Gradient-type penalty method with inertial effects for solving constrained convex optimization problems with smooth data

Radu Ioan Boţ [*†] Ernö Robert Csetnek [‡] Nimit Nimana [§]

June 2, 2017

Abstract. We consider the problem of minimizing a smooth convex objective function subject to the set of minima of another differentiable convex function. In order to solve this problem, we propose an algorithm which combines the gradient method with a penalization technique. Moreover, we insert in our algorithm an inertial term, which is able to take advantage of the history of the iterates. We show weak convergence of the generated sequence of iterates to an optimal solution of the optimization problem, provided a condition expressed via the Fenchel conjugate of the constraint function is fulfilled. We also prove convergence for the objective function values to the optimal objective value. The convergence analysis carried out in this paper relies on the celebrated Opial Lemma and generalized Fejér monotonicity techniques. We illustrate the functionality of the method via a numerical experiment addressing image classification via support vector machines.

Key Words. gradient method, penalization, Fenchel conjugate, inertial algorithm

AMS subject classification. 47H05, 65K05, 90C25

1 Introduction and preliminaries

Let H be a real Hilbert space with the norm and inner product given by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, respectively, and f and g be convex functions acting on H, which we assume for simplicity to be everywhere defined and (Fréchet) differentiable. The object of our investigation is the optimization problem

$$\min_{x\in\operatorname{argmin} g} f(x). \tag{1}$$

[*]University of Vienna, Faculty of Mathematics, Oskar-Morgenstern-Platz 1, A-1090 Vienna, Austria, email: radu.bot@univie.ac.at.

[†]Invited Associate Professor, Babeş-Bolyai University, Faculty of Mathematics and Computer Sciences, Str. M. Kogălniceanu nr. 1, 400084 Cluj-Napoca, Romania

[‡]University of Vienna, Faculty of Mathematics, Oskar-Morgenstern-Platz 1, A-1090 Vienna, Austria, email: ernoe.robert.csetnek@univie.ac.at. Research supported by FWF (Austrian Science Fund), Lise Meitner Programme, project M 1682-N25.

[§]Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok 65000, Thailand, email: nimitn@hotmail.com. Research done during the two months' stay of the author in Spring 2016 at the Faculty of Mathematics of the University of Vienna. The author is thankful to the Royal Golden Jubilee PhD Program for financial support.

We assume that

$$\mathcal{S} := \operatorname{argmin}\{f(x) : x \in \operatorname{argmin} g\} \neq \emptyset$$

and that the gradients ∇f and ∇g are Lipschitz continuous operators with constants L_f and L_g, respectively.

The work [5] of Attouch and Czarnecki has attracted since its appearance a huge interest from the research community, since it undertakes a qualitative analysis of the optimal solutions of (1) from the perspective of a penalty-term based dynamical system. This represented the starting point for the design and development of numerical algorithms for solving the minimization problem (1), several variants of it involving also nonsmooth data up to monotone inclusions that are related to optimality systems of constrained optimization problems. We refer the reader to [4–8, 10, 13–15, 20–23, 33, 35] and the references therein for more insights into this research topic.

A key assumption used in this context in order to guarantee the convergence properties of the numerical algorithms is the condition

$$\sum_{n=1}^{\infty} \lambda_n \beta_n \left[g^* \left(\frac{p}{\beta_n} \right) - \sigma_{\operatorname{argmin} g} \left(\frac{p}{\beta_n} \right) \right] < +\infty \ \forall p \in \operatorname{ran}(N_{\operatorname{argmin} g}),$$

where $\{\lambda_n\}_{n=1}^{\infty}$ and $\{\beta_n\}_{n=1}^{\infty}$ are positive sequences, $g^* : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is the Fenchel conjugate of g:

$$g^*(p) = \sup_{x \in \mathcal{H}} \{\langle p, x \rangle - g(x)\} \ \forall p \in \mathcal{H};$$

$\sigma_{\operatorname{argmin} g} : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is the support function of the set $\operatorname{argmin} g$:

$$\sigma_{\operatorname{argmin} g}(p) = \sup_{x \in \operatorname{argmin} g} \langle p, x \rangle \ \forall p \in \mathcal{H};$$

and $N_{\operatorname{argmin} g}$ is the normal cone to the set $\operatorname{argmin} g$, defined by

$$N_{\operatorname{argmin} g}(x) = \{p \in \mathcal{H} : \langle p, y - x \rangle \leq 0 \ \forall y \in \operatorname{argmin} g\}$$

for $x \in \operatorname{argmin} g$ and $N_{\operatorname{argmin} g}(x) = \emptyset$ for $x \notin \operatorname{argmin} g$. Finally, $\operatorname{ran}(N_{\operatorname{argmin} g})$ denotes the range of the normal cone $N_{\operatorname{argmin} g}$, that is, $p \in \operatorname{ran}(N_{\operatorname{argmin} g})$ if and only if there exists $x \in \operatorname{argmin} g$ such that $p \in N_{\operatorname{argmin} g}(x)$. Let us notice that for $x \in \operatorname{argmin} g$ one has $p \in N_{\operatorname{argmin} g}(x)$ if and only if $\sigma_{\operatorname{argmin} g}(p) = \langle p, x \rangle$. We also assume without loss of generality that $\min g = 0$.

In this paper we propose a numerical algorithm for solving (1) that combines the gradient method with penalization strategies also by employing inertial and memory effects. Algorithms of inertial type result from the time discretization of differential inclusions of second order type (see [1, 3]) and were first investigated in the context of the minimization of a differentiable function by Polyak in [36] and Bertsekas in [12]. The resulting iterative schemes share the feature that the next iterate is defined by means of the last two iterates, a fact which induces the inertial effect in the algorithm. Since the works [1, 3], one can notice an increasing number of research efforts dedicated to algorithms of inertial type (see [1–3, 9, 16–19, 24–28, 30–32, 34]).

In this paper we consider the following inertial algorithm for solving (1):

Algorithm 1 *Initialization: Choose the positive sequences* $\{\lambda_n\}_{n=1}^{\infty}$ *and* $\{\beta_n\}_{n=1}^{\infty}$*, and a positive constant parameter* $\alpha \in (0, 1)$*. Take arbitrary* $x_0, x_1 \in H$*.*

Iterative step: For given current iterates $x_{n-1}, x_n \in H$ *(*$n \geq 1$*), define* $x_{n+1} \in H$ *by*

$$x_{n+1} := x_n + \alpha(x_n - x_{n-1}) - \lambda_n \nabla f(x_n) - \lambda_n \beta_n \nabla g(x_n).$$

We notice that in the above iterative scheme $\{\lambda_n\}_{n=1}^{\infty}$ represents the sequence of step sizes, $\{\beta_n\}_{n=1}^{\infty}$ the sequence of penalty parameters, while α controls the influence of the inertial term.

For every $n \geq 1$ we denote by $\Omega_n := f + \beta_n g$, which is also a (Fréchet) differentiable function, and notice that $\nabla\Omega_n$ is $L_n := L_f + \beta_n L_g$-Lipschitz continuous.

