A variable smoothing algorithm for solving convex optimization problems

Radu Ioan Boţ [*] Christopher Hendrich [†]

March 16, 2014

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 them to convex and differentiable functions with Lipschitz continuous gradients 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

$$\inf_{x\in\mathcal{H}} \{f(x) + g(Kx)\}, \tag{1.1}$$

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 [24–26], we approximate (1.1) 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 [8,9,20], which assumes the smoothing of the Fenchel-dual problem to (1.1) to an optimization problem with a strongly convex and differentiable objective function with Lipschitz continuous gradient. There, the smoothed dual problem is solved via an appropriate fast gradient method (cf. [27]) and a primal

[*] University of Vienna, Faculty of Mathematics, Nordbergstraße 15, A-1090 Vienna, Austria, e-mail: radu.bot@univie.ac.at. Research partially supported by DFG (German Research Foundation), project BO 2516/4-1.

[†] Chemnitz University of Technology, Department of Mathematics, D-09107 Chemnitz, Germany, e-mail: christopher.hendrich@mathematik.tu-chemnitz.de. Research supported by a Graduate Fellowship of the Free State Saxony, Germany.

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. [28]) for convex optimization problems involving variable smoothing parameters which are updated in each iteration. For the implementation of the provided iterative scheme one only needs to determine the proximal points of the functions f and g (respectively, of their conjugate functions), while the operator K and its adjoint are involved via forward evaluations. For a large class of problems arising in location theory, machine learning, clustering, signal and image processing, etc., exact formulae for the proximal point mappings are available (see, for instance, [15, 17, 19]).

The algorithmic scheme we propose in this paper 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 (1.1). In the case of constant smoothing variables the approach provides a sequence of iterates which solves the problem (1.1) with a given a priori 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. More than that, when implementing the variable smoothing scheme, different to the constant smoothing one, it is not necessary to know the Lipschitz constants of the functions f and g in advance.

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 (1.1)

$$\inf_{x\in\mathcal{H}}\left\{f(x)+\sum_{i=1}^{m}g_i(K_ix)\right\}, \tag{1.2}$$

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

We would like to notice that variable smoothing parameters have been recently considered in [29] for the PRISMA algorithm in relation to nonsmooth optimization problems having as objective the sum of three convex functions with different properties. However, our approach allows considering compositions with linear continuous operators as summands in the objective, an aspect which is relevant in many practical applications, as it is emphasized by the numerical experiments considered in the last section. On the other hand, when comparing it to the popular augmented Lagrangian method (ALM) and alternating direction method of multipliers (ADMM) (see [12]), our method has the advantage that the linear continuous operators (and, respectively, their adjoints) are evaluated via forward steps, while for the nondifferentiable functions separate proximal steps are performed. In contrast to this splitting philosophy, the numerical schemes which arise in the implementation of augmented Lagrangian methods build on (possibly) expensive linear operator inversions. Thus, from the point of view of the implementation, the variable smoothing method shares similarities with the recently introduced class of primal-dual algorithms (see, for instance, [10,13,15,18,32]).

The accelerated gradient method of Nesterov from [28] has been employed also in the context of solving optimization problems of type (1.1) in finite-dimensional spaces by Beck and Teboulle in [2] in order to obtain improved rates of convergence, however, under the restrictive assumption that g is convex and differentiable with a Lipschitz continuous gradient, but by allowing f to be a proper, convex and lower semicontinuous function. It also worth to notice that in the setting in which our approach is introduced, one can smooth the functions also by using other appropriate methods (see, for instance, [3]). We opted for the use of the Moreau envelope in this scope, not only because its gradient is Lipschitz continuous, but also because it can be expressed by means of the proximal points of the function in discussion. In the light of Moreau's decomposition formula, this fact also allows a unitary treatment of the involved functions and of their conjugates.

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 (1.1) 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 $f \,\Box\, g : \mathcal{H} \to \overline{\mathbb{R}}$, $(f \,\Box\, g)(x) = \inf_{y\in\mathcal{H}} \{f(y) + g(x-y)\}$ for all $x \in \mathcal{H}$. When $f, g : \mathcal{H} \to \overline{\mathbb{R}}$ are proper and convex, then

$$(f+g)^* = f^* \,\Box\, g^* \tag{2.1}$$

provided that f (or g) is continuous at a point belonging to $\operatorname{dom} f \cap \operatorname{dom} g$. For other qualification conditions guaranteeing (2.1) we refer the reader to [6,14,22,31].

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 (see [23])

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

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

$$\inf_{y \in \mathcal{H}} \left\{ f(y) + \frac{1}{2\gamma} \|y - x\|^2 \right\}. \tag{2.2}$$

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

$$\|\operatorname{Prox}_{\gamma f}(x) - \operatorname{Prox}_{\gamma f}(y)\|^2 + \|(x - \operatorname{Prox}_{\gamma f}(x)) - (y - \operatorname{Prox}_{\gamma f}(y))\|^2 \leq \|x - y\|^2 \; \forall x, y \in \mathcal{H}. \tag{2.3}$$

