Subgradient Methods for Convex Minimization

by

Angelia Nedić

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2002

Author .....
Department of Electrical Engineering and Computer Science
May 20, 2002

Certified by .....
Dimitri P. Bertsekas
Professor
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Subgradient Methods for Convex Minimization

by

Angelia Nedić

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 2002, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering

Abstract

Many optimization problems arising in various applications require minimization of an objective cost function that is convex but not differentiable. Such a minimization arises, for example, in model construction, system identification, neural networks, pattern classification, and various assignment, scheduling, and allocation problems. To solve convex but not differentiable problems, we have to employ special methods that can work in the absence of differentiability, while taking the advantage of convexity and possibly other special structures that our minimization problem may possess. In this thesis, we propose and analyze some new methods that can solve convex (not necessarily differentiable) problems. In particular, we consider two classes of methods: incremental and variable metric.

Thesis Supervisor: Dimitri P. Bertsekas
Title: Professor

Acknowledgments

I acknowledge my deepest gratitude to my thesis supervisor, Professor Dimitri Bertsekas, for his ceaseless guidance and support through the course of this work, and for taking me as his student in the first place. His extraordinary insights and expertise in the field of optimization have immensely influenced my work and contributed to my professional development.

I would also like to thank the members of my thesis committee, Professor Sanjoy Mitter and John Tsitsiklis, for their insights and suggestions. I also thank all of my friends at LIDS, without whom my life at MIT would not be as enjoyable as it was. In particular, I would like to thank Asu Ozdaglar, Anand Ganti, Emre Koksal, Haixia Lin, Jinane Abounadi, Jun Sun, Karen Sigurd, May Liu, Soosan Beheshti Sean Warnick, Tengo Saengudomlert, and Vijay Konda. I am especially grateful to Asu for being more than a friend. I also thank my parents and my son, for their love and support.

My thesis research was supported by NSF under Grant ACI-9873339.

Subgradient Methods for Convex Minimization

Contents

1. Introduction . . . . . p. 9
1.1. Purpose of the Thesis . . . . . p. 9
1.2. Subgradient Methods . . . . . p. 10
1.3. Incremental Approach . . . . . p. 10
1.4. Variable Metric Approach . . . . . p. 13
1.5. Outline of the Thesis . . . . . p. 14

PART I: INCREMENTAL SUBGRADIENT METHODS

2. An Incremental Subgradient Method . . . . . p. 19
2.1. The Method . . . . . p. 20
2.2. Assumptions and Some Basic Relations . . . . . p. 20
2.3. Constant Stepsize Rule . . . . . p. 22
2.4. Diminishing Stepsize Rule . . . . . p. 26
2.5. Dynamic Stepsize Rule for Known f^* . . . . . p. 30
2.6. Dynamic Stepsize Rule for Unknown f^* . . . . . p. 34
2.7. Effects of Fixed Cyclic Processing Order . . . . . p. 41

3. An Incremental Subgradient Method with Randomization . . . . . p. 45
3.1. The Method . . . . . p. 45
3.2. Assumptions and Some Basic Relations . . . . . p. 46
3.3. Constant Stepsize Rule . . . . . p. 47
3.4. Diminishing Stepsize Rule . . . . . p. 53
3.5. Dynamic Stepsize Rule for Known f^* . . . . . p. 54
3.6. Dynamic Stepsize Rule for Unknown f^* . . . . . p. 61
3.7. Effects of Randomized Processing Order . . . . . p. 67
3.8. Experimental results . . . . . p. 70
3.9. Distributed Asynchronous Incremental Subgradient Method . . . . . p. 76

4. Extensions of the Incremental Subgradient Method . . . . . p. 103
4.1. An Incremental Subgradient Method with Weights . . . . . p. 103
4.2. An Incremental Approximate Subgradient Method . . . . . p. 107

PART II: VARIABLE METRIC SUBGRADIENT METHODS

5. A Variable Metric Subgradient Method . . . . . p. 119
5.1. The Method . . . . . p. 120
5.2. Assumptions and Some Basic Relations . . . . . p. 120
5.3. Constant and Diminishing Stepsize Rules . . . . . p. 123
5.4. Dynamic Stepsize Rules . . . . . p. 126

6. Space Dilation Methods . . . . . p. 137
6.1. Dilation Along Subgradients . . . . . p. 138
6.2. Properties of Dilation Transformations . . . . . p. 140
6.3. Assumptions and Some Basic Relations . . . . . p. 144
6.4. Convergence Properties of the Method with Dilation along Subgradients . . p. 147
6.5. Dilation Along Other Directions . . . . . p. 155
6.6. Assumptions and Some Basic Relations . . . . . p. 156
6.7. Convergence Properties of the Method with Dilation along Other Directions . p. 160

References . . . . . p. 169

1

Introduction

1.1. PURPOSE OF THE THESIS

Many optimization problems arising in various applications require minimization of an objective cost function that is convex but not differentiable. Such a minimization arises, for example, in model construction, system identification, neural networks, pattern classification, and various assignment, scheduling, and allocation problems. To solve convex but not differentiable problems, we have to employ special methods that can work in the absence of differentiability, while taking the advantage of convexity and possibly other special structures that our minimization problem may possess.

In this thesis, we propose and analyze some new methods that can solve convex (not necessarily differentiable) problems. In particular, we consider two classes of methods: incremental and variable metric. In the first part of the thesis, we discuss the incremental methods, which are applicable to problems where the objective cost function has an additive structure. These methods combine the ideas of the incremental approach with those of the standard methods for convex minimization. We propose and analyze several versions of the incremental method, including some that are stochastic, as well as some with special features such as weights. We study convergence of the method for various stepsize rules and for synchronous and asynchronous computational setting. Our convergence analysis and computational results indicate that the incremental methods can perform far better than their nonincremental counterparts.

In the second part of the thesis, we consider variable metric methods, which are applicable to unconstrained convex minimization problems. These methods are particularly well suited for poorly scaled problems, for which the standard methods of convex minimization are typically very slow. The variable metric methods combine the principles of variable metric approach with those of the standard methods for convex minimization. We discuss a variable metric method in a general form and a more specific method that employs space dilation.

1.2. SUBGRADIENT METHODS

Subgradient methods are the principal methods used in convex nondifferentiable minimization. This type of minimization arises in many applications, as well as in the context of duality, and various general solution strategies such as penalty function methods, regularization methods, and decomposition methods. Most notably, subgradient methods are used in the context of duality arising from Lagrangian relaxation, where they are referred to as *dual ascent methods* or *dual methods*.

Subgradient methods were first introduced in the Soviet Union in the middle sixties by N. Z. Shor. Since then, they have been extensively studied, and in general two major classes of subgradient methods have been developed: descent-based methods and nondescent methods. The descent-based subgradient methods are based on the principal idea of the function descent, which lies in the framework of gradient-type minimization. Nondescent subgradient methods are based on the idea of the distance decrease (distance from the set of minima), and their implementation is simpler than that of descent-based methods.

For nondescent subgradient methods, the early work of Ermoliev [Erm66] and Polyak [Pol67] was particularly influential. Due to their simple implementation, the nondescent subgradient methods have drawn a lot of attention, and the literature on these methods is very rich. An extensive treatment of these subgradient methods can be found in the textbooks by Dem'yanov and Vasil'ev [DeV85], Shor [Sho85], Minoux [Min86], Polyak [Pol87], Hiriart-Urruty and Lemaréchal [HiL93] Shor [Sho98], and Bertsekas [Ber99].

Our work is in the framework of the nondescent subgradient methods, for which we study, separately, the merits of the incremental approach and variable metric approach. In the next two sections, we describe these approaches and discuss the contributions of the thesis.

1.3. INCREMENTAL APPROACH

1.3.1 Problem Formulation

In the first part of the thesis, we consider an incremental approach for minimizing a function

that consists of the sum of a large number of component functions:

$$\begin{aligned} &\text{minimize} \quad f(x) = \sum_{i=1}^{m} f_i(x) \\ &\text{subject to} \quad x \in X, \end{aligned} \tag{1.1}$$

where the functions $f_i : \Re^n \mapsto \Re$, $i = 1, \ldots, m$, are convex, and the set $X \subset \Re^n$ is nonempty, closed, and convex. The most prominent example of this type of minimization is the linear least-squares problem arising in a broad class of practical problems such as model construction, system identification, neural networks, and pattern classification. This type of minimization also arises from Lagrangian relaxation of coupling constraints of large-scale separable problems in the domains of integer programming and combinatorial optimization, including various assignment, scheduling, and allocation problems such as: job assignment, job-shop scheduling, file allocation, bank accounts location, location of plants, concentrator location (in network design), and warehouse location.

Our primary interest is in the problem where f is nondifferentiable. Nondifferentiability is typical for problem (1.1) arising from Lagrangian relaxation, and solving this problem, possibly within a branch-and-bound or some heuristic context, is one of the most important and challenging algorithmic areas of optimization. When a branch-and-bound algorithm involves subproblems of the form (1.1), solving such subproblems quickly and with high accuracy results in a faster and more efficient search, thus improving the overall performance of the branch-and-bound algorithm. Hence, it is important to have an algorithm that quickly yields a good approximation of the optimal cost for problem (1.1).

A classical method for solving problem (1.1) is the subgradient method, whereby at each iteration we take a step in the opposite direction of a subgradient of f and we obtain a new iterate by projecting on the constraint set X (for the time being, we can think of a subgradient as a substitute for a gradient in the absence of the differentiability of f). This classical subgradient method, however, does not exploit the additive structure of the function f. To take advantage of the special structure of f, we consider an incremental approach. The idea of this approach is to perform iterations incrementally, by sequentially taking steps along the negative subgradients of the component functions, with intermediate adjustment of the variables after processing each component function. An iteration of the incremental subgradient method can be visualized an a long cycle consisting of m steps, whereby at each step we process only one component f_i such that all components $f_1, \ldots, f_m$ are processed exactly once within the cycle.

1.3.2 Previous Work

Incremental gradient methods for *differentiable* unconstrained problems have a long tradition, most notably in the training of neural networks, where they are known as *backpropagation methods*. These methods are related to the Widrow–Hoff algorithm [WiH60] and to stochastic gradient/stochastic approximation methods. The incremental gradient methods have been extensively studied, most recently by Luo [Luo91], Gaivoronski [Gai94], Grippo [Gri94], Luo and

Tseng [LuT94], Mangasarian and Solodov [MaS94], Bertsekas and Tsitsiklis [BeT96], Bertsekas [Ber97], Tseng [Tse98], Bertsekas and Tsitsiklis [BeT00]. It has been experimentally observed that incremental gradient methods often converge much faster than the steepest descent.

Incremental subgradient methods are similarly motivated by rate of convergence considerations. Despite the success of the incremental gradient methods, the merits of the incremental approach for solving *nondifferentiable* convex problems had not been properly evaluated prior to this work. Previous work on the incremental subgradient method is rather scarce. The method was proposed and analyzed by Kibardin in [Kib79], and later, also analyzed by Solodov and Zavriev [SoZ98]. In both of these references, convergence properties and some convergence rate estimates were established only for a cyclic nonrandomized processing order of function components $f_1, \ldots, f_m$ and for a stepsize diminishing to zero.

1.3.3 Our Work and Contributions

Our work has two general contributions:

(a) *New Algorithms.* The development of fast and simple methods, with low overhead per iteration, and in particular, the development of the incremental method with randomization, and the development of distributed and asynchronous incremental methods.

(b) *Unifying Analysis.* A line of analysis that provides a new framework for the convergence analysis of the whole class of nondescent subgradient methods including the methods with variable metric. This unifies the existing theory of nondescent subgradient methods.

Our work also has many specific contributions, such as new stepsize rules, convergence results, and convergence rate estimates. In particular, in the first part of the thesis, we give an exhaustive study of the incremental subgradient method and its versions using randomization, weights, and approximate subgradients. For the method and each of its versions, we provide a number of convergence and convergence rate results under four different stepsize rules:

(1) Constant stepsize rule, where the stepsize is fixed to a positive scalar.

(2) Diminishing stepsize, where the stepsize diminishes to zero.

(3) Dynamic stepsize with known optimal cost value, where the stepsize is proportional to the difference between the function value at the current iterate and the optimal cost value.

(4) Dynamic stepsize with unknown optimal cost value, which is a modification of the stepsize rule (3) obtained by using an estimate of the optimal cost value. For this stepsize, we give two estimate update procedures.

Even though the incremental approach performs well in centralized computation, it may perform even better in parallel computation especially for typical problems where computation of the component subgradients is relatively costly. For such problems, we propose and analyze a distributed asynchronous incremental subgradient method, where the computation of the component subgradients is distributed among a set of processors that communicate only with a coordinator. Our distributed methods are motivated by the parallel asynchronous deterministic

and stochastic gradient methods of Tsitsiklis, Bertsekas, and Athans [TBA86], and Bertsekas and Tsitsiklis [BeT89]. However, our algorithms do not fit in the framework of the general algorithmic models of Chapter 6 and 7 Bertsekas and Tsitsiklis [BeT89], and therefore our algorithms are not covered by the line of analysis of this reference.

We discovered that the performance of the incremental method depends not only on stepsize choice, but also on the processing order for the component functions. The dependence on the stepsize choice is captured by our convergence rate analysis. However, the dependence on the processing order for processing the component functions turned out to be very complex, and we were not able to capture it analytically. We also discovered that randomization can alleviate possibly detrimental effects of unfavorable processing orders, which we were able to prove analytically and to support by computational results.

Our convergence results are more general and by far richer than those of Kibardin [Kib79], and Solodov and Zavriev [SoZ98]. Our analytic approach is motivated by that of Dem'yanov and Vasil'ev [DeV85], Polyak [Pol87], and Correa and Lemaréchal [CoL93] for the ordinary subgradient method. Since the ordinary subgradient method is a special case of the incremental subgradient method where $m = 1$, our convergence results can be viewed as a generalization of the corresponding convergence results for the ordinary subgradient method, which can be found in the textbooks by Dem'yanov and Vasil'ev [DeV85], Shor [Sho85] and [Sho98], Minoux [Min86], Polyak [Pol87], Hiriart-Urruty and Lemaréchal [HiL93], Bertsekas [Ber99].

Most of the thesis work on incremental methods was previously published in the journal papers by Nedić, Bertsekas, and Borkar [NBB01], and Nedić and Bertsekas [NeB01a], [NeB01b].

1.4. VARIABLE METRIC APPROACH

1.4.1 Problem Formulation

In the second part of the thesis, we focus on an unconstrained problem

$$\text{minimize} \quad f(x), \tag{1.2}$$

where the function $f : \Re^n \mapsto \Re$ is convex. Such a minimization arises in parameter estimation, model construction, pattern classification, etc. Furthermore, such a minimization also arises when applying a penalty approach to a problem involving a convex constraint set.

We are particularly interested in poorly scaled problems, where the changes of the objective function are rapid along some directions and very slow along other directions. For such problems, standard subgradient method is very slow, since the subgradients are almost orthogonal to the directions pointing toward the set of minima. Thus, along the subgradient directions, the function improvements are insignificant and the method jamms. In this case, changing the stepsize generally cannot improve the situation, since the difficulties are associated with bad subgradient directions. However, by transforming the variables (i.e., varying the metric), we can modify subgradient directions and obtain a faster method.

1.4.2 Previous Work

Variable metric approach has traditionally been applied to poorly scaled *differentiable* problems. The idea of this approach is to rescale the original problem by varying the metric. The most common methods using this approach are diagonally scaled steepest descent method, Newton method, quasi-Newton method and its various modifications. The literature on variable metric methods for differentiable problems is vast (see, for example, the textbook by Bertsekas [Be99]).

Variable metric approach has also been used for solving *nondifferentiable* problem (1.2). The first variable metric subgradient method was proposed and studied by Shor [Sho70a], [Sho70b], [Sho77a], [Sho77b] (see also the books by Shor [Sho85] and [Sho98]). Shor suggested to use special linear transformations based on space dilation along a subgradient and along the difference of the two successive subgradients. According to some experimental results (see Lemaréchal [Lem82]), Shor's method using space dilation along the difference of the two successive subgradients is an excellent method, for which however convergence results have been established only for a stepsize with exact line search (cf. Shor [Sho85]). Based on Shor's work, Khachian [Kha79] has developed the celebrated ellipsoid method. Since then, most of the work on variable metric subgradient methods was closely related to ellipsoid method (see, for example, a review by Akgül [Akg84]). Some other variable metric subgradient methods, which are descent-based, have been proposed by Lemaréchal [Lem78] and Uryasev [Ury91].

1.4.3 Our Work

We here study two types of variable metric subgradient methods: with limited amount of space transformation and with dilation transformation. The methods rely only on the convexity property of the objective cost function, and they are not descent based. The method with limited amount of space transformation is new. Our analysis shows that this method converges for various stepsize rules. Our results are very general and include as a special case a diagonal scaling. However, there are still some open questions that need to be addressed, such as how to specifically choose the space transformations.

As for the dilation methods, we discuss dilation along subgradients and along other directions, including the direction of subgradient difference. For these methods, we propose a new dynamic stepsize rule that uses estimates of the optimal function value. The dilation method that can use dilation along directions other than subgradients is new. In a special case, this method is similar to Shor's method with dilation along subgradient difference. For this new dilation method, we establish convergence properties and convergence rate estimates using dynamic stepsize rules for known and unknown optimal function value.

1.5. OUTLINE OF THE THESIS

In the first part of the thesis, we consider the incremental approach. In Chapter 2, we formally introduce the incremental subgradient method, and we establish convergence properties and

convergence rate estimates for various stepsize rules. In Chapter 3, we propose and analyze the incremental subgradient method that uses randomization. By comparing the incremental randomized method with the nonrandomized method, both analytically and computationally, we demonstrate the advantages of randomization. We here also study the distributed asynchronous incremental method. In Chapter 4, we investigate the modifications of the incremental method involving weights and approximate subgradients (ϵ-subgradients).

In the second part of the thesis, we discuss the variable metric subgradient methods. In Chapter 5, we propose a method with limited amount of metric changes. We establish convergence properties of the method under several stepsize rules. In Chapter 6, we consider Shor's dilation method, and we also propose and analyze a new dilation method.

PART I:
Incremental Subgradient Methods

2

An Incremental Subgradient Method

Throughout the whole thesis, we view the elements of $\Re^n$ as column vectors. We use x' to denote the transpose of a vector x, and $\|\cdot\|$ to denote the standard Euclidean norm in $\Re^n$, i.e., $\|x\| = \sqrt{x'x}$. For a function $f : \Re^n \mapsto \Re$ and a constraint set $X \subset \Re^n$, we write f^* and X^* to denote, respectively, the minimum value of a f over X and the set of minima of f over X i.e.,

$$f^* = \inf_{x \in X} f(x), \qquad X^* = \big\{x \in X \mid f(x) = f^*\big\}.$$

We refer to the value f^* and the set X^*, respectively, as the optimal function value and the optimal solution set. We also write $dist(x, Y)$ to denote the distance between a vector x and a nonempty set Y, i.e.,

$$dist(x, Y) = \inf_{y \in Y} \|x - y\|.$$

In this part of the thesis, we consider the problem

$$\begin{aligned} &\text{minimize} \quad f(x) = \sum_{i=1}^{m} f_i(x) \\ &\text{subject to} \quad x \in X, \end{aligned} \tag{2.1}$$

where the function $f : \Re^n \mapsto \Re$ is convex, and the constraint set $X \subset \Re^n$ is a nonempty, closed, and convex.

2.1. THE METHOD

As mentioned earlier, a classical method for solving problem (2.1) is the subgradient method

$$x_{k+1} = \mathcal{P}_X \left[x_k - \alpha_k \sum_{i=1}^{m} d_{i,k} \right], \tag{2.2}$$

where $\mathcal{P}_X$ denotes the projection on the set X, α_k is a positive stepsize, and $d_{i,k}$ is a subgradient of f_i at x_k. In many important applications, the set X is simple enough so that the projection can be easily implemented. In particular, for the problems of the type (2.1) arising in the dual context from Lagrangian relaxation, the set X is either $\Re^n$ or the positive orthant in $\Re^n$ so that projecting on X is either not needed or not expensive.

The incremental subgradient method is similar to the standard subgradient method (2.2). The main difference is that at each iteration, x is changed incrementally, through a sequence of m steps. Each step is a subgradient iteration for a single component function f_i, and there is one step per component function. Thus, an iteration can be viewed as a cycle of m subiterations. If x_k is the vector obtained after k cycles, the vector x_{k+1} obtained after one more cycle is

$$x_{k+1} = \psi_{m,k}, \tag{2.3}$$

where $\psi_{m,k}$ is obtained after the m steps

$$\psi_{i,k} = \mathcal{P}_X \left[\psi_{i-1,k} - \alpha_k g_{i,k} \right], \qquad i = 1, \ldots, m, \tag{2.4}$$

starting with

$$\psi_{0,k} = x_k, \tag{2.5}$$

where $g_{i,k}$ is a subgradient of f_i at $\psi_{i-1,k}$. The updates described by Eq. (2.4) are referred to as the *subiterations* of the kth cycle.

Incremental subgradient methods that are somewhat different from the method (2.3)–(2.5) have been proposed by Kaskavelis and Caramanis [KaC98], and Zhao, Luh, and Wang [ZLW99]. Their methods share with ours the characteristic of computing a subgradient of only one component f_i per iteration, but differ from ours in that the direction used in an iteration is the sum of the (approximate) subgradients of all the components f_i. Thus, these methods essentially belong to the class of approximate subgradient (ϵ-subgradient) methods.

2.2. ASSUMPTIONS AND SOME BASIC RELATIONS

In our analysis here and in the subsequent chapters, we repeatedly use the defining property of a subgradient g of a convex function $h : \Re^n \mapsto \Re$ at a point x, which is

$$h(x) + g'(z - x) \leq h(z), \qquad \forall\ z \in \Re^n. \tag{2.6}$$

We denote by $\partial h(x)$ the subdifferential (set of all subgradients) of h at x.

We start with some assumptions and preliminary results that we frequently use in the forthcoming analysis. In particular, regarding the subgradients of the component functions f_i, we assume the following:

Assumption 2.1: (Subgradient Boundedness) There exists a positive scalar C such that

$$\|g\| \leq C, \qquad \forall\ g \in \partial f_i(x_k) \cup \partial f_i(\psi_{i-1,k}), \quad \forall\ i = 1, \ldots, m, \quad \forall\ k.$$

Since each component f_i is real-valued and convex over the entire space $\Re^n$, the subdifferential $\partial f_i(x)$ is nonempty and compact for all x and i. Therefore, if the set X is compact or the sequences $\{\psi_{i,k}\}$ are bounded, then Assumption 2.1 is satisfied since the set $\cup_{x \in Z} \partial f_i(x)$ is bounded for any bounded set Z (see e.g., Bertsekas [Ber99], Prop. B.24). We note that Assumption 2.1 is also satisfied if each f_i is a polyhedral function, i.e., f_i is the pointwise maximum of a finite number of affine functions. In this case, for every x, the set of subgradients $\partial f_i(x)$ is the convex hull of a finite number of vectors. In particular, often in problems (2.1) arising from Lagrangian relaxation, each f_i is a polyhedral function.

In the next lemma, we give a relation that captures the distance decrease property of the iterates generated by the incremental method (2.3)–(2.5). This lemma will play a crucial role in establishing all of our convergence results.

Lemma 2.1: Let Assumption 2.1 hold. Then, for the sequence $\{x_k\}$ generated by the incremental subgradient method, we have

$$\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k \big(f(x_k) - f(y)\big) + \alpha_k^2 m^2 C^2, \qquad \forall\ y \in X, \quad \forall\ k,$$

where C is as in Assumption 2.1.

Proof: Using the nonexpansion property of the projection, the subgradient boundedness (cf. Assumption 2.1), and the subgradient inequality (2.6) for each component function f_i, we obtain for all $y \in X$,

$$\begin{aligned}
\|\psi_{i,k} - y\|^2 &= \big\|\mathcal{P}_X[\psi_{i-1,k} - \alpha_k g_{i,k}] - y\big\|^2 \\
&\leq \|\psi_{i-1,k} - \alpha_k g_{i,k} - y\|^2 \\
&\leq \|\psi_{i-1,k} - y\|^2 - 2\alpha_k g_{i,k}'(\psi_{i-1,k} - y) + \alpha_k^2 C^2 \\
&\leq \|\psi_{i-1,k} - y\|^2 - 2\alpha_k \big(f_i(\psi_{i-1,k}) - f_i(y)\big) + \alpha_k^2 C^2, \qquad \forall\ i = 1, \ldots, m, \quad \forall\ k.
\end{aligned}$$

By adding the above inequalities over $i = 1, \ldots, m$, we see that for all $y \in X$ and k,

$$\begin{aligned}
\|x_{k+1} - y\|^2 &\leq \|x_k - y\|^2 - 2\alpha_k \sum_{i=1}^{m} \big(f_i(\psi_{i-1,k}) - f_i(y)\big) + \alpha_k^2 m C^2 \\
&= \|x_k - y\|^2 - 2\alpha_k \left(f(x_k) - f(y) + \sum_{i=1}^{m} \big(f_i(\psi_{i-1,k}) - f_i(x_k)\big)\right) + \alpha_k^2 m C^2.
\end{aligned}$$

By definition of the method [cf. Eqs. (2.3)–(2.5)] and Assumption 2.1, we have that $\|\psi_{i,k} - x_k\| \leq \alpha_k iC$ for all i and k. By using this relation, the subgradient inequality (2.6), and Assumption 2.1, we obtain for all i and k,

$$f_i(x_k) - f_i(\psi_{i-1,k}) \leq \|\tilde{g}_{i,k}\| \cdot \|\psi_{i-1,k} - x_k\| \leq C\|\psi_{i-1,k} - x_k\| \leq \alpha_k(i-1)C^2,$$

where $\tilde{g}_{i,k} \in \partial f_i(x_k)$. From this and the preceding relation, we see that for all $y \in X$ and k,

$$\begin{aligned}\|x_{k+1} - y\|^2 &\leq \|x_k - y\|^2 - 2\alpha_k\big(f(x_k) - f(y)\big) + \alpha_k^2\left(2\sum_{i=2}^{m}(i-1)C^2 + mC^2\right)\\ &= \|x_k - y\|^2 - 2\alpha_k\big(f(x_k) - f(y)\big) + \alpha_k^2 m^2 C^2.\end{aligned}$$

Q.E.D.

Among other things, Lemma 2.1 guarantees that given the current iterate x_k and some other point $y \in X$ with lower cost than x_k, the next iterate x_{k+1} will be closer to y than x_k, provided the stepsize α_k is sufficiently small [less than $2\big(f(x_k) - f(y)\big)/(mC)^2$]. We will use this fact repeatedly, with a variety of choices for y. Furthermore, when the optimal solution set X^* is nonempty, Lemma 2.1 yields a relation that plays an important role in our convergence rate analysis. This relation is given in the following lemma.

Lemma 2.2: Let Assumption 2.1 hold, and assume that the optimal solution set X^* is nonempty. Then, for the sequence $\{x_k\}$ generated by the incremental subgradient method, we have

$$\big(dist(x_{k+1}, X^*)\big)^2 \leq \big(dist(x_k, X^*)\big)^2 - 2\alpha_k\big(f(x_k) - f^*\big) + \alpha_k^2 m^2 C^2, \qquad \forall\ k. \tag{2.7}$$

Proof: Using Lemma 2.1 with $y = x^*$ for any $x^* \in X^*$, we see that

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - 2\alpha_k\big(f(x_k) - f^*\big) + \alpha_k^2 m^2 C^2, \qquad \forall\ x^* \in X^*, \quad \forall\ k,$$

and by taking the minimum over all $x^* \in X^*$, we obtain the desired relation. **Q.E.D.**

The assumption that the optimal solution set X^* is nonempty is satisfied, for example, when the constraint set X is compact. Moreover, it can be seen that this assumption is also satisfied when $\inf_{x\in X} f_i(x)$ is finite for each i, and at least one of the components f_i has bounded level sets (see Rockafellar [Roc70], Theorem 9.3).

2.3. CONSTANT STEPSIZE RULE

We here give convergence results and convergence rate estimates for the method using a constant stepsize rule. Our first result shows that, when f^* is finite, this method yields (in the

limit) a suboptimal function value with the approximation error $\alpha m^2 C^2/2$, and otherwise yields the optimal function value, as seen in the following proposition.

Proposition 2.1: Let Assumption 2.1 hold. Then, for the sequence $\{x_k\}$ generated by the incremental subgradient method with the stepsize α_k fixed to some positive constant α, we have:

(a) If $f^* = -\infty$, then

$$\liminf_{k\to\infty} f(x_k) = -\infty.$$

(b) If $f^* > -\infty$, then

$$\liminf_{k\to\infty} f(x_k) \leq f^* + \frac{\alpha m^2 C^2}{2},$$

where C is as in Assumption 2.1.

Proof: We prove (a) and (b) simultaneously. If the result does not hold, there must exist an $\epsilon > 0$ such that

$$\liminf_{k\to\infty} f(x_k) - \frac{\alpha m^2 C^2}{2} - 2\epsilon > f^*.$$

Let $\hat{y} \in X$ be such that

$$\liminf_{k\to\infty} f(x_k) \geq f(\hat{y}) + \frac{\alpha m^2 C^2}{2} + 2\epsilon,$$

and let k_0 be large enough so that for all $k \geq k_0$, we have

$$f(x_k) \geq \liminf_{k\to\infty} f(x_k) - \epsilon.$$

By combining the preceding two relations, we obtain

$$f(x_k) - f(\hat{y}) \geq \frac{\alpha m^2 C^2}{2} + \epsilon, \qquad \forall\ k \geq k_0.$$

Using Lemma 2.1, where $y = \hat{y}$, together with the preceding relation, we see that

$$\|x_{k+1} - \hat{y}\|^2 \leq \|x_k - \hat{y}\|^2 - 2\alpha\epsilon, \qquad \forall\ k \geq k_0,$$

implying that

$$\|x_{k+1} - \hat{y}\|^2 \leq \|x_k - \hat{y}\|^2 - 2\alpha\epsilon \leq \|x_{k-1} - \hat{y}\|^2 - 4\alpha\epsilon \leq \cdots \leq \|x_{k_0} - \hat{y}\|^2 - 2(k+1-k_0)\alpha\epsilon,$$

which cannot hold for k sufficiently large, a contradiction. **Q.E.D.**

The preceding bound is sharp within a constant, i.e., there exists a problem such that for any stepsize α, we can choose an initial point x_0 and a processing order for components f_i so that

$$\liminf_{k\to\infty} f(x_k) = f^* + \frac{1}{16}\frac{\alpha m^2 C^2}{2}, \qquad \forall\ i = 1, \ldots, m.$$

This is shown in the following example.

Example 2.1:

Consider the following problem

$$\begin{aligned} &\text{minimize} \quad f(x_1,x_2)=\sum_{i=1}^{p}\Big(|x_1+1|+2|x_1|+|x_1-1|+|x_2+1|+2|x_2|+|x_2-1|\Big)\\ &\text{subject to} \quad (x_1,x_2)\in\Re^2, \end{aligned}$$

where p is a positive integer. The optimal value is $f^*=4p$ and is attained at the point $(x_1^*,x_2^*)=(0,0)$, which is the only optimal solution. Given any positive stepsize α, we choose $(\alpha p,0)$ as an initial point, where $\alpha p\leq 1$, and we process the component functions of f in the following order: p components of the form $|x_2-1|$, p components of the form $|x_1|$ followed by p components of the form $|x_1+1|$, p components of the form $|x_2|$ followed by p components of the form $|x_2+1|$, then p components of the form $|x_1|$ followed by p components of the form $|x_1-1|$, and then the remaining p components of the form $|x_2|$. It can be seen that this processing order produces the iterates x_k such that

$$f(x_k)=f^*+2\alpha p^2, \qquad \forall\ i=1,\ldots,m, \quad \forall\ k.$$

Since $m=8p$ and $C=1$, it follows that

$$f^*+2\alpha p^2=f^*+\frac{1}{16}\frac{\alpha m^2C^2}{2},$$

which together with the preceding relation implies that

$$\lim_{k\to\infty}f(x_k)=f^*+\frac{1}{16}\frac{\alpha m^2C^2}{2}, \qquad \forall\ i=1,\ldots,m.$$

Furthermore, it can be seen that, even when subiterates $\psi_{i,k}$ are considered, the function values $f(\psi_{i,k})$ cannot attain values lower than this bound.

We next estimate the number K of iterations needed to have

$$\min_{0\leq k\leq K}f(x_k)\leq f^*+\frac{\alpha m^2C^2+\epsilon}{2},$$

where ϵ is a given error tolerance. In particular, we have the following result.

Proposition 2.2: Let Assumption 2.1 hold, and assume that the optimal solution set X^* is nonempty. Let the sequence $\{x_k\}$ be generated by the incremental subgradient method with the stepsize α_k fixed to some positive constant α. Then, for a positive scalar ϵ and the nonnegative integer K given by

$$K=\left\lfloor\frac{1}{\alpha\epsilon}\big(dist(x_0,X^*)\big)^2\right\rfloor,$$

we have

$$\min_{0\leq k\leq K}f(x_k)\leq f^*+\frac{\alpha m^2C^2+\epsilon}{2}. \tag{2.8}$$

Proof: To arrive at a contradiction, assume that

$$f(x_k) > f^* + \frac{\alpha m^2 C^2 + \epsilon}{2}, \qquad \forall\ k = 0, 1, \ldots, K.$$

By using this relation and Lemma 2.2, where α_k is replaced by α, we obtain

$$\begin{aligned}\big(dist(x_{k+1}, X^*)\big)^2 &\leq \big(dist(x_k, X^*)\big)^2 - 2\alpha\big(f(x_k) - f^*\big) + \alpha^2 m^2 C^2 \\ &\leq \big(dist(x_k, X^*)\big)^2 - (\alpha^2 m^2 C^2 + \alpha\epsilon) + \alpha^2 m^2 C^2 \\ &= \big(dist(x_k, X^*)\big)^2 - \alpha\epsilon, \qquad \forall\ k = 0, 1, \ldots, K.\end{aligned}$$

Adding the above inequalities over $k = 0, 1, \ldots, K$ yields

$$\big(dist(x_{K+1}, X^*)\big)^2 \leq \big(dist(x_0, X^*)\big)^2 - (K+1)\alpha\epsilon,$$

implying that

$$(K+1)\alpha\epsilon \leq \big(dist(x_0, X^*)\big)^2,$$

which contradicts the definition of K. **Q.E.D.**

The result of Prop. 2.2 indicates that a higher accuracy in the computation of the optimal function value f^*, (i.e., small stepsize α and tolerance level ϵ) corresponds to a larger number K. This is quite natural since higher accuracy usually requires more work.

We next show that, for a function f with sharp minima, the convergence rate of the method is linear for a sufficiently small stepsize. However, only convergence to a neighborhood of the optimal solution set X^* can be guaranteed.

Proposition 2.3: Let Assumption 2.1 hold, and assume that the optimal solution set X^* is nonempty. Assume further that there exists a positive scalar μ such that

$$f(x) - f^* \geq \mu\ \big(dist(x, X^*)\big)^2, \qquad \forall\ x \in X. \tag{2.9}$$

Then, for the sequence $\{x_k\}$ generated by the incremental subgradient method with the stepsize α_k fixed to some positive constant α, where $\alpha \leq \frac{1}{2\mu}$, we have

$$\big(dist(x_{k+1}, X^*)\big)^2 \leq (1 - 2\alpha\mu)^{k+1}\big(dist(x_0, X^*)\big)^2 + \frac{\alpha m^2 C^2}{2\mu}, \qquad \forall\ k.$$

Proof: By using Eq. (2.9) and Lemma 2.2 with α_k replaced by α, we obtain

$$\begin{aligned}\big(dist(x_{k+1}, X^*)\big)^2 &\leq \big(dist(x_k, X^*)\big)^2 - 2\alpha\big(f(x_k) - f^*\big) + \alpha^2 m^2 C^2 \\ &\leq (1 - 2\alpha\mu)\big(dist(x_k, X^*)\big)^2 + \alpha^2 m^2 C^2, \qquad \forall\ k.\end{aligned}$$

From this relation, by induction, we can see that

$$\big(dist(x_{k+1}, X^*)\big)^2 \leq (1-2\alpha\mu)^{k+1}\big(dist(x_0, X^*)\big)^2 + \alpha^2 m^2 C^2 \sum_{j=0}^{k}(1-2\alpha\mu)^j, \qquad \forall\ k,$$

which combined with

$$\sum_{j=0}^{k}(1-2\alpha\mu)^j \leq \frac{1}{2\alpha\mu}, \qquad \forall\ k,$$

yields the desired relation. **Q.E.D.**

2.4. DIMINISHING STEPSIZE RULE

We here consider the method that employs a diminishing stepsize. Our first result is a generalization of a classical convergence result for the ordinary subgradient method, which was obtained by Ermoliev [Erm66] and Polyak [Pol67], independently (see also Correa and Lemaréchal [CoL93]).

Proposition 2.4: Let Assumption 2.1 hold, and let the stepsize α_k be such that

$$\lim_{k\to\infty}\alpha_k = 0, \qquad \sum_{k=0}^{\infty}\alpha_k = \infty.$$

Then, for the sequence $\{x_k\}$ generated by the incremental method, we have

$$\liminf_{k\to\infty} f(x_k) = f^*.$$

Proof: Suppose to arrive at a contradiction that there exists an $\epsilon > 0$ such that

$$\liminf_{k\to\infty} f(x_k) + 2\epsilon > f^*.$$

Let $\hat{y} \in X$ be such that

$$\liminf_{k\to\infty} f(x_k) \geq f(\hat{y}) + 2\epsilon,$$

and let k_0 be large enough so that for all $k \geq k_0$, we have

$$f(x_k) \geq \liminf_{k\to\infty} f(x_k) - \epsilon.$$

From the preceding two relations it follows that

$$f(x_k) - f(\hat{y}) \geq \epsilon, \qquad \forall\ k \geq k_0.$$

Using Lemma 2.1, where $y = \hat{y}$, together with the preceding relation, we obtain

$$\|x_{k+1} - \hat{y}\|^2 \leq \|x_k - \hat{y}\|^2 - \alpha_k(2\epsilon - \alpha_k m^2 C^2), \qquad \forall\ k \geq k_0.$$

Because $\alpha_k \to 0$, without loss of generality, we may assume that k_0 is large enough so that $\epsilon \geq \alpha_k m^2 C^2$ for all $k \geq k_0$, implying that

$$\|x_{k+1} - \hat{y}\|^2 \leq \|x_k - \hat{y}\|^2 - \alpha_k \epsilon \leq \|x_{k-1} - \hat{y}\|^2 - \epsilon(\alpha_{k-1} + \alpha_k) \leq \cdots \leq \|x_{k_0} - \hat{y}\|^2 - \epsilon \sum_{j=k_0}^{k} \alpha_j.$$

Since $\sum_{k=0}^{\infty} \alpha_k = \infty$, this relation cannot hold for k sufficiently large, a contradiction. **Q.E.D.**

If we assume in addition that X^* is nonempty and bounded, the result of Prop. 2.4 can be strengthened, as seen in the forthcoming proposition. This proposition is also an extension of the convergence result obtained by Solodov and Zavriev [SoZ98], which was proved by a different line of analysis using the stronger assumption that X is a compact set.

Proposition 2.5: Let Assumption 2.1 hold, and assume that the optimal solution set X^* is nonempty and bounded. Assume further that the stepsize α_k is such that

$$\lim_{k \to \infty} \alpha_k = 0, \qquad \sum_{k=0}^{\infty} \alpha_k = \infty.$$

Then, for the sequence $\{x_k\}$ generated by the incremental subgradient method, we have

$$\lim_{k \to \infty} dist(x_k, X^*) = 0, \qquad \lim_{k \to \infty} f(x_k) = f^*.$$

Proof: The idea is to show that once x_k enters a certain level set, it cannot get too far out of that set. Fix an arbitrary $\gamma > 0$, and let k_0 be such that $\gamma \geq \alpha_k m^2 C^2$ for all $k \geq k_0$. We consider k for $k \geq k_0$ and we distinguish two cases:

Case 1: $f(x_k) > f^* + \gamma$. From Lemma 2.1 we obtain

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - 2\alpha_k \big(f(x_k) - f^*\big) + \alpha_k^2 m^2 C^2, \qquad \forall\ x^* \in X^*,$$

Hence,

$$\|x_{k+1} - x^*\|^2 < \|x_k - x^*\|^2 - 2\gamma\alpha_k + \alpha_k^2 m^2 C^2 \leq \|x_k - x^*\|^2 - \alpha_k \gamma, \qquad \forall\ x^* \in X^*,$$

so that

$$dist(x_{k+1}, X^*) \leq dist(x_k, X^*) - \alpha_k \gamma. \tag{2.10}$$

Case 2: $f(x_k) \leq f^* + \gamma$. This case must occur for infinitely many k, in view of Eq. (2.10) and the relation $\sum_{k=0}^{\infty} \alpha_k = \infty$. Since x_k belongs to the level set

$$L_\gamma = \big\{y \in X \mid f(y) \leq f^* + \gamma\big\},$$

which is bounded (in view of the boundedness of X^*), we have

$$dist(x_k, X^*) \leq d(\gamma) < \infty, \tag{2.11}$$

where we denote

$$d(\gamma) = \max_{y \in L_\gamma} dist(y, X^*).$$

From the iteration (2.3)–(2.5), we have $\|x_{k+1} - x_k\| \leq \alpha_k mC$ for any k, so that

$$\|x_{k+1} - x^*\| \leq \|x_k - x^*\| + \|x_{k+1} - x_k\| \leq \|x_k - x^*\| + \alpha_k mC, \qquad \forall\ x^* \in X^*.$$

By taking the minimum in this relation over $x^* \in X^*$ and by using Eq. (2.11), we obtain

$$dist(x_{k+1}, X^*) \leq d(\gamma) + \alpha_k mC. \tag{2.12}$$

Combining Eq. (2.10) which holds when $f(x_k) > f^* + \gamma$ (Case 1 above), with Eq. (2.12) which holds for the infinitely many k such that $f(x_k) \leq f^* + \gamma$ (Case 2 above), we see that

$$dist(x_k, X^*) \leq d(\gamma) + \alpha_k mC, \qquad \forall\ k \geq k_0.$$

Therefore, because $\alpha_k \to 0$, we have

$$\limsup_{k \to \infty} dist(x_k, X^*) \leq d(\gamma), \qquad \forall\ \gamma > 0.$$

Since in view of the continuity of f and the compactness of its level sets, we have that

$$\lim_{\gamma \to 0} d(\gamma) = 0,$$

it follows that $dist(x_k, X^*) \to 0$. This relation also implies that $f(x_k) \to f^*$. **Q.E.D.**

The assumption that X^* is nonempty and bounded holds, for example, when $\inf_{x \in X} f_i(x)$ is finite for all i, and at least one of the components f_i has bounded level sets (cf. Rockafellar [Roc70], Theorem 9.3). Prop. 2.5 does not guarantee convergence of the entire sequence $\{x_k\}$. However, with slightly different assumptions that include an additional mild restriction on the stepsize sequence, this convergence is guaranteed, as shown in the following proposition.

Proposition 2.6: Let Assumption 2.1 hold, and assume that the optimal solution set X^* is nonempty. Assume further that the stepsize α_k is such that

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

Then, the sequence $\{x_k\}$ generated by the incremental subgradient method converges to some optimal solution.

Proof: By Lemma 2.1, where $y = x^*$ with $x^* \in X^*$, we have

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - 2\alpha_k \big(f(x_k) - f^*\big) + \alpha_k^2 m^2 C^2, \qquad \forall\ x^* \in X^*, \quad \forall\ k. \tag{2.13}$$

Since $f(x_k) - f^* \geq 0$ for all k and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, it follows that the sequence $\{x_k\}$ is bounded. Furthermore, by Prop. 2.4, we have that

$$\liminf_{k\to\infty} f(x_k) = f^*.$$

Let $\{x_{k_j}\}$ be a subsequence of $\{x_k\}$ along which the above liminf is attained, so that

$$\lim_{j\to\infty} f(x_{k_j}) = f^*. \tag{2.14}$$

The sequence $\{x_{k_j}\}$ is bounded, so it has limit points. Let $\overline{x}$ be one of them, and without loss of generality we can assume that $x_{k_j} \to \overline{x}$. By continuity of f and Eq. (2.14), we have that $\overline{x} \in X^*$, so from Eq. (2.13) with $x^* = \overline{x}$, we obtain for any j and any $k \geq k_j$,

$$\|x_{k+1} - \overline{x}\|^2 \leq \|x_k - \overline{x}\|^2 + \alpha_k^2 m^2 C^2 \leq \cdots \leq \|x_{k_j} - \overline{x}\|^2 + m^2 C^2 \sum_{i=k_j}^{k} \alpha_i^2.$$

Taking first the limit as $k \to \infty$ and then the limit as $j \to \infty$, from the preceding relation we obtain

$$\limsup_{k\to\infty} \|x_{k+1} - \overline{x}\|^2 \leq \lim_{j\to\infty} \|x_{k_j} - \overline{x}\|^2 + m^2 C^2 \lim_{j\to\infty} \sum_{i=k_j}^{\infty} \alpha_i^2,$$

which by $x_{k_j} \to \overline{x}$ and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, implies that

$$\limsup_{k\to\infty} \|x_{k+1} - \overline{x}\|^2 = 0,$$

and consequently, $x_k \to \overline{x}$ with $\overline{x} \in X^*$. **Q.E.D.**

In Props. 2.4–2.6, we use the same stepsize α_k in all subiterations of a cycle. As shown by Kibardin in [Kib79] the convergence can be preserved if we vary the stepsize α_k within each cycle, provided that the variations of α_k within a cycle are suitably small. We will see this later on in Section 3.9 for a more general incremental method.

For a function f with a sharp minima, the convergence rate of the incremental subgradient method using the stepsize $\alpha_k = r/(k+1)$, with a positive scalar r, is sublinear. This convergence rate estimate is given by Nedić and Bertsekas in [NeB01b].

2.5. DYNAMIC STEPSIZE RULE FOR KNOWN f^*

To this end, we analyzed the constant and the diminishing stepsize choices. An interesting alternative for the ordinary subgradient method is the dynamic stepsize rule

$$\alpha_k = \gamma_k \frac{f(x_k) - f^*}{\|g_k\|^2}, \qquad 0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2,$$

where g_k is a subgradient of f at x_k. This stepsize rule was introduced by Polyak in [Pol69] (see also discussions in Shor [Sho85], [Sho98], Brännlund [Brä93], and Bertsekas [Ber99]). Clearly, this stepsize rule can be used only if f^* is finite, which we assume throughout this section. We would like to use such a stepsize rule for the incremental method, so we propose a variant of this stepsize where $\|g_k\|$ is replaced by an upper bound mC:

$$\alpha_k = \gamma_k \frac{f(x_k) - f^*}{m^2 C^2}, \qquad 0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2, \tag{2.15}$$

where C is as in Assumption 2.1. For this choice of stepsize, we have to be able to calculate the upper bound C, which can be done, for example, when the components f_i are polyhedral.

We first consider the case where f^* is known, and later we modify the stepsize so that f^* can be replaced by a dynamically updated estimate. Throughout this section, we assume that f^* is finite, which is evidently needed to use the stepsize (2.15). We next give convergence and convergence rate results for the method employing this stepsize.

Proposition 2.7: Let Assumption 2.1 hold. Then, for the sequence $\{x_k\}$ generated by the incremental subgradient method with the dynamic stepsize rule (2.15), we have

$$\liminf_{k \to \infty} f(x_k) = f^*.$$

Proof: Suppose to obtain a contradiction that there exists $\epsilon > 0$ such that

$$f^* + \frac{2\epsilon}{2 - \overline{\gamma} - \delta} < \liminf_{k \to \infty} f(x_k),$$

where $\delta > 0$ is a positive scalar such that $2 - \overline{\gamma} - \delta > 0$. Let k_0 be large enough so that

$$\frac{2\epsilon}{2 - \overline{\gamma} - \delta} \leq f(x_k) - f^*, \qquad \forall\ k \geq k_0, \tag{2.16}$$

and let a vector $\hat{y} \in X$ be such that

$$f(\hat{y}) - f^* \leq \epsilon.$$

Then, we have

$$f(\hat{y}) - f^* \leq \frac{2 - \overline{\gamma} - \delta}{2} \big(f(x_k) - f^*\big), \qquad \forall\ k \geq k_0. \tag{2.17}$$

By Lemma 2.1, where $y = \hat{y}$ and α_k is given by Eq. (2.15), it follows that for all $k \geq k_0$,

$$\begin{aligned}\|x_{k+1} - \hat{y}\|^2 &\leq \|x_k - \hat{y}\|^2 - \alpha_k \Big(2\big(f(x_k) - f(\hat{y})\big) - \gamma_k \big(f(x_k) - f^*\big)\Big) \\ &= \|x_k - \hat{y}\|^2 - \alpha_k \Big(2\big(f(x_k) - f^*\big) - 2\big(f(\hat{y}) - f^*\big) - \gamma_k \big(f(x_k) - f^*\big)\Big).\end{aligned}$$

By using Eq. (2.17) in this relation, we see that

$$\|x_{k+1} - \hat{y}\|^2 \leq \|x_k - \hat{y}\|^2 - \alpha_k(\overline{\gamma} + \delta - \gamma_k)\big(f(x_k) - f^*\big), \qquad \forall\ k \geq k_0.$$

Using the definition of the stepsize [cf. Eq. (2.15)] and Eq. (2.16), from the preceding inequality we obtain for all $k \geq k_0$,

$$\|x_{k+1} - \hat{y}\|^2 \leq \|x_k - \hat{y}\|^2 - \frac{4\underline{\gamma}\delta\epsilon^2}{m^2C^2(2 - \overline{\gamma} - \delta)^2} \leq \cdots \leq \|x_{k_0} - \hat{y}\|^2 - \frac{(k + 1 - k_0)4\underline{\gamma}\delta\epsilon^2}{m^2C^2(2 - \overline{\gamma} - \delta)^2},$$

which cannot hold for k sufficiently large, a contradiction. **Q.E.D.**

When the optimal solution set X^* is nonempty, the method converges to an optimal solution, which we show in the next proposition.

Proposition 2.8: Let Assumption 2.1 hold, and assume that the optimal solution set X^* is nonempty. Then, the sequence $\{x_k\}$ generated by the incremental subgradient method with the dynamic stepsize rule (2.15) converges to some optimal solution.

Proof: From Lemma 2.1 with $y = x^* \in X^*$, we have

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - 2\alpha_k\big(f(x_k) - f^*\big) + \alpha_k^2 m^2 C^2, \qquad \forall\ x^* \in X^*, \quad \forall\ k,$$

and by using the definition of α_k, we obtain

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \underline{\gamma}(2 - \overline{\gamma})\frac{\big(f(x_k) - f^*\big)^2}{m^2C^2}, \qquad \forall\ x^* \in X^*, \quad \forall\ k.$$

This implies that $\{x_k\}$ is bounded. Furthermore, $f(x_k) \to f^*$, since otherwise we would have $\|x_{k+1} - x^*\| \leq \|x_k - x^*\| - \epsilon$ for some suitably small $\epsilon > 0$ and infinitely many k. Hence, for any limit point $\overline{x}$ of $\{x_k\}$, we have $\overline{x} \in X^*$, and since the sequence $\big\{\|x_k - x^*\|\big\}$ is decreasing, it converges to $\|\overline{x} - x^*\|$ for every $x^* \in X^*$. If there are two limit points $\tilde{x}$ and $\overline{x}$ of $\{x_k\}$, we must have $\tilde{x} \in X^*$, $\overline{x} \in X^*$, and $\|\tilde{x} - x^*\| = \|\overline{x} - x^*\|$ for all $x^* \in X^*$, which is possible only if $\tilde{x} = \overline{x}$. **Q.E.D.**

We next give several convergence rate estimates for the method with the dynamic stepsize. In the next proposition, we present an asymptotic estimate for convergence rate of $f(x_k)$, which extends the estimate for the ordinary subgradient method given by Polyak in [Pol87], Theorem 2, p. 142. In the same proposition, we also estimate the number K of cycles required for

$$\min_{0 \leq k \leq K} f(x_k) \leq f^* + \epsilon$$

to hold, where the scalar ϵ is a prescribed error tolerance.

Proposition 2.9: Let Assumption 2.1 hold, and assume that the optimal solution set X^* is nonempty. Let the sequence $\{x_k\}$ be generated by the incremental subgradient method with the dynamic stepsize (2.15). Then, the following hold:

(a) We have
$$\liminf_{k\to\infty} \sqrt{k}\big(f(x_k) - f^*\big) = 0.$$

(b) For a positive scalar ϵ and the nonnegative integer K given by
$$K = \left\lfloor \frac{m^2 C^2}{\epsilon^2 \underline{\gamma}(2-\overline{\gamma})} \big(dist(x_0, X^*)\big)^2 \right\rfloor,$$
we have
$$\min_{0\le k\le K} f(x_k) \le f^* + \epsilon.$$

Proof: (a) Assume, to arrive at a contradiction, that $\liminf_{k\to\infty} \sqrt{k}\big(f(x_k) - f^*\big) = 2\beta$ for some $\beta > 0$. Then, for k_0 large enough, we have $f(x_k) - f^* \ge \frac{\beta}{\sqrt{k}}$ for all $k \ge k_0$, so that
$$\sum_{k=k_0}^{\infty} \big(f(x_k) - f^*\big)^2 \ge \beta^2 \sum_{k=k_0}^{\infty} \frac{1}{k} = \infty. \tag{2.18}$$

On the other hand, by using the definition of α_k and Lemma 2.2, for all $k \ge k_0$, we obtain
$$\big(dist(x_{k+1}, X^*)\big)^2 \le \big(dist(x_k, X^*)\big)^2 - \gamma_k(2-\gamma_k)\frac{\big(f(x_k) - f^*\big)^2}{m^2 C^2}, \tag{2.19}$$

so that
$$\sum_{k=0}^{\infty} \big(f(x_k) - f^*\big)^2 < \infty,$$

contradicting Eq. (2.18). Hence, we must have $\liminf_{k\to\infty} \sqrt{k}\big(f(x_k) - f^*\big) = 0$.

(b) To arrive at a contradiction, assume that
$$f(x_k) - f^* > \epsilon, \qquad \forall\ k = 0, 1, \ldots, K.$$

By using this relation in Eq. (2.19), since $\gamma_k \in [\underline{\gamma}, \overline{\gamma}]$ for all k, we obtain
$$\big(dist(x_{k+1}, X^*)\big)^2 \le \big(dist(x_k, X^*)\big)^2 - \underline{\gamma}(2-\overline{\gamma})\frac{\epsilon^2}{m^2 C^2}, \qquad \forall\ k = 0, 1, \ldots, K.$$

Adding these inequalities over $k = 0, 1, \ldots, K$ yields
$$\big(dist(x_{K+1}, X^*)\big)^2 \le \big(dist(x_0, X^*)\big)^2 - (K+1)\frac{\epsilon^2 \underline{\gamma}(2-\overline{\gamma})}{m^2 C^2},$$

implying that

$$(K+1)\frac{\epsilon^2\underline{\gamma}(2-\overline{\gamma})}{m^2C^2} \leq \big(dist(x_0, X^*)\big)^2,$$

contradicting the definition of K. **Q.E.D.**

Note that the bound on number K in Prop. 2.9(b), viewed as a function of $\underline{\gamma}$ and $\overline{\gamma}$, is smallest when $\underline{\gamma} = \overline{\gamma} = 1$. Note also that, for a practical use of this bound, we need some additional information about f or X such as, for example, an upper bound on $dist(x_0, X^*)$.

