Variable Smoothing for Weakly Convex Composite Functions

Axel Böhm* Stephen J. Wright†

March 17, 2020

We study minimization of a structured objective function, being the sum of a smooth function and a composition of a weakly convex function with a linear operator. Applications include image reconstruction problems with regularizers that introduce less bias than the standard convex regularizers. We develop a variable smoothing algorithm, based on the Moreau envelope with a decreasing sequence of smoothing parameters, and prove a complexity of $\mathcal{O}(\epsilon^{-3})$ to achieve an ϵ-approximate solution. This bound interpolates between the $\mathcal{O}(\epsilon^{-2})$ bound for the smooth case and the $\mathcal{O}(\epsilon^{-4})$ bound for the subgradient method. Our complexity bound is in line with other works that deal with structured nonsmoothness of weakly convex functions.

1 Introduction

We study minimization of the the sum of a smooth function h and a nonsmooth, weakly convex function g composed with a linear operator defined by the matrix $A \in \mathbb{R}^{n\times d}$, that is,

$$\min_{x\in\mathbb{R}^d} F(x) := h(x) + g(Ax). \tag{1}$$

For some $\rho \geq 0$, we say that

$$g : \mathbb{R}^n \to \overline{\mathbb{R}} \text{ is } \rho\textit{-weakly convex if } g + (\rho/2)\|\cdot\|^2 \textit{ is convex.}$$

Weakly convex functions share some properties with convex functions but include many interesting nonconvex cases; see Section 1.1. An example of a weakly convex *smooth*

*Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria. e-mail: axel.boehm@univie.ac.at. Research supported by the doctoral programme *Vienna Graduate School on Computational Optimization (VGSCO)*, FWF (Austrian Science Fund), project W 1260.

†Computer Sciences Department and Wisconsin Institute for Discovery, University of Wisconsin-Madison. e-mail: swright@cs.wisc.edu. Research supported by NSF Awards 1628384, 1634597, and 1740707; Subcontract 8F-30039 from Argonne National Laboratory; and Award N660011824020 from the DARPA Lagrange Program.

function is one in which the gradient is uniformly Lipschitz continuous. In this case, the weak convexity parameter ρ is the reciprocal of the Lipschitz constant.

Our approach makes use of a smooth approximation of g known as the *Moreau envelope* ${}^{\mu}g$, parametrized by a positive scalar μ, together with gradient descent. The Moreau envelope and proximal operator are defined as follows.

Definition 1.1. *For a proper, ρ-weakly convex and lower semicontinuous function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$, the Moreau envelope of g with the parameter $\mu \in (0, \rho^{-1})$ is the function from $\mathbb{R}^n$ to $\mathbb{R}$ defined by*

$$ {}^{\mu}g(y) := \inf_{z\in\mathbb{R}^n} \left\{ g(z) + \frac{1}{2\mu}\|z-y\|^2 \right\}. $$

The proximal operator of the function μg is the $\arg\min$ *of the right-hand side in this definition, that is,*

$$ \operatorname{prox}_{\mu g}(y) := \arg\min_{z\in\mathbb{R}^n} \left\{ g(z) + \frac{1}{2\mu}\|z-y\|^2 \right\}. \tag{2} $$

Note that $\operatorname{prox}_{\mu g}(y)$ is uniquely defined by (2) because the function being minimized is strongly convex.

Steps of the algorithm have the form

$$ x \leftarrow x - \gamma\nabla(h + {}^{\mu}g \circ A)(x), $$

for some steplength γ. Accelerated versions of these approaches have been proposed for convex problems in [4, 23, 7]. The use of acceleration makes the analysis more complicated than for the gradient case; see [3, 8].

1.1 Composite Problems

We discuss several instances of problems of the form (1).

Convex. The case of problems (1) in which g is nonsmooth and *convex* (with possible smooth and/or nonsmooth additive terms) has received a great deal of attention in the literature on convex optimization and applications; see for example [10, 24, 5, 6, 23]. The most notable applications are found in inverse problems involving images. In particular, (an)isotropic *Total Variation (TV)* denoising has the form

$$ \min_x \frac{1}{2}\|x-b\|^2 + \|\nabla x\|_1, \tag{3} $$

where b is the observed (noisy) image and ∇ denotes the discretized gradient in two or three dimensions. TV deblurring problems have the form

$$ \min_x \frac{1}{2}\|Ax-b\|^2 + \|\nabla x\|_1, \tag{4} $$

where A is the blurring operator; see [10, 9].

Other examples of convex problems of the form (1) include generalized convex feasibility [23] and support vector machine classification [7]. A typical formulation of the latter problem has $h(x) = (\lambda/2)\|x\|^2$ and $g(Ax) = \sum_{i=1}^n t(y_i a_i^T x)$, where $t(s) = \max(-s, 0)$ is the hinge loss and the rows of A are $y_i a_i^T$, $i = 1, 2, \ldots, n$, where $(y_i, a_i) \in \{-1, 1\} \times \mathbb{R}^d$ are the training points and their labels.

Weakly Convex Regularizers. Functions that are "sharp" around zero have a long history as sparsity-inducing regularizers. Foremost among such functions is the ℓ_1 norm $\|\cdot\|_1$, which is used for example in sparse least-squares regression (also known as LASSO):

$$\min_x \frac{1}{2}\|Ax - b\|^2 + \|x\|_1$$

or the anisotropic Total Variation denoising or deblurring problems (3) and (4). However, the use of the ℓ_1 regularizer tends to depress the magnitude of nonzero elements of the solution, resulting in *bias.* This phenomenon is a consequence of the fact that the proximal operator of the 1-norm, often called the *soft thresholding operator*, does not approach the identity for larger values of its argument. For this reason, nonconvex alternatives to $\|\cdot\|_1$ are often used to reduce bias. These include ℓ_p-norms (with $0 < p < 1$) which are not weakly convex, and the several weakly convex regularizers, which we now describe. The *minimax concave penalty (MCP)*, introduced in [25] and used in [21, 17], is a family of functions $r_{\lambda,\theta} : \mathbb{R} \to \mathbb{R}_+$ involving two positive parameters λ and θ, and defined by

$$r_{\lambda,\theta}(x) := \begin{cases} \lambda|x| - \frac{x^2}{2\theta}, & |x| \leq \theta\lambda, \\ \frac{\theta\lambda^2}{2}, & \text{otherwise.} \end{cases}$$