Hence, it is Lipschitz continuous with Lipschitz constant equal to 1. For a large class of functions arising in different fields of applications, the proximal point mappings are given by exact formulae, whereby it is often more convenient to calculate the proximal point mappings of the conjugates and then to deduce from here, via the formulae given bellow, the ones of the functions themselves (cf. [1, 17, 19]). We also have (cf. [1, Theorem 14.3])

$$^{\gamma}f(x) + {}^{\frac{1}{\gamma}} f^*(\tfrac{x}{\gamma}) = \frac{\|x\|^2}{2\gamma} \; \forall x \in \mathcal{H} \tag{2.4}$$

and the extended *Moreau's decomposition formula*

$$\operatorname{Prox}_{\gamma f}(x) + \gamma \operatorname{Prox}_{\frac{1}{\gamma} f^*} \left(\tfrac{x}{\gamma} \right) = x \; \forall x \in \mathcal{H}. \tag{2.5}$$

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])

$$\nabla(^{\gamma}f)(x) = \tfrac{1}{\gamma}(x - \operatorname{Prox}_{\gamma f}(x)) \; \forall x \in \mathcal{H}, \tag{2.6}$$

being in the light of (2.3) $\frac{1}{\gamma}$-Lipschitz continuous. For a nonempty, convex and closed set $C \subseteq \mathcal{H}$ and $\gamma \in \mathbb{R}_{++}$, it holds $\operatorname{Prox}_{\gamma \delta_C} = \mathcal{P}_C$, 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, 27, 28])

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_{\nabla f}}{2} \|y - x\|^2. \tag{2.7}$$

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 [5, Proposition 4.4.6], we have that

$$\operatorname{dom} f^* \subseteq L_f B_{\mathcal{H}} \text{ and } \operatorname{dom} g^* \subseteq L_g B_{\mathcal{K}}. \tag{2.8}$$

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, which is proper, convex and lower semicontinuous, via its Moreau envelope of parameter ρ, which yields ${}^{\rho}f : \mathcal{H} \to \mathbb{R}$, ${}^{\rho}f(x) = \left(f \,\square\, \frac{1}{2\rho}\|\cdot\|^2\right)(x)$ for every $x \in \mathcal{H}$. According to the Fenchel-Moreau Theorem and due to (2.1), 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. (2.6) and (2.5))

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

is $\frac{1}{\rho}$-Lipschitz continuous (see, for instance, [4, Proposition 3.4]).

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 $x \in \mathcal{H}$. The function ${}^{\mu}g \circ K$ is differentiable and its gradient $\nabla({}^{\mu}g \circ K) : \mathcal{H} \to \mathcal{H}$ fulfills (cf. (2.6) and (2.5))

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

Further, for every $x, y \in \mathcal{H}$, it holds (see (2.3))

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

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 $L(\rho,\mu)$-Lipschitz continuous gradient $\nabla F^{\rho,\mu} : \mathcal{H} \to \mathcal{H}$ given by

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

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

For $\rho_2 \ge \rho_1 > 0$ and every $x \in \mathcal{H}$, it holds (cf. (2.8))

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

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

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

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

$$F^{\rho_2,\mu_2}(x) \leq F^{\rho_1,\mu_1}(x) \leq F^{\rho_2,\mu_2}(x) + (\rho_2 - \rho_1)\frac{L_f^2}{2} + (\mu_2 - \mu_1)\frac{L_g^2}{2} \tag{3.1}$$

and

$$F^{\rho_2,\mu_2}(x) \leq F(x) \leq F^{\rho_2,\mu_2}(x) + \rho_2 \frac{L_f^2}{2} + \mu_2 \frac{L_g^2}{2}. \tag{3.2}$$

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. [28]) by using variable smoothing parameters, which we update in each iteration.

Algorithm 3.1. Let $y_1 = x_0 \in \mathcal{H}$, $(\rho_k)_{k\geq 1}$, $(\mu_k)_{k\geq 1} \subseteq \mathbb{R}_{++}$, let $t_1 = 1$, and set

$$(\forall k \geq 1) \left\lfloor \begin{array}{l} L_k = \frac{1}{\rho_k} + \frac{\|K\|^2}{\mu_k}, \\ x_k = y_k - \frac{1}{L_k}\left(\operatorname{Prox}_{\frac{1}{\rho_k}f^*}\left(\frac{y_k}{\rho_k}\right) + K^*\operatorname{Prox}_{\frac{1}{\mu_k}g^*}\left(\frac{Ky_k}{\mu_k}\right)\right), \\ t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}, \\ y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}). \end{array}\right. \tag{3.3}$$

The convergence of Algorithm 3.1 is proved by the following theorem.

Theorem 3.1. *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 3.1 generates a sequence $(x_k)_{k\geq 1} \subseteq \mathcal{H}$ satisfying

$$F(x_{k+1}) - F(x^*) \leq \frac{2(a + b\|K\|^2)}{k+2}\|x_0 - x^*\|^2 + \frac{2(1 + \ln(k+1))}{k+2}\left(\frac{L_f^2}{a} + \frac{L_g^2}{b}\right) \ \forall k \geq 1. \tag{3.4}$$

This yields 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

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

For any $k \geq 1$, it holds

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

and from here, it follows

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

Further, using (2.7), since $x_{k+1} = y_{k+1} - \frac{1}{L_{k+1}}\xi_{k+1}$, it follows

$$\begin{aligned} F^{k+1}(x_{k+1}) &\leq F^{k+1}(y_{k+1}) + \langle \xi_{k+1}, x_{k+1} - y_{k+1}\rangle + \frac{L_{k+1}}{2}\|x_{k+1} - y_{k+1}\|^2 \\ &= F^{k+1}(y_{k+1}) - \frac{1}{L_{k+1}}\|\xi_{k+1}\|^2 + \frac{1}{2L_{k+1}}\|\xi_{k+1}\|^2 \\ &= F^{k+1}(y_{k+1}) - \frac{1}{2L_{k+1}}\|\xi_{k+1}\|^2, \end{aligned} \tag{3.5}$$

