Working Paper

Decomposition via Alternating Linearization

Krzysztof C. Kiwiel
Charles H. Rosa
Andrzej Ruszczyński

WP-95-051
June 1995
Decomposition via Alternating Linearization

Krzysztof C. Kiwiel
Charles H. Rosa
Andrzej Ruszczyński

WP-95-051
June 1995

Working Papers are interim reports on work of the International Institute for Applied Systems Analysis and have received only limited review. Views or opinions expressed herein do not necessarily represent those of the Institute, its National Member Organizations, or other organizations supporting the work.
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:

\[
\text{minimize } F(x) := h(x) + f(x),
\]

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

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

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:

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

and

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

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:

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

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

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:

\[
\min_y \left\{ \varphi(y) + \sum_{j=1}^{N} \psi_j(y) \right\}
\]

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:

\[
\min_x \left\{ \sum_{j=1}^{N} \psi^*_j(x_j) + \varphi^* \left( -\sum_{j=1}^{N} x_j \right) \right\},
\]

where \( \varphi^* \) and \( \psi^*_j \) are the conjugates of \( \varphi \) and \( \psi_j \), 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 \( f^k \) and \( h^k \), 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 \( \rho_k \).

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

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

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

**Algorithm 2.1.**

**Step 0:** Choose \( z_j^0 \in \mathbb{R}^n \) and \( g_j^0 \in \partial f(z_j^0) \). Define \( \tilde{f}^1(\cdot) = f(z_j^0) + \langle g_j^0, \cdot - z_j^0 \rangle \). Set \( k = 1 \).

**Step 1:** Find the solution \( z_k^h \) of the problem:

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

Set

\[ g_h^k = -g_j^{k-1} - \rho(z_h^k - \bar{x}) \]  

and define

\[ \tilde{h}^k(\cdot) = h(z_h^k) + \langle g_h^k, \cdot - z_h^k \rangle. \]  

**Step 2:** Find the solution \( z_j^k \) of the problem

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

Set

\[ g_j^k = -g_j^{k-1} - \rho(z_j^k - \bar{x}) \]  

and define

\[ \tilde{f}^{k+1}(\cdot) = f(z_j^k) + \langle g_j^k, \cdot - z_j^k \rangle. \]  

**Step 3:** Increase \( k \) by 1 and go to Step 1.

Our objective is to prove that \( z_k^h \to p(\bar{x}) \).

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

\[ 0 \in \partial h(z_h^k) + g_j^{k-1} + \rho(z_h^k - \bar{x}), \]  

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

Let us denote by

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

and

\[ \eta_{k+1/2} = \tilde{h}^k(z_j^k) + f(z_j^k) + \frac{\rho}{2} |z_j^k - \bar{x}|^2 \]
the optimal values of (2.2) and (2.5), respectively. The way in which the successive linearizations \( \hat{f}^k \) and \( \hat{h}^k \) are generated ensures monotonicity of \( \{\eta_k\} \):

\[
\eta_k \leq \eta_{k+1/2} \leq \eta_{k+1}.
\] (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 \( \hat{h}^k(\cdot) \);

(b) replace \( f^k(\cdot) \) by \( f \).

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

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

\[
\min_x \left\{ Q_k(x, \mu) = \hat{h}^k(x) + (1 - \mu)(\alpha_p^k + \langle p^k, x \rangle) + \mu(\alpha_g^k + \langle g^k, x \rangle) + \frac{\rho}{2} |x - \hat{x}|^2 \right\},
\] (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 \( \hat{f}^k(\cdot) = \alpha_p^k + \langle p^k, \cdot \rangle \) 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 \( Q_k(\mu) \) of (2.11).

**Lemma 2.3.** The following inequalities hold for any \( g^k \in \partial f(z_h^k) \):

(i) \( \max_{\mu \in [0,1]} Q_k(\mu) - Q_k(0) \geq Q_k(\mu_k) - Q_k(0) \geq \bar{\mu}_k \delta_k/2 \),

(ii) \( \eta_{k+1} \geq \eta_{k+1/2} \geq \eta_k + \bar{\mu}_k \delta_k/2 \),

where \( \delta_k = F(z_h^k) - \hat{F}^k(z_h^k) \geq 0 \) and \( \bar{\mu}_k = \min \{1, \delta_k/|g^k - p^k|^2\} \).

