SIAM J. Optim.
Vol. 22, No. 2, pp. 557–580

SMOOTHING AND FIRST ORDER METHODS: A UNIFIED FRAMEWORK[*]

AMIR BECK[†] and MARC TEBOULLE[‡]

Abstract. We propose a unifying framework that combines smoothing approximation with fast first order algorithms for solving nonsmooth convex minimization problems. We prove that independently of the structure of the convex nonsmooth function involved, and of the given fast first order iterative scheme, it is always possible to improve the complexity rate and reach an $O(\varepsilon^{-1})$ efficiency estimate by solving an adequately smoothed approximation counterpart. Our approach relies on the combination of the notion of smoothable functions that we introduce with a natural extension of the Moreau-infimal convolution technique along with its connection to the smoothing mechanism via asymptotic functions. This allows for clarification and unification of several issues on the design, analysis, and potential applications of smoothing methods when combined with fast first order algorithms.

Key words. nonsmooth convex minimization, smoothing methods, convex minimization, first order proximal gradients, rate of convergence, infimal convolution, asymptotic functions

AMS subject classifications. 90C25, 90C30

DOI. 10.1137/100818327

1. Introduction.

A well-known methodology for designing solution techniques to nonsmooth optimization (NSO) problems is to replace the original problem by a sequence of *approximating smooth* problems, which hopefully can be solved more efficiently than by using direct and classical schemes such as subgradient and bundle type methods [21]. The basic idea is to transform the nondifferentiable problem into a smooth problem. Many researchers have proposed different smoothing approaches to various classes of NSO problems. Some earlier works on the subject can be found, for example, in [16, 11, 12, 10], and for a more recent account, see, for instance, [3] and references therein.

This work is motivated by a paper of Nesterov [17], where a new method for minimizing a nonsmooth convex function over a convex compact finite-dimensional set is proposed. The characteristic feature of Nesterov's method is that for a special class of nonsmooth convex functions which are given as specific "max" type functions (see section 4), adopting the smoothing methodology combined with a fast gradient scheme for minimizing smooth convex functions also developed there, i.e., a method that shares a rate of convergence $O(1/k^2)$ for function values, where k is the iteration counter, an ε-optimal solution of the original nonsmooth problem can be obtained within $O(1/\varepsilon)$ iterations by solving its smoothed counterpart. This clearly outperforms usual subgradient-based schemes which when minimizing a Lipschitz continuous nonsmooth convex function reach an ε-optimal solution within $O(1/\varepsilon^2)$ iterations. It should be stressed that convergence rates are with respect to the objective function values and not with respect to the iterates. The main difference which explains this improvement relies on the fact that, as opposed to classical subgradient schemes that

[*] Received by the editors December 15, 2010; accepted for publication (in revised form) February 22, 2012; published electronically June 5, 2012.
http://www.siam.org/journals/siopt/22-2/81832.html

[†] Department of Industrial Engineering and Management, Technion, 32000 Haifa, Israel (becka@ie.technion.ac.il).

[‡] School of Mathematical Sciences, Tel Aviv University, 69978 Tel Aviv, Israel (teboulle@post.tau.ac.il).

are black-box oriented and applicable to *any* convex function, in the approach developed in [17], the special structure of the function to be minimized is beneficially exploited when combined with a peculiar fast gradient scheme.

This paper can be viewed as a natural complement and extension of Nesterov's framework, thus clarifying and unifying several issues on the design, analysis, and potential applications of smoothing methods when combined with fast first order algorithms. Our main observation is that independently of the structure of the convex nonsmooth function involved, and of the given fast first order iterative scheme used, it is always possible to improve the complexity rate for a broad class of NSO problems by solving its corresponding smoothed problem via any given fast first order scheme. Roughly speaking, first we show that the underlying and restrictive max-structure assumption of the nonsmooth convex function [17] can be removed, and second, we show that given *any* fast first order iterative method that is capable of producing an $O(1/k^2)$ rate of convergence will then naturally induce a method capable of solving a convex NSO model via its smoothed counterpart, with the improved complexity rate $O(1/\varepsilon)$, rather than the usual $O(1/\varepsilon^2)$ obtained by using standard subgradient/bundle schemes.

In this paper we adopt a *partial smoothing* approach in which (possibly) only a part of the nonsmooth component of the objective function is actually smoothed. More precisely, we will consider minimization problems of the form

$$\min_{\mathbf{x}}\{F(\mathbf{x}) + h_1(\mathbf{x}) + h_2(\mathbf{x})\},$$

where F is smooth and h_1, h_2 are nonsmooth. (The precise setting is given in section 3.) In the standard smoothing methodology, or *full* smoothing, both h_1 and h_2 are replaced by approximate smoothing functions H_1 and H_2, respectively, giving rise to the smoothed problem

$$\text{(CS)} \quad \min_{\mathbf{x}}\{F(\mathbf{x}) + H_1(\mathbf{x}) + H_2(\mathbf{x})\}.$$

However, here we will consider the partial smoothing approach in which only one of the nonsmooth functions, say, h_1, is smoothed while the other (h_2) is kept untouched:

$$\text{(PS)} \quad \min_{\mathbf{x}}\{F(\mathbf{x}) + H_1(\mathbf{x}) + h_2(\mathbf{x})\}.$$

The motivation for such an approach is twofold. First, with respect to the design and algorithmic analysis of the corresponding scheme, it stems from the fact that minimization problems of composite functions of the form

$$\text{(C)} \quad \min_{\mathbf{x}}\{F(\mathbf{x}) + h(\mathbf{x})\},$$

where F is smooth and h is nonsmooth, can be solved by fast gradient-based methods with an $O(1/k^2)$ rate of convergence despite the apparent nonsmoothness of the objective function. That is, the presence of the nonsmooth function does not alter the complexity bound; see the recent algorithms described in [8, 18]. In these algorithms, in addition to gradient computations of the smooth function F, a proximal-type operator of the nonsmooth function h is evaluated at each iteration. Therefore, in a sense, these are only *conceptual* algorithms since evaluating a proximal-type operator can be as difficult as solving the original problem. For a recent analysis on the effects of approximate computation of such operators within fast gradient schemes, see [14].

Nevertheless, there are some classes of interesting nonsmooth functions h for which the proximal operation is simple; see, for instance, the recent work of Combettes and Pesquet [13], which provides a thorough review of proximal-based algorithms and an important list of computable proximity operators arising in many applications. Second, in many of these applications [9, 13], one of the nonsmooth terms in the model (PS), say, h_2, plays a key role in describing a desirable property of the decision variable $\mathbf{x}$ which otherwise could be destroyed by smoothing. Thus, when the nonmooth function h_2 in problem (PS) belongs to the aforementioned class and plays a central role in modeling the problem at hand, it can and should be kept untouched; see section 5, which illustrates such a situation.

To achieve the aforementioned goals, we introduce and characterize mathematically the broad concept of "smoothable functions" for general convex functions; see section 2. In section 3, we begin by introducing the formulation of the nonsmooth optimization problem of interest, which encapsulates a broad class of NSO problems. We then formalize what we call a fast iterative method $\mathcal{M}$ for composite convex minimization problems of the form (C), and we establish that when applied to a partially smoothed version of the original nonsmooth problem with an adequate smoothing parameter expressed in terms of the problem's data, we always obtain an improved scheme with complexity $O(1/\varepsilon)$. To apply our results, we need a smoothing procedure for a *general* convex function. This is developed in section 4, where we provide a unifying and general approach which naturally extends the so-called Moreau proximal regularization of a convex function [16], and to eventually make a connection with another well-known approach for smoothing which is based on asymptotic functions [3], thus closing the loop of various existing smoothing procedures. This also allows us to explain, recover, and extend the smoothing approach of [17]. Throughout, we illustrate our results with a variety of examples. Finally, section 5 contains a numerical example accompanied with a theoretical justification that illustrates the potential advantage of the proposed partial smoothing approach over the full standard smoothing methodology.

1.1. Notation.

Throughout the paper we consider finite-dimensional normed vector spaces, denoted by $\mathbb{E}, \mathbb{F}, \mathbb{V}$, etc. For a vector space $\mathbb{E}$, the endowed norm is denoted by $\|\cdot\|_{\mathbb{E}}$ and the space of linear functionals is denoted by $\mathbb{E}^*$. The dual norm is denoted by either $\|\cdot\|_{\mathbb{E}}^*$ or $\|\cdot\|_{\mathbb{E}^*}$ and is defined as usual as $\|\mathbf{x}\|_{\mathbb{E}}^* = \|\mathbf{x}\|_{\mathbb{E}^*} = \max\{\langle \mathbf{u}, \mathbf{x}\rangle : \|\mathbf{u}\|_{\mathbb{E}} \leq 1\}$ for any $\mathbf{x} \in \mathbb{E}^*$. Here $\langle \mathbf{u}, \mathbf{x}\rangle$ for $\mathbf{u} \in \mathbb{E}^*$ and $\mathbf{x} \in \mathbb{E}$ denotes the value of the functional $\mathbf{u}$ at $\mathbf{x}$. The norm of a linear transformation $A : \mathbb{E} \to \mathbb{V}$, where $\mathbb{E}$ and $\mathbb{V}$ are finite-dimensional vector spaces with endowed norms $\|\cdot\|_{\mathbb{E}}$ and $\|\cdot\|_{\mathbb{V}}$, respectively, is given by

$$\|A\|_{\mathbb{E},\mathbb{V}} = \max\{\|A\mathbf{x}\|_{\mathbb{V}} : \|\mathbf{x}\|_{\mathbb{E}} = 1\}.$$

The subscript indicating the vector spaces will be omitted when their identity is obvious from the context. The vector of all ones is denoted by $\mathbf{e}$ where the dimension of the vector will be clear from the context. The set $\Delta_n = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{z}^T\mathbf{e} = 1, \mathbf{z} \geq \mathbf{0}\}$ is the unit simplex set. For a set C, we denote by δ_C the indicator function of the set, that is, $\delta_C(\mathbf{x}) = 0$ if $\mathbf{x} \in C$ and ∞ otherwise. For any function f and $\mathbf{x} \in E$, we denote the gradient of f at $\mathbf{x}$ to be the vector $\nabla f(\mathbf{x}) \in E^*$ for which

$$\lim_{\|\mathbf{d}\|\to 0} \frac{f(\mathbf{x}+\mathbf{d}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{d}\rangle}{\|\mathbf{d}\|} = 0.$$

When we say that a function $f : X \to \mathbb{R}$ is continuously differentiable on any given subset $X \subseteq \mathbb{E}$, it should be understood that this implicity assumes that there exists an open set containing X on which the derivatives are defined as usual. We also often use the standard notation $C^{1,1}$ for a function which is continuously differentiable with Lipschitz gradient. Finally, further standard definitions or notations in convex analysis which are not explicitly mentioned here can be found in the classical book [19].

2. Smoothable convex functions.

We begin by defining the concept of a *smoothable* function and the corresponding *smooth approximation* of a given nondifferentiable convex function.[1]

DEFINITION 2.1 (smoothable functions). *Let $g : \mathbb{E} \to (-\infty, \infty]$ be a closed and proper convex function and let $X \subseteq \operatorname{dom} g$ be a closed convex set. The function g is called "(α, β, K)-smoothable" over X if there exist β_1, β_2 satisfying $\beta_1 + \beta_2 = \beta > 0$ such that for every $\mu > 0$ there exists a continuously differentiable convex function $g_\mu : \mathbb{E} \to (-\infty, \infty)$ such that the following hold:*

(i) *$g(\mathbf{x}) - \beta_1\mu \le g_\mu(\mathbf{x}) \le g(\mathbf{x}) + \beta_2\mu$ for every $\mathbf{x} \in X$.*

(ii) *The function g_μ has a Lipschitz gradient over X with Lipschitz constant which is less than or equal to $K + \frac{\alpha}{\mu}$. That is, there exists $K \ge 0, \alpha > 0$, such that*

$$\|\nabla g_\mu(\mathbf{x}) - \nabla g_\mu(\mathbf{y})\|^* \le \left(K + \frac{\alpha}{\mu}\right) \|\mathbf{x} - \mathbf{y}\| \text{ for every } \mathbf{x}, \mathbf{y} \in X. \tag{2.1}$$

The function g_μ is called a "μ-smooth approximation" of g over X with parameters (α, β, K).