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

$$\begin{aligned} \langle x^* - y_{k+1}, \xi_{k+1}\rangle &\leq F^{k+1}(x^*) - F^{k+1}(y_{k+1}) \\ &\overset{(3.5)}{\leq} F^{k+1}(x^*) - F^{k+1}(x_{k+1}) - \frac{1}{2L_{k+1}}\|\xi_{k+1}\|^2 \ \ \forall k \geq 1. \end{aligned} \tag{3.6}$$

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

$$\begin{aligned} \|\xi_{k+1}\|^2 &\overset{(3.5)}{\leq} 2L_{k+1}(F^{k+1}(y_{k+1}) - F^{k+1}(x_{k+1})) \\ &\leq 2L_{k+1}\left(F^{k+1}(x_k) - F^{k+1}(x_{k+1}) - \frac{1}{t_{k+1}}\langle \xi_{k+1}, p_k\rangle\right) \ \ \forall k \geq 1. \end{aligned} \tag{3.7}$$

Thus, as $t_{k+1}^2 - t_{k+1} = t_k^2$ and by making use of (3.1), for any $k \geq 1$, it yields

$$\|p_{k+1} - x_{k+1} + x^*\|^2 - \|p_k - x_k + x^*\|^2$$

$$
\begin{aligned}
&\overset{(3.6)}{\le} \frac{2(t_{k+1}-1)}{L_{k+1}}\langle p_k,\xi_{k+1}\rangle + \frac{2t_{k+1}}{L_{k+1}}(F^{k+1}(x^*)-F^{k+1}(x_{k+1})) + \frac{t_{k+1}^2-t_{k+1}}{L_{k+1}^2}\|\xi_{k+1}\|^2\\
&\overset{(3.7)}{\le} \frac{2t_{k+1}}{L_{k+1}}(F^{k+1}(x^*)-F^{k+1}(x_{k+1})) + \frac{2(t_{k+1}^2-t_{k+1})}{L_{k+1}}(F^{k+1}(x_k)-F^{k+1}(x_{k+1}))\\
&= \frac{2t_k^2}{L_{k+1}}(F^{k+1}(x_k)-F^{k+1}(x^*)) - \frac{2t_{k+1}^2}{L_{k+1}}(F^{k+1}(x_{k+1})-F^{k+1}(x^*))\\
&\overset{(3.1)}{\le} \frac{2t_k^2}{L_{k+1}}\left(F^k(x_k)-F^k(x^*)+(\rho_k-\rho_{k+1})\frac{L_f^2}{2}+(\mu_k-\mu_{k+1})\frac{L_g^2}{2}\right)\\
&- \frac{2t_{k+1}^2}{L_{k+1}}(F^{k+1}(x_{k+1})-F^{k+1}(x^*))\\
&= \frac{2t_k^2}{L_{k+1}}\left(F^k(x_k)-F^k(x^*)+\rho_k\frac{L_f^2}{2}+\mu_k\frac{L_g^2}{2}\right) - \frac{2t_{k+1}^2}{L_{k+1}}(F^{k+1}(x_{k+1})-F^{k+1}(x^*))\\
&- \frac{2t_k^2}{L_{k+1}}\left(\rho_{k+1}\frac{L_f^2}{2}+\mu_{k+1}\frac{L_g^2}{2}\right).
\end{aligned}
$$

By using (3.2), it follows that for any $k \ge 1$

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

thus

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

which implies that

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

Making again use of (3.2), this further yields for any $k \geq 1$

$$\begin{aligned}
&\frac{2t_{k+1}^2}{L_{k+1}} \left(F(x_{k+1}) - F(x^*)\right) \\
&\leq \frac{2t_{k+1}^2}{L_{k+1}} \left(F^{k+1}(x_{k+1}) - F^{k+1}(x^*) + \rho_{k+1}\frac{L_f^2}{2} + \mu_{k+1}\frac{L_g^2}{2}\right) + \|p_{k+1} - x_{k+1} + x^*\|^2 \\
&\leq \frac{2t_1^2}{L_1} \left(F^1(x_1) - F^1(x^*) + \rho_1\frac{L_f^2}{2} + \mu_1\frac{L_g^2}{2}\right) + \|p_1 - x_1 + x^*\|^2 \\
&+ \sum_{s=1}^{k} \frac{2t_{s+1}}{L_{s+1}} \left(\rho_{s+1}\frac{L_f^2}{2} + \mu_{s+1}\frac{L_g^2}{2}\right). \qquad (3.8)
\end{aligned}$$

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

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

we get

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

and this, together with (3.8), give rise to the following estimate

$$\frac{2t_{k+1}^2}{L_{k+1}} \left(F(x_{k+1}) - F(x^*)\right) \leq \|x_0 - x^*\|^2 + \sum_{s=1}^{k+1} \frac{t_s}{L_s}\left(\rho_s L_f^2 + \mu_s L_g^2\right). \qquad (3.9)$$

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

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

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

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

which concludes the proof. □

In the second part of this subsection we propose a variant of Algorithm 3.1 formulated with constant smoothing parameters:

