A new envelope function for nonsmooth DC optimization

Themelis, Andreas
Department of Electrical Engineering (ESAT-STADIUS) - KU Leuven

Hermans, Ben
MECO Research Team,Department of Mechanical Engineering

Patrinos, Panagiotis
Department of Electrical Engineering (ESAT-STADIUS) - KU Leuven

https://hdl.handle.net/2324/4399992

出版情報：2020 59th IEEE Conference on Decision and Control (CDC). 8, pp.4697-4702, 2021-05. IEEE
バージョン：

A new envelope function for nonsmooth DC optimization

Andreas Themelis,[1] Ben Hermans,[2] and Panagiotis Patrinos[1]

***Abstract*— Difference-of-convex (DC) optimization problems are shown to be equivalent to the minimization of a Lipschitz-differentiable "envelope". A gradient method on this surrogate function yields a novel (sub)gradient-free proximal algorithm which is inherently parallelizable and can handle fully nonsmooth formulations. Newton-type methods such as L-BFGS are directly applicable with a classical linesearch. Our analysis reveals a deep kinship between the novel DC envelope and the forward-backward envelope, the former being a smooth and convexity-preserving nonlinear reparametrization of the latter.**

I. Introduction

We consider difference-of-convex (DC) problems

$$\operatorname*{\mathbf{minimize}}_{s\in\mathbb{R}^p} \varphi(s) := g(s) - h(s), \tag{P}$$

where $g, h : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ are proper, convex, lsc functions (with the convention $\infty - \infty = \infty$). DC problems cover a very broad spectrum of applications; a well detailed theoretical and algorithmic analysis is presented in [24], where the nowadays textbook algorithm DCA is presented that interleaves subgradient evaluations $v \in \partial h(u)$, $u^+ \in \partial g^*(v)$, aiming at finding a *stationary* point u, that is, a point satisfying

$$\partial g(u) \cap \partial h(u) \neq \emptyset, \tag{1}$$

a relaxed version of the necessary condition $\partial h(u) \subseteq \partial g(u)$ [11]. As noted in [1], proximal *sub*gradient iterations are effective even in handling a nonsmooth nonconvex g and a nonsmooth concave $-h$. Alternative approaches use the identity $-f(x) = \mathbf{inf}_y \{f^*(y) - \langle x, y\rangle\}$ involving the convex conjugate f^* to include an additional convex function f as

$$\operatorname*{\mathbf{minimize}}_{x\in\mathbb{R}^n} g(x) - h(x) - f(x), \tag{2}$$

and then recast the problem as

$$\operatorname*{\mathbf{minimize}}_{x,y\in\mathbb{R}^n} \Phi(x, y) := \overbrace{g(x) + f^*(y)}^{G(x,y)} - \overbrace{\big(h(x) + \langle x, y\rangle\big)}^{H(x,y)}. \tag{3}$$

By adding and substracting suitably large quadratics, one can again obtain a decoupled DC formulation, showing that (P) is in fact as general as (2). When function h is smooth (differentiable with Lipschitz gradient), a cornerstone algorithm for the "convex+smooth" formulation (3) is forward-backward splitting (FBS), amounting to gradient evaluations of the smooth component $-h(s) - \langle s, t\rangle$ followed by proximal operations (possibly in parallel) on g and f^*.

Algorithm 1 Two-prox algorithm for the DC problem (P)

Select $\gamma > 0$ and $0 < \lambda < 2$, and starting from $s \in \mathbb{R}^p$, repeat

$$\begin{cases} u = \mathbf{prox}_{\gamma h}(s) \\ v = \mathbf{prox}_{\gamma g}(s) \end{cases}\!\Bigg]\ \text{(in parallel)} \qquad s^+ = s + \lambda(v - u) \tag{4}$$

Note: $s^+ = s - \lambda\gamma\nabla \mathrm{DCE}_\gamma^{g,h}(s)$, where $\mathrm{DCE}_\gamma^{g,h} = g^\gamma - h^\gamma$

Algorithm 2 Three-prox algorithm for the DC problem (2)

Select $0 < \gamma < 1 < \delta$, $0 < \lambda < 2(1-\gamma)$, and $0 < \mu < 2(1-\delta^{-1})$, and starting from $s, t \in \mathbb{R}^p$, repeat

$$\left\{\begin{array}{lll} u & = \mathbf{prox}_{\frac{\gamma\delta}{\delta-\gamma}h}\big(\tfrac{\delta s-\gamma t}{\delta-\gamma}\big) & \\ v & = \mathbf{prox}_{\gamma g}(s) & \text{(in parallel)} \\ z & = \mathbf{prox}_{\delta f}(t) & \\ s^+ & = s + \lambda(v - u) & \\ t^+ & = t + \mu(u - z) & \text{(in parallel)} \end{array}\right. \tag{5}$$

Note: $\binom{s^+}{t^+} = \binom{s}{t} - \begin{pmatrix}\gamma\lambda\mathrm{I} & \\ & \delta\mu\mathrm{I}\end{pmatrix}\nabla\Psi(s, t)$, where

$$\Psi(s, t) = g^\gamma(s) - f^\delta(t) - h^{\frac{\gamma\delta}{\delta-\gamma}}\big(\tfrac{\delta s-\gamma t}{\delta-\gamma}\big) + \tfrac{1}{2(\delta-\gamma)}\|s - t\|^2$$

A detailed overview on DC algorithms is beyond the scope of this paper; the interested reader is referred to the exhaustive surveys in [3,14,24] and references therein. Most related to our approach, [4] analyzes a Gauss-Seidel-type FBS in the spirit of the PALM algorithm [7], and [16] exploits the interpretation of FBS as a gradient-type algorithm on the *forward-backward envelope* (FBE) [17,22] to develop quasi-Newton methods for the nonsmooth and nonconvex problem (2). The gradient interpretation of splitting schemes originated in [20] with the proximal point algorithm and has recently been extended to several other schemes [10,17,18,23]. In this work we undertake a converse direction: first we design a smooth surrogate of the nonsmooth DC function in (P), and then derive a novel splitting algorithm from its gradient steps. Classical methods stemming from smooth minimization such as L-BFGS can conveniently be implemented, resulting in a method inherently robust against ill conditioning.

A. Contributions

a) Fully parallelizable splitting schemes: In this paper we propose the novel (sub)gradient-free proximal Algorithm 1 for the DC problem (P), and its fully parallelizable variant when applied to (2) synopsized in Algorithm 2 (see §II

[1] Andreas Themelis and Panagiotis Patrinos are with the Department of Electrical Engineering (ESAT-STADIUS) – KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium. This work was supported by the *Research Foundation Flanders (FWO)* research projects G086518N, G086318N, and G0A0920N; *Research Council KU Leuven* C1 project No. C14/18/068; *Fonds de la Recherche Scientifique — FNRS and the Fonds Wetenschappelijk Onderzoek — Vlaanderen* under EOS project no 30468160 (SeLMA).

