A variable smoothing algorithm for solving convex optimization problems

Radu Ioan Boţ * Christopher Hendrich †

July 13, 2012

Abstract. In this article we propose a method for solving unconstrained optimization problems with convex and Lipschitz continuous objective functions. By making use of the Moreau envelopes of the functions occurring in the objective, we smooth the latter to a convex and differentiable function with Lipschitz continuous gradient by using both variable and constant smoothing parameters. The resulting problem is solved via an accelerated first-order method and this allows us to recover approximately the optimal solutions to the initial optimization problem with a rate of convergence of order $\mathcal{O}(\frac{\ln k}{k})$ for variable smoothing and of order $\mathcal{O}(\frac{1}{k})$ for constant smoothing. Some numerical experiments employing the variable smoothing method in image processing and in supervised learning classification are also presented.

Keywords. Moreau envelope, regularization, variable smoothing, fast gradient method

AMS subject classification. 90C25, 90C46, 47A52

1 Introduction

In this paper we introduce and investigate the convergence properties of an efficient algorithm for solving nondifferentiable optimization problems of type

$$\inf_{x\in\mathcal{H}}\{f(x)+g(Kx)\}, \tag{1}$$

where $\mathcal{H}$ and $\mathcal{K}$ are real Hilbert spaces, $f:\mathcal{H}\to\mathbb{R}$ and $g:\mathcal{K}\to\mathbb{R}$ are convex and Lipschitz continuous functions and the operator $K:\mathcal{H}\to\mathcal{K}$ is linear and continuous. By replacing the functions f and g through their Moreau envelopes, approach which can be seen as part of the family of smoothing techniques introduced in [13–15], we approximate (1) by a convex optimization problem with a differentiable objective function with Lipschitz continuous gradient. This smoothing approach can be seen as the counterpart of the so-called double smoothing method investigated in [5, 6, 11], which assumes the smoothing of the Fenchel-dual problem to (1) to an optimization problem with a

*Faculty of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany, e-mail: radu.bot@mathematik.tu-chemnitz.de. Research partially supported by DFG (German Research Foundation), project BO 2516/4-1.

†Faculty of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany, e-mail: christopher.hendrich@mathematik.tu-chemnitz.de.

strongly convex and differentiable objective function with Lipschitz continuous gradient. There, the smoothed dual problem is solved via an appropriate fast gradient method (cf. [16]) and a primal optimal solution is reconstructed with a given level of accuracy. In contrast to that approach, which asks for the boundedness of the effective domains of f and g, determinant is here the boundedness of the effective domains of the conjugate functions f^* and g^*, which is automatically guaranteed by the Lipschitz continuity of f and g, respectively. For solving the resulting smoothed problem we propose an extension of the accelerated gradient method of Nesterov (cf. [17]) for convex optimization problems involving variable smoothing parameters which are updated in each iteration. This scheme yields for the minimization of the objective of the initial problem a rate of convergence of order $\mathcal{O}(\frac{\ln k}{k})$, while, in the particular case when the smoothing parameters are constant, the order of the rate of convergence becomes $\mathcal{O}(\frac{1}{k})$. Nonetheless, using variable smoothing parameters has an important advantage, although the theoretical rate of convergence is not as good as when these are constant. In the first case the approach generates a sequence of iterates $(x_k)_{k\geq 1}$ such that $(f(x_k) + g(Kx_k))_{k\geq 1}$ converges to the optimal objective value of (1). In the case of constant smoothing variables the approach provides a sequence of iterates which solves the problem (1) with an apriori given accuracy, however, the sequence $(f(x_k) + g(Kx_k))_{k\geq 1}$ may not converge to the optimal objective value of the problem to be solved.

In addition, we show, on the one hand, that the two approaches can be designed and keep the same convergence behavior also in the case when f is differentiable with Lipschitz continuous gradient and, on the other hand, that they can be employed also for solving the extended version of (1)

$$\inf_{x\in\mathcal{H}}\left\{f(x) + \sum_{i=1}^{m} g_i(K_i x)\right\}, \tag{2}$$

where $\mathcal{K}_i$ are real Hilbert spaces, $g_i : \mathcal{K}_i \to \mathbb{R}$ are convex and Lipschitz continuous functions and $K_i : \mathcal{H} \to \mathcal{K}_i$, $i = 1, \ldots, m$, are linear continuous operators.

The structure of this paper is as follows. In Section 2 we recall some elements of convex analysis and establish the working framework. Section 3 is mainly devoted to the description of the iterative methods for solving (1) and of their convergence properties for both variable and constant smoothing and to the presentation of some of their variants. In Section 4 numerical experiments employing the variable smoothing method in image processing and in supervised vector machines classification are presented.

2 Preliminaries of convex analysis and problem formulation

In the following we are considering the real Hilbert spaces $\mathcal{H}$ and $\mathcal{K}$ endowed with the inner product $\langle\cdot,\cdot\rangle$ and associated norm $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$. By $B_{\mathcal{H}} \subseteq \mathcal{H}$ and $\mathbb{R}_{++}$ we denote the *closed unit ball* of $\mathcal{H}$ and the set of strictly positive real numbers, respectively. The *indicator function* of the set $C \subseteq \mathcal{H}$ is the function $\delta_C : \mathcal{H} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ defined by $\delta_C(x) = 0$ for $x \in C$ and $\delta_C(x) = +\infty$, otherwise. For a function $f : \mathcal{H} \to \overline{\mathbb{R}}$ we denote by $\operatorname{dom} f := \{x \in \mathcal{H} : f(x) < +\infty\}$ its *effective domain*. We call f *proper* if $\operatorname{dom} f \neq \emptyset$ and $f(x) > -\infty$ for all $x \in \mathcal{H}$. The *conjugate function* of f is $f^* : \mathcal{H} \to \overline{\mathbb{R}}$,

$f^*(p) = \sup\{\langle p, x\rangle - f(x) : x \in \mathcal{H}\}$ for all $p \in \mathcal{H}$. The *biconjugate function* of f is $f^{**} : \mathcal{H} \to \overline{\mathbb{R}}$, $f^{**}(x) = \sup\{\langle x, p\rangle - f^*(p) : p \in \mathcal{H}\}$ and, when f is proper, convex and lower semicontinuous, according to the Fenchel-Moreau Theorem, one has $f = f^{**}$. The *(convex) subdifferential* of the function f at $x \in \mathcal{H}$ is the set $\partial f(x) = \{p \in \mathcal{H} : f(y) - f(x) \geq \langle p, y - x\rangle \ \forall y \in \mathcal{H}\}$, if $f(x) \in \mathbb{R}$, and is taken to be the empty set, otherwise. For a linear operator $K : \mathcal{H} \to \mathcal{K}$, the operator $K^* : \mathcal{K} \to \mathcal{H}$ is the *adjoint operator* of K and is defined by $\langle K^*y, x\rangle = \langle y, Kx\rangle$ for all $x \in \mathcal{H}$ and all $y \in \mathcal{K}$.

Having two functions $f, g : \mathcal{H} \to \overline{\mathbb{R}}$, their *infimal convolution* is defined by $f\square g : \mathcal{H} \to \overline{\mathbb{R}}$, $(f\square g)(x) = \inf_{y\in\mathcal{H}}\{f(y) + g(x - y)\}$ for all $x \in \mathcal{H}$. When $f, g : \mathcal{H} \to \overline{\mathbb{R}}$ are proper and convex, then

$$(f + g)^* = f^*\square g^* \tag{3}$$

provided that f (or g) is continuous at a point belonging to $\operatorname{dom} f \cap \operatorname{dom} g$. For other qualification conditions guaranteeing (3) we refer the reader to [3].

The *Moreau envelope* of parameter $\gamma \in \mathbb{R}_{++}$ of a proper, convex and lower semicontinuous function $f : \mathcal{H} \to \overline{\mathbb{R}}$ is the function ${}^{\gamma}f : \mathcal{H} \to \mathbb{R}$, defined as

$${}^{\gamma}f(x) := f\square\left(\frac{1}{2\gamma}\|\cdot\|^2\right)(x) = \inf_{y\in\mathcal{H}}\left\{f(y) + \frac{1}{2\gamma}\|x - y\|^2\right\} \ \forall x \in \mathcal{H}.$$

For every $x \in \mathcal{H}$ we denote by $\operatorname{Prox}_{\gamma f}(x)$ the *proximal point* of parameter γ of f at x, namely, the unique optimal solution of the optimization problem

$$\inf_{y\in\mathcal{H}}\left\{f(y) + \frac{1}{2\gamma}\|y - x\|^2\right\}. \tag{4}$$

Notice that $\operatorname{Prox}_{\gamma f} : \mathcal{H} \to \mathcal{H}$ is single-valued and firmly nonexpansive (cf. [1, Proposition 12.27]), i.e.,

$$\|\operatorname{Prox}_{\gamma f}(x) - \operatorname{Prox}_{\gamma f}(y)\|^2 + \|(x - \operatorname{Prox}_{\gamma f}(x)) - (y - \operatorname{Prox}_{\gamma f}(y))\|^2 \leq \|x - y\|^2 \ \forall x, y \in \mathcal{H}, \tag{5}$$

thus 1-Lipschitz continuous, i.e., Lipschitz continuous with Lipschitz constant equal to 1. We also have (cf. [1, Theorem 14.3])

$${}^{\gamma}f(x) + {}^{\frac{1}{\gamma}}f^*(\tfrac{x}{\gamma}) = \frac{\|x\|^2}{2\gamma} \ \forall x \in \mathcal{H} \tag{6}$$

and the extended *Moreau's decomposition formula*

$$\operatorname{Prox}_{\gamma f}(x) + \gamma\operatorname{Prox}_{\frac{1}{\gamma}f^*}\left(\tfrac{x}{\gamma}\right) = x \ \forall x \in \mathcal{H}. \tag{7}$$

The function ${}^{\gamma}f$ is (Fréchet) differentiable on $\mathcal{H}$ and its gradient $\nabla({}^{\gamma}f) : \mathcal{H} \to \mathcal{H}$ fulfills (cf. [1, Proposition 12.29])

$$\nabla({}^{\gamma}f)(x) = \tfrac{1}{\gamma}(x - \operatorname{Prox}_{\gamma f}(x)) \ \forall x \in \mathcal{H}, \tag{8}$$

being in the light of (5) $\frac{1}{\gamma}$-Lipschitz continuous. For a nonempty, convex and closed set $C \subseteq \mathcal{H}$ and $\gamma \in \mathbb{R}_{++}$ we have that $\operatorname{Prox}_{\gamma\delta_C} = \mathcal{P}_C$, where $\mathcal{P}_C : \mathcal{H} \to C$, $\mathcal{P}_C(x) = \arg\min_{z\in C}\|x - z\|$, denotes the *projection operator* on C.