Under some additional assumptions on f, we can obtain some different types of estimate of the convergence rate for the method with the dynamic stepsize. In deriving these estimates, we use the following result given by Polyak [Pol87], Lemma 6, p. 46.

Lemma 2.3: Let $\{u_k\}$ be a sequence of positive scalars satisfying

$$u_{k+1} \leq u_k - \beta_k u_k^{1+p}, \qquad \forall\ k,$$

where β_k are nonnegative scalars and p is a positive scalar. Then, we have

$$u_k \leq u_0 \left(1 + pu_0^p \sum_{j=0}^{k-1} \beta_j\right)^{-\frac{1}{p}}, \qquad \forall\ k.$$

In particular, if $\beta_k = \beta$ for all k, then

$$u_k \leq u_0(1 + pu_0^p\beta k)^{-\frac{1}{p}}, \qquad \forall\ k.$$

By using this lemma, we have the following.

Proposition 2.10: Let Assumption 2.1 hold, and assume that the optimal solution set X^* is nonempty. Let the sequence $\{x_k\}$ be generated by the incremental subgradient method with the dynamic stepsize (2.15). Then, the following hold:

(a) If for some positive scalar μ, the function f satisfies

$$f(x) - f^* \geq \mu\ dist(x, X^*), \qquad \forall\ x \in X,$$

we have

$$dist(x_k, X^*) \leq q^k\ dist(x_0, X^*), \qquad \forall\ k,$$

where

$$q = \sqrt{1 - \underline{\gamma}(2-\overline{\gamma})\frac{\mu^2}{m^2C^2}}.$$

(b) If for some positive scalars μ and p, the function f satisfies

$$f(x) - f^* \geq \mu\ \big(dist(x, X^*)\big)^{1+p}, \qquad \forall\ x \in X,$$

we have

$$dist(x_k, X^*) \leq \frac{dist(x_0, X^*)}{\big(1 + \overline{C}k\big)^{\frac{1}{2p}}}, \qquad \forall\ k,$$

where

$$\overline{C} = p\underline{\gamma}(2 - \overline{\gamma})\frac{\mu^2}{m^2C^2}\big(dist(x_0, X^*)\big)^{2p}.$$

Proof: (a) By applying Lemma 2.2 with α_k as in Eq. (2.15) and by using the given property of f, we obtain

$$\big(dist(x_{k+1}, X^*)\big)^2 \leq \left(1 - \underline{\gamma}(2 - \overline{\gamma})\frac{\mu^2}{m^2C^2}\right)\big(dist(x_k, X^*)\big)^2, \qquad \forall\ k,$$

from which, by induction, the desired relation follows.

(b) Similar to part (a), from Lemma 2.2 with α_k as in Eq. (2.15) and the given property of f, we obtain

$$\big(dist(x_{k+1}, X^*)\big)^2 \leq \big(dist(x_k, X^*)\big)^2 - \underline{\gamma}(2 - \overline{\gamma})\frac{\mu^2}{m^2C^2}\big(dist(x_k, X^*)\big)^{2(1+p)}, \qquad \forall\ k.$$

By denoting $u_k = (dist(x_k, X^*))^2$, we can rewrite the preceding inequality as follows

$$u_{k+1} \leq u_k - \beta u_k^{1+p}, \qquad \forall\ k,$$

where $\beta = \underline{\gamma}(2 - \overline{\gamma})(\mu/mC)^2$. By Lemma 2.3, we have that

$$u_k \leq \frac{u_0}{\big(1 + kp\beta u_0^p\big)^{\frac{1}{p}}}, \qquad \forall\ k,$$

and by using $u_k = \big(dist(x_k, X^*)\big)^2$ and $\beta = \underline{\gamma}(2 - \overline{\gamma})(\mu/mC)^2$, we obtain the desired relation. **Q.E.D.**

2.6. DYNAMIC STEPSIZE RULE FOR UNKNOWN f^*

In most practical problems the value f^* is not known. In this case, a popular modification of the dynamic stepsize rule for the ordinary subgradient method is

$$\alpha_k = \gamma_k \frac{f(x_k) - f_k^{\rm lev}}{\|g_k\|^2}, \qquad 0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2, \tag{2.20}$$

where $g_k \in \partial f(x_k)$, and $f_k^{\rm lev}$ is an estimate of f^* often referred to as a *target level*. This stepsize with a constant target level (i.e., $f_k^{\rm lev} = w$ for some scalar $w > 0$ and all k) was

first proposed by Polyak in [Pol69], and further analyzed by Allen, Helgason, Kennington, and Shetty [AHK87], and Kim and Um [KiU93]. The adjustment procedures for the target level $f_k^{\rm lev}$ in Eq. (2.20) that require knowledge of an underestimate of f^* are presented in Bazaraa and Sherali [BaS81], Kim, Ahn, and Cho [KAC91], Kiwiel [Kiw96a], [Kiw96b]. The procedures for $f_k^{\rm lev}$ that do not require any additional information about f^* are considered in Bertsekas [Ber99], Brännlund [Brä93], Goffin and Kiwiel [GoK99], Kiwiel, Larsson, and Lindberg [KLL98], Kulikov and Fazylov [KuF90].

Here, we consider a modification of the stepsize (2.20) of the form

$$\alpha_k = \gamma_k \frac{f(x_k) - f_k^{\rm lev}}{m^2 C^2}, \qquad 0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2. \tag{2.21}$$

We discuss two procedures for updating $f_k^{\rm lev}$ that do not require any additional information about f^*. In both procedures, the estimate $f_k^{\rm lev}$ is equal to the best function value achieved up to the kth iteration, i.e., $\min_{0\leq j\leq k} f(x_j)$, minus a positive amount δ_k which is adjusted based on the algorithm's progress. The first adjustment procedure is new, even when specialized to the ordinary subgradient method. This procedure is simple but is only guaranteed to yield a δ-optimal objective function value with δ positive and arbitrarily small (unless $f^* = -\infty$ in which case the procedure yields the optimal function value). The second adjustment procedure for $f_k^{\rm lev}$ is more complex but is guaranteed to yield the optimal value f^* in the limit.

In the first adjustment procedure, $f_k^{\rm lev}$ is given by

$$f_k^{\rm lev} = \min_{0\leq j\leq k} f(x_j) - \delta_k, \tag{2.22}$$

and δ_k is updated according to

$$\delta_{k+1} = \begin{cases} \lambda\delta_k & \text{if } f(x_{k+1}) \leq f_k^{\rm lev}, \\ \max\{\beta\delta_k, \delta\} & \text{if } f(x_{k+1}) > f_k^{\rm lev}, \end{cases} \tag{2.23}$$

where δ_0, δ, β, and λ are fixed positive constants with $\beta < 1$ and $\lambda \geq 1$. Thus, in this procedure, we essentially "aspire" to reach a target level that is smaller by δ_k over the best value achieved thus far. Whenever the target level is achieved, we increase δ_k or we keep it at the same value depending on the choice of ρ. If the target level is not attained at a given iteration, δ_k is reduced up to a threshold δ. This threshold guarantees that the stepsize α_k of Eq. (2.21) is bounded away from zero, since by Eq. (2.22), we have $f(x_k) - f_k^{\rm lev} \geq \delta$ and hence

$$\alpha_k \geq \underline{\gamma}\, \frac{\delta}{m^2 C^2}.$$

As a result, the method behaves similar to the one with a constant stepsize (cf. Prop. 2.1), as seen in the following proposition.

Proposition 2.11: Let Assumption 2.1 hold. Then, for the sequence $\{x_k\}$ generated by the incremental method and the dynamic stepsize rule (2.21)–(2.23), we have:

(a) If $f^* = -\infty$, then

$$\inf_{k\geq 0} f(x_k) = f^*.$$

(b) If $f^* > -\infty$, then

$$\inf_{k\geq 0} f(x_k) \leq f^* + \delta.$$

Proof: To arrive at a contradiction, assume that

$$\inf_{k\geq 0} f(x_k) > f^* + \delta. \tag{2.24}$$

Each time the target level is attained [i.e., $f(x_k) \leq f_{k-1}^{\rm lev}$], the current best function value $\min_{0\leq j\leq k} f(x_j)$ decreases by at least δ [cf. Eqs. (2.22) and (2.23)], so in view of Eq. (2.24), the target value can be attained only a finite number times. From Eq. (2.23) it follows that after finitely many iterations, δ_k is decreased to the threshold value and remains at that value for all subsequent iterations, i.e., there is an index $\overline{k}$ such that

$$\delta_k = \delta, \qquad \forall\ k \geq \overline{k}. \tag{2.25}$$

In view of Eq. (2.24), there exists $\overline{y} \in X$ such that $\inf_{k\geq 0} f(x_k) - \delta \geq f(\overline{y})$, so that by Eqs. (2.22) and (2.25), we have

$$f_k^{\rm lev} = \min_{0\leq j\leq k} f(x_j) - \delta \geq \inf_{k\geq 0} f(x_k) - \delta \geq f(\overline{y}), \qquad \forall\ k \geq \overline{k},$$

implying that

$$\alpha_k\big(f(x_k) - f(\overline{y})\big) \geq \alpha_k\big(f(x_k) - f_k^{\rm lev}\big) = \gamma_k \left(\frac{f(x_k) - f_k^{\rm lev}}{mC}\right)^2, \qquad \forall\ k \geq \overline{k}.$$

By Lemma 2.1 with $y = \overline{y}$, it follows that

$$\|x_{k+1} - \overline{y}\|^2 \leq \|x_k - \overline{y}\|^2 - 2\alpha_k\big(f(x_k) - f(\overline{y})\big) + \alpha_k^2 m^2 C^2, \qquad \forall\ k \geq 0.$$

Using the preceding two relations and the definition of α_k [cf. Eq. (2.21)], we obtain

$$\begin{aligned}
\|x_{k+1} - \overline{y}\|^2 &\leq \|x_k - \overline{y}\|^2 - 2\gamma_k \left(\frac{f(x_k) - f_k^{\rm lev}}{mC}\right)^2 + \gamma_k^2 \left(\frac{f(x_k) - f_k^{\rm lev}}{mC}\right)^2 \\
&= \|x_k - \overline{y}\|^2 - \gamma_k(2 - \gamma_k) \left(\frac{f(x_k) - f_k^{\rm lev}}{mC}\right)^2 \\
&\leq \|x_k - \overline{y}\|^2 - \underline{\gamma}(2 - \overline{\gamma}) \frac{\delta^2}{m^2 C^2}, \qquad \forall\ k \geq \overline{k},
\end{aligned}$$

where the last inequality follows from the relations $\gamma_k \in [\underline{\gamma}, \overline{\gamma}]$ and $f(x_k) - f_k^{\rm lev} \geq \delta$ for all k. Finally, by adding the above inequalities over k, we see that

$$\|x_{k+1} - \overline{y}\|^2 \leq \|x_{\overline{k}} - \overline{y}\|^2 - (k+1-\overline{k})\underline{\gamma}(2-\overline{\gamma})\frac{\delta^2}{m^2 C^2}, \qquad \forall\ k \geq \overline{k},$$

which cannot hold for large k, a contradiction. **Q.E.D.**

When $m = 1$, the incremental subgradient method (2.3)–(2.5) becomes the ordinary subgradient method

$$x_{k+1} = \mathcal{P}_X[x_k - \alpha_k g_k], \qquad \forall\ k.$$

Even for this method, the dynamic stepsize rule (2.21) with the adjustment procedure (2.22)–(2.23) (where mC is replaced by $\|g_k\|$), and the convergence result of Prop. 2.11 are new.

We now consider another procedure for adjusting $f_k^{\rm lev}$, which guarantees that $f_k^{\rm lev} \to f^*$, and convergence of the associated method to the optimum. This procedure is based on the ideas and algorithms of Brännlund [Brä93], and Goffin and Kiwiel [GoK99], where the parameter δ_k is reduced whenever the iterates "travel" a long distance without reaching the corresponding target level. The incremental method using such a procedure is described in the following algorithm.

Path-Based Incremental Target Level Algorithm

Step 0 (*Initialization*) Select x_0, $\delta_0 > 0$, and $b > 0$. Set $\sigma_0 = 0$, $f_{-1}^{\rm rec} = \infty$. Set $k = 0$, $l = 0$, and $k(l) = 0$ [$k(l)$ will denote the iteration number when the l-th update of $f_k^{\rm lev}$ occurs].

Step 1 (*Function evaluation*) Compute $f(x_k)$. If $f(x_k) < f_{k-1}^{\rm rec}$, then set $f_k^{\rm rec} = f(x_k)$. Otherwise set $f_k^{\rm rec} = f_{k-1}^{\rm rec}$ [so that $f_k^{\rm rec}$ keeps the record of the smallest value attained by the iterates that are generated so far, i.e., $f_k^{\rm rec} = \min_{0 \leq j \leq k} f(x_j)$].

Step 2 (*Sufficient descent*) If $f(x_k) \leq f_{k(l)}^{\rm rec} - \frac{\delta_l}{2}$, then set $k(l+1) = k$, $\sigma_k = 0$, $\delta_{l+1} = \delta_l$, increase l by 1, and go to Step 4.

Step 3 (*Oscillation detection*) If $\sigma_k > b$, then set $k(l+1) = k$, $\sigma_k = 0$, $\delta_{l+1} = \frac{\delta_l}{2}$, and increase l by 1.

Step 4 (*Iterate update*) Set $f_k^{\rm lev} = f_{k(l)}^{\rm rec} - \delta_l$. Select $\gamma_k \in [\underline{\gamma}, \overline{\gamma}]$ and compute x_{k+1} via Eqs. (2.3)–(2.5) with the stepsize (2.21).

Step 5 (*Path length update*) Set $\sigma_{k+1} = \sigma_k + \alpha_k mC$, increase k by 1, and go to Step 1.

The algorithm uses the same target level $f_k^{\rm lev} = f_{k(l)}^{\rm rec} - \delta_l$ for $k = k(l), k(l)+1, \ldots, k(l+1)-1$. The target level is updated only if sufficient descent or oscillation is detected (Step 2 or Step 3, respectively). It can be shown that the value σ_k is an upper bound on the length of the path traveled by iterates $x_{k(l)}, \ldots, x_k$ for $k < k(l+1)$. Whenever σ_k exceeds the prescribed upper bound b on the path length, the parameter δ_l is decreased, which (possibly) increases the target level $f_k^{\rm lev}$.

We will show that $\inf_{k\geq 0} f(x_k) = f^*$ even if f^* is not finite. First, we give a preliminary result showing that the target values $f_k^{\rm lev}$ are updated infinitely often (i.e., $l \to \infty$), and that $\inf_{k\geq 0} f(x_k) = -\infty$ when δ_l is bounded away from zero.

Lemma 2.4: Let Assumption 2.1 hold. Then, for the path-based incremental target level algorithm, we have $l \to \infty$, and either $\inf_{k\geq 0} f(x_k) = -\infty$ or $\lim_{l\to\infty} \delta_l = 0$.

Proof: Assume that l takes only a finite number of values, say $l = 0, 1, \ldots, \bar{l}$. In this case, we have $\sigma_k + \alpha_k C = \sigma_{k+1} \leq B$ for all $k \geq k(\bar{l})$, so that $\lim_{k\to\infty} \alpha_k = 0$. But this is impossible, since for all $k \geq k(\bar{l})$, we have

$$\alpha_k = \gamma_k \frac{f(x_k) - f_k^{\rm lev}}{m^2 C^2} \geq \underline{\gamma}\ \frac{\delta_{\bar{l}}}{m^2 C^2} > 0.$$

Hence, $l \to \infty$.

Let $\delta = \lim_{l\to\infty} \delta_l$. If $\delta > 0$, then from Steps 2 and 3 it follows that for all l large enough, we have $\delta_l = \delta$ and

$$f_{k(l+1)}^{\rm rec} - f_{k(l)}^{\rm rec} \leq -\frac{\delta}{2},$$

implying that $\inf_{k\geq 0} f(x_k) = -\infty$. **Q.E.D.**

We next give a convergence result for the path-based algorithm. In the special case of the ordinary subgradient method, this result coincides with that of Goffin and Kiwiel [GoK99].

Proposition 2.12: Let Assumption 2.1 hold. Then, for the path-based incremental target level algorithm, we have

$$\inf_{k\geq 0} f(x_k) = f^*.$$

Proof: If $\lim_{l\to\infty} \delta_l > 0$, then by Lemma 2.4, we have $\inf_{k\geq 0} f(x_k) = -\infty$ and we are done, so assume that $\lim_{l\to\infty} \delta_l = 0$. Let Λ be given by

$$\Lambda = \left\{ l \;\middle|\; \delta_l = \frac{\delta_{l-1}}{2},\ l \geq 1 \right\}.$$

Then, from Steps 3 and 5, we obtain

$$\sigma_k = \sigma_{k-1} + \alpha_{k-1} mC = \sum_{j=k(l)}^{k-1} \alpha_j mC,$$

so that $k(l+1) = k$ and $l+1 \in \Lambda$ whenever $\sum_{j=k(l)}^{k-1} \alpha_j mC > b$ at Step 3. Hence,

$$\sum_{j=k(l-1)}^{k(l)-1} \alpha_j > \frac{b}{mC}, \qquad \forall\ l \in \Lambda,$$

and since the cardinality of Λ is infinite, we have

$$\sum_{j=0}^{\infty} \alpha_j \geq \sum_{l \in \Lambda} \sum_{j=k(l-1)}^{k(l)-1} \alpha_j > \sum_{l \in \Lambda} \frac{b}{mC} = \infty. \tag{2.26}$$

Assume, to obtain a contradiction, that $\inf_{k \geq 0} f(x_k) > f^*$, so that for some $\hat{y} \in X$ and $\epsilon > 0$, we have

$$\inf_{k \geq 0} f(x_k) - \epsilon \geq f(\hat{y}). \tag{2.27}$$

Since $\delta_l \to 0$, there is a large enough $\hat{l}$ such that $\delta_l \leq \epsilon$ for all $l \geq \hat{l}$, implying that

$$f_k^{\rm lev} = f_{k(l)}^{\rm rec} - \delta_l \geq \inf_{k \geq 0} f(x_k) - \epsilon \geq f(\hat{y}), \qquad \forall\ k \geq k(\hat{l}).$$

Using this relation, Lemma 2.1 with $y = \hat{y}$, and the definition of α_k, we obtain

$$\begin{aligned}
\|x_{k+1} - \hat{y}\|^2 &\leq \|x_k - \hat{y}\|^2 - 2\alpha_k \big(f(x_k) - f(\hat{y})\big) + \alpha_k^2 m^2 C^2 \\
&\leq \|x_k - \hat{y}\|^2 - 2\alpha_k \big(f(x_k) - f_k^{\rm lev}\big) + \alpha_k^2 m^2 C^2 \\
&= \|x_k - \hat{y}\|^2 - \gamma_k(2 - \gamma_k) \frac{\big(f(x_k) - f_k^{\rm lev}\big)^2}{m^2 C^2} \\
&\leq \|x_k - \hat{y}\|^2 - \underline{\gamma}(2 - \overline{\gamma}) \frac{\big(f(x_k) - f_k^{\rm lev}\big)^2}{m^2 C^2}, \qquad \forall\ k \geq k(\hat{l}).
\end{aligned}$$

By adding these inequalities over $k \geq k(\hat{l})$, we see that

$$\frac{\underline{\gamma}(2 - \overline{\gamma})}{m^2 C^2} \sum_{k=k(\hat{l})}^{\infty} \big(f(x_k) - f_k^{\rm lev}\big)^2 \leq \|x_{k(\hat{l})} - \hat{y}\|^2,$$

yielding $\sum_{k=k(\hat{l})}^{\infty} \alpha_k^2 < \infty$ [see the definition of α_k in Eq. (2.21)], and consequently $\alpha_k \to 0$. This and the relation $\sum_{k=0}^{\infty} \alpha_k = \infty$ [cf. Eq. (2.26)] imply by Prop. 2.4 that

$$\liminf_{k \to \infty} f(x_k) = f^*,$$

contradicting Eq. (2.27). **Q.E.D.**

Let us note that there is no need to keep the path bound b fixed. Instead, as the method progresses, we can decrease b (possibly at Step 3) in such a way that $\sum_{l \in \Lambda} b_l = \infty$ [cf. Eq. (2.26)], which ensures that the convergence result of Prop. 2.12 is preserved.

In an attempt to improve the efficiency of the path-based incremental target level algorithm, one may introduce parameters $\beta, \tau \in (0,1)$ and $\rho \geq 1$ (whose values will be fixed at Step 0), and modify Steps 2 and 3 as follows:

Step 2′ If $f(x_k) \leq f^{\rm rec}_{k(l)} - \tau\delta_l$, then set $k(l+1) = k$, $\sigma_k = 0$, $\delta_{l+1} = \rho\delta_l$, increase l by 1, and go to Step 4.

Step 3′ If $\sigma_k > b$, then set $k(l+1) = k$, $\sigma_k = 0$, $\delta_{l+1} = \beta\delta_l$, and increase l by 1.

It can be seen that the result of Prop. 2.12 still holds for this modified algorithm. If we choose $\rho > 1$ at Step 3′, then in the proofs of Lemma 2.4 and Prop. 2.12 we have to replace $\lim_{l\to\infty}\delta_l$ with $\limsup_{l\to\infty}\delta_l$.

We next give a convergence rate estimate that applies to the stepsize rule (2.21)–(2.22) with any of the two adjustment procedures previously discussed.

Proposition 2.13: Let Assumption 2.1 hold, and assume that the optimal solution set X^* is nonempty. Let $\{x_k\}$ be the sequence generated by the incremental subgradient method with the dynamic stepsize (2.21)–(2.22). Then, for the largest positive integer K such that

$$\sum_{k=0}^{K-1}\delta_k^2 \leq \frac{m^2C^2}{\underline{\gamma}(2-\overline{\gamma})}\big(dist(x_0,X^*)\big)^2,$$

we have

$$\min_{0\leq k\leq K} f(x_k) \leq f^* + \max_{0\leq j\leq K}\delta_j.$$

Proof: Assume, to arrive at a contradiction, that

$$f(x_k) > f^* + \max_{0\leq j\leq K}\delta_j, \qquad \forall\ k = 0,1,\ldots,K,$$

which implies that

$$f_k^{\rm lev} = \min_{0\leq j\leq k} f(x_j) - \delta_k > f^* + \max_{0\leq j\leq K}\delta_j - \delta_k \geq f^*, \qquad \forall\ k = 0,1,\ldots,K. \tag{2.28}$$

From Lemma 2.2 and the definition of the stepsize, we obtain for all k,

$$\big(dist(x_{k+1},X^*)\big)^2 \leq \big(dist(x_k,X^*)\big)^2 - 2\gamma_k\frac{f(x_k)-f_k^{\rm lev}}{m^2C^2}\big(f(x_k)-f^*\big) + \gamma_k^2\frac{\big(f(x_k)-f_k^{\rm lev}\big)^2}{m^2C^2}.$$

By using $f(x_k) - f^* \geq f(x_k) - f_k^{\rm lev}$ [cf. Eq. (2.28)] and $f(x_k) - f_k^{\rm lev} \geq \delta_k$ for all k, in the preceding inequality, we see that

$$\big(dist(x_{k+1},X^*)\big)^2 \leq \big(dist(x_k,X^*)\big)^2 - \gamma_k(2-\gamma_k)\frac{\delta_k^2}{m^2C^2}, \qquad \forall\ k = 0,1,\ldots,K.$$

Summing these inequalities over $k = 0,1,\ldots,K$ and using $\gamma_k \in [\underline{\gamma},\overline{\gamma}]$ yields

$$\big(dist(x_{K+1},X^*)\big)^2 \leq \big(dist(x_0,X^*)\big)^2 - \underline{\gamma}(2-\overline{\gamma})\sum_{k=0}^{K}\frac{\delta_k^2}{m^2C^2},$$

implying that

$$\sum_{k=0}^{K} \delta_k^2 \leq \frac{m^2 C^2}{\underline{\gamma}(2-\overline{\gamma})} \big(dist(x_0, X^*)\big)^2,$$

thus contradicting the definition of K. **Q.E.D.**

The estimate of Prop. 2.13 is similar to the estimates for the ordinary subgradient method that are obtained by Kiwiel [Kiw96a], and Kulikov and Fazylov [KuF90]. For the adjustment procedure (2.23) with $\rho = 1$ and for the path-based adjustment procedure, the estimate of Prop. 2.13 holds with δ_0 in place of $\max_{0\leq j\leq K} \delta_j$.

2.6.1 Remarks

It can be verified that all the results of this section are valid for the incremental method that does not use projections within the cycles, but rather employs projections at the end of cycles:

$$\psi_{i,k} = \psi_{i-1,k} - \alpha_k g_{i,k}, \qquad g_{i,k} \in \partial f_i(\psi_{i-1,k}), \quad i = 1, \ldots, m,$$

where $\psi_{0,k} = x_k$ and the iterate x_{k+1} is given by

$$x_{k+1} = \mathcal{P}_X[\psi_{m,k}].$$

This method and its modifications are proposed and analyzed by Solodov and Zavriev [SoZ98], for the case of a compact set X and a diminishing stepsize rule.

The preceding convergence and convergence rate results hold assuming any order for processing the component functions f_i within a cycle, for as long as each component f_i is processed exactly once in every cycle. In particular, at the beginning of each cycle, we could reorder the components f_i by either shifting or reshuffling and then proceed with the calculations until the end of the cycle.

The convergence rate estimates of this section emphasize the role of the stepsize choice in the performance of incremental subgradient methods. These estimates do not capture the effects of processing order on the methods' performance, which can be significant, as we will see in the next section.

2.7. EFFECTS OF FIXED CYCLIC PROCESSING ORDER

Here, we will demonstrate by some examples the effects of a cyclic processing order on the performance of the incremental subgradient method. For this, we assume that the component functions $f_1, \ldots, f_m$ are processed in the same order within each cycle. We also assume that the optimal solution set X^* is nonempty, and we consider the method with a constant stepsize α, in which case the method exhibits an oscillatory behavior. To get some insights about this phenomenon, we will consider some simple examples where the oscillatory behavior is

displayed by the presence of limit cycles. We will introduce the size of a limit cycle, which in a sense measures the oscillations, and we will show that the size of oscillations depends on the processing order. As we will see, in the worst case (i.e., for the most unfavorable processing order), the size of oscillations can be proportional to $m\alpha$.

We define the size of a limit cycle to be the maximal distance from the cycle points to the optimal solution set X^*. In particular, if the points $\overline{\psi}_0, \overline{\psi}_1, \ldots, \overline{\psi}_{m-1}, \overline{\psi}_m$ (with $\overline{\psi}_0 = \overline{\psi}_m$) comprise the limit cycle, then the size of this cycle is

$$\max_{1 \leq i \leq m} dist\big(\overline{\psi}_i, X^*\big).$$

In the following example, we compute the size of limit cycles corresponding to the worst and best cyclic processing order, for the case where f is a nondifferentiable function.

Example 2.2: (Convex Nondifferentiable Function)

Assume that x is a scalar, and that the problem has the form

$$\text{minimize} \quad f(x) = \sum_{i=1}^{p} |x+1| + \sum_{i=1}^{p} |x-1| + \sum_{i=1}^{2rp} |x|$$

$$\text{subject to} \quad x \in \Re.$$

Here, the cost function f consists of p copies of the functions $|x-1|$ and $|x+1|$, and $2rp$ copies of the function $|x|$, where p and r are positive integers. We focus primarily on the case where p is large in order to exhibit most prominently the effects of processing order. For simplicity, we assume that x is unconstrained; a similar example can be constructed when there are constraints. The minimum of f is attained at $x^* = 0$, which is the unique optimal solution.

We consider the incremental subgradient method with a constant stepsize α. For x_k outside the interval $[-1, 1]$, which contains the minima of the component functions f_i, the subgradient of each component f_i is

$$g_{i,k} = \begin{cases} -1 & \text{if } x_k < -1, \\ 1 & \text{if } x_k > 1. \end{cases}$$

In this case, each of the steps

$$\psi_{i,k} = \psi_{i-1,k} - \alpha g_{i,k}, \qquad i = 1, \ldots, m$$

[cf. Eq. (2.4) with $\alpha_k = \alpha$] makes progress towards the minimum $x^* = 0$, and in fact we have

$$|\psi_{i,k}| = |\psi_{i-1,k}| - \alpha.$$

However, once the method enters the interval $[-1, 1]$, it exhibits oscillatory behavior, i.e., the iterates are trapped in a limit cycle.

For the method starting at the initial point $x_0 = p\alpha$ with $p\alpha \leq 1$, we determine the limit cycles and compute their sizes for the worst and the best processing orders. To keep the analysis simple, we assume that the subgradient $g = 0$ is used for component functions $|x|$ at $x = 0$.

Let the processing order be as follows: rp functions of the form $|x|$ followed by p functions of the form $|x+1|$, then rp functions of the form $|x|$ followed by p functions of the form $|x-$

$1|$. For this processing order, it can be seen that in each cycle, the first $p + rp$ iterates are $p\alpha, \ldots, \alpha, 0, -\alpha, \ldots, -p\alpha$, and the subsequent $p+rp$ iterates are $-p\alpha, \ldots, -\alpha, 0, \alpha, \ldots, p\alpha$. Hence, the size of the limit cycle is $p\alpha$, which is proportional to $m\alpha$ [here $m = 2(1+r)p$] and corresponds to the worst processing order.

Let now the processing order be as follows: a function of the form $|x+1|$ followed by a function of the form $|x-1|$ until all such components are processed, and then $2rp$ functions of the form $|x|$. This is the best processing order, because after the first cycle, we have $x_1 = 0 = x^*$, and each of the subsequent cycles generates the iterates $0, -\alpha, 0, \ldots, -\alpha, 0, 0, \ldots, 0$, which comprise the limit cycle. Thus, in this case, the size of the limit cycle is α, corresponding to the best processing order.

As seen in this example, the performance of the method is substantially affected by the *order* in which the functions f_i are processed within each cycle, and in the worst case, it can be proportional to $m\alpha$. Therefore, a stepsize converging to zero is needed to confine the limit cycle within the optimal solution set X^*, thus resulting in convergence of the method.

When f is a differentiable function, the incremental method with a constant stepsize exhibits a similar behavior, as seen in the following example.

Example 2.3: (Convex Differentiable Function)

Assume that x is a scalar, and that the problem has the form

$$\begin{aligned} &\text{minimize} \quad f(x) = \frac{1}{2}\left(\sum_{i=1}^{p}(x-1)^2 + \sum_{i=1}^{p}(x+1)^2\right) \\ &\text{subject to} \quad x \in \Re. \end{aligned}$$

Thus, the cost function consists of p copies of just two functions, $(x-1)^2$ and $(x+1)^2$, where p is a large positive integer. The minimum value of f is attained at $x^* = 0$, which is the unique optimal solution.

We next determine limit cycles corresponding to the worst and best processing orders for the method starting at $x_0 = \alpha/(2-\alpha)$, with the stepsize $\alpha < 2$. Consider first the case where, in each cycle, the p terms $(x-1)^2$ are processed first and the p terms $(x+1)^2$ are processed next. Let ψ_0 be the iterate at the beginning. Then, the first p iterates within the cycle are given by

$$\psi_i = \psi_{i-1} - \alpha(\psi_{i-1} - 1) = (1-\alpha)\psi_{i-1} + \alpha, \quad i = 1, \ldots, p,$$

leading to the mid-cycle iterate

$$\begin{aligned} \psi_p &= (1-\alpha)^p\psi_0 + \alpha\big(1 + (1-\alpha) + \cdots + (1-\alpha)^{p-1}\big) \\ &= (1-\alpha)^p\psi_0 + \big(1 - (1-\alpha)^p\big). \end{aligned} \tag{2.29}$$

The subsequent p iterates within the cycle are given by

$$\psi_{p+i} = \psi_{p+i-1} - \alpha(\psi_{p+i-1} + 1) = (1-\alpha)\psi_{p+i-1} - \alpha, \quad i = 1, \ldots, p,$$

leading similarly to the final iterate of the cycle

$$\psi_{2p} = (1-\alpha)^p\psi_p - \big(1 - (1-\alpha)^p\big). \tag{2.30}$$

Thus, by combining Eqs. (2.29) and (2.30), we have

$$\psi_m = (1-\alpha)^{2p}\psi_0 - \Big(1-(1-\alpha)^p\Big)^2.$$

Since $m = 2p$ and since we must have $\psi_m = \psi_0$ in the limit, we obtain that the limit cycle starts and ends at the point

$$\overline{\psi}_0 = -\frac{1-(1-\alpha)^{m/2}}{1+(1-\alpha)^{m/2}}.$$

It can be seen that the size of the limit cycle is

$$\frac{1-(1-\alpha)^{m/2}}{1+(1-\alpha)^{m/2}}. \tag{2.31}$$

Consider now the case where, in each cycle, we use the following processing order: a function of the form $(x+1)^2$ followed by a function of the form $(x-1)^2$. Since $x_0 = \alpha/(2-\alpha)$, in the first cycle, we have

$$\psi_1 = x_0 - \alpha(\psi_0 + 1) = (1-\alpha)\frac{\alpha}{2-\alpha} - \alpha = -\frac{\alpha}{2-\alpha},$$

$$\psi_2 = \psi_1 - \alpha(\psi_1 - 1) = -(1+\alpha)\frac{\alpha}{2-\alpha} + \alpha = \frac{\alpha}{2-\alpha},$$

and therefore, because $m = 2p$, the first cycle ends at the point $\psi_m = x_0$ and all subsequent cycles are the same as the first one. Hence, the size of the limit cycle is

$$\frac{\alpha}{2-\alpha}. \tag{2.32}$$

When α is small, this processing order is worse than the preceding one, as seen from Eqs. (2.31) and (2.32). However, when α is moderate and m is large enough, then this processing order is better than the preceding one.

The preceding examples illustrate several common characteristics of the incremental subgradient and incremental gradient methods, which tend to manifest themselves in some generality:

(a) When far from the optimal solution set, the incremental method makes progress comparable to that of the nonincremental method.

(b) When close to the optimal solution set X^*, the method can be trapped in an oscillatory region whose size depends on the stepsize and on the processing order for the component functions.

(c) The precise effect of the processing order is not fully understood at present, but it is interesting and substantial.

3

An Incremental Subgradient Method with Randomization

In this chapter, we consider a version of the incremental subgradient method that uses randomization. In particular, at each iteration we select randomly a component function whose subgradient is used in the iteration update, where each component function f_i is selected with the same probability $1/m$. We will here analyze the method for various stepsize choices, and we will see that this randomized method can have a better convergence rate than the nonrandomized incremental method (2.3)–(2.5) discussed in Chapter 2.

3.1. THE METHOD

As discussed in the preceding chapter, the processing order for the components f_i has a significant affect on the convergence rate of the method. Unfortunately, determining the most favorable order may be very difficult in practice. A popular technique for incremental gradient methods (differentiable components f_i) is to reshuffle randomly the order of the functions f_i at the beginning of each cycle. A variation of this method is to pick randomly a function f_i at each iteration rather than to pick each f_i exactly once in every cycle according to a randomized order. This variation can be viewed as a gradient method with random errors, as shown in Bertsekas and Tsitsiklis [BeT96] (see also Bertsekas and Tsitsiklis [BeT00]). Similarly, the corresponding incremental subgradient method at each step picks randomly a function f_i to be processed next. In this section, we will analyze the method for the constant, diminish-

ing, and dynamic stepsize rules. For the case of a diminishing stepsize, the convergence of the method follows from known stochastic subgradient convergence results (e.g., Ermoliev [Erm69], [Erm76], [Erm83], and [Erm88], and Polyak [Pol87]). The analysis for the constant and dynamic stepsize rules is new and has no counterpart in the available stochastic subgradient literature.

The formal description of the randomized order method is as follows:

$$x_{k+1} = \mathcal{P}_X\big[x_k - \alpha_k g(\omega_k, x_k)\big], \qquad k = 0, 1, \ldots, \tag{3.1}$$

where x_0 is an initial random vector, ω_k is a random variable taking equiprobable values from the set $\{1, \ldots, m\}$, and $g(\omega_k, x_k)$ is a subgradient of the component f_{ω_k} at x_k [i.e., if the random variable ω_k takes a value j, then the vector $g(\omega_k, x_k)$ is a subgradient of f_j at x_k].

3.2. ASSUMPTIONS AND SOME BASIC RELATIONS

In our analysis, we use the following assumption regarding the randomized method (3.1).

Assumption 3.1: Assume the following:

(a) The sequence $\{\omega_k\}$ is a sequence of independent random variables, each uniformly distributed over the set $\{1, \ldots, m\}$. Furthermore, the sequence $\{\omega_k\}$ is independent of the sequence $\{x_k\}$.

(b) The set of subgradients $\big\{g(\omega_k, x_k) \mid k = 0, 1, \ldots\big\}$ is bounded, i.e., there exists a positive constant C such that with probability 1

$$\|g(\omega_k, x_k)\| \leq C, \qquad \forall\ k.$$

Note that if the constraint set X is compact or the components f_i are polyhedral, then Assumption 3.1(b) is satisfied. In our analysis, along with the preceding assumption, we often use the assumption that the optimal solution set X^* is nonempty. Furthermore, the proofs of several propositions in this section rely on the Supermartingale Convergence Theorem as stated, for example, in Bertsekas and Tsitsiklis [BeT96], p. 148.

Theorem 3.1: (Supermartingale Convergence Theorem) Let $\{Y_k\}$, $\{Z_k\}$, and $\{W_k\}$ be sequences of random variables, and let $\mathcal{F}_k$, $k = 0, 1, 2, \ldots$, be sets of random variables such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k. Suppose that:

(a) The random variables Y_k, Z_k, and W_k are nonnegative, and are functions of the random variables in $\mathcal{F}_k$.

(b) For each k, we have $E\big\{Y_{k+1} \mid \mathcal{F}_k\big\} \leq Y_k - Z_k + W_k$.

(c) There holds $\sum_{k=0}^{\infty} W_k < \infty$.

Then, with probability 1, the sequence $\{Y_k\}$ converges to a nonnegative random variable and $\sum_{k=0}^{\infty} Z_k < \infty$.

3.3. CONSTANT STEPSIZE RULE

Here, we establish convergence properties and convergence rate estimates for the method using a constant stepsize α. We start with a basic result that we use in the analysis of this section and the next one, where the diminishing stepsize rule is considered.

Lemma 3.1: Let Assumption 3.1 hold. Then, for the sequence $\{x_k\}$ generated by the randomized method and a deterministic stepsize α_k, we have

$$E\big\{\|x_{k+1}-y\|^2 \mid \mathcal{F}_k\big\} \leq \|x_k-y\|^2 - \frac{2\alpha_k}{m}\big(f(x_k)-f^*\big) + \alpha_k^2 C^2, \qquad \forall\ y\in X, \quad \forall\ k,$$

where $\mathcal{F}_k = \{x_0, x_1, \ldots, x_k\}$.

Proof: By using Eq. (3.1), the nonexpansion property of the projection, and the boundedness of the subgradients $g(\omega_k, x_k)$, we have for any $y \in X$ and all k,

$$\begin{aligned}\|x_{k+1}-y\|^2 &= \big\|\mathcal{P}_X[x_k-\alpha_k g(\omega_k,x_k)]-y\big\|^2\\ &\leq \|x_k-\alpha_k g(\omega_k,x_k)-y\|^2\\ &\leq \|x_k-y\|^2 - 2\alpha_k g(\omega_k,x_k)'(x_k-y)+\alpha_k^2C^2.\end{aligned}$$

Since $g(\omega_k, x_k)$ is a subgradient of f_{ω_k} at x_k, it follows that

$$g(\omega_k,x_k)'(x_k-y) \geq f_{\omega_k}(x_k)-f_{\omega_k}(y),$$

so that

$$\|x_{k+1}-y\|^2 \leq \|x_k-y\|^2 - 2\alpha_k\big(f_{\omega_k}(x_k)-f_{\omega_k}(y)\big)+\alpha_k^2C^2, \qquad \forall\ y\in X, \quad \forall\ k.$$

By taking the expectation conditioned on $\mathcal{F}_k = \{x_0, x_1, \ldots, x_k\}$ in the preceding inequality and by using

$$E\big\{f_{\omega_k}(x_k)-f_{\omega_k}(y)\mid \mathcal{F}_k\big\} = \frac{1}{m}\sum_{i=1}^m\big(f_i(x_k)-f_i(y)\big) = \frac{1}{m}\big(f(x_k)-f(y)\big),$$

we obtain

$$E\big\{\|x_{k+1}-y\|^2 \mid \mathcal{F}_k\big\} \leq \|x_k-y\|^2 - \frac{2\alpha_k}{m}\big(f(x_k)-f(y)\big)+\alpha_k^2C^2, \qquad \forall\ y\in X, \quad \forall\ k.$$

Q.E.D.

For the method (3.1) using the constant stepsize rule, we have the following convergence result.

Proposition 3.1: Let Assumption 3.1 hold. Then, for the sequence $\{x_k\}$ generated by the randomized method and the stepsize α_k fixed to some positive constant α, with probability 1, we have:

(a) If $f^* = -\infty$, then

$$\inf_{k\geq 0} f(x_k) = f^*.$$

(b) If $f^* > -\infty$, then

$$\inf_{k\geq 0} f(x_k) \leq f^* + \frac{\alpha m C^2}{2}.$$

Proof: We prove (a) and (b) simultaneously. Let N be an arbitrary positive integer, and let $y_N \in X$ be such that

$$f(y_N) = \begin{cases} -N & \text{if } f^* = -\infty, \\ f^* + \frac{1}{N} & \text{if } f^* > -\infty. \end{cases}$$

Consider the level set L_N defined by

$$L_N = \left\{ x \in X \mid f(x) \leq f(y_N) + \frac{1}{N} + \frac{\alpha m C^2}{2} \right\},$$

and note that $y_N \in L_N$. Define a new process $\{\hat{x}_k\}$ as follows

$$\hat{x}_k = \begin{cases} x_k & \text{if } x_k \notin L_N, \\ y_N & \text{otherwise.} \end{cases}$$

Thus, the process $\{\hat{x}_k\}$ is identical to $\{x_k\}$, except that once x_k enters the level set L_N, the process $\{\hat{x}_k\}$ terminates with $\hat{x}_k = y_N$ (since $y_N \in L_N$). Using Lemma 3.1 with $y = y_N$, we have

$$E\big\{\|\hat{x}_{k+1} - y_N\|^2 \mid \mathcal{F}_k\big\} \leq \|\hat{x}_k - y_N\|^2 - \frac{2\alpha}{m}\big(f(\hat{x}_k) - f(y_N)\big) + \alpha^2 C^2, \qquad \forall\ k,$$

or equivalently,

$$E\big\{\|\hat{x}_{k+1} - y_N\|^2 \mid \mathcal{F}_k\big\} \leq \|\hat{x}_k - y_N\|^2 - z_k, \qquad \forall\ k, \tag{3.2}$$

where

$$z_k = \begin{cases} \frac{2\alpha}{m}\big(f(\hat{x}_k) - f(y_N)\big) - \alpha^2 C^2 & \text{if } \hat{x}_k \notin L_N, \\ 0 & \text{if } \hat{x}_k \in L_N. \end{cases}$$

For $\hat{x}_k \notin L_N$, we have

$$f(\hat{x}_k) - f(y_N) \geq \frac{1}{N} + \frac{\alpha m C^2}{2},$$

implying that

$$\frac{2\alpha}{m}\big(f(\hat{x}_k) - f(y_N)\big) - \alpha^2 C^2 \geq \frac{2\alpha}{mN}.$$

Hence, $z_k \geq 0$ if $\hat{x}_k \notin L_N$, and since otherwise $z_k = 0$, we see that $z_k \geq 0$ for all k. Therefore, by the Supermartingale Convergence Theorem, we obtain $\sum_{k=0}^{\infty} z_k < \infty$ with probability 1,

implying that $\hat{x}_k \in L_N$ for some k, with probability 1. Thus, in the original process, by the definitions of y_N and L_N, we have

$$\inf_{k\geq 0} f(x_k) \leq \begin{cases} -N + \frac{1}{N} + \frac{\alpha m C^2}{2} & \text{if } f^* = -\infty, \\ f^* + \frac{2}{N} + \frac{\alpha m C^2}{2} & \text{if } f^* > -\infty, \end{cases}$$

with probability 1, and by letting $N \to \infty$, we obtain the desired relations. **Q.E.D.**

The estimate in part (b) of the preceding proposition is sharp. For example, let $f_i(x) = C|x|$ for all $x \in \Re$ and $i = 1, \ldots, m$. For any α, choose the initial point $x_0 = \alpha C/2$. In this case, it can be seen that the iterates x_k generated by the method (3.1) take the values $\frac{\alpha C}{2}$ or $-\frac{\alpha C}{2}$, so that

$$f(x_k) = \frac{\alpha m C^2}{2}, \qquad \forall\ k.$$

To compare fairly the error bounds of Props. 2.1 and 3.1, we will assume that, in the randomized method, the function f is evaluated every m iterations. Thus, by Prop. 3.1, for the randomized method, we have

$$\inf_{k\geq 0} f(x_{mk}) \leq f^* + \frac{\alpha m C^2}{2}.$$

At the same time, by Prop. 2.1, for the nonrandomized incremental method, we have

$$\liminf_{k\to\infty} f(x_k) \leq f^* + \frac{\alpha m^2 C^2}{2}.$$

Thus, for the same value of the stepsize α, the error bound for the randomized method is smaller by a factor of m than that for the nonrandomized method (2.3)–(2.5). This indicates that when randomization is used, the stepsize α_k could generally be chosen larger than in the nonrandomized methods. Being able to use a larger stepsize suggests a potential rate of convergence advantage in favor of the randomized methods, which is consistent with our experimental results.

We next estimate the expected number of iterations needed to guarantee that, with probability 1, a solution is obtained with the approximation error ϵ. Such an estimate requires that the optimal solution set is nonempty.

Proposition 3.2: Let Assumption 3.1 hold, and assume that the optimal solution set X^* is nonempty. Let the sequence $\{x_k\}$ be generated by the randomized method with the stepsize α_k fixed to some positive constant α. Then, for any positive scalar ϵ, there exists a random nonnegative integer N such that

$$\min_{0\leq k\leq N} f(x_k) \leq f^* + \frac{\alpha m C^2 + \epsilon}{2},$$

with probability 1, and

$$E\{N\} \leq \frac{m}{\alpha\epsilon} E\Big\{ \big(dist(x_0, X^*)\big)^2 \Big\}.$$

Proof: From Lemma 3.1 with $y = x^*$ and $\alpha_k = \alpha$, we have

$$E\big\{\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\big\} \leq \|x_k - x^*\|^2 - \frac{2\alpha}{m}\big(f(x_k) - f^*\big) + \alpha^2 C^2, \qquad \forall\, x^* \in X^*, \quad \forall\, k.$$

By taking the minimum over $x^* \in X^*$ of both sides in this relation and by using the inequality

$$E\Big\{\big(dist(x_{k+1}, X^*)\big)^2 \;\Big|\; \mathcal{F}_k\Big\} \leq \min_{x^* \in X^*} E\big\{\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\big\},$$

we obtain

$$E\Big\{\big(dist(x_{k+1}, X^*)\big)^2 \;\Big|\; \mathcal{F}_k\Big\} \leq \big(dist(x_k, X^*)\big)^2 - \frac{2\alpha}{m}\big(f(x_k) - f^*\big) + \alpha^2 C^2, \qquad \forall\, k. \tag{3.3}$$

Let the level set L be given by

$$L = \left\{x \in X \;\middle|\; f(x) < f^* + \frac{\alpha m C^2 + \epsilon}{2}\right\},$$

and consider a new process $\{\hat{x}_k\}$ defined by

$$\hat{x}_k = \begin{cases} x_k & \text{if } x_k \notin L, \\ x^* & \text{otherwise,} \end{cases}$$

where $x^* \in X^*$ is some fixed vector. The process $\{\hat{x}_k\}$ is identical to $\{x_k\}$, except that once x_k enters the level set L the process $\{\hat{x}_k\}$ terminates at x^*. Thus, by Eq. (3.3), it follows that

$$\begin{aligned} E\Big\{\big(dist(\hat{x}_{k+1}, X^*)\big)^2 \;\Big|\; \mathcal{F}_k\Big\} &\leq \big(dist(\hat{x}_k, X^*)\big)^2 - \frac{2\alpha}{m}\big(f(\tilde{x}_k) - f^*\big) + \alpha^2 C^2 \\ &= \big(dist(\hat{x}_k, X^*)\big)^2 - z_k, \qquad \forall\, k, \end{aligned} \tag{3.4}$$

where

$$z_k = \begin{cases} \frac{2\alpha}{m}\big(f(\hat{x}_k) - f^*\big) - \alpha^2 C^2 & \text{if } \hat{x}_k \notin L, \\ 0 & \text{otherwise.} \end{cases}$$

We have

$$z_k \geq \frac{2\alpha}{m}\left(f^* + \frac{\alpha m C^2 + \epsilon}{2} - f^*\right) - \alpha^2 C^2 = \frac{\alpha\epsilon}{m}, \qquad \text{if} \quad \hat{x}_k \notin L, \tag{3.5}$$

and since otherwise $z_k = 0$, by the Supermartingale Convergence Theorem, it follows that $\sum_{k=0}^{\infty} z_k < \infty$ with probability 1. Hence, there exists a random nonnegative integer N such that $z_k = 0$ for all $k \geq N$, implying that $\hat{x}_N \in L$ with probability 1. Therefore, in the original process, with probability 1, we have

$$\min_{0 \leq k \leq N} f(x_k) \leq f^* + \frac{\alpha m C^2 + \epsilon}{2}.$$

Furthermore, by taking the total expectation in Eq. (3.4), we obtain

$$E\Big\{\big(dist(\tilde{x}_{k+1},X^*)\big)^2\Big\} \le E\Big\{\big(dist(\tilde{x}_k,X^*)\big)^2\Big\} - E\{z_k\}$$
$$\le E\Big\{\big(dist(x_0,X^*)\big)^2\Big\} - E\left\{\sum_{j=0}^{k} z_j\right\}, \qquad \forall\ k.$$

Therefore,

$$E\Big\{\big(dist(x_0,X^*)\big)^2\Big\} \ge E\left\{\sum_{k=0}^{\infty} z_k\right\} = E\left\{\sum_{k=0}^{N-1} z_k\right\} \ge \frac{\alpha\epsilon}{m}E\{N\},$$

where the last inequality above follows from Eq. (3.5), and the facts $x_k \notin L$ for $k < N$ and $z_k = 0$ for all $k \ge N$. **Q.E.D.**

To compare the result of Prop. 3.2 with that of Prop. 2.2, as a measure of complexity, we will consider the number of function evaluations. For the nonrandomized incremental method implemented with a constant stepsize α, we showed that (cf. Prop. 2.2)

$$\min_{0\le k\le K} f(x_k) \le f^* + \frac{\alpha m^2 C^2 + \epsilon}{2}$$

holds after K iterations, where

$$K = \left\lfloor \frac{1}{\alpha\epsilon}\big(dist(x_0,X^*)\big)^2 \right\rfloor.$$

Recall that, in the nonrandomized incremental method, we evaluate the function f at each iteration. Thus, the number K represents the number of function evaluations needed to guarantee that the optimal function value is achieved with error $(\alpha m^2 C^2 + \epsilon)/2$.

For a fair comparison, we assume that, in the randomized method, we use the same initial point x_0, the same tolerance level ϵ and stepsize α, and that we evaluate the function f every m iterations. Then, from Prop. 3.2 it follows that with probability 1,

$$\min_{0\le k\le K} f(x_{mk}) \le f^* + \frac{\alpha m C^2 + \epsilon}{2},$$

where the expected number of function evaluations K is such that

$$E\{K\} \le \frac{1}{\alpha\epsilon}\big(dist(x_0,X^*)\big)^2.$$

Thus, the bound on the number of function evaluations is the same for both nonrandomized and randomized method. However, the error term $\alpha m^2 C^2$ in the nonrandomized method is m times larger than the corresponding error term in the randomized method.

We now give a different estimate of the convergence rate for the randomized method with the constant stepsize rule, assuming that f has sharp minima.

Proposition 3.3: Let Assumption 3.1 hold, and assume that the optimal solution set X^* is nonempty. Assume further that for some positive scalar μ, with probability 1, we have

$$f(x) - f^* \geq \mu \left(dist(x, X^*)\right)^2, \qquad \forall\, x \in X.$$

Then, for the sequence $\{x_k\}$ generated by the randomized order method with a stepsize α_k fixed to some positive scalar α, we have

$$E\Big\{\left(dist(x_{k+1}, X^*)\right)^2\Big\} \leq \left(1 - \frac{2\alpha\mu}{m}\right)^{k+1} E\Big\{\left(dist(x_0, X^*)\right)^2\Big\} + \frac{\alpha m C^2}{2\mu}, \qquad \forall\, k.$$

Proof: By using Lemma 3.1 with $y = x^*$ and $\alpha_k = \alpha$ for all k, we can see that [cf. Eq. (3.3)]

$$E\Big\{\left(dist(x_{k+1}, X^*)\right)^2 \mid \mathcal{F}_k\Big\} \leq \left(dist(x_k, X^*)\right)^2 - \frac{2\alpha}{m}\big(f(x_k) - f^*\big) + \alpha^2 C^2, \qquad \forall k.$$

Then, by taking the total expectation and by using the given property of f, we obtain

$$\begin{aligned} E\Big\{\left(dist(x_{k+1}, X^*)\right)^2\Big\} &\leq E\Big\{\left(dist(x_k, X^*)\right)^2\Big\} - \frac{2\alpha}{m} E\big\{f(x_k) - f^*\big\} + \alpha^2 C^2 \\ &\leq \left(1 - \frac{2\alpha\mu}{m}\right) E\Big\{\left(dist(x_k, X^*)\right)^2\Big\} + \alpha^2 C^2, \qquad \forall\, k. \end{aligned}$$

Thus, by induction, we see that for all k,

$$E\Big\{\left(dist(x_{k+1}, X^*)\right)^2\Big\} \leq \left(1 - \frac{2\alpha\mu}{m}\right)^{k+1} E\Big\{\left(dist(x_0, X^*)\right)^2\Big\} + C^2\alpha^2 \sum_{j=0}^{k} \left(1 - \frac{2\alpha\mu}{m}\right)^j,$$

and by using the relation

$$\sum_{j=0}^{k} \left(1 - \frac{2\alpha\mu}{m}\right)^j \leq \frac{m}{2\alpha\mu},$$

we obtain the desired estimate. **Q.E.D.**

Let us compare, for the same initial vector x_0 and stepsize α, the estimate of Prop. 3.3 with that of Prop. 2.3. For the nonrandomized method, we have shown that (cf. Prop. 2.3)

$$\left(dist(x_{k+1}, X^*)\right)^2 \leq (1 - 2\alpha\mu)^{k+1} \left(dist(x_0, X^*)\right)^2 + \frac{\alpha m^2 C^2}{2\mu}, \qquad \forall\, k.$$

Thus, in both estimates, the error bound consists of two terms: the exponentially decreasing term and the asymptotic term. For the same value of the stepsize α, the asymptotic term in

the error bound for the nonrandomized method is m times larger than the asymptotic term in the error bound for the randomized method. However, if in the randomized method the stepsize α is replaced by $m\alpha$, then the asymptotic terms and the exponentially decreasing terms in the error bounds for both methods are the same. The main difference is that in the nonrandomized method, k represents the number of cycles (with m iterations per cycle), while in the randomized method, k represents the number of iterations. Therefore, for the same error level, the nonrandomized method requires a number of iterations that is m times larger than that of the randomized method.