[2] Ben Hermans is with the MECO Research Team, Department of Mechanical Engineering, KU Leuven, and Flanders Make - DMMS_M, Leuven, Belgium. His research benefits from KU Leuven-BOF PFV/10/002 Centre of Excellence: Optimization in Engineering (OPTEC), from project G0C4515N of the Research Foundation - Flanders (FWO - Flanders), from Flanders Make ICON: Avoidance of collisions and obstacles in narrow lanes, and from the KU Leuven Research project C14/15/067: B-spline based certificates of positivity with applications in engineering.

{andreas.themelis,ben.hermans2,panos.patrinos}@kuleuven.be

for the notation therein adopted). Our approach can be considered complementary to that in [16]. First, we propose a novel smooth DC envelope function (DCE) that shares minimizers and stationary points with the original nonsmooth DC function φ in (P), similarly to the FBE in [16]. Then, we show that a classical gradient descent on the DCE results in a novel (sub)gradient-free proximal algorithm that is particularly amenable to parallel implementations. In fact, even when specialized to problem (2) it involves operations on the three functions that can be done in parallel, differently from FBS-based approaches that prescribe serial (sub)gradient and proximal evaluations. Due to the complications of computing proximal steps in arbitrary metrics, this flexibility comes at the price of not being able to efficiently handle the composition of f in (2) with arbitrary linear operators, which is instead possible with FBS-based approaches such as [1,4,16].

b) Novel smooth DC reformulation: Thanks to the smooth gradient descent interpretation *it is possible to design classical linesearch strategies* to include directions stemming for instance from quasi-Newton methods, *without complicating the first-order algorithmic oracle*. In fact, differently from similar FBE-based quasi-Newton techniques in [16,17,22], no second-order derivatives are needed here and we actually allow for fully nonsmooth formulations. Moreover, being the difference of convex and Lipschitz-differentiable functions, the proposed envelope reformulation allows for the extension of the boosted DCA [2] to arbitrary DC problems.

c) A convexity-preserving nonlinear scaling of the FBE: When function h in (P) is smooth, we show that the DCE coincides with the FBE [17,22,26] after a nonlinear scaling. This change of variable overcomes some limitations of the FBE, such as preserving convexity when problem (P) is convex and being (Lipschitz) differentiable without additional requirements on function h.

B. Paper organization

The paper is organized as follows. Section II lists the adopted notational conventions and some known facts needed in the sequel. Section III introduces the DCE, a new envelope function for problem (P), and provides some of its basic properties and its connections with the FBE; the proofs of the lemmas in this section are deferred to the Appendix. Section IV shows that a classical gradient method on the DCE results in Algorithm 1, and establishes convergence results as a simple byproduct. Algorithm 2 is shown to be a *scaled* version of the parent Algorithm 1; for the sake of simplicity of presentation, some technicalities needed for this derivation are confined to this section. Section V shows the effect of L-BFGS acceleration on the proposed method on a sparse principal component analysis problem. Section VI concludes the paper.

II. Notation and known facts

The set of symmetric matrices in $\mathbb{R}^p$ is denoted as $\mathbf{sym}(\mathbb{R}^p)$; the subset of those which are positive definite is denoted as $\mathbf{sym}_{++}(\mathbb{R}^p)$. Any $M \in \mathbf{sym}_{++}(\mathbb{R}^p)$ induces the scalar product $(x, y) \mapsto x^\top My$ on $\mathbb{R}^p$, with corresponding norm $\|x\|_M = \sqrt{x^\top Mx}$. When $M = \mathrm{I}$, the identity matrix of suitable size, we will simply write $\|x\|$. id is the identity function on a suitable space. The subdifferential of a proper, lsc, convex function $f : \mathbb{R}^p \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ is

$$\partial f(x) := \{v \in \mathbb{R}^p \mid f(z) \geq f(x) + \langle v, z - x\rangle,\ \forall z\}.$$

The *effective domain* of f is $\mathbf{dom}\, f := \{x \in \mathbb{R}^p \mid f(x) < \infty\}$, while $f^*(y) := \mathbf{sup}_{x\in\mathbb{R}^p} \{\langle x, y\rangle - f(x)\}$ denotes the *Fenchel conjugate* of f, which is also proper, closed and convex. Properties of conjugate functions are well described for example in [5,13,19]. Among these we recall that

$$y \in \partial f(x) \Leftrightarrow \langle x, y\rangle = f(x) + f^*(y) \Leftrightarrow x \in \partial f^*(y). \tag{6}$$

The *proximal mapping* of f with stepsize $\gamma > 0$ is

$$\mathbf{prox}_{\gamma f}(x) := \operatorname*{\mathbf{arg\,min}}_{w\in\mathbb{R}^p} \Big\{f(w) + \tfrac{1}{2\gamma}\|w - x\|^2\Big\}, \tag{7}$$

while the value function of the above optimization problem defines the *Moreau envelope*

$$f^\gamma(x) := \inf_{w\in\mathbb{R}^p} \Big\{f(w) + \tfrac{1}{2\gamma}\|w - x\|^2\Big\}. \tag{8}$$

Properties of the Moreau envelope and the proximal mapping are well documented in the literature [5,8,9], some of which are summarized next.

Fact 1 (Proximal properties of convex functions)**.** Let f be proper, convex, and lsc. Then, for all $\gamma > 0$ and $s, s' \in \mathbb{R}^p$

(i) $\mathbf{prox}_{\gamma f}(s)$ is the unique point x such that $s \in x+\gamma\partial f(x)$.

(ii) $\|x-x'\|^2 \leq \langle x-x', s-s'\rangle \leq \|s-s'\|^2$, where $x = \mathbf{prox}_{\gamma f}(s)$ and $x' = \mathbf{prox}_{\gamma f}(s')$.

(iii) for $x = \mathbf{prox}_{\gamma f}(s)$ and $w \in \mathbb{R}^p$ it holds that $f^\gamma(s) \leq f(w) + \frac{1}{2\gamma}\|w - s\|^2 - \frac{1}{2\gamma}\|x - s\|^2$.

(iv) the Moreau envelope f^γ is convex and has $\frac{1}{\gamma}$-Lipschitz-continuous gradient $\nabla f^\gamma = \frac{1}{\gamma}(\mathrm{id} - \mathbf{prox}_{\gamma f})$.

III. The DC envelope

In this section we introduce a smooth DC reformulation of (P) that enables us to cast the nonsmooth and possibly extended-real valued DC problem into the unconstrained minimization of the DCE, a function with Lipschitz-continuous gradient. A classical gradient descent algorithm on this reformulation is then shown in Section IV to lead to the proposed Algorithms 1 and 2. In this sense, the DCE serves a similar role as the Moreau envelope for the proximal point algorithm [20], and the FBE and Douglas-Rachford envelopes for respectively FBS and Douglas-Rachford splitting (DRS) [18,22].

We begin by formalizing the DC setting of problem (P) dealt in the paper with the following list of requirements.