When $f : \mathcal{H} \to \mathbb{R}$ is convex and differentiable having an $L_{\nabla f}$-Lipschitz continuous gradient, then for all $x, y \in \mathcal{H}$ it holds (see, for instance, [1, 16, 17])

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_{\nabla f}}{2} \|y - x\|^2 . \tag{9}$$

The optimization problem that we investigate in this paper is

$$(P) \qquad \inf_{x \in \mathcal{H}} \{f(x) + g(Kx)\},$$

where $K : \mathcal{H} \to \mathcal{K}$ is a linear continuous operator and $f : \mathcal{H} \to \mathbb{R}$ and $g : \mathcal{K} \to \mathbb{R}$ are convex and L_f-Lipschitz continuous and L_g-Lipschitz continuous functions, respectively. According to [2, Proposition 4.4.6] we have that

$$\operatorname{dom} f^* \subseteq L_f B_{\mathcal{H}} \text{ and } \operatorname{dom} g^* \subseteq L_g B_{\mathcal{K}}. \tag{10}$$

3 The algorithm and its variants

3.1 The smoothing of the problem (P)

The algorithms we would like to introduce and analyze from the point of view of their convergence properties assume in a first instance an appropriate smoothing of the problem (P) which we are going to describe in the following.

For $\rho \in \mathbb{R}_{++}$ we smooth f via its Moreau envelope of parameter ρ, ${}^{\rho}f : \mathcal{H} \to \mathbb{R}$, ${}^{\rho}f(x) = \left(f \square \frac{1}{2\rho} \|\cdot\|^2\right)(x)$ for every $x \in \mathcal{H}$. According to the Fenchel-Moreau Theorem and due to (3), one has for $x \in \mathcal{H}$

$${}^{\rho}f(x) = \left(f^{**} \square \frac{1}{2\rho} \|\cdot\|^2\right)(x) = \left(f^* + \frac{\rho}{2} \|\cdot\|^2\right)^*(x) = \sup_{p \in \mathcal{H}} \left\{\langle x, p \rangle - f^*(p) - \frac{\rho}{2} \|p\|^2\right\}.$$

As already seen, ${}^{\rho}f$ is differentiable and its gradient (cf. (8) and (7))

$$\nabla({}^{\rho}f) : \mathcal{H} \to \mathcal{H}, \ \nabla({}^{\rho}f) = \tfrac{1}{\rho}(x - \operatorname{Prox}_{\rho f}(x)) = \operatorname{Prox}_{\frac{1}{\rho} f^*}\left(\frac{x}{\rho}\right) \ \forall x \in \mathcal{H},$$

is $\frac{1}{\rho}$-Lipschitz continuous.

For $\mu \in \mathbb{R}_{++}$ we smooth $g \circ K$ via ${}^{\mu}g \circ K : \mathcal{H} \to \mathbb{R}$, ${}^{\mu}g \circ K(x) = \left(g \square \frac{1}{2\mu} \|\cdot\|^2\right)(Kx)$ for every $x \in \mathcal{H}$. According to the Fenchel-Moreau Theorem and due to (3), one has

$$\begin{aligned} {}^{\mu}g \circ K(x) &= \left(g^{**} \square \frac{1}{2\mu} \|\cdot\|^2\right)(Kx) = \left(g^* + \frac{\mu}{2} \|\cdot\|^2\right)^*(Kx) \\ &= \sup_{p \in \mathcal{K}} \left\{\langle x, K^* p \rangle - g^*(p) - \frac{\mu}{2} \|p\|^2\right\} \ \forall x \in \mathcal{H}. \end{aligned}$$

The function ${}^{\mu}g \circ K$ is differentiable and its gradient $\nabla({}^{\mu}g \circ K) : \mathcal{H} \to \mathcal{H}$ fulfills (cf. (8) and (7))

$$\nabla({}^{\mu}g \circ K)(x) = K^* \nabla({}^{\mu}g)(Kx) = \tfrac{1}{\mu} K^*(Kx - \operatorname{Prox}_{\mu g}(Kx)) = K^* \operatorname{Prox}_{\frac{1}{\mu} g^*}\left(\tfrac{Kx}{\mu}\right) \ \forall x \in \mathcal{H}.$$

Further, for every $x, y \in \mathcal{H}$ it holds (see (5))

$$\begin{aligned}\|\nabla({}^{\mu}g \circ K)(x) - \nabla({}^{\mu}g \circ K)(y)\| &\leq \tfrac{1}{\mu}\|K\| \, \|(Kx - \mathrm{Prox}_{\mu g}(Kx)) - (Ky - \mathrm{Prox}_{\mu g}(Ky))\| \\ &\leq \frac{\|K\|^2}{\mu} \|x - y\|,\end{aligned}$$

which shows that $\nabla({}^{\mu}g \circ K)$ is $\frac{\|K\|^2}{\mu}$-Lipschitz continuous.

Finally, we consider as smoothing function for $f + g \circ K$ the function $F^{\rho,\mu} : \mathcal{H} \to \mathbb{R}$, $F^{\rho,\mu}(x) = {}^{\rho}f(x) + {}^{\mu}g \circ K(x)$, which is differentiable with Lipschitz continuous gradient $\nabla F^{\rho,\mu} : \mathcal{H} \to \mathcal{H}$ given by

$$\nabla F^{\rho,\mu}(x) = \mathrm{Prox}_{\frac{1}{\rho}f^*}\left(\tfrac{x}{\rho}\right) + K^* \mathrm{Prox}_{\frac{1}{\mu}g^*}\left(\tfrac{Kx}{\mu}\right) \ \forall x \in \mathcal{H},$$

having as Lipschitz constant $L(\rho, \mu) := \frac{1}{\rho} + \frac{\|K\|^2}{\mu}$.

For $\rho_2 \geq \rho_1 > 0$ and every $x \in \mathcal{H}$ it holds (cf. (10))

$$\begin{aligned}{}^{\rho_1}f(x) &= \sup_{p \in \mathrm{dom}\, f^*} \left\{ \langle x, p \rangle - f^*(p) - \frac{\rho_1}{2}\|p\|^2 \right\} \\ &\leq \sup_{p \in \mathrm{dom}\, f^*} \left\{ \langle x, p \rangle - f^*(p) - \frac{\rho_2}{2}\|p\|^2 \right\} + \sup_{p \in \mathrm{dom}\, f^*} \left\{ \frac{\rho_2 - \rho_1}{2}\|p\|^2 \right\} \\ &\leq {}^{\rho_2}f(x) + (\rho_2 - \rho_1)\frac{L_f^2}{2},\end{aligned}$$

which yields, letting $\rho_1 \downarrow 0$ (cf. [1, Proposition 12.32]),

$${}^{\rho_2}f(x) \leq f(x) \leq {}^{\rho_2}f(x) + \rho_2 \frac{L_f^2}{2}.$$

Similarly, for $\mu_2 \geq \mu_1 > 0$ and every $y \in \mathcal{K}$ it holds

$${}^{\mu_1}g(y) \leq {}^{\mu_2}g(y) + (\mu_2 - \mu_1)\frac{L_g^2}{2},$$

and

$${}^{\mu_2}g(y) \leq g(y) \leq {}^{\rho_2}g(y) + \rho_2 \frac{L_g^2}{2}.$$

Consequently, for $\rho_2 \geq \rho_1 > 0$, $\mu_2 \geq \mu_1 > 0$ and every $x \in \mathcal{H}$ we have

$$F^{\rho_2,\mu_2}(x) \leq F^{\rho_1,\mu_1}(x) \leq F^{\rho_2,\mu_2}(x) + (\rho_2 - \rho_1)\frac{L_f^2}{2} + (\mu_2 - \mu_1)\frac{L_g^2}{2} \tag{11}$$

and

$$F^{\rho_2,\mu_2}(x) \leq F(x) \leq F^{\rho_2,\mu_2}(x) + \rho_2 \frac{L_f^2}{2} + \mu_2 \frac{L_g^2}{2}. \tag{12}$$

3.2 The variable smoothing and the constant smoothing algorithms

Throughout this paper $F : \mathcal{H} \to \mathbb{R}$, $F(x) = f(x) + g(Kx)$, will denote the objective function of (P). The variable smoothing algorithm which we present at the beginning of this subsection can be seen as an extension of the accelerated gradient method of Nesterov (cf. [17]) by using variable smoothing parameters, which we update in each iteration.

$$\boxed{\begin{array}{l}\text{Initialization} : t_1 = 1,\ y_1 = x_0 \in \mathcal{H},\ (\rho_k)_{k\geq 1}, (\mu_k)_{k\geq 1} \subseteq \mathbb{R}_{++} \\ \quad \text{For } k \geq 1 : L_k = \dfrac{1}{\rho_k} + \dfrac{\|K\|^2}{\mu_k}, \\ \qquad\qquad x_k = y_k - \dfrac{1}{L_k}\left(\mathrm{Prox}_{\frac{1}{\rho_k} f^*}\left(\dfrac{y_k}{\rho_k}\right) + K^* \mathrm{Prox}_{\frac{1}{\mu_k} g^*}\left(\dfrac{Ky_k}{\mu_k}\right)\right), \\ \qquad\qquad t_{k+1} = \dfrac{1 + \sqrt{1 + 4t_k^2}}{2}, \\ \qquad\qquad y_{k+1} = x_k + \dfrac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})\end{array}} \tag{A1}$$

The convergence of the algorithm (A1) is proved by the following theorem.