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

\[
Q_k'(\mu) = (g^k - p^k, \hat{x}(\mu)) + \alpha_g^k - \alpha_p^k
= (g^k - p^k, \hat{x}(\mu) - \hat{x}(0)) + (\alpha_g^k + \langle g^k, \hat{x}(0) \rangle) - (\alpha_p^k + \langle p^k, \hat{x}(0) \rangle)
= (g^k - p^k, \hat{x}(\mu) - \hat{x}(0)) + F(z_h^k) - \hat{F}^k(z_h^k)
= -\frac{\mu |g^k - p^k|^2}{\rho} + \delta_k,
\]

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

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

Using the definition of \( \bar{\mu}_k \) yields (i). Assertion (ii) follows from (i) and (2.10). \( \square \)
Theorem 2.4. The sequences of points \( \{z^k\} \) and approximations \( \{\hat{F}^k\} \) generated by Algorithm 2.1 have the following properties:

(i) \( |z^k - p(\bar{x})| \leq \left\{ |F(z^k) - \hat{F}^k(z^k)|/\rho \right\}^{1/2} \) for \( k = 1, 2, \ldots \).

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

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

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

\[
F(p(\bar{x})) + \frac{\rho}{2} |p(\bar{x}) - \bar{x}|^2 \geq \hat{F}^k(p(\bar{x})) + \frac{\rho}{2} |p(\bar{x}) - \bar{x}|^2
\]

\[
\geq \hat{F}^k(z^k) + \frac{\rho}{2} |z^k - \bar{x}|^2 + \frac{\rho}{2} |p(\bar{x}) - z^k|^2. \quad (2.12)
\]

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

\[
F(z^k) + \frac{\rho}{2} |z^k - \bar{x}|^2 \geq F(p(\bar{x})) + \frac{\rho}{2} |p(\bar{x}) - \bar{x}|^2 + \frac{\rho}{2} |p(\bar{x}) - z^k|^2.
\]

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

\[
\frac{\rho}{2} |p(\bar{x}) - z^k|^2 \leq F(p(\bar{x})) + \frac{\rho}{2} |p(\bar{x}) - \bar{x}|^2 - \eta_k. \quad (2.13)
\]

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

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

we see that \( z^k \) and \( p^k \) are bounded. By (2.13), the sequence \( \{\eta_k\} \) is bounded from above, so Lemma 2.3 implies that it converges and \( \hat{p}_k \delta_k \to 0 \). Since \( \{\eta_k\} \) is bounded, assertion (ii) follows from the definition of \( \hat{p}_k \) (cf. Lemma 2.3). The final assertion is a consequence of (i) and (ii). \( \square \)

Remark 2.5. Algorithm 2.1 can be used in the implementable proximal point schemes of [Aus86, CoL93, EcB92, GoT89, Gul91, Lem89, Roc76b]. Indeed, Theorem 2.4 ensures that for every \( \epsilon > 0 \) we can find in finitely many steps a point \( z^k \) such that \( |z^k - p(\bar{x})| \leq \epsilon \). An alternative scheme will be presented in the next section.

3 The alternating linearization method

The algorithm 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 \( x^1 \in \text{dom} \ h \), \( z^1 \in \mathbb{R}^n \) and \( g^0 \in \partial f(z^0) \). Define \( \hat{f}^1(\cdot) = f(z^0) + \langle g^0, \cdot - z^0 \rangle \).

Choose parameters \( \rho_1 \geq \rho_{\text{min}} > 0, \kappa > 1, \beta_0 > 0, \beta_1 \in (0, 1) \). Set \( k = 1 \).
Step 1: Find the solution \( z_k^k \) of the problem

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

Set \( g_h^k = -g_f^k - \rho_k (z_h^k - x^k) \) 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

\[ v_k = \tilde{F}^k(z_h^k) - F(x^k). \]

If

\[ F(z_h^k) \leq F(x^k) + \beta_1 v_k, \]

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 \( \rho_{k+1} \in [\max\{\rho_{\text{min}}, \rho_k/\kappa\}, \rho_k] \). If \( x^{k+1} = x^k \) and

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

then choose \( \rho_{k+1} \geq \rho_k \), else set \( \rho_{k+1} = \rho_k \).

