Anisotropic Proximal Gradient

Emanuel Laude* Panagiotis Patrinos*

November 2, 2022

Abstract

This paper studies a novel algorithm for nonconvex composite minimization which can be interpreted in terms of dual space nonlinear preconditioning for the classical proximal gradient method. The proposed scheme can be applied to composite minimization problems whose smooth part exhibits an anisotropic descent inequality relative to a reference function. In the convex case this is a dual characterization of relative strong convexity in the Bregman sense. It is proved that the anisotropic descent property is closed under pointwise average if the dual Bregman distance is jointly convex and, more specifically, closed under pointwise conic combinations for the KL-divergence. We analyze the method's asymptotic convergence and prove its linear convergence under an anisotropic proximal gradient dominance condition. This is implied by anisotropic strong convexity, a recent dual characterization of relative smoothness in the Bregman sense. Applications are discussed including exponentially regularized LPs and logistic regression with nonsmooth regularization. In the LP case the method can be specialized to the Sinkhorn algorithm for regularized optimal transport and a classical parallel update algorithm for AdaBoost. Complementary to their existing primal interpretations in terms of entropic subspace projections this provides a new dual interpretation in terms of forward-backward splitting with entropic preconditioning.

Keywords: Bregman distance · proximal gradient method · duality · generalized convexity · DC-programming

AMS Subject Classification: 65K05 · 49J52 · 90C30

1 Introduction

1.1 Motivation and prior work

A classical first-order approach for the minimization of an additive composite problem is the celebrated proximal gradient algorithm. In particular, if the gradient mapping of the smooth part of the cost function is Lipschitz continuous the algorithm converges with a constant step-size which avoids a possibly expensive, in terms of function evaluations, linesearch procedure. However, there are many functions which do not exhibit a globally Lipschitz continuous gradient mapping and thus proximal gradient does not converge with a constant step-size.

*KU Leuven, Department of Electrical Engineering (ESAT-STADIUS), Leuven, Belgium (emanuel.laude,panos.patrinos@esat.kuleuven.be)

As a remedy Lipschitz smoothness was generalized to the notion of relative[1] smoothness [10, 6, 26]. Such functions exhibit a Bregman version of the classical descent inequality and can be optimized with a Bregman version of the proximal gradient method [10, 6, 26] using a constant step-size. For instance this allows one to consider functions whose Hessians grow like a polynomial in x or $1/x$ [26]. If, in addition, the function exhibits global lower bounds relative to the reference function the method converges even linearly [26, 5]. However, there exist many cost functions which are neither Lipschitz smooth nor Bregman smooth relative to a simple reference function. This leads us to consider functions that exhibit an anisotropic descent inequality [24, 23]. This gives rise to a different generalization of the classical proximal gradient algorithm which is studied in this paper. In the convex case the algorithm can be interpreted as a proximal extension of *dual space preconditioning* [27] for gradient descent.

The results of this work are summarized as follows:

1.2 Summary of results

- In section 2 we recapitulate the notion of anisotropic smoothness [24, 23] and introduce the proposed algorithm. In particular, we shall discuss the well-definedness of the iterates under certain constraint qualifications.

- In section 3 we recapitulate the basic notions of Φ-convexity and Φ-concavity and, following [33, 24], show that the anisotropic descent inequality can be formulated equivalently in terms of Φ-concavity. This is used heavily for developing its calculus and allows us to provide an equivalent interpretation of the algorithm in terms of a *difference of Φ-convex approach*.

- In section 4 we develop a calculus for the anisotropic descent property and provide examples. In particular we show that the anisotropic descent property is closed under pointwise average if the dual Bregman distance is jointly convex. This builds upon recent advances for the Bregman proximal average [43]. A refinement is proved for the exponential reference function which shows that in contrast to Euclidean Lipschitz smoothness the constant in the anisotropic descent property is invariant under pointwise conic combinations.

- In section 5 we show subsequential convergence of our method under anisotropic smoothness. As a measure of stationarity we employ a regularized gap function which vanishes at rate $\mathcal{O}(1/K)$. This technique allows one to show that $1/L$ is a feasible step-size even if both terms are nonconvex, as long as an anisotropic prox-boundedness condition holds true. To our knowledge this is also new in the Euclidean case where existing analyses require the step-size to be smaller than $1/L$. We also discuss a linesearch variant of the method if the smoothness constant is unknown. Furthermore, we prove the linear convergence of the anisotropic proximal gradient method under a non-Euclidean generalization of the proximal gradient dominated condition [21]. This is closely related to the PL-inequality. We prove that the anisotropic proximal gradient dominated condition is implied by anisotropic strong convexity of the smooth part, which was recently shown to be equivalent to dual relative smoothness in the Bregman sense [24].

[1] relative to a "simple" reference function whose inverse gradient can be evaluated easily

- In section 6 we examine an equivalent reformulation of the proposed algorithm in terms of a difference of Φ-convex approach. This is leveraged to prove linear convergence if the nonsmooth part of the objective is strongly convex in the anisotropic sense.

- In section 7 we consider applications based on the examples developed in section 4. For the exponential reference function we reveal connections to existing algorithms for regularized optimal transport [15, 9] and AdaBoost [18, 37, 13]. Exploiting the freedom of a nonzero composite term we provide examples beyond these existing applications.

1.3 Notation and preliminaries

We denote by $\langle \cdot, \cdot \rangle$ the standard Euclidean inner product on $\mathbb{R}^n$ and by $\|x\| := \sqrt{\langle x, x \rangle}$ for any $x \in \mathbb{R}^n$ the standard Euclidean norm on $\mathbb{R}^n$. In accordance with [30, 36] we extend the classical arithmetic on $\mathbb{R}$ to the extended real line $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$. We define upper addition $-\infty \dot{+} \infty = \infty$ and lower addition $-\infty \dot{+} \infty = -\infty$, and accordingly upper subtraction $\infty \dot{-} \infty = \infty$ and lower subtraction $\infty \dot{-} \infty = -\infty$. The effective domain of an extended real-valued function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is denoted by $\operatorname{dom} f := \{x \in \mathbb{R}^n : f(x) < \infty\}$, and we say that f is proper if $\operatorname{dom} f \neq \emptyset$ and $f(x) > -\infty$ for all $x \in \mathbb{R}^n$; lower semicontinuous (lsc) if $f(\bar{x}) \leq \liminf_{x \to \bar{x}} f(x)$ for all $\bar{x} \in \mathbb{R}^n$; super-coercive if $f(x)/\|x\| \to \infty$ as $\|x\| \to \infty$. We define by $\Gamma_0(\mathbb{R}^n)$ the class of all proper, lsc convex functions $f : \mathbb{R}^n \to \overline{\mathbb{R}}$. For any functions $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ we define the infimal convolution or epi-addition of f and g as $(f \boxplus g)(x) = \inf_{y \in \mathbb{R}^n} g(x - y) + f(y)$. For a proper function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $\lambda \geq 0$ we define the epi-scaling $(\lambda \star f)(x) = \lambda f(\lambda^{-1} x)$ for $\lambda > 0$ and $(\lambda \star f)(x) = \delta_{\{0\}}(x)$ otherwise. We adopt the operator precedence for epi-multiplication and epi-addition from the pointwise case. We denote by $\mathbb{R}^n_+ = [0, +\infty)^n$ the nonnegative orthant and by $\mathbb{R}^n_{++} = (0, +\infty)^n$ the positive orthant. We say that the function $\xi(x, y)$ is level-bounded in x locally uniformly in y in the sense of [36, Definition 1.16] if for every $\bar{y} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ we can find a neighborhood $V \in \mathcal{N}(\bar{y})$ along with a bounded set $B \subset \mathbb{R}^n$ such that $\{x \in \mathbb{R}^n \mid \xi(x, y) \leq \alpha\} \subseteq B$ for all $y \in V$. Let $g_- := g(-(\cdot))$ denote the reflection of g. Since convex conjugation and reflection commute we adopt the notation: $(g_-)^* = (g^*)_- =: g^*_-$. We adopt the notions of essential smoothness, essential strict convexity and Legendre type functions from [35, Section 26]: We say that a function $f \in \Gamma_0(\mathbb{R}^n)$ is *essentially smooth*, if $\operatorname{int}(\operatorname{dom} f) \neq \emptyset$ and f is differentiable on $\operatorname{int}(\operatorname{dom} f)$ such that $\|\nabla f(x^\nu)\| \to \infty$, whenever $\operatorname{int}(\operatorname{dom} f) \ni x^\nu \to x \in \operatorname{bdry} \operatorname{dom} f$, and *essentially strictly convex*, if f is strictly convex on every convex subset of $\operatorname{dom} \partial f$, and *Legendre type*, if f is both essentially smooth and essentially strictly convex. Otherwise we adopt the notation from [36].

Our concepts and algorithm involve a reference function ϕ which complies with the following standing requirement which is assumed valid across the entire paper unless stated otherwise:

(A1) $\phi \in \Gamma_0(\mathbb{R}^n)$ is of Legendre type with $\operatorname{dom} \phi = \mathbb{R}^n$.

The dual reference function ϕ^* is of Legendre type as well [35, Theorem 26.3] but does not necessarily have full domain. Instead, thanks to a conjugate duality between super-coercivity and full domain [7, Proposition 2.16], ϕ^* is super-coercive. The gradient mapping $\nabla \phi : \mathbb{R}^n \to \operatorname{int} \operatorname{dom} \phi^*$ is a diffeomorphism between $\mathbb{R}^n$ and $\operatorname{int} \operatorname{dom} \phi^*$ with inverse $(\nabla \phi)^{-1} = \nabla \phi^*$ [35, Theorem 26.5].

A leading example considered in this paper is the exponential reference function:

Example 1.1. Let $\phi : \mathbb{R}^n \to \mathbb{R}$ defined by $\phi(x) = \mathrm{Exp}(x) := \sum_{j=1}^n \exp(x_j)$. Then $\phi \in \Gamma_0(\mathbb{R}^n)$ is Legendre type with full domain and thus complies with our standing assumption. The convex conjugate ϕ^* is the von Neumann entropy defined by

$$\phi^*(x) = H(x) := \begin{cases} \sum_{j=1}^n x_j \log(x_j) - x_j & \text{if } x \in \mathbb{R}^n_+ \\ +\infty & \text{otherwise,} \end{cases}$$

where $0 \log(0) := 0$.

For the gradients we have $\nabla\phi(x) = (\exp(x_j))_{j=1}^n$ and $\nabla\phi^*(x) = (\log(x_j))_{j=1}^n$.

In the course of the paper we often consider the Bregman distance D_{ϕ^*} generated by the dual reference function ϕ^*:

Definition 1.2 (Bregman distance)**.** *We define the Bregman distance generated by ϕ^* as:*

$$D_{\phi^*}(x, y) := \begin{cases} \phi^*(x) - \phi^*(y) - \langle \nabla\phi^*(y), x - y \rangle & \text{if } x \in \operatorname{dom}\phi^*, y \in \operatorname{int}\operatorname{dom}\phi^* \\ +\infty & \text{otherwise.} \end{cases}$$

Thanks to [7, Theorem 3.7(iv)] we have that $D_{\phi^*}(x, y) = 0$ if and only if $x = y$.

For $\phi = \mathrm{Exp}$ being the exponential reference function as in Example 1.1 the Bregman distance D_{ϕ^*} is the classical KL-divergence. Thanks to [7, Theorem 3.7(v)] have the following key relation between D_{ϕ^*} and D_ϕ:

Lemma 1.3. *We have that the identity*

$$D_{\phi^*}(x, y) = D_\phi(\nabla\phi^*(y), \nabla\phi^*(x)),$$

holds true for any $x, y \in \operatorname{int}\operatorname{dom}\phi^*$.

2 Anisotropic proximal gradient

2.1 Anisotropic descent inequality

The algorithm that is developed in this paper is based on the following anisotropic descent property generalizing [24] to reference functions ϕ which are not necessarily super-coercive. Up to sign-flip it can be regarded as a globalization of anisotropic prox-regularity [23, Definition 2.13] specialized to smooth functions:

Definition 2.1 (anisotropic descent inequality)**.** *Let $f \in \mathcal{C}^1(\mathbb{R}^n)$ such that the following constraint qualification holds true*

$$\operatorname{rge}\nabla f \subseteq \operatorname{rge}\nabla\phi. \tag{1}$$

Then we say that f satisfies the anisotropic descent property relative to ϕ with constant $L > 0$ if for all $\bar{x} \in \mathbb{R}^n$

$$f(x) \leq f(\bar{x}) + \tfrac{1}{L} \star \phi(x - \bar{x} + L^{-1}\nabla\phi^*(\nabla f(\bar{x}))) - \tfrac{1}{L} \star \phi(L^{-1}\nabla\phi^*(\nabla f(\bar{x}))) \quad \forall x \in \mathbb{R}^n. \tag{2}$$

If $L = 1$ we say that f satisfies the anisotropic descent property relative to ϕ without further mention of L.

Remark 2.2. Since $\nabla(L^{-1} \star \phi)^* = L^{-1}\nabla\phi^*$, saying that f satisfies the anisotropic descent property relative to ϕ with constant $L > 0$ means equivalently that f satisfies the anisotropic descent property relative to $\frac{1}{L} \star \phi$ (with constant 1).

The constraint qualification (1) ensures that the expression $\nabla\phi^*(\nabla f(\bar{x})) \in \mathbb{R}^n$ is well defined. If we choose $x := \bar{x}$ the descent inequality (2) collapses to $f(\bar{x}) \leq f(\bar{x})$ and thus

$$x \mapsto f(\bar{x}) + \tfrac{1}{L} \star \phi(x - \bar{x} + L^{-1}\nabla\phi^*(\nabla f(\bar{x}))) - \tfrac{1}{L} \star \phi(L^{-1}\nabla\phi^*(\nabla f(\bar{x}))),$$

as a function in x is an upper bound or majorizer of f at $\bar{x}$. With some abuse of terminology we often refer to a function that only has anisotropic upper bounds as anisotropically smooth. This is different from [24] which assumes the existence of both, upper and lower bounds. For convex functions, however, the lower bound inequality holds automatically and thus the two notions coincide.

The right-hand side of the anisotropic descent inequality can be rewritten in terms of a proximal linearization of f similar to the Bregman descent inequality [10, 6, 26] where the arguments in the Bregman distance are shifted by the point $\bar{y} = \bar{x} - L^{-1}\nabla\phi^*(\nabla f(\bar{x}))$. In particular this guarantees the following shift-invariance property which is in general lacking in the Bregman descent inequality:

Remark 2.3 (shift-invariance and generalization of Euclidean descent lemma). Assume that f has the anisotropic descent property. By definition of the Bregman distance (2) can be rewritten as:

$$f(x) \leq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + D_{L^{-1}\star\phi}(x - \bar{y}, \bar{x} - \bar{y}),$$

for all $x, \bar{x} \in \mathbb{R}^n$ and $\bar{y} = \bar{x} - L^{-1}\nabla\phi^*(\nabla f(\bar{x}))$.

Choose $u, \bar{u} \in \mathbb{R}^n$ and define $x = u - a$ and $\bar{x} = \bar{u} - a$. Then the above inequality yields

$$f(u - a) \leq f(\bar{u} - a) + \langle \nabla f(\bar{u} - a), u - \bar{u} \rangle + D_{L^{-1}\star\phi}(u - \bar{w}, \bar{u} - \bar{w}),$$

for $\bar{w} = \bar{u} - L^{-1}\nabla\phi^*(\nabla f(\bar{u} - a))$. Thus the shifted version $f(\cdot - a)$ of f also has the anisotropic descent property. For $\phi = \frac{1}{2}\|\cdot\|^2$ the upper bound inequality (2) specializes to the classical descent lemma:

$$f(x) \leq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \tfrac{L}{2}\|x - \bar{x}\|^2.$$

Next we will introduce the algorithm and discuss the well-definedness of its iterates.

2.2 Definition, well-definedness and standing assumptions

In this paper our goal is to develop a novel algorithm for nonconvex composite minimization problems of the form:

$$\text{minimize } F := f + g, \tag{P}$$

where

(A2) $f \in \mathcal{C}^1(\mathbb{R}^n)$ has the anisotropic descent property with constant $L > 0$;

(A3) $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper lsc;

(A4) $\inf F > -\infty$,

holds without further mention unless stated otherwise.