Theorem 1. *Let $f : \mathcal{H} \to \mathbb{R}$ be a convex and L_f-Lipschitz continuous function, $g : \mathcal{K} \to \mathbb{R}$ a convex and L_g-Lipschitz continuous function, $K : \mathcal{H} \to \mathcal{K}$ a linear continuous operator and $x^* \in \mathcal{H}$ an optimal solution to (P). Then, when choosing*

$$\rho_k = \frac{1}{ak} \text{ and } \mu_k = \frac{1}{bk}\ \forall k \geq 1,$$

where $a, b \in \mathbb{R}_{++}$, algorithm (A1) *generates a sequence $(x_k)_{k\geq 1} \subseteq \mathcal{H}$ satisfying*

$$F(x_{k+1}) - F(x^*) \leq \frac{2(a + b\|K\|^2)}{k+2}\|x_0 - x^*\|^2 + \frac{2(1 + \ln(k+1))}{k+2}\left(\frac{L_f^2}{a} + \frac{L_g^2}{b}\right)\ \forall k \geq 1, \tag{13}$$

thus yielding a rate of convergence for the objective of order $\mathcal{O}(\frac{\ln k}{k})$.

Proof. For any $k \geq 1$ we denote $F^k := F^{\rho_k, \mu_k}$, $p_k := (t_k - 1)(x_{k-1} - x_k)$ and

$$\xi_k := \nabla F^k(y_k) = \mathrm{Prox}_{\frac{1}{\rho_k} f^*}\left(\frac{y_k}{\rho_k}\right) + K^* \mathrm{Prox}_{\frac{1}{\mu_k} g^*}\left(\frac{Ky_k}{\mu_k}\right).$$

For any $k \geq 1$ it holds

$$\begin{aligned} p_{k+1} - x_{k+1} &= (t_{k+1} - 1)(x_k - x_{k+1}) - x_{k+1} \\ &= (t_{k+1} - 1)x_k - t_{k+1}\left(y_{k+1} - \frac{1}{L_{k+1}}\nabla F^{k+1}(y_{k+1})\right) \\ &= p_k - x_k + \frac{t_{k+1}}{L_{k+1}}\nabla F^{k+1}(y_{k+1}) \end{aligned}$$

and from here it follows

$$\begin{aligned}
&\|p_{k+1} - x_{k+1} + x^*\|^2 \\
&= \|p_k - x_k + x^*\|^2 + 2\left\langle p_k - x_k + x^*, \frac{t_{k+1}}{L_{k+1}}\xi_{k+1}\right\rangle + \left(\frac{t_{k+1}}{L_{k+1}}\right)^2 \|\xi_{k+1}\|^2 \\
&= \|p_k - x_k + x^*\|^2 + \frac{2t_{k+1}}{L_{k+1}} \langle p_k, \xi_{k+1}\rangle \\
&+ \frac{2t_{k+1}}{L_{k+1}} \left\langle x^* - y_{k+1} - \frac{p_k}{t_{k+1}}, \xi_{k+1}\right\rangle + \left(\frac{t_{k+1}}{L_{k+1}}\right)^2 \|\xi_{k+1}\|^2 \\
&= \|p_k - x_k + x^*\|^2 + \frac{2(t_{k+1}-1)}{L_{k+1}} \langle p_k, \xi_{k+1}\rangle + \frac{2t_{k+1}}{L_{k+1}} \langle x^* - y_{k+1}, \xi_{k+1}\rangle + \left(\frac{t_{k+1}}{L_{k+1}}\right)^2 \|\xi_{k+1}\|^2 .
\end{aligned}$$

Further, using (9), since $x_{k+1} = y_{k+1} - \frac{1}{L_{k+1}}\xi_{k+1}$, it follows

$$\begin{aligned}
F^{k+1}(x_{k+1}) &\le F^{k+1}(y_{k+1}) + \langle \xi_{k+1}, x_{k+1} - y_{k+1}\rangle + \frac{L_{k+1}}{2} \|x_{k+1} - y_{k+1}\|^2 \\
&= F^{k+1}(y_{k+1}) - \frac{1}{L_{k+1}} \|\xi_{k+1}\|^2 + \frac{1}{2L_{k+1}} \|\xi_{k+1}\|^2 \\
&= F^{k+1}(y_{k+1}) - \frac{1}{2L_{k+1}} \|\xi_{k+1}\|^2 \qquad (14)
\end{aligned}$$

and, from here, by making use of the convexity of F^{k+1}, we have

$$\begin{aligned}
\langle x^* - y_{k+1}, \xi_{k+1}\rangle &\le F^{k+1}(x^*) - F^{k+1}(y_{k+1}) \\
&\overset{(14)}{\le} F^{k+1}(x^*) - F^{k+1}(x_{k+1}) - \frac{1}{2L_{k+1}} \|\xi_{k+1}\|^2 \ \ \forall k \ge 1. \qquad (15)
\end{aligned}$$

On the other hand, since $F^{k+1}(x_k) - F^{k+1}(y_{k+1}) \ge \langle \xi_{k+1}, x_k - y_{k+1}\rangle$, we obtain

$$\begin{aligned}
\|\xi_{k+1}\|^2 &\overset{(14)}{\le} 2L_{k+1}(F^{k+1}(y_{k+1}) - F^{k+1}(x_{k+1})) \\
&\le 2L_{k+1}\left(F^{k+1}(x_k) - F^{k+1}(x_{k+1}) - \frac{1}{t_{k+1}} \langle \xi_{k+1}, p_k\rangle\right) \ \ \forall k \ge 1. \qquad (16)
\end{aligned}$$

Thus, as $t_{k+1}^2 - t_{k+1} = t_k^2$ and by making use of (11), for any $k \ge 1$ it yields

$$\begin{aligned}
&\|p_{k+1} - x_{k+1} + x^*\|^2 - \|p_k - x_k + x^*\|^2 \\
&\overset{(15)}{\le} \frac{2(t_{k+1}-1)}{L_{k+1}} \langle p_k, \xi_{k+1}\rangle + \frac{2t_{k+1}}{L_{k+1}}(F^{k+1}(x^*) - F^{k+1}(x_{k+1})) + \frac{t_{k+1}^2 - t_{k+1}}{L_{k+1}^2} \|\xi_{k+1}\|^2 \\
&\overset{(16)}{\le} \frac{2t_{k+1}}{L_{k+1}}(F^{k+1}(x^*) - F^{k+1}(x_{k+1})) + \frac{2(t_{k+1}^2 - t_{k+1})}{L_{k+1}}(F^{k+1}(x_k) - F^{k+1}(x_{k+1})) \\
&\overset{(11)}{\le} \frac{2t_k^2}{L_{k+1}}\left(F^k(x_k) - F^k(x^*) + (\rho_k - \rho_{k+1})\frac{L_f^2}{2} + (\mu_k - \mu_{k+1})\frac{L_g^2}{2}\right) \\
&- \frac{2t_{k+1}^2}{L_{k+1}}(F^{k+1}(x_{k+1}) - F^{k+1}(x^*))
\end{aligned}$$

$$= \frac{2t_k^2}{L_{k+1}}\left(F^k(x_k) - F^k(x^*) + \rho_k\frac{L_f^2}{2} + \mu_k\frac{L_g^2}{2}\right) - \frac{2t_{k+1}^2}{L_{k+1}}(F^{k+1}(x_{k+1}) - F^{k+1}(x^*))$$
$$- \frac{2t_k^2}{L_{k+1}}\left(\rho_{k+1}\frac{L_f^2}{2} + \mu_{k+1}\frac{L_g^2}{2}\right).$$

By using (12) it follows that for any $k \geq 1$

$$F^k(x_k) - F^k(x^*) + \rho_k\frac{L_f^2}{2} + \mu_k\frac{L_g^2}{2} \geq F(x_k) - F^k(x^*) \geq F(x_k) - F(x^*) \geq 0,$$

thus

$$\begin{aligned}
&\|p_{k+1} - x_{k+1} + x^*\|^2 - \|p_k - x_k + x^*\|^2 \\
&\leq \frac{2t_k^2}{L_k}\left(F^k(x_k) - F^k(x^*) + \rho_k\frac{L_f^2}{2} + \mu_k\frac{L_g^2}{2}\right) - \frac{2t_{k+1}^2}{L_{k+1}}(F^{k+1}(x_{k+1}) - F^{k+1}(x^*)) \\
&- \frac{2t_k^2}{L_{k+1}}\left(\rho_{k+1}\frac{L_f^2}{2} + \mu_{k+1}\frac{L_g^2}{2}\right) \\
&= \frac{2t_k^2}{L_k}\left(F^k(x_k) - F^k(x^*) + \rho_k\frac{L_f^2}{2} + \mu_k\frac{L_g^2}{2}\right) - \frac{2t_{k+1}^2}{L_{k+1}}(F^{k+1}(x_{k+1}) - F^{k+1}(x^*)) \\
&- \frac{2t_{k+1}^2}{L_{k+1}}\left(\rho_{k+1}\frac{L_f^2}{2} + \mu_{k+1}\frac{L_g^2}{2}\right) + \frac{2t_{k+1}}{L_{k+1}}\left(\rho_{k+1}\frac{L_f^2}{2} + \mu_{k+1}\frac{L_f^g}{2}\right),
\end{aligned}$$

which implies that

$$\begin{aligned}
&\|p_{k+1} - x_{k+1} + x^*\|^2 + \frac{2t_{k+1}^2}{L_{k+1}}\left(F^{k+1}(x_{k+1}) - F^{k+1}(x^*) + \rho_{k+1}\frac{L_f^2}{2} + \mu_{k+1}\frac{L_g^2}{2}\right) \\
&\leq \|p_k - x_k + x^*\|^2 + \frac{2t_k^2}{L_k}\left(F^k(x_k) - F^k(x^*) + \rho_k\frac{L_f^2}{2} + \mu_k\frac{L_g^2}{2}\right) \\
&+ \frac{2t_{k+1}}{L_{k+1}}\left(\rho_{k+1}\frac{L_f^2}{2} + \mu_{k+1}\frac{L_g^2}{2}\right).
\end{aligned}$$

