Tight Analysis of Difference-of-Convex Algorithm (DCA) Improves Convergence Rates for Proximal Gradient Descent

Teodor Rotaru
KU Leuven, UCLouvain

Panagiotis Patrinos
KU Leuven

François Glineur
UCLouvain

Abstract

We investigate a difference-of-convex (DC) formulation where the second term is allowed to be weakly convex. We examine the precise behavior of a single iteration of the difference-of-convex algorithm (DCA), providing a tight characterization of the objective function decrease, distinguishing between six distinct parameter regimes. Our proofs, inspired by the performance estimation framework, are notably simplified compared to related prior research. We subsequently derive sublinear convergence rates for the DCA towards critical points, assuming at least one of the functions is smooth. Additionally, we explore the underexamined equivalence between proximal gradient descent (PGD) and DCA iterations, demonstrating how DCA, a parameter-free algorithm, without the need for a stepsize, serves as a tool for studying the exact convergence rates of PGD. Finally, we propose a method to optimize the DC decomposition to achieve optimal convergence rates, potentially transforming the subtracted function to become weakly convex.

1 INTRODUCTION

Consider the difference-of-convex formulation

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\, F(x) := f_1(x) - f_2(x), \qquad \text{(DC)}$$

where $f_1, f_2 : \mathbb{R}^d \to \mathbb{R}$ are proper, lower semicontinuous convex functions, and F is lower bounded.

A standard method to solve (DC) is the difference-of-convex algorithm (DCA), a versatile method with no

parameter that can find a *critical* point of F, defined as a point $x^* \in \mathbb{R}^d$ for which there exists subgradients $g_1^* \in \partial f_1(x^*)$ and $g_2^* \in \partial f_2(x^*)$ such that $g_1^* = g_2^*$. Stationary points of F are always critical, but the converse is not true. Extensive analyses of DCA are provided by Dinh and Thi (1997), Tao and An (1998), Horst and Thoai (1999), Le Thi and Pham Dinh (2018). DCA is also referred to as the convex-concave procedure (CCCP), as seen in the work of Yuille and Rangarajan (2001), Lanckriet and Sriperumbudur (2009), Lipp and Boyd (2016). Interestingly, convergence analysis of many methods can be reduced to the one of DCA; for example, the Frank-Wolfe algorithm (Yurtsever and Sra (2022)) or the proximal gradient descent (PGD) (Le Thi and Pham Dinh (2018, Section 3.3.4)). Conversely, Faust et al. (2023) show that DCA is an instance of the Bregman proximal point algorithm.

An extensive list of DCA applications is provided by Le Thi and Pham Dinh (2018). Notable examples include efficient formulations for clustering problems (Hoai An et al. (2014)), dictionary learning (Vo et al. (2015)), robust support vector regression (Wang et al. (2015)), multi-class support vector machines (MSVM) (Le Thi and Nguyen (2017)), sparse logistic regression (Yang and Qian (2016)), compressed sensing (Yin et al. (2015)), adversarial attack for adversarial robustness and approximate optimization of complex functions (Awasthi et al. (2024)) or Shallow Multilayer Perceptron (MLP) Neural Networks (Askarizadeh et al.(2024)). Sun et al. (2024) introduce the Negative ResNets, where f_2 is weakly convex.

We derive convergence rates to critical points when at least one of f_1 and f_2 is smooth (namely continuously differentiable, with Lipschitz gradient). Abbaszadehpeivasti et al. (2023) provide exact convergence rates of DCA when both functions are convex. Their approach, based on performance estimation (PEP) introduced by Drori and Teboulle (2014) and refined by Taylor et al. (2017a), leads to rigorous proofs for those rates. Their exactness is supported by strong numerical evidence and, in some cases, by the identification of instances matching those rates exactly.

In this work, we generalize the standard (DC) setting and consider the case where f_2 can be weakly convex (or hypoconvex). Some previous works also introduce weak convexity in either f_1 (Sun and Sun (2022)) or f_2 (Syrtseva et al. (2024)).

A key motivation for examining the case with f_2 weakly convex is that it mirrors the behavior of applying PGD with stepsizes larger than the inverse Lipschitz constant (see Section 5). Additionally, our generalized DCA framework provides a useful tool for analyzing exact rates for PGD, with the benefit of handling one fewer parameter - DCA involves four curvature parameters compared to PGD's four curvature parameters plus the stepsize. Therefore, due to the equivalence of the iterations, it is more convenient to use a DCA-like analysis. Our results follow the same line as Abbaszadehpeivasti et al. (2023), also relying on performance estimation. More precisely:

- We characterize in Theorem 1 the exact behavior of one iteration of (DCA) in the setting where f_2 may be weakly convex, providing a lower bound for the decrease of the objective expressed in terms of differences of subgradients.

- Theorem 1 describes a total of six distinct regimes, partitioning the parameters space based on smoothness and strong convexity of both functions. We conjecture that these bounds on the objective decrease are tight for all of those six regimes. Among them, only two were previously known and proved by Abbaszadehpeivasti et al. (2023), corresponding to the standard (DCA) setting (f_1 and f_2 convex) where in addition F is required to be both nonconvex and nonconcave.

- Corollary 1 proves that, in our setting allowing f_2 weakly convex, DCA converges sublinearly to critical points, with a $\mathcal{O}(\frac{1}{N})$ rate after N iterations, again with six distinct regimes. Based on strong numerical evidence, we conjecture that three of those rates are exact for any number of iterations.

- We show that a split of the objective F allowing weak convexity of f_2 can yield better rates than the standard DCA. Moreover, when both functions are smooth, a well-chosen DC splitting may surpass the celebrated gradient descent.

- As a direct consequence of our in-depth analysis, we can readily transfer the rates of specific regimes to the PGD setting.

We provide a GitHub repository to support the numerical conjectures and to reproduce all the simulations.

2 THEORETICAL BACKGROUND

Definition 1. *Let $L > 0$ and $\mu \leq L$. We say that a proper, lower semicontinuous function $f : \mathbb{R}^d \to \mathbb{R}$ belongs to the class $\mathcal{F}_{\mu,L}(\mathbb{R}^d)$ (or simply $\mathcal{F}_{\mu,L}$) if and only if it has both (i) upper curvature L, meaning that $\frac{L}{2}\|\cdot\|^2 - f$ is convex, and (ii) lower curvature μ, meaning that $f - \frac{\mu}{2}\|\cdot\|^2$ is convex. We also define the class $\mathcal{F}_{\mu,\infty}(\mathbb{R}^d)$ which requires only lower curvature μ.*

Intuitively, the curvature bounds μ and L correspond to the minimum and maximum eigenvalues of the Hessian for a function $f \in \mathcal{C}^2$. Functions in $\mathcal{F}_{\mu,L}$ must be smooth when $L < \infty$, while $\mathcal{F}_{\mu,\infty}$ also contains nonsmooth functions. Depending on the sign of the lower curvature μ, a function $f \in \mathcal{F}_{\mu,L}$ is categorized as: (i) weakly convex (or hypoconvex) when $\mu < 0$, (ii) convex when $\mu = 0$ or (iii) strongly convex for $\mu > 0$.

The subdifferential of a proper, lower semicontinuous convex (l.s.c.) function $f : \mathbb{R}^d \to \mathbb{R}$ at a point $x \in \mathbb{R}^d$ is defined as

$$\partial f(x) \coloneqq \{g \in \mathbb{R}^d \,|\, f(y) \geq f(x) + \langle g, y - x \rangle \; \forall y \in \mathbb{R}^d\}.$$

For weakly convex functions, the subdifferential can be defined as follows Bauschke et al. (2021, Proposition 6.3). Let $f \in \mathcal{F}_{\mu,\infty}$ be a weakly convex function with $\mu < 0$. Then $\tilde{f}(x) \coloneqq f(x) - \mu\frac{\|x\|^2}{2}$ is convex with a well-defined subdifferential $\partial \tilde{f}(x)$, and we let $\partial f(x) \coloneqq \{\tilde{g} + \mu x \,|\, \tilde{g} \in \partial \tilde{f}(x)\}$. Finally, if f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.

Assumption 1 (Objective and parameters)**.** *The objective function F in (DC) is lower bounded and can be written as $F = f_1 - f_2$, where $f_1 \in \mathcal{F}_{\mu_1,L_1}$ and $f_2 \in \mathcal{F}_{\mu_2,L_2}$, with parameters $\mu_1 \in [0,\infty)$, $L_1 \in (0,\infty]$, $\mu_2 \in (-\infty,\infty)$ and $L_2 \in (\mu_2,\infty]$, such that $\mu_1 < L_1$ and $\mu_2 < L_2$.*

Assumption 1 runs throughout the rest of this paper and it implies $F \in \mathcal{F}_{\mu_1 - L_2\,,\,L_1 - \mu_2}$. Allowing function f_2 to be concave is directly applicable to analyzing the PGD iteration on strongly convex functions with long stepsizes (see Section 5). We also denote $F_{lo} \coloneqq \inf_x F$.

The domain and range of function $f : \mathbb{R}^d \to \mathbb{R}$ are $\operatorname{dom} f \coloneqq \{x \in \mathbb{R}^d : f(x) < \infty\}$ and $\operatorname{range} f \coloneqq \{y \in \mathbb{R} : \exists \, x \in \operatorname{dom} f \text{ with } y = f(x)\}$, respectively. The domain and range of the subdifferential are $\operatorname{dom} \partial f = \{x \in \mathbb{R}^d : \partial f(x) \neq \emptyset\}$ and $\operatorname{range} \partial f = \cup \{\partial f(x) : x \in \operatorname{dom} \partial f\}$, respectively. The convex conjugate of a l.s.c. function f is defined as $f^*(y) \coloneqq \sup_{x \in \operatorname{dom} f}\{\langle y, x \rangle - f(x)\}$, where f^* is closed and convex.

DCA iteration.

1. Select $g_2 \in \partial f_2(x)$.
2. Select $x^+ \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}}\{f_1(w) - \langle g_2, w \rangle\}$.   (DCA)

With an abuse of notation, a more compact definition of the (DCA) iteration is $x^+ \in \partial f_1^*(\partial f_2(x))$.

The optimality condition in the definition of x^+ implies the existence of $g_1^+ \in \partial f_1(x^+)$ such that $g_1^+ = g_2$, where $g_2 \in \partial f_2(x)$. This is the only characterization of x^+ used in our derivations. Note that the sequence of iterates produced by (DCA) is not unique.

Assumption 2. *The subdifferentials of f_1 and f_2 satisfy the following conditions: $\varnothing \neq \operatorname{dom}\partial f_1 \subseteq \operatorname{dom}\partial f_2$ and $\operatorname{range}\partial f_2 \subseteq \operatorname{range}\partial f_1$.*

Proposition 1. *Under Assumption 2, the (DCA) iterations are well-defined, meaning there exists a sequence $\{x^k\}$, starting from $x^0 \in \operatorname{dom}\partial f_1$, generated by $x^{k+1} \in \partial f_1^*(\partial f_2(x^k))$.*

Tao and An (1998) note that DCA is typically well-defined, as for any l.s.c. function f, it holds $\operatorname{ri}(\operatorname{dom} f) \subseteq \operatorname{dom}\partial f \subseteq \operatorname{dom} f$, where $\operatorname{ri}(\operatorname{dom} f)$ is the relative interior of $\operatorname{dom} f$. The potential weak convexity of f_2 represents only a curvature adjustment in the subdifferential definition.

A critical point x^* satisfies $\partial f_2(x^*) \cap \partial f_1(x^*) \neq \emptyset$. When both functions f_1 and f_2 are smooth, any critical point x^* is clearly stationary as we have $\nabla F(x^*) = \nabla f_1(x^*) - \nabla f_2(x^*) = 0$. If only f_2 is smooth, we have $\partial F(x^*) = \partial f_1(x^*) - \nabla f_2(x^*)$ (Rockafellar and Wets (1998, Exercise 10.10)) and criticality also implies stationarity, since $0 \in \partial F(x^*)$. However, if only f_1 is smooth, we can only guarantee the inclusion $\partial(-f_2)(x^*) \subseteq -\partial f_2(x^*)$ (Rockafellar and Wets (1998, Corollary 9.21)), implying only $\partial F(x^*) \subseteq \nabla f_1(x^*) - \partial f_2(x^*)$, and critical points may not be stationary.

Proposition 2 (Sufficient condition for decrease). *Let $f_1 \in \mathcal{F}_{\mu_1,L_1}$ and $f_2 \in \mathcal{F}_{\mu_2,L_2}$. If $\mu_1 + \mu_2 \geq 0$, then the objective function $F = f_1 - f_2$ decreases after each iteration of (DCA). Moreover, if $\mu_1 + \mu_2 > 0$ that objective decrease is strict, unless $x^+ = x$.*

Proposition 2 is proved in Appendix A.1, inspired by Dinh and Thi (1997, Theorem 3, Proposition 2).

Remark 1. *Throughout this paper, we assume available the exact oracles of ∂f_1 and ∂f_1^*, thus we only focus on the progress of the iterations.*

Notation: Superscripts indicate the iteration index (e.g., x^k represents the k-th iterate).

3 CONVERGENCE ANALYSIS

Theorem 1 (One-step decrease). *Let $f_1 \in \mathcal{F}_{\mu_1,L_1}$ and $f_2 \in \mathcal{F}_{\mu_2,L_2}$ satisfy Assumptions 1 and 2, with at least f_1 or f_2 smooth, and assume $\mu_1 + \mu_2 > 0$ or $\mu_1 = \mu_2 = 0$. Then after one step of (DCA) we have*

$$F(x) - F(x^+) \geq \sigma_i \tfrac{1}{2}\|g_1 - g_2\|^2 + \sigma_i^+ \tfrac{1}{2}\|g_1^+ - g_2^+\|^2 \quad (1)$$

with $g_1 \in \partial f_1(x)$, $g_1^+ \in \partial f_1(x^+)$, $g_2 \in \partial f_2(x)$, $g_2^+ \in \partial f_2(x^+)$, and the expressions for $\sigma_i, \sigma_i^+ \geq 0$ correspond to one of the six regimes (indexed by $i = 1, \ldots, 6$) described in Table 1 according to the values of parameters L_1, L_2, μ_1, μ_2.

The six regimes appearing in Table 1 are illustrated in Figure 1; we refer to each p_i as one of the six regimes together with its corresponding expression. Notably, there is a striking symmetry between regimes p_1 and p_2, as well as between p_5 and p_6. Specifically, the formulas for p_2 and p_6 in Table 1 can be derived from those of p_1 and p_5 by swapping $L_1 \leftrightarrow L_2$, $\mu_1 \leftrightarrow \mu_2$, and $\sigma_i \leftrightarrow \sigma_i^+$. The proof of Theorem 1 is deferred to Appendix A.2.

Conjecture 1 (Tightest decrease after one iteration). *All six regimes outlined in Theorem 1 are tight, i.e., the corresponding lower bounds on the objective decrease cannot be improved.*

Conjecture 1 asserts that our set of six inequalities represents the tightest possible characterization after a single iteration. Specifically, there exist (separate) function examples for which each of these inequalities holds with equality.

Corollary 1 (DCA sublinear rates). *Let $f_1 \in \mathcal{F}_{\mu_1,L_1}$ and $f_2 \in \mathcal{F}_{\mu_2,L_2}$ satisfying Assumptions 1 and 2, assume at least f_1 or f_2 is smooth, and assume $\mu_1 + \mu_2 > 0$ or $\mu_1 = \mu_2 = 0$. Then after N iterations of (DCA) starting from x^0 we have*

$$\tfrac{1}{2}\min_{0 \leq k \leq N}\{\|g_1^k - g_2^k\|^2\} \leq \frac{F(x^0) - F(x^N)}{p_i(L_1, L_2, \mu_1, \mu_2)N}, \quad (2)$$

where $g_1^k \in \partial f_1(x^k)$ and $g_2^k \in \partial f_2(x^k)$ for all $k = 0, \ldots, N$ and $p_i = \sigma_i + \sigma_i^+$ is given in Table 1. Additionally, if F is nonconcave (i.e., $L_1 > \mu_2$):

$$\tfrac{1}{2}\min_{0 \leq k \leq N}\{\|g_1^k - g_2^k\|^2\} \leq \frac{F(x^0) - F_{lo}}{p_i(L_1, L_2, \mu_1, \mu_2)N + \frac{1}{L_1 - \mu_2}}.$$

Regimes p_1 and p_2 correspond in part to the standard setting of (DCA), where both functions are convex ($\mu_1 \geq 0$, $\mu_2 \geq 0$). Whether p_1 or p_2 holds depends on which is larger among L_1 and L_2. These regimes require the objective $F \in \mathcal{F}_{\mu_1 - L_2, \ L_1 - \mu_2}$ to be nonconvex ($L_2 > \mu_1$) and nonconcave ($L_1 > \mu_2$), and were first established by Abbaszadehpeivasti et al. (2023), using performance estimation. All other described regimes are novel.

