Conditional Subgradient Optimization — Theory and Applications

by

Torbjörn Larsson
Michael Patriksson†
Ann-Brith Strömberg

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

†Presently at Department of Mathematics, GN–50,
University of Washington, Seattle, WA, 98195, USA

To appear in European Journal of Operational Research

Abstract

We generalize the subgradient optimization method for nondifferentiable convex programming to utilize conditional subgradients. Firstly, we derive the new method and establish its convergence by generalizing convergence results for traditional subgradient optimization. Secondly, we consider a particular choice of conditional subgradients, obtained by projections, which leads to an easily implementable modification of traditional subgradient optimization schemes. To evaluate the subgradient projection method we consider its use in three applications: uncapacitated facility location, two-person zero-sum matrix games, and multicommodity network flows. Computational experiments show that the subgradient projection method performs better than traditional subgradient optimization; in some cases the difference is considerable. These results suggest that our simple modification may improve subgradient optimization schemes significantly. This finding is important as such schemes are very popular, especially in the context of Lagrangean relaxation.

Keywords: *Convex Programming, Nonsmooth Optimization, Subgradient Methods, Conditional Subgradients, Lagrangean Relaxation.*

Abbreviated title: *Conditional Subgradient Optimization*

1 Introduction and motivation

Subgradient optimization methods for minimizing a nondifferentiable convex function originate in a work by N.Z. Shor from 1962; see Shor (1991) for a review of the early history of nonsmooth optimization. For the case of unconstrained optimization, Ermol'ev (1966) proved convergence of the method using step lengths according to a divergent series, and Polyak (1967 and 1969) extended the method to the case of convex constraints and gave additional convergence results. These methods have been frequently and often successfully applied, particularly in connection with Lagrangean duality (e.g., Geoffrion, 1974, Held et al., 1974, and Fisher, 1981 and 1985); subgradient methods also are closely related to relaxation methods for solving systems of linear inequalities (e.g., Goffin, 1978). We present *conditional subgradient* optimization methods, which include as special cases the methods just mentioned and subgradient projection methods; the latter are shown to have better practical performances.

Let the function $f : \Re^n \mapsto \Re$ be convex and thus everywhere continuous and subdifferentiable. Further, let $X \subseteq \Re^n$ be a nonempty, closed and convex set, and assume that $\inf_{x \in X} f(x) = f(x^*) > -\infty$ for some $x^* \in X$. The program considered is

[P]

$$f^* = \min_{x \in X} f(x),$$

with the nonempty and convex set of optimal solutions $X^* = \{ x \in X \mid f(x) = f^* \}$. The stated properties of the program [P] are assumed to hold throughout the paper.

The *subdifferential* of f at x is defined as

$$\partial f(x) = \left\{ \gamma \in \Re^n \;\middle|\; f(y) \geq f(x) + \gamma^T (y - x), \quad \forall y \in \Re^n \right\},$$

which is a nonempty, convex and compact set (Rockafellar, 1970, Theorem 23.4). Letting t denote iteration number and $\gamma(x^t)$ be a *subgradient* of f at $x^t \in X$, i.e., $\gamma(x^t) \in \partial f(x^t)$, the subgradient optimization method for the program [P] is given by

$$x^{t+1} = P_X \left(x^t - \alpha_t \cdot \gamma(x^t) \right), \qquad t = 0, 1, \ldots, \tag{1}$$

where

$$P_X(x) = \arg\min_{y \in X} \|y - x\|^2 \tag{2}$$

denotes the Euclidean projection of a point $x \in \Re^n$ onto the set X, and the step length α_t is chosen according to some rule which guarantees convergence (see, e.g., Shor, 1985). When $X = \Re^n$, step lengths satisfying the divergent series conditions

$$\alpha_t > 0, \quad \forall t, \qquad \lim_{t \to \infty} \alpha_t = 0, \qquad \text{and} \qquad \sum_{t=0}^{\infty} \alpha_t = \infty \tag{3}$$

yield an infinite sequence $\{x^t\}$ of iteration points such that $\{f(x^t)\} \to f^*$ (Ermol'ev, 1966). When the optimal value of [P] is known a priori, one may use the step length formula (Polyak, 1969)

$$\alpha_t = \theta_t \cdot \frac{f(x^t) - f^*}{\|\gamma(x^t)\|^2}, \qquad 0 < \varepsilon_1 \leq \theta_t \leq 2 - \varepsilon_2 < 2, \qquad \forall t, \tag{4}$$

where θ_t is a step length (or relaxation) parameter.

Example 1 (Slow convergence of Polyak's method). The following (trivial) instance of the program [P] illustrates the possibly slow convergence of Polyak's method (1), (4). Let $X = \{ x \in \Re^2 \mid x_1 + x_2 = 1;\ x_1, x_2 \geq 0 \}$ and $f(x) = x_1 + 2x_2 - 1$, $x \in \Re^2$; the optimal solution is $x^* = (1,0)^T$, with $f^* = 0$. Applying the method (1), (4) with the starting point $x^0 = (0,1)^T$ and step length parameter $\theta_t \equiv 1$ gives step lengths $\alpha_t = f(x^t)/5$, for all t. The iterates then are $x^t = (1 - (9/10)^t, (9/10)^t)^T$, with objective values $f(x^t) = (9/10)^t$, for all t (see Figure 1). Clearly, for this instance, Polyak's method converges very slowly towards the optimal solution. □

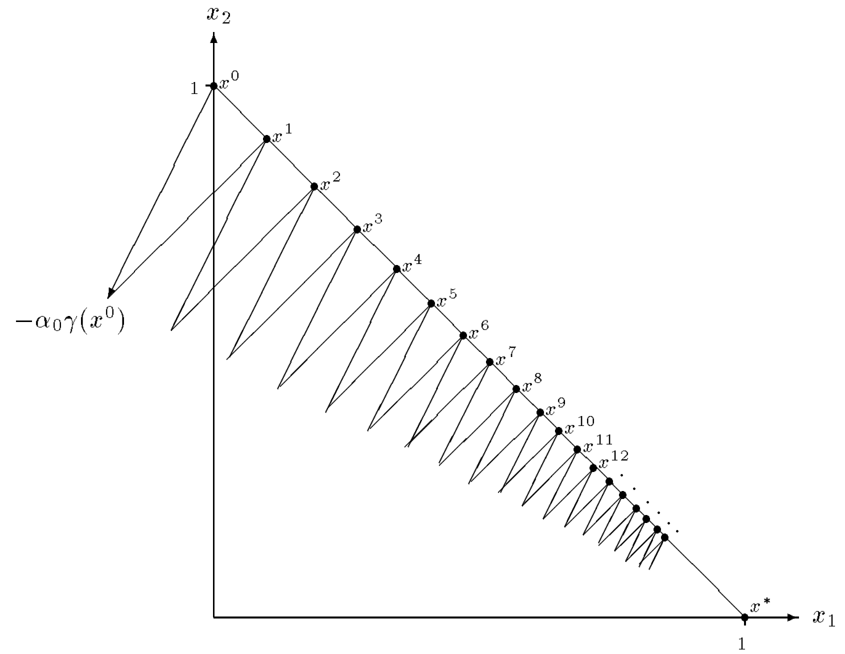

Figure 1: *Iteration point sequence for the example using Polyak's method.*

The zig-zagging phenomenon in this example is due to the fact that the subgradients are almost antiparallel to the normals of the feasible set. This behaviour was very pronounced in an application studied in Damberg (1990), and it motivated us to investigate the possibility of using projected subgradient directions in subgradient optimization schemes. It subsequently lead to the development of conditional subgradient optimization methods, of which subgradient projection methods are special cases. Because of the high computational cost of performing projections onto general convex sets, both Polyak's method (1), (4) and the subgradient projection method are most conveniently used when the feasible set is simple (for example a simplex).

To the best of our knowledge, subgradient projection approaches have only been studied by Svanberg and Bååthe (1985), and Kim and Um (1989). A more frequently utilized strategy for the minimization of nondifferentiable functions subject to linear constraints is to construct descent directions and perform line searches; see, e.g., Bertsekas and Mitter (1973), Polak et al. (1983), Strodiot et al. (1983), Bihain et al. (1987), Panier (1987), Sreedharan (1987), and Tamarit Goerlich (1988). For an application with general linear constraints it is likely that a descent method is preferable to a subgradient optimization method, since in that case both methods involve the solution of general quadratic programs, but the former requires fewer iterations.

The outline of the paper is as follows. The general conditional subgradient optimization method is derived in Section 2, together with convergence results for step length rules of the forms (3) and (4). In Section 3, we present the subgradient projection method. Implementation and evaluation of the subgradient projection method for applications to uncapacitated facility location, two-person zero-sum matrix games, and minimum cost multicommodity network flows is reported in Section 4. In Section 5, we make some concluding remarks and suggest further research.

2 Conditional subgradient optimization

2.1 Preliminaries

Dem'janov and Šomesova (1978 and 1980) generalize the definition of the subdifferential so that it takes the feasible set into account. This generalization is the *conditional subdifferential*

$$\partial^X f(x) = \left\{ \gamma \in \Re^n \;\middle|\; f(y) \geq f(x) + \gamma^T (y - x), \quad \forall y \in X \right\}, \qquad x \in X, \tag{5}$$

the elements of which will be referred to as *conditional subgradients.* Clearly, $\partial^X f(x) \supseteq \partial f(x)$ for all $x \in X$. The following proposition is immediate.

Proposition 1 (Properties of the conditional subdifferential)**.** *The conditional subdifferential $\partial^X f(x)$ is nonempty, closed and convex for all $x \in X$.* □

The *normal cone* of X at x is defined by

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

Dem'janov and Šomesova show the following result.

Proposition 2 (Characterization of the conditional subdifferential)**.**

$$\partial^X f(x) = \partial f(x) + N_X(x), \qquad x \in X.$$

□

Note that $\partial^X f(x)$ is unbounded whenever x is on the boundary of X.

In order to establish the relationships between traditional subgradient optimization and the conditional subgradient optimization method to be derived, we reformulate [P] as the unconstrained auxiliary program

[AP]

$$f^* = \min_{x \in \Re^n} f^X(x) \stackrel{\text{def}}{=} f(x) + \psi_X(x),$$

where

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

is the *indicator function* of the set X at x.

The auxiliary function f^X is lower semi-continuous, proper and convex, and its subdifferential is characterized in the below proposition, which follows from Theorem 5C in Rockafellar (1981) since $X \neq \emptyset$ and $\texttt{dom}\ f = \Re^n$ imply that $\texttt{int}\ (\texttt{dom}\ f) \cap \texttt{dom}\ \psi_X \neq \emptyset$. Here, `int` denotes interior and `dom` denotes effective domain.

Proposition 3 (Characterization of the subdifferential of the auxiliary function).

$$\partial f^X(x) = \partial f(x) + \partial \psi_X(x), \qquad x \in X.$$

□

Since $\partial \psi_X(x) = N_X(x)$ for all $x \in X$ (Rockafellar, 1981, Proposition 3J), Propositions 2 and 3 imply the following result.

