A REGULARIZATION INTERPRETATION OF THE PROXIMAL POINT METHOD FOR WEAKLY CONVEX FUNCTIONS

TIM HOHEISEL, MAXIME LABORDE, AND ADAM OBERMAN

Abstract. Empirical evidence and theoretical results suggest that the proximal point method can be computed approximately and still converge faster than the corresponding gradient descent method, in both the stochastic and exact gradient case. In this article we provide a perspective on this result by interpreting the method as gradient descent on a regularized function. This perspective applies in the case of weakly convex functions where proofs of the faster rates are not available. Using this analysis we find the optimal value of the regularization parameter in terms of the weak convexity.

1. Introduction

The importance of large scale optimization problems in machine learning has led to a resurgence of interest in first order optimization methods [Bec17], and stochastic gradient descent in particular [BCN16]. Proximal point methods [PB+14] are alternative to gradient descent methods, which first came to use in the setting where the proximal mapping can be computed exactly. Later, they were used in the stochastic setting where the proximal mapping can only be computed approximately [LMH15]. When the proximal point method parameter is tuned correctly, the proximal point method can converge faster than the corresponding stochastic gradient descent method [SRB11] [RVV14] [CDHS18]. However the optimal choice of the parameter depends on convexity parameters for the objective, which may not be available. Insight into the method is provided by a regularization interpretation: in [COO+18] the method was interpreted as gradient descent for a regularized objective function, which was the solution of a partial differential equation. The interpretation was also used to apply and tune the method in [YPOO18].

A heuristic explanation for the method motivated the implementation in [LMH15]: the proximal point method with parameter λ corresponds to implicit gradient descent with time step λ, which has a corresponding convergence rate. The convergence rate is much slower for stochastic gradient descent. However the proximal point method can be solved approximately, in a small number of iterations, even using stochastic gradients, since it is a strongly convex optimization problem. Thus stochastic proximal point method with parameter λ can converge as fast as *exact* gradient descent with time step λ. However, the challenge is to tune the parameter λ which depends on the unavailable weak-convexity parameter of the objective function.

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

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

research supported by: AFOSR FA9550-18-1-0167 (A.O.).

While many problems are non-convex, the model problem for analysis is convex, which allows for global analysis of convergence rates. Since many problems are ill-conditioned, accelerated methods [Pol64] [Nes13] [B+15] are used to improve convergence rates. Proximal stochastic methods can also be accelerated [PLD+18] [Nit14] [LMH15] (direct stochastic acceleration methods are also available [AZ17]).

Polyak's method [Pol64] provably accelerates strongly convex quadratic functions, but not general convex functions. However it has the advantage of a simple interpretation as the explicit Forward Euler discretization of a second order ODE. On the other hand, Nesterov's method provably accelerates convex functions, but defies such a simple interpretation. The influential paper [SBC14] provided an interpretation of Nesterov's method as the discretization of an ordinary differential equation (ODE), but with the gradients evaluated at an non-standard point in time. This interpretation was further studied in the quadratic case in [LRP16] and [FB15] as well as [SRBD17], using linear stability analysis. However, while linear stability analysis provides insight locally, it does not apply globally to convex functions.

In this paper we study two aspects of proximal methods. The first is a method which interpolates between gradient descent and the proximal point method. This may give insight into Nesterov's method, which involves a gradient evaluated at an intermediate point. In connection with the regularization interpretation of [COO+18], we study the convex analytical properties of the regularized function corresponding the proximal point method, in the weakly convex case. We also study the optimal parameter for the proximal step, in terms of the weak convexity of the objective function. In the article we focus on exact gradients rather than stochastic gradients. In this simpler setting we can study the weakly convex case using tools from convex analysis and optimization.