(Note that this function satisfies the definition of ρ-weak convexity with $\rho = \theta^{-1}$.) The proximal operator of this function (called *firm threshold* in [1]) can be written in the following closed form when $\theta > \gamma$:

$$\text{prox}_{\gamma r_{\lambda,\theta}}(x) = \begin{cases} 0, & |x| < \gamma\lambda, \\ \frac{x - \lambda\gamma\,\text{sgn}(x)}{1 - (\gamma/\theta)}, & \gamma\lambda \leq |x| \leq \theta\lambda, \\ x, & |x| > \theta\lambda. \end{cases}$$

The *fractional penalty function* (cf. [20, 17]) $\phi_a : \mathbb{R} \to \mathbb{R}_+$ (for parameter $a > 0$) is

$$\phi_a(x) := \frac{|x|}{1 + a|x|/2}.$$

The *smoothly clipped absolute deviation (SCAD)* [14] (cf. [17]) is defined for parameters $\lambda > 0$ and $\theta > 2$ as follows:

$$r_{\lambda,\theta}(x) = \begin{cases} \lambda|x|, & |x| \leq \lambda, \\ \frac{-x^2 + 2\theta\lambda|x| - \lambda^2}{2(\theta - 1)}, & \lambda < |t| \leq \theta\lambda, \\ \frac{(\theta+1)\lambda^2}{2}, & |t| > \theta\lambda. \end{cases}$$

(This function is $(\theta-1)^{-1}$-weakly convex.)

Since these functions approach (or attain) a finite value as their argument grows in magnitude, they do not introduce as much bias in the solution as does the ℓ_1 norm, and their proximal operators approach the identity for large arguments.

These regularizers have, however, mostly been used in the simple additive setting

$$\min_{x\in\mathbb{R}^d} h(x)+g(x)$$

for a smooth data fidelity term h and nonsmooth regularizer g, for example in least squares or logistic regression [21] and compressed sensing (cf. [1]).

Weakly Convex Composite Losses. The use of weakly convex functions composed with linear operators has been explored in the robust statistics literature. An early instance is the *Tukey biweight* function [2], in which $g(Ax)$ has the form

$$g(Ax)=\sum_{i=1}^{n}\phi(A_{i\cdot}x-b_i), \quad \text{where } \phi(\theta)=\frac{\theta^2}{1+\theta^2}. \tag{5}$$

This function behaves like the usual least-squares loss when $\theta^2 \ll 1$ but asymptotes at 1. It is ρ-weakly convex with $\rho=6$.

A different (but similar) definition of the Tukey biweight function appears in [18, Section 2.1]. This same reference also mentions another nonconvex loss function, the *Cauchy loss*, which has the form (5) except that ϕ is defined by

$$\phi(\theta)=\frac{\xi^2}{2}\log\left(1+\frac{\theta^2}{\xi^2}\right),$$

for some parameter ξ. This function is ρ-weakly convex with $\rho=6$.

1.2 Complexity Bounds for Weakly Convex Problems

To put our results in perspective, we provide a review of the literature on complexity bounds for optimization problems related to our formulation (1), including weakly convex functions. In all cases, these are bounds on the number of iterations required to find an approximately stationary point, where our measure of stationarity is based the norm of the gradient of the Moreau envelope (a smooth proxy).

The best known complexity for black box subgradient optimization for weakly convex functions is $\mathcal{O}(\epsilon^{-4})$. This result is proved for *stochastic* subgradient in [11], but as in the convex case, there is no known improvement in the deterministic setting. As in convex optimization, subgradient methods are quite general and implementable for weakly convex functions. However, when more structure is present in the function, algorithms that achieve better complexity can be devised. In particular, when the proximal operator of the nonsmooth weakly convex function can be calculated analytically, complexity bounds of $\mathcal{O}(\epsilon^{-2})$ can be proven (see Section 4), the same bounds as for steepest descent methods in the smooth nonconvex case. This means that the entire difficulty introduced by the nonsmoothness can be mitigated as long as it is treated by a proximal operator.

For convex optimization problems, bounds of $\mathcal{O}(\epsilon^{-1})$ can be obtained for gradient methods on smooth functions and $\mathcal{O}(\epsilon^{-1/2})$ for accelerated gradient methods. These same bounds can also be obtained for nonsmooth problems provided that the nonsmooth function is handled by a proximal operator. When the explicit proximal operator is not available and subgradient methods have to be used, the complexity reverts to $\mathcal{O}(\epsilon^{-2})$.

It is possible to keep the $\mathcal{O}(\epsilon^{-2})$ rate when just a local model of the weakly convex part is evaluated by a convex operator. The paper [13] studies optimization problems of the type

$$\min_x h(x) + g(c(x))$$

where h is convex, proper, and closed; g is convex and Lipschitz continuous; and c is smooth. (Under these assumptions, the composition $g \circ c$ is weakly convex.) The $\mathcal{O}(\epsilon^{-2})$ bound is proved for an algorithm in which the (convex) subproblem

$$\min_y h(y) + g(c(x) + \nabla c(x)(y - x)) + \frac{1}{2t}\|y - x\|^2 \tag{6}$$

is solved explicitly. In the more realistic case in which (6) must be solved by an iterative procedure, a bound of $\widetilde{\mathcal{O}}(\epsilon^{-3})$ is obtained in [13]. (The symbol $\widetilde{\mathcal{O}}$ hides logarithmic terms.)

Functions of the form $g(c(x))$ have also been studied in [16] for the case of a smooth nonlinear vector function c and a prox-regular g. This is more general than the formulations consider in this paper, both in the fact that all weakly convex functions are prox-regular, and in the nonlinearity of the inner map. The subproblems in [16] have a form similar to (6), and while convergence results are proved in the latter paper, it does not contain rate-of-convergence results or complexity results.

A different weakly convex structure is explored in [22], in which the weak convexity stems from a smooth saddle point problem. This paper studies the problem

$$\min_x \max_{y \in Y} l(x, y),$$

for a compact set $Y \subset \mathbb{R}^m$, where $l(x, \cdot)$ is concave, $l(\cdot, y)$ is nonconvex, and $l(\cdot, \cdot)$ is smooth. They prove a bound of $\widetilde{\mathcal{O}}(\epsilon^{-3})$ for a method that uses only gradient evaluations.