Step 4: Find the solution \( z_f^k \) of the problem

\[
\min_x \left\{ \tilde{f}^k(x) + f(x) + \frac{\rho_{k+1}}{2} |x - x^{k+1}|^2 \right\}.
\]

Set \( g_f^k = -g_f^k - \rho_{k+1} (z_f^k - x^{k+1}) \) 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

\[ \eta_k = \tilde{F}^k(z_h^k) + \frac{\rho_k}{2} |z_h^k - x^k|^2 \]

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 \( \tilde{F}^k \leq F \), so \( \eta_k \leq F(x^k) \) and \( v_k \leq 0 \). Thus (3.3) implies that \( \{F(x^k)\} \) is nonincreasing and \( \{x^k\} \subset \text{dom } F \). It will become clear that if \( v_k = 0 \) or \( \eta_k = F(x^k) \) then \( x^k \in \text{Argmin } F \).

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

(ii) \( \frac{\rho_{k+1}}{2} |z_f^k - x^{k+1}|^2 \leq F(x^{k+1}) - \eta_{k+1/2} \).
Proof. (3.2) and (3.5) yield \( F(x^k) + v_k \leq \eta_k \), and hence the right inequality of (i). Next, note that by construction (cf. Step 1)

\[
- \rho_k (z^k_h - x^k) = g^k_h + g^{k-1}_f \in \partial \tilde{F}^k(z^k_h),
\]

(4.1)
so the left inequality in (i) follows from the subgradient inequality, since

\[
-v_k = F(x^k) - \tilde{F}^k(z^k_h) \geq \tilde{F}^k(x^k) - \tilde{F}^k(z^k_h) \geq \rho_k |z^k_h - x^k|^2.
\]

Thus

\[
\eta_k = F(x^k) + v_k + \frac{\rho_k}{2} |z^k_h - x^k|^2 \leq F(x^k) + \frac{v_k}{2},
\]

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 \( x^k \in \text{Arg min} \, F \).

**Proof.** By Lemma 4.1(i) and (3.2), \( z^k_h = x^k \) and \( \tilde{F}^k(z^k_h) = F(z^k_h) = F(x^k) \). Then Theorem 2.4(i) yields \( x^k = z^k_h = \text{arg min} \, F + \rho_k \cdot \|\cdot\|^2/2 \), so \( x^k \in \text{Arg min} \, F \) [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 \( F(x^{x_0}) \), where \( x^{x_0} \) is the last point to which a descent step was made.

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

\[
\eta_{k+1} \geq \eta_k + \beta_1 \mu_k |v_k|/2,
\]

where \( \mu_k = \min \{1, \beta_1 |v_k|/|g_f(z^k_h) - g^{k-1}_f|^2\} \) for any \( g_f(z^k_h) \in \partial f(z^k_h) \).

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

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

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

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

**Lemma 4.5.** Suppose the set \( \mathcal{K} = \{k : x^{k+1} \neq x^k\} \) 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 \( k \in K, \) 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 \( k \in K. \) In view of Lemma 4.1, \( F(x^k) - \eta_k \to 0 \) for \( k \in K. \) 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 \( \rho_l \) to \( \rho_{l+1} \in [\rho_l/\kappa, \rho_l]; \)

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

Denote by \( \eta^{(b)}_l \) the resulting optimal value of (3.1) after partial modifications (a) and (b). By construction, \( g^l_h = g^l_h + g^l_{f-1} \in \partial \tilde{F}^l(x^{l+1}) \) is such that \( x^{l+1} - x^l = -g^l_h/\rho_l \) (cf. (4.1)) and

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

so

\[
\eta_l = \tilde{F}^l(x^{l+1}) + \frac{1}{2\rho_l} |g^l_h|^2 = \tilde{F}^l(x^l) - \frac{1}{2\rho_l} |g^l_h|^2.
\]

In a similar way,

\[
\eta^{(b)}_l = \min_x \left\{ \tilde{F}^l(x^{l+1}) + \langle g^l_h, x - x^{l+1} \rangle + \frac{\rho_{l+1}}{2} |x - x^{l+1}|^2 \right\} = \tilde{F}^l(x^{l+1}) - \frac{1}{2\rho_{l+1}} |g^l_h|^2.
\]

Therefore,

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

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

\[
\eta_{l+1} \geq \eta_{l+1/2} \geq \eta^{(b)}_l.
\]

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 \( \lim_{k \to \infty} F(x^{(k+1)}) - \eta_{(k+1)} = 0 \). 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 \( K = \{ k : x^{k+1} \neq x^k \} \) is infinite. If there exists a point \( \hat{x} \) such that \( F(x^k) \geq F(\hat{x}) \) for all \( k \), then \( \{ x^k \} \) converges to a point \( \hat{x} \in \text{dom } F \).