Algorithm 3.2. Let $y_1 = x_0 \in \mathcal{H}$, ρ, $\mu \in \mathbb{R}_{++}$, let $t_1 = 1$, $L(\rho,\mu) = \frac{1}{\rho} + \frac{\|K\|^2}{\mu}$, and set

$$(\forall k \geq 1) \left\lfloor \begin{array}{l} x_k = y_k - \frac{1}{L(\rho,\mu)}\left(\operatorname{Prox}_{\frac{1}{\rho}f^*}\left(\frac{y_k}{\rho}\right) + K^*\operatorname{Prox}_{\frac{1}{\mu}g^*}\left(\frac{Ky_k}{\mu}\right)\right), \\ t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}, \\ y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}). \end{array}\right. \tag{3.10}$$

Remark 3.1. Algorithm 3.2 is nothing else than the accelerated gradient method proposed by Nesterov in [28] employed to the minimization of the function $F^{\rho,\mu} = {}^{\rho}f + {}^{\mu}g \circ K$. It generates a sequence $(x_k)_{k\geq 1} \subseteq \mathcal{H}$ which provides a rate of convergence for the objective $F^{\rho,\mu}$ of order $\mathcal{O}(\frac{1}{k^2})$. In Theorem 3.2 we will discuss the rate of convergence of $(F(x_k))_{k\geq 1}$.

Remark 3.2. Constant smoothing parameters have been also used in [20] and [8, 9] 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 3.2. *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 3.2 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 3.1 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 (3.9). 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

$$F(x_{k+1}) - F(x^*) \leq \frac{2L(\rho,\mu)\|x_0 - x^*\|^2}{(k+2)^2} + \frac{\rho L_f^2 + \mu L_g^2}{2} \ \forall k \geq 1. \tag{3.11}$$

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 only have 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

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

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

Remark 3.3. The rate of convergence of Algorithm 3.1 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 3.2, despite of the fact that it approximates the optimal objective value with a better convergence rate, this sequence may not converge to the optimal objective value. Indeed, by taking into account (3.11), one can see that the right-hand side of this inequality is bounded from below by some strictly positive real number.

3.3 The case when f is differentiable with Lipschitz continuous gradient

In this subsection we show how Algorithm 3.1 and Algorithm 3.2 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 3.1 can be adapted to this framework as follows.

Algorithm 3.3. Let $y_1 = x_0 \in \mathcal{H}$, $(\mu_k)_{k\geq 1} \subseteq \mathbb{R}_{++}$, let $t_1 = 1$, and set

$$(\forall k \geq 1)\ \left\lfloor \begin{array}{l} L_k = L_{\nabla f} + \frac{\|K\|^2}{\mu_k},\\ x_k = y_k - \frac{1}{L_k}\left(\nabla f(y_k) + K^*\mathrm{Prox}_{\frac{1}{\mu_k}g^*}\left(\frac{Ky_k}{\mu_k}\right)\right),\\ t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2},\\ y_{k+1} = x_k + \frac{t_k-1}{t_{k+1}}(x_k - x_{k-1}). \end{array}\right. \tag{3.12}$$

Its convergence is furnished by the following theorem.

Theorem 3.3. *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 3.3 generates a sequence $(x_k)_{k\geq 1} \subseteq \mathcal{H}$ satisfying for any $k \geq 1$

$$F(x_{k+1}) - F(x^*) \leq \frac{2(L_{\nabla f} + b\left\|K\right\|^2)}{k+2}\left\|x_0 - x^*\right\|^2 + \frac{2(1+\ln(k+1))}{k+2}\frac{L_g^2(L_{\nabla f} + b\left\|K\right\|^2)}{b^2\left\|K\right\|^2}. \tag{3.13}$$

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

Proof. For any $k \geq 1$, we let $F^k : \mathcal{H} \to \mathbb{R}$, $F^k(x) = f(x) + {}^{\mu_k}g(Kx)$. For any $k \geq 1$ and every $x \in \mathcal{H}$, it holds $\nabla F^k(x) = \nabla f(x) + K^*\operatorname{Prox}_{\frac{1}{\mu_k}g^*}\left(\frac{Kx}{\mu_k}\right)$ 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 3.1, 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

$$\frac{2t_{k+1}^2}{L_{k+1}} \left(F(x_{k+1}) - F(x^*)\right) \leq \|x_0 - x^*\|^2 + \sum_{s=1}^{k+1} \frac{t_s L_g^2}{L_s}\mu_s. \tag{3.14}$$

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

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

By adapting Algorithm 3.3 to the framework considered in this subsection, we obtain the following algorithm with constant smoothing variables:

Algorithm 3.4. Let $y_1 = x_0 \in \mathcal{H}$, $\mu \in \mathbb{R}_{++}$, let $t_1 = 1$, $L(\mu) = L_{\nabla f} + \frac{\|K\|^2}{\mu}$, and set

$$(\forall k \geq 1) \quad \left\lfloor \begin{array}{l}
x_k = y_k - \frac{1}{L(\mu)}\left(\nabla f(y_k) + K^*\operatorname{Prox}_{\frac{1}{\mu}g^*}\left(\frac{Ky_k}{\mu}\right)\right), \\
t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}, \\
y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}).
\end{array}\right. \tag{3.15}$$

The convergence of Algorithm 3.4 is stated by the following theorem, which can be proved in the lines of the proof of Theorem 3.3.

Theorem 3.4. *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 3.4 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 $\mathcal{O}(\frac{1}{k})$.

