A new envelope function for nonsmooth DC optimization

Andreas Themelis,¹ Ben Hermans,² and Panagiotis Patrinos¹

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

\[ \min_{x \in \mathbb{R}^p} \varphi(s) := g(s) - h(s), \] (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 [23], where the nowadays textbook algorithm DCA is presented that inter-leaves subgradient evaluations \( v \in \partial h(u), u^* \in \partial g^*(v) \), aiming at finding a stationary point \( u \), that is, a point satisfying

\[ \partial g(u) \cap \partial h(u) \neq \emptyset, \] (1)

a relaxed version of the necessary condition \( \partial h(u) \subseteq \partial g(u) \) [11]. As noted in [1], proximal subgradient iterations are effective even in handling a nonsmooth nonconvex \( g \) and a non-smooth concave \( -h \). Alternative approaches use the identity

\[ -f(x) = \inf \left\{ f^*(y) - \langle x, y \rangle \right\} \]

involving the convex conjugate \( f^* \) to include an additional convex function \( f \) as

\[ \min_{x \in \mathbb{R}^p} g(x) - h(x) - f(x), \] (2)

and then recast the problem as

\[ \min_{x,y \in \mathbb{R}^p} \Phi(x,y) := \frac{G(x,y)}{2} + \frac{H(x,y)}{2} - (h(x) + \langle x, y \rangle). \] (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 \( f^* \).

A detailed overview on DC algorithms is beyond the scope of this paper; the interested reader is referred to the exhaustive surveys in [3,14,23] 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) [17,21] 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 [10,17,18,22]. 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

1) 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) is synopsized in Algorithm 2 (see §II

\[ \text{Algorithm 1 Two-prox algorithm for the DC problem (P)} \]

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

\[ \begin{cases} u = \text{prox}_{g/\gamma}(s) \\ v = \text{prox}_{\gamma}(s) \\ s^+ = s + \lambda(v - u) \end{cases} \] (in parallel) (4)

Note: \( s^+ = s - \lambda \nabla \text{env}^h(s) \), where \( \text{env}^h = g^* - h^* \)

\[ \text{Algorithm 2 Three-prox algorithm for the DC problem (2)} \]

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

\[ \begin{cases} u = \text{prox}_{\gamma}(s) \\ v = \text{prox}_{\gamma}(s) \\ z = \text{prox}_{\gamma}(t) \\ s^+ = s + \lambda(v - u) \\ t^+ = t + \mu(u - z) \end{cases} \] (in parallel) (5)

Note: \( \Psi(s,t) = g^*(s) - f^*(t) - h^* \left( \frac{\partial \Phi}{\partial s} (s,t) \right) + \frac{1}{2\gamma} \| s - t \|^2 \)

\[ \Psi(s,t) = g^*(s) - f^*(t) - h^* \left( \frac{\partial \Phi}{\partial s} (s,t) \right) + \frac{1}{2\gamma} \| s - t \|^2 \]
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 \( \varphi \) 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 [16,17,21], 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 [17,21,25] 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. 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 \( \text{sym}(\mathbb{R}^p) \); the subsets of those which are positive definite is denoted as \( \text{sym}_+(\mathbb{R}^p) \). Any \( M \in \text{sym}+(\mathbb{R}^p) \) induces the scalar product \( \langle x,y \rangle \mapsto x^T M y \) on \( \mathbb{R}^p \), with corresponding norm \( \|x\|_M = \sqrt{x^T M x} \). When \( M = I \), 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 \( \text{dom} f = \{ x \in \mathbb{R}^p \mid f(x) < \infty \} \), while \( f^*(y) := \sup_{x \in \mathbb{R}^p} \{ \langle x,y \rangle - f(x) \} \) 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

\[
y \in \partial f(x) \Leftrightarrow \langle x,y \rangle = f(x) + f^*(y) \Leftrightarrow x \in \partial f^*(y).
\]

The proximal mapping of \( f \) with stepsize \( \gamma > 0 \) is

\[
\text{prox}_f(x) := \arg \min_{w \in \mathbb{R}^p} \left\{ \frac{1}{2\gamma} \| w - x \|^2 + f(w) \right\},
\]

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

\[
f^\gamma(x) := \inf_{w \in \mathbb{R}^p} \left\{ \frac{1}{2\gamma} \| w - x \|^2 + f(w) \right\}.
\]

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) \( \text{prox}_f(s) \) is the unique point \( x \) such that \( s = x + \gamma \partial f(x) \).