This suggests the following iterative procedure for minimizing the cost function: In each iteration we minimize the anisotropic upper bound at the current iterate x^k to compute the next iterate x^{k+1}:

$$x^{k+1} = \operatorname*{arg\,min}_{x\in\mathbb{R}^n} f(x^k) + \tfrac{1}{L} \star \phi(x - x^k + L^{-1}\nabla\phi^*(\nabla f(x^k))) \\ - \tfrac{1}{L} \star \phi(L^{-1}\nabla\phi^*(\nabla f(x^k))) + g(x). \tag{3}$$

Choose $\lambda > 0$. Introducing a new variable $y^k := x^k - \lambda\nabla\phi^*(\nabla f(x^k))$ we can rewrite the update (3) in terms of a Gauss–Seidel like scheme listed in Algorithm 1.

Algorithm 1 Anisotropic proximal gradient

Require: Let $\lambda > 0$. Let $x^0 \in \mathbb{R}^n$.

for all $k = 0, 1, \ldots$ **do**

$$y^k = x^k - \lambda\nabla\phi^*(\nabla f(x^k)) \tag{4}$$

$$x^{k+1} \in \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \lambda \star \phi(x - y^k) + g(x), \tag{5}$$

end for

We refer to (4) as the anisotropic *forward-step* and (5) is the anisotropic *backward-step*.

Example 2.4. For $\phi \in \Gamma_0(\mathbb{R}^n)$ being a Euclidean norm defined as $\phi(x) = \frac{1}{2}\|x\|_M^2 := \frac{1}{2}\langle x, Mx\rangle$ with M symmetric positive definite we have $\nabla\phi^*(x) = M^{-1}x$. Thus the algorithm becomes a scaled version of the Euclidean proximal gradient method:

$$y^k = x^k - \lambda M^{-1}\nabla f(x^k) \tag{6}$$

$$x^{k+1} = \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \frac{1}{2\lambda}\|x - y^k\|_M^2 + g(x), \tag{7}$$

where the backward-step is a scaled Euclidean proximal mapping.

More generally, the backward-step amounts to the *left* anisotropic proximal mapping of g at y^k which is formally defined follows:

Definition 2.5 (left and right anisotropic proximal mapping and Moreau envelope)**.** *Let $\lambda > 0$ and $x \in \mathbb{R}^n$. Then the right anisotropic proximal mapping of g with parameter λ at x is defined as:*

$$\operatorname{aprox}_g^{\lambda\star\phi}(x) := \operatorname*{arg\,min}_{y\in\mathbb{R}^n} \lambda \star \phi(x - y) + g(y), \tag{8}$$

and the right anisotropic Moreau envelope is the infimal convolution of g and $\lambda \star \phi$:

$$(g \boxplus \lambda \star \phi)(x) = \inf_{y\in\mathbb{R}^n} \lambda \star \phi(x - y) + g(y). \tag{9}$$

The left anisotropic proximal mapping and Moreau envelope of g with parameter λ are defined as

$$y \mapsto \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \lambda \star \phi(x - y) + g(x) \quad \textit{and} \quad y \mapsto \inf_{x\in\mathbb{R}^n} \lambda \star \phi(x - y) + g(x). \tag{10}$$

The reflection ϕ_- of ϕ allows us to express the left anisotropic proximal mapping and Moreau envelope in terms of the right anisotropic proximal mapping and Moreau envelope aka infimal convolution:

$$\operatorname*{arg\,min}_{x\in\mathbb{R}^n} \lambda\star\phi(x-y)+g(x) = \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \lambda\star\phi_-(y-x)+g(x) = \operatorname{aprox}_g^{\lambda\star\phi_-}(y), \tag{11}$$

$$\inf_{x\in\mathbb{R}^n} \lambda\star\phi(x-y)+g(x) = \inf_{x\in\mathbb{R}^n} \lambda\star\phi_-(y-x)+g(x) = (g\boxplus\lambda\star\phi_-)(x). \tag{12}$$

If $\phi(x)=\frac{1}{2}\|x\|^2$ left and right anisotropic proximal mapping and Moreau envelope both coincide with the Euclidean proximal mapping and Moreau envelope.

The local continuity of the anisotropic proximal mapping and continuous differentiability of the anisotropic Moreau envelope was studied in [25, 23] under (anisotropic) prox-regularity however, using the pointwise scaling in place of the epi-scaling. For both scalings a sufficient condition for the non-emptyness of the proximal mapping and in particular the continuity of the anisotropic Moreau envelope aka infimal convolution $g\boxplus\lambda\star\phi_-$ is level-boundedness of $\xi(x,y):=g(x)+\lambda\star\phi(x-y)$ in x locally uniformly in y [36, Theorem 1.17].

Remark 2.6 (threshold of anisotropic prox-boundedness). Without loss of generality we can assume that $\phi(0)=0$ since a replacement of ϕ with the perturbed reference function $\phi(x)-\phi(0)$ does neither affect the descent inequality Definition 2.1 nor the updates in the algorithm. Then, thanks to [12, Lemma 4.4], $\lambda_1\star\phi(x-y)\leq\lambda_2\star\phi(x-y)$ for $\lambda_2<\lambda_1$ and thus uniform level-boundedness of $\xi(x,y;\lambda):=g(x)+\lambda\star\phi(x-y)$ for $\lambda=\lambda_0$ implies uniform level-boundedness of $\xi(x,y;\lambda)$ for any $\lambda<\lambda_0$. This leads us to define the threshold of anisotropic prox-boundedness:

$$\lambda_g := \sup\{\lambda>0 \mid \xi(x,y;\lambda) \text{ is level-bounded in } x \text{ locally uniformly in } y\}. \tag{13}$$

Note that $\lambda_g=+\infty$ if g (or ϕ) is bounded from below and ϕ (or g) is coercive.

We make the following additional assumption on λ_g and λ to ensure well-definedness of the proximal mapping and continuity of the anisotropic Moreau envelope:

(A5) Assume that $\lambda_g>0$ and choose $\lambda>0$ such that $\lambda<\lambda_g$ and $\lambda\leq 1/L$.

More specifically, if g is convex continuity of the anisotropic proximal mapping and Moreau envelope is implied by validity of a *constraint qualification* (CQ):

(A6) g is convex and $\emptyset\neq\operatorname{dom} g^*\cap\operatorname{int}(\operatorname{dom}\phi_-^*)$ and $\lambda\leq 1/L$.

Henceforth, whenever g is assumed to be convex we impose assumption (A6) in place of assumption (A5). In fact, assumption (A6) implies that $\lambda_g=+\infty$ as will be shown in Remark 2.8.

Note that for $g\equiv 0$ the CQ becomes $0\in\operatorname{int}\operatorname{dom}\phi^*$ and the algorithm reduces to $x^{k+1}=x^k-\lambda\nabla\phi^*(\nabla f(x^k))+\lambda\nabla\phi^*(0)$. This is equivalent to *dual space preconditioning* for gradient descent [27], where ϕ is replaced with $\phi-\langle\nabla\phi(0),\cdot\rangle$.

For g convex under assumption (A6) thanks to [14, Theorem 3.1(ii)] and [43, Proposition 2.16], originally due to [40, Theorem 3.2] we have the following generalization of Moreau's decomposition. This allows us to compute the anisotropic proximal mapping through known closed form expressions of the Bregman proximal mapping. In addition, the anisotropic Moreau envelope is $\mathcal{C}^1(\mathbb{R}^n)$ and the anisotropic proximal mapping is continuous as summarized in the following lemma:

Lemma 2.7 (Bregman–Moreau decomposition and continuity of anisotropic proximal mapping)**.** *Let* $h \in \Gamma_0(\mathbb{R}^n)$*. Assume that* $\emptyset \neq \operatorname{dom} h^* \cap \operatorname{int} \operatorname{dom} \phi^*$*. Define by* $\operatorname{bprox}_{h^*}^{\lambda\phi^*} := \arg\min_{y\in\mathbb{R}^n} h^*(y) + \lambda D_{\phi^*}(y,\cdot)$ *the left Bregman proximal mapping of* h^* *with parameter* λ^{-1} *and reference function* ϕ^**. Then the following statements hold:*

(i) $\operatorname{id} = \operatorname{aprox}_h^{\lambda\star\phi} + \lambda\nabla\phi^* \circ \operatorname{bprox}_{h^*}^{\lambda\phi^*} \circ \nabla\phi \circ \lambda^{-1} \operatorname{id}$*;*

(ii) $h \boxplus \lambda \star \phi \in \mathcal{C}^1(\mathbb{R}^n)$ *is convex with* $\nabla(h \boxplus \lambda \star \phi) = \nabla\phi \circ \lambda^{-1}(\operatorname{id} - \operatorname{aprox}_h^{\lambda\star\phi})$*;*

(iii) $\operatorname{bprox}_{h^*}^{\lambda\phi^*} \circ \nabla(\lambda \star \phi) = (\partial h^* + \lambda\nabla\phi^*)^{-1} = \nabla(h \boxplus \lambda \star \phi)$*;*

(iv) $\operatorname{aprox}_h^{\lambda\star\phi} = (\operatorname{id} + \lambda\nabla\phi^* \circ \partial h)^{-1}$.

The result is a slight refinement of [14, Theorem 3.1(ii)] and [43, Proposition 2.16]. For completeness a full proof is provided in Appendix A.1.

Remark 2.8. Assumption (A6) for convex g implies that $\xi(x,y)$ is level-bounded in x locally uniformly in y for any $\lambda > 0$ and thus in particular $\lambda_g = +\infty$, i.e., assumption (A5) holds true for any $\lambda \leq 1/L$.

This can be proved as follows: Let $\lambda > 0$. Rewrite

$$\begin{aligned}\xi(x,y) &= g(x) + \lambda \star \phi(x-y)\\ &= (g + \langle\nabla\phi(0),\cdot\rangle)(x) + \lambda \star (\phi - \langle\nabla\phi(0),\cdot\rangle)(x-y) - \langle\nabla\phi(0), y\rangle\\ &=: g_0(x) + \lambda \star \phi_0(x-y) - \langle\nabla\phi(0), y\rangle,\end{aligned}$$

which replaces ϕ with $\phi_0 := \phi - \langle\nabla\phi(0),\cdot\rangle$ and g with $g_0 := g + \langle\nabla\phi(0),\cdot\rangle$. Thus we have $\nabla\phi_0(0) = 0$. Since ϕ is Legendre type ϕ_0 is Legendre type as well and thus $\operatorname{rge}\nabla\phi_0 = \operatorname{int}(\operatorname{dom}\phi_0^*)$. This implies that $0 \in \operatorname{int}(\operatorname{dom}\phi_0^*)$ and in view of [36, Theorem 11.8(c)] ϕ_0 is level-bounded.

Assume that $\xi(x,y)$ is not level-bounded in x locally uniformly in y. This means that there is $\bar{y} \in \mathbb{R}^n$ and $\alpha > 0$ and we can find sequences $\|x^\nu\| \to \infty$ and $y^\nu \to \bar{y}$ such that

$$g_0(x^\nu) + \lambda \star \phi_0(x^\nu - y^\nu) - \langle\nabla\phi(0), y^\nu\rangle = \xi(x^\nu, y^\nu) \leq \alpha. \tag{14}$$

Let $\bar{v} \in \operatorname{dom} g^* \cap \operatorname{int}\operatorname{dom}\phi_-^*$ which exists by assumption. Since $g_0^* = g(\cdot - \nabla\phi(0))$, $\phi_0^* = \phi(\cdot + \nabla\phi(0))$ and $(\phi_0)_-^* = \phi_-^*(\cdot - \nabla\phi(0))$ this implies that $\bar{v}_0 := \bar{v} + \nabla\phi(0) \in \operatorname{dom} g_0^* \cap \operatorname{int}\operatorname{dom}(\phi_0)_-^*$. Since $\operatorname{int}\operatorname{dom}\phi_0^*$ is open $\bar{v}_0/\tau \in \operatorname{int}\operatorname{dom}(\phi_0)_-^*$ for $\tau < 1$ sufficiently large implying that $\bar{v}_0 \in \operatorname{int}\operatorname{dom}(\tau\star(\phi_0)_-^*) \cap \operatorname{dom} g_0^*$. Then we can invoke Lemma 2.7(ii) for $\tau(\phi_0)_-$ and g_0 and obtain that $g_0 \boxplus \lambda \star (\tau(\phi_0)_-)$ is continuous. Thus there is a local uniform lower bound $\beta \in \mathbb{R}$ such that $\beta \leq (g_0 \boxplus \lambda\star(\tau(\phi_0)_-))(y^\nu) = \inf_{x\in\mathbb{R}^n} g_0(x) + \tau(\lambda\star\phi_0)(x-y^\nu) \leq g_0(x^\nu) + \tau(\lambda\star\phi_0)(x^\nu - y^\nu)$ for ν sufficiently large. Summing with (14) we obtain

$$(1-\tau)(\lambda \star \phi_0)(x^\nu - y^\nu) - \langle\nabla\phi(0), y^\nu\rangle \leq \alpha - \beta.$$

Since ϕ_0 is level-bounded passing to $\nu \to \infty$ we obtain that $\infty \leq \alpha - \beta$, a contradiction.

3 Anisotropic smoothness and generalized concavity

In this section we recapitulate the existing notions of Φ-convexity and Φ-conjugacy [31] which are used heavily as tools in the remainder of the manuscript. Classically, these notions appear in the context of eliminating duality gaps in nonconvex and nonsmooth optimization or

optimal transport theory; see, e.g., [34, 33, 42, 2]. Following [33, 24], we will interpret the anisotropic descent inequality in terms of Φ-concavity where the infimal convolution operation and its inverse, the infimal deconvolution [20] form the underlying pair of Φ-conjugate transforms. The Φ-concavity interpretation of anisotropic smoothness is used in the development of its calculus and the discovery of practical examples. In addition, this allows us to provide an equivalent interpretation of the algorithm in terms of a difference of Φ-convex approach.

Definition 3.1 (Φ-convex and Φ-concave functions)**.** *Let X and Y be nonempty sets and $\Phi : X \times Y \to \mathbb{R}$ a real-valued coupling. Let $h_X : X \to \overline{\mathbb{R}}$ and $h_Y : Y \to \overline{\mathbb{R}}$. We say that h_X is Φ-convex on X if there is an index set $\mathcal{I}$ and parameters $(y_i, \beta_i) \in Y \times \overline{\mathbb{R}}$ for $i \in \mathcal{I}$ such that*

$$h_X(x) = \sup_{i \in \mathcal{I}} \Phi(x, y_i) - \beta_i \quad \forall x \in X. \tag{15}$$

If $\mathcal{I} = \emptyset$ we say that $h_X \equiv -\infty$ is Φ-convex on X by definition. We say that h_Y is Φ-convex on Y if there is an index set $\mathcal{J}$ and parameters $(x_j, \alpha_j) \in X \times \overline{\mathbb{R}}$ for $j \in \mathcal{J}$ such that

$$h_Y(y) = \sup_{j \in \mathcal{J}} \Phi(x_j, y) - \alpha_j \quad \forall y \in Y. \tag{16}$$

If $\mathcal{J} = \emptyset$ we say that $h_Y \equiv -\infty$ is Φ-convex on Y by definition. We say that h_X or h_Y is Φ-concave if $-h_X$ or $-h_Y$ is Φ-convex.

If $X = Y$ are indistinguishable and Φ is not symmetric we shall refer to h_X as left Φ-convex and h_Y as right Φ-convex.

We shall also introduce the notion of a Φ-conjugate and a Φ-biconjugate function:

Definition 3.2 (Φ-conjugate functions)**.** *Let X and Y be nonempty sets and $\Phi : X \times Y \to \mathbb{R}$ a real-valued coupling. Let $h_X : X \to \overline{\mathbb{R}}$. Then we define*

$$h_X^{\Phi}(y) = \sup_{x \in X} \Phi(x, y) - h_X(x), \tag{17}$$

as the Φ-conjugate of f on Y and

$$h_X^{\Phi\Phi}(x) = \sup_{y \in Y} \Phi(x, y) - h_X^{\Phi}(y), \tag{18}$$

as the Φ-biconjugate back on X. The definitions of h_Y^{Φ} and $h_Y^{\Phi\Phi}$ for $h_Y : Y \to \overline{\mathbb{R}}$ are parallel.

If $X = Y$ are indistinguishable and Φ is not symmetric we shall refer to h_X^{Φ} as the left Φ-conjugate and h_Y^{Φ} as the right Φ-conjugate. From the definition it is clear that h_X^{Φ} is Φ-convex on Y and $h_X^{\Phi\Phi}$ is Φ-convex back on X.