Proposition 4 (Equality of subdifferentials).

$$\partial^X f(x) = \partial f^X(x), \qquad x \in X.$$

□

Note that ∂f^X denotes the ordinary subdifferential of the auxiliary function f^X, whereas $\partial^X f$ denotes the conditional subdifferential of f with respect to to X. As a conditional subgradient of f is a subgradient of f^X, the conditional subgradient method to be derived below may be viewed as traditional subgradient optimization applied to the auxiliary unconstrained program [AP], with an additional projection of the point obtained onto $\mathtt{dom}\ f^X$. However, if traditional subgradient optimization is applied to the program [AP], then the existing convergence results can not be directly invoked since they require the objective function to be finite on the feasible set.

The optimality condition of the program [P] is given in the next proposition, which follows from Proposition 5A and Equation (5.5) in Rockafellar (1981).

Proposition 5 (An optimality condition).

$$x \in X^* \quad \Longleftrightarrow \quad 0 \in \partial^X f(x) \quad \Longleftrightarrow \quad -\partial f(x) \cap N_X(x) \neq \emptyset.$$

□

Note that if X is a polyhedral set, i.e., if $X = \{\, x \in \Re^n \mid a_i^T x \leq b_i,\ i = 1, \dots, m \,\}$, then

$$N_X(x) = \begin{cases} \{\nu \in \Re^n \mid \nu = \sum_{i=1}^m a_i \omega_i;\ \omega_i(a_i^T x - b_i) = 0;\ \omega_i \geq 0,\ i = 1, \dots, m\}, & x \in X, \\ \emptyset, & x \notin X, \end{cases} \tag{6}$$

and by defining an index set for the active constraints at x,

$$\mathcal{I}(x) = \{\, i \in \{1, \dots, m\} \mid a_i^T x = b_i \,\}, \tag{7}$$

the necessary and sufficient conditions for the optimality of x in [P] can be expressed as

$$\exists \gamma \in \partial f(x) \quad \text{and} \quad \omega_i \geq 0,\ i \in \mathcal{I}(x), \quad \text{such that} \quad \gamma = -\sum_{i \in \mathcal{I}(x)} a_i \omega_i$$

(cf. the Karush–Kuhn–Tucker conditions for a differentiable program).

The negative of the *shortest* conditional subgradient of f at x with respect to X equals the *conditional steepest descent direction* (Dem'yanov and Vasil'ev, 1985, Section 1.11.3); it is found by solving

$$\min \left\{ \|\gamma\|^2 \;\middle|\; \gamma \in \partial^X f(x) \equiv \partial f(x) + N_X(x) \right\}. \tag{8}$$

The resulting direction is feasible if X is polyhedral. From Proposition 5, $\gamma = 0$ is optimal in (8) if and only if x is optimal in [P]. The program (8) is the basis for the development of descent methods for [P] (e.g., Dem'yanov and Vasil'ev, 1985, Section 4.4), and also for our subgradient projection scheme (see Section 3). Its specialization to differentiable nonlinear programming has been utilized in the convergence analysis of descent methods (e.g., Burke and Moré, 1988). In addition, the step direction in Rosen's (1960) gradient projection method (e.g., Luenberger, 1984, Exercise 11.10:8) can be interpreted as a heuristic solution to a dual formulation of (8).

2.2 Method analysis

Let $\gamma^X(x^t)$ be a conditional subgradient of f at $x^t \in X$. The *conditional subgradient optimization method* is given by

$$x^{t+1} = P_X\left(x^t - \alpha_t \cdot \gamma^X(x^t)\right), \qquad t = 0, 1, \dots. \tag{9}$$

In the (unlikely) case that $\gamma^X(x^t) = 0$ for some t, x^t is optimal, and the procedure (9) is terminated.

Our main convergence result for the method (9), (3) establishes convergence towards the optimal set X^*; the distance from a point x to this set is defined as $d_{X^*}(x) = \min_{y \in X^*} \|y - x\|$. The below theorem generalizes a classical convergence result of Ermol'ev (1966, Section 9).

Theorem 1 (Convergence to optimal set using divergent series step lengths)**.** *Let $\{x^t\}$ be generated by the method (9), (3) applied to* [P]*. If X^* is bounded and $\sup_t\{\|\gamma^X(x^t)\|\} < \infty$, then $\{f(x^t)\} \to f^*$ and $\{d_{X^*}(x^t)\} \to 0$.*

Proof. The proof is a non-trivial extension of that of Ermol'ev's theorem. We show that the iterates will eventually belong to an arbitrarily small neighbourhood of the set of optimal solutions to [P].

Let $\delta > 0$ and $B^\delta = \{\, x \in \Re^n \mid \|x\| \leq \delta \,\}$. Since f is convex, X is nonempty, closed and convex, and X^* is bounded, it follows (from Theorem 27.2 of Rockafellar, 1970, applied to the lower semi-continuous, proper and convex function $f + \psi_X$) that there exists an $\varepsilon = \varepsilon(\delta) > 0$ such that the level set $\{\, x \in X \mid f(x) \leq f^* + \varepsilon \,\} \subseteq X^* + B^{\delta/2}$; this level set is denoted by X^ε. Moreover, since for all t, $\|\gamma^X(x^t)\| \leq \sup_s\{\|\gamma^X(x^s)\|\} < \infty$, and $\{\alpha_t\} \to 0$, there exists an $N(\delta)$ such that $\alpha_t\|\gamma^X(x^t)\|^2 \leq \varepsilon$ and $\alpha_t\|\gamma^X(x^t)\| \leq \delta/2$ for all $t \geq N(\delta)$.

The sequel of the proof is based on induction and is organized as follows. In the first part, we show that there exists a finite $t(\delta) \geq N(\delta)$ such that $x^{t(\delta)} \in X^* + B^\delta$. In the second part, we establish that if x^t belongs to $X^* + B^\delta$ for some $t \geq N(\delta)$ then so does x^{t+1}, by showing that $d_{X^*}(x^{t+1}) < d_{X^*}(x^t)$, or $x^t \in X^\varepsilon$ so that $x^{t+1} \in X^* + B^\delta$ since the step taken is not longer than $\delta/2$.

Let $x^* \in X^*$ be arbitrary. In every iteration t we then have

$$\begin{aligned} \left\|x^* - x^{t+1}\right\|^2 &= \left\|x^* - P_X\left(x^t - \alpha_t\gamma^X(x^t)\right)\right\|^2 \\ &\leq \left\|x^* - x^t + \alpha_t\gamma^X(x^t)\right\|^2 \\ &= \left\|x^* - x^t\right\|^2 + \alpha_t\left(2\,\gamma^X(x^t)^T\left(x^* - x^t\right) + \alpha_t\left\|\gamma^X(x^t)\right\|^2\right), \end{aligned} \tag{10}$$

where the inequality follows from the projection property. Now, suppose that

$$2\,\gamma^X(x^s)^T\left(x^* - x^s\right) + \alpha_s\left\|\gamma^X(x^s)\right\|^2 < -\varepsilon \tag{11}$$

for all $s \geq N(\delta)$. Then, using (10) repeatedly, we obtain that for any $t \geq N(\delta)$,

$$\left\|x^* - x^{t+1}\right\|^2 < \left\|x^* - x^{N(\delta)}\right\|^2 - \varepsilon \sum_{s=N(\delta)}^{t} \alpha_s,$$

and from (3) it follows that the right-hand-side of this inequality tends to minus infinity as $t \to \infty$, which clearly is impossible. Therefore,

$$2\,\gamma^X(x^t)^T\left(x^* - x^t\right) + \alpha_t\left\|\gamma^X(x^t)\right\|^2 \geq -\varepsilon \tag{12}$$

for at least one $t \geq N(\delta)$, say $t = t(\delta)$. From the definition of $N(\delta)$, it follows that $\gamma^X(x^{t(\delta)})^T(x^* - x^{t(\delta)}) \geq -\varepsilon$. From (5) we have that $f(x^*) - f(x^{t(\delta)}) \geq \gamma^X(x^{t(\delta)})^T(x^* - x^{t(\delta)})$, since $x^*, x^{t(\delta)} \in X$. Hence, $f(x^{t(\delta)}) \leq f^* + \varepsilon$, i.e., $x^{t(\delta)} \in X^\varepsilon \subseteq X^* + B^{\delta/2} \subset X^* + B^\delta$.

Now, suppose that $x^t \in X^* + B^\delta$ for some $t \geq N(\delta)$. If (11) holds, then, using (10), we have that $\|x^* - x^{t+1}\| < \|x^* - x^t\|$ for any $x^* \in X^*$. Hence,

$$d_{X^*}(x^{t+1}) \leq \left\| P_{X^*}(x^t) - x^{t+1} \right\| < \left\| P_{X^*}(x^t) - x^t \right\| = d_{X^*}(x^t) \leq \delta.$$

Thus, $x^{t+1} \in X^* + B^\delta$. Otherwise, (12) must hold and, using the same arguments as above, we obtain that $f(x^t) \leq f^* + \varepsilon$, i.e., $x^t \in X^\varepsilon \subseteq X^* + B^{\delta/2}$. As

$$\left\| x^{t+1} - x^t \right\| = \left\| P_X\left(x^t - \alpha_t \gamma^X(x^t)\right) - x^t \right\| \leq \left\| x^t - \alpha_t \gamma^X(x^t) - x^t \right\| = \alpha_t \left\| \gamma^X(x^t) \right\| \leq \frac{\delta}{2}$$

whenever $t \geq N(\delta)$, it follows that $x^{t+1} \in X^* + B^{\delta/2} + B^{\delta/2} = X^* + B^\delta$.

By induction with respect to $t \geq t(\delta)$, it follows that $x^t \in X^* + B^\delta$ for all $t \geq t(\delta)$. Since this holds for arbitrarily small values of $\delta > 0$ and f is continuous, the theorem follows. $\square$

From the proof of the theorem it can be seen that the condition $\sup_t\{\|\gamma^X(x^t)\|\} < \infty$ can be relaxed into $\{\alpha_t \|\gamma^X(x^t)\|^2\} \to 0$. However, we have chosen to impose the, generally, stronger condition $\sup_t\{\|\gamma^X(x^t)\|\} < \infty$ because it is both necessary and sufficient for the convergence of the important special case of conditional subgradient optimization studied in the next section.

By imposing an additional condition on the choice of step lengths, it is possible to establish convergence of the whole sequence of iterates to an optimal solution, even without the boundedness condition on X^* (see, e.g., Shepilov, 1976, for a corresponding result for traditional subgradient optimization).