Making again use of (12) this further yields for any $k \geq 1$

$$\begin{aligned}
&\frac{2t_{k+1}^2}{L_{k+1}}\left(F(x_{k+1}) - F(x^*)\right) \\
&\leq \frac{2t_{k+1}^2}{L_{k+1}}\left(F^{k+1}(x_{k+1}) - F^{k+1}(x^*) + \rho_{k+1}\frac{L_f^2}{2} + \mu_{k+1}\frac{L_g^2}{2}\right) + \|p_{k+1} - x_{k+1} + x^*\|^2 \\
&\leq \frac{2t_1^2}{L_1}\left(F^1(x_1) - F^1(x^*) + \rho_1\frac{L_f^2}{2} + \mu_1\frac{L_g^2}{2}\right) + \|p_1 - x_1 + x^*\|^2 \\
&+ \sum_{s=1}^{k}\frac{2t_{s+1}}{L_{s+1}}\left(\rho_{s+1}\frac{L_f^2}{2} + \mu_{s+1}\frac{L_g^2}{2}\right).
\end{aligned} \tag{17}$$

Since $x_1 = y_1 - \frac{1}{L_1}\nabla F^1(y_1)$ and

$$F^1(x^*) \geq F^1(y_1) + \left\langle \nabla F^1(y_1), x^* - y_1 \right\rangle$$
$$F^1(x_1) \leq F^1(y_1) + \left\langle \nabla F^1(y_1), x_1 - y_1 \right\rangle + \frac{L_1}{2}\|x_1 - y_1\|^2,$$

we get

$$\frac{2t_1^2}{L_1}\left(F^1(x_1) - F^1(x^*)\right) + \|p_1 - x_1 + x^*\|^2$$
$$\leq 2\langle x_1 - y_1, x^* - y_1\rangle - \|x_1 - y_1\|^2 + \|x_1 - x^*\|^2 = \|y_1 - x^*\|^2 = \|x_0 - x^*\|^2$$

and this, together with (17), give rise to the following estimate

$$\frac{2t_{k+1}^2}{L_{k+1}}\left(F(x_{k+1}) - F(x^*)\right) \leq \|x_0 - x^*\|^2 + \sum_{s=1}^{k+1}\frac{t_s}{L_s}\left(\rho_s L_f^2 + \mu_s L_g^2\right). \tag{18}$$

Furthermore, since $t_{k+1} \geq \frac{1}{2} + t_k$ for any $k \geq 1$, it follows that $t_{k+1} \geq \frac{k+2}{2}$, which, along with the fact that $L_k = \frac{1}{\rho_k} + \frac{\|K\|^2}{\mu_k} = (a + b\|K\|^2)k$, lead for any $k \geq 1$ to the following estimate

$$F(x_{k+1}) - F(x^*)$$
$$\leq \frac{2(a + b\|K\|^2)(k+1)}{(k+2)^2}\left(\|x_0 - x^*\|^2 + L_f^2\sum_{s=1}^{k+1}\frac{t_s\rho_s}{L_s} + L_g^2\sum_{s=1}^{k+1}\frac{t_s\mu_s}{L_s}\right)$$
$$\leq \frac{2(a + b\|K\|^2)}{k+2}\|x_0 - x^*\|^2 + \frac{2}{k+2}\sum_{s=1}^{k+1}\frac{t_s}{s^2}\left(\frac{L_f^2}{a} + \frac{L_f^2}{b}\right).$$

Using now that $t_{k+1} \leq 1 + t_k$ for any $k \geq 1$, it yields that $t_{k+1} \leq k + 1$ for any $k \geq 0$, thus

$$\sum_{s=1}^{k+1}\frac{t_s}{s^2} \leq \sum_{s=1}^{k+1}\frac{1}{s} \leq 1 + \sum_{s=2}^{k+1}\int_{s-1}^{s}\frac{1}{x}\,\mathrm{d}x = 1 + \int_1^{k+1}\frac{1}{x}\,\mathrm{d}x = 1 + \ln(k+1).$$

Finally, we obtain that

$$F(x_{k+1}) - F(x^*) \leq \frac{2(a + b\|K\|^2)}{k+2}\|x_0 - x^*\|^2 + \frac{2(1 + \ln(k+1))}{k+2}\left(\frac{L_f^2}{a} + \frac{L_g^2}{b}\right) \ \forall k \geq 1,$$

which concludes the proof. □

In the second part of this subsection we propose a variant of algorithm (A1) formu-

lated with constant smoothing parameters:

$$\boxed{\begin{array}{l}\text{Initialization}: t_1 = 1,\ y_1 = x_0 \in \mathcal{H},\ \rho, \mu \in \mathbb{R}_{++}, \\ \qquad\qquad L(\rho,\mu) = \dfrac{1}{\rho} + \dfrac{\|K\|^2}{\mu} \\ \text{For } k \geq 1: x_k = y_k - \dfrac{1}{L(\rho,\mu)}\left(\text{Prox}_{\frac{1}{\rho}f^*}\left(\dfrac{y_k}{\rho}\right) + K^*\text{Prox}_{\frac{1}{\mu}g^*}\left(\dfrac{Ky_k}{\mu}\right)\right), \\ \qquad\qquad t_{k+1} = \dfrac{1+\sqrt{1+4t_k^2}}{2}, \\ \qquad\qquad y_{k+1} = x_k + \dfrac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})\end{array}} \tag{A2}$$

Constant smoothing parameters have been also used in [11] and [5,6] within the framework of double smoothing algorithms, which assume the regularization in two steps of the Fenchel dual problem to (P) and, consequently, the solving of an unconstrained optimization problem with a strongly convex and differentiable objective function having a Lipschitz continuous gradient.

Theorem 2. *Let $f : \mathcal{H} \to \mathbb{R}$ be a convex and L_f-Lipschitz continuous function, $g : \mathcal{K} \to \mathbb{R}$ a convex and L_g-Lipschitz continuous function, $K : \mathcal{H} \to \mathcal{K}$ a linear continuous operator and $x^* \in \mathcal{H}$ an optimal solution to (P). Then, when choosing for $\varepsilon > 0$*

$$\rho = \frac{2\varepsilon}{3L_f^2} \text{ and } \mu = \frac{2\varepsilon}{3L_g^2},$$

algorithm (A2) *generates a sequence $(x_k)_{k\geq 1} \subseteq \mathcal{H}$ which provides an ε-optimal solution to (P) with a rate of convergence for the objective of order $\mathcal{O}(\frac{1}{k})$.*

Proof. In order to prove this statement, one has only to reproduce the first part of the proof of Theorem 1 when

$$\rho_k = \rho, \mu_k = \mu \text{ and } L_k = L(\rho,\mu) = \frac{1}{\rho} + \frac{\|K\|^2}{\mu}\ \forall k \geq 1,$$

fact which leads to (18). This inequality reads in this particular situation

$$F(x_{k+1}) - F(x^*) \leq \frac{L(\rho,\mu)\|x_0 - x^*\|^2}{2t_{k+1}^2} + \frac{\rho L_f^2 + \mu L_g^2}{2t_{k+1}^2}\sum_{s=1}^{k+1} t_s\ \forall k \geq 1.$$

Since $t_{k+1}^2 = t_k^2 + t_{k+1}$ for any $k \geq 1$, one can inductively prove that $t_{k+1}^2 = \sum_{s=1}^{k+1} t_s$, which, together with the fact that $t_{k+1} \geq \frac{k+2}{2}$ for any $k \geq 1$, yields

$$F(x_{k+1}) - F(x^*) \leq \frac{2L(\rho,\mu)\|x_0 - x^*\|^2}{(k+2)^2} + \frac{\rho L_f^2 + \mu L_g^2}{2}\ \forall k \geq 1.$$

In order to obtain ε-optimality for the objective of the problem (P), where $\varepsilon > 0$ is a given level of accuracy, we choose $\rho = \frac{2\varepsilon}{3L_f^2}$ and $\mu = \frac{2\varepsilon}{3L_g^2}$ and, thus, we have only to force

the first term in the right-hand side of the above estimate to be less than or equal to $\frac{\varepsilon}{3}$. Taking also into account that in this situation $L(\rho,\mu)=\frac{3L_f^2+3L_g^2\|K\|^2}{2\varepsilon}$, it holds

$$\begin{aligned}
&\frac{\varepsilon}{3} \geq \frac{2L(\rho,\mu)\left\|x_0-x^*\right\|^2}{(k+2)^2} = \frac{3\left(L_f^2+L_g^2\|K\|^2\right)\left\|x_0-x^*\right\|^2}{\varepsilon(k+2)^2} \\
\Leftrightarrow\ &\frac{\varepsilon^2}{9} \geq \frac{\left(L_f^2+L_g^2\|K\|^2\right)\left\|x_0-x^*\right\|^2}{(k+2)^2} \\
\Leftrightarrow\ &\frac{\varepsilon}{3} \geq \frac{\sqrt{L_f^2+L_g^2\|K\|^2}\left\|x_0-x^*\right\|}{k+2},
\end{aligned}$$

which shows that an ε-optimal solution to (P) can be provided with a rate of convergence for the objective of order $\mathcal{O}(\frac{1}{k})$. □

The rate of convergence of algorithm (A1) may not be as good as the one proved for the algorithm with constant smoothing parameters depending on a fixed level of accuracy $\varepsilon > 0$. However, the main advantage of the variable smoothing methods is given by the fact that the sequence of objective values $(f(x_k)+g(Kx_k))_{k\geq 1}$ converges to the optimal objective value of (P), whereas, when generated by algorithm (A2), despite of the fact that it approximates the optimal objective value with a better convergence rate, this sequence may not converge to this.

3.3 The case when f is differentiable with Lipschitz continuous gradient

In this subsection we show how the algorithms (A1) and (A2) for solving the problem (P) can be adapted to the situation when f is a differentiable function with Lipschitz continuous gradient. We provide iterative schemes with variable and constant smoothing variables and corresponding convergence statements. More precisely, we deal with the optimization problem

$$(P) \qquad \inf_{x\in\mathcal{H}}\{f(x)+g(Kx)\},$$

where $K:\mathcal{H}\to\mathcal{K}$ is a linear continuous operator, $f:\mathcal{H}\to\mathbb{R}$ is a convex and differentiable function with $L_{\nabla f}$-Lipschitz continuous gradient and $g:\mathcal{K}\to\mathbb{R}$ is a convex and L_g-Lipschitz continuous function.

