J Optim Theory Appl (2012) 153:123–138
DOI 10.1007/s10957-011-9936-x

Coupling the Gradient Method with a General Exterior Penalization Scheme for Convex Minimization

Juan Peypouquet

Received: 13 March 2011 / Accepted: 21 September 2011 / Published online: 18 October 2011

Abstract In this paper, we propose and analyze an algorithm that couples the gradient method with a general exterior penalization scheme for constrained or hierarchical minimization of convex functions in Hilbert spaces. We prove that a proper but simple choice of the step sizes and penalization parameters guarantees the convergence of the algorithm to solutions for the optimization problem. We also establish robustness and stability results that account for numerical approximation errors, discuss implementation issues and provide examples in finite and infinite dimension.

Keywords Convex optimization · Hierarchical minimization · Exterior penalization · Non-autonomous gradient-like systems

1 Introduction

This paper is concerned with the study of a class of gradient-type algorithms to solve constrained or hierarchical optimization problems in Hilbert spaces, using a fairly general exterior penalization procedure.

The main result is that any sequence generated by our *diagonal gradient scheme* (DGS) converges weakly to a solution of the constrained optimization problem; convergence being strong, if the objective function is strongly convex. Moreover, it is possible to account for numerical errors in the computation of the iterates, which may arise, for instance, from inaccurate evaluations of the functions and gradients.

Communicated by Jean-Pierre Crouzeix.

J. Peypouquet (✉)
Departamento de Matemática, Universidad Técnica Federico Santa María, Avenida España 1680, Valparaíso, Chile
e-mail: juan.peypouquet@usm.cl

Our method is based on an explicit discretization of the *multiscale asymptotic gradient* (MAG) differential inclusion introduced in [1]. The idea behind the exterior penalization approach is to add a high cost to constraint violation, forcing the trajectory towards the feasible set.

It is worth mentioning that in [2] the authors consider implicit discretizations of the MAG, which also produce solutions for the problem. This approach is closely related to the pioneer work [3] and also to [4, 5], and [6]. The fundamental advantage of the purely explicit scheme presented in this work is the simplicity in the computation of each iteration. In implicit schemes, each iteration has a considerably higher computational cost because one typically has to solve a nonlinear equation. Moreover, proximal schemes do not show better convergence properties when the functions involved are regular. In fact, it sometimes works in the opposite way (see [7]). Another advantage of gradient-type methods is the availability of different rules for the selection of the step sizes, which can accelerate the convergence. A forward–backward method is studied in [8], which is explicit with respect to the penalization and implicit with respect to the objective function.

The paper is organized as follows: Sect. 2 contains the description of the algorithm and the convergence results. Most technical aspects are gathered in Sect. 2.2. In Sect. 3, we discuss several implementation issues, namely: stability and robustness results which account for inexact computation of the iterates, step size selection and verification of the hypotheses. Finally, Sect. 4 contains some examples and applications in mathematical programming, best approximation, partial differential equations, and signal processing. We also provide a numerical illustration.

2 The Algorithm and Its Asymptotic Analysis

2.1 Preliminaries, Hypotheses and Main Result

Let H be a real Hilbert space with the norm and inner product given by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, respectively. Let Φ and Ψ be proper convex functions on H and assume for simplicity (see Sect. 3.3) that both are *everywhere* defined and differentiable. We consider the problem of finding a point in the set

$$\mathcal{S} := \operatorname{argmin}\{\Phi(x) : x \in \operatorname{argmin}(\Psi)\}$$

assuming $\mathcal{S}$, and thus $\operatorname{argmin}(\Psi)$, is nonempty. On the one hand, $\mathcal{S}$ can be interpreted as the set of solutions of a hierarchical optimization problem, where Ψ and Φ are primary and secondary criteria, respectively. On the other hand, any (convex and regular) constrained optimization problem of the form $\min\{\Phi(x) : x \in C\}$ can be expressed in this context by choosing, for instance, Ψ as the square of the distance to the set C. Another simple example is when $C := \{x \in H : g(x) \le 0\}$, where g is a differentiable convex function. In this case, one can take Ψ as the square of the positive part of g (see Sect. 4 for further details). In what follows, we write $C := \operatorname{argmin}(\Psi)$.

In order to approximate points in $\mathcal{S}$, we propose a *diagonal gradient scheme* (DGS) which generates a sequence in H by coupling the gradient method with an exterior penalization procedure with respect to Ψ, namely:

$$\text{(DGS)} \qquad \begin{cases} x^1 \in H, \\ x^{n+1} = x^n - \lambda_n \nabla \Omega_n(x^n), \quad \text{for } n \geq 1. \end{cases}$$

Here the *penalized function* Ω_n is given by $\Omega_n := \Phi + \beta_n \Psi$. The *step size* λ_n and the *penalization parameter* β_n are positive numbers. Throughout the paper, we assume that the gradients $\nabla\Phi$ and $\nabla\Psi$ are Lipschitz-continuous with constants L_Φ and L_Ψ, respectively. Therefore, $\nabla\Omega_n$ is Lipschitz-continuous with constant $L_n := L_\Phi + \beta_n L_\Psi$. This is a standard assumption for the convergence of gradient-type systems (see [9, Sect. 1.2]). We shall also assume, without any loss of generality, that $\min \Psi = 0$.

For the classical notation of Convex Analysis, see [10]. In particular, the *Fenchel conjugate* of Ψ is $\Psi^*(x^*) := \sup_{y\in H}\{\langle x^*, y\rangle - \Psi(y)\}$; the *support function* of C at x^* is $\sigma_C(x^*) := \sup_{c\in C}\langle x^*, c\rangle$; and the *normal cone* to C at x is $N_C(x) := \{x^* \in H : \langle x^*, c - x\rangle \leq 0 \text{ for all } c \in C\}$ if $x \in C$ and $\emptyset$ otherwise. We denote by $R(N_C)$ the range of the operator N_C. Consider the following hypotheses:

H₁: There exist $K, \delta > 0$ such that $\beta_{n+1} - \beta_n \leq K\lambda_{n+1}\beta_{n+1}$ and $\frac{L_n}{2} - \frac{1}{\lambda_n} \leq -(K + \delta)$ for all $n \geq 1$.
H₂: $\sum_{n\geq 1} \lambda_n\beta_n[\Psi^*(\frac{2p}{\beta_n}) - \sigma_C(\frac{2p}{\beta_n})] < \infty$ for all $p \in R(N_C)$.
H₃: $\liminf_{n\to\infty} \lambda_n\beta_n > 0$ and $\sum_{n\geq 1} \lambda_n = \infty$.