In light of the considerations above, the complexity bound of $\mathcal{O}(\epsilon^{-3})$ for our algorithm seems almost inevitable. It interpolates between the setting without structural assumptions about the nonsmoothness (black box subgradient) and the perfect structural knowledge of the nonsmoothness (explicit knowledge of the proximal operator).

In Section 4, we treat the simpler setting in which the linear operator from (1) is the identity, so that $F(x) = h(x) + g(x)$. Similar problems have been analyzed before, for example, in [21, 1]. However, it is assumed in [1] that convexity in the data fidelity term h compensates for nonconvexity in the regularizer g such that the overall objective function F remains convex. (We make no such assumption here.) The paper [21] does not make such restrictive assumptions and proves convergence but not complexity bounds.

2 Preliminaries

The concept of subgradient of a convex function can be adapted to weakly convex functions via the following definition.

Definition 2.1 (Fréchet subdifferential). *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a function and $\bar{y}$ a point such that $g(\bar{y})$ is finite. Then, the* Fréchet subdifferential *of g at $\bar{y}$, denoted by $\partial g(\bar{y})$, is the set of all vectors $v \in \mathbb{R}^n$ such that*

$$g(y) \geq g(\bar{y}) + \langle v, y - \bar{y} \rangle + o(\|y - \bar{y}\|) \quad \textit{as } y \to \bar{y}. \tag{7}$$

Modifying the convex case, in which subgradients are the slopes of linear functions that underestimate g but coincide with it at $\bar{y}$, Fréchet subgradients do so *up to first order*. This definition makes sense for arbitrary functions, but for lower semicontinuous ρ-weakly convex functions, more can be said. For example, for this class of function we know that subgradients satisfy the following stronger version of (7), for all $v \in \partial g(\bar{y})$,

$$g(y) \geq g(\bar{y}) + \langle v, y - \bar{y} \rangle - \frac{\rho}{2}\|y - \bar{y}\|^2, \quad \forall y \in \mathbb{R}^n.$$

Further, if we assume the weakly convex function to be continuous at a point y, then its subdifferential is nonempty at y. Both of these claims can be verified directly by adding $\frac{\rho}{2}\|\cdot\|^2$ to g and considering the convex subdifferential; see [12, Lemma 2.1].

Another nice property of weakly convex functions is that the definition of a Moreau envelope (see Definition 1.1) extends without modification to weakly convex functions, subject only to a restriction on the parameter μ. The proximal operator (2) also extends to this setting, and this operator and the Moreau envelope fulfil the same identity as in the convex setting.

Lemma 2.1 ([15, Corollary 3.4]). *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, ρ-weakly convex, and lower semicontinuous function, and let $\mu \in (0, \rho^{-1})$. Then the Moreau envelope ${}^{\mu}g(\cdot)$ is continuously differentiable on $\mathbb{R}^n$ with gradient*

$$\nabla({}^{\mu}g)(y) = \frac{1}{\mu}\left(y - \operatorname{prox}_{\mu g}(y)\right), \quad \textit{for all } y \in \mathbb{R}^n.$$

This gradient is $\max\left\{\mu^{-1}, \frac{\rho}{1-\rho\mu}\right\}$*-Lipschitz continuous. In particular, a gradient step with respect to the Moreau envelope corresponds to a proximal step, that is,*

$$y - \mu\nabla({}^{\mu}g)(y) = \operatorname{prox}_{\mu g}(y), \quad \textit{for all } y \in \mathbb{R}^n. \tag{8}$$

Lemma 2.1 not only clarifies the smoothness of the Moreau envelope, but also gives a way of computing its gradient via the prox operator. Obviously, a smooth representation whose gradient could not be computed would be of only limited usefulness from an algorithmic standpoint. The only difference between the weakly convex and convex settings is that the Moreau envelope need not be convex in the former case.

2.1 Stationarity

We say that a point $\bar{x}$ is a stationay point for a function if the Fréchet subdifferential of the function contains 0 at $\bar{x}$. The concept of *nearly stationary* is not quite so straightforward. We motivate our approach by looking first at the simple additive composite problem, also discussed in Section 4, which corresponds to setting $A = I$ in (1), that is,

$$\min_{x} h(x) + g(x). \tag{9}$$

Stationarity for (9) means that $0 \in \partial(h + g)(\bar{x})$, that is, $-\nabla h(\bar{x}) \in \partial g(\bar{x})$. A natural definition for ϵ-approximate stationarity would thus be

$$\operatorname{dist}(-\nabla h(x), \partial g(x)) \leq \epsilon. \tag{10}$$

However, since we are running gradient descent on the *smoothed* problem, our algorithm will naturally compute and detect points with that satisfy a threshold condition of the form

$$\|\nabla h(x) + \nabla^{\mu} g(x)\| \leq \epsilon. \tag{11}$$

The next lemma helps to clarify relationship between these two conditions.

Lemma 2.2. *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, ρ-weakly convex, and lower semicontinuous function, and let $\mu \in (0, \rho^{-1})$. Then*

$$\nabla^{\mu} g(x) \in \partial g(\operatorname{prox}_{\mu g}(x)). \tag{12}$$

Proof. From Definition 1.1, we have that

$$0 \in \partial g(\operatorname{prox}_{\mu g}(x)) + \frac{1}{\mu}(\operatorname{prox}_{\mu g}(x) - x),$$

from which the result follows when we use (8). □

(This result is proved for the case of g convex in [13, Lemma 2.1], with essentially the same proof.)

This lemma tells us that when (11) holds, then (10) is nearly satisfied, except that in the argument of ∂g, x is replaced by $\operatorname{prox}_{\mu g}(x)$. In general, however, $\operatorname{prox}_{\mu g}(x)$ might be arbitrarily far away from x. We can remedy this issue by requiring g to be Lipschitz continuous also.