In case $\alpha = 0$, Algorithm 1 collapses in the algorithm considered in [35] for solving (1). We prove weak convergence for the generated iterates to an optimal solution of (1), by making use of generalized Fejér monotonicity techniques and the Opial Lemma and by imposing the key assumption mentioned above as well as some mild conditions on the involved parameters. Moreover, the performed analysis allows us also to show the convergence of the objective function values to the optimal objective value of (1). As an illustration of the theoretical results, we present in the last section an application addressing image classification via support vector machines.

2 Convergence analysis

This section is devoted to the asymptotic analysis of Algorithm 1.

Assumption 2 *Assume that the following statements hold:*

(I) The function f is bounded from below;

(II) There exist positive constants $c > 1$ and $K > 0$ such that $\frac{L_n}{2} + \frac{\alpha-1}{\lambda_n} \leq -(c + (1+\alpha)K)$ and $\beta_{n+1} - \beta_n \leq K\lambda_{n+1}\beta_{n+1}$ for all $n \geq 1$;

(III) For every $p \in \mathrm{ran}(N_{\mathrm{argmin}\, g})$, we have $\sum_{n=1}^{\infty} \lambda_n\beta_n \left[g^*\left(\frac{p}{\beta_n}\right) - \sigma_{\mathrm{argmin}\, g}\left(\frac{p}{\beta_n}\right)\right] < +\infty$;

(IV) $\liminf_{n\to+\infty} \lambda_n\beta_n > 0$, $\left(\frac{1}{\lambda_{n+1}} - \frac{1}{\lambda_n}\right) \leq \frac{2}{\alpha}$ for all $n \geq 1$ and $\sum_{n=1}^{\infty} \lambda_n = +\infty$.

We would like to mention that in [21] we proposed a forward-backward-forward algorithm of penalty-type, endowed with inertial and memory effects, for solving monotone inclusion problems, which gave rise to a primal-dual iterative scheme for solving convex optimization problems with complex structures. However, we succeeded in proving only weak ergodic convergence for the generated iterates, while with the specific choice of the sequences $\{\lambda_n\}_{n=1}^{\infty}$ and $\{\beta_n\}_{n=1}^{\infty}$ in Assumption 2 we will be able to prove weak convergence of the iterates generated in Algorithm 1 to an optimal solution of (1).

Remark 3 The conditions in Assumption 2 slightly extend the ones considered in [35] in the noninertial case. The only differences are given by the first inequality in (II), which here involves the constant α which controls the inertial terms (for the corresponding condition in [35] one only has to take $\alpha = 0$), and by the inequality $\left(\frac{1}{\lambda_{n+1}} - \frac{1}{\lambda_n}\right) \leq \frac{2}{\alpha}$ for all $n \geq 1$.

We refer to Remark 12 for situations where the fulfillment of the conditions in Assumption 2 is guaranteed.

We start the convergence analysis with three technical results.

Lemma 4 *Let $\overline{x} \in \mathcal{S}$ and set $\overline{p} := -\nabla f(\overline{x})$. We have for all $n \geq 1$*

$$\begin{aligned}\varphi_{n+1} - \varphi_n - \alpha(\varphi_n - \varphi_{n-1}) + \lambda_n\beta_n g(x_n) \;\leq\; & \|x_{n+1} - x_n\|^2 + \alpha\|x_n - x_{n-1}\|^2 \\ & + \lambda_n\beta_n\left[g^*\left(\frac{2\overline{p}}{\beta_n}\right) - \sigma_{\mathrm{argmin}\, g}\left(\frac{2\overline{p}}{\beta_n}\right)\right],\end{aligned} \tag{2}$$

where $\varphi_n := \|x_n - \overline{x}\|^2$.

Proof. Since $\overline{x} \in \mathcal{S}$, we have according to the first-order optimality conditions that $0 \in \nabla f(\overline{x}) + N_{\operatorname{argmin} g}(\overline{x})$, thus $\overline{p} = -\nabla f(\overline{x}) \in N_{\operatorname{argmin} g}(\overline{x})$. Notice that for all $n \geq 1$

$$\nabla f(x_n) = \frac{y_n - x_{n+1}}{\lambda_n} - \beta_n \nabla g(x_n),$$

where $y_n := x_n + \alpha(x_n - x_{n-1})$. This, together with the monotonicity of ∇f, imply that

$$\left\langle \frac{y_n - x_{n+1}}{\lambda_n} - \beta_n \nabla g(x_n) + \overline{p}, x_n - \overline{x} \right\rangle = \langle \nabla f(x_n) - \nabla f(\overline{x}), x_n - \overline{x} \rangle \geq 0 \ \forall n \geq 1, \tag{3}$$

so

$$2\left\langle y_n - x_{n+1}, x_n - \overline{x} \right\rangle \geq 2\lambda_n\beta_n \langle \nabla g(x_n), x_n - \overline{x} \rangle - 2\lambda_n \langle \overline{p}, x_n - \overline{x} \rangle \ \forall n \geq 1. \tag{4}$$

On the other hand, since g is convex and differentiable, we have for all $n \geq 1$

$$0 = g(\overline{x}) \geq g(x_n) + \langle \nabla g(x_n), \overline{x} - x_n \rangle,$$

which means that

$$2\lambda_n\beta_n \langle \nabla g(x_n), x_n - \overline{x} \rangle \geq 2\lambda_n\beta_n g(x_n). \tag{5}$$

As for all $n \geq 1$

$$2\langle x_n - x_{n+1}, x_n - \overline{x} \rangle = \|x_{n+1} - x_n\|^2 + \varphi_n - \varphi_{n+1}$$

and

$$2\alpha\langle x_n - x_{n-1}, x_n - \overline{x} \rangle = \alpha\|x_n - x_{n-1}\|^2 + \alpha\left(\varphi_n - \varphi_{n-1}\right),$$

it follows

$$\begin{aligned} 2\langle y_n - x_{n+1}, x_n - \overline{x} \rangle &= 2\langle x_n - x_{n+1}, x_n - \overline{x} \rangle + 2\alpha\langle x_n - x_{n-1}, x_n - \overline{x} \rangle \\ &= \|x_{n+1} - x_n\|^2 + \alpha\|x_n - x_{n-1}\|^2 + \varphi_n - \varphi_{n+1} + \alpha\left(\varphi_n - \varphi_{n-1}\right). \end{aligned} \tag{6}$$

Combining (4), (5) and (6), we obtain that for each $n \geq 1$

$$\begin{aligned} &\varphi_{n+1} - \varphi_n - \alpha\left(\varphi_n - \varphi_{n-1}\right) + \lambda_n\beta_n g(x_n) \\ &\qquad \leq \|x_{n+1} - x_n\|^2 + \alpha\|x_n - x_{n-1}\|^2 - \lambda_n\beta_n g(x_n) + 2\lambda_n\langle \overline{p}, x_n \rangle - 2\lambda_n\langle \overline{p}, \overline{x} \rangle. \end{aligned} \tag{7}$$