3.4. DIMINISHING STEPSIZE RULE

We here analyze convergence of the randomized method (3.1) using a diminishing stepsize. In this case, the method exhibits the convergence similar to that of the stochastic subgradient method with the same stepsize, as seen in the following proposition.

Proposition 3.4: Let Assumption 3.1 hold, and assume that the optimal solution set X^* is nonempty. Assume further that the stepsize α_k is such that

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

Then, the sequence $\{x_k\}$ generated by the randomized method converges to some optimal solution with probability 1.

Proof: Using Lemma 3.1 with $y = x^* \in X^*$, we obtain

$$E\big\{\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\big\} \leq \|x_k - x^*\|^2 - \frac{2\alpha_k}{m}\big(f(x_k) - f^*\big) + \alpha_k^2 C^2, \qquad \forall\, x^* \in X^*, \quad \forall\, k.$$

By the Supermartingale Convergence Theorem, for each $x^* \in X^*$, with probability 1, we have

$$\sum_{k=0}^{\infty} \alpha_k \big(f(x_k) - f^*\big) < \infty, \tag{3.6}$$

and the sequence $\big\{\|x_k - x^*\|\big\}$ converges.

For each $x^* \in X^*$, let Ω_{x^*} denote the set of all sample paths for which Eq. (3.6) holds and $\big\{\|x_k - x^*\|\big\}$ converges. By convexity of f, the set X^* is convex, so there exist vectors $v_0, v_1, \ldots, v_p \in X^*$ that span the smallest affine set containing X^*, and are such that $v_i - v_0$, $i = 1, \ldots, p$, are linearly independent. The intersection

$$\Omega = \cap_{i=1}^{p} \Omega_{v_i}$$

has probability 1, and for each sample path in Ω, the sequences $\{\|x_k - v_i\|\}$, $i = 0, \ldots, p$, converge. Thus, with probability 1, $\{x_k\}$ is bounded, and therefore it has limit points. Furthermore, for each sample path in Ω, by Eq. (3.6) and the relation $\sum_{k=0}^{\infty} \alpha_k = \infty$, it follows that

$$\liminf_{k \to \infty} f(x_k) = f^*,$$

implying that $\{x_k\}$ has at least one limit point that belongs to X^* by continuity of f. For any sample path in Ω, let $\overline{x}$ and $\hat{x}$ be two limit points of $\{x_k\}$ such that $\overline{x} \in X^*$. Because $\{\|x_k - v_i\|\}$ converges for all $i = 0, \ldots, p$, we must have

$$\|\overline{x} - v_i\| = \|\hat{x} - v_i\|, \qquad \forall\ i = 0, 1, \ldots, p.$$

Moreover, since $\overline{x} \in X^*$, the preceding relation can hold only for $\overline{x} = \hat{x}$ by convexity of X^* and the choice of vectors v_i. Hence, for each sample path in Ω, the sequence $\{x_k\}$ has a unique limit point in X^*, implying that $\{x_k\}$ converges to some optimal solution with probability 1. **Q.E.D.**

When f has sharp minima, for a diminishing stepsize of the form $\alpha_k = r/(k+1)$ with a positive scalar r, the convergence rate of the method is sublinear, i.e., the expected value of $\big(dist(x_k, X^*)\big)^2$ converges to zero sublinearly. This is shown by Nedić and Bertsekas in [NeB01b].

3.5. DYNAMIC STEPSIZE RULE FOR KNOWN f^*

One possible version of the dynamic stepsize rule for the method (3.1) has the form

$$\alpha_k = \gamma_k \frac{f(x_k) - f^*}{mC^2}, \qquad 0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2,$$

where $\{\gamma_k\}$ is a deterministic sequence. This stepsize requires knowledge of the cost function value $f(x_k)$ at the current iterate x_k. However, it would be inefficient to compute $f(x_k)$ at each iteration since that iteration involves a single component f_i, while the computation of $f(x_k)$ requires all the components. We thus modify the dynamic stepsize rule so that the value of f and the parameter γ_k that are used in the stepsize formula are updated every M iterations rather than at each iteration, where M is any fixed positive integer. In particular, assuming f^* is known, we use the stepsize

$$\alpha_k = \gamma_p \frac{f(x_{Mp}) - f^*}{mMC^2}, \quad 0 < \underline{\gamma} \leq \gamma_p \leq \overline{\gamma} < 2, \quad k = Mp, \ldots, M(p+1) - 1, \quad p = 0, 1, \ldots, \tag{3.7}$$

where $\{\gamma_p\}$ is a deterministic sequence. Thus, we use the same stepsize within each block of M consecutive iterations. We can choose M greater than m, if m is relatively small, or we can select M smaller than m, if m is very large.

We next give a relation that will be used in the forthcoming convergence analysis.

Lemma 3.2: Let Assumption 3.1 hold, and let the sequence $\{x_k\}$ be generated by the randomized method with the stepsize α_k such that

$$\alpha_k = \alpha_{Mp}, \qquad k = Mp, \ldots, M(p+1)-1, \qquad p = 0, 1, \ldots.$$

Then, we have

$$E\big\{\|x_{M(p+1)}-y\|^2 \mid \mathcal{G}_p\big\} \le \|x_{Mp}-y\|^2 - \frac{2M\alpha_{Mp}}{m}\big(f(x_{Mp})-f(y)\big)+M^2\alpha_{Mp}^2C^2, \qquad \forall\, y\in X,\ \ \forall\, p,$$

where $\mathcal{G}_p = \big\{x_0, x_1, \ldots, x_{M(p+1)-1}\big\}$.

Proof: By adapting Lemma 2.1 to the case where f is replaced by f_{ω_k}, we have

$$\|x_{k+1}-y\|^2 \le \|x_k-y\|^2 - 2\alpha_k\big(f_{\omega_k}(x_k)-f_{\omega_k}(y)\big)+\alpha_k^2C^2, \qquad \forall\, y\in X, \quad k\ge 0.$$

Because $\alpha_k = \alpha_{Mp}$ for all $k = Mp, \ldots, M(p+1)-1$, by adding these inequalities over $k = Mp, \ldots, M(p+1)-1$, we obtain for all $y\in X$ and all p,

$$\|x_{M(p+1)}-y\|^2 \le \|x_{Mp}-y\|^2 - 2\alpha_{Mp}\sum_{k=Mp}^{M(p+1)-1}\big(f_{\omega_k}(x_k)-f_{\omega_k}(y)\big)+M\alpha_{Mp}^2C^2.$$

Taking the conditional expectation with respect to $\mathcal{G}_p = \big\{x_0, x_1, \ldots, x_{M(p+1)-1}\big\}$, we have for all $y\in X$ and all p,

$$\begin{aligned} E\big\{\|x_{M(p+1)}-y\|^2 \mid \mathcal{G}_p\big\} &\le \|x_{Mp}-y\|^2 - 2\alpha_{Mp}\sum_{k=Mp}^{M(p+1)-1}E\big\{\big(f_{\omega_k}(x_k)-f_{\omega_k}(y)\big)\mid x_k\big\} \\ &\quad + M^2\alpha_{Mp}^2C^2 \\ &\le \|x_{Mp}-y\|^2 - \frac{2\alpha_{Mp}}{m}\sum_{k=Mp}^{M(p+1)-1}\big(f(x_k)-f(y)\big)+M^2\alpha_{Mp}^2C^2. \end{aligned} \tag{3.8}$$

We now relate $f(x_{Mp})$ and $f(x_k)$ for $k = Mp, \ldots, M(p+1)-1$. We have for all $y\in X$,

$$\begin{aligned} f(x_k)-f(y) &= \big(f(x_k)-f(x_{Mp})\big)+\big(f(x_{Mp})-f(y)\big) \\ &\ge \tilde g_{Mp}'(x_k-x_{Mp})+f(x_{Mp})-f(y) \\ &\ge f(x_{Mp})-f(y)-mC\|x_k-x_{Mp}\|, \qquad \forall\, k = Mp, \ldots, M(p+1)-1, \quad \forall\, p, \end{aligned} \tag{3.9}$$

where $\tilde g_{Mp}$ is a subgradient of f at x_{Mp} and in the last inequality we use the fact

$$\|\tilde g_{Mp}\| = \left\|\sum_{i=1}^m \tilde g_{i,Mp}\right\| \le mC$$

[cf. Assumption 3.1(b)], with $\tilde{g}_{i,Mp}$ being a subgradient of f_i at x_{Mp}. Furthermore, by using Assumption 3.1(b), we see that

$$\|x_k - x_{Mp}\| \le \alpha_{Mp} \sum_{l=Mp}^{k-1} \|g(\omega_l, x_l)\| \le \big(k - Mp\big)\alpha_{Mp} C, \qquad \forall\, k = Mp, \ldots, M(p+1)-1, \quad \forall\, p,$$

which when substituted in Eq. (3.9) yields for all $y \in X$,

$$f(x_k) - f(y) \ge f(x_{Mp}) - f(y) - \big(k - Mp\big) m \alpha_{Mp} C^2, \qquad \forall\, k = Mp, \ldots, M(p+1)-1, \quad \forall\, p.$$

From this relation and Eq. (3.8) we obtain for all $y \in X$ and all p,

$$E\big\{\|x_{M(p+1)} - y\|^2 \mid \mathcal{G}_{p+1}\big\} \le \|x_{Mp} - y\|^2 - \frac{2M\alpha_{Mp}}{m}\big(f(x_{Mp}) - f(y)\big) + 2\alpha_{Mp}^2 C^2 \sum_{k=Mp}^{M(p+1)-1} \big(k - Mp\big) + M\alpha_{Mp}^2 C^2.$$

Since for all p,

$$2\alpha_{Mp}^2 C^2 \sum_{k=Mp}^{M(p+1)-1} \big(k - Mp\big) + M\alpha_{Mp}^2 C^2 = 2\alpha_{Mp}^2 C^2 \sum_{l=1}^{M-1} l + M\alpha_{Mp}^2 C^2 = M^2 \alpha_{Mp}^2 C^2,$$

it follows that for all $y \in X$ and all p,

$$E\big\{\|x_{M(p+1)} - y\|^2 \mid \mathcal{G}_p\big\} \le \|x_{Mp} - y\|^2 - \frac{2M\alpha_{Mp}}{m}\big(f(x_{Mp}) - f(y)\big) + M^2\alpha_{Mp}^2 C^2.$$

Q.E.D.

In the next proposition, assuming that the optimal solution set is nonempty, we show that the method (3.1) with the stepsize (3.7) converges to some optimal solution, with probability 1.

Proposition 3.5: Let Assumption 3.1 hold, and assume that the optimal solution set X^* is nonempty. Then, the sequence $\{x_k\}$ generated by the randomized method with the stepsize (3.7) converges to some optimal solution with probability 1.

Proof: Using Lemma 3.2 (with $y = x^*$) and the definition of α_k, we see that

$$E\big\{\|x_{M(p+1)} - x^*\|^2 \mid \mathcal{G}_p\big\} \le \|x_{Mp} - x^*\|^2 - \gamma_p\big(2 - \gamma_p\big)\frac{\big(f(x_{Mp}) - f^*\big)^2}{m^2 C^2}, \qquad \forall\, x^* \in X^*, \quad \forall\, p.$$

By the Supermartingale Convergence Theorem, it follows that with probability 1,

$$\sum_{p=0}^{\infty} \gamma_p \big(2-\gamma_p\big) \frac{\big(f(x_{Mp})-f^*\big)^2}{m^2C^2} < \infty,$$

and for each $x^* \in X^*$, the sequence $\big\{\|x_{Mp}-x^*\|\big\}$ converges. Because $\gamma_p \in [\underline{\gamma}, \overline{\gamma}] \subset (0,2)$, it follows that with probability 1,

$$\lim_{p\to\infty} f(x_{Mp}) = f^*.$$

For each $x^* \in X^*$, let Ω_{x^*} denote the set of all sample paths for which $f(x_{Mp}) \to f^*$ and $\big\{\|x_{Mp}-x^*\|\big\}$ converges. By convexity of f, it follows that X^* is convex, and therefore there exist the vectors $v_0, v_1, \ldots, v_r \in X^*$ that span the smallest affine set containing X^*, and are such that $v_i - v_0$, $i = 1, \ldots, r$ are linearly independent. The intersection

$$\Omega = \cap_{i=1}^{r} \Omega_{v_i}$$

has probability 1, and for each sample path in Ω, the sequences $\big\{\|x_{Mp}-v_i\|\big\}$, $i = 0, \ldots, r$ converge. Thus, $\{x_{Mp}\}$ is bounded and therefore has limit points, which belong to X^* by continuity of f and the relation $f(x_{Mp}) \to f^*$. For any sample path in Ω, if $\overline{x}$ and $\hat{x}$ are two limit points of $\{x_k\}$, then since $\big\{\|x_{Mp}-v_i\|\big\}$ converges for every $i = 0, \ldots, r$, we must have

$$\|\overline{x}-v_i\| = \|\hat{x}-v_i\|, \qquad \forall\ i = 0, 1, \ldots, r.$$

Since $\overline{x} \in X^*$ and $\hat{x} \in X^*$, the preceding relation can hold only for $\overline{x} = \hat{x}$ by our choice of the vectors v_i. Hence, for each sample path in Ω, the sequence $\big\{x_{Mp}\big\}$ has a unique limit point in X^*, implying that $\big\{x_{Mp}\big\}$ converges to some optimal solution with probability 1. Moreover, by Assumption 3.1(b), it follows that for all p and $k = Mp, \ldots, M(p+1)-1$,

$$\|x_k - x_{Mp}\| \le \alpha_{Mp} \sum_{l=Mp}^{k-1} \|g(\omega_l, x_l)\| \le \big(k-Mp\big)\alpha_{Mp}C \le M\alpha_{Mp}C.$$

By the definition of α_{Mp} and the relation $f(x_{Mp}) \to f^*$, we have $\alpha_{Mp} \to 0$. Therefore, since x_{Mp} converges to some optimal solution with probability 1, from preceding relation we see that the same is true for x_k. **Q.E.D.**

We next give some convergence rate estimates for the randomized method with the dynamic stepsize (3.7). We start with a preliminary result, which we will use here and in the next section for a dynamic stepsize with unknown optimal function value.

Lemma 3.3: Let Assumption 3.1 hold, and assume that the optimal solution set X^* is nonempty. Let the sequence $\{x_k\}$ be generated by the randomized method with the stepsize α_k such that

$$\alpha_k = \alpha_{Mp}, \qquad k = Mp, \ldots, M(p+1)-1, \qquad p = 0, 1, \ldots.$$

Then, we have for all p,

$$E\Big\{\big(dist(x_{M(p+1)},X^*)\big)^2 \mid \mathcal{G}_p\Big\} \leq \big(dist(x_{Mp},X^*)\big)^2 - \frac{2M\alpha_{Mp}}{m}\big(f(x_{Mp})-f^*\big) + M^2\alpha_{Mp}^2C^2,$$

where $\mathcal{G}_p = \big\{x_0, x_1, \ldots, x_{M(p+1)-1}\big\}$.

Proof: The relation follows from Lemma 3.2 with $y = x^*$, by taking the minimum over $x^* \in X^*$, and by using the following inequality

$$E\Big\{\big(dist(x_{M(p+1)},X^*)\big)^2 \mid \mathcal{G}_p\Big\} \leq \min_{x^*\in X^*} E\big\{\|x_{M(p+1)}-x^*\|^2 \mid \mathcal{G}_p\big\}.$$

Q.E.D.

For the method with the dynamic stepsize, we have the following convergence rate result.

Proposition 3.6: Let Assumption 3.1 hold, and assume that the optimal solution set X^* is nonempty. Let the sequence $\{x_k\}$ be generated by the randomized method with the dynamic stepsize (3.7). Then, the following hold:

(a) We have

$$\liminf_{p\to\infty} \sqrt{p}\, E\big\{f(x_{Mp})-f^*\big\} = 0.$$

(b) For any positive scalar ϵ, there exists a random nonnegative integer K such that

$$\min_{0\leq p\leq K} f(x_{Mp}) \leq f^* + \epsilon,$$

with probability 1, and

$$E\{K\} \leq \frac{m^2C^2}{\epsilon^2\underline{\gamma}(2-\overline{\gamma})}\; E\Big\{\big(dist(x_0,X^*)\big)^2\Big\}.$$

Proof: (a) Assume, to arrive at a contradiction, that for some $\epsilon > 0$,

$$\liminf_{p\to\infty} \sqrt{p}\, E\big\{f(x_{Mp})-f^*\big\} = 2\epsilon.$$

Then, there exists p_0 such that

$$E\big\{f(x_{MP})-f^*\big\} \geq \frac{\epsilon}{\sqrt{p}}, \qquad \forall\ p \geq p_0,$$

implying that

$$\sum_{p=p_0}^{\infty} E\Big\{\big(f(x_{Mp})-f^*\big)^2\Big\} \geq \epsilon^2 \sum_{p=p_0}^{\infty} \frac{1}{p} = \infty. \tag{3.10}$$

On the other hand, by using the definition of α_{Mp}, from Lemma 3.3 we have

$$E\Big\{\big(dist(x_{M(p+1)},X^*)\big)^2 \mid \mathcal{G}_p\Big\} \leq \big(dist(x_{Mp},X^*)\big)^2 - \frac{\gamma_p(2-\gamma_p)}{m^2C^2}\big(f(x_{Mp})-f^*\big)^2, \qquad \forall\, p, \tag{3.11}$$

from which by using the relation $\gamma_p \in [\underline{\gamma},\overline{\gamma}]$ for all p and by taking the total expectation, we can see that

$$\sum_{p=0}^{\infty} E\big\{f(x_{Mp})-f^*\big\}^2 < \infty,$$

which contradicts Eq. (3.10). Hence, we must have $\liminf_{p\to\infty}\sqrt{p}\,E\big\{f(x_{pM})-f^*\big\}=0$.

(b) Let the level set L be given by

$$L_\epsilon = \big\{x\in X \mid f(x)\leq f^*+\epsilon\big\},$$

and consider a new process $\{\hat{x}_k\}$ defined by

$$\hat{x}_k = \begin{cases} x_k & \text{if } x_{Mp}\notin L,\\ x^* & \text{otherwise,}\end{cases} \qquad k=Mp,\ldots,M(p+1)-1, \quad p=0,1,\ldots,$$

where $x^*\in X^*$ is some fixed vector. Thus, the process $\{\hat{x}_k\}$ is identical to $\{x_k\}$, except that once x_{Mp} enters the level set L_ϵ, the process $\{\hat{x}_k\}$ remains at the point x^*. Then, for the process $\{\hat{x}_k\}$ it can be seen that [cf. Eq. (3.11)],

$$\begin{aligned} E\Big\{\big(dist(\hat{x}_{M(p+1)},X^*)\big)^2 \mid \mathcal{G}_p\Big\} &\leq \big(dist(\hat{x}_{Mp},X^*)\big)^2 - \frac{\gamma_p(2-\gamma_p)}{m^2C^2}\big(f(\hat{x}_{Mp})-f^*\big)^2\\ &= \big(dist(\hat{x}_{Mp},X^*)\big)^2 - z_p, \qquad \forall\, p, \end{aligned} \tag{3.12}$$

where

$$z_p = \begin{cases} \frac{\gamma_p(2-\gamma_p)}{m^2C^2}\big(f(\hat{x}_{Mp})-f^*\big)^2 & \text{if } \hat{x}_{Mp}\notin L_\epsilon,\\ 0 & \text{otherwise.}\end{cases}$$

In the case where $\hat{x}_{Mp}\notin L_\epsilon$, by definition of α_{Mp}, we have $f(\hat{x}_{Mp})-f^*>\epsilon$, so that

$$z_p > \frac{\epsilon^2\underline{\gamma}(2-\overline{\gamma})}{m^2C^2}, \tag{3.13}$$

and since otherwise $z_p=0$, we see that $z_p\geq 0$ for all p. Therefore, by the Supermartingale Convergence Theorem, from Eq. (3.12) it follows that $\sum_{p=0}^{\infty} z_p<\infty$ with probability 1, implying that $z_p=0$ for all $p\geq K$, where K is a nonnegative random integer. Hence, $\hat{x}_{MK}\in L_\epsilon$ with probability 1, so that in the original process, we have with probability 1,

$$\min_{0\leq p\leq K} f(x_{Mp}) \leq f^*+\epsilon.$$

Furthermore, by taking the total expectation in Eq. (3.12), we obtain

$$E\Big\{\big(dist(\hat{x}_{M(p+1)},X^*)\big)^2\Big\} \le E\Big\{\big(dist(\hat{x}_{Mp},X^*)\big)^2\Big\} - E\{z_p\}$$
$$\le E\Big\{\big(dist(x_0,X^*)\big)^2\Big\} - E\left\{\sum_{j=0}^{p} z_j\right\}, \qquad \forall\ p.$$

Therefore,

$$E\Big\{\big(dist(x_0,X^*)\big)^2\Big\} \ge E\left\{\sum_{j=0}^{\infty} z_j\right\} = E\left\{\sum_{j=0}^{K-1} z_j\right\} \ge E\{K\}\frac{\epsilon^2\underline{\gamma}(2-\overline{\gamma})}{m^2C^2},$$

where the last inequality above follows from Eq. (3.13). **Q.E.D.**

When f has sharp minima, we can obtain a different estimate of the convergence rate for the randomized method with the dynamic stepsize, as seen in the following proposition.

Proposition 3.7: Let Assumption 3.1 hold, and assume that the optimal solution set X^* is nonempty. Assume further that for some positive scalar μ, with probability 1, we have

$$f(x) - f^* \ge \mu\ dist(x,X^*), \qquad \forall\ x \in X.$$

Then, for the sequence $\{x_k\}$ generated by the randomized method with the dynamic stepsize (3.7), we have

$$E\big\{dist(x_{Mp},X^*)\big\} \le r^p\ \sqrt{E\Big\{\big(dist(x_0,X^*)\big)^2\Big\}}, \qquad \forall\ p,$$

where

$$r = \sqrt{1-\underline{\gamma}(2-\overline{\gamma})\frac{\mu^2}{m^2C^2}}.$$

Proof: From Lemma 3.3, by using the definition of α_{Mp}, we obtain

$$E\Big\{\big(dist(x_{M(p+1)},X^*)\big)^2 \mid \mathcal{G}_p\Big\} \le \big(dist(x_{Mp},X^*)\big)^2 - \frac{\gamma_p(2-\gamma_p)}{m^2C^2}\big(f(x_{Mp})-f^*\big)^2, \qquad \forall\ p.$$

By taking the total expectation in this inequality, and by using the given property of f and the relation $\gamma_p \in [\underline{\gamma},\overline{\gamma}]$ for all p, we have

$$E\Big\{\big(dist(x_{M(p+1)},X^*)\big)^2\Big\} \le \left(1-\underline{\gamma}(2-\overline{\gamma})\frac{\mu^2}{m^2C^2}\right)E\Big\{\big(dist(x_{Mp},X^*)\big)^2\Big\}, \qquad \forall\ p,$$

from which the desired estimate follows by induction. **Q.E.D.**

It is difficult to compare the results of Props. 3.6 and 3.7 with the results of the corresponding Props. 2.9 and 2.10. Based on these results, if M is much smaller than m, then the convergence rate of the randomized method is superior. However, for a small M, there is an increased overhead associated with calculating the value of the dynamic stepsize.

3.6. DYNAMIC STEPSIZE RULE FOR UNKNOWN f^*

In the case where f^* is not known, we modify the dynamic stepsize (3.7) by replacing f^* with a target level estimate $f_p^{\rm lev}$. Thus, the stepsize is

$$\alpha_k = \gamma_p \frac{f(x_{Mp}) - f_p^{\rm lev}}{mMC^2}, \quad 0 < \underline{\gamma} \leq \gamma_p \leq \overline{\gamma} < 2, \quad k = Mp, \ldots, M(p+1) - 1, \quad p = 0, 1, \ldots . \tag{3.14}$$

To update the target values $f_p^{\rm lev}$, we may use any of the two adjustment procedures described in Chapter 2. Before we go into analysis of these procedures, let us first establish a preliminary result that applies to both of them.

Lemma 3.4: Let Assumption 3.1 hold and let the sequence $\{x_k\}$ be generated by the randomized method using the stepsize (3.14). Assume that the target values $f_p^{\rm lev}$ in Eq. (3.14) are such that

$$f_p^{\rm lev} = \min_{0 \leq j \leq p} f\big(x_{Mj}\big) - \delta_p, \qquad \forall\, p,$$

where the scalar sequence $\{\delta_p\}$ is positive and nonincreasing. Then, with probability 1, we have

$$\inf_{p \geq 0} f(x_{Mp}) \leq f^* + \lim_{p \to \infty} \delta_p.$$

Proof: Let $\hat{p}$ and N be arbitrary positive integers but fixed, and let $y_N \in X$ be such that

$$f(y_N) = \begin{cases} -N & \text{if } f^* = -\infty, \\ f^* + \frac{1}{N} & \text{if } f^* > -\infty. \end{cases}$$

Define the level set L by

$$L = \big\{x \in X \mid f(x) \leq f(y_N) + \delta_{\hat{p}}\big\},$$

and note that $y_N \in L$. In parallel with the process $\{x_k\}$, consider the process $\{\hat{x}_k\}$ defined by

$$\hat{x}_k = \begin{cases} x_k & \text{if } x_{Mp} \notin L, \\ y_N & \text{otherwise,} \end{cases} \qquad \forall\, k = Mp, \ldots, M(p+1) - 1, \quad \forall\, p.$$

Thus, the process $\{\hat{x}_k\}$ is the same as $\{x_k\}$ up to the time when x_{Mp} enters the level set L, in which case the process $\{\hat{x}_k\}$ remains at y_N.

In view of the definition of $\hat{x}_k$, similar to the proof of Lemma 3.2 with $y = y_N$, we can see that for all $p \geq \hat{p}$,

$$\begin{aligned} E\big\{\|\hat{x}_{M(p+1)} - y_N\|^2 \mid \hat{\mathcal{G}}_p\big\} &\leq \|\hat{x}_{Mp} - y_N\|^2 - \frac{2M\alpha_{Mp}}{m}\big(f(\hat{x}_{Mp}) - f(y_N)\big) + M^2\alpha_{Mp}^2 C^2 \\ &\leq \|\hat{x}_{Mp} - y_N\|^2 - z_p, \end{aligned} \tag{3.15}$$

where $\hat{\mathcal{G}}_p = \{x_{M\hat{p}}, x_{M\hat{p}+1}, \ldots, x_{M(p+1)-1}\}$ and

$$z_p = \begin{cases} \frac{2M\alpha_{Mp}}{m}\big(f(\hat{x}_{Mp}) - f(y_N)\big) - M^2\alpha_{Mp}^2 C^2 & \text{if } \hat{x}_{Mp} \notin L, \\ 0 & \text{if } \hat{x}_{Mp} \in L, \end{cases} \qquad \forall\ p \geq \hat{p}.$$

If $\hat{x}_{Mp} \notin L$ for $p \geq \hat{p}$, then we have by definition of the process $\{\hat{x}_k\}$ that

$$\hat{x}_{Mj} \notin L, \qquad \forall\ j = 0, 1, \ldots, p,$$

which by definition of the set L implies that

$$f\big(\hat{x}_{Mj}\big) > f(y_N) + \delta_{\hat{p}}, \qquad \forall\ j = 0, 1, \ldots, p.$$

Hence,

$$\min_{0\leq j\leq p} f\big(\hat{x}_{Mj}\big) - \delta_{\hat{p}} > f(y_N).$$

Since δ_p is nonincreasing, we have $\delta_p \leq \delta_{\hat{p}}$ for $p \geq \hat{p}$, so that

$$f_p^{\rm lev} = \min_{0\leq j\leq p} f\big(\hat{x}_{Mj}\big) - \delta_p \geq \min_{0\leq j\leq p} f\big(\hat{x}_{Mj}\big) - \delta_{\hat{p}} > f(y_N),$$

and therefore,

$$f(\hat{x}_{Mp}) - f(y_N) > f(\hat{x}_{Mp}) - f_p^{\rm lev}.$$

This relation, and the definitions of z_p and α_{Mp} yield

$$z_p > \gamma_p(2-\gamma_p)\frac{\big(f(\hat{x}_{Mp}) - f_p^{\rm lev}\big)^2}{m^2C^2} \geq \underline{\gamma}(2-\overline{\gamma})\frac{\delta_p^2}{m^2C^2} > 0.$$

Hence, if $\hat{x}_{Mp} \notin L$ for $p \geq \hat{p}$, then $z_p > 0$, and since otherwise $z_p = 0$, it follows that $z_p \geq 0$ for all $p \geq \hat{p}$. From Eq. (3.15) by the Supermartingale Convergence Theorem, we have that $\sum_{p=\hat{p}}^{\infty} z_p < \infty$ with probability 1, implying that $\hat{x}_{Mp} \in L$ for some $p \geq \hat{p}$, with probability 1. Therefore, in the original process, we have with probability 1,

$$\inf_{p\geq 0} f\big(x_{Mp}\big) \leq \begin{cases} -N + \delta_{\hat{p}} & \text{if } f^* = -\infty, \\ f^* + \frac{1}{N} + \delta_{\hat{p}} & \text{if } f^* > -\infty, \end{cases}$$

and by letting $N \to \infty$, we obtain

$$\inf_{p\geq 0} f\big(x_{Mp}\big) \leq \begin{cases} -\infty & \text{if } f^* = -\infty, \\ f^* + \delta_{\hat{p}} & \text{if } f^* > -\infty. \end{cases}$$

Finally, by letting $\hat{p} \to \infty$, the desired relation follows. **Q.E.D.**

We next consider the adjustment procedures of Section 2.6 that are adapted to the stepsize (3.14). In the first adjustment procedure, $f_p^{\rm lev}$ is given by

$$f_p^{\rm lev} = \min_{0\leq j\leq p} f\big(x_{Mj}\big) - \delta_p, \tag{3.16}$$

and δ_p is updated according to

$$\delta_{p+1} = \begin{cases} \delta_p & \text{if } f\big(x_{M(p+1)}\big) \leq f_p^{\text{lev}}, \\ \max\big\{\beta\delta_p, \delta\big\} & \text{if } f\big(x_{M(p+1)}\big) > f_p^{\text{lev}}, \end{cases} \tag{3.17}$$

where δ_0, δ, and β are fixed positive scalars with $\beta < 1$ [note here that the parameter ρ of Eq. (2.23) is set to 1; our results rely on this restriction]. Thus, all the parameters of the stepsize are updated every M iterations. Since the stepsize is bounded away from zero, the method behaves similar to the one with a constant stepsize (cf. Prop. 3.1), as seen in the following proposition.

Proposition 3.8: Let Assumption 3.1 hold. Then, for the sequence $\{x_k\}$ generated by the randomized method and the stepsize (3.14)–(3.17), with probability 1, we have:

(a) If $f^* = -\infty$, then

$$\inf_{p\geq 0} f(x_{Mp}) = f^*.$$

(b) If $f^* > -\infty$, then

$$\inf_{p\geq 0} f(x_{Mp}) \leq f^* + \delta.$$

Proof: We prove (a) and (b) simultaneously. By Lemma 3.4, it follows that with probability 1,

$$\inf_{p\geq 0} f(x_{Mp}) \leq f^* + \lim_{p\to\infty} \delta_p.$$

If $\lim_{p\to\infty} \delta_p = \delta$ with some probability π, then we have with probability π,

$$\inf_{p\geq 0} f(x_{Mp}) \leq f^* + \delta.$$

If $\lim_{p\to\infty} \delta_p > \delta$, which occurs with probability $1-\pi$, then the target level is achieved infinitely many times, i.e., $f\big(x_{M(p+1)}\big) \leq f_p^{\text{lev}}$ infinitely many times, with probability $1-\pi$. Since $\delta_p \geq \delta$ for all p, it follows that the function value is reduced by at least δ infinitely many times. Hence, in this case, we have with probability $1-\pi$,

$$\inf_{p\geq 0} f\big(x_{Mp}\big) = -\infty.$$

Thus, if $f^* = -\infty$, the relation in part (a) holds with probability 1. If $f^* > -\infty$, then we must have $\lim_{p\to\infty} \delta_p > \delta$ with probability 0, thereby implying that the relation in part (b) holds with probability 1. **Q.E.D.**

The target level f_p^{lev} can also be updated according to the second adjustment procedure discussed in Section 2.6, which we adjust for the randomized method that uses the same stepsize within each block of M consecutive iterations. We present this procedure in an algorithmic form, as follows.

Path-Based Randomized Target Level Algorithm

Step 0 (*Initialization*) Select x_0, $\delta_0 > 0$, and $b > 0$. Set $\sigma_0 = 0$, $f_{-1}^{\rm rec} = \infty$. Set $p = 0$, $l = 0$, and $p(l) = 0$ [$p(l)$ will denote the iteration number when the l-th update of $f_p^{\rm lev}$ occurs].

Step 1 (*Function evaluation*) Compute $f(x_{Mp})$. If $f(x_{Mp}) < f_{p-1}^{\rm rec}$, then set $f_p^{\rm rec} = f(x_{Mp})$. Otherwise set $f_p^{\rm rec} = f_{p-1}^{\rm rec}$ [so that $f_p^{\rm rec}$ keeps the record of the smallest value attained by the iterates at the end of each block of M iterations that are generated so far, i.e., $f_p^{\rm rec} = \min_{0\le j\le p} f(x_{Mj})$].

Step 2 (*Sufficient descent*) If $f(x_{Mp}) \le f_{p(l)}^{\rm rec} - \frac{\delta_l}{2}$, then set $p(l+1) = p$, $\sigma_p = 0$, $\delta_{l+1} = \delta_l$, increase l by 1, and go to Step 4.

Step 3 (*Oscillation detection*) If $\sigma_p > b$, then set $p(l+1) = p$, $\sigma_p = 0$, $\delta_{l+1} = \frac{\delta_l}{2}$, and increase l by 1.

Step 4 (*Iterate update*) Set $f_p^{\rm lev} = f_{p(l)}^{\rm rec} - \delta_l$. Select $\gamma_p \in [\underline{\gamma}, \overline{\gamma}]$ and for $k = Mp + 1, \ldots, M(p+1)$, calculate x_k via Eq. (3.1) with the stepsize (3.14).

Step 5 (*Path length update*) Set $\sigma_{p+1} = \sigma_p + \alpha_{Mp} MC$, increase p by 1, and go to Step 1.

For this algorithm, we have the following convergence result.

Proposition 3.9: Let Assumption 3.1 hold. Then, for the sequence $\{x_k\}$ generated by the path-based randomized incremental algorithm, with probability 1, we have

$$\inf_{k\ge 0} f(x_k) = f^*.$$

Proof: We first show that $l \to \infty$ with probability 1. To obtain a contradiction, we assume that l takes only a finite number of values, say $l = 0, 1, \ldots, \overline{l}$, with some positive probability. In this case, we have $\sigma_p + \alpha_{Mp} C = \sigma_{p+1} \le B$ for all $p \ge p(\overline{l})$, so that $\lim_{p\to\infty} \alpha_{Mp} = 0$, with some positive probability. But this is impossible, since for all $p \ge p(\overline{l})$, we have

$$\alpha_{Mp} = \gamma_p \frac{f(x_{Mp}) - f_p^{\rm lev}}{mMC^2} \ge \underline{\gamma}\, \frac{\delta_{\overline{l}}}{mMC^2} > 0.$$

Hence, $l \to \infty$ with probability 1.

By Lemma 3.4, it follows that with probability 1,

$$\inf_{p\ge 0} f(x_{Mp}) \le f^* + \lim_{p\to\infty} \delta_p.$$

If $\lim_{p\to\infty} \delta_p = 0$ with some probability π, then we have with probability π,

$$\inf_{p\ge 0} f(x_{Mp}) \le f^*.$$

If $\lim_{p\to\infty}\delta_p > 0$, which occurs with probability $1-\pi$, then from Steps 2 and 3, it follows that for all l large enough, we have $\delta_l = \delta$ and

$$f^{\rm rec}_{p(l+1)} - f^{\rm rec}_{p(l)} \leq -\frac{\delta}{2},$$

implying that with probability $1-\pi$,

$$\inf_{p\geq 0} f(x_{Mp}) = -\infty.$$

Hence, if $f^* = -\infty$, then the relation in part (a) holds with probability 1. If $f^* > -\infty$, then we must have $\lim_{p\to\infty}\delta_p > 0$ with probability 0, thereby implying that the relation in part (b) holds with probability 1. **Q.E.D.**

In the next proposition, we give a convergence rate result that applies to both procedures.

Proposition 3.10: Let Assumption 3.1 hold, and assume that the optimal solution set X^* is nonempty. Let the sequence $\{x_k\}$ generated by the randomized method with the stepsize (3.14), (3.16), where the scalar sequence $\{\delta_p\}$ is positive and nonincreasing. Then, there exists K is a random nonnegative integer such that

$$\min_{0\leq p\leq K} f(x_{Mp}) \leq f^* + \delta_0,$$

with probability 1, and

$$E\left\{\sum_{p=0}^{K-1}\delta_p^2\right\} \leq \frac{m^2C^2}{\underline{\gamma}(2-\overline{\gamma})}E\Big\{\big(dist(x_0,X^*)\big)^2\Big\}.$$

Proof: Let the level set L be given by

$$L = \big\{x\in X \mid f(x)\leq f^*+\delta_0\big\},$$

and consider a new process $\{\hat{x}_k\}$ defined as follows

$$\hat{x}_k = \begin{cases} x_k & \text{if } x_{Mp}\notin L,\\ x^* & \text{otherwise,}\end{cases} \qquad \forall\ k = Mp,\ldots,M(p+1)-1, \quad \forall\ p=0,1,\ldots,$$

where x^* is a fixed vector in X^*. The process $\{\hat{x}_k\}$ is identical to $\{x_k\}$, except that once x_{Mp} enters the level set L, the process $\{\hat{x}_k\}$ terminates at the point x^*. Then, for the process $\{\hat{x}_k\}$, by Lemma 3.3, we have for all p,

$$E\Big\{\big(dist(x_{M(p+1)},X^*)\big)^2 \mid \mathcal{G}_p\Big\} \leq \big(dist(x_{Mp},X^*)\big)^2 - \frac{2M\alpha_{Mp}}{m}\big(f(x_{Mp})-f(y)\big) + M^2\alpha_{Mp}^2C^2,$$

and by using the definition of α_{Mp}, we obtain

$$E\Big\{\big(dist(x_{M(p+1)},X^*)\big)^2 \mid \mathcal{G}_p\Big\} \le \big(dist(\hat{x}_{Mp},X^*)\big)^2 - z_p, \qquad \forall\, p, \tag{3.18}$$

where

$$z_p = \begin{cases} \frac{\gamma_p}{m^2C^2}\big(f(x_{Mp})-f_p^{\rm lev}\big)\big(f(x_{Mp})-f^*\big) - \frac{\gamma_p^2}{m^2C^2}\big(f(x_{Mp})-f_p^{\rm lev}\big)^2 & \text{if } \hat{x}_{Mp}\notin L, \\ 0 & \text{otherwise,} \end{cases} \qquad \forall\, p.$$

When $\hat{x}_{Mp}\notin L$, we have

$$f(\hat{x}_{Mj}) > f^*+\delta_0, \qquad j=0,\ldots,p,$$

so that

$$f_p^{\rm lev} = \min_{0\le j\le p} f(\hat{x}_{Mj}) - \delta_p \ge f^*+\delta_0-\delta_p \ge f^*,$$

where we use $\delta_0\ge\delta_p$ for all p. Therefore, it follows that

$$f(x_{Mp})-f^* \ge f(x_{Mp})-f_p^{\rm lev}.$$

Using this relation and $f(x_{Mp})-f_p^{\rm lev}\ge\delta_p$, we obtain

$$z_p \ge \frac{\gamma_p(2-\gamma_p)}{m^2C^2}\big(f(x_{Mp})-f_p^{\rm lev}\big)^2 \ge \frac{\delta_p^2\gamma_p(2-\gamma_p)}{m^2C^2} > 0. \tag{3.19}$$

Thus, $z_p>0$ if $\hat{x}_{Mp}\notin L$ and since otherwise $z_p=0$, we see that $z_p\ge 0$ for all p. By the Supermartingale Convergence Theorem, from Eq. (3.18) it follows that $\sum_{p=0}^{\infty} z_p<\infty$ with probability 1, implying that $z_p=0$ for all $p\ge K$, where K is a random nonnegative integer. Hence, $\hat{x}_{MK}\in L$ with probability 1, so that in the original process with probability 1,

$$\min_{0\le p\le K} f(x_{Mp}) \le f^*+\delta_0.$$

Furthermore, by taking the total expectation in Eq. (3.18), we have

$$\begin{aligned} E\Big\{\big(dist(\hat{x}_{M(p+1)},X^*)\big)^2\Big\} &\le E\Big\{\big(dist(\hat{x}_{Mp},X^*)\big)^2\Big\} - E\{z_p\} \\ &\le E\Big\{\big(dist(x_0,X^*)\big)^2\Big\} - E\left\{\sum_{j=0}^{p} z_j\right\}, \qquad \forall\, p. \end{aligned}$$

Using this relation, the definition of z_p, and Eq. (3.19), we see that

$$E\Big\{\big(dist(x_0,X^*)\big)^2\Big\} \ge E\left\{\sum_{k=0}^{\infty} z_k\right\} = E\left\{\sum_{k=0}^{K-1} z_k\right\} \ge E\left\{\sum_{k=0}^{K-1} \frac{\delta_p^2\gamma_p(2-\gamma_p)}{m^2C^2}\right\},$$

and since $\gamma_p \in [\underline{\gamma}, \overline{\gamma}]$ for all p, we obtain

$$E\left\{\sum_{k=0}^{K-1}\delta_p^2\right\} \leq \frac{m^2C^2}{\underline{\gamma}(2-\overline{\gamma})}E\Big\{\big(dist(x_0,X^*)\big)^2\Big\}.$$

Q.E.D.

We can compare the estimate of Prop. 3.10 with that of Prop. 2.13 for the nonrandomized method. For the same initial point x_0, when M is much smaller than m, the convergence rate of the randomized method is better. But for a small M, there is more overhead associated with function evaluation, which is needed to compute the dynamic stepsize.

In both adjustment procedures [cf. Eqs. (3.16) and (3.17), and the path-based algorithm], we have $\delta_0 \geq \delta_p$ for all p, since the scalars δ_p are nonincreasing in each procedure. Thus, the estimate of Prop. 3.10 applies to both procedures. In particular, based on this estimate, we can obtain another upper bound on $E\{K\}$ for the first adjustment procedure [cf. Eqs. (3.16) and (3.17)],

$$E\{K\} \leq \frac{m^2C^2}{\delta^2\underline{\gamma}(2-\overline{\gamma})}E\Big\{\big(dist(x_0,X^*)\big)^2\Big\}.$$

3.7. EFFECTS OF RANDOMIZED PROCESSING ORDER

In the preceding sections, for various stepsize rules, we compared convergence rates of the randomized method with that of the nonrandomized incremental method. To get some further insights into the effects of randomization, let us revisit the examples of Section 2.7.

For an easier reference, let us again state the problem considered in Example 2.2.

Example 3.1: (Convex Nondifferentiable Function)

The problem has the form

$$\begin{aligned} &\text{minimize} \quad f(x) = \sum_{i=1}^{p}|x+1| + \sum_{i=1}^{p}|x-1| + \sum_{i=1}^{2rp}|x| \\ &\text{subject to} \quad x \in \Re, \end{aligned} \tag{3.20}$$

where p and r are positive integers. Here, the optimal solution set consists of a single vector, $x^* = 0$. We consider the randomized method, as applied to this problem, using a small constant stepsize α such that $p\alpha \leq 1$, and the initial point $x_0 = p\alpha$. At each step, an index j is selected from the set $\big\{1,\ldots,2(1+r)p\big\}$ with probability $1/\big(2(1+r)p\big)$ and then the iteration

$$x_{k+1} = x_k - \alpha g_k$$

is executed, where g_k is a subgradient of f_j at the point x_k.

Since the starting point is an integer multiple of α, all the points generated by the algorithm are integer multiples of α. As a result, the algorithm can be modeled by a Markov chain of a random walk type. Let $N > 0$ and $\alpha = 1/N$. Define the states of the chain to be $-N\alpha, (-N+1)\alpha, \ldots, -\alpha, 0, \alpha, \ldots, (N-1)\alpha, N\alpha$. To simplify the notation, we will use i to denote the state $i\alpha$ for $i = -N, -N+1, \ldots, -1, 0, 1, \ldots, N-1, N$. It can be seen that the transition probabilities

$$P_{i,j} = P\{x_{k+1} = j \mid x_k = i\}$$

are given by

$$P_{-N,-N+1} = P_{N,N-1} = 1,$$

$$P_{i-N,i+1-N} = P_{N-i,N-(i+1)} = \frac{1+2r}{2(1+r)}, \qquad i = 1, \ldots, N-1,$$

$$P_{-i,-i-1} = P_{i,i+1} = \frac{1}{2(1+r)}, \qquad i = 0, \ldots, N-1.$$

The stationary probability distribution of the chain is

$$P\{x = i\} = P\{x = -i\} = \rho^i P\{x = 0\}, \qquad i = 1, \ldots, N,$$

$$P\{x = 0\} = \left(1 + 2\sum_{i=1}^{N} \rho^i\right)^{-1},$$

where $\rho = 1/(1+2r)$. For N large, we have $P\{x = 0\} \approx \frac{1-\rho}{1+\rho} = r/(1+r)$ and

$$\begin{aligned} E\{x^2\} &= 2P\{x = 0\} \sum_{i=1}^{N} \alpha^2 i^2 \rho^i \\ &\approx 2\alpha^2 P\{x = 0\} \frac{\rho + \rho^2}{(1-\rho)^3} \\ &\approx 2\alpha^2 \frac{\rho}{(1-\rho)^2} \\ &= 2\alpha^2 \frac{1+2r}{4r^2}, \end{aligned}$$

where we used the estimate $\sum_{i=1}^{N} i^2 q^i \approx \frac{q+q^2}{(1-q)^3}$ for a large scalar N and any scalar $q \in (0,1)$. Since $E\{x\} = 0$, we have

$$\sigma_x \approx \frac{\alpha}{2r}\sqrt{2(1+2r)}.$$

Thus, the standard deviation σ_x of x *does not depend on the value of* p and tends to 0 as r increases.

As seen in Example 2.2, when incremental method is used with a fixed cyclic order, the size of the limit cycle is $p\alpha$ for the worst processing order. If p is large then this distance is also large as compared to the standard deviation σ_x of the randomized method. Hence, the effects of a poor processing order of the components f_i within a cycle can be eliminated by randomization.

The most interesting aspect of the preceding example is that the randomized method has *qualitatively superior* convergence characteristics than the nonrandomized variant (unless the

order of processing the components f_i is favorably chosen). This type of behavior can be seen in other similar (but multidimensional) examples involving nondifferentiable cost functions, and is also evident in our computational experiments to be described in the next chapter.

In the preceding example, the functions f_i are nondifferentiable, and it may be that the behavior of the randomized order method was influenced by nondifferentiability of f_i. As we now revisit Example 2.3, we will see that the behavior is qualitatively different, although still interesting.

Example 3.2: (Convex Differentiable Function)

The problem is

$$\text{minimize} \quad f(x) = \frac{1}{2}\sum_{i=1}^{p}(x-1)^2 + \frac{1}{2}\sum_{i=1}^{p}(x+1)^2$$

$$\text{subject to} \quad x \in \Re.$$

The minimum value of f is attained at $x^* = 0$, which is the unique optimal solution. We assume that the stepsize is equal to some positive constant $\alpha < 2$.

Consider now the variant of the incremental gradient method that selects randomly (with equal probability $\frac{1}{2m}$) the index i from the set $\{1, 2, \ldots, 2m\}$ at each iteration. This method has the form

$$x_{k+1} = x_k - \alpha(x_k - w_k) = (1-\alpha)x_k + \alpha w_k,$$

where w_k takes the value 1 with probability $1/2$ [corresponding to the components with gradient $x-1$], and the value -1 with probability $1/2$ [corresponding to the components with gradient $x+1$]. The second moment of x_k obeys the recursion

$$E\{x_{k+1}^2\} = (1-\alpha)^2 E\{x_k^2\} + \alpha^2 E\{w_k^2\}.$$

The steady-state value of the second moment of x_k (which is also the steady-state value of the variance of x_k, since the expected value $E\{x_k\}$ converges to 0) is given by

$$\lim_{k\to\infty} E\{x_k^2\} = \frac{\alpha^2}{1-(1-\alpha)^2} = \frac{\alpha}{2-\alpha}.$$

Thus, for the standard deviation σ_k of x_k with k large, we have

$$\sigma_k \approx \sqrt{\frac{\alpha}{2-\alpha}}. \tag{3.21}$$

As seen in Example 2.3, for the incremental gradient method that, in each cycle, processes first all components of the form $(x-1)^2$ and then all components of the form $(x+1)^2$, the size of the limit cycle is given by

$$\frac{1-(1-\alpha)^{m/2}}{1+(1-\alpha)^{m/2}}. \tag{3.22}$$

While for the incremental gradient method that, in each cycle, processes a component of the form $(x+1)^2$ followed by a component of the form $(x-1)^2$ and so on until all components are processed, the size of the limit cycle is

$$\frac{\alpha}{2-\alpha}. \tag{3.23}$$

Thus, we see that for small values of α, the randomized method can be worse than the nonrandomized method (with the best processing order) in the sense that the standard deviation in the limit is larger than the size of the limit cycle [cf. Eqs. (3.21) and (3.22)]. For moderate values of α and large enough values of m, we see that the randomized method is as good as the nonrandomized method with the best processing order [cf. Eqs. (3.21) and (3.23)].

As seen from these examples, randomization can alleviate potentially detrimental effect of bad processing order. This is very important for practical problems, where typically the best processing order of the functions f_i cannot be determined.

3.8. EXPERIMENTAL RESULTS

We here present and interpret our experimental results. We first describe our test problem, and the stepsize and the order rules that we used in our experiments. We then compare the incremental subgradient method (2.3)–(2.5) with the ordinary subgradient method (2.2), and the nonrandomized incremental with the randomized incremental method [cf. Eqs. (2.3)–(2.5) and Eq. (3.1), respectively]. Our experimental results show that the randomized methods have substantially better performance, which is consistent with our analytical results of Chapter 4.

3.8.1 Test Problem

In this section, we report some of the numerical results for a certain type of test problem: the dual of a generalized assignment problem (see the book by Martello and Toth [MaT90], p. 189, or Bertsekas [Ber98], p. 362). The problem is to assign m jobs to n machines. If job i is to be performed at machine j, then it costs a_{ij} and requires p_{ij} time units. Given the total available time t_j at machine j, we want to find the minimum cost assignment of the jobs to the machines. Formally, the problem is

$$\begin{aligned}
&\text{minimize} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}y_{ij} \\
&\text{subject to} \quad \sum_{j=1}^{n} y_{ij} = 1, \qquad \forall\, i = 1,\ldots,m, \\
&\qquad\qquad \sum_{i=1}^{m} p_{ij}y_{ij} \le t_j, \qquad \forall\, j = 1,\ldots,n, \\
&\qquad\qquad y_{ij} = 0 \ \text{ or } \ y_{ij} = 1, \qquad \forall\, i = 1,\ldots,m, \quad \forall\, j = 1,\ldots,n,
\end{aligned}$$

where y_{ij} is the assignment variable, which is equal to 1 if the ith job is assigned to the jth machine and is equal to 0 otherwise. In our experiments, we chose n equal to 4 and m equal to one of the four values 500, 800, 4000, or 7000.

By relaxing the time constraints for the machines, we obtain the dual problem

$$\begin{aligned} &\text{maximize} \quad f(x) = \sum_{i=1}^{m} f_i(x) \\ &\text{subject to} \quad x \geq 0, \end{aligned} \tag{3.24}$$

where

$$f_i(x) = \min_{\substack{\sum_{j=1}^{n} y_{ij}=1 \\ y_{ij}=0 \text{ or } y_{ij}=1}} \sum_{j=1}^{n} (a_{ij} + x_j p_{ij}) y_{ij} - \frac{1}{m} \sum_{j=1}^{n} t_j x_j, \qquad \forall\ i = 1, \ldots, m.$$

Since $a_{ij} + x_j p_{ij} \geq 0$ for all i and j, we can easily evaluate $f_i(x)$ for each $x \geq 0$:

$$f_i(x) = a_{ij^*} + x_{j^*} p_{ij^*} - \frac{1}{m} \sum_{j=1}^{n} t_j x_j,$$

where j^* is such that

$$a_{ij^*} + x_{j^*} p_{ij^*} = \min_{1 \leq j \leq n} \{a_{ij} + x_j p_{ij}\}.$$

At the same time, at no additional cost, we obtain a subgradient g of f_i at x:

$$g = (g_1, \ldots, g_n)', \qquad g_j = \begin{cases} -\frac{t_j}{m} & \text{if } j \neq j^*, \\ p_{ij^*} - \frac{t_{j^*}}{m} & \text{if } j = j^*. \end{cases}$$

The experiments are divided in two groups, each with a different goal. The first group was designed to compare the performance of the ordinary subgradient method [cf. Eq. (2.2)] and the incremental subgradient method [cf. Eqs. (2.3)–(2.5)], as applied to the test problem (3.24), for different stepsize rules and a fixed cyclic processing order of the components f_i. The second group of experiments was designed to evaluate the incremental method for a fixed stepsize rule and different rules for the processing order of the components f_i.

3.8.2 Incremental vs. Ordinary Subgradient Method

In the first group of experiments, the data for the problems (i.e., the matrices $[a_{ij}], [p_{ij}]$) were generated randomly according to a uniform distribution over different intervals. The values t_j were calculated according to the formula

$$t_j = \frac{\overline{t}}{n} \sum_{i=1}^{m} p_{ij}, \qquad j = 1, \ldots, n, \tag{3.25}$$

with $\overline{t}$ taking one of the three values 0.5, 0.7, or 0.9. We used two stepsize rules:

(1) A diminishing stepsize of the form

$$\alpha_{kN} = \alpha_{kN+1} = \cdots = \alpha_{(k+1)N-1} = \frac{D}{k+1}, \qquad \forall\ k \geq 0,$$

where D is a positive scalar, and N is a positive integer representing the number of cycles during which the stepsize is kept at the same value. To guard against an unduly large value of D, we implemented an adaptive feature, whereby if within some (heuristically chosen) number S of consecutive iterations the current best cost function value is not improved, then the new iterate x_{k+1} is set equal to the iterate at which the current best value is attained.

(2) The dynamic stepsize rule given by

$$\alpha_k = \frac{f(x_k) - f_k^{\rm lev}}{\|g_k\|^2},$$

and its modification, where $f_k^{\rm lev}$ is adjusted according to the path-based procedure (cf. path-based incremental target level algorithm). In this procedure, the path bound is not fixed but rather the current value for B is multiplied by a certain factor $\xi \in (0,1)$ whenever an oscillation is detected (see the discussion following Prop. 2.12). The initial value for the path bound is $B_0 = r\|x_0 - x_1\|$ for some (heuristically chosen) positive scalar r.

In the forthcoming tables, we report the number of iterations required to achieve a given threshold cost $\tilde{f}$ for various methods and parameter choices. The notation used in the tables is as follows:

$> k \times 100$ for $k = 1, 2, 3, 4$: means that the value $\tilde{f}$ has been achieved or exceeded after $k \times 100$ iterations, but in less than $(k+1) \times 100$ iterations.

> 500: means that the value $\tilde{f}$ has not been achieved within 500 iterations.

