Working Paper

Decomposition via Alternating Linearization

Krzysztof C. Kiwiel
Charles H. Rosa
Andrzej Ruszczyński

WP-95-051
June 1995

IIASA International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 807 □ Fax: +43 2236 71313 □ E-Mail: info@iiasa.ac.at

Decomposition via Alternating Linearization

Krzysztof C. Kiwiel
Charles H. Rosa
Andrzej Ruszczyński

WP-95-051
June 1995

International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 807 □ Fax: +43 2236 71313 □ E-Mail: info@iiasa.ac.at

Abstract

A new approximate proximal point method for minimizing the sum of two convex functions is introduced. It replaces the original problem by a sequence of regularized subproblems in which the functions are alternately represented by linear models. The method updates the linear models and the prox center, as well as the prox coefficient. It is monotone in terms of the objective values and converges to a solution of the problem, if any. A dual version of the method is derived and analyzed. Applications of the methods to multistage stochastic programming problems are discussed and preliminary numerical experience presented.

Key words. Convex programming, large scale optimization, decomposition, proximal point methods, augmented Lagrangians, stochastic programming.

Decomposition via Alternating Linearization

*Krzysztof C. Kiwiel**
Charles H. Rosa
Andrzej Ruszczyński

1 Introduction

We present a method for solving structured convex optimization problems of the form:

$$\text{minimize} \quad F(x) := h(x) + f(x), \tag{1.1}$$

where $h : \mathbb{R}^n \to (-\infty, +\infty]$ and $f : \mathbb{R}^n \to \mathbb{R}$ are closed proper convex functions.

Our method is an approximate version of the proximal point algorithm [Mar70, Roc76b] which generates a sequence

$$x^{k+1} = \arg\min\{ F(x) + \frac{\rho_k}{2}|x - x^k|^2 \} \quad \text{for } k = 1, 2, \ldots, \tag{1.2}$$

starting from any point $x^1 \in \mathbb{R}^n$, where $|\cdot|$ is the Euclidean norm and $\{\rho_k\}$ is a sequence of positive numbers. To implement the iteration (1.2), our method employs a sequence of subproblems of the form:

$$\min\left\{ h(x) + \tilde{f}^k(x) + \frac{\rho_k}{2}|x - x^k|^2 \right\} \tag{1.3}$$

and

$$\min\left\{ \tilde{h}^k(x) + f(x) + \frac{\rho_k}{2}|x - x^k|^2 \right\}, \tag{1.4}$$

where $\tilde{f}^k$ and $\tilde{h}^k$ are linear models of f and h, respectively. This is the reason for baptizing our approach the *alternating linearization method.*

Our method makes it possible to exploit structural properties of h and f separately, which may be useful in many applications. Let us just mention two examples, which will be treated in more detail later.

Example 1.1. Consider the separable problem with linking constraints:

$$\min \sum_{j=1}^{N} \psi_j(x_j), \quad \text{s.t.} \sum_{j=1}^{N} A_j x_j = b,$$

*Systems Research Institute, Newelska 6, 01–447 Warsaw, Poland (kiwiel@ibspan.waw.pl)

where $\psi_j : \mathbb{R}^{n_j} \to (-\infty, +\infty]$ are closed proper convex functions and A_j are $m \times n_j$ matrices, $j = 1, \ldots, N$. Application of the *multiplier method* [Ber82, Hes69, Pow69, Roc76a] leads to subproblems of minimizing the augmented Lagrangian:

$$\min_x \left\{ \sum_{j=1}^{N} (\psi_j(x_j) - \langle \lambda, A_j x_j \rangle) + \langle \lambda, b \rangle + \frac{\rho}{2} |Ax - b|^2 \right\},$$

where $\lambda \in \mathbb{R}^m$ is the current vector of Lagrange multipliers, $\rho > 0$ is a penalty coefficient, $x = (x_1, \ldots, x_N)$ and $A = [A_1 \ \cdots \ A_n]$. This problem has the form (1.1) with $f(x) = \rho|Ax - b|^2/2$, in which (1.3) is decomposable into independent subproblems for each $j = 1, \ldots, N$, while (1.4) is just a least-squares problem.

Example 1.2. Let us now consider the decomposable problem with linking variables:

$$\min \left\{ \varphi(y) + \sum_{j=1}^{N} \psi_j(y) \right\}$$

with closed proper convex functions $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ and $\psi_j : \mathbb{R}^n \to (-\infty, +\infty]$, $j = 1, \ldots, N$. Splitting variables and dualization [BeT89, p. 231] leads to the problem:

$$\min_x \left\{ \sum_{j=1}^{N} \psi_j^*(x_j) + \varphi^* \left(-\textstyle\sum_{j=1}^{N} x_j \right) \right\},$$

where φ^* and ψ_j^* are the conjugates of φ and ψ_j, and $x_j \in \mathbb{R}^n$, $j = 1, \ldots, N$, are dual variables. This dual problem has the form (1.1), in which (1.3) decomposes into independent subproblems for $j = 1, \ldots, N$. All these subproblems and (1.4) are much easier to solve than the original formulation.

The general objective of our work has been pursued by many researchers; in particular the well-known *operator splitting methods* should be mentioned here (see [Eck94, EcB92, EcF94a, MOT95, MaT92, Spi85, Tse91, Tse90]). Their dual versions are known as *alternating direction methods* [BeT89, EcB92, EcF94b, Fuk92, Gab83]. Other related recent research is described in [ChT94, Tse94].

Our approach, although having parallel objectives, is fundamentally different. Contrary to earlier works, our method is *monotone* in terms of the values of the objective $F = h + f$. To achieve this, we employ two different types of updates of the models in (1.3) and (1.4). The first update changes only the approximations $\tilde{f}^k$ and $\tilde{h}^k$, while keeping x^k fixed; the second one updates x^k as well. In this way we ensure that $F(x^{k+1}) < F(x^k)$ whenever x^k is changed. We also allow changes in the value of the penalty coefficient ρ_k. On the other hand, our method is less general than some other ones because it requires that f be finite-valued; this, however, does not seem to limit its usefulness, at least in the applications that are of interest to us.

In §2 we present the main idea of the method: approximate implementation of the proximal step by using alternating linearizations. In §3 this idea is used within a descent algorithm for minimizing F. Its convergence is proved in §4. The dual version of the method is described in §5. In §6 we discuss applications to stochastic programming. Preliminary computational experience is reported in §7.

2 Proximal step by alternating linearization

Let us first describe and analyse an algorithm that employs subproblems (1.3)–(1.4) for finding an approximation to the proximal point

$$p(\bar{x}) = \arg\min\left\{h(x) + f(x) + \frac{\rho}{2}|x - \bar{x}|^2\right\}, \tag{2.1}$$

where $\bar{x} \in \mathbb{R}^n$ and $\rho > 0$ are fixed.

Algorithm 2.1.
Step 0: Choose $z_f^0 \in \mathbb{R}^n$ and $g_f^0 \in \partial f(z_f^0)$. Define $\tilde{f}^1(\cdot) = f(z_f^0) + \langle g_f^0, \cdot - z_f^0\rangle$. Set $k = 1$.

Step 1: Find the solution z_h^k of the problem:

$$\min_x\left\{h(x) + \tilde{f}^k(x) + \frac{\rho}{2}|x - \bar{x}|^2\right\}. \tag{2.2}$$

Set

$$g_h^k = -g_f^{k-1} - \rho(z_h^k - \bar{x}) \tag{2.3}$$

and define

$$\tilde{h}^k(\cdot) = h(z_h^k) + \langle g_h^k, \cdot - z_h^k\rangle. \tag{2.4}$$

Step 2: Find the solution z_f^k of the problem

$$\min_x\left\{\tilde{h}^k(x) + f(x) + \frac{\rho}{2}|x - \bar{x}|^2\right\}. \tag{2.5}$$

Set

$$g_f^k = -g_h^k - \rho(z_f^k - \bar{x}) \tag{2.6}$$

and define

$$\tilde{f}^{k+1}(\cdot) = f(z_f^k) + \langle g_f^k, \cdot - z_f^k\rangle. \tag{2.7}$$

Step 3: Increase k by 1 and go to Step 1.

Our objective is to prove that $z_h^k \to p(\bar{x})$.

Remark 2.2. The necessary and sufficient condition of optimality for (2.2) has the form

$$0 \in \partial h(z_h^k) + g_f^{k-1} + \rho(z_h^k - \bar{x}), \tag{2.8}$$

so the vector g_h^k (cf. (2.3)) is the element of $\partial h(z_h^k)$ which satisfies this condition. Hence $\tilde{h}^k \le h$ by the subgradient inequality. Similarly, the vector g_f^k (cf. (2.6)) is the element of $\partial f(z_f^k)$ which satisfies the optimality condition for (2.5): $0 \in g_h^k + \partial f(z_f^k) + \rho(z_f^k - \bar{x})$. Therefore, $\tilde{f}^{k+1} \le f$ and $\tilde{F}^k := h + \tilde{f}^k$ is a lower approximation of the objective $F = h + f$.

Let us denote by

$$\eta_k = h(z_h^k) + \tilde{f}^k(z_h^k) + \frac{\rho}{2}|z_h^k - \bar{x}|^2 \tag{2.9}$$

and

$$\eta_{k+1/2} = \tilde{h}^k(z_f^k) + f(z_f^k) + \frac{\rho}{2}|z_f^k - \bar{x}|^2$$

the optimal values of (2.2) and (2.5), respectively. The way in which the succesive linearizations $\tilde{f}^k$ and $\tilde{h}^k$ are generated ensures monotonicity of $\{\eta_k\}$:

$$\eta_k \leq \eta_{k+1/2} \leq \eta_{k+1}. \tag{2.10}$$

Indeed, the change from (2.2) to (2.5) at iteration k can be described in two steps:

(a) replace $h(\cdot)$ by $\tilde{h}^k(\cdot)$;

(b) replace $\tilde{f}^k(\cdot)$ by f.