Hypotheses $\mathbf{H_1}$ and $\mathbf{H_3}$ essentially refer to the relationship between the growth of (β_n) and the decay of (λ_n). Hypothesis $\mathbf{H_2}$ relates the parameter sequences to the shape of the function Ψ near the boundary of C. A more thorough discussion on the verification of these hypotheses, along with examples, is given in Sect. 3. The main result of this paper is the following:

Theorem 2.1
Assume Hypotheses $\mathbf{H_1}$–$\mathbf{H_3}$ *hold and let* (x^n) *satisfy* (DGS). *Then* (x^n) *converges weakly in* H *as* $n \to \infty$ *to some* $x^* \in \mathcal{S}$. *If, moreover,* Φ *is strongly convex, then* (x^n) *converges strongly in* H *as* $n \to \infty$ *to the unique* $x^* \in \mathcal{S}$.

2.2 Convergence

Let us denote by (x^n) an arbitrary sequence verifying (DGS) and provide some estimations.

Lemma 2.1 *Let* $\bar{x} \in \mathcal{S}$ *and set* $\bar{p} := -\nabla\Phi(\bar{x})$. *For each* $n \geq 1$, *we have*

$$\begin{aligned} &\|x^{n+1} - \bar{x}\|^2 - \|x^n - \bar{x}\|^2 + \lambda_n\beta_n\Psi(x^n) \\ &\quad \leq \|x^{n+1} - x^n\|^2 + \lambda_n\beta_n\left[\Psi^*\left(\frac{2\bar{p}}{\beta_n}\right) - \sigma_C\left(\frac{2\bar{p}}{\beta_n}\right)\right]. \end{aligned} \qquad (1)$$

Proof First, observe that $\bar{x} \in \mathcal{S}$ implies $0 \in \nabla\Phi(\bar{x}) + N_C(\bar{x})$ and so $\bar{p} \in N_C(\bar{x})$. Since

$$\frac{x^{n+1} - x^n}{\lambda_n} + \beta_n \nabla\Psi(x^n) = -\nabla\Phi(x^n),$$

the monotonicity of $\nabla\Phi$ gives

$$\left\langle \frac{x^n - x^{n+1}}{\lambda_n} - \beta_n \nabla\Psi(x^n) + \bar{p}, x^n - \bar{x} \right\rangle \geq 0, \tag{2}$$

and therefore,

$$2\langle x^n - x^{n+1}, x^n - \bar{x}\rangle \geq 2\lambda_n\beta_n\langle\nabla\Psi(x^n), x^n - \bar{x}\rangle + 2\lambda_n\langle\bar{p}, \bar{x} - x^n\rangle. \tag{3}$$

But the convexity of Ψ implies

$$0 = \Psi(\bar{x}) \geq \Psi(x^n) + \langle\nabla\Psi(x^n), \bar{x} - x^n\rangle, \tag{4}$$

whence

$$2\lambda_n\beta_n\langle\nabla\Psi(x^n), x^n - \bar{x}\rangle \geq 2\lambda_n\beta_n\Psi(x^n). \tag{5}$$

On the other hand, recall that

$$2\langle x^n - x^{n+1}, x^n - \bar{x}\rangle = \|x^{n+1} - x^n\|^2 + \|x^n - \bar{x}\|^2 - \|x^{n+1} - \bar{x}\|^2. \tag{6}$$

Combining (3), (5), and (6), we obtain

$$\|x^{n+1} - x^n\|^2 + \|x^n - \bar{x}\|^2 - \|x^{n+1} - \bar{x}\|^2 \geq 2\lambda_n\langle\bar{p}, \bar{x} - x^n\rangle + 2\lambda_n\beta_n\Psi(x^n),$$

which we rewrite as

$$\begin{aligned}
&\|x^{n+1} - \bar{x}\|^2 - \|x^n - \bar{x}\|^2 + \lambda_n\beta_n\Psi(x^n) \\
&\quad\leq \|x^{n+1} - x^n\|^2 + 2\lambda_n\langle\bar{p}, x^n\rangle - \lambda_n\beta_n\Psi(x^n) - 2\lambda_n\langle\bar{p}, \bar{x}\rangle.
\end{aligned}$$

Finally, observe that $\bar{p} \in N_C(\bar{x})$ if and only if $\sigma_C(\bar{p}) = \langle\bar{p}, \bar{x}\rangle$. Whence

$$\begin{aligned}
&2\lambda_n\langle\bar{p}, x^n\rangle - \lambda_n\beta_n\Psi(x^n) - 2\lambda_n\langle\bar{p}, \bar{x}\rangle \\
&\quad= \lambda_n\beta_n\left[\left\langle\frac{2\bar{p}}{\beta_n}, x^n\right\rangle - \Psi(x^n) - \left\langle\frac{2\bar{p}}{\beta_n}, \bar{x}\right\rangle\right] \\
&\quad\leq \lambda_n\beta_n\left[\Psi^*\left(\frac{2\bar{p}}{\beta_n}\right) - \left\langle\frac{2\bar{p}}{\beta_n}, \bar{x}\right\rangle\right] \\
&\quad= \lambda_n\beta_n\left[\Psi^*\left(\frac{2\bar{p}}{\beta_n}\right) - \sigma_C\left(\frac{2\bar{p}}{\beta_n}\right)\right],
\end{aligned}$$

which yields (1). □

If Φ is strongly convex, then the same argument leads to the following stronger estimation:

Lemma 2.2 *Let Φ be strongly convex with parameter $\alpha > 0$. Take $\bar{x} \in \mathcal{S}$ and set $\bar{p} := -\nabla\Phi(\bar{x})$. For each $n \geq 1$, we have*

$$\left\|x^{n+1} - \bar{x}\right\|^2 - \|x^n - \bar{x}\|^2 + \lambda_n\beta_n\Psi(x^n) + \alpha\lambda_n\|x^n - \bar{x}\|^2 \\ \leq \left\|x^{n+1} - x^n\right\|^2 + \lambda_n\beta_n\left[\Psi^*\left(\frac{2\bar{p}}{\beta_n}\right) - \sigma_C\left(\frac{2\bar{p}}{\beta_n}\right)\right].$$

Proof The strong monotonicity of $\nabla\Phi$ implies that inequality (2) can be reinforced to

$$\left\langle \frac{x^n - x^{n+1}}{\lambda_n} - \beta_n\nabla\Psi(x^n) + \bar{p}, x^n - \bar{x}\right\rangle \geq \alpha\|x^n - \bar{x}\|^2.$$