Assumption I. *The following hold in problem* (P)*:*

A1 $g, h : \mathbb{R}^p \to \overline{\mathbb{R}}$ *are proper, convex, and lsc;*

A2 φ *is lower bounded (with the convention $\infty - \infty = \infty$).*

Definition 2 (DC envelope)**.** *Suppose that Assumption I holds. Relative to problem* (P)*, the DC envelope (DCE) with stepsize $\gamma > 0$ is*

$$\mathrm{DCE}_\gamma^{g,h}(s) := g^\gamma(s) - h^\gamma(s).$$

Before showing that $\mathbf{DCE}_\gamma^{g,h}$ satisfies the anticipated smoothness properties and is tightly connected with the solutions of problem (P), we provide a characterization of stationary points in terms of the proximal mappings of the functions involved in the DC formulation. This will then be used to connect points that are stationary in the sense of (1) for (P) with points that are stationary in the classical sense for $\mathbf{DCE}_\gamma^{g,h}$.

Lemma 3 (Optimality conditions)**.** *Suppose that Assumption I holds. Then, any of the following is equivalent to stationarity at u in the sense of (1):*

(a) there exist $\gamma > 0$ and $s \in \mathbb{R}^p$ such that $u = \mathbf{prox}_{\gamma g}(s) = \mathbf{prox}_{\gamma h}(s)$;

(b) for any $\gamma > 0$ there exists $s \in \mathbb{R}^p$ such that $u = \mathbf{prox}_{\gamma g}(s) = \mathbf{prox}_{\gamma h}(s)$.

Lemma 4 (Basic properties of the DCE)**.** *Let Assumption I hold, and for notational conciseness given $s \in \mathbb{R}^p$ let $u := \mathbf{prox}_{\gamma h}(s)$ and $v := \mathbf{prox}_{\gamma g}(s)$. The following hold:*

(i) $\mathbf{DCE}_\gamma^{g,h}$ is $\frac{1}{\gamma}$-smooth with $\nabla \mathbf{DCE}_\gamma^{g,h} = \frac{1}{\gamma}(\mathbf{prox}_{\gamma h} - \mathbf{prox}_{\gamma g})$;

(ii) $\nabla \mathbf{DCE}_\gamma^{g,h}(s) = 0$ iff u is stationary (cf. (1));

(iii) $\varphi(v) + \frac{1}{2\gamma}\|v-u\|^2 \leq \mathbf{DCE}_\gamma^{g,h}(s) \leq \varphi(u) - \frac{1}{2\gamma}\|v-u\|^2$;

(iv) $\mathbf{arg\,min}\,\varphi = \mathbf{prox}_{\gamma h}(S_\star) = \mathbf{prox}_{\gamma g}(S_\star)$ and $\mathbf{inf}\,\varphi = \mathbf{inf}\,\mathbf{DCE}_\gamma^{g,h}$ for $S_\star = \mathbf{arg\,min}\,\mathbf{DCE}_\gamma^{g,h}$.

A. Connections with the forward-backward envelope

As will be detailed in Section IV-A, considering difference of hypoconvex functions in problem (P) leads to virtually no generalization. A more interesting scenario occurs when both h and $-h$ are hypoconvex functions, which amounts to h being L_h-smooth (differentiable with L_h-Lipschitz gradient). In order to elaborate on this property we first need to specialize Lemma 5 to smooth functions.

Lemma 5 (Proximal properties of smooth functions)**.** *Suppose that $f : \mathbb{R}^p \to \mathbb{R}$ is L_f-smooth. Then, there exist $\sigma_f, \sigma_{-f} \in [-L_f, L_f]$ with $L_f = \mathbf{max}\left\{|\sigma_f|, |\sigma_{-f}|\right\}$ such that $f - \frac{\sigma_f}{2}\|\cdot\|^2$ and $-f - \frac{\sigma_{-f}}{2}\|\cdot\|^2$ are convex functions. Then, for all $\gamma < 1/[\sigma_{-f}]_-$ (with the convention $1/0 = \infty$) and $s, s' \in \mathbb{R}^p$*

(i) $\mathbf{prox}_{-\gamma f}(s)$ is the unique u such that $s = u - \gamma\nabla f(u)$;

(ii) $\frac{1}{1-\gamma\sigma_f}\|s-s'\|^2 \leq \langle u-u', s-s'\rangle \leq \frac{1}{1+\gamma\sigma_{-f}}\|s-s'\|^2$, where $u = \mathbf{prox}_{-\gamma f}(s)$ and $u' = \mathbf{prox}_{-\gamma f}(s')$;

(iii) $(-f)^\gamma$ is differentiable with $\nabla(-f)^\gamma = \frac{\mathrm{id} - \mathbf{prox}_{-\gamma f}}{\gamma}$.

In the remainder of this subsection, suppose that h is smooth. Denoting $f := -h$, problem (P) reduces to

$$\underset{u \in \mathbb{R}^n}{\mathbf{minimize}}\, f(u) + g(u) = g(u) - (-f)(u) \tag{9}$$

with g convex and f smooth. A textbook algorithm for addressing such composite minimization problems is FBS, which interleaves proximal and gradient operations as

$$u^+ = \mathbf{prox}_{\gamma g}\left(u - \gamma\nabla f(u)\right). \tag{10}$$

By observing that $s = u - \gamma\nabla f(u)$ iff $u = \mathbf{prox}_{-\gamma f}(s)$ for $\gamma < 1/L_f$, one obtains the following curious connection among $\mathbf{DCE}_\gamma^{g,-f}$ and the forward-backward envelope [22, Eq. (2.3)]

$$\varphi_\gamma^{\text{FB}}(u) = f(u) - \tfrac{\gamma}{2}\|\nabla f(u)\|^2 + g^\gamma(u - \gamma\nabla f(u)). \tag{11}$$

Lemma 6. *In problem (9), suppose that f is L_f-smooth and g is proper, convex, and lsc. Then, for every $\gamma < 1/L_f$*

$$\varphi_\gamma^{\text{FB}} = \mathbf{DCE}_\gamma^{g,-f} \circ (\mathrm{id} - \gamma\nabla f) \text{ and } \mathbf{DCE}_\gamma^{g,-f} = \varphi_\gamma^{\text{FB}} \circ \mathbf{prox}_{-\gamma f}\,.$$

Moreover, $\mathbf{DCE}_\gamma^{g,-f}$ is $\frac{1-\gamma L_f}{\gamma}$-smooth, and if f is additionally convex then so is $\mathbf{DCE}_\gamma^{g,-f}$.

IV. The algorithm

Having assessed the $\frac{1}{\gamma}$-smoothness of $\mathbf{DCE}_\gamma^{g,h}$ and its connection with problem (P) in Lemma 4, the minimization of the nonsmooth DC function $\varphi = g - h$ can be carried out with a gradient descent with constant stepsize $\tau < 2\gamma$ on $\mathbf{DCE}_\gamma^{g,h}$. As shown in the next result, this is precisely Algorithm 1.