The following statement is standard in literature, see, e.g., [1, 17] and references therein:

Proposition 3.3. *Let X and Y be nonempty sets and $\Phi : X \times Y \to \mathbb{R}$ a real-valued coupling and $h_X : X \to \overline{\mathbb{R}}$. Then we have*

(i) *h_X^{Φ} is Φ-convex on Y and $h_X^{\Phi\Phi}$ is Φ-convex on X;*

(ii) *$h_X(x) + h_X^{\Phi}(y) \geq \Phi(x, y) \quad \forall x \in X, \quad \forall y \in Y$;*

(iii) *$h_X(x) \geq h_X^{\Phi\Phi}(x) \quad \forall x \in X$.*

In addition, $h_X^{\Phi\Phi}$ is the pointwise largest Φ-convex function below h_X. In particular, this means that h_X is Φ-convex on X if and only if $h_X(x) = h_X^{\Phi\Phi}(x)$ for all $x \in X$.

The statements for $h_Y : Y \to \overline{\mathbb{R}}$ are parallel.

We shall also introduce the notion of a Φ-subgradient.

Definition 3.4 (Φ-subgradients)**.** *Let X and Y be nonempty sets and $\Phi : X \times Y \to \mathbb{R}$ a real-valued coupling. Let $h_X : X \to \overline{\mathbb{R}}$. Then we say that $y \in Y$ is a Φ-subgradient of h_X at $\bar{x} \in X$ if*

$$h_X(x) \geq h_X(\bar{x}) + \Phi(x, y) - \Phi(\bar{x}, y) \quad \forall x \in X, \tag{19}$$

or in other words $\bar{x} \in \arg\max_{x \in X} \Phi(x, y) - h_X(x)$. We call the set $\partial_\Phi h_X(\bar{x})$ of all such $y \in Y$ the Φ-subdifferential of f at $\bar{x}$. The definitions for $h_Y : Y \to \overline{\mathbb{R}}$ and $\partial_\Phi h_Y(\bar{y})$ for some $\bar{y} \in Y$ are parallel.

The following equivalent statements are standard in literature, see, e.g., [1, 17] and references therein:

Proposition 3.5. *Let X and Y be nonempty sets and $\Phi : X \times Y \to \mathbb{R}$ a real-valued coupling. Let $h_X : X \to \overline{\mathbb{R}}$. Then for any $\bar{x} \in X$ and $\bar{y} \in Y$ the following statements are equivalent:*

(i) $\bar{y} \in \partial_\Phi h_X(\bar{x})$;

(ii) $h_X(\bar{x}) + h_X^\Phi(\bar{y}) = \Phi(\bar{x}, \bar{y})$;

(iii) $\bar{x} \in \arg\min_{x \in X} h_X(x) - \Phi(x, \bar{y})$;

where any of the above equivalent statements implies that $h_X(\bar{x}) = h_X^{\Phi\Phi}(\bar{x})$ and $\bar{x} \in \partial_\Phi h_X^\Phi(\bar{y})$. If, in addition, h_X is Φ-convex, any of the above statements is equivalent to $\bar{x} \in \partial_\Phi h_X^\Phi(\bar{y})$.

The statements for $h_Y : Y \to \overline{\mathbb{R}}$ are parallel.

For the remainder of this section we choose $X = Y = \mathbb{R}^n$ and the coupling

$$\Phi(x, y) := -\lambda \star \phi(x - y), \tag{20}$$

for $\lambda := L^{-1}$.

The anisotropic descent inequality can be understood in terms of a Φ-subgradient inequality due to the following correspondence between Φ-subgradients and classical gradients. This adapts [24] to reference functions ϕ which are not necessarily super-coercive:

Proposition 3.6 (correspondence between Φ-subgradients and gradients)**.** *Let $h \in \mathcal{C}^1(\mathbb{R}^n)$ and let $L > 0$. Define $\Phi(x, y) := -\lambda \star \phi(x - y)$ for $\lambda := L^{-1}$. Assume that $\operatorname{rge} \nabla h \subseteq \operatorname{rge} \nabla \phi$. Then the following statements are equivalent:*

(i) *h satisfies the anisotropic descent inequality relative to ϕ with constant L;*

(ii) $\partial_\Phi(-h) = \mathrm{id} - \lambda \nabla \phi^* \circ \nabla h$.

In particular, any of the above equivalent statements implies that $\partial_\Phi(-h)$ is single-valued.

The proof essentially follows the proof of [24, Proposition 3.9]. For completeness an adapted proof is provided in Appendix A.2.

In particular the above result implies that the forward-step (4) in Algorithm 1 is the unique Φ-subgradient of $-f$ at x^k.

Next we prove that anisotropic smoothness is equivalent to Φ-concavity of h which means that h has an infimal convolution expression. In the convex case we also show a conjugate duality between anisotropic smoothness and relative strong convexity in the Bregman sense refining [43, Lemma 4.2]. In the smooth case relative strong convexity was considered in [26, 5]. To this end we adapt [24] to reference functions ϕ which are not necessarily super-coercive.

Proposition 3.7 (infimal convolution representation of anisotropic smoothness)**.** *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper lsc and $L > 0$. Define $\Phi(x, y) := -\lambda \star \phi(x - y)$ for $\lambda := L^{-1}$. Then the following are equivalent:*

(i) *h has the anisotropic descent property relative to ϕ with constant L;*

(ii) *$h \in \mathcal{C}^1(\mathbb{R}^n)$ with $\operatorname{rge} \nabla h \subseteq \operatorname{rge} \nabla \phi$ and $-h$ is Φ-convex, i.e., $h = -\xi^\Phi = \xi \boxplus L^{-1} \star \phi$ for some proper $\xi : \mathbb{R}^n \to \overline{\mathbb{R}}$.*

Any of the above statements implies that the infimal convolution in the second statement can be taken to be exact for some $\xi : \mathbb{R}^n \to \overline{\mathbb{R}}$ lsc and the infimum in the definition of Φ-convexity is attained. If, in addition, h is convex the assumption $h \in \mathcal{C}^1(\mathbb{R}^n)$ is superfluous in the second statement, ξ can be taken to be convex lsc and the above statements are equivalent to

$$h^* = \psi + L^{-1}\phi^*,$$

for some $\psi \in \Gamma_0(\mathbb{R}^n)$ with $\operatorname{dom} \psi \cap \operatorname{int} \operatorname{dom} \phi^* \neq \emptyset$, i.e., h^* is strongly convex relative to ϕ^* with constant L^{-1} in the Bregman sense.

The proof combines and adapts the proofs of [24, Proposition 3.10, Corollary 4.6 and Theorem 4.7]. For completeness an adapted proof is provided in Appendix A.3.

As shown above anisotropic smoothness means that $-h$ is Φ-convex and equivalently $h = -(-h)^{\Phi\Phi} = (-h)^\Phi \boxplus \lambda \star \phi$ is an infimal convolution. This leads us to study the corresponding conjugate transform

$$(-h)^\Phi = -(-h) \boxplus \lambda \star \phi_- = \sup_{u \in \mathbb{R}^n} h(\cdot + u) - \lambda \star \phi(u) =: h \boxminus \lambda \star \phi, \tag{21}$$

which is also called the *infimal deconvolution* of h and ϕ due to [20].

For $h \in \Gamma_0(\mathbb{R}^n)$ we have the following convexity and uniqueness property of the deconvolution of h and ϕ.

Proposition 3.8 (convexity of infimal deconvolution)**.** *Let $h \in \Gamma_0(\mathbb{R}^n)$ such that $h = \xi \boxplus \lambda \star \phi$ for some $\xi : \mathbb{R}^n \to \overline{\mathbb{R}}$. Then we have $h \boxminus \lambda \star \phi = (h^* \dot{-} \lambda \phi^*)^* \in \Gamma_0(\mathbb{R}^n)$. If, in addition, ϕ is super-coercive and $\xi \in \Gamma_0(\mathbb{R}^n)$ then $\xi = h \boxminus \lambda \star \phi$, i.e., $\xi = h \boxminus \lambda \star \phi$ is the unique function among $\xi \in \Gamma_0(\mathbb{R}^n)$ for which we have $h = \xi \boxplus \lambda \star \phi$.*

Proof. Without loss of generality assume that $\lambda = 1$ by redefining ϕ with $\lambda \star \phi$.

Let $h \in \Gamma_0(\mathbb{R}^n)$ with $h = \xi \boxplus \phi$ for some $\xi : \mathbb{R}^n \to \overline{\mathbb{R}}$. By definition this means that $-h$ is Φ-convex relative to $\Phi(x, y) = -\phi(x - y)$. Invoking Proposition 3.3 this means that $h = -(-h)^{\Phi\Phi} = (-h)^\Phi \boxplus \phi$. We have thanks to Hiriart-Urruty [19, Theorem 2.2]

$$(-h)^\Phi = h \boxminus \phi = (h^* \dot{-} \phi^*)^* > -\infty,$$

where the last inequality follows since $-h = (-h)^{\Phi\Phi}$ and Φ are finite-valued. Since $(-h)^{\Phi\Phi} > -\infty$ we also have that $(-h)^{\Phi} \not\equiv +\infty$ and thus $(-h)^{\Phi}$ is proper. Since $(-h)^{\Phi} = (h^* \dot{-} \phi^*)^*$ we have that $(-h)^{\Phi} \in \Gamma_0(\mathbb{R}^n)$. Now assume that, in addition, ϕ is super-coercive and $\xi \in \Gamma_0(\mathbb{R}^n)$. Then $\operatorname{dom} \phi^* = \mathbb{R}^n$ and by taking convex conjugates on either side of the equations $h = \xi \boxplus \phi = (-h)^{\Phi} \boxplus \phi$ we obtain by definition of the infimal convolution that $h^* = \xi^* + \phi^* = ((-h)^{\Phi})^* + \phi^*$. This implies that for all $x \in \mathbb{R}^n$ we have $h^*(x) - \phi^*(x) = \xi^*(x) = ((-h)^{\Phi})^*(x)$, and thus, since $\xi, (-h)^{\Phi} \in \Gamma_0(\mathbb{R}^n)$ we have by taking convex conjugates again: $(h^* - \phi^*)^* = \xi = (-h)^{\Phi}$. □

The proof of the second part of the above result follows alternatively from [36, Theorem 3.34].

4 Calculus for anisotropic smoothness and examples

4.1 Basic calculus

As discussed in the previous section anisotropic smoothness is equivalent to the existence of an infimal convolution expression. In this section, we shall study operations that preserve this property.

To begin with we observe that like its Euclidean counterpart anisotropic smoothness is separable:

Proposition 4.1 (separability of anisotropic smoothness)**.** *Let $h_1, h_2 \in \mathcal{C}^1(\mathbb{R}^n)$ have the anisotropic descent property relative to ϕ_1, ϕ_2 with constants L_1, L_2. Then $h(x_1, x_2) := h_1(x_1) + h_2(x_2)$ has the anisotropic descent property relative to $\phi(x_1, x_2) := L_1^{-1} \star \phi_1(x_2) + L_2^{-1} \star \phi_2(x_2)$.*

Proof. The result follows by summing the individual descent inequalities. □

We have the following elementary property:

Proposition 4.2 (pointwise scaling and tilting)**.** *Let $h \in \mathcal{C}^1(\mathbb{R}^n)$ have the anisotropic descent property relative to ϕ with constant L and let $c \in \mathbb{R}^n$ and $\alpha > 0$. Then $\alpha h + \langle c, \cdot \rangle$ has the anisotropic descent property relative to $\alpha\phi + \langle c, \cdot \rangle$ with constant L.*

Proof. This follows by definition of the anisotropic descent property noting that $(\alpha\phi + \langle c, \cdot \rangle)^* = (\alpha \star \phi^*)(\cdot - c)$ and $\nabla(\alpha \star \phi^*)(\cdot - c) = \nabla\phi^*((\cdot)\alpha^{-1} - c)$. □

Based on the Φ-concavity interpretation of anisotropic smoothness we can prove the following elementary transitivity property of anisotropic smoothness:

Proposition 4.3 (transitivity of anisotropic smoothness)**.** *Let $h \in \mathcal{C}^1(\mathbb{R}^n)$ and $\phi_1, \phi_2 \in \Gamma_0(\mathbb{R}^n)$ be Legendre type with full domain. Assume that h is anisotropically smooth relative to ϕ_1 and let ϕ_1 be anisotropically smooth relative to ϕ_2. Then h has the anisotropic descent property relative to ϕ_2.*

Proof. Thanks to Proposition 3.7 anisotropic smoothness of h relative to ϕ_1 implies that $h = \xi \boxplus \phi_1$ for some $\xi : \mathbb{R}^n \to \overline{\mathbb{R}}$ proper lsc such that $\operatorname{rge} \nabla h \subseteq \operatorname{rge} \nabla\phi_1$. Anisotropic smoothness of ϕ_1 relative to ϕ_2 implies that $\phi_1 = \psi \boxplus \phi_2$ for some $\psi : \mathbb{R}^n \to \overline{\mathbb{R}}$ proper lsc such that $\operatorname{rge} \nabla\phi_1 \subseteq \operatorname{rge} \nabla\phi_2$. Thus we have $h = \xi \boxplus (\psi \boxplus \phi_2) = (\xi \boxplus \psi) \boxplus \phi_2$ and $\operatorname{rge} \nabla h \subseteq \operatorname{rge} \nabla\phi_1 \subseteq \operatorname{rge} \nabla\phi_2$. Again invoking Proposition 3.7 h is anisotropically smooth relative to ϕ_2. □

We show that anisotropic smoothness with constant L_1 is preserved for any $L_2 > L_1$.

Proposition 4.4 (epi-scaling preserves anisotropic smoothness)**.** *Let $h \in \mathcal{C}^1(\mathbb{R}^n)$ have the anisotropic descent property with constant $L_1 > 0$ relative to ϕ. Then h has the anisotropic descent property relative to ϕ for any $L_2 > L_1$, and in particular we have the following monotonicity property for any $\bar{x}, x \in \mathbb{R}^n$:*

$$\begin{aligned} h(x) &\leq h(\bar{x}) + \tfrac{1}{L_1} \star \phi(x - \bar{x} + {L_1}^{-1}\nabla\phi^*(\nabla h(\bar{x}))) - \tfrac{1}{L_1} \star \phi({L_1}^{-1}\nabla\phi^*(\nabla h(\bar{x}))) \\ &\leq h(\bar{x}) + \tfrac{1}{L_2} \star \phi(x - \bar{x} + {L_2}^{-1}\nabla\phi^*(\nabla h(\bar{x}))) - \tfrac{1}{L_2} \star \phi({L_2}^{-1}\nabla\phi^*(\nabla h(\bar{x}))). \end{aligned}$$

Proof. Let $L_2 > L_1 > 0$ and assume that h is anisotropically smooth relative to ϕ with constant L_1. Let $\bar{x}, x \in \mathbb{R}^n$. Define $\bar{y} := \bar{x} - L_1^{-1}\nabla\phi^*(\nabla h(\bar{x}))$ and $\xi_{\bar{x}} := L_1^{-1} \star \phi(\cdot - \bar{y})$. Then we have that $\xi_{\bar{x}}^* = L_1^{-1}\phi^* + \langle \cdot, \bar{y}\rangle = \psi_{\bar{x}} + L_2^{-1}\phi^*$ for $\psi_{\bar{x}} := (L_1^{-1} - L_2^{-1})\phi^* + \langle \cdot, \bar{y}\rangle \in \Gamma_0(\mathbb{R}^n)$. Invoking Proposition 3.7 we have that $\xi_{\bar{x}}$ is anisotropically smooth relative to $L_2^{-1} \star \phi$. By definition of $\bar{y}$ we have that $\nabla\xi_{\bar{x}}(\bar{x}) = \nabla\phi(L_1(\bar{x} - \bar{y})) = \nabla h(\bar{x})$. Thus the descent inequality for $\xi_{\bar{x}}$ reads

$$\begin{aligned} \xi_{\bar{x}}(x) &\leq \xi_{\bar{x}}(\bar{x}) + L_2^{-1} \star \phi(x - \bar{x} + L_2^{-1}\nabla\phi^*(\nabla\xi_{\bar{x}}(\bar{x}))) \\ &\quad - L_2^{-1} \star \phi(L_2^{-1}\nabla\phi^*(\nabla\xi_{\bar{x}}(\bar{x}))) \\ &= \xi_{\bar{x}}(\bar{x}) + L_2^{-1} \star \phi(x - \bar{x} + L_2^{-1}\nabla\phi^*(\nabla h(\bar{x}))) \\ &\quad - L_2^{-1} \star \phi(L_2^{-1}\nabla\phi^*(\nabla h(\bar{x}))). \end{aligned} \tag{22}$$

Then we have

$$\begin{aligned} h(x) &\leq h(\bar{x}) + \tfrac{1}{L_1} \star \phi(x - \bar{y}) - \tfrac{1}{L_1} \star \phi(\bar{x} - \bar{y}) = h(\bar{x}) + \xi_{\bar{x}}(x) - \xi_{\bar{x}}(\bar{x}) \\ &\leq h(\bar{x}) + \tfrac{1}{L_2} \star \phi(x - \bar{x} + {L_2}^{-1}\nabla\phi^*(\nabla h(\bar{x}))) - \tfrac{1}{L_2} \star \phi({L_2}^{-1}\nabla\phi^*(\nabla h(\bar{x}))), \end{aligned}$$

where the first inequality follows from the definition of $\bar{y}$ and anisotropic smoothness of f relative to ϕ with constant L_1 and the last inequality follows by (22). In particular this shows that f has the anisotropic descent property relative to ϕ for constant L_2. □

4.2 Pointwise addition under joint convexity of dual Bregman distance

Naturally, anisotropic smoothness is closed under epi-addition; see [24, Corollary 4.9]. In contrast, the pointwise sum of anisotropically smooth functions is in general not anisotropically smooth, see [24, Example 4.10], unless the dual Bregman divergence is jointly convex [43, Theorem 5.1(ii)]. In this section we will generalize this to nonconvex infimal convolutions and consider a refinement for the exponential reference function showing closedness under pointwise conic combinations. This is false in general for Euclidean Lipschitz smoothness with the same Lipschitz constant.