Algorithm (A1) can be adapted to this framework as follows:

$$\boxed{\begin{aligned}
&\text{Initialization}: t_1=1,\ y_1=x_0\in\mathcal{H},\ (\mu_k)_{k\geq 1}\subseteq\mathbb{R}_{++} \\
&\quad \text{For } k\geq 1: L_k = L_{\nabla f}+\frac{\|K\|^2}{\mu_k}, \\
&\qquad\qquad x_k = y_k-\frac{1}{L_k}\left(\nabla f(y_k)+K^*\mathrm{Prox}_{\frac{1}{\mu_k}g^*}\left(\frac{Ky_k}{\mu_k}\right)\right), \\
&\qquad\qquad t_{k+1}=\frac{1+\sqrt{1+4t_k^2}}{2}, \\
&\qquad\qquad y_{k+1}=x_k+\frac{t_k-1}{t_{k+1}}(x_k-x_{k-1})
\end{aligned}} \tag{A3}$$

while its convergence is furnished by the following theorem.

Theorem 3. *Let $f : \mathcal{H} \to \mathbb{R}$ be a convex and differentiable function with $L_{\nabla f}$-Lipschitz continuous gradient, $g : \mathcal{K} \to \mathbb{R}$ a convex and L_g-Lipschitz continuous function, $K : \mathcal{H} \to \mathcal{K}$ a nonzero linear continuous operator and $x^* \in \mathcal{H}$ an optimal solution to (P). Then, when choosing*

$$\mu_k = \frac{1}{bk} \ \forall k \geq 1,$$

where $b \in \mathbb{R}_{++}$, algorithm (A3) *generates a sequence $(x_k)_{k\geq 1} \subseteq \mathcal{H}$ satisfying for any $k \geq 1$*

$$F(x_{k+1}) - F(x^*) \leq \frac{2(L_{\nabla f} + b\|K\|^2)}{k+2}\|x_0 - x^*\|^2 + \frac{2(1+\ln(k+1))}{k+2}\frac{L_g^2(L_{\nabla f} + b\|K\|^2)}{b^2\|K\|^2}, \tag{19}$$

thus yielding a rate of convergence for the objective of order $\mathcal{O}(\frac{\ln k}{k})$.

Proof. For any $k \geq 1$ we denote by $F^k : \mathcal{H} \to \mathbb{R}$, $F^k(x) = f(x) + {}^{\mu_k}g(Kx)$. For any $k \geq 1$ and every $x \in \mathcal{H}$ it holds $\nabla F^k(x) = \nabla f(x) + K^*\mathrm{Prox}_{\frac{1}{\mu_k}g^*}\left(\frac{Kx}{\mu_k}\right)$ and ∇F^k is L_k-Lipschitz continuous, where $L_k = L_{\nabla f} + \frac{\|K\|^2}{\mu_k}$.

As in the proof of Theorem 1, by defining $p_k := (t_k - 1)(x_{k-1} - x_k)$, we obtain for any $k \geq 1$

$$\begin{aligned}
&\|p_{k+1} - x_{k+1} + x^*\|^2 - \|p_k - x_k + x^*\|^2 \\
&\leq \frac{2t_k^2}{L_{k+1}}\left(F^{k+1}(x_k) - F^{k+1}(x^*)\right) - \frac{2t_{k+1}^2}{L_{k+1}}(F^{k+1}(x_{k+1}) - F^{k+1}(x^*)) \\
&\leq \frac{2t_k^2}{L_{k+1}}\left(F^k(x_k) - F^{k+1}(x^*) + (\mu_k - \mu_{k+1})\frac{L_g^2}{2}\right) - \frac{2t_{k+1}^2}{L_{k+1}}(F^{k+1}(x_{k+1}) - F^{k+1}(x^*)) \\
&\leq \frac{2t_k^2}{L_{k+1}}\left(F^k(x_k) - F^k(x^*) + \mu_k\frac{L_g^2}{2}\right) - \frac{2t_{k+1}^2}{L_{k+1}}(F^{k+1}(x_{k+1}) - F^{k+1}(x^*)) - \frac{t_k^2}{L_{k+1}}\mu_{k+1}L_g^2 \\
&\leq \frac{2t_k^2}{L_k}\left(F^k(x_k) - F^k(x^*) + \mu_k\frac{L_g^2}{2}\right) - \frac{2t_{k+1}^2}{L_{k+1}}(F^{k+1}(x_{k+1}) - F^{k+1}(x^*)) - \frac{t_k^2}{L_{k+1}}\mu_{k+1}L_g^2 \\
&= \frac{2t_k^2}{L_k}\left(F^k(x_k) - F^k(x^*) + \mu_k\frac{L_g^2}{2}\right) - \frac{2t_{k+1}^2}{L_{k+1}}(F^{k+1}(x_{k+1}) - F^{k+1}(x^*)) \\
&- \frac{t_{k+1}^2L_g^2}{L_{k+1}}\mu_{k+1} + \frac{t_{k+1}L_g^2}{L_{k+1}}\mu_{k+1}
\end{aligned}$$

and, consequently,

$$\begin{aligned}
&\|p_{k+1} - x_{k+1} + x^*\|^2 + \frac{2t_{k+1}^2}{L_{k+1}}\left(F^{k+1}(x_{k+1}) - F^{k+1}(x^*) + \mu_{k+1}\frac{L_g^2}{2}\right) \\
&\leq \|p_k - x_k + x^*\|^2 + \frac{2t_k^2}{L_k}\left(F^k(x_k) - F^k(x^*) + \mu_k\frac{L_g^2}{2}\right) + \frac{t_{k+1}L_g^2}{L_{k+1}}\mu_{k+1}.
\end{aligned}$$

For any $k \geq 1$ it holds

$$\begin{aligned}
&\frac{2t_{k+1}^2}{L_{k+1}} \left(F(x_{k+1}) - F(x^*)\right) \\
&\leq \frac{2t_{k+1}^2}{L_{k+1}} \left(F^{k+1}(x_{k+1}) - F^{k+1}(x^*) + \mu_{k+1}\frac{L_g^2}{2}\right) + \|p_{k+1} - x_{k+1} + x^*\|^2 \\
&\leq \frac{2t_1^2}{L_1} \left(F^1(x_1) - F^1(x^*) + \mu_1\frac{L_g^2}{2}\right) + \|p_1 - x_1 + x^*\|^2 \\
&+ \sum_{s=1}^{k} \frac{t_{s+1}L_g^2}{L_{s+1}}\mu_{s+1},
\end{aligned}$$

which yields

$$\frac{2t_{k+1}^2}{L_{k+1}} \left(F(x_{k+1}) - F(x^*)\right) \leq \|x_0 - x^*\|^2 + \sum_{s=1}^{k+1} \frac{t_s L_g^2}{L_s}\mu_s. \tag{20}$$

For any $k \geq 1$, since $t_{k+1} \geq \frac{k+2}{2}$ and $L_k = L_{\nabla f} + \frac{\|K\|^2}{\mu_k} = L_{\nabla f} + b\|K\|^2 k$, it follows

$$\begin{aligned}
&F(x_{k+1}) - F(x^*) \\
&\leq \frac{2(L_{\nabla f} + b\|K\|^2 (k+1))}{(k+2)^2} \left(\|x_0 - x^*\|^2 + \sum_{s=1}^{k+1} \frac{t_s L_g^2}{(L_{\nabla f} + b\|K\|^2 s)sb}\right).
\end{aligned}$$

Thus, for any $k \geq 1$, since $t_k \leq k$, it yields

$$\begin{aligned}
&F(x_{k+1}) - F(x^*) \\
&\leq \frac{2(L_{\nabla f} + b\|K\|^2 (k+1))}{(k+2)^2} \left(\|x_0 - x^*\|^2 + \sum_{s=1}^{k+1} \frac{L_g^2}{(L_{\nabla f} + b\|K\|^2 s)b}\right) \\
&\leq \frac{2(L_{\nabla f} + b\|K\|^2 (k+1))}{(k+2)^2} \left(\|x_0 - x^*\|^2 + \sum_{s=1}^{k+1} \frac{L_g^2}{b^2\|K\|^2 s}\right) \\
&\leq \frac{2(L_{\nabla f} + b\|K\|^2 (k+1))}{(k+2)^2} \left(\|x_0 - x^*\|^2 + \frac{L_g^2}{b^2\|K\|^2}(1 + \ln(k+1))\right) \\
&\leq \frac{2(L_{\nabla f} + b\|K\|^2)}{k+2} \left(\|x_0 - x^*\|^2 + \frac{L_g^2}{b^2\|K\|^2}(1 + \ln(k+1))\right) \\
&\leq \frac{2(L_{\nabla f} + b\|K\|^2)}{k+2} \|x_0 - x^*\|^2 + \frac{2(1 + \ln(k+1))}{k+2} \frac{L_g^2(L_{\nabla f} + b\|K\|^2)}{b^2\|K\|^2}.
\end{aligned}$$

□

By adapting (A3) to the framework considered in this subsection we obtain the

following algorithm with constant smoothing variables:

$$\boxed{\begin{aligned}&\text{Initialization}: t_1 = 1,\ y_1 = x_0 \in \mathcal{H},\ \mu \in \mathbb{R}_{++},\\ &L(\mu) = L_{\nabla f} + \frac{\|K\|^2}{\mu}\\ &\text{For } k \geq 1: x_k = y_k - \frac{1}{L(\mu)}\left(\nabla f(y_k) + K^*\text{Prox}_{\frac{1}{\mu}g^*}\left(\frac{Ky_k}{\mu}\right)\right),\\ &t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2},\\ &y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})\end{aligned}} \tag{A4}$$

The convergence of algorithm (A4) is stated by the following theorem, which can be proved in the lines of the proof of Theorem 3.

