ON PROXIMAL POINT-TYPE ALGORITHMS FOR WEAKLY CONVEX FUNCTIONS AND THEIR CONNECTION TO THE BACKWARD EULER METHOD

TIM HOHEISEL, MAXIME LABORDE, AND ADAM OBERMAN

Abstract. In this article we study the connection between proximal point methods for nonconvex optimization and the backward Euler method from numerical Ordinary Differential Equations (ODEs). We establish conditions for which these methods are equivalent. In the case of weakly convex functions, for small enough parameters, the implicit steps can be solved using a strongly convex objective function. In practice, this method can be faster than gradient descent. In this paper we find the optimal value of the regularization parameter.

1. Introduction

Our motivation for this paper is the problem of local optimization of a weakly convex function $f : \mathbb{R}^n \to \mathbb{R}$ using a first order gradient oracle, cf. e.g. [Bec17]. Proximal point methods are alternative to gradient descent methods, see e.g. [PB+14]. Often proximal point methods are used when the objective splits into two functions, one of which is smooth and another one which is nonsmooth, but allows for an efficient computation or even an analytical formula for the proximal operator, as in the case with the l^1-regularization, see [PB+14]. Here, we primarily consider a smooth function, although at many places we nonsmooth mappings as well, and we use an approximate proximal point method as an inner loop in an optimization algorithm. The motivation is the empirical evidence that with the same number of gradient evaluations, the proximal method converges faster than the gradient descent method, see Figure 2 below. Similar ideas have been used in [LMH15, PLD+17]. In those works, proximal regularization replaces the convex objective function with a strongly convex auxiliary function.

We take a similar approach, but here we focus on the weakly convex case, and study the analytical properties of the proximal regularization. In particular we take the point of view - motivated by Partial Differential Equations (PDEs) - that the algorithm corresponds to gradient descent on a *regularized* function. This point of view was taken in [COO+17] where the emphasis was on the stochastic gradient case. The rate of convergence for the proximal point method for weakly convex functions has already been studied in [CDHS16]. Here we focus on exact gradients; in this simpler setting we can study the weakly convex case using tools from convex analysis and optimization.

2010 *Mathematics Subject Classification.* 26B25, 65K10, 90C25.

Key words and phrases. Proximal-point method, weak convexity, Moreau envelope, Euler method, θ-method.

It is well known in the numerical analysis field that implicit methods can take large time steps, whereas explicit methods require smaller ones, see e.g. [SH96]. The proximal point methods require the solution of a strongly convex optimization problem at each step, but allow for much longer time steps. Both the proximal point and the gradient descent methods can be interpreted as a time discretization of the Ordinary Differential Equation (ODE)

$$\frac{dx(t)}{dt} = -\nabla f(x(t)) \tag{GD-ODE}$$

We can also consider the non-differentiable case, where $\nabla f(x)$ is replaced by a subdifferential (regular/limiting/Clarke), see e.g. Section 2 or [RW98, Chapter 8], and we get the differential inclusion

$$\frac{dx(t)}{dt} \in -\partial f(x(t)).$$

Our starting point is a one-parameter family of discretizations, which appears in the numerical study of ODEs as the θ-method, cf. [SH96]. These methods are numerical discretizations of (GD-ODE) which interpolate between gradient descent (for $\theta = 0$) and the proximal point method (for $\theta = 1$). These are called the *forward* and *backward Euler method*, respectively, in the numerical ODE terminology.

Definition 1.1. The θ-method for (GD-ODE) corresponds to the time discretization

$$\frac{x_{n+1} - x_n}{\lambda} = -\nabla f((1-\theta)x_n + \theta x_{n+1}) \tag{1}$$

where λ is the time step. When $\theta = 0, 1$, the θ-method is called the explicit, implicit Euler method,

$$\frac{x_{n+1} - x_n}{\lambda} = -\nabla f(x_n), \qquad \frac{x_{n+1} - x_n}{\lambda} = -\nabla f(x_{n+1}),$$

respectively.

Note that we can generalize (1) to the nonsmooth case by

$$\frac{x_{n+1} - x_n}{\lambda} \in -\partial f((1-\theta)x_n + \theta x_{n+1}). \tag{2}$$

The θ-method from (1) and (2), respectively can be recovered from a proximal point-type iteration.

Lemma 1.2 (θ-method as θ-proximal point)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper (see below), $\theta \in (0, 1]$ and let $\{x_n\}$ be generated by*

$$x_{n+1} := \operatorname{argmin}_{y \in \mathbb{R}^n} \left\{ f((1-\theta)x_n + \theta y) + \frac{\theta}{2\lambda}\|x_n - y\|^2 \right\}. \tag{3}$$

Then $\{x_n\}$ satisfies (2).

Proof. We observe that the necessary optimality conditions for x_{n+1} read

$$0 \in \theta \partial f((1-\theta)x_n + \theta x_{n+1}) + \frac{\theta}{2\lambda}(x_{n+1} - x_n).$$

cf. [RW98, Exercise 8.8/10.7 and Theorem 8.15]. For $\theta \neq 0$ this is equivalent to (2). □

While the θ-method is implicit for $\theta \neq 0$ (meaning it requires the solution of a nonlinear equation or nonlinear optimization problem to find x_{n+1}), we can rewrite it as the gradient descent method on a modified function. In fact, defining the *θ-Moreau envelope* (see Section 3.1 for more details)

$$u^\theta(x,\lambda) := \inf_{y\in\mathbb{R}^n}\left\{ f((1-\theta)x+\theta y) + \frac{\theta}{2\lambda}\|x-y\|^2 \right\},$$

as we show below, for weakly convex f (see Section 2.1), the sequence (3) is also equivalent to

$$\frac{x_{n+1}-x_n}{\lambda} = -\nabla u^\theta(x_n;\lambda).$$

Remark 1.3 (PDE interpretation). Our analysis of the θ-Moreau envelope is based on direct arguments. An alternative approach is using the Hamilton-Jacobi PDE. It can be shown that the θ-Moreau envelope $u^\theta(x,\lambda)$ equals $v(x,\lambda)$ where $v(x,t)$ is the solution of the Hamilton-Jacobi equation

$$\partial_t v(x,t) = -\frac{\theta}{2}\|\nabla_x v(x,t)\|^2$$

along with initial data

$$v(x,0) = f(x).$$

In the special case $\theta = 1$, we recover the standard Hamilton-Jacobi equation for the Moreau envelope

$$\partial_t u(x,t) = -\frac{1}{2}\|\nabla_x u(x,t)\|^2,$$

see e.g. [Eva98]. Note that within this remark, we are using PDE notation and notion of weak solutions for Hamilton-Jacobi equations.

We define c-weak convexity below, and prove the following lemma.