Both the proximal point and the gradient descent methods can be interpreted as a time discretization of the Ordinary Differential Equation (ODE)

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

When the ODE is discretized using either the forward or backward Euler method, the resulting algorithm corresponds to the gradient descent and the proximal point method as we explain below. Our starting point is a one-parameter family of discretizations of the 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 (for $\theta = 1$). The proximal point methods require the solution of a strongly convex optimization problem at each step, but allow for much longer time steps. 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

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

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:

$$\bar{y} \in \partial f(\bar{x}) \iff \bar{x} \in \partial f^*(\bar{y}), \tag{1}$$

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. **Discretizations of ODEs.**

Definition 2.1. The θ-method for (GD-ODE) corresponds to the time discretization

$$\frac{x^{k+1} - x^k}{\lambda} = -\nabla f((1-\theta)x^k + \theta x^{k+1}) \tag{2}$$

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

$$\frac{x^{k+1} - x^k}{\lambda} = -\nabla f(x^k), \qquad \frac{x^{k+1} - x^k}{\lambda} = -\nabla f(x^{k+1}),$$

respectively.

Note that we can generalize (2) to the nonsmooth case by

$$\frac{x^{k+1} - x^k}{\lambda} \in -\partial f((1-\theta)x^k + \theta x^{k+1}). \tag{3}$$

The θ-method from (2) and (3), respectively can be recovered from a proximal point-type iteration.

Lemma 2.2 (θ-method as θ-proximal point)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper (see below), $\theta > 0$ and let $\{x^k\}$ be generated by*

$$x^{k+1} := \operatorname{argmin}_{y\in\mathbb{R}^n} \left\{ f((1-\theta)x^k + \theta y) + \frac{\theta}{2\lambda}\|x^k - y\|^2 \right\}. \tag{4}$$

Then $\{x^k\}$ satisfies (3).

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

$$0 \in \theta\partial f((1-\theta)x^k + \theta x^{k+1}) + \frac{\theta}{2\lambda}(x^{k+1} - x^k).$$

cf. [RW98, Exercise 8.8/10.7 and Theorem 8.15]. For $\theta \neq 0$ this is equivalent to (3). □

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^{k+1}), we can rewrite it as the gradient descent method on a modified function. In fact, defining the *θ-Moreau envelope* (see Section 3.2 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 2.2), the sequence (4) is also equivalent to

$$\frac{x^{k+1} - x^k}{\lambda} = -\nabla u^\theta(x^k;\lambda).$$

Remark 2.3 (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 $u^\theta(x,\lambda) = v(x,\lambda)$ where $v(x,t)$ is the weak (viscosity) solution of the Hamilton-Jacobi equation

$$\partial_t v(x,t) = -\frac{\theta}{2}\|\nabla_x v(x,t)\|^2, \qquad 0 \le t \le \lambda$$

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 [Eva98].

2.2. **Weakly convex functions.** The proximal point algorithm is based on the fixed-point iteration

$$xk + 1 := P_\lambda f(x^k) = \operatorname{argmin}_{u\in\mathbb{R}^n} \left\{ f(u) + \frac{1}{2\lambda}\|x^k - u\|^2 \right\}$$

for some $\lambda > 0$. This is in essence only tractable if the subproblem for computing the prox-operator is convex, and ideally has a unique solution. A natural class of functions that does this is the following; see also Proposition 3.1.

Definition 2.4 (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 $\Gamma_c \subset \Gamma_d$ whenever $c \leq d$.

Weakly convex functions can be further generalized to the class of *lower-C^2 functions* [RW98, Definition 10.29, Theorem 10.33], and many of the ideas and results in the sequel will hold for these kinds of functions too, but for simplicity we confine ourselves to weakly convex ones.

The class of weakly convex functions also contains the Lasry-Lions regularization [LL86],

$$(f_\lambda)^\mu(x) := \sup_u \inf_v \left\{ f(v) + \frac{1}{2\lambda}\|v-u\|^2 - \frac{1}{2\mu}\|u-x\|^2 \right\}.$$

This regularization is a $\mathcal{C}^{1,1}$ function and, in [AA93], the authors show that any lower semi-continuous function defined on a Hilbert space, quadratically minorized can be approximate by the Lasry-Lions regularization.

