangian values, the containment $\mathfrak{W}\left[(x_k)_{k\in\mathbb{N}}\right] \subset \boldsymbol{S}_{\mathscr{P}}$ ($\mathbb{P}$-a.s.), etc, do not require the separability imposed by (A$_9$). Likewise, something similar can be said for (A$_{10}$), which is only necessary for ensuring the strong convergence of the sequence of dual variables $(\mu_k)_{k\in\mathbb{N}}$ and can otherwise be omitted.

4.4 Applications

4.4.1 Stochastic Applications

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. Indeed, only (E.1) and (E.2) will be used for risk minimizaiton with increasing batch size while (E.3) and (E.4) will be needed for the results on risk minimization with variance reduction.

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

(E.2) For all η, the function $L(\cdot,\eta)$ is ω-smooth (see Definition 2.1.4) with ω nondecreasing.

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

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

Remark 4.4.1. In practical contexts, it's unrealistic that one will have access to the function f or knowledge of its regularity. To this end, we note that the assumptions (E.1) and (E.2), which depend only on the function $L(x,\eta)$, are sufficient to ensure that (E.3) holds and similarly for (E.4) if one adjusts (E.2) for Hölder-smoothness. Moreover, since Hölder-smoothness is a special case of ω-smoothness, (E.4) $\implies$ (E.3).

Remark 4.4.2. With the above choice for λ_k, the terms in $\nabla_x \mathcal{E}_k(x_k,\mu_k)$ coming from the augmented Lagrangian are computed exactly, however our analysis extends to the case where $\nabla_x\left(\frac{\rho_k}{2}\|Ax_k-b\|^2\right) = \rho_k A^*(Ax_k-b)$ is computed inexactly as well, as this function is always Lipschitz-continuous. We demonstrate this alternative choice in Section 4.5 by sampling the components $\rho_k A^*(Ax_k-b)^{(i)}$ in the numerical experiments.

For the sake of clarity, we demonstrate only the case where, for each $k\in\mathbb{N}$, $\lambda_k \equiv \lambda_k^f$ with $\lambda_k^f = \widehat{\nabla f}_k - \nabla f(x_k)$ and $\widehat{\nabla f}_k$ is our inexact computation of $\nabla f(x_k)$, to be defined in the following sections. As in the previous sections, all equalities/inequalities involving random variables should be understood to hold ($\mathbb{P}$-a.s.) even when it is not explicitly written for the sake of brevity.

4.4.1.1 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 4.4.3. *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 4.4.4. *Under (E.1) and (E.2), assume that the number of samples $n\left(k\right)$ at iteration k is lower bounded by $\left(\frac{\gamma_k}{\zeta(\gamma_k)}\right)^2$, i.e. for some $\alpha > 0$, $n\left(k\right) \geq \alpha\left(\frac{\gamma_k}{\zeta(\gamma_k)}\right)^2$. Then, the summability of the error in ($\mathbf{P}_8$) is satisfied; namely,*

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

Proof. By Lemma 4.4.3 we have, for each $k \in \mathbb{N}$,

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

The summability of $\zeta\left(\gamma_{k+1}\right)$ is given by $(\mathbf{P}_1)$ and thus $\gamma_{k=1}\mathbb{E}\left[\left\|\lambda_{k+1}^{f}\right\| \mid \mathcal{S}_k\right] \in \ell^1\left(\mathfrak{S}\right)$ □

Remark 4.4.5. The lower bound $n\left(k\right) \geq \alpha\left(\frac{\gamma_k}{\zeta(\gamma_k)}\right)^2$ is sufficient but not necessary; one can alternatively choose $n\left(k\right)$ 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 $(\mathbf{P}_1)$, the result will still hold.

4.4.1.2 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 section. For each $k \in \mathbb{N}$, let $\nu_k \in [0,1]$ and define

$$\widehat{\nabla f}_k \overset{\text{def}}{=} \left(1-\nu_k\right)\widehat{\nabla f}_{k-1} + \nu_k \nabla_x L\left(x_k, \eta_k\right) \tag{4.4.1}$$

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

Lemma 4.4.6. *Under (E.1) and (E.3), denote, for each $k \in \mathbb{N}$,*

$$\sigma_k^2 \overset{\text{def}}{=} \mathbb{E}\left[\left\|\nabla_x L\left(x_k, \eta_k\right) - \nabla f\left(x_k\right)\right\|^2 \mid \mathcal{S}_{k-1}\right] \tag{4.4.2}$$

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,

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

Proof. The proof of this theorem is inspired by a similar construction found in [88, Lemma 2]. By definition of λ_{k+1}^f and $\widehat{\nabla f}_{k+1}$, we have, for all $k \in \mathbb{N}$,

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

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

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

Applying the pythagoreas identity then gives,

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

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

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

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

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

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

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

We take the expectation on both sides, recalling the definition of σ_k (see (4.4.2)), 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 4.4.7. *Under (E.1) and (E.4), for each $k \in \mathbb{N}$, let $\widehat{\nabla f}_k$ be defined as in (4.4.1) 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,*

$$\left(\gamma_k^{1+\min\left\{\frac{\alpha}{2},\tau-\alpha\right\}}\right)_{k\in\mathbb{N}} \in \ell^1, \tag{4.4.3}$$

and, for k sufficiently large,

$$\frac{\gamma_k}{\gamma_{k+1}} \leq 1 + o\left(\gamma_k^\alpha\right), \tag{4.4.4}$$

then the summability condition in ($\mathbf{P}_8$) is satisfied; namely,

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

Proof. Since (E.4) $\implies$ (E.3), the assumptions (E.1) and (E.3) are satisfied and Lemma 4.4.6 gives, for all $k \in \mathbb{N}$,

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

By $(\mathbf{P}_5)$ 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 4.4.3, it holds that $\alpha < \min\{1, 2\tau - \alpha\}$, and the first condition of Lemma 2.2.5 is satisfied. Additionally, by (4.4.4), we have that the second condition, (2.2.3), of Lemma 2.2.5 is satisfied as well and we can apply Lemma 2.2.5 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 4.4.3 we have $\gamma_k^{1+\min\left\{\frac{\alpha}{2},\tau-\alpha\right\}} \in \ell^1$ and thus the summability condition of $(\mathbf{P}_8)$ is satisfied. □

Example 4.4.8. The condition (4.4.3) in Proposition 4.4.7 can be satisfied, for example, by taking $\gamma_k = \frac{1}{(k+1)^{1-b}}$. In this case, the condition (4.4.3) 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,

$$b < 1 - \left(1 + \min\left\{\frac{\alpha}{2}, \tau - \alpha\right\}\right)^{-1}. \tag{4.4.5}$$

The condition (4.4.4) in Proposition 4.4.7 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 4.3.1 and Theorem 4.3.2 under this choice of step size are,

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

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 $b < 1 - (4/3)^{-1} = \frac{1}{4}$

4.4.2 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 2.1.4) 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 4.4.1, 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 4.4.2). We will sweep, or cycle, through the functions f_i, taking the gradient of a single one at each iteration and recursively averaging with the past gradients. For notation, fixed n, we take $mod(k) \overset{\text{def}}{=} (k \mod n)$ 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 4.4.9. *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,*

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

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

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

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 $k\in\mathbb{N}$, $mod\,(k)\leq n$. Then,

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

□

Proposition 4.4.10. *Under (F.1) and (F.2), and assuming that* $\left(\gamma_k\omega\left(d_{\mathcal{C}}\gamma_k\right)\right)_{k\in\mathbb{N}}\in\ell^1$, *the summability condition of* $(\mathbf{P}_8)$ *holds; namely,*

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

Proof. By Lemma 4.4.9, we have, for all $k \geq 2n-1$,

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

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

4.5 Numerical Experiments

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

$$\min_{\substack{\|x\|_1\leq 1\\ Ax=0}} \frac{1}{2n}\|x-y\|^2, \tag{4.5.1}$$

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 4.4.2. Alternatively, we can let η be a random variable taking values in the set $\{1,\dots,n\}$ and write $L(x,\eta) = \frac{1}{2}(x_\eta - y_\eta)^2$ to cast the problem as risk minimization as in Section 4.4.1. 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 $(\mathbf{A}_1)$ to $(\mathbf{A}_8)$**(I)** 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 $(\mathbf{P}_1)$ to $(\mathbf{P}_7)$ are satisfied, as well as (F.2). In particular, to satisfy $(\mathbf{P}_8)$ 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 (4.4.3) in Proposition 4.4.7 is satisfied as was discussed in Example 4.4.8.

Since the problem (4.5.1) 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 4.1 and 4.2. 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 4.4.2), for two different step size and weight choices as well as different batch sizes (1, 64, or 256), displayed in Figures 4.1 and 4.2.

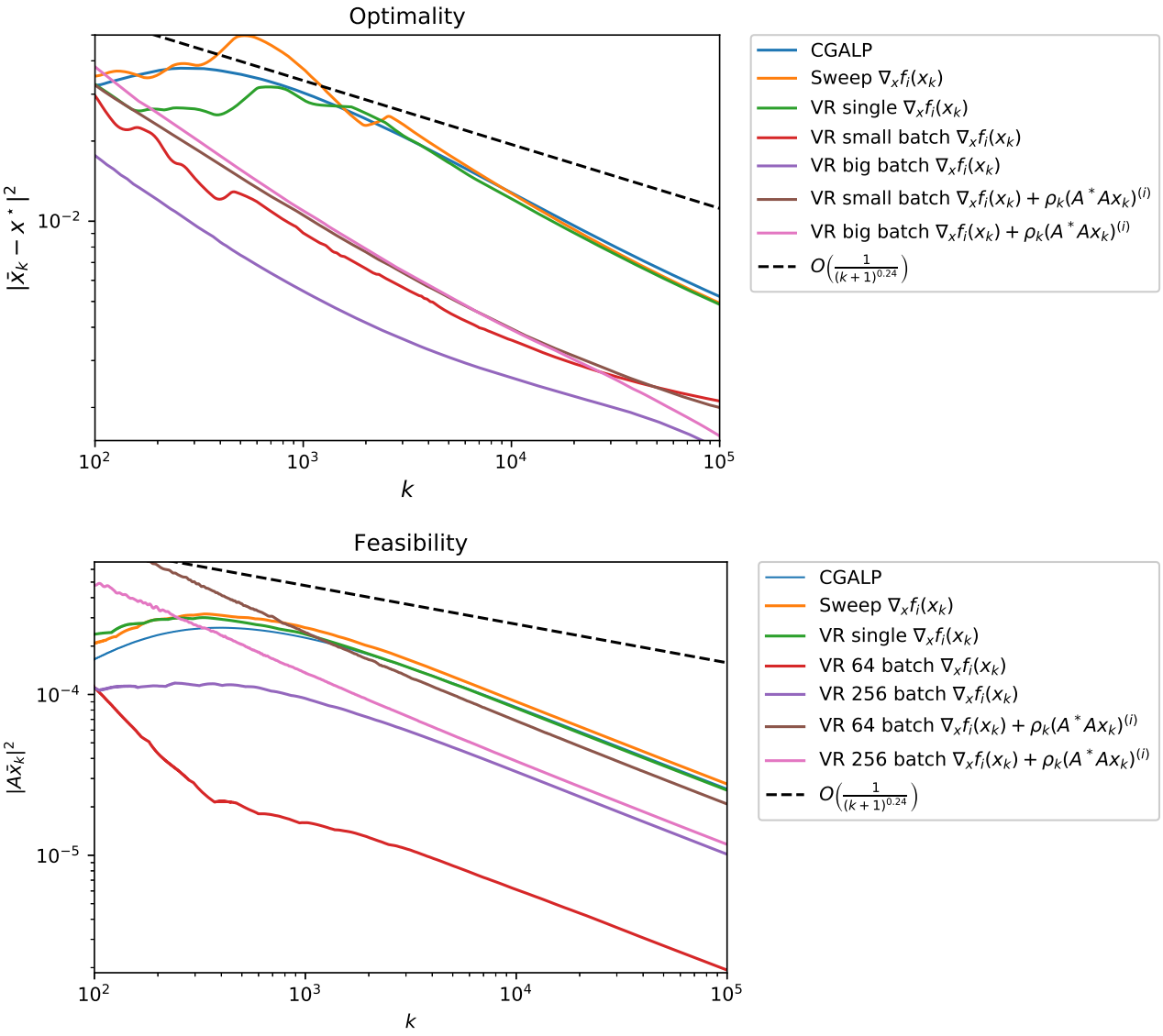

Figure 4.1: Ergodic convergence profiles for ICGALP applied to the projection problem (4.5.1) with $n = 1024$. The step size is, for each $k \in \mathbb{N}$, $\gamma_k = (k+1)^{-\left(1-\frac{1}{4}+0.01\right)}$ and the weight for variance reduction is, for each $k \in \mathbb{N}$, $\nu_k = \gamma_k^{2/3}$.

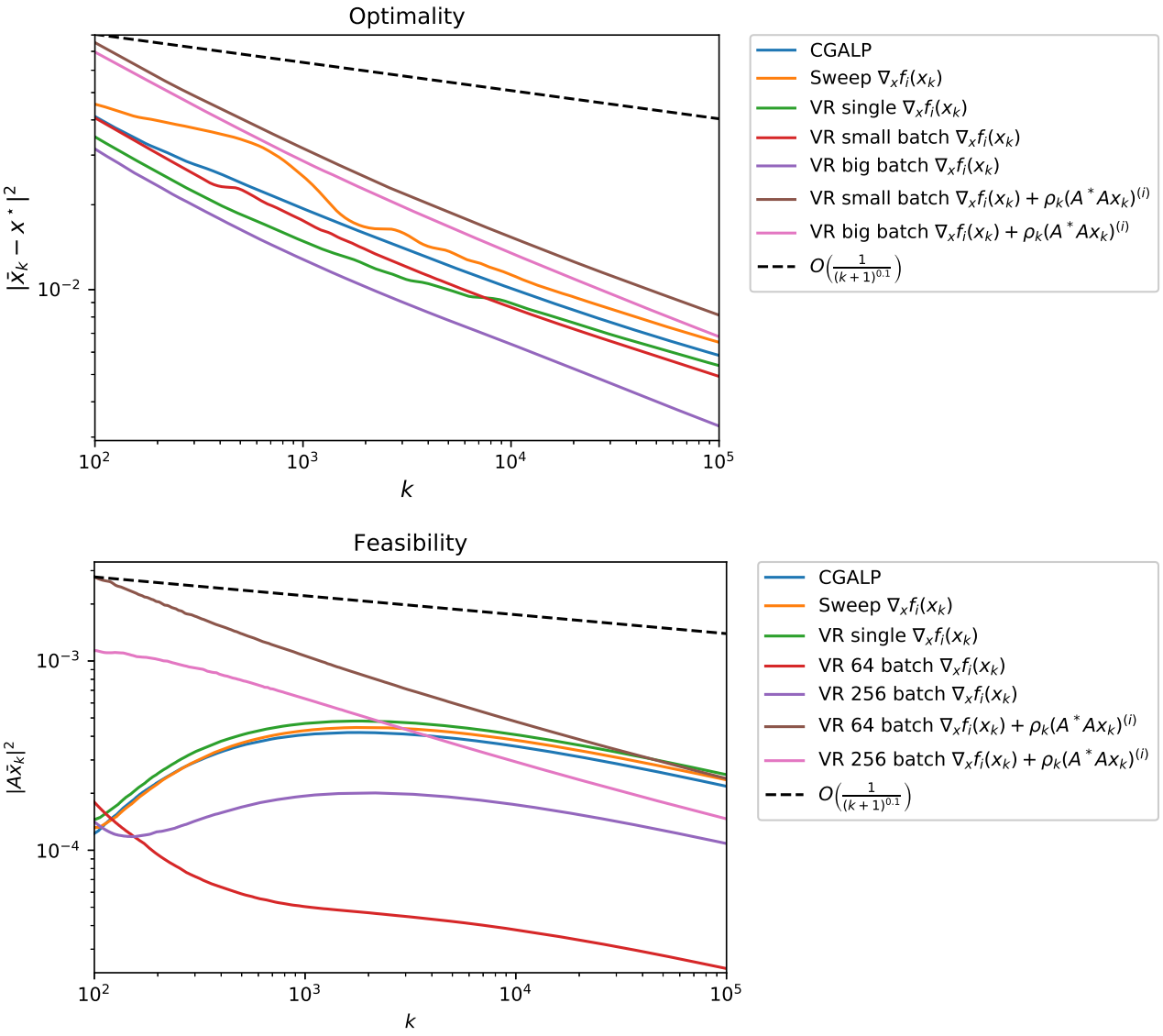

Figure 4.2: Ergodic convergence profiles for ICGALP applied to the projection problem (4.5.1) with $n = 1024$. The step size is, for each $k \in \mathbb{N}$, $\gamma_k = (k+1)^{-\left(1-\frac{1}{4}+0.15\right)}$ and the weight for variance reduction is, for each $k \in \mathbb{N}$, $\nu_k = \gamma_k^{2/3}$.

Chapter 5

Stochastic Bregman Primal-Dual Splitting

In this chapter, we propose and study a novel deterministic Bregman primal-dual algorithm and its inexact/stochastic extension, applicable to relatively smooth, saddle-point problems over reflexive Banach spaces $\mathcal{X}_p$ and $\mathcal{X}_d$ of the form

$$\min_{x \in \mathcal{C}_p \subset \mathcal{X}_p} \max_{\mu \in \mathcal{C}_d \subset \mathcal{X}_d} f(x) + g(x) - h^*(\mu) - l^*(\mu) + \langle Tx, \mu \rangle$$

with $f \in \Gamma_0(\mathcal{X}_p)$ relatively smooth over the non-empty convex set $\mathcal{C}_p$, $g \in \Gamma_0(\mathcal{X}_p)$ prox-friendly with respect to a certain Bregman divergence, $h^* \in \Gamma_0(\mathcal{X}_d)$ relatively smooth over the non-empty convex set $\mathcal{C}_d$, $l^* \in \Gamma_0(\mathcal{X}_d)$ prox-friendly with respect to a certain Bregman divergence, and T a linear continuous operator. Such a structured problem allows one to consider many practically relevant problems that were previously inaccessible. By introducing a new condition that quantifies the relationship between the Bregman divergence and the duality pairing, we are able to carry out a general analysis that doesn't require strong convexity of the entropies with respect to some norm. Our main contributions and findings can be summarized as follows:

Main contributions of this chapter

- Ergodic convergence in expectation of the Lagrangian optimality gap with a rate of $O(1/k)$ and that every almost sure weak cluster point of the ergodic sequence is a primal-dual optimal pair in expectation.
- Weak $\mathbb{P}$-almost sure convergence of the pointwise iterates to a primal-dual optimal pair under slightly stricter assumptions.
- Strong $\mathbb{P}$-almost sure convergence of the pointwise iterates to a primal-dual optimal pair under a relative strong convexity assumption on the objective functions and a total convexity assumption on the entropies.
- Applications: by requiring only relative smoothness over the sets $\mathcal{C}_p$ and $\mathcal{C}_d$, we are able to apply our algorithm to solve inverse problems, inspired by practical applications in data science, which use the Kullback-Leibler divergence as a data fidelity term over the simplex. The Kullback-Leibler divergence is not Lipschitz-smooth and so methods like [31] cannot be applied. We apply this novelty in the deterministic case to solve the trend filtering problem, verifying numerically our claimed convergence rates. We are also able to solve entropically regularized Wasserstein inverse problems, which we implement in the deterministic case numerically. Our new assumptions allow us to drastically reduce the dimensionality of the problem, allowing for compositions of the Wasserstein distance with linear operators and total variation constraints that were, previously, not practically solvable by methods like [41], [28], or [16]. These breakthroughs carry over to the Wasserstein barycenter case as well, which we outline in the applications section.

A paper with the content of this chapter is under preparation for submission to a journal.

Contents

5.1 Introduction

5.1.1 Problem Statement

The goal is to solve the following primal-dual, or saddle-point, problem over the real and reflexive Banach spaces $\mathcal{X}_p$ and $\mathcal{X}_d$:

$$\min_{x\in\mathcal{X}_p}\max_{\mu\in\mathcal{X}_d}\mathcal{L}(x,\mu)\,. \tag{$\mathscr{P}.\mathscr{D}.$}$$

where

$$\mathcal{L}(x,\mu)\overset{\text{def}}{=}f(x)+g(x)+\iota_p(x)+\langle Tx,\ \mu\rangle-h^*(\mu)-l^*(\mu)-\iota_d(\mu)$$

and ι_p and ι_d are the indicator functions of $\mathcal{C}_p$ and $\mathcal{C}_d$, respectively. We denote the primal and dual problems as

$$\min_{x\in\mathcal{C}_p}\left\{f(x)+g(x)+\left[\left(h\underset{\mathcal{C}_d}{\square}l\right)\circ T\right](x)\right\} \tag{$\mathscr{P}$}$$

$$\min_{\mu\in\mathcal{C}_d}\left\{h^*(\mu)+l^*(\mu)+\left[\left(f^*\underset{\mathcal{C}_p}{\square}g^*\right)\circ(-T^*)\right](\mu)\right\}, \tag{$\mathscr{D}$}$$

where $f^*\underset{\mathcal{C}_p}{\square}g^*(v)\overset{\text{def}}{=}\left(f+g+\iota_{\mathcal{C}_p}\right)^*(v)$ and similarly for $\underset{\mathcal{C}_d}{\square}$. In the case in which $\mathcal{C}_p$ and $\mathcal{C}_d$ are trivial constraints, i.e., the entire spaces $\mathcal{X}_p$ and $\mathcal{X}_d$, the corresponding primal and dual problems related to ($\mathscr{P}.\mathscr{D}.$) are

$$\min_{x\in\mathcal{X}_p}\left\{f(x)+g(x)+\left[(h\square l)\circ T\right](x)\right\}$$

$$\min_{\mu\in\mathcal{X}_d}\left\{h^*(\mu)+l^*(\mu)+\left[(f^*\square g^*)\circ(-T^*)\right](\mu)\right\},$$

where $\square$ denotes the classical *inf-convolution* defined by $(f\square g)^*\overset{\text{def}}{=}f^*+g^*$.

We again denote by $\Gamma_0(\mathcal{X})$ the space of proper, convex, and lower semicontinuous functions from $\mathcal{X}$ to $\mathbb{R} \cup \{+\infty\}$. We suppose the following general hypothesis on the problem, which we collectively denote by **(H)**

(H)
- **($\mathbf{H}_1$)** The Banach spaces $\mathcal{X}_p$ and $\mathcal{X}_d$ are real and reflexive, while $\mathcal{C}_p \subset \mathcal{X}_p$ and $\mathcal{C}_d \subset \mathcal{X}_d$ are non-empty, convex, and closed subsets.
- **($\mathbf{H}_2$)** The functions f and g belong to $\Gamma_0(\mathcal{X}_p)$, while l and h belong to $\Gamma_0(\mathcal{X}_d)$; the functions f and h^* are differentiable.
- **($\mathbf{H}_3$)** The operator $T : \mathcal{X}_p \to \mathcal{X}_d$ is linear and continuous.