Remark 2. *In the specific convex scenario $\mu_1 = \mu_2 = 0$, both regimes p_1 and p_2 hold, as outlined by Abbaszadehpeivasti et al. (2023, Corollary 3.1) and the one-step decrease is given by: $F(x) - F(x^+) \geq \frac{1}{2L_1}\|g_1 - g_2\|^2 + \frac{1}{2L_2}\|g_1^+ - g_2^+\|^2$. The same result is obtained using the Bregman proximal point algorithm perspective by Faust et al.(2023, Section 4.2).*

Table 1: Exact decrease after one iteration: $F(x) - F(x^+) \geq \sigma_i \frac{1}{2}\|g_1 - g_2\|^2 + \sigma_i^+ \frac{1}{2}\|g_1^+ - g_2^+\|^2$ (see Theorem 1), with $\sigma_i, \sigma_i^+ \geq 0$ and $p_i = \sigma_i + \sigma_i^+$, $i=1,...,6$. The domains satisfy the condition $\mu_1 + \mu_2 > 0$ or $\mu_1 = \mu_2 = 0$, with at least one between f_1 and f_2 smooth. Notation: $B := \mu_1^{-1} + \mu_2^{-1} + L_2^{-1}$ and $E := \frac{L_2+\mu_2}{L_1 L_2}\frac{L_2-L_1}{-\mu_2} + \mu_1^{-1} - L_1^{-1}$.

Regime	σ_i	σ_i^+	Domain			Description
p_1	$L_2^{-1}\dfrac{L_2-\mu_1}{L_1-\mu_1}$	$L_2^{-1}\left(1+\dfrac{L_2^{-1}-L_1^{-1}}{\mu_1^{-1}-L_1^{-1}}\right)$	$L_1 \geq L_2 \geq \mu_1 \geq 0$; $L_1 > \mu_2$		$\mu_2 \geq 0$	f_1, f_2 convex F nonconvex-nonconcave
					$\mu_2 < 0$ and $E \leq 0$	f_1 strongly convex, f_2 nonconvex F nonconvex-nonconcave
p_2	$L_1^{-1}\left(1+\dfrac{L_1^{-1}-L_2^{-1}}{\mu_2^{-1}-L_2^{-1}}\right)$	$L_1^{-1}\dfrac{L_1-\mu_2}{L_2-\mu_2}$	$L_2 \geq L_1 \geq \mu_2 \geq 0$; $L_2 > \mu_1$		$\mu_1 \geq 0$	f_1, f_2 convex F nonconvex-nonconcave
p_3	$\dfrac{L_1^{-1}(\mu_1^{-1}+\mu_2^{-1}+L_2^{-1})}{\mu_1^{-1}+\mu_2^{-1}+L_2^{-1}-L_1^{-1}}$	$\dfrac{1}{L_2+\mu_2}$	$\mu_2 < 0, \mu_1 > 0$; $L_2 > \mu_1$; $L_1 > \mu_2$	$B \leq 0$	$L_1 \geq L_2$ and $E \geq 0$	f_1 strongly convex, f_2 nonconvex F nonconvex-nonconcave
					$L_2 > L_1$	
p_4	0	$\dfrac{\mu_1+\mu_2}{\mu_2^2}$	$\mu_2 < 0, \mu_1 > 0$; $L_1 > \mu_2$;	$B > 0$	$L_2 > \mu_1 > 0$	f_1 strongly convex, f_2 nonconvex F nonconvex-nonconcave
					$0 < L_2 \leq \mu_1$	f_1 strongly convex, f_2 nonconvex F convex
				$B \leq 0$	$L_2 \leq 0$	f_1 strongly convex, f_2 concave F strongly convex
p_5	0	$\dfrac{L_2+\mu_1}{L_2^2}$	$L_1 > \mu_1 \geq L_2 > 0$; $L_1 > \mu_2$		$\mu_2 \geq 0$	f_1 strongly convex, f_2 convex F convex
					$\mu_2 < 0$ and $B \leq 0$	f_1 strongly convex, f_2 nonconvex F convex
p_6	$\dfrac{L_1+\mu_2}{L_1^2}$	0	$L_2 > \mu_2 \geq L_1 > \mu_1$		$\mu_1 \geq 0$	f_1 convex, f_2 strongly convex F concave

If f_1 is strongly convex, Theorem 1 actually extends regime p_1 beyond the difference-of-convex case, i.e., to situations where f_2 is weakly convex, such that $\mu_1 > 0 > -\mu_2$. This is valid up to a certain threshold determined by the sign of $E := \frac{L_2+\mu_2}{L_1 L_2}\frac{L_2-L_1}{-\mu_2} + \mu_1^{-1} - L_1^{-1}$. For p_1, the condition $E < 0$ holds, while for $E \geq 0$ regime p_3 emerges. Moreover, for $L_2 \geq L_1$ it always holds $E \geq 0$. Additionally, the boundary of regime p_3 is constrained by the threshold $B \leq 0$, where

$$B := \mu_1^{-1} + \mu_2^{-1} + L_2^{-1}. \qquad (3)$$

Regime p_4, emerging for $L_2 \cdot B > 0$, includes two cases: (i) when F is nonconvex-nonconcave; and (ii) when F is strongly convex (even containing f_2 concave with $L_2 \leq 0$). The threshold condition $B = 0$ (depicted by the red curve from Figure 1) distinguishes regime p_4 from p_3 and p_5. The later are separated by the condition $L_2 = \mu_1$, delineating the cases F nonconvex (for p_3) and F (strongly) convex (for p_5), respectively. For completeness of analysis, we also include regime p_6, arising for a (strongly) concave objective (and unbounded from below), with $\mu_2 \geq L_1$.

For the particular setup $L_1 = 2$ and $\mu_1 = 1$, in Figure 1 we show, as a contour plot, the values of denominators p_i depending on curvatures L_2 and μ_2.

Our numerical investigations show that the sublinear rates for $p_{4,5,6}$ in Corollary 1 are not tight beyond a single iteration. In the standard case of DCA with F being nonconvex and nonconcave ($\mu_1, \mu_2 < \min\{L_1, L_2\}$), the threshold condition $B = 0$ separates the tight and non-tight regimes in Corollary 1.

Conjecture 2 (Tightness of sublinear rates). *The DCA rates corresponding to regimes p_1, p_2 and p_3 from Corollary 1 are tight for any number of iterations N.*

Conjecture 2 asserts that regimes $p_{1,2,3}$ remain tight when exploiting the analysis for one iteration to obtain rates after an arbitrary number of iterations; on such functions, one recovers exactly the worst-case performance when applying DCA. In these cases, closed-from worst-case function examples can be derived. Regime p_2 is shown to be tight by Abbaszadehpeivasti et al. (2023, Example 3.1) for the specific decomposition $f_1 \in \mathcal{F}_{0,L_1}$ and $f_2 \in \mathcal{F}_{0,\infty}$. In Appendix B, we provide worst-case examples when both f_1 and f_2 are smooth, alongside PEP-based numerical evidences.

One Nonsmooth Term

All the above results hold when at least one of the functions f_1 and f_2 is smooth. When exactly one of them is smooth, i.e., when the other is nonsmooth, some expressions in Table 1 become simpler, and we give an explicit description below. In the standard use of DCA, the conjugate step is applied to f_1 nonsmooth.

Corollary 2. *Let $f_1 \in \mathcal{F}_{\mu_1,L_1}$ and $f_2 \in \mathcal{F}_{\mu_2,L_2}$, where exactly one function f_1 or f_2 is smooth, and assume*
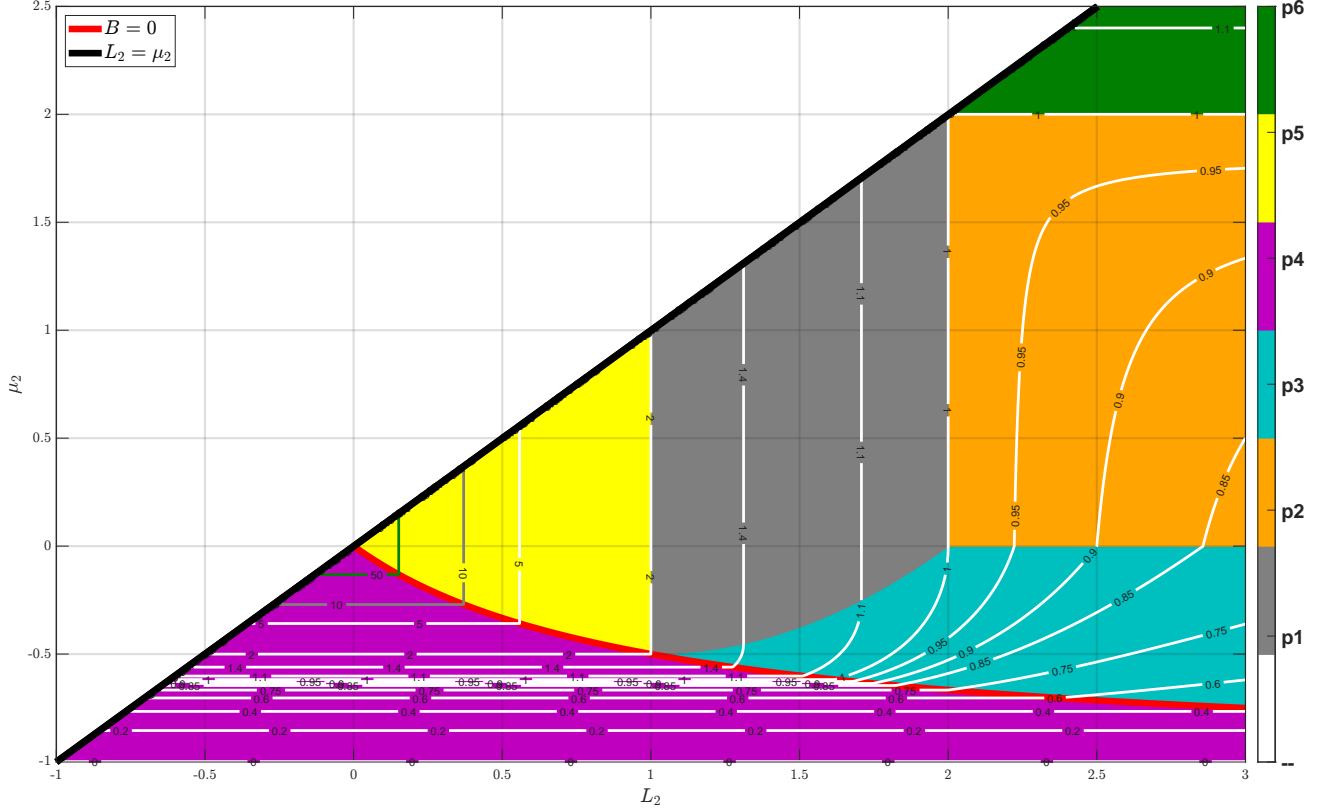
Figure 1: All regimes after one (DCA) iteration (Theorem 1), with $\mu_1 = 1$ and $L_1 = 2$. Contour lines of the denominators p_i vs. L_2 and μ_2 are shown. Regimes are bounded by $L_2 > \mu_2$ and $\mu_2 > -\mu_1 = -1$. Regimes p_1, p_2 and p_3 lie within the area delimited by the *threshold* $B = 0$ ((3)) and the conditions $L_2 > \mu_1 = 1$ and $L_1 = 2 > \mu_2$ (namely F is nonconvex-nonconcave), and are conjectured to be tight after N iterations of (DCA).

$\mu_1 + \mu_2 > 0$ or $\mu_1 = \mu_2 = 0$. *Consider N iterations of* (DCA) *starting from x^0. Then:*

$$\frac{1}{2} \min_{0 \leq k \leq N} \{\|g_1^k - g_2^k\|^2\} \leq \frac{F(x^0) - F(x^N)}{p_i(L_1, L_2, \mu_1, \mu_2)N},$$

where $g_1^k \in \partial f_1(x^k)$ and $g_2^k \in \partial f_2(x^k)$ for all $k = 0, \ldots, N$ and p_i is provided in Table 2. Additionally, if F is nonconcave (i.e., $L_1 > \mu_2$):

$$\frac{1}{2} \min_{0 \leq k \leq N} \{\|g_1^k - g_2^k\|^2\} \leq \frac{F(x^0) - F_{lo}}{p_i(L_1, L_2, \mu_1, \mu_2)N + \frac{1}{L_1 - \mu_2}}.$$

Corollary 2 is derived by setting $L_1 = \infty$ or $L_2 = \infty$ in Corollary 1 and in the corresponding entries from Table 1. It shows identical rates as Abbaszadehpeivasti et al. (2023, Corollary 3.1) for $\mu_1, \mu_2 \geq 0$, while extending them to scenarios involving weakly convex f_2, up to the threshold $B = 0$, beyond which regime p_4 emerges. Notably, we observe in Table 2 that regimes p_5 and p_6 condense to regimes p_1 and p_2, respectively. When $L_1 = \infty$, only regimes p_1 and p_4 hold, separated by the threshold $B = 0$. Conversely, when $L_2 = \infty$, p_2 covers the domain with $\mu_2 \geq 0$ and p_3 corresponds

Table 2: Exact decrease after one iteration: $F(x) - F(x^+) \geq \sigma_i \frac{1}{2}\|g_1 - g_2\|^2 + \sigma_i^+ \frac{1}{2}\|g_1^+ - g_2^+\|^2$, with f_1 or f_2 nonsmooth, $\sigma_i, \sigma_i^+ \geq 0$, $p_i = \sigma_i + \sigma_i^+$.

Regime	σ_i	σ_i^+	Domain
$p_{1,5}$	0	$\frac{L_2 + \mu_1}{L_2^2}$	$L_1 = \infty > L_2 \geq \mu_1 \geq 0$ $\mu_2(\mu_1^{-1} + \mu_2^{-1} + L_2^{-1}) \geq 0$
$p_{2,6}$	$\frac{L_1 + \mu_2}{L_1^2}$	0	$L_2 = \infty > L_1 \geq \mu_2 \geq 0$ $\mu_1(\mu_1^{-1} + \mu_2^{-1} + L_1^{-1}) \geq 0$
p_3	$\frac{\frac{1}{L_1}\left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right)}{\frac{1}{\mu_1} + \frac{1}{\mu_2} - \frac{1}{L_1}}$	0	$L_2 = \infty;\ \mu_1 > -\mu_2 > 0$
p_4	0	$\frac{\mu_1 + \mu_2}{\mu_2^2}$	$L_1 = \infty;\ \mu_1 > -\mu_2 > 0$ $0 < \mu_1^{-1} + \mu_2^{-1} + L_2^{-1}$

to $\mu_2 < 0$. For a graphical intuition of these regimes in the nonsmooth case, see Appendix D.

4 MOTIVATIONS FOR WEAKLY CONVEX f_2

We assume that the convex conjugate of f_1 and of any curvature adjustment $f_1 - \lambda \frac{\|\cdot\|^2}{2}$, where $\lambda \leq \mu_1$, can be computed efficiently. This leads to the natural

question: given a splitting $F = f_1 - f_2$, what is the optimal curvature shift λ in the decomposition $F = (f_1 - \lambda \frac{\|\cdot\|^2}{2}) - (f_2 - \lambda \frac{\|\cdot\|^2}{2})$? In fact, this is a standard approach for addressing weak convexity in the function f_2, by lifting it to a convex function, with some $\lambda \leq \mu_2 < 0$, and then applying the DCA iteratively.

We show that this approach is suboptimal and the sublinear rates, compared in terms of largest denominator p_i, can be improved. Specifically, the constant in sublinear rates for a nonconvex-nonconcave objective F (where $\max\{\mu_1, \mu_2\} < \min\{L_1, L_2\}$) is optimized with respect to the curvature shift λ by maximizing the denominator. In some cases, as indicated in the *Ratio* column from Table 3, the improvement is significant.

