A Bregman Divergence View on the Difference-of-Convex Algorithm

Oisín Faust
University of Cambridge

Hamza Fawzi
University of Cambridge

James Saunderson
Monash University

Abstract

The difference-of-convex (DC) algorithm is a conceptually simple method for the minimization of (non)convex functions that are expressed as the difference of two convex functions. An attractive feature of the algorithm is that it maintains a global overestimator on the function and does not require a choice of step size at each iteration. By adopting a Bregman divergence point of view, we simplify and strengthen many existing non-asymptotic convergence guarantees for the DC algorithm. We further present several sufficient conditions that ensure a linear convergence rate, namely a new DC Polyak-Łojasiewicz condition, as well as a relative strong convexity assumption. Importantly, our conditions do not require smoothness of the objective function. We illustrate our results on a family of minimization problems involving the quantum relative entropy, with applications in quantum information theory.

1 INTRODUCTION

Consider an optimization problem

$$\min_{x\in\mathbb{R}^n} \quad F(x) \tag{1}$$

where the function F can be expressed as a difference of two convex functions f_1 and f_2, namely

$$F(x) = f_1(x) - f_2(x). \tag{2}$$

(Although the formulation (1) does not explicitly impose constraints on the decision variable, it can model problems with convex constraints by allowing f_1 to take the value $+\infty$ outside of a convex set $\operatorname{dom} f_1 \subseteq \mathbb{R}^n$.)

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

A popular algorithm for solving such a problem is the DC algorithm, or DCA (Le Thi and Pham Dinh, 2018), which replaces f_2 by its linear approximation at each iteration, assuming f_2 is differentiable:

$$x^{k+1} \in \operatorname*{argmin}_{x\in\operatorname{int}\operatorname{dom} f_2} f_1(x) - [f_2(x^k) + \langle \nabla f_2(x^k), x - x^k\rangle]. \tag{3}$$

This algorithm is sometimes also called the Convex-Concave procedure, or CCCP (Yuille and Rangarajan, 2001). An attractive feature of the algorithm is that it maintains a global overestimator on the function and does not require a choice of step size at each iteration.

It is easy to see that the sequence of function values $(F(x^k))$ is monotonically nonincreasing. Furthermore, assuming both f_1 and f_2 are differentiable, it has been shown (Tao and An, 1997) that if (x^k) is bounded, then any of its limit points is a critical point for F, and the rate of convergence of $\min_{0\le k\le N} \|\nabla F(x^k)\|_2$ is $O(1/\sqrt{N})$. A linear rate of convergence under a Łojasiewicz gradient inequality has also been obtained in Le Thi et al. (2018).

1.1 Contributions

In this paper we consider the DC algorithm from the point of view of Bregman divergences. The main observation for this paper is that the DCA can be interpreted as the Bregman proximal point algorithm (Bregman PPA), with Bregman divergence associated to f_2, namely

$$D_{f_2}(x|y) = f_2(x) - (f_2(y) + \langle \nabla f_2(y), x - y\rangle).$$

This equivalence is reviewed in Section 2.3. This observation allows us to rederive and strengthen in a simplified way many existing convergence results for DCA. In particular it points out that a natural metric to measure convergence of the algorithm is the Bregman divergence D_{f_2}. All our convergence results involve this divergence, and existing results in the literature which are phrased in terms of the Euclidean metric follow by imposing a smoothness or strong convexity assumption on f_2.

In Section 3 we consider the case where the DC function F is convex. Surprisingly, we are not aware of

any prior result on the convergence rate of the DCA in the convex case. This is perhaps due to the existence of algorithms, such as the subgradient method, specifically designed for minimising nonsmooth convex functions. However, all global convergence guarantees for the subgradient method require F to be Lipschitz, and even then convergence can be no better than $O(1/\sqrt{k})$. On the other hand, when a DC decomposition (2) exists for which the DCA iterates can be computed, the DCA can perform much better. We offer two results in this direction. The first is an $O(1/k)$ convergence rate taken directly from the Bregman proximal point literature (and which does not require F to be Lipschitz). The second is a global linear convergence result which holds under a strong relative convexity assumption. The corresponding Bregman PPA result (Proposition 1) appears to be new.

In the nonconvex setting, we show in Section 4 how some nonasymptotic convergence results from the DCA literature can be recovered very naturally from the Bregman viewpoint. In some cases, we obtain improved results. We also introduce the notion of *DC PL inequalities*, and show that they imply linear convergence of the DCA. These inequalities are related to work of Bauschke et al. (2019) concerning the mirror descent algorithm.

In Section 5, we illustrate our results with the problem of computing the conjugate function and proximal operator of the quantum conditional entropy function which has applications in quantum information theory and quantum statistical mechanics.

2 PRELIMINARIES

After reviewing some terminology from convex analysis and the main technical assumptions throughout the paper, we present the equivalence between the DC algorithm and the Bregman proximal point algorithm.

2.1 Convex analysis

Define $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$. The *domain* of a convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is $\operatorname{dom} f = \{x \in \mathbb{R}^n \mid f(x) < \infty\}$. The function f is closed if each of its sublevel sets, $\{x \mid f(x) \leq t\}$ for $t \in \mathbb{R}$, is closed. The *subdifferential* of a convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ at $x \in \mathbb{R}^n$ is given by

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

Given a subset $C \subset \mathbb{R}^n$, the indicator function of C is the function $\mathbb{R}^n \to \overline{\mathbb{R}}$

$$\iota_C(x) = \begin{cases} 0 & x \in C \\ +\infty & \text{otherwise.} \end{cases}$$

If f is a convex function whose domain has nonempty interior, and which is differentiable in the interior of its domain, the *Bregman divergence* associated to f is

$$D_f(x|y) := f(x) - f(y) - \langle \nabla f(y), x - y\rangle.$$

Bregman divergences are nonnegative on $\operatorname{dom} f \times \operatorname{int dom} f$ and are convex in the first variable. When f is not differentiable, we can extend the notion of Bregman divergence by replacing $\nabla f(y)$ by any $v \in \partial f(y)$,

$$D_f^v(x|y) = f(x) - f(y) - \langle v, x - y\rangle. \tag{4}$$

Recall that for $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, its *convex conjugate* is $f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$ and defined by

$$f^*(v) = \sup_x \ \langle v, x\rangle - f(x). \tag{5}$$

A function f is L-smooth w.r.t. $\|\cdot\|$ if $\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|$ holds for all $x, y \in \operatorname{int dom} f$, and where $\|\cdot\|_*$ is the dual norm. Also, f is μ-strongly convex w.r.t. $\|\cdot\|$ if $D_f^v(x|y) \geq (\mu/2)\|x - y\|^2$ for all x, y and $v \in \partial f(y)$. If f is μ-strongly convex w.r.t. $\|\cdot\|$ then the conjugate f^* is $1/\mu$-smooth w.r.t. the dual norm $\|\cdot\|_*$, see e.g. (Beck, 2017, Chapter 5).

Finally we recall the notions of relative smoothness and relative strong convexity as introduced in Bauschke et al. (2017); Lu et al. (2018). Given a convex function g, we say that f is L-smooth with respect to g if $Lg - f$ is convex. Furthermore, we say that f is μ-strongly convex relative to g if $f - \mu g$ is convex. The standard Euclidean notions of smoothness and strong convexity are obtained in the special case $g(x) = \|x\|_2^2/2$.

