Low-Order Optimization Algorithms: Iteration Complexity and Applications

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Xiang Gao

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisor: Shuzhong Zhang

May, 2018

Acknowledgements

First, I would like to express my sincere gratitude to my advisor, Professor Shuzhong Zhang for his continuous support and invaluable guidance both in academic research and other aspects of life. This dissertation would not have been possible without him. It has been an honor to be his student, and I could not have imagined having a better advisor for my Ph.D. study.

Besides my advisor, I would like to thank the rest of my thesis committee: Professor Daniel Boley, Professor William L. Cooper, and Professor Zizhuo Wang for their precious time and insightful comments on my dissertation.

My gratitude extends to my colleagues and collaborators: Bo Jiang, Yangyang Xu, Xiaobo Li, Shaozhe Tao, and Junyu Zhang. I have received great helps from them. Several exciting joint works are originated from our numerous inspiring discussions.

My time at Minnesota was made most enjoyable in large part due to the friends of mine. My special thanks go to Shaozhe Tao, Xiaobo Li, Junyu Zhang, Xiao Chen, Guiyun Feng, Zeyang Wu, Ruizhi Shi, Qingwei Chen, Zhiyuan Xu and Kaiyu Wang.

Finally, I would like thank my parents for their endless love and support since I was born, and this dissertation is dedicated to them.

Abstract

Efficiency and scalability have become the new norms to evaluate optimization algorithms in the modern era of big data analytics. Despite its superior local convergence property, second or higher-order methods are often disadvantaged when dealing with large-scale problems arising from machine learning. The reason for this is that the second or higher-order methods require the amount of information, or to compute relevant quantities (e.g. Newton's direction), which is exceedingly large. Hence, they are not scalable, at least not in a naive way. Because of exactly the same reason, with substantially lower computational overhead per iteration, lower-order (first-order and zeroth-order) methods have received much attention and become popular in recent years. In this thesis, we present a systematic study of the lower-order algorithms for solving a wide range of different optimization models. As a starting point, the alternating direction method of multipliers (ADMM) will be studied and shown to be an efficient approach for solving large-scale separable optimization with linear constraint. However, the ADMM is originally designed for solving two-block optimization models and its subproblems are not always easy to solve. There are two possible ways to increase the scope of application for the ADMM: (1) to simplify its subroutines so as to fit a broader scheme of lower-order algorithms; (2) to extend it to solve a more general framework of multi-block problems. Depending on the informational structure of the underlying problem, we develop a suite of first-order and zeroth-order variants of the ADMM, where the trade-offs between the required information and the computational complexity are explicitly given. The new variants allow the method to be applicable to a much broader class of problems where only noisy estimations of the gradient or the function values are accessible. Moreover, we extend the ADMM framework to a general multi-block convex optimization model with coupled objective function and linear constraints. Based on a linearization scheme to decouple the objective function, several deterministic first-order algorithms have been developed for both two-block and multi-block problems. We show that, under suitable conditions, the sublinear convergence rate can be established for those methods. It is well known that the original ADMM may fail to converge when the number of blocks exceeds two. To overcome this difficulty, we propose a randomized primal-dual proximal block coordinate updating framework which includes several existing ADMM-type algorithms as special cases. Our result shows that if an appropriate

randomization procedure is used, then a sublinear rate of convergence in expectation can be guaranteed for multi-block ADMM, without assuming strong convexity or any additional conditions. The new approach is also extended to solve problems where only a stochastic approximation of the (sub-)gradient of the objective is available. Furthermore, we study various zeroth-order algorithms for both black-box optimizations and online learning problems. In particular, for the black-box optimization, we consider three different settings: (1) the stochastic programming with the restriction that only one random sample can be drawn at any given decision point; (2) a general nonconvex optimization framework with what we call the weakly pseudo-convex property; (3) an estimation of objective value with controllable noise is available. We further extend the idea to the stochastic bandit online learning problem, where the nonsmoothness of the loss function and the one random sample scheme are discussed.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Literature Review

Algorithm design is commonly considered as a central theme in the theory and practice of optimization. For continuous optimization, roughly speaking, algorithms can be classified into three types: (1) the high-order algorithms (which use the information of the Hessian or higher order derivatives of the objective function); (2) the first-order algorithms (which use no more than the gradient information of the objective function); (3) the zeroth-order algorithms (which only use the function value information). In this dissertation, we aim to present a study on the latter two types of algorithms for some specific optimization models. To distinguish from the high-order ones, let us loosely use the term *low-order* algorithms to represent the last two types of methods. High-order methods such as the interior point algorithms have proved to be extremely successful in solving optimization problems in general, as they typically only take a few steps to converge (cf. [6, 5]). However, there are applications arising from big data analytics that prevent high-order methods from being practical, as the computational complexity of performing one iteration of a high-order method may already be overly expensive. In those situations, the first-order methods become attractive since at each step their computational costs are substantially lower. Furthermore, in some applications only the function values are available for estimation. In such cases, the zeroth-order methods are the only choices to be considered.

As two subclasses of the lower-order algorithms, the first-order methods and the zeroth-order methods are closely related. In fact, as we will show in this dissertation, many zeroth-order methods can be derived from their well-designed first-order coun-

terpart. In the literature, two types of first-order methods are popular. The first type includes essentially the gradient algorithms and their variations, while the second type is based on the proximal gradient mappings. Consider

$$\min_{x\in\mathbb{R}^n} f(x) \tag{1.1}$$

where $f(x)$ is a smooth convex function. The gradient method is in the form of

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k), \tag{1.2}$$

where α_k is the step size. Moreover, for some problems, the objective function might be nonsmooth and there might be constraints and so the gradient method is not directly applicable. For example, consider

$$\min_{x\in\mathcal{X}} f(x) = f_0(x) + f_1(x), \tag{1.3}$$

where $\mathcal{X}$ is a convex set, and f_i is convex, $i = 0, 1$, and f_0 may be nonsmooth. This is a typical situation where proximal type method may be relevant. In particular, we define the *proximal operator* $\mathbf{prox}_{\lambda f}$ as

$$\mathbf{prox}_{\lambda f}(x) = \arg\min_{y\in\mathcal{X}} f(y) + \frac{1}{2\lambda}\|y - x\|^2.$$

For problem (1.3), the proximal point method can be described as the following iterative process

$$x^{k+1} = \mathbf{prox}_{\lambda_k f_0}(x^k), \tag{1.4}$$

and the proximal gradient method can be described as

$$x^{k+1} = \mathbf{prox}_{\lambda_k f_0}(x^k - \lambda_k \nabla f_1(x^k)). \tag{1.5}$$

There have been many variations originated from the proximal point and the proximal gradient methods, adapted for specific applications in various fields including engineering, statistics, and economics (cf. [11]). Besides, the aforementioned gradient-type methods can also serve as a starting point for many zeroth-order methods. When the gradient information is not readily available or impractical to obtain, based on some approximations of the gradient, the corresponding zeroth-order method can still be

applied.

Given the nature of the lower-order algorithm, it is particularly useful for problems that require higher computational efficiency. In general, we study its applications for the following areas: (1) large-scale block optimization; (2) stochastic and black-box optimization; (3) online learning and online optimization. In the previous gradient-type optimization methods, the vector x is treated as a single block of variables. But in large-scale optimization problems, the dimension n is large and it would be preferable to work with some smaller-sized subproblems at each step. In fact, there are plenty of problems including matrix/tensor factorization [65, 69], group LASSO [117, 128], SVM [22] etc., where x can be decomposed as $x = (x_1, x_2, \ldots, x_m)^\top$. With such kind of problems in mind, let us consider the following block-structured optimization model

$$\min_{x_i \in \mathcal{X}_i, i=1,\ldots,m} f(x_1, x_2, \ldots, x_m) + \sum_{i=1}^{m} u_i(x_i), \tag{1.6}$$

where $f(\cdot)$ is smooth, and $u_i(\cdot)$ may be nonsmooth. To solve this problem, it is intuitive to utilize the block structure so that at each step of the algorithm, we only need to deal with a smaller-sized problem. To this end, the *Block Coordinate Descent* (BCD) method is proposed for solving problem (1.6). Basically, the BCD method tries to minimize a single block variable x_i while all other blocks x_j, $j \neq i$ are fixed at each step by following a certain selection rule of the block (e.g. cyclic, randomized, etc.). By incorporating the proximal point or proximal gradient method and implementing different block updating rule, many variants of the BCD methods have been proposed. However, for some applications, for instance the robust PCA [14], (1.6) is still not general enough. Taking the constraints into account, let us consider the following model

$$\begin{aligned} \min_{x_i \in \mathcal{X}_i, i=1,\ldots,m} \quad & f(x_1, x_2, \ldots, x_m) + \sum_{i=1}^{m} u_i(x_i) \\ \text{s.t.} \quad & A_1 x_1 + A_2 x_2 + \cdots + A_m x_m = b, \end{aligned} \tag{1.7}$$

where $A_i, i = 1, \ldots, m$ are given matrices. For instance, by introducing a new variable, the well-known LASSO model [117] can be transformed into (1.7):

$$\min_{x \in \mathbb{R}^n} \quad \|Ax - b\|^2 + \lambda \|x\|_1 \quad \Rightarrow \quad \begin{aligned} \min_{x,y \in \mathbb{R}^n} \quad & \|Ax - b\|^2 + \lambda \|y\|_1 \\ \text{s.t.} \quad & x - y = 0. \end{aligned} \tag{1.8}$$

There are in fact many applications which can be formulated in the form of (1.7) including the consensus and sharing problems, basis pursuit, compressive sensing etc.; see [7, 81, 20, 29]. For this model, the ADMM-type methods based on augmented Lagrangian have received much attention recently, which will be discussed at length later in this thesis. Unlike the deterministic large-scale optimization, for many stochastic and black-box optimizations, the lower-order algorithm seems to be the only feasible solution. A stochastic optimization problem often assumes the following structure

$$\min_{x\in\mathcal{X}} f(x) := \mathsf{E}[F(x,\xi)], \tag{1.9}$$

where the expectation is taken over a random variable ξ. For many stochastic problems, the function F and the distribution of ξ are either very complex or unknown, which makes the higher-order information become unavailable. The black-box optimization further generalizes it into a nonparametric model where no functional form is assumed for the objective function. A good example is the hyperparameter tuning of machine learning algorithms, where the generalization error is used as the objective function. Clearly, for a given set of hyperparameters, the only possible information is the cross-validation or test error which is a noisy approximation of true generalization error. In light of this, both stochastic and black-box optimization can be viewed as the oracle-based optimization problem. For any query point x, depending on the informational structure of a problem, the corresponding feedback is given by an oracle. Moreover, the lower-order algorithm is even more powerful for solving a combination of the large-scale and the oracle-based optimization. As an extension of optimization to a changing environment, the online learning or online optimization also possesses a lower-order nature. In online learning, at each decision period $t \in \{1, 2, \ldots, T\}$, an online player chooses a feasible strategy x_t from a decision set $\mathcal{X} \subset \mathbb{R}^n$, and suffers a loss given by $f_t(x_t)$, where $f_t(\cdot)$ is a loss function. The key feature of this framework is that the player must make a decision for period t without knowing the loss function $f_t(\cdot)$. In this challenging setting with limited information, the lower-order algorithms become more appropriate.

Convergence or computational complexity analysis is an indispensable part of optimization theory. In general, the iteration complexity is a measure of how well the algorithm performs after a certain number of iterations. The way to measure the quality of the current iterate varies from problem to problem, but it mainly includes the distance between the iterate and the optimal solution set, the difference between the current

function value and the optimal function value, and the violation of the optimality conditions. For example, if an algorithm has the iteration complexity of the order $O(1/N)$ in terms of the objective function value, then this means that after k iterations, the current iterate x^k would satisfy $f(x^k) - f^* < \frac{C}{k}$ where C is a constant. For the two fundamental methods: the gradient method (1.2) and the proximal gradient method (1.5), the iteration complexities are well studied. In particular, the gradient method has been shown an iteration complexity of $O(1/N)$ (cf. [6]). Moreover, [86] shows that the complexity can be further improved to $O(1/N^2)$ by an acceleration procedure and this is the optimal rate that any gradient-type method can possibly achieve, and if the function is strongly convex then the method actually converges linearly. For the proximal gradient method, similar results hold: the $O(1/N)$ complexity of the original method, which can be accelerated to an $O(1/N^2)$ iteration complexity; see [3, 87, 88, 120]. For problem (1.6), extensive research of the iteration complexity of the BCD-type method has been reported in the literature. Under different conditions, the sublinear convergence rate $O(1/N)$ or the linear rate can be achieved for the block minimization BCD-type methods (cf. [79, 125, 60]). Furthermore, for the proximal gradient BCD-type method, [90, 100, 77, 4, 60] show that the similar convergence rate can still be established.

In this Ph.D. thesis, we study the first-order and zeroth-order methods for optimization models around the following themes: the iteration complexity analysis of different ADMM-type algorithms for solving various block optimization problems, and the analysis of lower-order gradient-type algorithms for solving oracle-based black-box optimization and online optimization.

Instead of aiming to solve the general multi-block model (1.7), the basic *Alternating Direction Method of Multipliers* (abbreviated as ADMM) is originally designed to solve the following two-block constrained convex optimization model

$$\begin{array}{ll} \min & f(x) + g(y) \\ \text{s.t.} & Ax + By = b, \\ & x \in \mathcal{X}, y \in \mathcal{Y} \end{array} \tag{1.10}$$

where $x \in \mathbb{R}^{n_x}$, $y \in \mathbb{R}^{n_y}$, $A \in \mathbb{R}^{m \times n_x}$, $B \in \mathbb{R}^{m \times n_y}$, $b \in \mathbb{R}^m$, and $\mathcal{X} \subseteq \mathbb{R}^{n_x}$, $\mathcal{Y} \subseteq \mathbb{R}^{n_y}$ are closed convex sets; f and g are convex functions.

An intensive recent research attention for solving problem (1.10) has been devoted to the ADMM, which is known to be a manifestation of the operator splitting method (cf. [30, 33, 46] and the references therein). Large-scale optimization problems in the

form of (1.10) can be found in many application domains including compressed sensing, imaging processing, and statistical learning. Due to the large-scale nature, it is often impossible to inquire the second order information (such as the Hessian of the objective function) or invoke any second order operations (such as inverting a full-scale matrix) in the solution process. In this context, the ADMM as a first order method is an attractive approach; see [10]. Specifically, for solving (1.10), a typical iteration of ADMM proposed in [46] runs as follows:

$$\begin{cases} x^{k+1} = \arg\min_{x\in\mathcal{X}} \mathcal{L}_\gamma(x, y^k, \lambda^k) \\ y^{k+1} = \arg\min_{y\in\mathcal{Y}} \mathcal{L}_\gamma(x^{k+1}, y, \lambda^k) \\ \lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b), \end{cases} \tag{1.11}$$

where $\mathcal{L}_\gamma(x, y, \lambda)$ is the augmented Lagrangian function for problem (1.10) defined as:

$$\mathcal{L}_\gamma(x, y, \lambda) = f(x) + g(y) - \lambda^\top(Ax + By - b) + \frac{\gamma}{2}\|Ax + By - b\|^2. \tag{1.12}$$

The convergence of the ADMM for (1.10) is actually a consequence of the convergence of the so-called Douglas-Rachford operator splitting method (see [45, 33]). However, the rate of convergence for ADMM is established only very recently: [57] shows that for problem (1.10) the ADMM converges at the rate of $O(1/N)$ where N is the number of total iterations. From the perspective of monotone inclusion, a similar iteration complexity is also obtained in [80] under different assumptions. Moreover, a non-ergodic $O(1/N)$ iteration complexity in terms of the infeasibility measure and the objective value are found very recently in [56, 74, 26]. Furthermore, by imposing additional conditions on the objective function or constraints, the ADMM can be shown to converge linearly; see [48, 28, 59, 8, 73]. Moreover, as we will show later in this thesis, the ADMM framework can be naturally extended to solve problems with more than two blocks of variables.

Besides the multi-block structure, we can take a different stance towards the applicability of the ADMM, depending on the prevailing information structure of the problem. Observe that to implement (1.11), it is necessary that $\arg\min_{x\in\mathcal{X}} \mathcal{L}_\gamma(x, y^k, \lambda^k)$ and $\arg\min_{y\in\mathcal{Y}} \mathcal{L}_\gamma(x^{k+1}, y, \lambda^k)$ can be solved efficiently at each iteration; i.e. the proximal mappings are assumed to be easy. While this is indeed the case for some classes of the problems (e.g. the lasso problem), it may also fail for many other applications. This triggers a natural question: Given the structure of the objective functions in the mini-

mization subroutines, can the multipliers' method be adapted accordingly? In Chapter 2, we study some variants of the ADMM based on incorporating the two basic first-order methods (1.2) and (1.5) to account for this informational structure of the objective functions. One possible scenario is that it is not easy to solve the subproblems of x and y in (1.11), in this case, we can introduce the gradient projection method (a special form of proximal gradient method) to replace the exact minimization subproblem. In particular, this leads to two algorithms: one is the GADM (Gradient-ADM, replacing the subproblem of one block by gradient method) and the other is GALM (Gradient-ALM, replacing the subproblems of both blocks by gradient method), and those two algorithms will further allow us to deal with different informational structure of the problem.

To take into account the informational structure, one natural way is to consider the stochastic setting of problem (1.10), where we can go beyond the deterministic informational structure of the problem. In stochastic programming (SP), the objective function is often in the form of expectation. In this case, even requesting its full gradient information is impractical. Historically, Robbins and Monro [102] introduced the so-called stochastic approximation (SA) approach to tackle this problem. Polyak and Juditsky [96, 97] proposed an SA method in which larger step-sizes are adopted and the asymptotical optimal rate of convergence is achieved; cf. [34, 37, 105, 104] for more details. Recently, there has been a renewed interest in SA, in the context of computational complexity analysis for convex optimization [85], which has focussed primarily on bounding the number of iterations required by the SA-type algorithms to ensure the expectation of the objective to be ϵ away from optimality. For instance, Nemirovski et al. [83] proposed a mirror descent SA method for the general nonsmooth convex stochastic programming problem attaining the optimal convergence rate of $O(1/\sqrt{N})$; Lan and his coauthors [43, 41, 40, 42, 68, 44] proposed various first-order methods for SP problems under suitable convex or non-convex settings. In [93], a stochastic version of problem (1.10) is considered. In this proposal we also consider the GADM and the GALM in the SP framework. Under the stochastic framework, the informational structure of the problem appears in a progressive way. To start with, we first assume that a noisy gradient information of the function is available. Thus, in our GADM method or GALM method, we can only use a noisy stochastic estimate of the gradient and we name those methods as SGADM (stochastic-ADM) and SGALM (stochastic-ALM). Furthermore, it is possible that even the noisy gradient information is not available. In

this case, we assume that we can only get the noisy estimation of the function value, and this gives us an avenue where the zeroth-order method can apply. Inspired by the work of Nesterov [89] for gradient-free minimization, we will propose a zeroth-order (gradient-free, a.k.a. direct) smoothing method for (1.10). Specifically, we show that the SGADM and the SGALM can be extended to the zeroth-order version by incorporating the zeroth-order gradient estimate into the algorithms.

So far we have discussed the different variants of ADMM for solving problem (1.10) which is a special case of the problem (1.7). Is it possible to extend the ADMM framework to the more general problem (1.7) where we have both multi-block variables and coupled the objective function $f(x_1, x_2, \ldots, x_m)$? The answer is positive, and we will discuss this thoroughly in Chapter 3 and Chapter 4. In fact, utilizing the multi-block structure of the problem, the multi-block ADMM updates the block variables sequentially. Specifically, it performs the following updates iteratively (by assuming the absence of the coupled functions f):

$$\begin{cases} x_1^{k+1} &= \arg\min_{x_1 \in \mathcal{X}_1} \mathcal{L}_\rho(x_1, x_2^k, \cdots, x_m^k, \lambda^k), \\ &\vdots \\ x_m^{k+1} &= \arg\min_{x_m \in \mathcal{X}_m} \mathcal{L}_\rho(x_1^{k+1}, \cdots, x_{m-1}^{k+1}, x_m, \lambda^k), \\ \lambda^{k+1} &= \lambda^k - \gamma(A_1 x_1^{k+1} + A_2 x_2^{k+1} + \cdots + A_m x_m^{k+1} - b), \end{cases} \tag{1.13}$$

where the augmented Lagrangian function is similarly defined as:

$$\mathcal{L}_\gamma(x, \lambda) = \sum_{i=1}^m u_i(x_i) - \lambda^\top \left(\sum_{i=1}^m A_i x_i - b \right) + \frac{\gamma}{2} \left\| \sum_{i=1}^m A_i x_i - b \right\|^2. \tag{1.14}$$

Although the multi-block ADMM scheme in (1.13) performs very well for many instances encountered in practice (e.g. [94, 116]), it may fail to converge for some instances if there are more than 2 blocks of variables, i.e., $m \geq 3$. In particular, an example was presented in [16] to show that the ADMM may even diverge with 3 blocks of variables, when solving a linear system of equations. Thus, some additional assumptions or modifications will have to be in place to ensure convergence of the multi-block ADMM. In fact, by incorporating some extra correction steps or changing the Gauss-Seidel updating rule, [27, 55, 54, 52, 124] show that the convergence can still be achieved for the multi-block ADMM. Moreover, if some part of the objective function is strongly convex or the objective has certain regularity property, then it can be shown

that the convergence holds under various conditions; see [17, 74, 13, 73, 70, 47, 123]. Using some other conditions including the error bound condition and taking small dual stepsizes, or by adding some perturbations to the original problem, authors of [59, 75] establish the rate of convergence results even without strong convexity. Not only for the problem with linear constraint, in [19, 71, 112] multi-block ADMM are extended to solve convex linear/quadratic conic programming problems. In a recent work [113], the authors propose a randomly permuted ADMM (RP-ADMM) that basically chooses a random permutation of the block indices and performs the ADMM update according to the order of indices in that permutation, and they show that the RP-ADMM converges in expectation for solving non-singular square linear system of equations.

In [58], the authors propose a block successive upper bound minimization method of multipliers (BSUMM) to solve problem (1.7). Essentially, at every iteration, the BSUMM replaces the nonseparable part $f(x)$ by an upper-bound function and works on that modified function in an ADMM manner. Under some error bound conditions and a diminishing dual stepsize assumption, the authors are able to show that the iterates produced by the BSUMM algorithm converge to the set of primal-dual optimal solutions. Along a similar direction, Cui et al. [23] introduces a quadratic upper-bound function for the nonseparable function f to solve 2-block problems; they show that their algorithm has an $O(1/N)$ convergence rate, where t is the number of total iterations. Moreover, [18] shows the convergence of the ADMM for 2-block problems by imposing quadratic structure on the coupled function $f(x)$ and also the convergence of RP-ADMM for multi-block case where all separable functions vanish (i.e. $u_i(x_i) = 0$, $\forall i$).

In Chapter 3, we study the ADMM and its variants for (1.7). (Some adaptations of the ADMM are particularly relevant if there is a coupling term in the objective, as the minimization subroutines required by the ADMM may become difficult to implement; see more discussions on this later.) Instead of using some upper-bound approximation (a.k.a. majorization-minimization), we work with the original objective function. In some applications, it is difficult or impossible to implement the ADMM iteration, because the augmented Lagrangian function in (1.11) may be difficult to optimize even if the other block of variables and the Lagrangian multipliers are fixed. This motivates us to propose the *Alternating Proximal Gradient Method of Multipliers* (APGMM), which essentially iterates between proximal gradient methods of each block variables before the multiplier is updated. If optimizing the augmented Lagrangian function for one block of variables is easy while optimizing the other block of variables is difficult, then a hybrid

between ADMM and APGMM is a natural choice. What if the gradient proximal subroutines are still too difficult to be implemented? One would then opt to compute the gradient projections. Hence, we propose the *Alternating Gradient Projection Method of Multipliers* (AGPMM), which replaces the proximal gradient steps in APGMM by the gradient projections. At this stage, all the methods mentioned above are considered in the context of the 2-block model. In general however, they can be extended to the multi-block model with a coupling term.

In Chapter 4, we study a more general model that includes (1.7) as a special case. Specifically, we introduce another set of variables y which has a similar mixed structure (coupled and separate) as x in (1.7). We propose a randomized primal-dual coordinate update algorithm by introducing randomization to the multi-block ADMM framework (1.13). Different from the random permutation scheme in [113, 18], a simpler variable selection rule based on uniform distribution is used. The randomization is crucial to the convergence of the algorithm. In fact, the randomization scheme enables us to establish the $O(1/N)$ convergence rate for this coupled multi-block model with mere convexity. The way we deal with the coupled objective function is to use proximal gradient approach, and it has two additional benefits. First, by incorporating a properly chosen proximal term, the variables can be decoupled and the algorithm can be done in parallel. Furthermore, the proximal gradient (linearization) scheme can be adapted to solve stochastic optimization problems. In fact, based on the informational structure of the coupled function f, an approximation of the gradient can be obtained. The randomized primal-dual coordinate update can still be readily implemented, given an unbiased approximation is available under some oracle-based models.

Besides the multi-block optimization problem, low-order algorithms are also suitable for solving black-box optimization and bandit online learning problem. For both of the problems, the key feature is that no higher-order information is available other than the function value estimation. In optimization, solution procedures using only objective values are often referred to as *direct methods*. In [99], Powell (1964) constructed a method based on conjugate directions for quadratic minimization; in [82], Nelder and Mead (1965) introduced the so-called simplex method for nonlinear optimization, while justifications for the simplex method in low dimensions can be found in [67, 66]. A modern account and historical notes of the direct methods can be found in a recent book ([21]) by Conn, Scheinberg, and Vicente. Our study however, builds on a relatively recent approach of *randomized* zeroth-order approximation of the gra-

dient, pioneered by Nesterov and Spokoiny [89]; some of the ideas in the approach can be traced back to Polyak [98]. As a different type of optimization method, Bayesian optimization [63] also provides a powerful tool for black-box optimization. In general, Bayesian optimization constructs a prior probabilistic distribution over the functional space. When the data are observed, it sequentially refines this model using Bayesian posterior distribution. The procedure finds the next query point by maximizing an acquisition function induced by the corresponding probabilistic model. For more details and applications about Bayesian Optimization, see, e.g., [12, 108, 111]. From a different perspective, there has been a great research interest in using the lower-order method for online learning problem. Several sub-linear cumulative regret bounds measured by stationary regret have been established in various papers in the literature. For example, [131] proposed an online gradient descent algorithm which achieves an regret bound of order $O(\sqrt{T})$ for convex loss functions. The order of the regret can be further improved to $O(\log T)$ if the loss functions are strongly convex (see [49]). Moreover, the bounds are shown to be tight for convex / strongly convex loss functions respectively in [1]. In the so-called *bandit online convex optimization*, where the online player is only supposed to know the function value $f_t(x_t)$ at x_t, instead of the entire function $f_t(\cdot)$. When the player can only observe the function value at a single point, [35] established an $O(T^{3/4})$ regret bound for general convex loss functions by constructing a zeroth-order approximation of the gradient. Assuming that the loss functions are smooth, the regret bound can be improved to $O(T^{2/3})$ by incorporating a self-concordant regularizer (see [106]). Alternatively, if multiple points can be inquired at the same time, [2] showed that the regrets can be further improved to $O(T^{1/2})$ and $O(\log T)$ for convex / strongly convex loss functions respectively. In this dissertation, we study various zeroth-order algorithms for the black-box optimization and online learning. In particular, we study the optimization models under three different settings. In the first setting, the model is basically stochastic programming with the following side restriction: similar to the online bandit learning framework, only one random sample can be drawn at any given decision point. In the second setting, we present a general nonconvex optimization framework (weakly pseudo convex), and develop a specialized zeroth-order normalized gradient method. In the third setting, the objective value can be estimated arbitrarily close to the true value, at a cost that is increasing with regard to the inverse of the precision desired. Furthermore, we extend the analysis to a general constrained model with a composite objective function, consisting of the original objective and a non-smooth

regularizer. The above-mentioned settings are considered respectively for that general case as well, extending the sample complexity analysis under a proximal gradient dominance assumption. In Chapter 6, we further extend the similar idea to the stochastic bandit online learning problem, where the nonsmoothness of the loss function and the one random sample scheme are discussed.

1.2 Overview and Organization

In Chapter 2, we present a suite of variants of the ADMM, including GADM, GALM, SGADM, SGALM, and the zeroth-order version of SGADM and SGALM. Clearly, the new variants allow the method to be applicable on a much broader class of problems where only noisy estimations of the gradient or the function values are accessible, yet the flexibility is achieved without sacrificing the computational complexity bounds. In fact, we will show that the rate of convergence of GADM and SGADM would be $O(1/N)$ and $O(1/\sqrt{N})$ respectively, and show SGALM admits a similar iteration complexity bound. Moreover, we will show that the zeroth-order SGADM and SGALM also have the $O(1/\sqrt{N})$ complexity.

In Chapter 3, we first study the 2-block case of the optimization model (1.7), where we analyze the proposed first-order algorithms to solve this model. First, the ADMM is extended, assuming that it is easy to optimize the augmented Lagrangian function with one block of variables at each time while fixing the other block. We prove that $O(1/N)$ iteration complexity bound holds under suitable conditions. If the subroutines of the ADMM cannot be implemented, then our APGMM, AGPMM, and the hybrids of them may be still applicable. Under suitable conditions, the $O(1/N)$ iteration complexity bound is shown to hold for all the newly proposed algorithms. Finally, we extend the analysis for the ADMM to the general multi-block case.

In Chapter 4, we propose a randomized primal-dual proximal block coordinate updating framework for a general multi-block convex optimization model with coupled objective function and linear constraints. Assuming mere convexity, we establish its $O(1/N)$ convergence rate in terms of the objective value and feasibility measure. Our analysis recovers and/or strengthens the convergence properties of several existing algorithms. In particular, Our result shows that a sublinear rate of convergence in expectation can be guaranteed for multi-block ADMM, without assuming any strong convexity. The new approach is also extended to solve problems where only a stochastic approxi-

mation of the (sub-)gradient of the objective is available, and we establish an $O(1/\sqrt{N})$ convergence rate of the extended approach for solving stochastic programming.

In Chapter 5, we first study an unconstrained stochastic optimization model where the objective can be allowed a single-sample at a point. The convergence analysis also extends to the star-convex functions. Moreover, we consider a class of nonconvex optimization model by introducing the so-called weak pseudo-convexity. For this model, we develop a zeroth-order normalized gradient descent method. For the aforementioned two models, we show the sublinear convergence rate of our zeroth-order methods. In addition, we study unconstrained optimization where only the objective function can be estimated, and the efforts required to estimate the function value depends on the precision. Finally, we extend our investigations to the constrained optimization with a regularization function. Linear convergence rate is derived for the latter two models under strong convexity and gradient dominance respectively.

In Chapter 6, we present the zeroth-order methods for solving online learning problem. Specifically, we study the online convex optimization with stochastic loss functions. The goal is to design some effective algorithms such that the total regret will be bounded above nontrivially by the time horizon T. In fact, we propose a stochastic gradient descent method under this setting. We prove the $O(\sqrt{T})$ regret bound for both smooth and non-smooth loss functions and the $O(T^{\frac{3}{4}})$ regret bound for non-smooth stochastic loss with one random sample restriction.

Chapter 2

Information-Adaptive Variants of the ADMM

2.1 Introduction

In this chapter, we consider the most basic ADMM model:

$$\begin{array}{ll} \min & f(x) + g(y) \\ \text{s.t.} & Ax + By = b, \\ & x \in \mathcal{X}, y \in \mathcal{Y} \end{array} \tag{2.1}$$

where $x \in \mathbb{R}^{n_x}$, $y \in \mathbb{R}^{n_y}$, $A \in \mathbb{R}^{m \times n_x}$, $B \in \mathbb{R}^{m \times n_y}$, $b \in \mathbb{R}^m$, and $\mathcal{X} \subseteq \mathbb{R}^{n_x}$, $\mathcal{Y} \subseteq \mathbb{R}^{n_y}$ are closed convex sets; f is a smooth convex function, and g is a convex function and possibly nonsmooth. We further assume that the gradient of f is Lipschitz continuous:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \ \forall\, x, y \in \mathcal{X}, \tag{2.2}$$

where L is a Lipschitz constant.

Recall, the augmented Lagrangian function for problem is defined as:

$$\mathcal{L}_\gamma(x, y, \lambda) = f(x) + g(y) - \lambda^\top (Ax + By - b) + \frac{\gamma}{2}\|Ax + By - b\|^2. \tag{2.3}$$

To bring out the hierarchy regarding the available information of the functions in question, let us first introduce the following definition.

Definition 1 *We call a convex function $f(x)$ to be* easy to minimize *with respect to x (f is hence said to be* **MinE** *as an abbreviation) if there exits some $H \succeq 0$ such that the proximal mapping $\arg\min_x f(x) + \frac{1}{2}\|x - z\|_H^2$ can be computed easily for any given z.*

Some remarks are in order here. If $\mathcal{L}_\gamma(x, y, \lambda)$ is **MinE** with respect to both x and y with $H = 0$, then the original ADMM (1.11) is readily applicable. For the cases where $\mathcal{L}_\gamma(x, y, \lambda)$ is **MinE** with respect to both x and y but H is nonzero, the convergence of different inexact ADMM-type methods have been studied in [51, 92, 53]. Moreover, [110, 15, 59] show that by incorporating various modifications of ADMM with the proximal method an $O(1/N)$ convergence rate can still be achieved. In the case that $\mathcal{L}_\gamma(x, y, \lambda)$ is **MinE** in x but not in y, Lin, Ma and Zhang [72] recently proposed an extra-gradient ADMM (EGADM) and showed an $O(1/N)$ iteration bound. In this chapter, we consider a simpler procedure of applying gradient only once in each iteration (to be named GADM):

$$\begin{cases} y^{k+1} = \arg\min_{y\in\mathcal{Y}} \mathcal{L}_\gamma(x^k, y, \lambda^k) + \frac{1}{2}\|y - y^k\|_H^2 \\ x^{k+1} = [x^k - \nabla_x \mathcal{L}_\gamma(x^k, y^{k+1}, \lambda^k)]_{\mathcal{X}} \\ \lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b), \end{cases} \tag{2.4}$$

where $[x]_{\mathcal{X}}$ denotes the projection of x onto $\mathcal{X}$. In fact, a variant of above procedure was considered in [72], where x is updated by taking the gradient of Lagrangian function rather than augmented Lagrangian function, and it was posed as an unsolved problem to determine the iteration complexity bound of this modified algorithm. In this chapter we prove that the GADM also has an iteration bound of $O(1/N)$. Moreover, our analysis does not require the optimal set to be bounded nor the coercivity of the objective function, which is key to obtaining iteration bounds in many previous works; see [32, 28, 72, 84]. In addition to the assumptions made at the beginning of this section, throughout the chapter we only assume:

Assumption 2.1.1 *The optimal solution set $\mathcal{X}^* \times \mathcal{Y}^*$ for problem* (2.1) *is non-empty.*

Under this assumption, naturally we have $\mathrm{dist}(x, \mathcal{X}^*)_M := \min_{u\in\mathcal{X}^*} \|x-u\|_M < \infty$, and $\mathrm{dist}(y, \mathcal{Y}^*)_M := \min_{v\in\mathcal{Y}^*} \|y - v\|_M < \infty$, for any given x, y and matrix $M \succeq 0$.

In this chapter we also consider (2.4) in the SP framework. We assume that a noisy gradient information of $\nabla\mathcal{L}_\gamma$ via the so-called *stochastic first order oracle* ($\mathcal{SFO}$) is available. Specifically, for a given x, instead of computing $\nabla f(x)$ we actually only get

a stochastic gradient $G(x,\xi)$ from the $\mathcal{SFO}$, where ξ is a random variable following a certain distribution. Formally we introduce:

Definition 2 *We call a function $f(x)$ to be* easy for gradient estimation *– denoted as* **GraE** *– if there is an $\mathcal{SFO}$ for f, which returns a stochastic gradient estimation $G(x,\xi)$ for ∇f at x, satisfying*

$$\mathsf{E}[G(x,\xi)] = \nabla f(x), \tag{2.5}$$

and

$$\mathsf{E}[\|G(x,\xi) - \nabla f(x)\|^2] \le \sigma^2. \tag{2.6}$$

If the exact gradient information is available then the $\mathcal{SFO}$ is deterministic. In general, the $\mathcal{SFO}$ is often stochastic and inaccurate. For instance, the stochastic gradient that is used in many machine learning algorithms can be viewed as a special case of $\mathcal{SFO}$ where the distribution is uniform on the dataset. As for problem (2.1), quite a few ADMM variants in stochastic and online optimization setting have been proposed recently; see [93, 121, 114, 115, 129, 130]. The basic idea in those works is to linearize the stochastic function and use the noisy gradient in the subproblem. In [93, 121, 114, 129], the $O(1/\sqrt{N})$ and $O(\ln N/N)$ iteration complexities have been shown for general convex function and strongly convex function respectively. Moreover, [130] shows an $O(1/N)$ iteration complexity can be achieved if an incremental approximation of the full gradient is used. By assuming both functions are strongly convex and smooth, a linear convergence is shown in [115]. In this chapter, when $\mathcal{L}_\gamma(x,y,\lambda)$ is **MinE** with respect to y, and $f(x)$ in (2.3) is **GraE**, we will then propose a stochastic gradient ADMM (SGADM), which alternates through one exact minimization step ADMM (1.11), one stochastic approximation iteration, and an update on the dual variables (multipliers). It is clear that the SGADM in the deterministic case is exactly GADM (2.4), and we will show that the rate of convergence of GADM and SGADM would be $O(1/N)$ and $O(1/\sqrt{N})$ respectively. n particular, if $f(x)$ is strongly convex, the complexity of SGADM can be improved to $O(\ln N/N)$. Moreover, if $f(x)$ and $g(y)$ in (2.3) are both **GraE**, then we propose a stochastic gradient augmented Lagrangian method (SGALM), and show that it admits a similar iteration complexity bound.

Furthermore, we are also interested in another class of stochastic problems, where even the noisy gradient information is not available; instead we assume that we can only get a noisy estimation of f via the so-called *stochastic zeroth-order oracle* ($\mathcal{SZO}$). Specifically, for any input x, by calling $\mathcal{SZO}$ once it returns a quantity $\mathcal{F}(x,\xi)$, which

is a noisy approximation of the true function value $f(x)$. More specifically,

Definition 3 *We call a function $f(x)$ to be* easy for function evaluation *– denoted as* **ValE** *– if there is an $\mathcal{SZO}$ for f, which returns a stochastic estimation for f at x if $\mathcal{SZO}$ is called, satisfying*

$$\mathsf{E}[\mathcal{F}(x,\xi)] = f(x), \tag{2.7}$$

$$\mathsf{E}[\nabla\mathcal{F}(x,\xi)] = \nabla f(x), \tag{2.8}$$

and

$$\mathsf{E}[\|\nabla\mathcal{F}(x,\xi) - \nabla f(x)\|^2] \leq \sigma^2. \tag{2.9}$$

Inspired by the work of Nesterov [89] for gradient-free minimization, in this chapter we will propose a zeroth-order (gradient-free, a.k.a. direct) smoothing method for (2.1). Instead of using the Gaussian smoothing scheme as in [89], which has an unbounded support set, we apply another smoothing scheme based on the $\mathcal{SZO}$ of f. To be specific, when $\mathcal{L}_\gamma(x,y,\lambda)$ is **MinE** with respect to y, and $f(x)$ in (2.3) is **ValE**, we will propose a zeroth-order gradient augmented Lagrangian method (zeroth-order GADM) and analyze its complexity. To summarize, according to the available informational structure of the objective functions, in this chapter we present suitable variants of the ADMM to account for the available information. In a nutshell, the details are in the following Table 2.1.

		Block x		
		MinE	**GraE**	**ValE**
Block y	**MinE**	ADMM	SGADM	zeroth-order GADM
	GraE	SGADM	SGALM	zeroth-order SGADM
	ValE	zeroth-order GADM	zeroth-order SGADM	zeroth-order GALM

Table 2.1: A summary of informational-hierarchic alternating direction of multiplier methods.

The rest of the chapter is organized as follows. In Section 2.2, we propose the stochastic gradient ADMM (SGADM) algorithm, and analyze its complexity. In Section 2.3, we present our stochastic gradient augmented Lagrangian method (SGALM) which uses gradient projection in both block variables, and analyze its convergence rate. In Section 2.4, we propose a zeroth-order GADM through a new smoothing scheme, and present a complexity result. Finally, we present some numerical experiment results in Section 2.5.

2.2 The Stochastic Gradient Alternating Direction of Multipliers

In this section, we assume $\mathcal{L}_\gamma(x, y, \lambda)$ to be **MinE** with respect to y, and $f(x)$ to be **GraE**. That is, for a given x, whenever we need $\nabla f(x)$, we can actually get a stochastic gradient $G(x, \xi)$ from the $\mathcal{SFO}$, where ξ is a random variable following a certain distribution. Moreover, $G(x, \xi)$ satisfies (2.5) and (2.6). By the definition of the augmented Lagrangian $\mathcal{L}_\gamma(x, y, \lambda)$, an $\mathcal{SFO}$ for $\mathcal{L}_\gamma(x, y, \lambda)$ can be constructed accordingly:

Definition 4 *Denote the $\mathcal{SFO}$ of $\nabla_x \mathcal{L}_\gamma(x, y, \lambda)$ as $G_L(x, y, \lambda, \xi)$, which is defined as:*

$$G_L(x, y, \lambda, \xi) := G(x, \xi) - A^\top \lambda + \gamma A^\top (Ax + By - b). \tag{2.10}$$

One example where such application arises is *stochastic lasso* problem:

$$\min_x \ \frac{1}{2}\mathsf{E}_\xi (a_\xi^\top x - b_\xi)^2 + \mu \|x\|_1,$$

where the sensing vector a_ξ as well as the sensing result b_ξ are given stochastically. The problem can be formulated as

$$\begin{array}{ll} \min & \frac{1}{2}\mathsf{E}_\xi (a_\xi^\top x - b_\xi)^2 + \mu \|y\|_1 \\ \text{s.t.} & x - y = 0. \end{array}$$

Assuming each time one sample is observed, we have $G(x, \xi) = a_\xi (a_\xi^\top x - b_\xi)$.

Our first algorithm to be introduced, SGADM, works as follows:

The Stochastic Gradient ADMM (SGADM)

Initialize $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}$ and λ^0

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

$y^{k+1} = \arg\min_{y \in \mathcal{Y}} \mathcal{L}_\gamma(x^k, y, \lambda^k) + \frac{1}{2}\|y - y^k\|_H^2$;

$x^{k+1} = [x^k - \alpha_k G_L(x^k, y^{k+1}, \lambda^k, \xi^{k+1})]_{\mathcal{X}}$;

$\lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)$.

end for

In the above notation, $[x]_{\mathcal{X}}$ denotes the projection of x onto $\mathcal{X}$, H is a pre-specified positive semidefinite matrix, α_k is the stepsize for the k-th iteration. In fact, matrix H is often used to cancel out the quadratic cross terms in the augmented Lagrangian,

in order for the resulting subproblem to be separable, or even to admit a closed-form solution. In our proposed algorithms, the H matrix can be set to 0 which recovers the original ADMM subproblem. It is easy to see that the deterministic version of SGADM is exactly GADM (2.4). The above SGADM is similar to the stochastic ADMM proposed in [93], where an $O(1/\sqrt{N})$ iteration complexity is shown. The difference lies in the fact that the SGADM linearizes the whole augmented Lagrangian and performs a gradient projection, while [93] linearizes the objective function in the augmented Lagrangian and minimizes the resulting function. We note that the OPG-ADMM in [114] also linearizes the whole augmented Lagrangian, but the order of updating blocks is different. In OPG-ADMM, it first updates the block with a gradient-type step and then updates the other block by exact minimization, while in our algorithm the order is reversed. In the following subsection, based on the measure of the constraint violation and the gap of objective value, we will show that the complexity of SGADM is $O(1/\sqrt{N})$ and the complexity of GADM is $O(1/N)$. Furthermore, if the function f is strongly convex, it can be shown that the complexity of SGADM is indeed $O(\ln N/N)$.

2.2.1 Convergence Rate Analysis of the SGADM

In this subsection, we shall analyze the convergence rate of SGADM algorithm. First, some notations and preliminaries are introduced to facilitate the discussion.

Preliminaries and Notations

Denote

$$u = \begin{pmatrix} y \\ x \end{pmatrix}, \quad w = \begin{pmatrix} y \\ x \\ \lambda \end{pmatrix}, \quad F(w) = \begin{pmatrix} -B^\top \lambda \\ -A^\top \lambda \\ Ax + By - b \end{pmatrix}, \tag{2.11}$$

$h(u) = f(x) + g(y)$, $\Omega = \mathcal{Y} \times \mathcal{X} \times \mathbb{R}^m$, and

$$Q_k = \begin{pmatrix} H & 0 & 0 \\ 0 & \frac{1}{\alpha_k} I_{n_x} & 0 \\ 0 & -A & \frac{1}{\gamma} I_m \end{pmatrix}, \quad P = \begin{pmatrix} I_{n_y} & 0 & 0 \\ 0 & I_{n_x} & 0 \\ 0 & -\gamma A & I_m \end{pmatrix}, \quad M_k = \begin{pmatrix} H & 0 & 0 \\ 0 & \frac{1}{\alpha_k} I_{n_x} & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix}. \tag{2.12}$$

Clearly, $Q_k = M_k P$. In addition to the sequence $\{w^k\}$ generated by the SGADM, we introduce an auxiliary sequence:

$$\tilde{w}^k := \begin{pmatrix} \tilde{y}^k \\ \tilde{x}^k \\ \tilde{\lambda}^k \end{pmatrix} = \begin{pmatrix} y^{k+1} \\ x^{k+1} \\ \lambda^k - \gamma(Ax^k + By^{k+1} - b) \end{pmatrix}. \tag{2.13}$$

Based on (2.13) and (2.12), the relationship between the new sequence $\{\tilde{w}^k\}$ and the original $\{w^k\}$ is

$$w^{k+1} = w^k - P(w^k - \tilde{w}^k). \tag{2.14}$$

The above succinct notations and analysis framework were originally introduced and used by He and Yuan in [57]. In this chapter, we adopt the same framework for analysis following that of [57]; in other words, our convergence result is also based on the auxiliary sequence $\tilde{w}^k$. Moreover, we denote $\delta_k \equiv G(x^{k-1}, \xi^k) - \nabla f(x^{k-1})$, which is the error of the noisy gradient generated by $\mathcal{SFO}$. The following lemma is straightforward.

Lemma 2.2.1 *For any $w^0, w^1, \cdots, w^{N-1}$, let F be defined in* (2.11) *and $\bar{w} = \frac{1}{N}\sum_{k=0}^{N-1} w^k$; then it holds*

$$(\bar{w} - w)^\top F(\bar{w}) = \frac{1}{N}\sum_{k=0}^{N-1}(w^k - w)^\top F(w^k).$$

Proof. Since $F(w) = \begin{pmatrix} 0 & 0 & -B \\ 0 & 0 & -A \\ A & B & 0 \end{pmatrix} \begin{pmatrix} y \\ x \\ \lambda \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ b \end{pmatrix}$, for any w_1 and w_2 we have

$$(w_1 - w_2)^\top (F(w_1) - F(w_2)) = 0. \tag{2.15}$$

Therefore,

$$\begin{aligned}(\bar{w}-w)^\top F(\bar{w}) &\overset{(2.15)}{=} (\bar{w}-w)^\top F(w)\\ &= \left(\frac{1}{N}\sum_{k=0}^{N-1} w^k - w\right)^\top F(w)\\ &= \frac{1}{N}\sum_{k=0}^{N-1}(w^k-w)^\top F(w)\\ &\overset{(2.15)}{=} \frac{1}{N}\sum_{k=0}^{N-1}(w^k-w)^\top F(w^k). \end{aligned} \tag{2.16}$$

□

The Complexity of SGADM without Strong Convexity

We now present the rate of convergence result for SGADM, which is $O(1/\sqrt{N})$. Denote $\Xi_k = (\xi_1, \xi_2, \ldots, \xi_k)$. In fact, the convergence rate is in the sense of the expectation taken over Ξ_k.

Theorem 2.2.2 *Suppose that $\mathcal{L}_\gamma(x, y, \lambda)$ is* **MinE** *with respect to y, and $f(x)$ is* **GraE**. *Given a fixed iteration number N, letting w^k be the sequence generated by the SGADM, and choosing $\eta_k = \sqrt{N}$, and $C > 0$ be a constant satisfying $CI_{n_x} - \gamma A^\top A - LI_{n_x} \succeq 0$, and $\alpha_k = \frac{1}{\eta_k + C} = \frac{1}{\sqrt{N}+C}$. Let*

$$\bar{w}_n := \frac{1}{n}\sum_{k=0}^{n-1}\tilde{w}^k, \tag{2.17}$$

where $\tilde{w}^k$ is defined in (2.13). *Then the following holds*

$$\begin{aligned}&\mathsf{E}_{\Xi_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right]\\ \leq\ & \frac{1}{2\sqrt{N}}(\sigma^2 + D_x^2) + \frac{1}{2N}\left(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right),\end{aligned} \tag{2.18}$$

where $D_x = \mathrm{dist}(x_0, \mathcal{X}^*)$, $D_{y,H} = \mathrm{dist}(y_0, \mathcal{Y}^*)_H$ and ρ is any given positive number.

As in [57], we first present a bound regarding the sequence $\{\tilde{w}^k\}$ in (2.13).

Proposition 2.2.3 *Let $\{\tilde{w}^k\}$ be defined by* (2.13), *and the matrices Q_k, M_k, and P be*

given in (2.12). *For any* $w \in \Omega$, *we have*

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq \; & (w - \tilde{w}^k)^\top Q_k(w^k - \tilde{w}^k) - (x - x^k)^\top \delta_{k+1} - \frac{\|\delta_{k+1}\|^2}{2\eta_k} - \frac{\eta_k + L}{2}\|x^k - \tilde{x}^k\|^2, \end{aligned} \tag{2.19}$$

where $\eta_k > 0$ *is any constant. Moreover, for any* $w \in \Omega$, *the term* $(w-\tilde{w}^k)^\top Q_k(w^k-\tilde{w}^k)$ *on the RHS of* (2.19) *can be further bounded below as follows*

$$\begin{aligned} & (w - \tilde{w}^k)^\top Q_k(w^k - \tilde{w}^k) \\ \geq \; & \frac{1}{2}\left(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2\right) + \frac{1}{2}(x^k - \tilde{x}^k)^\top \left(\frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A\right)(x^k - \tilde{x}^k). \end{aligned} \tag{2.20}$$

The proof of Proposition 2.2.3 involves several steps. In order not to distract the flow of presentation, we delegate its proof to the appendix.

Proof of Theorem 2.2.2

Proof. Recall that $CI_{n_x} - \gamma A^\top A - LI_{n_x} \succeq 0$ and $\alpha_k = \frac{1}{\eta_k + C}$. By (2.19) and (2.20),

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq \; & \frac{1}{2}\left(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2\right) + \frac{1}{2}(x^k - \tilde{x}^k)^\top \left(\frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A\right)(x^k - \tilde{x}^k) \\ & -(x - x^k)^\top \delta_{k+1} - \frac{\|\delta_{k+1}\|^2}{2\eta_k} - \frac{\eta_k + L}{2}\|x^k - \tilde{x}^k\|^2 \\ = \; & \frac{1}{2}\left(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2\right) \\ & + \frac{1}{2}(x^k - \tilde{x}^k)^\top \left(\frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A - (\eta_k + L) I_{n_x}\right)(x^k - \tilde{x}^k) \\ & -(x - x^k)^\top \delta_{k+1} - \frac{\|\delta_{k+1}\|^2}{2\eta_k} \\ \geq \; & \frac{1}{2}\left(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2\right) - (x - x^k)^\top \delta_{k+1} - \frac{\|\delta_{k+1}\|^2}{2\eta_k}. \end{aligned} \tag{2.21}$$

Using the definition of M_k, from (2.21) we have

$$\begin{aligned} & h(\tilde{u}^k) - h(u) + (\tilde{w}^k - w)^\top F(\tilde{w}^k) \\ \leq\ & \frac{1}{2}\left(\|y - y^k\|_H^2 - \|y - y^{k+1}\|_H^2\right) + \frac{1}{2\gamma}\left(\|\lambda - \lambda^k\|^2 - \|\lambda - \lambda^{k+1}\|^2\right) \\ & + \frac{\|x - x^k\|^2 - \|x - x^{k+1}\|^2}{2\alpha_k} + (x - x^k)^\top \delta_{k+1} + \frac{\|\delta_{k+1}\|^2}{2\eta_k}. \end{aligned} \tag{2.22}$$

Summing up the inequalities (2.22) for $k = 0, 1, \ldots, N-1$ we have

$$\begin{aligned} & h(\bar{u}_N) - h(u) + (\bar{w}_N - w)^\top F(\bar{w}_N) \\ \leq\ & \frac{1}{N}\sum_{k=0}^{N-1} h(\tilde{u}^k) - h(u) + \frac{1}{N}\sum_{k=0}^{N-1} (\tilde{w}^k - w)^\top F(\tilde{w}^k) \\ \leq\ & \frac{1}{2N}\sum_{k=0}^{N-1} \frac{\|x - x^k\|^2 - \|x - x^{k+1}\|^2}{\alpha_k} + \frac{1}{N}\sum_{k=0}^{N-1}\left[(x - x^k)^\top \delta_{k+1} + \frac{\|\delta_{k+1}\|^2}{2\eta_k}\right] \\ & + \frac{1}{2N}\left(\|y - y^0\|_H^2 + \frac{1}{\gamma}\|\lambda - \lambda^0\|^2\right), \end{aligned} \tag{2.23}$$

where the first inequality is due to the convexity of h and Lemma 2.2.1.

Note the above inequality is true for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $\lambda \in \mathbb{R}^m$, hence it is also true for any optimal solution x^*, y^*, and $\mathcal{B}_\rho = \{\lambda : \|\lambda\| \leq \rho\}$. As a result, by letting $w^* = (x^*, y^*, \lambda)^\top$, it follows that

$$\begin{aligned} & \sup_{\lambda \in \mathcal{B}_\rho}\left[h(\bar{u}_N) - h(u^*) + (\bar{w}_N - w^*)^\top F(\bar{w}_N)\right] \\ =\ & \sup_{\lambda \in \mathcal{B}_\rho}\Big[h(\bar{u}_N) - h(u^*) + (\bar{x}_N - x^*)^\top(-A^\top \bar{\lambda}_N) + (\bar{y}_N - y^*)^\top(-B^\top \bar{\lambda}_N) \\ & + (\bar{\lambda}_N - \lambda)^\top(A\bar{x}_N + B\bar{y}_N - b)\Big] \\ =\ & \sup_{\lambda \in \mathcal{B}_\rho}\left[h(\bar{u}_N) - h(u^*) + \bar{\lambda}_N^\top(Ax^* + By^* - b) - \lambda^\top(A\bar{x}_N + B\bar{y}_N - b)\right] \\ =\ & \sup_{\lambda \in \mathcal{B}_\rho}\left[h(\bar{u}_N) - h(u^*) - \lambda^\top(A\bar{x}_N + B\bar{y}_N - b)\right] \\ =\ & h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|, \end{aligned} \tag{2.24}$$

where $w^* = (x^*, y^*, \lambda)^\top$. Combining (2.23) and (2.24), we have

$$\begin{aligned} & h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\ \leq\ & \frac{1}{2N}\sum_{k=0}^{N-1}\frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k} + \frac{1}{N}\sum_{k=0}^{N-1}\left[(x^* - x^k)^\top\delta_{k+1} + \frac{\|\delta_{k+1}\|^2}{2\eta_k}\right] \\ & + \frac{1}{2N}\left(\|y^* - y^0\|_H^2 + \frac{1}{\gamma}\sup_{\lambda\in\mathcal{B}_\rho}\|\lambda - \lambda^0\|^2\right). \end{aligned} \tag{2.25}$$

Moreover, since $\alpha_k = \frac{1}{\eta_k + C} = \frac{1}{\sqrt{N}+C}$, it follows that

$$\begin{aligned} & \sum_{k=0}^{N-1}\frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k} \\ =\ & \sum_{k=0}^{N-1}(\sqrt{N}+C)(\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2) \\ \leq\ & (\sqrt{N}+C)\|x^* - x^0\|^2. \end{aligned} \tag{2.26}$$

Now, by plugging (2.26) into (2.25) and choosing x^*, y^* such that $D_x = \|x^* - x^0\|$ and $D_{y,H} = \|y^* - y^0\|_H$, it yields

$$\begin{aligned} & h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\ \leq\ & \frac{1}{N}\sum_{k=0}^{N-1}\left[(x^* - x^k)^\top\delta_{k+1} + \frac{\|\delta_{k+1}\|^2}{2\eta_k}\right] + \frac{1}{2\sqrt{N}}\|x^* - x^0\|^2 \\ & + \frac{1}{2N}\left(\|y^* - y^0\|_H^2 + C\|x^* - x^0\|^2 + \frac{1}{\gamma}\sup_{\lambda\in\mathcal{B}_\rho}\|\lambda - \lambda^0\|^2\right) \\ \leq\ & \frac{1}{N}\sum_{k=0}^{N-1}\left[(x^* - x^k)^\top\delta_{k+1} + \frac{\|\delta_{k+1}\|^2}{2\eta_k}\right] + \frac{D_x^2}{2\sqrt{N}} \\ & + \frac{1}{2N}\left(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right). \end{aligned} \tag{2.27}$$

Recall that $f(x)$ is **GraE**, so (2.5) and (2.6) hold. Consequently, $\mathsf{E}[\delta_{k+1}] = 0$. In addition, x_k is independent of ξ_{k+1}. Hence,

$$\mathsf{E}_{\Xi_{k+1}}[(x^* - x^k)^\top\delta_{k+1}] = 0. \tag{2.28}$$

Now, taking expectation over (2.27), and applying (2.6), we have

$$
\begin{aligned}
& \mathsf{E}_{\Xi_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\
\leq\ & \mathsf{E}_{\Xi_N}\left[\frac{1}{N}\sum_{k=0}^{N-1}((x^* - x^k)^\top \delta_{k+1} + \frac{\|\delta_{k+1}\|^2}{2\eta_k})\right] \\
& + \frac{D_x^2}{2\sqrt{N}} + \frac{1}{2N}\left(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right) \\
\overset{(2.6)}{\leq}\ & \frac{1}{N}\mathsf{E}_{\Xi_N}\left[\sum_{k=0}^{N-1}(x^* - x^k)^\top \delta_{k+1}\right] + \frac{\sigma^2}{2N}\sum_{k=0}^{N-1}\frac{1}{\eta_k} \\
& + \frac{D_x^2}{2\sqrt{N}} + \frac{1}{2N}\left(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right) \\
\overset{(2.41)}{=}\ & \frac{\sigma^2}{2N}\sum_{k=0}^{N-1}\frac{1}{\sqrt{N}} + \frac{D_x^2}{2\sqrt{N}} + \frac{1}{2N}\left(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right) \\
=\ & \frac{\sigma^2}{2\sqrt{N}} + \frac{D_x^2}{2\sqrt{N}} + \frac{1}{2N}\left(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right). \qquad (2.29)
\end{aligned}
$$

This completes the proof. □

Before concluding this section, some comments are in order here. Denoting $\hat{u}_N = \mathsf{E}_{\Xi_N}[\bar{u}_N]$, by Jensen's inequality it follows immediately that

$$
\begin{aligned}
& h(\hat{u}_N) - h(u^*) + \rho\|A\hat{x}_N + B\hat{y}_N - b\| \\
\leq\ & \frac{1}{2\sqrt{N}}(\sigma^2 + D_x^2) + \frac{1}{2N}\left(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right).
\end{aligned}
$$

That is to say, in the *ergodic sense*, in expectation the SGADM has a convergence rate of $O(1/\sqrt{N})$ when $f(x)$ is **GraE**. As we mentioned before, it is easy to slightly modify the proof for (2.18) to improve the complexity of GADM (i.e. the deterministic SGADM) to $O(1/N)$. In fact, when the exact gradient of f is available, σ in (2.6) and δ_k will be 0, and we can let $\eta_k = 1$ and constant stepsize $\alpha_k = \frac{1}{C+1}$. As a result,

$$
\sum_{k=0}^{N-1}\frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k} \leq (C+1)\|x^* - x^0\|^2.
$$

The iteration bound then improves to:

$$h(\hat{u}_N) - h(u^*) + \rho\|A\hat{x}_N + B\hat{y}_N - b\| \leq \frac{1}{2N}\left(D_{y,H}^2 + (C+1)D_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right), \tag{2.30}$$

and this establishes the $O(1/N)$ iteration complexity for the SGADM in the deterministic case. Moreover, in that case the stepsize α_k does not need to involve N at all.

Assuming the existence of the dual optimal solution λ^*, we can further assess the feasibility violation of the possibly infeasible solution $\hat{u}_N$ as in (2.30). Similar to Lemma 6 in [68] we introduce the following bound.

Lemma 2.2.4 *Assume that $\rho > 0$, and $\tilde{x} \in X$ is an approximate solution for the problem $f^* := \inf\{f(x) : Ax - b = 0, x \in X\}$ where f is convex, and X is a closed convex set, satisfying*

$$f(\tilde{x}) - f^* + \rho\|A\tilde{x} - b\| \leq \delta. \tag{2.31}$$

Suppose that an optimal Lagrange multiplier associated with the problem $\inf\{f(x) : Ax - b = 0, x \in X\}$ exists. Denote it to be y^*, satisfying $\|y^*\| < \rho$. Then, we have

$$\|A\tilde{x} - b\| \leq \frac{\delta}{\rho - \|y^*\|} \text{ and } f(\tilde{x}) - f^* \leq \delta$$

Proof. Define $v(u) := \inf\{f(x) : Ax - b = u, x \in X\}$, which is convex. Let y^* be such that $-y^* \in \partial v(0)$. Thus, we have

$$v(u) - v(0) \geq \langle -y^*, u\rangle \quad \forall u \in \mathbb{R}^m. \tag{2.32}$$

Let $u := A\tilde{x} - b$. Since $v(u) \leq f(\tilde{x})$ and $v(0) = f^*$, we have

$$-\|y^*\|\|u\| + \rho\|u\| \leq \langle -y^*, u\rangle + \rho\|u\| \leq v(u) - v(0) + \rho\|u\| \leq f(\tilde{x}) - f^* + \rho\|u\| \leq \delta.$$

Thus, $\|A\tilde{x} - b\| = \|u\| \leq \frac{\delta}{\rho - \|y^*\|}$, and $f(\tilde{x}) - f^* \leq \delta$. □

By (2.18) or (2.30), we know that the SGADM and the GADM achieve $h(\hat{u}_N) - h(u^*) + \rho\|A\hat{x}_N + B\hat{y}_N - b\| \leq \epsilon$ in $O(1/\epsilon^2)$ and $O(1/\epsilon)$ number of iterations respectively for any fixed $\rho > 0$. Lemma 2.2.4 further suggests that by choosing $\rho > \|\lambda^*\|$ we have

in fact established the error estimations

$$h(\hat{u}_N) - h(u^*) \leq O(\epsilon) \text{ and } \|A\hat{x}_N + B\hat{y}_N - b\| \leq O(\epsilon)$$

with the same iteration complexity. The same logic applies to all the subsequent convergence rate results.

2.2.2 The Complexity of SGADM under Strong Convexity

Under the assumption that f is strongly convex, the rate of convergence for SGADM can be improved to $O(\ln N/N)$. As before, the convergence rate is in the sense of the expectation taken over Ξ_k. Let's first introduce the notion of strong convexity.

Definition 5 *A function $f(x)$ is κ-strongly convex, if it satisfies the following*

$$f(y) \geq f(x) + \langle s, y - x\rangle + \frac{\kappa}{2}\|x - y\|^2 \quad \forall x, y \tag{2.33}$$

where $s \in \partial f(x)$ and $\partial f(x)$ is the subdifferential of f at x .

The main convengence rate result is presented as follows.