We introduce two functions, ϕ_p and ϕ_d, and, for simplicity, we denote by D_p and D_d their Bregman divergences (see 2.1.7), respectively. We denote by D the Bregman divergence associated to $\phi(x,\mu) \overset{\text{def}}{=} \phi_p(x) + \phi_d(\mu)$; namely, given $w_i \overset{\text{def}}{=} (x_i, \mu_i)$ with $(x_i, \mu_i) \in \mathcal{X}_p \times \mathcal{X}_d$ for $i \in \{1,2\}$,

$$D(w_1, w_2) \overset{\text{def}}{=} D_p(x_1, x_2) + D_d(\mu_1, \mu_2).$$

For brevity throughout the remainder of the chapter we employ the following notation

$$\mathcal{U}_p \overset{\text{def}}{=} \operatorname{intdom}(\phi_p) \cap \operatorname{dom}(\partial g), \quad \mathcal{U}_d \overset{\text{def}}{=} \operatorname{intdom}(\phi_d) \cap \operatorname{dom}(\partial l^*);$$
$$\tilde{\mathcal{U}}_p \overset{\text{def}}{=} \operatorname{dom}(\phi_p) \cap \operatorname{dom}(\partial g), \quad \tilde{\mathcal{U}}_d \overset{\text{def}}{=} \operatorname{dom}(\phi_d) \cap \operatorname{dom}(\partial l^*).$$

5.1.2 Algorithm

As before, we denote by $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space with set of events Ω, σ-algebra $\mathcal{F}$, and probability measure $\mathbb{P}$. Throughout, we assume that a Banach space $\mathcal{X}$ is endowed with its Borel σ-algebra, $\mathcal{B}(\mathcal{X})$.

We consider the possibility of some stochastic error in the computation of the gradients[1] ∇f and ∇h^* which we will denote for $\nabla f(x_k)$ as δ_k^p and for $\nabla h^*(\mu_k)$ as δ_k^d, i.e., δ_k^p and δ_k^d are measurable functions from Ω to $\mathcal{X}_p^*$ and $\mathcal{X}_d^*$ with their respective Borel σ-algebras. When it makes sense, we will also denote the combined error as Δ_k in the same way that we use w_k, e.g. $\langle \Delta_k, w - w_k \rangle \overset{\text{def}}{=} \langle \delta_k^p, x - x_k \rangle + \langle \delta_k^d, \mu - \mu_k \rangle$. The stochastic algorithm is given in Algorithm 10.

Algorithm 10: Stochastic Bregman Primal-Dual Splitting.

for $k = 0, 1, \ldots$ **do**

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{C}_p} \left\{ g(x) + \langle \nabla f(x_k) + \delta_k^p, x \rangle + \langle Tx, \bar{\mu}_k \rangle + \frac{1}{\lambda_k} D_p(x, x_k) \right\}$$

$$\mu_{k+1} = \operatorname*{argmin}_{\mu \in \mathcal{C}_d} \left\{ l^*(\mu) + \langle \nabla h^*(\mu_k) + \delta_k^d, \mu \rangle - \langle T\bar{x}_k, \mu \rangle + \frac{1}{\nu_k} D_d(\mu, \mu_k) \right\}$$

where $\bar{\mu}_k = \mu_k$ and $\bar{x}_k = 2x_{k+1} - x_k$.

Notice that, setting $\delta_k^p = 0$ for each $k \in \mathbb{N}$, the first step of the algorithm can be re-written in the following way:

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{x \in \mathcal{C}_p} \left\{ g(x) + f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \langle Tx, \bar{\mu}_k \rangle + \frac{1}{\lambda_k} D_p(x, x_k) \right\} \\ &= \left[\nabla \phi_p + \lambda_k \partial g \right]^{-1} \left[\nabla \phi_p - \lambda_k \nabla \left(f(\cdot) + \langle T\cdot, \bar{\mu}_k \rangle \right) \right] (x_k) \\ &= \left[\nabla \phi_p + \lambda_k \partial g \right]^{-1} \left(\nabla \phi_p(x_k) - \lambda_k \nabla f(x_k) - \lambda_k T^* \bar{\mu}_k \right). \end{aligned}$$

[1] The addition of stochastic error in the computation of D-prox operators associated to g or l^*, while interesting, is problematic for the algorithm in the sense that the monotone inclusions may no longer hold and the iterates themselves might not remain in $\mathcal{U}_p \times \mathcal{U}_d$ as desired.

Analogously, if $\delta_k^d = 0$ for all $k \in \mathbb{N}$,

$$\mu_{k+1} = \left[\nabla\phi_d + \nu_k \partial l^*\right]^{-1} \left(\nabla\phi_d\left(\mu_k\right) - \nu_k \nabla h^*\left(\mu_k\right) + \nu_k T\bar{x}_k\right).$$

A priori, the mappings $\left[\nabla\phi_p + \lambda_k\partial g\right]^{-1}$ and $\left[\nabla\phi_d + \nu_k\partial l^*\right]^{-1}$, sometimes referred to as D-prox mappings, may be empty, may not be single-valued, or may not map $\operatorname{intdom}\left(\phi_p\right)$ (resp. $\operatorname{intdom}\left(\phi_d\right)$) to $\operatorname{intdom}\left(\phi_p\right)$ (resp. $\operatorname{intdom}\left(\phi_d\right)$). In light of this, we will only consider ϕ_p and ϕ_d for which these mappings are well-defined and map from $\operatorname{intdom}\left(\phi_p\right)$ to $\operatorname{intdom}\left(\phi_p\right)$ and the analog for ϕ_d (See $(\mathbf{A}_1)$). In Section 5.1.3, we will elaborate on the class of Legendre functions in a reflexive Banach space given in [12][Definition 2.2] (see 2.1.8) which will help us to ensure that the D-prox mappings are well-defined.

5.1.3 Assumptions

For the remainder of the chapter, all equalities and inequalities involving random quantities should be understood as holding ($\mathbb{P}$-a.s.) unless explicitly written otherwise. Using the notation of Section 2.3, we denote the canonical filtration as $\mathfrak{S} \overset{\text{def}}{=} \left(\mathcal{S}_k\right)_{k\in\mathbb{N}}$ with $\mathcal{S}_k \overset{\text{def}}{=} \sigma\left\{\left(x_0,\mu_0\right),\left(x_1,\mu_1\right),\ldots,\left(x_k,\mu_k\right)\right\}$ such that all iterates up to $\left(x_k,\mu_k\right)$ are completely determined by $\mathcal{S}_k$.

Before the introduction of the main assumptions considered along the chapter, we define the following notation:

$$\begin{aligned}
\left(\frac{1}{\Lambda_k} - L\right) D\left(w_1,w_2\right) &\overset{\text{def}}{=} \left(\frac{1}{\lambda_k} - L_p\right) D_p\left(x_1,x_2\right) + \left(\frac{1}{\nu_k} - L_d\right) D_d\left(\mu_1,\mu_2\right); \\
\left(\frac{1}{\Lambda_\infty} - L\right) D\left(w_1,w_2\right) &\overset{\text{def}}{=} \left(\frac{1}{\lambda_\infty} - L_p\right) D_p\left(x_1,x_2\right) + \left(\frac{1}{\nu_\infty} - L_d\right) D_d\left(\mu_1,\mu_2\right); \\
M(w_1,w_2) &\overset{\text{def}}{=} \langle T(x_1-x_2), \mu_1-\mu_2\rangle,
\end{aligned} \tag{5.1.1}$$

where λ_k, ν_k are the step-sizes and L_p, L_d are the constants introduced in $(\mathbf{A}_1)$. Analogously to (5.1.1), we define also the following notation using the relative strong convexity constants from $(\mathbf{A}_{11})$:

$$\begin{aligned}
\left(\frac{1}{\Lambda_k} - m_{(f,h^*)}\right) D\left(w_1,w_2\right) &\overset{\text{def}}{=} \left(\frac{1}{\lambda_k} - m_f\right) D_p\left(x_1,x_2\right) + \left(\frac{1}{\nu_k} - m_{h^*}\right) D_d\left(\mu_1,\mu_2\right); \\
\left(\frac{1}{\Lambda_k} - m_{(g,l^*)}\right) D\left(w_1,w_2\right) &\overset{\text{def}}{=} \left(\frac{1}{\lambda_k} - m_g\right) D_p\left(x_1,x_2\right) + \left(\frac{1}{\nu_k} - m_{l^*}\right) D_d\left(\mu_1,\mu_2\right).
\end{aligned} \tag{5.1.2}$$

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

$$\begin{aligned}
\boldsymbol{S}_{\mathscr{P}} &\overset{\text{def}}{=} \operatorname*{argmin}_{x\in\mathcal{C}_p} \left\{\max_{\mu\in\mathcal{C}_d} \left\{f\left(x\right) + g\left(x\right) + \langle Tx,\mu\rangle - h^*\left(\mu\right) - l^*\left(\mu\right)\right\}\right\} \\
\boldsymbol{S}_{\mathscr{D}} &\overset{\text{def}}{=} \operatorname*{argmax}_{\mu\in\mathcal{C}_d} \left\{\min_{x\in\mathcal{C}_p} \left\{f\left(x\right) + g\left(x\right) + \langle Tx,\mu\rangle - h^*\left(\mu\right) - l^*\left(\mu\right)\right\}\right\}
\end{aligned} \tag{5.1.3}$$

and the notation for the weak cluster points of an arbitrary sequence $\left(x_k\right)_{k\in\mathbb{N}}$ in some Banach space $\mathcal{X}$ to be

$$\mathfrak{W}\left[\left(x_k\right)_{k\in\mathbb{N}}\right] \overset{\text{def}}{=} \left\{x \in \mathcal{X} : \exists \left(x_{k_j}\right)_{j\in\mathbb{N}}, x_{k_j} \rightharpoonup x\right\}. \tag{5.1.4}$$

We first state our assumptions and then remark on their motivations and common situations where they hold.

- $(\mathbf{A}_1)$ The two functions ϕ_p and ϕ_d are Legendre functions belonging to $\Gamma_0\left(\mathcal{X}_p\right)$ and $\Gamma_0\left(\mathcal{X}_d\right)$ with $\operatorname{dom}\left(\phi_p+\phi_d\right) = \mathcal{C}_p \times \mathcal{C}_d$ and with f and h^* being L_p and L_d - smooth w.r.t. ϕ_p and ϕ_d, respectively (see Definition 2.1.9). The D-prox mappings $\left[\nabla\phi_p + \lambda_k\partial g\right]^{-1}$ and $\left[\nabla\phi_d + \nu_k\partial l^*\right]^{-1}$ are well-defined (i.e., nonempty and single-valued) maps from $\operatorname{intdom}\left(\phi_p\right)$ and $\operatorname{intdom}\left(\phi_d\right)$ to $\operatorname{intdom}\left(\phi_p\right)$ and $\operatorname{intdom}\left(\phi_d\right)$, respectively.
- $(\mathbf{A}_2)$ The step size sequences $\left(\lambda_k\right)_{k\in\mathbb{N}}$ and $\left(\nu_k\right)_{k\in\mathbb{N}}$ are positive, nondecreasing, and bounded above with their limits denoted $\lim_{k\to\infty}\lambda_k = \lambda_\infty$ and $\lim_{k\to\infty}\nu_k = \nu_\infty$.
- $(\mathbf{A}_3)$ The step sizes satisfy $(\mathbf{A}_2)$ and one of the following holds:

(I) there is a function $d : (\mathcal{X}_p \times \mathcal{X}_d)^2 \to \mathbb{R}_+$ and $\varepsilon \geq 0$ such that

$$\inf_{\substack{w_1 \in \tilde{\mathcal{U}}_p \times \tilde{\mathcal{U}}_d,\ w_2 \in \mathcal{U}_p \times \mathcal{U}_d; \\ w_1 \neq w_2}} \frac{\left(\frac{1}{\Lambda_\infty} - L\right) D(w_1, w_2) - M(w_1, w_2)}{d(w_1, w_2)} \geq \varepsilon; \tag{5.1.5}$$

(II) the above holds with $\varepsilon > 0$.

($\mathbf{A}_4$) The error $(\Delta_k)_{k\in\mathbb{N}}$ is unbiased conditioned on the filtration $\mathfrak{S}$, i.e., for each $k \in \mathbb{N}$,

$$\mathbb{E}\left[\delta_k^p \mid \mathcal{S}_k\right] = \mathbb{E}\left[\delta_k^d \mid \mathcal{S}_k\right] = 0,$$

and one of the following holds (and analogously, although not necessarily the same case, for the dual):

(I) for each $k \in \mathbb{N}$, the stochastic error δ_k^p is zero almost surely;

(II) The following sequences satisfy,

$$\left(\mathbb{E}\left[\left\|\delta_k^p\right\| \mid \mathcal{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S}) \quad \text{and} \quad \left(\mathbb{E}\left[\left\|\delta_k^p\right\|\right]\right)_{k\in\mathbb{N}} \in \ell_+^1,$$

and the set $\mathcal{C}_p$ is bounded, i.e., $0 < \operatorname{diam}_{\mathcal{C}_p} < +\infty$;

(III) The entropies ϕ_p and ϕ_d are strongly convex with respect to $\|\cdot\|_p^2$ and $\|\cdot\|_d^2$ with modulus m_p m_d, respectively, the step sizes $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ satisfy **($\mathbf{A}_2$)** and

$$\nu_\infty \lambda_\infty < \frac{m_p m_d}{\|T\|_{p\to d^*}^2},$$

where $\|\cdot\|_{p\to d^*}$ is a standard operator norm between $\mathcal{X}_p$ and $\mathcal{X}_d^*$, and the following sequences satisfy

$$\mathbb{E}\left[\left\|\delta_k^p\right\|^2 \mid \mathcal{S}_k\right] \in \ell_+^1(\mathfrak{S}) \quad \text{and} \quad \mathbb{E}\left[\left\|\delta_k^p\right\|^2\right] \in \ell_+^1.$$

($\mathbf{A}_5$) For all sequences $(v_k)_{k\in\mathbb{N}}$ and $(z_k)_{k\in\mathbb{N}}$ in $\operatorname{intdom}(\phi)$

$$d(v_k, z_k) \to 0 \quad \Rightarrow \quad v_k - z_k \to 0. \tag{5.1.6}$$

($\mathbf{A}_6$) For every $w \overset{\text{def}}{=} (x, \mu) \in \operatorname{intdom}(\phi)$, at least one of $D(w, \cdot)$ or $d(w, \cdot)$ is coercive.

($\mathbf{A}_7$) The set-valued operator $\partial g + \nabla f + T^*$ is maximal monotone and similarly for $\partial l^* + \nabla h^* - T$.

($\mathbf{A}_8$) For any sequence $(w_k)_{k\in\mathbb{N}}$ with $w_k \in \operatorname{intdom}\phi$ for each $k \in \mathbb{N}$, if $w_{k+1} - w_k \to 0$, then

$$\nabla\phi_p(x_{k+1}) - \nabla\phi_p(x_k) \to 0 \ \text{ and } \ \nabla f(x_{k+1}) - \nabla f(x_k) \to 0;$$
$$\nabla\phi_d(\mu_{k+1}) - \nabla\phi_d(\mu_k) \to 0 \ \text{ and } \ \nabla h^*(\mu_{k+1}) - \nabla h^*(\mu_k) \to 0.$$

($\mathbf{A}_9$) For any sequence $(w_k)_{k\in\mathbb{N}}$ with $w_k \in \operatorname{intdom}\phi$, for each $k \in \mathbb{N}$, if $w_k \rightharpoonup w_\infty$, then

$$\nabla\phi(w_k) \rightharpoonup \nabla\phi(w_\infty).$$

($\mathbf{A}_{10}$) If $w_k \rightharpoonup 0$, then

$$\langle Tx_k, \mu_k\rangle \to 0.$$

($\mathbf{A}_{11}$) One of the functions f, g, or both are relatively strongly convex w.r.t. ϕ_p with constant m_f, m_g, or $m_f + m_g$, respectively (see Definition 2.1.11).

Remark 5.1.1. There are several, technical characterizations of sufficient conditions that ensure the latter half of **($\mathbf{A}_1$)** holds. They can be found, for instance, in [12][Theorem 3.18] for the reflexive Banach space case or in [13][Lemma 2] for the Euclidean case. In practice, these conditions are seldom violated and otherwise not used in the analysis of the algorithm. For **($\mathbf{A}_2$)**, it is sufficient to take the sequences $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ to simply be constant.

Remark 5.1.2. The infimum in **($\mathbf{A}_3$)** is taken with $w_1 \in \tilde{\mathcal{U}}_p \times \tilde{\mathcal{U}}_d$ and $w_2 \in \mathcal{U}_p \times \mathcal{U}_d$ because, a priori, a solution w^* may lie in the boundary of $\tilde{\mathcal{U}}_p \times \tilde{\mathcal{U}}_d$ even if the iterates $(w_k)_{k\in\mathbb{N}}$ themselves remain in $\mathcal{U}_p \times \mathcal{U}_d$.

Since the Bregman divergence is still well defined when the first argument (but not necessarily the second) is in $\mathrm{dom}\,(\phi) \setminus \mathrm{intdom}\,(\phi)$, there is no issue with taking the infimum over this set. Observe that $(\mathbf{A}_3)$ also entails that, for every $w_1 \in \tilde{\mathcal{U}}_p \times \tilde{\mathcal{U}}_d$ and $w_2 \in \mathcal{U}_p \times \mathcal{U}_d$, for each $k \in \mathbb{N}$,

$$\frac{1}{\Lambda_k} D\left(w_1, w_2\right) - M\left(w_1, w_2\right) \geq L D\left(w_1, w_2\right) + \varepsilon d\left(w_1, w_2\right) \geq 0. \tag{5.1.7}$$

Remark 5.1.3 (Example). Suppose that $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a convex, nondecreasing function with φ^* its positive conjugate and γ a finite, coercive gauge with domain $\mathbb{R}_+ \left(\mathcal{U}_p - \mathcal{U}_p\right) \subset \mathcal{X}_d$ (in the Minkowski sense) and polar γ°. Assume that the quantities defined by

$$\begin{aligned} \|T\|_{D_p} &\overset{\text{def}}{=} \sup_{x_1, x_2 \in \mathcal{U}_p;\ x_1 \neq x_2} \frac{\varphi(\gamma\left(T(x_2 - x_1)\right))}{D_p\left(x_1, x_2\right)}; \\ \|I\|_{D_d} &\overset{\text{def}}{=} \sup_{\mu_1, \mu_2 \in \mathcal{U}_d;\ \mu_1 \neq \mu_2} \frac{\varphi^*(\gamma^\circ\left(\mu_2 - \mu_1\right))}{D_d\left(\mu_1, \mu_2\right)} \end{aligned}$$

are finite. We use the notation $\|\cdot\|_{D_p}$ and $\|\cdot\|_{D_p}$, but notice that they may not be norms. If, moreover, we suppose that the step sizes verify, for each $k \in \mathbb{N}$, for some $\varepsilon_k \geq 0$,

$$\left(\frac{1}{\lambda_k} - L_p\right) \geq \|T\|_{D_p} + \varepsilon_k \ \text{ and } \ \left(\frac{1}{\nu_k} - L_d\right) \geq \|I\|_{D_d} + \varepsilon_k, \tag{5.1.8}$$

then $(\mathbf{A}_3)$ is satisfied with $d\left(w_1, w_2\right) = D\left(w_1, w_2\right)$. Indeed, for any pair $w_1, w_2 \in \mathcal{U}_p \times \mathcal{U}_d$, we have, for each $k \in \mathbb{N}$,

$$\begin{aligned} & \left(\frac{1}{\Lambda_k} - L\right) D\left(w_1, w_2\right) - M\left(w_1, w_2\right) \\ = \ & \left(\frac{1}{\lambda_k} - L_p\right) D_p\left(x_1, x_2\right) + \left(\frac{1}{\nu_k} - L_d\right) D_d\left(\mu_1, \mu_2\right) - \langle T(x_1 - x_2), \mu_1 - \mu_2 \rangle \\ \geq \ & \|T\|_{D_p} D_p\left(x_1, x_2\right) + \|I\|_{D_d} D_d\left(\mu_1, \mu_2\right) - \gamma\left(T(x_1 - x_2)\right) \gamma^\circ\left(\mu_1 - \mu_2\right) + \varepsilon_k D\left(w_1, w_2\right) \\ \geq \ & \varphi(\gamma\left(T(x_1 - x_2)\right)) + \varphi^*(\gamma^\circ\left(\mu_1 - \mu_2\right)) - \varphi(\gamma\left(T(x_1 - x_2)\right)) - \varphi^*(\gamma^\circ\left(\mu_1 - \mu_2\right)) + \varepsilon_k D\left(w_1, w_2\right) \\ = \ & \varepsilon_k D\left(w_1, w_2\right). \end{aligned} \tag{5.1.9}$$

Note that in the above example we have taken the action of T on the primal variables into the definition of $\|\cdot\|_{D_p}$. It is equally possible, and sometimes desirable, to the define things such that the action of the adjoint T^* on the dual variables is incorporated into $\|\cdot\|_{D_d}$ instead, which can drastically change the values (and consequently step sizes) in a non-Hilbertian setting.

Remark 5.1.4. Notice that, using Lemma 2.3.4, $(\mathbf{A}_4)$ (in any case) implies that $\left(\delta_k^p\right)_{k\in\mathbb{N}}$ and $\left(\delta_k^d\right)_{k\in\mathbb{N}}$ converge strongly (with respect to $\|\cdot\|_{p^*}$ and $\|\cdot\|_{d^*}$ respectively) to zero a.s. and, furthermore, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$, $\left(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1} \rangle\right]\right)_{k\in\mathbb{N}} \in \ell_+^1$ and $\left(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1} \rangle \mid \mathcal{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1$ $(\mathfrak{S})$ (see Lemma 5.2.5 for details). In $(\mathbf{A}_4)$**(III)**, the norm $\|\cdot\|_p$ can be arbitrary as long as ϕ_p is strongly convex with respect to its square.

Remark 5.1.5. In the case where $d\left(x, y\right)$ is the Bregman divergence induced by the Shannon-Boltzman entropy, the Hellinger, entropy, the fractional-power entropy, the Fermi-Dirac entropy, or the energy/euclidean entropy, $(\mathbf{A}_5)$ holds (see [13, Remark 4].

More generally, when $d = D_\psi$ for some entropy ψ, we have from [12][Example 4.10] that $(\mathbf{A}_5)$ is satisfied whenever one of the following holds

- ψ is uniformly convex on bounded sets;
- $\mathcal{X}_p \times \mathcal{X}_d$ is finite dimensional, $\operatorname{dom}(\psi)$ is closed, and $\psi\,|_{\operatorname{dom}(\psi)}$ is strictly convex and continuous.

Thus, if $\psi = \phi$, since ϕ is assumed to be Legendre by **(H)**, we require only $\operatorname{dom}(\phi)$ closed if $\mathcal{X}_p \times \mathcal{X}_d$ is finite dimensional.

Remark 5.1.6. Sufficient conditions for **($\mathbf{A}_6$)** to hold for Legendre functions in reflexive Banach spaces are given in [11][Lemma 7.3(viii) & (ix)]. For example, if ϕ_p is supercoercive and $x \in \operatorname{intdom}(\phi_p)$ then $D_p(x, \cdot)$ is coercive or if $\mathcal{X}_p$ is finite-dimensional, $\operatorname{dom}(\phi_p^*)$ is open, and $x \in \operatorname{intdom}(\phi_p)$ then $D_p(x, \cdot)$ is coercive (and similar conditions for ϕ_d).

Remark 5.1.7. There are numerous sufficient conditions for **($\mathbf{A}_7$)** to hold. In reflexive Banach spaces, we can impose one of the following (see [19])

- $\operatorname{intdom}(\partial g) \cap \operatorname{intdom}(\nabla f) \neq \emptyset$ and its analog for the dual.
- $\operatorname{dom}(\nabla f) \cap \operatorname{intdom}(\partial g) \neq \emptyset$ while $\operatorname{dom}(\nabla f)$ is closed and convex.
- $\operatorname{dom}(\nabla f)$ and $\operatorname{dom}(\partial g)$ are closed and convex and $0 \in \operatorname{core}(\operatorname{conv}(\operatorname{dom}(\nabla f) - \operatorname{dom}(\partial g)))$.