(ii) \( \| x-x' \|^2 \leq \langle x-x', s-s' \rangle \leq \| s-s' \|^2 \), where \( x = \text{prox}_f(s) \) and \( x' = \text{prox}_f(s') \).

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

(iv) the Moreau envelope \( f^\gamma \) is convex and has \( \frac{1}{\gamma} \) -Lipschitz-continuous gradient \( \nabla f^\gamma = \frac{1}{\gamma}(\text{id} - \text{prox}_f) \).

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 will then be 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 envelope respectively for FBS and the Douglas-Rachford splitting (DRS) [18,21].

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

Assumption 1. The following hold in problem (P): \( \lambda_1 g,h: \mathbb{R}^p \to \overline{\mathbb{R}} \) are proper, convex, and lsc;

\( \lambda_2 \varphi \) is lower bounded (with the convention \( \infty - \infty = \infty \)).

Definition 2 (DC envelope). Suppose that Assumption 1 holds. Relative to problem (P), the DC envelope (DCE) with stepsize \( \gamma > 0 \) is

\[
\text{en}_f^{\varphi,h}(s) := g^\gamma(s) - h^\gamma(s).
\]

Before showing that the DCE \( \text{en}_f^{\varphi,h} \) satisfies the anticipated smoothness properties and is tightly connected with
solutions of problem \((P)\), we provide a simple 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 \(\text{env}^\gamma_{\phi} \).

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

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

(b) for all \(\gamma > 0\) there exists \(s \in \mathbb{R}^p\) such that \(u = \text{prox}_{\gamma \phi}(s)\).

**Proof.** If \(u\) is stationary, then for every \(\gamma > 0\) and \(\xi \in \partial g(u) \cap \partial h(u) \neq 0\) it follows from Fact 1(i) that \(u = \text{prox}_{\gamma \phi}(s)\) for \(s = u + \gamma \xi\), proving 3(b) and thus 3(a). Conversely, if 3(a) holds then Fact 1(i) again implies \(s = u + \gamma \xi\) for all \(\gamma > 0\), proving that \(u\) is stationary. \(\square\)

**Lemma 4** (Basic properties of the DCE). Let Assumption 1 hold, and for notational conciseness given \(s \in \mathbb{R}^p\) let \(u := \text{prox}_{\gamma \phi}(s)\) and \(v := \text{prox}_{\gamma \phi}(s)\). The following hold:

(i) \(\text{env}_{\phi}^\gamma(h) \) is \(\frac{1}{\gamma}\)-smooth with \(\nabla \text{env}_{\phi}^\gamma(h) = \frac{1}{\gamma}(\text{prox}_{\gamma \phi}(h) - \text{prox}_{\gamma \phi}(g))\);

(ii) \(\nabla \text{env}_{\phi}^\gamma(s) = 0\) iff \(u\) is stationary (cf. \((1)\));

(iii) \(\varphi(v) + \frac{1}{\gamma^2}\|v - u\|^2 \leq \text{env}_{\phi}^\gamma(v) \leq \varphi(u) - \frac{1}{2\gamma^2}\|v - u\|^2\);

(iv) \(\inf \varphi = \text{prox}_{\gamma \phi}(S_s) = \text{prox}_{\gamma \phi}(S_s)\). \(\square\)

**Proof.**

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

\[
\left\langle \nabla \text{env}_{\phi}^\gamma(s) - \nabla \text{env}_{\phi}^\gamma(s'), s - s' \right\rangle \leq \frac{1}{\gamma}\|s - s'\|^2,
\]

proving that \(\nabla \text{env}_{\phi}^\gamma(h)\) is \(\gamma^{-1}\)-Lipschitz continuous.