2.2 Running Assumptions

Consider a DC function $F = f_1 - f_2$ where $f_1, f_2 : \mathbb{R}^n \to \overline{\mathbb{R}}$. Throughout, we make the following technical assumptions on f_1 and f_2. These ensure that the minimization problem for F is well-defined and that DCA iterates exist for all k:

1. $\varnothing \neq \operatorname{dom} f_1 \subseteq \operatorname{dom} f_2$,
2. F is bounded below on $\mathbb{R}^n$, and $F(x) = +\infty$ for $x \notin \operatorname{dom} f_2$,
3. $\operatorname{dom} f_1 \cap \operatorname{int dom} f_2 \neq \varnothing$,
4. f_2 is differentiable on $\operatorname{int dom} f_2$,
5. $\{\nabla f_2(x) \mid x \in \operatorname{int dom} f_2\} \subseteq \bigcup_{x \in \operatorname{int dom} f_2} \partial f_1(x)$.

Note that the final assumption ensures that iterates x^k satisfying (3) exist for all $k \in \mathbb{N}$, given $x^0 \in \operatorname{int dom} f_2$. Indeed, (3) is equivalent to $x^{k+1} \in \operatorname{int dom} f_2$ and

$$\partial f_1(x^{k+1}) \ni \nabla f_2(x^k). \tag{6}$$

2.3 DCA Is the Bregman Proximal Point Algorithm

Consider the problem of minimizing F over $\mathbb{R}^n$. Let g be a convex function which is differentiable in the interior of its domain, assumed nonempty. The *Bregman proximal point algorithm* (Bregman PPA) with respect to g is the iterative scheme:

$$x^{k+1} \in \arg\min_{x \in \operatorname{int}\operatorname{dom} g} \left\{ F(x) + t_k^{-1} D_g(x|x^k) \right\}, \quad (7)$$

where $t_k > 0$ for each k, given an initial $x^0 \in \operatorname{int}\operatorname{dom} g$.

Assuming $F = f_1 - f_2$, if we choose $g = f_2$ and $t_k = 1$ for all k then the update is

$$\begin{aligned} x^{k+1} &\in \arg\min_x f_1(x) - f_2(x) + D_{f_2}(x|x^k) \\ &= \arg\min_x f_1(x) - [f_2(x^k) + \langle \nabla f_2(x^k), x - x^k \rangle] \end{aligned}$$

which is exactly the DC algorithm (3).

Remark 1. *When f_1 (in addition to f_2) is differentiable, it can be shown that the DCA is also equivalent to* mirror descent *with Bregman divergence generated by f_1. Mirror descent replaces* (7) *by*

$$x^{k+1} \in \arg\min_x \{ \langle \nabla F(x^k), x \rangle + t_k^{-1} D_g(x|x^k) \}. \quad (8)$$

We recover DCA by setting $t_k \equiv 1$ and $g = f_1$. (Recall: to recover DCA from the Bregman PPA, we set $g = f_2$.) The DC decomposition $F = f_1 - f_2$ guarantees that F is 1-smooth relative *to f_1, i.e. $f_1 - F$ is convex. Mirror descent has been studied in the context of relative smoothness by Bauschke et al. (2017); Lu et al. (2018). We stress that the Bregman* PPA *view is more general, since it does not require f_1 to be differentiable.*

3 DCA GUARANTEES FOR CONVEX DC FUNCTIONS

In this section we assume that the function F is closed and convex. It is known that the iterates (7) of the Bregman proximal point algorithm for a convex function F, with respect to the divergence D_g satisfy (see, e.g., Censor and Zenios (1992); Chen and Teboulle (1993))

$$\begin{cases} F(x^{k+1}) - F(x^*) \leq \frac{1}{\sum_{i=0}^k t_i} D_g(x^*|x^0) \\ D_g(x^*|x^{k+1}) \leq D_g(x^*|x^k). \end{cases}$$

for any $k \geq 0$, and any x^* a minimizer of F. Applied to the DC setting, we immediately obtain

Corollary 1. *Suppose $F = f_1 - f_2$ is closed and convex, and let x^* be a minimizer of F. The DCA iterates* (3) *for F satisfy*

$$\begin{aligned} F(x^k) - F(x^*) &\leq \tfrac{1}{k} D_{f_2}(x^*|x^0) \\ D_{f_2}(x^*|x^{k+1}) &\leq D_{f_2}(x^*|x^k). \end{aligned}$$

The following proposition shows that if F is relatively strongly convex with respect to f_2, then the DCA iterates enjoy a linear rate of convergence. As far as we are aware, this result has not appeared before in the literature on the Bregman proximal point algorithm.

Since it is not clear what $(F - \mu f_2)(x)$ means for points $x \notin \operatorname{dom} f_2$ (and hence also $x \notin \operatorname{dom} F$), we define relative strong convexity as the existence of a decomposition $F = h + \mu f_2$, where $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex.

Proposition 1. *Suppose that F is μ-relatively strongly convex with respect to f_2, i.e., $F = f_1 - f_2 = h + \mu f_2$, for some $\mu \geq 0$ and convex function h. Let x^* be a minimizer of F. Then the iterates of the DCA satisfy*

$$D_{f_2}(x^*|x^{k+1}) \leq (1+\mu)^{-1} D_{f_2}(x^*|x^k).$$

Proof. We begin by lower bounding the progress of the iterates towards x^* as

$$\begin{aligned} &D_{f_2}(x^*|x^k) - D_{f_2}(x^*|x^{k+1}) \\ &= f_2(x^{k+1}) - f_2(x^k) + \langle \nabla f_2(x^{k+1}), x^* - x^{k+1} \rangle \\ &\quad - \langle \nabla f_2(x^k), x^* - x^k \rangle \\ &= D_{f_2}(x^{k+1}|x^k) + \langle x^* - x^{k+1}, \nabla f_2(x^{k+1}) - \nabla f_2(x^k) \rangle \\ &\geq \langle x^{k+1} - x^*, \nabla f_2(x^k) - \nabla f_2(x^{k+1}) \rangle \end{aligned}$$

where the last inequality holds because f_2 is convex.

Note that, since $F = h + \mu f_2$, necessarily $\operatorname{dom} F \subseteq \operatorname{dom} h$. Therefore $\operatorname{int}\operatorname{dom} f_2 \cap \operatorname{dom} h \neq \varnothing$, so by the Moreau-Rockafellar theorem (Mordukhovich and Nam, 2013, Corollary 2.45), $\partial F(x) = \partial h(x) + \mu \nabla f_2(x)$ holds for any $x \in \operatorname{int}\operatorname{dom} f_2$. Let $v^{k+1} \in \partial F(x^{k+1})$ be any element of the subgradient of F at x^{k+1}. Then

$$\begin{aligned} \langle x^{k+1} - x^*, v^{k+1} \rangle &\geq \mu \langle x^{k+1} - x^*, \nabla f_2(x^{k+1}) \rangle \\ &\quad + (F - \mu f_2)(x^{k+1}) - (F - \mu f_2)(x^*) \\ &= \mu D_{f_2}(x^*|x^{k+1}) + F(x^{k+1}) - F(x^*) \\ &\geq \mu D_{f_2}(x^*|x^{k+1}). \end{aligned}$$

The first inequality holds because $v^{k+1} - \mu \nabla f_2(x^{k+1}) \in \partial h(x^{k+1})$. The second inequality holds because x^* is a minimizer of F. The iterates of the DCA satisfy

$$\nabla f_2(x^k) \in \partial(F + f_2)(x^{k+1}) = \partial F(x^{k+1}) + \nabla f_2(x^{k+1}),$$

where the equality holds because $\varnothing \neq \operatorname{dom} F \cap \operatorname{int}\operatorname{dom} f_2$. Putting these observations together gives

$$\begin{aligned} &D_{f_2}(x^*|x^k) - D_{f_2}(x^*|x^{k+1}) \\ &\qquad \geq \langle x^{k+1} - x^*, \nabla f_2(x^k) - \nabla f_2(x^{k+1}) \rangle \\ &\qquad \geq \mu D_{f_2}(x^*|x^{k+1}) \end{aligned}$$

from which the result follows. □

Iterating this result and combining with the fact that

$$F(x^{k+1}) + D_{f_2}(x^{k+1}|x^k) \leq F(x^*) + D_{f_2}(x^*|x^k),$$

we obtain the following corollary.