Theorem 2.2.5 *Suppose that $\mathcal{L}_\gamma(x, y, \lambda)$ is* **MinE** *with respect to y, $f(x)$ is* **GraE** *and κ-strongly convex. Let w^k be the sequence generated by the SGADM, and choose $\eta_k = (k+1)\kappa$, and $C > 0$ be a constant satisfying $CI_{n_x} - \gamma A^\top A - LI_{n_x} \succeq 0$, and $\alpha_k = \frac{1}{\eta_k + C} = \frac{1}{(k+1)\kappa + C}$. Let*

$$\bar{w}_n := \frac{1}{n}\sum_{k=0}^{n-1} \tilde{w}^k, \tag{2.34}$$

where $\tilde{w}^k$ is defined in (2.13)*. Then the following holds*

$$\begin{aligned} & \mathsf{E}_{\Xi_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\ \leq \; & \frac{\sigma^2(\ln N + 1)}{2\kappa N} + \frac{1}{2N}\left(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right), \end{aligned} \tag{2.35}$$

where $D_x = \operatorname{dist}(x_0, \mathcal{X}^*)$, $D_{y,H} = \operatorname{dist}(y_0, \mathcal{Y}^*)_H$ and ρ is any given positive number.

Proof. Similar as in the proof of Proposition 2.2.3, using the κ-strong convexity of f,

we conclude that

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq\ & (w - \tilde{w}^k)^\top Q_k(w^k - \tilde{w}^k) - (x - x^k)^\top \delta_{k+1} \\ & -\frac{\|\delta_{k+1}\|^2}{2\eta_k} - \frac{\eta_k + L}{2}\|x^k - \tilde{x}^k\|^2 + \frac{\kappa}{2}\|x - x^k\|^2, \end{aligned} \tag{2.36}$$

where and $\eta_k > 0$ is any constant and matrices Q_k, M_k, and P are given in (2.12). Similar to (2.21), by (2.36) and (2.20) we have,

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq\ & \frac{1}{2}\left(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2\right) - (x - x^k)^\top \delta_{k+1} - \frac{\|\delta_{k+1}\|^2}{2\eta_k} + \frac{\kappa}{2}\|x - x^k\|^2. \end{aligned} \tag{2.37}$$

Following a similar line of arguments as in Theorem 2.2.2, we derive that

$$\begin{aligned} & h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\ \leq\ & \frac{1}{2N}\sum_{k=0}^{N-1}\left(\frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k} - \kappa\|x^* - x^k\|^2\right) \\ & + \frac{1}{N}\sum_{k=0}^{N-1}\left[(x^* - x^k)^\top \delta_{k+1} + \frac{\|\delta_{k+1}\|^2}{2\eta_k}\right] + \frac{1}{2N}\left(\|y^* - y^0\|_H^2 + \frac{1}{\gamma}\sup_{\lambda \in \mathcal{B}_\rho}\|\lambda - \lambda^0\|^2\right). \end{aligned} \tag{2.38}$$

Since $\alpha_k = \frac{1}{\eta_k + C} = \frac{1}{(k+1)\kappa + C}$, it follows that

$$\begin{aligned} & \sum_{k=0}^{N-1}\left(\frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k} - \kappa\|x^* - x^k\|^2\right) \\ =\ & \sum_{k=0}^{N-1}\left((k\kappa + C)\|x^* - x^k\|^2 - ((k+1)\kappa + C)\|x^* - x^{k+1}\|^2\right) \\ \leq\ & C\|x^* - x^0\|^2. \end{aligned} \tag{2.39}$$

Plugging (2.39) into (2.25) and choosing x^*, y^* such that $D_x = \|x^* - x^0\|$ and $D_{y,H} =$

$\|y^* - y^0\|_H$, it yields

$$\begin{aligned}
& h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\
\leq\ & \frac{1}{N}\sum_{k=0}^{N-1}\left[(x^* - x^k)^\top \delta_{k+1} + \frac{\|\delta_{k+1}\|^2}{2\eta_k}\right] \\
& + \frac{1}{2N}\left(\|y^* - y^0\|_H^2 + C\|x^* - x^0\|^2 + \frac{1}{\gamma}\sup_{\lambda\in\mathcal{B}_\rho}\|\lambda - \lambda^0\|^2\right) \\
\leq\ & \frac{1}{N}\sum_{k=0}^{N-1}\left[(x^* - x^k)^\top \delta_{k+1} + \frac{\|\delta_{k+1}\|^2}{2\eta_k}\right] + \frac{1}{2N}\left(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right).
\end{aligned} \tag{2.40}$$

Recall that $f(x)$ is **GraE**, so (2.5) and (2.6) hold. Consequently, $\mathsf{E}[\delta_{k+1}] = 0$. In addition, x_k is independent of ξ_{k+1}. Hence,

$$\mathsf{E}_{\Xi_{k+1}}[(x^* - x^k)^\top \delta_{k+1}] = 0. \tag{2.41}$$

Now, taking expectation over (2.40), and applying (2.6), we have

$$\begin{aligned}
& \mathsf{E}_{\Xi_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\
\leq\ & \mathsf{E}_{\Xi_N}\left[\frac{1}{N}\sum_{k=0}^{N-1}((x^* - x^k)^\top \delta_{k+1} + \frac{\|\delta_{k+1}\|^2}{2\eta_k})\right] \\
& + \frac{1}{2N}\left(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right) \\
\overset{(2.6)}{\leq}\ & \frac{1}{N}\mathsf{E}_{\Xi_N}\left[\sum_{k=0}^{N-1}(x^* - x^k)^\top \delta_{k+1}\right] + \frac{\sigma^2}{2N}\sum_{k=0}^{N-1}\frac{1}{\eta_k} \\
& + \frac{1}{2N}\left(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right) \\
\overset{(2.41)}{=}\ & \frac{\sigma^2}{2N}\sum_{k=0}^{N-1}\frac{1}{(k+1)\kappa} + \frac{1}{2N}\left(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right) \\
\leq\ & \frac{\sigma^2(\ln N + 1)}{2\kappa N} + \frac{1}{2N}\left(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right).
\end{aligned} \tag{2.42}$$

This completes the proof. □

2.3 The Stochastic Gradient Augmented Lagrangian Method

SGADM uses gradient projection for one block of variables and performs exact minimization for the other. However, there are cases where no exact minimization is possible at all for either of the block variables. For instance, the problem of estimating sparse additive models considered in [122] aims to solve the following stochastic minimization problem:

$$\min_{h_j, j=1,\cdots,d} \mathsf{E}[Y - \sum_{j=1}^{d} h_j(X_j)]^2 + \lambda \sum_{j=1}^{d} \sqrt{\mathsf{E}[h_j^2(X_j)]}$$

When h_js are all linear, we can introduce a linear constraint and get the following equivalent form:

$$\begin{array}{ll} \min_{z,h_j,j=1,\cdots,d} & \mathsf{E}[Y-z]^2 + \lambda \sum_{j=1}^{d} \sqrt{\mathsf{E}[h_j^2(X_j)]} \\ \text{s.t.} & \sum_{j=1}^{d} h_j(X_j) = z. \end{array}$$

Since both blocks of variables are involved in the expectation, the exact minimization for z or h_js is impossible. Therefore, it is natural to relax the exact minimization procedure of the other block variables to be replaced by gradient projection too. In this section, we assume both $f(x)$ and $g(y)$ in (2.3) are **GraE**; that is, we can only get stochastic gradients $S_f(x,\xi)$ and $S_g(y,\zeta)$ from the $\mathcal{SFO}$ for $\nabla f(x)$ and $\nabla g(y)$ respectively, where ξ and ζ are certain random variables. Recall the assumptions on **GraE**:

$$\mathsf{E}[S_f(x,\xi)] = \nabla f(x), \quad \mathsf{E}[S_g(y,\zeta)] = \nabla g(y), \tag{2.43}$$

and

$$\mathsf{E}[\|S_f(x,\xi) - \nabla f(x)\|^2] \le \sigma_1^2, \quad \mathsf{E}[\|S_g(y,\zeta) - \nabla g(y)\|^2] \le \sigma_2^2. \tag{2.44}$$

We now propose a stochastic gradient augmented Lagrangian method (SGALM). Given $\mathcal{SFO}$ for f and g, the $\mathcal{SFO}$ for $\nabla_x L_\gamma(x,y,\lambda)$ and $\nabla_y L_\gamma(x,y,\lambda)$ can be constructed as:

$$S_L^f(x,y,\lambda,\xi) := S_f(x,\xi) - A^\top \lambda + \gamma A^\top (Ax + By - b), \tag{2.45}$$

$$S_L^g(x,y,\lambda,\zeta) := S_g(y,\zeta) - B^\top \lambda + \gamma B^\top (Ax + By - b). \tag{2.46}$$

Our next algorithm, SGALM, works as follows:

The Stochastic Gradient Augmented Lagrangian Method (SGALM)

Initialize $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}$ and λ^0

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

$$y^{k+1} = [y^k - \beta_k S_L^g(x^k, y^k, \lambda^k, \zeta^{k+1})]_{\mathcal{Y}};$$
$$x^{k+1} = [x^k - \alpha_k S_L^f(x^k, y^{k+1}, \lambda^k, \xi^{k+1})]_{\mathcal{X}};$$
$$\lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b).$$

end for

Denote

$$\delta_{k+1}^f := S_f(x^k, \xi^{k+1}) - \nabla f(x^k),\ \delta_{k+1}^g := S_g(y^k, \zeta^{k+1}) - \nabla g(y^k).$$

Notice that in this section, the differentiability of function $g(y)$ is implicitly assumed. Moreover, throughout this section, we assume that the gradient ∇g is *also Lipschitz continuous*. For notational simplicity, we assume L is its Lipschitz constant too.

2.3.1 The Complexity of SGALM without Strong Convexity

Now, we are able to analyze the convergence rate of SGALM. Denote

$$\hat{Q}_k = \begin{pmatrix} H_k & 0 & 0 \\ 0 & \frac{1}{\alpha_k} I_{n_x} & 0 \\ 0 & -A & \frac{1}{\gamma} I_m \end{pmatrix}, \quad \hat{M}_k = \begin{pmatrix} H_k & 0 & 0 \\ 0 & \frac{1}{\alpha_k} I_{n_x} & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix} \tag{2.47}$$

where $H_k = \frac{1}{\beta_k} I_{n_y} - \gamma B^\top B$. The identity $\hat{Q}_k = \hat{M}_k P$ still holds where P is given according to (2.12).

Similar to Proposition 2.2.3, we have the following bounds regarding the sequence $\{\tilde{w}^k\}$ defined in (2.13), the proof of which is also delegated to the appendix.

Proposition 2.3.1 *Suppose that $\{\tilde{w}^k\}$ is given as in* (2.13)*, and the matrices $\hat{Q}_k$ and $\hat{M}_k$ are given as in* (2.47)*. For any $w \in \Omega$, we have*

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq\ & (w - \tilde{w}^k)^\top \hat{Q}_k (w^k - \tilde{w}^k) - (x - x^k)^\top \delta_{k+1}^f - (y - y^k)^\top \delta_{k+1}^g \\ & - \frac{\|\delta_{k+1}^f\|^2 + \|\delta_{k+1}^g\|^2}{2\eta_k} - \frac{\eta_k + L}{2} \left(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2 \right), \end{aligned} \tag{2.48}$$

where $\eta_k > 0$ is any prescribed sequence. Moreover, the term $(w - \tilde{w}^k)^\top \hat{Q}_k (w^k - \tilde{w}^k)$ on

the RHS can be further bounded as follows

$$\begin{aligned} & (w-\tilde{w}^k)^\top \hat{M}_k P(w^k-\tilde{w}^k) \\ \geq\ & \frac{1}{2}\left(\|w-w^{k+1}\|^2_{\hat{M}_k}-\|w-w^k\|^2_{\hat{M}_k}\right)+\frac{1}{2}(x^k-\tilde{x}^k)^\top\left(\frac{1}{\alpha_k}I_{n_x}-\gamma A^\top A\right)(x^k-\tilde{x}^k) \\ & +\frac{1}{2}(y^k-\tilde{y}^k)^\top\left(\frac{1}{\beta_k}I_{n_y}-\gamma B^\top B\right)(y^k-\tilde{y}^k), \quad \forall w\in\Omega, \end{aligned} \tag{2.49}$$

where by abusing the notation a bit we denote $\|x\|^2_A := x^\top Ax$ *with* A *being a symmetric matrix but not necessarily positive semidefinite.*

Now, we are in a position to present our main convergence rate result for the SGALM algorithm. Let us recycle the notation and denote $\Xi_k = (\xi_1, \xi_2, \ldots, \xi_k, \zeta_1, \zeta_2, \ldots, \zeta_k)$; the convergence rate will be in the expectation over Ξ_k.

Theorem 2.3.2 *Suppose both* $f(x)$ *and* $g(y)$ *in* (2.3) *are* **GraE**. *Given a fixed iteration number* N, *letting* w^k *be the sequence generated by the SGALM,* $\eta_k = \sqrt{N}$, *and* C *is a constant satisfying*

$$CI_{n_x} - \gamma A^\top A - LI_{n_x} \succeq 0 \text{ and } CI_{n_y} - \gamma B^\top B - LI_{n_y} \succeq 0,$$

and $\beta_k = \alpha_k = \frac{1}{\eta_k + C} = \frac{1}{\sqrt{N}+C}$. *For any integer* $n > 0$, *let*

$$\bar{w}_n = \frac{1}{n}\sum_{k=0}^{n-1} \tilde{w}^k, \tag{2.50}$$

where $\tilde{w}^k$ *is defined in* (2.13). *Then*

$$\begin{aligned} & \mathsf{E}_{\Xi_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\ \leq\ & \frac{\sigma_1^2+\sigma_2^2}{2\sqrt{N}} + \frac{D_x^2}{2\sqrt{N}} + \frac{D_y^2}{2\sqrt{N}} + \frac{1}{2N}\left(CD_x^2 + CD_y^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right), \end{aligned} \tag{2.51}$$

where $D_x = \mathrm{dist}(x_0, \mathcal{X}^*)$, $D_y = \mathrm{dist}(y_0, \mathcal{Y}^*)$ *and* ρ *is any fixed positive parameter.*

Proof. Similar to (2.21), by (2.48) and (2.49) we have

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq & \frac{1}{2}\left(\|w - w^{k+1}\|^2_{\hat{M}_k} - \|w - w^k\|^2_{\hat{M}_k}\right) \\ & -(x - x^k)^\top \delta^f_{k+1} - (y - y^k)^\top \delta^g_{k+1} - \frac{\|\delta^f_{k+1}\|^2 + \|\delta^g_{k+1}\|^2}{2\eta_k}. \end{aligned}$$

Following a similar line of arguments as in Theorem 2.2.2, we derive that

$$\begin{aligned} & h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\ \leq & \frac{1}{2N}\sum_{k=0}^{N-1} \frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k} + \frac{1}{2N}\sum_{k=0}^{N-1}\left(\|y^* - y^k\|^2_{H_k} - \|y^* - y^{k+1}\|^2_{H_k}\right) \\ & + \frac{1}{N}\sum_{k=0}^{N-1}\left[(x^* - x^k)^\top \delta^f_{k+1} + (y^* - y^k)^\top \delta^g_{k+1} + \frac{\|\delta^f_{k+1}\|^2 + \|\delta^g_{k+1}\|^2}{2\eta_k}\right] \\ & + \frac{1}{2N}\left(\frac{1}{\gamma}\sup_{\lambda \in \mathcal{B}_\rho} \|\lambda - \lambda^0\|^2\right). \end{aligned} \tag{2.52}$$

Compared to (2.25), the term $\sum_{k=0}^{N-1}(\|y^* - y^k\|^2_{H_k} - \|y^* - y^{k+1}\|^2_{H_k})$ is new. Since $\beta_k = \frac{1}{\eta_k + C} = \frac{1}{\sqrt{N}+C}$, we have $H_B := CI_{n_y} - \gamma B^\top B \succeq 0$. Thus,

$$\begin{aligned} & \sum_{k=0}^{N-1}\left(\|y^* - y^k\|^2_{H_k} - \|y^* - y^{k+1}\|^2_{H_k}\right) \\ = & \sum_{k=0}^{N-1}\sqrt{N}\left(\|y^* - y^k\|^2 - \|y^* - y^{k+1}\|^2\right) + \sum_{k=0}^{N-1}\left(\|y^* - y^k\|^2_{H_B} - \|y^* - y^{k+1}\|^2_{H_B}\right) \\ \leq & \sqrt{N}\|y^* - y^0\|^2 + \|y^* - y^0\|^2_{H_B} \\ \leq & (C + \sqrt{N})\|y^* - y^0\|^2. \end{aligned} \tag{2.53}$$

Moreover, according to (2.26), $\sum_{k=0}^{N-1} \frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k}$ is upper bounded by $(C + \sqrt{N})\|x^* - x^0\|^2$. Consequently, by choosing x^*, y^* such that $D_x = \|x^* - x^0\|$ and

$D_y = \|y^* - y^0\|$, we can further upper bound (2.52) as follows:

$$\begin{aligned} & h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\ \leq\ & \frac{1}{N}\sum_{k=0}^{N-1}\left[(x^* - x^k)^\top \delta_{k+1}^f + (y^* - y^k)^\top \delta_{k+1}^g + \frac{\|\delta_{k+1}^f\|^2 + \|\delta_{k+1}^g\|^2}{2\eta_k}\right] \\ & + \frac{D_x^2}{2\sqrt{N}} + \frac{D_y^2}{2\sqrt{N}} + \frac{1}{2N}\left(CD_y^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right). \end{aligned} \tag{2.54}$$

Recall that $\delta_{k+1}^f = S_f(x^k, \xi^{k+1}) - \nabla f(x^k)$, $\delta_{k+1}^g = S_g(y^k, \zeta^{k+1}) - \nabla g(y^k)$ and (2.43) holds. Since x_k is independent of ξ_{k+1} and y_k is independent of ζ_{k+1}, we have

$$\mathsf{E}_{\Xi_{k+1}}\left[(x^* - x^k)^\top \delta_{k+1}^f\right] = 0, \quad \mathsf{E}_{\Xi_{k+1}}\left[(y^* - y^k)^\top \delta_{k+1}^g\right] = 0. \tag{2.55}$$

Now, taking the expectation over (2.54), and applying (2.44), one has

$$\begin{aligned} & \mathsf{E}_{\Xi_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\ \leq\ & \mathsf{E}_{\Xi_N}\left[\frac{1}{N}\sum_{k=0}^{N-1}\left((x^* - x^k)^\top \delta_{k+1}^f + (y^* - y^k)^\top \delta_{k+1}^g + \frac{\|\delta_{k+1}^f\|^2 + \|\delta_{k+1}^g\|^2}{2\eta_k}\right)\right] \\ & + \frac{D_x^2}{2\sqrt{N}} + \frac{D_y^2}{2\sqrt{N}} + \frac{1}{2N}\left(CD_y^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right) \\ \leq\ & \frac{\sigma_1^2 + \sigma_2^2}{2\sqrt{N}} + \frac{D_x^2}{2\sqrt{N}} + \frac{D_y^2}{2\sqrt{N}} + \frac{1}{2N}\left(CD_y^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right). \end{aligned} \tag{2.56}$$

□

Therefore, the iteration complexity for the SGALM is the same as that of the SGADM: both are $O(1/\sqrt{N})$. Similar as before, in view of (2.51) it is easy to see that the complexity of SGALM for the deterministic setting would be $O(1/N)$, since in that case σ_1 and σ_2 in (2.44) are 0, and we can let $\eta_k = 1$ in Theorem 2.3.2 (thus the stepsize α_k and β_k are independent of N), leading to

$$h(\hat{u}_N) - h(u^*) + \rho\|A\hat{x}_N + B\hat{y}_N - b\| \leq \frac{1}{2N}\left((C+1)(D_x^2 + D_y^2) + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right), \tag{2.57}$$

which further leads to an $O(1/N)$ iteration complexity bound for the SGALM in the deterministic case.

2.3.2 The Complexity of SGALM under Strong Convexity

In this subsection, we show that strong convexity also leads to a lower complexity for SGALM as it does for SGADM. In fact, the iteration complexity becomes $O(\ln N/N)$ when both f and g are strongly convex. The main result is shown in the following theorem which will be in the expectation over $\Xi_k = (\xi_1, \xi_2, \ldots, \xi_k, \zeta_1, \zeta_2, \ldots, \zeta_k)$.

Theorem 2.3.3 *Suppose $f(x)$ is κ_f-strongly convex and $g(y)$ is κ_g-strongly convex, and both $f(x)$ and $g(y)$ in* (2.3) *are* **GraE**. *Let w^k be the sequence generated by the SGALM, $\eta_k = (k+1)\min(\kappa_f, \kappa_g)$, and C is a constant satisfying*

$$CI_{n_x} - \gamma A^\top A - LI_{n_x} \succeq 0 \text{ and } CI_{n_y} - \gamma B^\top B - LI_{n_y} \succeq 0,$$

and $\beta_k = \alpha_k = \frac{1}{\eta_k + C}$. For any integer $n > 0$, let

$$\bar{w}_n = \frac{1}{n}\sum_{k=0}^{n-1} \tilde{w}^k, \tag{2.58}$$

where $\tilde{w}^k$ is defined in (2.13). *Then*

$$\begin{aligned} & \mathsf{E}_{\Xi_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\ \leq\ & \frac{(\sigma_1^2 + \sigma_2^2)(\ln N + 1)}{2\min(\kappa_f, \kappa_g)N} + \frac{1}{2N}\left(CD_y^2 + CD_x^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right), \end{aligned} \tag{2.59}$$

where $D_x = \operatorname{dist}(x_0, \mathcal{X}^*)$, $D_y = \operatorname{dist}(y_0, \mathcal{Y}^*)$ and ρ is any fixed positive parameter.

Proof. Similar to the proof of Proposition 2.2.3, using the κ_f-strong convexity and κ_g-strong convexity of f and g, we conclude that

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq\ & (w - \tilde{w}^k)^\top \hat{Q}_k(w^k - \tilde{w}^k) - (x - x^k)^\top \delta_{k+1}^f - (y - y^k)^\top \delta_{k+1}^g \\ & - \frac{\|\delta_{k+1}^f\|^2 + \|\delta_{k+1}^g\|^2}{2\eta_k} - \frac{\eta_k + L}{2}\left(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2\right) \\ & + \frac{\kappa_f}{2}\|x - x^k\|^2 + \frac{\kappa_g}{2}\|y - y^k\|^2, \end{aligned} \tag{2.60}$$

where $\eta_k > 0$ is any prescribed sequence. Let $\kappa = \min(\kappa_f, \kappa_g)$, then similar to (2.21),

by (2.49) and (2.60) we have

$$\begin{aligned}
& h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\
\geq\ & \frac{1}{2}\left(\|w - w^{k+1}\|^2_{\hat{M}_k} - \|w - w^k\|^2_{\hat{M}_k}\right) - (x - x^k)^\top \delta^f_{k+1} - (y - y^k)^\top \delta^g_{k+1} \\
& - \frac{\|\delta^f_{k+1}\|^2 + \|\delta^g_{k+1}\|^2}{2\eta_k} + \frac{\kappa_f}{2}\|x - x^k\|^2 + \frac{\kappa_g}{2}\|y - y^k\|^2 \\
\geq\ & \frac{1}{2}\left(\|w - w^{k+1}\|^2_{\hat{M}_k} - \|w - w^k\|^2_{\hat{M}_k}\right) - (x - x^k)^\top \delta^f_{k+1} - (y - y^k)^\top \delta^g_{k+1} \\
& - \frac{\|\delta^f_{k+1}\|^2 + \|\delta^g_{k+1}\|^2}{2\eta_k} + \frac{\kappa}{2}(\|x - x^k\|^2 + \|y - y^k\|^2)
\end{aligned}$$

Following the same line of arguments as in Theorem 2.2.2, we derive that

$$\begin{aligned}
& h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\
\leq\ & \frac{1}{2N}\sum_{k=0}^{N-1}\left(\frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k} - \kappa\|x - x^k\|^2\right) \\
& + \frac{1}{2N}\sum_{k=0}^{N-1}\left(\|y^* - y^k\|^2_{H_k} - \|y^* - y^{k+1}\|^2_{H_k} - \kappa\|y - y^k\|^2\right) \\
& + \frac{1}{N}\sum_{k=0}^{N-1}\left[(x^* - x^k)^\top \delta^f_{k+1} + (y^* - y^k)^\top \delta^g_{k+1} + \frac{\|\delta^f_{k+1}\|^2 + \|\delta^g_{k+1}\|^2}{2\eta_k}\right] \\
& + \frac{1}{2N}\left(\frac{1}{\gamma}\sup_{\lambda \in \mathcal{B}_\rho}\|\lambda - \lambda^0\|^2\right).
\end{aligned} \tag{2.61}$$

Compared to (2.38), the term $\sum_{k=0}^{N-1}(\|y^* - y^k\|^2_{H_k} - \|y^* - y^{k+1}\|^2_{H_k})$ is new. Since

$\beta_k = \frac{1}{\eta_k + C} = \frac{1}{(k+1)\kappa + C}$ and $H_B := CI_{n_y} - \gamma B^\top B \succeq 0$, it holds that

$$\begin{aligned}
& \sum_{k=0}^{N-1} \left(\|y^* - y^k\|_{H_k}^2 - \|y^* - y^{k+1}\|_{H_k}^2 - \kappa \|y - y^k\|^2 \right) \\
= & \sum_{k=0}^{N-1} \left(k\kappa \|y^* - y^k\|^2 - (k+1)\kappa \|y^* - y^{k+1}\|^2 \right) \\
& + \sum_{k=0}^{N-1} \left(\|y^* - y^k\|_{H_B}^2 - \|y^* - y^{k+1}\|_{H_B}^2 \right) \\
\leq & \|y^* - y^0\|_{H_B}^2 \leq C\|y^* - y^0\|^2. \qquad (2.62)
\end{aligned}$$

Moreover, according to (2.39), $\sum_{k=0}^{N-1} \frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k}$ is upper bounded by $C\|x^* - x^0\|^2$. Consequently, by choosing x^*, y^* such that $D_x = \|x^* - x^0\|$ and $D_y = \|y^* - y^0\|$, we can further upper bound (2.61) as follows:

$$\begin{aligned}
& h(\bar{u}_N) - h(u^*) + \rho \|A\bar{x}_N + B\bar{y}_N - b\| \\
\leq & \frac{1}{N} \sum_{k=0}^{N-1} \left[(x^* - x^k)^\top \delta_{k+1}^f + (y^* - y^k)^\top \delta_{k+1}^g + \frac{\|\delta_{k+1}^f\|^2 + \|\delta_{k+1}^g\|^2}{2\eta_k} \right] \\
& + \frac{1}{2N} \left(CD_y^2 + CD_x^2 + \frac{1}{\gamma} \left(\rho + \|\lambda^0\| \right)^2 \right). \qquad (2.63)
\end{aligned}$$

Recall that $\delta_{k+1}^f = S_f(x^k, \xi^{k+1}) - \nabla f(x^k)$, $\delta_{k+1}^g = S_g(y^k, \zeta^{k+1}) - \nabla g(y^k)$ and (2.43) holds. Since x_k is independent of ξ_{k+1} and y_k is independent of ζ_{k+1}, we have

$$\mathsf{E}_{\Xi_{k+1}} \left[(x^* - x^k)^\top \delta_{k+1}^f \right] = 0, \quad \mathsf{E}_{\Xi_{k+1}} \left[(y^* - y^k)^\top \delta_{k+1}^g \right] = 0. \qquad (2.64)$$

Now, taking the expectation over (2.63), and applying (2.44), one has

$$\begin{aligned}
& \mathsf{E}_{\Xi_N} \left[h(\bar{u}_N) - h(u^*) + \rho \|A\bar{x}_N + B\bar{y}_N - b\| \right] \\
\leq & \mathsf{E}_{\Xi_N} \left[\frac{1}{N} \sum_{k=0}^{N-1} \left((x^* - x^k)^\top \delta_{k+1}^f + (y^* - y^k)^\top \delta_{k+1}^g + \frac{\|\delta_{k+1}^f\|^2 + \|\delta_{k+1}^g\|^2}{2\eta_k} \right) \right] \\
& + \frac{1}{2N} \left(CD_y^2 + CD_x^2 + \frac{1}{\gamma} \left(\rho + \|\lambda^0\| \right)^2 \right) \\
\leq & \frac{(\sigma_1^2 + \sigma_2^2)(\ln N + 1)}{2\kappa N} + \frac{1}{2N} \left(CD_y^2 + CD_x^2 + \frac{1}{\gamma} \left(\rho + \|\lambda^0\| \right)^2 \right). \qquad (2.65)
\end{aligned}$$

□

2.4 The Stochastic Zeroth-Order GADM

In this section, we consider another setting, where even the noisy gradient of $f(x)$ is not available. For example, the simulated-based inventory optimization problem studied in [42] definitely falls into this category. To be specific, we assume that $\mathcal{L}_\gamma(x, y, \lambda)$ is **MinE** with respect to y, and $f(x)$ is **ValE**. In other words, for any given x we can get a noisy approximation of the true function value $f(x)$ by calling an $\mathcal{SZO}$, which returns a quantity $\mathcal{F}(x, \xi)$ with ξ being a certain random variable. The $\mathcal{SZO}$ becomes relevant when a part of the objective contains the expectation of an unknown function, assuming only some sample realizations of the expectation are observable. This is the case, for instance, when we know nothing about the true nature of the randomness and how they are related to the objective in an explicit fashion; however, we can learn the objective by observation. Such problems are frequently encountered in management science; e.g., a demand function is often not explicitly accessible while its realizations are observable.

Now that we can access the $\mathcal{SZO}$, we shall use the smoothing scheme proposed in [89] to approximate the first order information of a given function f. The smoothing technique is to utilize the integration operator to promote the differentiability. More specifically, suppose that v is a random vector in $\mathbb{R}^n$ with density function ρ. A smooth approximation of f with the smoothing parameter μ is defined as:

$$f_\mu(x) = \int f(x + \mu v)\rho(v)dv. \tag{2.66}$$

Theoretically, one can choose to use any pre-specified smoothing distribution $\rho(v)$. For instance, in [89] Nesterov adopted the Gaussian distribution to simplify the computation. However, the Gaussian distribution has a support set of the whole space $\mathbb{R}^n$, which cannot be implemented for problems with constraints. To avoid using the entire space as the sample space, we shall use the smoothing scheme based on the uniform distribution over a (scalable) ball in $\mathbb{R}^n$ as introduced in [109].

Definition 6 *Let U_b be the uniform distribution over the unit Euclidean ball and B be the unit ball. Given $\mu > 0$, the smoothing function f_μ is defined as*

$$f_\mu(x) = \mathsf{E}_{\{v\sim U_b\}}[f(x + \mu v)] = \frac{1}{\alpha(n)} \int_B f(x + \mu v)dv \tag{2.67}$$

where $\alpha(n)$ is the volume of the unit ball in $\mathbb{R}^n$.

Some properties of the smoothing function are shown in the lemma below, which will be used in our forthcoming discussion; the proof of the lemma can be found in the appendix. In the following discussion, $C_L^1(\mathbb{R}^n)$ denotes the function class with Lipschitz continuous first-order derivative.

Lemma 2.4.1 *Suppose that $f \in C_L^1(\mathbb{R}^n)$. Let U_{S_p} be the uniform distribution over the unit Euclidean sphere, and S_p be the unit sphere in $\mathbb{R}^n$. Then we have:*

(a) The smoothing function f_μ is continuously differentiable, and its gradient is Lipschitz continuous with constant $L_\mu \leq L$ and

$$\nabla f_\mu(x) = \mathsf{E}_{\{v \sim U_{S_p}\}}\left[\frac{n}{\mu} f(x+\mu v) v\right] = \frac{1}{\beta(n)} \int_{v \in S_p} \frac{n}{\mu} \left[f(x+\mu v) - f(x)\right] v dv \tag{2.68}$$

where $\beta(n)$ is the surface area of the unit sphere in $\mathbb{R}^n$.

(b) For any $x \in \mathbb{R}^n$, we have

$$|f_\mu(x) - f(x)| \leq \frac{L\mu^2}{2}, \tag{2.69}$$

$$\|\nabla f_\mu(x) - \nabla f(x)\| \leq \frac{\mu n L}{2}, \tag{2.70}$$

$$\mathsf{E}_v\left[\left\|\frac{n}{\mu}[f(x+\mu v) - f(x)]v\right\|^2\right] \leq 2n\|\nabla f(x)\|^2 + \frac{\mu^2}{2} L^2 n^2. \tag{2.71}$$

(c) If f is convex, then f_μ is also convex.

We remark that the bounds in Part (b) are slightly sharper (up to some constant factor) than that of Gaussian smoothing scheme in [89]. Moreover, the new smoothing scheme will involve the sampling points in the μ-ball of x. This feature is important for the problems where the domain of f may only be slightly larger than $\mathcal{X}$, as we shall see from the oracle to be introduced next. Based on (2.68) we define the zeroth-order stochastic gradient of f at point x^k:

$$G_\mu(x^k, \xi_{k+1}, v) = \frac{n_x}{\mu}\left[\mathcal{F}(x^k + \mu v, \xi_{k+1}) - \mathcal{F}(x^k, \xi_{k+1})\right] v, \tag{2.72}$$

where v is the random vector uniformly distributed over the unit sphere in $\mathbb{R}^{n_x}$. The zeroth-order GADM algorithm is described as follows:

The Zeroth-Order GADM

Initialize $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}$ and λ^0

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

$y^{k+1} = \arg\min_{y\in\mathcal{Y}} \mathcal{L}_\gamma(x^k, y, \lambda^k) + \frac{1}{2}\|y - y^k\|_H^2$.

Call the $\mathcal{SZO}$ m_k times to obtain $G_\mu(x_k, \xi_{k+1,i}, v_{k+1,i}), i = 1, \cdots, m_k$.

Then set $G_{\mu,k} = \frac{1}{m_k}\sum_{i=1}^{m_k} G_\mu(x^k, \xi_{k+1,i}, v_{k+1,i})$, and compute

$x^{k+1} = [x^k - \alpha_k(G_{\mu,k} - A^\top\lambda^k + \gamma A^\top(Ax^k + By^{k+1} - b))]_{\mathcal{X}}$;

$\lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)$.

end for

Before conducting the complexity analysis for the algorithm above, we present some properties of the function $G(x^k, \xi_{k+1}) := \nabla_x \mathcal{F}(x^k, \xi_{k+1})$. Note that function f is **ValE**, i.e. (2.7) and (2.8) hold. This fact together with Lemma 2.4.1(a) leads to:

Lemma 2.4.2 *Suppose that $G_\mu(x^k, \xi_{k+1}, v)$ is defined as in* (2.72)*, and f is* **ValE***, i.e.* (2.7), (2.8) *and* (2.9) *hold. Then*

$$\mathsf{E}_{v,\xi_{k+1}}[G_\mu(x^k, \xi_{k+1}, v)] = \nabla f_\mu(x^k). \tag{2.73}$$

If we further assume $\|\nabla f(x)\| \le M$, $\forall x \in \mathcal{X}$, then the following holds

$$\mathsf{E}_{v,\xi_{k+1}}[\|G_\mu(x^k, \xi_{k+1}, v) - \nabla f_\mu(x^k)\|^2] \le \tilde{\sigma}^2, \tag{2.74}$$

where $\tilde{\sigma}^2 = 2n_x[M^2 + \sigma^2 + \mu^2 L^2 n_x]$.

Proof. The first statement is easy to verify. We shall focus on the second statement. Applying (2.71) and (2.9) to $\mathcal{F}(x^k, \xi_{k+1})$ and $G(x^k, \xi_{k+1})$, we have

$$\begin{aligned}
& \mathsf{E}_{v,\xi_{k+1}}\left[\|G_\mu(x^k, \xi_{k+1}, v)\|^2\right] \\
= \;& \mathsf{E}_{\xi_{k+1}}\left[\mathsf{E}_v\left[\|G_\mu(x^k, \xi_{k+1}, v)\|^2\right]\right] \\
\le \;& 2n_x\left[\mathsf{E}_{\xi_{k+1}}\left[\|G(x^k, \xi_{k+1})\|^2\right]\right] + \frac{\mu^2}{2}L^2 n_x^2 \\
\le \;& 2n_x\left\{\mathsf{E}_{\xi_{k+1}}[\|\nabla f(x^k)\|^2] + \mathsf{E}_{\xi_{k+1}}\left[\|G(x^k, \xi_{k+1}) - \nabla f(x^k)\|^2\right]\right\} + \mu^2 L^2 n_x^2 \\
\le \;& 2n_x\left\{\|\nabla f(x^k)\|^2 + \sigma^2\right\} + \mu^2 L^2 n_x^2.
\end{aligned} \tag{2.75}$$

Then from (2.75), (2.73), and $\|\nabla f(x^k)\| \leq M$, we have

$$\begin{aligned} & \mathsf{E}_{v,\xi_{k+1}}\left[\|G_\mu(x^k,\xi_{k+1},v)-\nabla f_\mu(x^k)\|^2\right] \\ = \;& \mathsf{E}_{v,\xi_{k+1}}\left[\|G_\mu(x^k,\xi_{k+1},v)\|^2\right]-\|\nabla f_\mu(x)\|^2 \\ \leq \;& 2n_x\left[M^2+\sigma^2+\mu^2L^2n_x\right]=\tilde{\sigma}^2. \end{aligned} \tag{2.76}$$

□

2.4.1 Convergence Rate of Zeroth-Order GADM

To establish the convergence rate, we refer the sequence $\tilde{w}^k$ to be the sequence defined in (2.13) with the corresponding iterates x^k, y^k, λ^k obtained from the zeroth-order GADM. We let $\delta_{\mu,k} = G_{\mu,k} - \nabla f_\mu(x_k)$, which plays a similar role as δ_k in SGADM. We have the following proposition, whose proof is almost identical to that of (2.48) in Proposition 2.3.1 except that δ_{k+1} is now replaced by $\delta_{\mu,k}$.

Proposition 2.4.3 *Suppose that $\mathcal{L}_\gamma(x,y,\lambda)$ is* **MinE** *with respect to y, and $f(x)$ is* **ValE**. *Let x^k, y^k, λ^k be obtained in the zeroth-order GADM, $\tilde{w}^k$ be specified as in* (2.13), *and $h_\mu(u) = f_\mu(x) + g(y)$. Then for any $w \in \Omega$, we have*

$$\begin{aligned} & h_\mu(u) - h_\mu(\tilde{u}^k) + (w-\tilde{w}^k)F(\tilde{w}^k) \\ \geq \;& (w-\tilde{w}^k)^\top Q_k(w^k-\tilde{w}^k) - (x-x^k)^\top\delta_{\mu,k} - \frac{\|\delta_{\mu,k}\|^2}{2\eta_k} - \frac{\eta_k+L}{2}\|x^k-\tilde{x}^k\|^2, \end{aligned} \tag{2.77}$$

where $\eta_k > 0$ can be any positive constant to be specified in the analysis later.

Now, we are ready to present the following theorem which leads to the convergence rate of the zeroth-order GADM. In the rest of this section, we denote $\Omega_n = (\xi_{k,i}, v_{k,i})$ for $k = 1,2,\ldots,n$ and $i = 1,2,\ldots,m_k$, the convergence rate will be considered in the expectation taken on Ω_N.

Theorem 2.4.4 *Let w^k be the sequence generated by the zeroth-order GADM, and C be a constant such that $CI_{n_x} - \gamma A^\top A - LI_{n_x} \succeq 0$, and $\alpha_k = \frac{1}{\eta_k+C}$. For any integer $n > 0$, let*

$$\bar{w}_n = \frac{1}{n}\sum_{k=0}^{n-1}\tilde{w}^k, \tag{2.78}$$

where $\tilde{w}^k$ is defined in (2.13)*. Then the following holds*

$$\begin{aligned}
& \mathsf{E}_{\Omega_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\
\leq\ & \frac{1}{2N}\sum_{k=1}^{N}\eta_k\left(\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2\right) \\
& + \frac{\tilde{\sigma}^2}{2N}\sum_{k=1}^{N}\frac{1}{m_k\eta_k} + \frac{1}{2N}\left(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}(\rho + \|\lambda_0\|)^2\right) + L\mu^2, \qquad (2.79)
\end{aligned}$$

where $D_x = \mathrm{dist}(x_0, \mathcal{X}^*)$ and $D_{y,H} = \mathrm{dist}(y_0, \mathcal{Y}^*)_H$, $\{\eta_k > 0\}$ and $\rho > 0$ are any given constants.

Proof. By (2.77) and (2.20), it follows that

$$\begin{aligned}
& h_\mu(u) - h_\mu(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\
\geq\ & \frac{1}{2}\left(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2\right) + \frac{1}{2}(x^k - \tilde{x}^k)^\top\left(\frac{1}{\alpha_k}I_{n_x} - \gamma A^\top A\right)(x^k - \tilde{x}^k) \\
& -(x - x^k)^\top\delta_{\mu,k} - \frac{\|\delta_{\mu,k}\|^2}{2\eta_k} - \frac{\eta_k + L}{2}\|x^k - \tilde{x}^k\|^2 \\
=\ & \frac{1}{2}(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2) \\
& + \frac{1}{2}(x^k - \tilde{x}^k)^\top\left(\frac{1}{\alpha_k}I_{n_x} - \gamma A^\top A - (\eta_k + L)I_{n_x}\right)(x^k - \tilde{x}^k) \\
& -(x - x^k)^\top\delta_{\mu,k} - \frac{\|\delta_{\mu,k}\|^2}{2\eta_k} \\
\geq\ & \frac{1}{2}\left(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2\right) - (x - x^k)^\top\delta_{\mu,k} - \frac{\|\delta_{\mu,k}\|^2}{2\eta_k}.
\end{aligned}$$

In similar vein as the proof of (2.25) in Theorem 2.2.2 (except that δ_{k+1} is replaced by $\delta_{\mu,k}$), we obtain:

$$\begin{aligned}
& h_\mu(\bar{u}_N) - h_\mu(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\
\leq\ & \frac{1}{2N}\sum_{k=0}^{N-1}\frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k} + \frac{1}{N}\sum_{k=0}^{N-1}\left[(x^* - x^k)^\top\delta_{\mu,k} + \frac{\|\delta_{\mu,k}\|^2}{2\eta_k}\right] \\
& + \frac{1}{2N}\left(\|y^* - y^0\|_H^2 + \frac{1}{\gamma}\sup_{\lambda\in\mathcal{B}_\rho}\|\lambda - \lambda^0\|^2\right). \qquad (2.80)
\end{aligned}$$

Recall that $\delta_{\mu,k} = G_{\mu,k} - \nabla f_\mu(x_k)$, which combined with (2.73) implies

$$\mathsf{E}_{\xi_{k+1},v_{k+1}}[\delta_{\mu,k}] = \mathsf{E}_{\xi_{k+1},v_{k+1}}[G_{\mu,k} - \nabla f_\mu(x_k)] = 0.$$

In addition, since ξ_{k+1} and v_{k+1} are independent to x_k, we have the following identity

$$\mathsf{E}_{\Omega_{k+1}}[(x^* - x^k)^\top \delta_{\mu,k}] = 0. \tag{2.81}$$

Now, taking expectation over (2.80), choosing x^*, y^* such that $D_x = \|x^* - x^0\|$ and $D_{y,H} = \|y^* - y^0\|_H$ and applying (2.74), we have

$$\begin{aligned}
& \mathsf{E}_{\Omega_N}\left[h_\mu(\bar{u}_N) - h_\mu(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\
\leq\ & \mathsf{E}_{\Omega_N}\left[\frac{1}{N}\sum_{k=0}^{N-1}\left((x^* - x^k)^\top \delta_{\mu,k} + \frac{\|\delta_{\mu,k}\|^2}{2\eta_k}\right)\right] \\
& + \frac{1}{2N}\sum_{k=0}^{N-1}\eta_k\left(\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2\right) \\
& + \frac{1}{2N}(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}(\rho + \|\lambda_0\|)^2) \\
\overset{(2.74)}{\leq}\ & \frac{\tilde{\sigma}^2}{N}\sum_{k=0}^{N-1}\frac{1}{m_k\eta_k} + \frac{1}{2N}\sum_{k=0}^{N-1}\eta_k\left(\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2\right) \\
& + \frac{1}{2N}(D_{y,H}^2 + CD_x^2 + \frac{1}{\gamma}(\rho + \|\lambda_0\|)^2).
\end{aligned} \tag{2.82}$$

By (2.69), we have $|(h_\mu(\bar{u}_N) - h_\mu(u^*)) - (h(\bar{u}_N) - h(u^*))| \leq L\mu^2$, and so

$$\mathsf{E}\left[h(\bar{u}_N) - h(u^*)\right] \leq \mathsf{E}\left[h_\mu(\bar{u}_N) - h_\mu(u^*)\right] + L\mu^2. \tag{2.83}$$

Finally, combining (2.82) and (2.83) yields the desired result. □

In Theorem 2.4.4, η_k and the batch sizes m_k are generic. It is possible to provide one choice of the parameters so as to yield an overall simpler iteration complexity bound.

Corollary 2.4.5 *Under the same assumptions as in Theorem 2.4.4, we let* $\eta_k = 1$ *for*

all $k = 1, 2, \ldots, N$, *and the batch sizes* $m_k = m$ *for all* $k = 1, 2, \ldots, N$. *Then*

$$\mathsf{E}_{\Omega_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \leq \frac{D_w^2}{2N} + \frac{n_x(M^2+\sigma^2)}{m} + \frac{\mu^2 L^2 n_x^2}{m} + L\mu^2.$$

Proof. It follows from (2.79), with the specified parameters, that

$$\begin{aligned}
& \mathsf{E}_{\Omega_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\
\leq\ & \frac{1}{2N}(D_{y,H}^2 + (C+1)D_x^2 + \frac{1}{\gamma}(\rho + \|\lambda_0\|)^2) + \frac{\tilde{\sigma}^2}{2m} + L\mu^2 \\
=\ & \frac{D_w^2}{2N} + \frac{\tilde{\sigma}^2}{2m} + L\mu^2 \\
=\ & \frac{D_w^2}{2N} + \frac{2n_x\left(M^2+\sigma^2+\mu^2L^2n_x\right)}{2m} + L\mu^2 \\
=\ & \frac{D_w^2}{2N} + \frac{n_x\left(M^2+\sigma^2+\mu^2L^2n_x\right)}{m} + L\mu^2 \\
=\ & \frac{D_w^2}{2N} + \frac{n_x\left(M^2+\sigma^2\right)}{m} + \frac{\mu^2L^2n_x^2}{m} + L\mu^2
\end{aligned}$$

where we denote $D_w^2 = D_{y,H}^2 + (C+1)D_x^2 + \frac{1}{\gamma}(\rho + \|\lambda_0\|)^2$. □

In the corollary above, the complexity bound is dependent on the sample size m, and the smoothing parameter μ. We shall further choose m and μ to obtain an explicit iteration bound.

Corollary 2.4.6 *Under the same assumptions as in Theorem 2.4.4 and Corollary 2.4.5, we have:*

(a) Given a fixed iteration number N, *if the smoothing parameter is chosen to be* $\mu \leq \sqrt{\frac{1}{N}}$, *and the number of calls to* $\mathcal{SZO}$ *at each iteration is* $m = N$, *then we have*

$$\mathsf{E}_{\Omega_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \leq \frac{1}{N}\left(\frac{D_w^2}{2} + n_x(M^2+\sigma^2) + L\right) + \frac{L^2n_x^2}{N^2}.$$

(b) Given a fixed number of calls to $\mathcal{SZO}$ *to be* $\bar{N}$, *choose the smoothing parameter* $\mu \leq \sqrt{\frac{1}{\bar{N}}}$ *and the number of calls to the* $\mathcal{SZO}$ *at each iteration to be*

$$m = \left\lfloor \min\left\{ \max\left\{ \frac{\sqrt{n_x(M^2+\delta^2)\bar{N}}}{\tilde{D}}, \frac{n_xL}{\tilde{D}} \right\}, \bar{N} \right\} \right\rfloor,$$

for some $\tilde{D} > 0$. *Then,* $N = \lfloor \frac{\bar{N}}{m} \rfloor$ *and*

$$\begin{aligned} & \mathsf{E}_{\Omega_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\ \leq \quad & \frac{L}{\bar{N}} + \frac{n_x L}{\bar{N}}\left(\tilde{D}\theta_2 + \frac{D_w^2}{\tilde{D}}\right) + \frac{\sqrt{n_x(M^2+\delta^2)}}{\sqrt{\bar{N}}}\left(\tilde{D}\theta_1 + \frac{D_w^2}{\tilde{D}}\right) \end{aligned}$$

where

$$\theta_1 = \max\left\{1, \frac{\sqrt{n_x(M^2+\delta^2)}}{\tilde{D}\sqrt{\bar{N}}}\right\} \text{ and } \theta_2 = \max\left\{1, \frac{n_x L}{\tilde{D}\bar{N}}\right\}. \tag{2.84}$$

Proof. Part (a). Since we have $m = N$, $\mu \leq \sqrt{\frac{1}{N}}$

$$\begin{aligned} & \mathsf{E}_{\Omega_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\ \leq \quad & \frac{D_w^2}{2N} + \frac{n_x(M^2+\sigma^2)}{m} + \frac{\mu^2 L^2 n_x^2}{m} + L\mu^2 \\ \leq \quad & \frac{D_w^2}{2N} + \frac{n_x(M^2+\sigma^2)}{N} + \frac{L^2 n_x^2}{N^2} + \frac{L}{N} \\ = \quad & \frac{1}{N}\left(\frac{D_w^2}{2} + n_x(M^2+\sigma^2) + L\right) + \frac{L^2 n_x^2}{N^2}. \end{aligned}$$

Part (b). The total number of $\mathcal{SZO}$ calls is now fixed to be $\bar{N}$. Under the assumption that at each iteration m times of $\mathcal{SZO}$ are called, we have $\bar{N}/2m \leq N \leq \bar{N}/m$, and so

$$\begin{aligned} & \mathsf{E}_{\Omega_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\ \leq \quad & \frac{D_w^2}{2N} + \frac{n_x(M^2+\sigma^2)}{m} + \frac{\mu^2 L^2 n_x^2}{m} + L\mu^2 \\ \leq \quad & \frac{D_w^2 m}{\bar{N}} + \frac{n_x(M^2+\sigma^2)}{m} + \frac{L^2 n_x^2}{m\bar{N}} + \frac{L}{\bar{N}} \\ \leq \quad & \frac{D_w^2 m}{\bar{N}} + \frac{n_x(M^2+\sigma^2)}{m} + \frac{L^2 n_x^2}{m\bar{N}} + \frac{L}{\bar{N}}. \end{aligned} \tag{2.85}$$

Now noting the definitions of θ_1, θ_2 in (2.84), we equivalently write m as

$$m = \left\lfloor \max\left\{\frac{\sqrt{n_x(M^2+\delta^2)\bar{N}}}{\tilde{D}\theta_1}, \frac{n_x L}{\tilde{D}\theta_2}\right\}\right\rfloor.$$

Finally,

$$\begin{aligned}
& \text{RHS of (2.85)} \\
\leq \quad & \frac{D_w^2\left(\frac{\sqrt{n_x(M^2+\delta^2)\bar{N}}}{\tilde{D}\theta_1}+\frac{n_xL}{\tilde{D}\theta_2}\right)}{\bar{N}}+\frac{\sqrt{n_x(M^2+\sigma^2)}\tilde{D}\theta_1}{\sqrt{\bar{N}}}+\frac{n_xL\tilde{D}\theta_2}{\bar{N}}+\frac{L}{\bar{N}} \\
\leq \quad & \frac{D_w^2}{\tilde{D}}\frac{\sqrt{n_x(M^2+\delta^2)}}{\sqrt{\bar{N}}}+\frac{D_w^2}{\tilde{D}}\frac{n_xL}{\bar{N}}+\frac{\sqrt{n_x(M^2+\sigma^2)}\tilde{D}\theta_1}{\sqrt{\bar{N}}}+\frac{n_xL\tilde{D}\theta_2}{\bar{N}}+\frac{L}{\bar{N}} \\
= \quad & \frac{L}{\bar{N}}+\frac{n_xL}{\bar{N}}\left(\tilde{D}\theta_2+\frac{D_w^2}{\tilde{D}}\right)+\frac{\sqrt{n_x\left(M^2+\delta^2\right)}}{\sqrt{\bar{N}}}\left(\tilde{D}\theta_1+\frac{D_w^2}{\tilde{D}}\right). \qquad (2.86)
\end{aligned}$$

□

Remark that the complexity bound of $O(1/N)$ in Part (a) of Corollary 2.4.6 is in terms of the iteration N. However, in the zeroth-order GADM algorithm we need to call $\mathcal{SZO}$ multiple times at each iteration. The complexity in terms of the total number of calls to $\mathcal{SZO}$ in Part (b) of Corollary 2.4.6 is denoted as $\bar{N}$, and this gives us a bound on the accuracy of $O(1/\sqrt{\bar{N}})$.

2.5 Numerical Experiments

In this section, we test the performance of the new SGADM algorithm on two problem instances: the fused logistic regression and the graph-guided regularized logistic regression, on which we compare the performance of SGADM with three existing stochastic ADMM-type algorithms: STOC-ADMM, OPG-ADMM, and RDA-ADMM. Specifically, STOC-ADMM proposed in [93] is the first stochastic ADMM-type algorithms; OPG-ADMM proposed in [114] is designed as online ADMM-type algorithms, but is also applicable for solving stochastic optimization. As shown in [114], the RDA-ADMM improves the performance of the online ADMM [121]. Hence, we do not include online ADMM in our test.

For both models, the tests are conducted on four binary classification datasets: *a9a, mushrooms, splice, w8a* [1], and the summary of those datasets are shown in Table 2.2. More details of those two experiments will be presented in the following subsections separately.

[1] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/binary.html

dataset	number of samples	dimensionality
a9a	32561	123
mushrooms	8124	112
splice	1000	60
w8a	64700	300

Table 2.2: Summary of datasets

2.5.1 Fused Logistic Regression

As suggested in [72], fused logistic regression, which incorporates a certain ordering information, is derived from the fused lasso problem and sparse logistic regression. Specifically, the sparse logistic regression problem (see [76]) is given by:

$$\min_{x\in\mathbb{R}^n} \quad l(x) + \beta\|x\|_1, \tag{2.87}$$

where $l(x) = \frac{1}{m}\sum_{i=1}^{m} \log(1+\exp(-b_i(a_i^\top x)))$, and $\{(a_i, b_i),\, i = 1,\ldots,m\}$ is a given training set with m samples $a_1, a_2, \ldots, a_m$ and $b_i \in \{\pm 1\}, i = 1,\ldots,m$ as the binary class labels. Combining requirements from the fused lasso [118] and the sparse logistic regression (2.87), the fused logistic regression that incorporates certain existed natural ordering features can be formulated as:

$$\min_{x\in\mathbb{R}^n} \quad l(x) + \beta\|x\|_1 + \rho\sum_{j=2}^{n} |x_j - x_{j-1}|. \tag{2.88}$$

If we further introduce a matrix $F \in \mathbb{R}^{(n-1)\times n}$, with all ones on the diagonal and negative ones on the super-diagonal and zeros elsewhere, then the problem boils down to

$$\min_{x\in\mathbb{R}^n} \quad l(x) + \beta\|x\|_1 + \rho\|Fx\|_1. \tag{2.89}$$

By introducing another variable y, and imposing the constraint $Fx = y$, this problem can be solved by stochastic ADMM. In fact, the total loss is in the form of expectation taken over the dataset under a uniform distribution.

In our experiments, the regularization parameters are set to be $\beta = 5\times 10^{-4}$ and $\rho = 5\times 10^{-3}$. For each dataset, we use 10-fold cross-validation for training and testing. Three different measures of the performance are shown in the comparison of those tested

algorithms, including *objective value*, *test loss*, *time cost* and *prediction error*. *Objective value* measures the function value of the optimization problem (2.89) evaluated on the training data samples, whereas *test loss* is the value of the logistic loss function evaluated on the test data sample. Besides, *prediction error* is the classification error rate evaluated on the test dataset. The number of epochs represents the number of passes that have been run by the algorithms on the whole training data samples. Figure 2.1 shows the results of those algorithms for solving the fused logistic regression problem, where the results are averaged over ten runs on ten folds. We observe that the new SGADM algorithm is competitive to other stochastic ADMM-type algorithms. In fact, SGADM consumes less computational time, and often achieves the best performance in terms of objective value and test loss on various datasets, although the performances of these methods are not drastically different.

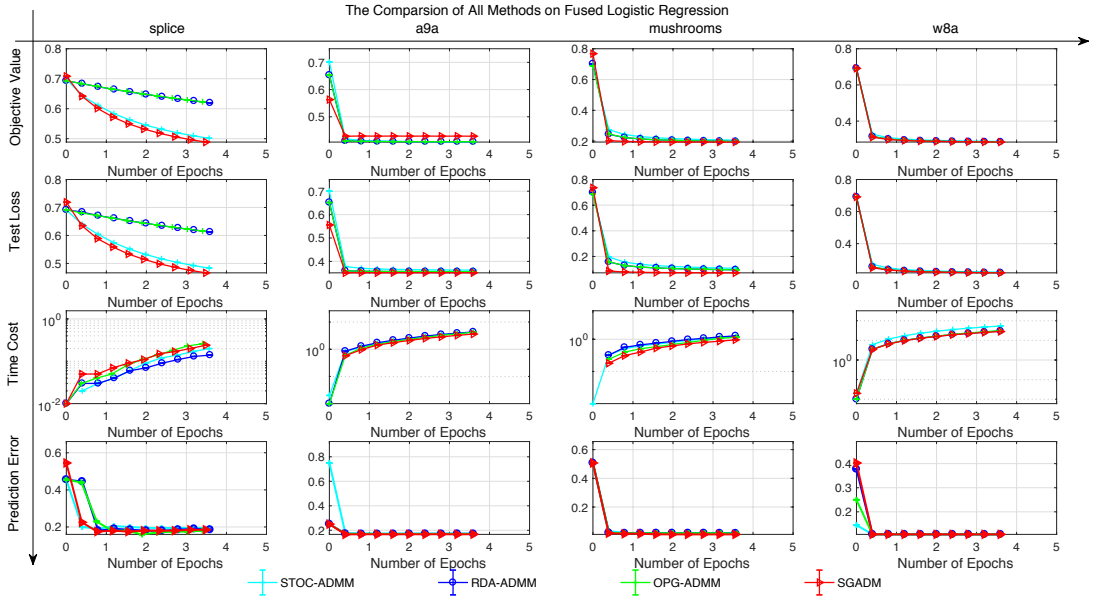

Figure 2.1: Comparison of SGADM, STOC-ADMM, RDA-ADMM, OPG-ADMM on Fused Logistic Regression. First Rows: objective values. Second Rows: test losses. Third Rows: time cost. Fourth Rows: Prediction Error.

2.5.2 Graph-guided Regularized Logistic Regression

In this subsection, we test those algorithms on the following graph-guided regularized logistic regression problem

$$\min_{x\in\mathbb{R}^n} \quad l(x) + \tfrac{\beta}{2}\|x\|_2^2 + \rho \sum_{(i,j)\in E} w_{ij}|x_i - x_j|, \tag{2.90}$$

where $l(x)$ is sum of logistic loss that is similarly defined as in (2.88), and E is the edge set in a certain graph. This model penalizes the difference between variables connected in the graph with different weight. By introducing a matrix $G \in \mathbb{R}^{n\times n}$ that captures the structure of the graph, the problem can be written as

$$\min_{x\in\mathbb{R}^n} \quad l(x) + \tfrac{\beta}{2}\|x\|_2^2 + \rho\|Gx\|_1. \tag{2.91}$$

With different loss functions, similar problems are considered in the graph-guided SVM [93] and generalized lasso [119]. Similar to fused logistic regression, by introducing another variable y and imposing the constraint $Gx = y$, this problem can also be solved by stochastic ADMM.

In our experiments, the regularization parameters are set as $\beta = 10^{-2}$ and $\rho = 10^{-5}$. For each dataset, we use 10-fold cross-validation for training and testing. Moreover, the matrix G in (2.91) is generated by sparse inverse covariance selection [107]. Figure 2.2 shows the results of those algorithms for solving the graph-guided regularized logistic regression problem. Similar to the fused logistic regression, the new SGADM algorithm is comparable and competitive to other stochastic type ADMM algorithms.

2.6 Conclusions

In this chapter, we considered the problem of minimizing the sum of two convex functions, subject to linear coupled constraints. In contrast to the original setting of the ADMM, we assume that only some noisy estimation of the objective function is possible. Therefore, the classical ADMM cannot be applied in this context. To account for the available (informational) structure, in this chapter we proposed a suite of adapted variants of the ADMM, and establish their iteration complexity bounds accordingly, under very mild conditions. For instance, we do not assume the boundness of the optimal set, nor the coecivity of the objective function. Therefore, the analysis in this chapter ac-

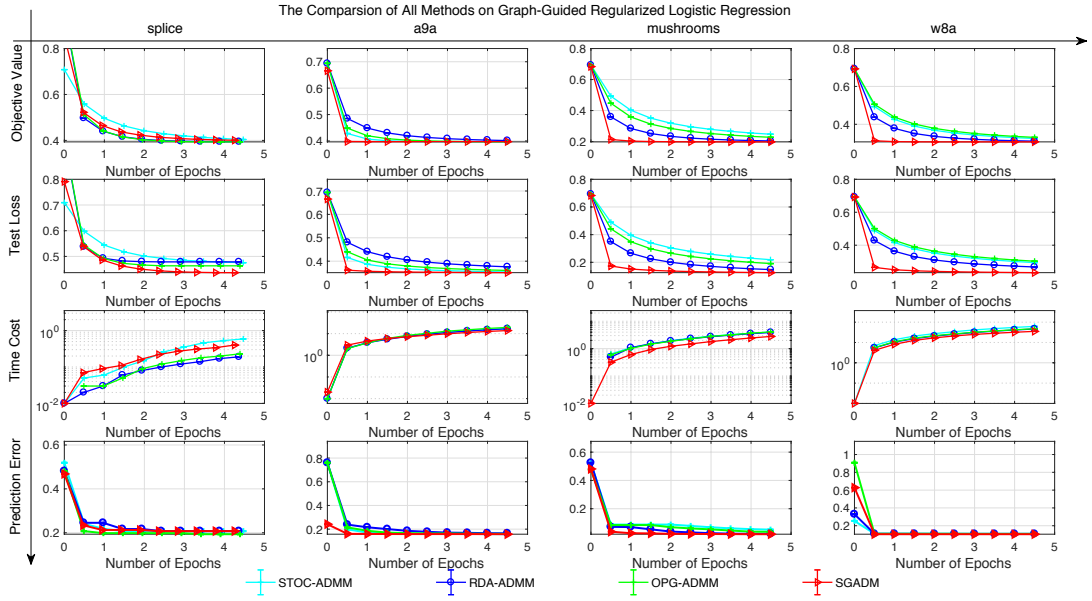

Figure 2.2: Comparison of SGADM, STOC-ADMM, RDA-ADMM, OPG-ADMM on Graph-guided Regularized Logistic Regression. First Rows: objective values. Second Rows: test losses. Third Rows: time cost. Fourth Rows: Prediction Error.

tually generalizes (and simplifies at the same time) the existing results on the iteration complexity bounds for the ADMM type algorithms in terms of the feasibility/objective measurements. Finally, we remark that the new zeroth-order smoothing oracle uses a bounded support set (specifically, a sphere with scalable radius), which is different from a more conventional normal distribution-based smoothing (e.g. Nesterov [89]) for which the support set would be the whole space. Obviously, the bounded-set smoothing is important for the constrained problems, as the points to be sampled must be in the domain of the objective function. Interestingly, we managed to prove that not only the bounded-set smoothing is feasible, but also the approximation bounds can be improved, up to some constant factors.

2.7 Technical Proofs

2.7.1 Proof of Proposition 2.2.3

Here we will prove Proposition 2.2.3. Before proceeding, let us present some technical lemmas without proof.

Lemma 2.7.1 *Suppose function f is smooth and its gradient is Lipschitz continuous, i.e.* (2.2) *holds, then we have*

$$f(x) \leq f(y) + \nabla f(y)^\top (x - y) + \frac{L}{2}\|x - y\|^2. \tag{2.92}$$

Lemma 2.7.2 *Suppose function f is smooth and convex, and its gradient is Lipschitz continuous with the constant L, i.e.* (2.2) *holds. Then we have*

$$(x - y)^\top \nabla f(z) \leq f(x) - f(y) + \frac{L}{2}\|z - y\|^2. \tag{2.93}$$

Furthermore, if f is κ-strongly convex, we have

$$(x - y)^\top \nabla f(z) \leq f(x) - f(y) + \frac{L}{2}\|z - y\|^2 - \frac{\kappa}{2}\|x - z\|^2. \tag{2.94}$$

Lemma 2.7.1 is also known as the descent lemma which is well known; one can find its proof in e.g. [6]. Lemma 2.7.2 is similar to Fact 1 in [31] which follows from the (strong) convexity of f and Lemma 2.7.1.