By construction of $\tilde{h}^k$ (cf. (2.4)), operation (a) does not change the solution and value of (2.2), since $\tilde{h}^k(z_h^k) = h(z_h^k)$. Operation (b) can only increase the optimal value, because $f \geq \tilde{f}^k$, so $\eta_{k+1/2} \geq \eta_k$. Similarly, replacing f by $\tilde{f}^{k+1}$ does not change the solution and value of (2.5), because g_f^k was chosen to satisfy the optimality conditions and $\tilde{f}^{k+1}(z_f^k) = f(z_f^k)$. Replacing $\tilde{h}^k$ by h can only increase the optimal value, so $\eta_{k+1} \geq \eta_{k+1/2}$.

To estimate this increase from η_k to $\eta_{k+1/2}$ for operation (b), consider the family of relaxations of (2.5) at iteration k:

$$\min_x \left\{ Q_k(x,\mu) = \tilde{h}^k(x) + (1-\mu)(\alpha_p^k + \langle p^k, x\rangle) + \mu(\alpha_g^k + \langle g^k, x\rangle) + \frac{\rho}{2}|x - \bar{x}|^2 \right\}, \tag{2.11}$$

where $\mu \in [0,1]$, $p^k = g_f^{k-1}$, $\alpha_p^k = f(z_f^{k-1}) - \langle p^k, z_f^{k-1}\rangle$ and $\alpha_g^k = f(z_h^k) - \langle g^k, z_h^k\rangle$ for an arbitrary $g^k = g_f(z_h^k) \in \partial f(z_h^k)$. Since $\tilde{f}^k(\cdot) = \alpha_p^k + \langle p^k, \cdot\rangle$ and $\alpha_g^k + \langle g^k, \cdot\rangle$ are lower approximations of f, (2.11) is a relaxation of (2.5) for all $\mu \in [0,1]$. For $\mu = 0$ the solution and value of (2.11) coincide with those of (2.2). Thus, the increase in the optimal value of (2.5) can be estimated from below by the increase in the optimal value $\hat{Q}_k(\mu)$ of (2.11).

Lemma 2.3. *The following inequalities hold for any $g^k \in \partial f(z_h^k)$:*
(i) $\max_{\mu\in[0,1]} \hat{Q}_k(\mu) - \hat{Q}_k(0) \geq \hat{Q}_k(\bar{\mu}_k) - \hat{Q}_k(0) \geq \bar{\mu}_k\delta_k/2$,
(ii) $\eta_{k+1} \geq \eta_{k+1/2} \geq \eta_k + \bar{\mu}_k\delta_k/2$,
where $\delta_k = F(z_h^k) - \tilde{F}^k(z_h^k) \geq 0$ and $\bar{\mu}_k = \min\left\{1, \delta_k\rho/|g^k - p^k|^2\right\}$.

Proof. Note that $\delta_k \geq 0$, since $f \geq \tilde{f}^k$, so $\bar{\mu}_k \in [0,1]$. By direct calculation, the solution of (2.11) has the form $\hat{x}(\mu) = \bar{x} - \left[g_h^k + p^k + \mu(g^k - p^k)\right]/\rho$. Therefore the derivative of $\hat{Q}_k$ can be expressed as follows:

$$\begin{aligned}
\hat{Q}_k'(\mu) &= \langle g^k - p^k, \hat{x}(\mu)\rangle + \alpha_g^k - \alpha_p^k \\
&= \langle g^k - p^k, \hat{x}(\mu) - \hat{x}(0)\rangle + \left(\alpha_g^k + \langle g^k, \hat{x}(0)\rangle\right) - \left(\alpha_p^k + \langle p^k, \hat{x}(0)\rangle\right) \\
&= \langle g^k - p^k, \hat{x}(\mu) - \hat{x}(0)\rangle + F(z_h^k) - \tilde{F}^k(z_h^k) \\
&= -\frac{\mu|g^k - p^k|^2}{\rho} + \delta_k,
\end{aligned}$$

where we used the fact that $\hat{x}(0) = z_h^k$. Thus

$$\hat{Q}_k(\bar{\mu}_k) - \hat{Q}_k(0) = \int_0^{\bar{\mu}_k} \hat{Q}_k'(\mu)d\mu = \bar{\mu}_k\left(\delta_k - \frac{\bar{\mu}_k|g^k - p^k|^2}{2\rho}\right).$$

Using the definition of $\bar{\mu}_k$ yields (i). Assertion (ii) follows from (i) and (2.10). □

Theorem 2.4. *The sequences of points* $\{z_h^k\}$ *and approximations* $\{\tilde{F}^k\}$ *generated by Algorithm* 2.1 *have the following properties*:

(i) $|z_h^k - p(\bar{x})| \leq \left\{[F(z_h^k) - \tilde{F}^k(z_h^k)]/\rho\right\}^{1/2}$ *for* $k = 1, 2, \ldots$.

(ii) $\lim_{k\to\infty} \left[F(z_h^k) - \tilde{F}^k(z_h^k)\right] = 0$.

(iii) $\lim_{k\to\infty} z_h^k = p(\bar{x})$.

Proof. Since $F \geq \tilde{F}^k$ and z_h^k solves the strongly convex problem (2.2), we have [Roc76b]

$$\begin{aligned} F(p(\bar{x})) + \frac{\rho}{2}|p(\bar{x}) - \bar{x}|^2 &\geq \tilde{F}^k(p(\bar{x})) + \frac{\rho}{2}|p(\bar{x}) - \bar{x}|^2 \\ &\geq \tilde{F}^k(z_h^k) + \frac{\rho}{2}|z_h^k - \bar{x}|^2 + \frac{\rho}{2}|p(\bar{x}) - z_h^k|^2. \end{aligned} \tag{2.12}$$

Similarly, $p(\bar{x})$ solves the strongly convex problem in (2.1), so

$$F(z_h^k) + \frac{\rho}{2}|z_h^k - \bar{x}|^2 \geq F(p(\bar{x})) + \frac{\rho}{2}|p(\bar{x}) - \bar{x}|^2 + \frac{\rho}{2}|p(\bar{x}) - z_h^k|^2.$$

Adding the last two inequalities and simplifying, we get $F(z_h^k) - \tilde{F}^k(z_h^k) \geq \rho|p(\bar{x}) - z_h^k|^2$, which proves assertion (i). Next, (2.12) can be equivalently written as (cf. (2.9))

$$\frac{\rho}{2}|p(\bar{x}) - z_h^k|^2 \leq F(p(\bar{x})) + \frac{\rho}{2}|p(\bar{x}) - \bar{x}|^2 - \eta_k. \tag{2.13}$$

By Lemma 2.3, $\{\eta_k\}$ is nondecreasing, so (2.13) implies that $\{z_h^k\}$ is bounded. Then $\{g^k\}$ is bounded as well, because $g^k \in \partial f(z_h^k)$ for all k and f is finite-valued (cf. [Roc70, Thm 24.7]). By an analogous argument, using the inequality

$$\frac{\rho}{2}|p(\bar{x}) - z_f^k|^2 \leq F(p(\bar{x})) + \frac{\rho}{2}|p(\bar{x}) - \bar{x}|^2 - \eta_{k+1/2},$$

we see that z_f^k and $p^k = g_f^k \in \partial f(z_f^k)$ are bounded. By (2.13), the sequence $\{\eta_k\}$ is bounded from above, so Lemma 2.3 implies that it converges and $\bar{\mu}_k\delta_k \to 0$. Since $\{|g^k - p^k|\}$ is bounded, assertion (ii) follows from the definition of $\bar{\mu}_k$ (cf. Lemma 2.3). The final assertion is a consequence of (i) and (ii). □

Remark 2.5. Algorithm 2.1 can be used in the implementable proximal point schemes of [Aus86, CoL93, EcB92, GoT89, Gül91, Lem89, Roc76b]. Indeed, Theorem 2.4 ensures that for every $\epsilon > 0$ we can find in finitely many steps a point z_h^k such that $|z_h^k - p(\bar{x})| \leq \epsilon$. An alternative scheme will be presented in the next section.

3 The alternating linearization method

The algoritm below employs a simple descent test for stopping the loop of Algorithm 2.1 in order to update the prox center.

Algorithm 3.1.

Step 0: Select $x^1 \in \operatorname{dom} h$, $z_f^0 \in \mathbb{R}^n$ and $g_f^0 \in \partial f(z_f^0)$. Define $\tilde{f}^1(\cdot) = f(z_f^0) + \langle g_f^0, \cdot - z_f^0\rangle$. Choose parameters $\rho_1 \geq \rho_{\min} > 0$, $\kappa > 1$, $\beta_0 > 0$, $\beta_1 \in (0, 1)$. Set $k = 1$.

Step 1: Find the solution z_h^k of the problem

$$\min_x \left\{ h(x) + \tilde{f}^k(x) + \frac{\rho_k}{2}|x - x^k|^2 \right\}. \tag{3.1}$$

Set $g_h^k = -g_f^{k-1} - \rho_k(z_h^k - x^k)$ and define $\tilde{h}^k(\cdot) = h(z_h^k) + \langle g_h^k, \cdot - z_h^k \rangle$.

Step 2: Let $\tilde{F}^k = h + \tilde{f}^k$. Set

$$v_k = \tilde{F}^k(z_h^k) - F(x^k). \tag{3.2}$$

If

$$F(z_h^k) \le F(x^k) + \beta_1 v_k, \tag{3.3}$$

then set $x^{k+1} = z_h^k$ (descent step); otherwise set $x^{k+1} = x^k$ (null step).

Step 3: If $x^{k+1} = z_h^k$, then choose $\rho_{k+1} \in [\max\{\rho_{\min}, \rho_k/\kappa\}, \rho_k]$. If $x^{k+1} = x^k$ and

$$\delta_k := F(z_h^k) - \tilde{F}(z_h^k) \ge \beta_0 \frac{|v_k|}{|z_h^k - x^k|},$$

then choose $\rho_{k+1} \ge \rho_k$, else set $\rho_{k+1} = \rho_k$.