The following result shows closedness under pointwise average if the dual Bregman distance is jointly convex:

Lemma 4.5 (closedness of anisotropic smoothness under pointwise average)**.** *Let $h_1, h_2 \in \mathcal{C}^1(\mathbb{R}^n)$ be anisotropically smooth relative to ϕ with constants L_1, L_2 respectively and assume that D_{ϕ^*} is jointly convex. Let $\alpha \in [0, 1]$. Then $\alpha h_1 + (1 - \alpha)h_2$ is anisotropically smooth relative to ϕ with constant $L := \max\{L_1, L_2\}$.*

Proof. Let $L := \max\{L_1, L_2\}$ and define $\Phi(x,y) := -L^{-1}\star\phi(x-y)$. Thanks to Proposition 4.4 h_1, h_2 are anisotropically smooth relative to ϕ with constant L. Invoking Proposition 3.7 we have $h_1(x) = \inf_{y\in\mathbb{R}^n} L^{-1}\star\phi(x-y)+(-h_1)^\Phi(y)$ and $h_2(x) = \inf_{u\in\mathbb{R}^n} L^{-1}\star\phi(x-u)+(-h_2)^\Phi(u)$.

Without loss of generality assume that $L = 1$. Let $\alpha \in [0,1]$ and $x \in \mathbb{R}^n$. Then we have

$$\begin{aligned}(\alpha h_1 + (1-\alpha)h_2)(x) &= \inf_{y\in\mathbb{R}^n} \alpha\phi(x-y) + \alpha(-h_1)^\Phi(y)) \\ &\quad + \inf_{u\in\mathbb{R}^n} (1-\alpha)\phi(x-u) + (1-\alpha)(-h_2)^\Phi(u) \\ &= \inf_{y\in\mathbb{R}^n}\inf_{u\in\mathbb{R}^n} \alpha\phi(x-y) + (1-\alpha)\phi(x-u) \\ &\quad + \alpha(-h_1)^\Phi(y) + (1-\alpha)(-h_2)^\Phi(u). \end{aligned} \tag{23}$$

Let $y, u \in \mathbb{R}^n$. Assuming that $\alpha\phi(\cdot-y)+(1-\alpha)\phi(\cdot-u) =: \xi(\cdot,y,u)$ is anisotropically smooth relative to ϕ (which is verified below) we have by Proposition 3.7 that

$$\alpha\phi(x-y) + (1-\alpha)\phi(x-u) = \inf_{v\in\mathbb{R}^n} \phi(x-v) + (-\xi(\cdot,y,u))^\Phi(v). \tag{24}$$

Define $\zeta(v) := \inf_{y\in\mathbb{R}^n}\inf_{u\in\mathbb{R}^n}(-\xi(\cdot;y,u))^\Phi(v)+\alpha(-h_1)^\Phi(y)+(1-\alpha)(-h_2)^\Phi(u)$. Substituting (24) and the expression of ζ in (23) an interchange of the order of minimization yields:

$$\begin{aligned}(\alpha h_1 + (1-\alpha)h_2)(x) &= \inf_{y\in\mathbb{R}^n}\inf_{u\in\mathbb{R}^n} \alpha\phi(x-y) + (1-\alpha)\phi(x-u) \\ &\quad + \alpha(-h_1)^\Phi(y) + (1-\alpha)(-h_2)^\Phi(u) \\ &= \inf_{y\in\mathbb{R}^n}\inf_{u\in\mathbb{R}^n}\inf_{v\in\mathbb{R}^n} \phi(x-v) + (-\xi(\cdot;y,u))^\Phi(v) \\ &\quad + \alpha(-h_1)^\Phi(y) + (1-\alpha)(-h_2)^\Phi(u) \\ &= \inf_{v\in\mathbb{R}^n} \phi(x-v) + \inf_{y\in\mathbb{R}^n}\inf_{u\in\mathbb{R}^n} (-\xi(\cdot;y,u))^\Phi(v) \\ &\quad + \alpha(-h_1)^\Phi(y) + (1-\alpha)(-h_2)^\Phi(u) \\ &= \inf_{v\in\mathbb{R}^n} \phi(x-v) + \zeta(v). \end{aligned}$$

By convexity of $\operatorname{rge}\nabla\phi = \operatorname{int}\operatorname{dom}\phi^*$ we have that $\operatorname{rge}\nabla(\alpha h_1 + (1-\alpha)h_2) \subseteq \alpha\operatorname{rge}\nabla h_1 + (1-\alpha)\operatorname{rge}\nabla h_2 \subseteq \operatorname{rge}\nabla\phi$. In view of Proposition 3.7 $\alpha h_1 + (1-\alpha)h_2$ is anisotropically smooth relative to ϕ.

It remains to show that $\alpha\phi(\cdot-y)+(1-\alpha)\phi(\cdot-u)$ is anisotropically smooth. We have that $\phi(\cdot-y) = (\langle\cdot,y\rangle+\phi^*)^*$ and $\phi(\cdot-u) = (\langle\cdot,u\rangle+\phi^*)^*$. Thus we can invoke [43, Lemma 4.2] and obtain that $\nabla\phi(\cdot-y)$ and $\nabla\phi(\cdot-u)$ are both $\nabla\phi^*$-firmly nonexpansive on $\mathbb{R}^n$ in the sense of [43, Definition 4.1]. Since D_{ϕ^*} is jointly convex, in view of [43, Remark 4.6] we can invoke [43, Lemma 4.4] and obtain that $\nabla(\alpha\phi(\cdot-y)+(1-\alpha)\phi(\cdot-u)) = \alpha\nabla\phi(\cdot-y)+(1-\alpha)\nabla\phi(\cdot-u)$ is $\nabla\phi^*$-firmly nonexpansive on $\mathbb{R}^n$ as well. By convexity of $\operatorname{dom}\phi^*$ we have that

$$\begin{aligned}\operatorname{dom}(\alpha\phi(\cdot-y)+(1-\alpha)\phi(\cdot-u))^* &= \operatorname{dom}\alpha\star(\langle\cdot,y\rangle+\phi^*) \\ &\quad + \operatorname{dom}(1-\alpha)\star(\langle\cdot,u\rangle+\phi^*) \\ &= \alpha\operatorname{dom}\phi^* + (1-\alpha)\operatorname{dom}\phi^* = \operatorname{dom}\phi^*.\end{aligned}$$

Thus we can again invoke [43, Lemma 4.2] to obtain that $\alpha\phi(\cdot-y)+(1-\alpha)\phi(\cdot-u) = \psi\boxplus\phi$ for some $\psi \in \Gamma_0(\mathbb{R}^n)$ and $\operatorname{rge}\nabla(\alpha\phi(\cdot-y)+(1-\alpha)\phi(\cdot-u)) \subseteq \operatorname{rge}\nabla\phi$. Again thanks to Proposition 3.7 $\alpha\phi(\cdot-y)+(1-\alpha)\phi(\cdot-u)$ is anisotropically smooth relative to ϕ as claimed. □

Remark 4.6. The above result generalizes directly to any finite weighted sum $h := \sum_{i=1}^m \alpha_i h_i$ for weights $\alpha_i \geq 0$ with $\sum_{i=1}^m \alpha_i = 1$: This follows by applying the previous result recursively noting that $\sum_{i=1}^m \alpha_i h_i = (1-\alpha_m)\sum_{i=1}^{m-1} \alpha_i/(1-\alpha_m) h_i + \alpha_m h_m$.

In the convex case and in particular the subsequent examples we often verify anisotropic smoothness via its dual characterization in terms of relative strong convexity in the Bregman sense; see Proposition 3.7. For that reason we shall prove the following lemma adopting the proof strategy of [43, Lemma 4.2]:

Lemma 4.7. *Let $h \in \Gamma_0(\mathbb{R}^n)$ assume that $\operatorname{dom} h \subseteq \operatorname{dom} \phi^*$ with $\operatorname{ri}\operatorname{dom} h \cap \operatorname{int}\operatorname{dom}\phi^* \neq \emptyset$. Then the following are equivalent:*

(i) *$h = \psi + \phi^*$ for some $\psi \in \Gamma_0(\mathbb{R}^n)$;*

(ii) *$h \dot{-} \phi^*$ is convex on $\operatorname{ri}\operatorname{dom} h$,*

where the first item implies that $\operatorname{dom}\psi \cap \operatorname{int}\operatorname{dom}\phi^* \neq \emptyset$ and ψ can be taken as $\operatorname{cl}\xi$ with $\xi(x) := h(x) - \phi^*(x) \in \mathbb{R}$ if $x \in \operatorname{ri}\operatorname{dom} h$ and $+\infty$ otherwise.

Proof. "4.7(i) $\Rightarrow$ 4.7(ii)": Let $\psi \in \Gamma_0(\mathbb{R}^n)$. Then $\psi \equiv h \dot{-} \phi^*$ is finite-valued and convex on $\operatorname{ri}\operatorname{dom} h \subseteq \operatorname{dom} h \subseteq \operatorname{dom}\phi^*$.

"4.7(ii) $\Rightarrow$ 4.7(i)": Define $\xi(x) := h(x) - \phi^*(x) \in \mathbb{R}$ if $x \in \operatorname{ri}\operatorname{dom} h$ and $+\infty$ otherwise. Then $\operatorname{dom}\xi = \operatorname{ri}\operatorname{dom} h$. Since ξ is proper, convex, by [35, Theorem 7.4] $\operatorname{cl}\xi \in \Gamma_0(\mathbb{R}^n)$ with $\xi \equiv \operatorname{cl}\xi$ on $\operatorname{ri}\operatorname{dom} h$. Next we prove that $h \equiv \operatorname{cl}\xi + \phi^*$ on $\operatorname{cl}\operatorname{dom} h$:

For every $x \in \operatorname{ri}\operatorname{dom} h \subseteq \operatorname{dom}\phi^*$ we have $(\operatorname{cl}\xi)(x) + \phi^*(x) = \xi(x) + \phi^*(x) = h(x)$. Now take $x_0 \in \operatorname{ri}\operatorname{dom} h \cap \operatorname{int}\operatorname{dom}\phi^*$ which exists by assumption and any $x \in \operatorname{cl}\operatorname{dom} h$. Then we have by the line segment principle [35, Theorem 6.1] that $(1-\tau)x_0 + \tau x \in \operatorname{ri}\operatorname{dom} h$ for every $\tau \in [0,1)$. In view of [35, Corollary 7.4.1] $\operatorname{ri}(\operatorname{dom}\operatorname{cl}\xi) = \operatorname{ri}\operatorname{dom}\xi = \operatorname{ri}\operatorname{dom} h$. Therefore $x_0 \in \operatorname{ri}(\operatorname{dom}\operatorname{cl}\xi) \cap \operatorname{int}\operatorname{dom}\phi^* = \operatorname{ri}(\operatorname{dom}\operatorname{cl}\xi \cap \operatorname{dom}\phi^*) = \operatorname{ri}(\operatorname{dom}(\operatorname{cl}\xi + \phi^*))$ by [35, Theorem 6.5] and $\operatorname{cl}\xi + \phi^* \in \Gamma_0(\mathbb{R}^n)$ by [35, Theorem 9.3]. Then we can invoke [35, Theorem 7.5] and obtain

$$\begin{aligned}(\operatorname{cl}\xi)(x) + \phi^*(x) &= \lim_{\tau \nearrow 1} (\operatorname{cl}\xi)((1-\tau)x_0 + \tau x) + \phi^*((1-\tau)x_0 + \tau x) \\ &= \lim_{\tau \nearrow 1} h((1-\tau)x_0 + \tau x) = h(x),\end{aligned}$$

where the second equality holds since $(1-\tau)x_0 + \tau x \in \operatorname{ri}\operatorname{dom} h$ and $\operatorname{cl}\xi + \phi^* \equiv h$ on $\operatorname{ri}\operatorname{dom} h$, and the last equality again by [35, Theorem 7.5] since $h \in \Gamma_0(\mathbb{R}^n)$ and $x_0 \in \operatorname{ri}\operatorname{dom} h$. This implies that $\operatorname{cl}\xi + \phi^* \equiv h$ on $\operatorname{cl}\operatorname{dom} h$. By [35, Theorem 7.4] $\operatorname{cl}\xi$ differs from ξ at most at relative boundary points of $\operatorname{dom}\xi = \operatorname{ri}\operatorname{dom} h$, i.e., at points $x \in \operatorname{cl}\operatorname{dom}\xi \setminus \operatorname{ri}\operatorname{dom}\xi = \operatorname{cl}\operatorname{dom} h \setminus \operatorname{ri}\operatorname{dom} h$. This implies that $\operatorname{cl}\xi \equiv \xi \equiv +\infty$ on $\mathbb{R}^n \setminus \operatorname{cl}\operatorname{dom} h$ and thus $\operatorname{cl}\xi + \phi^* \equiv h \equiv +\infty$ on $\mathbb{R}^n \setminus \operatorname{cl}\operatorname{dom} h$. Altogether this implies that $\psi + \phi^* = h$ for $\psi = \operatorname{cl}\xi$. This completes the reverse direction of the proof. Finally, since $h = \psi + \phi^*$ we have by assumption that $\emptyset \neq \operatorname{dom} h \cap \operatorname{int}\operatorname{dom}\phi^* = (\operatorname{dom}\psi \cap \operatorname{dom}\phi^*) \cap \operatorname{int}\operatorname{dom}\phi^* = \operatorname{dom}\psi \cap \operatorname{int}\operatorname{dom}\phi^*$. □

Using the calculus rules devised above we can show in the next example that the loss function in logistic regression is anisotropically smooth relative to a softmax approximation of the absolute value function:

Example 4.8. Let $f(x) = \frac{1}{m}\sum_{i=1}^m \ell(\langle a_i, x\rangle)$ with $a_i \in [-1,1]^n$ and $\ell(t) = \log(1+\exp(t))$. Then f is anisotropically smooth relative to the symmetrized logistic loss $\phi(x) = \sum_{j=1}^n h(x_j)$ for $h(t) = 2\log(1+\exp(t)) - t$ with constant $L = \max_{1\le i\le m} \|a_i\|^2$. Note that for $L \to \infty$ $\frac{1}{L} \star h$ converges pointwisely to the absolute value function.

For completeness a proof is provided in Appendix A.4.

Next we show an interesting refinement of the previous result for the exponential reference function $\phi = \operatorname{Exp}$ which reveals that the anisotropic descent property is closed under pointwise conic combinations. Note that this is false in general for Euclidean Lipschitz smoothness with the same Lipschitz constant.