Lemma 2.3. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a ρ-weakly convex function that is L_g-Lipschitz continuous, and let $\mu \in (0, \rho^{-1})$. Then the Moreau envelope ${}^{\mu}g$ is Lipschitz continuous with*

$$\|\nabla^{\mu} g(x)\| \leq L_g \tag{13}$$

and

$$\|x - \operatorname{prox}_{\mu g}(x)\| \leq \mu L_g, \quad \forall x \in \mathbb{R}^n. \tag{14}$$

Proof. Lipschitz continuity is equivalent to bounded subgradients [19], so by (12), we have for all $x \in \mathbb{R}^n$

$$\|\nabla^\mu g(x)\| \leq \sup \left\{ \|v\| : \, v \in \partial g(\operatorname{prox}_{\mu g}(x)) \right\} \leq L_g,$$

proving (13). The bound (14) follows immediately when we use $x - \operatorname{prox}_{\mu g}(x) = \mu \nabla(^\mu g)(x)$ from Lemma 2.1. $\square$

When $x \in \mathbb{R}^n$ satisfies (11), ∇h is $L_{\nabla h}$-Lipschitz continuous, g is L_g Lipschitz continuous, we have

$$\begin{aligned}
&\operatorname{dist}(-\nabla h(\operatorname{prox}_{\mu g}(x)), \partial g(\operatorname{prox}_{\mu g}(x))) \\
&\quad \leq \|\nabla h(\operatorname{prox}_{\mu g}(x)) - \nabla h(x)\| + \operatorname{dist}(-\nabla h(x), \partial g(\operatorname{prox}_{\mu g}(x))) \\
&\quad \leq L_{\nabla h} \|x - \operatorname{prox}_{\mu g}(x)\| + \epsilon && \text{(from (11) and (12))} \\
&\quad \leq L_{\nabla h} L_g \mu + \epsilon && \text{(from (14)).}
\end{aligned}$$

Thus, if μ is sufficiently small and x satisfies (11), then $\operatorname{prox}_{\mu g}(x)$ is near-stationary for (9).

2.2 Stationarity for the Composite Problem

It follows immediately from (12) in Lemma 2.2 that for $\mu \in (0, \rho^{-1})$, we have for all $x \in \mathbb{R}^d$

$$\nabla(^\mu g \circ A)(x) = A^* \nabla^\mu g(Ax) \in A^* \partial g(\operatorname{prox}_{\mu g}(Ax)). \tag{15}$$

Extending the results of the previous section to the case of a general linear operator A in (1) requires some work. Stationarity for (1) requires that $0 \in \nabla h(x) + A^* \partial g(Ax)$, so ϵ-near stationarity requires

$$\operatorname{dist}(-\nabla h(x), A^* \partial g(Ax)) \leq \epsilon. \tag{16}$$

Our method can compute a point x such that

$$\|\nabla h(x) + \nabla(^\mu g \circ A)(x)\| \leq \epsilon$$

which by (15) implies that

$$\operatorname{dist}(-\nabla h(x), A^* \partial g(z)) \leq \epsilon, \quad \text{where } \; z = \operatorname{prox}_{\mu g}(Ax), \tag{17}$$

where, provided that g is L_g-Lipschitz continuous, we have

$$\|Ax - z\| \leq L_g \mu. \tag{18}$$

The bound in (17) measures the criticality, while the bound in (18) concerns feasibility. The bounds (17), (18) are not a perfect match with (16), since the subdifferentials of h and $g \circ A$ are evaluated at different points.

Surjectivity of A. When A is surjective, we can perturb the x that satisfies (17), (18) to a nearby point x^* that satisfies a bound of the form (16). Since $z = \operatorname{prox}_{\mu g}(Ax)$ is in the range of A, we can define

$$x^* := \underset{x' \in \mathbb{R}^d}{\arg\min} \{\|x - x'\|^2 : \; Ax' = z\}, \tag{19}$$

which is given explicitly by

$$x^* = x - A^*(AA^*)^{-1}(Ax - z) = x - A^\dagger(Ax - z)$$

where $A^\dagger := A^*(AA^*)^{-1}$ is the pseudoinverse of A. The operator norm of the pseudoinverse is bounded by the inverse of the smallest singular value $\sigma_{\min}(A)$ of A, so when g is L_g-Lipschitz continuous, we have from (18) that

$$\|x - x^*\| \le \sigma_{\min}(A)^{-1}\|Ax - z\| \le \sigma_{\min}(A)^{-1} L_g \mu. \tag{20}$$

The point x^* is approximately stationary in the sense of (16), for μ sufficiently small, because

$$\begin{aligned}
&\operatorname{dist}(-\nabla h(x^*), A^*\partial g(Ax^*)) \\
&\quad \le \|\nabla h(x^*) - \nabla h(x)\| + \operatorname{dist}(-\nabla h(x), A^*\partial g(z)) && (\text{since } Ax^* = z = \operatorname{prox}_{\mu g}(Ax)) \\
&\quad \le L_{\nabla h}\|x - x^*\| + \epsilon && (\text{from (17)}) \\
&\quad \le L_{\nabla h}\sigma_{\min}(A)^{-1} L_g \mu + \epsilon && (\text{from (20)}).
\end{aligned} \tag{21}$$

By choosing μ small, x^* will be an approximate solution in the stronger sense (16) and not just the weaker notion of (17), (18), which we have to settle for if A is not surjective.

3 Variable Smoothing

We describe our variable smoothing approaches for the problem (1), where we assume that h is $L_{\nabla h}$-smooth; g is possibly nonsmooth, ρ-weakly convex, and L_g-Lipschitz continuous; and A is a nonzero linear continuous operator. For convenience, we define the smoothed approximation $F_k : \mathbb{R}^d \to \mathbb{R}$ based on the Moreau envelope with parameter μ_k as follows:

$$F_k(x) := h(x) + {}^{\mu_k}g(Ax).$$

We note from Lemma 2.1 and the chain rule that

$$\nabla F_k(x) = \nabla h(x) + \frac{1}{\mu_k} A^*(Ax - \operatorname{prox}_{\mu_k g}(Ax)). \tag{22}$$

The quantity L_k defined by

$$L_k := L_{\nabla h} + \|A\|^2 \max\left\{\mu_k^{-1}, \frac{\rho}{1 - \rho\mu_k}\right\} \tag{23}$$

is a Lipschitz constant of the gradient of ∇F_k, see Lemma 2.1. When $\rho\mu_k \le 1/2$, the maximum in (23) is achieved by μ_k^{-1}, so in this case we can define

$$L_k := L_{\nabla h} + \|A\|^2/\mu_k. \tag{24}$$

3.1 An Elementary Approach

Our first algorithm takes gradient descent steps on the smoothed problem, that is,

$$x_{k+1} = x_k - \gamma_k \nabla F_k(x_k), \tag{25}$$

for certain values of the parameter μ_k and step size γ_k. From (22), the formula (25) is equivalent to

$$x_{k+1} = x_k - \frac{\gamma_k}{\mu_k} A^*(Ax_k - \text{prox}_{\mu_k g}(Ax_k)) - \gamma_k \nabla h(x_k).$$

Our basic algorithm is described next.