Theorem 2 (Convergence to an optimal solution using divergent series step lengths). *Let $\{x^t\}$ be generated by the method (9), (3) applied to* [P] *with the step lengths also fulfilling $\sum_{t=0}^\infty \alpha_t^2 < \infty$. If $\sup_t\{\|\gamma^X(x^t)\|\} < \infty$, then $\{x^t\}$ converges to an element of X^*.*

Proof. Let $x^* \in X^*$ and $t \geq 1$. Repeated application of (10) yields that

$$\left\| x^* - x^t \right\|^2 \leq \left\| x^* - x^0 \right\|^2 + 2 \sum_{s=0}^{t-1} \alpha_s \gamma^X(x^s)^T(x^* - x^s) + \sum_{s=0}^{t-1} \alpha_s^2 \left\| \gamma^X(x^s) \right\|^2. \tag{13}$$

Since $x^* \in X^*$ and $\gamma^X(x^s) \in \partial^X f(x^s)$ for all $s \geq 0$ we obtain that

$$f(x^s) \geq f^* \geq f(x^s) + \gamma^X(x^s)^T(x^* - x^s), \qquad s \geq 0, \tag{14}$$

and hence that $\gamma^X(x^s)^T(x^* - x^s) \leq 0$ for all $s \geq 0$. Define $c = \sup_t\{\|\gamma^X(x^t)\|\}$ and $p = \sum_{t=0}^\infty \alpha_t^2$, so that $\|\gamma^X(x^s)\| \leq c$ for any $s \geq 0$ and $\sum_{s=0}^{t-1} \alpha_s^2 < p$. From (13) we then conclude that $\|x^* - x^t\|^2 < \|x^* - x^0\|^2 + pc^2$ for any $t \geq 1$, and thus that the sequence $\{x^t\}$ is bounded.

Assume now that there is no subsequence $\{x^{t_i}\}$ of $\{x^t\}$ with $\{\gamma^X(x^{t_i})^T(x^* - x^{t_i})\} \to 0$. Then there must exist an $\varepsilon > 0$ with $\gamma^X(x^s)^T(x^* - x^s) \leq -\varepsilon$ for all sufficiently large values of s. From (13) and the conditions on the step lengths it follows that $\{\|x^* - x^t\|\} \to -\infty$, which clearly is impossible. The sequence $\{x^t\}$ must therefore contain a subsequence $\{x^{t_i}\}$ such that $\{\gamma^X(x^{t_i})^T(x^* - x^{t_i})\} \to 0$. From (14) it follows that $\{f(x^{t_i})\} \to f^*$. The boundedness of $\{x^t\}$ implies the existence of an accumulation point of the subsequence $\{x^{t_i}\}$, say x^∞. From the continuity of f it follows that $x^\infty \in X^*$.

To show that x^∞ is the only accumulation point of $\{x^t\}$, let $\delta > 0$ and find an $M(\delta)$ such that $\|x^\infty - x^{M(\delta)}\|^2 \leq \delta/2$ and $\sum_{s=M(\delta)}^\infty \alpha_s^2 \leq \delta/(2c^2)$. Consider any $t > M(\delta)$. Analogously to the derivation of (13), and using (14), we then obtain that

$$\left\|x^{\infty}-x^{t}\right\|^{2} \leq \left\|x^{\infty}-x^{M(\delta)}\right\|^{2}+\sum_{s=M(\delta)}^{t-1} \alpha_{s}^{2}\left\|\gamma^{X}(x^{s})\right\|^{2}<\frac{\delta}{2}+\frac{\delta}{2c^{2}} c^{2}=\delta.$$

Since this holds for arbitrarily small values of $\delta > 0$, the theorem follows. □

The results of Theorems 1 and 2 can, through the *almost complete relaxation* strategy of Dem'yanov and Vasil'ev (1985, Section 3.4), be used to establish convergence for other step length selection rules; the following corollary is immediate.

Corollary 1 (Convergence under the almost complete relaxation condition). *Let $\{\beta_t\}$ be an arbitrary sequence of step lengths replacing $\{\alpha_t\}$ in the method (9). If there exist sequences $\{\underline{\alpha}_t\}$ and $\{\overline{\alpha}_t\}$ that both satisfy (3) and $\underline{\alpha}_t \leq \beta_t \leq \overline{\alpha}_t$ for all t, then the invocation of Theorem 1 is valid. If, in addition, $\sum_{t=0}^{\infty} \underline{\alpha}_t^2$ and $\sum_{t=0}^{\infty} \overline{\alpha}_t^2$ are finite, then the invocation of Theorem 2 is valid.* □

Since the elements of the sequences $\{\underline{\alpha}_t\}$ and $\{\overline{\alpha}_t\}$ can be made arbitrarily small and arbitrarily large, respectively, while satisfying (3) (e.g., $\underline{\alpha}_t = \underline{c}/t$ and $\overline{\alpha}_t = \overline{c}/t$, where $\underline{c}$ and $\overline{c}$ are very small and very large constants, respectively), the almost complete relaxation strategy may be used as a safe-guard in step length selections based on line searches or formulas involving estimates of the optimal value, among others, through an (almost never used) upwards or downwards rounding of the tentative step length. This safe-guarding strategy could be introduced to enforce convergence also in descent methods with line searches for nondifferentiable optimization, which may otherwise yield false convergence (e.g., Fletcher, 1987, Section 14.4).

We now consider step lengths

$$\alpha_t = \theta_t \cdot \frac{f(x^t)-f^*}{\|\gamma^X(x^t)\|^2}, \qquad 0 < \varepsilon_1 \leq \theta_t \leq 2-\varepsilon_2 < 2, \qquad \forall t. \tag{15}$$

The next convergence theorem generalizes Polyak's (1969, Theorem 1) convergence result for the method (1), (4).

Theorem 3 (Convergence to an optimal solution using Polyak step lengths). *Let $\{x^t\}$ be generated by the method (9), (15) applied to* [P]. *If $\sup_t\{\|\gamma^X(x^t)\|\} < \infty$, then $\{f(x^t)\} \to f^*$, while $\{x^t\}$ converges to an element of X^*.*

Proof. Identical to the proof of Polyak's (1969) Theorem 1, with γ replaced by γ^X. □

It is well known (e.g., Ermol'ev, 1966, and Poljak, 1967) that the method (1), (3) cannot exhibit a linear convergence rate; this is also true for the more general method (9), (3). For the method (9), (15) linear convergence is achieved if f satisfies a *weak sharpness condition*, i.e., if there is a $\mu > 0$ such that

$$f(x)-f^* \geq \mu\, d_{X^*}(x), \qquad x \in \left\{\, y \in X \;\middle|\; \left\|y-P_{X^*}(x^0)\right\| \leq d_{X^*}(x^0) \,\right\}. \tag{16}$$

Polyak (1969, Theorem 3) establishes this result for the method (1), (4). See Polyak (1969, 1979, and 1987), and Burke and Ferris (1993) for characterizations of the Condition (16).

Theorem 4 (Linear convergence under a weak sharpness condition). *Under the assumptions of Theorem 3 and the Condition (16), the method (9), (15) applied to* [P] *exhibits linear convergence towards an element $x^* \in X^*$, i.e., $\|x^t - x^*\| \leq K q^t$, where $K \geq 0$ and $q = (1-\varepsilon_1\varepsilon_2(\mu/c)^2)^{1/2} < 1$ with $c = \sup_t\{\|\gamma^X(x^t)\|\}$.*

Proof. Identical to the proof of Polyak's (1969) Theorem 3, with γ replaced by γ^X. $\square$

Analogously to Polyak (1969) it is also possible to show that in the method (9), (15) with f^* replaced by a lower bound $\underline{f} \leq f^*$ and with $\theta_t \equiv 1$ there exists, for any $\varepsilon > 0$, a finite $t(\varepsilon)$ such that $f(x^{t(\varepsilon)}) \leq 2f^* - \underline{f} + \varepsilon$.

A fundamental property of subgradients is that they define halfspaces containing X^*. In the case of conditional subgradients, these halfspaces may be given an interesting interpretation. For some $x \in X$, let $\gamma \in \partial f(x)$ and $\nu \in N_X(x)$. Then $X^* \subseteq \{ y \in \Re^n \mid \gamma^T(y-x) \leq 0 \}$ since $f(x^*) \leq f(x)$ for all $x^* \in X^*$, and $X^* \subseteq \{ y \in \Re^n \mid \nu^T(y-x) \leq 0 \}$ since $X^* \subseteq X$. By adding these two set-defining inequalities, the *surrogate inequality* $(\gamma+\nu)^T(y-x) \leq 0$ is obtained and, clearly, $X^* \subseteq \{ y \in \Re^n \mid (\gamma+\nu)^T(y-x) \leq 0 \}$. But this is also a consequence of the fact that $\gamma + \nu \in \partial^X f(x) = \partial f^X(x)$.

3 The subgradient projection method

3.1 Preliminaries

In many applications of subgradient optimization, for instance in Lagrangean relaxation schemes, only one subgradient of f at x is immediately available, while the feasible set is explicitly described by a set of (simple) constraints. The restriction of the projection problem (8) to the corresponding *partial* subdifferential of f^X at x is the *subgradient projection* problem

[SP]
$$\gamma^X(x) = \arg\min \left\{ \|\gamma\|^2 \;\middle|\; \gamma \in \gamma(x) + N_X(x) \right\}.$$

Henceforth, $\gamma^X(x)$ will denote a projected subgradient, and not an arbitrary finite conditional subgradient. The solution x is optimal in [P] if $\gamma^X(x) = 0$; the reverse is, however, not true in general.

The following characteristic of the projected subgradient is immediate from the projection property.

Lemma 1 (Projection property)**.** *Let $\gamma(x) \in \partial f(x)$, $x \in X$. Then, $\|\gamma^X(x)\| \leq \|\gamma(x)\|$.* $\square$

In order to give an alternative characterization of the projected subgradient we consider a dual formulation of [SP]. Here, $T_X(x)$ denotes the tangent cone of X at x (see Rockafellar, 1981, Chapter 2).

Lemma 2 (Dual characterization of the projected subgradient)**.** *The projected subgradient is alternatively obtained from*

[DSP]
$$\gamma^X(x) = \arg\min \left\{ \|\gamma - \gamma(x)\|^2 \;\middle|\; \gamma \in -T_X(x) \right\}.$$

Proof. The convexity of X yields that $T_X(x)$ is polar to $N_X(x)$ (Rockafellar, 1981, Chapter 2). Let $C = -N_X(x)$ and C° be its polar; then the Moreau decomposition (Zarantonello, 1971, Lemma 2.2) gives
$$\gamma(x) = P_C(\gamma(x)) + P_{C^\circ}(\gamma(x)).$$
But $P_C(\gamma(x)) = \gamma(x) - \gamma^X(x)$, which yields the desired result. $\square$

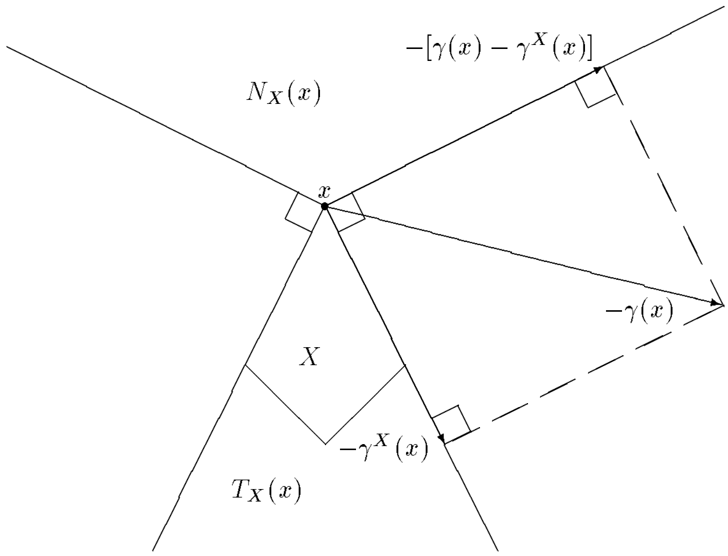

Figure 2: *Illustration of the Moreau decomposition.*

Hence, the projected subgradient $\gamma^X(x)$ equals the projection of $\gamma(x)$ onto $-T_X(x)$; this is illustrated in Figure 2. (Further, the program [DSP] is the *Wolfe dual* of [SP].)