Proof of (2.19) **in Proposition 2.2.3**

Proof. First, by the optimality condition of the two subproblems in SGADM, we have $\forall y \in \mathcal{Y}$

$$(y - y^{k+1})^\top \left(\partial g(y^{k+1}) - B^\top \left(\lambda^k - \gamma(Ax^k + By^{k+1} - b) \right) - H(y^k - y^{k+1}) \right) \geq 0,$$

and $\forall x \in \mathcal{X}$

$$(x - x^{k+1})^\top \left(x^{k+1} - \left(x^k - \alpha_k \left(G(x^k, \xi^{k+1}) - A^\top(\lambda^k - \gamma(Ax^k + By^{k+1} - b)) \right) \right) \right) \geq 0,$$

where $\partial g(y)$ is a subgradient of g at y. Using $\tilde{\lambda}^k = \lambda^k - \gamma(Ax^k + By^{k+1} - b)$ and the definition of $\tilde{w}^k$ in (2.13), the above two inequalities are equivalent to

$$(y - \tilde{y}^k)^\top \left(\partial g(\tilde{y}^k) - B^\top \tilde{\lambda}^k - H(y^k - y^{k+1}) \right) \geq 0, \quad \forall y \in \mathcal{Y}, \tag{2.95}$$

and

$$(x - \tilde{x}^k)^\top \left(\alpha_k \left(G(x^k, \xi^{k+1}) - A^\top \tilde{\lambda}^k \right) - (x^k - \tilde{x}^k) \right) \geq 0, \quad \forall x \in \mathcal{X}. \tag{2.96}$$

Moreover,

$$(A\tilde{x}^k + B\tilde{y}^k - b) - \left(-A(x^k - \tilde{x}^k) + \frac{1}{\gamma} \left(\lambda^k - \tilde{\lambda}^k \right) \right) = 0.$$

Thus

$$(\lambda - \tilde{\lambda}^k)^\top (A\tilde{x}^k + B\tilde{y}^k - b) = \left(\lambda - \tilde{\lambda}^k \right)^\top \left(-A \left(x^k - \tilde{x}^k \right) + \frac{1}{\gamma} \left(\lambda^k - \tilde{\lambda}^k \right) \right). \tag{2.97}$$

By the convexity of $g(y)$ and (2.95),

$$g(y) - g(\tilde{y}^k) + (y - \tilde{y}^k)^\top \left(-B^\top \tilde{\lambda}^k \right) \geq (y - \tilde{y}^k)^\top H(y^k - \tilde{y}^k), \quad \forall y \in \mathcal{Y}. \tag{2.98}$$

Since $\delta_{k+1} = G(x^k, \xi^{k+1}) - \nabla f(x^k)$, and by (2.96) we have $\forall x \in \mathcal{X}$.

$$(x - \tilde{x}^k)^\top \left(\alpha_k (\nabla f(x^k) - A^\top \tilde{\lambda}^k) + \alpha_k \delta_{k+1} - (x^k - \tilde{x}^k) \right) \geq 0, \quad \forall x \in \mathcal{X}$$

which leads to

$$(x - \tilde{x}^k)^\top \left(\alpha_k (\nabla f(x^k) - A^\top \tilde{\lambda}^k) \right) \geq (x - \tilde{x}^k)^\top \left(x^k - \tilde{x}^k \right) - \alpha_k \left(x - \tilde{x}^k \right)^\top \delta_{k+1},$$

Using (2.93), the above further leads to

$$\begin{aligned} & \alpha_k(f(x) - f(\tilde{x}^k)) + (x - \tilde{x}^k)^\top(-\alpha_k A^\top \tilde{\lambda}^k) \\ \geq\ & (x - \tilde{x}^k)^\top \left(x^k - \tilde{x}^k\right) - \alpha_k \left(x - \tilde{x}^k\right)^\top \delta_{k+1} - \frac{\alpha_k L}{2}\|x^k - \tilde{x}^k\|^2, \quad \forall x \in \mathcal{X}. \end{aligned} \tag{2.99}$$

Furthermore,

$$\begin{aligned} (x - \tilde{x}^k)^\top \delta_{k+1} &= (x - x^k)^\top \delta_{k+1} + (x^k - \tilde{x}^k)^\top \delta_{k+1} \\ &\leq (x - x^k)^\top \delta_{k+1} + \frac{\eta_k}{2}\|x^k - \tilde{x}^k\|^2 + \frac{\|\delta_{k+1}\|^2}{2\eta_k}. \end{aligned} \tag{2.100}$$

Substituting (2.100) in (2.99), and dividing both sides by α_k, we get

$$\begin{aligned} & f(x) - f(\tilde{x}^k) + (x - \tilde{x}^k)^\top(-A^\top \tilde{\lambda}^k) \\ \geq\ & \frac{(x - \tilde{x}^k)^\top(x^k - \tilde{x}^k)}{\alpha_k} - (x - x^k)^\top \delta_{k+1} - \frac{\|\delta_{k+1}\|^2}{2\eta_k} - \frac{\eta_k + L}{2}\|x^k - \tilde{x}^k\|^2. \end{aligned} \tag{2.101}$$

Finally, (2.19) follows by summing (2.101), (2.98), and (2.97). □

Now we show the second statement in Proposition 2.2.3.

Proof of (2.20) **in Proposition 2.2.3**

Proof. First, by (2.14), we have $P(w^k - \tilde{w}^k) = (w^k - w^{k+1})$, and so

$$(w - \tilde{w}^k)^\top Q_k(w^k - \tilde{w}^k) = (w - \tilde{w}^k)^\top M_k P(w^k - \tilde{w}^k) = (w - \tilde{w}^k)^\top M_k(w^k - w^{k+1}).$$

Applying the identity

$$(a - b)^\top M_k(c - d) = \frac{1}{2}\left(\|a - d\|_{M_k}^2 - \|a - c\|_{M_k}^2\right) + \frac{1}{2}\left(\|c - b\|_{M_k}^2 - \|d - b\|_{M_k}^2\right)$$

to the term $(w - \tilde{w}^k)^\top M_k(w^k - w^{k+1})$, we obtain

$$\begin{aligned} & (w - \tilde{w}^k)^\top M_k(w^k - w^{k+1}) \\ =\ & \frac{1}{2}\left(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2\right) + \frac{1}{2}\left(\|w^k - \tilde{w}^k\|_{M_k}^2 - \|w^{k+1} - \tilde{w}^k\|_{M_k}^2\right). \end{aligned} \tag{2.102}$$

Using (2.14) again, we have

$$\begin{aligned} & \|w^k - \tilde{w}^k\|_{M_k}^2 - \|w^{k+1} - \tilde{w}^k\|_{M_k}^2 \\ = \; & \|w^k - \tilde{w}^k\|_{M_k}^2 - \|(w^k - \tilde{w}^k) - (w^k - w^{k+1})\|_{M_k}^2 \\ = \; & \|w^k - \tilde{w}^k\|_{M_k}^2 - \|(w^k - \tilde{w}^k) - P(w^k - \tilde{w}^k)\|_{M_k}^2 \\ = \; & (w^k - \tilde{w}^k)^\top (2M_k P - P^\top M_k P)(w^k - \tilde{w}^k). \end{aligned} \tag{2.103}$$

Note that $Q_k = M_k P$ and the definition of those matrices (see (2.12)), we have

$$2M_k P - P^\top M_k P = 2Q_k - P^\top Q_k = \begin{pmatrix} H & 0 & 0 \\ 0 & \frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A & A^\top \\ 0 & -A & \frac{1}{\gamma} I_m \end{pmatrix}.$$

As a result,

$$\begin{aligned} & (w^k - \tilde{w}^k)^\top (2M_k P - P^\top M_k P)(w^k - \tilde{w}^k) \\ = \; & \|y^k - \tilde{y}^k\|_H^2 + \frac{1}{\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2 + (x^k - \tilde{x}^k)^\top \left(\frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A\right)(x^k - \tilde{x}^k) \\ \geq \; & (x^k - \tilde{x}^k)^\top \left(\frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A\right)(x^k - \tilde{x}^k). \end{aligned} \tag{2.104}$$

Combining (2.104), (2.103), and (2.102), the desired inequality (2.20) follows. □

2.7.2 Proof of Proposition 2.3.1

We first show the first part of Proposition 2.3.1.

Proof of (2.48) **in Proposition 2.3.1**

Proof. by the optimality condition of the two subproblems in SGALM, we have $\forall y \in \mathcal{Y}$

$$(y - y^{k+1})^\top \left(y^{k+1} - y^k + \beta_k \left(S_g(y^k, \zeta^{k+1}) - B^\top(\lambda^k - \gamma(Ax^k + By^k - b))\right)\right) \geq 0,$$

and also $\forall x \in \mathcal{X}$

$$(x - x^{k+1})^\top \left(x^{k+1} - x^k + \alpha_k \left(S_f(x^k, \xi^{k+1}) - A^\top(\lambda^k - \gamma(Ax^k + By^{k+1} - b))\right)\right) \geq 0.$$

Using $\tilde{\lambda}^k = \lambda^k - \gamma(Ax^k + By^{k+1} - b)$ and the definition of $\tilde{w}^k$, the above two inequalities

are equivalent to

$$(y-\tilde{y}^k)^\top\left(\beta_k\left(S_g(y^k,\zeta^{k+1})-B^\top\tilde{\lambda}^k\right)-(I_{n_y}-\beta_k\gamma B^\top B)(y^k-\tilde{y}^k)\right)\geq 0,\quad \forall y\in\mathcal{Y}, \tag{2.105}$$

and

$$(x-\tilde{x}^k)^\top\left(\alpha_k(S_f(x^k,\xi^{k+1})-A^\top\tilde{\lambda}^k)-(x^k-\tilde{x}^k)\right)\geq 0,\quad \forall x\in\mathcal{X}. \tag{2.106}$$

Also,

$$(\lambda-\tilde{\lambda}^k)^\top(A\tilde{x}^k+B\tilde{y}^k-b)=(\lambda-\tilde{\lambda}^k)^\top\left(-A(x^k-\tilde{x}^k)+\frac{1}{\gamma}(\lambda^k-\tilde{\lambda}^k)\right). \tag{2.107}$$

Since $\delta_{k+1}^f=S_f(x^k,\xi^{k+1})-\nabla f(x^k)$ and using (2.106), similar to (2.99) and (2.100) we have

$$\begin{aligned} & f(x)-f(\tilde{x}^k)+(x-\tilde{x}^k)^\top(-A^\top\tilde{\lambda}^k)\\ \geq\ & \frac{(x-\tilde{x}^k)^\top(x^k-\tilde{x}^k)}{\alpha_k}-(x-x^k)^\top\delta_{k+1}^f-\frac{\|\delta_{k+1}^f\|^2}{2\eta_k}-\frac{\eta_k+L}{2}\|x^k-\tilde{x}^k\|^2. \end{aligned} \tag{2.108}$$

Similarly, since $\delta_{k+1}^g=S_g(y^k,\zeta^{k+1})-\nabla g(y^k)$ and using (2.105), we also have

$$\begin{aligned} & g(y)-g(\tilde{y}^k)+(y-\tilde{y}^k)^\top(-B^\top\tilde{\lambda}^k)\\ \geq\ & (y-\tilde{y}^k)^\top\left(\frac{1}{\beta_k}I_{n_y}-\gamma B^\top B\right)(y^k-\tilde{y}^k)\\ & -(y-y^k)^\top\delta_{k+1}^g-\frac{\|\delta_{k+1}^g\|^2}{2\eta_k}-\frac{\eta_k+L}{2}\|y^k-\tilde{y}^k\|^2. \end{aligned} \tag{2.109}$$

Finally, (2.48) follows by summing (2.109), (2.108), and (2.107). □

Notice that $\hat{Q}_k = \hat{M}_k P$ and

$$2M_kP - P^\top M_k P = \begin{pmatrix} H_k & 0 & 0 \\ 0 & \frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A & A^\top \\ 0 & -A & \frac{1}{\gamma} I_m \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{\alpha_k} I_{n_x} - \gamma B^\top B & 0 & 0 \\ 0 & \frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A & A^\top \\ 0 & -A & \frac{1}{\gamma} I_m \end{pmatrix}.$$

Inequality (2.49) in Proposition 2.3.1 follows similarly as the derivation of (2.20) in Proposition 2.2.3.

2.7.3 Properties of the Smoothing Function

In this subsection, we will prove Lemma 2.4.1. Before that, we need some technical preparations which are summarized in the following lemma.

Lemma 2.7.3 *Let $\alpha(n)$ be the volume of the unit ball in $\mathbb{R}^n$, and $\beta(n)$ be the surface area of the unit sphere in $\mathbb{R}^n$. We also denote B, and S_p, to be the unit ball and unit sphere respectively.*

(a) If M_p is defined as $M_p = \frac{1}{\alpha(n)} \int_{v \in B} \|v\|^p dv$, we have

$$M_p = \frac{n}{n+p}. \tag{2.110}$$

(b) Let I be the identity matrix in $\mathbb{R}^{n \times n}$, then

$$\int_{S_p} vv^\top dv = \frac{\beta(n)}{n} I. \tag{2.111}$$

Proof. For (a), we can directly compute M_p by using the polar coordinates,

$$M_p = \frac{1}{\alpha(n)} \int_B \|v\|^p dv = \frac{1}{\alpha(n)} \int_0^1 \int_{S_p} r^p r^{n-1} dr d\theta = \frac{1}{n+p} \frac{\beta(n)}{\alpha(n)} = \frac{n}{n+p}.$$

For (b), Let $V = vv^\top$, then we know that $V_{ij} = v_i v_j$. Therefore, if $i \neq j$, by the symmetry of the unit sphere S_p (i.e. if $v \in S_p, v = (v_1, v_2, \ldots, v_n)$, then $w \in S_p$ for all

$w = (\pm v_1, \pm v_2, \ldots, \pm v_n))$, we have

$$\int_{S_p} V_{ij} dv = \int_{S_p} v_i v_j dv = \int_{S_p} -v_i v_j dv = \int_{S_p} -V_{ij} dv.$$

Thus, we obtain $\int_{S_p} V_{ij} dv = 0$.

If $i = j$, we know that $V_{ii} = v_i^2$. Since we already know that

$$\int_{S_p} (v_1^2 + v_2^2 + \cdots + v_n^2) dv = \int_{S_p} \|v\|^2 dv = \beta(n).$$

Then, by symmetry, we have

$$\int_{S_p} v_1^2 dv = \int_{S_p} v_2^2 dv = \cdots = \int_{S_p} v_n^2 dv = \frac{\beta(n)}{n}.$$

Thus we also have $\int_{S_p} V_{ii}^2 dv = \frac{\beta(n)}{n}$, for $i = 1, 2, \ldots, n$. Therefore, $\int_{S_p} vv^\top dv = \frac{\beta(n)}{n} I$. □

By the next three propositions, the part (b) of Lemma 2.4.1 is shown; for part (a) and (c) of Lemma 2.4.1, the proof can be found in [109].

Proposition 2.7.4 *If $f \in C_L^1(\mathbb{R}^n)$, then*

$$|f_\mu(x) - f(x)| \leq \frac{L\mu^2}{2}. \tag{2.112}$$

Proof. Since $f \in C_L^1(\mathbb{R}^n)$, we have

$$\begin{aligned}
|f_\mu(x) - f(x)| &= \left| \frac{1}{\alpha(n)} \int_B f(x + \mu v) dv - f(x) \right| \\
&= \left| \frac{1}{\alpha(n)} \int_B (f(x + \mu v) - f(x) - \nabla f(x)^\top \mu v) dv \right| \\
&\leq \int_B \left| (f(x + \mu v) - f(x) - \nabla f(x)^\top \mu v) \right| dv \\
&\leq \int_B \frac{L\mu^2}{2} \|v\|^2 dv \\
&\overset{(2.110)}{=} \frac{L\mu^2}{2} \frac{n}{n+2} \leq \frac{L\mu^2}{2}.
\end{aligned}$$

□

Proposition 2.7.5 *If $f \in C_L^1(\mathbb{R}^n)$, then*

$$\|\nabla f_\mu(x) - \nabla f(x)\| \leq \frac{\mu n L}{2}. \tag{2.113}$$

Proof.

$$\begin{aligned}
& \|\nabla f_\mu(x) - \nabla f(x)\| \\
= \quad & \left\| \frac{1}{\beta(n)} \left[\frac{n}{\mu} \int_{S_p} f(x+\mu v) v dv \right] - \nabla f(x) \right\| \\
\overset{(2.111)}{=} \quad & \left\| \frac{1}{\beta(n)} \left[\frac{n}{\mu} \int_{S_p} f(x+\mu v) v dv - \int_{S_p} \frac{n}{\mu} f(x) v dv - \int_{S_p} \frac{n}{\mu} \langle \nabla f(x), \mu v \rangle v dv \right] \right\| \\
\leq \quad & \frac{n}{\beta(n)\mu} \int_{S_p} |f(x+\mu v) - f(x) - \langle \nabla f(x), \mu v \rangle| \|v\| dv \\
\leq \quad & \frac{n}{\beta(n)\mu} \frac{L\mu^2}{2} \int_{S_p} \|v\|^3 dv = \frac{\mu n L}{2}.
\end{aligned}$$

□

Proposition 2.7.6 *If $f \in C_L^1(\mathbb{R}^n)$, and the $\mathcal{SZO}$ defined as $g_\mu(x) = \frac{n}{\mu}[f(x+\mu v) - f(x)]v$, then we have*

$$\mathsf{E}_v \left[\|g_\mu(x)\|^2 \right] \leq 2n \|\nabla f(x)\|^2 + \frac{\mu^2}{2} L^2 n^2. \tag{2.114}$$

Proof.

$$
\begin{aligned}
& \mathsf{E}_v[\|g_\mu(x)\|^2] \\
= \quad & \frac{1}{\beta(n)} \int_{S_p} \frac{n^2}{\mu^2} |f(x+\mu v) - f(x)|^2 \|v\|^2 dv \\
= \quad & \frac{n^2}{\beta(n)\mu^2} \int_{S_p} \left[f(x+\mu v) - f(x) - \langle \nabla f(x), \mu v \rangle + \langle \nabla f(x), \mu v \rangle \right]^2 dv \\
\leq \quad & \frac{n^2}{\beta(n)\mu^2} \int_{S_p} \left[2\left(f(x+\mu v) - f(x) - \langle \nabla f(x), \mu v \rangle \right)^2 + 2\left(\langle \nabla f(x), \mu v \rangle \right)^2 \right] dv \\
\leq \quad & \frac{n^2}{\beta(n)\mu^2} \left[\int_{S_p} 2 \left(\frac{L\mu^2}{2} \|v\|^2 \right)^2 dv + 2\mu^2 \int_{S_p} \nabla f(x)^\top v v^\top \nabla f(x) dv \right] \\
\overset{(2.111)}{=} \quad & \frac{n^2}{\beta(n)\mu^2} \left[\frac{L^2\mu^4}{2} \beta(n) + 2\mu^2 \frac{\beta(n)}{n} \|\nabla f(x)\|^2 \right] \\
= \quad & 2n\|\nabla f(x)\|^2 + \frac{\mu^2}{2} L^2 n^2.
\end{aligned}
$$

□

Chapter 3

First-Order Algorithms for Convex Optimization with Nonseparable Objective and Coupled Constraints

3.1 Preliminaries

Before we discuss the general multi-block optimization model (1.7), for the simplicity of presentation we first consider the following model:

$$\begin{array}{ll} \min & f(x,y)+h_1(x)+h_2(y) \\ \text{s.t.} & Ax+By=b, \\ & x\in\mathcal{X}, y\in\mathcal{Y} \end{array} \tag{3.1}$$

where $x\in\mathbb{R}^p$, $y\in\mathbb{R}^q$, $A\in\mathbb{R}^{m\times p}$, $B\in\mathbb{R}^{m\times q}$, $b\in\mathbb{R}^m$, $\mathcal{X},\mathcal{Y}$ are closed convex sets, f is a smooth jointly convex function, and h_1,h_2 are (possibly nonsmooth) convex functions. In this chapter, the augmented Lagrangian function for problem (3.1) is defined as

$$\mathcal{L}_\gamma(x,y,\lambda)=f(x,y)+h_1(x)+h_2(y)-\lambda^\top(Ax+By-b)+\frac{\gamma}{2}\|Ax+By-b\|^2,$$

where λ is the multiplier.

As we will show that under the assumptions that the gradient of the coupling func-

tion ∇f is Lipschitz continuous and one of h_1 and h_2 is strongly convex, then an $O(1/N)$ convergence rate of ADMM can still be assured. We also show that APGMM, AGPMM, and their hybrid version have a convergence rate of $O(1/N)$ if ∇f is Lipschitz continuous. Moreover, we show that ADMM can be extended to the multi-block model (1.7). Similarly, under the Lipschitz continuity of ∇f and the assumptions in [74], an $O(1/N)$ iteration bound still holds for the multi-block model.

The rest of this chapter is organized as follows. In Section 3.2, we introduce ADMM, APGMM, AGPMM and their hybrids. The results on the rate of convergence of these algorithms are presented in the subsections of the same section, while the detailed proofs of the convergence results are presented in Appendix 3.5. In Section 3.3, we extend our analysis of the ADMM to a general setting with multiple (more than 2) blocks of variables. Finally, we conclude the chapter in Section 3.4.

3.2 New Algorithms

Let us first introduce some notations that will be frequently used in the analysis later. The aggregated primal variables x, y and the primal-dual variables x, y, λ are respectively denoted by u and w, and the primal-dual mapping F; namely

$$u := \begin{pmatrix} x \\ y \end{pmatrix}, \quad w := \begin{pmatrix} x \\ y \\ \lambda \end{pmatrix}, \quad F(w) := \begin{pmatrix} -A^\top \lambda \\ -B^\top \lambda \\ Ax + By - b \end{pmatrix}, \tag{3.2}$$

and $h(u) := f(x, y) + h_1(x) + h_2(y)$.

Throughout this chapter, we assume f to be smooth and has a Lipschitz continuous gradient; i.e.

Assumption 3.2.1 *The coupling function f satisfies*

$$\|\nabla f(u_2) - \nabla f(u_1)\| \leq L\|u_2 - u_1\|, \ \forall\, u_1, u_2 \in \mathcal{X} \times \mathcal{Y}, \tag{3.3}$$

where L is a Lipschitz constant for ∇f.

For a function f satisfying Assumption 3.2.1, it is useful to note the following inequalities.

Lemma 3.2.1 *Suppose that function f satisfies* (3.3)*, then we have*

$$f(u_2) \leq f(u_1) + \nabla f(u_1)^\top (u_2 - u_1) + \frac{L}{2}\|u_2 - u_1\|^2, \tag{3.4}$$

for any u_1, u_2. In general, if f is also convex then

$$f(u_2) \leq f(u_1) + \nabla f(u_3)^\top (u_2 - u_1) + \frac{L}{2}\|u_2 - u_3\|^2, \tag{3.5}$$

for any u_1, u_2, u_3.

The proof of this lemma is similar to (2.7.2), we omit it here.

For convenience of analysis, we introduce some matrix notations. Let

$$Q := \begin{pmatrix} G & 0 & 0 \\ 0 & \gamma B^\top B & 0 \\ 0 & -B & \frac{1}{\gamma} I_m \end{pmatrix}, \quad P := \begin{pmatrix} I_p & 0 & 0 \\ 0 & I_q & 0 \\ 0 & -\gamma B & I_m \end{pmatrix}, \quad M := \begin{pmatrix} G & 0 & 0 \\ 0 & \gamma B^\top B & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix} \tag{3.6}$$

hence, $Q = MP$. Suppose the sequence $\{w^k\}$ is generated by an algorithm, we introduce an auxiliary sequence:

$$\tilde{w}^k := \begin{pmatrix} \tilde{x}^k \\ \tilde{y}^k \\ \tilde{\lambda}^k \end{pmatrix} = \begin{pmatrix} x^{k+1} \\ y^{k+1} \\ \lambda^k - \gamma(Ax^{k+1} + By^k - b) \end{pmatrix}. \tag{3.7}$$

Based on (3.7) and (3.6), the relationship between the new sequence $\{\tilde{w}^k\}$ and the original $\{w^k\}$ is

$$w^{k+1} = w^k - P(w^k - \tilde{w}^k). \tag{3.8}$$

3.2.1 The Alternating Direction Method of Multipliers

As we discussed earlier, the ADMM can be applied straightforwardly to solve (3.1), assuming that the augmented Lagrangian (with a proximal term) can be optimized for each block of variables, while other variables are fixed. This gives rise to the following scheme:

ADMM

Initialize $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}$ and λ^0

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

$x^{k+1} = \arg\min_{x\in\mathcal{X}} \mathcal{L}_\gamma(x, y^k, \lambda^k) + \frac{1}{2}\|x - x^k\|_G^2$;

$y^{k+1} = \arg\min_{y\in\mathcal{Y}} \mathcal{L}_\gamma(x^{k+1}, y, \lambda^k) + \frac{1}{2}\|y - y^k\|_H^2$;

$\lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)$.

end for

In the above algorithm, G and H are two pre-specified positive semidefinite matrices. The main result concerning its convergence and iteration complexity are summarized in the following theorem, whose proof can be found in Appendix 3.5.1.

Theorem 3.2.2 *Suppose that ∇f satisfies Lipschitz condition* (3.3)*, and $h_2(y)$ is strongly convex with parameter $\sigma > 0$, i.e.*

$$h_2(y) \geq h_2(z) + h_2'(z)^\top (y - z) + \frac{\sigma}{2}\|y - z\|^2 \tag{3.9}$$

where $h_2'(z) \in \partial h_2(z)$ is a subgradient of $h_2(z)$. Let $\{w^k\}$ be the sequence generated by the ADMM, and $G \succ 0, H \succ \left(L + \frac{L^2}{\sigma}\right) I_q$. Then the sequence $\{w^k\}$ generated by the ADMM converges to an optimal solution. Moreover, for any integer $n > 0$ letting

$$\bar{u}_n := \frac{1}{n}\sum_{k=1}^{n} u^k, \tag{3.10}$$

we have

$$\begin{aligned} & h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\ \leq \quad & \frac{1}{2N}\left(\mathrm{dist}(x^0, \mathcal{X}^*)_G^2 + \mathrm{dist}(y^0, \mathcal{Y}^*)_{\hat{H}}^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right), \end{aligned} \tag{3.11}$$

where $\mathcal{X}^* \times \mathcal{Y}^*$ is the optimal solution set, $\mathrm{dist}(x, S)_M := \inf_{y\in S}\|x - y\|_M$, and $\hat{H} := \gamma B^\top B + H$.

We quote Lemma 2.4 in [38] as follows:

> Assume that $\rho > 0$, and $\tilde{x} \in X$ is an approximate solution of the problem $f^* := \inf\{f(x) : Ax - b = 0, x \in X\}$ where f is convex, satisfying
>
> $$f(\bar{x}) - f^* + \rho\|A\bar{x} - b\| \leq \epsilon. \tag{3.12}$$

Then, we have

$$\|A\bar{x} - b\| \leq \frac{\epsilon}{\rho - \|\lambda^*\|} \text{ and } f(\bar{x}) - f^* \leq \epsilon$$

where λ^* is an optimal Lagrange multiplier associated with the constraint $Ax - b = 0$ in the problem $\inf\{f(x) : Ax - b = 0, x \in X\}$, assuming $\|\lambda^*\| < \rho$.

In other words, estimation (3.11) in Theorem 3.2.2 automatically establishes that

$$h(\bar{u}_N) - h(u^*) \leq O(1/N) \text{ and } \|A\bar{x}_N + B\bar{y}_N - b\| \leq O(1/N).$$

The same applies to all subsequent iteration complexity results presented in this section.

3.2.2 The Alternating Proximal Gradient Method of Multipliers

In some applications, the augmented Lagrangian function may be difficult to minimize for some block of variables, while fixing all others. In this subsection we consider an approach where we apply *proximal gradient* for each block of variables. The method bears some similarity to the Iterative Shrinkage-Thresholding (ISTA) Algorithm (cf. [3]), although we are dealing with multiple blocks of variables here. We shall call the new method *Alternating Proximal Gradient Method of Multipliers* (APGMM), presented as follows:

APGMM

Initialize $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}$ and λ^0

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

$$x^{k+1} = \arg\min_{x \in \mathcal{X}} \nabla_x f(x^k, y^k)^\top (x - x^k) + h_1(x) + \tfrac{\gamma}{2}\|Ax + By^k - b - \tfrac{1}{\gamma}\lambda^k\|^2; \\ + \tfrac{1}{2}\|x - x^k\|_G^2$$

$$y^{k+1} = \arg\min_{y \in \mathcal{Y}} \nabla_y f(x^k, y^k)^\top (y - y^k) + h_2(y) + \tfrac{\gamma}{2}\|Ax^{k+1} + By - b - \tfrac{1}{\gamma}\lambda^k\|^2; \\ + \tfrac{1}{2}\|y - y^k\|_H^2$$

$$\lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b).$$

end for

The convergence property and iteration complexity are summarized in the following theorem, whose proof is in Appendix 3.5.2.

Theorem 3.2.3 *Suppose that ∇f satisfies Lipschitz condition* (3.3)*. Let $\{w^k\}$ be the sequence generated by the APGMM, and $G \succ LI_p$ and $H \succ LI_q$. Then, the sequence*

$\{w^k\}$ generated by the APGMM converges to an optimal solution. Moreover, for any integer $n > 0$, letting

$$\bar{u}_n := \frac{1}{n}\sum_{k=1}^{n} u^k,$$

it holds that

$$\begin{aligned} & h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\ \leq \quad & \frac{1}{2N}\left(\mathrm{dist}(x^0, \mathcal{X}^*)_G^2 + \mathrm{dist}(y^0, \mathcal{Y}^*)_{\hat{H}}^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right), \end{aligned} \tag{3.13}$$

where $\mathcal{X}^* \times \mathcal{Y}^*$ is the optimal solution set, $\mathrm{dist}(x, S)_M := \inf_{y\in S}\|x - y\|_M$*, and* $\hat{H} := \gamma B^\top B + H$*.*

3.2.3 The Alternating Gradient Projection Method of Multipliers

Implementing proximal gradient step may still be difficult for some instances of applications. It is therefore natural to further simplify the step to *Gradient Projection*. Namely, for each block of variables we simply sequentially compute the projection of the gradient of the augmented Lagrangian function before updating the multipliers. The method is depicted as follows:

AGPMM

Initialize $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}$ and λ^0

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

$x^{k+1} = [x^k - \alpha(\nabla_x f(x^k, y^k) + \nabla_x h_1(x^k) - A^\top\lambda^k + A^\top(Ax^k + By^k - b))]_{\mathcal{X}}$;

$y^{k+1} = [y^k - \alpha(\nabla_y f(x^k, y^k) + \nabla_y h_2(y^k) - B^\top\lambda^k + B^\top(Ax^{k+1} + By^k - b))]_{\mathcal{Y}}$;

$\lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)$.

end for

---

where $[x]_{\mathcal{X}}$ denotes the projection of x onto $\mathcal{X}$, and $[y]_{\mathcal{Y}}$ denotes the projection of y onto $\mathcal{Y}$.

Note here that we used 'PG' as acronym for *Proximal Gradient*, and 'GP' as acronym for *Gradient Projection*. The acronyms are quite similar, and so some attention is needed not to confuse the two! Below we shall present the main convergence and the iteration complexity results for the above method; the proof of the theorem can be found in Appendix 3.5.3.

Theorem 3.2.4 *Suppose that ∇f satisfies Lipschitz condition (3.3). Let w^k be the sequence generated by the AGPMM, and $G := \gamma A^\top A + \frac{1}{\alpha} I_p$, $H := \frac{1}{\alpha} I_q - \gamma B^\top B$. Moreover, suppose that α is chosen to satisfy $H - 2LI_q \succ 0$, and $G - 2LI_p \succ 0$. Then, the sequence $\{w^k\}$ generated by the AGPMM converges to an optimal solution. For any integer $n > 0$, letting*

$$\bar{u}_n := \frac{1}{n} \sum_{k=1}^{n} u^k,$$

it holds that

$$\begin{aligned} & h(\bar{u}_N) - h(u^*) + \rho \|A\bar{x}_N + B\bar{y}_N - b\| \\ \leq \quad & \frac{1}{2N} \left(\mathrm{dist}(x^0, \mathcal{X}^*)_G^2 + \mathrm{dist}(y^0, \mathcal{Y}^*)_{\hat{H}}^2 + \frac{1}{\gamma} \left(\rho + \|\lambda^0\| \right)^2 \right), \end{aligned}$$

where $\mathcal{X}^* \times \mathcal{Y}^*$ is the optimal solution set, $\mathrm{dist}(x, S)_M := \inf_{y \in S} \|x - y\|_M$, and $\hat{H} = \gamma B^\top B + H$.

3.2.4 The Hybrids

There are instances where one part of the block variables is easy to deal with, while the other part is difficult, e.g. [72]. To take advantage of that situation, we propose the following two types of hybrid methods. The first one is to combine ADMM with Proximal Gradient in two blocks of variables:

ADM-PG

Initialize $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}$ and λ^0

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

$x^{k+1} = \arg\min_{x \in \mathcal{X}} \mathcal{L}_\gamma(x, y^k, \lambda^k) + \frac{1}{2}\|x - x^k\|_G^2$;

$y^{k+1} = \arg\min_{y \in \mathcal{Y}} \nabla_y f(x^{k+1}, y^k)^\top (y - y^k) + h_2(y) + \frac{\gamma}{2}\|Ax^{k+1} + By - b - \frac{1}{\gamma}\lambda^k\|^2$;
$+\frac{1}{2}\|y - y^k\|_H^2$

$\lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)$.

end for

The iteration complexity of the above method is as follows. The proof of the theorem can be found in Appendix 3.5.4.

Theorem 3.2.5 *Suppose that ∇f satisfies Lipschitz condition (3.3). Let w^k be the sequence generated by the ADM-PG, and $G \succ 0, H \succ LI_q$. Then, the sequence $\{w^k\}$ generated by the APGMM converges to an optimal solution. For any integer $n > 0$, letting*

$$\bar{u}_n := \frac{1}{n}\sum_{k=1}^{n} u^k, \tag{3.14}$$

it holds that

$$\begin{aligned} & h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\ \leq \quad & \frac{1}{2N}\left(\mathrm{dist}(x^0, \mathcal{X}^*)_G^2 + \mathrm{dist}(y^0, \mathcal{Y}^*)_{\hat{H}}^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right), \end{aligned}$$

where $\mathcal{X}^* \times \mathcal{Y}^*$ is the optimal solution set, $\mathrm{dist}(x, S)_M := \inf_{y\in S}\|x - y\|_M$, and $\hat{H} := \gamma B^\top B + H$.

Another possible approach is to combine ADMM with Gradient Projection, which works as follows:

ADM-GP

Initialize $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}$ and λ^0

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

$x^{k+1} = \arg\min_{x\in\mathcal{X}} \mathcal{L}_\gamma(x, y^k, \lambda^k) + \frac{1}{2}\|x - x^k\|_G^2$;

$y^{k+1} = [y^k - \alpha(\nabla_y f(x^{k+1}, y^k) + \nabla_y h_2(y^k) - B^\top\lambda^k + B^\top(Ax^{k+1} + By^k - b))]_{\mathcal{Y}}$;

$\lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)$.

end for

The main convergence result is as follows, and the proof of the theorem can be found in Appendix 3.5.4.

Theorem 3.2.6 *Let w^k be the sequence generated by the ADM-GP, $G \succ 0$ and $H := \frac{1}{\alpha}I_q - \gamma B^\top B$. Moreover, suppose that α is chosen to satisfy $H - LI_q \succ 0$. Then, the sequence $\{w^k\}$ generated by the ADM-GP converges to an optimal solution. For any integer $n > 0$, letting*

$$\bar{u}_n := \frac{1}{n}\sum_{k=1}^{n} u^k,$$

it holds that

$$\begin{aligned} & h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\ \leq \quad & \frac{1}{2N}\left(\mathrm{dist}(x^0, \mathcal{X}^*)_G^2 + \mathrm{dist}(y^0, \mathcal{Y}^*)_{\hat{H}}^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right), \end{aligned}$$

where $\mathcal{X}^* \times \mathcal{Y}^*$ *is the optimal solution set,* $\mathrm{dist}(x, S)_M := \inf_{y\in S}\|x - y\|_M$*, and* $\hat{H} := \gamma B^\top B + H$*.*

3.3 The General Multi-Block Model

Different variations of the ADMM have been a popular subject of study in the recent years, and the ADMM has been extended to solve general formulation with multiple blocks of variables; see [74] and the references therein for more information. In this section we shall discuss the iteration complexity of the ADMM for multi-block optimization with a nonseparable objective function. In particular, the problem that we consider is as follows:

$$\begin{array}{ll} \min & f(x_1, x_2, \ldots, x_n) + \sum_{i=1}^n h_i(x_i) \\ \text{s.t.} & A_1x_1 + A_2x_2 + \cdots + A_nx_n = b, \\ & x_i \in \mathcal{X}_i, i = 1, 2, \ldots, n \end{array} \tag{3.15}$$

where $A_i \in \mathbb{R}^{m\times p_i}$, $b \in \mathbb{R}^m$, $\mathcal{X}_i \subset \mathbb{R}^{p_i}$ are closed convex sets, and f, h_i $i = 1, \ldots, n$, are convex closed functions. Note that many important applications are in the form of (3.15), e.g. multi-stage stochastic programming. Accordingly, the ADMM algorithm for solving the problem (3.15) is:

The Multi-block ADMM

Initialize with $x_i^0 \in \mathcal{X}_i, i = 1, \ldots, n$, and λ^0

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

$$
\begin{aligned}
x_1^{k+1} &= \arg\min_{x_1 \in \mathcal{X}_1} \mathcal{L}_\gamma(x_1, x_2^k, \ldots, x_n^k, \lambda^k) + \tfrac{1}{2}\|x_1 - x_1^k\|_{H_1}^2; \\
x_2^{k+1} &= \arg\min_{x_2 \in \mathcal{X}_2} \mathcal{L}_\gamma(x_1^{k+1}, x_2, x_3^k \ldots, x_n^k, \lambda^k) + \tfrac{1}{2}\|x_2 - x_2^k\|_{H_2}^2; \\
&\vdots \\
x_i^{k+1} &= \arg\min_{x_i \in \mathcal{X}_i} \mathcal{L}_\gamma(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i, x_{i+1}^k \ldots, x_n^k, \lambda^k) + \tfrac{1}{2}\|x_i - x_i^k\|_{H_i}^2; \\
&\vdots \\
x_n^{k+1} &= \arg\min_{x_n \in \mathcal{X}_n} \mathcal{L}_\gamma(x_1^{k+1}, \ldots, x_{n-1}^{k+1}, x_n, \lambda^k) + \tfrac{1}{2}\|x_n - x_n^k\|_{H_n}^2; \\
\lambda^{k+1} &= \lambda^k - \beta(A_1 x_1^{k+1} + A_2 x_2^{k+1} + \cdots + A_n x_n^{k+1}).
\end{aligned}
$$

end for

where H_i, $i = 1, \ldots, n$, are pre-specified positive semidefinite matrices, γ is the augmented Lagrangian constant, and β is the dual stepsize. An $O(1/N)$ convergence rate of the ADMM can still be shown to hold for this general problem. In the following subsection, we sketch a convergence rate analysis highlighting the key components and steps. The details, however, will be omitted for succinctness.

Let us start with the assumptions.

Assumption 3.3.1 *The functions h_i, $i = 2, \ldots, n$, are strongly convex with parameters $\sigma_i > 0$:*

$$h_i(y) \geq h_i(x) + (y - x)^\top h_i'(x) + \frac{\sigma_i}{2}\|y - x\|^2,$$

where $h_i'(x) \in \partial h_i(x)$ is in subdifferential of $h_i(x)$.

Assumption 3.3.2 *The gradient of function $f(x_1, x_2, \ldots, x_n)$ is Lipschitz continuous with parameter $L \geq 0$:*

$$\|\nabla f(x_1', x_2', \ldots, x_n') - \nabla f(x_1, x_2, \ldots, x_n)\| \leq L\|(x_1' - x_1, x_2' - x_2, \ldots, x_n' - x_n)\|$$

for all $(x_1', x_2', \ldots, x_n'), (x_1, x_2, \ldots, x_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.

In all the following propositions and theorems, we denote $w^k = (x_1^k, \ldots, x_n^k, \lambda^k)$ to be the iterates generated by ADMM, and $u = (x_1, \ldots, x_n)$.

Proposition 3.3.1 *Suppose that there are γ, β and δ satisfying*

$$\frac{n-1}{2} \max_{2 \leq i \leq n} \left\{\lambda_{\max}(A_i^\top A_i)\right\} \gamma + \delta \leq \min_{2 \leq i \leq n} \sigma_i.$$

Moreover, suppose that the matrices H_i, $i = 2, \ldots, n$, satisfy

$$H_i^s := H_i - \left(L + \frac{(n-i+1)(n+i-2)L^2}{8\delta} \right) I_{p_i} \succeq 0 \quad \forall\, 2 \leq i \leq n.$$

Let $(x_1^{k+1}, \ldots, x_n^{k+1}, \lambda^{k+1}) \in \Omega$ be the sequence generated by ADMM. Then, for $u^* = (x_1^*, \ldots, x_n^*) \in \Omega^*$ and $\lambda \in \mathbb{R}^m$, the following inequality holds

$$\begin{aligned}
& h(u^*) - h(u^{k+1}) + \begin{pmatrix} x_1^* - x_1^{k+1} \\ \vdots \\ x_n^* - x_n^{k+1} \\ \lambda - \lambda^{k+1} \end{pmatrix}^\top \begin{pmatrix} -A_1^\top \lambda^{k+1} \\ \vdots \\ -A_n^\top \lambda^{k+1} \\ \sum_{i=1}^n A_i x_i^{k+1} - b \end{pmatrix} \\
& + \frac{\gamma}{2} \sum_{i=2}^n \left(\left\| \sum_{j=1}^{i-1} A_j x_j^* + \sum_{j=i}^n A_j x_j^k - b \right\|^2 - \left\| \sum_{j=1}^{i-1} A_j x_j^* + \sum_{j=i}^n A_j x_j^{k+1} - b \right\|^2 \right) \\
& + \frac{1}{2\beta} \left(\left\| \lambda - \lambda^k \right\|^2 - \left\| \lambda - \lambda^{k+1} \right\|^2 \right) + \frac{1}{2} \sum_{i=1}^n \left(\|x_i^* - x_i^k\|_{H_i}^2 - \|x_i^* - x_i^{k+1}\|_{H_i}^2 \right) \\
\geq \; & \left(\frac{\gamma - \beta}{2\beta^2} \right) \|\lambda^k - \lambda^{k+1}\|^2 + \frac{1}{2} \sum_{i=1}^3 \|x_i^{k+1} - x_i^k\|_{H_i^s}^2.
\end{aligned}$$

The following proposition exhibits an important relationship between two consecutive iterates w^k and w^{k+1} from which the convergence readily follows.

Proposition 3.3.2 *Let w^k be the sequence generated by the ADMM, then*

$$\begin{aligned}
& \frac{\gamma}{2} \sum_{i=2}^n \left(\|\mathcal{L}_i(w^*, w^k)\|^2 - \|\mathcal{L}_i(w^*, w^{k+1})\|^2 \right) \\
& + \|w^* - w^k\|_{\hat{\mathcal{M}}}^2 - \|w^* - w^{k+1}\|_{\hat{\mathcal{M}}}^2 - \|w^k - w^{k+1}\|_{\mathcal{H}}^2 \geq 0,
\end{aligned}$$

where $\mathcal{L}_i(w^*, w) := \sum_{j=1}^{i-1} A_j x_j^* + \sum_{j=i}^n A_j x_j - b$, $i = 2, ..., n$, and

$$\hat{\mathcal{M}} = \operatorname{diag}\left(\frac{1}{2} H_1, \ldots, \frac{1}{2} H_n, \frac{1}{\beta} I_m \right), \mathcal{H} = \operatorname{diag}\left(\frac{1}{2} H_1, \frac{1}{2} H_2^s, \ldots, \frac{1}{2} H_n^s, \frac{\gamma - \beta}{2\beta^2} I_m \right).$$

Propositions 3.3.1 and 3.3.2 lead to the following theorem:

Theorem 3.3.3 *Under the assumptions of Propositions 3.3.1 and 3.3.2, and*

$$\mathcal{H} = \operatorname{diag}\left(\frac{1}{2}H_1, \frac{1}{2}H_2^s, \ldots, \frac{1}{2}H_n^s, \frac{\gamma-\beta}{2\beta^2}I_m\right) \succ 0,$$

we conclude that the sequence $\{w^k\}$ *generated by the ADMM converges to an optimal solution. Moreover, for any integer* $t > 0$ *let*

$$\bar{w}_t := \frac{1}{t}\sum_{k=0}^{t-1} w^{k+1},$$

and for any $\rho > 0$ *we have*

$$\begin{aligned} & h(\bar{u}_N) - h(u^*) + \rho\left\|\sum_{i=1}^n A_i\bar{x}_N - b\right\| \\ \leq\ & \frac{1}{2N}\left(\gamma\sum_{i=1}^{n-1}\left\|\sum_{j=i+1}^n A_j(x_j^* - x_j^0)\right\|^2 + \sum_{i=1}^n \|x_i^* - x_i^0\|_{H_i}^2 + \frac{1}{\beta}\left(\rho + \|\lambda^0\|\right)^2\right). \end{aligned}$$

3.4 Concluding Remarks

In [9], the following model is considered

$$\min\ \ f(x) + g(y) + H(x,y), \tag{3.16}$$

which can be regarded as (3.1) without constraints, and the so-called *proximal alternating linearized minimization* (PALM) algorithm is proposed. The main focus of [9] is to analyze the convergence of PALM for a class of nonconvex problems based on the Kurdyka-Łojasiewicz property. In that regard, it has an entirely different aim. We note however, that PALM is similar to APGMM applied to (3.16) when there is no coupling linear constraint. On the linearized gradient part, one noticeable difference is that APGMM operates in a Jacobian fashion while PALM is Gauss-Seidel. If the computation of gradient is costly, then the Jacobian style is cheaper to implement. As shown in [9], PALM can be extended to allow multiple blocks. Similarly, APGMM is also extendable to solve (3.15). The same is true for the other variations of the ADMM proposed in this chapter. It remains a future research topic to establish the convergence rate of such types of first-order algorithms. Other future research topics include the

study of first-order algorithms for (3.1) where the objective is non-convex but satisfies the Kurdyka-Łojasiewicz property. It is also interesting to consider stochastic programming models studied in [38], but now allowing the objective function to be nonseparable.

3.5 Proofs of the Convergence Theorems

3.5.1 Proof of Theorem 3.2.2

We have $F(w) = \begin{pmatrix} 0 & 0 & -B \\ 0 & 0 & -A \\ A & B & 0 \end{pmatrix} \begin{pmatrix} y \\ x \\ \lambda \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ b \end{pmatrix}$, for any w_1 and w_2, and so

$$(w_1 - w_2)^\top (F(w_1) - F(w_2)) = 0.$$

Expanding on this identity, we have for any $w^0, w^1, \cdots, w^{t-1}$ and $\bar{w} = \frac{1}{t}\sum_{k=0}^{t-1} w^k$, that

$$(\bar{w} - w)^\top F(\bar{w}) = \frac{1}{t}\sum_{k=0}^{t-1} (w^k - w)^\top F(w^k). \tag{3.17}$$

We begin our analysis with the following property of the ADMM algorithm.

Proposition 3.5.1 *Suppose h_2 is strongly convex with parameter $\sigma > 0$. Let $\{\tilde{w}^k\}$ be defined by (3.7), and the matrices Q, M, P be given in (3.6). First of all, for any $w \in \Omega$, we have*

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq\ & (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) - \left(\left(\frac{L}{2} + \frac{L^2}{2\sigma} \right) \|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k) \right). \end{aligned} \tag{3.18}$$

Furthermore,

$$(w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) = \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2 \right) + \frac{1}{2}\|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2. \tag{3.19}$$

Proof. By the optimality condition of the two subproblems in ADMM, we have

$$\begin{aligned}
&(x - x^{k+1})^\top \Big[\nabla_x f(x^{k+1}, y^k) + h_1'(x^{k+1}) \\
&-A^\top(\lambda^k - \gamma(Ax^{k+1} + By^k - b)) + G(x^{k+1} - x^k)\Big] \\
\geq\ & 0 \quad \forall x \in \mathcal{X},
\end{aligned}$$

where $h_1'(x^{k+1}) \in \partial h_1(x^{k+1})$, and

$$\begin{aligned}
&(y - y^{k+1})^\top \Big[\nabla_y f(x^{k+1}, y^{k+1}) + h_2'(y^{k+1}) \\
&-B^\top(\lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)) + H(y^{k+1} - y^k)\Big] \\
\geq\ & 0 \quad \forall y \in \mathcal{Y}
\end{aligned}$$

where $h_2'(x^{k+1}) \in \partial h_2(x^{k+1})$.

Note that $\tilde{\lambda}^k = \lambda^k - \gamma(Ax^{k+1} + By^k - b)$. The above two inequalities can be rewritten as

$$(x - \tilde{x}^k)^\top \Big[\nabla_x f(\tilde{x}^k, y^k) + h_1'(\tilde{x}^k) - A^\top\tilde{\lambda}^k + G(\tilde{x}^k - x^k)\Big] \geq 0 \quad \forall x \in \mathcal{X}, \tag{3.20}$$

and

$$(y - \tilde{y}^k)^\top \Big[\nabla_y f(\tilde{x}^k, \tilde{y}^k) + h_2'(\tilde{y}^k) - B^\top\tilde{\lambda}^k + \gamma B^\top B(\tilde{y}^k - y^k) + H(\tilde{y}^k - y^k)\Big] \geq 0 \quad \forall y \in \mathcal{Y}. \tag{3.21}$$

Observe the following chain of inequalities

$$\begin{aligned}
& (x-\tilde{x}^k)^\top \nabla_x f(\tilde{x}^k, y^k) + (y-\tilde{y}^k)^\top \nabla_y f(\tilde{x}^k, \tilde{y}^k) \\
= \quad & (x-\tilde{x}^k)^\top \nabla_x f(\tilde{x}^k, y^k) + (y-\tilde{y}^k)^\top \nabla_y f(\tilde{x}^k, y^k) \\
& +(y-\tilde{y}^k)^\top (\nabla_y f(\tilde{x}^k, \tilde{y}^k) - \nabla_y f(\tilde{x}^k, y^k)) \\
\leq \quad & (x-\tilde{x}^k)^\top \nabla_x f(\tilde{x}^k, y^k) + (y-\tilde{y}^k)^\top \nabla_y f(\tilde{x}^k, y^k) + L\|y-\tilde{y}^k\|\|y^k-\tilde{y}^k\| \\
= \quad & (x-\tilde{x}^k)^\top \nabla_x f(\tilde{x}^k, y^k) + (y-y^k)^\top \nabla_y f(\tilde{x}^k, y^k) \\
& +(y^k-\tilde{y}^k)^\top \nabla_y f(\tilde{x}^k, y^k) + L\|y-\tilde{y}^k\|\|y^k-\tilde{y}^k\| \\
\leq \quad & f(x,y) - f(\tilde{x}^k, y^k) - (\tilde{y}^k - y^k)^\top \nabla_y f(\tilde{x}^k, y^k) + L\|y-\tilde{y}^k\|\|y^k-\tilde{y}^k\| \\
\overset{(3.4)}{\leq} \quad & f(x,y) - f(\tilde{x}^k, \tilde{y}^k) + \frac{L}{2}\|y^k-\tilde{y}^k\|^2 + L\|y-\tilde{y}^k\|\|y^k-\tilde{y}^k\| \\
\leq \quad & f(x,y) - f(\tilde{x}^k, \tilde{y}^k) + \frac{L}{2}\|y^k-\tilde{y}^k\|^2 + \frac{\sigma}{2}\|y-\tilde{y}^k\|^2 + \frac{L^2}{2\sigma}\|y^k-\tilde{y}^k\|^2. \qquad (3.22)
\end{aligned}$$

Since

$$(A\tilde{x}^k + B\tilde{y}^k - b) - B(\tilde{y}^k - y^k) - \frac{1}{\gamma}(\lambda^k - \tilde{\lambda}^k) = 0,$$

we have

$$(\lambda - \tilde{\lambda}^k)^\top \left(A\tilde{x}^k + B\tilde{y}^k - b\right) = (\lambda - \tilde{\lambda}^k)^\top \left(-B(y^k - \tilde{y}^k) + \frac{1}{\gamma}(\lambda^k - \tilde{\lambda}^k)\right). \qquad (3.23)$$

By the strong convexity of the function $h_2(y)$, we have

$$(y-\tilde{y}^k)^\top h_2'(\tilde{y}^k) \leq h_2(y) - h_2(\tilde{y}^k) - \frac{\sigma}{2}\|y-\tilde{y}^k\|^2. \qquad (3.24)$$

Because of the convexity of $h_1(x)$ and combining (3.24), (3.23), (3.22), (3.21) and (3.20), we have

$$\begin{aligned}
& h(u) - h(\tilde{u}^k) + \left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k-\tilde{y}^k\|^2 + (y-\tilde{y}^k)^\top H(\tilde{y}^k - y^k) \\
& + \begin{pmatrix} x-\tilde{x}^k \\ y-\tilde{y}^k \\ \lambda-\tilde{\lambda}^k \end{pmatrix}^\top \left[\begin{pmatrix} -A^\top\tilde{\lambda}^k \\ -B^\top\tilde{\lambda}^k \\ A\tilde{x}^k + B\tilde{y}^k - b \end{pmatrix} - \begin{pmatrix} G(x^k-\tilde{x}^k) \\ \gamma B^\top B(y^k-\tilde{y}^k) \\ -B(y^k-\tilde{y}^k) + \frac{1}{\gamma}(\lambda^k-\tilde{\lambda}^k) \end{pmatrix} \right] \geq 0
\end{aligned}$$

for any $w \in \Omega$ and $\tilde{w}^k$. By definition of Q, (3.18) of Proposition 3.5.1 follows. For (3.19), due to the similarity, we refer to Lemma 3.2 in [57] (noting the matrices Q, P and M).

□

The following theorem exhibits an important relationship between two consecutive iterates w^k and w^{k+1} from which the convergence would follow.

Proposition 3.5.2 *Let w^k be the sequence generated by the ADMM, $\tilde{w}^k$ be defined as in (3.7) and H satisfy $H_s := H - \left(L + \frac{L^2}{\sigma}\right) I_q \succeq 0$. Then the following holds*

$$\frac{1}{2}\left(\|w^* - w^k\|_{\hat{M}}^2 - \|w^* - w^{k+1}\|_{\hat{M}}^2\right) - \frac{1}{2}\|w^k - \tilde{w}^k\|_{H_d}^2 \geq 0, \tag{3.25}$$

where

$$\hat{H} = \gamma B^\top B + H, \quad \hat{M} = \begin{pmatrix} G & 0 & 0 \\ 0 & \hat{H} & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix}, \quad \text{and } H_d = \begin{pmatrix} G & 0 & 0 \\ 0 & H_s & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix}. \tag{3.26}$$

Proof. It follows from Proposition 3.5.1 that

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq\ & (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) - \left(\left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right) \\ =\ & \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2\right) + \frac{1}{2}\|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2 \\ & - \left(\left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right). \end{aligned} \tag{3.27}$$

Note that $H_s := H - (L + \frac{L^2}{\sigma})I_q \succeq 0$, we have the following

$$\begin{aligned} & \left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k) \\ =\ & \left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + \frac{1}{2}\left(\|y - y^k\|_H^2 - \|y - \tilde{y}^k\|_H^2 - \|y^k - \tilde{y}^k\|_H^2\right) \\ =\ & \frac{1}{2}\left(\|y - y^k\|_H^2 - \|y - \tilde{y}^k\|_H^2\right) - \frac{1}{2}\|y^k - \tilde{y}^k\|_{H_s}^2. \end{aligned} \tag{3.28}$$

Thus, combining (3.27) and (3.28) we have

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq \; & \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2\right) - \frac{1}{2}\left(\|y - y^k\|_H^2 - \|y - \tilde{y}^k\|_H^2\right) \\ & + \frac{1}{2}\|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2}\|y^k - \tilde{y}^k\|_{H_s}^2 + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2. \end{aligned} \tag{3.29}$$

By the definition of $\hat{M}$ and H_d according to (3.26), it follows from (3.29) that

$$h(\tilde{u}^k) - h(u) + (\tilde{w}^k - w)^\top F(\tilde{w}^k) \leq \frac{1}{2}\left(\|w - w^k\|_{\hat{M}}^2 - \|w - w^{k+1}\|_{\hat{M}}^2\right) - \frac{1}{2}\|w^k - \tilde{w}^k\|_{H_d}^2. \tag{3.30}$$

Letting $w = w^*$ in (3.30) we have

$$h(\tilde{u}^k) - h(u^*) + (\tilde{w}^k - w^*)^\top F(\tilde{w}^k) \leq \frac{1}{2}\left(\|w^* - w^k\|_{\hat{M}}^2 - \|w^* - w^{k+1}\|_{\hat{M}}^2\right) - \frac{1}{2}\|w^k - \tilde{w}^k\|_{H_d}^2. \tag{3.31}$$

By the monotonicity of F and using the optimality of w^*, we have

$$\begin{aligned} & \frac{1}{2}\left(\|w^* - w^k\|_{\hat{M}}^2 - \|w^* - w^{k+1}\|_{\hat{M}}^2\right) - \frac{1}{2}\|w^k - \tilde{w}^k\|_{H_d}^2 \\ \geq \; & h(\tilde{u}^k) - h(u^*) + (\tilde{w}^k - w^*)^\top F(\tilde{w}^k) \\ \geq \; & h(\tilde{u}^k) - h(u^*) + (\tilde{w}^k - w^*)^\top F(w^*) \\ \geq \; & 0, \end{aligned}$$

which completes the proof. □

Proof of Theorem 3.2.2.

Proof. First, according to (3.25), it holds that $\{w^k\}$ is bounded and

$$\lim_{k \to \infty} \|w^k - \tilde{w}^k\|_{H_d} = 0. \tag{3.32}$$

Thus, those two sequences have the same cluster points: For any $w^{k_n} \to w^\infty$, by (3.32) we also have $\tilde{w}^{k_n} \to w^\infty$. Applying inequality (3.18) to $\{w^{k_n}\}, \{\tilde{w}^{k_n}\}$ and taking the limit, it yields that

$$h(u) - h(u^\infty) + (w - w^\infty)^\top F(w^\infty) \geq 0. \tag{3.33}$$

Consequently, the cluster point w^∞ is an optimal solution. Since (3.25) is true for any

optimal solution w^*, it also holds for w^∞, and that implies w^k will converge to w^∞.

Recall (3.18) and (3.19) in Proposition 3.5.1, those would imply that:

$$\begin{aligned}
& h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\
\geq\ & (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) \\
& - \left(\left(\frac{L}{2} + \frac{L^2}{2\sigma} \right) \|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k) \right) \\
\geq\ & \frac{1}{2} \left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2 \right) \\
& - \left(\left(\frac{L}{2} + \frac{L^2}{2\sigma} \right) \|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k) \right).
\end{aligned} \tag{3.34}$$

Furthermore, since $H - \left(L + \frac{L^2}{\sigma} \right) I_q \succeq 0$, we have

$$\begin{aligned}
& \left(\frac{L}{2} + \frac{L^2}{2\sigma} \right) \|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k) \\
=\ & \left(\frac{L}{2} + \frac{L^2}{2\sigma} \right) \|y^k - \tilde{y}^k\|^2 + \frac{1}{2} \left(\|y - y^k\|_H^2 - \|y - \tilde{y}^k\|_H^2 - \|y^k - \tilde{y}^k\|_H^2 \right) \\
\leq\ & \frac{1}{2} \left(\|y - y^k\|_H^2 - \|y - \tilde{y}^k\|_H^2 \right).
\end{aligned} \tag{3.35}$$

Thus, combining (3.34) and (3.35) leads to

$$\begin{aligned}
& h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\
\geq\ & \frac{1}{2} \left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2 \right) - \frac{1}{2} \left(\|y - y^k\|_H^2 - \|y - \tilde{y}^k\|_H^2 \right).
\end{aligned} \tag{3.36}$$

By the definition of M in (3.6) and denoting $\hat{H} = \gamma B^\top B + H$, (3.36) leads to

$$\begin{aligned}
& h(\tilde{u}^k) - h(u) + (\tilde{w}^k - w)^\top F(\tilde{w}^k) \\
\leq\ & \frac{1}{2} \left(\|x - x^k\|_G^2 - \|x - x^{k+1}\|_G^2 \right) + \frac{1}{2} \left(\|y - y^k\|_{\hat{H}}^2 - \|y - y^{k+1}\|_{\hat{H}}^2 \right) \\
& + \frac{1}{2\gamma} \left(\|\lambda - \lambda^k\|^2 - \|\lambda - \lambda^{k+1}\|^2 \right).
\end{aligned} \tag{3.37}$$

Before proceeding, let us introduce $\bar{w}_n := \frac{1}{n} \sum_{k=0}^{n-1} \tilde{w}^k$. Moreover, recall the definition

of $\bar{u}_n$ in (3.14), we have

$$\bar{u}_n = \frac{1}{n}\sum_{k=1}^{n} u^k = \frac{1}{n}\sum_{k=0}^{n-1} \tilde{u}^k.$$

Now, summing the inequality (3.37) over $k = 0, 1, \ldots, N-1$ yields

$$\begin{aligned} & h(\bar{u}_N) - h(u) + (\bar{w}_N - w)^\top F(\bar{w}_N) \\ \leq\ & \frac{1}{N}\sum_{k=0}^{N-1} h(\tilde{u}^k) - h(u) + \frac{1}{N}\sum_{k=0}^{N-1} (\tilde{w}^k - w)^\top F(\tilde{w}^k) \\ \leq\ & \frac{1}{2N}\left(\|x - x^0\|_G^2 + \|y - y^0\|_{\hat{H}}^2 + \frac{1}{\gamma}\|\lambda - \lambda^0\|^2 \right), \end{aligned} \tag{3.38}$$

where the first inequality is due to the convexity of h and (3.17).

Note the above inequality is true for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $\lambda \in \mathbb{R}^m$, hence it is also true for any optimal solution x^*, y^*, and $\mathcal{B}_\rho = \{\lambda : \|\lambda\| \leq \rho\}$. As a result,

$$\begin{aligned} & \sup_{\lambda \in \mathcal{B}_\rho} \left\{ h(\bar{u}_N) - h(u^*) + (\bar{w}_N - w^*)^\top F(\bar{w}_N) \right\} \\ =\ & \sup_{\lambda \in \mathcal{B}_\rho} \Big\{ h(\bar{u}_N) - h(u^*) + (\bar{x}_N - x^*)^\top (-A^\top \bar{\lambda}_N) + (\bar{y}_N - y^*)^\top (-B^\top \bar{\lambda}_N) \\ & + (\bar{\lambda}_N - \lambda)^\top (A\bar{x}_N + B\bar{y}_N - b) \Big\} \\ =\ & \sup_{\lambda \in \mathcal{B}_\rho} \left\{ h(\bar{u}_N) - h(u^*) + \bar{\lambda}_N^\top (Ax^* + By^* - b) - \lambda^\top (A\bar{x}_N + B\bar{y}_N - b) \right\} \\ =\ & \sup_{\lambda \in \mathcal{B}_\rho} \left\{ h(\bar{u}_N) - h(u^*) - \lambda^\top (A\bar{x}_N + B\bar{y}_N - b) \right\} \\ =\ & h(\bar{u}_N) - h(u^*) + \rho \|A\bar{x}_N + B\bar{y}_N - b\|, \end{aligned} \tag{3.39}$$

which, combined with (3.38), implies that

$$\begin{aligned} & h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\ \leq\ & \frac{1}{2N}\left(\|x^* - x^0\|_G^2 + \|y^* - y^0\|_{\hat{H}}^2 + \frac{1}{\gamma} \sup_{\lambda \in \mathcal{B}_\rho} \|\lambda - \lambda^0\|^2 \right), \end{aligned}$$

and so by optimizing over $(x^*, y^*) \in \mathcal{X}^* \times \mathcal{Y}^*$ we have

$$\begin{aligned} & h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\ \leq \quad & \frac{1}{2N}\left(\mathrm{dist}(x^0, \mathcal{X}^*)_G^2 + \mathrm{dist}(y^0, \mathcal{Y}^*)_{\hat{H}}^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right). \end{aligned} \tag{3.40}$$

This completes the proof. □

3.5.2 Proof of Theorem 3.2.3

Similar to the analysis for ADMM, we need the following proposition in the analysis of APGMM.