**Proof.** Fix \( k \in K \). We have

\[
|x^{k+1} - \hat{x}|^2 = |x^k - \hat{x}|^2 + 2(x^{k+1} - \hat{x}, x^{k+1} - x^k) - |x^{k+1} - x^k|^2.
\]

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

\[
\rho_k(x^{k+1} - \hat{x}, x^{k+1} - x^k) = \langle \hat{x} - x^{k+1}, g^k_F \rangle \\
\leq \hat{F}^k(\hat{x}) - \hat{F}^k(x^{k+1}) \leq F(\hat{x}) - F(x^k) - v_k.
\]

Using this inequality in (4.3) yields

\[
|x^{k+1} - \hat{x}|^2 \leq |x^k - \hat{x}|^2 + 2|v_k|/\rho_k, \quad k \in K.
\]

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 \( \hat{x} \) by \( \bar{x} \) in the preceding argument, concluding that \( \bar{x} \) is the only accumulation point, since \( \sum_{k \in K, k \geq l} |v_k| \to 0 \) as \( l \to \infty \). □

**Lemma 4.7.** If there exists a point \( \hat{x} \) such that \( F(x^k) \geq F(\hat{x}) \) 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 \( \bar{x} \in \text{Arg min } F \).

**Proof.** By Lemmas 4.4–4.6, \( \{ x^k \} \) converges to some \( \bar{x} \in \text{dom } F \) 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 \( \hat{F}^k \leq F \),

\[
F(x^k) - \eta_k = F(x^k) - \min_x \left\{ \hat{F}^k(x) + \frac{\rho_k}{2} |x - x^k|^2 \right\} \\
\geq F(x^k) - \min_x \left\{ F(x) + \frac{\bar{\rho}}{2} |x - x^k|^2 \right\} \\
\geq F(x^k) - \min_x \left\{ F(x) + \frac{\bar{\rho}}{2} |x - x^k|^2 \right\} \geq 0.
\]

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 \( \bar{x} \in \text{Arg min } F \) (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
\[
\frac{\rho_k}{2} |z_k^h - x^h|^2 \leq |v_k| \to 0.
\] (4.4)

With $\rho_k \geq \rho_{\min}$ one must have $z_k^h - x^h \to 0$. In a similar way, $z_j^f - x^h \to 0$. Since $f$ is continuous over the domain of $h$,
\[
F(z_k^h) - \bar{F}(z_k^h) = F(z_k^h) - F(x^h) - v_k = f(z_k^h) - f(x^h) - v_k \to 0.
\] (4.5)
The penalty coefficient is increased infinitely many times, so (cf. Step 3) there must be a subsequence $K$ such that for $k \in K$
\[
F(z_k^h) - F(x^h) - v_k \geq \beta_0 |v_k| / |z_k^h - x^h|.
\] (4.6)
Dividing (4.4) by $|z_k^h - x^h|$ and using (4.6) and (4.5), we get $\rho_k |z_k^h - x^h| \to 0$. Therefore, using the definition of $g_k^h$ at Step 1,
\[
g_k^h + g_j^{k-1} \to 0, \quad k \in K.
\] (4.7)
Since $f$ is locally Lipschitz and $\{z_j^f\}$ is bounded, the vectors $g_j^f \in \partial f(z_j^f)$ are uniformly bounded. By the upper semicontinuity of $\partial f$ (cf. [Roc70, Thm 24.4]), we can restrict $K$ so that $g_j^{k-1} \to g_j(z) \in \partial f(z)$, $k \in K$. Then $g_k^h \to -g_j(z)$, $k \in K$. Consequently, $-g_j(z) \in \partial h(z)$, because $z_k^h \to z$ and $g_k^h \in \partial h(z_k^h)$. This proves that $0 \in \partial F(z)$.