When the set X satisfies the *Guignard constraint qualification* (Guignard, 1969), which in particular holds for polyhedral sets, the cones N_X and T_X are polyhedral (Gould and Tolle, 1971), and the problems [SP] and [DSP] are (dual) quadratic programs. If $X = \{ x \in \Re^n \mid a_i^T x \leq b_i,\ i = 1, \ldots, m \}$, then (6) and (7) can be used to explicitly state [SP] as the linearly constrained problem

[SPL]

$$\min_{\gamma,\omega} \left\{ \|\gamma\|^2 \;\middle|\; \gamma = \gamma(x) + \sum_{i \in \mathcal{I}(x)} a_i \omega_i; \quad \omega_i \geq 0, \quad i \in \mathcal{I}(x) \right\},$$

and [DSP] as

[DSPL]

$$\min_{\gamma} \left\{ \|\gamma - \gamma(x)\|^2 \;\middle|\; a_i^T \gamma \geq 0, \quad i \in \mathcal{I}(x) \right\}.$$

3.2 Method analysis

For an $x^t \in X$, letting $\gamma(x^t) \in \partial f(x^t)$ and $\gamma^X(x^t)$ be given by [DSP], the *subgradient projection method* is given by

$$x^{t+1} = P_X \left(x^t - \alpha_t \cdot \gamma^X(x^t) \right), \qquad t = 0, 1, \ldots. \tag{17}$$

Observe that if X is polyhedral then the direction $-\gamma^X(x^t)$ is feasible, and the projection (2) onto X is therefore not needed when the step length is small enough.

If

$$\sup_t \{ \|\gamma(x^t)\| \} < \infty, \tag{18}$$

then, from Lemma 1, $\sup_t\{\|\gamma^X(x^t)\|\} < \infty$. In particular, the Condition (18) holds if X is bounded or if f is piecewise linear with a finite number of segments; similarly to Shor (1985, p. 27) one may also show that for any $x^0 \in X$ there exists an $a(x^0) > 0$ such that if $\sup_t\{\alpha_t\} < a(x^0)$ then (18) holds.

Below, Corollary 2 follows from Theorems 1 and 2, and Lemmata 1 and 2.

Corollary 2 (Convergence using divergent series step lengths). *Let $\{x^t\}$ be generated by the method (17), (3) applied to* [P]. *If X^* is bounded and $\sup_t\{\|\gamma(x^t)\|\} < \infty$, then $\{f(x^t)\} \to f^*$ and $\{d_{X^*}(x^t)\} \to 0$. If $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$, then $\{x^t\}$ converges to an element of X^*.* □

From the continuity on $\Re^n$ of the convex function f, it follows that the sequence $\{x^t\}$ cannot be bounded if the sequence $\{\gamma(x^t)\}$ is unbounded. Hence, the sequence $\{x^t\}$ converges to X^* *if and only if* the Condition (18) holds.

Corollary 3 (Convergence using Polyak step lengths). *Let $\{x^t\}$ be generated by the method (17), (15) (with $\gamma^X(x^t)$ being a projected subgradient) applied to* [P]. *Then $\{f(x^t)\} \to f^*$, while $\{x^t\}$ converges to an element of X^*.*

Proof. The sequence $\{x^t\} \subseteq \{\, x \in X \mid \|x - x^*\| \leq \|x^0 - x^*\| \,\}$, where $x^* \in X^*$ (cf. Polyak, 1969, Theorem 1). Hence, $\{x^t\}$ is bounded and, from the continuity on $\Re^n$ of the convex function f, the Condition (18) holds. Apply Theorem 3 and Lemmata 1 and 2. □

If the weak sharpness Condition (16) is satisfied, then it follows from Theorem 4 and Lemmata 1 and 2 that the method (17), (15) converges linearly.

Corollary 4 (Linear convergence under a weak sharpness condition). *Under the assumptions of Corollary 3 and the Condition (16), the method (17), (15) (with $\gamma^X(x^t)$ being a projected subgradient) applied to* [P] *exhibits linear convergence towards an element $x^* \in X^*$, i.e., $\|x^t - x^*\| \leq Kq^t$, where $K \geq 0$ and $q = (1 - \varepsilon_1\varepsilon_2(\mu/c)^2)^{1/2} < 1$ with $c = \sup_t\{\|\gamma^X(x^t)\|\}$.* □

Example 2 (Enhanced convergence using projected subgradients). To show the effect of subgradient projection, we apply the method (17), (15) to the instance of Example 1, with $x^0 = (0,1)^T$ and $\theta_t \equiv 1$. Then $\gamma(x^0) = (1,2)^T$, $\gamma^X(x^0) = (-1/2, 1/2)^T$, and $\alpha_0 = 2$, which yields $x^1 = P_X((1,0)^T) = (1,0)^T = x^*$, i.e., the optimal solution is reached in one iteration. □

Kim and Um (1989) establish convergence of a subgradient projection method being the special case of the method (17), (15) in which the set X is required to be polyhedral and the value of the step length parameter is chosen as

$$\theta_t \in \begin{cases} [\varepsilon_1, \overline{\theta}_t]\,, & \text{if } \varepsilon_1 \leq \overline{\theta}_t \leq 2 - \varepsilon_2, \\ [\varepsilon_1, 2 - \varepsilon_2]\,, & \text{otherwise,} \end{cases} \tag{19}$$

where the value

$$\overline{\theta}_t = \min_i \left\{ \frac{a_i^T x^t - b_i}{a_i^T \gamma^X(x^t)} \cdot \left(\frac{f(x^t) - f^*}{\|\gamma^X(x^t)\|^2} \right)^{-1} \;\middle|\; i \notin \mathcal{I}(x^t) \text{ such that } a_i^T \gamma^X(x^t) < 0 \right\}$$

corresponds to the maximal feasible step length (15) from x^t in the direction $-\gamma^X(x^t)$. (Since, from (19), $\theta_t \in [\varepsilon_1, 2 - \varepsilon_2]$, convergence of their method follows from Corollary 3.) The reason for using (19) is to avoid performing the projection (2), and this is accomplished by choosing ε_1 small. It is, however, our experience (see Sections 4.2 and 4.3) that step lengths according to (19) may significantly deteriorate the performance of the method (17), (15). A theoretical explanation for this is that the rate of convergence given in Corollary 4 decreases as the value of ε_1 decreases.

The subgradient projection suggested by Svanberg and Bååthe (1985) is the solution to [DSPL] for the case when the feasible set is the nonnegative orthant $\Re^n_+$. They use the projected direction in the subgradient scheme of Camerini et al. (1975), where the step direction is a convex combination of the negative of a subgradient at the current iteration point and the previous step direction. Svanberg and Bååthe consider a Lagrangean relaxation application and observe that it is favourable to project the direction before making the convex combination.

4 Computational tests

In order to compare the performance of the proposed subgradient projection method (17), (15) to that of Polyak's (1969) and Kim and Um's (1989) methods, we have experimented with a Lagrangean relaxation scheme for the uncapacitated facility location problem, a nondifferentiable optimization approach to the two-person zero-sum matrix game, and a combined right-hand-side allocation and Lagrangean relaxation scheme for minimum cost multicommodity network flows. The first two applications were chosen for the purpose of illustration only, while in the third application the subgradient projection also leads to an improved performance of a well known, efficient solution approach. In all three applications, the projection problems [DSPL] and (2) have simple constraints and may thus be solved analytically or at a low computational cost. Therefore, the computational efforts per iteration of the three methods are similar, and the comparison is made with respect to the number of iterations. Further, in all three applications the convergence Condition (18) is fulfilled. In the first and the third application, the optimal values are not known a priori; when referring to the formulas (4) or (15), the optimal value f^* is for a minimization (maximization) problem then understood to be replaced by the best lower (upper) bound $\underline{f}$ ($\overline{f}$) found so far (e.g., Fisher, 1981).

4.1 The uncapacitated facility location problem

Assume that a number of facility locations is to be chosen among m alternatives and that each of n customers is to be assigned to exactly one of the located facilities. Let f_i be the cost for locating facility i, c_{ij} the cost for assigning customer j to facility i, and k and p be the lower and upper limits, respectively, on the number of located facilities. The problem of minimizing the sum of assignment and location costs subject to these restrictions is

[LOC]

$$\begin{array}{llcll}
\min & \displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} + \sum_{i=1}^{m} f_i y_i, & & & \\
\text{s.t.} & \displaystyle\sum_{i=1}^{m} x_{ij} & = & 1, & j=1,\dots,n, \\
& x_{ij} - y_i & \le & 0, & i=1,\dots,m, \quad j=1,\dots,n, \\
& k \le \displaystyle\sum_{i=1}^{m} y_i & \le & p, & \\
& x_{ij}, \quad y_i & \in & \{0,1\}, & i=1,\dots,m, \quad j=1,\dots,n.
\end{array}$$

We consider solving [LOC] heuristically by a standard Lagrangean relaxation scheme (e.g., Fisher, 1981). The variable upper bound constraints, $x_{ij} \le y_i$, are dualized with Lagrange multipliers u_{ij}, for all i, j. The Lagrangean dual problem is given by $\max_{u \ge 0} h(u)$, where $h : \Re^{mn}_+ \mapsto \Re$

is concave and piecewise linear with a finite number of segments (and the Condition (18) will therefore be fulfilled).

The method (1) applied to the dual problem consists of the following steps. Let $u^0 \in \Re^{mn}_+$. Given an iterate $u^t \geq 0$, evaluate $h(u^t)$, which gives a lower bound on the optimal value of [LOC]. A subgradient of $h(u^t)$ is given by $w_{ij}(u^t) = x_{ij}(u^t) - y_i(u^t)$, for all i, j, where $[x(u^t), y(u^t)]$ is found when evaluating $h(u^t)$. An upper bound on the optimal value is calculated by solving the restriction of [LOC] obtained from fixing $y = y(u^t)$. A step computed according to the formula (4) is taken in the direction of the subgradient $w(u^t)$, and the resulting point is projected onto $\Re^{mn}_+$, i.e., $u^{t+1}_{ij} = \max\{0, u^t_{ij} + \alpha_t w_{ij}(u^t)\}$ for all i, j. The procedure is then repeated from the new iterate u^{t+1}.

In the application of the subgradient projection method, the subgradient $w(u^t)$ is, in the above procedure, replaced by its projection onto the tangent cone of $\Re^{mn}_+$ at u^t, i.e.,

$$w^+_{ij}(u^t) = \begin{cases} \max\{0, w_{ij}(u^t)\}, & \text{if } u^t_{ij} = 0, \\ w_{ij}(u^t), & \text{if } u^t_{ij} > 0, \end{cases} \qquad \forall i, j.$$

The step length is computed according to the formula (15).