Proposition 3.5.3 *Let $\{\tilde{w}^k\}$ be defined by (3.7), and the matrices Q, M, P be given as in (3.6). For any $w \in \Omega$, we have*

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq \quad & (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) - \left(\frac{L}{2}(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2) + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right) \\ = \quad & \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2\right) + \frac{1}{2}\|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2 \\ & - \left(\frac{L}{2}\left(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2\right) + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right). \end{aligned} \tag{3.41}$$

Proof. First, by the optimality condition of the two subproblems in APGMM, we have

$$\begin{aligned} & (x - x^{k+1})^\top \Big[\nabla_x f(x^k, y^k) + h_1'(x^{k+1}) \\ & -A^\top(\lambda^k - \gamma(Ax^{k+1} + By^k - b)) + G(x^{k+1} - x^k)\Big] \\ \geq \quad & 0, \quad \forall x \in \mathcal{X}, \end{aligned}$$

and

$$\begin{aligned} & (y - y^{k+1})^\top \Big[\nabla_y f(x^k, y^k) + h_2'(y^{k+1}) \\ & -B^\top(\lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)) + H(y^{k+1} - y^k)\Big] \\ \geq \quad & 0, \quad \forall y \in \mathcal{Y}. \end{aligned}$$

Note that $\tilde{\lambda}^k = \lambda^k - \gamma(Ax^{k+1} + By^k - b)$, and by the definition of $\tilde{w}^k$, the above two

inequalities are equivalent to

$$(x-\tilde{x}^k)^\top \left[\nabla_x f(x^k,y^k) + h_1'(\tilde{x}^k) - A^\top\tilde{\lambda}^k + G(\tilde{x}^k - x^k)\right] \geq 0 \quad \forall x \in \mathcal{X}, \tag{3.42}$$

and

$$(y-\tilde{y}^k)^\top \left[\nabla_y f(x^k,y^k) + h_2'(\tilde{y}^k) - B^\top\tilde{\lambda}^k + \gamma B^\top B(\tilde{y}^k - y^k) + H(\tilde{y}^k - y^k)\right] \geq 0, \forall y \in \mathcal{Y}. \tag{3.43}$$

Notice that

$$\begin{aligned}
& (x-\tilde{x}^k)^\top\nabla_x f(x^k,y^k) + (y-\tilde{y}^k)^\top\nabla_y f(x^k,y^k) \\
= \quad & (x-x^k)^\top\nabla_x f(x^k,y^k) + (y-y^k)^\top\nabla_y f(x^k,y^k) \\
& +(x^k-\tilde{x}^k)^\top\nabla_x f(x^k,y^k) + (y^k-\tilde{y}^k)^\top\nabla_y f(x^k,y^k) \\
\leq \quad & f(x,y) - f(x^k,y^k) - (\tilde{x}^k - x^k)^\top\nabla_x f(x^k,y^k) - (\tilde{y}^k - y^k)^\top\nabla_y f(x^k,y^k) \\
\overset{(3.4)}{\leq} \quad & f(x,y) - f(\tilde{x}^k,\tilde{y}^k) + \frac{L}{2}\left(\|x^k-\tilde{x}^k\|^2 + \|y^k-\tilde{y}^k\|^2\right).
\end{aligned} \tag{3.44}$$

Besides, we also have

$$(A\tilde{x}^k + B\tilde{y}^k - b) - B(\tilde{y}^k - y^k) - \frac{1}{\gamma}\left(\lambda^k - \tilde{\lambda}^k\right) = 0.$$

Thus

$$(\lambda-\tilde{\lambda}^k)^\top(A\tilde{x}^k + B\tilde{y}^k - b) = (\lambda-\tilde{\lambda}^k)^\top\left(-B(y^k-\tilde{y}^k) + \frac{1}{\gamma}(\lambda^k-\tilde{\lambda}^k)\right). \tag{3.45}$$

By the convexity of $h_1(x)$ and $h_2(y)$, combining (3.45), (3.44), (3.43) and (3.42), we have

$$\begin{aligned}
& h(u) - h(\tilde{u}^k) + \frac{L}{2}\left(\|x^k-\tilde{x}^k\|^2 + \|y^k-\tilde{y}^k\|^2\right) + (y-\tilde{y}^k)^\top H(\tilde{y}^k - y^k) \\
& + \begin{pmatrix} x-\tilde{x}^k \\ y-\tilde{y}^k \\ \lambda-\tilde{\lambda}^k \end{pmatrix}^\top \left[\begin{pmatrix} -A^\top\tilde{\lambda}^k \\ -B^\top\tilde{\lambda}^k \\ A\tilde{x}^k + B\tilde{y}^k - b \end{pmatrix} - \begin{pmatrix} G(x^k-\tilde{x}^k) \\ \gamma B^\top B(y^k-\tilde{y}^k) \\ -B(y^k-\tilde{y}^k) + \frac{1}{\gamma}(\lambda^k-\tilde{\lambda}^k) \end{pmatrix}\right] \geq 0
\end{aligned}$$

for any $w \in \Omega$ and $\tilde{w}^k$.

By definition of Q, we have shown (3.41) in Proposition 3.5.3. The equality directly

follows from (3.19) in Proposition 3.5.1. □

With Proposition 3.5.3 in place, we can show Theorem 3.2.3 by exactly following the same steps as in the proof of Theorem 3.2.2, noting of course the altered assumptions on the matrices G and H. In the meanwhile, we also point out the following proposition which is similar to Proposition 3.5.2. Since most steps of the proofs are almost identical to that of the previous theorems, we omit the details for succinctness.

Proposition 3.5.4 *Let w^k be the sequence generated by the APGMM, and $\tilde{w}^k$ be as defined in* (3.7)*, and H and G are chosen so as to satisfy $H_s := H - LI_q \succ 0$ and $G_s := G - LI_p \succ 0$. Then the following holds*

$$\frac{1}{2}\left(\|w^* - w^k\|_{\hat{M}}^2 - \|w^* - w^{k+1}\|_{\hat{M}}^2\right) - \frac{1}{2}\|w^k - \tilde{w}^k\|_{H_d}^2 \geq 0,$$

where

$$\hat{M} = \begin{pmatrix} G & 0 & 0 \\ 0 & \hat{H} & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix}, \; H_d = \begin{pmatrix} G_s & 0 & 0 \\ 0 & H_s & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix}$$

and $\hat{H} = \gamma B^\top B + H$.

Theorem 3.2.3 follows from the above propositions.

3.5.3 Proof of Theorem 3.2.4

Similar to the analysis for APGMM, we do not need any strong convexity here, but we do need to assume that the gradients $\nabla_x h_1(x)$ and $\nabla_y h_2(y)$ are Lipschitz continuous. Without loss of generality, we further assume that the Lipschitz constant is the same as $\nabla f(x, y)$ which is L; that is,

$$\begin{aligned} &\|\nabla_x h_1(x_2) - \nabla_x h_1(x_1)\| \leq L\|x_2 - x_1\|, \; \forall\, x_1, x_2 \in \mathcal{X}, \\ &\|\nabla_y h_2(y_2) - \nabla_y h_2(y_1)\| \leq L\|y_2 - y_1\|, \; \forall\, y_1, y_2 \in \mathcal{Y}. \end{aligned} \tag{3.46}$$

Proposition 3.5.5 *Let $\{\tilde{w}^k\}$ be defined by* (3.7)*, and the matrices Q, M, P be as given in* (3.6)*, and $G := \gamma A^\top A + \frac{1}{\alpha} I_p$, $H := \frac{1}{\alpha} I_q - \gamma B^\top B \succeq 0$. First of all, for any $w \in \Omega$,*

we have

$$\begin{aligned}
& h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\
\geq\ & (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) - \Big(L(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2) + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\Big) \\
=\ & \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2\right) + \frac{1}{2}\|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2 \\
& - \Big(L(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2) + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\Big). \qquad (3.47)
\end{aligned}$$

Proof. First, by the optimality condition of the two subproblems in AGPMM, we have

$$\begin{aligned}
& (x - x^{k+1})^\top \Big[x^{k+1} - x^k + \alpha(\nabla_x f(x^k, y^k) + \nabla_y h_1(x^k) \\
& -A^\top(\lambda^k - \gamma(Ax^k + By^k - b)))\Big] \\
\geq\ & 0 \quad \forall x \in \mathcal{X},
\end{aligned}$$

and

$$\begin{aligned}
& (y - y^{k+1})^\top \Big[y^{k+1} - y^k + \alpha(\nabla_y f(x^k, y^k) + \nabla_y h_2(y^k) \\
& -B^\top(\lambda^k - \gamma(Ax^{k+1} + By^k - b)))\Big] \\
\geq\ & 0 \quad \forall y \in \mathcal{Y}.
\end{aligned}$$

Noting $\tilde{\lambda}^k = \lambda^k - \gamma(Ax^{k+1} + By^k - b)$ and the definition of $\tilde{w}^k$, the above two inequalities are respectively equivalent to

$$\begin{aligned}
& (x - \tilde{x}^k)^\top \Big[\nabla_x f(x^k, y^k) + \nabla_x h_2(x^k) \\
& -A^\top\tilde{\lambda}^k + \gamma A^\top A(\tilde{x}^k - x^k) + \frac{1}{\alpha}(\tilde{x}^k - x^k)\Big] \geq 0 \quad \forall x \in \mathcal{X}, \qquad (3.48)
\end{aligned}$$

and

$$(y - \tilde{y}^k)^\top \left[\nabla_y f(x^k, y^k) + \nabla_y h_2(y^k) - B^\top\tilde{\lambda}^k + \frac{1}{\alpha}(\tilde{y}^k - y^k)\right] \geq 0 \quad \forall y \in \mathcal{Y}. \qquad (3.49)$$

Similar to Proposition 3.5.3, we have

$$\begin{aligned} & (x-\tilde{x}^k)^\top \nabla_x f(x^k,y^k) + (y-\tilde{y}^k)^\top \nabla_y f(x^k,y^k) \\ \overset{(3.4)}{\leq} \quad & f(x,y) - f(\tilde{x}^k,\tilde{y}^k) + \frac{L}{2}\left(\|x^k-\tilde{x}^k\|^2 + \|y^k-\tilde{y}^k\|^2\right). \end{aligned} \tag{3.50}$$

Moreover, by (3.5) we have

$$\begin{aligned} (x-\tilde{x}^k)^\top \nabla_x h_1(x^k) &\leq h_1(x) - h_1(\tilde{x}^k) + \frac{L}{2}\|x^k-\tilde{x}^k\|^2 \\ (y-\tilde{y}^k)^\top \nabla_y h_2(y^k) &\leq h_2(y) - h_2(\tilde{y}^k) + \frac{L}{2}\|y^k-\tilde{y}^k\|^2. \end{aligned} \tag{3.51}$$

Besides,

$$(A\tilde{x}^k + B\tilde{y}^k - b) - B(\tilde{y}^k - y^k) - \frac{1}{\gamma}\left(\lambda^k - \tilde{\lambda}^k\right) = 0.$$

Thus

$$(\lambda-\tilde{\lambda}^k)^\top (A\tilde{x}^k + B\tilde{y}^k - b) = (\lambda-\tilde{\lambda}^k)^\top \left(-B(y^k-\tilde{y}^k) + \frac{1}{\gamma}\left(\lambda^k-\tilde{\lambda}^k\right)\right). \tag{3.52}$$

Combining (3.52), (3.51), (3.50), (3.49), and (3.48), and noticing that $G := \gamma A^\top A + \frac{1}{\alpha} I_p$, $H := \frac{1}{\alpha} I_q - \gamma B^\top B$, we have, for any $w \in \Omega$ and $\tilde{w}^k$, that

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + L(\|x^k-\tilde{x}^k\|^2 + \|y^k-\tilde{y}^k\|^2) + (y-\tilde{y}^k)^\top H(\tilde{y}^k - y^k) \\ & + \begin{pmatrix} x-\tilde{x}^k \\ y-\tilde{y}^k \\ \lambda-\tilde{\lambda}^k \end{pmatrix}^\top \left\{ \begin{pmatrix} -A^\top\tilde{\lambda}^k \\ -B^\top\tilde{\lambda}^k \\ A\tilde{x}^k + B\tilde{y}^k - b \end{pmatrix} - \begin{pmatrix} G(x^k-\tilde{x}^k) \\ \gamma B^\top B(y^k-\tilde{y}^k) \\ -B(y^k-\tilde{y}^k) + \frac{1}{\gamma}(\lambda^k-\tilde{\lambda}^k) \end{pmatrix} \right\} \geq 0. \end{aligned}$$

Using the definition of Q, (3.47) follows. In view of (3.19) in Proposition 3.5.1, the equality also readily follows. □

With Proposition 3.5.5, similar as before, we can show Theorem 3.2.4 by following the same approach as in the proof of Theorem 3.2.2. We skip the details here for succinctness.

Proposition 3.5.6 *Let w^k be the sequence generated by the AGPMM, $\tilde{w}^k$ be defined in (3.7) and $G := \gamma A^\top A + \frac{1}{\alpha} I_p$, $H := \frac{1}{\alpha} I_q - \gamma B^\top B$. Suppose that α satisfies that*

$H_s := H - 2LI_q \succ 0$ *and* $G_s := G - 2LI_p \succ 0$. *Then the following holds*

$$\frac{1}{2}\left(\|w^* - w^k\|^2_{\hat{M}} - \|w^* - w^{k+1}\|^2_{\hat{M}}\right) - \frac{1}{2}\|w^k - \tilde{w}^k\|^2_{H_d} \geq 0,$$

where

$$\hat{M} = \begin{pmatrix} G & 0 & 0 \\ 0 & \hat{H} & 0 \\ 0 & 0 & \frac{1}{\gamma}I_m \end{pmatrix},\ H_d = \begin{pmatrix} G_s & 0 & 0 \\ 0 & H_s & 0 \\ 0 & 0 & \frac{1}{\gamma}I_m \end{pmatrix},$$

and $\hat{H} = \gamma B^\top B + H$.

Theorem 3.2.4 now follows from the above propositions.

3.5.4 Proofs of Theorems 3.2.5 and 3.2.6

Proposition 3.5.7 *Let* $\{\tilde{w}^k\}$ *be defined by* (3.7)*, and the matrices* Q*,* M*,* P *be given in* (3.6)*. For any* $w \in \Omega$*, we have*

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq\ & (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) - \left(\frac{L}{2}\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right) \\ =\ & \frac{1}{2}\left(\|w - w^{k+1}\|^2_M - \|w - w^k\|^2_M\right) + \frac{1}{2}\|x^k - \tilde{x}^k\|^2_G + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2 \\ & - \left(\frac{L}{2}\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right). \end{aligned} \tag{3.53}$$

Proof. First, by the optimality condition of the two subproblems in ADM-PG, we have

$$\begin{aligned} & (x - x^{k+1})^\top \Big[\nabla_x f(x^{k+1}, y^k) + h_1'(x^{k+1}) \\ & -A^\top(\lambda^k - \gamma(Ax^{k+1} + By^k - b)) + G(x^{k+1} - x^k)\Big] \\ \geq\ & 0 \quad \forall x \in \mathcal{X}, \end{aligned}$$

and

$$\begin{aligned} & (y - y^{k+1})^\top \Big[\nabla_y f(x^{k+1}, y^k) + h_2'(y^{k+1}) \\ & -B^\top(\lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)) + H(y^{k+1} - y^k)\Big] \\ \geq\ & 0 \quad \forall y \in \mathcal{Y}. \end{aligned}$$

Noting $\tilde{\lambda}^k = \lambda^k - \gamma(Ax^{k+1} + By^k - b)$ and the definition of $\tilde{w}^k$, the above two inequalities are equivalent to

$$(x - \tilde{x}^k)^\top \left[\nabla_x f(\tilde{x}^k, y^k) + \nabla_x h_1(\tilde{x}^k) - A^\top \tilde{\lambda}^k + G(\tilde{x}^k - x^k)\right] \geq 0 \quad \forall x \in \mathcal{X}, \tag{3.54}$$

and

$$(y - \tilde{y}^k)^\top \left[\nabla_y f(\tilde{x}^k, y^k) + g_2(\tilde{y}^k) - B^\top \tilde{\lambda}^k + \gamma B^\top B(\tilde{y}^k - y^k) + H(\tilde{y}^k - y^k)\right] \geq 0, \forall y \in \mathcal{Y}. \tag{3.55}$$

Moreover,

$$\begin{aligned} & (x - \tilde{x}^k)^\top \nabla_x f(\tilde{x}^k, y^k) + (y - \tilde{y}^k)^\top \nabla_y f(\tilde{x}^k, y^k) \\ = \quad & (x - \tilde{x}^k)^\top \nabla_x f(\tilde{x}^k, y^k) + (y - y^k)^\top \nabla_y f(\tilde{x}^k, y^k) + (y^k - \tilde{y}^k)^\top \nabla_y f(\tilde{x}^k, y^k) \\ \leq \quad & f(x, y) - f(\tilde{x}^k, y^k) - (\tilde{y}^k - y^k)^\top \nabla_y f(\tilde{x}^k, y^k) \\ \overset{(3.4)}{\leq} \quad & f(x, y) - f(\tilde{x}^k, \tilde{y}^k) + \frac{L}{2}\|y^k - \tilde{y}^k\|^2. \end{aligned} \tag{3.56}$$

Besides,

$$(A\tilde{x}^k + B\tilde{y}^k - b) - B(\tilde{y}^k - y^k) - \frac{1}{\gamma}\left(\lambda^k - \tilde{\lambda}^k\right) = 0,$$

and so

$$(\lambda - \tilde{\lambda}^k)^\top (A\tilde{x}^k + B\tilde{y}^k - b) = (\lambda - \tilde{\lambda}^k)^\top \left(-B(y^k - \tilde{y}^k) + \frac{1}{\gamma}(\lambda^k - \tilde{\lambda}^k)\right). \tag{3.57}$$

By the convexity of $h_1(x)$ and $h_2(y)$, combining (3.57), (3.56), (3.55), and (3.54), we have

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + \frac{L}{2}\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k) \\ & + \begin{pmatrix} x - \tilde{x}^k \\ y - \tilde{y}^k \\ \lambda - \tilde{\lambda}^k \end{pmatrix}^\top \left[\begin{pmatrix} -A^\top \tilde{\lambda}^k \\ -B^\top \tilde{\lambda}^k \\ A\tilde{x}^k + B\tilde{y}^k - b \end{pmatrix} - \begin{pmatrix} G(x^k - \tilde{x}^k) \\ \gamma B^\top B(y^k - \tilde{y}^k) \\ -B(y^k - \tilde{y}^k) + \frac{1}{\gamma}(\lambda^k - \tilde{\lambda}^k) \end{pmatrix} \right] \geq 0 \end{aligned}$$

for any $w \in \Omega$ and $\tilde{w}^k$.

By similar derivations as in the proofs for Proposition 3.5.5, (3.53) follows. □

With Proposition 3.5.7 in place, we can prove Theorem 3.2.5 similarly as in the proof

of Theorem 3.2.2. We skip the details here for succinctness.

For ADM-GP we do not need strong convexity, but we do need to assume that the gradient $\nabla_y h_2(y)$ of $h_2(y)$ is Lipschitz continuous. Without loss of generality, we further assume that the Lipschitz constant of $\nabla_y h_2(y)$ is the same as $\nabla f(x,y)$ which is L:

$$\|\nabla_y h_2(y_2) - \nabla_y h_2(y_1)\| \leq L\|y_2 - y_1\|, \ \forall\, y_1, y_2 \in \mathcal{Y}. \tag{3.58}$$

Proposition 3.5.8 *Let $\{\tilde{w}^k\}$ be defined by (3.7), and the matrices Q, M, P be given in (3.6), and $H := \frac{1}{\alpha} I_q - \gamma B^\top B \succeq 0$. For any $w \in \Omega$, we have*

$$\begin{aligned}
& h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\
\geq\ & (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) - \left(L\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k) \right) \\
=\ & \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2 \right) + \frac{1}{2}\|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2 \\
& - \left(L\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k) \right).
\end{aligned} \tag{3.59}$$

Proof. By the optimality condition of the two subproblems in ADMM, we have

$$\begin{aligned}
& (x - x^{k+1})^\top \Big[\nabla_x f(x^{k+1}, y^k) + h_1'(x^{k+1}) \\
& -A^\top(\lambda^k - \gamma(Ax^{k+1} + By^k - b)) + G(x^{k+1} - x^k) \Big] \\
\geq\ & 0, \quad \forall x \in \mathcal{X}
\end{aligned}$$

and

$$\begin{aligned}
& (y - y^{k+1})^\top \Big[y^{k+1} - y^k + \alpha(\nabla_y f(x^{k+1}, y^k) + \nabla_y h_2(y^k) \\
& -B^\top(\lambda^k - \gamma(Ax^{k+1} + By^k - b))) \Big] \\
\geq\ & 0, \quad \forall y \in \mathcal{Y}.
\end{aligned}$$

Noting $\tilde{\lambda}^k = \lambda^k - \gamma(Ax^{k+1} + By^k - b)$ and the definition of $\tilde{w}^k$, the above two inequalities are equivalent to

$$(x - \tilde{x}^k)^\top \left[\nabla_x f(\tilde{x}^k, y^k) + h_1'(\tilde{x}^k) - A^\top \tilde{\lambda}^k + G(\tilde{x}^k - x^k) \right] \geq 0 \quad \forall x \in \mathcal{X}, \tag{3.60}$$

and

$$(y-\tilde{y}^k)^\top\left[\nabla_y f(\tilde{x}^k,y^k)+\nabla_y h_2(y^k)-B^\top\tilde{\lambda}^k+\frac{1}{\alpha}(\tilde{y}^k-y^k)\right]\geq 0\quad\forall y\in\mathcal{Y}. \tag{3.61}$$

Therefore,

$$\begin{aligned}
& (x-\tilde{x}^k)^\top\nabla_x f(\tilde{x}^k,y^k)+(y-\tilde{y}^k)^\top\nabla_y f(\tilde{x}^k,y^k)\\
= \;& (x-\tilde{x}^k)^\top\nabla_x f(\tilde{x}^k,y^k)+(y-y^k)^\top\nabla_y f(\tilde{x}^k,y^k)+(y^k-\tilde{y}^k)^\top\nabla_y f(\tilde{x}^k,y^k)\\
\leq \;& f(x,y)-f(\tilde{x}^k,y^k)-(\tilde{y}^k-y^k)^\top\nabla_y f(\tilde{x}^k,y^k)\\
\leq \;& f(x,y)-f(\tilde{x}^k,\tilde{y}^k)+\frac{L}{2}\|y^k-\tilde{y}^k\|^2.
\end{aligned} \tag{3.62}$$

Moreover, by (3.5), we have

$$(y-\tilde{y}^k)^\top\nabla_y h_2(y^k)\leq h_2(y)-h_2(\tilde{y}^k)+\frac{L}{2}\|y^k-\tilde{y}^k\|^2. \tag{3.63}$$

Since

$$A\tilde{x}^k+B\tilde{y}^k-b-B(\tilde{y}^k-y^k)-\frac{1}{\gamma}\left(\lambda^k-\tilde{\lambda}^k\right)=0,$$

we have

$$(\lambda-\tilde{\lambda}^k)^\top(A\tilde{x}^k+B\tilde{y}^k-b)=(\lambda-\tilde{\lambda}^k)^\top\left(-B(y^k-\tilde{y}^k)+\frac{1}{\gamma}(\lambda^k-\tilde{\lambda}^k)\right). \tag{3.64}$$

By the convexity of $h_1(x)$, combining (3.64), (3.63), (3.62), (3.61), (3.60), and noticing $H:=\frac{1}{\alpha}I_q-\gamma B^\top B$ for any $w\in\Omega$ and $\tilde{w}^k$ we have

$$\begin{aligned}
& h(u)-h(\tilde{u}^k)+L\|y^k-\tilde{y}^k\|^2+(y-\tilde{y}^k)^\top H(\tilde{y}^k-y^k)\\
& +\begin{pmatrix} x-\tilde{x}^k\\ y-\tilde{y}^k\\ \lambda-\tilde{\lambda}^k\end{pmatrix}^\top\left\{\begin{pmatrix} -A^\top\tilde{\lambda}^k\\ -B^\top\tilde{\lambda}^k\\ A\tilde{x}^k+B\tilde{y}^k-b\end{pmatrix}-\begin{pmatrix} G(x^k-\tilde{x}^k)\\ \gamma B^\top B(y^k-\tilde{y}^k)\\ -B(y^k-\tilde{y}^k)+\frac{1}{\gamma}(\lambda^k-\tilde{\lambda}^k)\end{pmatrix}\right\}\geq 0.
\end{aligned}$$

As a result, (3.59) follows. □

The proof of Theorem 3.2.6 follows a similar line of derivation as in the proof of Theorem 3.2.2, and we omit the details here.

Chapter 4

Randomized Primal-Dual Proximal Block Coordinate Updates

4.1 Introduction

In this chapter, we consider the following multi-block structured convex optimization model

$$\begin{aligned} \min_{x,y} \;\; & f(x_1, \cdots, x_N) + \sum_{i=1}^{N} u_i(x_i) + g(y_1, \cdots, y_M) + \sum_{j=1}^{M} v_j(y_j) \\ \text{s.t.} \; & \sum_{i=1}^{N} A_i x_i + \sum_{j=1}^{M} B_j y_j = b \\ & x_i \in \mathcal{X}_i,\, i = 1, \ldots, N;\;\; y_j \in \mathcal{Y}_j,\, j = 1, \ldots, M, \end{aligned} \tag{4.1}$$

where the variables $x = (x_1; \cdots; x_N)$ and $y = (y_1; \cdots; y_M)$ are naturally partitioned into N and M blocks respectively, $A = (A_1, \cdots, A_N)$ and $B = (B_1, \cdots, B_M)$ are block matrices, $\mathcal{X}_i$'s and $\mathcal{Y}_j$'s are some closed convex sets, f and g are smooth convex functions, and u_i's and v_j's are proper closed convex (possibly nonsmooth) functions.

4.1.1 Motivating examples

Optimization problems in the form of (4.1) have many emerging applications from various fields. For example, the constrained lasso (classo) problem that was first studied

by James *et al.* [61] as a generalization of the lasso problem, can be formulated as

$$\begin{array}{ll} \min\limits_{x} & \frac{1}{2}\|Ax-b\|_2^2+\tau\|x\|_1 \\ \text{s.t.} & Cx \le d, \end{array} \tag{4.2}$$

where $A \in \mathbb{R}^{m\times p}$, $b \in \mathbb{R}^m$ are the observed data, and $C \in \mathbb{R}^{n\times p}$, $d \in \mathbb{R}^n$ are the predefined data matrix and vector. Many widely used statistical models can be viewed as special cases of (4.2), including the monotone curve estimation, fused lasso, generalized lasso, and so on [61]. By partitioning the variable x into blocks as $x = (x_1; \cdots; x_K)$ where $x_i \in \mathbb{R}^{p_i}$ as well as other matrices and vectors in (4.2) correspondingly, and introducing another slack variable y, the classo problem can be transformed to

$$\begin{array}{ll} \min\limits_{x,y} & \frac{1}{2}\left\|\sum\limits_{i=1}^{K} A_i x_i - b\right\|_2^2 + \tau \sum\limits_{i=1}^{K} \|x_i\|_1 \\ \text{s.t.} & \sum\limits_{i=1}^{K} C_i x_i + y = d,\ y \ge 0, \end{array} \tag{4.3}$$

which is in the form of (4.1).

Another interesting example is the extended linear-quadratic programming [103] that can be formulated as

$$\begin{array}{ll} \min\limits_{x} & \frac{1}{2}x^\top P x + a^\top x + \max\limits_{s\in\mathcal{S}} \left\{(d-Cx)^\top s - \frac{1}{2}s^\top Q s\right\}, \\ \text{s.t.} & Ax \le b, \end{array} \tag{4.4}$$

where P and Q are symmetric positive semidefinite matrices, and $\mathcal{S}$ is a polyhedral set. Apparently, (4.4) includes quadratic programming as a special case. In general, its objective is a piece-wise linear-quadratic convex function. Let $g(s) = \frac{1}{2}s^\top Q s + \iota_S(s)$, where $\iota_{\mathcal{S}}$ denotes the indicator function of S. Then

$$\max_{s\in S}\left\{(d-Cx)^\top s - \frac{1}{2}s^\top Q s\right\} = g^*(d-Cx),$$

where g^* denotes the convex conjugate of g. Replacing $d - Cx$ by y and introducing slack variable z, we can equivalently write (4.4) into the form of (4.1):

$$\begin{array}{ll} \min\limits_{x,y,z} & \frac{1}{2}x^\top P x + a^\top x + g^*(y), \\ \text{s.t.} & Ax + z = b,\ z \ge 0,\ Cx + y = d, \end{array} \tag{4.5}$$

for which one can further partition the x-variable into a number of disjoint blocks.

Many other interesting applications in various areas can be formulated as optimization problems in the form of (4.1), including those arising from signal processing, image processing, machine learning and statistical learning; see [58, 23, 18, 39] and the references therein.

Finally, we mention that computing a point on the central path for a generic convex programming in block variables $(x_1; \cdots; x_N)$:

$$\begin{array}{ll} \min\limits_{x} & f(x_1, \cdots, x_N) \\ \text{s.t.} & \sum_{i=1}^{N} A_i x_i \leq b,\ x_i \geq 0,\ i = 1, 2, ..., N \end{array}$$

boils down to

$$\begin{array}{ll} \min\limits_{x,y} & f(x_1, \cdots, x_N) - \mu e^\top \ln x - \mu e^\top \ln y \\ \text{s.t.} & \sum_{i=1}^{N} A_i x_i + y = b, \end{array}$$

where $\mu > 0$ and $e^\top \ln v$ indicates the sum of the logarithm of all the components of v. This model is again in the form of (4.1).

4.1.2 Related works in the literature

One well-known approach for solving a linear constrained problem in the form of (4.1) is the augmented Lagrangian method, which iteratively updates the primal variable (x, y) by minimizing the augmented Lagrangian function in (4.7) and then the multiplier λ through dual gradient ascent. However, the linear constraint couples $x_1, \ldots, x_N$ and $y_1, \ldots, y_M$ all together, it can be very expensive to minimize the augmented Lagrangian function simultaneously with respect to all block variables.

It is very natural then, to use the multi-block structure of the problem. In fact, the multi-block ADMM updates the block variables sequentially, one at a time with the others fixed to their most recent values, followed by the update of multiplier. Specifically, for (4.1), it performs the following updates iteratively (by assuming the absence of the

coupled functions f and g):

$$\left\{\begin{array}{rcl} x_1^{k+1} & = & \arg\min_{x_1\in\mathcal{X}_1} \mathcal{L}_\rho(x_1, x_2^k, \cdots, x_N^k, y^k, \lambda^k), \\ & \vdots & \\ x_N^{k+1} & = & \arg\min_{x_N\in\mathcal{X}_N} \mathcal{L}_\rho(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N, y^k, \lambda^k), \\ y_1^{k+1} & = & \operatorname{argmin}_{y_1\in\mathcal{Y}_1} \mathcal{L}_\rho(x^{k+1}, y_1, y_2^k, \cdots, y_M^k, \lambda^k), \\ & \vdots & \\ y_M^{k+1} & = & \operatorname{argmin}_{y_M\in\mathcal{Y}_M} \mathcal{L}_\rho(x^{k+1}, y_1^{k+1}, \cdots, y_{M-1}^{k+1}, y_M, \lambda^k), \\ \lambda^{k+1} & = & \lambda^k - \rho(Ax^{k+1} + By^{k+1} - b), \end{array}\right. \tag{4.6}$$

where the augmented Lagrangian function is defined as:

$$\mathcal{L}_\rho(x, y, \lambda) = \sum_{i=1}^N u_i(x_i) + \sum_{j=1}^M v_j(y_j) - \lambda^\top (Ax + By - b) + \frac{\rho}{2} \|Ax + By - b\|^2 . \tag{4.7}$$

Besides the multi-block ADMM, our work also relates to another popular topic: the first-order primal-dual method for bilinear saddle-point problems. Below we briefly review the method and it convergence results. More complete discussion on the connections to our method will be provided after presenting our algorithm.

Primal-dual method for bilinear saddle-point problems

Recently, the work [24] generalizes the first-order primal-dual method in [15] to a randomized method for solving a class of saddle-point problems in the following form:

$$\min_{z\in Z} \left\{ h(z) + \max_{x\in X} \left\langle z, \sum_{i=1}^N A_i x_i \right\rangle - \sum_{i=1}^N u_i(x_i) \right\}, \tag{4.8}$$

where $x = (x_1; \ldots; x_N)$ and $X = X_1 \times \cdots \times X_N$. Let $Z = \mathbb{R}^p$ and $h(z) = -b^\top z$. Then it is easy to see that (4.8) is a saddle-point reformulation of the multi-block structured optimization problem

$$\min_{x\in X} \sum_{i=1}^N u_i(x_i), \text{s.t.} \sum_{i=1}^N A_i x_i = b,$$

which is a special case of (4.1) without y variable or the coupled function f.

At each iteration, the algorithm in [24] chooses one block of x-variable uniformly at random and performs a proximal update to it, followed by another proximal update to

the z-variable. More precisely, it iteratively performs the updates:

$$x_i^{k+1} = \begin{cases} \operatorname{argmin}_{x_i \in X_i} \langle -\bar{z}^k, A_i x_i \rangle + u_i(x_i) + \frac{\tau}{2}\|x_i - x_i^k\|_2^2, & \text{if } i = i_k, \\ x_i^k, & \text{if } i \neq i_k, \end{cases} \tag{4.9a}$$

$$z^{k+1} = \operatorname*{argmin}_{z \in Z} h(z) + \langle z, Ax^{k+1} \rangle + \frac{\eta}{2}\|z - z^k\|_2^2, \tag{4.9b}$$

$$\bar{z}^{k+1} = q(z^{k+1} - z^k) + z^{k+1}, \tag{4.9c}$$

where i_k is a randomly selected block, and τ, η and q are certain parameters[1]. When there is only one block of x-variable, i.e., $N = 1$, the scheme in (4.9) becomes exactly the primal-dual method in [15]. Assuming the boundedness of the constraint sets X and Z, [24] shows that under weak convexity, $O(1/t)$ convergence rate result of the scheme can be established by choosing appropriate parameters, and if u_i's are all strongly convex, the scheme can be accelerated to have $O(1/t^2)$ convergence rate by adapting the parameters.

4.1.3 Contributions and organization

- We propose a randomized primal-dual coordinate update algorithm to solve problems in the form of (4.1). The key feature is to introduce randomization as done in (4.9) to the multi-block ADMM framework (4.6). Unlike the random permutation scheme as previously investigated in [113, 18], we simply choose a subset of blocks of variables based on the uniform distribution. In addition, we perform a proximal update to that selected subset of variables. With appropriate proximal terms (e.g., the setting in (4.15)), the selected block variables can be decoupled, and thus the updates can be done in parallel.

- More general than (4.6), we can accommodate coupled terms in the objective function in our algorithm by linearizing such terms. By imposing Lipschitz continuity condition on the partial gradient of the coupled functions f and g and using proximal terms, we show that our method has an expected $O(1/t)$ convergence rate for solving problem (4.1) under mere convexity assumption.

- We show that our algorithm includes several existing methods as special cases

[1] Actually, [24] presents its algorithm in a more general way with the parameters adaptive to the iteration. However, its convergence result assumes constant values of these parameters for the weak convexity case.

such as the scheme in (4.9) and the proximal Jacobian ADMM in [27]. Our result indicates that the $O(1/t)$ convergence rate of the scheme in (4.9) can be shown without assuming boundedness of the constraint sets. In addition, the same order of convergence rate of the proximal Jacobian ADMM can be established in terms of a better measure.

- Furthermore, the linearization scheme allows us to deal with stochastic objective function, for instance, when the function f is given in a form of expectation $f = \mathbb{E}_\xi[f_\xi(x)]$ where ξ is a random vector. As long as an unbiased estimator of the (sub-)gradient of f is available, we can extend our method to the stochastic problem and an expected $O(1/\sqrt{t})$ convergence rate is achievable.

The rest of the chapter is organized as follows. In Section 4.2, we introduce our algorithm and present some preliminary results. In Section 4.3, we present the sublinear convergence rate results of the proposed algorithm. Depending on the multi-block structure of y, different conditions and parameter settings are presented in Subsections 4.3.1, 4.3.2 and 4.3.3, respectively. In Section 4.4, we present an extension of our algorithm where the objective function is assumed not to be even exactly computable, instead only some first-order stochastic approximation is available. The convergence analysis is extended to such settings accordingly. Numerical results are shown in Section 4.5. In Section 4.6, we discuss the connections of our algorithm to other well-known methods in the literature. The proofs for the technical lemmas are presented in Section 4.8, and the proofs for the main theorems are in Section 4.9.

4.2 Randomized Primal-Dual Block Coordinate Update Algorithm

In this section, we first present some notations and then introduce our algorithm as well as some preliminary lemmas.

4.2.1 Notations

We denote $X = X_1 \times \cdots \times X_N$ and $Y = Y_1 \times \cdots \times Y_M$. For any symmetric positive semidefinite matrix W, we define $\|z\|_W = \sqrt{z^\top W z}$. Given an integer $\ell > 0$, $[\ell]$ denotes the set $\{1, 2, \cdots, \ell\}$. We use I and J as index sets, while I is also used to denote

the identity matrix; we believe that the intention is evident in the context. Given $I = \{i_1, i_2, \cdots, i_n\}$, we denote:

- Block-indexed variable: $x_I = (x_{i_1}; x_{i_2}; \cdots; x_{i_n})$;
- Block-indexed set: $X_I = X_{i_1} \times \cdots \times X_{i_n}$;
- Block-indexed function: $u_I(x_I) = u_{i_1}(x_{i_1}) + u_{i_2}(x_{i_2}) + \cdots + u_{i_n}(x_{i_n})$;
- Block-indexed gradient: $\nabla_I f(x) = (\nabla_{i_1} f(x); \nabla_{i_2} f(x); \cdots; \nabla_{i_n} f(x))$;
- Block-indexed matrix: $A_I = \big[A_{i_1}, A_{i_2}, \cdots, A_{i_n}\big]$.

4.2.2 Algorithm

Our algorithm is rather general. Its major ingredients are randomization in selecting block variables, linearization of the coupled functions f and g, and adding proximal terms. Specifically, at each iteration k, it first randomly samples a subset I_k of blocks of x, and then a subset J_k of blocks of y according to the uniform distribution over the indices. The randomized sampling rule is as follows:

> **Randomization Rule (U):** For the given integers $n \leq N$ and $m \leq M$, it randomly chooses index sets $I_k \subset [N]$ with $|I_k| = n$ and $J_k \subset [M]$ with $|J_k| = m$ *uniformly*; i.e., for any subsets $\{i_1, i_2, \ldots, i_n\} \subset [N]$ and $\{j_1, j_2, \ldots, j_m\} \subset [M]$, the following holds
>
> $$\text{Prob}\big[I_k = \{i_1, i_2, \ldots, i_n\}\big] = 1/\begin{pmatrix} N \\ n \end{pmatrix},$$
> $$\text{Prob}\big[J_k = \{j_1, j_2, \ldots, j_m\}\big] = 1/\begin{pmatrix} M \\ m \end{pmatrix}.$$

After those subsets have been selected, it performs a prox-linear update to those selected blocks based on the augmented Lagrangian function, followed by an update of the Lagrangian multiplier. The details of the method are summarized in Algorithm 1 below.

In Algorithm 1, P^k and Q^k are predetermined positive semidefinite matrices with appropriate dimensions. For the selected blocks in I_k and J_k, instead of implementing the exact minimization of the augmented Lagrangian function, we perform a block proximal gradient update. In particular, before minimization, we first linearize the

Algorithm 1: **R**andomized **P**rimal-**D**ual **B**lock Coordinate **U**pdate Method (RPDBU)

Initialization: choose x^0, y^0 and $\lambda^0 = 0$; let $r^0 = Ax^0 + By^0 - b$; choose ρ, ρ_x, ρ_y

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

Randomly select $I_k \subset [N]$ and $J_k \subset [M]$ with $|I_k| = n$ and $|J_k| = m$ according to **(U)**.

Let $x_i^{k+1} = x_i^k$, $\forall i \notin I_k$ and $y_j^{k+1} = y_j^k$, $\forall j \notin J_k$.

For $I = I_k$, perform the update

$$x_I^{k+1} = \underset{x_I \in \mathcal{X}_I}{\operatorname{argmin}} \langle \nabla_I f(x^k) - A_I^\top \lambda^k, x_I \rangle + u_I(x_I) + \frac{\rho_x}{2} \|A_I(x_I - x_I^k) + r^k\|^2 + \frac{1}{2} \|x_I - x_I^k\|_{P^k}^2, \tag{4.10}$$

$$r^{k+\frac{1}{2}} = r^k + A_I(x_I^{k+1} - x_I^k). \tag{4.11}$$

For $J = J_k$, perform the update

$$y_J^{k+1} = \underset{y_J \in \mathcal{Y}_J}{\operatorname{argmin}} \langle \nabla_J g(y^k) - B_J^\top \lambda^k, y_J \rangle + v_J(y_J) + \frac{\rho_y}{2} \|B_J(y_J - y_J^k) + r^{k+\frac{1}{2}}\|^2 + \frac{1}{2} \|y_J - y_J^k\|_{Q^k}^2, \tag{4.12}$$

$$r^{k+1} = r^{k+\frac{1}{2}} + B_J(y_J^{k+1} - y_J^k). \tag{4.13}$$

Update the multiplier by

$$\lambda^{k+1} = \lambda^k - \rho r^{k+1}. \tag{4.14}$$

coupled functions f, g, and add some proximal terms to it. Note that one can always select all blocks, i.e., $I_k = [N]$ and $J_k = [M]$. Empirically however, the block coordinate update method usually outperforms the full coordinate update method if the problem possesses certain structures; see [95] for an example. To decouple the selected x blocks and also y blocks, we choose the matrices P^k and Q^k in Algorithm 1 as follows:

$$P^k = \hat{P}_{I_k} - \rho_x A_{I_k}^\top A_{I_k}, \qquad Q^k = \hat{Q}_{J_k} - \rho_y B_{J_k}^\top B_{J_k}, \tag{4.15}$$

where $\hat{P}$ and $\hat{Q}$ are symmetric positive semidefinite and block diagonal matrices, $\hat{P}_{I_k}$ denotes the diagonal blocks of $\hat{P}$ indexed by I_k. With such setting of P^k and Q^k, (4.10) and (4.12) respectively become

$$x_I^{k+1} = \operatorname*{argmin}_{x_I \in \mathcal{X}_I} \left\langle \nabla_I f(x^k) - A_I^\top(\lambda^k - \rho_x r^k), x_I \right\rangle + u_I(x_I) + \frac{1}{2}\|x_I - x_I^k\|_{\hat{P}_I}^2, \tag{4.16}$$

$$y_J^{k+1} = \operatorname*{argmin}_{y_J \in \mathcal{Y}_J} \left\langle \nabla_J g(y^k) - B_J^\top(\lambda^k - \rho_y r^{k+\frac{1}{2}}), y_J \right\rangle + v_J(y_J) + \frac{1}{2}\|y_J - y_J^k\|_{\hat{Q}_J}^2. \tag{4.17}$$

Due to the block diagonal structure of $\hat{P}$ and $\hat{Q}$, both x and y-updates can be computed in parallel.

4.2.3 Preliminaries

Let w be the aggregated primal-dual variables and $H(w)$ the primal-dual linear mapping; namely

$$w = \begin{pmatrix} x \\ y \\ \lambda \end{pmatrix}, \quad H(w) = \begin{pmatrix} -A^\top\lambda \\ -B^\top\lambda \\ Ax + By - b \end{pmatrix}, \tag{4.18}$$

and also let

$$u(x) = \sum_{i=1}^N u_i(x_i), \quad v(y) = \sum_{j=1}^M v_j(y_j),$$
$$F(x) = f(x) + u(x), \quad G(y) = g(y) + v(y), \quad \Phi(x,y) = F(x) + G(y).$$

The point (x^*, y^*) is a solution to (4.1) *if and only if* there exists λ^* such that

$$\Phi(x,y) - \Phi(x^*,y^*) + (w-w^*)^\top H(w^*) \geq 0,\ \forall (x,y) \in \mathcal{X}\times\mathcal{Y},\ \forall \lambda, \tag{4.19a}$$

$$Ax^* + By^* = b \tag{4.19b}$$

$$x^* \in \mathcal{X}, \quad y^* \in \mathcal{Y}. \tag{4.19c}$$

The following lemmas will be used in our subsequent analysis, whose proofs are elementary and thus are omitted here.

Lemma 4.2.1 *For any two vectors w and $\tilde{w}$, it holds*

$$(w-\tilde{w})^\top H(w) = (w-\tilde{w})^\top H(\tilde{w}). \tag{4.20}$$

Lemma 4.2.2 *For any two vectors u, v and a positive semidefinite matrix W:*

$$u^\top W v = \frac{1}{2}\big(\|u\|_W^2 + \|v\|_W^2 - \|u-v\|_W^2\big). \tag{4.21}$$

Lemma 4.2.3 *For any nonzero positive semidefinite matrix W, it holds for any $\mathbf{z}$ and $\hat{\mathbf{z}}$ of appropriate size that*

$$\|\mathbf{z}-\hat{\mathbf{z}}\|^2 \geq \frac{1}{\|W\|_2}\|\mathbf{z}-\hat{\mathbf{z}}\|_W^2, \tag{4.22}$$

where $\|W\|_2$ denotes the matrix operator norm of W.

The following lemma presents a useful property of $H(w)$, which essentially follows from (4.20).

Lemma 4.2.4 *For any vectors $\{w^k\}_1^t$, and sequence of positive numbers $\{\beta^k\}_1^t$, it holds that*

$$\left(\frac{\sum_{k=0}^t \beta^k w^k}{\sum_{k=0}^t \beta^k} - w\right)^\top H\left(\frac{\sum_{k=0}^t \beta^k w^k}{\sum_{k=0}^t \beta^k}\right) = \frac{1}{\sum_{k=0}^t \beta^k}\sum_{k=0}^t \beta^k (w^k - w)^\top H(w^k). \tag{4.23}$$

4.3 Convergence Rate Results

In this section, we establish sublinear convergence rate results of Algorithm 1 for three different cases. We differentiate those cases based on whether or not y in problem (4.1) also has the multi-block structure. In the first case where y is a multi-block variable, it requires $\frac{n}{N} = \frac{m}{M}$ where n and m are the cardinalities of the subsets of x and y selected in our algorithm respectively. Since the analysis only requires weak convexity, we can ensure the condition to hold by adding *zero* component functions if necessary, in such a way that $N = M$ and then choosing $n = m$. The second case is that y is treated as a single-block variable, and this can be reflected in our algorithm by simply selecting all y-blocks every time, i.e. $m = M$. The third case assumes no y-variable at all. It falls into the first and second cases, and we discuss this case separately since it requires weaker conditions to guarantee the same convergence rate. In particular, we make the following assumptions:

Assumption 4.3.1 (Convexity) *For* (4.1)*, $\mathcal{X}_i$'s and $\mathcal{Y}_j$'s are some closed convex sets, f and g are smooth convex functions, and u_i's and v_j's are proper closed convex function.*

Assumption 4.3.2 (Existence of an optimal solution) *There is at least one point $w^* = (x^*, y^*, \lambda^*)$ satisfying the conditions in* (4.19)*.*

Assumption 4.3.3 (Lipschitz continuous partial gradient) *There exist constants L_f and L_g such that for any subset I of $[N]$ with $|I| = n$ and any subset J of $[M]$ with $|J| = m$, it holds that*

$$\|\nabla_I f(x + U_I\tilde{x}) - \nabla_I f(x)\| \leq L_f\|\tilde{x}_I\|, \forall x, \tilde{x}, \tag{4.24a}$$

$$\|\nabla_J g(y + U_J\tilde{y}) - \nabla_J g(y)\| \leq L_g\|\tilde{y}_J\|, \forall y, \tilde{y}, \tag{4.24b}$$

where $U_I\tilde{x}$ keeps the blocks of $\tilde{x}$ that are indexed by I and zero elsewhere.

Before presenting the main convergence rate result, we first establish a few key lemmas.

Lemma 4.3.1 (One-step analysis) *Let $\{(x^k, y^k, r^k, \lambda^k)\}$ be the sequence generated from Algorithm 1 with matrices P^k and Q^k defined as in* (4.15)*. Then the following inequalities hold*

$$\mathbb{E}_{I_k}\Big[F(x^{k+1}) - F(x) + (x^{k+1} - x)^\top(-A^\top\lambda^{k+1})$$

$$
\begin{aligned}
&\quad + (\rho_x - \rho)(x^{k+1} - x)^\top A^\top r^{k+1} - \rho_x (x^{k+1} - x)^\top A^\top B(y^{k+1} - y^k)\Big] \\
&\quad + \mathbb{E}_{I_k}(x^{k+1} - x)^\top (\hat{P} - \rho_x A^\top A)(x^{k+1} - x^k) - \frac{L_f}{2}\mathbb{E}_{I_k}\|x^k - x^{k+1}\|^2 \\
&\leq \left(1 - \frac{n}{N}\right)\left[F(x^k) - F(x) + (x^k - x)^\top(-A^\top \lambda^k) + \rho_x (x^k - x)^\top A^\top r^k\right], \quad (4.25)
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}_{J_k}\left[G(y^{k+1}) - G(y) + (y^{k+1} - y)^\top(-B^\top \lambda^{k+1}) + (\rho_y - \rho)(y^{k+1} - y)^\top B^\top r^{k+1}\right] \\
&+ \mathbb{E}_{J_k}(y^{k+1} - y)^\top(\hat{Q} - \rho_y B^\top B)(y^{k+1} - y^k) - \frac{L_g}{2}\mathbb{E}_{J_k}\|y^k - y^{k+1}\|^2 \\
&- \left(1 - \frac{m}{M}\right)\rho_y (y^k - y)^\top B^\top A(x^{k+1} - x^k) \\
\leq\ &\left(1 - \frac{m}{M}\right)\left[G(y^k) - G(y) + (y^k - y)^\top(-B^\top \lambda^k) + \rho_y (y^k - y)^\top B^\top r^k\right], \qquad (4.26)
\end{aligned}
$$

where $\mathbb{E}_{I_k}$ denotes expectation over I_k and conditional on all previous history.

Note that for any feasible point (x, y) (namely, $x \in X, y \in Y$ and $Ax + By = b$),

$$
\begin{aligned}
Ax^{k+1} - Ax &= \frac{1}{\rho}(\lambda^k - \lambda^{k+1}) - (By^{k+1} - b) + (By - b) \\
&= \frac{1}{\rho}(\lambda^k - \lambda^{k+1}) - B(y^{k+1} - y) \qquad (4.27)
\end{aligned}
$$

and

$$
\begin{aligned}
By^k - By &= \frac{1}{\rho}(\lambda^{k-1} - \lambda^k) - (Ax^k - b) + (Ax - b) \\
&= \frac{1}{\rho}(\lambda^{k-1} - \lambda^k) - A(x^k - x). \qquad (4.28)
\end{aligned}
$$

Then, using (4.21) we have the following result.

Lemma 4.3.2 *For any feasible point (x, y) and integer t, it holds*

$$
\begin{aligned}
&\sum_{k=0}^{t}(x^{k+1} - x)^\top A^\top B(y^{k+1} - y^k) \\
=\ &\frac{1}{\rho}\sum_{k=0}^{t}(\lambda^k - \lambda^{k+1})^\top B(y^{k+1} - y^k) \\
&-\frac{1}{2}\left(\|y^{t+1} - y\|^2_{B^\top B} - \|y^0 - y\|^2_{B^\top B} + \sum_{k=0}^{t}\|y^{k+1} - y^k\|^2_{B^\top B}\right) \qquad (4.29)
\end{aligned}
$$

and

$$\begin{aligned}
&\sum_{k=0}^{t}(y^k-y)^\top B^\top A(x^{k+1}-x^k)\\
=\;&\frac{1}{\rho}\sum_{k=0}^{t}(\lambda^{k-1}-\lambda^k)^\top A(x^{k+1}-x^k)\\
&+\frac{1}{2}\left(\|x^0-x\|^2_{A^\top A}-\|x^{t+1}-x\|^2_{A^\top A}+\sum_{k=0}^{t}\|x^{k+1}-x^k\|^2_{A^\top A}\right). \qquad (4.30)
\end{aligned}$$

Lemma 4.3.3 *Given a continuous function h, for a random vector $\hat{w}=(\hat{x},\hat{y},\hat{\lambda})$, if for any feasible point $w=(x,y,\lambda)$ that may depend on $\hat{w}$, we have*

$$\mathbb{E}\big[\Phi(\hat{x},\hat{y})-\Phi(x,y)+(\hat{w}-w)^\top H(w)\big]\le\mathbb{E}[h(w)], \qquad (4.31)$$

then for any $\gamma>0$ and any optimal solution (x^*,y^*) to (4.1) we also have

$$\mathbb{E}\big[\Phi(\hat{x},\hat{y})-\Phi(x^*,y^*)+\gamma\|A\hat{x}+B\hat{y}-b\|\big]\le\sup_{\|\lambda\|\le\gamma}h(x^*,y^*,\lambda).$$

Noting

$$\Phi(x,y)-\Phi(x^*,y^*)+(w-w^*)^\top H(w^*)=\Phi(x,y)-\Phi(x^*,y^*)-(\lambda^*)^\top(Ax+By-b),$$

we can easily show the following lemma by the optimality of (x^*,y^*,λ^*) and the Cauchy-Schwarz inequality.

Lemma 4.3.4 *Assume (x^*,y^*,λ^*) satisfies (4.19). Then for any point $(\hat{x},\hat{y})\in X\times Y$, we have*

$$\Phi(\hat{x},\hat{y})-\Phi(x^*,y^*)\ge-\|\lambda^*\|\cdot\|A\hat{x}+B\hat{y}-b\|. \qquad (4.32)$$

The following lemma shows a connection between different convergence measures, and it can be simply proved by using (4.32). If both w and $\hat{w}$ are deterministic, it implies Lemma 2.4 in [38].

Lemma 4.3.5 *Assume that (x^*,y^*,λ^*) satisfies the optimality conditions in (4.19). Let γ be any number that is larger than $\|\lambda^*\|$. If a random vector $(\hat{x},\hat{y})$ satisfies*

$$\mathbb{E}\big[\Phi(\hat{x},\hat{y})-\Phi(x^*,y^*)+\gamma\|A\hat{x}+B\hat{y}-b\|\big]\le\epsilon,$$

then

$$\mathbb{E}\|A\hat{x} + B\hat{y} - b\| \leq \frac{\epsilon}{\gamma - \|\lambda^*\|} \text{ and } \mathbb{E}\big[\big|\Phi(\hat{x},\hat{y}) - \Phi(x^*,y^*)\big|\big] \leq \left(\frac{2\|\lambda^*\|}{\gamma - \|\lambda^*\|} + 1\right)\epsilon.$$

The convergence analysis for Algorithm 1 requires slightly different parameter settings under different structures. In fact, the underlying analysis and results also differ. To account for the differences, we present in the next three subsections the corresponding convergence results. The first one assumes there is no y part at all; the second case assumes a single block on the y side; the last one deals with the general case where the ratios n/N is assumed to be equal to m/M.

4.3.1 Multiple x blocks and no y variable

We first consider a special case with no y-variable, namely, $g = v = 0$ and $B = 0$ in (4.1). This case has its own importance. It is a parallel block coordinate update version of the linearized augmented Lagrangian method (ALM).

Theorem 4.3.6 (Sublinear ergodic convergence I) *Assume $g(y) = 0, v_j(y_j) = 0$, $\forall j$ and $B = 0$ in (4.1). Let $\{(x^k, y^k, \lambda^k)\}$ be the sequence generated from Algorithm 1 with $y^k \equiv y^0$. Assume $\frac{n}{N} = \theta$, $\rho = \theta\rho_x$, and*

$$\hat{P}_{I_k} \succeq L_f I + \rho_x A_{I_k}^\top A_{I_k}, \forall k. \tag{4.33}$$

Let

$$\hat{x}^t = \frac{x^{t+1} + \theta\sum_{k=1}^t x^k}{1+\theta t}. \tag{4.34}$$

Then, under Assumptions 4.3.1, 4.3.2 and 4.3.3, we have

$$\begin{aligned} & \max\left\{\mathbb{E}\left|F(\hat{x}^t) - F(x^*)\right|, \mathbb{E}\|A\hat{x}^t - b\|\right\} \\ \leq\ & \frac{2}{1+\theta t}\left[(1-\theta)\left(F(x^0) - F(x^*) + \frac{\rho_x}{2}\|r^0\|^2\right)\right. \\ & \left. + \frac{1}{2}\|x^0 - x^*\|_{\hat{P}}^2 + \frac{\max\{(0.5 + \|\lambda^*\|)^2, 9\|\lambda^*\|^2\}}{2\rho_x}\right] \end{aligned} \tag{4.35}$$

where (x^*, λ^*) is an arbitrary primal-dual solution.

Our result recovers the convergence of the proximal Jacobian ADMM introduced in [27]. In fact, the above theorem strengthens the convergence result in [27] by establishing

an $O(1/t)$ rate of convergence in terms of the feasibility measure and the objective value. If strong convexity is assumed on the objective function, the algorithm can be accelerated to have the rate $O(1/t^2)$ as shown in [127].

4.3.2 Multiple x blocks and a single y block

When the y-variable is simple to update, it could be beneficial to renew the whole of it at every iteration, such as the problem (4.3). In this subsection, we consider the case that there are multiple x-blocks but a single y-block (or equivalently, $m = M$), and we establish a sublinear convergence rate result with a different technique of dealing with the y-variable.

Theorem 4.3.7 (Sublinear ergodic convergence II) *Let $\{(x^k, y^k, \lambda^k)\}$ be the sequence generated from Algorithm 1 with $m = M$ and $\rho = \rho_y = \theta\rho_x$, where $\theta = \frac{n}{N}$. Assume*

$$\hat{P} \succeq L_f I + \rho_x A^\top A, \qquad \hat{Q} \succeq \frac{L_g}{\theta} I + \left(\frac{\rho}{\theta^4} - \frac{\rho}{\theta^2} + \rho_y\right) B^\top B. \tag{4.36}$$

Let

$$\hat{x}^t = \frac{x^{t+1} + \theta \sum_{k=1}^t x^k}{1+\theta t}, \quad \hat{y}^t = \frac{\tilde{y}^{t+1} + \theta \sum_{k=1}^t y^k}{1+\theta t} \tag{4.37}$$

where

$$\tilde{y}^{t+1} = \operatorname*{argmin}_{y\in\mathcal{Y}} \langle \nabla g(y^t) - B^\top \lambda^t, y\rangle + v(y) + \frac{\rho_x}{2}\|Ax^{t+1} + By - b\|^2 + \frac{\theta}{2}\|y - y^t\|^2_{\hat{Q}-\rho_y B^\top B}. \tag{4.38}$$

Then, under Assumptions 4.3.1, 4.3.2 and 4.3.3, we have

$$\begin{aligned} & \max\left\{\mathbb{E}\left|\Phi(\hat{x}^t, \hat{y}^t) - \Phi(x^*, y^*)\right|, \mathbb{E}\|A\hat{x}^t + B\hat{y}^t - b\|\right\} \qquad (4.39)\\ \leq\ & \frac{2}{1+\theta t}\left[(1-\theta)\left(\Phi(x^0, y^0) - \Phi(x^*, y^*) + \frac{\rho_x}{2}\|r^0\|^2\right) + \frac{1}{2}\|x^0 - x^*\|^2_{\hat{P}}\right.\\ & \left. + \frac{1}{2}\|y^0 - y^*\|^2_{\theta\hat{Q}+(\rho_x-\theta\rho_y)B^\top B} + \frac{\max\{(0.5 + \|\lambda^*\|)^2, 9\|\lambda^*\|^2\}}{2\rho_x}\right] \end{aligned}$$

where (x^*, y^*, λ^*) is an arbitrary primal-dual solution.

Remark 4.3.8 *It is easy to see that if $\theta = 1$, the result in Theorem 4.3.7 becomes exactly the same as that in Theorem 4.3.9 below. In general, they are different because the conditions in* (4.36) *on $\hat{P}$ and $\hat{Q}$ are different from those in* (4.41a).

4.3.3 Multiple x and y blocks

In this subsection, we consider the most general case where both x and y have multi-block structure. Assuming $\frac{n}{N} = \frac{m}{M}$, we can still have the $O(1/t)$ convergence rate. The assumption can be made without losing generality, e.g., by adding zero components if necessary (which is essentially equivalent to varying the probabilities of the variable selection).

Theorem 4.3.9 (Sublinear ergodic convergence III) *Let $\{(x^k, y^k, \lambda^k)\}$ be the sequence generated from Algorithm 1 with the parameters satisfying*

$$\rho = \frac{n\rho_x}{N} = \frac{m\rho_y}{M} > 0. \tag{4.40}$$

Assume $\frac{n}{N} = \frac{m}{M} = \theta$, and $\hat{P}, \hat{Q}$ satisfy one of the following conditions

$$\hat{P} \succeq (2-\theta)\left(\frac{1-\theta}{\theta^2}+1\right)\rho_x A^\top A + L_f I, \qquad \hat{Q} \succeq \frac{(2-\theta)}{\theta^2}\rho_y B^\top B + L_g I. \tag{4.41a}$$

$$\hat{P}_i \succeq (2-\theta)\left(\frac{1-\theta}{\theta^2}+1\right) n\rho_x A_i^\top A_i + L_f I,\ \forall i, \quad \hat{Q}_j \succeq \frac{(2-\theta)}{\theta^2} m\rho_y B_j^\top B_j + L_g I,\ \forall j. \tag{4.41b}$$

Let

$$\hat{x}^t = \frac{x^{t+1} + \theta\sum_{k=1}^t x^k}{1+\theta t}, \quad \hat{y}^t = \frac{y^{t+1} + \theta\sum_{k=1}^t y^k}{1+\theta t}. \tag{4.42}$$

Then, under Assumptions 4.3.1, 4.3.2 and 4.3.3, we have

$$\begin{aligned} & \max\left\{\mathbb{E}\left|\Phi(\hat{x}^t, \hat{y}^t) - \Phi(x^*, y^*)\right|, \mathbb{E}\|A\hat{x}^t + B\hat{y}^t - b\|\right\} \qquad (4.43)\\ \leq\ & \frac{2}{1+\theta t}\left[(1-\theta)\left(\Phi(x^0, y^0) - \Phi(x^*, y^*) + \rho_x\|r^0\|^2\right) + \frac{1}{2}\|x^0 - x^*\|^2_{\hat{P}-\theta\rho_x A^\top A}\right.\\ & \left. + \frac{1}{2}\|y^0 - y^*\|^2_{\hat{Q}} + \frac{\max\{(0.5+\|\lambda^*\|)^2, 9\|\lambda^*\|^2\}}{2\rho_x}\right] \end{aligned}$$

where (x^*, y^*, λ^*) is an arbitrary primal-dual solution.

Remark 4.3.10 *When $N = M = 1$, the two conditions in (4.41) become the same. However, in general, neither of the two conditions in (4.41) implies the other one. Roughly speaking, for the case of $n \approx N$ and $m \approx M$, the one in (4.41a) can be weaker, and for the case of $n \ll N$ and $m \ll M$, the one in (4.41b) is more likely weaker. In*

addition, (4.41b) *provides an explicit way to choose block diagonal* $\hat{P}$ *and* $\hat{Q}$ *by simply setting* $\hat{P}_i$ *and* $\hat{Q}_j$*'s to the lower bounds there.*

4.4 Randomized Primal-Dual Coordinate Approach for Stochastic Programming

In this section, we extend our method to solve a stochastic optimization problem where the objective function involves an expectation. Specifically, we assume the coupled function to be in the form of $f(x) = \mathbb{E}_\xi f_\xi(x)$ where ξ is a random vector. For simplicity we assume $g = v = 0$, namely, we consider the following problem

$$\begin{aligned} \min_x \ & \mathbb{E}_\xi f_\xi(x) + \sum_{i=1}^N u_i(x_i), \\ \text{s.t.} & \sum_{i=1}^N A_i x_i = b, \ x_i \in \mathcal{X}_i, \ i = 1, 2, ..., N. \end{aligned} \tag{4.44}$$

One can easily extend our analysis to the case where $g \neq 0, v \neq 0$ and g is also stochastic. An example of (4.44) is the penalized and constrained regression problem [62] that includes (4.2) as a special case.