Let $\tilde{f}_i^\lambda := f_i - \lambda \frac{\|\cdot\|^2}{2}$, with $i = \{1, 2\}$, be the curvature adjusted functions, and $P_\lambda := P(L_1, \mu_1, L_2, \mu_2, \lambda)$ be the denominator corresponding to one of the six possible regimes p_i, determined by the initial splitting curvatures and the parameter λ. Initially, $\lambda = 0$ and the denominator is P_0. In Table 3 we use p_i, along with its value, to represent the regime before and after the curvature adjustment. Given the analytical expressions for all six regimes, we can easily numerically compute $\lambda^* = \arg\max_\lambda P_\lambda$.

The examples in Table 3 suggest the following observations. First, consider both functions to be (strongly) convex. If $\mu_1 \leq \mu_2$, then the best splitting is achieved by making $\tilde{f}_1$ convex, hence $\lambda = \mu_1$. Notably, when $\mu_1 = \mu_2$, both functions become convex, which surprisingly implies that the initial strong convexity may actually slow down the algorithm. If $\mu_1 > \mu_2$, the optimal splitting occurs when f_2 is shifted to a weakly convex function. Furthermore, even when starting with f_2 as weakly convex, the best curvature maintains this weak convexity. Additionally, in case of a *bad* decomposition where $\mu_1 + \mu_2 < 0$ (providing no convergence guarantee for DCA iterations), an appropriate λ can ensure convergence. In fact, the optimal denominator is reached for some $\mu_2 < 0$.

This collection of seemingly surprising results involving weak convexity of f_2 is explained by the equivalence with the proximal gradient descent (PGD) (see Section 5) and its faster convergence when using a stepsize larger than the inverse Lipschitz constant.

Benefits on Smooth Functions. The objective F is smooth when both f_1 and f_2 are smooth, raising the question of why to not use gradient descent (GD) directly. We compare the rates of DCA, having the iteration $x^+ = \nabla f_1^*(\nabla f_2(x))$, to GD, whose iteration with stepsize $\gamma \in (0, \frac{2}{L_F})$ reads $x^+ = x - \gamma \nabla F(x)$. We use the criteria $\frac{1}{2}\min\{\|\nabla F(x)\|^2, \|\nabla F(x^+)\|^2\} \leq \frac{F(x) - F(x^+)}{p}$, where p denotes the worst-case denomi-

nator for DCA or GD. In Example 1 we show that, for smooth nonconvex functions, the optimal rate for DCA applied can surpass the optimal rate of GD.

Example 1 (*F* smooth-nonconvex). *Let $L_F = -\mu_F = 1$, and the decomposition $F = f_1 - f_2$, with parameters $\mu_1 = 1.5$, $L_1 = 2$, $\mu_2 = 1$ and $L_2 = 2.5$. This corresponds to regime p_2, with $p_2 = 0.9167$. The optimal stepsize for GD, $\gamma^* = \frac{2}{\sqrt{3}}$ (Abbaszadehpeivasti et al. (2021, Theorem 3)), yields the denominator $p_{GD} = 1.5396$, which is 67% larger than for the initial DCA splitting. However, the best DCA splitting achieved by subtracting curvature $\lambda^* = 1.0091$, leading to $F = \tilde{f}_1^\lambda - \tilde{f}_2^\lambda$ with $\tilde{f}_1^\lambda \in \mathcal{F}_{0.4009, 0.9009}$ and $\tilde{f}_2^\lambda \in \mathcal{F}_{-0.0991, 1.4009}$, corresponds to regime p_3, with $p_{DCA^*} = 1.724$, which is 12% larger than p_{GD}. Thus, with appropriately chosen curvatures in DCA, we can improve convergence rates even in the smooth case.*

5 IMPROVED CONVERGENCE RATES OF PGD

In this section we demonstrate that convergence rates for PGD can be directly derived from its iterate equivalence with DCA, a connection that is often underemphasized in the literature. While DCA is typically applied to nonconvex-nonconcave objective functions, it can also serve as a valuable tool to establish rates in the cases involving (strongly) convex costs.

Assumption 3 (PGD splitting). *Consider the composite objective function $F = \varphi + h$, where $\varphi \in \mathcal{F}_{\mu_\varphi, L_\varphi}$ is smooth, with $L_\varphi > 0$ and $\mu_\varphi \in (-\infty, L_\varphi)$, and $h : \mathbb{R}^d \to (-\infty, \infty]$ is proper, closed and convex, $h \in \mathcal{F}_{\mu_h, L_h}$, such that $0 \leq \mu_h \leq L_h$.*

The PGD iteration with stepsize $\gamma > 0$ is given by:

$$x^+ := \text{prox}_{\gamma\varphi}[x - \gamma \nabla \varphi(x)]$$
$$= \underset{w \in \mathbb{R}^d}{\arg\min} \left\{ h(w) + \frac{1}{2\gamma} \|w - x + \gamma \nabla \varphi(x)\|^2 \right\}. \quad \text{(PGD)}$$

To the best of our knowledge, the only tight rate for φ nonconvex is given by Abbaszadehpeivasti in (2024, Proposition 8.11) for $\mu_\varphi = -L_\varphi$ and limited stepsize domain. For φ convex, refer to Taylor et al. (2018), employing different performance metrics.

Proposition 3 (PGD is equivalent to DCA; Le Thi and Pham Dinh (2018), Section 3.3.4). *Starting from $x \in \mathbb{R}^d$, one iteration of (PGD) with stepsize $\gamma > 0$ on the composite objective function $F = \varphi + h$, under Assumption 3, yields the same point x^+ as a DCA iteration applied to the splitting $F = f_1 - f_2$, where $f_1 = h + \frac{1}{2\gamma}\|\cdot\|^2$ and $f_2 = \frac{1}{2\gamma}\|\cdot\|^2 - \varphi$. Furthermore, the curvatures are related as follows: $\mu_1 = \gamma^{-1} + \mu_h$; $L_1 = \gamma^{-1} + L_h$; $\mu_2 = \gamma^{-1} - L_\varphi$; $L_2 = \gamma^{-1} - \mu_\varphi$.*

Table 3: Improvement of DCA convergence rates for nonconvex-nonconcave objectives F ($\max\{\mu_1, \mu_2\} < \min\{L_1, L_2\}$) using curvature shifting with optimal λ^*. Relative improvement ($Ratio$) is defined as $\frac{P(\lambda^*) - P(0)}{P(0)}$, where $P(0)$ is the initial denominator before splitting and $P(\lambda^*)$ is the optimal (largest) one.

Setup	μ_1	L_1	μ_2	L_2	$P(0)$	λ^*	$P(\lambda^*)$	Ratio
	0.2	3	0.1	4	$p_2 = 0.4535$	0.1221	$p_3 = 0.6031$	33%
$\mu_1 > \mu_2$	0.2	1000	0.1	3	$p_1 = 0.3255$	0.1494	$p_1 = p_3 = 0.358$	10%
	1	2	0.5	1.5	$p_1 = 0.76$	0.6733	$p_3 = 2.0321$	167%
$\mu_1 = \mu_2$	1	4	1	3	$p_1 = 0.4583$	$\mu_1 = 1$	$p_1 = p_2 = 0.8333$	81%
$\mu_1 < \mu_2$	0.1	3	0.2	4	$p_2 = 0.4539$	$\mu_1 = 0.1$	$p_2 = 0.602$	32%
	0.001	4	0.002	3	$p_1 = 0.4531$	$\mu_1 = 0.001$	$p_1 = 0.5835$	28%
$\mu_1 > 0; \mu_1 + \mu_2 > 0$	2	4	-1.75	3	$p_4 = 0.4688$	-0.4855	$p_3 = 0.52$	11%
	2.99	4	-2.9	3	$p_4 = 0.4994$	-0.935	$p_4 = 0.5076$	1.6%
$\mu_1 > 0; \mu_1 + \mu_2 < 0$	1	2	-1.5	1.5	$-$	-0.6526	$p_3 = 0.8516$	$-$

Proposition 4. *The convergence measure based on the (sub)gradients residual norm is identical when applying PGD to the decomposition $F = \varphi + h$ or DCA to the decomposition $F = f_1 - f_2$, where f_1 and f_2 are defined in Proposition 3.*

Propositions 3 and 4 are proved in Appendix A.4. Proposition 4 shows that for any iteration x it holds $\|\nabla\varphi(x) + g_h\|^2 = \|g_1 - \nabla f_2(x)\|^2$, where $g_h \in \partial h(x)$ and $g_1 \in \partial f_1(x)$. To determine convergence rates for PGD, one can substitute the curvature values from the DCA convergence rate expressions in Section 3 with the corresponding curvatures defined in terms of PGD parameters as specified in Proposition 3. Moreover, the smoothness of φ implies the smoothness of f_2, ensuring that the critical points are also stationary.

Particular cases of PGD. With $h = 0$, PGD reduces to the celebrated gradient descent. If $h = \delta_C$, the indicator function of a non-empty, closed and convex set C, then PGD becomes the projected gradient descent. Setting $\gamma = \frac{1}{L}$ yields the iterative shrinkage thresholding algorithm (ISTA). If φ is a constant function, PGD becomes the proximal point method.

We focus on the typical PGD setup, where φ is smooth and h is convex and nonsmooth, thus $\mu_h = 0$, $L_h = \infty$. Hence, the DCA-like curvatures are $\mu_1 = \gamma^{-1}$, $L_1 = \infty$, $\mu_2 = \gamma^{-1} - L_\varphi$, $L_2 = \gamma^{-1} - \mu_\varphi$. However, if additional information about h is available, it can be similarly incorporated to derive improved rates.

From the equivalence of curvatures, it follows that large stepsizes $\gamma > \frac{1}{L_\varphi}$ correspond to negative μ_2, indicating that f_2 is weakly convex. Furthermore, the condition $\mu_1 + \mu_2 > 0$ in the DCA setting, ensuring the decrease in the objective after one iteration (see Proposition 2), translates to the standard upper bound on the stepsize for PGD, which is $\gamma < \frac{2}{L_\varphi}$.

In Table 4 we summarize the corresponding DCA curvatures for various notable cases in the PGD setting.

For the stepsize $\gamma = \frac{1}{L_\varphi}$, commonly used in convergence analysis, we have $\mu_2 = 0$. Moreover, the unusual case where $L_2 \leq 0$, meaning that f_2 is concave, corresponds to large stepsizes $\gamma \in \left[\frac{2}{L_\varphi + \mu_\varphi}, \frac{2}{L_\varphi}\right)$ applied to strongly convex objectives.

Theorem 2 (One-step decrease PGD). *Let $\varphi \in \mathcal{F}_{\mu_\varphi, L_\varphi}$ and $h \in \mathcal{F}_{0,\infty}$, with φ smooth. Consider one iteration of (PGD) with stepsize $\gamma \in \left(0, \frac{2}{L_\varphi}\right)$. Then:*

$$F(x) - F(x^+) \geq \sigma^+(L_\varphi, \mu_\varphi, \gamma)\frac{1}{2}\|\nabla\varphi(x^+) + g_h^+\|^2,$$

with $g_h^+ \in \partial h(x^+)$ and $\sigma^+(L_\varphi, \mu_\varphi, \gamma) \geq 0$ defined as:

$$\begin{cases} \frac{\gamma(2-\gamma\mu_\varphi)}{(1-\gamma\mu_\varphi)^2}, & \text{if } \mu_\varphi \leq 0 \text{ and either} \begin{cases} \gamma \in \left(0, \frac{1}{L_\varphi}\right]; \\ \gamma \in \left(\frac{1}{L_\varphi}, \frac{2}{L_\varphi}\right) \\ \quad\text{and } B \leq 0; \end{cases} \\ \frac{\gamma(2-\gamma L_\varphi)}{(1-\gamma L_\varphi)^2}, & \text{if } \gamma \in \left(\frac{1}{L_\varphi}, \frac{2}{L_\varphi}\right) \text{ and } B > 0, \end{cases}$$

where $B = 1 + \frac{1}{1-\gamma L_\varphi} + \frac{1}{1-\gamma\mu_\varphi}$.

Theorem 2 is derived by substituting the curvature expressions from the DCA splitting, as outlined in Proposition 3, into Theorem 1.

Corollary 3 (Sublinear PGD Rates). *Let $\varphi \in \mathcal{F}_{\mu_\varphi, L_\varphi}$ and $h \in \mathcal{F}_{0,\infty}$, with φ smooth, as in Assumption 3. Consider N iterations of (PGD) with stepsize $\gamma \in \left(0, \frac{2}{L_\varphi}\right)$, starting from x^0. Then:*

$$\frac{1}{2} \min_{0 \leq k \leq N}\{\|\nabla\varphi(x^k) + g_h^k\|^2\} \leq \frac{F(x^0) - F_{lo}}{\sigma^+(L_1, L_2, \mu_1, \mu_2)N},$$

where $g_h^k \in \partial h(x^k)$ for all $k = 1, \ldots, N$ and $\sigma^+(L_1, L_2, \mu_1, \mu_2) \geq 0$ is given in Theorem 2.

Corollary 3 follows immediately from Corollary 2. Conjecture 2 on tightness implies that only the first branch in Corollary 3 leads to an exact rate. This corresponds to non-strongly convex functions φ, with $\mu_\varphi \leq 0$, and stepsizes γ less than some threshold $\bar{\gamma}$

Table 4: PGD settings in DCA curvatures: $\mu_h = 0$, $L_h = \infty$ imply $\mu_1 = \gamma^{-1} > 0$ and $L_1 = \infty$ (see Proposition 3)

Convexity of φ	Stepsize γ	$\mu_2 = \gamma^{-1} - L_\varphi$	$L_2 = \gamma^{-1} - \mu_\varphi$	Regime
nonconvex $\mu_\varphi < 0$	$\gamma \in \left(0, \frac{1}{L_\varphi}\right)$	$\mu_2 > 0$	$L_2 > \mu_1$	p_1
	$\gamma = \frac{1}{L_\varphi}$	$\mu_2 = 0$		
	$\gamma \in \left(\frac{1}{L_\varphi}, \frac{2}{L_\varphi}\right)$	$\mu_2 < 0$		p_1 or p_4
convex $\mu_\varphi = 0$	$\gamma \in \left(0, \frac{1}{L_\varphi}\right)$	$\mu_2 > 0$	$L_2 = \mu_1$	$p_1 = p_5$
	$\gamma = \frac{1}{L_\varphi}$	$\mu_2 = 0$		
	$\gamma \in \left(\frac{1}{L_\varphi}, \frac{2}{L_\varphi}\right)$	$\mu_2 < 0$		p_4
strongly convex $\mu_\varphi > 0$	$\gamma \in \left(0, \frac{1}{L_\varphi}\right)$	$\mu_2 > 0$	$|\mu_2| < L_2 < \mu_1$	p_5
	$\gamma = \frac{1}{L_\varphi}$	$\mu_2 = 0$		
	$\gamma \in \left(\frac{1}{L_\varphi}, \frac{2}{L_\varphi + \mu_\varphi}\right)$	$\mu_2 < 0$		p_4 or p_5
	$\gamma \in \left[\frac{2}{L_\varphi + \mu_\varphi}, \frac{2}{L_\varphi}\right)$		$L_2 \leq -\mu_2 < \mu_1$	p_4

which cancels the expression of the threshold B. Further analysis is needed to complete the picture of exact rates for any stepsize γ and any curvature of φ.

6   PROOF TECHNIQUE

Performance estimation problem (PEP) and interpolation. PEP, introduced by Drori and Teboulle (2014) and further refined by Taylor et al. (2017a), provides a framework for analyzing tight convergence rates in various optimization methods. It casts the search for the worst-case function as an optimization problem itself, using specific inequalities which are necessary and sufficient to interpolate functions from the target class, as in Theorem 3. The problem is then reformulated as a convex semidefinite program, numerically solvable. Abbaszadehpeivasti et al. (2023) provide an in-depth guide on setting the PEP for DCA.