For the computational experiments we generated 10 instances of [LOC], each with 50 possible facility locations and 100 customers. The coordinates of the facility and customer sites are uniformly distributed between 0 and 100 and the assignment costs are the Euclidean distances between sites. The number of located facilities is between $k = 10$ and $p = 15$. We constructed two test series by setting the location costs to 25 and 50, respectively, for each instance.

In a preparatory experiment, we calibrated the step length parameter θ_t for each method and test series. Then, each instance was solved by each of the three methods and in every iteration the lower and upper bounds, normalized with the optimal values (which were computed in advance), were recorded. Finally, for each test series and every iteration, we computed the average values of these normalized bounds over the 10 instances.

Figure 3 shows the normalized bounds versus iterations for the test series with location costs $f_i = 25$, for all i. Here, for all t, θ_t equals 1.9 in Polyak's method, 1.0 in the subgradient projection method, and 1.1 in the method by Kim and Um. (As did Kim and Um, 1989, we used $\varepsilon_1 = 10^{-6}$.) The relative difference between the average values of the normalized bounds becomes less than 1.0% at iteration 50 of the subgradient projection method and at iteration 53 of Kim and Um's method. After 100 iterations of Polyak's method, the relative difference 1.3% was obtained. For this test series, the average number of located facilities was 14.6.

For the test series with location costs $f_i = 50$, for all i, we obtained the result shown in Figure 4. Here, for all t, θ_t equals 1.9 in Polyak's method, 1.1 in the subgradient projection method, and 1.5 in the method by Kim and Um. The relative difference between the average values of the normalized bounds reaches below 1.0% at iteration 73 of the subgradient projection method and at iteration 84 of Kim and Um's method. For Polyak's method, the relative difference did not reach below 5.4% during the first 100 iterations. For these location costs, on the average 10.8 facilities were located.

Figures 3 and 4 show that the subgradient projection strategy in this application significantly enhances the convergence rate of subgradient optimization. The poor performance of Polyak's method is due to the large number of active nonnegativity constraints at an optimal dual solution, which results in a zig-zagging behaviour similar to that illustrated in Figure 1.

4.2 The two-person zero-sum matrix game

Consider a two-person zero-sum matrix game in which the two players have n and m choices, respectively, which they choose with certain probabilities. If player one chooses alternative j and player two chooses alternative i, then player one looses a_{ij} monetary units to player two;

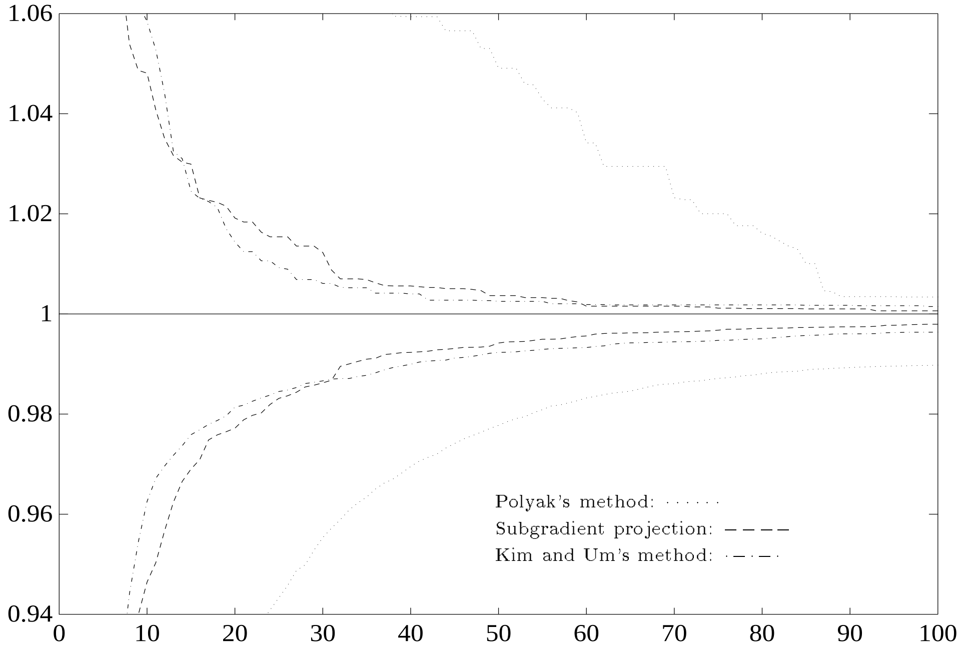

Figure 3: *Uncapacitated facility location with the lower location costs.*

the $m \times n$-matrix $A = (a_{ij})$ is referred to as the loss matrix of player one. Assume that player one chooses alternative j with the probability x_j, $j = 1, \ldots, n$, and that he/she wishes to find a strategy $x = (x_j)$ which minimizes the expected loss, irrespective of the strategy of player two; similarly, player two tries to minimize his/her expected loss irrespective of player one's strategy. The problems of the two players define a primal-dual pair of linear programs. Further, if A is skew symmetric, i.e., if $A^T = -A$, then these dual programs are equivalent, with optimal values zero; such games are called fair.

We consider formulating and solving the game problem as a nondifferentiable optimization problem. Introducing the unit simplex $S = \{ x \in \Re^n \mid \sum_{j=1}^n x_j = 1;\ x \geq 0 \}$, the problem of player one is (e.g., Tamarit Goerlich, 1988)

[PLAY]

$$f^* = \min_{x \in S} f(x),$$

where $f(x) = \max_{i \in \{1,\ldots,m\}} a_i^T x$, with a_i^T denoting the i:th row of A. The function $f : \Re^n \mapsto \Re$ is convex and piecewise linear with at most m segments (and the Condition (18) will therefore be fulfilled). The subdifferential of f at x is given by $\partial f(x) = \texttt{conv}\ \{ a_i,\ i \in \{1, \ldots, m\} \mid a_i^T x = f(x) \}$, where conv denotes the convex hull, and can easily be obtained when evaluating f at x. We consider fair games, so that $f^* = 0$.

Solution methods for two-person zero-sum matrix games based on nondifferentiable optimization formulations have been proposed earlier. Korpelevič (1979) considers non-fair games and applies Polyak's method to a problem (with optimal value zero) in the product space of the two players. Tamarit Goerlich (1988) discusses a descent method applied to [PLAY].

Polyak's method (1), (4) applied to [PLAY] results in the following steps. Given an iterate $x^t \in S$, find an i_t such that $a_{i_t}^T x^t = f(x^t)$, and let $x^{t+\frac{1}{2}} = x^t - \theta_t f(x^t)\, a_{i_t}/\|a_{i_t}\|^2$. The next iterate is $x^{t+1} = P_S(x^{t+\frac{1}{2}})$. This projection can be performed through a Lagrangean dualization of the convexity constraint (e.g., Held et al., 1974), with multiplier u, say. The evaluation of the corresponding Lagrangean dual objective function $h(u)$ yields $y(u)$, defined by $y_j(u) = \max\{0, x_j^{t+\frac{1}{2}} - u/2\}$, for all j. The optimal multiplier value is found by solving the one-

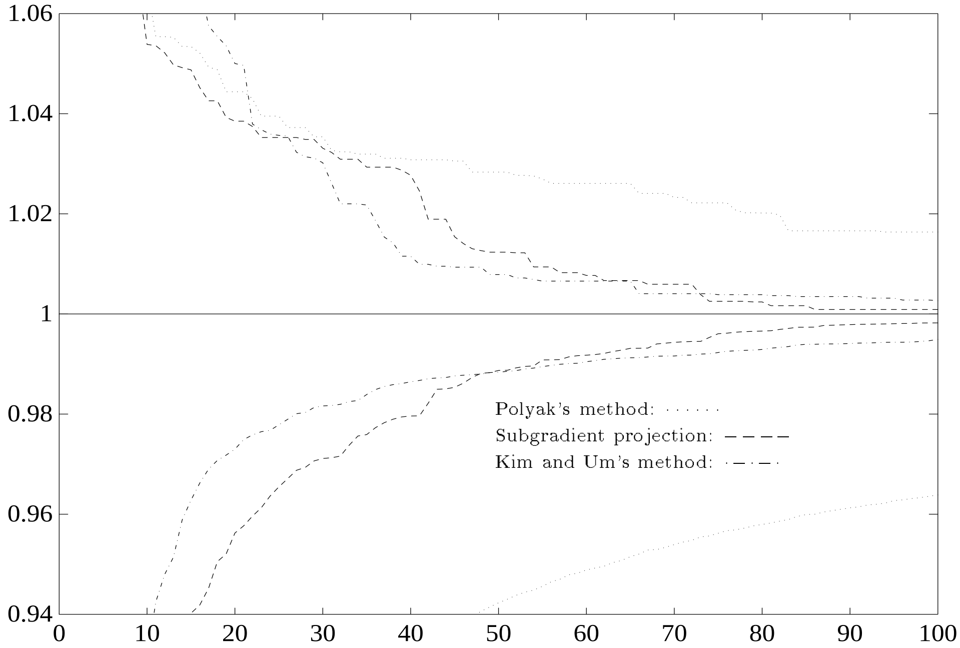

Figure 4: *Uncapacitated facility location with the higher location costs.*

dimensional equation $h'(u) = 0$, where the derivative $h'(u) = \sum_{j=1}^{n} y_j(u) - 1$ is a nonincreasing and piecewise linear function, by a simple line search based on the computation of the break-points. (This line search can be done in linear time by using a median search with respect to the break-points; see, e.g., Brucker, 1984.) The solution u^* yields the projection $x^{t+1} = y(u^*)$.

The projection $v^S(x^t)$ of the subgradient a_{i_t} is found by solving [DSPL], which in this application is given by

$$v^S(x^t) = \arg\min \left\{ \|y - a_{i_t}\|^2 \;\middle|\; \sum_{j=1}^{n} y_j = 0; \quad y_j \le 0, \quad j \in \mathcal{J}(x^t) \right\},$$

where $\mathcal{J}(x^t) = \{\, j \mid x_j^t = 0 \,\}$. This projection problem is solved analogously to the one above by dualizing the equality constraint, with multiplier u, say. The optimal multiplier u^* is obtained by solving $h'(u) = 0$, where $h'(u) = \sum_{j=1}^{n} y_j(u)$ and $y : \Re \mapsto \Re^n$ is given by

$$y_j(u) = \begin{cases} \min\{0, a_{i_t j} - u/2\}, & j \in \mathcal{J}(x^t), \\ a_{i_t j} - u/2, & j \notin \mathcal{J}(x^t). \end{cases}$$

The subgradient projection is given by $v^S(x^t) = y(u^*)$. The step length in the direction $-v^S(x^t)$ is in the subgradient projection method calculated according to (15).

We now turn to the generation of the test problems, and first observe that if the upper triangular part of A contains only positive elements, then the optimal solution is given by a pure strategy (i.e., an extreme point of S). If, on the other hand, the elements of the upper triangular part of A are uniformly distributed on an interval which is symmetric around zero, then an optimal solution is (as indicated by experiments) a mixed strategy with around 50% of the variables being positive. We created instances of [PLAY] with varying problem characteristics by taking convex combinations of randomly generated loss matrices of these two extreme types. We chose $m = n = 100$, and generated 20 instances of each problem characteristic; the portion (in percentage) of positive variables in an optimal solution, taken as the average of the 20 instances, is denoted by p.