Relaxed conditions exist for finite dimension (see [32]).

Remark 5.1.8. Note that **($\mathbf{A}_8$)**, **($\mathbf{A}_9$)**, and **($\mathbf{A}_{10}$)** all hold in the case where $\mathcal{X}_p \times \mathcal{X}_d$ is finite dimensional. Indeed, in finite dimension not only do strong and weak convergence coincide but also $\nabla\phi_p$, $\nabla\phi_d$, ∇f, and ∇h^* are all continuous on the interior of their domains by [104][Corollary 9.20] since $\phi_p, f \in \Gamma_0(\mathcal{X}_p)$ and $\phi_d, h^* \in \Gamma_0(\mathcal{X}_d)$.

5.1.4 Organization of the Chapter

The rest of the chapter is divided into four sections.

In Section 5.2, we establish the main estimation of Lemma 5.2.1 under **(H)** and **($\mathbf{A}_1$)**-**($\mathbf{A}_3$)** that will be used in the convergence analysis of the ergodic, pointwise, and relatively strongly convex cases. The key idea is to utilize the descent lemma given by relative smoothness along with the usual inequalities for $\Gamma_0(\mathcal{X})$ functions to estimate the optimality gap $\mathcal{L}(x_{k+1}, \mu) - \mathcal{L}(x, \mu_{k+1})$ in terms of the Bregman divergences induced by the entropies ϕ_p and ϕ_d. The proof of the estimation here is similar in spirit to the proof of the main estimation in [29], with the main difference being that we are unable to use Young's inequality to deal with the M terms, which we handle using **($\mathbf{A}_3$)**. There are also some lemmas involving **(H)** and **($\mathbf{A}_1$)**-**($\mathbf{A}_4$)** regarding the stochastic error, culminating in a summability result for the sequences $(\mathbb{E}[\langle \Delta_k, w - w_{k+1}\rangle])_{k\in\mathbb{N}}$ and $(\mathbb{E}[\langle \Delta_k, w - w_{k+1}\rangle \mid \mathcal{S}_k])_{k\in\mathbb{N}}$.

In Section 5.3, we use the estimation developed in Section 5.2 along with **(H)** and **($\mathbf{A}_1$)**-**($\mathbf{A}_{11}$)** regarding the entropies ϕ_p and ϕ_d and the regularity of their induced Bregman divergences to show convergence of the algorithm; first convergence of the expectation of the Lagrangian optimality gap in the ergodic sense under **(H)** and **($\mathbf{A}_1$)**-**($\mathbf{A}_4$)**, then almost sure weak convergence of the iterates in the pointwise sense under **(H)** and **($\mathbf{A}_1$)**-**($\mathbf{A}_{10}$)**, and finally we examine the case where **($\mathbf{A}_{11}$)** holds, i.e., there is relative strong covexity of the objective functions with respect to the entropies, and total conexity of the entropies themselves. For the ergodic analysis, we show that every almost sure weak sequential cluster point of the ergodic primal-dual sequence $((\bar{x}_k, \bar{\mu}_k))_{k\in\mathbb{N}}$ is a primal-dual optimal pair in expectation, $\mathbb{E}[(x_\infty, \mu_\infty)] = (x^\star, \mu^\star)$, and also convergence of the expectation of the Lagrangian optimality gap $\mathbb{E}[\mathcal{L}(\bar{x}_{k+1}, \mu) - \mathcal{L}(x, \bar{\mu}_{k+1})]$ with a rate of $O(1/k)$. For the pointwise analysis, we begin by showing an almost sure asymptotic regularity result for the primal-dual sequence (w_k). With this, we are then able to adapt the well known Opial's lemma (see [89]) to the Bregman primal-dual setting to establish almost sure weak convergence of the primal-dual sequence $(w_k)_{k\in\mathbb{N}}$ to a primal-dual optimal pair $w^\star$. In the final part of this section, we establish almost sure strong convergence of the primal-dual sequence $(w_k)_{k\in\mathbb{N}}$ to a primal-dual optimal pair $w^\star$ under **($\mathbf{A}_{11}$)** and total convexity of the entropies.

Finally, in Section 5.4, we explore potential applications of the algorithm and demonstrate numerically its effectiveness when applied to three different problems. The first is a simple linear inverse problem on the simplex. The second is a type of simplex regression with so-called trend filtering using the Kullback-Leibler divergence. The last problem is an application optimal transport involving the entropically regularized Wasserstein distance and inverse problems. There is also a discussion of other possible applications of the algorithm to entropic Wasserstein barycenter problems.

5.2 Preliminary Estimations

The following results provide the main estimations that will be instrumental in the convergence analysis of Algorithm 10. We start with an energy estimation of in the first section and then move on to estimating the error, Δ_k, in the second section.

5.2.1 Main Energy Estimation

Lemma 5.2.1. *Recall the notation of (5.1.1) and (5.1.2). Assume that (H), ($\mathbf{A}_1$), ($\mathbf{A}_2$), and ($\mathbf{A}_3$) hold, then we have the following energy estimation. For every $w \overset{\text{def}}{=} (x,\mu) \in \mathcal{C}_p \times \mathcal{C}_d$, for each $k \in \mathbb{N}$,*

$$\begin{aligned} &\mathcal{L}(x_{k+1},\mu) - \mathcal{L}(x,\mu_{k+1}) + \left[\frac{1}{\Lambda_{k+1}} D(w,w_{k+1}) - M(w,w_{k+1})\right] + \langle w_{k+1} - w, \Delta_k\rangle \\ &+ \varepsilon d(w_{k+1},w_k) \leq \left[\frac{1}{\Lambda_k} D(w,w_k) - M(w,w_k)\right]. \end{aligned} \tag{5.2.1}$$

If, moreover, ($\mathbf{A}_{11}$) holds for the primal and the dual, we have (using the notation of (5.1.2)) for every $w \overset{\text{def}}{=} (x,\mu) \in \mathcal{C}_p \times \mathcal{C}_d$, for each $k \in \mathbb{N}$,

$$\begin{aligned} &\mathcal{L}(x_{k+1},\mu) - \mathcal{L}(x,\mu_{k+1}) + \left[\left(\frac{1}{\Lambda_{k+1}} + m_{(g,l^*)}\right) D(w,w_{k+1}) - M(w,w_{k+1})\right] + \langle w_{k+1} - w, \Delta_k\rangle \\ &+ \varepsilon d(w_{k+1},w_k) \leq \left[\left(\frac{1}{\Lambda_k} - m_{(f,h^*)}\right) D(w,w_k) - M(w,w_k)\right]. \end{aligned} \tag{5.2.2}$$

Proof. For any $(x,\mu) \in \mathcal{C}_p \times \mathcal{C}_d$, the following holds by the definitions of x_{k+1} and μ_{k+1} in Algorithm 10 and the relative strong convexity of g and l^* with respect to D_p and D_d with constants m_g and m_{l^*}, respectively:

$$\begin{aligned} D_p(x,x_k) &\geq \lambda_k \left[g(x_{k+1}) - g(x) + \langle \nabla f(x_k) + \delta_k^p,\ x_{k+1} - x\rangle + \langle T(x_{k+1} - x),\ \bar{\mu}_k\rangle\right] \\ &\quad + (1 + m_g\lambda_k) D_p(x,x_{k+1}) + D_p(x_{k+1},x_k); \\ D_d(\mu,\mu_k) &\geq \nu_k \left[l^*(\mu_{k+1}) - l^*(\mu) + \langle \nabla h^*(\mu_k) + \delta_k^d,\ \mu_{k+1} - \mu\rangle - \langle T\bar{x}_k,\ \mu_{k+1} - \mu\rangle\right] \\ &\quad + (1 + m_{l^*}\nu_k) D_d(\mu,\mu_{k+1}) + D_d(\mu_{k+1},\mu_k). \end{aligned} \tag{5.2.3}$$

Moreover, from the relative smoothness asumed in ($\mathbf{A}_1$) and the consequent generalized descent lemma (2.1.9), we have, for each $k \in \mathbb{N}$,

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k),\ x_{k+1} - x_k\rangle + L_p D_p(x_{k+1},x_k); \\ h^*(\mu_{k+1}) &\leq h^*(\mu_k) + \langle \nabla h^*(\mu_k),\ \mu_{k+1} - \mu_k\rangle + L_d D_d(\mu_{k+1},\mu_k). \end{aligned} \tag{5.2.4}$$

By convexity (or relative strong-convexity in the case $m_f > 0$ or $m_{h^*} > 0$) of f and $h^\star$ (see Definition 2.1.11), we have, for each $(x,\mu) \in \mathcal{C}_p \times \mathcal{C}_d$, for each $k \in \mathbb{N}$,

$$\begin{aligned} f(x) &\geq f(x_k) + \langle \nabla f(x_k),\ x - x_k\rangle + m_f D_p(x,x_k); \\ h^*(\mu) &\geq h^*(\mu_k) + \langle \nabla h^*(\mu_k),\ \mu - \mu_k\rangle + m_{h^*} D_d(\mu,x_k). \end{aligned} \tag{5.2.5}$$

Summing (5.2.4) and (5.2.5), we obtain, for each $(x,\mu) \in \mathcal{C}_p \times \mathcal{C}_d$, for each $k \in \mathbb{N}$,

$$\begin{aligned} f(x_{k+1}) &\leq f(x) + \langle \nabla f(x_k),\ x_{k+1} - x\rangle + L_p D_p(x_{k+1},x_k) - m_f D_p(x,x_k); \\ h^*(\mu_{k+1}) &\leq h^*(\mu) + \langle \nabla h^*(\mu_k),\ \mu_{k+1} - \mu\rangle + L_d D_d(\mu_{k+1},\mu_k) - m_{h^*} D_d(\mu,\mu_k). \end{aligned}$$

Summing the latter with (5.2.3), we have, for each $(x,\mu)\in\mathcal{C}_p\times\mathcal{C}_d$, for each $k\in\mathbb{N}$,

$$\lambda_k\left[f(x_{k+1})+g(x_{k+1})-f(x)-g(x)+\langle T\left(x_{k+1}-x\right),\,\bar{\mu}_k\rangle+\left\langle x_{k+1}-x,\delta_k^p\right\rangle\right]+\left(1+m_g\lambda_k\right)D_p\left(x,x_{k+1}\right)$$
$$+(1-L_p\lambda_k)D_p\left(x_{k+1},x_k\right)\leq\left(1-m_f\lambda_k\right)D_p\left(x,x_k\right);$$

$$\nu_k\left[h^*(\mu_{k+1})+l^*(\mu_{k+1})-h^*(\mu)-l^*(\mu)-\langle T\bar{x}_k,\;\mu_{k+1}-\mu\rangle+\left\langle\mu_{k+1}-\mu,\delta_k^d\right\rangle\right]+\left(1+m_{l^*}\nu_k\right)D_d\left(\mu,\mu_{k+1}\right)$$
$$+(1-L_d\nu_k)D_d\left(\mu_{k+1},\mu_k\right)\leq\left(1-m_{h^*}\nu_k\right)D_d\left(\mu,\mu_k\right).$$

Recall the notations of (5.1.1), (5.1.2), and that

$$\langle w_1-w_2,\Delta_k\rangle\overset{\text{def}}{=}\left\langle x_1-x_2,\delta_k^p\right\rangle+\left\langle\mu_1-\mu_2,\delta_k^d\right\rangle.$$

Then, for each $(x,\mu)\in\mathcal{C}_p\times\mathcal{C}_d$, for each $k\in\mathbb{N}$,

$$\begin{aligned}&\mathcal{L}\left(x_{k+1},\mu\right)-\mathcal{L}\left(x,\mu_{k+1}\right)+\langle T\left(x_{k+1}-x\right),\bar{\mu}_k\rangle-\langle T\bar{x}_k,\mu_{k+1}-\mu\rangle+\langle w_{k+1}-w,\Delta_k\rangle\\&+\left(\frac{1}{\Lambda_k}+m_{(g,l^*)}\right)D\left(w,w_{k+1}\right)-\left(\frac{1}{\Lambda_k}-m_{(f,h^*)}\right)D\left(w,w_k\right)+\left(\frac{1}{\Lambda_k}-L\right)D\left(w_{k+1},w_k\right)\\\leq\;&\langle Tx_{k+1},\mu\rangle-\langle Tx,\mu_{k+1}\rangle.\end{aligned}$$

Rearranging the terms, we have, for each $(x,\mu)\in\mathcal{C}_p\times\mathcal{C}_d$, for each $k\in\mathbb{N}$,

$$\begin{aligned}&\mathcal{L}\left(x_{k+1},\mu\right)-\mathcal{L}\left(x,\mu_{k+1}\right)+\left(\frac{1}{\Lambda_k}+m_{(g,l^*)}\right)D\left(w,w_{k+1}\right)-\left(\frac{1}{\Lambda_k}-m_{(f,h^*)}\right)D\left(w,w_k\right)\\&+\left(\frac{1}{\Lambda_k}-L\right)D\left(w_{k+1},w_k\right)+\langle w_{k+1}-w,\Delta_k\rangle\\\leq\;&\langle Tx_{k+1},\mu-\bar{\mu}_k\rangle+\langle T(\bar{x}_k-x),\mu_{k+1}\rangle+\langle Tx,\bar{\mu}_k\rangle-\langle T\bar{x}_k,\mu\rangle\\=\;&\langle T(x_{k+1}-x),\mu-\bar{\mu}_k\rangle+\langle T(\bar{x}_k-x),\mu_{k+1}-\mu\rangle.\end{aligned}$$

Finally, for each $(x,\mu)\in\mathcal{C}_p\times\mathcal{C}_d$, for each $k\in\mathbb{N}$,

$$\begin{aligned}&\mathcal{L}\left(x_{k+1},\mu\right)-\mathcal{L}\left(x,\mu_{k+1}\right)+\left(\frac{1}{\Lambda_k}+m_{(g,l^*)}\right)D\left(w,w_{k+1}\right)-\left(\frac{1}{\Lambda_k}-m_{(f,h^*)}\right)D\left(w,w_k\right)\\&+\left(\frac{1}{\Lambda_k}-L\right)D\left(w_{k+1},w_k\right)+\langle w_{k+1}-w,\Delta_k\rangle\\\leq\;&\langle T(x_{k+1}-x),\mu-\bar{\mu}_k\rangle+\langle T(\bar{x}_k-x),\mu_{k+1}-\mu\rangle.\end{aligned}$$

Now we use the choice of $\bar{x}_k=2x_{k+1}-x_k$ and $\bar{\mu}_k=\mu_k$, to obtain, for each $(x,\mu)\in\mathcal{C}_p\times\mathcal{C}_d$, for each $k\in\mathbb{N}$,

$$\begin{aligned}&\mathcal{L}\left(x_{k+1},\mu\right)-\mathcal{L}\left(x,\mu_{k+1}\right)+\left(\frac{1}{\Lambda_k}+m_{(g,l^*)}\right)D\left(w,w_{k+1}\right)-\left(\frac{1}{\Lambda_k}-m_{(f,h^*)}\right)D\left(w,w_k\right)\\&+\left(\frac{1}{\Lambda_k}-L\right)D\left(w_{k+1},w_k\right)+\langle w_{k+1}-w,\Delta_k\rangle\\\leq\;&\langle T(x_{k+1}-x),\mu-\mu_k\rangle+\langle T(x_{k+1}-x),\mu_{k+1}-\mu\rangle+\langle T(x_{k+1}-x_k),\mu_{k+1}-\mu\rangle\\=\;&\left[\langle T(x_{k+1}-x_k),\mu_{k+1}-\mu_k\rangle+\langle T(x-x_{k+1}),\mu-\mu_{k+1}\rangle-\langle T(x-x_k),\mu-\mu_k\rangle\right].\end{aligned}$$

Equivalently, recalling that $M(w_1,w_2)\overset{\text{def}}{=}\langle T(x_1-x_2),\mu_1-\mu_2\rangle$, we have for each $(x,\mu)\in\mathcal{C}_p\times\mathcal{C}_d$, for each $k\in\mathbb{N}$,

$$\begin{aligned}&\mathcal{L}\left(x_{k+1},\mu\right)-\mathcal{L}\left(x,\mu_{k+1}\right)+\langle w_{k+1}-w,\Delta_k\rangle+\left[\left(\frac{1}{\Lambda_k}+m_{(g,l^*)}\right)D\left(w,w_{k+1}\right)-M\left(w,w_{k+1}\right)\right]\\&-\left[\left(\frac{1}{\Lambda_k}-m_{(f,h^*)}\right)D\left(w,w_k\right)-M\left(w,w_k\right)\right]+\left[\left(\frac{1}{\Lambda_k}-L\right)D\left(w_{k+1},w_k\right)-M\left(w_{k+1},w_k\right)\right]\leq 0.\end{aligned}\tag{5.2.6}$$

Recall that, by (A$_2$), $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ are nondecreasing sequences. Then, for each $k \in \mathbb{N}$, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\frac{1}{\Lambda_{k+1}} D\left(w, w_{k+1}\right) \leq \frac{1}{\Lambda_k} D\left(w, w_{k+1}\right). \tag{5.2.7}$$

Finally, combining (5.2.6) with (5.2.7) and (A$_3$)(5.1.5) applied at the points w_{k+1} and w_k, we get (5.2.1). □

5.2.2 Estimations of the Error Δ_k

Lemma 5.2.2. *Assume (H), (A$_1$), (A$_2$), and (A$_3$) hold and, for each $k \in \mathbb{N}$, denote by $\widehat{w}_{k+1}$ the exact update of the algorithm, i.e.*

$$\widehat{w}_{k+1} = \begin{pmatrix} \widehat{x}_{k+1} \\ \widehat{\mu}_{k+1} \end{pmatrix} = \begin{pmatrix} \left[\nabla\phi_p + \lambda_k \partial g\right]^{-1} \left(\nabla\phi_p\left(x_k\right) - \lambda_k\left(\nabla f\left(x_k\right)\right) - \lambda_k T^* \mu_k\right) \\ \left[\nabla\phi_d + \nu_k \partial l^*\right]^{-1} \left(\nabla\phi_d\left(\mu_k\right) - \nu_k\left(\nabla h^*\left(\mu_k\right)\right) + \nu_k T\left(2\widehat{x}_{k+1} - x_k\right)\right) \end{pmatrix}. \tag{5.2.8}$$

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

$$\left\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1} \right\rangle \geq \frac{1}{\Lambda_k}\left(D\left(\widehat{w}_{k+1}, w_{k+1}\right) + D\left(w_{k+1}, \widehat{w}_{k+1}\right)\right) - 2M\left(\widehat{w}_{k+1}, w_{k+1}\right) \geq 0. \tag{5.2.9}$$

Proof. By design of the algorithm, the following monotone inclusions hold, for each $k \in \mathbb{N}$,

$$\begin{aligned} \nabla\phi_p\left(x_k\right) - \lambda_k\left(\nabla f\left(x_k\right) - T^*\mu_k\right) - \nabla\phi_p\left(\widehat{x}_{k+1}\right) &\in \lambda_k \partial g\left(\widehat{x}_{k+1}\right); \\ \nabla\phi_p\left(x_k\right) - \lambda_k\left(\nabla f\left(x_k\right) + \delta_k^p - T^*\mu_k\right) - \nabla\phi_p\left(x_{k+1}\right) &\in \lambda_k \partial g\left(x_{k+1}\right). \end{aligned} \tag{5.2.10}$$

and similarly for the dual

$$\begin{aligned} \nabla\phi_d\left(\mu_k\right) - \nu_k\left(\nabla h^*\left(\mu_k\right) + T\left(2\widehat{x}_{k+1} - x_k\right)\right) - \nabla\phi_d\left(\widehat{\mu}_{k+1}\right) &\in \nu_k \partial l^*\left(\widehat{\mu}_{k+1}\right); \\ \nabla\phi_d\left(\mu_k\right) - \nu_k\left(\nabla h^*\left(\mu_k\right) + \delta_k^d + T\left(2x_{k+1} - x_k\right)\right) - \nabla\phi_d\left(\mu_{k+1}\right) &\in \nu_k \partial l^*\left(\mu_{k+1}\right). \end{aligned} \tag{5.2.11}$$

By monotonicity of the operators ∂l^* and ∂g combined with (5.2.11) and (5.2.10), we then have, for each $k \in \mathbb{N}$,

$$\begin{aligned} \left\langle \widehat{x}_{k+1} - x_{k+1}, \delta_k^p \lambda_k - \nabla\phi_p\left(\widehat{x}_{k+1}\right) + \nabla\phi_p\left(x_{k+1}\right) \right\rangle &\geq 0; \\ \left\langle \widehat{\mu}_{k+1} - \mu_{k+1}, \delta_k^d \nu_k - \nabla\phi_d\left(\widehat{\mu}_{k+1}\right) + \nabla\phi_d\left(\mu_{k+1}\right) + 2\nu_k T\left(\widehat{x}_{k+1} - x_{k+1}\right) \right\rangle &\geq 0. \end{aligned} \tag{5.2.12}$$

We can rewrite the above using Definition 2.1.7 to have, for each $k \in \mathbb{N}$,

$$\begin{aligned} \left\langle \widehat{x}_{k+1} - x_{k+1}, \delta_k^p \right\rangle &\geq \frac{1}{\lambda_k}\left(D_p\left(\widehat{x}_{k+1}, x_{k+1}\right) + D_p\left(x_{k+1}, \widehat{x}_{k+1}\right)\right); \\ \left\langle \widehat{\mu}_{k+1} - \mu_{k+1}, \delta_k^d \right\rangle &\geq \frac{1}{\nu_k}\left(D_d\left(\widehat{\mu}_{k+1}, \mu_{k+1}\right) + D_d\left(\mu_{k+1}, \widehat{\mu}_{k+1}\right)\right) - 2\left\langle T\left(\widehat{x}_{k+1} - x_{k+1}\right), \widehat{\mu}_{k+1} - \mu_{k+1} \right\rangle. \end{aligned} \tag{5.2.13}$$

Adding the above inequalities together gives, for each $k \in \mathbb{N}$,

$$\left\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1} \right\rangle \geq \frac{1}{\Lambda_k}\left(D\left(\widehat{w}_{k+1}, w_{k+1}\right) + D\left(w_{k+1}, \widehat{w}_{k+1}\right)\right) - 2M\left(\widehat{w}_{k+1}, w_{k+1}\right). \tag{5.2.14}$$

Using (A$_3$) and (5.1.7), and the fact that M is symmetric w.r.t. its arguments, for each $k \in \mathbb{N}$,

$$\frac{1}{\Lambda_k}\left(D\left(\widehat{w}_{k+1}, w_{k+1}\right) + D\left(w_{k+1}, \widehat{w}_{k+1}\right)\right) - 2M\left(\widehat{w}_{k+1}, w_{k+1}\right) \geq 0.$$

□

Lemma 5.2.3. *Assume (H), (A$_1$), (A$_2$), (A$_3$), that the entropies ϕ_p and ϕ_d are strongly convex with respect to $\|\cdot\|_p^2$ and $\|\cdot\|_d^2$ with modulus m_p and m_d, respectively, and that the step size limits λ_∞ and ν_∞ satisfy*

$$\nu_\infty \lambda_\infty < \frac{m_p m_d}{\|T\|_{p\to d^*}^2}.$$

One can choose $a > 0$ so that, for each $k \in \mathbb{N}$,

$$\frac{m_p}{\lambda_k} - \frac{\|T\|^2_{p\to d^*}}{a} > 0 \text{ and } \frac{m_d}{\nu_k} - a > 0$$