Theorem 3 ($\mathcal{F}_{\mu,L}$-interpolation). *Given an index set $\mathcal{I}$, let $\mathcal{T} := \{(x^i, g^i, f^i)\}_{i \in \mathcal{I}} \subseteq \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ be a set of triplets. There exists $f \in \mathcal{F}_{\mu,L}$, with $L > 0$ and $\mu \leq L$, such that $f(x^i) = f^i$ and $g^i \in \partial f(x^i)$ for all $i \in \mathcal{I}$, iff the following interpolation inequality holds for every pair of indices (i, j), with $i, j \in \mathcal{I}$:*

$$f^i - f^j - \langle g^j, x^i - x^j \rangle \geq \frac{1}{2L} \|g^i - g^j\|^2 + \\ + \frac{\mu}{2L(L-\mu)} \|g^i - g^j - L(x^i - x^j)\|^2. \quad (Q_{i,j})$$

Theorem 3 is introduced by Taylor et al. (2017a, Theorem 4) for $\mu \geq 0$ and extended to negative μ by Rotaru et al. (2022, Theorem 3.1). The following general performance estimation setup for DCA can be easily integrated in one of the specialized software packages PESTO (in Matlab; Taylor et al. (2017b)) or PEPit

(in Python; Goujaud et al. (2024)):

$$\text{maximize} \quad \frac{\frac{1}{2} \min_{0 \leq k \leq N} \{\|g_1^k - g_2^k\|^2\}}{F(x^0) - F(x^N)}$$
$$\text{subject to} \quad \{(x^k, g_{1,2}^k, f_{1,2}^k)\}_{k \in \mathcal{I}} \text{ satisfy } (Q_{i,j})$$
$$g_1^{k+1} = g_2^k \quad k \in \{0, \ldots, N-1\}. \quad (4)$$

The decision variables are $x^k, g_1^k, g_2^k, f_1^k, f_2^k$, with $k \in \mathcal{I} = \{0, \ldots, N\}$. The numerical solutions of (4) aided our derivations, enabling verification of the rates from Corollary 1, conjecturing their tightness, and selecting for the proofs the necessary interpolation inequalities from the complete set given in Theorem 3.

While employing PEP can sometimes make proofs difficult to comprehend, we offer a clearer demonstration of the various regimes in Appendix A.2. This involves combining interpolation inequalities for consecutive iterations, with each of the six regimes proof requiring a distinct weighted combination of them.

Conjectures 1 and 2 are supported by extensive numerical evidence obtained through the solution of a large number of performance estimation problems (4), spanning the range of parameters L_1, L_2, μ_1, μ_2. For Conjecture 1, addressing tightness after one iteration, Figure 1 is generated by fixing μ_1 and L_1, creating a sampling grid for μ_2 and L_2, and solving the instances of (4) for $N = 1$. The results align with the *exact* values of $p_{1,\ldots,6}$ in their respective regions. Conjecture 2, regarding tightness of regimes p_1, p_2 and p_3 after N iterations, was established through a similar sampling grid with $N \leq 20$, confirming that these regimes yield tight sublinear rates.

7   EXPERIMENTS: SPARSE PCA

We provide numerical evidence supporting the efficiency of curvature shifting, aligning with the optimal splittings proposed in Section 4, by exploring the sparse principal component analysis (SPCA) problem. Following Journée et al. (2010) and Themelis et al. (2020a), we include elastic-net regularization and solve

$$\underset{x \in \bar{B}(0,1)}{\text{minimize}} \, F(x) := \kappa\|x\|_1 + \eta\frac{\|x\|^2}{2} - \frac{1}{2}x^T \Sigma x,$$

where $\bar{B}(0,1) := \{x | \|x\|_2 \leq 1\}$ is the closed Euclidean unit ball, $\Sigma = A^T A$ the sample covariance matrix, and κ, η are the l_1-(sparsity inducing) and l_2-regularization parameters, respectively. We denote $f_1(x) := \kappa\|x\|_1 + \eta\frac{\|x\|^2}{2} + \delta_{\bar{B}(0,1)}(x)$ and $f_2(x) := \frac{1}{2}x^T \Sigma x$, with $\delta_C(x)$ as the indicator function of a set C. Note that $\mu_1 = \eta$, $L_1 = \infty$, $\mu_2 = \min\{\Lambda(\Sigma)\}$ and $L_2 = \max\{\Lambda(\Sigma)\}$, where $\Lambda(\Sigma)$ denotes the eigenvalues of Σ. Further, we consider the splittings $\tilde{f}_i^\lambda(x) := f_i(x) - \lambda\frac{\|x\|^2}{2}$, with $i = \{1, 2\}$, parametrized by λ (curvature shift); note that $F(x)$ remains unchanged. The subdifferential of the convex conjugate of $\tilde{f}_1^\lambda$ can be computed in closed form for any $\lambda \leq \eta$ (see Appendix F). We generated a sparse random matrix $A \in \mathbb{R}^{20n \times n}$ ($n = 200$), with 10% density, and normalized $\Sigma = AA^T$ by its maximum eigenvalue, obtaining $f_2 \in \mathcal{F}_{\mu_2=0.3882, L_2=1}$. Parameters κ and η were selected to ensure a sparse but non-trivial solution, and our focus is to demonstrate the impact of curvature shifting on convergence.

We use $M = 1000$ random initial points within the unit ball, and only keep the $\bar{M}$ runs converging to the same solution (non-trivial and with desired sparsity). We compare convergence rates for various splittings, namely $\{0, \pm\lambda^*, \pm 0.5\lambda^*, \lambda_{\max}\}$, where λ^* is the optimal curvature shift defined in Section 4 and $\lambda_{\max} := \frac{\mu_1 + \min\{\mu_1, \mu_2\}}{2}$ is the maximum shift guaranteeing convergence. We report in Table 5 the average number of iterations N_ε required to reach a specific accuracy level $\varepsilon \in \{10^{-1}, 10^{-2}, \ldots, 10^{-12}\}$ of the squared (sub)gradient norm of $F(x^k)$ (which is equal to $\text{argmin}_{0 \leq k \leq N} \|\tilde{g}_1^k - \nabla \tilde{f}_2^\lambda(x^k)\|^2$), for $\varepsilon = \{10^{-1}, 10^{-2}, \ldots, 10^{-12}\}$ and $\tilde{g}_1^k \in \partial \tilde{f}_1^\lambda(x^k)$. Complete tables are provided in Appendix F.

Case 1: $\eta > \mu_2$. Parameters: $\eta = \mu_1 = 0.5$, $\kappa = 0.02$ (400 runs kept out of 1000). Here, $\lambda^* = 0.4413$ and $\lambda_{\max} = 0.4441$. With $\lambda > \mu_2 = 0.3882$, $\tilde{f}_2^\lambda$ becomes weakly convex. Using λ^* achieves at least a twofold acceleration compared to $\lambda = 0$. Higher λ values, increasing the nonconvexity of $\tilde{f}_2^\lambda$, further improve the convergence. Conversely, adding curvature to both functions ($\lambda < 0$) slows the convergence. The improved convergence rates observed when $\tilde{f}_2^\lambda$ be-

Table 5:   Average number of iterations N_ε for $\kappa = 0.02$ and (Case 1) $\eta = 0.5 = \mu_1 > \mu_2$; (Case 2) $\eta = 0.2 = \mu_1 < \mu_2$, for $\{0, \pm\lambda^*, \pm 0.5\lambda^*, \lambda_{\max}\}$.

Case	λ \ ε	10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-10}
	0	5.74	46.70	150.17	336.27	544.63
	0.4413	3.65	21.49	70.75	154.93	247.93
1	-0.2207	6.83	59.32	189.31	426.39	693.40
	-0.4413	8.05	72.30	229.13	521.24	844.69
	0.2207	4.54	33.91	108.43	244.05	394.92
	0.4441	3.63	21.23	70.38	153.78	246.61
	0	5.88	49.85	153.46	339.67	547.61
	0.2	4.32	37.10	115.90	256.33	412.39
2	-0.1	6.48	55.52	172.24	382.11	616.29
	-0.2	7.06	61.44	192.46	426.00	686.04
	0.1	5.26	43.47	135.39	298.14	479.93

comes weakly convex further highlight the significance of studying DCA for weakly convex functions.

Case 2: $\eta < \mu_2$. Parameters: $\eta = \mu_1 = 0.2$, $\kappa = 0.02$ (703 runs kept out of 1000). Here, $\lambda^* = \lambda_{\max} = \mu_1 = 0.2$. Using λ^* to make $\tilde{f}_1^\lambda$ convex improves convergence by $\approx 20\%$ compared to the initial splitting.

In conclusion, λ^* from worst-case analysis is also effective for practical curvature shifting, despite better-than-predicted performance. Reducing convexity in both functions improves the rates, similar to larger stepsizes enhancing PGD convergence.

8   CONCLUSION

This work thoroughly examines the behavior of a single (DCA) iteration applied to the DC framework extended to accommodate one weakly convex function. We characterized six distinct regimes for the objective decrease based on subgradient differences and conjecture, based on numerical observations, that certain regimes remain tight across multiple iterations. The convergence rate results as a corollary of this analysis.

The (DC) structure of the objective facilitates curvature shifting by adding/subtracting $\frac{\lambda}{2}\|x\|^2$ to both terms. However, applying the (DCA) iteration on a split with one weakly convex function can yield better rates than using a modified split that achieves convexity for both functions, as shown by several examples and confirmed by numerical experiments.

We highlight the strong link between PGD and DCA, leveraging their iteration equivalence to translate DCA rates to PGD setups. Further work is necessary for obtaining tight rates for any stepsize and curvature choice in PGD, aided by the simpler DCA analysis.

Acknowledgements

References

Abbaszadehpeivasti, H. (2024). *Performance Analysis of Optimization Methods for Machine Learning.* CentER dissertation. CentER, Tilburg University.

Abbaszadehpeivasti, H., de Klerk, E., and Zamani, M. (2021). The exact worst-case convergence rate of the gradient method with fixed step lengths for l-smooth functions. *Optimization Letters.*

Abbaszadehpeivasti, H., de Klerk, E., and Zamani, M. (2023). On the rate of convergence of the difference-of-convex algorithm (DCA). *Journal of Optimization Theory and Applications.*

Askarizadeh, M., Morsali, A., Tofigh, S., and Nguyen, K. K. (2024). Convex-concave programming: An effective alternative for optimizing shallow neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, pages 1–14.

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. (2024). Dc-programming for neural network optimizations. *Journal of Global Optimization.*

Bauschke, H. H., Moursi, W. M., and Wang, X. (2021). Generalized monotone operators and their averaged resolvents. *Mathematical Programming*, 189(1):55–74.

Dinh, T. P. and Thi, H. A. L. (1997). Convex analysis approach to d.c. programming: Theory, algorithm and applications. In *Acta Matehmatica Vietnamica*, volume 22.

Drori, Y. and Teboulle, M. (2014). Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1-2):451–482.

Faust, O., Fawzi, H., and Saunderson, J. (2023). A bregman divergence view on the difference-of-convex algorithm. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 3427–3439. PMLR.

Goujaud, B., Moucer, C., Glineur, F., Hendrickx, J. M., Taylor, A. B., and Dieuleveut, A. (2024). Pepit: computer-assisted worst-case analyses of first-order optimization methods in python. *Mathematical Programming Computation*, 16(3):337–367.

Hoai An, L. T., Hoai Minh, L., and Tao, P. D. (2014). New and efficient dca based algorithms for minimum sum-of-squares clustering. *Pattern Recognition*, 47(1):388–401.

Horst, R. and Thoai, N. V. (1999). DC Programming: Overview. *Journal of Optimization Theory and Applications*, 103(1):1–43.

Journée, M., Nesterov, Y., Richtárik, P., and Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(15):517–553.

Lanckriet, G. and Sriperumbudur, B. K. (2009). On the convergence of the concave-convex procedure. *Advances in Neural Information Processing Systems*, 22.

Le Thi, H. A. and Nguyen, M. C. (2017). Dca based algorithms for feature selection in multi-class support vector machine. *Annals of Operations Research*, 249(1):273–300.

Le Thi, H. A. and Pham Dinh, T. (2018). DC programming and DCA: thirty years of developments. *Mathematical Programming*, 169(1):5–68.

Lipp, T. and Boyd, S. (2016). Variations and extension of the convex–concave procedure. *Optimization and Engineering*, 17(2):263–287.

Rockafellar, R. and Wets, R. J.-B. (1998). *Variational Analysis.* Springer Verlag, Heidelberg, Berlin, New York.

Rotaru, T., Glineur, F., and Patrinos, P. (2022). Tight convergence rates of the gradient method on smooth hypoconvex functions. *arXiv preprint arXiv:2203.00775.*

Sun, K. and Sun, X. A. (2022). Algorithms for difference-of-convex programs based on difference-of-moreau-envelopes smoothing. *INFORMS Journal on Optimization*, 5(4):321–339.

Sun, Y., Liu, Y., and Niu, Y.-S. (2024). Understand the effectiveness of shortcuts through the lens of dca. *arXiv:2412.09853.*

Syrtseva, K., de Oliveira, W., Demassey, S., and van Ackooij, W. (2024). Minimizing the difference of convex and weakly convex functions via bundle method. *PJO.*

Tao, P. D. and An, L. T. H. (1998). A d.c. optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8(2):476–505.

Taylor, A., Hendrickx, J. T., and Glineur, F. (2017a). Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. (2017b). Performance estimation toolbox (PESTO): Automated worst-case analysis of first-order optimization methods. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1278–1283.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. (2018). Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. *Journal of Optimization Theory and Applications*, 178(2):455–476.

Themelis, A., Hermans, B., and Patrinos, P. (2020a). A new envelope function for nonsmooth dc optimization. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 4697–4702.

Themelis, A., Hermans, B., and Patrinos, P. (2020b). A new envelope function for nonsmooth dc optimization. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 4697–4702.

Vo, X. T., An Le Thi, H., Dinh, T. P., and Nguyen, T. B. T. (2015). Dc programming and dca for dictionary learning. In Núñez, M., Nguyen, N. T., Camacho, D., and Trawiński, B., editors, *Computational Collective Intelligence*, pages 295–304, Cham. Springer International Publishing.

Wang, K., Zhu, W., and Zhong, P. (2015). Robust support vector regression with generalized loss function and applications. *Neural Processing Letters*, 41(1):89–106.

Yang, L. and Qian, Y. (2016). A sparse logistic regression framework by difference of convex functions programming. *Applied Intelligence*, 45(2):241–254.

Yin, P., Lou, Y., He, Q., and Xin, J. (2015). Minimization of l1-l2 for compressed sensing. *SIAM Journal on Scientific Computing*, 37(1):A536–A563.

Yuille, A. L. and Rangarajan, A. (2001). The concave-convex procedure (cccp). In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Yurtsever, A. and Sra, S. (2022). Cccp is frank-wolfe in disguise. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 35352–35364. Curran Associates, Inc.

Checklist

1. For all models and algorithms presented, check if you include:

    (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**]

    (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**]

    (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes**; the source code is provided with the supplementary material GitHub repository.]

2. For any theoretical claim, check if you include:

    (a) Statements of the full set of assumptions of all theoretical results. [**Yes**]

    (b) Complete proofs of all theoretical results. [**Yes**; due to space constraints, they are provided in the appendix.]

    (c) Clear explanations of any assumptions. [**Yes**]

3. For all figures and tables that present empirical results, check if you include:

    (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**; the source code to generate our examples, figures and tables is provided with the supplementary material.]

    (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Not Applicable**]

    (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Not Applicable**]

    (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Not Applicable**]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

    (a) Citations of the creator If your work uses existing assets. [**Not Applicable**]

    (b) The license information of the assets, if applicable. [**Not Applicable**]

    (c) New assets either in the supplemental material or as a URL, if applicable. [**Not Applicable**]

    (d) Information about consent from data providers/curators. [**Not Applicable**]

    (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

    (a) The full text of instructions given to participants and screenshots. [**Not Applicable**]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]

A PROOFS

A.1 Proof of Proposition 2

We use the following Lemma.

Lemma 1 (Rotaru et al. (2022), Lemma 2.5). *Let $L > 0$ and $\mu \leq L$, and let $f \in \mathcal{F}_{\mu,L}$. Then $\forall x, y \in \mathbb{R}^d$, with $g \in \partial f(y)$, we have the following:*

$$\frac{\mu}{2}\|x - y\|^2 \leq f(x) - f(y) - \langle g, x - y \rangle \leq \frac{L}{2}\|x - y\|^2.$$

Proof of Proposition 2. Using Lemma 1 on function f_1 for the pair $(x, y) = (x, x^+)$ and on function f_2 with $(x, y) = (x^+, x)$, and summing the inequalities, we obtain:

$$\frac{\mu_1 + \mu_2}{2}\|x^+ - x\|^2 \leq f_1(x) - f_1(x^+) - \langle g_1^+, x - x^+ \rangle +$$

$$f_2(x^+) - f_2(x) - \langle g_2, x^+ - x \rangle \leq \frac{L_1 + L_2}{2}\|x^+ - x\|^2,$$

where $g_1^+ \in \partial f_1(x^+)$ and $g_2 \in \partial f_2(x)$. By using $F(x) = f_1(x) - f_2(x)$ and $g_1^+ = g_2$ we get:

$$\frac{\mu_1 + \mu_2}{2}\|x - x^+\|^2 \leq F(x) - F(x^+) \leq \frac{L_1 + L_2}{2}\|x - x^+\|^2,$$

enough to prove both parts of the proposition. □

A.2 Proof of Theorem 1

In Table 6 we summarize the coefficients for all regimes, along with the corresponding multipliers α_i, which are utilized in the subsequent proofs of each regime in part.

