A Unified Framework of Descent Algorithms for Nonlinear Programs and Variational Inequalities

by

Michael Patriksson

Department of Mathematics
Linköping Institute of Technology
S-581 83 Linköping
Sweden

LiTH-MAT-R-92-XX

Revised April 3, 1995

To my sisters and their marvellous kids

Contents

Preface

One dark evening in February, 1990, I made my usual visit to the local University library, to look for new journal issues. A new number of 'Current Mathematical Publications' had arrived, and I skimmed through it, ready with a pen and some order forms. I had already filled in a few, when I came across a reference to an article by G. Auchmuty. Although the title did not strike me as very interesting, I ordered a copy of it just in case. It turned out, however, to be an extremely interesting paper, where a class of equivalent formulations of asymmetric variational inequalities as minimization problems was presented. What interested me most, however, was the lack of both references to earlier works on the same topic, and also of a discussion on the possible use of the new merit functions in descent algorithms for the variational inequality problem. Some days later, a preprint of a survey by P. Harker and J.-S. Pang on variational inequalities and complementarity problems reached me. In that paper the question of whether or not there exist differentiable merit functions for asymmetric variational inequalities was raised; an affirmative answer to the question was easily given, by using appropriately chosen functions in Auchmuty's class of formulations. I set out to further develop Auchmuty's work; during Spring several results concerning the use of these merit functions in descent algorithms were obtained by myself and my supervisor, Torbjörn Larsson, and a report on the development was submitted to Mathematical Programming. During Autumn the same year, I played with the formulations of the saddle functions involved in Auchmuty's merit functions, and tried to figure out what sense they made when applied to nonlinear programming. Some rather interesting connections evolved; when applying Auchmuty's framework to nonlinear programs, the subproblem that defines the merit functions is equivalent to the one which is solved as a direction finding subproblem in the regularized Frank-Wolfe method, as well as in a generalization of a partial linearization method; both of these algorithms were developed by my colleague Sakis Migdalas. Interestingly, these results brought me back to my Master's Thesis—supervised by Sakis—which dealt with dual methods for certain partial linearization subproblems, in an application to traffic assignment. Nothing is new under the sun, indeed. An article by M. Fukushima on descent methods for constrained, nondifferentiable problems prompted me to further generalize these methods to the nondifferentiable case, and a report on general partial linearization methods was finalized in early Spring of 1991. A large number of well known iterative methods can be viewed as special cases of the general method, including Newton methods, Jacobi and Gauss-Seidel algorithms, and proximal point methods, among others; a report devoted to a unification of methods for traffic assignment problems was written during Spring. In the Summer of 1991 I had the opportunity to participate in the 7th EURO Summer School in Cetraro, Italy. I presented the unified framework at the summer school, and had some interesting discussions on it with Jiahao Wu. His forthcoming dissertation dealt with a generalization of regularized Frank-Wolfe methods to variational inequalities, so we had a lot of research interests in common. He also mentioned this Guy Cohen, who had presented a class of iterative methods for variational inequalities. I took a closer look at his work after my return to Sweden. It turned out that his algorithm class was quite similar to my framework, although it was developed from a different standpoint; he also gave several examples of its possible use. Some time during Autumn 1991 two other important connections became clear. While sitting in a Linköping café, I came to realize that S. Dafermos' class of iterative methods for variational inequalities was essentially equivalent to mine, and that the same was true for a class of methods recently presented by P. Tseng. At that point (due to all the possible interpretations and connections), I could no longer find a suitable name for the method class! When, in the beginning of 1992, M. Fukushima published a paper on a descent algorithm for variational inequalities based on a differentiable merit function, which could be identified as a special case from Auchmuty's class, the thought of presenting all these, seemingly unknown, connections in a thesis became more concrete, and the work on it started in Spring 1992. Soon after, I finally decided to use the term *cost approximation* as a description of the basic principle, because of its natural interpretation. Up to the time when this is written, several more interesting relationships have been revealed, for instance those to deflected gradient methods for unconstrained optimization, which include standard methods such as steepest descent. Hopefully, also the relationships among the algorithms for nonlinear programming and variational inequalities that are mentioned above will become clear to the reader of this thesis. To finally quote G. Cohen, "we let the reader play the game by himself with any algorithm he is aware of and convince himself whether or not our objectives are met".

Acknowledgements

My deepest and most sincere thanks go to my supervisor, Torbjörn Larsson, for all the help and support he has given me over the years. Under his guidance, the process of learning both optimization theory and research methodology has been a great pleasure, and I only hope to live up to his expectations in the future.

Many heartfelt thanks go to Sakis Migdalas, for introducing me to the subject matter of the thesis, for the inspiration his work has given me and for his help with some of the obstacles in my projects.

Johan Hellstrand has shared with me not only doctoral courses and teaching, but also common interests outside the academic world. I thank him for being a true friend, and for interrupting my work now and then.

Thanks are also due to those others with whom I have conducted research during these years. I would especially like to mention Ann-Brith Strömberg, with whom I have had many stimulating discussions.

Olof Damberg helped me to cure some of LaTeX's worst diseases. Any remaining flaws are probably due to the reader's fastidious tastes.

Valuable comments on a draft version were kindly provided by many people at the Division of Optimization. Thanks go to the whole group for a friendly and stimulating working atmosphere.

Professor R. Tyrrell Rockafellar provided constructive criticism on a previous version of the thesis.

Thanks to Pamela Vang for comments on the English.

A high level of service from the local library is important for research work to carry on smoothly. Not only is our local university library one of best in the field of optimization, but the assistance of its staff has been beyond the call of duty.

Finally, I would like to express my appreciation to the musicians—some of them sadly missed—who have "bent a string" in my earphones during the completion of the thesis, thereby helping me through the tedious work of forcing LaTeX to fulfill my wishes. In particular, I would like to mention Duane Allman, Jeff Beck, Mike Bloomfield, Peter Buck, Eric Clapton, Ry Cooder, Steve Cropper, The Edge, Robben Ford, David Gilmour, Peter Green, BB King, Freddie King, David Lindley, Johnny Marr, Jimmy Page, Robbie Robertson, and Ricky Wilson.

Linköping, June 1993

Michael Patriksson

Abstract

We present a framework of algorithms for the solution of continuous optimization and variational inequality problems. In the general algorithm, a search direction finding auxiliary problem is obtained by replacing the original cost function with an approximating monotone cost function. Alternately, a line search is made in the search direction in order to obtain a decrease in a merit function. The proposed framework encompasses algorithm classes presented earlier by Cohen, Dafermos, Migdalas, and Tseng, and includes numerous descent and successive approximation type methods, such as Newton methods, Jacobi and Gauss-Seidel type decomposition methods for problems defined over Cartesian product sets, and proximal point methods, among others. The auxiliary problem of the general algorithm also induces equivalent optimization reformulations and descent methods for asymmetric variational inequalities.

We study the convergence properties of the general algorithm when applied to unconstrained optimization, nondifferentiable optimization, constrained differentiable optimization, and variational inequalities; the emphasis of the convergence analyses is placed on basic convergence results, convergence using different line search strategies and truncated subproblem solutions, and convergence rate results. This analysis offers a unification of known results; moreover, it provides strengthenings of convergence results for many existing algorithms, and indicates possible improvements of their realizations.

Some notations

$\Re^n_+$, $\Re^n_-$	Nonnegative and nonpositive orthants of $\Re^n$
2^X	Set of all subsets of X
$\lvert\mathcal{C}\rvert$	Cardinality of a finite set $\mathcal{C}$
$X \times Y$, $\prod_{i\in\mathcal{C}} X_i$	Cartesian product of X and Y and of X_i, $i \in \mathcal{C}$
$[a, b]$	Closed interval $\{x \in \Re \mid a \le x \le b\}$
$\lceil x \rceil$	Upper integer part of x
$\mathtt{int}\ (X)$, $\mathtt{rint}\ (X)$	Interior and relative interior of X
$\mathtt{aff}\ (X)$, $\mathtt{lin}\ (X)$	Affine hull and lineality of X
X°	Polar set of X
P_X, $P_X^{\mathbf{G}}$	Projection onto X, Euclidean and w.r.t. the $\mathbf{G}$-norm
δ_X	Indicator function of X
N_X, T_X	Normal and tangent cones of X
X_F, X_{F^*}	A face of X and optimal face
$\nabla_i f$	Gradient of f with respect to $\mathbf{x}_i$
$\nabla_{\mathbf{x}}\varphi(\mathbf{x}, \mathbf{y})$	Gradient of φ with respect to its first component $\mathbf{x}$
$\nabla^X f$, $\mathbf{F}^X$	Projected gradient and mapping
$\boldsymbol{\xi}_f(\mathbf{x}) \in \partial f(\mathbf{x})$	A subgradient of f at $\mathbf{x}$
$f'(\mathbf{x}, \mathbf{d})$	Directional derivative of f at $\mathbf{x}$ in the direction $\mathbf{d}$
f°	Conjugate function of f
f^*	Optimal value of f
$\mathtt{dom}\ (f)$	Effective domain of f
$\mathtt{range}\ (f)$	Range of f
L_X^f	Level set of f with respect to X
$f \in C^p$	f is p times continuously differentiable
$f \in C^p$ on X	f is p times continuously differentiable on an open neighbourhood of X
u.s.c., l.s.c.	Upper and lower semicontinuity
$\mathbf{x}_{i_-}$, $\mathbf{x}_{i_+}$	Elements $(x_1, \dots, x_{i-1})^{\mathrm{T}}$ and $(x_{i+1}, \dots, x_n)^{\mathrm{T}}$ of $\mathbf{x}$
$\mathbf{e}_i$	Unit vector
$\mathbf{I}$	Identity matrix
symm $(\mathbf{B})$	Symmetric part of $\mathbf{B}$
$\lVert\cdot\rVert$	Euclidean vector norm and induced matrix norm
$\lVert\cdot\rVert_{\mathbf{B}}$	Matrix norm defined by $\mathbf{B}$
$\lvert\lvert\lvert\cdot\rvert\rvert\rvert$	Operator norm induced by the Euclidean vector norm
M_f	Lipschitz continuity constant
m_f	Strong convexity (or monotonicity) constant

Chapter 1

Introduction

Anthony raised a hand for silence. "Fear not. Of course you can fit other theories. One always can; that's the devil of this bloodhound business. The only way to work is to pick a likely-looking path and go down it. I've chosen one to get on with."

PHILIP MACDONALD
The Rasp (1924)

Let X be a nonempty, closed and convex set in $\Re^n$ and let there be given two mappings (operators), $\mathbf{F}$ and $\mathbf{U}$, on X. While the mapping $\mathbf{F}$ is assumed to be point-to-point (so it is a map from X to $\Re^n$), $\mathbf{U}$ is a monotone point-to-set map (from X to $2^{\Re^n}$). Consider the problem of finding an $\mathbf{x}^* \in X$ and a $\mathbf{u}^* \in \mathbf{U}(\mathbf{x}^*)$ such that

[**GVIP**] $$[\mathbf{F}(\mathbf{x}^*) + \mathbf{u}^*]^{\mathrm{T}} (\mathbf{x} - \mathbf{x}^*) \geq 0, \ \ \forall \mathbf{x} \in X.$$

This problem is known as the *generalized variational inequality problem* [FaP82], and as the *nonlinear variational inequality problem* [Noo75, Noo82a, Noo82b, Noo91b]. The problem [**GVIP**] (and its special cases) has a large variety of applications in the mathematical and engineering sciences, for example in partial differential equations [HaS66, DuL72, EkT76, CGL80, KiS80, GLT81, BaC84, Cra84, Rod87, KiO88], equilibrium problems in games, economics and transportation analysis [Kar69a, Kar69b, Smi79, Daf80, BKS80, AhH82, BaC84, Flo86, Mat87, Daf90, Pat90], and nonlinear programming [Roc69a, Sta69, Kar69b, MaS72, Aus76, Roc80, BeT89, HaP90b].

The purpose of the thesis is to present and investigate a framework of descent algorithms for the solution of instances of [**GVIP**]. One iteration of the basic algorithm consists of two main steps, which can briefly be described as follows.

(1) Given a feasible point, a feasible search direction is defined through the (possibly inexact) solution of an approximation of the original problem, in which the mapping $\mathbf{F}$ is approximated by a monotone mapping.

(2) The direction defined by the solution to the above described subproblem is a feasible direction of descent with respect to a merit function whose minima coincide with the set of solutions to [**GVIP**]. A (possibly inexact) line search is made with respect to this merit function in the obtained direction, and the resulting step length defines a new point with a reduced value of the merit function.

Characteristic features of an algorithm included in the framework are hence that it generates a sequence of feasible solutions to the problem, with a monotonically decreasing value of a merit function, and that the descent directions are given through the solution of subproblems, defined by certain approximations of the original problem defining map.

This chapter gives an overview of the contents of the thesis. In the next section, we introduce the instances of [**GVIP**] that are given specific attention in the thesis, and describe relationships among them. In Section 1.2, we present the algorithm framework for the solution of [**GVIP**]. The structure of the remainder of the thesis is outlined in Section 1.3, together with the most important results obtained. In Section 1.4, we summarize the contributions of the thesis.

1.1 Instances of the problem

We begin by assuming that $\mathbf{U}$ is the subdifferential mapping of a lower semicontinuous (l.s.c.), proper and convex function $u : \Re^n \mapsto \Re \cup \{+\infty\}$. Under the additional assumption

$$\texttt{rint}\,(\texttt{dom}\,\mathbf{U}) \cap \texttt{rint}\,X \neq \emptyset, \tag{1.1}$$

one may show, using [Roc66, Roc70b, Roc70c] and [Roc70a, Th. 27.4], that [**GVIP**] then is equivalent to the problem of finding $\mathbf{x}^* \in X$ such that

$$\mathbf{F}(\mathbf{x}^*)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^*) + u(\mathbf{x}) - u(\mathbf{x}^*) \geq 0, \ \forall \mathbf{x} \in X. \tag{1.2}$$

If $u \equiv 0$, we obtain the *variational inequality problem* of finding $\mathbf{x}^* \in X$ such that

[**VIP**] $$\mathbf{F}(\mathbf{x}^*)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^*) \geq 0, \ \forall \mathbf{x} \in X.$$

It is also known as the *stationary point problem* [Eav78b]. By further assuming that $\mathbf{F}$ is the gradient mapping of a continuously differentiable function $f : X \mapsto \Re$, [**VIP**] defines the first-order necessary optimality conditions for the *constrained nonlinear program* [BaS79, Th. 4.1.3]

[**CNLP**] $$\min_{\mathbf{x} \in X} \ f(\mathbf{x}).$$

Under the same assumption on $\mathbf{F}$, the variational inequality (1.2) defines the necessary optimality conditions for the constrained nondifferentiable program [Cea78, Th. 2.3]

$$\min_{\mathbf{x} \in X} \ \{f(\mathbf{x}) + u(\mathbf{x})\}. \tag{1.3}$$

So far, the feasible set X has been treated explicitly. By introducing an indicator function, the set X can instead be treated implicitly, as demonstrated below. This fact is important for the establishing of relations among instances of [**GVIP**].

We first restate [**GVIP**] as a *generalized equation* [Rob79, Rob82, Rob83]. Let δ_X be the *indicator function* of X, i.e.,

$$\delta_X(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in X \\ +\infty, & \mathbf{x} \notin X. \end{cases}$$

It is well known that its subdifferential, $\partial\delta_X(\mathbf{x})$, at $\mathbf{x} \in \Re^n$ equals the *normal cone*, $N_X(\mathbf{x})$, to X at $\mathbf{x}$ [Roc70a, p. 215], which is defined by

$$N_X(\mathbf{x}) = \begin{cases} \left\{ \mathbf{z} \in \Re^n \;\middle|\; \mathbf{z}^{\mathrm{T}}(\mathbf{y}-\mathbf{x}) \leq 0, \ \forall \mathbf{y} \in X \right\}, & \mathbf{x} \in X \\ \emptyset, & \mathbf{x} \notin X. \end{cases}$$

Using this fact, [**GVIP**] may, under the assumption that (1.1) holds, be equivalently reformulated as the generalized equation[1]

$$\mathbf{0} \in \mathbf{F}(\mathbf{x}^*) + \mathbf{U}(\mathbf{x}^*) + N_X(\mathbf{x}^*). \tag{1.4}$$

The generalized equations corresponding to (1.2) and [**VIP**] are

$$\mathbf{0} \in \mathbf{F}(\mathbf{x}^*) + \partial u(\mathbf{x}^*) + N_X(\mathbf{x}^*) \tag{1.5}$$

and

$$\mathbf{0} \in \mathbf{F}(\mathbf{x}^*) + N_X(\mathbf{x}^*), \tag{1.6}$$

respectively.

Now consider [**GVIP**] with $X = \Re^n$. The corresponding generalized equation is

$$\mathbf{0} \in \mathbf{F}(\mathbf{x}^*) + \mathbf{U}(\mathbf{x}^*), \tag{1.7}$$

i.e., the problem of finding a zero of the sum of two operators (see [Bre73, LiM79, Tse91a, EcB92], and the references cited therein). Again, assuming that $\mathbf{U}$ is a subdifferential mapping as defined above, the generalized equation (1.7) reduces to

$$\mathbf{0} \in \mathbf{F}(\mathbf{x}^*) + \partial u(\mathbf{x}^*), \tag{1.8}$$

which, under the assumption that $\mathbf{F} \equiv \nabla f$, reduces to the necessary optimality conditions for the *nondifferentiable nonlinear program*

[**NNLP**]
$$\min_{\mathbf{x} \in \Re^n} \; \{f(\mathbf{x}) + u(\mathbf{x})\}.$$

Whenever $u \equiv 0$, the Inclusion (1.8) further reduces to the nonlinear equation system

$$\mathbf{F}(\mathbf{x}^*) = \mathbf{0}; \tag{1.9}$$

this system defines the first-order optimality conditions for the *nonlinear program*

[**NLP**]
$$\min_{\mathbf{x} \in \Re^n} \; f(\mathbf{x})$$

whenever $\mathbf{F} \equiv \nabla f$.

The linear and nonlinear complementarity problems [Kar69a, Kar69b, Eav71, Kar71, Kar72, Sai76, Mur88, HaP90b, CPS92, Pan93a] are defined by letting X in [**GVIP**] be a closed, convex cone. In what follows, these problems are not treated separately.

[1] Solving [**GVIP**] can therefore be interpreted as finding an $\mathbf{x}^* \in X$ and a $\mathbf{u}^* \in \mathbf{U}(\mathbf{x}^*)$ such that the vector $\mathbf{F}(\mathbf{x}^*) + \mathbf{u}^*$ is an inward normal to X at $\mathbf{x}^*$.

From the above discussion it is clear that constraints can either be modelled explicitly, or implicitly by redefining $\mathbf{U}$ to include a normal cone operator. Hence, whenever the mapping $\mathbf{U} \not\equiv \mathbf{0}$, one may, without loss of generality, assume that $X = \Re^n$.

In Figure 1.1 we summarize the above development of relations among instances of [**GVIP**]. In the figure, a problem $\mathbf{A}$ is understood to mean the problem of finding a solution to a variational inequality, a set inclusion, or a minimization problem. The relation $\mathbf{A} \subset \mathbf{B}$ states that the class of problems that can be represented in the form of $\mathbf{A}$ is included in that of $\mathbf{B}$, while the relation $\mathbf{A} = \mathbf{B}$ states that the reverse is true also, i.e., that $\mathbf{A}$ and $\mathbf{B}$ are equivalent problem classes.

$$
\begin{array}{ccccccccc}
[\mathbf{NLP}] & \subset & [\mathbf{NNLP}] & = & (1.3) & \supset & [\mathbf{CNLP}] & \supset & [\mathbf{NLP}] \\
\cap & & \cap & & \cap & & \cap & & \cap \\
(1.9) & \subset & (1.8) & = & (1.2) & \supset & [\mathbf{VIP}] & \supset & (1.9) \\
 & & \cap & & \cap & & & & \\
 & & (1.7) & = & [\mathbf{GVIP}] & & & &
\end{array}
$$

Figure 1.1: Relationships among problem classes

1.2 The cost approximation algorithm

In this section we present the algorithm framework for the solution of [**GVIP**]. The main idea is to iteratively approximate the cost mapping[2] $\mathbf{F}+\mathbf{U}$, in order to obtain a variational inequality problem which is more easily solved. The solution to this problem defines a search direction, in which a step is taken so as to decrease a merit function, whose set of minimizers coincides with the set of solutions to [**GVIP**].

Let $\mathbf{x}$ be a point in X. Introduce a monotone *cost approximating mapping* $\mathbf{\Phi} : X \mapsto \Re^n$. If, at $\mathbf{x} \in X$, the mapping $\mathbf{F}$ is replaced by $\mathbf{\Phi}$, the error made in the approximation obviously is $\mathbf{F} - \mathbf{\Phi}$. This error is taken into account by adding to $\mathbf{\Phi}$ the fixed error term $\mathbf{F}(\mathbf{x}) - \mathbf{\Phi}(\mathbf{x})$. [Alternatively, the approximation made can be seen as a fixation of the second term of the original cost $[\mathbf{\Phi}+\mathbf{U}]+[\mathbf{F}-\mathbf{\Phi}]$, at $\mathbf{x}$.] Thus, we arrive at a generalized variational inequality subproblem, in which a point $\mathbf{y}(\mathbf{x}) \in X$ and a $\mathbf{u}(\mathbf{y}(\mathbf{x})) \in \mathbf{U}(\mathbf{y}(\mathbf{x}))$ are sought such that[3]

$$
[\mathbf{GVIP}_{\mathbf{\Phi}}] \qquad [\mathbf{\Phi}(\mathbf{y}(\mathbf{x})) + \mathbf{u}(\mathbf{y}(\mathbf{x})) + \mathbf{F}(\mathbf{x}) - \mathbf{\Phi}(\mathbf{x})]^{\mathrm{T}} (\mathbf{y} - \mathbf{y}(\mathbf{x})) \geq 0, \ \forall \mathbf{y} \in X.
$$

We let $Y(\mathbf{x})$ denote the set of solutions $\mathbf{y}(\mathbf{x})$ to $[\mathbf{GVIP}_{\mathbf{\Phi}}]$.

By construction, the value of the problem defining mapping of $[\mathbf{GVIP}_{\mathbf{\Phi}}]$ coincides with that of the original cost map at the point of approximation. This fact is important, since it provides a termination criterion for the algorithm: if $\mathbf{x}$ solves the subproblem $[\mathbf{GVIP}_{\mathbf{\Phi}}]$, defined at $\mathbf{x}$, then it immediately follows that $\mathbf{x}$ also solves the original problem

[2] This term is taken from the important applications of [**GVIP**] in equilibrium problems, where $\mathbf{F}+\mathbf{U}$ usually represents a cost vector.

[3] Note that a fixed cost may be added to $\mathbf{\Phi}$ without changing the subproblem $[\mathbf{GVIP}_{\mathbf{\Phi}}]$.

[**GVIP**]. Observe that the point-to-set map $\mathbf{U}$ is not involved in the approximation. This is because otherwise neither is the above termination criterion valid, nor does the solution $\mathbf{y}(\mathbf{x})$ yield a descent direction in general. [An example of this fact is in Section 3.2 given for the problem (1.3).]

Remark 1.1 Under the assumptions that $\mathbf{U}$ is *maximal monotone* [Min62, Bro68] and (1.1) holds, the mapping $\mathbf{T} \stackrel{\text{def}}{=} \mathbf{\Phi} + \mathbf{U} + N_X$ is also maximal monotone, and the solution set, $Y(\mathbf{x})$, to [$\mathbf{GVIP}_{\mathbf{\Phi}}$] may be written as

$$Y(\mathbf{x}) = [\mathbf{\Phi} + \mathbf{U} + N_X]^{-1} [\mathbf{\Phi} - \mathbf{F}] (\mathbf{x}), ^4 \tag{1.10}$$

where $\mathbf{T}^{-1}$ is the inverse mapping of $\mathbf{T}$.[5] The characterization (1.10) of $Y(\mathbf{x})$ is useful in the convergence analysis of some special cases of cost approximation methods.

If $\mathbf{U}$ is a subdifferential mapping as defined in Section 1.1, and if further $\mathbf{\Phi}$ is chosen as the gradient mapping of a continuously differentiable convex function $\varphi : X \mapsto \Re$, the subproblem [$\mathbf{GVIP}_{\mathbf{\Phi}}$] amounts to solving the convex minimization problem

$$\min_{\mathbf{y} \in X} \left\{ \varphi(\mathbf{y}) + u(\mathbf{y}) + [\mathbf{F}(\mathbf{x}) - \nabla\varphi(\mathbf{x})]^{\mathrm{T}} \mathbf{y} \right\}. \tag{1.11}$$

Thus, by applying the cost approximation concept, [**GVIP**] can be solved as a sequence of optimization problems. Indeed, in most of the iterative methods for the solution of [**GVIP**] that we will identify as special cases from the class of cost approximation algorithms, subproblems of the form (1.11) are solved.

In general, we can not expect the original problem to be solved by $\mathbf{y}(\mathbf{x})$. A new iteration point is therefore defined through a line search in the direction $\mathbf{y}(\mathbf{x}) - \mathbf{x}$ with respect to a merit function, $\psi : X \mapsto \Re \cup \{+\infty\}$, whose minima coincide with the set of solutions to [**GVIP**]. When the problem [**GVIP**] corresponds to a mathematical program, in the sense that the map $\mathbf{F} + \mathbf{U}$ is the subdifferential of a function $f + u$, it is natural to choose $\psi = f + u$ as merit function. If this is not the case, another merit function must be identified; it will be shown that a merit function for [**VIP**] is inherent in the subproblem corresponding to (1.11) and can, under certain conditions, be utilized for the solution of this problem.

At the new point, the original mapping is again approximated—perhaps with another mapping $\mathbf{\Phi}$—and the algorithm proceeds until some stopping criterion is fulfilled.

Below, we summarize the different steps of the general algorithm.

[4] Note that, although the subproblem (1.10) bears resemblance to those of proximal point methods [Min62, Mar70, Roc76b], and forward-backward splitting methods [Bre73, Eck89, EcB92, ChR92] for finding a zero of a sum of two maximal monotone maps, the class of cost approximation algorithms is far more general; for instance, the mapping $\mathbf{\Phi}$ is not necessarily affine or strongly monotone and can be chosen to adapt to problem structures (such as separability in $\mathbf{U}$, $\mathbf{F}$ or X), and a (possibly inexact) line search is embedded in the algorithm.

[5] Since, by definition, $\texttt{range}\ \mathbf{T} = \texttt{dom}\ \mathbf{T}^{-1}$, $Y(\mathbf{x})$ is nonempty whenever $[\mathbf{\Phi} - \mathbf{F}](\mathbf{x}) \in \texttt{dom}\ \mathbf{T}^{-1}$; furthermore, it is locally bounded if $[\mathbf{\Phi} - \mathbf{F}](\mathbf{x}) \in \texttt{int}\ (\texttt{dom}\ \mathbf{T}^{-1})$ [Roc69b]. A sufficient condition for the subproblem solution to be attained and bounded is that $\mathbf{T}$ is *weakly coercive* [Zei90, Def. 32.34.b], since this condition implies that $\texttt{dom}\ \mathbf{T}^{-1} = \Re^n$ [Zei90, Cor. 32.35] (see also [Bro68]).

The cost approximation algorithm

A sequence $\{\mathbf{\Phi}^k\}$ of monotone cost approximating mappings[6] and a merit function ψ are assumed to be given.

Step 0 (*Initial guess*) Choose an initial point $\mathbf{x}^0 \in X$, and let $k = 0$.

Step 1 (*Search direction generation*) Find a $\mathbf{y}^k \in X$ and a $\mathbf{u}^k \in \mathbf{U}(\mathbf{y}^k)$ such that

$$[\mathbf{GVIP}^k_{\mathbf{\Phi}^k}] \quad \left[\mathbf{\Phi}^k\left(\mathbf{y}^k\right) + \mathbf{u}^k + \mathbf{F}\left(\mathbf{x}^k\right) - \mathbf{\Phi}^k\left(\mathbf{x}^k\right)\right]^{\mathrm{T}} \left(\mathbf{y} - \mathbf{y}^k\right) \geq 0, \ \forall \mathbf{y} \in X.$$

The resulting search direction is $\mathbf{d}^k = \mathbf{y}^k - \mathbf{x}^k$.

Step 2 (*Convergence check*) If $\mathbf{x}^k$ solves $[\mathbf{GVIP}^k_{\mathbf{\Phi}^k}] \to$ Stop ($\mathbf{x}^k$ solves [**GVIP**]). Otherwise, continue.

Step 3 (*Line search*) Find a step length, t_k, which solves the one-dimensional problem

$$t_k \in \arg \min_{\substack{t \geq 0 \\ \mathbf{x}^k + t\mathbf{d}^k \in X}} \psi\left(\mathbf{x}^k + t\mathbf{d}^k\right).$$

Step 4 (*Update*) Let $\mathbf{x}^{k+1} = \mathbf{x}^k + t_k\mathbf{d}^k$, and $k := k + 1$.

Step 5 (*Convergence check*) If $\mathbf{x}^k$ is acceptable as a solution[7] $\to$ Stop. Otherwise, go to Step 1.

Although both the subproblem solution (Step 1) and the line search (Step 3) are, in the above description, performed exactly, it will in the sequel be shown that they can both be done inexactly, while still ensuring convergence of the algorithm. (The step length rules considered are described in the Appendix.) This is of great practical importance, since neither of the two steps can, in general, be performed exactly. It will also be shown that the termination criterion in Step 2 can, for all the instances of [**GVIP**] studied in the thesis, be checked without any effort.

Many well known algorithms for [**GVIP**] can be identified as special cases from the class of cost approximation methods, by appropriately choosing $\{\mathbf{\Phi}^k\}$ and ψ. In many of these algorithms, the new iterate is given by the solution to the subproblem. Such algorithm are often referred to as *successive approximation algorithms*, and may, in the algorithm framework, be identified as algorithms where a unit step length is taken in the direction $\mathbf{d}^k$. There are two main motivations for instead considering step lengths given by line searches. First, convergence can in general be established under weaker conditions. Second, introducing a line search may enhance both the theoretical and practical convergence rates of the algorithm.

The convergence of many iterative algorithms relies on a monotone decrease in the Euclidean distance to the set of solutions to [**GVIP**], Ω. (This is especially the case for

[6] Note, however, that each mapping may also be constructed adaptively, given $\mathbf{x}^k$, instead of being chosen *a priori*.

[7] We will not discuss choices of stopping criteria in any detail, but only give some examples for the instances of [**GVIP**] studied.

successive approximation algorithms.) Such algorithms can also be included in the above algorithm framework, by letting $\psi(\mathbf{x}) = \inf_{\mathbf{y}\in\Omega} \|\mathbf{x}-\mathbf{y}\|^2$. We are, however, interested in merit functions that are possible to evaluate and use in a line search; this is obviously not possible with the Euclidean distance measure.

1.3 Outline of the thesis

The scope of this thesis is algorithm theory in the fields of nonlinear programming and variational inequalities.[8] The purpose is to present a unified framework of descent algorithms for the solution of instances of [**GVIP**], and investigate its convergence characteristics when applied to four interesting problem classes: unconstrained optimization (Chapter 2), nondifferentiable optimization (Chapter 3), constrained differentiable optimization (Chapter 4), and variational inequalities (Chapter 5). These problem classes represent the most important instances of [**GVIP**].[9]

The reason for studying each problem class separately is twofold. First, the presentation and analysis of the algorithm framework becomes self-contained for each problem class; second, the establishment of results for each problem class separately simplifies comparisons of convergence criteria, etc., among the problem classes, as well as among different algorithms within one particular problem class. To facilitate a coherent analysis of the cost approximation methods for the four problem classes considered, the chapters are given a common structure. Below, we describe this structure. Each item in the list corresponds to a section; for each section we discuss its main scope and purpose, and summarize the most important results obtained.

Optimality conditions. For the first three problems (unconstrained optimization, nondifferentiable optimization, and differentiable constrained optimization) there is a natural choice of merit function ψ for Step 3 of the cost approximation algorithm, since, from the discussion made in Section 1.1, the corresponding formulation of [**GVIP**] defines the first-order necessary optimality conditions of the minimization problems [**NLP**], [**NNLP**] and [**CNLP**], respectively. A brief discussion on the relations between [**GVIP**] and the corresponding minimization problem is the topic of this section.

For variational inequality problems, a merit function is not directly available, due to the nonintegrability of the problem defining map $\mathbf{F}$. The subject matter of Sections 5.5 and 5.6 is equivalent reformulations of [**VIP**] as minimization problems. We show that a merit function is defined by the solution to the subproblem corresponding to (1.11), for any choice of monotone and integrable cost approximating mapping $\mathbf{\Phi}$, i.e., a merit function for use in Step 3 is implicitly defined by the solution of a convex program. As a byproduct of this development, we show that most known optimization formulations of [**VIP**] are instances of the class of reformulations defined in this fashion.

The cost approximation algorithm. A description of the general cost approximation algorithm, as applied to the given problem.

Instances of cost approximation algorithms. The class of cost approximation

[8] The reader is assumed to be familiar with the basic concepts in these areas; some of the more commonly used ones are listed in Appendix A.

[9] We choose to treat nondifferentiable problems in the form [**NNLP**], although the problem (1.3) and [**NNLP**] are equivalent (cf. Figure 1.1). Further, although the class [**CNLP**] defines a subset of [**NNLP**], its importance in applications, and the fact that many results for [**CNLP**] do not hold in the case of nondifferentiability, motivate a separate study of [**CNLP**].

algorithms includes many well known algorithms as special cases, which are derived by appropriately choosing the sequence $\{\mathbf{\Phi}^k\}$ and ψ. The possibility of characterizing algorithms by particular choices of cost approximating mappings and merit functions enables a simple comparison among them. To mention some examples, for [**NLP**] we identify the steepest descent, quasi-Newton and Gauss-Seidel and Jacobi methods as instances of cost approximation; for [**NNLP**], we here mention subgradient optimization; for [**CNLP**], we mention the Frank-Wolfe, Newton, gradient projection, proximal point, and regularization algorithms; for [**VIP**], finally, we mention the class of linearization methods, diagonalization methods, and gap decreasing algorithms. The fact that the proximal point method is included in the framework is interesting in its own right, since it implies that many well known duality based methods, such as augmented Lagrangean methods, and splitting algorithms, that can be identified as instances of the proximal point algorithm, are *a fortiori* cost approximation algorithms.

In this section we also demonstrate through examples that cost approximation algorithms can effectively utilize problem structures. This is an important property of this class of algorithms, particularly for their applicability to large-scale problems.

Algorithmic equivalence results. One of the main purposes of the thesis is to demonstrate the close connections among classes of algorithms for the solution of instances of [**GVIP**]. The established relations enable a cross-fertilization among the areas where these method classes are presented. In this section we show that, for special choices of $\{\mathbf{\Phi}^k\}$, the search direction finding subproblem, corresponding to [$\mathbf{GVIP}^k_{\mathbf{\Phi}^k}$], is equivalent to those of a number of earlier presented algorithm classes: for [**NLP**], the class of deflected gradient methods; for [**NNLP**], the auxiliary problem principle of Cohen; for [**CNLP**], the method classes of Migdalas, Dussault *et al.*, and Tseng; for [**VIP**], the algorithm classes of Cohen, Dafermos, Wu *et al.*, and Noor. Important to mention here is that the class of cost approximation methods—where a method is characterized by its direction finding subproblems—encompasses all the mentioned algorithm classes.

Analysis of the direction finding subproblem. In this section we analyze the descent and continuity properties of the search directions generated by the subproblems. We validate the termination criterion of Step 2, and establish that descent directions are obtained also when the solution of the subproblem is truncated prematurely. For the first three problems considered, the analysis is essentially uniform, while for variational inequalities, further conditions must be imposed on the cost approximation to yield descent directions with the respect to the merit function.

Basic convergence of cost approximation algorithms. Using the continuity and descent properties obtained in the previous section, we establish in this section a basic convergence theorem, which states that the sequence of iterates converges to the set of solutions to the problem considered provided that all calculations are carried out exactly. The convergence analysis is made within the framework of Zangwill [Zan69a, Zan69b], and is essentially uniform for all the problem classes considered. The choice of Zangwill's framework is made for two reasons. Firstly, Zangwill's Theorem A is a classical and widely utilized convergence result, and therefore it constitutes a basis for comparisons of convergence analyses among iterative algorithms. Secondly, the properties of composite algorithmic maps have been thoroughly studied within his framework [Zan67, Zan69b]; this makes it possible to derive sufficient conditions for convergence from the properties of the original data.[10]

[10] Extensions of Zangwill's results have been proposed, see e.g. [Pol63, MeP71, Pol71, Dub73, Hua75,

For unconstrained programs, the characterization of the cost approximation algorithms as being gradient related provides an additional convergence result, for the case when the exact line search is replaced by a Goldstein step length rule; under slightly stronger assumptions on the problem data, we show that Armijo and predetermined step lengths may also be used.

Convergence under alternative step length rules. In this section we establish that also for the other three problems simple step length rules may be utilized, while still ensuring convergence. We concentrate on two rules: the Armijo rule and predetermined step lengths. Summarizing the results, the conditions on $\{\mathbf{\Phi}^k\}$ needed to obtain convergence become stronger, as the step length rule used becomes simpler. The validation of simple step length rules is very important for the practical realization of cost approximation algorithms, since an exact line search can not be made in practice. Also, it can be of practical advantage to avoid evaluating the merit function too often; these calculations can be avoided completely by using predetermined step lengths. The analysis made in this section is uniform for the problem classes.

A comparison of convergence conditions. In this section we apply the convergence conditions obtained in the previous two sections to some important instances of cost approximation algorithms. Perhaps somewhat surprisingly—considering the large size of the algorithm framework—we are able to show that often the conditions reduce to those obtained in earlier made specialized convergence analyses, and in some cases actually weaken them. This is particularly the case when comparing our analysis to those of Cohen, Dussault *et al.*, Migdalas, and Tseng for related method classes; in their analyses, such basic algorithms as the Frank-Wolfe method and Newton's method are not included. In our opinion, the algorithm framework is validated by the fact that the convergence analysis is strong enough to reproduce or strengthen many results previously obtained.

A truncated version of cost approximation algorithms. We here consider the possibility of truncating the solution of the direction finding subproblems prematurely. Together with the section on inexact line searches, the analysis made in this section enables the practical realization of cost approximation algorithms. Convergence is established under arbitrarily early truncations of the solutions of the subproblems, provided that the algorithm chosen for the inexact solution of the subproblem has a closed algorithmic map, and that at least one iteration is performed on each subproblem. The similarity in the properties of the search directions makes it possible to perform the analysis in a uniform fashion for all the problem classes. For the first three problem areas we also show that it is possible to combine inexact solutions of the subproblems and line searches.

Linear convergence of cost approximation algorithms. For unconstrained optimization, the characterization of cost approximation algorithms as gradient related methods is used to establish a linear rate of convergence under various step length rules. By choosing affine cost approximating maps, superlinearly or quadratically convergent quasi-Newton algorithms are obtained through previous analyses of deflected gradient methods.

For the other three problem classes, the above analysis can not be extended, due to either the nondifferentiability or to the presence of constraints. However, for affine cost approximating maps, a linear or superlinear rate of convergence is established for all the three problems, using various step length rules.

Finite identification of active constraints. For the efficient solution of large-

Mey76, Tis77, FiH79, Hua79, Mey79, TPK79, Szi90, HiS92]; although some of these may be applied to the cost approximation algorithm, we will not pursue the most general convergence conditions possible.

scale constrained problems, it is of advantage to be able to identify the set of constraints that are active at a solution, within a finite number of iterations. If the constraints are linear, this property implies that the iterates eventually lie in an essentially unconstrained set, and fast local convergence is achieved. Using recent results on the relations between such identification properties and the concept of projected gradients, the cost approximation algorithm is shown to finitely identify the active constraint set, under a nondegeneracy assumption on the limit point. Similar results have earlier been established for some instances of cost approximation algorithms, using different nondegeneracy conditions. In order to compare the previous results to those obtained here, we establish relationships among the different nondegeneracy conditions; the condition used here is the weakest among those used previously, and, as a consequence, several earlier identification results are strengthened. The above results are essentially equivalent for differentiable constrained optimization and variational inequalities; they have, however, not yet been investigated for nondifferentiable problems.

Finite convergence of cost approximation algorithms. By strengthening the nondegeneracy condition used in the previous section, the sequence of subproblem solutions is, for differentiable constrained optimization and variational inequalities, shown to reach a solution finitely. The strengthening essentially requires that the limit point must be an extreme point of the feasible set. The finiteness result is also extended to nondifferentiable problems; here, the condition is that the limit point is a point of nondifferentiability, and that the merit function grows faster than linear from this point.

A posteriori error bounds and lower bounds. For use in termination criteria in Step 5 of the algorithm, it is of interest to obtain a calculable measure of the distance to the set of solutions. In this section we establish two types of such measures. The first is an upper bound on the Euclidean distance to a solution. This measure is proportional to the distance between a point and a solution to the corresponding subproblem. The analysis is essentially uniform for all problem classes.

The second measure is a lower bound on the optimal value of the merit function. This measure is a generalization of the well known Frank-Wolfe bound for convex problems. For variational inequality problems, the merit functions considered have a known minimal value of zero.

Jacobi and Gauss-Seidel type methods. Many important applications of [**GVIP**] have a Cartesian product structure in their feasible sets. In this section we show that the cost approximation algorithms can effectively exploit such a problem structure, by choosing the cost approximating maps separable with respect to the feasible set. The resulting independent subproblems can either be solved in sequence or in parallel, and the cost approximation algorithm can therefore be tailored for use on different computer architectures. In the Jacobi type parallel algorithm, the independent subproblems are solved simultaneously by parallel processors. The convergence characteristics of this version of the cost approximation algorithm is the same as those of the original one. In the Gauss-Seidel type sequential algorithm, the independent sets of variables are iterated upon in a sequential manner. These two principles enable the solution of large-scale problems, through a decomposition into sequences of problems in smaller dimensions. They are considered for all problem classes; the convergence analysis is uniform, and includes the use of simple step length rules as well as truncated subproblems. In a third algorithm, we consider a partially asynchronous parallel version of the cost approximation algorithm for differentiable optimization problems. In such an algorithm, the information used in the solution of a subproblem is allowed to be outdated. Convergence is established for the use of predetermined step lengths; the maximal allowed step length is inverse proportional to an upper bound on the degree to which any information may be outdated.

Summary and further research. In this section we make a summary of the most important results obtained in each chapter, and discuss possibilities for further research, including both theoretical developments and practical investigations.

We conclude with some remarks about the presentation.

The cost approximating mapping will be used interchangeably in three different forms. In the most general case, a sequence $\{\mathbf{\Phi}^k\}$ of mappings, defined on X, is used. The second most general case is where this sequence is defined by a continuous cost approximating mapping on $X \times X$ of the form $\mathbf{\Phi}(\mathbf{x}, \mathbf{y})$ [giving $\mathbf{\Phi}^k(\mathbf{x}) = \mathbf{\Phi}(\mathbf{x}, \mathbf{x}^k)$]. The least general form is where the cost approximating mapping is a fixed function on X, i.e., a mapping of the form $\mathbf{\Phi}(\mathbf{x})$. Whenever the iteration dependency is not crucial to the analysis, the superscript k will be suppressed. Occasionally, we will simplify also other notations, for instance by dropping arguments when they are not of any significance in the context.

Most results are derived using a sequence of functions $\{\varphi^k\}$, rather than a (more general) sequence of cost approximating maps $\{\mathbf{\Phi}^k\}$, the reason being that it enables a more simple comparison with earlier performed analyses, and also because it simplifies the presentation. It should be noted that results that do not rely on the integrability of the gradients $\nabla\varphi^k$ can be more generally stated using a sequence $\{\mathbf{\Phi}^k\}$; the corresponding statements may be established with few or no additional assumptions.[11]

Most convergence results obtained demand the sequences $\{\mathbf{x}^k\}$ and $\{\mathbf{y}^k\}$ to be bounded. This is ensured by using different assumptions in different contexts, for instance compact feasible sets or level sets, or coercivity of φ^k. The reader should note that it is often possible to replace one such compactness assumption with another one, without affecting the result.

The cost approximation methods are generally only asymptotically convergent; therefore, we shall frequently (but implicitly) presume that the sequence of iterates is infinite. Further, we do not study the behaviour of the algorithms when a solution does not exist.

We limit ourselves to the finite-dimensional case; we point out, however, that the results obtained in this thesis can be extended to infinite-dimensional vector spaces by appropriate modifications.

1.4 Contributions of the thesis—a summary

Unifying frameworks are of significant scientific value, both as a means of summarizing existing knowledge and to provide tools for investigating relationships and performing analyses. This thesis presents a framework of descent methods for continuous mathematical programming problems. A common property of the methods included in the framework is that the search direction finding subproblems can be characterized by cost approximating mappings, which replace the original cost mapping in an iterative manner. The number of methods that can be described within the framework is large, and included are many important—and quite diverse—algorithms.

[11] The two most important exceptions are the results on truncated methods for the subproblems, and on descent methods for variational inequalities.

The main contribution of the thesis is the construction of the framework of cost approximation methods. This new concept contributes to the understanding and interpretation of—seemingly unrelated—iterative methods for different problem classes, and is, in our view, a natural and very elegant way of describing them. Benefits that are gained from the construction of the framework are the possibilities of relating algorithms in several ways, e.g., with respect to requirements for guaranteeing convergence, and a cross-fertilization among the problem areas where the methods included have originally been developed.

The framework encompasses many classes of algorithms which have previously been presented for nonlinear programs and variational inequalities, such as those given by Cohen, Dafermos, Migdalas, and Tseng; the establishment of relationships among these method classes is another major contribution of the thesis, and shows that researchers in different fields of optimization have been following parallel paths.

In the extensive convergence analysis made, the prerequisites and assumptions are based on the use of standard concepts in optimization theory and convex analysis, which have the advantage of being, in many cases, easily verified. Furthermore, the use of standard concepts facilitates comparisons with previous analyses. It is important to observe that, despite the size of the framework, a comparison with the results obtained both for the above mentioned method classes and from specialized analyses reveals that our analysis is strong enough to either reproduce previous results, or, in some cases, strengthen them. Because of the unification of the results, it is possible to easily compare both the convergence conditions for different algorithms for a particular problem, and the convergence conditions for a particular method when applied to problems in different areas.

Inherent in the framework are also other unifications of existing results. For example, we are able to show that results on the finite identification of optimal active constraints, and the finite convergence, of many iterative methods are special cases of general results for the cost approximation algorithm; we actually strengthen some of the results previously obtained. As a byproduct of this analysis, we establish some relationships among non-degeneracy and regularity conditions on stationary points. An additional benefit of the construction of the framework is a unification of hitherto proposed optimization reformulations of asymmetric variational inequalities, and of descent methods based on these reformulations.

We present several important results concerning practical realizations of cost approximation methods. We show that convergence is ensured also when the process of determining a search direction is truncated prematurely, and that simple step length rules can be used in place of an exact line search. These features are of great practical interest since, in the general case, subproblems are only asymptotically solvable, and line searches can not be performed exactly. Also, the, in some applications, computationally expensive objective function evaluations, can, in fact, be eliminated entirely, by using predetermined step lengths. The importance of these results is emphasized by the fact that they can be exploited to improve existing, frequently applied, methods which are instances of the cost approximation algorithm.

A particularly useful property of the class of cost approximation methods is its ability to adapt to problem structures; this is crucial to these methods' capability of solving large-scale applications. An important example is the Cartesian product structure of the feasible set, which is inherent in many applications of network and economic equilibria. For this type of problem it is possible to construct efficient decomposition methods by choosing the cost approximating mapping to match the separability of the feasible set.

We show how such decomposition schemes can be implemented on different computer architectures with varying degrees of parallelism and synchronization.

As has already been mentioned, the algorithm framework of G. Cohen, the auxiliary problem principle, has a very strong relationship to the class of cost approximation methods, and is, in fact, included in it. Among the algorithm classes to which we establish close relationships, the auxiliary problem principle is by far the most well studied (excluding deflected gradient methods), and in a series of papers (which are remarkably scarcely referred to), Cohen and associates have made considerable contributions in algorithm theory in a wide variety of areas, such as in decomposition/coordination methods in differentiable and nondifferentiable optimization [Coh72, Coh76, CoJ76, Coh78, Coh80, CoB80, Coh82, Zhu82, CoZ83, Coh84, CoZ84, Zhu87, CoM90], saddle point problems [Coh76, Coh78, Coh80, CoB80], decentralized optimal control and dynamical systems [CoJ76, Coh77, Coh78, JoC78], stochastic optimization [CoC85, CoC86b, CoC87, CuI87, CuC90], Nash equilibria [CoC85, CoC86a, Coh87], and variational inequalities [Coh88, CoC88]. Although it is far beyond the scope of this thesis to place all these results within the frame of cost approximation, we would like to stress that the close relationship between these two frameworks, together with the results obtained in this thesis, bring forward a large number of potential applications, extensions and interesting further research topics in the mentioned areas.

The scope of the thesis is the theoretical development of the framework of cost approximation algorithms; consequently, the thesis does not contain any computational investigations, although these new theoretical findings certainly lead to many opportunities for interesting future research along those lines.

The thesis is partly based on, and continues the work in, the following articles.

[LMP89] T. Larsson, A. Migdalas, and M. Patriksson, *A partial linearization method for the traffic assignment problem.* Optimization (1993) (in press).

[LaP90] T. Larsson and M. Patriksson, *A class of gap functions for variational inequalities.* To appear in Mathematical Programming.

[Pat91a] M. Patriksson, *Partial linearization methods in nonlinear programming*, Journal of Optimization Theory and Applications, 78 (1993) (in press).

[Pat91b] ——, *A unified description of iterative algorithms for traffic equilibria*, European Journal of Operational Research (1993) (in press).

"A splendid time is guaranteed for all ..."

Chapter 2

Unconstrained optimization

"It was therefore deducible that he must have left (a) by way of a rope to the roof, or (b) by means of climbing up inside the chimney to the roof. This was a simple mathematical certainty. If $PQ = pq$, it is therefore quite obvious that $PQ = pq + p\beta + qa + a\beta$."

JOHN DICKSON CARR
The Three Coffins (1935)

In this chapter we study [**GVIP**] in its simplest form, where $X = \Re^n$, $\mathbf{U} \equiv \mathbf{0}$ and $\mathbf{F}$ is a continuous gradient mapping, i.e., there exists a function $f : \Re^n \mapsto \Re$ in C^1 with $\mathbf{F} \equiv \nabla f$. Under these assumptions [**GVIP**] reduces to the problem of finding an $\mathbf{x}^* \in \Re^n$ such that

$$\nabla f(\mathbf{x}^*) = \mathbf{0}. \tag{2.1}$$

We let Ω denote the set of solutions to (2.1). It is well known (see, e.g., [OrR70, Th. 4.1.3]) that (2.1) defines the *necessary* optimality conditions for the problem

[**NLP**] $$\min_{\mathbf{x} \in \Re^n} f(\mathbf{x}),$$

and that it is also *sufficient* for the global optimality of $\mathbf{x}^*$ in [**NLP**] whenever f is pseudoconvex [BaS79, p. 106]. These facts motivate the use of $\psi \equiv f$ as merit function in the cost approximation algorithm.

2.1 The cost approximation algorithm

Below we specialize the algorithm of Section 1.2 for [**GVIP**] to the problem (2.1).

The cost approximation algorithm

A sequence $\{\mathbf{\Phi}^k\}$ of monotone cost approximating mappings is assumed to be given.

Step 0 (*Initial guess*) Choose an initial point $\mathbf{x}^0 \in \Re^n$, and let $k = 0$.

Step 1 (*Search direction generation*) Find a solution $\mathbf{y}^k \in \Re^n$ to the equation system

$$\mathbf{\Phi}^k\left(\mathbf{y}^k\right) + \nabla f\left(\mathbf{x}^k\right) - \mathbf{\Phi}^k\left(\mathbf{x}^k\right) = \mathbf{0}. \tag{2.2}$$

The resulting search direction is $\mathbf{d}^k = \mathbf{y}^k - \mathbf{x}^k$.

Step 2 (*Convergence check*) If $\mathbf{x}^k$ satisfies (2.2) $\rightarrow$ Stop ($\mathbf{x}^k \in \Omega$). Otherwise, continue.

Step 3 (*Line search*) Find a step length, t_k, which solves the one-dimensional problem

$$t_k \in \arg\min_{t \geq 0} \; f\left(\mathbf{x}^k + t\mathbf{d}^k\right).$$

Step 4 (*Update*) Let $\mathbf{x}^{k+1} = \mathbf{x}^k + t_k\mathbf{d}^k$, and $k := k+1$.

Step 5 (*Convergence check*) If $\mathbf{x}^k$ is acceptable as a solution[1] $\rightarrow$ Stop. Otherwise, go to Step 1.

In most known algorithms for [**NLP**] that we identify as special cases of cost approximation methods, the search direction finding subproblem (2.2) corresponds to a convex minimization problem. Introducing a sequence $\{\varphi^k\}$ of convex functions $\varphi^k : \Re^n \mapsto \Re$ in C^1, with $\nabla\varphi^k \equiv \mathbf{\Phi}^k$, the subproblem (2.2) reduces to the problem of finding

$$[\mathbf{NLP}^k_{\varphi^k}] \qquad \mathbf{y}^k \in \arg\min_{\mathbf{x} \in \Re^n} \; f_{\varphi^k}(\mathbf{x}),$$

where

$$f_{\varphi^k}(\mathbf{x}) = \varphi^k(\mathbf{x}) + f\left(\mathbf{x}^k\right) - \varphi^k\left(\mathbf{x}^k\right) + \left[\nabla f\left(\mathbf{x}^k\right) - \nabla\varphi^k\left(\mathbf{x}^k\right)\right]^{\mathrm{T}}\left(\mathbf{x} - \mathbf{x}^k\right).$$

Recall from Section 1.2 that the cost mapping of [$\mathbf{GVIP}^k_{\mathbf{\Phi}^k}$] is obtained through a fixation of an additive part of the original cost mapping at the current point, i.e., the cost $[\mathbf{\Phi}^k + \mathbf{U}] + [\mathbf{F} - \mathbf{\Phi}^k]$ is approximated by $[\mathbf{\Phi}^k + \mathbf{U}] + [\mathbf{F}(\mathbf{x}^k) - \mathbf{\Phi}^k(\mathbf{x}^k)]$. The corresponding interpretation of the approximation of the objective f is the following. Writing the objective f as $\varphi^k + [f - \varphi^k]$, the approximate objective f_{φ^k} is obtained through a first-order Taylor expansion of the second term around $\mathbf{x}^k$. This interpretation of the cost approximation algorithm as a *partial linearization* method was made in [Pat91a, Pat91b].

2.2 Instances of cost approximation algorithms

We show below that some well known methods for unconstrained nonlinear programming problems can be obtained through appropriate choices of functions φ^k.

Choosing $\varphi^k(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, and using that the optimality conditions (2.2) of [**NLP**] is

$$\nabla f\left(\mathbf{x}^k\right) = \nabla\varphi^k\left(\mathbf{x}^k\right) - \nabla\varphi^k\left(\mathbf{y}^k\right), \tag{2.3}$$

results in $\mathbf{d}^k = -\nabla f(\mathbf{x}^k)$, i.e., the search direction in the method of *steepest descent*. This method was introduced by Cauchy [Cau47], and has since then been extensively studied.

Choosing $\varphi^k(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2_{\mathbf{B}_k} = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{B}_k\mathbf{x}$, for some symmetric and positive definite matrix $\mathbf{B}_k$, results in the search direction $\mathbf{d}^k = -\mathbf{B}_k^{-1}\nabla f(\mathbf{x}^k)$. This is the step direction in the class of *deflected* (or *scaled*) *gradient* methods [Pol63, OrR70, Pol71, PsD78, Coh81, Ber82a]. Two important special cases of deflected gradient methods are the steepest descent method ($\mathbf{B}_k = \mathbf{I}$, $\forall k$) and *Newton's method* ($\mathbf{B}_k = \nabla^2 f(\mathbf{x}^k)$, $\forall k$).

[1] Typical stopping criteria are a small step length, a small decrease in the objective value between two iterations, a small relative error in the objective, or a gradient of short length.

Remark 2.1 Newton's method illustrates that $\{\varphi^k\}$ may depend on f, although no such relation will be assumed in our convergence analyses. (In Newton's method, the convexity requirement of φ^k implies that f must be a convex function in C^2.) For certain instances of cost approximation methods, conditions on f may therefore be implicitly imposed through the conditions on φ^k.

Since *any* deflected gradient direction may be obtained through the appropriate choice of the matrix $\mathbf{B}_k$ in the definition of φ^k above, the following algorithmic equivalence result is immediate.

Theorem 2.1 (An algorithmic equivalence result) [2] *Let $\{\varphi^k\}$ be a sequence of functions defined by $\varphi^k(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{B}_k\mathbf{x}$, where each matrix $\mathbf{B}_k$ is symmetric and positive definite. Then the resulting class of cost approximation algorithms is equivalent to the class of deflected gradient methods, in the sense that the search directions obtained are the same.*

This equivalence result implies that the class of cost approximation methods includes superlinearly and quadratically convergent methods (see Theorem 2.5).

2.3 Analysis of the direction finding subproblem

Lemma 2.1 (Descent lemma) *Let $\varphi : \Re^n \mapsto \Re$ be a convex function in C^1. Let $\mathbf{x} \in \Re^n$, $Y(\mathbf{x})$ be the (possibly empty) set of optimal solutions to $[\mathbf{NLP}_\varphi]$, $\mathbf{y} \in Y(\mathbf{x})$, and $\mathbf{d} = \mathbf{y} - \mathbf{x}$.*

(a) $\mathbf{x} \in \Omega \Leftrightarrow \mathbf{x} \in Y(\mathbf{x}) \Leftrightarrow f_\varphi(\mathbf{y}) = f_\varphi(\mathbf{x})$.

(b) *Let $\overline{\mathbf{x}} \in \Re^n$ be any point such that $f_\varphi(\overline{\mathbf{x}}) < f_\varphi(\mathbf{x})$. Then $\nabla f(\mathbf{x})^{\mathrm{T}}(\overline{\mathbf{x}} - \mathbf{x}) < 0$. Especially, for $\overline{\mathbf{x}} = \mathbf{y}$, if $\mathbf{x} \notin \Omega$, then $\nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d} < 0$.*

(c) *Let φ be strictly convex. If $Y(\mathbf{x})$ is nonempty, then it is singleton.*

(d) *Let φ be strongly convex with modulus m_φ. Then $Y(\mathbf{x})$ is nonempty and singleton, and*

$$\nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d} \leq -m_\varphi \|\mathbf{d}\|^2. \tag{2.4}$$

(e) *Let φ be strongly convex, and $\nabla\varphi$ be Lipschitz continuous with modulus $M_{\nabla\varphi}$. If $\mathbf{x} \notin \Omega$, then*

$$-\frac{\nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d}}{\|\nabla f(\mathbf{x})\|\,\|\mathbf{d}\|} \geq \rho, \tag{2.5}$$

where $\rho = m_\varphi / M_{\nabla\varphi}$.

Proof

(a) $\mathbf{x} \in \Omega \Leftrightarrow \nabla f(\mathbf{x}) = \mathbf{0} \Leftrightarrow \nabla f(\mathbf{x}) = \nabla\varphi(\mathbf{x}) - \nabla\varphi(\mathbf{x}) \Leftrightarrow \mathbf{x} \in Y(\mathbf{x}) \Leftrightarrow f_\varphi(\mathbf{y}) = f_\varphi(\mathbf{x})$.

[2] It is also possible to show that the obtained search directions are deflected gradients for arbitrary, strongly convex functions φ^k in C^2 with Lipschitzian gradients; in this more general case, however, the corresponding matrices $\mathbf{B}_k$ are not necessarily symmetric (cf. (2.3) and [OrR70, Sec. 3.2]).

(b) By the convexity of φ, $f_\varphi(\overline{\mathbf{x}}) < f_\varphi(\mathbf{x})$ implies $\nabla f(\mathbf{x})^{\mathrm{T}}(\overline{\mathbf{x}} - \mathbf{x}) < 0$. The point $\mathbf{y}$ is a special case of such an $\overline{\mathbf{x}}$.

(c) Follows from the strict convexity of f_φ, resulting in a unique solution to $[\mathbf{NLP}_\varphi]$.

(d) The first result is immediate from the strong convexity of f_φ. We also have

$$\begin{aligned} \nabla f(\mathbf{x})^{\mathrm{T}}(\mathbf{y} - \mathbf{x}) &= [\nabla\varphi(\mathbf{x}) - \nabla\varphi(\mathbf{y})]^{\mathrm{T}}(\mathbf{y} - \mathbf{x}) \quad (2.3) \\ &\leq -m_\varphi \|\mathbf{y} - \mathbf{x}\|^2. \quad \quad (\mathrm{A.11}) \end{aligned}$$

(e) From (2.3) and the Lipschitz continuity assumption it follows that $\|\nabla f(\mathbf{x})\| \leq M_{\nabla\varphi}\|\mathbf{d}\|$. This inequality, together with (2.4), gives that (2.5) holds. □

Property **(a)** validates the termination criterion of Step 2, and provides a *fixed point* characterization of the set of points satisfying (2.1). The descent properties **(b)** and **(d)** are especially important in inexact implementations. While **(b)** enables inexact solutions to $[\mathbf{NLP}_\varphi]$,[3] **(d)** enables inexact line search strategies, such as Armijo-type rules or predetermined step lengths. The Inequality (2.5) in the result **(e)** characterizes cost approximation methods as being *gradient related* [OrR70, Def. 14.3.1]; see also [Pol71, Coh81]. This characterization is used in Theorems 2.3.a and 2.4.a–b to establish convergence results for the class of cost approximation methods.

Below, we show that the search direction mapping $\mathbf{x} \mapsto D(\mathbf{x}) \stackrel{\text{def}}{=} Y(\mathbf{x}) - \mathbf{x}$ is closed on $\Re^n$, whenever φ is a continuous function of the form $\varphi(\mathbf{x}, \mathbf{y})$, convex and differentiable in $\mathbf{x}$. Under a level set compactness assumption, using Zangwill's [Zan69b] classical Theorem A, it is then possible to show that any accumulation point of the sequence $\{\mathbf{x}^k\}$ must satisfy (2.1) [see Theorem 2.2].

Lemma 2.2 (Continuity properties of the search direction) *Let $\varphi : \Re^{2n} \mapsto \Re$ be a continuous function on $\Re^{2n}$ of the form $\varphi(\mathbf{x}, \mathbf{y})$, convex and in C^1 with respect to $\mathbf{x}$.*

(a) *The mapping $\mathbf{x} \mapsto D(\mathbf{x})$ is closed on $\Re^n$.*

(b) *The set $Y(\mathbf{x})$ is nonempty and bounded*[4] *if and only if*

$$-\nabla f(\mathbf{x}) + \nabla_{\mathbf{x}}\varphi(\mathbf{x}, \mathbf{x}) \in \mathtt{int}\ (\mathtt{dom}\ \varphi^\circ(\cdot, \mathbf{x})), \tag{2.6}$$

where $\varphi^\circ(\cdot, \mathbf{x})$ is the conjugate of $\varphi(\cdot, \mathbf{x})$.

(c) *If, in addition to the assumption that (2.6) holds for every $\mathbf{x} \in \Re^n$, $\varphi(\cdot, \mathbf{x})$ is strictly convex for every $\mathbf{x} \in \Re^n$, then the mapping $\mathbf{x} \mapsto D(\mathbf{x})$ is continuous on $\Re^n$.*

Proof

(a) For an arbitrary $\mathbf{x} \in \Re^n$, let $\{\mathbf{x}^k\} \to \mathbf{x}$ and $\{\mathbf{y}^k\} \to \mathbf{y}$, with $\mathbf{y}^k \in Y(\mathbf{x}^k)$. For all k we have [cf. (2.3)],

$$\nabla f\left(\mathbf{x}^k\right) = \nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^k, \mathbf{x}^k\right) - \nabla_{\mathbf{x}}\varphi\left(\mathbf{y}^k, \mathbf{x}^k\right).$$

[3] See Theorem 4.4 for a convergence analysis.
[4] From **(a)** the mapping $\mathbf{x} \mapsto D(\mathbf{x})$ is u.s.c. on any compact subset of $\Re^n$; see [Nik68, Le. 4.4].

In the limit when $k \to \infty$ we obtain, from the continuity of ∇f and $\nabla_{\mathbf{x}}\varphi$,

$$\nabla f(\mathbf{x}) = \nabla_{\mathbf{x}}\varphi(\mathbf{x}, \mathbf{x}) - \nabla_{\mathbf{x}}\varphi(\mathbf{y}, \mathbf{x}), \tag{2.7}$$

which, by the convexity of $\varphi(\cdot, \mathbf{x})$, yields that $\mathbf{y} \in Y(\mathbf{x})$. It follows that $\mathbf{x} \mapsto Y(\mathbf{x})$ is closed on $\Re^n$.