Due to the expectation form of f, it is natural that the exact gradient of f is not available or very expensive to compute. Instead, we assume that its stochastic gradient is readily accessible. By some slight abuse of the notation, we denote

$$w = \begin{bmatrix} x \\ \lambda \end{bmatrix}, \quad H(w) = \begin{bmatrix} -A^\top \lambda \\ Ax - b \end{bmatrix}, \quad F(x) := \mathbb{E}_\xi f_\xi(x) + \sum_{i=1}^N u_i(x_i). \tag{4.45}$$

A point x^* is a solution to (4.44) *if and only if* there exists λ^* such that

$$F(x) - F(x^*) + (w - w^*)^\top H(w^*) \geq 0, \ \forall w, \tag{4.46a}$$

$$Ax^* = b, \ x^* \in \mathcal{X}. \tag{4.46b}$$

Modifying Algorithm 1 to (4.44), we present the stochastic primal-dual coordinate update method of multipliers, summarized in Algorithm 2, where G^k is a stochastic approximation of $\nabla f(x^k)$. The strategy of block coordinate update with stochastic gradient information was first proposed in [25, 126], which considered problems without

linear constraint.

Algorithm 2: **R**andomized **P**rimal-**D**ual **B**lock Coordinate **U**pdate Method for **S**tochastic Programming (RPDBUS)

Initialization: choose x^0, λ^0 and set parameters ρ, α_k's

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

Randomly select $I_k \subset [N]$ with $|I_k| = n$ according to **(U)**.

Let $x_i^{k+1} = x_i^k$, $\forall i \notin I_k$, and for $I = I_k$, do the update

$$x_I^{k+1} = \operatorname*{argmin}_{x_I \in \mathcal{X}_I} \langle G_I^k - A_I^\top \lambda^k, x_I \rangle + u_I(x_I) + \frac{\rho}{2}\|A_I(x_I - x_I^k) + r^k\|^2 + \frac{1}{2}\|x_I - x_I^k\|^2_{P^k + \frac{I}{\alpha_k}}. \tag{4.47}$$

Update the residual $r^{k+1} = r^k + A_I(x_I^{k+1} - x_I^k)$.

Update the multiplier by

$$\lambda^{k+1} = \lambda^k - \left(1 - \frac{(N-n)\alpha_{k+1}}{N\alpha_k}\right)\rho r^{k+1}. \tag{4.48}$$

We make the following assumption on the stochastic gradient G^k.

Assumption 4.4.1 *Let $\delta^k = G^k - \nabla f(x^k)$. There exists a constant σ such that for all k,*

$$\mathbb{E}[\delta^k \,|\, x^k] = \mathbf{0}, \tag{4.49a}$$

$$\mathbb{E}\|\delta^k\|^2 \le \sigma^2. \tag{4.49b}$$

Following the proof of Lemma 4.3.1 and also noting

$$\mathbb{E}_{I_k}\left[(x_{I_k} - x_{I_k}^{k+1})^\top \delta_{I_k}^k \,|\, x^k\right] = \mathbb{E}_{I_k}(x^k - x^{k+1})^\top \delta^k, \tag{4.50}$$

we immediately have the following result.

Lemma 4.4.1 (One-step analysis) *Let $\{(x^k, r^k, \lambda^k)\}$ be the sequence generated from Algorithm 2 where P^k is given in* (4.15) *with $\rho_x = \rho$. Then*

$$\begin{aligned}
&\mathbb{E}_{I_k}\left[F(x^{k+1}) - F(x) + (x^{k+1} - x)^\top(-A^\top\lambda^k) + \rho(x^{k+1} - x)^\top A^\top r^{k+1}\right] \\
&+\mathbb{E}_{I_k}(x^{k+1} - x)^\top \left(\hat{P} - \rho A^\top A + \frac{I}{\alpha_k}\right)(x^{k+1} - x^k) \\
&-\frac{L_f}{2}\mathbb{E}_{I_k}\|x^k - x^{k+1}\|^2 + \mathbb{E}_{I_k}(x^{k+1} - x^k)^\top \delta^k
\end{aligned}$$

$$\leq \quad \left(1-\frac{n}{N}\right)\left[F(x^k)-F(x)+(x^k-x)^\top(-A^\top\lambda^k)+\rho(x^k-x)^\top A^\top r^k\right]. \quad (4.51)$$

The following theorem is a key result, from which we can choose appropriate α_k to obtain the $O(1/\sqrt{t})$ convergence rate.

Theorem 4.4.2 *Let $\{(x^k,\lambda^k)\}$ be the sequence generated from Algorithm 2. Let $\theta=\frac{n}{N}$ and denote*

$$\beta_k=\frac{\alpha_k}{\left(1-\frac{\alpha_k(1-\theta)}{\alpha_{k-1}}\right)\rho},\forall k.$$

Assume $\alpha_k>0$ is nonincreasing, and

$$Ax^0=b,\quad \lambda^0=0, \quad (4.52a)$$

$$\hat{P}\succeq L_f I+\rho A^\top A, \quad (4.52b)$$

$$\frac{\alpha_{k-1}\beta_k}{2\alpha_k}+\frac{(1-\theta)\beta_{k+1}}{2}-\frac{\alpha_k\beta_{k+1}}{2\alpha_{k+1}}-\frac{(1-\theta)\beta_k}{2}\geq 0,\ \forall k \quad (4.52c)$$

$$\frac{\alpha_t}{2\rho}\geq\left|\frac{\alpha_{t-1}\beta_t}{\alpha_t}-(1-\theta)\beta_t-\frac{\alpha_t}{\rho}\right|,\ \textit{for some } t. \quad (4.52d)$$

Let

$$\hat{x}^t=\frac{\alpha_{t+1}x^{t+1}+\theta\sum_{k=1}^t\alpha_k x^k}{\alpha_{t+1}+\theta\sum\limits_{k=1}^t\alpha_k}. \quad (4.53)$$

Then, under Assumptions 4.3.1, 4.3.2, 4.3.3 and 4.4.1, we have

$$\begin{aligned}&(\alpha_{t+1}+\theta\sum_{k=1}^t\alpha_k)\mathbb{E}\left[F(\hat{x}^t)-F(x^*)+\gamma\|A\hat{x}^t-b\|\right]\\ \leq\ &(1-\theta)\alpha_0\left[F(x^0)-F(x^*)\right]+\frac{\alpha_0}{2}\|x^0-x^*\|^2_{\hat{P}-\rho A^\top A}+\frac{1}{2}\|x^0-x^*\|^2\\ &+\left|\frac{\alpha_0\beta_1}{2\alpha_1}-\frac{(1-\theta)\beta_1}{2}\right|\gamma^2+\sum_{k=0}^t\frac{\alpha_k^2}{2}\mathbb{E}\|\delta^k\|^2. \quad (4.54)\end{aligned}$$

The following proposition gives sublinear convergence rate of Algorithm 2 by specifying the values of its parameters. The choice of α_k depends on whether we fix the total number of iterations.

Proposition 4.4.3 *Let $\{(x^k,\lambda^k)\}$ be the sequence generated from Algorithm 2 with P^k given in* (4.15)*, $\hat{P}$ satisfying* (4.52b)*, and the initial point satisfying $Ax^0=b$ and $\lambda^0=\mathbf{0}$.*

Let C_0 be

$$C_0 = (1-\theta)\alpha_0\big[F(x^0) - F(x^*)\big] + \frac{1}{2}\|x^0 - x^*\|_{D_x}^2 + \frac{\alpha_0}{2\rho}\max\{(0.5 + \|\lambda^*\|)^2, 9\|\lambda^*\|^2\}, \tag{4.55}$$

where (x^*, λ^*) is a primal-dual solution, and $D_x := \alpha_0(\hat{P} - \rho A^\top A) + I$.

1. *If $\alpha_k = \frac{\alpha_0}{\sqrt{k}}, \forall k \geq 1$ for a certain $\alpha_0 > 0$, then for $t \geq 2$,*

$$\max\left\{\mathbb{E}\left|F(\hat{x}^t) - F(x^*)\right|, \mathbb{E}\|A\hat{x}^t - b\|\right\} \leq \frac{2C_0}{\theta\alpha_0\sqrt{t}} + \frac{\alpha_0(\log t + 2)\sigma^2}{\theta\sqrt{t}}. \tag{4.56}$$

2. *If the number of maximum number of iteration is fixed a priori, then by choosing $\alpha_k = \frac{\alpha_0}{\sqrt{t}}$, $\forall k \geq 1$ with any given $\alpha_0 > 0$, we have*

$$\max\left\{\mathbb{E}\left|F(\hat{x}^t) - F(x^*)\right|, \mathbb{E}\|A\hat{x}^t - b\|\right\} \leq \frac{2C_0}{\theta\alpha_0\sqrt{t}} + \frac{2\alpha_0\sigma^2}{\theta\sqrt{t}}. \tag{4.57}$$

Proof. When $\alpha_k = \frac{\alpha_0}{\sqrt{k}}$, we can show that (4.52c) and (4.52d) hold for $t \geq 2$; see Appendix 4.8.4. Hence, the result in (4.56) follows from (4.54), the convexity of F, Lemma 4.3.5 with $\gamma = \max\{1 + \|\lambda^*\|, 2\|\lambda^*\|\}$, and the inequalities

$$\sum_{k=1}^{t} \frac{1}{\sqrt{k}} \geq \sqrt{t}, \quad \sum_{k=1}^{t} \frac{1}{k} \leq \log t + 1.$$

When α_k is a constant, the terms on the left hand side of (4.52c) and on the right hand side of (4.52d) are both zero, so they are satisfied. Hence, the result in (4.57) immediately follows by noting $\sum_{k=1}^{t} \alpha_k = \alpha_0\sqrt{t}$ and $\sum_{k=0}^{t} \alpha_k^2 \leq 2\alpha_0^2$. □

The sublinear convergence result of Algorithm 2 can also be shown if f is nondifferentiable convex and Lipschitz continuous. Indeed, if f is Lipschtiz continuous with constant L_c, i.e.,

$$\|f(x) - f(y)\| \leq L_c\|x - y\|, \forall x, y,$$

then $\|\tilde{\nabla} f(x)\| \leq L_c$, $\forall x$, where $\tilde{\nabla} f(x)$ is a subgradient of f at x. Hence,

$$\begin{aligned}
& \mathbb{E}_{I_k}(x_{I_k} - x_{I_k}^{k+1})^\top \tilde{\nabla}_{I_k} f(x^k) \\
= \; & \mathbb{E}_{I_k}(x_{I_k} - x_{I_k}^{k})^\top \tilde{\nabla}_{I_k} f(x^k) + \mathbb{E}_{I_k}(x_{I_k}^{k} - x_{I_k}^{k+1})^\top \tilde{\nabla}_{I_k} f(x^k) \\
= \; & \frac{n}{N}(x - x^k)^\top \tilde{\nabla} f(x^k) + \mathbb{E}_{I_k}(x^k - x^{k+1})^\top \tilde{\nabla} f(x^{k+1}) \\
& + \mathbb{E}_{I_k}(x^k - x^{k+1})^\top \big(\tilde{\nabla} f(x^k) - \tilde{\nabla} f(x^{k+1})\big) \\
\leq \; & \frac{n}{N}(f(x) - f(x^k)) + \mathbb{E}_{I_k}[f(x^k) - f(x^{k+1})] \\
& + \mathbb{E}_{I_k}(x^k - x^{k+1})^\top \big(\tilde{\nabla} f(x^k) - \tilde{\nabla} f(x^{k+1})\big) \\
= \; & \frac{n-N}{N}(f(x) - f(x^k)) + \mathbb{E}_{I_k}[f(x) - f(x^{k+1})] \\
& + \mathbb{E}_{I_k}(x^k - x^{k+1})^\top \big(\tilde{\nabla} f(x^k) - \tilde{\nabla} f(x^{k+1})\big).
\end{aligned}$$

Now following the proof of Lemma 4.3.1, we can have a result similar to (4.51), and then through the same arguments as those in the proof of Theorem 4.4.2, we can establish sublinear convergence rate of $O(1/\sqrt{t})$.

4.5 Numerical Experiments

In this section, we test the proposed randomized primal-dual method on solving the nonnegativity constrained quadratic programming (NCQP):

$$\min_{x \in \mathbb{R}^n} F(x) \equiv \frac{1}{2} x^\top Q x + c^\top x, \text{s.t.} Ax = b, x_i \geq 0, i = 1, \ldots, n, \tag{4.58}$$

where $A \in \mathbb{R}^{m \times n}$, and $Q \in \mathbb{R}^{n \times n}$ is a symmetric positive semidefinite (PSD) matrix. There is no y-variable, and it falls into the case in Theorem 4.3.6. We perform two experiments on a Macbook Pro with 4 cores. The first experiment demonstrates the parallelization performance of the proposed method, and the second one compares it to other methods.

Parallelization. This test is to illustrate the power unleashed in our new method, which is flexible in terms of parallel and distributive computing. We set $m = 200, n = 2000$ and generate $Q = HH^\top$, where the components of $H \in \mathbb{R}^{n \times n}$ follow the standard Gaussian distribution. The matrix A and vectors b, c are also randomly generated. We treat every component of x as one block, and at every iteration we select and update p blocks, where p is the number of used cores. Figure 4.1 shows the running time by using 1, 2, and 4 cores, where the optimal value $F(x^*)$ is obtained by calling Matlab

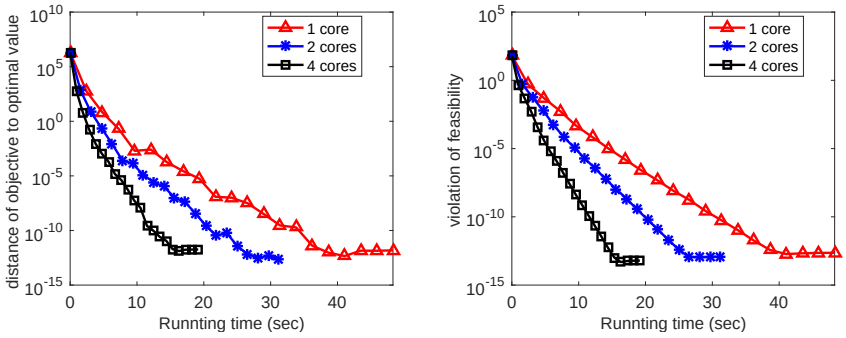

Figure 4.1: Nearly linear speed-up performance of the proposed primal-dual method for solving (4.58) on a 4-core machine. Left: distance of objective to optimal value $|F(x^k) - F(x^*)|$; Right: violation of feasibility $\|Ax^k - b\|$.

function `quadprog` with tolerance 10^{-16}. From the figure, we see that our proposed method achieves nearly linear speed-up.

Comparison to other methods. In this experiment, we compare the proposed method to the linearized ALM and the cyclic linearized ADMM methods. We set $m = 1000, n = 5000$ and generate $Q = HH^\top$, where the components of $H \in \mathbb{R}^{n\times(n-50)}$ follow standard Gaussian distribution. Note that Q is singular, and thus (4.58) is not strongly convex. We partition the variable into 100 blocks, each with 50 components. At each iteration of our method, we randomly select one block variable to update. Figure 4.2 shows the performance by the three compared methods, where one epoch is equivalent to updating 100 blocks once. From the figure, we see that our proposed method is comparable to the cyclic linearized ADMM and significantly better than the linearized ALM. Although the cyclic ADMM performs well on this example, in general it can diverge if the problem has more than two blocks; see [16].

4.6 Connections to Existing Methods

In this section, we discuss how Algorithms 1 and 2 are related to several existing methods in the literature, and we also compare their convergence results. It turns out that the proposed algorithms specialize to several known methods or their variants in the literature under various specific conditions. Therefore, our convergence analysis recovers some existing results as special cases, as well as provides new convergence results for certain existing algorithms such as the Jacobian proximal parallel ADMM and the

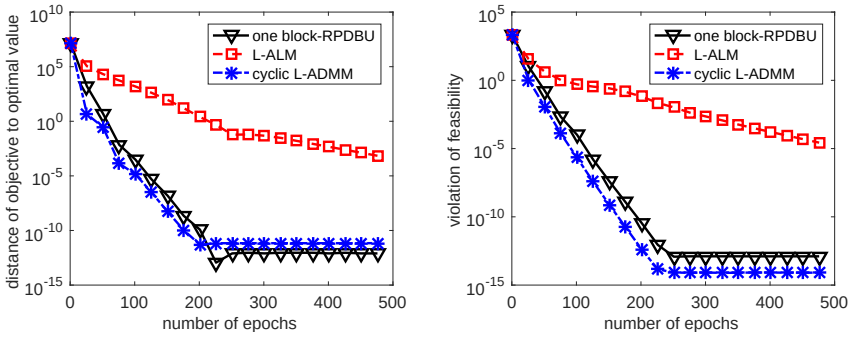

Figure 4.2: Comparison of the proposed method (RPDBU) to the linearized augmented Lagrangian method (L-ALM) and the cyclic linearized alternating direction method of multipliers (L-ADMM) on solving (4.58). Left: distance of objective to optimal value $|F(x^k) - F(x^*)|$; Right: violation of feasibility $\|Ax^k - b\|$.

primal-dual scheme in (4.9).

4.6.1 Randomized proximal coordinate descent

The randomized proximal coordinate descent (RPCD) was proposed in [90], where smooth convex optimization problems are considered. It was then extended in [100, 78] to deal with nonsmooth problems that can be formulated as

$$\min_x f(x_1, \cdots, x_N) + \sum_{i=1}^{N} u_i(x_i), \tag{4.59}$$

where $x = (x_1; \ldots; x_N)$. Toward solving (4.59), at each iteration k, the RPCD method first randomly selects one block i_k and then performs the update:

$$x_i^{k+1} = \begin{cases} \operatorname{argmin}_{x_i} \langle \nabla_i f(x^k), x_i \rangle + \frac{L_i}{2} \|x_i - x_i^k\|_2^2 + u_i(x_i), & \text{if } i = i_k, \\ x_i^k, & \text{if } i \neq i_k, \end{cases} \tag{4.60}$$

where L_i is the Lipschitz continuity constant of the partial gradient $\nabla_i f(x)$. With more than one blocks selected every time, (4.60) has been further extended into parallel coordinate descent in [101].

When there is no linear constraint and no y-variable in (4.1), then Algorithm 1 reduces to the scheme in (4.60) if $I_k = \{i_k\}$, i.e., only one block is chosen, and $P^k =$

$L_{i_k} I, \lambda^k = 0, \forall k$, and to the parallel coordinate descent in [101] if $I_k = \{i_k^1, \cdots, i_k^n\}$ and $P^k = \text{blkdiag}(L_{i_k^1} I, \cdots, L_{i_k^n} I), \lambda^k = 0, \forall k$. Although the convergence rate results in [100, 78, 101] are non-ergodic, we can easily strengthen our result to a non-ergodic one by noticing that (4.25) implies nonincreasing monotonicity of the objective if Algorithm 1 is applied to (4.59).

4.6.2 Stochastic block proximal gradient

For solving the problem (4.59) with a stochastic f, [25] proposes a stochastic block proximal gradient (SBPG) method, which iteratively performs the update in (4.60) with $\nabla_i f(x^k)$ replaced by a stochastic approximation. If f is Lipschitz differentiable, then an ergodic $O(1/\sqrt{t})$ convergence rate was shown. Setting $I_k = \{i_k\}, \forall k$, we reduce Algorithm 2 to the SBPG method, and thus our convergence results in Proposition 4.4.3 recover that in [25].

4.6.3 Multi-block ADMM

Without coupled functions or proximal terms, Algorithm 1 can be regarded as a randomized variant of the multi-block ADMM scheme in (4.6). While multi-block ADMM can diverge if the problem has three or more blocks, our result in Theorem 4.3.6 shows that $O(1/t)$ convergence rate is guaranteed if at each iteration, one randomly selected block is updated, followed by an update to the multiplier. Note that in the case of no coupled function and $n = 1$, (4.33) indicates that we can choose $P^k = 0$, i.e. without proximal term. Hence, randomization is a key to convergence.

When there are only two blocks, ADMM has been shown (e.g., [74]) to have an ergodic $O(1/t)$ convergence rate. If there are no coupled functions, (4.36) and (4.41a) both indicate that we can choose $\hat{P} = \rho_x A^\top A, \hat{Q} = \rho_y B^\top B$ if $\theta = 1$, i.e., all x and y blocks are selected. Thus according to (4.15), we can set $P^k = 0, Q^k = 0, \forall k$, in which case Algorithm 1 reduces to the classic 2-block ADMM. Hence, our results in Theorems 4.3.7 and 4.3.9 both recover the ergodic $O(1/t)$ convergence rate of ADMM for two-block convex optimization problems.

4.6.4 Proximal Jacobian parallel ADMM

In [27], the proximal Jacobian parallel ADMM (Prox-JADMM) was proposed to solve the linearly constrained multi-block separable convex optimization model

$$\min_x \sum_{i=1}^N u_i(x_i), \text{s.t.} \sum_{i=1}^N A_i x_i = b. \tag{4.61}$$

At each iteration, the Prox-JADMM method performs the updates for $i = 1, \ldots, n$ in parallel:

$$x_i^{k+1} = \operatorname*{argmin}_{x_i} u_i(x_i) - \langle \lambda^k, A_i x_i \rangle + \frac{\rho}{2} \left\| A_i x_i + \sum_{j \neq i} A_j x_j^k - b \right\|_2^2 + \frac{1}{2} \|x_i - x_i^k\|_{P_i}^2, \tag{4.62}$$

and then updates the multiplier by

$$\lambda^{k+1} = \lambda^k - \gamma\rho \left(\sum_{i=1}^N A_i x_i^{k+1} - b \right), \tag{4.63}$$

where $P_i \succ 0, \forall i$ and $\gamma > 0$ is a damping parameter. By choosing approapriate parameters, [27] established convergence rate of order $1/t$ based on norm square of the difference of two consecutive iterates.

If there is no y-variable or the coupled function f in (4.1), setting $I_k = [N], P^k = \text{blkdiag}(\rho_x A_1^\top A_1 + P_1, \cdots, \rho_x A_N^\top A_N + P_N) - \rho_x A^\top A \succeq 0$, $\forall k$, where P_i's are the same as those in (4.62), then Algorithm 1 reduces to the Prox-JADMM with $\gamma = 1$, and Theorem 4.3.6 provides a new convergence result in terms of the objective value and the feasibility measure.

4.6.5 Randomized primal-dual scheme in (4.9)

In this subsection, we show that the scheme in (4.9) is a special case of Algorithm 1. Let g be the convex conjugate of $g^* := h + \iota_Z$, namely, $g(y) = \sup_z \langle y, z \rangle - h(z) - \iota_Z(z)$. Then (4.8) is equivalent to the optimization problem:

$$\min_{x \in X} \sum_{i=1}^N u_i(x_i) + g(-Ax),$$

which can be further written as

$$\min_{x\in X, y} \sum_{i=1}^{N} u_i(x_i) + g(y), \text{s.t.} Ax + y = 0. \tag{4.64}$$

Proposition 4.6.1 *The scheme in* (4.9) *is equivalent to the following updates:*

$$x_i^{k+1} = \begin{cases} \underset{x_i\in X_i}{\operatorname{argmin}} \langle -z^k, A_i x_i\rangle + u_i(x_i) \\ +\frac{q}{2\eta}\|A_i(x_i - x_i^k) + r^k\|^2 + \frac{1}{2}\|x_i - x_i^k\|_{\tau I - \frac{q}{\eta}A_i^\top A_i}, & i = i_k, \\ x_i^k, & i \neq i_k, \end{cases} \tag{4.65a}$$

$$y^{k+1} = \underset{y}{\operatorname{argmin}}\, g(y) - \langle y, z^k\rangle + \frac{1}{2\eta}\|y + Ax^{k+1}\|^2, \tag{4.65b}$$

$$z^{k+1} = z^k - \frac{1}{\eta}(Ax^{k+1} + y^{k+1}), \tag{4.65c}$$

where $r^k = Ax^k + y^k$. *Therefore, it is a special case of Algorithm 1 applied to* (4.64) *with the setting of* $I_k = \{i_k\}, \rho_x = \frac{q}{\eta}, \rho_y = \rho = \frac{1}{\eta}$ *and* $P^k = \tau I - \frac{q}{\eta}A_{i_k}^\top A_{i_k}, Q^k = 0, \forall k$.

While the sublinear convergence rate result in [24] requires the boundedness of X and Z, the result in Theorem 4.3.7 indicates that the boundedness assumption can be removed if we add one proximal term to the y-update in (4.65b).

4.7 Concluding Remarks

We have proposed a randomized primal-dual coordinate update algorithm, called RPDBU, for solving linearly constrained convex optimization with multi-block decision variables and coupled terms in the objective. By using a randomization scheme and the proximal gradient mappings, we show a sublinear convergence rate of the RPDBU method. In particular, without any assumptions other than convexity on the objective function and without imposing any restrictions on the constraint matrices, an $O(1/t)$ convergence rate is established. We have also extended RPDBU to solve the problem where the objective is stochastic. If a stochastic (sub-)gradient estimator is available, then we show that by adaptively choosing the parameter α_k in the added proximal term, an $O(1/\sqrt{t})$ convergence rate can be established. Furthermore, if there is no coupled function f, then we can remove the proximal term, and the algorithm reduces to a randomized multi-

block ADMM. Hence, the convergence of the original randomized multi-block ADMM follows as a consequence of our analysis. Remark also that by taking the sampling set I_k as the whole set and P^k as some special matrices, our algorithm specializes to the proximal Jacobian ADMM. Finally, we pose as an open problem to decide whether or not a deterministic counterpart of the RPDBU exists, retaining similar convergence properties for solving problem (4.1). For instance, it would be interesting to know if the algorithm would still be convergent if a deterministic cyclic update rule is applied while a proper proximal term is incorporated.

4.8 Proofs of Lemmas

We give proofs of several lemmas that are used to show our main results.

4.8.1 Proof of Lemma 4.3.1

We prove (4.25), and (4.26) can be shown by the same arguments. By the optimality of $x_{I_k}^{k+1}$ from (4.10) or equivalently (4.16), we have for any $x_{I_k} \in X_{I_k}$,

$$(x_{I_k} - x_{I_k}^{k+1})^\top \left(\nabla_{I_k} f(x^k) - A_{I_k}^\top \lambda^k + \rho_x A_{I_k}^\top r^k + \tilde{\nabla} u_{I_k}(x_{I_k}^{k+1}) + \hat{P}_{I_k}(x_{I_k}^{k+1} - x_{I_k}^k) \right) \geq 0, \tag{4.66}$$

where $\tilde{\nabla} u_{I_k}(x_{I_k}^{k+1})$ is a subgradient of u_{I_k} at $x_{I_k}^{k+1}$, and we have used the formula of $r^{k+\frac{1}{2}}$ given in (4.11). We compute the expectation of each term in (4.66) in the following. First, we have

$$\begin{aligned}
& \mathbb{E}_{I_k}(x_{I_k} - x_{I_k}^{k+1})^\top \nabla_{I_k} f(x^k) \\
= \; & \mathbb{E}_{I_k}\left(x_{I_k} - x_{I_k}^k\right)^\top \nabla_{I_k} f(x^k) + \mathbb{E}_{I_k}(x_{I_k}^k - x_{I_k}^{k+1})^\top \nabla_{I_k} f(x^k) \\
= \; & \frac{n}{N}\left(x - x^k\right)^\top \nabla f(x^k) + \mathbb{E}_{I_k}(x^k - x^{k+1})^\top \nabla f(x^k) \qquad (4.67) \\
\leq \; & \frac{n}{N}\left(f(x) - f(x^k)\right) + \mathbb{E}_{I_k}\left[f(x^k) - f(x^{k+1}) + \frac{L_f}{2}\|x^k - x^{k+1}\|^2\right] \\
= \; & \frac{n-N}{N}\left(f(x) - f(x^k)\right) + \mathbb{E}_{I_k}\left[f(x) - f(x^{k+1}) + \frac{L_f}{2}\|x^k - x^{k+1}\|^2\right], \qquad (4.68)
\end{aligned}$$

where the last inequality is from the convexity of f and the Lipschitz continuity of $\nabla_{I_k} f(x)$. Secondly,

$$\begin{aligned} & \mathbb{E}_{I_k}(x_{I_k} - x_{I_k}^{k+1})^\top(-A_{I_k}^\top \lambda^k) \\ = \; & \mathbb{E}_{I_k}(x_{I_k} - x_{I_k}^{k})^\top(-A_{I_k}^\top \lambda^k) + \mathbb{E}_{I_k}(x_{I_k}^{k} - x_{I_k}^{k+1})^\top(-A_{I_k}^\top \lambda^k) \\ = \; & \frac{n}{N}\left(x - x^k\right)^\top(-A^\top\lambda^k) + \mathbb{E}_{I_k}(x^k - x^{k+1})^\top(-A^\top\lambda^k) \\ = \; & \frac{n-N}{N}\left(x - x^k\right)^\top(-A^\top\lambda^k) + \mathbb{E}_{I_k}(x - x^{k+1})^\top(-A^\top\lambda^k). \end{aligned} \tag{4.69}$$

Similarly,

$$\rho_x \mathbb{E}_{I_k}(x_{I_k} - x_{I_k}^{k+1})^\top A_{I_k}^\top r^k = \frac{n-N}{N}\rho_x(x - x^k)^\top A^\top r^k + \rho_x \mathbb{E}_{I_k}(x - x^{k+1})^\top A^\top r^k. \tag{4.70}$$

For the fourth term of (4.66), we have

$$\begin{aligned} & \mathbb{E}_{I_k}(x_{I_k} - x_{I_k}^{k+1})^\top \tilde{\nabla} u_{I_k}(x_{I_k}^{k+1}) \\ \leq \; & \mathbb{E}_{I_k}\left[u_{I_k}(x_{I_k}) - u_{I_k}(x_{I_k}^{k+1})\right] \\ = \; & \frac{n}{N}u(x) - \mathbb{E}_{I_k}[u(x^{k+1}) - u(x^k) + u_{I_k}(x_{I_k}^k)] \\ = \; & \frac{n}{N}\left[u(x) - u(x^k)\right] + \mathbb{E}_{I_k}[u(x^k) - u(x^{k+1})] \\ = \; & \frac{n-N}{N}\left[u(x) - u(x^k)\right] + \mathbb{E}_{I_k}[u(x) - u(x^{k+1})], \end{aligned} \tag{4.71}$$

where the inequality is from the convexity of u_{I_k}. Finally, we have

$$\mathbb{E}_{I_k}(x_{I_k} - x_{I_k}^{k+1})^\top \hat{P}_{I_k}(x_{I_k}^{k+1} - x_{I_k}^k) = \mathbb{E}_{I_k}(x - x^{k+1})^\top \hat{P}(x^{k+1} - x^k). \tag{4.72}$$

Plugging (4.68) through (4.72) into (4.66) and recalling $F(x) = f(x) + u(x)$, by rearranging terms we have

$$\begin{aligned} & \mathbb{E}_{I_k}\left[F(x^{k+1}) - F(x) + (x^{k+1} - x)^\top(-A^\top\lambda^k) + \rho_x(x^{k+1} - x)^\top A^\top r^k\right] \\ & + \mathbb{E}_{I_k}\left(x^{k+1} - x\right)^\top \hat{P}(x^{k+1} - x^k) - \frac{L_f}{2}\mathbb{E}_{I_k}\|x^k - x^{k+1}\|^2 \\ \leq \; & \frac{N-n}{N}\left[F(x^k) - F(x) + (x^k - x)^\top(-A^\top\lambda^k) + \rho_x(x^k - x)^\top A^\top r^k\right]. \end{aligned} \tag{4.73}$$

Note

$$(x^{k+1} - x)^\top(-A^\top\lambda^k) + \rho_x(x^{k+1} - x)^\top A^\top r^k$$

$$
\begin{aligned}
= \quad & (x^{k+1}-x)^\top(-A^\top\lambda^k) + \rho_x(x^{k+1}-x)^\top A^\top r^{k+1} \\
& -\rho_x(x^{k+1}-x)^\top A^\top A(x^{k+1}-x^k) - \rho_x(x^{k+1}-x)^\top A^\top B(y^{k+1}-y^k) \\
\overset{(4.14)}{=} \quad & (x^{k+1}-x)^\top(-A^\top\lambda^{k+1}) + (\rho_x-\rho)(x^{k+1}-x)^\top A^\top r^{k+1} \\
& -\rho_x(x^{k+1}-x)^\top A^\top A(x^{k+1}-x^k) - \rho_x(x^{k+1}-x)^\top A^\top B(y^{k+1}-y^k).
\end{aligned}
$$

Hence, we can rewrite (4.73) equivalently into (4.25). Through the same arguments, one can show (4.26), thus completing the proof.

4.8.2 Proof of Lemma 4.3.3

Letting $x = x^*, y = y^*$ in (4.31), we have for any λ that

$$
\begin{aligned}
& \mathbb{E}[h(x^*, y^*, \lambda)] \\
\geq \quad & \mathbb{E}\Big[\Phi(\hat{x},\hat{y}) - \Phi(x^*,y^*) + (\hat{x}-x^*)^\top(-A^\top\hat{\lambda}) \\
& +(\hat{y}-y^*)^\top(-B^\top\hat{\lambda}) + (\hat{\lambda}-\lambda)^\top(A\hat{x}+B\hat{y}-b)\Big] \\
= \quad & \mathbb{E}\Big[\Phi(\hat{x},\hat{y}) - \Phi(x^*,y^*) + \langle\hat{\lambda}, Ax^*+By^*-b\rangle - \langle\lambda, A\hat{x}+B\hat{y}-b\rangle\Big] \\
= \quad & \mathbb{E}\left[\Phi(\hat{x},\hat{y}) - \Phi(x^*,y^*) - \langle\lambda, A\hat{x}+B\hat{y}-b\rangle\right], \qquad (4.74)
\end{aligned}
$$

where the last equality follows from the feasibility of (x^*, y^*). For any $\gamma > 0$, restricting λ in $\mathcal{B}_\gamma$, we have

$$
\mathbb{E}[h(x^*, y^*, \lambda)] \leq \sup_{\lambda\in B_\gamma} h(x^*, y^*, \lambda).
$$

Hence, letting $\lambda = -\frac{\gamma(A\hat{x}+B\hat{y}-b)}{\|A\hat{x}+B\hat{y}-b\|} \in \mathcal{B}_\gamma$ in (4.74) gives the desired result.

4.8.3 Proof of Lemma 4.3.5

In view of (4.32), we have

$$
\mathbb{E}\big[(\gamma - \|\lambda^*\|)\|A\hat{x}+B\hat{y}-b\|\big] \leq \mathbb{E}\big[\Phi(\hat{x},\hat{y}) - \Phi(x^*,y^*) + \gamma\|A\hat{x}+B\hat{y}-b\|\big] \leq \epsilon,
$$

which implies

$$
\mathbb{E}\|A\hat{x}+B\hat{y}-b\| \leq \frac{\epsilon}{\gamma - \|\lambda^*\|}, \text{ and } \mathbb{E}\big[\Phi(\hat{x},\hat{y}) - \Phi(x^*,y^*)\big] \leq \epsilon. \qquad (4.75)
$$

Denote $a^- = \max(0, -a)$ for a real number a. Then from (4.32) and (4.75), it follows that

$$\mathbb{E}\big(\Phi(\hat{x},\hat{y}) - \Phi(x^*,y^*)\big)^- \le \|\lambda^*\| \cdot \mathbb{E}\|A\hat{x} + B\hat{y} - b\| \le \frac{\|\lambda^*\|\epsilon}{\gamma - \|\lambda^*\|}.$$

Noting $|a| = a + 2a^-$ for any real number a, we have

$$\begin{aligned} & \mathbb{E}\big|\Phi(\hat{x},\hat{y}) - \Phi(x^*,y^*)\big| \\ = \;& \mathbb{E}\big[\Phi(\hat{x},\hat{y}) - \Phi(x^*,y^*)\big] + 2\mathbb{E}\big(\Phi(\hat{x},\hat{y}) - \Phi(x^*,y^*)\big)^- \\ \le \;& \left(\frac{2\|\lambda^*\|}{\gamma - \|\lambda^*\|} + 1\right)\epsilon. \end{aligned} \tag{4.76}$$

4.8.4 Proof of Inequalities (4.52c) and (4.52d) with $\alpha_k = \frac{\alpha_0}{\sqrt{k}}$

We have $\beta_k = \frac{\alpha_0}{\rho\left(\sqrt{k}-(1-\theta)\sqrt{k-1}\right)}$, and

$$\begin{aligned} & \frac{\alpha_{k-1}}{\alpha_k}\beta_k + (1-\theta)\beta_{k+1} - \frac{\alpha_k}{\alpha_{k+1}}\beta_{k+1} - (1-\theta)\beta_k \\ = \;& \frac{\alpha_0}{\rho}\left[\left(\frac{\sqrt{k}}{\sqrt{k-1}} - (1-\theta)\right)\frac{1}{\left(\sqrt{k} - (1-\theta)\sqrt{k-1}\right)}\right. \\ & \left. - \left(\frac{\sqrt{k+1}}{\sqrt{k}} - (1-\theta)\right)\frac{1}{\left(\sqrt{k+1} - (1-\theta)\sqrt{k}\right)}\right] \\ =: \;& \frac{\alpha_0}{\rho}[\psi(k) - \psi(k+1)]. \end{aligned}$$

By elementary calculus, we have

$$\begin{aligned}
\psi'(k) &= \frac{\frac{\sqrt{k-1}}{\sqrt{k}} - \frac{\sqrt{k}}{\sqrt{k-1}}}{2(k-1)} \frac{1}{\left(\sqrt{k} - (1-\theta)\sqrt{k-1}\right)} \\
&\quad + \left(\frac{\sqrt{k}}{\sqrt{k-1}} - (1-\theta)\right) \frac{-1}{2\left(\sqrt{k} - (1-\theta)\sqrt{k-1}\right)^2} \left(\frac{1}{\sqrt{k}} - \frac{1-\theta}{\sqrt{k-1}}\right) \\
&= \frac{1}{2(k-1)(\sqrt{k} - (1-\theta)\sqrt{k-1})} \left[\frac{\sqrt{k-1}}{\sqrt{k}} - \frac{\sqrt{k}}{\sqrt{k-1}}\right. \\
&\quad \left. -\sqrt{k-1}\left(\frac{1}{\sqrt{k}} - \frac{1-\theta}{\sqrt{k-1}}\right)\right] \\
&= \frac{1}{2(k-1)(\sqrt{k} - (1-\theta)\sqrt{k-1})} \left((1-\theta) - \frac{\sqrt{k}}{\sqrt{k-1}}\right) < 0.
\end{aligned}$$

Hence, $\psi(k)$ is decreasing with respect to k, and thus (4.52c) holds.

When $\alpha_k = \frac{\alpha_0}{\sqrt{k}}$, (4.52d) becomes

$$\frac{\alpha_0}{2\rho\sqrt{t}} \geq \left| \left(\frac{\sqrt{t}}{\sqrt{t-1}} - (1-\theta)\right) \frac{\alpha_0}{\rho(\sqrt{t} - (1-\theta)\sqrt{t-1})} - \frac{\alpha_0}{\rho\sqrt{t}} \right|,$$

which is equivalent to

$$\frac{1}{2} \geq \frac{\sqrt{t}}{\sqrt{t-1}} - 1 \Longleftrightarrow t \geq \frac{9}{5}.$$

This completes the proof.

4.9 Proofs of Theorems

In this section, we give the technical details for showing all theorems. For simplicity of notation, throughout the proofs of this section, we define $\tilde{P}$ and $\tilde{Q}$ as follows:

$$\tilde{P} = \hat{P} - \rho_x A^\top A, \qquad \tilde{Q} = \hat{Q} - \rho_y B^\top B. \tag{4.77}$$

4.9.1 Proof of Theorem 4.3.6

Taking expectation over both sides of (4.25) and summing it over $k = 0$ through t, we have

$$\begin{aligned}
&\mathbb{E}\big[F(x^{t+1}) - F(x) + (x^{t+1} - x)^\top(-A^\top\lambda^{t+1})\big] + (1-\theta)\rho_x\mathbb{E}(x^{t+1} - x)^\top A^\top r^{t+1} \\
&+\theta\sum_{k=0}^{t-1}\mathbb{E}\big[F(x^{k+1}) - F(x) + (x^{k+1} - x)^\top(-A^\top\lambda^{k+1})\big] \\
&-\sum_{k=0}^{t}\rho_x\mathbb{E}(x^{k+1} - x)^\top A^\top B(y^{k+1} - y^k) \\
&+\sum_{k=0}^{t}\mathbb{E}(x^{k+1} - x)^\top\tilde{P}(x^{k+1} - x^k) - \frac{L_f}{2}\sum_{k=0}^{t}\mathbb{E}\|x^k - x^{k+1}\|^2 \\
\leq\ & (1-\theta)\left[F(x^0) - F(x) + (x^0 - x)^\top(-A^\top\lambda^0) + \rho_x(x^0 - x)^\top A^\top r^0\right], \qquad (4.78)
\end{aligned}$$

where we have used $\frac{n}{N} = \theta$, the condition in (4.40) and the definition of $\tilde{P}$ in (4.77). Similarly, taking expectation over both sides of (4.26), summing it over $k = 0$ through t, we have

$$\begin{aligned}
&\mathbb{E}\big[G(y^{t+1}) - G(y) + (y^{t+1} - y)^\top(-B^\top\lambda^{t+1})\big] + (1-\theta)\rho_y\mathbb{E}(y^{t+1} - y)^\top B^\top r^{t+1} \\
&+\theta\sum_{k=0}^{t-1}\mathbb{E}\big[G(y^{k+1}) - G(y) + (y^{k+1} - y)^\top(-B^\top\lambda^{k+1})\big] \\
&+\sum_{k=0}^{t}\mathbb{E}(y^{k+1} - y)^\top\tilde{Q}(y^{k+1} - y^k) - \frac{L_g}{2}\sum_{k=0}^{t}\mathbb{E}\|y^k - y^{k+1}\|^2 \\
\leq\ & (1-\theta)\left[G(y^0) - G(y) + (y^0 - y)^\top(-B^\top\lambda^0) + \rho_y(y^0 - y)^\top B^\top r^0\right] \qquad (4.79) \\
&+(1-\theta)\sum_{k=0}^{t}\mathbb{E}\rho_y(y^k - y)^\top B^\top A(x^{k+1} - x^k).
\end{aligned}$$

Recall $\lambda^{k+1} = \lambda^k - \rho r^{k+1}$, thus

$$(\lambda^{k+1} - \lambda)^\top r^{k+1} = -\frac{1}{\rho}(\lambda^{k+1} - \lambda)^\top(\lambda^{k+1} - \lambda^k), \qquad (4.80)$$

where λ is an arbitrary vector and possibly random. Denote $\tilde{\lambda}^{t+1} = \lambda^t - \rho_x r^{t+1}$. Then

similar to (4.80), we have

$$(\tilde{\lambda}^{t+1}-\lambda)^\top r^{t+1} = -\frac{1}{\rho_x}(\tilde{\lambda}^{t+1}-\lambda)^\top(\tilde{\lambda}^{t+1}-\lambda^t). \tag{4.81}$$

Summing (4.78) and (4.79) together and using (4.80) and (4.81), we have:

$$\begin{aligned}
&\mathbb{E}\left[\Phi(x^{t+1},y^{t+1})-\Phi(x,y)+(\tilde{w}^{t+1}-w)^\top H(\tilde{w}^{t+1})+\frac{1}{\rho_x}(\tilde{\lambda}^{t+1}-\lambda)^\top(\tilde{\lambda}^{t+1}-\lambda^t)\right]\\
&+\theta\sum_{k=0}^{t-1}\mathbb{E}\left[\Phi(x^{k+1},y^{k+1})-\Phi(x,y)+(w^{k+1}-w)^\top H(w^{k+1})\right.\\
&\left.+\frac{1}{\rho}(\lambda^{k+1}-\lambda)^\top(\lambda^{k+1}-\lambda^k)\right]\\
\leq\ &(1-\theta)\left[F(x^0)-F(x)+(x^0-x)^\top(-A^\top\lambda^0)+\rho_x(x^0-x)^\top A^\top r^0\right]\\
&+(1-\theta)\left[G(y^0)-G(y)+(y^0-y)^\top(-B^\top\lambda^0)+\rho_y(y^0-y)^\top B^\top r^0\right]\\
&+\sum_{k=0}^{t}\rho_x\mathbb{E}(x^{k+1}-x)^\top A^\top B(y^{k+1}-y^k)\\
&+(1-\theta)\sum_{k=0}^{t}\rho_y\mathbb{E}(y^k-y)^\top B^\top A(x^{k+1}-x^k)\\
&-\sum_{k=0}^{t}\mathbb{E}(x^{k+1}-x)^\top\tilde{P}(x^{k+1}-x^k)+\frac{L_f}{2}\sum_{k=0}^{t}\mathbb{E}\|x^k-x^{k+1}\|^2\\
&-\sum_{k=0}^{t}\mathbb{E}(y^{k+1}-y)^\top\tilde{Q}(y^{k+1}-y^k)+\frac{L_g}{2}\sum_{k=0}^{t}\mathbb{E}\|y^k-y^{k+1}\|^2,
\end{aligned}\tag{4.82}$$

where we have used $\Phi(x,y)=F(x)+G(y)$ and the definition of H given in (4.18).

When $B=0$ and $y^k\equiv y^0$, (4.82) reduces to

$$\begin{aligned}
&\mathbb{E}\left[F(x^{t+1})-F(x)+(\tilde{w}^{t+1}-w)^\top H(\tilde{w}^{t+1})+\frac{1}{\rho_x}(\tilde{\lambda}^{t+1}-\lambda)^\top(\tilde{\lambda}^{t+1}-\lambda^t)\right]\\
&+\theta\sum_{k=0}^{t-1}\mathbb{E}\left[F(x^{k+1})-F(x)+(w^{k+1}-w)^\top H(w^{k+1})+\frac{1}{\rho}(\lambda^{k+1}-\lambda)^\top(\lambda^{k+1}-\lambda^k)\right]\\
\leq\ &(1-\theta)\left[F(x^0)-F(x)+(x^0-x)^\top(-A^\top\lambda^0)+\rho_x(x^0-x)^\top A^\top r^0\right]\\
&-\sum_{k=0}^{t}\mathbb{E}(x^{k+1}-x)^\top\tilde{P}(x^{k+1}-x^k)+\frac{L_f}{2}\sum_{k=0}^{t}\mathbb{E}\|x^k-x^{k+1}\|^2.
\end{aligned}$$

Using (4.21) and noting $\theta=\frac{\rho}{\rho_x}$, from the above inequality after cancelling terms we

have

$$\begin{aligned}
&\mathbb{E}\left[F(x^{t+1}) - F(x) + (\tilde{w}^{t+1} - w)^\top H(\tilde{w}^{t+1})\right] \\
&+\theta \sum_{k=0}^{t-1} \mathbb{E}\left[F(x^{k+1}) - F(x) + (w^{k+1} - w)^\top H(w^{k+1})\right] \\
&+\frac{1}{2\rho_x}\mathbb{E}\left[\|\tilde{\lambda}^{t+1} - \lambda\|^2 - \|\lambda^0 - \lambda\|^2 + \|\tilde{\lambda}^{t+1} - \lambda^t\|^2 + \sum_{k=0}^{t-1}\|\lambda^{k+1} - \lambda^k\|^2\right] \\
\leq\ & (1-\theta)\left[F(x^0) - F(x) + (x^0 - x)^\top(-A^\top\lambda^0) + \rho_x(x^0 - x)^\top A^\top r^0\right] \\
&-\frac{1}{2}\mathbb{E}\left[\|x^{t+1} - x\|_{\tilde{P}}^2 - \|x^0 - x\|_{\tilde{P}}^2 + \sum_{k=0}^{t}\|x^{k+1} - x^k\|_{\tilde{P}}^2\right] + \frac{L_f}{2}\sum_{k=0}^{t}\mathbb{E}\|x^k - x^{k+1}\|^2.
\end{aligned} \tag{4.83}$$

For any feasible x, we note $\tilde{\lambda}^{t+1} - \lambda^t = \rho_x A(x^{t+1} - x)$ and thus

$$\frac{1}{\rho_x}\|\tilde{\lambda}^{t+1} - \lambda^t\|^2 = \rho_x\|x^{t+1} - x\|_{A^\top A}^2. \tag{4.84}$$

In addition, since x^{k+1} and x^k differ only on the index set I_k, we have by recalling $\tilde{P} = \hat{P} - \rho_x A^\top A$ that

$$\|x^{k+1} - x^k\|_{\tilde{P}}^2 - L_f\|x^{k+1} - x^k\|^2 = \|x_{I_k}^{k+1} - x_{I_k}^k\|_{\hat{P}_{I_k}}^2 - \|x_{I_k}^{k+1} - x_{I_k}^k\|_{\rho_x A_{I_k}^\top A_{I_k}}^2 - L_f\|x_{I_k}^{k+1} - x_{I_k}^k\|^2. \tag{4.85}$$

Plugging (4.84) and (4.85) into (4.83), and using (4.33) leads to

$$\begin{aligned}
&\mathbb{E}\left[F(x^{t+1}) - F(x) + (\tilde{w}^{t+1} - w)^\top H(\tilde{w}^{t+1})\right] \\
&+\theta \sum_{k=0}^{t-1} \mathbb{E}\left[F(x^{k+1}) - F(x) + (w^{k+1} - w)^\top H(w^{k+1})\right] \\
\leq\ & (1-\theta)\left[F(x^0) - F(x) + (x^0 - x)^\top(-A^\top\lambda^0) + \rho_x(x^0 - x)^\top A^\top r^0\right] \\
&+\frac{1}{2\rho_x}\mathbb{E}\|\lambda^0 - \lambda\|^2 + \frac{1}{2}\|x^0 - x\|_{\tilde{P}}^2.
\end{aligned}$$

The desired result follows from $\lambda^0 = 0$, and Lemmas 4.3.3 and 4.3.5 with $\gamma = \max\{0.5 + \|\lambda^*\|, 3\|\lambda^*\|\}$.

4.9.2 Proof of Theorem 4.3.7

It follows from (4.26) with $\rho_y = \rho$ and $m = M$ that (recall the definition of $\tilde{Q}$ in (4.77)) for any $y \in Y$,

$$G(y^{k+1}) - G(y) - \frac{L_g}{2}\|y^k - y^{k+1}\|^2 + (y^{k+1} - y)^\top(-B^\top \lambda^{k+1}) + (y^{k+1} - y)^\top \tilde{Q}(y^{k+1} - y^k) \leq 0. \tag{4.86}$$

Similar to (4.86), and recall the definition of $\tilde{y}^{t+1}$, we have for any $y \in Y$,

$$G(\tilde{y}^{t+1}) - G(y) - \frac{L_g}{2}\|\tilde{y}^{t+1} - y^t\|^2 + (\tilde{y}^{t+1} - y)^\top(-B^\top \tilde{\lambda}^{t+1}) + \theta(\tilde{y}^{t+1} - y)^\top \tilde{Q}(y^{t+1} - y^t) \leq 0, \tag{4.87}$$

where

$$\tilde{\lambda}^{t+1} = \lambda^t - \rho_x(Ax^{t+1} + B\tilde{y}^{t+1} - b). \tag{4.88}$$

Adding (4.86) and (4.87) to (4.78) and using the formula of λ^k gives

$$\begin{aligned}
& \mathbb{E}\left[F(x^{t+1}) - F(x) + (x^{t+1} - x)^\top(-A^\top \tilde{\lambda}^{t+1})\right] \\
& + \mathbb{E}\left(\tilde{\lambda}^{t+1} - \lambda\right)^\top \left(Ax^{t+1} + B\tilde{y}^{t+1} - b + \frac{1}{\rho_x}(\tilde{\lambda}^{t+1} - \lambda^t)\right) \\
& + \mathbb{E}\left[G(\tilde{y}^{t+1}) - G(y) + (\tilde{y}^{t+1} - y)^\top(-B^\top \tilde{\lambda}^{t+1}) + \theta(\tilde{y}^{t+1} - y)^\top \tilde{Q}(\tilde{y}^{t+1} - y^k)\right] \\
& + \theta \sum_{k=0}^{t-1} \mathbb{E}\left[F(x^{k+1}) - F(x) + (x^{k+1} - x)^\top(-A^\top \lambda^{k+1})\right] - \frac{L_g}{2}\mathbb{E}\|\tilde{y}^{t+1} - y^t\|^2 \\
& - \sum_{k=0}^{t-1} \rho_x \mathbb{E}(x^{k+1} - x)^\top A^\top B(y^{k+1} - y^k) - \rho_x \mathbb{E}(x^{t+1} - x)^\top A^\top B(\tilde{y}^{t+1} - y^t) \\
& + \theta \sum_{k=0}^{t-1} \mathbb{E}\left[G(y^{k+1}) - G(y) - \frac{L_g}{2}\|y^k - y^{k+1}\|^2\right. \\
& \left. + (y^{k+1} - y)^\top(-B^\top \lambda^{k+1}) + (y^{k+1} - y)^\top \tilde{Q}(y^{k+1} - y^k)\right] \\
& + \theta \sum_{k=0}^{t-1} \mathbb{E}(\lambda^{k+1} - \lambda)^\top \left(r^{k+1} + \frac{1}{\rho}(\lambda^{k+1} - \lambda^k)\right) \\
\leq\ & (1 - \theta)\left[F(x^0) - F(x) + (x^0 - x)^\top(-A^\top \lambda^0) + \rho_x(x^0 - x)^\top A^\top r^0\right] \\
& - \sum_{k=0}^{t} \mathbb{E}(x^{k+1} - x)^\top \tilde{P}(x^{k+1} - x^k) + \frac{L_f}{2}\sum_{k=0}^{t} \mathbb{E}\|x^k - x^{k+1}\|^2.
\end{aligned} \tag{4.89}$$

By the notation in (4.18) and using (4.29), (4.89) can be written into

$$
\begin{aligned}
& \mathbb{E}\left[\Phi(x^{t+1},\tilde{y}^{t+1})-\Phi(x,y)+(\tilde{w}^{t+1}-w)^\top H(\tilde{w}^{t+1})\right] \\
& +\theta\sum_{k=0}^{t-1}\mathbb{E}\left[\Phi(x^{k+1},y^{k+1})-\Phi(x,y)+(w^{k+1}-w)^\top H(w^{k+1})\right] \\
& +\theta\sum_{k=0}^{t-1}\mathbb{E}(y^{k+1}-y)^\top\tilde{Q}(y^{k+1}-y^k)+\theta\mathbb{E}(\tilde{y}^{t+1}-y)^\top\tilde{Q}(\tilde{y}^{t+1}-y^t) \\
& -\sum_{k=0}^{t-1}\rho_x\mathbb{E}\left(\frac{1}{\rho}(\lambda^k-\lambda^{k+1})^\top B(y^{k+1}-y^k)-(y^{k+1}-y)^\top B^\top B(y^{k+1}-y^k)\right) \\
& -\rho_x\mathbb{E}\left(\frac{1}{\rho_x}(\lambda^t-\tilde{\lambda}^{t+1})^\top B(\tilde{y}^{t+1}-y^t)-(\tilde{y}^{t+1}-y)^\top B^\top B(\tilde{y}^{t+1}-y^t)\right) \\
& +\frac{\theta}{\rho}\mathbb{E}(\tilde{\lambda}^{t+1}-\lambda)^\top\left(\tilde{\lambda}^{t+1}-\lambda^t\right)+\frac{\theta}{\rho}\sum_{k=0}^{t-1}\mathbb{E}(\lambda^{k+1}-\lambda)^\top\left(\lambda^{k+1}-\lambda^k\right) \\
\leq\ & (1-\theta)\left[F(x^0)-F(x)+(x^0-x)^\top(-A^\top\lambda^0)+\rho_x(x^0-x)^\top A^\top r^0\right] \\
& -\sum_{k=0}^{t}\mathbb{E}(x^{k+1}-x)^\top\tilde{P}(x^{k+1}-x^k)+\frac{L_f}{2}\sum_{k=0}^{t}\mathbb{E}\|x^k-x^{k+1}\|^2 \\
& +\frac{\theta L_g}{2}\sum_{k=0}^{t-1}\mathbb{E}\|y^k-y^{k+1}\|^2+\frac{L_g}{2}\mathbb{E}\|\tilde{y}^{t+1}-y^t\|^2.
\end{aligned}
$$

Now use (4.21) to derive from the above inequality that

$$
\begin{aligned}
& \mathbb{E}\left[\Phi(x^{t+1},\tilde{y}^{t+1})-\Phi(x,y)+(\tilde{w}^{t+1}-w)^\top H(\tilde{w}^{t+1})\right] \\
& +\theta\sum_{k=0}^{t-1}\mathbb{E}\left[\Phi(x^{k+1},y^{k+1})-\Phi(x,y)+(w^{k+1}-w)^\top H(w^{k+1})\right] \\
& +\frac{\theta}{2}\left(\mathbb{E}\|\tilde{y}^{t+1}-y\|_{\tilde{Q}}^2-\|y^0-y\|_{\tilde{Q}}^2\right)+\frac{\theta}{2}\sum_{k=0}^{t-1}\mathbb{E}\|y^{k+1}-y^k\|_{\tilde{Q}}^2+\frac{\theta}{2}\mathbb{E}\|\tilde{y}^{t+1}-y^t\|_{\tilde{Q}}^2 \\
& +\frac{\rho_x}{2}\left(\mathbb{E}\|\tilde{y}^{t+1}-y\|_{B^\top B}^2-\|y^0-y\|_{B^\top B}^2\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{\rho_x}{2}\sum_{k=0}^{t-1}\mathbb{E}\|y^{k+1}-y^k\|^2_{B^\top B}+\frac{\rho_x}{2}\mathbb{E}\|\tilde{y}^{t+1}-y^t\|^2_{B^\top B}\\
&-\sum_{k=0}^{t-1}\mathbb{E}\frac{\rho_x}{\rho}\left(\lambda^k-\lambda^{k+1}\right)^\top B(y^{k+1}-y^k)-\mathbb{E}(\lambda^t-\tilde{\lambda}^{t+1})^\top B(\tilde{y}^{t+1}-y^t)\\
&+\frac{\theta}{2\rho}\left(\mathbb{E}\|\tilde{\lambda}^{t+1}-\lambda\|^2-\|\lambda^0-\lambda\|^2\right)+\frac{\theta}{2\rho}\sum_{k=0}^{t-1}\mathbb{E}\|\lambda^{k+1}-\lambda^k\|^2+\frac{\theta}{2\rho}\mathbb{E}\|\tilde{\lambda}^{t+1}-\lambda^t\|^2\\
\leq\ & (1-\theta)\left[F(x^0)-F(x)+(x^0-x)^\top(-A^\top\lambda^0)+\rho_x(x^0-x)^\top A^\top r^0\right]\\
&-\frac{1}{2}\left[\mathbb{E}\|x^{t+1}-x\|^2_{\tilde{P}}-\|x^0-x\|^2_{\tilde{P}}+\sum_{k=0}^{t}\mathbb{E}\|x^k-x^{k+1}\|^2_{\tilde{P}}\right]+\frac{L_f}{2}\sum_{k=0}^{t}\mathbb{E}\|x^k-x^{k+1}\|^2\\
&+\frac{\theta L_g}{2}\sum_{k=0}^{t-1}\mathbb{E}\|y^k-y^{k+1}\|^2+\frac{L_g}{2}\mathbb{E}\|\tilde{y}^{t+1}-y^t\|^2. \qquad (4.90)
\end{aligned}
$$

Note that for $k\leq t-1$,

$$
-\frac{\rho_x}{\rho}(\lambda^k-\lambda^{k+1})^\top B(y^{k+1}-y^k)+\frac{\theta}{2\rho}\|\lambda^{k+1}-\lambda^k\|^2\geq-\frac{\rho}{2\theta^3}\|y^{k+1}-y^k\|^2_{B^\top B}
$$

and

$$
-(\lambda^t-\tilde{\lambda}^{t+1})^\top B(\tilde{y}^{t+1}-y^t)+\frac{\theta}{2\rho}\|\tilde{\lambda}^{t+1}-\lambda^t\|^2\geq-\frac{\rho}{2\theta}\|\tilde{y}^{t+1}-y^t\|^2_{B^\top B}.
$$

Because $\tilde{P},\tilde{Q}$ and ρ satisfy (4.36), we have from (4.90) that

$$
\begin{aligned}
&\mathbb{E}\left[\Phi(x^{t+1},\tilde{y}^{t+1})-\Phi(x,y)+(\tilde{w}^{t+1}-w)^\top H(\tilde{w}^{t+1})\right]\\
&+\theta\sum_{k=0}^{t-1}\mathbb{E}\left[\Phi(x^{k+1},y^{k+1})-\Phi(x,y)+(w^{k+1}-w)^\top H(w^{k+1})\right]\\
\leq\ & (1-\theta)\left[F(x^0)-F(x)+(x^0-x)^\top(-A^\top\lambda^0)+\rho_x(x^0-x)^\top A^\top r^0\right]\\
&+\frac{1}{2}\|x^0-x\|^2_{\tilde{P}}+\frac{\theta}{2}\|y^0-y\|^2_{\tilde{Q}}+\frac{\rho}{2\theta}\|y^0-y\|^2_{B^\top B}+\frac{\theta}{2\rho}\mathbb{E}\|\lambda^0-\lambda\|^2.
\end{aligned}
$$

Similar to Theorem 4.3.9, from the convexity of Φ and (4.23), we have

$$
\begin{aligned}
&(1+\theta t)\mathbb{E}\big[\Phi(\hat{x}^t,\hat{y}^t)-\Phi(x,y)+(\hat{w}^{t+1}-w)^\top H(w)\big]\\
\leq\ & (1-\theta)\big[F(x^0)-F(x)+(x^0-x)^\top(-A^\top\lambda^0)+\rho_x(x^0-x)^\top A^\top r^0\big]\\
&+\frac{1}{2}\|x^0-x\|^2_{\tilde{P}}+\frac{\theta}{2}\|y^0-y\|^2_{\tilde{Q}}+\frac{\rho}{2\theta}\|y^0-y\|^2_{B^\top B}+\frac{\theta}{2\rho}\mathbb{E}\|\lambda^0-\lambda\|^2. \qquad (4.91)
\end{aligned}
$$

Noting $\lambda^0 = 0$ and $(x^0 - x)^\top A^\top r^0 \leq \frac{1}{2}\left[\|x^0 - x\|_{A^\top A} + \|r^0\|^2\right]$, and using Lemmas 4.3.3 and 4.3.5 with $\gamma = \max\{0.5 + \|\lambda^*\|, 3\|\lambda^*\|\}$, we obtain the result (4.39).