Whenever the optimal value of the program [P] is known a priori, the methods (1), (4) and (17), (15) are (according to Theorem 3 of Polyak, 1969, and our Corollary 4, respectively)

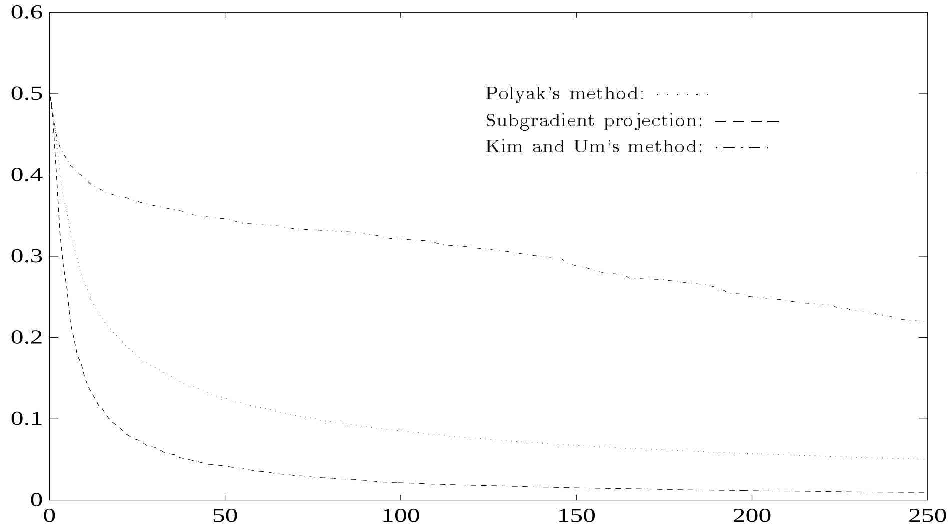

Figure 5: *Two-person zero-sum matrix game, $p = 5$.*

linearly convergent, with maximal rates of convergence for $\varepsilon_1 = \varepsilon_2 = 1$; in this application we therefore chose $\theta_t \equiv 1$. As did Kim and Um (1989), we used $\varepsilon_1 = 10^{-6}$ in their method.

As an example of the convergence behaviour of the three methods, Figure 5 shows the average objective values over 20 instances (with $p = 5$) versus iterations; the methods were initiated at $x_j^0 = 1/n$, $j = 1, \ldots, n$. Here, the subgradient projection method performs much better than the other two; Kim and Um's method in particular exhibits very slow convergence. The subgradient projection method needed on the average 87 iterations to reach below 5% of the initial objective value. For Polyak's method and the method by Kim and Um this number was 834 and 4476, respectively.

In order to investigate how the three methods perform for varying problem characteristics, we chose ten values of p in the interval $[4, 52]$ and recorded after 250 iterations, for each of these choices and each method, the average objective values over 20 instances; the methods were again initiated at $x_j^0 = 1/n$, $j = 1, \ldots, n$. The average objective values obtained by Polyak's and Kim and Um's methods were normalized with respect to those obtained by the subgradient projection method. Figure 6 shows these normalized average values versus the portion p.

Clearly, the subgradient projection method becomes increasingly better than the other two methods as the optimal solutions tend towards pure strategies. For small values of p, almost all of the nonnegativity constraints will be active in an optimal solution, and, as in the facility location application, Polyak's method then suffers from zig-zagging. In Kim and Um's method, the steps taken will become very short when an optimal solution is approached from a starting point located in the relative interior of the unit simplex. For large values of p, relatively few nonnegativity constraints are active in an optimal solution. Further, a typical row of a loss matrix which gives a large value of p is nearly orthogonal to a normal of the convexity constraint; therefore, a subgradient of f then defines a near-feasible direction with respect to this constraint. Thus, for large values of p, the step directions as well as the asymptotic performances of the three methods are similar.

To investigate the stability of the methods with respect to their initialization, we then changed the initial solution for each instance to its best pure strategy. The result is shown in Figure 7. The notable difference from the result of the previous experiment is in the performance of

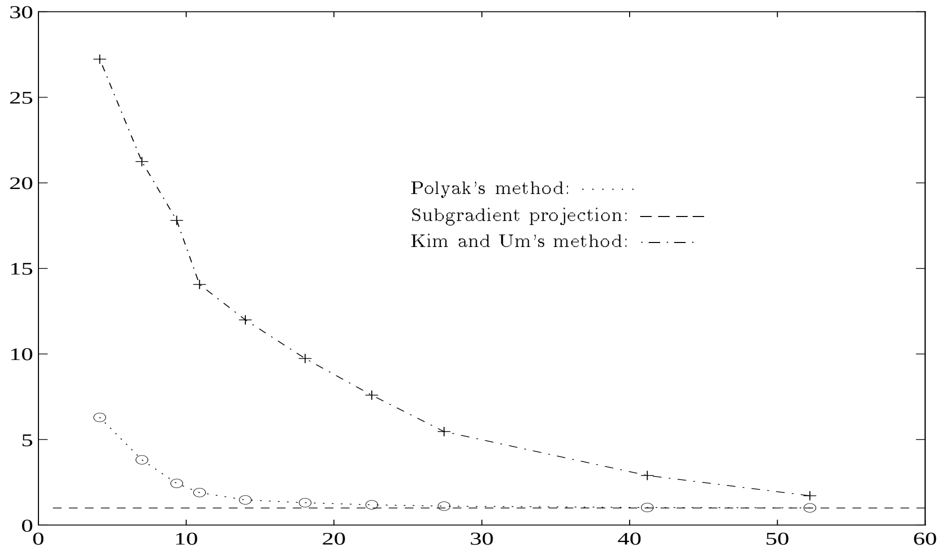

Figure 6: *Normalized objective values, initial solution at $x_j^0 = 1/n$ for all j.*

Kim and Um's method. In that experiment, Kim and Um's method suffered from taking very small steps when approaching the relative boundary. In this experiment, a similar phenomenon occurred when trying to leave the relative boundary. Note that the negative effects of the step length truncations now seem to be even more severe, since the performance is poor for all values of p.

We conclude that the subgradient projection method is robust with respect to the choice of starting point, and that it has a higher overall efficiency than the other two methods. Concerning Kim and Um's method, our experience is that the shortening of the step, done in order to reduce the computations in each iteration, is the reason for its poor performance.

4.3 The minimum cost multicommodity network flow problem

Let $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ denote a directed network, with $\mathcal{N}$ and $\mathcal{A}$ being its sets of nodes and directed arcs, respectively. Further, let $\mathcal{K}$ be a set of flow commodities, d_k be the demand vector for commodity $k \in \mathcal{K}$, and x_k be the vector of arc flows for commodity $k \in \mathcal{K}$. Associated with the commodity flows are vectors of unit transportation costs c_k, $k \in \mathcal{K}$, and the total arc flows are upper bounded by b. Letting A denote the node-arc incidence matrix for the network $\mathcal{G}$, the minimum cost multicommodity network flow problem is

[MCNFP]

$$
\begin{aligned}
f^* = \min \quad & \sum_{k \in \mathcal{K}} c_k^T x_k, \\
\text{s.t.} \quad & A x_k = d_k, \qquad k \in \mathcal{K}, \\
& \sum_{k \in \mathcal{K}} x_k \leq b, \\
& x_k \geq 0, \qquad k \in \mathcal{K}.
\end{aligned}
$$

Kennington and Shalaby (1977) suggest the following solution procedure, which is based on

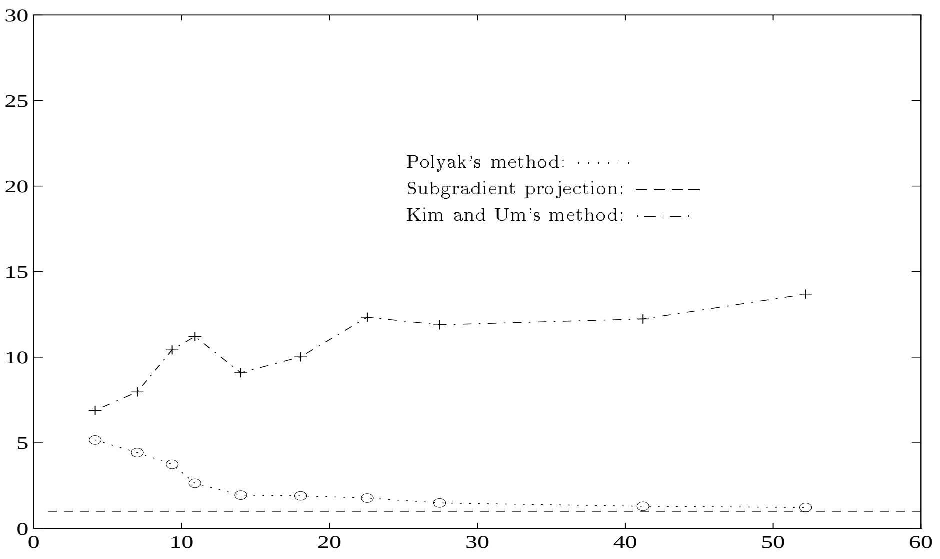

Figure 7: *Normalized objective values, initial solution at the best pure strategy.*

allocations of the capacities of the total arc flows to separate commodities. (See, e.g., Chapter 9 of Lasdon, 1970, for a general description of the right-hand-side allocation technique.)

Let $y = (y_k)_{k\in\mathcal{K}}$, where $y_k \in \Re_+^{|\mathcal{A}|}$, $k \in \mathcal{K}$, denote a commodity capacity allocation. Then [MCNFP] can be equivalently formulated as the master problem

[MP]

$$\min_{y\in Y} g(y),$$

where $Y = \{\, y \in \Re^{|\mathcal{A}|\cdot|\mathcal{K}|} \mid \sum_{k\in\mathcal{K}} y_k = b;\ y \geq 0 \,\}$ is the set of feasible allocations, and $g : \Re_+^{|\mathcal{A}|\cdot|\mathcal{K}|} \mapsto \Re$ is the implicit function

[SUB]

$$g(y) = \sum_{k\in\mathcal{K}} \min \left\{\, c_k^T x_k \;\middle|\; A x_k = d_k; \quad 0 \leq x_k \leq y_k \,\right\}, \qquad y \in \Re_+^{|\mathcal{A}|\cdot|\mathcal{K}|}.$$

Its evaluation amounts to the solution of $|\mathcal{K}|$ single-commodity minimum cost network flow problems. We presume, without any loss of generality, that g is finite on $\Re_+^{|\mathcal{A}|\cdot|\mathcal{K}|} \supseteq Y$; this can always be achieved by including artificial arcs with associated large costs and capacities in [MCNFP]. (However, g is not finite outside $\Re_+^{|\mathcal{A}|\cdot|\mathcal{K}|}$.) Letting $u_k \in \Re^{|\mathcal{N}|}$ and $v_k \in \Re_-^{|\mathcal{A}|}$, $k \in \mathcal{K}$, be the linear programming dual variables associated with the flow conservation and arc capacity constraints, respectively, of the $|\mathcal{K}|$ single-commodity problems of [SUB], and letting $(u_k^p, v_k^p)_{p\in\mathcal{P}_k}$, $k \in \mathcal{K}$, be the extreme points of the corresponding dual feasible sets, the function g may be expressed as

[DSUB]

$$g(y) = \sum_{k\in\mathcal{K}} \max_{p\in\mathcal{P}_k} \left\{\, d_k^T u_k^p + y_k^T v_k^p \,\right\}, \qquad y \in \Re_+^{|\mathcal{A}|\cdot|\mathcal{K}|}.$$

Hence, this function is convex and piecewise linear with a finite number of segments. From this expression of g it is also clear that if, for some $y \in Y$, $(u_k(y_k), v_k(y_k))$, $k \in \mathcal{K}$, are optimal dual extreme points, then $v(y) = (v_k(y_k))_{k\in\mathcal{K}}$ is a subgradient of g at y.