If a function is smoothable over the entire vector space $\mathbb{E}$, then it will just be called (α, β, K)-smoothable.

Remark 2.1. The choice of the decomposition of β as $\beta_1 + \beta_2$ is arbitrary. In fact, if

$$g(\mathbf{x}) - \beta_1\mu \le g_\mu(\mathbf{x}) \le g(\mathbf{x}) + \beta_2\mu \text{ for every } \mathbf{x} \in X,$$

then for every $\gamma \in \mathbb{R}$,

$$g(\mathbf{x}) - (\beta_1 - \gamma)\mu \le g_\mu(\mathbf{x}) + \gamma\mu \le g(\mathbf{x}) + (\beta_2 + \gamma)\mu \text{ for every } \mathbf{x} \in X,$$

which together with the fact that the Lipschitz constants of the gradients of $g_\mu + \gamma\mu$ and g_μ are the same implies that $g_\mu + \gamma\mu$ is also a μ-smooth approximation with parameters (α, β, K), but with $\beta_1 - \gamma, \beta_2 + \gamma$ taking the role of β_1, β_2 in property (i) of the definition.

It is quite easy to see that the following algebraic rules apply for smoothable functions.

LEMMA 2.1 (sum of smoothable functions). *Let γ_1, γ_2 be nonnegative constants and let g_1, g_2 be (α_1, β_1, K_1)- and (α_2, β_2, K_2)-smoothable functions, respectively, over some closed convex set X. Then $\gamma_1 g_1 + \gamma_2 g_2$ is a $(\gamma_1\alpha_1 + \gamma_2\alpha_2, \gamma_1\beta_1 + \gamma_2\beta_2, \gamma_1 K_1 + \gamma_2 K_2)$-smoothable function over X.*

Proof. Straightforward from the definition of smoothable functions. □

It is also important to understand the effect of a linear transformation of the variables on the parameters of a smoothable function.

[1] Note that the definition and properties discussed in this section remain valid for any closed proper function.

LEMMA 2.2 (linear transformation of a smoothable function). *Let $A : \mathbb{E} \to \mathbb{V}$ be a linear transformation. Let g be a (α, β, K)-smoothable function over a closed convex set $X \subseteq \mathbb{V}$, and let $\mathbf{b} \in \mathbb{V}$.*

Then the function $q : \mathbb{E} \to (-\infty, \infty)$ defined by $q(\mathbf{x}) = g(A\mathbf{x} + \mathbf{b})$ is a $(\alpha\|A\|^2, \beta, K\|A\|^2)$-smoothable function over $A^{-1}(X - \mathbf{b})$, where

$$\|A\| \equiv \|A\|_{\mathbb{E},\mathbb{V}} = \max\{\|A\mathbf{x}\|_{\mathbb{V}} : \|\mathbf{x}\|_{\mathbb{E}} = 1\}$$

and where A^{-1} is the inverse linear mapping defined by

$$A^{-1}(S) \equiv \{\mathbf{x} \in \mathbb{E} : A\mathbf{x} = \mathbf{s} \text{ for some } \mathbf{s} \in \mathrm{S}\}$$

for every $S \subseteq \mathbb{V}$.

Proof. First, let g_μ be a μ-smooth approximation of g with parameters (α, β, K) over X. Then there exists β_1, β_2 such that $\beta = \beta_1 + \beta_2$, for which it holds that

$$g(\mathbf{y}) - \beta_1\mu \le g_\mu(\mathbf{y}) \le g(\mathbf{y}) + \beta_2\mu$$

for any $\mathbf{y} \in X$. Making the change of variables $\mathbf{y} = A\mathbf{x} + \mathbf{b}$, where $\mathbf{x} \in A^{-1}(X - \mathbf{b})$, it follows that

$$g(A\mathbf{x} + \mathbf{b}) - \beta_1\mu \le g_\mu(A\mathbf{x} + \mathbf{b}) \le g(A\mathbf{x} + \mathbf{b}) + \beta_2\mu,$$

implying that

$$q(\mathbf{x}) - \beta_1\mu \le g_\mu(A\mathbf{x} + \mathbf{b}) \le q(\mathbf{x}) + \beta_2\mu$$

for any $\mathbf{x} \in A^{-1}(X - \mathbf{b})$. Therefore, property (i) of Definition 2.1 is satisfied with $q_\mu(\mathbf{x}) \equiv g_\mu(A\mathbf{x} + \mathbf{b})$. To verify property (ii) of the same definition, note that the gradient of q_μ is given by $A^*\nabla g_\mu(A\mathbf{x} + \mathbf{b})$ and for any $\mathbf{x}, \mathbf{y} \in A^{-1}(X - \mathbf{b})$ we have

$$\begin{aligned}
\|\nabla q_\mu(\mathbf{x}) - \nabla q_\mu(\mathbf{y})\|_{\mathbb{E}}^* &= \|A^*\nabla g_\mu(A\mathbf{x} + \mathbf{b}) - A^*\nabla g_\mu(A\mathbf{y} + \mathbf{b})\|_{\mathbb{E}}^* \\
&\le \|A\|\|\nabla g_\mu(A\mathbf{x} + \mathbf{b}) - \nabla g_\mu(A\mathbf{y} + \mathbf{b})\|_{\mathbb{V}}^* \\
&\le \left[\left(\frac{\alpha}{\mu} + K\right)\|A\|\right]\|A\mathbf{x} + \mathbf{b} - A\mathbf{y} - \mathbf{b}\|_{\mathbb{V}} \\
&\le \left[\left(\frac{\alpha}{\mu} + K\right)\|A\|^2\right]\|\mathbf{x} - \mathbf{y}\|_{\mathbb{E}},
\end{aligned}$$

establishing the desired result. ▯

There exist several ways to generate smooth approximations of nondifferentiable convex functions. This issue will be addressed in section 4.4, where we present a general framework for generating such smooth approximations. In the next section, we present a smoothing-based general minimization scheme.

3. A smoothing-based fast first order method.

We are interested in solving the convex problem (G) given by

$$(G) \quad H^* = \min\{H(\mathbf{x}) \equiv g(\mathbf{x}) + f(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}, \tag{3.1}$$

where the assumptions on the underlying functions are

- $h : \mathbb{E} \to (-\infty, \infty]$ is an extended valued closed proper convex function which is subdifferentiable over its domain which is denoted by $X = \operatorname{dom} h$;

- $f : X \to (-\infty, \infty)$ is a continuously differentiable function over X whose gradient is Lipschitz with constant L_f;
- $g : X \to (\infty, \infty]$ is a (α, β, K)-smoothable function over X.

Problem (G) is rich enough to cover many interesting generic optimization models by appropriate choices of (f, g, h). At a first glance, the use of three functions, with *two* being nonsmooth, and one smooth, might appear redundant. However, it is relevant since, as mentioned in the introduction, we will invoke what we call *partial smoothing*, namely, only the function g is smoothed, while the function h remains unchanged. Later on, in section 5, we will demonstrate the advantage of the partial smoothing approach in comparison to the "full" smoothing methodology, i.e., in which h would also be smoothed.

The *partially smoothed* problem is thus

$$(G_\mu) \quad H^*_\mu = \min\{H_\mu(\mathbf{x}) \equiv g_\mu(\mathbf{x}) + f(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}, \tag{3.2}$$

where g_μ is a μ-smooth approximation of g over X with parameters (α, β, K) for an appropriately chosen μ. Note that (G_μ) remains a nonsmooth problem, due to the presence of the nonsmooth function h.

The idea is now to be able to use any adequate algorithm for solving (G_μ). For that purpose we introduce the formal definition of a *fast* iterative method for solving the convex NSO problem,

$$(C) \qquad \min\{D(\mathbf{x}) \equiv F(\mathbf{x}) + h(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}, \tag{3.3}$$

which is assumed to have an optimal solution $\mathbf{x}^* \in \mathbb{E}$, and $D_* := D(\mathbf{x}^*)$ denotes its optimal value.

This problem is called the *input convex optimization model* and is characterized by the following data:

- h is an extended valued closed convex function which is subdifferentiable over its domain $\operatorname{dom} h$.
- F is a continuously differentiable convex function over $\operatorname{dom} h$ whose gradient is Lipschitz with constant L_F.

The input convex optimization model (C) is thus characterized by the triplet (F, h, L_F) satisfying the above premises.

Definition 3.1 (fast iterative method). *Let (F, h, L_F) be a given input convex optimization model with an optimal solution $\mathbf{x}^*$, and let $\mathbf{x}_0 \in \mathbb{E}$ be any given initial point. An iterative method $\mathcal{M}$ for solving problem* (C) *is called a fast method with constant $0 < \Lambda < \infty$, which possibly depends on $(\mathbf{x}_0, \mathbf{x}^*)$, if it generates a sequence $\{\mathbf{x}_k\}_{k \geq 0}$ satisfying for all $k \geq 1$,*

$$D(\mathbf{x}_k) - D^* \leq \frac{L_F \Lambda}{k^2}. \tag{3.4}$$

It is important to stress that within such a general setting, we are not preoccupied with the specific computations that are necessary—and that can be quite involved—to build the method $\mathcal{M}$; see also the remarks and discussion at the end of this section. Here, our interest and main observation is to establish that by applying *any* fast method $\mathcal{M}$ on the partially smoothed problem (G_μ) with an appropriately chosen smoothing parameter μ, an ε optimal solution can be obtained in no more than $O(1/\varepsilon)$ iterations, which, as mentioned in the introduction, is much better than the standard bound $O(1/\varepsilon^2)$ that would be obtained by using a usual black-box nonsmooth subgradient/bundle schemes on a nonsmooth problem (F, h, L_F) with a nontrivial function h.

The next result shows that the important idea of Nesterov [17, Theorem 3], to combine smoothing with a smooth optimization algorithm for specially structured max-type problems for reducing the complexity rate, can now be extended thanks to the concept of smoothable functions introduced in section 2 and independently of the given smooth optimization algorithm.

THEOREM 3.1. *Let $\{\mathbf{x}_k\}$ be the sequence generated by a fast iterative method $\mathcal{M}$ when applied to problem (G_μ), that is, to the input optimization problem $(f + g_\mu, h, L_{f+g_\mu})$. Suppose that the smoothing parameter is chosen as*

$$\mu = \sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta + (L_f + K)\varepsilon}}. \tag{3.5}$$

Then for

$$k \geq 2\sqrt{\alpha\beta\Lambda}\frac{1}{\varepsilon} + \sqrt{(L_f + K)\Lambda}\frac{1}{\sqrt{\varepsilon}}, \tag{3.6}$$

it holds that $H(\mathbf{x}_k) - H^* \leq \varepsilon$.