Corollary 2. *Suppose that $F = f_1 - f_2$ is μ-strongly convex relative to f_2. Let x^* be a minimizer of F. Then the DCA iterates for* (1) *satisfy*

$$F(x^{k+1}) - F(x^*) \leq D_{f_2}(x^*|x^k) \leq (1+\mu)^{-k} D_{f_2}(x^*|x^0).$$

Remark 2. *It is known (Lu et al., 2018, Theorem 3.1) that* relative strong convexity *(i.e., $F - mg$ is convex) and* relative smoothness *(i.e., $Lg - F$ is convex) conditions for $L \geq m > 0$ are enough to guarantee a global linear convergence rate of $1 - \frac{m}{L}$ for mirror descent* (8) *w.r.t. $D_{f_1}(x^*|x^k)$. Interpreting the DCA as mirror descent, as in Remark 1, this implies that if $F - mf_1$ is convex, then the DCA has a linear rate of convergence of $1 - m$ (recall that F is 1-relatively smooth with respect to f_1). Since $F = f_1 - f_2$, the condition that $F - mf_1$ is convex is the same as $F - \mu f_2$ being convex with $(1+\mu)^{-1} = 1 - m$. This is precisely the rate of convergence in Proposition 1. However, Proposition 1 guarantees linear convergence w.r.t. D_{f_2}, not D_{f_1}, which may be advantageous if $D_{f_2}(x^*|x^0) \ll D_{f_1}(x^*|x^0)$. Another advantage of Proposition 1 is that it does not require differentiability of f_1.*

3.1 Linear Convergence Via Local Relative Strong Convexity

In many practical examples, linear convergence in Bregman divergence $D_{f_2}(x^*|x^k)$ is observed even when the function $F - \epsilon f_2$ is nonconvex for every $\epsilon > 0$. As stated, Proposition 1 is insufficient to explain this behaviour. To rectify the situation, we propose the following generalization of Proposition 1.

Proposition 2. *Let S be a convex set containing the sublevel set $\{x \mid F(x) \leq F(x^0)\}$. Suppose that F is μ-strongly convex relative to f_2* on *S, in the sense that $F + \iota_S \equiv h + \mu f_2$ for some convex function h and some $\mu \geq 0$. Let x^* be a minimizer of F. Then $F(x^{k+1}) - F(x^*) \leq D_{f_2}(x^*|x^k) \leq (1+\mu)^{-1} D_{f_2}(x^*|x^0)$.*

Proof. See Appendix A.1. □

4 DCA GUARANTEES FOR NONCONVEX DC FUNCTIONS

This section begins with a collection of existing convergence rate results for the DCA in the nonconvex setting. We then prove Theorem 5, which can be used to recover (and sometimes improve) these existing results. Finally, we present some theory concerning linear convergence in the nonconvex setting.

4.1 Existing Guarantees on DCA

Throughout this subsection, let

$$F_* := \inf_{x \in \mathbb{R}^n} F(x).$$

We outline below some of the existing convergence results for the DCA. The first guarantee on the convergence of DCA appears in (Tao and An, 1997, Proposition 2).

Theorem 1 ((Tao and An, 1997, Proposition 2))**.** *Consider DCA applied to $F(x) = f_1(x) - f_2(x)$, where f_1, f_2 are convex. Let $\mu_1, \mu_2 \geq 0$ such that $f_i - (\mu_i/2)\|\cdot\|_2^2$ are convex, and assume that $\mu_1 + \mu_2 > 0$. Then after N iterations of DCA, we have*

$$\min_{0 \leq k \leq N-1} \|x^{k+1} - x^k\|_2^2 \leq \frac{2(F(x^0) - F_*)}{(\mu_1 + \mu_2)N}. \quad (9)$$

When f_1 and f_2 are smooth, a guarantee on the magnitude of the gradient is given in Corollary 3.1 of Abbaszadehpeivasti et al. (2021):

Theorem 2 ((Abbaszadehpeivasti et al., 2021, Corollary 3.1))**.** *Consider the DCA applied to $F(x) = f_1(x) - f_2(x)$, where f_1, f_2 are convex and differentiable. Assume that ∇f_1 and ∇f_2 are respectively L_1, L_2-Lipschitz. Then*

$$\min_{0 \leq k \leq N} \|\nabla F(x^k)\|_2^2 \leq \frac{2L_1 L_2 (F(x^0) - F_*)}{(L_1 + L_2)N + \max(L_1, L_2)}. \quad (10)$$

The following guarantee is from Corollary 3.2 of Yurtsever and Sra (2022), where it was proved by reducing DCA to the Frank-Wolfe algorithm. Note that it requires neither smoothness nor strong convexity of F.

Theorem 3. *Consider DCA applied to $F(x) = f_1(x) - f_2(x)$ where f_1, f_2 are convex and f_2 is differentiable. Let x^k be the sequence of iterates. Then for any $N \geq 1$ there exists $0 \leq k < N$ such that*

$$f_1(x^k) - f_1(x) - \langle \nabla f_2(x^k), x^k - x \rangle \leq \frac{F(x^0) - F_*}{N}$$

for all x.

Remark 3. *The quantity appearing on the left-hand side is reminiscent of a Bregman divergence. In fact, it* is *a Bregman divergence if $\nabla f_2(x^k) \in \partial f_1(x)$. Later, we will prove Theorem 5, a result which* is *stated naturally in terms of Bregman divergences and which is stronger than Theorem 3. This is evidence that the Bregman PPA view of the DCA is both natural and powerful.*

Abbaszadehpeivasti et al. (2021) show, in Theorem 5.1 of their paper, that a Polyak-Łojasiewicz type inequality implies linear convergence. For ease of exposition we

state the result when both f_1 and f_2 are smooth, although their result requires only one of them to be smooth. See also Le Thi et al. (2018) for a similar result, but which also assumes strong convexity.

Theorem 4. *Suppose that f_1, f_2 are L_1, L_2 smooth on $\mathbb{R}^n$. Suppose also that for some $\mu > 0$, the PL inequality*

$$\tfrac{1}{2}\|\nabla F(x)\|_2^2 \geq \mu\left(F(x) - F_*\right)$$

holds on $\{x \mid F(x) \leq F(x^0)\}$. Then for all $k \geq 0$

$$F(x^{k+1}) - F_* \leq \left(\frac{1-\mu/L_1}{1+\mu/L_2}\right)\left(F(x^k) - F_*\right).$$

4.2 Main Theorem and Recovering Existing Results

The following main theorem gives a guarantee on the iterates of the DC algorithm in terms of the Bregman divergences D_{f_1} and D_{f_2}. As we show below, this theorem subsumes all the sublinear rates from Theorems 1 to 3.

Theorem 5. *The sequence of iterates of the DC algorithm satisfies, for $N \geq 1$,*[1]

$$\min_{0\leq k\leq N-1}\left\{D_{f_1}^{\nabla f_2(x^k)}(x^k|x^{k+1}) + D_{f_2}(x^{k+1}|x^k)\right\} \leq \frac{F(x^0) - F(x^N)}{N}.$$

If f_1 is differentiable, then for each $0 \leq k < N$, the term inside the minimization equals

$$D_{f_1^*}^{x^k}(\nabla f_2(x^k)|\nabla f_1(x^k)) + D_{f_2^*}^{x^{k+1}}(\nabla f_1(x^{k+1})|\nabla f_2(x^{k+1})).$$

Proof. The iterates of the DC algorithm satisfy:

$$D_{f_1}^{\nabla f_2(x^k)}(x^k|x^{k+1}) + D_{f_2}(x^{k+1}|x^k) = F(x^k) - F(x^{k+1}).$$

Summing, and lower bounding the average by the minimum, we get the inequality in the theorem above.