Lemma 1.4. *Suppose f is c-weakly convex (see Section 2.1). Then x_{n+1}, solution of (3), can be found as the solution of a convex optimization problem, provided $\lambda,\theta > 0$ satisfy the following (generalized CFL condition/time step restriction):*

$$c\lambda\theta \leq 1.$$

Proof. To prove the CFL condition, notice that, since f is c-weakly convex, for all $x \in \mathbb{R}^n$, the function $y \mapsto f((1-\theta)x+\theta y)$ is $\theta^2 c$-weakly convex. Then for λ satisfying

$$\frac{\theta}{\lambda} \geq \theta^2 c \quad \Longleftrightarrow \quad c\lambda\theta \leq 1,$$

the mapping $y \mapsto f((1-\theta)x+\theta y) + \frac{\theta}{2\lambda}\|x-y\|^2$ is convex. □

Notation: The notation used is standard and widely consistent with the one used in [RW98]. However, here we use $\|\cdot\|$ to denote the Euclidean norm.

For $f : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ we define

$$\operatorname{argmin}_x f(x) := \left\{ \bar{x} \in \mathbb{R}^n \;\middle|\; f(\bar{x}) = \inf_x f(x) \right\}.$$

In order to indicate that a function F maps vectors in $\mathbb{R}^n$ to subsets in $\mathbb{R}^m$ we write $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ and call F *set-valued*. The domain of F is defined by

$$\operatorname{dom} F := \{ x \in \mathbb{R}^n \mid F(x) \neq \emptyset \}.$$

2. Preliminaries

We first recall standard concepts from nonsmooth analysis where, see [RW98]. A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}(:= \mathbb{R} \cup \{\pm\infty\})$ is called *closed* if its *epigraph*

$$\operatorname{epi} f := \{(x, \alpha) \mid f(x) \leq \alpha\}$$

is a closed set. We call it *proper if* $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and its *domain*

$$\operatorname{dom} f := \{x \mid f(x) < +\infty\}$$

is nonempty. Moreover, we call f *convex* if $\operatorname{epi} f$ is a convex set.

For a function $f : \mathbb{R} \to \overline{\mathbb{R}}$ its *(regular) subdifferential* at $\bar{x}$ with $f(\bar{x}) \in \mathbb{R}$ is defined by

$$\partial f(\bar{x}) := \{v \in \mathbb{R}^n \mid f(\bar{x}) + \langle v,\, x - \bar{x}\rangle + o(\|x - \bar{x}\|) \leq f(x)\ (x \in \mathbb{R}^n)\}.$$

If $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, convex it is well known that we have

$$\partial f(\bar{x}) := \{v \in \mathbb{R}^n \mid f(\bar{x}) + \langle v,\, x - \bar{x}\rangle \leq f(x)\ (x \in \mathbb{R}^n)\},$$

cf. e.g. [RW98, Proposition 8.12]. For $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ its *(Fenchel) conjugate* is the function $f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$ defined by

$$f^*(y) := \sup_{x \in \mathbb{R}^n} \{\langle x,\, y\rangle - f(x)\}.$$

If f is proper and has an affine minorant its conjugate f^* is always closed, proper, convex, see e.g. [RW98, Theorem 11.1] and notice that f is proper and has an affine minorant if and only if its *convex hull* is proper.

For $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ closed, proper, convex, the subdifferential and the conjugate function interact in the following way:

$$\bar{y} \in \partial f(\bar{x}) \iff \bar{x} \in \partial f^*(\bar{y}), \tag{4}$$

see e.g. [RW98, Proposition 11.3].

Given $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $\lambda > 0$, the *proximal mapping* or *prox-operator* is the set-valued map $P_\lambda f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined by

$$P_\lambda f(x) = \operatorname{argmin}_u \left\{ f(u) + \frac{1}{2\lambda}\|x - u\|^2 \right\},$$

while the *Moreau envelope* $e_\lambda f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is given by

$$e_\lambda f(x) = \inf_u \left\{ f(u) + \frac{1}{2\lambda}\|x - u\|^2 \right\}.$$

2.1. Weakly convex functions.

We next introduce a large class of functions for which $P_\lambda f(x)$ is a singleton for any $x \in \mathbb{R}^n$, hence in particular

$$e_\lambda f(x) = f(P_\lambda f(x)) = \frac{1}{2\lambda}\|P_\lambda f(x) - x\|^2 \quad (x \in \mathbb{R}^n).$$

Definition 2.1 (Weakly and strongly convex functions)**.** A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is called *c-weakly convex* if $f + \frac{c}{2}\|\cdot\|^2$ is closed, proper, convex. We denote by Γ_c the c-weakly convex functions, i.e.

$$\Gamma_c := \left\{ f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \;\middle|\; f + \frac{c}{2}\|\cdot\|^2 \text{ closed, proper, convex} \right\}.$$

A c-weakly convex function with $c < 0$ is called *c-strongly convex*.

Clearly, Γ_0 is the cone of closed, proper, convex functions. Moreover, we have

$$\Gamma_c \subset \Gamma_d \quad (c \leq d).$$

The following result which is due to [KT98] illustrates the richness of the class of weakly convex function. We would also like to refer to the similar result [RW98, Theorem 10.33].

Proposition 2.2 ([KT98, Proposition 1])**.** *Let $O \subset \mathbb{R}^n$ be open, convex and let $f_i \in C^{1,1}(O)$ $(i \in I)$ such that there exists $\bar{x} \in \bigcap_{i\in I} \operatorname{dom} f_i \cap O \neq 0$ such that $\sup_{i\in I} \|\nabla f_i(\bar{x})\| < \infty$. Moreover, assume that ∇f_i is L_i-Lipschitz on O with $L := \sup_{i\in I} L_i < \infty$. Then the function*

$$f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}, \quad f(x) := \sup_{i\in I} f_i(x)$$

is L-weakly convex and finite on O.

The central property of weakly convex functions is that if we add a "large enough" strongly convex term, the sum becomes *strongly convex*, hence both *coercive*, i.e.

$$\lim_{\|x\|\to\infty} \frac{f(x)}{\|x\|} \to \infty,$$

in particular, *level-bounded* and also *strictly convex*. We state this formally below.

Lemma 2.3. *Let $c > 0$ and $f \in \Gamma_c$. Then function*

$$\phi_\lambda := f + \frac{1}{2\lambda}\|\cdot\|^2 \quad \left(0 < \lambda < \frac{1}{c}\right), \tag{5}$$

is strongly convex, hence coercive and strictly convex.

Proof. This follows readily from the fact that any closed, proper, convex function has an affine minorant, see [RW98, Proposition 8.12]. □

The next result is clear from an elementary sum rule.

Proposition 2.4. *Let $c > 0$ and $f \in \Gamma_c$. Then for $0 < \lambda < \frac{1}{c}$ we have*

$$\partial f(x) = \partial\phi_\lambda(x) - \frac{x}{\lambda} \quad (x \in \operatorname{dom} f),$$

where ϕ_λ is given by (5).