Proof. Let $\mu > 0$ and $F := f + g_\mu$. Using Definition 2.1(ii), one has $L_F = L_f + K + \frac{\alpha}{\mu}$. Therefore, the sequence generated by the method $\mathcal{M}$ when applied to (G_μ) satisfies for all $k \geq 1$

$$H_\mu(\mathbf{x}_k) - H_\mu^* \leq \left(L_f + K + \frac{\alpha}{\mu}\right)\frac{\Lambda}{k^2}. \tag{3.7}$$

Since g_μ is a μ-smooth approximation of g with parameters (α, β, K), by Definition 2.1(i) there exist β_1, β_2 satisfying $\beta_1 + \beta_2 = \beta > 0$ for which

$$H(\mathbf{x}) - \beta_1\mu \leq H_\mu(\mathbf{x}) \leq H(\mathbf{x}) + \beta_2\mu \text{ for any } \mathbf{x} \in \mathbb{E}.$$

Thus, in particular, the following inequalities hold:

$$H^* \geq H_\mu^* - \beta_2\mu \text{ and } H(\mathbf{x}_k) \leq H_\mu(\mathbf{x}_k) + \beta_1\mu, \quad k = 1, 2, \ldots,$$

and hence, together with (3.7) we obtain

$$H(\mathbf{x}_k) - H^* \leq H_\mu(\mathbf{x}_k) - H_\mu^* + (\beta_1 + \beta_2)\mu \leq (L_f + K)\frac{\Lambda}{k^2} + \left(\frac{\alpha\Lambda}{k^2}\right)\frac{1}{\mu} + \beta\mu. \tag{3.8}$$

Minimizing the right-hand side of (3.8) with respect to $\mu > 0$ we obtain

$$\mu = \sqrt{\frac{\alpha\Lambda}{\beta}}\frac{1}{k}. \tag{3.9}$$

Plugging the above expression for μ in (3.8), we obtain

$$H(\mathbf{x}_k) - H^* \leq (L_f + K)\frac{\Lambda}{k^2} + 2\sqrt{\alpha\beta\Lambda}\frac{1}{k}.$$

Thus, given $\varepsilon > 0$, to obtain an ε-optimal solution satisfying $H(\mathbf{x}_k) - H^* \leq \varepsilon$, it remains to find values of k for which

$$(L_f + K)\Lambda\frac{1}{k^2} + 2\sqrt{\alpha\beta\Lambda}\frac{1}{k} \leq \varepsilon. \tag{3.10}$$

Denoting $t := \sqrt{\Lambda}\frac{1}{k}$, the above inequality reduces to

$$(L_f + K)t^2 + 2\sqrt{\alpha\beta}t - \varepsilon \leq 0,$$

which is equivalent to (recall that $t > 0$)

$$\sqrt{\Lambda}\frac{1}{k} = t \leq \frac{-\sqrt{\alpha\beta} + \sqrt{\alpha\beta + (L_f + K)\varepsilon}}{L_f + K} = \frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta + (L_f + K)\varepsilon}}.$$

Using the value of the upper bound just established for $\sqrt{\Lambda}\frac{1}{k}$ in (3.9), we obtain the desired expression of μ stated in (3.5). We have thus shown that by choosing μ as in (3.5) and k satisfying

$$k \geq \frac{\sqrt{\alpha\beta\Lambda} + \sqrt{\alpha\beta\Lambda + (L_f + K)\varepsilon\Lambda}}{\varepsilon}, \tag{3.11}$$

we have $H(\mathbf{x}_k) - H^* \leq \varepsilon$. To complete the proof and obtain the desired lower bound for k as given in (3.6), note that for any $A, B \geq 0$, the following inequality holds:

$$\sqrt{A} + \sqrt{A + B} \leq 2\sqrt{A} + \sqrt{B}. \tag{3.12}$$

By invoking (3.12) with

$$A := \frac{\alpha\beta\Lambda}{\varepsilon^2}, B := \frac{(L_f + K)\Lambda}{\varepsilon},$$

together with (3.11), the desired result (3.6) follows. □

Remark 3.1. Note that the "optimal" smoothing parameter (3.5) does not depend on the constant Λ of the method. Nonetheless, this constant does appear in the expression for the bound on the number of iterations required to obtain an ε-optimal solution.

This paper is not concerned with the development and applications of fast iterative schemes and we refer the reader to the cited references below for more details, analysis, and applications. We end this section with a brief discussion on such schemes within our result established in Theorem 3.1. Current prominent methods that satisfy the premises of a fast iterative method $\mathcal{M}$ for solving problem (C) are *first order proximal gradient schemes*, i.e., methods that use information on the gradient of F and on the proximal mapping of h, or simply first order gradient schemes in the case $h \equiv 0$; see [18, 8] for the former and [4, 17] for the latter. All these methods share the same theoretical complexity rate $O(1/k^2)$ but are quite different with respect to their analysis and computational demands. For instance, the methods in [17, 18] requires two proximal steps based on two different proximal terms per iteration and accumulated memory of previous gradients, while the methods described in [4, 8] request only one proximal-based computation per iteration, thus providing computational saving. Finally, note that in all these first order methods, the complexity bounds involve an expression on some kind of distance between the initial point $\mathbf{x}_0 \in \mathbb{E}$ and an optimal solution $\mathbf{x}^*$, which has been quantified by the number Λ in Definition 3.1. For example, in the Euclidean setting, one has

$$\Lambda := \|\mathbf{x}^* - \mathbf{x}_0\|^2. \tag{3.13}$$

When $\operatorname{dom} h$ is assumed bounded, for the aforementioned first order schemes, it can be shown that the constant Λ is a positive finite real number. However, even when

$\operatorname{dom} h$ is not bounded, it is still possible to employ the smoothing-based $\mathcal{M}$ scheme, since as shown in Theorem 3.1 the smoothing parameter μ given in (3.5) is in fact *independent on the constant* Λ of the method $\mathcal{M}$.

In order to apply a smoothing-based algorithm, it is necessary to specify the smoothing approximation which is being used. This is done in the next section, which provides a unified and general approach for smoothing a *general* nonsmooth convex function.

4. Smoothing convex functions.

Nondifferentiable convex functions can be approximated by smooth functions by various techniques. One natural tool for generating an approximate smooth functional is through the use of the so-called proximal map introduced by Moreau [16]. One can construct a smoothed approximation to a given nonsmooth convex function f by taking its *infimal convolution* with the quadratic norm $||\cdot||^2$. Another general smoothing mechanism is obtained by using asymptotic (recession) functions [3]. Building on these fundamental tools, we propose a natural and unifying framework to smooth a general class of nonsmooth convex functions. This allows us to extend and connect these techniques as well as to recover recent smoothing proposed in [17]. Before continuing onto details, for the convenience of the reader we first recall some basic convex analysis results which will be essential for the analysis below.

4.1. Some convex analysis preliminaries.

The material in this section can be found in the standard convex analysis literature, e.g., [19, 20]. The notion of *conjugate function* is fundamental in our analysis and is recalled below.

Definition 4.1. *For any extended real-valued function $f : \mathbb{E} \to (-\infty, \infty]$, its convex conjugate $f^* : \mathbb{E}^* \to (-\infty, \infty]$ is defined by*

$$f^*(\mathbf{y}) = \sup_{\mathbf{x}\in\mathbb{E}} \{\langle \mathbf{x}, \mathbf{y}\rangle - f(\mathbf{x})\} = \sup \{\langle \mathbf{x}, \mathbf{y}\rangle - f(\mathbf{x}) : \; \mathbf{x} \in \operatorname{dom} f\}.$$

Moreover, if f is closed, proper, and convex on $\mathbb{E}$, then so is f^* and $f^{**} = f$.

Recall that a proper function $f : \mathbb{E} \to (-\infty, +\infty]$ is called σ-strongly convex with respect to $\|\cdot\|$ if there exists a constant $\sigma > 0$ (often called the modulus of strong convexity) such that

$$f((1-t)\mathbf{x} + t\mathbf{y}) \leq (1-t)f(\mathbf{x}) + tf(\mathbf{y}) - \frac{\sigma}{2}t(1-t)\|\mathbf{x}-\mathbf{y}\|^2 \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{E},\ t \in (0,1).$$

We will use an important equivalence between differentiability of a convex function and strong convexity of its conjugate, which is also known as the Baillon–Haddad theorem; see, e.g., [20, section 12H] and [5].

Lemma 4.1. *Let $\sigma > 0$. The following statements are equivalent:*

(a) *$h : \mathbb{E} \to \mathbb{R}$ is convex differentiable with ∇h which is Lipschitz continuous with respect to $\|\cdot\|_{\mathbb{E}}$ with constant $\frac{1}{\sigma}$.*

(b) *The conjugate $h^* : \mathbb{E}^* \to (-\infty, \infty]$ is σ-strongly convex with respect to $\|\cdot\|_{\mathbb{E}}^*$.*

4.2. The Moreau proximal smoothing.

One of the most popular approaches in the Euclidean setting (that is, when $\mathbb{E}$ is an Euclidean space with norm $\|\cdot\| = \sqrt{\langle,\cdot,\cdot\rangle}$) is the celebrated Moreau proximal approximation [16] that yields a family of approximations $\{g_\mu^{\text{px}}\}_{\mu>0}$ via

$$g_\mu^{\text{px}}(\mathbf{x}) = \inf_{\mathbf{u}\in\mathbb{E}} \left\{ g(\mathbf{u}) + \frac{1}{2\mu}\|\mathbf{u}-\mathbf{x}\|^2 \right\}, \tag{4.1}$$

where $g : \mathbb{E} \to (-\infty, \infty]$ is a closed and proper convex function.

As proved by Moreau [16], for any $\mu > 0$, the function g_μ^{px} enjoys several remarkable properties: it is convex continuous, finite-valued, and differentiable with gradient $\nabla g_\mu^{\mathrm{px}}$ which is Lipschitz continuous with constant $1/\mu$.

The Moreau approximation of a convex function is the so-called infimal convolution of f with the quadratic function $q(\mathbf{x}) = \frac{1}{2\mu}\|\mathbf{x}\|^2$, i.e.,

$$g_\mu^{\mathrm{px}}(\mathbf{x}) = \inf_{\mathbf{x}_1,\mathbf{x}_2}\{g(\mathbf{x}_1) + q(\mathbf{x}_2) : \ \mathbf{x}_1 + \mathbf{x}_2 = \mathbf{x}\} = \inf_{\mathbf{u}\in\mathbb{E}}\{g(\mathbf{u}) + q(\mathbf{x}-\mathbf{u})\} \equiv (g\square q)(\mathbf{x}).$$

Smoothing techniques that share resemblance to the infimal convolution operation where the quadratic squared distance is replaced by other distance-like functions can be found in other works. In particular we mention the work of Attouch and Wets [1], which was probably one of the first studies in that direction and inspired, for instance, the variants proposed in [23]. However, all the nice properties of the Moreau approximation alluded to above are not preserved in these generalizations.

A less popular though quite useful representation of Moreau proximal smoothing can be obtained through its dual formulation (see, e.g., [12, Proposition 3.4, p. 171]), which is simply derived by a direct application of the Fenchel duality theorem [19]:

$$g_\mu^{\mathrm{px}}(\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{E}^*}\left\{\langle \mathbf{y}, \mathbf{x}\rangle - g^*(\mathbf{y}) - \frac{\mu}{2}\|\mathbf{y}\|^2\right\}. \tag{4.2}$$

In essence, the above shows that Moreau smoothing is a natural tool to also smooth *conjugate functions*. This provides the starting point of the forthcoming results.

Here, we will consider a natural and simple extension of Moreau approximation of a convex function. It also relies on the notion of infimal convolution and allows us to derive a broad family of smooth approximations of convex functions that preserve the fundamental properties established by Moreau. But first, we consider below the smoothing approach given by Nesterov [17] developed for a class of nonsmooth functions that admit a "max representation" and show its direct relation to the Moreau proximal smoothing.

4.3. Nesterov's smoothing. Let us briefly recall the class of nonsmooth functions considered by Nesterov [17]. Let $\mathbb{E}, \mathbb{V}$ be finite-dimensional vector spaces, $Q \subseteq \mathbb{V}^*$ compact and convex, and ϕ some continuous convex function on $Q \subseteq \operatorname{dom}\phi$. The class of nonsmooth convex functions considered in [17] are given by

$$q(\mathbf{x}) = \max\left\{\langle \mathbf{u}, A\mathbf{x}\rangle - \phi(\mathbf{u}) : \mathbf{u} \in Q\right\}, \quad \mathbf{x} \in \mathbb{E}, \tag{4.3}$$

where $A : \mathbb{E} \to \mathbb{V}$ is *a linear map*.