Finally, since $\overline{x} \in \operatorname{argmin} g$, we have that for all $n \geq 1$

$$\begin{aligned} 2\lambda_n\langle \overline{p}, x_n \rangle - \lambda_n\beta_n g(x_n) - 2\lambda_n\langle \overline{p}, \overline{x} \rangle &= \lambda_n\beta_n \left[\left\langle \frac{2\overline{p}}{\beta_n}, x_n \right\rangle - g(x_n) - \left\langle \frac{2\overline{p}}{\beta_n}, \overline{x} \right\rangle \right] \\ &\leq \lambda_n\beta_n \left[g^*\left(\frac{2\overline{p}}{\beta_n}\right) - \left\langle \frac{2\overline{p}}{\beta_n}, \overline{x} \right\rangle \right] \\ &= \lambda_n\beta_n \left[g^*\left(\frac{2\overline{p}}{\beta_n}\right) - \sigma_{\operatorname{argmin} g}\left(\frac{2\overline{p}}{\beta_n}\right) \right], \end{aligned}$$

which completes the proof. ■

Lemma 5 *We have for all* $n \geq 1$

$$\begin{aligned}\Omega_{n+1}(x_{n+1}) &\leq \Omega_n(x_n) + (\beta_{n+1} - \beta_n)g(x_{n+1}) \\ &\quad + \left[\frac{L_n}{2} + \frac{\alpha}{2\lambda_n} - \frac{1}{\lambda_n}\right]\|x_{n+1} - x_n\|^2 + \frac{\alpha}{2\lambda_n}\|x_n - x_{n-1}\|^2.\end{aligned} \tag{8}$$

Proof. From the descent Lemma and the fact that $\nabla\Omega_n$ is L_n-Lipschitz continuous, we get that

$$\Omega_n(x_{n+1}) \leq \Omega_n(x_n) + \langle \nabla\Omega_n(x_n), x_{n+1} - x_n\rangle + \frac{L_n}{2}\|x_{n+1} - x_n\|^2 \ \forall n \geq 1.$$

Since $\nabla\Omega_n(x_n) = -\frac{x_{n+1}-y_n}{\lambda_n}$, it holds for all $n \geq 1$

$$\begin{aligned}f(x_{n+1}) + \beta_n g(x_{n+1}) &\leq f(x_n) + \beta_n g(x_n) \\ &\quad - \left\langle \frac{x_{n+1} - y_n}{\lambda_n}, x_{n+1} - x_n \right\rangle + \frac{L_n}{2}\|x_{n+1} - x_n\|^2\end{aligned}$$

and then

$$\begin{aligned}f(x_{n+1}) + \beta_{n+1} g(x_{n+1}) &\leq f(x_n) + \beta_n g(x_n) + (\beta_{n+1} - \beta_n)g(x_{n+1}) \\ &\quad - \frac{1}{\lambda_n}\|x_{n+1} - x_n\|^2 + \frac{\alpha}{\lambda_n}\langle x_n - x_{n-1}, x_{n+1} - x_n\rangle \\ &\quad + \frac{L_n}{2}\|x_{n+1} - x_n\|^2,\end{aligned}$$

which is nothing else than

$$\begin{aligned}\Omega_{n+1}(x_{n+1}) &\leq \Omega_n(x_n) + (\beta_{n+1} - \beta_n)g(x_{n+1}) + \left[\frac{L_n}{2} - \frac{1}{\lambda_n}\right]\|x_{n+1} - x_n\|^2 \\ &\quad + \frac{\alpha}{\lambda_n}\langle x_n - x_{n-1}, x_{n+1} - x_n\rangle.\end{aligned} \tag{9}$$

By the Cauchy-Schwarz inequalty it holds that

$$\langle x_n - x_{n-1}, x_{n+1} - x_n\rangle \leq \frac{1}{2}\|x_{n-1} - x_n\|^2 + \frac{1}{2}\|x_{n+1} - x_n\|^2,$$

hence, (9) becomes

$$\begin{aligned}\Omega_{n+1}(x_{n+1}) &\leq \Omega_n(x_n) + (\beta_{n+1} - \beta_n)g(x_{n+1}) + \frac{\alpha}{2\lambda_n}\|x_{n-1} - x_n\|^2 \\ &\quad + \left[\frac{L_n}{2} - \frac{1}{\lambda_n} + \frac{\alpha}{2\lambda_n}\right]\|x_{n+1} - x_n\|^2 \ \forall n \geq 1.\end{aligned}$$

■

For $\overline{x} \in \mathcal{S}$ and all $n \geq 1$, we set

$$\begin{aligned}\Gamma_n &:= f(x_n) + (1 - K\lambda_n)\beta_n g(x_n) + K\varphi_n \\ &= \Omega_n(x_n) - K\lambda_n\beta_n g(x_n) + K\varphi_n,\end{aligned}$$

and, for simplicity, we denote

$$\delta_n := \left(\frac{1}{2\lambda_n} + K\right)\alpha + c.$$

Lemma 6 *Let $\overline{x} \in \mathcal{S}$ and set $\overline{p} := -\nabla f(\overline{x})$. We have for all $n \geq 2$*

$$\begin{aligned}\Gamma_{n+1} - \Gamma_n - \alpha(\Gamma_n - \Gamma_{n-1}) \;\leq\; & -\delta_n\|x_{n+1} - x_n\|^2 + \alpha\left(\frac{1}{2\lambda_n} + K\right)\|x_n - x_{n-1}\|^2 \\ & + K\lambda_n\beta_n\left[g^*\left(\frac{2\overline{p}}{\beta_n}\right) - \sigma_{\operatorname{argmin} g}\left(\frac{2\overline{p}}{\beta_n}\right)\right] \\ & + \alpha\left(\Omega_{n-1}(x_{n-1}) - \Omega_n(x_n)\right) \\ & + \alpha K\left(\lambda_n\beta_n g(x_n) - \lambda_{n-1}\beta_{n-1}g(x_{n-1})\right). \end{aligned} \tag{10}$$

Proof. According to Lemma 5 and Assumption 2(II), (8) becomes for all $n \geq 1$

$$\begin{aligned}\Omega_{n+1}(x_{n+1}) - \Omega_n(x_n) - K\lambda_{n+1}\beta_{n+1}g(x_{n+1}) \;\leq\; & -(K + \delta_n)\|x_{n+1} - x_n\|^2 \\ & + \frac{\alpha}{2\lambda_n}\|x_n - x_{n-1}\|^2. \end{aligned} \tag{11}$$

On the other hand, after multiplying (2) by K, we obtain for all $n \geq 1$

$$\begin{aligned}& K\varphi_{n+1} - K\varphi_n - \alpha\left(K\varphi_n - K\varphi_{n-1}\right) + K\lambda_n\beta_n g(x_n) \\ & \qquad \leq K\|x_{n+1} - x_n\|^2 + K\alpha\|x_n - x_{n-1}\|^2 + K\lambda_n\beta_n\left[g^*\left(\frac{2\overline{p}}{\beta_n}\right) - \sigma_{\operatorname{argmin} g}\left(\frac{2\overline{p}}{\beta_n}\right)\right]. \end{aligned} \tag{12}$$

After summing up the relations (11) and (12) and adding on both sides of the resulting inequality the expressions $\alpha\left(\Omega_{n-1}(x_{n-1}) - \Omega_n(x_n)\right)$ and $\alpha\left(K\lambda_n\beta_n g(x_n) - K\lambda_{n-1}\beta_{n-1}g(x_{n-1})\right)$ for all $n \geq 2$, we obtain the required statement. ■

The following proposition will play an essential role in the convergence analysis (see also [1–3, 16]).