Proof. See e.g. [RW98, Exercise 10.10]. □

We also point out that weakly convex functions are *Clarke regular* (see Definition [RW98, Definition 7.25]) hence their regular and limiting subdifferential coincide. In particular, for a (finite-valued) weakly convex functions, the (regular) subdifferential is equal to Clarke's subdifferential, i.e. can be computed as

$$\partial f(\bar{x}) = \operatorname{conv} \left\{ v \in \mathbb{R}^n \;\middle|\; \exists \{x^k \in D_f\} : \nabla f(x^k) \to v \right\} \tag{6}$$

where D_f is the (full measure) set of differentiability of f and conv is the convex hull-operator. We use (6) in Example 3.6.

2.2. **DC functions.** It is easily seen that Γ_c ($c \geq 0$) is contained in the (much larger class) of DC functions where a function f is called a *DC (difference of convex) function* if $f = g - h$ for some $g, h \in \Gamma_0$. We recall the central duality result for DC optimization.

Proposition 2.5 (Toland-Singer duality)**.** *Let $g, h \in \Gamma_0$. Then the following hold:*

a) $\inf g - h = \inf h^* - g^*$.
b) *If $\bar{x} \in \operatorname{argmin} g - h$ and $\bar{y} \in \partial h(\bar{x})$ then $\bar{y} \in \operatorname{argmin} h^* - g^*$.*
c) *If $\bar{y} \in \operatorname{argmin} h^* - g^*$ and $\bar{x} \in \partial g^*(\bar{y})$ the $\bar{x} \in \operatorname{argmin} g - h$.*

We point out that item a) and b) in Proposition 2.5 remain valid even if the convexity of g is dropped.

3. The prox-operator and Moreau envelope for weakly convex functions

In this section we study the Moreau envelope and proximal mapping for weakly convex functions. Many of the properties follow from more general results in variational analysis, see [RW98]. We will point out where this is the case. However, we present a vastly self-contained account only built on convex analysis (except when the nonconvex subdifferential is involved) and improve some of the existing results along the way.

Note that, given $f \in \Gamma_c$, ϕ_λ throughout denotes the function defined in Lemma 2.3.

Proposition 3.1 (Prox-operator of weakly convex functions)**.** *We have:*

a) *$P_\lambda f$ is a single-valued mapping $\mathbb{R}^n \to \mathbb{R}^n$.*
b) *$P_\lambda f = (\partial \phi_\lambda)^{-1}(\frac{\cdot}{\lambda}) = (\nabla \phi_\lambda^*)(\frac{\cdot}{\lambda})$ (which is single-valued).*
c) *$0 \in \partial f(x)$ if and only if $P_\lambda f(x) = x$, i.e. the critical points of f are exactly the fixed points of the prox-operator of $P_\lambda f$.*

Proof. a) By definition we have

$$P_\lambda f(x) = \operatorname{argmin}_u \left\{ f(u) + \frac{1}{2\lambda}\|x-u\|^2 \right\} = \operatorname{argmin}_u \left\{ \phi_\lambda(u) - \frac{1}{\lambda}\langle x, u\rangle \right\} \quad (x \in \mathbb{R}^n).$$

The function $u \mapsto \phi_\lambda(u) - \frac{1}{\lambda}\langle x, u\rangle$ is strongly convex for every $x \in \mathbb{R}^n$, see Lemma 2.3. Hence, the argmin set above is always a singleton.

b) We have

$$\begin{aligned} y \in (\partial\phi_\lambda)^{-1}\left(\frac{x}{\lambda}\right) &\iff \frac{x}{\lambda} \in \partial\phi_\lambda(y) \\ &\iff 0 \in \partial\left(\phi_\lambda - \frac{1}{\lambda}\langle x, \cdot\rangle\right)(y) \\ &\iff y \in P_\lambda f(x), \end{aligned}$$

where the second equivalence uses the convexity of ϕ_λ and the third one follows from the consideration above in a).

This proves the first equivalence in b). The second one then follows from (4).

c) We have

$$0 \in \partial f(x) \iff \frac{x}{\lambda} \in \partial\phi(x) \iff x = (\partial\phi)^{-1}\left(\frac{x}{\lambda}\right).$$

Here the first equivalence is due to Proposition 2.4. Part b) now gives the claim. □

Note that part b) is in a similar form given in [RW98, Proposition 12.19].

The following result constitutes a generalization of [BC11, Proposition 12.26] and its proof follows the same pattern. It is the key for establishing improved Lipschitz constants for the Moreau envelope.

Lemma 3.2. *Let $c > 0$ and $f \in \Gamma_c$, $x \in \mathbb{R}^n$, $0 < \lambda c < 1$, and put $p := P_\lambda f(x)$. Then*

$$f(p) + \frac{1}{\lambda}\langle x - p,\, y - p\rangle \leq f(y) + \frac{c}{2}\|p - y\|^2 \quad (y \in \mathbb{R}^n).$$

Proof. Let $y \in \mathbb{R}^n$ and put $p_\alpha := \alpha y + (1-\alpha)p$ for $\alpha \in (0,1)$. Then by the definition of the prox-operator we have

$$f(p) + \frac{1}{2\lambda}\|x - p\|^2 \leq f(p_\alpha) + \frac{1}{2\lambda}\|x - p_\alpha\|^2.$$

This is equivalent to

$$\begin{aligned} f(p) + \frac{c}{2}\|p\|^2 + \left(\frac{1}{2\lambda} - \frac{c}{2}\right)\|x - p\|^2 - c\langle x,\, p\rangle \\ \leq \quad f(p_\alpha) + \frac{c}{2}\|p_\alpha\|^2 + \left(\frac{1}{2\lambda} - \frac{c}{2}\right)\|x - p_\alpha\|^2 - c\langle x,\, p_\alpha\rangle, \end{aligned}$$

which, for $\phi := f + \frac{c}{2}\|\cdot\|^2 \in \Gamma_0$, is equivalent to

$$\phi(p) \leq \phi(p_\alpha) + \frac{1}{2}\left(\frac{1}{\lambda} - c\right)\left(\|(x-p) - \alpha(y-p)\|^2 - \|x-p\|^2\right) + c\langle x,\, p - p_\alpha\rangle.$$

The convexity of ϕ and the definition of p_α then imply

$$\phi(p) \leq \alpha\phi(y) + (1-\alpha)\phi(p) + \frac{1}{2}\left(\frac{1}{\lambda} - c\right)\left(\alpha^2\|y-p\|^2 - 2\alpha\langle x-p,\, y-p\rangle\right) + \alpha c\langle x,\, p - y\rangle.$$

Therefore, we have

$$\alpha\phi(p) + \frac{\alpha}{\lambda}\langle x - p,\, y - p\rangle \leq \alpha\phi(y) + \alpha c\langle p,\, p - y\rangle + \frac{\alpha^2}{2}\left(\frac{1}{\lambda} - c\right)\|y - p\|^2.$$

Dividing by $\alpha > 0$ and letting $\alpha \downarrow 0$ hence yields

$$f(p) + \frac{c}{2}\|p\|^2 + \frac{1}{\lambda}\langle x - p,\, y - p\rangle \leq f(y) + \frac{c}{2}\|y\|^2 + c\langle p,\, p - y\rangle,$$

i.e.

$$\begin{aligned} f(p) + \frac{1}{\lambda}\langle x - p,\, y - p\rangle \quad &\leq \quad f(y) + \frac{c}{2}\left(\|y\|^2 - \|p\|^2 + 2\langle p,\, p - y\rangle\right) \\ &= \quad f(y) + \frac{c}{2}\|p - y\|^2. \end{aligned}$$

□

From Lemma 3.2 we infer the following.