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

$$\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle \leq \left(\frac{m_p}{\lambda_k} - \frac{\|T\|^2_{p\to d^*}}{a}\right)^{-1} \left\|\delta_k^p\right\|^2_{p^*} + \left(\frac{m_d}{\nu_k} - a\right)^{-1} \left\|\delta_k^d\right\|^2_{d^*}.$$

Proof. It follows from the strong convexity of ϕ_p and ϕ_d that, for each $k \in \mathbb{N}$,

$$\begin{aligned}\frac{1}{\Lambda_k}\left(D\left(w_{k+1}, \widehat{w}_{k+1}\right) + D\left(\widehat{w}_{k+1}, w_{k+1}\right)\right) &= \frac{1}{\Lambda_k}\left\langle \nabla\phi\left(w_{k+1}\right) - \nabla\phi\left(\widehat{w}_{k+1}\right), w_{k+1} - \widehat{w}_{k+1}\right\rangle \\ &\geq \frac{m_p}{\lambda_k}\left\|\widehat{x}_{k+1} - x_{k+1}\right\|_p^2 + \frac{m_d}{\nu_k}\left\|\widehat{\mu}_{k+1} - \mu_{k+1}\right\|_d^2.\end{aligned} \tag{5.2.15}$$

Substituting this result into Lemma 5.2.2 (5.2.9) and applying Young's inequality with $a > 0$ we get, for each $k \in \mathbb{N}$,

$$\begin{aligned}\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle &\geq \frac{m_p}{\lambda_k}\left\|\widehat{x}_{k+1} - x_{k+1}\right\|_p^2 + \frac{m_d}{\nu_k}\left\|\widehat{\mu}_{k+1} - \mu_{k+1}\right\|_d^2 - 2M\left(\widehat{w}_{k+1}, w_{k+1}\right) \\ &= \frac{m_p}{\lambda_k}\left\|\widehat{x}_{k+1} - x_{k+1}\right\|_p^2 + \frac{m_d}{\nu_k}\left\|\widehat{\mu}_{k+1} - \mu_{k+1}\right\|_d^2 - 2\left(\left\langle T\left(\widehat{x}_{k+1} - x_{k+1}\right), \widehat{\mu}_{k+1} - \mu_{k+1}\right\rangle\right) \\ &\geq \frac{m_p}{\lambda_k}\left\|\widehat{x}_{k+1} - x_{k+1}\right\|_p^2 + \frac{m_d}{\nu_k}\left\|\widehat{\mu}_{k+1} - \mu_{k+1}\right\|_d^2 - \frac{\|T\|^2_{p\to d^*}}{a}\left\|\widehat{x}_{k+1} - x_{k+1}\right\|_p^2 - a\left\|\widehat{\mu}_{k+1} - \mu_{k+1}\right\|_d^2 \\ &= \left(\frac{m_p}{\lambda_k} - \frac{\|T\|^2_{p\to d^*}}{a}\right)\left\|\widehat{x}_{k+1} - x_{k+1}\right\|_p^2 + \left(\frac{m_d}{\nu_k} - a\right)\left\|\widehat{\mu}_{k+1} - \mu_{k+1}\right\|_d^2.\end{aligned} \tag{5.2.16}$$

Then, since the step size sequences $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ are bounded and nondecreasing by $(\mathbf{A}_2)$, and furthemore chosen small enough to satisfy

$$\nu_\infty\lambda_\infty < \frac{m_p m_d}{\|T\|^2_{p\to d^*}},$$

one can choose $a > 0$ so that

$$\frac{m_p}{\lambda_\infty} - \frac{\|T\|^2_{p\to d^*}}{a} > 0 \text{ and } \frac{m_d}{\nu_\infty} - a > 0$$

and, by extension under $(\mathbf{A}_2)$, for each $k \in \mathbb{N}$,

$$\frac{m_p}{\lambda_k} - \frac{\|T\|^2_{p\to d^*}}{a} > 0 \text{ and } \frac{m_d}{\nu_k} - a > 0$$

Finally, we apply Young's inequality twice to the following to find, for each $k \in \mathbb{N}$,

$$\begin{aligned}\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle &= \left\langle \delta_k^p, \widehat{x}_{k+1} - x_{k+1}\right\rangle + \left\langle \delta_k^d, \widehat{\mu}_{k+1} - \mu_{k+1}\right\rangle \\ &\leq \frac{1}{2}\left(\frac{m_p}{\lambda_k} - \frac{\|T\|^2_{p\to d^*}}{a}\right)^{-1}\left\|\delta_k^p\right\|^2_{p^*} + \frac{1}{2}\left(\frac{m_p}{\lambda_k} - \frac{\|T\|^2_{p\to d^*}}{a}\right)\left\|\widehat{x}_{k+1} - x_{k+1}\right\|_p^2 \\ &\quad + \frac{1}{2}\left(\frac{m_d}{\nu_k} - a\right)^{-1}\left\|\delta_k^d\right\|^2_{d^*} + \frac{1}{2}\left(\frac{m_d}{\nu_k} - a\right)\left\|\widehat{\mu}_{k+1} - \mu_{k+1}\right\|_d^2 \\ &\leq \frac{1}{2}\left(\frac{m_p}{\lambda_k} - \frac{\|T\|^2_{p\to d^*}}{a}\right)^{-1}\left\|\delta_k^p\right\|^2_{p^*} + \frac{1}{2}\left(\frac{m_d}{\nu_k} - a\right)^{-1}\left\|\delta_k^d\right\|^2_{d^*} \\ &\quad + \frac{1}{2}\left\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\right\rangle\end{aligned}$$

and the desired claim follows. □

Remark 5.2.4. In Lemma 5.2.3, one can instead choose to use $\|T^*\|^2_{d\to p^*}$ to have, for each $k \in \mathbb{N}$,

$$\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle \leq \left(\frac{m_p}{\lambda_k} - \frac{1}{a}\right)\left\|\delta_k^p\right\|^2_{p^*} + \left(\frac{m_d}{\nu_k} - a\left\|T^*\right\|^2_{d\to p^*}\right)\left\|\delta_k^d\right\|^2_{d^*}$$

which could be useful if there is asymmtery in the size of m_p and m_d.

In the event that just ϕ_p is strongly convex with respect to $\|\cdot\|_p^2$ but the analog doesn't hold for ϕ_d, we can make the following argument. Take (5.2.13) from Lemma 5.2.2 and use strong convexity,

$$\left\langle \widehat{x}_{k+1} - x_{k+1}, \delta_k^p\right\rangle \geq \frac{1}{\lambda_k}\left(D_p\left(\widehat{x}_{k+1}, x_{k+1}\right) + D_p\left(x_{k+1}, \widehat{x}_{k+1}\right)\right) \geq \frac{m_p}{\lambda_k}\left\|\widehat{x}_{k+1} - x_{k+1}\right\|_p^2$$

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

$$\left\langle \delta_k^p, \widehat{x}_{k+1} - x_{k+1}\right\rangle \leq \frac{\lambda_k}{m_p}\left\|\delta_k^p\right\|^2_{p^*}$$

without the restriction on λ_∞ and ν_∞ imposed in Lemma 5.2.3 because we no longer to need control the term $2M\left(\widehat{w}_{k+1}, w_{k+1}\right)$. This term, $2M\left(\widehat{w}k+1, w_{k+1}\right)$, is a result of the way we have defined $\widehat{\mu}_{k+1}$ to depend on $\widehat{x}_{k+1}$, which is necessary to keep $\widehat{w}_{k+1}$ deterministic conditioned on the filtration $\mathcal{S}_k$. Thus, if only one of the entropies can be chosen to be strongly convex, one is inclined to formulate the problem in such a way that the primal problem has the strongly convex entropy, and to deal with the dual problem using **(A$_4$)(I)** or **(A$_4$)(II)**.

Lemma 5.2.5. *Under* **(H)**, **(A$_1$)**, **(A$_2$)**, **(A$_3$)**, *and* **(A$_4$)**, *the following sequences satisfy, for any fixed* $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\left(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle \mid \mathcal{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1\left(\mathfrak{S}\right) \quad \textit{and} \quad \left(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle\right]\right)_{k\in\mathbb{N}} \in \ell_+^1.$$

Proof. The assumption **(A$_4$)** has three cases with the first, **(A$_4$)(I)**, corresponding to the deterministic setting, i.e., there is nothing to show. For the following two cases, we note that, by Lemma 5.2.2, for each $k \in \mathbb{N}$, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle \mid \mathcal{S}_k\right] = \mathbb{E}\left[\langle \Delta_k, w - \widehat{w}_{k+1}\rangle + \langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle \mid \mathcal{S}_k\right] = \mathbb{E}\left[\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle \mid \mathcal{S}_k\right] \geq 0 \tag{5.2.17}$$

since, due to **(A$_4$)**, Δ_k is unbiased conditioned on the filtration $\mathcal{S}_k$. By the law of total expectation applied to the above, it follows that, for each $k \in \mathbb{N}$, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle\right] = \mathbb{E}\left[\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle\right] \geq 0$$

and thus the following sequences satisfy, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\left(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle \mid \mathcal{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+\left(\mathfrak{S}\right) \quad \text{and} \quad \left(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle\right]\right)_{k\in\mathbb{N}} \in \ell_+.$$

For the second case, **(A$_4$)(II)**, recall that, for each $k \in \mathbb{N}$,

$$\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle \overset{\text{def}}{=} \left\langle \delta_k^p, \widehat{x}_{k+1} - x_{k+1}\right\rangle + \left\langle \delta_k^d, \widehat{\mu}_{k+1} - \mu_{k+1}\right\rangle.$$

By **(A$_4$)(II)**, the sets $\mathcal{C}_p$ and $\mathcal{C}_d$ are bounded and thus have finite diameters, $\mathrm{diam}_{\mathcal{C}_p}$ and $\mathrm{diam}_{\mathcal{C}_d}$ respectively. Then, for each $k \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\left[\left\langle \delta_k^p, \widehat{x}_{k+1} - x_{k+1}\right\rangle \mid \mathcal{S}_k\right] &\leq \mathbb{E}\left[\left\|\delta_k^p\right\|_{p^*}\left\|\widehat{x}_{k+1} - x_{k+1}\right\|_p \mid \mathcal{S}_k\right] \leq \mathrm{diam}_{\mathcal{C}_p}\mathbb{E}\left[\left\|\delta_k^p\right\|_{p^*} \mid \mathcal{S}_k\right]; \\
\mathbb{E}\left[\left\langle \delta_k^d, \widehat{\mu}_{k+1} - \mu_{k+1}\right\rangle \mid \mathcal{S}_k\right] &\leq \mathbb{E}\left[\left\|\delta_k^d\right\|_{d^*}\left\|\widehat{\mu}_{k+1} - \mu_{k+1}\right\|_d \mid \mathcal{S}_k\right] \leq \mathrm{diam}_{\mathcal{C}_d}\mathbb{E}\left[\left\|\delta_k^d\right\|_{d^*} \mid \mathcal{S}_k\right].
\end{aligned}$$

Since $\left(\mathbb{E}\left[\left\|\delta_k^p\right\|_{p^*} \mid \mathcal{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1\left(\mathfrak{S}\right)$ and $\left(\mathbb{E}\left[\left\|\delta_k^d\right\|_{d^*} \mid \mathcal{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1\left(\mathfrak{S}\right)$ by **(A$_4$)(II)**, and noting (5.2.17), it holds that, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\left(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle \mid \mathcal{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1\left(\mathfrak{S}\right).$$

Using the same argument with the law of total expectation together with the fact that $\left(\mathbb{E}\left[\left\|\delta_k^p\right\|_{p^*}\right]\right)_{k\in\mathbb{N}} \in \ell_+^1$ and $\left(\mathbb{E}\left[\left\|\delta_k^d\right\|_{d^*}\right]\right)_{k\in\mathbb{N}}$ by **(A$_4$)(II)**, it then follows that, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\left(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle\right]\right)_{k\in\mathbb{N}} \in \ell_+^1.$$

Finally, in the case of **(A$_4$)(III)**, we assume that the entropies ϕ_p and ϕ_d are strongly convex with respect to $\|\cdot\|_p$ and $\|\cdot\|_d$ respectively. Using Lemma 5.2.3 and taking expectation conditioned on $\mathcal{S}_k$, we have, for each $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle \mid \mathcal{S}_k\right] &\leq \left(\frac{m_p}{\lambda_k} - \frac{\|T\|_{p\to d^*}^2}{a}\right)^{-1} \mathbb{E}\left[\left\|\delta_k^p\right\|_{p^*}^2 \mid \mathcal{S}_k\right] + \left(\frac{m_d}{\nu_k} - a\right)^{-1} \mathbb{E}\left[\left\|\delta_k^d\right\|_{d^*}^2 \mid \mathcal{S}_k\right] \\ &\leq \left(\frac{m_p}{\lambda_\infty} - \frac{\|T\|_{p\to d^*}^2}{a}\right)^{-1} \mathbb{E}\left[\left\|\delta_k^p\right\|_{p^*}^2 \mid \mathcal{S}_k\right] + \left(\frac{m_d}{\nu_\infty} - a\right)^{-1} \mathbb{E}\left[\left\|\delta_k^d\right\|_{d^*}^2 \mid \mathcal{S}_k\right]\end{aligned}$$

and thus by the summability assumption of **(A$_4$)(III)**, we have

$$\left(\mathbb{E}\left[\left\|\delta_k^p\right\|_{p^*}^2 \mid \mathcal{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S}) \quad \text{and} \quad \left(\mathbb{E}\left[\left\|\delta_k^d\right\|_{d^*}^2 \mid \mathcal{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S})$$

and so, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\left(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle \mid \mathcal{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S}).$$

Similarly, taking Lemma 5.2.3 with total expectation and the summability assumption of **(A$_4$)(III)** yields, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\left(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle\right]\right)_{k\in\mathbb{N}} \in \ell_+^1.$$

□

5.3 Convergence Analysis

5.3.1 Ergodic Convergence

Define, for each $k \in \mathbb{N}$, the ergodic iterates $\bar{x}_k \overset{\text{def}}{=} \frac{1}{k}\sum_{i=1}^{k} x_i$ and $\bar{\mu}_k \overset{\text{def}}{=} \frac{1}{k}\sum_{i=1}^{k} \mu_i$.

Theorem 5.3.1. *Let* **(H)**, **(A$_1$)**, **(A$_2$)**, **(A$_3$)**, *and* **(A$_4$)** *hold. Then we have the following convergence rate: for each $k \in \mathbb{N}$, for every $(x, \mu) \in \mathcal{C}_p \times \mathcal{C}_d$,*

$$\mathbb{E}\left[\mathcal{L}(\bar{x}_k, \mu) - \mathcal{L}(x, \bar{\mu}_k)\right] \leq \frac{\frac{1}{\Lambda_0} D(w, w_0) - M(w, w_0) + \sum_{i=0}^{+\infty} \mathbb{E}\left[\langle \Delta_i, w - w_{i+1}\rangle\right]}{k}. \tag{5.3.1}$$

In particular, every almost sure weak sequential cluster point of $(\bar{w}_k)_{k\in\mathbb{N}}$ is optimal in mean; if $\bar{w}_{k_j} \rightharpoonup w_\infty$ almost surely, then $\mathbb{E}(w_\infty)$ is a primal-dual optimal pair.

Proof. Beginning with Lemma 5.2.1, we have for every $(x, \mu) \in \mathcal{C}_p \times \mathcal{C}_d$, for each $k \in \mathbb{N}$,

$$\begin{aligned}&\mathcal{L}(x_{k+1}, \mu) - \mathcal{L}(x, \mu_{k+1}) + \left[\frac{1}{\Lambda_{k+1}} D(w, w_{k+1}) - M(w, w_{k+1})\right] \\ &+ \varepsilon d(w_{k+1}, w_k) \leq \left[\frac{1}{\Lambda_k} D(w, w_k) - M(w, w_k)\right] + \langle \Delta_k, w - w_{k+1}\rangle.\end{aligned} \tag{5.3.2}$$

Taking the the total expectation and summing up from 0 to $k-1$, discarding positive terms on the left hand side, we have, for every $(x, \mu) \in \mathcal{C}_p \times \mathcal{C}_d$, for each $k \in \mathbb{N}$,

$$\begin{aligned}\sum_{i=0}^{k-1} \mathbb{E}\left[\mathcal{L}(x_{i+1}, \mu) - \mathcal{L}(x, \mu_{i+1})\right] &\leq \frac{1}{\Lambda_0} D(w, w_0) - M(w, w_0) + \sum_{i=0}^{k-1} \mathbb{E}\left[\langle \Delta_i, w - w_{i+1}\rangle\right] \\ &\leq \frac{1}{\Lambda_0} D(w, w_0) - M(w, w_0) + \sum_{i=0}^{\infty} \mathbb{E}\left[\langle \Delta_i, w - w_{i+1}\rangle\right].\end{aligned} \tag{5.3.3}$$

Using Jensen's inequality with the convex-concave function $\mathcal{L}$, we have (5.3.1), noting that

$$\sum_{i=0}^{\infty} \mathbb{E}\left[\langle \Delta_i, w - w_{i+1} \rangle\right] < +\infty$$

by $(\mathbf{A}_4)$ and Lemma 5.2.5.

Now let $(\bar{x}_{k_j}, \bar{\mu}_{k_j}) \rightharpoonup (x_\infty, \mu_\infty)$ almost surely. Then, for every $(x, \mu) \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\begin{aligned} \mathcal{L}\left(\mathbb{E}\left(x_\infty\right), \mu\right) - \mathcal{L}\left(x, \mathbb{E}\left(\mu_\infty\right)\right) &\leq \mathbb{E}\left[\mathcal{L}\left(x_\infty, \mu\right) - \mathcal{L}\left(x, \mu_\infty\right)\right] \\ &\leq \mathbb{E}\left[\liminf_{j \to \infty} \left[\mathcal{L}\left(\bar{x}_{k_j}, \mu\right) - \mathcal{L}\left(x, \bar{\mu}_{k_j}\right)\right]\right] \\ &\leq \liminf_{j \to \infty} \mathbb{E}\left[\mathcal{L}\left(\bar{x}_{k_j}, \mu\right) - \mathcal{L}\left(x, \bar{\mu}_{k_j}\right)\right] \\ &\leq 0, \end{aligned} \tag{5.3.4}$$

where we used Jensen's inequality, weak lower semicontinuity of $\mathcal{L}$, Fatou's Lemma and (5.3.1) with $(\mathbf{A}_4)$ and Lemma 5.2.5. So $\left(\mathbb{E}\left(x_\infty\right), \mathbb{E}\left(\mu_\infty\right)\right)$ is a primal-dual optimal pair for $\mathcal{L}$. □

5.3.2 Asymptotic Regularity

Theorem 5.3.2. *Let* $(\mathbf{H})$, $(\mathbf{A}_1)$, $(\mathbf{A}_3)(\mathbf{II})$, $(\mathbf{A}_4)$ *and* $(\mathbf{A}_5)$ *hold. Then the primal-dual sequence* $(x_k, \mu_k)_{k \in \mathbb{N}}$ *is almost surely asymptotically regular, meaning that* $x_{k+1} - x_k \to 0$ *and* $\mu_{k+1} - \mu_k \to 0$ *almost surely.*

Proof. Use again Lemma 5.2.1 with w equal to a primal-dual optimal pair $w^\star$ and take the total expectation: for each $k \in \mathbb{N}$,

$$\begin{aligned} &\mathbb{E}\left[\mathcal{L}\left(x_{k+1}, \mu^\star\right) - \mathcal{L}\left(x^\star, \mu_{k+1}\right)\right] + \mathbb{E}\left[\frac{1}{\Lambda_{k+1}} D\left(w^\star, w_{k+1}\right) - M\left(w^\star, w_{k+1}\right)\right] \\ &+ \varepsilon \mathbb{E}\left[d\left(w_{k+1}, w_k\right)\right] \leq \mathbb{E}\left[\frac{1}{\Lambda_k} D\left(w^\star, w_k\right) - M\left(w^\star, w_k\right)\right] + \mathbb{E}\left[\langle \Delta_k, w^\star - w_{k+1} \rangle\right]. \end{aligned} \tag{5.3.5}$$

By the definition of primal-dual optimal pair, we know that, for each $k \in \mathbb{N}$,

$$\mathcal{L}\left(x_k, \mu^\star\right) - \mathcal{L}\left(x^\star, \mu_k\right) \geq 0$$

and so, from Lemma 2.2.1 with $(\mathbf{A}_4)$, Lemma 5.2.5, and $(\mathbf{A}_3)(\mathbf{II})$,

$$\mathbb{E}\left[d\left(w_{k+1}, w_k\right)\right] \in \ell^1_+.$$

So, by Lemma 2.3.4, $d\left(w_{k+1}, w_k\right) \to 0$ almost surely. In view of $(\mathbf{A}_5)$, we get that, almost surely,

$$w_{k+1} - w_k \to 0, \tag{5.3.6}$$

i.e., the primal-dual sequence $(w_k)_{k \in \mathbb{N}}$ is almost surely asymptotically regular. □

5.3.3 Pointwise Convergence

The main result of this section is the pointwise convergence of the primal-dual sequence $(x_k, \mu_k)_{k \in \mathbb{N}}$ to a primal-dual optimal pair. Note that in the case of finite-dimensional spaces $\mathcal{X}_p$ and $\mathcal{X}_d$, the assumptions $(\mathbf{A}_8)$, $(\mathbf{A}_9)$, $(\mathbf{A}_{10})$ can be removed as they are trivially satisfied. We will also impose the following conditions, which are only necessary for this particular section in the stochastic case and can be dropped for the deterministic case or the other sections.

$(\mathbf{PW}_1)$ The spaces $\mathcal{X}_p$ and $\mathcal{X}_d$ are separable.

$(\mathbf{PW}_2)$ The Bregman divergence D satisfies the following property: if there exists $\tilde{\Omega} \in \mathcal{F}$ such that, for any $\omega \in \tilde{\Omega}$, for any $w^\star \in \boldsymbol{S}_{\mathscr{P}} \times \boldsymbol{S}_{\mathscr{D}}$ with a sequence $(s_n)_{n \in \mathbb{N}}$ in $\boldsymbol{S}_{\mathscr{P}} \times \boldsymbol{S}_{\mathscr{D}}$ such that $s_n \to w^\star$ and satisfying

$$\lim_{k \to \infty} \Lambda_k^{-1} D\left(s_n, w_k\left(\omega\right)\right) - M\left(s_n, w_k\left(\omega\right)\right) = r_{s_n}\left(\omega\right) \in [0, +\infty[,$$

then there exists a $[0,+\infty[$-valued random variable $r_{w^\star}$ such that, for any $\omega \in \tilde{\Omega}$,

$$\lim_{k\to\infty} \Lambda_k^{-1} D\left(w^\star, w_k\left(\omega\right)\right) - M\left(w^\star, w_k\left(\omega\right)\right) = r_{w^\star}\left(\omega\right).$$

Proposition 5.3.3. *Let* **(H)**, **(A$_1$)**, **(A$_2$)**, **(A$_3$)(II)**, **(A$_4$)**, **(A$_5$)**, **(A$_6$)**, **(A$_7$)**, *and* **(A$_8$)** *hold. Then* $\left(\left(x_k, \mu_k\right)\right)_{k\in\mathbb{N}}$ *is almost surely bounded and, recalling the notation of* (5.1.4) *and* (5.1.3), $\mathfrak{W}\left[\left(w_k\right)_{k\in\mathbb{N}}\right] \subset \boldsymbol{S}_{\mathscr{P}} \times \boldsymbol{S}_{\mathscr{D}}$ $(\mathbb{P}$*-a.s.*$)$.