Theorem 4. *Let $f : \mathcal{H} \to \mathbb{R}$ be a convex and differentiable function with $L_{\nabla f}$-Lipschitz continuous gradient, $g : \mathcal{K} \to \mathbb{R}$ a convex and L_g-Lipschitz continuous function, $K : \mathcal{H} \to \mathcal{K}$ a nonzero linear continuous operator and $x^* \in \mathcal{H}$ an optimal solution to (P). Then, when choosing for $\varepsilon > 0$*

$$\mu = \frac{\varepsilon}{L_g^2},$$

algorithm (A4) *generates a sequence $(x_k)_{k\geq 1} \subseteq \mathcal{H}$ which provides an ε-optimal solution to (P) with a rate of convergence for the objective of order $\mathcal{O}(\frac{1}{k})$.*

3.4 The optimization problem with the sum of more than two functions in the objective

We close this section by discussing the employment of the algorithmic schemes presented in the previous two subsections to the optimization problem (2)

$$\inf_{x\in\mathcal{H}}\left\{f(x) + \sum_{i=1}^{m} g_i(K_i x)\right\},$$

where $\mathcal{H}$ and $\mathcal{K}_i$, $i = 1, ..., m$, are real Hilbert spaces, $f : \mathcal{H} \to \mathbb{R}$ is a convex and either L_f-Lipschitz continuous or differentiable with $L_{\nabla f}$-continuous gradient function, $g_i : \mathcal{K}_i \to \mathbb{R}$ are convex and L_{g_i}-Lipschitz continuous functions and $K_i : \mathcal{H} \to \mathcal{K}_i$, $i = 1, ..., m$, are linear continuous operators. By endowing $\mathcal{K} := \mathcal{K}_1 \times ... \times \mathcal{K}_m$ with the inner product defined as

$$\langle y, z\rangle = \sum_{i=1}^{m}\langle y_i, z_i\rangle\ \forall y, z \in \mathcal{K},$$

and with the corresponding norm and by defining $g : \mathcal{K} \to \mathbb{R}, g(y_1, ..., y_m) = \sum_{i=1}^m g_i(y_i)$ and $K : \mathcal{H} \to \mathcal{K}, Kx = (K_1x, ..., K_mx)$, problem (2) can be equivalently written as

$$\inf_{x\in\mathcal{H}}\{f(x) + g(Kx)\}$$

and, consequently, solved via one of the variable or constant smoothing algorithms introduced in the subsections 3.2 and 3.3, depending on the properties the function f is endowed with.

In the following we determine the elements related to the above constructed function g which appear in these iterative schemes and in the corresponding convergence statements. Obviously, the function g is convex and, since for every $(y_1, ..., y_m), (z_1, ..., z_m) \in \mathcal{K}$

$$|g(y_1, ..., y_m) - g(z_1, ..., z_m)| \leq \sum_{i=1}^{m} L_{g_i} \|y_i - z_i\| \leq \left(\sum_{i=1}^{m} L_{g_i}^2 \right)^{\frac{1}{2}} \|(y_1, ..., y_m) - (z_1, ..., z_m)\|,$$

it is $\left(\sum_{i=1}^{m} L_{g_i}^2\right)^{\frac{1}{2}}$-Lipschitz continuous. On the other hand, for each $\mu \in \mathbb{R}_{++}$ and $(y_1, ..., y_m) \in \mathcal{K}$ it holds

$$^{\mu}g(y_1, ..., y_m) = \sum_{i=1}^{m} {}^{\mu}g_i(y_i),$$

thus

$$\begin{aligned} \nabla(^{\mu}g)(y_1, ..., y_m) &= (\nabla(^{\mu}g_1)(y_1), ..., \nabla(^{\mu}g_m)(y_m)) \\ &= \left(\mathrm{Prox}_{\frac{1}{\mu}g_1^*}\left(\frac{y_1}{\mu}\right), ..., \mathrm{Prox}_{\frac{1}{\mu}g_m^*}\left(\frac{y_m}{\mu}\right) \right). \end{aligned}$$

Since $K^*(y_1, ..., y_m) = \sum_{i=1}^{m} K_i^* y_i$, for every $(y_1, ..., y_m) \in \mathcal{K}$, we have

$$\begin{aligned} \nabla(^{\mu}g \circ K)(x) = K^* \nabla(^{\mu}g)(K_1 x, ..., K_m x) &= \sum_{i=1}^{m} K_i^* \nabla(^{\mu}g_i)(K_i x) \\ &= \sum_{i=1}^{m} K_i^* \mathrm{Prox}_{\frac{1}{\mu}g_i^*}\left(\frac{K_i x}{\mu}\right) \ \forall x \in \mathcal{H}. \end{aligned}$$

Finally, we notice that for arbitrary $x, y \in \mathcal{H}$ one has

$$\begin{aligned} \|\nabla(^{\mu}g \circ K)(x) - \nabla(^{\mu}g \circ K)(y)\| &= \left\| \sum_{i=1}^{m} K_i^* \nabla(^{\mu}g_i)(K_i x) - \sum_{i=1}^{m} K_i^* \nabla(^{\mu}g_i)(K_i y) \right\| \\ &\leq \sum_{i=1}^{m} \|K_i\| \, \|\nabla(^{\mu}g_i)(K_i x) - \nabla(^{\mu}g_i)(K_i y)\| \\ &\leq \sum_{i=1}^{m} \frac{\|K_i\|}{\mu} \|K_i x - K_i y\| \leq \frac{\sum_{i=1}^{m} \|K_i\|^2}{\mu} \|x - y\|, \end{aligned}$$

which shows that the Lipschitz constant of $\nabla(^{\mu}g \circ K)$ is $\frac{\sum_{i=1}^{m} \|K_i\|^2}{\mu}$.

4 Numerical experiments

4.1 Image processing

The first numerical experiment involving the variable smoothing algorithm concerns the solving of an extremely ill-conditioned linear inverse problem which arises in the field

of signal and image processing, by basically solving the regularized nondifferentiable convex optimization problem

$$\inf_{x\in\mathbb{R}^n}\{\|Ax-b\|_1+\lambda\|Wx\|_1\}, \tag{21}$$

where $b\in\mathbb{R}^n$ is the blurred and noisy image, $A:\mathbb{R}^n\to\mathbb{R}^n$ is a blurring operator, $W:\mathbb{R}^n\to\mathbb{R}^n$ is the discrete Haar wavelet transform with four levels and $\lambda>0$ is the regularization parameter. The blurring operator is constructed by making use of the Matlab routines `imfilter` and `fspecial` as follows:

```
H=fspecial('gaussian',9,4); % gaussian blur of size 9 times 9
                            % and standard deviation 4
B=imfilter(X,H,'conv','symmetric'); % B=observed blurred image
                                    % X=original image
```

The function `fspecial` returns a rotationally symmetric Gaussian lowpass filter of size 9×9 with standard deviation 4, the entries of H being nonnegative and their sum adding up to 1. The function `imfilter` convolves the filter H with the image X and furnishes the blurred image B. The boundary option "symmetric" corresponds to reflexive boundary conditions. Thanks to the rotationally symmetric filter H, the linear operator A defined via the routine `imfilter` is symmetric, too. By making use of the real spectral decomposition of A, it shows that $\|A\|^2=1$. Furthermore, since W is an orthogonal wavelet, it holds $\|W\|^2=1$.

The optimization problem (21) can be written as

$$\inf_{x\in\mathbb{R}^n}\{f(x)+g_1(Ax)+g_2(Wx)\},$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is taking to be $f\equiv 0$ with the Lipschitz constant of its gradient $L_{\nabla f}=0$, $g_1:\mathbb{R}^n\to\mathbb{R}$, $g_1(y)=\|y-b\|_1$ is convex and $\sqrt{n}$-Lipschitz continuous and $g_2:\mathbb{R}^n\to\mathbb{R}$, $g_2(y)=\lambda\|y\|_1$ is convex and $\lambda\sqrt{n}$-Lipschitz continuous. For every $p\in\mathbb{R}^n$ it holds $g_1^*(p)=\delta_{[-1,1]^n}(p)+p^Tb$ and $g_2^*(p)=\delta_{[-\lambda,\lambda]^n}(p)$ (see, for instance, [3]). We solved this problem, by using also the considerations made in Subsection 3.4, with algorithm (A3) and computed to this aim for $\mu\in\mathbb{R}_{++}$ and $x\in\mathbb{R}^n$

$$\begin{aligned}\mathrm{Prox}_{\frac{1}{\mu}g_1^*}\left(\frac{Ax}{\mu}\right)&=\operatorname*{arg\,min}_{p\in\mathbb{R}^n}\left\{\frac{1}{\mu}g_1^*(p)+\frac{1}{2}\left\|\frac{Ax}{\mu}-p\right\|^2\right\}=\operatorname*{arg\,min}_{p\in[-1,1]^n}\left\{\frac{1}{\mu}p^Tb+\frac{1}{2}\left\|\frac{Ax}{\mu}-p\right\|^2\right\}\\&=\operatorname*{arg\,min}_{p\in[-1,1]^n}\left\{\frac{1}{2}\left\|\frac{Ax}{\mu}-p\right\|^2-\left(\frac{Ax}{\mu}-p\right)^T\frac{b}{\mu}+\frac{\|b\|^2}{2\mu^2}-\frac{\|b\|^2}{2\mu^2}+\frac{(Ax)^Tb}{\mu^2}\right\}\\&=\operatorname*{arg\,min}_{p\in[-1,1]^n}\left\{\frac{1}{2}\left\|\frac{Ax-b}{\mu}-p\right\|^2\right\}-\frac{\|b\|^2}{2\mu^2}+\frac{(Ax)^Tb}{\mu^2}=\mathcal{P}_{[-1,1]^n}\left(\frac{Ax-b}{\mu}\right)\end{aligned}$$

and

$$\begin{aligned}\mathrm{Prox}_{\frac{1}{\mu}g_2^*}\left(\frac{Wx}{\mu}\right)=\operatorname*{arg\,min}_{p\in\mathbb{R}^n}\left\{\frac{1}{\mu}g_2^*(p)+\frac{1}{2}\left\|\frac{Wx}{\mu}-p\right\|^2\right\}&=\operatorname*{arg\,min}_{p\in[-\lambda,\lambda]^n}\frac{1}{2}\left\|\frac{Wx}{\mu}-p\right\|^2\\&=\mathcal{P}_{[-\lambda,\lambda]^n}\left(\frac{Wx}{\mu}\right).\end{aligned}$$