Proposition 3.3 (Expansiveness bound of prox)**.** *Let $c > 0$ and $f \in \Gamma_c$. Then for $0 < \lambda c < 1$ we have*

$$\|P_\lambda f(x) - P_\lambda f(y)\|^2 \leq \frac{1}{1 - c\lambda}\langle x - y,\, P_\lambda f(x) - P_\lambda f(y)\rangle \quad (x, y \in \mathbb{R}^n).$$

Proof. Let $x, y \in \mathbb{R}^n$ and put $p := P_\lambda f(x)$ and $q := P_\lambda f(y)$. By Lemma 3.2 we have

$$f(p) + \frac{1}{\lambda}\langle x - p,\, q - p\rangle \leq f(q) + \frac{c}{2}\|q - p\|^2$$

and

$$f(q) + \frac{1}{\lambda}\langle y - q,\, p - q\rangle \leq f(p) + \frac{c}{2}\|q - p\|^2.$$

Adding the above inequalities yields

$$\frac{1}{\lambda}\langle p - q - (x - y),\, p - q\rangle \leq c\|p - q\|^2.$$

Rearranging gives the desired inequality. □

As an immediate consequence of Proposition 3.3 we recover the well-known result, see [RW98, Proposition 12.19] that $P_\lambda f$ is $\frac{1}{1-\lambda c}$-Lipschitz continuous for any $f \in \Gamma_c$ and $0 < c\lambda$.

We now turn our attention to the Moreau envelope. We point out that the Lipschitz constant for the gradient of the Moreau envelope is, to the best of our knowledge, sharper than what can be found in the literature.

Corollary 3.4 (Moreau envelope)**.** *Let $c > 0$ and $f \in \Gamma_c$. Then the following hold for $0 < \lambda c < 1$:*

a) $e_\lambda f = \frac{1}{2\lambda}\|\cdot\|^2 - (f + \frac{1}{2\lambda}\|\cdot\|^2)^*\left(\frac{\cdot}{\lambda}\right)$.
b) $\nabla e_\lambda f = \frac{1}{\lambda}\left(\mathrm{id} - P_\lambda f\right)$ *is L-Lipschitz with*

$$L = \begin{cases} \frac{c}{1-c\lambda} & \text{if} \quad \frac{1}{2} \leq c\lambda < 1, \\ \frac{1}{\lambda} & \text{if} \quad 0 < c\lambda < \frac{1}{2}. \end{cases}$$

.
c) $\inf f = \inf e_\lambda f$.
d) $0 \in \partial f(x)$ *if and only if* $\nabla e_\lambda f(x) = 0$, *i.e. the stationary points of f and $e_\lambda f$ coincide.*
e) $\operatorname{argmin} f = \operatorname{argmin} e_\lambda f$.

Proof. Put $\phi_\lambda := f + \frac{1}{2\lambda}\|\cdot\|^2$.

a) We observe that

$$\begin{aligned} e_\lambda f(x) &= \frac{1}{2\lambda}\|x\|^2 - \sup_u \left\{\frac{1}{\lambda}\langle x,\, u\rangle - \phi_\lambda(u)\right\} \\ &= \frac{1}{2\lambda}\|x\|^2 - \phi_\lambda^*\left(\frac{x}{\lambda}\right). \end{aligned}$$

b) By Proposition 3.1, ϕ_λ is strongly convex, hence ϕ_λ^* is continuously differentiable with Lipschitz gradient, see e.g. [RW98, Proposition 12.60]. Thus, by a), we have

$$\nabla e_\lambda f = \frac{1}{\lambda}\left(\mathrm{id} - \nabla\phi_\lambda^*\left(\frac{\cdot}{\lambda}\right)\right).$$

Since, by Proposition 3.1 c), $P_\lambda f = (\partial\phi_\lambda)^{-1}(\frac{\cdot}{\lambda}) = (\nabla\phi_\lambda^*)(\frac{\cdot}{\lambda})$, this gives the formula for $\nabla e_\lambda f$.

The Lipschitz modulus can be seen as follows: By Proposition 3.3, we have

$$\begin{aligned}&\|(x-p)-(y-q)\|^2\\ &= \|x-y\|^2+\left(\frac{1}{1-c\lambda}-2\right)\langle p-q,\,x-y\rangle+\|p-q\|^2-\frac{1}{1-c\lambda}\langle p-q,\,x-y\rangle\\ &\leq \|x-y\|^2+\left(\frac{1}{1-c\lambda}-2\right)\langle p-q,\,x-y\rangle\end{aligned}$$

Now observe that

$$\frac{1}{1-c\lambda}-2\geq 0 \quad\Longleftrightarrow\quad \frac{1}{2}\leq c\lambda.$$

As $\langle p-q,\,x-y\rangle\geq 0$ (cf. Proposition 3.3), we thus have

$$\|(x-p)-(y-q)\|^2\leq\|x-y\|^2 \quad \left(\frac{1}{2}\geq c\lambda\right).$$

On the other hand, for $\frac{1}{2}\leq c\lambda$, we can continue the sequence of inequalities from above using Proposition 3.3 and Cauchy-Schwarz to find

$$\begin{aligned}\|(x-p)-(y-q)\|^2 &\leq \|x-y\|^2+\left(\frac{1}{1-c\lambda}-2\right)\langle p-q,\,x-y\rangle\\ &\leq \|x-y\|^2+\left(\frac{1}{1-c\lambda}-2\right)\|p-q\|\cdot\|x-y\|\\ &\leq \|x-y\|^2+\frac{1}{1-c\lambda}\left(\frac{1}{1-c\lambda}-2\right)\|x-y\|^2\\ &= \left(\frac{c\lambda}{1-c\lambda}\right)^2\|x-y\|^2,\end{aligned}$$

All in all, putting

$$M:=\begin{cases}\left(\frac{c\lambda}{1-c\lambda}\right)^2 & \text{if}\quad \frac{1}{2}\leq c\lambda<1,\\ 1 & \text{if}\quad 0<c\lambda<\frac{1}{2}\end{cases}$$

we see that

$$\|\nabla e_\lambda f(x)-\nabla e_\lambda f(y)\|=\frac{1}{\lambda}\|(x-p)-(y-q)\|\leq\frac{1}{\lambda}\sqrt{M}\|x-y\|,$$

which proves the desired Lipschitz constant.