The method of Kennington and Shalaby employs a subgradient optimization scheme for solving the nondifferentiable program [MP]. From the finiteness of the sets $\mathcal{P}_k$, $k \in \mathcal{K}$, it follows that the sequence $\{v(y^t)\}$ generated in the solution process is bounded; therefore, the Condition (18) will be fulfilled.

Letting x^0 be the solution to the relaxation of [MCNFP] obtained by replacing the set of constraints $\sum_{k\in\mathcal{K}} x_k \le b$ by $x_k \le b$, $k \in \mathcal{K}$, Kennington and Shalaby's procedure is initialized at $y^0 = P_Y(x^0)$. Given an iterate $y^t \in Y$, the value $g(y^t)$ defines an upper bound on the optimal value of [MCNFP]. A step is taken in the direction $-v(y^t)$, with step length according to the formula (4), and the point obtained, $y^{t+\frac{1}{2}} = y^t - \alpha_t v(y^t)$, is projected onto the set Y. This projection separates into single-arc problems, each of which is solved by the dual line search method outlined in Section 4.2.

Our modification of the above scheme is to replace the subgradient $v(y^t)$ by its projection, found by solving [DSPL]; this projection separates into single-arc problems which are solved by the dual method outlined in Section 4.2. The step length is computed according to the formula (15).

In order to obtain strong lower bounds on the optimal value of [MCNFP], Allen (1985) combines the procedure of Kennington and Shalaby (1977) with Lagrangean relaxation of the capacity constraints. In our application of Allen's relaxation scheme, we include in [MCNFP] the redundant commodity capacity constraints $x_k \le b$, $k \in \mathcal{K}$, thereby obtaining a stronger Lagrangean subproblem. (Allen considers a version of [MCNFP] which includes commodity capacity constraints $x_k \le b_k$, $k \in \mathcal{K}$.) Introducing a vector u of multipliers, the Lagrangean dual problem then is to find $\max_{u\ge 0} h(u)$, where

$$h(u) = \sum_{k\in\mathcal{K}} \min\left\{ (c_k+u)^T x_k \;\middle|\; Ax_k = d_k; \quad 0 \le x_k \le b \right\} \;-\; b^T u.$$

The function $h : \Re_+^{|\mathcal{A}|} \mapsto \Re$ is concave and piecewise linear with a finite number of segments (and the Condition (18) will therefore be fulfilled), and its evaluation amounts to solving $|\mathcal{K}|$ single-commodity problems.

Let $u^0 \in \Re_+^{|\mathcal{A}|}$. Given an iterate $u^t \ge 0$, the value $h(u^t)$ is a lower bound on the optimal value of [MCNFP]. A subgradient of h at u^t is given by $w(u^t) = \sum_{k\in\mathcal{K}} x_k(u^t) - b$, where $x_k(u^t)$, $k \in \mathcal{K}$, is found when calculating $h(u^t)$. A step is taken in the direction $w(u^t)$, with step length according to the formula (4), and the resulting point is projected onto the nonnegative orthant (cf. Section 4.1). In Allen's combination of this dual scheme with the procedure of Kennington and Shalaby, each main iteration consists of one dual subgradient step, which gives u^{t+1} and a lower bound on f^*, and one primal subgradient step, resulting in y^{t+1} and an upper bound on f^*.

Our modification of the dual scheme is to project the subgradient $w(u^t)$ onto the tangent cone of $\Re_+^{|\mathcal{A}|}$ at u^t (cf. Section 4.1). The step length is computed according to the formula (15).

The test problems have a transportation network structure and are randomly generated in accordance with the principles in Kennington and Shalaby (1977). We generated 15 instances, each having 10 sources, 10 sinks and 10 commodities. The supply/demand for each source/sink and commodity is in the range 0–300, arc capacities in the range 100–400, and arc costs in the range 0–100. All data is uniformly distributed. The transportation subproblems were solved using the minimum cost network flow code RELAX (Bertsekas and Tseng, 1988).

For each of the three methods and for each instance and iteration we recorded current upper and lower bounds normalized by the corresponding optimal value (which was computed by the linear programming package XMP; see Marsten, 1981). Then, for each method and iteration, we computed the average value of these normalized bounds over the 15 instances.

In our first experiment we investigated the difference in quality of the step directions of Polyak's (1969) method and the subgradient projection method, and the effect of the shortening of the step (19) made in the method by Kim and Um (1989). In order to make the comparison unbiased we chose $\theta_t \equiv 1$ in both the primal and the dual application of each of the three meth-

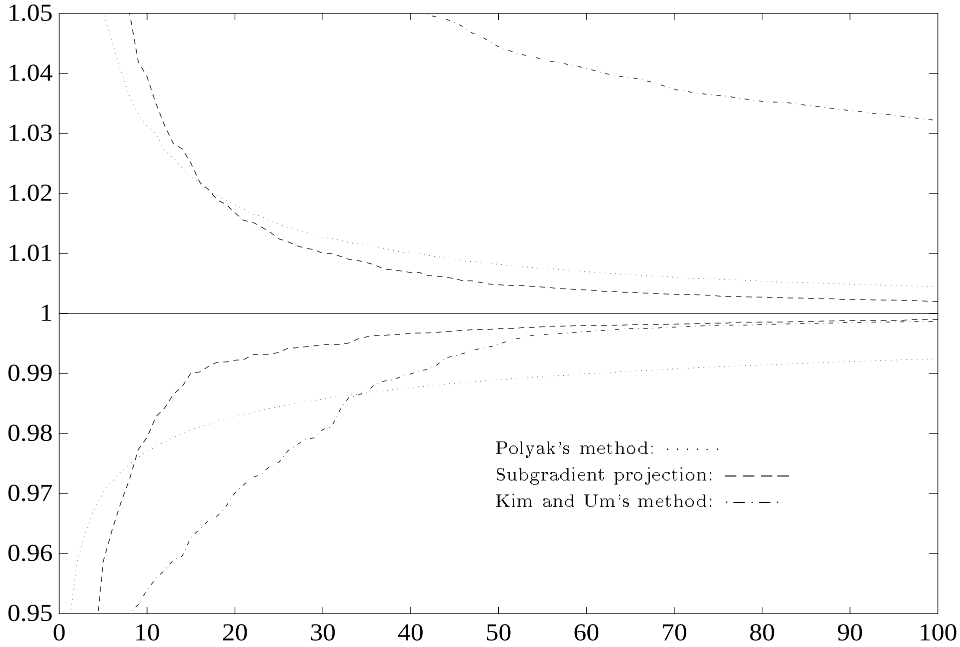

Figure 8: *Minimum cost multicommodity network flow, $\theta_t \equiv 1$ for all three methods.*

ods. (As did Kim and Um, 1989, we used $\varepsilon_1 = 10^{-6}$ in their method.) Figure 8 shows average values of normalized upper and lower bounds for the three methods versus iterations. The relative difference between the average values of the normalized bounds became less than 2.0% at iteration 48 of Polyak's method and at iteration 25 of the subgradient projection method. For Kim and Um's method, this difference did not reach below 3.3% during the first 100 iterations. A first observation is that the introduction of the projection [DSPL] yields significantly better step directions. Secondly, a shortening of the step length deteriorates the performance of the subgradient projection scheme.

In the second experiment we compared the three methods' best possible performances. For this purpose we calibrated the primal and dual step length parameters for each of the three methods. This resulted in choosing, for all t, θ_t equal to 1.9 in both the primal and the dual application of Polyak's method, 1.2 and 0.3 in the primal and dual applications of the subgradient projection method respectively, and correspondingly, 1.0 and 0.3 in the applications of Kim and Um's method. The result of the second experiment is shown in Figure 9, which was generated in the same way as Figure 8. The relative difference between the average values of the normalized bounds became less than 1.0% at iteration 63 of Polyak's method and at iteration 33 of the subgradient projection method. In the primal application of Kim and Um's method, the step length reduction resulted in a poor performance for all feasible choices of the step length parameter θ_t.

5 Conclusion and Discussion

The contribution of this paper is twofold. The theoretical contribution is the introduction of the conditional subgradient optimization method and the special case of subgradient projection, and the establishment of these methods' convergence under several step length selection rules. The computational contribution is found in the results of our experiments with three applications; these results justify the use of conditional subgradient optimization by showing that the practical rate of convergence of a subgradient optimization scheme may be significantly improved by the

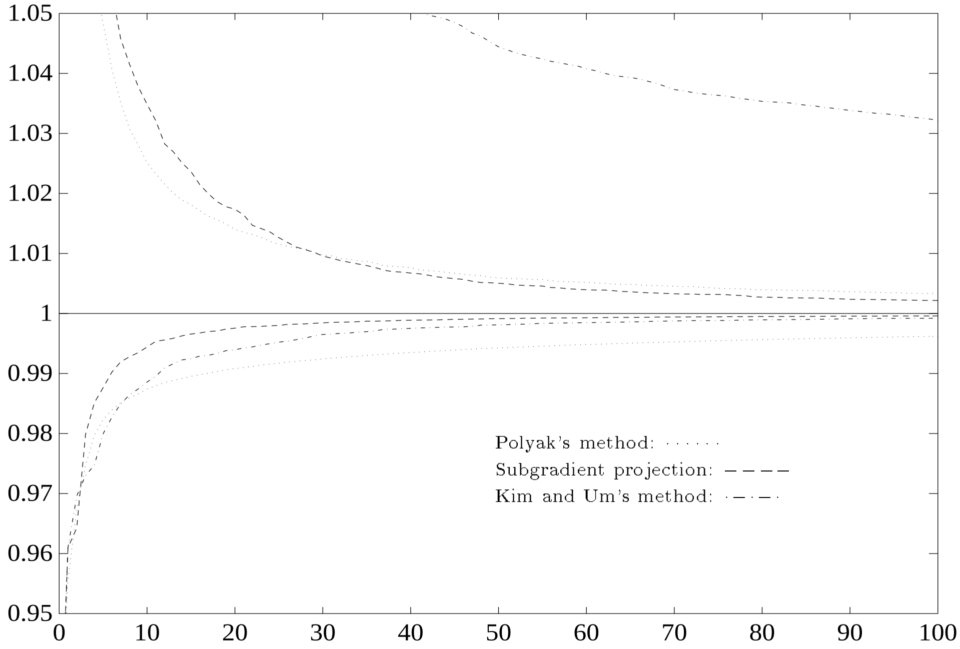

Figure 9: *Minimum cost multicommodity network flow, θ_t calibrated for each method.*

introduction of projected subgradients.