Step 4: Find the solution z_f^k of the problem

$$\min_x \left\{ \tilde{h}^k(x) + f(x) + \frac{\rho_{k+1}}{2}|x - x^{k+1}|^2 \right\}. \tag{3.4}$$

Set $g_f^k = -g_h^k - \rho_{k+1}(z_f^k - x^{k+1})$ and define $\tilde{f}^{k+1}(\cdot) = f(z_f^k) + \langle g_f^k, \cdot - z_f^k \rangle$.

Step 5: Increase k by 1 and go to Step 1.

We shall preserve the notation of the previous section, with only necessary changes. So

$$\eta_k = \tilde{F}^k(z_h^k) + \frac{\rho_k}{2}|z_h^k - x^k|^2 \tag{3.5}$$

will denote the optimal value of (3.1), and $\eta_{k+1/2}$ that of (3.4).

By construction (cf. Remark 2.2), $g_f^k \in \partial f(z_f^k)$ and $\tilde{F}^k \le F$, so $\eta_k \le F(x^k)$ and $v_k \le 0$. Thus (3.3) implies that $\{F(x^k)\}$ is nonincreasing and $\{x^k\} \subset \operatorname{dom} F$. It will become clear that if $v_k = 0$ or $\eta_k = F(x^k)$ then $x^k \in \operatorname{Arg\,min} F$.

4 Convergence

Let us first make a simple observation concerning the optimal values of (3.1) and (3.4).

Lemma 4.1. *The following inequalities are true for all* $k = 1, 2, \ldots$:

(i) $\frac{\rho_k}{2}|z_h^k - x^k|^2 \le \frac{|v_k|}{2} \le F(x^k) - \eta_k \le |v_k|$,

(ii) $\frac{\rho_{k+1}}{2}|z_f^k - x^{k+1}|^2 \le F(x^{k+1}) - \eta_{k+1/2}$.

Proof. (3.2) and (3.5) yield $F(x^k)+v_k \le \eta_k$, and hence the right inequality of (i). Next, note that by construction (cf. Step 1)

$$-\rho_k(z_h^k - x^k) = g_h^k + g_f^{k-1} \in \partial \tilde{F}^k(z_h^k), \tag{4.1}$$

so the left inequality in (i) follows from the subgradient inequality, since

$$-v_k = F(x^k) - \tilde{F}^k(z_h^k) \ge \tilde{F}^k(x^k) - \tilde{F}^k(z_h^k) \ge \rho_k |z_h^k - x^k|^2.$$

Thus

$$\eta_k = F(x^k) + v_k + \frac{\rho_k}{2}|z_h^k - x^k|^2 \le F(x^k) + \frac{v_k}{2},$$

which completes the proof of (i). Assertion (ii) can be obtained similarly. □

The following result is a simple consequence of Lemma 4.1 and Theorem 2.4.

Corollary 4.2. *If $v_k = 0$ then $x^k \in \operatorname{Arg\,min} F$.*

Proof. By Lemma 4.1(i) and (3.2), $z_h^k = x^k$ and $\tilde{F}^k(z_h^k) = F(z_h^k) = F(x^k)$. Then Theorem 2.4(i) yields $x^k = z_h^k = \arg\min F + \rho_k|\cdot - x^k|^2/2$, so $x^k \in \operatorname{Arg\,min} F$ [Roc76b]. □

We split our convergence analysis into several stages, starting from the case of an infinite series of null steps. Our objective is to prove that in this case the optimal values of (3.1) and (3.4) converge to $F(x^{k_0})$, where x^{k_0} is the last point to which a descent step was made.

Lemma 4.3. *If a null step is made at iteration k then*

$$\eta_{k+1} \ge \eta_k + \beta_1 \bar{\mu}_k |v_k|/2,$$

where $\bar{\mu}_k = \min\{1, \beta_1|v_k|/|g_f(z_h^k) - g_f^{k-1}|^2\}$ for any $g_f(z_h^k) \in \partial f(z_h^k)$.

Proof. If (3.3) fails, then $\delta_k = F(z_h^k) - \tilde{F}(z_h^k) \ge \beta_1|v_k|$. Hence if $\rho_{k+1} = \rho_k$ then Lemma 2.3(ii) yields $\eta_{k+1/2} \ge \eta_k + \beta_1\bar{\mu}_k|v_k|/2$. When $\rho_{k+1} > \rho_k$, the minimum value of (3.4) can only be greater. Next, $\eta_{k+1} \ge \eta_{k+1/2}$, by the same argument as in Lemma 2.3. □

Lemma 4.4. *If the set $\mathcal{K} = \{k : x^{k+1} \ne x^k\}$ is finite, then $v_k \to 0$.*

Proof. By assumption, there is k_0 such that $x^k = x^{k_0}$ for all $k \ge k_0$. By Lemma 4.3, $\{\eta_k\}$ is nondecreasing for $k \ge k_0$, hence convergent, because $\eta_k \le F(x^{k_0})$, so $\eta_{k+1} - \eta_k \to 0$ and $\bar{\mu}_k|v_k| \to 0$. Since $\rho_k \ge \rho_{\min} > 0$ for all k, and $\{x^k\}$ is bounded, so are $\{z_h^k\}$ and $\{z_f^k\}$ (cf. Lemma 4.1), and hence also $g_f(z_h^k) \in \partial f(z_h^k)$ and $g_f^k \in \partial f(z_f^k)$, because f is locally Lipschitz (cf. [Roc70, Thm 24.7]). Thus, using the definition of $\bar{\mu}_k$, we get $v_k \to 0$. □

Let us now pass to the case of infinitely many descent steps.

Lemma 4.5. *Suppose the set $\mathcal{K} = \{k : x^{k+1} \ne x^k\}$ is infinite and $\inf F > -\infty$. Then:*
(i) $\sum_{k\in\mathcal{K}} |v_k| < \infty$;

(ii) $\lim_{k\to\infty} v_k = 0$;

(iii) $\lim_{k\to\infty} \left[F(x^k) - \eta_k\right] = 0$;

(iv) $\lim_{k\to\infty} \left[F(x^{k+1}) - \eta_{k+1/2}\right] = 0$.

Proof. For each $k \in \mathcal{K}$, a descent step occurs with $F(x^k) - F(x^{k+1}) \geq -\beta_1 v_k \geq 0$. Summing these inequalities over k and using monotonicity and boundedness of $\{F(x^k)\}$, we get (i) and $v_k \to 0$ for $k \in \mathcal{K}$. In view of Lemma 4.1, $F(x^k) - \eta_k \to 0$ for $k \in \mathcal{K}$. To show convergence of the whole sequences, let us denote by $l(k)$ the number of the last iteration with a descent step preceding iteration k. By Lemma 4.3,

$$0 \leq F(x^k) - \eta_k \leq F(x^{l(k)+1}) - \eta_{l(k)+1}. \tag{4.2}$$

¿From (i) and Lemma 4.1 we obtain $F(x^{l(k)}) - \eta_{l(k)} \to 0$. It remains to relate $F(x^{l(k)+1}) - \eta_{l(k)+1}$ to $F(x^{l(k)}) - \eta_{l(k)}$. The changes in (3.1) at a descent step at iteration $l = l(k)$ can be decomposed into the following operations:

(a) the shift of the regularizing point x^l to $x^{l+1} = z_h^l$;

(b) the change of the penalty parameter ρ_l to $\rho_{l+1} \in [\rho_l/\kappa, \rho_l]$;

(c) replacement of $\tilde{f}^l$ by $\tilde{f}^{l+1}$.

Denote by $\eta_l^{(b)}$ the resulting optimal value of (3.1) after partial modifications (a) and (b). By construction, $g_{\tilde{F}}^l = g_h^l + g_f^{l-1} \in \partial \tilde{F}^l(x^{l+1})$ is such that $x^{l+1} - x^l = -g_{\tilde{F}}^l/\rho_l$ (cf. (4.1)) and

$$\eta_l = \min_x \left\{ \tilde{F}^l(x^{l+1}) + \langle g_{\tilde{F}}^l, x - x^{l+1} \rangle + \frac{\rho_l}{2}|x - x^l|^2 \right\},$$

so

$$\eta_l = \tilde{F}^l(x^{l+1}) + \frac{1}{2\rho_l}|g_{\tilde{F}}^l|^2 = \tilde{F}^l(x^l) - \frac{1}{2\rho_l}|g_{\tilde{F}}^l|^2.$$

In a similar way,

$$\eta_l^{(b)} = \min_x \left\{ \tilde{F}^l(x^{l+1}) + \langle g_{\tilde{F}}^l, x - x^{l+1} \rangle + \frac{\rho_{l+1}}{2}|x - x^{l+1}|^2 \right\} = \tilde{F}^l(x^{l+1}) - \frac{1}{2\rho_{l+1}}|g_{\tilde{F}}^l|^2.$$

Therefore,

$$\tilde{F}^l(x^l) - \eta_l = \frac{1}{2\rho_l}|g_{\tilde{F}}^l|^2 = \frac{\rho_{l+1}}{\rho_l}\left[\tilde{F}^l(x^{l+1}) - \eta_l^{(b)}\right] \geq \frac{1}{\kappa}\left[\tilde{F}^l(x^{l+1}) - \eta_l^{(b)}\right].$$

Finally, operation (c) is a hypothetical null step, so by Lemma 2.3

$$\eta_{l+1} \geq \eta_{l+1/2} \geq \eta_l^{(b)}.$$

Combining the last two relations and noting that at descent steps $F(x^{l+1}) \leq F(x^l) = \tilde{F}^l(x^{l+1}) + |v_l|$, we obtain for each descent step $l(k)$ the relation

$$F(x^{l(k)+1}) - \eta_{l(k)+1} \leq \kappa\left[F(x^{l(k)}) - \eta_{l(k)}\right] + |v_{l(k)}|.$$