To prove that also $\mathbf{x} \mapsto D(\mathbf{x})$ is closed on $\Re^n$, we let

$$\texttt{graph}\ Y \stackrel{\text{def}}{=} \left\{(\mathbf{x}, \mathbf{y}) \in \Re^{2n} \mid \mathbf{y} \in Y(\mathbf{x})\right\}$$

denote the graph of the subproblem solution defining mapping $Y : \Re^n \mapsto 2^{\Re^n}$. From [Roc70a, Th. 24.4], this set is closed in $\Re^{2n}$, and since the graph of the direction finding map $D : \Re^n \mapsto 2^{\Re^n}$ defined by $\mathbf{x} \mapsto D(\mathbf{x}) = Y(\mathbf{x}) - \mathbf{x}$,

$$\texttt{graph}\ D \stackrel{\text{def}}{=} \left\{(\mathbf{x}, \mathbf{y} - \mathbf{x}) \in \Re^{2n} \mid \mathbf{y} \in Y(\mathbf{x})\right\},$$

is obtained from $\texttt{graph}\ Y$ through an affine transformation, it then follows from [Ber63, p. 111] or [Nik68, p. 65] that $\mathbf{x} \mapsto D(\mathbf{x})$ is closed on $\Re^n$, as desired.

(b) $Y(\mathbf{x})$ is the set of points $\mathbf{y}$ satisfying (2.7). From [Roc70a, Cor. 23.5.1 and Th. 27.1.b],

$$Y(\mathbf{x}) = \partial\varphi^\circ\left(-\nabla f(\mathbf{x}) + \nabla_{\mathbf{x}}\varphi(\mathbf{x}, \mathbf{x}), \mathbf{x}\right)$$

[see also (1.10)]. By [Roc70a, Ths. 23.4 and 27.1.d], $Y(\mathbf{x})$ is nonempty and bounded if and only if (2.6) holds (see also Remark 1.1).

(c) Follows directly from **(a)**, and from the fact that the mapping is point-to-point, since $Y(\mathbf{x})$ is singleton (cf. Lemma 2.1.c). □

2.4 Basic convergence of cost approximation algorithms

We begin by quoting two lemmas, which will be used in the convergence analysis.

Lemma 2.3 (Convergence toward a set) *Let $S \subseteq T \subset \Re^n$, where S and T are nonempty sets, and T is compact. Let $\{\mathbf{x}^k\}$ be a sequence in T, such that every accumulation point belongs to S. Then the sequence $\{\mathbf{x}^k\}$ converges to S, in the sense that*

$$\left\{\inf_{\mathbf{x}\in S} \left\|\mathbf{x}^k - \mathbf{x}\right\|\right\} \to 0. \tag{2.8}$$

Proof See [OrR70, Th. 14.1.4]. □

Lemma 2.4 (Convergence toward a point) *Assume that the sequence $\{\mathbf{x}^k\}$ is bounded, and that the set of accumulation points, Ω, consists of a finite number of points. Then if $\{\mathbf{x}^{k+1} - \mathbf{x}^k\} \to \mathbf{0}$, the sequence $\{\mathbf{x}^k\}$ converges to one of the points in Ω.*

Proof See [Ost66, Th. 28.1] or [OrR70, Th. 14.1.5]. □

Below, we state our first convergence result for cost approximation methods. Using the closedness of the direction finding map $\mathbf{x} \mapsto D(\mathbf{x})$ [Lemma 2.2.a], Zangwill's well known Theorem A is applied to show that, under the exact line search Rule M, any accumulation point $\mathbf{x}^\infty$ of $\{\mathbf{x}^k\}$ satisfies (2.1), i.e., $\mathbf{x}^\infty \in \Omega$.

Theorem 2.2 (Basic convergence of cost approximation algorithms) *Let $\varphi : \Re^{2n} \mapsto \Re$ be a continuous function on $\Re^{2n}$ of the form $\varphi(\mathbf{x}, \mathbf{y})$, convex and in C^1 with respect to $\mathbf{x}$. Assume that the point $\mathbf{x}^0$ is chosen so that the level set $L^f_{\Re^n}(\mathbf{x}^0)$ is bounded, and assume further that the problem $[\mathbf{NLP}_\varphi]$ is well defined, in the sense that $Y(\mathbf{x})$ is nonempty and bounded for every $\mathbf{x} \in \Re^n$. Then, under Rule M, $\{f(\mathbf{x}^k)\} \to f(\overline{\mathbf{x}})$ for some $\overline{\mathbf{x}} \in \Omega$, any accumulation point $\mathbf{x}^\infty$ of the sequence $\{\mathbf{x}^k\}$ (at least one such point exists) lies in Ω, and*

$$\left\{ \inf_{\mathbf{x} \in \Omega \cap L^f_{\Re^n}(\mathbf{x}^0)} \left\| \mathbf{x}^k - \mathbf{x} \right\| \right\} \to 0. \tag{2.9}$$

Proof In order to apply Zangwill's Theorem A [Zan69b, Sec. 4.5], we first identify the *solution set* with Ω, the *descent function* with f, and the *algorithmic map* with the composite map $A = MD$, where D is the direction finding map and M is the exact line search map. To fulfill the assumptions of Theorem A, we next show that **(1)** the sequence $\{\mathbf{x}^k\}$ lies in a compact set, **(2)** $f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$ if $\mathbf{x}^k \notin \Omega$ and the algorithm terminates if $\mathbf{x}^k \in \Omega$, and **(3)** the mapping A is closed at all points not in Ω.

(1) Because of the level set compactness assumption and the descent property of the algorithm, the sequence $\{\mathbf{x}^k\}$ lies in a compact set.

(2) If $\mathbf{x}^k \notin \Omega$, then the descent property follows from Lemma 2.1.b. The termination property ($\mathbf{x}^k \in \Omega$) follows from Lemma 2.1.a.

(3) The mapping D is closed on $L^f_{\Re^n}(\mathbf{x}^0)$ from Lemma 2.2.a, and the same is true for the mapping M [Zan69b, Le. 5.1]. Since $L^f_{\Re^n}(\mathbf{x}^0)$ is bounded, [Zan69b, Cor. 4.2.1] yields that the map $A = MD$ is closed on $L^f_{\Re^n}(\mathbf{x}^0)$.

The assumptions of Theorem A are thus fulfilled, and we conclude that $\{f(\mathbf{x}^k)\}$ converges to $f(\overline{\mathbf{x}})$ for some $\overline{\mathbf{x}} \in \Omega$, and any accumulation point $\mathbf{x}^\infty$ of $\{\mathbf{x}^k\}$ lies in Ω. (The existence of at least one such point follows from the boundedness of $\{\mathbf{x}^k\}$.) The last statement follows from Lemma 2.3, identifying T with the level set $L^f_{\Re^n}(\mathbf{x}^0)$, and S with the set $\Omega \cap L^f_{\Re^n}(\mathbf{x}^0)$. □

From several points of view the condition of compactness imposed on the subproblem solutions, characterized in Lemma 2.2.b, is quite natural. One example of a sufficient condition for (2.6) to hold is that $\varphi(\cdot, \mathbf{x})$ is *coercive* (see Definition A.3.b). Then, it is also *cofinite* [Roc70a, p. 116 and 259], and $\texttt{dom}\,\varphi^\circ(\cdot, \mathbf{x}) = \Re^n$ holds [Roc70a, Cor. 13.3.1]; thus, (2.6) follows (see also [HaS66, Mor74] and Remark 1.1). In addition, it does not seem to be possible to weaken this compactness assumption within Zangwill's framework. We finally note that this requirement has the clear advantage of being sufficient also for the more complex problems under study in the thesis, and it therefore enables a consistent presentation.

Another sufficient condition for the subproblems to be well defined is that the functions φ^k are strongly convex (since strong convexity implies coercivity). Based on this assumption, we show below that convergence of the cost approximation algorithm is ensured, by using known results from gradient related methods. We also establish convergence using various step length rules. For definitions of the step length rules, see Appendix A.

Theorem 2.3 (Basic convergence of cost approximation algorithms) [5] *Let $\varphi^k : \Re^n \mapsto \Re$ be a strongly convex function in C^2, and let $\nabla\varphi^k$ be Lipschitz continuous, for all k. Let $0 < m_\varphi \leq m_{\varphi^k}$ and $M_{\nabla\varphi^k} \leq M_{\nabla\varphi} < +\infty$, $\forall k$.*

(a) *Assume that the point* $\mathbf{x}^0$ *is chosen so that the level set* $L^f_{\Re^n}(\mathbf{x}^0)$ *is bounded. Then, under Rules M and G, any accumulation point* $\mathbf{x}^\infty$ *of the sequence* $\{\mathbf{x}^k\}$ *(at least one such point exists) lies in* Ω. *Furthermore,* $\{\nabla f(\mathbf{x}^k)\} \to \mathbf{0}$, *and (2.9) holds. If further* Ω *is finite, and the step lengths* t_k *are upper bounded,* $\{\mathbf{x}^k\}$ *converges to one point in* Ω.

(b) *Assume that* f *is bounded from below and that* ∇f *is Lipschitz continuous. Then, from any starting point* $\mathbf{x}^0 \in \Re^n$, *under Rules M and A,* $\{\nabla f(\mathbf{x}^k)\} \to \mathbf{0}$. *Furthermore, whenever* Ω *is nonempty, (2.9) holds. If further* $\{\mathbf{x}^k\}$ *is bounded,* Ω *is finite, and the step lengths* t_k *are upper bounded,* $\{\mathbf{x}^k\}$ *converges to one point in* Ω. *For Rule P, the same conclusions hold if either* t_k *belongs to a compact subset of* $(0, \frac{2m_{\varphi^k}}{M_{\nabla f}})$, *or* $t_k \equiv t \in (0, \frac{2m_\varphi}{M_{\nabla f}})$.

Proof

(a) The search directions are gradient related with $\rho \geq \frac{m_\varphi}{M_{\nabla\varphi}}$ (cf. Lemma 2.1.e). Then $\{\nabla f(\mathbf{x}^k)\} \to \mathbf{0}$ follows from [OrR70, Th. 14.2.10 and p. 495] (Rule M) and [OrR70, Th. 14.2.14 and p. 495] (Rule G).

The validity of (2.9) follows from Lemma 2.3.

From $\{\nabla f(\mathbf{x}^k)\} \to \mathbf{0}$, (2.3), and the lower bound $m_\varphi > 0$, $\{\mathbf{d}^k\} \to \mathbf{0}$ follows. Convergence to one point in Ω then follows from Lemma 2.4 and the inequality

$$\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\| = t_k \left\|\mathbf{d}^k\right\| \leq \bar{t}\left\|\mathbf{d}^k\right\|. \tag{2.10}$$

(b) We begin by showing that for all the three rules we must have that $\{\mathbf{d}^k\} \to \mathbf{0}$.

Rule M From the Lipschitz continuity of ∇f and the strong convexity of φ^k, we obtain that, for $t \geq 0$,

$$f\left(\mathbf{x}^{k+1}\right) \leq f\left(\mathbf{x}^k + t\mathbf{d}^k\right) \leq f\left(\mathbf{x}^k\right) + \left(-tm_{\varphi^k} + \frac{M_{\nabla f}}{2}t^2\right)\left\|\mathbf{d}^k\right\|^2. \tag{2.11}$$

Minimizing the right-hand side of this inequality, we obtain that $t = \bar{t} = \frac{m_{\varphi^k}}{M_{\nabla f}}$,[6] and hence that

$$f\left(\mathbf{x}^{k+1}\right) - f\left(\mathbf{x}^k\right) \leq -\frac{m_{\varphi^k}^2}{2M_{\nabla f}}\left\|\mathbf{d}^k\right\|^2.$$

Rule A From the Lipschitz continuity of ∇f and the strong convexity of φ^k, we obtain that (cf. Lemma 4.4)

$$f\left(\mathbf{x}^{k+1}\right) - f\left(\mathbf{x}^k\right) \leq -\frac{2\beta\alpha(1-\alpha)m_{\varphi^k}^2}{M_{\nabla f}}\left\|\mathbf{d}^k\right\|^2.$$

Rule P From (2.11) we obtain a decrease in f whenever t_k belongs to a compact subset of $(0, \frac{2m_{\varphi^k}}{M_{\nabla f}})$, or $t_k \equiv t \in (0, \frac{2m_\varphi}{M_{\nabla f}})$.

[5] A related convergence result, for functions $f \in C^2$ and using a variant of Rule N, is given in [OrR70, Th. 14.2.9]. The Lipschitz condition for Rule P in **(b)** is slightly relaxed in [BeT89, Exc. 3.2.1]. Related results have been obtained for other step length rules, see e.g. [OrR70, Ber82a].

[6] It can be shown, through a second-order Taylor expansion of f around $t = t_k$, that the step length obtained in Rule M is bounded from below by this constant.

Since f is lower bounded and, from the above, $\{f(\mathbf{x}^k)\}$ is decreasing, $\{f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k)\} \to 0$. Thus, in all the three rules we obtain that $\{\mathbf{d}^k\} \to \mathbf{0}$.

From (2.3) and the upper bound on $M_{\nabla\varphi^k}$, as $\{\mathbf{d}^k\} \to \mathbf{0}$,

$$\left\{\nabla f\left(\mathbf{x}^k\right)\right\} = \left\{\nabla\varphi^k\left(\mathbf{x}^k\right) - \nabla\varphi^k\left(\mathbf{y}^k\right)\right\} \to \mathbf{0}. \tag{2.12}$$

Assume now that $\{\mathbf{x}^k\}$ has an accumulation point, $\mathbf{x}^\infty$, corresponding to a convergent subsequence $\mathcal{K}$. From (2.12) it then follows that $\nabla f(\mathbf{x}^\infty) = \mathbf{0}$. Thus, $\mathbf{x}^\infty \in \Omega$.

The validity of (2.9) follows from Lemma 2.3. Convergence to one point in Ω follows as in **(a)**. □

2.5 Linear convergence of cost approximation algorithms

Below, we show that for strongly convex functions f in C^2 with a Lipschitz continuous gradient, the sequence $\{\mathbf{x}^k\}$ converges at least linearly (see Definition A.9.a), by utilizing the relationships to the gradient related methods (Rules M and G) and, for quadratic functions φ^k, to the deflected gradient methods (Rules A and P).

Theorem 2.4 (Linear convergence of cost approximation algorithms) [7, 8] *Let f and φ^k, $\forall k$, be strongly convex in C^2 with Lipschitz continuous gradients. Let $0 < m_\varphi \leq m_{\varphi^k}$ and $M_{\nabla\varphi^k} \leq M_{\nabla\varphi} < +\infty$, $\forall k$, and let $\rho = m_\varphi / M_{\nabla\varphi}$.*

(a) *Under Rule M, $\{\mathbf{x}^k\}$ converges at least linearly to the unique minimizer, $\mathbf{x}^*$, of f, with the ratio*

$$q = \sqrt{1 - \left(\rho\frac{m_f}{M_{\nabla f}}\right)^2}.$$

(b) *The conclusion in* **(a)** *holds for Rule G, with $\mu_1 = 1 - \mu_2 = \alpha \in (0, \frac{1}{2})$, with the ratio*

$$q = \sqrt{1 - \left(2\alpha\rho\frac{m_f}{M_{\nabla f}}\right)^2}.$$

(c) *Furthermore, let $\{\varphi^k\}$ be a sequence of functions defined by $\varphi^k(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{B}_k\mathbf{x}$, where each matrix $\mathbf{B}_k$ is symmetric and positive definite. The conclusion in* **(a)** *then holds for Rule A,*[9] *with the ratio*

$$q = \sqrt{1 - 2\gamma m_f^2 M_{\nabla f}},$$

where $\gamma = \min\{\alpha\varepsilon m_\varphi, 2\alpha(1-\alpha)\beta\rho^2/M_{\nabla f}\}$.[10]

(d) *Let $r > 0$ and $R < +\infty$ denote the smallest and largest eigenvalue of $\mathbf{B}_k^{-1}\nabla^2 f\left(\mathbf{x}^k\right)$ over all k, respectively.*[11] *The conclusion in* **(a)** *then holds for Rule P, if t_k belongs to a compact subset of $(0, \frac{2}{R})$. In particular, if $t_k \equiv 2/(R+r)$, then the ratio is*

$$q = \frac{R - r}{R + r}.$$

Proof

(a) Follows from the characterization (2.5) of cost approximation methods as gradient related methods, and the results in [Pol63, Wol69] or [Pol71, pp. 242–246].

(b) Follows as in **(a)**, and from the results in [Wol69] or [Pol71, pp. 247–248].

(c) Follows from the equivalence to deflected gradient methods (Theorem 2.1), and the results in [Coh81] or [PsD78, Ths. II.1.2 and II.1.6].

(d) Follows as in **(c)**, and from the results in [PsD78, p. 55], [BeT89, Prop. 3.2.4] or [CrC55, Gol62, Pol63, Coh81]. □

These convergence results are probably the strongest possible given the assumptions.[12] By imposing further conditions on the functions φ^k, however, so that the obtained search direction approaches the Newton direction in the limit, giving *quasi-Newton* methods, stronger convergence rate results may be achieved. [For these methods, the convergence rate analysis made above is too conservative, since $\{\mathbf{B}_k\}$ is not related to $\nabla^2 f(\mathbf{x}^*)$.] For comparison reasons, we give below the known results for these methods.

Theorem 2.5 (Convergence of quasi-Newton algorithms) [13, 14] *Let f be a strongly convex function in C^2, with a Lipschitz continuous gradient. Let $\{\varphi^k\}$ be a sequence of functions defined by $\varphi^k(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{B}_k\mathbf{x}$, where each matrix $\mathbf{B}_k$ is symmetric and positive definite.*

(a) *Assume that $\{\mathbf{x}^k\} \to \mathbf{x}^*$, where $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Then $\{\mathbf{x}^k\}$ converges superlinearly if and only if*

$$\left\{ \frac{\left\| \left[\frac{1}{t_k}\mathbf{B}_k - \nabla^2 f(\mathbf{x}^*)\right] \left(\mathbf{x}^{k+1} - \mathbf{x}^k\right) \right\|}{\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|} \right\} \to 0. \tag{2.13}$$

If Rule G or A is used, with initial step size 1, there exists a K (depending on $\mathbf{x}^0$) so that for all $k \geq K$ no reduction of the initial step will take place.

---

[7] Since convergence rate results are asymptotic results, the assumptions on f may be relaxed to hold only in a neighbourhood of $\mathbf{x}^*$; cf. [OrR70, Th. 14.3.7 and Exc. E 14.3-7] and [Coh81].

[8] Cohen [Coh81] has obtained tighter bounds on the rate of convergence of gradient related and deflected gradient methods under the additional assumption that $f \in C^3$. Using his results, the convergence ratios for Rules M and G (as $\alpha \to \frac{1}{2}$) can be improved to

$$q = \frac{(M_{\nabla f} - m_f) - (M_{\nabla f} + m_f)\sqrt{1 - (m_\varphi / M_{\nabla\varphi})^2}}{(M_{\nabla f} + m_f) + (M_{\nabla f} - m_f)\sqrt{1 - (m_\varphi / M_{\nabla\varphi})^2}};$$

linear convergence can also be established under Rule N (giving the same ratio). For quadratic functions φ^k, Rules M, G, N, and a combination of Rule A and a secant-type step can, under the assumption that $f \in C^3$, be shown to yield the ratio $(R - r)/(R + r)$ [the Kantorovitch ratio].

[9] Note that linear convergence is ensured also for nonquadratic functions φ^k, under a slight modification of the Armijo rule; cf. [Coh81].

[10] Note that, in the case of a quadratic function φ^k, the strong convexity constant m_{φ^k} and the Lipschitz continuity constant $M_{\nabla\varphi^k}$ can be identified with the smallest and largest eigenvalue of $\mathbf{B}_k$, respectively.

[11] Estimates of r and R are given in e.g. [Coh81].

[12] In fact, in [Aka59] it is shown that the ratio $(R - r)/(R + r)$ is the exact convergence rate of the steepest descent method.

(b) *Assume that* $\{\mathbf{B}_k\}$ *is chosen so that*

$$\left\{\mathbf{B}_k - \nabla^2 f(\mathbf{x}^*)\right\} \to \mathbf{0}. \tag{2.14}$$

Then, $\{\mathbf{x}^k\}$ *converges superlinearly if and only if* $\{t_k\} \to 1$.

(c) *Assume that* $\nabla^2 f$ *is Lipschitz continuous in a neighbourhood of* $\mathbf{x}^*$. *If, for some* $\eta \geq 0$, $\{\mathbf{B}_k\}$ *satisfies*

$$\left\|\mathbf{B}_k - \nabla^2 f(\mathbf{x}^*)\right\| \leq \eta \left\|\mathbf{x}^k - \mathbf{x}^*\right\|, k = 0, 1, \ldots, \tag{2.15}$$

then convergence is quadratic.

(d) *Let further* $\mathbf{B}_k = \nabla^2 f(\mathbf{x}^k), \forall k$. *Then there is a* K *(depending on* $\mathbf{x}^0$*) so that we may take* $t_k = 1, \forall k \geq K$, *and quadratic convergence is achieved.*

Proof

(a) See [DeM74b, GoP67] and [Ber82a, Prop. 1.15].

(b) See [Fle87, p. 126].

(c) See [DeM77].

(d) See e.g. [OrR70, Th. 10.2.2]. □

For nonquadratic problems [**NLP**] finite convergence can not be obtained in general.[15] For quadratic problems, however, finite convergence of cost approximation methods may be achieved by choosing $\{\varphi^k\}$ such that the sequence of iterates coincide with that of a finitely convergent quasi-Newton or conjugate gradient method (cf. Theorem 2.1). Examples of such algorithms are the DFP method [Dav59, FlP63], the BFGS method [Bro69, Bro70, Fle70, Gol70, Sha70] and the Broyden family of methods [Bro65].

2.6 A posteriori error bounds and lower bounds

Below we give upper bounds on the Euclidean distance from any point $\mathbf{x} \in \Re^n$ to an optimal solution $\mathbf{x}^*$ of [**NLP**], and a lower bound on the optimal objective value.

Theorem 2.6 (A posteriori error bounds) *Let* f *be strongly convex. Let* $\varphi : \Re^n \mapsto \Re$ *be a convex function in* C^1, *with a Lipschitz continuous gradient.*

(a) *For any* $\mathbf{x} \in \Re^n$ *and* $\mathbf{y} \in Y(\mathbf{x})$,

$$\|\mathbf{x} - \mathbf{x}^*\| \leq \frac{M_{\nabla\varphi}}{m_f} \|\mathbf{y} - \mathbf{x}\|. \tag{2.16}$$

[13] Some of these results hold for line search rules other than Rule M, see e.g. [OrR70, Th. 14.4.3] and [PsD78, Ths. II.2.1–2]. Also, the conditions may be relaxed to hold only locally around $\mathbf{x}^*$.

[14] The Conditions (2.13) through (2.15) are successively stronger. Condition (2.13) states that $\{\mathbf{B}_k\}$ converges to $\nabla^2 f(\mathbf{x}^*)$ along the direction $\mathbf{x}^{k+1} - \mathbf{x}^k$, i.e., that the search direction converges to the Newton direction in the limit, while Condition (2.14) requires this convergence to hold uniformly. The Condition (2.15) requires this convergence to be sufficiently fast.

[15] To obtain finite convergence, a *sharpness* property (see Definitions A.8.c and A.8.d) must be imposed on the set of solutions. See Theorems 3.6, 4.7, 4.9, and 5.17 for such results. The problem [**NLP**] can, however, not have such a property [Pol87].

(b) *Assume further that $f \in C^2$ has a Lipschitz continuous gradient. Then for any $\mathbf{x} \in \Re^n$,*

$$\|\mathbf{x} - \mathbf{x}^*\| \leq \frac{\|\nabla f(\mathbf{x})\|}{m_f}. \tag{2.17}$$

Proof

(a) For any $\mathbf{x} \in \Re^n$ and $\mathbf{y} \in Y(\mathbf{x})$, we have that

$$\begin{aligned} M_{\nabla\varphi} \|\mathbf{y} - \mathbf{x}\| &\geq \|\nabla\varphi(\mathbf{y}) - \nabla\varphi(\mathbf{x})\| && (\nabla\varphi \text{ Lipschitz}) \\ &= \|\nabla f(\mathbf{x}^*) - \nabla f(\mathbf{x})\| && (2.3) \\ &\geq m_f \|\mathbf{x}^* - \mathbf{x}\|. && (f \text{ strongly convex}) \end{aligned}$$

(b) See [PsD78, Th. II.1.2]. □

From (2.3), the inequality $\|\nabla f(\mathbf{x})\| \leq M_{\nabla\varphi} \|\mathbf{y} - \mathbf{x}\|$ holds; the bound (2.17) is hence in general stronger than (2.16). However, the first part of the theorem may be extended to nondifferentiable and constrained problems (see Theorems 3.7 and 4.10).

Below we establish a lower bound on f^*.

Theorem 2.7 (Lower bound) *Let $\varphi : \Re^n \mapsto \Re$ be a convex function in C^1. Assume that the function $f - \varphi$ is convex. Then, for any $\mathbf{x} \in \Re^n$ and $\mathbf{y} \in Y(\mathbf{x})$,*

$$f_\varphi(\mathbf{y}) \leq f^*. \tag{2.18}$$

Furthermore, equality holds in (2.18) if and only if $\mathbf{x}$ is a minimizer of f.

Proof We have, for any $\mathbf{x} \in \Re^n$, $\mathbf{y} \in Y(\mathbf{x})$, and minimizer $\mathbf{x}^*$,

$$\begin{aligned} f_\varphi(\mathbf{y}) &\leq \varphi(\mathbf{x}^*) + f(\mathbf{x}) - \varphi(\mathbf{x}) + [\nabla f(\mathbf{x}) - \nabla\varphi(\mathbf{x})]^{\mathrm{T}} (\mathbf{x}^* - \mathbf{x}) && [\mathbf{y} \in Y(\mathbf{x})] \\ &\leq f^*. && [f - \varphi \text{ convex}] \end{aligned}$$

To establish the second part of the theorem we note that, from Lemma 2.1.a, $\mathbf{x}$ minimizes $f \Leftrightarrow f_\varphi(\mathbf{y}) = f_\varphi(\mathbf{x})$, which is equivalent to $f_\varphi(\mathbf{y}) = f_\varphi(\mathbf{x}) = f(\mathbf{x}) = f^*$. □

This result extends that of [LaM90] for strictly convex problems over compact Cartesian product sets. It may be extended to the nondifferentiable case (see Theorem 3.8), and to constrained problems (see Theorem 4.11).

2.7 Jacobi and Gauss-Seidel type methods

The classical Jacobi and Gauss-Seidel methods for linear and nonlinear equations (see [Var62, OrR70] and the relevant references cited therein) are instances of the general cost approximation algorithm, as we show below. Rewrite [**NLP**] as follows.

[**NLP**]
$$\min_{\mathbf{x} \in \Re^n} f\left(\mathbf{x}_{i_-}, x_i, \mathbf{x}_{i_+}\right)$$

The Jacobi iteration, at $\mathbf{x}^k \in \Re^n$, is defined by n simultaneous one-dimensional problems, which for variable x_i is

$$[\mathbf{J}_i^k] \qquad x_i^{k+1} \in \arg\min_{x_i \in \Re} f\left(\mathbf{x}_{i_-}^k, x_i, \mathbf{x}_{i_+}^k\right).$$

To show that the Jacobi algorithm is an instance of the general cost approximation method, we introduce the function $\varphi : \Re^{2n} \mapsto \Re$ as

$$\varphi(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \varphi_i(x_i, \mathbf{y}) = \sum_{i=1}^{n} f(\mathbf{y}_{i_-}, x_i, \mathbf{y}_{i_+}). \tag{2.19}$$

It can easily be seen that $\nabla_{\mathbf{x}} \varphi(\mathbf{x}, \mathbf{y})|_{\mathbf{x}=\mathbf{y}} = \nabla f(\mathbf{y})$, and

$$f_\varphi\left(\mathbf{x}, \mathbf{x}^k\right) = \sum_{i=1}^{n} f\left(\mathbf{x}_{i_-}^k, x_i, \mathbf{x}_{i_+}^k\right),$$

so that $[\mathbf{NLP}_\varphi^k]$ equals $[\mathbf{J}_i^k]$, $i = 1, \ldots, n$, by letting $x_i^{k+1} = y_i^k$ for all i, i.e., the Jacobi method is an instance of the cost approximation methods, where unit steps are used.

This result implies that the Jacobi algorithm, for which convergence requires strong conditions (see, e.g., [BeT89, Prop. 3.3.10]), can both be made convergent under weaker conditions and more efficient in practice, by the introduction of a line search.[16] We also note here that a decomposition into independent subproblems is obtained from any choice of mapping $\mathbf{\Phi}^k$, which is separable with respect to $\Re^n$.

The Gauss-Seidel iteration is defined through the following n one-dimensional problems, solved sequentially.

$$[\mathbf{GS}_i^k] \qquad x_i^{k+1} \in \arg\min_{x_i \in \Re} f\left(\mathbf{x}_{i_-}^{k+1}, x_i, \mathbf{x}_{i_+}^k\right)$$

To show that the Gauss-Seidel method is a special case of cost approximation, we introduce the mapping $\mathbf{\Phi} : \Re^{2n} \mapsto \Re^n$, with ith component

$$\mathbf{\Phi}_i(\mathbf{x}, \mathbf{y}) = \frac{\partial f\left(\mathbf{x}_{i_-}, x_i, \mathbf{y}_{i_+}\right)}{\partial x_i}. \tag{2.20}$$

From (2.2), the cost approximation subproblem may be written as

$$\mathbf{\Phi}\left(\mathbf{y}^k, \mathbf{x}^k\right) + \nabla f\left(\mathbf{x}^k\right) - \mathbf{\Phi}\left(\mathbf{x}^k, \mathbf{x}^k\right) = \mathbf{0}. \tag{2.21}$$

But $\mathbf{\Phi}(\mathbf{x}, \mathbf{x}) = \nabla f(\mathbf{x})$ for any $\mathbf{x} \in \Re^n$, and (2.21) reduces to $\mathbf{\Phi}(\mathbf{y}^k, \mathbf{x}^k) = \mathbf{0}$, which, by (2.20), may be written as

$$\left.\frac{\partial f\left(\mathbf{y}_{i_-}^k, y_i, \mathbf{x}_{i_+}^k\right)}{\partial y_i}\right|_{y_i = y_i^k} = 0, \; i = 1, \ldots, n. \tag{2.22}$$

Since f is minimized in each step during iteration k we have that $x_i^{k+1} = y_i^k$, and (2.22) is equivalent to the sequence of problems $[\mathbf{GS}_i^k]$.

The fact that Gauss-Seidel can be included in the same unified description as deflected gradient methods is interesting in its own right, since these two methods are generally considered to be totally different in character [Lue84, p. 229].

We introduce a relaxation parameter, $\omega > 0$, in $[\mathbf{GS}_i^k]$, and let $x_i^{k+1} = x_i^k + \omega(y_i^k - x_i^k)$. The basic convergence results for Gauss-Seidel methods are stated below.

[16]Note that the convexity requirement on $\varphi(\cdot, \mathbf{x})$ yields that f is implicitly assumed to be convex in each variable component. This is an example of the implicit relations between requirements on φ and f that may appear in certain instances of cost approximation algorithms; see Remark 2.1.

Theorem 2.8 (Convergence of the Gauss-Seidel method)

(a) *Let f be pseudoconvex and let $\mathbf{x}^0$ be such that the level set $L^f_{\Re^n}(\mathbf{x}^0)$ is bounded. For $\omega = 1$, any accumulation point of the sequence $\{\mathbf{x}^k\}$ (at least one such point exists) is optimal in* [**NLP**].

(b) *Let f be convex on $\Re^n$, and strictly convex with respect to each variable x_i, $i = 1, \dots, n$. Assume further that f is bounded from below on $\Re^n$ and that the sequence $\{\mathbf{x}^k\}$, with $\omega = 1$, is well defined, in the sense that each problem* [$\mathbf{GS}^k_i$] *has a solution. Then any accumulation point of the sequence $\{\mathbf{x}^k\}$ is optimal in* [**NLP**].

(c) *Assume that f is strongly convex. Then for $\omega \in (0, 1]$, the sequence $\{\mathbf{x}^k\}$ converges to the unique solution of* [**NLP**].

Proof

(a) See [Zad69].

(b) See [BeT89, Prop. 3.2.5].

(c) See [Sch62, Sch68] or [OrR70, Th. 14.6.7]. □

In the following, we shall present a generalization of the Gauss-Seidel algorithm, in which each subproblem [$\mathbf{GS}^k_i$] is replaced by a (more general) separable cost approximation subproblem, and a global line search step. This algorithm may also be viewed as a version of the original cost approximation algorithm, where only a subset of the variables are updated in each iteration.

The general cyclic cost approximation scheme is as follows. Introduce the partitioning $\mathbf{x}^{\mathrm{T}} = (\mathbf{x}_1^{\mathrm{T}}, \dots, \mathbf{x}_{|\mathcal{C}|}^{\mathrm{T}})$, where $\mathbf{x}_i \in \Re^{n_i}$ and $\sum_{i \in \mathcal{C}} n_i = n$, and let $\mathbf{\Phi} : \Re^{2n} \mapsto \Re^n$ be given by

$$\mathbf{\Phi}(\mathbf{x}, \mathbf{y})^{\mathrm{T}} = \left[\mathbf{\Phi}_1(\mathbf{x}_1, \mathbf{y})^{\mathrm{T}}, \dots, \mathbf{\Phi}_{|\mathcal{C}|}\left(\mathbf{x}_{|\mathcal{C}|}, \mathbf{y}\right)^{\mathrm{T}} \right],$$

where $\mathbf{\Phi}_i(\mathbf{x}_i, \mathbf{y})$ is monotone as a function of $\mathbf{x}_i$. For simplicity, we assume that each function $\mathbf{\Phi}_i(\mathbf{x}_i, \mathbf{y})$ is the gradient of a function $\varphi_i(\mathbf{x}_i, \mathbf{y})$, which then is convex and in C^1 with respect to $\mathbf{x}_i$. In iteration k, the index $i_k \in \mathcal{C}$ is chosen. If a cyclic rule is used, $i_k = k \pmod{|\mathcal{C}|} + 1$. We then solve ($i = i_k$)

$$[\mathbf{NLP}^k_{\varphi_i}] \qquad \mathbf{y}^k_i \in \arg \min_{\mathbf{x}_i \in \Re^{n_i}} \left\{ \varphi_i\left(\mathbf{x}_i, \mathbf{x}^k\right) + \left[\nabla_i f\left(\mathbf{x}^k\right) - \nabla \varphi_i\left(\mathbf{x}^k_i, \mathbf{x}^k\right)\right]^{\mathrm{T}} \left(\mathbf{x}_i - \mathbf{x}^k_i\right) \right\},$$

and define $\mathbf{y}^k_j = \mathbf{x}^k_j$, for all $j \neq i$. Let $\mathbf{x}^{k+1} = \mathbf{x}^k + t_k(\mathbf{y}^k - \mathbf{x}^k)$, i.e.,

$$\mathbf{x}^{k+1}_j = \begin{cases} \mathbf{x}^k_i + t_k\left(\mathbf{y}^k_i - \mathbf{x}^k_i\right), & j = i \\ \mathbf{x}^k_j, & j \neq i, \end{cases}$$

where

$$t_k \in \arg \min_{t \geq 0} \ f\left(\mathbf{x}^k_{i_-}, \mathbf{x}^k_i + t\left(\mathbf{y}^k_i - \mathbf{x}^k_i\right), \mathbf{x}^k_{i_+}\right).$$

If $|\mathcal{C}| = 1$, the algorithm reduces to the original cost approximation algorithm, while, with $|\mathcal{C}| = n$ and φ chosen as in (2.19), it reduces to the classical Gauss-Seidel scheme.

Theorem 2.9 (Convergence of cyclic cost approximation algorithms) *Let each function $\varphi_i : \Re^{n_i} \times \Re^n \mapsto \Re$ be a convex function in C^1 with respect to $\mathbf{x}_i$ and continuous on $\Re^{n_i} \times \Re^n$. Let the indices be chosen in cyclic order, and let $\mathbf{x}^0$ be chosen such that $L^f_{\Re^n}(\mathbf{x}^0)$ is bounded. Assume that the problem $[\mathbf{NLP}_{\varphi_i}]$ is well defined, in the sense that its solution set $Y_i(\mathbf{x})$ is nonempty and bounded for any $\mathbf{x} \in \Re^n$. Assume further that the sequence $\{\mathbf{y}^k\}$ is bounded. Then, under Rule M, $\{f(\mathbf{x}^k)\} \to f(\overline{\mathbf{x}})$ for some $\overline{\mathbf{x}} \in \Omega$, any accumulation point $\mathbf{x}^\infty$ of the sequence $\{\mathbf{x}^k\}$ (at least one such point exists) lies in Ω, and (2.9) holds.*

Proof The proof of the theorem is a slight extension of the proof of Theorem 2.2; we only describe the major changes.

(1) The fact that $\{\mathbf{x}^k\}$ lies in a compact set follows from the level set compactness assumption and from the descent property below.

(2) The descent property follows from Lemma 2.1.b. The algorithm terminates if, for some k, $\mathbf{x}^k \in \Omega$, since this implies that $\nabla_i f(\mathbf{x}^k) = \mathbf{0}$ for all $i \in \mathcal{C}$, which, in turn, implies, from the cyclic rule and Lemma 2.1.a, that $\mathbf{x}^k_{i_{k+l}} \in Y_{i_{k+l}}(\mathbf{x}^k)$ holds for $l = 0, \dots, |\mathcal{C}| - 1$.

(3) The algorithmic map may be decomposed into $A = MD_{|\mathcal{C}|} \cdot \ldots \cdot MD_1$, where D_i is the direction finding map corresponding to $[\mathbf{NLP}_{\varphi_i}]$. Each map D_i is closed from Lemma 2.2.a. The closedness of M, together with the boundedness of $\{\mathbf{y}^k\}$ and [Zan69b, Cor. 4.2.1], then implies that A is closed.

The final statement follows from Lemma 2.3. □

The function φ does not need to be separable for the theorem to be applicable. In the analysis of cost approximation algorithms for problems defined over Cartesian product sets, we shall extend the above results in several directions; under a strong convexity assumption on φ, the exact line search may be replaced by simpler step length rules, and the indices may be chosen in a more flexible fashion. The results obtained for such problems in Section 4.15 may be applied to [**NLP**] directly.

2.8 Summary

In this chapter we have studied the cost approximation method applied to unconstrained optimization. For this problem class, it is shown that the search directions are highly connected to gradient related and deflected gradient directions (Lemma 2.1.e and Theorem 2.1, respectively). Based on these observations, convergence results are derived (Theorems 2.3–2.5). Convergence is also established under slightly less restrictive assumptions (Theorem 2.2). An important conclusion is that coordinate descent methods are instances of the class of cost approximation methods, thus making it possible to unify seemingly unrelated descent methods in nonlinear programming. The Jacobi and Gauss-Seidel type versions of the cost approximation algorithm enable the solution of large-scale problems, through a decomposition into sequences of problems of smaller dimensions.

An interesting question for further research is if it is possible to extend the convergence rate results for quadratic functions φ^k (Theorems 2.4.c–d and 2.5) to nonquadratic ones. The convergence rates (practical as well as theoretical) that can be obtained from the decomposition versions of the cost approximation algorithms (Section 2.7) should also be further investigated.

Chapter 3

Nondifferentiable optimization

"Don't mind him," he said, wheezing contemptuously. "He's a stickler for things. Worst of all, the man's a mathematician. Pah! *A mathematician," repeated Dr. Fell, glaring at his salad as though he expected to find a binomial theorem lurking in the lettuce.*

JOHN DICKSON CARR
Hag's Nook (1933)

In this chapter we study the cost approximation algorithm for nondifferentiable programs. Consider the generalized equation (1.7). Assuming that $\mathbf{U} \equiv \partial u$, where $u : \Re^n \mapsto \Re \cup \{+\infty\}$ is l.s.c., proper and convex, and that $\mathbf{F} \equiv \nabla f$, where $f : \Re^n \mapsto \Re \cup \{+\infty\}$ is in C^1 on $\texttt{dom}\, u$, (1.7) reduces to the problem of finding an $\mathbf{x}^* \in \Re^n$ such that

$$-\nabla f(\mathbf{x}^*) \in \partial u(\mathbf{x}^*). \tag{3.1}$$

(These properties of u and f are standing assumptions throughout this chapter.) We let Ω denote the set of solutions to (3.1).

We first examine under what assumptions (3.1) defines necessary and sufficient conditions for the nondifferentiable nonlinear program[1]

$$[\mathbf{NNLP}] \qquad \min_{\mathbf{x} \in \Re^n} T(\mathbf{x}) \stackrel{\text{def}}{=} f(\mathbf{x}) + u(\mathbf{x}).$$

3.1 Optimality conditions

We first note that u is continuous on $\texttt{rint}\,(\texttt{dom}\, u)$ [Roc70a, Th. 10.1], and Lipschitz continuous on any closed and bounded subset of $\texttt{rint}\,(\texttt{dom}\, u)$ [Roc70a, Th. 10.4].

[1] Recall from Section 1.1 that the constrained program (1.3) can be expressed in the form [**NNLP**] through the use of an indicator function. The convergence analysis made in this chapter is therefore directly applicable to (1.3), by replacing u with $w + \delta_X$, where $w : \Re^n \mapsto \Re \cup \{+\infty\}$ is a l.s.c., proper and convex function, and δ_X is the indicator function of X. (The function δ_X of a nonempty convex set X is proper and convex, and l.s.c. if and only if X also is closed [Phe89, p. 40].) In this case, $\texttt{dom}\, u = \texttt{dom}\, w \cap X$; whenever $\texttt{dom}\, u$ is nonempty, $u = w + \delta_X$ is a l.s.c., proper and convex function (see [vTi84, Sec. 5.4] and [Roc70a, Th. 5.2]). Also, $\partial u(\mathbf{x}) = \partial w(\mathbf{x}) + N_X(\mathbf{x})$ holds for any $\mathbf{x} \in \Re^n$ if $\texttt{rint}\,(\texttt{dom}\, u) \cap \texttt{rint}\, X \neq \emptyset$ [Roc70a, Th. 23.8], see also [Roc81, Th. 5C]. (If X is polyhedral, $\texttt{rint}\,(\texttt{dom}\, u) \cap X \neq \emptyset$ suffices.)

When $\mathtt{dom}\ u = \Re^n$, u is *locally Lipschitz* and T inherits this property, since f is continuously differentiable on an open neighbourhood of $\mathtt{dom}\ u$. From [Cla75] it then follows that T has a *generalized gradient*, $\partial^* T$, everywhere. For any $\mathbf{x} \in \Re^n$, $\partial^* T(\mathbf{x}) = \nabla f(\mathbf{x}) + \partial u(\mathbf{x})$ holds, and the necessary optimality condition for [**NNLP**] is [Cla76]

$$\mathbf{0} \in \partial^* T\left(\mathbf{x}^*\right), \tag{3.2}$$

which is equivalent to (3.1). Points $\mathbf{x}^*$ satisfying (3.2) are in [MiF81] referred to as *critical points*, and are characterized by the inequality

$$T'\left(\mathbf{x}^*; \mathbf{x} - \mathbf{x}^*\right) \geq 0, \forall \mathbf{x} \in \Re^n, \tag{3.3}$$

where $T'(\mathbf{x}; \mathbf{d})$ is the (one-sided) directional derivative of T at $\mathbf{x}$ in the direction $\mathbf{d}$, i.e.,

$$T'(\mathbf{x}; \mathbf{d}) \stackrel{\text{def}}{=} \lim_{\lambda \downarrow 0} \frac{1}{\lambda} \left[T(\mathbf{x} + \lambda \mathbf{d}) - T(\mathbf{x})\right]. \tag{3.4}$$

When $\mathtt{dom}\ u \neq \Re^n$, the Condition (3.1) is not generally necessary for the optimality of $\mathbf{x}^*$ in [**NNLP**]. However, under the following additional assumption on u, it is shown in [MiF81] that (3.1) is indeed a necessary optimality condition for [**NNLP**].[2]

Assumption 3.1 (Lower semicontinuity of $u'(\mathbf{x}; \cdot)$) *For any* $\mathbf{x} \in \mathtt{dom}\ u$ *and* $\mathbf{d} \in \Re^n$,

$$u'(\mathbf{x}; \mathbf{d}) = \liminf_{\mathbf{e} \to \mathbf{d}}\ u'(\mathbf{x}; \mathbf{e}).$$

A direct consequence of this assumption is that the directional derivative of u, for any $\mathbf{x} \in \mathtt{dom}\ u$ and $\mathbf{d} \in \Re^n$, may be calculated as [Roc66], [Roc70a, Th. 23.2]

$$u'(\mathbf{x}; \mathbf{d}) = \sup\left\{ \boldsymbol{\xi}_u^{\mathrm{T}} \mathbf{d} \;\middle|\; \boldsymbol{\xi}_u \in \partial u(\mathbf{x}) \right\}. \tag{3.5}$$

The Condition (3.1) is *sufficient* for the global optimality of $\mathbf{x}^*$ if f is convex, or if f is pseudoconvex and $u \equiv \delta_X$, for a nonempty, closed and convex set $X \subseteq \Re^n$.

3.2 The cost approximation algorithm

The above established relations between (3.1) and [**NNLP**] motivate the choice $\psi \equiv T$ as merit function.

The cost approximation algorithm

A sequence $\{\boldsymbol{\Phi}^k\}$ of monotone cost approximating mappings is assumed given.

Step 0 (*Initial guess*) Choose an initial point $\mathbf{x}^0 \in \Re^n$, and let $k = 0$.

Step 1 (*Search direction generation*) Find a solution $\mathbf{y}^k \in \Re^n$ to the set inclusion

$$\mathbf{0} \in \boldsymbol{\Phi}^k\left(\mathbf{y}^k\right) + \mathbf{U}\left(\mathbf{y}^k\right) + \nabla f\left(\mathbf{x}^k\right) - \boldsymbol{\Phi}^k\left(\mathbf{x}^k\right). \tag{3.6}$$

The resulting search direction is $\mathbf{d}^k = \mathbf{y}^k - \mathbf{x}^k$.

[2] Note that Assumption 3.1 is always fulfilled for $\mathbf{x} \in \mathtt{rint}\ (\mathtt{dom}\ u)$ [Roc70a, Th. 23.4]. Further, if $u \equiv \delta_X$ for a nonempty, closed and convex set $X \subseteq \Re^n$, this assumption is not needed to ensure the necessity of (3.1), since, in this case, (3.1) reduces to (4.2), which defines the necessary optimality conditions for $\mathbf{x}^*$ without any further conditions on X.

Step 2 (*Convergence check*) If $\mathbf{x}^k$ satisfies (3.6) $\rightarrow$ Stop ($\mathbf{x}^k \in \Omega$). Otherwise, continue.

Step 3 (*Line search*) Find a step length, t_k, which solves the one-dimensional problem

$$t_k \in \arg\min_{t \geq 0} \, T\left(\mathbf{x}^k + t\mathbf{d}^k\right).$$

Step 4 (*Update*) Let $\mathbf{x}^{k+1} = \mathbf{x}^k + t_k \mathbf{d}^k$, and $k := k + 1$.

Step 5 (*Convergence check*) If $\mathbf{x}^k$ is acceptable as a solution[3]$\rightarrow$ Stop. Otherwise, go to Step 1.

A major difference between applying cost approximation methods to [**NNLP**] and to the continuously differentiable problem [**NLP**] is that a cost approximation involving u cannot be shown to yield descent directions in general.[4] When applying the cost approximation algorithms to [**NNLP**], we therefore restrict ourselves to linearizing only a differentiable part of the objective.[5] With this choice of linearization strategy, subproblem solutions yield descent directions, and the convergence analysis becomes unified with that of the differentiable case.

As in the differentiable case, most known instances of the class of cost approximation algorithms are defined by a search direction finding minimization subproblem. To derive the minimization problem corresponding to (3.6), we let $\varphi^k : \Re^n \mapsto \Re$ be a convex function in C^1 on $\mathtt{dom}\, u$. At $\mathbf{x}^k \in \mathtt{dom}\, u$, the objective T is written as $[\varphi^k(\mathbf{x}) + u(\mathbf{x})] + [f(\mathbf{x}) - \varphi^k(\mathbf{x})]$, and the subproblem [$\mathbf{NNLP}^k_{\varphi^k}$] is defined through a linearization of the second term, i.e.,

$$[\mathbf{NNLP}^k_{\varphi^k}] \qquad \mathbf{y}^k \in \arg\min_{\mathbf{x} \in \Re^n} \, T_{\varphi^k}(\mathbf{x}),$$

where

$$T_{\varphi^k}(\mathbf{x}) = \varphi^k(\mathbf{x}) + u(\mathbf{x}) + f\left(\mathbf{x}^k\right) - \varphi^k\left(\mathbf{x}^k\right) + \left[\nabla f\left(\mathbf{x}^k\right) - \nabla \varphi^k\left(\mathbf{x}^k\right)\right]^{\mathrm{T}} \left(\mathbf{x} - \mathbf{x}^k\right). \quad (3.7)$$

This algorithm framework has little in common with the two most well known algorithmic approaches in nondifferentiable programming. In the first, predetermined step lengths are taken in the negative direction of arbitrary subgradients (or simple combinations thereof) at each iteration point, see e.g. [Pol69, Sho83, Sho85]. These algorithms will not yield a monotonically decreasing sequence of objective values; instead, convergence is based on a decrease in the distance from iteration points to the set of optimal solutions. The second approach is based on line searches in directions of descent, obtained from (approximate) shortest ε-subgradients, see e.g. [PMW83, PoM85, Kiw85]. The framework developed here is also based on descent direction generation and line searches, but the basis for the auxiliary problem [$\mathbf{NNLP}^k_{\varphi^k}$] is different from the above approach.

Due to the nondifferentiability of T_{φ^k} it may be of practical interest to terminate the solution of [$\mathbf{NNLP}^k_{\varphi^k}$] prematurely. We will validate such an approach, and in fact prove that the solution of [$\mathbf{NNLP}^k_{\varphi^k}$] may be truncated almost arbitrarily early, provided that the algorithm used for solving it has a closed algorithmic map.

---

[3] Typical stopping criteria are a small step length, a small decrease in the objective value between two iterations, or a small relative error in the objective.

[4] As an example, with $f \equiv 0$, using $\varphi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$ yields $\mathbf{d}^k = -\boldsymbol{\xi}_u(\mathbf{x}^k)$, i.e., the direction of the negative of a subgradient of u at $\mathbf{x}^k$, which is well known not to be a descent direction with respect to u in general.

[5] Note that the additive decomposition of T into f and u is *never* unique; adding an arbitrary convex function in C^1 to u and subtracting it from f leaves T unaltered.

3.3 Instances of cost approximation algorithms

Mine and Fukushima [MiF81] present an algorithm for [**NNLP**], in which the differentiable part f of T is linearized; their algorithm is, in our framework, obtained by choosing $\varphi^k \equiv 0$, $\forall k$, and using Rule M. Convergence is proved under conditions similar to those of Theorem 3.2, but with an additional (unnecessary) strict convexity condition on u.

In [Kiw86], a method similar to that of [MiF81] is proposed. The main difference is that here, in order to reduce the computational work, a polyhedral approximation to u is constructed from collected subgradients of u. (The truncation strategy for [$\mathbf{NNLP}^k_{\varphi^k}$] developed below can be seen as an alternative to this approach.)

The algorithm of [MiF81] is extended in [FuM81], in which the function φ^k is chosen equal to $\frac{c_k}{2}\|\mathbf{x}\|^2$, where $c_k > 0$, $\forall k$, and an Armijo-type step length rule, which is closely related to the Armijo rule adopted here, is utilized. This algorithm is similar to the proximal point algorithm [Min62, Mor65, Mar70, Mar72, Roc76a, Roc76b, Ber82a, BeT89], and actually reduces to it in the cases where $f \equiv 0$ and the line search is replaced by a unit step. The method of [FNS91] introduces an updating rule for the parameters c_k, based on the trust region concept, which eliminates the need for a line search to obtain a sufficient decrease in the objective in each step. Other proximal point type methods for nondifferentiable programs can be found in e.g. [Aus84, Fuk84].

As remarked earlier, replacing u with an affine function, $u(\mathbf{x}^k) + \boldsymbol{\xi}_u^{k\,\mathrm{T}}(\mathbf{x} - \mathbf{x}^k)$, based on an arbitrary subgradient $\boldsymbol{\xi}_u^k$ of u at $\mathbf{x}^k$, does not generally yield a direction of descent with respect to T. The resulting subproblem objective, however, has a desirable property of being differentiable, which is not the case for [$\mathbf{NNLP}^k_{\varphi^k}$]. In the following, we extend the cost approximation concept to also allow for linearizations of u; by means of examples we show that the corresponding algorithm class includes well known instances from both the above described traditional approaches for nondifferentiable optimization.

Consider the problem (1.3), with $f \equiv 0$. Let $\varphi^k = \varepsilon_k\varphi$, where $\varphi : \Re^n \mapsto \Re$ is a strongly convex function in C^1 on X, and where the sequence $\{\varepsilon_k\}$ fulfills

$$\varepsilon_k > 0, \; \sum_{k=1}^{\infty} \varepsilon_k = +\infty, \; \sum_{k=1}^{\infty} \varepsilon_k^2 < +\infty. \tag{3.8}$$

In the cost approximation method, with $t_k = 1$, $\forall k$, the sequence of iterates is defined by

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in X} \left\{ \varphi(\mathbf{x}) + \left[\varepsilon_k \boldsymbol{\xi}_u^k - \nabla\varphi\left(\mathbf{x}^k\right)\right]^{\mathrm{T}} \left(\mathbf{x} - \mathbf{x}^k\right) \right\},$$

where $\boldsymbol{\xi}_u^k$ is an arbitrary subgradient of u at $\mathbf{x}^k$. Cohen and Zhu [Zhu82, CoZ83, CoZ84, Zhu87] present this algorithm for the solution of (1.3), and show that, under boundedness properties of ∂u of the form $\|\boldsymbol{\xi}_u(\mathbf{x})\| \leq M_1\|\mathbf{x}\| + M_2$, $\forall \boldsymbol{\xi}_u(\mathbf{x}) \in \partial u(\mathbf{x})$, $\forall \mathbf{x}$, for some $M_1, M_2 > 0$, any accumulation point of $\{\mathbf{x}^k\}$ is optimal.

A very interesting special case of algorithms within this framework is obtained by letting $\varphi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$. Then $\mathbf{x}^{k+1}$ can be shown to be given by

$$\mathbf{x}^{k+1} = P_X\left(\mathbf{x}^k - \varepsilon_k \boldsymbol{\xi}_u^k\right),$$

i.e., we recover the standard *subgradient optimization method* [Erm66, Pol67, Sho85]. In [CoZ84] it is shown that the well known alternative to the step length rule (3.8),

$$\varepsilon_k = \gamma_k \frac{u\left(\mathbf{x}^k\right) - u^*}{\left\|\boldsymbol{\xi}_u^k\right\|^2}, \; 0 < \gamma \leq \gamma_k \leq 2 - \gamma$$

(see [Ere65, Pol69, Pol78]), may also be used, and that it may be extended to the more general functions φ, by replacing the latter condition by $0 < \gamma \leq \gamma_k \leq m_\varphi - \gamma$. (Also other extensions are discussed in [CoZ83, CoZ84].) Convergence is established through the monotone decrease of the generalized norm

$$\psi(\mathbf{x}) = \varphi(\mathbf{x}^*) - \varphi(\mathbf{x}) - \nabla\varphi(\mathbf{x})^{\mathrm{T}}(\mathbf{x}^* - \mathbf{x}), \tag{3.9}$$

for any minimizer $\mathbf{x}^*$; note that this merit function reduces to the familiar $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|^2$ for the special case $\varphi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$. Linear convergence may also be established under conditions similar to those of [Pol69] for the special case of the subgradient optimization method.[6]

In the second approach the subgradient $\boldsymbol{\xi}_u^k$ is chosen such that the subproblem solution yields descent with respect to T. One example of such an algorithm is the extension of the steepest descent method to nondifferentiable problems [Lue70, BeM73], which can be identified as a special case of cost approximation by the choices $\varphi^k(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, and $\boldsymbol{\xi}_u^k = \arg\min\{\boldsymbol{\xi}_u \mid \boldsymbol{\xi}_u \in \partial u(\mathbf{x}^k)\}$, $\forall k$. (To ensure convergence of this type of scheme, the search directions must however be chosen from larger sets, e.g., from *ε-subdifferentials* [PMW83, PoM85].)

To summarize, we have shown that both the traditional subgradient optimization method and the extension of the steepest descent method to nondifferentiable optimization are instances of a cost approximation method which allows for linearizations of u. The above results may be extended in various directions, for instance including a differentiable part f and more general choices of φ^k; the convergence properties of these extensions of cost approximation methods are interesting topics for further research.

3.4 An algorithmic equivalence result

In a series of papers Cohen [Coh76, Coh78, Coh80] develops a class of iterative methods (the auxiliary problem principle) for the problem (1.3). Given $\mathbf{x}^k \in X$, $\mathbf{x}^{k+1}$ is given by

$$\mathbf{x}^{k+1} \in \arg\min_{\mathbf{x}\in X} \left\{\varphi^k(\mathbf{x}) + \varepsilon_k u(\mathbf{x}) + \left[\varepsilon_k \nabla f\left(\mathbf{x}^k\right) - \nabla\varphi^k\left(\mathbf{x}^k\right)\right]^{\mathrm{T}} \mathbf{x}\right\}, \tag{3.10}$$

where $\varepsilon_k > 0$. Although this algorithm class differs in spirit from that of the cost approximation methods (for instance, Cohen does not consider line searches), the subproblems are equivalent in the sense of the below theorem.

Theorem 3.1 (An algorithmic equivalence result) *Consider (1.3). If each cost approximating map $\boldsymbol{\Phi}^k$ is a gradient of a convex function φ^k, then the resulting subproblem of the cost approximation method,*

$$\mathbf{y}^k \in \arg\min_{\mathbf{x}\in X} T_{\varphi^k}(\mathbf{x}), \tag{3.11}$$

where T_{φ^k} is given by (3.7), is equivalent to (3.10).

Proof Dividing the objective of (3.10) by ε_k, replacing $\frac{1}{\varepsilon_k}\varphi^k$ with φ^k, adding necessary constants, and letting $\mathbf{x}^{k+1} = \mathbf{y}^k$ we obtain (3.11). □

[6]These conditions include, in particular, the weak sharp minimum property (see Definition A.8.c), under which we are able to establish finite convergence for cost approximation methods (see Section 4.13).

3.5 Analysis of the direction finding subproblem

We begin by providing generalizations of Lemma 2.1 and 2.2.

Lemma 3.1 (Descent lemma) *Let $\varphi : \Re^n \mapsto \Re$ be a convex function in C^1 on $\mathtt{dom}\, u$. Let $\mathbf{x} \in \mathtt{dom}\, u$, $Y(\mathbf{x})$ be the (possibly empty) set of optimal solutions to* $[\mathbf{NNLP}_\varphi]$, $\mathbf{y} \in Y(\mathbf{x})$, *and $\mathbf{d} = \mathbf{y} - \mathbf{x}$. Let Assumption 3.1 hold.*

(a) $\mathbf{x} \in \Omega \Leftrightarrow \mathbf{x} \in Y(\mathbf{x}) \Leftrightarrow T_\varphi(\mathbf{y}) = T_\varphi(\mathbf{x})$.

(b) *Let $\overline{\mathbf{x}} \in \Re^n$ be any point such that $T_\varphi(\overline{\mathbf{x}}) < T_\varphi(\mathbf{x})$. Then $T'(\mathbf{x}; \overline{\mathbf{x}} - \mathbf{x}) < 0$. Especially, for $\overline{\mathbf{x}} = \mathbf{y}$, if $\mathbf{x} \notin \Omega$, then $T'(\mathbf{x}; \mathbf{d}) < 0$.*

(c) *Let φ or u be strictly convex on $\mathtt{dom}\, u$. If $Y(\mathbf{x})$ is nonempty, then it is singleton.*

(d) *Let φ be strongly convex on $\mathtt{dom}\, u$. Then $Y(\mathbf{x})$ is nonempty and singleton, and*

$$T'(\mathbf{x}; \mathbf{d}) \leq -m_\varphi \|\mathbf{d}\|^2. \tag{3.12}$$

(e) *Let φ be strongly convex on $\mathtt{dom}\, u$, and let $\nabla\varphi$ be Lipschitz continuous on $\mathtt{dom}\, u$. If $\mathbf{x} \notin \Omega$, then*

$$-\frac{T'(\mathbf{x}; \mathbf{d})}{\|\nabla f(\mathbf{x}) + \boldsymbol{\xi}_u\| \cdot \|\mathbf{d}\|} \geq \rho, \tag{3.13}$$

where $\rho = m_\varphi / M_{\nabla\varphi}$, and $\boldsymbol{\xi}_u \in \partial u(\mathbf{y})$.

Proof

(a) $\mathbf{x} \in \Omega \Leftrightarrow -\nabla f(\mathbf{x}) \in \partial u(\mathbf{x}) \Leftrightarrow -\nabla\varphi(\mathbf{x}) - \nabla f(\mathbf{x}) + \nabla\varphi(\mathbf{x}) \in \partial u(\mathbf{x}) \Leftrightarrow \mathbf{x} \in Y(\mathbf{x}) \Leftrightarrow T_\varphi(\mathbf{y}) = T_\varphi(\mathbf{x})$.

(b) By the convexity of φ, $T_\varphi(\overline{\mathbf{x}}) < T_\varphi(\mathbf{x})$ implies that $\nabla f(\mathbf{x})^{\mathrm{T}}(\overline{\mathbf{x}} - \mathbf{x}) < u(\mathbf{x}) - u(\overline{\mathbf{x}})$. Then,

$$\begin{aligned} T'(\mathbf{x}; \overline{\mathbf{x}} - \mathbf{x}) &= \nabla f(\mathbf{x})^{\mathrm{T}}(\overline{\mathbf{x}} - \mathbf{x}) + \sup\left\{\boldsymbol{\xi}_u^{\mathrm{T}}(\overline{\mathbf{x}} - \mathbf{x}) \,\middle|\, \boldsymbol{\xi}_u \in \partial u(\mathbf{x})\right\} \\ &< u(\mathbf{x}) - u(\overline{\mathbf{x}}) + \sup\left\{\boldsymbol{\xi}_u^{\mathrm{T}}(\overline{\mathbf{x}} - \mathbf{x}) \,\middle|\, \boldsymbol{\xi}_u \in \partial u(\mathbf{x})\right\}, \end{aligned} \tag{3.14}$$

and the right hand side of (3.14) is nonpositive, since u is convex.

(c) Follows from the strict convexity of T_φ.

(d) The first result is immediate from the strong convexity of T_φ. Using the optimality condition

$$-\nabla\varphi(\mathbf{y}) - \nabla f(\mathbf{x}) + \nabla\varphi(\mathbf{x}) \in \partial u(\mathbf{y}) \tag{3.15}$$

of $[\mathbf{NNLP}_\varphi]$, and the monotonicity of ∂u [Roc70a, Cor. 31.5.2] we have that

$$[-\nabla\varphi(\mathbf{y}) - \nabla f(\mathbf{x}) + \nabla\varphi(\mathbf{x}) - \boldsymbol{\xi}_u]^{\mathrm{T}}(\mathbf{y} - \mathbf{x}) \geq 0, \forall \boldsymbol{\xi}_u \in \partial u(\mathbf{x}). \tag{3.16}$$

But (3.16), together with the strong convexity of φ, then implies

$$\begin{aligned} T'(\mathbf{x}; \mathbf{y} - \mathbf{x}) &= \nabla f(\mathbf{x})^{\mathrm{T}}(\mathbf{y} - \mathbf{x}) + \sup\left\{\boldsymbol{\xi}_u^{\mathrm{T}}(\mathbf{y} - \mathbf{x}) \,\middle|\, \boldsymbol{\xi}_u \in \partial u(\mathbf{x})\right\} \\ &\leq [\nabla\varphi(\mathbf{x}) - \nabla\varphi(\mathbf{y})]^{\mathrm{T}}(\mathbf{y} - \mathbf{x}) \\ &\leq -m_\varphi \|\mathbf{y} - \mathbf{x}\|^2. \end{aligned}$$

(e) From (3.15) and the Lipschitz continuity of $\nabla\varphi$, it follows that $\|\nabla f(\mathbf{x}) + \boldsymbol{\xi}_u\| \leq M_{\nabla\varphi}\|\mathbf{y}-\mathbf{x}\|$. Together with (3.12) it gives that (3.13) holds. □

The remarks made on Lemma 2.1 apply here, too. The property (3.13) is a generalization of the gradient relatedness property (2.5).

The set $\mathtt{dom}\ u$ need not be closed, even if u is l.s.c. (cf. [Roc70a, p. 52]). When extending the continuity properties of the direction finding map $\mathbf{x} \mapsto D(\mathbf{x}) = Y(\mathbf{x}) - \mathbf{x}$ from Lemma 2.2, we must therefore sometimes replace it with a closed subset $S \subseteq \mathtt{dom}\ u$. (In Theorem 3.2, the level set $L^T_{\Re^n}(\mathbf{x}^0)$ is one such set S.)