Table 6: Exact decrease after one step: $F(x) - F(x^+) \geq \sigma_i \frac{1}{2}\|g_1 - g_2\|^2 + \sigma_i^+ \frac{1}{2}\|g_1^+ - g_2^+\|^2$ (see Theorem 1), where $\sigma_i, \sigma_i^+ \geq 0$ and $p_i = \sigma_i + \sigma_i^+$, with $i = 1, \ldots, 6$. Scalar $\alpha_i \geq 0$ is a parameter of the proofs.

Regime	σ_i	σ_i^+	α_i
p_1	$L_2^{-1}\dfrac{L_2 - \mu_1}{L_1 - \mu_1}$	$L_2^{-1}\left(1 + \dfrac{L_2^{-1} - L_1^{-1}}{\mu_1^{-1} - L_1^{-1}}\right)$	$\dfrac{\mu_1}{L_2}\dfrac{L_1 - L_2}{L_1 - \mu_1}$
p_2	$L_1^{-1}\left(1 + \dfrac{L_1^{-1} - L_2^{-1}}{\mu_2^{-1} - L_2^{-1}}\right)$	$L_1^{-1}\dfrac{L_1 - \mu_2}{L_2 - \mu_2}$	$\dfrac{\mu_2}{L_1}\dfrac{L_2 - L_1}{L_2 - \mu_2}$
p_3	$\dfrac{L_1^{-1}\left(\mu_1^{-1} + \mu_2^{-1} + L_2^{-1}\right)}{\mu_1^{-1} + \mu_2^{-1} + L_2^{-1} - L_1^{-1}}$	$\dfrac{1}{L_2 + \mu_2}$	$\dfrac{-\mu_2}{L_2 + \mu_2}$
p_4	0	$\dfrac{\mu_1 + \mu_2}{\mu_2^2}$	$\dfrac{\mu_1 + \mu_2}{-\mu_2}$
p_5	0	$\dfrac{L_2 + \mu_1}{L_2^2}$	$\dfrac{\mu_1}{L_2}$
p_6	$\dfrac{L_1 + \mu_2}{L_1^2}$	0	$\dfrac{\mu_2}{L_1}$

Proof of Theorem 1. Let $g_1 \in \partial f_1(x)$, $g_1^+ \in \partial f_1(x^+)$, $g_2 \in \partial f_2(x)$ and $g_2^+ \in \partial f_2(x^+)$, with $g_1^+ = g_2$. We use the notation: $\Delta x := x - x^+$, $G := g_1 - g_2$, $G^+ := g_1^+ - g_2^+$ and $\Delta F(x) := F(x) - F(x^+)$. By writing $(Q_{i,j})$ for function f_1 with the iterates (x, x^+) we obtain:

$$f_1(x) - f_1(x^+) - \langle g_1^+, \Delta x \rangle \geq \frac{1}{2L_1}\|G\|^2 + \frac{\mu_1}{2L_1(L_1 - \mu_1)}\|G - L_1\Delta x\|^2 \tag{5}$$

and for function f_2 with the iterates (x^+, x) we get:

$$f_2(x^+) - f_2(x) + \langle g_2, \Delta x \rangle \geq \frac{1}{2L_2} \|G^+\|^2 + \frac{\mu_2}{2L_2(L_2 - \mu_2)} \|G^+ - L_2 \Delta x\|^2. \tag{6}$$

Summing them up and performing simplifications we get:

$$\Delta F(x) \geq \frac{1}{2L_1} \|G\|^2 + \frac{\mu_1}{2L_1(L_1 - \mu_1)} \|G - L_1 \Delta x\|^2 + \frac{1}{2L_2} \|G^+\|^2 + \frac{\mu_2}{2L_2(L_2 - \mu_2)} \|G^+ - L_2 \Delta x\|^2. \tag{7}$$

By writing $(Q_{i,j})$ for function f_2 with the iterates (x, x^+):

$$f_2(x) - f_2(x^+) - \langle g_2^+, \Delta x \rangle \geq \frac{1}{2L_2} \|G^+\|^2 + \frac{\mu_2}{2L_2(L_2 - \mu_2)} \|G^+ - L_2 \Delta x\|^2 \tag{8}$$

and summing it up with (6) we get:

$$\langle G^+, \Delta x \rangle \geq \frac{1}{L_2} \|G^+\|^2 + \frac{\mu_2}{L_2(L_2 - \mu_2)} \|G^+ - L_2 \Delta x\|^2. \tag{9}$$

Similarly, by writing $(Q_{i,j})$ for function f_1 with the iterates (x^+, x) and summing it up with (5) we get:

$$\langle G, \Delta x \rangle \geq \frac{1}{L_1} \|G\|^2 + \frac{\mu_1}{L_1(L_1 - \mu_1)} \|G - L_1 \Delta x\|^2. \tag{10}$$

The proofs only involve adjusting the right-hand side of (7) using either inequality (9) or inequality (10), weighted by scalars $\alpha > 0$, which depend on the curvatures (see Table 6). Specifically, to establish regimes $p_{1,3,4,5}$ we substitute α in:

$$\Delta F(x) \geq \frac{1}{2L_1} \|G\|^2 + \frac{\mu_1}{2L_1(L_1 - \mu_1)} \|G - L_1 \Delta x\|^2 +$$
$$\frac{1 + 2\alpha}{2L_2} \|G^+\|^2 + \frac{\mu_2(1 + 2\alpha)}{2L_2(L_2 - \mu_2)} \|G^+ - L_2 \Delta x\|^2 - \alpha \langle G^+, \Delta x \rangle \tag{11}$$

and to demonstrate regimes $p_{2,6}$ we plug in α in:

$$\Delta F(x) \geq \frac{1}{2L_2} \|G^+\|^2 + \frac{\mu_2}{2L_2(L_2 - \mu_2)} \|G^+ - L_2 \Delta x\|^2$$
$$\frac{1 + 2\alpha}{2L_1} \|G\|^2 + \frac{\mu_1(1 + 2\alpha)}{2L_1(L_1 - \mu_1)} \|G - L_1 \Delta x\|^2 - \alpha \langle G, \Delta x \rangle.$$

Since the proof is entirely based on algebraic manipulations, by exploiting the symmetry on the right-hand side of the two inequalities we focus on demonstrating the regimes p_1, p_3, p_4, p_5. Regimes p_2 and p_6 are complementary to p_1 and p_5, respectively, under the condition $\mu_1 \geq 0$; specifically, $p_1 \leftrightarrow p_2$ and $p_5 \leftrightarrow p_6$ (see also Table 1). Their proofs can be obtained by interchanging: (i) the curvature indices 1 and 2; and (ii) G and G^+.

Regime p_1: It corresponds to $L_1 \geq L_2 > \mu_1 \geq 0$; in particular, $L_1 \geq \max\{L_2, \mu_1, \mu_2\}$. By setting $\alpha = \frac{\mu_1}{L_2} \frac{L_1 - L_2}{L_1 - \mu_1}$ in (11), after simplifications and building the squares we get:

$$\Delta F(x) \geq \frac{L_2 - \mu_1}{L_2(L_1 - \mu_1)} \frac{\|G\|^2}{2} + \frac{\mu_1}{L_2(L_1 - \mu_1)} \frac{\|G - L_2 \Delta x\|^2}{2} + \frac{1}{L_2} \left[1 + \frac{\mu_1(L_1 - L_2)}{L_2(L_1 - \mu_1)} \right] \frac{\|G^+\|^2}{2} +$$
$$+ \frac{\mu_1 \frac{L_1}{L_2} \mu_2 [\mu_1^{-1} + \mu_2^{-1} + L_2^{-1} - L_1^{-1}(2 + \frac{L_2}{\mu_2})]}{(L_1 - \mu_1)(L_2 - \mu_2)} \frac{\|G^+ - L_2 \Delta x\|^2}{2}.$$

The weight of $\|G - L_2 \Delta x\|^2$ is positive $(\mu_1 \geq 0)$, while the weight of $\|G^+ - L_2 \Delta x\|^2$ is positive if either (i) $\mu_2 \geq 0$ or (ii) $\mu_2 < 0$ and $\mu_1^{-1} + \mu_2^{-1} + L_2^{-1} - L_1^{-1} \leq L_1^{-1}(1 + \frac{L_2}{\mu_2})$. Under these two cases, by neglecting both mixed terms we get:

$$\Delta F(x) \geq \frac{L_2 - \mu_1}{L_2(L_1 - \mu_1)} \frac{\|G\|^2}{2} + \frac{1}{L_2} \left(1 + \frac{L_2^{-1} - L_1^{-1}}{\mu_1^{-1} - L_1^{-1}} \right) \frac{\|G^+\|^2}{2},$$

with equality only if $G = G^+ = L_2 \Delta x$.

Regime p_3: It corresponds to $L_2 \geq \mu_1 \geq 0$ and negativity of the weight of $\|G^+ - L_2 \Delta x\|^2$ from the proof of p_1, i.e., $\mu_2 < 0$ and $\mu_1^{-1} + \mu_2^{-1} + L_2^{-1} - L_1^{-1} > L_1^{-1}(1 + \frac{L_2}{\mu_2})$. Since $L_2 \geq \mu_1 > -\mu_2$, we have $L_2 + \mu_2 > 0$. By setting $\alpha = \frac{-\mu_2}{L_2 + \mu_2}$ in (11), after simplifications and building up a square including all weight of $L_2 \Delta x$ we get:

$$\Delta F(x) \geq \frac{L_1^{-1}(\mu_1^{-1} + \mu_2^{-1} + L_2^{-1})}{\mu_1^{-1} + \mu_2^{-1} + L_2^{-1} - L_1^{-1}} \frac{\|G\|^2}{2} + \frac{1}{L_2 + \mu_2} \frac{\|G^+\|^2}{2} +$$

$$+ \frac{\mu_1 L_1 \mu_2 (\mu_1^{-1} + \mu_2^{-1} + L_2^{-1} - L_1^{-1})}{L_2(L_1 - \mu_1)(L_2 - \mu_2)} \frac{\left\| \frac{G}{\frac{L_1}{L_2} + \frac{\mu_2(L_1 - \mu_1)}{\mu_1(L_2 + \mu_2)}} - L_2 \Delta x \right\|^2}{2}.$$

Since $\mu_2 < 0$, the weight of the mixed term is nonnegative only if $\mu_1^{-1} + \mu_2^{-1} + L_2^{-1} - L_1^{-1} \leq 0$. Further on, this implies that the weight of $\|G\|^2$ is nonnegative only if $\mu_1^{-1} + \mu_2^{-1} + L_2^{-1} \leq 0$, which is exactly the threshold condition $B \leq 0$, with B defined in (3). Finally, by neglecting the mixed term we get:

$$\Delta F(x) \geq \frac{L_1^{-1}(\mu_1^{-1} + \mu_2^{-1} + L_2^{-1})}{\mu_1^{-1} + \mu_2^{-1} + L_2^{-1} - L_1^{-1}} \frac{\|G\|^2}{2} + \frac{1}{L_2 + \mu_2} \frac{\|G^+\|^2}{2},$$

with equality only if $G = L_2 \left[\frac{L_1}{L_2} + \frac{\mu_2(L_1 - \mu_1)}{\mu_1(L_2 + \mu_2)} \right] \Delta x$.

Regime p_4: With $\mu_2 < 0$, if $\mu_1^{-1} + \mu_2^{-1} + L_2^{-1} > 0$, then the proof for regime p_3 breaks. Then we set $\alpha = \frac{\mu_1 + \mu_2}{-\mu_2} > 0$ in (11). After simplifications and building up the squares we get:

$$\Delta F(x) \geq \frac{\mu_1 + \mu_2}{\mu_2^2} \frac{\|G^+\|^2}{2} + \frac{1}{L_1 - \mu_1} \frac{\|G - \mu_1 \Delta x\|^2}{2} + \frac{\mu_1 L_2}{-\mu_2} \frac{\mu_1^{-1} + \mu_2^{-1} + L_2^{-1}}{L_2 - \mu_2} \frac{\|G^+ - \mu_2 \Delta x\|^2}{2}.$$

The second mixed term has a positive weight if $L_2(\mu_1^{-1} + \mu_2^{-1} + L_2^{-1}) > 0$. Note that this condition allows $L_2 \leq 0$, bounded bellow through the necessary condition $L_2 > \mu_2 > -\mu_1$. After disregarding the mixed squares with positive weights, we obtain:

$$\Delta F(x) \geq \frac{\mu_1 + \mu_2}{\mu_2^2} \frac{\|G^+\|^2}{2},$$

which holds with equality only if $G = \mu_1 \Delta x$ and $G^+ = \mu_2 \Delta x$.

Regime p_5: Assume $0 < L_2 \leq \mu_1$, i.e., F is (strongly) convex, and $\mu_2(\mu_1^{-1} + \mu_2^{-1} + L_2^{-1}) \geq 0$, where either $\mu_2 \geq 0$ or $\mu_1 > -\mu_2 > 0$. By setting $\alpha = \frac{\mu_1}{L_2}$ in (11), after simplifications and building up the squares we get:

$$\Delta F(x) \geq \frac{\mu_1 + L_2}{L_2^2} \frac{\|G^+\|^2}{2} + \frac{1}{L_1 - \mu_1} \frac{\|G - \mu_1 \Delta x\|^2}{2} + \frac{\mu_1 \mu_2}{L_2} \frac{\mu_1^{-1} + \mu_2^{-1} + L_2^{-1}}{L_2 - \mu_2} \frac{\|G^+ - L_2 \Delta x\|^2}{2},$$

where the mixed term can be disregarded to obtain:

$$\Delta F(x) \geq \frac{L_2 + \mu_1}{L_2^2} \frac{\|G^+\|^2}{2},$$

which holds with equality only if $G = \mu_1 \Delta x$ and $G^+ = L_2 \Delta x$. $\qquad \square$

A.3 Proof of Corollary 1

Proof of Corollary 1. From Theorem 1, by taking the minimum between the (sub)gradients residual norms in (1) we get:

$$F(x) - F(x^+) \geq \sigma_i \frac{1}{2} \|g_1 - g_2\|^2 + \sigma_i^+ \frac{1}{2} \|g_1^+ - g_2^+\|^2$$

$$\geq p_i \frac{1}{2} \min\{\|g_1 - g_2\|^2, \|g_1^+ - g_2^+\|^2\},$$

where $p_i = \sigma_i + \sigma_i^+$, for $i = 1, \ldots, 6$, are given in Table 1. The rate (2) results by telescoping the above inequality for N iterations and taking the minimum among all (sub)gradients residual norms. A rate with respect to F_{lo}

is obtained either by applying the trivial bound $F(x^0) - F(x^N) \geq F(x^0) - F_{lo}$ or, if $L_1 > \mu_2$, by using a tighter bound like the one demonstrated by Abbaszadehpeivasti et al. (2023, Lemma 2.1):

$$F(x^N) - F_{lo} \geq \frac{1}{2(L_1 - \mu_2)} \|g_1^N - g_2^N\|^2.$$

By incorporating this into the telescoped sum and once again taking the minimum over all (sub)gradients residual norms, we obtain the second rate from Corollary 1. Furthermore, note that a necessary condition for the tightness of these sublinear rates is that the subgradients residual norms $\|g_1^k - g_2^k\|$ must be equal for any $k = 0, \dots, N$. □

A.4 Proofs of Proposition 3 and Proposition 4

Proof of Proposition 3. The approach is to reformulate a PGD iteration as a DCA one, starting from the proximal step definition:

$$
\begin{aligned}
x^+ &= \mathrm{prox}_{\gamma h}\left[x - \gamma \nabla \varphi(x)\right] \\
&= \underset{w \in \mathbb{R}^d}{\mathrm{argmin}}\left\{h(w) + \frac{1}{2\gamma}\|w - x + \gamma \nabla \varphi(x)\|^2\right\} \\
&= \underset{w \in \mathbb{R}^d}{\mathrm{argmin}}\left\{h(w) + \frac{\|w\|^2}{2\gamma} + \langle \nabla \varphi(x) - \frac{1}{\gamma}x, w\rangle\right\} \\
&= \underset{w \in \mathbb{R}^d}{\mathrm{argmin}}\left\{\left[h(w) + \frac{\|w\|^2}{2\gamma}\right] - \langle \nabla\left[\frac{\|x\|^2}{2\gamma} - \varphi(x)\right], w\rangle\right\}.
\end{aligned}
$$

