STUDIES ON
BLOCK COORDINATE GRADIENT METHODS
FOR NONLINEAR OPTIMIZATION PROBLEMS
WITH SEPARABLE STRUCTURE

XIAOQIN HUA

STUDIES ON BLOCK COORDINATE GRADIENT METHODS FOR NONLINEAR OPTIMIZATION PROBLEMS WITH SEPARABLE STRUCTURE

by

XIAOQIN HUA

Submitted in partial fulfillment of
the requirement for the degree of
DOCTOR OF INFORMATICS
(Applied Mathematics and Physics)

KYOTO UNIVERSITY
KYOTO 606–8501, JAPAN
DECEMBER 2014

Preface

In this thesis, we study the block coordinate gradient methods for two kinds of the nonlinear optimization problems with separable structure: the classical optimization problem and the online optimization problem. These two optimization problems are highly constructed models arising from practical problems in science and engineering, and the corresponding applications are typically built on large scales. Hence, proposing efficient and practical solution methods to solve them is a worth studying topic.

Due to the large scales, the classical second order methods, such as the Newton method, the interior point method, can not be applied successfully. Practical experiments indicate that the first order methods are more efficient. The block coordinate descent (BCD) method is one of the oldest first order methods, whereby a part of the variable is updated at each iteration. It is very efficient for large scale problems. However, its convergence requires restrictive conditions, such as the strict convexity and the differentiability. Thus, this method did not get too much attention in the mathematical optimization field. Recently, as many large scale problems in engineering arise, such as the machine learning, the image reconstruction, etc., the block type methods have been revived. Although there are some existing results on the block type methods for large scale problems, there still remain unknown problems. For example, most of the existing results are established on the assumption that the subproblem is solved exactly. This assumption is difficult to satisfy in practice. How about the case where the subproblem is solved inexactly? Moreover, the convergence rate of the BCD method for the nonsmooth problem is still unknown.

The main contribution of this thesis is to propose efficient block coordinate gradient methods for solving large scale nonlinear optimization problems with separable structure. We propose two novel classes of methods. One is the inexact coordinate descent method, where we give a new criterion for the inexact solution of the subproblem and only require an approximate solution at each iteration. The other method is a class of block coordinate proximal gradient methods with variable Bregman functions. In this class of methods, using the variable kernels is the innovation, which offers great advantages to both algorithm analysis and practical implementation. We establish its global and R-linear convergence rate for the nonconvex nonsmooth problem. With special kernels, we even show the R-linear convergence rate of the (inexact) BCD method, which is the first result on the linear convergence of the BCD method for the nonsmooth problem. Moreover, for both proposed methods, some numerical experiments have been carried out, which demonstrate the excellent performances of the proposed methods.

Another contribution of this thesis is to propose a new Lipschitz continuity-like definition, called the "block lower triangular Lipschitz continuous", which helps us to supplement and improve the theoretical analysis of the block coordinate gradient (BCG) method. In particular, we obtain a tighter iteration complexity bound of the BCG method for the nonlinear convex optimization problem with separable structure and improve the convergence rate of the BCG method for the online and the stochastic optimization problem.

The author hopes that the results in this thesis will contribute for the further studies on the block type solution methods for the nonlinear optimization problems and their related problems.

Xiaoqin Hua
December 2014

Acknowledgments

I would like to show my deepest gratitude to my supervisor, Professor Nobuo Yamashita of Kyoto university. It is my honor to be a student of this excellent optimization laboratory. I thank him for providing opportunity for my family reunion and the patient guidance, encouragement and advice throughout my time as his student. During these years, he spent a lot of time answering my trivial questions and guided me on how to carry out the research. He read my draft manuscript carefully and gave me constructive comments. His meticulous attitude in scientific studies has impressed me deeply, which will benefit me in my life.

I would like to express my heartfelt thanks to Professor Masao Fukushima of Nanzan University, who accepted me as a visiting student in this laboratory, and led me into the world of optimization. He also gave me a lot of advice on the research approach. My family and I are really grateful to him.

I am also thankful to Assistant Professor Ellen Hidemi Fukuda of Kyoto University for all her kindness and help. When I was depressed and frustrated in the study, she encouraged me greatly. She also helped me with my Japanese and corrected my pronunciations and grammars.

I would express my acknowledgement to Professor Shunsuke Hayashi of Tohoku University. With his kindness and help, I could share the excellent cuisine in the journey of attending the society meetings. I am thankful for Professor Zhi-Quan Luo of Department of Electrical and Computer Engineering in University of Minnesota for giving me a chance to study in his laboratory. I am also greatly indebted to Professor Jiang and other teachers of Jiangsu University of Science and Technology for their understanding and full support for my research.

I would like to tender my acknowledgement to all the members of Yamashita Laboratory and former members of Fukushima Laboratory, Dr. Yuya Yamakawa, Mr. Takayuki Okuno, Mr. Masataka Nishimori, and others for their direct and indirect help to me. Especially, I thank Ms. Fumie Yagura and Ms. Yuiko Igura, who have rendered me much help on my work and life.

I must express my gratitude to my husband and my lovely daughter for their considerations and great confidence in me through these years. Without their support, I could not complete this thesis. I also would like to deliver my gratitude to my parents and my

parents-in-law, relatives and friends for their continued support and encouragement.

I would like to express my deep thanks to my Pastor and his wife, and all brothers and sisters of the Kyoto Missionary Church. They gave my family and me a lot of encouragement, prayer and support.

Finally, I would like to give back all the glory, honor, praise, majesty, power and authority to Jesus Christ our LORD. For without Him, all these would not have been possible and meaningful.

Contents

List of Figures

List of Tables

List of Notations

$\|x\|$	the absolute value of number x
$\lfloor x \rfloor$	the largest integer not greater than number x
$\mathcal{N}$	the set of $\{1, 2, \cdots, n\}$
J	a subset of $\mathcal{N}$
$\overline{J}$	the complement set of set J with respect to $\mathcal{N}$
$\|J\|$	the cardinality of set J
$\mathrm{int}S$	the interior of set S
$\mathrm{dom}f$	the effective domain of function f
x_i	the i-th coordinate of vector x
x_J	the subvector $(x_j \mid j \in J)$
$x_{\overline{J}}$	the subvector $(x_j \mid j \in \overline{J})$
$\\|x\\|_1$	the 1-norm of vector x
$\\|x\\|_2$	the 2-norm of vector x
$\\|x\\|_G$	the G-norm of vector x
$\langle \cdot, \cdot \rangle$	the Euclidean inner product
∇h	the gradient of h
$\nabla^2 h$	the Hessian matrix of h
$\nabla_i h$	the i-th coordinate of gradient ∇h
∂h	the subdifferential of function h
$\partial_J h$	the subdifferential of function h with respect to x_J
$\exp^x$	the exponential function
A^T	the transpose of matrix A
A_j	the j-th column matrix A
A_{ij}	the element at the i-th row and j-th column of matrix A
$A \succeq 0$	matrix A is positive semidefinite
$A \succeq B$	matrix $A - B$ is positive semidefinite

Chapter 1

Introduction

The *mathematical optimization* is one of the mature areas of the applied mathematics, which aims at finding a solution to a given model by optimization algorithms or methods. It was introduced by professor Robert Dorfman in 1940s, and became more and more active with the rapid development of the computer technology. Recently, its theories and techniques are used widely in the industrial designs, the computer science and the economics management.

In this chapter, we give an overview of the research problems, their characteristics, the research motivations and contributions. We refer to some basic optimization terminologies and names of the algorithms directly, whose precise definitions are given in Chapter 2.

1.1 Nonlinear optimization problem

The *nonlinear optimization problem* [8, 13, 41] is a main subfield of the mathematical programming, which has the following general form.

$$\begin{aligned} \text{minimize} \quad & F(x) \\ \text{subject to} \quad & h_i(x) = 0, i \in \{1, \dots, p\}, \\ & g_j(x) \leq 0, j \in \{1, \dots, m\}, \end{aligned} \tag{1.1.1}$$

where $x \in \mathcal{R}^n$, function $F : \mathcal{R}^n \to \mathcal{R}$ is the objective function (alternatively, the loss function or the cost function), and functions h_i, $g_j : \mathcal{R}^n \to \mathcal{R}$, $i = 1, \dots, p$, $j = 1, \dots, m$, are the constraint functions. The set

$$\mathcal{F} := \{x \in \mathcal{R}^n \mid h_i(x) = 0, g_j(x) \leq 0, i = 1, \dots, p, j = 1, \dots, m\} \tag{1.1.2}$$

is called the feasible set.

Note that problem (1.1.1) may or may not be a convex problem. In general, there are significant differences in the characteristics of the solutions between the convex and the nonconvex cases. The *convex optimization problem* [6, 13] is relatively simple, in which both objective and constraint are convex functions. It has the following two important characteristics [8]. One is that any local minimum of the convex optimization problem is also the global minimum. Of course, its local minimum solution may not be unique. Another is that the first order optimality condition (Theorems 2.2.1 and 2.2.2 in Subsection 2.2.3) is also sufficient for guaranteeing the optimality. These two characteristics are the foundations of the analysis of the theories and algorithms for the convex optimization problems. Moreover, if additional conditions are satisfied for problem (1.1.1), the existence and uniqueness of the optimal solutions can be guaranteed. For example, if the objective function is strictly convex and the feasible set is convex, there exists at most one optimal minimum. Additionally, if the feasible set is compact, there exists only one minimum. See [8, Proposition A.8] for details.

For the general *nonconvex optimization problem* [15, 34], in which the objective function is nonlinear and/or the feasible region is determined by nonlinear constraints, the optimal solution may not always exist, or it may have multiple locally optimal solutions. Hence, the solutions in this case are more complicated. Generally, we study the stationary points instead.

In addition, according to the differentiability of the functions in problem (1.1.1), the nonlinear optimization problem can be divided into the smooth optimization problem and the nonsmooth optimization problem. The most well known algorithm for the nonlinear smooth optimization problem is the Newton method, which is extremely powerful in general. Even today, Newton method is still the most widely used and studied algorithm. However, this method is not perfect. For example, at each iteration we need to compute the Hessian of the objective function, which needs $O(n^2)$ computation, generally. As the scale n becomes large, the computation at each iteration will be very expensive. For this reason, it is not worth to be applied to the large scale problems directly. For the nonsmooth optimization problems, we usually try to find the source of the "nonsmoothness", and further develop some techniques for its particular structure.

1.2 Nonlinear optimization problems with separable structure

In this thesis, we consider the *nonlinear optimization problems with separable structure*, which are highly structured optimization models. In these models, we minimize the sum of a smooth function and a "simple" nonsmooth convex function, where the simple convex function has block separable structure. Hence, they belong to the nonsmooth subfield of the nonlinear optimization problem. In this thesis, we study two kinds of these nonlinear optimization problems: the *classical nonlinear optimization problem* and the *online optimization problem*, which are briefly outlined in the subsequent subsections.

1.2.1 Nonlinear optimization problem with separable structure

The *nonlinear optimization problem with separable structure*, considered in this thesis, has the following form.

$$\underset{x}{\text{minimize}} \;\; F(x) := f(x) + \tau\psi(x), \tag{1.2.1}$$

where function f is smooth on an open subset of $\mathcal{R}^n$ containing $\operatorname{dom}\psi := \{x \in \mathcal{R}^n \mid \psi(x) < \infty\}$, τ is a positive constant, and $\psi : \mathcal{R}^n \to (-\infty, \infty]$ is a proper, convex and lower semi-continuous (l.s.c.) function with the block separable structure [1].

Such problems arise in various practical problems of science and engineering, such as the machine learning [35, 77], the data mining [55] and the network routing [19]. When function f in problem (1.2.1) is convex, problem (1.2.1) becomes a convex optimization problem. In this thesis, we propose several efficient methods for problem (1.2.1) with particular forms, including the convex and nonconvex cases.

1.2.2 Online optimization problem with separable structure

The *online optimization problem* is a powerful learning model, which has attracted great attention in many large scale optimization fields, such as the machine learning [3], the network routing [3], and the investment decisions [27]. By this model, a decision maker makes a sequence of accurate decisions for his/her practical problems, where his/her possible options are given as a convex set in advance. The precise definition of the online convex optimization problem is recalled by Definition 2.2.8 in Subsection 2.2.4. Roughly speaking, its main characteristics include the following two aspects in contrast to the classical nonlinear optimization problem.

- We minimize a sequence of dynamically generated loss functions $\{F^t(x),\, t = 1, 2, \dots\}$ in the online optimization problem, where t denotes the time step when new function is generated.

- We must make a decision at the time step t, denoted by x^t, before getting the true loss function $F^t(x)$.

In this thesis, we consider an online convex optimization problem with separable structure, whose loss function $F^t : \Omega \to \mathcal{R}$ at time step t is given as follows.

$$F^t(x) := f^t(x) + \tau\psi(x), \;\; t = 1, 2, \dots, \tag{1.2.2}$$

[1] We say that function ψ is block separable with respect to nonempty subset $J \subseteq \{1, 2, \dots, n\}$ if there exist some proper, convex, l.s.c. functions $\psi_J : \mathcal{R}^{|J|} \to \mathcal{R}$ and $\psi_{\overline{J}} : \mathcal{R}^{|\overline{J}|} \to \mathcal{R}$ such that $\psi(x) = \psi_J(x_J) + \psi_{\overline{J}}(x_{\overline{J}})$ holds for all $x \in \mathcal{R}^n$.

where $\Omega \subseteq \bigcap_{t=1}^{\infty} \operatorname{dom} F^t$, $f^t : \Omega \to \mathcal{R}$ is smooth and convex, τ is a positive constant, and $\psi : \Omega \to (-\infty, \infty]$ is a proper, convex and lower semicontinuous (l.s.c.) function with the block separable structure.

From the characteristics of the online convex optimization problem, we know that it is impossible to select a point x^t that exactly minimizes the loss function $F^t(x)$ at the t-th time step, because we do not know the true loss function $F^t(x)$ until the prediction x^t is determined. Instead, the researchers, who study the online optimization problem, focus on proposing an algorithm to generate predictions $\{x^t, t = 1, 2, \dots\}$, with which, for given $T > 0$, the practical total loss $\sum_{t=1}^{T} F^t(x^t)$ is not much larger than the ideal total loss $\sum_{t=1}^{T} F^t(x^*)$, where x^* is an optimal solution in some sense, e.g., $x^* \in \operatorname{argmin}_{x\in\Omega} \frac{1}{T} \sum_{t=1}^{T} F^t(x)$ [81]. For convenience, we call the difference between these two values the "regret" [81], which is formally defined by Definition 2.2.9 in Subsection 2.2.4. Hence, the goal of the online convex optimization problem is to construct an algorithm, with which the generating decisions make us to achieve a regret as low as possible. We say that an algorithm is a no internal regret algorithm if the regret $R(T)$ is an infinitesimal of higher order than T [27].

1.2.3 Applications

The problems (1.2.1) and (1.2.2) appear in many applications. Usually, functions f and f^t represent as empirical loss functions, and function $\psi(x)$ acts as a regularization term to introduce additional information or to prevent overfitting. Some examples of functions f and f^t in applications are described as follows.

(1) In the compressed sensing [77], which is a classical nonlinear optimization problem, function f represents the error between the noiseless signal and the transformation of the elementary signals. In this application, function f can be written by a quadratic function with

$$f(x) = \frac{1}{2}\|Ax - y\|_2^2,$$

where $A \in \mathcal{R}^{m\times n}$, the set $\{A_j, j = 1, 2, \dots, n\}$ comprises the elementary signals, and $y \in \mathcal{R}^m$ denotes the noiseless signal.

In the online regression problem [14, 53], function f^t is used for estimating the relationships among variables. In the simple linear regression model, function f^t can be represented by a quadratic function.

$$f^t(x) = \sum_{i=1}^{m} \left(b_i^t - v - \langle A_i^t, w\rangle\right)^2,$$

where $x = (w, v) \in \mathcal{R}^n$ with $w \in \mathcal{R}^{n-1}$ and $v \in \mathcal{R}$. For any $t > 0$, and $i = 1, \dots, m$, $b_i^t \in \mathcal{R}$, $A^t \in \mathcal{R}^{(n-1)\times m}$. The set $\{(A_i^t, b_i^t), i = 1, \dots, m\}$ denotes the data points at the t-th time step.

(2) In the data classification [35] and the data mining [55], which belong to the classical nonlinear optimization (1.1.1), function f is used for predicting the outcome of a categorical dependent variable based on many predictor variables or features. In these applications, function f is given by the logistic function.

$$f(x) = \frac{1}{m}\sum_{j=1}^{m}\log\Big(1 + \exp\big(-(w^T q^j + vp^j)\big)\Big), \tag{1.2.3}$$

where $x = (w, v) \in \mathcal{R}^n$ and $q^j = p^j z^j$. Moreover, $\{(z^j, p^j) \in \mathcal{R}^{n-1} \times \{-1, 1\}, j = 1, 2, \ldots, m\}$ is a set of training examples.

In the sequential investment problem [14], which is an online optimization problem (1.2.2), f^t denotes the logistic wealth ratio, given by a logistic function.

$$f^t(x) = -\log\sum_{i=1}^{n}(x_i^t z_i^t),$$

where vector $z^t \in \mathcal{R}^n$ with $z_i^t > 0$, $i = 1, \ldots, n$, represents the price relatives for the trading period t, and $x^t \in \mathcal{R}^n$ denotes the investment proportions on the period t, such that $\sum_{i=1}^n x_i^t = 1, x_i^t > 0, i = 1, \ldots, n$.

The most common variants of function $\psi(x)$ are listed as follows.

(1) l_1-regularization [35, 38, 64, 65, 72], i.e.,

$$\psi(x) = \|x\|_1.$$

(2) Elastic net regularization [82], i.e.,

$$\psi(x) = \lambda_1\|x\|_1 + \lambda_2\|x\|_2^2.$$

(3) Block l_2-regularization [47, 78], i.e.,

$$\psi(x) = \sum_{i=1}^{N}\|x_{\mathcal{J}^i}\|_2,$$

where $\{x_{\mathcal{J}^i},\, i = 1, 2, \ldots, N\}$ denote the disjoint subvectors of vector x.

(4) Mixed norm penalty [33, 36], i.e.,

$$\psi(x) = \|x\|_1 + \sum_{i=1}^{N}\|x_{\mathcal{J}^i}\|_2,$$

where $\{x_{\mathcal{J}^i},\, i = 1, 2, \ldots, N\}$ denote the disjoint subvectors of vector x.

(5) Indicator function. For the smooth optimization problem with simple separable constraints, e.g., box constraints [48] and separable simplex constraints [19], ψ can be rewritten as an indicator function with respect to closed separable convex set X, i.e.,

$$\psi(x) = \begin{cases} 0 & \text{if } x \in X, \\ \infty & \text{otherwise.} \end{cases} \tag{1.2.4}$$

Note that these regularization terms have their own respective characteristics in the applications. The l_1-regularization [65] helps us to get relatively sparse solutions, i.e., many elements of the variable x are 0. This is a good technique for obtaining sparse solutions, but not perfect. If the solutions are strongly correlated, the l_1-regularization does not work well. In this case, the elastic net regularization is better [82]. The block l_2-regularization [47] is an extension of the l_1-regularization. Its sparsity is obtained at the group level, that is to say, a group is picked or dropped. But it does not yield sparsity within a group. The mixed norm penalty [33, 36] yields solutions, which are sparse at both group and individual elements. Apparently, the above regularization terms have different forms. However, from the optimization point of view, all of them are special forms of models (1.2.1) and (1.2.2).

1.3 Solution methods

The applications of the separable optimization problems (1.2.1) and (1.2.2) are mostly built on a large scale. In general, the number of the variables is of order 10^4 or even higher. Roughly speaking, there are two approaches to deal with them. One is to solve an equivalent reformulation problem, which is called the indirect solution method. The other is to solve the original problem directly, which is referred to as the direct solution method. Next, we take problem (1.2.1) as an example to introduce some existing solution methods.

For the indirect solution methods, when $\varphi(x)$ is an indicator function with respect to a convex set, problem (1.2.1) is equivalent to a constrained smooth optimization. Then it can be solved by some efficient methods, such as the gradient projection method, the trust region method, the active set method, etc.

When problem (1.2.1) is an unconstrained l_1-regularization problem, i.e., $\varphi(x) = \|x\|_1$, it can be reformulated as a $2n$-dimensional bounded constrained smooth optimization problem with the following form [69].

$$\begin{aligned} \underset{y,z}{\text{minimize}} \quad & f(y-z) + \tau\langle e, y+z\rangle, \\ \text{subject to} \quad & y \geq 0, \\ & z \geq 0, \end{aligned} \tag{1.3.1}$$

where vector $e \in \mathcal{R}^n$ is defined by $e = (1, \ldots, 1)^T$. For the reformulated problem (1.3.1), many effective methods or softwares have been developed, such as the L-BFGS method [80] and the MINOS software [49]. The drawback of this type of reformulation is that the dimension of the reformulated problem (1.3.1) is $2n$, a double size of the original problem, which is unfavorable for the large scale problems.

When problem (1.2.1) is an unconstrained problem with a general regularization term, it can be reformulated as an $(n+1)$-dimensional smooth optimization problem over a closed convex set by some optimization techniques [69]. The new reformulation has the following form.

$$\begin{aligned} \underset{x,\vartheta}{\text{minimize}} \quad & F(x) := f(x) + \tau\vartheta, \\ \text{subject to} \quad & \varphi(x) \leq \vartheta. \end{aligned} \tag{1.3.2}$$

Then the reformulated problem (1.3.2) can be solved by some state of art methods theoretically, such as the interior point method [46] and the sequential quadratic programming algorithm [21]. Yet, such a reformulation still has its drawbacks. The main disadvantage is that the existing methods can not exploit the block separable structure of the original problem (1.2.1). Moreover, as the size n becomes large, the storage and the computation of the Hessian will become huge, which shows that the second order methods [21, 46] can hardly be carried out.

For the direct solution methods, in consideration of the computation time and storage, first order methods are shown to be more efficient. The existing results on algorithms or methods are developed from the following two aspects.

(a) Global convergence, i.e., the generated sequence converges to a solution of the separable problems (1.2.1) and (1.2.2) in some sense. This is the most basic topic, and the global convergence property ensures to obtain a solution from an arbitrary initial point.

(b) Convergence speed. Commonly, we evaluate the convergence speed of an algorithm for problem (1.2.1) by the following two approaches. One is the local convergence rate, such as the linear convergence, super linear convergence, and quadratic convergence, which describes the speed of obtaining a solution when the generated point is near the solution. However, the local convergence rate does not care about the whole performance of the iterative method from the initial point. The other approach is the iteration complexity, by which we estimate the order of the iterations required by the proposed method to find a solution within ε error tolerance, such as $O(1/\varepsilon)$, $O(1/\varepsilon^2)$. In contrast to the convergence rate, the iteration complexity focuses on entire information from the initial point, rather than the local behavior near the solution. For the online optimization problem (1.2.2), we evaluate the proposed algorithm for the regret by the iteration complexity in this thesis, since every decision $x^t, t = 1, 2, \ldots, T$, is important during the T time steps.

In the next subsections, we introduce existing results on the block coordinate gradient solution methods for problems (1.2.1) and (1.2.2), respectively.

1.3.1 Existing methods for problem (1.2.1)

In this subsection, we introduce existing work on the efficient solution methods for problem (1.2.1), including the latest results on their convergence rates and iteration complexities.

When functions f and φ in problem (1.2.1) have particular forms, some special solution methods are proposed. For the l_1-regularized least square problem in the compressed sensing, the software SpaRSA is well developed for finding the sparse approximate solution, which can be downloaded from *http://www.lx.it.pt/∼mtf/SpaRSA/*. For the Group LASSO, which is a generalization of the lasso for group-wise variable selection, a software program, called R package gglasso, has been implemented, which is publicly available from *http://cran.r-project.org/web/packages/gglasso*.

For a more general setting problem, due to its large size, practical experiments indicate that first order methods are more suitable.

Among them, the (*block*) *coordinate descent* (*CD*) *method*, also called the nonlinear Gauss-Seidel method, is an attractive one. In this method, the variable is partitioned into several blocks and we only update one of the blocks at every iteration, while the other blocks are held fixed. In particular, at the r-th iteration, we choose a nonempty set $J \subseteq \{1, 2, \ldots, n\}$ and obtain an update for the block vector x_J^{r+1} by

$$x_J^{r+1} = \underset{x_J \in \mathcal{R}^{|J|}}{\operatorname{argmin}} F(x_J, x_{\bar{J}}^r). \tag{1.3.3}$$

The subproblem (1.3.3) is a $|J|$-dimensional problem. When $|J| \ll n$, its computation is much cheaper than the batch type method [2]. When $|J| = 1$ for each r, the block coordinate descent method reduces to the coordinate descent method, and it can be solved quickly by some second order methods, such as the Newton method and the quasi-Newton method. When $|J| = n$, problem (1.3.3) is the same as the original optimization problem.

The greatest advantage of the (block) CD method is that the storage requirement of the calculation is small. In some special cases, it can be implemented in parallel. Due to these properties, the block CD has been used for various large scale problems [7, 38, 43, 58, 59, 67, 69, 72, 78].

However, the global convergence of the (block) CD method requires restrictive conditions. Mainly, it depends on two factors. One is the order (alternatively, the rule) of choosing the block J at each iteration. The typical rules to choose block J are the Gauss-Seidel rule

[2] The optimization method is said to be a batch type method if it updates all elements of the variable together at a time.

[67], the Gauss-Southwell rule [69] and the random rule [50]. For details, see Section 2.4. Note that the Gauss-Southwell rule is less appealing than the Gauss-Seidel rule, because it requires the knowledge of the full gradient. For the random rule, the global convergence of an algorithm is obtained in terms of statistic.

Another critical factor is the inherent property of the objective function F. Generally, the global convergence of the CD method can not be guaranteed even for the smooth or convex optimization problem. When $\psi(x) = 0$ for any $x \in \operatorname{dom} F$, i.e., problem (1.2.1) is a smooth problem, and the function f is not (pseudo) convex, Powell gave an example to show that the CD method may not approach any stationary point [57]. When the cost function is not differentiable, Auslender showed that the CD method may stagnate at a nonstationary point even when it is a convex problem [1]. Therefore, in general, it is difficult to show the global convergence of the CD method for an optimization problem, when it is neither convex nor smooth. The existing results on the convergence of the (block) CD method are mostly developed for some particular cases. For example, for the smooth optimization, if the cost function is a strictly convex (or quasiconvex or hemivariate) function, it is shown in [43] that the CD method is convergent. For the nonsmooth problem, when the nondifferentiable part of the cost function is separable, Tseng [67] proved that the block coordinate descent method is convergent under certain convexity and regularity assumptions. When problem (1.2.1) is a convex problem, and function ψ is an indicator function with respect to a special box constraint $x \geq 0$, Luo and Tseng [43] proved that the block CD method has global and linear convergence rate. For general separable problem (1.2.1), its convergence rate is still unknown.

Moreover, the existing results on the global convergence of the CD method are established on the assumption that the subproblem (1.3.3) is solved exactly [43, 67]. It is possible for special problems, such as the l_1-l_2 problem, but hard for the general separable optimization problem (1.2.1), even if it is a l_1-regularized convex problem. To get around this difficulty, some variants of the CD method have been proposed, such as the inexact block coordinate descent method [11], the block coordinate gradient descent (BCGD) method [69] and the block coordinate proximal point method [74]. The BCGD method is executed with one step of the gradient method for the subproblem (1.3.3), while the method [74] exploits the proximal point method to find an approximate solution. Thus, they are regarded as the inexact CD methods. Bonettini [11] proposed an inexact version of the CD method. He gave appropriate conditions about the inexactness of the solution for the subproblem (1.3.3), and has shown that the proposed method with the proposed conditions has global convergence. However, he only focused on the smooth optimization problem, i.e., $\psi(x) = 0$, for all $x \in \operatorname{dom} F$, and did not show the rate of the convergence of the proposed method.

In addition to the block CD method, the *proximal gradient* (*PG*) *method* is also an efficient method, because it only requires the evaluation of the gradient at each iteration. The search direction of the PG method at point $x^r \in \mathcal{R}^n$ is defined by

$$d_{\eta^r}(x^r) = \operatorname*{argmin}_{d \in \mathcal{R}^n} \{\langle \nabla f(x^r), d\rangle + B_{\eta^r}(x^r + d, x^r) + \tau\psi(x^r + d)\}, \tag{1.3.4}$$

where function $B_{\eta^r}(\cdot, \cdot) : X \times \text{int}X \to \mathcal{R}$ is called the Bregman function defined by

$$B_{\eta^r}(x, y) := \eta^r(x) - \eta^r(y) - \langle \nabla\eta^r(y), x - y\rangle, \tag{1.3.5}$$

where function $\eta^r : X \to \mathcal{R}$, called the "kernel of B_{η^r}", is assumed to be convex and continuously differentiable on int$X$, and $X \subseteq \text{dom}\, F \subseteq \mathcal{R}^n$ is a closed convex set. The common selections for the kernel $\eta(x)$ include $\frac{1}{2}\|x\|^2$, $\frac{1}{2}x^T\nabla^2 f(x^r)x$, $x \ln x$, etc. For different regularization term $\psi(x)$, the proximal gradient method reduces to many well known methods with suitable kernel $\eta(x)$. See Table 4.1 in Chapter 4 for details.

The proximal gradient method [37] has been widely studied on its convergence rate and its iteration complexity. When function f in problem (1.2.1) is convex, the kernel $\eta(x) = \frac{1}{2}\|x\|^2$, and the step size is set with the fix constant $1/L_f$ or chosen by the line search, it is shown in [52, Theorem 2.1.14] and [70] that

$$F(x^r) - \inf F \le O(\frac{L_f}{r}), \tag{1.3.6}$$

where L_f is the Lipschitz constant for ∇f, and $\inf F$ denotes the infimum of function F. Hence, the proximal gradient method has $O(\frac{L_f}{\varepsilon})$ iteration complexity in the convex case, where ε donotes the approximation accuracy. Moreover, under the "local Lipschitz error bound" assumption, its convergence rate can be further improved. It is shown that the proximal gradient method with the quadratic kernel has the R-linear convergence rate [69] even if problem (1.2.1) is nonconvex. Additionally, a series of accelerated proximal gradient methods have been proposed. See [66, 68, 71] and references therein for details.

Although the proximal gradient method is very efficient for large scale problems, there still have lots of problems for further improvement. For example, since the subproblem (1.3.4) for getting the search direction is an $|n|$-dimensional problem, it is still time consuming as n becomes very large.

Motivated by this, the *block coordinate gradient descent* (*BCGD*) *method* is proposed [69]. Namely, this method is a hybrid of the gradient method and the block coordinate descent method. As mentioned before, the requirement of the convergence of the block CD method is restrictive. Hence, to show the global convergence of the block coordinate gradient descent method is nontrivial. In [69], Tseng and Yun studied a block coordinate gradient descent method with the quadratic kernel $\eta^r(x) = \frac{1}{2}x^T H^r x$, where matrix $H^r \in \mathcal{R}^{n\times n}$ is symmetric and positive definite, and matrix H^r is often chosen to be an approximation to the Hessian

$\nabla^2 f(x^r)$. In [69], at each iteration, we first choose a nonempty block $J \subseteq \{1, 2, \ldots, n\}$. Then we get a search direction by solving the following subproblem.

$$d_H(x; J) = \operatorname*{argmin}_{d \in \mathcal{R}^n} \left\{ \langle \nabla f(x), d \rangle + \frac{1}{2} d^T H d + \tau \psi(x + d) \middle| \, d_{\bar{J}} = 0 \right\}. \tag{1.3.7}$$

Under the "local Lipschitz error bound" assumption, Tseng and Yun [69] showed that the block coordinate proximal gradient method with the Gauss-Seidel rule or the Gauss-Southwell rule also has the R-linear convergence rate for general separable optimization problem (1.2.1).

Additionally, the topic of the iteration complexity of this method has also been extensively discussed. For smooth convex problems, Beck and Tetruashvili [7] showed that the block coordinate gradient projection (BCGP) method with a cyclic rule has the $O(\frac{NL_f}{\varepsilon})$ iteration complexity, where L_f is the Lipschitz constant for ∇f, N is the number of blocks, and $\varepsilon > 0$ is the approximation accuracy. In [32], Hong et al. proposed a general block coordinate descent (BCD) type method for general separable optimization (1.2.1), and proved that it obtains an ε-accurate solution in $O(\frac{NL_f}{\varepsilon})$ iterations when the blocks are updated with the cyclic rule. However, for some special cases of problem (1.2.1), the iteration complexity bound can be sharpened. For example, Saha and Tewari [62] showed that the iteration complexity of the coordinate descent (CD) method for the l_1-regularized problem can be improved to $O(\frac{L_f}{\varepsilon})$ under an isotonicity assumption. It is worth noting that this upper bound does not depend on the number N of blocks and the size n.

1.3.2 Existing methods for the online problem (1.2.2)

The applications of the online optimization problem are mostly built on a large scale. Some researchers have studied the performance of the gradient methods for the online convex optimization problem (1.2.2) [73, 81]. When $\psi(x)$ in (1.2.1) is an indicator function, Zinkevich [81] proved that the greedy projection method for the online convex optimization problem has a regret $O(\sqrt{T})$. When $\psi(x)$ in (1.2.1) is a general regularization function, Xiao [73] proposed a dual averaging method, which is first proposed by Nesterov for classical convex optimization problems. He showed that the proposed method achieves the same regret $O(\sqrt{T})$ for the online optimization problem. However, both of these two methods are full gradient methods, i.e., they update all components of the variable x at each iteration. When the scale of the problem becomes very large, the evaluation for updating the gradient of each iteration would take much time.

Recently, the "block" type methods are becoming very popular, especially for the large scale classical optimization problems [59, 66, 67]. Compared to the full gradient methods, the block type methods can reduce the calculation time at each iteration. Quite recently, Xu and Yin [75] proposed a block coordinate stochastic gradient method with the cyclic rule for a regularized stochastic optimization problem, which relates to the online optimization problem. Under the Lipschitz continuity-like assumption, they showed that the proposed method converges with $O(\frac{1+\log T}{\sqrt{1+T}}N)$, where N is the number of blocks. Note that, as the number of blocks reduces to 1, i.e., $N = 1$, this upper bound reduces to $O(\frac{1+\log T}{\sqrt{1+T}})$, which is still bigger than the average regret $\frac{R(T)}{T} = O(\frac{1}{\sqrt{T}})$ of the greedy projection method [81].

1.4 Motivations and contributions

As mentioned above, the block type methods are verified to be very efficient for the large scale optimization problems [7, 38, 43, 58, 59, 67, 69, 72, 78]. Although this type of methods have been widely studied, there still exist unknown problems. For example, does the (block) CD method converge linearly for the general nonsmooth optimization problem? Is there any tighter iteration complexity bound for the BCPG method for the nonsmooth problem? These problems motivate us to join the research on the block coordinate gradient methods.

In this thesis, we carry out our study from the following two aspects, one is to propose new algorithms for problem (1.2.1), the other is to supplement and improve the theoretical analysis of the existing block coordinate gradient methods. In particular, the contributions of this thesis are itemized as follows.

(1) We present a new inexact CD method with an inexactness description for a class of weighted l_1-regularized convex optimization problem with a box constraint. Under the same assumptions in [43], we show that the proposed inexact CD method is not only globally convergent but also with at least R-linear convergence rate under the almost cycle rule. At each iteration step, we only need to find an approximate solution for a one dimensional problem, which raises the possibility to solve general l_1-regularized convex problems.

(2) We propose a novel class of block coordinate proximal gradient (BCPG) methods with variable Bregman functions for solving the general nonsmooth nonconvex problem (1.2.1). For the proposed methods, we establish their global convergence and R-linear convergence rate with the Gauss-Seidel rule. The idea of using the variable kernels is the innovation, which enables us to obtain many well-known algorithms from the proposed BCPG methods, including the (inexact) BCD method. Moreover, some special kernels allow the proposed BCPG methods to adopt the fixed step size, and help us to construct accelerated algorithms.

(3) We improve the iteration complexity of the block coordinate gradient descent (BCGD) method with the cyclic rule for the convex separable optimization (1.2.1). The great point of the improvement lies in proposing a new Lipschitz continuity-like assumption.

Furthermore, we study the relations between the proposed assumption and the Lipchitz continuity, and show that $M \leq \sqrt{N} L_f$ or $M \leq 2L_f$, where M is the constant given in the proposed assumption, and L_f is the Lipschitz constant. These results yield that the iteration complexity bound derived in this thesis is sharper than existing results.

(4) We investigate the performance of the block coordinate gradient (BCG) method with the cyclic rule for the online and stochastic optimization problems. For the separable online optimization problem (1.2.2), we show that the proposed method has the same regret as the greedy projection (GP) method, where the GP method is a full gradient projection method. For the stochastic optimization problem, the results in this thesis are shown to be tighter than the existing results.

For convenience, we use the following abbreviations of some well known methods in the subsequent sections.

Table 1.1: Abbreviations for the well known methods

Methods with full name	Abbreviation
coordinate descent method	CD method
inexact coordinate descent method	ICD method
block coordinate descent method	BCD method
proximal gradient method	PG method
coordinate gradient descent method [69]	CGD method
block coordinate gradient descent method [69]	BCGD method
block coordinate proximal gradient method	BCPG method

1.5 Overview of the thesis

This thesis is organized as follows.

In Chapter 2, we introduce some notations, basic definitions, the proximal gradient methods and preliminary results, which will be used in the subsequent chapters.

In Chapter 3, we propose an inexact CD method with a new inexactness description for a class of weighted l_1-regularized convex optimization problem with a box constraint, and show that the proposed method has global and R-linear convergence rate. Moreover we propose a specific ICD algorithm, and report numerical results on the comparison of the proposed algorithm and the CGD method.

In Chapter 4, we propose a class of block coordinate proximal gradient (BCPG) methods with variable Bregman functions for solving the general nonsmooth nonconvex problem

(1.2.1). We establish their global convergence and R-linear convergence rate with the Gauss-Seidel rule. Moreover, we propose a specific algorithm of the BCPG methods with variable kernels for a convex problem with separable simplex constraints. The numerical results on the proposed algorithm and the algorithm with a fixed kernel are reported.

In Chapter 5, we investigate the iteration complexity of the block coordinate gradient descent (BCGD) method with the cyclic rule for the convex separable optimization (1.2.1). With the new Lipschitz continuity-like assumption, we improve the iteration complexity of the BCGD method.

In Chapter 6, we investigate the performance of the block coordinate gradient (BCG) method with the cyclic rule for the online separable optimization problem and the stochastic optimization problem. We show that the proposed method has the same regret as the greedy projection method [81] for the online optimization problem (1.2.2). Moreover, we extend our results to the regularized stochastic optimization problem, and show that the results in this thesis are tighter than that in [75].

Finally, in Chapter 7, we summarize this thesis and mention some issues for the future research.

Chapter 2

Preliminaries

In this chapter, we introduce some notations, definitions, some versions of the proximal gradient methods and preliminary results, which will be used in the subsequent chapters.

2.1 Notations

For any vectors $x, y \in \mathcal{R}^n$, the Euclidean inner product $\langle x, y \rangle$ is defined by

$$\langle x, y \rangle := x_1 y_1 + x_2 y_2 + \cdots + x_n y_n.$$

For a vector $x \in \mathcal{R}^n$ and a matrix $G \in \mathcal{R}^{n \times n}$, $G \succeq 0$, the norms $\|x\|_1$, $\|x\|_2$, and $\|x\|_G$ are defined as follows.

$$\begin{aligned}
&\|x\|_1 := |x_1| + \cdots + |x_n|, \\
&\|x\|_2 := \sqrt{\langle x, x \rangle} = \sqrt{x_1^2 + \cdots + x_n^2}, \\
&\|x\|_G := \sqrt{\langle x, Gx \rangle}.
\end{aligned}$$

Unless otherwise stated, we let $\| \cdot \|$ denote the norm $\| \cdot \|_2$. For a matrix $A \in \mathcal{R}^{n \times n}$, norm $\|A\|$ is defined by

$$\|A\| := \max_{x \neq 0,\, x \in \mathcal{R}^n} \frac{\|Ax\|}{\|x\|}.$$

For a differentiable function $h : \mathcal{R}^n \to \mathcal{R}$, the gradient $\nabla h(x) \in \mathcal{R}^n$ is defined by

$$\nabla h(x) := \begin{pmatrix} \frac{\partial h(x)}{\partial x_1} \\ \vdots \\ \frac{\partial h(x)}{\partial x_n} \end{pmatrix}.$$

Moreover, if function h is twice differentiable, the Hessian matrix $\nabla^2 h(x) \in \mathcal{R}^{n\times n}$ is defined by

$$\nabla^2 h(x) := \begin{pmatrix} \frac{\partial^2 h(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 h(x)}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 h(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 h(x)}{\partial x_n^2} \end{pmatrix}.$$

2.2 Definitions

In this section, we introduce some basic definitions, which will be used in this thesis. For more details, see [13, 52, 60].

2.2.1 Convexity

In this subsection, we give the definitions and relevant properties related to the convexity.

Definition 2.2.1. *Let $X \subseteq \operatorname{dom} f$ be a convex set and $f : X \to \mathcal{R}$ be a scalar function.*

(1) *Function f is convex if it holds that*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y),\ \forall x, y \in X, t \in (0,1).$$

(2) *Function f is strictly convex if it holds that*

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y),\ \forall x, y \in X, t \in (0,1).$$

(3) *Function f is μ_f-strongly convex on X, $\mu_f > 0$, if it holds that*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{1}{2}\mu_f t(1-t)\|x-y\|_2^2,\ \forall x, y \in X, t \in (0,1).$$

Additionally, if function f is differentiable, the (strong) convexity can be described as follows.

Lemma 2.2.1 ([13, Section 3.1.3], [52, Theorem 2.1.9])**.** *Let $X \subseteq \operatorname{dom} f$ be a convex set and $f : X \to \mathcal{R}$ be a differentiable function.*

(1) *Function f is convex if and only if*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle,\ \forall x, y \in X.$$

(2) *Function f is μ_f-strongly convex on X, $\mu_f > 0$, if and only if*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu_f}{2}\|y-x\|_2^2,\ \forall x, y \in X.$$

In other words, the convex differentiable function is lower bounded by its first order Taylor approximation, while the strongly convex function is lower bounded by a quadratic function.

When the function is nondifferentiable and convex, it has a similar character as Lemma 2.2.1 (1), where the gradient is replaced by the "subgradient".

Definition 2.2.2. *Let function $f : \mathcal{R}^n \to \mathcal{R}$ be convex. A vector $\xi \in \mathcal{R}^n$ is a subgradient of function f at point $x \in \operatorname{dom} f$ if*

$$f(y) \geq f(x) + \langle \xi, y - x \rangle, \forall y \in \operatorname{dom} f.$$

The set of all subgradients of function f at point $x \in \operatorname{dom} f$, denoted by $\partial f(x)$, is called the subdifferential of function f at point $x \in \operatorname{dom} f$, i.e.,

$$\partial f(x) := \{\xi \mid f(y) \geq f(x) + \langle \xi, y - x \rangle, \forall y \in \operatorname{dom} f\}.$$

When function f is proper, convex and $x \in \operatorname{int} \operatorname{dom} f$, subdifferential $\partial f(x)$ is nonempty, bounded and convex [52, Theorem 3.1.13]. In addition, when function f is convex and differentiable at x, the element of the subdifferential $\partial f(x)$ is unique, and $\partial f(x) = \{\nabla f(x)\}$. We let $\partial_J f \in \mathcal{R}^{|J|}$ denote the subdifferential of function f with respect to variable x_J.

2.2.2 Lipschitz continuity

In this subsection, we introduce the definition of some types of the Lipschitz continuities and their related results.

Definition 2.2.3. *Let function $f : \mathcal{R}^n \to \mathcal{R}$ be continuously differentiable and let $\{\mathcal{J}^i, i = 1, \dots, N\}$ be a partition of the set $\mathcal{N} = \{1, \dots, n\}$* [1]*. The gradient ∇f is said to be block-wise Lipschitz continuous with respect to blocks $\{\mathcal{J}^i, i = 1, \dots, N\}$ if for any $i \in \{1, \dots, N\}$ there exists a positive constant L_i such that*

$$\|\nabla_{\mathcal{J}^i} f(y) - \nabla_{\mathcal{J}^i} f(x)\| \leq L_i \|y - x\|, \forall x, y \in \operatorname{dom} f \text{ with } y_{\bar{\mathcal{J}}^i} = x_{\bar{\mathcal{J}}^i}. \tag{2.2.1}$$

The constant L_i is called the Lipschitz constant of gradient ∇f with respect to block $\mathcal{J}^i$.

Note that when the number N of blocks reduces to 1, Definition 2.2.3 reduces to the standard Lipschitz continuity. In this thesis, we let constant $L_f > 0$ denote the Lipschiz constant of ∇f with respect to the whole variable x.

The Lipschitz continuity plays an important role for the linear convergence or the iteration complexity. Next, we recall a lemma, which states the properties of function f with the Lipschitz continuity.

[1] A family of sets $\{\mathcal{J}^i, i = 1, 2, \dots, N\}$ is said to be a partition of set $\mathcal{N} = \{1, 2, \dots, N\}$ if (i) $\mathcal{J}^i \subseteq \mathcal{N}$, $i = 1, 2, \dots, N$, is nonempty. (ii) $\bigcup_{i=1}^N \mathcal{J}^i = \{1, 2, \dots, n\}$. (iii) $\mathcal{J}^i \cap \mathcal{J}^j = \varnothing, \forall i, j \in \{1, 2, \dots, N\}, i \neq j$.

Lemma 2.2.2 ([7, Lemma 3.2], [50, Lemma 2]). *Let $\{\mathcal{J}^i,\ i = 1, \dots, N\}$ be a partition of the set $\mathcal{N} = \{1, \dots, n\}$. Suppose that the gradient ∇f is block-wise Lipschitz continuous with respect to blocks $\{\mathcal{J}^i, i = 1, \dots, N\}$. Then, the following statements hold.*

(i) *For any $i \in \{1, \dots, N\}$ and any $y, x \in \operatorname{dom} f$ with $x_{\bar{\mathcal{J}}^i} = y_{\bar{\mathcal{J}}^i}$, $f(y) \leq f(x) + \langle \nabla_{\mathcal{J}^i} f(x), y_{\mathcal{J}^i} - x_{\mathcal{J}^i}\rangle + \frac{L_i}{2}\|y_{\mathcal{J}^i} - x_{\mathcal{J}^i}\|^2$.*

(ii) *There exists a positive constant L_f such that $L_f \leq \displaystyle\sum_{i=1}^{N} L_i$ and $\|\nabla f(y) - \nabla f(x)\| \leq L_f\|y - x\|$ hold for any $y, x \in \operatorname{dom} f$.*

Lemma 2.2.2 (i) implies that the gradient Lipschitz continuous function is upper bounded by a quadratic function. If function f is both gradient block Lipschitz continuous and strongly convex, it follows from Lemma 2.2.1 (2) and Lemma 2.2.2 (i) that $\mu_f \leq L_i$ holds for any $i \in \{1, \dots, N\}$. Lemma 2.2.2 (ii) states the relations between the Lipschitz constants L_f and L_i, $i \in \{1, \dots, N\}$.

Next, we give a new Lipschitz continuity-like definition, which helps us to improve the iteration complexity of the block type methods.