Proof. Evaluating Lemma 5.2.1 at a primal-dual optimal pair $w = w^\star$ and taking expectation conditioned on the filtration $\mathcal{S}_k$, we get, for each $k \in \mathbb{N}$,

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{L}\left(x_{k+1}, \bar{\mu}\right) - \mathcal{L}\left(\bar{x}, \mu_{k+1}\right) \mid \mathcal{S}_k\right] + \mathbb{E}\left[\frac{1}{\Lambda_{k+1}} D\left(w^\star, w_{k+1}\right) - M\left(w^\star, w_{k+1}\right) \mid \mathcal{S}_k\right] \\
&+ \varepsilon\mathbb{E}\left[d\left(w_{k+1}, w_k\right) \mid \mathcal{S}_k\right] \leq \left[\frac{1}{\Lambda_k} D\left(w^\star, w_k\right) - M\left(w^\star, w_k\right)\right] + \mathbb{E}\left[\left\langle \Delta_k, w^\star - w_{k+1}\right\rangle \mid \mathcal{S}_k\right].
\end{aligned}$$

Then, by **(A$_4$)**, Lemma 5.2.5, and Lemma 2.3.2, $\left(\Lambda_k^{-1} D\left(w^\star, w_k\right) - M\left(w^\star, w_k\right)\right)_{k\in\mathbb{N}}$ is almost surely convergent to some $r \in [0,+\infty[$. In particular, from **(A$_3$)** and (5.1.7), both $\left(D\left(w^\star, w_k\right)\right)_{k\in\mathbb{N}}$ and $\left(d\left(w^\star, w_k\right)\right)_{k\in\mathbb{N}}$ are almost surely bounded and the coercivity condition **(A$_6$)** entails that the sequence $\left(w_k\right)_{k\in\mathbb{N}}$ is almost surely bounded in $\operatorname{int}(\operatorname{dom}(\phi))$. Let $w_\infty = (x_\infty, \mu_\infty)$ be an almost sure weak sequential cluster point of $\left(w_k\right)_{k\in\mathbb{N}}$, i.e., there is a subsequence $\left(w_{k_i}\right)_{i\in\mathbb{N}}$ such that $w_{k_i} \rightharpoonup w_\infty$ almost surely. The updates of Algorithm 10 are equivalent to the following monotone inclusions,

$$\begin{aligned}
&\begin{pmatrix} \frac{\nabla\phi_p(x_{k_i}) - \nabla\phi_p(x_{k_i+1})}{\lambda_k} + \left(\nabla f(x_{k_i+1}) - \nabla f(x_{k_i}) - \delta_{k_i}^p\right) + T^*(\mu_{k_i+1} - \mu_{k_i}) \\ \frac{\nabla\phi_d(\mu_{k_i}) - \nabla\phi_d(\mu_{k_i+1})}{\nu_k} + \left(\nabla h^*(\mu_{k_i+1}) - \nabla h^*(\mu_{k_i}) - \delta_{k_i}^d\right) + T(x_{k_i+1} - x_{k_i}) \end{pmatrix} \\
\in\ &\begin{pmatrix} \partial g + \nabla f & 0 \\ 0 & \partial l^* + \nabla h^* \end{pmatrix} \begin{pmatrix} x_{k_i+1} \\ \mu_{k_i+1} \end{pmatrix} + \begin{pmatrix} 0 & T^* \\ -T & 0 \end{pmatrix} \begin{pmatrix} x_{k_i+1} \\ \mu_{k_i+1} \end{pmatrix}.
\end{aligned}$$

The operator on the right hand side is maximally monotone (hence weak-strong sequentially closed) by **(A$_7$)**. Recall that, by **(A$_4$)** and Remark 5.1.4, $\left(\delta_k^p\right)_{k\in\mathbb{N}}$ and $\left(\delta_k^d\right)_{k\in\mathbb{N}}$ converge to zero almost surely. From Theorem 5.3.2 and the fact that $w_{k_i} \rightharpoonup w_\infty$, we have also that $\left(\left(x_{k_i+1}, \mu_{k_i+1}\right)\right)_{i\in\mathbb{N}}$ converges weakly to (x_∞, μ_∞) almost surely. In addition, by **(H)**, T is linear (and bounded) which, combined with Theorem 5.3.2, yields

$$T(x_{k_i+1} - x_{k_i}) \to 0 \ \text{ and } \ T^*(\mu_{k_i+1} - \mu_{k_i}) \to 0$$

almost surely. From **(A$_8$)** combined with Theorem 5.3.2, we deduce that, almost surely,

$$\begin{aligned}
\nabla\phi_p(x_{k_i+1}) - \nabla\phi_p(x_{k_i}) \to 0 \ \text{ and } \ \nabla f(x_{k_i+1}) - \nabla f(x_{k_i}) \to 0 \\
\nabla\phi_d(\mu_{k_i+1}) - \nabla\phi_d(\mu_{k_i}) \to 0 \ \text{ and } \ \nabla h^*(\mu_{k_i+1}) - \nabla h^*(\mu_{k_i}) \to 0.
\end{aligned}$$

Altogether, recalling that, by **(A$_2$)**, both $\left(\lambda_k\right)_{k\in\mathbb{N}}$ and $\left(\nu_k\right)_{k\in\mathbb{N}}$ are bounded, we conclude that, almost surely,

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \partial g + \nabla f & T^* \\ -T & \partial l^* + \nabla h^* \end{pmatrix} \begin{pmatrix} x_\infty \\ \mu_\infty \end{pmatrix},$$

whence it follows that each weak sequential cluster point of $\left(w_k\right)_{k\in\mathbb{N}}$ is a primal-dual optimal pair almost surely. □

Proposition 5.3.4. *Let* **(H)**, **(A$_1$)**, **(A$_2$)**, **(A$_3$)(II)**, **(A$_4$)**, **(A$_5$)**, **(A$_6$)**, **(A$_7$)**, **(A$_8$)**, **(A$_9$)**, *and* **(A$_{10}$)** *hold as well as* **(PW$_1$)** *and* **(PW$_2$)**. *Then, there exists* $\tilde{\Omega} \in \mathcal{F}$ *such that* $\mathbb{P}\left(\tilde{\Omega}\right) = 1$ *and, for every* $\omega \in \tilde{\Omega}$, *for every* $w^\star \in \boldsymbol{S}_{\mathscr{P}} \times \boldsymbol{S}_{\mathscr{D}}$, *the sequence* $\left(D\left(w^\star, w_k\left(\omega\right)\right) - M\left(w^\star, w_k\left(\omega\right)\right)\right)_{k\in\mathbb{N}}$ *converges with limit in* $[0,+\infty[$.

Proof. By **(PW$_1$)**, there exists a countable set S such that $\bar{S} = \boldsymbol{S}_{\mathscr{P}} \times \boldsymbol{S}_{\mathscr{D}}$. Once again, as in the proof of Proposition 5.3.3, for every $w^\star \in \boldsymbol{S}_{\mathscr{P}} \times \boldsymbol{S}_{\mathscr{D}}$ there exist $\Omega_{w^\star} \in \mathcal{F}$ such that $\mathbb{P}\left(\Omega_{w^\star}\right) = 1$ and, for every $\omega \in \Omega_{w^\star}$, it holds

$$\lim_{k\to\infty} \Lambda_k^{-1} D\left(w^\star, w_k\left(\omega\right)\right) - M\left(w^\star, w_k\left(\omega\right)\right) = r^\star\left(\omega\right) \in [0,+\infty[.$$

Let $\tilde{\Omega} = \bigcap_{s\in S} \Omega_s$ and notice that $\mathbb{P}\left(\tilde{\Omega}\right) = 1$ since, by countability of S we have,

$$\mathbb{P}\left(\tilde{\Omega}\right) = 1 - \mathbb{P}\left(\tilde{\Omega}^c\right) = 1 - \mathbb{P}\left(\bigcup_{s\in S} \Omega_s^c\right) \geq 1 - \sum_{s\in S} \mathbb{P}\left(\Omega_s^c\right) = 1.$$

Fix a particular $w^\star \in \boldsymbol{S}_{\mathscr{P}} \times \boldsymbol{S}_{\mathscr{D}}$; since $\bar{S} = \boldsymbol{S}_{\mathscr{P}} \times \boldsymbol{S}_{\mathscr{D}}$, there exists a sequence $(s_n)_{n\in\mathbb{N}}$ in S such that $s_n \to w^\star$. At the same time, for each $n \in \mathbb{N}$, there exists r_n, a $[0, +\infty[$-valued random variable such that, for each $\omega \in \tilde{\Omega}$,

$$\lim_{k\to\infty} \Lambda_k^{-1} D\left(s_n, w_k\left(\omega\right)\right) - M\left(s_n, w_k\left(\omega\right)\right) = r_n\left(\omega\right) \in [0, +\infty[.$$

Applying now $(\mathbf{PW}_2)$, we find that, for any $\omega \in \tilde{\Omega}$, for any $w^\star \in \boldsymbol{S}_{\mathscr{P}} \times \boldsymbol{S}_{\mathscr{D}}$,

$$\lim_{k\to\infty} \Lambda_k^{-1} D\left(w^\star, w_k\left(\omega\right)\right) - M\left(w^\star, w_k\left(\omega\right)\right) = r_{w^\star}\left(\omega\right) \in [0, \infty[.$$

□

Theorem 5.3.5. *Let* $(\mathbf{H})$, $(\mathbf{A}_1)$, $(\mathbf{A}_2)$, $(\mathbf{A}_3)$**(II)**, $(\mathbf{A}_4)$, $(\mathbf{A}_5)$, $(\mathbf{A}_6)$, $(\mathbf{A}_7)$, $(\mathbf{A}_8)$, $(\mathbf{A}_9)$, *and* $(\mathbf{A}_{10})$ *hold as well as* $(\mathbf{PW}_1)$ *and* $(\mathbf{PW}_2)$. *Then, there exists* $\bar{w}$, *a* $\boldsymbol{S}_{\mathscr{P}} \times \boldsymbol{S}_{\mathscr{D}}$*-valued random variable, such that* $(w_k)_{k\in\mathbb{N}} \rightharpoonup \bar{w}$ $(\mathbb{P}$*-a.s.*$)$.

Proof. To show global convergence, we use a reasoning similar to Opial's lemma (see [89], [37]). We recall the notation of (5.1.4) for the set of weak cluster points of a sequence. By the assumptions and Proposition 5.3.3, there exists $\Omega' \in \mathcal{F}$ with $\mathbb{P}\left(\Omega'\right) = 1$ such that, for any $\omega \in \Omega'$, the following holds

$$\mathfrak{W}\left[\left(w_k\left(\omega\right)\right)\right] \subset \boldsymbol{S}_{\mathscr{P}} \times \boldsymbol{S}_{\mathscr{D}}$$

and the sequence $\left(w_k\left(\omega\right)\right)_{k\in\mathbb{N}}$ is bounded. Furthermore, by Proposition 5.3.4, there exists $\Omega'' \in \mathcal{F}$ with $\mathbb{P}\left(\Omega''\right) = 1$ such that, for any $\omega \in \Omega''$, for any $w^\star \in \boldsymbol{S}_{\mathscr{P}} \times \boldsymbol{S}_{\mathscr{D}}$, it holds

$$\lim_{k\to\infty} \Lambda_k^{-1} D\left(w^\star, w_k\left(\omega\right)\right) - M\left(w^\star, w_k\left(\omega\right)\right) = r_{w^\star}\left(\omega\right) \in [0, +\infty[$$

Let $\tilde{\Omega} = \Omega' \cap \Omega''$, for any $\omega \in \tilde{\Omega}$ we let $w^1\left(\omega\right) \in \mathfrak{W}\left[\left(w_k\left(\omega\right)\right)_{k\in\mathbb{N}}\right]$ and $w^2\left(\omega\right) \in \mathfrak{W}\left[\left(w_k\left(\omega\right)\right)_{k\in\mathbb{N}}\right]$ be two weak sequential cluster points of $\left(w_k\left(\omega\right)\right)_{k\in\mathbb{N}}$, i.e., there exists two subsequences $\left(w_{k_i}\left(\omega\right)\right)_{i\in\mathbb{N}}$ and $\left(w_{k_j}\left(\omega\right)\right)_{j\in\mathbb{N}}$ such that $w_{k_i}\left(\omega\right) \rightharpoonup w^1\left(\omega\right)$ and $w_{k_j}\left(\omega\right) \rightharpoonup w^2\left(\omega\right)$ almost surely. Since $\mathfrak{W}\left[\left(w_k\left(\omega\right)\right)_{k\in\mathbb{N}}\right] \subset \boldsymbol{S}_{\mathscr{P}} \times \boldsymbol{S}_{\mathscr{D}}$, $w^1\left(\omega\right)$ and $w^2\left(\omega\right)$ are primal-dual optimal pairs.Thus, there exist $r_{w^1}\left(\omega\right), r_{w^2}\left(\omega\right) \in [0, +\infty[$ such that,

$$\lim_{k\to\infty} \left(\Lambda_k^{-1} D(w^1\left(\omega\right), w_k\left(\omega\right)) - M(w^1\left(\omega\right), w_k\left(\omega\right))\right) = r_{w^1}\left(\omega\right)$$

and

$$\lim_{k\to\infty} \left(\Lambda_k^{-1} D(w^2\left(\omega\right), w_k\left(\omega\right)) - M(w^2\left(\omega\right), w_k\left(\omega\right))\right) = r_{w^2}\left(\omega\right).$$

Using the three point identity, we have, for each $i \in \mathbb{N}$,

$$\begin{aligned}
&\Lambda_{k_i}^{-1} D(w^1\left(\omega\right), w_{k_i}\left(\omega\right)) - M(w^1\left(\omega\right), w_{k_i}\left(\omega\right)) - \Lambda_{k_i}^{-1} D(w^2\left(\omega\right), w_{k_i}\left(\omega\right)) + M(w^2\left(\omega\right), w_{k_i}\left(\omega\right)) \\
&= \Lambda_{k_i}^{-1} \left(D(w^1\left(\omega\right), w_{k_i}\left(\omega\right)) - D(w^2\left(\omega\right), w_{k_i}\left(\omega\right))\right) - \left(M(w^1\left(\omega\right), w_{k_i}\left(\omega\right)) - M(w^2\left(\omega\right), w_{k_i}\left(\omega\right))\right) \\
&= \Lambda_{k_i}^{-1} \left(D(w^1\left(\omega\right), w^2\left(\omega\right)) - \left\langle \nabla\phi(w_{k_i}\left(\omega\right)) - \nabla\phi(w^2\left(\omega\right)), w^1\left(\omega\right) - w^2\left(\omega\right)\right\rangle\right) \\
&\qquad - \left(M(w^1\left(\omega\right), w_{k_i}\left(\omega\right)) - M(w^2\left(\omega\right), w_{k_i}\left(\omega\right))\right).
\end{aligned} \tag{5.3.7}$$

Recall that, by $(\mathbf{A}_2)$, both $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ are nondecreasing and bounded above with limits λ_∞ and ν_∞, respectively. We denote $\Lambda_\infty \overset{\text{def}}{=} (\lambda_\infty, \nu_\infty)$. Then, recalling $(\mathbf{A}_9)$ and $(\mathbf{A}_{10})$ and passing to the limit in (5.3.7)

we get

$$\begin{aligned} r_{w^1}(\omega) - r_{w^2}(\omega) &= \Lambda_\infty^{-1}\left(D(w^1(\omega), w^2(\omega)) - \left\langle \nabla\phi(w^1(\omega)) - \nabla\phi(w^2(\omega)), w^1(\omega) - w^2(\omega)\right\rangle\right) \\ &\qquad + M(w^2(\omega), w^1(\omega)) \\ &= \Lambda_\infty^{-1}\left(D(w^1(\omega), w^2(\omega)) - D(w^1(\omega), w^2(\omega)) - D(w^2(\omega), w^1(\omega))\right) \\ &\qquad + M(w^2(\omega), w^1(\omega)) \\ &= -\Lambda_\infty^{-1} D(w^2(\omega), w^1(\omega)) + M(w^2(\omega), w^1(\omega)). \end{aligned}$$

Repeating this argument, replacing $w_{k_i}(\omega)$ by $w_{k_j}(\omega)$ above, we furthermore have

$$r_{w^1}(\omega) - r_{w^2}(\omega) = \Lambda_\infty^{-1} D(w^1(\omega), w^2(\omega)) - M(w^1(\omega), w^2(\omega)),$$

which shows that

$$\left[\Lambda_\infty^{-1} D(w^1(\omega), w^2(\omega)) - M(w^1(\omega), w^2(\omega))\right] + \left[\Lambda_\infty^{-1} D(w^2(\omega), w^1(\omega)) - M(w^2(\omega), w^1(\omega))\right] = 0.$$

By $(\mathbf{A}_3)$**(II)** and (5.1.7), we get that

$$L\left[D(w_1(\omega), w_2(\omega)) + D(w_2(\omega), w_1(\omega))\right] + \varepsilon\left[d(w_1(\omega), w_2(\omega)) + d(w_2(\omega), w_1(\omega))\right] = 0$$

and finally

$$D(w_1(\omega), w_2(\omega)) = D(w_2(\omega), w_1(\omega)) = 0.$$

Then, as ϕ_p and ϕ_d are Legendre by **(H)**, $w_1(\omega) = w_2(\omega)$ as claimed. Thus the sequence $(w_k(\omega))_{k\in\mathbb{N}}$ converges to some $\bar{w}(\omega) \in \mathfrak{W}\left[(w_k(\omega))_{k\in\mathbb{N}}\right] \subset \boldsymbol{S}_{\mathscr{P}} \times \boldsymbol{S}_{\mathscr{D}}$. Since this holds for all $\omega \in \tilde{\Omega}$ with $\mathbb{P}\left(\tilde{\Omega}\right) = 1$, we are done. □

5.3.4 Relatively Strongly Convex Case

We assume that either f, g, or both are relatively strongly convex (see Definition 2.1.11) with respect to ϕ_p with constant m_f, m_g, or $m_f + m_g$, respectively, as in $(\mathbf{A}_{11})$. For brevity, we analyze only the primal case but all of the analogous convergence results will hold for the dual case by making the corresponding assumptions on h^*, l^*, and ϕ_d, as in $(\mathbf{A}_{11})$. In addition, if the assumptions made here on the primal functions and entropies hold for the corresponding dual functions and entropies, we will have convergence results for the primal-dual sequence $(w_k)_{k\in\mathbb{N}}$. We also assume that ϕ_p is sequentially consistent and totally convex, which we now go on to define. The following definitions come from [25] although earlier notions of total convexity and its modulus exist.

Definition 5.3.6. Define, for all $x \in \text{intdom}\phi_p$ and $t \in [0, \infty[$,

$$\Theta_{\phi_p}(x, t) \overset{\text{def}}{=} \inf\left\{D_p(x', x) : \|x - x'\| = t\right\}.$$

The function Θ is called the modulus of total convexity and it is clearly nondecreasing in t (see [25][Page 18]). We call a function ϕ_p totally convex at a point x iff $\Theta_{\phi_p}(x, t) > 0$ for any $t > 0$. We say the function ϕ_p is totally convex on a set X iff it is totally convex for each $x \in X$.

Total convexity is a sort of generalization of strict convexity to functions defined on Banach spaces. Indeed, for finite-dimensional spaces, a function ϕ is strictly convex at every point $x \in \text{intdom}(\phi)$ iff it is totally convex at each $x \in \text{intdom}(\phi)$ (see [26]). Examples of totally convex functions include the Shannon-Boltzmann entropy, the Hellinger entropy, the Fermi-Dirac entropy, the Helinger entropy, the energy/euclidean entropy, and any strongly convex function as well. We point out that $(\mathbf{A}_{11})$ ensures that there is a unique solution $x^\star$ to $(\mathscr{P})$ (and similarly if we have $(\mathbf{A}_{11})$ on the dual).

Definition 5.3.7. A function ϕ_p is called sequentially consistent on a set X iff for any bounded subset $V \subseteq X$, for any $t > 0$, we have

$$\inf_{x\in V} \Theta_{\phi_p}(w, t) > 0.$$

Lemma 5.3.8. *Assume* (H), ($\mathbf{A}_1$), ($\mathbf{A}_2$), ($\mathbf{A}_3$), ($\mathbf{A}_4$), *and* ($\mathbf{A}_{11}$). *Then, for the solution* $x^\star$, $D_p(x^\star, x_k) \to 0$ ($\mathbb{P}$-*a.s.*). *Similarly, if we have* ($\mathbf{A}_{11}$) *and the corresponding assumptions for* ϕ_d, *it holds for the solution* $\mu^\star$, $D_d(\mu^\star, \mu_k) \to 0$ ($\mathbb{P}$-*a.s.*).

Proof. By Lemma 5.2.1 evaluated at $w = w^\star$ we have, for each $k \in \mathbb{N}$,

$$\left(\frac{1}{\Lambda_k} - m_{(f,h^*)}\right) D(w^\star, w_k) - M(w^\star, w_k) + \langle \Delta_k, w^\star - w_{k+1}\rangle \geq \left(\frac{1}{\Lambda_{k+1}} + m_{(g,l^*)}\right) D(w^\star, w_{k+1}) - M(w^\star, w_{k+1})$$

which we rewrite as, for each $k \in \mathbb{N}$,

$$\frac{1}{\Lambda_k} D(w^\star, w_k) - \frac{1}{\Lambda_{k+1}} D(w^\star, w_{k+1}) - M(w^\star, w_k) + M(w^\star, w_{k+1}) + \langle \Delta_k, w^\star - w_{k+1}\rangle \geq m_g D_p(x^\star, x_{k+1}) + m_f D_p(x^\star, x_k).$$

We now break the proof into two cases based on whether $m_g > 0$ or $m_f > 0$, starting with $m_f > 0$. Taking the expectation conditioned on the filtration, we have, for each $k \in \mathbb{N}$,

$$\begin{aligned} m_f D_p(x^\star, x_k) \leq \frac{1}{\Lambda_k} D(w^\star, w_k) - \frac{1}{\Lambda_{k+1}} \mathbb{E}[D(w^\star, w_{k+1}) \mid \mathcal{S}_k] - M(w^\star, w_k) \\ + \mathbb{E}[M(w^\star, w_{k+1}) \mid \mathcal{S}_k] + \mathbb{E}[\langle \Delta_k, w^\star - w_{k+1}\rangle \mid \mathcal{S}_k] \end{aligned} \tag{5.3.8}$$

Applying Remark 2.3.2 to (5.3.8) along with the assumption that $m_f > 0$ and ($\mathbf{A}_4$) with 5.2.5, we find that $(D(x^\star, x_k))_{k\in\mathbb{N}} \in \ell^1_+(\mathfrak{S})$ and, in particular, $D(x^\star, x_k) \to 0$ almost surely.