Lemma 3.2 (Continuity properties of the search direction) *Let $\varphi : \Re^{2n} \mapsto \Re$ be a continuous function on $\mathtt{dom}\ u \times \mathtt{dom}\ u$ of the form $\varphi(\mathbf{x},\mathbf{y})$, convex and in C^1 on $\mathtt{dom}\ u$ with respect to $\mathbf{x}$. Let $S \subseteq \mathtt{dom}\ u$ be an arbitrary, closed set in $\Re^n$. Let Assumption 3.1 hold.*

(a) *The mapping $\mathbf{x} \mapsto D(\mathbf{x})$ is closed on S.*

(b) *The set $Y(\mathbf{x})$ is nonempty and bounded* [7] *if and only if*

$$-\nabla f(\mathbf{x}) + \nabla_{\mathbf{x}}\varphi(\mathbf{x},\mathbf{x}) \in \mathtt{int}\ (\mathtt{dom}\ (u + \varphi(\cdot,\mathbf{x}))^\circ). \tag{3.17}$$

(c) *If, in addition to the assumptions that (3.17) holds for every $\mathbf{x} \in S$, u is strictly convex on S or $\varphi(\cdot,\mathbf{x})$ is strictly convex on S for every $\mathbf{x} \in S$, then the mapping $\mathbf{x} \mapsto D(\mathbf{x})$ is continuous on S.*

Proof

(a) Let $\{\mathbf{x}^k\} \subseteq S$ and $\{\mathbf{y}^k\}$ be sequences in $\mathtt{dom}\ u$ with $\{\mathbf{x}^k\} \to \mathbf{x}$, $\{\mathbf{y}^k\} \to \mathbf{y}$ and $\mathbf{y}^k \in Y(\mathbf{x}^k), \forall k$. Then (3.15) holds for all k, i.e., $-\nabla_{\mathbf{x}}\varphi(\mathbf{y}^k,\mathbf{x}^k) - \nabla f(\mathbf{x}^k) + \nabla_{\mathbf{x}}\varphi(\mathbf{x}^k,\mathbf{x}^k)$ is a subgradient of u at $\mathbf{y}^k$. Using [Roc70a, Th. 24.4] and the continuity of ∇f and $\nabla_{\mathbf{x}}\varphi$ on $\mathtt{dom}\ u$ and $\mathtt{dom}\ u \times \mathtt{dom}\ u$, respectively, we may conclude that

$$-\nabla_{\mathbf{x}}\varphi(\mathbf{y},\mathbf{x}) - \nabla f(\mathbf{x}) + \nabla_{\mathbf{x}}\varphi(\mathbf{x},\mathbf{x}) \in \partial u(\mathbf{y}), \tag{3.18}$$

which, by the convexity of u and $\varphi(\cdot,\mathbf{x})$, yields that $\mathbf{y} \in Y(\mathbf{x})$. Hence, the map $\mathbf{x} \mapsto Y(\mathbf{x})$ is closed on S.

The proof of the closedness of $\mathbf{x} \mapsto D(\mathbf{x})$ then follows from precisely the same arguments as in Lemma 2.2.a.

(b) Follows from the same line of argument as in Lemma 2.2.b, using the characterization (3.18) of $Y(\mathbf{x})$.

(c) Follows directly from **(a)**, and from the fact that the mapping is point-to-point, since $Y(\mathbf{x})$ is singleton (cf. Lemma 3.1.c). □

3.6 Basic convergence of cost approximation algorithms

Below, we extend the basic convergence Theorem 2.2 to nondifferentiable programming.

[7] From **(a)** the mapping $\mathbf{x} \mapsto D(\mathbf{x})$ is u.s.c. on any compact subset of $\mathtt{dom}\ u$; see [Nik68, Le. 4.4].

Theorem 3.2 (Basic convergence of cost approximation algorithms) *Let u be continuous on* $\mathtt{dom}\ u$. *Let $\varphi : \Re^{2n} \mapsto \Re$ be a continuous function on* $\mathtt{dom}\ u \times \mathtt{dom}\ u$ *of the form $\varphi(\mathbf{x}, \mathbf{y})$, convex and in C^1 on* $\mathtt{dom}\ u$ *with respect to* $\mathbf{x}$. *Assume that the point $\mathbf{x}^0$ is chosen so that the level set $L^T_{\Re^n}(\mathbf{x}^0)$ is bounded, and assume further that the problem* [$\mathbf{NNLP}_\varphi$] *is well defined, in the sense that $Y(\mathbf{x})$ is nonempty and bounded for every $\mathbf{x} \in$* $\mathtt{dom}\ u$. *Assume that Assumption 3.1 holds. Then, under Rule M, $\{T(\mathbf{x}^k)\} \to T(\overline{\mathbf{x}})$ for some $\overline{\mathbf{x}} \in \Omega$, any accumulation point $\mathbf{x}^\infty$ of the sequence $\{\mathbf{x}^k\}$ (at least one such point exists) lies in Ω, and*

$$\left\{ \inf_{\mathbf{x} \in \Omega \cap L^T_{\Re^n}(\mathbf{x}^0)} \left\| \mathbf{x}^k - \mathbf{x} \right\| \right\} \to 0. \tag{3.19}$$

Proof The proof of the theorem is a slight extension of the proof of Theorem 2.2; we only describe the major changes.

(1) Because of the level set boundedness assumption (closedness of the level set follows from e.g. [StW70, p. 134]) and the descent property of the algorithm, the sequence $\{\mathbf{x}^k\}$ lies in a compact set.

(2) The descent property follows from Lemma 3.1.b. The termination property ($\mathbf{x}^k \in \Omega$) follows from Lemma 3.1.a.

(3) The mapping D is closed on $L^T_{\Re^n}(\mathbf{x}^0)$ from Lemma 3.2.a, and the same is true for the mapping M, since T is continuous on $\mathtt{dom}\ u$ [Zan69b, Le. 5.1]. Since $L^T_{\Re^n}(\mathbf{x}^0)$ is bounded, [Zan69b, Cor. 4.2.1] yields that the map $A = MD$ is closed on $L^T_{\Re^n}(\mathbf{x}^0)$.

The last statement follows from Lemma 2.3, as in Theorem 2.2. □

For the special case $u \equiv w + \delta_X$, where X is a compact convex set, a direct proof not using Zangwill's Theorem A is given in [Pat91a].

Note that it is essential that u is continuous on $L^T_{\Re^n}(\mathbf{x}^0)$, since otherwise the line search map M need not be closed on this set.

There are several conditions ensuring the well-posedness of [$\mathbf{NNLP}_\varphi$]. If $u + \varphi(\cdot, \mathbf{x})$ is coercive, then it is also cofinite (cf. Definition A.3.b); from [Roc70a, Cor. 13.3.1] it then follows that $\mathtt{dom}\ (u + \varphi(\cdot, \mathbf{x}))^\circ = \Re^n$, and Condition (3.17) holds. Another sufficient condition for (3.17) to hold is that $\mathtt{dom}\ u$ is bounded (an important special case is $u \equiv w + \delta_X$, with X compact); T_φ is then effectively minimized on a compact set.

The most natural assumption ensuring well-posedness, and which leads to stronger convergence results, is that the functions φ^k are strongly convex on $\mathtt{dom}\ u$. Under this assumption, we will strengthen the convergence result obtained in Theorem 3.2. We also show that the convergence conditions for alternative step length rules are very similar to those obtained in the differentiable case.

3.7 Convergence under alternative step length rules

The rules studied here are a variant of the Armijo Rule A and the predetermined step length Rule P.

Due to the nondifferentiability of T the Armijo rule must be modified. First, since u is allowed to include an indicator function of a convex set, the unit step is the largest step that is always guaranteed to be valid. Second, the fact that the generalized gradient can not be Lipschitz continuous makes it necessary to replace the directional derivative in Rule A with a monotonicity measure in $\nabla\varphi^k$.

Rule A' Let $\alpha \in (0,1)$ and $t_k = \beta^{\bar{\imath}}$, where $\beta \in (0,1)$ and $\bar{\imath}$ is the smallest nonnegative integer i such that

$$\alpha \leq \frac{T\left(\mathbf{x}^k + \beta^i \mathbf{d}^k\right) - T\left(\mathbf{x}^k\right)}{\beta^i \left[\nabla\varphi^k\left(\mathbf{x}^k\right) - \nabla\varphi^k\left(\mathbf{y}^k\right)\right]^{\mathrm{T}} \mathbf{d}^k}. \tag{3.20}$$

As is easily seen from Equation (2.3), Rule A' reduces to the original Armijo Rule A whenever $u \equiv 0$.

Lemma 3.3 (Decrease of T in the direction $\mathbf{d}$) *Let ∇f be Lipschitz continuous on* $\mathtt{dom}\, u$. *Let $\varphi : \Re^n \mapsto \Re$ be a convex function in C^1 on* $\mathtt{dom}\, u$. *Let Assumption 3.1 hold,* $\mathbf{x} \in \mathtt{dom}\, u$, $\mathbf{y} \in Y(\mathbf{x})$, *and* $\mathbf{d} = \mathbf{y} - \mathbf{x}$. *Then, for* $t \in [0,1]$,

$$T(\mathbf{x} + t\mathbf{d}) - T(\mathbf{x}) \leq -t\left[\nabla\varphi(\mathbf{y}) - \nabla\varphi(\mathbf{x})\right]^{\mathrm{T}} \mathbf{d} + \frac{M_{\nabla f}}{2} t^2 \left\|\mathbf{d}\right\|^2. \tag{3.21}$$

Proof Taylor's formula and the Lipschitz continuity of ∇f yields, for $t \in [0,1]$,

$$\begin{aligned} f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x}) &= \int_0^t \nabla f(\mathbf{x} + s\mathbf{d})^{\mathrm{T}}\mathbf{d}\, ds \\ &\leq t\nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d} + \int_0^t sM_{\nabla f} \left\|\mathbf{d}\right\|^2 \, ds \\ &= t\nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d} + \frac{M_{\nabla f}}{2} t^2 \left\|\mathbf{d}\right\|^2. \end{aligned} \tag{3.22}$$

For $t \in [0,1]$,

$$\begin{aligned} u(\mathbf{x} + t\mathbf{d}) - u(\mathbf{x}) &= \int_0^t u'(\mathbf{x} + s\mathbf{d}; \mathbf{d})\, ds && \text{([Roc70a, Cor. 24.2.1])} \\ &\leq -tu'(\mathbf{x} + \mathbf{d}; -\mathbf{d}) && \text{([Roc70a, Th. 24.1])} \\ &= t\inf\left\{\boldsymbol{\xi}_u^{\mathrm{T}}\mathbf{d} \,\middle|\, \boldsymbol{\xi}_u \in \partial u(\mathbf{y})\right\} && \text{([Roc66])} \\ &\leq t\boldsymbol{\xi}_u^{\mathrm{T}}\mathbf{d}, \ \forall \boldsymbol{\xi}_u \in \partial u(\mathbf{y}). \end{aligned} \tag{3.23}$$

Combining (3.22) and (3.23), choosing $\boldsymbol{\xi}_u \in \partial u(\mathbf{y})$ such that [cf. (3.15)]

$$-\nabla\varphi(\mathbf{y}) - \nabla f(\mathbf{x}) + \nabla\varphi(\mathbf{x}) = \boldsymbol{\xi}_u \tag{3.24}$$

holds, we then have

$$\begin{aligned} T(\mathbf{x} + t\mathbf{d}) - T(\mathbf{x}) &= \left[f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})\right] + \left[u(\mathbf{x} + t\mathbf{d}) - u(\mathbf{x})\right] \\ &\leq t\left[\nabla f(\mathbf{x}) + \boldsymbol{\xi}_u\right]^{\mathrm{T}} \mathbf{d} + \frac{M_{\nabla f}}{2} t^2 \left\|\mathbf{d}\right\|^2 \\ &= t\left[\nabla\varphi(\mathbf{x}) - \nabla\varphi(\mathbf{y})\right]^{\mathrm{T}} \mathbf{d} + \frac{M_{\nabla f}}{2} t^2 \left\|\mathbf{d}\right\|^2. \end{aligned}$$

The lemma is proved. □

Lemma 3.4 (Validity of Rule A') *Under the assumptions of Lemma 3.3, and the additional conditions that φ is strictly convex on* $\mathtt{dom}\, u$ *and that* $\mathbf{x} \notin \Omega$, *there exists a finite* $\bar{\imath}$ *satisfying (3.20).*

Proof Using (3.21) and the strict monotonicity of $\nabla\varphi$, by choosing

$$t \leq 2(\alpha-1)\frac{\left[\nabla\varphi(\mathbf{x})-\nabla\varphi(\mathbf{y})\right]^{\mathrm{T}}\mathbf{d}}{M_{\nabla f}\left\|\mathbf{d}\right\|^2},$$

we obtain

$$T(\mathbf{x}+t\mathbf{d})-T(\mathbf{x}) \leq \alpha t\left[\nabla\varphi(\mathbf{x})-\nabla\varphi(\mathbf{y})\right]^{\mathrm{T}}\mathbf{d}. \tag{3.25}$$

Replacing t with $\beta^{\bar{\imath}}$, we see that $\bar{\imath}$ is the smallest integer i to satisfy

$$\beta^i \leq 2(\alpha-1)\frac{\left[\nabla\varphi(\mathbf{x})-\nabla\varphi(\mathbf{y})\right]^{\mathrm{T}}\mathbf{d}}{M_{\nabla f}\left\|\mathbf{d}\right\|^2}. \tag{3.26}$$

Hence, the Inequality (3.20) will be satisfied after a finite number of steps. □

Lemmas 3.3 and 3.4 were first proved in [Pat91a] for the problem (1.3).

Remark 3.1 From (3.26) we conclude that the step length $\tilde{t}$ given by Rule A' satisfies

$$\frac{1}{\beta}\tilde{t} = \beta^{(\bar{\imath}-1)} > 2(\alpha-1)\frac{\left[\nabla\varphi(\mathbf{x})-\nabla\varphi(\mathbf{y})\right]^{\mathrm{T}}\mathbf{d}}{M_{\nabla f}\left\|\mathbf{d}\right\|^2},$$

i.e.,

$$\tilde{t} > 2\beta(\alpha-1)\frac{\left[\nabla\varphi(\mathbf{x})-\nabla\varphi(\mathbf{y})\right]^{\mathrm{T}}\mathbf{d}}{M_{\nabla f}\left\|\mathbf{d}\right\|^2}. \tag{3.27}$$

From (3.26) we may also find conditions under which the unit step, i.e., $\bar{\imath}=0$, is accepted by Rule A'. Especially, for a strongly convex function φ, the Armijo Rule A' yields a unit step ($\bar{\imath}=0$) if

$$\frac{2(1-\alpha)m_\varphi}{M_{\nabla f}} \geq 1. \tag{3.28}$$

Note that, from (3.27), the step length given by Rule A' always satisfies

$$t \geq \min\left\{1, \frac{2\beta(1-\alpha)\left[\nabla\varphi(\mathbf{y})-\nabla\varphi(\mathbf{x})\right]^{\mathrm{T}}\mathbf{d}}{M_{\nabla f}\left\|\mathbf{d}\right\|^2}\right\},$$

which reduces to

$$t \geq \min\left\{1, \frac{2\beta(1-\alpha)m_\varphi}{M_{\nabla f}}\right\}, \tag{3.29}$$

whenever φ is strongly convex.[8]

Theorem 3.3 (Convergence under various step length rules) [9, 10] *Let ∇f be Lipschitz continuous on* $\texttt{dom}\,u$*, and assume that T is bounded from below on $\Re^n$. Let $\varphi^k:\Re^n\mapsto\Re$ be a strongly convex function in C^2 on* $\texttt{dom}\,u$ *with a Lipschitz continuous gradient on* $\texttt{dom}\,u$*. Let $0<m_\varphi\leq m_{\varphi^k}$ and $M_{\nabla\varphi^k}\leq M_{\nabla\varphi}<+\infty$, $\forall k$. Let Assumption 3.1 hold. Then, from any starting point $\mathbf{x}^0\in$* $\texttt{dom}\,u$*, under Rules M and A' any accumulation point $\mathbf{x}^\infty$ of $\{\mathbf{x}^k\}$ lies in Ω. Furthermore, if $\{\mathbf{x}^k\}$ is bounded and Ω is nonempty, then (3.19) holds. If further Ω is finite, and the step lengths t_k are upper bounded, $\{\mathbf{x}^k\}$ converges to one point in Ω. For Rule P, the same conclusions hold if $t_k\leq 1$, and either t_k belongs to a compact subset of $(0,\frac{2m_{\varphi^k}}{M_{\nabla f}})$, or $t_k\equiv t\in(0,\frac{2m_\varphi}{M_{\nabla f}})$, $\forall k$.*

[8]The results apply also to the differentiable case, since $u\equiv 0$ is allowed.

Proof We begin by showing that for all the three rules we must have that $\{\mathbf{d}^k\} \to \mathbf{0}$.

Rule M From (3.21) and the strong convexity of φ^k we obtain that, for $t \in [0,1]$,

$$T\left(\mathbf{x}^{k+1}\right) \leq T\left(\mathbf{x}^k + t\mathbf{d}^k\right) \leq T\left(\mathbf{x}^k\right) + \left(-tm_{\varphi^k} + \frac{M_{\nabla f}}{2}t^2\right)\left\|\mathbf{d}^k\right\|^2. \quad (3.30)$$

Minimizing the right-hand side of this inequality, we obtain that $t = \bar{t} = \frac{m_{\varphi^k}}{M_{\nabla f}}$. Hence, if $\bar{t} \leq 1$, we have

$$T\left(\mathbf{x}^{k+1}\right) - T\left(\mathbf{x}^k\right) \leq -\frac{m^2_{\varphi^k}}{2M_{\nabla f}}\left\|\mathbf{d}^k\right\|^2,$$

while, if $\bar{t} > 1$, we have

$$T\left(\mathbf{x}^{k+1}\right) - T\left(\mathbf{x}^k\right) \leq \left(-m_{\varphi^k} + \frac{M_{\nabla f}}{2}\right)\left\|\mathbf{d}^k\right\|^2,$$

where $-m_{\varphi^k} + \frac{M_{\nabla f}}{2} < 0$.

Rule A' From (3.25), (3.27) and the strong convexity of φ^k, we have

$$\begin{aligned}
T\left(\mathbf{x}^{k+1}\right) - T\left(\mathbf{x}^k\right) &\leq -\alpha t_k\left[\nabla\varphi^k\left(\mathbf{y}^k\right) - \nabla\varphi^k\left(\mathbf{x}^k\right)\right]^{\mathrm{T}}\mathbf{d}^k \\
&\leq -\frac{2\beta\alpha(1-\alpha)}{M_{\nabla f}\|\mathbf{d}^k\|^2}\left(\left[\nabla\varphi^k\left(\mathbf{y}^k\right) - \nabla\varphi^k\left(\mathbf{x}^k\right)\right]^{\mathrm{T}}\mathbf{d}^k\right)^2 \\
&\leq -\frac{2\beta\alpha(1-\alpha)m^2_{\varphi^k}}{M_{\nabla f}}\left\|\mathbf{d}^k\right\|^2.
\end{aligned}$$

Rule P From (3.30) we obtain a decrease in T whenever $t_k \leq 1$ and either t_k belongs to a compact subset of $(0, \frac{2m_{\varphi^k}}{M_{\nabla f}})$, or $t_k \equiv t \in (0, \frac{2m_{\varphi}}{M_{\nabla f}})$.

[9] The assumptions for Rule A' are in [Pat91a] slightly relaxed for functions u with compact effective domains; if φ is of the form $\varphi(\mathbf{x},\mathbf{y})$, strong convexity in C^2 on `dom` u of φ may be replaced by only strict convexity in C^1 on `dom` u with respect to $\mathbf{x}$.

[10] Under the additional assumptions of convexity of f (so that T is convex, and (3.1) is sufficient for global optimality) and boundedness of $\{\mathbf{x}^k\}$, the theorem also holds when the functions φ^k are in C^1 on `dom` u. Indeed, by tracing the proof we obtain that $\{\mathbf{d}^k\} \to \mathbf{0}$, for all the three step length rules. We then have, for any optimal solution $\mathbf{x}^*$, and any k, that

$$\begin{aligned}
T(\mathbf{x}^*) - T\left(\mathbf{y}^k\right) &\geq \nabla f\left(\mathbf{y}^k\right)^{\mathrm{T}}\left(\mathbf{x}^* - \mathbf{y}^k\right) + u\left(\mathbf{x}^*\right) - u\left(\mathbf{y}^k\right) \quad [f \text{ convex}] \\
&\geq \left[\nabla f\left(\mathbf{y}^k\right) - \nabla f\left(\mathbf{x}^k\right)\right]^{\mathrm{T}}\left(\mathbf{x}^* - \mathbf{y}^k\right) \\
&\quad + \left[\nabla\varphi^k\left(\mathbf{x}^k\right) - \nabla\varphi^k\left(\mathbf{y}^k\right)\right]^{\mathrm{T}}\left(\mathbf{x}^* - \mathbf{y}^k\right). \quad \left[\mathbf{y}^k \in Y\left(\mathbf{x}^k\right)\right]
\end{aligned}$$

From the Lipschitz continuity of ∇f and $\nabla\varphi^k$, and the boundedness of $\{\mathbf{x}^k\}$ (it is easy to show that this, together with the strong convexity of the φ^k, implies boundedness of $\{\mathbf{y}^k\}$), the right hand side of the above inequality tends to zero. Let $\mathbf{x}^\infty$ be an accumulation point of $\{\mathbf{x}^k\}$, corresponding to a convergent subsequence $\mathcal{K}$. Then, from the above inequality, the l.s.c. property of T, and the fact that $\{\mathbf{d}^k\} \to \mathbf{0}$,

$$T\left(\mathbf{x}^*\right) \geq \lim_{k \in \mathcal{K}} T\left(\mathbf{y}^k\right) \geq T\left(\mathbf{x}^\infty\right).$$

This shows that $\mathbf{x}^\infty$ is an optimal solution to [**NNLP**]. The result here obtained can be immediately used for [**NLP**], i.e., in Theorem 2.3.b, and extends directly to constrained programs [**CNLP**].

Since T is lower bounded and, from the above, $\{T(\mathbf{x}^k)\}$ is decreasing, $\{T(\mathbf{x}^{k+1})-T(\mathbf{x}^k)\} \to 0$. Thus, in all the three rules we obtain that $\{\mathbf{d}^k\} \to \mathbf{0}$.

From (3.24) and the upper bound on $M_{\nabla\varphi^k}$, as $\{\mathbf{d}^k\} \to \mathbf{0}$,

$$\left\{\nabla f\left(\mathbf{x}^k\right) + \boldsymbol{\xi}_u^k\right\} = \left\{\nabla\varphi^k\left(\mathbf{x}^k\right) - \nabla\varphi^k\left(\mathbf{y}^k\right)\right\} \to \mathbf{0}. \tag{3.31}$$

Assume now that $\{\mathbf{x}^k\}$ has an accumulation point, $\mathbf{x}^\infty$, corresponding to a convergent subsequence $\mathcal{K}$. Since $\{\mathbf{y}^k - \mathbf{x}^k\} = \{\mathbf{d}^k\} \to \mathbf{0}$, we have that $\{\mathbf{y}^k\}_{k\in\mathcal{K}} \to \mathbf{x}^\infty$. From the closedness of ∂u [Roc70a, Th. 24.4] and the continuity of ∇f, (3.31) yields $\{\boldsymbol{\xi}_u^k\}_{k\in\mathcal{K}} \to \boldsymbol{\xi}_u^\infty \in \partial u(\mathbf{x}^\infty)$, and, again appealing to (3.31), $\nabla f(\mathbf{x}^\infty) + \boldsymbol{\xi}_u^\infty = \mathbf{0}$. Thus, $\mathbf{x}^\infty \in \Omega$.

The result (3.19) follows from Lemma 2.3, and convergence toward one point in Ω follows from (2.10) and Lemma 2.4. □

For the problem (1.3), some of the results for Rule P are obtained in [Coh80].

Remark 3.2 To ensure that a unit step implies convergence in Rule P, we may choose φ such that

$$\frac{2m_\varphi}{M_{\nabla f}} > 1. \tag{3.32}$$

This condition is actually approached by Condition (3.28) when $\alpha \to 0$.[11]

The maximal allowed step length in Rule P is, from the above theorem, bounded by the constant $\frac{2m_{\varphi^k}}{M_{\nabla f}}$. Now, suppose that f may be written as $f = f_1 + f_2$, where f_1 and f_2 are convex with Lipschitz continuous gradients, and that f_1 is strongly convex. For the sake of this example, we further assume that f_1 is quadratic. To enhance the speed of convergence of the cost approximation algorithm, we consider redefining φ^k as $\varphi^k := \varphi^k + f_1$. (This operation corresponds to applying cost approximation to the function $[u+f_1]+f_2$, instead of to $u+f$.) Indeed, this operation yields an increased maximal allowed step length, since

$$\frac{2m_{\varphi^k+f_1}}{M_{\nabla f_2}} = \frac{2(m_{\varphi^k} + m_{f_1})}{M_{\nabla(f-f_1)}} > \frac{2m_{\varphi^k}}{M_{\nabla f}}.$$

It also yields a faster convergence for the other step length rules; for Rule A', cf. the change in the linear convergence ratio in Theorem 3.5.

From this example, we are lead to conclude that, in order to achieve the best possible rate of convergence, the function $\varphi^k := \varphi^k + f_1$ should be constructed such that $f_2 = f - f_1$ is *not* strongly convex. (The desire to obtain a high rate of convergence must of course be weighed against the computational difficulty of the subproblems.)

Similar observations can be made for the other instances of [**GVIP**] under study in this thesis.

Chen and Rockafellar [ChR92] study splitting methods for the problem of finding a zero of the sum of two maximal monotone operators. They argue that all the strong monotonicity that is inherent in a problem defining mapping should be kept in the mapping defining the *forward step*; this result corresponds to the above for this special case of cost approximation (see Section 4.3).

[11] The result applies also to the differentiable case, since $u \equiv 0$ is allowed.

3.8 A truncated version of the cost approximation algorithms

From Lemma 3.1.b we know that $\overline{\mathbf{x}}-\mathbf{x}^k$ is a descent direction with respect to T whenever $T_\varphi(\overline{\mathbf{x}}) < T_\varphi(\mathbf{x}^k)$. The idea behind the *truncated cost approximation method* is to reduce the work performed on [$\mathbf{NNLP}_\varphi^k$] by limiting the number of iterations performed solving it with a descent algorithm. This strategy introduces a trade-off between the computational effort spent on solving the subproblem and the quality of the search direction obtained.

Theorem 3.4 (Convergence of truncated cost approximation algorithms) *Let u be continuous on* $\mathtt{dom}\, u$*. Let $\varphi : \Re^{2n} \mapsto \Re$ be a continuous function on* $\mathtt{dom}\, u \times \mathtt{dom}\, u$ *of the form $\varphi(\mathbf{x}, \mathbf{y})$, convex and in C^1 on* $\mathtt{dom}\, u$ *with respect to* $\mathbf{x}$*. Furthermore, let Assumption 3.1 hold. Assume that $\mathbf{x}^0$ is chosen such that $L^T_{\Re^n}(\mathbf{x}^0)$ is bounded, and that the algorithm used for solving* [$\mathbf{NNLP}_\varphi^k$] *is a descent algorithm with a closed algorithmic map. If the termination criteria chosen for* [$\mathbf{NNLP}_\varphi^k$] *are such that $1 \le l^k < \infty$ iterations are performed, $\forall k$, the sequence $\{\mathbf{y}^k\}$ of approximate subproblem solutions is bounded, and if Rule M is used, then $\{T(\mathbf{x}^k)\} \to T(\overline{\mathbf{x}})$ for some $\overline{\mathbf{x}} \in \Omega$, any accumulation point $\mathbf{x}^\infty$ of the sequence $\{\mathbf{x}^k\}$ (at least one such point exists) lies in Ω, and (3.19) holds. Under the additional assumptions that ∇f is Lipschitz continuous on* $\mathtt{dom}\, u$ *and that φ is strictly convex on* $\mathtt{dom}\, u$*, the same conclusions hold under Rule A'.*

Proof From the compactness of the level set and the descent property of the algorithm, $\{\mathbf{x}^k\}$ is bounded, and since $\{\mathbf{y}^k\}$ is also bounded, we conclude that there exists a subsequence $\mathcal{K}$ such that $\{\mathbf{x}^k\}_{k\in\mathcal{K}} \to \mathbf{x}^\infty$ and $\{\mathbf{y}^k\}_{k\in\mathcal{K}} \to \mathbf{y}^\infty$.

From the use of Rule M, we may invoke the proof of [Pat91a, Th. 4.2] to conclude that there exists a $\boldsymbol{\xi}_u^\infty \in \partial u(\mathbf{x}^\infty)$ such that

$$[\nabla f(\mathbf{x}^\infty) + \boldsymbol{\xi}_u^\infty]^{\mathrm{T}} (\mathbf{y}^\infty - \mathbf{x}^\infty) \ge 0, \tag{3.33}$$

i.e., in the limit no descent direction can be obtained.

From the relation

$$T_\varphi\left(\mathbf{y}^k\right) < T_\varphi\left(\mathbf{x}^k\right) \tag{3.34}$$

it follows, in the limit, using the convexity of φ, that, for all $\boldsymbol{\xi}_u \in \partial u(\mathbf{x}^\infty)$,

$$[\nabla f(\mathbf{x}^\infty) + \boldsymbol{\xi}_u]^{\mathrm{T}} (\mathbf{y}^\infty - \mathbf{x}^\infty) \le 0, \tag{3.35}$$

so that, by combining (3.33) and (3.35) we obtain

$$[\nabla f(\mathbf{x}^\infty) + \boldsymbol{\xi}_u^\infty]^{\mathrm{T}} (\mathbf{y}^\infty - \mathbf{x}^\infty) = 0. \tag{3.36}$$

From (3.34) we have, in the limit, that

$$T_\varphi(\mathbf{y}^\infty) \le T_\varphi(\mathbf{x}^\infty). \tag{3.37}$$

We also have, since $\boldsymbol{\xi}_u^\infty \in \partial u(\mathbf{x}^\infty)$,

$$\begin{aligned} T_\varphi(\mathbf{y}^\infty) &\ge T_\varphi(\mathbf{x}^\infty) + [\nabla f(\mathbf{x}^\infty) + \boldsymbol{\xi}_u^\infty]^{\mathrm{T}} (\mathbf{y}^\infty - \mathbf{x}^\infty) \quad && (T_\varphi \text{ convex}) \\ &= T_\varphi(\mathbf{x}^\infty). && (3.36) \end{aligned} \tag{3.38}$$

Thus,

$$T_\varphi(\mathbf{y}^\infty) = T_\varphi(\mathbf{x}^\infty). \tag{3.39}$$

By observing that the boundedness of the sequence $\{\mathbf{y}^k\}$ implies, by [Zan69b, Cor. 4.2.1], that the algorithmic map MD is closed, the rest of the proof is exactly like that of [Pat91b, Th. 3.3], and we conclude that $\mathbf{y}^\infty \in Y(\mathbf{x}^\infty)$.[12] From (3.39) it then follows that $\mathbf{x}^\infty \in Y(\mathbf{x}^\infty)$, i.e., that $\mathbf{x}^\infty \in \Omega$.

Assuming further that ∇f is Lipschitz continuous on $\mathtt{dom}\, u$ and that φ is strictly convex on $\mathtt{dom}\, u$, Rule A' is valid. Tracing the proof of [Pat91a, Th. 4.3], we may conclude that (3.33) holds, and the rest of the proof follows exactly as above. □

The first part of the theorem extends that of [Pat91b] for [**CNLP**].

3.9 Linear convergence of cost approximation algorithms

The analysis of the asymptotic convergence rate of cost approximation methods is not as easily extended from the differentiable case as are the basic convergence results. This is because the classical analysis for gradient related and deflected gradient methods rely heavily on the Lipschitz continuity of the gradient of the function to be minimized. It does not seem to be possible to replace that condition with a corresponding one for the differentiable part of T. We instead analyze the cost approximation algorithm by means of some elements in operator theory, using in particular results on the nonexpansiveness of operators (see Definition A.5.a). In the analysis, we shall limit ourselves to studying quadratic functions φ. (It is an open question if linear convergence can be obtained for nonquadratic functions.)

Lemma 3.5 (Characterization of the subproblem solution) *Let $\varphi : \Re^n \mapsto \Re$ be given by $\varphi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}$, where $\mathbf{B}$ is a symmetric and positive definite matrix on $\mathtt{dom}\, u$ with eigenvalues in the interval $[m_\varphi, M_{\nabla\varphi}]$. Let Assumption 3.1 hold. Then, for any $\mathbf{x} \in \mathtt{dom}\, u$,*

$$\mathbf{y} = \mathbf{y}(\mathbf{x}) = [\mathbf{B} + \partial u]^{-1}[\mathbf{B} - \nabla f](\mathbf{x}). \tag{3.40}$$

Proof From the optimality conditions (3.15) of [$\mathbf{NNLP}_\varphi$] we have that

$$[\mathbf{B} - \nabla f](\mathbf{x}) \in [\mathbf{B} + \partial u](\mathbf{y}(\mathbf{x})) \tag{3.41}$$

(see also Remark 1.1). Since $\mathbf{y}(\mathbf{x})$ is unique, the operator $[\mathbf{B} + \partial u]^{-1}$ must be single-valued, and we obtain from (3.41) the expression (3.40). □

This lemma is a generalization of [FuM81, Prop. 3.1], which is given for the special case $\mathbf{B} = \frac{1}{\gamma}\mathbf{I}$, $\gamma > 0$.

In order to prove a linear rate of convergence, we express (3.40) as $\mathbf{y}(\mathbf{x}) = \mathbf{PQ}(\mathbf{x})$, where

$$\mathbf{P} = \left[\frac{1}{m_\varphi}(\mathbf{B} + \partial u)\right]^{-1} \text{ and } \mathbf{Q} = \left[\frac{1}{m_\varphi}(\mathbf{B} - \nabla f)\right].$$

This rewriting of the characterization of $\mathbf{y}(\mathbf{x})$ is done, because the operator $\mathbf{P}$ can be shown to be nonexpansive.

[12]Thus, the subproblems are solved to optimality in the limit, although $\mathbf{y}^k \notin Y(\mathbf{x}^k)$ in general.

Lemma 3.6 (Nonexpansiveness of $\mathbf{P}$) *Under the conditions of Lemma 3.5, the operator* $\mathbf{P}$ *is nonexpansive on* $\texttt{dom}\, u$.

Proof The proof follows closely that of [EcB92, Th. 2], which is given for the special case of the proximal point algorithm (i.e., $\mathbf{B} = \frac{1}{\gamma}\mathbf{I}$, $\gamma > 0$).

From the monotonicity of ∂u [Roc70a, Cor. 31.5.2] and the positive definiteness of $\mathbf{B}$ we have, for all $\mathbf{x}, \mathbf{y} \in \texttt{dom}\, u$, and $\boldsymbol{\xi}_u(\mathbf{x}) \in \partial u(\mathbf{x})$, $\boldsymbol{\xi}_u(\mathbf{y}) \in \partial u(\mathbf{y})$,

$$\begin{aligned} [\boldsymbol{\xi}_u(\mathbf{x}) - \boldsymbol{\xi}_u(\mathbf{y})]^{\mathrm{T}} (\mathbf{x} - \mathbf{y}) &\geq 0, \text{ and} \\ (\mathbf{x} - \mathbf{y})^{\mathrm{T}} \mathbf{B} (\mathbf{x} - \mathbf{y}) &\geq m_\varphi \|\mathbf{x} - \mathbf{y}\|^2 . \end{aligned}$$

Dividing both the above inequalities by m_φ, and adding them, we obtain

$$\left(\left[\frac{1}{m_\varphi} (\mathbf{B} + \boldsymbol{\xi}_u) \right] (\mathbf{x}) - \left[\frac{1}{m_\varphi} (\mathbf{B} + \boldsymbol{\xi}_u) \right] (\mathbf{y}) \right)^{\mathrm{T}} (\mathbf{x} - \mathbf{y}) \geq \|\mathbf{x} - \mathbf{y}\|^2 .$$

But this inequality yields that the operator $\mathbf{P}$ is firmly nonexpansive (cf. Definition A.5.b). Moreover, firmly nonexpansive operators are nonexpansive, see e.g. [Roc76b, EcB92]. □

The importance of this lemma is twofold. First, the convergence analysis is simplified, since, using Lemma 3.1.a, we immediately have that for any $\mathbf{x}^*$ satisfying (3.1),

$$\|\mathbf{y}(\mathbf{x}) - \mathbf{x}^*\| = \|\mathbf{PQ}(\mathbf{x}) - \mathbf{PQ}(\mathbf{x}^*)\| \leq \|\mathbf{Q}(\mathbf{x}) - \mathbf{Q}(\mathbf{x}^*)\| ,$$

and it suffices to study the operator norm of $\mathbf{Q}$ in the vicinity of $\mathbf{x}^*$. Second, we may note that Lemma 3.6 does not utilize the fact that ∇f is integrable; hence, the convergence rate results obtained here carry directly over to variational inequalities (see Theorem 5.15).

Theorem 3.5 (Linear convergence of cost approximation algorithms) *Let* ∇f *be Lipschitz continuous on* $\texttt{dom}\, u$. *Let* $\varphi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}$, *where* $\mathbf{B}$ *is a symmetric and positive definite matrix on* $\texttt{dom}\, u$ *with eigenvalues in the interval* $[m_\varphi, M_{\nabla\varphi}]$. *Let Assumption 3.1 hold. Assume that, using Rule A' or P with* t_k *chosen in a compact subset of* $(0, \frac{2m_\varphi}{M_{\nabla f}})$, $\forall k$, $\{\mathbf{x}^k\} \to \mathbf{x}^* \in \Omega$, *where* f *is twice continuously differentiable and* $\nabla^2 f(\mathbf{x}^*)$ *is positive definite. Then if* $\mathbf{B}$ *is chosen so that*

$$\sigma \stackrel{\text{def}}{=} \frac{M_{\nabla\varphi}}{m_\varphi} \left\| \mathbf{I} - \mathbf{B}^{-1} \nabla^2 f(\mathbf{x}^*) \right\| < 1, \tag{3.42}$$

$\{\mathbf{x}^k\}$ *converges at least linearly with the ratio*

$$q \leq \max\left\{ \sigma, 1 - \frac{2\beta(1-\alpha)m_\varphi}{M_{\nabla f}}(1-\sigma) \right\}, \quad \textit{for Rule A', and}$$

$$q \leq \max\{\sigma, 1 - \underline{t}(1-\sigma)\} \qquad \textit{for Rule P,}$$

where $\underline{t} = \inf_k\{t_k\} > 0$.

Proof The proof is similar to that of [FuM81, Th. 5.1], given for the special case $\mathbf{B} = \frac{1}{\gamma}\mathbf{I}$, $\gamma > 0$, and for a step length rule similar to A'.

From Lemma 3.1.a and the nonexpansiveness property of $\mathbf{P}$, we have

$$\begin{aligned} \left\|\mathbf{y}^k - \mathbf{x}^*\right\| &= \left\|\mathbf{PQ}\left(\mathbf{x}^k\right) - \mathbf{PQ}(\mathbf{x}^*)\right\| \\ &\leq \left\|\mathbf{Q}\left(\mathbf{x}^k\right) - \mathbf{Q}(\mathbf{x}^*) - \nabla\mathbf{Q}(\mathbf{x}^*)\left(\mathbf{x}^k - \mathbf{x}^*\right)\right\| + \left\|\nabla\mathbf{Q}(\mathbf{x}^*)\left(\mathbf{x}^k - \mathbf{x}^*\right)\right\| , \end{aligned}$$

where $\nabla \mathbf{Q}(\mathbf{x}^*) = \frac{1}{m_\varphi}(\mathbf{B} - \nabla^2 f(\mathbf{x}^*))$.

Let $\varepsilon > 0$ be arbitrary. Since $\{\mathbf{x}^k\} \to \mathbf{x}^*$, we have for a sufficiently large k that

$$\left\| \mathbf{Q}\left(\mathbf{x}^k\right) - \mathbf{Q}\left(\mathbf{x}^*\right) - \nabla \mathbf{Q}\left(\mathbf{x}^*\right)\left(\mathbf{x}^k - \mathbf{x}^*\right) \right\| \leq \varepsilon \left\| \mathbf{x}^k - \mathbf{x}^* \right\|.$$

We also have that $\|\nabla \mathbf{Q}(\mathbf{x}^*)\| \leq \frac{M_{\nabla\varphi}}{m_\varphi} \|\mathbf{I} - \mathbf{B}^{-1}\nabla^2 f(\mathbf{x}^*)\| = \sigma$, so that

$$\left\| \mathbf{y}^k - \mathbf{x}^* \right\| \leq (\sigma + \varepsilon) \left\| \mathbf{x}^k - \mathbf{x}^* \right\|.$$

Then,

$$\begin{aligned} \left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\| &= \left\| \mathbf{x}^k + t_k \mathbf{d}^k - \mathbf{x}^* \right\| \\ &\leq t_k \left\| \mathbf{y}^k - \mathbf{x}^* \right\| + (1 - t_k) \left\| \mathbf{x}^k - \mathbf{x}^* \right\| \\ &\leq [1 - t_k(1 - \sigma - \varepsilon)] \left\| \mathbf{x}^k - \mathbf{x}^* \right\|. \end{aligned}$$

For Rule A', from (3.29) we then have, for a large enough k,

$$\left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\| \leq q_\varepsilon \left\| \mathbf{x}^k - \mathbf{x}^* \right\|,$$

where

$$q_\varepsilon = \max\left\{\sigma + \varepsilon, 1 - \frac{2\beta(1-\alpha)m_\varphi}{M_{\nabla f}}(1 - \sigma - \varepsilon)\right\},$$

while, for Rule P, the same holds with

$$q_\varepsilon = \max\left\{\sigma + \varepsilon, 1 - \underline{t}(1 - \sigma - \varepsilon)\right\}.$$

Since ε was arbitrary we have asymptotically

$$q = \limsup_{k \to \infty} \frac{\left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\|}{\left\| \mathbf{x}^k - \mathbf{x}^* \right\|} \leq q_0,$$

where

$$q_0 = \max\left\{\sigma, 1 - \frac{2\beta(1-\alpha)m_\varphi}{M_{\nabla f}}(1 - \sigma)\right\},$$

in the case of Rule A', and

$$q_0 = \max\left\{\sigma, 1 - \underline{t}(1 - \sigma)\right\},$$

for Rule P. If $\sigma < 1$, then $q_0 < 1$, and the theorem is proved. □

Remark 3.3 If (3.32) holds, $t_k \equiv 1$ is a valid step length in Rule P. Then, $q \leq \sigma$, and the choice $\mathbf{B} = \nabla^2 f(\mathbf{x}^*)$ yields *superlinear* convergence. For Rule A', the same conclusion may be drawn, provided (3.28) holds.

Corollary 3.1 (Linear convergence of some instances of cost approximation algorithms) *Let the assumptions of Theorem 3.5 hold. Furthermore, let* $\mathbf{B} = \frac{1}{\gamma}\mathbf{I}$, *for some* $\gamma > 0$.

(a) *The Criterion (3.42) for linear convergence under Rule A' or P reduces to the Condition (3.32). Thus, for Rule P the criterion for convergence under unit steps implies linear convergence.*

(b) *Assume that Rule P is used, with step $t_k \equiv 1$. Then the convergence ratio is*

$$q \leq \max\{|1-\gamma m_f|, |1-\gamma M_{\nabla f}|\}. \tag{3.43}$$

(c) *Assume further that $u \equiv \delta_X$, where X is a nonempty, closed and convex set in $\Re^n$. Then the algorithm is equivalent to the gradient projection algorithm of [Gol64, LeP66]. Moreover, the Criterion (3.32) and Ratio (3.43) coincide with those of this algorithm.*

(d) *Assume further that $u \equiv 0$. Then the algorithm is equivalent to the steepest descent method with fixed step length γ [Pol63].*

Proof

(a) Noting that $m_\varphi = M_{\nabla\varphi} = \frac{1}{\gamma}$, the Inequality (3.32) reduces to $\gamma < \frac{2}{\|\nabla^2 f(\mathbf{x}^*)\|}$. For any symmetric matrix $\mathbf{A}$, $\|\mathbf{I} - \gamma\mathbf{A}\| < 1$ if $\gamma < \frac{2}{\|\mathbf{A}\|}$, and (3.32) implies (3.42).

(b) With $t = 1$ we have that $q \leq \sigma = \|\mathbf{I} - \gamma\nabla^2 f(\mathbf{x}^*)\| = \max\{|1-m|, |1-M|\}$, where m and M are, respectively, the smallest and largest eigenvalue of $\gamma\nabla^2 f(\mathbf{x}^*)$.

(c) Straightforward calculations yield $\mathbf{y}^k = P_X(\mathbf{x}^k - \gamma\nabla f(\mathbf{x}^k))$. Since $t_k \equiv 1$, $\mathbf{x}^{k+1} = \mathbf{y}^k$, and we obtain the gradient projection algorithm. The convergence Criterion (3.32) and Ratio (3.43) can be found in [LeP66].

(d) Follows from **(c)**, with $X = \Re^n$. □

With the choice $\mathbf{B} = \frac{1}{\gamma}\mathbf{I}$ and a step length rule similar to A', we obtain the algorithm of [FuM81]. Note that it reduces to the proximal point method whenever $f \equiv 0$, and the step lengths $t_k \equiv 1$ are used. The method of [FuM81] is shown to be linearly convergent; it also follows from Corollary 3.1. For the proximal point algorithm, it is shown in [Roc76b] that in the continuously differentiable case superlinear convergence is obtained by letting $\{\gamma_k\} \to +\infty$. It is remarked in [FuM81] that superlinear convergence is not possible to obtain in their algorithm, since letting $\{\gamma_k\} \to +\infty$ would destroy the validity of their step length rule. It is interesting to note that, by the analysis made above, superlinear convergence may be obtained also in the nondifferentiable case, by introducing second-order information from the differentiable part of T in φ. This is consistent with the results in the differentiable case where, in Theorem 2.5.a, superlinear convergence is characterized by convergence of search directions toward Newton directions.

3.10 Finite convergence of cost approximation algorithms

We show in this section that, for problems with sharp minima (cf. Definition A.8.d), the class of cost approximation methods exhibits finite convergence.

Lemma 3.7 [Pol87, Le. 5.3.3] (Equivalent definitions of a sharp minimum) [13] *Let f be convex on* `dom` *u. Let Assumption 3.1 hold. The following statements are equivalent.*

(1) $\mathbf{x}^*$ *is a sharp minimum of* T *on* $\Re^n$.

(2) $T'(\mathbf{x}^*; \mathbf{d}) \geq \alpha \|\mathbf{d}\|, \forall \mathbf{d} \in \Re^n$.

(3) $-\nabla f(\mathbf{x}^*) \in \texttt{int}\ \partial u(\mathbf{x}^*)$.

Theorem 3.6 (Finite convergence of cost approximation algorithms) *Let the assumptions of Theorem 3.2 hold, and let both* f *and* $\varphi(\cdot, \mathbf{x})$, $\forall \mathbf{x} \in \texttt{dom}\ u$, *have Lipschitz continuous gradients on* $\texttt{dom}\ u$. *Assume further that* f *is convex, that* $\{\mathbf{x}^k\} \to \mathbf{x}^*$, *and that* $\mathbf{x}^*$ *is a sharp minimum. Then,* $\mathbf{y}^k = \mathbf{x}^*$, *for some finite* k.

Proof We will show that for a sufficiently large, but finite, k, $\mathbf{x}^*$ solves [$\mathbf{NNLP}_\varphi^k$].
From the Lipschitz continuity assumptions, we have that

$$\left\| \left[\nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^k, \mathbf{x}^k\right) - \nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^*, \mathbf{x}^k\right)\right] + \left[\nabla f(\mathbf{x}^*) - \nabla f\left(\mathbf{x}^k\right)\right] \right\| \leq (M_{\nabla\varphi} + M_{\nabla f}) \left\|\mathbf{x}^k - \mathbf{x}^*\right\|.$$

Using this result, and the property **(3)** of Lemma 3.7, we have for a sufficiently large k,

$$\begin{aligned} -\nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^*, \mathbf{x}^k\right) - \nabla f\left(\mathbf{x}^k\right) + \nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^k, \mathbf{x}^k\right) &= \\ \left[\nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^k, \mathbf{x}^k\right) - \nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^*, \mathbf{x}^k\right)\right] + \left[\nabla f(\mathbf{x}^*) - \nabla f\left(\mathbf{x}^k\right)\right] - \nabla f(\mathbf{x}^*) &\in \partial u(\mathbf{x}^*), \end{aligned}$$

i.e., $\mathbf{x}^*$ solves [$\mathbf{NNLP}_\varphi^k$]. □

An alternative proof of the theorem is to show that $T'_\varphi(\mathbf{x}^*; \mathbf{d}) \geq 0, \forall \mathbf{d} \in \Re^n$, for a large enough k. This approach shows also how close $\mathbf{x}^k$ must be to $\mathbf{x}^*$ when termination occurs.

Alternative proof

$$\begin{aligned} T'_\varphi(\mathbf{x}^*; \mathbf{d}) &= u'(\mathbf{x}^*; \mathbf{d}) + \left[\nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^*, \mathbf{x}^k\right) + \nabla f\left(\mathbf{x}^k\right) - \nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^k, \mathbf{x}^k\right)\right]^{\mathrm{T}} \mathbf{d} \\ &= T'(\mathbf{x}^*; \mathbf{d}) + \left[\nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^*, \mathbf{x}^k\right) - \nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^k, \mathbf{x}^k\right)\right]^{\mathrm{T}} \mathbf{d} + \left[\nabla f\left(\mathbf{x}^k\right) - \nabla f(\mathbf{x}^*)\right]^{\mathrm{T}} \mathbf{d} \\ &\geq \left(\alpha - (M_{\nabla\varphi} + M_{\nabla f}) \left\|\mathbf{x}^k - \mathbf{x}^*\right\|\right) \|\mathbf{d}\| \\ &\geq 0, \forall \mathbf{d} \in \Re^n, \end{aligned}$$

whenever[14, 15]

$$\left\|\mathbf{x}^k - \mathbf{x}^*\right\| \leq \frac{\alpha}{M_{\nabla\varphi} + M_{\nabla f}}.$$

This proves the theorem. □

[13]The parameter α is a measure of the degree of sharpness of the minimizer; from **(2)** it may be interpreted as the angle of the kink of T at $\mathbf{x}^*$.

[14]It is apparent from this inequality, that faster convergence is achieved for larger values of α, i.e., the sharper the minimum is (see [Pol87, p. 137, Fig. 17]). When u is the indicator function of a nonempty, closed and convex set X, this result can be rephrased as faster convergence resulting from a more pointed extreme point $\mathbf{x}^*$ of X.

[15]This inequality yields an upper bound on the total number of iterations necessary, provided the distance to $\mathbf{x}^*$ is strictly decreasing. Such a measure is in fact obtainable from the linear convergence Theorem 3.5; it is easy to show that for Rule P with $t_k \equiv 1$, the total number of iterations is not greater than $\lceil \log_q L \rceil + K + 1$, where q is given by (3.42), K is the first iteration in which the relation $\|\mathbf{x}^{k+1} - \mathbf{x}^*\| \leq q\|\mathbf{x}^k - \mathbf{x}^*\|$ holds, and $L = \frac{\alpha}{(M_{\nabla\varphi} + M_{\nabla f})\|\mathbf{x}^K - \mathbf{x}^*\|}$.

3.11 A posteriori error bounds and lower bounds

The following a posteriori error bound follows immediately from [Coh80, Th. 2.2].

Theorem 3.7 (A posteriori error bound) [16] *Let f be strongly convex on* $\mathtt{dom}\ u$*, with a Lipschitz continuous gradient on* $\mathtt{dom}\ u$*. Let $\varphi : \Re^n \mapsto \Re$ be a strongly convex function in C^1 on* $\mathtt{dom}\ u$ *with a Lipschitz continuous gradient on* $\mathtt{dom}\ u$*. Then, for any* $\mathbf{x} \in \mathtt{dom}\ u$*,*

$$\|\mathbf{y}(\mathbf{x}) - \mathbf{x}^*\| \leq \frac{M_{\nabla\varphi} + M_{\nabla f}}{m_f} \|\mathbf{y}(\mathbf{x}) - \mathbf{x}\|. \tag{3.44}$$

The lower bounding results of Theorem 2.7 may be extended to the nondifferentiable case as follows. The proof is straightforward and is therefore omitted.

Theorem 3.8 (Lower bound) *Let $\varphi : \Re^n \mapsto \Re$ be a convex function in C^1 on* $\mathtt{dom}\ u$*. Assume that the function $f - \varphi$ is convex. Then, for any* $\mathbf{x} \in \mathtt{dom}\ u$ *and* $\mathbf{y} \in Y(\mathbf{x})$*,*

$$T_\varphi(\mathbf{y}) \leq T^*. \tag{3.45}$$

Furthermore, equality holds in (3.45) if and only if $\mathbf{x}$ *is a minimizer of* T*.*

3.12 Gauss-Seidel type methods

The Gauss-Seidel method for [**NLP**] was in Section 2.7 shown to be an instance of the cost approximation method. Applying this algorithm to [**NNLP**], we first let $f \equiv 0$. Then,

$$[\mathbf{GS}_i^k] \qquad x_i^{k+1} \in \arg\min_{x_i \in \Re} u\left(\mathbf{x}_{i_-}^{k+1}, x_i, \mathbf{x}_{i_+}^k\right).$$

The convergence properties of Gauss-Seidel is well known from its many applications.[17] When u is nondifferentiable, the Gauss-Seidel method may terminate at a nonoptimal point, even when u is strictly convex with bounded level sets, see e.g. [BaS79, pp. 273–274] and [BeT89, p. 209]. The strongest convergence results known are summarized below.

Theorem 3.9 (Convergence of Gauss-Seidel methods) *Let $u : \Re^n \mapsto \Re$ be continuous.*

(a) *Let u be strictly convex* [18] *with respect to each variable x_i, and let $\mathbf{x}^0$ be such that the level set $L^u_{\Re^n}(\mathbf{x}^0)$ is bounded. Then, for any accumulation point, $\mathbf{x}^\infty$, of the sequence $\{\mathbf{x}^k\}$ (at least one such point exists), the following inequalities hold.*

$$u\left(\mathbf{x}_{i_-}^\infty, x_i^\infty, \mathbf{x}_{i_+}^\infty\right) \leq u\left(\mathbf{x}_{i_-}^\infty, x_i, \mathbf{x}_{i_+}^\infty\right), \quad \forall x_i \in \Re, \ i = 1, \dots, n \tag{3.46}$$

(b) *Let u be strongly convex, and let x_i^{k+1} be given by $x_i^{k+1} = x_i^k + \omega(y_i^k - x_i^k)$, where y_i^k solves $[\mathbf{GS}_i^k]$ and $\omega \in (0,1]$. Then the same conclusion as in* **(a)** *holds.*

---

[16] Note that (3.44) does not reduce to (2.16) if $u \equiv 0$; the result is therefore not a true generalization of Theorem 2.6.a.

[17] A particularly well studied application is that of linear and convex network flows, where [**NNLP**] corresponds to an unconstrained dual formulation, see [Ber81, Ber85a, Ber85b, Tse86, BHT87, BeE88, BeT88a, BeT88b, Ber91a, Ber91b].

Proof

(a) See [FiH75, FiH79].

(b) See [Aus70, Aus71, Aus76].

Thus, even for strongly convex functions u the convergence result guarantees only that in an accumulation point the function can not be further reduced along the coordinate directions. This deficiency is well documented, and various anti-jamming procedures have been devised to proceed from such points, see e.g. [Erl78] and [BaS79, Sec. 8.4].

In order to establish global convergence of the Gauss-Seidel method, we must impose restrictions on u such that (3.46) implies the optimality condition

$$u'(\mathbf{x}^*; \mathbf{d}) \geq 0, \ \forall \mathbf{d} \in \Re^n. \tag{3.47}$$

Assumption 3.2 (Separability of $u'(\mathbf{x};\cdot)$)[19] *For all* $\mathbf{x} \in \mathtt{dom}\ u$ *and* $\mathbf{d} \in \Re^n$,

$$u'(\mathbf{x}; \mathbf{d}) = \sum_{i=1}^{n} u'(\mathbf{x}; d_i \mathbf{e}_i). \tag{3.48}$$

Lemma 3.8 (A convergence criterion) *Let $u : \Re^n \mapsto \Re$ be continuous. Under Assumption 3.2, (3.46) implies (3.47).*

Proof Assume that $\mathbf{x}^*$ satisfies (3.46). For all $\mathbf{d} \in \Re^n$ we then have

$$u'(\mathbf{x}^*; \mathbf{d}) = \sum_{i=1}^{n} u'(\mathbf{x}^*; d_i \mathbf{e}_i) = \sum_{i=1}^{n} \lim_{\lambda \downarrow 0} \frac{1}{\lambda} \left[u\left(\mathbf{x}^*_{i_-}, x^*_i + \lambda d_i, \mathbf{x}^*_{i_+}\right) - u(\mathbf{x}^*) \right] \geq 0.$$

The lemma is proved. □

Thus, with Assumption 3.2 at hand, Theorem 3.9 may be strengthened to ensure global convergence for convex functions u.[20]

To show that the Gauss-Seidel method applied to a nondifferentiable problem is an instance of the cost approximation method, we introduce $\mathbf{\Phi} : \Re^{2n} \mapsto 2^{\Re}$ with

$$\mathbf{\Phi}_i(\mathbf{x}, \mathbf{y}) = \partial_i u\left(\mathbf{x}_{i_-}, x_i, \mathbf{y}_{i_+}\right), \tag{3.49}$$

where $\partial_i u$ denotes the ith component of ∂u. Noting that $\mathbf{\Phi}(\mathbf{x}, \mathbf{x}) = \partial u(\mathbf{x})$, the subproblem of the cost approximation method, given a point $\mathbf{x}^k \in \Re^n$, is $\mathbf{0} \in \mathbf{\Phi}(\mathbf{y}^k, \mathbf{x}^k)$, which, by (3.49), may be written as

$$0 \in \partial_i u\left(\mathbf{y}^k_{i_-}, y^k_i, \mathbf{x}^k_{i_+}\right), \ i = 1, \ldots, n. \tag{3.50}$$

[18] The strict convexity assumption is relaxed in [FiH75, FiH79]. The result of the theorem is then weakened to (essentially) guaranteeing (3.46) to hold for those coordinates i that occur an infinite number of times in the building of the subsequence, converging to $\mathbf{x}^\infty$.

[19] Assumption 3.2 is always satisfied when $u'(\mathbf{x};\cdot)$ is linear; this is a characterization of *Gâteaux-differentiable* convex functions, see e.g. [RoV73, p. 113], [Fle80, Sec. 4.1], [Gil82, p. 118], [Phe89, p. 3]. A more interesting case arises from the introduction of a differentiable function f, and assuming that the objective has the partially separable form $T(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^n u_i(x_i)$. (The fact that T fulfills Assumption 3.2 is immediate.) This case is considered in Gauss-Seidel methods in [CeG73, Cea78], and for the special case of $u_i(x_i) = \alpha_i |x_i|$, $\alpha_i \geq 0$, $\forall i$, also in [Aus70, Aus71, GLT81].

[20] The author is not aware of the existence of results corresponding to case **(a)**; the case **(b)** is covered in [Aus70, Aus71, Aus76].

Now, for any fixed i, and for all $d \in \Re$ and $\boldsymbol{\xi}_u \in \partial u(\mathbf{y}^k_{i_-}, y^k_i, \mathbf{x}^k_{i_+})$ we have that

$$\begin{aligned} u'\left(\left(\mathbf{y}^k_{i_-}, y^k_i, \mathbf{x}^k_{i_+}\right); d\mathbf{e}_i\right) &\geq d\boldsymbol{\xi}_u^{\mathrm{T}}\mathbf{e}_i \quad ([\text{Roc70a, Th. } 23.2]) \\ &= 0, \qquad (3.50) \end{aligned}$$

which, by the convexity of u, is equivalent to y^k_i solving [$\mathbf{GS}^k_i$]. Letting $x^{k+1}_i = y^k_i$ for all $i = 1, \dots, n$, we conclude that the Gauss-Seidel method is a cost approximation method, through the choice (3.49) of the cost approximating map $\mathbf{\Phi}$.

The termination criterion (Step 2) of the algorithm, which for Gauss-Seidel is (3.46), is valid (in the sense that the solution obtained is optimal) under Assumption 3.2.

Extending the block updating cost approximation algorithm of Section 2.7 from the differentiable case, we introduce a partitioning $\mathbf{x}^{\mathrm{T}} = (\mathbf{x}_1^{\mathrm{T}}, \dots, \mathbf{x}^{\mathrm{T}}_{|\mathcal{C}|})$, where $\mathbf{x}_i \in \Re^{n_i}$ and where $\sum_{i\in\mathcal{C}} n_i = n$. We assume that the objective T has the form[21]

$$T(\mathbf{x}) = f(\mathbf{x}) + \sum_{i\in\mathcal{C}} u_i(\mathbf{x}_i), \tag{3.51}$$

where $u_i : \Re^{n_i} \mapsto \Re \cup \{+\infty\}$ is l.s.c., proper and convex and where $f : \Re^n \mapsto \Re \cup \{+\infty\}$ is in C^1 on $\prod_{i\in\mathcal{C}}(\texttt{dom}\ u_i)$. Introduce the function $\varphi : \Re^{2n} \mapsto \Re$, with

$$\varphi(\mathbf{x}, \mathbf{y}) = \sum_{i\in\mathcal{C}} \varphi_i(\mathbf{x}_i, \mathbf{y}), \tag{3.52}$$

where each φ_i is a convex function in C^1 on $\texttt{dom}\ u$ with respect to $\mathbf{x}_i$.

The algorithm is extended from the differentiable case, with [$\mathbf{NLP}^k_{\varphi_i}$] replaced by

$$[\mathbf{NNLP}^k_{\varphi_i}] \qquad \mathbf{y}^k_i \in \arg\min_{\mathbf{x}_i\in\Re^{n_i}} \left\{ u_i(\mathbf{x}_i) + \varphi_i\left(\mathbf{x}_i, \mathbf{x}^k\right) + \left[\nabla_i f\left(\mathbf{x}^k\right) - \nabla\varphi_i\left(\mathbf{x}^k_i, \mathbf{x}^k\right)\right]^{\mathrm{T}} \mathbf{x}_i \right\}.$$

Theorem 3.10 (Convergence of cyclic cost approximation algorithms) *Let the objective T in [**NNLP**] be given by (3.51), and let φ be given by (3.52). Let each function u_i be continuous on $\texttt{dom}\ u_i$. Let Assumption 3.1 hold. Let the indices be chosen in cyclic order, and let $\mathbf{x}^0 \in \prod_{i\in\mathcal{C}}(\texttt{dom}\ u_i)$ be such that $L^T_{\Re^n}(\mathbf{x}^0)$ is bounded. Assume that the problem [$\mathbf{NNLP}_{\varphi_i}$] is well defined, in the sense that its solution set $Y_i(\mathbf{x})$ is nonempty and bounded for all $\mathbf{x} \in \prod_{i\in\mathcal{C}}(\texttt{dom}\ u_i)$. Assume further that the sequence $\{\mathbf{y}^k\}$ is bounded. Then, under Rule M, $\{T(\mathbf{x}^k)\} \to T(\overline{\mathbf{x}})$ for some $\overline{\mathbf{x}} \in \Omega$, any accumulation point $\mathbf{x}^\infty$ of the sequence $\{\mathbf{x}^k\}$ (at least one such point exists) lies in Ω, and (3.19) holds.*

Proof A straightforward extension of the proof of Theorem 2.9. □

Note that for $|\mathcal{C}| = 1$, the algorithm reduces to the basic algorithm of this chapter. The above result is valid for a general, non-additive, u, if an extension of Assumption 3.2 to variable blocks is assumed to hold.

[21] An interesting special case of [**NNLP**] with objective (3.51) is the following. Let each function u_i be an indicator function of a nonempty, closed and convex set $X_i \subseteq \Re^{n_i}$; then [**NNLP**] becomes the problem of minimizing f on $\prod_{i\in\mathcal{C}} X_i$. This problem is given further consideration in Section 4.15.

3.13 Summary and further research

In this chapter we have studied the class of cost approximation methods for nondifferentiable optimization. We summarize below some of the most important contributions and possible directions of further research.

Convex constrained problems with partially differentiable objectives arise when solving nonlinear continuous problems using Benders decomposition techniques [Ben62, FlN74, HoJ84]. Other applications are to be found in dual approaches for large-scale linearly constrained convex programming, where the dual problem may be formulated as a convex and partially differentiable program, and for use in exact penalty methods for differentiable convex programming. It is a nice feature of this class of algorithms that second order methods may be defined for such problems, by applying second-order approximations to the differentiable part of the objective.

To make the results general enough to include constrained nondifferentiable programs, the function u was allowed to have a domain not equal to $\Re^n$. It would be interesting to investigate whether or not the convergence results can be strengthened under the assumption that $\mathtt{dom}\ u = \Re^n$, i.e., for locally Lipschitzian functions. The linear convergence results obtained are restricted to quadratic functions φ; this is in contrast to the linear convergence results obtained for the differentiable problem [**NLP**], which hold under more general conditions. Whether this is due to the nondifferentiability, and the results obtained cannot be strengthened, or it is possible to extend the results to nonquadratic functions by using some other form of analysis, should be investigated.

The potential difficulty of solving the convex nondifferentiable subproblem [$\mathbf{NNLP}_\varphi$] can to some extent be avoided by using the truncation strategy, validated in Theorem 3.4. The trade-off, between the work spent on solving [$\mathbf{NNLP}_\varphi$] and the quality of the search directions obtained, offered by using this strategy, should be further investigated; this includes both the practical efficiency and the theoretical convergence rates that are achieved under different strategies of truncation.