We also point the reader to [KT98, Proposition 1] for another class of functions which are weakly convex restricted to some open set.

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 2.5. *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. Strong convexity is clear (by the c-weak convexity of f) and implies both coercivity (using an affine minorization argument, see [RW98, Proposition 8.12]) and obviously strict convexity. □

The next result is clear from an elementary sum rule.

Proposition 2.6. *Let $c > 0$ and $f \in \Gamma_c$. Then for $0 < \lambda < \frac{1}{c}$ we have*

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

where ϕ_λ is given by (5). In particular, $\partial f(x)$ is compact for every $x \in \operatorname{int}(\operatorname{dom} f)$, and $\operatorname{gph} \partial f = \{(x, g) \mid g \in \partial f(x)\}$ is closed in $\mathbb{R}^n \times \mathbb{R}^n$.

Proof. See e.g. [RW98, Exercise 10.10] for the representation of the subdifferential. The remainder follows from that and the fact that the respective statements hold for convex functions. □

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

Lemma 2.7. *Suppose f is c-weakly convex. Then x^{k+1}, solution of (4), 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

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

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

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

2.3. **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 2.8 (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 2.8 remain valid even if the convexity of g is dropped.

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

3.1. **The Moreau envelopes.** 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 and montone operator theory, see [BC11, 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.

Throughout this section we use the following:

Assumption 1:

(1) $f \in \Gamma_c \quad (c > 0)$;
(2) $\phi_\lambda := f + \frac{1}{2\lambda}\|\cdot\|^2 \quad (0 < \lambda < 1/c)$.

Proposition 3.1 (Prox-operator of weakly convex functions)**.** *Let f and ϕ_λ as in Assumption 1. Then the following hold:*

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, for all $x \in \mathbb{R}^n$,

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

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 2.5. 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 (1).

c) We have

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

Here the first equivalence is due to Proposition 2.6. 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 slight generalization of [BC11, Proposition 12.26] and its self-contained proof follows the same pattern.

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*

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

Proof. Let $y \in \mathbb{R}^n$ and ϕ_λ defined by (5). Using the $\left(\frac{1}{\lambda} - c\right)$-strong convexity of ϕ_λ and noticing that $\frac{x}{\lambda} \in \partial\phi_\lambda(p)$ by Proposition 3.1 b), we obtain

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

Therefore, we have

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

which is equivalent to

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

□

From Lemma 3.2 we infer the following property of the prox-operator, which in the literature is known as *cocoercivity*, and which can be derived as a consequence of the *Baillon-Haddad Theorem* [BH77, BC10, BC11] as ϕ_λ is $(\frac{1}{\lambda} - c)$-strongly convex, and $\nabla\phi_\lambda^* = P_\lambda f(\lambda(\cdot))$, see Proposition 3.1. Our proof is, however, self-contained as it only builds on Lemma 3.2 which itself is self-contained.

Proposition 3.3 (Cocoercivity of the prox-operator)**.** *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) $\nabla e_\lambda f = \frac{1}{\lambda}\,(\mathrm{id} - P_\lambda f)$ *is L-Lipschitz with*

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

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

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

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

where $p = P_\lambda f(x)$ and $q = P_\lambda f(y)$. Now observe that

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

First, considering the case $\frac{1}{2} \geq c\lambda$, as $\langle p-q, x-y\rangle \geq 0$ (cf. Proposition 3.3), we thus have

$$\|(x-p)-(y-q)\|^2 \leq \|x-y\|^2.$$

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 2.8) and the last equality is due to a).

d) 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 2.8 to $g := \lambda\phi_\lambda$ and $h := \frac{1}{2}\|\cdot\|^2$: Since $\nabla h = \mathrm{id}$, Proposition 2.8 b) gives the '$\subset$'-inclusion immediately.