The method suggested in [17] proposes the following smoothing methodology. A function d is called a *prox-function* of a given compact set C if $C \subseteq \operatorname{dom} d$ and d is a σ-strongly convex continuous function over the compact set C. The *prox center* is defined by $\mathbf{u}_0 = \operatorname{argmin}_{\mathbf{u}\in C} d(\mathbf{u})$ and it can be assumed without loss of generality that $\mathbf{u}_0 = \mathbf{0}$. In this setting it can be shown that $d(\mathbf{u}) \geq \frac{\sigma}{2}\|\mathbf{u} - \mathbf{u}_0\|_{\mathbb{V}^*}^2$ for every $\mathbf{u} \in C$.

The smooth approximation of q suggested in [17] is given by the convex function

$$q_\mu(\mathbf{x}) = \max\left\{\langle \mathbf{u}, A\mathbf{x}\rangle - \phi(\mathbf{u}) - \mu d(\mathbf{u}) : \mathbf{u} \in Q\right\}, \quad \mathbf{x} \in \mathbb{E}, \tag{4.4}$$

where $d(\cdot)$ is a prox-function for Q. It was then proved in [17, Theorem 1] that the convex function q_μ is $C^{1,1}(\mathbb{E})$ with Lipschitz continuous gradient with constant

$L_\mu = \|A\|^2/\sigma\mu$ and with gradient $\nabla q_\mu(\mathbf{x}) = A^* u_\mu(\mathbf{x})$, where $u_\mu(\mathbf{x})$ is the unique minimizer of (4.4).

A close inspection of the above result indicates that the smoothing procedure of [17] is a natural non-Euclidean extension of the *dual* Moreau smoothing approximation (4.2) of a conjugate function at the point $A\mathbf{x}$, whereby the squared norm in (4.2) has been replaced by a prox-function $d(\cdot)$ defined on $Q \subseteq \operatorname{dom}\phi$.

The result obtained in [17] and within the above interpretation appears to limit the class of convex functions that can be smoothed to be exclusively of the form of (4.3), i.e., conjugate-like convex functions. However, as we shall see now, this is not the case, and in the non-Euclidean setting, we will show that the infimal convolution operation of a given nonsmooth convex function with a properly defined smooth convex function remains the key player in smoothing *any* convex function without requiring any a priori special structure of the function to be smoothed.

4.4. The inf-conv smoothing technique. We are now ready to define the inf-conv μ-smooth approximation of a proper, closed, and convex function which is essentially an infimal convolution with a $C^{1,1}$ convex function.

Definition 4.2 (inf-conv μ-smooth approximation). *Let $g : \mathbb{E} \to (-\infty, \infty]$ be a closed proper convex function and let $\omega : \mathbb{E} \to \mathbb{R}$ be a $C^{1,1}$ convex function with Lipschitz gradient constant $1/\sigma$ ($\sigma > 0$). Suppose that for any $\mu > 0$ and any $\mathbf{x} \in \mathbb{E}$, the following infimal convolution is finite:*

$$g_\mu^{\mathrm{ic}}(\mathbf{x}) = \inf_{\mathbf{u}\in\mathbb{E}} \left\{ g(\mathbf{u}) + \mu\omega\left(\frac{\mathbf{x}-\mathbf{u}}{\mu}\right)\right\} = (g \square \omega_\mu)(\mathbf{x}), \tag{4.5}$$

where

$$\omega_\mu(\cdot) \equiv \mu\omega\left(\frac{\cdot}{\mu}\right). \tag{4.6}$$

Then g_μ^{ic} is called the inf-conv μ-smooth approximation of g.

Note that the assumption that $g\square\omega_\mu$ is a finite-valued function is satisfied, for example, when ω has bounded level sets and g satisfies that $\min_{\mathbf{x}\in\mathbb{E}} g(\mathbf{x}) > -\infty$. It is also satisfied when $\omega(\cdot) = c\|\cdot\|_{\mathbb{E}}^2$ for any constant $c > 0$.

We are now ready to recall some of the main properties of the inf-conv μ-smooth approximation of a convex function. A self-contained simple proof is given in the appendix for the sake of completeness; see also Remark 4.1 below.

Theorem 4.1. *Consider the setting of Definition 4.2. Then,*

(a) *the following "dual" formulation for g_μ^{ic} holds:*

$$g_\mu^{\mathrm{ic}}(\mathbf{x}) = (g^* + \omega_\mu^*)^*(\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{E}^*} \{\langle \mathbf{y}, \mathbf{x}\rangle - g^*(\mathbf{y}) - \mu\omega^*(\mathbf{y})\}; \tag{4.7}$$

(b) *g_μ^{ic} is differentiable and with gradient $\nabla g_\mu^{\mathrm{ic}}$ which is Lipschitz with constant $\frac{1}{\sigma\mu}$;*

(c) *let $\mathbf{x} \in \mathbb{E}$, and suppose that the minimum in* (4.5) *is attained at the point $\mathbf{u}_\mu(\mathbf{x})$; then*

$$\nabla g_\mu^{\mathrm{ic}}(\mathbf{x}) = \nabla\omega\left(\frac{\mathbf{x}-\mathbf{u}_\mu(\mathbf{x})}{\mu}\right) = \nabla\omega_\mu(\mathbf{x} - \mathbf{u}_\mu(\mathbf{x})). \tag{4.8}$$

Proof. See Appendix A. □

Remark 4.1. As noted earlier, the use of infimal convolution to smooth convex functions is very well known and originates from the seminal work of Moreau [16]. For a detailed and modern account of such results, see the recent comprehensive monograph of Bauchke and Combettes [6] and relevant references therein. The proof of the differentiability of g_μ and part (c) of Theorem 4.1 can also be found in [6, Proposition 18.7]. We also note that properties (a) and (b) of Theorem 4.1 were shown in Theorem 1 of [17].

Remark 4.2. It should be noted that part (a) of the theorem, namely, the dual formulation, is always true for any proper closed convex function g and *any* finite valued convex function ω.

Remark 4.3. Note that since ω^* is strongly convex, it follows that the maximization problem in (4.7) has a unique maximizer. Denote this maximizer by $\mathbf{y}_\mu(\mathbf{x})$. Then by [19, Corollary 23.5.1] it follows that $\mathbf{y}_\mu(\mathbf{x}) = \nabla g_\mu^{\text{ic}}(\mathbf{x})$, which combined with part (c) of Theorem 4.1 implies that

$$\mathbf{y}_\mu(\mathbf{x}) = \nabla\omega_\mu(\mathbf{x} - \mathbf{u}_\mu(\mathbf{x})).$$

So far we have shown that g_μ^{ic} satisfies property (ii) of Definition 2.1 of a μ-smooth approximation, and in fact we have shown that the Lipschitz condition there (2.1) is satisfied for all $\mathbf{x} \in \mathbb{E}$ and not specifically on a certain closed convex subset $X \subseteq \mathbb{E}$. It remains to detect conditions under which g_μ^{ic} also satisfies property (i) of Definition 2.1.

LEMMA 4.2. *Consider the setting of Definition* 4.2 *and let* $X \subseteq \mathbb{E}$ *be a closed convex set. Suppose that* g *is subdifferentiable over* X. *Then for any* $\mu > 0$ *and* $\mathbf{x} \in X$ *the following holds:*

$$g(\mathbf{x}) - \mu\omega^*(\gamma_{\mathbf{x}}) \le g_\mu^{\text{ic}}(\mathbf{x}) \le g(\mathbf{x}) + \mu\omega(\mathbf{0}),$$

where $\gamma_{\mathbf{x}} \in \partial g(\mathbf{x})$ *is a subgradient of* g *at* $\mathbf{x}$.

Proof. By the definition of g_μ^{ic} one has

$$\begin{aligned} g_\mu^{\text{ic}}(\mathbf{x}) &= \inf_{\mathbf{u}\in\mathbb{E}} \left\{ g(\mathbf{u}) + \mu\omega\left(\frac{\mathbf{x}-\mathbf{u}}{\mu}\right)\right\} \\ &\le g(\mathbf{x}) + \mu\omega\left(\frac{\mathbf{x}-\mathbf{x}}{\mu}\right) = g(\mathbf{x}) + \mu\omega(\mathbf{0}). \end{aligned}$$

For the opposite inequality, we can use the subgradient inequality for g to obtain that for every $\mathbf{x} \in X$

$$\begin{aligned} g_\mu^{\text{ic}}(\mathbf{x}) - g(\mathbf{x}) &= \min_{\mathbf{u}\in\mathbb{E}} \left\{ g(\mathbf{u}) - g(\mathbf{x}) + \mu\omega\left(\frac{\mathbf{x}-\mathbf{u}}{\mu}\right)\right\} \\ &\ge \min_{\mathbf{u}\in\mathbb{E}} \left\{ \langle\gamma_{\mathbf{x}}, \mathbf{u}-\mathbf{x}\rangle + \mu\omega\left(\frac{\mathbf{x}-\mathbf{u}}{\mu}\right)\right\} \\ &= \min_{\mathbf{z}\in\mathbb{E}} \{-\langle\gamma_{\mathbf{x}}, \mathbf{z}\rangle + \omega_\mu(\mathbf{z})\} \\ &= -\max_{\mathbf{z}\in\mathbb{E}} \{\langle\gamma_{\mathbf{x}}, \mathbf{z}\rangle - \omega_\mu(\mathbf{z})\} \\ &= -\mu\omega^*(\gamma_{\mathbf{x}}). \qquad \square \end{aligned}$$

The above result readily implies that if $\max_{\mathbf{x}\in X} \omega^*(\gamma_{\mathbf{x}}) < \infty$, then property (i) of a smooth approximation as given in Definition 2.1 will be satisfied. This is recorded in Corollary 4.1, which states essentially that *any* convex function is smoothable over closed convex sets on which its subgradients are bounded.

Corollary 4.1. *Consider the setting of Definition* 4.2 *and let $X \subseteq \mathbb{E}$ be a closed convex set. Suppose that*

$$D[g,\omega^*] = \sup_{\mathbf{x}\in X} \sup_{\mathbf{d}\in\partial g(\mathbf{x})} \omega^*(\mathbf{d}) < \infty. \tag{4.9}$$

Then for any $\mu > 0$, g_μ^{ic} is a μ-smooth approximation of g over X with parameters

$$\left(\frac{1}{\sigma}, D[g,\omega^*] + \omega(\mathbf{0}), 0\right).$$

Remark 4.4. Note that (4.9) could also be defined by replacing the supremum over $\mathbf{d} \in \partial g(\mathbf{x})$ by an infimum.

Going back to the special form of nonsmooth functions q considered by [17] and given in (4.3), let us show how the smoothing (4.4) can be recovered. First, note that the function to be smoothed can be written as $q(\mathbf{x}) = g(A\mathbf{x})$, where

$$g := (\tilde{\phi})^* \text{ and } \tilde{\phi} := \phi + \delta_Q.$$

Now, let $\tilde{d} := d + \delta_Q$. Since d is given σ-strongly convex, so is $\tilde{d}$, and by Lemma 4.1 it follows that $(\tilde{d})^* \in C^{1,1}(\mathbb{E})$. Defining $\omega = (\tilde{d})^*$, we can invoke Theorem 4.1 to get the corresponding inf-conv μ-smooth approximation:

(4.10)

$$g_\mu^{\text{ic}}(\mathbf{x}) \overset{(4.7)}{=} (g^* + \omega_\mu^*)^*(\mathbf{x}) = (\tilde{\phi} + \mu\tilde{d})^*(\mathbf{x}) = \max_{\mathbf{u}} \{\langle \mathbf{u}, \mathbf{x}\rangle - \phi(\mathbf{u}) - \mu d(\mathbf{u}) : \mathbf{u} \in Q\}.$$

By Corollary 4.1 (and the identity (4.9)), it follows that g_μ^{ic} is a μ-smooth approximation of g with parameters $(\frac{1}{\sigma}, D, 0)$, where $D = \max\{d(\mathbf{u}) : \mathbf{u} \in Q\}$. (Note that here $\omega(\mathbf{0}) = (\tilde{d})^*(\mathbf{0}) = 0$, by definition of the prox center for d.) Now clearly, the formulation (4.10) implies that q_μ given in (4.4) is nothing else but $q_\mu(\mathbf{x}) = g_\mu^{\text{ic}}(A\mathbf{x})$, and hence by Lemma 2.2 it follows that q_μ is a μ-smooth approximation of q with parameters $(\frac{\|A\|^2}{\sigma}, D, 0)$, where $\|A\| = \max\{\|A\mathbf{x}\|_{\mathbb{V}} : \|\mathbf{x}\|_{\mathbb{E}} = 1\}$, thus recovering the result of [17].