This explains the additional term $\alpha\lambda_n\|x^n - \bar{x}\|^2$ on the left-hand side of the inequality. □

We now turn our attention to the values of the penalized function $\Omega_n = \Phi + \beta_n\Psi$, whose gradient is Lipschitz-continuous with constant $L_n = L_\Phi + \beta_n L_\Psi$. Observe that from [9, Proposition A.24] we deduce that

$$\Omega_n(y) \leq \Omega_n(x) + \langle\nabla\Omega_n(x), y - x\rangle + \frac{L_n}{2}\|x - y\|^2 \quad \text{for all } x, y \text{ in } H. \tag{7}$$

Lemma 2.3 *For each $n \geq 1$, we have*

$$\Omega_{n+1}\left(x^{n+1}\right) - \Omega_n(x^n) \leq (\beta_{n+1} - \beta_n)\Psi\left(x^{n+1}\right) + \left[\frac{L_n}{2} - \frac{1}{\lambda_n}\right]\left\|x^{n+1} - x^n\right\|^2.$$

Proof Recall that $-\frac{x^{n+1}-x^n}{\lambda_n} = \nabla\Omega_n(x^n)$. By inequality (7), we have

$$\Phi\left(x^{n+1}\right) + \beta_n\Psi\left(x^{n+1}\right) \\ \leq \Phi(x^n) + \beta_n\Psi(x^n) - \left\langle\frac{x^{n+1} - x^n}{\lambda_n}, x^{n+1} - x^n\right\rangle + \frac{L_n}{2}\left\|x^{n+1} - x^n\right\|^2.$$

We conclude by adding $\beta_{n+1}\Psi(x^{n+1})$ to both sides and rearranging the terms. □

For $\bar{x} \in \mathcal{S}$ write

$$\xi_n := \Phi(x^n) + (1 - K\lambda_n)\beta_n\Psi(x^n) + K\|x^n - \bar{x}\|^2 \\ = \Omega_n(x^n) - K\lambda_n\beta_n\Psi(x^n) + K\|x^n - \bar{x}\|^2.$$

Corollary 2.1 *Let $\bar{x} \in \mathcal{S}$, set $\bar{p} = -\nabla\Phi(\bar{x})$ and assume Hypothesis $\mathbf{H_1}$ holds. Then for each $n \geq 1$ we have*

$$\xi_{n+1} - \xi_n + \delta\left\|x^{n+1} - x^n\right\|^2 \leq K\lambda_n\beta_n\left[\Psi^*\left(\frac{2\bar{p}}{\beta_n}\right) - \sigma_C\left(\frac{2\bar{p}}{\beta_n}\right)\right].$$

Proof Hypothesis $\mathbf{H_1}$ and Lemma 2.3 together imply

$$\Omega_{n+1}\big(x^{n+1}\big) - \Omega_n(x^n) \le K\lambda_{n+1}\beta_{n+1}\Psi\big(x^{n+1}\big) - (K+\delta)\big\|x^{n+1} - x^n\big\|^2.$$

Now multiply inequality (1) by K to obtain

$$\begin{aligned} &K\big\|x^{n+1} - \bar{x}\big\|^2 - K\|x^n - \bar{x}\|^2 + K\lambda_n\beta_n\Psi(x^n) \\ &\quad \le K\big\|x^{n+1} - x^n\big\|^2 + K\lambda_n\beta_n\left[\Psi^*\left(\frac{2\bar{p}}{\beta_n}\right) - \sigma_C\left(\frac{2\bar{p}}{\beta_n}\right)\right]. \end{aligned}$$

The result follows upon adding the last two inequalities. □

We shall use the following elementary fact concerning the convergence of real sequences. A proof can be found, for instance, in [11, Lemma 3.1] or [8, Lemma 2].

Lemma 2.4 *Let (ζ_n) be bounded from below and let (δ_n) be non-negative. Assume*

$$\zeta_{n+1} - \zeta_n + \delta_n \le \varepsilon_n$$

for all $n \ge 1$ and $\sum_{n\ge1}\varepsilon_n < \infty$. Then $\lim_{n\to\infty}\zeta_n$ exists and $\sum_{n\ge1}\delta_n < \infty$.

Proposition 2.1 *Let $\bar{x} \in \mathcal{S}$ and let Hypotheses $\mathbf{H_1}$ and $\mathbf{H_2}$ hold. Then*

- (i) $\lim_{n\to\infty}\xi_n$ *exists and* $\sum_{n\ge1}\|x^{n+1} - x^n\|^2 < \infty$.
- (ii) $\lim_{n\to\infty}\|x^n - \bar{x}\|$ *exists and* $\sum_{n\ge1}\lambda_n\beta_n\Psi(x^n) < \infty$.
- (iii) $\lim_{n\to\infty}\Omega_n(x^n)$ *exists.*
- (iv) *If, moreover,* $\liminf_{n\to\infty}\lambda_n\beta_n > 0$ *then* $\lim_{n\to\infty}\Psi(x^n) = 0$ *and every weak cluster point of the sequence* (x^n) *lies in* C.

Proof For (i) set $\zeta_n = \xi_n$, $\delta_n = \delta\|x^{n+1} - x^n\|^2$ and $\varepsilon_n = K\lambda_n\beta_n[\Psi^*(\frac{2\bar{p}}{\beta_n}) - \sigma_C(\frac{2\bar{p}}{\beta_n})]$, where $\bar{p} = -\nabla\Phi(\bar{x})$. The second inequality in Hypothesis $\mathbf{H_1}$ implies $1 - K\lambda_n > 0$. This fact and the convexity of Φ yield

$$\begin{aligned} \xi_n &\ge \Phi(x^n) + K\|x^n - \bar{x}\|^2 \\ &\ge \Phi(\bar{x}) + \langle\nabla\Phi(\bar{x}), x^n - \bar{x}\rangle + K\|x^n - \bar{x}\|^2 \\ &\ge \Phi(\bar{x}) - \|\bar{p}\|\,\|x^n - \bar{x}\| + K\|x^n - \bar{x}\|^2 \\ &\ge \Phi(\bar{x}) - \frac{\|\bar{p}\|^2}{4K}, \end{aligned}$$

and so the sequence (ξ_n) is bounded from below. Since $\bar{p} \in N_C(\bar{x})$, Hypothesis $\mathbf{H_2}$ implies $\sum_{n\ge1}\varepsilon_n < \infty$. Corollary 2.1 and Lemma 2.4 then give the result.