c) We have

$$\begin{aligned}\inf_x f(x) &= \inf_x\left\{\phi_\lambda(x)-\frac{1}{2\lambda}\|x\|^2\right\}\\ &= \frac{1}{\lambda}\inf_x\left\{(\lambda\phi_\lambda)(x)-\frac{1}{2}\|x\|^2\right\}\\ &= \frac{1}{\lambda}\inf_y\left\{\frac{1}{2}\|y\|^2-(\lambda\phi_\lambda)^*(y)\right\}\\ &= \frac{1}{\lambda}\inf_y\left\{\frac{1}{2}\|y\|^2-\lambda\phi_\lambda^*\left(\frac{y}{\lambda}\right)\right\}\\ &= \inf_y\left\{\frac{1}{2\lambda}\|y\|^2-\phi_\lambda^*\left(\frac{y}{\lambda}\right)\right\}\\ &= \inf_y e_\lambda f(y).\end{aligned}$$

Here the third equality uses Toland-Singer duality (see Proposition 2.5) and the last equality is due to a).

d) Follows from b) and Proposition 3.1 c).

e) Let $\lambda > 0$ such that $\lambda c < 1$. Then $\phi_\lambda = f + \frac{1}{2\lambda}\|\cdot\|^2 \in \Gamma_0$. Using the same arguments as in c) we find that

$$\operatorname{argmin} f = \operatorname{argmin} \lambda\phi_\lambda - \frac{1}{2}\|\cdot\|^2 \quad \text{and} \quad \operatorname{argmin} e_\lambda f = \operatorname{argmin} \frac{1}{2}\|\cdot\|^2 - (\lambda\phi_\lambda)^*.$$

We now apply Proposition 2.5 to $g := \lambda\phi_\lambda$ and $h := \frac{1}{2}\|\cdot\|^2$: Since $\nabla h = \mathrm{id}$, Proposition 2.5 b) gives the '$\subset$'-inclusion immediately.

In turn, let $\bar{y} \in \operatorname{argmin} e_\lambda f$. Combining (4) and Proposition 3.1 b), we observe that $\partial g^*(\bar{y}) = P_\lambda f(\bar{y})$. Therefore, by Proposition 2.5 c), $P_\lambda f(\bar{y}) \in \operatorname{argmin} f$. But every minimizer of f is a fixed point of the prox-operator, cf. Proposition 3.1 c), and therefore $\bar{y} = P_\lambda f(\bar{y}) \in \operatorname{argmin} f$, which proves the remaining inclusion.

□

The fact in Corollary 3.4 c) and e) that the optimal value and minimizers, respectively, of f and its Moreau envelope conincide is well-known, and valid under even weaker assumptions, see [RW98, Example 1.46]. However, our technique of proof via DC duality theory remains in the convex realm and merits presentation of said proof.

Remark 3.5 (Optimal parameter choice for λ). Corollary 3.4 b) provides us with an "optimal choice" for the parameter λ: Suppose that

$$c := \inf\left\{c \geq 0 \;\middle|\; f + \frac{c}{2}\|\cdot\|^2 \text{ convex}\right\} > 0$$

Then $\lambda = \frac{1}{2c}$ yields $L = 2c$ which is as small as the Lipschitz constant can be for a given f.

In view of the Lipschitz constant for $\nabla e_\lambda f$ derived in Corollary 3.4 b) the question as to whether this constant can be improved generally in the class Γ_c arises naturally. The following example gives an illustration of Corollary 3.4 and also provides a negative answer to this question, in that it presents a Γ_c-function for which the Lipschitz constant provided by Corollary 3.4 is sharp in either case.

Example 3.6 (Piecewise quadratic). For $0 < a < b$ consider $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) := \max\left\{\frac{a}{2}(1-x^2), \frac{b}{2}(x^2-1)\right\} = \begin{cases} \frac{a}{2}(1-x^2) & \text{if} \quad |x| \leq 1, \\ \frac{b}{2}(x^2-1) & \text{if} \quad |x| > 1. \end{cases}$$

Then, clearly, f is a-weakly convex. Using (6) we find that

$$\partial f(x) = \begin{cases} -ax & \text{if} \quad |x| < 1, \\ [-a, b] & \text{if} \quad x = 1, \\ [-b, a] & \text{if} \quad x = -1, \\ bu & \text{if} \quad |x| > 1, \end{cases} \qquad P_\lambda f(x) = \begin{cases} \frac{x}{1-\lambda a} & \text{if} \quad |x| < 1-\lambda a, \\ 1 & \text{if} \quad x \in [1-\lambda a, 1+\lambda b], \\ -1 & \text{if} \quad x \in [-(1+\lambda b), \lambda a - 1)], \\ \frac{x}{1+\lambda b} & \text{if} \quad |x| > 1+\lambda b. \end{cases}$$

Therefore we have

$$e_\lambda f(x) = \begin{cases} \frac{a}{2}\left(1 - \frac{x^2}{(1-\lambda a)^2}\right) + \frac{1}{2\lambda}\left(x - \frac{x}{1-\lambda a}\right)^2 & \text{if} \quad |x| < 1-\lambda a, \\ \frac{1}{2\lambda}(x-1)^2 & \text{if} \quad x \in [1-\lambda a, 1+\lambda b], \\ \frac{1}{2\lambda}(x+1)^2 & \text{if} \quad x \in [-(1+\lambda b), \lambda a - 1], \\ \frac{b}{2}\left(\frac{x^2}{(1+\lambda b)^2} - 1\right) + \frac{1}{2\lambda}\left(x - \frac{x}{1+\lambda b}\right)^2 & \text{if} \quad |x| > 1+\lambda b \end{cases}$$

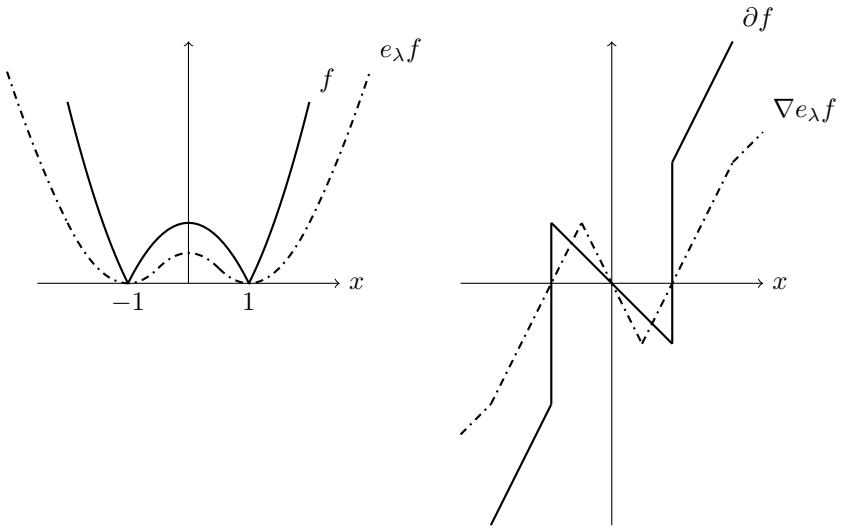

FIGURE 1. Illustration of Example 3.6

and

$$\nabla e_\lambda f(x) = \begin{cases} -\frac{a}{1-\lambda a}x & \text{if} \quad |x| < 1-\lambda a, \\ \frac{x-1}{\lambda} & \text{if} \quad x \in [1-\lambda a, 1+\lambda b], \\ \frac{x+1}{\lambda} & \text{if} \quad x \in [-(1+\lambda b), \lambda a - 1], \\ \frac{b}{1+\lambda b}x & \text{if} \quad |x| > 1+\lambda b. \end{cases}$$

In particular, we see that the Lipschitz constant

$$L = \begin{cases} \frac{a}{1-a\lambda} & \text{if} \quad \frac{1}{2} \le a\lambda, \\ \frac{1}{\lambda} & \text{if} \quad \frac{1}{2} > a\lambda \end{cases}$$

for $\nabla e_\lambda f$ provided by Corollary 3.4 b) is sharp.