In turn, let $\bar{y} \in \operatorname{argmin} e_\lambda f$. Combining (1) and Proposition 3.1 b), we observe that $\partial g^*(\bar{y}) = P_\lambda f(\bar{y})$. Therefore, by Proposition 2.8 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.

e) Follows from b) and Proposition 3.1 c). □

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

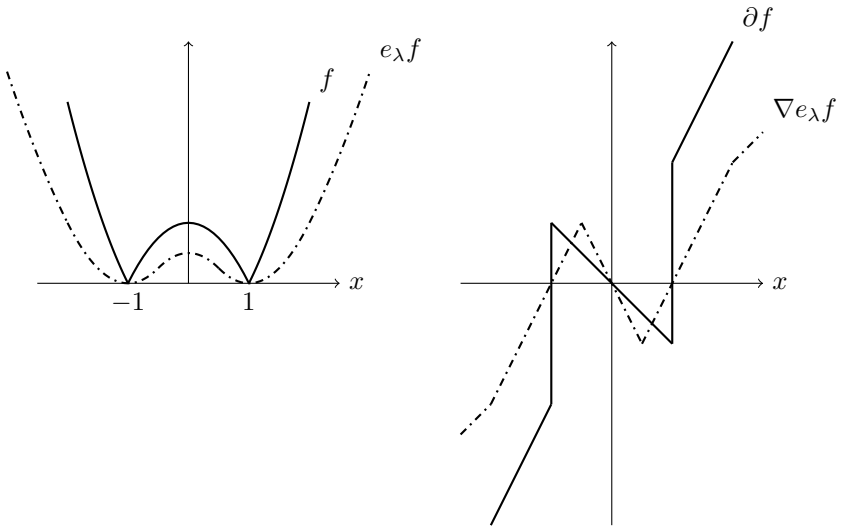

FIGURE 1. Illustration of Example 3.6

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

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

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

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

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.

3.2. **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 operator* is the map $P_\lambda^\theta f : \mathbb{R}^n \to \mathbb{R}^n$ given by

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

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

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

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

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*

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

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. Proximal-point as gradient descent

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

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

Note that in the following, θ can be chosen bigger than 1. In the implicit case ($\theta = 1$), the decrease of f along the iteration is straightforward and in Proposition 4.1, we extend this result for (7). We do not discuss about the rate of convergence and we refer to [AB09, MP10].

By Lemma 2.2, the method described in (7) is equivalent to the θ-proximal point method

$$x^{k+1} = P_\lambda^\theta f(x^k).$$

To simplify, we assume that f is differentiable but the results can be generalized to the nonsmooth case. We say that ∇f is *one-sided L_f-Lipschitz* if

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \leqslant L_f\|x-y\|^2 \quad (x, y \in \mathbb{R}^n)$$

which is equivalent to

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

It is easy to check that if ∇f is a L_f-Lipschitz function then ∇f is one-sided L_f-Lipschitz. In addition, if f is convex then the two conditions are equivalent.

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

$$f(x^{k+1}) - f(x^k) \leq \left(\frac{L_f(1-\theta)^2 + c\theta^2}{2} - \frac{1}{\lambda}\right)|x^{k+1} - x^k|^2 \quad (k \in \mathbb{N}). \tag{9}$$

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 using (7). □

Remark 4.2. Note that in the convex or strongly convex case i.e. $-L_f \leqslant c \leqslant 0$ and for all $\theta \in \left[\frac{\sqrt{L_f}}{\sqrt{L_f}+\sqrt{-c}}, \frac{\sqrt{L_f}}{\sqrt{L_f}-\sqrt{-c}}\right]$, we recover the fact that the descent of f is guaranteed for all $\lambda > 0$.

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 4.3. *Let $f \in \Gamma_c$, $\theta > 0$, $\lambda > 0$ 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 $\lambda \leqslant \frac{1}{L}$, the sequence $\{e_\lambda^\theta f(x^k)\}$ decreases.