$D/N/S/iter$: gives the values of the parameters D, N, and S for the diminishing stepsize rule (1), while $iter$ is the number of iterations (or cycles) needed to achieve or exceed $\tilde{f}$.

$r/\xi/\delta_0/iter$: describes the values of the parameters and number of iterations for the target level stepsize rule (2).

Tables 1 and 2 show the results of applying the ordinary and incremental subgradient methods to problem (3.24) with $n = 4$, $m = 800$, and $\overline{t} = 0.5$ in Eq. (3.25). The optimal value of the problem is $f^* \approx 1578.47$. The threshold value is $\tilde{f} = 1578$. The tables show when the value $\tilde{f}$ was attained or exceeded.

Ordinary subgradient method		
Initial point x_0	Diminishing $D/N/S/iter$	Target level $r/\xi/\delta_0/iter$
(0,0,0,0)	$0.08/2/7/ > 500$	$0.03/0.97/12 \times 10^5/ > 500$
(0,0,0,0)	$0.1/2/7/ > 500$	$0.5/0.98/2 \times 10^4/ > 500$
(0,0,0,0)	$0.07/3/10/ > 500$	$0.5/0.95/3 \times 10^4/ > 500$
(0,0,0,0)	$0.01/10/7/ > 500$	$0.3/0.95/5 \times 10^4/ > 400$
(0,0,0,0)	$0.09/1/7/ > 500$	$0.1/0.9/10^6/ > 200$
(0,0,0,0)	$0.03/5/500/ > 500$	$0.2/0.93/5 \times 10^4/ > 300$
(0,0,0,0)	$0.08/4/7/ > 500$	$0.8/0.97/12 \times 10^3/ > 500$
(0,0,0,0)	$0.09/5/10/ > 500$	$0.03/0.95/10^6/ > 500$
(1.2,1.1,2,1.04)	$0.005/2/5/ > 500$	$0.4/0.975/2 \times 10^4/ > 200$
(1.2,1.1,2,1.04)	$0.009/1/5/ > 500$	$0.5/0.97/4 \times 10^3/ > 50$
$(0.4, 0.2, 1.4, 0.1)$	$0.009/2/5/ > 500$	$0.4/0.8/2700/ > 500$
$(0.4, 0.2, 1.4, 0.1)$	$0.005/5/500/ > 500$	$0.5/0.9/1300/ > 500$

Table 1. $n = 4,\ m = 800,\ f^* \approx 1578.47,\ \tilde{f} = 1578.$

Incremental subgradient method		
Initial point x_0	Diminishing $D/N/S/iter$	Target level $r/\xi/\delta_0/iter$
(0,0,0,0)	$0.05/3/500/99$	$3/0.7/5 \times 10^6/97$
(0,0,0,0)	$0.09/2/500/ > 100$	$2/0.6/55 \times 10^5/ > 100$
(0,0,0,0)	$0.1/1/500/99$	$0.7/0.8/55 \times 10^5/ > 100$
(0,0,0,0)	$0.1/1/10/99$	$0.4/0.95/10^7/80$
(0,0,0,0)	$0.05/5/7/ > 100$	$0.3/0.93/10^7/ > 100$
(0,0,0,0)	$0.07/3/10/ > 100$	$0.5/0.9/10^7/ > 200$
(0,0,0,0)	$0.01/7/7/ > 500$	$0.3/0.93/15 \times 10^6/30$
(0,0,0,0)	$0.009/5/7/ > 500$	$2/0.8/5 \times 10^6/ > 100$
(1.2,1.1,2,1.04)	$0.05/1/500/40$	$0.4/0.97/12 \times 10^6/ > 100$
(1.2,1.1,2,1.04)	$0.04/3/500/35$	$0.3/0.975/10^7/27$
(0.4,0.2,1.4,0.1)	$0.07/1/500/48$	$0.4/0.975/12 \times 10^6/100$
(0.4,0.2,1.4,0.1)	$0.048/1/500/39$	$0.5/0.94/12 \times 10^6/ > 100$

Table 2. $n = 4,\ m = 800,\ f^* \approx 1578.47,\ \tilde{f} = 1578.$

Tables 3 and 4 show the results of applying the ordinary and incremental subgradient methods to problem (3.24) with $n = 4$, $m = 4000$, and $\overline{t} = 0.7$ in Eq. (3.25). The optimal value of the problem is $f^* \approx 6832.3$ and the threshold value is $\tilde{f} = 6831.5$. The tables show the number of iterations needed to attain or exceed the value $\tilde{f} = 6831.5$.

Ordinary subgradient method		
Initial point x_0	Diminishing $D/N/S/iter$	Target level $r/\xi/\delta_0/iter$
(0,0,0,0)	$0.01/2/7/ > 500$	$1/0.9/5000/58$
(0,0,0,0)	$0.001/5/7/ > 300$	$2/0.99/5500/ > 100$
(0,0,0,0)	$0.0008/5/10/ > 300$	$1.3/0.98/4800/54$
(0,0,0,0)	$0.0005/5/7/ > 200$	$1.5/0.98/2000/88$
(0,0,0,0)	$0.0001/5/10/99$	$0.5/0.8/4000/99$
(0,0,0,0)	$0.0001/2/500/ > 100$	$0.4/0.9/4000/89$
(0,0,0,0)	$0.0001/5/10/ > 200$	$0.5/0.9/3000/88$
(0,0,0,0)	$0.00009/5/500/100$	$0.5/0.95/2000/98$
(0.5,0.9,1.3,0.4)	$0.0005/3/500/ > 100$	$0.5/0.98/2000/95$
(0.5,0.9,1.3,0.4)	$0.0002/7/7/ > 100$	$0.4/0.97/3000/98$
(0.26,0.1,0.18,0.05)	$0.0002/5/7/100$	$0.3/0.98/3000/90$
(0.26,0.1,0.18,0.05)	$0.00005/7/7/30$	$0.095/0.985/10/50$

Table 3. $n = 4,\ m = 4000,\ f^* \approx 6832.3,\ \tilde{f} = 6831.5.$

Incremental subgradient method		
Initial point x_0	Diminishing $D/N/S/iter$	Target level $r/\xi/\delta_0/iter$
(0,0,0,0)	$0.005/2/500/46$	$5/0.99/10^6/7$
(0,0,0,0)	$0.007/1/500/37$	$8/0.97/11 \times 10^5/5$
(0,0,0,0)	$0.001/2/500/95$	$2/0.99/7 \times 10^5/ > 100$
(0,0,0,0)	$0.0008/1/500/30$	$0.8/0.4/9 \times 10^5/6$
(0,0,0,0)	$0.0002/2/500/21$	$0.7/0.4/10^6/7$
(0,0,0,0)	$0.0005/2/500/40$	$0.1/0.9/10^6/15$
(0,0,0,0)	$0.0002/2/7/21$	$0.08/0.9/15 \times 10^5/18$
(0,0,0,0)	$0.0003/1/500/21$	$0.25/0.9/2 \times 10^6/20$
(0.5,0.9,1.3,0.4)	$0.001/1/500/40$	$0.07/0.9/10^6/7$
(0.5,0.9,1.3,0.4)	$0.0004/1/500/30$	$0.04/0.9/10^6/26$
(0.26,0.1,0.18,0.05)	$0.00045/1/500/20$	$0.04/0.9/15 \times 10^5/10$
(0.26,0.1,0.18,0.05)	$0.00043/1/7/20$	$0.045/0.91/1.55 \times 10^6/10$

Table 4. $n = 4,\ m = 4000,\ f^* \approx 6832.3,\ \tilde{f} = 6831.5.$

Tables 1 and 2 demonstrate that the incremental subgradient method performs substantially better than the ordinary subgradient method. As m increases, the performance of the incremental method improves as indicated in Tables 3 and 4. The results obtained for other problems that we tested are qualitatively similar and consistently show substantially and often dramatically faster convergence for the incremental method.

3.8.3 Nonrandomized vs. Randomized Incremental Method

We suspected that the random generation of the problem data induced a behavior of the (nonrandomized) incremental method that is similar to the one of the randomized version. Consequently, for the second group of experiments, the coefficients $\{a_{ij}\}$ and $\{p_{ij}\}$ were generated as before and then they were sorted in an nonincreasing order, so as to create a sequential dependence among the data. In all runs, we used the diminishing stepsize choice (as described earlier) with $S = 500$, while the processing order for the components f_i was changed according to the following three rules:

(1) *Fixed Order*. The components are processed in the fixed order $1, 2, \ldots, m$.

(2) *Cyclically Shifted Order*. In the first cycle, the components are processed in the order $1, 2, \ldots, m$. If in the kth cycle, the components are processed in the order $i_1, \ldots, i_m$, then in the $k+1$st cycle, they are processed in the order $i_{K+1}, \ldots, i_m, i_1, \ldots, i_K$, where K is a positive integer K.

(3) *Random Order*. The index of the component to be processed is chosen randomly, with each component equally likely to be selected.

To compare fairly the randomized methods with the other methods, we count as an "iteration" the processing of m consecutively and randomly chosen components f_i. In this way, an "iteration" of the randomized method is equally time-consuming as a cycle or "iteration" of any of the nonrandomized methods.

Table 5 below shows the results of applying the incremental subgradient method with order rules (1)–(3) for solving the problem (3.24) with $n = 4$, $m = 800$, and $\overline{t} = 0.9$ in Eq. (3.25). The optimal value is $f^* \approx 1672.44$ and the threshold value is $\tilde{f} = 1672$. The table gives the number of iterations needed to attain or exceed $\tilde{f}$.

Incremental subgradient method / Diminishing stepsize			
Initial point x_0	Fixed order $D/N/iter$	Cyclically shifted order $D/N/K/iter$	Random order $D/N/iter$
(0,0,0,0)	$0.005/1/ > 500$	$0.007/1/9/ > 500$	0.0095/4/5
(0,0,0,0)	$0.0045/1/ > 500$	$0.0056/1/13/ > 500$	0.08/1/21
(0,0,0,0)	$0.003/2/ > 500$	$0.003/2/7/ > 500$	0.085/1/7
(0,0,0,0)	$0.002/3/ > 500$	$0.002/2/29/ > 500$	0.091/1/17
(0,0,0,0)	$0.001/5/ > 500$	$0.001/6/31/ > 500$	0.066/1/18
(0,0,0,0)	$0.006/1/ > 500$	$0.0053/1/3/ > 500$	0.03/2/18
(0,0,0,0)	$0.007/1/ > 500$	$0.00525/1/11/ > 500$	0.07/1/18
(0,0,0,0)	$0.0009/7/ > 500$	$0.005/1/17/ > 500$	0.054/1/17
(0.2,0.4,0.8,3.6)	$0.001/1/ > 500$	$0.001/1/17/ > 500$	0.01/1/13
(0.2,0.4,0.8,3.6)	$0.0008/3/ > 500$	$0.0008/3/7/ > 500$	0.03/1/8
(0,0.05,0.5,2)	$0.0033/1/ > 400$	$0.0037/1/7/ > 400$	0.033/1/7
(0,0.05,0.5,2)	$0.001/4/ > 500$	$0.0024/2/13/ > 500$	0.017/1/8

Table 5. $n = 4$, $m = 800$, $f^* \approx 1672.44$, $\tilde{f} = 1672$.

The following table 6 shows the results of applying the incremental subgradient method with order rules (1)–(3) for solving the problem (3.24) with $n = 4$, $m = 7000$, and $\overline{t} = 0.5$ in Eq. (3.25). The optimal value is $f^* \approx 14601.38$ and the threshold value is $\tilde{f} = 14600$. The table gives when the value $\tilde{f}$ was attained or exceeded.

Incremental subgradient method / Diminishing stepsize			
Initial point x_0	Fixed order $D/N/iter$	Cyclically shifted order $D/N/K/iter$	Random order $D/N/iter$
(0,0,0,0)	$0.0007/1/ > 500$	$0.0007/1/3/ > 500$	$0.047/1/18$
(0,0,0,0)	$0.0006/1/ > 500$	$0.0006/1/59/ > 500$	$0.009/1/10$
(0,0,0,0)	$0.00052/1/ > 500$	$0.00052/1/47/ > 500$	$0.008/1/2$
(0,0,0,0)	$0.0008/1/ > 500$	$0.0005/1/37/ > 500$	$0.023/1/34$
(0,0,0,0)	$0.0004/2/ > 500$	$0.0004/2/61/ > 500$	$0.0028/1/10$
(0,0,0,0)	$0.0003/2/ > 500$	$0.0003/2/53/ > 500$	$0.06/1/22$
(0,0,0,0)	$0.00025/3/ > 500$	$0.00025/3/11/ > 500$	$0.05/1/18$
(0,0,0,0)	$0.0009/1/ > 500$	$0.00018/3/79/ > 500$	$0.007/1/10$
(0,0.1,0.5,2.3)	$0.0005/1/ > 500$	$0.0005/1/79/ > 500$	$0.004/1/10$
(0,0.1,0.5,2.3)	$0.0003/1/ > 500$	$0.0003/1/51/ > 500$	$0.0007/1/18$
(0,0.2,0.6,3.4)	$0.0002/1/ > 500$	$0.0002/1/51/ > 500$	$0.001/1/10$
(0,0.2,0.6,3.4)	$0.0004/1/ > 500$	$0.00007/2/93/ > 500$	$0.0006/1/10$

Table 6. $n = 4$, $m = 7000$, $f^* \approx 14601.38$, $\tilde{f} = 14600$.

Tables 5 and 6 show how an unfavorable fixed order can have a dramatic effect on the performance of the incremental subgradient method. Note that shifting the components at the beginning of every cycle did not improve the convergence rate of the method. However, the randomization of the processing order resulted in fast convergence. The results for the other problems that we tested are qualitatively similar and also demonstrated the superiority of the randomized method.

3.9. DISTRIBUTED ASYNCHRONOUS INCREMENTAL SUBGRADIENT METHOD

To this end, we considered the incremental subgradient methods in centralized computation. However, for problems where the computation of subgradients of some of the component functions is relatively costly, it is better to parallelize the subgradient computations. For such problems, we here propose and analyze distributed asynchronous incremental subgradient methods, where the computation of the component subgradients is distributed among a set of processors which communicate only with a coordinator. We will first introduce the method and describe the distributed computing system. We will then present convergence results and give their proofs.

3.9.1 The Method

We develop the method departing from the classical subgradient iteration

$$x(t+1) = \mathcal{P}_X \left[x(t) - \alpha(t) \sum_{i=1}^{m} g_i(t) \right], \tag{3.26}$$

where $\alpha(t)$ is a positive stepsize, $g_i(t)$ is a subgradient of f_i at $x(t)$, and $x(0) \in X$ is an initial vector. The most straightforward way to parallelize the above iteration is to use multiple processors to compute in parallel the component subgradients $g_i(t)$. Once all of these components have been computed, they can be collected at a single processor, called the *updating processor*, which will execute the update of the vector $x(t)$ using iteration (3.26). The updating processor will then distribute (broadcast) the new iterate $x(t+1)$ to the subgradient-computing processors which will collectively compute the new subgradients for the subsequent iteration. The parallel method just described is more efficient than the serial method (3.26). It can be termed *synchronous*, in the sense that there is clear division between the computations of successive iterations, i.e., all computation relating to iteration t must be completed before iteration $t+1$ can begin.

A more general method is the parallel method that uses subgradient components not necessarily computed at the same vector $x(t)$. Such a method, termed *asynchronous*, is far more interesting and useful in the situations where some subgradient components $g_i(t)$ are not available at time t, which can be due to, for example, communication delay or excessive computation of some subgradients. In such situations, to avoid further delay in executing the update of $x(t)$, the most recently computed components $g_i\big(\tau_i(t)\big)$ can be used in the iteration (3.26) in place of the missing components $g_i(t)$. A method of this form is the following:

$$x(t+1) = \mathcal{P}_X \left[x(t) - \alpha(t) \sum_{i=1}^{m} g_i\big(\tau_i(t)\big) \right], \tag{3.27}$$

where $\tau_i(t) \leq t$ for all i and the difference $t - \tau_i(t)$ represents the "delay". This method was proposed and analyzed by Kiwiel and Lindberg in [KiL01].

We will here consider a more general method given by

$$x(t+1) = \mathcal{P}_X \left[x(t) - \alpha(t) \sum_{i \in I(t)} g_i\big(\tau_i(t)\big) \right], \tag{3.28}$$

where $I(t)$ is a nonempty subset of the index set $\{1, \ldots, m\}$, $g_i\big(\tau_i(t)\big)$ is a subgradient of f_i computed at $x\big(\tau_i(t)\big)$ with $\tau_i(t) \leq t$ for all i. To visualize the execution of this iteration, it is useful to think of the computing system as consisting of two parts: the *updating system* (US for short), and the *subgradient-computing system* (GCS for short). The US executes iteration (3.28) at each time t and delivers the corresponding iterate $x(t)$ to the GCS. The GCS uses the values $x(t)$ obtained from the US, computes subgradient components $g_i(\tau_i(t))$, and deposits them in

a queue from where they can be accessed by the US. There is no synchronization between the operations of the CS and the GCS. Furthermore, while the GCS may involve multiple processors that compute subgradient components in parallel, the characteristics of the GCS (e.g., shared memory, message passing, communication architecture, number of processors, synchronization rules, etc.) are not material to the description of our algorithm.

The motivation for considering iteration (3.28) rather than its special case (3.27) is twofold. First, it makes sense to keep the US busy with updates while the GCS is computing subgradient components. This is particularly so if the computation of g_i is much more time consuming for some i than for others, thereby creating a synchronization bottleneck. Second, it appears that updating the value of $x(t)$ as quickly as possible and using it in the calculation of the component subgradients has a beneficial effect in the convergence rate of the subgradient method. As seen from the preceding sections, this is the main characteristic of the incremental subgradient methods, which we may view as a special case of the iteration (3.28) where $\tau_i(t) = t$ for all i and the set $I(t)$ consists of a *single* index.

We believe that the incremental structure that is inherent in our proposed parallel subgradient method (3.28) results in convergence and rate of convergence properties that are similar to those of incremental subgradient methods. In particular, we expect an enhanced convergence rate over the nonincremental version given by Eq. (3.27).

In what follows, we will analyze a version of the method (3.28), where the set $I(t)$ in the iteration (3.28) consists of a single element denoted $i(t)$. In particular, we consider the following method

$$x(t+1) = \mathcal{P}_X\Big[x(t) - \alpha(t) g_{i(t)}\big(\tau(t)\big)\Big]. \tag{3.29}$$

For $X = \Re^n$, analysis of this simplified version does not involve an essential loss of generality since an iteration involving multiple component function subgradients may be broken down into several iterations each involving a single component function subgradient. When $X \neq \Re^n$, our analysis can be extended for the more general iteration (3.28). The most important assumptions in our analysis are:

(a) The stepsize $\alpha(t)$ is either constant or is diminishing to 0, and satisfies some common technical conditions such as $\sum_{t=0}^{\infty} \alpha(t) = \infty$ (see a more precise statement later). In the case of a constant stepsize, we only show convergence to optimality within an error which depends on the length of the stepsize.

(b) The "delay" $t - \tau(t)$ is bounded from above by some (unknown) positive integer D, so that our algorithm belongs to the class of *partially asynchronous methods*, as defined by Bertsekas and Tsitsiklis [BeT89].

(c) All the component functions f_i are used with the same "long-term frequency" by the algorithm. Precise methods to enforce this assumption are given later, but basically what we mean is that if $n_i(t)$ is the number of times a subgradient of the component f_i is used by the algorithm up to time t, then the ratios $n_i(t)/t$ should all be asymptotically equal to $1/m$ (as $t \to \infty$).

(d) The subgradients $g_{i(t)}(\tau(t))$ used in the method are uniformly bounded.

The restriction (c) can be easily enforced in a number of ways, by regulating the frequency of the indices of subgradient components computed by the subgradient-computing system. We will consider one specific approach, whereby we first select a sequence of indexes $\{j(t)\}$ according to one of two rules:

(1) *Cyclic Rule.* The sequence $\{j(t)\}$ is obtained by a permutation of each of the periodic blocks $\{1, 2, \ldots, m\}$ in the periodic sequence $\{1, 2, \ldots, m, 1, 2, \ldots, m, \ldots\}$.

(2) *Random rule.* The sequence $\{j(t)\}$ consists of independent identically distributed random variables, each taking the values $1, 2, \ldots, m$ with equal probability $1/m$.

Given a sequence $\{j(t)\}$ obtained by the cyclic or the random rule, the sequence $\{i(t)\}$ used in the iteration (3.29) is given by

$$i(t) = j\big(\pi(t)\big), \tag{3.30}$$

where $\pi(\cdot)$ is a permutation mapping that maps the set $\{0, 1, \ldots\}$ into itself such that for some positive integer T, we have

$$|\pi(t) - t| \leq T, \qquad \forall\ t = 0, 1, \ldots. \tag{3.31}$$

The permutation mapping $\pi(\cdot)$ captures the asynchronous character of the algorithm, whereby component function subgradients are offered to the updating system in the order of $\{j\big(\pi(t)\big)\}$, which is different than the order of $\{j(t)\}$ in which their computation was initiated within the subgradient-computing system. Note that when $\pi(t) = t$ for all t and there is no delay (i.e., $\tau(t) = t$ for all t), then the method (3.29) reduces to the incremental subgradient method.

A version of the algorithm that does not work in this setting is when the component subgradients $g_{i(t)}\big(\tau(t)\big)$ are normalized by multiplying with $1/\|g_{i(t)}(\tau(t))\|$, which may be viewed as a weight associated with the component $f_{i(t)}$ at time t. Unless these weights are asymptotically equal, this modification would effectively alter the "long-term frequency" by which the components f_i are selected, thereby violating a fundamental premise for the validity of our algorithm.

We note that our proposed parallel algorithms (3.28) and (3.29) do not fit in the framework of the general algorithmic models of Chapters 6 and 7 of Bertsekas and Tsitsiklis [BeT89], so these algorithms are not covered by the line of analysis of this reference. In the algorithmic models of Bertsekas and Tsitsiklis [BeT89], at each time t, only *some of the components of* x are updated using an equation that (within our subgradient method context) would depend on *all components* f_i (perhaps with communication delays). By contrast in the present paper, at each time t, *all components of* x are updated using an equation that involves *some of the components* f_i.

The proof ideas of this section are related to those of parallel asynchronous deterministic and stochastic gradient methods as discussed in Tsitsiklis, Bertsekas, and Athans [TBA86], and Bertsekas and Tsitsiklis [BeT89], as well as the proof ideas of incremental deterministic and randomized subgradient methods as discussed in the preceding sections. In particular, the key proof idea is to view the parallel asynchronous method as an iterative method with deterministic or stochastic errors, the effects of which are controlled with an appropriate mechanism, such as a stepsize selection. An alternative approach is possible based on differential inclusions

that extend the "ODE" approach for the analysis of stochastic approximation algorithms (see Benveniste, Metivier, and Priouret [BMP90], Borkar [Bor98], and Kushner and Yin [KuY97]).

3.9.2 Convergence Results for Cyclic Selection Rule

We here give convergence results for the method (3.29) with the cyclic selection rule, under the following assumption.

Assumption 3.2:

(a) There exists a positive constant C such that

$$\|g\| \leq C, \qquad \forall\, g \in \partial f_i\big(x(\tau(t))\big) \cup \partial f_i\big(x(t)\big), \qquad \forall\, i = 1, \ldots, m, \qquad \forall\, t,$$

where $\partial f_i(x)$ denotes the set of subgradients of f_i at a vector x.

(b) There exists a positive integer D such that

$$t - \tau(t) \leq D, \qquad \forall\, t.$$

Note that if the components f_i are polyhedral or if the set X is compact, then Assumption 3.2(a) holds. Assumption 3.2(b) is natural, since our algorithm does not use the value of the bound D.

For the method using a constant stepsize, we have the following result.

Proposition 3.11: Let Assumption 3.2 hold. Then, for the sequence $\{x(t)\}$ generated by the method with the cyclic selection rule and the stepsize fixed to some positive scalar α, we have:

(a) If $f^* = -\infty$, then

$$\liminf_{t\to\infty} f\big(x(t)\big) = -\infty.$$

(b) If f^* is finite, then

$$\liminf_{t\to\infty} f\big(x(t)\big) \leq f^* + mC^2\left(\frac{1}{2} + m + 2D + T\right)\alpha.$$

When $T = 0$ and $D = 0$, in which case the method (3.29) reduces to the incremental subgradient method, the order of the error in part (b) is $m^2C^2\alpha$, thus coinciding with that of the error in Prop. 2.1(b) for the incremental subgradient method.

We next consider a diminishing stepsize that satisfies the following assumption.

Assumption 3.3: The stepsize $\alpha(t)$ is given by

$$\alpha(t) = \frac{r_0}{(l + r_1)^q}, \qquad \forall\, t = \sigma_l, \sigma_l + 1, \ldots, \sigma_{l+1} - 1, \quad \forall\, l = 0, 1, \ldots,$$

where r_0, r_1, and q are some positive scalars with $0 < q \leq 1$, and the sequence $\{\sigma_l\}$ is increasing nd is such that for some positive integer S,

$$\sigma_{l+1} - \sigma_l \leq S, \qquad \forall\, l.$$

For the method (3.29) using this stepsize, we have the following convergence result.

Proposition 3.12: Let Assumptions 3.2 and 3.3 hold. Then, for the sequence $\{x(t)\}$ generated by the method with the cyclic selection rule, we have 1

$$\liminf_{t\to\infty} f\big(x(t)\big) = f^*.$$

When the optimal solution set is nonempty, under a mild additional restriction on the stepsize, we can strengthen the result of Prop. 3.12 by showing that the entire sequence $\{x(t)\}$ converges to some optimal solution. This stronger convergence result is established in the next proposition.

Proposition 3.13: Let Assumptions 3.2 and 3.3 hold, where $1/2 < q \leq 1$ in Assumption 3.3. Assume further that the optimal solution set X^* is nonempty. Then, the sequence $\{x(t)\}$ generated by the method with the cyclic selection rule converges to some optimal solution.

3.9.3 Convergence Results for Random Selection Rule

In this section, we present convergence results for the method (3.29) with the random selection rule. We assume that the stepsize sequence $\{\alpha(t)\}$ deterministic. We also assume the following.

Assumption 3.4:

(a) Assumption 3.2 holds.

(b) Assumption 3.3 holds with a scalar q such that $3/4 < q \leq 1$.

(c) The sequence $\{j(t)\}$ is a sequence of independent random variables each of which is uniformly distributed over the set $\{1, \ldots, m\}$. Furthermore, the sequence $\{j(t)\}$ is independent of the sequence $\{x(t)\}$.

We next give the convergence result for a diminishing stepsize.

Proposition 3.14: Let Assumption 3.4 hold. Then, for the sequence $\{x(t)\}$ generated by the method with the random selection rule, we have with probability 1,

$$\liminf_{t\to\infty} f\big(x(t)\big) = f^*.$$

When the optimal solution set is nonempty, we can strengthen this result, as shown in the following proposition.

Proposition 3.15: Let Assumption 3.4 hold, and assume that the optimal solution set X^* is nonempty. Then, the sequence $\{x(t)\}$ generated by the method with the random selection rule converges to some optimal solution with probability 1.

We note that when the underlying set X is compact, it can be shown that the preceding result holds using a wider range of values for q in the stepsize rule [cf. Assumption 3.4(b)]. In particular, the result is valid for $1/2 < q \leq 1$, which we discuss in more detail in Section 3.9.5.

3.9.4 Convergence Proofs for Cyclic Selection Rule

Here and in the next section, we give the proofs for the convergence results for the selection rules of Sections 3.9.2 and 3.9.3, respectively. The material in these two sections is rather technical, and the reader who is not interested in the proofs can safely skip them.

The proofs are complicated and long, so we break them down into several steps. For notational convenience, we define

$$\alpha(t) = \alpha(0), \qquad \forall\ t < 0,$$

$$t_k = km, \qquad x_k = x(t_k), \qquad \forall\ k \geq 0.$$

We first provide some estimates of the progress of the method in terms of the distances of the iterates to an arbitrary point in the constraint set and in terms of the objective function values. These estimates are given in the subsequent Lemma 3.6. Some preliminary results that are used in the proof of this lemma are given below.

Lemma 3.5: Let Assumption 3.2 hold. Then, we have:

(a) For any $y \in X$ and all t,

$$\begin{aligned}\|x(t+1) - y\|^2 \leq \|x(t) - y\|^2 &- 2\alpha(t)\Big(f_{j(t)}\big(x(t)\big) - f_{j(t)}(y)\Big) \\ &+ C^2(1+4D)\alpha^2(t-D) \\ &+ 2\alpha(t)\sum_{l=1}^{m}\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l\big(x(t)\big) - f_l(y)\Big),\end{aligned}$$

where δ^l_i is the Kronecker symbol (i.e., $\delta^l_i = 1$ if $l = i$ and $\delta^l_i = 0$ otherwise).

(b) For any $y \in X$, and all N and K with $N \geq K$,

$$\sum_{l=1}^{m}\sum_{t=K}^{N}\alpha(t)\Big(\delta^l_{j(t)}-\delta^l_{i(t)}\Big)\Big(f_l\big(x(t)\big)-f_l(y)\Big)\le C^2T\sum_{t=K}^{N}\alpha^2(t-T)$$
$$+\max\big\{C,G(y)\big\}\sum_{t=K}^{N}\Big(\alpha(t-T)-\alpha(t+T)\Big)\left(C\sum_{r=0}^{t}\alpha(r)+\|x_0-y\|\right)$$
$$+c(y)\Big(\alpha^2(K)+\alpha(K)+\alpha^2(N+1-T)+\beta\alpha(N+1-T)\Big)$$
$$+\left(\alpha(K)\|x(K)-y\|^2+\frac{1}{\beta}\alpha(N+1-T)\|x(N+1)-y\|^2\right),$$

where β is an arbitrary positive scalar, and

$$G(y)=\max\big\{\|g\|\mid g\in\partial f_l(y),\ l=1,\ldots,m\big\},\tag{3.32}$$

$$c(y)=\max\left\{CT^2\big(C+G(y)\big),\frac{T^2}{2}\big(C^2+G^2(y)\big)\right\}.\tag{3.33}$$

Proof: (a) From the definition of $x(t+1)$ [cf. Eq. (3.29)], the nonexpansion property of the projection, and the subgradient boundedness [cf. Assumption 3.2(a)], we have

$$\begin{aligned}\|x(t+1)-y\|^2&\le\|x(t)-y\|^2-2\alpha(t)g_{i(t)}\big(\tau(t)\big)'\big(x(t)-y\big)+C^2\alpha^2(t)\\&\le\|x(t)-y\|^2-2\alpha(t)\Big(f_{i(t)}\big(x(t)\big)-f_{i(t)}(y)\Big)\\&\quad+4C\alpha(t)\big\|x(t)-x\big(\tau(t)\big)\big\|+C^2\alpha^2(t),\qquad\forall\ y\in X,\quad\forall\ t.\end{aligned}\tag{3.34}$$

where in the last inequality we use

$$g_{i(t)}\big(\tau(t)\big)'\big(x(t)-y\big)\ge f_{i(t)}\big(x(t)\big)-f_{i(t)}(y)-2C\big\|x(t)-x\big(\tau(t)\big)\big\|,\qquad\forall\ y\in X,\quad\forall\ t,$$

which can be obtained from the fact $x(t)=x\big(\tau(t)\big)+\Big(x(t)-x\big(\tau(t)\big)\Big)$, the convexity of $f_{i(t)}$, and the subgradient boundedness. Furthermore, from the relation

$$\|x(t)-x(\hat t)\|\le C\sum_{s=\hat t}^{t-1}\alpha(s),\qquad\forall\ t,\hat t,\ \ t\ge\hat t,\tag{3.35}$$

and the relations $t-D\le\tau(t)\le t$ and $\alpha(r)\le\alpha(t-D)$ for $r=t-D,\ldots,t-1$ and all t, we obtain

$$\big\|x(t)-x\big(\tau(t)\big)\big\|\le C\sum_{r=t-D}^{t-1}\alpha(r)\le CD\alpha(t-D),\qquad\forall\ t.$$

By using this estimate and the fact $\alpha(t) \leq \alpha(t-D)$ for all t, from Eq. (3.34) we see that for any $y \in X$ and all t,

$$\|x(t+1)-y\|^2 \leq \|x(t)-y\|^2 - 2\alpha(t)\Big(f_{i(t)}\big(x(t)\big) - f_{i(t)}(y)\Big) + C^2(1+4D)\alpha^2(t-D).$$

Finally, by adding and subtracting $2\alpha(t)\Big(f_{j(t)}\big(x(t)\big) - f_{j(t)}(y)\Big)$, and by using the Kronecker symbol, we obtain for any $y \in X$ and all t,

$$\begin{aligned}\|x(t+1)-y\|^2 \leq{}& \|x(t)-y\|^2 - 2\alpha(t)\Big(f_{j(t)}\big(x(t)\big) - f_{j(t)}(y)\Big)\\ &+ C^2(1+4D)\alpha^2(t-D)\\ &+ 2\alpha(t)\sum_{l=1}^{m}\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l\big(x(t)\big) - f_l(y)\Big).\end{aligned}$$

(b) For each K and N with $N \geq K$, we introduce the following sets:

$$M_{K,N} = \Big\{t \in \{K,\ldots,N\} \mid j(t) = i\big(p(t)\big) \text{ with } p(t) \in \{K,\ldots,N\}\Big\},$$

$$P_{K,N} = \Big\{t \in \{K,\ldots,N\} \mid j(t) = i\big(p(t)\big) \text{ with } p(t) < K \text{ or } p(t) > N\Big\}, \tag{3.36}$$

$$Q_{K,N} = \Big\{t \in \{K,\ldots,N\} \mid i(t) = j\big(\pi(t)\big) \text{ with } \pi(t) < K \text{ or } \pi(t) > N\Big\}, \tag{3.37}$$

where $p(t)$ is the inverse of the permutation mapping $\pi(t)$, i.e., $p(t) = \pi^{-1}(t)$. Note that, since $|\pi(t)-t| \leq T$ for all t [cf. Eq. (3.31)], for the inverse mapping $p(t)$ we have

$$|p(t)-t| \leq T, \qquad \forall\, t = 0,1,\ldots.$$

The set $M_{K,N}$ contains all $t \in \{K,\ldots,N\}$ for which the subgradient $g_{j(t)}$ of $f_{j(t)}$ is used in an update of $x(t)$ at some time between K and N. Similarly, the set $P_{K,N}$ contains all $t \in \{K,\ldots,N\}$ for which the subgradient $g_{j(t)}$ of $f_{j(t)}$ is used in an update $x(t)$ at some time before K or after N [i.e., $j(t) = i\big(p(t)\big) \notin \big\{i(K),\ldots,i(N)\big\}$]. The set $Q_{K,N}$ contains all $t \in \{K,\ldots,N\}$ for which the subgradient $g_{i(t)}$ of $f_{i(t)}$ is used in an update $x(t)$ at some time between K and N, but the $j\big(\pi(t)\big)$ corresponding to $i(t)$ does not belong to the set $\big\{j(K),\ldots,j(N)\big\}$. By using the above defined sets, we have

$$\begin{aligned}\sum_{l=1}^{m}\sum_{t=K}^{N}\alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l\big(x(t)\big) - f_l(y)\Big) ={}& \sum_{l=1}^{m}\sum_{t\in M_{K,N}}\alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l\big(x(t)\big) - f_l(y)\Big)\\ &+ \sum_{l=1}^{m}\sum_{t\in P_{K,N}}\alpha(t)\delta^l_{j(t)}\Big(f_l\big(x(t)\big) - f_l(y)\Big)\\ &- \sum_{l=1}^{m}\sum_{t\in Q_{K,N}}\alpha(t)\delta^l_{i(t)}\Big(f_l\big(x(t)\big) - f_l(y)\Big).\end{aligned} \tag{3.38}$$

We now estimate each of the terms in the preceding relation. According to the definition of $M_{K,N}$, we have $j(t) = i\big(p(t)\big)$ for all $t \in M_{K,N}$ [i.e., $g_{j(t)}$ is used at time $p(t)$ with the corresponding step $\alpha\big(p(t)\big)$], so that

$$\begin{aligned}
&\sum_{l=1}^{m}\sum_{t\in M_{K,N}} \alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l\big(x(t)\big) - f_l(y)\Big) \\
&\qquad = \sum_{l=1}^{m}\sum_{t\in M_{K,N}} \delta^l_{j(t)}\Big[\alpha(t)\Big(f_l\big(x(t)\big) - f_l(y)\Big) - \alpha\big(p(t)\big)\Big(f_l\big(x\big(p(t)\big)\big) - f_l(y)\Big)\Big] \\
&\qquad = \sum_{l=1}^{m}\sum_{t\in M_{K,N}} \delta^l_{j(t)}\alpha\big(p(t)\big)\Big(f_l\big(x(t)\big) - f_l\big(x\big(p(t)\big)\big)\Big) \\
&\qquad + \sum_{l=1}^{m}\sum_{t\in M_{K,N}} \delta^l_{j(t)}\Big(\alpha(t) - \alpha\big(p(t)\big)\Big)\Big(f_l\big(x(t)\big) - f_l(y)\Big).
\end{aligned}$$

By using the convexity of each f_l, the subgradient boundedness, the monotonicity of $\alpha(t)$, and the facts $|p(t) - t| \le T$ and $\sum_{l=1}^{m} \delta^l_{j(t)} = 1$ for all t, from the preceding relation we obtain

$$\begin{aligned}
\sum_{l=1}^{m}\sum_{t\in M_{K,N}} \alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l\big(x(t)\big) - f_l(y)\Big) &\le C\sum_{t=K}^{N}\alpha(t-T)\big\|x(t) - x\big(p(t)\big)\big\| \\
&\quad + \max\big\{C, G(y)\big\}\sum_{t=K}^{N}\Big(\alpha(t-T) - \alpha(t+T)\Big)\|x(t) - y\|,
\end{aligned}$$

where $G(y)$ is given by

$$G(y) = \max\big\{\|g\| \mid g \in \partial f_l(y),\ l = 1, \ldots, m\big\}.$$

Furthermore, we have

$$\big\|x(t) - x\big(p(t)\big)\big\| \le CT\alpha(t-T),$$

$$\|x(t) - y\| \le C\sum_{r=0}^{t}\alpha(r) + \|x_0 - y\|,$$

where in the first relation we use the monotonicity of $\alpha(t)$ and the fact $|p(t) - t| \le T$, while in the second relation we use Eq. (3.35). By substituting these relations in the preceding inequality, we have

$$\begin{aligned}
\sum_{l=1}^{m}\sum_{t\in M_{K,N}} \alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l\big(x(t)\big) - f_l(y)\Big) &\le C^2T\sum_{t=K}^{N}\alpha^2(t-T) \\
&\quad + \max\{C, G(y)\}\sum_{t=K}^{N}\Big(\alpha(t-T) - \alpha(t+T)\Big)\left(C\sum_{r=0}^{t}\alpha(r) + \|x_0 - y\|\right).
\end{aligned} \tag{3.39}$$

We next consider the second term on the right hand side of Eq. (3.38). For $t = K, \ldots, N$, we may have $j(t) \notin \{i(K), \ldots, i(N)\}$ possibly at times $t = K, \ldots, K+T-1$ and $t = N+1-T, \ldots, N$. Therefore, from the convexity of each f_l, the subgradient boundedness, and the fact $\sum_{l=1}^m \delta_{j(t)}^l = 1$ for all t, we obtain

$$\sum_{l=1}^{m} \sum_{t \in P_{K,N}} \alpha(t)\delta_{j(t)}^l \Big(f_l\big(x(t)\big) - f_l(y)\Big) \leq C \sum_{t=K}^{K-1+T} \alpha(t)\|x(t)-y\| + C \sum_{t=N+1-T}^{N} \alpha(t)\|x(t)-y\|.$$

By using Eq. (3.35), the triangle inequality, and the monotonicity of $\alpha(t)$, we have

$$\begin{aligned} C \sum_{t=K}^{K-1+T} \alpha(t)\|x(t)-y\| &\leq C \sum_{t=K}^{K-1+T} \alpha(t)\Big(\|x(t)-x(K)\| + \|x(K)-y\|\Big) \\ &\leq C^2T^2\alpha^2(K) + CT\alpha(K)\|x(K)-y\| \\ &\leq C^2T^2\alpha^2(K) + \frac{\alpha(K)}{2}\Big(C^2T^2 + \|x(K)-y\|^2\Big), \end{aligned}$$

where in the last inequality we use the relation $2ab \leq a^2 + b^2$ valid for any scalars a and b. Similarly, it can be seen that

$$\begin{aligned} C \sum_{t=N+1-T}^{N} \alpha(t)\|x(t)-y\| &\leq C \sum_{t=N+1-T}^{N} \alpha(t)\Big(\|x(t)-x(N+1)\| + \|x(N+1)-y\|\Big) \\ &\leq C^2T^2\alpha^2(N+1-T) + CT\alpha(N+1-T)\|x(N+1)-y\| \\ &\leq C^2T^2\alpha^2(N+1-T) \\ &\quad + \frac{\alpha(N+1-T)}{2}\left(\beta C^2T^2 + \frac{1}{\beta}\|x(N+1)-y\|^2\right), \end{aligned}$$

where the last inequality follows from the fact $2ab \leq \beta a^2 + \frac{1}{\beta}b^2$ for any scalars a,b, and β with $\beta > 0$. Therefore,

$$\begin{aligned} \sum_{l=1}^{m} \sum_{t \in P_{K,N}} \alpha(t)\delta_{j(t)}^l \Big(f_l\big(x(t)\big) - f_l(y)\Big) &\leq C^2T^2\Big(\alpha^2(K) + \alpha^2(N+1-T)\Big) \\ &\quad + \frac{C^2T^2}{2}\Big(\alpha(K) + \beta\alpha(N+1-T)\Big) \\ &\quad + \frac{1}{2}\left(\alpha(K)\|x(K)-y\|^2 + \frac{1}{\beta}\alpha(N+1-T)\|x(N+1)-y\|^2\right). \end{aligned} \tag{3.40}$$

Finally, we estimate the last term in Eq. (3.38). For $t = K, \ldots, N$, we may have $i(t) \notin \{j(K), \ldots, j(N)\}$ possibly at times $t = K, \ldots, K+T-1$ and $t = N+1-T, \ldots, N$. Therefore,

similar to the preceding analysis, it can be seen that

$$-\sum_{l=1}^{m}\sum_{t\in Q_{K,N}}\alpha(t)\delta^l_{i(t)}\Big(f_l\big(x(t)\big)-f_l(y)\Big)\leq G(y)CT^2\Big(\alpha^2(K)+\alpha^2(N+1-T)\Big)$$
$$+\frac{G^2(y)T^2}{2}\Big(\alpha(K)+\beta\alpha(N+1-T)\Big)$$
$$+\frac{1}{2}\left(\alpha(K)\|x(K)-y\|^2+\frac{1}{\beta}\alpha(N+1-T)\|x(N+1)-y\|^2\right),$$

where $G(y)$ is given by Eq. (3.32). By substituting Eqs. (3.39)-(3.40) and the preceding relation in the equality (3.38), and by using the definition of $c(y)$ [cf. Eq. (3.33)], we obtain the desired relation. **Q.E.D.**

Lemma 3.6: Let Assumption 3.2 hold. Then, we have:

(a) For any $y\in X$, and all k_0 and $\hat{k}$ with $\hat{k}>k_0$,

$$\left(1-\frac{2}{\beta}\,\alpha(t_{\hat{k}}-W)\right)\|x_{\hat{k}}-y\|^2\leq\Big(1+2\alpha(t_{k_0})\Big)\|x_{k_0}-y\|^2$$
$$-2\sum_{k=k_0}^{\hat{k}-1}\alpha(t_k)\Big(f(x_k)-f(y)\Big)+2\tilde{C}\sum_{k=k_0}^{\hat{k}-1}\alpha^2(t_k-W)$$
$$+2K(y)\sum_{k=k_0}^{\hat{k}-1}\Big(\alpha(t_k-W)-\alpha(t_{k+1}+W)\Big)\left(C\sum_{r=0}^{t_{k+1}}\alpha(r)+\|x_0-y\|\right)$$
$$+2c(y)\Big(\alpha^2(t_{k_0})+\alpha(t_{k_0})+\alpha^2(t_{\hat{k}}-W)+\beta\alpha(t_{\hat{k}}-W)\Big),$$

where $W=\max\{T,D\}$, β is an arbitrary positive scalar,

$$K(y)=mC+m\max\big\{C,G(y)\big\},$$
$$\tilde{C}=mC^2\left(\frac{1}{2}+m+2D+T\right),\tag{3.41}$$

and $G(y)$ and $c(y)$ are defined by Eqs. (3.32) and (3.33), respectively.

(b) For any $y\in X$ and all $\hat{k}\geq 1$,

$$\frac{\sum_{k=0}^{\hat{k}-1}\alpha(t_k)f(x_k)}{\sum_{k=0}^{\hat{k}-1}\alpha(t_k)}\leq f(y)+\frac{\Big(1+2\alpha(0)\Big)\|x_0-y\|^2}{2\sum_{k=0}^{\hat{k}-1}\alpha(t_k)}+\tilde{C}\frac{\sum_{k=0}^{\hat{k}-1}\alpha^2(t_k-W)}{\sum_{k=0}^{\hat{k}-1}\alpha(t_k)}$$
$$+K(y)\frac{\sum_{k=0}^{\hat{k}-1}\Big(\alpha(t_k-W)-\alpha(t_{k+1}+W)\Big)\Big(C\sum_{r=0}^{t_{k+1}}\alpha(r)+\|x_0-y\|\Big)}{\sum_{k=0}^{\hat{k}-1}\alpha(t_k)}$$
$$+c(y)\frac{\Big(\alpha^2(0)+\alpha(0)+\alpha^2(t_{\hat{k}}-W)+\beta\alpha(t_{\hat{k}}-W)\Big)}{\sum_{k=0}^{\hat{k}-1}\alpha(t_k)},$$

where $\beta \geq 2\alpha(0)$.

Proof: (a) By using the convexity of each $f_{j(t)}$, the subgradient boundedness, the monotonicity of $\alpha(t)$, and the following relation [cf. Eq. (3.35)]

$$\|x(t) - x(\hat{t})\| \leq C \sum_{s=\hat{t}}^{t-1} \alpha(s), \qquad \forall\ t, \hat{t},\ t \geq \hat{t},$$

we have for any $t \in \{t_k, \ldots, t_{k+1} - 1\}$,

$$f_{j(t)}\big(x(t)\big) \geq f_{j(t)}(x_k) + g_{j(t)}(t_k)'\big(x(t) - x_k\big) \geq f_{j(t)}(x_k) - mC^2\alpha(t_k),$$

where $g_{j(t)}(t_k)$ is a subgradient of $f_{j(t)}$ at x_k. By substituting this relation in Lemma 3.5(a) and by summing over $t = t_k, \ldots, t_{k+1} - 1$, we obtain

$$\begin{aligned} \|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 &- 2 \sum_{t=t_k}^{t_{k+1}-1} \alpha(t)\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \\ &+ mC^2(1 + 2m + 4D)\alpha^2(t_k - D) \\ &+ 2\sum_{l=1}^{m} \sum_{t=t_k}^{t_{k+1}-1} \alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l\big(x(t)\big) - f_l(y)\Big), \end{aligned} \tag{3.42}$$

where we also use $\alpha(t_k) \leq \alpha(t_k - D)$ and

$$\sum_{t=t_k}^{t_{k+1}-1} \alpha^2(t - D) \leq m\alpha^2(t_k - D), \qquad \sum_{t=t_k}^{t_{k+1}-1} \alpha(t) \leq m\alpha(t_k),$$

which follow from the monotonicity of $\alpha(t)$ and the fact $t_{k+1} - t_k = m$ for all k.

We now estimate the second term on the right hand side in the inequality (3.42). For this we define

$$I_k^+(y) = \Big\{t \in \{t_k, \ldots, t_{k+1} - 1\} \mid f_{j(t)}(x_k) - f_{j(t)}(y) \geq 0\Big\},$$

$$I_k^-(y) = \Big\{t \in \{t_k, \ldots, t_{k+1} - 1\} \mid f_{j(t)}(x_k) - f_{j(t)}(y) < 0\Big\}.$$

Since $\alpha(t_k) \geq \alpha(t) \geq \alpha(t_{k+1})$ for all t with $t_k \leq t < t_{k+1}$, we have for $t \in I_k^+(y)$,

$$\alpha(t)\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \geq \alpha(t_{k+1})\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big),$$

and for $t \in I_k^-(y)$,

$$\alpha(t)\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \geq \alpha(t_k)\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big).$$

Hence, for all t with $t_k \leq t < t_{k+1}$,

$$\begin{aligned}\sum_{t=t_k}^{t_{k+1}-1} \alpha(t)\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) &\geq \alpha(t_{k+1}) \sum_{t\in I_k^+(y)} \Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \\ &\quad + \alpha(t_k) \sum_{t\in I_k^-(y)} \Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \\ &= \alpha(t_k) \sum_{t=t_k}^{t_{k+1}-1} \Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \\ &\quad - \Big(\alpha(t_k) - \alpha(t_{k+1})\Big) \sum_{t\in I_k^+(y)} \Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big).\end{aligned} \tag{3.43}$$

Furthermore, by using the convexity of each $f_{j(t)}$, the subgradient boundedness, and Eq. (3.35), we can see that

$$f_{j(t)}(x_k) - f_{j(t)}(y) \leq C\big(\|x_k - x_0\| + \|x_0 - y\|\big) \leq C\left(C\sum_{r=0}^{t_k}\alpha(r) + \|x_0 - y\|\right),$$

and, since the cardinality of $I_k^+(y)$ is at most m, we obtain

$$\sum_{t\in I_k^+(y)} \Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) \leq mC\left(C\sum_{r=0}^{t_k}\alpha(r) + \|x_0 - y\|\right).$$

For the cyclic rule, we have $\{j(t_k), \ldots, j(t_{k+1}-1)\} = \{1, \ldots, m\}$, so that

$$\sum_{t=t_k}^{t_{k+1}-1} \Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) = f(x_k) - f(y).$$

By using the last two relations, from Eq. (3.43) we obtain

$$\begin{aligned}\sum_{t=t_k}^{t_{k+1}-1} \alpha(t)\Big(f_{j(t)}(x_k) - f_{j(t)}(y)\Big) &\geq \alpha(t_k)\Big(f(x_k) - f(y)\Big) \\ &\quad - mC\Big(\alpha(t_k) - \alpha(t_{k+1})\Big)\left(C\sum_{r=0}^{t_k}\alpha(r) + \|x_0 - y\|\right),\end{aligned}$$

which when substituted in Eq. (3.42) yields for all $y \in X$ and k

$$\begin{aligned}\|x_{k+1} - y\|^2 &\leq \|x_k - y\|^2 - 2\alpha(t_k)\Big(f(x_k) - f(y)\Big) + mC^2(1 + 2m + 4D)\alpha^2(t_k - D) \\ &\quad + 2mC\Big(\alpha(t_k) - \alpha(t_{k+1})\Big)\left(C\sum_{r=0}^{t_k}\alpha(r) + \|x_0 - y\|\right) \\ &\quad + 2\sum_{l=1}^{m}\sum_{t=t_k}^{t_{k+1}-1} \alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l(x(t)) - f_l(y)\Big).\end{aligned}$$

By adding these inequalities over $k = k_0, \ldots, \hat{k}-1$, and by using the facts $t_k < t_{k+1}$, $\alpha(t_k) \leq \alpha(t_k - W)$, $\alpha(t_{k+1}) \geq \alpha(t_{k+1} + W)$, $\alpha^2(t_k - D) \leq \alpha^2(t_k - W)$, we obtain

$$\begin{aligned}\|x_{\hat{k}} - y\|^2 \leq \|x_{k_0} - y\|^2 &- 2\sum_{k=k_0}^{\hat{k}-1} \alpha(t_k)\Big(f(x_k) - f(y)\Big) \\ &+ mC^2(1 + 2m + 4D)\sum_{k=k_0}^{\hat{k}-1} \alpha^2(t_k - W) \\ &+ 2mC\sum_{k=k_0}^{\hat{k}-1}\Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big)\left(C\sum_{r=0}^{t_{k+1}} \alpha(r) + \|x_0 - y\|\right) \\ &+ 2\sum_{k=k_0}^{\hat{k}-1}\sum_{l=1}^{m}\sum_{t=t_k}^{t_{k+1}-1} \alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l(x(t)) - f_l(y)\Big),\end{aligned} \tag{3.44}$$

with $W = \max\{D, T\}$. Note that the last term in the preceding relation can be written as the sum over $l = 1, \ldots, m$ and over $t = t_{k_0}, \ldots, t_{\hat{k}} - 1$, so that by using the monotonicity of $\alpha(t)$, and Lemma 3.5(b) with $K = t_{k_0}$ and $N = t_{\hat{k}} - 1$, we have

$$\begin{aligned}\sum_{l=1}^{m}\sum_{t=t_{k_0}}^{t_{\hat{k}}-1} \alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l\big(x(t)\big) - f_l(y)\Big) &\leq C^2T\sum_{t=t_{k_0}}^{t_{\hat{k}}-1} \alpha^2(t - W) \\ &+ \max\big\{C, G(y)\big\}\sum_{t=t_{k_0}}^{t_{\hat{k}}-1}\Big(\alpha(t - W) - \alpha(t + W)\Big)\left(C\sum_{r=0}^{t} \alpha(r) + \|x_0 - y\|\right) \\ &+ c(y)\Big(\alpha^2(t_{k_0}) + \alpha(t_{k_0}) + \alpha^2(t_{\hat{k}} - W) + \beta\alpha(t_{\hat{k}} - W)\Big) \\ &+ \left(\alpha(t_{k_0})\|x(t_{k_0}) - y\|^2 + \frac{1}{\beta}\alpha(t_{\hat{k}} - W)\|x_{\hat{k}} - y\|^2\right).\end{aligned} \tag{3.45}$$

Furthermore, by the monotonicity of $\alpha(t)$, it follows that

$$\sum_{t=t_{k_0}}^{t_{\hat{k}}-1} \alpha^2(t - W) = \sum_{k=k_0}^{\hat{k}-1}\sum_{t=t_k}^{t_{k+1}-1} \alpha^2(t - W) \leq m\sum_{k=k_0}^{\hat{k}-1} \alpha^2(t_k - W),$$

and

$$\begin{aligned}\sum_{t=t_{k_0}}^{t_{\hat{k}}-1}\Big(\alpha(t - W) - \alpha(t + W)\Big)\left(C\sum_{r=0}^{t} \alpha(r) + \|x_0 - y\|\right) \\ \leq m\sum_{k=k_0}^{\hat{k}-1}\Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big)\left(C\sum_{r=0}^{t_{k+1}} \alpha(r) + \|x_0 - y\|\right).\end{aligned}$$

The desired relation follows from Eq. (3.44) by using Eq. (3.45) and the preceding two relations.

(b) The desired relation follows from part (a), where $k_0 = 0$, by dividing with $2\sum_{k=0}^{\hat{k}-1}\alpha(t_k)$, and by using the relation $\beta \geq 2\alpha(0) \geq 2\alpha(t)$ for all t. **Q.E.D.**

In the forthcoming proofs, we also use the following lemma.