Then, from $f_1 = h + \frac{1}{2\gamma}\|\cdot\|^2$ and $f_2 = \frac{1}{2\gamma}\|\cdot\|^2 - \varphi$, we recover the (DCA) iteration for f_2 smooth. □

Proof of Proposition 4. For any $x \in \mathbb{R}^d$, let $g_h \in \partial h(x)$ and define $G := \nabla \varphi(x) + g_h \in \partial F(x)$. By Proposition 3, we have that $g_1 := g_h + \frac{x}{\gamma} \in \partial f_1(x)$. Moreover, the smoothness of φ implies the one of f_2. Then the following holds:

$$G = \nabla \varphi(x) + g_h = \left[\frac{x}{\gamma} - \nabla f_2(x)\right] + \left[g_1 - \frac{x}{\gamma}\right] = g_1 - \nabla f_2(x).$$

Consequently, $\|G\|^2 = \|\nabla \varphi(x) + g_h\|^2 = \|g_1 - \nabla f_2(x)\|^2$. Since F is unchanged between the two splittings, the conclusion follows. □

B TIGHTNESS ANALYSIS (see Conjectures 1 and 2)

For each of the six regimes denoted p_i, with $i = 1, \dots, 6$, we provide numerical examples of the corresponding worst-case values, defined as $\mathrm{wc}(N) = \frac{1}{2}\min_{0 \leq k \leq N}\{\|g_1^k - \nabla f_2(x^k)\|^2\}$, where $g_1^k \in \partial f_1(x^k)$ and the initial condition is fixed to $F(x^0) - F(x^N) \leq 1$. See Figure 2 for regime p_1 when f_2 is convex or weakly convex, Figure 3 for regimes p_2 and p_3, Figure 4 for regime p_4 with $L_2 > 0$ or $L_2 < 0$ and Figure 5 for regimes p_5 and p_6. The rates from Corollary 1 are predicted to be sublinear, with $\mathrm{wc}(N) = \frac{1}{p_i N}$. This holds for regimes p_1, p_2 and p_3, as stated in Conjecture 2. For regimes p_4, p_5 and p_6, however, the tight rates can be improved and further investigation is required.
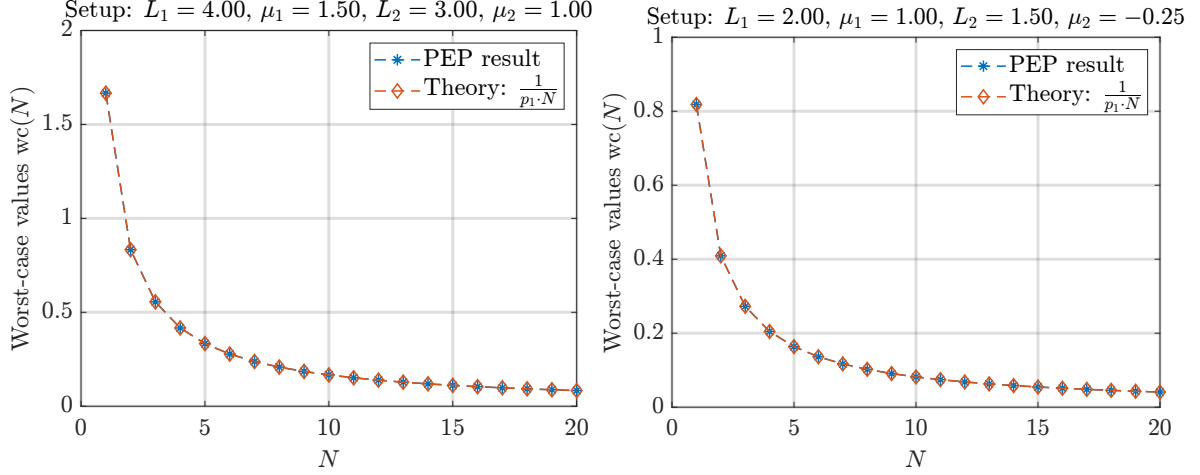
Figure 2: Examples for regime p_1 with f_2 convex (*left*) and f_2 weakly convex (*right*), showing exactness of our expressions.
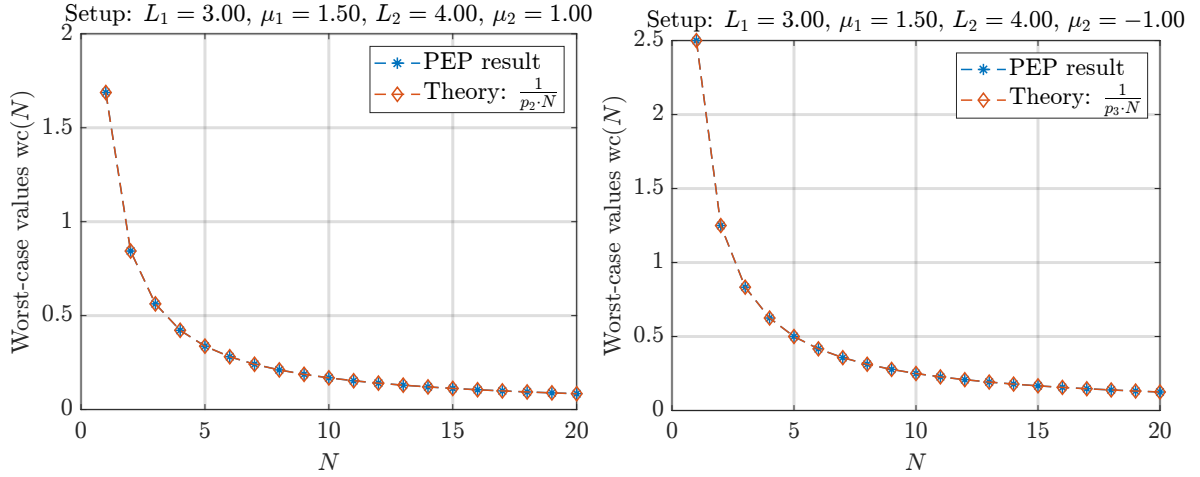

Figure 3: Examples for regimes p_2 (*left*) and p_3 (*right*), showing exactness of our expressions.
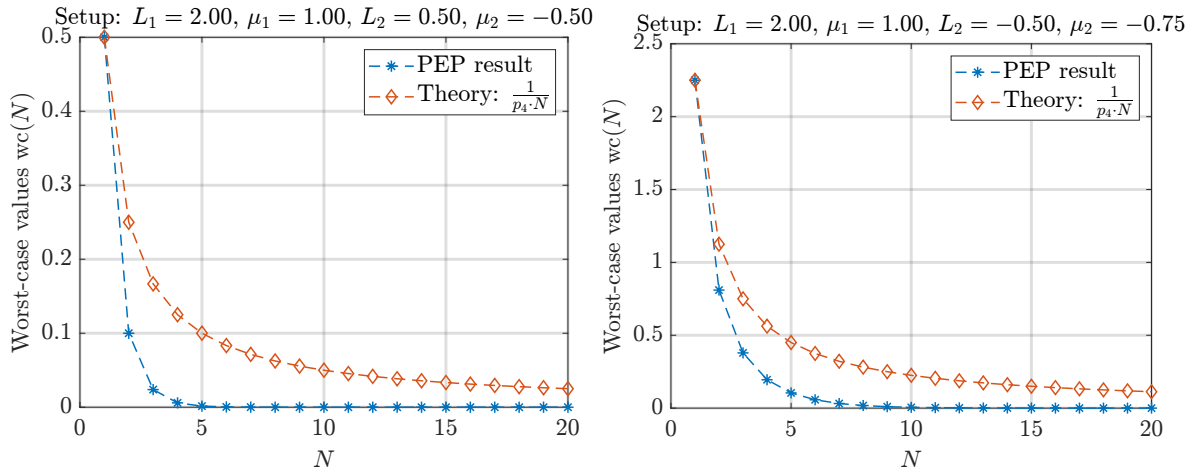

Figure 4: Examples for regime p_4, with $L_2 > 0$ (*left*) and $L_2 < 0$ (*right*), showing that our expressions are non-tight for $N > 1$ iterations.
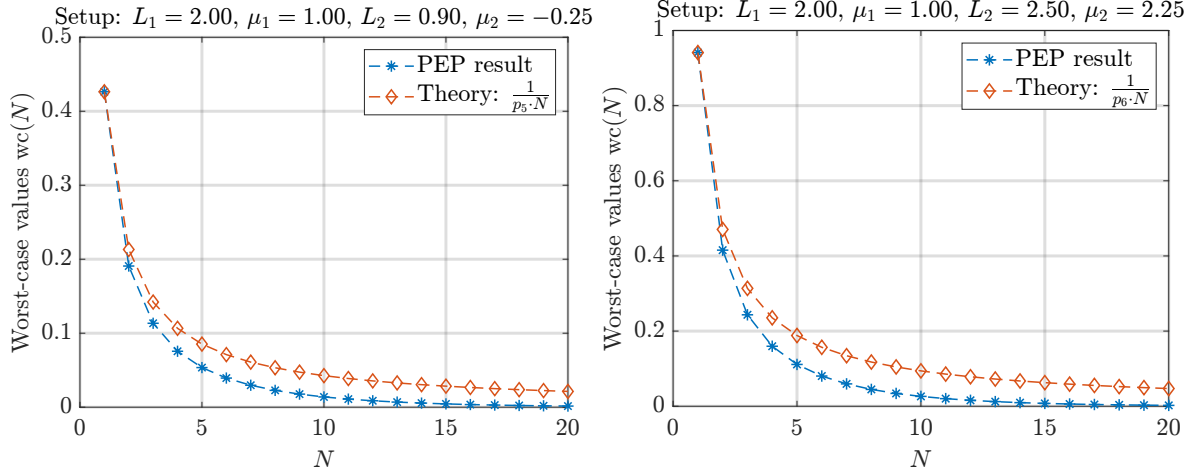
Figure 5: Examples for regimes p_5 (*left*) and p_6 (*right*), showing that our expressions are non-tight for $N > 1$ iterations.

The proof of Theorem 1 (see Appendix A.2) provides necessary conditions for the worst-cases corresponding to each of the six regimes. Using these, we elaborate on Conjecture 2 about the tightness of regimes p_1 and p_2, providing analytical worst-case functions for the cases when both f_1 and f_2 are smooth. The nonsmooth case is discussed in Abbaszadehpeivasti et al. (2023, Example 3.1) for regime p_2, with $f_1 \in \mathcal{F}_{0,L_1}$ and $f_2 \in \mathcal{F}_{0,\infty}$.

B.1 Worst-Case Function Examples for Regime p_1

Proposition 5 (Tightness of p_1). *Let f_1^0, f_2^0, g_2^0, x^0 be some real numbers such that $\Delta := f_1^0 - f_2^0 > 0$ and let $N \geq 1$ be an integer. Consider μ_1, L_1, μ_2, L_2 as belonging to the domain of regime p_1 from Table 1, where the conditions are $0 \leq \mu_1 < L_2 \leq L_1 < \infty$, $\mu_2 < L_1$ and either (i) $\mu_2 \geq 0$ or (ii) $\mu_1 > -\mu_2 > 0$ and $E = \frac{L_2 + \mu_2}{L_1 L_2} \frac{L_2 - L_1}{-\mu_2} + \mu_1^{-1} - L_1^{-1} \leq 0$. Let $U := -\sqrt{\frac{2\Delta}{p_1 N}}$, with expression of p_1 given in Table 1. Consider the functions $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ defined as:*

$$f_2(x) := \tfrac{1}{2} L_2 (x - x^0)^2 + g_2^0 (x - x^0) + f_2^0$$

$$f_1(x) := \begin{cases} \tfrac{1}{2} L_1 (x - x^0)^2 + g_1^0 (x - x^0) + f_1^0 & x \in (-\infty, x^0]; \\[2mm] \tfrac{1}{2} L_1 (x - x^k)^2 + g_1^k (x - x^k) + f_1^k & x \in [x^k, \bar{x}^k]; \\[2mm] \tfrac{1}{2} \mu_1 (x - x^{k+1})^2 + g_1^{k+1} (x - x^{k+1}) + f_1^{k+1} & x \in [\bar{x}^k, x^{k+1}]; \\[2mm] \tfrac{1}{2} L_1 (x - x^N)^2 + g_1^N (x - x^N) + f_1^N & x \in [x^N, \infty), \end{cases}$$

with $k = 0, 1, \ldots, N-1$, $\nabla f_2(x^0) = g_2^0$, $f_1(x^k) = f_1^k$, $\nabla f_1(x^k) = g_1^k$ and:

$$x^k = x^0 - k \frac{U}{L_2}, \quad \forall k = 0, \ldots, N;$$

$$g_1^k = g_2^0 - (k-1) U, \quad \forall k = 0, \ldots, N;$$

$$f_1^k = f_2(x^k) + \frac{N-k}{N} \Delta, \quad \forall k = 0, \ldots, N;$$

$$\bar{x}^k = x^k - \frac{L_2 - \mu_1}{L_1 - \mu_1} \frac{U}{L_2}, \quad \forall k = 0, \ldots, N-1.$$

Then by performing N iterations of (DCA) on the function $F(x) = f_1(x) - f_2(x)$, starting from x^0, the following result holds:

$$\tfrac{1}{2} \min_{0 \leq k \leq N} \left\{ \| \nabla f_1(x^k) - \nabla f_2(x^k) \|^2 \right\} = \frac{\Delta}{p_1(L_1, L_2, \mu_1, \mu_2) N}.$$
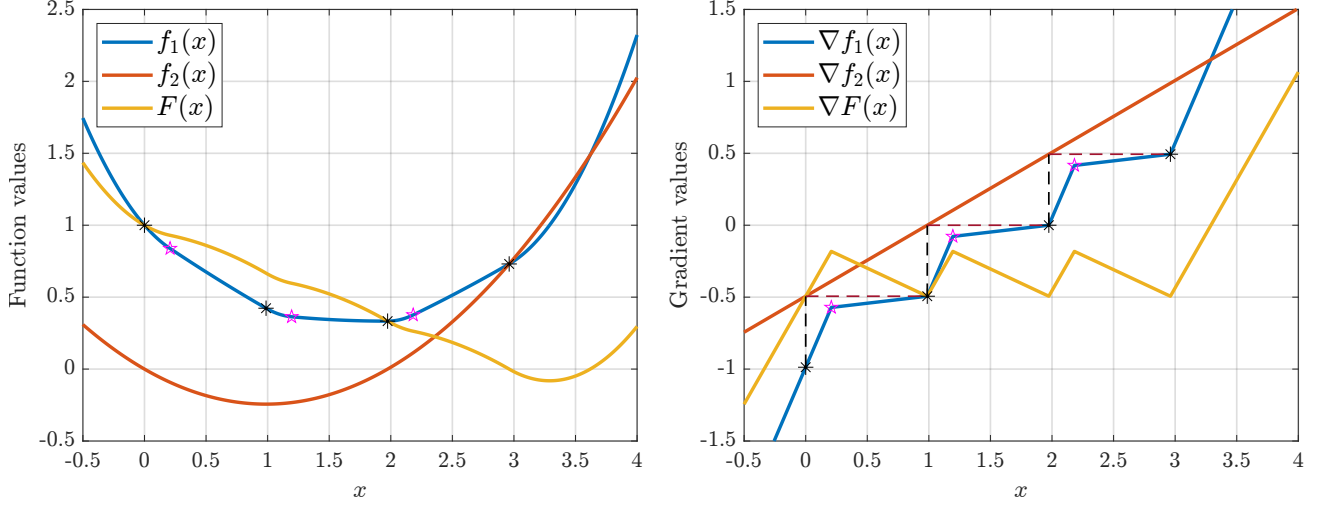
Figure 6: Worst-case example for regime p_1 after $N = 3$ iterations, showing the function values (*left*) and the gradient values (*right*). Setup: $L_1 = 2$, $\mu_1 = 0.1$, $L_2 = 0.5$, $\mu_2 = -0.01$. The initial condition is $\Delta = f_1(x^0) - f_2(x^0) = 1$. The iterations begin at $x^0 = 0$ where $f_1(x^0) = 1$ and $f_2(x^0) = 0$. The *black* stars represent the DCA iterations x^k, while the *magenta* pentagrams indicate the inflection points of f_1, denoted by $\bar{x}^k$, placed between consecutive iterations x^k.