Since the right side of the above inequality converges to 0, and the left side is nonnegative, we must have $\lim_{k\to\infty} F(x^{l(k)+1}) - \eta_{l(k)+1} = 0$. Using this relation in (4.2) we conclude that $F(x^k) - \eta_k \to 0$ and $F(x^{k+1}) - \eta_{k+1/2} \to 0$, i.e., (iii) and (iv) hold. Assertion (ii) follows from Lemma 4.1. □

Lemma 4.6. *Suppose the set $\mathcal{K} = \{k : x^{k+1} \neq x^k\}$ is infinite. If there exists a point $\tilde{x}$ such that $F(x^k) \geq F(\tilde{x})$ for all k, then $\{x^k\}$ converges to a point $\bar{x} \in \operatorname{dom} F$.*

Proof. Fix $k \in \mathcal{K}$. We have

$$|x^{k+1} - \tilde{x}|^2 = |x^k - \tilde{x}|^2 + 2\langle x^{k+1} - \tilde{x}, x^{k+1} - x^k\rangle - |x^{k+1} - x^k|^2. \tag{4.3}$$

By (4.1), $g^k_{\tilde{F}} = g^k_h + g^{k-1}_f = -\rho_k(x^{k+1} - x^k) \in \partial F^k(x^{k+1})$, so

$$\begin{aligned}\rho_k\langle x^{k+1} - \tilde{x}, x^{k+1} - x^k\rangle &= \langle \tilde{x} - x^{k+1}, g^k_{\tilde{F}}\rangle \\ &\leq \tilde{F}^k(\tilde{x}) - \tilde{F}^k(x^{k+1}) \leq F(\tilde{x}) - F(x^k) - v_k.\end{aligned}$$

Using this inequality in (4.3) yields

$$|x^{k+1} - \tilde{x}|^2 \leq |x^k - \tilde{x}|^2 + 2|v_k|/\rho_k, \quad k \in \mathcal{K}.$$

Since $\{\rho_k\}$ is bounded away from 0 by construction, the last inequality and assertion (i) of Lemma 4.5 imply that the sequence $\{x^k\}$ is bounded. Hence, it has an accumulation point $\bar{x}$. By monotonicity of $\{F(x^k)\}$ and closedness of F, $F(\bar{x}) \leq F(x^k)$ for all k, so we can replace $\tilde{x}$ by $\bar{x}$ in the preceding argument, concluding that $\bar{x}$ is the only accumulation point, since $\sum_{k\in\mathcal{K}, k\geq l} |v_k| \to 0$ as $l \to \infty$. □

Lemma 4.7. *If there exists a point $\tilde{x}$ such that $F(x^k) \geq F(\tilde{x})$ for all k, then:*
(i) *$v_k \to 0$, $F(x^k) - \eta_k \to 0$ and $F(x^{k+1}) - \eta_{k+1/2} \to 0$, as $k \to \infty$;*
(ii) *The sequence $\{x^k\}$ converges to a point $\bar{x} \in \operatorname{Arg\,min} F$.*

Proof. By Lemmas 4.4–4.6, $\{x^k\}$ converges to some $\bar{x} \in \operatorname{dom} F$ and assertion (i) holds. Let us consider two cases.

Case 1: There exists $\bar{\rho}$ such that $\rho_k \leq \bar{\rho}$ for all k. Since $\tilde{F}^k \leq F$,

$$\begin{aligned}F(x^k) - \eta_k &= F(x^k) - \min_x \left\{\tilde{F}^k(x) + \frac{\rho_k}{2}|x - x^k|^2\right\} \\ &\geq F(x^k) - \min_x \left\{F(x) + \frac{\rho_k}{2}|x - x^k|^2\right\} \\ &\geq F(x^k) - \min_x \left\{F(x) + \frac{\bar{\rho}}{2}|x - x^k|^2\right\} \geq 0.\end{aligned}$$

With $F(x^k) - \eta_k \to 0$ and $x^k \to \bar{x}$, passing to the limit and using the closedness of F one obtains (cf. [HUL93, Thm XV.4.1.4]) $F(\bar{x}) = \min_x \left\{F(x) + \frac{\bar{\rho}}{2}|x - \bar{x}|^2\right\}$, which is equivalent to $\bar{x} \in \operatorname{Arg\,min} F$ (see, e.g., [HUL93, Thm XV.4.1.7]).

Case 2: $\limsup \rho_k = +\infty$. Since $v_k \to 0$, Lemma 4.1(i) yields

$$\frac{\rho_k}{2}|z_h^k - x^k|^2 \le |v_k| \to 0. \tag{4.4}$$

With $\rho_k \ge \rho_{\min}$ one must have $z_h^k - x^k \to 0$. In a similar way, $z_f^k - x^k \to 0$. Since f is continuous over the domain of h,

$$F(z_h^k) - \tilde{F}(z_h^k) = F(z_h^k) - F(x^k) - v_k = f(z_h^k) - f(x^k) - v_k \to 0. \tag{4.5}$$

The penalty coefficient is increased infinitely many times, so (cf. Step 3) there must be a subsequence $\mathcal{K}$ such that for $k \in \mathcal{K}$

$$F(z_h^k) - F(x^k) - v_k \ge \beta_0 |v_k|/|z_h^k - x^k|. \tag{4.6}$$

Dividing (4.4) by $|z_h^k - x^k|$ and using (4.6) and (4.5), we get $\rho_k |z_h^k - x^k| \to 0$. Therefore, using the definition of g_h^k at Step 1,

$$g_h^k + g_f^{k-1} \to 0, \ k \in \mathcal{K}. \tag{4.7}$$

Since f is locally Lipschitz and $\{z_f^k\}$ is bounded, the vectors $g_f^k \in \partial f(z_f^k)$ are uniformly bounded. By the upper semicontinuity of ∂f (cf. [Roc70, Thm 24.4]), we can restrict $\mathcal{K}$ so that $g_f^{k-1} \to g_f(\bar{x}) \in \partial f(\bar{x})$, $k \in \mathcal{K}$. Then $g_h^k \to -g_f(\bar{x})$, $k \in \mathcal{K}$. Consequently, $-g_f(\bar{x}) \in \partial h(\bar{x})$, because $z_h^k \to \bar{x}$ and $g_h^k \in \partial h(z_h^k)$. This proves that $0 \in \partial F(\bar{x})$. □

Our results can be summarized as follows.

Theorem 4.8. *Algorithm* 3.1 *generates a sequence* $\{x^k\}$ *with the following properties:*
(i) $F(x^k) \downarrow \inf F$.
(ii) *If* $\operatorname{Arg\,min} F \ne \emptyset$ *then* $\{x^k\}$ *converges to a point* $\hat{x} \in \operatorname{Arg\,min} F$.
(iii) *If* $\operatorname{Arg\,min} F = \emptyset$ *then* $|x^k| \to \infty$.
(iv) *If* $\operatorname{Arg\,min} F \ne \emptyset$ *and the sequence* $\{\rho_k\}$ *is bounded, then the sequences* $\{g_f^k\}$ *and* $\{g_h^k\}$ *are bounded,* $g_h^k + g_f^{k-1} \to 0$, $g_h^k + g_f^k \to 0$, *and every accumulation point* $(\hat{g}_f, \hat{g}_h)$ *of* $\{(g_f^k, g_h^k)\}$ *satisfies the relations:* $\hat{g}_f \in \partial f(\hat{x})$, $\hat{g}_h \in \partial h(\hat{x})$ *and* $\hat{g}_f + \hat{g}_h = 0$.

Proof. If $\operatorname{Arg\,min} F$ contains a point $\tilde{x}$, one has $F(x^k) \ge F(\tilde{x})$ for all k. Then by Lemma 4.7, $x^k \to \hat{x} \in \operatorname{Arg\,min} F$, and $F(x^k) \downarrow F(\hat{x}) = \inf F$, which proves (i)–(ii) in this case.

Suppose now that $\operatorname{Arg\,min} F = \emptyset$. If there existed $\tilde{x}$ such that $F(x^k) \ge F(\tilde{x})$ for all k, then Lemma 4.7 would imply convergence of $\{x^k\}$ to a minimizer of F, a contradiction. Therefore for every $\tilde{x}$ we can find k such that $F(x^k) < F(\tilde{x})$. This implies that $F(x^k) \downarrow \inf F$ in this case, too, i.e., (i) is true. Moreover, if $\{x^k\}$ had a bounded subsequence, then (by the closedness of F) each of its accumulation points would minimize F, another contradiction. Therefore (iii) must be true.

Let us now consider in more detail the case when $\operatorname{Arg\,min} F \ne \emptyset$ and the sequence $\{\rho_k\}$ is bounded. We already know that $x^k \to \hat{x} \in \operatorname{Arg\,min} F$. By Lemma 4.7, $F(x^k) - \eta_k \to 0$ and $F(x^{k+1}) - \eta_{k+1/2} \to 0$. Then Lemma 4.1 implies that $z_h^k \to \hat{x}$ and $z_f^k \to \hat{x}$. Since $g_f^k \in \partial f(z_f^k)$ and f is locally Lipschitz, the sequence $\{g_f^k\}$ is bounded and each its accumulation point is in $\partial f(\hat{x})$. Next, by the definitions of g_f^k and g_h^k, $g_f^k + g_h^k = \rho_{k+1}(z_f^k - x^{k+1}) \to 0$ and $g_f^{k-1} + g_h^k = \rho_k(z_h^k - x^k) \to 0$. Thus $\{g_h^k\}$ must be bounded, too, and the required result follows. □

Remark 4.9. Without boundedness of $\{\rho_k\}$ we obtain (iv) only on some subsequence, as follows from (4.7).