Definition 2.2.4. *Let $\{\mathcal{J}^i,\ i = 1, \dots, N\}$ be a partition of the set $\mathcal{N} = \{1, \dots, n\}$. We say that gradient ∇f is block lower triangular Lipschitz continuous with respect to blocks $\{\mathcal{J}^i,\ i = 1, 2, \dots, N\}$, if there exists a nonnegative constant M such that*

$$\|g(x, y) - \nabla f(x)\| \leq M\|y - x\|, \forall x, y \in \operatorname{dom} f, \tag{2.2.2}$$

where $g : \mathcal{R}^{n+n} \to \mathcal{R}^n$ with

$$g_{\mathcal{J}^i}(x, y) = \nabla_{\mathcal{J}^i} f(y_{\mathcal{J}^1}, \dots, y_{\mathcal{J}^{i-1}}, x_{\mathcal{J}^i} \dots, x_{\mathcal{J}^N}), \;\; i = 1, \dots, N. \tag{2.2.3}$$

Note that the constant M in inequality (2.2.2) of Definition 2.2.4 is different from the Lipschitz constant L_f. The relation between the constants M and L_f is summarized by the following remark.

Remark 2.2.1. *When $N = 1$, we have $g(x, y) = \nabla f(x)$, which yields that $M = 0$ in (2.2.2). When $N > 1$, it is shown in Section 5.4 that $M \leq 2L_f$ holds for many classes of functions f. For general continuously differentiable function f, we have $M \leq \sqrt{N}L_f$, which is proven in Section 5.4.*

2.2.3 Optimal solution and optimal conditions

In this subsection, we give the definitions of optimal solution, stationary point, as well as the first order optimality condition of problem (1.2.1). For details, see [8, 13, 52, 60] and references therein.

Definition 2.2.5. **(1)** *A vector $x^* \in \operatorname{dom} F$ is a locally optimal solution of problem (1.2.1) if there exists a scalar $R > 0$ such that*

$$F(x^*) \leq F(x),\ \forall x \in \{x \mid \|x - x^*\| \leq R, x \in dom F\}.$$

Function value $F(x^*)$ is called the local minimum of problem (1.2.1).

(2) *A vector $x^* \in \operatorname{dom} F$ is a globally optimal solution of problem (1.2.1) if it holds that*

$$F(x^*) \leq F(x),\ \forall x \in dom F.$$

Function value $F(x^*)$ is called the global minimum of problem (1.2.1).

When problem (1.2.1) is a smooth problem, i.e, function $\psi(x) = 0$ for any $x \in \operatorname{dom} F$, we provide its first order necessary optimality condition as follows.

Theorem 2.2.1 ([8, Propositions 1.1.1 and 1.1.2])**.** *Suppose that $\psi(x) = 0$ in problem (1.2.1) for any $x \in \operatorname{dom} F$. If vector $x^* \in \operatorname{dom} F$ is a locally optimal solution, then we have*

$$\nabla f(x^*) = 0.$$

Moreover, if function f is convex, then condition $\nabla f(x^*) = 0$ is a necessary and sufficient condition for a vector $x^* \in \operatorname{dom} F$ to be a globally optimal solution.

When problem (1.2.1) is a nonsmooth convex problem, its optimality condition can be described as follows.

Theorem 2.2.2 ([52, Theorem 3.1.15], [12, Propositions 2.1.1 and 2.1.2])**.** *Suppose that problem (1.2.1) is a convex problem. A vector $x^* \in \operatorname{dom} F$ is a locally optimal solution of problem (1.2.1) if and only if one of the following statements holds.*

(1) *$0 \in \partial F(x^*)$, i.e., vector $0 \in \mathcal{R}^n$ is a subgradient of F at x^*.*

(2) *$F'(x^*; d) \geq 0$ holds for any $d \in \mathcal{R}^n$, where $F'(x^*; d)$ is a direction derivative at the vector x^* with respect to the direction $d \in \mathcal{R}^n$, i.e., $F'(x^*; d) := \lim\limits_{t \to 0^+} \dfrac{F(x^* + td) - F(x^*)}{t}$.*

The condition $0 \in \partial F(x^*)$ reduces to $\nabla f(x^*) = 0$ if $\varphi(x) = 0$ for any $x \in \operatorname{dom} F$. When problem (1.2.1) is a general nonconvex nonsmooth problem, the vector $x^* \in \mathcal{R}^n$ satisfying the condition in Theorem 2.2.2 (2) is referred to as a *stationary point.*

The next theorem states the optimality condition of the constrained nonsmooth convex problem.

Theorem 2.2.3 ([8, Proposition B.24 (f)])**.** *Let $F : \mathcal{R}^n \to \mathcal{R}$ in problem (1.2.1) be a convex function. A vector $x^* \in \operatorname{dom} F$ minimizes F over a convex set $\Omega \subseteq \mathcal{R}^n$ if and only if there exists a subgradient $\eta \in \partial F(x^*)$ such that*

$$\langle \eta, x - x^* \rangle \geq 0,\ \forall x \in \Omega.$$

2.2.4 Convergence rate and regret

In this subsection, we give several definitions on the convergence speed, such as the linear convergence and the iteration complexity. Moreover, we give a precise definition of the online optimization problem and the related definition regret. See [52, 56, 81] and references therein for details.

Definition 2.2.6. *Let $\{x^r\}$ be a sequence generated by an iterative method. Suppose that the sequence $\{x^r\}$ converges to the vector x^*.*

(1) *The sequence $\{x^r\}$ is said to converge Q-linearly, if there exists a constant $\mu \in (0,1)$ such that*

$$\lim_{r\to\infty} \frac{\|x^{r+1} - x^*\|}{\|x^r - x^*\|} = \mu.$$

In this case, we also say that the iterative method has Q-linear convergence rate.

(2) *The sequence $\{x^r\}$ is said to converge R-linearly, if there exist constants $\mu \in (0,1)$ and $c > 0$ such that*

$$\|x^r - x^*\| \leq c\mu^r.$$

In this case, we also say that the iterative method has R-linear convergence rate.

Unless otherwise stated, “linear convergence” means the “Q-linear convergence” in this thesis. The following theorem states the relation between the Q-linear and R-linear convergence.

Theorem 2.2.4 ([56, Proposition 1.3])**.** *If sequence $\{x^r\}$ converges Q-linearly, then it converges R-linearly, but not vice versa.*

The iteration complexity can be defined formally as follows.

Definition 2.2.7. *Let $\{x^r\}$ be the sequence generated by an iterative method. Let F^* be an optimal value and $\varepsilon > 0$ be an approximation accuracy. The global iteration complexity bound of the iterative method is an iteration number $r_0 \geq 1$ such that, for any $r > r_0$,*

$$F(x^r) - F^* \leq \varepsilon.$$

We also say that the iterative method has r_0 iteration complexity.

For example, if there exist constants $c > 0$ and $p > 0$ such that

$$F(x^r) - F^* \leq cr^{-\frac{1}{p}},$$

by setting $r_0 = (\frac{c}{\varepsilon})^p$, for any $r \geq r_0$, we have $F(x^r) - F^* \leq \varepsilon$. Hence, the corresponding iterative method has $(\frac{c}{\varepsilon})^p$ iteration complexity. For simplicity, we say that the iterative method has $O(\frac{1}{\varepsilon^p})$ iteration complexity, or say that the sequence $\{F(x^r)\}$ converges to F^* with $O(\frac{1}{\varepsilon^p})$.

Next, we introduce the formal definition of the online optimization problem.

Definition 2.2.8 ([81]). *An online convex optimization problem consists of a feasible set $\Omega \subseteq \mathcal{R}^n$ and an infinite sequence $\{F^1, F^2, \dots\}$, where $F^t : \Omega \to \mathcal{R}$ is a convex function. At each time step t, an algorithm selects a vector $x^t \in \Omega$. After the vector is selected, it receives the loss function F^t.*

As mentioned in Chapter 1, for the online optimization problem, a primary concept is the regret, which is defined as follows.

Definition 2.2.9. *For the online optimization problem (1.2.2), for given $T > 0$, the regret, denoted by $R(T)$, is defined by*

$$R(T) := \sum_{t=1}^{T} F^t(x^t) - \sum_{t=1}^{T} F^t(x^*), \tag{2.2.4}$$

where $\{x^1, \dots, x^T\}$ are decision vectors generated by an iterative method, and x^* is an optimal solution in some sense.

Note that the common selection of the optimal solution x^* for the online problem is $x^* \in \operatorname{argmin}_{x\in\Omega} \sum_{t=1}^{T} F^t(x)$ [81].

2.3 (Block) proximal gradient methods and related results

In this section, we take problem (1.2.1) as an example to introduce several existing versions of the (block) proximal gradient methods with Bregman functions and the related results. The (block) proximal gradient method with Bregman functions for the online problem (1.2.2) can be described similarly.

First, we introduce the framework of the proximal gradient method with Bregman functions without block for problem (1.2.1).

Proximal gradient (PG) method with Bregman functions:

Step 0: Choose an initial point $x^0 \in \mathrm{int}X$ and choose a kernel function η. Let $r = 0$.

Step 1: Solve the following subproblem and obtain a direction $d^r = d_\eta(x^r)$.

$$d_\eta(x^r) = \operatorname*{argmin}_{d \in \mathcal{R}^n} \left\{ \langle \nabla f(x^r), d \rangle + B_\eta(x^r + d, x^r) + \tau\psi(x^r + d) \right\}, \tag{2.3.1}$$

Step 2: Choose a stepsize $\alpha^r > 0$. Set $x^{r+1} = x^r + \alpha^r d^r$ and $r = r + 1$. Go to Step 1.

The most common and easy setting for the kernel η is the quadratic function $\frac{1}{2}\|x\|^2$. The following existing methods can be looked on as special cases of the above PG method.

(1) The gradient projection method, i.e., when function $\psi(x)$ is an indicator function with respect to a convex set X, the sequence $\{x^r\}$ is generated by

$$x^{r+1} = P_X(x^r - \alpha_r \nabla f(x^r)),$$

where P_X is a projection operator, defined by $P_X(x) = \operatorname*{argmin}_{u \in X} \|u - x\|_2^2$.

This method is included in the PG method with kernel $\eta(x) = \|x\|^2$.

(2) The soft-thresholding method, i.e., when function $\psi(x) = \|x\|_1$, the sequence $\{x^r\}$ is generated by

$$x^{r+1} = T_{\alpha_r\tau}(x^r - \alpha_r \nabla f(x^r)),$$

where mapping $T_{\alpha_r\tau} : \mathcal{R}^n \to \mathcal{R}^n$ is the soft-thresholding operator, defined as follows.

$$T_{\alpha_r\tau}(u)_i := \begin{cases} u_i - \alpha_r\tau & \text{if } u_i \geq \alpha_r\tau, \\ 0 & \text{if } -\alpha_r\tau \leq u_i \leq \alpha_r\tau, \\ u_i + \alpha_r\tau & \text{if } u_i \leq -\alpha_r\tau. \end{cases} \tag{2.3.2}$$

This method can be deduced from the above PG method with kernel $\eta(x) = \|x\|^2$.

(3) The exponentiated gradient method, i.e., when function $\psi(x)$ is an indicator function with respect to a simplex $\sum_{i=1}^n x_i = 1$, $x_i \geq 0$, $i = 1, \ldots, n$, the sequence $\{x^r\}$ is generated by

$$x_i^{r+1} = \frac{x_i^r e^{-\alpha_r \nabla_i f(x^r)}}{\sum_{i=1}^n x_i^r e^{-\alpha_r \nabla_i f(x^r)}}.$$

This method is a special PG method with kernel $\eta(x) = \sum_{i=1}^n x_i \ln x_i$.

Note that in the case (1), subproblem (1.3.4) must be solved by an iterative solution method. In the cases (2) and (3), subproblem (1.3.4) has a closed form solution (alternatively, analytic solution).

Next, we introduce the existing block coordinate gradient descent method for problem (1.2.1), which is proposed in [69]. The particular framework is described as follows.

Block coordinate gradient descent (BCGD) method:

Step 0: Choose an initial point $x^0 \in \operatorname{dom} F$ and let $r = 0$.

Step 1: Choose a nonempty $J^r \subseteq \{1, 2, \ldots, n\}$ and a symmetric matrix $H^r \in \mathcal{R}^{n\times n}$, $H^r \succ 0$.

Step 2: Solve the following subproblem and obtain a direction $d^r = d_{\eta^r}(x^r)$.

$$d_{\eta^r}(x^r) = \operatorname*{argmin}_{d\in\mathcal{R}^n} \left\{ \langle \nabla f(x^r), d\rangle + \frac{1}{2} d^T H^r d + \tau\psi(x^r + d) \,\middle|\, d_{\bar{J}^r} = 0 \right\}. \tag{2.3.3}$$

Step 3: Choose a stepsize $\alpha^r > 0$. Set $x^{r+1} = x^r + \alpha^r d^r$ and $r = r + 1$. Go to Step 1.

Note that matrix $H^r \succ 0$ is usually chosen to be an approximation of the Hessian $\nabla^2 f(x^r)$. This BCGD method is closely related to the PG method. If we choose kernel $\eta^r(x) = \frac{1}{2}x^T H^r x$ in the PG method, we have $B_{\eta^r}(x^r + d, x^r) = \frac{1}{2}d^T H^r d$. With different matrix H^r, the directions $d_{\eta^r}(x^r)$ in (2.3.3) are different. The following existing methods can be regarded as special cases of this BCGD method.

(1) The Newton method, i.e., when function $\psi(x) = 0$ for any $x \in \operatorname{dom} f$, the sequence $\{x^r\}$ is generated by

$$x^{r+1} = x^r - \alpha^r (\nabla^2 f(x^r))^{-1} \nabla f(x^r).$$

This method can be regarded as a special case of the BCGD method with $H^r = \nabla^2 f(x^r)$ and $J^r = \{1, 2, \ldots, n\}$.

(2) The regularized Newton method, i.e., when function $\psi(x) = 0$ for any $x \in \operatorname{dom} f$, the sequence $\{x^r\}$ is generated by

$$x^{r+1} = x^r - \alpha^r (\nabla^2 f(x^r) + \mu I)^{-1} \nabla f(x^r).$$

This method can be deduced from the BCGD method with $H^r = \nabla^2 f(x^r) + \mu I$ and $J^r = \{1, 2, \ldots, n\}$.

In the existing works on the convergence of the (block) proximal gradient method, the following lemma is necessary, which is called the “three-point property”.

Lemma 2.3.1 ([68, Property 1]). *For any proper l.s.c. convex function $\varphi : X \to (-\infty, \infty]$ and any $z \in X$, if*

$$z_+ := \operatorname*{argmin}_{x\in X} \left\{ \varphi(x) + B_\eta(x, z) \right\},$$

where $B_\eta(x,z) := \eta(x) - \eta(z) - \langle \nabla\eta(z), x-z\rangle$ *and* $\eta : X \to (-\infty, \infty]$ *is differentiable at* z_+, *then*

$$\varphi(x) + B_\eta(x,z) \geq \varphi(z_+) + B_\eta(z_+,z) + B_\eta(x,z_+), \forall x \in X.$$

The existing results on the iteration complexity of (block) proximal gradient methods are summarized in Table 2.1, where the constant N denotes the number of blocks, L_f is the Lipschitz constant for ∇f, and ε is the approximation accuracy.

Table 2.1: The existing iteration complexity of the (block) proximal gradient methods when f is convex or strongly convex

Method	Complexity (convex)	Complexity (strongly convex)
proximal gradient method[45]	$O(\frac{L_f}{\varepsilon})$	$O(\log\frac{1}{\varepsilon})$
BCGD method [7, 50, 59, 70]	$O(\frac{NL_f}{\varepsilon})$ or $O(\frac{NL_f}{\varepsilon}\log\frac{1}{\varepsilon})$	$O(N\log\frac{1}{\varepsilon})$

2.4 Rules for choosing blocks

In this subsection, we introduce the rules to select a block for the block type method.

The following generalized Gauss-Seidel rule [69, 71] is an extension of the classical cycle rule [41].

Generalized Gauss-Seidel rule: Choose $\{J^r\}$ to satisfy the following condition. There exists an integer $B \geq 1$ such that $J^0, J^1, \ldots$ collectively cover the set $\mathcal{N} = \{1, 2, \ldots, n\}$ for every B consecutive iterations.

This rule implies that the index set J^r satisfies

$$J^r \bigcup J^{r+1} \bigcup \cdots \bigcup J^{r+B-1} = \mathcal{N}, \forall r = 0, 1, \ldots. \tag{2.4.1}$$

Note that the blocks $\{J^{r+j},\ j = 0, 1, \ldots, B-1\}$ in the generalized Gauss-Seidel rule can be overlapping.

The following restricted Gauss-Seidel rule [69] is a special case of the generalized Gauss-Seidel rule.

Restricted Gauss-Seidel rule: Choose $\{J^r\}$ to satisfy the following condition. There exists a subsequence $\Gamma \subseteq \{0, 1, \ldots\}$ such that

$$0 \in \Gamma,\ \ \mathcal{N} = \left(\text{disjoint union of } J^r,\ J^{r+1},\ \ldots,\ J^{\varphi(r)-1}\right), \forall r \in \Gamma, \tag{2.4.2}$$

where $\varphi(r)$ is defined as $\varphi(r) := \min\{r' \in \Gamma \mid r' > r\}$.

It is worth mentioning that the blocks $\{J^{r+j}, j = 0, 1, \dots, \varphi(r)-1, r \in \Gamma\}$ in the restricted Gauss-Seidel rule cannot be overlapping. Thus, if the convex function ψ in problem (1.2.1) is block separable with respect to each block J^r, then $\psi(x^r)$ can be rewritten as

$$\psi(x^r) = \sum_{i=r}^{\varphi(r)-1} \psi_{J^i}(x^r_{J^i}), \ \forall r \in \Gamma. \tag{2.4.3}$$

The following cyclic rule [41] is the simplest Gauss-Seidel rule, which is a special case of the restricted Gauss-Seidel rule with $\varphi(r) = r + N$ and $\Gamma = \{0, N, 2N, \dots\}$.

Cyclic rule: Let $\{\mathcal{J}^i,\ i = 1, \dots, N\}$ be a partition of the set $\mathcal{N} = \{1, \dots, n\}$. Choose blocks in a cyclic order.

The Gauss-Southwell rule [69] is a general name for the Gauss-Southwell-r rule and the Gauss-Southwell-q rule, which are described as follows.

Gauss-Southwell-r rule: Choose $\{J^r\}$ to satisfy the following condition.

$$\|d_{D^r}(x^r; J^r)\|_\infty \geq \nu \|d_{D^r}(x^r; \mathcal{N})\|_\infty,$$

where direction $d_{D^r}(x^r; J^r)$ is defined by (1.3.7) with $H = D^r \in \mathcal{R}^{n\times n}$ such that D^r is diagonal with $D^r \succ 0$ and constant $\nu \in (0, 1]$.

Gauss-Southwell-q rule: Choose $\{J^r\}$ to satisfy the following condition.

$$q_{D^r}(x^r; J^r) \geq \nu q_{D^r}(x^r; \mathcal{N}),$$

where value $q_H(x; J)$ is defined by

$$q_H(x; J) := \Big(\langle \nabla f(x), d\rangle + \frac{1}{2} d^T H d + \tau\psi(x + d)\Big)_{d = d_H(x;J)} - \tau\psi(x),$$

direction $d_H(x; J)$ is defined by (1.3.7), constant $\nu \in (0, 1]$, and matrix $D^r \in \mathcal{R}^{n\times n}$ is diagonal with $D^r \succ 0$.

Note that in the Gauss-Southwell-r rule and Gauss-Southwell-q rule, we need to compute $d_{D^r}(x^r; \mathcal{N})$, which needs to solve an n-dimensional problem.

The random rule [59] is described as follows.

Random rule: Let $\{\mathcal{J}^i,\ i = 1, \dots, N\}$ be a partition of set $\mathcal{N} = \{1, \dots, n\}$, and let $\{p_i, i = 1, \dots, N\}$ be a set of probability vector such that, for any $i \in \{1, 2, \dots, N\}$, $p_i > 0$, and $\sum_{i=1}^{N} p_i = 1$. At each iteration, we choose a block $\mathcal{J}^i \in \{\mathcal{J}^1, \mathcal{J}^2, \dots, \mathcal{J}^N\}$ with probability p_i.

Chapter 3

An inexact coordinate descent method for the weighted l_1-regularized convex optimization problem

3.1 Introduction

In this chapter, we consider the following weighted l_1-regularized convex optimization problem with box constraints.

$$\begin{aligned} &\text{minimize} \;\; F(x) := g(Ax) + \langle b, x \rangle + \sum_{i=1}^{n} \tau_i |x_i| \\ &\text{subject to} \;\; l \leq x \leq u, \end{aligned} \tag{3.1.1}$$

where $g : \mathcal{R}^m \to (-\infty, \infty]$ is a *strictly convex* and continuously differentiable function, $A \in \mathcal{R}^{m \times n}$ and $b \in \mathcal{R}^n$. Moreover, τ, l and u are n-dimensional vectors such that $l_i \in [-\infty, \infty)$, $u_i \in (-\infty, \infty], \tau_i \in [0, \infty)$ and $l_i < u_i$ for each $i = 1, \ldots, n$. The nonnegative scalar constant τ_i is called the weight and the term $\sum_{i=1}^{n} \tau_i |x_i|$ is called the l_1-regularization function. For convenience, we denote the differentiable term of F by f, that is, $f(x) := g(Ax) + \langle b, x \rangle$.

When $l_i = -\infty$, $u_i = \infty$ and $\tau_i = \tau$ hold for any $i = 1, 2, \ldots, n$, problem (3.1.1) reduces to the unconstrained separable optimization problem (1.2.1). Additionally, it is worth mentioning that problem (3.1.1) is a convex problem since function g is assumed to be strictly convex. However, the optimal solutions are possibly not unique because the matrix A may not have the full column rank.

As described in Subsection 1.2.3, the applications [29, 35, 55, 77] of problem (3.1.1) typically have large scales, and the CD method [38, 43, 67, 72] is shown to be an efficient method to solve it. Luo and Tseng [43] proved that it has global and linear convergence for a smooth problem, that is, $\tau_i = 0$ for all i. For more complicate regularization problem, in 2001, Tseng [67] showed the global convergence of a block coordinate descent (BCD) method for minimizing a nondifferentiable function with certain separability. However, its convergence rate is still unknown. Moreover, in the most of the existing works, we assume that the exact minimizers of the subproblem can be found at each iteration in [67, 43]. It is possible for the l_1-l_2 problem, while usually it is hard for the general l_1-regularized convex problem.

In order to improve the applicability of the CD method, some inexact CD methods are proposed [11, 69, 74], such as the inexact block coordinate descent method [11], the coordinate gradient descent (CGD) method [69] and the coordinate proximal point method [74]. The CGD method is executed with one step of the gradient method for the subproblem of the CD method, while the method [74] exploits the proximal point method to find an approximate solution. Thus they are regarded as the inexact CD methods. Bonettini [11] proposed an inexact version of the CD method. He gave some appropriate conditions about the inexactness of the solution for the subproblem, and has shown that the proposed method with these conditions has global convergence. However, he only focused on a smooth optimization problem, i.e., $\tau_i = 0$, for all i, and did not show the rate of convergence of the proposed method.

In this chapter, we present a new inexact coordinate descent (ICD) method with a new inexactness description, which is an extension of the result of Luo and Tseng [43]. In particular, we extend in the following three aspects.

- The smooth convex problem is extended to that with the l_1-regularized function.
- At each iteration, we accept an inexact solution of the subproblem instead of the exact solution.
- The linear convergence rate is proven for the nonsmooth problem.

Under the same assumptions in [43], we show that the proposed ICD method is not only globally convergent but also with at least R-linear convergence rate under the almost cycle rule (Theorem 3.4.2 in Section 3.4 for details).

This chapter is organized as follows. In Section 3.2, we derive optimality conditions for problem (3.1.1) and also define ε-optimality conditions which are related to an inexact solution. In Section 3.3, we present a framework of the ICD method and make some assumptions for the “inexact solutions”. The global convergence and linear convergence rate are established in Section 3.4. In Section 3.5, we report some numerical experiments for the proposed ICD method and show the comparison with the coordinate gradient descent (CGD) method [69]. Finally, we conclude this chapter in Section 3.6.

3.2 Preliminaries

Throughout the chapter, we make the following basic assumptions for problem (3.1.1).

Assumption 3.2.1. *For problem (3.1.1), we assume that*

(a) A_j is a nonzero vector for all $j \in \{1, 2, \dots, n\}$.

(b) $l_i < 0 < u_i$ for all $i \in \{1, 2, \dots, n\}$.

(c) The set of the optimal solutions, denoted by X^*, is nonempty.

(d) The effective domain of g, denoted by $\operatorname{dom} g$, is nonempty and open.

(e) g is twice continuously differentiable on $\operatorname{dom} g$.

(f) $\nabla^2 g(Ax^*)$ is positive definite for every optimal solution $x^* \in X^*$.

We make a few remarks on these assumptions. In Part (a), if A_j is zero, then x_j^* of the optimal solution x^* can be easily determined. Thus we can remove x_j from problem (3.1.1). Part (b) is just for simplification. If both l_i and u_i are positive for some $i \in \{1, 2, \dots, n\}$, we may replace x_i, l_i and u_i by $\bar{x}_i + \frac{l_i + u_i}{2}$, $\frac{l_i - u_i}{2}$ and $\frac{u_i - l_i}{2}$. Then problem (3.1.1) is reformulated into the case without l_1-regularized term for the index i. If g is strongly convex and twice differentiable on $\operatorname{dom} g$, then Parts (e) and (f) are satisfied automatically. For example, a quadratic function, an exponential function, and even some complicate functions in the l_1-regularized logistic regression problem satisfy (e) and (f). Note that we do not assume the boundedness of the optimal solution set X^*.

Next, we present some properties under Assumption 3.2.1 that are used in the subsequent sections. From (e) and (f) in Assumption 3.2.1, there exists a sufficiently small closed neighborhood $B(Ax^*)$ of Ax^* such that $B(Ax^*) \subseteq \operatorname{dom} g$ and $\nabla^2 g$ is positive definite in $B(Ax^*)$. Furthermore, it implies that g is strongly convex in $B(Ax^*)$, i.e., there exists a scalar $\mu_g > 0$ such that

$$g(y) - g(z) - \langle \nabla g(z), y - z \rangle \geq \frac{\mu_g}{2} \|y - z\|^2, \ \forall y, z \in B(Ax^*). \tag{3.2.1}$$

3.2.1 Optimality conditions

The KKT conditions [60] for problem (3.1.1) are described as follows.

$$\begin{aligned} &\nabla_i f(x) + \tau_i \partial |x_i| - \mu_i + \nu_i \ni 0, \\ &x_i \geq l_i, \mu_i \geq 0, \mu_i (x_i - l_i) = 0, \quad i = 1, \dots, n, \\ &x_i \leq u_i, \nu_i \geq 0, \nu_i (u_i - x_i) = 0, \end{aligned} \tag{3.2.2}$$

where $\partial|\cdot|$ is the subdifferential of the absolute value function. Since problem (3.1.1) is convex, x satisfying (3.2.2) is an optimal solution of problem (3.1.1). The KKT conditions (3.2.2) can be rewritten as follows.

Lemma 3.2.1. *A vector x is an optimal solution of problem (3.1.1) if and only if one of the following statements holds for each $i \in \{1, \dots, n\}$.*

(i) $\nabla_i f(x) \geq \tau_i$ *and* $x_i = l_i$.

(ii) $\nabla_i f(x) = \tau_i$ *and* $l_i \leq x_i \leq 0$.

(iii) $|\nabla_i f(x)| \leq \tau_i$ *and* $x_i = 0$.

(iv) $\nabla_i f(x) = -\tau_i$ *and* $0 \leq x_i \leq u_i$.

(v) $\nabla_i f(x) \leq -\tau_i$ *and* $x_i = u_i$.

Next, we represent these conditions as a fixed point of some operator. To this end, we first use the soft-thresholding operator, given in Section 2.3, to define a mapping $T_\tau : \mathcal{R}^n \to \mathcal{R}^n$ as

$$T_\tau(x)_i := (|x_i| - \tau_i)_+ \mathrm{sgn}(x_i), \tag{3.2.3}$$

where the scalar function $(a)_+$ is defined by $(a)_+ := \max(0, a)$, and $\mathrm{sgn}(a)$ is a sign function defined as follows.

$$\mathrm{sgn}(a) := \begin{cases} -1 & \text{if } a < 0, \\ 0 & \text{if } a = 0, \\ 1 & \text{if } a > 0. \end{cases}$$

It can be verified that mapping T_τ is nonexpansive, i.e., $\|T_\tau(y) - T_\tau(z)\| \leq \|y - z\|$, for any $y, z \in \mathrm{dom}\, F$.

Let $[x]^+_{[l,u]}$ denote the orthogonal projection of a vector x onto the box $[l, u]$. This projection is also nonexpansive and its i-th coordinate can be written as $[x_i]^+_{[l_i,u_i]} := \mathrm{mid}\{x_i, l_i, u_i\}$, where $\mathrm{mid}\{x_i, l_i, u_i\}$ is defined by $\mathrm{mid}\{x_i, l_i, u_i\} := \max\{l_i, \min\{u_i, x_i\}\}$.

By using the mappings T_τ and $[\cdot]^+_{[l,u]}$, we define a mapping $P_{\tau,l,u}(x) : \mathcal{R}^n \to \mathcal{R}^n$ by

$$P_{\tau,l,u}(x) := [T_\tau(x - \nabla f(x))]^+_{[l,u]}. \tag{3.2.4}$$

Since $[x]^+_{[l,u]}$ and T_τ are nonexpansive, we have that

$$\|P_{\tau,l,u}(y) - P_{\tau,l,u}(z)\| \leq \|y - z - \nabla f(y) + \nabla f(z)\|, \ \forall y, z \in \mathrm{dom}\, F. \tag{3.2.5}$$

Now, the optimal solutions can be described as a fixed point of the mapping $P_{\tau,l,u}$.

Theorem 3.2.1. *For problem (3.1.1), a vector x belongs to the optimal solution set X^* if and only if $x = P_{\tau,l,u}(x)$, i.e., $X^* = \{x | \ x \in \operatorname{dom} g, x = P_{\tau,l,u}(x)\}$.*

Proof. This theorem is a direct consequence of Theorem 3.2.2 that will be shown in Subsection 2.2. ■

Since the solution set X^* is not necessarily bounded, the level set of F may not be bounded. Nevertheless, as an extension of [43, Lemma 3.3], we can show the compactness of the set $\Omega(\zeta) := \{t | \ t = Ax, F(x) \leq \zeta, x \in [l, u]\}$.

Lemma 3.2.2. *For a given constant value ζ, the set $\Omega(\zeta)$ is a compact subset of* $\operatorname{dom} g$.

Proof. The l_1-regularized convex problem (3.1.1) can be transformed into a smooth optimization problem with box constraints.

$$\begin{aligned} &\underset{x^+,\, x^- \in \mathcal{R}^n}{\text{minimize}} \quad \bar{F}(x^+, x^-) := g(Ax^+ - Ax^-) + \langle b, x^+ - x^- \rangle + \sum_{i=1}^{n} \tau_i (x_i^+ + x_i^-) \\ &\text{subject to } 0 \leq x_i^+ \leq u_i, i = 1, \dots, n, \\ &\qquad\qquad\quad 0 \leq x_i^- \leq |l_i|, i = 1, \dots, n. \end{aligned} \tag{3.2.6}$$

Note that if (x^+, x^-) is feasible for problem (3.2.6), then $x = x^+ - x^-$ is also feasible for problem (3.1.1) due to $l \leq x \leq u$.

Let $\bar{\Omega}(\zeta)$ be defined as follows.

$$\begin{aligned} \bar{\Omega}(\zeta) :=& \{Ax^+ - Ax^- | \ \bar{F}(x^+, x^-) \leq \zeta, x^+ \in [0, u], x^- \in [0, |l|]\} \\ =& \{Ax | \ x = x^+ - x^-, \bar{F}(x^+, x^-) \leq \zeta, x^+ \in [0, u], x^- \in [0, |l|]\}, \end{aligned}$$

where $|l| = (|l_1|, \dots, |l_n|)^T$. Then $\bar{\Omega}(\zeta)$ is a compact set of $\operatorname{dom} g$ from Appendix in [43].

In the rest part, we only need to show $\bar{\Omega}(\zeta) = \Omega(\zeta)$. In fact, for every $t \in \bar{\Omega}(\zeta)$, there exists (x, x^+, x^-) such that $t = Ax, x = x^+ - x^-, \bar{F}(x^+, x^-) \leq \zeta, x^+ \in [0, u]$, and $x^- \in [0, |l|]$. Then we have $x \in [l, u]$ and $\zeta \geq \bar{F}(x^+, x^-) \geq F(x)$. It further implies that $t \in \Omega(\zeta)$, i.e., $\bar{\Omega}(\zeta) \subseteq \Omega(\zeta)$.

Conversely, for every $t \in \Omega(\zeta)$, there exists a vector x such that $t = Ax, F(x) \leq \zeta$, and $x \in [l, u]$. Let $x_i^+ := \max\{x_i, 0\}$ and $x_i^- := \max\{-x_i, 0\}$ for each $i = 1, \dots, n$. Then we have $x^+ \in [0, u], x^- \in [0, |l|]$, $x = x^+ - x^-$, and $\bar{F}(x^+, x^-) = F(x)$. Therefore, we deduce that $t \in \bar{\Omega}(\zeta)$, which implies that $\Omega(\zeta) \subseteq \bar{\Omega}(\zeta)$. Consequently, the relation $\bar{\Omega}(\zeta) = \Omega(\zeta)$ holds. ■

Next, we show that ∇g is Lipschitz continuous on some compact set including $\Omega(\zeta)$. For this purpose, we define a set $\Omega(\zeta) + B(\epsilon_0)$ as $\Omega(\zeta) + B(\epsilon_0) := \{p + v | \ p \in \Omega(\zeta), \ \|v\| \leq \epsilon_0\}$, where ϵ_0 is a positive constant. It is easy to see that the set $\Omega(\zeta) + B(\epsilon_0)$ is compact.

Lemma 3.2.3. *There exist constants $L_g > 0$ and $\epsilon_0 > 0$ such that $\Omega(\zeta) + B(\epsilon_0) \subseteq \operatorname{dom} g$ and $\|\nabla g(y) - \nabla g(z)\| \leq L_g\|y - z\|$ for all $y,\ z \in \Omega(\zeta) + B(\epsilon_0)$.*

Proof. Since set $\Omega(\zeta)$ is closed from Lemma 3.2.2 and $\operatorname{dom} g$ is open, there exists a positive constant ϵ_0 such that $\Omega(\zeta) + B(\epsilon_0) \subseteq \operatorname{dom} g$. Furthermore, since g is twice continuously differentiable on $\operatorname{dom} g$, and $\Omega(\zeta) + B(\epsilon_0)$ is compact, we have that $\nabla^2 g(x)$ is bounded in $\Omega(\zeta) + B(\epsilon_0)$, that is, there exists a constant $L_g > 0$ such that $\|\nabla^2 g(x)\| \leq L_g$ for all $x \in \Omega(\zeta) + B(\epsilon_0)$. Then, this lemma holds from the mean value theorem. ■

Similar to [44, Lemma 2.1], we can prove the following invariant property of the optimal solution set X^*. For simplicity, we omit the proof here.

Lemma 3.2.4. *For any $x^*, y^* \in X^*$, we have $Ax^* = Ay^*$.*

3.2.2 ε-optimality conditions

In this subsection, we give a definition of the relaxed optimality conditions, and show a relation between the conditions and the mapping $P_{\tau,l,u}$.

Definition 3.2.1. *We say that the ε-optimality conditions for problem (3.1.1) hold at x if one of the following statements holds for each i.*

(i) $\nabla_i f(x) - \tau_i \geq -\varepsilon$ *and* $|x_i - l_i| \leq \varepsilon$.

(ii) $|\nabla_i f(x) - \tau_i| \leq \varepsilon$ *and* $l_i - \varepsilon \leq x_i \leq \varepsilon$.

(iii) $|\nabla_i f(x)| \leq \tau_i + \varepsilon$ *and* $|x_i| \leq \varepsilon$.

(iv) $|\nabla_i f(x) + \tau_i| \leq \varepsilon$ *and* $-\varepsilon \leq x_i \leq u_i + \varepsilon$.

(v) $\nabla_i f(x) + \tau_i \leq \varepsilon$ *and* $|x_i - u_i| \leq \varepsilon$.

Definition 3.2.2. *We say that x is an ε-approximate solution of problem (3.1.1) if the ε-optimality conditions hold at x.*

Note that the optimality conditions in Lemma 3.2.1 can be obtained by Definition 3.2.1 with $\varepsilon = 0$.

For convenience, we define the following five index sets.

$$\begin{aligned}
J_1(x,\varepsilon) &:= \{i|\ \nabla_i f(x) - \tau_i \geq -\varepsilon, |x_i - l_i| \leq \varepsilon\};\\
J_2(x,\varepsilon) &:= \{i|\ |\nabla_i f(x) - \tau_i| \leq \varepsilon, l_i - \varepsilon \leq x_i \leq \varepsilon\};\\
J_3(x,\varepsilon) &:= \{i|\ |\nabla_i f(x)| \leq \tau_i + \varepsilon, |x_i| \leq \varepsilon\};\\
J_4(x,\varepsilon) &:= \{i|\ |\nabla_i f(x) + \tau_i| \leq \varepsilon, -\varepsilon \leq x_i \leq u_i + \varepsilon\};
\end{aligned}$$

$$J_5(x,\varepsilon) := \{i|\ \nabla_i f(x) + \tau_i \le \varepsilon, |x_i - u_i| \le \varepsilon\}.$$

Then the ε-optimality conditions hold at x if and only if $\bigcup_{i=1}^5 J_i(x,\varepsilon) = \{1,2,\dots,n\}$.

Throughout the chapter, for simplicity, we assume that

$$\varepsilon < \frac{1}{2} \min_{i=1,\dots,n} \{-l_i, u_i\}. \tag{3.2.7}$$

The next theorem gives an equivalent description of the ε-optimality conditions, which will be used for constructing an inexact CD method and investigating its convergence properties.

Theorem 3.2.2. *The ε-optimality conditions hold at x if and only if $|x_i - P_{\tau,l,u}(x)_i| \le \varepsilon$ holds for each $i \in \{1,2,\dots,n\}$.*

Proof. By the definitions of $T_\tau(x)$ and $P_{\tau,l,u}(x)$ in (3.2.3) and (3.2.4), we have that

$$\begin{aligned} |x_i - P_{\tau,l,u}(x)_i| &= |x_i - \mathrm{mid}\{l_i, u_i, \max\{0, |x_i - \nabla_i f(x)| - \tau_i\}\mathrm{sgn}(x_i - \nabla_i f(x))\}| \\ &= \begin{cases} |x_i - l_i| & \text{if } x_i - \nabla_i f(x) \in (-\infty, l_i - \tau_i], \\ |\nabla_i f(x) - \tau_i| & \text{if } x_i - \nabla_i f(x) \in (l_i - \tau_i, -\tau_i], \\ |x_i| & \text{if } x_i - \nabla_i f(x) \in (-\tau_i, \tau_i], \\ |\nabla_i f(x) + \tau_i| & \text{if } x_i - \nabla_i f(x) \in (\tau_i, u_i + \tau_i], \\ |x_i - u_i| & \text{if } x_i - \nabla_i f(x) \in (u_i + \tau_i, \infty). \end{cases} \end{aligned} \tag{3.2.8}$$

We firstly consider the “ if ” part of this theorem. It is sufficient to show that if $|x_i - P_{\tau,l,u}(x)_i| \le \varepsilon$ holds for each $i \in \{1,2,\dots,n\}$, then for each $i \in \{1,2,\dots,n\}$ there exists a $j \in \{1,2,\dots,5\}$ such that $i \in J_j(x,\varepsilon)$. We can prove this according to the distinct cases in (3.2.8). If $x_i - \nabla_i f(x) \in (-\infty, l_i - \tau_i]$, then it follows from $|x_i - P_{\tau,l,u}(x)_i| \le \varepsilon$ and (3.2.8) that $|x_i - P_{\tau,l,u}(x)_i| = |x_i - l_i| \le \varepsilon$, that is, $x_i - l_i \ge -\varepsilon$. Moreover, since $x_i - \nabla_i f(x) \in (-\infty, l_i - \tau_i]$ implies that $\nabla_i f(x) - \tau_i \ge x_i - l_i$, we have $\nabla_i f(x) - \tau_i \ge -\varepsilon$. Therefore, $i \in J_1(x,\varepsilon)$ holds. Similarly, we can show that if $x_i - \nabla_i f(x)$ is located in other intervals, the corresponding results also hold.

Conversely, suppose that x is an ε-approximate solution, i.e., for each $i \in \{1,2,\dots,n\}$, there exists a $j \in \{1,2,\dots,5\}$ such that $i \in J_j(x,\varepsilon)$. Thus, it is sufficient to show that for each i and j such that $i \in J_j(x,\varepsilon)$, the inequality $|x_i - P_{\tau,l,u}(x)_i| \le \varepsilon$ holds.

Case 1: $i \in J_1(x,\varepsilon)$ **or** $i \in J_5(x,\varepsilon)$. First suppose that $i \in J_1(x,\varepsilon)$. Then we have

$$\nabla_i f(x) - \tau_i \ge -\varepsilon \text{ and } |x_i - l_i| \le \varepsilon. \tag{3.2.9}$$

They imply that $x_i - \nabla_i f(x) \le l_i - \tau_i + 2\varepsilon$. It then follows from (3.2.7) that $x_i - \nabla_i f(x) \in (-\infty, -\tau_i)$. Thus, we focus on (3.2.8) in two intervals $(-\infty, l_i - \tau_i]$ and $(l_i - \tau_i, -\tau_i]$. If

$x_i - \nabla_i f(x) \in (-\infty, l_i - \tau_i]$, it follows from (3.2.8) that $|x_i - P_{\tau,l,u}(x)_i| = |x_i - l_i|$. Then the inequality $|x_i - P_{\tau,l,u}(x)_i| \leq \varepsilon$ holds due to (3.2.9). If $x_i - \nabla_i f(x) \in (l_i - \tau_i, -\tau_i]$, then we have $\nabla_i f(x) - \tau_i < x_i - l_i$ and $|x_i - P_{\tau,l,u}(x)_i| = |\nabla_i f(x) - \tau_i|$, which together with (3.2.9) imply $|x_i - P_{\tau,l,u}(x)_i| \leq \varepsilon$. A symmetric argument can prove the case with $i \in J_5(x, \varepsilon)$.

Case 2: $i \in J_2(x, \varepsilon)$ **or** $i \in J_4(x, \varepsilon)$. First suppose that $i \in J_2(x, \varepsilon)$. Then we have

$$|\nabla_i f(x) - \tau_i| \leq \varepsilon \text{ and } l_i - \varepsilon \leq x_i \leq \varepsilon. \tag{3.2.10}$$

We obtain $-\tau_i - \varepsilon \leq -\nabla_i f(x) \leq \varepsilon - \tau_i$ from the first inequality. Adding these inequalities and the second inequalities of (3.2.10), we have $l_i - \tau_i - 2\varepsilon \leq x_i - \nabla_i f(x) \leq 2\varepsilon - \tau_i$. With the assumption (3.2.7) on ε, we have $x_i - \nabla_i f(x) \in [l_i - \tau_i - 2\varepsilon, u_i)$. Now we show $|x_i - P_{\tau,l,u}(x)_i| \leq \varepsilon$ from (3.2.8) and (3.2.10) by dividing the interval $[l_i - \tau_i - 2\varepsilon, u_i)$ into $[l_i - \tau_i - 2\varepsilon, l_i - \tau_i]$, $(l_i - \tau_i, -\tau_i]$, $(-\tau_i, \tau_i]$ and (τ_i, u_i).

(i) If $x_i - \nabla_i f(x) \in (l_i - \tau_i - 2\varepsilon, l_i - \tau_i]$, it follows from (3.2.8) that $|x_i - P_{\tau,l,u}(x)_i| = |x_i - l_i|$. Meanwhile, we obtain $x_i - l_i \leq \nabla_i f(x) - \tau_i$. Then we have $x_i - l_i \leq \varepsilon$ from the first inequality in (3.2.10). On the other hand, we have $x_i - l_i \geq -\varepsilon$ from the inequalities $l_i - \varepsilon \leq x_i \leq \varepsilon$ in (3.2.10). Hence, the inequality $|x_i - P_{\tau,l,u}(x)_i| \leq \varepsilon$ holds.

(ii) If $x_i - \nabla_i f(x) \in (l_i - \tau_i, -\tau_i]$, then the inequality $|x_i - P_{\tau,l,u}(x)_i| \leq \varepsilon$ holds due to (3.2.8) and (3.2.10).

(iii) If $x_i - \nabla_i f(x) \in (-\tau_i, \tau_i]$, then we have $|x_i - P_{\tau,l,u}(x)_i| = |x_i|$ by (3.2.8). Moreover, it yields $x_i \geq \nabla_i f(x) - \tau_i$. It then follows from the inequality $|\nabla_i f(x) - \tau_i| \leq \varepsilon$ in (3.2.10) that $x_i \geq -\varepsilon$. Furthermore, we have $x_i \leq \varepsilon$ from (3.2.10). Hence, $|x_i - P_{\tau,l,u}(x)_i| = |x_i| \leq \varepsilon$.

(iv) If $x_i - \nabla_i f(x) \in [\tau_i, u_i]$, we have $|x_i - P_{\tau,l,u}(x)_i| = |\nabla_i f(x) + \tau_i|$ from (3.2.8). Thus, $|x_i - P_{\tau,l,u}(x)_i| \leq \varepsilon$ is equivalent to $-\tau_i - \varepsilon \leq \nabla_i f(x) \leq \varepsilon - \tau_i$. First, we have $\nabla_i f(x) \leq x_i - \tau_i \leq \varepsilon - \tau_i$, where the first inequality follows from the assumption $x_i - \nabla_i f(x) \in [\tau_i, u_i]$, and the second inequality follows from (3.2.10). Next, we obtain $\nabla_i f(x) \geq -\varepsilon + \tau_i \geq -\varepsilon - \tau_i$, where the first inequality follows from (3.2.10), and the second inequality holds due to $\tau_i \geq 0$.

In the case where $i \in J_4(x, \varepsilon)$, a similar analysis shows $|x_i - P_{\tau,l,u}(x)_i| \leq \varepsilon$.

Case 3: $i \in J_3(x, \varepsilon)$. Then we have

$$|\nabla_i f(x)| \leq \tau_i + \varepsilon \quad \text{and} \quad |x_i| \leq \varepsilon. \tag{3.2.11}$$

These inequalities imply $-\tau_i-2\varepsilon \le x_i-\nabla_i f(x) \le \tau_i+2\varepsilon$. Moreover, we have by (3.2.7) that $l_i - \tau_i < x_i - \nabla_i f(x) < u_i + \tau_i$. Then we prove $|x_i - P_{\tau,l,u}(x)_i| \le \varepsilon$ by dividing the interval $(l_i - \tau_i, u_i + \tau_i)$ into the following three intervals: $(l_i - \tau_i, -\tau_i], (-\tau_i, \tau_i]$ and $(\tau_i, u_i + \tau_i)$.

(i) If $l_i - \tau_i \le x_i - \nabla_i f(x) \le -\tau_i$, then we have $|x_i - P_{\tau,l,u}(x)_i| = |\nabla_i f(x) - \tau_i|$ from (3.2.8). Thus, $|x_i - P_{\tau,l,u}(x)_i| \le \varepsilon$ is equivalent to $\tau_i - \varepsilon \le \nabla_i f(x) \le \tau_i + \varepsilon$. We first have $\tau_i - \varepsilon \le \nabla_i f(x)$ from (3.2.11) and the ineqaulity $x_i - \nabla_i f(x) \le -\tau_i$. Next, we have $\nabla_i f(x) \le \tau_i + \varepsilon$ since the inequality $|\nabla_i f(x)| \le \tau_i + \varepsilon$ in (3.2.11) holds.