The following two examples illustrate well-known instances of smooth approximation.

Example 4.1 (Euclidean norm function). Let $\mathbb{E} = \mathbb{R}^n$ endowed with the l_2 norm $\|\cdot\| = \|\cdot\|_2$. Consider the setting

$$g(\mathbf{x}) = \|\mathbf{x}\|, X = \mathbb{E}, \omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2.$$

Then

$$\omega(\mathbf{0}) = 0, D[g,\omega] = \frac{1}{2}$$

and ω is 1-strongly convex, which implies by Corollary 4.1 that g_μ^{ic}, which in this case is the same as g_μ^{px}, is a μ-smooth approximation of g (over $\mathbb{R}^n$) with parameters $\left(1, \frac{1}{2}, 0\right)$. It is easy to see that for every $\mu > 0$ the smooth approximation is given by

$$g_\mu^{\text{px}}(\mathbf{x}) = \min_{\mathbf{u}} \left\{\|\mathbf{u}\| + \frac{1}{2\mu}\|\mathbf{u} - \mathbf{x}\|^2\right\} = \begin{cases} \frac{\|\mathbf{x}\|^2}{2\mu}, & \|\mathbf{x}\| \le \mu, \\ \|\mathbf{x}\| - \frac{\mu}{2} & \text{else,} \end{cases} \tag{4.11}$$

which is the so-called Huber function in $\mathbb{R}^n$ [15].

Example 4.2 (l_1 norm function). Consider the same vector space $\mathbb{E}$ and underlying function ω as in the previous example and let $g(\mathbf{x}) = \|\mathbf{x}\|_1$. Then $\omega(\mathbf{0}) = 0, D[g,\omega] = \frac{n}{2}$, and hence $g_\mu^{\rm px}$, which in this case is the sum of Huber functions on each of the components

$$g_\mu^{\rm px}(\mathbf{x}) = \sum_{i=1}^n H_\mu(x_i), \qquad \left(H_\mu(y) \equiv \begin{cases} \frac{y^2}{2\mu} & |y| \le \mu, \\ |y| - \frac{\mu}{2} & \text{else}, \end{cases} \right),$$

is a μ-smooth approximation of g (over $\mathbb{R}^n$) with parameters $(1, \frac{n}{2}, 0)$.

The next example describes a function which is smoothable (via the prox operation) only over bounded sets of the space.

Example 4.3. Consider the setting $\mathbb{E} = \mathbb{R}, \|\cdot\| = |\cdot|$ and let $g(x) = \max\{|x|, x^2\}$. The subdifferential set of the function is given by

$$\partial g(x) = \begin{cases} [-1,1], & x = 0, \\ \{\text{sign}(x)\}, & 0 < |x| < 1, \\ [1,2], & x = 1, \\ [-2,-1], & x = -1, \\ \{2x\}, & |x| > 1. \end{cases}$$

Clearly, the subgradients of g are not bounded over the entire real line $\mathbb{R}$. We will consider the prox-based smooth approximation over $X = [-2,2]$ which is now computed:

$$\begin{aligned} g_\mu^{\rm px}(x) &= \min_u \left\{ \max\{|u|, u^2\} + \frac{1}{2\mu}(u-x)^2 \right\} \\ &= \min_{\alpha \ge 0} \min_{u:|u|=\alpha} \left\{ \max\{\alpha, \alpha^2\} + \frac{1}{2\mu}(\alpha^2 - 2xu + x^2) \right\} \\ &= \min_{\alpha \ge 0} \left\{ \max\{\alpha, \alpha^2\} + \frac{1}{2\mu}(\alpha^2 - 2\alpha|x| + x^2) \right\} \\ &= \begin{cases} \frac{x^2}{2\mu}, & |x| < \mu, \\ |x| - \frac{\mu}{2}, & \mu \le |x| < \mu + 1, \\ 1 + \frac{1}{2\mu}(1 - |x|)^2, & \mu + 1 \le |x| < 2\mu + 1, \\ \frac{x^2}{2\mu+1}, & |x| \ge 2\mu + 1. \end{cases} \end{aligned}$$

By Corollary 4.1, the above function is a μ-smooth approximation of g over X with parameters $(1,2,0)$. The function and its approximations $g_{0.5}^{\rm px}, g_{0.1}^{\rm px}$ are shown in Figure 1.

More examples in the *non-Euclidean* setting are given in the next section.

4.5. Smoothing via asymptotic functions. Another general approach to smooth nondifferentiable functions is via the concept of recession or asymptotic functions. This approach was introduced in [10], where it was observed that many optimization problems may be formulated in the form

$$(K) \qquad \inf\{u_\infty(f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})) : \ \mathbf{x} \in \mathbb{E}\},$$

where u_∞ is the asymptotic function of some given function u. This was broadly extended with more general results in [2]; see also [3, Chapter 3] for an overview and more references.

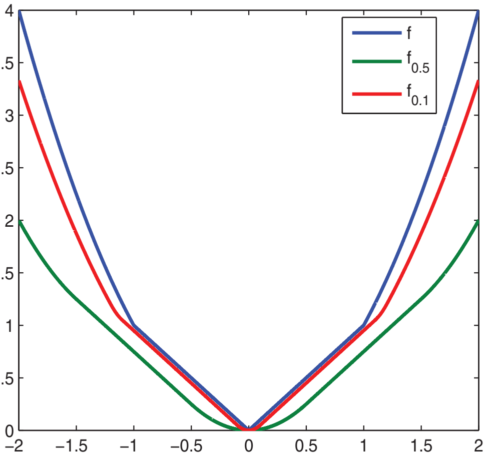

FIG. 1. *The function* $g(x) = \max\{|x|, x^2\}$ *and its smooth approximations* $g_{0.5}^{\mathrm{px}}, g_{0.1}^{\mathrm{px}}$.

Here, we will show that there exists an interesting close relation between the asymptotic function-based smoothing and the inf-conv μ-smooth approximation of a convex function discussed in the previous section.

First, we briefly recall the notion of an asymptotic function; see, e.g., [19, 3]. For $u : \mathbb{E} \to (-\infty, +\infty]$ proper, closed, and convex, its asymptotic function u_∞ is a closed proper convex function on $\mathbb{E}$ which is given by

$$u_\infty(d) = \lim_{\mu \to 0^+} \left\{ u_\mu(\mathbf{d}) := \mu u\left(\frac{\mathbf{d}}{\mu}\right) \right\} \text{ for every } \mathbf{d} \in \operatorname{dom} u. \tag{4.12}$$

The asymptotic function u_∞ is positively homogeneous[2] with $u_\infty(\mathbf{0}) = 0$.

Relation (4.12) was the basis used in [10] to naturally suggest approximating the problem (K) whereby u_∞ is replaced by u_μ.

We will now show that the inf-conv μ-smooth approximation of a convex function has a simple and special structure when the function g to be smoothed is the asymptotic function of the finite-valued convex function $\omega : \mathbb{E} \to (\infty, +\infty)$ satisfying the premises of Definition 4.2.

Before stating our result, we record in the next lemma the following fundamental property of the conjugate of an asymptotic function; see, e.g., [3, Theorem 2.5.4(b), p. 55] for a proof.

LEMMA 4.3. *Let* $u : \mathbb{E} \to (-\infty, +\infty]$ *be a closed proper convex function, and let* u^* *be its convex conjugate. Then* $(u_\infty)^* = \delta_{\mathrm{cl}\,\mathrm{dom} u^*}$, *where* cl *stands for the closure operation.*

Following [10], we make the following assumption.

Assumption 1. For any $\mu > 0$

$$\mu\omega\left(\frac{\mathbf{x}}{\mu}\right) \geq \omega_\infty(\mathbf{x}) \text{ for all } \mathbf{x}.$$

We are ready to state our result.

[2] A function p is positively homogeneous on $\mathbb{E}$ if $\mathbf{0} \in \operatorname{dom} p$ and $p(t\mathbf{u}) = tp(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{E}$ and all $t > 0$.

THEOREM 4.2. *Let $\omega : \mathbb{E} \to \mathbb{R}$ be a $C^{1,1}$ convex function with Lipschitz gradient constant $1/\sigma$, and let g be a convex finite-valued function over $\mathbb{E}$. Suppose that Assumption 1 holds and let $g = \omega_\infty$. Then for any $\mu > 0$,*

$$g_\mu^{\mathrm{ic}}(\mathbf{x}) = \mu\omega\left(\frac{\mathbf{x}}{\mu}\right) \text{ for every } \mathbf{x} \in \mathbb{E}.$$

Moreover, the function g_μ^{ic} is a μ-smooth approximation of g with parameters $(\frac{1}{\sigma\mu}, \omega(\mathbf{0}), 0)$.

Proof. With $g = \omega_\infty$, using Theorem 4.1(a) we have

$$\begin{aligned} g_\mu^{\mathrm{ic}}(\mathbf{x}) &= (g^* + \omega_\mu^*)^*(\mathbf{x}) = ((\omega_\infty)^* + \omega_\mu^*)^*(\mathbf{x}), \\ &= (\delta_{\mathrm{cl\,dom}\,\omega^*} + \omega_\mu^*)^* \text{ [by Lemma 4.3]}, \\ &= (\omega_\mu^*)^*(\mathbf{x}) \text{ [since } \mathrm{dom}\,\omega_\mu^* = \mathrm{dom}\,\omega^*], \\ &= \omega_\mu(\mathbf{x}) = \mu\omega\left(\frac{\mathbf{x}}{\mu}\right) \text{ [since } \omega_\mu \text{ is continuous convex]}, \end{aligned}$$

proving the claimed formula for g_μ^{ic} in the theorem. Now, by Theorem 4.1 it follows that $\nabla g_\mu^{\mathrm{ic}}$ is Lipschitz with constant $\frac{1}{\sigma\mu}$. Invoking Lemma 4.2 and using Assumption 1, it follows that for any $\mathbf{x} \in \mathbb{E}$,

$$g(\mathbf{x}) = \omega_\infty(\mathbf{x}) \le g_\mu^{\mathrm{ic}}(\mathbf{x}) \le g(\mathbf{x}) + \mu\omega(\mathbf{0}),$$

and the desired result is proved. □

4.6. Examples. We now illustrate our results within some interesting examples which have been commonly used and are well known in the smoothing literature; see [12, 10, 2, 3] and references therein. The last example, Example 4.9, illustrates a situation where Nesterov's smoothing framework is not applicable, since the function to be smoothed cannot be represented in the max-formulation of (4.3) through a linear map A.