For (ii) set $\zeta_n = \|x^n - \bar{x}\|^2$, $\delta_n = \lambda_n\beta_n\Psi(x^n)$, $\varepsilon_n = \|x^{n+1} - x^n\|^2 + \lambda_n\beta_n[\Psi^*(\frac{2\bar{p}}{\beta_n}) - \sigma_C(\frac{2\bar{p}}{\beta_n})]$ and use inequality (1) along with part (i) and Lemma 2.4.

For (iii) just notice that

$$\Omega_n(x^n) = \xi_n + K\lambda_n\beta_n\Psi(x^n) - K\|x^n - \bar{x}\|^2$$

and use parts (i) and (ii).

Finally, (iv) follows immediately from (ii). □

Proposition 2.2 *Let $\bar{x} \in \mathcal{S}$ and assume Hypotheses* $\mathbf{H_1}$ *and* $\mathbf{H_2}$ *hold. Then*

$$\sum_{n\geq 1} \lambda_n [\Omega_n(x^n) - \Phi(\bar{x})] < +\infty \quad (\textit{possibly } -\infty).$$

Proof The convexity of Φ gives

$$\Phi(\bar{x}) \geq \Phi(x^n) + \langle \nabla\Phi(x^n), \bar{x} - x^n \rangle.$$

This inequality and (4) together give

$$\Phi(\bar{x}) - \Omega_n(x^n) \geq \langle \nabla\Omega_n(x^n), \bar{x} - x^n \rangle = \left\langle \frac{x^n - x^{n+1}}{\lambda_n}, \bar{x} - x^n \right\rangle.$$

Using (6), we deduce that

$$2\lambda_n [\Omega_n(x^n) - \Phi(\bar{x})] \leq \left\| x^n - x^{n+1} \right\|^2 + \|x^n - \bar{x}\|^2 - \left\| x^{n+1} - \bar{x} \right\|^2,$$

and so

$$2\sum_{n\geq 1} \lambda_n [\Omega_n(x^n) - \Phi(\bar{x})] \leq \left\| x^1 - \bar{x} \right\|^2 + \sum_{n\geq 1} \left\| x^{n+1} - x^n \right\|^2 < +\infty$$

as required. □

Now we are in position to prove the main result:

Proof of Theorem 2.1. By Opial's Lemma [12] and part (ii) of Proposition 2.1, it suffices to prove that every weak cluster point of the sequence $\{x_n\}$ lies in $\mathcal{S}$. Let (x^{k_n}) converge weakly to x^∞ as $n \to \infty$. But $\sum_{n\geq 1} \lambda_n = \infty$ by the second statement in Hypothesis $\mathbf{H_3}$. Therefore, part (iii) in Proposition 2.1 and Proposition 2.2 together imply $\lim_{n\to\infty} \Omega_n(x^n) \leq \Phi(\bar{x})$ for every $\bar{x} \in \mathcal{S}$. In view of the weak lower-semicontinuity of Φ, we have

$$\Phi(x^\infty) \leq \liminf_{n\to\infty} \Phi(x^{k_n}) \leq \lim_{n\to\infty} \Omega_n(x^n) \leq \Phi(\bar{x})$$

for every $\bar{x} \in \mathcal{S}$. But x^∞ must belong to C by the first statement in Hypothesis $\mathbf{H_3}$ and part (iv) in Proposition 2.1. This implies $x^\infty \in \mathcal{S}$ and proves the weak convergence. For the strong convergence in the strongly convex case, we use Lemma 2.2, observing that the right-hand side of the inequality is summable by Hypothesis $\mathbf{H_2}$ and part (i) in Proposition 2.1. Lemma 2.4 then implies that $\sum_{n\geq 1} \lambda_n \|x^n - \bar{x}\|^2 < \infty$, whence $\liminf_{n\to\infty} \|x^n - \bar{x}\| = 0$. Since $\lim_{n\to\infty} \|x^n - \bar{x}\|$ exists, the sequence (x^n) must converge strongly to $\bar{x}$. □

3 Implementation Issues

In this section, we discuss some ideas leading to the practical use of this method. First, we present some stability and robustness properties, which in particular allow computing the iterates inexactly. Next we comment on the selection of the parameter sequences (λ_n) and (β_n) in order to satisfy the hypotheses of Theorem 2.1. We also describe a complementary heuristic for the step size selection. Finally, we mention some facts about the differentiability assumptions that explain how they can be weakened.

3.1 Stability and Robustness

We now derive some stability properties of the algorithm described in the preceding sections with respect to perturbations of the initial data.

For $n \geq 1$ and $x \in H$, write $P_n(x) = x - \lambda_n \nabla \Omega_n(x)$ so that the sequences generated by (DGS) verify $x^{n+1} = P_n(x^n)$.

Lemma 3.1 *Assume $\lambda_n L_n \leq 2$. Then the function P_n is nonexpansive.*

Proof For $x, y \in H$, we have

$$\begin{aligned}\|P_n(x) - P_n(y)\|^2 &= \|(x-y) - \lambda_n(\nabla\Omega_n(x) - \nabla\Omega_n(y))\|^2 \\ &= \|x-y\|^2 + \lambda_n^2\|\nabla\Omega_n(x) - \nabla\Omega_n(y)\|^2 \\ &\quad - 2\lambda_n\langle x-y, \nabla\Omega_n(x) - \nabla\Omega_n(y)\rangle.\end{aligned}$$

Since $\nabla\Omega_n$ is L_n-Lipschitz, we deduce from [13, Corollary 10] that

$$\langle x-y, \nabla\Omega_n(x) - \nabla\Omega_n(y)\rangle \geq \frac{1}{L_n}\|\nabla\Omega_n(x) - \nabla\Omega_n(y)\|^2.$$

Whence

$$\|P_n(x) - P_n(y)\|^2 \leq \|x-y\|^2 + \lambda_n\left[\lambda_n - \frac{2}{L_n}\right]\|\nabla\Omega_n(x) - \nabla\Omega_n(y)\|^2 \leq \|x-y\|^2,$$

and so P_n is nonexpansive. □

Observe that the second inequality in Hypothesis $\mathbf{H_1}$ implies $\lambda_n L_n \leq 2$. We have the following result concerning the stability of the sequence and its weak limits:

Proposition 3.1 *Let (x_1^n) and (x_2^n) satisfy (DGS) starting from x_1^1 and x_2^1, respectively.*

(i) *If $\lambda_n L_n \leq 2$ for all $n \geq 1$ then $\|x_1^n - x_2^n\| \leq \|x_1^1 - x_2^1\|$ for all $n \geq 1$.*
(ii) *If, moreover, $x_1^n \rightharpoonup x_1^\infty$ and $x_2^n \rightharpoonup x_2^\infty$ as $n \to \infty$ then $\|x_1^\infty - x_2^\infty\| \leq \|x_1^1 - x_2^1\|$.*

Finally, we prove that convergence can still be granted if the sequence is computed approximately with sufficiently small errors.