(ii) If $-\tau_i < x_i - \nabla_i f(x) \le \tau_i$, then we have $|x_i - P_{\tau,l,u}(x)_i| = |x_i|$ from (3.2.8). It then follows from (3.2.11) that $|x_i - P_{\tau,l,u}(x)_i| \le \varepsilon$.

(iii) If $\tau_i \le x_i - \nabla_i f(x) \le \tau_i + u_i$, then we have $|x_i - P_{\tau,l,u}(x)_i| = |\nabla_i f(x) + \tau_i|$ from (3.2.8). Meanwhile, $\nabla_i f(x) \le x_i - \tau_i$ holds. Then the inequality $\nabla_i f(x) \le \varepsilon - \tau_i$ holds due to $x_i \le \varepsilon$ in (3.2.11). Moreover, we have $\nabla_i f(x) \ge -\tau_i - \varepsilon$ by (3.2.11). Hence the inequality $|x_i - P_{\tau,l,u}(x)_i| \le \varepsilon$ holds.

Upon the preceding proof, the necessary condition of this theorem is confirmed. ■

3.3 Inexact coordinate descent (ICD) method

In this section, we first present a framework for the ICD method, and then give some assumptions for the "inexact solutions".

A general framework of the ICD method can be described as follows.

Inexact coordinate descent (ICD) method:

Step 0: Choose an initial point $x^0 \in [l, u]$ and let $r = 0$.

Step 1: If some termination condition holds, then stop.

Step 2: Choose an index $i(r) \in \{1, \ldots, n\}$, and get an approximate solution $x_{i(r)}^{r+1}$ of the following one dimensional subproblem:

$$\underset{x_{i(r)} \in \{l_{i(r)} \le x_{i(r)} \le u_{i(r)}\}}{\text{minimize}} \quad F(x_1^r, x_2^r, \ldots, x_{i(r)-1}^r, x_{i(r)}, x_{i(r)+1}^r, \ldots, x_n^r). \tag{3.3.1}$$

Step 3: Set $x_j^{r+1} = x_j^r$ for all $j \in \{1, \ldots, n\}$ such that $j \neq i(r)$, and let $r = r + 1$. Go to Step 1.

Note that the exact solution of the subproblem (3.3.1) is unique from Assumption 3.2.1(a) and the strict convexity of g. We use the notation $i(r)$ for the index chosen at the r-th iteration. For simplicity, we use i instead of $i(r)$ when $i(r)$ is clear from the context.

For the global convergence of the ICD method, it is important to define the inexactness of the approximate solutions of the subproblem (3.3.1) and to choose an appropriate index $i(r)$ in Step 2.

For the inexactness, we require the following assumptions.

Assumption 3.3.1. *We assume that the following statements hold:*

(i) $F(x_1^r, x_2^r, \ldots, x_{i-1}^r, x_i^{r+1}, x_{i+1}^r, \ldots, x_n^r) \leq \min\limits_{x_i \in \{l_i, 0, u_i, x_i^r\}} F(x_1^r, x_2^r, \ldots, x_{i-1}^r, x_i, x_{i+1}^r, \ldots, x_n^r)$.

(ii) x_i^{r+1} *is feasible, i.e.,* $x_i^{r+1} \in [l_i, u_i]$.

(iii) x_i^{r+1} *is an* ε^{r+1}*-approximate solution of the subproblem (3.3.1).*

(iv) ***Conditions on*** ε^{r+1}*:* $\varepsilon^{r+1} \leq \min\{\delta_r, \alpha_r |x_i^{r+1} - x_i^r|, \varepsilon^r\}$*, where* $\{\delta_r\}$ *is a monotonically decreasing sequence such that* $\lim\limits_{r\to\infty} \delta_r = 0$*, and* $\alpha_r \in [0, \bar{\alpha}]$ *holds with a positive constant* $\bar{\alpha}$*.*

(v) ***Conditions on*** α_r*:* $\alpha_r < \dfrac{\mu_g \min\limits_j \|A_j\|^2}{2L_g \max\limits_j \|A_j\|^2 + 2}$ *holds for sufficiently large* r*, where* μ_g *is a positive constant defined in (3.2.1), and* L_g *is the Lipschitz constant of* ∇g *given in Lemma 3.2.3.*

Here we make a simple explanation. Part (i) enforces not only that $\{F(x^r)\}$ is decreasing but also that $\{F(x^{r+1})\}$ is less than $F(x_1^r, x_2^r, \ldots, x_{i-1}^r, x_i, x_{i+1}^r, \ldots, x_n^r)$ at the point where F is nonsmooth. This condition is easy to check when computing. It also plays a key role for the convergence of $\{x^r\}$ when the objective function is not differentiable. In Part (iii), recall that the ε-optimality conditions for the one dimensional subproblem (3.3.1) is that one of (i)-(v) in Definition 3.2.1 holds at $x_{i(r)}$. The assumptions (i)-(iv) are necessary for the global convergence while the assumption (v) on α_r is used to guarantee the linear convergence rate of $\{x^r\}$.

Note that if we obtain the exact solution of the subproblem (3.3.1) at each iteration, then the sequence $\{x^r\}$ satisfies Assumption 3.3.1 automatically. Hence, the classical CD method is a special case of the ICD method.

For the choice of the coordinate $i(r)$ in Step 2, we adopt the "generalized Gauss-Seidel rule" [69, 71] with $|J^r| = 1, r = 1, 2, \ldots,$ which is precisely defined in Section 2.4. For simplicity, in this chapter, we call it the "almost cyclic rule", which is described as follows.

Almost cyclic rule:
There exists an integer $B \geq n$, such that every coordinate is iterated upon at least once every B successive iterations.

In the next section, we will show the ICD method with the almost cycle rule converges R-linearly to a solution under Assumptions 3.2.1 and 3.3.1.

3.4 Global and linear convergence

In this section, we show the global and linear convergence of the ICD method. Compared with the classical exact CD method, the ICD method has many "inexact" factors. Thus we need some preparations.

First of all, we illustrate a brief outline of the proof.

(1) $\lim_{r\to\infty}\{x^{r+1}-x^r\}=0$. (Lemma 3.4.3)

(2) $Ax^r \to Ax^*$, where x^* is one of the optimal solutions. (Theorem 3.4.1)

(3) Sufficient decreasing: $F(x^r)-F(x^{r+1}) \geq \eta\|x^r-x^{r+1}\|^2$ for some positive constant η. (Lemma 3.4.8)

(4) Error bound: $\|Ax^r-Ax^*\| \leq \kappa\|x^r-P_{\tau,l,u}(x^r)\|$ for some κ. (Lemma 3.4.9)

(5) Linear convergence. (Theorems 3.4.2 and 4.5.2)

Note that since it is not necessary for the matrix A to have full column rank, $Ax^r \to Ax^*$ (Theorem 3.4.1) does not imply $x^r \to x^*$.

For convenience, we define two vectors $\tilde{x}^{r+1}$ and x^{r+1} as follows.

$$\tilde{x}^{r+1} := (x_1^r, x_2^r, \ldots, x_{i(r)-1}^r, \tilde{x}_{i(r)}^{r+1}, x_{i(r)+1}^r, \ldots, x_n^r), \tag{3.4.1}$$

and

$$x^{r+1} := (x_1^r, x_2^r, \ldots, x_{i(r)-1}^r, x_{i(r)}^{r+1}, x_{i(r)+1}^r, \ldots, x_n^r), \tag{3.4.2}$$

where $x_{i(r)}^{r+1}$ and $\tilde{x}_{i(r)}^{r+1}$ are an ε^{r+1}-approximate solution and the exact solution of the subproblem (3.3.1), respectively.

In the first part of this section, we show $\lim_{r\to\infty}\{F(\tilde{x}^r)-F(x^r)\}=0$ and $\lim_{r\to\infty}\{x^{r+1}-x^r\}=0$. To this end, we need the following function $h_i : \mathcal{R}^n \times \mathcal{R}^n \to \mathcal{R}$ and Lemma 3.4.1.

$$\begin{aligned} h_i(y,z) &:= \nabla_i f(z)(y_i-z_i)+\tau_i(|y_i|-|z_i|) \\ &= \begin{cases} (\nabla_i f(z)+\tau_i)(y_i-z_i) & \text{if } y_i \geq 0, z_i \geq 0, \\ \nabla_i f(z)(y_i-z_i)+\tau_i(y_i+z_i) & \text{if } y_i \geq 0, z_i \leq 0, \\ \nabla_i f(z)(y_i-z_i)+\tau_i(-y_i-z_i) & \text{if } y_i \leq 0, z_i \geq 0, \\ (\nabla_i f(z)-\tau_i)(y_i-z_i) & \text{if } y_i \leq 0, z_i \leq 0. \end{cases} \end{aligned} \tag{3.4.3}$$

Lemma 3.4.1. *There exists a positive constant $\mathcal{M}$ such that $|x_{i(r)}^{r+1}-\tilde{x}_{i(r)}^{r+1}| \leq \frac{2\mathcal{M}}{\|A_{i(r)}\|}$ for all r.*

Proof. By lemma 3.2.2, we have that the set $\Omega(F(x^0))$ is compact. Since $\{Ax^{r+1}\}$, $\{A\tilde{x}^{r+1}\} \subseteq \Omega(F(x^0))$ holds, we further obtain that $\{Ax^{r+1}\}$ and $\{A\tilde{x}^{r+1}\}$ are bounded, that is, there exists a constant $\mathcal{M} > 0$ such that $\|Ax^{r+1}\|$, $\|Ax^r\| \leq \mathcal{M}$ for all r. Then we deduce

$$\|A_{i(r)}\||x^{r+1}_{i(r)} - \tilde{x}^{r+1}_{i(r)}| = \|Ax^{r+1} - A\tilde{x}^{r+1}\| \leq \|Ax^{r+1}\| + \|A\tilde{x}^{r+1}\| \leq 2\mathcal{M},$$

which implies the conclusion since A_i is nonzero for all i. ■

Lemma 3.4.2. $\lim\limits_{r\to\infty}\{F(\tilde{x}^r) - F(x^r)\} = 0.$

Proof. Since $\tilde{x}^{r+1}_{i(r)}$ is the exact solution of subproblem (3.3.1), the inequality

$$F(\tilde{x}^{r+1}) - F(x^{r+1}) \leq 0 \tag{3.4.4}$$

always holds. On the other hand, by the convexity of f, we have

$$\begin{aligned} F(\tilde{x}^{r+1}) - F(x^{r+1}) &\geq \nabla_{i(r)} f(x^{r+1})(\tilde{x}^{r+1}_{i(r)} - x^{r+1}_{i(r)}) + \tau_{i(r)}(|\tilde{x}^{r+1}_{i(r)}| - |x^{r+1}_{i(r)}|) \\ &= h_{i(r)}(\tilde{x}^{r+1}, x^{r+1}). \end{aligned} \tag{3.4.5}$$

Let index sets Z^A and Z^B be defined by

$$Z^A := \{r|\, |\tilde{x}^r_{i(r)} - x^r_{i(r)}| \leq \varepsilon^r\},\ Z^B := \{r|\, |\tilde{x}^r_{i(r)} - x^r_{i(r)}| > \varepsilon^r\},$$

respectively. First we consider the subsequence $\{x^{r+1}\}_{Z^A}$ of $\{x^r\}$. Since $\{Ax^r\}$ is bounded, $\{\nabla f(x^r)\}$ is also bounded from the continuity of ∇g. It then follows from (3.4.4), (3.4.5) and $\varepsilon^{r+1} \to 0$ that $\lim\limits_{r\to\infty,\, r\in Z^A}\{F(\tilde{x}^{r+1}) - F(x^{r+1})\} = 0$.

Next we consider the subsequence $\{x^{r+1}\}_{Z^B}$. We will show the following inequality

$$h_{i(r)}(\tilde{x}^{r+1}, x^{r+1}) \geq -P\varepsilon^{r+1}, \forall\, r+1 \in Z^B \tag{3.4.6}$$

holds, where $P = \frac{2\mathcal{M}}{\|A_{i(r)}\|} + 2\tau_{i(r)} + 2\varepsilon^{r+1}$. Then it is easy to show $\lim\limits_{r\to\infty,\, r\in Z^B}\{F(\tilde{x}^{r+1}) - F(x^{r+1})\} = 0$ from (3.4.4), (3.4.5), (3.4.6) and $\varepsilon^r \to 0$.

Recall that $x^{r+1}_{i(r)}$ is an ε^{r+1}-approximate solution of the subproblem (3.3.1), i.e., there exists a $j \in \{1, 2, \ldots, 5\}$ such that $i(r) \in J_j(x^{r+1}, \varepsilon^{r+1})$. Suppose that $r + 1 \in Z^B$. In the rest part, we show that (3.4.6) holds for $i(r) \in J_j(x^{r+1}, \varepsilon^{r+1})$, $j \in \{1, 2, \ldots, 5\}$. For simplicity, we only show the cases $i(r) \in J_j(x^{r+1}, \varepsilon^{r+1})$, $j \in \{1, 2, 3\}$. The cases $j \in \{4, 5\}$ can be deduced in a similar way.

Case 1: $i(r) \in J_1(x^{r+1}, \varepsilon^{r+1})$. We have $\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)} \geq -\varepsilon^{r+1}$ and $|x^{r+1}_{i(r)} - l_{i(r)}| \leq \varepsilon^{r+1}$.

Since $\varepsilon^{r+1} \leq \dfrac{1}{2}\min\{-l_{i(r)}, u_{i(r)}\}$, the inequality $x^{r+1}_{i(r)} < 0$ holds.

(a) If $\tilde{x}_{i(r)}^{r+1} \geq 0$, then it follows from (3.4.3), Lemma 3.4.1 and $\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)} \geq -\varepsilon^{r+1}$ that

$$\begin{aligned} h_{i(r)}(\tilde{x}^{r+1}, x^{r+1}) &= (\nabla_{i(r)} f(x^{r+1}) + \tau_{i(r)})\tilde{x}_{i(r)}^{r+1} - (\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)})x_{i(r)}^{r+1} \\ &\geq (2\tau_{i(r)} - \varepsilon^{r+1})\tilde{x}_{i(r)}^{r+1} - x_{i(r)}^{r+1}(-\varepsilon^{r+1}) \\ &\geq -\varepsilon^{r+1}(\tilde{x}_{i(r)}^{r+1} - x_{i(r)}^{r+1}) \\ &\geq -\varepsilon^{r+1}\frac{2\mathcal{M}}{\|A_{i(r)}\|}. \end{aligned}$$

(b) If $\tilde{x}_{i(r)}^{r+1} < 0$, then $\tilde{x}_{i(r)}^{r+1} - x_{i(r)}^{r+1} > 0$ holds by $|x_{i(r)}^{r+1} - l_{i(r)}| \leq \varepsilon^{r+1}$ and $r+1 \in Z^B$. We further have $h_{i(r)}(\tilde{x}^{r+1}, x^{r+1}) = (\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)})(\tilde{x}_{i(r)}^{r+1} - x_{i(r)}^{r+1}) \geq -\varepsilon^{r+1}\frac{2\mathcal{M}}{\|A_{i(r)}\|}$ from (3.4.3), $\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)} \geq -\varepsilon^{r+1}$ and Lemma 3.4.1. Therefore, the inequality (3.4.6) holds when $i(r) \in J_1(x^{r+1}, \varepsilon^{r+1})$.

Case 2: $i(r) \in J_2(x^{r+1}, \varepsilon^{r+1})$. We have $|\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)}| \leq \varepsilon^{r+1}$ and $l_{i(r)} - \varepsilon^{r+1} \leq x_{i(r)}^{r+1} \leq \varepsilon^{r+1}$. Now,

$$h_{i(r)}(\tilde{x}^{r+1}, x^{r+1}) = (\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)})(\tilde{x}_{i(r)}^{r+1} - x_{i(r)}^{r+1}) + T(x_{i(r)}^{r+1}, \tilde{x}_{i(r)}^{r+1}, \tau_{i(r)}), \quad (3.4.7)$$

where

$$\begin{aligned} T(x_{i(r)}^{r+1}, \tilde{x}_{i(r)}^{r+1}, \tau_{i(r)}) &:= \tau_{i(r)}\left(\tilde{x}_{i(r)}^{r+1} + |\tilde{x}_{i(r)}^{r+1}| - x_{i(r)}^{r+1} - |x_{i(r)}^{r+1}|\right) \\ &= \begin{cases} 0 & \text{if } \tilde{x}_{i(r)}^{r+1} \leq 0,\ x_{i(r)}^{r+1} \leq 0, \\ 2\tau_{i(r)}\tilde{x}_{i(r)}^{r+1} & \text{if } 0 < \tilde{x}_{i(r)}^{r+1},\ x_{i(r)}^{r+1} \leq 0, \\ -2\tau_{i(r)}x_{i(r)}^{r+1} & \text{if } \tilde{x}_{i(r)}^{r+1} \leq 0,\ 0 < x_{i(r)}^{r+1}, \\ 2\tau_{i(r)}\left(\tilde{x}_{i(r)}^{r+1} - x_{i(r)}^{r+1}\right) & \text{if } 0 < \tilde{x}_{i(r)}^{r+1},\ 0 < x_{i(r)}^{r+1}. \end{cases} \end{aligned} \quad (3.4.8)$$

Suppose first that one of $\tilde{x}_{i(r)}^{r+1}$ and $x_{i(r)}^{r+1}$ is nonpositive. It is easy to see that $T(x_{i(r)}^{r+1}, \tilde{x}_{i(r)}^{r+1}, \tau_{i(r)})$ is no less than $-2\tau_{i(r)}\varepsilon^{r+1}$. It then follows from $|\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)}| \leq \varepsilon^{r+1}$, Lemma 3.4.1 and (3.4.7) that

$$h_{i(r)}(\tilde{x}^{r+1}, x^{r+1}) \geq -\varepsilon^{r+1}(\frac{2\mathcal{M}}{\|A_{i(r)}\|} + 2\tau_{i(r)}).$$

Next suppose that both $\tilde{x}_{i(r)}^{r+1}$ and $x_{i(r)}^{r+1}$ are positive. Then

$$\begin{aligned} h_{i(r)}(\tilde{x}^{r+1}, x^{r+1}) &= (\nabla_{i(r)} f(x^{r+1}) + \tau_{i(r)})\tilde{x}_{i(r)}^{r+1} - x_{i(r)}^{r+1}(\nabla_{i(r)} f(x^{r+1}) + \tau_{i(r)}) \\ &\geq (2\tau_i - \varepsilon^{r+1})\tilde{x}_{i(r)}^{r+1} - x_{i(r)}^{r+1}(2\tau_{i(r)} + \varepsilon^{r+1}) \\ &\geq -\varepsilon^{r+1}(\frac{2\mathcal{M}}{\|A_{i(r)}\|} + x_{i(r)}^{r+1}) - x_{i(r)}^{r+1}(2\tau_{i(r)} + \varepsilon^{r+1}) \end{aligned}$$

$$\geq -(\frac{2\mathcal{M}}{\|A_{i(r)}\|} + 2\tau_{i(r)} + 2\varepsilon^{r+1})\varepsilon^{r+1},$$

where the first inequality follows from $|\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)}| \leq \varepsilon^{r+1}$, $\tilde{x}^{r+1}_{i(r)} > 0$ and $x^{r+1}_{i(r)} > 0$, the second inequality follows from $\tilde{x}^{r+1}_{i(r)} > 0$ and Lemma 3.4.1, and the last inequality follows from $0 \leq x^{r+1}_{i(r)} \leq \varepsilon^{r+1}$. Thus, the inequality (3.4.6) is confirmed.

Case 3: $i(r) \in J_3(x^{r+1}, \varepsilon^{r+1})$. We have $|\nabla_{i(r)} f(x^{r+1})| \leq \tau_{i(r)} + \varepsilon^{r+1}$ and $|x^{r+1}_{i(r)}| \leq \varepsilon^{r+1}$. Moreover, we deduce $\nabla_{i(r)} f(x^{r+1}) + \tau_{i(r)} \in [-\varepsilon^{r+1}, 2\tau_i + \varepsilon^{r+1}]$ from the first inequality. Next we only show that the inequality (3.4.6) holds when $0 \leq x^{r+1}_{i(r)} \leq \varepsilon^{r+1}$. A symmetric argument can prove the case $-\varepsilon^{r+1} \leq x^{r+1}_{i(r)} \leq 0$.

(a) Suppose that $\tilde{x}^{r+1}_{i(r)} \geq 0$. If $\nabla_i f(x^{r+1}) + \tau_{i(r)} \in [-\varepsilon^{r+1}, 0)$, then we have from Lemma 3.4.1 that

$$\begin{aligned} h_{i(r)}(\tilde{x}^{r+1}, x^{r+1}) &= (\nabla_{i(r)} f(x^{r+1}) + \tau_{i(r)})(\tilde{x}^{r+1}_{i(r)} - x^{r+1}_{i(r)}) \\ &\geq -|\nabla_{i(r)} f(x^{r+1}) + \tau_{i(r)}||\tilde{x}^{r+1}_{i(r)} - x^{r+1}_{i(r)}| \\ &\geq -\varepsilon^{r+1}\frac{2\mathcal{M}}{\|A_{i(r)}\|}. \end{aligned}$$

If $\nabla_i f(x^{r+1}) + \tau_{i(r)} \in [0, 2\tau_{i(r)} + \varepsilon^{r+1}]$, then $\tilde{x}^{r+1}_{i(r)}(\nabla_{i(r)} f(x^{r+1}) + \tau_{i(r)}) \geq 0$. Since $0 \leq x^{r+1}_{i(r)} \leq \varepsilon^{r+1}$, we have

$$\begin{aligned} h_{i(r)}(\tilde{x}^{r+1}, x^{r+1}) &= \tilde{x}^{r+1}_{i(r)}(\nabla_{i(r)} f(x^{r+1}) + \tau_{i(r)}) - x^{r+1}_{i(r)}(\nabla_{i(r)} f(x^{r+1}) + \tau_{i(r)}) \\ &\geq -\varepsilon^{r+1}(\varepsilon^{r+1} + 2\tau_{i(r)}). \end{aligned}$$

(b) Suppose that $\tilde{x}^{r+1}_{i(r)} < 0$. Then it follows from $|\nabla_{i(r)} f(x^{r+1})| \leq \tau_{i(r)} + \varepsilon^{r+1}$, $0 \leq x^{r+1}_{i(r)} \leq \varepsilon^{r+1}$ and Lemma 3.4.1 that

$$\begin{aligned} h_{i(r)}(\tilde{x}^{r+1}, x^{r+1}) &= (\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)})\tilde{x}^{r+1}_{i(r)} - x^{r+1}_{i(r)}(\nabla_i f(x^{r+1}) + \tau_{i(r)}) \\ &\geq \varepsilon^{r+1}\tilde{x}^{r+1}_{i(r)} - x^{r+1}_{i(r)}(2\tau_{i(r)} + \varepsilon^{r+1}) \\ &= \varepsilon^{r+1}(\tilde{x}^{r+1}_{i(r)} - x^{r+1}_{i(r)}) - 2\tau_{i(r)}x^{r+1}_{i(r)} \\ &\geq -\varepsilon^{r+1}\left(\frac{2\mathcal{M}}{\|A_{i(r)}\|} + 2\tau_{i(r)}\right). \end{aligned}$$

It is clear that $h_{i(r)}(\tilde{x}^{r+1}, x^{r+1})$ in both cases (a) and (b) satisfies (3.4.6). ■

Using the above lemmas, we can show that $\{x^{r+1} - x^r\}$ converges to 0.

Lemma 3.4.3. *For the sequence $\{x^r\}$ generated by the ICD method, we have* $\lim_{r\to\infty}\{x^{r+1} - x^r\} = 0.$

Proof. We argue it by contradiction. Suppose that $x^{r+1}-x^r \nrightarrow 0$. Then there exists at least one coordinate $i \in \{1,2,\ldots,n\}$, a scalar $\gamma > 0$ and an infinite subset $\tilde{Z}$ of nonnegative integers such that $|x_i^{r+1} - x_i^r| \geq \gamma$ for all $r \in \tilde{Z}$. Since $\gamma > 0$, the index i is the index $i(r)$ chosen in Step 2 of the ICD method at the r-th step. Therefore, for any $j \neq i(r)$, we have $x_j^{r+1} = x_j^r$, which together with the assumption $|x_{i(r)}^{r+1} - x_{i(r)}^r| \geq \gamma$ implies that

$$\|A(x^{r+1} - x^r)\| = \|A_{i(r)}\||x_{i(r)}^{r+1} - x_{i(r)}^r| \geq \|A_{i(r)}\|\gamma, \;\; \forall r \in \tilde{Z}. \tag{3.4.9}$$

Since $\{Ax^r\}$ is bounded, there exist $t^{1,\infty},\ t^{2,\infty} \in \mathcal{R}^n$ and an infinite set $\mathcal{H} \subseteq \tilde{Z}$ such that

$$\lim_{r\to\infty,\, r\in\mathcal{H}} Ax^r = t^{1,\infty}, \lim_{r\to\infty,\, r\in\mathcal{H}} Ax^{r+1} = t^{2,\infty}. \tag{3.4.10}$$

Note that $t^{1,\infty} \neq t^{2,\infty}$ due to (3.4.9). It then follows from the continuity of g on $\Omega(F(x^0))$ and (3.4.10) that

$$\lim_{r\to\infty,\, r\in\mathcal{H}} g(Ax^r) = g(t^{1,\infty}), \lim_{r\to\infty,\, r\in\mathcal{H}} g(Ax^{r+1}) = g(t^{2,\infty}). \tag{3.4.11}$$

Since $F(x^r)$ is monotonically decreasing from Assumption 3.3.1(i) and $F(x^r) \geq F(x^*)$ holds for any optimal solution x^*, the sequence $\{F(x^r)\}$ is convergent. Let F^∞ be its limit. Then we have

$$\lim_{r\to\infty,\, r\in\mathcal{H}} F(x^r) = F^\infty, \lim_{r\to\infty,\, r\in\mathcal{H}} F(x^{r+1}) = F^\infty. \tag{3.4.12}$$

Moreover, by Lemma 3.4.2 and (3.4.12), we obtain

$$\lim_{r\to\infty,\, r\in\mathcal{H}} F(\tilde{x}^{r+1}) = \lim_{r\to\infty,\, r\in\mathcal{H}} F(x^{r+1}) - \lim_{r\to\infty,\, r\in\mathcal{H}} (F(x^{r+1}) - F(\tilde{x}^{r+1})) = F^\infty, \tag{3.4.13}$$

where $\tilde{x}^{r+1}$ is defined in (3.4.1). Since F is convex and $F(\tilde{x}^{r+1}) \leq F(x^{r+1}) \leq F(x^r)$ hold, we have

$$F(\tilde{x}^{r+1}) \leq F\left(\frac{x^r + x^{r+1}}{2}\right) \leq \frac{1}{2}F(x^r) + \frac{1}{2}F(x^{r+1}) \leq F(x^r).$$

Taking a limit on these inequalities, we obtain

$$\lim_{r\to\infty,\, r\in\mathcal{H}} F\left(\frac{x^{r+1} + x^r}{2}\right) = F^\infty. \tag{3.4.14}$$

On the other hand,

$$\begin{aligned} &\lim_{r\to\infty,\, r\in\mathcal{H}} F\left(\frac{x^{r+1} + x^r}{2}\right) \\ \leq &\lim_{r\to\infty,\, r\in\mathcal{H}} g(\frac{Ax^{r+1} + Ax^r}{2}) + \limsup_{r\to\infty,\, r\in\mathcal{H}} \left\{\langle b, \frac{x^{r+1} + x^r}{2}\rangle + \sum_{i=1}^n \tau_{i(r)} \left|\frac{x_{i(r)}^{r+1} + x_{i(r)}^r}{2}\right|\right\} \end{aligned}$$

$$
\begin{aligned}
&\le g(\frac{t^{1,\infty}+t^{2,\infty}}{2}) + \frac{1}{2}\limsup_{r\to\infty,\ r\in\mathcal{H}}\{\langle b, x^r\rangle + \sum_{i=1}^{n}\tau_{i(r)}|x^r_{i(r)}|\} + \frac{1}{2}\limsup_{r\to\infty,\ r\in\mathcal{H}}\{\langle b, x^{r+1}\rangle + \sum_{i=1}^{n}\tau_{i(r)}|x^{r+1}_{i(r)}|\}\\
&= g(\frac{t^{1,\infty}+t^{2,\infty}}{2}) + \frac{1}{2}\limsup_{r\to\infty,\ r\in\mathcal{H}}\{F(x^r) - g(Ax^r)\} + \frac{1}{2}\limsup_{r\to\infty,\ r\in\mathcal{H}}\{F(x^{r+1}) - g(Ax^{r+1})\}\\
&= g(\frac{t^{1,\infty}+t^{2,\infty}}{2}) + \frac{1}{2}(F^\infty - g(t^{1,\infty})) + \frac{1}{2}(F^\infty - g(t^{2,\infty}))\\
&< \frac{1}{2}(g(t^{1,\infty}) + g(t^{2,\infty})) + \frac{1}{2}(F^\infty - g(t^{1,\infty})) + \frac{1}{2}(F^\infty - g(t^{2,\infty}))\\
&= F^\infty,
\end{aligned}
$$

where the second inequality follows from the continuity of g and (3.4.10), the first equality follows from the definition of F, the second equality follows from (3.4.11) and (3.4.12), and the third inequality follows from the strict convexity of g and $t^{1,\infty} \neq t^{2,\infty}$. But this inequality contradicts (3.4.14). Thus $\lim_{r\to\infty}\{x^{r+1} - x^r\} = 0$. ∎

In the second part of this section, we will show the convergence of $\{Ax^r\}$. Since $\{Ax^r\}$ is bounded, there exist $t^\infty \in \mathcal{R}^n$ and an infinite set $\mathcal{X}$ such that

$$\lim_{r\to\infty,\ r\in\mathcal{X}} Ax^r = t^\infty. \tag{3.4.15}$$

Then with the continuity of ∇g, we have

$$\lim_{r\to\infty,\ r\in\mathcal{X}} \nabla f(x^r) = d^\infty, \tag{3.4.16}$$

where

$$d^\infty := A^T\nabla g(t^\infty) + b. \tag{3.4.17}$$

For the set $\mathcal{X}$, we have the following result with Lemma 3.4.3, which provides an interesting property associated with $\{\nabla f(x^r)\}$.

Lemma 3.4.4. *For any $s \in \{0, 1, \ldots, B-1\}$, where B is the integer defined in the almost cycle rule, we have $\lim_{r\to\infty,\ r\in\mathcal{X}} \nabla f(x^{r-s}) = d^\infty$.*

Proof. For any $s \in \{0, 1, \ldots, B-1\}$, we have $Ax^{r-s} = \sum_{k=0}^{s-1} A(x^{r-s+k} - x^{r-s+k+1}) + Ax^r$. It then follows from Lemma 3.4.3 and (3.4.15) that

$$\lim_{r\to\infty,\ r\in\mathcal{X}} Ax^{r-s} = \lim_{r\to\infty,\ r\in\mathcal{X}}\sum_{k=0}^{s-1} A(x^{r-s+k} - x^{r-s+k+1}) + \lim_{r\to\infty, r\in\mathcal{X}} Ax^r = t^\infty.$$

From the continuity of ∇g, we have $\lim_{r\to\infty,\ r\in\mathcal{X}} \nabla f(x^{r-s}) = \lim_{r\to\infty,\ r\in\mathcal{X}} A^T\nabla g(Ax^{r-s}) + b = A^T\nabla g(t^\infty) + b$, which together with (3.4.17) shows this lemma. ∎

Lemma 3.4.4 implies that for each $i \in \{1, 2, \ldots, n\}$, and $s \in \{0, 1, \ldots, B-1\}$, we have

$$\lim_{r\to\infty,\, r\in\mathcal{X}} \nabla_i f(x^{r-s}) = d_i^\infty. \tag{3.4.18}$$

For a fixed coordinate i, let $\varphi(r, i)$ denote the largest integer $\bar{r}$, which does not exceed r, such that the i-th coordinate of x is iterated upon at the $\bar{r}$-th iteration, that is, for all $r \in \mathcal{X}$, we have

$$x_i^r = x_i^{\varphi(r,i)}. \tag{3.4.19}$$

Since the coordinate is chosen by the almost cycle rule, the relation $r - B + 1 \leq \varphi(r, i) \leq r$ holds for all $r \in \mathcal{X}$. From (3.4.18), we further obtain

$$\lim_{r\to\infty,\, r\in\mathcal{X}} \nabla_i f(x^{\varphi(r,i)}) = d_i^\infty. \tag{3.4.20}$$

Now we define the following six index sets associated with d_i^∞ as

$$\begin{aligned}
J_1^\infty &:= \{i|\ d_i^\infty > \tau_i\};\\
J_2^\infty &:= \{i|\ d_i^\infty < -\tau_i\};\\
J_3^\infty &:= \{i|\ |d_i^\infty| < \tau_i\};\\
J_4^\infty &:= \{i|\ d_i^\infty = \tau_i,\ \tau_i > 0\};\\
J_5^\infty &:= \{i|\ d_i^\infty = -\tau_i,\ \tau_i > 0\};\\
J_6^\infty &:= \{i|\ d_i^\infty = 0,\ \tau_i = 0\}.
\end{aligned}$$

Note that $\bigcup_{i=1}^6 J_i^\infty = \{1, 2, \ldots, n\}$. Next two lemmas give sufficient conditions under which $\{x_i^r\}_\mathcal{X}$ is fixed or lies in some interval.

Lemma 3.4.5. *Suppose that Assumption 3.3.1(i) and (iii) hold. Let L_g and ε_0 be the constants given in Lemma 3.2.3. If $\varepsilon^{\varphi(r,i)} < \varepsilon_0$, then the following statements hold for any fixed i:*

(i) *If $\nabla_i f(x^{\varphi(r,i)}) - \tau_i > L_g\|A_i\|^2\varepsilon^{\varphi(r,i)}$ and $x_i^{\varphi(r,i)} \leq \varepsilon^{\varphi(r,i)} + l_i$ hold, then $x_i^{\varphi(r,i)} = l_i$.*

(ii) *If $\nabla_i f(x^{\varphi(r,i)}) + \tau_i < -L_g\|A_i\|^2\varepsilon^{\varphi(r,i)}$ and $u_i - \varepsilon^{\varphi(r,i)} \leq x_i^{\varphi(r,i)}$ hold, then $x_i^{\varphi(r,i)} = u_i$.*

(iii) *If $\nabla_i f(x^{\varphi(r,i)}) + \tau_i > L_g\|A_i\|^2\varepsilon^{\varphi(r,i)}$ and $|x_i^{\varphi(r,i)}| \leq \varepsilon^{\varphi(r,i)}$ hold, then $x_i^{\varphi(r,i)} \leq 0$.*

(iv) *If $\nabla_i f(x^{\varphi(r,i)}) - \tau_i < -L_g\|A_i\|^2\varepsilon^{\varphi(r,i)}$ and $|x_i^{\varphi(r,i)}| \leq \varepsilon^{\varphi(r,i)}$ hold, then $x_i^{\varphi(r,i)} \geq 0$.*

Proof. Here, we only show (i) and (iii). The rest can be obtained similarly.

To show (i), we argue by contradiction. If it is not true, then we have $l_i < x_i^{\varphi(r,i)} \leq \varepsilon^{\varphi(r,i)} + l_i$ by Assumption 3.3.1(ii). From the Lipschitz continuity of ∇g in Lemma 3.2.3, we obtain that $|\nabla_i f(\hat{x}^{\varphi(r,i)}) - \nabla_i f(x^{\varphi(r,i)})| \le L_g\|A_i\|^2|l_i - x_i^{\varphi(r,i)}|$, where $\hat{x}^{\varphi(r,i)} := (x_1^r, \ldots, x_{i-1}^r, l_i, x_{i+1}^r, \ldots, x_n^r)$. We further can ensure $\nabla_i f(\hat{x}^{\varphi(r,i)}) - \tau_i \ge -L_g\|A_i\|^2\varepsilon^{\varphi(r,i)} + \nabla_i f(x^{\varphi(r,i)}) - \tau_i > 0$ with the assumptions $l_i < x_i^{\varphi(r,i)} \le \varepsilon^{\varphi(r,i)} + l_i$ and $\nabla_i f(x^{\varphi(r,i)}) - \tau_i > L_g\|A_i\|^2\varepsilon^{\varphi(r,i)}$. It then follows from the KKT conditions in Lemma 3.2.1 that l_i is the exact solution of the subproblem (3.3.1). Since the solution of the subproblem (3.3.1) is unique, we have $F(x^{\varphi(r,i)}) - F(\hat{x}^{\varphi(r,i)}) > 0$, which contradicts Assumption 3.3.1(i). Therefore, we have $x_i^{\varphi(r,i)} = l_i$.

For (iii), we also prove by contradiction. Suppose that the contray holds, i.e., $x_i^{\varphi(r,i)} \in (0, \varepsilon^{\varphi(r,i)}]$. Let $\tilde{x}^{\varphi(r,i)} := (x_1^r, \ldots, x_{i-1}^r, 0, x_{i+1}^r, \ldots, x_n^r)$. Then, by Lemma 3.2.3 and the assumption $x_i^{\varphi(r,i)} \in (0, \varepsilon^{\varphi(r,i)}]$, we have

$$|\nabla_i f(\tilde{x}^{\varphi(r,i)}) - \nabla_i f(x^{\varphi(r,i)})| \le L_g\|A_i\|^2|0 - x_i^{\varphi(r,i)}| \le L_g\|A_i\|^2\varepsilon^{\varphi(r,i)},$$

which implies

$$-L_g\|A_i\|^2\varepsilon^{\varphi(r,i)} + \nabla_i f(x^{\varphi(r,i)}) \le \nabla_i f(\tilde{x}^{\varphi(r,i)}).$$

By the convexity of f, $0 < x_i^{\varphi(r,i)} \le \varepsilon^{\varphi(r,i)}$ and $\nabla_i f(x^{\varphi(r,i)}) + \tau_i > L_g\|A_i\|^2\varepsilon^{\varphi(r,i)}$, we further have that

$$F(x^{\varphi(r,i)}) - F(\tilde{x}^{\varphi(r,i)}) \ge \nabla_i f(\tilde{x}^{\varphi(r,i)})(x_i^{\varphi(r,i)} - 0) + \tau_i x_i^{\varphi(r,i)} > 0, \tag{3.4.21}$$

which contradicts Assumption 3.3.1(i). ■

Lemma 3.4.6. *Suppose that Assumption 3.3.1 holds. Then, for sufficiently large r, we have*

$$\{x_i^r\}_{\mathcal{X}} = l_i, \forall i \in J_1^\infty; \tag{3.4.22}$$
$$\{x_i^r\}_{\mathcal{X}} = u_i, \forall i \in J_2^\infty; \tag{3.4.23}$$
$$\{x_i^r\}_{\mathcal{X}} = 0, \forall i \in J_3^\infty; \tag{3.4.24}$$
$$l_i \le \{x_i^r\}_{\mathcal{X}} \le 0, \forall i \in J_4^\infty; \tag{3.4.25}$$
$$0 \le \{x_i^r\}_{\mathcal{X}} \le u_i, \forall i \in J_5^\infty; \tag{3.4.26}$$
$$l_i \le \{x_i^r\}_{\mathcal{X}} \le u_i, \forall i \in J_6^\infty. \tag{3.4.27}$$

Proof. Here we only show (3.4.22) and (3.4.25). Since the rest part can be shown in a similar way, we omit the proof.

Case 1: $i \in J_1^\infty$. To show (3.4.22), it is sufficient to show

$$\{x_i^{\varphi(r,i)}\}_{\mathcal{X}} = l_i, \tag{3.4.28}$$

since $x_i^r = x_i^{\varphi(r,i)}$ holds by (3.4.19). From (3.4.20), we have that for $\bar{\varepsilon} = \frac{d_i^\infty - \tau_i}{2} > 0$, $i \in J_1^\infty$, there exists a nonnegative integer $\bar{r}$ such that

$$d_i^\infty - \bar{\varepsilon} \le \nabla_i f(x^{\varphi(r,i)}) \le d_i^\infty + \bar{\varepsilon},\ \forall r \ge \bar{r}, r \in \mathcal{X}.$$

It is easy to see that $d_i^{\infty} - \tau_i - \bar{\varepsilon}$ is positive. Then we have

$$\nabla_i f(x^{\varphi(r,i)}) - \tau_i \geq d_i^{\infty} - \tau_i - \bar{\varepsilon} > \max\{1, L_g\|A_i\|^2\}\varepsilon^{\varphi(r,i)} \geq \varepsilon^{\varphi(r,i)} \tag{3.4.29}$$

for sufficiently large r, since $\varepsilon^r \to 0$ and $\nabla_i f(x^{\varphi(r,i)}) \to d_i^{\infty}$ hold. Furthermore, we ensure $i \in J_1(x^{\varphi(r,i)}, \varepsilon^{\varphi(r,i)})$, since $x_i^{\varphi(r,i)}$ is an $\varepsilon^{\varphi(r,i)}$-approximate solution of the subproblem (3.3.1). It implies that $|x_i^{\varphi(r,i)} - l_i| \leq \varepsilon^{\varphi(r,i)}$. Then by the Assumption 3.3.1(ii) and (3.2.7), we have

$$l_i \leq x_i^{\varphi(r,i)} \leq \varepsilon^{\varphi(r,i)} + l_i < 0. \tag{3.4.30}$$

Thus, the equality (3.4.28) follows from (3.4.29), (3.4.30) and Lemma 3.4.5(i), and hence (3.4.22) holds.

Case 2: $i \in J_4$. In this case, we have $d_i^{\infty} = \tau_i$ and $\tau_i > 0$. Let $\tilde{\varepsilon} = \frac{\tau_i}{2}$. It then follows from (3.4.20) that there exists an $\tilde{r}$, such that $\frac{1}{2}\tau_i < \nabla_i f(x^{\varphi(r,i)}) < \frac{3}{2}\tau_i$ hold for all $r \in \mathcal{X}$, $r \geq \tilde{r}$. Then for sufficiently large r, the inequalities

$$\nabla_i f(x^{\varphi(r,i)}) + \tau_i > \frac{3}{2}\tau_i > \max\{1, L_g\|A_i\|^2\}\varepsilon^{\varphi(r,i)} \geq \varepsilon^{\varphi(r,i)} \tag{3.4.31}$$

hold due to $\varepsilon^r \to 0$. We further obtain $i \in \bigcup_{j=1}^{3} J_j(x^{\varphi(r,i)}, \varepsilon^{\varphi(r,i)})$ from Definition 3.2.1. Therefore, we have

$$x_i^{\varphi(r,i)} \in [l_i, \varepsilon^{\varphi(r,i)}]. \tag{3.4.32}$$

It finally follows from (3.4.31), (3.4.32) and Lemma 3.4.5(iii) that $x_i^{\varphi(r,i)} \in [l_i, 0]$. Then, (3.4.25) holds from (3.4.19).

■

Next, we will show that $Ax^r \to Ax^*$, where x^* is an arbitrary optimal solution of problem (3.1.1). For this purpose, we recall Hoffman's error bound [31].

Lemma 3.4.7. *Let $B \in \mathcal{R}^{k\times n}$, $C \in \mathcal{R}^{k\times n}$ and $e \in \mathcal{R}^k$, $d \in \mathcal{R}^k$. Suppose that the linear system $By = e, Cy \leq d$ is consistent. Then there exists a scalar $\theta > 0$ depending only on B and C such that, for any $\bar{x} \in [l, u]$, l, $u \in \mathcal{R}^n$, there is a point $\bar{y} \in \mathcal{R}^n$ satisfying $B\bar{y} = e, C\bar{y} \leq d$ and $\|\bar{x} - \bar{y}\| \leq \theta(\|B\bar{x} - e\| + \|(C\bar{x} - d)_+\|)$, where $(x_i)_+ := \max\{0, x_i\}$.*

Theorem 3.4.1. *Let x^* be an optimal solution of problem (3.1.1). Then we have*

$$\lim_{r\to\infty} Ax^r = Ax^*.$$

Proof. In the first step, we show that $Ax^r \to Ax^*$ holds for $r \in \mathcal{X}$, where $\mathcal{X}$ is an infinite set given in (3.4.15). To this end, we consider the following linear system of y:

$$Ay = Ax^r,\ \ y_i = x_i^r\ (i \in J_1^\infty \cup J_2^\infty \cup J_3^\infty),\ y_i \le 0\ (i \in J_4^\infty),\ \text{and}\ y_i \ge 0\ (i \in J_5^\infty),\ y \in [l,u].$$

It follows from (3.4.22)-(3.4.27) that x^r is a solution of this system for sufficiently large r, that is, the system is consistent. For any fixed point $\bar{x}$ in $[l,u]$, by Lemma 3.4.7, there exists a solution $y^r \in [l,u]$ of the above system and a constant θ, which is independent of x^r, such that

$$\|y^r - \bar{x}\| \le \theta \left(\|A\bar{x} - Ax^r\| + \sum_{i \in J_1 \cup J_2 \cup J_3} |\bar{x}_i - x_i^r| + \sum_{i \in J_4} \max\{0, \bar{x}_i\} + \sum_{i \in J_5} \max\{0, -\bar{x}_i\} \right).$$

From the boundedness of $\{Ax^r\}$ and (3.4.22)-(3.4.24), we further have that the right-hand side of this inequality is bounded. It implies that $\{y^r\}_{\mathcal{X}}$ is also bounded, and hence it has at least one accumulation point. We denote it by y^∞. Furthermore, from (3.4.15) and Lemma 3.4.6, we have that y^∞ satisfies the following system:

$$Ay^\infty = t^\infty,\ \ y_i^\infty = l_i\ (i \in J_1),\ \ y_i^\infty = u_i\ (i \in J_2),\ \ y_i^\infty = 0\ (i \in J_3),$$

$$l_i \le y_i^\infty \le 0\ (i \in J_4),\ 0 \le y_i^\infty \le u_i\ (i \in J_5), l_i \le y_i^\infty \le u_i\ (i \in J_6).$$

It then follows from (3.4.17) that $\nabla f(y^\infty) = A^T \nabla g(Ay^\infty) + b = d^\infty$. Moreover, the relation $y^\infty = P_{\tau,l,u}(y^\infty)$ holds from the above system and Lemma 3.2.1. Thus, y^∞ is an optimal solution of problem (3.1.1) by Lemma 3.2.1. From Lemma 3.2.4, we have $Ay^\infty = Ax^*$, i.e., $t^\infty = Ax^*$.

In the second step, we show $\lim_{r\to\infty} Ax^r = Ax^*$. Since $\{Ax^r\}$ is bounded, it is sufficient to show that any accumulation point of $\{Ax^r\}$ is Ax^*. Let $\hat{\mathcal{X}}$ be any subset of nonnegative integers such that $\{Ax^r\}$ is convergent, and let $\hat{t}^\infty$ be a limit of $\{Ax^r\}_{\hat{\mathcal{X}}}$. Then we can show that $\hat{t}^\infty = Ax^*$ holds for the set $\hat{\mathcal{X}}$ as Lemmas 3.4.4-3.4.6. Moreover, the first step of the current proof, i.e., $\{Ax^r\}_{\hat{\mathcal{X}}} \to Ax^*$ holds. Thus, $\{Ax^r\} \to Ax^*$ holds for $r \to \infty$. ■

Theorem 3.4.1 implies that there exists a scalar $\bar{r} > 0$, such that $Ax^r \in B(Ax^*)$ for any $r \ge \bar{r}$, where $B(Ax^*)$ is the closed ball defined before (3.2.1). Note that g is strongly convex on $B(Ax^*)$.

In the third part of this section, we show the sufficient decreasing of $\{F(x^r)\}$ for sufficiently large r.