\(\bullet\) 4(ii) Follows from assertion 4(ii) and Lemma 3.

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

\(\bullet\) 4(iv) Follows from assertion 4(iii), Lemma 3, and the fact that global minimizers for \(\varphi\) are stationary. \(\square\)

A. Connections with the forward-backward envelope

As it will be detailed in Section IV-A, considering difference of hypocoercive functions in problem \((P)\) leads to virtually no generalization. A more interesting scenario occurs when both \(h\) and \(h\) are hypocoercive functions, which amounts to \(h\) being \(L_0\)-smooth (differentiable with \(L_0\)-Lipschitz gradient). In order to elaborate 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 \(L_1\)-smooth. Then, there exist \(\sigma_f, \sigma_{-f} \in [-L_f, L_f]\) with \(L_f = \max \left\{\|\sigma_f\|, \|\sigma_{-f}\|\right\}\) such that \(f - \frac{\sigma_f}{2} \|\cdot\|^2\) and \(-f - \frac{\sigma_{-f}}{2} \|\cdot\|^2\) are convex functions. Then, for all \(\gamma < \frac{1}{\|\sigma_f\|}\) (with the convention \(\|h\| = \infty\)) and \(s, s' \in \mathbb{R}^p\):

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

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

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

**Proof.** The claim on 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 functions, and that \(f\) is \(L_f\)-smooth iff so is \(-f\). All other claims then follow from Fact 1 applied to the convex function \(f = -f - \frac{\sigma_{-f}}{2} \|\cdot\|^2\), in light of the identity \(\text{prox}_{\gamma \Phi_f} = \text{prox}_{\gamma \Phi_{-\Phi_f}}\) \(\circ \frac{\text{id}}{\gamma}\) [5, Prop. 24.8(i)]. \(\square\)

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

\[
\begin{align*}
\text{minimize} \quad & f(u) + g(u) = g(u) - (-f)(u) \\
\text{subject to} & \quad u \in \text{convex and smooth}.
\end{align*}
\]

A textbook algorithm for addressing such composite minimization problems is FBS, which interleaves proximal and gradient operations as

\[
u_t = \text{proj}_{\gamma \phi}(u_t - \gamma \nabla f(u_t)).
\]

By observing that \(s = u - \gamma \nabla f(u)\) iff \(u = \text{proj}_{\gamma \phi}(s)\) for \(\gamma < \frac{1}{\|\sigma_f\|}\), one obtains the following curious connection among \(\text{env}_{\phi}^\gamma\) and the forward-backward envelope [21, Eq. (2.3)]

\[
\varphi^\gamma(u) = f(u) - \frac{1}{2\gamma}\|\nabla f(u)\|^2 + g^\gamma(u - \gamma \nabla f(u)).
\]

Moreover, \(\text{env}_{\phi}^\gamma\) is \(\frac{1}{\gamma^2}\)-smooth, and if \(f\) is additionally convex then so is \(\text{env}_{\phi}^\gamma\).

**Proof.** Let \(u \in \mathbb{R}^p\) and \(\gamma \in (0, \frac{1}{\|\sigma_f\|})\) be fixed, and for notational conciseness let \(u = \text{proj}_{\gamma \phi}(s)\). Then, \(s = u - \gamma \nabla f(u)\) and \((-f)^\gamma(s) = -f(u) + \|u - s\|^2\), hence

\[
\text{env}_{\phi}^\gamma(s) = g^\gamma(u - \gamma \nabla f(u)) + f(u) - \frac{1}{2\gamma^2}\|u - s\|^2 = f(u) - \frac{1}{2\gamma^2}\|\nabla f(u)\|^2 + g^\gamma(u - \gamma \nabla f(u)),
\]

which is exactly \(\varphi^\gamma(u)\), cf. (12). By using Lemma 5(iii) for \(h = -f\), the bounds in (9) become

\[
\frac{\sigma_f}{2\gamma} \leq \langle \nabla \text{env}_{\phi}^\gamma(s) - \nabla \text{env}_{\phi}^\gamma(s'), s - s' \rangle \leq \frac{\gamma^2\|s - s'\|^2}{\gamma^2 - \sigma_f^2}.
\]