Proposition 7 *Let $\{a_n\}_{n=1}^{\infty}$, $\{b_n\}_{n=1}^{\infty}$ and $\{c_n\}_{n=1}^{\infty}$ be real sequences and $\alpha \in [0, 1)$ be given. Assume that $\{a_n\}_{n=1}^{\infty}$ is bounded from below, $\{b_n\}_{n=1}^{\infty}$ is nonnegative and $\sum_{n=1}^{\infty} c_n < +\infty$ such that*

$$a_{n+1} - a_n - \alpha(a_n - a_{n-1}) + b_n \leq c_n \;\forall n \geq 1.$$

Then the following statements hold:

(i) $\sum_{n=1}^{\infty}[a_n - a_{n-1}]_+ < +\infty$, where $[t]_+ := \max\{t, 0\}$;

(ii) $\{a_n\}_{n=1}^{\infty}$ converges and $\sum_{n=1}^{\infty} b_n < +\infty$.

The following lemma collects some convergence properties of the sequences involved in our analysis.

Lemma 8 *Let $\overline{x} \in \mathcal{S}$. Then the following statements are true:*

(i) The sequence $\{\Gamma_n\}_{n=1}^{\infty}$ is bounded from below.

(ii) $\sum_{n=1}^{\infty}\|x_{n+1} - x_n\|^2 < +\infty$ and $\lim_{n\to+\infty}\Gamma_n$ exists.

(iii) $\lim_{n\to+\infty}\|x_n - \overline{x}\|$ exists and $\sum_{n=1}^{\infty}\lambda_n\beta_n g(x_n) < +\infty$.

(iv) $\lim_{n\to+\infty}\Omega_n(x_n)$ exists.

(v) $\lim_{n\to+\infty} g(x_n) = 0$ and every sequential weak cluster point of the sequence $\{x_n\}_{n=1}^{\infty}$ lies in $\operatorname{argmin} g$.

Proof. We set $\overline{p} := -\nabla f(\overline{x})$ and recall that $g(\overline{x}) = \min g = 0$.

(i) Since f is convex and differentiable, it holds for all $n \geq 1$

$$\begin{aligned}\Gamma_n &= f(x_n) + (1 - K\lambda_n)\beta_n g(x_n) + K\varphi_n \\ &\geq f(x_n) + K\|x_n - \overline{x}\|^2 \\ &\geq f(\overline{x}) + \langle \nabla f(\overline{x}), x_n - \overline{x}\rangle + K\|x_n - \overline{x}\|^2 \geq f(\overline{x}) - \frac{\|\overline{p}\|^2}{4K},\end{aligned}$$

which means that $\{\Gamma_n\}_{n=1}^{\infty}$ is bounded from below. Notice that the first inequality in the above relation is a consequence of Assumption 2(II), since $\frac{1-\alpha}{\lambda_n} \geq c + (1+\alpha)K \geq K$, thus $\lambda_n K \leq 1 - \alpha \leq 1$ for all $n \geq 1$.

(ii) For all $n \geq 2$, we may set

$$\mu_n := \Gamma_n - \alpha\Gamma_{n-1} + \alpha\left(\frac{1}{2\lambda_n} + K\right)\|x_n - x_{n-1}\|^2$$

and

$$u_n := \Omega_{n-1}(x_{n-1}) - \Omega_n(x_n) + K\lambda_n\beta_n g(x_n) - K\lambda_{n-1}\beta_{n-1}g(x_{n-1}).$$

We fix a natural number $N_0 \geq 2$. Then

$$\sum_{n=2}^{N_0} u_n = f(x_1) + (1 - K\lambda_1)\beta_1 g(x_1) - f(x_{N_0}) - (1 - K\lambda_{N_0})\beta_{N_0} g(x_{N_0}).$$

Since f is bounded from below and $g(x_{N_0}) \geq g(\overline{x}) = 0$, it follows that $\sum_{n=2}^{\infty} u_n < +\infty$.

We notice that $-\delta_n + \alpha\left(\frac{1}{2\lambda_{n+1}} + K\right) = \frac{\alpha}{2}\left(\frac{1}{\lambda_{n+1}} - \frac{1}{\lambda_n}\right) - c$ and, since $\left(\frac{1}{\lambda_{n+1}} - \frac{1}{\lambda_n}\right) \leq \frac{2}{\alpha}$, we have for all $n \geq 1$

$$-\delta_n + \alpha\left(\frac{1}{2\lambda_{n+1}} + K\right) \leq 1 - c. \tag{13}$$

Thus, according Lemma 6, we get for all $n \geq 2$

$$\begin{aligned}\mu_{n+1} - \mu_n &= \Gamma_{n+1} - \Gamma_n - \alpha(\Gamma_n - \Gamma_{n-1}) + \alpha\left(\frac{1}{2\lambda_{n+1}} + K\right)\|x_{n+1} - x_n\|^2 \\ &\quad -\alpha\left(\frac{1}{2\lambda_n} + K\right)\|x_n - x_{n-1}\|^2 \\ &\leq -\delta_n\|x_{n+1} - x_n\|^2 + K\lambda_n\beta_n\left[g^*\left(\frac{2\overline{p}}{\beta_n}\right) - \sigma_{\operatorname{argmin} g}\left(\frac{2\overline{p}}{\beta_n}\right)\right] \\ &\quad +\alpha u_n + \alpha\left(\frac{1}{2\lambda_{n+1}} + K\right)\|x_{n+1} - x_n\|^2 \\ &\leq (1-c)\|x_{n+1} - x_n\|^2 + K\lambda_n\beta_n\left[g^*\left(\frac{2\overline{p}}{\beta_n}\right) - \sigma_{\operatorname{argmin} g}\left(\frac{2\overline{p}}{\beta_n}\right)\right] + \alpha u_n.\end{aligned}$$

We fix another natural number $N_1 \geq 2$ and sum up the last inequality for $n = 2, \ldots, N_1$. We

obtain

$$\begin{aligned}\mu_{N_1+1}-\mu_2 &\leq (1-c)\sum_{n=2}^{N_1}\|x_{n+1}-x_n\|^2\\ &\quad +K\sum_{n=2}^{N_1}\lambda_n\beta_n\left[g^*\left(\frac{2\overline{p}}{\beta_n}\right)-\sigma_{\operatorname{argmin} g}\left(\frac{2\overline{p}}{\beta_n}\right)\right]\\ &\quad +\alpha\sum_{n=2}^{N_1}u_n, \end{aligned}\tag{14}$$

which, by taking into account Assumption 2(III), means that $\{\mu_n\}_{n=2}^{\infty}$ is bounded from above by a positive number that we denote by M. Consequently, for all $n\geq 2$ we have

$$\Gamma_{n+1}-\alpha\Gamma_n\leq\mu_{n+1}\leq M,$$

so

$$\Gamma_{n+1}\leq\alpha\Gamma_n+M,$$

which further implies that

$$\Gamma_n\leq\alpha^{n-2}\Gamma_2+M\sum_{k=1}^{n-2}\alpha^{k-1}\leq\alpha^{n-2}\Gamma_2+\frac{M}{1-\alpha}\ \forall n\geq 3.$$