Now, assuming $m_g > 0$ gives, for each $k \in \mathbb{N}$,

$$m_g D_p(x^\star, x_{k+1}) \leq \frac{1}{\Lambda_k} D(w^\star, w_k) - \frac{1}{\Lambda_{k+1}} D(w^\star, w_{k+1}) - M(w^\star, w_k) + M(w^\star, w_{k+1}) + \langle \Delta_k, w^\star - w_{k+1}\rangle.$$

Taking the expectation then leads to, for each $k \in \mathbb{N}$,

$$m_g \mathbb{E}[D_p(x^\star, x_{k+1})] \leq \frac{1}{\Lambda_k} \mathbb{E}[D(w^\star, w_k)] - \frac{1}{\Lambda_{k+1}} \mathbb{E}[D(w^\star, w_{k+1})] - \mathbb{E}[M(w^\star, w_k)] + \mathbb{E}[M(w^\star, w_{k+1})] + \mathbb{E}[\langle \Delta_k, w^\star - w_{k+1}\rangle].$$

Then, by Lemma 2.2.1 with the assumption $m_g > 0$, ($\mathbf{A}_4$) and Lemma 5.2.5, we have that $(\mathbb{E}[D_p(x^\star, x_k)])_{k\in\mathbb{N}} \in \ell^1_+$ and so, by Lemma 2.3.4, we have that $D_p(x^\star, x_k) \to 0$ almost surely. □

Theorem 5.3.9. *Assume* (H), ($\mathbf{A}_1$), ($\mathbf{A}_2$), ($\mathbf{A}_3$), ($\mathbf{A}_4$), *and* ($\mathbf{A}_{11}$) *hold, that* ϕ_p *is totally convex and sequentially consistent, and let* $x^\star$ *be the solution to the primal problem. Then, if the sublevel sets of* $D_p(x^\star, \cdot)$ *are bounded, the sequence* $(x_k)_{k\in\mathbb{N}}$ *converges strongly to the solution* $x^\star$ *almost surely. Furthermore, if additionally the analog of* ($\mathbf{A}_{11}$) *holds for the dual,* ϕ_d *is totally convex and sequentially consistent, and the sublevel sets of* $D_d(\mu^\star, \cdot)$ *are bounded, then the sequence* $(w_k)_{k\in\mathbb{N}}$ *converges strongly to the primal-dual optimal pair* $w^\star$ *almost surely.*

Proof. Under these assumptions, Lemma 5.3.8 ensures $D_p(x^\star, x_k) \to 0$ almost surely. By the boundedness of the sublevel sets of $D(w^\star, \cdot)$, we have that the sublevel sets of $D_p(x^\star, \cdot)$ are bounded and thus the sequence $(x_k)_{k\in\mathbb{N}}$ is bounded, i.e., there exists $U_p \subseteq \mathcal{U}_p$ a bounded set such that, for each $k \in \mathbb{N}$, $x_k \in U_p$. Since ϕ_p is totally convex and sequentially consistent on $\mathcal{U}_p$, we have, for any $t > 0$,

$$\inf_{x\in U_p} \Theta_{\phi_p}(x, t) > 0.$$

Assume now that $(x_k)_{k\in\mathbb{N}}$ does not converge to $x^\star$. Then there exists a subsequence $\left(x_{k_j}\right)_{j\in\mathbb{N}}$, $\epsilon > 0$, and $K \in \mathbb{N}$ such that for all $j > K$ it holds,

$$\left\|x_{k_j} - x^\star\right\| > \epsilon.$$

Since $\left(x_{k_j}\right)_{j\in\mathbb{N}}$ is a subsequence of $(x_k)_{k\in\mathbb{N}}$, $\left(D_p\left(x^\star, x_{k_j}\right)\right)_{j\in\mathbb{N}}$ is a subsequence of $\left(D_p\left(x^\star, x_k\right)\right)_{k\in\mathbb{N}}$ and so its limit is 0. Since ϕ_p is both totally convex and sequentially consistent, and $\left\|x_{k_j} - x^\star\right\| > \epsilon$, the following is true, for any $j > K$,

$$D_p\left(x^\star, x_{k_j}\right) \geq \Theta_{\phi_p}\left(x_{k_j}, \left\|x_{k_j} - x^\star\right\|\right) > \Theta_{\phi_p}\left(x_{k_j}, \epsilon\right) \geq \inf_{x\in U_p} \Theta_{\phi_p}\left(x, \epsilon\right) > 0, \tag{5.3.9}$$

which contradicts the fact that $\lim_{j\to\infty} D_p\left(x^\star, x_{k_j}\right) = 0$ since the positive lower bound $\inf_{x\in U_p} \Theta_{\phi_p}\left(x, \epsilon\right)$ does not depend on j. Thus such a subsequence $\left(x_{k_j}\right)_{j\in\mathbb{N}}$ cannot exist and the desired claim follows.

Repeating this argument for the dual gives convergence of $(\mu_k)_{k\in\mathbb{N}}$ to the solution of the dual problem $\mu^\star$ and thus, if $(\mathbf{A}_{11})$ holds for the primal and the dual, we have that $(w_k)_{k\in\mathbb{N}}$ converges to a primal-dual optimal pair $w^\star$. □

Remark 5.3.10. The assumption that the sublevel sets of the the Bregman divergence be bounded, used in Theorem 5.3.9, holds for a wide class of entropies which includes the Shannon-Boltzmann entropy, the Hellinger entropy, the Fermi-Dirac entropy, the fractional power entropy, and energy/euclidean entropy (see [13, Remark 4]).

5.4 Applications and Numerical Experiments

The following results will be useful throughout the applications section, particularly when it comes to satisfying $(\mathbf{A}_3)$.

Lemma 5.4.1. *Assume that $\gamma > 0$, (H) holds, $\phi_p(x) = \sum_{i=1}^{n} x_i \log(x_i)$, and $\phi_d(\mu) = \frac{1}{2}\|\mu\|_2^2$. If $(\mathbf{A}_1)$ and $(\mathbf{A}_2)$ hold with $g(x) = \iota_{\{1\}}\left(x^T\mathbb{1}\right)$ and*

$$\lambda_\infty \leq \frac{1}{L_p + \gamma\|T\|_2^2} \text{ and } \nu_\infty \leq \frac{1}{L_d + \gamma^{-1}}$$

then $(\mathbf{A}_3)$ is satisfied with $\epsilon = \frac{1}{2}$ and

$$d(w_1, w_2) = \left(\frac{1}{\lambda_\infty} - L_p - \gamma\|T\|_2^2\right)\|x_1 - x_2\|_1^2 + \left(\frac{1}{\nu_\infty} - L_d - \frac{1}{\gamma}\right)\|\mu_1 - \mu_2\|_2^2.$$

Proof. By definition (see (5.1.1)), for any $w \in \tilde{\mathcal{U}}_p \times \tilde{\mathcal{U}}_d$ and $w' \in \mathcal{U}_p \times \mathcal{U}_d$, for each $k \in \mathbb{N}$,

$$\left(\frac{1}{\Lambda_\infty} - L\right) D(w, w') - M(w, w') = \left(\frac{1}{\lambda_\infty} - L_p\right) D_p(x, x') + \left(\frac{1}{\nu_\infty} - L_d\right)\frac{1}{2}\|\mu - \mu'\|_2^2 - \left\langle T\left(x - x'\right), \mu - \mu'\right\rangle.$$

Using Lemma 2.2.6, it holds for any $x \in \tilde{\mathcal{U}}_p$ and $x' \in \mathcal{U}_p$,

$$D_p(x, x') \geq \frac{1}{2}\|x - x'\|_1^2.$$

By Young's inequality, for any $\gamma > 0$, we also have, for any x and μ,

$$-\left\langle T\left(x - x'\right), \mu - \mu'\right\rangle \geq -\frac{\gamma}{2}\left\|T\left(x - x'\right)\right\|_2^2 - \frac{1}{2\gamma}\|\mu - \mu'\|_2^2$$

Combining the two and using the fact that $\|\cdot\|_2^2 \leq \|\cdot\|_1^2$, we find, for any $w \in \tilde{\mathcal{U}}_p \times \tilde{\mathcal{U}}_d$ and $w' \in \mathcal{U}_p \times \mathcal{U}_d$, for each $k \in \mathbb{N}$,

$$\begin{aligned}\left(\frac{1}{\Lambda_\infty} - L\right) D(w, w') - M(w, w') &\geq \frac{1}{2}\left[\left(\frac{1}{\lambda_\infty} - L_p - \gamma\|T\|_2^2\right)\|x - x'\|_2^2 + \left(\frac{1}{\nu_\infty} - L_d - \frac{1}{\gamma}\right)\|\mu - \mu'\|_2^2\right] \\ &\geq \frac{1}{2}\left[\left(\frac{1}{\lambda_\infty} - L_p - \gamma\|T\|_2^2\right)\|x - x'\|_1^2 + \left(\frac{1}{\nu_\infty} - L_d - \frac{1}{\gamma}\right)\|\mu - \mu'\|_2^2\right],\end{aligned}$$

where $\|T\|_2^2$ is the square of the classical operator norm and the desired claim follows. □

5.4.1 Linear Inverse Problems on the Simplex

In [31], the problem of least squares regression on the simplex was considered as an application of the Chambolle-Pock algorithm. A natural extension for Algorithm 10 is to replace the euclidean norm with the Kullback-Leibler divergence. The Kullback-Leibler divergence is not Lipschitz-smooth and so the Chambolle-Pock algorithm of [29] and [31] cannot be applied, although [31] does allow one to use an entropy in computing the D-proximal mapping associated to g.

Consider the problem,

$$\min_{\substack{x\in\mathbb{R}^n_+ \\ x^T\mathbb{1}=1}} D_K\left(Ax,b\right)+\beta\left\|\nabla x\right\|_1 \tag{5.4.1}$$

where $A:\mathbb{R}^n\to\mathbb{R}^m$ is a linear operator, $b\in\mathbb{R}^m_{++}$, K is the Shannon-Boltzmann entropy,

$$K\left(x\right)=\sum_{i=1}^n x_i\log\left(x_i\right),$$

and $\nabla:\mathbb{R}^n\to\mathbb{R}^{n-1}$ is the linear operator given by

$$\nabla x=\begin{pmatrix} x_2-x_1 \\ \vdots \\ x_n-x_{n-1}\end{pmatrix}.$$

Rewriting (5.4.1), the associated saddle-point problem is given by,

$$\min_{x\in\mathbb{R}^n_+}\max_{\mu\in\mathbb{R}^{n-1}} D_K\left(Ax,y\right)+\iota_{\{1\}}\left(x^T\mathbb{1}\right)+\left\langle\nabla x,\mu\right\rangle-\iota_{\mathcal{B}^\infty_\beta}\left(\mu\right).$$

We can apply Algorithm 10 with the following choices,

$$f\left(x\right)=D_K\left(Ax,y\right),\quad g\left(x\right)=\iota_{\{1\}}\left(x^T\mathbb{1}\right),\quad T=\nabla,\quad h^*\equiv 0,\quad l^*\left(\mu\right)=\iota_{\mathcal{B}^\infty_\beta}\left(\mu\right),$$
$$\mathcal{C}_p=\mathbb{R}^n_+,\ \text{ and }\ \mathcal{C}_d=\mathbb{R}^{n-1}.$$

We choose ϕ_p and ϕ_d to be

$$\phi_p\left(x\right)=\sum_{i=1}^n x_i\log\left(x_i\right)\ \text{ and }\ \phi_d\left(\mu\right)=\frac{1}{2}\left\|\mu\right\|_2^2$$

which induces the divergences D_p and D_d

$$D_p\left(x,x'\right)=\sum_{i=1}^n x_i\log\left(\frac{x_i}{x_i'}\right)-x_i+x_i'\ \text{ and }\ D_d\left(\mu,\mu'\right)=\frac{1}{2}\left\|\mu-\mu'\right\|_2^2.$$

This gives us the following D-prox operator for our problem,

$$\operatorname{prox}^{D_p}_{\lambda_k g}\left(x\right)\overset{\text{def}}{=}\operatorname*{argmin}_{u\in\mathcal{C}_p}\left\{\lambda_k g\left(u\right)+D_p\left(u,x\right)\right\}=\operatorname*{argmin}_{\substack{u\in\mathbb{R}^n_+ \\ u^T\mathbb{1}=1}}\left\{D_p\left(u,x\right)\right\}=\left(\frac{\exp\left(x_i\right)}{\sum_{j=1}^n\exp\left(x_j\right)}\right)_{i=1}^n.$$

The main hypothesis **(H)** is clearly satisfied in this problem. In order to satisfy **(A_1)**, we must find a constant $L_p>0$ such that $L_p\phi_p\left(x\right)-f\left(x\right)$ is convex for all $x\in\operatorname{int}\left(\operatorname{dom}\phi_p\right)=\mathbb{R}^n_{++}$. This is precisely what is shown in [13][Lemma 8], which we include here for clarity.

Lemma 5.4.2. *Let $\phi_p\left(x\right)=\sum_{i=1}^n x_i\log\left(x_i\right)$, $f\left(x\right)=D_K\left(Ax,y\right)$, and $A\in\mathbb{R}^{m\times n}_+$ such that none of the columns or rows of A are completely 0. Then, for any L_p such that*

$$L_p\geq\max_{1\leq j\leq m}\left(\sum_{i=1}^n A_{i,j}\right),$$

$L_p\phi_p-f$ is convex on $\mathbb{R}^n_{++}$.

Proof. See [13][Lemma 8] □

It remains to choose step sizes $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ such that $(\mathbf{A}_2)$ and $(\mathbf{A}_3)$ are satisfied, for which we refer to Lemma 5.4.1.

Remark 5.4.3. Notice that the constant $\gamma > 0$ in Lemma 5.4.1 is arbitrary. For the experiments, we took $\gamma = \|\nabla\|_2^{-1}$ to have symmetric step sizes,

$$\lambda_k = \frac{1}{L_p + \|\nabla\|_2} \text{ and } \nu_k = \frac{1}{L_d + \|\nabla\|_2}$$

since $L_d = 0$ in this problem.

We now apply Algorithm 10 to solve (5.4.1) using the step size and entropy choices discussed above. We take $n = 100$, $m = 100$, generate A with random i.i.d. uniformly distributed entries in $[0.01, 1.01]$, and generate b with random i.i.d. uniformly distributed entries in $[0, 1]$. We initialize with $x_0 = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ and $\mu_0 = 0$. We take the constant step sizes $\lambda_k = \frac{1}{L_p+\|\nabla\|_2}$ and $\nu_k = \frac{1}{\|\nabla\|_2}$. The lagrangian optimality gap is presented, for the ergodic and pointwise iterates, in Figure 5.1. To plot this gap, we first run the algorithm for a high number of iterations (1 million) to find the (approximate) primal-dual optimal pair $(x^\star, \mu^\star)$ and then rerun the algorithm for 80% of the number of inital iterations, computing the gap at each iteration.

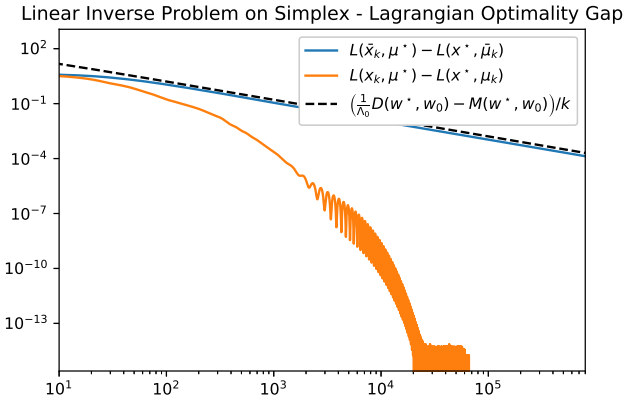

Figure 5.1: Ergodic and pointwise convergence profiles for Algorithm 10 applied to the linear inverse problem on the simplex, $n = 100$ and $\beta = 1$.

5.4.2 Trend Filtering on the Simplex

In the following we consider a variant of the so-called ℓ^1 trend filtering problem, introduced in [70] as a way to analyze time series data, applied to a new setting with simplex constraints. Let $Y \in \mathbb{R}^{n\times l}_{++}$, $X \in \mathbb{R}^{n\times m}_{+}$, and define

$$\mathcal{D}(X, Y) \stackrel{\text{def}}{=} \sum_{i=1}^{n} D_K\left(A^i x_i, y_i\right)$$

where K, as in the previous section, is the Shannon-Boltzmann entropy and, for each $i \in \{1, \ldots, n\}$, $A^i \in \mathbb{R}^{l\times m}_{+}$ is a linear operator, and $x_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}^l$ are the ith rows of X and Y respectively. We assume that the matrices A^i do not contain any rows which are completely zero. Next, define the following linear operator $\nabla : \mathbb{R}^{n\times m} \to \mathbb{R}^{m(n-1)}$ to be

$$\nabla X \stackrel{\text{def}}{=} (x_2 - x_1, \ldots, x_n - x_{n-1})$$

such that we have

$$\|\nabla X\|_1 = \sum_{i=1}^{n-1}\sum_{j=1}^{m} |X_{i+1,j} - X_{i,j}|.$$

The idea behind this choice of regularizer is to enforce a piecewise-constant structure on the columns of X. Meanwhile, we will constrain the rows of X to lie in the simplex. We formulate the problem as

$$\min_{\substack{X\in\mathbb{R}_+^{n\times m}\\ X\mathbb{1}_m=\mathbb{1}_n}} \mathcal{D}(X,Y)+\beta\|\nabla X\|_1 \tag{5.4.2}$$

where $\mathbb{1}_n$ is the length n column vector of all 1s, and its associated saddle-point problem,

$$\min_{X\in\mathbb{R}_+^{n\times m}}\max_{\mu\in\mathbb{R}^{m(n-1)}} \mathcal{D}(X,Y)+\iota_{\mathbb{1}_n}(X\mathbb{1}_m)+\langle\nabla X,\mu\rangle-\iota_{\mathcal{B}_\beta^\infty}(\mu).$$

Thus, it is possible to apply the algorithm with the following choices,

$$f(X)=\mathcal{D}(X,Y),\quad g(X)=\iota_{\mathbb{1}_n}(X\mathbb{1}_m),\quad T=\nabla,\quad h^*\equiv 0,\quad l^*(\mu)=\iota_{\mathcal{B}_\beta^\infty}(\mu),$$
$$\mathcal{C}_p=\mathbb{R}_+^{n\times m},\quad \text{and}\quad \mathcal{C}_d=\mathbb{R}^{m(n-1)}.$$

We take the entropies

$$\phi_p(X)=\sum_{i=1}^n\sum_{j=1}^m X_{i,j}\log(X_{i,j})\ \text{ and }\ \phi_d(\mu)=\frac{1}{2}\|\mu\|_2^2$$

which induce the divergences

$$D_p(X,X')=\sum_{i=1}^n\sum_{j=1}^m X_{i,j}\log\left(\frac{X_{i,j}}{X'_{i,j}}\right)-X_{i,j}+X'_{i,j}\ \text{ and }\ D_d(\mu,\mu')=\frac{1}{2}\|\mu-\mu'\|_2^2.$$

It is clear that **(H)** holds here. Once again we must find a constant $L_p>0$ such that $L_p\phi_p(x)-f(x)$ is convex for all $x\in\operatorname{int}(\operatorname{dom}\phi_p)$ to satisfy **($\mathbf{A}_1$)**.

Lemma 5.4.4. *For each $i\in\{1,\dots,n\}$, let $L_i\geq\max\limits_{1\leq q\leq m}\sum\limits_{p=1}^l A_{p,q}^i$ and let $L_p=\max\limits_{1\leq i\leq n}L_i$. Then $L_p\phi_p(X)-f(X)$ is convex for all $X\in\operatorname{int}(\operatorname{dom}\phi_p)$.*

Proof. Recall from [13][Lemma 8] that, for each $i\in\{1,\dots,n\}$, taking $L_i\geq\max\limits_{1\leq q\leq m}\sum\limits_{p=1}^l A_{p,q}^i$ implies that the function $\psi_i(X,L)$ defined by

$$\psi_i(X,L)\overset{\text{def}}{=}L\sum_{j=1}^m X_{i,j}\log(X_{i,j})-D_K\left(A^ix_i,y_i\right)$$

is convex for all $X\in\operatorname{int}(\operatorname{dom}\phi_p)$ when $L\geq L_i$. We can write $L_p\phi_p(X)-f(X)$ as

$$L_p\phi_p(X)-f(X)=\sum_{i=1}^n\psi_i(X,L_p)$$

and thus taking taking $L_p=\max\limits_{1\leq i\leq n}L_i$ ensures the desired result. □

As before, it remains to choose step sizes $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ such that **($\mathbf{A}_2$)** and **($\mathbf{A}_3$)** are satisfied. We refer again to Lemma 5.4.1, since the choice of entropy here is essentially the same as before due to separability with respect to the components of X.

We display below the results of applying Algorithm 10 to solve (5.4.2) using the step size and entropy choices from above. We take $n=100$, $m=3$, and A^i to be the identity for all $i\in\{1,\dots,n\}$. The rows of Y are generated with random i.i.d. Dirichlet distributed entries lying in the simplex. We initialize X_0 with each row $x_i=\left(\frac{1}{m},\dots,\frac{1}{m}\right)$ and $\mu_0=0$. We take the constant step sizes, as in the previous problem, for each $k\in\mathbb{N}$, $\lambda_k=\frac{1}{L_p+\|\nabla\|_2}$ and $\nu_k=\frac{1}{\|\nabla\|_2}$. The Lagrangian optimality gap is shown, for the ergodic and pointwise iterates, in Figure 5.2. As before, we run the algorithm for a high number of iterations to find the (approximate) primal-dual optimal pair $(x^\star,\mu^\star)$ and then rerun the algorithm for 80% of the number of initial iterations, computing the gap at each iteration. We also show the recovered trends or columns of X in Figure 5.3 and observe the effect of penalty parameter β on the recovered trends in Figure 5.4.

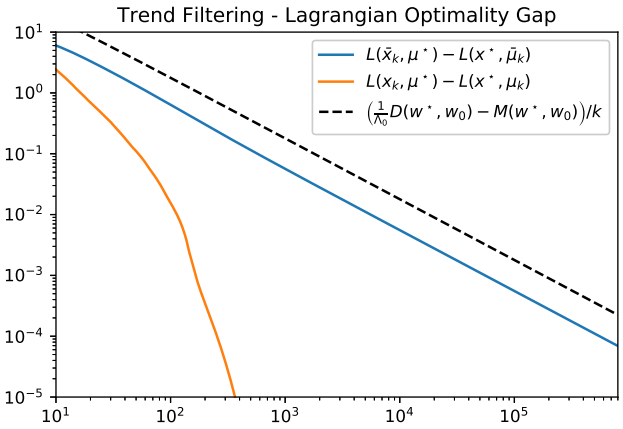

Figure 5.2: Ergodic and pointwise convergence profiles for Algorithm 10 applied to the trend filtering problem with $n = 100$, $m = 3$, and $\beta = 1$.