Proposition 3.2 *Let Hypotheses* $\mathbf{H_1}$–$\mathbf{H_3}$ *hold and assume* (x_n) *satisfies*

$$\left\|x^{n+1} - P_n(x^n)\right\| \le \varepsilon_n$$

for all $n \ge 1$. *The conclusions of Theorem 2.1 remain valid provided* $\sum_{n\ge 1} \varepsilon_n < \infty$.

Proof It suffices to apply [14, Proposition 6.2] in view of Lemma 3.1. □

3.2 The Hypotheses and Their Verification

One simple way to build sequences (λ_n) and (β_n) satisfying Hypotheses $\mathbf{H_1}$ and $\mathbf{H_3}$ is the following: Take any $K > 0$, $\delta > 0$, $0 < \gamma < \frac{2}{L_\Psi}$ and $0 < q \le 1$. Then set

$$\beta_n = \frac{\gamma[L_\Phi + 2(K + \delta)]}{2 - \gamma L_\Psi} + \gamma K n^q \quad \text{and} \quad \lambda_n = \frac{\gamma}{\beta_n}. \tag{8}$$

Then, clearly, $\beta_{n+1} - \beta_n = \gamma K[(n+1)^q - n^q] \le \gamma K = K\lambda_{n+1}\beta_{n+1}$. This is the first inequality in Hypothesis $\mathbf{H_1}$. On the other hand, for all $n \ge 1$ one has

$$\beta_n \ge \frac{\gamma[L_\Phi + 2(K + \delta)]}{2 - \gamma L_\Psi}.$$

Since $2 - \gamma L_\Psi > 0$, we have $2\beta_n - \gamma\beta_n L_\Psi \ge \gamma[L_\Phi + 2(K + \delta)]$, and so $2\beta_n \ge \gamma[L_n + 2(K + \delta)]$. Dividing by 2γ and rearranging the terms, we obtain the second inequality in Hypothesis $\mathbf{H_1}$. Hypothesis $\mathbf{H_3}$ is straightforward since $0 < q \le 1$.

The following heuristic—based on the *exact minimization rule*[1] (see [9, Sect. 1.2])—can complement (8) as a criterion for the selection of the parameters. Recall that $\Omega_n(x) = \Phi(x) + \beta_n\Psi(x)$ and write $T_n = \nabla\Omega_n$ so that $x^{n+1} = x^n - \lambda_n T_n(x^n)$. For $\lambda > 0$, write $\theta_n(\lambda) := \Omega_n(x^n - \lambda T_n(x^n))$ and let λ_n be any minimizer of θ_n, provided such minimizers exist (for instance, if Ω_n is coercive). Since θ_n is differentiable one must have $\theta_n'(\lambda_n) = 0$. In other words, λ_n solves

$$\langle T_n(x^n - \lambda_n T_n(x^n)), T_n(x^n)\rangle = 0.$$

If T_n is replaced by a linear approximation $\widetilde{T}_n$ near x^n it seems reasonable to choose

$$\lambda_n = \frac{\|\widetilde{T}_n(x^n)\|^2}{\langle \widetilde{T}_n^2(x^n), \widetilde{T}_n(x^n)\rangle}.$$

Regarding Hypothesis $\mathbf{H_2}$, we begin by pointing out that it is the discrete version of Hypothesis (H_1) in [1] and was already introduced in [2]. Next, observe that all the terms in the sum are nonnegative. Indeed, since Ψ is bounded from above by the indicator function of the set C, the reverse inequality holds for their Fenchel conjugates, whence $\Psi^*(p) - \sigma_C(p) \ge 0$ for all $p \in H$. On the other hand, if Ψ has quadratic growth, Hypothesis $\mathbf{H_2}$ can be granted under a very simple assumption on

[1] Other alternatives are the *limited minimization rule*, the *Armijo rule*, and the *Goldstein rule*.

the parameters. More precisely, suppose that $\Psi(\cdot) \geq \frac{a}{2}\mathrm{dist}(\cdot, C)^2$ for some $a > 0$.[2] Then $\Psi^*(p) - \sigma_C(p) \leq \frac{1}{2a}\|p\|^2$ for all $p \in R(N_C)$. In that case,

$$\lambda_n \beta_n \left[\Psi^* \left(\frac{2p}{\beta_n} \right) - \sigma_C \left(\frac{2p}{\beta_n} \right) \right] \leq \frac{2\lambda_n \|p\|^2}{a\beta_n},$$

and so the summability of the sequence $(\frac{\lambda_n}{\beta_n})$ is sufficient for $\mathbf{H_2}$. Notice that it is also necessary if $\Psi(\cdot) = \frac{a}{2}\mathrm{dist}(\cdot, C)^2$. Further, observe that if $\liminf_{n\to\infty} \lambda_n \beta_n > 0$—which holds under Hypothesis $\mathbf{H_3}$—then the summability of (λ_n^2) is sufficient for the summability of $(\frac{\lambda_n}{\beta_n})$.

3.3 The Regularity of Φ and Ψ

We shall comment briefly on two remarks concerning the Lipschitz-continuity assumption on the gradients of Φ and Ψ.

Global to Local

One realizes a posteriori that a local Lipschitz-continuity assumption on the gradients is sufficient for the convergence of the method. In practice, the problem is that the parameter sequences (λ_n) and (β_n) depend on the Lipschitz constants. In particular instances it would be possible to use local Lipschitz constants on appropriate sublevel sets.

Restricted Domain

The functions Φ and Ψ need only be defined and regular on a convex domain $D \subset H$, provided the sequence (x^n) is well-defined in the sense that $x^n - \lambda_n \nabla \Omega_n(x^n) \in D$ for all $n \geq 1$. A more careful selection of the step sizes may be necessary. This seems an interesting line for future research.

4 Examples

In this section, we describe several simple instances where this method can be applied. They appear in different contexts in science and engineering problems, such as optimal control of linear systems, mathematical programming, domain decomposition methods for PDEs, transport, imaging and signal processing (see [2, 8, 16] or [17]), among others. As an illustration, we provide a numerical example in signal reconstruction from partial information.