The good overall performance of the subgradient projection method compared to that of Polyak's and Kim and Um's methods may be explained by the following observations. The method by Kim and Um (1989) performs like an active set method in the sense that the set of active constraints changes very slowly; this behaviour degrades the practical performance of the method and makes it sensitive to changes in the location of the starting solution. In the subgradient projection method, the active set is taken into account only when computing the step direction. Longer steps are therefore allowed, thus enabling a more rapid change in the active set; this property makes the method less sensitive to the choice of starting solution. (In some tests, not reported here, we found that the performance of Kim and Um's method can be improved by increasing the value of ε_1 considerably, so that their method tends to perform as the subgradient projection method.) Rapid changes in the active set are typical also for Polyak's (1969) method. However, when approaching an optimal solution, the zig-zagging phenomenon illustrated in Figure 1 occurs. (As can be seen from the results of the experiments reported in Sections 4.1 and 4.3, the best performance of Polyak's method was always obtained from large values of the step length parameter, the reason being that the use of large step lengths compensate, to some extent, for the effects of the zig-zagging behaviour.)

We conclude with a discussion of some suggestions for continued research.

The method of Kennington and Shalaby (1977) for multicommodity network flows is claimed (Minoux, 1986, Section 8.5.2) to be efficient for obtaining good approximate solutions to [MC-NFP]. It is our experience that the combination of this method with Allen's (1985) lower bounding procedure improves on its performance, and further that the introduction of the subgradient projection enhances the performance of the combined scheme. A more sophisticated implementation of the combined subgradient projection scheme may therefore yield a highly efficient tool for the approximate solution of large-scale problems; it may also be applied to yield an advanced starting solution for an optimizing method (e.g., Saviozzi, 1986).

As discussed in Section 1, the subgradient projection method is most beneficially used when the feasible set is simple, and this may seem to limit its applicability. It would, however, be possible to embed the method in a simplicial decomposition scheme (e.g., von Hohenbalken, 1977), which would convert any linearly constrained nondifferentiable optimization problem into a sequence of nondifferentiable programs over unit simplices, and our method may then be applied to each of these programs. In a current endeavour to establish such a scheme, we are exploiting an ergodic sequence of directions constructed from the subgradients used within the conditional subgradient optimization method. (These directions can be used to calculate lower bounds on the optimal value.) Any limit point of this sequence of directions is both a subgradient and an element of the negative of the normal cone at an optimal solution (cf. Proposition 5); the latter property provides a means for defining a new nondifferentiable program over a unit simplex (in a higher dimension).

Acknowledgements

The research leading to this report was financially supported by the Swedish Research Council for Engineering Sciences (TFR). We thank Dr. Zhuangwei Liu for letting us use her implementation of the dual part of the multicommodity network flow algorithm, and Dr. Ulf Brännlund for drawing our attention to the report by Svanberg and Bååthe. We also thank two anonymous referees for giving constructive suggestions leading to an improved presentation, and one of them in particular for also finding a flaw in a proof.

References

Allen, E.P. (1985), *Using Two Sequences of Pure Network Problems to Solve the Multicommodity Network Flow Problem*, Dissertation, School of Engineering and Applied Science, Southern Methodist University, Dallas, TX.

Bertsekas, D.P., and Mitter, S.K. (1973), "A descent method for optimization problems with nondifferentiable cost functionals", *SIAM Journal on Control* 11, 637–652.

Bertsekas, D.P., and Tseng, P. (1988), "The RELAX codes for linear minimum cost network flow problems", *Annals of Operations Research* 13, 125–190.

Bihain, A., Nguyen, V.H., and Strodiot, J.-J. (1987), "A reduced subgradient algorithm", *Mathematical Programming Study* 30, 127–149.

Brucker, P. (1984), "An $\mathcal{O}(n)$ algorithm for quadratic knapsack problems", *Operations Research Letters* 3, 163–166.

Burke, J.V., and Ferris, M.C. (1993), "Weak sharp minima in mathematical programming", *SIAM Journal on Control and Optimization* 31, 1340–1359.

Burke, J.V., and Moré, J.J. (1988), "On the identification of active constraints", *SIAM Journal on Numerical Analysis* 25, 1197–1211.

Camerini, P.M., Fratta, L., and Maffioli, F. (1975), "On improving relaxation methods by modified gradient techniques", *Mathematical Programming Study* 3, 26–34.

Damberg, O. (1990), Solving the traffic assignment problem — a new approach using the gap function, Masters Thesis, LiTH-MAT-EX-90-06, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden.

Dem'janov, V.F., and Šomesova, V.K. (1978), "Conditional subdifferentials of convex functions", *Soviet Mathematics Doklady* 19, 1181–1185.

Dem'yanov, V.F., and Shomesova, V.K. (1980), "Subdifferentials of functions on sets", *Cybernetics* 16/1, 24–31.

Dem'yanov, V.F., and Vasil'ev, L.V. (1985), *Nondifferentiable Optimization*, Optimization Software, New York, NY.

Ermol'ev, Yu.M. (1966), "Methods for solving nonlinear extremal problems", *Kibernetika* 2/4, 1–17 (in Russian). Translated in *Cybernetics* 2/4, 1–14.

Fisher, M.L. (1981), "The Lagrangian relaxation method for solving integer programming problems", *Management Science* 27, 1–18.

Fisher, M.L. (1985), "An applications oriented guide to Lagrangian relaxation", *Interfaces* 15, 10–21.
Fletcher, R. (1987), *Practical Methods of Optimization*, Second Edition, John Wiley & Sons, Chichester, England.
Geoffrion, A.M. (1974), "Lagrangean relaxation for integer programming", *Mathematical Programming Study* 2, 82–114.
Goffin, J.L. (1978), "Nondifferentiable optimization and the relaxation method", in: C. Lemarechal and R. Mifflin (eds.), *Nonsmooth Optimization*, Proceedings of a IIASA Workshop, 1977, Pergamon Press, Oxford, 31–49.
Gould, F.J., and Tolle, W. (1971), "A necessary and sufficient qualification for constrained optimization", *SIAM Journal on Applied Mathematics* 20, 164–172.
Guignard, M. (1969), "Generalized Kuhn–Tucker conditions for mathematical programming problems in a Banach space", *SIAM Journal on Control* 10, 93–98.
Held, M., Wolfe, P., and Crowder, H.P. (1974), "Validation of subgradient optimization", *Mathematical Programming* 6, 62–88.
von Hohenbalken, B. (1977), "Simplicial decomposition in nonlinear programming algorithms", *Mathematical Programming* 13, 49–68.
Kennington, J., and Shalaby, M. (1977), "An effective subgradient procedure for minimal cost multicommodity flow problems", *Management Science* 23, 994–1004.
Kim, S., and Um, B.-S. (1989), "Polyak's subgradient method with simplified projection for nondifferentiable optimization with linear constraints", *Optimization* 20, 451–456.
Korpelevič, G.M. (1979), "A relaxation method of solution for matrix games", *Matematičeskie Metody Rešenija Ekonomičeskie Zadač*, Moscow, 8, 21–31 (in Russian).
Lasdon, L.S. (1970), *Optimization Theory for Large Systems*, Macmillan, New York, NY.
Luenberger, D.G. (1984), *Linear and Nonlinear Programming*, Second Edition, Addison-Wesley, Reading, MA.
Marsten, R.E. (1981), "The design of the XMP linear programming library", *ACM Transactions on Mathematical Software* 7, 481–497.
Minoux, M. (1986), *Mathematical Programming: Theory and Algorithms*, John Wiley & Sons, Chichester, England.
Panier, E. (1987), "An active set method for solving linearly constrained nonsmooth optimization problems", *Mathematical Programming* 37, 269–292.
Polak, E., Mayne, D.Q., and Wardi, Y. (1983), "On the extension of constrained optimization algorithms from differentiable to nondifferentiable problems", *SIAM Journal on Control and Optimization* 21, 179–203.
Poljak, B.T. (1967), "A general method of solving extremum problems", *Doklady Akademii Nauk SSSR* 174/1, 33–36 (in Russian). Translated in *Soviet Mathematics Doklady* 8/3, 593–597.
Polyak, B.T. (1969), "Minimization of unsmooth functionals", *USSR Computational Mathematics and Mathematical Physics* 9, 14–29.
Polyak, B.T. (1979), "Sharp minima", Lecture Notes, Institute of Control Sciences, Moscow, USSR. Presented at the IIASA Workshop on Generalized Lagrangians and their Applications, IIASA, Laxenburg, Austria, 1979.
Polyak, B.T. (1987), *Introduction to Optimization*, Optimization Software Inc., Publications Division, New York, NY.
Rockafellar, R.T. (1970), *Convex Analysis*, Princeton University Press, Princeton, NJ.
Rockafellar, R.T. (1981), *The Theory of Subgradients and its Applications to Problems of Optimization: Convex and Nonconvex Functions*, Research and Education in Mathematics 1, Heldermann Verlag, Berlin.
Rosen, J.B. (1960), "The gradient projection method for nonlinear programming, part I: linear constraints", *Journal of the Society for Industrial and Applied Mathematics* 8, 181–217.
Saviozzi, G. (1986), "Advanced start for the multicommodity network flow problem", *Mathematical Programming Study* 26, 221–224.

Shepilov, M.A. (1976), "Method of the generalized gradient for finding the absolute minimum of a convex function", *Cybernetics* 12, 547–553.

Shor, N.Z. (1985), *Minimization Methods for Non-Differentiable Functions*, translated from the Russian by K.C. Kiwiel and A. Ruszczyński, Springer-Verlag, Berlin.

Shor, N.Z. (1991), "The development of numerical methods for nonsmooth optimization in the USSR", in: J.K. Lenstra, A.H.G. Rinnoy Kan and A. Schrijver (eds.), *History of Mathematical Programming: A Collection of Personal Reminiscences*, CWI and North-Holland, Amsterdam, The Netherlands, 135–139.

Sreedharan, V.P. (1987), "ε-subgradient projection algorithm", *Journal of Approximation Theory* 51, 27–46.

Strodiot, J.-J., Nguyen, V.H., and Heukemes, N. (1983), "ε-optimal solutions in nondifferentiable convex programming and some related questions", *Matematical Programming* 25, 307–328.

Svanberg, K., and Bååthe, O. (1985), "Computational experiments with different subgradient techniques", TRITA-MAT-1985-3, The Swedish Institute of Applied Mathematics, Sweden.

Tamarit Goerlich, J.M. (1988), "Improving feasible directions for a class of nondifferentiable functions", *European Journal of Operational Research* 34, 99–104.

Zarantonello, E.H. (1971), "Projections on convex sets in Hilbert space and spectral theory", in: E.H. Zarantonello (ed.), *Contributions to Nonlinear Functional Analysis*, Proceedings of a Symposium Conducted by the Mathematics Research Center, University of Wisconsin, Madison, WI, April 1971, Academic Press, New York, NY, 237–424.