Since \(\text{env}_{\phi}^\gamma\) is smooth, this is exactly \(\varphi^\gamma(s)\), proving convexity of \(\text{env}_{\phi}^\gamma\). \(\square\)

**IV. THE ALGORITHM**

Having assessed the \(\frac{1}{\gamma}\)-smoothness of \(\text{env}_{\phi}^\gamma\) 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 \(\text{env}_{\phi}^\gamma\). As shown in the next result, this is precisely Algorithm 1.

**Theorem 7.** Suppose that Assumption 1 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 \(s^{k+1} = s^k - \gamma \lambda \nabla \text{env}_{\phi}^\gamma(s^k)\) and

\[
\text{env}_{\phi}^\gamma(s^{k+1}) \leq \text{env}_{\phi}^\gamma(s^k) - \frac{\lambda^2 - 2\lambda}{2\gamma}\|u^k - v^k\|^2.
\]

In particular:
Proof. That \( s^{k+1} = s^k + \alpha y \) follows from Lemma 4(ii). The proof is now standard, see e.g., [6]: \( \frac{1}{\gamma} \)-smoothness implies the upper bound

\[
\text{env}^y(s^{k+1}) \leq \text{env}^y(s^k) + \langle \nabla \text{env}^y(s^k), s^{k+1} - s^k \rangle + \frac{1}{2\gamma} \|s^{k+1} - s^k\|^2
\]

which is (13). We now show the numbered claims.

- **7(i)**) By telescoping (13) and using the fact that \( \inf \text{env}^y(s^k) = \inf \varphi \) owing to Lipschitz continuity of the proximal mappings.
- **7(ii)**) That the sequences have same cluster points follows from assertion 7(i). Moreover, (13) and the lower boundedness of \( \text{env}^y(s) \) imply that the sequence \( \text{env}^y(s^k) \) monotonically decreases to a finite value, be it \( \varphi_\star \). Continuity of \( \text{env}^y(s) \) then implies that \( \text{env}^y(s^k) = \varphi_\star \) for every limit point \( s_\star \) of \( (s^k) \). If \( (s^k) \) is bounded, then so are \( (u^k) \) and \( (v^k) \) and \( \varphi_\star \) is continuous. Moreover, for every \( k \) one has \( s^k = u^k + \gamma v^k = v^k + (\gamma - 1)u^k \), hence the claim.

**7(iii)**) Boundedness of \( (s^k) \) follows from the fact that \( \text{env}^y(s^k) \leq \text{env}^y(s^k) \) for all \( k \), owing to (13). In turn, boundedness of \( (u^k) \) and \( (v^k) \) follows from Lipschitz continuity of the proximal mappings.

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 \( \Phi \) in problem (3). This simple issue is addressed next.

A. Strongly and hypoconvex functions

Clearly, adding a same quantity to both functions \( g \) and \( h \) leaves problem (P) unchanged. In particular, the convexity setting of Assumption 1 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 \( f \) being \#-convex it holds that \( \text{prox}_{\gamma f}(s) = \text{prox}_{\gamma f}(s + \frac{1}{\gamma} \|s\|^2) \) for

\[
\gamma = \frac{\gamma}{1+\gamma} [5, \text{Prop. 24.8(i)}] \]. Therefore, as long as there exists \( \mu \in \mathbb{R} \) such that both \( g + \frac{\mu}{2} \|s\|^2 \) and \( h + \frac{\mu}{2} \|s\|^2 \) are convex functions, one can apply iterations (4) to the minimization of \( g + \frac{\mu}{2} \|s\|^2 - (h + \frac{\mu}{2} \|s\|^2) \) to obtain