Example 4.4 (smoothing of the max function). Consider the space $\mathbb{E} = \mathbb{R}^n$ with the endowed norm $\|\cdot\| = \|\cdot\|_\infty$. The function $g(\mathbf{x}) = \max\{x_1, \ldots, x_n\}$ is an asymptotic function of $\omega(\mathbf{x}) = \log\left(\sum_{i=1}^n e^{x_i}\right)$. The gradient of ω, $\nabla\omega$, is Lipschitz with constant 1. To show this, simply notice that $\omega^*(\mathbf{y}) \equiv \sum_{i=1}^n y_i \log y_i$ with $\mathrm{dom}\,\omega^* = \Delta_n$, which is a 1-strongly convex function with respect to the l_1 norm (see, e.g., [7]) and therefore $\omega = \omega^{**}$ has a gradient which is Lipschitz with respect to the l_∞ norm with constant 1. In addition, $\omega(0) = \log(n)$ and it is easy to see [10] that for any $\mu > 0$ and $\mathbf{x} \in \mathbb{E}$

$$g(\mathbf{x}) \le \mu\omega\left(\frac{\mathbf{x}}{\mu}\right),$$

and thus, invoking Theorem 4.2, we have that $\mu\omega(\frac{\mathbf{x}}{\mu}) = \mu\log(\sum_{i=1}^n e^{x_i/\mu})$ is a μ-smooth approximation of $\max\{x_1, \ldots, x_n\}$ with parameters $(1, \log n, 0)$.

Example 4.5 (max of linear functions). Recall that by Lemma 2.2 smoothability is preserved under linear transformations of the variables. For instance, if we consider a function which is the maximum of affine functions $g(\mathbf{x}) = \max\{\mathbf{a}_1^T\mathbf{x} + b_1, \ldots, \mathbf{a}_m^T\mathbf{x} + b_m\}$ over the space $\mathbb{R}^n$ with the endowed norm $\|\cdot\|_1$, then by Example 4.4 and Lemma 2.2, a μ-smooth approximation of this function will be

$$g_\mu(\mathbf{x}) = \mu\log\left(\sum_{i=1}^m e^{(\mathbf{a}_i^T\mathbf{x} + b_i)/\mu}\right).$$

The parameters of the above μ-smooth approximation (computed with respect to the l_∞ norm) are $(\|\mathbf{A}\|^2, \log m, 0)$, where $\mathbf{A}$ is the $m \times n$ matrix whose rows are $\mathbf{a}_i^T$ and

$$\|A\| = \max\{\|\mathbf{A}\mathbf{x}\|_\infty : \|\mathbf{x}\|_1 = 1\} = \max_{i,j} |A_{i,j}|.$$

The above smoothing of the max function gives rise to a smoothing of the absolute value function which can be also rewritten as the max function $g(x) = |x| = \max\{x, -x\}$, that is, $g(x) = q(\mathbf{A}x)$, where $q(x_1, x_2) := \max\{x_1, x_2\}$ and $\mathbf{A} = \binom{1}{-1}$. The corresponding μ-smooth approximation is $g_\mu(x) = q_\mu(Ax) = \mu \log(\frac{1}{2}(e^{x/\mu} + e^{-x/\mu}))$ with parameters $(1, \log 2, 0)$ since $\|A\| = 1$.

Example 4.6 (smoothing of the l_1 norm). Consider the space $\mathbb{E} = \mathbb{R}^n$ with the endowed norm $\|\cdot\|_2$. Let $g(\mathbf{x}) = \|\mathbf{x}\|_1$ and $\omega(\mathbf{x}) = \sum_{i=1}^n \sqrt{1 + x_j^2}$. Then g is an asymptotic function of ω. The gradient $\nabla\omega$ is Lipschitz with constant 1, $\omega(\mathbf{0}) = n$, and again (cf. [10]) we have for any $\mu > 0$ and $\mathbf{x} \in \mathbb{E}$ that

$$g(\mathbf{x}) \le \mu\omega\left(\frac{\mathbf{x}}{\mu}\right),$$

which, invoking Theorem 4.2, implies that the function $\sum_{i=1}^n \sqrt{\mu^2 + x_j^2}$ is a μ-smooth approximation of $\|\mathbf{x}\|_1$ with parameters $(1, n, 0)$.

Example 4.7. Consider the setting $\mathbb{E} = \mathbb{R}, \|\cdot\|_\mathbb{E} = |\cdot|$. By the previous example, $\sqrt{y^2 + \mu^2}$ is a μ-smooth approximation of the absolute values with parameters $(1, 1, 0)$. Overall, we have encountered three μ-smooth approximations of the absolute value function: $\sqrt{y^2 + \mu^2}$, $\mu \log\left(\frac{1}{2}\left(e^{y/\mu} + e^{-y/\mu}\right)\right)$, and Huber's function $\begin{cases} \frac{y^2}{2\mu}, & |y| \le \mu, \\ |y| - \frac{\mu}{2} & \text{else} \end{cases}$ with parameters

$$(1, 1, 0), (1, \log 2, 0), (1, 0.5, 0),$$

respectively. Therefore, it is not surprising that, as can be seen in Figure 2, Huber's function is the best approximation and the square-root-based approximation is the worst (has the largest β).

Example 4.8. Let $g(\mathbf{x}) = \|\mathbf{x}\|$ and $\omega(\mathbf{x}) = \sqrt{1 + \|\mathbf{x}\|^2}$ over the space $\mathbb{E} = \mathbb{R}^n$ with the endowed l_2 norm. Then $\omega^*(\mathbf{y}) = -\sqrt{1 - \|\mathbf{y}\|^2}$ with $\operatorname{dom}\omega^* = \{\mathbf{y} : \|\mathbf{y}\| \le 1\}$ and g is of course the asymptotic function of ω. In addition, ω satisfies Assumption 1 and thus by Theorem 4.2 it follows that $\mu\omega(\frac{\mathbf{x}}{\mu}) = \sqrt{\mu + \|\mathbf{x}\|^2}$ is a μ-smooth approximation of $\|\mathbf{x}\|$ with parameters $(1, 1, 0)$, which is a slightly worse approximation than Huber's function recalled in Example 4.1.

Example 4.9 (maximum of convex functions). For a given integer $m > 1$, consider the convex function

$$g(\mathbf{x}) = \max\left\{f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})\right\},$$

where $f_1, \ldots, f_m$ are m continuously differentiable convex functions over a compact convex set $X \subseteq \mathbb{E}$ with Lipschitz gradients over X with constants $L_{f_1}, \ldots, L_{f_m}$, respectively. The vector space $\mathbb{E}$ has a norm denoted by $\|\cdot\|_\mathbb{E}$. Clearly, the function g can also be rewritten as

$$g(\mathbf{x}) = \max_{\lambda \in \Delta_m} \sum_{i=1}^m \lambda_i f_i(\mathbf{x});$$

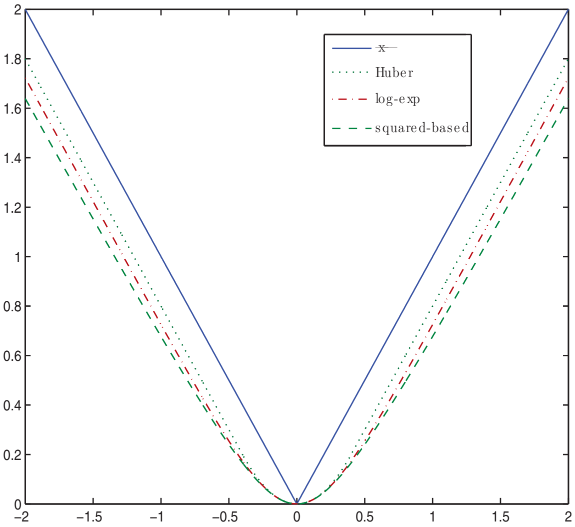

FIG. 2. *The function* $|y|$ *and its smooth approximations with* $\mu = 0.2$. *"Huber" stands for the Huber function given in* (4.11), *"log-exp" is the function* $\mu \log(\frac{1}{2}(e^{y/\mu}+e^{-y/\mu}))$, *and "squared-based" is the function* $\sqrt{y^2+\mu^2}-\mu$.

however, since $f_i(\cdot)$ are *nonlinear*, the smoothing framework of [17] (cf. (4.3)) cannot be applied.

Denoting the max function by

$$h(\mathbf{z}) \equiv \max\{z_1, \dots, z_m\}, \ (\mathbf{z} \in \mathbb{R}^m),$$

the function g can be rewritten as $g(\mathbf{x}) = h(f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))$. Recall that by Example 4.4,

$$h_\mu(\mathbf{z}) = \mu \log\left(\sum_{i=1}^{m} e^{z_i/\mu}\right)$$

is a μ-smooth convex approximation of the max function h with parameters $(1, \log(m), 0)$ over the space $\mathbb{V}$ which is defined as $\mathbb{R}^m$ endowed with the norm $\|\cdot\|_\infty$.

The next result shows that a μ-smooth convex approximation of the function g over X is given by[3]

$$g_\mu(\mathbf{x}) = h_\mu(f_1(\mathbf{x}), \dots, f_m(\mathbf{x})) = \mu \log\left(\sum_{i=1}^{m} e^{f_i(\mathbf{x})/\mu}\right).$$

Proposition 4.1 below computes the parameters (α, β, K) for which this approximation g_μ will satisfy the premises of Definition 2.1. The proof of the proposition, which is rather technical, is given in Appendix B.

PROPOSITION 4.1. *Let* $f_1, \dots, f_m$ *be* m *continuously differentiable convex functions over a compact convex set* $X \subseteq \mathbb{E}$ *whose gradients are Lipschitz over* X *with constants* $L_{f_1}, \dots, L_{f_m}$, *respectively. Let*

$$g(\mathbf{x}) = \max\{f_1(\mathbf{x}), \dots, f_m(\mathbf{x})\}.$$

[3] Recall that the convexity of a composite function $h(f_1, \dots, f_m)$ is preserved when f_i are convex and h is convex and isotone, i.e., $u_i \le v_i, \ i = 1, \dots, m$, implies $h(\mathbf{u}) \le h(\mathbf{v})$.

Then for every $\mu > 0$ the function

$$g_\mu(\mathbf{x}) = \mu \log\left(\sum_{i=1}^{m} e^{f_i(\mathbf{x})/\mu}\right)$$

is a μ-smooth approximation of g with parameters

$$\left(\max_{i=1,\ldots,m} M_{f_i}^2, \log(m), \max_{i=1,\ldots,m} L_{f_i}\right),$$

where $M_{f_i} := \max\{\|\nabla f_i(\mathbf{x})\|_{\mathbb{E}}^* : \mathbf{x} \in X\}, i = 1, \ldots, m.$

5. To smooth or not to smooth?

In this section we will demonstrate through a numerical example the advantage of *partial* smoothing in comparison to *full* smoothing. Consider the following $l_1 - l_1$ least fitting problem on the vector space $\mathbb{R}^n$ endowed with the norm $\|\cdot\|_1$:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{M(\mathbf{x}) \equiv \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1 + \|\mathbf{x}\|_1\}, \tag{5.1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$. This problem does not possess any smooth component, so that $f \equiv 0$ in the model (G) given in (3.1). Note that the objective function is a sum of two l_1 norms, namely, the componentwise sum of absolute value functions which will be smoothed using Huber's function defined by

$$H_\mu(y) \equiv \begin{cases} \frac{y^2}{2\mu}, & |y| \le \mu, \\ |y| - \frac{\mu}{2} & \text{else.} \end{cases}$$

The function H_μ is a μ-smooth approximation of the absolute value function $|y|$ with parameters $(1, 0.5, 0)$; see Example 4.7. There are (at least) two possible smoothing approaches for this problem within our model (G):

A. Full smoothing. Take $g(\mathbf{x}) \equiv \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1 + \|\mathbf{x}\|_1$ and $h \equiv 0$.

B. Partial smoothing. Take $g(\mathbf{x}) \equiv \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1, h(\mathbf{x}) = \|\mathbf{x}\|_1$.