If f_1 is differentiable, then $\nabla f_1(x^{k+1}) = \nabla f_2(x^k)$. So, using the fact that $f_2^*(\nabla f_2(x)) = \langle x, \nabla f_2(x)\rangle - f_2(x)$,

$$\begin{aligned} D_{f_2}(x^{k+1}|x^k) &= D_{f_2^*}^{x^{k+1}}(\nabla f_2(x^k)|\nabla f_2(x^{k+1})) \\ &= D_{f_2^*}^{x^{k+1}}(\nabla f_1(x^{k+1})|\nabla f_2(x^{k+1})). \end{aligned}$$

Similarly, since f_1 is now assumed differentiable,

$$\begin{aligned} D_{f_1}^{\nabla f_2(x^k)}(x^k|x^{k+1}) &= D_{f_1}(x^k|x^{k+1}) \\ &= D_{f_1^*}^{x^k}(\nabla f_1(x^{k+1})|\nabla f_1(x^k)) \\ &= D_{f_1^*}^{x^k}(\nabla f_2(x^k)|\nabla f_1(x^k)). \end{aligned}$$

□

[1] Note that by definition of the DC iterates, we have $\nabla f_2(x^k) \in \partial f_1(x^{k+1})$, which justifies the validity of the term $D_{f_1}^{\nabla f_2(x^k)}(x^k|x^{k+1})$.

By putting further assumptions on f_1, f_2 (L-smoothness, etc.) we can recover the previous results from Theorems 1 to 3, as we show next.

Recovering Theorem 1 If f_1 is μ_1-strongly convex and f_2 is μ_2-strongly convex then we have $D_{f_i}^v(a|b) \geq \frac{\mu_i}{2}\|a-b\|_2^2$ for $v \in \partial f_i(b)$, $i = 1, 2$. Therefore

$$\begin{aligned} &\min_{0\leq k\leq N-1} \frac{\mu_1+\mu_2}{2}\|x^k - x^{k+1}\|_2^2 \\ &\leq \min_{0\leq k\leq N-1}\left\{D_{f_1}^{\nabla f_2(x^k)}(x^k|x^{k+1}) + D_{f_2}(x^{k+1}|x^k)\right\} \\ &\leq \frac{F(x^0) - F(x^N)}{N}, \end{aligned}$$

which implies Theorem 1 because $F(x^0) - F_*$ is larger than $F(x^0) - F(x^N)$.

Recovering Theorem 2 If f_1 is L_1-smooth and f_2 is L_2-smooth, then f_1^* (respectively f_2^*) is L_1^{-1}-strongly (respectively L_2^{-1}-strongly) convex. Thus

$$\begin{aligned} \frac{1}{2L_1}\|\nabla F(x^k)\|_2^2 + \frac{1}{2L_2}\|\nabla F(x^{k+1})\|_2^2 &\leq D_{f_1^*}^{x^k}(\nabla f_2(x^k)|\nabla f_1(x^k)) \\ &\quad + D_{f_2^*}^{x^{k+1}}(\nabla f_1(x^{k+1})|\nabla f_2(x^{k+1})). \end{aligned}$$

For any $0 \leq k < N$ left hand side is at least

$$\tfrac{1}{2}(L_1^{-1} + L_2^{-1})\ \min_{0\leq i\leq N}\{\|\nabla F(x^i)\|_2^2\}.$$

Therefore, taking the minimum of the right-hand side over k and applying Theorem 5, we get the bound

$$\min_{0\leq k\leq N}\|\nabla F(x^k)\|_2^2 \leq \frac{2L_1L_2}{(L_1+L_2)N}(F(x^0) - F(x^N)).$$

Note that if only one function is smooth this bound still makes sense.

Recovering Theorem 3 Note that, by definition of x^{k+1}, we have that for any x

$$f_1(x^k) - f_1(x) - \langle \nabla f_2(x^k), x^k - x\rangle \leq D_{f_1}^{\nabla f_2(x^k)}(x^k|x^{k+1}).$$

Applying Theorem 5, we in fact obtain the *stronger* result that for some $0 \leq k < N$,

$$f_1(x^k) - f_1(x) - \langle \nabla f_2(x^k), x^k - x\rangle + D_{f_2}(x^{k+1}|x^k) \leq \frac{F(x^0) - F_*}{N}.$$

4.3 DC PL Inequalities

In optimization, an important class of differentiable functions are those obeying a Polyak-Łojasiewicz (PL) inequality. We say that F satisfies a PL inequality on $\mathcal{X}$ with respect to $F_* \in \mathbb{R}$ if there exists $\mu > 0$ such that

$$\frac{1}{2}\|\nabla F(x)\|_2^2 \geq \mu\left(F(x) - F_*\right), \tag{11}$$

for every $x \in \mathcal{X}$. PL inequalities are useful in the convergence analysis of, for example, (proximal) gradient descent (Karimi et al., 2016).

We now introduce a DC-specific modification of the PL inequality which (a) is adapted to the specific DC decomposition of F and (b) measures the size of $\nabla F(x)$ in Bregman divergence rather than via the Euclidean norm.

Definition 1. *Let f_1 and f_2 be differentiable. We say that F satisfies the* DC PL *inequality on $\mathcal{X}$ with respect to F_* if there exist $\eta_1 \geq 0$ and $\eta_2 \geq 0$ with $\eta_1 + \eta_2 > 0$ such that for all $x \in \mathcal{X}$,*[2]

$$\begin{cases} \eta_1(F(x) - F_*) \leq D^x_{f_1^*}(\nabla f_2(x)|\nabla f_1(x)) \\ \eta_2(F(x) - F_*) \leq D^x_{f_2^*}(\nabla f_1(x)|\nabla f_2(x)). \end{cases} \tag{12}$$

The DC PL inequality above allows us to prove a linear convergence rate on the function value.

Lemma 1. *Suppose that F satisfies a DC PL inequality on the sublevel set $\{x : F(x) \leq F(x^0)\}$, with respect to $F_* = \inf_x F(x)$. Then for any $k \geq 0$*

$$F(x^{k+1}) - F_* \leq \frac{1-\eta_1}{1+\eta_2}\left(F(x^k) - F_*\right).$$

Proof. See Appendix A.2. □

Remark 4. *Bauschke et al. (2019) introduced an inequality (which they call GD2) which is identical to the first inequality in* (12)*. They considered the problem of minimizing F via mirror descent* (8) *with Bregman divergence generated by f_1. Recall from Remark 1 that this algorithm is identical to the DCA when f_1 and f_2 are both differentiable. Our work provides an improvement in two ways. First, we show that linear convergence follows from GD2 alone, without any extra assumptions concerning the symmetry of the divergence $D_{f_1}(\cdot|\cdot)$—see Remark 4.1(a) in (Bauschke et al., 2019). Next, the second inequality in* (12)*, which involves $D_{f_2^*}(\cdot|\cdot)$ and also implies linear convergence, is new.*

Recovering Theorem 4 Suppose that f_1 is L_1-smooth and f_2 is L_2-smooth and that the regular PL inequality (11) holds for all $x \in \mathcal{X} = \{x : F(x) \leq F(x^0)\}$. Then, the DC PL inequality (12) holds with $\eta_1 = \mu/L_1$ and $\eta_2 = \mu/L_2$. Indeed, from the L_i-smoothness of f_i it follows that f_i^* is L_i^{-1}-strongly convex. Thus $D^x_{f_1^*}(\nabla f_2(x)|\nabla f_1(x)) \geq \frac{L_1^{-1}}{2}\|\nabla F(x)\|_2^2$ and $D_{f_2^*}(x|y) \geq \frac{L_2^{-1}}{2}\|\nabla F(x)\|_2^2$. Therefore

$$F(x) - F_* \leq (\mu/L_1)^{-1} D_{f_1^*}(\nabla f_2(x)|\nabla f_1(x)) \quad \text{and}$$
$$F(x) - F_* \leq (\mu/L_2)^{-1} D_{f_2^*}(\nabla f_1(x)|\nabla f_2(x))$$

for all $x \in \mathcal{X}$. Lemma 1 then tells us that

$$\frac{F(x^{k+1}) - F_*}{F(x^k) - F_*} \leq \frac{1-\eta/L_1}{1+\eta/L_2},$$

for all $k \geq 0$, which recovers Theorem 4.