We have for all $n\geq 2$

$$\mu_{n+1}\geq f(\overline{x})-\frac{\|\overline{p}\|^2}{4K}-\alpha\Gamma_n,$$

hence

$$-\mu_{n+1}\leq\alpha\Gamma_n-f(\overline{x})+\frac{\|\overline{p}\|^2}{4K}\leq\alpha^{n-1}\Gamma_2+\frac{\alpha M}{1-\alpha}-f(\overline{x})+\frac{\|\overline{p}\|^2}{4K}.\tag{15}$$

Consequently, for the arbitrarily chosen natural number $N_1\geq 2$, we have (see (14))

$$\begin{aligned}(c-1)\sum_{n=2}^{N_1}\|x_{n+1}-x_n\|^2 &\leq -\mu_{N_1+1}+\mu_2\\ &\quad +K\sum_{n=2}^{N_1}\lambda_n\beta_n\left[g^*\left(\frac{2\overline{p}}{\beta_n}\right)-\sigma_{\operatorname{argmin} g}\left(\frac{2\overline{p}}{\beta_n}\right)\right]+\alpha\sum_{n=2}^{N_1}u_n,\end{aligned}$$

which together with (15) and the fact that $c>1$ implies that

$$\sum_{n=1}^{\infty}\|x_{n+1}-x_n\|^2<+\infty.$$

On the other hand, due to (13) we have $\delta_{n+1}\leq\delta_n+1$ for all $n\geq 1$. Consequently, using also that $c>1$, (10) implies that

$$\begin{aligned}\Gamma_{n+1}-\Gamma_n-\alpha(\Gamma_n-\Gamma_{n-1}) &\leq -\delta_n\|x_{n+1}-x_n\|^2+(\delta_n-c)\|x_n-x_{n-1}\|^2\\ &\quad +K\lambda_n\beta_n\left[g^*\left(\frac{2\overline{p}}{\beta_n}\right)-\sigma_{\operatorname{argmin} g}\left(\frac{2\overline{p}}{\beta_n}\right)\right]+\alpha u_n\\ &\leq -\delta_n\|x_{n+1}-x_n\|^2+\delta_{n-1}\|x_n-x_{n-1}\|^2\\ &\quad +K\lambda_n\beta_n\left[g^*\left(\frac{2\overline{p}}{\beta_n}\right)-\sigma_{\operatorname{argmin} g}\left(\frac{2\overline{p}}{\beta_n}\right)\right]+\alpha u_n\ \forall n\geq 1.\end{aligned}$$

According to Proposition 7 and by taking into account that $\{\Gamma_n\}_{n=1}^{\infty}$ is bounded from below, we obtain that $\lim_{n\to+\infty}\Gamma_n$ exists.

(iii) By Lemma 4 and Proposition 7, $\lim_{n\to+\infty}\varphi_n$ exists and $\sum_{n=1}^{\infty}\lambda_n\beta_n g(x_n) < +\infty$.

(iv) Since $\Omega_n(x_n) = \Gamma_n - K\varphi_n + K\lambda_n\beta_n g(x_n)$ for all $n \geq 1$, by using (ii) and (iii), we get that $\lim_{n\to+\infty}\Omega_n(x_n)$ exists.

(v) Since $\liminf_{n\to+\infty}\lambda_n\beta_n > 0$, we also obtain that $\lim_{n\to+\infty} g(x_n) = 0$. Let w be a sequential weak cluster point of $\{x_n\}_{n=1}^{\infty}$ and assume that the subsequence $\{x_{n_j}\}_{j=1}^{\infty}$ converges weakly to w. Since g is weak lower semicontinuous, we have

$$g(w) \leq \liminf_{j\to+\infty} g(x_{n_j}) = \lim_{n\to+\infty} g(x_n) = 0,$$

which implies that $w \in \operatorname{argmin} g$. This completes the proof. ■

In order to show also the convergence of the sequence $(f(x_n))_{n=1}^{\infty}$, we prove first the following result.

Lemma 9 *Let $\overline{x} \in \mathcal{S}$ be given. We have*

$$\sum_{n=1}^{\infty}\lambda_n\left[\Omega_n(x_n) - f(\overline{x})\right] < +\infty.$$

Proof. Since f is convex and differentiable, we have for all $n \geq 1$

$$f(\overline{x}) \geq f(x_n) + \langle\nabla f(x_n), \overline{x} - x_n\rangle.$$

Since g is convex and differentiable, we have for all $n \geq 1$

$$0 \geq \beta_n g(x_n) + \langle\beta_n\nabla g(x_n), \overline{x} - x_n\rangle,$$

which together imply that

$$\begin{aligned} f(\overline{x}) &\geq \Omega_n(x_n) + \langle\nabla\Omega_n(x_n), \overline{x} - x_n\rangle \\ &= \Omega_n(x_n) + \left\langle\frac{y_n - x_{n+1}}{\lambda_n}, \overline{x} - x_n\right\rangle \ \forall n \geq 1. \end{aligned}$$

From here we obtain for all $n \geq 1$ (see (6))

$$\begin{aligned} 2\lambda_n\left[\Omega_n(x_n) - f(\overline{x})\right] &\leq 2\langle y_n - x_{n+1}, x_n - \overline{x}\rangle \\ &= \|x_{n+1} - x_n\|^2 + \varphi_n - \varphi_{n+1} + \alpha(\varphi_n - \varphi_{n-1}) + \alpha\|x_n - x_{n-1}\|^2. \end{aligned}$$

Hence, by using the previous lemma, the required result holds. ■

The Opial Lemma that we recall below will play an important role in the proof of the main result of this paper.

Proposition 10 (Opial Lemma) *Let H be a real Hilbert space, $C \subseteq H$ a nonempty set and $\{x_n\}_{n=1}^{\infty}$ a given sequence such that:*

(i) For every $z \in C$, $\lim_{n\to+\infty}\|x_n - z\|$ exists.

(ii) Every sequential weak cluster point of $\{x_n\}_{n=1}^{\infty}$ lies in C.

Then the sequence $\{x_n\}_{n=1}^{\infty}$ converges weakly to a point in C.

Theorem 11 *(i) The sequence $\{x_n\}_{n=1}^{\infty}$ converges weakly to a point in $\mathcal{S}$.*

(ii) The sequence $(f(x_n))_{n=1}^{\infty}$ converges to the optimal objective value of the optimization problem (1).