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 $\text{Arg min} \ F \neq \emptyset$ then $\{x^k\}$ converges to a point $\hat{x} \in \text{Arg min} \ F$.

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

(iv) If $\text{Arg min} \ F \neq \emptyset$ and the sequence $\{\rho_k\}$ is bounded, then the sequences $\{g_j^f\}$ and $\{g_k^h\}$ are bounded, $g_k^h + g_j^{k-1} \to 0$, $g_k^h + g_j^f \to 0$, and every accumulation point $(\hat{g}_f, \hat{g}_h)$ of $\{(g_j^f, g_k^h)\}$ 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 $\text{Arg min} \ F$ contains a point $\hat{x}$, one has $F(x^k) \geq F(\hat{x})$ for all $k$. Then by Lemma 4.7, $x^k \to \hat{x} \in \text{Arg min} \ F$, and $F(x^k) \downarrow F(\hat{x}) = \inf F$, which proves (i)–(ii) in this case.

Suppose now that $\text{Arg min} \ F = \emptyset$. If there existed $\hat{x}$ such that $F(x^k) \geq F(\hat{x})$ for all $k$, then Lemma 4.7 would imply convergence of $\{x^k\}$ to a minimizer of $F$, a contradiction. Therefore for every $\hat{x}$ we can find $k$ such that $F(x^k) < F(\hat{x})$. 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 $\text{Arg min} \ F \neq \emptyset$ and the sequence $\{\rho_k\}$ is bounded. We already know that $x^k \to \hat{x} \in \text{Arg min} \ F$. 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 $z_k^h \to \hat{x}$ and $z_j^f \to \hat{x}$. Since $g_j^f \in \partial f(z_j^f)$ and $f$ is locally Lipschitz, the sequence $\{g_j^f\}$ is bounded and each of its accumulation point is in $\partial f(\hat{x})$. Next, by the definitions of $g_j^f$ and $g_k^h$, $g_j^f + g_k^h = \rho_k(z_j^f - x^{k+1}) \to 0$ and $g_j^{k-1} + g_k^h = \rho_k(z_k^h - x^k) \to 0$. Thus $\{g_k^h\}$ 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:

\[
\inf \{ \varphi(y) + \psi(My) \}
\]

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

\[
\begin{align*}
\inf \{ \varphi(y) + \psi(w) \} \\
w - My = 0,
\end{align*}
\]

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

\[
\sup_x \left\{ L_D(x) = \inf_{y, w} L(y, w, x) \right\}
\]

can be equivalently written as

\[
\inf_x \left\{ F(x) = \psi^*(x) + \varphi^*(-M^T x) \right\},
\]

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 \( \text{dom } f = \mathbb{R}^n \). 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 \( x^1 \in \text{dom } h \) and calculate \( F(x^1) = h(x^1) + f(x^1) \). Choose \( z_f^0 \in \mathbb{R}^n \). Calculate

\[
f(z_f^0) = -\min_y \{ \varphi(y) + \langle z_f^0, My \rangle \}.
\]

Choose a minimizer \( y^0 \) in the problem above. Select \( \rho_1 \geq \rho_{\text{min}} > 0 \), \( \kappa > 1 \), \( \beta_0 > 0 \), \( \beta_1 \in (0, 1) \). 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
\[ z^k_h = x^k - (w^k - My^{k-1})/\rho_k. \tag{5.6} \]

Step 2: Calculate
\[ h(z^k_h) = \langle w^k, z^k_h \rangle - \psi(w^k), \tag{5.7} \]
\[ f(z^k_h) = - \min_y \left\{ \varphi(y) + \langle z^k_h, My \rangle \right\}, \tag{5.8} \]
\[ \tilde{f}^k(z^k_h) = - \left\{ \varphi(y^{k-1}) + \langle z^k_h, My^{k-1} \rangle \right\}. \tag{5.9} \]
Set \( F(z^k_h) = h(z^k_h) + f(z^k_h) \) and \( \tilde{F}^k(z^k_h) = h(z^k_h) + \tilde{f}^k(z^k_h) \). Set \( v_k = \tilde{F}^k(z^k_h) - F(x^k) \). If \( F(z^k_h) \leq F(x^k) + \beta tv_k \), then set \( x^{k+1} = z^k_h \); otherwise set \( x^{k+1} = x^k \).