For differentiable, constrained optimization and variational inequalities, finite identification of active constraints by the cost approximation algorithms can be established (see Sections 4.12 and 5.13). Such results are scarce for nondifferentiable, constrained optimization (see [BuM88, Fla92]), but are obviously just as important. It would be interesting to extend these results to the nondifferentiable case.

Finite convergence was established under a sharp minimum condition. An extension to weak sharp minima is left for future research (although it seems straightforward).

Chapter 4

Constrained optimization

"In short, I never yet encountered
the mere mathematician who could be trusted
out of equal roots, or one who did not clandestinely
hold it as a point of his faith that $x^2 + px$ was
absolutely and unconditionally equal to q.
Say to one of these gentlemen, by way of experiment,
if you please, that you believe occasions may occur
where $x^2 + px$ is not *altogether equal to q,*
and, having made him understand what you mean, get
out of his reach as speedily as convenient, for,
beyond doubt, he will endeavour to knock you down."

EDGAR ALLAN POE
The Purloined Letter (1841)

In this chapter we study the cost approximation algorithm for differentiable, constrained optimization. In [**GVIP**], assume that $\mathbf{U} \equiv \mathbf{0}$, and that $\mathbf{F}$ is the gradient of some function $f : X \mapsto \Re$ in C^1 on X, where X is a nonempty, closed and convex set in $\Re^n$. (These properties of X and f are standing assumptions throughout this chapter.)

4.1 Optimality conditions

The above assumptions yield

$$\nabla f(\mathbf{x}^*)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^*) \geq 0, \ \forall \mathbf{x} \in X. \tag{4.1}$$

We let Ω denote the set of solutions to (4.1). Using the normal cone operator, we may equivalently express (4.1) as

$$-\nabla f(\mathbf{x}^*) \in N_X(\mathbf{x}^*). \tag{4.2}$$

These conditions are the first-order *necessary* optimality conditions for

[**CNLP**] $$\min_{\mathbf{x} \in X} f(\mathbf{x}),$$

[BaS79, Th. 4.1.3]; they are *sufficient* for the global optimality of $\mathbf{x}^*$ under a pseudoconvexity assumption on f. Under an appropriate constraint qualification [BGS72, Pet73, BaS76], (4.1) can be expressed in the form of the Karush-Kuhn-Tucker (KKT) conditions. Clearly, the relation between (4.1) and [**CNLP**] motivates the choice $\psi \equiv f$ as merit function.

Remark 4.1 In view of the fact that the set X can be represented by its indicator δ_X, cost approximation methods for [**CNLP**] can be analyzed as a special case of unconstrained nondifferentiable programs [**NNLP**]. There are, however, several reasons for making a separate analysis for [**CNLP**]; an explicit treatment of the set X offers a simplification of the presentation and analysis of the cost approximation methods for [**CNLP**], as well as a simplified comparison with other algorithms for this problem. Furthermore, certain results can only be obtained under a differentiability assumption on f.

4.2 The cost approximation algorithm

The cost approximation algorithm

A sequence $\{\mathbf{\Phi}^k\}$ of monotone cost approximating mappings is assumed to be given.

Step 0 (*Initial guess*) Choose an initial point $\mathbf{x}^0 \in X$, and let $k = 0$.

Step 1 (*Search direction generation*) Find a $\mathbf{y}^k \in X$ such that

$$\left[\mathbf{\Phi}^k\left(\mathbf{y}^k\right) + \nabla f\left(\mathbf{x}^k\right) - \mathbf{\Phi}^k\left(\mathbf{x}^k\right)\right]^{\mathrm{T}}\left(\mathbf{y} - \mathbf{y}^k\right) \geq 0, \ \forall \mathbf{y} \in X. \tag{4.3}$$

The resulting search direction is $\mathbf{d}^k = \mathbf{y}^k - \mathbf{x}^k$.

Step 2 (*Convergence check*) If $\mathbf{x}^k$ solves (4.3) $\to$ Stop ($\mathbf{x}^k \in \Omega$). Otherwise, continue.

Step 3 (*Line search*) Find a step length, t_k, which solves the one-dimensional problem

$$t_k \in \arg \min_{\substack{t \geq 0 \\ \mathbf{x}^k + t\mathbf{d}^k \in X}} f\left(\mathbf{x}^k + t\mathbf{d}^k\right).$$

Step 4 (*Update*) Let $\mathbf{x}^{k+1} = \mathbf{x}^k + t_k\mathbf{d}^k$, and $k := k+1$.

Step 5 (*Convergence check*) If $\mathbf{x}^k$ is acceptable as a solution[1] $\to$ Stop. Otherwise, go to Step 1.

In most known algorithms for [**CNLP**] that can be identified as special cases of cost approximation methods, the search direction finding subproblem (4.3) is a convex minimization problem. Introducing a sequence $\{\varphi^k\}$ of convex functions $\varphi^k : \Re^n \mapsto \Re$ in C^1, with $\nabla\varphi^k \equiv \mathbf{\Phi}^k$, the subproblem (4.3) reduces to

$$[\mathbf{CNLP}^k_{\varphi^k}] \qquad \mathbf{y}^k \in \arg\min_{\mathbf{x} \in X} f_{\varphi^k}(\mathbf{x}),$$

where

$$f_{\varphi^k}(\mathbf{x}) = \varphi^k(\mathbf{x}) + f\left(\mathbf{x}^k\right) - \varphi^k\left(\mathbf{x}^k\right) + \left[\nabla f\left(\mathbf{x}^k\right) - \nabla\varphi^k\left(\mathbf{x}^k\right)\right]^{\mathrm{T}}\left(\mathbf{x} - \mathbf{x}^k\right).$$

[1] Typical stopping criteria are a small step length, a small decrease in the objective value between two iterations, or a small relative error in the objective.

The subproblem [$\mathbf{CNLP}^k_{\varphi^k}$] may be interpreted as follows. At $\mathbf{x}^k \in X$, f is approximated by φ^k. The error thus introduced, $f - \varphi^k$, is then linearized at $\mathbf{x}^k$.

With the special case of cost approximating mapping $\mathbf{\Phi} : X \times X \mapsto \Re^n$ of the form $\mathbf{\Phi}(\mathbf{x}, \mathbf{y})$, continuous on $X \times X$ and monotone with respect to $\mathbf{x}$, the subproblem (4.3) reduces to the problem of finding a point $\mathbf{y}(\mathbf{x}) \in X$ such that

$$[\mathbf{\Phi}(\mathbf{y}(\mathbf{x}), \mathbf{x}) + \nabla f(\mathbf{x}) - \mathbf{\Phi}(\mathbf{x}, \mathbf{x})]^{\mathrm{T}} (\mathbf{y} - \mathbf{y}(\mathbf{x})) \geq 0, \ \forall \mathbf{y} \in X. \tag{4.4}$$

This subproblem will later be identified as equivalent to those of the algorithm frameworks of Dafermos and Tseng (see Theorem 4.1.a).

Before analyzing the basic algorithm we present below some well known instances, and provide some interesting algorithmic equivalence results.

4.3 Instances of cost approximation algorithms

Choosing $\varphi^k \equiv 0$ results in the subproblem of minimizing $\nabla f(\mathbf{x}^k)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^k)$ over $\mathbf{x} \in X$, i.e., the subproblem in the *Frank-Wolfe* (or *conditional gradient*) method [FrW56, DeR65, LeP66, DeR70].

Choosing $\varphi^k(\mathbf{x}) = \frac{1}{2\gamma_k}\mathbf{x}^{\mathrm{T}}\mathbf{B}_k\mathbf{x}$, where $\gamma_k > 0$ and $\mathbf{B}_k$ is a symmetric and positive semidefinite matrix on X, yields a subproblem where

$$\nabla f\left(\mathbf{x}^k\right)^{\mathrm{T}} \left(\mathbf{x} - \mathbf{x}^k\right) + \frac{1}{2\gamma_k} \left(\mathbf{x} - \mathbf{x}^k\right)^{\mathrm{T}} \mathbf{B}_k \left(\mathbf{x} - \mathbf{x}^k\right)$$

is minimized over $\mathbf{x} \in X$. If the matrix $\mathbf{B}_k$ is positive definite on X, $\mathbf{y}^k$ is given by

$$\mathbf{y}^k = P_X^{\mathbf{B}_k} \left(\mathbf{x}^k - \gamma_k \mathbf{B}_k^{-1} \nabla f\left(\mathbf{x}^k\right)\right). \tag{4.5}$$

This is the subproblem of the class of (*deflected*) *gradient projection* methods, including, as special cases, the *gradient projection method* ($\mathbf{B}_k \equiv \mathbf{I}$) [Gol64, LeP66] and *Newton's method* ($\mathbf{B}_k = \nabla^2 f(\mathbf{x}^k), \gamma_k \equiv 1$) [LeP66, Dan69, Dan70, PsD78, Dun80b].[2]

A variety of *regularization methods* are included in the framework. Let $\varphi^k(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{\gamma_k} D(\mathbf{x}, \mathbf{x}^k)$, where $\gamma_k > 0$ and where the function $D(\mathbf{x}, \mathbf{y})$ is continuous on $X \times X$ and convex and in C^1 on X with respect to $\mathbf{x}$. Furthermore, if we assume that $\nabla_{\mathbf{x}} D(\mathbf{x}, \mathbf{x}) \stackrel{\text{def}}{=} \nabla_{\mathbf{x}} D(\mathbf{x}, \mathbf{y})\,|_{\mathbf{y}=\mathbf{x}} = \mathbf{0}$ holds, then

$$f_{\varphi^k}(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{\gamma_k} D(\mathbf{x}, \mathbf{x}^k), \tag{4.6}$$

which is a generalization of the subproblem objective in regularization methods [Tik63, LeP66, TiA77, Pol87].

In particular, if we let D have the form

$$D_r(\mathbf{x}, \mathbf{y}) = r(\mathbf{x}) - r(\mathbf{y}) - \nabla r(\mathbf{y})^{\mathrm{T}}(\mathbf{x} - \mathbf{y}),$$

[2] Modified versions of these methods, where the step length γ_k is determined through a line search [McT72, Gol74, Ber76b, Dun81, GaB82, Phe86, CaM87, Dun87, BuM88], or where the matrix $\mathbf{B}_k$, with respect to which norm the projection is made, is replaced by some other symmetric and positive definite matrix [Ber82a, Ber82c, GaB84, GaD88], are, strictly speaking, not included in this framework.

where $r : X \mapsto \Re$ is a *Bregman function*, then D_r is the D-function of Bregman [Bre66, Bre67], also studied in [CeL81, DeI86, CeZ92, Teb92, Eck93, ChT93]. (Usually, $\gamma_k \equiv 1$ is used.)

An extensively studied special case is the result of choosing $r(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, which yields

$$f_{\varphi^k}(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{2\gamma_k}\left\|\mathbf{x} - \mathbf{x}^k\right\|^2,$$

i.e., the subproblem objective of the *proximal point* method [Min62, Mor65, Mar70, Mar72, PoT74, Roc73, Roc76a, Roc76b, Ber82a, Luq84a, BeT89].[3]

Closely related to the proximal point algorithm is the class of *splitting methods* [Bre73, LiM79, Tse91a, EcB92] for finding a zero of the sum of two (maximal) monotone operators. (For an objective of the form $f = g + h$, in a splitting algorithm the function φ^k is given by $\varphi^k(\mathbf{x}) = g(\mathbf{x}) + \frac{1}{2\gamma_k}\|\mathbf{x} - \mathbf{x}^k\|^2$.) It has been shown [LaS87, Eck88, Eck89, EcB92] that a number of algorithms in this class, such as Lions-Mercier splitting [LiM79, Pas79] and Douglas-Rachford splitting [DoR56] are instances of the proximal point algorithm. It is also known [Gab83, Tse91a] that the method of alternating directions [GlM75, GaM76] is a special case of Lions-Mercier splitting, and that the same holds for the method of partial inverses [Spr83, Spr85] and the method of Han and Lou [HaL88]. It has been observed that *augmented Lagrangean* methods (e.g. the method of multipliers [Hes69, Pow69, HaB70, Ber82a]) define special cases of proximal point methods for dual programs, see for instance [Roc73, Roc76a, Ber82a], [Lue84, Sec. 13.4] and [BeT89, Sec. 3.4]. All the above mentioned methods are *a fortiori* cost approximation methods.[4, 5]

In the special case where f is quadratic, with $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{q}^{\mathrm{T}}\mathbf{x}$, for some symmetric matrix $\mathbf{Q} \in \Re^{n\times n}$, positive semidefinite on X, and $\mathbf{q} \in \Re^n$, and X polyhedral, [**CNLP**] is a convex quadratic program. The above splitting methods then reduce to the *matrix splitting* methods [Hil57, OrR70, LiP87, Man90, LuT91, Pan93b] for linear complementarity problems and quadratic programs. A *splitting* $(\mathbf{A}, \mathbf{B})$ of $\mathbf{Q}$ is given by $\mathbf{Q} = \mathbf{A} + \mathbf{B}$; it is termed *regular* if $\mathbf{A} - \mathbf{B}$ is positive definite. Assuming, for the sake of this example, that $\mathbf{A}$ is symmetric and letting $\varphi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, we obtain

$$\mathbf{y}^k = \arg\min_{\mathbf{x}\in X}\left\{\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \left[\mathbf{B}\mathbf{x}^k + \mathbf{q}\right]^{\mathrm{T}}\mathbf{x}\right\},$$

[3] The main reason for considering regularization methods is the wish to strictly convexify a nonconvex objective; regularization methods are usually studied under the additional assumption that D is strongly convex. More general D-functions for algorithms devised to find a point in the intersection of convex sets are studied in [Bre66, Bre67]. Several other variations and extensions of the proximal point method have been presented, see, e.g., [KoB72, GoT74, Ber75b, GoT75, Kor75a, Kor75b, Ber76a, KoB76, GoT79, Ber82a, Luq84b, Luq86a, Luq86b, Ha90, Gul92].

[4] By using the algorithmic equivalence Theorem 4.1.a below, [Tse91b, Exs. 6.1 and 6.7] illustrate the fact that augmented Lagrangean methods are cost approximation methods.

[5] Apparently, several specialized algorithms for Lagrangean dual formulations of convex programs can be identified as cost approximation methods. Apart from those mentioned, classical multiplier methods such as Uzawa's method [Uza58] and the Arrow-Hurwicz algorithm [ArH60], and the price coordination methods of Takahara [Tak64, CoJ76] and Bernhard [Coh72, Ber75a] are included. Cohen [CoJ76, Coh78, Coh80, CoB80, CoZ84, CoM90] discusses at length dual decomposition/coordination methods (see e.g. [ArH60, Tak64, BLP65, LaS65, Las70, MMT70]) in the context of the auxiliary problem principle, and gives several examples of how it can be applied in two-level algorithms and price coordination schemes; from Theorem 4.1.c, this discussion applies to the class of cost approximation methods as well. However, a further analysis of applications to dual programs is beyond the scope of this thesis.

which is equivalent to the subproblem of the matrix splitting method. Thus, matrix splitting methods are special cases of cost approximation methods.

A good example of utilization of both objective and constraint structure in a cost approximation algorithm is the method of Evans [Eva76], for the combined distribution and assignment problem. Here, the objective is to minimize

$$f(\mathbf{x}, \mathbf{d}) = \sum_{a \in \mathcal{A}} \int_0^{x_a} t_a(s) ds + \frac{1}{\gamma} \sum_{(p,q) \in \mathcal{C}} d_{pq} \left(\log d_{pq} - 1 \right),$$

where $\mathbf{x} = (x_a)_{a \in \mathcal{A}}$ is the vector of arc flows in a transportation network, $\mathbf{d} = (d_{pq})_{(p,q) \in \mathcal{C}}$ is the vector of trip demands between zones in the network, and γ is a positive parameter. In Evans' algorithm, the objective is linearized with respect to the $\mathbf{x}$ variables, resulting in a subproblem that may be solved in two stages. First, a shortest route pattern is calculated with $\mathbf{t}(\mathbf{x}^k)$ as arc costs; then, a set of auxiliary demands $\mathbf{e}^k$ is calculated through the solution of a doubly constrained gravity model, using the shortest route costs previously calculated. The assignment of the demands obtained onto the minimum cost routes yields an auxiliary flow $\mathbf{y}^k$. A line search is finally made between $(\mathbf{x}^k, \mathbf{d}^k)$ and $(\mathbf{y}^k, \mathbf{e}^k)$ to obtain a new solution. That this algorithm is an instance of the cost approximation algorithm is easily verified by choosing $\varphi^k(\mathbf{x}, \mathbf{d}) = \frac{1}{\gamma} \sum_{(p,q) \in \mathcal{C}} d_{pq} \left(\log d_{pq} - 1 \right)$, $\forall k$.

In Section 4.15, it will be shown how the class of cost approximation methods can effectively utilize separability in the constraints to induce decomposition methods of the Gauss-Seidel and Jacobi type.

4.4 Algorithmic equivalence results

For differentiable unconstrained programming [**NLP**] we established a strong relationship between cost approximation methods and the classes of gradient related methods (see Lemma 2.1.e) and deflected gradient methods (see Theorem 2.1). For the constrained case, we will establish results relating cost approximation methods to the algorithm classes of Cohen [Coh78, Coh80], Dussault *et al.* [DFL86], Larsson and Migdalas [Mig88, LaM90], Dafermos [Daf83], Migdalas [Mig90], and Tseng [Tse91b]. We begin by stating the basic subproblems of these methods, as applied to [**CNLP**].

The algorithm class of Cohen

Let $\varphi^k : X \mapsto \Re$ be a strongly convex function in C^1 on X, and let $\varepsilon_k > 0$. Given $\mathbf{x}^k \in X$, the following subproblem determines $\mathbf{x}^{k+1}$ in the *auxiliary problem principle.*

$$\min_{\mathbf{y} \in X} \left\{ \varphi^k(\mathbf{y}) + \left[\varepsilon_k \nabla f\left(\mathbf{x}^k\right) - \nabla \varphi^k\left(\mathbf{x}^k\right) \right]^{\mathrm{T}} \mathbf{y} \right\} \tag{4.7}$$

The algorithm class of Dussault et al., and Larsson and Migdalas

Let $\varphi : X \mapsto \Re$ be a strictly convex function in C^1 on X. Given $\mathbf{x}^k \in X$, the following subproblem is solved.

$$\min_{\mathbf{y} \in X} \left\{ \varphi(\mathbf{y}) + \left[\nabla f\left(\mathbf{x}^k\right) - \nabla \varphi\left(\mathbf{x}^k\right) \right]^{\mathrm{T}} \mathbf{y} \right\} \tag{4.8}$$

The algorithm class of Dafermos

Let $\mathbf{g} : X \times X \mapsto \Re^n$ be a continuous mapping satisfying

(1) $\mathbf{g}(\mathbf{x}, \mathbf{x}) = \nabla f(\mathbf{x}), \ \forall \mathbf{x} \in X$,

(2) $\nabla_{\mathbf{x}} \mathbf{g}(\mathbf{x}, \mathbf{y})$ is symmetric and positive definite for any fixed $\mathbf{x}, \mathbf{y} \in X$.

Given $\mathbf{x}^k \in X$, the point $\mathbf{x}^{k+1}$ is obtained by solving the variational inequality subproblem

$$\mathbf{g}\left(\mathbf{x}^{k+1}, \mathbf{x}^k\right)^{\mathrm{T}} \left(\mathbf{x} - \mathbf{x}^{k+1}\right) \geq 0, \ \forall \mathbf{x} \in X. \tag{4.9}$$

The algorithm class of Migdalas

Let $h : X \times X \mapsto \Re$ be a continuous function of the form $h(\mathbf{x}, \mathbf{y})$, strictly convex and in C^1 on X with respect to $\mathbf{x}$. Assume further that

$$\nabla_{\mathbf{x}} h(\mathbf{x}, \mathbf{x}) \stackrel{\text{def}}{=} \nabla_{\mathbf{x}} h(\mathbf{x}, \mathbf{y}) \left.\right|_{\mathbf{y}=\mathbf{x}} = \mathbf{0}. \tag{4.10}$$

Given $\mathbf{x}^k \in X$, the subproblem in the *regularized Frank-Wolfe algorithm* is

$$\min_{\mathbf{y} \in X} \left\{ \nabla f\left(\mathbf{x}^k\right)^{\mathrm{T}} \mathbf{y} + h\left(\mathbf{y}, \mathbf{x}^k\right) \right\}. \tag{4.11}$$

The algorithm class of Tseng

Let $\mathbf{H} : X \times X \mapsto \Re$ be strictly *proximal related*, i.e.,

(1) $\mathbf{H}(\mathbf{x}, \mathbf{x}) = \mathbf{0}, \ \forall \mathbf{x} \in X$,

(2) $\mathbf{H}(\mathbf{x}, \mathbf{y})^{\mathrm{T}}(\mathbf{x} - \mathbf{y}) > 0, \ \forall \mathbf{x}, \mathbf{y} \in X, \ \mathbf{x} \neq \mathbf{y}$.

Given $\mathbf{x}^k \in X$, the subproblem in the *nonlinear proximal descent method* is the following variational inequality.

$$\left[\mathbf{H}\left(\mathbf{y}^k, \mathbf{x}^k\right) + \nabla f\left(\mathbf{x}^k\right)\right]^{\mathrm{T}} \left(\mathbf{y} - \mathbf{y}^k\right) \geq 0, \ \forall \mathbf{y} \in X \tag{4.12}$$

In the following theorem we relate the subproblems of the above algorithm classes to each other, and to that of the class of cost approximation methods. When stating that two subproblems are equivalent, we mean that any instance of subproblems of the first type can be reformulated as a subproblem of the second type, and vice versa.

Theorem 4.1 (Algorithmic equivalence results)

(a) *If $\{\mathbf{\Phi}^k\}$ is of the form $\mathbf{\Phi} : X \times X \mapsto \Re^n$, then the resulting subproblem (4.4) is equivalent to the subproblems (4.9) of [Daf83], and (4.12) of [Tse91b].*

(b) *If $\{\varphi^k\}$ is of the form $\varphi : X \times X \mapsto \Re$, then the resulting subproblem*

$$\min_{\mathbf{y} \in X} \left\{ \varphi(\mathbf{y}, \mathbf{x}) + f(\mathbf{x}) - \varphi(\mathbf{x}, \mathbf{x}) + \left[\nabla f(\mathbf{x}) - \nabla_{\mathbf{x}} \varphi(\mathbf{x}, \mathbf{x})\right]^{\mathrm{T}} (\mathbf{y} - \mathbf{x}) \right\} \tag{4.13}$$

is equivalent to the subproblem (4.11) of [Mig90].

(c) *If $\{\varphi^k\}$ is of the form $\varphi^k : X \mapsto \Re$, then the resulting subproblem* $[\mathbf{CNLP}^k_{\varphi^k}]$ *is equivalent to the subproblem (4.7) of [Coh78, Coh80].*

(d) *If* $\{\varphi^k\}$ *is of the form* $\varphi : X \mapsto \Re$, *then the resulting subproblem* [**CNLP**$_\varphi^k$] *is equivalent to the subproblem (4.8) of [DFL86, Mig88, LaM90].*

Proof

(a) The proof is based on a sequence of implications, forming a cycle.

[(4.12) $\Rightarrow$ (4.9)] For a given mapping $\mathbf{g}$ in the definition of (4.9), choose $\mathbf{H}(\mathbf{x}, \mathbf{y}) = \mathbf{g}(\mathbf{x}, \mathbf{y}) - \nabla f(\mathbf{y})$. From the property **(1)** of $\mathbf{g}$, $\mathbf{H}(\mathbf{x}, \mathbf{x}) = \mathbf{0}$, $\forall \mathbf{x} \in X$, so that a subproblem of the type (4.12) is defined.

[(4.9) $\Rightarrow$ (4.4)] For a given mapping $\mathbf{\Phi}$ in the definition of (4.4), choose $\mathbf{g}(\mathbf{x}, \mathbf{y}) = \mathbf{\Phi}(\mathbf{x}, \mathbf{y}) + \nabla f(\mathbf{y}) - \mathbf{\Phi}(\mathbf{y}, \mathbf{y})$. Then $\mathbf{g}(\mathbf{x}, \mathbf{x}) = \nabla f(\mathbf{x})$, $\forall \mathbf{x} \in X$, so that we recover a subproblem of the type (4.9).

[(4.4) $\Rightarrow$ (4.12)] If the mapping $\mathbf{\Phi}$ in the definition of (4.4) is chosen such that $\mathbf{\Phi}(\mathbf{x}, \mathbf{x}) = \mathbf{0}$, $\forall \mathbf{x} \in X$, the problem (4.4) reduces to (4.12), with $\mathbf{H} \equiv \mathbf{\Phi}$.

(b) Follows immediately from the variational inequality characterization of (4.11), which is equivalent to (4.12), with $\mathbf{H} \equiv \nabla_{\mathbf{x}} h$.

(c) Follows from Theorem 3.1, by letting $u \equiv 0$.

(d) Follows from **(c)**, by letting $\varphi^k \equiv \varphi$. □

The results in **(a)** and **(b)** first appeared in [Pat91b]; the result in **(c)** first appeared in [LaP90], there applied to [**VIP**].

The reader should note that this algorithmic equivalence result does not state that the entire schemes are equivalent; only the equivalence of the direction finding subproblems is shown. Furthermore, some of these algorithm classes are originally stated for more general problems than [**CNLP**]. Cohen [Coh78, Coh80] considers (1.3); he extends his scheme to variational inequalities in [Coh88]. Also Dafermos [Daf83] considers variational inequalities in her analysis. The above theorem remains true for this more general case, by simply replacing ∇f by a more general mapping $\mathbf{F}$. For further discussions on variational inequalities, we refer to Section 5.3.

The above theorem shows the strong relationships among independently developed algorithm classes. We will frequently utilize results obtained for the above mentioned classes; some of these will be considerably strengthened.

The relations among the above described method classes are shown in Figure 4.1. From this figure, it is clear that the most general cost approximation algorithm, that is, the one given by a subproblem of the type (4.3), defines a generalization of *all* the method classes discussed in this section.

4.5 Analysis of the direction finding subproblem

In this section we extend the results obtained in Sections 2.3 and 3.5 to [**CNLP**].

Lemma 4.1 (Descent lemma) *Let* $\varphi : \Re^n \mapsto \Re$ *be a convex function in* C^1 *on* X. *Let* $\mathbf{x} \in X$, $Y(\mathbf{x})$ *be the (possibly empty) set of optimal solutions to* [**CNLP**$_\varphi$], $\mathbf{y} \in Y(\mathbf{x})$, *and* $\mathbf{d} = \mathbf{y} - \mathbf{x}$.

$$
\begin{tikzcd}
{[\mathbf{CNLP}_{\varphi^k}^k]} \arrow[d, Leftrightarrow] & (4.3) \arrow[l, Rightarrow] \arrow[d, Rightarrow] & \\
\begin{bmatrix}\text{Coh78}\\ \text{Coh80}\end{bmatrix} \arrow[d, Rightarrow] & (4.4) \arrow[r, Leftrightarrow] \arrow[d, Rightarrow] & \begin{bmatrix}\text{Daf83}\\ \text{Tse91b}\end{bmatrix} \\
{[\text{Mig90}]} \arrow[r, Leftrightarrow] \arrow[d, Rightarrow] & (4.13) & \\
\begin{bmatrix}\text{DFL86}\\ \text{Mig88}\\ \text{LaM90}\end{bmatrix} & &
\end{tikzcd}
$$

Figure 4.1: Relations among method classes

(a) $\mathbf{x} \in \Omega \Leftrightarrow \mathbf{x} \in Y(\mathbf{x}) \Leftrightarrow f_\varphi(\mathbf{y}) = f_\varphi(\mathbf{x})$.

(b) *Let* $\overline{\mathbf{x}} \in X$ *be any point such that* $f_\varphi(\overline{\mathbf{x}}) < f_\varphi(\mathbf{x})$. *Then* $\nabla f(\mathbf{x})^{\mathrm{T}}(\overline{\mathbf{x}} - \mathbf{x}) < 0$. *Especially, for* $\overline{\mathbf{x}} = \mathbf{y}$, *if* $\mathbf{x} \notin \Omega$, *then* $\nabla f(\mathbf{x})^{\mathrm{T}}(\mathbf{y} - \mathbf{x}) < 0$.

(c) *Let* φ *be strictly convex on* X. *If* $Y(\mathbf{x})$ *is nonempty, then it is singleton.*

(d) *Let* φ *be strongly convex on* X. *Then* $Y(\mathbf{x})$ *is nonempty and singleton, and*

$$\nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d} \le -m_\varphi \|\mathbf{d}\|^2 . \tag{4.14}$$

(e) *Let* φ *be strongly convex on* X, *and let* $\nabla\varphi$ *be Lipschitz continuous on* X. *If* $\mathbf{x} \notin \Omega$, *then*

$$-\frac{\nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d}}{\|\nabla f(\mathbf{x}) + \mathbf{v}\| \, \|\mathbf{d}\|} \ge \rho, \tag{4.15}$$

where $\rho = m_\varphi / M_{\nabla\varphi}$, *and* $\mathbf{v} \in N_X(\mathbf{y})$.

Proof

(a) $\mathbf{x} \in \Omega \Leftrightarrow -\nabla f(\mathbf{x}) \in N_X(\mathbf{x}) \Leftrightarrow -\nabla\varphi(\mathbf{x}) - \nabla f(\mathbf{x}) + \nabla\varphi(\mathbf{x}) \in N_X(\mathbf{x}) \Leftrightarrow \mathbf{x} \in Y(\mathbf{x}) \Leftrightarrow f_\varphi(\mathbf{y}) = f_\varphi(\mathbf{x})$.

(b) By the convexity of φ, $f_\varphi(\overline{\mathbf{x}}) < f_\varphi(\mathbf{x})$ implies $\nabla f(\mathbf{x})^{\mathrm{T}}(\overline{\mathbf{x}} - \mathbf{x}) < 0$. The point $\mathbf{y}$ is a special case of such an $\overline{\mathbf{x}}$.

(c) Follows from the strict convexity of f_φ, resulting in a unique solution to $[\mathbf{CNLP}_\varphi]$.

(d) The first result is immediate from the strong convexity of f_φ. We also have

$$
\begin{aligned}
\nabla f(\mathbf{x})^{\mathrm{T}}(\mathbf{y} - \mathbf{x}) &\le [\nabla\varphi(\mathbf{x}) - \nabla\varphi(\mathbf{y})]^{\mathrm{T}}(\mathbf{y} - \mathbf{x}) && (4.3)\\
&\le -m_\varphi \|\mathbf{y} - \mathbf{x}\|^2 . && (\mathrm{A}.11)
\end{aligned}
$$

(e) The variational inequality (4.3) can be equivalently written as the set inclusion

$$-\nabla\varphi(\mathbf{y}) - \nabla f(\mathbf{x}) + \nabla\varphi(\mathbf{x}) \in N_X(\mathbf{y}). \tag{4.16}$$

Then there must be a $\mathbf{v} \in N_X(\mathbf{y})$ such that

$$\nabla\varphi(\mathbf{y}) + \nabla f(\mathbf{x}) - \nabla\varphi(\mathbf{x}) + \mathbf{v} = \mathbf{0}. \tag{4.17}$$

The Lipschitz continuity of $\nabla\varphi$ yields

$$\|\nabla f(\mathbf{x}) + \mathbf{v}\| \leq M_{\nabla\varphi} \|\mathbf{y} - \mathbf{x}\|.$$

Together with (4.14) this inequality gives that (4.15) holds. □

The continuity properties of the direction finding map $\mathbf{x} \mapsto D(\mathbf{x}) \stackrel{\text{def}}{=} Y(\mathbf{x}) - \mathbf{x}$ are analyzed below.

Lemma 4.2 (Continuity properties of the search direction) *Let $\varphi : X \times X \mapsto \Re$ be a continuous function on $X \times X$ of the form $\varphi(\mathbf{x}, \mathbf{y})$, convex and in C^1 on X with respect to $\mathbf{x}$.*

(a) *The mapping $\mathbf{x} \mapsto D(\mathbf{x})$ is closed on X.*

(b) *The set $Y(\mathbf{x})$ is nonempty and bounded* [6] *if and only if*

$$-\nabla f(\mathbf{x}) + \nabla_{\mathbf{x}}\varphi(\mathbf{x}, \mathbf{x}) \in \texttt{int}\ (\texttt{dom}\ (\delta_X + \varphi(\cdot, \mathbf{x}))^\circ). \tag{4.18}$$

(c) *If, in addition to the assumption that (4.18) holds for every $\mathbf{x} \in X$, $\varphi(\cdot, \mathbf{x})$ is strictly convex on X for every $\mathbf{x} \in X$, then the mapping $\mathbf{x} \mapsto D(\mathbf{x})$ is continuous on X.*

Proof

(a) Let $\{\mathbf{x}^k\}$ and $\{\mathbf{y}^k\}$ be sequences in X with $\{\mathbf{x}^k\} \to \mathbf{x}$, $\{\mathbf{y}^k\} \to \mathbf{y}$ and $\mathbf{y}^k \in Y(\mathbf{x}^k), \forall k$. From (4.16) it follows that, for all k, $-\nabla\varphi(\mathbf{y}^k, \mathbf{x}^k) - \nabla f(\mathbf{x}^k) + \nabla\varphi(\mathbf{x}^k, \mathbf{x}^k)$ is an element of $N_X(\mathbf{y}^k)$. Using [Roc70a, Th. 24.4] and the continuity of ∇f and $\nabla_{\mathbf{x}}\varphi$ on X and $X \times X$, respectively, we may conclude that

$$-\nabla\varphi(\mathbf{y}, \mathbf{x}) - \nabla f(\mathbf{x}) + \nabla\varphi(\mathbf{x}, \mathbf{x}) \in N_X(\mathbf{y}), \tag{4.19}$$

i.e., $\mathbf{y} \in Y(\mathbf{x})$. Hence, the map $\mathbf{x} \mapsto Y(\mathbf{x})$ is closed on X.

The proof of the closedness of $\mathbf{x} \mapsto D(\mathbf{x})$ then follows from precisely the same arguments as in Lemma 2.2.a.

(b) Follows from the same line of argument as in Lemma 2.2.b, using the characterization (4.19) of $Y(\mathbf{x})$.

(c) Follows from **(a)**, and the fact that the mapping is point-to-point, since $Y(\mathbf{x})$ is singleton (cf. Lemma 4.1.c). □

[6] From **(a)** the mapping $\mathbf{x} \mapsto D(\mathbf{x})$ is u.s.c. on any compact subset of X; see [Nik68, Le. 4.4].

4.6 Basic convergence of cost approximation algorithms

In this section convergence is established for continuous functions of the form $\varphi(\mathbf{x}, \mathbf{y})$, under exact solutions of both the subproblem [$\mathbf{CNLP}_\varphi$] and the line search problem. All these conditions will subsequently be relaxed.

The following theorem extends Theorem 2.2, and specializes Theorem 3.2 to constrained, differentiable programs.

Theorem 4.2 (Basic convergence of cost approximation algorithms) *Let $\varphi : X \times X \mapsto \Re$ be a continuous function on $X \times X$ of the form $\varphi(\mathbf{x}, \mathbf{y})$, convex and in C^1 on X with respect to $\mathbf{x}$. Assume that the point $\mathbf{x}^0$ is chosen so that the level set $L_X^f(\mathbf{x}^0)$ is bounded, and further that [$\mathbf{CNLP}_\varphi$] is well defined, in the sense that $Y(\mathbf{x})$ is nonempty and bounded for every $\mathbf{x} \in X$. Then, under Rule M, $\{f(\mathbf{x}^k)\} \to f(\overline{\mathbf{x}})$ for some $\overline{\mathbf{x}} \in \Omega$, any accumulation point $\mathbf{x}^\infty$ of the sequence $\{\mathbf{x}^k\}$ (at least one such point exists) lies in Ω, and*

$$\left\{ \inf_{\mathbf{x} \in \Omega \cap L_X^f(\mathbf{x}^0)} \left\| \mathbf{x}^k - \mathbf{x} \right\| \right\} \to 0. \tag{4.20}$$

Proof The proof of the theorem is a slight extension of the proof of Theorem 2.2; we only describe the major changes.

(1) Because of the level set compactness assumption and the descent property of the algorithm, the sequence $\{\mathbf{x}^k\}$ lies in a compact set.

(2) The descent property follows from Lemma 4.1.b. The termination property ($\mathbf{x}^k \in \Omega$) follows from Lemma 4.1.a.

(3) The algorithmic map is closed; the proof is similar to that of Theorem 2.2, and follows from the closedness property of D, established in Lemma 4.2.a.

The last statement follows from Lemma 2.3, as in Theorem 2.2. □

A direct proof, which does not utilize Zangwill's theorem, is given in [Pat91a].

4.7 Convergence under alternative step length rules

In the next convergence result, analogous to Theorem 2.3 for [**NLP**] and Theorem 3.3 for [**NNLP**], we introduce inexact line searches through the step length Rules A and P. For simplicity, Rule A is used with a unit initial step (i.e., $\varepsilon = 1$).

Lemma 4.3 (Validity of Rule A) *Let ∇f be Lipschitz continuous on X. Let $\mathbf{x} \in X$, $\mathbf{y} \in Y(\mathbf{x})$, and $\mathbf{d} = \mathbf{y} - \mathbf{x}$. If $\mathbf{x} \notin \Omega$, there exists a finite integer $\bar{\imath}$ satisfying (A.37).*

Proof Taylor's formula and the Lipschitz continuity of ∇f yield, for $t \in [0, t^{\max}]$,

$$f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x}) \leq t \nabla f(\mathbf{x})^{\mathrm{T}} \mathbf{d} + \frac{M_{\nabla f}}{2} t^2 \left\| \mathbf{d} \right\|^2 \tag{4.21}$$

[see also (3.22)]. From (4.21) and Lemma 4.1.b we obtain, by choosing

$$t \leq 2(\alpha - 1)\frac{\nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d}}{M_{\nabla f}\left\|\mathbf{d}\right\|^2},$$

that

$$f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x}) \leq \alpha t \nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d}. \tag{4.22}$$

Replacing t with $\beta^{\bar{\imath}}$, we see that $\bar{\imath}$ is the smallest integer i to satisfy

$$\beta^i \leq 2(\alpha - 1)\frac{\nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d}}{M_{\nabla f}\left\|\mathbf{d}\right\|^2}.$$

Hence, the Inequality (A.37) will be satisfied after a finite number of steps. □

Remark 4.2 Note that, denoting by $\tilde{t}$ the step given by Rule A, we have

$$\frac{1}{\beta}\tilde{t} = \beta^{(\bar{\imath}-1)} > 2(\alpha - 1)\frac{\nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d}}{M_{\nabla f}\left\|\mathbf{d}\right\|^2},$$

i.e.,

$$\tilde{t} > 2\beta(\alpha - 1)\frac{\nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d}}{M_{\nabla f}\left\|\mathbf{d}\right\|^2}. \tag{4.23}$$

The above lemma was first established in [Pat91a]. From Remark 3.1, by choosing φ strongly convex and such that (3.28) holds, the Armijo rule [i.e., Inequality (A.37)] can be used with $t_k = 1$, $\forall k$.

We next establish when, under Rules M, A, and P, the search directions tend to zero.

Lemma 4.4 (Search directions tend to zero) *Let f be bounded from below on X and ∇f Lipschitz continuous on X. Let $\varphi^k : X \mapsto \Re$ be a strongly convex function in C^1 on X with a Lipschitz continuous gradient on X, and let $0 < m_\varphi \leq m_{\varphi^k}$ and $M_{\nabla\varphi^k} \leq M_{\nabla\varphi} < +\infty$, $\forall k$. Then, from any starting point $\mathbf{x}^0 \in X$, under Rules M and A,*

$$\left\{\mathbf{d}^k\right\} \to \mathbf{0}. \tag{4.24}$$

For Rule P,[7] the same conclusion holds if $t_k \leq t_k^{\max}$, and either t_k belongs to a compact subset of $(0, \frac{2m_{\varphi^k}}{M_{\nabla f}})$, or $t_k \equiv t \in (0, \frac{2m_\varphi}{M_{\nabla f}})$.

Proof The proof is similar to that of Theorem 3.3; we begin by showing that under each rule a sufficient decrease in f is obtained in each step.

Rule M From (4.21) and the strong convexity of φ^k we obtain that, for $t \in [0, t_k^{\max}]$,

$$f\left(\mathbf{x}^{k+1}\right) \leq f\left(\mathbf{x}^k + t\mathbf{d}^k\right) \leq f\left(\mathbf{x}^k\right) + \left(-tm_{\varphi^k} + \frac{M_{\nabla f}}{2}t^2\right)\left\|\mathbf{d}^k\right\|^2. \tag{4.25}$$

[7] Observe that $t_k^{\max} \geq 1$, $\forall k$, so that Rule P can always be applied.

Minimizing the right-hand side of this inequality, we obtain that $t = \bar{t} = \frac{m_{\varphi^k}}{M_{\nabla f}}$. Hence, if $\bar{t} \leq t_k^{\max}$, we have

$$f\left(\mathbf{x}^{k+1}\right) - f\left(\mathbf{x}^k\right) \leq -\frac{m_{\varphi^k}^2}{2M_{\nabla f}} \left\|\mathbf{d}^k\right\|^2,$$

while, if $\bar{t} > t_k^{\max}$, we have

$$f\left(\mathbf{x}^{k+1}\right) - f\left(\mathbf{x}^k\right) \leq t_k^{\max}\left(-m_{\varphi^k} + \frac{M_{\nabla f}}{2} t_k^{\max}\right) \left\|\mathbf{d}^k\right\|^2,$$

where $-m_{\varphi^k} + \frac{M_{\nabla f}}{2} t_k^{\max} < -\frac{m_{\varphi^k}}{2} < 0$.[8]

Rule A From (4.22), (4.23) and the strong convexity of φ^k, we have

$$\begin{aligned} f\left(\mathbf{x}^{k+1}\right) - f\left(\mathbf{x}^k\right) &\leq \alpha t_k \nabla f\left(\mathbf{x}^k\right)^{\mathrm{T}} \mathbf{d}^k \\ &\leq -\frac{2\beta\alpha(1-\alpha)}{M_{\nabla f}\left\|\mathbf{d}^k\right\|^2}\left(\nabla f\left(\mathbf{x}^k\right)^{\mathrm{T}} \mathbf{d}^k\right)^2 \\ &\leq -\frac{2\beta\alpha(1-\alpha)m_{\varphi^k}^2}{M_{\nabla f}}\left\|\mathbf{d}^k\right\|^2. \end{aligned}$$

Rule P From (4.25) we obtain a decrease in f whenever $t_k \leq t_k^{\max}$ and either t_k belongs to a compact subset of $(0, \frac{2m_{\varphi^k}}{M_{\nabla f}})$, or $t_k \equiv t \in (0, \frac{2m_{\varphi}}{M_{\nabla f}})$.

Since f is lower bounded on X and, from the above, $\{f(\mathbf{x}^k)\}$ is decreasing, $\{f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k)\} \to 0$. Thus, in all the three rules we obtain (4.24). □

We are now ready to state the main result of this section.

Theorem 4.3 (Convergence under various step length rules)[9, 10] *Let f be bounded from below on X and ∇f Lipschitz continuous on X.*

(a) *Let $\varphi : X \times X \mapsto \Re$ be a continuous function on $X \times X$ of the form $\varphi(\mathbf{x}, \mathbf{y})$, convex and in C^1 on X with respect to $\mathbf{x}$, and assume that X is a bounded set. Then, from any starting point $\mathbf{x}^0 \in X$, under Rule A any accumulation point $\mathbf{x}^\infty$ of the sequence $\{\mathbf{x}^k\}$ (at least one such point exists) lies in Ω, and (4.20) holds.*

(b) *Let $\varphi^k : X \mapsto \Re$ be a strongly convex function in C^2 on X with a Lipschitz continuous gradient on X. Let $0 < m_\varphi \leq m_{\varphi^k}$ and $M_{\nabla\varphi^k} \leq M_{\nabla\varphi} < +\infty$, $\forall k$. Then from any starting point $\mathbf{x}^0 \in X$, under Rules M and A any accumulation point $\mathbf{x}^\infty$ of the sequence $\{\mathbf{x}^k\}$ lies in Ω. Furthermore, if $\{\mathbf{x}^k\}$ is bounded and Ω is nonempty, then (4.20) holds. If further Ω is finite, and the step lengths t_k are upper bounded, $\{\mathbf{x}^k\}$ converges to one point in Ω. For Rule P, the same conclusions hold if $t_k \leq t_k^{\max}$ and either t_k belongs to a compact subset of $(0, \frac{2m_{\varphi^k}}{M_{\nabla f}})$, or $t_k \equiv t \in (0, \frac{2m_\varphi}{M_{\nabla f}})$.*

Proof

[8] Note here that $t_k \geq \min\{t_k^{\max}, \frac{m_{\varphi^k}}{M_{\nabla f}}\}$ [compare the proof of Theorem 2.3.b].

[9] From the proof for Rule A in Lemma 4.4, convergence is actually guaranteed for *any* step length rule that results in a larger reduction in f at each iteration than the Armijo rule. The same holds for Rule P, and it also holds for the corresponding results in [**NLP**] (Theorem 2.3.b), [**NNLP**] (Theorem 3.3), and [**VIP**] (Theorem 5.13).

[10] As noted in Section 3.7, part **(b)** of the theorem holds for functions φ^k that are in C^1 on X only, if, in addition to the other assumptions, we impose convexity on f and assume that $\{\mathbf{x}^k\}$ is bounded.

(a) See [Pat91a, Th. 2.3]. The result (4.20) follows from Lemma 2.3, by identifying T with X and S with $\Omega \cap X$.

(b) We first note that (4.24) holds, from Lemma 4.4. The rest of the proof then follows from using the same line of arguments as in the proof of Theorem 3.3, using the characterization (4.17) of $Y(\mathbf{x}^k)$. We conclude that any accumulation point $\mathbf{x}^\infty$ of $\{\mathbf{x}^k\}$ lies in Ω. The validity of (4.20) follows as in **(a)**. Finally, convergence toward one point in Ω follows from (2.10) and Lemma 2.4. □

4.8 A comparison of convergence conditions

In Theorem 4.1 we established an equivalence among the generic subproblems of the cost approximation method and the algorithm classes of Cohen, Dafermos, Migdalas, and Tseng. The assumptions put on the objective f, the feasible set X, the subproblem defining maps, and the step length rules used are, however, quite different from each other, a main reason being that the motivations behind the formulations of these algorithm classes and the contexts in which they are presented are different.[11] One important advantage of our convergence analysis is that it includes the classical Frank-Wolfe and Newton methods, unlike either of the mentioned classes. Below we describe the convergence conditions of the classes of Cohen, Migdalas, and Tseng, and compare them to the above analysis.

Cohen [Coh80] uses Rule P, and proves that any accumulation point of $\{\mathbf{x}^k\}$ (at least one exists) is optimal in [**CNLP**], under the assumptions that f is a convex and weakly coercive[12] (see Definition A.3.a) function in C^1 on X with a Lipschitz continuous gradient on X, that $\{\varphi^k\}$ is a sequence of strongly convex functions in C^1 on X with Lipschitz continuous gradients on X, and, finally, that either, with $t_k \equiv 1$, the sequence $\{\varepsilon_k\}$ defining (4.7) satisfies

$$a_1 \leq \varepsilon_k \leq \frac{2m_{\varphi^k}}{M_{\nabla f} + a_2}, \quad a_1, a_2 > 0, \ \forall k, \tag{4.26}$$

or, with $\varepsilon_k \equiv 1$, the sequence $\{t_k\}$ satisfies $t_k \leq t_k^{\max}$ and (4.26).

Remark 4.3 It is interesting to note that the convergence conditions for sequences $\{\mathbf{x}^k\}$ defined by a scaling of φ^k with ε_k (see Theorem 3.1) and using unit steps, and sequences $\{\mathbf{x}^k\}$ defined by unscaled subproblems and predetermined step lengths (see Theorem 4.3.b) are *equivalent*, in the sense that the scaling ε_k in the first method and the step length t_k in the second are both subject to the same restrictions, namely the interval constraint (4.26); the sequences generated by the two methods are, however, not necessarily the same, since the subproblems differ. Note that the Condition (4.26) and the step length restriction of e.g. Theorem 4.3.b are equivalent.

[11] Migdalas introduces a strictly convex function in order to avoid extreme point subproblem solutions, thereby speeding up the slow convergence of the Frank-Wolfe method. Cohen considers optimal control problems, which, in many instances, have objectives that are expensive to evaluate; hence the choice of a predetermined step rather than a time-consuming line search.

[12] The weak coercivity condition on f is, in our analysis of Rule P, replaced by the less restrictive lower boundedness of f on X. The weak coercivity condition, however, has the interesting consequence that, since $\{f(\mathbf{x}^k)\}$ converges, the sequence $\{\mathbf{x}^k\}$ must be bounded, which in turn ensures that there exists an accumulation point of this sequence. Replacing lower boundedness with weak coercivity in the conditions of Theorem 4.3.b we immediately have that (4.20) holds.

Dussault *et al.* [DFL86], as well as Larsson and Migdalas [Mig88, LaM90], prove optimality of any accumulation point of $\{\mathbf{x}^k\}$, under Rule M, assuming that X is bounded, f is convex on X, and φ is strictly convex on X.

Migdalas [Mig90] uses Rule M and proves that any accumulation point of $\{\mathbf{x}^k\}$ is optimal in [**CNLP**], under the assumptions that X is bounded, f is pseudoconvex on X and h is strictly convex on X with respect to $\mathbf{x}$.

Tseng [Tse91b] uses a step length rule similar to Rule A, and proves that any accumulation point of $\{\mathbf{x}^k\}$ is optimal in [**CNLP**], under the assumptions that f is a convex and lower bounded function in C^1 on X, and $\mathbf{H}$ is strongly proximal related; the latter assumption may be replaced by only strictly proximal relatedness if it is further assumed that X is bounded.

Comparing the above convergence conditions with those imposed in Theorem 4.2 and 4.3, we may conclude that it is possible to remove the boundedness of X or the strict convexity of φ or h (or strict/strong proximal relatedness of $\mathbf{H}$), and still obtain the desired result, i.e., that the convergence conditions of the mentioned algorithm classes are more restrictive than necessary.

An important feature of the algorithm framework developed in this thesis is that, when applying the general convergence analysis of cost approximation methods to special cases, results obtained from previous specially made analyses are reproduced. Below we give a few examples of this fact.

4.9 Application of the convergence analysis

We first consider the Frank-Wolfe method, resulting from choosing $\varphi \equiv 0$. Under the assumptions that X is bounded and that Rule M is used, any accumulation point of $\{\mathbf{x}^k\}$ is a stationary point, i.e., (4.1) is satisfied [DeR65, DeR67, DeR70, Mar75]. This result also follows from Theorem 4.2. Further assuming that f is lower bounded and ∇f is Lipschitz continuous, the same conclusion holds using Rule A [PsD78, Le. III.3.2]. Here, Theorem 4.3.a is used. Convergence using the descent approach can, however, seemingly not be proved for predetermined step lengths, since φ is only convex.[13] Further note that since Cohen, Migdalas and Tseng all impose assumptions that correspond to $\mathbf{\Phi}$ being (at least) strictly monotone, the classical Frank-Wolfe algorithm is not included in their convergence analyses.

Next, we study (deflected) gradient projection methods, where $\varphi^k(\mathbf{x}) = \frac{1}{2\gamma_k}\mathbf{x}^{\mathrm{T}}\mathbf{B}_k\mathbf{x}$ for some $\gamma_k > 0$ and a symmetric and positive (semi)definite matrix $\mathbf{B}_k$. The gradient projection method ($\mathbf{B}_k \equiv \mathbf{I}$) is studied in [Gol64, LeP66]. With $m_{\varphi^k} = \frac{1}{\gamma_k}$ and $t_k \equiv 1$ in Rule P, the step length restriction reduces to γ_k belonging to a compact subset of $(0, \frac{2}{M_{\nabla f}})$. This requirement on γ_k is equivalent to that imposed in the convergence theorems of [Gol64, LeP66]. (See also Corollary 3.1.)

If $\mathbf{B}_k = \nabla^2 f(\mathbf{x}^k)$ and $\gamma_k = 1$, then Newton's method is obtained. Convergence studies for Newton methods are often made under strong convexity assumptions on f, see e.g. [Dan69, Dan70, PsD78, Pol87], or for combinations with relaxation/active set methods, see e.g. [Fle87]. To apply Theorems 4.2 and 4.3 it is sufficient to require that f is a convex function in C^2 on X. Results similar to those of Theorem 4.3.a are established for a step

[13] Sufficient conditions for convergence of a Frank-Wolfe method using predetermined step lengths are given e.g. in [PoS82].

length rule similar to Rule A in [PsD78, Th. III.3.3], and for Rule G in [Dun80b]. It is interesting to note that Newton's method is not covered by the convergence analyses of Cohen, Migdalas or Tseng.

In splitting methods for quadratic programs, usually, $\mathbf{x}^{k+1} = \mathbf{y}^k$ is set. Applying the convergence results for Rule P in Theorem 4.3.b to this special case, the conditions for convergence using unit steps are equivalent to the condition for regularity of the splitting $(\mathbf{A}, \mathbf{B})$. (See also [Pan82, Pan84a].) Further, according to Theorem 4.3.b, the splitting can be varied over the iterations, provided that the eigenvalues of $\mathbf{A}_k - \mathbf{B}_k$ stay bounded (see also [Man77, Pan86b, LuT92a, LuT92d, Ius93]).

4.10 A truncated version of cost approximation algorithms

The subproblem [$\mathbf{CNLP}_\varphi$] is, for a general choice of φ, potentially as difficult to solve as [**CNLP**], and a truncated version of the cost approximation method is therefore of practical interest. The truncation of the solution of [$\mathbf{CNLP}_\varphi$] is done in the same manner as for [$\mathbf{NNLP}_\varphi$], that is, the only requirement is that at least one iteration is made on a descent method with a closed algorithmic map. The specialization of Theorem 3.4 to [**CNLP**] is given below.

Theorem 4.4 (Convergence of truncated cost approximation algorithms) *Let $\varphi : X \times X \mapsto \Re$ be a continuous function on $X \times X$ of the form $\varphi(\mathbf{x}, \mathbf{y})$, convex and in C^1 on X with respect to $\mathbf{x}$. Assume that $\mathbf{x}^0$ is chosen such that $L_X^f(\mathbf{x}^0)$ is bounded, and that the algorithm used for solving [$\mathbf{CNLP}_\varphi^k$] is a descent algorithm with a closed algorithmic map. If the termination criteria chosen for [$\mathbf{CNLP}_\varphi^k$] are such that $1 \leq l^k < \infty$ iterations are performed, $\forall k$, the sequence $\{\mathbf{y}^k\}$ of approximate subproblem solutions is bounded, and if Rule M is used, then $\{f(\mathbf{x}^k)\} \to f(\overline{\mathbf{x}})$ for some $\overline{\mathbf{x}} \in \Omega$, any accumulation point $\mathbf{x}^\infty$ of the sequence $\{\mathbf{x}^k\}$ (at least one such point exists) lies in Ω, and (4.20) holds. Under the additional assumption that ∇f is Lipschitz continuous on X, the same conclusions hold under Rule A.*

Proof See [Pat91b]. The second part of the theorem is similar to that of Theorem 3.4, and utilizes [Pat91a, Th. 2.3]. □

Remark 4.4 Observe that the theorem is valid for the unconstrained program [**NLP**], and therefore constitutes a basis for validating truncation strategies for this problem.

Migdalas [Mig90] proposes to use a truncated Frank-Wolfe method on the subproblem in the regularized Frank-Wolfe algorithm (see Section 4.4). An advantage of the use of Frank-Wolfe for [$\mathbf{CNLP}_\varphi^k$] is the availability of a termination criterion through its lower bound (4.43); the solution is terminated if, for some $\varepsilon_k > 0$, a point $\mathbf{y}_{\varepsilon_k}^k \in X$ is found, satisfying

$$\nabla f_\varphi \left(\mathbf{y}_{\varepsilon_k}^k\right)^{\mathrm{T}} \left(\mathbf{y} - \mathbf{y}_{\varepsilon_k}^k\right) \geq -\varepsilon_k, \ \forall \mathbf{y} \in X, \tag{4.27}$$

i.e., $\mathbf{y}_{\varepsilon_k}^k$ is an ε_k-optimal solution to [$\mathbf{CNLP}_\varphi^k$] [DeV85, Le. 1.8.2]. The validity of (4.27) is checked automatically in the Frank-Wolfe algorithm, since the left hand side of (4.27) is minimized over $\mathbf{y} \in X$ in its subproblem phase. To ensure convergence, Migdalas requires $\{\varepsilon_k\} \to 0$ to hold, so that the subproblem is solved exactly in the limit.[14, 15]

4.11 Linear convergence of cost approximation algorithms

In Section 3.9, linear convergence is established for certain cost approximation methods applied to [**NNLP**], where the function φ is defined by a fixed quadratic form and the step length Rules A' and P are used; the convergence results obtained are directly applicable to [**CNLP**]. It is not possible to include Rule M in the analysis of [**NNLP**], since a lower bound on the step taken is not available. In the differentiable case, however, the step taken in an exact line search may be lower bounded (cf. the proof of Lemma 4.4), and Rule M may therefore be included in the analysis below.

Theorem 4.5 (Linear convergence of cost approximation algorithms) *Let ∇f be Lipschitz continuous on X. Let $\varphi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}$, where $\mathbf{B}$ is a symmetric and positive definite matrix on X with eigenvalues in the interval $[m_\varphi, M_{\nabla\varphi}]$. Assume that, using Rule M, A, or Rule P with t_k chosen in a compact subset of $(0, \frac{2m_\varphi}{M_{\nabla f}})$, $\forall k$, $\{\mathbf{x}^k\} \to \mathbf{x}^* \in \Omega$, where f is twice continuously differentiable and $\nabla^2 f(\mathbf{x}^*)$ is positive definite. Then, if $\mathbf{B}$ is chosen so that*

$$\sigma \stackrel{\text{def}}{=} \frac{M_{\nabla\varphi}}{m_\varphi} \left\| \mathbf{I} - \mathbf{B}^{-1}\nabla^2 f(\mathbf{x}^*) \right\| < 1, \tag{4.28}$$

$\{\mathbf{x}^k\}$ converges at least linearly with the ratio

$$q \le \max\left\{\sigma, 1 - \frac{m_\varphi}{M_{\nabla f}}(1-\sigma)\right\}, \qquad \textit{for Rule M,}$$

$$q \le \max\left\{\sigma, 1 - \frac{2\beta(1-\alpha)m_\varphi}{M_{\nabla f}}(1-\sigma)\right\}, \qquad \textit{for Rule A, and}$$

$$q \le \max\{\sigma, 1 - \underline{t}(1-\sigma)\}, \qquad \textit{for Rule P,}$$

where $\underline{t} = \inf_k\{t_k\} > 0$.

Proof The results for Rules A and P follow from the same line of arguments as in the proof of Theorem 3.5. For Rule M, the analysis is similar, where we use the lower bound $t_k \ge \min\{1, \frac{m_\varphi}{M_{\nabla f}}\}$. □

Remark 4.5 Corollary 3.1 establishes linear convergence of gradient projection methods under Rules A and P; from the above theorem, this holds also for Rule M.

Remark 4.6 For Rules A and P, a superlinear rate of convergence may be obtained (cf. Remark 3.3); for Rule M, however, this is not possible without imposing strong assumptions on the eigenvalue structure of $\nabla^2 f$.

[14] From the proof of the above theorem, the subproblem [$\mathbf{CNLP}^k_\varphi$] will, in the truncated version of the cost approximation algorithm, also be solved exactly in the limit, although this is not imposed explicitly.

[15] For the subproblem in Newton's method, [DeT88] also apply a truncated Frank-Wolfe algorithm. By imposing further conditions on $\{\varepsilon_k\}$, convergence rate results are derived [DeT84, DeT85, DeT88]. (See also [DES82, Nas82, DeS83, Nas84, Nas85a, Nas85b] for similar results in the unconstrained case.)

4.12 Finite identification of active constraints

To be able to finitely identify the set of active constraints at an optimal solution (or the optimal face) in an iterative algorithm is of practical importance; the reason is that once this set is identified the problem becomes essentially unconstrained. Then, convergence rate results for unconstrained programming can be applied.

In this section we shall show that if a strict complementarity condition holds at a point in Ω, the class of cost approximation methods will finitely identify the set of active constraints at this point. This result generalizes identification results earlier obtained for instances of cost approximation methods.

The finiteness results developed in this section rely on the asymptotic properties of the *projected gradient*,

$$\nabla^X f(\mathbf{x}) \stackrel{\text{def}}{=} \arg \min_{\mathbf{v} \in T_X(\mathbf{x})} \|\nabla f(\mathbf{x}) + \mathbf{v}\|, \ \mathbf{x} \in X, \tag{4.29}$$

i.e., the projection of $-\nabla f(\mathbf{x})$ onto the tangent cone $T_X(\mathbf{x})$. Using Moreau decomposition [Mor62], see also [Zar71, Le. 2.2], which states that $\mathbf{d} \in \Re^n$ can be uniquely expressed as

$$\mathbf{d} = P_K(\mathbf{d}) + P_{K^\circ}(\mathbf{d}), \tag{4.30}$$

where K is a closed and convex cone, we may restate the projected gradient as

$$\nabla^X f(\mathbf{x}) = \arg \min_{-\mathbf{z} \in \nabla f(\mathbf{x}) + N_X(\mathbf{x})} \|\mathbf{z}\|, \ \mathbf{x} \in X, \tag{4.31}$$

i.e., as the negative of the shortest element in the subdifferential of $f + \delta_X$ at $\mathbf{x} \in X$ (see also [BuM88]). Yet another equivalent definition of the projected gradient, which shows its strong relation to the projected gradient method [which is defined through (4.5)] is ([Zar71, Le. 4.6], [McT72])

$$\nabla^X f(\mathbf{x}) = \lim_{t \downarrow 0} \frac{1}{t} \left[P_X \left(\mathbf{x} - t \nabla f(\mathbf{x}) \right) - \mathbf{x} \right], \ \mathbf{x} \in X. \tag{4.32}$$

Considering the above expressions, a more appropriate term for $\nabla^X f(\mathbf{x})$ might be the *projected steepest descent direction* [BMT90]. The notion of projected gradient is a specialization to differentiable programming of the negative of the *shortest conditional subgradient* [DeV85, Secs. 1.6.4 and 1.11.3], see also [DeS78, DeS80].[16]

The projected gradient has been used as a step direction in descent algorithms [McT72, DeT83] (see also [Wol72, DuF93]),[17] but more commonly it is used as a tool in convergence analyses of other methods, see e.g. [DeT84, DeT85, BMT90]. In particular, the (deflected) gradient projection method (4.5) has been analyzed extensively using the concept of projected gradients [GaB84, Phe85, Phe86, CaM87, BuM88].

The identification results established in this section are based on the characterization (see (4.31) and [CaM87])

$$\nabla^X f(\mathbf{x}^*) = \mathbf{0} \Leftrightarrow \mathbf{x}^* \in \Omega. \tag{4.33}$$

[16]Methods based on (approximate) solutions of the nondifferentiable version of (4.31) [$\nabla f(\mathbf{x})$ replaced by (an approximation of) $\partial f(\mathbf{x})$] are given in [BeM73, CHL82, PMW83, Kiw85, BNS87, Sre87, KiU89, LPS92].

[17]The step direction in Rosen's gradient projection method [Ros60, Ros61] can be interpreted as the result of a heuristic solution of (4.29), see e.g. [Lue84, Exc. 11.10.8].

Although $\nabla^X f$ is discontinuous in general,[18] the gradient projection method has been shown to force the projected gradient to zero if ∇f is uniformly continuous and f bounded below on X [CaM87, BuM88]. In fact, it is possible to show that the subproblems defined in the class of cost approximation methods force the projected gradient to zero under similar conditions (see Lemma 4.6 below).

Burke and Moré [BuM88] show that the sequences that make the projected gradient tend to zero characterize those sequences that finitely identify the optimal face, i.e., the set of active constraints. The assumptions put on the limit point $\mathbf{x}^*$ of the sequence of iterates are that it satisfies the nondegeneracy Condition **(a)** [see Definition A.8.a], and that it lies in the relative interior of a *quasi-polyhedral face* X_F of X. (A face X_F of X is said to be quasi-polyhedral if $\mathtt{aff}\ X_F = \mathbf{x} + \mathtt{lin}\ T_X(\mathbf{x})$, for any $\mathbf{x} \in \mathtt{rint}\ X_F$.)[19]

We give their results below.

Lemma 4.5 [BuM88] *Assume that $\{\mathbf{z}^k\}$ is a sequence in X which converges to a point $\mathbf{z}^* \in \Omega$ where Condition* **(a)** *holds. If X_F is a quasi-polyhedral face of X with $\mathbf{z}^* \in \mathtt{rint}\ X_F$, then $\mathbf{z}^k \in \mathtt{rint}\ X_F$ for all k sufficiently large if and only if $\{\nabla^X f(\mathbf{z}^k)\} \to \mathbf{0}$.*

Lemma 4.5 is particularly interesting in the case of polyhedral sets X, since it then can be expressed in terms of active constraints. Let $X = \{\mathbf{x} \in \Re^n \mid \mathbf{a}_i^{\mathrm{T}}\mathbf{x} \leq b_i,\ i = 1, \ldots, m\}$, and let $\mathcal{I}(\mathbf{x}) = \{i \mid \mathbf{a}_i^{\mathrm{T}}\mathbf{x} = b_i\}$ denote the set of active constraints at $\mathbf{x}$.