Algorithm 1 Variable Smoothing

Require: $x_1 \in \mathbb{R}^d$;
  for $k = 1, 2, 3, \ldots$ **do**
    Set $\mu_k \leftarrow (2\rho)^{-1} k^{-1/3}$, define L_k as in (24), set $\gamma_k \leftarrow 1/L_k$;
    Set $x_{k+1} \leftarrow x_k - \gamma_k \nabla F_k(x_k)$;
  end for

We now state the convergence result for Algorithm 1. This result and later results make use of a quantity

$$F^* := \liminf_{k\to\infty} F_k(x_k), \tag{26}$$

which is finite if F is bounded below (and possibly in other circumstances too). When $F^* = -\infty$, the claim of the theorem is vacuously true.

Theorem 3.1. *Suppose that Algorithm 1 is applied to the problem* (1)*, where g is ρ-weakly convex and ∇h and g are Lipschitz continuous with constants $L_{\nabla h}$ and L_g, respectively. We have*

$$\begin{aligned}\min_{1\le j\le k} \text{dist}(-\nabla h(x_j), & A^*\partial g(\text{prox}_{\mu_j g}(Ax_j))) \\ & \le k^{-1/3}\sqrt{L_{\nabla h} + 2\rho\|A\|^2}\sqrt{F_1(x_1) - F^* + (2\rho)^{-1}L_g^2},\end{aligned}$$

where

$$\|Ax_j - \text{prox}_{\mu_j g}(Ax_j)\| \le j^{-1/3}(2\rho)^{-1}L_g,$$

and F^* is defined as in (26)*. If A is also surjective, then for $x_k^* := x_k - A^\dagger(Ax_k - \text{prox}_{\mu_k g}(Ax_k))$, we have*

$$\begin{aligned}\min_{1\le j\le k} & \text{dist}(-\nabla h(x_j^*), A^*\partial g(Ax_j^*)) \\ & \le k^{-1/3}\left(\sqrt{L_{\nabla h} + (2\rho)\|A\|^2}\sqrt{F_1(x_1) - F^* + (2\rho)^{-1}L_g^2} + L_{\nabla h}\sigma_{\min}(A)^{-1}L_g\right)\end{aligned}$$

and $\|x_j - x_j^*\| \le \sigma_{\min}(A)^{-1}L_g\mu_j = \sigma_{\min}(A)^{-1}L_g(2\rho)^{-1}j^{-1/3}$.

Before proving this theorem, we state and prove a lemma that relates the function values of two Moreau envelopes with two different smoothing parameters. In the convex case, such statements are well known, but in the nonconvex case this result is novel.

Lemma 3.1. *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, closed, and ρ-weakly convex function, and let μ_2 and μ_1 be parameters such that $0 < \mu_2 \leq \mu_1 < \rho^{-1}$. Then, we have*

$$ {}^{\mu_2}g(y) \leq {}^{\mu_1}g(y) + \frac{1}{2}\frac{\mu_1 - \mu_2}{\mu_2}\mu_1 \|\nabla^{\mu_1} g(y)\|^2. $$

If, in addition, g is L_g-Lipschitz continuous, we have

$$ {}^{\mu_2}g(y) \leq {}^{\mu_1}g(y) + \frac{1}{2}\frac{\mu_1 - \mu_2}{\mu_2}\mu_1 L_g^2. $$

Proof. By using the definition of the Moreau envelope, together with Lemma 2.1, we obtain

$$ \begin{aligned} {}^{\mu_2}g(y) &= \min_{u \in \mathbb{R}^n} \left\{ g(u) + \frac{1}{2\mu_2}\|y - u\|^2 \right\} \\ &= \min_{u \in \mathbb{R}^n} \left\{ g(u) + \frac{1}{2\mu_1}\|y - u\|^2 + \frac{1}{2}\left(\frac{1}{\mu_2} - \frac{1}{\mu_1}\right)\|y - u\|^2 \right\} \\ &\leq g(\operatorname{prox}_{\mu_1 g}(y)) + \frac{1}{2\mu_1}\|y - \operatorname{prox}_{\mu_1 g}(y)\|^2 + \frac{1}{2}\left(\frac{1}{\mu_2} - \frac{1}{\mu_1}\right)\|y - \operatorname{prox}_{\mu_1 g}(y)\|^2 \\ &= {}^{\mu_1}g(y) + \frac{1}{2}\left(\frac{\mu_1 - \mu_2}{\mu_2}\right)\mu_1 \|\nabla^{\mu_1} g(y)\|^2, \end{aligned} $$

proving the first claim. The second claim follows immediately from (13). □

Proof of Theorem 3.1. Since $L_k = 1/\gamma_k$ is the Lipschitz constant of ∇F_k, we have for any $k = 1, 2, \ldots$ that

$$ F_k(x_{k+1}) \leq F_k(x_k) + \langle \nabla F_k(x_k), x_{k+1} - x_k \rangle + \frac{1}{2\gamma_k}\|x_{k+1} - x_k\|^2. $$

By substituting the definition of x_{k+1} from (25), we have

$$ F_k(x_{k+1}) \leq F_k(x_k) - \frac{\gamma_k}{2}\|\nabla F_k(x_k)\|^2. \tag{27} $$

From Lemma 3.1, we have for all $x \in \mathbb{R}^d$

$$ F_{k+1}(x) \leq F_k(x) + \frac{1}{2}(\mu_k - \mu_{k+1})\frac{\mu_k}{\mu_{k+1}}\|(\nabla^{\mu_k} g)(Ax)\|^2 \leq F_k(x) + (\mu_k - \mu_{k+1})L_g^2, $$

where we used in the second inequality that $\frac{\mu_k}{\mu_{k+1}} \leq 2$. We set $x = x_{k+1}$ and substitute into (27) to obtain

$$ F_{k+1}(x_{k+1}) \leq F_k(x_k) - \frac{\gamma_k}{2}\|\nabla F_k(x_k)\|^2 + (\mu_k - \mu_{k+1})L_g^2. $$