Lemma 4.9 (closedness under pointwise conic combinations for exponential smoothness)**.** *Let $h_1, h_2 \in \mathcal{C}^1(\mathbb{R}^n)$ be anisotropically smooth relative to $\phi = \operatorname{Exp}$ with constants L_1, L_2. Let $\alpha, \beta \ge 0$ such that either $\alpha > 0$ or $\beta > 0$. Then $\alpha h_1 + \beta h_2$ is anisotropically smooth relative to ϕ with constant $L := \max\{L_1, L_2\}$.*

Proof. First note that $h \equiv 0$ is not anisotropically smooth since $\operatorname{dom} h^* = \{0\} \not\subseteq \operatorname{int}(\operatorname{dom}\phi^*)$. Let $L := \max\{L_1, L_2\}$ and define $\Phi(x,y) := -L^{-1} \star \operatorname{Exp}(x-y)$. Thanks to Proposition 4.4 h_1, h_2 are anisotropically smooth relative to Exp with constant L. Thanks to Proposition 3.7 we have $h_1(x) = \inf_{y\in\mathbb{R}^n} L^{-1} \star \phi(x-y) + (-h_1)^\Phi(y)$ and $h_2(x) = \inf_{u\in\mathbb{R}^n} L^{-1} \star \phi(x-u) + (-h_2)^\Phi(u)$. Without loss of generality assume that $L = 1$.

Let $x \in \mathbb{R}^n$. Then we have

$$\begin{aligned}(\alpha h_1 + \beta h_2)(x) &= \inf_{y\in\mathbb{R}^n} \alpha\phi(x-y) + \alpha(-h_1)^\Phi(y)) \\ &\quad + \inf_{u\in\mathbb{R}^n} \beta\phi(x-u) + \beta(-h_2)^\Phi(u) \\ &= \inf_{y\in\mathbb{R}^n}\inf_{u\in\mathbb{R}^n} \alpha\phi(x-y) + \beta\phi(x-u) \\ &\quad + \alpha(-h_1)^\Phi(y) + \beta(-h_2)^\Phi(u). \end{aligned} \tag{25}$$

Let $y, u \in \mathbb{R}^n$. It holds that

$$\begin{aligned}\alpha\phi(x-y) + \beta\phi(x-u) &= \sum_{i=1}^n \alpha\exp(x_i - y_i) + \beta\exp(x_i - u_i) \\ &= \sum_{i=1}^n \exp(x_i)(\alpha\exp(-y_i) + \beta\exp(-u_i)) \\ &= \sum_{i=1}^n \exp(x_i + \log(\alpha\exp(-y_i) + \beta\exp(-u_i))) \\ &= \inf_{v\in\mathbb{R}^n} \sum_{i=1}^n \exp(x_i - v_i) + \xi(v; y, u) \\ &= \inf_{v\in\mathbb{R}^n} \phi(x-v) + \xi(v; y, u),\end{aligned}$$

where the third equality holds since either $\alpha > 0$ or $\beta > 0$ and thus $\alpha\exp(-y_i) + \beta\exp(-u_i)) > 0$ and the second last equality involves the expression

$$\xi(v; y, u) = \begin{cases} 0 & \text{if } v_i := -\log(\alpha\exp(-y_i) + \beta\exp(-u_i)) \text{ for all } 1 \le i \le n, \\ +\infty & \text{otherwise.}\end{cases} \tag{26}$$

Define $\zeta(v) := \inf_{y\in\mathbb{R}^n} \inf_{u\in\mathbb{R}^n} \xi(v;y,u) + \alpha(-h_1)^\Phi(y) + \beta(-h_2)^\Phi(u) = \inf\{\alpha(-h_1)^\Phi(y) + \beta(-h_2)^\Phi(u) \mid y,u\in\mathbb{R}^n : -\log(\alpha\exp(-y_i)+\beta\exp(-u_i)) = v_i\}$. Substituting the expression for ζ in (25) an interchange of the order of minimization leads to:

$$\begin{aligned}(\alpha h_1+\beta h_2)(x) &= \inf_{y\in\mathbb{R}^n}\inf_{u\in\mathbb{R}^n} \alpha\phi(x-y)+\beta\phi(x-u)+\alpha(-h_1)^\Phi(y)\\ &\quad +\beta(-h_2)^\Phi(u)\\ &= \inf_{y\in\mathbb{R}^n}\inf_{u\in\mathbb{R}^n}\inf_{v\in\mathbb{R}^n} \phi(x-v)+\xi(v;y,u)+\alpha(-h_1)^\Phi(y)\\ &\quad +\beta(-h_2)^\Phi(u)\\ &= \inf_{v\in\mathbb{R}^n}\phi(x-v)+\inf_{y\in\mathbb{R}^n}\inf_{u\in\mathbb{R}^n}\xi(v;y,u)+\alpha(-h_1)^\Phi(y)\\ &\quad +\beta(-h_2)^\Phi(u)\\ &= \inf_{v\in\mathbb{R}^n}\phi(x-v)+\zeta(v).\end{aligned}$$

By assumption we have that $\operatorname{rge}\nabla h_1 \subseteq \mathbb{R}^n_{++}$ and $\operatorname{rge}\nabla h_2 \subseteq \mathbb{R}^n_{++}$ and since either $\alpha>0$ or $\beta>0$ we have $\operatorname{rge}(\alpha\nabla h_1+\beta\nabla h_2)\subseteq\mathbb{R}^n_{++} = \operatorname{rge}\nabla\phi$. Again thanks to Proposition 3.7 $\alpha h_1+\beta h_2$ is anisotropically smooth relative to ϕ as claimed. □

Example 4.10. Let $f(x)=\sum_{i=1}^m \sigma\star\exp(\langle a_i,x\rangle - b_i)$ with $a_i\in\mathbb{R}^n_{++}$ and $\sigma>0$. Then f is anisotropically smooth relative to $\phi(x)=\operatorname{Exp}(x)$ with constant $L=\max_{1\le i\le m}\|a_i\|_1/\sigma$.

For completeness a full proof is provided in Appendix A.5.

Remark 4.11. The assumption $a_i\in\mathbb{R}^n_{++}$ can be relaxed to $a_i\in\mathbb{R}^n_+$ as long as $\sum_{i=1}^m a_i\in\mathbb{R}^n_{++}$. Identifying the vectors a_i as the rows of a matrix $A\in\mathbb{R}^{m\times n}_+$ this means equivalently that each column of A has at least one nonzero element. This is proved using a continuity argument: Choose $x,\bar{x}\in\mathbb{R}^n$. Let $\varepsilon>0$ and denote by $\mathbb{1}\in\mathbb{R}^n$ the vector of all ones. Then it holds that $a_i+\varepsilon\mathbb{1}\in\mathbb{R}_{++}$ and for $f_\varepsilon := \sum_{i=1}^m \sigma\star\exp(\langle a_i+\varepsilon\mathbb{1},\cdot\rangle - b_i)$ and $L_\varepsilon := \max_{1\le i\le m}\|a_i+\varepsilon\mathbb{1}\|_1/\sigma = \max_{1\le i\le m}\|a_i\|_1/\sigma + \varepsilon n/\sigma$ we have thanks to the derivation above that

$$f_\varepsilon(x)\le f_\varepsilon(\bar{x}) + \tfrac{1}{L_\varepsilon}\star\phi(x-\bar{x}+L_\varepsilon^{-1}\nabla\phi^*(\nabla f_\varepsilon(\bar{x}))) - \tfrac{1}{L_\varepsilon}\star\phi(L_\varepsilon^{-1}\nabla\phi^*(\nabla f_\varepsilon(\bar{x}))).$$

By continuity we have for $\varepsilon\to 0$ that $L_\varepsilon\to L=\max_{1\le i\le m}\|a_i\|_1/\sigma$ and $\mathbb{R}^n_{++}\ni\nabla f_\varepsilon(\bar{x}) = \sum_{i=1}^m (a_i+\varepsilon\mathbb{1})(\sigma\star\exp)(\langle a_i+\varepsilon\mathbb{1},\bar{x}\rangle - b_i)\to\nabla f(\bar{x}) = \sum_{i=1}^m a_i(\sigma\star\exp)(\langle a_i,\bar{x}\rangle - b_i)\in\mathbb{R}^n_{++}$ since by assumption $\sum_{i=1}^m a_i\in\mathbb{R}_{++}$. Passing to $\varepsilon\to 0$ in the previous inequality since $\nabla\phi^*$ is continuous relative to $\mathbb{R}^n_{++}$ we deduce that the descent inequality holds at the chosen points $x,\bar{x}$. Since $x,\bar{x}\in\mathbb{R}^n$ are arbitrary we obtain the claimed result.

5 Analysis of the anisotropic proximal gradient method

5.1 Subsequential convergence and linesearch

Recall validity of assumptions (A1) to (A5) and the definition of the threshold of anisotropic prox-boundedness λ_g and $\lambda<\lambda_g$. The main goal of this section is to analyse the subsequential convergence of the anisotropic proximal gradient method. For that purpose we shall introduce the following regularized gap function as a measure of stationarity:

$$\mathcal{G}_F^\phi(\bar{x},\lambda) := \frac{1}{\lambda}\big(F(\bar{x}) - F_\lambda(\bar{x})\big), \tag{27}$$

for any $\lambda < \lambda_g$ and

$$F_\lambda(\bar{x}) := \inf_{x\in\mathbb{R}^n} f(\bar{x}) + \lambda \star \phi(x - \bar{x} + \lambda\nabla\phi^*(\nabla f(\bar{x}))) \tag{28}$$
$$- \lambda \star \phi(\lambda\nabla\phi^*(\nabla f(\bar{x}))) + g(x)$$
$$= f(\bar{x}) + (g \boxplus \lambda \star \phi_-)(\bar{x} - \lambda\nabla\phi^*(\nabla f(\bar{x}))) - \lambda \star \phi(\lambda\nabla\phi^*(\nabla f(\bar{x}))), \tag{29}$$

is an anisotropic generalization of the forward-backward envelope [32, 38]. The regularized gap function is a generalization of [21, Equation 13] to the anisotropic case and is closely related to regularized gap functions for solving variational inequalities [22]. The particular scaling $\frac{1}{\lambda}$ will be important subsequently for proving a linear convergence result under anisotropic proximal gradient dominance.

Next we show the following key properties of $\mathcal{G}_F^\phi(\bar{x}, \lambda)$ to justify its choice as a measure of stationarity:

Lemma 5.1. *For any $\lambda < \lambda_g$ the gap function $\mathcal{G}_F^\phi(\cdot, \lambda) \geq 0$ is lsc and for any $x^\star \in \mathbb{R}^n$ with $\mathcal{G}_F^\phi(x^\star, \lambda) = 0$ we have that $x^\star$ is a stationary point of F, i.e., $0 \in \nabla f(x^\star) + \partial g(x^\star)$. If, in addition, $g \equiv 0$, the gap simplifies to $\mathcal{G}_F^\phi(\cdot, \lambda) = D_{\phi^*}(0, \nabla f(\cdot))$.*

Proof. Let $\lambda < \lambda_g$. Let $\bar{x} \in \mathbb{R}^n$ and define

$$\xi(x, \bar{x}) := f(\bar{x}) + \lambda \star \phi(x - \bar{x} + \lambda\nabla\phi^*(\nabla f(\bar{x}))) - \lambda \star \phi(\lambda\nabla\phi^*(\nabla f(\bar{x}))) + g(x).$$

Then $F_\lambda(\bar{x}) = \inf_{x\in\mathbb{R}^n} \xi(x, \bar{x}) \leq \xi(\bar{x}, \bar{x}) = F(\bar{x})$ implying that $\mathcal{G}_F^\phi(\bar{x}, \lambda) \geq 0$. Let $x^\star \in \mathbb{R}^n$ and assume that $\mathcal{G}_F^\phi(x^\star, \lambda) = 0$, i.e., $F(x^\star) = F_\lambda(x^\star)$. Therefore $F_\lambda(x^\star) = \inf_{x\in\mathbb{R}^n} \xi(x, x^\star) \leq \xi(x^\star, x^\star) = F(x^\star) = F_\lambda(x^\star)$ and thus the infimum in the definition of $F_\lambda(x^\star)$ is attained at $x^\star$. By Fermat's rule [36, Theorem 10.1] we have

$$0 \in \partial\xi(\cdot, x^\star)(x^\star) = \nabla\phi(\nabla\phi^*(\nabla f(x^\star))) + \partial g(x^\star) = \nabla f(x^\star) + \partial g(x^\star).$$

Since by assumption $h(x, y) = g(x) + \lambda \star \phi(x - y)$ is level-bounded in x locally uniformly in y, thanks to [36, Theorem 1.17], $g \boxplus \lambda \star \phi_-$ is continuous and thus F_λ as a composition of continuous functions is continuous as well. Since $F = f + g$ is lsc we have that $\mathcal{G}_F^\phi(\bar{x}, \lambda) := \frac{1}{\lambda}(F(\bar{x}) - F_\lambda(\bar{x}))$ is lsc.

If, in particular, $g \equiv 0$ we have for any $\bar{x} \in \mathbb{R}^n$

$$\begin{aligned}
\mathcal{G}_F^\phi(\bar{x}, \lambda) &= -\frac{1}{\lambda} \min_{x\in\mathbb{R}^n} \lambda \star \phi(x - \bar{x} + \lambda\nabla\phi^*(\nabla f(\bar{x}))) - \lambda \star \phi(\lambda\nabla\phi^*(\nabla f(\bar{x}))) \\
&= \phi(\nabla\phi^*(\nabla f(\bar{x}))) - \min_{x\in\mathbb{R}^n} \phi(\lambda^{-1}(x - \bar{x}) + \nabla\phi^*(\nabla f(\bar{x}))) \\
&= \phi(\nabla\phi^*(\nabla f(\bar{x}))) - \phi(\nabla\phi^*(0)) \\
&\quad - \langle \nabla\phi(\nabla\phi^*(0)), \nabla\phi^*(\nabla f(\bar{x})) - \nabla\phi^*(0)\rangle \\
&= D_\phi(\nabla\phi^*(\nabla f(\bar{x})), \nabla\phi^*(0)) = D_{\phi^*}(0, \nabla f(\bar{x})),
\end{aligned}$$

where the identities hold due to the constraint qualification $0 \in \operatorname{int}\operatorname{dom}\phi^*$ and the fact that $\hat{x}$ minimizes $x \mapsto \phi(\lambda^{-1}(x-\bar{x})+\nabla\phi^*(\nabla f(\bar{x})))$ if and only if $\nabla\phi^*(0) = \lambda^{-1}(\hat{x}-\bar{x})+\nabla\phi^*(\nabla f(\bar{x}))$, the identity $\nabla\phi(\nabla\phi^*(0)) = 0$ and the identity in Lemma 1.3. □

We first prove the following sufficient decrease property:

Lemma 5.2. *Let $\{x^k\}_{k\in\mathbb{N}_0}$ be the sequence of backward-steps generated by Algorithm 1. Then the following sufficient decrease property holds true for all $k \in \mathbb{N}_0$:*

$$F(x^{k+1}) \leq F(x^k) - \lambda \mathcal{G}_F^\phi(x^k, \lambda). \tag{30}$$

Proof. We have:

$$\begin{aligned}
F(x^{k+1}) &= f(x^{k+1}) + g(x^{k+1}) \\
&\leq f(x^k) + \tfrac{1}{L} \star \phi(x^{k+1} - x^k + L^{-1}\nabla\phi^*(\nabla f(x^k))) \\
&\quad - \tfrac{1}{L} \star \phi(L^{-1}\nabla\phi^*(\nabla f(x^k))) + g(x^{k+1}) \\
&\leq f(x^k) + \lambda \star \phi(x^{k+1} - x^k + \lambda\nabla\phi^*(\nabla f(x^k))) \\
&\quad - \lambda \star \phi(\lambda\nabla\phi^*(\nabla f(x^k))) + g(x^{k+1}) \\
&= F_\lambda(x^k) = F(x^k) - (F(x^k) - F_\lambda(x^k)) = F(x^k) - \lambda \mathcal{G}_F^\phi(x^k, \lambda),
\end{aligned}$$

where the first inequality follows by the anisotropic descent inequality, the second inequality by Proposition 4.4 and the choice $\lambda^{-1} \geq L$ and the second equality by the update of x^{k+1} and the definition of $F_\lambda(x^k)$. □

Theorem 5.3. *Let $\{x^k\}_{k\in\mathbb{N}_0}$ be the sequence of backward-steps generated by Algorithm 1. The sequence $\{F(x^k)\}_{k\in\mathbb{N}_0}$ is nonincreasing and convergent. In addition, every limit point $x^\star$ of the sequence of iterates $\{x^k\}_{k\in\mathbb{N}_0}$ is a stationary point of F, i.e., $0 \in \nabla f(x^\star) + \partial g(x^\star)$. In particular the minimum over the past regularized gaps vanishes sublinearly at rate $\mathcal{O}(1/K)$:*

$$\min_{k\in\{0,1,\dots,K-1\}} \mathcal{G}_F^\phi(x^k, \lambda) \leq \frac{1}{\lambda K}\big(F(x^0) - \inf F\big).$$

Proof. Thanks to Lemma 5.2 we have

$$F(x^{k+1}) \leq F(x^k) - \lambda \mathcal{G}_F^\phi(x^k, \lambda).$$

This implies that $F(x^k)$ is monotonically decreasing. Since $-\infty < \inf F \leq F(x^k)$ is bounded from below this means that $F(x^k) \to F^*$ converges to some F^*. Summing the inequality we obtain since $-\infty < \inf F \leq F(x^K)$ is bounded from below:

$$\begin{aligned}
-\infty < \inf F - F(x^0) &\leq F(x^K) - F(x^0) \\
&= \sum_{k=0}^{K-1} F(x^{k+1}) - F(x^k) \leq -\lambda \sum_{k=0}^{K-1} \mathcal{G}_F^\phi(x^k, \lambda).
\end{aligned} \tag{31}$$

This implies that $\{\sum_{k=0}^{K-1} \mathcal{G}_F^\phi(x^k, \lambda)\}_{K\in\mathbb{N}}$ is convergent and thus we have

$$\lim_{k\to\infty} \mathcal{G}_F^\phi(x^k, \lambda) = 0. \tag{32}$$

Let $x^\star$ be a limit point of $\{x^k\}_{k\in\mathbb{N}_0}$. Let $x^{k_j} \to x^\star$ be a corresponding subsequence. In view of Lemma 5.1 $\mathcal{G}_F^\phi(\cdot, \lambda)$ is lsc and thus we have

$$0 \leq \mathcal{G}_F^\phi(x^\star, \lambda) \leq \lim_{j\to\infty} \mathcal{G}_F^\phi(x^{k_j}, \lambda) = 0,$$

where the last equality follows from (32). Then invoking Lemma 5.1 again this implies that $x^\star$ is a stationary point of $f+g$, i.e., $0 \in \nabla f(x^\star) + \partial g(x^\star)$.