[2] This holds, for instance, if $C = \{x \in H : Ax = b\}$ and $\Psi(x) = \|Ax - b\|_Z^2$, where $A : H \to Z$ is a bounded linear operator whose range is closed in Z (see, for example, [15, Paragraph II.7]).

4.1 Relaxed Feasibility

The convex *feasibility problem* consists in finding a point in the intersection of nonempty closed convex sets $C_1, \dots, C_M$. This can be expressed as

$$\text{(F)} \qquad \min \sum_{m=1}^{M} \delta_{C_m}(x),$$

where δ_{C_m} denotes the indicator function of C_m. Due to possible inaccuracy in the description of the sets $C_1, \dots, C_M$, the intersection may be empty, and so problem (F) may not have a solution. A *relaxed* form is

$$\min \Psi(x), \quad \text{where} \quad \Psi(x) = \frac{1}{2}\sum_{m=1}^{M} w_m \operatorname{dist}(x, C_m)^2$$
$$\text{with} \quad w_m > 0 \text{ for } m = 1, \dots, M.$$

This gives exact solutions of (F), if there are any and approximate solutions otherwise. Observe that $\operatorname{dist}(x, C_m) = \|x - P_m(x)\|$, where P_m denotes the projection operator onto C_m. Since

$$\|x + h - P_m(x + h)\| \le \|x + h - P_m(x)\|,$$

a simple computation shows that

$$\operatorname{dist}(x + h, C_m)^2 - \operatorname{dist}(x, C_m)^2 - 2\langle x - P_m(x), h\rangle \le \|h\|^2$$

for each m. Whence Ψ is differentiable and

$$\nabla\Psi(x) = \sum_{m=1}^{M} w_m\big(x - P_m(x)\big).$$

The function Φ can be incorporated as a criterion for selecting particular feasible points.

4.2 Convex Inequality Constraints

Consider the mathematical programming problem

$$\min\{\Phi(x) : x \in C\}, \quad \text{where} \quad C = \big\{x \in \mathbf{R}^N :\ g_j(x) \le 0,\ \text{for } j = 1, \dots, J\big\},$$

where Φ and the g_j's are proper differentiable convex functions on H. Let $[r]_+$ denote the positive part of $r \in \mathbf{R}$. Take $\Psi(x) = \frac{1}{2}\sum_{j=1}^{J}[g_j(x)]_+^2$ so that $C = \operatorname{argmin}(\Psi)$. If each g_j is differentiable, then so is Ψ and

$$\nabla\Psi(x) = \sum_{j=1}^{J}[g_j(x)]_+ \nabla g_j(x).$$

4.3 Realizing the Distance Between Two Closed Affine Subspaces

For $i=1,2$, consider a point b_i in a Hilbert space Y_i, a bounded linear operator $A_i : H \to Y_i$ and set $P_i = \{x \in H : A_i x = b_i\}$. The distance between P_1 and P_2 can be expressed as

$$\min\{\Phi(x_1,x_2) : (x_1,x_2) \in \operatorname{argmin}(\Psi)\},$$

where $\Phi(x_1,x_2) = \frac{1}{2}\|x_1 - x_2\|^2$ and $\Psi(x_1,x_2) = \frac{1}{2}\|A_1x_1 - b_1\|^2 + \frac{1}{2}\|A_2x_2 - b_2\|^2$. Here

$$\begin{cases} x_1^{n+1} = (1-\lambda_n)x_1^n + \lambda_n x_2^n - \lambda_n\beta_n A_1^*(A_1x_1^n - b_1), \\ x_2^{n+1} = \lambda_n x_1^n + (1-\lambda_n)x_2^n - \lambda_n\beta_n A_2^*(A_2x_2^n - b_2). \end{cases}$$

Observe that this can be seen as a two-step iteration where one first computes a barycenter of x_1^n and x_2^n and then performs a steepest descent step with respect to Ψ.

4.4 Structured Optimization with Coupling

Consider the minimization problem

$$\min\{F_1(x_1) + Q(x_1,x_2) + F_2(x_2) : A_1x_1 = A_2x_2,\ (x_1,x_2) \in H_1 \times H_2\}, \tag{9}$$

where H_1, H_2, and Z are real Hilbert spaces, each A_i is a bounded linear (or affine) operator from H_i to Z, each F_i is a proper differentiable convex function on H_i and Q is a positive semidefinite quadratic function. Here $\Phi(x_1,x_2) = F_1(x_1) + Q(x_1,x_2) + F_2(x_2)$ and $\Psi(x_1,x_2) = \|A_1x_1 - A_2x_2\|^2$. In this case,

$$\begin{cases} x_1^{n+1} = x_1^n + \lambda_n \nabla F_1(x_1^n) + \lambda_n \nabla_{x_1} Q(x_1^n, x_2^n) - \lambda_n\beta_n A_1^*(A_1x_1^n - A_2x_2^n), \\ x_2^{n+1} = x_2^n + \lambda_n \nabla F_2(x_2^n) + \lambda_n \nabla_{x_2} Q(x_1^n, x_2^n) - \lambda_n\beta_n A_2^*(A_2x_2^n - A_1x_1^n). \end{cases}$$

Proximal-type algorithms often require computations of resolvents of sums for these kinds of problems (see [16] or [18]). Exceptions are the predictor–corrector methods, as studied in [19].

4.5 Stokes Equation

The following formulation has been taken from [20, Chapter IV, Section 2.5] and pointed out by F. Álvarez. Let Ω be a bounded domain in $\mathbf{R}^d$ and let $f \in L^2(\Omega;\mathbf{R}^d)$. Consider the problem of finding a velocity $u \in H_0^1(\Omega;\mathbf{R}^d)$ and a pressure $p \in L^2(\Omega;\mathbf{R})$ such that

$$\text{(S)} \qquad \begin{cases} -\Delta u + \nabla p = f & \text{on } \Omega, \\ \operatorname{div}(u) = 0 & \text{on } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

We shall express (S) as a variational problem in the product space framework described in Sect. 4.4 with $H_1 = H_0^1(\Omega;\mathbf{R}^d)$[3], $H_2 = Z = L^2(\Omega;\mathbf{R})$, $A_1u = \operatorname{div}(u)$,

[3] We can use $\|v\|_{H_0^1} = \|\nabla v\|_{L^2}$, by virtue of Poincaré Inequality.