Step 3: Choose \( \rho_{k+1} \) as at Step 3 of Algorithm 3.1.

Step 4: Calculate
\[ y^k \in \arg \min_y \left\{ \varphi(y) + \langle x^{k+1}, My \rangle + \frac{1}{2\rho_{k+1}} |w^k - My|^2 \right\}. \tag{5.10} \]

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:
\( x^* \in \partial f(x) \), \( x \in \partial f^*(x^*) \), \( f(x) + f^*(x^*) = \langle x, x^* \rangle \), \( x \in \operatorname{Argmin}\{f(\cdot) - \langle x^*, \cdot \rangle\} \).

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 \( \operatorname{Argmin} F \neq \emptyset \) then \( \{x^k\} \) converges to a point \( \hat{x} \in \operatorname{Argmin} F \).
(iii) If \( \operatorname{Argmin} F = \emptyset \) then \( |x^k| \to \infty \).
(iv) If \( \operatorname{Argmin} F \neq \emptyset \) 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
\[ y^0 \in \partial \varphi^*(-M^Tz^0_j). \]
Therefore, by Fact 5.2, \( -My^0 \in \partial f(z^0_j) \) and we can define \( g^0_j = -My^0 \).

We shall use induction. Assume that for some \( k \) we have
\[ y^{k-1} \in \partial \varphi^*(-M^Tz^{k-1}_j) \tag{5.11} \]
and

\[ g_f^{k-1} = -M y^{k-1}. \]  \hspace{1cm} (5.12)

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

\[ \min_x \left\{ \psi^*(x) - \langle M y^{k-1}, x \rangle + \frac{\rho_k}{2} |x - x^k|^2 \right\}. \]  \hspace{1cm} (5.13)

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

\[ z^k = x^k - \frac{w^k - M y^{k-1}}{\rho_k} \in \partial \psi(w^k), \]  \hspace{1cm} (5.14)

which by Fact 5.2 is equivalent to

\[ w^k \in \partial \psi^*(z^k). \]  \hspace{1cm} (5.15)

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

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

Combining this relation with (5.12) we obtain

\[ f^k(z^k) = f(z_j^{k-1}) + \langle g_f^{k-1}, z^k - z_j^{k-1} \rangle \]

\[ = -\varphi(y^{k-1}) - \langle M^T z_j^{k-1}, y^{k-1} \rangle - \langle M y^{k-1}, z^k - z_j^{k-1} \rangle, \]

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

\[ g^k_h = -g_f^{k-1} - \rho_k (z^k - x^k) = w^k. \]  \hspace{1cm} (5.16)

Therefore, problem (3.4) can be written as

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

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

\[ z_j^k = x^{k+1} - \frac{w^k - M y^k}{\rho_{k+1}}, \]  \hspace{1cm} (5.18)

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

\[-M^T z_j^k = -M^T x^{k+1} + M^T (w^k - M y^k)/\rho_{k+1} \in \partial \varphi(y^k), \]  \hspace{1cm} (5.19)

which by Fact 5.2 is equivalent to the relation \( y^k \in \partial \varphi^*(-M^T z_j^k) \), i.e., (5.11) holds for \( k \). The last relation is equivalent to \( -M y^k \in \partial f(z_j^k) \) (Fact 5.2). Substitution of \( M y^k \) from
(5.18) yields the optimality condition for (5.17): 
\[-w^k - \rho_{k+1}(z_j^k - x^{k+1}) \in \partial f(z_j^k).\]
Finally, from (5.16) and (5.18) we get
\[g_j^k = -g_h^k - \rho_{k+1}(z_j^k - x^{k+1}) = -My^k,\]
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,
\[w^k - My^{k-1} \to 0\]
and \(w^k - My^{k-1} \to 0\). To complete the proof of (iv), let \((w^k, y^k) \to (\hat{w}, \hat{y})\), \(k \in K\). Taking limits in (5.14) and (5.19), we obtain \(\hat{x} \in \partial \psi(\hat{w}), -M^T \hat{x} \in \partial \phi(\hat{y})\) and, by (5.21), \(\hat{w} - M\hat{y} = 0\). This proves the optimality of \((\hat{w}, \hat{y})\) in (5.2). \(\square\)