\[
\begin{align*}
\hat{u}^k &= \text{prox}_{\gamma b \hat{\theta}}(\hat{s}^k) \\
\hat{v}^k &= \text{prox}_{\gamma b \hat{\theta}}(\hat{s}^k) \\
\hat{s}^{k+1} &= \hat{s}^k + \bar{\lambda} (\hat{v}^k - \hat{u}^k)
\end{align*}
\]

where \( \bar{\lambda} := \frac{\gamma}{1+\gamma} \), and \( \lambda \) is defined as above. By observing that \( \frac{\gamma}{1+\gamma} \) ranges in \( (0,1) \) for \( \gamma \in (0,\infty) \) (with the convention \( 1/0 = \infty \)), and that \( \lambda = \lambda(1 - \gamma \mu) \), we obtain the following.

***Remark 8*** (Strongly convex and hypoconvex functions). If \( \mu \in \mathbb{R} \) is such that both \( g + \frac{\mu}{2} \|s\|^2 \) and \( h + \frac{\mu}{2} \|s\|^2 \) are 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 \( \delta \) and \( \mu_\star \) in Algorithm 1.

B. Matrix stepsize and relaxation

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

\[
\text{prox}_{\Gamma f}(x) = \text{arg min}_w \{ f(w) + \frac{1}{2} \|w - x\|^2 \}
\]

and

\[
f^{\Gamma f}(x) = \text{min}_w \{ f(w) + \frac{1}{2} \|w - x\|^2 \}
\]

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

\[
0 \leq \langle \nabla f^{\Gamma f}(s), s - s\star \rangle \leq \|s - s\star\|^2.
\]

**Remark 9** (Matrix stepsizes and relaxations). Under Assumption 1, given a diagonal stepsize \( \Gamma \in \text{sym}_+(\mathbb{R}^p) \) and a diagonal relaxation \( \Lambda \in \text{sym}_+(\mathbb{R}^p) \) the iterations

\[
\begin{align*}
\hat{u}^k &= \text{prox}_{\gamma f\Gamma}(\hat{s}^k) \\
\hat{v}^k &= \text{prox}_{\gamma f\Gamma}(\hat{s}^k) \\
\hat{s}^{k+1} &= \hat{s}^k + \lambda (\hat{v}^k - \hat{u}^k)
\end{align*}
\]

produce a sequence such that

\[
\text{env}_{\gamma f}(s^{k+1}) \leq \text{env}_{\gamma f}(s^k) - \frac{1}{2} \|u^k - v^k\|^2 \|2I - \Lambda I\|^2 \Lambda.
\]

In particular, all the numbered claims of Theorem 7 still hold when \( 0 < \Lambda < 2I \).}

Notice that the optimality condition for minimization problem (14) reads \( 0 \in \partial f(w) + \Gamma^{\top}(w - x) \). Equivalently,

\[
w = \text{prox}_{\gamma f\Gamma}(x) \iff x \in \arg min f(w) + \Gamma w - x.
\]

By using this fact, if a symmetric matrix \( M \) is such that the function \( f = f + \frac{1}{2} \left( \cdot, M \cdot \right) \) is convex, one can express its

\[1\text{ Although similar claims can be made for more general positive definite matrices, the diagonal requirement guarantees the symmetry of } (2I - \Lambda I)\Lambda^{-1}\Lambda \text{ and thus its positive definiteness for } \Lambda \text{ as prescribed above.}
proximal map in terms of that of $f$ in a similar fashion as the scalar case considered in §IV-A, namely,

$$
\text{prox}_f^\Gamma = \text{prox}^\Gamma \circ (I - \Gamma M)
$$

with $\Gamma = (I^{-1} + M)^{-1}$. 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 convex, then the sequence produced by (16) satisfies all the numbered claims of Theorem 7 as long as $0 < \Lambda < 2(1 - \Gamma M)$. □