Lemma 3.4.8. *Under Assumption 3.3.1, there exists a scalar $\eta > 0$ such that $F(x^r) - F(x^{r+1}) \ge \eta\|x^r - x^{r+1}\|^2$ holds for sufficiently large r.*

Proof. Note that $Ax^r, Ax^{r+1} \in B(Ax^*)$ holds for sufficiently large r. It then follows from Assumption 3.2.1 that g is strongly convex in $B(Ax^*)$. Furthermore, we have

$$\begin{aligned}F(x^r) - F(x^{r+1}) &= g(Ax^r) - g(Ax^{r+1}) - \langle A^T\nabla g(Ax^{r+1}), x^r - x^{r+1}\rangle \\ &\quad + \langle \nabla f(x^{r+1}), x^r - x^{r+1}\rangle + \tau_{i(r)}|x^r_{i(r)}| - \tau_{i(r)}|x^{r+1}_{i(r)}| \\ &\geq \frac{\mu_g}{2}\|A(x^r - x^{r+1})\|^2 + \langle \nabla_{i(r)} f(x^{r+1}), x^r_{i(r)} - x^{r+1}_{i(r)}\rangle + \tau_{i(r)}\left(|x^r_{i(r)}| - |x^{r+1}_{i(r)}|\right) \\ &= \frac{\mu_g}{2}\|A_{i(r)}\|^2|x^r_{i(r)} - x^{r+1}_{i(r)}|^2 + h_{i(r)}(x^r, x^{r+1}) \\ &\geq \frac{\mu_g}{2}\min_j \|A_j\|^2\|x^r - x^{r+1}\|^2 + h_{i(r)}(x^r, x^{r+1}),\end{aligned}$$

where $h_{i(r)}$ is defined in (3.4.3), and $i(r)$ denotes the index chosen on the r-th step.

Next, we show the inequality

$$h_{i(r)}(x^r, x^{r+1}) \geq -\alpha_r\tilde{L}(x^r_{i(r)} - x^{r+1}_{i(r)})^2, \tag{3.4.33}$$

where $\tilde{L} := \max_j\{1, L_g\|A_j\|^2\}$, and α_r is given in Assumption 3.3.1(v). Note that $\tilde{L} \geq 1$.

We show it by considering 6 cases: $i(r) \in J^\infty_j$, $j = 1, 2, \ldots, 6$. First, we have from Lemma 3.4.6 that

$$h_{i(r)}(x^r, x^{r+1}) = 0, \ \forall\ i(r) \in \bigcup_{j=1}^{3} J^\infty_j.$$

Hence, (3.4.33) holds for $i(r) \in J^\infty_j$, $j = 1, 2, 3$. Then, we only need to consider the other three cases $i \in J^\infty_4$, $i \in J^\infty_5$ and $i \in J^\infty_6$. Here, for simplicity, we only show the case $i \in J^\infty_4$. The rest two cases can be obtained in a similar way.

If $i(r) \in J^\infty_4$, then it follows from Lemma 3.4.6 that for the sufficiently large r, $x^r_{i(r)}, x^{r+1}_{i(r)} \in [l_{i(r)}, 0]$ holds. Then we have

$$\begin{aligned}h_{i(r)}(x^r, x^{r+1}) &= \langle \nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)}, x^r_{i(r)} - x^{r+1}_{i(r)}\rangle \\ &\geq -|\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)}||x^r_{i(r)} - x^{r+1}_{i(r)}|.\end{aligned} \tag{3.4.34}$$

From the proof of (3.4.25) in Lemma 3.4.6, we have $i(r) \in \bigcup_{j=1}^{3} J_j(x^{r+1}, \varepsilon^{r+1})$. Thus we show (3.4.33) by considering the following three distinct cases.

Case 1: $i(r) \in J_1(x^{r+1}, \varepsilon^{r+1})$. We have by Assumption 3.3.1(ii) that

$$\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)} \geq -\varepsilon^{r+1} \text{ and } l_{i(r)} \leq x^{r+1}_{i(r)} \leq l_{i(r)} + \varepsilon^{r+1}. \tag{3.4.35}$$

The first inequality means that $\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)} \in [-\varepsilon^{r+1}, \infty) = [-\varepsilon^{r+1}, \tilde{L}\varepsilon^{r+1}] \cup (\tilde{L}\varepsilon^{r+1}, \infty)$. First suppose that $\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)} \in [-\varepsilon^{r+1}, \tilde{L}\varepsilon^{r+1}]$. It then follows from (3.4.34) and Assumption 3.3.1(iv) that $h_{i(r)}(x^r, x^{r+1}) \geq -\tilde{L}\varepsilon^{r+1}|x^r_{i(r)} - x^{r+1}_{i(r)}| \geq -\alpha_r \tilde{L}|x^r_{i(r)} - x^{r+1}_{i(r)}|^2$, which satisfies (3.4.33).

Next suppose that $\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)} \in (\tilde{L}\varepsilon^{r+1}, \infty)$. Then $x^{r+1}_{i(r)} = l_{i(r)}$ holds from $l_{i(r)} \leq x^{r+1}_{i(r)} \leq l_{i(r)} + \varepsilon^{r+1}$ and Lemma 3.4.5(i). Therefore, we get $h_{i(r)}(x^r, x^{r+1}) = \langle \nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)}, x^r_{i(r)} - l_{i(r)} \rangle \geq 0$, which implies (3.4.33) obviously.

Case 2: $i(r) \in J_2(x^{r+1}, \varepsilon^{r+1})$. In this case, we have $|\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)}| \leq \varepsilon^{r+1}$ and $l_{i(r)} \leq x^{r+1}_{i(r)} \leq 0$. From Assumption 3.3.1(iv) and (3.4.34), we have $h_{i(r)}(x^r, x^{r+1}) \geq -\varepsilon^{r+1}|x^r_{i(r)} - x^{r+1}_{i(r)}| \geq -\alpha_r |x^r_{i(r)} - x^{r+1}_{i(r)}|^2$, which also implies (3.4.33).

Case 3: $i(r) \in J_3(x^{r+1}, \varepsilon^{r+1})$. We have $|\nabla_{i(r)} f(x^{r+1})| \leq \tau_{i(r)} + \varepsilon^{r+1}$ and $-\varepsilon^{r+1} \leq x^{r+1}_{i(r)} \leq 0$, hence we have $\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)} \in [-2\tau_{i(r)} - \varepsilon^{r+1}, \varepsilon^{r+1}]$. If $\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)} \in [-\tilde{L}\varepsilon^{r+1}, \varepsilon^{r+1}]$, then (3.4.33) holds from Assumption 3.3.1(iv). If $\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)} \in [-2\tau_{i(r)} - \varepsilon^{r+1}, -\tilde{L}\varepsilon^{r+1})$, then we have $x^{r+1}_{i(r)} = 0$ from Lemma 3.4.5 and $x^{r+1}_{i(r)} \in [-\varepsilon^{r+1}, 0]$. Hence, we have $h_{i(r)}(x^r, x^{r+1}) = (\nabla_{i(r)} f(x^{r+1}) - \tau_{i(r)}) x^r_{i(r)} \geq 0 \geq -\alpha_r \tilde{L}(x^r_{i(r)} - x^{r+1}_{i(r)})^2$.

Consequently, the inequality (3.4.33) holds.

The sequence $\{\alpha_r\}$ satisfies $\alpha_r < \dfrac{\mu_g \min\limits_j \|A_j\|^2}{2\max\limits_j\{1, L_g\|A_j\|^2\}}$ for sufficiently large r from the Assumption 3.3.1(v). Then the inequality of this theorem holds for sufficiently large r with $\eta = \frac{\mu_g}{2}\min\limits_j \|A_j\|^2 - \alpha_r \max\limits_j\{1, L_g\|A_j\|^2\} > 0$. ■

In the last part of this section, before showing the global and linear convergence of $\{x^r\}$, we first recall a kind of the Lipschitz error bound in [66, 67, 44].

Lemma 3.4.9. *There exists a scalar constant $\kappa > 0$ such that*

$$\|Ax^r - Ax^*\| \leq \kappa \|x^r - P_{\tau,l,u}(x^r)\| \tag{3.4.36}$$

holds for any $Ax^r \in B(Ax^*)$.

Proof. Since g is strongly convex on $B(Ax^*)$ and ∇g is Lipschitz continuous, there exists a constant $\hat{\kappa} > 0$ such that $\|x^r - x^*(r)\| \leq \hat{\kappa}\|x^r - P_{\tau,l,u}(x^r)\|$, where $x^*(r)$ is a nearest solution from x^r [44, Lemma 4.4]. It then follows from Lemma 3.2.4 and $\|Ax^r - Ax^*\| \leq \|A\|\|x^r - x^*\|$ that (3.4.36) holds with $\kappa := \|A\|\hat{\kappa}$. ■

The following result is a direct extension of [43, Lemma 4.5(a)] to problem (3.1.1).

Lemma 3.4.10. *Under Assumption 3.3.1, there exists a constant $\omega > 0$ such that the inequality $\|Ax^r - Ax^*\| \leq \omega \sum\limits_{h=r}^{r+B-1} \|x^h - x^{h+1}\|$ holds for sufficiently large r.*

Proof. To show this lemma, by Lemmas 3.4.9, it is sufficient to show that there exists a constant $\hat{\omega} > 0$ such that $\|x^r - P_{\tau,l,u}(x^r)\| \leq \hat{\omega}\sum_{h=r}^{r+B-1} \|x^h - x^{h+1}\|$. Since $\|x^r - P_{\tau,l,u}(x^r)\| \leq \sqrt{n} \max_i |x_i^r - P_{\tau,l,u}(x^r)_i|$, we only need to show that there exists a constant $\tilde{\omega} > 0$ such that $|x_i^r - P_{\tau,l,u}(x^r)_i| \leq \tilde{\omega}\sum_{h=r}^{r+B-1} \|x^h - x^{h+1}\|$ holds for each $i \in \{1, 2, \ldots, n\}$.

Note that $Ax^r \in B(Ax^*)$ for sufficiently large r. For any fixed index $i \in \{1, 2, \ldots, n\}$, let $\psi(r, i)$ be the smallest integer $\mathbb{N}$ ($\mathbb{N} \geq r$) such that x_i^r is updated on the $\mathbb{N}$-th step. Then, we have

$$\begin{aligned}
&|x_i^r - P_{\tau,l,u}(x^r)_i| \\
=&\left|\sum_{h=r}^{\psi(r,i)-1} \left[(x_i^h - P_{\tau,l,u}(x^h)_i) - (x_i^{h+1} - P_{\tau,l,u}(x^{h+1})_i)\right] + (x_i^{\psi(r,i)} - P_{\tau,l,u}(x^{\psi(r,i)})_i)\right| \\
\leq& \sum_{h=r}^{\psi(r,i)-1} \left|\left[(x_i^h - P_{\tau,l,u}(x^h)_i) - (x_i^{h+1} - P_{\tau,l,u}(x^{h+1})_i)\right]\right| + \left|x_i^{\psi(r,i)} - P_{\tau,l,u}(x^{\psi(r,i)})_i\right|,
\end{aligned}$$

where the inequality follows from the triangle inequality.

It then follows from the the nonexpansive property (3.2.5) of the projection $P_{\tau,l,u}(x)$, Assumption 3.3.1(iv) and Theorem 3.2.2 that

$$|x_i^r - P_{\tau,l,u}(x^r)_i| \leq \sum_{h=r}^{\psi(r,i)-1} \left(2\left|x_i^h - x_i^{h+1}\right| + \left|\nabla_i f(x^h) - \nabla_i f(x^{h+1})\right|\right) + \alpha_r \left|x_i^{\psi(r,i)} - x_i^{\psi(r,i)-1}\right|.$$

Since $r + 1 \leq \psi(r, i) \leq r + B$ holds by the almost cycle rule, we obtain

$$|x_i^r - P_{\tau,l,u}(x^r)_i| \ \leq \sum_{h=r}^{r+B-1} \left(2\left|x_i^h - x_i^{h+1}\right| + \left|\nabla_i f(x^h) - \nabla_i f(x^{h+1})\right|\right) + \alpha_r \left|x_i^{\psi(r,i)} - x_i^{\psi(r,i)-1}\right|.$$

It then follows from the Lipschitz continuity of ∇g and Assumption 3.3.1 that

$$\begin{aligned}
|x_i^r - P_{\tau,l,u}(x^r)_i| &\leq (2 + \|A\|^2 L_g) \sum_{h=r}^{r+B-1} \left\|x^h - x^{h+1}\right\| + \alpha_r \left\|x^{\psi(r,i)} - x^{\psi(r,i)-1}\right\| \\
&\leq \left(2 + \|A\|^2 L_g + \frac{\mu_g \min_j \|A_j\|^2}{2\max_j\{1, L_g\|A_j\|^2\}}\right) \sum_{h=r}^{r+B-1} \left\|x^h - x^{h+1}\right\|,
\end{aligned}$$

where the first inequality follows from $\|x^h - x^{h+1}\| \geq |x_i^h - x_i^{h+1}|$.

Let $\tilde{\omega} = 2 + \|A\|^2 L_g + \frac{\mu_g \min_j \|A_j\|^2}{2\max_j\{1, L_g\|A_j\|^2\}}$. Then it is easy to see that $\tilde{\omega} > 0$. Thus the inequality of this lemma holds with $\omega = \kappa\sqrt{n}\tilde{\omega}$, where κ is given in Lemma 3.4.9. ■

Now we are ready to show the linear convergence of $\{F(x^r)\}$ and $\{x^r\}$.

Theorem 3.4.2. *Suppose that $\{x^r\}$ is generated by the ICD method with the almost cycle rule. Let F^* denote the optimal value of problem (3.1.1). Then $\{F(x^r)\}$ converges to F^* at least B-step Q-linearly.*

Proof. In the first step, we show the global convergence of the sequence $\{F(x^r)\}$. Let x^* be an optimal solution of problem (3.1.1). Then we have $F^* = F(x^*)$. It follows from the mean value theorem that there exists $\xi \in \mathcal{R}^n$, which is on the line segment that joins x^r with x^*, such that $g(Ax^r) - g(Ax^*) = \langle A^T\nabla g(A\xi), x^r - x^*\rangle$.

Since $Ax^r \to Ax^*$and $\nabla f(x^r) \to d^\infty$ hold, we have

$$d^\infty = \lim_{x\to\infty} \nabla f(x^r) = \lim_{x\to\infty} A^T\nabla g(Ax^r) + b = A^T\nabla g(Ax^*) + b = \nabla f(x^*). \tag{3.4.37}$$

Thus, we have

$$\begin{aligned}
&F(x^r) - F^* \\
&= \langle A^T\nabla g(A\xi) - A^T\nabla g(Ax^*), x^r - x^*\rangle + \langle A^T\nabla g(Ax^*) + b, x^r - x^*\rangle + \sum_{i=1}^n \tau_i(|x_i^r| - |x_i^*|) \\
&\le L_g\|A\xi - Ax^*\|\|A(x^r - x^*)\| + \langle A^T\nabla g(Ax^*) + b, x^r - x^*\rangle + \sum_{i=1}^n \tau_i(|x_i^r| - |x_i^*|) \\
&\le L_g\|A(x^r - x^*)\|^2 + \langle d^\infty, x^r - x^*\rangle + \sum_{i=1}^n \tau_i(|x_i^r| - |x_i^*|) \\
&= L_g\|A(x^r - x^*)\|^2 + \sum_{i=1}^n \left[d_i^\infty(x_i^r - x_i^*) + \tau_i(|x_i^r| - |x_i^*|)\right],
\end{aligned} \tag{3.4.38}$$

where the first inequality follows from the Lipschitz continuity of ∇g, and the second inequality follows from (3.4.37).

With the special structure of problem (3.1.1), we can show that for sufficiently large r,

$$d_i^\infty(x_i^r - x_i^*) + \tau_i(|x_i^r| - |x_i^*|) = 0, \ \forall i \in \{1, 2, \ldots, n\}. \tag{3.4.39}$$

We prove this by considering the distinct cases about the index sets J_j^∞, $j = \{1, 2, \ldots, 6\}$ since $\{1, 2, \ldots, n\} = \bigcup_{j=1}^6 J_j^\infty$. For simplicity, we only prove the cases $i \in J_1^\infty$ and $i \in J_4^\infty$. The other cases can be shown in a similar way. If $i \in J_1^\infty$, i.e., $d_i^\infty > \tau_i$, then it follows from Lemma 3.4.6 that $x_i^r = l_i$ for sufficiently large r. On the other hand, we have $\nabla_i f(x^*) > \tau_i$ by (3.4.37). It then follows from Lemma 3.2.1 that $x_i^* = l_i$. These two relations imply that (3.4.39) holds. If $i \in J_4$, i.e., $d_i^\infty = \tau_i$, it then follows from Lemma 3.4.6 that for sufficiently large r, $l_i \le x_i^r \le 0$. On the other hand, we have $\tau_i = \nabla_i f(x^*)$ by (3.4.37). It further implies that $l_i \le x^* \le 0$ from Lemma 3.2.1. Combining these three relations, we have that (3.4.39) holds.

Consequently, we have $0 \leq F(x^r) - F^* \leq L_g\|A(x^r - x^*)\|^2$ by (3.4.38) and (3.4.39). It implies $F(x^r) \to F^*$, since $Ax^r \to Ax^*$ holds, that is, $\{F(x^r)\}$ is globally convergent.

In the second step, we show the B-step Q-linear convergence rate of $\{F(x^r)\}$. To this end, we need to ensure that there exists a constant $c \in (0, 1)$ such that

$$F(x^{r+B}) - F^* \leq c\,(F(x^r) - F^*)\,. \tag{3.4.40}$$

From (3.4.38), (3.4.39) and Lemma 3.4.10, we have

$$F(x^r) - F^* \leq L_g\omega^2 \left(\sum_{h=r}^{r+B-1} \|x^h - x^{h+1}\| \right)^2 .$$

Letting $k = h - r + 1$, we further have that

$$\begin{aligned} F(x^r) - F^* &\leq L_g\omega^2 \left(\sum_{k=1}^{B} \|x^{k+r-1} - x^{k+r}\| \right)^2 \\ &\leq L_g\omega^2 B \sum_{k=1}^{B} \left(\|x^{k+r-1} - x^{k+r}\| \right)^2 . \end{aligned}$$

It then follows from Lemma 3.4.8 that

$$\begin{aligned} F(x^r) - F^* &\leq \frac{L_g\omega^2 B}{\eta} \sum_{k=1}^{B} \left(F(x^{k+r-1}) - F(x^{k+r}) \right) \\ &= \frac{L_g\omega^2 B}{\eta} \left(F(x^r) - F(x^{r+B}) \right) . \end{aligned}$$

By rearranging the items of the above inequality, we have

$$F(x^{r+B}) - F^* \leq c\,(F(x^r) - F^*)\,, \tag{3.4.41}$$

where $c = 1 - \frac{\eta}{L_g\omega^2 B}$. Since $\frac{\eta}{L_g\omega^2 B} > 0$ and $c < 1$, it means that $\{F(x^r)\}$ converges to F^* at least B-step Q-linearly. ■

Theorem 3.4.3. *Suppose that $\{x^r\}$ is generated by the ICD method with the almost cycle rule. Then $\{x^r\}$ converges to an optimal solution of problem (3.1.1) at least R-linearly.*

Proof. First we show that $\{x^r\}$ is convergent. Let F^* be the optimal value of problem (3.1.1). Since $F(x^r)$ converges to F^* at least Q-linearly by Theorem 3.4.2, we have that $F(x^r)$ converges to F^* at least R-linearly, that is, there exist constants $K > 0$ and $\hat{c} \in (0, 1)$ such that

$$F(x^r) - F^* \leq K\hat{c}^r. \tag{3.4.42}$$

From Lemma 3.4.8, we have for sufficiently large r,

$$0 \le \|x^r - x^{r+1}\|^2 \le \frac{1}{\eta}\left(F(x^r) - F^*\right) + \frac{1}{\eta}\left(F^* - F(x^{r+1})\right) \le \frac{1}{\eta}\left(F(x^r) - F^*\right), \qquad (3.4.43)$$

where the last inequality holds since $F^* - F(x^{r+1}) \le 0$.

By combining (3.4.42) with (3.4.43), we have that $\|x^r - x^{r+1}\|^2 \le \frac{K}{\eta}\hat{c}^r$, that is, $\|x^r - x^{r+1}\| \le \sqrt{\frac{K}{\eta}}\hat{c}^{\frac{r}{2}}$. Let $\bar{c} := \hat{c}^{\frac{1}{2}}$. Then, we have $\bar{c} \in (0,1)$. Moreover, we obtain, for any positive integer m, n and $m > n$,

$$\|x^m - x^n\| \le \sum_{k=0}^{m-n-1} \|x^{m-k} - x^{m-k-1}\| \le \sqrt{\frac{K}{\eta}} \sum_{k=0}^{m-n-1} \bar{c}^{m-k-1} = \sqrt{\frac{K}{\eta}\frac{\bar{c}^n - \bar{c}^m}{1-\bar{c}}} \le \sqrt{\frac{K}{\eta}\frac{\bar{c}^n}{1-\bar{c}}},$$

which implies that $\{x^r\}$ is a cauchy sequence due to $0 < \bar{c} < 1$. Therefore, $\{x^r\}$ is convergent.

In the rest, we show that $\{x^r\}$ converges to an optimal solution at least R-linearly. Let x^∞ denote the limit point of $\{x^r\}$. Since $\|x^m - x^n\| \le \sqrt{\frac{K}{\eta}\frac{\bar{c}^n - \bar{c}^m}{1-\bar{c}}}$, we have

$$\|x^\infty - x^n\| = \lim_{m\to\infty}\|x^m - x^n\| \le \lim_{m\to\infty}\sqrt{\frac{K}{\eta}\frac{\bar{c}^n - \bar{c}^m}{1-\bar{c}}} = \sqrt{\frac{K}{\eta}\frac{\bar{c}^n}{1-\bar{c}}},$$

which implies that $\{x^r\}$ converges to x^∞ at least R-linearly since $0 < \bar{c} < 1$ holds.

Finally, we complete the proof by showing that the x^∞ is an optimal solution. With the continuity of F, we have $\lim_{r\to\infty} F(x^r) = F(x^\infty)$. It then follows from $F(x^r) \to F^*$ in Theorem 3.4.2 that $F(x^\infty) = F^*$, that is, x^∞ is also an optimal solution of problem (3.1.1). ■

3.5 Numerical experiments

In this section, we present some numerical experiments of the ICD method (the proposed method) for the following unconstrained l_1-regularized logistic regression problem.

$$\underset{w\in\mathcal{R}^{n-1}, v\in\mathcal{R}}{\text{minimize}} \quad F(x) := \frac{1}{m}\sum_{j=1}^{m}\log(1+\exp(-(w^Tq^j + vp^j))) + \mu\|w\|_1, \qquad (3.5.1)$$

where $x = (w, v) \in \mathcal{R}^n$ and $q^j = p^j z^j$. Moreover, $(z^j, p^j) \in \mathcal{R}^{n-1} \times \{-1, 1\}, j = 1, 2, \dots, m$ are a set of training examples. For simplicity, we let $f(x) = \frac{1}{m}\sum_{j=1}^{m}\log(1+\exp(-(w^Tq^j + vp^j)))$ and $\tau = (\mu, \dots, \mu, 0)^T \in \mathcal{R}^n$. Note that the computational costs of evaluating $f(x)$, $\nabla_i f(x)$ and $\nabla^2_{ii} f(x)$ are $O(m)$ if we update only one variable x_i on each step and store $\beta = Bx$, where $B = [Q, p]$ with $Q^T = [q^1, \dots, q^m]$ and $p = (p^1, \dots, p^m)^T \in \mathcal{R}^m$. This is because $f(x) = \frac{1}{m}\sum_{j=1}^{m}\log(1+\exp(-\beta_j))$ and $\beta^{\text{new}} = \beta^{\text{old}} + (x_i^{\text{new}} - x_i^{\text{old}})B_i$.

We report some numerical results on randomly generated problems for various inexact criteria satisfying Assumption 3.3.1. We also show the comparison with the CGD method [69].

3.5.1 Implementations

We exploit the following gradient method with line search to solve the one dimensional subproblem (3.3.1) in the ICD method.

Algorithm 3-1:

Step 0: Let $i := i(r)$ and $G_i := \nabla_i f(x^r)$. If $\mathrm{mid}\{G_i + \tau_i, G_i - \tau_i, x_i^r\} = 0$, then set $x_i^{r+1} = x_i^r$ and return. Otherwise let $k = 0$, $G_i^0 = G_i$, and $y^0 = x^r$. Go to step 1.

Step 1: Choose a scaling factor $s_{ii}^k > 0$. Calculate a search direction d^k as follows.

$$d^k = \operatorname*{argmin}_{d \in \mathcal{R}} \left\{ G_i^k d + \tau_i |y_i^k + d| + \frac{s_{ii}^k}{2} d^2 \right\}.$$

Step 2: Determine a stepsize α^k by the Armijo rule in [69] with $\gamma = 0$.

Step 3: Set $y_i^{k+1} = y_i^k + \alpha^k d^k$, $y_j^{k+1} = x_j^r$ for all $j \neq i$, and $G_i^{k+1} = \nabla_i f(y^{k+1})$. If the inexact criterion is satisfied, then set $x_i^{r+1} = y_i^{k+1}$ and return. Otherwise let $k = k + 1$. Go to Step 1.

The difference between the ICD method and the CGD method [69] lies in Step 3 of Algorithm 3-1. The CGD method does not check the inexact criterion in Step 3 and always returns to the main algorithm with $k = 0$. On the other hand, the ICD method returns to the main algorithm only when the inexact criterion holds. Note that if the criteria are weak, then the ICD method may be regarded as the CGD method.

In the numerical experiments, we choose the scaling factor s_{ii}^k in Step 1 according to the following three options.

(i) $s_{ii}^k = \nabla_{ii}^2 f(y^k)$;

(ii) $s_{ii}^k = 1$;

(iii) $s_{ii}^0 = 1$ and $s_{ii}^k = \frac{G_i^k - G_i^{k-1}}{y_i^k - y_i^{k-1}}$ for $k \geq 1$.

The choice (i) corresponds to the Newton method, while choice (ii) conforms to the steepest descent method. The option (iii) is motivated by the quasi-Newton method.

Additionally, we exploit the under/over-relaxation technique in the numerical experiments. Note that $P_{\tau,l,u}(x) = T_\tau(x - \nabla f(x))$ when $l = -\infty$ and $u = +\infty$. Let x_i^{r+1} be an ε^{r+1}-approximate solution of subproblem (3.3.1), i.e., $|x_i^{r+1} - T_\tau(x^{r+1} - \nabla f(x^{r+1}))_i| \leq \varepsilon^{r+1}$, and $\bar{x}^{r+1}$ be an under/over-relaxation estimator to x^{r+1} with parameter ω such that

$$\begin{aligned} \bar{x}_i^{r+1} &= \omega x_i^{r+1} + (1 - \omega) x_i^r, \\ \bar{x}_j^{r+1} &= x_j^{r+1}, \forall j \neq i. \end{aligned} \tag{3.5.2}$$

If the gradient of the function f in (3.5.1) is Lipschitz continuous with Lipschitz constant L_f, we have

$$\begin{aligned}|\bar{x}_i^{r+1} - T_\tau(\bar{x}^{r+1} - \nabla f(\bar{x}^{r+1}))_i| &\le \left|x_i^{r+1} - T_\tau(x^{r+1} - \nabla f(x^{r+1}))_i\right| + (2 + L_f)\left|(\omega - 1)(x_i^{r+1} - x_i^r)\right| \\ &\le \varepsilon^{r+1} + (2 + L_f)|\omega - 1|\left|x_i^{r+1} - x_i^r\right| \\ &\le \left(a_r + (2 + L_f)|\omega - 1|\right)\left|x_i^{r+1} - x_i^r\right|,\end{aligned}$$

where the last inequality follows from Assumption (3.3.1). Let $\bar{a}_r = a_r + (2 + L_f)|\omega - 1|$. If $\delta_r > \bar{a}_r|x_i^{r+1} - x_i^r|$, then $\bar{x}_i^{r+1}$ is an $\bar{\varepsilon}^{r+1}$-approximate solution, where $\bar{\varepsilon}^{r+1} = \min\{\delta_r, \bar{a}_r|x_i^{r+1} - x_i^r|\}$. This condition usually holds when δ_r slowly converges to 0, e.g., $\delta_r = O(\frac{1}{r})$.

3.5.2 Test problems

We generate the training examples randomly as in [35]. In our implementation, we have generated 8 random problems. Four of them have the scale of $n = 1001, m = 100$, and the others are $n = 101, m = 1000$. All training examples have an equal number of positive ($p^j = 1$) and negative ($p^j = -1$) training examples. Each feature q_i^j of positive (negative) examples q^j obeys independent and identical distribution. In our implementation, we adopt the normal distribution $\mathcal{N}(v, 1)$, where the mean v is drawn from a uniform distribution on $[0, 1]$ for positive examples ($[-1, 0]$ for negative examples).

We choose the regularized parameter μ based on $\mu_{\max} = \frac{1}{m}\left\|\frac{m_-}{m}\mu_{g_{p^j=1}}q^j + \frac{m_+}{m}\mu_{g_{p^j=-1}}q^j\right\|_\infty$, where m_- denotes the number of negative examples, and m_+ denotes the number of positive examples. It is shown in [35] that the vector $x = 0 \in \mathcal{R}^n$ is the optimal solution of problem (3.5.1) if $\mu \ge \mu_{\max}$. In our implementation, we let $\mu = 0.1\mu_{\max}$ or $\mu = 0.01\mu_{\max}$.

3.5.3 Numerical results

In this subsection, we give some numerical examples to illustrate the performances of the ICD method. The algorithm is implemented in MATLAB (Version 7.10.0), and run on an Intel(R) Core(TM)2 Duo CPU E6850 @3.00GHz. We terminate the algorithms when

$$\|x^r - T_\tau(x^r - \nabla f(x^r))\|_\infty \le 10^{-3}. \tag{3.5.3}$$

To save the CPU time, we check the termination condition in every 100 iterations. Throughout the experiments, we choose all initial points $x^0 = 0$, and adopt the simple cycle rule to choose i for the ICD method and the CGD method.

Investigation of the inexact criteria

To see the performances of the ICD method on various inexact criteria, we solve two random problems with

$$\varepsilon^r = \min\{\frac{10}{r^{\lfloor \frac{r}{n} \rfloor}}, a^{\lfloor \frac{r}{n} \rfloor}|x_i^{r+1} - x_i^r|\}, \tag{3.5.4}$$

where a varies from 0.1 to 0.8. Here, we use $\lfloor \frac{r}{n} \rfloor$ to reduce its sensitivity to r. In these experiments, we choose $s_{ii}^k = \nabla_{ii}^2 f(y^k)$ in Step 2 of Algorithm 3-1. We also use the same s_{ii}^k for the CGD method.

Table 3.1: Performances of the ICD method with various a in (3.5.4) and the CGD method.

	ICD ($a = 0.1$)	ICD ($a = 0.3$)	ICD ($a = 0.5$)	ICD ($a = 0.8$)	CGD
Problem 1	$n = 1001, m = 100, \mu = 0.01\mu_{\max}$				
iteration	9200	9200	9200	9200	9200
♯ of g	10334	9974	9856	9856	9199
♯ of f	2002	1485	1316	1316	1316
CPU time (s)	1.3125	1.1875	1.0469	1.0468	0.9531
Problem 2	$n = 101, m = 1000, \mu = 0.1\mu_{\max}$				
iteration	3300	3300	3300	3300	3300
♯ of g	9904	9299	8634	6014	3299
♯ of f	14262	13493	12235	5432	5432
CPU time (s)	3.9218	3.5781	3.3593	1.9062	1.6406

Table 3.1 presents the total number of evaluating G_i^k and f, the iteration r, and the CPU time (in seconds) for these two problems. From Table 3.1, we find that the ICD method performs better when a approaches to 1, yet it is worse than the CGD method. The results indicate that the solution of the subproblem (3.3.1) with high accuracy does not always improve the convergence. Note that the number of the gradient evaluations for the ICD method is larger than that for the CGD method. This is because the ICD method evaluates both $G_i^0 = \nabla_i f(y^0)$ and $G_i^1 = \nabla_i f(y^1)$ even if the Algorithm 3-1 is terminated at Step 3 with $k = 0$. However, the CGD method only evaluates G_i^0 at each iteration.

Comparison of the ICD method and the CGD method

We first show some numerical results for the ICD method and the CGD method with the Hessian information, that is, $s_{ii}^k = \nabla_{ii}^2 f(y^k)$. The ICD method is implemented with under the relaxation technique ($\omega = 0.5 \sim 1.0$ in (3.5.2)) and $\varepsilon^r = \max\{10^{-4}, \min\{10/r^{\lfloor \frac{r}{n} \rfloor}, 0.8^{\lfloor \frac{r}{n} \rfloor}|x_i^{r+1} -$

$x_i^r|\}\}$. Table 3.2 reports the numerical results for four instances. From Table 3.2, we see that the performances on the ICD method with $\omega = 1.0$ and the CGD method are roughly same since both of them exploit the Hessian information. The ICD method with appropriate relaxation factor ($\omega < 1.0$) is faster than the CGD method for some problems. The performances of the ICD method with over relaxation, i.e., $\omega > 1$, is worse for these four instances and hence we omit them.

Next we consider the case where the Hessian $\nabla^2_{ii} f(y^k)$ is not available. Then we may choose s^k_{ii} as in the steepest descent method ($s^k_{ii} = 1$) or in the quasi-Newton method. Note that the CGD method can not adopt the quasi-Newton method since it returns with $k = 0$ in Algorithm 3-1. Table 3.3 reports the performances of the ICD method combined with the quasi-Newton method and the CGD method with $s^k_{ii} = 1$. We also give results for the CGD method with $s^k_{ii} = \nabla^2_{ii} f(y^k)$ for the better understanding.

From Table 3.3, we find that the ICD method combined with the quasi-Newton method performs similarly as the CGD method with $s^k_{ii} = \nabla^2_{ii} f(y^k)$, but much better than the CGD method with $s^k_{ii} = 1$. Hence, if the Hessian computation for the function f is expensive, then the ICD method combined with the quasi-Newton method is an efficient alternative approach.

3.6 Conclusion

In this chapter, we have presented a framework of the ICD method for solving l_1-regularized convex optimization (3.1.1). We also have established the R-linear convergence rate of this method with the almost cycle rule. The key to the ICD method lies in Assumption 3.3.1 for the "inexact solutions". At each iteration, we only need to find an approximate solution, which raises the possibility to solve general l_1-regularized convex problems.

If we set $\varepsilon = 0$ in Assumption 3.3.1, then the ICD method reduces to the classical CD method. It then follows from Theorem 4.5.2 that the classical CD method has R-linear convergence rate for the l_1-regularized optimization problem (3.1.1) as well.

Table 3.2: Comparison of the ICD method and the CGD method for $s_{ii}^k = \nabla_{ii}^2 f(y^k)$.

	ICD	ICD	ICD	ICD	ICD	ICD
	$(\omega = 0.5)$	$(\omega = 0.6)$	$(\omega = 0.7)$	$(\omega = 0.8)$	$(\omega = 1.0)$	CGD
Problem 3	$n = 1001, m = 100, \mu = 0.01\mu_{\max}$					
iteration	13200	12200	9200	11200	13200	13200
$\sharp$ of g	15299	13945	10377	12620	14247	13199
$\sharp$ of f	4064	3494	2358	2844	2098	2098
CPU time (s)	1.8281	1.6406	1.2187	1.5468	1.5000	1.4375
Problem 4	$n = 1001, m = 100, \mu = 0.1\mu_{\max}$					
iteration	28400	28400	31000	31000	34100	34100
$\sharp$ of g	32504	32174	34507	34115	36086	34099
$\sharp$ of f	8212	7552	7018	6234	3976	3976
CPU time (s)	3.6093	3.5156	3.7968	3.562	3.7031	3.6406
Problem 5	$n = 101, m = 1000, \mu = 0.01\mu_{\max}$					
iteration	400	500	500	600	700	700
$\sharp$ of g	747	915	899	1070	1204	699
$\sharp$ of f	698	834	802	944	1012	1012
CPU time (s)	0.2656	0.2968	0.3906	0.3437	0.5156	0.3750
Problem 6	$n = 101, m = 1000, \mu = 0.1\mu_{\max}$					
iteration	1700	1900	1900	2600	2700	2700
$\sharp$ of g	3191	3570	3554	4896	4882	2699
$\sharp$ of f	2986	3344	3312	4596	4368	4368
CPU time (s)	1.1093	1.2812	1.2656	1.6718	1.6250	1.4687

Table 3.3: Performances of the ICD method and the CGD method when $\nabla^2_{ii} f(y^k)$ is not available.

	ICD (quasi-Newton)	CGD ($s^k_{ii} = 1$)	CGD ($s^k_{ii} = \nabla^2_{ii} f(y^k)$)
Problem 7	$n = 1001, m = 100, \mu = 0.1\mu_{\max}$		
iteration	33100	95000	34100
$\sharp$ of g	37909	94999	34099
$\sharp$ of f	8922	17384	3976
CPU time (s)	3.9843	9.9218	3.5937
Problem 8	$n = 101, m = 1000, \mu = 0.01\mu_{\max}$		
iteration	700	4400	700
$\sharp$ of g	1815	4399	699
$\sharp$ of f	1730	8800	1012
CPU time (s)	0.6406	2.1875	0.3437

Chapter 4

Block coordinate proximal gradient methods with variable Bregman functions for nonsmooth separable optimization problem

4.1 Introduction

In this chapter, we consider the following nonsmooth nonconvex optimization problem.

$$\underset{x}{\text{minimize}}\, F(x) := f(x) + \tau\psi(x), \tag{4.1.1}$$

where $\psi : \mathcal{R}^n \to (-\infty, \infty]$ is a proper, convex, and l.s.c. function with a block separable structure, $f : \mathcal{R}^n \to \mathcal{R}$ is smooth on an open subset of $\mathcal{R}^n$ containing $\operatorname{dom}\psi = \{x \in \mathcal{R}^n \mid \psi(x) < \infty\}$, and τ is a positive constant.

Throughout this chapter, we do not assume that the function f is convex, and hence, we are only concerned about obtaining the stationary points of problem (4.1.1).

As described in Subsection 1.2.3, the applications [29, 35, 55, 77] of problem (4.1.1) are mostly built on large scales. In general, the number of the variables is of order 10^4 or even higher. Hence, the classical second order methods can not be applied efficiently. Recently, "block" type first order methods have been investigated extensively for solving these large-scale problems. Tseng [67] proposed a block coordinate descent (BCD) method to solve a nondifferentiable nonconvex optimization problem with certain separability of the objective function. He proved that the BCD method has a global convergence property under appropriate assumptions. However, the convergence rate of the BCD method remains unknown. Tseng and Yun [69] proposed a block coordinate gradient descent (BCGD) method for problem (4.1.1), which may be viewed as a hybrid of the BCD and gradient methods.

In [69], the global convergence and the R-linear convergence rate of the BCGD method are established. Another related method, called the accelerated block coordinate relaxation (ABCR) method, has been proposed by Wright [71]. One of his significant contributions is that he adopted the reduced Newton step to achieve rapid convergence.

In this chapter, we propose a class of block coordinate proximal gradient (BCPG) methods for solving the nonsmooth nonconvex problem (4.1.1). As presented in Subsection 1.3.1, the search direction of the BCPG method at the r-th step is generated by

$$d_{\eta^r}(x^r; J^r) = \operatorname*{argmin}_{d\in\mathcal{R}^n} \left\{ \langle \nabla f(x^r), d\rangle + B_{\eta^r}(x^r+d, x^r) + \tau\psi(x^r+d) \middle|\, d_{\bar{J}^r} = 0 \right\}, \tag{4.1.2}$$

where $J^r \subseteq \mathcal{N}$ is the index set selected at the r-th step, $B_{\eta^r}(\cdot,\cdot) : X \times \mathrm{int}X \to \mathcal{R}$ is the Bregman function, defined by (1.3.5) in Chapter 1.3.1, and function $\eta^r : X \to \mathcal{R}$, called the "kernel of B_{η^r}", is assumed to be convex and continuously differentiable on $\mathrm{int}X$ and $X \subseteq \mathrm{dom}\, F$ is a closed convex set. For simplicity, we use the notation d^r instead of $d_{\eta^r}(x^r; J^r)$ in this chapter when it is clear from the context.

It is worth mentioning that kernels $\{\eta^r\}$ are not fixed for different iterations in this chapter, which yield at least three advantages.

- They allow us to obtain many well-known algorithms from the proposed BCPG methods. See Table 4.1, Subsections 4.3.1 and 4.6.1 for details.
- Some special kernels enable the BCPG methods to adopt a fixed step size. See Lemma 4.5.1, Theorems 4.6.1 and 4.6.5, and Algorithm 4-1 for details. This property is appealing when the evaluations of the functions in the line search are expensive.
- We may obtain accelerated algorithms by changing kernels when the iteration point is close to a solution. See Algorithm 4-1 in Section 4.7 for details.

For the proposed BCPG methods, we first prove their global convergence (Theorem 4.4.1 in Section 4.4) with the generalized Gauss-Seidel rule and establish their R-linear convergence rate (Theorem 4.5.2 in Section 4.5) under certain additional assumptions. As a consequence of this result, the (inexact) BCD method is shown to have at least an R-linear convergence rate for solving nonsmooth problem (4.1.1) (Theorems 4.6.2 and 4.6.5 in Section 4.6). To our knowledge, this is the first result on the linear convergence of the BCD type methods for nonsmooth problem (4.1.1). Finally, we propose a specific algorithm of the BCPG methods with variable kernels for a convex problem with separable simplex constraints (Algorithm 4-1 in Section 4.7). The numerical results show that the proposed algorithm performs better than the algorithm with a fixed kernel.

This chapter is organized as follows. In Section 4.2, we introduce some basic concepts and properties, which will be used in the subsequent analysis. In Section 4.3, we present a framework of the BCPG methods and introduce propositions about the stationary points. Then, we investigate their global convergence in Section 4.4 and determine the linear convergence rate in Section 4.5. Some special BCPG methods are further discussed in Section 4.6, and the numerical experiments are presented in Section 4.7. Finally, we conclude this chapter in Section 4.8.

Table 4.1: Reduced BCPG methods with special kernels

Special Kernel	Reduced BCPG Methods
$\eta^r(v) := f(v_{J^r}, v^r_{\bar{J}^r}) + \frac{1}{2}\lvert v_{\bar{J}^r}\rvert^2$ $J^r = \{(r \bmod n) + 1\}$	Coordinate descent method[1]
$\eta^r(v) := \frac{1}{2}v^T H^r v,\ H^r \succeq 0$	BCGD method [69]
$\eta^r(v)$ is fixed for all r, $J^r = \mathcal{N}$	Proximal gradient method [66] Steepest descent method Proximal point method Exponentiated gradient method
$\eta^r(v) := \frac{1}{2}v^T B_r v,\ J^r = \mathcal{N}$	Quasi-Newton method Newton method Regularized Newton method

4.2 Preliminaries

In this section, we introduce some useful properties for the Bregman function $B_{\eta^r}(\cdot,\cdot)$ defined in (1.3.5), and present some important properties for a convex function. First, we define a class of kernel functions (strongly convex functions) for the Bregman function $B_{\eta^r}(\cdot,\cdot)$.

Definition 4.2.1. *For a given positive constant $\underline{\mu}$, let $\Phi(X;\underline{\mu})$ denote a set of functions $\eta : X \to \mathcal{R}$ such that the following conditions hold.*

(i) *The function η is a closed proper differentiable function on* int$X$*;*

(ii) *The function η is μ_η-strongly convex on X, i.e., $\eta(y) \geq \eta(x) + \langle \nabla\eta(x), y - x\rangle + \frac{\mu_\eta}{2}\|y - x\|^2$ holds for any $y \in X$, $x \in$ int$X$, where $\mu_\eta \geq \underline{\mu} > 0$.*

[1] For the coordinate descent method, we assume that function f is strongly convex with respect to each element. Note that function f can be nonconvex with respect to the whole variable in this case. It is shown in [74] that the block coordinate descent method is convergent when f is only block wise strongly convex.

For simplicity, we define a subset $\Psi(X;\underline{\mu},\overline{L})$ of $\Phi(X;\underline{\mu})$ as follows.

$$\Psi(X;\underline{\mu},\overline{L}) := \{\eta \in \Phi(X;\underline{\mu}) \mid \|\nabla\eta(x) - \nabla\eta(y)\| \leq L_\eta\|x-y\|, 0 < L_\eta \leq \overline{L}, \forall x, y \in \mathrm{int}X\}.$$

Note that any function in $\Psi(X;\underline{\mu},\overline{L})$ is not only strongly convex but also gradient Lipschitz continuous. A simple example in the class $\Psi(X;\underline{\mu},\overline{L})$ is $\eta^r(x) = \frac{1}{2}\langle H^r x, x\rangle + \langle b, x\rangle + a$, where a, $b \in \mathcal{R}^n$ and $H^r \in \mathcal{R}^{n\times n}$ is a symmetric positive definite matrix such that $\overline{L}I \succeq H^r \succeq \underline{\mu}I$. The inequality $B_{\eta^r}(x,y) \geq 0$ holds for all x, $y \in \mathrm{int}X$ since $\eta^r \in \Phi(X;\underline{\mu})$ is convex. For convenience, we use $B_\eta(x,y)$ instead of $B_{\eta^r}(x,y)$ when it is clear from the context.

Next, we state some useful properties related to the Bregman function $B_\eta(\cdot,\cdot)$. Let $\nabla_1 B_\eta(\cdot,\cdot)$ denote the gradient of $B_\eta(\cdot,\cdot)$ with respect to the first variable, i.e.,

$$\nabla_1 B_\eta(x,y) = \nabla\eta(x) - \nabla\eta(y). \tag{4.2.1}$$

The following lemma, called the“three-point identity theorem”, is originally presented in [17, Lemma 3.1].

Lemma 4.2.1. *For any $\eta \in \Phi(X;\underline{\mu})$, a, b∈ int$X$, and $c \in X$, we have*

$$B_\eta(c,a) + B_\eta(a,b) - B_\eta(c,b) = \langle\nabla\eta(b) - \nabla\eta(a), c-a\rangle, \tag{4.2.2}$$

$$B_\eta(a,b) - B_\eta(c,b) \leq -\langle\nabla_1 B_\eta(a,b), c-a\rangle. \tag{4.2.3}$$

Proof. By the definition of B_η in (1.3.5), we can verify equality (4.2.2) easily. Inequality (4.2.3) holds because of (4.2.1) and $B_\eta(c,a) \geq 0$. ■

Lemma 4.2.2. *For any $\eta \in \Psi(X;\underline{\mu},\overline{L})$, the following two inequalities hold.*

$$\tfrac{\overline{L}}{2}\|x-y\|^2 \geq B_\eta(x,y) \geq \tfrac{\underline{\mu}}{2}\|x-y\|^2, \forall x,\ y \in \mathrm{int}X, \tag{4.2.4}$$

$$\langle\nabla_1 B_\eta(x,y), x-y\rangle \geq \underline{\mu}\|x-y\|^2, \forall x,\ y \in \mathrm{int}X. \tag{4.2.5}$$

Proof. The first inequality in (4.2.4), originally presented in [52, Theorem 2.1.5], follows from the Lipschitz continuity of $\nabla\eta$. The second inequality in (4.2.4) and inequality (4.2.5) hold because of the strong convexity of η. ■

For the global convergence of the proposed BCPG methods, the variable kernels in this chapter are required to satisfy the following condition.