4.9.3 Proof of Theorem 4.3.9

Using (4.29) and (4.30), applying (4.21) to the cross terms, and also noting the definition of $\tilde{P}$ and $\tilde{Q}$ in (4.77), we have

$$
\begin{aligned}
& -\frac{\theta}{\rho}\mathbb{E}\left[\sum_{k=0}^{t-1}(\lambda^{k+1} - \lambda)^\top(\lambda^{k+1} - \lambda^k) + (\tilde{\lambda}^{t+1} - \lambda)^\top(\tilde{\lambda}^{t+1} - \lambda^t)\right] \\
& + \sum_{k=0}^{t}\rho_x\mathbb{E}(x^{k+1} - x)^\top A^\top B(y^{k+1} - y^k) \\
& + (1-\theta)\sum_{k=0}^{t}\rho_y\mathbb{E}(y^k - y)^\top B^\top A(x^{k+1} - x^k) \\
& - \sum_{k=0}^{t}\mathbb{E}(x^{k+1} - x)^\top \tilde{P}(x^{k+1} - x^k) + \frac{L_f}{2}\sum_{k=0}^{t}\mathbb{E}\|x^k - x^{k+1}\|^2 \\
& - \sum_{k=0}^{t}\mathbb{E}(y^{k+1} - y)^\top \tilde{Q}(y^{k+1} - y^k) + \frac{L_g}{2}\sum_{k=0}^{t}\mathbb{E}\|y^k - y^{k+1}\|^2 \\
= \; & -\frac{\theta}{2\rho}\mathbb{E}\left[\|\tilde{\lambda}^{t+1} - \lambda\|^2 - \|\lambda^0 - \lambda\|^2 + \sum_{k=0}^{t-1}\|\lambda^{k+1} - \lambda^k\|^2 + \|\tilde{\lambda}^{t+1} - \lambda^t\|^2\right] \\
& + \frac{\rho_x}{\rho}\sum_{k=0}^{t}\mathbb{E}(\lambda^k - \lambda^{k+1})^\top B(y^{k+1} - y^k) \\
& + \frac{(1-\theta)\rho_y}{\rho}\sum_{k=0}^{t}\mathbb{E}(\lambda^{k-1} - \lambda^k)^\top A(x^{k+1} - x^k) \\
& - \frac{\theta\rho_y}{2}\mathbb{E}\left(\|x^0 - x\|^2_{A^\top A} - \|x^{t+1} - x\|^2_{A^\top A}\right) + \frac{(2-\theta)\rho_y}{2}\sum_{k=0}^{t}\mathbb{E}\|x^{k+1} - x^k\|^2_{A^\top A} \\
& - \frac{1}{2}\mathbb{E}\left(\|x^{t+1} - x\|^2_{\hat{P}} - \|x^0 - x\|^2_{\hat{P}} + \sum_{k=0}^{t}\|x^{k+1} - x^k\|^2_{\hat{P}}\right) + \frac{L_f}{2}\sum_{k=0}^{t}\mathbb{E}\|x^k - x^{k+1}\|^2 \\
& - \frac{1}{2}\mathbb{E}\left(\|y^{t+1} - y\|^2_{\hat{Q}} - \|y^0 - y\|^2_{\hat{Q}} + \sum_{k=0}^{t}\|y^{k+1} - y^k\|^2_{\hat{Q}}\right) + \frac{L_g}{2}\sum_{k=0}^{t}\mathbb{E}\|y^k - y^{k+1}\|^2,
\end{aligned}
\tag{4.92}
$$

where we have used the conditions in (4.40).

By Young's inequality, we have that for $0 \leq k \leq t$,

$$\begin{aligned} & \frac{\rho_x}{\rho}(\lambda^k - \lambda^{k+1})^\top B(y^{k+1} - y^k) - \frac{\theta}{2\rho}\frac{1}{2-\theta}\|\lambda^{k+1} - \lambda^k\|^2 \\ \leq \; & \frac{\rho}{\theta}\frac{2-\theta}{2}\frac{\rho_x^2}{\rho^2}\|B(y^{k+1} - y^k)\|^2 \overset{(4.40)}{=} \frac{(2-\theta)\rho_y}{2\theta^2}\|y^{k+1} - y^k\|^2_{B^\top B}, \end{aligned} \tag{4.93}$$

and for $1 \leq k \leq t$,

$$\begin{aligned} & \frac{(1-\theta)\rho_y}{\rho}(\lambda^{k-1} - \lambda^k)^\top A(x^{k+1} - x^k) - \frac{\theta}{2\rho}\frac{1-\theta}{2-\theta}\|\lambda^{k-1} - \lambda^k\|^2 \\ \leq \; & (1-\theta)\frac{\rho}{\theta}\frac{(2-\theta)\rho_y^2}{2\rho^2}\|A(x^{k+1} - x^k)\|^2 \overset{(4.40)}{=} \frac{(1-\theta)(2-\theta)}{2\theta^2}\rho_x\|x^{k+1} - x^k\|^2_{A^\top A}. \end{aligned} \tag{4.94}$$

Plugging (4.93) and (4.94) and also noting $\|\tilde{\lambda}^{t+1} - \lambda^t\|^2 \geq \|\lambda^{t+1} - \lambda^t\|^2$, we can upper bound the right hand side of (4.92) by

$$\begin{aligned} & -\frac{\theta}{2\rho}\mathbb{E}\left[\|\tilde{\lambda}^{t+1} - \lambda\|^2 - \|\lambda^0 - \lambda\|^2\right] - \frac{\theta\rho_y}{2}\mathbb{E}\left(\|x^0 - x\|^2_{A^\top A} - \|x^{t+1} - x\|^2_{A^\top A}\right) \\ & + \left(\frac{(1-\theta)(2-\theta)}{2\theta^2}\rho_x + \frac{(2-\theta)\rho_y}{2}\right)\sum_{k=0}^{t}\mathbb{E}\|x^{k+1} - x^k\|^2_{A^\top A} \\ & + \frac{(2-\theta)\rho_y}{2\theta^2}\sum_{k=0}^{t}\mathbb{E}\|y^{k+1} - y^k\|^2_{B^\top B} \\ & - \frac{1}{2}\mathbb{E}\left(\|x^{t+1} - x\|^2_{\hat{P}} - \|x^0 - x\|^2_{\hat{P}} + \sum_{k=0}^{t}\|x^{k+1} - x^k\|^2_{\hat{P}}\right) \\ & + \frac{L_f}{2}\sum_{k=0}^{t}\mathbb{E}\|x^k - x^{k+1}\|^2 \\ & - \frac{1}{2}\mathbb{E}\left(\|y^{t+1} - y\|^2_{\hat{Q}} - \|y^0 - y\|^2_{\hat{Q}} + \sum_{k=0}^{t}\|y^{k+1} - y^k\|^2_{\hat{Q}}\right) \\ & + \frac{L_g}{2}\sum_{k=0}^{t}\mathbb{E}\|y^k - y^{k+1}\|^2 \\ \overset{(4.41a)}{\leq} \; & \frac{1}{2}\left(\|x^0 - x\|^2_{\hat{P} - \theta\rho_x A^\top A} + \|y^0 - y\|^2_{\hat{Q}}\right) + \frac{\theta}{2\rho}\mathbb{E}\|\lambda^0 - \lambda\|^2. \end{aligned} \tag{4.95}$$

In addition, note that

$$\theta\|x^{t+1}-x\|^2_{A^\top A}=\frac{n}{N}\left\|\sum_{i=1}^N A_i(x_i^{t+1}-x_i)\right\|^2\le n\sum_{i=1}^N\|x_i^{t+1}-x_i\|^2_{A_i^\top A_i}$$
$$\|x^k-x^{k+1}\|^2_{A^\top A}=\left\|\sum_{i\in I_k} A_i(x_i^k-x_i^{k+1})\right\|^2\le n\sum_{i=1}^N\|x_i^k-x_i^{k+1}\|^2_{A_i^\top A_i}$$
$$\|y^k-y^{k+1}\|^2_{B^\top B}=\left\|\sum_{j\in J_k} B_j(y_j^k-y_j^{k+1})\right\|^2\le m\sum_{j=1}^M\|y_j^k-y_j^{k+1}\|_{B_j^\top B_j}.$$

Hence, if $\hat{P}$ and $\hat{Q}$ satisfy (4.41b), then (4.95) also holds.

Combining (4.82), (4.92) and (4.95) yields

$$\begin{aligned}
&\mathbb{E}\left[\Phi(x^{t+1},y^{t+1})-\Phi(x,y)+(\tilde{w}^{t+1}-w)^\top H(\tilde{w}^{t+1})\right]\\
&+\theta\sum_{k=0}^{t-1}\mathbb{E}\left[\Phi(x^{k+1},y^{k+1})-\Phi(x,y)+(w^{k+1}-w)^\top H(w^{k+1})\right]\\
\le\ &(1-\theta)\left[\Phi(x^0,y^0)-\Phi(x,y)\right]\\
&+(1-\theta)\Big[(x^0-x)^\top(-A^\top\lambda^0)+\rho_x(x^0-x)^\top A^\top r^0\\
&+(y^0-y)^\top(-B^\top\lambda^0)+\rho_y(y^0-y)^\top B^\top r^0\Big]\\
&+\frac{1}{2}\left(\|x^0-x\|^2_{\hat{P}-\theta\rho_x A^\top A}+\|y^0-y\|^2_{\hat{Q}}\right)+\frac{\theta}{2\rho}\mathbb{E}\|\lambda^0-\lambda\|^2.
\end{aligned}\tag{4.96}$$

Applying the convexity of Φ and the properties (4.23) of H, we have

$$\begin{aligned}
&(1+\theta t)\mathbb{E}\left[\Phi(\hat{x}^t,\hat{y}^t)-\Phi(x,y)+(\hat{w}^{t+1}-w)^\top H(w)\right]\\
\overset{(4.20)}{=}\ &(1+\theta t)\mathbb{E}\left[\Phi(\hat{x}^t,\hat{y}^t)-\Phi(x,y)+(\hat{w}^{t+1}-w)^\top H(\hat{w}^{t+1})\right]\\
\overset{(4.23)}{\le}\ &\mathbb{E}\left[\Phi(x^{t+1},y^{t+1})-\Phi(x,y)+(\tilde{w}^{t+1}-w)^\top H(\tilde{w}^{t+1})\right]\\
&+\theta\sum_{k=0}^{t-1}\mathbb{E}\left[\Phi(x^{k+1},y^{k+1})-\Phi(x,y)+(w^{k+1}-w)^\top H(w^{k+1})\right].
\end{aligned}\tag{4.97}$$

Now combining (4.97) and (4.96), we have

$$\begin{aligned}
&(1+\theta t)\mathbb{E}\left[\Phi(\hat{x}^t,\hat{y}^t)-\Phi(x,y)+(\hat{w}^{t+1}-w)^\top H(w)\right]\\
\le\ &(1-\theta)\left[\Phi(x^0,y^0)-\Phi(x,y)\right]
\end{aligned}$$

$$
\begin{aligned}
&+(1-\theta)\Big[(x^0-x)^\top(-A^\top\lambda^0)+\rho_x(x^0-x)^\top A^\top r^0\\
&+(y^0-y)^\top(-B^\top\lambda^0)+\rho_y(y^0-y)^\top B^\top r^0\Big]\\
&+\frac{1}{2}\left(\|x^0-x\|^2_{\hat{P}-\theta\rho_x A^\top A}+\|y^0-y\|^2_{\hat{Q}}\right)+\frac{\theta}{2\rho}\mathbb{E}\|\lambda^0-\lambda\|^2. \qquad (4.98)
\end{aligned}
$$

By Lemmas 4.3.3 and 4.3.5 with $\gamma=\max\{0.5+\|\lambda^*\|,3\|\lambda^*\|\}$, we have the desired result.

4.9.4 Proof of Theorem 4.4.2

From the nonincreasing monotonicity of α_k, one can easily show the following result.

Lemma 4.9.1 *Assume $\lambda^{-1}=\lambda^0$. It holds that*

$$
\begin{aligned}
&\sum_{k=0}^{t}\frac{(1-\theta)\beta_k}{2}\left[\|\lambda^k-\lambda\|^2-\|\lambda^{k-1}-\lambda\|^2+\|\lambda^k-\lambda^{k-1}\|^2\right]\\
&-\sum_{k=0}^{t-1}\frac{\alpha_k\beta_{k+1}}{2\alpha_{k+1}}\left[\|\lambda^{k+1}-\lambda\|^2-\|\lambda^k-\lambda\|^2+\|\lambda^{k+1}-\lambda^k\|^2\right]\\
\le\ &-\sum_{k=0}^{t-1}\frac{\beta_{k+1}}{2}\|\lambda^{k+1}-\lambda^k\|^2+\sum_{k=1}^{t}\frac{(1-\theta)\beta_k}{2}\|\lambda^k-\lambda^{k-1}\|^2\\
&-\sum_{k=0}^{t-1}\frac{\alpha_k\beta_{k+1}}{2\alpha_{k+1}}\|\lambda^{k+1}-\lambda\|^2-\sum_{k=1}^{t}\frac{(1-\theta)\beta_k}{2}\|\lambda^{k-1}-\lambda\|^2\\
&+\frac{\alpha_0\beta_1}{2\alpha_1}\|\lambda^0-\lambda\|^2+\sum_{k=1}^{t-1}\frac{\alpha_k\beta_{k+1}}{2\alpha_{k+1}}\|\lambda^k-\lambda\|^2+\sum_{k=1}^{t}\frac{(1-\theta)\beta_k}{2}\|\lambda^k-\lambda\|^2\\
=\ &-\sum_{k=0}^{t-1}\frac{\theta\beta_{k+1}}{2}\|\lambda^{k+1}-\lambda^k\|^2+\left(\frac{\alpha_0\beta_1}{2\alpha_1}-\frac{(1-\theta)\beta_1}{2}\right)\|\lambda^0-\lambda\|^2\\
&-\left(\frac{\alpha_{t-1}\beta_t}{2\alpha_t}-\frac{(1-\theta)\beta_t}{2}\right)\|\lambda^t-\lambda\|^2\\
&-\sum_{k=1}^{t-1}\left(\frac{\alpha_{k-1}\beta_k}{2\alpha_k}+\frac{(1-\theta)\beta_{k+1}}{2}-\frac{\alpha_k\beta_{k+1}}{2\alpha_{k+1}}-\frac{(1-\theta)\beta_k}{2}\right)\|\lambda^k-\lambda\|^2. \qquad (4.99)
\end{aligned}
$$

By the update formula of λ in (4.48), we have from (4.51) that

$$
\mathbb{E}\left[F(x^{k+1})-F(x)+(x^{k+1}-x)^\top(-A^\top\lambda^{k+1})+(\lambda^{k+1}-\lambda)^\top r^{k+1}\right]
$$

$$
\begin{aligned}
&+\mathbb{E}\left[\frac{(\lambda^{k+1}-\lambda)^\top(\lambda^{k+1}-\lambda^k)}{\left(1-\frac{(1-\theta)\alpha_{k+1}}{\alpha_k}\right)\rho}+\frac{(1-\theta)\alpha_{k+1}}{\alpha_k}\rho(x^{k+1}-x)^\top A^\top r^{k+1}\right]\\
&+\mathbb{E}(x^{k+1}-x)^\top\left(\tilde{P}+\frac{I}{\alpha_k}\right)(x^{k+1}-x^k)-\frac{L_f}{2}\mathbb{E}\|x^k-x^{k+1}\|^2+\mathbb{E}(x^{k+1}-x^k)^\top\delta^k\\
\leq\ & (1-\theta)\mathbb{E}\Big[F(x^k)-F(x)+(x^k-x)^\top(-A^\top\lambda^k)\\
&+(\lambda^k-\lambda)^\top r^k+\frac{(\lambda^k-\lambda)^\top(\lambda^k-\lambda^{k-1})}{\left(1-\frac{(1-\theta)\alpha_k}{\alpha_{k-1}}\right)\rho}\Big]\\
&+(1-\theta)\rho\mathbb{E}(x^k-x)^\top A^\top r^k, \qquad (4.100)
\end{aligned}
$$

where similar to (4.77), we have defined $\tilde{P}=\hat{P}-\rho A^\top A$.

Multiplying α_k to both sides of (4.100) and using (4.45) and (4.21), we have

$$
\begin{aligned}
&\alpha_k\mathbb{E}\Big[F(x^{k+1})-F(x)+(w^{k+1}-w)^\top H(w^{k+1})\Big]\\
&+\frac{\alpha_k\beta_{k+1}}{2\alpha_{k+1}}\mathbb{E}\Big[\|\lambda^{k+1}-\lambda\|^2-\|\lambda^k-\lambda\|^2+\|\lambda^{k+1}-\lambda^k\|^2\Big]\\
&+\mathbb{E}\Big[(1-\theta)\alpha_{k+1}\rho(x^{k+1}-x)^\top A^\top r^{k+1}\Big]\\
&+\frac{\alpha_k}{2}\mathbb{E}\big[\|x^{k+1}-x\|_{\tilde{P}}^2-\|x^k-x\|_{\tilde{P}}^2+\|x^{k+1}-x^k\|_{\tilde{P}}^2\big]\\
&+\frac{1}{2}\mathbb{E}\Big[\|x^{k+1}-x\|^2-\|x^k-x\|^2+\|x^{k+1}-x^k\|^2\Big]\\
&-\frac{\alpha_k L_f}{2}\mathbb{E}\|x^k-x^{k+1}\|^2+\alpha_k\mathbb{E}(x^{k+1}-x^k)^\top\delta^k\\
\leq\ & (1-\theta)\alpha_k\mathbb{E}\Big[F(x^k)-F(x)+(w^k-w)^\top H(w^k)\Big]\\
&+\frac{(1-\theta)\beta_k}{2}\mathbb{E}\Big[\|\lambda^k-\lambda\|^2-\|\lambda^{k-1}-\lambda\|^2+\|\lambda^k-\lambda^{k-1}\|^2\Big]\\
&+\alpha_k(1-\theta)\rho\mathbb{E}(x^k-x)^\top A^\top r^k. \qquad (4.101)
\end{aligned}
$$

Denote $\tilde{\lambda}^{t+1}=\lambda^t-\rho r^{t+1}$. Then for $k=t$, it is easy to see that (4.101) becomes

$$
\begin{aligned}
&\alpha_t\mathbb{E}\Big[F(x^{t+1})-F(x)+(\tilde{w}^{t+1}-w)^\top H(\tilde{w}^{t+1})\Big]\\
&+\frac{\alpha_t}{2\rho}\mathbb{E}\Big[\|\tilde{\lambda}^{t+1}-\lambda\|^2-\|\lambda^t-\lambda\|^2+\|\tilde{\lambda}^{t+1}-\lambda^t\|^2\Big]\\
&+\frac{\alpha_t}{2}\mathbb{E}\Big[\|x^{t+1}-x\|_{\tilde{P}}^2-\|x^t-x\|_{\tilde{P}}^2+\|x^{t+1}-x^t\|_{\tilde{P}}^2\Big]\\
&+\frac{1}{2}\mathbb{E}\left[\|x^{t+1}-x\|^2-\|x^t-x\|^2+\|x^{t+1}-x^t\|^2\right]\\
&-\frac{\alpha_t L_f}{2}\mathbb{E}\|x^t-x^{t+1}\|^2+\alpha_t\mathbb{E}(x^{t+1}-x^t)^\top\delta^t
\end{aligned}
$$

$$
\begin{aligned}
&\leq (1-\theta)\alpha_t \mathbb{E}\left[F(x^t) - F(x) + (w^t - w)^\top H(w^t)\right] \\
&\quad + \frac{(1-\theta)\beta_t}{2}\mathbb{E}\left[\|\lambda^t - \lambda\|^2 - \|\lambda^{t-1} - \lambda\|^2 + \|\lambda^t - \lambda^{t-1}\|^2\right] \\
&\quad + \alpha_t(1-\theta)\mathbb{E}\rho(x^t - x)^\top A^\top r^t. \qquad (4.102)
\end{aligned}
$$

By the nonincreasing monotonicity of α_k, summing (4.101) from $k = 0$ through $t-1$ and (4.102) and plugging (4.99) gives

$$
\begin{aligned}
&\alpha_t \mathbb{E}\left[F(x^{t+1}) - F(x) + (\tilde{w}^{t+1} - w)^\top H(\tilde{w}^{t+1})\right] \\
&+\theta\alpha_{k+1}\sum_{k=0}^{t-1}\mathbb{E}\left[F(x^{k+1}) - F(x) + (w^{k+1} - w)^\top H(w^{k+1})\right] \\
&+\frac{\alpha_t}{2\rho}\mathbb{E}\left[\|\tilde{\lambda}^{t+1} - \lambda\|^2 - \|\lambda^t - \lambda\|^2 + \|\tilde{\lambda}^{t+1} - \lambda^t\|^2\right] \\
&+\frac{\alpha_{t+1}}{2}\mathbb{E}\|x^{t+1} - x\|_{\tilde{P}}^2 + \sum_{k=0}^{t}\frac{\alpha_k}{2}\mathbb{E}\|x^{k+1} - x^k\|_{\tilde{P}}^2 \\
&+\frac{1}{2}\mathbb{E}\big[\|x^{t+1} - x\|^2 - \|x^0 - x\|^2 + \sum_{k=0}^{t}\|x^{k+1} - x^k\|^2\big] \\
&-\sum_{k=0}^{t}\frac{\alpha_k L_f}{2}\mathbb{E}\|x^k - x^{k+1}\|^2 + \sum_{k=0}^{t}\alpha_k\mathbb{E}(x^{k+1} - x^k)^\top\delta^k \\
\leq\ &(1-\theta)\alpha_0\mathbb{E}\left[F(x^0) - F(x) + (w^0 - w)^\top H(w^0)\right] \\
&+\alpha_0(1-\theta)\rho(x^0 - x)^\top A^\top r^0 + \frac{\alpha_0}{2}\|x^0 - x\|_{\tilde{P}}^2 \\
&-\sum_{k=0}^{t-1}\frac{\theta\beta_{k+1}}{2}\mathbb{E}\|\lambda^{k+1} - \lambda^k\|^2 + \left(\frac{\alpha_0\beta_1}{2\alpha_1} - \frac{(1-\theta)\beta_1}{2}\right)\mathbb{E}\|\lambda^0 - \lambda\|^2 \\
&-\left(\frac{\alpha_{t-1}\beta_t}{2\alpha_t} - \frac{(1-\theta)\beta_t}{2}\right)\mathbb{E}\|\lambda^t - \lambda\|^2 \\
&-\sum_{k=1}^{t-1}\left(\frac{\alpha_{k-1}\beta_k}{2\alpha_k} + \frac{(1-\theta)\beta_{k+1}}{2} - \frac{\alpha_k\beta_{k+1}}{2\alpha_{k+1}} - \frac{(1-\theta)\beta_k}{2}\right)\mathbb{E}\|\lambda^k - \lambda\|^2.
\end{aligned} \qquad (4.103)
$$

From (4.52d), we have

$$
\frac{\alpha_t}{2\rho}\left[\|\tilde{\lambda}^{t+1} - \lambda\|^2 - \|\lambda^t - \lambda\|^2 + \|\tilde{\lambda}^{t+1} - \lambda^t\|^2\right] \geq -\left(\frac{\alpha_{t-1}\beta_t}{2\alpha_t} - \frac{(1-\theta)\beta_t}{2}\right)\|\lambda^t - \lambda\|^2.
$$

In addition, from Young's inequality, it holds that

$$\frac{1}{2}\|x^{k+1}-x^k\|^2+\alpha_k\mathbb{E}(x^{k+1}-x^k)^\top\delta^k\geq\frac{\alpha_k^2}{2}\|\delta\|^2.$$

Hence, dropping negative terms on the right hand side of (4.103), from the convexity of Φ and (4.23), we have

$$\begin{aligned}
&\left(\alpha_{t+1}+\theta\sum_{k=1}^{t}\alpha_k\right)\mathbb{E}\left[F(\hat{x}^t)-F(x)+(\hat{w}^t-w)^\top H(\hat{w}^t)\right]\\
&\alpha_t\mathbb{E}\left[F(x^{t+1})-F(x)+(\tilde{w}^{t+1}-w)^\top H(\tilde{w}^{t+1})\right]\\
&+\theta\alpha_{k+1}\sum_{k=0}^{t-1}\mathbb{E}\left[F(x^{k+1})-F(x)+(w^{k+1}-w)^\top H(w^{k+1})\right]\\
\leq\;&(1-\theta)\alpha_0\left[F(x^0)-F(x)+(w^0-w)^\top H(w^0)\right]\\
&+(1-\theta)\alpha_0\rho(x^0-x)^\top A^\top r^0+\frac{\alpha_0}{2}\|x^0-x\|_P^2+\frac{1}{2}\|x^0-x\|^2\\
&+\left(\frac{\alpha_0\beta_1}{2\alpha_1}-\frac{(1-\theta)\beta_1}{2}\right)\mathbb{E}\|\lambda^0-\lambda\|^2+\sum_{k=0}^{t}\frac{\alpha_k^2}{2}\mathbb{E}\|\delta^k\|^2. \qquad (4.104)
\end{aligned}$$

Using Lemma 4.3.3 and the properties of H, we derive the desired result.

4.9.5 Proof of Proposition 4.6.1

Let $(I+\partial\phi)^{-1}(x):=\operatorname{argmin}_z\,\phi(z)+\frac{1}{2}\|z-x\|_2^2$ denote the proximal mapping of ϕ at x. Then the update in (4.9b) can be written to

$$z^{k+1}=\left(I+\partial\left(\frac{g^*}{\eta}\right)\right)^{-1}\left(z^k-\frac{1}{\eta}Ax^{k+1}\right).$$

Define y^{k+1} as that in (4.65b). Then

$$\begin{aligned}
\frac{1}{\eta}y^{k+1} &= \frac{1}{\eta}\left\{\operatorname*{argmin}_y g(y) - \langle y, z^k\rangle + \frac{1}{2\eta}\|y + Ax^{k+1}\|^2\right\} \\
&= \frac{1}{\eta}\left\{\operatorname*{argmin}_y g(y) + \frac{\eta}{2}\|\frac{1}{\eta}y - (z^k - \frac{1}{\eta}Ax^{k+1})\|^2\right\} \\
&= \operatorname*{argmin}_y g(\eta y) + \frac{\eta}{2}\|y - (z^k - \frac{1}{\eta}Ax^{k+1})\|^2 \\
&= \left(I + \partial\left(\frac{1}{\eta}g(\eta\cdot)\right)\right)^{-1}\left(z^k - \frac{1}{\eta}Ax^{k+1}\right).
\end{aligned}$$

Hence, using the fact that the conjugate of $\frac{1}{\eta}g^*$ is $\frac{1}{\eta}g(\eta\cdot)$ and the Moreau's identity $(I+\partial\phi)^{-1} + (I+\partial\phi^*)^{-1} = I$ for any convex function ϕ, we have

$$\begin{aligned}
& z^k - \frac{1}{\eta}Ax^{k+1} \\
= \quad & \left(I + \partial\left(\frac{g^*}{\eta}\right)\right)^{-1}\left(z^k - \frac{1}{\eta}Ax^{k+1}\right) + \left(I + \partial\left(\frac{1}{\eta}g(\eta\cdot)\right)\right)^{-1}\left(z^k - \frac{1}{\eta}Ax^{k+1}\right).
\end{aligned}$$

Therefore, (4.65c) holds, and thus from (4.9c) it follows

$$\bar{z}^{k+1} = z^{k+1} - \frac{q}{\eta}(Ax^{k+1} + y^{k+1}).$$

Substituting the formula of $\bar{z}^k$ into (4.9a), we have for $i = i_k$,

$$\begin{aligned}
x_i^{k+1} &= \operatorname*{argmin}_{x_i\in X_i}\langle -\bar{z}^k, A_i x_i\rangle + u_i(x_i) + \frac{\tau}{2}\|x_i - x_i^k\|^2 \\
&= \operatorname*{argmin}_{x_i\in X_i}\langle -z^k, A_i x_i\rangle + \frac{q}{\eta}\langle Ax^k + y^k, A_i x_i\rangle + u_i(x_i) + \frac{\tau}{2}\|x_i - x_i^k\|^2 \\
&= \operatorname*{argmin}_{x_i\in X_i}\langle -z^k, A_i x_i\rangle + u_i(x_i) + \frac{q}{2\eta}\|A_i x_i + A_{\neq i}x_{\neq i}^k + y^k\|^2 + \frac{1}{2}\|x_i - x_i^k\|_{\tau I - \frac{q}{\eta}A_i^\top A_i},
\end{aligned}$$

which is exactly (4.65a). Hence, we complete the proof.

Chapter 5

Zeroth-Order Algorithms for Black-Box Optimization

5.1 Introduction

In this chapter, we consider a black-box optimization in the form of

$$\begin{array}{lll} (P) & \min & f(x) + h(x) \\ & \text{s.t.} & x \in \mathcal{X}, \end{array} \tag{5.1}$$

where $\mathcal{X}$ is a closed convex set, $h(x)$ is a regularization function, which is typically non-smooth. The key feature of this model however, is that the exact formulation of $f(x)$ is unknowable. Instead, only some noisy estimation of $f(x)$ is possible. This rules out any high-order solution methods, leaving only zeroth-order methods as a solution choice. In this context, we consider two settings. In the first setting, $f(x)$ is an expectation of $F(x, \xi)$ with unknown distribution ξ. Unlike the usual stochastic programming model, where the classical Sample Averaging Approximation (SAA) is applicable (cf. [83]), we assume here that for each given x only one sample point can be collected for $F(x, \xi)$. This is the case, for instance, when sampling is *information sensitive*. This is particularly relevant for design problems, in which one cannot copy exactly the same design and test the responses without worrying that the responses are influenced by the previous sampling events. As an example, designing questions for standard tests such as SAT or ACT is information sensitive, for there can be no two exams that contain exactly the same questions. In fact, this setting resembles the bandit online learning model,

where one can obtain one feedback at a given point. In a second setting, the objective is not necessarily stochastic. In fact, it may not be stochastic at all. However, the objective function maybe expensive to evaluate. Examples of such optimization model include the design problems where the design variables are initial and/or boundary conditions of a differential equation, and the objective value depends on the solution of the differentiable equation. Given the initial/boundary conditions, the evaluation of the objective function reduces to solving a differential equation. Therefore, one can only evaluate the quality of the design variables approximately. However, the solution can be made arbitrarily precise if one is willing to invest more time and effort.

In the above described black-box optimization models, no higher order information is possible. We are left with some approximative zeroth-order subroutines. This chapter sets out to explore the convergence rate of zeroth-order algorithms, assuming the objective to have convexity or convexity-like (e.g. star-convexity and weak pseudo-convexity) property. The analysis is then extended to the setting where some regularization function is included in the objective. This leads to exploring zeroth-order proximal-gradient type solution procedures. Our emphasis is placed on analyzing the overall sample complexity, which essentially means the total amount of 'efforts', in order to reach an ϵ-optimal solution.

This chapter is organized as follows. In Section 5.2, we study an unconstrained stochastic optimization model where the objective can be allowed a single-sample at a point. The convergence study also extends to the star-convex functions. In Section 5.3, we study a class of nonconvex optimization model by introducing the so-called weak pseudo-convexity. For this model, we develop a zeroth-order normalized gradient descent method. In Section 5.4, we study unconstrained optimization where the objective function can only be estimated. Moreover, the efforts required to estimate the function value depends on the precision. Under the convexity assumption, sample complexity bounds (to reach an ϵ-optimal solution) are derived for the zeroth-order methods based on the coordinate-gradient method and the ellipsoid method respectively. In Section 5.5 we extend our investigations to the constrained optimization with a regularization function. Sample complexities are derived for all the afore-mentioned methods. Finally, we present the numerical experiments by comparing with the Bayesian optimization on two practical problems.

5.2 Stochastic Programming: One Sample at a Point

In this section, we consider model (5.1) where the regularization term h does not exist, and the function f is of the following form

$$f(x) = \mathsf{E}[F(x, \xi)], \tag{5.2}$$

where the expectation is taken over the random vector ξ. For a query of the function value, a sample $F(x, \xi)$ is revealed. Furthermore, we assume that only one single sample is possible for every query x. In other words, for queries of the function value at x_1 and x_2, it returns two samples $F(x_1, \xi^1)$ and $F(x_2, \xi^2)$ with different realizations of the random vectors ξ^1 and ξ^2. For this stochastic optimization framework, we first present some assumptions of $F(x, \xi)$ as well as some definitions.

Assumption 5.2.1 *Suppose the function $f(x)$ is given in the form (5.2), then we assume that $F(x, \xi)$ satisfies*

$$\mathsf{E}[F(x, \xi)] = f(x), \tag{5.3}$$

$$\mathsf{E}[\nabla F(x, \xi)] = \nabla f(x), \tag{5.4}$$

and

$$\mathsf{E}[\|F(x, \xi) - f(x)\|^2] \leq \theta_0^2, \tag{5.5}$$

$$\mathsf{E}[\|\nabla F(x, \xi) - \nabla f(x)\|^2] \leq \theta_1^2. \tag{5.6}$$

Definition 7 *We denote $C^i_{L_i}(\mathcal{D})$ be the class of functions which are i-th order Lipschitz continuous in the domain $\mathcal{D}$ with the corresponding Lipschitz constant L_i, namely*

$$C^i_{L_i}(\mathcal{D}) = \{f \mid \|\nabla^{(i)} f(x) - \nabla^{(i)} f(y)\| \leq L_i \|x - y\|, \forall x, y \in \mathcal{D}\}.$$

There are two cases that are of special interest: $i = 0$ and $i = 1$. We simply denote L_1 by L. In other words, $C^1_L(\mathcal{D})$ is the class of functions of Lipschitz continuous gradients, i.e. for $f \in C^1_L(\mathcal{D})$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \ \ \forall x, y \in \mathcal{D}. \tag{5.7}$$

Now, we introduce a smoothing scheme and its properties which lead to the stochastic zeroth-order oracle ($\mathcal{SZO}$).

Definition 8 *Let U_b be the uniform distribution over the unit Euclidean ball and B be the unit ball. Given $\mu > 0$, the smoothing function f^μ is defined as*

$$f^\mu(w) = \mathsf{E}_{\{v\sim U_b\}}[f(w+\mu v)] = \frac{1}{\alpha(n)}\int_B f(w+\mu v)dv \tag{5.8}$$

where $\alpha(n)$ is the volume of the unit ball in $\mathbb{R}^n$.

Some properties of the smoothing function are shown in the lemma below, the proof of the lemma can be found in [38].

Lemma 5.2.1 *Suppose that $f \in C_L^1(\mathbb{R}^n)$. Let U_{S_p} be the uniform distribution over the unit Euclidean sphere, and S_p be the unit sphere in $\mathbb{R}^n$. Then we have:*

(a) The smoothing function f^μ is continuously differentiable, and its gradient is Lipschitz continuous with constant $L_\mu \le L$ and

$$\nabla f^\mu(w) = \mathsf{E}_{\{v\sim U_{S_p}\}}\left[\frac{n}{\mu}f(w+\mu v)v\right] = \frac{1}{\beta(n)}\int_{v\in S_p}\frac{n}{\mu}\left[f(w+\mu v)-f(w)\right]vdv \tag{5.9}$$

where $\beta(n)$ is the measure of the unit sphere in $\mathbb{R}^n$.

(b) For any $w \in \mathbb{R}^n$, we have

$$|f^\mu(w) - f(w)| \le \frac{L\mu^2}{2}, \tag{5.10}$$

$$\|\nabla f^\mu(w) - \nabla f(w)\| \le \frac{\mu n L}{2}, \tag{5.11}$$

$$\mathsf{E}_v\left[\left\|\frac{n}{\mu}[f(w+\mu v)-f(w)]v\right\|^2\right] \le 2n\|\nabla f(w)\|^2 + \frac{\mu^2}{2}L^2n^2. \tag{5.12}$$

(c) If f is convex then f^μ is convex.

Based on (5.9) we introduce a single-sampling zeroth-order stochastic gradient ($\mathcal{SSZO}$) of f at x:

$$G_\mu(x,\xi^{1,2},v) := \frac{n}{\mu}\left[F(x+\mu v,\xi^1) - F(x,\xi^2)\right]v, \tag{5.13}$$

where v is the random vector uniformly distributed over the unit sphere in $\mathbb{R}^n$. Note that ξ^1 and ξ^2 are i.i.d. samples.

The following lemma shows the unbiasedness and boundedness of the $\mathcal{SSZO}$.

Lemma 5.2.2 *Suppose that $G_\mu(x, \xi^{1,2}, v)$ is defined as in (5.13), and f satisfies Assumption 5.2.1. Then*

$$\mathsf{E}_{v,\xi}[G_\mu(x, \xi^{1,2}, v)] = \nabla f^\mu(x). \tag{5.14}$$

If we further assume $f \in C^0_{L_0}(\mathcal{X})$ and $F(x, \xi) \in C^1_L(\mathbb{R}^n)$ for all ξ, then the following holds

$$\mathsf{E}_{v,\xi}[\|G_\mu(x, \xi^{1,2}, v)\|^2] \le 4nN + \mu^2 L^2 n^2 + 4\frac{n^2}{\mu^2}\theta_0^2, \tag{5.15}$$

where $N = L_0^2 + \theta_1^2$.

Proof. The first equation is easy to verify. We prove the second inequality. Applying (5.12) and (5.63) to $F(x, \xi)$, we have

$$\begin{aligned}
& \mathsf{E}_{v,\xi^{1,2}}\left[\|G_\mu(x, \xi^{1,2}, v)\|^2\right] \\
= \;& \mathsf{E}_{\xi^{1,2}}\left[\mathsf{E}_v\left[\|G_\mu(x, \xi^{1,2}, v)\|^2\right]\right] \\
= \;& \mathsf{E}_{\xi^{1,2}}\left[\mathsf{E}_v\left[\left\|\frac{n}{\mu}\left[F(x+\mu v, \xi^1) - F(x, \xi^2)\right] v\right\|^2\right]\right] \\
\le \;& \mathsf{E}_{\xi^{1,2}}\left[\mathsf{E}_v\left[2\left\|\frac{n}{\mu}\left[F(x+\mu v, \xi^1) - F(x, \xi^1)\right] v\right\|^2\right]\right. \\
& \left. +\mathsf{E}_v\left[2\left\|\frac{n}{\mu}\left[F(x, \xi^1) - F(x, \xi^2)\right] v\right\|^2\right]\right] \\
\overset{(5.12)}{\le} \;& 4n\left[\mathsf{E}_{\xi^1}\left[\|\nabla F(x, \xi)\|^2\right]\right] + \mu^2 L^2 n^2 + 2\frac{n^2}{\mu^2}\mathsf{E}_{\xi^{1,2}}\left[|F(x, \xi^1) - F(x, \xi^2)|^2\right] \\
\le \;& 4n\left\{\mathsf{E}_\xi[\|\nabla f(x)\|^2] + \mathsf{E}_\xi\left[\|\nabla F(x, \xi) - \nabla f(x)\|^2\right]\right\} + \mu^2 L^2 n^2 + 4\frac{n^2}{\mu^2}\theta_0^2 \\
\le \;& 4n\left(\|\nabla f(x)\|^2 + \theta_1^2\right) + \mu^2 L^2 n^2 + 4\frac{n^2}{\mu^2}\theta_0^2 \\
\le \;& 4nN + \mu^2 L^2 n^2 + 4\frac{n^2}{\mu^2}\theta_0^2.
\end{aligned} \tag{5.16}$$

□

5.2.1 Convex Optimization

In this subsection, we apply a single-sampling approach to stochastic optimization problem. The solution framework is depicted as follows, where $\mathrm{Proj}_{\mathcal{X}}(x) := \operatorname*{argmin}_{y \in \mathcal{X}} \|y - x\|$.

Stochastic Single-Sampling Zeroth-Order Gradient Descent

Parameters: $\eta, \mu > 0$ and a convex set $\mathcal{X} \subset \mathbb{R}^n$.

Initialization: $x_1 = 0$.

for $t = 1, \cdots, T$,

Pick $v^k \sim U_{S_p}$;

At $x^k + \delta v^k$ and x^k, receive $F(x^k + \mu v^k, \xi_k^1), F(x^k, \xi_k^2)$ respectively;

Assemble $\mathcal{SSZO}$ as $G_\mu(x^k, \xi_k^{1,2}, v^k) = \frac{n}{\mu}\left[F(x^k + \delta v^k, \xi_k^1) - F(x^k, \xi_k^2)\right] v^k$;

Update $x_{t+1} = \mathrm{Proj}_{\mathcal{X}}\left(x^k - \eta G_\mu(x^k, \xi_k^{1,2}, v^k)\right)$.

end for

The following theorem shows an expected $O(T^{-1/3})$ rate of convergence for the stochastic single-sampling zeroth-order gradient descent algorithm. The expectation is taken over the σ-field generated by the random variables $\{\xi_k^{1,2}, v^k\}_{k=1}^T$.

Theorem 5.2.3 *Suppose $f(x)$ is convex and $f \in C_{L_0}^0(\mathbb{R}^n)$ satisfies Assumption 5.2.1, $F(x, \xi) \in C_L^1(\mathbb{R}^n)$ for all ξ. Let $\{x^k\}$ be the sequence produced by the stochastic single-sampling zeroth-order gradient descent algorithm. Furthermore, we define an averaging sequence as*

$$\bar{x}_T = \frac{1}{T}\sum_{k=1}^{T} x^k. \tag{5.17}$$

Then, the following inequality holds

$$\mathsf{E}\left[f(\bar{x}_T) - f(x^*)\right] \leq \frac{\|x^1 - x^*\|^2}{2\eta T} + \frac{\eta}{2}\left(4nN + \mu^2 L^2 n^2 + 4\frac{n^2}{\mu^2}\theta_0^2\right) + L\mu^2. \tag{5.18}$$

In particular, if we set $\eta = T^{-2/3}$, and $\mu = T^{-1/6}$, then an $O(T^{-1/3})$ rate of convergence follows.

Proof. Let $z_k := \|x^k - x^*\|$. Then

$$\begin{aligned} z_{k+1}^2 &= \|x^{k+1} - x^*\|^2 \\ &= \left\|\mathrm{Proj}_{\mathcal{X}}\left(x^k - \eta G_\mu(x^k, \xi_k^{1,2}, v^k)\right) - x^*\right\|^2 \\ &\leq \left\|x^k - \eta G_\mu(x^k, \xi_k^{1,2}, v^k) - x^*\right\|^2 \\ &= z_k^2 + \eta^2 \|G_\mu(x^k, \xi_k^{1,2}, v^k)\|^2 - 2\eta G_\mu(x^k, \xi_k^{1,2}, v^k)^\top (x^k - x^*). \end{aligned}$$

Equivalently, we have

$$G_\mu(x^k, \xi_k^{1,2}, v^k)^\top (x^k - x^*) \leq \frac{1}{2\eta}\left(z_k^2 - z_{k+1}^2\right) + \frac{\eta}{2}\|G_\mu(x^k, \xi_k^{1,2}, v^k)\|^2. \tag{5.19}$$

Notice that $\mathsf{E}_{\xi_k^{1,2}, v^k}\left[G_\mu(x^k, \xi_k^{1,2}, v^k) \mid x^k\right] = \nabla f^\mu(x^k)$ which is shown in (5.14), we have

$$\nabla f^\mu(x^k)^\top (x^k - x^*) \leq \frac{1}{2\eta}\left(\mathsf{E}[z_k^2 \mid x^k] - \mathsf{E}[z_{k+1}^2 \mid x^k]\right) + \frac{\eta}{2}\|G_\mu(x^k, \xi_k^{1,2}, v^k)\|^2.$$

Summing up the above inequalities and using the convexity of f^μ, we have

$$\mathsf{E}\left[\sum_{k=1}^{T}(f^\mu(x^k) - f^\mu(x^*))\right] \leq \frac{1}{2\eta}\|z_1\|_2^2 + \frac{\eta}{2}\sum_{k=1}^{T}\mathsf{E}\left[\|G_\mu(x^k, \xi_k^{1,2}, v^k)\|^2\right]. \tag{5.20}$$

By (5.10), we have

$$f(x^k) - f(x^*) \leq f^\mu(x^k) - f^\mu(x^*) + L\mu^2. \tag{5.21}$$

Combining (5.21) and (5.20) leads to

$$\mathsf{E}\left[\sum_{k=1}^{T}(f(x^k) - f(x^*))\right] \leq \frac{1}{2\eta}\|z_1\|_2^2 + \frac{\eta}{2}\sum_{k=1}^{T}\mathsf{E}\left[\|G_\mu(x^k, \xi_k^{1,2}, v^k)\|^2\right] + L\mu^2 T. \tag{5.22}$$

Based on (5.15), we can further bound $\sum_{k=1}^{T}\mathsf{E}\left[\|G_\mu(x^k, \xi_k^{1,2}, v^k)\|^2\right]$ as

$$\sum_{k=1}^{T}\mathsf{E}\left[\|G_\mu(x^k, \xi_k^{1,2}, v^k)\|^2\right] \leq \left(4nN + \mu^2 L^2 n^2 + 4\frac{n^2}{\mu^2}\theta_0^2\right) T.$$

Due to the convexity of $f(x)$, we have

$$\begin{aligned} & \mathsf{E}\left[f(\bar{x}_T) - f(x^*)\right] \\ \leq\ & \mathsf{E}\left[\frac{1}{T}\sum_{k=1}^{T}(f(x^k) - f(x^*))\right] \\ \leq\ & \frac{\|z_1\|^2}{2\eta T} + \frac{\eta}{2}\left(4nN + \mu^2 L^2 n^2 + 4\frac{n^2}{\mu^2}\theta_0^2\right) + L\mu^2. \end{aligned} \tag{5.23}$$

If we set $\eta = T^{-2/3}$, and $\mu = T^{-1/6}$, then an $O(T^{-1/3})$ rate of convergence follows. □

5.2.2 Optimization with Star-Convexity

Definition 9 *(Star-convex functions). A function $f : \mathbb{R}^n \to \mathbb{R}$ is star-convex if there is $x^* \in \operatorname{argmin}_{x \in \mathcal{X}} f(x)$ such that for all $\alpha \in [0, 1]$ and $x \in \mathcal{X}$,*

$$f((1-\alpha)x^* + \alpha x) \leq (1-\alpha)f(x^*) + \alpha f(x). \tag{5.24}$$

The following lemma characterizes the differentiable star-convex functions.

Lemma 5.2.4 *For a differentiable function f, the star convexity condition* (5.24) *is equivalent to the following condition*

$$f(x) - f(x^*) \leq \nabla f(x)^\top (x - x^*), \tag{5.25}$$

where $x^* = \operatorname{argmin}_{x \in \mathcal{X}} f(x)$.

Proof. Suppose (5.24) holds, then we have

$$f(x) - f(x^*) \leq \frac{f(x) - f((1-\alpha)x^* + \alpha x)}{1-\alpha}, \tag{5.26}$$

for all $\alpha \in [0, 1]$. Note that

$$\lim_{\alpha \to 1^-} \frac{f(x) - f((1-\alpha)x^* + \alpha x)}{1-\alpha} = \nabla f(x)^\top (x - x^*),$$

which implies (5.25). Conversely, suppose that (5.25) holds. Let us denote

$$d(\alpha) := f((1-\alpha)x^* + \alpha x) - f(x^*).$$

Clearly, (5.24) is equivalent to

$$d(\alpha) \leq \alpha d(1), \text{ for all } 0 \leq \alpha \leq 1. \tag{5.27}$$

It remains to show that if f is differentiable then (5.25) implies (5.27). In fact, (5.25)

leads to

$$f((1-\alpha)x^* + \alpha x) - f(x^*) \leq \alpha \nabla f((1-\alpha)x^* + \alpha x)^\top (x - x^*),$$

or,

$$d(\alpha) \leq \alpha d'(\alpha).$$

Hence,

$$\left(\frac{d(\alpha)}{\alpha}\right)' = \frac{\alpha d'(\alpha) - d(\alpha)}{\alpha^2} \geq 0,$$

for all $0 < \alpha \leq 1$, implying that $\frac{d(\alpha)}{\alpha}$ is a nondecreasing function for $\alpha \in (0, 1]$. Therefore,

$$\frac{d(\alpha)}{\alpha} \leq \frac{d(1)}{1},$$

which proves (5.27) for $\alpha \in (0, 1]$. Since $d(0) = f(x^*) = 0$, (5.27) in fact holds for all $\alpha \in [0, 1]$. □

Theorem 5.2.5 *Suppose $f(x)$ is star-convex and $f(x) \in C_L^1(\mathbb{R}^n)$ satisfies Assumption 5.2.1, $F(x, \xi) \in C_L^1(\mathbb{R}^n)$ for all ξ. Let $\{x^k\}$ be the sequence produced by the stochastic single-sampling zeroth-order gradient descent algorithm. Furthermore, after T iteration, we define a random output $\hat{x}_T$ as follows*

$$\text{Prob}\left(\hat{x}_T = x^k\right) = \frac{1}{T}, \quad \textit{for } k = 1, 2, \ldots, T. \tag{5.28}$$

Then, the following inequality holds

$$\mathsf{E}\left[f(\hat{x}_T) - f(x^*)\right] \leq \frac{\|x^1 - x^*\|^2}{2\eta T} + \frac{\eta}{2}\left(4nN + \mu^2 L^2 n^2 + 4\frac{n^2}{\mu^2}\theta_0^2\right) + \mu n L R, \tag{5.29}$$

where $\mathcal{X}$ is assumed to be bounded with $\sup_{x \in \mathcal{X}} \|x\| \leq R$. In particular, if we set $\eta = T^{-3/4}$, and $\mu = T^{-1/4}$, then we have an $O(T^{-1/4})$ rate of convergence for $\mathsf{E}\left[f(\hat{x}_T) - f(x^*)\right]$.

Proof. Let $z_k := \|x^k - x^*\|$. Similar to the proof of Theorem 5.3.7, we have

$$G_\mu(x^k, \xi_k^{1,2}, v^k)^\top (x^k - x^*) \leq \frac{1}{2\eta}\left(z_k^2 - z_{k+1}^2\right) + \frac{\eta}{2}\|G_\mu(x^k, \xi_k^{1,2}, v^k)\|^2. \tag{5.30}$$

Notice that $\mathsf{E}_{\xi_k^{1,2},v^k}[G_\mu(x^k,\xi_k^{1,2},v^k)\mid x^k]=\nabla f^\mu(x^k)$ which is shown in (5.14), we have

$$\nabla f^\mu(x^k)^\top(x^k-x^*)\le\frac{1}{2\eta}\left(\mathsf{E}[z_k^2\mid x^k]-\mathsf{E}[z_{k+1}^2\mid x^k]\right)+\frac{\eta}{2}\|G_\mu(x^k,\xi_k^{1,2},v_k)\|^2.$$

Based on (5.11), it follows

$$\nabla f(x^k)^\top(x^k-x^*)\le\frac{1}{2\eta}\left(\mathsf{E}[z_k^2\mid x^k]-\mathsf{E}[z_{k+1}^2\mid x^k]\right)+\frac{\eta}{2}\|G_\mu(x^k,\xi_k^{1,2},v^k)\|^2+\frac{\mu nL}{2}\|x^k-x^*\|.$$

Summing up the above inequalities and recall the weak star convexity of f, we have

$$\begin{aligned}&\mathsf{E}\left[\sum_{k=1}^T(f(x^k)-f(x^*))\right]\\ \le\quad&\frac{1}{2\eta}\|z_1\|_2^2+\frac{\eta}{2}\sum_{k=1}^T\mathsf{E}\left[\|G_\mu(x^k,\xi_k^{1,2},v^k)\|^2\right]+\frac{\mu nL}{2}\sum_{k=1}^T\mathsf{E}\left[\|x^k-x^*\|\right].\end{aligned}\tag{5.31}$$

By the boundedness of $\mathcal{X}$, we have $\|x^k-x^*\|\le 2R$. Thus,

$$\mathsf{E}\left[\sum_{k=1}^T(f(x^k)-f(x^*))\right]\le\frac{1}{2\eta}\|z_1\|_2^2+\frac{\eta}{2}\sum_{k=1}^T\mathsf{E}\left[\|G_\mu(x^k,\xi_k^{1,2},v^k)\|^2\right]+\mu nLRT.\tag{5.32}$$

Based on (5.15), we can bound $\sum\limits_{k=1}^T\mathsf{E}\left[\|G_\mu(x^k,\xi_k^{1,2},v^k)\|^2\right]$ as

$$\sum_{k=1}^T\mathsf{E}\left[\|G_\mu(x^k,\xi_k^{1,2},v^k)\|^2\right]\le\left(4nN+\mu^2L^2n^2+4\frac{n^2}{\mu^2}\theta_0^2\right)T.$$

Base on the definition of $\hat{x}_T$, we have

$$\begin{aligned}&\mathsf{E}\left[f(\hat{x}_T)-f(x^*)\right]\\ =\quad&\mathsf{E}\left[\frac{1}{T}\sum_{k=1}^T(f(x^k)-f(x^*))\right]\\ \le\quad&\frac{\|z_1\|^2}{2\eta T}+\frac{\eta}{2}\left(4nN+\mu^2L^2n^2+4\frac{n^2}{\mu^2}\theta_0^2\right)+\mu nLR.\end{aligned}\tag{5.33}$$

Setting $\eta=T^{-3/4}$, and $\mu=T^{-1/4}$, an $O(T^{-1/4})$ convergence rate in expectation follows. □

5.3 Optimization with Weakly Pseudo-Convex Objective

In this section, we introduce a notion of weak pseudo-convexity (WPC) which further generalizes the star-convexity. Despite the similarity to the so-called strictly locally quasi-convexity (SLQC) in [50], the WPC is in fact a weaker assumption. Our algorithm is also based on the normalized gradient descent method. In particular, the key of our zeroth-order algorithm is to build a novel estimation of the normalized gradient.

5.3.1 Problem Setup

We consider the following form of the problem (5.1)

$$\min_{x\in\mathcal{X}} f(x) \tag{5.34}$$

where $\mathcal{X} \subset \mathbb{R}^n$ is a bounded convex set i.e., there exists $R > 0$ such that $\|x\| \leq R$ for all $x \in \mathcal{X}$, and $f \in C_L^1(\mathcal{X})$. In addition, in this section, we present the following definitions regarding the function f.

Definition 10 (Bounded Gradient) *A function $f(\cdot)$ is said to have bounded gradient if there exists a finite positive value M such that for all $x \in \mathcal{X}$, it holds that $\|\nabla f(x)\| \leq M$.*

Note that if $f(\cdot)$ has bounded gradient, then it is also Lipschitz continuous with Lipschitz constant M on the set $\mathcal{X}$.

Definition 11 (Weak Pseudo-Convexity) *A function $f(\cdot)$ is weakly pseudo convex (WPC) if there exists $K > 0$ such that*

$$f(x) - f(x^*) \leq K \frac{\nabla f(x)^\top (x - x^*)}{\|\nabla f(x)\|},$$

holds for all $x \in \mathcal{X}$, with the convention that $\frac{\nabla f(x)}{\|\nabla f(x)\|} = 0$ if $\nabla f(x) = 0$, where x^* is one optimal solution, i.e., $x^* \in \arg\min_{x\in\mathcal{X}} f(x)$.

Here we discuss some implications of the weak pseudo-convexity. If a differentiable function $f(\cdot)$ is Lipschitz continuous and pseudo-convex, then we have (see similar derivation in [87])

$$f(x) - f(y) \leq M \frac{\nabla f(x)^\top (x - y)}{\|\nabla f(x)\|},$$

for all y, x with $f(x) \geq f(y)$, where M is Lipschitz constant. Therefore, we can simply let $K = M$, and the function is also weakly pseudo-convex. Moreover, as another example, the star-convex function proposed by [91] is weakly pseudo-convex.

Proposition 5.3.1 *If $f(\cdot)$ is star-convex and smooth with bounded gradient in $\mathcal{X}$, then $f(\cdot)$ is weakly pseudo-convex.*

In light of Lemma 5.2.4, the proposition is obvious. We next introduce a property that is essentially the same as the SLQC property introduced in [50].

Definition 12 (Acute Angle) *Gradient of $f(\cdot)$ is said to satisfy the **acute angle** condition if there exists a positive value Z such that*

$$\begin{aligned} \cos(\nabla f(x), x - x^*) &= \frac{\nabla f(x)^\top (x - x^*)}{\|\nabla f(x)\| \cdot \|x - x^*\|} \\ &\geq Z > 0, \end{aligned}$$

holds for all $x \in \mathcal{X}$, with the convention that $\frac{\nabla f(x)}{\|\nabla f(x)\|} = 0$ if $\nabla f(x) = 0$, where x^* is one optimal solution, i.e., $x^* \in \arg\min_{x \in \mathcal{X}} f(x)$.

The following proposition shows that the acute angle condition together with the Lipschitz continuity implies the weak pseudo-convexity.

Proposition 5.3.2 *If $f(\cdot)$ has bounded gradient and satisfies the acute angle condition, then $f(\cdot)$ is weakly pseudo-convex.*

The proof of Proposition 5.3.2 is straightforward, hence we omit it here. The class of weakly pseudo-convex functions certainly go beyond the acute angle condition. For example, below is another class of functions satisfying the WPC.

Proposition 5.3.3 *If $f(\cdot)$ has bounded gradient and satisfy the α-homogeneity with respect to its minimum, i.e., there exists $\alpha > 0$ satisfying*

$$f(t(x - x^*) + x^*) - f(x^*) = t^\alpha (f(x) - f(x^*)),$$

for all $x \in \mathcal{X}$ and $t \geq 0$ where $x^* = \arg\min_{x \in \mathcal{X}} f(x)$, then $f(\cdot)$ is weak pseudo-convex.

Proof. By taking the derivative of the equation (5.3.3) with respective to t and letting $t = 1$, we have

$$\nabla f(x)^\top (x - x^*) = \alpha (f(x) - f(x^*)).$$

Therefore, we have

$$\begin{aligned} f(x) - f(x^*) &= \frac{1}{\alpha}\nabla f(x)^\top (x - x^*) \\ &\leq \frac{M}{\alpha}\frac{\nabla f(x)^\top (x - x^*)}{\|\nabla f(x)\|}, \end{aligned}$$

which satisfies the weak pseudo-convexity condition with $K = \frac{M}{\alpha}$. □

Proposition 5.3.3 suggests that all non-negative homogeneous polynomial satisfies WPC with respect to 0. Take $f(x) = (x_1^2+x_2^2)^2+10(x_1^2-x_2^2)^2$ as an example. It is easy to verify that $f(\cdot)$ satisfies the condition in Proposition 5.3.3, and thus is weakly pseudo-convex. In Figure 5.1, the curvature of $f(x)$ and a sub-level set of this function are plotted. The function is not quasi-convex since the sub-level set is non-convex. However, this function satisfies the acute-angle condition in 12.

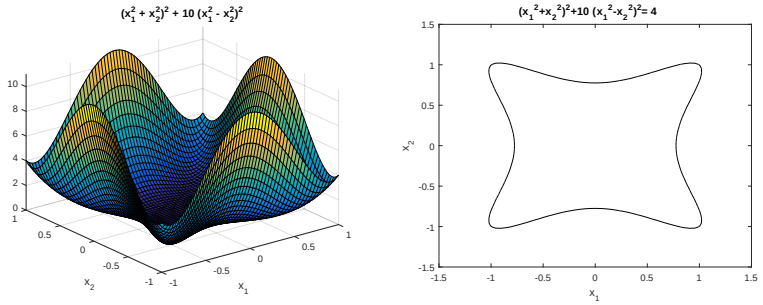

Figure 5.1: Plot of a WPC function that is not quasi-convex.

Note that if $f_i(x)$ is α_i-homogeneous with respect to the shared minimum x^* for all $1 \leq i \leq I$ with $\alpha_i \geq \alpha > 0$, and the gradient of f_i is uniformly bounded over a set $\mathcal{X}$, then $\sum_{i=1}^{I} f_i(x)$ is WPC. As a result, we can construct functions that are WPC but do not satisfy the acute-angle condition. Consider a two-dimensional function $f(x) = x_1^2 + |x_2|^{3/2}$, and suppose that $\mathcal{X}$ is the unit disc centered at the origin. Clearly, $f(x)$ is differentiable and Lipchitz continuous in $\mathcal{X}$. Also, it is the sum of a 2-homogeneous function and a 3/2-homogeneous function with a shared minimum $(0, 0)$. Thus $f(x)$ is

WPC. We compute that

$$\begin{aligned}\cos(\nabla f(x), x - x^*) &= \frac{\nabla f(x)^\top (x - x^*)}{\|\nabla f(x)\| \cdot \|x - x^*\|} \\ &= \frac{2x_1^2 + \frac{3}{2}|x_2|^{3/2}}{\sqrt{(4x_1^2 + \frac{9}{4}|x_2|)(x_1^2 + x_2^2)}}.\end{aligned}$$

Consider a parameterized path $(x_1, x_2) = (t^{1/2}, t^{2/3})$ with $t > 0$. On this path, we have

$$\begin{aligned}\cos(\nabla f(x), x - x^*) &= \frac{2x_1^2 + \frac{3}{2}|x_2|^{3/2}}{\sqrt{(4x_1^2 + \frac{9}{4}|x_2|)(x_1^2 + x_2^2)}} \\ &= \frac{7t}{2\sqrt{(4t + \frac{9}{4}t^{2/3})(t + t^{4/3})}} \\ &= \frac{7t^{1/6}}{2\sqrt{(4t^{1/3} + \frac{9}{4})(1 + t^{1/3})}}.\end{aligned}$$

Therefore, along the path, as t approaches to 0, we have $\cos(\nabla f(x), x - x^*) \to 0$. This example shows that a WPC function may fail to satisfy the acute angle condition.

Definition 13 (Error Bound) *There exists $D > 0$ and $0 < \gamma \leq 1$ such that*

$$\|x - x^*\| \leq D\|\nabla f(x)\|^\gamma,$$

for all $x \in \mathcal{X}$, where x^* is the optimal solution to $f(x)$, i.e., $x^* = \arg\min_{x \in \mathcal{X}} f(x)$.

Since $\mathcal{X}$ is a compact set, the error bound condition is essentially the requirement for a unique optimal solution and no local minimum. We further introduce some notations that will be used in subsequent analysis.

- $S(n)$: the unit sphere in $\mathbb{R}^n$;
- $m(A)$: the measure of set $A \subset \mathbb{R}^n$;
- β_n: the area of the unit sphere $S(n)$;
- dS_n: the differential unit on the unit sphere $S(n)$;
- $\mathbf{1}_A(x)$: the indicator function of set A;
- $\text{sign}(\cdot)$: the sign function.

5.3.2 The Zeroth-Order Normalized Gradient Descent

In this part, we assume that only the function value information $f(x)$ is available for a given query point x. In zeroth-order setting, the main technique we used so far is to construct a zeroth-order approximation of the gradient of a smoothed function. That smoothed function is often created by integrating the original loss function with a chosen probability distribution. By querying some random samples of the function value according to a probability distribution, the player is able to create an unbiased zeroth-order approximation of the gradient of the smoothed function. This is, however, not applicable in our normalized gradient descent algorithm since what we need is the direction of the gradient. Therefore, we shall first develop a new type of zeroth-order oracle that can approximate the gradient direction without averaging multiple samples of gradients when the norm of the gradient is not too small.

Before we present the main results, several lemmas are in order. The first lemma considers some geometric properties of the unit sphere.

Lemma 5.3.4 *For any non-zero vector $d \in \mathbb{R}^n$ and $\delta < 1$, let S_δ^x be defined as*

$$S_\delta^x := \left\{ v \in S(n) |\ \text{s.t. } |d^\top v| < \delta^2 \right\}.$$

If $\|d\| \geq \delta$, then there exists a constant $C_n > 0$, such that

$$m(S_\delta^x) < C_n \delta.$$

Proof. We have

$$m(S_\delta^x) = \int\limits_{v \in S(n) \cap S_\delta^x} dS_n.$$

By the symmetry of $S(n)$, we may assume w.l.o.g. that $d = (0, \ldots, 0, \|d\|)^\top$. Let $a = \frac{\delta^2}{\|d\|}$.

Since $a < 1$, we have

$$\begin{aligned}
& m(S_\delta^x) \\
= & \int\limits_{v\in S(n)} \mathbf{1}_{\left\{-\frac{\delta^2}{\|d\|}\le v_n\le \frac{\delta^2}{\|d\|}\right\}}(v)dS_n \\
= & \ 2\int\limits_{1-a^2\le v_1^2+\cdots+v_{n-1}^2\le 1} \frac{1}{\sqrt{1-v_1^2-\cdots-v_{n-1}^2}}dv_1\cdots dv_{n-1} \\
= & \ 2\int\limits_{\sqrt{1-a^2}\le r\le 1} \frac{r^{n-2}}{\sqrt{1-r^2}}dr\cdot dS_{n-1} \\
= & \ 2\beta_{n-1}\int\limits_{\sqrt{1-a^2}\le r\le 1} \frac{r^{n-2}}{\sqrt{1-r^2}}dr \\
\le & \ 2\beta_{n-1}\int\limits_{\sqrt{1-a^2}\le r\le 1} \frac{1}{\sqrt{1-r^2}}dr \\
= & \ 2\beta_{n-1}\left(\frac{\pi}{2}-\arcsin(\sqrt{1-a^2})\right) \\
= & \ 2\beta_{n-1}(\arcsin a) < 2\beta_{n-1}\frac{\pi}{2}a = \pi\beta_{n-1}\frac{\delta^2}{\|d\|} \le \pi\beta_{n-1}\delta.
\end{aligned}$$

By setting $C_n = \pi\beta_{n-1}$, the desired result follows. □

The next lemma leads to an unbiased first-order estimator of the direction of a vector.

Lemma 5.3.5 *Suppose $d \in \mathbb{R}^n$, and $d \neq 0$. Then,*

$$\int\limits_{v\in S(n)} \operatorname{sign}(d^\top v)vdS_n = P_n\frac{d}{\|d\|},$$

where P_n is a constant.