Corollary 4.1 [BuM88] *Let X be a nonempty polyhedron in $\Re^n$, and assume that $\{\mathbf{z}^k\}$ is a sequence in X which converges to a point $\mathbf{z}^* \in \Omega$ where Condition* **(a)** *holds. Then $\mathcal{I}(\mathbf{z}^k) = \mathcal{I}(\mathbf{z}^*)$ for all k sufficiently large if and only if $\{\nabla^X f(\mathbf{z}^k)\} \to \mathbf{0}$.*

We now prove that the projected gradient tends to zero in the cost approximation methods.

Lemma 4.6 (The cost approximation algorithms force $\nabla^X f$ to zero) *Let $\varphi : X \times X \mapsto \Re$ be a continuous function on $X \times X$ of the form $\varphi(\mathbf{x}, \mathbf{y})$, convex and in C^1 on X with respect to $\mathbf{x}$. Further, let ∇f and $\nabla_{\mathbf{x}}\varphi$ be uniformly continuous on X and $X \times X$, respectively.*[20] *Let $\{\mathbf{x}^k\} \to \mathbf{x}^*$, and assume that (4.24) holds. Then,*

$$\left\{\nabla^X f\left(\mathbf{y}^k\right)\right\} \to \mathbf{0}. \tag{4.34}$$

Proof The necessary optimality conditions of $[\mathbf{CNLP}_\varphi^k]$ yield

$$-\left[\nabla_{\mathbf{x}}\varphi\left(\mathbf{y}^k, \mathbf{x}^k\right) + \nabla f\left(\mathbf{x}^k\right) - \nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^k, \mathbf{x}^k\right)\right] \in N_X\left(\mathbf{y}^k\right). \tag{4.35}$$

From the Moreau decomposition (4.30) we have that

$$\left\|\nabla^X f\left(\mathbf{y}^k\right)\right\| = \inf_{\mathbf{v} \in N_X\left(\mathbf{y}^k\right)} \left\|\nabla f\left(\mathbf{y}^k\right) + \mathbf{v}\right\|.$$

[18] However, $\|\nabla^X f\|$ is l.s.c. on X [CaM87], and $\nabla^X f$ is continuous on the relative interior of any face of X [BuM88].

[19] The relative interior of a convex set is a quasi-polyhedral face [BuM88]. In the case when X is a polyhedron, any face of X is quasi-polyhedral [BuM88, Th. 2.6], and, as a consequence, any point in X lies in the relative interior of a quasi-polyhedral face of X. The concept of *open facet* [Dun87] is equivalent to that of the relative interior of a quasi-polyhedral face [BuM88].

[20] If the sequences $\{\mathbf{x}^k\}$ and $\{\mathbf{y}^k\}$ are bounded, this assumption can be removed.

From (4.35) it then follows that

$$\left\|\nabla^X f\left(\mathbf{y}^k\right)\right\| \leq \left\|\left[\nabla f\left(\mathbf{y}^k\right) - \nabla f\left(\mathbf{x}^k\right)\right] + \left[\nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^k, \mathbf{x}^k\right) - \nabla_{\mathbf{x}}\varphi\left(\mathbf{y}^k, \mathbf{x}^k\right)\right]\right\|.$$

But from the uniform continuities of ∇f and $\nabla_{\mathbf{x}}\varphi$ and (4.24), the right hand side of the above inequality tends to zero, and we obtain (4.34). □

Combining this lemma with the results of [BuM88] above, we may state the following results regarding the identification properties of the cost approximation methods.

Theorem 4.6 (Finite identification of optimal face and active constraints) *Let $\varphi : X \times X \mapsto \Re$ be a continuous function on $X \times X$ of the form $\varphi(\mathbf{x}, \mathbf{y})$, convex and in C^1 on X with respect to $\mathbf{x}$. Further, let ∇f and $\nabla_{\mathbf{x}}\varphi$ be uniformly continuous on X and $X \times X$, respectively. Let $\{\mathbf{x}^k\} \to \mathbf{x}^*$, and assume that (4.24) holds. Assume further that $\mathbf{x}^* \in \Omega$ and that Condition* (**a**) *holds.*

(**a**) *If $\mathbf{x}^* \in \texttt{rint}\ X_F$, where X_F is a quasi-polyhedral face of X, then $\mathbf{y}^k \in \texttt{rint}\ X_F$ for all k sufficiently large.*

(**b**) *Assume further that X is polyhedral. Then $\mathcal{I}(\mathbf{y}^k) = \mathcal{I}(\mathbf{x}^*)$ for all k sufficiently large.*

Let X be polyhedral. If the sequence $\{\mathbf{x}^k\}$ can be shown to identify the optimal face finitely, the algorithm eventually reduces to its unconstrained version on the manifold of active constraints. The asymptotic convergence rate is then governed by the eigenvalue structure of $\nabla^2 f$ on the manifold $\{\mathbf{x} \in \Re^n \mid \mathbf{a}_i^{\mathrm{T}}\mathbf{x} = b_i,\ \forall i \in \mathcal{I}(\mathbf{x}^*)\}$.

Next, we give a sufficient condition for $\{\mathbf{x}^k\}$ to finitely identify the optimal face.

Corollary 4.2 (Finite identification of active constraints) [21, 22] *Under the assumptions of Theorem 4.6.b, by further assuming that*

$$\mathcal{I}\left(\mathbf{x}^k\right) \supseteq \mathcal{I}\left(\mathbf{y}^k\right) \tag{4.36}$$

holds for some k large enough for $\mathcal{I}(\mathbf{y}^k) = \mathcal{I}(\mathbf{x}^*)$ to hold, and $t_k \leq 1$ for all subsequent k, then

$$\mathcal{I}\left(\mathbf{x}^k\right) = \mathcal{I}\left(\mathbf{x}^*\right) \tag{4.37}$$

also holds for all such k.

Proof If $X \supset \{\mathbf{x}^k\} \to \mathbf{x}^*$, then for k sufficiently large,

$$\mathcal{I}\left(\mathbf{x}^k\right) \subseteq \mathcal{I}\left(\mathbf{x}^*\right). \tag{4.38}$$

But under the assumptions of Theorem 4.6.b,

$$\mathcal{I}\left(\mathbf{y}^k\right) = \mathcal{I}\left(\mathbf{x}^*\right) \tag{4.39}$$

eventually holds. If therefore (4.36) holds for some k sufficiently large for (4.39) to hold, and $t_k \leq 1$ for all subsequent k, then (4.36) must hold for *all* such k. Together with (4.38) and (4.39), this fact then implies that (4.37) also holds for all such k. $\square$

Remark 4.7 The Inclusion (4.36) may always be fulfilled, for instance by choosing $\{\varphi^k\}$ such that it is possible to choose $t_k = 1$ [cf. (3.32)], for some sufficiently large k.

By imposing conditions stronger than those leading to finite identification of the optimal face, the class of cost approximation methods can be shown to converge finitely. This is the subject of the next section. The conditions imposed in this case are highly related to the strict complementarity condition, and conditions that have been used in the analysis of finiteness of other methods. The relations among these regularity conditions have, however, not earlier been thoroughly investigated; in the below lemma, some of their relations are given. In Figure 4.2, Conditions **(a)**–**(l)** correspond to Definitions A.8.a–A.8.l. Further, P denotes that X is polyhedral, BP that X is also bounded, and QP that [**CNLP**] is a quadratic program.

Lemma 4.7 (Relations among regularity conditions) [23, 24] *Let f be convex on X, and $\Omega = \arg\min_{\mathbf{x}\in X} f(\mathbf{x})$ be nonempty. Then the relations described in Figure 4.2 hold among the regularity Conditions* **(a)**–**(l)**.

$$
\begin{tikzcd}
 & & \text{(d)} \arrow[r, Leftrightarrow] \arrow[d, Rightarrow] & \text{(g)} \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \text{(e)} \arrow[r, Leftrightarrow] \arrow[d, Rightarrow] & \text{(f)} \\
\text{(h)} & \text{(j)} \arrow[l, Rightarrow] \arrow[r, Leftrightarrow, "\text{(h)}"] & \text{(a)} & \text{(c)} \arrow[l, Rightarrow, "QP"'] \arrow[r, Leftrightarrow, "P"] & \text{(b)} & \\
\text{(i)} \arrow[u, Rightarrow] & \text{(k)} \arrow[u, Rightarrow, "P"] \arrow[r, Leftrightarrow, "BP"] & \text{(l)} \arrow[u, Rightarrow] & & &
\end{tikzcd}
$$

Figure 4.2: Relations among regularity conditions

[21] The Inclusion (4.36) amounts to $\mathbf{d}^k \in \{\mathbf{x} \in \Re^n \mid \mathbf{a}_i^{\mathrm{T}}\mathbf{x} = 0,\ \forall i \in \mathcal{I}(\mathbf{x}^*)\}$.

[22] As a corollary to Lemma 4.6 it can be shown that $\{P_X(\mathbf{x}^k - \nabla f(\mathbf{x}^k)) - \mathbf{x}^k\} \to \mathbf{0}$ (see also [Tse91b]). If Lipschitz continuity of $\nabla_{\mathbf{x}}\varphi$ is further assumed, it can be shown that

$$\left\| P_X\left(\mathbf{x}^k - \nabla f\left(\mathbf{x}^k\right)\right) - \mathbf{x}^k \right\| \leq (2 + M_{\nabla\varphi}) \left\|\mathbf{y}^k - \mathbf{x}^k\right\|,$$

for all k. Using this inequality, and the measurements of decrease in f provided in the proof of Lemma 4.4, it is possible to prove that the Inequalities (3.1) and (3.3) of [LuT92d] hold. In order to prove linear convergence for linearly constrained problems, it remains to be shown that also (3.2) of [LuT92d] holds (see [LuT92d, Th. 3.1]). Proving (3.2), however, is the most difficult part of the convergence analysis, as remarked in [LuT92d, p. 416]. From the proof of [LuT92d, Le. 3.2], we see that proving (3.2) is roughly the same as proving that $\mathbf{x}^k$ is in the optimal face for large enough k; but then linear convergence also follows from the fact that the problem becomes essentially unconstrained.

Proof

[**(d)** $\Longleftrightarrow$ **(g)**] It follows from [Dun79, Th. 3.5] that (A.33) is equivalent to

$$-\nabla f(\mathbf{x}^*) \in \mathtt{int}\; N_X(\mathbf{x}^*). \tag{4.40}$$

But, from [Pol87, Le. 7.1.1], this is equivalent to (A.30).

[**(e)** $\Longleftrightarrow$ **(f)**] Follows from [Dun79, Th. 3.3].

[**(b)** $\Longleftarrow$ **(c)**] Follows from [BuF92, Th. 4.2].

[**(b)** $\stackrel{P}{\Longrightarrow}$ **(c)**] Follows from [BuF92, Th. 4.2].[25]

[**(a)** $\stackrel{QP}{\Longleftarrow}$ **(c)**] Follows from [FeM92, Le. 5 and Th. 6].

[**(a)** $\stackrel{\mathbf{(h)}}{\Longleftrightarrow}$ **(j)**] Follows from [BuM88].

[**(j)** $\stackrel{P}{\Longleftarrow}$ **(k)**] Follows from [MaD89].

[**(k)** $\stackrel{BP}{\Longleftrightarrow}$ **(l)**] Follows from [DuM89, Cor. 2.4].

[**(h)** $\Longleftarrow$ **(j)**] Follows from the fact that Guignard's constraint qualification is the weakest possible condition ensuring that a point in Ω is a Karush-Kuhn-Tucker point [GoT71].

[**(h)** $\Longleftarrow$ **(i)**] Follows from, e.g., [Pet73].

The remainder of the relationships follow from inspection. □

In the special case of (deflected) gradient projection methods [see (4.5)] with $\mathbf{B}_k \equiv \mathbf{B}$, $\gamma_k \equiv \gamma$ and unit steps (e.g., the basic Goldstein-Levitin-Polyak method), the finite identification of the optimal face under Condition **(a)** follows directly from Theorem 4.6.b.

In [Ber76b], finite identification is established for the basic gradient projection method ($\mathbf{B}_k \equiv \mathbf{I}$), when $X = \Re^n_+$ and γ_k in (4.5) is determined through an Armijo-type line search, under a strict complementarity condition, stronger than **(j)**, and an additional second-order sufficiency condition. This result is later extended in various directions. In [CaM87] it is extended (using the projected gradient) to general polyhedral sets, while also replacing the second-order condition with **(i)**. It is further extended to general convex sets in [Dun87, BuM88], assuming only **(a)** and that the limit point lies in an open facet, and in the relative interior of a quasi-polyhedral face, respectively.

In [Ber82a, Prop. 1.37] and [Ber82c], the results of [Ber76b] are extended to hold for certain partly diagonal matrices $\mathbf{B}_k$, for feasible sets $X = \Re^n_+$; for general polyhedra the

[23] Guignard's constraint qualification **(h)** always holds for polyhedral sets. The regularity Conditions **(k)** and **(l)** are implied by a slightly more restrictive regularity condition than **(j)**, in which *all* optimal Lagrange multipliers for active constraints are assumed positive [MaD89, DuM89]. Under a strong convexity assumption on f, the Condition **(k)** implies the regularity condition of [Rob80, Rob83], see [MaD89]. The latter condition is also implied by **(i)** and an additional second-order condition [Rob80].

[24] From the definition of Condition **(b)** and the lemma it follows that under either one of Conditions **(b)**–**(g)** there must be at least one optimal *extreme point*; under Conditions **(d)** and **(g)** it must be unique [BuM88, AlK91]. A sufficient condition for the relation **(a)** $\stackrel{QP}{\Longrightarrow}$ **(c)** to hold is provided in [FeM92]; further, under this condition the existence of an extreme point optimum is ensured also under Conditions **(a)**, **(j)**, **(k)**, and **(l)**. Any solvable linear program has the weak sharp minimum property **(c)** [MaM79, Pol87]; this is in [BuF92] extended to hold for proper, polyhedral convex functions. Unique optimal solutions to linear programs are sharp minima, i.e., **(d)** holds [Pol87, Le. 10.1.5].

[25] The equivalence of **(b)** and **(c)** for quadratic programs was earlier established by [FeM92, Th. 13].

result is shown in [Ber82c] under the additional Condition **(i)**. In [GaD88], a general matrix $\mathbf{B}(\mathbf{x})$ is used; their result requires a strict complementarity condition stronger than **(j)**, and in addition it requires **(i)** to hold.

Two-metric projection methods, in which the matrix $\mathbf{B}_k^{-1}$, defining the projection operator in (4.5), is replaced by some other symmetric and positive definite matrix, are studied in [GaB84]; the set X is assumed polyhedral, and conditions similar to **(i)** and **(j)** are used to establish finite identification results.[26]

It is interesting to note that a result closely related to Theorem 4.6.b is obtained in [MaD89, DuM89] for variational inequality problems. They consider Dafermos' class of algorithms, which is equivalent to the cost approximation methods in the sense of Theorem 4.1.a. Under the Conditions **(k)** and **(l)**, respectively, they show that the sequence of subproblem solutions attains the optimal face finitely.

Remark 4.8 The reader should observe that the finite identification results obtained for the class of cost approximation methods are based on the weakest regularity condition among those compared in the above lemma.

Remark 4.9 Virtually all results obtained in this section, as well as those in the next one, carry over immediately to variational inequalities, since the fact that ∇f is integrable is not used in the analysis. In particular, Theorem 4.6 can be shown to hold also for variational inequality problems (see Theorem 5.16). This result then generalizes those of [MaD89, DuM89], since, in their analyses, the set X is assumed polyhedral, unit steps are taken (whereas we only assume $\{\mathbf{d}^k\} \to \mathbf{0}$), and since both Conditions **(k)** and **(l)** are stronger than Condition **(a)**. For further discussions on this topic, see Section 5.13.

4.13 Finite convergence of cost approximation algorithms

The finite identification results of the previous section will here be sharpened; we show that the class of cost approximation algorithms terminate finitely at an optimal solution, under Conditions **(c)** or **(d)**. This result extends those earlier obtained for the Frank-Wolfe, gradient projection, Newton and proximal point methods.

Recall from Theorem 4.6.a that if $\mathbf{x}^*$ is in the relative interior of a quasi-polyhedral face of X and Condition **(a)** holds, then so is $\mathbf{y}^k$, for a sufficiently large k. Burke and Moré [BuM88] show that if the normal cone $N_X(\mathbf{x}^*)$ has a nonempty interior, then $\{\mathbf{x}^*\}$ is a quasi-polyhedral face. The implication is that $\mathbf{y}^k = \mathbf{x}^*$ for all sufficiently large k, i.e., finite convergence is obtained [BuM88, Cor. 3.5]. (This result may also be obtained as a corollary to Theorem 4.6.a; if $\mathtt{int}\, N_X(\mathbf{x}^*)$ is nonempty, then $\mathtt{rint}\, N_X(\mathbf{x}^*) = \mathtt{int}\, N_X(\mathbf{x}^*)$, which yields from (4.40) that Conditions **(a)** and **(d)** are equivalent.)

Finite convergence under Condition **(d)** has also been obtained in Theorem 3.6, which is directly applicable to [**CNLP**].

We specialize Lemma 3.7 to [**CNLP**], before establishing finite convergence.

Lemma 4.8 [Pol87, Le. 7.1.1] (Equivalent definitions of a sharp minimum) *Let f be convex on X. The following statements are equivalent.*

[26] Finite identification properties, assuming Condition **(a)**, have also been obtained for sequential quadratic programming [Wri89, BuM88] and trust region methods [BMT90].

(1) $\mathbf{x}^*$ *is a sharp minimum of* f *on* X.

(2) $\nabla f(\mathbf{x}^*)^{\mathrm{T}}(\mathbf{x}-\mathbf{x}^*) \geq \alpha \|\mathbf{x}-\mathbf{x}^*\|, \forall \mathbf{x} \in X$.

(3) $-\nabla f(\mathbf{x}^*) \in \mathtt{int}\ N_X(\mathbf{x}^*)$.

Theorem 4.7 (Finite convergence of cost approximation algorithms) *Let* f *be convex on* X. *Let* $\varphi : X \times X \mapsto \Re$ *be a continuous function on* $X \times X$ *of the form* $\varphi(\mathbf{x}, \mathbf{y})$, *convex and in* C^1 *on* X *with respect to* $\mathbf{x}$. *Let* $\{\mathbf{x}^k\} \to \mathbf{x}^*$, *and assume that* $\mathbf{x}^*$ *is a sharp minimum.*

(a) *Let* ∇f *and* $\nabla_{\mathbf{x}}\varphi$ *be uniformly continuous on* X *and* $X \times X$, *respectively, and assume that (4.24) holds. Then* $\mathbf{y}^k = \mathbf{x}^*$ *holds for all sufficiently large* k.

(b) *Let* ∇f *and* $\nabla_{\mathbf{x}}\varphi$ *be Lipschitz continuous on* X *and* $X \times X$, *respectively. Then* $\mathbf{y}^k = \mathbf{x}^*$ *holds for all* k, *such that* [27]

$$\left\|\mathbf{x}^k - \mathbf{x}^*\right\| \leq \frac{\alpha}{M_{\nabla\varphi} + M_{\nabla f}}. \tag{4.41}$$

Proof

(a) Follows from Theorem 4.6.a and the fact that $\mathtt{rint}\ X_F = \{\mathbf{x}^*\}$.

(b) If (4.41) holds, then, for any $\mathbf{x} \in X$, letting $\mathbf{d} = \mathbf{x} - \mathbf{x}^*$,

$$\begin{aligned}
\left[\nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^*, \mathbf{x}^k\right) + \nabla f\left(\mathbf{x}^k\right) - \nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^k, \mathbf{x}^k\right)\right]^{\mathrm{T}} \mathbf{d} &= \\
\left[\nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^*, \mathbf{x}^k\right) - \nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^k, \mathbf{x}^k\right)\right]^{\mathrm{T}} \mathbf{d} + \left[\nabla f\left(\mathbf{x}^k\right) - \nabla f(\mathbf{x}^*)\right]^{\mathrm{T}} \mathbf{d} + \nabla f(\mathbf{x}^*)^{\mathrm{T}} \mathbf{d} &\geq \\
\left(-\left\|\nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^*, \mathbf{x}^k\right) - \nabla_{\mathbf{x}}\varphi\left(\mathbf{x}^k, \mathbf{x}^k\right)\right\| - \left\|\nabla f\left(\mathbf{x}^k\right) - \nabla f(\mathbf{x}^*)\right\| + \alpha\right) \|\mathbf{d}\| &\geq \\
\left(-\left(M_{\nabla\varphi} + M_{\nabla f}\right)\left\|\mathbf{x}^k - \mathbf{x}^*\right\| + \alpha\right) \|\mathbf{d}\| &\geq 0,
\end{aligned}$$

i.e., from (4.4), $\mathbf{x}^*$ solves $[\mathbf{CNLP}_{\varphi}^k]$. The first inequality follows from Schwartz' Inequality and Lemma 4.8, and the second from the Lipschitz continuity. □

The theorem generalizes several earlier obtained results. The fact that the Frank-Wolfe algorithm is finite under Condition **(d)** is not surprising, since $\mathbf{x}^*$ must then be an extreme point of X (see Lemma 4.7), and the subproblem in the Frank-Wolfe algorithm always generates extreme points. Finiteness of this method has earlier been established in [Dun79, Dun80a] under (the equivalent) Condition **(g)** and Lipschitz continuity of ∇f. Finiteness of the gradient projection method is shown in [Pol79, Dun81] under Conditions **(d)** and **(g)**, respectively, and for Newton's method in [Dun80b] under Condition **(g)**. Finiteness of the proximal point algorithm is established in [Roc76a, Roc76b] under Condition **(d)**.

Remark 4.10 In the proof of Theorem 4.7, convexity of f is only used in the definition of a sharp minimum. Therefore, if the sharpness condition is relaxed to assume that f satisfies Condition **(2)** of Lemma 4.8 only *locally* around $\mathbf{x}^*$ (which then is a local minimum [Pol87, Th. 7.1.2] which further is isolated under a constraint qualification [AlK91, ShA93]), then Theorem 4.7 shows that cost approximation methods converge

[27] The upper bound on the number of iterations needed to fulfill (4.41), provided in the proof of Theorem 3.6, is directly applicable here.

finitely also for nonconvex problems. This result also provides an extension of that in [AlK91]; under the above local sharpness assumption, finite convergence is established for any convergent algorithm that invokes a Frank-Wolfe subproblem regularly. Clearly, this result extends to the inclusion of *any* cost approximation subproblem, satisfying the conditions of Theorem 4.7, in a convergent algorithm. The extension of this finiteness result to nondifferentiable problems is also valid (cf. the proof of Theorem 3.6).

Some algorithms included in the class of cost approximation methods have been shown to be finite under conditions slightly less restrictive than Condition **(d)**. (Note that this condition implies that $\mathbf{x}^*$ must be an extreme point.) A recent result of Burke and Ferris [BuF92] can be used to show that Theorem 4.7.a holds under the slightly less restrictive Condition **(c)**.

Theorem 4.8 [BuF92, Th. 4.7] *Let f be convex on X. Let Ω be a set of weak sharp minima for* [**CNLP**]. *If $\{\mathbf{z}^k\} \subset X$ is such that $\{\inf_{\mathbf{z}\in\Omega} \|\mathbf{z}^k - \mathbf{z}\|\} \to 0$ and ∇f is uniformly continuous on an open set containing $\{\mathbf{z}^k\}$, then $\mathbf{z}^k \in \Omega$ for all k sufficiently large if and only if $\{\nabla^X f(\mathbf{z}^k)\} \to \mathbf{0}$.*

This result can be applied directly to the cost approximation method, as follows.

Theorem 4.9 (Finite convergence of cost approximation algorithms) *Let f be convex on X. Let $\varphi : X \times X \mapsto \Re$ be a continuous function on $X \times X$ of the form $\varphi(\mathbf{x}, \mathbf{y})$, convex and in C^1 on X with respect to $\mathbf{x}$. Let $\{\mathbf{x}^k\} \to \mathbf{x}^*$, and assume that $\mathbf{x}^*$ is a weak sharp minimum. Let ∇f and $\nabla_{\mathbf{x}}\varphi$ be uniformly continuous on X and $X \times X$, respectively, and assume that (4.24) holds. Then $\mathbf{y}^k = \mathbf{x}^*$ holds for all sufficiently large k.*

Proof Immediate from Lemma 4.6 and Theorem 4.8. □

Not only does the above theorem sharpen the finiteness results for the class of cost approximation methods, but it also brings forward a strong relation between the identification and finiteness properties of this class of methods.

Theorem 4.9 extends the results obtained, under Condition **(c)**, for the proximal point algorithm, see [Luq84a, Fer91] and [BeT89, Prop. 3.4.1.d]. Since all solvable linear programs satisfy Condition **(c)** [MaM79], cost approximation methods applied to linear programs exhibit finite convergence; this has earlier been established for the special case of the proximal point method, see e.g. [PoT74, Ber75c]. Polyak [Pol87, Exc. 7.2.2] states in an exercise that the basic Goldstein-Levitin-Polyak gradient projection method has finite convergence under Condition **(c)**; this also follows from Theorem 4.9. In [BuF92], the idea developed in [AlK91] is extended from Condition **(d)** to **(c)**. Invoking a Frank-Wolfe subproblem regularly results in finite termination of any convergent algorithm applied to a problem with weak sharp minimum; clearly, this also holds if any cost approximation subproblem, satisfying the conditions of Theorem 4.9, is used in place of the Frank-Wolfe subproblem.

An extension of Theorem 4.9 to nonconvex problems is valid (similarly to an extension of Theorem 4.7, see Remark 4.10), if the set of local weak sharp minima is assumed to be regular in the sense of [BuF92]; cf. the proof of [BuF92, Cor. 4.5.b].

4.14 A posteriori error bounds and lower bounds

The following a posteriori error bound specializes Theorem 3.7 to [**CNLP**].

Theorem 4.10 (A posteriori error bound) *Let f and $\varphi : X \mapsto \Re$ be strongly convex on X, with Lipschitz continuous gradients on X. Then, for any $\mathbf{x} \in X$,*

$$\|\mathbf{y}(\mathbf{x}) - \mathbf{x}^*\| \leq \frac{M_{\nabla\varphi} + M_{\nabla f}}{m_f} \|\mathbf{y}(\mathbf{x}) - \mathbf{x}\|. \tag{4.42}$$

The lower bound result of Theorem 2.7 is directly applicable to [**CNLP**].

Theorem 4.11 (Lower bound) *Let $\varphi : X \mapsto \Re$ be a convex function in C^1 on X. Assume that the function $f - \varphi$ is convex on X. Then, for any $\mathbf{x} \in X$ and $\mathbf{y} \in Y(\mathbf{x})$,*

$$f_\varphi(\mathbf{y}) \leq f^*. \tag{4.43}$$

Furthermore, equality holds in (4.43) if and only if $\mathbf{x}$ is a minimizer of f on X.

A well known lower bound is a special case of the above result; for convex programs, the Frank-Wolfe subproblem objective value defines a lower bound on the optimal value of [**CNLP**], see e.g. [Mar75, p. 245]. Stopping criteria for a cost approximation method can be based on (4.43); one possibility is to terminate when the relative error $[f(\mathbf{x}^k) - f_\varphi(\mathbf{y}^k)]/f_\varphi(\mathbf{y}^k)$ becomes small enough.

4.15 Jacobi and Gauss-Seidel type methods

In this section we consider the important special case of [**CNLP**] where the feasible set is a Cartesian product of nonempty, closed and convex sets of smaller dimensions, i.e.,

$$X = \prod_{i \in \mathcal{C}} X_i, \; X_i \subseteq \Re^{n_i}, \; \sum_{i \in \mathcal{C}} n_i = n. \tag{4.44}$$

This special structure is inherent in many applications of network programming (in which case $\mathcal{C}$ is an index set of commodities or groups of users of the network), for instance in market and traffic equilibrium models [She85, Pat90, Nag93]. It is also obtained in subproblems in relaxation schemes for problems with *block-angular* constraint structures, that is, where, in addition to the constraints defining (4.44), coupling constraints are present [DMZ89, Rus89, PiZ92], and in Lagrangean dual formulations of strictly convex programs [BeT89, Tse90, TsB91].

We shall describe how the cost approximation methods can be made to utilize this structure, and validate specialized versions of the basic algorithm, with varying degrees of parallelism and asynchronism. In the first results we specialize the Gauss-Seidel type methods for [**NNLP**] discussed in Section 3.12 to the problem under consideration here; we validate the use of both exact and inexact line searches as well as the use of truncated subproblem solutions. By strengthening the conditions on φ, we then extend the cyclic rule for choosing indices to the *essentially* cyclic rule; we also introduce predetermined step length rules in the analysis.

The above mentioned methods are inherently sequential; Jacobi-type versions of the cost approximation method, in which subproblems are allowed to be solved in parallel, are therefore introduced and validated. This strategy may speed-up the practical convergence rate, but the need for a synchronizing step in the algorithm (the exchange of subproblem solutions among processors, and possibly a global line search) may still deteriorate the performance, since faster processors (subproblems) must wait for slower ones before the information exchange can be made. We therefore consider also asynchronous versions of the Jacobi-type methods, in which processors do not wait to receive the latest information available. The advantage of such an approach is a minimal delay in communication, which may speed-up the convergence compared to synchronous algorithms. Convergence still holds, provided that the information that any particular processor utilizes is not arbitrarily outdated.

We introduce $\varphi : X \times X \mapsto \Re$ as a separable function

$$\varphi(\mathbf{x}, \mathbf{y}) = \sum_{i \in \mathcal{C}} \varphi_i(\mathbf{x}_i, \mathbf{y}), \tag{4.45}$$

where $\varphi_i : X_i \times X \mapsto \Re$ is a continuous function on $X_i \times X$ and convex and in C^1 on X_i with respect to $\mathbf{x}_i$. The algorithm is defined as in Section 2.7 for [**NLP**] and Section 3.12 for [**NNLP**]. In iteration k, the index $i_k \in \mathcal{C}$ is chosen. If the cyclic rule is used,

$$i_k = k \ (\mathrm{mod}\, |\mathcal{C}|) + 1. \tag{4.46}$$

We then solve ($i = i_k$)

$$[\mathbf{CNLP}^k_{\varphi_i}] \qquad \mathbf{y}^k_i \in \arg\min_{\mathbf{x}_i \in X_i} \left\{ \varphi_i\left(\mathbf{x}_i, \mathbf{x}^k\right) + \left[\nabla_i f\left(\mathbf{x}^k\right) - \nabla\varphi_i\left(\mathbf{x}^k_i, \mathbf{x}^k\right)\right]^{\mathrm{T}} \left(\mathbf{x}_i - \mathbf{x}^k_i\right) \right\},$$

and define $\mathbf{y}^k_j = \mathbf{x}^k_j$, for all $j \neq i$. Let $\mathbf{x}^{k+1} = \mathbf{x}^k + t_k(\mathbf{y}^k - \mathbf{x}^k)$, i.e.,

$$\mathbf{x}^{k+1}_j = \begin{cases} \mathbf{x}^k_j + t_k\left(\mathbf{y}^k_j - \mathbf{x}^k_j\right), & j = i \\ \mathbf{x}^k_j, & j \neq i, \end{cases}$$

where

$$t_k \in \arg \min_{\mathbf{x}^k_i + t\left(\mathbf{y}^k_i - \mathbf{x}^k_i\right) \in X_i} f\left(\mathbf{x}^k_{i_-}, \mathbf{x}^k_i + t\left(\mathbf{y}^k_i - \mathbf{x}^k_i\right), \mathbf{x}^k_{i_+}\right)$$

if Rule M is used.

If $|\mathcal{C}| = 1$, the algorithm reduces to the original cost approximation method, while with $|\mathcal{C}| = n$ and φ chosen as in (2.19), it reduces to the classical Gauss-Seidel scheme for [**CNLP**].

The following theorem specializes Theorem 3.10 to [**CNLP**].

Theorem 4.12 (Convergence of cyclic cost approximation algorithms) [28] *Let X be defined by (4.44). Let $\varphi : X \times X \mapsto \Re$ be given by (4.45). Let the indices be chosen in cyclic order. Let $\mathbf{x}^0 \in X$ be chosen such that $L^f_X(\mathbf{x}^0)$ is bounded. Assume that the problem [$\mathbf{CNLP}_{\varphi_i}$] is well defined, in the sense that its solution set $Y_i(\mathbf{x})$ is nonempty and bounded for all $\mathbf{x} \in X$. Assume further that the sequence $\{\mathbf{y}^k\}$ is bounded. Then, under Rule M, $\{f(\mathbf{x}^k)\} \to f(\overline{\mathbf{x}})$ for some $\overline{\mathbf{x}} \in \Omega$, any accumulation point $\mathbf{x}^\infty$ of the sequence $\{\mathbf{x}^k\}$ (at least one such point exists) lies in Ω, and (4.20) holds.*

Proof A straightforward extension of the proof of Theorem 2.9. □

This basic algorithm will now be extended in several directions. We first show that Rule M can be replaced by the Armijo rule if the functions φ_i are strictly convex.

[28] If the Condition (4.46) on $\{i_k\}$ is replaced by the condition that every index $i \in \mathcal{C}$ is chosen an infinite number of times in $\{i_k\}$, we obtain the *free-steering* order [Sch62, Sch68, Elk68, OrR70, Mie71, FiH79]. Under the assumptions of the theorem, under Rules M and A (in which case we further assume that X

Theorem 4.13 (Convergence under the Armijo rule) *In addition to the assumptions of Theorem 4.12, let X be bounded, and $\nabla_i f$ Lipschitz continuous on X_i and $\varphi_i : X_i \times X \mapsto \Re$ strictly convex on X_i with respect to $\mathbf{x}_i$, for all $i \in \mathcal{C}$. Replacing Rule M with Rule A in the above cyclic decomposition scheme, the conclusions of Theorem 4.12 are still valid.*

Proof We first note that the Inequality (3.22) here becomes

$$f\left(\mathbf{x} + t\left(\left(\mathbf{x}_{i_-}, \mathbf{y}_i, \mathbf{x}_{i_+}\right) - \mathbf{x}\right)\right) - f(\mathbf{x}) \leq t \nabla_i f(\mathbf{x})^{\mathrm{T}} (\mathbf{y}_i - \mathbf{x}_i) + \frac{M_{\nabla_i f}}{2} t^2 \|\mathbf{y}_i - \mathbf{x}_i\|^2, \quad (4.47)$$

where $M_{\nabla_i f}$ satisfies

$$\|\nabla_i f(\mathbf{x}) - \nabla_i f(\mathbf{y})\| \leq M_{\nabla_i f} \|\mathbf{x}_i - \mathbf{y}_i\|, \ \forall \mathbf{x}_i, \mathbf{y}_i \in X_i, \mathbf{x}_j = \mathbf{y}_j, j \neq i.$$

Using (4.47) it is easy to show that Rule A is valid, and we obtain, for any k,

$$f\left(\mathbf{x}^{k+1}\right) - f\left(\mathbf{x}^k\right) \leq -2\beta\alpha(1-\alpha) \frac{\left[\nabla_{i_k} f\left(\mathbf{x}^k\right)^{\mathrm{T}} \left(\mathbf{y}_{i_k}^k - \mathbf{x}_{i_k}^k\right)\right]^2}{M_{\nabla_{i_k} f} \left\|\mathbf{y}_{i_k}^k - \mathbf{x}_{i_k}^k\right\|^2}.$$

Letting D_i be the diameter of X_i, and defining $D = \max_{i \in \mathcal{C}} D_i$ and $M = \max_{i \in \mathcal{C}} M_{\nabla_i f}$,

$$f\left(\mathbf{x}^{k+1}\right) - f\left(\mathbf{x}^k\right) \leq -\frac{2\beta\alpha(1-\alpha)}{MD^2} \left[\nabla f\left(\mathbf{x}^k\right)^{\mathrm{T}} \left(\mathbf{y}^k - \mathbf{x}^k\right)\right]^2.$$

Since f is lower bounded on X, this implies that $\{\nabla f(\mathbf{x}^k)^{\mathrm{T}}(\mathbf{y}^k - \mathbf{x}^k)\} \to 0$.

Let $\mathbf{x}^\infty$ be any accumulation point of the sequence $\{\mathbf{x}^k\}$, corresponding to a convergent subsequence $\{\mathbf{x}^k\}_{k \in \mathcal{K}}$, and $\mathbf{y}^\infty$ any accumulation point of $\{\mathbf{y}^k\}_{k \in \mathcal{K}}$, corresponding to a convergent subsequence $\{\mathbf{y}^k\}_{k \in \hat{\mathcal{K}}}$. We then have

$$\nabla f(\mathbf{x}^\infty)^{\mathrm{T}} (\mathbf{y}^\infty - \mathbf{x}^\infty) = 0. \quad (4.48)$$

We shall first show that $\mathbf{y}^\infty = \mathbf{x}^\infty$ holds. Choose an arbitrary index $i \in \mathcal{C}$. If it not occurs an infinite number of times in $\{i_k\}_{k \in \hat{\mathcal{K}}}$, $\mathbf{y}_i^k = \mathbf{x}_i^k$ holds for all sufficiently large k, and $\mathbf{y}_i^\infty = \mathbf{x}_i^\infty$ holds immediately. Assuming that it does occur an infinite number of times, we consider the subsequence of $\{i_k\}_{k \in \hat{\mathcal{K}}}$ that corresponds to this index. In the limit of this subsequence, the variational inequality characterization of [$\mathbf{CNLP}_{\varphi_i}^k$] yields

$$\left[\nabla \varphi_i (\mathbf{y}_i^\infty, \mathbf{x}^\infty) + \nabla_i f(\mathbf{x}^\infty) - \nabla \varphi_i (\mathbf{x}_i^\infty, \mathbf{x}^\infty)\right]^{\mathrm{T}} (\mathbf{x}_i - \mathbf{y}_i^\infty) \geq 0, \ \forall \mathbf{x}_i \in X_i.$$

Letting $\mathbf{x}_i = \mathbf{x}_i^\infty$, then if $\mathbf{y}_i^\infty \neq \mathbf{x}_i^\infty$, the strict convexity of φ_i would yield that

$$\nabla_i f(\mathbf{x}^\infty)^{\mathrm{T}} (\mathbf{y}_i^\infty - \mathbf{x}_i^\infty) < 0.$$

---

is bounded and $\nabla_i f$ is Lipschitz continuous on X_i for all i),

$$\nabla_i f(\mathbf{x}^\infty)^{\mathrm{T}} (\mathbf{x}_i - \mathbf{x}_i^\infty) \geq 0, \ \forall \mathbf{x}_i \in X_i$$

holds for those indices that occur infinitely many times in building the subsequence converging to $\mathbf{x}^\infty$. This result is analogous to those obtained earlier for general fixed point problems in [FiH75, FiH79], and can probably not be strengthened without introducing either restrictive convexity conditions on f or further restrictions on the way in which indices are chosen. (Convergence under the free-steering rule has been established for the Gauss-Seidel method applied to specially structured strictly convex problems [Pan84b, BHT87, BeT89] and strongly convex unconstrained programs [Sch62, Sch68, Elk68, OrR70].)

This in turn would contradict (4.48), and we conclude that $\mathbf{y}_i^\infty = \mathbf{x}_i^\infty$. Since i was arbitrary in $\mathcal{C}$, we obtain $\mathbf{y}^\infty = \mathbf{x}^\infty$.

Now, since

$$\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\| = t_k \left\|\mathbf{y}_{i_k}^k - \mathbf{x}_{i_k}^k\right\| \leq \left\|\mathbf{y}^k - \mathbf{x}^k\right\|,$$

$\{\mathbf{x}^{k+1} - \mathbf{x}^k\}_{k\in\hat{\mathcal{K}}} \to \mathbf{0}$, and $\{\mathbf{x}^{k+1}\}_{k\in\hat{\mathcal{K}}} \to \mathbf{x}^\infty$ follows. Repeated use of this argument (applying a technique similar to that of the proof of [Tse91b, Prop. 4.1]) shows that $\{\mathbf{x}^{k+j}\}_{k\in\hat{\mathcal{K}}} \to \mathbf{x}^\infty$ for all $j = 0, \dots, |\mathcal{C}| - 1$. Without loss of generality, we now assume (passing into a further subsequence if necessary) that $i_{k+j-1} = j$ for all $k \in \hat{\mathcal{K}}$. (That is, we let $\hat{\mathcal{K}}$ correspond to the subsequence of iterations where a fixed index was chosen.) From the optimality of $\mathbf{y}_j^{k+j-1}$ in $[\mathbf{CNLP}_{\varphi_j}^{k+j-1}]$ and the strict convexity of φ_j for all j, in the limit of $\hat{\mathcal{K}}$ we obtain that

$$\nabla_j f(\mathbf{x}^\infty)^{\mathrm{T}} \left(\mathbf{x}_j - \mathbf{x}_j^\infty\right) \geq 0, \ \forall \mathbf{x}_j \in X_j, \ j = 1, \dots, |\mathcal{C}|,$$

i.e., that $\mathbf{x}^\infty \in \Omega$.

The final statement follows from Lemma 2.3. □

Remark 4.11 The strict convexity of φ_i is essential for the theorem to hold. This is in contrast to the basic algorithm ($|\mathcal{C}| = 1$), where convexity is sufficient.

Remark 4.12 Using the same method of proof, the slightly more general *essentially cyclic rule* [Tse90, Tse91b, LuT92c] for choosing indices can be shown to yield convergence. This rule is an intermediate between the cyclic and free-steering rules, in which every index i is assumed to be chosen at least once every B successive iterations, i.e., there is a $B \geq |\mathcal{C}|$ such that[29]

$$\{1, \dots, |\mathcal{C}|\} \subseteq \{i_k, \dots, i_{k+B-1}\}, \ \forall k \geq 1. \tag{4.49}$$

Tseng [Tse91b] considers the nonlinear proximal descent algorithm (which is equivalent to the cost approximation methods, in the sense of Theorem 4.1.a) in a decomposition version analogous to that given in this section. Convergence is established for a step length rule similar to Rule A, under the strongly proximal relatedness of the maps $\mathbf{H}_i$ (see Section 4.4) and, when X is bounded, also for only strictly proximal related maps $\mathbf{H}_i$. The latter result is similar to that of the above theorem. When φ_i is assumed strongly convex, $\forall i$, the algorithm may, however, be simplified, to allow for the use of predetermined step lengths, and Rule A can be applied without the compactness assumption on X.

Theorem 4.14 (Convergence of essentially cyclic cost approximation algorithms) *Let X be defined by (4.44). Let $\varphi : X \times X \mapsto \Re$ be given by (4.45). Further, let f be lower bounded on X, and assume that $\nabla_i f$ is Lipschitz continuous on X_i and the function*

[29] The possibility of executing the algorithm with the essentially cyclic rule brings forward possibilities for devising rules for choosing indices that may speed up the practical convergence, compared to the cyclic rule. The basis for this is the fact that the bound B can be made arbitrarily large; within this bound the indices may be chosen according to any specified rule. A good strategy may be to generalize the Gauss-Southwell (remotest) order [Cen81, Lue84], in which the index chosen in a particular iteration is the one corresponding to the variable block being, in a certain sense, farthest from the set of optimal solutions. (For the traffic assignment problem, such rules, based on the Frank-Wolfe lower bound, are provided in [Pet75, Pat91b].)

$\varphi_i : X_i \times X \mapsto \Re$ strongly convex on X_i with respect to $\mathbf{x}_i$ and $\nabla\varphi_i$ Lipschitz continuous on X_i, for all $i \in \mathcal{C}$. Let indices be chosen according to the essentially cyclic rule (4.49). Then, from any starting point $\mathbf{x}^0 \in X$, under Rules M and A any accumulation point $\mathbf{x}^\infty$ of the sequence $\{\mathbf{x}^k\}$ lies in Ω. Furthermore, if $\{\mathbf{x}^k\}$ is bounded and Ω is nonempty, then (4.20) holds. If further Ω is finite, and the step lengths t_k are upper bounded, $\{\mathbf{x}^k\}$ converges to one point in Ω. For Rule P, the same conclusions hold if $t_k^i \leq t_k^{\max}$ and either t_k^i belongs to a compact subset of $(0, \frac{2m_{\varphi_i}}{M_{\nabla_i f}})$, or $t_k^i \equiv t^i \in (0, \frac{2m_{\varphi_i}}{M_{\nabla_i f}})$, $\forall i \in \mathcal{C}$, $\forall k$.

Proof From Lemma 4.4 it follows that in each of the three Rules M, A and P, (4.24) holds. Let $\mathbf{x}^\infty$ be any accumulation point of $\{\mathbf{x}^k\}$ and $\{\mathbf{x}^k\}_{k\in\mathcal{K}}$ any subsequence converging to it. From (4.24) we have that $\{\mathbf{y}^k\}_{k\in\mathcal{K}} \to \mathbf{x}^\infty$. Tracing the proof of [Tse91b, Prop. 4.1] we may conclude that $\mathbf{x}^\infty \in \Omega$. The validity of (4.20) follows from Lemma 2.3. Convergence toward one point in Ω follows from (2.10) and Lemma 2.4. □

Since X is a Cartesian product, step lengths in Rule P may be chosen individually for each $i \in \mathcal{C}$. Cohen [Coh78, Coh80] discusses Rule P in a cyclic decomposition version of the auxiliary problem principle (which is equivalent to the cost approximation methods, in the sense of Theorem 4.1.c). Under somewhat stronger conditions than those imposed in the above theorem (for instance, φ is assumed iteration independent and the cyclic rule is used) he establishes convergence for [cf. (4.26)]

$$\varepsilon_k^i \equiv \varepsilon^i \in \left(0, \frac{2m_{\varphi_i}}{M_{\nabla_i f}}\right), \ \forall i \in \mathcal{C}. \tag{4.50}$$

(This condition can be restated in terms of allowed step lengths t_k^i, cf. Remark 4.3; the corresponding step length bound is the same as for $t_k^i \equiv t^i$, $\forall i$, in the above theorem.)

When executing the decomposition version of the cost approximation method on a sequential computer, its Gauss-Seidel variant (i.e., the cyclic or essentially cyclic rules are used) is expected to converge faster than its Jacobi variant, since the most recent information is utilized when constructing the next subproblem. Cohen [Coh80] gives an intuitive explanation for the faster convergence of the cyclic version in the context of the auxiliary problem principle. If ∇f is Lipschitz continuous and φ strongly convex, then

$$M_{\nabla f} \geq \max_{i\in\mathcal{C}} \{M_{\nabla_i f}\}, \ m_\varphi \leq \min_{i\in\mathcal{C}} \{m_{\varphi_i}\}$$

holds, and, as a result, a larger interval for ε^i is obtained in (4.50) than for ε_k in (4.26). The consequence of this becomes even more clear if we consider the allowed step lengths; from the above we may conclude that longer steps are allowed in the cyclic version.

For the remainder of this section, we assume that we have access to a parallel computer with $|\mathcal{C}|$ independent processors, each one responsible for its own subproblem.[30] The cost approximation method can then be implemented in a way such that all subproblems $[\mathbf{CNLP}_{\varphi_i}]$, $i \in \mathcal{C}$, are solved simultaneously. (This parallel, Jacobi-type, decomposition scheme may be more efficient than its cyclic counterpart due to the concurrency, although it is expected to need more main iterations to yield a given accuracy, cf. the above discussion.) Obviously, the resulting method has the same convergence properties as the

[30] If less than $|\mathcal{C}|$ processors are available, each processor can be assigned to several variable blocks; the preceding results apply to this case also.

original cost approximation method (with the exception that individual step lengths are allowed in Rule P).[31]

To make sure that the sequence of iterates generated by this algorithm agrees with that given by the original (sequential) method, the implementation requires a *synchronization mechanism* [HwB85, HoJ88, BeT89, Gre89], by which the processors are coordinated to operate on the correct data and in the correct sequence. The possible inefficiencies resulting from such a requirement are briefly discussed below.

Assume that Rule P is used, and consider one particular processor, i. In a tightly coupled computer system [HwB85, Sec. 7.1.2] processor i calculates $\mathbf{y}_i^k$ and writes it on the main memory. Before it can resume calculations on a new subproblem, processor i must then wait for all the other processors to finish calculations on their respective subproblems; only then can processor i gain access to the necessary information ($\mathbf{y}_j^k$, $j \neq i$) to create a new subproblem. (The discussion is similar for distributed systems, where instead the solutions $\mathbf{y}_i^k$ are distributed among the processors through some message passing system [BeT89].)

It is obvious that the speed of computation in a synchronized parallel algorithm is governed by the slowest processor (or the most difficult subproblem). The efficiency loss resulting from this is of course highly dependent on the difference in the amount of computations needed for solving the subproblems.[32] The efficiency can be further degraded by memory conflicts (tightly coupled systems [Kun76]) or slow communication channels (distributed systems [BeT89]). Also, if Rule P is replaced by some (inexact) line search, the efficiency is degraded due to an increase in serial computations.

To avoid some of these inefficiencies we may consider applying load balancing strategies [Cyb87, BeT89]. Another way by which we can hope to gain in efficiency is to instead remove the synchronization of the processors, and thereby enable the faster processors to execute more iterations, since they are not required to wait for the most recent results to become available. Because the speed of computations and communications can vary among the processors, and communication delays can be substantial, the processors will, as a result, perform the calculations out of phase with each other. Thus, the advantage of a reduced synchronization penalty is paid for by an increase in interprocessor communications, a use of outdated information that may be counterproductive if certain conditions are not met, and a more difficult convergence detection, see [BeT89]. (Certainly, the convergence analysis also becomes much more complicated.) Recent numerical experiments indicate, however, that the introduction of *asynchronous* computations can substantially enhance the efficiency of parallel iterative methods [ElB89, BeC91, ChZ91].

The convergence of asynchronous iterative methods have been studied by several researchers, see e.g. [ChM69, Mie75a, Mie75b, Bau78, Ber83, BTA84, Tsi87, BeT89, UrD89]. While *totally asynchronous*, or *chaotic*, algorithms allow arbitrarily large communication delays and differences in the frequency of computation of different processors, other algorithms, which are referred to as *partially asynchronous*, can not be established convergent,

[31] The first presentation of a Jacobi algorithm including line searches is [FeM88]; the methods of [ChM88, LMP89] are also highly related to the Jacobi method; see [Pat91b] for a further discussion on Jacobi-type methods for problems over product sets.

[32] Consider for instance large-scale traffic assignment problems, where [$\mathbf{CNLP}_{\varphi_i}$] is a nonlinear single-commodity flow problem [Pat91b, Tse91b]; then, the locations in the network of the origin and destination nodes highly influence the complexity of the subproblem. In the other extreme, we may consider Lagrangean duals of constrained matrix problems [LaS81, ScZ90], where the subproblems [$\mathbf{CNLP}_{\varphi_i}$] are very uniform, analytically solvable one-dimensional problems.

unless an upper bound on these delays is imposed. Examples of totally asynchronous algorithms include dynamic programming methods, fixed point methods, and coordinate search methods for linear and convex network flows [Ber82b, Ber83, Ber86, BeE87, BeE88, BeT89, Ber92]; proposed partially asynchronous algorithms have mainly been of the gradient projection type [Tsi84, TsB86, TBA86, BeT89, TBT90, Tse92]. Partially asynchronous versions of instances of the cost approximation algorithm have been studied only to a limited extent; deflected gradient methods in unconstrained programming [BeT89, Sec. 7.5], gradient projection methods [Tsi84, TsB86, TBA86, BeT89, Tse92], and coordinate ascent methods for strictly convex network flows [TBT90] are the only examples to date.

We end this section by establishing the convergence of a partially asynchronous version of the cost approximation method. The model of partial asynchronism is adapted from [BeT89, Sec. 7.1]. We introduce

(1) a set T^i of times at which $\mathbf{x}_i$ is updated, and

(2) a variable $\tau_j^i(l)$ for each $i, j \in \mathcal{C}$ and $l \in T^i$, measuring the amount by which the information used in an update of $\mathbf{x}_i$ is outdated, satisfying $0 \leq \tau_j^i(l) \leq l$.

To ensure convergence, we must impose a bound on the delays $l - \tau_j^i(l)$.

Assumption 4.1 (Partial asynchronism) *There exists a positive integer B such that*

(a) *for each $i \in \mathcal{C}$ and $l \geq 0$, at least one of the elements of $\{l, l+1, \dots, l+B-1\}$ belongs to T^i, and*

(b) *$l - B < \tau_j^i(l) \leq l$ holds for all $i, j \in \mathcal{C}$ and all $l \geq 0$ belonging to T^i.*

(c) *$\tau_i^i(l) = l$, for all $i \in \mathcal{C}$ and $l \in T^i$.*

Parts **(a)** and **(b)** of the assumption state that both communication delays and processor idle times are bounded; part **(c)** states that processor i always uses the most recent value of its own component $\mathbf{x}_i$ of $\mathbf{x}$. For further discussions on the assumption, we refer to [BeT89, TBT90]; we only remark that they are often easily enforced in a practical implementation.

The iterate $\mathbf{x}(l)$ is defined by the vector of $\mathbf{x}_i(l)$, where $\mathbf{x}_i(l)$ is updated by processor i according to

$$\mathbf{x}_i(l+1) = \mathbf{x}_i(l) + t\left(\mathbf{y}_i(l) - \mathbf{x}_i(l)\right), \ l \in T^i, \tag{4.51}$$

that is, we assume that Rule P is used. We also define $\mathbf{d}_i(l) \stackrel{\text{def}}{=} \mathbf{y}_i(l) - \mathbf{x}_i(l) = \mathbf{0}$, for $l \notin T^i$. Processor i has knowledge of a possibly outdated version of $\mathbf{x}(l)$; we let

$$\mathbf{x}^i(l)^{\mathrm{T}} = \left[\mathbf{x}_1\left(\tau_1^i(l)\right)^{\mathrm{T}}, \dots, \mathbf{x}_{|\mathcal{C}|}\left(\tau_{|\mathcal{C}|}^i(l)\right)^{\mathrm{T}}\right]$$

denote this vector.

Theorem 4.15 (Convergence of partially asynchronous cost approximation algorithms) [33] *Let X be defined by (4.44). Let $\varphi : X \times X \mapsto \Re$ be given by (4.45). Further, let f be lower bounded on X, and assume that $\nabla_i f$ is Lipschitz continuous on X_i and the function*

$\varphi_i : X_i \times X \mapsto \Re$ strongly convex on X_i with respect to $\mathbf{x}_i$ and $\nabla\varphi_i$ Lipschitz continuous on X_i, for all $i \in \mathcal{C}$. Let $m_\varphi = \min_{i\in\mathcal{C}} m_{\varphi_i}$. In the above partially asynchronous algorithm, assume that in Rule P,

$$t \in \left(0, \frac{m_\varphi}{M_{\nabla f}\left(\frac{1}{2} + (|\mathcal{C}| + 1)\, B\right)}\right), \tag{4.52}$$

and $t \leq 1$. Then, from any starting point $\mathbf{x}(0) \in X$, any accumulation point $\mathbf{x}^\infty$ of the sequence $\{\mathbf{x}(l)\}$ lies in Ω. Furthermore, if $\{\mathbf{x}(l)\}$ is bounded and Ω is nonempty, then (4.20) holds. If further Ω is finite, then $\{\mathbf{x}(l)\}$ converges to one point in Ω.

Proof The proof is similar to that of [BeT89, Prop. 7.5.3] for the special case of the partially asynchronous Goldstein-Levitin-Polyak gradient projection algorithm.

From (4.21), the strong convexity of the φ_i and the Lipschitz continuity of $\nabla_i f$, we obtain that

$$\begin{aligned}
f(\mathbf{x}(l+1)) - f(\mathbf{x}(l)) &\leq t\sum_{i\in\mathcal{C}} \mathbf{d}_i(l)^{\mathrm{T}}\nabla_i f(\mathbf{x}(l)) + \frac{M_{\nabla f}}{2}t^2\|\mathbf{d}(l)\|^2 \\
&= t\sum_{i\in\mathcal{C}} \mathbf{d}_i(l)^{\mathrm{T}}\nabla_i f\left(\mathbf{x}^i(l)\right) + \frac{M_{\nabla f}}{2}t^2\|\mathbf{d}(l)\|^2 \\
&\quad + t\sum_{i\in\mathcal{C}} \mathbf{d}_i(l)^{\mathrm{T}}\left(\nabla_i f(\mathbf{x}(l)) - \nabla_i f\left(\mathbf{x}^i(l)\right)\right) \\
&\leq -t m_\varphi \sum_{i\in\mathcal{C}}\|\mathbf{d}_i(l)\|^2 + \frac{M_{\nabla f}}{2}t^2\|\mathbf{d}(l)\|^2 \\
&\quad + M_{\nabla f}t\sum_{i\in\mathcal{C}}\|\mathbf{d}_i(l)\|\left\|\mathbf{x}(l) - \mathbf{x}^i(l)\right\|.
\end{aligned} \tag{4.53}$$

Next, we bound $\|\mathbf{x}(l) - \mathbf{x}^i(l)\|$. From Rule P and Assumption 4.1.b,

$$\left\|\mathbf{x}^i_j(l) - \mathbf{x}_j(l)\right\| = \left\|\mathbf{x}_j\left(\tau^i_j(l)\right) - \mathbf{x}_j(l)\right\| = t\left\|\sum_{\tau=\tau^i_j(l)}^{l-1}\mathbf{d}_j(\tau)\right\| \leq t\sum_{\tau=l-B}^{l-1}\|\mathbf{d}_j(\tau)\|.$$

Using the triangle inequality, we obtain

$$\left\|\mathbf{x}^i(l) - \mathbf{x}(l)\right\| \leq t\sum_{\tau=l-B}^{l-1}\|\mathbf{d}(\tau)\|. \tag{4.54}$$

[33]This result can be slightly strengthened, if we take into account the coupling among the sets X_i in the objective [BeT89, p. 533]. We refine the Lipschitz continuity assumption of ∇f, and assume that there is a constant M with

$$\|\nabla_i f(\mathbf{x}) - \nabla_i f(\mathbf{y})\| \leq M\|\mathbf{x} - \mathbf{y}\|, \ \forall i, \ \forall \mathbf{x}, \mathbf{y} \in X \text{ with } \mathbf{x}_i = \mathbf{y}_i.$$

If f is separable with respect to $\mathbf{x}_i$, then $M = 0$. Hence, a small value of M indicates that the problem is weakly coupled. The interval of allowed step lengths now becomes

$$t \in \left(0, \frac{m_\varphi}{\frac{M_{\nabla f}}{2} + (|\mathcal{C}| + 1)\, MB}\right).$$

Using (4.54) in (4.53), we obtain

$$f(\mathbf{x}(l+1))-f(\mathbf{x}(l)) \leq -t\left(m_{\varphi}-\frac{M_{\nabla f}}{2}t\right)\|\mathbf{d}(l)\|^2 + M_{\nabla f}t^2\sum_{i\in\mathcal{C}}\|\mathbf{d}_i(l)\|\sum_{\tau=l-B}^{l-1}\|\mathbf{d}(\tau)\|.$$

We then use the inequality

$$\|\mathbf{d}_i(l)\|\,\|\mathbf{d}(\tau)\| \leq \|\mathbf{d}_i(l)\|^2 + \|\mathbf{d}(\tau)\|^2,$$

to obtain

$$\begin{aligned} f(\mathbf{x}(l+1)) - f(\mathbf{x}(l)) &\leq -t\left(m_{\varphi}-\frac{M_{\nabla f}}{2}t\right)\|\mathbf{d}(l)\|^2 \\ &\quad + M_{\nabla f}t^2\sum_{i\in\mathcal{C}}\sum_{\tau=l-B}^{l-1}\left(\|\mathbf{d}_i(l)\|^2 + \|\mathbf{d}(\tau)\|^2\right) \\ &= -t\left(m_{\varphi}-\frac{M_{\nabla f}}{2}t\right)\|\mathbf{d}(l)\|^2 \\ &\quad + M_{\nabla f}t^2\left(B\|\mathbf{d}(l)\|^2 + |\mathcal{C}|\sum_{\tau=l-B}^{l-1}\|\mathbf{d}(\tau)\|^2\right) \\ &= -t\left(m_{\varphi}-\frac{M_{\nabla f}}{2}t - M_{\nabla f}Bt\right)\|\mathbf{d}(l)\|^2 \\ &\quad + |\mathcal{C}|\,M_{\nabla f}t^2\sum_{\tau=l-B}^{l-1}\|\mathbf{d}(\tau)\|^2. \end{aligned} \tag{4.55}$$

Adding (4.55) over all l yields

$$\begin{aligned} f(\mathbf{x}(l+1)) - f(\mathbf{x}(0)) &\leq -t\left(m_{\varphi}-\frac{M_{\nabla f}}{2}t - M_{\nabla f}Bt - |\mathcal{C}|\,M_{\nabla f}Bt\right)\sum_{\tau=0}^{l}\|\mathbf{d}(\tau)\|^2 \\ &\quad + |\mathcal{C}|\,M_{\nabla f}Bt^2\sum_{\tau=0}^{l}\|\mathbf{d}(\tau)\|^2 \\ &= -t\left(m_{\varphi}-\frac{M_{\nabla f}}{2}t - M_{\nabla f}Bt - |\mathcal{C}|\,M_{\nabla f}Bt\right)\sum_{\tau=0}^{l}\|\mathbf{d}(\tau)\|^2. \end{aligned}$$

If (4.52) holds, the right hand side of the inequality is negative and, since f is lower bounded on X, we must have that

$$\{\mathbf{d}(l)\} \to \mathbf{0}.$$

We also have from (4.51) that

$$\{\mathbf{x}(l+1) - \mathbf{x}(l)\} \to \mathbf{0}, \tag{4.56}$$

and from (4.54) that

$$\left\{\mathbf{x}^i(l) - \mathbf{x}(l)\right\} \to \mathbf{0}. \tag{4.57}$$

Let $\mathbf{x}^\infty$ be an accumulation point of the sequence $\{\mathbf{x}(l)\}$, corresponding to a convergent subsequence $\{l_k\}$, and let τ_k be such that $|l_k - \tau_k| \leq B$ and $\tau_k \in T^i$. Then, (4.56) and (4.57) imply that $\{\mathbf{x}(\tau_k)\} \to \mathbf{x}^\infty$ and $\{\mathbf{x}^i(\tau_k)\} \to \mathbf{x}^\infty$ holds. From (4.51) it finally holds that $\mathbf{y}_i^\infty = \mathbf{x}_i^\infty$, i.e., (4.1) holds. The last statements follow from Lemma 2.3, and (2.10) and Lemma 2.4, respectively. □

Remark 4.13 Interesting observations can be made regarding the relations among the maximal step length, the allowed amount of asynchronism, the number of processors involved in the calculations, and the properties of the given problem, that is brought forward in the expression (4.52). Some observations are made in [BeT89, Sec. 7.5.2] for the special case of Goldstein-Levitin-Polyak gradient projection; their conclusions apply also to this more general case.

Remark 4.14 The increasing degree of parallelism and asynchronism from the sequential, Gauss-Seidel-type schemes to the partially asynchronous algorithms correspond to decreasing intervals of allowed step lengths. [Compare e.g. (4.52) with the allowed steps in the sequential algorithm (Theorem 4.3.b).] This is a consequence of the decreasing quality of the step directions, resulting from the usage of more outdated information.

4.16 Summary and further research

In this chapter we have studied the class of cost approximation methods for constrained, differentiable programs. An algorithmic equivalence result was established, linking the cost approximation concept to the algorithm classes of Dafermos, Cohen, Migdalas and Tseng (Theorem 4.1). The results obtained in Sections 4.6 and 4.7, however, strengthen their analyses in several directions; most notably, in neither of the mentioned analyses the basic Frank-Wolfe and Newton methods are included.

Through examples, such as regularization methods, Evans' algorithm, and the decomposition methods of Section 4.15, the class of cost approximation methods was shown to be able to effectively adapt to problem structures, both in the objective (e.g., strictly convexify a nonconvex objective) and in the constraints (e.g., induce separability in nonseparable problems). For problems over product sets, algorithms with an increasing degree of parallelism and asynchronism were established. In many instances of the cost approximation method unit steps are used, in particular for regularization methods; the analysis made in this chapter shows that the conditions for convergence can be relaxed considerably at the cost of an additional (but possibly inexact) line search step.

The convergence rate analysis made for [**CNLP**] is limited, compared to that made for unconstrained programs. For example, linear convergence is established only for quadratic functions φ. To strengthen these results, one could impose growth conditions on the function defined in (A.26); such conditions are used by Dunn in convergence rate analyses of the Frank-Wolfe method, gradient projection and Newton methods [Dun79, Dun80a, Dun80b, Dun81, Dun87]. Another possibility is to find conditions under which the linear convergence rate analysis in [LuT92d] can be used. This approach seems more straightforward, and some preliminary results are already available (see the discussions in Section 4.12). Also, using the results of Theorem 4.6, stronger convergence results are available from unconstrained programs (cf. the discussion following the theorem). A fourth alternative is to invoke conditions which relate $\{\varphi^k\}$ to f in the convergence rate analysis. This could enable us to obtain stronger results for Newton and regularization type methods.