Sufficient conditions for the DC PL condition We now outline a few sufficient conditions that guarantee the DC PL condition (12).

- *Strong convexity and smoothness with respect to a norm:* If F is μ-strongly convex w.r.t. a norm $\|\cdot\|$ and the f_i's are L_i-smooth w.r.t. $\|\cdot\|$, then the DC PL inequality (12) automatically holds with constant $\eta_i = \mu/L_i$. Indeed, if F is strongly convex w.r.t. $\|\cdot\|$, then we have, for any x, $F(x) - F_* \leq 1/(2\mu)\|\nabla F(x)\|_*^2$. Furthermore, if f_i is L_i-smooth with respect to $\|\cdot\|$ then $D^x_{f_i^*}(v|\nabla f_i(x)) \geq \frac{1}{2L_i}\|v - \nabla f_i(x)\|_*^2$ for all x, v. It follows that (with $j \neq i$)

  $$D^x_{f_i^*}(\nabla f_j(x)|\nabla f_i(x)) \geq \frac{1}{2L_i}\|\nabla f_i(x) - \nabla f_j(x)\|_*^2.$$

  Then, for $j \neq i$, we obtain the desired inequalities

  $$\tfrac{\mu}{L_i}\left(F(x) - F_*\right) \leq D^x_{f_i^*}(\nabla f_j(x)|\nabla f_i(x)).$$

- *Self-Concordance and Strong Geodesic Convexity* Suppose that $\operatorname{dom} f_2$ is open, and that f_2 is twice differentiable and strictly convex. Then $\nabla^2 f_2(x)$ defines a Riemann metric $\|v\|_x = \sqrt{v^\top \nabla^2 f_2(x) v}$ for $x \in \operatorname{dom} f_2$. Write $\mathcal{M}$ for the Riemannian manifold $(\operatorname{dom} f_2, \nabla^2 f_2)$. Recall that f is *geodesically μ-strongly convex* on $\mathcal{M}$ if $f \circ \gamma : [0,1] \to \overline{\mathbb{R}}$ is $\mu L(\gamma)^2$-strongly convex (in the usual sense) for every geodesic $\gamma : [0,1] \to \mathcal{M}$ of length $L(\gamma)$ (Boumal, 2023, Definition 11.5). Then one can prove the following:

  Proposition 3. *Suppose that F is geodesically μ-strongly convex on $\mathcal{M}$, and that f_2 is strongly non-degenerate self-concordant (Renegar, 2001, §2.5). Then a DC PL inequality with $\eta_2 = \frac{\mu}{2}$ holds on $\{x \mid \|\nabla F(x)\|_{x,*} < 1\}$, where $\|\cdot\|_{x,*} := \sqrt{\cdot^\top \nabla^2 f_2(x)^{-1} \cdot}$ is the norm dual to $\|\cdot\|_x$.*

  Proof. See Appendix A.3. □

[2] Note that $x \in \partial f_i^*(\nabla f_i(x))$ which justifies the Bregman divergences in (12)

DC PL Inequality Implies Local Strong Relative Convexity Under certain regularity conditions, a DC PL inequality implies local strong relative convexity. We state the result here, and prove it in Appendix A.

Proposition 4. *Suppose f_1 and f_2 are C^2 and strictly convex on an open set $\mathcal{X}$. Let x^* be a local minimum of F such that $\nabla^2 F(x^*) \succ 0$. Suppose a DC PL inequality holds with constants η_1, η_2 on $\mathcal{X}$ w.r.t. $F(x^*)$. Then*

$$\nabla^2 F(x^*) \succeq \max\{\frac{\eta_1}{1-\eta_1}, \eta_2\} \nabla^2 f_2(x^*).$$

Proof. See Appendix A.4. □

5 Application to the Conjugate and Proximal Operators of the Quantum Conditional Entropy Function

The state of a quantum system is represented by a $d \times d$ density matrix X, a Hermitian positive semidefinite matrix with unit trace. The *von Neumann entropy* of X is defined by

$$S(X) = -\mathrm{tr}(X \log X)$$

and is a concave function of X. If the quantum system represented by X is composed of two parts A and B of local dimensions d_A and d_B respectively, then $d = d_A d_B$. One can define the conditional entropy of A given B by the difference of concave expression:

$$S(A|B)_X = S(X) - S(\mathrm{tr}_A X)$$

where tr_A is a linear map which corresponds to marginalizing out the A subsystem. Remarkably, the conditional entropy is itself concave because we can write $S(A|B)_X = -D(X\|I_A \otimes \mathrm{tr}_A X)$ where

$$D(X\|W) := \mathrm{Tr}[X(\log X - \log W)] \tag{13}$$

is the quantum relative entropy, which is jointly convex in (X, W) (Lieb and Ruskai, 1973).

The quantum conditional entropy function plays a fundamental role in quantum information, and many quantities can be expressed as optimization problems involving this function, such as capacities of quantum channels (Wilde, 2013), or in quantum statistical mechanics for the estimation of quantum partition functions (Poulin and Hastings, 2011; Bravyi et al., 2021). To solve these problems for large matrix sizes $d = d_A d_B$, first-order proximal methods are very natural candidate algorithms, provided the proximal operator of the conditional entropy function can be computed efficiently.

In this section we show that the DCA indeed allows us to compute the conjugate function, as well as the proximal operator very accurately in a small number of iterations, for large problem sizes.

It is tempting to compare the performance of the DCA with that of (accelerated) projected gradient descent. However, because the conditional entropy function does not have a Lipschitz gradient, there are no guarantees for projected gradient descent. Moreover, these methods tend to produce iterates that lie on the boundary of the feasible set, where the gradient of the function in question is not defined. Instead, we compare the DCA with the open-source interior-point solver `Hypatia` (Coey et al., 2021) which is the only interior point solver we are aware of which can optimise directly over the quantum relative entropy cone.

5.1 Conjugate of the Quantum Conditional Entropy

The convex conjugate of the negative conditional entropy function evaluated at some $d \times d$ Hermitian matrix H is

$$\min_{X \succeq 0, \mathrm{tr} X = 1} \left\{ \mathrm{tr}(HX) - S(A|B)_X \right\}. \tag{14}$$

This is a difference-of-convex optimization problem, with $f_1(X) := \mathrm{tr}(HX) - S(X) + \iota_{\{\mathrm{tr}(X)=1\}}$ and $f_2(X) = -S(\mathrm{tr}_A X)$. Moreover $F := f_1 - f_2$ is itself convex. The DCA iterates take the form

$$X^{k+1} = \frac{1}{Z} \exp(-H + I_A \otimes \log \mathrm{tr}_A X^k) \tag{15}$$

where $Z = \mathrm{tr} \exp(-H + I_A \otimes \log \mathrm{tr}_A X^k)$.

Corollary 1 holds for this problem, giving the guarantee

$$\begin{aligned} F(X^k) - F_* &\leq \tfrac{1}{k} D(\mathrm{tr}_A X^* \| \mathrm{tr}_A X^0) \\ &\leq \tfrac{1}{k} \log(d_B) \end{aligned} \tag{16}$$

when $X^0 = I/d$ is the scaled identity matrix.