Proof. By the symmetry of $S(n)$, again we may assume $d = (0,\ldots,0,\|d\|)^\top$, and

$$\int\limits_{v\in S(n)} \operatorname{sign}(d^\top v)vdS_n = 2\int\limits_{v\in S(n)} \mathbf{1}_{v_n\ge 0}(v)vdS_n.$$

Notice that if $v \in S(n)$, then $u = (-v_1, -v_2, \ldots, -v_{n-1}, v_n)^\top$ is also in $S(n)$. As a result,

the above integral will be on the direction of $\frac{d}{\|d\|} = (0, 0, \ldots, 0, 1)^\top$, and its length is given by

$$\begin{aligned}
& 2 \int_{v \in S(n)} \mathbf{1}_{v_n \geq 0}(v) v_n dS_n \\
= \; & 2 \int_{0 \leq v_1^2 + \cdots + v_{n-1}^2 \leq 1} \sqrt{1 - v_1^2 - \cdots - v_{n-1}^2} dS_n \\
= \; & 2 \int_{0 \leq v_1^2 + \cdots + v_{n-1}^2 \leq 1} \frac{\sqrt{1 - v_1^2 - \cdots - v_{n-1}^2}}{\sqrt{1 - v_1^2 - \cdots - v_{n-1}^2}} dv_1 \ldots dv_{n-1} \\
= \; & 2 \int_{0 \leq r \leq 1} r^{n-2} dr dS_{n-1} \\
= \; & \frac{2\beta_{n-1}}{n-1} := P_n.
\end{aligned}$$

□

Using the previous lemmas, we have the following result which constructs a zeroth-order estimator for the normalized gradient.

Theorem 5.3.6 *Suppose $f(x) \in C_L^1(\mathbb{R}^n)$ and $\|\nabla f(x)\| \geq \delta$ at x. Let $\epsilon = \frac{\delta^2}{L}$. Then we have*

$$\left\| \mathsf{E}_{S(n)} \left[\text{sign}(f(x + \epsilon v) - f(x)) v \right] - Q_n \frac{\nabla f(x)}{\|\nabla f(x)\|} \right\| \leq 2 D_n \delta$$

where v is a random vector uniformly distributed over $S(n)$, and $Q_n = \frac{P_n}{\beta_n}$ and $D_n = \frac{C_n}{\beta_n}$.

Proof. Since f has Lipschitz gradient, we have

$$\begin{aligned}
& |f(x + \epsilon v) - f(x) - \epsilon \nabla f(x)^\top v| \leq \frac{\epsilon L}{2} \|v\|^2 \iff \\
& \nabla f(x)^\top v - \frac{\epsilon}{2} L \leq \frac{f(x + \epsilon v) - f(x)}{\epsilon} \leq \nabla f(x)^\top v + \frac{\epsilon}{2} L.
\end{aligned}$$

Since $|\nabla f(x)^\top v| \geq \delta^2$ for $v \in S(n) \setminus S_\delta^x$, if we let $\epsilon = \frac{\delta^2}{L}$, we have

$$\nabla f(x)^\top v - \frac{\delta^2}{2} \leq \frac{f(x + \epsilon v) - f(x)}{\epsilon} \leq \nabla f(x)^\top v + \frac{\delta^2}{2}.$$

Thus,

$$\begin{aligned} & \operatorname{sign}\left(\nabla f(x)^\top v\right) = \operatorname{sign}\left(\nabla f(x)^\top v - \frac{\delta^2}{2}\right) \\ \leq\ & \operatorname{sign}\left(\frac{f(x+\epsilon v)-f(x)}{\epsilon}\right) \leq \operatorname{sign}\left(\nabla f(x)^\top v + \frac{\delta^2}{2}\right) \\ =\ & \operatorname{sign}\left(\nabla f(x)^\top v\right), \end{aligned}$$

implying $\operatorname{sign}(\nabla f(x)^\top v) = \operatorname{sign}\left(\frac{f(x+\epsilon v)-f(x)}{\epsilon}\right)$. Therefore,

$$\begin{aligned} & \beta_n \mathsf{E}_{S(n)}\left[\operatorname{sign}(f(x+\epsilon v)-f(x))v\right] \\ =\ & \int\limits_{v\in S(n)\setminus S_\delta^x} \left[\operatorname{sign}(f(x+\epsilon v)-f(x))v\right] dS(n) + \int\limits_{v\in S_\delta^x} \left[\operatorname{sign}(f(x+\epsilon v)-f(x))v\right] dS(n) \\ =\ & \int\limits_{v\in S(n)\setminus S_\delta^x} \left[\operatorname{sign}(\nabla f(x)^\top v)v\right] dS(n) + \int\limits_{v\in S_\delta^x} \left[\operatorname{sign}(f(x+\epsilon v)-f(x))v\right] dS(n) \\ =\ & \int\limits_{v\in S(n)} \left[\operatorname{sign}(\nabla f(x)^\top v)v\right] dS(n) - \int\limits_{v\in S_\delta^x} \left[\operatorname{sign}(\nabla f(x)^\top v)v\right] dS(n) \\ & + \int\limits_{v\in S_\delta^x} \left[\operatorname{sign}(f(x+\epsilon v)-f(x))v\right] dS(n) \\ =\ & P_n \frac{\nabla f(x)}{\|\nabla f(x)\|} - \int\limits_{v\in S_\delta^x} \left[\operatorname{sign}(\nabla f(x)^\top v)v\right] dS(n) \\ & + \int\limits_{v\in S_\delta^x} \left[\operatorname{sign}(f(x+\epsilon v)-f(x))v\right] dS(n), \end{aligned}$$

where the last equality is due to Lemma 5.3.5.

Putting the estimations together, we have

$$\begin{aligned} & \left\| \mathsf{E}_{S(n)}\left[\operatorname{sign}(f(x+\epsilon v)-f(x))v\right] - \frac{P_n}{\beta_n}\frac{\nabla f(x)}{\|\nabla f(x)\|} \right\| \\ \leq\ & \frac{1}{\beta_n} \int\limits_{v\in S_\delta^x} \left\| \operatorname{sign}(\nabla f(x)^\top v)v \right\| dS(n) + \frac{1}{\beta_n} \int\limits_{v\in S_\delta^x} \|\operatorname{sign}(f(x+\epsilon v)-f(x))v\|\, dS(n) \\ \leq\ & \frac{2m(S_\delta^x)}{\beta_n} \leq \frac{2C_n\delta}{\beta_n}. \end{aligned}$$

Note that $Q_n = \frac{P_n}{\beta_n}$ and $D_n = \frac{C_n}{\beta_n}$, the theorem is proved. □

Based on Theorem 5.3.6, for a given $\delta > 0$ we have a zeroth-order estimator for the normalized gradient given as:

$$G(x, v) = \frac{\text{sign}(f(x + \epsilon v) - f(x))}{Q_n} v, \tag{5.35}$$

where $\epsilon = \delta^2/L$ and v is an uniformly distributed random vector over $S(n)$. Theorem 5.3.6 implies that the distance between the estimator and the normalized gradient can be controlled up to a factor of δ. Specifically, the Zeroth-Order Normalized Gradient Descent (ZNGD) algorithm is as follows.

Algorithm 3: Zeroth-Order Normalized Gradient Descent (ZONGD)

Input: feasible set $\mathcal{X}$, number of iterations T, δ
Initialization: $x_1 \in \mathcal{X}$, $\epsilon = \delta^2/L$
for $k = 1$ **to** T **do**
  Sample v^k uniformly over $S(n) \subset \mathbb{R}^n$;
  play x^k and $x^k + \epsilon v^k$;
  receive feedbacks $f(x^k)$ and $f(x^k + \epsilon v^k)$;
  set $G(x^k, v^k) = \frac{\text{sign}(f(x^k + \epsilon v^k) - f(x^k))}{Q_n} v^k$;
  update $x^{k+1} = \text{Proj}_{\mathcal{X}}\left(x^k - \eta G(x^k, v^k)\right)$.
end for

Note that Algorithm 3 actually outputs a random sequence of vectors $\{x^k\}_1^T$, and the following theorem shows an expected $O(T^{-1/2})$ rate of convergence for the zeroth-order normalized gradient descent algorithm. The expectation is taken over the σ-field generated by the random variables $\{v^k\}_{k=1}^T$.

Theorem 5.3.7 *Suppose $f \in C_L^1(\mathcal{X})$, and satisfies the error bound condition (Definition 13) and is weakly pseudo-convex with bounded gradient. Let $\{x^k\}_1^T$ be the sequence produced by the zeroth-order normalized gradient descent algorithm. Furthermore, we define an output solution as*

$$\hat{x}_T = \underset{x \in \{x^k\}_1^T}{\text{argmin}} f(x) \tag{5.36}$$

Then, the following inequality holds

$$\mathsf{E}\left[f(\hat{x}_T) - f(x^*)\right] \leq \frac{K}{2\eta T}\left(4R^2 + \frac{T\eta^2}{Q_n^2}\right) + U\delta^\gamma. \tag{5.37}$$

In particular, if we set $\eta = \frac{2Q_nR}{T}$ *and* $\delta = \min\{T^{-\frac{1}{2\gamma}}, T^{-\frac{1}{4}}\}$ *where* $Q_n = \frac{P_n}{\beta_n}$ *and* P_n *is a constant, then an* $O(T^{-1/2})$ *rate of convergence follows.*

Proof. Let $z_k := \|x^k - x^*\|$. Then,

$$\begin{aligned} z_{k+1}^2 &= \|x^{k+1} - x^*\|^2 = \left\| \prod_{\mathcal{X}} \left(x^k - \eta G(x^k, v^k) \right) - x^* \right\|^2 \\ &\leq \left\| x^k - \eta G(x^k, v^k) - x^* \right\|^2 = z_k^2 + \eta^2 \|G(x^k, v^k)\|^2 - 2\eta G(x^k, v^k)^\top (x^k - x^*) \\ &\leq z_k^2 + \frac{\eta^2}{Q_n^2} - 2\eta G(x^k, v^k)^\top (x^k - x^*). \end{aligned}$$

By rearranging the terms, we have:

$$KG(x^k, v^k)^\top (x^k - x^*) \leq \frac{K}{2\eta} \left(z_k^2 - z_{k+1}^2 + \frac{\eta^2}{Q_n^2} \right).$$

Now based on $\|\nabla f(x^k)\|$, we have two different cases:

- $\|\nabla f(x^k)\| \geq \delta$. In this case, by Theorem 5.3.6, we have

$$\|\mathsf{E}[G(x^k, v^k)|x^k] - \frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}\| \leq \frac{2D_n}{Q_n}\delta.$$

Therefore,

$$\begin{aligned} f(x^k) - f(x^*) &\leq K \frac{\nabla f(x^k)^\top (x^k - x^*)}{\|\nabla f(x^k)\|} \\ &\leq K \mathsf{E}[G(x^k, v^k)|x^k]^\top (x^k - x^*) + \frac{2D_n K}{Q_n}\delta \|x^k - x^*\| \\ &= \frac{K}{2\eta} \left(\mathsf{E}[z_k^2|x^k] - \mathsf{E}[z_{k+1}^2|x^k] + \frac{\eta^2}{Q_n^2} \right) + \frac{2D_n K}{Q_n}\delta \|x^k - x^*\| \\ &\leq \frac{K}{2\eta} \left(\mathsf{E}[z_k^2|x^k] - \mathsf{E}[z_{k+1}^2|x^k] + \frac{\eta^2}{Q_n^2} \right) + \frac{4D_n K}{Q_n} R\delta. \end{aligned} \tag{5.38}$$

- $\|\nabla f(x^k)\| < \delta$. In this case, by the error bound property (Definition 13) we have

$$\|x^k - x^*\| \leq D\|\nabla f(x^k)\|^\gamma < D\delta^\gamma.$$

Therefore, due to the boundedness of gradient

$$f(x^k) - f(x^*) \leq M\|x^k - x^*\| \leq MD\delta^\gamma, \tag{5.39}$$

and

$$\begin{aligned} 0 &\leq \frac{K}{2\eta}\left(\mathsf{E}[z_k^2|x^k] - \mathsf{E}[z_{k+1}^2|x^k] + \frac{\eta^2}{Q_n^2}\right) - K\mathsf{E}[G(x^k, v^k)|x^k]^\top(x^k - x^*) \\ &\leq \frac{K}{2\eta}\left(\mathsf{E}[z_k^2|x^k] - \mathsf{E}[z_{k+1}^2|x^k] + \frac{\eta^2}{Q_n^2}\right) + K\frac{\beta_n}{Q_n}D\delta^\gamma. \end{aligned}$$

Adding (5.39) with (5.40), it follows that

$$\begin{aligned} & f(x^k) - f(x^*) \\ \leq\ & \frac{K}{2\eta}\left(\mathsf{E}[z_k^2|x^k] - \mathsf{E}[z_{k+1}^2|x^k] + \frac{\eta^2}{Q_n^2}\right) + \left(K\frac{\beta_n}{Q_n}D + MD\right)\delta^\gamma. \end{aligned} \tag{5.40}$$

In view of (5.38) and (5.40), if we let $U = \max\left\{\frac{4C_nK}{P_n}R, (K\frac{\beta_n}{Q_n}D + MD)\right\}$, then in either case the following inequality holds:

$$f(x^k) - f(x^*) \leq \frac{K}{2\eta}\left(\mathsf{E}[z_k^2|x^k] - \mathsf{E}[z_{k+1}^2|x^k] + \frac{\eta^2}{Q_n^2}\right) + U\delta^\gamma.$$

Summing these inequalities over $k = 1, ..., T$, we have

$$\begin{aligned} & \mathsf{E}\left[f(\hat{x}_T) - f(x^*)\right] \\ \leq\ & \frac{1}{T}\mathsf{E}\left[\sum_{k=1}^{T}(f(x^k) - f(x^*))\right] \\ \leq\ & \frac{K}{2\eta T}\left(\mathsf{E}[z_1^2] - \mathsf{E}[z_{T+1}^2] + \frac{T\eta^2}{Q_n^2}\right) + U\delta^\gamma \\ \leq\ & \frac{K}{2\eta T}\left(4R^2 + \frac{T\eta^2}{Q_n^2}\right) + U\delta^\gamma. \end{aligned}$$

By choosing $\eta = \frac{2Q_nR}{\sqrt{T}}$, and $\delta = \min\{T^{-\frac{1}{2\gamma}}, T^{-\frac{1}{4}}\}$, we have

$$\mathsf{E}\left[f(\hat{x}_T) - f(x^*)\right] \leq \frac{2KR}{Q_n}\frac{1}{\sqrt{T}} + \frac{U}{\sqrt{T}} \leq O(\frac{1}{\sqrt{T}}).$$

□

Compared with the result in [50] where a first-order method is considered, we show the similar convergence rate for a zeroth-order method under a more general condition (weakly pseudo-convex). Moreover, the zeroth-order estimator for the normalized gradient could be of interest on its own.

5.4 Optimization with a Controllably Noisy Objective

In this section, we consider model (5.1) without h and $\mathcal{X} = \mathbb{R}^n$ as following

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in \mathbb{R}^n. \end{array} \tag{5.41}$$

Moreover, the evaluation of the objective function f is assumed can be made to any degree of precision at the cost of paying an increased effort. To be precise, we have the following assumption.

Assumption 5.4.1 *For all $\rho > 0$, by an effort of $\mathrm{eff}(\rho)$, the feedback $\hat{f}_\rho(x)$ (which could be either stochastic or deterministic) can be made to be no more than ρ away from the true value $f(x)$, i.e.*

$$|\hat{f}_\rho(x) - f(x)| < \rho.$$

For simplicity we assume that eff is independent of x. For problem (5.41), we study how to achieve the best overall precision by optimizing total efforts. We choose to work with a finite difference approach, and consider the following deterministic zeroth-order oracle.

Definition 14 *Let e_i, $i = 1, \ldots, n$, be the standard orthonormal basis of $\mathbb{R}^n$. Then for $\forall \mu > 0, \rho > 0$, we define*

$$g_i(x, \mu, \rho) = \frac{\hat{f}_\rho(x + \mu e_i) - \hat{f}_\rho(x)}{\mu}, \tag{5.42}$$

and $g(x, \mu, \rho) = \sum_{i=1}^{n} g_i(x, \mu, \rho) e_i$.

In this section, we assume that the objective function $f(x)$ is strongly convex defined as follows:

Definition 15 *A smooth function f is said to be σ-strongly convex if*

$$\frac{\sigma}{2}\|x-y\|^2 \le f(y) - f(x) - \nabla f(x)^{\mathrm{T}}(y-x), \ \ \forall x, y \in \mathbb{R}^n.$$

Note that we do not impose any constraint on $\hat{f}_\rho(x)$. The following lemma shows that if ρ is small enough, the difference between $g(x,\mu,\rho)$ and $\nabla f(x)$ can be bounded.

Lemma 5.4.1 *For $f \in C^1_L(\mathbb{R}^n)$ and $x \in \mathbb{R}^n$, we have:*

$$\begin{aligned} \|g(x,\mu,\rho) - \nabla f(x)\| &\le \frac{\sqrt{2n}\mu L}{2} + \frac{2\sqrt{2n}\rho}{\mu}, \\ \|g(x,\mu,\rho)\|^2 &\le 2\|\nabla f(x)\|^2 + \frac{n\mu^2 L^2}{2} + \frac{8n\rho^2}{\mu^2}. \end{aligned}$$

Proof. Since $f \in C^1_L(\mathbb{R}^n)$, from descent lemma we have

$$\begin{array}{rcccl} -\frac{L}{2}\mu^2 + \frac{\partial f}{\partial x_i}(x^k)\mu & \le & f(x^k+\mu e_i) - f(x^k) & \le & \frac{L}{2}\mu^2 + \frac{\partial f}{\partial x_i}(x^k)\mu, \\ -\frac{L}{2}\mu & \le & \frac{1}{\mu}\left(f(x^k+\mu e_i) - f(x^k)\right) - \frac{\partial f}{\partial x_i}(x^k) & \le & \frac{L}{2}\mu. \end{array}$$

Therefore,

$$\begin{aligned} & \|g(x,\mu,\rho) - \nabla f(x)\|^2 \\ = & \sum_{i=1}^n \left(\frac{1}{\mu}\left(\hat{f}_\rho(x^k+\mu e_i) - \hat{f}_\rho(x^k)\right) - \frac{\partial f}{\partial x_i}(x^k)\right)^2 \\ \le & \sum_{i=1}^n 2\left\{ \left(\frac{1}{\mu}\left(f(x^k+\mu e_i) - f(x^k)\right) - \frac{\partial f}{\partial x_i}(x^k)\right)^2 \right. \\ & \left. + \frac{1}{\mu^2}\left(\hat{f}_\rho(x^k+\mu e_i) - f(x^k+\mu e_i) + f(x^k) - \hat{f}_\rho(x^k)\right)^2 \right\} \\ \le & \frac{n}{2}\mu^2 L^2 + \frac{8n\rho^2}{\mu^2}, \end{aligned}$$

which prove the first inequality. For the second inequality, we have

$$\begin{aligned}
& \|g(x,\mu,\rho)\|^2 \\
= & \sum_{i=1}^{n}\left(\frac{1}{\mu}\left(\hat{f}_\rho(x^k+\mu e_i)-\hat{f}_\rho(x^k)\right)\right)^2 \\
= & \sum_{i=1}^{n}\frac{1}{\mu^2}\left(\left(f(x^k+\mu e_i)-f(x^k)\right)\right. \\
& \left.+\left(-\hat{f}_\rho(x^k+\mu e_i)+f(x^k+\mu e_i)-f(x^k)+\hat{f}_\rho(x^k)\right)\right)^2 \\
= & \sum_{i=1}^{n}\frac{1}{\mu^2}\left(2\left(f(x^k+\mu e_i)-f(x^k)\right)^2\right. \\
& \left.+2\left(-\hat{f}_\rho(x^k+\mu e_i)+f(x^k+\mu e_i)-f(x^k)+\hat{f}_\rho(x^k)\right)^2\right) \\
\leq & \sum_{i=1}^{n}\left(2\frac{\mu^2L^2}{4}+2\frac{\partial f}{\partial x_i}(x^k)^2+\frac{8\rho^2}{\mu^2}\right) \\
= & \frac{n\mu^2L^2}{2}+2\|\nabla f(x)\|^2+\frac{8n\rho^2}{\mu^2}.
\end{aligned}$$

□

5.4.1 The Zeroth-Order Gradient Descent Method

In this section, we present a zeroth order algorithm that achieves the linear convergence rate. The algorithm is as follows.

Zeroth-order Gradient Descent with Dynamically Increasing Precision (ZGDDIP)

Parameters: $\gamma > 0, \beta \in (0,1), \epsilon > 0, \alpha > 0, L > 0$.

Initialization: $y^0 = 0$, $\rho = 1/2$.

for $t = 0, 1, \cdots, T$,

$x^0 = y^t$;

$\mu = \frac{2\sqrt{\rho}}{\sqrt{L}}$;

for $k = 0, 1, \cdots, K_t$,

$g(x^k, \mu, \rho) = \frac{\sum_{i=1}^{n} (\hat{f}_\rho(x^k + \mu e_i) - \hat{f}_\rho(x^k)) e_i}{\mu}$;

$x^{k+1} = x^k - \alpha g(x^k, \mu, \rho)$.

end for

Set $\bar{k}_t = \arg\min_k \hat{f}_\rho(x^k)$;

$y^{t+1} = x^{\bar{k}_t}$;

$\rho = \beta\rho$.

end while

Note that in the ZGDDIP algorithm, there are two layers of iterations; the outer loop which generates the $\{y^t\}$ sequence, and the inner loop which uses gradient descent with the increasing approximation precision ρ for $g(x^k, \mu, \rho)$. Moreover, in practice, as the original function f is not known, the selection of y^{t+1} only depends on the noisy estimation $\hat{f}_\rho$.

For the ease of later reference, we introduce two parameters :

$$C_1 = \frac{1}{32nL} \text{ and } C_2 = \frac{2}{\sqrt{L}},$$

where L is the Lipschitz constant for gradient ∇f.

The following proposition establishes the sufficient decrease for the inner iterations.

Proposition 5.4.2 *Suppose f is σ-strongly convex (Definition 15) and $f \in C_L^1(\mathbb{R}^n)$. Set $\alpha = \frac{32}{133}$ in ZGDDIP. For all ρ, if $\|\nabla f(x^k)\| \geq \sqrt{\frac{\rho}{C_1}}$ for all steps $k = 1, ..., K_t$, then*

$$f(x^{k+1}) - f(x^*) \leq \left(1 - \frac{32\sigma}{133L}\right)\left(f(x^k) - f(x^*)\right).$$

Proof. If $\|\nabla f(x^k)\| \geq \sqrt{\frac{\rho}{C_1}}$, we have $\rho \leq C_1 \|\nabla f(x^k)\|^2$

and $\mu = C_2\sqrt{\rho} \leq C_2\sqrt{C_1}\|\nabla f(x^k)\|$.

By Definition 15, we have:

$$\begin{aligned}
& f(x^{k+1}) - f(x^k) \\
\leq\ & \frac{L\alpha^2}{2}\|g(x^k,\mu,\rho)\|^2 - \alpha\nabla f(x^k)^{\mathrm{T}} g(x^k,\mu,\rho) \\
=\ & \frac{L\alpha^2}{2}\|g(x^k,\mu,\rho)\|^2 - \alpha\nabla f(x^k)^{\mathrm{T}}\left(g(x^k,\mu,\rho) - \nabla f(x^k)\right) - \alpha\|\nabla f(x^k)\|^2 \\
\leq\ & \frac{L\alpha^2}{2}\left(\frac{n\mu^2L^2}{2} + 2\|\nabla f(x^k)\|^2 + \frac{8n\rho^2}{\mu^2}\right) \\
& +\alpha\|\nabla f(x^k)\|\left(\frac{\sqrt{2n}\mu L}{2} + \frac{2\sqrt{2n}\rho}{\mu}\right) - \alpha\|\nabla f(x^k)\|^2 \\
\leq\ & \frac{L\alpha^2}{2}\|\nabla f(x^k)\|^2\left(\frac{nC_2^2C_1L^2}{2} + 2 + \frac{8nC_1}{C_2^2}\right) \\
& +\alpha\|\nabla f(x^k)\|^2\left(\frac{\sqrt{2nC_1}C_2L}{2} + \frac{2\sqrt{2nC_1}}{C_2}\right) - \alpha\|\nabla f(x^k)\|^2 \\
\leq\ & \frac{L\alpha^2}{2}\|\nabla f(x^k)\|^2\left(\frac{nC_2^2C_1L^2}{2} + 2 + \frac{8nC_1}{C_2^2}\right) \\
& -\alpha\|\nabla f(x^k)\|^2\left(1 - \frac{\sqrt{2nC_1}C_2L}{2} - \frac{2\sqrt{2nC_1}}{C_2}\right) \\
\leq\ & \frac{133L\alpha^2}{128}\|\nabla f(x^k)\|^2 - \frac{1}{2}\alpha\|\nabla f(x^k)\|^2.
\end{aligned}$$

Since we have chosen $\alpha = \frac{32}{133L}$, it follows that

$$f(x^{k+1}) - f(x^k) \leq -\frac{16}{133L}\|\nabla f(x^k)\|^2.$$

By the strong convexity, we have:

$$f(x^*) - f(x^k) \geq -\frac{1}{2\sigma}\|\nabla f(x^k)\|^2.$$

Therefore, we have

$$\begin{aligned}
f(x^{k+1}) - f(x^*) &\leq f(x^k) - f(x^*) - \frac{16}{133L}\|\nabla f(x^k)\|^2 \\
&\leq \left(1 - \frac{32\sigma}{133L}\right)\left(f(x^k) - f(x^*)\right).
\end{aligned}$$

□

The next proposition shows that, with a constant K_t, the precision can be improved by a constant factor β.

Proposition 5.4.3 *Suppose f is σ-strongly convex (Definition 15) and $f \in C_L^1(\mathbb{R}^n)$. In ZGDDIP, set $\alpha = \frac{32}{133L}$, and*

$$K_0 = \left\lfloor \frac{-\ln\left(4C_1\sigma\left(f(0) - f(x^*)\right)\right)}{\ln\left(1 - \frac{32\sigma}{133L}\right)} \right\rfloor + 1$$

and

$$K_t = \left\lfloor \frac{\ln\beta - \ln\left(\frac{1}{2C_1\sigma} + 2\right)}{\ln\left(1 - \frac{32\sigma}{133L}\right)} \right\rfloor + 1, \forall \rho < 1/2.$$

Then we have

$$f(y^t) - f(x^*) \leq \left(\frac{1}{2C_1\sigma} + 2\right)\rho, \forall t = 0, ..., T. \tag{5.43}$$

Proof. In the first outer iteration, $t = 0$, and

$$K_0 > \frac{-\ln\left(4C_1\sigma\left(f(0) - f(x^*)\right)\right)}{\ln\left(1 - \frac{32\sigma}{133L}\right)}.$$

Based on Proposition 5.4.2, we either have:

$$f(x^{K_0}) - f(x^*) \leq \left(1 - \frac{32\sigma}{133L}\right)^{K_0}\left(f(0) - f(x^*)\right) \leq \frac{\rho}{2C_1\sigma},$$

or $\|\nabla f(x^k)\| \geq \sqrt{\frac{\rho}{C_1}}$ for some k. In the latter case, we have

$$f(x^k) - f(x^*) \leq \frac{1}{2\sigma}\|\nabla f(x^k)\|^2 \leq \frac{\rho}{2C_1\sigma}.$$

Therefore, we have

$$\begin{aligned}
\min_{k=0,\ldots,K_t} f(x^k) - f(x^*) &\leq \frac{\rho}{2C_1\sigma}, \\
\min_{k=0,\ldots,K_t} \hat{f}_\rho(x^k) - f(x^*) &\leq \frac{\rho}{2C_1\sigma} + \rho, \\
\hat{f}_\rho(x^{\bar{k}}) - f(x^*) &\leq \frac{\rho}{2C_1\sigma} + \rho, \\
f(x^{\bar{k}}) - f(x^*) &\leq \left(\frac{1}{2C_1\sigma} + 2\right)\rho.
\end{aligned}$$

Suppose (5.43) holds for $0, ..., t-1$, i.e.

$$f(y^s) - f(x^*) \leq \left(\frac{1}{2C_1\sigma} + 2\right)\frac{1}{2}\beta^{s-1},$$

for all $s = 0, ..., t-1$. By Proposition 5.4.2, we will either have:

$$\begin{aligned}
f(x^{K_t}) - f(x^*) &\leq \left(1 - \frac{32\sigma}{133L}\right)^{K_t}\left(f(y^{t-1}) - f(x^*)\right) \\
&\leq \frac{\beta}{1 + 4C_1\sigma}\left(\frac{1}{2C_1\sigma} + 2\right)\frac{1}{2}\beta^{t-1} \\
&\leq \frac{\rho}{2C_1\sigma}
\end{aligned}$$

or $\|\nabla f(x^k)\| \geq \sqrt{\frac{\rho}{C_1}}$ for some k; hence

$$f(y^t) - f(x^*) \leq \left(\frac{1}{2C_1\sigma} + 2\right)\rho.$$

□

Therefore, with a constant step K_t, we are able to reduce the optimality gap from $\left(\frac{1}{2C_1\sigma} + 2\right)\rho$ to $\left(\frac{1}{2C_1\sigma} + 2\right)\beta\rho$. We summarize the sample complexity and the total effort needed, assuming that the effort function is $\text{eff}(\rho) = O\left(\rho^{-\kappa}\right)$ for some $\kappa > 0$.

Theorem 5.4.4 *Suppose f is σ-strongly convex (Definition 15) and $f \in C_L^1(\mathbb{R}^n)$. Let α, K_0, K_t be as defined in Proposition 5.4.3. In at most*

$$M = \left\lfloor \frac{\ln(4C_1\sigma\epsilon) - \ln(1 + 4C_1\sigma)}{\ln\beta} \right\rfloor + 2$$

outer iterations, we would have $f(x^{\bar{k}}) - f(x^*) \leq \epsilon$. *Moreover, the total number of function evaluations is of the order* $O(n(\ln n)^2 \ln(1/\epsilon))$ *(taking* σ, L, β *as constants). In addition, if* $\mathrm{eff}(\rho) = O(\rho^{-\kappa})$ *for some* $\kappa > 0$, *the total efforts we need to spend is* $\mathrm{TEF} = O(n(\ln n)^2 \mathrm{eff}(\epsilon))$.

Proof. Let $\left(\frac{1}{2C_1\sigma} + 2\right)\rho = \epsilon$, we have $\rho = \frac{2C_1\sigma\epsilon}{1+4C_1\sigma}$. Since in each outer iteration, ρ is reduced to $\beta\rho$. The total number of outer iterations required to reduce ρ from $1/2$ to $\frac{2C_1\sigma\epsilon}{1+4C_1\sigma}$ is $\frac{\ln(4C_1\sigma\epsilon)-\ln(1+4C_1\sigma)}{\ln\beta}$, thus proving the first claim. From Proposition 5.4.3, we know that the total effort required is upper bounded by

$$n\left\lfloor \frac{-\ln\left(4C_1\sigma\left(f(0)-f(x^*)\right)\right)}{\ln\left(1-\frac{32\sigma}{133L}\right)} \right\rfloor + n\frac{\ln(4C_1\sigma\epsilon)-\ln(1+4C_1\sigma)}{\ln\beta}\left\lfloor \frac{\ln\beta - \ln\left(\frac{1}{2C_1\sigma}+2\right)}{\ln\left(1-\frac{32\sigma}{133L}\right)} \right\rfloor + 2n,$$

which is of the order $O(n\ln n\ln(1/\epsilon))$.

We may specify the total efforts to spend to be

$$\begin{aligned} \mathrm{TEF} &\leq n\left\lfloor \tfrac{-\ln(4C_1\sigma(f(0)-f(x^*)))}{\ln\left(1-\frac{32\sigma}{133L}\right)} \right\rfloor \mathrm{eff}(1/2) + n\textstyle\sum_{i=1}^{M-1} \mathrm{eff}(0.5\beta^i)\left\lfloor \tfrac{\ln\beta-\ln\left(\frac{1}{2C_1\sigma}+2\right)}{\ln\left(1-\frac{32\sigma}{133L}\right)} \right\rfloor \\ &= n\left\lfloor \tfrac{-\ln(4C_1\sigma(f(0)-f(x^*)))}{\ln\left(1-\frac{32\sigma}{133L}\right)} \right\rfloor \mathrm{eff}(1/2) + O\left(n2^{\kappa}\left\lfloor \tfrac{\ln\beta-\ln\left(\frac{1}{2C_1\sigma}+2\right)}{\ln\left(1-\frac{32\sigma}{133L}\right)} \right\rfloor \tfrac{\beta^{-M\kappa}-\beta^{-\kappa}}{\beta^{-\kappa}-1}\right). \end{aligned}$$

Since $M \leq \frac{\ln\epsilon}{\ln\beta} + \frac{\ln(4C_1\sigma)-\ln(1+4C_1\sigma)}{\ln\beta} + 3$, we have $\mathrm{TEF} = O(n\ln n\,\mathrm{eff}(\epsilon))$. □

Remark: If we use the highest precision from the beginning, i.e., we let $\rho = \frac{2C_1\sigma\epsilon}{1+4C_1\sigma}$ in the first outer iteration, it is not hard to verify that the number of iterations required is $\frac{\ln\epsilon-\ln(f(0)-f(x^*))}{\ln(133L-32\sigma)-\ln(133L)}$. Therefore, the total sample complexity is of the order $O(n\ln n\ln(1/\epsilon))$ but the total effort required is $O(n\ln n\,\epsilon^{-\kappa}\ln(1/\epsilon))$. By dynamically increasing the precision, we can reduce the total effort by a factor of $\ln(1/\epsilon)$.

5.4.2 The Zeroth-Order Ellipsoid Method

In this section, we present another algorithm that incorporate the ellipsoid method. We show that this algorithm also achieves the linear convergence rate. We further show that one can relax the assumption of strongly convexity. We assume in this subsection that

function $f(x)$ is Lipschitz continuous. We start by showing that if μ is small enough, $g(x, \mu, \rho)$ provides a supporting hyperplane.

Lemma 5.4.5 *Suppose f is σ-strongly convex (Definition 15) and $f \in C^1_L(\mathbb{R}^n)$. If $\mu \leq \frac{\epsilon\sigma}{2L\sqrt{2n}}$, and $\rho = \frac{\mu^2 L}{4}$, then it holds that*

$$f(x) - f(x^*) \leq \frac{2L}{\sigma} g(x, \mu, \rho)^\top (x - x^*), \tag{5.44}$$

as long as $\|x - x^*\| \geq \epsilon$, where x^* is the minimum point of f.

Proof. By Lemma 5.4.1, we have

$$\begin{aligned} g(x, \mu, \rho)^\top (x - x^*) &\geq \nabla f(x)^\top (x - x^*) - \left(\frac{\sqrt{2n}\mu L}{2} + \frac{2\sqrt{2n}\rho}{\mu} \right) \|x - x^*\| \\ &\geq \sigma \|x - x^*\|^2 - \sqrt{2n}\mu L \|x - x^*\| \\ &= \frac{\sigma}{2} \|x - x^*\|^2 + \|x - x^*\| \left(\frac{\sigma}{2} \|x - x^*\| - \sqrt{2n}\mu L \right). \end{aligned}$$

Clearly, as long as $\|x - x^*\| \geq \epsilon$, we have $g(x, \mu, \rho)^\top (x - x^*) \geq \frac{\sigma}{2} \|x - x^*\|^2$, and therefore,

$$\begin{aligned} f(x) - f(x^*) &\leq \nabla f(x)^\top (x - x^*) \\ &\leq L \|x - x^*\|^2 \\ &\leq \frac{2L}{\sigma} g(x, \mu, \rho)^\top (x - x^*). \end{aligned} \tag{5.45}$$

□

Suppose that an initial ball with radius R is found to contain x^*. At step k, we have an iterative point x^k. We then spend effort to compute a ρ-accurate search direction at $x^k + \mu e_i$. With that direction, we proceed with the ellipsoid method. Formally, we present our zeroth-order ellipsoid algorithm as follows.

The zeroth-order ellipsoid algorithm

Parameters: $\rho, \mu > 0$.

Initialization: $x^0 = a^0 = 0$, $A^0 = R^2 I$.

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

$$d^k = g(x^k, \mu, \rho) = \frac{\sum_{i=1}^{n} (\hat{f}_\rho(x^k + \mu e_i) - \hat{f}_\rho(x^k)) e_i}{\mu};$$

$$b^k = A^k d^k / \sqrt{(d^k)^{\mathrm{T}} A^k d^k};$$

$$A^{k+1} = \frac{n^2}{n^2-1} \left(A^k - \frac{2}{n+1} b^k (b^k)^{\mathrm{T}} \right);$$

$$x^{k+1} = x^k - \frac{1}{n+1} b^k.$$

end for

Denote the ellipsoid at the k-th iteration be $E(A^k; x^k) := \{x \mid (x - x^k)^\top A_k^{-1} (x - x^k) \leq 1\}$. Inequality (5.45) stipulates that as long as $\|x^k - x^*\| \geq \epsilon$, the half-ellipsoid $E(A^k; x^k) \cap \{x \mid (d^k)^\top (x - x^k) \leq 0\}$ contains the optimal solution x^*. Below, we shall present a convergence analysis without resorting to the geometric insights of the ellipsoid algorithm; see also [36].

Theorem 5.4.6 *Set $0 < \epsilon \leq 1$, $\mu = \frac{\epsilon\sigma}{2L\sqrt{2n}}$, and $\rho = \frac{\mu^2 L}{4}$. Suppose f is σ-strongly convex (Definition 15) and $f \in C^0_{L_0}(\mathbb{R}^n) \cap C^1_L(\mathbb{R}^n)$. Let us run the zeroth-order ellipsoid algorithm for k iterations. Then, either $\min_{0 \leq \ell \leq k} \|x^\ell - x^*\| < \epsilon$, or*

$$\min_{0 \leq \ell \leq k} f(x^\ell) - f(x^*) \leq \frac{4\sqrt{2} L R \max(L_0, \sigma)}{\sigma} \exp\left(-\frac{k}{2n^2} \right).$$

Proof. Denote

$$\delta_k := \frac{\sqrt{g(x^k, \mu, \rho)^\top A_k g(x^k, \mu, \rho)}}{\|g(x, \mu, \rho)\|},$$

and $\delta'_k := \min_{1 \leq \ell \leq k} \delta_\ell$.

If $\|x^k - x^*\| \geq \epsilon$, then by Lemma 5.4.5 we have

$$\begin{aligned} f(x^k) - f(x^*) &\leq \frac{2L}{\sigma} (d^k)^{\mathrm{T}} (x^k - x^*) \\ &\leq \frac{2L}{\sigma} \left[(d^k)^{\mathrm{T}} x^k - \min\{(d^k)^{\mathrm{T}} x \mid x \in E(A^k; x^k)\} \right] \\ &= \frac{2L}{\sigma} \sqrt{(d^k)^{\mathrm{T}} A^k d^k} \\ &= \frac{2L \|d^k\|}{\sigma} \delta_k. \end{aligned} \tag{5.46}$$

Denote $\tau := \frac{n^2}{n^2-1}$ and $\zeta := \frac{2}{n+1}$. By the Sherman-Morrison formula,

$$\begin{aligned}(A^{k+1})^{-1} &= \frac{1}{\tau}\left[(A^k)^{-1} + \frac{\zeta}{1-\zeta}\cdot\frac{d^k(d^k)^{\mathrm{T}}}{(d^k)^{\mathrm{T}}A^k d^k}\right],\\ \det(A^{k+1}) &= \tau^n(1-\zeta)\det(A^k).\end{aligned}$$

This leads to

$$\begin{aligned}\det((A^{k+1})^{-1}) &= \left([\tau^n(1-\zeta)]^{k+1}\det(A^0)\right)^{-1} = \left([\tau^n(1-\zeta)]^{k+1}R^{2n}\right)^{-1},\\ \mathrm{tr}\left((A^{k+1})^{-1}\right) &= \frac{1}{\tau}\mathrm{tr}\left((A^k)^{-1}\right) + \frac{\zeta}{\tau(1-\zeta)}\delta_k^{-2}\\ &= \frac{1}{\tau^{k+1}}\mathrm{tr}\left((A^0)^{-1}\right) + \frac{\zeta}{1-\zeta}\sum_{\ell=0}^{k}\frac{1}{\tau^{k-\ell+1}}\cdot\frac{1}{\delta_\ell^2}.\end{aligned}$$

Using the inequality

$$n\left(\det(A^{k+1})^{-1}\right)^{\frac{1}{n}} \le \mathrm{tr}\,\left((A^{k+1})^{-1}\right)$$

we have

$$\frac{n}{R^2\tau^{k+1}\left((1-\zeta)^{\frac{1}{n}}\right)^{k+1}} \le \frac{n}{R^2\tau^{k+1}} + \frac{\zeta}{1-\zeta}\sum_{\ell=0}^{k}\frac{1}{\tau^{k-\ell+1}}\cdot\frac{1}{\delta_\ell^2},$$

leading to

$$\frac{\zeta}{1-\zeta}\sum_{\ell=0}^{k}\frac{\tau^\ell}{\delta_\ell^2} \ge \frac{n}{R^2}\left[\frac{1}{(1-\zeta)^{\frac{k+1}{n}}} - 1\right].$$

Hence,

$$(\delta_k')^2 \le \frac{\zeta R^2}{n(1-\zeta)(\tau-1)} \times \frac{\tau^{k+1}-1}{\left(\frac{1}{(1-\zeta)^{\frac{1}{n}}}\right)^{k+1} - 1}. \tag{5.47}$$

Because

$$\tau(1-\zeta)^{\frac{1}{n}} = \left[\left(\frac{n}{n-1}\right)^{n-1}\left(\frac{n}{n+1}\right)^{n+1}\right]^{\frac{1}{n}},$$

and

$$\left(\frac{n}{n-1}\right)^{n-1}\left(\frac{n}{n+1}\right)^{n+1} < \exp\left(-\frac{1}{n}\right),$$

it follows from (5.47) that

$$(\delta_k')^2 \leq \frac{\zeta R^2}{n(1-\zeta)(\tau-1)} \exp\left(-\frac{k}{n^2}\right) = \frac{2(n+1)R^2}{n} \exp\left(-\frac{k}{n^2}\right) \leq 4R^2 \exp\left(-\frac{k}{n^2}\right). \tag{5.48}$$

Since $0 < \epsilon \leq 1$, by Lemma 5.4.1, we have

$$\begin{aligned} \|d^k\| &\leq \sqrt{2\|\nabla f(x)\|^2 + \frac{n\mu^2 L^2}{2} + \frac{8n\rho^2}{\mu^2}} \\ &\leq \sqrt{2}\|\nabla f(x)\| + \frac{\sqrt{n}\mu L}{\sqrt{2}} + \frac{\sqrt{8n}\rho}{\mu} \\ &\leq \sqrt{2}L_0 + \frac{\epsilon\sigma}{2} \\ &\leq \sqrt{2}\max(L_0, \sigma). \end{aligned}$$

Therefore, combining (5.46) and (5.48) we have

$$\min_{0\leq\ell\leq k} f(x^\ell) - f(x^*) \leq \frac{2L\|d^k\|}{\sigma}\delta_k' \leq \frac{4\sqrt{2}LR\max(L_0,\sigma)}{\sigma} \exp\left(-\frac{k}{2n^2}\right).$$

□

Theorem 5.4.6 implies that in at most $k = 2n^2 \ln\left(\frac{4\sqrt{2}R\max(L_0,\sigma)}{\sigma\epsilon^2}\right)$ iterations, the zeroth order ellipsoid algorithm will ensure that

$$\min_{0\leq\ell\leq k} f(x^\ell) - f(x^*) \leq L\epsilon^2.$$

The total efforts required is $2n^3 \ln\left(\frac{4\sqrt{2}R\max(L_0,\sigma)}{\sigma\epsilon^2}\right) \mathrm{eff}\left(\frac{\sigma^2\epsilon^2}{32nL}\right)$.

Remark: If $f(x)$ is convex but not strongly convex, by assuming that $f \in C_{L_0}^0(\mathbb{R}^n) \cap C_L^1(\mathbb{R}^n)$, then one may perturb the objective function to be

$$f(x) + \frac{\epsilon}{R^2}\|x\|^2.$$

The function is now $\frac{2\epsilon}{R^2}$-strongly convex, and its minimum is ϵ away from that of the original function. Replacing σ by ϵ, the total effort is $2n^3 \ln\left(\frac{4\sqrt{2}RL_0}{\epsilon^3}\right) \mathrm{eff}\left(\frac{\epsilon^4}{32nL}\right)$.

5.5 Regularized Optimization with Controllable Accuracies

In this section, we consider the general model (5.1), or

$$\min_{x\in\mathcal{X}} \Phi(x) := f(x) + h(x) \tag{5.49}$$

where $\mathcal{X} \in \mathbb{R}^n$ is bounded convex set contained in the Euclidean ball with radius R, f is a smooth function but possibly non-convex, and h is a convex function but possibly non-smooth. Furthermore, we introduce some definitions and assumptions regarding the objective function.

For the function $\Phi(x)$, we define the proximal gradient mapping and proximal gradient.

Definition 16 *For a given x, the proximal gradient mapping x^+ of Φ is defined as*

$$x^+ = \operatorname*{argmin}_{y\in\mathcal{X}} \left\{ \langle \nabla f(x), y\rangle + h(y) + \frac{1}{2\gamma}\|y - x\|^2 \right\}. \tag{5.50}$$

Moreover, denote

$$\tilde{\nabla}\Phi(x;\gamma) := \frac{1}{\gamma}(x - x^+).$$

Based on the proximal gradient, we can similarly define the proximal gradient dominant condition of the composite function $\Phi(x)$.

Assumption 5.5.1 ***(Proximal Gradient Dominance)*** *The function $\Phi(x)$ in (5.49) is said to be proximal gradient dominant if there exist an α such that the following holds*

$$\Phi(x) - \Phi^* \le \chi\|\tilde{\nabla}\Phi(x;\alpha)\|^2, \ \ \forall x \in \mathcal{X} \tag{5.51}$$

where χ is a positive constant.

5.5.1 Basics of the Proximal Gradient Mapping

In this subsection, we discuss some properties of the proximal gradient mapping. The use of the gradient $\nabla f(x)$ in (5.50) is not necessary, and it can be replaced by any vector

$d \in \mathbb{R}^n$. The proximal mapping can be similarly defined as

$$x^+ = \underset{y \in \mathcal{X}}{\operatorname{argmin}} \left\{ \langle d, y \rangle + h(y) + \frac{1}{2\gamma} \|y - x\|^2 \right\}. \tag{5.52}$$

Moreover, the generalized proximal gradient can still be defined as

$$G(x, d, \gamma) := \frac{1}{\gamma}(x - x^+), \tag{5.53}$$

where x^+ is given by (5.52).

The first lemma shows the monotonicity of the norm of the generalized proximal gradient.

Lemma 5.5.1 *Let $G(x, d, \gamma)$ be given in* (5.53) *for fixed x and d. Then, the norm $\|G(x, d, \gamma)\|$ is a non-increasing function of γ.*

Proof. We denote

$$m(\gamma) := \min_{y \in \mathcal{X}} \left\{ M(y, \gamma) := \langle d, y \rangle + h(y) + \frac{1}{2\gamma} \|y - x\|^2 \right\}. \tag{5.54}$$

Since $M(y, \gamma)$ is jointly convex with respect to y and γ, $m(\gamma)$ is also convex. Moreover, based on the definition of x^+, we have

$$\frac{dm(\gamma)}{d\gamma} = -\frac{\|x - x^+\|^2}{2\gamma^2}. \tag{5.55}$$

Due to the convexity of $m(\gamma)$, the $\frac{dm(\gamma)}{d\gamma}$ is a non-decreasing function. That results in $\|G(x, d, \gamma)\|$ is a non-increasing function of γ. □

Besides its simplicity, the Lemma 5.5.1 has an important implication. It shows that if Assumption 5.5.1 is satisfied with a specific α, then it would also hold for any γ such that $0 < \gamma \leq \alpha$.

The following lemmas present some geometric properties of the proximal gradient, and their proofs can be found in [44]. The first lemma shows the magnitude of $G(x, d, \gamma)$ and its angle between d can be bounded, and the second lemma establishes the non-expansiveness of the proximal gradient.

Lemma 5.5.2 *Let x^+ be given in* (5.52)*. Then, the following inequality holds*

$$\langle d, G(x,d,\gamma)\rangle \geq \|G(x,d,\gamma)\|^2 + \frac{1}{\gamma}\left(h(x^+) - h(x)\right).$$

Lemma 5.5.3 *Suppose $G(x,d_1,\gamma)$ and $G(x,d_2,\gamma)$ are the generalized proximal gradient with d_1 and d_2 in* (5.52) *respectively. Then, we have the inequality below*

$$\|G(x,d_1,\gamma) - G(x,d_2,\gamma)\| \leq \|d_1 - d_2\|. \tag{5.56}$$

5.5.2 Deterministic Zeroth-Order Algorithm

In this section, we propose a zeroth-order ISTA algorithm for problem (5.49). Similarly, we assume the controllable noisy function evaluation is available. For all $\rho > 0$, with the effort $\mathrm{eff}(\rho)$, the noisy estimation is at most ρ away, i.e. $|\hat{f}_\rho(x) - f(x)| < \rho$ for all x.

It is clear that for function $f \in C_L^1(\mathcal{X})$, the bounds in Lemma 5.4.1 still hold. Let $g(x,\mu,\rho)$ be defined as in Definition 14.

Lemma 5.5.4 *Suppose $f \in C_L^1(\mathcal{X})$, i.e. it satisfies* (5.7)*, then the following inequalities hold*

$$\begin{aligned}
\|g(x,\mu,\rho) - \nabla f(x)\| &\leq \frac{\sqrt{2n}\mu L}{2} + \frac{2\sqrt{2n}\rho}{\mu}, \\
\|g(x,\mu,\rho)\|^2 &\leq 2\|\nabla f(x)\|^2 + \frac{n\mu^2 L^2}{2} + \frac{8n\rho^2}{\mu^2}.
\end{aligned}$$

Based on the oracle defined above, we propose the deterministic zeroth-order ISTA algorithm.

Deterministic zeroth-order ISTA

Parameters: $\gamma, \mu_k, \rho_k > 0$.

Initialization: $x^0 = 0$.

for $k = 0, 1, \cdots,$

Set $\mu_k = \sqrt{\rho_k}$;

$g(x^k,\mu_k,\rho_k) = \frac{\sum_{i=1}^n (\hat{f}_{\rho_k}(x^k+\mu_k e_i) - \hat{f}_{\rho_k}(x^k))e_i}{\mu_k}$;

$x^{k+1} = \underset{y\in\mathcal{X}}{\mathrm{argmin}} \left\{ \langle g(x^k,\mu_k,\rho_k), y\rangle + h(y) + \frac{1}{2\gamma}\|y - x^k\|^2 \right\}.$

end for

Notice that the accuracy parameter ρ_k and the finite difference parameter μ_k are allowed to change over time. Under the proximal gradient dominant condition, the following theorem shows the relationship between consecutive iterates of the deterministic zeroth-order ISTA. Moreover, the generalized proximal gradient is denoted as $g^k = \frac{1}{\gamma}(x^k - x^{k+1})$.

Theorem 5.5.5 *Suppose $f \in C_L^1(\mathcal{X})$ and $\Phi(x)$ satisfies Assumption 5.5.1, and $\gamma \leq \min\{1/L, \alpha\}$. Let $\{x^k\}$ be the sequence produced by the deterministic zeroth-order ISTA algorithm. The following holds*

$$\begin{aligned}\Phi(x^{k+1}) - \Phi^* \quad \leq \quad & \left(1 - \frac{\gamma - \frac{L\gamma^2}{2}}{2\chi}\right)(\Phi(x^k) - \Phi^*) \\ & + \left(\gamma - \frac{L\gamma^2}{2}\right)\left(\frac{n\mu_k^2 L^2}{2} + \frac{8n\rho_k^2}{\mu_k^2}\right) + 2R\left(\frac{\sqrt{2n}\mu_k L}{2} + \frac{2\sqrt{2n}\rho_k}{\mu_k}\right),\end{aligned}$$

where R is the radius of $\mathcal{X}$, i.e. $\|x\| \leq R\ \forall x \in \mathcal{X}$.

Proof. It follows from Assumption 10 that

$$\begin{aligned} f(x^{k+1}) \quad \leq \quad & f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\ = \quad & f(x^k) - \langle \nabla f(x^k), \gamma g^k \rangle + \frac{L\gamma^2}{2}\|g^k\|^2 \\ = \quad & f(x^k) - \gamma\langle g(x^k, \mu_k, \rho_k), g^k \rangle + \frac{L\gamma^2}{2}\|g^k\|^2 + \gamma\langle g(x^k, \mu_k, \rho_k) - \nabla f(x^k), g^k \rangle \\ \leq \quad & f(x^k) - \gamma\|g^k\|^2 + h(x^k) - h(x^{k+1}) + \frac{L\gamma^2}{2}\|g^k\|^2 \\ & + \gamma\langle g(x^k, \mu_k, \rho_k) - \nabla f(x^k), g^k \rangle. \end{aligned}$$

The first inequality follows from descent lemma, whereas the second inequality is due

to Lemma 5.5.2. Based on the definition of Φ, we have

$$\begin{aligned}
\Phi(x^{k+1}) &\leq \Phi(x^k) - \left(\gamma - \frac{L\gamma^2}{2}\right)\|g^k\|^2 + \gamma\langle g(x^k,\mu_k,\rho_k) - \nabla f(x^k), g^k\rangle \\
&\leq \Phi(x^k) - \left(\gamma - \frac{L\gamma^2}{2}\right)\frac{\|\tilde{\nabla}\Phi(x^k;\gamma)\|^2}{2} + \left(\gamma - \frac{L\gamma^2}{2}\right)\|\tilde{\nabla}\Phi(x^k;\gamma) - g^k\|^2 \\
&\quad +\gamma\langle g(x^k,\mu_k,\rho_k) - \nabla f(x^k), g^k\rangle \\
&\leq \Phi(x^k) - \left(\gamma - \frac{L\gamma^2}{2}\right)\frac{\|\tilde{\nabla}\Phi(x^k;\gamma)\|^2}{2} \\
&\quad + \left(\gamma - \frac{L\gamma^2}{2}\right)\|g(x^k,\mu_k,\rho_k) - \nabla f(x^k)\|^2 + \gamma\langle g(x^k,\mu_k,\rho_k) - \nabla f(x^k), g^k\rangle.
\end{aligned}$$

Since $\|g^k\| = \frac{1}{\gamma}\|x^k - x^{k+1}\| \leq \frac{2R}{\gamma}$, invoking the proximal gradient dominant condition (5.51) and $\gamma \leq \min\{1/L, \alpha\}$, we have

$$\begin{aligned}
&\Phi(x^{k+1}) - \Phi^* \\
\leq\ & \Phi(x^k) - \Phi^* - \frac{\gamma - \frac{L\gamma^2}{2}}{2\chi}(\Phi(x^k) - \Phi^*) \\
&+ \left(\gamma - \frac{L\gamma^2}{2}\right)\|g(x^k,\mu_k,\rho_k) - \nabla f(x^k)\|^2 + \gamma\langle g(x^k,\mu_k,\rho_k) - \nabla f(x^k), g^k\rangle \\
\leq\ & \left(1 - \frac{\gamma - \frac{L\gamma^2}{2}}{2\chi}\right)(\Phi(x^k) - \Phi^*) \\
&+ \left(\gamma - \frac{L\gamma^2}{2}\right)\left(\frac{n\mu_k^2L^2}{2} + \frac{8n\rho_k^2}{\mu_k^2}\right) + 2R\left(\frac{\sqrt{2n}\mu_k L}{2} + \frac{2\sqrt{2n}\rho_k}{\mu_k}\right).
\end{aligned}$$

□

The following corollary shows the linear convergence rate with constant μ and ρ.

Corollary 5.5.6 *Suppose $f \in C_L^1(\mathcal{X})$ and $\Phi(x)$ satisfies Assumption 5.5.1, and $\gamma \leq \min\{1/L, \alpha\}$, $\mu_k = \mu$, $\rho_k = \rho = \mu^2$. Let $\{x^k\}$ be the sequence produced by the deterministic zeroth-order ISTA algorithm, then the following holds*

$$\Phi(x^k) - \Phi^* \leq p^k(\Phi(x^0) - \Phi^*) + \frac{q}{1-p}, \tag{5.57}$$

where $p = 1 - \frac{\gamma - \frac{L\gamma^2}{2}}{2\chi}$ and $q = \left(\gamma - \frac{L\gamma^2}{2}\right)\left(\frac{n\mu^2L^2}{2} + \frac{8n\rho^2}{\mu^2}\right) + 2R\left(\frac{\sqrt{2n}\mu L}{2} + \frac{2\sqrt{2n}\rho}{\mu}\right)$, where

R satisfies $\|x\| \leq R \ \forall x \in \mathcal{X}$.

Proof. Let $e_k := \Phi(x^k) - \Phi^*$, $p = 1 - \frac{\gamma - \frac{L\gamma^2}{2}}{2\chi}$ and $q = \left(\gamma - \frac{L\gamma^2}{2}\right)\left(\frac{n\mu^2 L^2}{2} + \frac{8n\rho^2}{\mu^2}\right) + 2R\left(\frac{\sqrt{2n}\mu L}{2} + \frac{2\sqrt{2n}\rho}{\mu}\right)$. It follows from Theorem 5.5.5 that

$$e_{k+1} \leq pe_k + q.$$

Hence,

$$e_k \leq p^k e_0 + q\sum_{i=0}^{k-1} p^i \leq p^k e_0 + \frac{q}{1-p},$$

and that proves (5.57). □

Remark: With the choice of γ, we have $p := 1 - \frac{\gamma - \frac{L\gamma^2}{2}}{2\chi} < 1$. In view of (5.57), if μ is chosen of the order $O(\epsilon)$, the deterministic zeroth-order ISTA can reach ϵ accuracy of the objective function Φ^* in $O(\ln(1/\epsilon))$ iterations. However, since $\rho = \mu^2 = O(\epsilon^2)$ and there are n function value queries at each iteration, the total effort is $O(n\ln(1/\epsilon))\text{eff}(\epsilon^2)$.

For a particular effort function $\text{eff}(\rho) = D\rho^{-\kappa}$ where $\kappa > 0$ and $D > 0$, the following corollary shows that the dynamically increased precision ρ_k could be beneficial.

Corollary 5.5.7 *Suppose* $f \in C_L^1(\mathcal{X})$ *and* $\Phi(x)$ *satisfies Assumption 5.5.1, and* $\gamma \leq \min\{1/L, \alpha\}$, $\mu_k = q^k$, $\rho_k = q^{2k}$ *where* $1 - \frac{\gamma - \frac{L\gamma^2}{2}}{2\chi} < q < 1$. *Let* $\{x^k\}$ *be the sequence produced by the deterministic zeroth-order ISTA algorithm, then the following holds*

$$\Phi(x^k) - \Phi^* \leq p^k(\Phi(x^0) - \Phi^*) + \frac{Uq^k}{q-p}, \tag{5.58}$$

where $p = 1 - \frac{\gamma - \frac{L\gamma^2}{2}}{2\chi}$, *and R satisfies* $\|x\| \leq R \ \forall x \in \mathcal{X}$, *and U is constant depends on* γ, L *and* n.

Moreover, to achieve the ϵ *accuracy of* Φ^*, *the total effort is of the order* $O(n\epsilon^{-2\kappa})$.

Proof. Let $e_k := \Phi(x^k) - \Phi^*$, $p = 1 - \frac{\gamma - \frac{L\gamma^2}{2}}{2\chi}$ and

$$r_k = \left(\gamma - \frac{L\gamma^2}{2}\right)\left(\frac{n\mu_k^2 L^2}{2} + \frac{8n\rho_k^2}{\mu_k^2}\right) + 2R\left(\frac{\sqrt{2n}\mu_k L}{2} + \frac{2\sqrt{2n}\rho_k}{\mu_k}\right).$$

It follows from Theorem 5.5.5 that

$$e_{k+1} \leq pe_k + r_k.$$

Since $\rho_k = q^{2k}$ and $\mu_k = q^k$ where $p < q < 1$, there exists a constant U such that

$$r_k \leq Uq^k, \ \ \forall k.$$

From the above bound, it is clear that the following holds

$$e_k \leq p^k e_0 + Uq^{k-1} \sum_{i=0}^{k-1} \left(\frac{p}{q}\right)^i \leq p^k e_0 + \frac{Uq^k}{q-p}, \tag{5.59}$$

and that proves (5.58).

Since the dominating term in (5.59) is $\frac{Uq^k}{q-p}$, to achieve ϵ-accuracy of $e_k := \Phi(x^k) - \Phi^*$, it requires $N = O(\ln \epsilon / \ln q)$ iterations. With the effort function $\text{eff}(\rho) = D\rho^{-\kappa}$ and $\rho_k = q^{2k}$, the total effort is

$$\begin{aligned} n \sum_{k=0}^{N} \text{eff}(\rho_k) &= n \sum_{k=0}^{N} Dq^{-2k\kappa} \\ &= nD \frac{q^{-2\kappa(N+1)} - 1}{q^{-2\kappa} - 1}. \end{aligned}$$

Note that N is of the order $O(\ln \epsilon / \ln q)$, as a result, the total effort is of the order $O(n \, \text{eff}^2(\epsilon))$. □

Remark: With the choice of γ, we have $p := 1 - \frac{\gamma - \frac{L\gamma^2}{2}}{2\chi} < q < 1$. In view of (5.58), the deterministic zeroth-order ISTA can reach ϵ accuracy of the objective function Φ^* in $O(\ln(1/\epsilon))$ iterations. However, the total effort is of the order $O(n \, \text{eff}^2(\epsilon))$. It reduces the effort compared with that $O(n \ln(1/\epsilon) \, \text{eff}^2(\epsilon))$ when ρ_k's are set to be constantly ρ.

5.5.3 Stochastic Zeroth-Order Algorithm

In this section, we consider the stochastic objective function. Specifically, in problem (5.49), the function f is assumed of the following form

$$f(x) = \mathsf{E}[F(x, \xi)] \tag{5.60}$$

where expectation is taken over the random vector ξ. Compared with Assumption 5.2.1, we make slightly different assumptions of $F(x, \xi)$.

Assumption 5.5.2 *Suppose the loss function $f(x)$ is given in the form* (5.60)*, then we assume that $F(x, \xi)$ satisfies*

$$\mathsf{E}[F(x, \xi)] = f(x), \tag{5.61}$$

$$\mathsf{E}[\nabla F(x, \xi)] = \nabla f(x), \tag{5.62}$$

and

$$\mathsf{E}[\|\nabla F(x, \xi) - \nabla f(x)\|^2] \leq \theta_1^2. \tag{5.63}$$

Note that in Assumption 5.5.2, the variance of $F(x, \xi)$ is no longer assumed to be bounded. The smoothing scheme which is the same as (5.8) leads to the stochastic zeroth-order oracle ($\mathcal{SZO}$).

The proximal gradient dominant condition (5.51) can be extended to the smoothing function.