Proof. Follows immediately from (8). □

From Proposition 4.1, we deduce that every accumulation point of $\{x^k\}$ is a stationary point of f.

Proposition 4.4. *Let $c > 0$ and $f \in \Gamma_c$. In addition, assume that f is a $\mathcal{C}^1$, coercive bounded from below function satisfying (8). Now, let $\{x^k\}$ be generated by (7). Then, for all λ such that*

$$0 < \lambda < \frac{2}{L_f(1-\theta)^2 + c\theta^2},$$

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

Proof. The proof is similar to [AB09, Proposition 1]. Using Proposition 4.1, observe that $\{f(x^k)\}$ is decreasing and, since f is bounded from below, $f(x^k)$ converges to f^*. By coercivity, there exists a subsequence $x^{\varphi(k)}$ which converges to x_∞. From (9), we obtain

$$\sum_{i=0}^{+\infty} |x^{k+1} - x^k|^2 \leqslant \frac{1}{\rho}(f(x^0) - \inf f) < +\infty,$$

where $\rho = \frac{1}{\lambda} - \frac{L_f(1-\theta)^2+c\theta^2}{2} > 0$, and then $|x^{k+1} - x^k| \to 0$. Therefore we deduce that $\nabla f(x_\infty) = 0$ by (7).

□

Remark 4.5. The $\mathcal{C}^1$ condition can be relaxed to a lower semicontinuity assumption using the *limiting subdifferential*, see [AB09].

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

Example 4.6 (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 4.7 (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. Perspectives on the proximal point method for weakly convex functions

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

5.1. **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)

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

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).

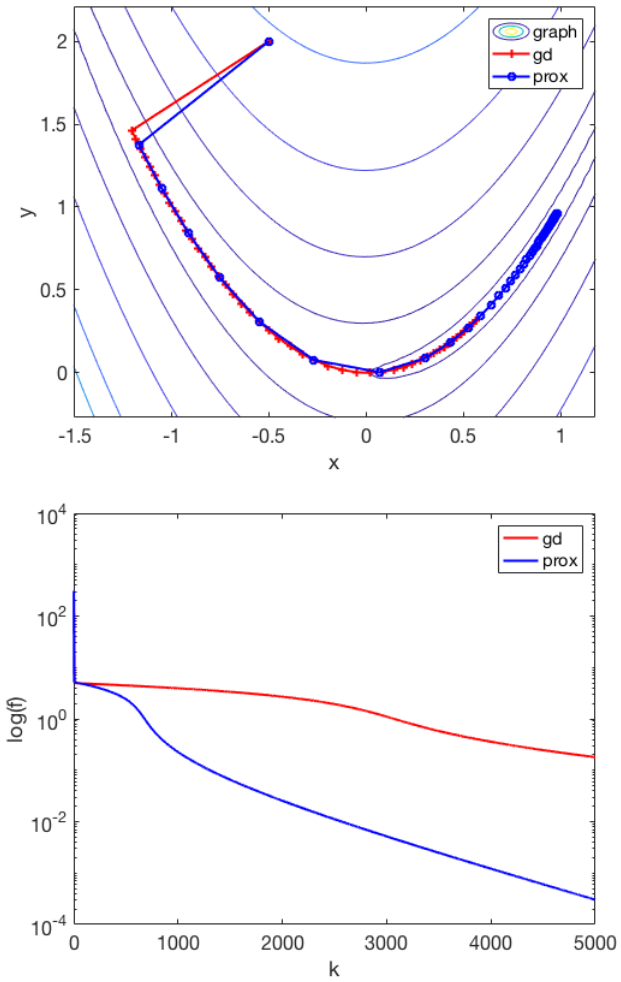

FIGURE 2. Top: Iterations of gradient descent and the proximal point method. The proximal point method gets closer to the global mininum with the same number of total gradient evaluations. Bottom: function values at each iteration of gradient descent and at each outer proximal point iteration for the 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.