Hence, choosing $\mu_k = \frac{1}{ak}$, for some parameter $a \in \mathbb{R}_{++}$ and taking into account that $L_k = \frac{\|A\|^2+\|W\|^2}{\mu_k} = 2ak$, for $k \geq 1$, the iterative scheme (A3) with starting point $b \in \mathbb{R}^n$ becomes

$$\begin{aligned}
&\text{Initialization}: t_1 = 1,\ y_1 = x_0 = b \in \mathbb{R}^n,\ a > 0,\\
&\quad \text{For } k \geq 1: \mu_k = \frac{1}{ak},\ L_k = 2ak,\\
&\qquad x_k = y_k - \frac{1}{L_k}\left(A\mathcal{P}_{[-1,1]^n}\left(\frac{Ay_k - b}{\mu_k}\right) + W\mathcal{P}_{[-\lambda,\lambda]^n}\left(\frac{Wy_k}{\mu_k}\right)\right),\\
&\qquad t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2},\\
&\qquad y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})
\end{aligned}$$

We considered the 256×256 cameraman test image, which is part of the image processing toolbox in Matlab, that we vectorized (to a vector of dimension $n = 256^2 = 65536$) and normalized, in order to make pixels range in the closed interval from 0 (pure black) to 1 (pure white). In addition, we added normally distributed white Gaussian noise with standard deviation 10^{-3} and set the regularization parameter to $\lambda =$ 2e-5. The original and observed images are shown in Figure 4.1. When measuring the quality of

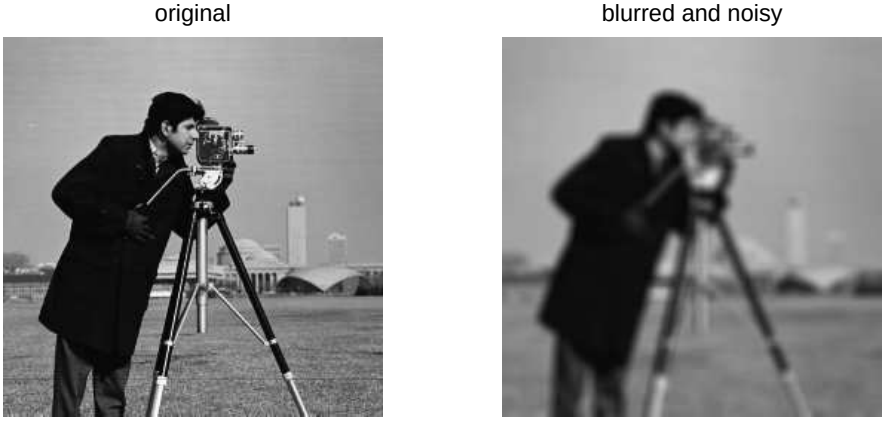

Figure 4.1: The 256 × 256 cameraman test image

the restored images, we made use of the *improvement in signal-to-noise ratio (ISNR)*, which is defined as

$$\text{ISNR}_k = 10\log_{10}\left(\frac{\|x-b\|^2}{\|x-x_k\|^2}\right),$$

where x, b and x_k denote the original, the observed and the estimated image at iteration $k \geq 1$, respectively. We tested several values for $a \in \mathbb{R}_{++}$ and we obtained after 100 iterations the objective values and the ISNR values presented in Table 4.1. In the context of solving the problem (21) we compared the variable smoothing approach (VS) for $a =$ 1e-1 with the operator-splitting algorithm based on skew splitting (SS) proposed in [8, 10] with parameters $\varepsilon = \frac{1}{2(\sqrt{2}+1)}$ and $\gamma_k = \gamma = \frac{\varepsilon}{2} + \frac{1-\varepsilon}{2\sqrt{2}}$, for any $k \geq 1$,

a	1e-4	1e-3	1e-2	1e-1	1	1e+1	1e+2	1e+3
fval	164.621	80.915	55.763	53.669	53.579	63.754	208.413	531.022
ISNR	1.282	3.839	5.241	5.352	5.337	4.351	1.180	0.199

Table 4.1: Objective values (fval) and ISNR values (higher is better) after 100 iterations.

and with the primal-dual algorithm (PD) from [9] with parameters $\theta = 1$, $\sigma = 0.01$ and $\tau = 49.999$. The parameters considered for the three approaches provide the best results when solving (21). The output of these three algorithms after 100 iterations,

Figure 4.2: Results furnished by the primal-dual (PD), the skew splitting (SS) and the variable smoothing (VS) algorithms after 100 iterations.

along with the corresponding objective values, can be seen in Figure 4.2 and they show that the variable smoothing approach outperforms the other two methods. Figure 4.3 shows the evolution of the values of the objective function and of the improvement in signal-to-noise ratio within the first 100 iterations.

4.2 Support vector machines classification

The second numerical experiment we consider for the variable smoothing algorithm concerns the solving of the problem of classifying images via support vector machines classification, an approach which belong to the class of kernel based learning methods.

The given data set consisting of 5268 images of size 200×50 was taken from a real-world problem a supplier of the automotive industry was faced with by establishing a computer-aided quality control for manufactured devices at the end of the manufacturing process (see [4] for more details on this data set). The overall task is to classify fine and defective components which are labeled by $+1$ and -1, respectively.

The classifier functional `f` is assumed to be an element of the *Reproducing Kernel Hilbert Space (RHKS)* $\mathcal{H}_\kappa$, which in our case is induced by the symmetric and finitely positive definite Gaussian kernel function

$$\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R},\ \kappa(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right).$$

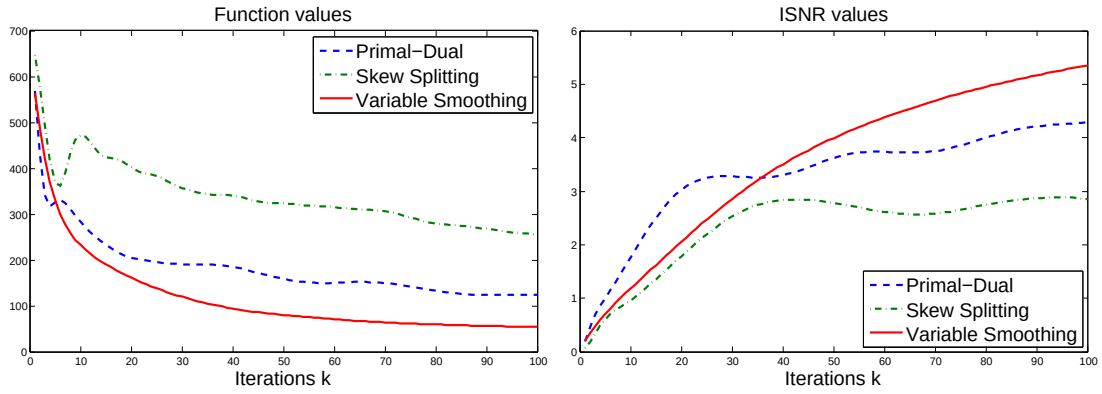

Figure 4.3: The evolution of the values of the objective function and of the ISNR for the primal-dual (PD), the skew splitting (SS) and the variable smoothing (VS) algorithms after 100 iterations.

Let $\langle \cdot, \cdot \rangle_\kappa$ denote the inner product on $\mathcal{H}_\kappa$, $\| \cdot \|_\kappa$ the corresponding norm and $K \in \mathbb{R}^{n \times n}$ the *Gram matrix* with respect to the training data set $\mathcal{Z} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \subseteq \mathbb{R}^d \times \{+1, -1\}$, namely the symmetric and positive definite matrix with entries $K_{ij} = \kappa(X_i, X_j)$ for $i, j = 1, \ldots, n$. Within this example we make use of the *hinge loss* $v : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $v(x, y) = \max\{1 - xy, 0\}$, which penalizes the deviation between the predicted value $\mathtt{f}(x)$ and the true value $y \in \{+1, -1\}$. The smoothness of the decision function $\mathtt{f} \in \mathcal{H}_\kappa$ is employed by means of the *smoothness functional* $\Omega : \mathcal{H}_\kappa \to \mathbb{R}$, $\Omega(f) = \|\mathtt{f}\|_\kappa^2$, taking high values for non-smooth functions and low values for smooth ones. The decision function $\mathtt{f}$ we are looking for is the optimal solution of the *Tikhonov regularization problem*

$$\inf_{\mathtt{f} \in \mathcal{H}_\kappa} \left\{ \frac{1}{2} \Omega(\mathtt{f}) + C \sum_{i=1}^{n} v(\mathtt{f}(X_i), Y_i) \right\}, \tag{22}$$

where $C > 0$ denotes the regularization parameter controlling the tradeoff between the loss function and the smoothness functional.