Theorem 7. *Suppose that Assumption I holds, and starting from $s^0 \in \mathbb{R}^n$ consider the iterates $(s^k, u^k, v^k)_{k\in\mathbb{N}}$ generated by Algorithm 1 with $\gamma > 0$ and $\lambda \in (0,2)$. Then, for every $k \in \mathbb{N}$ it holds that $s^{k+1} = s^k - \gamma\lambda\nabla\mathbf{DCE}_\gamma^{g,h}(s^k)$ and*

$$\mathbf{DCE}_\gamma^{g,h}(s^{k+1}) \leq \mathbf{DCE}_\gamma^{g,h}(s^k) - \tfrac{\lambda(2-\lambda)}{2\gamma}\|u^k - v^k\|^2. \tag{12}$$

In particular:

(i) the fixed-point residual vanishes with $\mathbf{min}_{i\leq k}\|u^i - v^i\| = o(1/\sqrt{k})$;

(ii) $(u^k)_{k\in\mathbb{N}}$ and $(v^k)_{k\in\mathbb{N}}$ have the same set of cluster points, Ω; when $(s^k)_{k\in\mathbb{N}}$ is bounded, every $u_\star \in \Omega$ is stationary for φ (in the sense of (1)) and φ is constant on Ω, the value being the (finite) limit of the sequences $(\mathbf{DCE}_\gamma^{g,h}(s^k))_{k\in\mathbb{N}}$ and $(\varphi(v^k))_{k\in\mathbb{N}}$;

(iii) if φ is coercive, then $(u^k, v^k)_{k\in\mathbb{N}}$ is bounded; if, additionally, $\mathbf{dom}\,h = \mathbb{R}^p$, then also $(s^k)_{k\in\mathbb{N}}$ is bounded.

Proof. That $s^{k+1} = s^k - \lambda\gamma\nabla\mathbf{DCE}_\gamma^{g,h}(s^k)$ follows from Lemma 4*(i)*. The proof is now standard, see e.g., [6]: $\frac{1}{\gamma}$-smoothness implies the upper bound

$$\begin{aligned}\mathbf{DCE}_\gamma^{g,h}(s^{k+1}) &\leq \mathbf{DCE}_\gamma^{g,h}(s^k) + \langle\nabla\mathbf{DCE}_\gamma^{g,h}(s^k), s^{k+1} - s^k\rangle \\ &\quad + \tfrac{1}{2\gamma}\|s^{k+1} - s^k\|^2 \\ &= \mathbf{DCE}_\gamma^{g,h}(s^k) - \tfrac{\lambda(2-\lambda)}{2\gamma}\|u^k - v^k\|^2,\end{aligned}$$

which is (12). We now show the numbered claims.

♠ 7*(i)* By telescoping (12) and using the fact that $\mathbf{inf}\,\mathbf{DCE}_\gamma^{g,h} = \mathbf{inf}\,\varphi > -\infty$ owing to Lemma 4*(iv)* and requirement I.A2, we obtain that the sequence of squared residuals $(\|u^k - v^k\|^2)_{k\in\mathbb{N}}$ has finite sum, hence the claim.

♠ 7*(ii)* That the sequences have the same cluster points follows from assertion 7*(i)*. Moreover, (12) and the lower boundedness of $\mathbf{DCE}_\gamma^{g,h}$ imply that the sequence $(\mathbf{DCE}_\gamma^{g,h}(s^k))_{k\in\mathbb{N}}$ monotonically decreases to a finite value $\varphi_\star$. Continuity of $\mathbf{DCE}_\gamma^{g,h}$ then implies that $\mathbf{DCE}_\gamma^{g,h}(s_\star) = \varphi_\star$ for every limit point $s_\star$ of $(s^k)_{k\in\mathbb{N}}$. If $(s^k)_{k\in\mathbb{N}}$ is bounded, then so are $(u^k)_{k\in\mathbb{N}}$ and $(v^k)_{k\in\mathbb{N}}$ owing to Lipschitz continuity of the proximal mappings. Moreover, for every k one has $s^k = u^k + \gamma\xi^k = v^k + \gamma\eta^k$ for some $\xi^k \in \partial h(u^k)$ and $\eta^k \in \partial g(v^k)$. Necessarily, the sequences of subgradients are bounded, and for any limit point $u_\star$ of $(u^k)_{k\in\mathbb{N}}$ we have that $u_\star = \mathbf{prox}_{\gamma h}(s_\star) = \mathbf{prox}_{\gamma g}(s_\star)$ for some cluster point $s_\star$ of $(s^k)_{k\in\mathbb{N}}$. By invoking Lemma 3 we conclude that $\varphi(u_\star) = \varphi_\star$.

♠ 7(*iii*) Boundedness of $(v^k)_{k\in\mathbb{N}}$ follows from the fact that $\varphi(v^k) \leq \mathbf{DCE}_{\gamma}^{g,h}(s^k) \leq \mathbf{DCE}_{\gamma}^{g,h}(s^0)$ for all k, owing to Lemma 4(*iii*) and (12); in turn, that of $(u^k)_{k\in\mathbb{N}}$ follows from assertion 7(*i*). If $\operatorname{\mathbf{dom}} h = \mathbb{R}^p$, since $(u^k)_{k\in\mathbb{N}}$ is bounded we may invoke [21, Ex. 9.14] to infer that h is L-Lipschitz continuous on a set containing $(u^k)_{k\in\mathbb{N}}$ for some $L \geq 0$, hence $\gamma^{-1}(s^k - u^k) \in \partial h(u^k) \subseteq \overline{\mathbf{B}}(0; L)$, and boundedness of $(s^k)_{k\in\mathbb{N}}$ follows. □

The remainder of the section is devoted to deriving Algorithm 2 as a special instance of Algorithm 1 applied to the problem reformulation (3). In order to formalize this derivation, we first need to address a minor technicality arising because of the nonconvexity of function H therein, which prevents a direct application of Algorithm 1 to the function decomposition $G-H$. Fortunately however, by simply adding a quadratic term to both G and H the desired DC formulation is obtained without actually changing the cost function Φ in problem (3). This simple issue is addressed next.