Proof. By construction, $f_1 \in \mathcal{F}_{\mu_1, L_1}$ and $f_2 \in \mathcal{F}_{\mu_2, L_2}$, where μ_2 can be any real number such that $\mu_2 < L_2$. Moreover, for all $k = 0, 1, \ldots, N - 1$ we have $g_2^k = g_1^{k+1}$, which is the necessary condition of solving the DCA iteration. Consequently, the quantity $\|g_1^k - g_2^k\|^2$ is exactly $U^2 = \frac{2\Delta}{p_1 N}$, where $\Delta = F(x^0) - F(x^N)$. Given that f_2 is quadratic and f_1 is piecewise quadratic, starting from the initial point x^0, the iterations x^k are uniquely determined by applying (DCA). One can verify that x^k and $\bar{x}^k$ are inflection points for f_1, the curvature changing between μ_1 and L_1. The inflection points $\bar{x}^k$ result from ensuring continuity of both the function and its gradient values and one can verify that $\bar{x}^k \in [x^k, x^{k+1}]$ for all $k = 0, \ldots, N - 1$. $\qquad\square$

The worst-case example from Proposition 5 builds on the proof of regime p_1 (see Appendix A.2). Firstly, condition $G = G^+ = L_2 \Delta x$ implies $g_1^k - g_2^k = L_2(x^k - x^{k+1})$ at each iteration $k = 0, \ldots, N - 1$. Additionally, the worst-case after N iterations implies all gradient norms equal with value of $|U|$ from Proposition 5.

DCA iterations. The plot on the *right* in Figure 6 provides an intuitive illustration of the DCA iteration process. Starting from $x^0 = 0$, the iteration moves vertically along the y-axis until intersecting the graph of ∇f_2. Next, it moves horizontally along the x-axis until reaching the graph of ∇f_1 at x^1. This procedure is then repeated for each subsequent iteration.

B.2 Worst-Case Function Examples for Regime p_2

Proposition 6 (Tightness of p_2). *Let f_1^0, f_2^0, g_2^0, x^0 be some real numbers such that $\Delta := f_1^0 - f_2^0 > 0$ and let $N \geq 1$ be an integer. Consider μ_1, L_1, μ_2, L_2 as belonging to the domain of regime p_2 from Table 1, where the conditions are $\mu_1, \mu_2 \geq 0$ and $\max\{\mu_1, \mu_2\} < L_1 < L_2 < \infty$. Let $U := -\sqrt{\frac{2\Delta}{p_2 N}}$, where the expression of p_2 is*
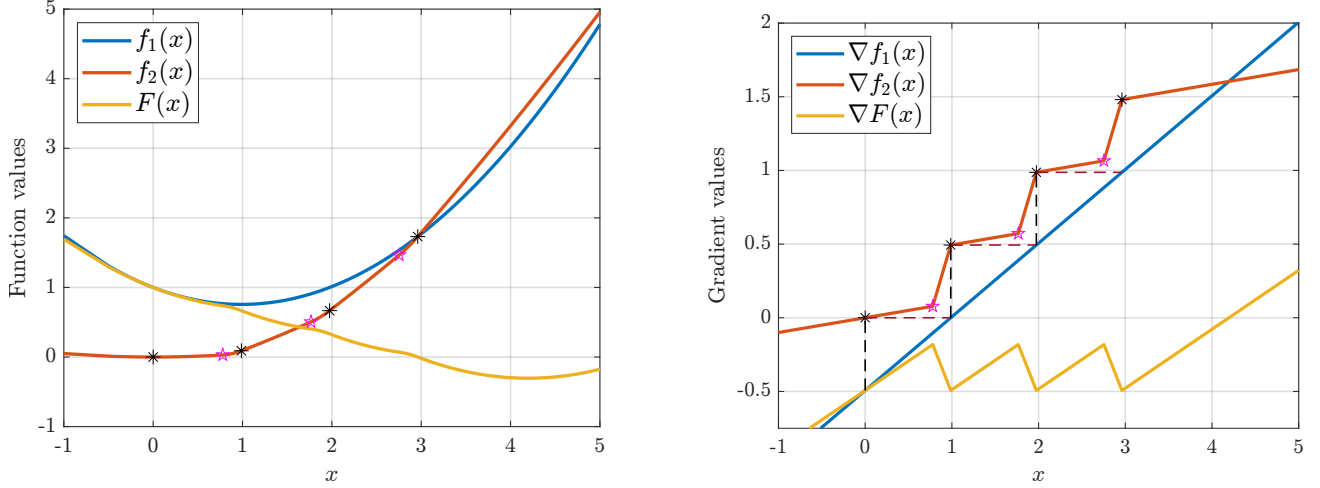
Figure 7: Worst-case example for regime p_2 after $N = 3$ iterations, showing the function values (*left*) and the gradient values (*right*). Setup: $L_1 = 1.5$, $\mu_1 = 0.25$, $L_2 = 2$, $\mu_2 = 1$. The initial condition is $\Delta = f_1(x^0) - f_2(x^0) = 1$. The iterations begin at $x^0 = 0$ where $f_1(x^0) = 1$ and $f_2(x^0) = 0$. The *black* stars represent the DCA iterations x^k, while the *magenta* pentagrams indicate the inflection points of f_2, denoted by $\bar{x}^k$, placed between consecutive iterations x^k. The dashed lines show the DCA iterations $x^{k+1} = \nabla f_1^*(\nabla f_2(x^k))$.

given in Table 1. *Consider the functions* $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ *defined as:*

$$f_1(x) := \tfrac{1}{2}L_1(x - x^0)^2 + g_1^0(x - x^0) + f_1^0$$

$$f_2(x) := \begin{cases} \tfrac{1}{2}\mu_2(x - x^0)^2 + g_2^0(x - x^0) + f_2^0 & x \in (-\infty, x^0]; \\[2mm] \tfrac{1}{2}\mu_2(x - x^k)^2 + g_2^k(x - x^k) + f_2^k & x \in [x^k, \bar{x}^k]; \\[2mm] \tfrac{1}{2}L_2(x - x^{k+1})^2 + g_2^{k+1}(x - x^{k+1}) + f_2^{k+1} & x \in [\bar{x}^k, x^{k+1}]; \\[2mm] \tfrac{1}{2}\mu_2(x - x^N)^2 + g_2^N(x - x^N) + f_2^N & x \in [x^N, \infty), \end{cases} \tag{12}$$

where for all $k = 0, 1, \ldots, N - 1$ *it holds* $\nabla f_1(x^0) = g_1^0$, $f_2(x^k) = f_2^k$, $\nabla f_2(x^k) = g_2^k$ *and:*

$$x^k = x^0 - k\frac{U}{L_1}, \quad \forall k = 0, \ldots, N;$$

$$g_2^k = g_1^0 - (k+1)U, \quad \forall k = 0, \ldots, N;$$

$$f_2^k = f_1(x^k) - \frac{N-k}{N}\Delta, \quad \forall k = 0, \ldots, N;$$

$$\bar{x}^k = x^k - \frac{L_2 - L_1}{L_2 - \mu_2}\frac{U}{L_1}, \quad \forall k = 0, \ldots, N - 1.$$

Then by performing N iterations of (DCA) *on the function* $F(x) = f_1(x) - f_2(x)$, *starting from* x^0, *the following result holds:*

$$\tfrac{1}{2}\min_{0 \le k \le N}\{\|\nabla f_1(x^k) - \nabla f_2(x^k)\|^2\} = \frac{\Delta}{p_2(L_1, L_2, \mu_1, \mu_2)N}. \tag{13}$$

The proof of Proposition 6 is similar to the one of Proposition 5.

Regime p_2 can be understood as a transformation of regime p_1. Specifically, regime p_2 appears when the (DCA) iterations are applied in reverse on the function $-F(x) = f_2(x) - f_1(x)$, with the roles of f_1 and f_2 interchanged. Thus, the dynamic of regime p_1 is mirrored in regime p_2.

C   DECISION TREE FOR IDENTIFYING THE ACTIVE REGIME

Figure 8 shows a decision tree outlining how to identify the corresponding regime from Theorem 1 based on the given curvature parameters.
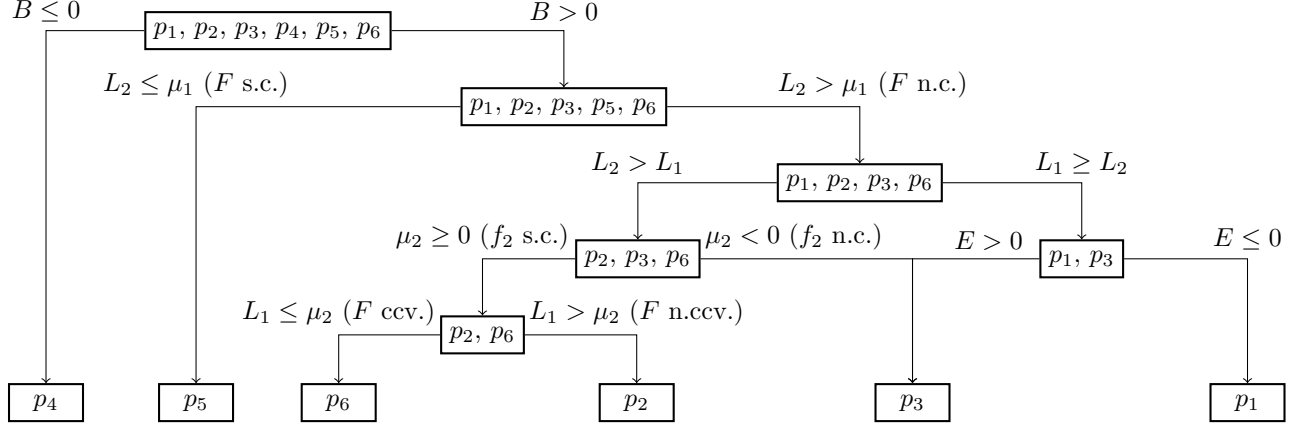

Figure 8: Decision tree on selecting regimes after one iteration defined in Theorem 1. Abbreviations – s.c.: strongly convex; n.c.: nonconvex; ccv.: concave; n.ccv.: nonconcave. Recall that $F = f_1 - f_2$, where we assume $f_1 \in \mathcal{F}_{\mu_1, L_1}$ and $f_2 \in \mathcal{F}_{\mu_2, L_2}$, with $\mu_1 + \mu_2 > 0$ or $\mu_1 = \mu_2 = 0$, and use the notation: $B := \mu_1^{-1} + \mu_2^{-1} + L_2^{-1}$ and $E := \frac{L_2 + \mu_2}{L_1 L_2} \frac{L_2 - L_1}{-\mu_2} + \mu_1^{-1} - L_1^{-1}$.

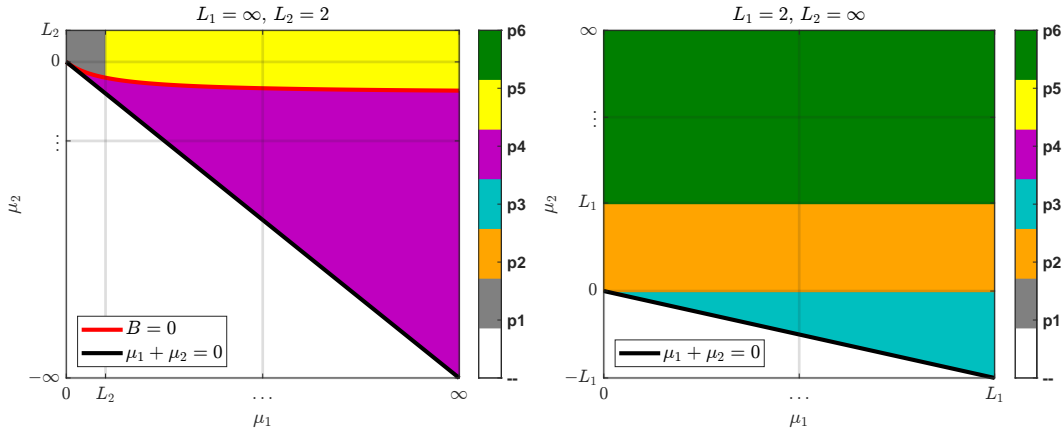## D   SKETCH OF REGIMES IN THE NONSMOOTH CASES (see Corollary 2)

Figure 9: All regimes after one (DCA) iteration when either f_1 or f_2 is nonsmooth (see Table 2 for precise expressions). The denominator values p_i are identical for regimes p_1 and p_5, as well as for regimes p_2 and p_6. Furthermore, when $L_2 = \infty$ (indicating F (strongly) concave), the threshold $B = \mu_1^{-1} + \mu_2^{-1} + L_2^{-1} = 0$ and the limit line $\mu_1 + \mu_2 = 0$ coincide.

E   ON THE BEST CURVATURE SPLITTING PROBLEM (see Section 4)

In Section 4 we examined how to achieve the optimal splitting for a given objective $F = f_1 - f_2$ by subtracting the same curvature term $\lambda \frac{\|\cdot\|^2}{2}$ in both functions. We concluded that if $\mu_1 \leq \mu_2$, the best splitting is achieved by shifting the lower curvature of f_1 to 0, i.e., make it convex. Conversely, when $\mu_1 > \mu_2$, the best splitting is achieved for some f_2 weakly convex, with lower curvature $\mu_2 - \lambda < 0$.

Within this section, we provide additional numerical experiments. In Figure 10 we provide an example with fixed curvature bounds of a nonconvex-nonconcave objective function F, namely $\mu_F = -0.5$ and $L_F = 1.5$, and illustrate all possible regimes after one iteration. Note that $\mu_F = \mu_1 - L_2$ and $L_F = L_1 - \mu_2$; further on, we examine the regimes based on the ranges of L_2 and μ_2. The condition $\mu_1 \geq 0$ implies that $L_2 = \mu_1 - \mu_F \geq -\mu_F$, while the condition $\mu_1 + \mu_2 > 0$ that $\mu_2 > -\mu_1 = L_2 + \mu_F$. The contour lines represent the values of the denominators p_i, with $i = 1, 2, 3, 4$. The *red* points mark the initial curvature values of L_2 and μ_2, whereas the *green* dots indicate the points with the largest possible p_i obtained through the optimal choice of λ. Since these shifts are linear in λ, the dashed lines connecting the dots have a slope of one.