Figure 1 shows the running time of DCA for a randomly chosen H normalized to have unit operator norm, as a function of the problem dimension, where we fix $d_A = 2$, and take d_B a growing power of 2. We used the condition $2\|\nabla F(X^k)\|_\sigma \leq 10^{-7}$ as our stopping criterion (where $\|\cdot\|_\sigma$ is the operator norm), which guarantees that $F(X^k) - F^* \leq 10^{-7}$ since $\|X - Y\|_1 \leq 2$ for any two density matrices X, Y. The DCA is at least 2 orders of magnitude faster than the interior-point solver `Hypatia`. On the machine we used for this experiment, `Hypatia` failed to solve problems with $d_B \geq 64$ due to the machine having insufficient RAM. We note that complexity of each IPM iteration scales at least as $\Omega(d^5)$ (Coey et al., 2022, Table 1), whereas the per-iteration complexity of the DCA is $O(d^3)$, dominated by the matrix exponential and matrix logarithm operations. Of

Figure 1: Time Taken By DCA vs IPM To Solve (14)

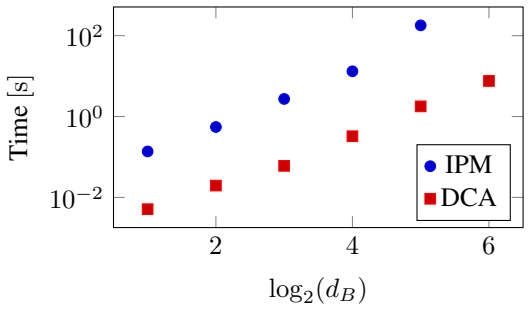


Figure 2: Convergence of the DCA (15) for $d_B = 64$

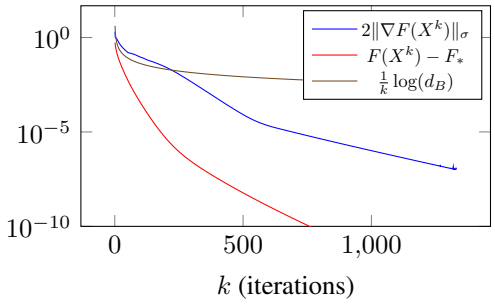


Figure 3: Time Taken By DCA vs IPM To Solve (17)

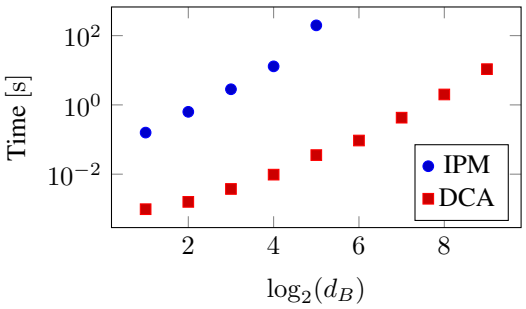
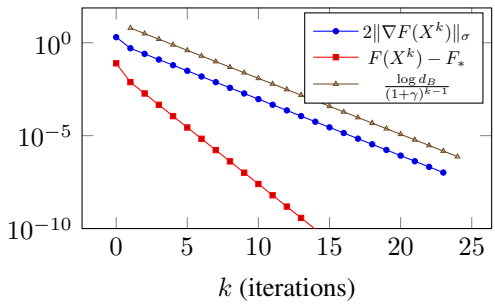


Figure 4: Convergence of the DCA (18) for $d_B = 512$



course, the IPM usually requires fewer iterations than the DCA to achive the same accuracy, which is why we use seconds rather than iteration count to compare the performance of the two algorithms.

Figure 2 illustrates the convergence of the DCA for a particular random H with $d_B = 64$. Although the convergence is much faster than the bound (16) suggests, we cannot prove a global rate of linear convergence, as F is not globally strongly convex relative to f_2.

5.2 Bregman Proximal Operator of the Negative Quantum Conditional Entropy

Next, we consider the proximal operator of the negative conditional entropy function with respect to the quantum relative entropy function $D(\cdot\|\cdot)$ in (13). Computing this operator corresponds to solving problems of the form:

$$\min_{X\succeq 0, \mathrm{tr}X=1} \{\mathrm{tr}(HX) - S(A|B)_X + \gamma D(X\|X_0)\} \tag{17}$$

where $\gamma > 0$ and X_0 is a positive definite density matrix. This is again a DC problem, with $f_1(X) := \mathrm{tr}(HX) - S(X) + D(X\|X_0) + \iota_{\{\mathrm{tr}(X)=1\}}$ and $f_2(X) = -S(\mathrm{tr}_A X)$. The DCA iteration is

$$X^{k+1} = \frac{1}{Z}\exp\left(t[I_A \otimes \log X_B^k - H] + (1-t)X_0\right) \tag{18}$$

where $t = \frac{1}{1+\gamma}$, $X_B^k = \mathrm{tr}_A X^k$ and Z is chosen so that $\mathrm{tr}(X^{k+1}) = 1$. Since $D(X\|X_0) + S(\mathrm{tr}_A X) = D(X\|I_A \otimes \mathrm{tr}_A X) + \mathrm{tr}[X(\log X_0)]$ is convex, $f_1 - (1+\gamma)f_2$ is convex for this problem. Therefore Corollary 2 applies, giving a global linear convergence rate of $(1+\gamma)^{-1}$, indeed with $X^0 = I/d$:

$$\begin{aligned} F(X^{k+1}) - F_* &\leq (1+\gamma)^{-k} D(\mathrm{tr}_A X^* \| \mathrm{tr}_A X^0) \\ &\leq (1+\gamma)^{-k}\log(d_B). \end{aligned} \tag{19}$$

Results are displayed in Figure 3 and Figure 4 for $\gamma = 1$ and $X_0 = I/d$. Note that, whereas the interior point method `Hypatia` fails to solve problems with $d_B \geq 64$, the DCA easily solves problems with $d_B = 512$, i.e. problems involving density matrices of size 1024×1024.

5.3 Standard Proximal Operator of the Negative Quantum Conditional Entropy

The DCA can be used to compute the Euclidean proximal operator of the negative conditional entropy function, namely:

$$\min_{\substack{X\succeq 0,\\ \mathrm{tr}X=1}} \left\{\mathrm{tr}(HX) - S(A|B)_X + \frac{\gamma}{2}\|X - X_0\|_F^2\right\}. \tag{20}$$

Compared with (17), the regularization term $D(X\|X_0)$ is replaced by $\frac{\gamma}{2}\|X - X_0\|_F^2$. Here, the DCA iteration is

$$X^{k+1} = \frac{1}{\gamma} W\left[\frac{1}{Z}\exp\left(I_A \otimes \log X_B^k - H + \gamma X_0\right)\right] \tag{21}$$

where W is the matrix Lambert W function, i.e. the inverse of $X \mapsto X\exp(X)$, and Z is a scalar which must be chosen so that $\mathrm{tr}(X^{k+1}) = 1$. We note that, since efficient implementations of the scalar Lambert W function exist, this update can be efficiently computed by means of an eigendecomposition of the matrix $-H + I_A \otimes \log \mathrm{tr}_A X^k + \gamma X_0$, combined with a scalar root-finding subroutine to determine Z.

In Appendix B, we illustrate the behaviour of this iteration with some numerical experiments, and we notice once again that the DCA is orders of magnitude faster than the IPM. We note however that unlike in Section 5.2, linear convergence is not guaranteed since $S(\mathrm{tr}_A X)$ is not strongly concave.

6 CONCLUSION AND FURTHER COMMENTS

In this paper, we revisited the difference-of-convex algorithm through the lens of Bregman divergences.

For DC functions which are convex, sublinear convergence results from the Bregman proximal point algorithm literature immediately translate to the DCA setting. When the DC decomposition of the (convex) objective function satisfies a strong relative convexity condition, we proved a linear convergence result, which we believe is a new result for the Bregman PPA.