Definition 4.2.2. *Let $\{\eta^r\} \subseteq \Phi(X;\underline{\mu})$. The group of kernels $\{\eta^r\}$ is (ϱ, $\bar{\eta}$)-upper bounded on X if there exists a function $\bar{\eta}(x) \in \Phi(X;\underline{\mu})$ and $\varrho > 1$ such that the inequality*

$$B_{\eta^r}(x,y) \leq \varrho B_{\bar{\eta}}(x,y) \tag{4.2.6}$$

holds for any $x \in X$, $y \in \mathrm{int}X$ and $r > 0$.

Definition 4.2.2 is an extension of that in [19, Definition 5.1]. The condition (4.2.6) is weaker than gradient Lipschitz continuity assumption on $\{\eta^r\}$, that is, $\eta^r \in \Psi(X; \underline{\mu}, \overline{L})$ for all r. The following two examples demonstrate this assertion further.

Example 4.2.1. *Let $\{\eta^r\} \subseteq \Psi(X; \underline{\mu}, \overline{L})$. It can be easily verified that $\bar{\eta}(x) = \frac{\mu}{2}\|x\|^2 \in \Phi(X; \underline{\mu})$ and $B_{\bar{\eta}}(x, y) = \frac{\mu}{2}\|x - y\|^2$. Then $\{\eta^r\}$ is $(\frac{\overline{L}}{\underline{\mu}}, \bar{\eta})$-upper bounded on X from (4.2.4) in Lemma 4.2.2.*

Example 4.2.2. *Let $X = \{x \in \mathcal{R}^n \mid x_i \geq 0, \sum_{i=1}^{n} x_i = 1\}$, $\eta^1(x) = \sum_{i=1}^{n} x_i \ln x_i$, and $\eta^r(x) \in \Psi(X; 1, \overline{L})$, $r = 2, 3, \ldots$. It is shown in [6, Proposition 5.1] that $\eta^1(x) \in \Phi(X; 1)$, that is,*

$$B_{\eta^1}(x, y) \geq \frac{1}{2}\|x - y\|^2.$$

Moreover, it follows from (4.2.4) in Lemma 4.2.2 that for any $r \geq 2$,

$$B_{\eta^r}(x, y) \leq \frac{\overline{L}}{2}\|x - y\|^2 \leq \overline{L} B_{\eta^1}(x, y).$$

Hence, for any $r \geq 1$, we have that

$$B_{\eta^r}(x, y) \leq \max\{1, \overline{L}\} B_{\eta^1}(x, y).$$

Then the kernels $\{\eta^r\}$ is $(\max\{1, \overline{L}\}, \eta^1)$-upper bounded on X.

The following lemma shows some elementary inequalities on a convex function ψ.

Lemma 4.2.3. *Let $\psi : \mathcal{R}^n \to \mathcal{R}$ be convex. Then, the following inequalities hold for any x, $y \in \mathcal{R}^n$ and $t \in [0, 1]$.*

$$\psi(x + ty) - \psi(x) \leq t[\psi(x + y) - \psi(x)], \tag{4.2.7}$$

$$\psi(x + ty) - \psi(x + y) \leq (t - 1)[\psi(x + y) - \psi(x)]. \tag{4.2.8}$$

Proof. Since $x + ty = t(x + y) + (1 - t)x$, from the convexity of ψ, we have

$$\psi(x + ty) \leq t\psi(x + y) + (1 - t)\psi(x), \ \forall t \in [0, 1],$$

which yields the desired results. ■

4.3 Block coordinate proximal gradient (BCPG) methods

In this section, we first present the BCPG methods for solving problem (4.1.1). Then, we prove that the search direction d^r defined in (4.1.2) is a feasible descent direction of problem (4.1.1). Finally, we show the explicit rules to select the block J^r and the step size α^r.

4.3.1 The proposed BCPG methods

Block coordinate proximal gradient (BCPG) methods:

Step 0: Select an initial point $x^0 \in \mathrm{int}X$, and let $r = 0$.

Step 1: If some termination condition holds, then stop.

Step 2: Select a block $J^r \subseteq \{1, \dots, n\}$ by one of the Gauss-Seidel rules and select a strongly convex function $\eta^r : X \to \mathcal{R}$ as a kernel.

Step 3: Solve the following subproblem with the variable d to obtain a search direction d^r.

$$\begin{aligned} &\text{minimize} \ \ \langle \nabla f(x^r), d\rangle + B_{\eta^r}(x^r + d, x^r) + \tau\psi(x^r + d) \\ &\text{subject to} \ \ d_{\bar{J}^r} = 0. \end{aligned} \tag{4.3.1}$$

Step 4: Determine step size α^r by the Armijo rule.

Step 5: Set $x^{r+1}_{J^r} = x^r_{J^r} + \alpha^r d^r_{J^r}$, $x^{r+1}_{\bar{J}^r} = x^r_{\bar{J}^r}$, and $r = r + 1$. Go to Step 1.

Note that we can adopt a different kernel function at each iteration. In the remainder of this chapter, for the global convergence, we assume that there exists a constant $\underline{\mu} > 0$ such that $\eta^r \in \Phi(X; \underline{\mu})$ for all r. The Gauss-Seidel rules in Step 2 and the Armijo rule in Step 4 are presented in Section 2.4 and Subsection 4.3.3, respectively.

Remark 4.3.1. *Steps 3 and 4 of the above BCPG methods are different from those in [71]. The search direction in [71] conforms to the gradient projection method with the Armijo rule along the projection arc. Such an approach in [71] tends to obtain sparse (or active) solutions, whereas it requires solving subproblem (4.3.1) repeatedly. We can also construct the BCPG methods with the step size rule in [71], and show the same convergence properties.*

As mentioned in the introduction of this chapter, the BCPG methods include many well-known optimization methods. Among them, the following two methods are of particular interest in this chapter.

(i) A Block Coordinate Descent method

Suppose that function f is strongly convex with respect to each block. Let

$$\eta^r(x, x^r_{\bar{J}^r}) := f(x_{J^r}, x^r_{\bar{J}^r}) + \frac{1}{2}\|x_{\bar{J}^r} - x^r_{\bar{J}^r}\|^2. \tag{4.3.2}$$

Then, it can be verified that η^r is also strongly convex. Moreover, we have $\langle \nabla f(x^r), d\rangle + B_{\eta^r}(x^r + d, x^r) = f(x^r_{J^r} + d_{J^r}, x^r_{\bar{J}^r}) - f(x^r) + \langle \nabla_{\bar{J}^r} f(x^r_{J^r}, x^r_{\bar{J}^r}), d^r_{\bar{J}^r}\rangle + \frac{1}{2}\|d_{\bar{J}^r}\|^2$. Hence, subproblem (4.3.1) is equivalent to the problem

$$\underset{d_{J^r}}{\text{minimize}} \ \ f(x^r_{J^r} + d_{J^r}, x^r_{\bar{J}^r}) + \tau\psi_{J^r}(x^r_{J^r} + d_{J^r}). \tag{4.3.3}$$

Note that, in (4.3.3), $d_{J^r} \in \mathcal{R}^{|J^r|}$, where $|J^r|$ denotes the number of elements in J^r. Replacing subproblem (4.3.1) by (4.3.3), the BCPG methods are reduced to the classical block coordinate descent method [67].

(ii) Inexact BCPG methods

When subproblem (4.3.1) is difficult to solve exactly, we accept an approximate solution as a compromise. Next, we give a definition for the approximate solution of subproblem (4.3.1), under which the inexact BCPG methods are also regarded as certain "exact" BCPG methods.

Definition 4.3.1. *We say that d^r is an approximate solution of subproblem (4.3.1) with error ε^r if the pair $(d^r, \varepsilon^r) \in \mathcal{R}^n \times \mathcal{R}^n$ satisfies*

$$\begin{cases} \nabla_{J^r} f(x^r) + \nabla_{J^r}\eta(x^r + d^r) - \nabla_{J^r}\eta(x^r) + \varepsilon^r_{J^r} \in -\tau\partial_{J^r}\psi(x^r + d^r), \\ d^r_{\bar{J}^r} = 0, \varepsilon^r_{\bar{J}^r} = 0. \end{cases} \tag{4.3.4}$$

Note that if d^r satisfies (4.3.4) with $\varepsilon^r = 0$, then d^r is the exact solution of (4.3.1).

Now, suppose that $(d^r, \varepsilon^r) \in \mathcal{R}^n \times \mathcal{R}^n$ satisfies (4.3.4). Let $E^r \in \mathcal{R}^{n\times n}$ be the diagonal matrix, for which E^r_{ii} is given by

$$E^r_{ii} := \begin{cases} \frac{\varepsilon^r_i}{d^r_i} & \text{if } d^r_i \neq 0, \\ 0 & \text{if } d^r_i = 0. \end{cases} \tag{4.3.5}$$

Then, we have $\varepsilon^r = E^r d^r$. Combining it with (4.3.4), we have

$$\begin{cases} \nabla_{J^r} f(x^r) + \nabla_{J^r}\eta(x^r + d^r) - \nabla_{J^r}\eta(x^r) + (E^r d^r)_{J^r} \in -\tau\partial_{J^r}\psi(x^r + d^r), \\ d^r_{\bar{J}^r} = 0, \end{cases} \tag{4.3.6}$$

which are equivalent to the optimality conditions of subproblem (4.3.1) whose kernel $\eta^r(x)$ is replaced by $\eta^r(x) + x^T E^r x$. Hence, when ε^r is sufficiently small, the inexact version BCPG methods (inexactness is described as (4.3.4)) are reformulated as the proposed BCPG methods with the kerenl

$$\tilde{\eta}^r(x) = \eta^r(x) + x^T E^r x. \tag{4.3.7}$$

The conditions on ε^r for global convergence are given in Section 4.6.

4.3.2 A feasible descent property of d^r

In this subsection, we show the descent property of the search direction d^r given in Step 3, and present elementary results about the stationary points of subproblem (4.3.1) and the original problem (4.1.1). The next lemma states that the direction d^r is a feasible descent direction of F, which is a natural generalization of [69, Lemma 1].

Lemma 4.3.1. *For any $x^r \in X$ and $J^r \subseteq \mathcal{N}$, we have*

$$F(x^r + td^r) \le F(x^r) + t\Theta(x^r, d^r) + o(t), \ \forall t \in (0, 1], \tag{4.3.8}$$

where $\Theta(x^r, d^r) := \langle \nabla f(x^r), d^r\rangle + \tau\psi(x^r + d^r) - \tau\psi(x^r)$. Moreover, if $\eta^r \in \Phi(X; \underline{\mu})$, then we have

$$\Theta(x^r, d^r) \le -\langle \nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r\rangle \le 0. \tag{4.3.9}$$

In particular, $\Theta(x^r, d^r) < 0$ holds if $d^r \neq 0$.

Proof. Inequality (4.3.8) follows from [69, Lemma 1]. Next, we show inequalities (4.3.9). For any $t \in (0, 1)$, using (4.2.3) with $a = x^r + td^r$, $b = x^r$ and $c = x^r + d^r$, we have

$$B_{\eta^r}(x^r + td^r, x^r) - B_{\eta^r}(x^r + d^r, x^r) \le -(1-t)\langle \nabla_1 B_{\eta^r}(x^r + td^r, x^r), d^r\rangle. \tag{4.3.10}$$

Since d^r is a solution of subproblem (4.3.1), we obtain

$$\begin{aligned}&\langle \nabla f(x^r), d^r\rangle + B_{\eta^r}(x^r + d^r, x^r) + \tau\psi(x^r + d^r)\\ &\qquad \le\ t\langle \nabla f(x^r), d^r\rangle + B_{\eta^r}(x^r + td^r, x^r) + \tau\psi(x^r + td^r).\end{aligned}$$

Then, it follows from (4.2.8) and (4.3.10) that

$$(1-t)\langle \nabla f(x^r), d^r\rangle + (1-t)\left(\tau\psi(x^r + d^r) - \tau\psi(x^r)\right) \le -(1-t)\langle \nabla_1 B_{\eta^r}(x^r + td^r, x^r), d^r\rangle.$$

Since $1 - t > 0$ for any $t \in (0, 1)$, dividing the above inequality by $1 - t$, we have

$$\langle \nabla f(x^r), d^r\rangle + \tau(\psi(x^r + d^r) - \psi(x^r)) \le -\langle \nabla_1 B_{\eta^r}(x^r + td^r, x^r), d^r\rangle.$$

We obtain the first inequality of (4.3.9) by letting $t \to 1$ in the above inequality. The second inequality of (4.3.9) can be proved easily from Lemma 4.2.2. Further, if $d^r \neq 0$, then it follows from (4.2.5) that $\langle \nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r\rangle \ge \underline{\mu}\|d^r\|^2 > 0$, which yields $\Theta(x^r, d^r) < 0$ from (4.3.9). ■

Next, we present the definition of the stationary point and its sufficient and necessary conditions.

Definition 4.3.2. *We say that $x^* \in \operatorname{dom} F$ is a stationary point of F with respect to the block J if $F'(x^*; d) \ge 0$ holds for all $d \in \mathcal{R}^n$ with $d_{\bar{J}} = 0$, where $F'(x^*; d)$ denotes the directional derivatives of F at the vector x^* with respect to the direction d.*

Lemma 4.3.2. *For any $\eta \in \Phi(X; \underline{\mu})$, $d_\eta(x^*, J) = 0$ holds if and only if the vector $x^* \in \operatorname{dom} F$ is a stationary point of F with respect to the block J.*

Proof. First, we prove the "if " part by contradiction. Suppose that $d_\eta(x^*, J) \neq 0$ holds. Then it follows from Lemma 4.3.1 that

$$\begin{aligned} F'(x^*, d_\eta(x^*, J)) &= \lim_{t\downarrow 0} \frac{F(x^* + td_\eta(x^*, J)) - F(x^*)}{t} \\ &\leq \lim_{t\downarrow 0} \frac{t\Theta(x^*, d_\eta(x^*, J)) + o(t)}{t} \\ &= \Theta(x^*, d_\eta(x^*, J)) \\ &< 0, \end{aligned}$$

where the first inequality follows from (4.3.8) and the second inequality follows from (4.3.9) and the assumption $d_\eta(x^*, J) \neq 0$. However, it contradicts with Definition 4.3.2.

Conversely, if $d_\eta(x^*, J) = 0$ holds, for all $t > 0$, $y \in \mathcal{R}^n$ with $y_{\bar{J}} = 0$, we have from (4.1.2)

$$t\langle \nabla f(x^*), y\rangle + B_\eta(x^* + ty, x^*) + \tau\psi(x^* + ty) \geq \tau\psi(x^*). \tag{4.3.11}$$

Then, for any $y \in \mathcal{R}^n$ such that $y_{\bar{J}} = 0$, we have

$$\begin{aligned} F'(x^*, y) &= \lim_{t\downarrow 0} \frac{f(x^* + ty) - f(x^*) + \tau\psi(x^* + ty) - \tau\psi(x^*)}{t} \\ &\geq \lim_{t\downarrow 0} \frac{f(x^* + ty) - f(x^*) - t\langle \nabla f(x^*), y\rangle - B_\eta(x^* + ty, x^*)}{t} \\ &= \lim_{t\downarrow 0} \frac{-\eta(x^* + ty) + \eta(x^*) + t\langle \nabla \eta(x^*), y\rangle}{t} \\ &= 0, \end{aligned}$$

where the first inequality follows from (4.3.11), the second equality follows from the definition of B_η and the differentiability of f, and the last equality follows from the differentiability of η. Therefore, x^* is a stationary point of F with respect to the block J from Definition 4.3.2. ■

The following corollary follows immediately from Lemma 4.3.2 by setting $J = \mathcal{N}$.

Corollary 4.3.1. *For any $\eta \in \Phi(X; \underline{\mu})$, $d_\eta(x^*, \mathcal{N}) = 0$ holds if and only if x^* is a stationary point of problem (4.1.1).*

4.3.3 Gauss-Seidel rules and Armijo rule

For establishing the global convergence, we assume that the rules to select a block in Step 2 is the generalized Gauss-Seidel rule. To show the linear convergence rate, we employ the restricted Gauss-Seidel rule [69]. For the formal definitions, see Section 2.4 for details.

To find the step size α^r in Step 4, we adopt the following generalized Armijo rule [69].

Armijo rule:

Select any scalar $\alpha^r_{init} > 0$ with $\sup_r \alpha^r_{init} < \infty$, and let α^r be the largest element of the sequence $\{\alpha^r_{init}\beta^j\}_{j=0,1,\ldots}$ such that

$$F(x^r + \alpha^r d^r) \le F(x^r) + \alpha^r \sigma \Delta(x^r + d^r), \tag{4.3.12}$$

where $\beta,\ \sigma \in (0,1)$, $\gamma \in [0,1)$, and

$$\Delta(x^r + d^r) := \langle \nabla f(x^r), d^r\rangle + \gamma\langle \nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r\rangle + \tau\psi(x^r + d^r) - \tau\psi(x^r). \tag{4.3.13}$$

We make a few remarks on this Armijo rule. The assumption $\sup_r \alpha^r_{init} < \infty$ implies that both $\{\alpha^r_{init}\}$ and $\{\alpha^r\}$ are bounded. However, it is still possible that $\{\alpha^r\} \to 0$ as $r \to \infty$. If the smooth part f of problem (4.1.1) is gradient Lipschitz continuous, then we can ensure that $\inf_r \alpha^r > 0$. This assertion will be shown by Lemma 4.5.2 in Section 4.5.

By the definition of $\Theta(x^r, d^r)$ in Lemma 4.3.1, $\Delta(x^r + d^r)$ in (4.3.13) can be rewritten as

$$\Delta(x^r + d^r) = \Theta(x^r, d^r) + \gamma\langle \nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r\rangle. \tag{4.3.14}$$

The following lemma states that the term $\Delta(x^r + d^r)$ is nonpositive, which is important for the validity of the above Armijo rule. The proof can be easily deduced by using (4.2.5), (4.3.9), and (4.3.14).

Lemma 4.3.3. *For any $\eta^r \in \Phi(X; \underline{\mu})$, $\gamma \in [0,1)$, and $d^r \in \mathcal{R}^n$, we have*

$$\Delta(x^r + d^r) \le (\gamma - 1)\langle \nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r\rangle \le (\gamma - 1)\underline{\mu}\|d^r\|^2 \le 0. \tag{4.3.15}$$

Furthermore, $\Delta(x^r + d^r) < 0$ holds if $d^r \neq 0$.

The following two remarks further illustrate the role of Lemma 4.3.3.

Remark 4.3.2. *Lemma 4.3.3 implies that α^r in the Armijo rule is well defined. It is illustrated by Lemma 4.3.1 that the nonzero d^r is a descent direction of F. Thus, there exists a constant $t > 0$ such that $F(x^r + td^r) \le F(x^r) + t\sigma\Delta(x^r + d^r)$ since $\sigma \in (0,1)$ and $\Theta(x^r, d^r) \le \Delta(x^r + d^r) < 0$ hold for any nonzero d^r. Hence, the existence of α^r satisfying (4.3.13) is confirmed.*

Remark 4.3.3. *The sequence $\{F(x^r)\}$ generated by the Armijo rule is not increasing since $\Delta(x^r + d^r) < 0$ holds for any nonzero d^r. Therefore, we have either $\{F(x^r)\} \downarrow -\infty$ or $\{F(x^r)\} > -\infty$. If $\{F(x^r)\} > -\infty$, then the limit of $\{F(x^r)\}$ exists.*

4.4 Global convergence

In this section, we show the global convergence of the BCPG methods. Since the proof is an extension of that for the BCGD method in [69], we refer some lemmas from [69] and omit the corresponding proofs. Throughout this section, $\{x^r\}$, $\{\alpha^r\}$, and $\{d^r\}$ denote the sequences generated by the BCPG methods.

The following lemma, corresponding to Theorem 1(b) in [69], shows that the search direction d^r vanishes when $\{x^r\}$ is bounded. It can be proved by replacing ${d^k}^T H^k d^k$ by $\langle \nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r \rangle$ in the proof of [69, Theorem 1(b)].

Lemma 4.4.1. *If there exists an infinite set $\mathcal{X} \subseteq \{0, 1, \dots\}$ and a vector $\bar{x}$ such that $\lim_{r\to\infty,\, r\in\mathcal{X}} x^r = \bar{x} \in X$, then the following statements hold.*

(i) $\lim_{r\to\infty} \alpha^r \Delta(x^r + d^r) = 0.$

(ii) $\lim_{r\to\infty,\, r\in\mathcal{X}} d^r = 0.$

Next, we prove the global convergence of $\{x^r\}$.

Theorem 4.4.1. *Suppose that ψ is block separable with respect to each block J^r, and that kernels $\{\eta^r\}$ are $(\varrho,\, \bar{\eta})$-upper bounded on X with $\varrho > 1$ and $\bar{\eta} \in \Phi(X; \underline{\mu})$. Let $\{x^r\}$ be generated by the BCPG methods. Then, any accumulation point of $\{x^r\}$ is a stationary point of problem (4.1.1).*

Proof. Let $\bar{x}$ be an accumulation point of $\{x^r\}$. Then, there exists an infinite set $\mathcal{Z} \subseteq \{0, 1, \dots\}$ such that $\lim_{r\to\infty,\, r\in\mathcal{Z}} x^r = \bar{x}$.

To prove this theorem, first, we show that there exists a block denoted by Q^0 such that $\bar{x}$ is a stationary point of F with respect to the block Q^0, i.e.,

$$F'(\bar{x}, d) \geq 0, \forall d \in \mathcal{R}^n \text{ with } d_{\overline{Q^0}} = 0. \tag{4.4.1}$$

From Lemma 4.3.2, it is equivalent to show that $d_{\widehat{\eta}}(\bar{x}; Q^0) = 0$ for certain $\widehat{\eta} \in \Phi(X; \underline{\mu})$.

In fact, at each iteration, there are limited alternatives (no more than $2^n - 1$) for selecting a block. Hence, we can suppose that there exists a block Q^0 and an infinite subset $\mathcal{Z}_0 \subseteq \mathcal{Z}$ such that $J^r = Q^0$ for all $r \in \mathcal{Z}_0$. Further, since $\lim_{r\to\infty,\, r\in\mathcal{Z}} x^r = \bar{x}$ and $\mathcal{Z}_0 \subseteq \mathcal{Z}$, we have

$$\lim_{r\to\infty,\, r\in\mathcal{Z}_0} x^r = \bar{x}. \tag{4.4.2}$$

Then, for any $r \in \mathcal{Z}_0$, d^r, and $x \in \mathcal{R}^n$ such that $x_{\overline{Q^0}} = x^r_{\overline{Q^0}}$, we have

$$\begin{aligned}
& \langle \nabla f(x^r), d^r \rangle + \frac{\mu}{2}\|d^r\|^2 + \tau\psi(x^r + d^r) \\
\leq\ & \langle \nabla f(x^r), d^r \rangle + B_{\eta^r}(x^r + d^r, x^r) + \tau\psi(x^r + d^r) \\
\leq\ & \langle \nabla f(x^r), x - x^r \rangle + B_{\eta^r}(x, x^r) + \tau\psi(x) \\
\leq\ & \langle \nabla f(x^r), x - x^r \rangle + \varrho B_{\bar{\eta}}(x, x^r) + \tau\psi(x),
\end{aligned} \tag{4.4.3}$$

where the first inequality follows from Lemma 4.2.2, the second inequality holds since d^r is a solution of subproblem (4.3.1), and the last inequality follows from Definition 4.2.2.

From (4.4.2) and Lemma 4.4.1(ii), we have $\lim\limits_{r\to\infty,\, r\in\mathcal{Z}_0} d^r = 0$. Further, letting $r \to \infty$ with $r \in \mathcal{Z}_0$ in (4.4.3), we have

$$\tau\psi(\bar{x}) \leq \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \varrho B_{\bar{\eta}}(x, x^r) + \tau\psi(x),\ \forall x \in \mathcal{R}^n \text{ with } x_{\overline{Q^0}} = \bar{x}_{\overline{Q^0}},$$

which implies that $d_{\widehat{\eta}}(\bar{x}; Q^0) = 0$, where $\widehat{\eta}(x) = \varrho\bar{\eta}(x)$. It then follows from $\varrho > 1$ that $\widehat{\eta}(x) \in \Phi(X; \varrho\underline{\mu}) \subseteq \Phi(X; \underline{\mu})$.

In the second step, we show that there exist other $B-1$ blocks Q^i, $i = 1, \ldots, B-1$, such that $F'(\bar{x}, d) \geq 0$ holds for any $d \in \mathcal{R}^n$ with $d_{\overline{Q^i}} = 0$ and $\sum_{i=0}^{B-1} Q^i = \mathcal{N}$.

Since α^r is bounded, we have from (4.4.2) and Lemma 4.4.1 that

$$\lim_{r\to\infty,\, r\in\mathcal{Z}_0} x^{r+1} = \lim_{r\to\infty,\, r\in\mathcal{Z}_0} \{x^r + \alpha^r d^r\} = \bar{x}.$$

By the same argument as in the first step, there exists a block Q^1 and an infinite subset $\mathcal{Z}_1 \subseteq \mathcal{Z}_0$ such that $J^{r+1} = Q^1$ for all $r \in \mathcal{Z}_1$. Further, we obtain $d_{\widehat{\eta}}(\bar{x}; Q^1) = 0$ with $\widehat{\eta}(x) = \varrho\bar{\eta}(x) \in \Phi(X; \underline{\mu})$, i.e., $\bar{x}$ is a stationary point of F with respect to block Q^1. From Definition 4.3.2, we have

$$F'(\bar{x}, d) \geq 0, \forall d \in \mathcal{R}^n \text{ with } d_{\overline{Q^1}} = 0. \tag{4.4.4}$$

Similarly, there exist $B-2$ blocks Q^i, $i = 2, \ldots, B-1$ such that

$$F'(\bar{x}, d) \geq 0, \forall d \in \mathcal{R}^n \text{ with } d_{\overline{Q^i}} = 0. \tag{4.4.5}$$

From the generalized Gauss-Seidel rule, we have $\mathcal{N} = \bigcup_{i=0}^{B-1} J^{r+i} = \bigcup_{i=0}^{B-1} Q^i$.

Finally, we show that $\bar{x}$ is a stationary point of problem (4.1.1). From Corollary 4.3.1, we only need to show $F'(\bar{x}, d) \geq 0$ for all $d \in \mathcal{R}^n$. In fact, we can obtain B disjoint blocks $\tilde{J}^i \subseteq Q^i$, $i = 0, 1, \ldots, B-1$, such that $\bigcup_{i=1}^{B-1} \tilde{J}^i = \mathcal{N}$ and $\tilde{J}^i \cap \tilde{J}^j = \emptyset$ for any i, j such that $i \neq j$. Hence, we have $\langle \nabla f(\bar{x}), d\rangle = \sum_{i=0}^{B-1} \langle \nabla_{\tilde{J}^i} f(\bar{x}), d_{\tilde{J}^i}\rangle$ and $\psi(\bar{x} + td) = \sum_{i=0}^{B-1} \psi_{\tilde{J}^i}(\bar{x}_{\tilde{J}^i} + td_{\tilde{J}^i})$ for any $d \in \mathcal{R}^n$. For convenience, we denote $d^o_{\tilde{J}^i} = (0, \ldots, 0, d^T_{\tilde{J}^i}, 0, \ldots, 0)^T \in \mathcal{R}^n$. Note that

$$F'(\bar{x}, d^o_{\tilde{J}^i}) \geq 0 \tag{4.4.6}$$

holds for all i from (4.4.1), (4.4.4), and (4.4.5).

Then, we obtain

$$\begin{aligned} F'(\bar{x}, d) &= \lim_{t\downarrow 0} \frac{f(\bar{x} + td) - f(\bar{x}) + \tau\psi(\bar{x} + td) - \tau\psi(\bar{x})}{t} \\ &= \sum_{i=0}^{B-1} \langle \nabla_{\tilde{J}^i} f(\bar{x}), d_{\tilde{J}^i}\rangle + \tau \lim_{t\downarrow 0} \frac{\sum\limits_{i=0}^{B-1} \psi_{\tilde{J}^i}(\bar{x}_{\tilde{J}^i} + td_{\tilde{J}^i}) - \psi_{\tilde{J}^i}(\bar{x}_{\tilde{J}^i})}{t} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=0}^{B-1} F'(\bar{x}, d^{o}_{\tilde{J}i}) \\
&\geq 0,
\end{aligned}
$$

where the second equality follows from the differentiability of f and the block separability of ψ, the third equality follows from the definition of directional derivative, and the inequality follows from (4.4.6). Hence, the proof is completed. ■

4.5 Linear convergence rate

In this section, we establish the linear convergence rate for the BCPG methods with the restricted Gauss-Seidel rule. Throughout this section, $\{x^r\}$, $\{\alpha^r\}$, and $\{d^r\}$ are sequences generated by the BCPG methods with the restricted Gauss-Seidel rule.

We make some further assumptions on the objective function F and the smooth part f for linear convergence. For convenience, we denote the set of stationary points of problem (4.1.1) by $\bar{X}$ in the rest of this chapter.

Assumption 4.5.1. *The gradient ∇f is L_f-Lipschitz continuous, i.e., $\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\|$ holds for any $x, y \in \mathcal{R}^n$.*

Assumption 4.5.2. *There exists a scalar $\delta > 0$ such that $\|x - y\| \geq \delta$, whenever $x, y \in \bar{X}$ and $F(x) \neq F(y)$.*

Assumption 4.5.2 means that the stationary point set $\bar{X}$ has a separable property. It holds when $\bar{X}$ contains only a finite number of values, or the components of $\bar{X}$ are properly separable. Note that, if the objective function F is convex, then Assumption 4.5.2 automatically holds.

Assumption 4.5.3. *The set $\bar{X}$ is nonempty. Moreover, for any $\zeta \in \mathcal{R}$, there exist scalars $\kappa > 0$ and $\epsilon > 0$ such that $\operatorname{dist}(x, \bar{X}) \leq \kappa \|d_{\bar{\eta}}(x; \mathcal{N})\|$, whenever $F(x) \leq \zeta$ and $\|d_{\bar{\eta}}(x; \mathcal{N})\| \leq \epsilon$. Here, $\bar{\eta}(x) := \frac{1}{2}\underline{\mu} x^T x$ and $\operatorname{dist}(x, \bar{X}) = \min\limits_{\bar{x} \in \bar{X}} \|x - \bar{x}\|$.*

Assumption 4.5.3 is called the "local Lipschitz error bound assumption" in [69]. Note that it holds whenever one of the following conditions holds (See [66, 69] for details).

(C1) $f = \frac{1}{2}x^T E x + \langle q, x\rangle$ for all $x \in \mathcal{R}^n$, where $E \in \mathcal{R}^{n \times n}$, $q \in \mathcal{R}^n$, and ψ is polyhedral.

(C2) $f(x) = g(Ex) + \langle q, x\rangle$ for all $x \in \mathcal{R}^n$, where $E \in \mathcal{R}^{m \times n}$, $q \in \mathcal{R}^n$, and g is a strongly convex differentiable function on $\mathcal{R}^m$ with ∇g Lipschitz continuous on $\mathcal{R}^m$. ψ is polyhedral.

(C3) $f(x) = \max_{y \in Y}\{\langle Ex, y\rangle - g(y)\} + \langle q, x\rangle$ for all $x \in \mathcal{R}^n$, where Y is a polyhedral set in $\mathcal{R}^m$, $E \in \mathcal{R}^{m\times n}$, $q \in \mathcal{R}^n$, and g is a strongly convex differentiable function on $\mathcal{R}^m$ with ∇g Lipschitz continuous on $\mathcal{R}^m$. ψ is polyhedral.

(C4) f is strongly convex and ∇f is Lipschitz continuous.

(C5) $f(x) = g(Ex)$ for all $x \in \mathcal{R}^n$, where $E \in \mathcal{R}^{m\times n}$, g is a strongly convex differentiable function on $\mathcal{R}^m$ and ∇g Lipschitz continuous on $\mathcal{R}^m$. $\psi := \sum_{i=1}^{N} \omega_i \|x_{\mathcal{J}^i}\|_2$.

Moreover, it can be easily verified that the functions in (C1)-(C5) satisfy Assumption 4.5.1. When matrix E in (C1) is symmetric and positive semidefinite, the functions in (C1)-(C5) satisfy Assumption 4.5.2 as well.

Before presenting the main theorem of this section, we introduce some technical lemmas.

Under Assumption 4.5.1, we have the following two lemmas, which are originally given in [69, Lemma 5(b)] and [69, Theorem 1(f)]. We omit their proofs. Lemma 4.5.1 gives an estimation for the step size α^r generated by the Armijo rule, and Lemma 4.5.2 shows that the direction d^r is globally convergent to 0, which is sharper than Lemma 4.4.1(ii).

Lemma 4.5.1. *Suppose that Assumption 4.5.1 holds. For any* $\eta^r \in \Phi(X; \underline{\mu})$, $\sigma \in (0,1)$, *and* $\alpha \in [0, \min\{1,\ 2\underline{\mu}(1-\sigma+\sigma\gamma)/L_f\}]$, *we have*

$$F(x^r + \alpha d^r) - F(x^r) \le \sigma\alpha\Delta(x^r + d^r). \qquad \square$$

Note that Step 4 can adopt $\alpha^r_{init} = 1$ and $\alpha = \min\{1,\ 2\underline{\mu}(1-\sigma+\sigma\gamma)/L_f\}$ without evaluating F when L_f is known previously. It is useful when the evaluation of F is computationally expensive. In the rest of this chapter, for simplicity, we assume that $\alpha^r_{init} = 1$.

Lemma 4.5.2. *Suppose that Assumption 4.5.1 holds. Then,* $\inf_r \alpha^r > 0$. *Further,* $\lim_{r\to\infty} d^r = 0$ *holds if* $\lim_{r\to\infty} F(x^r) > -\infty$.

From the above lemmas, we find that step size α^r is away from zero.

The next lemma shows a relation between the distinct directions generated with respect to different kernels.

Lemma 4.5.3. *For any* $\eta^r \in \Psi(X; \underline{\mu}, \overline{L})$, *let* $\bar{\eta}(x) = \frac{1}{2}\underline{\mu}x^T x$, $d^r := d^r_{\eta^r}(x^r; J^r)$, *and* $\bar{d}^r := d^r_{\bar{\eta}}(x^r; J^r)$. *Then, we have*

$$\|\bar{d}^r\| \le w_1 \|d^r\|, \tag{4.5.1}$$

where $w_1 = \frac{\sqrt{(\overline{L}+\underline{\mu})^2 - 4\underline{\mu}^2}}{2\underline{\mu}} + \frac{\overline{L}+\underline{\mu}}{2\underline{\mu}} > 0.$

Proof. For simplicity, we only show (4.5.1) with $J^r = \mathcal{N}$. Using Fermat's rule, we obtain

$$d^r \in \operatorname*{argmin}_u \{\langle \nabla f(x^r) + \nabla_1 B_{\eta^r}(x^r + d^r, x^r), u\rangle + \tau\psi(x^r + u)\}, \tag{4.5.2}$$

$$\bar{d}^r \in \operatorname*{argmin}_u \left\{\langle \nabla f(x^r) + \nabla_1 B_{\bar{\eta}}(x^r + \bar{d}^r, x^r), u\rangle + \tau\psi(x^r + u)\right\}. \tag{4.5.3}$$

These two relations imply that

$$\begin{aligned}\langle \nabla f(x^r) + &\nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r\rangle + \tau\psi(x^r + d^r)\\ &\leq \langle \nabla f(x^r) + \nabla_1 B_{\eta^r}(x^r + d^r, x^r), \bar{d}^r\rangle + \tau\psi(x^r + \bar{d}^r),\\ \langle \nabla f(x^r) + &\nabla_1 B_{\bar{\eta}}(x^r + \bar{d}^r, x^r), \bar{d}^r\rangle + \tau\psi(x^r + \bar{d}^r)\\ &\leq \langle \nabla f(x^r) + \nabla_1 B_{\bar{\eta}}(x^r + \bar{d}^r, x^r), d^r\rangle + \tau\psi(x^r + d^r).\end{aligned}$$

Summing these two inequalities and rearranging it, we have

$$\begin{aligned}&\langle \nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r\rangle + \langle \nabla_1 B_{\bar{\eta}}(x^r + \bar{d}^r, x^r), \bar{d}^r\rangle\\ &- \langle \nabla_1 B_{\eta^r}(x^r + d^r, x^r), \bar{d}^r\rangle - \langle \nabla_1 B_{\bar{\eta}}(x^r + \bar{d}^r, x^r), d^r\rangle \leq 0.\end{aligned} \tag{4.5.4}$$

From Lemma 4.2.2, we have that

$$\langle \nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r\rangle \geq \underline{\mu}\|d^r\|^2, \quad \langle \nabla_1 B_{\bar{\eta}}(x^r + \bar{d}^r, x^r), \bar{d}^r\rangle \geq \underline{\mu}\|\bar{d}^r\|^2.$$

Since $\nabla\eta^r$ and $\nabla\bar{\eta}$ are Lipschitz continuous, from the Hölder inequality we have

$$\langle \nabla_1 B_{\eta^r}(x^r + d^r, x^r), \bar{d}^r\rangle \leq \overline{L}\|d^r\|\|\bar{d}^r\|, \quad \langle \nabla_1 B_{\bar{\eta}}(x^r + \bar{d}^r, x^r), d^r\rangle \leq \underline{\mu}\|d^r\|\|\bar{d}^r\|.$$

Combining them with (4.5.4), we have

$$\underline{\mu}\|d^r\|^2 - (\overline{L} + \underline{\mu})\|d^r\|\|\bar{d}^r\| + \underline{\mu}\|\bar{d}^r\|^2 \leq 0.$$

Dividing both sides by $\underline{\mu}$ and completing the square for the first two terms, we have

$$\left(\|\bar{d}^r\| - \frac{\overline{L} + \underline{\mu}}{2\underline{\mu}}\|d^r\|\right)^2 \leq \left[\frac{(\overline{L} + \underline{\mu})^2 - 4\underline{\mu}^2}{4\underline{\mu}^2}\right]\|d^r\|^2.$$

Since $(\overline{L} + \underline{\mu})^2 \geq 4\overline{L}\underline{\mu} \geq 4\underline{\mu}^2$, we have

$$\|\bar{d}^r\| \leq \left(\frac{\sqrt{(\overline{L} + \underline{\mu})^2 - 4\underline{\mu}^2}}{2\underline{\mu}} + \frac{\overline{L} + \underline{\mu}}{2\underline{\mu}}\right)\|d^r\|.$$

Let $w_1 = \frac{\sqrt{(\overline{L}+\underline{\mu})^2 - 4\underline{\mu}^2}}{2\underline{\mu}} + \frac{\overline{L}+\underline{\mu}}{2\underline{\mu}}$. Then, we obtain the desired inequality. ∎

The next lemma illustrates the Lipschitz continuity of the search direction d with respect to ∇f.

Lemma 4.5.4. *Suppose that φ is block-separable with respect to the block J^r. For any $x^r \in \mathrm{int}X$, and $a \in \mathcal{R}^n$, we define*

$$d^r(a) = \underset{d}{\mathrm{argmin}} \left\{ \langle a, d\rangle + \frac{1}{2}\underline{\mu} d^T d + \tau\psi(x^r + d) \mid d_{\bar{J}^r} = 0 \right\}.$$

Then, we have

$$\|d^r(a) - d^r(b)\| \leq \frac{1}{\underline{\mu}}\|a - b\|, \ \textit{for all} \ a, b \in \mathcal{R}^n. \tag{4.5.5}$$

Proof. This lemma follows immediately from [69, Lemma 4] with $h(u) = \frac{1}{2}\underline{\mu}u^T u$, $p = 2$, and $\rho = \underline{\mu}$. ■

The following technical lemma is an extension of [69, Lemma 5(a)], which will be used for Lemma 4.5.8.

Lemma 4.5.5. *Suppose that φ is block-separable with respect to the block J^r. For any $\eta^r \in \Phi(X; \underline{\mu})$, $x^r \in \mathrm{int}X$, $\gamma \in [0, 1)$, and $t \in (0, 1]$, we have*

$$\begin{aligned}\langle \nabla f(x^r) + &\nabla_1 B_{\eta^r}(x^r + d^r, x^r), \hat{x}^r(t) - \bar{x}^r\rangle + \tau\psi(\hat{x}^r(t)) - \tau\psi(\bar{x}^r) \\ &\leq \ (t - 1)\left[(1 - \gamma)\langle \nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r\rangle + \Delta(x^r + d^r)\right],\end{aligned}$$

where $\hat{x}^r(t) := x^r + td^r$, $d^r_{\bar{J}^r} = 0$, and $\bar{x}^r \in \mathcal{R}^n$ with $\bar{x}^r_{\bar{J}^r} = x^r_{\bar{J}^r}$.

Proof. Since d^r is the solution of (4.1.2), by Fermat's rule, we get that

$$d^r \in \underset{d \in \mathcal{R}^n}{\mathrm{argmin}} \left\{ \langle \nabla f(x^r) + \nabla_1 B_{\eta^r}(x^r + d^r, x^r), d\rangle + \tau\psi(x^r + d) | \ d_{\bar{J}^r} = 0\right\},$$

which implies that

$$\begin{aligned}\langle \nabla f(x^r) + &\nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r\rangle + \tau\psi(x^r + d^r) \\ &\leq \langle \nabla f(x^r) + \nabla_1 B_{\eta^r}(x^r + d^r, x^r), \bar{x}^r - x^r\rangle + \tau\psi(\bar{x}^r)\end{aligned}$$

holds for all $\bar{x}^r \in \mathcal{R}^n$ with $\bar{x}^r_{\bar{J}^r} = x^r_{\bar{J}^r}$, i.e.,

$$\langle \nabla f(x^r) + \nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r - \bar{x}^r + x^r\rangle \leq \tau\psi(\bar{x}^r) - \tau\psi(x^r + d^r). \tag{4.5.6}$$

Hence, we obtain

$$\begin{aligned}&\langle \nabla f(x^r) + \nabla_1 B_{\eta^r}(x^r + d^r, x^r), \hat{x}^r(t) - \bar{x}^r\rangle + \tau\psi(\hat{x}^r(t)) - \tau\psi(\bar{x}^r) \\ &= \ \langle \nabla f(x^r) + \nabla_1 B_{\eta^r}(x^r + d^r, x^r), x^r + td^r - \bar{x}^r\rangle + \tau\psi(x^r + td^r) - \tau\psi(\bar{x}^r) \\ &= \ \langle \nabla f(x^r) + \nabla_1 B_{\eta^r}(x^r + d^r, x^r), x^r + d^r - \bar{x}^r\rangle + \tau\psi(x^r + td^r) - \tau\psi(\bar{x}^r) \\ &\quad + (t - 1)\langle \nabla f(x^r) + \nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r\rangle\end{aligned}$$

$$
\begin{aligned}
&\le \ \tau\psi(\bar{x}^r) - \tau\psi(x^r + d^r) + (t-1)\langle \nabla f(x^r) + \nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r\rangle + \tau\psi(x^r + td^r) - \tau\psi(\bar{x}^r) \\
&\le \ (t-1)\left[\tau\psi(x^r + d^r) - \tau\psi(x^r) + \langle \nabla f(x^r) + \nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r\rangle\right] \\
&= \ (t-1)\left[(1-\gamma)\langle \nabla_1 B_{\eta^r}(x^r + d^r, x^r), d^r\rangle + \Delta(x^r + d^r)\right],
\end{aligned}
$$

where the first inequality follows from (4.5.6), the second inequality follows from (4.2.8), and the last inequality follows from (4.3.13). ■

The next lemma presents a relation of the directions generated with the blocks $\mathcal{N}$ and J. Recall that $\varphi(r)$ and Γ are defined in the restricted Gauss-Seidel rule.

Lemma 4.5.6. *Suppose that Assumption 4.5.1 holds. Let $\bar{\eta}(x) = \frac{1}{2}\underline{\mu}x^T x$. Then, the following statements hold for any $\eta^r \in \Psi(X; \underline{\mu}, \overline{L})$.*

(a) $\|d_{\bar{\eta}}(x^r; \mathcal{N})\| \le w_2 \displaystyle\sum_{i=r}^{\varphi(r)-1} \|d_{\eta^i}(x^i; J^i)\|$ *for all $r \in \Gamma$, where* $w_2 = \frac{\sqrt{(\overline{L}+\underline{\mu})^2 - 4\underline{\mu}^2}}{2\underline{\mu}} + \frac{\overline{L}+\underline{\mu}}{2\underline{\mu}} + \frac{L_f}{\underline{\mu}} \sup_r \alpha^r > 0.$

(b) *If* $\displaystyle\lim_{r\to\infty} F(x^r) > -\infty$, *then* $\displaystyle\lim_{r\to\infty, r\in\Gamma} \sum_{i=r}^{\varphi(r)-1} \|d_{\eta^i}(x^i; J^i)\| = 0.$

(c) *If* $\displaystyle\lim_{r\to\infty} F(x^r) > -\infty$, *then* $\displaystyle\lim_{r\to\infty, r\in\Gamma} d_{\bar{\eta}}(x^r; \mathcal{N}) = 0.$

Proof. (a) By using Lemmas 4.5.1, 4.5.3, and 4.5.4, we can show this inequality in a similar manner to [69, Theorem 2(a)]. Here, we omit the details.

(b) Under Assumption 4.5.1 and the assumption $\displaystyle\lim_{r\to\infty} F(x^r) > -\infty$, we have from Lemma 4.5.2 that $\displaystyle\lim_{r\to\infty} d_{\eta^r}(x^r; J^r) = 0$. Moreover, we obtain $\displaystyle\lim_{r\to\infty, r\in\Gamma} \sum_{i=r}^{\varphi(r)-1} \|d_{\eta^i}(x^i; J^i)\| = 0$ since $\varphi(r) - r \le n$ for all $r \in \Gamma$.

(c) Combining (b) with (a) in this lemma, we have $\displaystyle\lim_{r\to\infty,\ r\in\Gamma} d_{\bar{\eta}}(x^r; \mathcal{N}) = 0.$ ■

Suppose that $\displaystyle\lim_{r\to\infty} F(x^r) > -\infty$. From Lemma 4.5.6 (b), we have $\displaystyle\lim_{r\to\infty, r\in\Gamma} \sum_{i=r}^{\varphi(r)-1} \|d_{\eta^i}(x^i; J^i)\| = 0$, which yields

$$
\lim_{r\to\infty,\ r\in\Gamma} \{x^{\varphi(r)} - x^r\} = 0, \tag{4.5.7}
$$

since $x^{\varphi(r)} - x^r = \sum_{i=r}^{\varphi(r)-1} \alpha^r d^r$ and $0 \le \alpha^r \le \sup_r \alpha^r_{init} < \infty$.

Further, from Assumptions 4.5.1, 4.5.3, and Lemma 4.5.6 (c), we obtain $\displaystyle\lim_{r\to\infty,\ r\in\Gamma} \operatorname{dist}(x^r, \bar{X}) = 0$, i.e.,

$$
\lim_{r\to\infty,\ r\in\Gamma} \|x^r - \bar{x}^r\| = 0, \tag{4.5.8}
$$

where $\bar{x}^r$ denotes the stationary point nearest to x^r. Then, it yields $\lim\limits_{r\to\infty,\, r\in\Gamma}\{x^{\varphi(r)} - x^r - \bar{x}^{\varphi(r)} + \bar{x}^r\} = 0$. Hence, we have from (4.5.7) that

$$\lim_{r\to\infty,\, r\in\Gamma}\{\bar{x}^{\varphi(r)} - \bar{x}^r\} = 0. \tag{4.5.9}$$

Then, it follows from Assumption 4.5.2 that there exist scalars $\bar{r} > 0$ and $F^* \in \mathcal{R}$ such that

$$F(\bar{x}^r) = F^*, \ \forall r \in \Gamma,\ r \geq \bar{r}. \tag{4.5.10}$$

The following lemma states that the value F^* defined in (4.5.10) is a lower bound for the sequence $\{F(x^r)\}$.