It would also be interesting to find conditions under which subproblems can be solved inexactly, while still ensuring fast convergence, and the convergence rates (theoretical and practical) that result from different choices of truncation strategy. Such analyses have been made for Newton's method (see the references in Section 4.10), but should also be possible in a more general setting.

Finite convergence was established for problems where Condition **(c)** or **(d)** holds. (For instance, the class of cost approximation methods exhibits finite convergence for problems with polyhedral objectives (e.g., linear programs), since such problems satisfy Condition **(c)**. The issue of whether or not it is possible to extend this result to problems with a mildly nonlinear (or nonpolyhedral) objective is highly related to nonlinear perturbation results for linear [PoT74, TiA77, MaM79, Man84] and convex programs [FeM91], as well as finiteness results in exact penalty methods [HaM79, Fle87, Bur91].) However, these conditions imply that there is an optimal extreme point, and this does not hold in many applications. It would be of interest to investigate whether or not it is possible to enlarge the set of problems for which the cost approximation methods exhibit finite convergence.

The identification property of gradient projection methods has led to the development of combined algorithms, where a gradient projection method is used to predict the active constraint set $\mathcal{I}(\mathbf{x}^*)$, and an unconstrained programming algorithm is used to explore the face of X described by a current set of active constraints, see e.g. [Ber76b, Ber82c, CaM87, CGT88a, Mor88, Toi88, Mor89, MoT89, BMT90]. An advantage of projection methods is that, in contrast to standard active set type methods, they can add and drop many constraints from the active set in each iteration; this advantage is supported by some numerical experiments on large-scale problems with simple constraints [CGT88b, MoT89]. It should be possible to devise algorithms, where gradient projection is replaced by the more general cost approximation concept (cf. Theorem 4.6); the choice of φ^k to match problem structure offers possibilities for a further gain in practical efficiency of such combined schemes. It would also be of interest to extend the finite identification property to nondifferentiable constrained problems, and to study extensions of the obtained results to nonconvex problems, where the regularity conditions are replaced by local versions.

It is well known that the Frank-Wolfe algorithm often exhibits a slow convergence, due to the linearity of the subproblem. This drawback is, to some extent, inherited by the simplicial decomposition methods [Hol74, vHo75, vHo77, HLV85, HLV87, LaP92, VeH93], especially when column dropping is applied on problems with an optimal face of high dimension (see [HLV87] for examples). Consider a modification of the basic simplicial decomposition scheme, where the linear (Frank-Wolfe) subproblem is replaced by the subproblem of the cost approximation algorithm, given some function φ^k. This modified method extends both the original simplicial decomposition method (which is obtained when $\varphi^k \equiv 0$, $\forall k$) and the original cost approximation method (which is obtained if the number of retained points in the master problems is always one).[34] The use of a nonlinear cost approximation subproblem may undoubtedly enhance the performance of a simplicial decomposition scheme. (The increase in subproblem computations can be partly circumvented by choosing φ^k to exploit the problem structure, and applying truncation strategies.) If this scheme includes column dropping strategies, we obtain a cost approximation algorithm which utilizes master problems only when it is necessary in order to make sufficient progress. The theoretical and practical properties of the above described combination of cost approximation and simplicial decomposition are subjects of ongoing research efforts, including applications to other instances of [**GVIP**].

[34] The idea of using non-vertex points in master problems of decomposition schemes is suggested in [KiN91] for the Dantzig-Wolfe method in linear programming.

Chapter 5

Variational inequalities

"You see, Barley, a lot of people around here would attach a heap of suspicion to a gap like that," Brady said.

JOHN LE CARRÉ
The Russia House (1989)

In this chapter, we study cost approximation methods for the variational inequality problem. In [**GVIP**], let $\mathbf{U} \equiv \mathbf{0}$. The problem then is to find an $\mathbf{x}^* \in X$ such that

$$[\textbf{VIP}] \qquad \mathbf{F}(\mathbf{x}^*)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^*) \geq 0, \ \forall \mathbf{x} \in X,$$

where $\mathbf{F} : X \mapsto \Re^n$ is continuous on the nonempty, closed and convex set X. (These properties of X and $\mathbf{F}$ are standing assumptions throughout this chapter.) We let Ω denote the set of solutions to [**VIP**]. (Sufficient conditions for the existence of a solution to [**VIP**] can be found e.g. in [KiS80, GLT81, HaP90b].)

It is well known (see e.g. [BaS79, Th. 4.2]) that the problem [**VIP**] defines the first-order *necessary* optimality conditions for the optimization problem

$$[\textbf{CNLP}] \qquad \min_{\mathbf{x} \in X} f(\mathbf{x}) \stackrel{\text{def}}{=} \oint_{\mathbf{0}}^{\mathbf{x}} \mathbf{F}(\mathbf{s})^{\mathrm{T}} d\mathbf{s},$$

whenever $\mathbf{F}$ is the gradient of a function $f : X \mapsto \Re$. If further $\mathbf{F}$ is in C^1 on X, this property is equivalent to the Jacobian, $\nabla \mathbf{F}$, of $\mathbf{F}$ being symmetric on X [OrR70, Th. 4.1.6], see also [Car77].

If, however, $\mathbf{F}$ is not the gradient of any function, the integral in [**CNLP**] is not unambiguously defined, and [**VIP**] can in this case not be converted into an equivalent optimization problem of this form. (In analogy with the above symmetry characterization, we then say that the problem is *asymmetric*.)

Traditionally, methods proposed for the solution of [**VIP**] have been based on the concepts of successive approximation and contractive mappings. It is the purpose of this chapter to demonstrate that the subproblem in the cost approximation methods induces merit functions ψ for asymmetric variational inequalities, to investigate their properties, and devise descent methods based on them. Interesting connections with earlier proposed merit functions and descent methods, as well as with the successive approximation methods will also be established.

5.1 The cost approximation algorithm

The cost approximation algorithm

A sequence $\{\mathbf{\Phi}^k\}$ of monotone cost approximating mappings and a merit function ψ are assumed to be given.

Step 0 (*Initial guess*) Choose an initial point $\mathbf{x}^0 \in X$, and let $k = 0$.

Step 1 (*Search direction generation*) Find $\mathbf{y}^k \in X$ such that

$$[\mathbf{VIP}^k_{\mathbf{\Phi}^k}] \qquad \left[\mathbf{\Phi}^k\left(\mathbf{y}^k\right) + \mathbf{F}\left(\mathbf{x}^k\right) - \mathbf{\Phi}^k\left(\mathbf{x}^k\right)\right]^{\mathrm{T}}\left(\mathbf{y} - \mathbf{y}^k\right) \geq 0, \ \forall \mathbf{y} \in X.$$

The resulting search direction is $\mathbf{d}^k = \mathbf{y}^k - \mathbf{x}^k$.

Step 2 (*Convergence check*) If $\mathbf{x}^k$ solves $[\mathbf{VIP}^k_{\mathbf{\Phi}^k}] \to$ Stop ($\mathbf{x}^k \in \Omega$). Otherwise, continue.

Step 3 (*Line search*) Find a step length, t_k, which solves the one-dimensional problem

$$t_k \in \arg \min_{\substack{t \geq 0 \\ \mathbf{x}^k + t\mathbf{d}^k \in X}} \psi\left(\mathbf{x}^k + t\mathbf{d}^k\right).$$

Step 4 (*Update*) Let $\mathbf{x}^{k+1} = \mathbf{x}^k + t_k\mathbf{d}^k$, and $k := k + 1$.

Step 5 (*Convergence check*) If $\mathbf{x}^k$ is acceptable as a solution $\to$ Stop. Otherwise, go to Step 1.

If $\mathbf{\Phi}^k \equiv \nabla\varphi^k$ for some convex function φ^k in C^1 on X, $[\mathbf{VIP}^k_{\mathbf{\Phi}^k}]$ reduces to the necessary and sufficient optimality conditions for

$$[\mathbf{CNLP}^k_{\varphi^k}] \qquad \min_{\mathbf{x} \in X} \left\{\varphi^k(\mathbf{x}) + f\left(\mathbf{x}^k\right) - \varphi^k\left(\mathbf{x}^k\right) + \left[\mathbf{F}\left(\mathbf{x}^k\right) - \nabla\varphi^k\left(\mathbf{x}^k\right)\right]^{\mathrm{T}}\left(\mathbf{x} - \mathbf{x}^k\right)\right\}$$

(see also Section 4.2).

For general asymmetric variational inequalities, due to the nonexistence of an unambiguous integral in [**CNLP**] above, there exists no obvious choice of merit function to be used in the line search in Step 3.[1] In a vast majority of the iterative methods developed

[1] If the problem [**VIP**] arises from a saddle point problem [vNe28, Dan67, Roc70a, DeM74a, BeT89], i.e., a problem in which, for a given function $L : X \times Y \mapsto \Re$, a pair $(\mathbf{x}^*, \mathbf{y}^*) \in X \times Y$ is sought such that

$$L(\mathbf{x}^*, \mathbf{y}) \leq L(\mathbf{x}^*, \mathbf{y}^*) \leq L(\mathbf{x}, \mathbf{y}^*), \ \forall \mathbf{x} \in X, \ \forall \mathbf{y} \in Y,$$

then a natural merit function is $\psi(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - g(\mathbf{y})$, where

$$f(\mathbf{x}) = \sup_{\mathbf{y} \in Y} L(\mathbf{x}, \mathbf{y}), \qquad g(\mathbf{y}) = \inf_{\mathbf{x} \in Y} L(\mathbf{x}, \mathbf{y}).$$

The approach we take in this chapter to construct a merit function is to reformulate [**VIP**] as an equivalent saddle point *type* problem.

for [**VIP**], convergence is based on a decrease in the (artificial) merit function

$$\psi(\mathbf{x}) = \min_{\mathbf{x}^* \in \Omega} \|\mathbf{x} - \mathbf{x}^*\|, \tag{5.1}$$

where $\|\cdot\|$ is some appropriate vector norm. Clearly, a line search can not be made with respect to this merit function. Convergence in this measure is ensured, however, by showing that the underlying algorithmic map, $\mathbf{x}^{k+1} \in A(\mathbf{x}^k)$, satisfies the two conditions:

(1) the algorithmic map is a fixed point map, i.e.,

$$\mathbf{x} \in \Omega \Leftrightarrow \mathbf{x} \in A(\mathbf{x}), \tag{5.2}$$

(2) A is contractive with respect to the norm defined in (5.1) [see Definition A.5.c].

It will subsequently be shown that a large number of well known methods for [**VIP**] in this class can be identified by $A \equiv Y$ and $t_k \equiv 1$, where $\mathbf{x} \mapsto Y(\mathbf{x})$ is the map that defines the set of optimal solutions to [$\mathbf{VIP}^k_{\mathbf{\Phi}^k}$], that is, they can be seen as methods where the algorithmic map corresponds to a cost approximation subproblem.

The validity of (5.2) for $A \equiv Y$ is established in the following lemma, extending the result of Lemma 4.1.a to variational inequalities.

Lemma 5.1 (A fixed point characterization of Ω) *Let $\mathbf{\Phi} : X \mapsto \Re^n$ be continuous on X. Then,*

$$\mathbf{x} \in \Omega \Leftrightarrow \mathbf{x} \in Y(\mathbf{x}).$$

Proof $\mathbf{x} \in \Omega \Leftrightarrow -\mathbf{F}(\mathbf{x}) \in N_X(\mathbf{x}) \Leftrightarrow -\mathbf{\Phi}(\mathbf{x}) - \mathbf{F}(\mathbf{x}) + \mathbf{\Phi}(\mathbf{x}) \in N_X(\mathbf{x}) \Leftrightarrow \mathbf{x} \in Y(\mathbf{x})$. □

This fixed point characterization extends well known characterizations of solutions to [**VIP**]. Choosing $\mathbf{\Phi} \equiv \mathbf{0}$ yields $Y(\mathbf{x}) = \arg\min_{\mathbf{y} \in X} \mathbf{F}(\mathbf{x})^{\mathrm{T}}(\mathbf{y} - \mathbf{x})$, i.e., $\mathbf{x} \mapsto Y(\mathbf{x})$ is the point-to-set map corresponding to a Frank-Wolfe type subproblem. This special case of Lemma 5.1 is established in [ZPP69b], see also [Hea82, Evt85, Mar85]. Choosing $\mathbf{\Phi}(\mathbf{x}) = \mathbf{Q}\mathbf{x}$, where $\mathbf{Q}$ is symmetric and positive definite on X, yields $Y(\mathbf{x}) = \{\mathbf{y}(\mathbf{x})\}$, where $\mathbf{y}(\mathbf{x}) = P_X^{\mathbf{Q}}(\mathbf{x} - \mathbf{Q}^{-1}\mathbf{F}(\mathbf{x}))$. This special case is established in [Eav71], see also [Daf80].

5.2 Instances of successive cost approximation algorithms

We show in this section that the successive approximation version of the class of cost approximation methods, i.e., where $t_k \equiv 1$, includes many well known methods for [**VIP**] as special cases.[2]

We distinguish between linear and nonlinear approximation methods.

[2]Successive approximation versions of all the methods discussed in Section 4.3 have been extended to variational inequality problems. It is easily shown that, by replacing ∇f with $\mathbf{F}$ and formulating subproblems in terms of variational inequalities, these methods are instances of the successive cost approximation methods for [**VIP**].

In linear approximation methods, the map $\mathbf{F}$ is approximated by an affine map,

$$\mathbf{F}^k(\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{F}\left(\mathbf{x}^k\right) + \mathbf{A}_k\left(\mathbf{x} - \mathbf{x}^k\right), \tag{5.3}$$

where $\mathbf{A}_k$ is a positive semidefinite matrix in $\Re^{n\times n}$. Choosing $\mathbf{\Phi}^k(\mathbf{x}) = \mathbf{A}_k\mathbf{x}$ yields that the map defining $[\mathbf{VIP}^k_{\mathbf{\Phi}^k}]$ has the form (5.3), i.e., the class of linear approximation methods are cost approximation methods where the maps $\mathbf{\Phi}^k$ are affine. Whenever $\mathbf{A}_k$ is symmetric, the resulting variational inequality $[\mathbf{VIP}^k_{\mathbf{\Phi}^k}]$ reduces to a convex, quadratic minimization problem. In Table 5.1 we list some well known examples of linear approximation methods, which are instances of the class of cost approximation methods. (For $\mathbf{F}$ in C^1 on X, we let its Jacobian at $\mathbf{x} \in X$ be decomposed as $\nabla\mathbf{F}(\mathbf{x}) = \mathbf{L}(\mathbf{x}) + \mathbf{D}(\mathbf{x}) + \mathbf{U}(\mathbf{x})$, where $\mathbf{D}(\mathbf{x})$ is the diagonal part of $\nabla\mathbf{F}(\mathbf{x})$, and $\mathbf{L}(\mathbf{x})$ and $\mathbf{U}(\mathbf{x})$ its lower and upper triangular part, respectively. Further, we let $0 < \omega < 2$, and $\mathbf{Q}$ be a symmetric and positive definite matrix in $\Re^{n\times n}$.)

Choice of $\mathbf{A}_k$	Resulting method	Basic references
$\nabla\mathbf{F}\left(\mathbf{x}^k\right)$	Newton	[Eav78a, Jos79a]
$\approx \nabla\mathbf{F}\left(\mathbf{x}^k\right)$	Quasi-Newton	[Jos79b]
$\frac{1}{2}\left(\nabla\mathbf{F}\left(\mathbf{x}^k\right) + \nabla\mathbf{F}\left(\mathbf{x}^k\right)^{\mathrm{T}}\right)$	Symmetrized Newton	[Daf80, HaM87]
$\mathbf{L}\left(\mathbf{x}^k\right) + \mathbf{D}\left(\mathbf{x}^k\right)/\omega$ or $\mathbf{U}\left(\mathbf{x}^k\right) + \mathbf{D}\left(\mathbf{x}^k\right)/\omega$	SOR; Linearized Gauss-Seidel ($\omega = 1$)	[OrR70, PaC82]
$\mathbf{D}\left(\mathbf{x}^k\right)$	Linearized Jacobi	[PaC82]
$\mathbf{Q}$	Projection	[Sib70, Daf80]

Table 5.1: Examples of linear approximation methods

In nonlinear approximation methods, the nonlinearity of $\mathbf{F}$ is preserved in the approximation, while (typically) separability or symmetry is induced to obtain a variational inequality which is more easily solved.[3]

Let $\mathbf{\Phi}^k$ be of the form $\mathbf{\Phi}(\mathbf{x}, \mathbf{y})$, continuous on $X \times X$. Then if X is of the form (4.44), we obtain a Jacobi method [Ahn79, AhH82, PaC82, Pan85, Daf83] for [**VIP**] by choosing

$$\mathbf{\Phi}_i(\mathbf{x}, \mathbf{y}) = \mathbf{F}_i\left(\mathbf{y}_{i_-}, \mathbf{x}_i, \mathbf{y}_{i_+}\right), \ i \in \mathcal{C}, \tag{5.4}$$

and a Gauss-Seidel scheme [PaC82, BeT89, Pan85] by the choice

$$\mathbf{\Phi}_i(\mathbf{x}, \mathbf{y}) = \mathbf{F}_i\left(\mathbf{x}_{i_-}, \mathbf{x}_i, \mathbf{y}_{i_+}\right), \ i \in \mathcal{C} \tag{5.5}$$

(compare (2.19) and (2.20), respectively).

[3] Such methods are of particular interest in the many applications of [**VIP**] where the feasible set is a Cartesian product. Applications include various competitive equilibrium models, such as traffic equilibrium [Daf80, Pat90, Nag93], spatial economic equilibrium [FaP80, ToF83, DaN87], and Nash equilibrium [Kar69b, GaM80, Coh87]; see also [Pan85, BeT89, HaP90b].

Regularization methods are obtained by choosing $\mathbf{\Phi}^k(\mathbf{x}) = \mathbf{F}(\mathbf{x}) + \frac{1}{\gamma_k}\mathbf{R}(\mathbf{x}, \mathbf{x}^k)$, where $\gamma_k > 0$ and $\mathbf{R}(\mathbf{x}, \mathbf{y})$ is strongly monotone in $\mathbf{x}$; the corresponding subproblem map becomes $\mathbf{F} + \frac{1}{\gamma_k}\mathbf{R}(\cdot, \mathbf{x}^k)$, see [Bro66, BaP74] and also Section 4.3. Special cases of regularization methods are the proximal point algorithms [Mar70, Roc76b], obtained from choosing $\mathbf{R}(\mathbf{x}, \mathbf{y}) = \mathbf{x}$, and splitting methods [Gab79, Gab83, Tse91a].

5.3 Algorithmic equivalence results

In this section, we derive algorithmic equivalence results among the cost approximation method and the method classes of Dafermos [Daf83], Cohen [Coh88], Wu *et al.* [WFM90], and Noor [Noo92]. We first give a short description of the respective subproblems.

The algorithm class of Cohen

Let $\varphi : X \mapsto \Re$ be a strongly convex function in C^1 on X, and let $\varepsilon_k > 0$. Given $\mathbf{x}^k \in X$, the following subproblem determines $\mathbf{x}^{k+1}$ in the *auxiliary problem principle.*

$$\min_{\mathbf{y} \in X} \left\{ \varphi(\mathbf{y}) + \left[\varepsilon_k \mathbf{F}\left(\mathbf{x}^k\right) - \nabla\varphi\left(\mathbf{x}^k\right)\right]^{\mathrm{T}} \mathbf{y} \right\} \tag{5.6}$$

The algorithm class of Noor

Let $\mathbf{\Phi} : X \mapsto \Re^n$ be a strongly monotone, continuous map, and let $\varepsilon > 0$. Given $\mathbf{x}^k \in X$, the following subproblem determines $\mathbf{x}^{k+1}$.

$$\left[\mathbf{\Phi}\left(\mathbf{x}^{k+1}\right) + \varepsilon\mathbf{F}\left(\mathbf{x}^k\right) - \mathbf{\Phi}\left(\mathbf{x}^k\right)\right]^{\mathrm{T}} \left(\mathbf{x} - \mathbf{x}^{k+1}\right) \geq 0, \ \forall \mathbf{x} \in X \tag{5.7}$$

The algorithm class of Dafermos

Let $\mathbf{g} : X \times X \mapsto \Re^n$ be a continuous mapping satisfying

(1) $\mathbf{g}(\mathbf{x}, \mathbf{x}) = \mathbf{F}(\mathbf{x}), \ \forall \mathbf{x} \in X$,

(2) $\nabla_{\mathbf{x}} \mathbf{g}(\mathbf{x}, \mathbf{y})$ is symmetric and positive definite for any fixed $\mathbf{x}, \mathbf{y} \in X$.

Given $\mathbf{x}^k \in X$, the point $\mathbf{x}^{k+1}$ is obtained by solving the variational inequality subproblem

$$\mathbf{g}\left(\mathbf{x}^{k+1}, \mathbf{x}^k\right)^{\mathrm{T}} \left(\mathbf{x} - \mathbf{x}^{k+1}\right) \geq 0, \ \forall \mathbf{x} \in X. \tag{5.8}$$

The algorithm class of Wu et al.

Let $h : X \times X \mapsto \Re$ be a continuous function of the form $h(\mathbf{x}, \mathbf{y})$, strictly convex and in C^1 on X with respect to $\mathbf{x}$. Assume further that $\nabla_{\mathbf{x}} h(\mathbf{x}, \mathbf{x}) \stackrel{\text{def}}{=} \nabla_{\mathbf{x}} h(\mathbf{x}, \mathbf{y})\,|_{\mathbf{y}=\mathbf{x}} = \mathbf{0}$. Given $\mathbf{x}^k \in X$, the subproblem is

$$\min_{\mathbf{y} \in X} \left\{ \mathbf{F}\left(\mathbf{x}^k\right)^{\mathrm{T}} \mathbf{y} + h\left(\mathbf{y}, \mathbf{x}^k\right) \right\}. \tag{5.9}$$

We note that (5.6) is equivalent to (4.7), with the exception that the function φ does not vary over iterations in the former case; the subproblem (5.7) is equivalent to (5.6), if $\varepsilon_k \equiv \varepsilon$. The subproblem (5.8) is equivalent to (4.9), and, finally, subproblem (5.9) is equivalent to (4.11). Based on these observations, we give below an algorithmic equivalence theorem for variational inequality methods, extending Theorem 4.1.

Theorem 5.1 (Algorithmic equivalence results)

(a) *If $\{\boldsymbol{\Phi}^k\}$ is of the form $\boldsymbol{\Phi} : X \times X \mapsto \Re^n$, then the resulting subproblem $[\mathbf{VIP}^k_{\boldsymbol{\Phi}^k}]$ is equivalent to the subproblem (5.8) of [Daf83].*

(b) *If $\{\boldsymbol{\Phi}^k\}$ is of the form $\boldsymbol{\Phi} : X \mapsto \Re^n$, then the resulting subproblem $[\mathbf{VIP}^k_{\boldsymbol{\Phi}^k}]$ is equivalent to the subproblem (5.7) of [Noo92].*

(c) *If $\{\varphi^k\}$ is of the form $\varphi^k(\mathbf{x}) = \varepsilon_k \varphi(\mathbf{x})$, where $\varepsilon_k > 0$ and $\varphi : X \mapsto \Re$, then the resulting subproblem $[\mathbf{CNLP}^k_{\varphi^k}]$ is equivalent to the subproblem (5.6) of [Coh88].*

(d) *If $\{\varphi^k\}$ is of the form $\varphi : X \times X \mapsto \Re$, then the resulting subproblem $[\mathbf{CNLP}^k_{\varphi^k}]$ is equivalent to the subproblem (5.9) of [WFM90].*

Proof

(a) See [Pat91b].

(b) See [LaP90].

(c) See [LaP90].

(d) See [Pat91b]. $\square$

The relations among the subproblems of the above described method classes are given in Figure 5.1. From this figure it is clear that the subproblem of the general cost approximation algorithm defines a generalization of those of *all* the method classes discussed in this section.

$$
\begin{array}{ccccc}
[\text{Coh88}] & \Longleftarrow & [\mathbf{CNLP}^k_{\varphi^k}] & \Longleftarrow & [\mathbf{VIP}^k_{\boldsymbol{\Phi}^k}] \\
\Downarrow & & \Downarrow & & \Downarrow \\
[\text{LaP90}] & \Longleftarrow & [\text{WFM90}] & \Longleftarrow & \begin{bmatrix} \text{Daf83} \\ \text{Pat91b} \end{bmatrix} \\
 & & & & \Downarrow \\
 & & & & [\text{Noo92}]
\end{array}
$$

Figure 5.1: Relations among method classes

5.4 Convergence of successive cost approximation algorithms

Here we will state global convergence results for the successive cost approximation methods; we draw on the algorithmic equivalence results of Theorem 5.1.

Theorem 5.2 (Convergence of successive cost approximation algorithms)

(a) *Let* $\mathbf{F}$ *be strongly monotone and Lipschitz continuous on* X, *and let* $\varphi : X \mapsto \Re$ *be a strongly convex function in* C^1 *on* X. *If* φ *is chosen such that*

$$2m_\varphi > \frac{M_{\mathbf{F}}^2}{m_{\mathbf{F}}}, \tag{5.10}$$

then, with $t_k \equiv 1, \{\mathbf{x}^k\} \to \mathbf{x}^*$, *where* $\mathbf{x}^*$ *is the unique solution of* [**VIP**].

(b) *Let* X *be bounded and assume that* $\mathbf{F}$ *is in* C^1 *on* X. *Let* $\mathbf{\Phi} : X \times X \mapsto \Re^n$ *be a map in* C^1 *on* $X \times X$ *of the form* $\mathbf{\Phi}(\mathbf{x}, \mathbf{y})$, *where* $\nabla_{\mathbf{x}}\mathbf{\Phi}(\mathbf{x}, \mathbf{y})$ *is positive definite for all* $\mathbf{x}, \mathbf{y} \in X$. *Then, if* [4]

$$\left|\!\left|\!\left| \mathrm{symm}\left(\nabla_{\mathbf{x}}\mathbf{\Phi}\left(\mathbf{x}^1, \mathbf{y}^1\right)^{-1/2}\right) \cdot \left[\nabla_{\mathbf{y}}\mathbf{\Phi}\left(\mathbf{x}^2, \mathbf{y}^2\right) + \mathbf{F}\left(\mathbf{y}^2\right) - \nabla_{\mathbf{y}}\mathbf{\Phi}\left(\mathbf{y}^2, \mathbf{y}^2\right)\right] \cdot \right.\right.\right.$$
$$\left.\left.\left. \mathrm{symm}\left(\nabla_{\mathbf{x}}\mathbf{\Phi}\left(\mathbf{x}^3, \mathbf{y}^3\right)^{-1/2}\right) \right|\!\right|\!\right| < 1, \tag{5.11}$$

$\forall \mathbf{x}^1, \mathbf{y}^1, \mathbf{x}^2, \mathbf{y}^2, \mathbf{x}^3, \mathbf{y}^3 \in X$, *with* $t_k \equiv 1$, $\{\mathbf{x}^k\} \to \mathbf{x}^*$, *where* $\mathbf{x}^*$ *is the unique solution of* [**VIP**].

Proof

(a) Follows from [Coh88, Th. 2.2] and Theorem 5.1.c, where we choose $\varepsilon_k \equiv 1$.

(b) Follows from [Daf83, Th. 2.1 and Rem. 2.2], Theorem 5.1.a, and the identification $\mathbf{g}(\mathbf{x}, \mathbf{y}) = \mathbf{\Phi}(\mathbf{x}, \mathbf{y}) + \mathbf{F}(\mathbf{y}) - \mathbf{\Phi}(\mathbf{y}, \mathbf{y})$. □

Remark 5.1 Comparing (5.10) to the Condition (3.32) for convergence using unit steps in the symmetric case, we see that they coincide if $m_{\mathbf{F}} = M_{\mathbf{F}}$; in general $m_{\mathbf{F}} < M_{\mathbf{F}}$ holds, and (5.10) is then more severe than (3.32), see also [Coh88, Rem. 2.1].

Remark 5.2 Cohen [Coh88] extends the auxiliary problem principle (validated in **(a)** above) to variational inequalities with a point-to-set map $\mathbf{U}$, i.e., a problem [**GVIP**] with $\mathbf{F} \equiv \mathbf{0}$. In this extended algorithm, the mapping $\mathbf{U}$ is iteratively approximated by an arbitrary element of $\mathbf{U}(\mathbf{x}^k)$; this approximation strategy extends that for [**NNLP**] where the convex function u is iteratively linearized using an arbitrary subgradient $\boldsymbol{\xi}_u^k \in \partial u(\mathbf{x}^k)$ [see Section 3.3]. The convergence conditions for Cohen's method for [**GVIP**] are similar to those of this algorithm for [**NNLP**], and include boundedness conditions on $\mathbf{U}$ and the Condition (3.8) on the sequence $\{\varepsilon_k\}$ defining (5.6). Both for the approach in **(a)** and its extension, the generalized norm (3.9) is utilized as merit function.

Specialized algorithms for various equilibrium problems have been analyzed within Dafermos' framework (see [Zha88, Daf90, ZhD91] and references cited therein). The algorithm of [Noo92] is a special case from this class, with $\mathbf{\Phi}(\mathbf{x}, \mathbf{y}) \equiv \mathbf{\Phi}(\mathbf{x})$. This algorithm has in [Noo85, Noo87, Noo91a] been considered for variational inequality problems of special form, mainly as a tool for establishing the existence of a nonempty solution set Ω; this is a potential use of the whole class of cost approximation methods (see also the next section).

[4] This condition implies that $\mathbf{F}$ is strictly monotone on X [Daf83]. Convergence can also be established for unbounded feasible sets X, provided that (5.11) holds uniformly and that $\mathbf{\Phi}$ is strongly monotone on X with respect to $\mathbf{x}$, see [Daf83].

For the Jacobi method [see (2.19)] Dafermos [Daf83] shows that (5.11) holds under conditions similar to those obtained in specialized analyses [Daf82, PaC82, Pan85].

The most frequently studied method for variational inequality problems is the projection algorithm (cf. Table 5.1), see [BrS69, GaK70, Klu70, Sib70, BaP74, Gol75, Aga78, Daf80, BeG82, Fan82, FiN82, PaC82, Daf83]. An application of Theorem 5.2.a to this method shows that convergence holds for matrices $\mathbf{Q}$, positive definite on X with minimal eigenvalue μ satisfying

$$2\mu > \frac{M_{\mathbf{F}}^2}{m_{\mathbf{F}}}. \tag{5.12}$$

This condition slightly relaxes that of [PaC82, Cor. 2.11], in which the left hand side of (5.12) is replaced by the smaller (unless $\mathbf{Q}$ is the identity matrix) constant $\frac{2\mu^2}{\|\mathbf{Q}\|}$.

From these examples, Theorem 5.2 seems to be strong enough to be able to reproduce earlier obtained convergence conditions for special cases of cost approximation methods for variational inequalities. There are, however, two important deficiencies of the successive approximation version of the class of cost approximation methods. First, the conditions for convergence require either that $\mathbf{F}$ is strongly monotone, or they are very difficult to verify in practice [cf. Inequality (5.11)]. This is in contrast to the very mild (and in many cases easily checked) conditions that we have found to guarantee convergence of cost approximation methods in the case of nonlinear programming. Second, it is not possible to use the highly artificial merit function (5.1) in a line search, which undoubtedly would enhance the methods' practical performance. If a useful merit function can be constructed, the inclusion of a line search may improve the performance as well as facilitate a weakening of the conditions for convergence. This is the topic of the following sections.

5.5 A class of merit functions for variational inequalities

The merit functions to be studied are all of the type given by the following definition.

Definition 5.1 (Gap function) *Let Ω be the set of solutions to* [**VIP**]. *A function $\gamma : X \mapsto \Re \cup \{-\infty, +\infty\}$ is a gap function for* [**VIP**] *if*

(1) *γ is restricted in sign on X, and*

(2) *$\gamma(\mathbf{x}) = 0 \Leftrightarrow \mathbf{x} \in \Omega$.*

A gap function provides a measure of the violation of [**VIP**] at any point $\mathbf{x} \in X$, and by minimizing γ over X (assuming that γ is nonnegative on X), a point in Ω is obtained. It may therefore be used as a merit function for variational inequalities. One class of gap functions is inherent in the cost approximation concept. In the following we will describe and analyze these gap functions in detail, and validate descent methods based on them.

We begin by showing that the cost approximation subproblems define gap functions for [**VIP**]. Let $\varphi : X \mapsto \Re$ be a convex function in C^1 on X, and define

$$\psi(\mathbf{x}) \stackrel{\text{def}}{=} \sup_{\mathbf{y} \in X} L(\mathbf{y}, \mathbf{x}), \tag{5.13}$$

where

$$L(\mathbf{y},\mathbf{x}) \stackrel{\text{def}}{=} \varphi(\mathbf{x}) - \varphi(\mathbf{y}) + [\mathbf{F}(\mathbf{x}) - \nabla\varphi(\mathbf{x})]^{\mathrm{T}}(\mathbf{x}-\mathbf{y}). \tag{5.14}$$

The problem defining $\psi(\mathbf{x})$ is the subproblem [$\mathbf{CNLP}_\varphi$], defined in Section 5.1.

Lemma 5.2 [LaP90, Pat91b] (ψ is a gap function) [5] *Let $\varphi : X \mapsto \Re$ be a convex function in C^1 on X. Then $\psi : X \mapsto \Re_+ \cup \{+\infty\}$ is a gap function for* [**VIP**].

Proof For any $\mathbf{x} \in X$, $\psi(\mathbf{x}) \geq 0$ holds, since $L(\mathbf{x},\mathbf{x}) = 0$.

Next, assume that $\mathbf{x}^* \in \Omega$. Then

$$\begin{aligned} L(\mathbf{y},\mathbf{x}^*) &\leq \mathbf{F}(\mathbf{x}^*)^{\mathrm{T}}(\mathbf{x}^* - \mathbf{y}) \quad (\varphi \text{ convex}) \\ &\leq 0, \ \forall \mathbf{y} \in X. \qquad (\mathbf{x}^* \in \Omega) \end{aligned}$$

Since $L(\mathbf{x}^*,\mathbf{x}^*) = 0$ holds, we have that $\psi(\mathbf{x}^*) = 0$.

Conversely, let $\psi(\mathbf{x}^*) = 0$, for some $\mathbf{x}^* \in X$. Since, from (5.13), $L(\mathbf{y},\mathbf{x}^*) \leq 0$, $\forall \mathbf{y} \in X$, $\mathbf{x}^*$ must be a solution to the problem defining $\psi(\mathbf{x}^*)$. Replacing $\mathbf{y}(\mathbf{x}^*)$ with $\mathbf{x}^*$ in the necessary optimality conditions for (5.13),

$$[\nabla\varphi(\mathbf{y}(\mathbf{x}^*)) + \mathbf{F}(\mathbf{x}^*) - \nabla\varphi(\mathbf{x}^*)]^{\mathrm{T}}(\mathbf{y} - \mathbf{y}(\mathbf{x}^*)) \geq 0, \ \forall \mathbf{y} \in X \tag{5.18}$$

then yields that $\mathbf{x}^*$ solves [**VIP**]. □

This lemma states that the subproblem [$\mathbf{CNLP}_\varphi$] of the cost approximation method, resulting from a *symmetric* approximation of the, possibly asymmetric, mapping $\mathbf{F}$, defines a merit function for [**VIP**]. Obviously, the better the gradient $\nabla\varphi$ approximates $\mathbf{F}$, the closer a solution $\mathbf{y}(\mathbf{x})$ of [$\mathbf{CNLP}_\varphi$] will be to Ω.

---

[5] There is another way of establishing this result, utilizing conjugate duality and *Fenchel's inequality* [Fen49]; this result states that

$$g(\mathbf{x}) + g^\circ(\mathbf{z}) - \mathbf{x}^{\mathrm{T}}\mathbf{z} \geq 0, \ \forall \mathbf{x},\mathbf{z} \in \Re^n, \tag{5.15}$$

and that equality holds in (5.15) if and only if $\mathbf{z} \in \partial g(\mathbf{x})$, where $g : \Re^n \mapsto \Re \cup \{+\infty\}$ is a proper, convex function, and the conjugate g° of g is given by

$$g^\circ(\mathbf{w}) = \sup_{\mathbf{y}} \{\mathbf{y}^{\mathrm{T}}\mathbf{w} - g(\mathbf{y})\}. \tag{5.16}$$

To show that Fenchel's inequality yields the desired property of ψ, we identify g with $\delta_X + \varphi$ and let $\mathbf{z} = \nabla\varphi(\mathbf{x}) - \mathbf{F}(\mathbf{x})$. Then, from (5.16),

$$\begin{aligned} \psi(\mathbf{x}) &= \delta_X(\mathbf{x}) + \varphi(\mathbf{x}) + \sup_{\mathbf{y}}\{[\nabla\varphi(\mathbf{x}) - \mathbf{F}(\mathbf{x})]^{\mathrm{T}}\mathbf{y} - \delta_X(\mathbf{y}) - \varphi(\mathbf{y})\} - [\nabla\varphi(\mathbf{x}) - \mathbf{F}(\mathbf{x})]^{\mathrm{T}}\mathbf{x} \\ &= g(\mathbf{x}) + g^\circ(\nabla\varphi(\mathbf{x}) - \mathbf{F}(\mathbf{x})) - [\nabla\varphi(\mathbf{x}) - \mathbf{F}(\mathbf{x})]^{\mathrm{T}}\mathbf{x} \\ &= g(\mathbf{x}) + g^\circ(\mathbf{z}) - \mathbf{x}^{\mathrm{T}}\mathbf{z}, \end{aligned} \tag{5.17}$$

which shows that the value $\psi(\mathbf{x})$ may be interpreted as the gap in Fenchel's inequality, for a special choice of convex function g. From this relationship, (5.15) then yields that $\psi(\mathbf{x}) \geq 0$, $\forall \mathbf{x} \in X$ and that

$$\psi(\mathbf{x}) = 0 \Leftrightarrow \nabla\varphi(\mathbf{x}) - \mathbf{F}(\mathbf{x}) \in N_X(\mathbf{x}) + \nabla\varphi(\mathbf{x}) \Leftrightarrow -\mathbf{F}(\mathbf{x}) \in N_X(\mathbf{x}) \Leftrightarrow \mathbf{x} \in \Omega,$$

i.e., ψ is a gap function. (The latter result may also be established from the relation $\mathbf{z} \in \partial g(\mathbf{x}) \Leftrightarrow \mathbf{x} \in \partial g^\circ(\mathbf{z}) = Y(\mathbf{x})$, as in the proof of the lemma.) Based on the relation (5.17), further properties of ψ may be established; see [Auc89] and [Aub93, pp. 329–330].

Furthermore, assume that $\nabla\mathbf{F}$ is symmetric and positive semidefinite on X. We may then choose the function φ such that $\nabla\varphi \equiv \mathbf{F}$, and [**CNLP**$_\varphi$] reduces to [**VIP**]. (Then, $\psi(\mathbf{x}) = f(\mathbf{x}) + c$, where $c = -\inf_{\mathbf{x}\in X} f(\mathbf{x})$, so that minimizing ψ is equivalent to minimizing f over X.) This possibility of choosing $\nabla\varphi$ so as to approximate $\mathbf{F}$ exactly for symmetric variational inequalities suggests a symmetrization strategy for the asymmetric case. In that case, the non-existence of the optimization problem [**CNLP**] is due to the fact that the value of the line integral in [**CNLP**] is dependent of the curve, Γ, taken from $\mathbf{0}$ to $\mathbf{x} \in X$. To define an unambiguous integral, we may choose the curve Γ as the straight line from $\mathbf{0}$ to $\mathbf{x}$, and hence introduce a parametrization $\mathbf{s} = \mathbf{s}(t) = t\mathbf{x}$, with $t \in [0, 1]$. We then define

$$\varphi(\mathbf{x}) \stackrel{\text{def}}{=} \int_\Gamma \mathbf{F}(\mathbf{s})^{\mathrm{T}} d\mathbf{s} = \int_0^1 \Psi(t)dt, \tag{5.19}$$

where $\Psi : \Re \mapsto \Re$ is given by $\Psi(t) = \mathbf{F}(t\mathbf{x})^{\mathrm{T}}\mathbf{x}$. Clearly, $\varphi \equiv f$ in [**CNLP**], see also [HaP90b, p. 165], if $\mathbf{F} \equiv \nabla f$. The following example shows that a well known symmetrization of affine variational inequalities is obtained from the application of (5.19).

Example 5.1 Consider the asymmetric and affine variational inequality, i.e., $\mathbf{F}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, where $\mathbf{A}$ is an asymmetric matrix in $\Re^{n\times n}$. Choosing φ as in (5.19), we obtain

$$\varphi(\mathbf{x}) = \int_0^1 (t\mathbf{A}\mathbf{x} - \mathbf{b})^{\mathrm{T}}\mathbf{x}dt = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} - \mathbf{b}^{\mathrm{T}}\mathbf{x},$$

which indeed defines a symmetrization of $\nabla\mathbf{F}$, since $\nabla^2\varphi = \frac{1}{2}\left(\mathbf{A} + \mathbf{A}^{\mathrm{T}}\right)$ is the symmetric part of $\nabla\mathbf{F} = \mathbf{A}$. □

The class of gap functions defined through symmetric cost approximations of $\mathbf{F}$ is actually equivalent to that proposed by Auchmuty [Auc89]. He applies to [**VIP**] earlier made studies on merit functions for operator equations and initial value problems [Auc83, Auc86, Auc88], which in turn define a development of previous related works in [MoF53, Gur64, Ton72, BrE76a, BrE76b, Ton84]. In these applications, the problems of minimizing the merit function are least squares and energy minimization problems. The purpose of devising merit functions for these problems lies not so much in the computing of solutions as in establishing the existence and uniqueness of solutions (i.e., showing that there must exist some (possibly unique) point $\mathbf{x}^* \in X$ with $\psi(\mathbf{x}^*) = 0$). This usage of the proposed merit functions could be seen as an alternative to the approach of Noor for establishing existence of solutions (see Section 5.4); furthermore, the descent methods for [**VIP**], based on ψ, developed here could be used to compute solutions to these problems.

Below we summarize the results of Auchmuty [Auc89] on the gap function ψ.

Theorem 5.3 [Auc89] (Saddle points of L are in Ω) *Let $\varphi : X \mapsto \Re$ be a convex function in C^1 on X. If $(\mathbf{x}^*, \mathbf{y}^*) \in X \times X$ satisfies*

$$L(\mathbf{x}^*, \mathbf{y}) \le L(\mathbf{x}^*, \mathbf{y}^*) \le L(\mathbf{x}, \mathbf{y}^*), \tag{5.20}$$

for all $(\mathbf{x}, \mathbf{y}) \in X \times X$, then $\mathbf{x}^*$ solves [**VIP**].

Since $L(\mathbf{x}, \mathbf{y})$ is not convex in the $\mathbf{x}$ argument, it is, in general, not a true saddle function, and the point $\mathbf{y}^*$ does, in general, not solve [**VIP**].

The optimization formulation

$$[\mathbf{CNLP}_\psi] \qquad \inf_{\mathbf{x}\in X} \psi(\mathbf{x}),$$

where ψ is defined by (5.13), corresponds to the rightmost inequality in (5.20).[6] Auchmuty shows that ψ is l.s.c. on X, and finite if X is bounded. The finiteness of ψ can be ensured also for unbounded sets by for instance imposing strong convexity on φ. Since a minimum of ψ over X always exists whenever ψ is finite, [$\mathbf{CNLP}_\psi$] can be used to detect the non-existence of a solution to [**VIP**], since, in this case, $\min_{\mathbf{x}\in X} \psi(\mathbf{x}) > 0$; the properties, with respect to [**VIP**], of such minimizers are subjects for further research.

5.6 Some related merit functions

In this section we show that the class of gap functions defined by ψ includes well known optimization formulations of variational inequality problems.

Let $\varphi \equiv 0$ in (5.14). Exploring the rightmost inequality of (5.20), we obtain the *primal gap function* $G : X \mapsto \Re_+ \cup \{+\infty\}$, defined by

$$G(\mathbf{x}) \stackrel{\text{def}}{=} \sup_{\mathbf{y}\in X} \mathbf{F}(\mathbf{x})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}), \tag{5.21}$$

which was introduced by Zuhovickiĭ *et al.* [ZPP69a, ZPP69b, ZPP70, ZPP73]. Some important properties of G are summarized below.

Theorem 5.4 (Properties of G) *For any $\mathbf{x} \in X$, let $Y(\mathbf{x})$ denote the (possibly empty) set of optimal solutions to the problem defined in (5.21).*

(a) *G is a gap function.*

(b) *G is l.s.c. on X.*

(c) *If X is bounded and $\mathbf{F} \in C^1$ on X, then G is Lipschitz continuous on X.*

(d) *If $\mathbf{F} \in C^1$ on X, then G is differentiable at $\mathbf{x} \in X$ if $Y(\mathbf{x}) = \{\mathbf{y}\}$, with $\nabla G(\mathbf{x}) = \mathbf{F}(\mathbf{x}) + \nabla \mathbf{F}(\mathbf{x})^{\mathrm{T}}(\mathbf{x}-\mathbf{y})$.*

(e) *If $\mathbf{F} \in C^1$ on X, and monotone on X, then if $\mathbf{x} \notin \Omega$ and $Y(\mathbf{x}) = \{\mathbf{y}\}$, the direction $\mathbf{d} = \mathbf{y}-\mathbf{x}$ is a feasible direction of descent with respect to G at $\mathbf{x}$, and the directional derivative satisfies $G'(\mathbf{x};\mathbf{d}) = \nabla G(\mathbf{x})^{\mathrm{T}}\mathbf{d} \le -G(\mathbf{x})$.*

(f) *G is convex on X if $\mathbf{F}(\mathbf{x})^{\mathrm{T}}\mathbf{x}$ is convex on X and each component of $\mathbf{F}$ is concave on X.*

---

[6] Auchmuty also considers the leftmost inequality of (5.20) in a formulation of a dual optimization principle for [**VIP**],

$$[\mathbf{CNLP}_g] \qquad \sup_{\mathbf{y}\in X} g(\mathbf{y}), \text{ where } g(\mathbf{y}) \stackrel{\text{def}}{=} \inf_{\mathbf{x}\in X} L(\mathbf{y},\mathbf{x}).$$

Although it can be shown [Auc89, Th. 4] that g is nonpositive on X, and that a point $\mathbf{y} \in X$ with $g(\mathbf{y}) = 0$ is a solution to [**VIP**], a solution to [**VIP**] is not necessarily a zero of g even for strongly monotone maps $\mathbf{F}$ [LaP90, Ex. 2.1], and g is therefore not a gap function. (However, for pseudo-monotone maps $\mathbf{F}$ and the choice $\varphi \equiv 0$, g is a gap function; see [ZPP69a, ZPP69b, DeP72, ZPP73, Aus76, Kar76, Hea78, Hea82, HeN82, HLN84, NgD84, HaM87, HeL89, LaP90] for further discussions on this optimization formulation.)

(g) (A fixed point characterization of Ω) $\mathbf{x} \in \Omega \Leftrightarrow \mathbf{x} \in Y(\mathbf{x})$.

(h) (A stationary point characterization of Ω) *Under the conditions on* $\mathbf{F}$ *in* **(e)**, $\mathbf{x} \in \Omega \Leftrightarrow G'(\mathbf{x}; \mathbf{y} - \mathbf{x}) \geq 0,\ \forall \mathbf{y} \in X$.

Proof

(a) See [ZPP69a, ZPP69b, ZPP70, ZPP73]. See also [DeR70, Sec. 3.1] and [Hea78, Hea82].

(b) See [Dun80a].

(c) See [Mar85].

(d) See [DeP72]. See also [ZPP73, Aus76, Hea78, Hea82, HLN84].

(e) See [DeP72]. See also [Hea82] for the first result, and [Mar85].

(f) See [Hea78, Hea82, HLN84].

(g) See [ZPP69b]. See also [Evt85, Mar85].

(h) See [Mar85]. □

When X is polyhedral, (5.21) reduces to a linear program, which, in the symmetric case, is equivalent to that solved in the Frank-Wolfe algorithm. The primal gap function G has been used as a merit function in several algorithms for [**VIP**], see [ZPP70, ZPP73, DeP72, Pri75, LaH84, Mar85, Mar86, MaD85, MaD87, MaG88, DuM89, MaD89].

If the feasible set X is a closed and convex cone, then [**VIP**] reduces to the *nonlinear complementarity problem*, i.e., to find an $\mathbf{x}^* \in X$ such that

[**NCP**] $$\mathbf{F}(\mathbf{x}^*) \in X^\circ \text{ and } \mathbf{F}(\mathbf{x}^*)^{\mathrm{T}} \mathbf{x}^* = 0.$$

In the special case of $X = \Re^n_+$, Cottle [Cot66] (see also [Kar71, CPV89]) shows that $\mathbf{x}^*$ solves [**NCP**] if and only if $\mathbf{x}^*$ minimizes $\mathbf{F}(\mathbf{x})^{\mathrm{T}}\mathbf{x}$ on the set $\{\mathbf{x} \geq \mathbf{0} \mid \mathbf{F}(\mathbf{x}) \geq \mathbf{0}\}$, and $\mathbf{F}(\mathbf{x}^*)^{\mathrm{T}} \mathbf{x}^* = 0$. Auchmuty shows that [$\mathbf{CNLP}_G$] extends this optimization reformulation to a general nonlinear complementarity problem, by establishing that [$\mathbf{CNLP}_G$] in fact is equivalent to minimizing $\mathbf{F}(\mathbf{x})^{\mathrm{T}}\mathbf{x}$ on the set $\{\mathbf{x} \in X \mid \mathbf{F}(\mathbf{x}) \in X^\circ\}$, and that $\mathbf{x}^*$ solves [**NCP**] if and only if $\mathbf{F}(\mathbf{x}^*)^{\mathrm{T}} \mathbf{x}^* = 0$ (i.e., if and only if $G(\mathbf{x}^*) = 0$). Noting that the above development corresponds to the choice $\varphi \equiv 0$, we may conclude that the gap function ψ provides a means for making optimization reformulations of nonlinear complementarity problems; the one by Cottle is indeed a very special case of such optimization reformulations. (Thus, new descent algorithms for [**NCP**] may be constructed, by specializing those algorithms developed here that are valid for variational inequalities over unbounded feasible sets.)

The family of gap functions developed by Smith [Smi83a, Smi83b, Smi84] is defined by

$$G^p(\mathbf{x}) \stackrel{\text{def}}{=} \sum_{i \in \mathcal{E}} \left(\left[\mathbf{F}(\mathbf{x})^{\mathrm{T}} \left(\mathbf{x} - \hat{\mathbf{y}}^i \right) \right]_+ \right)^p, \quad \mathbf{x} \in X,$$

where $p > 0$ and $\{\hat{\mathbf{y}}^i | i \in \mathcal{E}\}$ is the set of extreme points of the bounded polyhedron X. These functions have the following properties, similar to the ones for ψ: for all $p \geq 1$, $G^p(\mathbf{x}) > 0,\ \forall\ \mathbf{x} \notin \Omega$ and $G^p(\mathbf{x}^*) = 0 \Leftrightarrow \mathbf{x}^* \in \Omega$. Further studies of the properties of

$G^p(\mathbf{x})$ are made in [HLN84]. They show that the primal gap function G is a limiting case of this class of functions, since, for all $\mathbf{x} \in X$,

$$\lim_{p \to +\infty} [G^p(\mathbf{x})]^{1/p} = \max_{i \in \mathcal{E}} \mathbf{F}(\mathbf{x})^{\mathrm{T}} \left(\mathbf{x} - \hat{\mathbf{y}}^i\right) = G(\mathbf{x}).$$

They proceed to show that $G^p(\mathbf{x})$ is differentiable for $2 \leq p < +\infty$, and convex for $p \geq 2$ and $p = +\infty$ e.g. if $\mathbf{F}$ is affine and monotone. When $p < +\infty$, in order to evaluate $G^p(\mathbf{x})$, the extreme points must be known explicitly; the practical use of these gap functions is therefore limited to polyhedral sets, in combination with simplicial decomposition/column generation techniques, see [Smi83a, Smi83b, Pat90].

Fukushima [Fuk92] introduces a gap function, which can be identified as a special case of ψ, by using $\varphi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}$, where $\mathbf{Q}$ is a symmetric and positive definite matrix in $\Re^{n \times n}$. The corresponding subproblem (5.13) is the same as in the projection method (see Table 5.1), i.e.,

$$\mathbf{y}(\mathbf{x}) = P_X^{\mathbf{Q}} \left(\mathbf{x} - \mathbf{Q}^{-1}\mathbf{F}(\mathbf{x})\right). \tag{5.22}$$

Fukushima [Fuk92] establishes convergence for a descent method which uses line searches with respect to the corresponding gap function, ψ_{Fu}, in the direction $\mathbf{y}(\mathbf{x}) - \mathbf{x}$, where $\mathbf{y}(\mathbf{x})$ is given by (5.22). In [TFI93], this gap function is utilized in a descent method which is based on Newton subproblems; the resulting method is a natural extension of the Newton methods of [MaD85, MaD87, MaD89] which utilize the primal gap function G. The former approach for Newton methods is related to recently proposed Newton-type descent methods for variational inequalities and complementarity problems based on B-differentiable equations [Pan86a, HaP90a, HaX90, Pan90a, Pan90b, Pan91, HPR92].

5.7 Further properties of merit functions for variational inequalities

We now further investigate the properties of the gap function ψ and the optimization formulation [**CNLP**$_\psi$]. We will in particular provide a characterization of Ω in terms of fixed points and stationary points, and give sufficient conditions for the continuity, continuous differentiability and convexity of ψ. These properties will then be used in the validation of descent methods, based on the merit function ψ.

The first result is a direct extension of Lemma 4.1.a for [**CNLP**]; it is also an extension of Theorem 5.4.g, for the primal gap function G.

Theorem 5.5 [LaP90] (A fixed point characterization of Ω) *Let $\varphi : X \mapsto \Re$ be a convex function in C^1 on X. For any $\mathbf{x} \in X$, let $Y(\mathbf{x})$ denote the (possibly empty) set of optimal solutions to (5.13). Then,*

$$\mathbf{x} \in \Omega \Leftrightarrow \mathbf{x} \in Y(\mathbf{x}).$$

Note the difference to Lemma 5.1, which does not include any monotonicity assumption on $\mathbf{\Phi}$; this difference is due to the fact that in Theorem 5.5 the solution set $Y(\mathbf{x})$ is described in terms of the variational inequality [**VIP**$_{\mathbf{\Phi}}$].

The next result establishes that ψ is continuous whenever it is everywhere finite; continuity of ψ is important if a line search is to be well defined (i.e., closed in the sense of Zangwill [Zan69b]; see Definition A.7).

Theorem 5.6 [LaP90] (Continuity properties of ψ and Y) *Let $\varphi : X \mapsto \Re$ be a convex function in C^1 on X.*

(a) *For any $\mathbf{x} \in X$, the set $Y(\mathbf{x})$ is closed and convex, the mapping $\mathbf{x} \mapsto Y(\mathbf{x})$ is closed and ψ is l.s.c. on X.*

(b) *If X is bounded or if φ is strongly convex on X, then for any $\mathbf{x} \in X$, $Y(\mathbf{x})$ is nonempty and bounded, and ψ is continuous on X.*

(c) *If X is bounded and $\mathbf{F} \in C^1$ on X, and $\varphi \in C^2$ on X, then ψ is Lipschitz continuous on X.*

(d) *If X is bounded and φ is strictly convex on X, or if φ is strongly convex on X, then, for any $\mathbf{x} \in X$, $Y(\mathbf{x})$ is a singleton and $\mathbf{x} \mapsto Y(\mathbf{x})$ is continuous.*

Under additional assumptions on $\mathbf{F}$ and φ, the mapping $\mathbf{x} \mapsto Y(\mathbf{x})$ can be shown to also be Lipschitz continuous, see Lemma 5.3. This property is used in the validation of inexact line searches with respect to ψ.

Theorem 5.7 [LaP90] (Differentiability of ψ) *Let $\mathbf{F}$ be in C^1 on X, and let $\varphi : X \mapsto \Re$ be in C^2 on X. If X is bounded and φ is strictly convex on X, or if φ is strongly convex on X, then $\psi \in C^1$ on X, with*

$$\nabla\psi(\mathbf{x}) = \mathbf{F}(\mathbf{x}) + \left[\nabla\mathbf{F}(\mathbf{x})^{\mathrm{T}} - \nabla^2 f(\mathbf{x})\right](\mathbf{x} - \mathbf{y}(\mathbf{x})), \tag{5.23}$$

where $\mathbf{y}(\mathbf{x})$ is the unique solution to (5.13).

Remark 5.3 The theorem gives an affirmative answer to a question put by Harker and Pang [HaP90b, p. 165], of whether or not it is possible to equivalently formulate asymmetric variational inequalities as differentiable optimization problems.[7]

Remark 5.4 From the strict convexity of φ, the fixed point property of Theorem 5.5, and (5.23), it follows that

$$\nabla\psi(\mathbf{x}^*) = \mathbf{F}(\mathbf{x}^*), \ \forall \mathbf{x}^* \in \Omega. \tag{5.24}$$

This relation provides another justification of the symmetric approximations of $\mathbf{F}$ that constitute the basis for the construction of the merit functions ψ. When $\mathbf{F}$ is the gradient of a convex function, it is known [Man88, BuF91] that $\mathbf{F}(\mathbf{x})$ is constant on Ω, in which case (5.24) shows that this is also true for $\nabla\psi$ although ψ is not convex in general. The invariance property of $\mathbf{F}$ on Ω does, however, not extend to asymmetric monotone variational inequalities in general.[8]

[7] One may also note that differentiability does in general not hold for G, unless the feasible set is *strongly convex* in the sense of [ZPP69b]; this condition excludes applications to polyhedral feasible sets.

The next theorem characterizes solutions to [**VIP**] as stationary points to the problem [$\mathbf{CNLP}_\psi$], for different monotonicity properties of $\mathbf{F}$ and, correspondingly, different convexity properties of φ. The theorem validates descent methods for [$\mathbf{CNLP}_\psi$].

Theorem 5.8 [LaP90] (A stationary point characterization of Ω) *Let* $\mathbf{F}$ *be in* C^1 *on* X, *and* $\varphi : X \mapsto \Re$ *be a convex function in* C^2 *on* X. *Under either one of the following three additional assumptions,* $\mathbf{x}^*$ *solves* [**VIP**] *if and only if* $\mathbf{x}^*$ *is a stationary point of the problem* [$\mathbf{CNLP}_\psi$], *i.e.,*

$$\mathbf{x}^* \in \Omega \Leftrightarrow \psi'(\mathbf{x}^*; \mathbf{x} - \mathbf{x}^*) \geq 0, \ \forall \mathbf{x} \in X.$$

(1) *X is bounded, and the matrix $\nabla\mathbf{F}(\mathbf{x}^*)^{\mathrm{T}} - \nabla^2\varphi(\mathbf{x}^*)$ is positive definite.*

(2) *X is bounded, φ is strictly convex on X, and the matrix $\nabla\mathbf{F}(\mathbf{x}^*)^{\mathrm{T}} - \nabla^2\varphi(\mathbf{x}^*)$ is positive semidefinite.*

(3) *$\mathbf{F}$ is strongly monotone on X, φ is strongly convex on X, $\nabla\varphi$ is Lipschitz continuous on X, and*

$$M_{\nabla\varphi} - m_\varphi < m_{\mathbf{F}}. \tag{5.26}$$

Remark 5.5 The Inequality (5.26) is possible to fulfill through a proper scaling of any strongly convex function φ with a Lipschitz continuous gradient.

As a direct consequence of this theorem, if a feasible point $\mathbf{x}$ is not a solution to [**VIP**], then there is a $\mathbf{y} \in Y(\mathbf{x})$ such that the direction $\mathbf{y} - \mathbf{x}$ is a feasible direction of descent with respect to ψ.

---

[8] Consider the linear program

$$[\mathbf{LP}] \qquad \min \ \mathbf{c}^{\mathrm{T}}\mathbf{x} \quad \text{s.t.} \ \ \mathbf{x} \in X = \{\mathbf{x} \in \Re^n_+ \mid \mathbf{A}\mathbf{x} \geq \mathbf{b}\}.$$

Through a saddle point formulation of [**LP**], its optimality conditions may be expressed as (e.g. [Mur83, Th. 4.16])

$$\begin{aligned} \nabla_{\mathbf{x}} L(\mathbf{x}^*, \mathbf{y}^*)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^*) \geq 0, \quad \forall \mathbf{x} \in \Re^n_+ \\ -\nabla_{\mathbf{y}} L(\mathbf{x}^*, \mathbf{y}^*)^{\mathrm{T}}(\mathbf{y} - \mathbf{y}^*) \geq 0, \quad \forall \mathbf{y} \in \Re^m_+, \end{aligned} \tag{5.25}$$

where $L(\mathbf{x}, \mathbf{y}) = \mathbf{c}^{\mathrm{T}}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ and $(\mathbf{x}^*, \mathbf{y}^*)$ is a primal-dual pair of solutions to [**LP**]. Formulating the variational inequality [**VIP**] on $\Re^n_+ \times \Re^m_+$, corresponding to the optimality conditions (5.25), yields a problem defining map $\mathbf{F} : \Re^n_+ \times \Re^m_+ \mapsto \Re^n \times \Re^m$, given by

$$\mathbf{F}(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} \nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{y}) \\ -\nabla_{\mathbf{y}} L(\mathbf{x}, \mathbf{y}) \end{pmatrix} = \begin{pmatrix} \mathbf{c} - \mathbf{A}^{\mathrm{T}}\mathbf{y} \\ \mathbf{A}\mathbf{x} - \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & -\mathbf{A}^{\mathrm{T}} \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} + \begin{pmatrix} \mathbf{c} \\ -\mathbf{b} \end{pmatrix}.$$

Note that the matrix

$$\begin{pmatrix} \mathbf{0} & -\mathbf{A}^{\mathrm{T}} \\ \mathbf{A} & \mathbf{0} \end{pmatrix}$$

is *skew-symmetric* (i.e., $\mathbf{Q}^{\mathrm{T}} = -\mathbf{Q}$), and therefore positive semidefinite; hence, $\mathbf{F}$ is monotone on $\Re^n_+ \times \Re^m_+$ and [**VIP**] is a monotone variational inequality.

Now, assume that [**LP**] has a degenerate optimal solution $\mathbf{x}^*$. Then, the dual problem may have multiple optimal solutions, and it follows that $\mathbf{F}$ is *not* constant on the primal-dual solution set.

Theorem 5.9 [LaP90] (Descent theorem) *Let* $\mathbf{F}$ *be in* C^1 *on* X*, and* $\varphi : X \mapsto \Re$ *be a convex function in* C^2 *on* X*.*

(a) *If* X *is bounded,* $\mathbf{x} \notin \Omega$*, and the matrix* $\nabla \mathbf{F}(\mathbf{x})^{\mathrm{T}} - \nabla^2 \varphi(\mathbf{x})$ *is positive definite, then there exists a* $\mathbf{y} \in Y(\mathbf{x})$ *such that*

$$\psi'(\mathbf{x}; \mathbf{y} - \mathbf{x}) < 0.$$

(b) *If* X *is bounded,* φ *is strictly convex on* X*,* $\mathbf{x} \notin \Omega$ *and the matrix* $\nabla \mathbf{F}(\mathbf{x})^{\mathrm{T}} - \nabla^2 \varphi(\mathbf{x})$ *is positive semidefinite, then*

$$\nabla \psi(\mathbf{x})^{\mathrm{T}}(\mathbf{y}(\mathbf{x}) - \mathbf{x}) < 0.$$

(c) *Assume that* $\mathbf{F}$ *is strongly monotone on* X*, and that* φ *is strongly convex with a Lipschitz continuous gradient on* X*. Then, for any* $\mathbf{x} \in X$*,*

$$\nabla \psi(\mathbf{x})^{\mathrm{T}}(\mathbf{y}(\mathbf{x}) - \mathbf{x}) \le -(m_{\mathbf{F}} + m_{\varphi} - M_{\nabla\varphi}) \|\mathbf{y}(\mathbf{x}) - \mathbf{x}\|^2. \tag{5.27}$$

For an affine map $\mathbf{F}$ and $\varphi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}$, the conditions in **(b)** reduce to the positive semidefiniteness of the matrix $\mathbf{A}^{\mathrm{T}} - \mathbf{Q}$ on X. The symmetrization defined in Example 5.1 has the descent property whenever $\mathbf{A}$ is positive definite on X.

Using the stationary point property we have thus established that the solution to (5.13), obtained when evaluating $\psi(\mathbf{x})$, defines a feasible descent direction with respect to ψ, whenever $\mathbf{x} \notin \Omega$. In the next two sections we will show that, using an exact line search or an Armijo step length rule, the sequence of iteration points converges toward the unique solution $\mathbf{x}^*$, from any feasible starting point.

The descent methods to be presented here are, in general, only asymptotically convergent, and it is therefore of practical interest to obtain a measure of the distance from an approximate solution point to the set Ω. The following theorem establishes an *a posteriori* error bound for strongly monotone variational inequalities, measured in terms of the value of the function ψ.