A. *Hypoconvex functions*

Clearly, adding a same quantity to both functions g and h leaves problem (P) unchanged. In particular, the convexity setting of Assumption I can also be achieved when g and h are *hypoconvex*, in the sense that they are convex up to adding a suitably large quadratic function. Recall that for $\tilde{f} = f + \frac{\mu}{2}\|\cdot\|^2$ it holds that $\mathbf{prox}_{\tilde{\gamma}\tilde{f}}(s) = \mathbf{prox}_{\gamma f}(\frac{s}{1+\tilde{\gamma}\mu})$ for $\gamma = \frac{\tilde{\gamma}}{1+\tilde{\gamma}\mu}$ [5, Prop. 24.8(i)]. Therefore, as long as there exists $\mu \in \mathbb{R}$ such that both $g + \frac{\mu}{2}\|\cdot\|^2$ and $h + \frac{\mu}{2}\|\cdot\|^2$ are convex functions, one can apply iterations (4) to the minimization of $g + \frac{\mu}{2}\|\cdot\|^2 - \left(h + \frac{\mu}{2}\|\cdot\|^2\right)$ to obtain

$$\begin{cases} u^k &= \mathbf{prox}_{\tilde{\gamma}h}(\tilde{s}^k) \\ v^k &= \mathbf{prox}_{\tilde{\gamma}g}(\tilde{s}^k) \\ \tilde{s}^{k+1} &= \tilde{s}^k + \tilde{\lambda}(v^k - u^k), \end{cases}$$

where $\tilde{\gamma} := \frac{\gamma}{1+\gamma\mu}$, $\tilde{s}^k := \frac{1}{1+\gamma\mu}s^k$, and $\tilde{\lambda} := \frac{1}{1+\gamma\mu}\lambda$. By observing that $\frac{\gamma}{1+\gamma\mu}$ ranges in $(0, {}^1\!/\!_\mu)$ for $\gamma \in (0, \infty)$ (with the convention ${}^1\!/\!_0 = \infty$), and that $\tilde{\lambda} = \lambda(1 - \tilde{\gamma}\mu)$, we obtain the following.

Remark 8 (*Hypo*convex functions)**.** If $\mu \in \mathbb{R}$ is such that both $g + \frac{\mu}{2}\|\cdot\|^2$ and $h + \frac{\mu}{2}\|\cdot\|^2$ are proper, lsc, convex functions, then all the numbered claims of Theorem 7 still hold provided that $0 < \lambda < 2(1 - \gamma\mu)$. □

As a final step towards the analysis of Algorithm 2, in the next subsection we motivate the presence of the two additional parameters δ and μ missing in Algorithm 1.

B. *Matrix stepsize and relaxation*

A substantial degree of flexibility can be introduced by replacing the quadratic term $\frac{1}{2\gamma}\|w - \cdot\|^2$ appearing in the definition (7) of the proximal mapping with the squared norm $\frac{1}{2}\|w - \cdot\|^2_{\Gamma^{-1}}$ induced by a matrix $\Gamma \in \mathbf{sym}_{++}(\mathbb{R}^p)$. The scalar stepsize γ is achieved by considering $\Gamma = \gamma\mathrm{I}$; in general, we may thus think of Γ as a matrix stepsize. Denoting

$$\mathbf{prox}_f^{\Gamma}(x) = \operatorname*{\mathbf{arg\,min}}_w \left\{f(w) + \tfrac{1}{2}\|w - x\|^2_{\Gamma^{-1}}\right\} \tag{13}$$

and

$$f^{\Gamma}(x) = \operatorname*{\mathbf{min}}_w \left\{f(w) + \tfrac{1}{2}\|w - x\|^2_{\Gamma^{-1}}\right\} \tag{14}$$

the corresponding Moreau envelope, as shown in [12, Thm. 4.1.4] we have that $\nabla f^{\Gamma} = \Gamma^{-1}(\mathrm{id} - \mathbf{prox}_f^{\Gamma})$ satisfies

$$0 \leq \langle \nabla f^{\Gamma}(s) - \nabla f^{\Gamma}(s'), s - s' \rangle \leq \|s - s'\|^2_{\Gamma^{-1}}.$$

Remark 9 (Matrix stepsizes and relaxations)**.** Under Assumption I, given a diagonal stepsize $\Gamma \in \mathbf{sym}_{++}(\mathbb{R}^p)$ and a diagonal relaxation $\Lambda \in \mathbf{sym}_{++}(\mathbb{R}^p)$ the iterations

$$\begin{cases} u^k &= \mathbf{prox}_h^{\Gamma}(s^k) \\ v^k &= \mathbf{prox}_g^{\Gamma}(s^k) \\ s^{k+1} &= s^k + \Lambda(v^k - u^k) \end{cases} \tag{15}$$

produce a sequence such that

$$\mathbf{DCE}_{\Gamma}^{g,h}(s^{k+1}) \leq \mathbf{DCE}_{\Gamma}^{g,h}(s^k) - \tfrac{1}{2}\|u^k - v^k\|^2_{(2\mathrm{I}-\Lambda)\Gamma^{-1}\Lambda}.$$

In particular, all the numbered claims of Theorem 7 still hold when $0 \prec \Lambda \prec 2\mathrm{I}$.[1] □

Notice that the optimality condition for the minimization problem (13) reads $0 \in \partial f(w) + \Gamma^{-1}(w - x)$. Equivalently,

$$w = \mathbf{prox}_f^{\Gamma}(x) \quad \Leftrightarrow \quad x \in w + \Gamma\partial f(w). \tag{16}$$

By using this fact, if a symmetric matrix M is such that the function $\tilde{f} = f + \frac{1}{2}\langle \cdot\,, M\,\cdot \rangle$ is convex, one can express its proximal map in terms of that of f in a similar fashion as the scalar case considered in §IV-A, namely,

$$\mathbf{prox}_{\tilde{f}}^{\tilde{\Gamma}} = \mathbf{prox}_f^{\Gamma} \circ (\mathrm{I} - \Gamma M)$$

with $\Gamma = (\tilde{\Gamma}^{-1} + M)^{-1}$.[2] It is thus possible to combine Remarks 8 and 9 as follows, where again for simplicity we restrict the case to diagonal matrices.

Remark 10. If a diagonal matrix M is such that both functions $g + \frac{1}{2}\langle \cdot\,, M\,\cdot \rangle$ and $h + \frac{1}{2}\langle \cdot\,, M\,\cdot \rangle$ are proper, lsc, convex, then the sequence produced by (15) satisfies all the numbered claims of Theorem 7 as long as $0 \prec \Lambda \prec 2(\mathrm{I} - \Gamma M)$. □

C. *A parallel three-prox splitting*

After the generalization documented in Remark 10 we are ready to address the formulation (2) and express Algorithm 2 as a "scaled" variant of Algorithm 1. We begin by rigorously framing the problem setting.

Assumption II. *In problem* (2)

A1 $f, g, h : \mathbb{R}^n \to \overline{\mathbb{R}}$ *are proper, lsc, and convex;*

A2 φ *is lower bounded (with the convention $\infty - \infty = \infty$).*

Theorem 11. *Let Assumption II hold, and starting from $(s^0, t^0) \in \mathbb{R}^n \times \mathbb{R}^n$ consider the iterates $(s^k, t^k, u^k, v^k, z^k)_{k\in\mathbb{N}}$ generated by Algorithm 2 with $0 < \gamma < 1 < \delta$, $0 < \lambda < 2(1 - \gamma)$ and $0 < \mu < 2(1 - \delta^{-1})$. Then, denoting*

$$\begin{aligned} \Psi(s, t) &= \mathbf{DCE}_{\Gamma}^{G,H}(s, t/\delta) \\ &= g^{\gamma}(s) - f^{\delta}(t) - h^{\frac{\gamma\delta}{\delta-\gamma}}\left(\tfrac{\delta s - \gamma t}{\delta - \gamma}\right) + \tfrac{1}{2(\delta-\gamma)}\|s - t\|^2, \end{aligned} \tag{17}$$

for every $k \in \mathbb{N}$ it holds that

$$\begin{pmatrix} s^{k+1} \\ t^{k+1} \end{pmatrix} = \begin{pmatrix} s^k \\ t^k \end{pmatrix} - \begin{pmatrix} \gamma\lambda\mathrm{I} & \\ & \delta\mu\mathrm{I} \end{pmatrix} \nabla\Psi(s^k, t^k). \tag{18}$$

[1] Although similar claims can be made for more general positive definite matrices, the diagonal requirement guarantees the symmetry of $(2\mathrm{I}-\Lambda)\Gamma^{-1}\Lambda$ and thus its positive definiteness for Λ as prescribed above.