Lemma 4.5.7. *Suppose that Assumptions 4.5.1-4.5.3 hold. If* $\lim\limits_{r\to\infty} F(x^r) > -\infty$*, then* $\lim\limits_{r\to\infty,\, r\in\Gamma} F(x^r) \geq F^*$*, where* F^* *is defined in (4.5.10).*

Proof. The existence of the limit of $\{F(x^r)\}_{r\in\Gamma}$ is guaranteed by Remark 4.3.3. Next, we only need to show $\lim\limits_{r\to\infty,\, r\in\Gamma} F(x^r) \geq F^*$.

Let $\bar{x}^r$ be a stationary point of problem (4.1.1). From Corollary 4.3.1 , we have $F'(\bar{x}^r, d) \geq 0$ for any $d \in \mathcal{R}^n$. Then, for any x^r satisfying (4.5.8), we get

$$\langle \nabla f(\bar{x}^r), x^r - \bar{x}^r\rangle + \tau\psi(x^r) - \tau\psi(\bar{x}^r) \geq 0. \tag{4.5.11}$$

Since f is smooth, using the mean value theorem, we have

$$f(x^r) - f(\bar{x}^r) = \langle \nabla f(\xi^r), x^r - \bar{x}^r\rangle, \tag{4.5.12}$$

where ξ^r lies on the line segment joining x^r and $\bar{x}^r$.

Then, we obtain

$$\begin{aligned} F^* - F(x^r) &= f(\bar{x}^r) - f(x^r) + \tau\psi(\bar{x}^r) - \tau\psi(x^r) \\ &\leq \langle \nabla f(\xi^r) - \nabla f(\bar{x}^r), \bar{x}^r - x^r\rangle \\ &\leq \|\nabla f(\xi^r) - \nabla f(\bar{x}^r)\|\|\bar{x}^r - x^r\| \\ &\leq L_f\|\bar{x}^r - x^r\|^2, \end{aligned}$$

where the first inequality follows from (4.5.11) and (4.5.12), and the last inequality follows from Assumption 4.5.1 and the inequality $\|\bar{x}^r - \xi^r\| \leq \|\bar{x}^r - x^r\|$.

Letting $r \to \infty$ and $r \in \Gamma$, we get the desired inequality by (4.5.8). ■

The next result presents an estimator for the distance between $F(x^r)$ and F^*. It is a modification of [69, (40) on Page 408] and plays a key role to prove the convergence rate of the BCPG methods for the nonsmooth problem (4.1.1).

Lemma 4.5.8. *Suppose that Assumptions 4.5.1-4.5.3 hold. Then, there exists a positive constant* ϖ *such that* $F(x^{\varphi(r)}) - F^* \leq -\varpi \sum_{i=r}^{\varphi(r)-1} \Delta(x^i + d^i)$ *holds for any sufficiently large* $r \in \Gamma$*, where* $\varphi(r)$ *is defined in the restricted Gauss-Seidel rule, and* F^* *is defined in (4.5.10).*

Proof. Let $\bar{x}^r$ be a stationary point of problem (4.1.1). From the restricted Gauss-Seidel rule, we have

$$\psi(\bar{x}^r) = \sum_{i=r}^{\varphi(r)-1} \psi_{J^i}(\bar{x}^r_{J^i}), \ \forall r \in \Gamma. \tag{4.5.13}$$

Then, for a sufficiently large $r \in \Gamma$, we have

$$\begin{aligned}
& F(x^{\varphi(r)}) - F^* \\
= \ & f(x^{\varphi(r)}) - f(\bar{x}^r) + \tau\psi(x^{\varphi(r)}) - \tau\psi(\bar{x}^r) \\
= \ & \langle \nabla f(\bar{\xi}^r), x^{\varphi(r)} - \bar{x}^r \rangle + \tau \sum_{i=r}^{\varphi(r)-1} \left[\psi_{J^i}(x^{\varphi(r)}_{J^i}) - \psi_{J^i}(\bar{x}^r_{J^i}) \right] \\
= \ & \left\{ \langle \nabla f(\bar{\xi}^r) - \nabla f(x^r), x^{\varphi(r)} - \bar{x}^r \rangle \right\} + \left\{ \sum_{i=r}^{\varphi(r)-1} \langle \nabla_{J^i} f(x^r) - \nabla_{J^i} f(x^i), x^{\varphi(r)}_{J^i} - \bar{x}^r_{J^i} \rangle \right\} \\
& + \left\{ \sum_{i=r}^{\varphi(r)-1} \left[\langle \nabla_{J^i} f(x^i) + \left(\nabla_1 B_{\eta^i}(x^i + d^i, x^i) \right)_{J^i}, x^{\varphi(r)}_{J^i} - \bar{x}^r_{J^i} \rangle + \tau\psi_{J^i}(x^{\varphi(r)}_{J^i}) - \tau\psi_{J^i}(\bar{x}^r_{J^i}) \right] \right\} \\
& + \left\{ \sum_{i=r}^{\varphi(r)-1} \langle \nabla_1 B_{\eta^i}(x^i + d^i, x^i)_{J^i}, \bar{x}^r_{J^i} - x^{\varphi(r)}_{J^i} \rangle \right\} \\
:= \ & S_1 + S_2 + S_3 + S_4,
\end{aligned}$$

where the second equality follows from (4.5.13) and the mean value theorem ($\bar{\xi}^r$ lies on the line segment joining $x^{\varphi(r)}$ and $\bar{x}^r$), and S_i, $i = 1, \dots, 4$, denotes the four terms in the above braces, respectively.

For S_1, S_2, and S_4, we can show that there exists a positive constant $w_3 = nL_f(\kappa w_2 + 2n) + n\overline{L}$ such that $S_1, S_2, S_4 \leq w_3 \sum_{i=r}^{\varphi(r)-1} \|d^i\|^2$ from Lemma 4.5.6 (a) and Assumptions 4.5.1 and 4.5.3.

For S_3, we get for all $\gamma \in [0, 1)$ that

$$\begin{aligned}
S_3 & \leq \sum_{i=r}^{\varphi(r)-1} (\alpha^i - 1)[(1-\gamma)\langle \nabla_1 B_{\eta^i}(x^i + d^i, x^i), d^i \rangle + \Delta(x^i + d^i)] \\
& \leq \sum_{i=r}^{\varphi(r)-1} (\alpha^i - 1)\Delta(x^i + d^i),
\end{aligned}$$

where the first inequality follows from Lemma 4.5.5, and the second inequality follows from $\langle \nabla_1 B_{\eta^i}(x^i + d^i, x^i), d^i \rangle \geq 0$, $\alpha^i \leq 1$, and $\gamma \in [0, 1)$.

Thus, we obtain

$$\begin{aligned} F(x^{\varphi(r)}) - F^* \leq & \; 3w_3 \sum_{i=r}^{\varphi(r)-1} \|d^i\|^2 + \sum_{i=r}^{\varphi(r)-1} (\alpha^i - 1)\Delta(x^i + d^i) \\ \leq & -\left(\frac{3w_3}{(1-\gamma)\underline{\mu}} + 1 \right) \sum_{i=r}^{\varphi(r)-1} \Delta(x^i + d^i), \end{aligned}$$

where the second inequality follows from Lemma 4.3.3 and $\alpha^i > 0$. Setting $\varpi = \frac{3w_3}{(1-\gamma)\underline{\mu}} + 1$, we complete the proof. ∎

Now, we show the linear convergence rates for $\{F(x^r)\}_\Gamma$ and $\{x^r\}_\Gamma$.

Theorem 4.5.1. *Suppose that Assumptions 4.5.1-4.5.3 hold. Let $\eta^r \in \Psi(X; \underline{\mu}, \overline{L})$ and $\{x^r\}$ be generated by the BCPG methods with the restricted Gauss-Seidel rule. Then, we have either $\{F(x^r)\} \downarrow -\infty$ as $r \to \infty$ or $\{F(x^r)\}_\Gamma$ converges to F^* at least Q-linearly, where F^* is defined in (4.5.10).*

Proof. From Remark 4.3.3, we have either $\{F(x^r)\} \downarrow -\infty$ or $\{F(x^r)\} > -\infty$ as $r \to \infty$. Next, we only suppose that $\{F(x^r)\} > -\infty$.

From the Armijo rule and Lemma 4.5.1, for any $r \in \Gamma$, we have

$$F(x^{i+1}) - F(x^i) \leq \sigma\alpha^i \Delta(x^i + d^i) \leq \sigma\Delta(x^i + d^i), \; \forall i \in \{r, r+1, \ldots, \varphi(r) - 1\}.$$

Summing over $i = r, r+1, \ldots, \varphi(r) - 1$, we have

$$F(x^{\varphi(r)}) - F(x^r) \leq \sigma \sum_{i=r}^{\varphi(r)-1} \Delta(x^i + d^i), \; \forall r \in \Gamma. \tag{4.5.14}$$

Using (4.5.14), Lemmas 4.5.7 and 4.5.8, we obtain for any sufficiently large $r \in \Gamma$ that

$$0 \leq F(x^{\varphi(r)}) - F^* \leq -\varpi \sum_{i=r}^{\varphi(r)-1} \Delta(x^i + d^i) \leq w_4(F(x^r) - F(x^{\varphi(r)})),$$

where $w_4 = \frac{\varpi}{\sigma} > 0$.

Rearranging the above inequalities, we have

$$F(x^{\varphi(r)}) - F^* \leq \frac{w_4}{1 + w_4}(F(x^r) - F^*), \tag{4.5.15}$$

which implies that $\{F(x^r)\}_\Gamma$ converges to F^* at least Q-linearly. ∎

Remark 4.5.1. *From (4.5.15), we find that constant w_4 has impact on the convergence of $\{F(x^r)\}$. By w_2, w_3 and ϖ in Lemmas 4.5.6 (a) and 4.5.8, we have that $w_4 = \frac{1}{\sigma}\Big\{1 + \frac{1}{(1-\gamma)\underline{\mu}}$ $[6n^2L_f + 3n\overline{L} + 3nL_f\kappa(\frac{\sqrt{(\overline{L}+\underline{\mu})^2-4\underline{\mu}^2}}{2\underline{\mu}} + \frac{\overline{L}+\underline{\mu}}{2\underline{\mu}} + \frac{L_f}{\underline{\mu}})]\Big\}$. If constant w_4 is not very big for some $\underline{\mu}, \overline{L} > 0$, $\underline{\mu} \le \overline{L}$, it follows from (4.5.15) that we may achieve good convergence by choosing kernels such that $\underline{\mu}$ is set as high as possible, and $\bar{L}$ is set as low as possible. If w_4 is sufficiently large for any $\underline{\mu}, \overline{L} > 0$, $\underline{\mu} \le \overline{L}$, $\frac{w_4}{1+w_4}$ approaches to constant 1. Then the influence of constants $\underline{\mu}$ and $\overline{L}$ is negligible.*

Theorem 4.5.2. *Suppose that Assumptions 4.5.1-4.5.3 hold. Let $\eta^r \in \Psi(X; \underline{\mu}, \overline{L})$ and $\{x^r\}$ be generated by the BCPG methods with the restricted Gauss-Seidel rule. If $\{F(x^r)\} > -\infty$, then $\{x^r\}_\Gamma$ converges to a stationary point of problem (4.1.1) at least R-linearly.*

Proof. First, we show the global convergence of $\{x^r\}_\Gamma$. For convenience, we denote $\Gamma = \{k_1, k_2, \dots\}$. Since $\{F(x^r)\}_\Gamma$ converges to F^* at least Q-linearly by Theorem 4.5.1, $\{F(x^{k_t})\}$ also converges to F^* at least R-linearly, where F^* is defined in (4.5.10). Thus, there exist constants $K > 0$, $\hat{c} \in (0,1)$, and an integer $\hat{t}$ such that

$$F(x^{k_t}) - F^* \le K\hat{c}^t \tag{4.5.16}$$

for any $t > \hat{t}$. Using (4.5.14) and Lemma 4.3.3, we have

$$F(x^{k_t}) - F(x^{k_{t+1}}) \ge -\sigma \sum_{i=k_t}^{k_{t+1}-1} \Delta(x^i + d^i) \ge -\sigma(\gamma - 1)\underline{\mu} \sum_{i=k_t}^{k_{t+1}-1} \|d^i\|^2.$$

Since $x^{i+1} = x^i + \alpha^i d^i$ and $\sup_r \alpha^r \le 1$, we further have

$$F(x^{k_t}) - F(x^{k_{t+1}}) \ge \sigma(1-\gamma)\underline{\mu} \sum_{i=k_t}^{k_{t+1}-1} \|x^{i+1} - x^i\|^2. \tag{4.5.17}$$

Hence, we obtain

$$\begin{aligned}
\|x^{k_{t+1}} - x^{k_t}\| &\le \sqrt{(k_{t+1} - k_t) \sum_{i=k_t}^{k_{t+1}-1} \|x^{i+1} - x^i\|^2} \\
&\le \sqrt{\frac{n}{\sigma(1-\gamma)\underline{\mu}}(F(x^{k_t}) - F(x^{k_{t+1}}))} \\
&= \sqrt{\frac{n}{\sigma(1-\gamma)\underline{\mu}}(F(x^{k_t}) - F^* + F^* - F(x^{k_{t+1}}))} \\
&\le \sqrt{\frac{n}{\sigma(1-\gamma)\underline{\mu}}(F(x^{k_t}) - F^*)} \\
&\le \sqrt{\frac{n}{\sigma(1-\gamma)\underline{\mu}}K\hat{c}^t},
\end{aligned}$$

where the second inequality follows from $k_{t+1} - k_t \leq n$ and (4.5.17), the third inequality follows from Lemma 4.5.7, and the last inequality follows from (4.5.16).

We denote $p := \sqrt{\frac{n}{\sigma(1-\gamma)\underline{\mu}}K}$ and $\bar{c} := \hat{c}^{\frac{1}{2}}$. Then, for any positive integers m, n with $m > n$, we have

$$\|x^{k_m} - x^{k_n}\| \leq \sum_{l=0}^{k_m-k_n-1} \|x^{k_m-l} - x^{k_m-l-1}\| \leq p \sum_{l=0}^{k_m-k_n-1} \bar{c}^{k_m-l-1} = p\frac{\bar{c}^{k_n} - \bar{c}^{k_m}}{1-\bar{c}} \leq p\frac{\bar{c}^{k_n}}{1-\bar{c}},$$

which implies that $\{x^{k_t}\}$ is a Cauchy sequence because of $0 < \bar{c} = \hat{c}^{\frac{1}{2}} < 1$. Hence, $\{x^{k_t}\}$ is convergent, i.e., $\{x^r\}_\Gamma$ is convergent. Let x^∞ denote the limit of $\{x^r\}_\Gamma$. Then, x^∞ is a stationary point of F from Theorem 4.4.1.

Finally, we complete the proof by showing the linear convergence rate of $\{x^r\}_\Gamma$.

Since $\|x^{k_m} - x^{k_n}\| \leq p\frac{\bar{c}^{k_n}-\bar{c}^{k_m}}{1-\bar{c}}$, by letting $m \to \infty$, we have

$$\|x^\infty - x^{k_n}\| = \lim_{m\to\infty} \|x^{k_m} - x^{k_n}\| \leq \lim_{m\to\infty} p\frac{\bar{c}^{k_n} - \bar{c}^{k_m}}{1-\bar{c}} = p\frac{\bar{c}^{k_n}}{1-\bar{c}},$$

which implies that $\{x^{k_t}\}$ converges to x^∞ at least R-linearly, since $0 < \bar{c} < 1$. ■

4.6 Special BCPG methods

In this section, we discuss three special BCPG methods: the block coordinate descent method, the inexact BCPG methods and an inexact block coordinate descent method. We present some sufficient conditions for their global convergence.

4.6.1 The block coordinate descent (BCD) method

In this subsection, first, we show that the unit step size is acceptable for the BCD method presented in Subsection 4.3.1. Then, we note that the requirement for its linear convergence in Theorem 4.5.2 can be weakened. To this end, we require the following definitions.

Definition 4.6.1. *A function $f : X \to \mathcal{R}$ is strongly convex with respect to the block J if the inequality $f(y_J, x_{\bar{J}}) - f(x_J, x_{\bar{J}}) - \langle \nabla_J f(x_J, x_{\bar{J}}), y_J - x_J \rangle \geq \frac{\mu_f}{2}\|y_J - x_J\|^2$ holds for any $x_J, y_J \subseteq \mathcal{R}^{|J|}$, and $(x_J, x_{\bar{J}})$, $(y_J, x_{\bar{J}}) \in X$, where μ_f is a positive constant.*

Note that, if f is strongly convex with respect to a block J, then the solution of the subproblem (4.3.3) is unique.

Remark 4.6.1. *If Definition 4.6.1 does not hold for a certain block J, we may take $f(x) + \mu\|x\|^2$ ($\mu > 0$) instead of $f(x)$ in (4.3.2). In this case, the BCPG methods reduce to the block coordinate proximal point method [74].*

Based on the Definitions 4.6.1 and 2.2.3, we assume that the smooth part f satisfies the following conditions.

Assumption 4.6.1. *(1) The function f is strongly convex with respect to each block.*

(2) The gradient ∇f is block-wise Lipschitz continuous with respect to each block, which is precisely defined by Definition 2.2.3 in Subsection 2.2.2.

For convenience, we let the scalar μ_{min} denote the smallest strongly convex parameter for different blocks, and let the scalar L_{max} denote the largest Lipschitz constant of f for different blocks.

The following theorem states that the unit step size ($\alpha^r = 1$) can be adapted for the BCD method.

Theorem 4.6.1. *Suppose that Assumption 4.6.1 holds. Then, the unit step size $\alpha^r = 1$ is acceptable for the BCD method.*

Proof. Without loss of generality, we consider the case at the r-th iteration. Thus, we obtain $d^r_{\bar{J}^r} = 0$. Since the BCD method is a special case of the BCPG method, we can prove the following inequality by a similar argument for (4.3.9) in Lemma 4.3.1.

$$\tau\psi(x^r + d^r) - \tau\psi(x^r) \leq -\langle \nabla f(x^r + d^r), d^r \rangle. \tag{4.6.1}$$

Then, $\Delta(x^r + d^r)$ in (4.3.13) can be rewritten as follows.

$$\Delta(x^r + d^r) = \langle \nabla f(x^r), d^r \rangle + \gamma\langle \nabla f(x^r + d^r) - \nabla f(x^r), d^r \rangle + \tau\psi(x^r + d^r) - \tau\psi(x^r). \tag{4.6.2}$$

Hence,

$$\begin{aligned}
& F(x^r + d^r) - F(x^r) - \sigma\Delta(x^r + d^r) \\
= \;& f(x^r + d^r) - f(x^r) - \sigma(1-\gamma)\langle \nabla f(x^r), d^r \rangle - \sigma\gamma\langle \nabla f(x^r + d^r), d^r \rangle \\
& + (1-\sigma)[\tau\psi(x^r + d^r) - \tau\psi(x^r)] \\
\leq \;& f(x^r + d^r) - f(x^r) - \langle \nabla f(x^r + d^r), d^r \rangle - \sigma(1-\gamma)\langle \nabla f(x^r) - \nabla f(x^r + d^r), d^r \rangle \\
\leq \;& \left(-\frac{\mu_{min}}{2} + \sigma(1-\gamma)L_{max}\right)\|d^r\|^2,
\end{aligned}$$

where the first equality follows from (4.6.2), the first inequality follows from (4.6.1), and the last inequality follows from the block strong convexity and block gradient Lipschitz continuity of f.

Note that the parameters μ_{min} and L_{max} are fixed constants. By selecting appropriate parameters σ and γ in the Armijo rule, we can ensure that $-\frac{\mu_{min}}{2} + \sigma(1-\gamma)L_{max} < 0$ for all r, i.e., $\alpha^r = 1$ is acceptable for the BCD method. ■

The next theorem states the convergence rate of the BCD method, which is a direct corollary of Theorem 4.5.2.

Theorem 4.6.2. *Suppose that function f in problem (4.1.1) is strongly convex and Assumption 4.5.1 holds. Let $\{J^r\}$ be selected by the restricted Gauss-Seidel rule. Then $\{x^r\}_\Gamma$ generated by the BCD method converges to an optimal solution of problem (4.1.1) at least R-linearly.*

Note that, Theorem 4.6.2 holds even if f is block strongly convex and block gradient Lipschitz continuous for each block. To the best of our knowledge, this is the first linear convergence result on the classical block coordinate descent method for the nonsmooth problem (4.1.1).

4.6.2 Inexact BCPG methods

As described in Subsection 4.3.1, the inexact BCPG methods with the criterion (4.3.4) can be regarded as the special cases of the proposed BCPG methods with the kernel $\tilde{\eta}^r(x) = \eta(x) + x^T E^r x$. Next, we present a sufficient condition on the error ε^r and the direction d^r for the convergence of the inexact BCPG methods.

Lemma 4.6.1. *Let $\eta^r \in \Psi(X; \underline{\mu}, \overline{L})$, $\delta_\mu \in (0, \underline{\mu})$, and $\delta_L \in (0, \infty)$. Suppose that $(d^r, \varepsilon^r) \in \mathcal{R}^n \times \mathcal{R}^n$ satisfies (4.3.4) and*

$$|\varepsilon_i^r| \le \min\{\delta_\mu, \delta_L\} |d_i^r|. \tag{4.6.3}$$

Then, for all r, the kernel $\tilde{\eta}^r(x)$ belongs to the set $\Psi(X; \hat{\mu}, \hat{L})$, where $\tilde{\eta}^r(x)$ is defined by (4.3.7), $\hat{\mu} = \underline{\mu} - \delta_\mu > 0$, and $\hat{L} = \overline{L} + \delta_L$.

Proof. First, we show that $\tilde{\eta}^r(x) = \eta^r(x) + x^T E^r x \in \Phi(X; \hat{\mu})$. It is equivalent to show that $\langle \nabla\tilde{\eta}^r(y) - \nabla\tilde{\eta}^r(x), y - x\rangle - \hat{\mu}\|y - x\|^2 \ge 0$ for any $x, y \in \text{int}X$.

In fact, we have

$$\begin{aligned}
&\langle \nabla\tilde{\eta}^r(y) - \nabla\tilde{\eta}^r(x), y - x\rangle - \hat{\mu}\|y - x\|^2 \\
= \; & \langle \nabla\eta^r(y) - \nabla\eta^r(x), y - x\rangle + (y - x)^T (E^r - \hat{\mu} I)(y - x) \\
\ge \; & (y - x)^T (\underline{\mu} I + E^r - \hat{\mu} I)(y - x) \\
\ge \; & 0,
\end{aligned}$$

where $I \in \mathcal{R}^{n\times n}$ is an identity matrix, and the first inequality follows from strong convexity of function η^r, and the last inequality holds from (4.3.5) and (4.6.3). Hence, the first part is proved.

In the next part, we prove that $\nabla\tilde{\eta}^r(x)$ is $\hat{L}$-Lipschitz continuous. It is equivalent to show that $\|\nabla\tilde{\eta}^r(y) - \nabla\tilde{\eta}^r(x)\| \le \hat{L}\|y - x\|$ for any $x, y \in \text{int}X$. In fact, we have

$$\|\nabla\tilde{\eta}^r(y) - \nabla\tilde{\eta}^r(x)\| \le \|\nabla\eta^r(y) - \nabla\eta^r(x)\| + \|E^r(y - x)\|$$

$$\begin{aligned}&\leq (\overline{L} + \max|E^r_{ii}|)\|y - x\| \\ &\leq \hat{L}\|y - x\|,\end{aligned}$$

where the second inequality holds since $\nabla\eta^r$ is $\overline{L}$-Lipschitz continuous and the matrix E^r is diagonal, and the last inequality follows from (4.3.5) and (4.6.3). ■

Remark 4.6.2. *The error ε^r may be explicitly given by a subgradient of function F. If $\psi(x) = 0$, i.e., problem (4.1.1) is a smooth optimization problem, then we may set $\varepsilon^r_j = -\nabla_j f(x^r) - \nabla_j\eta(x^r + d^r) + \nabla_j\eta(x^r)$, $j \in J^r$. If $\psi(x)$ is nondifferentiable, then we have to consider the subdifferential $\partial_{J^r}\psi(x^r + d^r)$. In some applications, $\partial_{J^r}\psi$ is explicitly given (For example, when $\psi(x) = \sum_{i=1}^n |x_i|$ and $x^r_j + d^r_j = 0, \partial_j\psi(x^r + d^r) = [-1, 1]$). Then we may set $\varepsilon^r_j = \operatorname*{argmin}_{\xi\in\partial_j\psi(x^r+d^r)} \{-\tau\xi - \nabla_j f(x^r) - \nabla_j\eta(x^r + d^r) + \nabla_j\eta(x^r)\}$, $j \in J^r$.*

Lemma 4.6.1 shows that the inexact BCPG methods are reduced to the exact BCPG methods with $\tilde{\eta}^r \in \Psi(X; \hat{\mu}, \hat{L})$ in Step 2. Combining Lemma 4.6.1 and Theorem 4.5.2, we obtain the following theorem immediately, which states the linear convergence rate of the inexact BCPG methods.

Theorem 4.6.3. *Let $\eta^r \in \Psi(X; \underline{\mu}, \overline{L})$. Suppose that Assumptions 4.5.1-4.5.3 hold and that (d^r, ε^r) satisfies (4.3.4) and (4.6.3) for any $r > 0$. Then, $\{x^r\}_\Gamma$ generated by the inexact BCPG methods with the restricted Gauss-Seidel rule converges to a stationary point of problem (4.1.1) at least R-linearly if $\{F(x^r)\} > -\infty$.*

4.6.3 An inexact block coordinate descent (BCD) method

Letting the kernels of the inexact BCPG methods be the functions defined by (4.3.2), the inexact BCPG methods reduce to an inexact BCD method. In this subsection, we establish a practical criterion for the inexactness, and propose a specific inexact BCD algorithm with unit step size for solving problem (4.1.1). We show that the proposed algorithm has R-linear convergence rate as well.

By the definition of the approximate solution in (4.3.4), we say that $d^r_{J^r}$ is an approximate solution of subproblem (4.3.3) with error $\varepsilon^r_{J^r}$ if the pair $(d^r_{J^r}, \varepsilon^r_{J^r})$ satisfies

$$\nabla_{J^r} f(x^r_{J^r} + d^r_{J^r}, x^r_{\bar{J}^r}) + \varepsilon^r_{J^r} \in -\tau\partial\psi_{J^r}(x^r_{J^r} + d^r_{J^r}). \tag{4.6.4}$$

Then the condition (4.6.3) holds if the direction $d^r_{J^r}$satisfies the following inequality.

$$\min_{\xi\in\partial\psi_{J^r}(x^r_{J^r}+d^r_{J^r})} \|\nabla_{J^r} f(x^r_{J^r} + d^r_{J^r}, x^r_{\bar{J}^r}) + \tau\xi\| \leq \min\{\delta_\mu, \delta_L\}\|d^r_{J^r}\|, \tag{4.6.5}$$

where $\delta_\mu \in (0, \underline{\mu})$, and $\delta_L \in (0, \infty)$. In the following part, we adopt inequality (4.6.5) as a criterion for the inexact BCD method.

The following theorem shows that the unit step size ($\alpha^r = 1$) is also acceptable for the inexact BCD method with (4.6.5).

Theorem 4.6.4. *Suppose that Assumption 4.6.1 holds, and that the direction d^r satisfies (4.6.5) for any r. Then, the unit step size $\alpha = 1$ is acceptable for the inexact BCD method with (4.6.5).*

Proof. To show this theorem, it is sufficient to show that step size $\alpha^r = 1$ satisfies inequality (4.3.12) in Armijo rule with the kernel (4.3.2). Let $\varepsilon^r_{J^r}$ denote the error corresponding to the direction $d^r_{J^r}$. From (4.6.4), we obtain

$$\tau\psi_{J^r}(x^r_{J^r} + d^r_{J^r}) - \tau\psi_{J^r}(x^r_{J^r}) \le -\langle \nabla_{J^r} f(x^r_{J^r} + d^r_{J^r}, x^r_{\bar{J}^r}) + \varepsilon^r_{J^r}, d^r_{J^r} \rangle.$$

By a similar deduction to the proof of Theorem 4.6.1, we get

$$\begin{aligned}
& F(x^r_{J^r} + d^r_{J^r}, x^r_{\bar{J}^r}) - F(x^r) - \sigma\Delta(x^r_{J^r} + d^r_{J^r}, x^r_{\bar{J}^r}) \\
\le\ & f(x^r_{J^r} + d^r_{J^r}, x^r_{\bar{J}^r}) - f(x^r) - \langle \nabla_{J^r} f(x^r_{J^r} + d^r_{J^r}, x^r_{\bar{J}^r}), d^r_{J^r} \rangle \\
& - \sigma(1-\gamma)\langle \nabla_{J^r} f(x^r) - \nabla_{J^r} f(x^r_{J^r} + d^r_{J^r}, x^r_{\bar{J}^r}), d^r_{J^r} \rangle - (1-\sigma)\langle \varepsilon^r_{J^r}, d^r_{J^r} \rangle \\
\le\ & \Big(-\frac{\mu_{min}}{2} + \sigma(1-\gamma)L_{max} + (1-\sigma)\min\{\delta_\mu, \delta_L\} \Big) \|d^r_{J^r}\|^2,
\end{aligned}$$

where the last inequality follows from Assumption 4.6.1 and (4.6.3). Note that there exist parameters $\sigma \in (0,1)$ and $\gamma \in [0,1)$ such that $-\frac{\mu_{min}}{2} + \sigma(1-\gamma)L_{max} + (1-\sigma)\min\{\delta_\mu, \delta_L\} < 0$. Hence, $\alpha^r = 1$ is acceptable. ■

Now we describe the inexact BCD algorithm as follows.

```
An inexact block coordinate descent algorithm:
Step 0: Select an initial point $x^0 \in \text{int}X$, and let $r = 0$.
Step 1: If a termination condition holds, then stop.
Step 2: Select a block $J^r$ by the restricted Gauss-Seidel rule.
Step 3: Solve subproblem (4.3.3) by a proper method to get a search direction $d^r$ satisfying (4.6.5).
Step 4: Set $x^{r+1}_{J^r} = x^r_{J^r} + d^r_{J^r}$, $x^{r+1}_{\bar{J}^r} = x^r_{\bar{J}^r}$, and $r = r + 1$. Go to Step 1.
```

We would like to emphasize that the inexact BCD algorithm does not use the line search. Hence, it is suitable for large scale problems, whose objective function values are expensive to evaluate.

The following theorem shows the linear convergence rate of the inexact BCD algorithm, which follows from Theorems 4.6.3 and 4.6.4 immediately.

Theorem 4.6.5. *Suppose that function f in problem (4.1.1) is strongly convex and that Assumption 4.5.1 holds. Then, $\{x^r\}_\Gamma$ generated by the inexact BCD algorithm converges to an optimal point of problem (4.1.1) at least R-linearly.*

As noted in Subsection 4.6.1, Theorem 4.6.5 also holds under Assumption 4.6.1.

4.7 Numerical experiments

In this section, we propose a new algorithm for a convex optimization problem with separable simplex constraints, which is an inexact BCPG method with variable kernels. We also report numerical results for the proposed algorithm and compare it with the exponentiated gradient algorithm, which is one of the standard solvers in the machine learning community.

4.7.1 The Log-linear dual problem

Let $\{(y_i, z_i),\ i = 1, 2, \ldots, l\}$ be given data, where $y_i \in \mathcal{X}$ and $z_i \in \mathcal{Y} := \{1, 2, \ldots, m\}$ represent features and a label (class) of data, respectively. Some of the structured prediction problems in supervised machine learning [19] can be written as follows.

$$\underset{w}{\text{minimize}} \quad -\sum_{i}^{l} \ln p(z_i|y_i; w) + \frac{C}{2}\|w\|^2, \tag{4.7.1}$$

where $C > 0$ is a regularization constant, $w \in \mathcal{R}^d$ is a decision parameter, and function $p(z_i|y_i; w)$ is the conditional distribution defined by

$$p(z_i|y_i; w) = \frac{1}{\displaystyle\sum_{j=1}^{m} \exp^{\langle w, \phi(y_i, j)\rangle}} \exp^{\langle w, \phi(y_i, z_i)\rangle},$$

where function $\phi(u, v) : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}^d$ maps data (u, v) to feature vectors.

Collins et al. [19] show that problem (4.7.1) can be transformed into the following convex dual problem, which is called “ the Log-linear dual” in [19].

$$\begin{aligned} &\text{minimize} \ \ \tilde{F}(x) := \frac{1}{2}x^T A x + \sum_{i}^{n} x_i \ln x_i \\ &\text{subject to} \ \ x \in \Delta^l, \end{aligned} \tag{4.7.2}$$

where $x \in \mathcal{R}^n$, $n = lm$, and Δ^l is the Cartesian product of Δ, i.e., $\Delta^l = \Delta \times \cdots \times \Delta$, Δ denotes the simplex of distributions over a classification, i.e.,

$$\Delta = \{x \in \mathcal{R}^m \mid x_i \geq 0, \sum_{i=1}^{m} x_i = 1\}.$$

Moreover, A is an $\mathcal{R}^{n\times n}$ matrix given by

$$A_{tm+i,hm+j} := \frac{1}{C}\left\langle \varphi_{t+1,i}, \varphi_{h+1,j} \right\rangle ,$$

where $\varphi_{t+1,i} = \phi(y_{t+1}, z_{t+1}) - \phi(y_{t+1}, z_i)$, $t, h \in \{0, 1, \dots, l-1\}$ and $i, j \in \{1, \dots, m\}$.

Note that matrix A in (4.7.2) is symmetric and positive semidefinite. For convenience, we use the following notations in this section. The k-th $m \times m$ diagonal block of matrix A is denoted by $A_{[kk]} \in \mathcal{R}^{m\times m}$. The vector $e_{[n]} \in \mathcal{R}^n$ denotes the vector, whose components are all ones. Moreover, we choose blocks $\{J^r\}$ in this section as follows.

$$J^r = \{k(r)m + 1, k(r)m + 2, \dots, k(r)m + m\}, \tag{4.7.3}$$

where $k(r) = r \bmod l$, that is, we choose blocks by the cyclic rule, defined in Section 2.4, with $N = l$, $\mathcal{J}^i = \{i, i + 1, \dots, i + m - 1\}$, and $i = 1, \dots, l$.

4.7.2 Block type exponentiated gradient (B-EG) algorithm

The exponentiated gradient (EG) algorithm [6, 19] is a very useful method for solving (4.7.2), a problem over unit simplices, since it has an exact closed-form solution on each iteration. Moreover, as described in Table 4.1, the EG algorithm is a special BCPG method with $\eta^r(v) = \frac{1}{t_r}\sum_i v_i \ln v_i$. By easy computing, the solution of subproblem (4.3.1) can be written as follows.

$$d^r_j = \frac{x^r_j \exp^{(-t_r\nabla_j\tilde{F}(x^r))}}{\sum\limits_{j\in J^r} x^r_j \exp^{(-t_r\nabla_j\tilde{F}(x^r))}} - x^r_j, \quad \forall j \in J^r. \tag{4.7.4}$$

It is shown in [19, Theorem 1] that the “batch” EG algorithm (see [19, Figure 1] for details) with a fixed step size $t \in (0, \frac{1}{l\|A\|_\infty}]$ converges linearly for problem (4.7.2). This result can be similarly extended to the following “block” type EG algorithm.

Block type exponentiated gradient (B-EG) algorithm:

Step 0: Select an initial point $x^0 \in \Delta^l$ such that $x^0_i > 0$ for all i, and let $r = 0$.

Step 1: Let $t_r = \frac{1}{\|A\|_\infty}$ and determine the block J^r by (4.7.3). Compute $d^r_{J^r}$ by (4.7.4).

Step 2: Set $x^{r+1}_{J^r} = x^r_{J^r} + d^r_{J^r}$, $x^{r+1}_{\bar{J}^r} = x^r_{\bar{J}^r}$ and $r = r + 1$. Go to Step 1.

Remark 4.7.1. *Note that the B-EG algorithm is a special BCPG method with the fixed kernel $\eta^r(v) = \|A\|_\infty\sum_i v_i \ln v_i \in \Phi(\Delta^n, \|A\|_\infty)$. Using [19, Lemma 2], we can verify that inequality (4.3.12) holds with the unit step size $\alpha^r = 1$. Hence, it follows from Theorem 4.4.1 that the B-EG algorithm is a globally convergent algorithm for problem (4.7.2).*

4.7.3 An inexact BCPG algorithm with variable kernels

Although the EG algorithm is very useful for problem (4.7.2), its convergence speed is still slow. This fact motivates us to try other kernels. We propose a hybrid method of the EG method and the inexact BCD method.

The EG method is studied in Subsection 4.7.2. Next we analyze the case of the inexact BCD method, that is, we need to get an approximate solution to the following problem.

$$\begin{aligned} &\underset{d_{J^r}}{\text{minimize}} \;\; \tilde{F}(x^r_{J^r} + d_{J^r}, x^r_{\bar{J}^r}) \\ &\text{subject to} \;\; x^r_{J^r} + d_{J^r} \in \Delta. \end{aligned} \tag{4.7.5}$$

We consider the Newton method. However, the inequality constraints $x^r_i + d_i \geq 0$ in the simplex cause the difficulty to solve the Newton subproblem. Hence, we ignore the constraint $x^r_i + d_i \geq 0$ and solve the following problem with one equality constraint.

$$\begin{aligned} &\text{minimize} \;\; \langle \nabla_{J^r} \tilde{F}(x^r), d_{J^r} \rangle + \frac{1}{2} d^T_{J^r} \nabla^2_{[rr]} \tilde{F}(x^r) d_{J^r} \\ &\text{subject to} \;\; \sum_{i=1}^{m} d_i = 0, \end{aligned} \tag{4.7.6}$$

where $\nabla^2_{[rr]} \tilde{F}(x^r) \in \mathcal{R}^{m \times m}$ denotes the corresponding diagonal block of $\nabla^2 \tilde{F}(x^r)$ to the block J^r. By the KKT condition for (4.7.6), we have

$$\begin{pmatrix} d^r_{J^r} \\ \lambda \end{pmatrix} = \begin{pmatrix} \nabla^2_{[rr]} \tilde{F}(x^r) & e_{[m]} \\ e^T_{[m]} & 0 \end{pmatrix}^{-1} \begin{pmatrix} \nabla_{J^r} \tilde{F}(x^r) \\ 0 \end{pmatrix}, \tag{4.7.7}$$

where $\lambda \in \mathcal{R}$ denotes a Lagrange multiplier. Note that if $x^r_{J^r} + d^r_{J^r} > 0$, then the solution $d^r_{J^r}$ is also a solution of the Newton subproblem for (4.7.5). Let $\Xi := \{d \in \mathcal{R}^m \mid \sum_{i=1}^m d_i = 0\}$. Then it can be verified that $\partial \delta_\Xi(d) = \{\gamma e_{[m]} \mid \gamma \in \mathcal{R}\}$, and

$$\min_{\xi \in \partial \delta_\Xi(d)} \|\nabla_{J^r} \tilde{F}(x^r_{J^r} + d^r_{J^r}, x^r_{\bar{J}^r}) + \tau \xi\| = \left\| -\nabla_{J^r} \tilde{F}(x^r + d^r) + \frac{1}{m} \sum_{i \in J^r} \nabla_i \tilde{F}(x^r_{J^r} + d^r_{J^r}, x^r_{\bar{J}^r}) e_{[m]} \right\|.$$

With (4.6.5), we adopt the following conditions as the criterion for the approximate solution of subproblem (4.3.3).

$$\begin{cases} \left\| -\nabla_{J^r} \tilde{F}(x^r + d^r) + \frac{1}{m} \sum\limits_{i \in J^r} \nabla_{J^r} \tilde{F}(x^r_{J^r} + d^r_{J^r}, x^r_{\bar{J}^r}) e_{[m]} \right\| \leq (\|A\| + 1) \|d^r_{J^r}\| \\ x^r_{J^r} + d^r_{J^r} > 0. \end{cases} \tag{4.7.8}$$

Note that criterion (4.7.8) holds for sufficiently large r, since $\tilde{F}$ is strongly convex on Δ^l and the solution x^* of problem (4.7.2) satisfies $x^*_i > 0$.

Now we formally present the proposed algorithm as follows.

```
Algorithm 4-1: A hybrid method of B-EG method and inexact BCD method
Step 0: Select an initial point $x^0 \in \Delta^l$ such that $x_i^0 > 0$ for all $i$ and let $r = 0$.
Step 1: Determine the block $J^r$ by (4.7.3).
Step 2: Get a direction $d^r_{J^r}$ by (4.7.7). If criterion (4.7.8) holds, then set $x^{r+1}_{J^r} = x^r_{J^r} + d^r_{J^r}$, $x^{r+1}_{\bar{J}^r} = x^r_{\bar{J}^r}$, and $r = r + 1$. Go to Step 1. Otherwise go to Step 3.
Step 3: Compute $d^r_{J^r}$ by (4.7.4) with $t_r = \frac{1}{\|A\|_\infty}$. Set $x^{r+1}_{J^r} = x^r_{J^r} + d^r_{J^r}$, $x^{r+1}_{\bar{J}^r} = x^r_{\bar{J}^r}$ and $r = r + 1$. Go to Step 1.
```

Note that, on each iteration of Algorithm 4-1, we must calculate the Newton direction (4.7.7). Although criterion (4.7.8) holds for sufficiently large r, it may fail in the early steps of Algorithm 4-1. Hence, the calculation of (4.7.7) is redundant for some steps. To reduce the wasted calculations, we may exploit the local error bound or some identification techniques as a switch. Here, we omit such techniques for simplicity.

Moreover, for given $\nabla_{J^r}\tilde{F}(x^r)$, the iteration complexity of (4.7.7) is $O(m^3)$. If we calculate the eigenvalue decomposition of $A_{[kk]}$ in advance, it can be reduced to $O(m^2)$. On the other hand, the calculation of $\nabla_{J^r}\tilde{F}(x^r)$ for (4.7.4) at each iteration is $O(mn)$. Therefore, if $m \ll n$, then the burden of Algorithm 4-1 is the calculation of $\nabla_{J^r}\tilde{F}(x^r)$, and hence the CPU time of one iteration on Algorithm 4-1 is almost same as that of the B-EG algorithm.

Remark 4.7.2. *It follows from Theorems 4.4.1, 4.6.4 and Remark 4.7.1 that Algorithm 4-1 is a globally convergent algorithm for problem (4.7.2).*

4.7.4 Results

In this subsection, we report numerical results of Algorithm 4-1 and compare it with the B-EG algorithm. The algorithms are implemented in MATLAB (version 8.3.0.532 (R2014a)) and running on an Intel(R) Core(TM) i5-3470 CPU @3.20GHz. In our implementation, we let matrix $A = \tilde{A}^T\tilde{A}$, where $\tilde{A}$ is an $a \times n$ matrix, whose elements are generated randomly with uniform distribution in the interval $(-\frac{1}{2}, \frac{1}{2})$. Note that matrix A is singular, when $a < n$. Besides, we choose

$$x^0 = \frac{1}{m}(1, 1, \ldots, 1)^T \in \Delta.$$

We present numerical results on Algorithm 4-1 and the B-EG algorithm for different size problems: $n = 1000, 2000, 5000$ and $m = 10, 50$ in Figures 1-6, which show plots of the objective function values versus the iteration number of $\frac{r}{l}$, that is, the number of the iterations for the whole variable. From Figures 1-6, we can see that Algorithm 4-1 converges significantly faster than the B-EG algorithm.

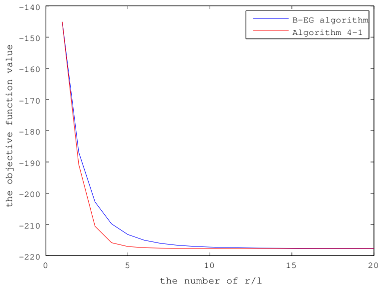

Figure 4.1: $n = 1000, a = 200, m = 10$

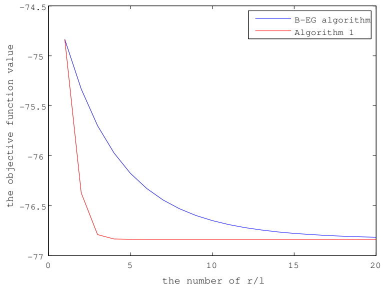

Figure 4.2: $n = 1000, a = 200, m = 50$

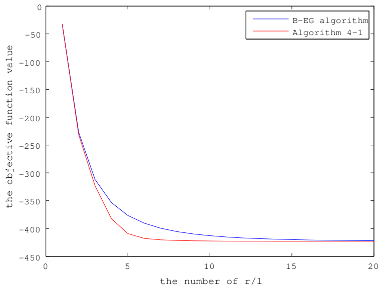

Figure 4.3: $n = 2000, a = 500, m = 10$

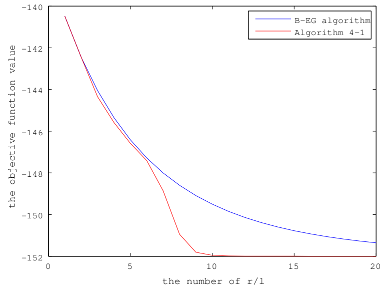

Figure 4.4: $n = 2000, a = 500, m = 50$

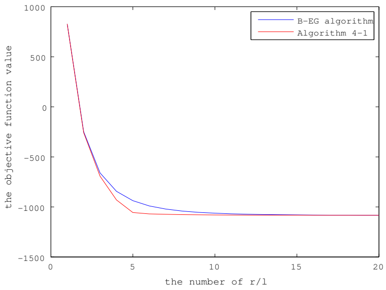

Figure 4.5: $n = 5000, a = 1000, m = 10$

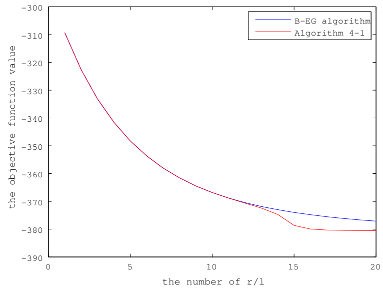

Figure 4.6: $n = 5000, a = 1000, m = 50$

4.8 Conclusion

In this chapter, we have presented a class of block coordinate proximal gradient (BCPG) methods for solving the structured nonconvex optimization problem (4.1.1). For the proposed methods, we have established their global convergence and R-linear convergence rate under some appropriate assumptions. The idea of using the variable kernels is the innovation of this chapter, which enables us to obtain many well-known algorithms from the proposed BCPG methods, including the (inexact) BCD method. Moreover, some special kernels allow the proposed BCPG methods to adopt the fixed step size. Finally, they help us to construct accelerated algorithms.

There are many issues for the future research.

(1) To extend the linear convergence rate of the proposed BCPG method with the generalized Gauss-Seidel (G-G-S) rule. The essential difference between the G-G-S rule and the restricted Gauss-Seidel rule lies in whether the blocks can be overlapping or not. In Chapter 3, we have shown the linear convergence of the CD method with the G-G-S rule. Thus we think that it may be possible to show the linear convergence of the BCPG method with the G-G-S rule.

(2) To give a convergence speed analysis of the proposed Algorithm 4-1. On step 3 of Algorithm 4-1, we adopt the entropy kernel $\eta^r = \frac{1}{t_r} \sum_{i=1}^{l} v_i \ln v_i$, which is not gradient Lipschitz continuous as v_i approaches to the bound of the simplex Δ. Then entropy kernel function does not belong to the function set $\Psi(\Delta^l; \underline{\mu}, \overline{L})$. Thus we can not yield the R-linear convergence rate of the Algorithm 4-1 from Theorem 4.5.2. However, in [19], it is shown that the B-EG method with the random rule has an exponential convergence rate. Note that the cyclic rule can be looked on as a special case of the random rule. The numerical results in this chapter show that the proposed algorithm converge faster than the B-EG method. Hence, it may be possible to show that Algorithm 4-1 converges at leat exponentially.