Theorem 5.10 [LaP90] (A posteriori error bound) *Let* $\mathbf{F}$ *be strongly monotone on* X*, and let* $\varphi : X \mapsto \Re$ *be a function in* C^2 *on* X*, with a Lipschitz continuous gradient on* X*. If* φ *is chosen such that* $M_{\nabla\varphi} < 2m_{\mathbf{F}}$*, then for any* $\mathbf{x} \in X$*,*

$$\|\mathbf{x} - \mathbf{x}^*\|^2 \le \frac{2\psi(\mathbf{x})}{2m_{\mathbf{F}} - M_{\nabla\varphi}}. \tag{5.28}$$

Remark 5.6 The error bound (5.28) generalizes those earlier obtained in [Pan87] for $\varphi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$ (projection measure, Theorem 3.1) and $\varphi \equiv 0$ (linear programming measure, Theorem 4.1), and that obtained in [TFI93] for the choice of $\varphi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_{\mathbf{Q}}^2$. Notice that neither Lipschitz continuity (cf. [Pan87]), nor differentiability, of $\mathbf{F}$ is necessary, and that the result holds regardless of any convexity properties of φ.

Corollary 5.1 [LaP90] (Coercivity of ψ) *Under the assumptions of the theorem,* ψ *has bounded level sets and is weakly coercive on* X*.*

Although ψ is nonconvex in general, the corollary establishes properties which ψ shares with convex functions. (Weak coercivity and boundedness of level sets are well known properties of *strongly* convex functions.)

For affine variational inequalities, ψ can be constructed strictly convex.

Theorem 5.11 [LaP90] (Convexity of ψ) *Let* $\mathbf{F}$ *be affine, i.e.,* $\mathbf{F}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$. *Let* $\varphi : X \mapsto \Re$ *be given by* $\varphi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} - \mathbf{q}^{\mathrm{T}}\mathbf{x}$, *where* $\mathbf{Q}$ *is symmetric and positive definite on* X. *Then* ψ *is convex on* X *if the matrix* $\mathbf{A} + \mathbf{A}^{\mathrm{T}} - \mathbf{Q}$ *is positive semidefinite on* X, *and strictly convex if this matrix is positive definite on* X.

Under the assumptions of the theorem, the saddle function L defined in (5.14) is convex-concave; hence, under these conditions, [**VIP**] may be equivalently reformulated as a true saddle point problem.

The choice of φ as the symmetrization made in Example 5.1 leads to a strictly convex gap function whenever $\mathbf{A}$ is positive definite on X. Moreover, this property leads to the stationary point characterization, see Theorem 5.8.2. Whether or not ψ can be constructed convex for non-affine mappings $\mathbf{F}$ is an open question.

5.8 A descent algorithm for variational inequalities

In this section we utilize the descent Theorem 5.9 to establish global convergence of the cost approximation method where ψ is used in an exact line search. The possible use of an inexact Armijo step length rule is established in the next section.

Theorem 5.12 [LaP90] (Basic convergence of cost approximation algorithms) *Let* $\mathbf{F}$ *be in* C^1 *on* X, *and* $\varphi : X \mapsto \Re$ *be a convex function in* C^2 *on* X. *Let* $\mathbf{x}^0$ *be arbitrary in* X, *and let* $\{\mathbf{x}^k\}$ *be the sequence generated by the cost approximation algorithm under Rule M. Consider the following sets of assumptions.*

(1) X *is bounded, the matrix* $\nabla\mathbf{F}(\mathbf{x})^{\mathrm{T}} - \nabla^2\varphi(\mathbf{x})$ *is positive definite on* X, *and the point* $\mathbf{y}^k$ *is chosen as in Theorem 5.9.a.*

(2) X *is bounded,* φ *is strictly convex on* X, *and the matrix* $\nabla\mathbf{F}(\mathbf{x})^{\mathrm{T}} - \nabla^2\varphi(\mathbf{x})$ *is positive semidefinite on* X.

(3) $\mathbf{F}$ *is strongly monotone on* X, φ *is strongly convex with a Lipschitz continuous gradient on* X *and chosen such that (5.26) holds. Further,* X *is bounded or* $M_{\nabla\varphi} < 2m_{\mathbf{F}}$ *holds.*

Under either one of the above sets of assumptions, [**VIP**] *has a unique solution, and the sequence* $\{\mathbf{x}^k\}$ *converges to it.*

Remark 5.7 Under the condition that $M_{\nabla\varphi} < 2m_{\mathbf{F}}$ holds, the function ψ has bounded level sets, and convergence can be established without having to resort to boundedness of the feasible set X. This enables the descent algorithms developed here to be applied to complementarity problems.

5.9 Convergence under alternative step length rules

The use of the Armijo step length rule requires that $\nabla\psi$ is Lipschitz continuous; this property is a consequence of the Lipschitz continuity of the map $\mathbf{x} \mapsto Y(\mathbf{x})$, which is established in the below lemma.

Lemma 5.3 [LaP90] (Lipschitz continuity of Y) *Let* $\mathbf{F}$ *be Lipschitz continuous on* X, *and* $\varphi : X \mapsto \Re$ *a strongly convex function in* C^1 *on* X, *with a Lipschitz continuous gradient on* X. *Then the mapping* $\mathbf{x} \mapsto Y(\mathbf{x})$ *is Lipschitz continuous on* X *with modulus* $\frac{M_{\mathbf{F}}+M_{\nabla\varphi}}{m_\varphi}$.

Using this lemma, and the additional conditions that $\nabla\mathbf{F}$ and $\nabla^2\varphi$ are Lipschitz continuous on X, the Lipschitz continuity of $\nabla\psi$ on X follows from (5.23), see also [Fuk92]. From Lemma 4.3 it then follows that the Armijo rule is valid for the inexact line minimization of ψ. Using also Lemma 4.4 we now establish the convergence of the algorithm.

Theorem 5.13 (Convergence under the Armijo rule) [9] *Let the assumptions of Lemma 5.3 and Theorem 5.12.3 hold. Furthermore, let* $\nabla\mathbf{F}$ *and* $\nabla^2\varphi$ *be Lipschitz continuous on* X. *Then, under Rule A, the conclusions of Theorem 5.12.3 hold.*

Proof From Lemma 4.4 it follows that

$$\psi\left(\mathbf{x}^{k+1}\right) - \psi\left(\mathbf{x}^k\right) \leq -\frac{2\beta\alpha(1-\alpha)\mu^2}{M_{\nabla\psi}}\left\|\mathbf{d}^k\right\|^2,$$

where $\mu \stackrel{\text{def}}{=} m_{\mathbf{F}} + m_\varphi - M_{\nabla\varphi} > 0$. Since ψ is bounded from below, we conclude that

$$\left\{\mathbf{d}^k\right\} \to \mathbf{0}. \tag{5.29}$$

The existence of an accumulation point $\mathbf{x}^\infty$ of the sequence $\{\mathbf{x}^k\}$ is ensured by the compactness of either X or of the level sets of ψ. From (5.29), $\mathbf{x}^\infty = Y(\mathbf{x}^\infty)$, which by Theorem 5.5 implies that $\mathbf{x}^\infty \in \Omega$. The strong monotonicity of $\mathbf{F}$ then implies that $\mathbf{x}^\infty$ is the unique accumulation point. □

Remark 5.8 The subproblem solution $\mathbf{y}^{k+1}$ is obtained in the Armijo line search for t_k, and is therefore available at $\mathbf{x}^{k+1}$ with no extra computation.

The Armijo rule adopted here is slightly different from that used in [LaP90] (there, the term $\nabla\psi(\mathbf{x})^{\mathrm{T}}(\mathbf{y}(\mathbf{x})-\mathbf{x})$ in (A.37) is replaced by $-\|\mathbf{y}(\mathbf{x})-\mathbf{x}\|^2$), and actually requires slightly less restrictive assumptions on $\mathbf{F}$ and φ.

[9] From the same arguments as for Rule A, Rule P is valid, and yields global convergence provided that $t_k \leq t_k^{\max} \stackrel{\text{def}}{=} \max\{t \mid \mathbf{x}^k + t\mathbf{d}^k \in X\}$ and t_k belongs to a compact subset of $(0, \frac{2\mu}{M_{\nabla\psi}})$, $\forall k$. The difficulty of finding a measure of the constant $M_{\nabla\psi}$ is, however, even greater than for the corresponding constant $M_{\nabla f}$ in the symmetric case; we therefore consider this rule less useful in this context.

5.10 A truncated version of cost approximation algorithms

In this section, we validate truncation strategies for the solution of the subproblems. The idea is to truncate the solution of (5.13) prematurely, in the same fashion as that outlined in Section 4.10 for [**CNLP**], i.e., for some $\varepsilon > 0$ a point $\mathbf{y}_\varepsilon(\mathbf{x}) \in X$ is found, satisfying

$$\nabla_{\mathbf{y}} L\left(\mathbf{y}_\varepsilon(\mathbf{x}), \mathbf{x}\right)^{\mathrm{T}}\left(\mathbf{y} - \mathbf{y}_\varepsilon(\mathbf{x})\right) \leq \varepsilon, \ \forall \mathbf{y} \in X. \tag{5.30}$$

This implies, by the concavity of L with respect to $\mathbf{y}$, that $\psi(\mathbf{x}) \leq L\left(\mathbf{y}_\varepsilon(\mathbf{x}), \mathbf{x}\right) + \varepsilon$, i.e., that $\mathbf{y}_\varepsilon(\mathbf{x})$ is an ε-optimal solution to (5.13). The validity of (5.30) can be inspected in the way described in Section 4.10. In the next theorem, it is established that if ε is chosen small enough, then the direction $\mathbf{y}_\varepsilon(\mathbf{x}) - \mathbf{x}$ defines a descent direction for ψ.

Theorem 5.14 [LaP90] (Descent obtained from truncated subproblems) *Let the assumptions of Theorem 5.9.c be satisfied. Then, for any* $\mathbf{x} \in X$ *and* $\varepsilon \geq 0$*,*

$$\left\|\mathbf{y}(\mathbf{x}) - \mathbf{y}_\varepsilon(\mathbf{x})\right\|^2 \leq \frac{\varepsilon}{m_f}. \tag{5.31}$$

Further, if $\mathbf{x} \notin \Omega$*, (5.26) holds, and* $\varepsilon > 0$ *is chosen small enough (dependent of* $\mathbf{x}$*), then*

$$\nabla\psi(\mathbf{x})^{\mathrm{T}}\left(\mathbf{y}_\varepsilon(\mathbf{x}) - \mathbf{x}\right) < 0. \tag{5.32}$$

Remark 5.9 By further imposing Lipschitz continuity of φ on X, the term $\|\mathbf{y}(\mathbf{x}) - \mathbf{x}\|$ may be estimated both from above and below by bounded positive measures in $\mathbf{x}$, ε and $\mathbf{y}_\varepsilon(\mathbf{x})$ only. As shown in [LaP90], this enables (5.32) to be easily checked during the solution of (5.13), and the resulting estimates may therefore be used in a termination criterion for the solution of the subproblem. Further, these estimates tend to zero when $\mathbf{x}$ approaches Ω, so that, in effect, the conditions in the termination criteria amount to solving (5.13) accurately in the limit. Introducing a sequence $\{\varepsilon_k\}$, converging to zero, the global convergence of the overall scheme can be established from Theorem 5.12.

5.11 A comparison of convergence conditions

We begin by noting that, when the cost approximation method is used in its successive approximation version, i.e., step lengths $t_k \equiv 1$ are used, and convergence relies on contractive arguments, convergence is ensured by choosing the function φ so that it, in a sense, dominates the original mapping $\mathbf{F}$ [cf. (5.10)], while in the descent approach the convergence conditions amount to precisely the opposite [cf. (5.26)]. In both cases, convergence can be ensured from any choice of a strongly convex function $\varphi : X \mapsto \Re^n$ in C^1 on X, with a Lipschitz continuous gradient on X, by a proper scaling. However, the convergence conditions of the descent approach are weaker in general, see Theorems 5.12.1 and 5.12.2.

An interesting observation is that cost approximation methods based on an iteration dependent map $\mathbf{\Phi}$ (or φ) seem to have the inherent property of severe convergence conditions, which are also very difficult to verify in practice. This may be seen, for instance, from Condition (5.11) in the successive approximation version. Wu *et al.* [WFM90] develop a descent approach, based on the subproblem (5.9). [In their convergence analysis, however, the possible choices of functions h are much more limited.] Their basic convergence result requires that the function h is of the form $\frac{1}{2}(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{Q}(\mathbf{y})(\mathbf{x}-\mathbf{y})$, i.e., a direct generalization of Fukushima's choice of function φ, and, further, that the matrix $\nabla\mathbf{Q}(\mathbf{x})(\mathbf{y}(\mathbf{x})-\mathbf{x})$ is positive definite on X. In a second result, $\nabla_{\mathbf{x}}h(\mathbf{x},\mathbf{y})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) = -\nabla_{\mathbf{y}}h(\mathbf{x},\mathbf{y})^{\mathrm{T}}(\mathbf{x}-\mathbf{y})$ is required to hold for all $\mathbf{x},\mathbf{y}\in X$. The former requirement is very difficult to verify, while, in essence, the latter reduces the function h to be of the form used by Fukushima. We conclude that the conditions for convergence are, for iteration dependent maps $\mathbf{\Phi}$ and φ, severe, both for the successive approximation and descent approaches.

We next turn to Fukushima's [Fuk92] convergence analysis for the special case $\varphi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}$. For such a choice of φ it is possible to streamline the proofs of Theorems 5.8, 5.9, and 5.12; as a consequence, they can be proved under slightly milder conditions. Note, however, that the compactness assumption of X used by Fukushima is unnecessary (cf. Theorem 5.12.3). The Armijo-type step length rule used by Fukushima requires that its acceptance parameter α is chosen so that $\alpha < m_{\mathbf{F}}$. This requirement is, however, not necessary for the use of Rule A. Further, our development includes the possibility of truncating the subproblem solution process.

In a series of papers, Marcotte develops descent methods based on the primal gap function G [Mar85, MaD85, Mar86, MaD87, MaG88, DuM89, MaD89]. The primal gap function is used, in particular, to prove global convergence of a Newton-type approach, in which the Newton subproblem (see Table 5.1) defines a search direction for G. Convergence is shown under a monotonicity assumption on $\mathbf{F}$, and local quadratic convergence under standard strong monotonicity and regularity assumptions. Taji *et al.* [TFI93] extend the descent algorithm of Fukushima [Fuk92] to a Newton method based on the merit function ψ_{Fu}. The conditions on $\mathbf{F}$ are more restrictive than in the case of merit function G (strong monotonicity is required); on the other hand, the set X does not need to be bounded. In the method of [Mar85], search directions are generated through polyhedral approximations of the set of solutions to (5.21), in a similar way to ε-approximation schemes in nondifferentiable programming [Lem78]. A similar strategy could be used in the method established in Theorem 5.12.a.

5.12 Linear convergence of cost approximation algorithms

As noted in the discussion following Lemma 3.6, the linear convergence rate result proved for [**NNLP**] (as well as for [**CNLP**]), does not rely on the fact that ∇f is integrable. We state below the corresponding convergence result for variational inequalities.

Theorem 5.15 (Linear convergence of cost approximation algorithms) *Let $\mathbf{F}$ be strongly monotone in C^1 on X, and let $\mathbf{F}$ and $\nabla\mathbf{F}$ be Lipschitz continuous on X. Let $\varphi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}$, where $\mathbf{B}$ is a symmetric and positive definite matrix on X with eigenvalues in the interval $[m_\varphi, M_{\nabla\varphi}]$. Assume that, using Rule M, A, or Rule P with t_k chosen in a compact subset of $(0, \frac{2\mu}{M_{\nabla\psi}})$, $\forall k$, $\{\mathbf{x}^k\}\to\mathbf{x}^*\in\Omega$. Then if $\mathbf{B}$ is chosen so that*

$$\sigma \stackrel{\text{def}}{=} \frac{M_{\nabla\varphi}}{m_\varphi}\left\|\mathbf{I}-\mathbf{B}^{-1}\nabla\mathbf{F}(\mathbf{x}^*)\right\| < 1,$$

$\{\mathbf{x}^k\}$ *converges at least linearly with the ratio*

$$q \leq \max\left\{\sigma, 1 - \frac{\mu}{M_{\nabla\psi}}(1-\sigma)\right\}, \qquad \textit{for Rule M,}$$

$$q \leq \max\left\{\sigma, 1 - \frac{2\beta(1-\alpha)\mu}{M_{\nabla\psi}}(1-\sigma)\right\}, \quad \textit{for Rule A, and}$$

$$q \leq \max\left\{\sigma, 1 - \underline{t}(1-\sigma)\right\}, \qquad \textit{for Rule P,}$$

where $\mu = m_{\mathbf{F}} + m_{\varphi} - M_{\nabla\varphi} > 0$ *and* $\underline{t} = \inf_k\{t_k\} > 0$.

Proof Follows from the same line of argument as in the proof of Theorem 4.5. □

The theorem may be used to establish linear convergence of Fukushima's [Fuk92] special case of gap decreasing algorithm.[10]

5.13 Finite identification of active constraints

We shall here briefly describe extensions of the results of Section 4.12 to [**VIP**].

The notion of a projected gradient [see (4.29)] is here extended to

$$\mathbf{F}^X(\mathbf{x}) \stackrel{\text{def}}{=} \arg\min_{\mathbf{v}\in T_X(\mathbf{x})} \|\mathbf{F}(\mathbf{x}) + \mathbf{v}\|, \ \mathbf{x} \in X. \tag{5.33}$$

Replacing ∇f with $\mathbf{F}$ in the analysis made for [**CNLP**], we immediately have

$$\mathbf{F}^X(\mathbf{x}) = \mathbf{0} \Leftrightarrow \mathbf{x} \in \Omega,$$

and the following results follow easily from Lemma 4.6 and Theorem 4.6, respectively.

Lemma 5.4 (The cost approximation algorithms force $\mathbf{F}^X$ to zero) *Let* $\mathbf{\Phi} : X \times X \mapsto \Re^n$ *be a continuous function on* $X \times X$ *of the form* $\mathbf{\Phi}(\mathbf{x}, \mathbf{y})$, *monotone on* X *with respect to* $\mathbf{x}$. *Further, let* $\mathbf{F}$ *and* $\mathbf{\Phi}$ *be uniformly continuous on* X *and* $X \times X$, *respectively. Let* $\{\mathbf{x}^k\} \to \mathbf{x}^*$, *and assume that (5.29) holds. Then,*

$$\left\{\mathbf{F}^X\left(\mathbf{y}^k\right)\right\} \to \mathbf{0}. \tag{5.34}$$

Theorem 5.16 (Finite identification of optimal face and active constraints) *Let* $\mathbf{\Phi}$: $X \times X \mapsto \Re^n$ *be a continuous function on* $X \times X$ *of the form* $\mathbf{\Phi}(\mathbf{x}, \mathbf{y})$, *monotone on* X *with respect to* $\mathbf{x}$. *Further, let* $\mathbf{F}$ *and* $\mathbf{\Phi}$ *be uniformly continuous on* X *and* $X \times X$, *respectively. Let* $\{\mathbf{x}^k\} \to \mathbf{x}^*$, *and assume that (5.29) holds. Assume further that* $\mathbf{x}^* \in \Omega$ *and that*

$$-\mathbf{F}(\mathbf{x}^*) \in \texttt{rint}\ N_X(\mathbf{x}^*). \tag{5.35}$$

(a) *If* $\mathbf{x}^* \in \texttt{rint}\ X_F$, *where* X_F *is a quasi-polyhedral face of* X, *then* $\mathbf{y}^k \in \texttt{rint}\ X_F$ *for all* k *sufficiently large.*

[10] He establishes convergence toward the unique solution for Rule M and an Armijo step length rule (which actually requires slightly stronger assumptions than Rule A, since the size of the parameter α in the Armijo rule must be taken into account).

(b) *Assume further that X is polyhedral. Then $\mathcal{I}(\mathbf{y}^k) = \mathcal{I}(\mathbf{x}^*)$ for all k sufficiently large.*

Note that, from Section 4.12, the identification results of [DuM89, MaD89] define special cases of those given in Theorem 5.16.

If the sequence $\{\mathbf{x}^k\}$ identifies the optimal face finitely, the cost approximation algorithm reduces to its unconstrained version (i.e., for nonlinear equations) on the manifold of active constraints. We give below sufficient conditions for this result to hold.

Corollary 5.2 (Finite identification of active constraints) *Under the assumptions of Theorem 5.16.b, by further assuming that*

$$\mathcal{I}\left(\mathbf{x}^k\right) \supseteq \mathcal{I}\left(\mathbf{y}^k\right)$$

holds for some k large enough for $\mathcal{I}(\mathbf{y}^k) = \mathcal{I}(\mathbf{x}^*)$ to hold, and $t_k \leq 1$ for all subsequent k, then

$$\mathcal{I}\left(\mathbf{x}^k\right) = \mathcal{I}(\mathbf{x}^*)$$

also holds for all such k.

Proof Follows from the same line of arguments as in the proof of Corollary 4.2. □

The regularity conditions given in Definition A.8 and related to each other in Lemma 4.7 can all be cast in terms of variational inequalities. For all but Condition **(c)**, this fact is immediate from the definitions, or from the lemma. In order to extend Condition **(c)** to variational inequalities, we can use the result of [BuF92, Cor. 2.7.c] that Ω is a set of weak sharp minima of f over X if and only if, for an arbitrary $\mathbf{x}^* \in \Omega$,

$$-\nabla f(\mathbf{x}^*) \in \mathtt{int} \bigcap_{\mathbf{x} \in \Omega} \left[T_X(\mathbf{x}) \cap N_\Omega(\mathbf{x})\right]^\circ .$$

(This inclusion reduces to (4.40) whenever Ω is a singleton.) Replacing ∇f with $\mathbf{F}$, we may use this inclusion as a definition of *weak sharp solutions* to variational inequalities. With these extended definitions at hand, it follows immediately that Lemma 4.7 extends to [**VIP**].

5.14 Finite convergence of cost approximation algorithms

Here, we extend the finite convergence results obtained for [**CNLP**] in Section 4.13.

The following result is a direct generalization of Theorem 4.7.

Theorem 5.17 (Finite convergence of cost approximation algorithms) *Let $\mathbf{\Phi} : X \times X \mapsto \Re^n$ be a continuous function on $X \times X$ of the form $\mathbf{\Phi}(\mathbf{x}, \mathbf{y})$, monotone on X with respect to $\mathbf{x}$. Let $\{\mathbf{x}^k\} \to \mathbf{x}^*$, and assume that $\mathbf{x}^*$ is a sharp solution, i.e.,*

$$-\mathbf{F}(\mathbf{x}^*) \in \mathtt{int}\, N_X(\mathbf{x}^*) . \tag{5.36}$$

(a) *Let $\mathbf{F}$ and $\mathbf{\Phi}$ be uniformly continuous on X and $X \times X$, respectively, and assume that (5.29) holds. Then $\mathbf{y}^k = \mathbf{x}^*$ holds for all sufficiently large k.*

(b) *Let* $\mathbf{F}$ *and* $\mathbf{\Phi}$ *be Lipschitz continuous on* X *and* $X \times X$*, respectively. Then* $\mathbf{y}^k = \mathbf{x}^*$ *holds for all sufficiently large* k*, such that*

$$\left\| \mathbf{x}^k - \mathbf{x}^* \right\| \leq \frac{\alpha}{M_{\mathbf{\Phi}} + M_{\mathbf{F}}}.$$

In [AlK91], it is shown that under the Condition (5.36), by invoking a Frank-Wolfe type subproblem regularly in a convergent algorithm, finite convergence is obtained. Clearly, this result can be extended to the inclusion of *any* cost approximation subproblem, satisfying the conditions of Theorem 5.17, in a convergent algorithm.

The finite convergence Theorem 4.9 for [**CNLP**], based on the weak sharp minimum property, was established by applying the analysis in [BuF92, Th. 4.7]. It is, however, not possible to extend this result to variational inequalities with weak sharp solutions, since the proof of the latter result utilizes the invariance of ∇f on Ω, which does not extend to asymmetric variational inequalities (see Remark 5.4).

5.15 Jacobi and Gauss-Seidel type methods

When the feasible set X is of the form (4.44), the cost approximating mapping $\mathbf{\Phi}$ can, as is demonstrated in the previous chapters, be chosen to match this structure. If the cost approximating function $\mathbf{\Phi} : X \mapsto \Re^n$ is of the form $\mathbf{\Phi}_i(\mathbf{x}) = \mathbf{\Phi}_i(\mathbf{x}_i)$, then the variational inequality subproblem [$\mathbf{VIP}_{\mathbf{\Phi}}$] decomposes into $|\mathcal{C}|$ independent variational inequalities. The corresponding choice of φ is the additively separable form

$$\varphi(\mathbf{x}) = \sum_{i \in \mathcal{C}} \varphi_i(\mathbf{x}_i). \tag{5.37}$$

In the latter case, the evaluation of $\psi(\mathbf{x})$ amounts to solving $|\mathcal{C}|$ convex problems. Since such a decomposition takes place in the line search phase also, the computations involved in the descent methods developed in this chapter could effectively utilize parallel computing environments.

A variety of important applications are variational inequality problems over product sets, including traffic equilibrium problems [Daf80, Pat90, Nag93], spatial economic equilibrium [FaP80, ToF83, DaN87], and Nash equilibrium [Kar69b, GaM80, Coh87]; see also [Pan85, BeT89, HaP90b]. Further, any variational inequality defined on a polyhedral set, where the map $\mathbf{F}$ is invertible, can be cast in a dual form, where the feasible set is a product set [Mos72, Pie84, Pan85].

For strongly monotone variational inequalities, we demonstrate below that the decomposition algorithm developed in Sections 2.7, 3.12, and 4.15 can be extended to the gap decreasing algorithms of this chapter.

Assume that, in iteration k, index $i_k \in \mathcal{C}$ is chosen. We then evaluate $\psi(\mathbf{x}^k)$ by solving (5.13), keeping the variables $\mathbf{x}_j$, $j \neq i_k$, fixed. [This problem is equivalent to [$\mathbf{CNLP}^k_{\varphi_{i_k}}$], with ∇f replaced by $\mathbf{F}$ and $\varphi_i(\mathbf{x}_i, \mathbf{y}) \equiv \varphi_i(\mathbf{x}_i)$.] Letting $\mathbf{y}_j^k = \mathbf{x}_j^k$ for all $j \neq i_k$, the search direction for then is $\mathbf{y}^k - \mathbf{x}^k$. A line search with respect to ψ gives a step length t_k, and

$$\mathbf{x}_j^{k+1} = \begin{cases} \mathbf{x}_j^k + t_k \left(\mathbf{y}_j^k - \mathbf{x}_j^k \right), & j = i_k \\ \mathbf{x}_j^k, & j \neq i_k. \end{cases}$$

In order to ensure that the search directions are descent directions for ψ, we add the following assumptions. Assume that there exists a constant $m_{\mathbf{F}_i} > 0$ such that

$$\mathbf{y}_i^{\mathrm{T}} \left(\nabla_i \mathbf{F}(\mathbf{x}) - m_{\mathbf{F}_i} \mathbf{I} \right) \mathbf{y}_i \geq 0, \ \forall \mathbf{x} \in X, \ \forall \mathbf{y}_i \in \Re^{n_i}, \tag{5.38}$$

where $\nabla_i \mathbf{F}$ is the $n_i \times n_i$ matrix of partial derivatives of $\mathbf{F}$ with respect to $\mathbf{x}_i$ [compare (A.18)]. We further assume that each function φ_i is strongly convex in C^2 on X_i, with a Lipschitz continuous gradient on X_i. It is then straightforward to prove that the following holds for each $i \in \mathcal{C}$.

$$\nabla \psi(\mathbf{x})^{\mathrm{T}}(\mathbf{y}(\mathbf{x}) - \mathbf{x}) = \nabla_i \psi(\mathbf{x})^{\mathrm{T}}(\mathbf{y}_i(\mathbf{x}) - \mathbf{x}_i) \leq -\left(m_{\mathbf{F}_i} + m_{\varphi_i} - M_{\nabla \varphi_i} \right) \|\mathbf{y}_i(\mathbf{x}) - \mathbf{x}_i\|^2$$

Hence, choosing φ_i such that

$$M_{\nabla \varphi_i} - m_{\varphi_i} < m_{\mathbf{F}_i} \tag{5.39}$$

holds, ensures that $\mathbf{y}(\mathbf{x}) - \mathbf{x}$ is a direction of descent for ψ. (Note that if $|\mathcal{C}| = 1$, the Inequality (5.39) reduces to (5.26), and the algorithm reduces to the basic one.)

The algorithm will be proven convergent under Rules M, A, and P, using the essentially cyclic rule; the below result is an extension of Theorem 4.14 from [**CNLP**].

Theorem 5.18 (Convergence of essentially cyclic cost approximation algorithms) *Let $\mathbf{F}$ be in C^1 on X, where X is defined by (4.44). Let further $\mathbf{F}$ and $\nabla \mathbf{F}$ be Lipschitz continuous on X, and let (5.38) hold for each $i \in \mathcal{C}$. Let $\varphi : X \mapsto \Re$ be given by (5.37), where each function $\varphi_i : X_i \mapsto \Re$ is a strongly convex function in C^2 on X_i, and let $\nabla \varphi_i$ and $\nabla^2 \varphi_i$ be Lipschitz continuous on X_i. Further, let (5.39) hold for each $i \in \mathcal{C}$, and assume that either X is bounded or $M_{\nabla \varphi_i} < 2m_{\mathbf{F}}$ holds for each $i \in \mathcal{C}$. Let the sequence $\{\mathbf{x}^k\}$ be defined by the above decomposition based gap decreasing algorithm, where indices are chosen according to the essentially cyclic rule (4.49). Then, from any starting point $\mathbf{x}^0 \in X$, under Rules M and A the sequence $\{\mathbf{x}^k\}$ converges to the unique solution $\mathbf{x}^*$ of* [**VIP**]*. For Rule P, the same conclusion holds if $t_k^i \leq t_k^{\max}$ and t_k^i belongs to a compact subset of $(0, \frac{2\mu_i}{M_{\nabla_i \psi}})$, where $\mu_i \overset{\text{def}}{=} m_{\mathbf{F}_i} + m_{\varphi_i} - M_{\nabla \varphi_i} > 0$, $\forall i \in \mathcal{C}$.*

Proof From the assumptions, the Rules M, A, and P are valid. Replacing f with ψ in the proof of Theorem 4.14, the result follows. $\square$

5.16 Summary and further research

In this chapter we have considered the class of cost approximation methods for variational inequalities. Interesting algorithmic equivalence results were obtained; in particular, Dafermos' framework was shown to be equivalent to a successive cost approximation algorithm. Furthermore, many well known iterative methods based on successive approximation were shown to be instances of cost approximation algorithms. In the development of descent methods for [**VIP**], we demonstrated that any variational inequality can be equivalently reformulated as a differentiable, and in some cases convex, nonlinear program; these optimization formulations arise very naturally from symmetric cost approximations, and unify previously proposed reformulations of asymmetric variational inequalities as nonlinear programs.

As is shown in this chapter, as well as in previous works, descent methods based on gap functions require, in general, less restrictive assumptions than their successive approximation counterparts. Gap functions therefore constitute an interesting way of extending the range of applicability of traditional methods for variational inequalities; this is a main motive for considering methods involving gap functions. The descent methods developed in this chapter provide a unification of hitherto proposed descent methods. Furthermore, the results obtained extend earlier ones in several ways. In particular, we show that descent methods can be applied to problems defined on unbounded sets. The methods can be applied in an inexact fashion, in the sense that the search direction generation or line search step can be performed inexactly. To the best of our knowledge, the investigation made in Section 5.15 is the first study of decomposition type descent methods for variational inequalities. Such methods are of practical interest, due to the large number of applications of [**VIP**] to problems over Cartesian product sets.

A question for further study regards the properties of ψ. There are, seemingly, connections between the descent, convexity and coercivity properties of ψ that are not covered by existing theory, and it would be interesting to try to characterize these connections. A related, interesting topic for further studies is the relationships between [**VIP**], saddle point problems (see Section 5.1 and Theorems 5.3 and 5.11), and Fenchel's inequality (see Lemma 5.2), that are brought forward in the analysis of merit functions.

Harker [Har88] introduces an acceleration step, which corresponds to a line search with respect to the variational inequality, in the Jacobi and projection methods. This strategy could be seen as an alternative to the more costly line search with respect to a gap function; is it possible to extend this strategy to cost approximation methods other than those considered by Harker?

It is shown (Theorem 5.15) that if φ is chosen as a positive definite quadratic form, then linear convergence is achieved. It would be interesting to investigate the convergence rates obtained from non-quadratic functions φ, in particular in connection with truncated subproblems. The combination of truncated subproblems and inexact line searches should also be investigated. Also, the performance of cost approximation methods using (inexact) line searches should be compared to that of successive approximation methods.

Two special cases of gap functions have been used to establish global convergence in connection with Newton's method, [MaD85, MaD87, MaG88, DuM89, MaD89, TFI93]. Is it possible to extend these results to a more general choice of gap function?

The optimization formulation [$\mathbf{CNLP}_\psi$] is defined through a minimax formulation. It is therefore possible to consider methods for the minimization of the merit function ψ, based on minimax or saddle point algorithms. Other possibilities include gradient methods based on $\nabla\psi$.

For affine variational inequalities, error bounds (local and global) have been obtained assuming only monotonicity of $\mathbf{F}$ [LuT92a, LuT92b, LuT92d, Man92]; see also [Rob81]. These error bounds have been used to prove linear convergence of matrix splitting methods for these problems. The error bounds proposed so far for nonlinear maps $\mathbf{F}$ require strong monotonicity (see Theorem 5.10 and [Pan86a, Pan87, TFI93]). A successful attempt to weaken this assumption could lead to a similar weakening of the convergence conditions and strengthening of the convergence rate results for many of the iterative methods discussed in this chapter.

An interesting observation is that gap functions can be used to establish the existence of (or detect the lack of) a solution to a variational inequality. This would constitute an alternative approach to that of Noor for establishing the existence of a solution. In the case of nonexistence of a solution, the property of the minimizers of a gap function can perhaps be used as an instrument to learn more about the nature of the underlying problem.

An interesting extension of the gap decreasing algorithms of this chapter is their application to [**GVIP**]. If, in [**GVIP**], the point-to-set map $\mathbf{U}$ is a subdifferential mapping of a convex function, it is possible to show (using the same arguments as in Lemma 5.2) that the convex subproblem (1.11) defines a gap function for [**GVIP**]. We conjecture that gap decreasing algorithms based on this subproblem can be shown to yield convergence under appropriate assumptions. Such results would imply that well known methods for instances of [**GVIP**], such as splitting algorithms, could be furnished with accelerating line search steps with respect to a merit function.

Another possible extension of gap decreasing algorithms is to quasi-variational inequality problems, in which an $\mathbf{x}^* \in X(\mathbf{x}^*)$ is sought such that

$$[\mathbf{QVIP}] \qquad \mathbf{F}(\mathbf{x}^*)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^*) \geq 0, \ \forall \mathbf{x} \in X(\mathbf{x}^*)$$

(see [BGL73, BeL73, Ben74, BaC84, Har91]), for which few algorithms have been presented.

The combined extension, the generalized quasi-variational inequality problem [ChP82, Yao91], should also be studied in this context.

We finally remark that there are still few applications of gap functions in descent methods for variational inequalities. The potential use of such methods to enhance the convergence, as well as weaken the convergence conditions, of the traditional successive approximation schemes, indicate that these new methods deserve much further attention.

Appendix A

Definitions

In this appendix we collect the definitions and abbreviations of various concepts used in the thesis.

Properties of functions

Definition A.1 (Convexity) *Let X be a nonempty, closed and convex set in $\Re^n$, and let f be a function from X to $\Re$.*

(a) *$f \in C^1$ on X is pseudoconvex on X if*

$$\nabla f(\mathbf{y})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) \geq 0 \Rightarrow f(\mathbf{x}) \geq f(\mathbf{y}),\ \forall \mathbf{x}, \mathbf{y} \in X. \tag{A.1}$$

(b) *f is convex on X if*

$$f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y}),\ \forall \mathbf{x}, \mathbf{y} \in X, \forall \lambda \in [0,1]. \tag{A.2}$$

If $f \in C^1$ on X, then (A.2) is equivalent to both the following statements (see [BaS79, Th. 3.3.3] and [BaS79, Th. 3.3.4], respectively).

$$f(\mathbf{x}) \geq f(\mathbf{y})+\nabla f(\mathbf{y})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}),\ \forall \mathbf{x}, \mathbf{y} \in X \tag{A.3}$$

$$[\nabla f(\mathbf{x})-\nabla f(\mathbf{y})]^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) \geq 0,\ \forall \mathbf{x}, \mathbf{y} \in X \tag{A.4}$$

If $f \in C^2$ on X, then (A.2) is equivalent to the following [BaS79, Th. 3.3.6].

$$\mathbf{y}^{\mathrm{T}}\nabla^2 f(\mathbf{x})\mathbf{y} \geq 0,\ \forall \mathbf{x} \in X, \forall \mathbf{y} \in \Re^n \tag{A.5}$$

(c) *f is strictly convex on X if*

$$f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y}),\ \forall \mathbf{x}, \mathbf{y} \in X, \mathbf{x} \neq \mathbf{y}, \forall \lambda \in (0,1). \tag{A.6}$$

If $f \in C^1$ on X, then (A.6) is equivalent to both the following statements (see [BaS79, Th. 3.3.3] and [BaS79, Th. 3.3.4], respectively).

$$f(\mathbf{x}) > f(\mathbf{y})+\nabla f(\mathbf{y})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}),\ \forall \mathbf{x}, \mathbf{y} \in X, \mathbf{x} \neq \mathbf{y} \tag{A.7}$$

$$[\nabla f(\mathbf{x})-\nabla f(\mathbf{y})]^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) > 0,\ \forall \mathbf{x}, \mathbf{y} \in X, \mathbf{x} \neq \mathbf{y} \tag{A.8}$$

(d) *f is strongly (uniformly) convex (with modulus m_f) on X if there exists a positive constant m_f such that*

$$f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y})-\frac{m_f}{2}\lambda(1-\lambda)\left\|\mathbf{x}-\mathbf{y}\right\|^2, \tag{A.9}$$

$\forall \mathbf{x},\mathbf{y} \in X, \forall \lambda \in [0,1]$.

If $f \in C^1$ on X, then (A.9) is equivalent to both the following statements (see [Pol87, Sec. 1.1.4, Le. 3] and [LeP66], respectively).

$$f(\mathbf{x}) \geq f(\mathbf{y})+\nabla f(\mathbf{y})^{\mathrm{T}}(\mathbf{x}-\mathbf{y})+\frac{m_f}{2}\left\|\mathbf{x}-\mathbf{y}\right\|^2,\ \forall \mathbf{x},\mathbf{y} \in X \tag{A.10}$$

$$\left[\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\right]^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) \geq m_f\left\|\mathbf{x}-\mathbf{y}\right\|^2,\ \forall \mathbf{x},\mathbf{y} \in X \tag{A.11}$$

If $f \in C^2$ on X, then (A.9) is equivalent to the following [Pol87, Sec. 1.1.4].

$$\mathbf{y}^{\mathrm{T}}\left(\nabla^2 f(\mathbf{x})-m_f\mathbf{I}\right)\mathbf{y} \geq 0,\ \forall \mathbf{x} \in X, \forall \mathbf{y} \in \Re^n \tag{A.12}$$

Definition A.2 (Monotonicity) *Let X be a nonempty, closed and convex set in $\Re^n$, and $\mathbf{F}$ a function from X to $\Re^n$.*

(a) *$\mathbf{F}$ is pseudomonotone on X if*

$$\mathbf{F}(\mathbf{y})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) \geq 0 \Rightarrow \mathbf{F}(\mathbf{x})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) \geq 0,\ \forall \mathbf{x},\mathbf{y} \in X. \tag{A.13}$$

(b) *$\mathbf{F}$ is monotone on X if*

$$\left[\mathbf{F}(\mathbf{x})-\mathbf{F}(\mathbf{y})\right]^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) \geq 0,\ \forall \mathbf{x},\mathbf{y} \in X. \tag{A.14}$$

If $\mathbf{F}$ is in C^1 on X, then (A.14) is equivalent to the following [OrR70, Th. 5.4.3].

$$\mathbf{y}^{\mathrm{T}}\nabla\mathbf{F}(\mathbf{x})\mathbf{y} \geq 0,\ \forall \mathbf{x} \in X, \forall \mathbf{y} \in \Re^n \tag{A.15}$$

(c) *$\mathbf{F}$ is strictly monotone on X if*

$$\left[\mathbf{F}(\mathbf{x})-\mathbf{F}(\mathbf{y})\right]^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) > 0,\ \forall \mathbf{x},\mathbf{y} \in X, \mathbf{x} \neq \mathbf{y}. \tag{A.16}$$

(d) *$\mathbf{F}$ is strongly (uniformly) monotone on X if there exists a positive constant $m_{\mathbf{F}}$ such that*

$$\left[\mathbf{F}(\mathbf{x})-\mathbf{F}(\mathbf{y})\right]^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) \geq m_{\mathbf{F}}\left\|\mathbf{x}-\mathbf{y}\right\|^2,\ \forall \mathbf{x},\mathbf{y} \in X. \tag{A.17}$$

If $\mathbf{F}$ is in C^1 on X, then (A.17) is equivalent to the following [OrR70, Th. 5.4.3].

$$\mathbf{y}^{\mathrm{T}}\left(\nabla\mathbf{F}(\mathbf{x})-m_{\mathbf{F}}\mathbf{I}\right)\mathbf{y} \geq 0,\ \forall \mathbf{x} \in X, \forall \mathbf{y} \in \Re^n \tag{A.18}$$

Definition A.3 (Coercivity) *Let X be a nonempty, closed and convex set in $\Re^n$. Let f be a function from X to $\Re$ and $\mathbf{F}$ a function from X to $\Re^n$.*

(a) *f is weakly coercive on X if*

$$\lim_{\substack{\mathbf{x}\in X \\ \|\mathbf{x}\|\to+\infty}} f(\mathbf{x}) = +\infty. \tag{A.19}$$

(b) *f is coercive on X if*

$$\lim_{\substack{\mathbf{x}\in X \\ \|\mathbf{x}\|\to+\infty}} \frac{f(\mathbf{x})}{\|\mathbf{x}\|} = +\infty. \tag{A.20}$$

(c) *$\mathbf{F}$ is coercive on X if there exists a vector $\mathbf{x}^0 \in X$ such that*

$$\lim_{\substack{\mathbf{x}\in X \\ \|\mathbf{x}\|\to+\infty}} \frac{\mathbf{F}(\mathbf{x})^{\mathrm{T}}(\mathbf{x}-\mathbf{x}^0)}{\|\mathbf{x}\|} = +\infty. \tag{A.21}$$

Definition A.4 (Lipschitz continuity) *A function $f : X \mapsto \Re^m$ is Lipschitz continuous (with modulus M_f) on X if there exists a nonnegative constant M_f such that*

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq M_f \|\mathbf{x}-\mathbf{y}\|, \ \forall \mathbf{x},\mathbf{y} \in X. \tag{A.22}$$

Definition A.5 (Nonexpansiveness) *Let X be a nonempty, closed and convex set in $\Re^n$. Let f be a mapping from X to X.*

(a) *The mapping f is nonexpansive if*

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq \|\mathbf{x}-\mathbf{y}\|, \ \forall \mathbf{x},\mathbf{y} \in X. \tag{A.23}$$

(b) *The mapping f is firmly nonexpansive if*

$$\|f(\mathbf{x}) - f(\mathbf{y})\|^2 \leq [f(\mathbf{x}) - f(\mathbf{y})]^{\mathrm{T}}(\mathbf{x}-\mathbf{y}), \ \forall \mathbf{x},\mathbf{y} \in X. \tag{A.24}$$

(c) *The mapping f is contractive with modulus $\alpha \in [0,1)$ if*

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq \alpha \|\mathbf{x}-\mathbf{y}\|, \ \forall \mathbf{x},\mathbf{y} \in X. \tag{A.25}$$

Definition A.6 [Roc70a] (Semicontinuity) *Let $f : \Re^n \mapsto \Re \cup \{-\infty,+\infty\}$ be a given function.*

(a) *f is lower semicontinuous (l.s.c.) on $\Re^n$ if*

$$f(\mathbf{x}) = \liminf_{\mathbf{y}\to\mathbf{x}} f(\mathbf{y}), \ \forall \mathbf{x} \in \Re^n.$$

(b) *f is upper semicontinuous (u.s.c.) on $\Re^n$ if*

$$f(\mathbf{x}) = \limsup_{\mathbf{y}\to\mathbf{x}} f(\mathbf{y}), \ \forall \mathbf{x} \in \Re^n.$$

Definition A.7 [Nik68, Hog73] (Closedness) *Let $f : X \mapsto 2^Y$ be a point-to-set map. f is closed at $\mathbf{x} \in X$ if*

$$\left.\begin{array}{r} \{\mathbf{x}^k\} \to \mathbf{x} \\ \mathbf{y}^k \in f\left(\mathbf{x}^k\right), \{\mathbf{y}^k\} \to \mathbf{y} \end{array}\right\} \Rightarrow \mathbf{y} \in f(\mathbf{x}).$$

If f is closed at every $\mathbf{x} \in X$, f is said to be closed on X.

Properties of optimal sets

Several researchers have developed regularity conditions on the problem data, under which convergent algorithms for constrained optimization problems either finitely identify the optimal active constraint set, or converge finitely. Below we list some regularity conditions for the differentiable program [**CNLP**] which will be discussed in the context of finiteness properties of cost approximation methods. Their relationships are established in Lemma 4.7.

Definition A.8 (Regularity conditions) *Let $f : \Re^n \mapsto \Re$ be a convex function in C^1 on X, where X is a nonempty, closed and convex set in $\Re^n$. Let Ω be the (nonempty) set of solutions to* [**CNLP**]*, and let $\mathbf{x}^* \in \Omega$. Further, let*

$$W(\mathbf{x}) = \arg\min_{\mathbf{y}\in X} \nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{y},$$

$$a(\sigma) = \inf_{\substack{\mathbf{y}\in X \\ \|\mathbf{y}-\mathbf{x}^*\|\geq\sigma}} \nabla f(\mathbf{x}^*)^{\mathrm{T}}\,\mathbf{y}. \tag{A.26}$$

(a) $\mathbf{x}^*$ *is a nondegenerate minimum [Dun87] if*

$$-\nabla f(\mathbf{x}^*) \in \mathtt{rint}\; N_X(\mathbf{x}^*). \tag{A.27}$$

(b) $\mathbf{x}^*$ *satisfies a minimum principle sufficiency condition [FeM92] if*

$$W(\mathbf{x}^*) \subseteq \Omega. \tag{A.28}$$

(c) $\mathbf{x}^*$ *is a weak sharp minimum [Pol69, Pol87, Fer88, Fer91] if, for some $\alpha > 0$,*

$$f(\mathbf{x}) \geq f^* + \alpha\,\|\mathbf{x} - P_\Omega(\mathbf{x})\|\,, \quad \forall \mathbf{x} \in X. \tag{A.29}$$

(d) $\mathbf{x}^*$ *is a sharp minimum [Pol79, Pol87] if, for some $\alpha > 0$,*

$$f(\mathbf{x}) \geq f^* + \alpha\,\|\mathbf{x} - \mathbf{x}^*\|\,, \quad \forall \mathbf{x} \in X. \tag{A.30}$$

(e) $\mathbf{x}^*$ *is nonsingular [Dun79] if*

$$W(\mathbf{x}^*) = \{\mathbf{x}^*\}\,. \tag{A.31}$$

(f) $\mathbf{x}^*$ *is strongly nonsingular [Dun79] if (A.31) holds, and*

$$a(\sigma) > 0, \quad \forall \sigma > 0. \tag{A.32}$$

(g) $\mathbf{x}^*$ *is strongly regular [Dun79] if (A.31) holds, and, for some $\alpha > 0$,*

$$a(\sigma) \geq \alpha\sigma, \quad \forall \sigma > 0. \tag{A.33}$$

Let further $X = \{\mathbf{x} \in \Re^n \mid g_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m\}$, where each $g_i : \Re^n \mapsto \Re$ is in C^1, and let $\mathcal{I}(\mathbf{x})$ be the index set of binding constraints at $\mathbf{x}$.

(h) $\mathbf{x}^*$ *satisfies the Guignard constraint qualification [Gui69] if*

$$T_X(\mathbf{x}^*) = \left\{\mathbf{v} \in \Re^n \;\middle|\; \mathbf{v}^{\mathrm{T}}\nabla g_i(\mathbf{x}^*) \geq 0,\ i \in \mathcal{I}(\mathbf{x}^*)\right\}. \tag{A.34}$$

(i) $\mathbf{x}^*$ *satisfies the linear independence constraint qualification [ArU60] if* $\nabla g_i(\mathbf{x}^*)$, $i \in \mathcal{I}(\mathbf{x}^*)$, *are linearly independent.*

(j) $\mathbf{x}^*$ *satisfies the strict complementarity slackness condition if there is a set of optimal Lagrange multipliers with* $\lambda_i^* > 0$, $\forall i \in \mathcal{I}(\mathbf{x}^*)$.

Let further X *be polyhedral, i.e.,* $g_i(\mathbf{x}) = \mathbf{a}_i^{\mathrm{T}}\mathbf{x} - b_i$, $\forall i$.

(k) $\mathbf{x}^*$ *is geometrically stable [MaD89] if*

$$\nabla f(\mathbf{x}^*)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^*) = 0 \Rightarrow \mathbf{x} \in X_{F^*}. \tag{A.35}$$

(l) $\mathbf{x}^*$ *is geometrically regular [DuM89] if (A.27) holds, and*

$$\Omega \subseteq \texttt{rint}\ X_{F^*}. \tag{A.36}$$

Algorithmic definitions

The following line search rules will be used in the convergence analysis (listed in the order of increasing computational simplicity). Assume that $\mathbf{d}^k$ is the search direction in iteration k, and that t_k is the chosen step length. The line search rules are given for the case where the feasible set $X = \Re^n$. When the set X is bounded in the direction $\mathbf{d}^k$ (assumed locally feasible), an upper bound on t_k must be introduced.

Rule M (Exact minimization) Choose t_k such that

$$f\left(\mathbf{x}^k + t_k\mathbf{d}^k\right) = \min_{t \geq 0} f\left(\mathbf{x}^k + t\mathbf{d}^k\right).$$

Rule G (Goldstein [Gol62]) Let $0 < \mu_1 \leq \mu_2 < 1$[1] and choose t_k such that

$$\mu_1 \leq \frac{f\left(\mathbf{x}^k + t_k\mathbf{d}^k\right) - f\left(\mathbf{x}^k\right)}{t_k \nabla f\left(\mathbf{x}^k\right)^{\mathrm{T}} \mathbf{d}^k} \leq \mu_2.$$

Rule A (Armijo [Arm66]) Let $\alpha \in (0,1)$ and $t_k = \varepsilon\beta^{\bar{\imath}}$, where $\beta \in (0,1), \varepsilon > 0$ and $\bar{\imath}$ is the smallest nonnegative integer i such that

$$\alpha \leq \frac{f\left(\mathbf{x}^k + \varepsilon\beta^i\mathbf{d}^k\right) - f\left(\mathbf{x}^k\right)}{\varepsilon\beta^i \nabla f\left(\mathbf{x}^k\right)^{\mathrm{T}} \mathbf{d}^k}. \tag{A.37}$$

Rule N (Newton step) Choose $t_k = -\dfrac{\nabla f\left(\mathbf{x}^k\right)^{\mathrm{T}} \mathbf{d}^k}{\mathbf{d}^{k\mathrm{T}} \nabla^2 f\left(\mathbf{x}^k\right) \mathbf{d}^k}$.

Rule P (Predetermined steps) Let $0 < \omega < \Omega$. Choose t_k arbitrary in the interval $[\omega, \Omega - \omega]$. (A particular example is the fixed step length formula, $t_k = t \in (0, \Omega), \forall k$.)

---

[1] An important special case results from using $\mu_1 = \alpha$ and $\mu_2 = 1 - \alpha$, where $\alpha \in (0, \frac{1}{2})$. See also [OrR70, Sec. 8.3].

Definition A.9 [BaS79, Lue84] (Convergence rate) *Let the sequence* $\{\mathbf{x}^k\}$ *converge to* $\mathbf{x}^*$. *The rate of convergence of the sequence is the supremum of the nonnegative numbers* p *satisfying*

$$\limsup_{k\to\infty} \frac{\left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\|}{\left\|\mathbf{x}^k - \mathbf{x}^*\right\|^p} = q < +\infty. \tag{A.38}$$

(a) *If* $p = 1$ *and* $q < 1$, *the sequence has a linear rate of convergence (or the rate of geometrical progression) with ratio* q.

(b) *If* $p > 1$, *or if* $p = 1$ *and* $q = 0$, *the sequence has a superlinear rate of convergence (or faster than any geometric progression).*

(c) *If* $p = 2$ *the sequence has a quadratic rate of convergence.*

References

[Aga78] M. Aganagic, *Variational inequalities and generalized complementarity problems*, Technical Report SOL 78-11, Systems Optimization Laboratory, Department of Operations Research, Stanford University, Stanford, CA, 1978.

[Ahn79] B. H. Ahn, *Computation of Market Equilibria for Policy Analysis: The Project Independence Evaluation Study (PIES) Approach*, Garland, New York, NY, 1979.

[AhH82] B. H. Ahn and W. W. Hogan, *On convergence of the PIES algorithm for computing equilibria*, Operations Research, 30 (1982), pp. 281–300.

[Aka59] H. Akaike, *On the successive transformation of probability distributions and its application to the analysis of the optimum gradient method*, Annals of the Institute of Mathematical Statistics, 11 (1959), pp. 1–17.

[AlK91] F. Al-Khayyal and J. Kyparisis, *Finite convergence of algorithms for nonlinear programs and variational inequalities*, Journal of Optimization Theory and Applications, 70 (1991), pp. 319–332.

[Arm66] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific Journal of Mathematics, 16 (1966), pp. 1–3.

[ArH60] K. J. Arrow and L. Hurwicz, *Decentralization and computation in resource allocation*, in Essays in Economics and Econometrics, R. W. Pfouts, ed., University of North Carolina Press, Rayleigh, NC, 1960, pp. 34–104.

[ArU60] K. J. Arrow and H. Uzawa, *Constraint qualifications in maximization problems, II*, Technical Report 84, Institute of Mathematical Studies in Social Sciences, Stanford University, Stanford, CA, 1960.

[Aub93] J.-P. Aubin, *Optima and Equilibria: An Introduction to Nonlinear Analysis*, Springer-Verlag, Berlin, 1993.

[Auc83] G. Auchmuty, *Duality for non-convex variational principles*, Journal of Differential Equations, 50 (1983), pp. 80–145.

[Auc86] ——, *Variational principles for periodic solutions of autonomous ordinary differential equations*, in Proceedings of a Conference on Nonlinear Oscillations in Chemistry and Biology, vol. 66 of Lecture Notes in Biomathematics, Springer-Verlag, New York, NY, 1986, pp. 252–260.

[Auc88] ——, *Variational principles for operator equations and initial value problems*, Nonlinear Analysis, Theory, Methods & Applications, 12 (1988), pp. 531–564.

[Auc89] ——, *Variational principles for variational inequalities*, Numerical Functional Analysis and Optimization, 10 (1989), pp. 863–874.

[Aus70] A. Auslender, *Une méthode générale pour le décomposition et la minimisation de fonctions non différentiables*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 271 (1970), pp. 1078–1081.

[Aus71] ——, *Méthodes numériques pour la décomposition et la minimisation de fonctions non différentiables*, Numerische Mathematik, 18 (1971), pp. 213–223.

[Aus76] ——, *Optimisation: Méthodes Numériques*, Masson, Paris, 1976.

[Aus84] ——, *Méthode de descente en programmation convexe non différentiable*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 298 (1984), pp. 529–532.

[BaC84] C. Baiocchi and A. Capelo, *Variational and Quasivariational Inequalities: Applications to Free Boundary Problems*, John Wiley & Sons, New York, NY, 1984.

[BaP74] A. B. Bakušinskiĭ and B. T. Poljak, *On the solution of variational inequalities*, Soviet Mathematics Doklady, 15 (1974), pp. 1705–1710.

[Bau78] G. M. Baudet, *Asynchronous iterative methods for multiprocessors*, Journal of the ACM, 25 (1978), pp. 226–244.

[BGS72] M. S. Bazaraa, J. J. Goode, and C. M. Shetty, *Constraint qualifications revisited*, Management Science, 18 (1972), pp. 567–573.

[BaS76] M. S. Bazaraa and C. M. Shetty, *Foundations of Optimization*, vol. 122 of Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1976.

[BaS79] ——, *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, New York, NY, 1979.

[Ben62] J. F. Benders, *Partitioning procedures for solving mixed variables programming problems*, Numerische Mathematik, 4 (1962), pp. 238–252.

[Ben74] A. Bensoussan, *Points de Nash dans le cas de fonctionelles quadratiques et jeux différentiels linéaires a N personnes*, SIAM Journal on Control, 12 (1974), pp. 460–499.

[BGL73] A. Bensoussan, M. Goursat, and J. L. Lions, *Contrôle impulsionnel et inéquations quasi-variationnelles stationnaires*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 276 (1973), pp. 1279–1284.

[BKS80] A. Bensoussan, P. Kleindorfer, and C. S. Tapiero, eds., *Applied Stochastic Control in Econometrics and Management Science*, vol. 130 of Contributions to Economic Analysis, North-Holland, Amsterdam, 1980.

[BeL73] A. Bensoussan and J. L. Lions, *Nouvelle formulation de problemes de contrôle impulsionnel et applications*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 276 (1973), pp. 1189–1192.

[Ber63] C. Berge, *Topological Spaces*, Oliver & Boyd, Edinburgh, 1963.

[Ber75a] P. Bernhard, *Commande Optimale, Décentralisation, et Jeux Dynamiques*, Dunod, Paris, 1975.

[Ber75b] D. P. Bertsekas, *Combined primal-dual and penalty methods for constrained optimization*, SIAM Journal on Control, 13 (1975), pp. 521–544.

[Ber75c] ——, *Necessary and sufficient conditions for a penalty method to be exact*, Mathematical Programming, 9 (1975), pp. 87–99.

[Ber76a] ——, *Multiplier methods: a survey*, Automatica, 12 (1976), pp. 133–145.

[Ber76b] ——, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Transactions on Automatic Control, AC-21 (1976), pp. 174–184.

[Ber81] ——, *A new algorithm for the assignment problem*, Mathematical Programming, 21 (1981), pp. 152–171.

[Ber82a] ——, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, San Diego, CA, 1982.

[Ber82b] ——, *Distributed dynamic programming*, IEEE Transactions on Automatic Control, AC-27 (1982), pp. 610–616.

[Ber82c] ——, *Projected Newton methods for optimization problems with simple constraints*, SIAM Journal on Control and Optimization, 20 (1982), pp. 221–246.

[Ber83] ———, *Distributed asynchronous computation of fixed points*, Mathematical Programming, 27 (1983), pp. 107–120.

[Ber85a] ———, *A distributed asynchronous relaxation algorithm for the assignment problem*, in Proceedings of the 24th IEEE Conference on Decision and Control, Fort Lauderdale, FL, 1985, pp. 1703–1704.

[Ber85b] ———, *A unified framework for primal-dual methods in minimum cost network flow problems*, Mathematical Programming, 32 (1985), pp. 125–145.

[Ber86] ———, *Distributed asynchronous relaxation methods for linear network flow problems*, Report LIDS-P-1606, Department of Electrical Engineering and Computer Science, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1986.

[Ber91a] ———, *An auction algorithm for shortest paths*, SIAM Journal on Optimization, 1 (1991), pp. 425–447.

[Ber91b] ———, *Linear Network Optimization: Algorithms and Codes*, MIT Press, Cambridge, MA, 1991.

[Ber92] ———, *Auction algorithms for network flow problems: a tutorial introduction*, Computational Optimization and Applications, 1 (1992), pp. 7–66.

[BeC91] D. P. Bertsekas and D. A. Castañon, *Parallel synchronous and asynchronous implementations of the auction algorithm*, Parallel Computing, 17 (1991), pp. 707–732.

[BeE88] D. P. Bertsekas and J. Eckstein, *Dual coordinate step methods for linear network flow problems*, Mathematical Programming, 42 (1988), pp. 203–243.

[BeE87] D. P. Bertsekas and D. El Baz, *Distributed asynchronous relaxation methods for convex network flow problems*, SIAM Journal on Control and Optimization, 25 (1987), pp. 74–85.

[BeG82] D. P. Bertsekas and E. M. Gafni, *Projection methods for variational inequalities with application to the traffic equilibrium problem*, Mathematical Programming Study, 17 (1982), pp. 139–159.

[BHT87] D. P. Bertsekas, P. A. Hosein, and P. Tseng, *Relaxation methods for network flow problems with convex arc costs*, SIAM Journal on Control and Optimization, 25 (1987), pp. 1219–1243.

[BeM73] D. P. Bertsekas and S. K. Mitter, *A descent numerical method for optimization problems with nondifferentiable cost functionals*, SIAM Journal on Control, 11 (1973), pp. 637–652.

[BeT88a] D. P. Bertsekas and P. Tseng, *RELAX: a computer code for minimum cost network flow problems*, Annals of Operations Research, 13 (1988), pp. 127–190.

[BeT88b] ———, *Relaxation methods for minimum cost ordinary and generalized network flow problems*, Operations Research, 36 (1988), pp. 93–114.

[BeT89] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, London, 1989.

[BTA84] D. P. Bertsekas, J. N. Tsitsiklis, and M. Athans, *Convergence theories of distributed iterative processes: a survey*, Report LIDS-P-1412, Department of Electrical Engineering and Computer Science, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[BNS87] A. Bihain, V. H. Nguyen, and J.-J. Strodiot, *A reduced subgradient algorithm*, Mathematical Programming Study, 30 (1987), pp. 127–149.

[Bre66] L. M. Bregman, *A relaxation method of finding a common point of convex sets and its application to problems of optimization*, Soviet Mathematics Doklady, 7 (1966), pp. 1578–1581.

[Bre67] ——, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Computational Mathematics and Mathematical Physics, 7 (1967), pp. 200–217.

[Bre73] H. Brézis, *Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert*, North-Holland, Amsterdam, 1973.

[BrE76a] H. Brézis and I. Ekeland, *Un principe variational associé à certaines équations paraboliques: le cas independent du temps*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 282 (1976), pp. 971–974.

[BrE76b] ——, *Un principe variational associé à certaines équations paraboliques: le cas dependent du temps*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 282 (1976), pp. 1197–1198.

[BrS69] H. Brézis and M. Sibony, *Méthodes d'approximation et d'itération pour les opérateurs monotones*, Archive for Rational Mechanics and Analysis, 27 (1969), pp. 59–82.

[BLP65] C. B. Brosilow, L. S. Lasdon, and J. D. Pearson, *Feasible optimization methods for interconnected systems*, in Proceedings of the Joint Automatic Control Conference, Troy, New York, NY, 1965.

[Bro66] F. E. Browder, *Existence and approximation of solutions of nonlinear variational inequalities*, Proceedings of the National Academy of Sciences, U.S.A., 56 (1966), pp. 1080–1086.

[Bro68] ——, *Nonlinear maximal monotone mappings in Banach spaces*, Mathematische Annalen, 175 (1968), pp. 81–113.

[Bro65] C. G. Broyden, *A class of methods for solving nonlinear simultaneous equations*, Mathematics of Computation, 19 (1965), pp. 577–593.

[Bro69] ——, *A new double-rank minimization algorithm*, Notices of the American Mathematical Society, 16 (1969), p. 670.

[Bro70] ——, *The convergence of single-rank quasi-Newton methods*, Mathematics of Computation, 24 (1970), pp. 365–382.

[Bur91] J. V. Burke, *An exact penalty viewpoint of constrained optimization*, SIAM Journal on Control and Optimization, 29 (1991), pp. 968–998.

[BuF91] J. V. Burke and M. C. Ferris, *Characterization of solution sets of convex programs*, Operations Research Letters, 10 (1991), pp. 57–60.

[BuF92] ——, *Weak sharp minima in mathematical programming*, unpublished report, Department of Mathematics, University of Washington, Seattle, WA, 1992.

[BuM88] J. V. Burke and J. J. Moré, *On the identification of active constraints*, SIAM Journal on Numerical Analysis, 25 (1988), pp. 1197–1211.

[BMT90] J. V. Burke, J. J. Moré, and G. Toraldo, *Convergence properties of trust region methods for linear and convex constraints*, Mathematical Programming, 47 (1990), pp. 305–336.

[CaM87] P. H. Calamai and J. J. Moré, *Projected gradient methods for linearly constrained problems*, Mathematical Programming, 39 (1987), pp. 93–116.

[Car77] M. Carey, *Integrability and mathematical programming models: a survey and a parametric approach*, Econometrica, 45 (1977), pp. 1957–1976.

[Cau47] A. Cauchy, *Méthode générale pour la résolution des systèmes d'équations simultanées*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 25 (1847), pp. 536–538.

[Cea78] J. Céa, *Optimization: Theory and Algorithms*, Springer-Verlag, Berlin, 1978.

[CeG73] J. Céa and R. Glowinski, *Sur des méthodes d'optimisation par relaxation*, Revue Française d'Automatique, Informatique et Recherche Opérationelle, R-3 (1973), pp. 5–31.

[Cen81] Y. Censor, *Row-action methods for huge and sparse systems and their applications*, SIAM Review, 23 (1981), pp. 444–466.