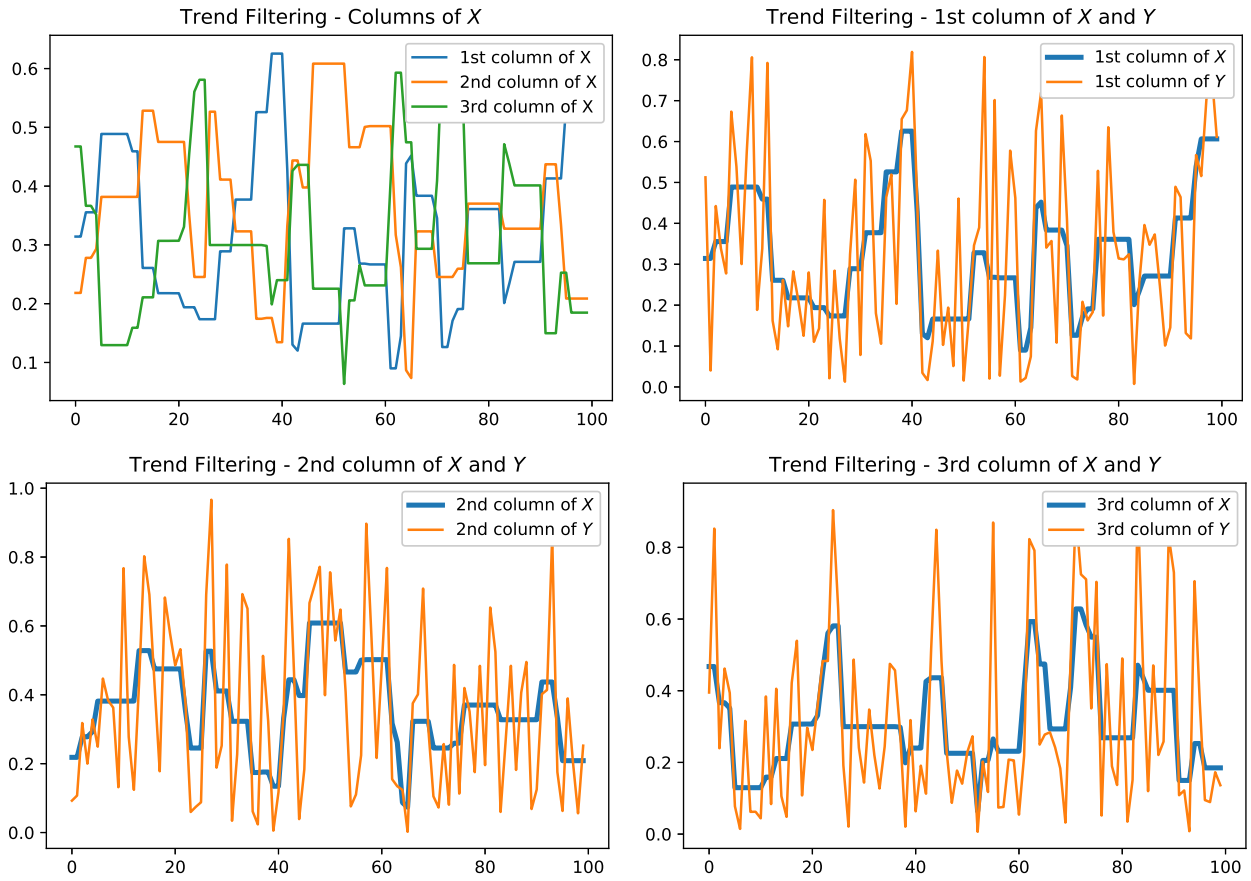

Figure 5.3: The recovered trends, i.e., columns of X from the trend filtering problem with $\beta = 1$.

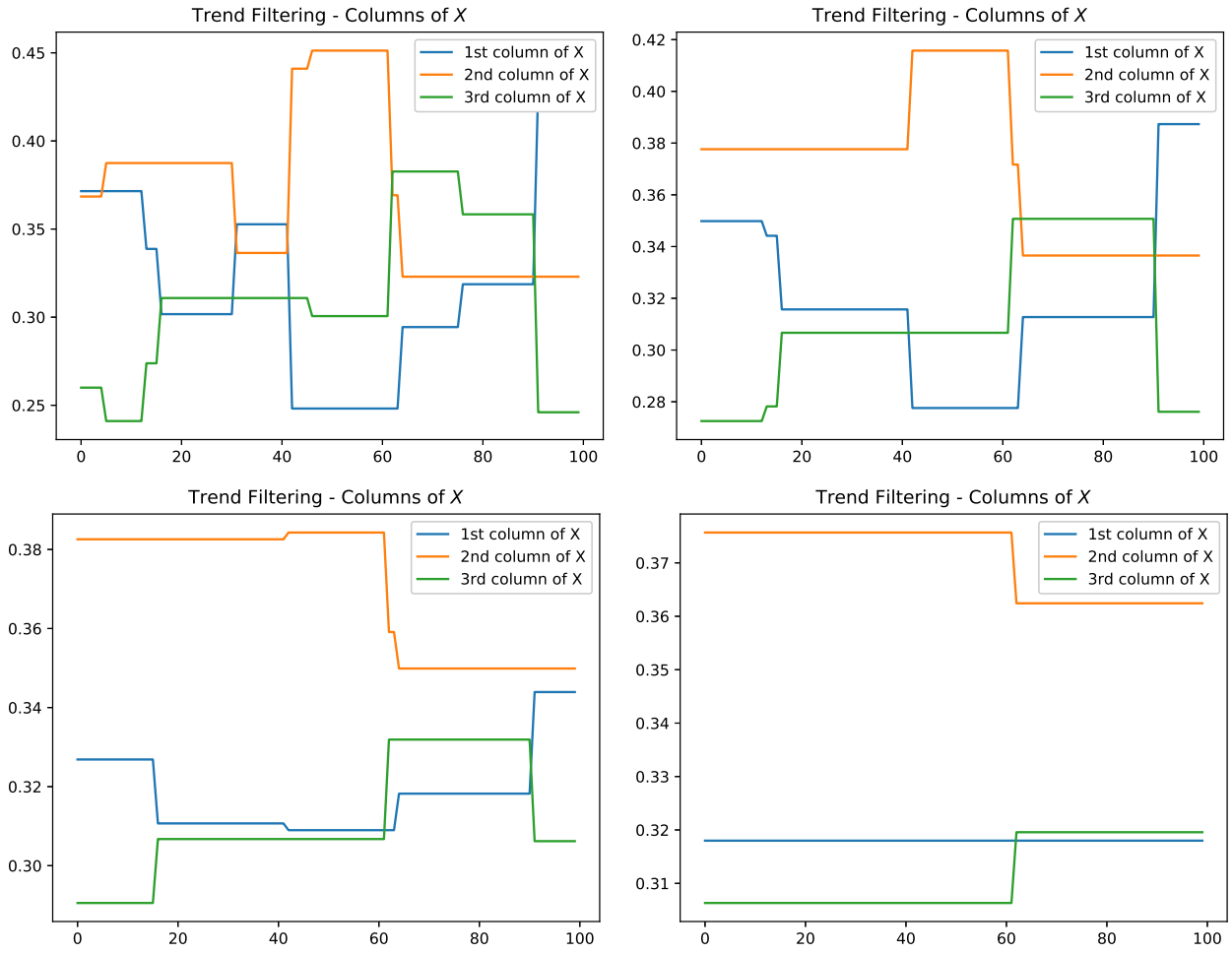

Figure 5.4: The recovered trends for different values of β; $\beta = 3, 4, 5$ and 6 starting at the top left.

5.4.3 Variational Problems with the Entropic Wasserstein Distance

Consider the optimal transport problem between two discrete measures, ρ and θ, defined on two metric spaces $\mathcal{X}$ and $\mathcal{Y}$. Let $C \in \mathbb{R}^{n\times m}$ be the ground cost on $\mathcal{X}\times\mathcal{Y}$. The cost C is typically application-dependent, and reflects some prior knowledge on the data to be processed. We regularize the optimal transport problem by subtracting in the objective the entropy of the transport plan π,

$$E(\pi) = -\sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{i,j}\log(\pi_{i,j})$$

The idea of regularizing the optimal transport problem by including the entropy of the transport plan π is not new, popularized by [40] and then explored, for example, in [41] for computing entropic Wasserstein barycenters, in [93] for approximating entropic Wasserstein gradient flows, in [42] for variational Wasserstein problems, in [43], etc. For $\gamma > 0$, the entropic regularization of the Kantorovich formulation of optimal transport can be written as the convex optimization problem

$$W_\gamma(\rho,\theta) \overset{\text{def}}{=} \inf_{\pi\in\Pi(\rho,\theta)} \left\{ \langle C,\pi\rangle + \gamma\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{i,j}\log(\pi_{i,j}) = \gamma\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{i,j}\log\left(\frac{\pi_{i,j}}{\xi_{i,j}}\right)\right\}, \tag{5.4.3}$$

where $\Pi(\rho,\theta) \overset{\text{def}}{=} \left\{\pi\in\mathbb{R}_+^{n\times m} : \pi\mathbb{1} = \rho, \pi^T\mathbb{1} = \theta\right\}$ is the so-called transportation polytope and $\xi_{i,j} \overset{\text{def}}{=} \exp(\frac{-C_{i,j}}{\gamma})$ is the Gibbs Kernel. When $\mathcal{X} = \mathcal{Y}$, $\gamma = 0$ and $C = d^p$, where d is a distance on $\mathcal{X}$, then $W_0^{1/p}$ is the well-known p-Wasserstein distance.

We consider solving the following variational problem over discrete measures, i.e. vectors in the simplex $\Sigma^n \overset{\text{def}}{=} \left\{x : x \geq 0, x^T\mathbb{1} = 1\right\}$,

$$\min_{\rho\in\Sigma^n} W_\gamma(F\rho,\theta) + J\circ A(\rho), \tag{5.4.4}$$

where $J\in\Gamma_0(\mathbb{R}^p)$, $F : \Sigma^n \to \Sigma^m$ and $A : \mathbb{R}^n \to \mathbb{R}^p$ are both linear operators. Seen as a matrix, F is typically column-stochastic while $\rho\in\Sigma^n$ is a discrete measure over the metric space $\mathcal{X}$ and $\theta\in\Sigma^m$ is the fixed, observed discrete measure over the metric space $\mathcal{Y}$.

Problem (5.4.4) is a natural way to solve inverse problems on discrete measures where one assumes that

$$\theta \approx F\rho_0,$$

where ρ_0 is an unknown discrete measure over $\mathcal{Y}$ to recover from the observed θ. When $F = \mathrm{Id}$ and $\gamma = 0$, (5.4.4) is closely related to computing the Wasserstein gradient (aka JKO [69]) flow of $J\circ A$. The JKO flow was first studied in [69] as it relates to the Fokker-Planck equation before being generalized (cf. [4], [106]). Entropic regularization, i.e., with $\gamma > 0$, was studied in [93] to compute Wasserstein gradient flows over spaces of probability distributions with the topology induced by the Wasserstein metric.

Applying Fenchel-Rockafellar duality to (5.4.3) (see [94][Proposition 2.4] for the unregularized case and [41][Section 5.1] for the entropic case), it is straightforward to see that problem (5.4.4) reads also

$$\min_{\rho\in\Sigma^n}\sup_{\tau\in\mathbb{R}^m,\eta\in\mathbb{R}^m} \langle\tau,F\rho\rangle + \langle\eta,\theta\rangle - \gamma\sum_{j=1}^{m}\sum_{i=1}^{m}\exp\left(\frac{\tau_i+\eta_j-C_{i,j}}{\gamma}\right) + J\circ A(\rho). \tag{5.4.5}$$

Taking the supremum over η, one can easily show that (see also [55, Proposition 2.1]),

$$\min_{\rho\in\Sigma^n}\sup_{\tau\in\mathbb{R}^m} \langle\tau,F\rho\rangle - \gamma\sum_{j=1}^{m}\theta_j\log\left(\sum_{i=1}^{m}\exp\left(\frac{\tau_i-C_{i,j}}{\gamma}\right)\right) + J\circ A(\rho). \tag{5.4.6}$$

Remark 5.4.5. Observe in (5.4.6) that the smooth term in τ (excluding the inner product $\langle\tau,F\rho\rangle$) is actually a log-sum-exp smooth approximation of the $\max$ function, which would appear naturally when marginalizing with respect to η in the case $\gamma = 0$.

Now, dualizing on J, we finally get that (5.4.4) is equivalent to

$$\min_{\rho\in\mathbb{R}^n_+}\sup_{\tau\in\mathbb{R}^m,\zeta\in\mathbb{R}^p}\iota_{\{1\}}(\rho^T\mathbb{1})+\langle(\tau,\zeta),(F\rho,A\rho)\rangle-\gamma\sum_{j=1}^m\theta_j\log\left(\sum_{i=1}^m\exp\left(\frac{\tau_i-C_{i,j}}{\gamma}\right)\right)-J^*(\zeta). \quad (5.4.7)$$

Remark 5.4.6. A chief advantage of (5.4.7), in contrast to optimizing with respect to the transport plan π, is the significant difference in computational complexity, since the former is operating over $n+m+p$ variables only rather than nm.

The problem in (5.4.7) is a saddle-point problem which can be solved with Algorithm 10 by taking

$$\mathcal{C}_p=\mathbb{R}^n_+,\quad \mathcal{C}_d=\mathbb{R}^{m+p},\quad T(\rho)=(F\rho,A\rho),\quad f(\rho)=0,\quad g(\rho)=\iota_{\{1\}}\left(\rho^T\mathbb{1}\right),$$
$$l^*(\mu)=l^*(\zeta)=J^*(\zeta),\ \text{ and }\ h^*(\mu)=h^*(\tau)=\gamma\sum_{j=1}^m\theta_j\log\left(\sum_{i=1}^m\exp\left(\frac{\tau_i-C_{i,j}}{\gamma}\right)\right).$$

The natural choice for the entropies is, again,

$$\phi_p(x)=\sum_{i=1}^n x_i\log(x_i)\ \text{ and }\ \phi_d(\mu)=\frac{1}{2}\|\mu\|_2^2.$$

Lemma 5.4.7. *The function $h^*(\mu)$ is L_d Lipschitz-smooth for $L_d\geq\gamma^{-1}\sum_{j=1}^m\theta_j=\gamma^{-1}$.*

Proof. The log-sum-exp function (with temperature constant γ),

$$lse_\gamma(x)\overset{\text{def}}{=}\gamma\log\left(\sum_{i=1}^n\exp\left(\frac{x_i}{\gamma}\right)\right),$$

is C^2 and convex on $\mathbb{R}^n$ (See [54][Lemma 4], [104][Example 2.16, page 48]) and thus so is $h^*(\tau,\zeta)$. The gradient, $\nabla_x lse_\gamma(x)$, is given, component-wise, for each $k\in\{1,\ldots,n\}$ by

$$(\sigma_\gamma(x))^{(k)}=\frac{\exp(x_k/\gamma)}{\sum_{i=1}^n\exp(x_i/\gamma)}.$$

The function $\sigma_\gamma(x)$ is called the softmax function with temperature constant γ and is Lipschitz-continuous in the euclidean norm with Lipschitz constant γ^{-1} (see [54][Proposition 4]). Thus, to see that the function h^* is Lipschitz-smooth, denote the jth column of C as $C_{\cdot,j}$ and notice

$$h^*(\mu)=h^*(\tau)=\sum_{j=1}^m\theta_j lse_\gamma(\tau-C_{\cdot,j})\implies\nabla h^*(\mu)=\nabla h^*(\tau)=\sum_{j=1}^m\theta_j\sigma_\gamma(\tau-C_{\cdot,j}).$$

With this we write,

$$\begin{aligned}\left\|\nabla h^*(\mu)-\nabla h^*(\mu')\right\|_2&=\left\|\sum_{j=1}^m\theta_j\left(\sigma_\gamma(\tau-C_{\cdot,j})-\sigma_\gamma(\tau'-C_{\cdot,j})\right)\right\|_2\\&\leq\left(\sum_{j=1}^m\theta_j\right)\left\|\sigma_\gamma(\tau-C_{\cdot,j})-\sigma_\gamma(\tau'-C_{\cdot,j})\right\|_2\\&\leq\gamma^{-1}\left(\sum_{j=1}^m\theta_j\right)\left\|\tau-\tau'\right\|_2\end{aligned}$$

and the desired claim follows. □

It is clear that **(H)** holds in this setting. It remains to find suitable step sizes $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ to satisfy **(A_2)** and **(A_3)**. Since the entropies here are exactly the same as in the linear inverse problem on the simplex, we

refer again to Lemma 5.4.1. With these step sizes, we consider a one-dimensional instance of the problem with $n = 108$, $C_{i,j} = \frac{1}{2}\|i-j\|_2^2$, F a convolution operator with kernel $K(x) = \exp\left(-\frac{1}{1-x^2}\right)$ for $x \in]-,1,1[$ and 0 otherwise, $J \circ A$ the total variation, and $F(\rho_0)$ corrupted by Dirichlet distributed noise, which we denote as $\widetilde{F}(\rho_0)$. We take $x_0 = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ and $\mu_0 = 0$. The results are displayed in Figure 5.5.

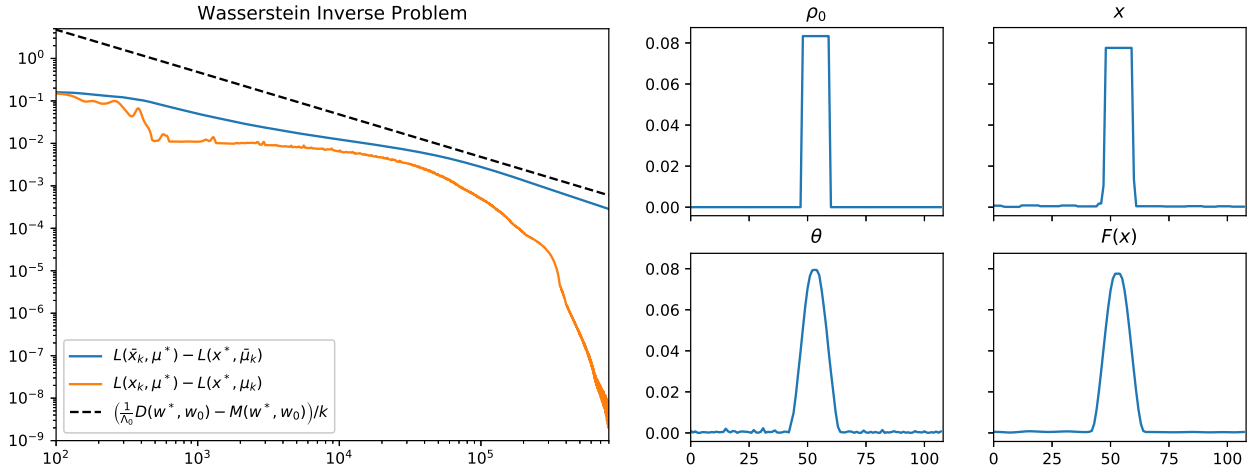

Figure 5.5: (Left) Ergodic and pointwise convergence profiles for Algorithm 10 applied to the Wasserstein inverse problem with entropic regularization parameter $\gamma = 1$, total variation regularization parameter $\beta = 1$, and $n = 108$. (Right) The ground truth measure ρ_0, the recovered measure x, the corrupted observation θ, and the image of the recovered measure $F(x)$.

Remark 5.4.8. Note that, although we considered here only a simple Wasserstein inverse problem involving a single observed measure, Algorithm 10 and our problem framework readily extend to more complicated regularized Wasserstein barycenter problems. Wasserstein barycenter problems were first introduced in [1] without entropic regularization of the Wasserstein distance. Later, the use of entropic regularization of the Wasserstein distance to speed up computation of barycenters was put forth in [41], however the barycenter itself was not regularized; such developments would come later, e.g., [28], [16], etc, and even then the problems considered did not include the possibility of observing the image of the measure θ under a linear operator F rather than observing the measure θ itself.

Let $q \in \mathbb{N}$ and consider q reference measures $\theta^i \in \mathbb{R}^{n_i}$ with $n_i \in \mathbb{N}$ for each $1 \leq i \leq q$, each having been observed through some linear operator $F^i : \mathbb{R}^n \to \mathbb{R}^{n_i}$ applied to an unknown discrete measure $\rho^i \in \Sigma^n$, i.e. $\theta^i \approx F^i \rho^i$. Then we can write the regularized Wasserstein barycenter problem as,

$$\min_{\rho \in \Sigma^n} \sum_{k=1}^{q} \alpha_q W_{\gamma_k}\left(F^k \rho, \theta^k\right) + \sum_{r=1}^{q'} J^r \circ A^r(\rho)$$

which is equivalent to the following,

$$\min_{\rho \in \Sigma^n} \sup_{\substack{\tau^1 \in \mathbb{R}^{n_1}, \ldots, \tau^q \in \mathbb{R}^{n_q} \\ \zeta^1 \in \mathbb{R}^{m_1}, \ldots, \zeta^{q'} \in \mathbb{R}^{m_{q'}}}} \sum_{k=1}^{q} \left[\langle \alpha_k \tau_k, F_k \rho \rangle - \alpha_k \gamma_k \sum_{j=1}^{n_k} \theta_j^k \log\left(\sum_{i=1}^{n_k} \exp\left(\frac{\tau_i^k - C_{i,j}^k}{\gamma_k} \right) \right) \right]$$
$$+ \sum_{r=1}^{q'} \left[\langle \zeta^r, A^r \rho \rangle - (J^r)^*(\zeta^r) \right].$$

This formulation of the problem can be solved with with Algorithm 10 by taking

$$\mathcal{C}_p = \mathbb{R}^n_+, \quad \mathcal{C}_d = \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_q} \times \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_{q'}}, \quad f(\rho) = 0, \quad g(\rho) = \iota_{\{\sum_{i=1}^{n} \rho_i = 1\}}(\rho),$$

$$l^*(\mu) = l^*(\zeta_1, \ldots, \zeta_{q'}) = \sum_{l=1}^{q'} J_l^*(\zeta_l), \text{ and}$$

$$h^*(\mu) = h^*(\tau^1, \ldots, \tau^q) = \sum_{k=1}^{q} \alpha_k \gamma_k \sum_{j=1}^{n_k} \theta_j^k \log\left(\sum_{i=1}^{n_k} \exp\left(\frac{\tau_i^k - C_{i,j}^k}{\gamma_k}\right)\right),$$

with the same entropy choices as we took for (5.4.7).

Remark 5.4.9. Consider the same setup as in the previous remark with $(\theta^1, \ldots, \theta^q)$ and let $\beta \in \mathbb{R}_+$. Another interesting formulation of the regularized Wasserstein barycenter problem that can be solved using Algorithm 10 is the following.

$$\min_{\rho} \min_{\rho_1, \ldots, \rho_q} \sum_{i=1}^{q} \left[W_{\gamma_i}(\theta^i, F^i \rho_i) + J \circ A(\rho^i)\right] + \beta \sum_{i=1}^{q} \alpha_i W_{\gamma_i}(\rho_i, \rho).$$

This problem is simultaneously solving the Wasserstein inverse problem for each observed measure θ^i while also finding a barycenter ρ among the proposed solutions ρ^i of the Wasserstein inverse problems.

Chapter 6

Conclusion

6.1 Summary

We have proposed two novel first-order algorithms for large-scale optimization. These algorithms will allow a new class of problems, previously inaccessible to the existing algorithms in the literature, to be solved efficiently. We have achieved this by analyzing our algorithms under very general assumptions that don't exclude key cases that arise in practice, such as in matrix completion or problems on the simplex. We have provided a rigorous treatment of the convergence properties of these algorithms, their associated convergence rates, and numerical experiments which demonstrate their validity. We have made these arguments with a high level of generality such that they apply even to infinite-dimensional problems. Our assumptions have been outlined in such a way that future researchers can easily understand in which way they contribute to the arguments we make.

In addition to this, we have also extended both of these algorithms to the stochastic setting where we have retained essentially all of the results from before in a ($\mathbb{P}$-a.s.) sense, with a few only carrying over in expectation. Such extensions allow for the practical utilization of these algorithms in large-scale settings, for example in machine learning, computer vision, inverse problems, signal processing, econometrics, operations research, etc.

We summarize the main conclusions to be drawn from our work:

(i) It is possible to show convergence of the Lagrangian and asymptotic feasibility of the iterates for generalized conditional gradient algorithms combined with an augmented Lagrangian type of penalization on the affine constraint under general smoothness conditions that broadly extend the class of Hölder smooth functions. Thus it is possible to have the best of all worlds; we can use the gradient, the proximal operator, or the linear minimization oracle depending on whichever is most accessible to us to maximize efficiency in practice, as was shown in the numerics of Chapter 3. There were two keys to allowing such a general problem formulation. The first was our fusion of relative smoothness with the curvature constant of [68]. This inspired a type of generalized curvature constant and smoothness, yielding a new descent lemma. The second key was the use of the augmented Lagrangian to handle the affine constraint. Having an affine constraint in the problem formulation allows one to solve problems involving multiple constraint sets and multiple nonsmooth functions g simultaneously, as was outlined in Chapter 3.