(3) To study the error bound further. In this chapter, the local error bound, Assumption 4.5.3, is the key for establishing the convergence rate of the BCPG methods. In [74], Kurdyka-Lojasiewiez (KL) inequality is shown to be the central for the BCD method. It is interesting to study the relation between the KL inequality and the local error bound in the future.

(4) To extend the BCPG methods to the more general constrained problems, such as the SVM problem. It is a challenging topic.

Chapter 5

Iteration complexity of a block coordinate gradient descent method for convex optimization problem

5.1 Introduction

In this chapter, we consider the following nonsmooth convex optimization problem.

$$\underset{x}{\text{minimize}}\, F(x) := f(x) + \tau\psi(x), \tag{5.1.1}$$

where f is smooth and *convex* on an open subset of $\mathcal{R}^n$ containing $\operatorname{dom}\psi := \{x \in \mathcal{R}^n \mid \psi(x) < \infty\}$, τ is a positive constant, and $\psi : \mathcal{R}^n \to (-\infty, \infty]$ is a proper, convex and l.s.c. function with a block separable structure.

Note that problem (5.1.1) considered in this chapter is a convex problem, since both functions f and ψ are assumed to be convex here.

It is known that (block) coordinate descent-type methods are very efficient for large scale problems [7, 58, 59, 69, 78]. Recently, the topic of the iteration complexity of these methods has been extensively discussed [7, 32, 62]. In most of the existing results, the block coordinate descent (BCD)-type methods with the cyclic rule have $O(\frac{NL_f}{\varepsilon})$ iteration complexity, where L_f is the Lipschitz constant for ∇f, N is the number of blocks, and $\varepsilon > 0$ is the approximation accuracy. For details, see [7, 32] and references therein. However, when problem (5.1.1) is an l_1-regularized problem, it is shown in [62] that the iteration complexity of the coordinate descent (CD) method can be improved to $O(\frac{L_f}{\varepsilon})$ under an isotonicity assumption. It is worth noting that this upper bound does not depend on the dimension n. This result implies that the existing results $O(\frac{NL_f}{\varepsilon})$ on the block type method may be too loose, since the CD method is a special case of the block coordinate descent method.

In this chapter, we further improve the iteration complexity of a block coordinate gradient descent (BCGD) method with a cyclic rule for problem (5.1.1), and show that the complexity bound is potentially independent of the number N of blocks. In particular, we make our research on the following two aspects.

- Based on the Lipschitz continuity-like assumption (Definition 2.2.4 in Subsection 2.2.2), we prove that the iteration complexity of the proposed BCGD method may be improved to $O(\frac{\max\{M, L_f\}}{\varepsilon})$ (Corollary 5.3.1 in Section 5.3), where M is the constant given in the proposed assumption.

- We analyze the relation between the constant M and the Lipschitz constant L_f. We show that $M \le \sqrt{N} L_f$ holds for general functions (Theorem 5.4.1 in Section 5.4), and list some special functions (Theorems 5.4.2–5.4.4 in Section 5.4) such that $M \le 2L_f$. These relations yield a sharper iteration complexity bound than those in [7] and [32]. See Table 5.1 in Section 5.5 for details.

This chapter is organized as follows. In Section 5.2, we introduce some basic assumptions and relevant properties. Section 5.3 presents the proposed algorithm, states our new assumption, and derives the resulting iteration complexity. The relations between M and L_f are discussed in Section 5.4. Finally, we conclude this chapter in Section 5.5.

5.2 Preliminaries

In this section, we introduce some basic assumptions and relevant properties.

Throughout this chapter, we assume that the optimal solution set, denoted by X^*, is nonempty. Since problem (5.1.1) is convex, every local minimum is also a global minimum, denoted by F^*. Let $\{\mathcal{J}^1, \mathcal{J}^2, \dots, \mathcal{J}^N\}$ be a partition of the set $\{1, 2, \dots, n\}$. Hence, $x^T = (x^T_{\mathcal{J}^1}, \dots, x^T_{\mathcal{J}^N})$. Moreover, the function ψ in problem (5.1.1) is block separable with respect to each block $\mathcal{J}^i$, that is, there exist N functions $\psi_i : \mathcal{R}^{|\mathcal{J}^i|} \to \mathcal{R}$, $i = 1, \dots, N$, such that $\psi(x) = \sum_{i=1}^N \psi_i(x_{\mathcal{J}^i})$.

For the smooth function f in problem (5.1.1), we assume that

Assumption 5.2.1. *The gradient ∇f is block-wise Lipschitz continuous with positive constants $\{L_1, \dots, L_N\}$, which is precisely defined by Definition 2.2.3 in Subsection 2.2.2.*

Under Assumption 5.2.1, we have the following lemma, which is given in [7, Lemma 3.2] and [50, Lemma 2]. For simplicity, we omit its proof here.

Lemma 5.2.1. *Suppose that Assumption 5.2.1 holds. Then, the following statements hold.*

(i) *For any $y, x \in \operatorname{dom} f$ with $x_{\bar{\mathcal{J}}^i} = y_{\bar{\mathcal{J}}^i}$, $f(y) \le f(x) + \langle \nabla_{\mathcal{J}^i} f(x), y_{\mathcal{J}^i} - x_{\mathcal{J}^i} \rangle + \frac{L_i}{2} \|y_{\mathcal{J}^i} - x_{\mathcal{J}^i}\|^2$.*

(ii) *There exists a positive constant L_f such that $L_f \le \sum\limits_{i=1}^N L_i$ and $\|\nabla f(y) - \nabla f(x)\| \le L_f \|y - x\|$ for any $y, x \in \operatorname{dom} f$.*

Additionally, we assume that the level set satisfies the following assumption.

Assumption 5.2.2. *For any initial point x^0, the measure of the level set*

$$R_* := \max_{y} \max_{x^* \in X^*} \left\{ \|y - x^*\| \mid F(y) \leq F(x^0) \right\}$$

is bounded.

5.3 Iteration complexity analysis of the BCGD method

In this section, we mainly focus on establishing the iteration complexity of a BCGD method with a cyclic rule for solving (5.1.1). First, we introduce the specified algorithm in Subsection 5.3.1. We then propose a new assumption and present several technical lemmas in Subsection 5.3.2. Finally, in Subsection 5.3.3, we give our main results on the iteration complexity for the cases in which the smooth function f is convex and the cost function F is strongly convex.

5.3.1 The BCGD method

The proposed algorithm proceeds as follows.

```
Algorithm 5-1. BCGD method with the cyclic rule:
Step 0: Choose an initial point $x^0 \in \operatorname{dom} F$, and let $r = 0$.
Step 1: If some termination condition holds, then stop.
Step 2-0: Let $x^{r,0} = x^r$ and $i = 1$.
Step 2-1: Solve the following subproblem with $\widehat{L_k} \in [L_k, +\infty)$, and find a search direction $d^{r,i}$.
    $$d^{r,i} = \operatorname*{argmin}_{d \in \mathcal{R}^n} \left\{ \langle \nabla f(x^{r,i-1}), d \rangle + \frac{1}{2} \sum_{k=1}^{N} \widehat{L_k} \|d_{\mathcal{J}^k}\|^2 + \tau \psi(x^{r,i-1} + d) \,\middle|\, d_{\bar{\mathcal{J}}^i} = 0 \right\}.$$
Step 2-2: Set $x^{r,i}_{\mathcal{J}^i} = x^{r,i-1}_{\mathcal{J}^i} + d^{r,i}_{\mathcal{J}^i}$, $x^{r,i}_{\bar{\mathcal{J}}^i} = x^{r,i-1}_{\bar{\mathcal{J}}^i}$, and $i = i + 1$. If $i = N + 1$, then go to Step 3. Otherwise, go to Step 2-1.
Step 3: Let $x^{r+1} = x^{r,N}$, and $r = r + 1$. Go to Step 1.
```

The sequence $\{x^r\}$ generated by Algorithm 5-1 has the following properties.

Lemma 5.3.1. *For any $i \in \{1, 2, \ldots, N\}$ and $r \geq 0$, we have*

$$x^r_{\mathcal{J}^i} = x^{r,0}_{\mathcal{J}^i} = x^{r,j}_{\mathcal{J}^i}, \quad \forall\, i > j \geq 1. \tag{5.3.1}$$

$$x^{r+1}_{\mathcal{J}^i} = x^{r,N}_{\mathcal{J}^i} = x^{r,j}_{\mathcal{J}^i}, \quad \forall\, i \leq j \leq N. \tag{5.3.2}$$

$$x^{r+1}_{\mathcal{J}^i} - x^r_{\mathcal{J}^i} = x^{r,N}_{\mathcal{J}^i} - x^{r,0}_{\mathcal{J}^i} = d^{r,i}_{\mathcal{J}^i}. \tag{5.3.3}$$

5.3.2 Technical results

In this subsection, we give a necessary Lipschitz continuity-like assumption and present several useful lemmas. Among these, Lemma 5.3.6 plays a key role in the final results.

Assumption 5.3.1. *Let $\{\mathcal{J}^i,\ i = 1, \ldots, N\}$ be a partition of the set $\mathcal{N} = \{1, \ldots, n\}$. The gradient ∇f is block lower triangular Lipschitz continuous with respect to blocks $\{\mathcal{J}^i,\ i = 1, 2, \ldots, N\}$, which is formally defined by Definition 2.2.4 in Section 2.2.2.*

The relations between constants M and L_f are discussed in Section 5.4.

Throughout this chapter, we denote

$$L_{\min} := \min_{i=1,\ldots,N} \widehat{L}_i, \tag{5.3.4}$$

$$L_{\max} := \max_{i=1,\ldots,N} \widehat{L}_i, \tag{5.3.5}$$

where $\{L_i,\ i = 1, \ldots, N\}$ are block Lipschitz constants, given in Algorithm 5-1.

Moreover, we let $g^r \in \mathcal{R}^n$ with

$$g^r_{\mathcal{J}^i} := \nabla_{\mathcal{J}^i} f(x^{r,i-1}),\ \ i = 1, \ldots, N, \tag{5.3.6}$$

and define

$$v^r := \operatorname*{argmin}_{d\in\mathcal{R}^n} \left\{ \langle g^r, d\rangle + \frac{1}{2}\sum_{k=1}^{N} \widehat{L_k} \|d_{\mathcal{J}^k}\|^2 + \tau\psi(x^r + d) \right\}, \tag{5.3.7}$$

$$u^r := \operatorname*{argmin}_{d\in\mathcal{R}^n} \left\{ \langle \nabla f(x^r), d\rangle + \frac{1}{2}\sum_{k=1}^{N} \widehat{L_k} \|d_{\mathcal{J}^k}\|^2 + \tau\psi(x^r + d) \right\}. \tag{5.3.8}$$

It then follows from Algorithm 5-1 and (5.3.3) that

$$v^r = x^{r+1} - x^r. \tag{5.3.9}$$

Since $g^r = g(x^r, x^{r+1})$ from (2.2.3) and (5.3.6), under Assumption 5.3.1, we have that

$$\|g^r - \nabla f(x^r)\| \le M\|x^{r+1} - x^r\| = M\|v^r\|. \tag{5.3.10}$$

Next, we define an approximation of F at point x^r with direction u^r by

$$Q_F(x^r, u^r) := f(x^r) + \langle \nabla f(x^r), u^r\rangle + \frac{1}{2}\sum_{k=1}^{N} \widehat{L_k} \|u^r_{\mathcal{J}^k}\|^2 + \tau\psi(x^r + u^r). \tag{5.3.11}$$

The following lemma shows a relation between $F(x^{r+1})$ and its approximation $Q_F(x^r, u^r)$.

Lemma 5.3.2. *Suppose that Assumption 5.2.1 holds. Then, we have*

$$F(x^{r+1}) \le Q_F(x^r, u^r) + \langle g^r - \nabla f(x^r), u^r\rangle, \ \ \forall r \ge 0. \tag{5.3.12}$$

Proof. For any $r \geq 0$ and $i \in \{1, \ldots, N\}$, we obtain

$$\begin{aligned} F(x^{r,i}) - F(x^{r,i-1}) &= f(x^{r,i}) - f(x^{r,i-1}) + \tau\psi(x^{r,i}) - \tau\psi(x^{r,i-1}) \\ &\leq \langle \nabla_{\mathcal{J}^i} f(x^{r,i-1}), d^{r,i}_{\mathcal{J}^i} \rangle + \frac{\widehat{L}_i}{2} \|d^{r,i}_{\mathcal{J}^i}\|^2 + \tau\psi_{\mathcal{J}^i}(x^{r,i-1}_{\mathcal{J}^i} + d^{r,i}_{\mathcal{J}^i}) - \tau\psi_{\mathcal{J}^i}(x^{r,i-1}_{\mathcal{J}^i}) \\ &\leq \langle \nabla_{\mathcal{J}^i} f(x^{r,i-1}), u^r_{\mathcal{J}^i} \rangle + \frac{\widehat{L}_i}{2} \|u^r_{\mathcal{J}^i}\|^2 + \tau\psi_{\mathcal{J}^i}(x^{r,0}_{\mathcal{J}^i} + u^r_{\mathcal{J}^i}) - \tau\psi_{\mathcal{J}^i}(x^{r,0}_{\mathcal{J}^i}), \end{aligned}$$

where the first inequality follows from Lemma 2.2.2 (i), and the second inequality holds since $d^{r,i}$ is a solution of its subproblem.

Summing over $i = 1, \ldots, N$, we get

$$F(x^{r,N}) - F(x^{r,0}) \leq \langle g^r, u^r \rangle + \frac{1}{2}\sum_{k=1}^{N} \widehat{L_k} \|u^r_{\mathcal{J}^k}\|^2 + \tau\psi(x^{r,0} + u^r) - \tau\psi(x^{r,0}).$$

It then follows from (5.3.1) and (5.3.2) that

$$F(x^{r+1}) \leq f(x^r) + \langle g^r, u^r \rangle + \frac{1}{2}\sum_{k=1}^{N} \widehat{L_k} \|u^r_{\mathcal{J}^k}\|^2 + \tau\psi(x^r + u^r),$$

which, together with the definition of $Q_F(x^r, u^r)$ in (5.3.11), leads to (5.3.12). ■

The next lemma helps us to investigate the relation between the direction v^r given by the BCGD method and the direction u^r in the proximal gradient method [66].

Lemma 5.3.3. *For any $r \geq 0$ and $a \in \mathcal{R}^n$, let*

$$d^r_a = \operatorname*{argmin}_d \left\{ \langle a, d \rangle + \frac{1}{2}\sum_{k=1}^{N} \widehat{L_k} \|d_{\mathcal{J}^k}\|^2 + \tau\psi(x^r + d) \right\}.$$

Then, for any $a, b \in \mathcal{R}^n$, we have

$$\|d^r_a - d^r_b\| \leq \frac{1}{L_{\min}} \|a - b\|.$$

Proof. This result follows immediately from [69, lemma 4] with $h(u) = \frac{1}{2}\sum_{k=1}^{N} \widehat{L_k} \|u_{\mathcal{J}^k}\|^2$, $p = q = 2$, and $\rho = L_{\min}$. ■

Letting $a = g^r$ and $b = \nabla f(x^r)$ in Lemma 5.3.3, we have that

$$\|v^r - u^r\| \leq \frac{1}{L_{\min}} \|g^r - \nabla f(x^r)\|. \tag{5.3.13}$$

Then, an upper bound for $\langle g^r - \nabla f(x^r), u^r \rangle$, which appears in (5.3.12), can be established.

Lemma 5.3.4. *Suppose that Assumptions 5.2.1 and 5.3.1 hold. Then, for any $r \geq 0$, we have*

$$\langle g^r - \nabla f(x^r), u^r \rangle \leq M \left(\frac{M}{L_{\min}} + 1 \right) \|v^r\|^2,$$

where M is the constant given in Assumption 5.3.1.

Proof. For any $r \geq 0$, we have

$$\langle g^r - \nabla f(x^r), u^r \rangle \leq \|g^r - \nabla f(x^r)\| \|u^r\| \leq \|g^r - \nabla f(x^r)\| (\|u^r - v^r\| + \|v^r\|).$$

Combining this with (5.3.10) and (5.3.13), we obtain the desired inequality. ■

Lemma 5.3.5. *Suppose that Assumption 5.2.1 holds. Then, for any $r \geq 0$, we have*

$$F(x^r) - F(x^{r+1}) \geq \frac{1}{2} L_{\min} \|v^r\|^2. \tag{5.3.14}$$

Proof. Using Lemma 2.3.1 with $\varphi(x) = \langle \nabla f(x^{r,i-1}), x - x^{r,i-1} \rangle + \tau\psi(x)$, $B_\eta(x,z) = \frac{1}{2}\sum_{k=1}^N \widehat{L_k} \|x_{\mathcal{J}^k} - z_{\mathcal{J}^k}\|^2$, $z_+ = x^{r,i}$, and $z = x^{r,i-1}$, we have that, for any $x \in \operatorname{dom} F$ and $i \in \{1, \ldots, N\}$,

$$\begin{aligned}
&\langle \nabla f(x^{r,i-1}), x - x^{r,i-1} \rangle + \tau\psi(x) + \frac{1}{2}\sum_{k=1}^N \widehat{L_k} \|x_{\mathcal{J}^k} - x^{r,i-1}_{\mathcal{J}^k}\|^2 \\
&\geq \langle \nabla f(x^{r,i-1}), x^{r,i} - x^{r,i-1} \rangle + \tau\psi(x^{r,i}) + \frac{1}{2}\sum_{k=1}^N \widehat{L_k} \| x^{r,i}_{\mathcal{J}^k} - x^{r,i-1}_{\mathcal{J}^k}\|^2 + \frac{1}{2}\sum_{k=1}^N \widehat{L_k} \| x_{\mathcal{J}^k} - x^{r,i}_{\mathcal{J}^k}\|^2.
\end{aligned}$$

Setting $x = x^{r,i-1}$, we obtain

$$\tau\psi(x^{r,i-1}) \geq \langle \nabla_{\mathcal{J}^i} f(x^{r,i-1}), d^{r,i}_{\mathcal{J}^i} \rangle + \tau\psi(x^{r,i}) + \widehat{L}_i \|d^{r,i}_{\mathcal{J}^i}\|^2, \tag{5.3.15}$$

which, together with Lemma 2.2.2 (i), implies that

$$\begin{aligned}
F(x^{r,i}) &\leq f(x^{r,i-1}) + \langle \nabla_{\mathcal{J}^i} f(x^{r,i-1}), d^{r,i}_{\mathcal{J}^i} \rangle + \frac{\widehat{L}_i}{2} \|d^{r,i}_{\mathcal{J}^i}\|^2 + \tau\psi(x^{r,i}) \\
&\leq F(x^{r,i-1}) - \frac{\widehat{L}_i}{2} \|d^{r,i}_{\mathcal{J}^i}\|^2.
\end{aligned}$$

Summing over $i = 1, \ldots, N$, we obtain

$$F(x^{r+1}) \leq F(x^r) - \frac{1}{2}\sum_{k=1}^N \widehat{L_k} \|x^{r+1}_{\mathcal{J}^k} - x^r_{\mathcal{J}^k}\|^2.$$

Using (5.3.9), we have the desired result. ■

It is shown in [69, Lemma 2] that x is a stationary point of problem (5.1.1) if and only if the direction $u^r = 0$ in (5.3.8) with $x^r = x$. The following remark describes the convergence rate of $\{\|u^r\|\}$ to zero, where function f is not necessarily convex.

Remark 5.3.1. *Suppose that Assumptions 5.2.1 and 5.3.1 hold. Then, for any $r \geq 0$, we have*

$$\min_{k=0,\ldots,r} \|u^k\| \leq \frac{1}{\sqrt{r+1}} \sqrt{\frac{2(F(x^0) - F^*)}{L_{\min}}} \left(\frac{M}{L_{\min}} + 1 \right), \tag{5.3.16}$$

where F^* is the optimal value of problem (5.1.1).

Proof. From (5.3.10) and (5.3.13), we have

$$\|u^r\| \leq \|u^r - v^r\| + \|v^r\| \leq \frac{1}{L_{\min}} \|g^r - \nabla f(x^r)\| + \|v^r\| \leq \left(\frac{M}{L_{\min}} + 1 \right) \|v^r\|. \tag{5.3.17}$$

Moreover, from Lemma 5.3.5, we get

$$F(x^0) - F^* \geq F(x^0) - F(x^{r+1}) \geq \frac{1}{2}(r+1) L_{\min} \min_{k=0,\ldots,r} \|v^k\|^2,$$

which, together with (5.3.17), gives the desired inequality. ■

Next, we show the key results of this chapter.

Lemma 5.3.6. *Suppose that Assumptions 5.2.1 and 5.3.1 hold. Then, for any $r \geq 0$, we have*

$$F(x^{r+1}) \leq Q_F(x^r, u^r) + \omega_1 (F(x^r) - F(x^{r+1})), \tag{5.3.18}$$

where $\omega_1 = \frac{2M}{L_{\min}} (\frac{M}{L_{\min}} + 1)$.

Proof. From Lemmas 5.3.4 and 5.3.5, we have

$$\langle g^r - \nabla f(x^r), u^r \rangle \leq \omega_1 (F(x^r) - F(x^{r+1})),$$

which, together with Lemma 5.3.2, gives the desired inequality. ■

Since $\{F(x^r)\}$ is nonincreasing according to Lemma 5.3.5, it follows from Assumption 5.2.2 that the distance $R_r := \|x^r - x^*\|$ is bounded for any $x^* \in X^*$ and any $r > 0$. Let

$$\Delta^r := F(x^r) - F^*,$$

where F^* is the global minimum. The following lemma presents an estimate for $Q_F(x^r, u^r) - F^*$, which is also shown in [59, Lemma 4, Lemma 6]. For simplicity, we omit its proof here.

Lemma 5.3.7. *Suppose that Assumption 5.2.2 holds. Then, the following statements hold.*

(i) *If function f is convex, then for any $r \geq 0$, we have*

$$Q_F(x^r, u^r) - F^* \leq \begin{cases} \left(1 - \frac{\Delta^r}{2L_{\max}R_r^2}\right)\Delta^r, & \text{if } \Delta^r \leq L_{\max}R_r^2, \\ \frac{1}{2}L_{\max}R_r^2 < \frac{1}{2}\Delta^r, & \text{otherwise.} \end{cases}$$

(ii) *If $\mu_f + \tau\mu_\psi > 0$, where μ_f and μ_ψ are strongly convex parameters of functions f and ψ, respectively, then for any $r \geq 0$, we have*

$$Q_F(x^r, u^r) - F^* \leq \frac{L_{\max} - \mu_f}{L_{\max} + \tau\mu_\psi}\Delta^r.$$

5.3.3 Iteration complexity analysis

Theorem 5.3.1. *Suppose that Assumptions 5.2.1, 5.2.2, and 5.3.1 hold. Then, we have*

$$F(x^r) - F^* \leq \varepsilon \quad \text{whenever} \quad r \geq \frac{\zeta(1+\omega_1)}{\varepsilon},$$

where $\zeta = 2\max\{L_{\max}R_*^2, \Delta^0\}$, $\omega_1 = \frac{2M}{L_{\min}}(\frac{M}{L_{\min}} + 1)$.

Proof. From Lemma 5.3.7 (i), we have

$$Q_F(x^r, u^r) - F^* \leq \max\left\{\frac{1}{2}, 1 - \frac{\Delta^r}{2L_{\max}R_r^2}\right\}\Delta^r. \tag{5.3.19}$$

From Lemma 5.3.5, $\Delta^0 \geq \Delta^1 \geq \cdots \geq \Delta^r > 0$. It can then be verified that

$$\max\left\{\frac{1}{2}, 1 - \frac{\Delta^r}{2L_{\max}R_r^2}\right\} \leq 1 - \frac{\Delta^r}{\zeta}. \tag{5.3.20}$$

Combining (5.3.19), (5.3.20), and Lemma 5.3.6 yields

$$\Delta^{r+1} \leq (1 - \frac{\Delta^r}{\zeta})\Delta^r + \omega_1(\Delta^r - \Delta^{r+1}).$$

Simplifying this, we have

$$\Delta^{r+1} \leq \Delta^r - \frac{(\Delta^r)^2}{\zeta(1+\omega_1)}.$$

Dividing both sides by $\Delta^r\Delta^{r+1}$, we get

$$\frac{1}{\Delta^r} \leq \frac{1}{\Delta^{r+1}} - \frac{1}{\zeta(1+\omega_1)}\frac{\Delta^r}{\Delta^{r+1}}. \tag{5.3.21}$$

Using the fact that $\Delta^r \geq \Delta^{r+1} > 0$, we obtain $\frac{\Delta^r}{\Delta^{r+1}} \geq 1$, which, together with (5.3.21), implies

$$\frac{1}{\Delta^{r+1}} - \frac{1}{\Delta^r} \geq \frac{1}{\zeta(1+\omega_1)}.$$

Summing over r, we get

$$\frac{1}{\Delta^r} - \frac{1}{\Delta^0} \geq \frac{r}{\zeta(1+\omega_1)}.$$

Since $\Delta^0 > 0$, we have $F(x^r) - F^* = \Delta^r \leq \frac{\zeta(1+\omega_1)}{r}$. Hence, the result follows. ■

We can deduce an immediate consequence of Theorem 5.3.1 using special settings for $\widehat{L_k}$.

Corollary 5.3.1. *Suppose that Assumptions 5.2.1, 5.2.2, and 5.3.1 hold. If we set* $\widehat{L_k} = \max\{M, L_f\}$ *for* $k = 1, \ldots, N$, *then* $F(x^r) - F^* \leq \varepsilon$ *whenever* $r \geq \frac{10\max\{MR_*^2, L_f R_*^2, \Delta^0\}}{\varepsilon}$.

Proof. The proof follows directly from $L_{\min} = L_{\max} = \max\{M, L_f\}$ and Theorem 5.3.1. ■

Ignoring the constant Δ^0 in Corollary 5.3.1, we can see that Algorithm 5-1 has the $O(\frac{\max\{M,L_f\}}{\varepsilon})$ iteration complexity. The next theorem shows that Algorithm 5-1 converges linearly when the cost function F is strongly convex.

Theorem 5.3.2. *Suppose that Assumptions 5.2.1, 5.2.2, and 5.3.1 hold, and that* $\mu_f + \tau\mu_\psi > 0$. *Then,*

$$F(x^r) - F^* \leq \left(\frac{\omega_1 + \omega_2}{1+\omega_1}\right)^r (F(x^0) - F^*), \tag{5.3.22}$$

where $\omega_1 = \frac{2M}{L_{\min}}(\frac{M}{L_{\min}} + 1)$ *and* $\omega_2 = \frac{L_{\max} - \mu_f}{L_{\max} + \tau\mu_\psi} < 1$.

Proof. From Lemmas 5.3.6 and 5.3.7 (ii), we obtain

$$\Delta^{r+1} \leq \omega_2 \Delta^r + \omega_1(\Delta^r - \Delta^{r+1}).$$

Hence, we have

$$\Delta^r \leq \frac{\omega_2 + \omega_1}{1+\omega_1}\Delta^{r-1} \leq \left(\frac{\omega_2+\omega_1}{1+\omega_1}\right)^r \Delta^0,$$

which proves the desired inequality. ■

A direct result of Theorems 5.3.1 and 5.3.2 is that $\{F(x^r)\}$ has a linear convergence rate if either f or ψ is strongly convex.

5.4 Relations between M and L_f

In this section, we study the relation between constants M and L_f, which, together with Corollary 5.3.1, yields that the iteration complexity deduced in this chapter is sharper than those in [7, 62, 32]. Particularly, we first prove that $M \leq \sqrt{N} L_f$ for general functions. Then, we list some special functions for which $M \leq 2L_f$.

Theorem 5.4.1. *Suppose that Assumption 5.2.1 holds. Then, we have $M \le \sqrt{N} L_f$, where M is the constant in Assumption 5.3.1, and L_f is the Lipschitz constant of ∇f.*

Proof. From Assumption 5.3.1, we have

$$\|g(x,y) - \nabla f(x)\| = \sqrt{\sum_{i=1}^{N} \|g_{\mathcal{J}^i}(x,y) - \nabla_{\mathcal{J}^i} f(x)\|^2} \le \sqrt{\sum_{i=1}^{N} \|\nabla f(z^i) - \nabla f(x)\|^2},$$

where $z^i \in \mathcal{R}^n$ with $z^i = (y_{\mathcal{J}^1}^T, \dots, y_{\mathcal{J}^{i-1}}^T, x_{\mathcal{J}^i}^T, \dots, x_{\mathcal{J}^N}^T)^T$.

Using Lemma 2.2.2 (ii), we obtain

$$\|\nabla f(z^i) - \nabla f(x)\| \le L_f \|z^i - x\| \le L_f \|y - x\|.$$

Then, we have that

$$\|g(x,y) - \nabla f(x)\| \le \sqrt{N} L_f \|y - x\|.$$

Hence, we have $M \le \sqrt{N} L_f$. ■

Remark 5.4.1. *From Corollary 5.3.1 and Theorem 5.4.1, we have that the iteration complexity of Algorithm 5-1 actually is $O(\frac{\sqrt{N}L_f}{\varepsilon})$ for the nonsmooth minimization problem (5.1.1). This bound is sharper than the results in [7, 32].*

Next, we will give several functions such that $M \le 2L_f$. We let $L_f = \max_{x \in \mathcal{R}^n} \|\nabla^2 f(x)\|$ and show the proofs for Theorems 5.4.2-5.4.3 only for the case $N = n$. The proof for the block case, i.e., $N < n$, can be deduced in a similar way.

Assume that $\nabla^2 f(x)$ is decomposed into a strictly lower triangular matrix $P(x)$, diagonal matrix $\Lambda(x)$, and upper triangular matrices $(P(x))^T$, i.e.,

$$\nabla^2 f(x) = P(x) + \Lambda(x) + (P(x))^T.$$

Moreover, let $y \in \mathcal{R}^n$ and let $z^i \in \mathcal{R}^n$ with $z^i = (y_1, \dots, y_{i-1}, x_i, \dots, x_n)^T$. Then, for any $r \ge 0$ and $i \in \{1, \dots, n\}$, we have

$$\begin{aligned} g_i(x,y) - \nabla_i f(x) =& \nabla_i f(z^i) - \nabla_i f(x) \\ =& \int_0^1 \sum_{j=1}^{n} \nabla_{i,j}^2 f(x + \tau(z^i - x))(z_j^i - x_j) d\tau \\ =& \int_0^1 \sum_{j=1}^{n} P_{i,j}(x + \tau(z^i - x))(y_j - x_j) d\tau, \end{aligned}$$

where the last equality follows from the fact that $x_k = z_k^i$ for any $k \ge i$.

Letting $\hat{P}(\tau, x, z) \in \mathcal{R}^{n\times n}$ with

$$\hat{P}_{i,j}(\tau, x, z) = P_{i,j}(x + \tau(z^i - x)), \forall i, j \in \{1, 2, \ldots, n\}, \tag{5.4.1}$$

we obtain

$$g(x, y) - \nabla f(x) = \int_0^1 \hat{P}(\tau, x, z)(y - x)d\tau,$$

which implies that

$$\|g(x, y) - \nabla f(x)\| \le \int_0^1 \|\hat{P}(\tau, x, z)\|\|y - x\|d\tau \le \max_{\tau\in[0,1]}\|\hat{P}(\tau, x, z)\|\|y - x\|. \tag{5.4.2}$$

Hence, we obtain

$$M \le \max_{\tau\in[0,1]}\|\hat{P}(\tau, x, z)\|. \tag{5.4.3}$$

Theorem 5.4.2. *Suppose that f is twice differentiable, and that $\nabla^2 f(x)$ is a tridiagonal matrix. Then, we have $M \le L_f$.*

Proof. Since $\nabla^2 f(x)$ is assumed to be tridiagonal, we have

$$\|\hat{P}(\tau, x, z)\| = \max_{i=1,\ldots,n-1}|P_{i+1,i}(\tau, x, z)| \le \max_{\substack{i=1,\ldots,n-1\\ x\in\mathcal{R}^n}} |P_{i+1,i}(x)|. \tag{5.4.4}$$

Using the property that $\|A\| \ge \max_{i,j}|A_{i,j}|$ for any $A \in \mathcal{R}^{n\times n}$, we have $L_f = \max_{x\in\mathcal{R}^n}\|\nabla^2 f(x)\| \ge \max_{\substack{i=1,\ldots,n-1\\ x\in\mathcal{R}^n}} |P_{i+1,i}(x)|$. Combined with (5.4.3) and (5.4.4), this proves the desired inequality. ■

Theorem 5.4.3. *Suppose that, for any $i = 1, \ldots, n$ and $x \in \mathcal{R}^n$, there exists a submatrix $E^i_{sub}(x)$ of $\nabla^2 f(x)$ such that $\|E^i_{sub}(x)\| \ge \sum_{j=1}^{i-1} |P_{i,j}(x)|$ holds. Then, we have $M \le \sqrt{2}L_f$.*

Proof. For any $i = 1, \ldots, n$, we have

$$\|E^i_{sub}(x)\|^2 \ge \left(\sum_{j=1}^{i-1} |P_{i,j}(x)|\right)^2 \ge \sum_{j=1}^{i-1} |P_{i,j}(x)|^2. \tag{5.4.5}$$

Moreover, for the strictly lower triangular matrix $\hat{P}(\tau, x, z)$ defined by (5.4.1), we get for any $y \in \mathcal{R}^n$ that

$$\|\hat{P}(\tau, x, z)y\|^2 = \sum_{i=2}^{n} \left|\sum_{j=1}^{i-1} \hat{P}_{i,j}(\tau, x, z)y_i\right|^2 \le 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} |\hat{P}_{i,j}(\tau, x, z)|^2|y_i|^2.$$

Using the fact that the variables in the i-th row of matrix $\hat{P}(\tau, x, z)$ are the same, we have from (5.4.5) that

$$\|\hat{P}(\tau,x,z)y\|^2 \leq 2\sum_{i=2}^{n} \left\|E^i_{sub}(x+\tau(z^i-x))\right\|^2 |y_i|^2 \leq 2 \max_{\substack{i=1,\ldots,n-1\\ x\in\mathcal{R}^n}} \|E^i_{sub}(x)\|^2\|y\|^2.$$

Recalling the property that $\|A\| \geq \max_{i=1,\ldots,n} \{\|A^i_{sub}\|\}$ for any matrix $A \in \mathcal{R}^{n\times n}$, we have from (5.4.3) that

$$M \leq \|\hat{P}(\tau,x,z)\| = \max_{x\neq 0} \frac{\|\hat{P}(\tau,x,z)y\|}{\|x\|} \leq \sqrt{2} \max_{\substack{i=1,\ldots,n-1\\ x\in\mathcal{R}^n}} \|E^i_{sub}(x)\| \leq \sqrt{2}\max_{x\in\mathcal{R}^n}\|\nabla^2 f(x)\|,$$

which proves the desired inequality. ■

The following corollary follows immediately from Theorem 5.4.3.

Corollary 5.4.1. *Suppose that the Hessian matrix $\nabla^2 f(x)$ is row diagonally dominant, i.e., $|\nabla^2_{ii} f(x)| \geq \sum_{j\neq i} |\nabla^2_{ij} f(x)|$ for any $i = 1, \ldots, n$. Then, we have $M \leq \sqrt{2}L_f$.*

Next, we give a sufficient condition when f is quadratic. The similar condition is assumed in [62] to show the iteration complexity $O(\frac{L_f}{\varepsilon})$.

Theorem 5.4.4. *Suppose that f is a quadratic function with $f(x) = \frac{1}{2}x^T Ex$, $E \in \mathcal{R}^{n\times n}$ is symmetric and positive semidefinite, and all nonzero elements in $\{E_{ij},\ i,j = 1,\ldots,n,\ i \neq j\}$ have the same signs. Then, we have $M \leq 2L_f$.*

Proof. Throughout the proof, we let $L_f = \|E\|$ and let E be decomposed into a block diagonal matrix Λ^B and strictly lower and upper triangular matrices P^B and $(P^B)^T$, i.e.,

$$E = P^B + \Lambda^B + (P^B)^T,$$

where $\Lambda^B := \text{Diag}\{\Lambda^B_1, \ldots, \Lambda^B_N\}$, $\Lambda^B_i \in \mathcal{R}^{|\mathcal{J}^i|\times|\mathcal{J}^i|}$.

A simple computation gives that

$$g(x,y) - \nabla f(x) = P^B(y-x),$$

which implies that

$$M \leq \|P^B\|. \tag{5.4.6}$$

From the assumption on matrix E, it can be easily verified that

$$\|P^B\| = \max_{x\neq 0}\frac{\|P^B x\|}{\|x\|} = \max_{x\geq 0, x\neq 0}\frac{\|P^B x\|}{\|x\|} \leq \max_{x\geq 0, x\neq 0}\frac{\|(P^B+(P^B)^T)x\|}{\|x\|} = \|P^B + (P^B)^T\|,$$

which yields that

$$M \leq \|P^B\| \leq \|P^B + (P^B)^T\| = \|E - \Lambda^B\| \leq \|E\| + \|\Lambda^B\|. \tag{5.4.7}$$

Thus, we need only show that

$$\|\Lambda^B\| = \max_{i=1,\ldots,N}\{\|\Lambda^B_{\mathcal{J}^i}\|\} \leq \|E\|. \tag{5.4.8}$$

Then, by (5.4.7), we can conclude that the theorem holds. In fact, we have

$$\|\Lambda^B x\|^2 = \sum_{i=1}^{N} \|\Lambda^B_{\mathcal{J}^i} x_{\mathcal{J}^i}\|^2 \leq \sum_{i=1}^{N} \|\Lambda^B_{\mathcal{J}^i}\|^2 \|x_{\mathcal{J}^i}\|^2 \leq \max_{i=1,\ldots,N}\{\|\Lambda^B_{\mathcal{J}^i}\|^2\}\|x\|^2.$$

Then, we get

$$\|\Lambda^B\| = \max_{x\neq 0}\frac{\|\Lambda^B x\|}{\|x\|} \leq \max_{i=1,\ldots,N}\{\|\Lambda^B_{\mathcal{J}^i}\|\}.$$

On the other hand, using the property that $\|\Lambda^B\| \geq \|\Lambda^B_{\mathcal{J}^i}\|$ for any $i = 1, \ldots, N$, we have

$$\|\Lambda^B\| \geq \max_{i=1,\ldots,N}\{\|\Lambda^B_{\mathcal{J}^i}\|\}.$$

Hence, we obtain $\|\Lambda^B\| = \max_{i=1,\ldots,N}\{\|\Lambda^B_{\mathcal{J}^i}\|\}$. Moreover, $\max_{i=1,\ldots,N}\|\Lambda^B_i\| \leq \|E\|$ holds because $\|\Lambda^B_{\mathcal{J}^i}\| \leq \|E\|$ for any $i \in \{1, 2, \ldots, N\}$. ■

Remark 5.4.2. *Functions in Theorems 5.4.2–5.4.4 satisfies Assumption 5.3.1 with $M \leq 2L_f$. Combined with Corollary 5.3.1, this yields that the iteration complexity of Algorithm 5-1 can be improved to $O(\frac{L_f}{\varepsilon})$. This bound is independent of the number of blocks.*

Remark 5.4.3. *As a comparison to [62], matrix E with $E_{ij} \leq 0$ for any $j \neq i$ in Theorem 5.4.4 meets the isotonicity assumption in [62]. However, the results in this chapter applies not only to l_1-regularized loss minimization problems, but also to much more optimization problems. Another favorable point is that Theorem 5.4.4 does not require a special initial point as in [62].*

5.5 Conclusion

In this chapter, we have studied the iteration complexity of the BCGD method with a cyclic rule for solving nonsmooth convex optimization problem (5.1.1). We have proposed a new Lipschitz continuity-like assumption, and improved the iteration complexity to $O(\frac{\max\{M, L_f\}}{\varepsilon})$, where M is the constant given in the proposed assumption. Furthermore, we have studied the relation between M and L_f. Theorems 5.4.1–5.4.4 show that $M \leq \sqrt{N}L_f$ or $M \leq 2L_f$, and this implies that the iteration complexity bound derived in this chapter is sharper than existing results (see Table 5.1 for details).

Table 5.1: Comparison of BCD methods with the cyclic rule for convex problem (5.1.1)

Algorithm	$\psi(x)$	$\widehat{L_k}$	Complexity bound	Complexity
Algorithm 5-1 (in this chapter)	Separable	$\max\{M, L_f\}$	$\frac{10\max\{MR_*^2, L_fR_*^2, \Delta^0\}}{\varepsilon}$	$O(\frac{\sqrt{N}L_f}{\varepsilon})$ $O(\frac{L_f}{\varepsilon})$ for problems in Th. 5.4.2–5.4.4
Beck et al. [7]	0	L_f	$\frac{4L_f(1+N)R_*^2}{\varepsilon} - \frac{8}{N}$	$O(\frac{NL_f}{\varepsilon})$
Hong et al. [32]	Separable	L_f	$\frac{8tNL_fR_*^2}{\varepsilon}$	$O(\frac{NL_f}{\varepsilon})$

In future work, it would be interesting to find more functions for which the corresponding constant M is independent of the number N of blocks. Currently, we have not found a counter example where $N^\sigma L \leq M$ for a positive constant σ.

Additionally, Assumption 5.2.2 in this chapter seems a little strict. Recently, for the BCGD method with the random rule, some iteration complexity results have been established without Assumption 5.2.2 [39, 40]. In the future, it would be challenging to study the iteration complexity of the BCGD method with the cyclic rule without Assumption 5.2.2.

Chapter 6

Regret analysis of a block coordinate gradient method for online convex optimization problem

6.1 Introduction

In this chapter, we consider an online convex optimization problem with a separable structure, whose loss function $F^t : \Omega \to \mathcal{R}$ at time step t is given as follows.

$$F^t(x) := f^t(x) + \tau\psi(x), \quad t = 1, 2, \ldots, \tag{6.1.1}$$

where $f^t : \Omega \to \mathcal{R}$ is smooth and convex, $\Omega \subseteq \bigcap_{t=1}^{\infty} \operatorname{dom} F^t$ is a nonempty convex set, τ is a positive constant, and $\psi : \Omega \to (-\infty, \infty]$ is a proper, convex and l.s.c. function with the block separable structure.

As described in Subsection 1.2.2, it is impossible to select a point x^t that exactly minimize the loss function $F^t(x)$ at the t-th time step for the online optimization problems, and the goal of the online convex optimization problem is to propose an algorithm, with which the generating decisions make us to achieve a regret as low as possible. The definition of the "regret" is formally given by Definition 2.2.8 in Subsection 2.2.4.

Moreover, as mentioned in Subsection 1.2.3, the applications of the online optimization problems [3, 14, 27] are mostly built on large scales. Some researchers have studied the performances of the gradient methods for the online convex optimization problems [73, 81]. When $\psi(x)$ in (6.1.1) is an indicator function, Zinkevich [81] proved that the projected gradient method for the online convex optimization problem has a regret $O(\sqrt{T})$. When $\psi(x)$ in (6.1.1) is a general regularization function, Xiao [73] proposed a dual averaging method, which is first proposed by Nesterov for classical convex optimization problem. He showed that the proposed method achieves the same regret $O(\sqrt{T})$ as [81]. However, both of these two methods are full gradient methods, i.e., they update all components of the variable x at each iteration. When the scale of the problem becomes very large, the evaluations for updating the gradients of each iteration would take much time.

Recently, the "block" type methods are becoming very popular, especially for the large scale problems [59, 66, 67]. Compared to the full gradient methods, the block type methods can reduce the calculation time at each iteration. Quite recently, Xu and Yin [75] proposed a block coordinate stochastic gradient method with the cyclic rule for a regularized stochastic optimization problem, which is related to the online optimization problem (6.1.1). Under the Lipschitz continuity-like assumption, they showed that the proposed method converges with $O(\frac{1+\log T}{\sqrt{1+T}}N)$, where N is the number of blocks. Furthermore, as the number N of the blocks reduces to 1, i.e., $N = 1$, this iteration compklexity bound reduces to $O(\frac{1+\log T}{\sqrt{1+T}})$, which is bigger than the average regret $\frac{R(T)}{T} = O(\frac{1}{\sqrt{T}})$ of the greedy projection method [81].

In this chapter, we propose a block coordinate gradient method with the cyclic rule (C-BCG) for the online convex optimization problem with the loss function (6.1.1). For the proposed methods, we make our research on the following two aspects.

- We establish its regret bound. In particular, we show that the C-BCG method has a regret $O(\sqrt{T})$. See Theorem 6.4.1 for details.

- We extend the C-BCG method to the convex stochastic optimization problem. See Theorem 6.5.1 for details.

Note that the regret bound of the C-BCG method in this chapter is independent of the number N of blocks under proper assumptions, although we solve N subproblems at each step. When the total number of blocks reduces to one, and the function ψ is set to be an indicator function, the regret of the proposed method reduce to the same result in [81]. Hence, it is a natural extension of greedy projection method [81].

Additionally, although the C-BCG method proposed in this chapter is essentially same as the block coordinate stochastic gradient method with the cyclic rule [75], by different analysis, we show that the ergodic convergence upper bound of the C-BCG method is tighter than that in [75].

This chapter is organized as follows. In Section 6.2, we introduce the algorithms of the block coordinate gradient methods with the cyclic rule. Then we introduce some basic assumptions and present revelent properties in Section 6.3. In Section 6.4, we investigate the regret analysis for the proposed method. Finally, we conclude this chapter in Section 6.5.

6.2 The BCG method

In this chapter, we propose a block coordinate gradient method with the cyclic rule for the online convex optimization problem (6.1.1). For convenience, we start with recalling several important results in [81].

The greedy projection method proposed in [81], can be described as follows.

Greedy projection method:
Step 0: Choose an initial point $x^1 \in \Omega$ and set a sequence of constants $\lambda_1, \lambda_2, ... > 0$.
Step 1: Update the vector x^t according to

$$\begin{aligned} x^{t+1} &= x^t + d^t\,, \\ d^t &= P_\Omega(x^t - \lambda_t \nabla f^t(x^t)) - x^t, \end{aligned}$$

where $P_\Omega(\cdot)$ denotes a projection onto the set Ω. Constants $\{\lambda_t,\, t = 1, 2, \dots\}$ are called the "learning rates".

Before proposing the block coordinate gradient algorithms, we present several basic assumptions. Throughout this chapter, the variable x is assumed to be partitioned into N blocks, denoted by $x^T = (x^T_{\mathcal{J}^1}, \dots, x^T_{\mathcal{J}^N})$.

We also assume that ψ is block separable with respect to each block $\mathcal{J}^i, i = \{1, 2, \dots, n\}$. For the set Ω in (6.1.1), we suppose that $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_N$, where operator "×" denotes the Cartesian product. For given ∇f^t at time step t, we consider the direction, which is defined by

$$d(x; J, t, \lambda) = \operatorname*{argmin}_{d \in \mathcal{R}^n} \left\{ \langle \nabla f^t(x), d\rangle + \frac{1}{2\lambda}\|d\|^2 + \tau\psi(x + d) \middle|\, d_{\bar{J}} = 0, x + d \in \Omega \right\}. \quad (6.2.1)$$

As defined in the greedy projection method [81], constant λ in (6.2.1) is called the learning rate. When $J = \{1, 2, \dots, n\}$ and $\psi = 0$, $d(x; J, t, \lambda)$ reduces to the direction d^t in Step 1 of the greedy projection method. Note that direction $d(x; J, t, \lambda)$ given by (6.2.1) is well defined, since the minimizer of the corresponding optimization problem always exists and is unique [60, Theorem 31.5]. Moreover, the direction $d(x; J, t, \lambda)$ is a descent direction for the loss function F^t.