5 Dual application

Let us now discuss in more detail the application of the alternating linearization method to structured problems of the form:

$$\inf \{ \varphi(y) + \psi(My) \} \tag{5.1}$$

with closed proper convex functions $\varphi : \mathbb{R}^m \to (-\infty, +\infty]$, $\psi : \mathbb{R}^n \to (-\infty, +\infty]$, and an $n \times m$ matrix M. Splitting variables yields the problem

$$\inf \{ \varphi(y) + \psi(w) \}, \tag{5.2a}$$

$$w - My = 0, \tag{5.2b}$$

with the Lagrangian $L(y, w, x) = \varphi(y) + \psi(w) + \langle x, My - w \rangle$, where $x \in \mathbb{R}^n$ are dual variables. The dual problem

$$\sup_x \Big\{ L_D(x) = \inf_{y,w} L(y, w, x) \Big\}$$

can be equivalently written as

$$\inf_x \Big\{ F(x) = \psi^*(x) + \varphi^*(-M^T x) \Big\}, \tag{5.3}$$

using the conjugates $\varphi^*(\cdot) = \sup_y \{\langle \cdot, y \rangle - \varphi(y)\}$, $\psi^*(\cdot) = \sup_w \{\langle \cdot, w \rangle - \psi(w)\}$. The dual problem (5.3) has the form (1.1), with

$$h(x) = \psi^*(x)$$

and

$$f(x) = \varphi^*(-M^T x).$$

Let us assume that $\varphi^* \circ (M^T)$ is finite-valued. Then both f and h are closed proper convex functions [Roc70, Thm 12.2] and $\operatorname{dom} f = \mathbb{R}^n$. Therefore problem (5.3) satisfies all the assumptions required for applying the alternating linearization method.

The algorithm below will be shown to constitute a dual version of Algorithm 3.1.

Algorithm 5.1.
Step 0: Select $x^1 \in \operatorname{dom} h$ and calculate $F(x^1) = h(x^1) + f(x^1)$. Choose $z_f^0 \in \mathbb{R}^n$. Calculate

$$f(z_f^0) = -\min_y \Big\{ \varphi(y) + \langle z_f^0, My \rangle \Big\}. \tag{5.4}$$

Choose a minimizer y^0 in the problem above. Select $\rho_1 \geq \rho_{\min} > 0$, $\kappa > 1$, $\beta_0 > 0$, $\beta_1 \in (0, 1)$. Set $k = 1$.

Step 1: Calculate

$$w^k = \arg\min_w \left\{ \psi(w) - \langle x^k, w \rangle + \frac{1}{2\rho_k} |w - My^{k-1}|^2 \right\}, \tag{5.5}$$

and set

$$z_h^k = x^k - (w^k - My^{k-1})/\rho_k. \tag{5.6}$$

Step 2: Calculate

$$h(z_h^k) = \langle w^k, z_h^k \rangle - \psi(w^k), \tag{5.7}$$

$$f(z_h^k) = -\min_y \left\{ \varphi(y) + \langle z_h^k, My \rangle \right\}, \tag{5.8}$$

$$\tilde{f}^k(z_h^k) = -\left\{ \varphi(y^{k-1}) + \langle z_h^k, My^{k-1} \rangle \right\}. \tag{5.9}$$

Set $F(z_h^k) = h(z_h^k) + f(z_h^k)$ and $\tilde{F}^k(z_h^k) = h(z_h^k) + \tilde{f}^k(z_h^k)$. Set $v_k = \tilde{F}^k(z_h^k) - F(x^k)$. If $F(z_h^k) \le F(x^k) + \beta_1 v_k$, then set $x^{k+1} = z_h^k$; otherwise set $x^{k+1} = x^k$.

Step 3: Choose ρ_{k+1} as at Step 3 of Algorithm 3.1.

Step 4: Calculate

$$y^k \in \operatorname{Arg\,min}_y \left\{ \varphi(y) + \langle x^{k+1}, My \rangle + \frac{1}{2\rho_{k+1}} |w^k - My|^2 \right\}. \tag{5.10}$$

Step 5: Increase k by 1 and go to Step 1.

The analysis of Algorithm 5.1 will be based on the following fact [Roc70, Thm 23.5].

Fact 5.2. *For a proper convex closed function f the following conditions are equivalent:* $x^* \in \partial f(x)$, $x \in \partial f^*(x^*)$, $f(x) + f^*(x^*) = \langle x, x^* \rangle$, $x \in \operatorname{Arg\,min}\{f(\cdot) - \langle x^*, \cdot \rangle\}$.

Theorem 5.3. *Algorithm* 5.1 *generates sequences* $\{x^k\}$, $\{y^k\}$ *and* $\{w^k\}$ *with the following properties*:

(i) $F(x^k) \downarrow \inf F$.

(ii) *If* $\operatorname{Arg\,min} F \neq \emptyset$ *then* $\{x^k\}$ *converges to a point* $\hat{x} \in \operatorname{Arg\,min} F$.

(iii) *If* $\operatorname{Arg\,min} F = \emptyset$ *then* $|x^k| \to \infty$.

(iv) *If* $\operatorname{Arg\,min} F \neq \emptyset$ *and the sequence* $\{\rho_k\}$ *is bounded, then the sequences* $\{My^k\}$ *and* $\{w^k\}$ *are bounded,* $w^k - My^k \to 0$ *and* $w^k - My^{k-1} \to 0$. *Further, each accumulation point* $\hat{y}$ *of* $\{y^k\}$ *is a solution of* (5.1).

Proof. We shall prove that Algorithm 5.1 is equivalent to Algorithm 3.1 applied to the dual problem (5.3).

First, let us note that the minimizer y^0 in (5.4) chosen at Step 0 (which exists because $\varphi^* \circ (M^T)$ is finite-valued) satisfies the relation $y^0 \in \partial\varphi^*(-M^T z_f^0)$. Therefore, by Fact 5.2, $-My^0 \in \partial f(z_f^0)$ and we can define $g_f^0 = -My^0$.

We shall use induction. Assume that for some k we have

$$y^{k-1} \in \partial\varphi^*(-M^T z_f^{k-1}) \tag{5.11}$$

and

$$g_f^{k-1} = -My^{k-1}. \tag{5.12}$$

By (5.12), problem (3.1) can be formulated as follows:

$$\min_x \left\{ \psi^*(x) - \langle My^{k-1}, x \rangle + \frac{\rho_k}{2}|x - x^k|^2 \right\}. \tag{5.13}$$

We now show that (5.5)–(5.6) define its solution z_h^k. Indeed, the optimality condition for (5.5) yields:

$$z_h^k = x^k - (w^k - My^{k-1})/\rho_k \in \partial\psi(w^k), \tag{5.14}$$

which by Fact 5.2 is equivalent to

$$w^k \in \partial\psi^*(z_h^k). \tag{5.15}$$

Using (5.6) we can rewrite the last relation as $My^{k-1} - \rho_k(z_h^k - x^k) \in \partial\psi^*(z_h^k)$, which is necessary and sufficient for the optimality of z_h^k in (5.13). From (5.15), using Fact 5.2, we obtain $\psi^*(z_h^k) = \langle w^k, z_h^k \rangle - \psi(w^k)$, which validates (5.7). Relation (5.8) follows directly from the definition. Next, (5.11) and Fact 5.2 yield

$$f(z_f^{k-1}) = \varphi^*(-M^T z_f^{k-1}) = -\varphi(y^{k-1}) - \langle M^T z_f^{k-1}, y^{k-1} \rangle.$$

Combining this relation with (5.12) we obtain

$$\begin{aligned} \tilde{f}^k(z_h^k) &= f(z_f^{k-1}) + \left\langle g_f^{k-1}, z_h^k - z_f^{k-1} \right\rangle \\ &= -\varphi(y^{k-1}) - \langle M^T z_f^{k-1}, y^{k-1} \rangle - \left\langle My^{k-1}, z_h^k - z_f^{k-1} \right\rangle, \end{aligned}$$

which is equivalent to (5.9). The remaining part of Step 2 and Step 3 are identical to those in Algorithm 3.1.

By direct calculation, using (5.12) and (5.6), we obtain

$$g_h^k = -g_f^{k-1} - \rho_k(z_h^k - x^k) = w^k. \tag{5.16}$$

Therefore, problem (3.4) can be written as

$$\min_x \left\{ \langle w^k, x \rangle + \varphi^*(-M^T x) + \frac{\rho_{k+1}}{2}|x - x^{k+1}|^2 \right\}. \tag{5.17}$$

We now show that the point z_f^k, the solution of (5.17), has the form

$$z_f^k = x^{k+1} - (w^k - My^k)/\rho_{k+1}, \tag{5.18}$$

where y^k is given by (5.10). Indeed, the optimality condition for (5.10) reads

$$-M^T z_f^k = -M^T x^{k+1} + M^T(w^k - My^k)/\rho_{k+1} \in \partial\varphi(y^k), \tag{5.19}$$

which by Fact 5.2 is equivalent to the relation $y^k \in \partial\varphi^*(-M^T z_f^k)$, i.e., (5.11) holds for k. The last relation is equivalent to $-My^k \in \partial f(z_f^k)$ (Fact 5.2). Substitution of My^k from

(5.18) yields the optimality condition for (5.17): $-w^k - \rho_{k+1}(z_f^k - x^{k+1}) \in \partial f(z_f^k)$. Finally, from (5.16) and (5.18) we get

$$g_f^k = -g_h^k - \rho_{k+1}(z_f^k - x^{k+1}) = -My^k, \tag{5.20}$$

which proves (5.12) for k and completes the induction.