3.1. **The θ-envelopes.** We now generalize the notion of the proximal point mapping and Moreau envelope by embedding them in a parameterized family of proximal mappings and envelopes, respectively.

Definition 3.7 (θ-Moreau envelopes)**.** Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $\theta, \lambda > 0$. The *θ-proximal point method* is the map $P^\theta_\lambda f : \mathbb{R}^n \to \mathbb{R}^n$ given by

$$P^\theta_\lambda f(x) = \operatorname{argmin}_{y \in \mathbb{R}^n} \left\{ f\left((1-\theta)x + \theta y\right) + \frac{\theta}{2\lambda}\|x-y\|^2 \right\}.$$

The *θ-Moreau envelope* is the function $e^\theta_\lambda f : \mathbb{R}^n \to \overline{\mathbb{R}}$ defined by

$$e^\theta_\lambda f(x) := \inf_{y \in \mathbb{R}^n} \left\{ f((1-\theta)x + \theta y) + \frac{\theta}{2\lambda}\|x-y\|^2 \right\}.$$

The following result shows the intimate relation of the θ-envelope and the θ-method objects to the Moreau envelope and the prox-operator.

Lemma 3.8. *Let $\theta, \lambda > 0$ and $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. Then the following hold:*

a) $e^\theta_\lambda f = e_{\lambda\theta} f$;

b) $P_\lambda^\theta f = \frac{P_{\lambda\theta} f - (1-\theta)\mathrm{id}}{\theta}$, *i.e.*

$$P_{\lambda\theta} f(x) = (1-\theta)x + \theta P_\lambda^\theta f(x) \quad (x \in \mathbb{R}^n).$$

Proof. Let $\bar{x} \in \mathbb{R}^n$ be fixed. The mapping

$$y \mapsto (1-\theta)\bar{x} + \theta y$$

is bijective on $\mathbb{R}^n$. Therefore, we observe that

$$\begin{aligned} e_\lambda^\theta f(\bar{x}) &= \inf_{y\in\mathbb{R}^n} \left\{ f((1-\theta)\bar{x} + \theta y) + \frac{\theta}{2\lambda}\|\bar{x} - y\|^2 \right\} \\ &= \inf_{u\in\mathbb{R}^n} \left\{ f(u) + \frac{\theta}{2\lambda} \left\| \bar{x} - \frac{u - (1-\theta)\bar{x}}{\theta} \right\|^2 \right\} \\ &= \inf_{u\in\mathbb{R}^n} \left\{ f(u) + \frac{\theta}{2\lambda} \left\| \frac{\bar{x} - u}{\theta} \right\|^2 \right\} \\ &= \inf_{u\in\mathbb{R}^n} \left\{ f(u) + \frac{1}{2\lambda\theta} \|\bar{x} - u\|^2 \right\} \\ &= e_{\lambda\theta} f(\bar{x}). \end{aligned}$$

This proves a). In order to prove b) just revisit the above reasoning and observe that

$$y \in \operatorname{argmin}_{y\in\mathbb{R}^n} \left\{ f((1-\theta)\bar{x} + \theta y) + \frac{\theta}{2\lambda}\|\bar{x} - y\|^2 \right\}$$

if and only if

$$(1-\theta)\bar{x} + \theta y \in \operatorname{argmin}_{u\in\mathbb{R}^n} \left\{ f(u) + \frac{1}{2\lambda\theta} \|\bar{x} - u\|^2 \right\}.$$

□

We readily infer the following result.

Corollary 3.9 (θ-envelope)**.** *For $c > 0$ let $f \in \Gamma_c$ and $\theta, \lambda > 0$ such that $0 < c\theta\lambda < 1$. Then the following hold:*

a) $e_\lambda^\theta f = \frac{1}{2\lambda\theta}\|\cdot\|^2 - \left(f + \frac{1}{2\lambda\theta}\|\cdot\|\right)^* \left(\frac{\cdot}{\lambda\theta}\right)$.

b) $\nabla e_\lambda^\theta f = \frac{1}{\lambda\theta}(\mathrm{id} - P_{\lambda\theta} f) = \frac{1}{\lambda}\left(\mathrm{id} - P_\lambda^\theta f\right)$ *is L-Lipschitz with*

$$L = \begin{cases} \frac{c}{1-c\lambda\theta} & \text{if} \quad \frac{1}{2} \le c\lambda\theta < 1, \\ \frac{1}{\lambda\theta} & \text{if} \quad 0 < c\lambda\theta < \frac{1}{2}. \end{cases}$$

c) $\inf f = \inf e_\lambda^\theta f$.

d) $0 \in \partial f(x)$ *if and only if* $\nabla e_\lambda^\theta f(x) = 0$.

e) $\operatorname{argmin} f = \operatorname{argmin} e_\lambda^\theta f$.

Proof. Follows immediately from combining Corollary 3.4 with Lemma 3.8. □

4. Guaranteed decrease of f

In this section, we study the behavior of a differentiable function $f \in \Gamma_c$ in the θ-method discretization for the gradient descent

$$\frac{x_{k+1} - x_k}{\lambda} = -\nabla f((1-\theta)x_k + \theta x_{k+1}). \tag{7}$$

We say that ∇f is *one-sided L_f-Lipschitz* if

$$f(x) \leq f(y) + \langle \nabla f(y), x - y\rangle + \frac{L_f}{2}\|x-y\|^2 \quad (x, y \in \mathbb{R}^n). \tag{8}$$

Proposition 4.1. *$f \in \Gamma_c$ be such that ∇f is one-sided L_f-Lipschitz and let $\{x^k\}$ be generated by (7) for $\theta \in [0,1]$. Then*

$$f(x_{k+1}) - f(x_k) \leq \left(\frac{L_f(1-\theta)^2 + c\theta^2}{2} - \frac{1}{\lambda}\right)|x_{k+1} - x_k|^2 \quad (k \in \mathbb{N}).$$

In particular, if $\lambda \in \left(0, \frac{2}{L_f(1-\theta)^2 + c\theta^2}\right)$, the sequence $\{f(x_k)\}$ is decreasing.