For example, in the case where $L_2 = 1$ and $\mu_2 = 0.75$, we have $\mu_1 < \mu_2$ and the best splitting is obtained by shifting to the lowest possible value of L_2, corresponding to $\mu_1 = 0$. In all other examples, $\mu_1 > \mu_2$ and the optimal splittings are found within regime p_3, where $\mu_2 < 0$.
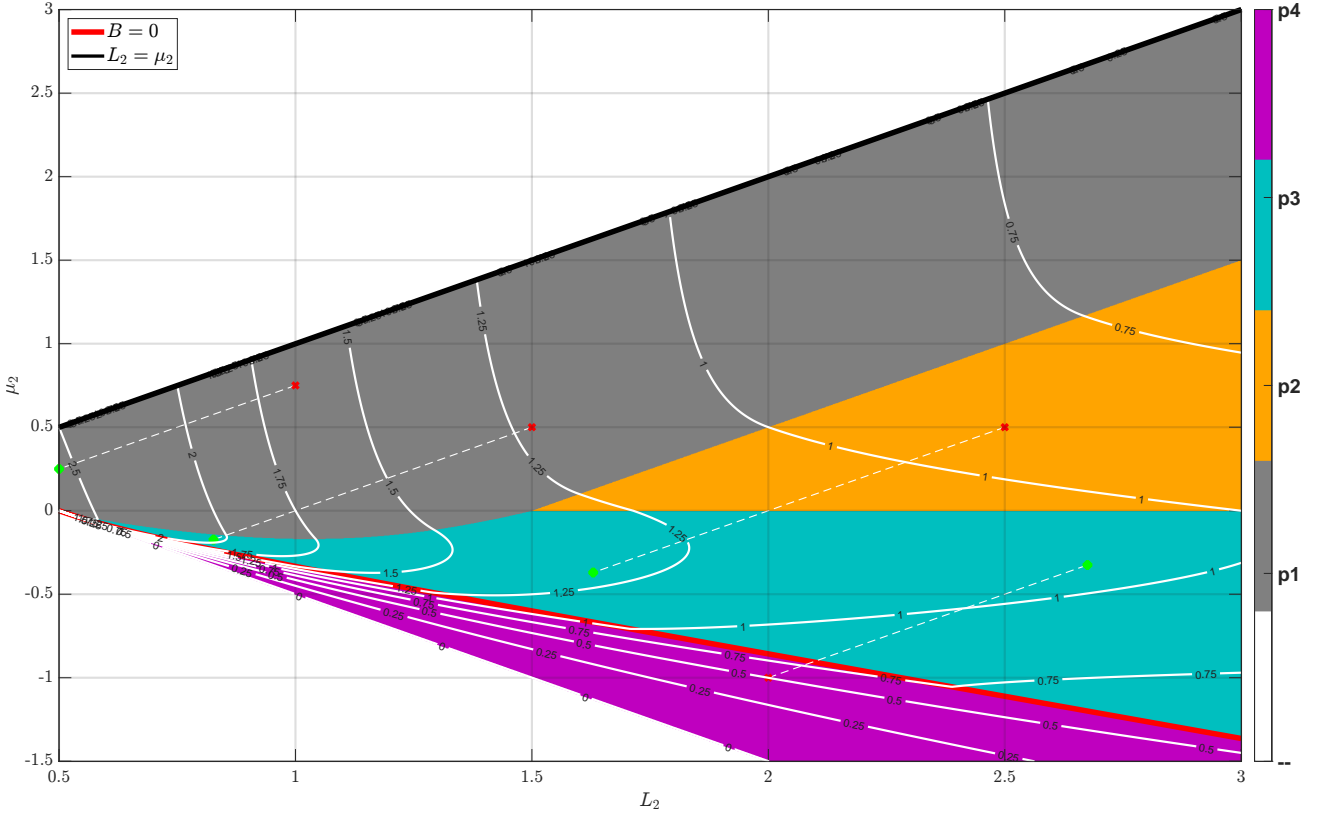

Figure 10: All regimes for a fixed objective function F with $\mu_F = -0.5$ and $L_F = 1.5$, along with several mappings of the optimal splittings, shown as transitions from *red* to *green* dots along dashed lines with a slope of 1.

F   NUMERICAL EXPERIMENTS (extends Section 7)

Recall the for any $x \in \mathbb{R}^n$ we define $f_1(x) = \kappa\|x\|_1 + \eta\frac{\|x\|^2}{2} + \delta_{\bar{B}(0,1)}(x)$ and $f_2(x) := \frac{1}{2}x^T\Sigma x$, where $\kappa > 0$, $\eta \geq 0$, $\delta_{\bar{B}(0,1)}$ is the indicator function over the closed Euclidean unit ball, $\Sigma = A^T A$ is the sample covariance matrix, where $A \in \mathbb{R}^{20n \times n}$, with $n = 200$ and 10% sparsity. Note that $\mu_1 = \eta$, $L_1 = \infty$, $\mu_2 = \min\{\Lambda(\Sigma)\}$ and $L_2 = \max\{\Lambda(\Sigma)\}$, where $\Lambda(\Sigma)$ denotes the eigenvalues of Σ. Recall $\tilde{f}_i^\lambda = f_i - \lambda\frac{\|\cdot\|^2}{2}$, with $i = \{1, 2\}$, where $\lambda \in (-\infty, \eta]$ is the curvature shift, being the curvature adjusted functions.

F.1   Closed-Form Expression of the $\partial\tilde{f}_1^{\lambda^*}(y)$

Firstly, for each $y \in \mathbb{R}^n$ we compute $\partial f_1^*(y) = \mathrm{argmin}_{x \in \mathbb{R}^n}\{\langle x, y\rangle - f_1(x)\}$. Let $f(x) := \kappa\|x\|_1 + \delta_{\bar{B}(0,1)}(x)$, such that $f_1 = f + \eta\frac{\|\cdot\|^2}{2}$ and $\tilde{f}_1^\lambda = f + (\eta - \lambda)\frac{\|\cdot\|^2}{2}$. From Themelis et al. (2020b, Section V) we have:

$$\mathrm{prox}_{\gamma f}(y) = \frac{\mathrm{sgn}(y) \odot [|y| - \kappa\gamma\mathbf{1}]_+}{\max\left\{1, \,\|\,[|y| - \kappa\gamma\mathbf{1}]_+\,\|_2\right\}},$$

where $\gamma > 0$, $[\cdot]_+$ denotes the positive part, $\odot$ is the elementwise multiplication, and $\mathbf{1}$ is the vector of all ones.

Case $\eta - \lambda > 0$. $\tilde{f}_1^\lambda$ is strongly convex and the subdifferential of the convex conjugate is a singleton. Using the properties of the proximal operator, we get:

$$\nabla\tilde{f}_1^{\lambda^*}(y) = \mathrm{prox}_{(\eta - \lambda)^{-1}f}\left((\eta - \lambda)^{-1}y\right).$$

Case $\eta - \lambda = 0$. We get $\tilde{f}_1^\lambda = f$, which is convex. We obtain

$$\partial f^*(y) = \begin{cases} \mathbf{0}, & \text{if } y = \mathbf{0} \\ \mathbf{0} \cup \left\{x \in \bar{B}(0,1) : \mathrm{sgn}(x_i) = \mathrm{sgn}(y_i), \forall i = 1, \dots, n\right\}, & \text{if } |y_i| = \kappa, \forall i = 1, \dots, n; \\ \frac{\mathrm{sgn}(y)\odot[|y| - \kappa\mathbf{1}]_+}{\|[|y| - \kappa\mathbf{1}]_+\|_2}, & \text{otherwise}, \end{cases}$$

where $\mathbf{0}$ is the vector of all zeros. Putting together, for all $\lambda \leq \eta$ it holds:

$$\partial\tilde{f}_1^{\lambda^*}(y) = \begin{cases} \mathbf{0}, & \text{if } \lambda \leq \eta, \quad y = \mathbf{0}; \\ \mathbf{0} \cup \left\{x \in \bar{B}(0,1) : \mathrm{sgn}(x_i) = \mathrm{sgn}(y_i), \forall i = 1, \dots, n\right\}, & \text{if } \lambda = \eta, \quad |y_i| = \kappa, \forall i = 1, \dots, n; \\ \frac{\mathrm{sgn}(y)\odot[|y| - \kappa\mathbf{1}]_+}{\max\{\eta - \lambda, \|[|y| - \kappa\mathbf{1}]_+\|_2\}}, & \text{otherwise}. \end{cases}$$

The subdifferential is a singleton unless $\lambda = \eta$ and y is on the boundary of the l_∞-ball of radius κ. In this case, in the numerical experiments we select the subgradient $\mathbf{0} \in \partial\tilde{f}_1^{\lambda^*}(y)$.

F.2   Extended Experiment Results

Let $\bar{M} \leq M = 1000$ be the number of runs converging to the same solution (non-trivial and with the desired sparsity) and denote by N_ε the average number of iterations to reach a certain accuracy $\varepsilon = \{10^{-1}, 10^{-2}, \dots, 10^{-12}\}$, defined as $N_\varepsilon := \frac{1}{\bar{M}}\sum_{d=1}^{\bar{M}}\mathrm{argmin}_{0 \leq k \leq N}\{\|\tilde{g}_1^k - \nabla\tilde{f}_2^\lambda(x^k)\|^2 \leq \varepsilon\}$.

Case 1: $\eta > \mu_2$. Parameters: $\eta = \mu_1 = 0.5$, $\kappa = 0.02$ (400 runs kept out of 1000). Here, $\lambda^* = 0.4413$ and the maximum *theoretical* shift is $\lambda_{\max} = \frac{\mu_1 + \mu_2}{2} = 0.4441$. With $\lambda > \mu_2 = 0.3882$, $\tilde{f}_2^\lambda$ becomes weakly convex. The results are reported in Table 7. Using λ^* achieves at least a twofold acceleration compared to $\lambda = 0$. Higher λ values, as $\lambda_{\max}$, making $\tilde{f}_2^\lambda$ even more nonconvex, further improve the algorithm. Conversely, adding curvature to both functions ($\lambda < 0$) slows the convergence.

Case 2: $\eta < \mu_2$. Parameters: $\eta = 0.2$, $\kappa = 0.02$ (703 runs kept out of 1000). We get $\lambda^* = \mu_1 = 0.2$. Using λ^* to make $\tilde{f}_1^\lambda$ convex improves the convergence speed by approximately 20% compared to the initial splitting. Generally, decreasing the curvature of f_2 improves the convergence.

Table 7: **(Case 1)** Average number of iterations N_ε to reach ε accuracy of the minimum subgradient norm in the setting $\eta = \mu_1 = 0.5$, $\kappa = 0.02$, corresponding to $\mu_1 > \mu_2$. Lowest values are obtained for $\lambda = \lambda_{\max} = 0.4441$, whereas $\lambda^* = 0.4413$ performs the second best.

ε \ λ	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}	10^{-11}	10^{-12}
0	1.00	5.74	17.73	46.70	96.70	150.17	242.57	336.27	439.46	544.63	651.33	758.74
$\lambda^* = 0.4413$	1.00	3.65	8.44	21.49	44.76	70.75	112.57	154.93	200.98	247.93	295.43	343.06
$-0.5\lambda^* = -0.2207$	1.00	6.83	22.34	59.32	122.73	189.31	307.17	426.39	558.51	693.40	830.17	967.74
$-\lambda^* = -0.4413$	1.00	8.05	27.00	72.30	147.77	229.13	375.34	521.24	681.09	844.69	1010.64	1177.50
$0.5\lambda^* = 0.2207$	1.00	4.54	13.06	33.91	70.19	108.43	176.35	244.05	318.63	394.92	472.42	550.15
$\lambda_{\max} = 0.4441$	1.00	3.63	8.35	21.23	44.01	70.38	111.91	153.78	199.95	246.61	293.76	340.99

Table 8: **(Case 2)** Average number of iterations N_ε to reach ε accuracy of the minimum subgradient norm in the setting $\eta = 0.2$, $\kappa = 0.02$, corresponding to $\eta < \mu_2$. Lowest values are obtained for $\lambda^* = \mu_1 = 0.2$.

| ε \ λ	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}	10^{-11}	10^{-12}
0	1.04	5.88	19.21	49.85	99.16	153.46	245.40	339.67	442.23	547.61	654.99	763.01
$\lambda^* = 0.2$	1.04	4.32	14.23	37.10	75.09	115.90	185.48	256.33	332.81	412.39	492.94	574.18
$-0.5\lambda^* = -0.1$	1.04	6.48	21.62	55.52	110.98	172.24	276.85	382.11	498.04	616.29	737.17	859.14
$-\lambda^* = -0.2$	1.04	7.06	23.87	61.44	123.41	192.46	307.80	426.00	554.21	686.04	820.61	956.00
$0.5\lambda^* = 0.1$	1.04	5.26	16.83	43.47	87.38	135.39	215.03	298.14	387.83	479.93	574.02	668.76
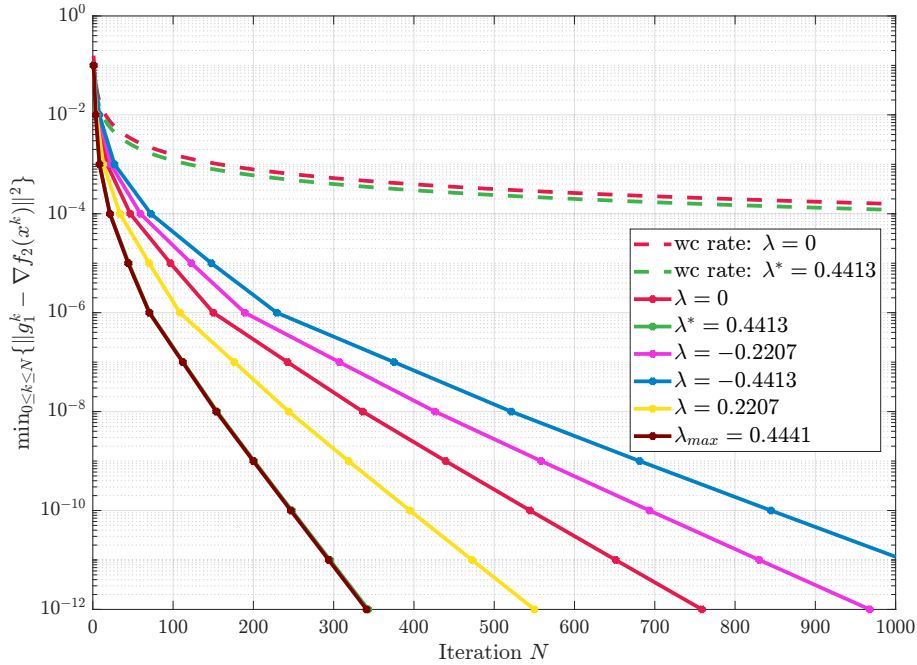
Figure 11: **Case 1.** Setting: $\eta = 0.5 > \mu_2$, $\kappa = 0.02$, $\lambda = \{0, \pm\lambda^*, \pm0.5\lambda^*, \lambda_{\max}\}$. Lowest values are obtained for $\lambda_{\max} = 0.4441$. Average number of iterations N_ε to reach ε accuracy of the best subgradient norm $\|g_1^k - \nabla f_2(x^k)\|$, where $g_1^k \in \partial f_1(x^k)$. Although the actual rate exceeds theoretical guarantees (*wc rate*), there is correlation between optimizing the worst-case for λ and better performance in the experiment.

G GITHUB REPOSITORY

We provide in the GitHub repository MATLAB scripts to support the numerical conjectures and to reproduce all the simulations, figures and tables presented in the paper.
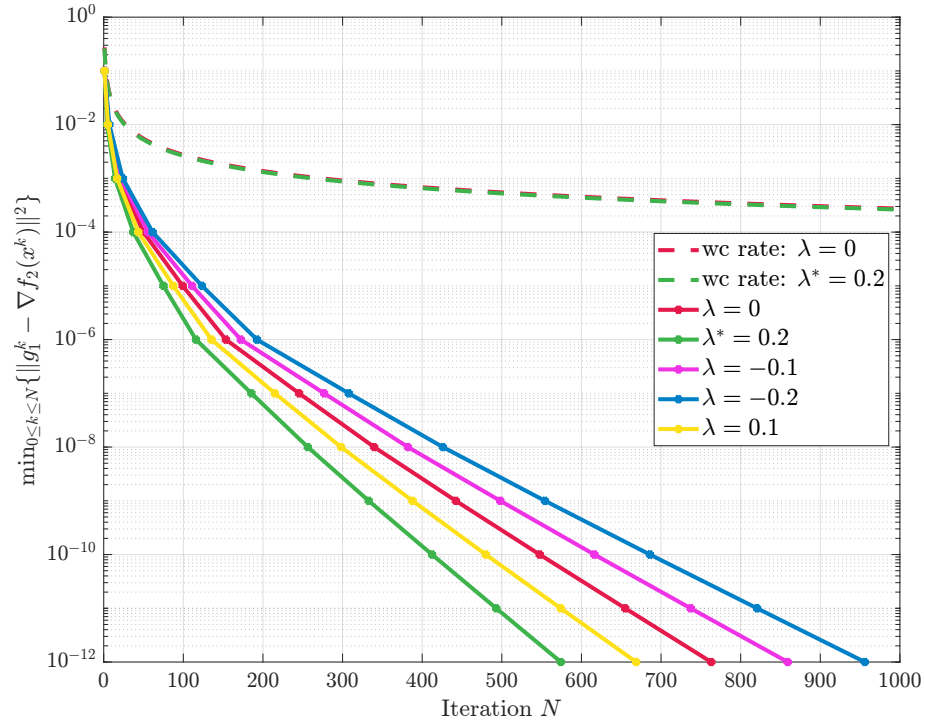
Figure 12: **Case 2.** Setting: $\eta = 0.2 = \mu_1 < \mu_2$, $\kappa = 0.02$, $\lambda = \{0, \pm\lambda^*, \pm 0.5\lambda^*\}$. Average number of iterations N_ε to reach ε accuracy of the best subgradient norm $\|g_1^k - \nabla f_2(x^k)\|$, where $g_1^k \in \partial f_1(x^k)$. Lowest values are obtained for $\lambda^* = \lambda_{\max} = \mu_1 = 0.2$. There is correlation between optimizing the worst-case for λ and better performance in the experiment.