Therefore, assertions (i)–(iii) follow from those of Theorem 4.8. To show (iv), observe that from (5.16) and (5.20), by Theorem 4.8(iv), the sequences $\{My^k\}$ and $\{w^k\}$ are bounded,

$$w^k - My^k \to 0 \tag{5.21}$$

and $w^k - My^{k-1} \to 0$. To complete the proof of (iv), let $(w^k, y^k) \to (\hat{w}, \hat{y})$, $k \in \mathcal{K}$. Taking limits in (5.14) and (5.19), we obtain $\hat{x} \in \partial\psi(\hat{w})$, $-M^T\hat{x} \in \partial\varphi(\hat{y})$ and, by (5.21), $\hat{w} - M\hat{y} = 0$. This proves the optimality of $(\hat{w}, \hat{y})$ in (5.2). □

As mentioned in §1–2, the alternating linearization method fits in the framework of inexact proximal point algorithms and bears some resemblance to the operator splitting methods. Therefore it is not surprising that its dual version, Algorithm 5.1, is intimately related to augmented Lagrangian methods and alternating direction methods of multipliers [BeT89, DLMK+94, EcB92, EcF94b, Fuk92, Gab83].

Specifically, consider the augmented Lagrangian for (5.2):

$$\Lambda_\rho(y, w, x) = \varphi(y) + \psi(w) - \langle x, w - My\rangle + \frac{\rho}{2}|w - My|^2, \tag{5.22}$$

where $x \in \mathbb{R}^n$ is the vector of multipliers and $\rho > 0$ is a penalty coefficient. Assuming that in Algorithm 5.1 the points x^k remain fixed at x and the penalty coefficients ρ_k fixed at ρ, we see that (5.5) and (5.10) implement the Gauss-Seidel method for minimizing the augmented Lagrangian (5.22). Note, however, that in the alternating direction method the multipliers are updated after each Gauss-Seidel iteration. In Algorithm 5.1, the classical update (cf. (5.6))

$$x^{k+1} = x^k - (w^k - My^{k-1})/\rho_k$$

takes place only under the descent conditions of Step 2. Moreover, the penalty coefficient is allowed to change within the "Gauss-Seidel" loop as well as after the multiplier update.

Example 5.4. Let us consider the problem

$$\min\left\{\varphi(y) + \sum_{j=1}^N \psi_j(y)\right\},$$

with closed proper convex functions $\varphi : \mathbb{R}^m \to (-\infty, +\infty)$ and $\psi_j : \mathbb{R}^m \to (-\infty, +\infty]$, $j = 1, \ldots, N$. This is a special case of (5.1) with $My = (y, y, \ldots, y)$, $\psi(w) = \sum_{j=1}^N \psi_j(w_j)$ and $n = Nm$. The key operations of Algorithm 5.1 can be substantially simplified in this case. With $x = (x_1, \ldots, x_N) \in \mathbb{R}^{Nm}$ problem (5.5) solved at Step 1 decomposes into parallel subproblems for $j = 1, \ldots, N$:

$$w_j^k = \arg\min_{w_j}\left\{\psi_j(w_j) - \langle x_j^k, w_j\rangle + \frac{1}{2\rho_k}|w_j - y^{k-1}|^2\right\},$$

$$(z_h^k)_j = x_j^k - (w_j^k - y^{k-1})/\rho_k,$$

while (5.10) takes the form:

$$y^k = \arg\min_y \left\{ \varphi(y) + \langle \textstyle\sum_{j=1}^N x_j^{k+1}, y\rangle + \dfrac{1}{2\rho_{k+1}} \displaystyle\sum_{j=1}^N |w_j^k - y|^2 \right\}.$$

We easily recognize some similarities with the algorithms of [HaL88, MNS91, Tse91], but our approach has different rules for updating the multipliers and a variable penalty coefficient.

6 Applications to stochastic programming

We now consider an important class of optimization models known as multistage stochastic programming problems.

We use the modeling methodology developed in [RoW91] (see also [ChR94, MuR95, Rob91]). The basic object in the model is the *scenario tree*, whose levels $1, \ldots, T$ (counted from the root to the leaves) correspond to time stages and each path from the root to the leaves (scenarios) has exactly T nodes. With each scenario path j ($j = 1, \ldots, N$) the following objects are associated: the decision subvector

$$w_j = (w_j(1), \ldots, w_j(T)) \in \mathbb{R}^{q_1} \times \cdots \times \mathbb{R}^{q_T},$$

the closed convex cost function $\psi_j : \mathbb{R}^{q_1} \times \cdots \times \mathbb{R}^{q_T} \to (-\infty, +\infty]$ and the probability p_j. The entire decision vector $w = (w_1, \ldots, w_N) \in \mathbb{R}^{qN}$, where $q = q_1 + \ldots + q_T$, must satisfy the *nonanticipativity* constraint: for all $t = 1, \ldots, T-1$ and for all pairs (i, j) of scenarios (paths) with identical first t nodes, one must have

$$w_i(\tau) - w_j(\tau) = 0, \ \tau = 1, \ldots, t.$$

All these constraints (or a sufficient subset of them) can be put into one linear equation $Aw = \sum_{j=1}^N A_j w_j = 0$, where $A = [A_1 \cdots A_N]$ has dimension $m_A \times qN$. The entire problem can be formulated as follows:

$$\min \sum_{j=1}^N p_j \psi_j(w_j), \tag{6.1a}$$

$$\text{s.t.} \sum_{j=1}^N A_j w_j = 0. \tag{6.1b}$$

6.1 Augmented Lagrangian Decomposition

Consider the augmented Lagrangian for (6.1)

$$\Lambda(w, \lambda) = \sum_{j=1}^N p_j \psi_j(w_j) + \left\langle \lambda, \textstyle\sum_{j=1}^N A_j w_j \right\rangle + \dfrac{\rho}{2} \left| \textstyle\sum_{j=1}^N A_j w_j \right|^2, \tag{6.2}$$

where $\lambda \in \mathbb{R}^{m_A}$ and $\rho > 0$ is a penalty parameter. A solution of (6.1) can be obtained by the following method of multipliers (cf. [Ber82, Hes69, Pow69, Roc76a]).

Algorithm 6.1.
Step 0: Choose $\lambda^1 \in \mathbb{R}^{m_A}$. Set $l = 1$.
Step 1: Find $w^l \in \operatorname{Arg\,min}_w \Lambda(w, \lambda^l)$.
Step 2: Set $\lambda^{l+1} = \lambda^l + \rho A w^l$, increase l by 1 and go to Step 1.

It remains to determine an efficient method for minimizing (6.2). In fact, the alternating linearization algorithm is a good candidate. To see this, note that the problem in question is nearly identical to that presented in Example 1.1. In particular, we have:

$$h(w) = \sum_{j=1}^{N} \{p_j \psi_j(w_j) + \langle \lambda, A_j w_j \rangle\}$$

and

$$f(w) = \frac{\rho}{2} |Aw|^2 .$$

The functions h and f meet all the properties required by the alternating linearization algorithm. The separability of h means that Step 1 of Algorithm 3.1 can be decomposed into parallel subproblems for $j = 1, \ldots, N$:

$$z_{h,j}^k = \arg\min_{w_j} \left\{ p_j \psi_j(w_j) + \left\langle \lambda + \rho A z_f^k, A_j w_j \right\rangle + \frac{\rho_k}{2} \left| w_j - w_j^k \right|^2 \right\},$$

whereas Step 4 requires solving the least squares problem:

$$z_f^k = \arg\min_{w} \left\{ \langle g_h^k, w \rangle + \frac{\rho}{2} |Aw|^2 + \frac{\rho_{k+1}}{2} \sum_{j=1}^{N} \left| w_j - w_j^{k+1} \right|^2 \right\}.$$

6.2 Dual Strategy

All non-anticipative vectors $w = (w_1, \ldots, w_N)$ form a linear subspace $\mathcal{L}$ of $\mathbb{R}^{qN}$. The orthogonal projection on $\mathcal{L}$ will be denoted $\Pi_{\mathcal{L}}$. Given w, its projection $u = \Pi_{\mathcal{L}} w$ can be calculated as follows (see [RoW91]). For every $j = 1, \ldots, N$ and $t = 1, \ldots, T$, we find the set of scenarios indistinguishable from scenario j till stage t:

$$I_j(t) = \{i : \ \nu_\tau(i) = \nu_\tau(j), \ \tau = 1, \ldots, t\},$$

and we average $w_i(t)$ over this subset:

$$u_j(t) = \frac{1}{|I_j(t)|} \sum_{i \in I_j(t)} w_i(t).$$

Using the indicator function $\delta_{\mathcal{L}}$ of $\mathcal{L}$ we can formulate (6.1) equivalently as:

$$\min \left\{ \delta_{\mathcal{L}}(w) + \sum_{j=1}^{N} p_j \psi_j(w_j) \right\}. \tag{6.3}$$

Let r majorize the Euclidean norm of a solution to (6.1) and let $\mathcal{B} = \{y \in \mathbb{R}^{qN} : \ |y| \leq r\}$. With

$$\varphi(w) = \delta_{\mathcal{L} \cap \mathcal{B}}(w)$$

and

$$\psi(w) = \sum_{j=1}^{N} p_j \psi_j(w_j),$$

we can regard problem (6.3) as an instance of (5.1), where $M = I$ (the identity). For $x = (x_1, \ldots, x_N) \in \mathbb{R}^{qN}$, we have

$$h(x) = -\sum_{j=1}^{N} \inf_{w_j} \{p_j \varphi_j(w_j) - \langle x_j, w_j \rangle\}, \tag{6.4a}$$

$$f(x) = \max_{y} \{\langle -x, y \rangle : \ |y| \leq r, \ y \in \mathcal{L}\} = r|\Pi_{\mathcal{L}} x|, \tag{6.4b}$$

and the entire algorithm simplifies as follows.

Algorithm 6.2.
Step 0: Select $x^1 \in \mathbb{R}^{qN}$ and calculate $F(x^1) = h(x^1) + f(x^1)$, using (6.4). Choose $z_f^0 \in \mathbb{R}^{qN}$. Calculate $f(z_f^0) = r|\Pi_{\mathcal{L}} z_f^0|$ and $y^0 = -r\Pi_{\mathcal{L}} z_f^0 / |\Pi_{\mathcal{L}} z_f^0|$ ($u^0 = 0$ if $z_f^0 \perp \mathcal{L}$). Choose $\rho_1 \geq \rho_{\min} > 0$, $\kappa > 1$, $\beta_0 > 0$, $\beta_1 \in (0,1)$. Set $k = 1$.