In the partial smoothing setting, the problem to be solved is

$$\min_{\mathbf{x}} \left\{\sum_{i=1}^{m} H_\mu(\mathbf{A}_i\mathbf{x} - b_i) + \|\mathbf{x}\|_1 \equiv g_\mu(\mathbf{x}) + \|\mathbf{x}\|_1\right\}, \tag{PS$_\mu$}$$

where $\mathbf{A}_i$ is the ith row of the matrix $\mathbf{A}$. Since $\sum_{i=1}^m H_\mu(y_i)$ is a μ-smooth approximation of the l_1 norm function $\|\mathbf{y}\|_1 = \sum_{i=1}^m |y_i|$ with parameters $(1, \frac{m}{2}, 0)$ (see Example 4.2), it follows by Lemma 2.2 that here g_μ is a μ-smooth approximation of g with parameters $(\|\mathbf{A}\|^2, \frac{m}{2}, 0)$. In the full smoothing setting, the smooth problem to be solved is

$$\min_{\mathbf{x}} \left\{\sum_{i=1}^{m} H_\mu(\mathbf{A}_i\mathbf{x} - b_i) + \sum_{j=1}^{n} H_\mu(x_j) \equiv h_\mu(\mathbf{x})\right\}. \tag{FS$_\mu$}$$

By Lemma 2.1 it follows that here h_μ is a μ-smooth approximation of h with parameters $(\|\mathbf{A}\|^2 + 1, \frac{m+n}{2}, 0)$. Note that this is a worse approximation than the one given in the partial smoothing setting since both parameters α, β (corresponding to the proximity between the function and its approximation and the Lipschitz constant of the gradient) are larger.

What is important here is that there is really no need to smooth the l_1 part $\|\mathbf{x}\|_1$, since in that case one can directly invoke a fast proximal gradient method, like FISTA devised in [8], and still achieve the $O(1/k^2)$ rate of convergence. The proximal mapping of the h part is trivial in both settings: in the partial smoothing setting ($h(\mathbf{x}) \equiv \|\mathbf{x}\|_1$) it is equal to the soft thresholding operation, and in the full smoothing setting ($h \equiv 0$) it is simply the identity mapping; see [8] for the detailed algorithms. Note also that it would not be advisable to consider the partial smoothing approach in the opposite way, that is, to smooth the l_1 norm function $\|\mathbf{x}\|_1$ and keep the l_1 fitting term $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1$ untouched. The reason for this is that computing a proximal mapping of the l_1 fitting term seems to be as difficult as solving the original problem.

To compare the two approaches, we performed Monte Carlo runs, where in each run the components of the matrix $\mathbf{A}$ and the vector $\mathbf{b}$ were randomly and independently generated from a standard normal distribution. The dimensions are set to $m = 15, n = 30$. For each realization of $\mathbf{A}$ and $\mathbf{b}$, we ran FISTA for $N = 100$, 200, 400 iterations on both (PS$_\mu$) and (FS$_\mu$), where the parameter μ was chosen according to Theorem 3.1 with $\varepsilon = 0.1$. Note that the choice of μ according to (3.5) is different for the two problems (PS$_\mu$) and (FS$_\mu$). To compare the errors obtained by the two methods, we also found for each realization of $\mathbf{A}$ and $\mathbf{b}$ the optimal value of problem (5.1) using SeDuMi [22]. The results are summarized in the table below. The second and third columns (Err-FS and Err-PS, respectively) contain the average over the 100 realizations of the errors in function values $M(\mathbf{x}_N) - M^*$ (M^* is the optimal value of the problem computed using SeDuMi and $\mathbf{x}_N$ is the output of the corresponding method) when using the full and partial smoothing methodologies, respectively. The fourth column contains the average over the 100 realizations of the ratio of errors $(M(\mathbf{x}_{\text{FS}}) - M^*)/(M(\mathbf{x}_{\text{PS}}) - M^*))$, where $\mathbf{x}_{\text{FS}}$ and $\mathbf{x}_{\text{PS}}$ are the outputs of FISTA in the full and partial smoothing settings, respectively.

N	Err-FS	Err-PS	Err-FS/Err-PS
100	3.2951	1.3722	2.7152
200	1.0009	0.2740	5.0633
400	0.1741	0.0284	22.4585

Clearly, the partial smoothing approach is superior to the full smoothing approach, as it reaches better accuracies for a given number of iterations. Furthermore, the error in function values of the full smoothing setting is more than 22 times the error obtained by the partial smoothing setting when 400 iterations of FISTA are performed.

The above empirical observation has a theoretical justification which is now explained. Suppose that we wish to solve a problem of the form

$$\min_{\mathbf{x}} \{g(\mathbf{x}) + q(\mathbf{x})\},$$

where both g and q are convex and nonsmooth functions and where as usual we assume that the optimal set of this problem is nonempty and bounded. Assume that g_μ is a μ-smooth approximation of g with parameters (α_g, β_g, K_g) and q_μ is a μ-smooth approximation of q with parameters (α_q, β_q, K_q). In the full smoothing approach, the problem to be solved via the fast method is

$$\min_{\mathbf{x}} \{g_\mu(\mathbf{x}) + q_\mu(\mathbf{x})\}, \tag{5.2}$$

while in the partial smoothing approach the relevant problem is

$$\min_{\mathbf{x}}\{g_\mu(\mathbf{x})+q(\mathbf{x})\}. \tag{5.3}$$

Let $(\mathbf{x}^*_{ps},\mathbf{x}^*_{fs})$ be the corresponding optimal solutions of problems (5.3)–(5.2), respectively. Applying a fast iterative method $\mathcal{M}$ on both problems with initial $\mathbf{x}_0=0$ and setting $\Lambda=\max\{\|\mathbf{x}^*_{ps}\|^2,\|\mathbf{x}^*_{fs}\|^2\}$ (cf., for example, (3.13)), it follows that by Theorem 3.1, a lower bound on the number of iterations required to obtain an ε-optimal solution via the fast method $\mathcal{M}$ to problem (5.2) is

$$N_1\equiv 2\sqrt{(\alpha_g+\alpha_q)(\beta_g+\beta_q)\Lambda}\frac{1}{\varepsilon}+\sqrt{(K_g+K_q)\Lambda}\frac{1}{\sqrt{\varepsilon}},$$

while the lower bound on the number of iterations required to obtain an ε-optimal solution of problem (5.3) via the same fast method is given by

$$N_2\equiv 2\sqrt{\alpha_g\beta_g\Lambda}\frac{1}{\varepsilon}+\sqrt{K_g\Lambda}\frac{1}{\sqrt{\varepsilon}}.$$

Obviously, $N_1>N_2$, meaning that at least with respect to the lower bound on the number of iterations, finding an optimal solution of the partially smooth problem (5.3) is easier than finding an optimal solution of the fully smooth problem (5.2).

To conclude from this example, the answer to the question in the title of this section is the following: smoothing is a valuable approach for tackling nonsmooth problems, but it should be used “moderately” and only when truly necessary!

Appendix A. Proof of Theorem 4.1. (a) Let $\mathbf{x}\in\mathbb{E}$ and $\mu>0$. Define $s_1(\mathbf{u})\equiv g(\mathbf{u})$ and $s_2(\mathbf{u})\equiv\mu\omega((\mathbf{x}-\mathbf{u})/\mu)=\omega_\mu(\mathbf{x}-\mathbf{u})$. Then by definition we have

$$g^{\mathrm{ic}}_\mu(\mathbf{x})=\inf_{\mathbf{u}\in\mathbb{E}}\{s_1(\mathbf{u})+s_2(\mathbf{u})\}. \tag{A.1}$$

Since here $\operatorname{dom}\omega=\mathbb{E}$, it follows that $\operatorname{dom}s_2=\operatorname{dom}\omega_\mu=\mathbb{E}$ and hence that $\operatorname{ri}(\operatorname{dom}s_1)\cap\operatorname{ri}(\operatorname{dom}s_2)\neq\emptyset$. Therefore, by Fenchel's duality theorem [19, Theorem 31.1], the expression (A.1) also reads as

$$g^{\mathrm{ic}}_\mu(\mathbf{x})=\max_{\mathbf{y}\in\mathbb{E}^*}\{-s_1^*(\mathbf{y})-s_2^*(-\mathbf{y})\}=\max_{\mathbf{y}\in\mathbb{E}^*}\{-g^*(\mathbf{y})-s_2^*(-\mathbf{y})\}. \tag{A.2}$$

Finally, by the definition of s_2 and ω_μ it follows that $s_2^*(-\mathbf{y})=\omega^*_\mu(\mathbf{y})-\langle\mathbf{y},\mathbf{x}\rangle$ and $\omega^*_\mu(\mathbf{y})=\mu\omega^*(\mathbf{y})$, which after substitution in (A.2) proves formula (4.7).

(b) Since $\nabla\omega$ is Lipschitz with constant $\frac{1}{\sigma}$, and since by definition, $\omega_\mu(\mathbf{x})=\mu\omega(\frac{\mathbf{x}}{\mu})$, it follows that $\nabla\omega_\mu$ is Lipschitz with constant $\frac{1}{\sigma\mu}$. Therefore, by Lemma 4.1, it follows that ω^*_μ, and hence also $g^*+\omega^*_\mu$, is strongly convex with parameter $\sigma\mu$. Invoking again Lemma 4.1, we conclude that $g^{\mathrm{ic}}_\mu=(g^*+\omega^*_\mu)^*$ is differentiable with a Lipschitz gradient with constant $\frac{1}{\sigma\mu}$.

(c) Let $\mathbf{x}\in\mathbb{E}$ be such that there exists a minimizer $\mathbf{u}_\mu(\mathbf{x})$ of (4.5), namely,

$$g^{\mathrm{ic}}_\mu(\mathbf{x})=g(\mathbf{u}_\mu(\mathbf{x}))+\omega_\mu(\mathbf{x}-\mathbf{u}_\mu(\mathbf{x})). \tag{A.3}$$

For convenience, define $\mathbf{z}\equiv\nabla\omega_\mu(\mathbf{x}-\mathbf{u}_\mu(\mathbf{x}))$. Our objective is to show that $\nabla g^{\mathrm{ic}}_\mu(\mathbf{x})=\mathbf{z}$. By standard calculus this means that we have to show that for any $\boldsymbol{\xi}\in\mathbb{E}$,

$\lim_{\|\xi\|\to 0} |\phi(\boldsymbol{\xi})|/\|\boldsymbol{\xi}\| = 0$, where $\phi(\boldsymbol{\xi}) \equiv g_\mu^{\mathrm{ic}}(\mathbf{x}+\boldsymbol{\xi}) - g_\mu^{\mathrm{ic}}(\mathbf{x}) - \langle \boldsymbol{\xi}, \mathbf{z}\rangle$. Using the definition of g_μ^{ic} we obtain

$$g_\mu^{\mathrm{ic}}(\mathbf{x}+\boldsymbol{\xi}) \le g(\mathbf{u}_\mu(\mathbf{x})) + \omega_\mu(\mathbf{x}+\boldsymbol{\xi}-\mathbf{u}_\mu(\mathbf{x})),$$

and combining the latter inequality with (A.3) we get

$$\begin{aligned}
\phi(\boldsymbol{\xi}) &\le \omega_\mu(\mathbf{x}+\boldsymbol{\xi}-\mathbf{u}_\mu(\mathbf{x})) - \omega_\mu(\mathbf{x}-\mathbf{u}_\mu(\mathbf{x})) - \langle \boldsymbol{\xi}, \mathbf{z}\rangle,\\
&\le \langle \boldsymbol{\xi}, \nabla\omega_\mu(\mathbf{x}+\boldsymbol{\xi}-\mathbf{u}_\mu(\mathbf{x}))\rangle - \langle \boldsymbol{\xi}, \mathbf{z}\rangle \quad \text{[by the gradient inequality for } \omega_\mu\text{]},\\
&= \langle \boldsymbol{\xi}, \nabla\omega_\mu(\mathbf{x}+\xi-\mathbf{u}_\mu(\mathbf{x})) - \nabla\omega_\mu(\mathbf{x}-\mathbf{u}_\mu(\mathbf{x}))\rangle \quad \text{[substitution of } \mathbf{z}\text{]},\\
&\le \|\boldsymbol{\xi}\|\|\nabla\omega_\mu(\mathbf{x}+\boldsymbol{\xi}-\mathbf{u}_\mu(\mathbf{x})) - \nabla\omega_\mu(\mathbf{x}-\mathbf{u}_\mu(\mathbf{x}))\|^* \quad \text{[Cauchy–Schwarz inequality]},\\
&\le \frac{1}{\mu\sigma}\|\boldsymbol{\xi}\|^2 \quad \text{[Lipschitz constant of } \nabla\omega_\mu \text{ is } 1/\mu\sigma\text{]}.
\end{aligned}$$

To complete the proof, it remains to show that we also have $\phi(\boldsymbol{\xi}) \ge -\frac{1}{\mu\sigma}\|\boldsymbol{\xi}\|^2$. Since g_μ^{ic} is convex, so is ϕ, which along the fact that $\phi(\mathbf{0}) = 0$ implies that $\phi(\boldsymbol{\xi}) \ge -\phi(-\boldsymbol{\xi})$, and hence the desired result follows.