Proof. Denote $x_\theta = (1-\theta)x_k + \theta x_{k+1}$. By weak convexity and (8), we obtain

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= f(x_{k+1}) - f(x_\theta) + f(x_\theta) - f(x_k) \\ &\leq \langle \nabla f(x_\theta), x_{k+1} - x_\theta\rangle + \frac{L_f}{2}|x_{k+1} - x_\theta|^2 \\ &+ \langle \nabla f(x_\theta), x_\theta - x_k\rangle + \frac{c}{2}|x_\theta - x_k|^2. \end{aligned}$$

By definition of x_θ, we have

$$x_{k+1} - x_\theta = (1-\theta)(x_{k+1} - x_k) \text{ and } x_\theta - x_k = \theta(x_{k+1} - x_k),$$

which then yields the desired inequality. □

In addition, given $f \in \Gamma_c$, we have already seen that the sequence $\{x_k\}$ generated by (7) can be interpreted as a sequence obtained from applying the gradient descent to the θ-envelope $e_\lambda^\theta f$. By Corollary 3.9, we know that $\nabla e_\lambda^\theta f$ is L-Lipschitz which implies that $e_\lambda^\theta f$ satisfies (8) and thus the following result follows readily.

Proposition 4.2. *Let $f \in \Gamma_c$, $\theta \in (0,1]$, $\lambda > 0$ such that $0 < c\theta\lambda < 1$, and let $\{x^k\}$ be generated by (7). Then*

$$e_\lambda^\theta f(x_{k+1}) - e_\lambda^\theta(x_k) \leq \left(L - \frac{1}{\lambda}\right)\|x_{k+1} - x_k\|^2,$$

where $L > 0$ is the Lipschitz constant in Corollary 3.9. In particular, if $\lambda < \frac{1}{L}$, the sequence $\{e_\lambda^\theta f(x_k)\}$ decreases.

Proof. Follows immediately from (8). □

5. Perspectives on the proximal point method for weakly convex functions

In this section we present different interpretations of the proximal point method, namely as gradient descent, DC algorithm and proximal-gradient method, all of which provide different insights.

5.1. Proximal point as gradient descent.

Proposition 5.1. *Let $c > 0$ and $f \in \Gamma_c$. Moreover, let $\lambda, \varepsilon \geq 0$ such that*

$$\max\left\{\varepsilon, \frac{1}{2c}\right\} \leq \lambda \leq \left(\frac{2-\varepsilon}{3-\varepsilon}\right)\frac{1}{c}. \tag{9}$$

Now, let $\{x^k\}$ be generated by the proximal point method with constant step-size λ, i.e.

$$x^{k+1} := P_\lambda f(x^k) \quad (k \in \mathbb{N}), \quad x^0 \in \mathbb{R}^n.$$

Then every accumulation point of $\{x^k\}$ is a stationary point of f.

Proof. Using Proposition 3.4 b), observe that

$$x^{k+1} = x^k + \big(P_\lambda f(x^k) - x^k\big) = x^k - \lambda \nabla e_\lambda f(x^k) \quad (k \in \mathbb{N}),$$

i.e. $\{x^k\}$ is in fact generated by the gradient method with constant step-size λ for the function $e_\lambda f$. Condition (9) ensures on the one hand that $\nabla e_\lambda f$ has Lipschitz constant $L := \frac{c}{1-c\lambda}$, cf. Corollary 3.4 b). On the other it guarantees that $\varepsilon \leq \lambda \leq (2-\varepsilon)\frac{1}{L}$. Therefore, [Ber99, Proposition1.2.3] is applicable and implies that every accumulation point of $\{x^k\}$ is a stationary point of $\nabla e_\lambda f$, which by Proposition 3.4 d) gives the desired statement. □

We illustrate the above result by two examples. In the first one we revisit Example (3.6).

Example 5.2 (Piecewise quadratic). In Example 3.6, for $a = 1, b = 2$, the function

$$f(x) = \max\left\{\frac{1}{2}(1-x^2), (x^2-1)\right\},$$

is 1-weakly convex. By Remark 3.5, the optimal parameter choice is $\lambda = \frac{1}{2}$. Then the proximal point method $x_{k+1} = P_{1/2} f(x_k)$ is explicit:

- if $x_0 = 0, 1, -1$ then the sequence is constantly equal to $0, 1$ and -1, respectively;
- if $x_0 \in \left[\frac{1}{2^{K+1}}, \frac{1}{2^K}\right) \cup (2^K, 2^{K+1}]$, for a fixed $K \in \mathbb{N}$, then the algorithm converges in $K+1$ steps to 1,
- if $x_0 \in \left(-\frac{1}{2^{K+1}}, -\frac{1}{2^K}\right] \cup [-2^K, -2^{K+1})$, for a fix $K \in \mathbb{N}$, then the algorithm converges in $K+1$ steps to -1.

The second example concerns the classical *Rosenbrock function*.

Example 5.3 (Rosenbrock function). Consider the Rosenbrock function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x, y) = (x-1)^2 + 100(y - x^2)^2.$$

In Figure 2, we plot the iterations for the gradient descent and the proximal point method with the optimal parameter choice $\lambda = \frac{1}{2c}$. In addition, we observe the decay of the Rosenbrock function.

5.2. **Proximal point method as DC algorithm.** A very popular and powerful algorithm for solving DC optimization problems of the form

$$\min f = g - h \tag{10}$$

with $g, h \in \Gamma_0$ is the so-called *DC Algorithm*, *DCA* for short, which goes back to An and Tao, see e.g. [AT97]. In its simplified version (which coincides with the original version in our setting) it reads as follows:

(1) Choose $x^0 \in \operatorname{dom} \partial h$;
(2) Compute $y^k \in \partial h(x^k)$;
(3) Compute $x^{k+1} \in \partial g^*(y^k)$.