Lemma 3.7: Let $\{\phi_k\}$ and $\{\mu_k\}$ be scalar sequences such that $\mu_k > 0$ for all k and $\sum_{k=0}^{\infty}\mu_k = \infty$. Then, we have

$$\liminf_{k\to\infty}\phi_k \leq \liminf_{k\to\infty}\frac{\sum_{j=0}^{k}\mu_j\phi_j}{\sum_{j=0}^{k}\mu_j} \leq \limsup_{k\to\infty}\frac{\sum_{j=0}^{k}\mu_j\phi_j}{\sum_{j=0}^{k}\mu_j} \leq \limsup_{k\to\infty}\phi_k.$$

In particular, if $\lim_{k\to\infty}\phi_k$ exists, then

$$\lim_{k\to\infty}\phi_k = \lim_{k\to\infty}\frac{\sum_{j=0}^{k}\mu_j\phi_j}{\sum_{j=0}^{k}\mu_j}.$$

Proof: Let ϵ be an arbitrary positive scalar. Then, there exists $\tilde{k}$ large enough so that

$$\liminf_{k\to\infty}\phi_k \leq \phi_j + \epsilon, \qquad \forall\ j \geq \tilde{k},$$

implying that

$$\liminf_{k\to\infty}\phi_k \leq \inf_{j\geq\tilde{k}}\phi_j + \epsilon \leq \frac{\sum_{j=\tilde{k}}^{k}\mu_j\phi_j}{\sum_{j=\tilde{k}}^{k}\mu_j} + \epsilon, \qquad \forall\ k \geq \tilde{k}. \tag{3.46}$$

We further have for all k,

$$\frac{\sum_{j=\tilde{k}}^{k}\mu_j\phi_j}{\sum_{j=\tilde{k}}^{k}\mu_j} = \frac{\sum_{j=0}^{k}\mu_j\phi_j}{\sum_{j=0}^{k}\mu_j}\,\frac{\sum_{j=0}^{k}\mu_j}{\sum_{j=\tilde{k}}^{k}\mu_j} - \frac{\sum_{j=0}^{\tilde{k}-1}\mu_j\phi_j}{\sum_{j=\tilde{k}}^{k}\mu_j},$$

and since by $\sum_{k=0}^{\infty}\mu_k = \infty$, we have that

$$\lim_{k\to\infty}\frac{\sum_{j=\tilde{k}}^{k}\mu_j}{\sum_{j=0}^{k}\mu_j} = 1, \qquad \lim_{k\to\infty}\frac{\sum_{j=0}^{\tilde{k}-1}\mu_j\phi_j}{\sum_{j=\tilde{k}}^{k}\mu_j} = 0,$$

it follows that

$$\liminf_{k\to\infty}\frac{\sum_{j=\tilde{k}}^{k}\mu_j\phi_j}{\sum_{j=\tilde{k}}^{k}\mu_j} = \liminf_{k\to\infty}\frac{\sum_{j=0}^{k}\mu_j\phi_j}{\sum_{j=0}^{k}\mu_j}.$$

From the preceding relation and Eq. (3.46) we see that

$$\liminf_{k\to\infty} \phi_k \leq \liminf_{k\to\infty} \frac{\sum_{j=0}^{k} \mu_j \phi_j}{\sum_{j=0}^{k} \mu_j} + \epsilon,$$

and by letting $\epsilon \to 0$, we have

$$\liminf_{k\to\infty} \phi_k \leq \liminf_{k\to\infty} \frac{\sum_{j=0}^{k} \mu_j \phi_j}{\sum_{j=0}^{k} \mu_j}.$$

Replacing ϕ_k with $-\phi_k$ in the preceding relation, it follows that

$$\limsup_{k\to\infty} \frac{\sum_{j=0}^{k} \mu_j \phi_j}{\sum_{j=0}^{k} \mu_j} \leq \limsup_{k\to\infty} \phi_k.$$

Q.E.D.

We now prove Prop. 3.11.

Proof of Prop. 3.11: It suffices to show (a) and (b) for the sequence $\{x_k\}$. Since $\alpha(t) = \alpha$ for $t \in (-\infty, \infty)$ [recall that $\alpha(t) = \alpha(0)$ for $t < 0$], from Lemma 3.6(b) we obtain

$$\frac{1}{\hat{k}} \sum_{k=0}^{\hat{k}-1} f(x_k) \leq f(y) + \frac{(1+2\alpha)\|x_0 - y\|^2}{2\alpha\hat{k}} + \tilde{C}\alpha + c(y)\frac{1+2\alpha+\beta}{\alpha\hat{k}}, \qquad \forall\ y \in X, \quad \forall\ \hat{k} \geq 1.$$

By letting $\hat{k} \to \infty$ and by using Lemma 3.7 with $\phi_k = f(x_k)$ and $\mu_k = 1/k$, we see that

$$\liminf_{\hat{k}\to\infty} f(x_{\hat{k}}) \leq f(y) + \tilde{C}\alpha, \qquad \forall\ y \in X,$$

from which the desired results follow by taking the minimum over $y \in X$ and by using $\tilde{C} = mC^2(1/2 + m + 2D + T)$ [cf. Eq. (3.41)]. **Q.E.D.**

In the proofs of Props. 3.12 and 3.13, we use some special properties of the stepsize $\alpha(t)$ satisfying Assumption 3.3. These properties are given in the following lemma.

Lemma 3.8: Let the stepsize $\alpha(t)$ satisfy Assumption 3.3. Then, we have

$$\lim_{k\to\infty} \frac{\alpha^2(t_k - W)}{\alpha(t_k)} = 0, \qquad \lim_{k\to\infty} \frac{\alpha(t_k - W) - \alpha(t_{k+1} + W)}{\alpha(t_k)} \sum_{t=0}^{t_{k+1}} \alpha(t) = 0,$$

$$\sum_{k=0}^{\infty} \alpha(t_k) = \infty, \qquad \sum_{k=0}^{\infty} \Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big) < \infty,$$

where $t_k = mk$ and W is a nonnegative integer. In addition, for $1/2 < q \leq 1$, we have

$$\sum_{k=0}^{\infty} \alpha^2(t_k - W) < \infty, \qquad \sum_{k=0}^{\infty} \Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big) \sum_{t=0}^{t_{k+1}} \alpha(t) < \infty.$$

Proof: Let $0 < q \leq 1$. The stepsize $\alpha(t)$ is smallest when $S = 1$, so that

$$\sum_{k=0}^{\infty} \alpha(t_k) \geq r_0 \sum_{k=0}^{\infty} \frac{1}{(km + r_1)^q} = \infty.$$

Let $\{l_k\}$ be a sequence of nonnegative integers such that

$$\alpha(t_k - W) = \frac{r_0}{(l_k + r_1)^q}, \qquad \forall\ k. \tag{3.47}$$

Note that $l_k \to \infty$ as $k \to \infty$. Given the value of $\alpha(t_k - W)$, the values of $\alpha(t_k)$ and $\alpha(t_{k+1} + W)$ are smallest if we decrease the stepsize $\alpha(t)$ at each time t for $t > t_k - W$. Therefore,

$$\alpha(t_k) \geq \frac{r_0}{(l_k + W + r_1)^q}, \qquad \forall\ k, \tag{3.48}$$

$$\alpha(t_{k+1} + W) \geq \frac{r_0}{(l_k + m + 2W + r_1)^q}, \qquad \forall\ k, \tag{3.49}$$

where in the last inequality above we use the fact $t_k = mk$. By combining Eqs. (3.47) and (3.48), we see that

$$\lim_{k \to \infty} \frac{\alpha^2(t_k - W)}{\alpha(t_k)} = 0.$$

Moreover, from Eqs. (3.47) and (3.49) we obtain

$$\begin{aligned} \alpha(t_k - W) - \alpha(t_{k+1} + W) &= r_0 \frac{(l_k + m + 2W + r_1)^q - (l_k + r_1)^q}{(l_k + r_1)^q (l_k + m + 2W + r_1)^q} \\ &\leq \frac{r_0 q (m + 2W)}{(l_k + r_1)(l_k + W + r_1)^q}, \qquad \forall\ k, \end{aligned} \tag{3.50}$$

where in the last inequality above we use $l_k + m + 2W + r_1 \geq l_k + W + r_1$ and

$$b^q - a^q = q \int_a^b \frac{dx}{x^{1-q}} \leq \frac{q}{a^{1-q}} \int_a^b dx = \frac{q(b-a)}{a^{1-q}}, \qquad \forall\ a, b,\ 0 < a \leq b, \quad \forall\ q,\ 0 < q \leq 1.$$

In particular, the relation (3.50) implies that

$$\alpha(t_k - W) - \alpha(t_{k+1} + W) \leq \frac{r_0 q (m + 2W)}{(l_k + r_1)^{1+q}}, \qquad \forall\ k, \tag{3.51}$$

so that $\sum_{k=0}^{\infty}\Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big) < \infty$. Furthermore, by combining Eqs. (3.48) and (3.50), we obtain

$$\frac{\alpha(t_k - W) - \alpha(t_{k+1} + W)}{\alpha(t_k)} \leq \frac{q(m + 2W)}{l_k + r_1}, \qquad \forall\ k. \tag{3.52}$$

We now estimate $\sum_{t=0}^{t_{k+1}} \alpha(t)$. By using the definition and the monotonicity of $\alpha(t)$, and Eq. (3.47), we have for all k large enough (so that $t_k - W > 0$),

$$\sum_{t=0}^{t_{k+1}} \alpha(t) \leq \sum_{t=0}^{t_k - W} \alpha(t) + (1 + m + W)\alpha(t_k - W) \leq \sum_{l=0}^{l_k} \frac{Sr_0}{(l + r_1)^q} + (1 + m + W)\alpha(t_k - W).$$

Since

$$\sum_{l=0}^{l_k} \frac{1}{(l + r_1)^q} \leq \begin{cases} \frac{1}{r_1} + \ln(l_k + r_1) & \text{if } q = 1, \\ \frac{1}{r_1^q} + \frac{(l_k + r_1)^{1-q}}{1-q} & \text{if } 0 < q < 1, \end{cases}$$

from the preceding relation we obtain for all k large enough,

$$\sum_{t=0}^{t_{k+1}} \alpha(t) \leq u_k, \tag{3.53}$$

where

$$u_k = \begin{cases} O\Big(\ln(l_k + r_1)\Big) & \text{if } q = 1, \\ O\Big((l_k + r_1)^{1-q}\Big) & \text{if } 0 < q < 1. \end{cases} \tag{3.54}$$

This together with Eq. (3.52) implies that

$$\lim_{k \to \infty} \frac{\alpha(t_k - W) - \alpha(t_{k+1} + W)}{\alpha(t_k)} \sum_{t=0}^{t_{k+1}} \alpha(t) = 0.$$

Let now $1/2 < q \leq 1$. Then, by using the definition of $\alpha(t)$, we have for K large enough (so that $t_k - W > 0$ for all $k \geq K$)

$$\sum_{k=K}^{\infty} \alpha^2(t_k - W) \leq \sum_{t=K}^{\infty} \alpha^2(t - W) \leq \sum_{s=0}^{\infty} \alpha^2(s) \leq \sum_{l=0}^{\infty} \frac{Sr_1^2}{(l + r_1)^{2q}},$$

implying that $\sum_{k=0}^{\infty} \alpha^2(t_k - W)$ is finite. Furthermore, by combining Eqs. (3.51), (3.53), and (3.54), we obtain

$$\alpha(t_k - W) - \alpha(t_{k+1} + W) \sum_{t=0}^{t_{k+1}} \alpha(t) \leq v_k, \qquad \forall\ k \geq K,$$

where

$$v_k = \begin{cases} O\left(\frac{\ln(l_k+r_1)}{(l_k+r_1)^2}\right) & \text{if } q = 1, \\ O\left(\frac{1}{(l_k+r_1)^{2q}}\right) & \text{if } 0 < q < 1. \end{cases}$$

Hence, $\sum_{k=0}^{\infty} \alpha(t_k - W) - \alpha(t_{k+1} + W) \sum_{t=0}^{t_{k+1}} \alpha(t)$ is finite for $1/2 < q \leq 1$. **Q.E.D.**

We are now ready to prove Props. 3.12 and 3.13.

Proof of Prop. 3.12: It suffices to show that $\liminf_{k\to\infty} f(x_k) = f^*$. From Lemma 3.6(b) by letting $\hat{k} \to \infty$ and by using the relation $\sum_{k=0}^{\infty} \alpha(t_k) = \infty$ [cf. Lemma 3.8], we obtain for all $y \in X$,

$$\liminf_{\hat{k}\to\infty} \frac{\sum_{k=0}^{\hat{k}-1} \alpha(t_k) f(x_k)}{\sum_{k=0}^{\hat{k}-1} \alpha(t_k)} \leq f(y) + \tilde{C} \lim_{\hat{k}\to\infty} \frac{\sum_{k=0}^{\hat{k}-1} \alpha^2(t_k - W)}{\sum_{k=0}^{\hat{k}-1} \alpha(t_k)} + K(y)C \lim_{\hat{k}\to\infty} \frac{\sum_{k=0}^{\hat{k}-1} \Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big) \sum_{r=0}^{t_{k+1}} \alpha(r)}{\sum_{k=0}^{\hat{k}-1} \alpha(t_k)}.$$

Since by Lemma 3.8 we have

$$\lim_{k\to\infty} \frac{\alpha^2(t_k - W)}{\alpha(t_k)} = 0,$$

$$\lim_{k\to\infty} \frac{\alpha(t_k - W) - \alpha(t_{k+1} + W)}{\alpha(t_k)} \sum_{r=0}^{t_{k+1}} \alpha(r) = 0,$$

it follows from Lemma 3.7 that

$$\lim_{\hat{k}\to\infty} \frac{\sum_{k=0}^{\hat{k}-1} \alpha^2(t_k - W)}{\sum_{k=0}^{\hat{k}-1} \alpha(t_k)} = 0,$$

$$\lim_{\hat{k}\to\infty} \frac{\sum_{k=0}^{\hat{k}-1} \Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big) \sum_{r=0}^{t_{k+1}} \alpha(r)}{\sum_{k=0}^{\hat{k}-1} \alpha(t_k)} = 0.$$

Hence,

$$\liminf_{\hat{k}\to\infty} \frac{\sum_{k=0}^{\hat{k}-1} \alpha(t_k) f(x_k)}{\sum_{k=0}^{\hat{k}-1} \alpha(t_k)} \leq f(y), \qquad \forall\ y \in X.$$

Using Lemma 3.7 and taking the minimum over $y \in X$, we see that

$$\liminf_{k\to\infty} f(x_k) \leq \liminf_{\hat{k}\to\infty} \frac{\sum_{k=0}^{\hat{k}-1} \alpha(t_k) f(x_k)}{\sum_{k=0}^{\hat{k}-1} \alpha(t_k)} \leq f^*,$$

thus implying that $\liminf_{k\to\infty} f(x_k) = f^*$. **Q.E.D.**

We next prove Prop. 3.13.

Proof of Prop. 3.13: It suffices to show that $\{x_k\}$ converges to some optimal solution. ¿From Lemma 3.6(a) by letting $\beta = 1$ and $y = x^*$ for some $x^* \in X^*$, and by dropping the nonpositive term involving $f(x_k) - f^*$, we obtain for all $x^* \in X^*$ and $\hat{k} > k_0$,

$$\begin{aligned}\Big(1 - 2\alpha(t_{\hat{k}} - W)\Big)\|x_{\hat{k}} - x^*\|^2 \le \Big(1 + 2\alpha(t_{k_0})\Big)\|x_{k_0} - x^*\|^2 + 2\tilde{C}\sum_{k=k_0}^{\hat{k}-1}\alpha^2(t_k - W)\\ + 2K(x^*)\sum_{k=k_0}^{\hat{k}-1}\Big(\alpha(t_k - W) - \alpha(t_{k+1} + W)\Big)\left(C\sum_{r=0}^{t_{k+1}}\alpha(r) + \|x_0 - x^*\|\right)\\ + 2c(x^*)\Big(\alpha^2(t_{k_0}) + \alpha(t_{k_0}) + \alpha^2(t_{\hat{k}} - W) + \alpha(t_{\hat{k}} - W)\Big),\end{aligned} \tag{3.55}$$

As $\hat{k} \to \infty$, by using Lemma 3.8 with $1/2 < q \le 1$, it follows that

$$\limsup_{\hat{k}\to\infty} \|x_{\hat{k}} - x^*\| < \infty,$$

implying that $\{x_k\}$ is bounded. Furthermore, according to Prop. 3.12, we have

$$\liminf_{k\to\infty} f(x_k) = f^*,$$

so that by continuity of f and by boundedness of $\{x_k\}$, there exists a subsequence $\{x_{k_j}\} \subset \{x_k\}$ and a vector $\hat{x}^* \in X^*$ such that

$$\lim_{j\to\infty} \|x_{k_j} - \hat{x}^*\| = 0.$$

Set $x^* = \hat{x}^*$ and $k_0 = k_j$ for some j in Eq. (3.55). In the resulting relation, by first letting $\hat{k} \to \infty$ and then $j \to \infty$, and by using Lemma 3.8 and the fact $x_{k_j} \to \hat{x}^*$, we obtain

$$\limsup_{\hat{k}\to\infty} \|x_{\hat{k}} - \hat{x}^*\| = 0.$$

Q.E.D.

3.9.5 Convergence Proofs for Random Selection Rule

In this section, we give proofs of Props. 3.14 and 3.15. The proofs rely on the martingale convergence theorem (see, for example, Gallager [Gal96], p. 256).

Theorem 3.2: (Martingale Convergence Theorem) Let $\{Z_k\}$ be a martingale such that $E\left\{Z_k^2\right\} \leq M$ for some positive scalar M and all k. Then, there exists a random variable Z such that with probability 1,

$$\lim_{k\to\infty} Z_k = Z.$$

In the proofs, we also use some properties of the stepsize $\alpha(t)$ that are given in the following lemma.

Lemma 3.9: Let Assumption 3.3 hold with $3/4 < q \leq 1$. Then, we have

$$\sum_{t=0}^{\infty} \alpha^2(t-W) < \infty, \qquad \sum_{t=0}^{\infty} \big(\alpha(t-T) - \alpha(t+T)\big) < \infty, \qquad \sum_{t=0}^{\infty} \alpha(t) = \infty,$$

$$\sum_{t=0}^{\infty} \big(\alpha(t-T) - \alpha(t+T)\big) \sum_{r=0}^{t} \alpha(r) < \infty, \qquad \sum_{t=0}^{\infty} \alpha^2(t) \left(\sum_{r=0}^{t} \alpha(r)\right)^2 < \infty, \tag{3.56}$$

where W is a nonnegative integer.

Proof: We show the last relation in Eq. (3.56). The rest can be shown similar to the proof of Lemma 3.8. Note that $\alpha(t)$ is largest when we change the step every S iterations, i.e., $\sigma_{l+1} - \sigma_l = S$ for all l, so that

$$\alpha(t) \leq \frac{r_0}{(l+r_1)^q}, \qquad t = lS, \ldots, (l+1)S-1, \quad l = 0, 1, \ldots,$$

and consequently,

$$\sum_{r=0}^{t} \alpha(r) \leq Sr_0 \sum_{k=0}^{l} \frac{1}{(k+r_1)^q} \leq \begin{cases} Sr_0 \left(\frac{1}{r_1} + \ln(l+r_1)\right) & \text{if } q = 1, \\ Sr_0 \left(\frac{1}{r_1^q} + \frac{(l+r_1)^{1-q}}{1-q}\right) & \text{if } 0 < q < 1. \end{cases}$$

Therefore,

$$\alpha^2(t) \left(\sum_{r=0}^{t} \alpha(r)\right)^2 \leq w_l, \qquad t = lS, \ldots, (l+1)S-1,$$

where

$$w_l = \begin{cases} O\left(\frac{\ln^2(l+1+r_1)}{(l+r_1)^2}\right) & \text{if } q = 1, \\ O\left(\frac{1}{(l+r_1)^{4q-2}}\right) & \text{if } 0 < q < 1. \end{cases}$$

Hence, $\sum_{t=0}^{\infty} \alpha^2(t) \left(\sum_{r=0}^{t} \alpha(r)\right)^2$ is finite for $3/4 < q \leq 1$. **Q.E.D.**

In the next lemma, we give some relations that are crucial for the subsequent proofs of Props. 3.14 and 3.15.

Lemma 3.10: Let Assumption 3.4 hold. Then, we have:

(a) For any $y \in X$ and all t,

$$\begin{aligned}\|x(t+1)-y\|^2 \le \|x(t)-y\|^2 &- \frac{2\alpha(t)}{m}\Big(f\big(x(t)\big)-f(y)\Big)+2\Big(z_y(t)-z_y(t-1)\Big)\\ &+C^2(1+4D)\alpha^2(t-D)\\ &+2\alpha(t)\sum_{l=1}^{m}\Big(\delta^l_{j(t)}-\delta^l_{i(t)}\Big)\Big(f_l\big(x(t)\big)-f_l(y)\Big),\end{aligned} \tag{3.57}$$

where δ^l_i is the Kronecker symbol, $z_y(-1)=0$, and

$$z_y(t)=\sum_{r=0}^{t}\alpha(r)\left(\frac{1}{m}\Big(f\big(x(r)\big)-f(y)\Big)-\Big(f_{j(r)}\big(x(r)\big)-f_{j(r)}(y)\Big)\right), \qquad \forall\ t\ge 0. \tag{3.58}$$

(b) For any $y \in X$, and all N and K with $N \ge K$,

$$\begin{aligned}\|x(N+1)-y\|^2 \le \|x(K)-y\|^2 &- \frac{2}{m}\sum_{t=K}^{N}\alpha(t)\Big(f\big(x(t)\big)-f(y)\Big)\\ &+2\Big(z_y(N)-z_y(K-1)\Big)\\ &+C^2\Big(1+4D+2T\Big)\sum_{t=K}^{N}\alpha^2(t-W)\\ &+2\max\big\{G(y),C\big\}\sum_{t=K}^{N}\Delta(t)\left(C\sum_{r=0}^{t}\alpha(r)+\|x(0)-y\|\right)\\ &+2c(y)\Big(\alpha^2(K)+\alpha(K)+\alpha^2(N+1-T)+\alpha(N+1-T)\Big)\\ &+2\Big(\alpha(K)\|x(K)-y\|^2+\alpha(N+1-T)\|x(N+1)-y\|^2\Big),\end{aligned}$$

where $W=\max\{D,T\}$, $G(y)$ and $c(y)$ are given by Eqs. (3.32) and (3.33), respectively, and $\Delta(t)=\alpha(t-T)-\alpha(t+T)$ for all t.

(c) For any $y \in X$, the sequence $\{z_y(t)\}$ defined by Eq. (3.58) is a convergent martingale with probability 1.

Proof: (a) From Lemma 3.5(a) by adding and subtracting $\frac{2\alpha(t)}{m}\Big(f(x_k)-f(y)\Big)$, and by using the definition of $z_y(t)$ [cf. Eq. (3.58)], we obtain the relation (3.57).

(b) Summing the inequalities (3.57) over $t = K, \ldots, N$ yields for any $y \in X$,

$$\begin{aligned}\|x(N+1) - y\|^2 \leq \|x(K) - y\|^2 &- \frac{2}{m}\sum_{t=K}^{N} \alpha(t)\Big(f\big(x(t)\big) - f(y)\Big) \\ &+ 2\Big(z_y(N) - z_y(K-1)\Big) + C^2(1+4D)\sum_{t=K}^{N}\alpha^2(t-D) \\ &+ 2\sum_{t=K}^{N}\sum_{l=1}^{m}\alpha(t)\Big(\delta^l_{j(t)} - \delta^l_{i(t)}\Big)\Big(f_l\big(x(t)\big) - f_l(y)\Big).\end{aligned}$$

The desired relation follows by using Lemma 3.5(b) where $\beta = 1$ and $x_0 = x(0)$.

(c) Let $y \in X$ be fixed. We first show that the sequence $\{z_y(t)\}$ is a martingale. By using the definition of $z_y(t)$ [cf. Eq. (3.58)], we have

$$\begin{aligned}E\big\{z_y(t) \mid z_y(t-1)\big\} &= z_y(t-1) + \alpha(t)E\left\{\frac{1}{m}\Big(f\big(x(t)\big) - f(y)\Big) - \Big(f_{j(t)}\big(x(t)\big) - f_{j(t)}(y)\Big)\right\} \\ &= z_y(t-1),\end{aligned}$$

where in the last equality we use the iterated expectation rule and

$$E\left\{\frac{1}{m}\Big(f\big(x(t)\big) - f(y)\Big) - \Big(f_{j(t)}\big(x(t)\big) - f_{j(t)}(y)\Big) \mid x(t)\right\} = 0,$$

which follows from the properties of $\big\{j(t)\big\}$ [cf. Assumption 3.4(c)]. Hence, $z_y(t)$ is indeed a martingale.

We next show that $E\big\{z_y^2(t)\big\}$ is bounded. From the definition of $z_y(t)$ it follows that

$$E\Big\{z_y^2(t)\Big\} = \sum_{r=0}^{t}\alpha^2(r)E\left\{\left(\frac{1}{m}\Big(f(x(r)) - f(y)\Big) - \Big(f_{j(r)}(x(r)) - f_{j(r)}(y)\Big)\right)^2\right\}, \quad \forall\ t \geq 0. \tag{3.59}$$

This is because the expected values of the cross terms appearing in $z_y^2(t)$ are equal to 0, which can be seen by using the iterated expectation rule [i.e., by conditioning on the values $x(s)$ and $x(r)$ for $s, r \leq t$] and by exploiting the properties of $j(r)$ [cf. Assumption 3.4(c)]. Furthermore, by using convexity of each f_i, the triangle inequality, and the following relation [cf. Eq. (3.35)]

$$\|x(t) - x(\hat{t})\| \leq C\sum_{s=\hat{t}}^{t-1}\alpha(s), \qquad \forall\ t, \hat{t},\ t \geq \hat{t},$$

for every r, we have

$$\begin{aligned}\Big(\frac{1}{m}\Big(f\big(x(r)\big)-f(y)\Big)-\Big(f_{j(r)}\big(x(r)\big)-f_{j(r)}(y)\Big)\Big)^2 &\le \frac{2}{m^2}\Big(f\big(x(r)\big)-f(y)\Big)^2\\ &\quad+2\Big(f_{j(r)}\big(x(r)\big)-f_{j(r)}(y)\Big)^2\\ &\le 4\Big(\max\big\{G(y),C\big\}\Big)^2\|x(r)-y\|^2\\ &\le 4\Big(\max\big\{G(y),C\big\}\Big)^2\left(C\sum_{s=0}^{r}\alpha(s)+\|x(0)-y\|\right)^2\\ &\le 8\Big(\max\big\{G(y),C\big\}\Big)^2\left[C^2\left(\sum_{s=0}^{r}\alpha(s)\right)^2+\|x(0)-y\|^2\right].\end{aligned}$$

By using this inequality and Lemma 3.9, from Eq. (3.59) it follows that $E\big\{z_y^2(t)\big\}$ is bounded. Thus, by the Martingale Convergence Theorem, the sequence $\{z_y(t)\}$ converges to some random variable with probability 1. **Q.E.D.**

We now prove Prop. 3.14.

Proof of Prop. 3.14: Let $\epsilon>0$ be arbitrary and let $\hat{y}\in X$ be such that

$$f(\hat{y})\le\begin{cases}f^*+\epsilon & \text{if } f^* \text{ is finite,}\\ -\frac{1}{\epsilon} & \text{otherwise.}\end{cases}$$

Fix a sample path, denoted by $\mathcal{P}$, for which the martingale $\big\{z_{\hat{y}}(t)\big\}$ is convergent [cf. Lemma 3.10(c)]. From Lemma 3.10(b), where $K=0$ and $y=\hat{y}$, we have for the path $\mathcal{P}$ and all N sufficiently large,

$$\begin{aligned}\frac{2}{m}\sum_{t=0}^{N}\alpha(t)\Big(f(x(t))-f(\hat{y})\Big) &\le \Big(1+2\alpha(0)\Big)\|x(0)-\hat{y}\|^2+2z_{\hat{y}}(N)\\ &\quad+C^2\Big(1+4D+2T\Big)\sum_{t=0}^{N}\alpha^2(t-W)\\ &\quad+2\max\{C,G(\hat{y})\}\sum_{t=0}^{N}\Delta(t)\left(C\sum_{r=0}^{t}\alpha(r)+\|x(0)-\hat{y}\|\right)\\ &\quad+2c(\hat{y})\Big(\alpha^2(0)+\alpha(0)+\alpha^2(N+1-T)+\alpha(N+1-T)\Big),\end{aligned}$$

where $\Delta(t)=\alpha(t-T)-\alpha(t+T)$, we use the fact $z_{\hat{y}}(-1)=0$, and we take N sufficiently large so that $1-2\alpha(N+1-T)\ge 0$. Since $z_{\hat{y}}(N)$ converges, by dividing the above inequality with $(2/m)\sum_{t=0}^{N}\alpha(t)$, by letting $N\to\infty$, and by using Lemma 3.9, we obtain

$$\liminf_{N\to\infty}\frac{\sum_{t=0}^{N}\alpha(t)f\big(x(t)\big)}{\sum_{t=0}^{N}\alpha(t)}\le f(\hat{y}).$$

By Lemma 3.7, we have

$$\liminf_{N\to\infty} f\big(x(N)\big) \leq \liminf_{N\to\infty} \frac{\sum_{t=0}^{N} \alpha(t) f\big(x(t)\big)}{\sum_{t=0}^{N} \alpha(t)},$$

so that for the path $\mathcal{P}$,

$$\liminf_{N\to\infty} f\big(x(N)\big) \leq f(\hat{y}).$$

Therefore, $\liminf_{t\to\infty} f\big(x(t)\big) \leq f(\hat{y})$ with probability 1, implying by definition of $\hat{y}$ that with probability 1,

$$\liminf_{t\to\infty} f\big(x(t)\big) \leq \begin{cases} f^* + \epsilon & \text{if } f^* \text{ is finite,} \\ -\frac{1}{\epsilon} & \text{otherwise.} \end{cases}$$

Since ϵ is arbitrary, it follows that $\liminf_{t\to\infty} f(x(t)) = f^*$ with probability 1. **Q.E.D.**

We next give the proof of Prop. 3.15.

Proof of Prop. 3.15: For each $x^* \in X^*$, let Ω_{x^*} denote the set of all sample paths for which the sequence $\{z_{x^*}(t)\}$ is a convergent martingale and

$$\liminf_{t\to\infty} f(x(t)) = f^*. \tag{3.60}$$

By Lemma 3.10(c) and Prop. 3.14, the set Ω_{x^*} has probability 1 for each $x^* \in X^*$. Since f and X are convex, the set X^* is also convex, so there exist vectors $v_0, v_1, \ldots, v_p \in X^*$ that span the smallest affine set containing X^*, and are such that $v_i - v_0$, $i = 1, \ldots, p$, are linearly independent. The intersection

$$\Omega = \cap_{i=1}^{p} \Omega_{v_i}$$

has probability 1.

We now fix a sample path $\mathcal{P} \in \Omega$, for which by definition of Ω, every martingale $z_{v_i}(t)$ is convergent and the relation (3.60) holds. Furthermore, we fix an $i \in \{0, \ldots, p\}$. Let K_0 be a positive integer large enough so that

$$1 - 2\alpha(K - T) > 0, \qquad \forall\ K \geq K_0.$$

By using Lemma 3.10(b) with $y = v_i$ and $N > K \geq K_0$, and by dropping the nonnegative term involving $f\big(x(t)\big) - f(v_i)$, we obtain

$$\begin{aligned}\Big(1 - 2\alpha(N+1-T)\Big)\|x(N+1) - v_i\|^2 \leq{}& \Big(1 + 2\alpha(K)\Big)\|x(K) - v_i\|^2 + 2\Big(z_s(N) - z_s(K-1)\Big) \\ &+ C^2\Big(1 + 4D + 2T\Big)\sum_{t=K}^{N} \alpha^2(t - W) \\ &+ 2\max\{G(v_i), C\}\sum_{t=K}^{N} \Delta(t)\left(C\sum_{r=0}^{t} \alpha(r) + \|x(0) - v_i\|\right) \\ &+ 2c(v_i)\Big(\alpha^2(K) + \alpha(K) + \alpha^2(N+1-T) + \alpha(N+1-T)\Big).\end{aligned}$$

By using Lemma 3.9 and the convergence of the martingale $\{z_{v_i}(t)\}$, from the preceding relation we obtain

$$\limsup_{N\to\infty} \|x(N+1) - v_i\|^2 \leq \liminf_{K\to\infty} \|x(K) - v_i\|^2.$$

Because i is arbitrary, $\lim_{t\to\infty} \|x(t) - v_i\|$ exists for all $i = 0, \ldots, p$. Furthermore, $\{x(t)\}$ is bounded so it has limit points, at least one of which must belong to X^* by Eq. (3.60) and continuity of f. Let $\overline{x} \in X^*$ be such a limit point. If $\hat{x}$ is another limit point of $\{x(t)\}$, then since $\{\|x_k - v_i\|\}$ converges for all $i = 0, \ldots, p$, we must have

$$\|\overline{x} - v_i\| = \|\hat{x} - v_i\|, \qquad \forall\ i = 0, 1, \ldots, p.$$

Since $\overline{x} \in X^*$, by convexity of X^* and the choice of vectors v_i, the preceding relation can hold only for $\overline{x} = \hat{x}$. Hence, for the path $\mathcal{P}$, the sequence $\{x(t)\}$ has a unique limit point in X^*, implying that $\{x(t)\}$ converges to some optimal solution with probability 1. **Q.E.D.**

When the constraint set X is compact, the result of Prop. 3.15 holds under Assumption 3.4(b) with $1/2 < q \leq 1$ instead of $3/4 < q \leq 1$. This can be seen similar to the preceding analysis by using the fact that the martingale $\{z_y(t)\}$ is convergent for $1/2 < q \leq 1$ [cf. Lemma 3.10(c)]. In particular, for $1/2 < q \leq 1$, we can show that

$$E\{z_y^2(t)\} \leq 4\Big(\max\{G(y), C\}\Big)^2 \sup_{x\in X} \|x - y\|^2 \sum_{r=0}^{t} \alpha^2(r)$$

[see the proof of Lemma 3.10(c)].

4

Extensions of the Incremental Subgradient Method

In this chapter, we consider the incremental subgradient method with some special features. In particular, in Sections 4.1 and 4.2, respectively, we discuss two variants of the incremental method: one with weights and one with approximate subgradients (ϵ-subgradients).

4.1. AN INCREMENTAL SUBGRADIENT METHOD WITH WEIGHTS

Here, we consider an incremental subgradient method with weights. Just like the pure incremental method of Section 2.1, this method also operates in cycles, but it uses directions that are different from that of the pure incremental method. In particular, at the ith subiteration of a cycle, the direction used is a weighted sum of a subgradient of a newly selected component f_i and the subgradients of components $f_1, f_2 \ldots, f_{i-1}$ that have already been processed within the current cycle. More precisely, in a typical cycle, the method starts with

$$\psi_{0,k} = x_k, \tag{4.1}$$

performs m subiterations

$$\psi_{i,k} = \mathcal{P}_X \left[\psi_{i-1,k} - \alpha_k \sum_{j=1}^{i} w_{i,j}^k g_{j,k} \right], \qquad i = 1, \ldots, m, \tag{4.2}$$

where the scalar α_k is a positive stepsize, the scalars $w^k_{i,1}, \ldots, w^k_{i,i}$ are nonnegative weights, and the vector $g_{j,k}$ is a subgradient of f_j at $\psi_{j-1,k}$. The last of these subiterations is the beginning of a new cycle

$$x_{k+1} = \psi_{m,k}. \tag{4.3}$$

The incremental method of Section 2.1 is a special case of this method corresponding to the case where

$$w^k_{i,1} = 0, \ldots, w^k_{i,i-1} = 0, w^k_{i,i} = 1, \qquad \forall\, i = 1, \ldots, m, \quad \forall\, k.$$

4.1.1 Assumptions and Basic Relation

Regarding the method (4.1)–(4.3), we assume the following:

Assumption 4.1:

(a) The weights $w^k_{i,j}$ are nonnegative and

$$\sum_{i=j}^{m} w^k_{i,j} = 1, \qquad \forall\, j = 1, \ldots, m, \quad \forall\, k.$$

(b) There exists a positive scalar C such that

$$\|g\| \leq C, \qquad \forall\, g \in \partial f_i(x_k) \cup \partial f_i(\psi_{i-1,k}), \quad \forall\, i = 1, \ldots, m, \quad \forall\, k.$$

Assumption 4.1(a) says that in each cycle k, the sum of all weights corresponding to the same subgradient $g_{j,k}$ is equal to 1. A possible choice of such weights is the one where the weights corresponding to the same subgradient $g_{j,k}$ are all equal, i.e.,

$$w^k_{i,j} = \frac{1}{m-j+1}, \qquad \forall\, j = 1, \ldots, m, \quad \forall\, i = j, \ldots, m, \quad \forall\, k.$$

In the next lemma, we establish an important relation between the iterates obtained at the beginning and the end of a cycle.

Lemma 4.1: Let Assumption 4.1 hold and let $\{x_k\}$ be the sequence generated by the incremental subgradient method with weights. We then have

$$\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k \big(f(x_k) - f(y)\big) + \alpha_k^2 m^2 C^2, \qquad \forall\, y \in X, \quad \forall\, k,$$

where C is as in Assumption 4.1(b).

Proof: By using the nonexpansion property of the projection, we obtain for all $y \in X$, and all i and k,

$$\begin{aligned}\|\psi_{i,k} - y\|^2 &= \left\| \mathcal{P}_X \left[\psi_{i-1,k} - \alpha_k \sum_{j=1}^{i} w_{i,j}^k g_{j,k} \right] - y \right\|^2 \\ &\leq \left\| \psi_{i-1,k} - \alpha_k \sum_{j=1}^{i} w_{i,j}^k g_{j,k} - y \|^2 \right\| \qquad (4.4) \\ &\leq \|\psi_{i-1,k} - y\|^2 - 2\alpha_k \sum_{j=1}^{i} w_{i,j}^k g'_{j,k}(\psi_{i-1,k} - y) + \alpha_k^2 C^2 \left(\sum_{j=1}^{i} w_{i,j}^k \right)^2,\end{aligned}$$

where in the last inequality we use the following relation

$$\left\| \sum_{j=1}^{i} w_{i,j}^k g_{j,k} \right\|^2 \leq \left(\sum_{j=1}^{i} w_{i,j}^k \|g_{j,k}\| \right)^2$$

and the subgradient boundedness [cf. Assumption 4.1(b)].

For $i = 1$, from the relation $\psi_{0,k} = x_k$ [cf. Eq. (4.1)] and the subgradient inequality

$$g'_{1,k}(x_k - y) \geq f_1(x_0) - f_1(y),$$

we have

$$\|\psi_{1,k} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k w_{1,1}^k \big(f_1(x_0) - f_1(y)\big) + \alpha_k^2 C^2 \big(w_{1,1}^k\big)^2, \qquad \forall\, y \in X, \quad \forall\, k. \quad (4.5)$$

For $i = 2, \ldots, m$, we next estimate the term $\sum_{j=1}^{i} w_{i,j}^k g'_{j,k}(\psi_{i-1,k} - y)$ in the right hand side of Eq. (4.4). In particular, by using the subgradient inequality and the subgradient boundedness assumption, we obtain

$$\begin{aligned}\sum_{j=1}^{i} w_{i,j}^k g'_{j,k}(\psi_{i-1,k} - y) &\geq \sum_{j=1}^{i} w_{i,j}^k \big(g'_{j,k}(\psi_{j-1,k} - y) + g'_{j,k}(\psi_{i-1,k} - \psi_{j-1,k}) \big) \\ &\geq \sum_{j=1}^{i} w_{i,j}^k \big(f_j(\psi_{j-1,k}) - f_j(y) - C\|\psi_{i-1,k} - \psi_{j-1,k}\| \big) \\ &\geq \sum_{j=1}^{i} w_{i,j}^k \Big(\big(f_j(x_k) - f_j(y)\big) + \big(f_j(\psi_{j-1,k}) - f_j(x_k)\big) - C\|\psi_{i-1,k} - \psi_{j-1,k}\| \Big).\end{aligned}$$

By convexity of each f_j and by Assumption 4.1(b), it follows that

$$f_j(\psi_{j-1,k}) - f_j(x_k) \geq \tilde{g}'_{j,k}(\psi_{j-1,k} - x_k) \geq -C\|\psi_{j-1,k} - x_k\|,$$

where $\tilde{g}_{j,k}$ is a subgradient of f_j at x_k. Combining the preceding two relations, we see that

$$\sum_{j=1}^{i} w_{i,j}^k g'_{j,k}(\psi_{i-1,k} - y) \geq \sum_{j=1}^{i} w_{i,j}^k \Big(\big(f_j(x_k) - f_j(y)\big) - C\big(\|\psi_{i-1,k} - \psi_{j-1,k}\| + \|\psi_{j-1,k} - x_k\|\big)\Big).$$

Furthermore, from the subiterate definition [cf. Eq. (4.2)] and the subgradient boundedness assumption, we have for all p and s with $1 \leq s < p$,

$$\|\psi_{p,k} - \psi_{s,k}\| \leq \|\psi_{p,k} - \psi_{p-1,k}\| + \cdots + \|\psi_{s+1,k} - \psi_{s,k}\| \leq \alpha_k C \left(\sum_{l=1}^{p} w_{p,l}^k + \cdots + \sum_{l=1}^{s+1} w_{s+1,l}^k \right).$$

Using this and the relation $x_k = \psi_{0,k}$, it can be seen that for all $i = 2, \ldots, m$ and $j = 1, \ldots, i$,

$$\|\psi_{i-1,k} - \psi_{j-1,k}\| + \|\psi_{j-1,k} - x_k\| \leq \alpha_k C \left(\sum_{l=1}^{i-1} w_{i-1,l}^k + \cdots + w_{1,1} \right).$$

Therefore,

$$\sum_{j=1}^{i} w_{i,j}^k g'_{j,k}(\psi_{i-1,k} - y) \geq \sum_{j=1}^{i} w_{i,j}^k \left[\big(f_j(x_k) - f_j(y)\big) - \alpha_k C^2 \left(\sum_{l=1}^{i-1} w_{i-1,l}^k + \cdots + w_{1,1} \right) \right].$$

Substituting the preceding relation in Eq. (4.4), we obtain for all $y \in X$, $i = 2, \ldots, m$, and all k,

$$\begin{aligned} \|\psi_{i,k} - y\|^2 \leq \|\psi_{i-1,k} - y\|^2 &- 2\alpha_k \sum_{j=1}^{i} w_{i,j}^k \big(f_j(x_k) - f_j(y)\big) \\ &+ 2\alpha_k^2 C^2 \left(\sum_{j=1}^{i} w_{i,j}^k \right) \left(\sum_{l=1}^{i-1} w_{i-1,l}^k + \cdots + w_{1,1} \right) + \alpha_k^2 C^2 \left(\sum_{j=1}^{i} w_{i,j}^k \right)^2, \end{aligned}$$

or equivalently,

$$\begin{aligned} \|\psi_{i,k} - y\|^2 \leq \|\psi_{i-1,k} - y\|^2 &- 2\alpha_k \sum_{j=1}^{i} w_{i,j}^k \big(f_j(x_k) - f_j(y)\big) \\ &+ \alpha_k^2 C^2 \left[\left(\sum_{j=1}^{i} w_{i,j}^k + \cdots + w_{1,1} \right)^2 - \left(\sum_{l=1}^{i-1} w_{i-1,l}^k + \cdots + w_{1,1} \right)^2 \right]. \end{aligned}$$

By adding these inequalities for $i = 2, \ldots, m$ and then adding Eq. (4.5) for $i = 1$, and by using $\psi_{m,k} = x_{k+1}$ [cf. Eq. (4.3)], we have for all $y \in X$ and all k,

$$\|x_{k+1} - y\|^2 \leq \|x_k - y\| - 2\alpha_k \sum_{i=1}^{m} \sum_{j=1}^{i} w_{i,j}^k \big(f_j(x_k) - f_j(y)\big) + \alpha_k^2 C^2 \left(\sum_{j=1}^{m} w_{m,j}^k + \cdots + w_{1,1} \right)^2. \tag{4.6}$$

Finally, since

$$\sum_{i=j}^{m} w_{i,j}^k = 1, \qquad \forall\, j = 1, \ldots, m, \quad \forall\, k.$$

[cf. Assumption 4.1(b)], it follows that

$$\sum_{i=1}^{m}\sum_{j=1}^{i} w_{i,j}^k \big(f_j(x_k) - f_j(y)\big) = \sum_{j=1}^{m} \big(f_j(x_k) - f_j(y)\big) \sum_{i=j}^{m} w_{i,j}^k = \sum_{j=1}^{m} \big(f_j(x_k) - f_j(y)\big) = f(x_k) - f(y),$$

and

$$\sum_{j=1}^{m} w_{m,j}^k + \cdots + w_{1,1} = \sum_{i=1}^{m}\sum_{j=1}^{i} w_{i,j}^k = \sum_{j=1}^{m} \left(\sum_{i=j}^{m} w_{i,j}^k \right) = \sum_{j=1}^{m} 1 = m.$$

Using the preceding two relations in Eq. (4.6), we obtain for all $y \in X$ and all k,

$$\|x_{k+1} - y\|^2 \leq \|x_k - y\| - 2\alpha_k \big(f(x_k) - f(y)\big) + \alpha_k^2 m^2 C^2.$$

Q.E.D.

The relation established in Lemma 4.1 is the same at the relation given in Lemma 2.1 for the pure incremental subgradient method (cf. Section 2.2). Since all convergence and convergence rate results of Chapter 2 are based on Lemma 2.1, all these results apply to the incremental subgradient method with weights.

4.2. AN INCREMENTAL APPROXIMATE SUBGRADIENT METHOD

The incremental approximate subgradient method is similar to the incremental subgradient method of Section 2.1. The only difference is that subgradients are replaced by approximate subgradients. Thus, at a typical iteration, the method starts with

$$\psi_{0,k} = x_k, \tag{4.7}$$

performs m subiterations

$$\psi_{i,k} = \mathcal{P}_X[\psi_{i-1,k} - \alpha_k \overline{g}_{i,k}], \qquad i = 1, \ldots, m, \tag{4.8}$$

where the scalar α_k is a positive stepsize and the vector $\overline{g}_{i,k}$ is an $\epsilon_{i,k}$-subgradient of f_i at $\psi_{i-1,k}$ with $\epsilon_{i,k} \geq 0$. The last of these subiterations is the new iteration

$$x_{k+1} = \psi_{m,k}. \tag{4.9}$$

In our analysis, for an $\epsilon \geq 0$, we use the defining property of an ϵ-subgradient $\overline{g}$ of a convex function $h : \Re^n \mapsto \Re$ at a point x, which is

$$h(x) + \overline{g}'(z - x) \leq h(z) + \epsilon, \qquad \forall\ z \in \Re^n. \tag{4.10}$$

We denote by $\partial_\epsilon h(x)$ the ϵ-subdifferential (set of all ϵ-subgradients) of h at x. Regarding the method (4.7)–(4.9), we define

$$\epsilon_k = \epsilon_{1,k} + \cdots + \epsilon_{m,k}, \qquad \forall\ k, \tag{4.11}$$

$$\varepsilon = \limsup_{k \to \infty} \epsilon_k,$$

and we assume the following:

Assumption 4.2:

(a) $\varepsilon < \infty$.

(b) There exists a positive scalar C such that

$$\|\overline{g}\| \leq C, \qquad \forall\ \overline{g} \in \partial f_i(x_k) \cup \partial_{\epsilon_{i,k}} f_i(\psi_{i-1,k}), \quad \forall\ i = 1, \ldots, m, \quad \forall\ k.$$

When the sequence $\{\epsilon_{i,k}\}$ is bounded, i.e., there exists a scalar $\bar{\epsilon} \geq 0$ such that

$$\epsilon_{i,k} \leq \bar{\epsilon}, \qquad \forall\ i, k,$$

then Assumption 4.2(a) is satisfied. In this case, if the constraint set X is compact or the sequence $\{\psi_{i,k}\}$ is bounded, then Assumption 4.2(b) is also satisfied, since the set $\cup_{x \in Z} \partial_\epsilon f_i(x)$ is bounded for any bounded set Z and any positive scalar ϵ (see e.g., Bertsekas, Nedić, and Ozdaglar [BNO02], Exercise 4.11). Furthermore, Assumption 4.2(b) is also satisfied if each f_i is a polyhedral function, i.e., f_i is the pointwise maximum of a finite number of affine functions.

We next establish a relation that is a basis for the forthcoming convergence results.

Lemma 4.2: Let Assumption 4.2 hold and let $\{x_k\}$ be the sequence generated by the incremental approximate subgradient method. Then, we have

$$\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k \big(f(x_k) - f(y)\big) + \alpha_k^2 m^2 C^2 + 2\alpha_k \epsilon_k, \qquad \forall\ y \in X, \quad \forall\ k,$$

where ϵ_k is given by Eq. (4.11), and C is as in Assumption 4.2(b).

Proof: Using the nonexpansion property of the projection, the $\epsilon_{i,k}$-subgradient boundedness [cf. Assumption 4.2(b)], and the ϵ-subgradient inequality (4.10) for each component function f_i with $\epsilon = \epsilon_{i,k}$, we obtain for all $y \in X$, and all i and k,

$$\begin{aligned}
\|\psi_{i,k} - y\|^2 &= \big\|\mathcal{P}_X[\psi_{i-1,k} - \alpha_k \overline{g}_{i,k}] - y\big\|^2 \\
&\leq \|\psi_{i-1,k} - \alpha_k \overline{g}_{i,k} - y\|^2 \\
&\leq \|\psi_{i-1,k} - y\|^2 - 2\alpha_k \overline{g}_{i,k}'(\psi_{i-1,k} - y) + \alpha_k^2 C^2 \\
&\leq \|\psi_{i-1,k} - y\|^2 - 2\alpha_k \big(f_i(\psi_{i-1,k}) - f_i(y)\big) + 2\alpha_k \epsilon_{i,k} + \alpha_k^2 C^2.
\end{aligned}$$

By adding the above inequalities over $i = 1, \ldots, m$ and by using $\epsilon_k = \epsilon_{1,k} + \cdots + \epsilon_{m,k}$ [cf. Eq. (4.11)], we see that for all $y \in X$ and k,

$$\begin{aligned}\|x_{k+1} - y\|^2 &\leq \|x_k - y\|^2 - 2\alpha_k \sum_{i=1}^{m} \big(f_i(\psi_{i-1,k}) - f_i(y)\big) + 2\alpha_k \epsilon_k + \alpha_k^2 m C^2 \\ &= \|x_k - y\|^2 - 2\alpha_k \left(f(x_k) - f(y) + \sum_{i=1}^{m} \big(f_i(\psi_{i-1,k}) - f_i(x_k)\big) \right) \\ &\quad + 2\alpha_k \epsilon_k + \alpha_k^2 m C^2.\end{aligned}$$

By definition of the method [cf. Eqs. (4.7)–(4.9)] and Assumption 4.2(b), we have that $\|\psi_{i,k} - x_k\| \leq \alpha_k i C$ for all i and k. Using this relation, the subgradient inequality, and Assumption 4.2(b), we obtain for all i and k,

$$f_i(x_k) - f_i(\psi_{i-1,k}) \leq \|\tilde{g}_{i,k}\| \cdot \|\psi_{i-1,k} - x_k\| \leq C\|\psi_{i-1,k} - x_k\| \leq \alpha_k (i-1) C^2,$$

where $\tilde{g}_{i,k} \in \partial f_i(x_k)$. From this and the preceding relation, we see that for all $y \in X$ and k,

$$\begin{aligned}\|x_{k+1} - y\|^2 &\leq \|x_k - y\|^2 - 2\alpha_k \big(f(x_k) - f(y)\big) + \alpha_k^2 \left(2 \sum_{i=2}^{m} (i-1) C^2 + m C^2 \right) + 2\alpha_k \epsilon_k \\ &= \|x_k - y\|^2 - 2\alpha_k \big(f(x_k) - f(y)\big) + \alpha_k^2 m^2 C^2 + 2\alpha_k \epsilon_k.\end{aligned}$$

Q.E.D.

Lemma 4.2 guarantees that given the current iterate x_k and some other point $y \in X$ whose cost is lower than $f(x_k) - \epsilon_k$, the next iterate x_{k+1} will be closer to y than x_k, provided the stepsize α_k is sufficiently small [less than $2\big(f(x_k) - \epsilon_k - f(y)\big)/(mC)^2$]. This fact, with appropriate choices for y, will be used in the analysis of the method (4.7)–(4.9).

4.2.1 Constant and Diminishing Stepsize Rules

In this section, we give convergence results for the method (4.7)–(4.9) using either a constant or a diminishing stepsize. The analysis here is similar to that of the incremental subgradient method (cf. Sections 2.3 and 2.4).

For the method with a constant stepsize, we have the following result.

Proposition 4.1: Let Assumption 4.2 hold. Then, for the sequence $\{x_k\}$ generated by the incremental approximate subgradient method with the stepsize α_k fixed to some positive constant α, we have:

(a) If $f^* = -\infty$, then

$$\liminf_{k \to \infty} f(x_k) = -\infty.$$

(b) If $f^* > -\infty$, then

$$\liminf_{k\to\infty} f(x_k) \leq f^* + \frac{\alpha m^2 C^2}{2} + \varepsilon,$$

where ε and C are as in Assumption 4.2.

Proof: We prove (a) and (b) simultaneously. If the result does not hold, there must exist a $\varrho > 0$ such that

$$\liminf_{k\to\infty} f(x_k) - \frac{\alpha m^2 C^2}{2} - \varepsilon - 3\varrho > f^*.$$

Let $\hat{y} \in X$ be such that

$$\liminf_{k\to\infty} f(x_k) \geq f(\hat{y}) + \frac{\alpha m^2 C^2}{2} + \varepsilon + 3\varrho,$$

and let k_0 be large enough so that for all $k \geq k_0$, we have

$$f(x_k) \geq \liminf_{k\to\infty} f(x_k) - \varrho.$$

By combining the preceding two relations, we obtain

$$f(x_k) - f(\hat{y}) \geq \frac{\alpha m^2 C^2}{2} + \varepsilon + 2\varrho, \qquad \forall\ k \geq k_0.$$

Since $\varepsilon = \limsup_{k\to\infty} \epsilon_k$, we may assume without loss of generality that k_0 is large enough so that

$$\varepsilon + \varrho \geq \epsilon_k, \qquad \forall\ k \geq k_0,$$

implying that

$$f(x_k) - f(\hat{y}) \geq \frac{\alpha m^2 C^2}{2} + \epsilon_k + \varrho, \qquad \forall\ k \geq k_0.$$

Using Lemma 4.2, where $y = \hat{y}$ and $\alpha_k = \alpha$, together with the preceding relation, we see that

$$\|x_{k+1} - \hat{y}\|^2 \leq \|x_k - \hat{y}\|^2 - 2\alpha\varrho, \qquad \forall\ k \geq k_0.$$

Therefore,

$$\|x_{k+1} - \hat{y}\|^2 \leq \|x_{k-1} - \hat{y}\|^2 - 4\alpha\varrho \leq \cdots \leq \|x_{k_0} - \hat{y}\|^2 - 2(k+1-k_0)\alpha\varrho,$$

which cannot hold for k sufficiently large, a contradiction. **Q.E.D.**

We next give a result for the method that employs a diminishing stepsize.