[2] These expressions in terms of the new stepsize Γ use the matrix identities $(\mathrm{I} + \tilde{\Gamma}M)^{-1}\tilde{\Gamma} = (\tilde{\Gamma}^{-1} + M)^{-1}$ and $(\mathrm{I} + \tilde{\Gamma}M)^{-1} = \mathrm{I} - \Gamma M$ for $\Gamma = (\mathrm{I} + \tilde{\Gamma}M)^{-1}\tilde{\Gamma}$.

Moreover

(i) the fixed-point residual vanishes with $\mathbf{min}_{i\leq k}\left\|\left(\begin{smallmatrix}u^i-v^i\\u^i-z^i\end{smallmatrix}\right)\right\| = o(1/\sqrt{k})$*;*

(ii) $(u^k)_{k\in\mathbb{N}}$ $(v^k)_{k\in\mathbb{N}}$ *and* $(z^k)_{k\in\mathbb{N}}$ *have the same set of cluster points, be it* Ω*; when* $(s^k)_{k\in\mathbb{N}}$ *is bounded, every* $u_\star\in\Omega$ *satisfies the stationarity condition*

$$\emptyset\neq\partial g(u_\star)\cap(\partial f(u_\star)+\partial h(u_\star))\subseteq\partial g(u_\star)\cap\partial(f+h)(u_\star)$$

and φ *is constant on* Ω*, the value being the (finite) limit of the sequence* $(\varphi(v^k))_{k\in\mathbb{N}}$*;*

(iii) if φ *is coercive, then* $(u^k,v^k,z^k)_{k\in\mathbb{N}}$ *is bounded; if, additionally,* $\mathbf{dom}\,h=\mathbb{R}^p$*, then* $(s^k,t^k)_{k\in\mathbb{N}}$ *is also bounded.*

Proof. Let Φ, G and H be as in (3), and $\Gamma:=\left(\begin{smallmatrix}\gamma\mathrm{I}&\\&\delta^{-1}\mathrm{I}\end{smallmatrix}\right)$. Under Assumption II, G is convex and one can easily verify that

$$(v_s,v_t)=\mathbf{prox}_G^\Gamma(s,t)\Leftrightarrow\begin{cases}v_s=\mathbf{prox}_{\gamma g}(s)\\v_t=t-\delta^{-1}\,\mathbf{prox}_{\delta f}(\delta t)\end{cases}$$

in light of the Moreau identity $\mathbf{prox}_{f^*/\delta}(t)=t-\delta^{-1}\,\mathbf{prox}_{\delta f}(\delta t)$, see [5, Thm. 14.3(ii)]. Furthermore, from (16) we have

$$\begin{aligned}(u_s,u_t)=\mathbf{prox}_H^\Gamma(s,t)&\Leftrightarrow\begin{cases}s\in u_s+\gamma\partial h(u_s)+\gamma u_t\\t=u_t+u_s/\delta\end{cases}\\&\Leftrightarrow\begin{cases}\frac{s-\gamma t}{1-\gamma/\delta}\in u_s+\frac{\gamma}{1-\gamma/\delta}\partial h(u_s)\\u_t=t-u_s/\delta\end{cases}\\&\Leftrightarrow\begin{cases}u_s=\mathbf{prox}_{\frac{\gamma\delta}{\delta-\gamma}h}\left(\frac{\delta s-\gamma\delta t}{\delta-\gamma}\right)\\u_t=t-u_s/\delta.\end{cases}\end{aligned}$$

In particular,

$$\left(\begin{smallmatrix}s\\\delta t\end{smallmatrix}\right)+\left(\begin{smallmatrix}\lambda\mathrm{I}&\\&\delta\mu\mathrm{I}\end{smallmatrix}\right)\left(\mathbf{prox}_G^\Gamma\left(\begin{smallmatrix}s\\t\end{smallmatrix}\right)-\mathbf{prox}_H^\Gamma\left(\begin{smallmatrix}s\\t\end{smallmatrix}\right)\right)=\left(\begin{smallmatrix}s+\lambda(v_s-u_s)\\\delta t+\mu(u_s-\mathbf{prox}_{\delta f}(\delta t))\end{smallmatrix}\right).$$

Apparently, iterations (5) correspond to those in (15) with $\Lambda:=\left(\begin{smallmatrix}\lambda\mathrm{I}&\\&\mu\mathrm{I}\end{smallmatrix}\right)$ after the scaling $t\leftarrow t/\delta$. From these computations and using the fact that $(f^*)^{1/\delta}\circ\mathrm{id}/\delta=\frac{1}{2\delta}\|\cdot\|^2-f^\delta$, see [5, Thm. 14.3(i)], the expressions in (17) and (18) are obtained. Since function $H+\frac{1}{2}\|\cdot\|^2$ is convex — that is, the setting of Remark 10 is satisfied with $M=\mathrm{I}$ — and the condition $0\prec\Lambda\prec2(\mathrm{I}-\Gamma)$ holds when $\gamma,\delta,\lambda,\mu$ are as in the statement, it only remains to discuss boundedness and properties of the limit points as in assertion 11*(ii)*, as the rest of the proof follows from Theorem 7*(i)* and Remark 10. Since $\left(\begin{smallmatrix}v^k-u^k\\u^k-z^k\end{smallmatrix}\right)=\left(\begin{smallmatrix}s^{k+1}-s^k\\t^{k+1}-t^k\end{smallmatrix}\right)\to0$ the sequences $(u^k)_{k\in\mathbb{N}}$ $(v^k)_{k\in\mathbb{N}}$ and $(z^k)_{k\in\mathbb{N}}$ have the same cluster points, and all are bounded provided at least one is. Since

$$\Phi(x,y)\geq\inf_{y'}\Phi(x,y')=\varphi(x)$$

and $\Psi(v_s^k,v_t^k)\leq\mathrm{DCE}_\Gamma^{G,H}(s^k,t^k)\leq\mathrm{DCE}_\Gamma^{G,H}(s^0,t^0)$ (cf. Lemma 4*(iii)*), if φ is coercive then $(v_s^k=v^k)_{k\in\mathbb{N}}$ is bounded.