Thanks to (31) we have

$$K \cdot \min_{k\in\{0,1,\dots,K-1\}} \mathcal{G}_F^\phi(x^k,\lambda) \le \sum_{k=0}^{K-1} \mathcal{G}_F^\phi(x^k,\lambda) \le \frac{1}{\lambda}(F(x^0) - \inf F).$$

Dividing by K we obtain the claimed sublinear rate. □

Remark 5.4. To our knowledge feasibility of the step-size $\lambda = 1/L$ for nonconvex g is also new in the Euclidean case as existing results typically require $\lambda < 1/L$. In the Euclidean case uniform level-boundedness is implied by prox-boundedness with some threshold λ_g; see [36, Theorem 1.25]. In the anisotropic case this holds for any $\lambda > 0$ if for example g (or ϕ) is bounded from below and ϕ (or g) is coercive in which case we have $\lambda_g = \infty$.

If L is unknown (or only an over-estimate is available) we can still apply our algorithm by estimating a (sharper) step-size λ_k in each iteration k using a backtracking linesearch: Given $0 < \lambda_{\max} < \lambda_g$ and a constant $0 < \alpha < 1$ we can backtrack $\lambda_k = \alpha^t \lambda_{\max}$ until for some $t \in \mathbb{N}_0$ the descent inequality (2) with constant $L = 1/\lambda_k$ is valid at x^k and x^{k+1}. The complete scheme is listed in Algorithm 2.

Algorithm 2 Linesearch anisotropic proximal gradient

Require: Let $0 < \alpha < 1$ and choose some $\lambda_g > \lambda_{\max} > 0$. Let $x^0 \in \mathbb{R}^n$.
  for all $k = 0, 1, \dots$ **do**
    $t \leftarrow 0$
    repeat
      $\lambda_k \leftarrow \alpha^t \cdot \lambda_{\max}$
      $y^k \leftarrow x^k - \lambda_k \nabla\phi^*(\nabla f(x^k))$
      $x^{k+1} \leftarrow \arg\min_{x\in\mathbb{R}^n} \ \lambda_k \star \phi(x - y^k) + g(x)$
      $t \leftarrow t + 1$
    until $f(x^{k+1}) \le \lambda_k \star \phi(x^{k+1} - y^k) - \lambda_k \star \phi(\lambda_k \nabla\phi^*(\nabla f(x^k))) + f(x^k)$
  end for

In the following theorem we show that Algorithm 2 converges subsequentially:

Theorem 5.5 (asymptotic convergence under backtracking linesearch)**.** *Let $\{x^k\}_{k\in\mathbb{N}_0}$ be the sequence of iterates generated by Algorithm 2 and denote by $\{\lambda_k\}_{k\in\mathbb{N}_0}$ the corresponding sequence of step-sizes. Then the sequence $\{F(x^k)\}_{k\in\mathbb{N}_0}$ is nonincreasing and convergent. In addition, every limit point $x^\star$ of $\{x^k\}_{k\in\mathbb{N}_0}$ is a stationary point of F, i.e., $0 \in \nabla f(x^\star) + \partial g(x^\star)$. In particular the minimum over the past regularized gaps vanishes sublinearly at rate $\mathcal{O}(1/K)$:*

$$\min_{k\in\{0,1,\dots,K-1\}} \mathcal{G}_F^\phi(x^k,\lambda_{\min}) \le \frac{1}{K\lambda_{\min}}\big(F(x^0) - \inf F\big),$$

where $\lambda_{\min} := \min\{\lambda_{\max}, \alpha/L\}$.

Proof. Thanks to Proposition 4.4 in each iteration $k \in \mathbb{N}_0$ the linesearch terminates with some λ_k with $\lambda_{\min} \le \lambda_k \le \lambda_{\max}$ such that

$$f(x^{k+1}) \le f(x^k) + \lambda_k \star \phi(x^{k+1} - x^k + \lambda_k \nabla\phi^*(\nabla f(x^k))) - \lambda_k \star \phi(\lambda_k \nabla\phi^*(\nabla f(x^k))).$$

Adding $g(x^{k+1})$ to both sides of the inequality we have:

$$\begin{aligned}
F(x^{k+1}) &= f(x^{k+1}) + g(x^{k+1}) \\
&\leq f(x^k) + \lambda_k \star \phi(x^{k+1} - x^k + \lambda_k \nabla\phi^*(\nabla f(x^k))) \\
&\quad - \lambda_k \star \phi(\lambda_k \nabla\phi^*(\nabla f(x^k))) + g(x^{k+1}) \\
&= \inf_{x\in\mathbb{R}^n} f(x^k) + \lambda_k \star \phi(x - x^k + \lambda_k \nabla\phi^*(\nabla f(x^k))) \\
&\quad - \lambda_k \star \phi(\lambda_k \nabla\phi^*(\nabla f(x^k))) + g(x) \\
&\leq \inf_{x\in\mathbb{R}^n} f(x^k) + \lambda_{\min} \star \phi(x - x^k + \lambda_{\min} \nabla\phi^*(\nabla f(x^k))) \\
&\quad - \lambda_{\min} \star \phi(\lambda_{\min} \nabla\phi^*(\nabla f(x^k))) + g(x) \\
&= F_{\lambda_{\min}}(x^k) = F(x^k) - (F(x^k) - F_{\lambda_{\min}}(x^k)) \\
&= F(x^k) - \lambda_{\min} \mathcal{G}_F^\phi(x^k, \lambda_{\min}),
\end{aligned}$$

where the second equality holds by the x-update and the last inequality holds by Proposition 4.4 and the inequality $\lambda_{\min}^{-1} \geq \lambda_k^{-1}$ and the last equalities by definition of $F_{\lambda_{\min}}$.

Summing the inequality from $k = 0$ to $k = K - 1$ and adapting the proof of Theorem 5.3 we obtain the claimed result. □

5.2 Linear convergence under anisotropic proximal gradient dominance

The following definition and subsequent theorem is a generalization of the proximal PL-inequality due to [21] to the anisotropic case:

Definition 5.6 (anisotropic proximal gradient dominance)**.** *We say that $F := f + g$ satisfies the anisotropic proximal gradient dominance condition relative to ϕ with constant $\mu > 0$ for parameter $0 < \lambda \leq \min\{\mu^{-1}, \lambda_g\}$ if for all $\bar{x} \in \mathbb{R}^n$*

$$\mu(F(\bar{x}) - \inf F) \leq \mathcal{G}_F^\phi(\bar{x}, \lambda).$$

The linear convergence under anisotropic proximal gradient dominance follows immediately from Lemma 5.2:

Theorem 5.7. *Let $\{x^k\}_{k\in\mathbb{N}_0}$ be the sequence of backward-steps generated by Algorithm 1. Let $F := f + g$ satisfy the anisotropic proximal gradient dominance condition relative to ϕ with constant $0 < \mu < L$ for parameter $1/L < \lambda_g$. Let $\lambda = 1/L$. Then $\{F(x^k)\}_{k\in\mathbb{N}_0}$ converges linearly, in particular,*

$$F(x^k) - \inf F \leq \left(1 - \frac{\mu}{L}\right)^k (F(x^0) - \inf F).$$

Proof. Thanks to Lemma 5.2 we have for any $k \in \mathbb{N}$ that

$$F(x^k) - \inf F \leq F(x^{k-1}) - \inf F - \frac{1}{L}\mathcal{G}_F^\phi(x^{k-1}, L^{-1}).$$

The anisotropic gradient dominance condition yields that

$$\begin{aligned}
F(x^k) - \inf F &\leq F(x^{k-1}) - \inf F - \frac{\mu}{L}(F(x^{k-1}) - \inf F) \\
&= \left(1 - \frac{\mu}{L}\right)(F(x^{k-1}) - \inf F).
\end{aligned}$$

Iterating the inequality we obtain the claimed linear convergence. □

If $g \equiv 0$ our method specializes to dual space preconditioning for gradient descent [27] for which the linear convergence was established under different conditions [27, Theorem 3.9]. As shown in [27, Proposition 3.3] these conditions are equivalent to Bregman strong convexity and smoothness of f^* relative to ϕ^*. As we shall see next this is a special case of our setting if $g \equiv 0$.

Next we show that $F = f + g$ satisfies the anisotropic proximal gradient dominance condition if f is anisotropically strongly convex defined as follows:

Definition 5.8 (anisotropic strong convexity)**.** *Let $h \in \Gamma_0(\mathbb{R}^n)$ such that $\operatorname{rge} \partial h \supseteq \operatorname{rge} \nabla \phi$. Then we say that h is anisotropically strongly convex relative to ϕ with constant μ if for all $(\bar{x}, \bar{v}) \in \operatorname{gph} \partial h \cap (\mathbb{R}^n \times \operatorname{int} \operatorname{dom} \phi^*)$ the following inequality holds true*

$$h(x) \geq h(\bar{x}) + \tfrac{1}{\mu} \star \phi(x - \bar{x} + \mu^{-1} \nabla \phi^*(\bar{v})) - \tfrac{1}{\mu} \star \phi(\mu^{-1} \nabla \phi^*(\bar{v})), \quad \forall x \in \mathbb{R}^n. \tag{33}$$

Next we show that anisotropic strong convexity of f implies anisotropic proximal gradient dominance of $F = f + g$. For that purpose we shall prove the following epi-scaling property of the gap function $\mathcal{G}_F^\phi(\cdot, \lambda)$ adapting the proof of [21, Lemma 1] to the anisotropic case:

Lemma 5.9 (scaling property of gap function)**.** *Let $g \in \Gamma_0(\mathbb{R}^n)$. For any $0 < \lambda_2 \leq \lambda_1$ we have that*

$$\mathcal{G}_F^\phi(\bar{x}, \lambda_2) \geq \mathcal{G}_F^\phi(\bar{x}, \lambda_1).$$

Proof. Without loss of generality let $\bar{x} \in \operatorname{dom} g$. We rewrite

$$\begin{aligned}
\mathcal{G}_F^\phi(\bar{x}, \lambda) &= \frac{1}{\lambda}\big(F(\bar{x}) - F_\lambda(\bar{x})\big) \\
&= \phi(\nabla \phi^*(\nabla f(\bar{x}))) \\
&\quad - \min_{x \in \mathbb{R}^n} \left\{ \phi(\lambda^{-1}(x - \bar{x}) + \nabla \phi^*(\nabla f(\bar{x}))) + \lambda^{-1}(g(x) - g(\bar{x})) \right\} \\
&= \phi(\nabla \phi^*(\nabla f(\bar{x}))) \\
&\quad - \min_{u \in \mathbb{R}^n} \left\{ \phi(u + \nabla \phi^*(\nabla f(\bar{x}))) + \lambda^{-1}(g(\lambda u + \bar{x}) - g(\bar{x})) \right\} \qquad (34) \\
&= \phi(\nabla \phi^*(\nabla f(\bar{x}))) - \min_{u \in \mathbb{R}^n} \left\{ \phi(u + \nabla \phi^*(\nabla f(\bar{x}))) + \lambda^{-1} \star h(u) \right\}, \qquad (35)
\end{aligned}$$

where (34) follows by the change of variable $u := \lambda^{-1}(x - \bar{x})$ and (35) follows by defining $h := g(\cdot + \bar{x}) - g(\bar{x})$. Thus $\mathcal{G}_F^\phi(\bar{x}, \lambda_2) \geq \mathcal{G}_F^\phi(\bar{x}, \lambda_1)$ is implied by $\frac{1}{\lambda_2} \star h(u) \leq \frac{1}{\lambda_1} \star h(u)$, which holds by convexity of h and $h(0) = 0$ thanks to [12, Lemma 4.4] also see the proof of [21, Lemma 1]. □

Now we are ready to prove that anisotropic strong convexity of f implies anisotropic proximal gradient dominance of $F = f + g$:

Proposition 5.10 (anisotropic strong convexity implies anisotropic proximal gradient dominance)**.** *Let f be anisotropically strongly convex relative to ϕ with constant μ and $g \in \Gamma_0(\mathbb{R}^n)$. Then $\operatorname{rge} \nabla f = \operatorname{rge} \nabla \phi$ and $F := f + g$ is coercive and strictly convex relative to its effective domain implying that it has a unique minimizer $x^\star = \arg\min F$. In particular, F satisfies the anisotropic proximal gradient dominance condition relative to ϕ with constant μ for any parameter $0 < \lambda \leq \mu^{-1}$, i.e.,*

$$\mu(F(\bar{x}) - F(x^\star)) \leq \mathcal{G}_F^\phi(\bar{x}, \lambda).$$

Proof. Since f is anisotropically strongly convex we have $\operatorname{rge}\nabla f \supseteq \operatorname{rge}\nabla\phi$. By anisotropic smoothness of f we also have $\operatorname{rge}\nabla f \subseteq \operatorname{rge}\nabla\phi$. Combining the inclusions we have $\operatorname{rge}\nabla f = \operatorname{rge}\nabla\phi = \operatorname{int}\operatorname{dom}\phi^*$. Let $\bar{x}\in\mathbb{R}^n$. The anisotropic strong convexity inequality yields

$$\begin{aligned} f(x) &\geq f(\bar{x}) + \tfrac{1}{\mu}\star\phi(x-\bar{x}+\mu^{-1}\nabla\phi^*(\nabla f(\bar{x}))) - \tfrac{1}{\mu}\star\phi(\mu^{-1}\nabla\phi^*(\nabla f(\bar{x}))) \\ &> f(\bar{x}) + \langle\nabla f(\bar{x}), x-\bar{x}\rangle, \end{aligned}$$

for all $x\in\mathbb{R}^n$ with $x\neq\bar{x}$, where the second inequality follows due to strict convexity of ϕ. Thus f is strictly convex and therefore $F=f+g$ is strictly convex relative to $\operatorname{dom}F=\operatorname{dom}g$ implying that F^* is essentially smooth. Thanks to Lemma 2.7(i) the convex lower bound

$$\xi_{\bar{x}}(x) := f(\bar{x}) + \tfrac{1}{\mu}\star\phi(x-\bar{x}+\mu^{-1}\nabla\phi^*(\nabla f(\bar{x}))) - \tfrac{1}{\mu}\star\phi(\mu^{-1}\nabla\phi^*(\nabla f(\bar{x}))) + g(x),$$

of F has the unique minimizer $\arg\min_{x\in\mathbb{R}^n}\xi_{\bar{x}}(x) = \operatorname{aprox}_g^{\mu\star\phi_-}(\bar{x}-\mu^{-1}\nabla\phi^*(\nabla f(\bar{x})))$. Invoking [36, Theorem 11.8(b)] we know that $0\in\operatorname{dom}\partial\xi_{\bar{x}}^*$. Since $\xi_{\bar{x}}$ is strictly convex relative to $\operatorname{dom}g$ we have that $\xi_{\bar{x}}^*$ is essentially smooth implying that $0\in\operatorname{dom}\partial\xi_{\bar{x}}^* = \operatorname{int}(\operatorname{dom}\xi_{\bar{x}}^*)$. In view of [36, Theorem 11.8(c)] this means that $\xi_{\bar{x}}$ is coercive. Since $\xi_{\bar{x}}\leq F$ the cost F is coercive too. Again invoking [36, Theorem 11.8(c)] we know that $0\in\operatorname{int}\operatorname{dom}F^*$. Since F^* is essentially smooth we have that $0\in\operatorname{int}\operatorname{dom}F^* = \operatorname{dom}\partial F^*$. Invoking [36, Theorem 11.8(b)] F^* has a unique minimizer $x^\star$. By anisotropic strong convexity we have:

$$f(x) \geq f(\bar{x}) + \tfrac{1}{\mu}\star\phi(x-\bar{x}+\mu^{-1}\nabla\phi^*(\nabla f(\bar{x}))) - \tfrac{1}{\mu}\star\phi(\mu^{-1}\nabla\phi^*(\nabla f(\bar{x}))).$$