Proposition 4.2: Let Assumption 4.2 hold, and let the stepsize α_k be such that

$$\lim_{k\to\infty} \alpha_k = 0, \qquad \sum_{k=0}^{\infty} \alpha_k = \infty.$$

Then, for the sequence $\{x_k\}$ generated by the incremental approximate subgradient method, we have:

(a) If $f^* = -\infty$, then

$$\liminf_{k\to\infty} f(x_k) = -\infty.$$

(b) If $f^* > -\infty$, then

$$\liminf_{k\to\infty} f(x_k) \leq f^* + \varepsilon.$$

Proof: Suppose to arrive at a contradiction that there exists an $\varrho > 0$ such that

$$\liminf_{k\to\infty} f(x_k) + \varepsilon + 3\varrho > f^*.$$

Let $\hat{y} \in X$ be such that

$$\liminf_{k\to\infty} f(x_k) \geq f(\hat{y}) + \varepsilon + 3\varrho,$$

and let k_0 be large enough so that

$$f(x_k) \geq \liminf_{k\to\infty} f(x_k) - \varrho, \qquad \forall\ k \geq k_0,$$

$$\varepsilon + \varrho \geq \epsilon_k, \qquad \forall\ k \geq k_0.$$

From the preceding three relations it follows that

$$f(x_k) - f(\hat{y}) \geq \epsilon_k + \varrho, \qquad \forall\ k \geq k_0.$$

Using Lemma 4.2, where $y = \hat{y}$, together with the preceding relation, we obtain

$$\|x_{k+1} - \hat{y}\|^2 \leq \|x_k - \hat{y}\|^2 - \alpha_k(2\varrho - \alpha_k m^2 C^2), \qquad \forall\ k \geq k_0.$$

Because $\alpha_k \to 0$, without loss of generality, we may assume that k_0 is large enough so that $\varrho \geq \alpha_k m^2 C^2$ for all $k \geq k_0$, implying that

$$\|x_{k+1} - \hat{y}\|^2 \leq \|x_k - \hat{y}\|^2 - \alpha_k \varrho \leq \|x_{k-1} - \hat{y}\|^2 - \varrho(\alpha_{k-1} + \alpha_k) \leq \cdots \leq \|x_{k_0} - \hat{y}\|^2 - \varrho \sum_{j=k_0}^{k} \alpha_j.$$

Since $\sum_{k=0}^{\infty} \alpha_k = \infty$, this relation cannot hold for k sufficiently large, a contradiction. **Q.E.D.**

Let us now consider the case where $\varepsilon = 0$. In this case, the results of Props. 4.1 and 4.2 coincide with those of Props. 2.1 and 2.4 for the incremental subgradient method (cf. Sections 2.3 and 2.4). In addition, we have the following two results.

Proposition 4.3: Let Assumption 4.2 hold with $\varepsilon = 0$, and let the optimal solution set X^* be nonempty and bounded. Assume further that the stepsize α_k is such that

$$\lim_{k\to\infty} \alpha_k = 0, \qquad \sum_{k=0}^{\infty} \alpha_k = \infty.$$

Then, for the sequence $\{x_k\}$ generated by the incremental approximate subgradient method, we have

$$\lim_{k\to\infty} dist(x_k, X^*) = 0, \qquad \lim_{k\to\infty} f(x_k) = f^*.$$

Proof: Use Lemma 4.2 and a line of analysis similar to that of Prop. 2.5 (cf. Section 2.4). **Q.E.D.**

Proposition 4.4: Let Assumption 4.2 hold with $\varepsilon = 0$, and let the optimal solution set X^* be nonempty. Assume further that the stepsize α_k is such that

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

Then, the sequence $\{x_k\}$ generated by the incremental approximate subgradient method converges to some optimal solution.

Proof: Use Lemma 4.2 and a line of analysis similar to that of Prop. 2.6 (cf. Section 2.4). **Q.E.D.**

4.2.2 Dynamic Stepsize Rules

Here, we consider the method using dynamic stepsize rules. We start with the dynamic stepsize rule for known f^*, where

$$\alpha_k = \gamma_k \frac{f(x_k) - f^*}{m^2 C^2}, \qquad 0 < \underline{\gamma} \le \gamma_k \le \overline{\gamma} < 2, \qquad \forall\ k. \tag{4.12}$$

For this stepsize, we have the following result.

Proposition 4.5: Let Assumption 4.2 hold. Then, for the sequence $\{x_k\}$ generated by the incremental approximate subgradient method with the dynamic stepsize rule (4.12), we have

$$\liminf_{k\to\infty} f(x_k) \le f^* + \frac{2\varepsilon}{2-\overline{\gamma}}.$$

Proof: Suppose to obtain a contradiction that there exists $\varrho > 0$ such that

$$f^* + \frac{2(\varrho+\varepsilon)}{2-\overline{\gamma}} < \liminf_{k\to\infty} f(x_k).$$

Let k_0 be large enough so that

$$\frac{2(\varrho+\epsilon_k)}{2-\overline{\gamma}} \le f(x_k) - f^*, \qquad \forall\ k \ge k_0, \tag{4.13}$$

and let a vector $\hat{y} \in X$ be such that

$$f(\hat{y}) - f^* \leq \frac{\varrho}{2}.$$

Then, we have

$$2\big(f(\hat{y}) - f^* + \epsilon_k\big) \leq 2(\varrho + \epsilon_k) - \varrho \leq (2 - \overline{\gamma})\big(f(x_k) - f^*\big) - \varrho, \qquad \forall\ k \geq k_0. \tag{4.14}$$

By Lemma 4.2, where $y = \hat{y}$ and α_k is given by Eq. (4.12), it follows that for all $k \geq k_0$,

$$\begin{aligned}\|x_{k+1} - \hat{y}\|^2 &\leq \|x_k - \hat{y}\|^2 - \alpha_k\Big(2\big(f(x_k) - f(\hat{y})\big) - \gamma_k\big(f(x_k) - f^*\big) - 2\epsilon_k\Big)\\ &= \|x_k - \hat{y}\|^2 - \alpha_k\Big(2\big(f(x_k) - f^*\big) - 2\big(f(\hat{y}) - f^*\big) - \gamma_k\big(f(x_k) - f^*\big) - 2\epsilon_k\Big).\end{aligned}$$

By using Eq. (4.14) in this relation and the fact $\gamma_k \leq \overline{\gamma}$, we see that

$$\|x_{k+1} - \hat{y}\|^2 \leq \|x_k - \hat{y}\|^2 - \alpha_k\Big((\overline{\gamma} - \gamma_k)\big(f(x_k) - f^*\big) + \varrho\Big) \leq \|x_k - \hat{y}\|^2 - \alpha_k\,\varrho, \qquad \forall\ k \geq k_0.$$

By the definition of the stepsize [cf. Eq. (4.12)] and Eq. (4.13), from the preceding inequality we obtain for all $k \geq k_0$,

$$\|x_{k+1} - \hat{y}\|^2 \leq \|x_k - \hat{y}\|^2 - \frac{2\gamma\varrho^2}{m^2C^2(2-\overline{\gamma})} \leq \cdots \leq \|x_{k_0} - \hat{y}\|^2 - \frac{(k+1-k_0)2\gamma\varrho^2}{m^2C^2(2-\overline{\gamma})},$$

which cannot hold for k sufficiently large, a contradiction. **Q.E.D.**

If $\varepsilon = 0$ (or equivalently $\epsilon_k \to 0$), then the result of Prop. 4.5 coincides with that of Prop. 2.7 for the incremental subgradient method (cf. Section 2.5). Furthermore, if the optimal solution set X^* is nonempty and ϵ_k tend to zero fast enough, then $\{x_k\}$ converges to some optimal solution. This result is shown in the next proposition. A similar result, for the (nonincremental) approximate subgradient method, was shown by Brännlund [Brä93] in Theorem 3.1, p. 41.

Proposition 4.6: Let Assumption 4.2(b) hold, and assume that for some scalar c with $c \in (0,1)$,

$$\epsilon_k \leq c\,\frac{2-\gamma_k}{2}\big(f(x_k) - f^*\big), \qquad \forall\ k.$$

Assume further that the optimal solution set X^* is nonempty. Then, the sequence $\{x_k\}$ generated by the incremental approximate subgradient method with the dynamic stepsize rule (4.12) converges to some optimal solution.

Proof: By Lemma 4.2, where $y = x^*$ and α_k is given by Eq. (4.12), it follows that

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \alpha_k(2-\gamma_k)\big(f(x_k) - f^*\big) + 2\alpha_k\epsilon_k, \qquad \forall\ x^* \in X^*, \quad \forall\ k.$$

By our assumption, we have that for $c \in (0,1)$,

$$2\epsilon_k \leq c\,(2-\gamma_k)\big(f(x_k)-f^*\big), \qquad \forall\ k,$$

implying that

$$\|x_{k+1}-x^*\|^2 \leq \|x_k-x^*\|^2 - \alpha_k(1-c)(2-\gamma_k)\big(f(x_k)-f^*\big), \qquad \forall\ x^*\in X^*, \quad \forall\ k.$$

Therefore, by using the definition of the stepsize α_k, we obtain

$$\|x_{k+1}-x^*\|^2 \leq \|x_k-x^*\|^2 - \underline{\gamma}(1-c)(2-\overline{\gamma})\frac{\big(f(x_k)-f^*\big)^2}{m^2C^2}, \qquad \forall\ x^*\in X^*, \quad \forall\ k.$$

Thus, for any $x^*\in X^*$, the sequence $\big\{\|x_k-x^*\|\big\}$ is nonincreasing, and therefore bounded. Furthermore, from the preceding relation it follows that

$$\|x_{k+1}-x^*\|^2 \leq \|x_0-x^*\|^2 - \frac{\underline{\gamma}(1-c)(2-\overline{\gamma})}{m^2C^2}\sum_{j=0}^{k}\big(f(x_j)-f^*\big)^2, \qquad \forall\ x^*\in X^*, \quad \forall\ k,$$

implying that $\sum_{j=0}^{\infty}\big(f(x_j)-f^*\big)^2$ is finite. Hence, $f(x_k)\to f^*$, and by continuity of f, we have $\overline{x}\in X^*$ for any limit point $\overline{x}$ of $\{x_k\}$. Since the sequence $\big\{\|x_k-x^*\|\big\}$ is nonincreasing, it converges to $\|\overline{x}-x^*\|$ for every $x^*\in X^*$. If there are two limit points $\tilde{x}$ and $\overline{x}$ of $\{x_k\}$, we must have $\tilde{x}\in X^*$, $\overline{x}\in X^*$, and $\|\tilde{x}-x^*\| = \|\overline{x}-x^*\|$ for all $x^*\in X^*$, which is possible only if $\tilde{x}=\overline{x}$. **Q.E.D.**

We next consider the dynamic stepsize rule with unknown f^*,

$$\alpha_k = \gamma_k\frac{f(x_k)-f_k^{\rm lev}}{m^2C^2}, \qquad 0<\underline{\gamma}\leq\gamma_k\leq\overline{\gamma}<2, \tag{4.15}$$

where the estimates $f_k^{\rm lev}$ are given by

$$f_k^{\rm lev} = \min_{0\leq j\leq k} f(x_j) - \delta_k, \tag{4.16}$$

while δ_k is updated using procedures similar to those of Section 2.6. We start with the adjustment procedure where δ_k is updated according to the following rule:

$$\delta_{k+1} = \begin{cases} \lambda\delta_k & \text{if } f(x_{k+1})\leq f_k^{\rm lev}, \\ \max\big\{\beta\delta_k,\delta\big\} & \text{if } f(x_{k+1})> f_k^{\rm lev}, \end{cases} \tag{4.17}$$

where δ_0, δ, β, and λ are fixed positive scalars with $\beta<1$ and $\lambda\geq 1$.

For the method using the stepsize (4.15)–(4.17), we have the following result.

Proposition 4.7: Let Assumption 4.2 hold. Then, for the sequence $\{x_k\}$ generated by the incremental approximate subgradient method and the dynamic stepsize rule (4.15)–(4.17), we have:

(a) If $f^* = -\infty$, then

$$\inf_{k\geq 0} f(x_k) = f^*.$$

(b) If $f^* > -\infty$, then

$$\inf_{k\geq 0} f(x_k) \leq f^* + \delta + \varepsilon.$$

Proof: To arrive at a contradiction, assume that

$$\inf_{k\geq 0} f(x_k) - \delta - \epsilon > f^*. \tag{4.18}$$

Each time the target level is attained [i.e., $f(x_k) \leq f_{k-1}^{\rm lev}$], the current best function value $\min_{0\leq j\leq k} f(x_j)$ decreases by at least δ [cf. Eqs. (4.16) and (4.17)], so in view of Eq. (4.18), the target value can be attained only a finite number times. From Eq. (4.17) it follows that after finitely many iterations, δ_k is decreased to the threshold value and remains at that value for all subsequent iterations, i.e., there is an index $\overline{k}$ such that

$$\delta_k = \delta, \qquad \forall\ k \geq \overline{k}. \tag{4.19}$$

In view of Eq. (4.18), there exists a $\overline{y} \in X$ such that

$$\inf_{k\geq 0} f(x_k) - \delta - \varepsilon > f(\overline{y}).$$

Without loss of generality, we may assume that $\overline{k}$ is large enough so that

$$\inf_{k\geq 0} f(x_k) - \delta - \epsilon_k \geq f(\overline{y}), \qquad \forall\ k \geq \overline{k}.$$

Thus, by Eqs. (4.16) and (4.19), we have

$$f_k^{\rm lev} - \epsilon_k = \min_{0\leq j\leq k} f(x_j) - \delta - \epsilon_k \geq \inf_{k\geq 0} f(x_k) - \delta - \epsilon_k \geq f(\overline{y}), \qquad \forall\ k \geq \overline{k}, \tag{4.20}$$

By Lemma 4.2 with $y = \overline{y}$ and α_k as in Eq. (4.15), it follows that

$$\begin{aligned}
\|x_{k+1} - \overline{y}\|^2 &\leq \|x_k - \overline{y}\|^2 - 2\alpha_k\big(f(x_k) - f(\overline{y})\big) + \alpha_k^2 m^2 C^2 + 2\alpha_k \epsilon_k \\
&= \|x_k - \overline{y}\|^2 - \alpha_k\Big(2\big(f(x_k) - f(\overline{y})\big) - \gamma_k\big(f(x_k) - f_k^{\rm lev}\big) - 2\epsilon_k\Big) \\
&= \|x_k - \overline{y}\|^2 - \alpha_k\Big((2-\gamma_k)\big(f(x_k) - f_k^{\rm lev}\big) + 2\big(f_k^{\rm lev} - f(\overline{y}) - \epsilon_k\big)\Big), \qquad \forall\ k \geq 0.
\end{aligned}$$

Using Eq. (4.20) in the preceding relation and the definition of α_k [cf. Eq. (4.15)], we obtain

$$\begin{aligned}
\|x_{k+1} - \overline{y}\|^2 &\leq \|x_k - \overline{y}\|^2 - \alpha_k(2-\gamma_k)\big(f(x_k) - f_k^{\rm lev}\big) \\
&\leq \|x_k - \overline{y}\|^2 - \gamma_k(2-\gamma_k)\left(\frac{f(x_k) - f_k^{\rm lev}}{mC}\right)^2 \\
&\leq \|x_k - \overline{y}\|^2 - \underline{\gamma}(2-\overline{\gamma})\frac{\delta^2}{m^2C^2}, \qquad \forall\ k \geq \overline{k},
\end{aligned}$$

where the last inequality follows from the relations $\gamma_k \in [\underline{\gamma}, \overline{\gamma}]$ and $f(x_k) - f_k^{\rm lev} \geq \delta$ for all k. Finally, by adding the above inequalities over k, we see that

$$\|x_{k+1} - \overline{y}\|^2 \leq \|x_{\overline{k}} - \overline{y}\|^2 - (k+1-\overline{k})\underline{\gamma}(2-\overline{\gamma})\frac{\delta^2}{m^2 C^2}, \qquad \forall\ k \geq \overline{k},$$

which cannot hold for sufficiently large k, a contradiction. **Q.E.D.**

We now describe the method that employs the stepsize (4.15)-(4.16), where the parameters δ_k are adjusted according to the path-based procedure. The method is given in the following algorithm.

Path-Based Incremental Approximate Subgradient Algorithm

Step 0 (*Initialization*) Select x_0, $\delta_0 > 0$, and $b > 0$. Set $\sigma_0 = 0$, $f_{-1}^{\rm rec} = \infty$. Set $k = 0$, $l = 0$, and $k(l) = 0$ [$k(l)$ will denote the iteration number when the l-th update of $f_k^{\rm lev}$ occurs].

Step 1 (*Function evaluation*) Compute $f(x_k)$. If $f(x_k) < f_{k-1}^{\rm rec}$, then set $f_k^{\rm rec} = f(x_k)$. Otherwise set $f_k^{\rm rec} = f_{k-1}^{\rm rec}$ [so that $f_k^{\rm rec}$ keeps the record of the smallest value attained by the iterates that are generated so far, i.e., $f_k^{\rm rec} = \min_{0 \leq j \leq k} f(x_j)$].

Step 2 (*Sufficient descent*) If $f(x_k) \leq f_{k(l)}^{\rm rec} - \frac{\delta_l}{2}$, then set $k(l+1) = k$, $\sigma_k = 0$, $\delta_{l+1} = \delta_l$, increase l by 1, and go to Step 4.

Step 3 (*Oscillation detection*) If $\sigma_k > b$, then set $k(l+1) = k$, $\sigma_k = 0$, $\delta_{l+1} = \frac{\delta_l}{2}$, and increase l by 1.

Step 4 (*Iterate update*) Set $f_k^{\rm lev} = f_{k(l)}^{\rm rec} - \delta_l$. Select $\gamma_k \in [\underline{\gamma}, \overline{\gamma}]$ and compute x_{k+1} via Eqs. (4.7)–(4.9) with the stepsize (4.15).

Step 5 (*Path length update*) Set $\sigma_{k+1} = \sigma_k + \alpha_k mC$, increase k by 1, and go to Step 1.

The interpretation of the parameters b and σ_k is the same as in the path-based incremental algorithm of Section 2.6.

For the preceding algorithm, we have the following convergence result.

Proposition 4.8: Let Assumption 4.2 hold. Then, for the path-based incremental approximate subgradient algorithm, we have:

(a) If $f^* = -\infty$, then

$$\inf_{k \geq 0} f(x_k) = f^*.$$

(b) If $f^* > -\infty$, then

$$\inf_{k \geq 0} f(x_k) \leq f^* + \varepsilon.$$

Proof: Use Lemma 4.2, and a line of analysis similar to that of Prop. 2.12 of Section 2.6. **Q.E.D.**

PART II:

Variable Metric Subgradient Methods

5

A Variable Metric Subgradient Method

We here propose a new subgradient method that uses a variable metric. This method combines the principles of the variable metric approach with those of subgradient methods. The main idea is the same as in variable metric methods for differentiable problems, namely, to transform the original space coordinates in order to get better convergence rates. This is important, in particular, for problems where subgradients are almost perpendicular to the directions pointing toward the set of minima, in which case the ordinary subgradient method is slow. Changing the stepsize cannot improve the method's progress, since poor performance is due to bad subgradient directions. In this case, it is better to modify subgradient directions, which can be done by transforming the space coordinates.

Our method is applicable to unconstrained convex problems

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && x \in \Re^n, \end{aligned}$$

where the function $f : \Re^n \mapsto \Re$ is convex but not necessarily differentiable. In Section 5.1, we introduce the method, and in Section 5.2, we establish its basic properties. In Sections 5.3 and 5.4, we analyze its convergence for the three stepsize rules: constant, diminishing, and dynamic.

5.1. THE METHOD

At a typical iteration of a variable metric subgradient method, we have

$$x_{k+1} = x_k - \alpha_k B_k B_k' g_k, \tag{5.1}$$

where α_k is a positive stepsize, B_k is an $n \times n$ invertible matrix representing a linear transformation of the space coordinates, and g_k is a subgradient of f at x_k. In what follows, we use the terms "matrix" and "linear transformation" interchangeably.

An important property of subgradients under a linear transformation, thanks to which variable metric subgradient methods work, is the following: A linear transformation maps a subgradient of f in the original space into a subgradient of some function in the transformed space. In particular, let $y = B^{-1}x$ for $x \in \Re^n$, so that y-space is the transformed space. Let $\overline{x} \in \Re^n$ and $\overline{y} = B^{-1}\overline{x}$, and let g be a subgradient of f at $\overline{x}$. In the y-space, consider the function F given by

$$F(y) = f(By),$$

which is convex since f is convex (see [Roc70], Theorem 5.7). Furthermore, for the vector $B'g$ and any $y \in \Re^n$, we have by using $\overline{y} = B^{-1}\overline{x}$,

$$F(\overline{y}) + (B'g)'(y - \overline{y}) = f(\overline{x}) + g'(By - \overline{x}).$$

Because g is a subgradient of f at $\overline{x}$, it follows that for any $y \in \Re^n$,

$$F(\overline{y}) + (B'g)'(y - \overline{y}) \leq f(By) = F(y),$$

thus showing that $B'g$ is a subgradient of F at $\overline{y}$.

Using this property of subgradients and appropriate linear transformations, we can ensure that the distance between the iterates x_k and the set of minima is bounded. In particular, this can be guaranteed by allowing a *limited total amount of space transformation* in some sense, which we describe in the next section.

5.2. ASSUMPTIONS AND SOME BASIC RELATIONS

We first establish a basic relation for the iterates generated by the method (5.1). This relation is given in the following lemma and is used repeatedly in our convergence analysis.

Lemma 5.1: Let $\{x_k\}$ be the sequence generated by the variable metric method. Then, for any $y \in \Re^n$ and all k, we have

$$\left\| B_{k+1}^{-1}(x_{k+1} - y) \right\|^2 \leq \|B_{k+1}^{-1} B_k\|^2 \Big(\left\| B_k^{-1}(x_k - y) \right\|^2 - 2\alpha_k \big(f(x_k) - f(y) \big) + \alpha_k^2 \|B_k' g_k\|^2 \Big).$$

Proof: Let $y \in \Re^n$ be arbitrary. By the iterate definition [cf. Eq. (5.1)], we have for all k,

$$B_{k+1}^{-1}(x_{k+1} - y) = B_{k+1}^{-1}(x_k - \alpha_k B_k B_k' g_k - y) = B_{k+1}^{-1} B_k \big(B_k^{-1}(x_k - y) - \alpha_k B_k' g_k\big).$$

From this relation, we obtain for all k,

$$\begin{aligned}\big\|B_{k+1}^{-1}(x_{k+1} - y)\big\|^2 &\leq \|B_{k+1}^{-1} B_k\|^2 \big\|B_k^{-1}(x_k - y) - \alpha_k B_k' g_k\big\|^2 \\ &= \|B_{k+1}^{-1} B_k\|^2 \Big(\big\|B_k^{-1}(x_k - y)\big\|^2 - 2\alpha_k g_k'(x_k - y) + \alpha_k^2 \|B_k' g_k\|^2\Big).\end{aligned}$$

Since g_k is a subgradient of f at x_k, we have

$$g_k'(x_k - y) \geq f(x_k) - f(y),$$

implying that for all k,

$$\big\|B_{k+1}^{-1}(x_{k+1} - y)\big\|^2 \leq \|B_{k+1}^{-1} B_k\|^2 \Big(\big\|B_k^{-1}(x_k - y)\big\|^2 - 2\alpha_k \big(f(x_k) - f(y)\big) + \alpha_k^2 \|B_k' g_k\|^2\Big).$$

Q.E.D.

When the function f has a nonempty set of minima over $\Re^n$, the result of Lemma 5.1 can be strengthened, as seen in the following lemma.

Lemma 5.2: Let $\{x_k\}$ be the sequence generated by the variable metric method, and assume that the optimal solution set X^* is nonempty. Then, we have for all k,

$$\begin{aligned}\min_{x^* \in X^*} \big\|B_{k+1}^{-1}(x_{k+1} - x^*)\big\|^2 \leq \|B_{k+1}^{-1} B_k\|^2 \Bigg(&\min_{x^* \in X^*} \big\|B_k^{-1}(x_k - x^*)\big\|^2 \\ &- 2\alpha_k \big(f(x_k) - f^*\big) + \alpha_k^2 \|B_k' g_k\|^2\Bigg).\end{aligned}$$

Proof: By Lemma 5.1, where $y = x^*$, we obtain for any $x^* \in X^*$ and all k,

$$\big\|B_{k+1}^{-1}(x_{k+1} - x^*)\big\|^2 \leq \|B_{k+1}^{-1} B_k\|^2 \Big(\big\|B_k^{-1}(x_k - x^*)\big\|^2 - 2\alpha_k \big(f(x_k) - f(x^*)\big) + \alpha_k^2 \|B_k' g_k\|^2\Big),$$

from which the result follows by taking the minimum over all x^* in X^*. **Q.E.D.**

In our convergence analysis, we use the following assumption.

Assumption 5.1:

(a) There exists a positive scalar $\bar{C}$ such that

$$\|B_k' g_k\| \leq \bar{C}, \qquad \forall\ k.$$

(b) The linear transformations B_k are such that

$$\|B_{k+1}^{-1}B_k\| \geq 1, \qquad \forall\ k,$$

$$\prod_{k=0}^{\infty} \|B_{k+1}^{-1}B_k\|^2 < \infty.$$

Assumption 5.1(a) says that the subgradients in the transformed space are bounded. Assumption 5.1(b) allows us to guarantee that the distance between the iterates x_k and the set of minima is bounded in some appropriate norm. In particular, the second condition is the one that poses a limit on the total amount of space transformation.

Let us now discuss the cases where Assumption 5.1 is satisfied. When f a is polyhedral function, Assumption 5.1 is satisfied, for example, when B_k is the identity matrix for all sufficiently large k. The assumption can also be satisfied with diagonal matrices B_k. However, it is hard to give a general rule for selecting the diagonal entries of B_k. The selection of the diagonal entries may be based on some additional information about the function f.

Assumption 5.1 can also be satisfied with the matrices B_k generated as in Shor's space dilation method (cf. Shor [Sho70a], Shor and Zhurbenko [ShZ71], see also Shor [Sho85] and [Sho98]), where

$$B_k = B_{k-1}R_{\rho_k}(\xi_k), \qquad \forall\ k \geq 1,$$

B_0 is some invertible initial matrix, for example, $B_0 = I$. The linear transformation $R_{\rho_k}(\xi_k)$ is given by

$$R_{\rho_k}(\xi_k) = I + (\rho_k - 1)\xi_k\xi_k', \qquad \forall\ k \geq 1,$$

for a positive scalar ρ_k and the vector ξ_k defined by

$$\xi_k = \frac{B_{k-1}'d_k}{\|B_{k-1}'d_k\|},$$

where $d_k = g_k$ or $d_k = g_k - g_{k-1}$. It can be shown that (cf. Shor [Sho85] and [Sho98]) for a positive $\rho \in \Re$ and $\xi \in \Re^n$ with $\|\xi\| = 1$,

$$\big(R_\rho(\xi)\big)^{-1} = R_{\frac{1}{\rho}}(\xi), \qquad \big\|R_\rho(\xi)\big\| = \max\{1, \rho\}. \tag{5.2}$$

Thus, if the parameters ρ_k are chosen such that

$$\prod_{k=1}^{\infty} \max\{1, \rho_k\} < \infty,$$

then for all $k \geq 1$

$$\|B_k'g_k\| = \|R_{\rho_k}'(\xi_k)\cdots R_{\rho_1}'(\xi_1)g_k\| \leq \left(\prod_{i=1}^{k} \max\{1, \rho_i\}\right)\|g_k\|,$$

implying that Assumption 5.1(a) is satisfied for a polyhedral function f.

Furthermore, by using the relations in (5.2), we obtain

$$\|B_{k+1}^{-1}B_k\| = \max\left\{1, \frac{1}{\rho_{k+1}}\right\} \geq 1,$$

showing that, for B_k generated as in Shor's space dilation method, the first condition of Assumption 5.1(b) is always satisfied. In order to have $\prod_{k=0}^{\infty}\|B_{k+1}^{-1}B_k\|^2 < \infty$, it suffices to choose the parameters ρ_k such that

$$\prod_{k=0}^{\infty}\max\left\{1, \frac{1}{\rho_{k+1}^2}\right\} < \infty.$$

5.3. CONSTANT AND DIMINISHING STEPSIZE RULES

Here, we give convergence results for the method using the constant and diminishing stepsize rules. Our first result is for the constant stepsize rule, where the stepsize α_k is fixed to some positive scalar α. In this case, as seen in our next proposition, the function values $f(x_k)$ along a subsequence of $\{x_k\}$ converge to f^* within an error proportional to α.

Proposition 5.1: Let Assumption 5.1 hold, and let the sequence $\{x_k\}$ be generated by the variable metric method with the stepsize α_k fixed to some positive scalar α. We have:

(a) If $f^* = -\infty$, then

$$\liminf_{k\to\infty} f(x_k) = -\infty.$$

(b) If $f^* > -\infty$, then

$$\liminf_{k\to\infty} f(x_k) \leq f^* + \frac{\alpha\bar{C}^2}{2},$$

where $\bar{C}$ is as in Assumption 5.1.

Proof: We prove (a) and (b) simultaneously. If the result does not hold, there must exist an $\epsilon > 0$ such that

$$\liminf_{k\to\infty} f(x_k) - \epsilon - \frac{\alpha\bar{C}^2}{2} > f^*.$$

Let $\hat{y} \in \Re^n$ be such that

$$\liminf_{k\to\infty} f(x_k) - \epsilon - \frac{\alpha\bar{C}^2}{2} > f(\hat{y}),$$

and let k_0 be large enough so that for all $k \geq k_0$, we have

$$f(x_k) - \epsilon - \frac{\alpha\bar{C}^2}{2} \geq f(\hat{y}).$$

Using Lemma 5.1, where $y = \hat{y}$ and $\alpha_k = \alpha$, together with the preceding relation, we see that

$$\left\|B_{k+1}^{-1}(x_{k+1} - \hat{y})\right\|^2 \leq \|B_{k+1}^{-1}B_k\|^2\left(\left\|B_k^{-1}(x_k - \hat{y})\right\|^2 - 2\alpha\epsilon - \alpha^2\bar{C}^2 + \alpha^2\|B_k'g_k\|^2\right), \qquad \forall\ k \geq k_0.$$

By Assumption 5.1(a), we have $\|B_k'g_k\| \leq \bar{C}$ for all k, implying that

$$\left\|B_{k+1}^{-1}(x_{k+1} - \hat{y})\right\|^2 \leq \|B_{k+1}^{-1}B_k\|^2\left(\left\|B_k^{-1}(x_k - \hat{y})\right\|^2 - 2\alpha\epsilon\right), \qquad \forall\ k \geq k_0.$$

Since, by Assumption 5.1(b), we have $\|B_{k+1}^{-1}B_k\| \geq 1$ for all k, it follows that for all $k \geq k_0$,

$$\begin{aligned}\left\|B_{k+1}^{-1}(x_{k+1} - \hat{y})\right\|^2 &\leq \|B_{k+1}^{-1}B_k\|^2\left\|B_k^{-1}(x_k - \hat{y})\right\|^2 - 2\alpha\epsilon \\ &\leq \left(\prod_{i=k_0}^{k}\|B_{i+1}^{-1}B_i\|^2\right)\left\|B_{k_0}^{-1}(x_{k_0} - \hat{y})\right\|^2 - 2(k+1-k_0)\alpha\epsilon.\end{aligned}$$

Furthermore, by the same assumption, $\prod_{i=k_0}^{k}\|B_{i+1}^{-1}B_i\|^2 \leq \prod_{i=0}^{\infty}\|B_{i+1}^{-1}B_i\|^2 < \infty$, implying that the preceding relation cannot hold for k sufficiently large, a contradiction. **Q.E.D.**

When the optimal solution set X^* is nonempty, we can estimate the number of iterations required to achieve the optimal function value f^* within some error not greater than $(\alpha\bar{C}^2 + \epsilon)/2$, where ϵ is an arbitrarily small positive scalar. In particular, we have the following result.

Proposition 5.2: Let Assumption 5.1 hold, and assume that the optimal solution set X^* is nonempty. Let the sequence $\{x_k\}$ be generated by the variable metric method with the stepsize α_k fixed to some positive scalar α. Then, for a positive scalar ϵ and the smallest positive integer K such that

$$\alpha\epsilon K \leq \left(\prod_{i=0}^{K-1}\|B_{i+1}^{-1}B_i\|^2\right)\min_{x^* \in X^*}\left\|B_0^{-1}(x_0 - x^*)\right\|^2,$$

we have

$$\min_{0 \leq k \leq K} f(x_k) \leq f^* + \frac{\alpha\bar{C}^2 + \epsilon}{2}.$$

Proof: To arrive at a contradiction, assume that

$$f(x_k) > f^* + \frac{\alpha\bar{C}^2 + \epsilon}{2}, \qquad \forall\ k = 0, 1, \ldots, K.$$

By using this relation and Lemma 5.2, where α_k is replaced by α, we obtain for $k = 0, 1, \ldots, K$,

$$\min_{x^* \in X^*}\left\|B_{k+1}^{-1}(x_{k+1} - x^*)\right\|^2 \leq \|B_{k+1}^{-1}B_k\|^2\left(\min_{x^* \in X^*}\left\|B_k^{-1}(x_k - x^*)\right\|^2 - \alpha\epsilon - \alpha^2\bar{C}^2 + \alpha^2\|B_k'g_k\|^2\right)$$

By Assumption 5.1, we have $\|B_k' g_k\| \leq \bar{C}$ and $\|B_{k+1}^{-1} B_k\| \geq 1$ for all k, implying that for $k = 0, 1, \ldots, K$,

$$\begin{aligned}\min_{x^* \in X^*} \left\| B_{k+1}^{-1}(x_{k+1} - x^*) \right\|^2 &\leq \|B_{k+1}^{-1} B_k\|^2 \left(\min_{x^* \in X^*} \left\| B_k^{-1}(x_k - x^*) \right\|^2 - \alpha\epsilon \right) \\ &\leq \|B_{k+1}^{-1} B_k\|^2 \min_{x^* \in X^*} \left\| B_k^{-1}(x_k - x^*) \right\|^2 - \alpha\epsilon.\end{aligned}$$

Hence, for $k = 0, 1, \ldots, K$,

$$\min_{x^* \in X^*} \left\| B_{k+1}^{-1}(x_{k+1} - x^*) \right\|^2 \leq \left(\prod_{i=0}^{k} \|B_{i+1}^{-1} B_i\|^2 \right) \min_{x^* \in X^*} \left\| B_0^{-1}(x_0 - x^*) \right\|^2 - (k+1)\alpha\epsilon.$$

In particular, we have that for $k = K - 2$,

$$\min_{x^* \in X^*} \left\| B_{K-1}^{-1}(x_{K-1} - x^*) \right\|^2 \leq \left(\prod_{i=0}^{K-2} \|B_{i+1}^{-1} B_i\|^2 \right) \min_{x^* \in X^*} \left\| B_0^{-1}(x_0 - x^*) \right\|^2 - (K-1)\alpha\epsilon,$$

implying that

$$(K-1)\alpha\epsilon \leq \left(\prod_{i=0}^{K-2} \|B_{i+1}^{-1} B_i\|^2 \right) \min_{x^* \in X^*} \left\| B_0^{-1}(x_0 - x^*) \right\|^2,$$

contradicting the definition of K. **Q.E.D.**

We now consider the method that uses a diminishing stepsize. In this case, the function values $f(x_k)$, along a subsequence of $\{x_k\}$, converge to the optimal function value f^*, as seen in our next proposition.

Proposition 5.3: Let Assumption 5.1 hold, and let the stepsize α_k be such that

$$\lim_{k \to \infty} \alpha_k = 0, \qquad \sum_{k=0}^{\infty} \alpha_k = \infty.$$

Then, for the sequence $\{x_k\}$ generated by the variable metric method, we have

$$\liminf_{k \to \infty} f(x_k) = f^*.$$

Proof: Suppose to arrive at a contradiction that there exists an $\epsilon > 0$ such that

$$\liminf_{k \to \infty} f(x_k) - \epsilon > f^*.$$

Let $\hat{y} \in \Re^n$ be such that

$$\liminf_{k \to \infty} f(x_k) - \epsilon > f(\hat{y}),$$

and let k_0 be large enough so that

$$f(x_k) - \epsilon \geq f(\hat{y}), \qquad \forall\ k \geq k_0.$$

Using Lemma 5.1, where $y = \hat{y}$, together with the preceding relation, we obtain

$$\left\|B_{k+1}^{-1}(x_{k+1} - \hat{y})\right\|^2 \leq \|B_{k+1}^{-1}B_k\|^2\left(\left\|B_k^{-1}(x_k - \hat{y})\right\|^2 - \alpha_k\left(2\epsilon\ - \alpha_k\|B_k'g_k\|^2\right)\right), \qquad \forall\ k \geq k_0.$$

Because $\alpha_k \to 0$ and $\|B_k'g_k\| \leq \bar{C}$ for all k [cf. Assumption 5.1(a)], without loss of generality, we may assume that k_0 is large enough so that $\epsilon \geq \alpha_k\bar{C}^2$ for all $k \geq k_0$, implying that

$$\left\|B_{k+1}^{-1}(x_{k+1} - \hat{y})\right\|^2 \leq \|B_{k+1}^{-1}B_k\|^2\left(\left\|B_k^{-1}(x_k - \hat{y})\right\|^2 - \alpha_k\epsilon\ \right), \qquad \forall\ k \geq k_0.$$

By Assumption 5.1(b), we have $\|B_{k+1}^{-1}B_k\| \geq 1$ for all k, and therefore

$$\begin{aligned}\left\|B_{k+1}^{-1}(x_{k+1} - \hat{y})\right\|^2 &\leq \|B_{k+1}^{-1}B_k\|^2\left\|B_k^{-1}(x_k - \hat{y})\right\|^2 - \alpha_k\epsilon \\ &\leq \left(\prod_{i=k_0}^{k}\|B_{i+1}^{-1}B_i\|^2\right)\left\|B_{k_0}^{-1}(x_{k_0} - \hat{y})\right\|^2 - \epsilon\sum_{j=k_0}^{k}\alpha_j, \qquad \forall\ k \geq k_0.\end{aligned}$$

However, this relation cannot hold for sufficiently large k, since $\sum_{j=0}^{\infty}\alpha_j = \infty$ and since by Assumption 5.1, we have

$$\prod_{i=k_0}^{k}\|B_{i+1}^{-1}B_i\|^2 \leq \prod_{i=0}^{\infty}\|B_{i+1}^{-1}B_i\|^2 < \infty,$$

a contradiction. **Q.E.D.**

5.4. DYNAMIC STEPSIZE RULES

Here, we discuss the method using a dynamic stepsize rule. We first consider the dynamic stepsize rule for known f^*, where

$$\alpha_k = \gamma_k\frac{f(x_k) - f^*}{\|B_k'g_k\|^2}, \qquad 0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2, \qquad \forall\ k. \tag{5.3}$$

For this stepsize, we have the following convergence result.

Proposition 5.4: Let Assumption 5.1 hold, and let the sequence $\{x_k\}$ be generated by the variable metric method with the dynamic stepsize (5.3). Then, we have

$$\liminf_{k\to\infty} f(x_k) = f^*.$$

Proof: To obtain a contradiction, suppose that there exists scalar $\epsilon > 0$ such that

$$\liminf_{k\to\infty} f(x_k) > f^* + \frac{2\epsilon}{2-\overline{\gamma}},$$

implying that for some large enough k_0,

$$f(x_k) \geq f^* + \frac{2\epsilon}{2-\overline{\gamma}}, \qquad \forall\ k \geq k_0. \tag{5.4}$$

Let a vector $\hat{y} \in \Re^n$ be such that

$$f^* + \frac{\epsilon}{2} \geq f(\hat{y}).$$

Using Lemma 5.1, where $y = \hat{y}$ and α_k is as in Eq. (5.3), we obtain for all $k \geq k_0$,

$$\left\|B_{k+1}^{-1}(x_{k+1}-\hat{y})\right\|^2 \leq \|B_{k+1}^{-1}B_k\|^2\left[\left\|B_k^{-1}(x_k-\hat{y})\right\|^2 - \alpha_k\Big(2\big(f(x_k)-f(\hat{y})\big) - \gamma_k\big(f(x_k)-f^*\big)\Big)\right].$$

From Eq. (5.4) and the fact $f^* + \epsilon/2 \geq f(\hat{y})$, we see that for all $k \geq k_0$,

$$2\big(f(x_k)-f(\hat{y})\big) - \gamma_k\big(f(x_k)-f^*\big) = (2-\gamma_k)\big(f(x_k)-f^*\big) + 2\big(f^*-f(\hat{y})\big) \geq (2-\gamma_k)\frac{2\epsilon}{2-\overline{\gamma}} - \epsilon,$$

and since $\gamma_k \leq \overline{\gamma}$ for all k, it follows that

$$2\big(f(x_k)-f(\hat{y})\big) - \gamma_k\big(f(x_k)-f^*\big) \geq 2\epsilon - \epsilon = \epsilon, \qquad \forall\ k \geq k_0.$$

Therefore, for all $k \geq k_0$,

$$\left\|B_{k+1}^{-1}(x_{k+1}-\hat{y})\right\|^2 \leq \|B_{k+1}^{-1}B_k\|^2\Big(\left\|B_k^{-1}(x_k-\hat{y})\right\|^2 - \alpha_k\epsilon\Big).$$

By using the definition of α_k [cf. Eq. (5.3)], the relation (5.4), and the boundedness of $\|B_k' g_k\|$ [cf. Assumption 5.1(a)], we obtain

$$\alpha_k = \gamma_k\frac{f(x_k)-f^*}{\|B_k' g_k\|^2} \geq \frac{2\epsilon\underline{\gamma}}{(2-\overline{\gamma})\bar{C}^2}, \qquad \forall\ k \geq k_0.$$

Since $\|B_{k+1}^{-1}B_k\| \geq 1$ for all k by Assumption 5.1(b), the preceding two relations imply that for all $k \geq k_0$,

$$\begin{aligned}\left\|B_{k+1}^{-1}(x_{k+1}-\hat{y})\right\|^2 &\leq \|B_{k+1}^{-1}B_k\|^2\left\|B_k^{-1}(x_k-\hat{y})\right\|^2 - \frac{2\epsilon^2\underline{\gamma}}{(2-\overline{\gamma})\bar{C}^2}\\ &\leq \left(\prod_{i=k_0}^{k}\|B_{i+1}^{-1}B_i\|^2\right)\left\|B_{k_0}^{-1}(x_{k_0}-\hat{y})\right\|^2 - (k+1-k_0)\frac{2\epsilon^2\underline{\gamma}}{(2-\overline{\gamma})\bar{C}^2}.\end{aligned}$$

But this relation cannot hold for k sufficiently large, since by Assumption 5.1(b), we have

$$\prod_{i=k_0}^{k} \|B_{i+1}^{-1}B_i\|^2 \leq \prod_{i=0}^{\infty} \|B_{i+1}^{-1}B_i\|^2 < \infty,$$

a contradiction. **Q.E.D.**

Assuming that the optimal solution set X^* is nonempty, we can estimate the number of iterations required to achieve the optimal function value f^* within a given error, as seen in the following proposition.

Proposition 5.5: Let Assumption 5.1 hold, and assume that the optimal solution set X^* is nonempty. Let $\{x_k\}$ be the sequence generated by the variable metric method with the stepsize (5.3). Then, for a positive scalar ϵ and the smallest positive integer K such that

$$K\frac{\epsilon^2\underline{\gamma}(2-\overline{\gamma})}{\bar{C}^2} \leq \left(\prod_{i=0}^{K-1} \|B_{i+1}^{-1}B_i\|^2\right) \min_{x^*\in X^*} \left\|B_0^{-1}(x_0-x^*)\right\|^2,$$

we have

$$\min_{0\leq k\leq K} f(x_k) \leq f^* + \epsilon.$$

Proof: By using Lemma 5.2, with α_k given by Eq. (5.3), we obtain for all k,

$$\min_{x^*\in X^*} \left\|B_{k+1}^{-1}(x_{k+1}-x^*)\right\|^2 \leq \|B_{k+1}^{-1}B_k\|^2 \left(\min_{x^*\in X^*} \left\|B_k^{-1}(x_k-x^*)\right\|^2 - \gamma_k(2-\gamma_k)\frac{\left(f(x_k)-f^*\right)^2}{\|B_k'g_k\|^2}\right).$$

By the definition of α_k, we have $\gamma_k \in [\underline{\gamma},\overline{\gamma}]$, while by Assumption 5.1, we have $\|B_k'g_k\| \leq \bar{C}$ and $\|B_{k+1}^{-1}B_k\| \geq 1$, implying that for all k,

$$\min_{x^*\in X^*} \left\|B_{k+1}^{-1}(x_{k+1}-x^*)\right\|^2 \leq \|B_{k+1}^{-1}B_k\|^2 \min_{x^*\in X^*} \left\|B_k^{-1}(x_k-x^*)\right\|^2 - \underline{\gamma}(2-\overline{\gamma})\frac{\left(f(x_k)-f^*\right)^2}{\bar{C}^2}.$$

Assume now, to arrive at a contradiction, that

$$f(x_k) > f^* + \epsilon, \qquad \forall\ k = 0,1,\ldots,K.$$

Combining this with the preceding relation, we obtain for all $k = 0,1,\ldots,K$,

$$\begin{aligned}\min_{x^*\in X^*} \left\|B_{k+1}^{-1}(x_{k+1}-x^*)\right\|^2 &\leq \|B_{k+1}^{-1}B_k\|^2 \min_{x^*\in X^*} \left\|B_k^{-1}(x_k-x^*)\right\|^2 - \underline{\gamma}(2-\overline{\gamma})\frac{\epsilon^2}{\bar{C}^2} \\ &\leq \left(\prod_{i=0}^{k} \|B_{i+1}^{-1}B_i\|^2\right) \min_{x^*\in X^*} \left\|B_0^{-1}(x_0-x^*)\right\|^2 - (k+1)\underline{\gamma}(2-\overline{\gamma})\frac{\epsilon^2}{\bar{C}^2}.\end{aligned}$$

Thus, in particular for $k = K-2$, we have

$$\min_{x^* \in X^*} \left\| B_{K-1}^{-1}(x_{K+1} - x^*) \right\|^2 \leq \left(\prod_{i=0}^{K-2} \|B_{i+1}^{-1} B_i\|^2 \right) \min_{x^* \in X^*} \left\| B_0^{-1}(x_0 - x^*) \right\|^2 - (K-1)\underline{\gamma}(2-\overline{\gamma}) \frac{\epsilon^2}{\bar{C}^2},$$

implying that

$$(K-1)\frac{\epsilon^2 \underline{\gamma}(2-\overline{\gamma})}{\bar{C}^2} \leq \left(\prod_{i=0}^{K-2} \|B_{i+1}^{-1} B_i\|^2 \right) \min_{x^* \in X^*} \left\| B_0^{-1}(x_0 - x^*) \right\|^2,$$

contradicting the definition of K. **Q.E.D.**

We now consider the dynamic stepsize rule with unknown f^*, where

$$\alpha_k = \gamma_k \frac{f(x_k) - f_k^{\rm lev}}{\|B_k' g_k\|^2}, \qquad 0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2, \tag{5.5}$$

the estimates $f_k^{\rm lev}$ are given by

$$f_k^{\rm lev} = \min_{0 \leq j \leq k} f(x_j) - \delta_k, \tag{5.6}$$

for positive scalars δ_k. We discuss two adjustment procedures for updating δ_k, which are modifications of the procedures discussed in Section 2.6. We start with the first adjustment procedure, where δ_k is updated according to the following procedure

$$\delta_{k+1} = \begin{cases} \lambda \delta_k & \text{if } f(x_{k+1}) \leq f_k^{\rm lev}, \\ \max\{\beta \delta_k, \delta\} & \text{if } f(x_{k+1}) > f_k^{\rm lev}, \end{cases} \tag{5.7}$$

where δ_0, δ, β, and λ are fixed positive scalars with $\beta < 1$ and $\lambda \geq 1$.

For the method using the stepsize (5.5)–(5.7), we have the following result.

Proposition 5.6: Let Assumption 5.1 hold, and let the sequence $\{x_k\}$ be generated by the variable metric method with the dynamic stepsize (5.5)–(5.7).

(a) If $f^* = -\infty$, then

$$\inf_{k \geq 0} f(x_k) = f^*.$$

(b) If $f^* > -\infty$, then

$$\inf_{k \geq 0} f(x_k) \leq f^* + \delta.$$

Proof: We prove (a) and (b) simultaneously. To arrive at a contradiction, assume that

$$\inf_{k \geq 0} f(x_k) - \delta > f^*. \tag{5.8}$$

Each time the target level is attained [i.e., $f(x_k) \leq f_{k-1}^{\rm lev}$], the current best function value $\min_{0\leq j\leq k} f(x_j)$ decreases by at least δ [cf. Eqs. (5.6) and (5.7)], so in view of Eq. (5.8), the target value can be attained only a finite number times. From Eq. (5.7) it follows that after finitely many iterations, δ_k is decreased to the threshold value and remains at that value for all subsequent iterations, i.e., there is a $\hat{k}$ such that

$$\delta_k = \delta, \qquad \forall\ k \geq \hat{k}. \tag{5.9}$$

In view of Eq. (5.8), there exists $\hat{y} \in \Re^n$ such that

$$\inf_{k\geq 0} f(x_k) - \delta \geq f(\hat{y}).$$

Thus, by Eqs. (5.6) and (5.9), we have

$$f_k^{\rm lev} = \min_{0\leq j\leq k} f(x_j) - \delta \geq \inf_{k\geq 0} f(x_k) - \delta \geq f(\hat{y}), \qquad \forall\ k \geq \hat{k}, \tag{5.10}$$

By Lemma 5.1, with $y = \hat{y}$ and α_k as in Eq. (5.5), it follows that

$$\left\|B_{k+1}^{-1}(x_{k+1}-\hat{y})\right\|^2 \leq \|B_{k+1}^{-1}B_k\|^2\Big[\left\|B_k^{-1}(x_k-\hat{y})\right\|^2 - \alpha_k\Big(2\big(f(x_k)-f(\hat{y})\big) - \gamma_k\big(f(x_k)-f_k^{\rm lev}\big)\Big)\Big].$$

Using the fact $f_k^{\rm lev} \geq f(\hat{y})$ for all $k \geq \hat{k}$ [cf. Eq. (5.10)] and the definition of α_k [cf. Eq. (5.5)], from the preceding relation we obtain for all $k \geq \hat{k}$,

$$\begin{aligned}\left\|B_{k+1}^{-1}(x_{k+1}-\hat{y})\right\|^2 &\leq \|B_{k+1}^{-1}B_k\|^2\Big(\left\|B_k^{-1}(x_k-\hat{y})\right\|^2 - \alpha_k(2-\gamma_k)\big(f(x_k)-f_k^{\rm lev}\big)\Big)\\ &= \|B_{k+1}^{-1}B_k\|^2\left(\left\|B_k^{-1}(x_k-\hat{y})\right\|^2 - \gamma_k(2-\gamma_k)\frac{\big(f(x_k)-f_k^{\rm lev}\big)^2}{\|B_k'g_k\|^2}\right).\end{aligned}$$

Since $\gamma_k \in [\underline{\gamma},\overline{\gamma}]$ and $f(x_k)-f_k^{\rm lev} \geq \delta_k \geq \delta$ for all k [cf. Eqs. (5.5)–(5.7)], and since $\|B_k'g_k\| \leq \bar{C}$ [cf. Assumption 5.1(a)], it follows that

$$\left\|B_{k+1}^{-1}(x_{k+1}-\hat{y})\right\|^2 \leq \|B_{k+1}^{-1}B_k\|^2\left(\left\|B_k^{-1}(x_k-\hat{y})\right\|^2 - \underline{\gamma}(2-\overline{\gamma})\frac{\delta^2}{\bar{C}^2}\right), \qquad \forall\ k \geq \hat{k}.$$

Furthermore, by Assumption 5.1(b), we have $\|B_{k+1}^{-1}B_k\| \geq 1$ for all k, implying that for all $k \geq \hat{k}$,

$$\begin{aligned}\left\|B_{k+1}^{-1}(x_{k+1}-\hat{y})\right\|^2 &\leq \|B_{k+1}^{-1}B_k\|^2\left\|B_k^{-1}(x_k-\hat{y})\right\|^2 - \underline{\gamma}(2-\overline{\gamma})\frac{\delta^2}{\bar{C}^2}\\ &\leq \left(\prod_{i=\hat{k}}^{k}\|B_{i+1}^{-1}B_i\|^2\right)\left\|B_{\hat{k}}^{-1}(x_{\hat{k}}-\hat{y})\right\|^2 - (k+1-\hat{k})\underline{\gamma}(2-\overline{\gamma})\frac{\delta^2}{\bar{C}^2}.\end{aligned}$$

However, by Assumption 5.1(b), we have $\prod_{i=\hat{k}}^{k} \|B_{i+1}^{-1}B_i\|^2 \leq \prod_{i=0}^{\infty} \|B_{i+1}^{-1}B_i\|^2 < \infty$, so that the preceding relation cannot hold for k sufficiently large, a contradiction. **Q.E.D.**

We now describe the method that employs the stepsize (5.5)-(5.6), where the parameters δ_k are adjusted according to a path-based procedure. The idea here is to adjust δ_k only if a sufficient descent occurs or if the iterates travel a path longer than some prescribed path bound b. Thus, the idea is the same as in the path-bound procedure of Section 2.6, however, there is a difference: we here measure the path length of the iterates in a variable metric. The method (5.1) using the path-based procedure is given in the following algorithm.

Path-Based Variable Metric Algorithm

Step 0 (*Initialization*) Select x_0, B_0, $\delta_0 > 0$, and $b > 0$. Set $\sigma_0 = 0$, $f_{-1}^{\rm rec} = \infty$. Set $k = 0$, $l = 0$, and $k(l) = 0$ [$k(l)$ will denote the iteration number when the l-th update of $f_k^{\rm lev}$ occurs].

Step 1 (*Function evaluation*) Calculate $f(x_k)$. If $f(x_k) < f_{k-1}^{\rm rec}$, then set $f_k^{\rm rec} = f(x_k)$. Otherwise set $f_k^{\rm rec} = f_{k-1}^{\rm rec}$ [so that $f_k^{\rm rec}$ keeps the record of the smallest value attained by the iterates that are generated so far, i.e., $f_k^{\rm rec} = \min_{0\leq j\leq k} f(x_j)$].

Step 2 (*Sufficient descent*) If $f(x_k) \leq f_{k(l)}^{\rm rec} - \frac{\delta_l}{2}$, then set $k(l+1) = k$, $\sigma_k = 0$, $\delta_{l+1} = \delta_l$, increase l by 1, and go to Step 4.

Step 3 (*Oscillation detection*) If $\sigma_k > b$, then set $k(l+1) = k$, $\sigma_k = 0$, $\delta_{l+1} = \frac{\delta_l}{2}$, and increase l by 1.

Step 4 (*Iterate update*) Set $f_k^{\rm lev} = f_{k(l)}^{\rm rec} - \delta_l$. Select B_k and $\gamma_k \in [\underline{\gamma}, \overline{\gamma}]$. Compute a subgradient g_k of f at x_k, and compute x_{k+1} via Eq. (5.1) with the stepsize (5.5).

Step 5 (*Path length update*) Set $\sigma_{k+1} = \sigma_k + \alpha_k\|B_k'g_k\|$, increase k by 1, and go to Step 1.