If $(s^k)_{k\in\mathbb{N}}$ is bounded, arguing as in the proof of Theorem 7*(ii)* we have that if $u^k\to u_\star$ as $k\in K$ for an infinite set of indices $K\subseteq\mathbb{N}$, necessarily also $v^k\to u_\star$ as $k\in K$, and $(s^k,t^k)\to(s_\star,t_\star)$ as $k\in K$ for some $s_\star,t_\star$ such that

$$\mathbf{prox}_{\frac{\gamma\delta}{\delta-\gamma}h}\left(\frac{\delta s_\star-\gamma t_\star}{\delta-\gamma}\right)=\mathbf{prox}_{\gamma g}(s_\star)=\mathbf{prox}_{\delta f}(t_\star).$$

We then conclude from Fact 1*(i)* that

$$\frac{\frac{\delta s_\star-\gamma t_\star}{\delta-\gamma}-u_\star}{\frac{\gamma\delta}{\delta-\gamma}}\in\partial h(u_\star),\quad\frac{s_\star-u_\star}{\gamma}\in\partial g(u_\star),\quad\frac{t_\star-u_\star}{\delta}\in\partial f(u_\star),$$

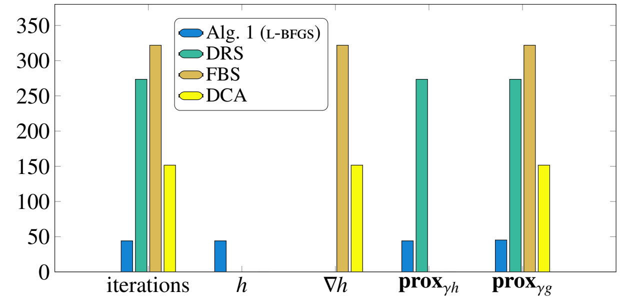

Fig. 1. Iteration comparison for random instances of (19).

which gives

$$\frac{s_\star-u_\star}{\gamma}\in\partial g(u_\star)\cap(\partial f(u_\star)+\partial h(u_\star)),$$

and the claimed stationarity condition follows from the inclusion $\partial f+\partial h\subseteq\partial(f+h)$, see [19, Thm. 23.8]. □

V. Simulations

We study the performance of Algorithm 1 applied to a sparse principal component analysis (SPCA) problem. Following [15, §2.1], an SPCA problem can be formulated as

$$\mathbf{minimize}\;-\tfrac{1}{2}s^\top\Sigma s+\kappa\|s\|_1\quad\mathbf{subject\ to}\;s\in\overline{\mathbf{B}}(0;1)\qquad(19)$$

with $\overline{\mathbf{B}}(0;1):=\{s\mid\|s\|\leq1\}$, $\Sigma=A^\top A$ the sample covariance matrix, and κ a sparsity inducing parameter. This problem can be identified as a DC problem of type (P) by denoting $g(s)=\kappa\|s\|_1+\delta_{\overline{\mathbf{B}}(0;1)}(s)$ and $h(s)=\frac{1}{2}s^\top\Sigma s$, where δ_C denotes the indicator function of a (nonempty closed convex) set C, namely $\delta_C(x)=0$ if $x\in C$ and ∞ otherwise. Then,

$$\mathbf{prox}_{\gamma h}(s)=(\mathrm{I}+\gamma\Sigma)^{-1}s,\quad\text{and}$$

$$\mathbf{prox}_{\gamma g}(s)=\frac{\mathbf{sgn}(s)\odot[|s|-\kappa\gamma\mathbf{1}]_+}{\mathbf{max}\{1,\|[|s|-\kappa\gamma\mathbf{1}]_+\|\}},$$

with $\odot$ the elementwise multiplication, $|\cdot|$ the elementwise absolute value, and $\mathbf{1}$ the $\mathbb{R}^n$-vector of all ones.

To (19) we applied FBS, DRS, DCA and Algorithm 1 (gradient descent on the DCE) with L-BFGS steps and Wolfe backtracking. Sparse random matrices $A\in\mathbb{R}^{20n\times n}$ with 10% nonzeros were generated for 11 values of n on a linear scale between 100 and 1000, with a sufficiently small κ [15, §2.1]. The mean number of iterations required by the solvers over these instances is reported in the first column of Figure 1. A stepsize $\gamma=0.9\lambda_{\max}^{-1}(\Sigma)$ was selected for Algorithm 1 and FBS, and $\gamma=0.45\lambda_{\max}^{-1}(\Sigma)$ for DRS consistently with the nonconvex analysis in [25]. Stepsize tuning might lead to a better performance of these algorithms but was not considered here. The termination criterion $\|\mathbf{prox}_{\gamma h}(s)-\mathbf{prox}_{\gamma g}(s)\|\leq10^{-6}$ was used for all solvers. Plain Algorithm 1 (without L-BFGS) always exceeded 1000 iterations.

Figure 1 also shows complexity in terms of function calls. Evaluating h and ∇h requires a matrix-vector product, which is $O(n^2)$ operations. By factorizing $\mathrm{I}+\gamma\Sigma$ once offline, each backsolve to compute $\mathbf{prox}_{\gamma h}$ also requires $O(n^2)$ operations. Finally, $\mathbf{prox}_{\gamma g}$ requires $2n$ comparisons and a norm-operation, and is clearly the least expensive operation. DCA and FBS need one ∇h and one $\mathbf{prox}_{\gamma g}$ (or similar) operation, and DRS one $\mathbf{prox}_{-\gamma h}$ (work equivalent to $\mathbf{prox}_{\gamma h}$) and one $\mathbf{prox}_{\gamma g}$ operation per iteration. Algorithm 1 requires one $\mathbf{prox}_{\gamma h}$ and one $\mathbf{prox}_{\gamma g}$ operation per iteration, and L-BFGS needs additionally one call to h, $\mathbf{prox}_{\gamma h}$ and $\mathbf{prox}_{\gamma g}$ per trial stepsize in the linesearch. However, as h and $\mathbf{prox}_{\gamma h}$ involve linear operations for this problem, only one evaluation is required in the linesearch. Furthermore, it was observed that a stepsize of 1 was almost always accepted. From Figure 1 it follows, therefore, that Algorithm 1 with L-BFGS requires less work to converge than the other methods, disregarding the one time factorization cost not present in FBS and DCA.

VI. Conclusions

By reshaping nonsmooth DC problems into the minimization of the smooth DC envelope function (DCE), a gradient method yields a new algorithm for DC programming. The algorithm is of the splitting type, involving (subgradient-free, proximal) operations on each component which, additionally, can be carried out in parallel at each iteration. The smooth reinterpretation naturally leads to the possibility of Newton-type acceleration techniques which can significantly affect the convergence speed. The DCE has also a theoretical appeal in its deep kinship with the forward-backward envelope, as it is shown to be a reparametrization with more favorable reguarity properties. We believe that this connection may be a valuable tool for relaxing assumptions in FBE-based algorithms, which is planned for future work.