Proof. (i) According to Lemma 8, $\lim_{n\to+\infty}\|x_n-\overline{x}\|$ exists for all $\overline{x}\in\mathcal{S}$. Let w be a sequential weak cluster point of $\{x_n\}_{n=1}^{\infty}$. Then there exists a subsequence $\{x_{n_j}\}_{j=1}^{\infty}$ of $\{x_n\}_{n=1}^{\infty}$ such that x_{n_j} converges weakly to w as $j\to+\infty$. By Lemma 8, we have that $w\in\operatorname{argmin} g$. This means that in order to come to the conclusion it suffices to show that $f(w)\leq f(x)$ for all $x\in\operatorname{argmin} g$. From Lemma 9, Lemma 8 and the fact that $\sum_{n=1}^{\infty}\lambda_n=+\infty$, it follows that $\lim_{n\to\infty}[\Omega_n(x_n)-f(\overline{x})]\leq 0$ for all $\overline{x}\in\mathcal{S}$. Thus,

$$f(w)\leq\liminf_{j\to+\infty}f(x_{n_j})\leq\lim_{n\to+\infty}\Omega_n(x_n)\leq f(\overline{x})\ \forall\overline{x}\in\mathcal{S},$$

which shows that $w\in\mathcal{S}$. Hence, thanks to Opial Lemma, $\{x_n\}_{n=1}^{\infty}$ converges weakly to a point in $\mathcal{S}$.

(ii) The statement follows easily from the above considerations. ■

In the end of this section we present some situations where Assumption 2 is verified.

Remark 12 Let $\alpha\in(0,1)$, $c\in(1,+\infty)$, $q\in(0,1)$ and $\gamma\in\left(0,\frac{2}{L_g}\right)$ be arbitrarily chosen. We set

$$K:=\frac{2}{\alpha}>0,$$

$$\beta_n:=\frac{\gamma[L_f+2((1+\alpha)K+c)]}{2-\gamma L_g}+(1-\alpha)\gamma Kn^q,$$

and

$$\lambda_n:=\frac{(1-\alpha)\gamma}{\beta_n},$$

for all $n\geq 1$.

(i) Since $\beta_n\geq\frac{\gamma[L_f+2((1+\alpha)K+c)]}{2-\gamma L_g}$, we have $\beta_n(2-\gamma L_g)\geq\gamma[L_f+2((1+\alpha)K+c)]$, which implies that $\frac{L_n}{2}+\frac{\alpha-1}{\lambda_n}\leq-(c+(1+\alpha)K)$ for all $n\geq 1$.

(ii) For all $n\geq 1$ it holds

$$\beta_{n+1}-\beta_n=(1-\alpha)\gamma K[(n+1)^q-n^q]\leq(1-\alpha)\gamma K=K\lambda_{n+1}\beta_{n+1}.$$

(iii) It holds $\liminf_{n\to+\infty}\lambda_n\beta_n=\liminf_{n\to+\infty}(1-\alpha)\gamma>0$.

(iv) For all $n\geq 1$ we have

$$\frac{1}{\lambda_{n+1}}-\frac{1}{\lambda_n}=\frac{1}{(1-\alpha)\gamma}(\beta_{n+1}-\beta_n)=K\left((n+1)^q-n^q\right)\leq K=\frac{2}{\alpha}.$$

(v) Since $q\in(0,1)$, we have $\sum_{n=1}^{\infty}\frac{1}{\beta_n}=+\infty$, which implies that $\sum_{n=1}^{\infty}\lambda_n=+\infty$.

(vi) Finally, as $g\leq\delta_{\operatorname{argmin} g}$, we have $g^*\geq(\delta_{\operatorname{argmin} g})^*=\sigma_{\operatorname{argmin} g}$ and this implies that $g^*-\sigma_{\operatorname{argmin} g}\geq 0$. We present a situation where Assumption 2(III) holds and refer to [10] for further examples. For instance, if $g(x)\geq\frac{a}{2}\operatorname{dist}^2(x,\operatorname{argmin} g)$ where $a>0$, then $g^*(x)-\sigma_{\operatorname{argmin} g}(x)\leq\frac{1}{2a}\|x\|^2$ for every $x\in H$. Thus, for $p\in\operatorname{ran}(N_{\operatorname{argmin} g})$, we have

$$\lambda_n\beta_n\left[g^*\left(\frac{p}{\beta_n}\right)-\sigma_{\operatorname{argmin} g}\left(\frac{p}{\beta_n}\right)\right]\leq\frac{\lambda_n}{2a\beta_n}\|p\|^2.$$

Hence $\sum_{n=1}^{\infty}\lambda_n\beta_n\left[g^*\left(\frac{p}{\beta_n}\right)-\sigma_{\operatorname{argmin} g}\left(\frac{p}{\beta_n}\right)\right]$ converges, if $\sum_{n=1}^{\infty}\frac{\lambda_n}{\beta_n}$ converges or, equivalently, if $\sum_{n=1}^{\infty}\frac{1}{\beta_n^2}$ converges. This holds for the above choices of $\{\beta_n\}_{n=1}^{\infty}$ and $\{\lambda_n\}_{n=1}^{\infty}$ when $q\in\left(\frac{1}{2},1\right)$.

3 Numerical example: image classification via support vector machines

In this section we employ the algorithm proposed in this paper in the context of image classification via support vector machines.

Having a set of training data $a_i \in \mathbb{R}^n,\ i = 1, \ldots, k$, belonging to one of two given classes denoted by "-1" and "$+1$", the aim is to construct by using this information a decision function given in the form of a separating hyperplane, which assigns every new data to one of the two classes with a misclassification rate as low as possible. In order to be able to handle the situation when a full separation is not possible, we make use of non-negative slack variables $\xi_i \geq 0,\ i = 1, \ldots, k$; thus the goal will be to find $(s, r, \xi) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^k_+$ as optimal solution of the following optimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|s\|^2 + \frac{C}{2}\|\xi\|^2 \\ \text{subject to} & d_i(a_i^\top s + r) \geq 1 - \xi_i, \forall i = 1, \ldots, k \\ & \xi_i \geq 0, \forall i = 1, \ldots, k, \end{array}$$

where, for $i = 1, \ldots, k$, d_i is equal to -1 if a_i belongs to the class "-1" and it is equal to $+1$, otherwise. Each new data $a \in \mathbb{R}^n$ will by assigned to one of the two classes by means of the resulting decision function $z(a) = a^\top s + r$, namely, a will be assigned to the class "-1", if $z(a) < 0$, and to the class "$+1$", otherwise. For more theoretical insights in support vector machines we refer the reader to [29].