5.2. **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 5.1. *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

[AA93] H. Attouch and D. Azé. Approximation and regularization of arbitrary functions in Hilbert spaces by the Lasry-Lions method. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 10(3):289–312, 1993. URL: https://doi.org/10.1016/S0294-1449(16)30214-1, doi:10.1016/S0294-1449(16)30214-1.

[AB09] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.*, 116(1-2, Ser. B):5–16, 2009. URL: https://doi.org/10.1007/s10107-007-0133-5, doi:10.1007/s10107-007-0133-5.

[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.

[AZ17] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1200–1205. ACM, 2017.

[B+15] S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[BC10] H. H. Bauschke and P. L. Combettes. The Baillon-Haddad theorem revisited. *J. Convex Anal.*, 17(3-4):781–787, 2010.

[BC11] H. H. Bauschke and P. 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.

[BCN16] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*, 2016.

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

[BH77] J.-B. Baillon and G. Haddad. Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones. *Israel J. Math.*, 26(2):137–150, 1977. URL: https://doi.org/10.1007/BF03007664, doi:10.1007/BF03007664.

[CDHS18] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM J. Optim.*, 28(2):1751–1772, 2018. URL: https://doi.org/10.1137/17M1114296, doi:10.1137/17M1114296.

[COO+18] P. Chaudhari, A. Oberman, S. Osher, S. Soatto, and G. Carlier. Deep relaxation: partial differential equations for optimizing deep neural networks. *Research in the Mathematical Sciences*, 5(3):30, Jun 2018. URL: https://doi.org/10.1007/s40687-018-0148-y, doi:10.1007/s40687-018-0148-y.

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

[FB15] N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pages 658–695, 2015.

[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.

[LL86] J.-M. Lasry and P.-L. Lions. A remark on regularization in Hilbert spaces. *Israel J. Math.*, 55(3):257–266, 1986. URL: https://doi.org/10.1007/BF02765025, doi:10.1007/BF02765025.

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

[LRP16] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[MP10] B. Merlet and M. Pierre. Convergence to equilibrium for the backward Euler scheme and applications. *Commun. Pure Appl. Anal.*, 9(3):685–702, 2010. URL: https://doi.org/10.3934/cpaa.2010.9.685, doi:10.3934/cpaa.2010.9.685.

[Nes13] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[Nit14] Atsushi Nitanda. Stochastic proximal gradient descent with acceleration techniques. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1574–1582. Curran Associates, Inc., 2014. URL: http://papers.nips.cc/paper/5610-stochastic-proximal-gradient-descent-with-acceleration-techniques.pdf.

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

[PLD+18] C. Paquette, H. Lin, D. Drusvyatskiy, J. Mairal, and Z. Harchaoui. Catalyst for gradient-based nonconvex optimization. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 613–622, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR. URL: http://proceedings.mlr.press/v84/paquette18a.html.

[Pol64] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[RVV14] Lorenzo Rosasco, Silvia Villa, and Bang Công Vũ. Convergence of stochastic proximal gradient algorithm. *arXiv preprint arXiv:1403.5074*, 2014.

[RW98] R. T. Rockafellar and R. 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.

[SBC14] W. Su, S. Boyd, and E. Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

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

[SRB11] Mark Schmidt, Nicolas L Roux, and Francis R Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in neural information processing systems*, pages 1458–1466, 2011.

[SRBD17] D. Scieur, V. Roulet, F. Bach, and A. D'Aspremont. Integration methods and accelerated optimization algorithms. *arXiv preprint arXiv:1702.06751*, 2017.

[YPOO18] Penghang Yin, Minh Pham, Adam Oberman, and Stanley Osher. Stochastic backward euler: An implicit gradient descent algorithm for k-means clustering. *Journal of Scientific Computing*, 77(2):1133–1146, 2018.

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

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

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