By summing both sides of this expression over $k = 1, 2, \ldots, K$, and telescoping, we deduce

$$\sum_{k=1}^{K} \frac{\gamma_k}{2} \|\nabla F_k(x_k)\|^2 \le F_1(x_1) - F_K(x_K) + (\mu_1 - \mu_K) L_g^2 \le F_1(x_1) - F^* + \mu_1 L_g^2. \quad (28)$$

Since

$$\gamma_k = \frac{\mu_k}{\mu_k L_{\nabla h} + \|A\|^2} \ge k^{-1/3} \frac{(2\rho)^{-1}}{(2\rho)^{-1} L_{\nabla h} + \|A\|^2} = k^{-1/3} \frac{1}{L_{\nabla h} + 2\rho\|A\|^2}.$$

we have from (28) that

$$\frac{1}{L_{\nabla h} + 2\rho\|A\|^2} \min_{1\le j\le K} \|\nabla F_j(x_j)\|^2 \frac{1}{2} \sum_{k=1}^{K} k^{-1/3} \le F_1(x_1) - F^* + (2\rho)^{-1} L_g^2. \quad (29)$$

Now we observe that

$$\begin{aligned}\sum_{k=1}^{K} k^{-1/3} &\ge \sum_{k=1}^{K} \int_k^{k+1} x^{-1/3}\, \mathrm{d}x = \int_1^{K+1} x^{-1/3}\, \mathrm{d}x = \frac{3}{2}\left((K+1)^{2/3} - 1\right) \\ &\ge (K+1)^{2/3} - 1 \ge \frac{1}{2} K^{2/3}, \quad K = 1, 2, \ldots,\end{aligned}$$

where the final inequality can be checked numerically. Therefore, by substituting into (29), we have

$$\min_{1\le j\le K} \|\nabla F_j(x_j)\|^2 \le 4 \frac{L_{\nabla h} + (2\rho)\|A\|^2}{K^{2/3}} \left(F_1(x_1) - F^* + (2\rho)^{-1} L_g^2\right)$$

and so

$$\min_{1\le j\le K} \|\nabla F_j(x_j)\| \le \frac{C}{K^{1/3}},$$

where $C := 2\sqrt{L_{\nabla h} + (2\rho)\|A\|^2}\sqrt{F_1(x_1) - F^* + (2\rho)^{-1} L_g^2}$. By combining this bound with (17), and defining $z_j := \mathrm{prox}_{\mu_j g}(Ax_j)$, we obtain

$$\min_{1\le j\le k} \mathrm{dist}(-\nabla h(x_j), A^* \partial g(z_j)) \le \min_{1\le j\le k} \|\nabla F_j(x_j)\| \le \frac{C}{k^{1/3}}, \quad (30)$$

where we deduce from (14) that

$$\|Ax_j - z_j\| \le \frac{(2\rho)^{-1} L_g}{j^{1/3}}, \quad \text{for all } j \ge 1.$$

The second statement concerning surjectivity of A follows from the consideration made in (19) to (21). □

There is a mismatch between the two bounds in this theorem. The first bound (the criticality bound) indicates that during the first $k = O(\epsilon^{-3})$ iterations, we will encounter an iteration j at which the first-order optimality condition is satisfied to within a tolerance of ϵ. However, this bound could have been satisfied at an early iteration (that is, $j \ll \epsilon^{-3}$), for which value the second (feasiblity) bound, on $\|Ax_j - \mathrm{prox}_{\mu_j g}(Ax_j)\|$, may not be particularly small. The next section describes an algorithm that remedies this defect.

3.2 An Epoch-Wise Approach with Improved Convergence Guarantees

We describe a variant of Algorithm 1 in which the steps are organized into a series of epochs, each of which is twice as long as the one before. We show that there is some iteration $j = O(\epsilon^{-3})$ such that both $\mathrm{dist}(-\nabla h(x_j), A^*\partial g(\mathrm{prox}_{\mu_j g}(Ax_j)))$ and $\|Ax_j - \mathrm{prox}_{\mu_j g}(Ax_j)\|$ are smaller than the given tolerance ϵ.

Algorithm 2 Variable Smoothing with Epochs

Require: $x_1 \in \mathbb{R}^d$ and tolerance $\epsilon > 0$;
  for $l = 0, 1, \ldots$ **do**
    Set $S_l \leftarrow \infty$, Set $j_l \leftarrow 2^l$;
    for $k = 2^l, 2^l + 1, \ldots, 2^{l+1} - 1$ **do**
      Set $\mu_k \leftarrow (2\rho)^{-1} k^{-1/3}$, define L_k as in (24), set $\gamma_k \leftarrow 1/L_k$;
      Set $x_{k+1} \leftarrow x_k - \gamma_k \nabla F_k(x_k)$;
      if $\|\nabla F_{k+1}(x_{k+1})\| \leq S_l$ **then**
        Set $S_l \leftarrow \|\nabla F_{k+1}(x_{k+1})\|$; Set $j_l \leftarrow k + 1$;
        if $S_l \leq \epsilon$ and $\|Ax_{k+1} - \mathrm{prox}_{\mu_{k+1} g}(Ax_{k+1})\| \leq \epsilon$ **then**
          STOP;
        end if
      end if
    end for
  end for

Theorem 3.2. *Consider solving* (1) *using Algorithm 2, where h and g satisfy the assumptions of Theorem 3.1 and F^* defined in* (26) *is finite. For a given tolerance $\epsilon > 0$, Algorithm 2 generates an iterate x_j for some $j = O(\epsilon^{-3})$ such that*

$$\mathrm{dist}(-\nabla h(x_j), A^*\partial g(z_j)) \leq \epsilon \quad \textit{and} \quad \|Ax_j - z_j\| \leq \epsilon, \quad \textit{where } z_j = \mathrm{prox}_{\mu_j g}(Ax_j).$$

Proof. As in (28), by using monotonicity of $\{F_k(x_k)\}$ and discarding nonnegative terms, we have that

$$\sum_{k=2^l}^{2^{l+1}-1} \frac{\gamma_k}{2} \|\nabla F_k(x_k)\|^2 \leq F_1(x_1) - F^* + (2\rho)^{-1} L_g^2.$$