Step 1: For scenarios $j = 1, \ldots, N$, calculate:

$$w_j^k = \arg\min_{w_j} \{p_j \psi_j(w_j) - \langle x_j^k, w_j \rangle + \frac{1}{2\rho_k} \left| w_j - y_j^{k-1} \right|^2\},$$

and set $z_h^k = x^k - (w^k - y^{k-1})/\rho_k$.

Step 2: Calculate

$$h(z_h^k) = \sum_{j=1}^{N} \left\{ \left\langle w_j^k, (z_h^k)_j \right\rangle - p_j \varphi_j(w_j^k) \right\},$$

$$f(z_h^k) = r|\Pi_{\mathcal{L}} z_h^k|,$$

$$\tilde{f}^k(z_h^k) = -\langle z_h^k, y^{k-1} \rangle.$$

Set $F(z_h^k) = h(z_h^k) + f(z_h^k)$ and $\tilde{F}^k(z_h^k) = h(z_h^k) + \tilde{f}^k(z_h^k)$. Set $v_k = \tilde{F}^k(z_h^k) - F(x^k)$. If $F(z_h^k) \leq F(x^k) + \beta_1 v_k$, then set $x^{k+1} = z_h^k$; otherwise set $x^{k+1} = x^k$.

Step 3: Choose ρ_{k+1} as at Step 3 of Algorithm 3.1.

Step 4: Calculate y^k as the orthogonal projection of $\tilde{y}^k = \Pi_{\mathcal{L}}(w^k - \rho_{k+1} x^{k+1})$ on the ball $\{y : \ |y| \leq r\}$.

Step 5: Increase k by 1 and go to Step 1.

To justify Step 4 of Algoritm 6.2 we note that

$$\begin{aligned} \arg\min_{y} \left\{ \varphi(y) + \langle x^{k+1}, y \rangle + \frac{1}{2\rho_{k+1}} |w^k - y|^2 \right\} \\ = \arg\min \left\{ \langle x^{k+1}, y \rangle + \frac{1}{2\rho_{k+1}} |w^k - y|^2 : \ |y| \leq r, \ y \in \mathcal{L} \right\} \\ = \arg\min \left\{ |w^k - \rho_{k+1} x^{k+1} - y|^2 : \ |y| \leq r, \ y \in \mathcal{L} \right\}. \end{aligned}$$

Major loop (l)	Alternating steps (k)	Descent steps	Null steps	$\|Aw^l\|^2/2$	$\frac{\|v^k\|^2}{1+\|F(x^k)\|^2}$
1	10	6	4	1284	1.90E-3
2	431	256	175	1.429	7.91E-7
3	24	11	13	0.276	4.48E-7
4	11	5	6	0.133	1.96E-7
5	13	9	4	0.104	1.62E-7
6	107	76	31	0.076	1.21E-7
7	1	1	0	0.049	1.18E-7

Table 7.1: Results for the augmented Lagrangian decomposition method

Algorithm 6.2 bears some similarities to the scenario aggregation method of [RoW91], which is a special version of the alternating direction method of multipliers. There are differences, though, in the way the multipliers x^k are updated and in the variable penalty coefficient. It is worth noting that the descent test in the dual space (Step 2) does not require much work, because the values of $F = h + f$ are easily available.

7 Numerical illustration

We consider a multistage stochastic macroeconomic energy model described in detail in [Ros94]. The model has the form (6.1) with $N = 8$, $n = 610$ and $m_A = 3240$. Each function ψ_j has a simple analytic form, but its domain is defined by 398 constraints, out of which 25 are nonlinear (with 85 "nonlinear" variables). Thus, out of 4880 variables in the entire model, 680 are "nonlinear" variables. The scenario model was formulated in GAMS [BKM92] and MINOS [MuS82] was used to solve scenario subproblems (with default parameters).

7.1 Augmented Lagrangian decomposition

Algorithm 6.1 was run with $\rho = 1$ and $\lambda^1 = 0$. At Step 1 we used Algorithm 3.1 with the following parameters: $\kappa = 2$, $\beta_0 = 1$, $\beta_1 = 0.1$, $\rho_1 = \rho$, $\rho_{\min} = \rho/1000$. It started from $x^1 = \arg\min\{h(x) + |x|^2/2\}$ at $l = 1$ and from w^{l-1} otherwise, and terminated when $\max\{|v_k|, |z_h^k - x^k|^2/2\} \leq 0.1|Aw^{l-1}|^2/2$ (with $w^0 = x^1$).

Seven major iterations of Algorithm 6.1 were made; the accuracy of the final solution was comparable with that obtained by other methods [RoR94, Rus95]. Table 7.1 illustrates our results. The relative accuracy in the inner loop was estimated by $|v_k|/(1 + |F(x^k)|)$.

The progress of the alternating linearization method at major iterations 2 and 6 is illustrated in Figures 7.1 and 7.2. The absolute error in the objective value was calculated as $F(x^k) - F(x^{k_*}) + v_{k_*}$, where k_* refers to the final iteration of Algorithm 3.1. We see that the algorithm can attain relatively high accuracy.

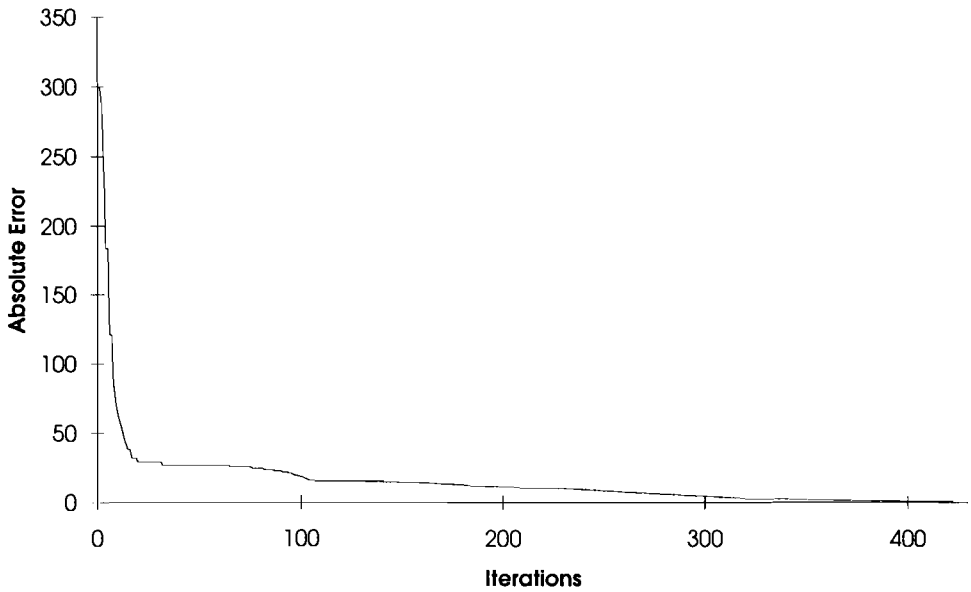

Figure 7.1: Absolute error in the objective value: Major iteration 2

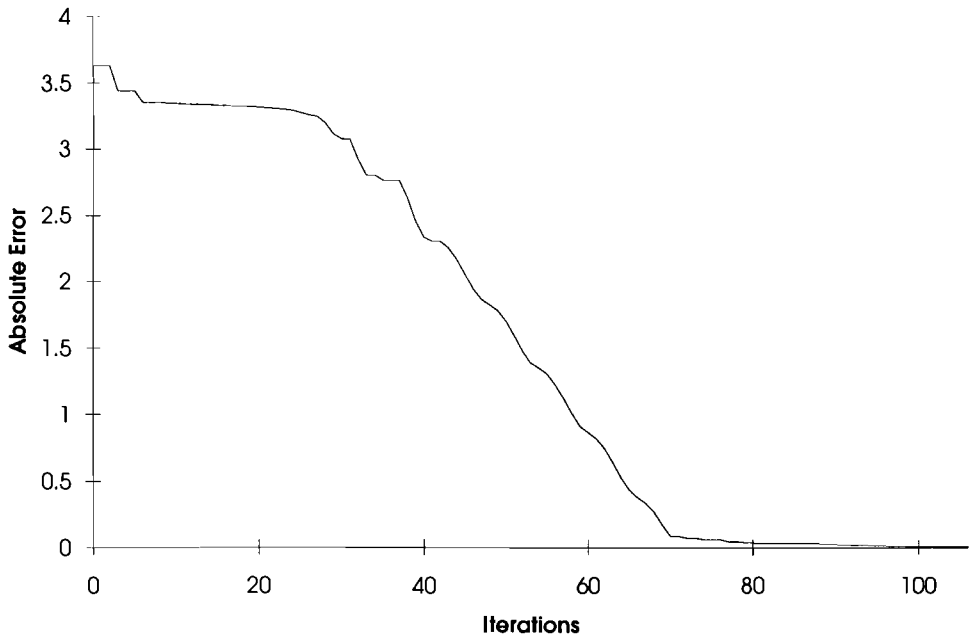

Figure 7.2: Absolute error in the objective value: Major iteration 6

7.2 Dual strategy

We chose $r = 3 \times 10^3$ large enough to majorize the solution obtained by other methods, so f (which may be interpreted as an exact penalty function) had rather steep walls. Accordingly, in Algorithm 6.2 we used a larger value of $\rho_1 = 10^6$. The other parameters were the same as in §7.1. The starting point was $x^1 = 0$.