For DC functions which may not be convex, we showed how existing sublinear convergence results can be understood, proved, and sometimes strengthened, by using Bregman divergences. We remark that some of the previously known guarantees on the DCA had rather technical proofs – for example Abbaszadehpeivasti et al. (2021) where the proofs are computer-assisted. We introduced a DC-specific version of the Polyak-Łojasiewicz inequality, and established sufficient conditions for these DC-PL inequalities to hold.

To illustrate some of the convergence results presented earlier in the paper, we considered a class of optimization problems over the set of positive semidefinite matrices which arise in quantum information theory. Specifically, the convex conjugate and the proximal operator of the negative quantum conditional entropy function have a natural DC structure. We observed that the DCA compares favourably with an implementation of an interior point algorithm capable of solving linear programs over the quantum relative entropy cone.

Another application where this framework is natural is maximum likelihood estimation of multivariate Gaussians subject to convex constraints on the covariance. In covariance parameterization, the negative log likelihood function is DC. When the domain is restricted to matrices sufficiently close (in an appropriate spectral sense) to the sample covariance, the objective function becomes convex Zwiernik et al. (2017), and even relatively strongly convex with respect to f_2 in the DC decomposition. A further example is computing the Brascamp-Lieb constant, which has a DC formulation which also happens to be geodesically convex Sra et al. (2018).

Acknowledgements

JS is the recipient of an Australian Research Council Discovery Early Career Researcher Award (project DE210101056) funded by the Australian Government.

References

Abbaszadehpeivasti, H., de Klerk, E., and Zamani, M. (2021). On the rate of convergence of the difference-of-convex algorithm (DCA). *arXiv preprint arXiv:2109.13566*. 4, 9

Bauschke, H., Bolte, J., Chen, J., Teboulle, M., and Wang, X. (2019). On linear convergence of non-Euclidean gradient methods without strong convexity and Lipschitz gradient continuity. *Journal of Optimization Theory and Applications*, 182. 2, 6

Bauschke, H. H., Bolte, J., and Teboulle, M. (2017). A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operational Research*, 42:330–348. 2, 3

Beck, A. (2017). *First-order methods in optimization*. SIAM. 2

Boumal, N. (2023). *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press. 6, 11

Bravyi, S., Chowdhury, A., Gosset, D., and Wocjan, P. (2021). On the complexity of quantum partition functions. *arXiv preprint arXiv:2110.15466*. 7

Censor, Y. and Zenios, S. A. (1992). Proximal minimization algorithm with D-functions. *J. Optim. Theory Appl.*, 73(3):451–464. 3

Chen, G. and Teboulle, M. (1993). Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM J. Optim.*, 3:538–543. 3

Coey, C., Kapelevich, L., and Vielma, J. P. (2021). Solving natural conic formulations with Hypatia.jl. 7

Coey, C., Kapelevich, L., and Vielma, J. P. (2022). Performance enhancements for a generic conic interior

point algorithm. *Mathematical Programming Computation*. 7

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In *European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 9851*, ECML PKDD 2016, page 795–811, Berlin, Heidelberg. Springer-Verlag. 6

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

Le Thi, H. A., Van Ngai, H., and Dinh, T. (2018). Convergence analysis of difference-of-convex algorithm with subanalytic data. *Journal of Optimization Theory and Applications*, 179. 1, 5

Lieb, E. H. and Ruskai, M. B. (1973). Proof of the strong subadditivity of quantum-mechanical entropy. *Journal of Mathematical Physics*, 14(12):1938–1941. 7

Lu, H., Freund, R. M., and Nesterov, Y. (2018). Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354. 2, 3, 4

Mordukhovich, B. S. and Nam, N. M. (2013). *An Easy Path to Convex Analysis and Applications*. Morgan & Claypool Publishers, 1st edition. 3

Nesterov, Y. and Nemirovskii, A. (1994). *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics. 12

Poulin, D. and Hastings, M. B. (2011). Markov entropy decomposition: a variational dual for quantum belief propagation. *Physical review letters*, 106(8):080403. 7

Renegar, J. (2001). A mathematical view of interior-point methods in convex optimization. In *MPS-SIAM series on optimization*. 6, 12

Sra, S., Vishnoi, N. K., and Yildiz, O. (2018). On Geodesically Convex Formulations for the Brascamp-Lieb Constant. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*, volume 116 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 25:1–25:15. 9

Tao, P. D. and An, L. H. (1997). Convex analysis approach to DC programming: theory, algorithms and applications. *Acta mathematica vietnamica*, 22(1):289–355. 1, 4

Wilde, M. M. (2013). *Quantum information theory*. Cambridge University Press. 7

Yuille, A. L. and Rangarajan, A. (2001). The concave-convex procedure (CCCP). In *Advances in Neural Information Processing Systems*, volume 14. MIT Press. 1

Yurtsever, A. and Sra, S. (2022). CCCP is Frank-Wolfe in disguise. *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. 4

Zwiernik, P., Uhler, C., and Richards, D. (2017). Maximum likelihood estimation for linear Gaussian covariance models. *Journal of the Royal Statistical Society Series B*, 79(4):1269–1292. 9

A PROOFS

A.1 Proof of Proposition 2

Proof of Proposition 2. We claim that the sequence (x^k) is a valid sequence of iterates for the DCA applied to the function $F + \iota_S$ with DC decomposition $(f_1 + \iota_S) - f_2$. First, note that $\varnothing \neq \operatorname{dom}(f_1 + \iota_S) \subseteq \operatorname{dom} f_2$, since we already had $\varnothing \neq \operatorname{dom} f_1 \subseteq \operatorname{dom} f_2$ by assumption $x^* \in S \cap \operatorname{dom} f_1$.

Next, we need to show that for each x^{k+1},

$$\partial(f_1 + \iota_S)(x^{k+1}) \supseteq \partial f_1(x^{k+1}).$$

This is certainly true whenever $x^{k+1} \in S$. Note that since DCA is a descent method, all iterates x^{k+1} remain in $\{x \mid F(x) \leq F(x^0)\}$ and hence in S.

We can now apply Corollary 2 to $F + \iota_S$, which completes the proof. □

A.2 Proof of Lemma 1

Proof of Lemma 1. It suffices to prove the claim for $k = 0$. From the DC Bregman PL inequality with $x = x^1 \in \mathcal{X}$ (because the DC algorithm is a descent method) we have that

$$\begin{aligned}
\eta_2(F(x^1) - F_*) &\leq D^{x^1}_{f_2^*}(\nabla f_1(x^1) | \nabla f_2(x^1)) \\
&= D^{x^1}_{f_2^*}(\nabla f_2(x^0) | \nabla f_2(x^1)) \\
&= f_2^*(\nabla f_2(x^0)) - f_2^*(\nabla f_2(x^1)) - \langle x^1, \nabla f_2(x^0) - \nabla f_2(x^1) \rangle \\
&= \langle \nabla f_2(x^0), x^0 \rangle - f_2(x^0) - \langle \nabla f_2(x^1), x^1 \rangle + f_2(x^1) - \langle x^1, \nabla f_2(x^0) - \nabla f_2(x^1) \rangle \\
&= D_{f_2}(x^1 | x^0) \\
&= (F(x^0) - F(x^1) - D_{f_1}(x^0 | x^1)).
\end{aligned}$$

We now apply the DC Bregman PL inequality with $x = x^0$ to obtain

$$\begin{aligned}
(F(x^0) - F(x^1) - D_{f_1}(x^0 | x^1)) &= ((F(x^0) - F_*) - (F(x^1) - F_*) - D^{x^0}_{f_1^*}(\nabla f_1(x^1) | \nabla f_1(x^0))) \\
&= ((F(x^0) - F_*) - (F(x^1) - F_*) - D^{x^0}_{f_1^*}(\nabla f_2(x^0) | \nabla f_1(x^0))) \\
&\leq ((F(x^0) - F_*) - (F(x^1) - F_*) - \eta_1(F(x^0) - F_*)).
\end{aligned}$$

Overall, then,

$$(\eta_2 + 1)(F(x^1) - F_*) \leq (1 - \eta_1)(F(x^0) - F_*)$$

which implies that

$$\frac{F(x^1) - F_*}{F(x^0) - F_*} \leq \frac{1 - \eta_1}{1 + \eta_2}.$$

□

A.3 Proof of Proposition 3

Proof of Proposition 3. It is known (Boumal, 2023, Lemma 11.28) that geodesic strong convexity of F implies

$$\|\operatorname{grad} F(x)\|_x^2 \geq 2\mu(F(x) - F_*). \tag{22}$$

Here $\operatorname{grad} F(x) = \nabla^2 f_2^*(x) \nabla F(x) = \nabla^2 f_2(x)^{-1} \nabla F(x)$ is the *Riemannian gradient* of F.