The *representer theorem* (cf. [18]) ensures the existence of a vector of coefficients $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ such that the minimizer $\mathtt{f}$ of (22) can be expressed as a kernel expansion in terms of the training data, i.e., $\mathtt{f}(\cdot) = \sum_{i=1}^n c_i \kappa(\cdot, X_i)$. Thus, the smoothness functional becomes $\Omega(\mathtt{f}) = \|\mathtt{f}\|_\kappa^2 = \langle \mathtt{f}, \mathtt{f} \rangle_\kappa = \sum_{i=1}^n \sum_{j=1}^n c_i c_j \kappa(X_i, X_j) = c^T K c$ and for $i = 1, \ldots, n$, it holds $\mathtt{f}(X_i) = \sum_{j=1}^n c_j \kappa(X_i, X_j) = (Kc)_i$. Hence, in order to determine the decision function one has to solve the convex optimization problem

$$\inf_{c \in \mathbb{R}^n} \left\{ f(c) + C \sum_{i=1}^{n} g_i(Kc) \right\}, \tag{23}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $f(c) = \frac{1}{2} c^T K c$, and $g_i : \mathbb{R}^n \to \mathbb{R}$, $g_i(c) = C v(c_i, Y_i)$ for $i = 1, \ldots, n$. The function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable and it fulfills $\nabla f(c) = Kc$ for every $c \in \mathbb{R}^n$, thus ∇f is Lipschitz continuous with Lipschitz constant $L_{\nabla f} = \|K\|$. For any $i = 1, ..., n$ the function $g_i : \mathbb{R}^n \to \mathbb{R}$ is convex and C-Lipschitz continuous, properties which allowed us to solve the problem (23) with algorithm (A3), by using

also the considerations made in Subsection 3.4. For any $i = 1, ..., n$ and every $p = (p_1, ..., p_n)^T \in \mathbb{R}^n$ it holds (see, also, [4, 7])

$$\begin{aligned} g_i^*(p) &= \sup_{c\in\mathbb{R}^n} \{\langle p, c\rangle - Cv(c_i, Y_i)\} = C \sup_{c\in\mathbb{R}^n} \left\{ \left\langle \frac{p}{C}, c \right\rangle - v(c_i, Y_i) \right\} \\ &= \begin{cases} C(v(\cdot, Y_i))^* \left(\dfrac{p_i}{C}\right), \text{ if } p_j = 0,\ i \neq j, \\ +\infty, \text{ otherwise,} \end{cases} \\ &= \begin{cases} p_i Y_i, \text{ if } p_j = 0,\ i \neq j \text{ and } p_i Y_i \in [-C, 0], \\ +\infty, \text{ otherwise.} \end{cases} \end{aligned}$$

Thus, for $\mu \in \mathbb{R}_{++}$, $c = (c_1, ..., c_n)^T$ and $i = 1, ..., n$ we have

$$\begin{aligned} \text{Prox}_{\frac{1}{\mu} g_i^*} \left(\frac{c}{\mu}\right) &= \arg\min_{p\in\mathbb{R}^n} \left\{ \frac{1}{\mu} g_i^*(p) + \frac{1}{2} \left\| \frac{c}{\mu} - p \right\|^2 \right\} \\ &= \arg\min_{\substack{p_i Y_i \in [-C,0] \\ p_j = 0, j\neq i}} \left\{ \frac{p_i Y_i}{\mu} + \frac{1}{2} \left(\frac{c_i}{\mu} - p_i \right)^2 \right\} \\ &= \arg\min_{\substack{p_i Y_i \in [-C,0] \\ p_j = 0, j\neq i}} \left\{ p_i Y_i + \frac{\mu}{2} \left(\frac{c_i}{\mu} - p_i \right)^2 \right\}. \end{aligned}$$

For $Y_i = 1$ we have

$$\text{Prox}_{\frac{1}{\mu} g_i^*} \left(\frac{c}{\mu}\right) = \arg\min_{\substack{p_i Y_i \in [-C,0] \\ p_j = 0, j\neq i}} \left\{ p_i + \frac{\mu}{2} \left(\frac{c_i}{\mu} - p_i \right)^2 \right\} = \left(0, \ldots, \mathcal{P}_{[-C,0]} \left(\frac{c_i - 1}{\mu} \right), \ldots, 0\right)^T,$$

while for $Y_i = -1$, it holds

$$\text{Prox}_{\frac{1}{\mu} g_i^*} \left(\frac{c}{\mu}\right) = \arg\min_{\substack{p_i Y_i \in [-C,0] \\ p_j = 0, j\neq i}} \left\{ -p_i + \frac{\mu}{2} \left(\frac{c_i}{\mu} - p_i \right)^2 \right\} = \left(0, \ldots, \mathcal{P}_{[0,C]} \left(\frac{c_i + 1}{\mu} \right), \ldots, 0\right)^T.$$

Summarizing, it follows

$$\text{Prox}_{\frac{1}{\mu} g_i^*} \left(\frac{c}{\mu}\right) = \left(0, \ldots, \mathcal{P}_{Y_i[-C,0]} \left(\frac{c_i - Y_i}{\mu} \right), \ldots, 0\right)^T.$$

Thus, for every $c = (c_1, ..., c_n)^T$ we have

$$\begin{aligned} \nabla \left(\sum_{i=1}^n ({}^{\mu} g_i \circ K) \right)(c) &= \sum_{i=1}^n \nabla ({}^{\mu} g_i \circ K)(c) = \sum_{i=1}^n K \text{Prox}_{\frac{1}{\mu} g_i^*} \left(\frac{Kc}{\mu} \right) \\ &= K \left(\mathcal{P}_{Y_1[-C,0]} \left(\frac{(Kc)_1 - Y_1}{\mu} \right), ..., \mathcal{P}_{Y_n[-C,0]} \left(\frac{(Kc)_n - Y_n}{\mu} \right) \right)^T. \end{aligned}$$

Using the nonexpansiveness of the projection operator, we obtain for every $c, d \in \mathbb{R}^n$

$$\left\| \nabla \left(\sum_{i=1}^n (g_i^{\mu} \circ K) \right)(c) - \nabla \left(\sum_{i=1}^n (g_i^{\mu} \circ K) \right)(d) \right\| \leq \|K\| \left\| \frac{Kc - Kd}{\mu} \right\| \leq \frac{\|K\|^2}{\mu} \|c - d\|.$$

Choosing $\mu_k = \frac{1}{ak}$, for some parameter $a \in \mathbb{R}_{++}$ and taking into account that $L_k = \|K\| + ak\|K\|^2$, for $k \geq 1$, the iterative scheme (A3) with starting point $x_0 = 0 \in \mathbb{R}^n$ becomes

Initialization : $t_1 = 1,\ y_1 = x_0 = 0 \in \mathbb{R}^n,\ a \in \mathbb{R}_{++}$,

For $k \geq 1$: $\mu_k = \dfrac{1}{ak},\ L_k = \|K\| + ak\|K\|^2$,

$$x_k = y_k - \frac{1}{L_k}\left(Ky_k + K\left(\mathcal{P}_{Y_i[-C,0]}\left(\frac{(Kc)_i - Y_i}{\mu}\right)\right)^T_{i=\overline{1,n}}\right),$$

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2},$$

$$y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})$$

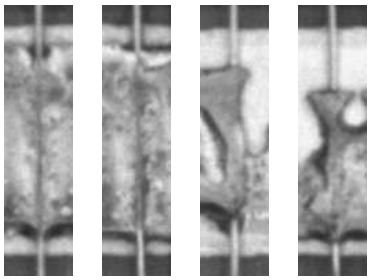

Figure 4.4: Example of two fine and two defective devices.

Coming to the real-data set, we denote by $\mathcal{D} = \{(X_i, Y_i), i = 1, \ldots, 5268\} \subseteq \mathbb{R}^{10000} \times \{+1, -1\}$ the set of all data available consisting of 2682 images of class $+1$ and 2586 images of class -1. Notice that two examples of each class are shown in Figure 4.4. Due to numerical reasons, the images have been normalized (cf. [12]) by dividing each of them by the quantity $\left(\frac{1}{5268}\sum_{i=1}^{5268}\|X_i\|^2\right)^{\frac{1}{2}}$. We considered as regularization parameter

a	1e-5	1e-4	1e-3	1e-2	1e-1	1	1e+1	1e+2	1e+3
err	0.4176	0.3037	0.2278	0.2468	0.3986	0.5315	0.5125	1.5945	48.9561

Table 4.2: Average classification errors in percentage.

$C = 100$ and as kernel parameter $\sigma = 0.5$, which are the optimal values reported in [4] for this data set from a given pool of parameter combinations, tested different values for $a \in \mathbb{R}_{++}$ and performed for each of those choices a 10-fold cross validation on $\mathcal{D}$. We terminated the algorithm after a fixed number of 10000 iterations was reached, the average classification errors being presented in Table 4.2. For $a =$ 1e-3 we obtained the lowest missclassification rate of 0.2278 percentage. In other words, from 527 images belonging to the test data set an average of 1.2 were not correctly classified.

References

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

[2] J.M. Borwein and J.D. Vanderwerff. *Convex Functions: Constructions, Characterizations and Counterexamples*. Cambridge University Press, 2010.

[3] R.I. Boţ. *Conjugate Duality in Convex Optimization*. Lecture Notes in Economics and Mathematical Systems, Vol. 637, Springer-Verlag Berlin Heidelberg, 2010.

[4] R.I. Boţ, A. Heinrich and G. Wanka. Employing different loss functions for the classification of images via supervised learning. *Preprint, Chemnitz University of Technology, Faculty of Mathematics*, 2012.

[5] R.I. Boţ and C. Hendrich. A double smoothing technique for solving unconstrained nondifferentiable convex optimization problems. *arXiv:1203.2070v1 [math.OC]*, 2012.

[6] R.I. Boţ and C. Hendrich. On the acceleration of the double smoothing technique for unconstrained convex optimization problems. *arXiv:1205.0721v1 [math.OC]*, 2012.

[7] R.I. Boţ and N. Lorenz. Optimization problems in statistical learning: Duality and optimality conditions. *European Journal of Operational Research*, 213(2):395–404, 2011.

[8] L.M. Briceño-Arias and P.L. Combettes. A Monotone + Skew Splitting Model for Composite Monotone Inclusions in Duality. *SIAM Journal on Optimization*, 21(4):1230–1250, 2011.

[9] A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[10] P.L. Combettes and J.-C. Pesquet. Primal-Dual Splitting Algorithm for Solving Inclusions with Mixtures of Composite, Lipschitzian, and Parallel-Sum Type Monotone Operators. *Set-Valued and Variational Analysis*, 20(2):307–330, 2012.

[11] O. Devolder, F. Glineur and Y. Nesterov. Double Smoothing Technique for Large-Scale Linearly Constrained Convex Optimization. *SIAM Journal on Optimization*, 22(2):702–727, 2012.

[12] T.N. Lal, O. Chapelle and B. Schölkopf. Combining a Filter Method with SVMs. *Studies in Fuzziness and Soft Computing*, 207:439–445, 2006.

[13] Y. Nesterov. Excessive gap technique in nonsmooth convex optimization. *SIAM Journal of Optimization*, 16(1):235–249, 2005.

[14] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[15] Y. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming*, 110(2):245–259, 2005.

[16] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers Dordrecht, 2004.

[17] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269:543–547, 1983.

[18] J. Shawe-Taylor and N. Christianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.