The rule to choose blocks is also important for convergence. In this chapter, we choose blocks in the order of $\mathcal{J}^1, \mathcal{J}^2, \dots, \mathcal{J}^N$ cyclically, i.e., we adopt the cyclic rule, which is precisely defined in Section 2.4.

Next, we describe the frameworks of the BCG methods with the cyclic rule.

```
Algorithm 6-1. A BCG method with the cyclic rule (C-BCG):
Step 0: Choose an initial point $x^1 \in \operatorname{int}\Omega$. Let $t = 1$.
Step 1: If some termination condition holds, then stop.
Step 2-0: Let $x^{t,0} = x^t$ and $i = 1$.
Step 2-1: Set the learning rate $\lambda_{t,i} \in (0, +\infty)$. Solve the subproblem (6.2.1) with $x = x^{t,i-1}$,
    $J = \mathcal{J}^i$, $\lambda = \lambda_{t,i}$ and get a direction $d^{t,i} = d(x^{t,i-1}; \mathcal{J}^i, t, \lambda_{t,i})$.
Step 2-2: Set $x^{t,i} = x^{t,i-1} + d^{t,i}$ and $i = i + 1$. If $i = N + 1$, then go to Step 3. Otherwise,
    go to Step 2-1.
Step 3: Set $x^{t+1} = x^{t,N}$. Let $t = t + 1$ and go to Step 1.
```

At each time step, the C-BCG method solves subproblem (6.2.1) N times. When $N = 1$ and function ψ is an indicator function, this method reduces to the greedy projection method [81].

6.3 Basic assumptions

In this section, we introduce basic assumptions and present relevant properties, which will be used in the subsequent sections.

Given a constant $T > 0$, we denote the set of all optimal solutions of the problem $\min\limits_{x\in\Omega} \sum\limits_{t=1}^{T} F^t(x)$ by $X^{*,[T]}$ in the rest of this chapter.

For the loss functions f^t and ψ, we make the following assumptions, where Assumptions 6.3.1 and 6.3.2 are also used in [81].

Assumption 6.3.1. *The feasible set Ω for loss functions $\{F^t,\ t = 1, 2, \dots\}$ in (6.1.1) is nonempty and compact.*

For convenience, we define

$$D := \max_{x,y\in\Omega} \|x - y\|. \tag{6.3.1}$$

It follows from Assumption 6.3.1 that $D < \infty$.

Assumption 6.3.2. *There exists a positive constant G such that $\|\nabla f^t(x)\| \le G$ and $\|\partial\psi(x)\| \le G$ hold for all $t > 0$ and $x \in \Omega$.*

Note that when f^t is a linear function, that is, $f^t(x) = \langle a^t, x\rangle + b^t$ with some $a^t \in \mathcal{R}^n$, $b^t \in \mathcal{R}$, we have $\nabla f^t(x) = a^t$. Then, $\|\nabla f^t(x)\| \le G$ means that $\|a^t\| \le G$ holds for any $t > 0$. When f^t is a quadratic function with $f^t(x) = x^T A^t x$, we get $\nabla f^t(x) = A^t x$. From Assumption 6.3.1, $\|\nabla f^t(x)\| \le G$ is equivalent to that $\|A^t\| \le \frac{G}{D}$. For the function $\psi(x)$, when $\psi(x) = \|x\|$, we have $\|\partial\psi(x)\| \le 1$.

It is worth mentioning that Assumptions 6.3.1 and 6.3.2 are a little restrictive in this chapter. In fact, we only need to assume that there exists a compact set $\tilde{\Omega} \subseteq \mathcal{R}^n$ such that the iterations $\{x^t\} \subseteq \tilde{\Omega}$, $X^{*,[T]} \subseteq \tilde{\Omega}$, and that $\|\nabla f^t(x^t)\|$, $\|\partial\psi(x^t)\| \leq G$ for all $x^t \in \tilde{\Omega}$. For simplicity, we adopt Assumptions 6.3.1 and 6.3.2 in this chapter, which are in accord with the assumptions in [81].

The following assumption is a Lipschitz continuity-like assumption, which is defined in Subsection 2.2.2.

Assumption 6.3.3. *Let $\{\mathcal{J}^i,\ i = 1, \ldots, N\}$ be a partition of the set $\mathcal{N} = \{1, \ldots, n\}$. The gradient ∇f is block lower triangular Lipschitz continuous with respect to blocks $\{\mathcal{J}^i,\ i = 1, 2, \ldots, N\}$, which is formally defined by Definition 2.2.4 in Subsection 2.2.2.*

For the vector $g(x, y)$, defined by (2.2.3), in the following chapter, we use the notation g^t instead of $g^t(x^t, x^{t+1})$ when it is clear from the context. The next remark states several particular cases of constant M in Assumption 6.3.3.

Remark 6.3.1. *When $N = 1$ or function f^t is separable with respect to the blocks $\{\mathcal{J}^i,\ i = 1, \ldots, N\}$, we have that $g^t(x, y) = \nabla f^t(x)$, which yields that $M = 0$ in (2.2.2). When $N > 1$, it is shown in Section 5.4 that $M \leq 2\max\{L_{f^1}, \ldots, L_{f^N}\}$ holds for many classes of functions f^t. For example, when functions f^t have the same forms with $f(x)$, and the Hessian matrix $\nabla^2 f(x)$ is tridiagonal or row diagonal dominant, we have that $M \leq 2L_f$.*

Under Assumptions 6.3.1-6.3.3, we can show that the vector g^t is bounded for all t.

Lemma 6.3.1. *Suppose that Assumptions 6.3.1-6.3.3 hold. Then we have*

$$\|g^t\|^2 = \sum_{i=1}^{N} \|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2 \leq \bar{G}^2, \tag{6.3.2}$$

where $\bar{G} = MD + G$. Moreover, when $N = 1$ or function f^t is separable with respect to the blocks $\{\mathcal{J}^i,\ i = 1, \ldots, N\}$, we have $\|g^t\| \leq G$.

Proof. Since we denote $g^t = g^t(x^t, x^{t+1})$, from the definition of $g^t(x, y)$ in (2.2.3), we have that

$$\|g^t\|^2 = \sum_{i=1}^{N} \|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2.$$

Moreover, from Assumptions 6.3.1-6.3.3, we have that

$$\|g^t\| \leq \|g^t - \nabla f^t(x^t)\| + \|\nabla f^t(x^t)\| \leq M\|x^{t+1} - x^t\| + G \leq MD + G.$$

Hence, the relation (6.3.2) holds.

When $N = 1$ or function f^t is separable with respect to the blocks $\{\mathcal{J}^i,\ i = 1, \dots, N\}$, we have $\|g^t\| = \|\nabla f^t(x^t)\|$ from Remark 6.3.1, which together with Assumption 6.3.2 yields that $\|g^t\| \le G$. ■

For the learning rate λ in the proposed methods, we make the following assumption.

Assumption 6.3.4. *The learning rate $\lambda_{t,i}$ in the C-BCG method is given by $\lambda_{t,i} = \frac{c\beta_i}{\sqrt{t}}$, $t = 1, 2, \dots$, $i = 1, 2, \dots, N$, where $c > 0$, and $\beta_i \in [\underline{\beta}, \bar{\beta}]$, $\bar{\beta} \ge \underline{\beta} > 0$.*

The constants $\{\beta_i,\ i = 1, 2, \dots, N\}$ in Assumption 6.3.4 act as scaling factors of the learning rates on different blocks. When $\beta_i = 1$, $i = 1, 2, \dots, N$, and $c = 1$, we have $\lambda_{t,i} = \frac{1}{\sqrt{t}}$, which reduces to the case in [81].

Next, we recall the regret of the greedy projection method, which is given in [81].

Theorem 6.3.1. *Suppose that Assumptions 6.3.1-6.3.2 hold. Let $\lambda_t = \frac{1}{\sqrt{t}}$ and $x^{*,[T]} \in X^{*,[T]}$. Then, the regret $R(T)$ of the greedy projection method satisfies*

$$R(T) \le \frac{\sqrt{T}}{2} D^2 + \frac{(2\sqrt{T} - 1)}{2} G^2, \tag{6.3.3}$$

where constant D is defined by (6.3.1), and constant G is given in Assumption 6.3.2.

6.4 Regret of the BCG method

In this section, we give the regret analysis of the C-BCG method for the online convex optimization problem (6.1.1). Throughout this subsection, the sequence $\{x^t\}$ denotes the sequence generated by the C-BCG method.

We first introduce several technical lemmas. The following lemma presents main characteristics of the sequence $\{x^t\}$, which can be verified easily, and hence, we omit the proof here.

Lemma 6.4.1. *For the sequence $\{x^t\}$, we have*

$$\begin{aligned}
&x^t_{\mathcal{J}^i} = x^{t,0}_{\mathcal{J}^i} = x^{t,j}_{\mathcal{J}^i}, \ \forall\, i = 1, 2, \dots, N,\ 1 \le j < i. \\
&x^{t+1}_{\mathcal{J}^i} = x^{t,N}_{\mathcal{J}^i} = x^{t,j}_{\mathcal{J}^i}, \ \forall\, i = 1, 2, \dots, N,\ i \le j \le N. \\
&x^{t+1}_{\mathcal{J}^i} - x^t_{\mathcal{J}^i} = x^{t,N}_{\mathcal{J}^i} - x^{t,0}_{\mathcal{J}^i} = d^{t,i}_{\mathcal{J}^i}, \ \forall\, i = 1, 2, \dots, N.
\end{aligned}$$

The following lemma states that each movement is closely related to the learning rate defined in Assumption 6.3.4.

Lemma 6.4.2. *Suppose that Assumptions 6.3.1-6.3.4 hold. Then, for any $t > 0$, we have*

$$\|x^{t+1} - x^t\| \leq \frac{c\tilde{G}}{\sqrt{t}},$$

where $\tilde{G} = \bar{\beta}\sqrt{2\bar{G}^2 + 2\tau^2 G^2}$, and $\bar{G} = MD + G$.

Proof. Since function $\psi(x)$ in problem (6.1.1) is block separable, from (6.2.1) the subvector $d^{t,i}_{\mathcal{J}^i}$ can be rewritten as

$$d^{t,i}_{\mathcal{J}^i} = \underset{x^{t,i-1}_{\mathcal{J}^i} + d_{\mathcal{J}^i} \in \Omega_i}{\operatorname{argmin}} \left\{ \frac{1}{2\lambda_{t,i}} \|d_{\mathcal{J}^i} + \lambda_{t,i} \nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2 + \tau \psi_i(x^{t,i-1}_{\mathcal{J}^i} + d_{\mathcal{J}^i}) \right\}. \tag{6.4.1}$$

From the first order optimality condition, we have

$$\langle \frac{1}{\lambda_{t,i}} d^{t,i}_{\mathcal{J}^i} + \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}) + \tau \eta^{t,i}_{\mathcal{J}^i}, d_{\mathcal{J}^i} - d^{t,i}_{\mathcal{J}^i} \rangle \geq 0, \forall d_{\mathcal{J}^i} \text{ such that } x^{t,i-1}_{\mathcal{J}^i} + d_{\mathcal{J}^i} \in \Omega_i, \tag{6.4.2}$$

where $\eta^{t,i}_{\mathcal{J}^i} \in \partial \psi_i(x^{t,i}_{\mathcal{J}^i})$. Since $x^{t,i-1}_{\mathcal{J}^i} \in \Omega_i$, we let $d_{\mathcal{J}^i} = 0$ in (6.4.2) and get

$$\langle \frac{1}{\lambda_{t,i}} d^{t,i}_{\mathcal{J}^i} + \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}) + \tau \eta^{t,i}_{\mathcal{J}^i}, -d^{t,i}_{\mathcal{J}^i} \rangle \geq 0, \tag{6.4.3}$$

which implies that

$$\|d^{t,i}_{\mathcal{J}^i}\|^2 \leq \lambda_{t,i} \langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}) + \tau \eta^{t,i}_{\mathcal{J}^i}, -d^{t,i}_{\mathcal{J}^i} \rangle \leq \lambda_{t,i} \|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1}) + \tau \eta^{t,i}_{\mathcal{J}^i}\| \|d^{t,i}_{\mathcal{J}^i}\|.$$

Dividing by $\|d^{t,i}_{\mathcal{J}^i}\|$ on both sides and squaring it, we get

$$\begin{aligned} \|d^{t,i}_{\mathcal{J}^i}\|^2 &\leq \lambda^2_{t,i} \left(\|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\| + \tau \|\eta^{t,i}_{\mathcal{J}^i}\| \right)^2 \\ &\leq 2\frac{c^2 \beta_i^2}{t} \left(\|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2 + \tau^2 \|\eta^{t,i}_{\mathcal{J}^i}\|^2 \right). \end{aligned}$$

Summing this inequality over i from 1 to N, we obtain

$$\|x^{t+1} - x^t\|^2 = \sum_{i=1}^N \|d^{t,i}_{\mathcal{J}^i}\|^2 \leq 2\frac{c^2 \bar{\beta}^2}{t} \left(\sum_{i=1}^N \|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2 + \tau^2 \sum_{i=1}^N \|\eta^{t,i}_{\mathcal{J}^i}\|^2 \right). \tag{6.4.4}$$

Since $(\eta^{t,1}_{\mathcal{J}^1}, \dots, \eta^{t,N}_{\mathcal{J}^N}) \in \partial \psi(x^{t+1})$, it follows from Assumption 6.3.2 that $\sum_{i=1}^N \|\eta^{t,i}_{\mathcal{J}^i}\|^2 \leq G^2$, which together with Lemma 6.3.1 and (6.4.4) proves the desired result. ■

The next result presents an estimator between $F^t(x^t)$ and $F^t(x)$, which plays a key role for the final regret analysis.

Lemma 6.4.3. *For any $t > 0$, we have*

$$F^t(x^t) - F^t(x) \leq S^{1,t} + S^{2,t}(x) + S^{3,t}(x),$$

where $S^{1,t}$, $S^{2,t}(x)$ and $S^{3,t}(x)$ are defined as follows.

$$S^{1,t} := \sum_{i=1}^{N} \left[-\frac{1}{2\lambda_{t,i}} \|x^{t,i}_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\|^2 - \langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}), x^{t,i}_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i} \rangle \right]; \tag{6.4.5}$$

$$S^{2,t}(x) := \sum_{i=1}^{N} \frac{1}{2\lambda_{t,i}} \left[\|x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\|^2 - \|x_{\mathcal{J}^i} - x^{t,i}_{\mathcal{J}^i}\|^2 \right]; \tag{6.4.6}$$

$$S^{3,t}(x) := \sum_{i=1}^{N} \langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}) - \nabla_{\mathcal{J}^i} f^t(x^t), x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i} \rangle + \tau[\psi(x^t) - \psi(x^{t+1})]. \tag{6.4.7}$$

Proof. Using Lemma 2.3.1 with $\varphi(x) = \langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}), x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i} \rangle + \tau\psi_i(x_{\mathcal{J}^i})$, $\lambda = \lambda_{t,i}$, $z_+ = x^{t,i}_{\mathcal{J}^i}$, $z = x^{t,i-1}_{\mathcal{J}^i}$, $i \in \{1, \ldots, N\}$, we have

$$\begin{aligned}\tau\psi_i(x^{t,i}_{\mathcal{J}^i}) - \tau\psi_i(x_{\mathcal{J}^i}) \leq & \frac{1}{2\lambda_{t,i}} \|x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\|^2 - \frac{1}{2\lambda_{t,i}} \|x^{t,i}_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\|^2 - \frac{1}{2\lambda_{t,i}} \|x_{\mathcal{J}^i} - x^{t,i}_{\mathcal{J}^i}\|^2 \\ & + \langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}), x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i} \rangle - \langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}), x^{t,i}_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i} \rangle.\end{aligned} \tag{6.4.8}$$

Moreover, from the convexity of function f^t, we obtain

$$f^t(x^t) - f^t(x) \leq -\langle x - x^t, \nabla f^t(x^t) \rangle. \tag{6.4.9}$$

Then we have

$$\begin{aligned}& F^t(x^t) - F^t(x) \\ & = f^t(x^t) - f^t(x) + \tau[\psi(x^t) - \psi(x)] \\ & \leq -\langle x - x^t, \nabla f^t(x^t) \rangle + \tau[\psi(x^{t+1}) - \psi(x)] + \tau[\psi(x^t) - \psi(x^{t+1})] \\ & = \sum_{i=1}^{N} \left\{ -\langle x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}, \nabla_{\mathcal{J}^i} f^t(x^t) \rangle + \tau[\psi_i(x^{t,i}_{\mathcal{J}^i}) - \psi_i(x_{\mathcal{J}^i})] \right\} + \tau[\psi(x^t) - \psi(x^{t+1})],\end{aligned}$$

where the inequality follows from (6.4.9), and the last equality follows from Lemma 6.4.1. Combining with inequality (6.4.8), we obtain the desired inequality. ■

Next, we establish upper bounds for $S^{1,t}$, $S^{2,t}(x)$ and $S^{3,t}(x)$, respectively.

Lemma 6.4.4. *Suppose that Assumptions 6.3.1-6.3.4 hold. Then, for any $t > 0$, we get*

$$S^{1,t} \leq \frac{c\bar{\beta}}{2\sqrt{t}} \bar{G}^2.$$

Proof. For any $t > 0$ and $i \in \{1, \ldots, N\}$, we have

$$\begin{aligned}
&-\frac{1}{2\lambda_{t,i}}\|x^{t,i}_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\|^2 - \langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}), x^{t,i}_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\rangle \\
&= -\frac{1}{2\lambda_{t,i}}\left\|x^{t,i}_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i} + \lambda_{t,i}\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\right\|^2 + \frac{\lambda_{t,i}}{2}\|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2 \\
&\le \frac{\lambda_{t,i}}{2}\|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2.
\end{aligned}$$

Summing this inequality over i from 1 to N, we get

$$S^{1,t} \le \sum_{i=1}^{N} \frac{\lambda_{t,i}}{2}\|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2 \le \frac{c\bar{\beta}}{2\sqrt{t}}\|g^t\|^2 \le \frac{c\bar{\beta}}{2\sqrt{t}}\bar{G}^2,$$

where the last two inequalities follow from Lemma 6.3.1. ■

For convenience, we define a diagonal matrix $B \in \mathcal{R}^{n\times n}$ with

$$B_{jj} = \beta_i,\ \forall j \in \mathcal{J}^i, i = 1, 2, \ldots, N, \qquad (6.4.10)$$

where $\{\beta_i,\ i = 1, 2, \ldots, N\}$ are constants given in Assumption 6.3.4. Since we assume $\beta_i \ge \underline{\beta} > 0$ for any $i = 1, \ldots, N$, matrix B is invertible.

For $S^{2,t}(x)$, it follows from (6.4.6), Assumption 6.3.4, and the definition of the norm $\|\cdot\|_{B^{-1}}$ that

$$S^{2,t}(x) = \frac{\sqrt{t}}{2c}\left[\|x - x^t\|^2_{B^{-1}} - \|x - x^{t+1}\|^2_{B^{-1}}\right]. \qquad (6.4.11)$$

A bound for $S^{3,t}(x)$ is given by the following lemma.

Lemma 6.4.5. *Suppose that Assumptions 6.3.1-6.3.4 hold. Then, for any $t > 0$, we have*

$$S^{3,t}(x) \le \frac{c}{\sqrt{t}}M\tilde{G}\|x - x^{t+1}\| + \tau[\psi(x^t) - \psi(x^{t+1})],$$

where $\tilde{G} = \bar{\beta}\sqrt{2\bar{G}^2 + 2\tau^2G^2}$, and $\bar{G} = MD + G$.

Proof. In contrast to (6.4.7), we only need to show that $\sum_{i=1}^{N}\langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}) - \nabla_{\mathcal{J}^i} f^t(x^t), x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\rangle \le \frac{c}{\sqrt{t}}M\tilde{G}\|x - x^t\|$ holds for any $t > 0$. In fact, we have

$$\begin{aligned}
&\sum_{i=1}^{N}\langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}) - \nabla_{\mathcal{J}^i} f^t(x^t), x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\rangle \\
&= \langle g^t - \nabla f^t(x^t), x - x^t\rangle
\end{aligned}$$

$$
\begin{aligned}
&\leq \|g^t - \nabla f^t(x^t)\| \|x - x^t\| \\
&\leq M \|x^{t+1} - x^t\| \|x - x^t\| \\
&\leq M \frac{c}{\sqrt{t}} \tilde{G} \|x - x^t\|,
\end{aligned}
$$

where the second inequality follows from Assumption 6.3.3, and the last inequality follows from Lemma 6.4.2. Thus, this completes the proof. ■

Now we show the regret of the BCG method with the cyclic rule.

Theorem 6.4.1. *Suppose that Assumptions 6.3.1-6.3.4 hold. Let $\{x^r\}$ be generated by the C-BCG method for the online optimization problem (6.1.1). Then, for any $x^{*,[T]} \in X^{*,[T]}$, we have*

$$
R(T) \leq \left(\frac{c\bar{\beta}\bar{G}^2}{2} + \frac{D^2}{2c\underline{\beta}} + M\tilde{G}Dc \right) (2\sqrt{T} - 1) + \tau DG,
$$

where $\tilde{G} = \bar{\beta}\sqrt{2\bar{G}^2 + 2\tau^2 G^2}$, and $\bar{G} = MD + G$.

Proof. It follows from Lemma 6.4.3 that

$$
R(T) = \sum_{t=1}^{T} \{F^t(x) - F^t(x^{*,[T]})\} = \sum_{t=1}^{T} S^{1,t} + \sum_{t=1}^{T} S^{2,t}(x^{*,[T]}) + \sum_{t=1}^{T} S^{3,t}(x^{*,[T]}).
$$

Moreover, we have

$$
\sum_{t=1}^{T} \frac{c}{\sqrt{t}} \leq c + c\int_{t=1}^{T} \frac{dt}{\sqrt{t}} \leq c + 2c\sqrt{T} - 2c = c(2\sqrt{T} - 1), \tag{6.4.12}
$$

which, together with Lemma 6.4.4, yields that

$$
\sum_{t=1}^{T} S^{1,t} \leq \sum_{t=1}^{T} \frac{c\bar{\beta}}{2\sqrt{t}} \bar{G}^2 = \frac{c\bar{\beta}\bar{G}^2}{2} (2\sqrt{T} - 1).
$$

For $S^{2,t}(x^{*,[T]})$, it follows from (6.4.11) that

$$
\begin{aligned}
\sum_{t=1}^{T} S^{2,t}(x^{*,[T]}) =& \frac{1}{2c} \|x^{*,[T]} - x^1\|_{B^{-1}}^2 - \frac{\sqrt{T}}{2c} \|x^{*,[T]} - x^{T+1}\|_{B^{-1}}^2 + \sum_{t=2}^{T} (\frac{\sqrt{t}}{2c} - \frac{\sqrt{t-1}}{2c}) \|x^{*,[T]} - x^t\|_{B^{-1}}^2 \\
\leq& \frac{1}{2c\underline{\beta}} D^2 + \sum_{t=2}^{T} (\frac{\sqrt{t}}{2c} - \frac{\sqrt{t-1}}{2c}) \frac{1}{\underline{\beta}} D^2 \\
=& \frac{D^2}{2c\underline{\beta}} \sqrt{T}.
\end{aligned}
$$

Let $\eta \in \partial\psi(x^1)$. Then it follows from Assumption 6.3.2 that $\|\eta\| \leq G$. For $S^{3,t}(x^{*,[T]})$, from Lemma 6.4.5, we have

$$\begin{aligned}
\sum_{t=1}^{T} S^{3,t}(x^{*,[T]}) =& \sum_{t=1}^{T} \left\{ \frac{c}{\sqrt{t}} M\tilde{G}\|x^{*,[T]} - x^{t+1}\| + \tau[\psi(x^t) - \psi(x^{t+1})] \right\} \\
=& \sum_{t=1}^{T} \frac{c}{\sqrt{t}} M\tilde{G}\|x^{*,[T]} - x^{t+1}\| + \tau[\psi(x^1) - \psi(x^{T+1})] \\
\leq& M\tilde{G}D \sum_{t=1}^{T} \frac{c}{\sqrt{t}} + +\tau\langle \eta, x^1 - x^{T+1} \rangle \\
\leq& M\tilde{G}Dc(2\sqrt{T} - 1) + \tau DG, \qquad (6.4.13)
\end{aligned}$$

where the first inequality follows from Assumption 6.3.1 and the convexity of function ψ, and the last inequality follows from (6.4.12) and Assumptions 6.3.1-6.3.2.

Hence, we get

$$\begin{aligned}
R(T) \leq& \frac{c\bar{\beta}\bar{G}^2}{2}(2\sqrt{T} - 1) + \frac{D^2}{2c\underline{\beta}}\sqrt{T} + M\tilde{G}Dc(2\sqrt{T} - 1) + \tau DG \\
\leq& \left(\frac{c\bar{\beta}\bar{G}^2}{2} + \frac{D^2}{2c\underline{\beta}} + M\tilde{G}Dc \right)(2\sqrt{T} - 1) + \tau DG,
\end{aligned}$$

where the second inequality follows from the fact $\sqrt{T} \leq 2\sqrt{T} - 1$, $T \geq 1$. ■

Note that the regret bound of the C-BCG method in Theorem 6.4.1 is independent of the number N of blocks and the dimension n. Moreover, Theorem 6.4.1 implies that $\frac{R(T)}{T} \leq O(\frac{1}{\sqrt{T}})$, and the C-BCG method is a no internal regret algorithm. The next remark states that the regret bound of the C-BCG method in Theorem 6.4.1 is an extension of bound of the greedy projection method [81].

Remark 6.4.1. *When $\psi = 0$, we have $\psi(x^1) = \psi(x^{t+1}) = 0$. Then the evaluation for $\sum_{t=1}^{T} S^{3,t}(x^{*,[T]})$ in (6.4.13) reduces to $\sum_{t=1}^{T} S^{3,t}(x^{*,[T]}) \leq M\tilde{G}Dc(2\sqrt{T} - 1)$. Moreover, if we let $N = 1$, $\beta_i = 1$, $i = 1, 2, \ldots, N$, and $c = 1$, from Lemma 6.3.1 and Remark 6.3.1, we have $\bar{G} = G$ and $M = 0$. Hence, the regret of the C-BCG method reduces to $R(T) \leq \frac{G^2}{2}(2\sqrt{T} - 1) + \frac{D^2}{2}\sqrt{T}$, which is the same as Theorem 6.3.1. Therefore, the C-BCG method is an extension of the greedy projection method.*

6.5 Extension to the stochastic optimization problem

In this section, we develop the ergodic convergence of the proposed C-BCG method for the stochastic optimization problem.

We consider the following regularized convex stochastic optimization problem.

$$\underset{x}{\text{minimize}}\, \tilde{F}(x) := E_z[f(x,z)] + \tau\psi(x), \tag{6.5.1}$$

where $z = (u, v) \in \mathcal{R}^{n+n}$ is an input-output pair of the data drawn from an unknown underlying distribution, $f(x, z)$ is the loss function of using u with parameter x to predict v, $E_z[f(x, z)]$ denotes the expected value of the loss function $f(x, z)$ with respect to the selection pair z.

A common way to solve stochastic optimization problem (6.5.1) is to approximate the expectation of the whole loss $E_z[f(x, z)]$ by using a finite set of independent observations $z_1, \dots, z_t$, and solve the following problem instead.

$$\underset{x}{\text{minimize}}\, \frac{1}{T}\sum_{t=1}^{T} f(x, z_t) + \tau\psi(x). \tag{6.5.2}$$

This problem can be regarded as the batch optimization problem for the online optimization problem with $F^t = f(x, z_t) + \tau\psi(x)$. Let $x^* \in \underset{x}{\operatorname{argmin}}\, \tilde{F}(x)$. The corresponding regret can be written as follows.

$$R(T) = \sum_{t=1}^{T}\big\{f(x^t, z_t) + \tau\psi(x^t)\big\} - \sum_{t=1}^{T}\big\{f(x^*, z_t) + \tau\psi(x^*)\big\}. \tag{6.5.3}$$

Theorem 6.5.1. *Suppose that Assumptions 6.3.1-6.3.4 hold. Let $\{x^t\}$ be generated by the C-BCG method for the convex stochastic optimization problem (6.5.1), and $\bar{x}^T = \frac{1}{T}\sum_{r=1}^{T} x^r$, $T \geq 1$. Then, for any $x^* \in \underset{x}{\operatorname{argmin}}\, \tilde{F}(x)$, we have*

$$E_{z_{[T]}}[\tilde{F}(\bar{x}^T)] - \tilde{F}(x^*) \leq \left(\frac{c\bar{\beta}\bar{G}^2}{2} + \frac{D^2}{2c\underline{\beta}} + M\tilde{G}Dc\right)\frac{2\sqrt{T}-1}{T} + \frac{1}{T}\tau DG.$$

Proof. Let $z_{[T]} := \{z_1, \dots, z_T\}$, where $\{z_i,\ 1 \leq i \leq T\}$ follow the independent and isotonical distribution. Note that the variable x^t, $1 \leq t \leq T$, is dependent on the random variables $\{z_1, \dots, z_{t-1}\}$, but independent on the random variables $\{z_t, \dots, z_T\}$.

Hence, we have

$$\begin{aligned} &E_{z_{[T]}}[f(x^t, z_t)] + \tau\psi(x^t) = E_{z_{[t-1]}}\big[E_{z_{[t]}}[f(x^t, z_t)] + \tau\psi(x^t)\big] = E_{z_{[t-1]}}[\tilde{F}(x^t)],\\ &E_{z_{[T]}}[f(x^*, z_t)] + \tau\psi(x) = E_{z_{[t]}} f(x^*, z_t) + \tau\psi(x^*) = \tilde{F}(x^*). \end{aligned}$$

Combing with (6.5.3), we get

$$0 \leq E_{z_{[T]}}[R(T)] = \sum_{t=1}^{T}\Big(E_{z_{[T]}}[\tilde{F}(x^t)] - \tilde{F}(x^*)\Big). \tag{6.5.4}$$

On the other hand, by the convexity of the function $\tilde{F}(x)$, we have

$$\tilde{F}(\bar{x}^T) = \tilde{F}(\frac{1}{T}\sum_{t=1}^{T} x^t) \leq \frac{1}{T}\sum_{t=1}^{T} \tilde{F}(x^t). \tag{6.5.5}$$

Subtracting the optimal value $\tilde{F}(x^*)$ on both sides and taking the expectation with respect to the random variables $z_{[T]}$, we get

$$E_{z_{[T]}}[\tilde{F}(\bar{x}^T)] - \tilde{F}(x^*) \leq \frac{1}{T}\sum_{t=1}^{T}\left(E_{z_{[T]}}[\tilde{F}(x^t)] - \tilde{F}(x^*)\right) = \frac{1}{T}E_{z_{[T]}}R(T).$$

From Theorem 6.4.1, we prove the desired result. ■

Note that Corollary 6.5.1 implies that $E_{z_{[T]}}[\tilde{F}(\bar{x}^T)] - \tilde{F}(x^*) \leq O(\frac{1}{\sqrt{T}})$, where the upper bound $O(\frac{1}{\sqrt{T}})$ is sharper than $O(\frac{1+\log T}{\sqrt{1+T}}N)$ given in [75].

6.6 Conclusion

In this chapter, we have proposed a block coordinate gradient (BCG) method for the online convex optimization problem (6.1.1). We have shown that the proposed method has a regret $O(\sqrt{T})$, which is the same as [81]. Moreover, we have extended our results to the regularized stochastic optimization problem, and have shown that the results in this chapter are tighter than that in [75].

In Chapter 4, an extension of the BCG method, called the "block coordinate proximal gradient (BCPG) methods with variable Bregman functions", has been studied, where the quadratic term $\frac{1}{2}\|d\|^2$ in (6.2.1) is replaced by the Bregman distance $B_\eta(x, x+d)$. It is shown that the proposed BCPG methods have the same convergence rate with the block coordinate gradient descent (BCGD) method for the classical separable optimization problems. Hence, it may be possible to obtain a similar convergence of the BCPG methods with variable Bregman functions for the online and stochastic optimization problems as the results in this chapter.

Chapter 7

Conclusion

In this thesis, we have proposed two classes of block coordinate gradient methods for solving the classical separable optimization problem, and introduced a new concept, called the "block lower triangular Lipschitz continuous", with which the theoretical analysis of the block coordinate gradient method is supplemented and improved for various optimization problems, including the online and the stochastic optimization problems. The results obtained in this thesis are summarized as follows.

(a) In Chapter 3, we have proposed an inexact coordinate descent (ICD) method for a class of weighted l_1-regularized convex optimization problem with a box constraint, where we have given a new criterion for the "inexact solution" of the subproblem and only required an approximate solution at each iteration. For the proposed method, we have established its R-linear convergence rate with the almost cycle rule, and have examined its efficiency by numerical experiments on the comparison of the proposed method and the coordinate gradient descent method.

(b) In Chapter 4, we have proposed a class of block coordinate proximal gradient (BCPG) methods with variable Bregman functions for solving the general nonsmooth nonconvex problem. We have established the global convergence and R-linear convergence rate for the proposed methods with the Gauss-Seidel rule. The idea of using the variable kernels is the innovation of these methods, which enabled us to obtain many well-known algorithms from the proposed BCPG methods, including the (inexact) BCD method. Some special kernels even allowed the proposed BCPG methods to adopt the fixed step size, and helped us to construct accelerated algorithms. Finally, the numerical results on the proposed algorithm and the algorithm with a fixed kernel proved the efficiency.

(c) In Chapter 5, we have improved the iteration complexity of the block coordinate gradient descent (BCGD) method with the cyclic rule for the convex separable optimization. The improvement lies in the new Lipschitz continuity-like assumption. In particular, we have proven that the BCGD method with the cyclic rule converges with $O(\frac{\max\{M, L_f\}}{\varepsilon})$, where M is the constant given in the proposed assumption, and L_f is the Lipschitz constant. Furthermore, we have studied the relation between M and L_f, and showed that $M \leq \sqrt{N}L_f$ or $M \leq 2L_f$, which implies that the iteration complexity bound derived in this thesis is sharper than existing results.

(d) In Chapter 6, we have investigated the performance of a block coordinate gradient (BCG) method with the cyclic rule for the online optimization problem and the stochastic optimization problem. We firstly have shown that the proposed method has the same regret as the greedy projection (GP) method, where the GP method is a full gradient projection method. Moreover, we have extended the BCG method for the stochastic optimization problem, and have shown that the the result in this thesis is tighter than the existing results.

As we summarized above, we have made several contributions on the block coordinate gradient methods for the separable optimization problems. However, there are many problems that still remain unknown. In what follows, we mention some main issues based on our current achievements.

(a) In Chapter 4, we have presented a class of block coordinate proximal gradient (BCPG) methods for the separable nonsmooth nonconvex optimization problem. For showing the convergence rate of the proposed methods, we assumed that the Lipachitz local error bound assumption holds for the original probelm. In [74], another assumption, called the "Kurdyka-Lojasiewiez (KL) inequality", is established for the BCD method. It is interesting to study the relation between the KL inequality and the local error bound in the future. Moreover, extending the BCPG methods to the more general constrained problems, such as the support vector machine (SVM) problem, is also a challenging topic.

(b) In this thesis, we have proposed a new Lipschitz continuity-like definition, and have studied the relation between two constants M and L_f in Chapter 5, where M is the constant given in the proposed definition, and L_f is the common Lipschitz constant. Although we have shown that $M \leq \sqrt{N}L_f$ holds for any function, and have found several classes of functions such that the relation $M \leq 2L_f$ holds. In the future, it would be interesting to find more functions for which the corresponding constant M is independent of the number N of blocks. Currently, we have not found a counterexample where $N^\sigma L_f \leq M$ for a positive constant σ.

Bibliography

[1] A. Auslender, *Optimisation, Méthodes Numéiques*, Masson, Paris, 1976.

[2] A. Bagirov and A.N. Ganjehlou, *An approximate subgradient algorithm for unconstrained nonsmooth, nonconvex optimization*, Mathematical Methods of Operations Research, 67 (2008), pp. 187–206.

[3] N. Bansal, A. Blum, S. Chawla, and A. Meyerson, *Online oblivious routing*, Proceedings of the 15th Annual ACM Symposium on Parallel Algorithms and Architectures, (2003), pp. 44–49.

[4] P.L. Bartlett, M. Collins, B. Taskar, and D. McAllester, *Exponentiated gradient algorithms for large-margin structured classification*, Proceedings of Advances in Neural Information Processing Systems 17, 2004.

[5] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202.

[6] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters, 31 (2003), pp. 167–175.

[7] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM Journal on Optimization, 23 (2013), pp. 2037–2060.

[8] D.P. Bertsekas, *Nonlinear Programming*, 2nd edition, Athena Scientific, Belmont, 1999.

[9] E.G. Birgin, R. Biloti, M. Tygel, and L.T. Santos, *Restricted optimization: A clue to a fast and accurate implementation of the common reflection surface stack method*, Journal of Applied Geophysics, 42 (1999), pp. 143–155.

[10] J. Blitzer, M. Dredze, and F. Pereira, *Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification*, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, (2007), pp. 440–447.

[11] S. Bonettini, *Inexact block coordinate descent methods with application to non-negative matrix factorization*, IMA Journal of Numerical Analysis, 31 (2011), pp. 1431–1452.

[12] J.M. Borwein and A.S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, Canadian Mathematical Society Books in Mathematics, spinger, New York, 2000.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[14] S. Bubeck, *Introduction to Online Optimization*, Lecture Notes, 2011, http://www.princeton.edu/∼sbubeck/BubeckLectureNotes.pdf.

[15] J.V. Burke, A.S. Lewis, and M.L. Overton, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM Journal on Optimization, 15 (2005), pp. 751–779.

[16] N. Cesa-Bianchi and S. Kakade, *An optimal algorithm for linear bandits*, (2011), arXiv: 1110.4322v1.

[17] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM Journal on Optimization, 3 (1993), pp. 538–543.

[18] S.S. Chen, D.L. Donoho, and M.A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Journal on Scientific Computing, 20 (1998), pp. 33–61.

[19] M. Collins, A. Globerson, T. Koo, X. Carreras, and P.L. Bartlett, *Exponentiated gradient algorithms for conditional random fields and Max-Margin Markov networks*, The Journal of Machine Learning Research, 9 (2008), pp. 1775–1822.

[20] T.M. Cover, *Universal portfolios*, Mathematical Finance, 1 (1991), pp. 1–29.

[21] F.E. Curtis and M.L. Overton, *A sequential quadratic programming algorithm for nonconvex, nonsmooth constrained optimization*, SIAM Journal on Optimization, 22 (2012), pp. 474–500.

[22] V. Dani, S.M. Kakade, and T.P. Hayes, *The price of bandit information for online optimization*, Proceedings of Advances in Neural Information Processing Systems 20, 2007.

[23] I. Daubechies, M. Defrise, and C.D. Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Communications on Pure and Applied Mathematics, 57 (2004), pp. 1413–1457.

[24] J. Duchi, E. Hazan, and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, The Journal of Machine Learning Research, 12 (2011), pp. 2121–2159.

[25] J. Duchi and Y. Singer, *Efficient online and batch learning using forward backward splitting*, The Journal of Machine Learning Research, 10 (2009), pp. 2899–2934.

[26] M.A.T. Figueiredo, R.D. Nowak, and S.J. Wright, *Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems*, IEEE Journal of Selected Topics in Signal Processing, 1 (2007), pp. 586–597.

[27] D.P. Foster and R. Vohra, *Regret in the on-line decision problem*, Games and Economic Behavior, 29 (1999), pp. 7–35.

[28] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, *Pathwise coordinate optimization*, Annals of Applied Statistics, 1 (2007), pp. 302–332.

[29] A. Gholami and H.R. Siahkoohi, *Regularization of linear and non-linear geophysical ill-posed problems with joint sparsity constraints*, Geophysical Journal International, 180 (2010), pp. 871–882.

[30] E. Hazan, A. Kalai, S. Kale, and A. Agarwal, *Logarithmic regret algorithms for online convex optimization*, Machine Learning, 69 (2007), pp. 169–192.

[31] A.J. Hoffman, *On approximate solutions of systems of linear inequalities*, Journal of Research of the National Bureau of Standards, 49 (1952), pp. 263–265.

[32] M. Hong, X. Wang, M. Razaviyayn, and Z-Q. Luo, *Iteration complexity analysis of block coordinate descent methods*, Technical Report, University of Minnesota, USA, 2013, http://arxiv.org/abs/1310.6957.

[33] J. Huang, T. Zhang, and D. Metaxas, *Learning with structured sparsity*, The Journal of Machine Learning Research, 12 (2001), pp. 3371–3412.

[34] A. Kaplan and R. Tichatschke, *Proximal point methods and nonconvex optimization*, Journal of Global Optimization, 13 (1998), pp. 389–406.

[35] K. Koh, S.J. Kim, and S. Boyd, *An interior-point method for large-scale l_1-regularized logistic regression*, Journal of Machine Learning Research, 8 (2007), pp. 1519–1555.

[36] M. Kowalski and B. Torrésani, *Structured sparsity: From mixed norms to structured shrinkage*, Workshop on Signal Processing with Adaptive Sparse Representations, 2009.

[37] P.L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM Journal on Numerical Analysis, 16 (1979), pp. 964–979.

[38] H. Liu, M. Palatucci, and J. Zhang, *Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery*, Proceedings of the 26th Annual International Conference on Machine Learning, (2009), pp. 649–656.

[39] J. Liu and S.J. Wright, *Asynchronous stochastic coordinate descent: parallelism and convergence properties*, (2014), arXiv:1403.3862.

[40] Z. Lu and L. Xiao, *On the complexity analysis of randomized block-coordinate descent methods*, (2013), arXiv: 1305.4723.

[41] D.G. Luenberger, *Introduction to Linear and Nonlinear Programming*, Addision Wesley, 1973.

[42] L. Lukšan and J. Vlček, *A bundle-Newton method for nonsmooth unconstrained minimization*, Mathematical Programming, 83 (1998), pp. 373–391.

[43] Z.Q. Luo and P. Tseng, *On the convergence of the coordinate descent method for convex differentiable minimization*, Journal of Optimization and Applications, 72 (1992), pp. 7–35.

[44] Z. Q. Luo and P. Tseng, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM Journal on Control and Optimization, 30 (1992), pp. 408–425.

[45] J. Mairal, *Optimization with first-order surrogate functions*, (2013), arXiv: 1305.3120

[46] S. Mehrotra, *On the implementation of a primal-dual interior point method*, SIAM Journal on Optimization, 2 (1992), pp. 575–601.

[47] L. Meier, S.V.D. Geer, and P. Bühlmann, *The group lasso for logistic regression*, Journal of the Royal Statistical Society: Series B, 70 (2008), pp. 53–71.

[48] J.J. Moré and G. Toraldo, *On the solution of large quadratic programming problems with bound constraints*, SIAM Journal on Optimization, 1 (1991), pp. 93–113.

[49] B.A. Murtagh and M.A. Saunders, *MINOS 5.5 user's guide. Report SOL 83–20R*, Department of Operations Research, Stanford University, Stanford, 1983.

[50] Y. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM Journal on Optimization, 22 (2012), pp. 341–362.

[51] Y. Nesterov, *Gradient methods for minimizing composite objective function*, CORE Report, 2007, http://www.ecore.be/DPs/dp 1191313936.pdf.

[52] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publisher, Boston, 2004.

[53] F. Orabona, K. Crammer, and N. Cesa-Bianchi, *A generalized online mirror descent with applications to classification and regression*, (2013), arXiv: 1304.2994.

[54] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[55] M.Y. Park and T. Hastie, *L1-regularization path algorithm for generalized linear models*, Journal of the Royal Statistical Society: Series B, 69 (2007), pp. 659–677.

[56] F.A. Potra, *On Q-order and R-order of convergence*, Journal of Optimization Theory and Applications, 63 (1989), pp. 415–431.

[57] M.J.D. Powell, *On search directions for minimization algorithms*, Mathematical Programming, 4 (1973), pp. 193–201.

[58] Z. Qin, K. Scheinberg, and D. Goldfarb, *Efficient block-coordinate descent algorithms for the Group Lasso*, Mathematical Programming Computation, 5 (2013), pp. 143–169.

[59] P. Richtárik and M. Takáč, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Mathematical Programming, 144 (2014), pp. 1–38.

[60] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1970.

[61] R.T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.

[62] A. Saha and A. Tewari, *On the nonasymptotic convergence of cyclic coordinate descent methods*, SIAM Journal on Optimization, 23 (2013), pp. 576–601.

[63] N. Schraudolph, J. Yu, and S. Günter, *A stochastic quasi-Newton method for online convex optimization*, Proceedings of 11th International Conference on Artificial Intelligence and Statistics, (2009), pp. 436–443.

[64] H. Taylor, S. Bank, and J. McCoy, *Deconvolution with the l_1 norm*, Geophysics, 44 (1979), pp. 39–52.

[65] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B, 58 (1996), pp. 267–288.

[66] P. Tseng, *Approximation accuracy, gradient methods, and error bound for structured convex optimization*, Mathematical Programming, 125 (2010), pp. 263–295.

[67] P. Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization*, Journal of Optimization Theory and Applications, 109 (2001), pp. 475–494.

[68] P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization*, Technical report, Department of Mathematics, University of Washington, 2008.

[69] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Mathematical Programming, 117 (2009), pp. 387–423.

[70] P. Tseng and S. Yun, *Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization*, Journal of Optimization Theory and Applications, 140 (2009), pp. 513–535.

[71] S.J. Wright, *Accelerated block-coordinate relaxation for regularized optimization*, SIAM Journal on Optimization, 22 (2012), pp. 159–186.

[72] T.T. Wu and K. Lange, *Coordinate descent algorithms for lasso penalized regression*, The Annals of Applied Statistics, 2 (2008), pp. 224–244.

[73] L. Xiao, *Dual averaging methods for regularized stochastic learning and online optimization*, Proceedings of Advances in Neural Information Processing Systems 22, 2009.

[74] Y. Xu and W. Yin, *A block coordinate descent method for multi-convex optimization with applications to nonnegative tensor factorization and completion*, Technical Report, Department of Computational and Applied Mathematics, Rice University, 2012.

[75] Y. Xu and W. Yin, *Block stochastic gradient iteration for convex and nonconvex optimization*, (2014), arXiv: 1408.2597.

[76] G.B. Ye, Y.F. Chen, and X. Xie, *Efficient variable selection in support vector machines via the alternating direction method of multipliers*, Proceedings of 14th International Conference on Atificial Intelligence and Statistics, (2011), pp. 832–840.

[77] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, *Bregman iterative algorithms for l_1-minimization with applications to compressed sensing*, SIAM Journal on Imaging Sciences, 1 (2008), pp. 143–168.

[78] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society: Series B, 68 (2006), pp. 49–67.

[79] J. Zhou, W.H.K. Lam, and B.G. Heydecker, *The generalized Nash equilibrium model for oligopolistic transit market with elastic demand*, Transportation Research Part B: Methodological, 39 (2005), pp. 519–544.

[80] C. Zhu, R.H. Byrd, and J. Nocedal, *L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization*, ACM Transactions on Mathematical Software, 23 (1997), pp. 550–560.

[81] M. Zinkevich, *Online convex programming and generalized infinitesimal gradient ascent*, Proceedings of 20th International Conference on Machine Learning, (2003), pp. 928–936.

[82] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: Series B, 67 (2005), pp. 301–320.