and $A_2 \equiv 0$. However, we shall see that the problem can be completely decoupled and expressed as two simpler problems, one on each factor space. First, define $F_1(u) = \frac{1}{2}\|\nabla u\|_{L^2}^2 - \langle f, u\rangle_{L^2}$ and consider the problem

$$\text{(P)} \qquad \min\{F_1(u) : u \in H_1 \text{ and } \operatorname{div}(u) = 0\} = \min\{F_1(u) + \delta_{\{0\}}(A_1 u) : u \in H_1\}.$$

If we define the Lagrangian function

$$L(u, p) = F_1(u) + \langle p, A_1 u\rangle_{L^2},$$

then the dual of (P) in the sense of Fenchel–Rockafellar is

$$\text{(D)} \qquad \min\{F_1^*(-A_1^* p) + \delta_{\{0\}}^*(p) : p \in L^2\} = \min\{F_1^*(-A_1^* p) : p \in H_2\}.$$

Here the pressure p can be interpreted as a Lagrange multiplier for the incompressibility condition $\operatorname{div}(u) = 0$. Observe that

$$F_1^*(-A_1^* p) = \sup_{v \in H_1} \{\langle -A_1^* p, v\rangle_{H_1^*, H_1} - F_1(v)\}. \tag{10}$$

For each $p \in H_2$, the optimization problem above has a unique solution v_p, which is also the unique function in H_1 satisfying $-\Delta v_p = f + \nabla p$ in the sense of distributions. Moreover,

$$F_1^*(-A_1^* p) = \frac{1}{2}\|\nabla v_p\|_{L^2}^2.$$

The reader can verify that (P) and (D) have solutions. Observe that, if u^* is a solution of (P) and p^* is a solution of (D), then $\operatorname{div}(u^*) = 0$ and

$$F_1(u^*) + F_1^*(-A_1^* p^*) = \langle -A_1^* p^*, u^*\rangle = -\langle p^*, A_1 u^*\rangle = 0.$$

This also implies that $u^* = v_{p^*}$ by uniqueness of solution of the optimization problem in (10). Whence $-\Delta u^* + \nabla(-p^*) = f$ in the sense of distributions, and so the pair $(u^*, -p^*)$ is a weak solution for (S). Moreover, (u^*, p^*) is a saddle point of L (see [21, Sect. 8.4.4]). Setting $F_2(p) = \frac{1}{2}\|v_p\|_{H_1}^2$, $\Phi(u, p) = F_1(u) + F_2(p)$, and $\Psi(u, p) = \frac{1}{2}\|\operatorname{div}(u)\|_{L^2}^2$, the solutions of

$$\min\{\Phi(u, p) : (u, p) \in \operatorname{argmin}(\Psi)\}$$

are weak solutions for Stokes equation (S) and can be approximated using our (DGS). The complete decoupling makes this equivalent to solving one problem on each space H_1 and H_2. Observe that F_1 is strongly convex but F_2 is not. Whence the velocities converge strongly.

4.6 Signal Reconstruction

Let $H = L^2(\Omega; \mathbf{R})$, where Ω is a bounded domain in $\mathbf{R}^N$. A signal $x \in H$ is to be reconstructed from partial information given by a set of observations and a priori information on the signal itself. This is related to the stable signal recovery problem (see [22]). As an example, suppose that the support of the signal is known to be contained in some set $\Omega_0 \subset \Omega$ (a priori information) and a finite number of its Fourier coefficients have been computed (observations). Let $w_1, \dots, w_J$ be selected Fourier coefficients with respect to the normalized functions $\widehat{e}_1, \dots, \widehat{e}_J$. Assume they have been computed approximately with a tolerance $\varepsilon > 0$. Then we have

$$C = \{x \in H : \operatorname{supp}(x) \subset \Omega_0 \text{ and } |\langle \widehat{e}_j, x\rangle - w_j| \le \varepsilon \text{ for } j = 1, \dots, J\}.$$

In order to find the *least-energy* function satisfying the constraints, one can define

$$\Phi(x) = \frac{1}{2}\|x\|^2_{L^2(\Omega;\mathbf{R})} \quad \text{and}$$

$$\Psi(x) = \frac{1}{2}\|x\|^2_{L^2(\Omega\setminus\Omega_0;\mathbf{R})} + \frac{1}{2}\sum_{j=1}^{J}\Big[|\langle \widehat{e}_j, x\rangle - w_j| - \varepsilon\Big]_+^2.$$

Here

$$\nabla\Phi(x) = x, \quad \text{and} \quad \nabla\Psi(x) = x\mathbf{1}_{\Omega\setminus\Omega_0} + \sum_{j=1}^{J}\rho_j(x)\widehat{e}_j,$$

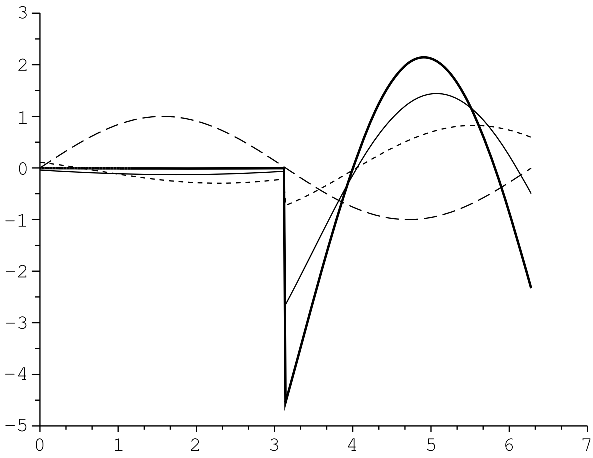
where $\mathbf{1}$ is the characteristic function and

$$\rho_j(x) = \begin{cases} \langle \widehat{e}_j, x\rangle - w_j - \varepsilon & \text{if } \langle \widehat{e}_j, x\rangle - w_j > \varepsilon, \\ \langle \widehat{e}_j, x\rangle - w_j + \varepsilon & \text{if } \langle \widehat{e}_j, x\rangle - w_j < -\varepsilon, \\ 0 & \text{otherwise,} \end{cases}$$