Upon each change of the target level $f_k^{\rm lev}$, which occurs at $k = k(l)$ (see Steps 2 and 3), the parameter σ_k is reset to zero so as to keep track of the path length traveled by the subsequent iterates $x_{k(l)}, x_{k(l)+1}, \ldots$. As seen from Step 5 and the iterate definition $x_{k+1} = x_k - \alpha_k B_k B_k' g_k$, this path is measured in a variable metric

$$\sigma_{k+1} = \sigma_k + \alpha_k\|B_k'g_k\| = \sum_{j=k(l)}^{k} \alpha_j\|B_j'g_j\| = \sum_{j=k(l)}^{k} \left\|B_j^{-1}(x_{j+1} - x_j)\right\|.$$

We next prove the correctness of the algorithm. We first give a preliminary result showing that the target values $f_k^{\rm lev}$ are updated infinitely often (i.e., $l \to \infty$), and that $\inf_{k\geq 0} f(x_k) = -\infty$ when the sequence $\{\delta_l\}$ is bounded away from zero.

Lemma 5.3: Let Assumption 5.1(a) hold. Then, for the path-based variable metric algorithm, we have $l \to \infty$, and either $\inf_{k\geq 0} f(x_k) = -\infty$ or $\lim_{l\to\infty} \delta_l = 0$.

Proof: Assume that l takes only a finite number of values, say $l = 0, 1, \ldots, \bar{l}$. In this case, we have

$$\sigma_{k+1} = \sigma_k + \alpha_k \|B_k' g_k\| \leq b, \qquad \forall\ k \geq k(\bar{l}),$$

implying that

$$\lim_{k \to \infty} \alpha_k \|B_k' g_k\| = 0.$$

But this is impossible, since by the definition of α_k and Assumption 5.1(a), we have

$$\alpha_k \|B_k' g_k\| = \gamma_k \frac{f(x_k) - f_k^{\text{lev}}}{\|B_k' g_k\|} \geq \frac{\underline{\gamma}\, \delta_{\bar{l}}}{\bar{C}}, \qquad \forall\ k \geq k(\bar{l}).$$

Hence, $l \to \infty$.

Let $\delta = \lim_{l \to \infty} \delta_l$. If $\delta > 0$, then from Steps 2 and 3 it follows that for all l large enough, we have $\delta_l = \delta$ and

$$f^{\text{rec}}_{k(l+1)} - f^{\text{rec}}_{k(l)} \leq -\frac{\delta}{2},$$

implying that $\inf_{k \geq 0} f(x_k) = -\infty$. **Q.E.D.**

For the algorithm, we have the following convergence result.

Proposition 5.7: Let Assumption 5.1 hold. Then, for the path-based variable metric algorithm, we have

$$\inf_{k \geq 0} f(x_k) = f^*.$$

Proof: If $\lim_{l \to \infty} \delta_l > 0$, then by Lemma 5.3, we have $\inf_{k \geq 0} f(x_k) = -\infty$ and we are done, so assume that $\lim_{l \to \infty} \delta_l = 0$. Let Λ be given by

$$\Lambda = \left\{ l \;\middle|\; \delta_l = \frac{\delta_{l-1}}{2},\ l \geq 1 \right\}.$$

Then, from Steps 3 and 5, we obtain

$$\sigma_k = \sigma_{k-1} + \alpha_{k-1} \|B_{k-1}' g_{k-1}\| = \sum_{j=k(l)}^{k-1} \alpha_j \|B_j' g_j\|,$$

so that $k(l+1) = k$ and $l + 1 \in \Lambda$ whenever $\sum_{j=k(l)}^{k-1} \alpha_j \|B_j' g_j\| > b$ at Step 3. Thus, by using $\|B_k' g_k\| \leq \bar{C}$ [cf. Assumption 5.1(a)], we see that

$$\sum_{j=k(l-1)}^{k(l)-1} \alpha_j \geq \frac{1}{\bar{C}} \sum_{j=k(l-1)}^{k(l)-1} \alpha_j \|B_j' g_j\| > \frac{b}{\bar{C}}, \qquad \forall\ l \in \Lambda.$$

Since $\delta_l \to 0$, it follows that the cardinality of Λ is infinite, and therefore we have

$$\sum_{j=0}^{\infty} \alpha_j \geq \sum_{l \in \Lambda} \sum_{j=k(l-1)}^{k(l)-1} \alpha_j > \sum_{l \in \Lambda} \frac{b}{\bar{C}} = \infty. \tag{5.11}$$

To obtain a contradiction, suppose that $\inf_{k \geq 0} f(x_k) > f^*$, so that for some $\hat{y} \in \Re^n$ and $\epsilon > 0$, we have

$$\inf_{k \geq 0} f(x_k) - \epsilon \geq f(\hat{y}). \tag{5.12}$$

Since $\delta_l \to 0$, there is a large enough $\hat{l}$ such that $\delta_l \leq \epsilon$ for all $l \geq \hat{l}$, implying that

$$f_k^{\rm lev} = f_{k(l)}^{\rm rec} - \delta_l \geq \inf_{k \geq 0} f(x_k) - \epsilon \geq f(\hat{y}), \qquad \forall\ k \geq k(\hat{l}).$$

Using this relation and the definition of α_k in Lemma 5.1,where $y = \hat{y}$, we obtain for all $k \geq k(\hat{l})$,

$$\begin{aligned}
\left\|B_{k+1}^{-1}(x_{k+1} - \hat{y})\right\|^2 &\leq \|B_{k+1}^{-1}B_k\|^2 \Big[\left\|B_k^{-1}(x_k - \hat{y})\right\|^2 \\
&\quad - \alpha_k\Big(2\big(f(x_k) - f(\hat{y})\big) - \gamma_k\big(f(x_k) - f_k^{\rm lev}\big)\Big)\Big] \\
&\leq \|B_{k+1}^{-1}B_k\|^2\Big(\left\|B_k^{-1}(x_k - \hat{y})\right\|^2 - \alpha_k(2-\gamma_k)\big(f(x_k) - f_k^{\rm lev}\big)\Big) \\
&= \|B_{k+1}^{-1}B_k\|^2\left(\left\|B_k^{-1}(x_k - \hat{y})\right\|^2 - \gamma_k(2-\gamma_k)\frac{\big(f(x_k) - f_k^{\rm lev}\big)^2}{\|B_k' g_k\|^2}\right).
\end{aligned}$$

Because $\|B_{k+1}^{-1}B_k\| \geq 1$ and $\gamma_k \in [\underline{\gamma}, \overline{\gamma}]$ for all k [cf. Assumption 5.1(b) and Eq. (5.5), respectively], we have for all $k \geq k(\hat{l})$,

$$\begin{aligned}
\left\|B_{k+1}^{-1}(x_{k+1} - \hat{y})\right\|^2 &\leq \|B_{k+1}^{-1}B_k\|^2\left\|B_k^{-1}(x_k - \hat{y})\right\|^2 - \underline{\gamma}(2-\overline{\gamma})\frac{\big(f(x_k) - f_k^{\rm lev}\big)^2}{\|B_k' g_k\|^2} \\
&\leq \left(\prod_{i=k(\hat{l})}^{k} \|B_{i+1}^{-1}B_i\|^2\right)\left\|B_{k(\hat{l})}^{-1}(x_{k(\hat{l})} - \hat{y})\right\|^2 \\
&\quad - \underline{\gamma}(2-\overline{\gamma})\sum_{j=k(\hat{l})}^{k} \frac{\big(f(x_j) - f_j^{\rm lev}\big)^2}{\|B_j' g_j\|^2}.
\end{aligned}$$

Since by Assumption 5.1(b), we have $\prod_{i=k(\hat{l})}^{k} \|B_{i+1}^{-1}B_i\|^2 \leq \prod_{i=0}^{\infty} \|B_{i+1}^{-1}B_i\|^2 < \infty$, from the preceding relation we obtain

$$\sum_{k=k(\hat{l})}^{\infty} \frac{\big(f(x_k) - f_k^{\rm lev}\big)^2}{\|B_k' g_k\|^2} < \infty.$$

This relation and the definition of α_k [cf. Eq. (5.5)] imply that $\sum_{k=k(\hat{l})}^{\infty} \alpha_k^2 < \infty$, and consequently $\alpha_k \to 0$. Furthermore, we have already shown that $\sum_{k=0}^{\infty} \alpha_k = \infty$ [cf. Eq. (5.11)], so that by Prop. 5.3, we have

$$\liminf_{k \to \infty} f(x_k) = f^*,$$

contradicting Eq. (5.12). **Q.E.D.**

For the dynamic stepsize with unknown optimal function value f^*, under assumption that the set X^* of optimal solutions is nonempty, we can estimate the number of iterations needed to guarantee achievement of f^* within some error. In particular, we have the following.

Proposition 5.8: Let Assumption 5.1 hold, and assume that the optimal solution set X^* is nonempty. Let the sequence $\{x_k\}$ be generated by the variable metric method with the dynamic stepsize of the form (5.5)–(5.6). Then, for the smallest positive integer K such that

$$\frac{\underline{\gamma}(2-\overline{\gamma})}{\bar{C}^2} \sum_{k=0}^{K-1} \delta_k^2 \leq \left(\prod_{i=0}^{K-1} \|B_{i+1}^{-1} B_i\|^2 \right) \min_{x^* \in X^*} \left\| B_0^{-1}(x_0 - x^*) \right\|^2,$$

we have

$$\min_{0 \leq k \leq K} f(x_k) \leq f^* + \max_{0 \leq j \leq K} \delta_j.$$

Proof: The proof is similar to that of Prop. 5.5, where to arrive at a contradiction, we assume that

$$f(x_k) > f^* + \max_{0 \leq j \leq k} \delta_j, \qquad \forall\ k = 0, 1, \ldots, K.$$

Q.E.D.

In particular, if in the dynamic stepsize (5.5), we use

$$f_k^{\text{lev}} = \min_{0 \leq j \leq k} f(x_j) - \delta, \qquad \forall\ k,$$

with a scalar $\delta > 0$, then from Prop. 5.8 it follows that for the nonnegative integer K given by

$$\frac{\underline{\gamma}(2-\overline{\gamma})}{\bar{C}^2} K\delta \leq \left(\prod_{i=0}^{K-1} \|B_{i+1}^{-1} B_i\|^2 \right) \min_{x^* \in X^*} \left\| B_0^{-1}(x_0 - x^*) \right\|^2,$$

we have

$$\min_{0 \leq k \leq K} f(x_k) \leq f^* + \delta.$$

Furthermore, if

$$f_k^{\text{lev}} = \min_{0 \leq j \leq k} f(x_j) - \delta_k, \qquad \forall\ k,$$

and δ_k are adjusted by using either Eq. (5.7) with $\lambda = 1$, or the path-based procedure, then δ_k is nonincreasing. Thus, in Prop. 5.8, we have

$$\min_{0 \leq k \leq K} f(x_k) \leq f^* + \delta_0.$$

Let us note that all the results of this section (cf. Props. 5.4–5.8 and Lemma 5.3) can hold if instead assuming that the sequence $\{B'_k g_k\}$ is bounded [cf. Assumption 5.1(a)], we assume that the sequence $\{B_k\}$ of linear transformations is bounded. The reason for this lies in the fact that, for the stepsizes of the form (5.3) or (5.5), the sequence $\{x_k\}$ is bounded, which implies the boundedness of the subgradients g_k. Thus, for such stepsizes, in order to have bounded $\|B'_k g_k\|$, it suffices to assume that $\|B_k\|$ is bounded.

6

Space Dilation Methods

In this chapter, we discuss a special class of variable metric methods where the metric is changed through space dilations, aiming at accelerated convergence. In particular, we will consider methods with two types of space dilations: along subgradient directions and along directions that can differ from subgradient directions.

As mentioned earlier, poor performance of subgradient methods is most notable in the cases where subgradient directions are almost orthogonal to a direction pointing toward a minimum. In such cases, typically, the subgradient components that are orthogonal to the directions pointing toward the minima are very small as compared to the other subgradient components. Thus, by moving along subgradients, the advances toward the set of minima are insignificant, and if the stepsize is small, the method can jam. This situation can be avoided by scaling the subgradients appropriately, and one such scaling can be carried out through space dilations along subgradients. The method with space dilations along subgradients was proposed and analyzed by N. Z. Shor. However, for this method, we here give new convergence results, including a new stepsize choice.

The situation where the subgradients are almost orthogonal to the directions pointing toward the set of minima can be alternatively viewed as the case where the cone of subgradient directions is too wide. In this case, convergence can be accelerated by transforming this cone into a narrower cone, which can be done, for example, by using space dilation along the difference of the two successive subgradients. To include this case, as well as other choices that may be appropriate for specific problems, we here propose and analyze a rather general dilation method.

Our methods are applicable to unconstrained minimization problems

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in \Re^n, \end{aligned}$$

where $f : \Re^n \mapsto \Re$ is a convex function.

This chapter is organized as follows: In Section 6.1, we introduce and interpret dilation along subgradients. In Section 6.2, we establish some important properties of space dilation transformations in a general form. In Section 6.3 and 6.4, respectively, we give some basic relations and we discuss the convergence properties of the dilation method along subgradients. In Sections 6.5–6.7, we introduce and analyze the general dilation method that can use directions different from subgradient directions.

6.1. DILATION ALONG SUBGRADIENTS

In this section, we introduce a method using dilation along subgradients, where the coordinate transformations (changes in metric) are carried out through successive dilations. At a typical iteration of the method, we have the current iterate x_k, a subgradient g_k of f at x_k, and a matrix B_k. We next compute the new iterate according to the following rule

$$x_{k+1} = x_k - \alpha_k \frac{B_k B_k' g_k}{\|B_k' g_k\|}, \tag{6.1}$$

where the scalar α_k is a positive stepsize, and the initial transformation is $B_0 = I$. We then update B_k as follows:

$$B_{k+1} = B_k R_{\rho_k}(\xi_k), \tag{6.2}$$

where

$$R_{\rho_k}(\xi_k) = I + (\rho_k - 1)\xi_k \xi_k', \tag{6.3}$$

for the scalar ρ_k is positive and the vector ξ_k is given by

$$\xi_k = \frac{B_k' g_k}{\|B_k' g_k\|}. \tag{6.4}$$

This method was proposed by Shor [Sho70a] (see also Shor [Sho70b], [Sho77a], [Sho77b], [Sho83], [Sho85], and [Sho98]). We will refer to it as *dilation method along subgradients.* Let us mention that the celebrated ellipsoid method, which is due to Khachian [Kha79] (see also Bertsimas and Tsitsiklis [BeT97], p. 363), is just a special case of dilation method along subgradients (cf. Shor [Sho77a] and [Sho98]).

To interpret the method, let us first take a closer look at the transformation

$$R_\rho(\xi) = I + (\rho - 1)\xi\xi',$$

where ρ is a positive scalar and ξ is a vector with $\|\xi\| = 1$. If a vector x is orthogonal to ξ, then the transformation $R_\rho(\xi)$ leaves the vector x unchanged, i.e., $R_\rho(\xi)x = x$. If the vector x is parallel to ξ, then $R_\rho(\xi)$ scales the vector x by the factor ρ, i.e., $R_\rho(\xi)x = \rho x$. Thus, the transformation $R_\rho(\xi)$ does not change the components of a vector that are orthogonal to ξ and scales the other components of the vector by the factor ρ.

Let us now interpret the method. For this, consider the coordinate transformation given by

$$y = B_k^{-1}x, \qquad x \in \Re^n,$$

(as we will see later each transformation B_k is invertible). The vector $B_k'g_k$ is a subgradient of the function $F(y) = f(B_k y)$ at the point $y_k = B_k^{-1}x_k$, as discussed in Section 5.1, Suppose now that, starting at y_k, we move in the opposite direction of the normalized subgradient $B_k'g_k$, which gives

$$y_{k+1} = y_k - \alpha_k \frac{B_k'g_k}{\|B_k'g_k\|}.$$

Thus, the iteration (6.1) corresponds to an iteration of the subgradient method applied to the function $F(y) = f(B_k y)$.

Assume, for the sake of simplicity, that the set of minima consists of a single vector x^*. Let y^* be the vector in the y-space corresponding to x^*, and note that y^* is the minimum of $F(y)$. Then, by using $y = B_k^{-1}x$, we have

$$(B_k'g_k)'(y_k - y^*) = g_k'(x_k - x^*) = 0.$$

Thus, if the subgradient g_k is almost orthogonal to the direction $x_k - x^*$, then, in the y-space, the same is true for the subgradient $B_k'g_k$ and the direction $y_k - y^*$. Then, in order to have the subsequent subgradient better pointed toward the minimum, we would like to scale down the subgradient components that are orthogonal to $y_k - y^*$, while leaving the other directions unchanged. For this, since $B_k'g_k$ is almost orthogonal to $y_k - y^*$, we can use a contraction along $B_k'g_k$, which is formally done via Eqs. (6.2)–(6.4).

Let us mention that there is an alternative implementation form of dilation method along subgradients that was given by Skokov in [Sko74], where

$$x_{k+1} = x_k - \alpha_k \frac{H_k g_k}{\sqrt{g_k' H_k g_k}}, \tag{6.5}$$

with $H_k = B_k B_k'$, and H_k is updated via the following formula

$$H_{k+1} = H_k + (\rho_k^2 - 1)\frac{H_k g_k g_k' H_k}{g_k' H_k g_k}. \tag{6.6}$$

These iteration formulas bear some similarities with those of quasi-Newton method (see, for example the textbook by Bertsekas [Ber99], p. 149), but other than this, there is no useful connection between these two methods.

The implementation of dilation method along subgradients in the form (6.5)-(6.6) is computationally more efficient than its implementation in the form (6.1)-(6.4). However, the method in the form (6.5)-(6.6) is more sensitive to computational errors. In particular, theoretically, the matrices H_k are positive definite. However, due to computational errors, the approximation of H_k that is actually used instead of H_k may not be positive definite, and this can distort the method.

6.2. PROPERTIES OF DILATION TRANSFORMATIONS

In this section, we give basic properties of the transformation $R_\rho(\xi)$, which is the building block for the transformations B_k. We then examine the properties of B_k for a general case where the directions ξ_k are given by $\xi_k = B_k' d_k / \|B_k' d_k\|$ with nonzero vectors d_k.

In the following lemma, we list some important features of the transformation $R_\rho(\xi)$. The proof of this lemma can be found in Shor [Sho98], p. 73.

Lemma 6.1: Let the linear transformation $R_\rho(\xi) : \Re^n \mapsto \Re^n$ be given by

$$R_\rho(\xi) = I + (\rho - 1)\xi\xi',$$

where ξ is a vector with $\|\xi\| = 1$, ρ is a scalar, and I is the identity matrix. We then have:

(a) For any nonzero scalar ν,

$$R_\rho(\xi) R_\nu(\xi) = R_{\rho\nu}(\xi).$$

In particular, if $\rho \neq 0$, then the linear transformation $R_\rho(\xi)$ is invertible and its inverse is $R_{\frac{1}{\rho}}(\xi)$, i.e.,

$$R_\rho(\xi) R_{\frac{1}{\rho}}(\xi) = I.$$

(b) For any $x \in \Re$,

$$\big\|R_\rho(\xi)x\big\|^2 = \|x\|^2 + (\rho^2 - 1)(\xi' x)^2.$$

(c) The norm of $R_\rho(\xi)$ is given by

$$\big\|R_\rho(\xi)\big\| = \max\big\{1, |\rho|\big\}.$$

We next establish basic properties of transformations B_k of the form $B_{k+1} = B_k R_{\rho_k}(\xi_k)$, where the direction ξ_k is not necessarily the same as in dilation method along subgradients [cf. Eq.(6.4)]. By considering more general directions ξ_k, we can bring to the surface the properties of B_k that are independent of any iterative process. This, in turn, will allow us to consider the methods that use dilation along directions other than the subgradient directions.

The basic properties of transformations B_k are given in the following lemma. The proof of this lemma exploits some ideas of Nesterov [Nes84].

Lemma 6.2: Let the sequence $\{B_k\}$ of linear transformations be such that

$$B_{k+1} = B_k R_{\rho_k}(\xi_k), \qquad \text{with} \qquad \xi_k = \frac{B_k' d_k}{\|B_k' d_k\|}, \qquad \forall\ k,$$

where $B_0 = I$, d_k are nonzero vectors, and ρ_k are scalars. Assume that the vectors d_k are bounded, i.e., there exists a positive scalar C such that

$$\|d_k\| \leq C, \qquad \forall\ k.$$

Assume further that the scalars ρ_k are such that

$$0 < \underline{\rho} \leq \rho_k \leq \overline{\rho} < 1, \qquad \forall\ k.$$

We then have:

(a) $\lim_{k\to 0} \|B_k' d_k\| = 0$.

(b) For all k,

$$\min_{0\leq j\leq k} \|B_j' d_j\| \leq C\, \frac{\overline{\rho}^{(k+1)/n}}{\underline{\rho}} \sqrt{\frac{(k+1)}{n} \frac{(1-\underline{\rho}^2)}{\left(1-\overline{\rho}^{2(k+1)/n}\right)}}.$$

Proof: (a) We will prove that $B_k B_k' d_k \to 0$, and then conclude that $B_k' d_k \to 0$. For this, we define

$$H_k = B_k B_k', \qquad \forall\ k,$$

and we note that the matrix H_k is symmetric and positive definite for all k. We next establish a relation between H_{k+1} and H_k, which will be important for our proof. By definition of B_{k+1}, we have

$$H_{k+1} = B_k R_{\rho_k}(\xi_k) R_{\rho_k}'(\xi_k) B_k', \qquad \forall\ k.$$

Since $R_{\rho_k}(\xi_k)$ is symmetric, by using Lemma 6.1(a) with $\rho = \nu = \rho_k$, we obtain

$$H_{k+1} = B_k R_{\rho_k^2}(\xi_k) B_k', \qquad \forall\ k.$$

Therefore,

$$H_{k+1} = B_k \left(I + (\rho_k^2 - 1)\xi_k \xi_k'\right) B_k' = H_k + (\rho_k^2 - 1) B_k \xi_k \xi_k' B_k', \qquad \forall\ k.$$

By using $\xi_k = B_k' d_k / \|B_k' d_k\|$ in this relation, we see that

$$H_{k+1} = H_k + (\rho_k^2 - 1) \frac{H_k d_k d_k' H_k}{d_k' H_k d_k}, \qquad \forall\ k, \tag{6.7}$$

We now show that $B_k B_k' d_k \to 0$ by estimating the trace of the matrix H_{k+1}, denoted by $\mathrm{Tr}(H_{k+1})$. In view of Eq. (6.7), we have

$$\mathrm{Tr}(H_{k+1}) = \mathrm{Tr}(H_k) - (1-\rho_k^2) \frac{\mathrm{Tr}(H_k d_k d_k' H_k)}{d_k' H_k d_k}, \qquad \forall\ k.$$

Since H_k is positive definite, we must have $d_k' H_k d_k > 0$. Furthermore, because $\rho_k < 1$, by Lemma 6.1(c), it can be seen that $\|H_k\| \leq 1$, which together with our assumption that $\|d_k\| \leq C$ yields

$$0 < d_k' H_k d_k \leq \|H_k\| \|d_k\|^2 \leq C^2, \qquad \forall\ k.$$

Moreover, since H_k is symmetric, we have $\mathrm{Tr}(H_k d_k d_k' H_k) = \|H_k d_k\|^2$. Therefore, for all k,

$$\mathrm{Tr}(H_{k+1}) \leq \mathrm{Tr}(H_k) - (1-\rho_k^2)\frac{\|H_k d_k\|^2}{C^2} \leq \mathrm{Tr}(H_0) - \sum_{j=0}^{k}(1-\rho_j^2)\frac{\|H_j d_j\|^2}{C^2}.$$

For each k, all eigenvalues of H_k are positive because H_k is a positive definite. Since the trace of a matrix is equal to the sum of its eigenvalues, we have that $\mathrm{Tr}(H_k) > 0$ for all k. Furthermore, since $H_0 = I$, we have $\mathrm{Tr}(H_0) = n$. By using these relations and the inequality $\rho_k \leq \overline{\rho}$, we obtain

$$0 < \mathrm{Tr}(H_{k+1}) \leq n - \frac{(1-\overline{\rho}^2)}{C^2}\sum_{j=0}^{k}\|H_j d_j\|^2, \qquad \forall\ k,$$

implying, by $1-\overline{\rho} > 0$, that the sum $\sum_{j=0}^{\infty}\|H_j d_j\|^2$ is finite, and hence

$$\lim_{k\to\infty}\|H_k d_k\| = 0.$$

We now have by the definition of H_k,

$$\|B_k' d_k\|^2 = d_k' H_k d_k \leq \|d_k\|\|H_k d_k\|, \qquad \forall\ k.$$

By our assumption, the vectors d_k are bounded, so that

$$\limsup_{k\to\infty}\|B_k' d_k\|^2 \leq 0,$$

implying that $\|B_k' d_k\| \to 0$.

(b) We will derive the desired relation by estimating the determinant of the matrix H_{k+1}^{-1}, denoted by $\det\big(H_{k+1}^{-1}\big)$. We have by Eq. (6.7),

$$\big(H_{k+1}\big)^{-1} = \left(H_k + (\rho_k^2 - 1)\frac{H_k d_k d_k' H_k}{d_k' H_k d_k}\right)^{-1}, \qquad \forall\ k.$$

According to Sherman-Morrison formula (cf. Golub and Van Loan [GoV84], p. 3), we have

$$(A + uv')^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{1 + v'A^{-1}u},$$

for any invertible matrix A, and any vectors u and v. By using this formula with the following identifications:

$$A = H_k, \qquad u = (\rho_k^2 - 1)\frac{H_k d_k}{d_k' H_k d_k}, \qquad v = H_k d_k,$$

after some algebra, we obtain

$$H_{k+1}^{-1} = H_k^{-1} + \frac{(1-\rho_k^2)}{\rho_k^2}\frac{d_k d_k'}{d_k' H_k d_k}, \qquad \forall\ k. \tag{6.8}$$

Therefore, for all k,

$$\begin{aligned}\det\big(H_{k+1}^{-1}\big) &= \det\left(H_k^{-1}\left(I + \frac{(1-\rho_k^2)}{\rho_k^2}\frac{H_k d_k d_k'}{d_k' H_k d_k}\right)\right)\\ &= \det\big(H_k^{-1}\big)\det\left(I + \frac{(1-\rho_k^2)}{\rho_k^2}\frac{H_k d_k d_k'}{d_k' H_k d_k}\right).\end{aligned}$$

Since $\det(I+uv') = 1+v'u$ for any vectors u and v (see Golub and Van Loan [GoV84], p. 43), the last determinant in the relation above is equal to $1/\rho_k^2$, implying that for all k,

$$\det\big(H_{k+1}^{-1}\big) = \det\big(H_k^{-1}\big)\frac{1}{\rho_k^2} = \det\big(H_0^{-1}\big)\frac{1}{\prod_{j=0}^k \rho_j^2} \geq \frac{1}{\overline{\rho}^{2(k+1)}}, \tag{6.9}$$

where the inequality above follows from $H_0 = I$ and $\rho_j \leq \overline{\rho}$.

The determinant $\det\big(H_{k+1}^{-1}\big)$ is equal to the product $\prod_{i=1}^n \lambda_i$ of its eigenvalues λ_i, while the trace 2(k+1) $\mathrm{Tr}\big(H_{k+1}^{-1}\big)$ is equal to the sum $\sum_{i=1}^n \lambda_i$, where all eigenvalues λ_i are positive since H_{k+1} is a positive definite matrix and so is H_{k+1}^{-1}. Furthermore, because the geometric mean $\left(\prod_{i=1}^n \lambda_i\right)^{1/n}$ of any positive scalars $\lambda_1, \ldots, \lambda_n$, $n \geq 1$, is smaller than their arithmetic mean $(\lambda_1 + \cdots + \lambda_n)/n$, it follows that

$$\Big(\det\big(H_{k+1}^{-1}\big)\Big)^{\frac{1}{n}} \leq \frac{1}{n}\mathrm{Tr}\big(H_{k+1}^{-1}\big), \qquad \forall\ k.$$

This relation and Eq. (6.9) yield

$$\frac{n}{\overline{\rho}^{2(k+1)/n}} \leq n\Big(\det\big(H_{k+1}^{-1}\big)\Big)^{\frac{1}{n}} \leq \mathrm{Tr}\big(H_{k+1}^{-1}\big), \qquad \forall\ k.$$

In view of Eq. (6.8), we have

$$\begin{aligned}\mathrm{Tr}\big(H_{k+1}^{-1}\big) &= \mathrm{Tr}\big(H_k^{-1}\big) + \frac{(1-\rho_k^2)}{\rho_k^2}\frac{\mathrm{Tr}(d_k d_k')}{d_k' H_k d_k}\\ &= \mathrm{Tr}\big(H_k^{-1}\big) + \frac{(1-\rho_k^2)}{\rho_k^2}\frac{\|d_k\|^2}{d_k' H_k d_k}.\end{aligned}$$

Hence, for all k,

$$\frac{n}{\overline{\rho}^{2(k+1)/n}} \leq \mathrm{Tr}\big(H_k^{-1}\big) + \frac{(1-\rho_k^2)}{\rho_k^2}\frac{\|d_k\|^2}{d_k' H_k d_k} \leq n + \sum_{j=0}^k \frac{(1-\rho_j^2)}{\rho_j^2}\frac{\|d_j\|^2}{d_j' H_j d_j}.$$

Since $\rho_k \geq \overline{\rho} > 0$, we have that $(1-\rho_k^2)/\rho_k^2 \leq (1-\underline{\rho}^2)/\underline{\rho}^2$ for all k. Furthermore, by our assumption, we have $\|d_k\| \leq C$ for all k, implying that

$$\begin{aligned}\frac{n}{\overline{\rho}^{2(k+1)/n}} &\leq n + C^2\frac{(1-\underline{\rho}^2)}{\underline{\rho}^2}\sum_{j=0}^{k}\frac{1}{d_j'H_jd_j}\\ &\leq n + C^2\frac{(1-\underline{\rho}^2)}{\underline{\rho}^2}\frac{(k+1)}{\min_{0\leq j\leq k} d_j'H_jd_j}, \qquad \forall\ k.\end{aligned}$$

After some algebra, from this relation we obtain

$$\min_{0\leq j\leq k} d_j'H_jd_j \leq C^2\frac{(k+1)}{n}\frac{(1-\underline{\rho}^2)}{\underline{\rho}^2}\frac{\overline{\rho}^{2(k+1)/n}}{\left(1-\overline{\rho}^{2(k+1)/n}\right)}, \qquad \forall\ k,$$

from which the desired estimate follows by using the relation $d_j'H_jd_j = \|B_jd_j\|^2$. **Q.E.D.**

Lemma 6.2 has important consequences for the dilation method along subgradients. In particular, by part (a) of this lemma (where $d_k = g_k$), we have $B_k'g_k \to 0$, which we will use to prove convergence of the method. Furthermore, by using part (b) of the lemma, we will be able to assess the convergence rate of the method. Moreover, the lemma is also important for the analysis of the dilation method that we consider later in Section 6.4.

6.3. ASSUMPTIONS AND SOME BASIC RELATIONS

Here, we give our assumptions and a key relation between the two successive iterates of the dilation method (6.1). The assumption that we use in assessing convergence is the following:

Assumption 6.1:

(a) There exists a positive scalar C such that

$$\|g_k\| \leq C, \qquad \forall\ k.$$

(b) There exists a positive scalar μ such that for every nonempty level set $\{y \in \Re^n \mid f(y) \leq \omega\}$ and every vector x with $f(x) \geq \omega$, we have

$$f(x) - \omega \geq \mu\ dist(x, L_\omega),$$

where L_ω is the level set $\{y \in \Re^n \mid f(y) \leq \omega\}$.

Assumption 6.1 is satisfied, for example, when f is a polyhedral function, i.e.,

$$f(x) = \max\{a_1'x + b_1, \ldots, a_m'x + b_m\},$$

for some vectors a_i (not all equal to zero) and scalars b_i. In this case, we have

$$C = \max_{1\leq i\leq m} \|a_i\|, \qquad \mu = \min_{1\leq i\leq m} \big\{\|a_i\| \mid a_i \neq 0\big\}.$$

Under Assumption 6.1, we can establish a basic relation for the iterates generated by dilation method along subgradients, as seen in the following lemma.

Lemma 6.3: Let Assumption 6.1 hold, and let the parameters ρ_k be such that

$$0 < \rho_k \leq 1, \qquad \forall\ k.$$

Let further $\{x_k\}$ be the sequence generated by dilation method along subgradients. Then, for any k, and any scalar ω such that $f(x_k) \geq \omega$ and the level set $L_\omega = \big\{y \in \Re^n \mid f(y) \leq \omega\big\}$ is nonempty, we have

$$\begin{aligned}\min_{y\in L_\omega} \big\|B_{k+1}^{-1}(x_{k+1}-y)\big\|^2 \leq \min_{y\in L_\omega} \big\|B_k^{-1}(x_k-y)\big\|^2 &+ \frac{(1-\rho_k^2)}{\rho_k^2}\frac{C^2}{\mu^2}\frac{\big(f(x_k)-\omega\big)^2}{\|B_k'g_k\|^2}\\ &- 2\frac{\alpha_k}{\rho_k^2}\frac{\big(f(x_k)-\omega\big)}{\|B_k'g_k\|} + \frac{\alpha_k^2}{\rho_k^2},\end{aligned}$$

where C and μ are as in Assumption 6.1.

Proof: Let k be arbitrary but fixed. By the definition of x_{k+1} and B_{k+1} [cf. Eqs. (6.1) and (6.2), respectively], we have for any y,

$$\begin{aligned}B_{k+1}^{-1}(x_{k+1}-y) &= \big(R_{\rho_k}(\xi_k)\big)^{-1} B_k^{-1}\left(x_k - y - \alpha_k \frac{B_k B_k' g_k}{\|B_k'g_k\|}\right)\\ &= \big(R_{\rho_k}(\xi_k)\big)^{-1}\left(B_k^{-1}(x_k-y) - \alpha_k\frac{B_k'g_k}{\|B_k'g_k\|}\right),\end{aligned}$$

Since by Lemma 6.1(a), we have $\big(R_{\rho_k}(\xi_k)\big)^{-1} = R_{1/\rho_k}(\xi_k)$, and by Lemma 6.1(b), we have

$$\big\|R_\rho(\xi)x\big\|^2 = \|x\|^2 + (\rho^2-1)(\xi'x)^2, \qquad \forall\ x,$$

it follows that for any y,

$$\begin{aligned}\big\|B_{k+1}^{-1}(x_{k+1}-y)\big\|^2 &= \left\|R_{\frac{1}{\rho_k}}(\xi_k)\left(B_k^{-1}(x_k-y) - \alpha_k\frac{B_k'g_k}{\|B_k'g_k\|}\right)\right\|^2\\ &= \left\|B_k^{-1}(x_k-y) - \alpha_k\frac{B_k'g_k}{\|B_k'g_k\|}\right\|^2\\ &\quad + \left(\frac{1}{\rho_k^2}-1\right)\left(\xi_k'B_k^{-1}(x_k-y) - \alpha_k\xi_k'\frac{B_k'g_k}{\|B_k'g_k\|}\right)^2.\end{aligned}$$

By using the relation $\xi_k = B_k' g_k / \|B_k' g_k\|$ [cf. Eq. (6.4)] and by expanding the terms on the right hand-side in the preceding relation, we obtain for any y,

$$\begin{aligned}\left\|B_{k+1}^{-1}(x_{k+1}-y)\right\|^2 &= \left\|B_k^{-1}(x_k-y)\right\|^2 - 2\alpha_k \frac{g_k'(x_k-y)}{\|B_k' g_k\|} + \alpha_k^2 \\ &\quad + \frac{1-\rho_k^2}{\rho_k^2}\left(\frac{\big(g_k'(x_k-y)\big)^2}{\|B_k' g_k\|^2} - 2\alpha_k \frac{g_k'(x_k-y)}{\|B_k' g_k\|} + \alpha_k^2\right) \\ &= \left\|B_k^{-1}(x_k-y)\right\|^2 + \frac{(1-\rho_k^2)}{\rho_k^2}\frac{\big(g_k'(x_k-y)\big)^2}{\|B_k' g_k\|^2} \\ &\quad - 2\frac{\alpha_k}{\rho_k^2}\cdot\frac{g_k'(x_k-y)}{\|B_k' g_k\|} + \frac{\alpha_k^2}{\rho_k^2}.\end{aligned}$$

By Schwartz inequality, we have

$$\big(g_k'(x_k-y)\big)^2 \le \|g_k\|^2\|x_k-y\|^2,$$

while by the subgradient inequality, we have

$$f(x_k) - f(y) \le g_k'(x_k-y).$$

From the preceding three relations, since $1-\rho_k^2 \ge 0$, we obtain for any y with $f(x_k) \ge \omega \ge f(y)$,

$$\begin{aligned}\left\|B_{k+1}^{-1}(x_{k+1}-y)\right\|^2 &\le \left\|B_k^{-1}(x_k-y)\right\|^2 + \frac{(1-\rho_k^2)}{\rho_k^2}\frac{\|g_k\|^2\|x_k-y\|^2}{\|B_k' g_k\|^2} \\ &\quad - 2\frac{\alpha_k}{\rho_k^2}\frac{\big(f(x_k)-\omega\big)}{\|B_k' g_k\|} + \frac{\alpha_k^2}{\rho_k^2}.\end{aligned}$$

Using subgradient boundedness [cf. Assumption 6.1(a)] and taking the minimum over all y in the level set $L_\omega = \big\{y \in \Re^n \mid f(y) \le \omega\big\}$, we see that

$$\begin{aligned}\min_{y\in L_\omega}\left\|B_{k+1}^{-1}(x_{k+1}-y)\right\|^2 &\le \min_{y\in L_\omega}\left\|B_k^{-1}(x_k-y)\right\|^2 \\ &\quad + \frac{(1-\rho_k^2)}{\rho_k^2}\frac{C^2\big(dist(x_k,L_\omega)\big)^2}{\|B_k' g_k\|^2} \\ &\quad - 2\frac{\alpha_k}{\rho_k^2}\frac{\big(f(x_k)-\omega\big)}{\|B_k' g_k\|} + \frac{\alpha_k^2}{\rho_k^2}.\end{aligned}$$

Finally, by Assumption 6.1(b), we have

$$dist(x_k, L_\omega) \le \frac{1}{\mu}\big(f(x_k)-\omega\big),$$

which when substituted in the preceding inequality yields the desired relation. **Q.E.D.**

6.4. CONVERGENCE PROPERTIES OF THE METHOD WITH DILATION ALONG SUBGRADIENTS

In this section, we discuss the convergence properties of the method (6.1)–(6.4) using dynamic stepsize rules. Even though this method belongs to the class of variable metric methods, its behavior and its analysis are different from those of the variable metric methods of Chapter 5.

Let us briefly outline our analytical approach. Assuming, for the time being, that the optimal solution set consists of a single vector x^*, we note that, based on the subgradient defining property, the following relation holds

$$f(x_k) - f^* \leq g_k'(x_k - x^*), \qquad \forall\ k.$$

We then change the space coordinates appropriately, so that

$$f(x_k) - f^* \leq (B_k' g_k)' B_k^{-1}(x_k - x^*) \leq \|B_k' g_k\| \left\| B_k^{-1}(x_k - x^*) \right\|.$$

Thus, if the vectors $B_k' g_k$ converge to zero and the distances $\left\| B_k^{-1}(x_k - x^*) \right\|$ are bounded, then the function values $f(x_k)$ will converge to the optimal function value f^*. This, precisely, describes the basis of our analysis. In particular, we establish the convergence of the vectors $B_k' g_k$ by using Lemma 6.2. The harder task is to show that the distances $\left\| B_k^{-1}(x_k - x^*) \right\|$ are bounded, and that will be done by using Lemma 6.3 with appropriate stepsize choices.

6.4.1 Dynamic Stepsize Rule for known f^*

We establish here the convergence properties of the method using a dynamic stepsize rule for known f^*, where

$$\alpha_k = \gamma_k \frac{f(x_k) - f^*}{\|B_k' g_k\|}, \qquad 0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2, \quad \forall\ k. \tag{6.10}$$

For this stepsize, we have the following result.

Proposition 6.1: Let Assumption 6.1 hold, and let the parameters ρ_k be such that

$$0 < \underline{\rho} \leq \rho_k \leq \overline{\rho} < 1, \qquad \forall\ k,$$

$$(1 - \rho_k^2)\frac{C^2}{\mu^2} \leq \gamma_k(2 - \gamma_k), \qquad \forall\ k.$$

Assume further that the optimal solution set X^* is nonempty. Then, for the sequence $\{x_k\}$ generated by dilation method along subgradients and the dynamic stepsize (6.10), we have

$$\lim_{k \to \infty} f(x_k) = f^*.$$

Furthermore,

$$\min_{0\leq j\leq k} f(x_j) \leq f^* + C\,\frac{\overline{\rho}^{(k+1)/n}}{\underline{\rho}}\sqrt{\frac{(k+1)}{n}\,\frac{(1-\underline{\rho}^2)}{\left(1-\overline{\rho}^{2(k+1)/n}\right)}}\,dist(x_0,X^*).$$

Proof: We first prove that the distances $\min_{x^*\in X^*}\left\|B_k^{-1}(x_k-x^*)\right\|$ are nonincreasing. From Lemma 6.3, where $\omega = f^*$, we see that

$$\begin{aligned}\min_{x^*\in X^*}\left\|B_{k+1}^{-1}(x_{k+1}-x^*)\right\|^2 \leq \min_{x^*\in X^*}\left\|B_k^{-1}(x_k-x^*)\right\|^2 &\\ + \frac{(1-\rho_k^2)}{\rho_k^2}\frac{C^2}{\mu^2}\frac{\left(f(x_k)-f^*\right)^2}{\|B_k'g_k\|^2} &\\ - 2\frac{\alpha_k}{\rho_k^2}\frac{\left(f(x_k)-f^*\right)}{\|B_k'g_k\|} + \frac{\alpha_k^2}{\rho_k^2}, \qquad \forall\ k.&\end{aligned}$$

By using the definition of the stepsize α_k [cf. Eq. (6.10)], we obtain for all k,

$$\begin{aligned}\min_{x^*\in X^*}\left\|B_{k+1}^{-1}(x_{k+1}-x^*)\right\|^2 \leq \min_{x^*\in X^*}\left\|B_k^{-1}(x_k-x^*)\right\|^2 &\\ + \frac{\left(f(x_k)-f^*\right)^2}{\rho_k^2\|B_k'g_k\|^2}\left((1-\rho_k^2)\frac{C^2}{\mu^2} - \gamma_k(2-\gamma_k)\right).&\end{aligned}$$

By our assumption that $(1-\rho_k^2)\frac{C^2}{\mu^2} - \gamma_k(2-\gamma_k) \leq 0$ for all k, it follows that the scalar sequence $\left\{\min_{x^*\in X^*}\left\|B_k^{-1}(x_k-x^*)\right\|\right\}$ is nonincreasing, and since $B_0 = I$, we have that

$$\min_{x^*\in X^*}\left\|B_k^{-1}(x_k-x^*)\right\| \leq \min_{x^*\in X^*}\left\|B_0^{-1}(x_0-x^*)\right\| = dist(x_0,X^*), \qquad \forall\ k.$$

We now show that $f(x_k)$ converges to f^*. By using the subgradient inequality, we have for all $x^*\in X^*$ and all k,

$$\begin{aligned}f(x_k)-f^* &\leq g_k'(x_k-x^*)\\ &= (B_k'g_k)'B_k^{-1}(x_k-x^*)\\ &\leq \|B_k'g_k\|\left\|B_k^{-1}(x_k-x^*)\right\|.\end{aligned}$$

Taking the minimum over $x^*\in X^*$ in this inequality and using the preceding relation, we obtain

$$f(x_k)-f^* \leq \|B_k'g_k\|\ dist(x_0,X^*), \qquad \forall\ k. \tag{6.11}$$

Letting $k\to\infty$ in the preceding relation, we see that

$$\limsup_{k\to\infty} f(x_k) - f^* \leq \lim_{k\to\infty}\|B_k'g_k\|\ dist(x_0,X^*).$$

Since $\|B_k' g_k\| \to 0$ by Lemma 6.2(a), where $d_k = g_k$, it follows that

$$\limsup_{k\to\infty} f(x_k) - f^* \leq 0,$$

thus implying that $f(x_k) \to f^*$.

We next prove the given convergence rate estimate. In view of Eq. (6.11), it follows that

$$\min_{0\leq j\leq k} f(x_j) - f^* \leq \min_{0\leq j\leq k} \|B_j' g_j\|\ dist(x_0, X^*), \qquad \forall\ k.$$

By Lemma 6.2(b), where $d_k = g_k$, we have

$$\min_{0\leq j\leq k} \|B_j' g_j\| \leq C\,\frac{\overline{\rho}^{(k+1)/n}}{\underline{\rho}}\sqrt{\frac{(k+1)}{n}\,\frac{(1-\underline{\rho}^2)}{\left(1-\overline{\rho}^{2(k+1)/n}\right)}}, \qquad \forall\ k,$$

which when substituted in the preceding relation gives the desired estimate. **Q.E.D.**

The dilation method (6.1)–(6.4) with the dynamic stepsize rule for known f^* was proposed and analyzed by Shor in [Sho70b] (see also [Sho98], Theorem 50) for a function f satisfying a special growth condition. When f is convex, the growth condition that Shor assumed reduces to the following: there exists an optimal point x^* and positive scalars r and M, with $M > 1$, such that

$$g(x)'(x - x^*) \leq M\big(f(x) - f^*\big), \qquad \forall\ x \ \text{ with } \ \|x - x^*\| \leq r, \tag{6.12}$$

where $g(x)$ is a subgradient of f at x. Furthermore, the initial point x_0 was assumed to lie in the sphere centered at x^* with radius r. Thus, Shor had analyzed only a local behavior of the method. Furthermore, the dilation parameters ρ_k were such that

$$\rho_k = \rho \geq \frac{M-1}{M+1}, \qquad \forall\ k,$$

Therefore, to implement the method, we would need to know the value of M, as well as to choose the initial point x_0 within the distance r from x^* and to make sure that the condition (6.12) holds. This is practically impossible, even for a polyhedral function f.

Nesterov in [Nes84] also analyzed the method (6.1)–(6.4) with the dynamic stepsize rule for known f^*. He assumed that a condition (6.12) is satisfied for some optimal point x^* and all x, which from practical point of view is not any easier to verify than the original condition (6.12). However, some analytical ideas of Nesterov were applicable to the functions satisfying Assumption 6.1, and we have used these ideas in the proof of Prop. 6.1.

When ρ_k is fixed to some positive scalar ρ, then the convergence rate of $\min_{0\leq j\leq k} f(x_j)$ to f^* is at least as fast as

$$\rho^{\frac{k+1}{n}}\sqrt{\frac{k+1}{n}}.$$

Thus, we would like to use the smallest ρ satisfying the condition

$$(1-\rho)\frac{C^2}{\mu^2} \leq \gamma_k(2-\gamma_k), \qquad \forall\ k.$$

It can be seen, that such smallest ρ corresponds to the case where $\gamma_k = 1$, and it is given by

$$\rho = \sqrt{1 - \frac{\mu^2}{C^2}}.$$

Thus, in this case, the estimate of Prop. 6.1 shows that the values $\min_{0 \leq j \leq k} f(x_j)$ converge to f^* at least as fast as

$$\left(1 - \frac{\mu^2}{C^2}\right)^{\frac{k+1}{2n}} \sqrt{\frac{k+1}{n}}.$$

This convergence rate is somewhat slower than geometric.

In the next proposition, under a slightly stronger assumption on the parameters ρ_k, we give some more convergence rate estimates.

Proposition 6.2: Let Assumption 6.1 hold, and let the parameters ρ_k be such that

$$0 < \underline{\rho} \leq \rho_k \leq \overline{\rho} < 1, \qquad \forall\ k,$$

$$(1 - \rho_k^2)\frac{C^2}{\mu^2} \leq \gamma_k(2 - \gamma_k) - c, \qquad \forall\ k,$$

for a positive scalar c. Assume further that the optimal solution set X^* is nonempty, and let $\{x_k\}$ be the sequence generated by dilation method along subgradients and the dynamic stepsize (6.10). Then, the following hold:

(a) We have

$$\liminf_{k \to \infty} \sqrt{k+1}\, \frac{\big(f(x_k) - f^*\big)}{\|B_k' g_k\|} = 0.$$

(b) For a positive scalar ϵ and the smallest positive integer K such that

$$\sum_{k=0}^{K-1} \frac{c\epsilon^2}{\rho_k^2 \|B_k' g_k\|^2} \leq \big(dist(x_0, X^*)\big)^2,$$

we have

$$\min_{0 \leq k \leq K} f(x_k) \leq f^* + \epsilon.$$

Proof: (a) By using Lemma 6.3, with $\omega = f^*$, and the definition of the stepsize α_k, we can see that for all k,

$$\min_{x^* \in X^*} \big\|B_{k+1}^{-1}(x_{k+1} - x^*)\big\|^2 \leq \min_{x^* \in X^*} \big\|B_k^{-1}(x_k - x^*)\big\|^2 + \frac{\big(f(x_k) - f^*\big)^2}{\rho_k^2 \|B_k' g_k\|^2} \left((1 - \rho_k^2)\frac{C^2}{\mu^2} - \gamma_k(2 - \gamma_k)\right).$$

Since by our assumption, we have $(1-\rho_k^2)C^2/\mu^2 - \gamma_k(2-\gamma_k) \leq -c$ for a positive scalar c and all k, it follows that for all k,

$$\min_{x^* \in X^*} \left\| B_{k+1}^{-1}(x_{k+1} - x^*) \right\|^2 \leq \min_{x^* \in X^*} \left\| B_k^{-1}(x_k - x^*) \right\|^2 - c \frac{\big(f(x_k) - f^*\big)^2}{\rho_k^2 \|B_k' g_k\|^2}. \tag{6.13}$$

Therefore, since $\rho_k \leq \overline{\rho}$ for all k, it follows that

$$\sum_{k=0}^{\infty} \frac{\big(f(x_k) - f^*\big)^2}{\|B_k' g_k\|^2} < \infty. \tag{6.14}$$

Suppose now that

$$\liminf_{k \to \infty} \sqrt{k+1}\, \frac{\big(f(x_k) - f^*\big)}{\|B_k' g_k\|} > 0,$$

in which case there exist a positive scalar ε and a nonnegative integer k_0 such that

$$\sqrt{k+1}\, \frac{\big(f(x_k) - f^*\big)}{\|B_k' g_k\|} \geq \varepsilon, \qquad \forall\ k \geq k_0.$$

We then have

$$\sum_{k=0}^{\infty} \frac{\big(f(x_k) - f^*\big)^2}{\|B_k' g_k\|^2} \geq \sum_{k=0}^{\infty} \frac{\varepsilon^2}{k+1} = \infty,$$

contradicting Eq. (6.14).

(b) Suppose, to arrive at a contradiction, that

$$f(x_k) - f^* > \epsilon, \qquad \forall\ k = 0, \ldots, K.$$

By using this inequality, from Eq. (6.13) we obtain for $k = 0, \ldots, K$,

$$\begin{aligned} \min_{x^* \in X^*} \left\| B_{k+1}^{-1}(x_{k+1} - x^*) \right\|^2 &\leq \min_{x^* \in X^*} \left\| B_k^{-1}(x_k - x^*) \right\|^2 - \frac{c\,\epsilon^2}{\rho_k^2 \|B_k' g_k\|^2} \\ &\leq \min_{x^* \in X^*} \left\| B_0^{-1}(x_0 - x^*) \right\|^2 - \sum_{j=0}^{k} \frac{c\,\epsilon^2}{\rho_j^2 \|B_j' g_j\|^2}. \end{aligned}$$

Therefore, for $k = K-2$, since $B_0 = I$, it follows that

$$\sum_{j=0}^{K-2} \frac{c\,\epsilon^2}{\rho_j^2 \|B_j' g_j\|^2} \leq \min_{x^* \in X^*} \left\| B_0^{-1}(x_0 - x^*) \right\|^2 = \big(dist(x_0, X^*)\big)^2,$$

contradicting the definition of K. **Q.E.D.**

6.4.2 Dynamic Stepsize Rule for Unknown f^*

We here consider the dynamic stepsize rule with unknown f^*, where

$$\alpha_k = \gamma_k \frac{f(x_k) - f_k^{\rm lev}}{\|B_k' g_k\|}, \qquad 0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2. \tag{6.15}$$

The estimates $f_k^{\rm lev}$ have the form

$$f_k^{\rm lev} = \min_{0 \leq j \leq k} f(x_j) - \delta_k, \tag{6.16}$$

where the positive scalars δ_k are updated according to the following rule

$$\delta_{k+1} = \begin{cases} \lambda \delta_k & \text{if } f(x_{k+1}) \leq f_k^{\rm lev}, \\ \max\{\beta \delta_k, \delta\} & \text{if } f(x_{k+1}) > f_k^{\rm lev}. \end{cases} \tag{6.17}$$

The scalars δ_0, δ, β, and λ are positive, with $\beta < 1$ and $\rho \geq 1$.

For the method using the stepsize (6.15)–(6.17), we have the following result.