Remark 3.4. Algorithm 3.2 is the accelerated gradient method from [28] employed to the minimization of the function $F^\mu = f + {}^\mu g \circ K$, which provides a rate of convergence for the objective F^μ of order $\mathcal{O}(\frac{1}{k^2})$. Make use of the terminology in [3], one can notice that $g \circ K$ is $(\|K\|^2, \frac{L_g^2}{2}, 0)$-smoothable, thus, for the choice of the smoothing parameter μ in Algorithm 3.4, such that the generated sequence $(x_k)_{k\geq 1} \subseteq \mathcal{H}$ provides an ε-optimal solution to (P) with a rate of convergence for the objective of order $\mathcal{O}(\frac{1}{k})$, one can make use of [3, Theorem 3.1]. A simple calculation shows that this will lead to the same value for μ as given in Theorem 3.4.

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 (1.2)

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

where $\mathcal{H}$ and $\mathcal{K}_i$, $i = 1, \ldots, m$, 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 $K_i : \mathcal{H} \to \mathcal{K}_i$, $i = 1, \ldots, m$, are linear continuous operators. By endowing $\mathcal{K} := \mathcal{K}_1 \times \ldots \times \mathcal{K}_m$ 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 $g : \mathcal{K} \to \mathbb{R}, g(y_1, \ldots, y_m) = \sum_{i=1}^m g_i(y_i)$ and $K : \mathcal{H} \to \mathcal{K}, Kx = (K_1 x, \ldots, K_m x)$, problem (1.2) can equivalently be 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 $(y_1, \ldots, y_m), (z_1, \ldots, z_m) \in \mathcal{K}$

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

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 $(y_1, \dots, y_m) \in \mathcal{K}$, it holds

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

thus

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

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

$$\begin{aligned}\nabla(^{\mu}g \circ K)(x) = K^*\nabla(^{\mu}g)(K_1x, \dots, K_mx) &= \sum_{i=1}^m K_i^*\nabla(^{\mu}g_i)(K_ix) \\ &= \sum_{i=1}^m K_i^*\operatorname{Prox}_{\frac{1}{\mu}g_i^*}\left(\frac{K_ix}{\mu}\right) \ \forall x \in \mathcal{H}.\end{aligned}$$

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

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

which shows that $\nabla(^{\mu}g \circ K)$ is $\frac{\sum_{i=1}^m \|K_i\|^2}{\mu}$-Lipschitz continuous. In order to deduce the Lipschitz constant of $\nabla(^{\mu}g \circ K)$ one can alternatively apply [3, Lemma 2.1 and Lemma 2.2].

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 (see, for instance, [15, Subsection 6.2.2])

$$\inf_{x \in \mathbb{R}^n} \{\|Ax - u\|_1 + \lambda \|Wx\|_1\}, \tag{4.1}$$

where $u \in \mathbb{R}^n$ 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. Here we use a robust l_1 data fidelity term, which is contrast invariant (cf. [15]) and nonsmooth. The norms of the linear continuous operators involved are $\|A\| = 1$ and $\|W\| = 1$.

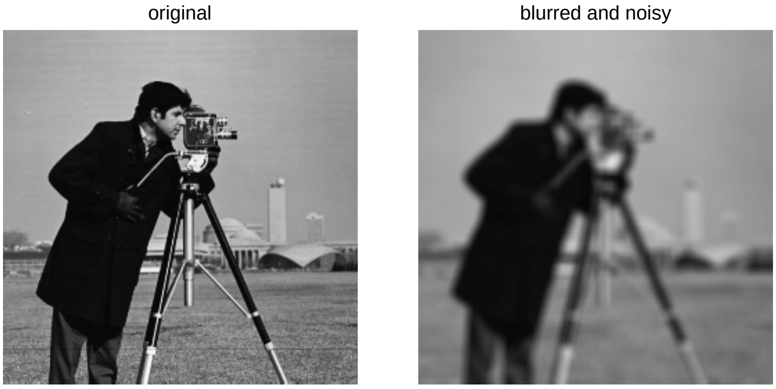

Figure 4.1: The 256 × 256 cameraman test image

The optimization problem (4.1) 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 $L_{\nabla f} = 0$, $g_1 : \mathbb{R}^n \to \mathbb{R}$, $g_1(y) = \|y - u\|_1$ 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 $g_1^*(p) = \delta_{[-1,1]^n}(p) + p^T u$ and $g_2^*(p) = \delta_{[-\lambda,\lambda]^n}(p)$ (see, for instance, [6]). By also using the considerations made in Subsection 3.4, we solved this problem with Algorithm 3.3 and computed to this aim for $\mu \in \mathbb{R}_{++}$ and $x \in \mathbb{R}^n$

$$\operatorname{Prox}_{\frac{1}{\mu}g_1^*}\left(\frac{Ax}{\mu}\right) = \operatorname*{arg\,min}_{p\in[-1,1]^n} \left\{\frac{1}{\mu}p^T u + \frac{1}{2}\left\|\frac{Ax}{\mu} - p\right\|^2\right\} = \mathcal{P}_{[-1,1]^n}\left(\frac{Ax-u}{\mu}\right),$$

and

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

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 in Algorithm 3.3 with starting point $u \in \mathbb{R}^n$ becomes

Algorithm 4.1. Let $y_1 = x_0 = u \in \mathbb{R}^n$, $a \in \mathbb{R}_{++}$, let $t_1 = 1$, and set

$$(\forall k \geq 1)\quad \left\lfloor \begin{array}{l} \mu_k = \frac{1}{ak},\ L_k = 2ak, \\ x_k = y_k - \frac{1}{L_k}\left(A^*\mathcal{P}_{[-1,1]^n}\left(\frac{Ay_k-u}{\mu_k}\right) + W^*\mathcal{P}_{[-\lambda,\lambda]^n}\left(\frac{Wy_k}{\mu_k}\right)\right), \\ t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}, \\ y_{k+1} = x_k + \frac{t_k-1}{t_{k+1}}(x_k - x_{k-1}). \end{array}\right. \tag{4.2}$$

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 (black) to 1 (white). In addition, we added 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.