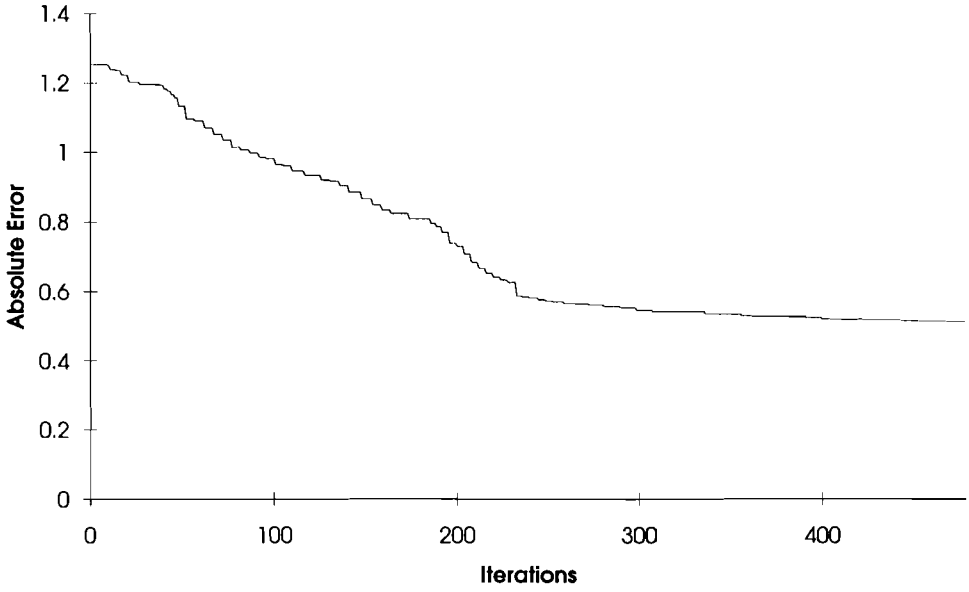

Figure 7.3: Dual method: absolute error in the objective value

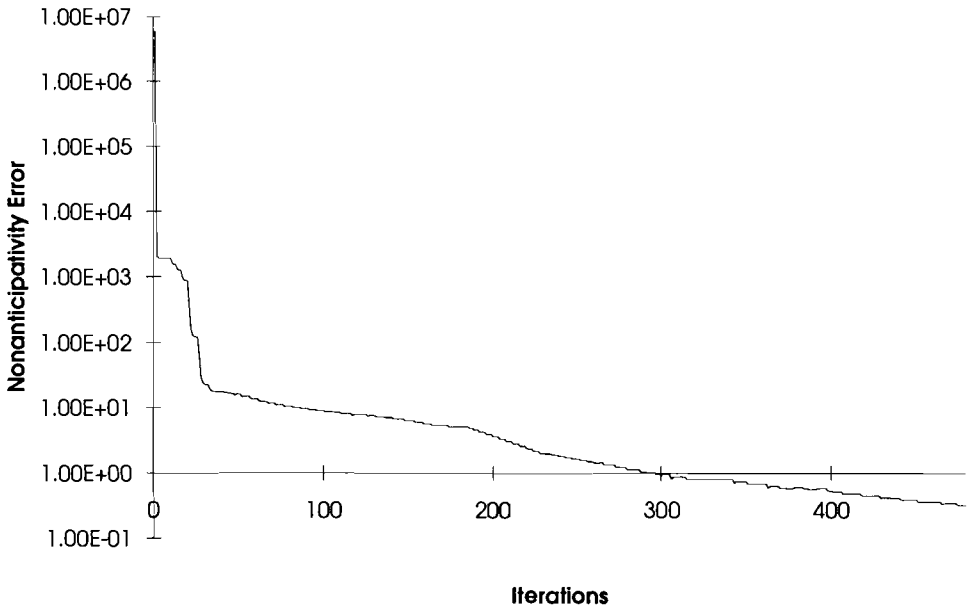

Figure 7.4: Dual method: nonanticipativity

Figure 7.3 illustrates the progress of the method in terms of the absolute error in the objective value: $\psi(w^k) - \psi_{\min}$ (where $\psi_{\min}$ is the known optimal value), and Figure 7.4

shows the decrease in the measure of nonanticipativity of the current solution: $|w^k - y^{k-1}|^2/2$. Again, we see that the method converges quickly at the initial stage, although the speed of convergence at the tail is not high, because of the essential nonsmoothness of f.

Summing up, this preliminary numerical experience indicates that the alternating linearization method, both in the primal and in the dual form, has a potential to become a useful tool for large-scale nonsmooth optimization.

References

[Aus86] A. Auslender, *Numerical methods for nondifferentiable convex optimization*, Math. Programming Stud. **30** (1986) 102–126.

[Ber82] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[BeT89] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[BKM92] A. Brooke, D. Kendrick and A. Meeraus, *GAMS: A User's Guide*, Scientific Press, San Francisco, 1992.

[ChR94] B. J. Chun and S. M. Robinson, *Scenario analysis via bundle decomposition*, Tech. report, Dept. of Ind. Eng., Univ. of Wisconsin, Madison, Wisconsin 53706, 1994.

[ChT94] G. Chen and M. Teboulle, *A proximal-based decomposition method for convex minimization problems*, Math. Programming **64** (1994) 81–101.

[CoL93] R. Correa and C. Lemaréchal, *Convergence of some algorithms for convex minimization*, Math. Programming **62** (1993) 261–275.

[DLMK+94] R. De Leone, R. R. Meyer, S. Kontogiorgis, Z. Zakarian and G. Zakeri, *Coordination in coarse-grained decomposition*, SIAM J. Optim. **4** (1994) 777–793.

[EcB92] J. Eckstein and D. P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Math. Programming **55** (1992) 293–318.

[EcF94a] J. Eckstein and M. C. Ferris, *Operator splitting methods for monotone affine variational inequalities, with a parallel application to optimal control*, Math. Prog. Tech. Report 94-??, Computer Sciences Dept., Univ. of Wisconsin, Madison, WI, Dec. 1994.

[EcF94b] J. Eckstein and M. Fukushima, *Some reformulations and applications of the alternating direction method of multipliers*, in Large Scale Optimization: State of the Art, W. W. Hager, D. W. Hearn and P. M. Pardalos, eds., Kluwer, Dordrecht, 1994, pp. 115–134.

[Eck94] J. Eckstein, *Some saddle-function splitting methods for convex programming*, Optimization Methods & Software **4** (1994) 75–83.

[Fuk92] M. Fukushima, *Application of the alternating direction method of multipliers to separable convex programming problems*, Comput. Optim. Appl. **1** (1992) 93–111.

[Gab83] D. Gabay, *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.

[GoT89] E. G. Golshtein and N. V. Tretyakov, *Modified Lagrange Functions; Theory and Optimization Methods*, Nauka, Moscow, 1989 (Russian).

[Gül91] O. Güler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Control Optim. **29** (1991) 403–419.

[HaL88] S.-P. Han and G. Lou, *A parallel algorithm for a class of convex programs*, SIAM J. Control Optim. **26** (1988) 345–355.

[Hes69] M. R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appl. **4** (1969) 303–320.

[HUL93] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.

[Lem89] B. Lemaire, *The proximal algorithm*, in New Methods in Optimization and Their Industrial Uses, J. P. Penot, ed., International Series of Numerical Mathematics 87, Birkhäuser, Basel, 1989, pp. 73–87.

[Mar70] B. Martinet, *Régularisation d'inéquations variationelles par approximations successives*, RAIRO Rech. Opér. **4(R3)** (1970) 154–158.

[MaT92] P. Mahey and P.-D. Tao, *Partial regularization of the sum of two maximal monotone operators*, Tech. report, Laboratoire ARTEMIS, IMAG, Grenoble, France, 1992.

[MNS91] K. Mouallif, V. H. Nguyen and J.-J. Strodiot, *A perturbed parallel decompositon method for a class of nonsmooth convex minimization problems*, SIAM J. Control Optim. **29** (1991) 829–847.

[MOT95] P. Mahey, S. Oualibouch and P.-D. Tao, *Proximal decomposition on the graph of a maximal monotone operator*, SIAM J. Optim. **5** (1995) ?–? To appear.

[MuR95] J. M. Mulvey and A. Ruszczyński, *A new scenario decomposition method for large scale stochastic optimization*, Oper. Res. **43** (1995) ?–? To appear.

[MuS82] B. A. Murtagh and M. A. Saunders, *A projected Lagrangian algorithm and its implementation for sparse nonlinear constraints*, Math. Programming Stud. **16** (1982) 84–117.

[Pow69] M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, London, 1969, pp. 283–298.

[Rob91] S. M. Robinson, *Extended scenario analysis*, Ann. Oper. Res. **31** (1991) 385–398.

[Roc70] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[Roc76a] ———, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res. **1** (1976) 97–116.

[Roc76b] ———, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim. **14** (1976) 877–898.

[RoR94] C. H. Rosa and A. Ruszczyński, *On augmented Lagrangian decomposition methods for multistage stochastic programs*, WP-94-05, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1994.

[Ros94] C. H. Rosa, *Pathways of economic development in an uncertain environment: A finite scenario approach to the u.s. region under carbon emission restrictions*, WP-94-41, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1994.

[RoW91] R. T. Rockafellar and R. J.-B. Wets, *Scenarios and policy aggregation in optimization under uncertainty*, Math. Oper. Res. **16** (1991) 1–23.

[Rus95] A. Ruszczyński, *On convergence of an augmented Lagrangian decomposition method for sparse convex optimization*, Math. Oper. Res. **?** (1995) ?–? To appear.

[Spi85] J. E. Spingarn, *Applications of the method of partial inverses to convex programming: Decomposition*, Math. Programming **32** (1985) 199–223.

[Tse90] P. Tseng, *Further applications of a splitting algorithm to decomposition in variational inequalities and convex programming*, Math. Programming **48** (1990) 249–263.

[Tse91] ———, *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities*, SIAM J. Control Optim. **29** (1991) 119–138.

[Tse94] ———, *Alternating projection-proximal methods for convex programming and variational inequalities*, Tech. report, Dept. of Mathematics, Univ. of Washington, Seattle, WA, Dec. 1994.