With the same arguments as in the earlier proof, we obtain

$$\sum_{k=2^l}^{2^{l+1}-1} k^{-1/3} \geq \sum_{k=2^l}^{2^{l+1}-1} \int_k^{k+1} x^{-1/3}\, \mathrm{d}x = \int_{2^l}^{2^{l+1}} x^{-1/3}\, \mathrm{d}x = \frac{3}{2}\left((2^{l+1})^{2/3} - (2^l)^{2/3} \right)$$
$$= \frac{3}{2}\left(2^{2/3} - 1\right)(2^l)^{2/3} \geq \frac{1}{2}(2^l)^{2/3}.$$

Therefore, we have

$$\min_{2^l \leq j \leq 2^{l+1}-1} \|\nabla F_j(x_j)\| \leq \frac{C}{(2^l)^{1/3}},$$

with C defined as before, that is, $C = 2\sqrt{L_{\nabla h} + (2\rho)\|A\|^2}\sqrt{F_1(x_1) - F^* + (2\rho)^{-1}L_g^2}$. Noting that $z_j := \text{prox}_{\mu_j g}(Ax_j)$, we have as in (30) that

$$\min_{2^l \le j \le 2^{l+1}-1} \text{dist}(-\nabla h(x_j), A^*\partial g(z_j)) \le \frac{C}{(2^l)^{1/3}}, \tag{31}$$

as previously. Further, we have for $2^l \le j \le 2^{l+1} - 1$ that

$$\|Ax_j - z_j\| \le L_g \mu \le \frac{(2\rho)^{-1}L_g}{j^{1/3}} \le \frac{(2\rho)^{-1}L_g}{(2^l)^{1/3}}. \tag{32}$$

From (31) and (32) we deduce that Algorithm 2 must terminate before the end of epoch l, that is, before 2^{l+1} iterations have been completed, where l is the first nonnegative integer such that

$$2^l \ge \max\{C^3, (2\rho)^{-3}L_g^3\}\epsilon^{-3}.$$

Thus, termination occurs after at most $2\max\{C^3, (2\rho)^{-3}L_g^3\}\epsilon^{-3}$ iterations. □

For the case of A surjective, we have the following stronger result.

Corollary 3.1. *Suppose that the assumptions of Theorem 3.2 hold, that A is also surjective, and that the condition $\|Ax_{k+1} - \text{prox}_{\mu_{k+1}g}(Ax_{k+1})\| \le \epsilon$ in Algorithm 2 is replaced by $\|x_{k+1} - x_{k+1}^*\| \le \epsilon$, for $x_j^* := x_j - A^\dagger(Ax_j - \text{prox}_{\mu_j g}(Ax_j))$. Then for some $j' = O(\epsilon^{-3})$, we have that*

$$\text{dist}\left(-\nabla h(x_{j'}^*), A^*\partial g(Ax_{j'}^*)\right) \le \epsilon$$

and $\|x_{j'} - x_{j'}^*\| \le \epsilon$.

Proof. With the considerations made in the previous proof as well as the one made in (19) to (21), we can choose l to be the smallest positive integer such that

$$2^{l+1} \ge 2\max\{C^3, \sigma_{\min}(A)^{-3}L_g^3(2\rho)^{-3}\}\epsilon^{-3}.$$

The claim then holds for some $j' \le 2^{l+1}$. □

Although Algorithm 2 seems more complicated than Algorithm 1, the steps are the same. The only difference is that for the second algorithm, we do not search for the iterate that minimizes criticality across *all* iterations but only across at most the last $k/2$ iterations, where k is the total number of iterations.

4 Proximal Gradient

Here we derive a complexity bound for the proximal gradient algorithm applied to the more elementary problem (9) studied in Section 2.1, that is,

$$\min_{x \in \mathbb{R}^d} F(x) := h(x) + g(x), \tag{33}$$

for $h : \mathbb{R}^d \to \mathbb{R}$ a $L_{\nabla h}$-smooth function and $g : \mathbb{R}^d \to \overline{\mathbb{R}}$ a possibly nonsmooth, ρ-weakly convex function. Such a bound has not been made explicit before, to the authors' knowledge, though it is a fairly straightforward consequence of existing results. The bound makes a interesting comparison with the result in Section 3, where the nonsmoothness issue becomes more complicated due to the composition with a linear operator. In this section, we assume that a closed-form proximal operator is available for g, and we show that the complexity bound of $\mathcal{O}(\epsilon^{-2})$ is the same order as for gradient descent applied to smooth nonconvex functions.

Standard proximal gradient for (33), given parameter $\lambda \in (0, \min\{\rho^{-1}/2, L_{\nabla h}^{-1}\}]$ and initial point x_1, is as follows:

$$\begin{aligned} x_{k+1} &:= \arg\min_{x\in\mathbb{R}^d} \left\{ g(x) + \langle \nabla h(x_k), x - x_k\rangle + \frac{1}{2\lambda}\|x - x_k\|^2 \right\}, \qquad (34)\\ &= \mathrm{prox}_{\lambda g}\left(x_k - \lambda\nabla h(x_k)\right), \quad k = 1, 2, \ldots, \end{aligned}$$

where the choice of λ ensures that the function to be minimized in (34) is $(\lambda^{-1} - \rho)$-strongly convex, so that x_{k+1} is uniquely defined.

We have the following convergence result.

Theorem 4.1. *Consider the algorithm defined by* (34) *applied to problem* (33)*, where we assume that g is ρ-weakly convex and that ∇h is Lipschitz continuous with constant $L_{\nabla h}$. Supposing that $\lambda \in (0, \min\{\rho^{-1}/2, L_{\nabla h}^{-1}\}]$, we have for all $k \geq 1$ that*

$$\min_{2\leq j\leq k+1} \mathrm{dist}(0, \partial(h+g)(x_j)) \leq k^{-1/2}\sqrt{2(F(x_1) - F^*)}\,\frac{\lambda^{-1} + L_{\nabla h}}{\sqrt{\lambda^{-1} - \rho}},$$

where F^* is defined in (26)*.*

Proof. Note first that the result is vacuous if $F^* = -\infty$, so we assume henceforth that F^* is finite. We have for every $x \in \mathbb{R}^d$ that

$$\begin{aligned} g(x_{k+1}) + h(x_k) + \langle \nabla h(x_k), x_{k+1} - x_k\rangle + \frac{1}{2\lambda}\|x_{k+1} - x_k\|^2 + \frac{1}{2}(\lambda^{-1} - \rho)\|x - x_{k+1}\|^2 \\ \leq g(x) + h(x_k) + \langle \nabla h(x_k), x - x_k\rangle + \frac{1}{2\lambda}\|x - x_k\|^2. \end{aligned}$$