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):
\[
\Lambda_{\rho}(y, w, x) = \phi(y) + \psi(w) - \langle x, w - My \rangle + \frac{\rho}{2} |w - My|^2,
\]
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 \(\rho_k\) fixed at \(\rho\), 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
\[
\min \left\{ \phi(y) + \sum_{j=1}^{N} \psi_j(y) \right\},
\]
with closed proper convex functions \(\phi : \mathbb{R}^m \to (-\infty, +\infty)\) 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 \(j = 1, \ldots, N\):
\[w_j^k = \arg \min_{w_j} \left\{ \psi_j(w_j) - \langle x_j^k, w_j \rangle + \frac{1}{2\rho_k} |w_j - y_j^{k-1}|^2 \right\},\]
\[
\text{14}
\]
\[(z^k)_j = x^k_j - (w^k_j - y^{k-1})/\rho_k,\]

while (5.10) takes the form:

\[y^k = \arg \min_y \left\{ \varphi(y) + \langle \sum_{j=1}^N x^{k+1}_j, y \rangle + \frac{1}{2\rho_{k+1}} \sum_{j=1}^N |w^k_j - y|^2 \right\}.\]

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

\[w_i(t) - w_j(t) = 0, \quad t = 1, \ldots, T.\]

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:

\[\begin{align*}
\min \sum_{j=1}^N p_j \psi_j(w_j), \quad (6.1a) \\
s.t. \sum_{j=1}^N A_j w_j = 0. \quad (6.1b)
\end{align*}\]

6.1 Augmented Lagrangian Decomposition

Consider the augmented Lagrangian for (6.1)

\[\Lambda(w, \lambda) = \sum_{j=1}^N p_j \psi_j(w_j) + \langle \lambda, \sum_{j=1}^N A_j w_j \rangle + \frac{\rho}{2} \sum_{j=1}^N |A_j w_j|^2, \quad (6.2)\]

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 $w^l \in \text{Arg}\min_w \Lambda(w, \lambda^l)$.

**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:

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

and

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

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 $j = 1, \ldots, N$:

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

whereas Step 4 requires solving the least squares problem:

$$z_j^k = \arg\min_w \left\{ (q_j^k, w) + \frac{\rho}{2} |A w|^2 + \frac{\rho_{k+1}}{2} \sum_{j=1}^{N} |w_j - w_j^{k+1}|^2 \right\}.$$

### 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 $t$:

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

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

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

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

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

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

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

16
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

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

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

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 $z_0^0 \in \mathbb{R}^{qN}$. Calculate $f(z_0^0) = r|\Pi_{\mathcal{L}}z_0^0|$ and $y_0^0 = -r\Pi_{\mathcal{L}}z_0^0/||\Pi_{\mathcal{L}}z_0^0||$ ($u_0^0 = 0$ if $z_0^0 \perp \mathcal{L}$). 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:

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

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

**Step 2:** Calculate

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

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

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

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

**Step 3:** Choose $\rho_{k+1}$ 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 Algorithm 6.2 we note that

$$\arg \min_{y} \left\{ \varphi(y) + \langle x^{k+1}, y \rangle + \frac{1}{2\rho_{k+1}} |w^k - y|^2 \right\}$$

$$= \arg \min \left\{ \langle x^{k+1}, y \rangle + \frac{1}{2\rho_{k+1}} |w^k - y|^2 : |y| \leq r, \ y \in \mathcal{L} \right\}$$

$$= \arg \min \left\{ |w^k - \rho_{k+1} x^{k+1} - y|^2 : |y| \leq r, \ y \in \mathcal{L} \right\}.$$
Ivli
l2
Major Alternating Descent
Null
IAw'I2/2
l+l~(zk)12

Loop (1) steps (K) steps steps
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 $\lambda^k$ 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 $\psi_j$ 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 $\max \{|v_k|, |z^k|/x^k|^2/2\} \leq 0.1|Aw^{l-1}|^2/2$ (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](image)

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

![Figure 7.4: Dual method: nonanticipativity](image)
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