Adding $g(x)$ to both sides of the inequality and minimizing both sides wrt x we have

$$\begin{aligned} F(x^\star) &\geq \inf_{x\in\mathbb{R}^n} \tfrac{1}{\mu}\star\phi(x-\bar{x}+\mu^{-1}\nabla\phi^*(\nabla f(\bar{x}))) - \tfrac{1}{\mu}\star\phi(\mu^{-1}\nabla\phi^*(\nabla f(\bar{x}))) + g(x) \\ &= F_{\mu^{-1}}(\bar{x}). \end{aligned}$$

This implies that $\mu^{-1}\mathcal{G}_F^\phi(\bar{x},\mu^{-1}) = F(\bar{x}) - F_{\mu^{-1}}(\bar{x}) \geq F(\bar{x}) - F(x^\star)$. Invoking Lemma 5.9 we have $\mu(F(\bar{x})-F(x^\star)) \leq \mathcal{G}_F^\phi(\bar{x},\mu^{-1}) \leq \mathcal{G}_F^\phi(\bar{x},\lambda)$, for any $\lambda\leq\mu^{-1}$. □

Next we prove a conjugate duality correspondence between anisotropic strong convexity and relative smoothness in the Bregman sense [10, 6, 26] generalizing [24] to reference functions which are not necessarily super-coercive. This turns out helpful in the subsequent Examples 5.14 and 5.16.

Proposition 5.11 (conjugate duality between anisotropic strong convexity and relative smoothness)**.** *Let $h\in\Gamma_0(\mathbb{R}^n)$. Assume that* $\operatorname{rge}\partial h \supseteq \operatorname{rge}\nabla\phi$. *Let $\mu>0$. Then the following are equivalent:*

(i) *h satisfies the anisotropic strong convexity inequality for all $(\bar{x},\bar{v})\in\operatorname{gph}\partial h\cap(\mathbb{R}^n\times\operatorname{int}\operatorname{dom}\phi^*)$, i.e., the following inequality holds true*

$$h(x) \geq h(\bar{x}) + \tfrac{1}{\mu}\star\phi(x-\bar{x}+\mu^{-1}\nabla\phi^*(\bar{v})) - \tfrac{1}{\mu}\star\phi(\mu^{-1}\nabla\phi^*(\bar{v})) \quad \forall x\in\mathbb{R}^n; \tag{36}$$

(ii) *h^* is smooth relative to ϕ^* in the Bregman sense with constant μ^{-1}, i.e., $\mu^{-1}\phi^* \mathbin{\dot{-}} h^*$ is convex on* $\operatorname{int}\operatorname{dom}\phi^*$.

The proof is an adaption of the proof of [24, Theorem 4.3]. For completeness an adapted proof is provided in Appendix A.6.

Remark 5.12. For super-coercive ϕ the inequality (56) holds with equality and thus the conjugate $h = h^{**}$ of a convex Bregman smooth function h^* is a so-called Klee envelope $h^{**}(x) = \sup_{y\in\mathbb{R}^n} \phi(x-y) \dot{-} g^*_-(y)$. As shown in [24] the converse, however, is false in general unless $\phi = \frac{1}{2}\|\cdot\|^2$, h is essentially smooth, or a finite-valued, one-dimensional function.

Remark 5.13. The constraint qualification $\operatorname{rge}\partial h \supseteq \operatorname{rge}\nabla\phi$ in the theorem above is equivalent to $\operatorname{dom} h^* \supseteq \operatorname{dom}\phi^*$ which is the standard constraint qualification in Bregman smoothness, see, e.g., [6].

This is proved as follows: $\operatorname{dom} h^* \supseteq \operatorname{dom}\phi^*$ implies that $\operatorname{int}\operatorname{dom} h^* \supseteq \operatorname{int}\operatorname{dom}\phi^*$ and thus $\operatorname{rge}\partial h = \operatorname{dom}\partial h^* \supseteq \operatorname{int}\operatorname{dom} h^* \supseteq \operatorname{int}\operatorname{dom}\phi^* = \operatorname{rge}\nabla\phi$.

To show the opposite implication let $x \in \operatorname{dom}\phi^*$ and $x_0 \in \operatorname{int}\operatorname{dom}\phi^*$. We have

$$\begin{aligned} h^*(x) &\le \liminf_{\tau\nearrow 1} h^*((1-\tau)x_0 + \tau x) \\ &\le \lim_{\tau\nearrow 1} h^*(x_0) + \langle \nabla h^*(x_0), (1-\tau)x_0 + \tau x - x_0\rangle + D_{\phi^*}((1-\tau)x_0 + \tau x, x_0) \\ &= h^*(x_0) + \langle \nabla h^*(x_0), x - x_0\rangle + D_{\phi^*}(x, x_0), \end{aligned}$$

where the first inequality holds since h^* is lsc, the second since in view of the proof of the above theorem, $\phi^* - h^*$ is convex and smooth on $\operatorname{int}\operatorname{dom}\phi^*$, and $(1-\tau)x_0 + \tau x \in \operatorname{int}\operatorname{dom}\phi^*$ for $\tau \in [0,1)$ by the line segment principle [36, Theorem 2.33] and the last equality holds due to [36, Theorem 2.35] and since $\phi^* \in \Gamma_0(\mathbb{R}^n)$. Since $x \in \operatorname{dom}\phi^*$ and $x_0 \in \operatorname{int}\operatorname{dom}\phi^*$ the right-hand side of the inequality is finite and thus $x \in \operatorname{dom} h^*$.

The next examples show that the absolute value function and the quadratic function are both anisotropically strongly convex relative to a softmax approximation of the absolute value function:

Example 5.14. Let $g(x) := \nu|x|$ with $\nu \ge 1$. Then g is anisotropically strongly convex relative to the symmetrized logistic loss $\phi(x) := 2\log(1+\exp(x)) - x$ with any constant $\mu > 0$.

Proof. The individual conjugates amount to $g^*(x) = \delta_{[-\nu,\nu]}(x)$ and $\phi^*(x) = (x+1)\log((x+1)/2) + (1-x)\log((1-x)/2)$ with $x\log x + (1-x)\log(1-x) = 0$ for $x \in \{0,1\}$ and $\operatorname{dom}\phi^* = [-1,1]$. Since $\operatorname{dom} g^* = [-\nu,\nu] \supseteq [-1,1]$ thanks to Proposition 5.11 anisotropic strong convexity of g is implied by convexity of $\frac{1}{\mu}\phi^* \dot{-} g^*$ on $(-1,1)$. Since $g^* \equiv 0$ on $(-1,1)$ this is valid for any $\mu > 0$. □

Remark 5.15 (Disclaimer). It should be noted that for $\nu \ge 1$ Example 5.14 is rather of theoretical interest: For the smooth part f we have the restriction $\operatorname{rge} f' \subseteq \operatorname{int}\operatorname{dom}\phi^* = (-1,1)$ and thus the first-order optimality condition $-f'(x^\star) \in \partial g(x^\star)$ is only valid at $x^\star = 0$.

Example 5.16. Let $g(x) := \frac{\nu}{2}x^2$ with $\nu > 0$. Then we show that g is anisotropically strongly convex relative to $\phi(x) := 2\log(1+\exp(x)) - x$ with constant $\mu := 2\nu$.

Proof. The individual conjugates amount to $g^*(x) = \frac{1}{2\nu}x^2$ and $\phi^*(x) = (x+1)\log((x+1)/2) + (1-x)\log((1-x)/2)$ with $x\log x + (1-x)\log(1-x) = 0$ for $x \in \{0,1\}$ and $\operatorname{dom}\phi^* = [-1,1]$. Since $\operatorname{dom} g^* = \mathbb{R}^n$ thanks to Proposition 5.11 anisotropic strong convexity of g is implied by convexity of $\frac{1}{\mu}\phi^* \dot{-} g^*$ on $(-1,1)$. This is implied by $(\phi^*)''(x) = 2/(1-x^2) \ge \frac{\mu}{\nu}$ for all $x \in (-1,1)$ which is valid for $\mu = 2\nu$. □

6 Difference of Φ-convex approach

In this section we provide an equivalent interpretation of the algorithm in terms of a *difference of Φ-convex approach* and show that the algorithm is invariant under a certain double-min duality (DC-duality). This allows us to transfer smoothness between the components f and g of the optimization problem. If g in place of f is anisotropically strongly convex this is used to show that the DC-dual problem has the anisotropic proximal gradient dominance condition and the algorithm attains linear convergence.

In view of Propositions 3.3 and 4.4 anisotropic smoothness of f relative to ϕ with constant L implies that $-f$ is Φ-convex relative to $\Phi(x,y) = -\lambda \star \phi(x-y)$ for any $\lambda \leq 1/L$ and thus that $(-f)(x) = (-f)^{\Phi\Phi}(x) = -\sup_{y\in\mathbb{R}^n} -\lambda \star \phi(x-y) - (-f)^{\Phi}(y)$ where the supremum is attained. By invoking the notion of the Φ-subdifferential this allows us to provide an alternative interpretation of our algorithm in terms of a difference of Φ-convex approach applied to $F = g - (-f)$: Thanks to Propositions 3.5 and 3.6 the updates in Algorithm 1 can be rewritten:

$$y^k = x^k - \lambda \nabla\phi^*(\nabla f(x^k)) \in \partial_\Phi(-f)(x^k) \tag{37}$$

$$x^{k+1} = \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \ \lambda \star \phi(x-y^k) + g(x) = \operatorname*{arg\,max}_{x\in\mathbb{R}^n} \ \Phi(x,y^k) - g(x) = \partial_\Phi g^\Phi(y^k). \tag{38}$$

Up to replacing the classical subdifferential with the Φ-subdifferential this algorithm exactly resembles the structure of the classical *difference of convex approach* (DCA).

Next we shall discuss a double-min duality also called DC-duality in the classical DCA; see, e.g., [39]. In particular we show that the algorithm is invariant under this duality.

Based on the identity $-f = (-f)^{\Phi\Phi}$ assuming that the optimization problem (P) has a minimizer we can rewrite (P) as:

$$\begin{aligned}
\min_{x\in\mathbb{R}^n} \ g(x) + f(x) &= \min_{x\in\mathbb{R}^n} \ g(x) - (-f)^{\Phi\Phi}(x) \\
&= \min_{x\in\mathbb{R}^n} \ g(x) - \max_{y\in\mathbb{R}^n} -\lambda \star \phi(x-y) - (-f)^{\Phi}(y) \\
&= \min_{x\in\mathbb{R}^n} \min_{y\in\mathbb{R}^n} \ g(x) + \lambda \star \phi(x-y) + (-f)^{\Phi}(y) \\
&= \min_{y\in\mathbb{R}^n} \min_{x\in\mathbb{R}^n} \ g(x) + \lambda \star \phi(x-y) + (-f)^{\Phi}(y) \qquad (39) \\
&= \min_{y\in\mathbb{R}^n} (-f)^{\Phi}(y) - \max_{x\in\mathbb{R}^n} -\lambda \star \phi(x-y) - g(x) \\
&= \min_{y\in\mathbb{R}^n} (-f)^{\Phi}(y) - g^{\Phi}(y), \qquad (40)
\end{aligned}$$

where reversing the order of minimization is also called double-min duality; see [36, Theorem 11.67]. In the spirit of classical DC-duality we refer to (40) as the DC-dual problem of (P).

Further rewriting $-g^\Phi$ respectively $(-f)^\Phi$ in terms of the infimal convolution respectively infimal deconvolution, introduced in section 3, the DC-dual problem of (P) is rewritten as

$$\text{minimize } G := (g \boxplus \lambda \star \phi_-) + (f \boxminus \lambda \star \phi). \tag{D}$$

The epi-graphical addition and subtraction of the reference function leads to a transfer of smoothness in the DC-dual and corresponds to pointwise addition and subtraction of the conjugate reference function in the Fenchel–Rockafellar dual. In the Euclidean case this is also related to primal and dual regularization in classical DC-programming [39, Section 5.2].

Next we show that up to interchanging the roles of forward- and backward-step our algorithm is invariant under double-min duality as is known for the classical DCA. The following result is a generalization of [23, Theorem 3.8] for the Euclidean case to the anisotropic case:

Proposition 6.1 (interchange of forward- and backward-step)**.** *Let g be convex. Then $g \boxplus \lambda \star \phi_-$ is anisotropically smooth relative to ϕ_- with constant λ^{-1} and for any $y \in \mathbb{R}^n$ a backward-step on g at y is a forward-step on $g \boxplus \lambda \star \phi_-$ at y, i.e.,*

$$y - \lambda \nabla \phi_-^*(\nabla(g \boxplus \lambda \star \phi_-)(y)) = \operatorname{aprox}_g^{\lambda \star \phi_-}(y).$$

For any $x \in \mathbb{R}^n$ a forward-step on f at x is backward-step on $f \boxminus \lambda \star \phi$ at x, i.e.,

$$\operatorname{aprox}_{f \boxminus \lambda \star \phi}^{\lambda \star \phi}(x) = x - \lambda \nabla \phi^*(\nabla f(x)).$$

If, in addition, f is convex we have $f \boxminus \lambda \star \phi \in \Gamma_0(\mathbb{R}^n)$ with $\operatorname{dom}(f \boxminus \lambda \star \phi)^* \cap \operatorname{int} \operatorname{dom} \phi^* \neq \emptyset$, i.e., the constraint qualification for the anisotropic proximal mapping in the convex case is satisfied.