for $j = 1, \dots, J$. One easily sees that $L_\Phi = 1$ and $L_\Psi = J + 1$. We provide a simple numerical simulation with $\Omega = [0, 2\pi] \subset \mathbf{R}$, $\Omega_0 = [\pi, 2\pi]$, $\widehat{e}_1(t) = \frac{1}{\sqrt{2\pi}}$, $\widehat{e}_2(t) = \frac{1}{\sqrt{\pi}}\cos(t)$, $\widehat{e}_3(t) = \frac{1}{\sqrt{\pi}}\sin(t)$, $w = (0, 1, -1)$, and $\varepsilon = 10^{-2}$. The following naive SCILAB implementation uses $\beta_n = n$ and $\lambda_n = \frac{1}{3n}$ starting from $x^1(t) = \sin(t)$.

```
N=1000; h=0.02; eps=0.01;
t=0:h:2*%pi; K=length(t); K2=(K-1)/2;
e1=ones(1,K); e2=cos(t); e3=sin(t); w=[0, 1, -1];
x=sin(t); y=zeros(1,K);
for n=1:N
  d1=(sqrt(2*%pi)/K)*sum(x)-w(1);
  d2=(2*sqrt(%pi)/K)*sum(e2.*x)-w(2);
  d3=(2*sqrt(%pi)/K)*sum(e3.*x)-w(3);
  if d1>eps then rho1=d1-eps; elseif d1<-eps then rho1=d1+eps; else
  rho1=0; end
  if d2>eps then rho2=d2-eps; elseif d2<-eps then rho2=d2+eps; else
  rho2=0; end
  if d3>eps then rho3=d3-eps; elseif d3<-eps then rho3=d3+eps; else
  rho3=0; end
  lambdan=1/(3*n); betan=n;
  y=x; for j=(K2+1):K y(1,j)=0; end
  z=(1/sqrt(2*%pi))*rho1*e1+(1/sqrt(%pi))*rho2*e2+(1/sqrt(%pi))*rho3*e3;
  x=(1-lambdan)*x-lambdan*betan*y-lambdan*betan*z;
end
```

Fig. 1 Approximate solutions x^1 (– –), x^{10} (- - -), x^{100} (—) and x^{1000} (**—**)



The processing time for 1000 iterations was 0.7 seconds on a personal computer with an E2200 Intel(R) Pentium(R) Dual CPU and 3 GB of RAM. Figure 1 shows the evolution of the approximate solutions.

Following Sect. 3.3, the energy may be replaced by different selection criteria, such as the Boltzmann–Shannon entropy. Its implementation goes beyond the scope of this paper, though.

5 Concluding Remarks

We have presented a *diagonal gradient scheme* inspired by previous works from [1, 2], and [8]. The algorithm couples the gradient method with a general exterior penalization procedure. We establish the weak or strong convergence according to the properties of the objective function. Next, we provide some guidelines for the implementation of the method. These include the selection of the parameters as well as stability and robustness properties. Finally, we discuss applications to relaxed feasibility, mathematical programming with convex inequality constraints, the distance between (possibly infinite-dimensional) closed affine subspaces of a Hilbert space, structured optimization with coupling, Stokes equation and signal reconstruction.

Acknowledgements Partly supported by FONDECYT grant 11090023 and Basal Proyect, CMM, Universidad de Chile. The author thanks H. Attouch for useful remarks.

References

1. Attouch, H., Czarnecki, M.-O.: Asymptotic behavior of coupled dynamical systems with multiscale aspects. J. Differ. Equ. **248**(6), 1315–1344 (2010)
2. Attouch, H., Czarnecki, M.-O., Peypouquet, J.: Prox-penalization and splitting methods for constrained variational problems. SIAM J. Optim. **21**(1), 149–173 (2011)
3. Auslender, A., Crouzeix, J.-P., Fedit, P.: Penalty-proximal methods in convex programming. J. Optim. Theory Appl. **55**(1), 1–21 (1987)

4. Alvarez, F., Cominetti, R.: Primal and dual convergence of a proximal point exponential penalty method for linear programming. Math. Program., Ser. A **93**(1), 87–96 (2002)
5. Cominetti, R., Courdurier, M.: Coupling general penalty schemes for convex programming with the steepest descent method and the proximal point algorithm. SIAM J. Optim. **13**, 745–765 (2002)
6. Kiwiel, K.: Exact penalty functions in proximal bundle methods for constrained convex nondifferentiable minimization. Math. Program., Ser. B **52**(2), 285–302 (1991)
7. Peypouquet, J.: Asymptotic convergence to the optimal value of diagonal proximal iterations in convex minimization. J. Convex Anal. **16**(1), 277–286 (2009)
8. Attouch, H., Czarnecki, M.-O., Peypouquet, J.: Coupling forward-backward with penalty schemes and parallel splitting for constrained variational inequalities. SIAM J. Optim. (in press)
9. Bertsekas, D.: Nonlinear Programming. Athena Scientific, Belmont (1999)
10. Rockafellar, R.-T.: Convex Analysis. Princeton University Press, Princeton NJ (1970)
11. Combettes, P.-L.: Quasi-Fejérian analysis of some optimization algorithms. In: Butnariu, D., Censor, Y., Reich, S. (eds.) Inherently Parallel Algorithms for Feasibility and Optimization, pp. 115–152. Elsevier, Amsterdam (2001)
12. Opial, Z.: Weak convergence of the sequence of successive approximations for nonexpansive mappings. Bull. Am. Math. Soc. **73**, 591–597 (1967)
13. Baillon, J.-B., Haddad, G.: Quelques propriétés des opérateurs angle-bornés et *n*-cycliquement monotones. Isr. J. Math. **26**(2), 137–150 (1977)
14. Alvarez, F., Peypouquet, J.: Asymptotic almost-equivalence and ergodic convergence of Lipschitz evolution systems in Banach spaces. Nonlinear Anal. **73**(9), 3018–3033 (2010)
15. Brézis, H.: Analyse fonctionnelle: théorie et applications. Dunod, Paris (1999)
16. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Alternating proximal algorithms for weakly coupled convex minimization problems, Applications to dynamical games and PDEs. J. Convex Anal. **15**(3), 485–506 (2008)
17. Attouch, H., Cabot, A., Frankel, P., Peypouquet, J.: Alternating proximal algorithms for constrained variational inequalities. Application to domain decomposition for PDE's. Nonlinear Anal., Theory, Methods Appl. **74**(18), 7455–7473 (2011)
18. Xu, M.-H.: Proximal alternating directions method for structured variational inequalities. J. Optim. Theory Appl. **134**, 107–117 (2007)
19. Chen, G., Teboulle, M.: A proximal-based decomposition method for convex minimization problems. Math. Program., Ser. A **64**(1), 81–101 (1994)
20. Ekeland, I., Temam, R.: Convex Analysis and Variational Problems. Classics in Applied Mathematics, vol. 28. SIAM, Philadelphia (1999)
21. Evans, L.: Partial Differential Equations, 2nd edn. AMS Graduate Studies in Mathematics. AMS, Providence (2002)
22. Candès, E.-J., Romberg, J.-K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)