Appendix B. Proof of Proposition 4.1. Since h_μ is a $(1, \log(m), 0)$-smooth approximation of h over $\mathbb{V}$, then by property (i) of Definition 2.1, it follows that there exists a decomposition $\log(m) = \beta_1 + \beta_2$ for which

$$h(\mathbf{z}) - \beta_1\mu \le h_\mu(\mathbf{z}) \le h(\mathbf{z}) + \beta_2\mu \text{ for every } \mathbf{z} \in \mathbb{V}.$$

Making the change of variables $\mathbf{z} = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))^T$ and within the restriction $\mathbf{x} \in X$, we obtain that

$$g(\mathbf{x}) - \beta_1\mu \le g_\mu(\mathbf{x}) \le g(\mathbf{x}) + \beta_2\mu \text{ for every } \mathbf{x} \in X,$$

thus proving property (i) with $\beta = \log(m)$. To find the other parameters α and K, we introduce some further notation. Let $\mathbf{f} := (f_1, \ldots, f_m)^T$, so that $g(\mathbf{x}) = h(\mathbf{f}(\mathbf{x}))$ and $g_\mu(\mathbf{x}) = h_\mu(\mathbf{f}(\mathbf{x}))$. The matrix $\mathbf{J}_\mathbf{f}(\mathbf{x})$ denotes the transpose of the Jacobian matrix $\mathbf{f}$ given by

$$\mathbf{J}_\mathbf{f}(\mathbf{x}) = (\nabla f_1(\mathbf{x}), \nabla f_2(\mathbf{x}), \ldots, \nabla f_m(\mathbf{x})),$$

and by the chain rule it follows that

$$\nabla g_\mu = \mathbf{J}_\mathbf{f}(\mathbf{x})\nabla h_\mu(\mathbf{f}(\mathbf{x})).$$

Now, for every $\mathbf{x}, \mathbf{y} \in X$ we have

$$\begin{aligned}
&\|\nabla g_\mu(\mathbf{x}) - \nabla g_\mu(\mathbf{y})\|_{\mathbb{E}}^*\\
&\quad = \|\mathbf{J}_\mathbf{f}(\mathbf{x})\nabla h_\mu(\mathbf{f}(\mathbf{x})) - \mathbf{J}_\mathbf{f}(\mathbf{y})\nabla h_\mu(\mathbf{f}(\mathbf{y}))\|_{\mathbb{E}}^*\\
&\quad = \|\mathbf{J}_\mathbf{f}(\mathbf{x})\left(\nabla h_\mu(\mathbf{f}(\mathbf{x})) - \nabla h_\mu(\mathbf{f}(\mathbf{y}))\right) + \left(\mathbf{J}_\mathbf{f}(\mathbf{x}) - \mathbf{J}_\mathbf{f}(\mathbf{y})\right)\nabla h_\mu(\mathbf{f}(\mathbf{y}))\|_{\mathbb{E}}^*\\
&\quad \le \|\mathbf{J}_\mathbf{f}(\mathbf{x})\left(\nabla h_\mu(\mathbf{f}(\mathbf{x})) - \nabla h_\mu(\mathbf{f}(\mathbf{y}))\right)\|_{\mathbb{E}}^* + \|\left(\mathbf{J}_\mathbf{f}(\mathbf{x}) - \mathbf{J}_\mathbf{f}(\mathbf{y})\right)\nabla h_\mu(\mathbf{f}(\mathbf{y}))\|_{\mathbb{E}}^*\\
&\quad \le \|\mathbf{J}_\mathbf{f}(\mathbf{x})\|_{\mathbb{V}^*,\mathbb{E}^*} \cdot \|\nabla h_\mu(\mathbf{f}(\mathbf{x})) - \nabla h_\mu(\mathbf{f}(\mathbf{y}))\|_{\mathbb{V}}^*\\
&\qquad + \|\mathbf{J}_\mathbf{f}(\mathbf{x}) - \mathbf{J}_\mathbf{f}(\mathbf{y})\|_{\mathbb{V}^*,\mathbb{E}^*} \cdot \|\nabla h_\mu(\mathbf{f}(\mathbf{y}))\|_{\mathbb{V}}^*
\end{aligned}$$

$$\text{(B.1)} \quad \le \frac{1}{\mu}\|\mathbf{J}_\mathbf{f}(\mathbf{x})\|_{\mathbb{V}^*,\mathbb{E}^*} \cdot \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\|_{\mathbb{V}} + \|\mathbf{J}_\mathbf{f}(\mathbf{x}) - \mathbf{J}_\mathbf{f}(\mathbf{y})\|_{\mathbb{V}^*,\mathbb{E}^*} \cdot \|\nabla h_\mu(\mathbf{f}(\mathbf{y}))\|_{\mathbb{V}}^*,$$

where the last inequality follows by the fact that ∇h_μ is Lipschitz with constant $\frac{1}{\mu}$. In addition, note that for every $\mathbf{z} \in \mathbb{R}^m$

$$\|\nabla h_\mu(\mathbf{z})\|_{\mathbb{V}}^* = \|\nabla h_\mu(\mathbf{z})\|_1 = \left\| \frac{1}{\sum_{j=1}^m e^{z_j/\mu}} \begin{pmatrix} e^{z_1/\mu} \\ \vdots \\ e^{z_m/\mu} \end{pmatrix} \right\|_1 = 1 \tag{B.2}$$

and that $M_{f_i} = \max\left\{\|\nabla f_i(\mathbf{x})\|_{\mathbb{E}}^* : \mathbf{x} \in X\right\}, i = 1, \ldots, m$. Thus,

$$\begin{aligned} \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\|_{\mathbb{V}} &= \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\|_\infty \\ &= \max_{i=1,\ldots,m} \{|f_i(\mathbf{x}) - f_i(\mathbf{y})|\} \le \left(\max_{i=1,\ldots,m} M_{f_i}\right) \|\mathbf{x} - \mathbf{y}\|_{\mathbb{E}}, \end{aligned} \tag{B.3}$$

$$\|\mathbf{J_f}(\mathbf{x})\|_{\mathbb{V}^*,\mathbb{E}^*} = \max\left\{ \left\| \sum_{i=1}^m v_i \nabla f_i(\mathbf{x}) \right\|_{\mathbb{E}}^* : \sum_{i=1}^m |v_i| \le 1 \right\} \le \max_{i=1,\ldots,m} M_{f_i}, \tag{B.4}$$

$$\|\mathbf{J_f}(\mathbf{x}) - \mathbf{J_f}(\mathbf{y})\|_{\mathbb{V}^*,\mathbb{E}^*} = \max_{i=1,\ldots,m} \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_{\mathbb{E}}^* \le \left(\max_{i=1,\ldots,m} L_{f_i}\right) \|\mathbf{x} - \mathbf{y}\|_{\mathbb{E}}. \tag{B.5}$$

To conclude, plugging (B.2)–(B.5) into (B.1) we get

$$\|\nabla g_\mu(\mathbf{x}) - \nabla g_\mu(\mathbf{y})\|_{\mathbb{E}}^* \le \left[\frac{1}{\mu}\left(\max_{i=1,\ldots,m} M_{f_i}^2\right) + \max_{i=1,\ldots,m} L_{f_i}\right] \|\mathbf{x} - \mathbf{y}\|_{\mathbb{E}},$$

showing that $g_\mu(\cdot) = h_\mu(\mathbf{f}(\cdot))$ is a μ-smooth approximation of $g(\cdot) = h(\mathbf{f}(\cdot))$ over X with parameters

$$\left(\max_{i=1,\ldots,m} M_{f_i}^2, \log(m), \max_{i=1,\ldots,m} L_{f_i}\right).$$

Acknowledgment. We would like to thank the associate editor and anonymous reviewers for their useful comments and additional references which helped to improve the presentation of the paper.

REFERENCES

[1] H. Attouch and R.J.-B. Wets, *Epigraphical analysis*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 6 (1989), pp. 73–100.

[2] A. Auslender, *Penalty and barrier methods: A unified framework*, SIAM J. Optim., 10 (1999), pp. 211–230.

[3] A. Auslender and M. Teboulle, *Asymptotic Cones and Functions in Optimization and Variational Inequalities*, Springer Monogr. Math., Springer, New York, 2003.

[4] A. Auslender and M. Teboulle, *Interior gradient and proximal methods for convex and conic optimization*, SIAM J. Optim., 16 (2006), pp. 697–725.

[5] H. H. Bauschke and P. L. Combettes, *The Baillon-Haddad theorem revisited*, J. Convex Anal., 17 (2010), pp. 781–787.

[6] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer-Verlag, New York, 2011.

[7] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Oper. Res. Lett., 31 (2003), pp. 167–175.

[8] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[9] A. Beck and M. Teboulle, *Gradient-based algorithms with applications to signal recovery problems*, in Convex Optimization in Signal Processing and Communications, D. Palomar and Y. Eldar, eds., Cambridge University Press, Cambridge, UK, 2009, pp. 139–162.

[10] A. Ben-Tal and M. Teboulle, *A smoothing technique for nondifferentiable optimization problems*, in Optimization, Lecture Notes in Math. 1405, S. Dolecki, ed., Springer-Verlag, New York, 1989, pp. 1–11.

[11] D. P. Bertsekas, *Nondifferentiable optimization via approximation*, Math. Program. Stud., 1975, pp. 1–25.

[12] D. P. Bertsekas, *Constrained Optimization and Lagrangian Multipliers*, Academic Press, New York, 1982.

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

[14] O. Devolder, F. Glineur, and Y. Nesterov, *First-Order Methods of Smooth Convex Optimization with Inexact Oracle*, http://www.optimization-online/DB_HTML/2010/12/2865.html.

[15] P. J. Huber, *Robust Statistics*, Wiley Ser. Probab. Math. Stat., John Wiley, New York, 1981.

[16] J. J. Moreau, *Proximité et dualité dans un espace Hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.

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

[18] Y. Nesterov, *Gradient Methods for Minimizing Composite Objective Function*, http://www.ecore.beDPs/dp1191313936.pdf (2007).

[19] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[20] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer-Verlag, Berlin, 1998.

[21] N. Z. Shor, *Minimization Methods for Nondifferentiable Functions*, Springer Ser. Comput. Math. 3, Springer-Verlag, New York, 1985.

[22] J. F. Sturm, *Using SeDuMi* 1.02*, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11–12 (1999), pp. 625–653.

[23] M. Teboulle, *Entropic proximal mappings with application to nonlinear programming*, Math. Oper. Res., 17 (1992), pp. 670–681.