Lemma 5.5.8 *Suppose function $\Phi(x)$ in* (5.49) *satisfies proximal gradient dominant condition, and $f^\mu(x)$ is given by* (5.8)*. Let $\Phi^\mu(x) := f^\mu(x) + h(x)$, where $f^\mu(x)$ is defined as in Definition 8. and x^* be the optimal solution to problem* (5.49)*. Then the following inequality holds*

$$|\Phi^\mu(x) - \Phi^\mu(x^*)| \leq 2\chi\|\tilde{\nabla}\Phi^\mu(x; \alpha)\|^2 + \left(L + \frac{n^2 L^2}{2}\right)\mu^2. \tag{5.64}$$

where $\tilde{\nabla}\Phi^\mu(x; \alpha)$ is defined as in (5.53) *with d being $\nabla f^\mu(x)$.*

Proof. Based on inequality (5.10), we have

$$\begin{aligned}
& |\Phi^\mu(x) - \Phi^\mu(x^*)| \\
= \; & |\Phi^\mu(x) - \Phi(x) - (\Phi^\mu(x^*) - \Phi(x^*)) + \Phi(x) - \Phi(x^*)| \\
\leq \; & |\Phi(x) - \Phi(x^*)| + |\Phi^\mu(x) - \Phi(x)| + |\Phi^\mu(x^*) - \Phi(x^*)| \\
\leq \; & |\Phi(x) - \Phi(x^*)| + L\mu^2 \\
\leq \; & \chi\|\tilde{\nabla}\Phi(x; \alpha)\|^2 + L\mu^2.
\end{aligned} \tag{5.65}$$

Moreover, from the non-expansiveness of the proximal gradient, we have

$$\|\tilde{\nabla}\Phi(x; \alpha) - \tilde{\nabla}\Phi^\mu(x; \alpha)\| \leq \|\nabla f^\mu(w) - \nabla f(w)\| \leq \frac{\mu n L}{2}.$$

The above inequality together with (5.65) leads to

$$\begin{aligned} & |\Phi^\mu(x) - \Phi^\mu(x^*)| \\ \leq\ & \chi\|\tilde{\nabla}\Phi(x;\alpha)\|^2 + L\mu^2 \\ \leq\ & \chi(2\|\tilde{\nabla}\Phi^\mu(x;\alpha)\|^2 + 2\|\tilde{\nabla}\Phi(x;\alpha) - \tilde{\nabla}\Phi^\mu(x;\alpha)\|^2) + L\mu^2 \\ \leq\ & 2\chi\|\tilde{\nabla}\Phi^\mu(x;\alpha)\|^2 + \frac{\mu^2 n^2 L^2}{2} + L\mu^2, \end{aligned}$$

which proves (5.64). □

Based on (5.9) we similarly define the stochastic zeroth-order gradient($\mathcal{SZO}$) of f at point x:

$$G_\mu(x,\xi,v) = \frac{n}{\mu}\left[F(x+\mu v,\xi) - F(x,\xi)\right] v, \tag{5.66}$$

where v is the random vector uniformly distributed over the unit sphere in $\mathbb{R}^n$.

The following lemma shows some properties of the $\mathcal{SZO}$. Note that function f satisfies Assumption 5.5.2, i.e. (5.61) and (5.62) hold. This fact together with Lemma 5.2.1(a) leads to:

Lemma 5.5.9 *Suppose that $G_\mu(x,\xi,v)$ is defined as in* (5.66)*, and f satisfies Assumption 5.5.2, i.e.* (5.61)*,* (5.62) *and* (5.63) *hold. Then*

$$\mathsf{E}_{v,\xi}[G_\mu(x,\xi,v)] = \nabla f^\mu(x). \tag{5.67}$$

Suppose $f(x) \in C^0_{L_0}(\mathcal{X})$ and $F(x,\xi) \in C^1_L(\mathbb{R}^n)$ for all ξ. Then the following holds

$$\mathsf{E}_{v,\xi}[\|G_\mu(x,\xi,v)\|^2] \leq 2nN + \mu^2 L^2 n^2, \tag{5.68}$$

where $N = L_0^2 + \theta_1^2$.

Proof. The first equation is easy to verify. We prove the second inequality. Apply-

ing (5.12) and (5.63) to $F(x,\xi)$, we have

$$\begin{aligned}
& \mathsf{E}_{v,\xi}\left[\|G_\mu(x,\xi,v)\|^2\right] \\
= \;& \mathsf{E}_{\xi}\left[\mathsf{E}_{v}\left[\|G_\mu(x,\xi,v)\|^2\right]\right] \\
\overset{(5.12)}{\leq} \;& 2n\left[\mathsf{E}_{\xi}\left[\|\nabla F(x,\xi)\|^2\right]\right] + \frac{\mu^2}{2}L^2n^2 \\
\leq \;& 2n\left\{\mathsf{E}_{\xi}[\|\nabla f(x)\|^2] + \mathsf{E}_{\xi}\left[\|\nabla F(x,\xi)-\nabla f(x)\|^2\right]\right\} + \mu^2L^2n^2 \\
\leq \;& 2n\left(\|\nabla f(x)\|^2 + \theta_1^2\right) + \mu^2L^2n^2 \\
\leq \;& 2nN + \mu^2L^2n^2. \qquad (5.69)
\end{aligned}$$

□

From Lemma 5.5.9, it also implies the boundedness of the variance of $G_\mu(x,\xi,v)$. In fact, the following inequality is straightforward

$$\mathsf{E}_{v,\xi}[\|G_\mu(x,\xi,v)-\nabla f^\mu(x)\|^2] \leq \mathsf{E}_{v,\xi}[\|G_\mu(x,\xi,v)\|^2] \leq 2nN + \mu^2L^2n^2. \qquad (5.70)$$

For the simplicity, we denote $\hat{\theta} := 2nN + \mu^2L^2n^2$. In order to control the variance, in the algorithm, we take m samples of the $\mathcal{SZO}$ at each iteration. Formally, we define

$$G_\mu(x^k,m) = \frac{1}{m}\sum_{t=1}^{m} G_\mu(x^k,\xi_k^t,v_k^t). \qquad (5.71)$$

Moreover, define $\delta^k := G_\mu(x,\xi,v) - \nabla f^\mu(x)$, it is clear that

$$\mathsf{E}_{v,\xi}[\|\delta\|^2] \leq \frac{\hat{\theta}^2}{m}. \qquad (5.72)$$

Now we are ready to show the stochastic zeroth-order ISTA algorithm for the problem (5.60).

Stochastic zeroth-order ISTA

Parameters: $\gamma, \mu > 0$.

Initialization: $x^0 = 0$.

for $k = 0, 1, \cdots,$

$$G_\mu(x^k, m) = \frac{1}{m}\sum_{t=1}^{m} G_\mu(x^k, \xi_k^t, v_k^t);$$

$$x^{k+1} = \operatorname*{argmin}_{y\in\mathcal{X}} \left\{ \langle G_\mu(x^k, m), y\rangle + h(y) + \tfrac{1}{2\gamma}\|y - x^k\|^2 \right\}.$$

end for

The following theorem shows the linear convergence for the algorithm. The expectation is taken over the σ-field generated by the random variables $\{\xi_k^t, v_k^t\}_{t=1}^m$, $k = 1, 2, \ldots$, and the generalized proximal gradient is defines as

$$g_\mu^k = \frac{1}{\gamma}(x^k - x^{k+1}).$$

Theorem 5.5.10 *Suppose $\Phi(x)$ satisfies Assumption 5.5.1 and Assumption 5.5.2 and $F(x, \xi) \in C_L^1(\mathbb{R}^n)$ for all ξ. Let $\{x^k\}$ be the sequence produced by the stochastic zeroth-order ISTA algorithm and $\gamma \le \min\{1/L, \alpha\}$. Then,*

$$\mathsf{E}[\Phi(x^k)] - \Phi^* \le p^k(\Phi(x^0) - \Phi^*) + \frac{q}{1-p} + 2L\mu^2, \tag{5.73}$$

where $p = 1 - \frac{\gamma - \frac{L\gamma^2}{2}}{4\chi}$ and $q = \frac{\gamma - \frac{L\gamma^2}{2}}{4\chi}\left(L + \frac{n^2L^2}{2}\right)\mu^2 + \left(2\gamma - \frac{L\gamma^2}{2}\right)\frac{\hat{\theta}^2}{m}$.

Proof. It follows from Assumption 10 and the property of the smoothing function that

$$\begin{aligned} f^\mu(x^{k+1}) &\le f^\mu(x^k) + \langle \nabla f^\mu(x^k), x^{k+1} - x^k\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\ &= f^\mu(x^k) - \langle \nabla f^\mu(x^k), \gamma g_\mu^k\rangle + \frac{L\gamma^2}{2}\|g_\mu^k\|^2 \\ &= f^\mu(x^k) - \gamma\langle G_\mu(x^k, m), g_\mu^k\rangle + \frac{L\gamma^2}{2}\|g_\mu^k\|^2 + \gamma\langle G_\mu(x^k, m) - \nabla f^\mu(x^k), g_\mu^k\rangle \\ &\le f^\mu(x^k) - \gamma\|g_\mu^k\|^2 + h(x^k) - h(x^{k+1}) + \frac{L\gamma^2}{2}\|g_\mu^k\|^2 + \gamma\langle \delta^k, g_\mu^k\rangle. \end{aligned} \tag{5.74}$$

The first inequality follows from the so-called descent lemma, whereas the second in-

equality is due to Lemma 5.5.2. Based on the definition of Φ^μ, we have

$$\begin{aligned}
\Phi^\mu(x^{k+1}) &\leq \Phi^\mu(x^k) - (\gamma - \frac{L\gamma^2}{2})\|g_\mu^k\|^2 + \gamma\langle\delta^k, g_\mu^k\rangle \\
&\leq \Phi^\mu(x^k) - \left(\gamma - \frac{L\gamma^2}{2}\right)\frac{\|\tilde{\nabla}\Phi^\mu(x^k;\gamma)\|^2}{2} \\
&\quad + \left(\gamma - \frac{L\gamma^2}{2}\right)\|\tilde{\nabla}\Phi^\mu(x^k;\gamma) - g_\mu^k\|^2 + \gamma\langle\delta^k, g_\mu^k\rangle \\
&\leq \Phi^\mu(x^k) - \left(\gamma - \frac{L\gamma^2}{2}\right)\frac{\|\tilde{\nabla}\Phi^\mu(x^k;\gamma)\|^2}{2} + \left(\gamma - \frac{L\gamma^2}{2}\right)\|\delta^k\|^2 \\
&\quad + \gamma\langle\delta^k, \tilde{\nabla}\Phi^\mu(x^k;\gamma)\rangle + \gamma\langle\delta^k, g_\mu^k - \tilde{\nabla}\Phi^\mu(x^k;\gamma)\rangle. \qquad (5.75)
\end{aligned}$$

Notice $\mathsf{E}[\langle\delta^k, \tilde{\nabla}\Phi^\mu(x^k;\gamma)\rangle|x^k] = 0$ and $\langle\delta^k, g_\mu^k - \tilde{\nabla}\Phi^\mu(x^k;\gamma)\rangle \leq \|\delta^k\|\|g_\mu^k - \tilde{\nabla}\Phi^\mu(x^k;\gamma)\| \leq \|\delta^k\|^2$. Taking expectation on both sides of (5.75), we have

$$\mathsf{E}[\Phi^\mu(x^{k+1})] \leq \mathsf{E}[\Phi^\mu(x^k)] - \left(\gamma - \frac{L\gamma^2}{2}\right)\frac{\mathsf{E}[\|\tilde{\nabla}\Phi^\mu(x^k;\gamma)\|^2]}{2} + \left(2\gamma - \frac{L\gamma^2}{2}\right)\mathsf{E}[\|\delta^k\|^2].$$

Invoking the proximal gradient dominant condition (5.64) and $\gamma \leq \min\{1/L, \alpha\}$, we have

$$\begin{aligned}
&\mathsf{E}[\Phi^\mu(x^{k+1})] - \Phi^\mu(x^*) \\
\leq\ & \mathsf{E}[\Phi^\mu(x^k)] - \Phi^\mu(x^*) + \left(2\gamma - \frac{L\gamma^2}{2}\right)\mathsf{E}[\|\delta^k\|^2] \\
& - \frac{\gamma - \frac{L\gamma^2}{2}}{4\chi}\left(\mathsf{E}[\Phi^\mu(x^k)] - \Phi^\mu(x^*) - \left(L + \frac{n^2L^2}{2}\right)\mu^2\right) \\
\leq\ & \left(1 - \frac{\gamma - \frac{L\gamma^2}{2}}{4\chi}\right)\left(\mathsf{E}[\Phi^\mu(x^k)] - \Phi^\mu(x^*)\right) \\
& + \frac{\gamma - \frac{L\gamma^2}{2}}{4\chi}\left(L + \frac{n^2L^2}{2}\right)\mu^2 + \left(2\gamma - \frac{L\gamma^2}{2}\right)\frac{\hat{\theta}^2}{m}.
\end{aligned}$$

Denote $e_k := \mathsf{E}[\Phi^\mu(x^k)] - \Phi^\mu(x^*)$, $p = 1 - \frac{(\gamma - \frac{L\gamma^2}{2})}{4\chi}$ and $q = \frac{\gamma - \frac{L\gamma^2}{2}}{4\chi}(L + \frac{n^2L^2}{2})\mu^2 + (2\gamma - \frac{L\gamma^2}{2})\frac{\hat{\theta}^2}{m}$. The above expression can be simplified to

$$e_{k+1} \leq pe_k + q.$$

Expanding recursively, we obtain

$$e_k \leq p^k e_0 + q \sum_{i=0}^{k-1} p^i \leq p^k e_0 + \frac{q}{1-p}.$$

Since $\Phi(x^k) - \Phi(x^*) - L\mu^2 \leq \Phi^\mu(x^k) - \Phi^\mu(x^*) \leq \Phi(x^k) - \Phi(x^*) + L\mu^2$, this leads to (5.73). □

Remark: With the choice of γ, we have $p := 1 - \frac{\gamma - \frac{L\gamma^2}{2}}{4\chi} < 1$. In view of (5.73), if μ is chosen of the order $O(\sqrt{\epsilon})$ and m is of the order $O(1/\epsilon)$, the stochastic zeroth-order ISTA can reach ϵ accuracy of the objective function Φ^* in $O(\ln(1/\epsilon))$ iterations. However, if we consider the total number of samples, the sample complexity is of the order $O(1/\epsilon \ln(1/\epsilon))$.

5.6 Numerical Experiments

In this section, we test the performance of the zeroth-order gradient descent algorithm on two problem instances: Branin-Hoo function and logistic regression classification on the popular MNIST data, on which we compare with the Bayesian optimization algorithms [108]. For Bayesian optimization, the Branin-Hoo function is a common benchmark test case [64]. It is defined over $x \in \mathbb{R}^2$ where $0 \leq x_1 \leq 15$ and $-5 \leq x_1 \leq 15$. We also test logistic regression classification task on the popular MNIST data. This is a typical application of the black-box optimization, where our goal is to find the best configuration of the hyperparameters in terms of the general misclassification error. Since the MNIST is a multi-class dataset, we use the multinomial logistic regression with L_1 regularization. The algorithm requires choosing three hyperparameters, the L_1 regularization parameter, between 0 and 2, the tolerance, from 1e-6 to 0.1 and the number of iterations, from 20 to 300. Specifically, we compare with the Bayesian optimization of two different acquisition functions, the expected improvement (EI) and the upper confidence bound (UCB), where both of them are based on the Gaussian process model. For each algorithm, the mean and standard error of every iteration are reported, and they are tested on the Branin-Hoo and logistic regression problems for 100 and 10 times respectively. The results of these analyses are presented in Figures 5.2 in terms of the number of iterations of each algorithm. On Branin-Hoo, all the algorithms are able to find the optimal solution, and our zeroth-order gradient descent converges

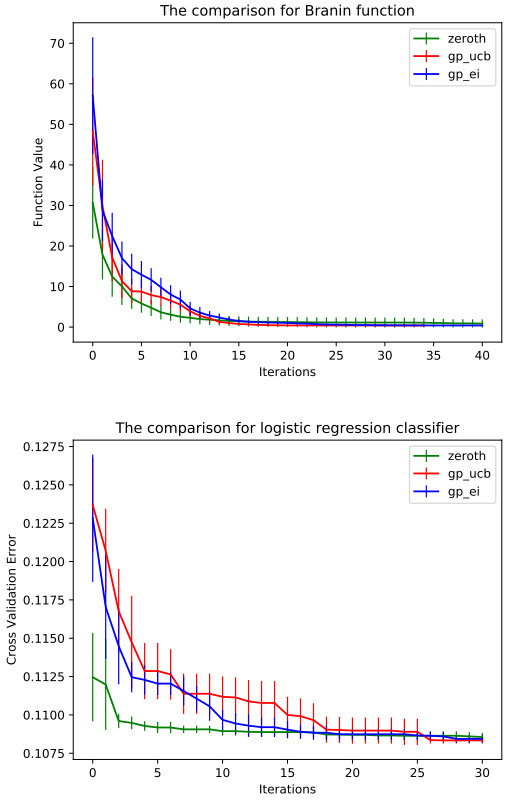

Figure 5.2: Upper: Comparisons on the Branin-Hoo function; Lower: Comparison on training logistic regression on MNIST.

slightly faster than the Bayesian optimization methods. For logistic regression, in terms of the quality of the solution, the Bayesian approaches slightly outperform our method. However, as the training process is very time consuming, the zeroth-order gradient descent has the advantage that it can find a relative good solution with fewer iterations.

5.7 Conclusion

In this chapter, we presented a suite of zeroth-order methods for solving various black-box optimization models. For problems with single objective function and composite objective function, under strong convexity and gradient dominance, we have established linear convergence rate for different zeroth-order methods. For stochastic block-box optimization problem, we considered a single-sample setting and we showed that zeroth-

order gradient descent method can still achieve a sublinear convergence rate. Moreover, for certain classes of nonconvex optimization problems including the star-convexity and weak pseudo-convexity, we proposed simple zeroth-order algorithms which converge to the optimal solution at the sublinear rate. In particular, for the weakly pseudo-convex optimization, we also developed a novel approximation scheme of the direction of the gradient which enables us to extend the applicability of the normalized gradient descent method to the zeroth-order setting. In addition, by comparing with the state-of-the-art Bayesian optimization method for solving some benchmark problems, our numerical results show the comparable practical performance of our algorithms as well.

Chapter 6

Zeroth-Order Algorithms for Online Learning

6.1 Preparation

In general, an online learning (online optimization) problem can be described as follows

General Online Learning Problem

Input: A convex set S

for $t = 1, 2, \ldots$

- predict a vector $x_t \in S$,
- receive a loss function $f_t : S \to \mathbb{R}$,
- suffer a loss $f_t(x_t)$.

As an illustrating example, the online linear regression works as follows: on each decision period, the learner first receives feature vector $w_t \in \mathbb{R}^d$, and then the learner needs to make a prediction p_t. After the true target $y_t \in \mathbb{R}$ is revealed, the learner suffers the loss $|p_t - y_t|$. Assuming the learner is using the linear predictors of the form $w_t \mapsto \langle x, w_t \rangle$, we can easily cast this online prediction problem into the online optimization framework. In particular, the learner should provide a vector x_t, which yields the prediction $p_t = \langle x_t, w_t \rangle$, and the loss function becomes to $f_t(x_t) = |p_t - y_t| = |\langle x_t, w_t \rangle - y_t|$.

In the online learning framework, at each period $t \in \{1, 2, \ldots, T\}$, an online player chooses a feasible strategy x_t from a decision set S, and suffers a loss given by $f_t(x_t)$, where $f_t(\cdot)$ is the loss function. One key feature of the online learning is that the player must make a decision for period t without knowing the loss function $f_t(\cdot)$. As a result, for an online learning algorithm, the performance is usually measured by the so-called *regret*. For a stationary strategy of playing a fixed u, the regret of an online algorithm up to time T with respect to u is defined as:

$$\text{Regret}_T(u) = \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(u). \tag{6.1}$$

where x_t are the predictions produced by the algorithm. For instance, the regret of the online linear regression problem with respect to a fixed linear predictor u is

$$\text{Regret}_T(u) = \sum_{t=1}^{T} |\langle x_t, x_t \rangle - y_t| - \sum_{t=1}^{T} |\langle u, x_t \rangle - y_t|. \tag{6.2}$$

The goal of the online learning is to design some efficient algorithms which can achieve a nontrivial regret bound, i.e. the regret should be bounded as $\text{Regret}_T(u) \leq O(T^\alpha)$, with $\alpha < 1$. When the loss functions f_t are convex and deterministic, there are several algorithms that have been shown that an $O(T^{\frac{1}{2}})$ regret bound is achievable, including the *Follow the Regularized Leader* (FoRel), *Online Gradient Descent*, and *Online Mirror Descent* and so on; see [109]. However, when the loss functions are stochastic, the research on the regret bound is still very limited and we will study that in this chapter.

6.2 Stochastic Loss Functions

We consider an online convex optimization problem where the loss function $f_t(x_t)$ is given in the following form

$$f_t(x_t) = \mathsf{E}_\xi \left[F_t(x_t, \xi) \right] \quad \forall t \geq 1, \tag{6.3}$$

where the expectation is taken over the random variable ξ. However, at each time t, after the learner chooses a decision vector x_t, the full information of $f_t(x_t)$ is not disclosed. Instead, we can observe an unbiased random sample $F_t(x_t, \xi_t)$. Based on

the sample information, an algorithm which outputs a random vector sequence $\{x_t\}_{t=1}^T$ would entail an expected regret as

$$\text{Regret}_T(u) = \mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t) - f_t(u))\right]. \tag{6.4}$$

As we will show in this section, the stochastic zeroth-order online gradient descent method achieves a regret bound in the order of $O(\sqrt{T})$.

6.2.1 The Stochastic Zeroth-Order Online Gradient Descent

Before we present the algorithm, we first introduce some assumptions of $F_t(x_t, \xi)$.

Assumption 6.2.1 *Suppose the loss function $f_t(x_t)$ is given in the form* (6.3)*, then we assume that $F_t(x_t, \xi)$ satisfies*

$$\mathsf{E}[F_t(x_t, \xi)] = f_t(x_t), \tag{6.5}$$

$$\mathsf{E}[\nabla F_t(x_t, \xi)] = \nabla f_t(x_t), \tag{6.6}$$

and

$$\mathsf{E}[\|\nabla F_t(x_t, \xi) - \nabla f_t(x_t)\|^2] \leq \sigma^2. \tag{6.7}$$

for all $t = 1, \ldots, T$.

Now, we introduce our smoothing scheme and point out the definition of $\mathcal{SZO}$ as well as its properties.

Definition 17 *Let U_b be the uniform distribution over the unit Euclidean ball and B be the unit ball. Given $\delta > 0$, the smoothing function f_t^δ is defined as*

$$f_t^\delta(x) = \mathsf{E}_{\{v \sim U_b\}}[f_t(x + \delta v)] = \frac{1}{\alpha(d)} \int_B f_t(x + \delta v) dv \tag{6.8}$$

where $\alpha(d)$ is the volume of the unit ball in $\mathbb{R}^d$.

Some basic properties of the smoothing function are shown in the lemma below, which has been used in previous chapters. For the sake of clarity, we present it here again.

Lemma 6.2.1 *Suppose that $f_t \in C_L^1(\mathbb{R}^d)$. Let U_{S_p} be the uniform distribution over the unit Euclidean sphere, and S_p be the unit sphere in $\mathbb{R}^d$. Then we have:*

(a) The smoothing function f_t^δ is continuously differentiable, and its gradient is Lipschitz continuous with constant $L_\delta \leq L$ and

$$\nabla f_t^\delta(w) = \mathsf{E}_{\{v \sim U_{S_p}\}}\left[\frac{d}{\delta} f_t(w+\delta v) v\right] = \frac{1}{\beta(d)} \int_{v \in S_p} \frac{d}{\delta} \left[f_t(w+\delta v) - f_t(w)\right] v dv \tag{6.9}$$

where $\beta(d)$ is the measure of the unit sphere in $\mathbb{R}^d$.

(b) For any $w \in \mathbb{R}^d$, we have

$$|f_t^\delta(w) - f_t(w)| \leq \frac{L\delta^2}{2}, \tag{6.10}$$

$$\|\nabla f_t^\delta(w) - \nabla f_t(w)\| \leq \frac{\delta d L}{2}, \tag{6.11}$$

$$\mathsf{E}_v\left[\left\|\frac{d}{\delta}[f_t(w+\delta v) - f_t(w)]v\right\|^2\right] \leq 2d\|\nabla f_t(w)\|^2 + \frac{\delta^2}{2} L^2 d^2. \tag{6.12}$$

(c) If f_t is convex, then f_t^δ is also convex.

Now based on (6.9) we define the zeroth-order stochastic gradient of f_t at point x_t:

$$G_\delta(x_t, \xi_t, v) = \frac{d}{\delta}\left[F_t(x_t + \delta v, \xi_t) - F_t(x_t, \xi_t)\right] v, \tag{6.13}$$

where v is the random vector uniformly distributed over the unit sphere in $\mathbb{R}^d$.

Before presenting the regret bound analysis for the algorithm , we first show some properties of the function $G(x_t, \xi_t) := \nabla_x F(x_t, \xi_t)$.

Lemma 6.2.2 *Suppose that $G_\delta(x_t, \xi_t, v)$ is defined as in* (6.13)*, and f_t satisfies Assumption 6.2.1, i.e.* (6.5)*,* (6.6) *and* (6.7) *hold. Then*

$$\mathsf{E}_{v,\xi_t}[G_\delta(x_t, \xi_t, v)] = \nabla f_t^\delta(x_t). \tag{6.14}$$

If we further assume $\|\nabla f_t(w)\| \leq M$, $\forall w \in S, t = 1, \ldots, T$, then the following holds

$$\mathsf{E}_{v,\xi_t}[\|G_\delta(x_t, \xi_t, v)\|^2] \leq 2dN + \delta^2 L^2 d^2, \tag{6.15}$$

where $N = M^2 + \sigma^2$.

Proof. The first statement is easy to verify. We shall focus on the second statement. Applying (6.12) and (6.7) to $F_t(x_t, \xi_t)$ and $G(x_t, \xi_t)$, we have

$$\begin{aligned}
& \mathsf{E}_{v,\xi_t}\left[\|G_\delta(x_t,\xi_t,v)\|^2\right] \\
= \;& \mathsf{E}_{\xi_t}\left[\mathsf{E}_v\left[\|G_\delta(x_t,\xi_t,v)\|^2\right]\right] \\
\le \;& 2d\left[\mathsf{E}_{\xi_t}\left[\|G(x_t,\xi_t)\|^2\right]\right] + \frac{\delta^2}{2}L^2d^2 \\
\le \;& 2d\left\{\mathsf{E}_{\xi_t}[\|\nabla f_t(x_t)\|^2] + \mathsf{E}_{\xi_t}\left[\|G(x_t,\xi_t)-\nabla f_t(x_t)\|^2\right]\right\} + \delta^2L^2d^2 \\
\le \;& 2d\left\{\|\nabla f_t(x_t)\|^2+\sigma^2\right\} + \delta^2L^2d^2 \\
\le \;& 2dN + \delta^2L^2d^2. \qquad (6.16)
\end{aligned}$$

□

For the online convex optimization problem (6.3), the stochastic zeroth-order online gradient descent method is as follows.

Stochastic Zeroth-order Online Gradient Descent

---

parameters: $\eta, \delta > 0$ and a convex set $S \subset \mathbb{R}^d$

initialize: $\theta_1 = 0$

for $t = 0, 1, \cdots T$,

let $x_t = \arg\min_{w\in S} \|w - \eta\theta_t\|_2$

pick $v_t \sim U_{S_p}$

predict $x_t + \delta v_t$ and x_t, receive $F_t(x_t + \delta v_t), F_t(x_t)$

compute $\mathcal{SZO}$ as $z_t = \frac{d}{\delta}\left[F_t(x_t + \delta v_t, \xi_t) - F_t(x_t, \xi_t)\right] v_t := G_\delta(x_t, \xi_t, v_t)$

update $\theta_{t+1} = \theta_t - z_t$

end for

One observation is in order here. Since we need to evaluate both $F_t(x_t + \delta v_t, \xi_t)$ and $F_t(x_t, \xi_t)$ at time t, we will incur two losses at time t, namely $f_t(x_t + \delta v_t)$ and $f_t(x_t)$. Thus the regret of this algorithm should be define as

$$\mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t) - f_t(u))\right] + \mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t+\delta v_t) - f_t(u))\right].$$

The following theorem shows the $O(\sqrt{T})$ regret bound for the algorithm. Moreover, we use Ξ_n to denote the σ-field generated by the random variables ξ_t, v_t, $t = 1, \ldots, n$.

Theorem 6.2.3 *Consider running the stochastic zeroth-order online gradient descent method on the loss functions $f_t \in C_L^1(\mathbb{R}^d)$ which satisfy Assumption 6.2.1. Let S be a convex set and define $B = \max_{u\in S} \|u\|$. Then, for all $u \in S$ we have*

$$\begin{aligned}&\mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t) - f_t(u))\right] + \mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t + \delta v_t) - f_t(u))\right]\\ \leq&\frac{1}{\eta}B^2 + 2\eta(2dN + \delta^2L^2d^2)T + (2L\delta^2 + M\delta)T. \end{aligned} \tag{6.17}$$

In particular, if we set $\eta = B/\sqrt{(2dN + \delta^2L^2d^2)2T}$, and $\delta = T^{-\frac{1}{2}}/(Ld)$, then we have the regret is bounded by $O(\sqrt{T})$.

Proof. In view of the Follow the Regularized Leader algorithm (cf. [109]), we have the following,

$$\sum_{t=1}^{T}\langle z_t, x_t - u\rangle \leq \frac{1}{2\eta}\|u\|_2^2 + \eta\sum_{t=1}^{T}\|z_t\|^2. \tag{6.18}$$

Take expectation, and notice that $\mathsf{E}_{\xi_t,v_t}[z_t|\Xi_{t-1}] = \nabla f_t^\delta(x_t)$ which is shown in (6.14), we have

$$\mathsf{E}\left[\sum_{t=1}^{T}\langle \nabla f_t^\delta(x_t), x_t - u\rangle\right] \leq \frac{1}{2\eta}\|u\|_2^2 + \eta\sum_{t=1}^{T}\mathsf{E}\left[\|z_t\|^2\right]. \tag{6.19}$$

Recall the convexity of f_t^δ, we have

$$\mathsf{E}\left[\sum_{t=1}^{T}(f_t^\delta(x_t) - f_t^\delta(u))\right] \leq \frac{1}{2\eta}\|u\|_2^2 + \eta\sum_{t=1}^{T}\mathsf{E}\left[\|z_t\|^2\right]. \tag{6.20}$$

From the properties of the smoothing function, we have the following chain of inequality

$$\begin{aligned}&f_t(x_t) - f_t(u) \leq f_t^\delta(x_t) - f_t^\delta(u) + L\delta^2,\\ &f_t(x_t + \delta v_t) - f_t(u) \leq f_t(x_t) - f_t(u) + M\delta \leq f_t^\delta(x_t) - f_t^\delta(u) + L\delta^2 + M\delta. \end{aligned} \tag{6.21}$$

By combining (6.21) and (6.20), we have

$$\begin{aligned}&\mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t)-f_t(u))\right]+\mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t+\delta v_t)-f_t(u))\right]\\ \leq&\frac{1}{\eta}\|u\|_2^2+2\eta\sum_{t=1}^{T}\mathsf{E}\left[\|z_t\|^2\right]+(2L\delta^2+M\delta)T. \end{aligned}\tag{6.22}$$

Now, from (6.15), we can bound $\sum_{t=1}^{T}\mathsf{E}\left[\|z_t\|^2\right]$ as

$$\sum_{t=1}^{T}\mathsf{E}\left[\|z_t\|^2\right]\leq(2dN+\delta^2L^2d^2)T.$$

Let $B=\max_{u\in S}\|u\|$, we have

$$\begin{aligned}&\mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t)-f_t(u))\right]+\mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t+\delta v_t)-f_t(u))\right]\\ \leq&\frac{1}{\eta}B^2+2\eta(2dN+\delta^2L^2d^2)T+(2L\delta^2+M\delta)T. \end{aligned}\tag{6.23}$$

In particular, if we set $\eta=B/\sqrt{(2dN+\delta^2L^2d^2)2T}$, and $\delta=T^{-\frac{1}{2}}/(Ld)$, then we have the regret is bounded by $O(\sqrt{T})$. □

6.3 Extensions Under Stochastic Loss Function Setting

6.3.1 Non-differentiability

Previously, we assume the function f_t is differentiable with Lipschitz gradient. In this section, we only assume the continuity of the function f_t and we further assume

Assumption 6.3.1 *Suppose the loss function $f_t(x_t)$ is given in the form* (6.3)*. Moreover, f_t is a continuous function which satisfies*

$$|f_t(x)-f_t(y)|\leq L\|x-y\|^{\beta},\quad\forall x,y\in S\tag{6.24}$$

where $\beta\leq 1$.

Under these assumptions, we have a similar lemma for the smoothed function f_t^{δ}, .

Lemma 6.3.1 *Suppose that f_t satisfies Assumption 6.3.1. Let U_{S_p} be the uniform distribution over the unit Euclidean sphere, and S_p be the unit sphere in $\mathbb{R}^d$. Then we have:*

(a) The smoothing function f_t^δ is continuously differentiable, and it also satisfies Assumption 6.3.1 with constant $L_\delta \leq L$ and

$$\nabla f_t^\delta(x) = \mathsf{E}_{\{v\sim U_{S_p}\}}\left[\frac{d}{\delta} f_t(x+\delta v)v\right] = \frac{1}{\beta(d)}\int_{v\in S_p}\frac{d}{\delta}\left[f_t(x+\delta v) - f_t(x)\right]v dv \tag{6.25}$$

where $\beta(d)$ is the measure of the unit sphere in $\mathbb{R}^d$.

(b) For any $x \in \mathbb{R}^d$, we have

$$|f_t^\delta(x) - f_t(x)| \leq L\delta^\beta \frac{d}{d+\beta} < L\delta^\beta, \tag{6.26}$$

$$\mathsf{E}_{\{v\sim U_{S_p}\}}\left[\left\|\frac{d}{\delta}[f_t(x+\delta v) - f_t(x)]v\right\|^2\right] \leq L^2 d^2 \delta^{2\beta-2}. \tag{6.27}$$

(c) If f_t is convex, then f_t^δ is also convex.

Now if we define the zeroth-order stochastic oracle $G_\delta(x_t, \xi_t, v)$ similar to (6.13), then we have following lemma.

Lemma 6.3.2 *Suppose that $G_\delta(x_t, \xi_t, v)$ is defined as in* (6.13)*, and f_t satisfies Assumption 6.3.1. Then*

$$\mathsf{E}_{v,\xi_t}[G_\delta(x_t, \xi_t, v)] = \nabla f_t^\delta(x_t). \tag{6.28}$$

Moreover, the following estimate holds

$$\mathsf{E}_{v,\xi_t}[\|G_\delta(x_t, \xi_t, v)\|^2] \leq L^2 d^2 \delta^{2\beta-2}, \tag{6.29}$$

Proof. The first statement is easy to verify. We shall focus on the second statement.

Applying (6.27), we have

$$\begin{aligned}
& \mathsf{E}_{v,\xi_t}\left[\|G_\delta(x_t,\xi_t,v)\|^2\right] \\
= \; & \mathsf{E}_{\xi_t}\left[\mathsf{E}_v\left[\|G_\delta(x_t,\xi_t,v)\|^2\right]\right] \\
= \; & \mathsf{E}_{\xi_t}\left[\mathsf{E}_v\left[\left\|\frac{d}{\delta}[F_t(x+\delta v,\xi_t)-F_t(x,\xi_t)]v\right\|^2\right]\right] \\
\leq \; & \mathsf{E}_{\xi_t}\left[L^2d^2\delta^{2\beta-2}\right] = L^2d^2\delta^{2\beta-2}. \qquad (6.30)
\end{aligned}$$

□

Under the non-differentiability, the regret bound for the stochastic zeroth-order online gradient descent algorithm is shown in the following theorem.

Theorem 6.3.3 *Consider running the stochastic zeroth-order online gradient descent method on the loss functions $f_t \in C_L^0(\mathbb{R}^d)$ which satisfy Assumption 6.3.1. Let S be a convex set and define $B = \max_{u\in S}\|u\|$. Then, for all $u \in S$ we have*

$$\begin{aligned}
& \mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t)-f_t(u))\right] + \mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t+\delta v_t)-f_t(u))\right] \\
\leq & \frac{1}{\eta}B^2 + 2\eta L^2d^2\delta^{2\beta-2}T + 5L\delta^\beta T. \qquad (6.31)
\end{aligned}$$

In particular, if we set $\eta \sim T^{-(1-\frac{\beta}{2})}$, and $\delta \sim T^{-\frac{1}{2}}$, then we have the regret is bounded by $O(T^{(1-\frac{\beta}{2})})$. When $\beta = 1$ we have the bound to be $O(T^{\frac{1}{2}})$.

Proof. In view of the Follow the Regularized Leader algorithm (cf. [109]), we have the following,

$$\sum_{t=1}^{T}\langle z_t, x_t - u\rangle \leq \frac{1}{2\eta}\|u\|_2^2 + \eta\sum_{t=1}^{T}\|z_t\|^2. \qquad (6.32)$$

Take expectation, and notice that $\mathsf{E}_{\xi_t,v_t}[z_t|\Xi_{t-1}] = \nabla f_t^\delta(x_t)$ which is shown in (6.28), we have

$$\mathsf{E}\left[\sum_{t=1}^{T}\langle\nabla f_t^\delta(x_t), x_t - u\rangle\right] \leq \frac{1}{2\eta}\|u\|_2^2 + \eta\sum_{t=1}^{T}\mathsf{E}\left[\|z_t\|^2\right]. \qquad (6.33)$$

Recall the convexity of f_t^δ, we have

$$\mathsf{E}\left[\sum_{t=1}^{T}(f_t^\delta(x_t) - f_t^\delta(u))\right] \leq \frac{1}{2\eta}\|u\|_2^2 + \eta\sum_{t=1}^{T}\mathsf{E}\left[\|z_t\|^2\right]. \tag{6.34}$$

From the property (6.26) of the smoothing function, we have the following chain of inequality

$$\begin{aligned} &f_t(x_t) - f_t(u) \leq f_t^\delta(x_t) - f_t^\delta(u) + 2L\delta^\beta, \\ &f_t(x_t + \delta v_t) - f_t(u) \leq f_t(x_t) - f_t(u) + L\delta^\beta \leq f_t^\delta(x_t) - f_t^\delta(u) + 3L\delta^\beta. \end{aligned} \tag{6.35}$$

By combining (6.35) and (6.34), we have

$$\begin{aligned} &\mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t) - f_t(u))\right] + \mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t + \delta v_t) - f_t(u))\right] \\ \leq &\frac{1}{\eta}\|u\|_2^2 + 2\eta\sum_{t=1}^{T}\mathsf{E}\left[\|z_t\|^2\right] + 5L\delta^\beta T. \end{aligned} \tag{6.36}$$

Now, from (6.29), we can bound $\sum_{t=1}^{T}\mathsf{E}\left[\|z_t\|^2\right]$ as

$$\sum_{t=1}^{T}\mathsf{E}\left[\|z_t\|^2\right] \leq L^2 d^2 \delta^{2\beta-2} T.$$

Let $B = \max_{u\in S}\|u\|$, we have

$$\begin{aligned} &\mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t) - f_t(u))\right] + \mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t + \delta v_t) - f_t(u))\right] \\ \leq &\frac{1}{\eta}B^2 + 2\eta L^2 d^2 \delta^{2\beta-2} T + 5L\delta^\beta T. \end{aligned} \tag{6.37}$$

In particular, if we set $\eta \sim T^{-(1-\frac{\beta}{2})}$, and $\delta \sim T^{-\frac{1}{2}}$, then we have the regret is bounded by $O(T^{(1-\frac{\beta}{2})})$. When $\beta = 1$ we have the bound to be $O(T^{\frac{1}{2}})$. □

6.3.2 One Random Sample at One Sample Point

In this section, we discuss another extension of the stochastic online learning. In previous discussions, the $\mathcal{SZO}$ (zeroth-order oracle) is in the form of

$$G_\delta(x_t, \xi_t, v) = \frac{d}{\delta}\left[F_t(x_t + \delta v, \xi_t) - F_t(x_t, \xi_t)\right] v, \tag{6.38}$$

where we implicitly assume that we can obtain two responses $F_t(x_t + \delta v, \xi_t)$, $F_t(x_t, \xi_t)$ at the same sample point ξ_t for two different query points. However, following the discussion in Chapter 5, in some cases, it is only possible to have one random sample at a point, i.e. we can only receive two responses $F_t(x_t + \delta v, \xi_t^1)$ and $F_t(x_t, \xi_t^2)$ based on different sample points ξ_t^1 and ξ_t^2. As a result, we can define our new $\mathcal{SZO}$ as follows

$$G_\delta(x_t, \xi_t^{1,2}, v) = \frac{d}{\delta}\left[F_t(x_t + \delta v, \xi_t^1) - F_t(x_t, \xi_t^2)\right] v, \tag{6.39}$$

where we assume that ξ_t^1 and ξ_t^2 are independent.

To facilitate our analysis, we make another assumption regarding the function $f_t(x)$.

Assumption 6.3.2 *Suppose the loss function $f_t(x)$ is given in the form* (6.3)*, we assume*

$$Var[F_t(x, \xi)] := \mathsf{E}_\xi\left[(F_t(x, \xi) - f_t(x))^2\right] \le \theta^2, \quad \forall w \in S \tag{6.40}$$

where $\theta > 0$ is a constant.

Under Assumption 6.3.1, we have the following result.

Lemma 6.3.4 *Suppose that $G_\delta(x_t, \xi_t^{1,2}, v)$ is defined as in* (6.39)*, and f_t satisfies Assumption 6.3.1 and Assumption 6.3.2. Then*

$$\mathsf{E}_{v, \xi_t^{1,2}}[G_\delta(x_t, \xi_t^{1,2}, v)] = \nabla f_t^\delta(x_t). \tag{6.41}$$

Moreover, the following estimate holds

$$\mathsf{E}_{v, \xi_t}[\|G_\delta(x_t, \xi_t, v)\|^2] \le 2L^2 d^2 \delta^{2\beta - 2} + 4\frac{d^2}{\delta^2}\theta^2, \tag{6.42}$$

Proof. The first statement is easy to verify. We shall focus on the second statement.

Applying (6.27), we have

$$\begin{aligned}
& \mathsf{E}_{v,\xi_t^{1,2}}\left[\|G_\delta(x_t,\xi_t^{1,2},v)\|^2\right] \\
= \quad & \mathsf{E}_{\xi_t^{1,2}}\left[\mathsf{E}_v\left[\|G_\delta(x_t,\xi_t^{1,2},v)\|^2\right]\right] \\
= \quad & \mathsf{E}_{\xi_t^{1,2}}\left[\mathsf{E}_v\left[\left\|\frac{d}{\delta}[F_t(x+\delta v,\xi_t^1)-F_t(x,\xi_t^2)]v\right\|^2\right]\right] \\
\leq \quad & \mathsf{E}_{\xi_t^{1,2}}\left[\mathsf{E}_v\left[2\left\|\frac{d}{\delta}[F_t(x+\delta v,\xi_t^1)-F_t(x,\xi_t^1)]v\right\|^2+2\left\|\frac{d}{\delta}[F_t(x,\xi_t^1)-F_t(x,\xi_t^2)]v\right\|^2\right]\right] \\
\leq \quad & \mathsf{E}_{\xi_t^{1,2}}\left[2L^2d^2\delta^{2\beta-2}+2\frac{d^2}{\delta^2}|F_t(x,\xi_t^1)-F_t(x,\xi_t^2)|^2\right] \\
\overset{(6.40)}{\leq} \quad & 2L^2d^2\delta^{2\beta-2}+4\frac{d^2}{\delta^2}\theta^2 \qquad (6.43)
\end{aligned}$$

□

Now based on the same idea, we analyze the regret bound for the stochastic zeroth-order online gradient descent algorithm with $G_\delta(x_t,\xi_t^{1,2},v)$ as the oracle.

Theorem 6.3.5 *Consider running the stochastic zeroth-order online gradient descent method on the loss functions $f_t \in C_L^0(\mathbb{R}^d)$ which satisfy Assumption 6.3.1. Let S be a convex set and define $B = \max_{u\in S}\|u\|$. Then, for all $u \in S$ we have*

$$\begin{aligned}
& \mathsf{E}\left[\sum_{t=1}^T(f_t(x_t)-f_t(u))\right]+\mathsf{E}\left[\sum_{t=1}^T(f_t(x_t+\delta v_t)-f_t(u))\right] \\
\leq & \frac{1}{\eta}B^2+\eta(4L^2d^2\delta^{2\beta-2}+8\frac{d^2}{\delta^2}\theta^2)T+5L\delta^\beta T. \qquad (6.44)
\end{aligned}$$

In particular, if we set $\eta \sim T^{-\frac{2+\beta}{2(1+\beta)}}$, and $\delta \sim T^{-\frac{1}{2(1+\beta)}}$, then the regret is bounded by $O(T^{\frac{2+\beta}{2(1+\beta)}})$. When $\beta = 1$ we have the bound to be $O(T^{\frac{3}{4}})$.

Proof. In view of the Follow the Regularized Leader algorithm (cf. [109]), we have the following,

$$\sum_{t=1}^T\langle z_t, x_t-u\rangle \leq \frac{1}{2\eta}\|u\|_2^2+\eta\sum_{t=1}^T\|z_t\|^2. \qquad (6.45)$$

Taking expectation, and noticing that $\mathsf{E}_{\xi_t^{1,2},v_t}[z_t|\Xi_{t-1}] = \nabla f_t^\delta(x_t)$ which is shown in (6.41),

we have

$$\mathsf{E}\left[\sum_{t=1}^{T}\langle\nabla f_t^{\delta}(x_t), x_t - u\rangle\right] \leq \frac{1}{2\eta}\|u\|_2^2 + \eta\sum_{t=1}^{T}\mathsf{E}\left[\|z_t\|^2\right]. \tag{6.46}$$

Recall the convexity of f_t^{δ}, we have

$$\mathsf{E}\left[\sum_{t=1}^{T}(f_t^{\delta}(x_t) - f_t^{\delta}(u))\right] \leq \frac{1}{2\eta}\|u\|_2^2 + \eta\sum_{t=1}^{T}\mathsf{E}\left[\|z_t\|^2\right]. \tag{6.47}$$

From the property (6.26) of the smoothing function, we have the following chain of inequality

$$\begin{aligned} &f_t(x_t) - f_t(u) \leq f_t^{\delta}(x_t) - f_t^{\delta}(u) + 2L\delta^{\beta}, \\ &f_t(x_t + \delta v_t) - f_t(u) \leq f_t(x_t) - f_t(u) + L\delta^{\beta} \leq f_t^{\delta}(x_t) - f_t^{\delta}(u) + 3L\delta^{\beta}. \end{aligned} \tag{6.48}$$

By combining (6.48) and (6.47), we have

$$\begin{aligned} &\mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t) - f_t(u))\right] + \mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t + \delta v_t) - f_t(u))\right] \\ \leq &\frac{1}{\eta}\|u\|_2^2 + 2\eta\sum_{t=1}^{T}\mathsf{E}\left[\|z_t\|^2\right] + 5L\delta^{\beta}T. \end{aligned} \tag{6.49}$$

Now, from (6.42), we can bound $\sum_{t=1}^{T}\mathsf{E}\left[\|z_t\|^2\right]$ as

$$\sum_{t=1}^{T}\mathsf{E}\left[\|z_t\|^2\right] \leq (2L^2d^2\delta^{2\beta-2} + 4\frac{d^2}{\delta^2}\theta^2)T.$$

Let $B = \max_{u\in S}\|u\|$, we have

$$\begin{aligned} &\mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t) - f_t(u))\right] + \mathsf{E}\left[\sum_{t=1}^{T}(f_t(x_t + \delta v_t) - f_t(u))\right] \\ \leq &\frac{1}{\eta}B^2 + \eta(4L^2d^2\delta^{2\beta-2} + 8\frac{d^2}{\delta^2}\theta^2)T + 5L\delta^{\beta}T. \end{aligned} \tag{6.50}$$

In order to achieve the best possible bound in terms of T, we need to choose η and δ carefully. Suppose we have $\eta \sim T^{-a}$, and $\delta \sim T^{-b}$, where $a, b \geq 0$. Then the order in

(6.50) becomes to

$$T^{\max\{a,1-a-(2\beta-2)b,1+2b-a,1-\beta b\}}.$$

Thus, we need to find the value of

$$\min_{a,b\geq 0} \max\left\{a, 1-a-(2\beta-2)b, 1+2b-a, 1-\beta b\right\}. \tag{6.51}$$

Since $0 < \beta \leq 1$, we have $1+2b-a \geq 1-a-(2\beta-2)b$, so the above problem becomes to

$$\min_{a,b\geq 0} \max\left\{a, 1+2b-a, 1-\beta b\right\}. \tag{6.52}$$

By inspection, we can find the optimal a, b as $a = \frac{2+\beta}{2(1+\beta)}$, $b = \frac{1}{2(1+\beta)}$. In particular, if we set $\eta \sim T^{-\frac{2+\beta}{2(1+\beta)}}$, and $\delta \sim T^{-\frac{1}{2(1+\beta)}}$, then the regret is bounded by $O(T^{\frac{2+\beta}{2(1+\beta)}})$ which is the best possible in terms of T. When $\beta = 1$, the regret bound becomes to $O(T^{\frac{3}{4}})$. □

Chapter 7

Conclusions and Discussions

In this dissertation, we studied the convergence properties and the applications of the first-order and the zeroth-order optimization algorithms. Our discussions include: the iteration complexity analysis of different ADMM-type algorithms for solving various multi-block optimization with linear constraint, and the analysis of lower-order gradient-type algorithms for solving oracle-based black-box optimization and online learning problem. From the theoretical point of view, without sacrificing the computational complexity bounds, the zeroth-order smoothing scheme enables different algorithms to be applicable on a much broader class of problems where only noisy estimations of the function values are available. Moreover, as we showed in Chapter 4, the randomization is really the key to establish the convergence rate result for the multi-block ADMM method. Together with a carefully selected proximal term, the parallelization makes our randomized algorithm even more efficient and powerful. From the practical point of view, our numerical studies also indicate that our proposed algorithms are indeed comparable to those state-of-the-art methods by means of evaluation using well-established standard benchmark problems, while the theoretical convergence rate is also achieved.

There are several directions for future research. In terms of the convergence rate, it is interesting to explore if it is possible to accelerate those lower-order ADMM-type algorithms, especially for the zeroth-order method. Moreover, the zeroth-order smoothing scheme is a powerful tool for constructing approximations of the gradient. Designing algorithms for an optimization model often amounts to maintaining a balance between the degree of information to request from the model on the one hand, and the computational speed to expect on the other hand. Naturally, the more information is available, the faster one can expect the algorithm to converge. Thus, how to generalize the zeroth-

order smoothing scheme to approximate higher-order derivatives so as to achieve faster convergence rate is another possible direction to explore.

References

[1] J. Abernethy, A. Agarwal, P. L. Bartlett, and A. Rakhlin. A Stochastic View of Optimal Regret through Minimax Duality. *arXiv preprint arXiv:0903.5328*, 2009.

[2] A. Agarwal, O. Dekel, and L. Xiao. Optimal Algorithms for Online Convex Optimization with Multi-Point Bandit Feedback. In *COLT*, pages 28–40. Citeseer, 2010.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[4] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

[5] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. Siam, 2001.

[6] D. P. Bertsekas. Nonlinear programming. 1999.

[7] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and distributed computation: numerical methods*, volume 23. Prentice-Hall, Inc., 1989.

[8] D. Boley. Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs. *SIAM Journal on Optimization*, 23(4):2183–2207, 2013.

[9] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146:459–494, 2014.

[10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends®in Machine Learning*, 3(1):1–122, 2011.

[11] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[12] E. Brochu, V. M. Cora, and N. De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

[13] X. Cai, D. Han, and X. Yuan. On the convergence of the direct extension of ADMM for three-block separable convex minimization models with one strongly convex function. *Computational Optimization and Applications*, 66(1):39–73, 2017.

[14] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[15] 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.

[16] C. Chen, B. He, Y. Ye, and X. Yuan. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, 155(1-2):57–79, 2016.

[17] C. Chen, Y. Shen, and Y. You. On the convergence analysis of the alternating direction method of multipliers with three blocks. In *Abstract and Applied Analysis*, volume 2013. Hindawi, 2013.

[18] C. H. Chen, M. Li, X. Liu, and Y. Y. Ye. On the convergence of multi-block alternating direction method of multipliers and block coordinate descent method. *arXiv preprint arXiv:1508.00193*, 2015.

[19] L. Chen, D. Sun, and K.-C. Toh. An efficient inexact symmetric Gauss–Seidel based majorized ADMM for high-dimensional convex composite conic programming. *Mathematical Programming*, 161(1-2):237–270, 2017.

[20] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998.

[21] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to derivative-free optimization*, volume 8. Siam, 2009.

[22] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[23] Y. Cui, X. Li, D. Sun, and K.-C. Toh. On the convergence properties of a majorized ADMM for linearly constrained convex optimization problems with coupled objective functions. *arXiv preprint arXiv:1502.00098*, 2015.

[24] C. Dang and G. Lan. Randomized first-order methods for saddle point optimization. *arXiv preprint arXiv:1409.8625*, 2014.

[25] C. D. Dang and G. Lan. Stochastic block mirror descent methods for nonsmooth and stochastic optimization. *SIAM Journal on Optimization*, 25(2):856–881, 2015.

[26] D. Davis and W. Yin. Convergence rate analysis of several splitting schemes. *arXiv preprint arXiv:1406.4834*, 2014.

[27] W. Deng, M. J. Lai, Z. Peng, and W. Yin. Parallel Multi-Block ADMM with o(1/k) Convergence. *Journal of Scientific Computing*, 71(2):712–736, 2017.

[28] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. 2012.

[29] D. L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[30] J. Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American mathematical Society*, 82(2):421–439, 1956.

[31] Y. Drori, S. Sabach, and M. Teboulle. A simple algorithm for a class of nonsmooth convex–concave saddle-point problems. *Operations Research Letters*, 43(2):209–214, 2015.

[32] A. DAspremont, O. Banerjee, and L. E. Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications*, 30(1):56–66, 2008.

[33] 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.

[34] Y. Ermoliev. Stochastic quasigradient methods and their application to system optimization. *An International Journal of Probability and Stochastic Processes*, 9(1-2):1–36, 1983.

[35] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394. Society for Industrial and Applied Mathematics, 2005.

[36] J. B. G. Frenk, J. Gromicho, and S. Zhang. A deep cut ellipsoid algorithm for convex programming: Theory and applications. *Mathematical Programming*, 63(1-3):83–108, 1994.

[37] A. A. Gaivoronskii. Nonstationary stochastic programming problems. *Cybernetics and Systems Analysis*, 14(4):575–579, 1978.

[38] X. Gao, B. Jiang, and S. Zhang. On the Information-Adaptive Variants of the ADMM: An Iteration Complexity Perspective. *Journal of Scientific Computing*, pages 1–37, 2017.

[39] X. Gao and S. Zhang. First-order algorithms for convex optimization with non-separable objective and coupled constraints. *Journal of the Operations Research Society of China*, 5(2):131–159, 2017.

[40] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[41] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

[42] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[43] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2015.

[44] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *arXiv preprint arXiv:1308.6594*, 155(1):267–305, 2014.

[45] R. Glowinski and P. Le Tallec. *Augmented Lagrangian and operator-splitting methods in nonlinear mechanics*, volume 9. SIAM, 1989.

[46] R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(2):41–76, 1975.

[47] D. Han and X. Yuan. A note on the alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, 155(1):227–238, 2012.

[48] D. Han and X. Yuan. Local linear convergence of the alternating direction method of multipliers for quadratic programs. *SIAM Journal on numerical analysis*, 51(6):3446–3457, 2013.

[49] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[50] E. Hazan, K. Levy, and S. Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems*, pages 1594–1602, 2015.

[51] B. He, L.-Z. Liao, D. Han, and H. Yang. A new inexact alternating directions method for monotone variational inequalities. *Mathematical Programming*, 92(1):103–118, 2002.

[52] B. He, M. Tao, and X. Yuan. Alternating direction method with Gaussian back substitution for separable convex programming. *SIAM Journal on Optimization*, 22(2):313–340, 2012.

[53] B. He, H.-K. Xu, and X. Yuan. On the proximal Jacobian decomposition of ALM for multiple-block separable convex minimization problems and its relationship to ADMM. *Journal of Scientific Computing*, 66(3):1204–1217, 2016.

[54] B. S. He, L. Hou, and X. Yuan. On full Jacobian decomposition of the augmented Lagrangian method for separable convex programming. 2013.

[55] B. S. He, M. Tao, and X. Yuan. Convergence rate and iteration complexity on the alternating direction method of multipliers with a substitution procedure for separable convex programming. *Math. Oper. Res., under revision*, 2012.

[56] B. S. He and X. Yuan. On non-ergodic convergence rate of douglas-rachford alternating direction method of multipliers. 2012.

[57] B. S. He and X. 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.

[58] M. Hong, T.-H. Chang, X. Wang, M. Razaviyayn, S. Ma, and Z.-Q. Luo. A Block Successive Upper Bound Minimization Method of Multipliers for Linearly Constrained Convex Optimization. *arXiv preprint arXiv:1401.7079*, 2014.

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

[60] M. Hong, X. Wang, M. Razaviyayn, and Z.-Q. Luo. Iteration complexity analysis of block coordinate descent methods. *arXiv preprint arXiv:1310.6957*, 2013.

[61] G. M. James, C. Paulson, and P. Rusmevichientong. The constrained lasso. In *Refereed Conference Proceedings*, volume 31, pages 4945–4950. Citeseer, 2012.

[62] G. M. James, C. Paulson, and P. Rusmevichientong. Penalized and constrained regression. Technical report, mimeo, Marshall School of Business, University of Southern California, 2013.

[63] V. T. Jonas Mockus and A. Zilinskas. The application of bayesian methods for seeking the extremum. *Towards Global Optimization*, 2:117–129, 1978.

[64] D. R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21(4):345–383, 2001.

[65] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[66] J. C. Lagarias, B. Poonen, and M. H. Wright. Convergence of the Restricted Nelder–Mead Algorithm in Two Dimensions. *SIAM Journal on Optimization*, 22(2):501–532, 2012.

[67] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on optimization*, 9(1):112–147, 1998.

[68] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

[69] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[70] M. Li, D. Sun, and K.-C. Toh. A convergent 3-block semi-proximal ADMM for convex minimization problems with one strongly convex block. *Asia-Pacific Journal of Operational Research*, 32(04):1550024, 2015.

[71] X. Li, D. Sun, and K.-C. Toh. A Schur complement based semi-proximal ADMM for convex quadratic conic programming and extensions. *Mathematical Programming*, 155(1-2):333–373, 2016.

[72] T. Lin, S. Ma, and S. Zhang. An extragradient-based alternating direction method for convex minimization. *Foundations of Computational Mathematics*, pages 1–25, 2015.

[73] T. Lin, S. Ma, and S. Zhang. On the Global Linear Convergence of the ADMM with Multi-Block Variables. *SIAM Journal on Optimization*, 25(3):1478–1497, 2015.

[74] T. Lin, S. Ma, and S. Zhang. On the Sublinear Convergence Rate of Multi-block ADMM. *Journal of the Operations Research Society of China*, 3(3):251–274, 2015.

[75] T. Lin, S. Ma, and S. Zhang. Iteration Complexity Analysis of Multi-block ADMM for a Family of Convex Minimization Without Strong Convexity. *Journal of Scientific Computing*, 69(1):52–81, 2016.

[76] J. Liu, J. Chen, and J. Ye. Large-scale sparse logistic regression. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–556. ACM, 2009.

[77] Z. Lu and L. Xiao. A Randomized Nonmonotone Block Proximal Gradient Method for a Class of Structured Nonlinear Programming. *arXiv preprint arXiv:1306.5918*, 2013.

[78] Z. Lu and L. Xiao. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152(1-2):615–642, 2015.

[79] Z.-Q. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992.

[80] R. D. C. Monteiro and B. F. Svaiter. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization*, 23(1):475–507, 2013.

[81] A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *Automatic Control, IEEE Transactions on*, 54(1):48–61, 2009.

[82] J. A. Nelder and R. Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[83] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[84] A. S. Nemirovski. Prox-method with rate of convergence O(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2005.

[85] A. S. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

[86] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. Akad. Nauk SSSR*, pages 543–547, 1983.

[87] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media, 2004.

[88] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[89] Y. Nesterov. Random gradient-free minimization of convex functions. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2011.

[90] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[91] Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[92] M. K. Ng, F. Wang, and X. Yuan. Inexact alternating direction methods for image recovery. *SIAM Journal on Scientific Computing*, 33(4):1643–1668, 2011.

[93] H. Ouyang, N. He, L. Tran, and A. Gray. Stochastic alternating direction method of multipliers. In *Proceedings of the 30th International Conference on Machine Learning*, pages 80–88, 2013.

[94] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2233–2246, 2012.

[95] Z. Peng, T. Wu, Y. Xu, M. Yan, and W. Yin. Coordinate friendly structures, algorithms and applications. *Annals of Mathematical Sciences and Applications*, 1(1):57–119, 2016.

[96] B. Polyak. New stochastic approximation type procedures. *Automat. i Telemekh*, 7(98-107):2, 1990.

[97] B. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[98] B. T. Polyak. Introduction to optimization. translations series in mathematics and engineering. *Optimization Software*, 1987.

[99] M. J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, 7(2):155–162, 1964.

[100] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

[101] P. Richtárik and M. Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1-2):433–484, 2016.

[102] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[103] R. T. Rockafellar. Large-scale extended linear-quadratic programming and multistage optimization. *Advances in Numerical Partial Differential Equations and Optimization*, pages 247–261, 1991.

[104] A. Ruszczyński and W. Syski. A method of aggregate stochastic subgradients with on-line stepsize rules for convex stochastic programming problems. In *Stochastic Programming 84 Part II*, pages 113–131. Springer, 1986.

[105] J. Sacks. Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, pages 373–405, 1958.

[106] A. Saha and A. Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *International Conference on Artificial Intelligence and Statistics*, pages 636–642, 2011.

[107] K. Scheinberg, S. Ma, and D. Goldfarb. Sparse inverse covariance selection via alternating linearization methods. In *Advances in neural information processing systems*, pages 2101–2109, 2010.

[108] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

[109] S. Shalev-Shwartz. Online Learning and Online Convex Optimization. *Foundations and Trends®in Machine Learning*, 4(2):107–194, 2011.

[110] R. Shefi and M. 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.

[111] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[112] D. Sun, K.-C. Toh, and L. Yang. A convergent 3-block semiproximal alternating direction method of multipliers for conic programming with 4-type constraints. *SIAM journal on Optimization*, 25(2):882–915, 2015.

[113] R. Sun, Z.-Q. Luo, and Y. Ye. On the expected convergence of randomly permuted ADMM. *arXiv preprint arXiv:1503.06387*, 2015.

[114] T. Suzuki. Dual Averaging and Proximal Gradient Descent for Online Alternating Direction Multiplier Method. In *Proceedings of the 30th International Conference on Machine Learning*, pages 392–400, 2013.

[115] T. Suzuki. Stochastic Dual Coordinate Ascent with Alternating Direction Method of Multipliers. In *Proceedings of the 31th International Conference on Machine Learning*, pages 736–744, 2014.

[116] M. Tao and X. Yuan. Recovering low-rank and sparse components of matrices from incomplete and noisy observations. *SIAM Journal on Optimization*, 21(1):57–81, 2011.

[117] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[118] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused LASSO. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108, 2005.

[119] R. J. Tibshirani and J. Taylor. The solution path of the generalized LASSO. *Annals of statistics*, 39(3):1335–1371, 2011.

[120] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *manuscript*, 2008.

[121] H. Wang and A. Banerjee. Online Alternating Direction Method. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[122] M. Wang, E. X. Fang, and H. Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017.

[123] Y. Wang, W. Yin, and J. Zeng. Global Convergence of ADMM in Nonconvex Nonsmooth Optimization. *arXiv preprint arXiv:1511.06324*, 2015.

[124] Y. Xu. Hybrid Jacobian and Gauss–Seidel Proximal Block Coordinate Update Methods for Linearly Constrained Convex Programming. *SIAM Journal on Optimization*, 28(1):646–670, 2018.

[125] Y. Xu and W. Yin. A Block Coordinate Descent Method for Regularized Multiconvex Optimization with Applications to Nonnegative Tensor Factorization and Completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013.

[126] Y. Xu and W. Yin. Block stochastic gradient iteration for convex and nonconvex optimization. *SIAM Journal on Optimization*, 25(3):1686–1716, 2015.

[127] Y. Xu and S. Zhang. Accelerated primal–dual proximal block coordinate updating methods for constrained convex optimization. *Computational Optimization and Applications*, 70(1):91–128, 2018.

[128] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[129] P. Zhao, J. Yang, T. Zhang, and P. Li. Adaptive Stochastic Alternating Direction Method of Multipliers. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 69–77, 2015.

[130] W. Zhong and J. T.-Y. Kwok. Fast Stochastic Alternating Direction Method of Multipliers. In *Proceedings of the 31th International Conference on Machine Learning*, pages 46–54, 2014.

[131] M. Zinkevich. Online Convex Programming and Generalized Infinitesimal Gradient Ascent. *Machine Learning*, 20(February):421–422, 2003.