By making use of the matrix

$$\mathbf{A} = \begin{bmatrix} d_1 a_1^\top & d_1 & 1 & 0 & \cdots & 0 \\ d_2 a_2^\top & d_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ d_k a_k^\top & d_k & 0 & 0 & \cdots & 1 \\ \mathbf{0}_{\mathbb{R}^n}^\top & 0 & 1 & 0 & \cdots & 0 \\ \mathbf{0}_{\mathbb{R}^n}^\top & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{\mathbb{R}^n}^\top & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{2k \times (n+1+k)}$$

the problem under investigation can be written as

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|s\|^2 + \frac{C}{2}\|\xi\|^2 \\ \text{subject to} & \mathbf{A}\begin{pmatrix} s \\ r \\ \xi \end{pmatrix} - \begin{pmatrix} \mathbf{1}_{\mathbb{R}^k} \\ \mathbf{0}_{\mathbb{R}^k} \end{pmatrix} \in \mathbb{R}^{2k}_+ \end{array}$$

or, equivalently,

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|s\|^2 + \frac{C}{2}\|\xi\|^2 \\ \text{subject to} & \begin{pmatrix} s \\ r \\ \xi \end{pmatrix} \in \arg\min \frac{1}{2}\text{dist}^2\left(\mathbf{A}(\cdot) - \begin{pmatrix} \mathbf{1}_{\mathbb{R}^k} \\ \mathbf{0}_{\mathbb{R}^k} \end{pmatrix}, \mathbb{R}^{2k}_+\right). \end{array}$$

By considering $f : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^k \to \mathbb{R}$ as $f\begin{pmatrix} s \\ r \\ \xi \end{pmatrix} := \frac{1}{2}\|s\|^2 + \frac{C}{2}\|\xi\|^2$, we have $\nabla f\begin{pmatrix} s \\ r \\ \xi \end{pmatrix} = \begin{pmatrix} s \\ 0 \\ C\xi \end{pmatrix}$ and notice that ∇f is $\max\{1, C\}$-Lipschitz continuous.

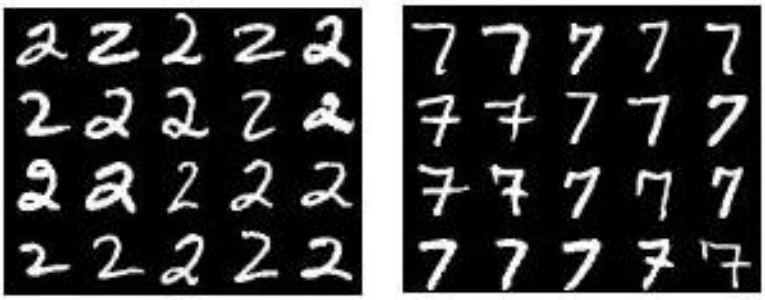

Figure 1: A sample of images belonging to the classes -1 and $+1$, respectively.

Figure 2: A sample of misclassified images.

Further, for $g\begin{pmatrix} s \\ r \\ \xi \end{pmatrix} := \frac{1}{2}\text{dist}^2\left(\mathbf{A}\begin{pmatrix} s \\ r \\ \xi \end{pmatrix} - \begin{pmatrix} \mathbf{1}_{\mathbb{R}^k} \\ \mathbf{0}_{\mathbb{R}^k} \end{pmatrix}, \mathbb{R}^{2k}_+\right)$, we have $\nabla g\begin{pmatrix} s \\ r \\ \xi \end{pmatrix} = \mathbf{A}^\top\left(I - \text{proj}_{\mathbb{R}^{2k}_+}\right)\left(\mathbf{A}\begin{pmatrix} s \\ r \\ \xi \end{pmatrix} - \begin{pmatrix} \mathbf{1}_{\mathbb{R}^k} \\ \mathbf{0}_{\mathbb{R}^k} \end{pmatrix}\right)$ and notice that ∇g is $\|\mathbf{A}\|^2$-Lipschitz continuous, where $\text{proj}_{\mathbb{R}^{2k}_+}$ denotes the projection operator on the set $\mathbb{R}^{2k}_+$.

For the numerical experiments we used a data set consisting of 6.000 training images and 2.060 test images of size 28×28 taken from the website http://www.cs.nyu.edu/~roweis/data.html representing the handwritten digits 2 and 7, labeled by -1 and $+1$, respectively (see Figure 1). We evaluated the quality of the resulting decision function on test data set by computing the percentage of misclassified images.

We denote by $\mathcal{D} = \{(X_i, Y_i), i = 1, \ldots, 6.000\} \subset \mathbb{R}^{784} \times \{-1, +1\}$ the set of available training data consisting of 3.000 images in the class -1 and 3.000 images in the class $+1$. Due to numerical reasons each image has been vectorized and normalized. We tested in MATLAB different combinations of parameters chosen as in Remark 12 by running the algorithm for 3.000 iterations. A sample of misclassified images is shown in Figure 2.

In Table 1 we present the misclassification rate in percentage for different choices for the parameters $\alpha \in (0, 1)$ (we recall that in this case we take $K := 2/\alpha$) and $C > 0$, while for $\alpha = 0$ which corresponds to the noninertial version of the algorithm we consider different choices of the parameter $K > 0$ and $C > 0$. We observe that when combining $\alpha = 0.1$ with each regularization parameters $C = 5, 10, 100$ leads to the lowest misclassification rate with 2.1845 %.

In Table 2 we present the misclassification rate in percentage for different choices of the parameters $C > 0$ and $c > 1$. The lowest classification rate of 2.1845% is obtained for each regularization parameter $C = 5, 10, 100$.

Finally, Table 3 shows the misclassification rate in percentage for different choices for the

α	$C = 0.1$	$C = 1$	$C = 2$	$C = 5$	$C = 10$	$C = 100$
0.1	2.2330	2.2330	2.2330	2.1845	2.1845	2.1845
0.3	2.2330	2.2816	2.2816	2.2816	2.2816	2.2816
0.5	2.2330	2.2330	2.2330	2.2816	2.2816	2.3301
0.7	2.3786	2.3786	2.3786	2.3786	2.3786	2.3786
0.9	2.9126	2.9126	2.9126	2.9126	2.8641	2.8155
0 (K=0.1)	3.1068	3.0583	3.0583	2.9612	2.9612	2.7184
0 (K=1)	2.2816	2.2330	2.2330	2.2330	2.2330	2.2330
0 (K=10)	2.2816	2.2330	2.2330	2.2330	2.2330	2.2330
0 (K=100)	2.2330	2.2330	2.2330	2.2330	2.2330	2.2330
0 (K=1000)	2.2330	2.2330	2.2330	2.2330	2.2330	2.2330

Table 1: Misclassification rate in percentage for different choices for the parameters α and C when $c = 2$ and $q = 0.9$.

C	$c = 1.1$	$c = 2$	$c = 5$	$c = 10$	$c = 100$
0.1	2.2330	2.2330	2.2330	2.2330	2.2330
1	2.2330	2.2330	2.2330	2.2330	2.2330
2	2.2330	2.2330	2.2330	2.2330	2.2330
5	2.1845	2.1845	2.1845	2.1845	2.1845
10	2.1845	2.1845	2.1845	2.1845	2.1845
100	2.1845	2.1845	2.1845	2.1845	2.1845

Table 2: Misclassification rate in percentage for different choices for the parameters C and $c > 1$ when $\alpha = 0.1$ and $q = 0.9$.