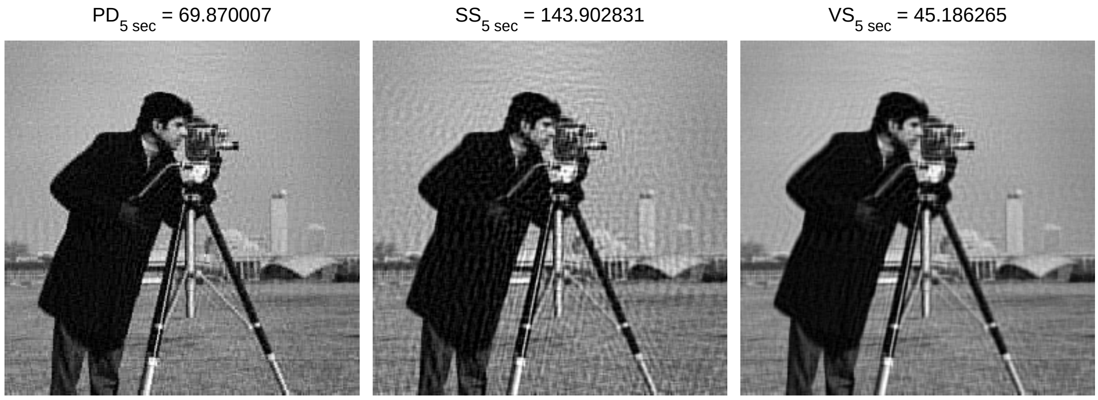
$\mathrm{PD}_{5\ \mathrm{sec}}$ = 69.870007
$\mathrm{SS}_{5\ \mathrm{sec}}$ = 143.902831
$\mathrm{VS}_{5\ \mathrm{sec}}$ = 45.186265

Figure 4.2: Results furnished by the primal-dual (PD), the skew splitting (SS) and the variable smoothing (VS) algorithms after 5 seconds of CPU time.

When measuring the quality of the restored images, we made use of the *improvement in signal-to-noise ratio* (ISNR, cf. [16]), which is defined as

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

where x, u 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.

a	1e-4	1e-3	1e-2	1e-1	1	1e+1	1e+2	1e+3
fval	129.560	65.392	47.711	45.316	44.534	49.257	168.110	472.602
ISNR	2.141	4.590	6.100	6.133	6.117	5.258	1.570	0.301

Table 4.1: Objective values (fval) and ISNR values (higher is better) after 5 seconds of CPU time.

In the context of solving the problem (4.1), we compared the variable smoothing approach (VS) for $a = $ 1e-1 with the operator-splitting algorithm based on skew splitting (SS) proposed in [13,18] with parameters $\varepsilon = \frac{1}{10(\sqrt{2}+1)}$ and $\gamma_k = \gamma = \frac{1-\varepsilon}{\sqrt{2}}$, for any $k \geq 1$, and with the primal-dual algorithm (PD) from [15] with parameters $\theta = 1$, $\sigma = 0.01$ and $\tau = 70.001$. The parameters considered for the three approaches provide the best results when solving (4.1). The output of these three algorithms after 5 seconds of CPU time, 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 5 seconds.

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 belongs to the class of kernel based learning methods.

The given data set consisting of 11339 training images and 1850 test images of size 28×28 was taken from the website http://www.cs.nyu.edu/~roweis/data.html. The

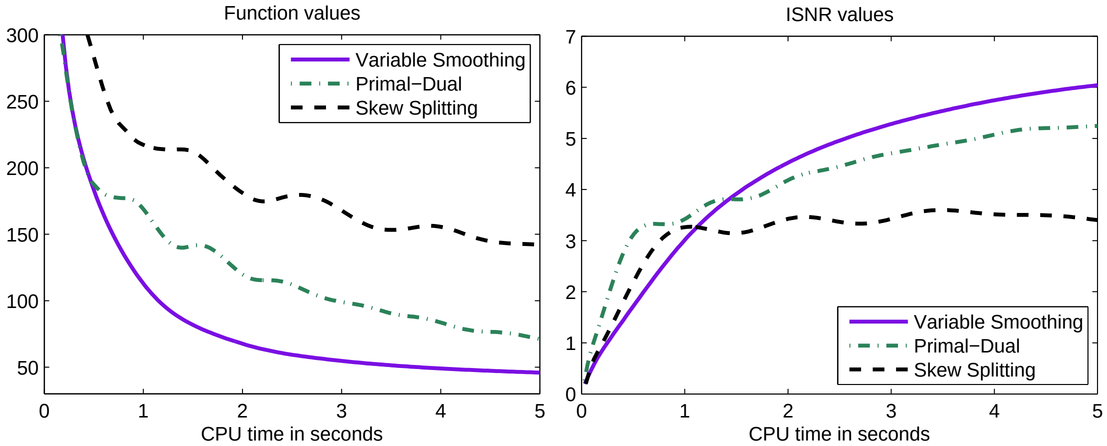

Figure 4.3: The evolution of the values of the objective function and of the ISNR for the primal-dual (PD), the skew splitting (SS) and the variable smoothing (VS) algorithms after 5 seconds of CPU time.

problem we consider is to determine a decision function based on a pool of handwritten digits showing either the number five or the number six, labeled by $+1$ and -1, respectively (see Figure 4.4). Subsequently, we evaluate the quality of the decision function on the test data set by computing the percentage of misclassified images. In order to reduce the computational effort, we used only half of the available images from the training data set.