[CeL81] Y. Censor and A. Lent, *An iterative row-action method for interval convex programming*, Journal of Optimization Theory and Applications, 34 (1981), pp. 321–353.

[CeZ92] Y. Censor and S. A. Zenios, *Proximal minimization algorithm with D-functions*, Journal of Optimization Theory and Applications, 73 (1992), pp. 451–464.

[ChZ91] E. D. Chajakis and S. A. Zenios, *Synchronous and asynchronous implementations of relaxation algorithms for nonlinear network optimization*, Parallel Computing, 17 (1991), pp. 873–894.

[ChP82] D. Chan and J.-S. Pang, *The generalized quasi-variational inequality problem*, Mathematics of Operations Research, 7 (1982), pp. 211–222.

[CHL82] J. Chatelon, D. Hearn, and T. J. Lowe, *A subgradient algorithm for certain minimax and minisum problems—the constrained case*, SIAM Journal on Control and Optimization, 20 (1982), pp. 455–469.

[ChM69] D. Chazan and W. Miranker, *Chaotic relaxation*, Linear Algebra and Its Applications, 2 (1969), pp. 199–222.

[ChT93] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM Journal on Optimization, 3 (1993).

[ChR92] G. H.-G. Chen and R. T. Rockafellar, *Forward-backward splitting methods in Lagrangian optimization*, unpublished report, Department of Applied Mathematics, University of Washington, Seattle, WA, 1992.

[ChM88] R. J. Chen and R. R. Meyer, *Parallel optimization for traffic assignment*, Mathematical Programming, 42 (1988), pp. 327–345.

[Cla75] F. H. Clarke, *Generalized gradients and applications*, Transactions of the American Mathematical Society, 205 (1975), pp. 247–262.

[Cla76] ——, *A new approach to Lagrange multipliers*, Mathematics of Operations Research, 1 (1976), pp. 165–174.

[Coh81] A. I. Cohen, *Stepsize analysis for descent methods*, Journal of Optimization Theory and Applications, 33 (1981), pp. 187–205.

[Coh72] G. Cohen, *Coordination of constrained optimization problems by resource allocation*, Internal Report A/49, Centre d'Automatique et Informatique, Ecole Nationale Supérieure des Mines de Paris, Fontainebleau, France, 1972.

[Coh76] ——, *Une approche unifiée des algorithmes d'optimisation par décomposition-coordination*, Internal Report E/51, Centre d'Automatique et Informatique, Ecole Nationale Supérieure des Mines de Paris, Fontainebleau, France, 1976.

[Coh77] ——, *On an algorithm of decentralized optimal control*, Journal of Mathematical Analysis and Applications, 59 (1977), pp. 242–259.

[Coh78] ——, *Optimization by decomposition and coordination: a unified approach*, IEEE Transactions on Automatic Control, AC-23 (1978), pp. 222–232.

[Coh80] ——, *Auxiliary problem principle and decomposition of optimization problems*, Journal of Optimization Theory and Applications, 32 (1980), pp. 277–305.

[Coh82] ——, *Two lemmas and their use in convergence analysis of some optimization algorithms*, Internal Report E/68, Centre d'Automatique et Informatique, Ecole Nationale Supérieure des Mines de Paris, Fontainebleau, France, 1982.

[Coh84] ——, *Décomposition et coordination en optimisation déterministe différentiable et non différentiable*, PhD thesis, U.E.R. Mathématiques de la Décision, Université de Paris-IX Dauphine, Paris, 1984.

[Coh87] ——, *Nash equilibria: gradient and decomposition algorithms*, Large Scale Systems, 12 (1987), pp. 173–184.

[Coh88] ——, *Auxiliary problem principle extended to variational inequalities*, Journal of Optimization Theory and Applications, 59 (1988), pp. 325–333.

[CoB80] G. Cohen and J. F. Balducchi, *A new algorithm in nondifferentiable optimization and its use for computing saddle points*, in Proceedings of the 18th Annual Allerton Conference on Communication, Control and Computing, Monticello, IL, 1980.

[CoC86a] G. Cohen and F. Chaplais, *Algorithmes numériques pour les équilibres de Nash*, Automatique-Productique Informatique Industrielle, 20 (1986), pp. 273–293.

[CoC88] ——, *Nested monotony for variational inequalities over product of spaces and convergence of iterative algorithms*, Journal of Optimization Theory and Applications, 59 (1988), pp. 369–390.

[CoC85] G. Cohen and J.-C. Culioli, *Algorithmes de décomposition en optimisation stochastique et pour les équilibres de Nash*, Final Report ATP CNRS/CNET 'Systèmes Technologiques Complexes', Centre d'Automatique et Informatique, Ecole Nationale Supérieure des Mines de Paris, Fontainebleau, France, 1985.

[CoC86b] ——, *Algorithmes de décomposition/coordination en optimisation stochastique*, Automatique-Productique Informatique Industrielle, 20 (1986), pp. 252–272.

[CoC87] ——, *Decomposition and coordination in stochastic optimization*, in Proceedings of the IFAC Symposium on Large Scale Systems: Theory and Applications, Zurich, Switzerland, 1986, H. P. Geering and M. Mansour, eds., IFAC Proceedings Series, Oxford, 1987, pp. 155–160.

[CoJ76] G. Cohen and G. Joalland, *Coordination methods by the prediction principle in large dynamic constrained optimization problems*, in Proceedings of the IFAC Symposium on Large-Scale Systems, G. Guardabassi and A. Locatelli, eds., Udine, Italy, 1976.

[CoM90] G. Cohen and B. Miara, *Optimization with an auxiliary constraint and decomposition*, SIAM Journal on Control and Optimization, 28 (1990), pp. 137–157.

[CoZ83] G. Cohen and D. L. Zhu, *Decomposition of nonsmooth optimization problems*, in Large Scale Systems: Theory and Applications 1983, Proceedings of the IFAC/IFORS Symposium, Warszaw, Poland, A. Straszak, ed., Pergamon Press, Oxford, 1983, pp. 339–344.

[CoZ84] ——, *Decomposition coordination methods in large scale optimization problems: the nondifferentiable case and the use of augmented Lagrangians*, in Advances in Large Scale Systems, Volume 1, J. B. Cruz, ed., JAI Press, Greenwich, CT, 1984, pp. 203–266.

[CGT88a] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Global convergence of a class of trust region algorithms for optimization with simple bounds*, SIAM Journal on Numerical Analysis, 25 (1988), pp. 433–460. Correction, ibid, 26 (1989), pp. 764–767.

[CGT88b] ——, *Testing a class of methods for solving minimization problems with simple bounds on the variables*, Mathematics of Computation, 50 (1988), pp. 399–430.

[Cot66] R. W. Cottle, *Nonlinear programs with positively bounded Jacobians*, SIAM Journal on Applied Mathematics, 14 (1966), pp. 147–158.

[CGL80] R. W. Cottle, F. Giannessi, and J.-L. Lions, eds., *Variational Inequalities and Complementarity Problems: Theory and Applications*, John Wiley & Sons, Chichester, U.K., 1980.

[CPS92] R. W. Cottle, J.-S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, New York, NY, 1992.

[CPV89] R. W. Cottle, J.-S. Pang, and V. Venkateswaran, *Sufficient matrices and the linear complementarity problem*, Linear Algebra and Its Applications, 114/115 (1989), pp. 231–249.

[Cra84] J. Crank, *Free and Moving Boundary Problems*, Clarendon Press, Oxford, U.K., 1984.

[CrC55] J. B. Crockett and H. Chernoff, *Gradient methods of maximization*, Pacific Journal of Mathematics, 5 (1955), pp. 33–50.

[Cul87] J.-C. Culioli, *Algorithmes de décomposition/coordination en optimisation stochastique*, PhD thesis, Centre d'Automatique et Informatique, Ecole Nationale Supérieure des Mines de Paris, 1987.

[CuC90] J.-C. Culioli and G. Cohen, *Decomposition/coordination algorithms in stochastic optimization*, SIAM Journal on Control and Optimization, 28 (1990), pp. 1372–1403.

[Cyb87] G. Cybenko, *Dynamic load balancing for distributed memory multiprocessors*, Technical Report 87-1, Department of Computer Science, Tufts University, Medford, MA, 1987.

[Daf80] S. Dafermos, *Traffic equilibrium and variational inequalities*, Transportation Science, 14 (1980), pp. 42–54.

[Daf82] ——, *Relaxation algorithms for the general asymmetric traffic equilibrium problem*, Transportation Science, 16 (1982), pp. 231–240.

[Daf83] ——, *An iterative scheme for variational inequalities*, Mathematical Programming, 26 (1983), pp. 40–47.

[Daf90] ——, *Exchange price equilibria and variational inequalities*, Mathematical Programming, 46 (1990), pp. 391–402.

[DaN87] S. Dafermos and A. Nagurney, *Oligopolistic and competitive behaviour of spatially separated markets*, Regional Science and Urban Economics, 17 (1987), pp. 245–254.

[Dan69] Y. M. Danilin, *On an approach to minimization problems*, Soviet Mathematics Doklady, 10 (1969), pp. 1274–1276.

[Dan70] ——, *Minimization methods based on approximation of the initial functional by a convex functional*, USSR Computational Mathematics and Mathematical Physics, 10 (1970), pp. 1–18.

[Dan67] J. M. Danskin, *The Theory of Max-Min*, Springer-Verlag, Berlin, 1967.

[Dav59] W. C. Davidon, *Variable metric method for minimization*, Report ANL-5990 Rev, Argonne National Laboratories, Argonne, IL, 1959.

[DeI86] A. R. De Pierro and A. N. Iusem, *A relaxed version of Bregman's method for convex programming*, Journal of Optimization Theory and Applications, 51 (1986), pp. 421–440.

[DES82] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM Journal on Numerical Analysis, 19 (1982), pp. 400–408.

[DMZ89] R. S. Dembo, J. M. Mulvey, and S. A. Zenios, *Large-scale nonlinear network models and their application*, Operations Research, 37 (1989), pp. 353–372.

[DeS83] R. S. Dembo and T. Steihaug, *Truncated-Newton algorithms for large-scale unconstrained optimization*, Mathematical Programming, 26 (1983), pp. 190–212.

[DeT83] R. S. Dembo and U. Tulowitzki, *On the minimization of quadratic functions subject to box constraints*, Working Paper Series B # 71, School of Organization and Management, Yale University, New Haven, CT, 1983.

[DeT84] ——, *Local convergence analysis for successive inexact quadratic programming methods*, Working Paper Series B # 78, School of Organization and Management, Yale University, New Haven, CT, 1984.

[DeT85] ——, *Sequential truncated quadratic programming methods*, in Numerical Optimization 1984, Proceedings of the SIAM Conference on Numerical Optimization, Boulder, CO, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., Society of Industrial and Applied Mathematics, Philadelphia, PA, 1985, pp. 83–101.

[DeT88] ——, *Computing equilibria on large multicommodity networks: an application of truncated quadratic programming algorithms*, Networks, 18 (1988), pp. 273–284.

[DeM74a] V. F. Dem'yanov and V. N. Malozemov, *Introduction to Minimax*, John Wiley & Sons, New York, NY, 1974.

[DeP72] V. F. Dem'yanov and A. B. Pevnyi, *Numerical methods for finding saddle points*, USSR Computational Mathematics and Mathematical Physics, 12 (1972), pp. 11–52.

[DeR65] V. F. Dem'yanov and A. M. Rubinov, *On the problem of minimization of a smooth functional with convex constraints*, Soviet Mathematics Doklady, 6 (1965), pp. 9–11.

[DeR67] ——, *The minimization of a smooth convex functional on a convex set*, SIAM Journal on Control, 5 (1967), pp. 280–294.

[DeR70] ——, *Approximate Methods in Optimization Problems*, American Elsevier, New York, NY, 1970.

[DeS78] V. F. Dem'yanov and V. K. Shomesova, *Conditional subdifferentials of convex functions*, Soviet Mathematics Doklady, 19 (1978), pp. 1181–1185.

[DeS80] ——, *Subdifferentials of functions on sets*, Cybernetics, 16 (1980), pp. 24–31.

[DeV85] V. F. Dem'yanov and L. V. Vasil'ev, *Nondifferentiable Optimization*, Optimization Software, New York, NY, 1985.

[DeM74b] J. E. Dennis and J. J. Moré, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Mathematics of Computation, 28 (1974), pp. 549–560.

[DeM77] ——, *Quasi-Newton methods, motivation and theory*, SIAM Review, 19 (1977), pp. 46–89.

[DoR56] J. Douglas and H. H. Rachford, *On the numerical solution of heat conduction problems in two and three space variables*, Transactions of the American Mathematical Society, 82 (1956), pp. 421–439.

[Dub73] J. Dubois, *Theorems of convergence for improved nonlinear programming algorithms*, Operations Research, 21 (1973), pp. 328–332.

[Dun79] J. C. Dunn, *Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals*, SIAM Journal on Control and Optimization, 17 (1979), pp. 187–211.

[Dun80a] ——, *Convergence rates for conditional gradient sequences generated by implicit step length rules*, SIAM Journal on Control and Optimization, 18 (1980), pp. 473–487.

[Dun80b] ——, *Newton's method and the Goldstein step-length rule for constrained minimization problems*, SIAM Journal on Control and Optimization, 18 (1980), pp. 659–674.

[Dun81] ——, *Global and asymptotic convergence rate estimates for a class of projected gradient processes*, SIAM Journal on Control and Optimization, 19 (1981), pp. 368–400.

[Dun87] ——, *On the convergence of projected gradient processes to singular critical points*, Journal of Optimization Theory and Applications, 55 (1987), pp. 203–216.

[DFL86] J.-P. Dussault, J. A. Ferland, and B. Lemaire, *Convex quadratic programming with one constraint and bounded variables*, Mathematical Programming, 36 (1986), pp. 90–104.

[DuF93] J. P. Dussault and G. Fournier, *On the convergence of the projected gradient method*, Journal of Optimization Theory and Applications, 77 (1993), pp. 197–208.

[DuM89] J.-P. Dussault and P. Marcotte, *Conditions de régularité géométrique pour les inéquations variationnelles*, Recherche opérationelle, 23 (1989), pp. 1–16.

[DuL72] G. Duvaut and J.-L. Lions, *Les Inéquations en Mécanique et en Physique*, Dunod, Paris, 1972.

[Eav71] B. C. Eaves, *On the basic theorem of complementarity*, Mathematical Programming, 1 (1971), pp. 68–75.

[Eav78a] ——, *Computing stationary points*, Mathematical Programming Study, 7 (1978), pp. 1–14.

[Eav78b] ——, *A locally quadratically convergent algorithm for computing stationary points*, Technical Report SOL-78-13, Department of Operations Research, Stanford University, Stanford, CA, 1978.

[Eck88] J. Eckstein, *The Lions-Mercier splitting algorithm and the alternating direction method are instances of the proximal point algorithm*, Report LIDS-P-1769, Department of Electrical Engineering and Computer Science, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1988.

[Eck89] ——, *Splitting methods for monotone operators with applications to parallel optimization*, PhD thesis, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA, 1989.

[Eck93] ——, *Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming*, Mathematics of Operations Research, 18 (1993), pp. 202–226.

[EcB92] J. Eckstein and D. P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Mathematical Programming, 55 (1992), pp. 293–318.

[EkT76] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[ElB89] D. El Baz, *A computational experience with distributed asynchronous iterative methods for convex network flow problems*, in Proceedings of the 28th IEEE Conference on Decision and Control, Tampa, FL, 1989, pp. 590–591.

[Elk68] R. Elkin, *Convergence theorems for Gauss-Seidel and other minimization algorithms*, PhD thesis, University of Maryland, College Park, MD, 1968.

[Ere65] I. I. Eremin, *The relaxation method of solving systems of inequalities with convex functions on the left sides*, Soviet Mathematics Doklady, 6 (1965), pp. 219–222.

[Erl78] D. Erlenkotter, *A dual-based procedure for uncapacitated facility location*, Operations Research, 26 (1978), pp. 992–1009.

[Erm66] Y. M. Ermol'ev, *Methods of solution of nonlinear extremal problems*, Cybernetics, 2 (1966), pp. 1–14.

[Eva76] S. P. Evans, *Derivation and analysis of some models for combining trip distribution and assignment*, Transportation Research, 10 (1976), pp. 37–57.

[Evt85] Y. G. Evtushenko, *Numerical Optimization Techniques*, Optimization Software, New York, NY, 1985.

[Fan82] S.-C. Fang, *Traffic equilibria on multiclass-user transportation networks analyzed via variational inequalities*, Tamkang Journal of Mathematics, 13 (1982), pp. 1–9.

[FaP80] S.-C. Fang and E. L. Peterson, *Economic equilibria on networks*, Mathematics Research Report 80-13, Department of Mathematics, University of Maryland, Baltimore County, MD, 1980.

[FaP82] ——, *Generalized variational inequalities*, Journal of Optimization Theory and Applications, 38 (1982), pp. 363–383.

[FeM88] B. Feijoo and R. R. Meyer, *Piecewise-linear approximation methods for nonseparable convex optimization*, Management Science, 34 (1988), pp. 411–419.

[Fen49] W. Fenchel, *On conjugate convex functions*, Canadian Journal of Mathematics, 1 (1949), pp. 73–77.

[Fer88] M. C. Ferris, *Weak sharp minima and penalty functions in mathematical programming*, Technical Report 779, Computer Sciences Department, University of Wisconsin, Madison, WI, 1988.

[Fer91] ——, *Finite termination of the proximal point algorithm*, Mathematical Programming, 50 (1991), pp. 359–366.

[FeM91] M. C. Ferris and O. L. Mangasarian, *Finite perturbation of convex programs*, Applied Mathematics and Optimization, 23 (1991), pp. 263–273.

[FeM92] ——, *Minimum principle sufficiency*, Mathematical Programming, 57 (1992), pp. 1–14.

[FiH75] J.-C. Fiorot and P. Huard, *Composition et réunion d'algorithmes généraux d'optimisation*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 280 (1975), pp. 1455–1458.

[FiH79] ——, *Composition and union of general algorithms of optimization*, Mathematical Programming Study, 10 (1979), pp. 69–85.

[FiN82] C. S. Fisk and S. Nguyen, *Solution algorithms for network equilibrium models with asymmetric user costs*, Transportation Science, 16 (1982), pp. 361–381.

[Fla92] S. D. Flåm, *On finite convergence and constraint identification of subgradient projection methods*, Mathematical Programming, 57 (1992), pp. 427–437.

[Fle70] R. Fletcher, *A new approach to variable metric algorithms*, Computer Journal, 13 (1970), pp. 317–322.

[Fle87] ——, *Practical Methods of Optimization*, John Wiley & Sons, Chichester, U.K., second ed., 1987.

[FlP63] R. Fletcher and M. J. D. Powell, *A rapidly convergent descent method for minimization*, Computer Journal, 6 (1963), pp. 163–168.

[Fle80] T. M. Flett, *Differential Analysis*, Cambridge University Press, Cambridge, 1980.

[Flo86] M. Florian, *Nonlinear cost network models in transportation analysis*, Mathematical Programming Study, 26 (1986), pp. 167–196.

[FlN74] M. Florian and S. Nguyen, *A method for computing network equilibrium with elastic demands*, Transportation Science, 8 (1974), pp. 321–332.

[FrW56] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Research Logistics Quarterly, 3 (1956), pp. 95–110.

[Fuk84] M. Fukushima, *A descent algorithm for nonsmooth convex optimization*, Mathematical Programming, 30 (1984), pp. 163–175.

[Fuk92] ——, *Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems*, Mathematical Programming, 53 (1992), pp. 99–110.

[FuM81] M. Fukushima and H. Mine, *A generalized proximal point algorithm for certain non-convex minimization problems*, International Journal of Systems Sciences, 12 (1981), pp. 989–1000.

[FNS91] M. Fukushima, V. H. Nguyen, and J.-J. Strodiot, *A globally convergent algorithm for a class of nonsmooth optimization problems and its application to parallel decomposition of convex programs*, Technical Report 91002, Department of Applied Mathematics and Physics, Kyoto University, Kyoto, Japan, 1991.

[Gab79] D. Gabay, *Méthodes numériques pour l'optimisation non linéaire*, PhD thesis, Université Pierre et Marie Curie, Paris, 1979.

[Gab83] ——, *Applications of the method of multipliers to variational inequalities*, in Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems, M. Fortin and R. Glowinski, eds., North-Holland, Amsterdam, 1983, pp. 299–331.

[GaM76] D. Gabay and B. Mercier, *A dual algorithm for the solution of a nonlinear variational problem via finite element approximation*, Computers and Mathematics with Applications, 2 (1976), pp. 17–40.

[GaM80] D. Gabay and H. Moulin, *On the uniqueness and stability of Nash-equilibria in noncooperative games*, in Applied Stochastic Control in Econometrics and Management Science, A. Bensoussan, P. Kleindorfer, and C. S. Tapiero, eds., North-Holland, Amsterdam, 1980, pp. 271–293.

[GaB82] E. M. Gafni and D. P. Bertsekas, *Convergence of a gradient projection method*, Report LIDS-P-1201, Department of Electrical Engineering and Computer Science, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1982.

[GaB84] ——, *Two-metric projection methods for constrained optimization*, SIAM Journal on Control and Optimization, 22 (1984), pp. 936–964.

[GaK70] H. Gajewski and R. Kluge, *Projektionsverfahren bei nichtlinearen Variationsungleichungen*, Mathematische Nachrichten, 46 (1970), pp. 363–373.

[GaD88] M. Gawande and J. C. Dunn, *Variable metric gradient projection processes in convex feasible sets defined by nonlinear inequalities*, Applied Mathematics and Optimization, 17 (1988), pp. 103–119.

[Gil82] J. R. Giles, *Convex Analysis with Application in Differentiation of Convex Functions*, vol. 58 of Research Notes in Mathematics, Pitman, Boston, MA, 1982.

[GLT81] R. Glowinski, J.-L. Lions, and R. Trémolières, *Numerical Analysis of Variational Inequalities*, North-Holland, Amsterdam, 1981.

[GlM75] R. Glowinski and A. Marrocco, *Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires*, Revue Française d'Automatique, Informatique et Recherche Opérationelle, R-2 (1975), pp. 41–76.

[Gol70] D. Goldfarb, *A family of variable-metric methods derived by variational means*, Mathematics of Computation, 24 (1970), pp. 23–26.

[Gol62] A. A. Goldstein, *Cauchy's method of minimization*, Numerische Mathematik, 4 (1962), pp. 146–150.

[Gol64] ——, *Convex programming in Hilbert space*, Bulletin of the American Mathematical Society, 70 (1964), pp. 709–710.

[Gol74] ——, *On gradient projection*, in Proceedings of the 12th Annual Allerton Conference on Circuit and System Theory, Monticello, IL, 1974, pp. 38–40.

[GoP67] A. A. Goldstein and J. B. Price, *An effective algorithm for minimization*, Numerische Mathematik, 10 (1967), pp. 184–189.

[Gol75] E. G. Gol'shtein, *Method of modification for monotone mappings*, Economics and Mathematical Methods, 11 (1975), pp. 1144–1159.

[GoT74] E. G. Gol'shtein and N. V. Tret'yakov, *Modified Lagrangian functions*, Economics and Mathematical Methods, 10 (1974), pp. 568–591.

[GoT75] ——, *The gradient method of minimization and algorithms of convex programming based on modified Lagrangian functions*, Economics and Mathematical Methods, 11 (1975), pp. 730–742.

[GoT79] ——, *Modified Lagrangians in convex programming and their generalizations*, Mathematical Programming Study, 10 (1979), pp. 86–97.

[GoT71] F. J. Gould and J. W. Tolle, *A necessary and sufficient qualification for constrained optimization*, SIAM Journal on Applied Mathematics, 20 (1971), pp. 164–172.

[Gre89] A. Greenbaum, *Synchronization costs on multiprocessors*, Parallel Computing, 10 (1989), pp. 3–14.

[Gui69] M. Guignard, *Generalized Kuhn-Tucker conditions for mathematical programming problems in a Banach space*, SIAM Journal on Control, 7 (1969), pp. 232–241.

[Gul92] O. Güler, *New proximal point algorithms for convex minimization*, SIAM Journal on Optimization, 2 (1992), pp. 649–664.

[Gur64] M. E. Gurtin, *Variational principles for linear initial-value problems*, Quarterly of Applied Mathematics, 22 (1964), pp. 252–256.

[Ha90] C. D. Ha, *A generalization of the proximal point algorithm*, SIAM Journal on Control and Optimization, 28 (1990), pp. 503–512.

[HaB70] P. C. Haarhoff and J. D. Buys, *A new method for the optimization of a nonlinear function subject to nonlinear constraints*, The Computer Journal, 13 (1970), pp. 178–184.

[HaM87] J. H. Hammond and T. L. Magnanti, *A contracting ellipsoid method for variational inequality problems*, Working Paper OR 160-87, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, 1987.

[HaL88] S.-P. Han and G. Lou, *A parallel algorithm for a class of convex programs*, SIAM Journal on Control and Optimization, 26 (1988), pp. 345–355.

[HaM79] S.-P. Han and O. L. Mangasarian, *Exact penalty functions in nonlinear programming*, Mathematical Programming, 17 (1979), pp. 251–269.

[HPR92] S.-P. Han, J.-S. Pang, and N. Rangaraj, *Globally convergent Newton methods for nonsmooth equations*, Mathematics of Operations Research, 17 (1992), pp. 586–607.

[Har88] P. T. Harker, *Accelerating the convergence of the diagonalization and projection algorithms for finite-dimensional variational inequalities*, Mathematical Programming, 41 (1988), pp. 29–59.

[Har91] ——, *Generalized Nash games and quasi-variational inequalities*, European Journal of Operational Research, 54 (1991), pp. 81–94.

[HaP90a] P. T. Harker and J.-S. Pang, *A damped-Newton method for the linear complementarity problem*, in Computational Solution of Nonlinear Systems of Equations, E. L. Allgower and K. Georg, eds., vol. 26 of Lectures in Applied Mathematics, American Mathematical Society, Providence, RI, 1990, pp. 265–284.

[HaP90b] ——, *Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications*, Mathematical Programming, 48 (1990), pp. 161–220.

[HaX90] P. T. Harker and B. Xiao, *Newton's method for the nonlinear complementarity problem: a B-differentiable equation approach*, Mathematical Programming, 48 (1990), pp. 339–357.

[HaS66] P. Hartman and G. Stampacchia, *On some non-linear elliptic differential-functional equations*, Acta Mathematica, 115 (1966), pp. 271–310.

[Hea78] D. W. Hearn, *Network aggregation in transportation planning, volume I: summary and survey*, Mathtec Final Report DOT-TSC-RSPD-78-8, I, Mathtec, Inc., Princeton, NJ, 1978.

[Hea82] ——, *The gap function of a convex program*, Operations Research Letters, 1 (1982), pp. 67–71.

[HeL89] D. W. Hearn and S. Lawphongpanich, *A dual ascent algorithm for traffic assignment problems*, in Dynamic Control and Flow Equilibrium, Proceedings of the Italy-U.S.A. Joint Seminar on Urban Traffic Networks, Naples and Capri, Italy, 1989, pp. 35–53.

[HLN84] D. W. Hearn, S. Lawphongpanich, and S. Nguyen, *Convex programming formulations of the asymmetric traffic assignment problem*, Transportation Research, 18B (1984), pp. 357–365.

[HLV85] D. W. Hearn, S. Lawphongpanich, and J. A. Ventura, *Finiteness in restricted simplicial decomposition*, Operations Research Letters, 4 (1985), pp. 125–130.

[HLV87] ——, *Restricted simplicial decomposition: computation and extensions*, Mathematical Programming Study, 31 (1987), pp. 99–118.

[HeN82] D. W. Hearn and S. Nguyen, *Dual and saddle functions related to the gap function*, Research Report 82-4, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 1982.

[Hes69] M. R. Hestenes, *Multiplier and gradient methods*, Journal of Optimization Theory and Applications, 4 (1969), pp. 303–320.

[HiS92] J. L. Higle and S. Sen, *On the convergence of algorithms with implications for stochastic and nondifferentiable optimization*, Mathematics of Operations Research, 17 (1992), pp. 112–131.

[Hil57] C. Hildreth, *A quadratic programming procedure*, Naval Research Logistics Quarterly, 4 (1957), pp. 79–85. Erratum, ibid., p. 361.

[HoJ88] R. W. Hockney and C. R. Jesshope, *Parallel Computers 2: Architecture, Programming and Algorithms*, Adam Hilger, Bristol, 1988.

[Hog73] W. W. Hogan, *Point-to-set maps in mathematical programming*, SIAM Review, 15 (1973), pp. 591–603.

[Hol74] C. A. Holloway, *An extension of the Frank and Wolfe method of feasible directions*, Mathematical Programming, 6 (1974), pp. 14–27.

[HoJ84] K. Holmberg and K. O. Jörnsten, *Cross decomposition applied to the stochastic transportation problem*, European Journal of Operational Research, 17 (1984), pp. 361–368.

[Hua75] P. Huard, *Optimization algorithms and point-to-set maps*, Mathematical Programming, 8 (1975), pp. 308–331.

[Hua79] ——, *Extensions of Zangwill's theorem*, Mathematical Programming Study, 10 (1979), pp. 98–103.

[HwB85] K. Hwang and F. A. Briggs, *Computer Architecture and Parallel Processing*, McGraw-Hill, Singapore, 1985.

[Ius93] A. I. Iusem, *On the convergence of iterative methods for symmetric linear complementarity problems*, Mathematical Programming, 59 (1993), pp. 33–48.

[JoC78] G. Joalland and G. Cohen, *Optimal control of a water distribution network by two multilevel methods*, in Proceedings of the 7th IFAC World Congress, Helsinki, Finland, A. Niemi, ed., Pergamon Press, London, 1978.

[Jos79a] N. H. Josephy, *Newton's method for generalized equations*, Technical Report 1965, Mathematics Research Center, University of Wisconsin, Madison, WI, 1979.

[Jos79b] ——, *Quasi-Newton methods for generalized equations*, Technical Report 1966, Mathematics Research Center, University of Wisconsin, Madison, WI, 1979.

[Kar69a] S. Karamardian, *The nonlinear complementarity problem with applications, part 1*, Journal of Optimization Theory and Applications, 4 (1969), pp. 87–98.

[Kar69b] ——, *The nonlinear complementarity problem with applications, part 2*, Journal of Optimization Theory and Applications, 4 (1969), pp. 167–181.

[Kar71] ——, *Generalized complementarity problem*, Journal of Optimization Theory and Applications, 8 (1971), pp. 161–168.

[Kar72] ——, *The complementarity problem*, Mathematical Programming, 2 (1972), pp. 107–129.

[Kar76] ——, *An existence theorem for the complementarity problem*, Journal of Optimization Theory and Applications, 18 (1976), pp. 445–454.

[KiO88] N. Kikuchi and J. T. Oden, *Contact Problems in Elasticity*, Society of Industrial and Applied Mathematics, Philadelphia, PA, 1988.

[KiN91] K. Kim and J. L. Nazareth, *The decomposition principle and algorithms for linear programming*, Linear Algebra and Its Applications, 152 (1991), pp. 119–133.

[KiU89] S. Kim and B.-S. Um, *Polyak's subgradient method with simplified projection for nondifferentiable optimization with linear constraints*, Optimization, 20 (1989), pp. 451–456.

[KiS80] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, NY, 1980.

[Kiw85] K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, vol. 1133 of Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1985.

[Kiw86] ——, *A method for minimizing the sum of a convex function and a continuously differentiable function*, Journal of Optimization Theory and Applications, 48 (1986), pp. 437–449.

[Klu70] R. Kluge, *Zur approximativen Lösung nichtlinearer Variationsungleichungen*, Deutsche Akademie der Wissenschaften zu Berlin. Monatsberichte, 12 (1970), pp. 120–134.

[Kor75a] B. W. Kort, *Combined primal-dual and penalty function algorithms for nonlinear programming*, PhD thesis, Department of Electrical Engineering, Stanford University, Stanford, CA, 1975.

[Kor75b] ——, *Rate of convergence of the method of multipliers with inexact minimization*, in Nonlinear Programming 2, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, NY, 1975, pp. 193–214.

[KoB72] B. W. Kort and D. P. Bertsekas, *A new penalty function method for constrained minimization*, in Proceedings of the 1972 IEEE Conference on Decision and Control, New Orleans, IEEE, New York, NY, 1972, pp. 162–166.

[KoB76] ——, *Combined primal-dual and penalty methods for convex programming*, SIAM Journal on Control and Optimization, 14 (1976), pp. 268–294.

[Kun76] H. T. Kung, *Synchronized and asynchronous parallel algorithms for multiprocessors*, in Algorithms and Complexity: New Directions and Recent Results, J. F. Traub, ed., Academic Press, New York, NY, 1976, pp. 153–200.

[LaS81] B. Lamond and N. F. Stewart, *Bregman's balancing method*, Transportation Research, 15B (1981), pp. 239–248.

[LaM90] T. Larsson and A. Migdalas, *An algorithm for nonlinear programs over Cartesian product sets*, Optimization, 21 (1990), pp. 535–542.

[LMP89] T. Larsson, A. Migdalas, and M. Patriksson, *A partial linearization method for the traffic assignment problem*, Report LiTH-MAT-R-1989-06, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1989. Optimization (1993) (in press).

[LaP90] T. Larsson and M. Patriksson, *A class of gap functions for variational inequalities*, Report LiTH-MAT-R-1990-18, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1990. To appear in Mathematical Programming.

[LaP92] ——, *Simplicial decomposition with disaggregated representation for the traffic assignment problem*, Transportation Science, 26 (1992), pp. 4–17.

[LPS92] T. Larsson, M. Patriksson, and A.-B. Strömberg, *Conditional subgradient optimization—theory and applications*, Report LiTH-MAT-R-1992-24, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1992.

[Las70] L. S. Lasdon, *Optimization Theory for Large Systems*, Macmillan, New York, NY, 1970.

[LaS65] L. S. Lasdon and J. D. Schoeffler, *A multi-level technique for optimization*, in Proceedings of the Joint Automatic Control Conference, Troy, New York, NY, 1965.

[LaH84] S. Lawphongpanich and D. W. Hearn, *Simplicial decomposition of the asymmetric traffic assignment problem*, Transportation Research, 18B (1984), pp. 123–133.

[LaS87] J. Lawrence and J. E. Spingarn, *On fixed points of non-expansive piecewise isometric mappings*, Proceedings of the London Mathematical Society, 55 (1987), pp. 605–624.

[Lem78] C. Lemaréchal, *Nonsmooth optimization and descent methods*, Research Report RR-78-4, IIASA, Laxenburg, Austria, 1978.

[LeP66] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, USSR Computational Mathematics and Mathematical Physics, 6 (1966), pp. 1–50.

[LiP87] Y. Y. Lin and J.-S. Pang, *Iterative methods for large convex quadratic programs: a survey*, SIAM Journal on Control and Optimization, 25 (1987), pp. 383–411.

[LiM79] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM Journal on Numerical Analysis, 16 (1979), pp. 964–979.

[Lue70] D. G. Luenberger, *Control problems with kinks*, IEEE Transactions on Automatic Control, AC-15 (1970), pp. 570–575.

[Lue84] ——, *Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, second ed., 1984.

[LuT91] Z.-Q. Luo and P. Tseng, *On the convergence of a matrix splitting algorithm for the symmetric monotone linear complementarity problem*, SIAM Journal on Control and Optimization, 29 (1991), pp. 1037–1060.

[LuT92a] ——, *Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem*, SIAM Journal on Optimization, 2 (1992), pp. 43–54.

[LuT92b] ——, *On a global error bound for a class of monotone affine variational inequality problems*, Operations Research Letters, 11 (1992), pp. 159–165.

[LuT92c] ——, *On the convergence of the coordinate descent method for convex differentiable minimization*, Journal of Optimization Theory and Applications, 72 (1992), pp. 7–35.

[LuT92d] ——, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM Journal on Control and Optimization, 30 (1992), pp. 408–425.

[Luq84a] F. J. Luque, *Asymptotic convergence analysis of the proximal point algorithm*, SIAM Journal on Control and Optimization, 22 (1984), pp. 277–293.

[Luq84b] J. Luque, *Nonlinear proximal point algorithms*, PhD thesis, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[Luq86a] ——, *The nonlinear proximal point algorithm*, Report LIDS-P-1598, Laboratory for Information and Decision Sciences, Massachusetts Institute of Technology, Cambridge, MA, 1986.

[Luq86b] ——, *The nonlinear proximal point algorithm and multiplier methods*, Report LIDS-P-1596, Laboratory for Information and Decision Sciences, Massachusetts Institute of Technology, Cambridge, MA, 1986.

[MaS72] O. G. Mancino and G. Stampacchia, *Convex programming and variational inequalities*, Journal of Optimization Theory and Applications, 9 (1972), pp. 3–23.

[Man77] O. L. Mangasarian, *Solution of symmetric linear complementarity problems by iterative methods*, Journal of Optimization Theory and Applications, 22 (1977), pp. 465–485.

[Man84] ——, *Normal solutions of linear programs*, Mathematical Programming Study, 22 (1984), pp. 206–216.

[Man88] ——, *A simple characterization of solution sets of convex programs*, Operations Research Letters, 7 (1988), pp. 21–26.

[Man90] ——, *Convergence of iterates of an inexact matrix splitting algorithm for the symmetric monotone linear complementarity problem*, SIAM Journal on Optimization, 1 (1991), pp. 114–122.

[Man92] ——, *Global error bounds for monotone affine variational inequality problems*, Linear Algebra and Its Applications, 174 (1992), pp. 153–163.

[MaM79] O. L. Mangasarian and R. R. Meyer, *Nonlinear perturbation of linear programs*, SIAM Journal on Control and Optimization, 17 (1979), pp. 745–752.

[Mar85] P. Marcotte, *A new algorithm for solving variational inequalities with application to the traffic assignment problem*, Mathematical Programming, 33 (1985), pp. 339–351.

[Mar86] ——, *Gap-decreasing algorithms for monotone variational inequalities.* Conference paper, presented at the ORSA/TIMS Joint National Meeting, Miami Beach, FL, 1986.

[MaD85] P. Marcotte and J.-P. Dussault, *A modified Newton method for solving variational inequalities*, in Proceedings of the 24th IEEE Conference on Decision and Control, Fort Lauderdale, FL, 1985, pp. 1433–1436.

[MaD87] ——, *A note on a globally convergent Newton method for solving monotone variational inequalities*, Operations Research Letters, 6 (1987), pp. 35–42.

[MaD89] ——, *A sequential linear programming algorithm for solving monotone variational inequalities*, SIAM Journal on Control and Optimization, 27 (1989), pp. 1260–1278.

[MaG88] P. Marcotte and J. Guélat, *Adaptation of a modified Newton method for solving the asymmetric traffic equilibrium problem*, Transportation Science, 22 (1988), pp. 112–124.

[Mar70] B. Martinet, *Regularisation d'inéquations variationnelles par approximations successives*, Revue Française d'Informatique et de Recherche Opérationelle, R-3 (1970), pp. 154–158.

[Mar72] ——, *Détermination approchée d'un point fixe d'une application pseudo-contractante*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 274 (1972), pp. 163–165.

[Mar75] B. Martos, *Nonlinear Programming Theory and Methods*, North-Holland, Amsterdam, 1975.

[Mat87] L. Mathiesen, *An algorithm based on a sequence of linear complementarity problems applied to a Walrasian equilibrium model: an example*, Mathematical Programming, 37 (1987), pp. 1–18.

[McT72] G. P. McCormick and R. A. Tapia, *The gradient projection method under mild differentiability conditions*, SIAM Journal on Control, 10 (1972), pp. 93–98.

[MMT70] M. D. Mesarovic, D. Macko, and Y. Takahara, *Theory of Hierarchical Multilevel Systems*, Academic Press, New York, NY, 1970.

[Mey79] G. G. L. Meyer, *Asymptotic properties of sequences iteratively generated by point-to-set maps*, Mathematical Programming Study, 10 (1979), pp. 115–127.

[MeP71] G. G. L. Meyer and E. Polak, *Abstract models for the synthesis of optimization algorithms*, SIAM Journal on Control, 9 (1971), pp. 547–560.

[Mey76] R. R. Meyer, *Sufficient conditions for the convergence of monotonic mathematical programming algorithms*, Journal of Computer and System Sciences, 12 (1976), pp. 108–121.

[Mie71] J. C. Miellou, *Méthodes de Jacobi, Gauss-Seidel, sur-(sous) relaxation par blocs: appliquées à une classe de problèmes non linéaires*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 273 (1971), pp. 1257–1260.

[Mie75a] ——, *Algorithmes de relaxation chaotique à retards*, Revue Française d'Automatique, Informatique et Recherche Opérationelle, R-1 (1975), pp. 55–82.

[Mie75b] ——, *Itérations chaotiques à retards: études de la convergence dans le cas d'espaces partiellment ordonnés*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 280 (1975), pp. 233–236.

[Mig88] A. Migdalas, *Mathematical programming techniques for analysis and design of communication and transportation networks*, PhD thesis, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1988.

[Mig90] ——, *A regularization of the Frank-Wolfe method*, Report LiTH-MAT-R-1990-10, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1990.

[MiF81] H. Mine and M. Fukushima, *A minimization method for the sum of a convex function and a continuously differentiable function*, Journal of Optimization Theory and Applications, 33 (1981), pp. 9–23.

[Min62] G. J. Minty, *Monotone (nonlinear) operators in Hilbert space*, Duke Mathematical Journal, 29 (1962), pp. 341–346.

[Mor74] J. J. Moré, *Coercivity conditions in nonlinear complementarity problems*, SIAM Review, 17 (1974), pp. 1–16.

[Mor88] ——, *Trust regions and projected gradients*, in System Modelling and Optimization, M. Iri and K. Yajima, eds., vol. 113 of Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, 1988, pp. 1–13.

[Mor89] ——, *Gradient projection techniques for large-scale optimization problems*, in Proceedings of the 28th IEEE Conference on Decision and Control, Tampa, FL, 1989, pp. 378–381.

[MoT89] J. J. Moré and G. Toraldo, *Algorithms for bound constrained quadratic programming problems*, Numerische Mathematik, 55 (1989), pp. 377–400.

[Mor62] J.-J. Moreau, *Décomposition orthogonale dans un espace hilbertien selon deux cônes mutuellement polaires*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 255 (1962), pp. 233–240.

[Mor65] ——, *Proximité et dualité dans un espace Hilbertien*, Bulletin de la Société Mathématique de France, 93 (1965), pp. 273–299.

[MoF53] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics, Part 1*, McGraw-Hill, New York, NY, 1953.

[Mos72] U. Mosco, *Dual variational inequalities*, Journal of Mathematical Analysis and Applications, 40 (1972), pp. 202–206.

[Mur83] K. G. Murty, *Linear Programming*, John Wiley & Sons, New York, NY, 1983.

[Mur88] ——, *Linear Complementarity, Linear and Nonlinear Programming*, Heldermann Verlag, Berlin, 1988.

[Nag93] A. Nagurney, *Network Economics: A Variational Inequality Approach*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.

[Nas82] S. G. Nash, *Truncated-Newton methods*, PhD thesis, Computer Science Department, Stanford University, Stanford, CA, 1982.

[Nas84] ——, *Newton-type minimization via the Lanczos method*, SIAM Journal on Numerical Analysis, 21 (1984), pp. 770–788.

[Nas85a] ——, *Preconditioning of truncated-Newton methods*, SIAM Journal on Scientific and Statistical Computing, 6 (1985), pp. 599–616.

[Nas85b] ——, *Solving nonlinear programming problems using truncated-Newton techniques*, in Numerical Optimization 1984, Proceedings of the SIAM Conference on Numerical Optimization, Boulder, CO, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., Society of Industrial and Applied Mathematics, Philadelphia, PA, 1985, pp. 119–136.

[NgD84] S. Nguyen and C. Dupuis, *An efficient method for computing traffic equilibria in networks with asymmetric transportation costs*, Transportation Science, 18 (1984), pp. 185–202.

[Nik68] H. Nikaido, *Convex Structures and Economic Theory*, Academic Press, New York, NY, 1968.

[Noo75] M. A. Noor, *On variational inequalities*, PhD thesis, Brunel University, U.K., 1975.

[Noo82a] ——, *Mildly nonlinear variational inequalities*, Mathematica, 24 (1982), pp. 99–110.

[Noo82b] ——, *Strongly nonlinear variational inequalities*, La Sociéte Royale du Canada. L'Academie des Sciences. Comptes Rendus Mathématiques, 4 (1982), pp. 213–218.

[Noo85] ——, *Variational inequalities related with a Signorini problem*, La Sociéte Royale du Canada. L'Academie des Sciences. Comptes Rendus Mathématiques, 7 (1985), pp. 267–272.

[Noo87] ——, *General nonlinear variational inequalities*, Journal of Mathematical Analysis and Applications, 126 (1987), pp. 78–84.

[Noo91a] ——, *A computational method for variational inequalities*, in Computational Modelling of Free and Moving Boundary Problems, Vol. 1, L. C. Wrobel and C. A. Brebbia, eds., Computational Mechanics, Southampton, 1991, pp. 359–366.

[Noo91b] ——, *Some classes of variational inequalities*, in Constantin Carathéodory: An International Tribute, T. M. Rassias, ed., World Scientific Publ. Co., Singapore, 1991, pp. 996–1019.

[Noo92] ——, *General algorithm for variational inequalities*, Journal of Optimization Theory and Applications, 73 (1992), pp. 409–413.

[OrR70] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, NY, 1970.

[Ost66] A. Ostrowski, *Solution of Equations and Systems of Equations*, Academic Press, New York, NY, second ed., 1966.

[Pan82] J.-S. Pang, *On the convergence of a basic iterative method for the implicit complementarity problem*, Journal of Optimization Theory and Applications, 37 (1982), pp. 149–162.

[Pan84a] ——, *Necessary and sufficient conditions for the convergence of iterative methods for the linear complementarity problem*, Journal of Optimization Theory and Applications, 42 (1984), pp. 1–17.

[Pan84b] ——, *On the convergence of dual ascent methods for large-scale linearly constrained optimization problems*, technical report, School of Management, University of Texas at Dallas, Richardson, TX, 1984.

[Pan85] ——, *Asymmetric variational inequality problems over product sets: applications and iterative methods*, Mathematical Programming, 31 (1985), pp. 206–219.

[Pan86a] ——, *Inexact Newton methods for the nonlinear complementarity problem*, Mathematical Programming, 36 (1986), pp. 54–71.

[Pan86b] ——, *More results on the convergence of iterative methods for the symmetric linear complementarity problem*, Journal of Optimization Theory and Applications, 49 (1986), pp. 107–134.

[Pan87] ——, *A posteriori error bounds for the linearly-constrained variational inequality problem*, Mathematics of Operations Research, 12 (1987), pp. 474–484.

[Pan90a] ——, *Newton's method for B-differentiable equations*, Mathematics of Operations Research, 15 (1990), pp. 311–341.

[Pan90b] ——, *Solution differentiability and continuation of Newton's method for variational inequality problems over polyhedral sets*, Journal of Optimization Theory and Applications, 66 (1990), pp. 121–135.

[Pan91] ——, *A B-differentiable equation-based, globally and locally quadratically convergent algorithm for nonlinear programs, complementarity and variational inequality problems*, Mathematical Programming, 51 (1991), pp. 101–131.

[Pan93a] ——, *Complementarity problems*, report, Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD, 1993.

[Pan93b] ——, *Convergence of splitting and Newton methods for complementarity problems: an application of some sensitivity results*, Mathematical Programming, 58 (1993), pp. 149–160.

[PaC82] J.-S. Pang and D. Chan, *Iterative methods for variational and complementarity problems*, Mathematical Programming, 24 (1982), pp. 284–313.

[Pas79] G. B. Passty, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, Journal of Mathematical Analysis and Applications, 72 (1979), pp. 383–390.

[Pat90] M. Patriksson, *The traffic assignment problem—theory and algorithms*, Report LiTH-MAT-R-90-29, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1990.

[Pat91a] ——, *Partial linearization methods in nonlinear programming*, Report LiTH-MAT-R-91-11, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1991. Journal of Optimization Theory and Applications, 78 (1993) (in press).

[Pat91b] ——, *A unified description of iterative algorithms for traffic equilibria*, Report LiTH-MAT-R-91-35, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1991. European Journal of Operational Research (1993) (in press).

[Pet75] E. R. Petersen, *A primal-dual traffic assignment algorithm*, Management Science, 22 (1975), pp. 87–95.

[Pet73] D. W. Peterson, *A review of constraint qualifications in finite-dimensional spaces*, SIAM Review, 15 (1973), pp. 639–654.

[Phe85] R. R. Phelps, *Metric projections and the gradient projection method in Banach spaces*, SIAM Journal on Control and Optimization, 23 (1985), pp. 973–977.

[Phe86] ——, *The gradient projection method using Curry's steplength*, SIAM Journal on Control and Optimization, 24 (1986), pp. 692–699.

[Phe89] ——, *Convex Functions, Monotone Operators and Differentiability*, vol. 1364 of Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1989.

[Pie84] G. Pierra, *Decomposition through formalization in a product space*, Mathematical Programming, 28 (1984), pp. 96–115.

[PiZ92] M. Ç. Pinar and S. A. Zenios, *Parallel decomposition of multicommodity network flows using a linear-quadratic penalty algorithm*, ORSA Journal on Computing, 4 (1992), pp. 235–249.

[Pol71] E. Polak, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, NY, 1971.

[PoM85] E. Polak and D. Q. Mayne, *Algorithm models for nondifferentiable optimization*, SIAM Journal on Control and Optimization, 23 (1985), pp. 477–491.

[PMW83] E. Polak, D. Q. Mayne, and Y. Wardi, *On the extension of constrained optimization algorithms from differentiable to nondifferentiable problems*, SIAM Journal on Control and Optimization, 21 (1983), pp. 179–203.

[Pol63] B. T. Polyak, *Gradient methods for the minimisation of functionals*, USSR Computational Mathematics and Mathematical Physics, 3 (1963), pp. 864–878.

[Pol67] ——, *A general method of solving extremum problems*, Soviet Mathematics Doklady, 8 (1967), pp. 593–597.

[Pol69] ——, *Minimization of unsmooth functionals*, USSR Computational Mathematics and Mathematical Physics, 9 (1969), pp. 14–29.

[Pol78] ——, *Subgradient methods: a survey of Soviet research*, in Nonsmooth Optimization, C. Lemaréchal and R. Mifflin, eds., Pergamon Press, Oxford, 1978, pp. 5–29.

[Pol79] ——, *Sharp minimum*. Conference paper, presented at the IIASA Workshop on Generalized Lagrangians and their Applications, IIASA, Luxenburg, Austria, 1979.

[Pol87] ——, *Introduction to Optimization*, Optimization Software, New York, NY, 1987.

[PoT74] B. T. Polyak and N. V. Tret'yakov, *An iterative method for linear programming and its economic interpretation*, Matekon, 10 (1974), pp. 81–100.

[Pow69] M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, NY, 1969, pp. 283–298.

[PoS82] W. B. Powell and Y. Sheffi, *The convergence of equilibrium algorithms with predetermined step sizes*, Transportation Science, 16 (1982), pp. 45–55.

[Pri75] M. E. Primak, *A computational process of search for equilibrium points*, Cybernetics, 9 (1975), pp. 106–113.

[PsD78] B. N. Pshenichny and Y. M. Danilin, *Numerical Methods in Extremal Problems*, MIR Publishers, Moscow, 1978.

[RoV73] A. W. Roberts and D. E. Varberg, *Convex functions*, Academic Press, New York, NY, 1973.

[Rob79] S. M. Robinson, *Generalized equations and their solutions, part I: basic theory*, Mathematical Programming Study, 10 (1979), pp. 128–141.

[Rob80] ——, *Strongly regular generalized equations*, Mathematics of Operations Research, 1 (1980), pp. 43–62.

[Rob81] ——, *Some continuity properties of polyhedral multifunctions*, Mathematical Programming Study, 14 (1981), pp. 206–214.

[Rob82] ——, *Generalized equations and their solutions, part II: applications to nonlinear programming*, Mathematical Programming Study, 19 (1982), pp. 200–221.

[Rob83] ——, *Generalized equations*, in Mathematical Programming: The State of the Art—Bonn 1982, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 346–367.

[Roc66] R. T. Rockafellar, *Characterization of the subdifferentials of convex functions*, Pacific Journal of Mathematics, 17 (1966), pp. 497–510.

[Roc69a] ——, *Convex functions, monotone operators and variational inequalities*, in Theory and Applications of Monotone Operators, Proceedings of the NATO Advanced Study Institute, Venice, Italy, A. Ghizzetti, ed., Edizioni Oderisi, Gubbio, Italy, 1969, pp. 35–65.

[Roc69b] ——, *Local boundedness of nonlinear, monotone operators*, Michigan Mathematical Journal, 16 (1969), pp. 397–407.

[Roc70a] ——, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[Roc70b] ——, *On the maximal monotonicity of subdifferential mappings*, Pacific Journal of Mathematics, 33 (1970), pp. 209–216.

[Roc70c] ——, *On the maximality of sums of nonlinear monotone operators*, Transactions of the American Mathematical Society, 149 (1970), pp. 75–88.

[Roc73] ——, *The multiplier method of Hestenes and Powell applied to convex programming*, Journal of Optimization Theory and Applications, 12 (1973), pp. 555–562.

[Roc76a] ——, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Mathematics of Operations Research, 1 (1976), pp. 97–116.

[Roc76b] ——, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.

[Roc80] ——, *Lagrange multipliers and variational inequalities*, in Variational Inequalities and Complementarity Problems: Theory and Applications, R. W. Cottle, F. Giannessi, and J.-L. Lions, eds., John Wiley & Sons, Chichester, U.K., 1980, ch. 20, pp. 303–322.

[Roc81] ——, *The Theory of Subgradients and its Applications to Problems of Optimization: Convex and Nonconvex Functions*, Heldermann Verlag, Berlin, 1981.

[Rod87] J.-F. Rodrigues, *Obstacle Problems in Mathematical Physics*, North-Holland, Amsterdam, 1987.

[Ros60] J. B. Rosen, *The gradient projection method for nonlinear programming, part I: linear constraints*, Journal of the Society of Industrial and Applied Mathematics, 8 (1960), pp. 181–217.

[Ros61] ——, *The gradient projection method for nonlinear programming, part II: nonlinear constraints*, Journal of the Society of Industrial and Applied Mathematics, 9 (1961), pp. 514–532.

[Rus89] A. Ruszczyński, *An augmented Lagrangian decomposition method for block diagonal linear programming problems*, Operations Research Letters, 8 (1989), pp. 287–294.

[Sai76] R. Saigal, *Extension of the generalized complementarity problem*, Mathematics of Operations Research, 1 (1976), pp. 260–266.

[Sch62] S. Schechter, *Iteration methods for nonlinear problems*, Transactions of the American Mathematical Society, 104 (1962), pp. 179–189.

[Sch68] ——, *Relaxation methods for convex problems*, SIAM Journal on Numerical Analysis, 5 (1968), pp. 601–612.

[ScZ90] M. H. Schneider and S. A. Zenios, *A comparative study of algorithms for matrix balancing*, Operations Research, 38 (1990), pp. 439–455.

[Sha70] D. F. Shanno, *Conditioning of quasi-Newton methods for function minimization*, Mathematics of Computation, 24 (1970), pp. 647–656.

[ShA93] A. Shapiro and F. Al-Khayyal, *First-order conditions for isolated locally optimal solutions*, Journal of Optimization Theory and Applications, 77 (1993), pp. 189–196.

[She85] Y. Sheffi, *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1985.

[Sho83] N. Z. Shor, *Generalized gradient methods of nondifferentiable optimization employing space dilatation operations*, in Mathematical Programming: The State of the Art—Bonn 1982, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 501–529.

[Sho85] ——, *Minimization Methods for Non-Differentiable Functions*, Springer-Verlag, Berlin, 1985.

[Sib70] M. Sibony, *Méthodes itératives pour les équations et inéquations aux dérivées partielles non linéares de type monotone*, Calcolo, 7 (1970), pp. 65–183.

[Smi79] M. J. Smith, *The existence, uniqueness and stability of traffic equilibria*, Transportation Research, 13B (1979), pp. 295–304.

[Smi83a] ——, *An algorithm for solving asymmetric equilibrium problems with a continuous cost-flow function*, Transportation Research, 17B (1983), pp. 365–371.

[Smi83b] ——, *The existence and calculation of traffic equilibria*, Transportation Research, 17B (1983), pp. 291–303.

[Smi84] ——, *A descent algorithm for solving monotone variational inequalities and monotone complementarity problems*, Journal of Optimization Theory and Applications, 44 (1984), pp. 485–496.

[Spr83] J. E. Springarn, *Partial inverse of a monotone operator*, Applied Mathematics and Optimization, 10 (1983), pp. 247–265.

[Spr85] ——, *Applications of the method of partial inverses to convex programming: decomposition*, Mathematical Programming, 32 (1985), pp. 199–223.

[Sre87] V. P. Sreedharan, *ε-Subgradient projection algorithm*, Journal of Approximation Theory, 51 (1987), pp. 27–46.

[Sta69] G. Stampacchia, *Variational inequalities*, in Theory and Applications of Monotone Operators, Proceedings of the NATO Advanced Study Institute, Venice, Italy, A. Ghizzetti, ed., Edizioni Oderisi, Gubbio, Italy, 1969, pp. 101–193.

[StW70] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimensions I*, Springer-Verlag, Berlin, 1970.

[Szi90] F. Szidarovszky, *On the convergence of nonstationary algorithms modeled by point-to-set maps*, Applied Mathematics and Computation, 39 (1990), pp. 107–110.

[TFI93] K. Taji, M. Fukushima, and T. Ibaraki, *A globally convergent Newton method for solving strongly monotone variational inequalities*, Mathematical Programming, 58 (1993), pp. 369–383.

[Tak64] Y. Takahara, *Multilevel approach to dynamic optimization*, Report SRC-50-C-64-18, Systems Research Center, Case Western Reserve University, Cleveland, OH, 1964.

[Teb92] M. Teboulle, *Entropic proximal mappings with applications to nonlinear programming*, Mathematics of Operations Research, 17 (1992), pp. 670–690.

[Tik63] A. N. Tikhonov, *Solution of incorrectly formulated problems and the regularization method*, Soviet Mathematics Doklady, 4 (1963), pp. 1035–1038.

[TiA77] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, John Wiley & Sons, New York, NY, 1977.

[Tis77] S. Tishyadhigama, *General convergence theorems: their relationships and applications*, PhD thesis, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, 1977.

[TPK79] S. Tishyadhigama, E. Polak, and R. Klessig, *A comparative study of several general convergence conditions for algorithms modeled by point-to-set maps*, Mathematical Programming Study, 10 (1979), pp. 172–190.

[ToF83] R. L. Tobin and T. L. Friesz, *Formulating and solving the spatial price equilibrium problem with transshipment in terms of arc variables*, Journal of Regional Science, 23 (1983), pp. 187–198.

[Toi88] P. L. Toint, *Global convergence of a class of trust region methods for nonconvex minimization in Hilbert space*, IMA Journal on Numerical Analysis, 8 (1988), pp. 231–252.

[Ton72] E. Tonti, *On the variational formulation for linear initial value problems*, Annali di Matematica pura ed applicata, XCV (1972), pp. 331–360.

[Ton84] ——, *Variational formulation for every nonlinear problem*, International Journal of Engineering Sciences, 22 (1984), pp. 1343–1371.

[Tse86] P. Tseng, *Relaxation methods for monotropic programming problems*, PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 1986.

[Tse90] ——, *Dual ascent methods for problems with strictly convex costs and linear constraints: a unified approach*, SIAM Journal on Control and Optimization, 28 (1990), pp. 214–242.

[Tse91a] ——, *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities*, SIAM Journal on Control and Optimization, 29 (1991), pp. 119–138.

[Tse91b] ——, *Decomposition algorithm for convex differentiable minimization*, Journal of Optimization Theory and Applications, 70 (1991), pp. 109–135.

[Tse92] ——, *On the rate of convergence of a partially asynchronous gradient projection algorithm*, SIAM Journal on Optimization, 1 (1992), pp. 603–619.

[TsB91] P. Tseng and D. P. Bertsekas, *Relaxation methods for problems with strictly convex costs and linear constraints*, Mathematics of Operations Research, 16 (1991), pp. 462–481.

[TBT90] P. Tseng, D. P. Bertsekas, and J. N. Tsitsiklis, *Partially asynchronous, parallel algorithms for network flow and other problems*, SIAM Journal on Control and Optimization, 28 (1990), pp. 678–710.

[Tsi84] J. N. Tsitsiklis, *Problems in decentralized decision making and computation*, PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[Tsi87] ——, *On the stability of asynchronous iterative processes*, Mathematical Systems Theory, 20 (1987), pp. 137–153.

[TsB86] J. N. Tsitsiklis and D. P. Bertsekas, *Distributed asynchronous optimal routing in data networks*, IEEE Transactions on Automatic Control, AC-31 (1986), pp. 325–332.

[TBA86] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Transactions on Automatic Control, AC-31 (1986), pp. 803–812.

[UrD89] A. Üresin and M. Dubois, *Sufficient conditions for the convergence of asynchronous iterations*, Parallel Computing, 10 (1989), pp. 83–92.

[Uza58] H. Uzawa, *Iterative methods for concave programming*, in Studies in Linear and Nonlinear Programming, K. J. Arrow, L. Hurwicz, and H. Uzawa, eds., Stanford University Press, Stanford, CA, 1958, pp. 154–165.

[vTi84] J. van Tiel, *Convex Analysis: An Introductory Text*, John Wiley & Sons, Chichester, U.K., 1984.

[Var62] R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[VeH93] J. A. Ventura and D. W. Hearn, *Restricted simplicial decomposition for convex constrained problems*, Mathematical Programming, 59 (1993), pp. 71–85.

[vHo75] B. von Hohenbalken, *A finite algorithm to maximize certain pseudoconcave functions on polytopes*, Mathematical Programming, 9 (1975), pp. 189–206.

[vHo77] ——, *Simplicial decomposition in nonlinear programming algorithms*, Mathematical Programming, 13 (1977), pp. 49–68.

[vNe28] J. von Neumann, *Zur Theorie der Gesellschaftsspiele*, Mathematische Annalen, 100 (1928), pp. 295–320.

[Wol69] P. Wolfe, *Convergence conditions for ascent methods*, SIAM Review, 11 (1969), pp. 226–235.

[Wol72] ——, *On the convergence of gradient methods under constraints*, IBM Journal of Research and Development, 16 (1972), pp. 407–411.

[Wri89] S. Wright, *Convergence of SQP-like methods for constrained optimization*, SIAM Journal on Control and Optimization, 27 (1989), pp. 13–26.

[WFM90] J. H. Wu, M. Florian, and P. Marcotte, *A general descent framework for the monotone variational inequality problem*, Publication 723, Centre de recherche sur les transports, Université de Montréal, Montréal, 1990.

[Yao91] J.-C. Yao, *The generalized quasi-variational inequality problem with applications*, Journal of Mathematical Analysis and Applications, 158 (1991), pp. 139–160.

[Zad69] N. Zadeh, *A note on the cyclic coordinate ascent method*, Management Science, 16 (1969/1970), pp. 642–644.

[Zan67] W. I. Zangwill, *Application of the convergence conditions*, Working Paper 231, Center for Research in Management Science, University of California, Berkeley, CA, 1967.

[Zan69a] ——, *Convergence conditions for nonlinear programming algorithms*, Management Science, 16 (1969), pp. 1–13.

[Zan69b] ——, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

[Zar71] E. H. Zarantonello, *Projections on convex sets in Hilbert space and spectral theory*, in Contributions to Nonlinear Functional Analysis, Proceedings of a Symposium Conducted by the Mathematics Research Center, The University of Wisconsin, Madison, E. H. Zarantonello, ed., Academic Press, New York, NY, 1971, pp. 237–424.

[Zei90] E. Zeidler, *Nonlinear Functional Analysis and its Applications II/B: Nonlinear Monotone Operators*, Springer-Verlag, New York, NY, 1990.

[Zha88] L. Zhao, *Variational inequalities in general equilibrium: analysis and computation*, PhD thesis, Division of Applied Mathematics, Brown University, Providence, RI, 1988.

[ZhD91] L. Zhao and S. Dafermos, *General economic equilibrium and variational inequalities*, Operations Research Letters, 10 (1991), pp. 369–376.

[Zhu82] D. L. Zhu, *Optimisation sous-differentiable et méthodes de décomposition*, PhD thesis, Centre d'Automatique et Informatique, Ecole Nationale Supérieure des Mines de Paris, Fontainebleau, France, 1982.

[Zhu87] ——, *Decomposition methods in large scale nonsmooth optimization problems*, Journal of Systems Science and Mathematical Sciences, 7 (1987), pp. 63–71.

[ZPP69a] S. I. Zuhovickiĭ, R. A. Polyak, and M. E. Primak, *Numerical methods of finding equilibrium points of n-person games*, in Proceedings of the First Winter School on Mathematical Programming at Drogobych, 1969, pp. 93–130.

[ZPP69b] ——, *Two methods of search for equilibrium points of n-person concave games*, Soviet Mathematics Doklady, 10 (1969), pp. 279–282.

[ZPP70] ——, *On an n-person concave game and a production model*, Soviet Mathematics Doklady, 11 (1970), pp. 522–526.

[ZPP73] ——, *Concave multiperson games: numerical methods*, Matekon, 9 (1973), pp. 10–30.