We point out that DCA applied to (10) is well-defined if (and only if)

$$\operatorname{dom} \partial g \subset \operatorname{dom} \partial h \text{ and } \operatorname{dom} \partial h^* \subset \operatorname{dom} \partial g^*, \tag{11}$$

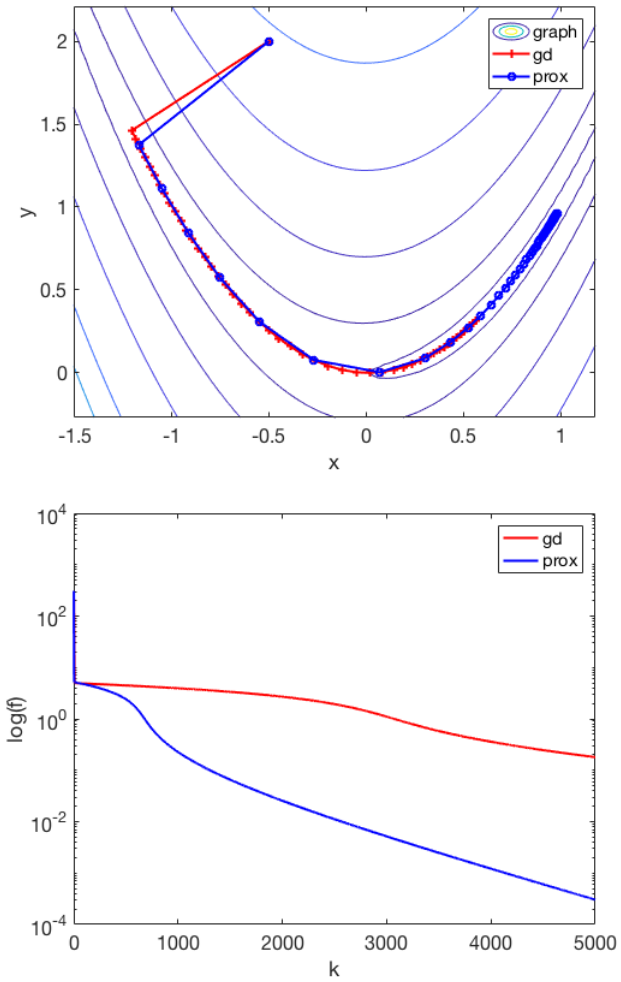

FIGURE 2. Top: Iterations of gradient descent and the proximal point method. The proximal point method gets closer to the global mininum with the same number of total gradient evaluations. Bottom: function values at each iteration of gradient descent and at each outer proximal point iteration for the Rosenbrock function (a fair comparison is used by counting total gradient evaluations, using 10 or 20 for the approximate proximal point). After 2000 gradient evaluations the function value for gradient descent is still order 1, while the proximal point method is order 10^{-2}, moreover the function values appear to decrease at a first order rate.

cf. [AT97, Lemma 1]. Now assume that $f \in \Gamma_c$. As was argued earlier, a natural DC decomposition of f is

$$f = \phi_\lambda - \frac{1}{2\lambda}\|\cdot\|^2 \quad (0 < \lambda c < 1),$$

where, as always, $\phi_\lambda = f + \frac{1}{2\lambda}\|\cdot\|^2$. Condition (11) is clearly satisfied. Hence, for any $x^0 \in \mathbb{R}^n$, the DCA is well-defined and generates the sequences

$$y^k = \frac{1}{\lambda}x^k \quad \text{and} \quad x^{k+1} = \nabla\phi_\lambda^*(y^k) = P_\lambda f(x^k),$$

cf. Proposition 3.1 b).

5.3. **Proximal point as proximal gradient.** Again, we consider the trivial decomposition

$$f = \phi_\lambda - \frac{1}{2\lambda}\|\cdot\|^2 \quad (0 < \lambda c < 1).$$

The proximal gradient iteration with $L_k := L = \frac{1}{\lambda}$, cf. [Bec17, Section 10.2], for this decomposition reads

$$x^{k+1} = P_{\frac{1}{L}}\phi_\lambda\left(x^k + \frac{1}{L}\nabla(\frac{1}{2\lambda}\|\cdot\|^2)(x^k)\right) = P_\lambda\phi_\lambda(2x^k).$$

On the other hand we have the following lemma.

Lemma 5.4. *For $f \in \Gamma_c$ we have*

$$P_\lambda\phi_\lambda(2x) = P_{\frac{\lambda}{2}}f(x) \quad (x \in \mathbb{R}^n,\ 0 < \lambda c < 1).$$

Proof. We have

$$\begin{aligned}
\{P_\lambda\phi_\lambda(2x)\} &= \operatorname{argmin}_y\left\{f(y) + \frac{1}{2\lambda}\|y\|^2 + \frac{1}{2\lambda}\|2x - y\|^2\right\}\\
&= \operatorname{argmin}_y\left\{f(y) + \frac{1}{\lambda}\|y\|^2 - \frac{2}{\lambda}\langle x, y\rangle + \frac{1}{2\lambda}\|2x\|^2\right\}\\
&= \operatorname{argmin}_y\left\{f(y) + \frac{1}{\lambda}\|y\|^2 - \frac{2}{\lambda}\langle x, y\rangle + \frac{1}{\lambda}\|x\|^2\right\}\\
&= \operatorname{argmin}_y\left\{f(y) + \frac{1}{\lambda}\|x - y\|^2\right\}\\
&= \left\{P_{\frac{\lambda}{2}}f(x)\right\}.
\end{aligned}$$

□

6. Final remarks

We studied proximal point-type methods for weakly convex functions where the main results were the following: We investigated the proximal mapping and Moreau envelope for weakly convex (not necessarily smooth) functions while establishing an optimal choice for the regularization parameter. In the smooth case we revealed a connection between the θ-proximal point method and the θ-method for gradient flows. Moreover, under an additional one-sided Lipschitz property we prove a guaranteed decrease of the regularized objective function for the θ-proximal point method. Finally, we gave three different interpretations of the proximal point method for (possibly nonsmooth) weakly convex functions, which provide new insights into the algorithm.

References

[AT97] L. T. H. An and P. D. Tao. Convex analysis approach to d.c. programming: theory, algorithms and applications. *Acta Math. Vietnam.*, 22(1):289–355, 1997.

[BC11] Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011. With a foreword by Hédy Attouch. URL: https://doi.org/10.1007/978-1-4419-9467-7.

[Bec17] Amir Beck. *First-Order Methods in Optimization*, volume 25. SIAM, 2017.

[Ber99] Dimitri P. Bertsekas. *Nonlinear programming*. Athena Scientific Optimization and Computation Series. Athena Scientific, Belmont, MA, second edition, 1999.

[CDHS16] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Accelerated Methods for Non-Convex Optimization. *ArXiv e-prints*, November 2016. arXiv:1611.00756.

[COO+17] Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto, and Guillame Carlier. Deep relaxation: partial differential equations for optimizing deep neural networks. *arXiv preprint arXiv:1704.04932*, 2017.

[Eva98] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 1998.

[KT98] A. Kaplan and R. Tichatschke. Proximal point methods and nonconvex optimization. *J. Global Optim.*, 13(4):389–406, 1998. Workshop on Global Optimization (Trier, 1997). URL: https://doi.org/10.1023/A:1008321423879.

[LMH15] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.

[PB+14] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[PLD+17] Courtney Paquette, Hongzhou Lin, Dmitriy Drusvyatskiy, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for gradient-based non-convex optimization, 2017. arXiv:arXiv:1703.10993.

[RW98] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998. URL: https://doi.org/10.1007/978-3-642-02431-3.

[SH96] AM Stuart and AR Humphries. Dynamical systems and numerical analysis, volume 2 of cambridge monographs on applied and computational mathematics, 1996.

E-mail address: tim.hoheisel@mcgill.ca

E-mail address: maxime.laborde@mail.mcgill.ca

E-mail address: adam.oberman@mcgill.ca