(ii) Similarly, it is possible to extend the primal-dual splitting framework, in a way which allows one to show convergence, to allow for highly composite saddle-point problems which are only relatively smooth and for which the proximal operators are calculated with respect to a Bregman divergence. This analysis does not require strong convexity of the objective or the entropies themselves in the general case but still guarantees convergence of the Lagrangian gap, with a $O\left(1/k\right)$ convergence rate for the ergodic iterates, and weak convergence of the pointwise iterates. The key to the analysis of this algorithm is the assumption we make which quantifies the relationship between the Bregman divergence $D\left(w, w'\right)$ and the linear term $M\left(w, w'\right)$. This assumption acts as a replacement for how one would use Young's inequality in typical primal-dual estimations found in, e.g., [31]. These results open the door to studying very challenging

composite structured optimization problems such as those appearing when solving entropically regularized Wasserstein inverse problems as well as a host of inverse problems involving the Kullback-Liebler divergence as a fidelity term, as shown in the applications of Chapter 5. These problems were previously inaccessible by first-order methods due to the lack of Lipschitz-smoothness.

(iii) Extension of algorithms which have been analyzed using arguments centered around quasi-Féjer monotonicity can be successfully performed through a stochastic perturbation analysis. Showing convergence results from such arguments boils down to showing that some summability condition involving the conditional expectation (or sometimes the total expectation) of the norm of the error is satisfied, as was done in Chapter 4. A model problem, previously inaccessible[1] to traditional stochastic conditional gradient methods, was solved numerically with the several different methods proposed, under varying batch sizes, and shown to agree with the proposed convergence analysis.

All the algorithms proposed here, and their theoretical results, have been verified numerically, in NumPy.

6.2 Future Work

Several paths appear to go further with the analysis of the algorithms proposed here.

Acceleration for Bregman Primal-Dual Splitting In the deterministic case, it would be interesting to further explore the effects that total convexity or strong convexity can have on the convergence rates of the algorithm. Is it possible to have some form of acceleration even with relative smoothness in place of Lipschitz-smoothness and without the entropies being strongly convex? For the closely related NoLips algorithm it's been shown that acceleration is not possible [47] in the general case without strong convexity, leading one to believe that a similar result is probably true for the SBPD algorithm. The answer to such a question, in the positive or the negative as we describe in the next paragraph, is not only enticing for theory reasons but also for practical purposes.

Optimal Complexity Analysis Analyzing the optimal complexity of the algorithms is another interesting path to be explored in the future. The idea, from [48], is to produce a lower bound for the convergence rate of the algorithm using semidefinite programming methods. This type of analysis was carried out for the NoLips algorithm utilizing relative smoothness assumptions in [47]. One can imagine that a similar analysis is possible, at least in principle, for both the CGALP and SBPD (in the deterministic case) algorithms.

Dynamical Systems Perspective Analyzing optimization algorithms as discretizations of continuous-time dynamical systems can yield fruitful results about the convergence and behavior of the optimization algorithm. This has been done, in Bregman settings in [18], in Riemannian settings in [3], and others like [7] and [110]. This perspective allows one access to all the tools of ordinary differential equation theory to answer questions regarding convergence to optimality and rates of convergence as well. Tying into the previous paragraph, analyzing the SBPD algorithm, in the deterministic case, using these methods will shed light on the how acceleration is possible, if at all.

CGALP for Banach Spaces Allowing one to solve problems over a real reflexive Banach space (and ultimately, if possible, a nonreflexive Banach space as motivated by problems involving Radon measures) would be useful for nonsmooth versions of the Beurling LASSO problem (see [23], [45] for generalized conditional gradient algorithms applied to the Beurling LASSO problem).

CGALP Beyond Bounded Sets Recently, it's been shown that, by imposing certain coercivity conditions on the differentiable function f to be minimized, one can construct a conditional gradient type algorithm which does not require the boundedness of the underlying constraint set $\mathcal{C}$. It is an open problem whether one can incorporate into CGALP these ideas for optimization over cones or other unbounded sets.

[1] Besides the contemporary work, [77], which handles the nonsmooth affine constraint through the Moreau envelope rather than an augmented Lagrangian approach.

Nonconvex CGALP Extending CGALP to nonconvex settings is an obvious path to follow. Some work has been done regarding general conditional gradient algorithms for nonconvex problems but, as far as we know, nothing has been said about CGALP type algorithms that invoke an augmented Lagrangian to handle the affine constraint (or even using the Moreau envelope, for that matter). Usually, to prove convergence under such assumptions one must choose the step-size carefully to ensure that the objective is decreasing sufficiently at each iteration; the current arguments used in the CGALP analysis do not allow one to specify a step-size in this way so new arguments would have to be presented.

Riemannian CGALP Similarly the previous paragraph, extending CGALP to allow for Riemannian optimization, i.e., optimization on a Riemannian manifold, is an obvious and practically useful path to follow. There has been work on extending classical conditional gradient methods in this direction, for instance in [113] and [114].

Stochastic Methods of Computing prox_g and ∇f Finally, for the ICGALP algorithm, the method approximating the gradient or proximal operator tends to restrict the parameters choices which, in turn, reduces the convergence rates for the feasibility and optimality. Developing a scheme which uses only one sample of the gradient at each iteration without imposing additional restrictions on the step sizes would be useful for practical purposes. It seems possible to extend the deterministic sweeping scheme presented here to allow for any sequence of random permutations of a finite sum of functions. The key to the analysis there is to have a lower bound on how frequently we sample each gradient, i.e., we must ensure a full pass of the data continues to happen regularly but we don't care in which order the pass is made. However, such an extension is limited to the finite sum minimization setting and cannot be applied to a true empirical risk minimization.

List of Publications

In preparation

(1) A. Silveti-Falls, C. Molinari, J. Fadili, *Stochastic Bregman Primal-Dual Splitting*.

Preprints

(2) A. Silveti-Falls, C. Molinari, and J. Fadili, *Inexact and Stochastic Generalized Conditional Gradient with Augmeneted Lagrangian and Proximal Step*, submitted to Journal of Nonsmooth Analysis and Optimization, *arXiv:2005.05158*, 2020. (Revised).

Journal Papers

(3) A. Silveti-Falls, C. Molinari, J. Fadili, *Generalized Conditional Gradient with Augmented Lagrangian and Proximal Step*, SIAM Journal on Optimization, 30(4):2687-2725, 2020.

(4) Gray, K., Hampton, B., Silveti-Falls, T., McConnell, A. and Bausell, C., *Comparison of Bayesian credible intervals to frequentist confidence intervals*, J. Mod. Appl. Statist. Meth. 14(8), 2015

Conference Papers

(5) A. Silveti-Falls, C. Molinari, and J. Fadili, *Generalized Conditional Gradient with Augmented Lagrangian for Composite Minimization.*, (**GRETSI**), 2019.

(6) A. Silveti-Falls, C. Molinari, and J. Fadili, *Generalized Conditional Gradient with Augmented Lagrangian for Composite Optimization*, Structure et Parcimonie pour la Représentation Adaptative de Signaux (**SPARS**), 2019 **(Winner of Best Student Paper Award)**.

List of Notations

General definitions

$\mathbb{R}$: the set of real numbers
$\mathbb{R}_+$: nonnegative real numbers
$\mathbb{R}_{++}$: positive real numbers
$\bar{\mathbb{R}}$: $]-\infty, +\infty[\cup\{+\infty\}$, the extended real values
ℓ_+^1: nonnegative summable sequence
$\mathbb{N}$: set of nonnegative integers
$\mathbb{N}_+$: set of positive integers
$\mathbb{R}^n, \mathbb{R}^m$: finite dimensional real Euclidean spaces
$\mathcal{H}, \mathcal{H}_p, \ldots$: real Hilbert spaces
$\mathcal{X}, \mathcal{X}_p, \ldots$: real reflexive Banach spaces
Id: identity operator on $\mathcal{H}$ or $\mathbb{R}^n$
T, A: bounded linear operators
$\mathbf{x}^{(\mathbf{i})}$: ith component of the vector $\mathbf{x}$
$\mathbb{1}$: vector of all 1s

Set related

$\mathcal{C}$: a convex (often compact) set
$\iota_{\mathcal{C}}(\cdot)$: indicator function for the set $\mathcal{C}$
$d_{\mathcal{C}}$: diameter of a convex set $\mathcal{C}$
$\sigma_{\mathcal{C}}(\cdot)$: support function of the set $\mathcal{C}$
$\mathrm{P}_{\mathcal{C}}(\cdot)$: projection operator onto $\mathcal{C}$
$\mathrm{int}\mathcal{C}$: interior of $\mathcal{C}$
$\bar{\mathcal{C}}$: closure of $\mathcal{C}$
$\mathrm{ri}(\mathcal{C})$: relative interior of $\mathcal{C}$
$\mathrm{aff}(\mathcal{C})$: smallest affine subspace that contains $\mathcal{C}$, a.k.a. affine hull of $\mathcal{C}$
$\mathrm{par}(\mathcal{C})$: the subspace parallel to $\mathcal{C}$
A^{-1}: inverse of A
$\mathrm{dom}(A)$: domain of A
$\mathrm{ran}(A)$: range of A
argmin: the set of minimizing arguments
$\mathbb{B}^r$: a ball centered at the origin with radius $r > 0$
$\mathcal{B}(\mathcal{H})$: the Borel σ-algebra on $\mathcal{H}$

Function related

$\Gamma_0(\mathcal{X}), \Gamma_0(\mathcal{H})$: the set of proper convex and lower semi-continuous functions on a Banach space $\mathcal{X}$, a Hilbert space $\mathcal{H}$, etc, respectively.
f, g, h, l: functions of $\Gamma_0(\mathcal{X})$, $\Gamma_0(\mathcal{H})$ or $\Gamma_0(\mathbb{R}^n)$
$f^\star$: the minimum value of the function f.

$\mathrm{lmo}_h(z)$: the linear minimization oracle associated to h, $\operatorname{argmin}_x \{\langle z, x\rangle + h(x)\}$

$\mathcal{L}$: the Lagrangian function

$\mathrm{dom}(J)$: domain of J

J^*: Fenchel conjugate of J

∇F: gradient of F

$\mathrm{prox}_{\gamma J}$: proximal operator of J with $\gamma > 0$

$J^\gamma(x)$: Moreau envelope of J parameterised by $\gamma > 0$

∂J: subdifferential of function J

$[\partial J(x)]^0$: minimal norm selection of ∂J at x

$(\gamma_k)_{k\in\mathbb{N}}$: a sequence indexed by k

D_ϕ: Bregman divergence associated to ϕ

$K_{(F,\zeta,\mathcal{C})}$: generalized curvature constant associated to (F, ζ) smoothness

$\mathbb{E}[x]$: total expectation of the random variable x

$\mathbb{E}[x \mid \mathcal{F}]$: expectation of the random variable x conditioned on the σ-algebra $\mathcal{F}$

$\mathbb{P}$: a probability measure

$\sigma(x_1, \ldots, x_n)$: the σ-algebra generated by $x_1, \ldots, x_n$.

List of Figures

Bibliography

[1] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

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

[3] Felipe Alvarez, Jérôme Bolte, and Olivier Brahic. Hessian riemannian gradient flows in convex programming. *SIAM journal on control and optimization*, 43(2):477–501, 2004.

[4] L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2008.

[5] Andreas Argyriou, Marco Signoretto, and J Suykens. Hybrid conditional gradient-smoothing algorithms with applications to sparse and low rank regularization. *Regularization, Optimization, Kernels, and Support Vector Machines*, pages 53–82, 2014.

[6] H. Attouch. *Variational convergence for functions and operators*. Applicable mathematics series. Pitman Advanced Publishing Program, 1984.

[7] Hedy Attouch, Zaki Chbani, Jalal Fadili, and Hassan Riahi. First-order optimization algorithms via inertial systems with hessian driven damping. *Mathematical Programming*, pages 1–43, 2020.

[8] F. Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.

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

[10] H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

[11] Heinz H. Bauschke, Jonathan M. Borwein, and Patrick L. Combettes. Essential smoothness, essential strict convexity, and legendre functions in banach spaces. *Communications in Contemporary Mathematics*, 03(04):615–647, 2001.

[12] Heinz H. Bauschke, Jonathan M. Borwein, and Patrick L. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization*, 42(2):596–636, 2003.

[13] H.H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Math. Oper. Res.*, 42(2):330–348, 2017.

[14] A. Beck, E. Pauwels, and S. Sabach. The cyclic block conditional gradient method for convex optimization problems. *SIAM Journal on Optimization*, 25, 02 2015.

[15] Amir Beck and Marc Teboulle. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Mathematical Methods of Operations Research*, 59(2):235–247, Jun 2004.

[16] Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Data-driven regularization of wasserstein barycenters with an application to multivariate density registration. *Information and Inference: A Journal of the IMA*, 8(4):719–755, 2019.

[17] Benjamin Birnbaum, Nikhil R Devanur, and Lin Xiao. Distributed algorithms via gradient descent for fisher markets. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 127–136, 2011.

[18] J. Bolte and M. Teboulle. Barrier operators and associated gradient-like dynamical systems for constrained minimization problems. *SIAM J. Control. Optim.*, 42:1266–1292, 2003.

[19] Jonathan M. Borwein. Maximality of sums of two maximal monotone operators in general banach space. *Proceedings of the American Mathematical Society*, 135(12):3917–3924, 2007.

[20] Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[21] K. Bredies and D. Lorenz. Iterated hard shrinkage for minimization problems with sparsity constraints. *SIAM Journal on Scientific Computing*, 30(2):657–683, 2008.

[22] K. Bredies, D. A. Lorenz, and P. Maass. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Computational Optimization and Applications*, 42(2):173–193, Mar 2009.

[23] K. Bredies and H.K. Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19:190–218, 2013.

[24] H. Brezis and A. Pazy. Convergence and approximation of semigroups of nonlinear operators in banach spaces. *J. Functional Analysis*, 9:63–74, 1972.

[25] Dan Butnariu and Alfredo N Iusem. *Totally convex functions for fixed points computation and infinite dimensional optimization*, volume 40. Springer Science & Business Media, 2000.

[26] Dan Butnariu, Alfredo N Iusem, and Constantin Zalinescu. On uniform convexity, total convexity and convergence of the proximal point and outer bregman projection algorithm in banach spaces. *Journal of Convex Analysis*, 10(1):35–62, 2003.

[27] P. Catala, V. Duval, and G. Peyré. A low-rank approach to off-the-grid sparse deconvolution. In *Journal of Physics: Conference Series*, volume 904, page 012015. IOP Publishing, 2017.

[28] Elsa Cazelles, Jérémie Bigot, and Nicolas Papadakis. Regularized barycenters in the wasserstein space. In *International Conference on Geometric Science of Information*, pages 83–90. Springer, 2017.

[29] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[30] Antonin Chambolle, Matthias J. Ehrhardt, Peter Richtárik, and Carola-Bibiane Schönlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

[31] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

[32] Liang-Ju Chu. On the sum of monotone operators. *Michigan Math. J.*, 43(2):273–289, 1996.

[33] P. L. Combettes. Quasi-Fejérian analysis of some optimization algorithms. *Studies in Computational Mathematics*, 8:115–152, 2001.

[34] P. L. Combettes and J. C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, 2011.

[35] P. L. Combettes and B. C. Vũ. Variable metric Forward–Backward splitting with applications to monotone inclusions in duality. *Optimization*, 63(9):1289–1318, 2014.

[36] Patrick L. Combettes and Jean-Christophe Pesquet. Primal-dual splitting algorithm for solving inclusions with mixtures of composite, lipschitzian, and parallel-sum type monotone operators. *Set-Valued and Variational Analysis*, 20(2):307–330, Jun 2012.

[37] Patrick L. Combettes and Jean-Christophe Pesquet. Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping, 2015.

[38] Patrick L Combettes and Valérie R Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[39] L. Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, pages 1–20, 2012.

[40] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[41] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. *Journal of Machine Learning Research*, 2014.

[42] Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

[43] Marco Cuturi and Gabriel Peyré. Semidual regularized optimal transport. *SIAM Review*, 60(4), Jan 2018.

[44] V. F. Dem'yanov and A. M. Rubinov. The minimization of a smooth convex functional on a convex set. *SIAM J. Control*, 5(2):280–294, 1967.

[45] Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The sliding frank–wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1):014001, dec 2019.

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

[47] Radu-Alexandru Dragomir, Adrien Taylor, Alexandre d'Aspremont, and Jérôme Bolte. Optimal complexity and certification of bregman first-order methods, 2019.

[48] Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1):451–482, Jun 2014.

[49] J.C. Dunn and S. Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432 – 444, 1978.

[50] J. Eckstein. Parallel alternating direction multiplier decomposition of convex programs. *Journal of Optimization Theory and Applications*, 80(1):39–62, 1994.

[51] J. Eckstein and D. P. Bertsekas. On the douglas–rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

[52] L.C. Evans. *Partial differential equations*. American Mathematical Society, Providence, R.I., 2010.

[53] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval research logitistics quarterly*, 3(1-2):95–110, 1956.

[54] Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning, 2017.

[55] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems 29*, pages 3440–3448. Curran Associates, Inc., 2016.

[56] G. Gidel, F. Pedregosa, and S. Lacoste-Julien. Frank-wolfe splitting via augmented lagrangian method. *10th NIPS Workshop on Optimization for Machine Learning*, 2018. in press (arXiv:1804.03176).

[57] Roland Glowinski and A Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 9(R2):41–76, 1975.

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

[59] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv preprint arXiv:1803.01401*, 2018.

[60] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Math. Program.*, 152(1-2):75–112, August 2015.

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

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

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

[64] Bingsheng He and Xiaoming Yuan. On the o(1/n) convergence rate of the douglas-rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

[65] C. Imbert. Convex analysis techniques for hopf-lax formulae in hamilton-jacobi equations. *Journal of Nonlinear and Convex Analysis*, 2(3):333–343, 2001.

[66] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 427–435, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[67] M. Jaggi, M. Sulovsk, et al. A simple algorithm for nuclear norm regularized problems. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 471–478, 2010.

[68] Martin Jaggi. *Sparse Convex Optimization Methods for Machine Learning*. PhD thesis, ETH Zurich, October 2011.

[69] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[70] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM review*, 51(2):339–360, 2009.

[71] N. Komodakis and J. Pesquet. Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems. *IEEE Signal Processing Magazine*, 32(6):31–54, 2015.

[72] S. Lacoste-Julien and M. Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *Advances in Neural Information Processing Systems*, pages 496–504, 2015.

[73] M. Laghdir and M. Volle. A general formula for the horizon function of a convex composite function. *Archiv der Mathematik*, 73(4):291–302, Oct 1999.

[74] VN Lebedev and NT Tynjanskiı. Duality theory of concave-convex games. In *Soviet Math. Dokl*, volume 8, pages 752–756, 1967.

[75] E.S. Levitin and B.T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1 – 50, 1966.

[76] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[77] 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.

[78] Haihao Lu. "relative continuity" for non-lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent. *INFORMS Journal on Optimization*, 1(4):288–303, 2019.

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

[80] Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[81] L. McLinden. An extension of fenchel's duality theorem to saddle functions and dual minimax problems. *Pacific J. Math.*, 50(1):135–158, 1974.

[82] 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.

[83] C. Molinari, J. Liang, and J. Fadili. Convergence rates of Forward–Douglas–Rachford splitting method. *ArXiv e-prints*, January 2018.

[84] J.-J. Moreau. Théorèmes "inf-sup,". *C. R. Acad. Sci. Paris Sér. A Math.*, 258:2720–2722, 1964.

[85] H. Narasimhan. Learning with complex loss functions and constraints. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1646–1654, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

[86] Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, Aug 2015.

[87] Yu. Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming*, 171(1):311–330, Sep 2018.

[88] 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.

[89] Z. Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bulletin of the American Mathematical Society*, 73(4):591–597, 1967.

[90] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[91] G. B. Passty. Ergodic convergence to a zero of the sum of monotone operators in hilbert space. *Journal of Mathematical Analysis and Applications*, 72(2):383–390, 1979.

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

[93] Gabriel Peyré. Entropic approximation of wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

[94] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[95] Mark Semenovich Pinsker. The information stability of gaussian random variables and processes (in russian)). In *Doklady Akademii Nauk*, volume 133, pages 28–30. Russian Academy of Sciences, 1960.

[96] E. Polak. An historical survey of computational methods in optimal control. *SIAM Review*, 15(2):553–584, 1973.

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

[98] H. Raguet, M. J. Fadili, and G. Peyré. Generalized forward-backward splitting. *SIAM Journal on Imaging Sciences*, 6(3):1199–1226, 2013.

[99] Julian Rasch and Chambolle Antonin. Inexact first-order primal–dual algorithms. *Computational Optimization and Applications*, 76(2):381–430, 2020.

[100] 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.

[101] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In Jagdish S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233 – 257. Academic Press, 1971.

[102] R. T. Rockafellar. Minimax theorems and conjugate saddle-functions. *Mathematica Scandinavica*, 14(2):151–173, 1964.

[103] R. T. Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

[104] R. T. Rockafellar and R. Wets. *Variational analysis*, volume 317. Springer Verlag, 1998.

[105] Shoham Sabach and Marc Teboulle. Lagrangian methods for composite optimization. In *Handbook of Numerical Analysis*, volume 20, pages 401–436. Elsevier, 2019.

[106] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

[107] Ron Shefi and Marc Teboulle. Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. *SIAM Journal on Optimization*, 24(1):269–297, 2014.

[108] Antonio Silveti-Falls, Cesare Molinari, and Jalal Fadili. Generalized conditional gradient with augmented lagrangian for composite minimization. *SIAM Journal on Optimization*, 30(4):2687–2725, 2020.

[109] Antonio Silveti-Falls, Cesare Molinari, and Jalal Fadili. Inexact and stochastic generalized conditional gradient with augmented lagrangian and proximal step, 2020. Journal of Nonsmooth Analysis and Optimization (Revised)).

[110] Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. *The Journal of Machine Learning Research*, 17(1):5312–5354, 2016.

[111] S. Vaiter, M. Golbabaee, J. Fadili, and G. Peyré. Model selection with low complexity priors. *Information and Inference: A Journal of the IMA (IMAIAI)*, 4(3):230–287, 2015.

[112] B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, pages 1–15, 2011.

[113] Melanie Weber and Suvrit Sra. Frank-wolfe methods for geodesically convex optimization with application to the matrix geometric mean. *arXiv preprint arXiv:1710.10770*, 2017.

[114] Melanie Weber and Suvrit Sra. Nonconvex stochastic optimization on manifolds via riemannian frank-wolfe methods. *arXiv preprint arXiv:1910.04194*, 2019.

[115] 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.

[116] A. Yurtsever, O. Fercoq, F. Locatello, and V. Cevher. A conditional gradient framework for composite convex minimization with applications to semidefinite programming. *ICML*, 80:5713–5722, 2018.

[117] A. Yurtsever, M. Udell, J. A Tropp, and V. Cevher. Sketchy decisions: Convex low-rank matrix optimization with optimal storage. *arXiv preprint arXiv:1702.06838*, 2017.

[118] X. Zhang, D. Schuurmans, and Y.L. Yu. Accelerated training for matrix-norm regularization: A boosting approach. In *Advances in Neural Information Processing Systems*, pages 2906–2914, 2012.