**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 $\varphi$ is lower bounded.

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

$$
\Psi(s, t) = \text{env}_G^{\Gamma}(s, t/\delta) = g(s) - f^\delta(t) - h\frac{\delta y - \varphi}{\delta y} + \frac{1}{2 \delta y} ||s - t||^2,
$$

(18)

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

$$
\left(\frac{s^k}{t^k}\right) = \left(\frac{s^k}{t^k}\right) - \left(\gamma \delta^{-1} \frac{\partial \varphi}{\partial y}\right) \nabla \Psi(s^k, t^k).
$$

(19)

Moreover

(i) the fixed-point residual vanishes with $\min_{\mathbb{R}^n} ||(u^k - v^k)|| = o(1/\sqrt{k})$;

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

$$
\theta \neq \partial g(u^*) \cap \partial f(u^*) \subseteq \partial g(u^*) \cap \partial (f + h)(u^*)
$$

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

(iii) if $\varphi$ is coercive, then $(s^k, t^k, u^k, v^k, z^k)_{k \in \mathbb{N}}$ is bounded.

**Proof.** Let $\Phi$, $G$, and $H$ be as in (3), and observe that

$$
\Phi(x, y) \geq \inf_x \Phi(x, y') = \varphi(x).
$$

In particular, if $\varphi$ is coercive then necessarily so is $\Phi$. Let $\Gamma = \left(\begin{smallmatrix} \delta & 1 \\ -1 & \delta \end{smallmatrix}\right)$. Under Assumption II, function $G$ is convex and one can easily verify that

$$
(v_s, v_t) = \text{prox}_G^\Gamma(s, t) \iff \begin{cases} 
 v_s = \text{prox}_f^\Gamma(s) \\
 v_t = t - \frac{1}{\delta} \text{prox}_{g^\delta}(\delta t)
\end{cases}
$$

in light of the Moreau identity $\text{prox}_{g^\delta}(s) = t - \delta^{-1} \text{prox}_{g^\delta}(\delta t)$, see [5, Thm. 14.3(ii)]. Furthermore, from (17) we have

$$
(u_s, u_t) = \text{prox}_H^\Gamma(s, t) \iff \begin{cases} 
 s \in u_t + \gamma \partial h(u_t) + \gamma u_t \\
 t = u_t + \delta u_s
\end{cases}
$$

$$
\Rightarrow \begin{cases} 
 s - \gamma \frac{1}{\delta y} \in u_t + \gamma \frac{1}{\delta y} \partial h(u_t) \\
 u_t = t - \delta u_s
\end{cases}
$$

$$
\Rightarrow \begin{cases} 
 u_s = \text{prox}_{\frac{\delta y - \varphi}{\delta y}} \left(\frac{1}{\delta y} \right) \\
 u_t = t - \delta u_s
\end{cases}
$$

In particular,

$$
\left(\frac{s^k}{t^k}\right) + \left(\gamma \delta^{-1} \frac{\partial \varphi}{\partial y}\right) \left(\frac{s^k}{t^k}\right) = \left(\frac{s^k}{t^k}\right) = \left(\frac{s^k}{t^k}\right).
$$

Therefore, (5) corresponds to those in (16) with $\Lambda = \left(\frac{1}{\delta y}\right)$ after the scaling $t \mapsto \delta y$. From these computations and using the fact that $(f^\delta)^{1/\delta} \circ id/\delta = \frac{1}{\delta y} ||\cdot||^2 - f^\delta$, see [5, Thm. 14.3(i)], the expressions in (18) and (19) are obtained. Since function $H + \frac{1}{2 \delta y} ||\cdot||^2$ is convex — that is, the setting of Remark 10 is satisfied with $M = I$ — and the condition $0 < \Lambda < 2(I - \Gamma)$ holds when $\gamma, \delta, \mu, \lambda$ are as in the state, it only remains to show that the limit points satisfy the stationarity condition of assertion 11(ii), as the rest of the proof follows from Theorem 7(i) and Remark 10. To this end, since $(\tilde{s}^0, \tilde{t}^0) = \left(\frac{s^0}{t^0}\right) \to 0$ 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. If $(s^k)_{k \in \mathbb{N}}$ is bounded, arguing as in the proof of Theorem 7(ii) we have that if $u^k \to u^*$ as $k \to \infty$ for an infinite set of indices $\mathbb{K} \subseteq \mathbb{N}$, necessarily also $v^k \to v^*$ as $k \to \infty$, and $(s^k, t^k) \to (s^*, t^*)$ as $k \to \infty$ for some $s^*, t^*$ such that