By applying the inequality

$$h(x_{k+1}) \leq h(x_k) + \langle \nabla h(x_k), x_{k+1} - x_k\rangle + \frac{1}{2\lambda}\|x_{k+1} - x_k\|^2, \quad \text{for all } x \in \mathbb{R}^d,$$

obtained from the Lipschitz continuity of ∇h and the fact that $\lambda \leq L_{\nabla h}^{-1}$, we deduce that

$$F(x_{k+1}) + \frac{1}{2}(\lambda^{-1} - \rho)\|x - x_{k+1}\|^2 \leq g(x) + h(x_k) + \langle \nabla h(x_k), x - x_k\rangle + \frac{1}{2\lambda}\|x - x_k\|^2,$$

for every $x \in \mathbb{R}^d$. By setting $x = x_k$, we obtain

$$F(x_{k+1}) + \frac{1}{2}(\lambda^{-1} - \rho)\|x_k - x_{k+1}\|^2 \leq F(x_k),$$

which shows, together with the definition (26), that

$$\sum_{k=1}^{\infty} \|x_k - x_{k+1}\|^2 \leq \frac{2(F(x_1) - F^*)}{\lambda^{-1} - \rho}. \tag{35}$$

From the optimality conditions for (34), we obtain

$$0 \in \nabla h(x_k) + \partial g(x_{k+1}) + \lambda^{-1}(x_{k+1} - x_k)$$

which also shows that

$$w_{k+1} := \frac{1}{\lambda}(x_k - x_{k+1}) + \nabla h(x_{k+1}) - \nabla h(x_k) \in \partial(h+g)(x_{k+1}), \tag{36}$$

so that

$$\|w_{k+1}\|^2 \leq \left(\lambda^{-1} + L_{\nabla h}\right)^2 \|x_k - x_{k+1}\|^2.$$

By combining this bound with (35), we obtain

$$\sum_{k=1}^{\infty} \|w_{k+1}\|^2 \leq 2(F(x_1) - F^*) \frac{\left(\lambda^{-1} + L_{\nabla h}\right)^2}{\lambda^{-1} - \rho}.$$

from which it follows that

$$\min_{1 \leq j \leq k} \|w_{j+1}\| \leq \sqrt{2(F(x_1) - F^*)}\, \frac{(\lambda^{-1} + L_{\nabla h})}{\sqrt{\lambda^{-1} - \rho}}.$$

The result now follows from (36), when we note that

$$\min_{1 \leq j \leq k} \operatorname{dist}(0, \partial(h+g)(x_{j+1})) \leq \min_{1 \leq j \leq k} \|w_{j+1}\|.$$

□

This theorem indicates that the proximal gradient algorithm requires at most $\mathcal{O}(\epsilon^{-2})$ to find an iterate with ϵ-approximate stationarity. This bound contrasts with the bound $\mathcal{O}(\epsilon^{-3})$ of Section 3 for the case of general A. Moreover, the $\mathcal{O}(\epsilon^{-2})$ bound has the same order as the bound for gradient descent applied to general smooth nonconvex optimization.

References

[1] Ilker Bayram. On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty. *IEEE Transactions on Signal Processing*, 64(6):1597–1608, 2015.

[2] A. E. Beaton and J. W. Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185, 1974.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[4] Radu Ioan Boţ and Axel Böhm. Variable smoothing for convex optimization problems using stochastic gradients. *arXiv preprint arXiv:1905.06553*, 2019.

[5] Radu Ioan Boţ and Ernö Robert Csetnek. On the convergence rate of a forward-backward type primal-dual splitting algorithm for convex optimization problems. *Optimization*, 64(1):5–23, 2015.

[6] Radu Ioan Boţ and Christopher Hendrich. Convergence analysis for a primal-dual monotone + skew splitting algorithm with applications to total variation minimization. *Journal of Mathematical Imaging and Vision*, 49(3):551–568, 2014.

[7] Radu Ioan Boţ and Christopher Hendrich. A variable smoothing algorithm for solving convex optimization problems. *TOP*, 23(1):124–150, 2015.

[8] Antonin Chambolle and Charles Dossal. On the convergence of the iterates of the “fast iterative shrinkage/thresholding algorithm”. *Journal of Optimization Theory and Applications*, 166(3):968–982, 2015.

[9] Antonin Chambolle, Matthias J. Ehrhardt, Peter Richtárik, and Carola-Bibiane Schönlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

[10] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1), 2011.

[11] Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018.

[12] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[13] Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178:503–558, 2019.

[14] Jianqing Fan. Comments on «Wavelets in Statistics: A Review» by A. Antoniadis. *Journal of the Italian Statistical Society*, 6(2):131, 1997.

[15] Tim Hoheisel, Maxime Laborde, and Adam Oberman. On proximal point-type algorithms for weakly convex functions and their connection to the backward Euler method, 2018.

[16] Adrian S. Lewis and Stephen J. Wright. A proximal method for composite minimization. *Mathematical Programming*, 158(1-2):501–546, 2016.

[17] Guoyin Li and Ting Kei Pong. Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of Computational Mathematics*, 18(5):1199–1232, 2018.

[18] Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(3):866–896, 2017.

[19] Boris S. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*, volume 330. Springer Science & Business Media, 2006.

[20] Ankit Parekh and Ivan W Selesnick. Convex denoising using non-convex tight frame regularization. *IEEE Signal Processing Letters*, 22(10):1786–1790, 2015.

[21] Xinyue Shen and Yuantao Gu. Nonconvex sparse logistic regression with weakly convex regularization. *IEEE Transactions on Signal Processing*, 66(12):3199–3211, 2018.

[22] Kiran Koshy Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. *arXiv preprint arXiv:1907.01543*, 2019.

[23] Quoc Tran-Dinh, Olivier Fercoq, and Volkan Cevher. A smooth primal-dual optimization framework for nonsmooth composite convex minimization. *SIAM Journal on Optimization*, 28(1):96–134, 2018.

[24] Băng Công Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3), 2013.

[25] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.