On the other hand,

$$D_{f_2^*}(\nabla f_1(x) | \nabla f_2(x)) = \nabla F(x)^\top \int_0^1 \int_0^t \nabla^2 f_2^*(\nabla f_2(x) + s \nabla F(x)) ds\, dt \nabla F(x).$$

Recall that f_2 is strongly nondegenerate self-concordant if, for every $x \in \operatorname{dom} f_2$ and every y satisfying $\|y-x\|_x < 1$, we have $y \in \operatorname{dom} f_2$ and

$$(1-\|y-x\|)^2\nabla^2 f_2(x) \preceq \nabla^2 f_2(y) \preceq \frac{1}{(1-\|y-x\|)^2}\nabla^2 f_2(x) \tag{23}$$

in the positive semidefinite order (Renegar, 2001, §2.5). From now on we follow Renegar in suppressing the words *strongly nondegenerate*. Recall also that a function is self-concordant if and only if its convex conjugate is (Nesterov and Nemirovskii, 1994, Theorem 2.4.1). Since f_2^* is self-concordant, we have that whenever $1 > \|\nabla F(x)\|_{x,*} = \|\operatorname{grad} F(x)\|_x$, then

$$\begin{aligned} D_{f_2^*}(\nabla f_1(x)|\nabla f_2(x)) &\geq \nabla F(x)^\top \int_0^1 \int_0^t (1-s\|\operatorname{grad} F(x)\|_x)^2 \nabla^2 f_2^*(x) ds\, dt\ \nabla F(x) \\ &\geq \frac{7-4\|\operatorname{grad} F(x)\|_x}{12}\|\operatorname{grad} F(x)\|_x^2 \\ &\geq \frac{1}{4}\|\operatorname{grad} F(x)\|_x^2 \\ &\geq \frac{\mu}{2}(F(x)-F_*). \end{aligned}$$

□

Remark 5. *The only place where we used strong geodesic convexity was to obtain the Riemannian PL inequality* (22). *This is in general weaker than strong geodesic convexity.*

A.4 Proof of Proposition 4

The following lemma will be used to prove Proposition 4.

Lemma 2. *Suppose that f_1 and f_2 are C^2 and strictly convex on an open convex set $S \subseteq \mathbb{R}^n$. Let $F := f_1 - f_2$, and let $x^* \in S$ be such that $\nabla F(x^*) = 0$. Then for $i = 1, 2$*

$$D_{f_i^*}(\nabla f_j(x)|\nabla f_i(x)) = \frac{1}{2}(x-x^*)^\top \nabla^2 F(x^*)\nabla^2 f_i(x^*)^{-1}\nabla^2 F(x^*)(x-x^*) + o(\|x-x^*\|^2),$$

where $j := 3-i$.

Proof. Since f_i is C^2 and strictly convex, $\nabla^2 f_i^*(\nabla f_i(x^*))$ exists and is equal to $\nabla^2 f_i(x^*)^{-1}$. Writing $v_1 = \nabla f_1(x) - \nabla f_1(x^*)$ and $v_2 = \nabla f_2(x) - \nabla f_2(x^*)$, we have by Taylor's theorem

$$\begin{aligned} D_{f_i^*}(\nabla f_j(x)|\nabla f_i(x)) &= f_i^*(\nabla f_j(x)) - f_i^*(\nabla f_i(x)) - \langle x, \nabla f_j(x) - \nabla f_i(x)\rangle \\ &= \langle x^*, v_j - v_i\rangle + \frac{1}{2}v_j^\top \nabla^2 f_i^*(\nabla f_i(x^*))v_j - \frac{1}{2}v_i^\top \nabla^2 f_i^*(\nabla f_j(x^*))v_i - \langle x, v_j - v_i\rangle \\ &\qquad + o(\|v_i\|^2, \|v_j\|^2) \\ &= \frac{1}{2}v_j^\top \nabla^2 f_i(x^*)^{-1}v_j - \frac{1}{2}v_i^\top \nabla^2 f_i(x^*)^{-1}v_i + \langle x-x^*, v_j - v_i\rangle + o(\|v_i\|^2, \|v_j\|^2). \end{aligned}$$

We have $v_1 = \nabla^2 f_1(x^*)(x-x^*) + o(\|x-x^*\|)$ and $v_2 = \nabla^2 f_2(x^*)(x-x^*) + o(\|x-x^*\|)$. Therefore

$$D_{f_i^*}(\nabla f_j(x)|\nabla f_i(x)) = \frac{1}{2}(x-x^*)^\top \nabla^2 F(x^*)\nabla^2 f_i(x^*)^{-1}\nabla^2 F(x^*)(x-x^*) + o(\|x-x^*\|^2).$$

□

Proof of Proposition 4. We have $F(x) - F(x^*) = \frac{1}{2}(x-x^*)^\top \nabla^2 F(x^*)(x-x^*) + o(\|x-x^*\|^2)$. By Lemma 2

$$D_{f_i^*}(\nabla f_j(x)|\nabla f_i(x)) = \frac{1}{2}(x-x^*)^\top \nabla^2 F(x^*)\nabla^2 f_i(x^*)^{-1}\nabla^2 F(x^*)(x-x^*) + o(\|x-x^*\|^2),$$

for $i = 1, 2$ where $j := 3 - i$. It follows that the DC PL inequality implies

$$\eta_i \nabla^2 F(x^*) \preceq \nabla^2 F(x^*) \nabla^2 f_i(x^*)^{-1} \nabla^2 F(x^*).$$

Since $\nabla^2 F(x^*)$ is positive definite, this is equivalent to

$$\eta_i \nabla^2 f_i(x^*) \preceq \nabla^2 F(x^*) = \nabla^2 f_1(x^*) - \nabla^2 f_2(x^*),$$

which completes the proof. □

B ADDITIONAL NUMERICAL EXPERIMENTS

Here we present the results of numerical experiments which we could not place in Section 5.3 due to space constraints. We consider the problem (20) and the DCA iteration (21) for solving it. Results are displayed in Figure 5 and Figure 6 for $\gamma = 1$ and $X_0 = I/d$.

Figure 5: Time Taken By DCA vs IPM To Solve (20)

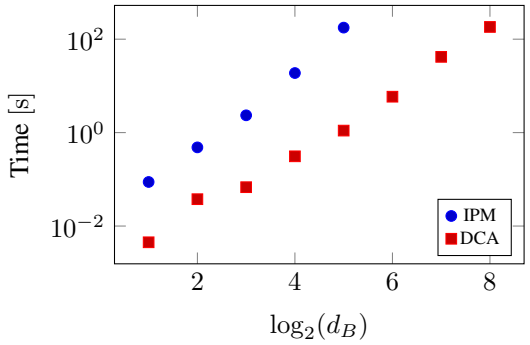


Figure 6: Convergence of the DCA for (20) with $d_B = 256$