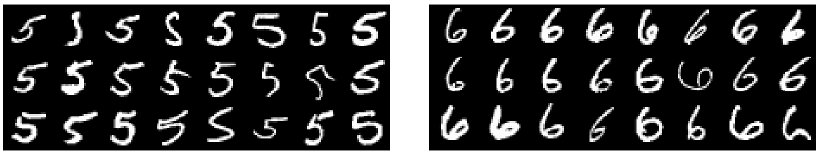

Figure 4.4: A sample of images belonging to the classes $+1$ and -1, respectively.

The classifier functional $\mathtt{f}$ 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

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

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 $\mathcal{Z} = \{(X_1,Y_1),\dots,(X_n,Y_n)\} \subseteq \mathbb{R}^d \times \{+1,-1\}$, namely the symmetric and positive definite matrix with entries $K_{ij} = \kappa(X_i,X_j)$ for $i,j = 1,\dots,n$. 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 nonsmooth 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*

$$\inf_{\mathtt{f}\in\mathcal{H}_\kappa}\left\{\frac{1}{2}\Omega(\mathtt{f})+C\sum_{i=1}^{n}v(\mathtt{f}(X_i),Y_i)\right\}, \tag{4.3}$$

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

The *representer theorem* (cf. [30]) ensures the existence of a vector of coefficients $c=(c_1,\dots,c_n)^T\in\mathbb{R}^n$ such that the minimizer $\mathtt{f}$ of (4.3) 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 $\Omega(\mathtt{f})=\|\mathtt{f}\|_\kappa^2=\langle\mathtt{f},\mathtt{f}\rangle_\kappa=\sum_{i=1}^n\sum_{j=1}^n c_ic_j\kappa(X_i,X_j)=c^TKc$ and for $i=1,\dots,n$, 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

$$\inf_{c\in\mathbb{R}^n}\left\{f(c)+\sum_{i=1}^n g_i(Kc)\right\}, \tag{4.4}$$

where $f:\mathbb{R}^n\to\mathbb{R}$, $f(c)=\frac{1}{2}c^TKc$, and $g_i:\mathbb{R}^n\to\mathbb{R}$, $g_i(c)=Cv(c_i,Y_i)$ for $i=1,\dots,n$. 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 $i=1,\dots,n$ the function $g_i:\mathbb{R}^n\to\mathbb{R}$ is convex and C-Lipschitz continuous, properties which allowed us to solve the problem (4.4) with Algorithm 3.3, by using also the considerations made in Subsection 3.4. For any $i=1,\dots,n$ and every $p=(p_1,\dots,p_n)^T\in\mathbb{R}^n$, it holds (see, also, [7, 11])

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

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

$$\operatorname{Prox}_{\frac{1}{\mu}g_i^*}\left(\frac{c}{\mu}\right)=\underset{\substack{p_iY_i\in[-C,0]\\ p_j=0,j\neq i}}{\arg\min}\left\{\frac{p_iY_i}{\mu}+\frac{1}{2}\left(\frac{c_i}{\mu}-p_i\right)^2\right\}.$$

For $Y_i=1$, we have

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

while for $Y_i=-1$, it holds

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

Summarizing, it follows

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

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

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

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

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

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 $k \ge 1$, the iterative scheme in Algorithm 3.3 with starting point $x_0 = 0 \in \mathbb{R}^n$ becomes

Algorithm 4.2. Let $y_1 = x_0 = 0 \in \mathbb{R}^n$, $a \in \mathbb{R}_{++}$, let $t_1 = 1$, and set

$$(\forall k \ge 1) \left\lfloor \begin{array}{l} \mu_k = \frac{1}{ak},\ L_k = \|K\| + ak\|K\|^2, \\ x_k = y_k - \frac{1}{L_k}\left(Ky_k + K^*\left(\mathcal{P}_{Y_i[-C,0]}\left(\frac{(Ky_k)_i - Y_i}{\mu_k}\right)\right)^T_{i=\overline{1,n}}\right), \\ t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}, \\ y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}). \end{array}\right. \tag{4.5}$$

We denote by $\mathcal{D} = \{(X_i, Y_i), i = 1, \ldots, 5670\} \subseteq \mathbb{R}^{784} \times \{+1, -1\}$ the set of available training data consisting of 2711 images in the class $+1$ and 2959 images in the class -1. Notice that a sample from each class of images is shown in Figure 4.4. Due to numerical reasons, the images have been normalized (cf. [21]) by dividing each of them by the quantity $\left(\frac{1}{5670}\sum_{i=1}^{5670}\|X_i\|^2\right)^{\frac{1}{2}}$.

C	$\sigma = 0.125$	$\sigma = 0.25$	$\sigma = 0.5$	$\sigma = 0.75$	$\sigma = 1$	$\sigma = 2$
0.1	1.0270	1.3514	1.3514	1.8919	2.1081	3.0270
1	1.0270	0.7027	0.7568	1.3514	1.4595	2.2162
10	1.0270	0.7568	0.9189	1.0811	1.1892	1.8378
100	1.0270	0.7568	0.8649	1.4054	1.2432	1.8378
1000	1.0270	0.7568	0.8649	1.4595	1.2432	1.8378

Table 4.2: Misclassification rate in percentage for different model parameters.