Proposition 6.3: Let Assumption 6.1 hold, and let the parameters ρ_k be such that

$$0 < \underline{\rho} \leq \rho_k \leq \overline{\rho} < 1, \qquad \forall\ k,$$

$$2(1 - \rho_k^2)\frac{C^2}{\mu^2} \leq \gamma_k(2 - \gamma_k), \qquad \forall\ k,$$

Then, for the sequence $\{x_k\}$ generated by dilation method along subgradients with the dynamic stepsize (6.15)–(6.17), we have:

(a) If $f^* = -\infty$, then

$$\inf_{k \geq 0} f(x_k) = f^*.$$

(b) If $f^* > -\infty$, then

$$\inf_{k \geq 0} f(x_k) \leq f^* + \delta.$$

Proof: We prove (a) and (b) simultaneously. To obtain a contradiction, assume that

$$\inf_{k \geq 0} f(x_k) - \delta > f^*, \tag{6.18}$$

Each time the target level is attained [i.e., $f(x_k) \leq f_{k-1}^{\rm lev}$], the current best function value $\min_{0 \leq j \leq k} f(x_j)$ decreases by at least δ [cf. Eqs. (6.16) and (6.17)], so in view of Eq. (6.18), the target value can be attained only a finite number times. From Eq. (6.17) it follows that after

finitely many iterations, δ_k is decreased to the threshold value and remains at that value for all subsequent iterations, i.e., there is a $\hat{k}$ such that

$$\delta_k = \delta, \qquad \forall\ k \geq \hat{k}.$$

Let the scalar ω be given by

$$\omega = \inf_{k \geq 0} f(x_k) - \delta.$$

Using the preceding two relations and the definition of $f_k^{\rm lev}$ [cf. Eq. (6.16)], we have

$$f_k^{\rm lev} = \min_{0 \leq j \leq k} f(x_j) - \delta \geq \omega, \qquad \forall\ k \geq \hat{k}, \tag{6.19}$$

$$f_k^{\rm lev} \downarrow \omega, \quad \text{as} \quad k \to \infty. \tag{6.20}$$

By Eq. (6.19), it follows that $f(x_k) \geq \omega$ for all $k \geq \hat{k}$. Furthermore, the level set $L_\omega = \{y \in \Re^n \mid f(y) \leq \omega\}$ is nonempty since $\omega > f^*$ [cf. Eq. (6.18)]. Thus, by Lemma 6.3 and the definition of the stepsize α_k [cf. Eq. (6.15)], we obtain for all $k \geq \hat{k}$,

$$\begin{aligned}\min_{y \in L_\omega} \left\| B_{k+1}^{-1}(x_{k+1} - y) \right\|^2 \leq{}& \min_{y \in L_\omega} \left\| B_k^{-1}(x_k - y) \right\|^2 \\ &+ \frac{(1-\rho_k^2)}{\rho_k^2} \frac{C^2}{\mu^2} \frac{\big(f(x_k) - \omega\big)^2}{\|B_k' g_k\|^2} \\ &- 2\frac{\gamma_k}{\rho_k^2} \frac{\big(f(x_k) - f_k^{\rm lev}\big)\big(f(x_k) - \omega\big)}{\|B_k' g_k\|^2} + \frac{\gamma_k^2}{\rho_k^2} \frac{\big(f(x_k) - f_k^{\rm lev}\big)^2}{\|B_k' g_k\|^2}.\end{aligned}$$

By using the estimate

$$\big(f(x_k) - \omega\big)^2 \leq 2\big(f(x_k) - f_k^{\rm lev}\big)^2 + 2\big(f_k^{\rm lev} - \omega\big)^2,$$

and by writing

$$f(x_k) - \omega = \big(f(x_k) - f_k^{\rm lev}\big) + \big(f_k^{\rm lev} - \omega\big),$$

we see that for all $k \geq \hat{k}$,

$$\begin{aligned}\min_{y \in L_\omega} \left\| B_{k+1}^{-1}(x_{k+1} - y) \right\|^2 \leq{}& \min_{y \in L_\omega} \left\| B_k^{-1}(x_k - y) \right\|^2 \\ &+ \left(2(1-\rho_k^2)\frac{C^2}{\mu^2} - 2\gamma_k + \gamma_k^2 \right) \frac{\big(f(x_k) - f_k^{\rm lev}\big)^2}{\rho_k^2 \|B_k' g_k\|^2} \\ &+ 2\frac{(1-\rho_k^2)}{\rho_k^2} \frac{C^2}{\mu^2} \frac{\big(f_k^{\rm lev} - \omega\big)^2}{\|B_k' g_k\|^2} - 2\frac{\gamma_k}{\rho_k^2} \frac{\big(f(x_k) - f_k^{\rm lev}\big)\big(f_k^{\rm lev} - \omega\big)}{\|B_k' g_k\|^2}.\end{aligned}$$

By our assumption, we have

$$2(1-\rho_k^2)\frac{C^2}{\mu^2} \leq 2\gamma_k + \gamma_k^2, \qquad \forall\ k,$$

which implies that

$$2(1-\rho_k^2)\frac{C^2}{\mu^2} \leq 2\gamma_k, \qquad \forall\ k.$$

Using these two inequalities in the preceding relation, we obtain for all $k \geq \hat{k}$,

$$\min_{y\in L_\omega} \left\|B_{k+1}^{-1}(x_{k+1}-y)\right\|^2 \leq \min_{y\in L_\omega} \left\|B_k^{-1}(x_k-y)\right\|^2 + 2\frac{\gamma_k}{\rho_k^2}\frac{\left(f_k^{\rm lev}-\omega\right)}{\|B_k' g_k\|^2}\Big(\left(f_k^{\rm lev}-\omega\right)-\left(f(x_k)-f_k^{\rm lev}\right)\Big).$$

Since $f_k^{\rm lev}$ monotonically decreases to ω as $k\to\infty$ [cf. Eq. (6.20)], without loss of generality, we may assume that $\hat{k}$ is large enough so that $f_k^{\rm lev}-\omega \leq \delta$ for all $k\geq\hat{k}$, which combined with the fact $f(x_k)-f_k^{\rm lev}\geq\delta$ for all k [cf. Eqs. (6.16) and (6.17)], yields

$$\left(f_k^{\rm lev}-\omega\right)-\left(f(x_k)-f_k^{\rm lev}\right)\leq 0, \qquad \forall\ k\geq\hat{k}.$$

Furthermore, since $f_k^{\rm lev}\geq\omega$ for all $k\geq\hat{k}$ [cf. Eq. (6.19)], we see that

$$2\frac{\gamma_k}{\rho_k^2}\frac{\left(f_k^{\rm lev}-\omega\right)}{\|B_k' g_k\|^2}\Big(\left(f_k^{\rm lev}-\omega\right)-\left(f(x_k)-f_k^{\rm lev}\right)\Big)\leq 0, \qquad \forall\ k\geq\hat{k},$$

implying that

$$\min_{y\in L_\omega} \left\|B_{k+1}^{-1}(x_{k+1}-y)\right\|^2 \leq \min_{y\in L_\omega} \left\|B_k^{-1}(x_k-y)\right\|^2, \qquad \forall\ k\geq\hat{k}.$$

Therefore, the sequence $\left\{\min_{y\in L_\omega}\left\|B_k^{-1}(x_k-y)\right\|\right\}$ is bounded.

By the subgradient inequality, we have for all k and all vectors y in the level set L_ω,

$$f(x_k)-\omega \leq f(x_k)-f(y)\leq g_k'(x_k-y).$$

By writing $g_k'(x_k-y)=(B_k'g_k)'B_k^{-1}(x_k-y)$ and by using Schwartz inequality, we see that

$$f(x_k)-\omega\leq\|B_k'g_k\|\left\|B_k^{-1}(x_k-y)\right\|, \qquad \forall\ k, \quad \forall\ y\in L_\omega.$$

Taking the minimum over $y\in L_\omega$ in this relation, we obtain

$$f(x_k)-\omega\leq\|B_k'g_k\|\min_{y\in L_\omega}\left\|B_k^{-1}(x_k-y)\right\|, \qquad \forall\ k.$$

By Lemma 6.2(a), where $d_k = g_k$, we have that $\|B'_k g_k\| \to 0$, so that by letting $k \to \infty$ in the preceding relation and by taking into the account that $\min_{y \in L_\omega} \left\|B_k^{-1}(x_k - y)\right\|$ is bounded, we see that

$$\limsup_{k \to \infty} f(x_k) - \omega \leq 0,$$

which is impossible because

$$\inf_{k \geq 0} f(x_k) - \omega = \delta > 0.$$

Q.E.D.

Since the stepsize rule considered in Prop. 6.3 is new, accordingly, the convergence result established in this proposition is new. Furthermore, this result is the first to show the convergence of the dilation method with a dynamic stepsize rule using estimates of the optimal function value f^* instead of the exact value f^*. In practice, it is not typical that we know the optimal function value f^*, so that the dynamic stepsize rule (6.15)–(6.17) is more practical than the dynamic stepsize rule with known f^*.

On a technical side, let us note that by taking a closer look at the proof of Prop. 6.3, we can see that the result of this proposition holds if

$$2(1 - \rho_k^2)\frac{C^2}{\mu^2} - \gamma_k(2 - \gamma_k) \leq 0,$$

for all sufficiently large k (instead of all k).

6.5. DILATION ALONG OTHER DIRECTIONS

In this section, we discuss the method that uses dilations along directions that may differ from subgradient directions. The method is similar to the method (6.1)–(6.4), with a major difference being the update formula for x_{k+1} and the choice of dilation direction ξ_k. In particular, at a typical iteration, we have the current iterate x_k, a subgradient g_k of f at x_k, and a matrix B_k. We compute B_{k+1} as follows:

$$B_{k+1} = B_k R_{\rho_k}(\xi_k), \tag{6.21}$$

with $B_0 = I$, and

$$R_{\rho_k}(\xi_k) = I + (\rho_k - 1)\xi_k \xi'_k, \tag{6.22}$$

$$\xi_k = \frac{B'_k d_k}{\|B'_k d_k\|}, \tag{6.23}$$

for a positive scalar ρ_k and a nonzero vector d_k. We next compute x_{k+1} according to the following rule

$$x_{k+1} = x_k - \alpha_k \frac{B_{k+1} B'_{k+1} g_k}{\|B'_{k+1} g_k\|}. \tag{6.24}$$

We refer to this method as *dilation method*. This method is rather general, and its interpretation depends on the choice for the directions d_k. For example, the method with $d_k = g_k$ can be related to Shor's method of Section 6.1. In particular, Shor's method can be viewed as a "delayed" dilation method (6.21)–(6.24), where $d_k = g_k$. To see this, note that in the dilation method, we first update B_{k+1}, and then we use it to compute x_{k+1}. In Shor's method, however, we first compute x_{k+1} using the transformation B_k from the preceding iteration, and then we update B_{k+1}. Thus, the use of B_{k+1} is delayed until the next iteration.

The dilation method along the difference of two successive subgradients (i.e., $d_k = g_k - g_{k-1}$), also known as r-algorithm, was proposed and analyzed by Shor and Zhurbenko [ShZ71] (see also, Shor [Sho85] and [Sho98]).

Before closing this section, let us mention that we can write the formulas (6.21)–(6.24) in an alternative form by introducing the matrix $H_k = B_k B_k'$. In this case, it can be seen that

$$H_{k+1} = H_k + (\rho_k^2 - 1)\frac{H_k d_k d_k' H_k}{d_k' H_k d_k},$$

$$x_{k+1} = x_k - \alpha_k \frac{H_{k+1} g_k}{\sqrt{g_k' H_{k+1} g_k}}.$$

These formulas are computationally more efficient than Eqs. (6.21)–(6.24). However, the update formula for H_{k+1} is more sensitive with respect to computational errors than the update formula for B_{k+1}.

6.6. ASSUMPTIONS AND SOME BASIC RELATIONS

We here give the assumption and the basic relation that we use throughout our analysis. In particular, the assumption that we use is the following:

Assumption 6.2:

(a) The sequence $\{d_k\}$ is bounded, i.e., there exist a positive scalar C such that

$$\|d_k\| \leq C, \qquad \forall\ k.$$

Furthermore, the directions d_k are such that

$$\|B_k' d_k\| \geq \|B_k' g_k\|, \qquad \forall\ k.$$

(b) There exists a positive scalar μ such that for every nonempty level set $\{y \in \Re^n \mid f(y) \leq \omega\}$ and every vector x with $f(x) \geq \omega$, we have

$$f(x) - \omega \geq \mu\ dist(x, L_\omega),$$

where L_ω is the level set $\{y \in \Re^n \mid f(y) \leq \omega\}$.

Assumption 6.2(b) is the same as Assumption 6.1(b), which we repeated here for an easier reference. When $d_k = g_k$, then Assumption 6.2 coincides with Assumption 6.1, but this case is not very interesting, since as discussed earlier, the dilation method with $d_k = g_k$ can be related to the dilation method (6.1)–(6.4) of Section 6.1.

We now discuss the possibility to use the direction of the difference of two successive subgradients, i.e., $d_k = g_k - g_{k-1}$. This choice for directions d_k is motivated by the instances where the angle formed by subgradients g_{k-1} and g_k is too wide, which can be an indication that the subgradient g_k is almost orthogonal to directions pointing toward the set of minima. In this case, to obtain the subgradient directions better pointed toward the set of minima, we can think of reducing the angle formed by g_{k-1} and g_k by applying a space contraction along the difference of the vectors g_k and g_{k-1}. By combining this idea with the idea of changing the metric, we can consider the directions d_k of the following form

$$d_k = \begin{cases} g_k - g_{k-1} & \text{if } g_k' B_k B_k' g_{k-1} < 0, \\ g_k & \text{otherwise.} \end{cases}$$

Let us verify that the conditions of Assumption 6.2(a) are satisfied for this choice. If the subgradients g_k are bounded, then

$$\|d_k\| \leq \|g_k\| + \|g_{k-1}\|, \qquad \forall\ k,$$

thus showing that the vectors d_k are also bounded. Hence, in this case, the first condition of Assumption 6.2(a) is satisfied with $C = 2\max_k \|g_k\|$. Furthermore, for the case where $g_k' B_k B_k' g_{k-1} < 0$, we have

$$\|B_k' d_k\|^2 = \|B_k' g_k\|^2 - 2 g_k' B_k B_k' g_{k-1} + \|B_k' g_{k-1}\|^2 \geq \|B_k' g_k\|^2.$$

Therefore, by the definition of d_k, it follows that

$$\|B_k' d_k\| \geq \|B_k' g_k\|, \qquad \forall\ k.$$

The method using directions d_k that we just discussed is not the same as the dilation method using $d_k = g_k - g_{k-1}$ in all iterations, which was proposed and analyzed by Shor and Zhurbenko [ShZ71] (see also, Shor [Sho85] and [Sho98]). Since, for our analysis, the assumption Assumption 6.2(a) is essential, our results do not apply to the method that uses $d_k = g_k - g_{k-1}$ in all iterations.

From now on, our focus is on the method (6.21)–(6.24) with directions d_k satisfying the requirements of Assumption 6.2(a). In our next lemma, we establish a basic relation for the iterates generated by the method. This relation will be repeatedly invoked in our convergence analysis.

Lemma 6.4: Let Assumption 6.2 hold, and let the parameters ρ_k be such that

$$0 < \rho_k \leq 1, \qquad \forall\ k.$$

Let further $\{x_k\}$ be the sequence generated by the dilation method. Then, for any k, and any scalar ω such that $f(x_k) \geq \omega$ and the level set $L_\omega = \{y \in \Re^n \mid f(y) \leq \omega\}$ is nonempty, we have

$$\min_{y \in L_\omega} \left\| B_{k+1}^{-1}(x_{k+1} - y) \right\|^2 \leq \min_{y \in L_\omega} \left\| B_k^{-1}(x_k - y) \right\|^2 + \frac{(1-\rho_k^2)}{\rho_k^2} \frac{C^2}{\mu^2} \frac{\big(f(x_k) - \omega\big)^2}{\|B'_{k+1} g_k\|^2} - 2\alpha_k \frac{\big(f(x_k) - \omega\big)}{\|B'_{k+1} g_k\|} + \alpha_k^2,$$

where C and μ are as in Assumption 6.2.

Proof: Let k be arbitrary but fixed. By the definition of x_{k+1} [cf. Eq. (6.24)], we have for any y,

$$\begin{aligned} \left\| B_{k+1}^{-1}(x_{k+1} - y) \right\|^2 &= \left\| B_{k+1}^{-1}(x_k - y) - \alpha_k \frac{B'_{k+1} g_k}{\|B'_{k+1} g_k\|} \right\|^2 \\ &= \left\| B_{k+1}^{-1}(x_k - y) \right\|^2 - 2\alpha_k \frac{g_k'(x_k - y)}{\|B'_{k+1} g_k\|} + \alpha_k^2. \end{aligned}$$

Furthermore, by using the subgradient inequality

$$f(x_k) - f(y) \leq g_k'(x_k - y), \qquad \forall\ y,$$

we obtain for all y,

$$\left\| B_{k+1}^{-1}(x_{k+1} - y) \right\|^2 \leq \left\| B_{k+1}^{-1}(x_k - y) \right\|^2 - 2\alpha_k \frac{\big(f(x_k) - f(y)\big)}{\|B'_{k+1} g_k\|} + \alpha_k^2. \tag{6.25}$$

We next estimate the term $\left\| B_{k+1}^{-1}(x_k - y) \right\|^2$. By the definition of B_{k+1} [cf. Eq. (6.21)], we have

$$B_{k+1}^{-1} = R_{\rho_k}^{-1}(\xi_k) B_k^{-1}.$$

Since by part (a) of Lemma 6.1, we have $\big(R_{\rho_k}(\xi_k)\big)^{-1} = R_{1/\rho_k}(\xi_k)$, and by part (b) of the same lemma, we have

$$\left\| R_\rho(\xi) x \right\|^2 = \|x\|^2 + (\rho^2 - 1)(\xi' x)^2, \qquad \forall\ x,$$

we obtain for any y,

$$\begin{aligned} \left\| B_{k+1}^{-1}(x_k - y) \right\|^2 &= \left\| B_k^{-1}(x_k - y) \right\|^2 + \left(\frac{1}{\rho_k^2} - 1 \right) \big(\xi_k' B_k^{-1}(x_k - y) \big)^2 \\ &= \left\| B_k^{-1}(x_k - y) \right\|^2 + \frac{(1-\rho_k^2)}{\rho_k^2} \frac{\big(d_k'(x_k - y)\big)^2}{\|B_k' d_k\|^2}. \end{aligned}$$

The last equality in the preceding relation follows from $\xi_k = B_k' d_k / \|B_k' d_k\|$ [cf. Eq. (6.23)]. By Schwartz inequality and assumption that $\|d_k\| \leq C$ [cf. Assumption 6.2(a)], it follows that

$$\big(d_k'(x_k - y)\big)^2 \leq \|d_k\|^2 \|x_k - y\|^2 \leq C^2 \|x_k - y\|^2.$$

Furthermore, since $0 < \rho_k \leq 1$, by Lemma 6.1(c), we have $\big\|R_{\rho_k}'(\xi_k)\big\| = 1$. This relation, together with the definition of B_{k+1} [cf. Eq. (6.21)] and our assumption that $\|B_k' d_k\| \geq \|B_k' g_k\|$, yields

$$\|B_k' d_k\| \geq \|B_k' g_k\| = \big\|R_{\rho_k}'(\xi_k)\big\| \|B_k' g_k\| \geq \big\|R_{\rho_k}'(\xi_k) B_k' g_k\big\| = \|B_{k+1}' g_k\|.$$

From the preceding three relations, we obtain

$$\big\|B_{k+1}^{-1}(x_k - y)\big\|^2 \leq \big\|B_k^{-1}(x_k - y)\big\|^2 + \frac{(1 - \rho_k^2)}{\rho_k^2} \frac{C^2 \|x_k - y\|^2}{\|B_{k+1}' g_k\|^2}.$$

By substituting the preceding relation in Eq. (6.25), we see that for all y,

$$\big\|B_{k+1}^{-1}(x_{k+1} - y)\big\|^2 \leq \big\|B_k^{-1}(x_k - y)\big\|^2 + \frac{(1 - \rho_k^2)}{\rho_k^2} \frac{C^2 \|x_k - y\|^2}{\|B_{k+1}' g_k\|^2} - 2\alpha_k \frac{\big(f(x_k) - f(y)\big)}{\|B_{k+1}' g_k\|} + \alpha_k^2.$$

Let ω be a scalar such that $f(x_k) \geq \omega$ and the level set $L_\omega = \big\{y \in \Re^n \mid f(y) \leq \omega\big\}$ is nonempty. Then, for all $y \in L_\omega$, we have

$$f(x_k) - f(y) \geq f(x_k) - \omega.$$

By first using this inequality in the preceding relation, and then taking the minimum over all $y \in L_\omega$, we obtain

$$\min_{y \in L_\omega} \big\|B_{k+1}^{-1}(x_{k+1} - y)\big\|^2 \leq \min_{y \in L_\omega} \big\|B_k^{-1}(x_k - y)\big\|^2 + \frac{(1 - \rho_k^2)}{\rho_k^2} \frac{C^2 \big(dist(x_k, L_\omega)\big)^2}{\|B_{k+1}' g_k\|^2} - 2\alpha_k \frac{\big(f(x_k) - \omega\big)}{\|B_{k+1}' g_k\|} + \alpha_k^2.$$

Finally, since by Assumption 6.2(b), we have

$$dist(x_k, L_\omega) \leq \frac{1}{\mu}\big(f(x_k) - \omega\big),$$

which when substituted in the preceding inequality gives the desired relation. **Q.E.D.**

6.7. CONVERGENCE PROPERTIES OF THE METHOD WITH DILATION ALONG OTHER DIRECTIONS

In our analysis here, we use the same idea as in Section 6.4 with some technical adjustments. In particular, by using the subgradient inequality and by changing the space coordinates, we can see that for an optimal solution x^*,

$$f(x_k) - f^* \leq g_k'(x_k - x^*) \leq \|B_{k+1}' g_k\| \left\| B_{k+1}^{-1}(x_k - x^*) \right\|, \qquad \forall\ k.$$

Then, under some conditions, Lemma 6.2(a) will easily yield convergence of vectors $B_{k+1}' g_k$ to zero. After that, the main analytical effort is to show that the distance $\left\| B_{k+1}^{-1}(x_k - x^*) \right\|$ is bounded, in which Lemma 6.4 will play a crucial role.

6.7.1 Dynamic Stepsize Rule for Known f^*

We here give convergence and convergence rate results for the dilation method using a dynamic stepsize rule for known f^*,

$$\alpha_k = \gamma_k \frac{f(x_k) - f^*}{\|B_{k+1}' g_k\|}, \qquad 0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2, \quad \forall\ k. \tag{6.26}$$

Our first result shows that the function values $f(x_k)$ converge to the optimal function value f^*, and that the values $\min_{0 \leq j \leq k} f(x_j)$ converge to f^* as fast as $\overline{\rho}^{\,(k+1)/n} \sqrt{(k+1)/n}$.

Proposition 6.4: Let Assumption 6.2 hold, and let the parameters ρ_k be such that

$$0 < \underline{\rho} \leq \rho_k \leq \overline{\rho} < 1, \qquad \forall\ k,$$

$$\frac{(1 - \rho_k^2)}{\rho_k^2} \frac{C^2}{\mu^2} \leq \gamma_k (2 - \gamma_k), \qquad \forall\ k.$$

Assume further that the optimal solution set X^* is nonempty. Then, for the sequence $\{x_k\}$ generated by the dilation method and the dynamic stepsize (6.26), we have

$$\lim_{k \to \infty} f(x_k) = f^*.$$

Furthermore,

$$\min_{0 \leq j \leq k} f(x_j) \leq f^* + C\, \frac{\overline{\rho}^{\,(k+1)/n}}{\underline{\rho}^2} \sqrt{\frac{(k+1)}{n} \frac{(1 - \underline{\rho}^2)}{\left(1 - \overline{\rho}^{\,2(k+1)/n}\right)}}\, dist(x_0, X^*).$$

Proof: We first show that the distances $\min_{x^* \in X^*} \left\| B_k^{-1}(x_k - x^*) \right\|$ are nonincreasing. From Lemma 6.4, where $\omega = f^*$, we see that for all k,

$$\min_{x^* \in X^*} \left\| B_{k+1}^{-1}(x_{k+1} - x^*) \right\|^2 \leq \min_{x^* \in X^*} \left\| B_k^{-1}(x_k - x^*) \right\|^2 + \frac{(1 - \rho_k^2)}{\rho_k^2} \frac{C^2}{\mu^2} \frac{\left(f(x_k) - f^*\right)^2}{\|B_{k+1}' g_k\|^2}$$
$$- 2\alpha_k \frac{\left(f(x_k) - f^*\right)}{\|B_{k+1}' g_k\|} + \alpha_k^2.$$

By using the definition of the stepsize α_k [cf. Eq. (6.26)], we obtain for all k,

$$\min_{x^* \in X^*} \left\| B_{k+1}^{-1}(x_{k+1} - x^*) \right\|^2 \leq \min_{x^* \in X^*} \left\| B_k^{-1}(x_k - x^*) \right\|^2 + \frac{\left(f(x_k) - f^*\right)^2}{\|B'_{k+1} g_k\|^2} \left(\frac{(1-\rho_k^2)}{\rho_k^2} \frac{C^2}{\mu^2} - \gamma_k(2-\gamma_k) \right).$$

By our assumption that

$$\frac{(1-\rho_k^2)}{\rho_k^2} \frac{C^2}{\mu^2} \leq \gamma_k(2-\gamma_k), \qquad \forall\ k,$$

it follows that the distances $\min_{x^* \in X^*} \left\| B_k^{-1}(x_k - x^*) \right\|$ are nonincreasing, and since $B_0 = I$, we have that

$$\min_{x^* \in X^*} \left\| B_k^{-1}(x_k - x^*) \right\| \leq \min_{x^* \in X^*} \left\| B_0^{-1}(x_0 - x^*) \right\| = dist(x_0, X^*), \qquad \forall\ k. \tag{6.27}$$

We now prove that $f(x_k) \to f^*$. By using the subgradient inequality, we have for all $x^* \in X^*$ and all k,

$$f(x_k) - f^* \leq g_k'(x_k - x^*) = (B'_{k+1} g_k)' B_{k+1}^{-1}(x_k - x^*) \leq \|B'_{k+1} g_k\| \left\| B_{k+1}^{-1}(x_k - x^*) \right\|. \tag{6.28}$$

Since $0 < \underline{\rho} \leq \rho_k \leq \overline{\rho} < 1$ and $B_{k+1} = B_k R_{\rho_k}(\xi_k)$, by using Lemma 6.1, it can be seen that

$$\left\| B_{k+1}^{-1}(x_k - y) \right\| \leq \left\| R_{\rho_k}^{-1}(\xi_k) \right\| \left\| B_k^{-1}(x_k - y) \right\| \leq \frac{1}{\underline{\rho}} \left\| B_k^{-1}(x_k - y) \right\|, \qquad \forall\ k, \tag{6.29}$$

$$\|B'_{k+1} g_k\| \leq \|R'_{\rho_k}(\xi_k)\| \|B'_k g_k\| \leq \|B'_k g_k\|, \qquad \forall\ k.$$

Furthermore, because by Assumption 6.2, we have $\|B'_k g_k\| \leq \|B'_k d_k\|$ for all k, from the preceding relation it follows that

$$\|B'_{k+1} g_k\| \leq \|B'_k d_k\|, \qquad \forall\ k. \tag{6.30}$$

By combining Eqs. (6.28)–(6.30), we obtain for all $x^* \in X^*$,

$$f(x_k) - f^* \leq \frac{1}{\underline{\rho}} \|B'_k d_k\| \left\| B_k^{-1}(x_k - y) \right\|, \qquad \forall\ k.$$

Taking the minimum over $x^* \in X^*$ in this inequality and using the relation (6.27), we obtain

$$f(x_k) - f^* \leq \frac{1}{\underline{\rho}} \|B'_k d_k\|\, dist(x_0, X^*), \qquad \forall\ k. \tag{6.31}$$

Since $\|B'_k d_k\| \to 0$ by Lemma 6.2(a), it follows that

$$\limsup_{k \to \infty} f(x_k) - f^* \leq \frac{1}{\underline{\rho}} \lim_{k \to \infty} \|B'_k d_k\|\, dist(x_0, X^*) = 0,$$

thus implying that $f(x_k) \to f^*$.

We next show the given convergence rate estimate. In view of Eq. (6.31), it follows that

$$\min_{0\le j\le k} f(x_j) - f^* \le \frac{1}{\underline{\rho}} \min_{0\le j\le k} \|B_j' d_j\|\, dist(x_0, X^*), \qquad \forall\ k.$$

By Lemma 6.2(b), we have

$$\min_{0\le j\le k} \|B_j' d_j\| \le C\, \frac{\overline{\rho}^{(k+1)/n}}{\underline{\rho}} \sqrt{\frac{(k+1)}{n} \frac{(1-\underline{\rho}^2)}{\left(1-\overline{\rho}^{2(k+1)/n}\right)}}, \qquad \forall\ k,$$

which when substituted in the preceding relation gives the desired estimate. **Q.E.D.**

Under a slightly more restrictive condition on the parameters ρ_k, we can give convergence rate estimates that do not depend on n, as seen in the following proposition.

Proposition 6.5: Let Assumption 6.2 hold, and let the parameters ρ_k be such that

$$0 < \underline{\rho} \le \rho_k \le \overline{\rho} < 1, \qquad \forall\ k,$$

$$\frac{(1-\rho_k^2)}{\rho_k^2} \frac{C^2}{\mu^2} \le \gamma_k(2-\gamma_k) - c, \qquad \forall\ k,$$

for a positive scalar c. Assume further that the optimal solution set X^* is nonempty, and let $\{x_k\}$ be the sequence generated by the dilation method and the dynamic stepsize (6.26). Then, the following hold:

(a) We have

$$\liminf_{k\to\infty} \sqrt{k+1}\, \big(f(x_k) - f^*\big) = 0.$$

(b) For a positive scalar ϵ and the nonnegative integer K given by

$$K = \left\lfloor \frac{C^2}{c\,\epsilon^2} \big(dist(x_0, X^*)\big)^2 \right\rfloor,$$

with C being an upper bound on the norms $\|d_k\|$, we have

$$\min_{0\le k\le K} f(x_k) \le f^* + \epsilon.$$

Proof: (a) By using Lemma 6.4, with $\omega = f^*$, and the definition of the stepsize α_k, we can see that for all k,

$$\min_{x^*\in X^*} \big\|B_{k+1}^{-1}(x_{k+1} - x^*)\big\|^2 \le \min_{x^*\in X^*} \big\|B_k^{-1}(x_k - x^*)\big\|^2 + \frac{\big(f(x_k) - f^*\big)^2}{\|B_{k+1}' g_k\|^2} \left(\frac{(1-\rho_k^2)}{\rho_k^2} \frac{C^2}{\mu^2} - \gamma_k(2-\gamma_k) \right).$$

Since by assumption, we have $(1-\rho_k^2)C^2/(\rho_k^2\mu^2) \leq \gamma_k(2-\gamma_k) - c$ for a positive scalar c and all k, it follows that

$$\min_{x^* \in X^*} \left\| B_{k+1}^{-1}(x_{k+1} - x^*) \right\|^2 \leq \min_{x^* \in X^*} \left\| B_k^{-1}(x_k - x^*) \right\|^2 - c \frac{\big(f(x_k) - f^*\big)^2}{\|B'_{k+1} g_k\|^2}, \qquad \forall\ k.$$

By using Assumption 6.2(a) and Lemma 6.1(c), we can see that

$$\|B'_{k+1} g_k\| \leq \|R'_{\rho_k}(\xi_k)\| \|B'_k g_k\| \leq \|B'_k g_k\| \leq \|B'_k d_k\| \leq \|B_k\| \|d_k\| \leq C, \qquad \forall\ k.$$

Hence,

$$\min_{x^* \in X^*} \left\| B_{k+1}^{-1}(x_{k+1} - x^*) \right\|^2 \leq \min_{x^* \in X^*} \left\| B_k^{-1}(x_k - x^*) \right\|^2 - c\, \frac{\big(f(x_k) - f^*\big)^2}{C^2}, \qquad \forall\ k. \tag{6.32}$$

implying that

$$\sum_{k=0}^{\infty} \big(f(x_k) - f^*\big)^2 < \infty \tag{6.33}$$

Suppose that

$$\liminf_{k \to \infty} \sqrt{k+1}\, \big(f(x_k) - f^*\big) > 0,$$

in which case there exist a positive scalar ε and a nonnegative integer k_0 such that

$$\sqrt{k+1}\, \big(f(x_k) - f^*\big) \geq \varepsilon, \qquad \forall\ k \geq k_0.$$

We then have

$$\sum_{k=0}^{\infty} \big(f(x_k) - f^*\big)^2 \geq \sum_{k=0}^{\infty} \frac{\varepsilon^2}{k+1} = \infty,$$

contradicting Eq. (6.33).

(b) Suppose, to arrive at a contradiction, that

$$f(x_k) - f^* > \epsilon, \qquad \forall\ k = 0, \ldots, K.$$

By using this inequality, from Eq. (6.13) we obtain for $k = 0, \ldots, K$,

$$\begin{aligned} \min_{x^* \in X^*} \left\| B_{k+1}^{-1}(x_{k+1} - x^*) \right\|^2 &\leq \min_{x^* \in X^*} \left\| B_k^{-1}(x_k - x^*) \right\|^2 - \frac{c\,\epsilon^2}{C^2} \\ &\leq \min_{x^* \in X^*} \left\| B_0^{-1}(x_0 - x^*) \right\|^2 - (k+1)\frac{c\,\epsilon^2}{C^2}. \end{aligned}$$

Therefore, for $k = K$, since $B_0 = I$, it follows that

$$(K+1)\frac{c\,\epsilon^2}{C^2} \leq \min_{x^* \in X^*} \left\| B_0^{-1}(x_0 - x^*) \right\|^2 = \big(dist(x_0, X^*)\big)^2,$$

contradicting the definition of K. **Q.E.D.**

6.7.2 Dynamic Stepsize Rule for Unknown f^*

We here consider the dynamic stepsize rule for unknown f^*, where

$$\alpha_k = \gamma_k \frac{f(x_k) - f_k^{\rm lev}}{\|B'_{k+1} g_k\|}, \qquad 0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2. \tag{6.34}$$

The estimates $f_k^{\rm lev}$ are given by

$$f_k^{\rm lev} = \min_{0 \leq j \leq k} f(x_j) - \delta_k, \tag{6.35}$$

where the positive scalars δ_k are updated according to the following rule

$$\delta_{k+1} = \begin{cases} \lambda \delta_k & \text{if } f(x_{k+1}) \leq f_k^{\rm lev}, \\ \max\{\beta \delta_k, \delta\} & \text{if } f(x_{k+1}) > f_k^{\rm lev}, \end{cases} \tag{6.36}$$

where δ_0, δ, β, and λ are fixed positive scalars with $\beta < 1$ and $\lambda \geq 1$.

For the method using this stepsize, we have the following result.

Proposition 6.6: Let Assumption 6.2 hold, and let the parameters ρ_k be such that

$$0 < \underline{\rho} \leq \rho_k \leq \overline{\rho} < 1, \qquad \forall\ k,$$

$$2 \frac{(1 - \rho_k^2)}{\rho_k^2} \frac{C^2}{\mu^2} \leq \gamma_k (2 - \gamma_k), \qquad \forall\ k,$$

Then, for the sequence $\{x_k\}$ generated by Shor's method with the dynamic stepsize (6.34)–(6.36), we have:

(a) If $f^* = -\infty$, then

$$\inf_{k \geq 0} f(x_k) = f^*.$$

(b) If $f^* > -\infty$, then

$$\inf_{k \geq 0} f(x_k) \leq f^* + \delta.$$

Proof: We prove (a) and (b) simultaneously. To obtain a contradiction, assume that

$$\inf_{k \geq 0} f(x_k) - \delta > f^*, \tag{6.37}$$

Each time the target level is attained [i.e., $f(x_k) \leq f_{k-1}^{\rm lev}$], the current best function value $\min_{0 \leq j \leq k} f(x_j)$ decreases by at least δ [cf. Eqs. (6.35) and (6.36)], so in view of Eq. (6.37), the target value can be attained only a finite number times. From Eq. (6.36) it follows that after

finitely many iterations, δ_k is decreased to the threshold value and remains at that value for all subsequent iterations, i.e., there is a $\hat{k}$ such that

$$\delta_k = \delta, \qquad \forall\ k \geq \hat{k}.$$

Let the scalar ω be given by

$$\omega = \inf_{k \geq 0} f(x_k) - \delta.$$

Using the preceding two relations and the definition of $f_k^{\rm lev}$ [cf. Eq. (6.35)], we have

$$f_k^{\rm lev} = \min_{0 \leq j \leq k} f(x_j) - \delta \geq \omega, \qquad \forall\ k \geq \hat{k}, \tag{6.38}$$

$$f_k^{\rm lev} \downarrow \omega, \quad \text{as} \quad k \to \infty. \tag{6.39}$$

By Eq. (6.38), it follows that $f(x_k) \geq \omega$ for all $k \geq \hat{k}$. Furthermore, the level set $L_\omega = \{y \in \Re^n \mid f(y) \leq \omega\}$ is nonempty since $\omega > f^*$ [cf. Eq. (6.37)]. Thus, by Lemma 6.4 and the definition of the stepsize α_k [cf. Eq. (6.34)], we obtain for all $k \geq \hat{k}$,

$$\begin{aligned}\min_{y \in L_\omega} \left\|B_{k+1}^{-1}(x_{k+1} - y)\right\|^2 \leq \min_{y \in L_\omega} \left\|B_k^{-1}(x_k - y)\right\|^2 + \frac{(1-\rho_k^2)}{\rho_k^2}\frac{C^2}{\mu^2}\frac{\big(f(x_k) - \omega\big)^2}{\|B'_{k+1}g_k\|^2} \\ - 2\gamma_k \frac{\big(f(x_k) - f_k^{\rm lev}\big)\big(f(x_k) - \omega\big)}{\|B'_{k+1}g_k\|^2} + \gamma_k^2 \frac{\big(f(x_k) - f_k^{\rm lev}\big)^2}{\|B'_{k+1}g_k\|^2}.\end{aligned}$$

By using the estimate

$$\big(f(x_k) - \omega\big)^2 \leq 2\big(f(x_k) - f_k^{\rm lev}\big)^2 + 2\big(f_k^{\rm lev} - \omega\big)^2,$$

and by writing

$$f(x_k) - \omega = \big(f(x_k) - f_k^{\rm lev}\big) + \big(f_k^{\rm lev} - \omega\big),$$

we see that for all $k \geq \hat{k}$,

$$\begin{aligned}\min_{y \in L_\omega} \left\|B_{k+1}^{-1}(x_{k+1} - y)\right\|^2 \leq \min_{y \in L_\omega} \left\|B_k^{-1}(x_k - y)\right\|^2 \\ + \left(2\frac{(1-\rho_k^2)}{\rho_k^2}\frac{C^2}{\mu^2} - 2\gamma_k + \gamma_k^2\right)\frac{\big(f(x_k) - f_k^{\rm lev}\big)^2}{\|B'_{k+1}g_k\|^2} \\ + 2\frac{(1-\rho_k^2)}{\rho_k^2}\frac{C^2}{\mu^2}\frac{\big(f_k^{\rm lev} - \omega\big)^2}{\|B'_{k+1}g_k\|^2} - 2\gamma_k \frac{\big(f(x_k) - f_k^{\rm lev}\big)\big(f_k^{\rm lev} - \omega\big)}{\|B'_{k+1}g_k\|^2}.\end{aligned}$$

By our assumption, we have

$$2\frac{(1-\rho_k^2)}{\rho_k^2}\frac{C^2}{\mu^2} - 2\gamma_k + \gamma_k^2 \leq 0, \qquad \forall\ k,$$

implying that

$$2\frac{(1-\rho_k^2)}{\rho_k^2}\frac{C^2}{\mu^2} \leq 2\gamma_k, \qquad \forall\ k.$$

Using these two inequalities in the preceding relation, we obtain for all $k \geq \hat{k}$,

$$\min_{y\in L_\omega}\left\|B_{k+1}^{-1}(x_{k+1}-y)\right\|^2 \leq \min_{y\in L_\omega}\left\|B_k^{-1}(x_k-y)\right\|^2 + 2\gamma_k\frac{\left(f_k^{\rm lev}-\omega\right)}{\|B'_{k+1}g_k\|^2}\Big(\left(f_k^{\rm lev}-\omega\right)-\left(f(x_k)-f_k^{\rm lev}\right)\Big).$$

Since $f_k^{\rm lev}$ decreases to ω [cf. Eq. (6.39)], without loss of generality, we may assume that $\hat{k}$ is large enough so that $f_k^{\rm lev}-\omega \leq \delta$ for all $k \geq \hat{k}$, which together with the fact $f(x_k)-f_k^{\rm lev} \geq \delta$ for all k [cf. Eqs. (6.35) and (6.36)], yields

$$\left(f_k^{\rm lev}-\omega\right)-\left(f(x_k)-f_k^{\rm lev}\right) \leq 0, \qquad \forall\ k \geq \hat{k}.$$

Furthermore, since $f_k^{\rm lev} \geq \omega$ for all $k \geq \hat{k}$ [cf. Eq. (6.38)], we see that

$$2\gamma_k\frac{\left(f_k^{\rm lev}-\omega\right)}{\|B'_{k+1}g_k\|^2}\Big(\left(f_k^{\rm lev}-\omega\right)-\left(f(x_k)-f_k^{\rm lev}\right)\Big) \leq 0, \qquad \forall\ k \geq \hat{k},$$

implying that

$$\min_{y\in L_\omega}\left\|B_{k+1}^{-1}(x_{k+1}-y)\right\|^2 \leq \min_{y\in L_\omega}\left\|B_k^{-1}(x_k-y)\right\|^2, \qquad \forall\ k \geq \hat{k}.$$

Therefore, the distances $\min_{y\in L_\omega}\left\|B_k^{-1}(x_k-y)\right\|$ are bounded.

By using the subgradient inequality, we have for all k and all vectors $y \in L_\omega$,

$$f(x_k)-\omega \leq f(x_k)-f(y) \leq g'_k(x_k-y) = (B'_{k+1}g_k)'B_{k+1}^{-1}(x_k-y) \leq \|B'_{k+1}g_k\|\left\|B_{k+1}^{-1}(x_k-y)\right\|. \tag{6.40}$$

Since $0 < \underline{\rho} \leq \rho_k \leq \overline{\rho} < 1$ and $B_{k+1} = B_k R_{\rho_k}(\xi_k)$, by Lemma 6.1(c), it can be seen that

$$\left\|B_{k+1}^{-1}(x_k-y)\right\| \leq \left\|R_{\rho_k}(\xi_k)^{-1}\right\|\left\|B_k^{-1}(x_k-y)\right\| \leq \frac{1}{\underline{\rho}}\left\|B_k^{-1}(x_k-y)\right\|, \qquad \forall\ k, \tag{6.41}$$

$$\|B'_{k+1}g_k\| \leq \|R'_{\rho_k}(\xi_k)\|\,\|B'_kg_k\| \leq \|B'_kg_k\|, \qquad \forall\ k.$$

Furthermore, because by Assumption 6.2, we have that $\|B'_kg_k\| \leq \|B'_kd_k\|$ for all k, from the preceding relation it follows that

$$\|B'_{k+1}g_k\| \leq \|B'_kd_k\|, \qquad \forall\ k. \tag{6.42}$$

By combining Eqs. (6.40)–(6.42), we obtain for all $y \in L_\omega$,

$$f(x_k)-\omega \leq \frac{1}{\underline{\rho}}\|B'_kd_k\|\left\|B_k^{-1}(x_k-y)\right\|, \qquad \forall\ k.$$

Taking the minimum over $y \in L_\omega$, we further obtain

$$f(x_k) - \omega \leq \frac{1}{\underline{\rho}} \|B_k' d_k\| \min_{y \in L_\omega} \left\| B_k^{-1}(x_k - y) \right\|, \qquad \forall\ k.$$

Since $\|B_k' d_k\| \to 0$ by Lemma 6.2(a), and since the distances $\min_{y \in L_\omega} \left\| B_k^{-1}(x_k - y) \right\|$ are bounded, by letting $k \to \infty$ in the preceding relation, we see that

$$\limsup_{k \to \infty} f(x_k) - \omega \leq 0,$$

which is impossible because

$$\inf_{k \geq 0} f(x_k) - \omega = \delta > 0.$$

Q.E.D.

A closer look at the preceding proof reveals that the result of Prop. 6.6 holds when the condition

$$2 \frac{(1 - \rho_k^2)}{\rho_k^2} \frac{C^2}{\mu^2} - \gamma_k (2 - \gamma_k) \leq 0,$$

holds for all sufficiently large k (instead of all k).

References

[**AHK87**] Allen, E., Helgason, R., Kennington, J., and Shetty, B., "A Generalization of Polyak's Convergence Result for Subgradient Optimization," Math. Programming, 37, 1987, pp. 309–317.

[**Akg84**] Akgül, M., Topics in Relaxation and Ellipsoidal Methods, Research Notes in Mathematics, Vol. 97, Pitman, 1984.

[**BaS81**] Bazaraa, M. S., and Sherali, H. D., "On the Choice of Step Size in Subgradient Optimization," European Journal of Operational Research, 7, 1981, pp. 380–388.

[**Ber97**] Bertsekas, D. P., "A New Class of Incremental Gradient Methods for Least Squares Problems," SIAM J. on Optim., Vol. 7, 1997, pp. 913–926.

[**Ber98**] Bertsekas, D. P., Network Optimization: Continuous and Discrete Models, Athena Scientific, Belmont, MA., 1998.

[**Ber99**] Bertsekas, D. P., Nonlinear Programming, (2nd edition), Athena Scientific, Belmont, MA, 1999.

[**BeT89**] Bertsekas, D. P., and Tsitsiklis, J. N., Parallel and Distributed Computation: Numerical Methods, Prentice-Hall, Inc., 1989.

[**BeT96**] Bertsekas, D. P., and Tsitsiklis, J. N., Neuro-Dynamic Programming, Athena Scientific, Belmont, MA., 1996.

[**BeT97**] Bertsimas, D., and Tsitsiklis, J. N., Introduction to Linear Optimization, Athena Scientific, Belmont, MA., 1997.

[**BeT00**] Bertsekas, D. P., and Tsitsiklis, J. N., "Gradient Convergence in Gradient Methods," SIAM J. on Optim., Vol. 10, No. 3, 2000, pp. 627–642.

[**BMP90**] Benveniste, A., Metivier, M., and Priouret, P., Adaptive Algorithms and Stochastic Approximations, Springer-Verlag, N. Y., 1990.

[**BNO02**] Bertsekas D. P., Nedic A., and Ozdaglar A. E., Convex Analysis and Optimization, Athena Scientific, Belmont, MA., to appear in fall 2002.

[**Bor98**] Borkar, V. S., "Asynchronous Stochastic Approximation," SIAM J. on Optim., Vol. 36, 1998, pp. 840–851.

[**Brä93**] Brännlund, U., "On Relaxation Methods for Nonsmooth Convex Optimization," Doctoral Thesis, Royal Institute of Technology, Stockholm, Sweden, 1993.

[**CoL93**] Correa, R., and Lemaréchal, C., "Convergence of Some Algorithms for Convex Minimization," Math. Programming, Vol. 62, 1993, pp. 261–275.

[**DeV85**] Dem'yanov, V. F., and Vasil'ev, L. V., Nondifferentiable Optimization, Optimization Software, N.Y., 1985.

[**Erm66**] Ermoliev, Yu. M., "Methods for Solving Nonlinear Extremal Problems," Kibernetika, Kiev, No. 4, 1966, pp. 1–17.

[**Erm69**] Ermoliev, Yu. M., "On the Stochastic Quasi-gradient Method and Stochastic Quasi-Feyer Sequences," Kibernetika, Kiev, No. 2, 1969, pp. 73–83.

[**Erm76**] Ermoliev, Yu. M., Stochastic Programming Methods, Nauka, Moscow, 1976.

[**Erm83**] Ermoliev, Yu. M., "Stochastic Quasigradient Methods and Their Application to System Optimization," Stochastics, Vol. 9, 1983, pp. 1–36.

[**Erm88**] Ermoliev, Yu. M., "Stochastic Quasigradient Methods," in Numerical Techniques for Stochastic Optimization, Eds., Ermoliev, Yu. M., and Wets, R. J-B., Springer-Verlag, 1988, pp. 141–185.

[**Gai94**] Gaivoronski, A. A., "Convergence Analysis of Parallel Backpropagation Algorithm for Neural Networks," Optim. Methods and Software, Vol. 4, 1994, pp. 117–134.

[**Gal96**] Gallager, R. G., Discrete Stochastic Processes, Kluwer Academic Publishers, 1996.

[**GoK99**] Goffin, J. L, and Kiwiel, K., "Convergence of a Simple Subgradient Level Method," Math. Programming, Vol. 85, 1999, pp. 207–211.

[**GoV84**] Golub, G. H., and Van Loan, C. F., Matrix Computations, The Johns Hopkins University Press, Baltimore, MD, 1984.

[**Gri94**] Grippo, L., "A Class of Unconstrained Minimization Methods for Neural Network Training," Optim. Methods and Software, Vol. 4, 1994, pp. 135–150.

[**HiL93**] Hiriart-Urruty, J.-B., and Lemaréchal, C., Convex Analysis and Minimization Algorithms, Vols. I and II, Springer-Verlag, Berlin and N.Y., 1993.

[**KAC91**] Kim, S., Ahn, H., and Cho, S.-C., "Variable Target Value Subgradient Method," Math. Programming, 49, 1991, pp. 359–369.

[**KaC98**] Kaskavelis, C. A., and Caramanis, M. C., "Efficient Lagrangian Relaxation Algorithms for Industry Size Job-Shop Scheduling Problems," IIE Transactions on Scheduling and Logistics, Vol. 30, 1998, pp. 1085–1097.

[**Kha79**] Khachian, L. G., "A Polynomial Algorithm in Linear Programming," Soviet Mathematics Doklady, No. 20, 1979, pp. 191–194.

[**Kib79**] Kibardin, V. M., "Decomposition into Functions in the Minimization Problem," Automation and Remote Control, Vol. 40, 1980, pp. 1311–1323.

[**KiL01**] Kiwiel, K. C., and Lindberg, P. O., "Parallel Subgradient Methods for Convex Optimization," in Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, Eds., Butnariu, D., Censor, Y., and Reich, S., Elsevier Science, Amsterdam, Netherlands, 2001, pp. 335–344.

[**KiU93**] Kim, S., and Um, B., "An Improved Subgradient Method for Constrained Nondifferentiable Optimization," Operations Research Letters, 14, 1993, pp. 61–64.

[**Kiw96a**] Kiwiel, K. C., "The Efficiency of Subgradient Projection Methods for Convex Optimization, Part I: General Level Methods," SIAM J. on Control and Optim., 34 (2), 1996, pp. 660–676.

[**Kiw96b**] Kiwiel, K. C., "The Efficiency of Subgradient Projection Methods for Convex Optimization, Part II: Implementations and Extensions," SIAM J. on Control and Optim., 34 (2), 1996, pp. 677–697.

[**KLL98**] Kiwiel, K. C., Larsson, T., and Lindberg, P. O., "The Efficiency of Ballstep Subgradient Level Methods for Convex Optimization," Working Paper LiTH-MAT-R-1998-22, Dept. of Mathematics, Linköpings Universitet, Sweden, 1998.

[**KuF90**] Kulikov, A. N., and Fazylov, V. R., "Convex Optimization with Prescribed Accuracy," USSR Computational Mathematics and Mathematical Physics, 30 (3), 1990, pp. 16–22.

[**KuY97**] Kushner, H. J., and Yin, G., Stochastic Approximation Algorithms and Applications, Springer-Verlag, N. Y., 1997.

[**Lem78**] Lemaréchal, C., "Nonsmooth Optimization and Descent Methods," IIASA Research Report RR–78–4, March 1978, Laxenburg, Austria.

[**Lem82**] Lemaréchal, C., "Numerical Experiments in Nonsmooth Optimization," Progress in Nondifferentiable Optimization, Ed., Nurminski E. A., Proceedings of IIASA, Laxenburg, Austria, 1982, pp. 61–84.

[**Luo91**] Luo, Z. Q., "On the Convergence of the LMS Algorithm with Adaptive Learning Rate for Linear Feedforward Networks," Neural Computation, Vol. 3, 1991, pp. 226–245.

[**LuT94**] Luo, Z. Q., and Tseng, P., "Analysis of an Approximate Gradient Projection Method with Applications to the Backpropagation Algorithm," Opt. Methods and Software, Vol. 4, 1994, pp. 85–101.

[**MaS94**] Mangasarian, O. L., and Solodov, M. V., "Serial and Parallel Backpropagation Convergence Via Nonmonotone Perturbed Minimization," Opt. Methods and Software, Vol. 4, 1994, pp. 103–116.

[**MaT90**] Martello, S., and Toth, P., Knapsack Problems, J. Wiley, N.Y., 1990.

[**Min86**] Minoux, M., Mathematical Programming: Theory and Algorithms, J. Wiley, N.Y., 1986.

[**NBB01**] Nedić, A., Bertsekas, D. P., and Borkar, V. S., "Distributed Asynchronous Incremental Subgradient Methods," in Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, Eds., Butnariu, D., Censor, Y., and Reich, S., Elsevier Science, Amsterdam, Netherlands, 2001, pp. 381–407.

[**NeB01a**] Nedić, A., and Bertsekas, D. P., "Incremental Subgradient Methods for Nondifferentiable Optimization," SIAM J. on Optim., Vol. 12, No. 1, 2001, pp. 109–138.

[**NeB01b**] Nedić, A., and Bertsekas, D. P., "Convergence Rate of Incremental Subgradient Algorithms," in Stochastic Optimization: Algorithms and Applications, Eds., Uryasev, S., and Pardalos, P. M., Kluwer Academic Publishers, Dordrecht, Netherlands, 2001, pp. 223–264.

[**Nes84**] Nesterov, Yu. E., "Methods of Minimization of Nonsmooth Convex and Quasiconvex Functions," (Russian) Ekonom. i Mat. Metody, Vol. 20, No. 3, 1984, pp. 519–531.

[**Pol67**] Polyak, B. T., "A General Method of Solving Extremum Problems," Soviet Math. Doklady, Vol. 8, No. 3, 1967, pp. 593–597.

[**Pol69**] Polyak, B. T., "Minimization of Unsmooth Functionals," Z. Vychisl. Mat. i Mat. Fiz., Vol. 9, No. 3, 1969, pp. 509–521.

[**Pol87**] Polyak, B. T., Introduction to Optimization, Optimization Software Inc., N.Y., 1987.

[**Roc70**] Rockafellar, R. T., Convex Analysis, Princeton Univ. Press, Princeton, N.J., 1970.

[**Sho70a**] Shor, N. Z., "An Application of the Operation of Space Dilation to the Problems of Minimizing Convex Functions," (in Russian) Kibernetika, Kiev, No. 1, 1970, pp. 6–12.

[**Sho70b**] Shor, N. Z., "On the Speed of Convergence of the Method of Generalized Gradient Descent with Space Dilation," (in Russian) Kibernetika, Kiev, No. 2, 1970, pp. 80–85.

[**Sho77a**] Shor, N. Z., "A Method of Section with Space Dilation for Solving Convex Programming Problems," (in Russian) Kibernetika, Kiev, No. 1, 1977, pp. 94–95.

[**Sho77b**] Shor, N. Z., "New Trends in the Development of Methods for Nondifferentiable Optimization," (in Russian) Kibernetika, Kiev, No. 6, 1977, pp. 87–91.

[**Sho83**] Shor, N. Z., "Generalized Gradient Methods of Nondifferentiable Optimization Employing Space Dilatation Operations," Mathematical Programming: The State of the Art, XI Inter. Symp. on Math. Programming., Eds., Bachem, A., Grötschel, M., and Korte, B., Springer-Verlag, 1983, pp. 501–529.

[**Sho85**] Shor, N. Z., Minimization Methods for Nondifferentiable Functions, Springer-Verlag, Berlin, 1985.

[**Sho98**] Shor, N. Z., Nondifferentiable Optimization and Polynomial Problems, Kluwer Academic Publishers, 1998.

[**ShZ71**] Shor, N. Z. and Zhurbenko, N. G., "A Minimization Method Using the Operation of Space Dilation in the Direction of the Difference of Two Successive Gradients," (in Russian) Kibernetika, Kiev, No. 3, 1971, pp. 51–59.

[**Sko74**] Skokov, V. A., "Note on Minimization Methods Employing Space Stretching, (in Russian) Kibernetika, Kiev, No. 4, 1974, pp. 115–117.

[**Sol98**] Solodov, M. V., "Incremental Gradient Algorithms with Stepsizes Bounded Away From Zero," Computational Opt. and Appl., Vol. 11, 1998, pp. 28–35.

[**SoZ98**] Solodov, M. V., and Zavriev, S. K., "Error Stability Properties of Generalized Gradient-Type Algorithms," J. Opt. Theory and Appl., Vol. 98, No. 3, 1998, pp. 663–680.

[**TBA86**] Tsitsiklis, J. N., Bertsekas, D. P., and Athans, M., "Distributed Asynchronous Deterministic and Stochastic Gradient Optimization Algorithms," IEEE Trans. on Automatic Control, AC-31, 1986, pp. 803–812.

[**Tse98**] Tseng, P., "An Incremental Gradient(-Projection) Method with Momentum Term and Adaptive Stepsize Rule," SIAM J. on Optim., Vol. 8, No. 2, 1998, 506–531.

[**Ury91**] Uryasev, S. P., "New Variable-Metric Algorithms for Nondifferentiable Optimization Problems," J. Opt. Theory and Appl., Vol. 71, No. 2, 1991, pp. 359–388.

[**WiH60**] Widrow, B., and Hoff, M. E., "Adaptive Switching Circuits," Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, part 4, 1960, pp. 96–104.

[**ZLW99**] Zhao, X., Luh, P. B., and Wang, J., "Surrogate Gradient Algorithm for Lagrangian Relaxation," J. Opt. Theory and Appl., Vol. 100, 1999, pp. 699–712.