C	$q = 0.6$	$q = 0.75$	$q = 0.9$
0.1	2.2816	2.3301	2.2330
1	2.2330	2.2816	2.2330
2	2.2816	2.2816	2.2330
5	2.2330	2.2816	2.1845
10	2.2330	2.2816	2.1845
100	2.2330	2.2330	2.1845

Table 3: Misclassification rate in percentage for different choices for the parameters C and $q \in (1/2, 1)$ when $\alpha = 0.1$ and $c = 2$.

parameters $C > 0$ and $q \in (1/2, 1)$. The lowest classification rate of 2.1845% is obtained when combining the value $q = 0.9$ with each regularization parameter $C = 5, 10, 100$.

Acknowledgements. The authors are thankful to two anonymous reviewers for hints and comments which improved the quality of the paper.

References

[1] F. Alvarez, *On the minimizing property of a second order dissipative system in Hilbert spaces*, SIAM Journal on Control and Optimization 38(4), 1102–1119, 2000

[2] F. Alvarez, *Weak convergence of a relaxed and inertial hybrid projection-proximal point algorithm for maximal monotone operators in Hilbert space*, SIAM Journal on Optimization 14(3), 773–782, 2004

[3] F. Alvarez, H. Attouch, *An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping*, Set-Valued Analysis 9, 3–11, 2001

[4] H. Attouch, A. Cabot, M.-O. Czarnecki, *Asymptotic behavior of nonautonomous monotone and subgradient evolution equations*, to appear in Transactions of the American Mathematical Society, arXiv:1601.00767, 2016

[5] H. Attouch, M.-O. Czarnecki, *Asymptotic behavior of coupled dynamical systems with multiscale aspects*, Journal of Differential Equations 248(6), 1315–1344, 2010

[6] H. Attouch, M.-O. Czarnecki, *Asymptotic behavior of gradient-like dynamical systems involving inertia and multiscale aspects*, Journal of Differential Equations 262(3), 2745–2770, 2017

[7] H. Attouch, M.-O. Czarnecki, J. Peypouquet, *Prox-penalization and splitting methods for constrained variational problems*, SIAM Journal on Optimization 21(1), 149–173, 2011

[8] H. Attouch, M.-O. Czarnecki, J. Peypouquet, *Coupling forward-backward with penalty schemes and parallel splitting for constrained variational inequalities*, SIAM Journal on Optimization 21(4), 1251–1274, 2011

[9] H. Attouch, J. Peypouquet, P. Redont, *A dynamical approach to an inertial forward-backward algorithm for convex minimization*, SIAM Journal on Optimization 24(1), 232–256, 2014

[10] S. Banert, R.I. Boţ, *Backward penalty schemes for monotone inclusion problems*, Journal of Optimization Theory and Applications 166(3), 930–948, 2015

[11] H.H. Bauschke, P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics, Springer, New York, 2011

[12] D.P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Cambridge, MA, 1999

[13] R.I. Boţ, E.R. Csetnek, *Forward-backward and Tseng's type penalty schemes for monotone inclusion problems*, Set-Valued and Variational Analysis 22, 313–331, 2014

[14] R.I. Boţ, E.R. Csetnek, *A Tseng's type penalty scheme for solving inclusion problems involving linearly composed and parallel-sum type monotone operators*, Vietnam Journal of Mathematics 42(4), 451–465, 2014

[15] R.I. Boţ, E.R. Csetnek, *Levenberg-Marquardt dynamics associated to variational inequalities*, Set-Valued and Variational Analysis, DOI: 10.1007/s11228-017-0409-8, 2017

[16] R.I. Boţ, E.R. Csetnek, *An inertial forward-backward-forward primal-dual splitting algorithm for solving monotone inclusion problems*, Numerical Algorithms 71, 519–540, 2016

[17] R.I. Boţ, E.R. Csetnek, *An inertial alternating direction method of multipliers*, Minimax Theory and its Applications 1(1), 29–49, 2016

[18] R.I. Boţ, E.R. Csetnek, *A hybrid proximal-extragradient algorithm with inertial effects*, Numerical Functional Analysis and Optimization 36(8), 951–963, 2015

[19] R.I. Boţ, E.R. Csetnek, *An inertial Tseng's type proximal algorithm for nonsmooth and nonconvex optimization problems*, Journal of Optimization Theory and Applications 171 (2), 600–616, 2016

[20] R.I. Boţ, E.R. Csetnek, *Approaching the solving of constrained variational inequalities via penalty term-based dynamical systems*, Journal of Mathematical Analysis and Applications 435, 1688–1700, 2016

[21] R.I. Boţ, E.R. Csetnek, *Penalty schemes with inertial effects for monotone inclusion problems*, Optimization 66(6), 965–982, 2017

[22] R.I. Boţ, E.R. Csetnek, *Second order dynamical systems associated to variational inequalities*, Applicable Analysis 96(5), 799–809, 2017

[23] R.I. Boţ, E.R. Csetnek, *A second order dynamical system with Hessian-driven damping and penalty term associated to variational inequalities*, arXiv:1608.04137, 2016

[24] R.I. Boţ, E.R. Csetnek, C. Hendrich, *Inertial Douglas-Rachford splitting for monotone inclusion problems*, Applied Mathematics and Computation 256, 472–487, 2015

[25] R.I. Boţ, E.R. Csetnek, S. László, *An inertial forward-backward algorithm for the minimization of the sum of two nonconvex functions*, EURO Journal on Computational Optimization 4, 3–25, 2016

[26] A. Cabot, P. Frankel, *Asymptotics for some proximal-like method involving inertia and memory aspects*, Set-Valued and Variational Analysis 19, 59–74, 2011

[27] C. Chen, R.H. Chan, S. MA, J. Yang, *Inertial proximal ADMM for linearly constrained separable convex optimization*, SIAM Journal on Imaging Sciences 8(4), 2239–2267, 2015

[28] C. Chen, S. MA, J. Yang, *A general inertial proximal point algorithm for mixed variational inequality problem*, SIAM Journal on Optimization 25(4), 2120–2142, 2015

[29] N. Cristianini, J.S. Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, 2000

[30] P.-E. Maingé, *Convergence theorems for inertial KM-type algorithms*, Journal of Computational and Applied Mathematics 219, 223–236, 2008

[31] P.-E. Maingé, A. Moudafi, *Convergence of new inertial proximal methods for dc programming*, SIAM Journal on Optimization 19(1), 397–413, 2008

[32] A. Moudafi, M. Oliny, *Convergence of a splitting inertial proximal method for monotone operators*, Journal of Computational and Applied Mathematics 155, 447–454, 2003

[33] N. Noun, J. Peypouquet, *Forward-backward penalty scheme for constrained convex minimization without inf-compactness*, Journal of Optimization Theory and Applications, 158(3), 787–795, 2013

[34] P. Ochs, Y. Chen, T. Brox, T. Pock, *iPiano: Inertial proximal algorithm for non-convex optimization*, SIAM Journal on Imaging Sciences 7(2), 1388–1419, 2014

[35] J. Peypouquet, *Coupling the gradient method with a general exterior penalization scheme for convex minimization*, Journal of Optimization Theory and Applications 153(1), 123–138, 2012

[36] B.T. Polyak, *Introduction to Optimization*, (Translated from the Russian) Translations Series in Mathematics and Engineering, Optimization Software, Inc., Publications Division, New York, 1987