In order to specify a good choice for the kernel parameter $\sigma \in \mathbb{R}_{++}$ and the tradeoff parameter $C \in \mathbb{R}_{++}$, we tested different combination of them with a MATLAB solver over a fixed number of 20000 iterations. Table 4.2 shows the misclassification rate in percentage for a selection of different model parameters, whereby the combination $\sigma = 0.25$ and $C = 1$ provides with 0.7027 % the lowest misclassification error. This means that among the 1870 images belonging to the test data set 13 were not correctly classified.

By using the variable smoothing algorithm proposed in this paper, we tested different values of $a \in \mathbb{R}_{++}$ for this combination. In Table 4.3 we present the number of iterations needed in order to guarantee that the misclassification rate reaches and remains below the optimal level of 0.7027 %. It turns out that for $a = 0.03$ the algorithm only needs 117 iterations to achieve this precision. Table 4.3 also gives an insight into the numerical sensitivity of the parameter a when using the variable smoothing approach. For example, by taking $a = 0.003$ or $a = 0.5$, the algorithm needs more than ten times the amount of iterations in order to reach the same precision.

a	0.003	0.005	0.01	0.03	0.05	0.1	0.3	0.5
iterations	1235	741	379	117	140	300	864	1475

Table 4.3: Number of iterations needed by the variable smoothing algorithm such that the misclassification rate remains below the misclassification rate of 0.7027 % for $\sigma = 0.25$ and $C = 1$.

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

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] A. Beck and M. Teboulle. A fast iterative shrinkage-tresholding algorithm for linear inverse problems. *SIAM J. Imag. Sci.* **2**(1), 183–202, 2009.

[3] A. Beck and M. Teboulle. Smoothing and first order methods: a unified framework. *SIAM J. Optim.* **22**(2), 557–580, 2012.

[4] D. Bertsekas. *Constrained optimization and Lagrange Multiplier Methods.* Athena Scientific, Belmont, 1996.

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

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

[7] R.I. Boţ, A. Heinrich and G. Wanka. Employing different loss functions for the classification of images via supervised learning. *Cent. Eur. J. Math.* **12**(2), 381–394, 2014.

[8] R.I. Boţ and C. Hendrich. A double smoothing technique for solving unconstrained nondifferentiable convex optimization problems. *Comput. Optim. Appl.* **54**(2), 239–262, 2013.

[9] R.I. Boţ and C. Hendrich. On the acceleration of the double smoothing technique for unconstrained convex optimization problems. *Optimization*, 2012. http://dx.doi.org/10.1080/02331934.2012.745530

[10] R.I. Boţ and C. Hendrich. A Douglas–Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators. *SIAM J. Optim.* **23**(4), 2541–2565, 2013.

[11] R.I. Boţ and N. Lorenz. Optimization problems in statistical learning: Duality and optimality conditions. *Eur. J. Oper. Res.* **213**(2), 395–404, 2011.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3**(1), 1–122, 2010.

[13] L.M. Briceño-Arias and P.L. Combettes. A monotone + skew splitting model for composite monotone inclusions in duality. *SIAM J. Optim.* **21**(4), 1230–1250, 2011.

[14] R.S. Burachik and V. Jeyakumar. A new geometric condition for Fenchel's duality in infinite dimensional spaces. *Math. Program.* **104**(2–3), 229–233, 2005.

[15] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.* **40**(1), 120–145, 2011.

[16] G. Chantas, N. Galatsanos, A. Likas and M. Saunders. Variational bayesian image restoration based on a product of t-distributions image prior. *IEEE Trans. Image Process.* **17**(10), 1795–1805, 2008.

[17] P.L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, Springer Ser. Optim. Appl. 49, H. H. Bauschke et al., eds., Springer, New York, pp. 185–212, 2011.

[18] 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 Var. Anal.* **20**(2), 307–330, 2012.

[19] P.L. Combettes and V.R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.* **4**(4), 1168–1200, 2005.

[20] O. Devolder, F. Glineur and Y. Nesterov. Double smoothing technique for large-scale linearly constrained convex optimization. *SIAM J. Optim.* **22**(2), 702–727, 2012.

[21] T.N. Lal, O. Chapelle and B. Schölkopf. Combining a filter method with SVMs. *Studies in Fuzziness and Soft Computing*, Springer, **207**, pp. 439–445, 2006.

[22] G. Li and K.F. Ng. On extension of Fenchel duality and its application. *SIAM J. Optim.* **19**(3), 1489–1509, 2008.

[23] J.J. Moreau. Proximité et dualitè dans un espace hilbertien *Bull. Soc. Math. Fr.* **93**, 273–299, 1965.

[24] Y. Nesterov. Excessive gap technique in nonsmooth convex optimization. *SIAM J. Optim.*, **16**(1), 235–249, 2005.

[25] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.* **103**(1), 127–152, 2005.

[26] Y. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Math. Program.* **110**(2), 245–259, 2005.

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

[28] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. *Doklady AN SSSR (translated as Soviet Math. Docl.)*, **269**, 543–547, 1983.

[29] F. Orabona, A. Argyriou and N. Srebro. PRISMA: PRoximal Iterative SMoothing Algorithm. *arXiv:1206.2372 [math.OC]*, 2012.

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

[31] S. Simons. *From Hahn–Banach to Monotonicity.* Springer, Berlin, 2008.

[32] B.C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Adv. Comp. Math.* **38**(3), 667–681, 2013.