$$
\text{prox}_{\frac{\delta y - \varphi}{\delta y}} \left(\frac{s^k}{t^k}\right) = \text{prox}_{\frac{\delta y - \varphi}{\delta y}} \left(\frac{s^k}{t^k}\right) = \text{prox}_{\frac{\delta y - \varphi}{\delta y}} \left(\frac{s^k}{t^k}\right).
$$

We then conclude from Fact 1(i) that

$$
\frac{s^k}{t^k} - u^* \in \partial g(u^*),
$$

$$
\frac{s^k}{t^k} - u^* \in \partial g(u^*),
$$

$$
\frac{s^k}{t^k} - u^* \in \partial g(u^*),
$$

and the claimed stationarity condition follows from the inclusion $\partial f + \partial h \subseteq \partial (f^\delta)$, 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

$$
\text{minimize} \ -\frac{1}{2} x^T \Sigma x + k ||s||_1 \quad \text{subject to} \ s \in \overline{\mathbb{R}}(0; 1)
$$

with $\overline{\mathbb{R}}(0; 1) := \{s \mid ||s||_1 \leq 1, \Sigma = A^T A \text{ the sample covariance matrix, and } k \text{ a sparsity inducing parameter}$. This problem can be identified as a DC problem of type (P) by denoting $g(s) = k ||s||_1 + \delta \overline{\mathbb{R}}(0; 1)(s)$ and $h(s) = \frac{1}{2} x^T \Sigma x$, where $\delta \overline{\mathbb{R}}$ denotes the indicator function of a nonempty closed convex set $C$, namely $\delta \overline{\mathbb{R}}(x) = 0$ if $x \in C$ and $\infty$ otherwise. Then,

$$
\text{prox}_{\gamma \delta \overline{\mathbb{R}}}(s) = (1 + \gamma \Sigma)^{-1} s, \text{ and }
$$

$$
\text{prox}_{\gamma k} s = \begin{cases} 
 \text{sgn}(s) \|s - k\|_1 \\
 \max \{1, \frac{1}{\|s - k\|_1} \}
\end{cases}
$$

with $\circ$ the elementwise multiplication, $\|\cdot\|$ the elementwise absolute value, and $\mathbf{1}$ the $\mathbb{R}^n$-vector of all ones.
To (20) we applied FBS, DRS, DCA and Algorithm 1 (gradient descent on the DCE) with L-BFGS steps and Wolfe backtracking. Sparse random matrices $A \in \mathbb{R}^{200 \times n}$ with 10% nonzeros were generated for 11 values of $n$ on a linear scale between 100 and 1000, with a sufficiently small $\kappa$ [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 $\gamma = 0.94 \min \{\Sigma\}$ was selected for Algorithm 1 and FBS, and $\gamma = 0.45 \min \{\Sigma\}$ for DRS consistently with the non-convex analysis in [24]. Stepsize tuning might lead to a better performance of these algorithms but was not considered here. The termination criterion $\|\text{prox}_{g}(s) - \text{prox}_{g}(\tau)\| \leq 10^{-6}$ was used for all solvers. Plain Algorithm 1 (without L-BFGS) always exceeded 1000 iterations.