Appendix

Proof of Lemma 3 *(Optimality conditions).* If u is stationary, for $\gamma > 0$ and $\xi \in \partial g(u) \cap \partial h(u) \neq \emptyset$ Fact 1*(i)* implies that $u = \mathbf{prox}_{\gamma g}(u + \gamma\xi) = \mathbf{prox}_{\gamma g}(u + \gamma\xi)$, proving 3*(b)* and thus 3*(a)*. Conversely, if 3*(a)* holds then Fact 1*(i)* again implies $\frac{s-u}{\gamma} \in \partial g(u)$ and $\frac{s-u}{\gamma} \in \partial h(u)$, proving that u is stationary. □

Proof of Lemma 4 *(Basic properties of the DCE).*

♠ 4*(i)* The expression of the gradient follows from Fact 1*(iv)*. The bounds in Fact 1*(ii)* imply that

$$\left|\langle \nabla \mathbf{DCE}_\gamma^{g,h}(s) - \nabla \mathbf{DCE}_\gamma^{g,h}(s'), s - s' \rangle\right| \leq \tfrac{1}{\gamma}\|s - s'\|^2, \tag{20}$$

proving that $\nabla \mathbf{DCE}_\gamma^{g,h}$ is γ^{-1}-Lipschitz continuous.

♠ 4*(ii)* Follows from assertion 4*(i)* and Lemma 3.

♠ 4*(iii)* Follows by applying the proximal inequalities of Fact 1*(iii)* with $w = u$ and $w = v$.

♠ 4*(iv)* Follows from assertion 4*(iii)*, Lemma 3, and the fact that global minimizers for φ are stationary. □

Proof of Lemma 5 *(Prox. properties of smooth functions).* The existence of $\sigma_{\pm f}$ comes from the fact that f is L_f-smooth iff $\frac{L_f}{2}\|\cdot\|^2 \pm f$ are convex, and that f is L_f-smooth iff so is $-f$. The proof now follows from Fact 1 applied to the convex function $\tilde{f} = -f - \frac{\sigma_{-f}}{2}\|\cdot\|^2$, by using the identity $\mathbf{prox}_{\gamma\tilde{f}} = \mathbf{prox}_{-\frac{\gamma}{1-\gamma\sigma_{-f}}f} \circ \frac{\mathrm{id}}{1-\gamma\sigma_{-f}}$ [5, Prop. 24.8(i)]. □

Proof of Lemma 6. Let $u \in \mathbb{R}^p$ and $\gamma \in (0, 1/L_f)$ be fixed, and for notational conciseness let $u = \mathbf{prox}_{-\gamma f}(s)$. Then, $s = u - \gamma\nabla f(u)$ and $(-f)^\gamma(s) = -f(u) + \frac{1}{2\gamma}\|u - s\|^2$, hence

$$\begin{aligned}\mathbf{DCE}_\gamma^{g,-f}(s) &= g^\gamma(u - \gamma\nabla f(u)) + f(u) - \tfrac{1}{2\gamma}\|u - s\|^2\\ &= f(u) - \tfrac{\gamma}{2}\|\nabla f(u)\|^2 + g^\gamma(u - \gamma\nabla f(u)),\end{aligned}$$

which is exactly $\varphi_\gamma^{\mathrm{FB}}(u)$, cf. (11). By using Lemma 5*(ii)* for $h = -f$, the bounds in (20) become

$$\tfrac{\sigma_f\|s-s'\|^2}{1-\gamma\sigma_f} \leq \langle \nabla \mathbf{DCE}_\gamma^{g,-f}(s) - \nabla \mathbf{DCE}_\gamma^{g,-f}(s'), s - s' \rangle \leq \tfrac{\gamma^{-1}\|s-s'\|^2}{1+\gamma\sigma_{-f}}.$$

Since $|\sigma_f|, |\sigma_{-f}| \leq L_f$, the claimed smoothness follows. Finally, if f is convex then σ_f is nonnegative and thus so is the lower bound above, proving convexity of $\mathbf{DCE}_\gamma^{g,-f}$. □

References

[1] N.T. An and N.M. Nam. Convergence analysis of a proximal point algorithm for minimizing differences of functions. *Optimization*, 66(1):129–147, 2017.

[2] F. Artacho, R. Fleming, and P.T. Vuong. Accelerating the DC algorithm for smooth functions. *Math. Prog.*, 169(1):95–118, 2018.

[3] M. Bačák and J. Borwein. On difference convexity of locally Lipschitz functions. *Optimization*, 60(8-9):961–978, 2011.

[4] S. Banert and R. Boţ. A general double-proximal gradient algorithm for DC programming. *Math. Prog.*, 178(1-2):301–326, 2019.

[5] H.H. Bauschke and P.L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2017.

[6] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2016.

[7] J. Bolte, S. Sabach, and M. Teboulle. Proximal Alternating Linearized Minimization for nonconvex and nonsmooth problems. *Math. Prog.*, 146(1–2):459–494, 2014.

[8] P.L. Combettes and JC. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, 2011.

[9] P.L. Combettes and V.R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[10] P. Giselsson and M. Fält. Envelope functions: Unifications and further properties. *JOTA*, 178(3):673–698, 2018.

[11] JB. Hiriart-Urruty. *From Convex Optimization to Nonconvex Optimization. Necessary and Sufficient Conditions for Global Optimality*, pages 219–239. Springer, 1989.

[12] JB. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms I: Fundamentals*, volume 305. Springer, 1993.

[13] JB. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2012.

[14] R. Horst and NV. Thoai. DC programming: overview. *JOTA*, 103(1):1–43, 1999.

[15] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *JMLR*, 11:517–553, 2010.

[16] T. Liu and TK. Pong. Further properties of the forward-backward envelope with applications to difference-of-convex programming. *Comput Optim Appl*, 67(3):489–520, Jul 2017.

[17] P. Patrinos and A. Bemporad. Proximal Newton methods for convex composite optimization. In *52nd IEEE CDC*, pages 2358–2363, 2013.

[18] P. Patrinos, L. Stella, and A. Bemporad. Douglas-Rachford splitting: Complexity estimates and accelerated variants. In *53rd IEEE CDC*, pages 4234–4239, 12 2014.

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

[20] R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Opt.*, 14(5):877–898, 1976.

[21] R.T. Rockafellar and R.J.-B. Wets. *Variational analysis*, volume 317. Springer, 2011.

[22] L. Stella, A. Themelis, and P. Patrinos. Forward-backward quasi-Newton methods for nonsmooth optimization problems. *Comput Optim Appl*, 67(3):443–487, Jul 2017.

[23] L. Stella, A. Themelis, and P. Patrinos. Newton-type alternating minimization algorithm for convex optimization. *IEEE TAC*, 2018.

[24] P.D. Tao and L.T.H. An. Convex analysis approach to DC programming: theory, algorithms and applications. *Acta mathematica vietnamica*, 22(1):289–355, 1997.

[25] A. Themelis and P. Patrinos. Douglas–Rachford splitting and ADMM for nonconvex optimization: Tight convergence results. *SIAM J. Opt.*, 30(1):149–181, 2020.

[26] A. Themelis, L. Stella, and P. Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM J. Opt.*, 28(3):2274–2303, 2018.