Figure 1 also lists the complexity in terms of function calls. Evaluating $h$ and $\nabla h$ requires a matrix-vector product, which is $O(n^2)$ operations. By factorizing $I + \gamma \Sigma$ once offline, each call to compute $\text{prox}_{gh}$ also requires $O(n^2)$ operations. Finally, $\text{prox}_{g}$ requires $2n$ comparisons and a norm-operation, and is clearly the least expensive operation.

DCA and FBS need one $\nabla h$ and one $\text{prox}_{gh}$ (or similar) operation, and DRS one $\text{prox}_{gh}$ (work equivalent to $\text{prox}_{gh}$) and one $\text{prox}_{g}$ operation per iteration. Algorithm 1 requires one $\text{prox}_{gh}$ and one $\text{prox}_{g}$ operation per iteration, and L-BFGS needs additionally one call to $h$, $\text{prox}_{gh}$ and $\text{prox}_{g}$ per trial stepsize in the linesearch. However, as $h$ and $\text{prox}_{gh}$ involve linear operations for this particular problem, only one evaluation is required during the whole linesearch. Furthermore, in practice, 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 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 regularity properties. We believe that this connection may be a valuable tool for relaxing assumptions in FBE-based algorithms, which is planned for future work.

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. *Mathematical Programming*, 169(1):95–118, 2018.
[3] M. Baćák and J. Borwein. On difference convexity of locally lipschitz functions. *Optimization*, 66(8-9):961–978, 2017.
[4] S. Banert and R. Bot. A general double-proximal gradient algorithm for DC programming. *Mathematical programming*, 178(1-2):301–326, 2019.
[5] H.H. Bauschke and P.L. Combettes. Convex analysis and monotone operator theory in Hilbert spaces. CMS Books in Mathematics. Springer, 2017.
[6] A. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2016.
[7] J. Bolte, S. Sabach, and M. Teboulle. Proximal Alternating Linearized Minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.
[8] P.L. Combettes and J.C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer New York, New York, NY, 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. *Journal of Optimization Theory and Applications*, 178(3):673–698, 2018.
[11] J.B. Hiriart-Urruty. From Convex Optimization to Nonconvex Optimization, Necessary and Sufficient Conditions for Global Optimality, pages 219–239. Springer US, Boston, MA, 1989.
[12] J.B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms I: Fundamentals*, volume 305. Springer, 1993.
[13] J.B. Hiriart-Urruty and C. Lemaréchal. Fundamentals of Convex Analysis, Grundlehren Text Editions. Springer Berlin Heidelberg, 2012.
[14] R. Horst and N. Thoai. DC programming: overview. *Journal of Optimization Theory and Applications*, 103(1):1–43, 1999.
[15] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(Feb):517–553, 2010.
[16] T. Liu and TK. Pong. Further properties of the forward-backward envelope with applications to difference-of-convex programming. *Computational Optimization and Applications*, 67(3):489–520, Jul 2017.
[17] P. Patrinos and A. Bemporad. Proximal Newton methods for convex composite optimization. In 52nd *IEEE Conference on Decision and Control*, pages 2358–2363, 2013.
[18] P. Patrinos, L. Stella, and A. Bemporad. Douglas-Rachford splitting: Complexity estimates and accelerated variants. In 53rd *IEEE Conference on Decision and Control*, pages 4234–4239, Dec 2014.
[19] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
[20] R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.
[21] L. Stella, A. Themelis, and P. Patrinos. Forward-backward quasi-Newton methods for nonsmooth optimization problems. *Computational Optimization and Applications*, 67(3):443–487, Jul 2017.
[22] L. Stella, A. Themelis, and P. Patrinos. Newton-type alternating minimization algorithm for convex optimization. *IEEE Transactions on Automatic Control*, 2018.
[23] 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.
[24] A. Themelis and P. Patrinos. Douglas–Rachford splitting and ADMM for nonconvex optimization: Tight convergence results. *SIAM Journal on Optimization*, 30(1):149–181, 2020.
[25] A. Themelis, L. Stella, and P. Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM Journal on Optimization*, 28(3):2274–2303, 2018.