Proof. In view of Lemma 2.7(ii) the following gradient formula for the anisotropic Moreau envelope holds true:

$$\nabla(g \boxplus \lambda \star \phi_-) = \nabla \phi_- \circ \lambda^{-1}(\operatorname{id} - \operatorname{aprox}_g^{\lambda \star \phi_-}), \tag{41}$$

and in particular $\operatorname{rge} \nabla(g \boxplus \lambda \star \phi_-) = \operatorname{rge} \nabla \phi_- \circ \lambda^{-1}(\operatorname{id} - \operatorname{aprox}_g^{\lambda \star \phi_-}) \subseteq \operatorname{rge} \nabla \phi_-$. Thus we can invoke Proposition 3.7 and obtain that $g \boxplus \lambda \star \phi_-$ has the anisotropic smoothness property relative ϕ_- with constant λ^{-1} and in particular we have for the forward-step applied to $g \boxplus \lambda \star \phi_-$ at any $y \in \mathbb{R}^n$:

$$\begin{aligned} y - \lambda \nabla \phi_-^*(\nabla(g \boxplus \lambda \star \phi_-)(y)) &= y - \lambda \nabla \phi_-^*(\nabla \phi_-(\lambda^{-1}(y - \operatorname{aprox}_g^{\lambda \star \phi_-}(y)))) \\ &= y - (y - \operatorname{aprox}_g^{\lambda \star \phi_-}(y)) = \operatorname{aprox}_g^{\lambda \star \phi_-}(y). \end{aligned}$$

Define $\Phi(x, y) := -\lambda \star \phi(x - y)$. Let $y \in \operatorname{aprox}_{f \boxminus \lambda \star \phi}^{\lambda \star \phi}(x) = \arg\max_{y \in \mathbb{R}^n} \Phi(x, y) - (-f)^\Phi(y)$ be the backward-step applied to $f \boxminus \lambda \star \phi = (-f)^\Phi$ at x. By definition of the Φ-subdifferential this is equivalent to $x \in \partial_\Phi(-f)^\Phi(y)$. Since f is anisotropically smooth relative to ϕ with constant L, by Proposition 4.4, f is anisotropically smooth relative to ϕ with constant $\lambda^{-1} \geq L$ as well. By Proposition 3.7 this means that $-f$ is Φ-convex and thus by Proposition 3.5 $x \in \partial_\Phi(-f)^\Phi(y)$ is equivalent to $y \in \partial_\Phi(-f)(x)$. Invoking Proposition 3.6 this implies that $\operatorname{aprox}_{f \boxminus \lambda \star \phi}^{\lambda \star \phi}(x) = \partial_\Phi(-f)(x) = (\operatorname{id} - \lambda \nabla \phi^* \circ \nabla f)(x) = x - \lambda \nabla \phi^*(\nabla f(x))$.

If, in addition, f is convex, thanks to Proposition 3.8 the deconvolution $f \boxminus \lambda \star \phi = (f^* \dot{-} \lambda \phi^*)^* \in \Gamma_0(\mathbb{R}^n)$. Since $\operatorname{rge} \nabla f \subseteq \operatorname{rge} \nabla \phi$ we have that $\operatorname{ri} \operatorname{dom} f^* \subseteq \operatorname{dom} \partial f^* \subseteq \operatorname{int} \operatorname{dom} \phi^*$. Since $\operatorname{dom}(f^* \dot{-} \lambda \phi^*) = \operatorname{dom} f^*$ we have $\operatorname{dom}(f^* \dot{-} \lambda \phi^*) \cap \operatorname{int} \operatorname{dom} \phi^* \neq \emptyset$. Since $(f \boxminus \lambda \star \phi)^* = (f^* \dot{-} \lambda \phi^*)^{**} \leq f^* \dot{-} \lambda \phi^*$ we have $\operatorname{dom}(f \boxminus \lambda \star \phi)^* \supseteq \operatorname{dom}(f^* \dot{-} \lambda \phi^*)$. This implies that $\operatorname{dom}(f \boxminus \lambda \star \phi)^* \cap \operatorname{int} \operatorname{dom} \phi^* \neq \emptyset$. □

Alternatively, our algorithm can be understood in terms of Gauss–Seidel minimization applied to (39) for which the same self-duality in the DC-sense is true.

Next, we exploit the transfer of smoothness in combination with the self-duality to show linear convergence wrt G if g is anisotropically strongly convex. Due to a lack of tilt-invariance a primal transfer of strong convexity as in Euclidean forward-backward splitting is in general not possible in the anisotropic case:

Lemma 6.2 (transfer of smoothness)**.** *Let g be anisotropically strongly convex with constant μ relative to ϕ_-. Then $g \boxplus \frac{1}{L} \star \phi_-$ is anisotropically strongly convex with constant $\sigma = 1/(L^{-1} + \mu^{-1}) < \mu$ and has the anisotropic descent property with constant L.*

Proof. Thanks to Proposition 5.11 anisotropic strong convexity of $g \boxplus L^{-1} \star \phi_-$ with constant σ is implied by Bregman smoothness of $(g \boxplus L^{-1} \star \phi_-)^* = g^* + L^{-1}\phi_-^*$ relative to $\sigma^{-1}\phi_-^*$, i.e., convexity of $\sigma^{-1}\phi_-^* \dot{-} (g \boxplus L^{-1} \star \phi_-)^* = (\sigma^{-1} - L)\phi_-^* \dot{-} g^*$, on $\operatorname{int}\operatorname{dom}\phi_-^*$ and $\operatorname{dom}\partial(g \boxplus L^{-1} \star \phi_-)^* \supseteq \operatorname{int}\operatorname{dom}\phi_-^*$. Since g is anisotropically strongly convex relative to ϕ_- we have $\operatorname{dom}\partial g^* \supseteq \operatorname{int}\operatorname{dom}\phi_-^*$. In light of Proposition 5.11, $\mu^{-1}\phi_-^* \dot{-} g^*$ is convex on $\operatorname{int}\operatorname{dom}\phi_-^*$. Thus $(\sigma^{-1} - L^{-1})\phi_-^* \dot{-} g^*$ is convex on $\operatorname{int}\operatorname{dom}\phi_-^*$ if $\sigma^{-1} \geq L^{-1} + \mu^{-1}$, i.e., $\sigma \leq 1/(L^{-1} + \mu^{-1})$. We also have thanks to the constraint qualification and the sum-rule and the inclusion $\operatorname{dom}\partial g^* \supseteq \operatorname{int}\operatorname{dom}\phi_-^*$ that $\operatorname{dom}\partial(g \boxplus L^{-1} \star \phi_-)^* = \operatorname{dom}(\partial g^* + \nabla\phi_-^*) = \operatorname{dom}\partial g^* \cap \operatorname{dom}\nabla\phi_-^* = \operatorname{dom}\partial g^* \supseteq \operatorname{int}\operatorname{dom}\phi_-^*$. This proves the claimed result. □

Thanks to the self-duality of our algorithm in the DC-sense and the previous proposition it can be shown that the method converges linearly wrt to the DC-dual cost $G = (g \boxplus \lambda \star \phi_-) + (f \boxminus \lambda \star \phi)$ evaluated at the forward-steps y^k:

Proposition 6.3. *Let f be convex and g be anisotropically strongly convex with constant μ relative to ϕ_-. Let $\{y^k\}_{k\in\mathbb{N}_0}$ be the sequence of forward-steps generated by Algorithm 1 with $\lambda = 1/L$. Then G has a unique minimizer $y^\star$ and $\{G(y^k)\}_{k\in\mathbb{N}_0}$ converges linearly, in particular,*

$$G(y^k) - G(y^\star) \leq \left(1 - \frac{\mu}{\mu + L}\right)^k (G(y^0) - G(y^\star)).$$

Proof. By Lemma 6.2 $g \boxplus \lambda\star\phi_-$ is anisotropically smooth with constant L and anisotropically strongly convex with constant $1/(L^{-1} + \mu^{-1})$. By Proposition 6.1 $f \boxminus \lambda \star \phi \in \Gamma_0(\mathbb{R}^n)$ with $\operatorname{dom}(f \boxminus \lambda \star \phi) \cap \operatorname{int}\operatorname{dom}\phi^* \neq \emptyset$ and therefore we can invoke Proposition 5.10 to show that G has a unique minimizer $y^\star$ and $G = (g \boxplus \lambda\star\phi_-) + (f \boxminus \lambda\star\phi)$ satisfies the anisotropic proximal gradient dominance condition relative to ϕ_- with constant $1/(L^{-1} + \mu^{-1})$ and parameter L. Thanks to Proposition 6.1 we have for every $k \in \mathbb{N}_0$ that $y^{k+1} = x^{k+1} - \lambda\nabla\phi^*(x^{k+1}) = \operatorname{aprox}_{f \boxminus \lambda\star\phi}^{\lambda\star\phi}(x^{k+1})$ and $x^{k+1} = \operatorname{aprox}_g^{\lambda\star\phi_-}(y^k) = y^k - \nabla\phi_-^*(\nabla(g \boxplus \phi_-))(y^k)$. In combination we have $y^{k+1} = \operatorname{aprox}_{f \boxminus \lambda\star\phi}^{\lambda\star\phi}(y^k - \nabla\phi_-^*(\nabla(g \boxplus \phi_-))(y^k))$. This is identical to Algorithm 1 with step-size λ, reference function ϕ_- and initial iterate y^0 applied to $G = (g \boxplus \lambda\star\phi_-) + (f \boxminus \lambda\star\phi)$. Thus we can apply Theorem 5.7 to the sequence $\{y^k\}_{k\in\mathbb{N}_0}$ to show that $\{G(y^k)\}_{k\in\mathbb{N}_0}$ converges linearly with the claimed constant. □

The above result can be translated to primal linear convergence by showing that $G(y^k) = F_\lambda(x^k)$ which relates the dual cost G and the forward-backward envelope F_λ to each other. This relation was observed previously in the Euclidean setting [41]:

Proposition 6.4 (DC-dual vs. forward-backward envelope)**.** *Let $\{x^k\}_{k\in\mathbb{N}_0}$ be the sequence of backward-steps and $\{y^k\}_{k\in\mathbb{N}_0}$ be the sequence of forward-steps generated by Algorithm 1. Then it holds that*

$$G(y^k) = (g \boxplus \lambda \star \phi_-)(y^k) + (f \boxminus \lambda \star \phi)(y^k) = F_\lambda(x^k),$$

for all $k \in \mathbb{N}_0$.

Proof. Let $k \in \mathbb{N}_0$. In view of Proposition 6.1 we have $y^k = x^k - \lambda\nabla\phi^*(\nabla f(x^k)) = \arg\min_{y\in\mathbb{R}^n} \lambda \star \phi(x^k - y) + (f \boxminus \lambda \star \phi)(y)$. Since f is anisotropically smooth relative to ϕ with constant L, by Proposition 4.4, f is anisotropically smooth relative to ϕ with constant $\lambda^{-1} \geq L$ as well. Owing to Proposition 3.7 $-f$ is Φ-convex relative to $\Phi(x, y) = -\lambda\star\phi(x-y)$ and thus by Proposition 3.3 $f = -(-f)^{\Phi\Phi} = \inf_{y\in\mathbb{R}^n} \lambda \star \phi(x^k - y) + (f \boxminus \lambda \star \phi)(y)$ implying that

$$(f \boxminus \lambda \star \phi)(y^k) + \lambda \star \phi(x^k - y^k) = f(x^k). \tag{42}$$

By definition of the forward-backward envelope since $x^k - y^k = \lambda\nabla\phi^*(\nabla f(x^k))$ we have that $F_\lambda(x^k) = (g \boxplus \lambda \star \phi_-)(y^k) + f(x^k) - \lambda \star \phi(x^k - y^k)$. In combination with (42) we obtain that

$$F_\lambda(x^k) = (g \boxplus \lambda \star \phi_-)(y^k) + (f \boxminus \lambda \star \phi)(y^k) = G(y^k) \qquad \square$$

Now we are ready to state the primal linear convergence:

Corollary 6.5. *Let f be convex and g be anisotropically strongly convex with constant μ relative to ϕ_-. Let $\{x^k\}_{k\in\mathbb{N}_0}$ be the sequence of backward-steps generated by Algorithm 1 with step-size $\lambda = L^{-1}$. Then $\{F(x^{k+1})\}_{k\in\mathbb{N}_0}$ converges linearly, in particular,*

$$F(x^{k+1}) - \inf F \leq \left(1 - \frac{\mu}{\mu + L}\right)^k (F(x^0) - \inf F).$$

Proof. Let $k \in \mathbb{N}_0$. Since f is anisotropically smooth the anisotropic descent inequality holds true which after adding $g(x^{k+1})$ to both sides of the inequality amounts to:

$$\begin{aligned} F(x^{k+1}) &\leq f(x^k) + \lambda \star \phi(x^{k+1} - y^k) - \lambda \star \phi(\lambda\nabla\phi^*(\nabla f(x^k))) + g(x^{k+1}) \\ &= F_\lambda(x^k), \end{aligned} \tag{43}$$

where the last equality follows by definition of F_λ and x^{k+1}. Define $\xi(x, x^k) := f(x^k) + \lambda \star \phi(x - x^k + \lambda\nabla\phi^*(\nabla f(x^k))) - \lambda \star \phi(\lambda\nabla\phi^*(\nabla f(x^k))) + g(x^{k+1})$. Then by definition $F_\lambda(x^k) = \inf_{x\in\mathbb{R}^n} \xi(x, x^k)$ and $\xi(x^k, x^k) = F(x^k)$. Therefore we also have

$$F_\lambda(x^k) \leq F(x^k). \tag{44}$$

Invoking Proposition 6.3 G has a unique minimizer $y^\star$ and we have after K iterations:

$$G(y^k) - G(y^\star) \leq \left(1 - \frac{\mu}{\mu + L}\right)^k (G(y^0) - G(y^\star)).$$

By double-min duality (39) we have $G(y^\star) = \inf F$. Invoking Proposition 6.4 we have that $F_\lambda(x^k) = G(y^k)$ for all $k \in \mathbb{N}_0$. Thanks to the identities (43) and (44) we have $F(x^{k+1}) \leq F_\lambda(x^k) = G(y^k)$ and $F(x^0) \geq F_\lambda(x^0) = G(y^0)$ and thus we obtain:

$$\begin{aligned} F(x^{k+1}) - \inf F \leq G(y^k) - G(y^\star) &\leq \left(1 - \frac{\mu}{\mu + L}\right)^k (G(y^0) - G(y^\star)) \\ &\leq \left(1 - \frac{\mu}{\mu + L}\right)^k (F(x^0) - \inf F). \end{aligned} \qquad \square$$

7 Applications

7.1 Regularized logistic regression

In this section we consider regularized logistic regression. Here, one is interested in the minimization of the following cost function:

$$\min_{x\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m} \log(1+\exp(-b_i\langle a_i, x\rangle)) + g(x), \tag{45}$$

for m input output training pairs $(a_i, b_i) \in [-1,1]^n \times \{-1,1\}$ and g is a possibly nonsmooth regularizer. Thanks to Example 4.8 $f(x) = \frac{1}{m}\sum_{i=1}^m \log(1+\exp(-b_i\langle a_i, x\rangle))$ is anisotropically smooth relative to the symmetrized logistic loss $\phi(x) = \sum_{j=1}^n h(x_j)$ with $h(t) = 2\log(1+\exp(t)) - t$. The conjugate amounts to $h^*(t) = (t+1)\log((t+1)/2) + (1-t)\log((1-t)/2)$ with $\operatorname{dom} h^* = [-1,1]$ where $t\log t + (1-t)\log(1-t) = 0$ for $t \in \{0,1\}$. Note that ϕ is a softmax approximation to the one-norm.

We discuss the following choices of regularizers $g = \nu\|\cdot\|_1$ and $g = \frac{\nu}{2}\|\cdot\|^2$. For both choices it holds that $\operatorname{dom} g^* \cap \operatorname{int}\operatorname{dom}\phi_-^* \neq \emptyset$ and thus the algorithm can be applied. Thanks to Example 5.16 and Corollary 6.5 the algorithm attains linear convergence wrt the DC-dual cost for $g = \frac{\nu}{2}\|\cdot\|^2$. For $g = \nu\|\cdot\|_1$ a closed form solution to the anisotropic proximal mapping is obtained via the Bregman–Moreau decomposition Lemma 2.7(i). Then the anisotropic proximal mapping can be computed in terms of the Bregman proximal mapping of $g^* = \delta_{[-\nu,\nu]^n}$ whose solution involves a simple clipping operation. The anisotropic proximal mapping of g thus amounts to the following soft-thresholding operation: $[\operatorname{aprox}_g^{\lambda\star\phi_-}(y)]_i = \operatorname{sign}(y_i)\max\{|y_i| - \rho, 0\}$, for $\rho = \lambda(h^*)'(\nu)$. For $g = \frac{\nu}{2}\|\cdot\|^2$ by a change of variable the anisotropic proximal mapping of g can be written as the Euclidean proximal mapping of the logistic loss $x \mapsto \log(1+\exp(x))$. Thanks to [11, Proposition 2] its solution can be obtained in closed form via the generalized Lambert W function [29, 28]. Also nonconvex choices for g are possible.

7.2 Exponential regularization for linear programs

In this section we consider exponential smoothing for linear programming. Given a *linear program* (LP) with linear inequality constraints $\langle a_i, x\rangle - b_i \leq 0$ we consider an unconstrained approximation to the original LP which takes the form

$$\min_{x\in\mathbb{R}^n} \ \langle c, x\rangle + \sum_{i=1}^{m} \sigma \star \exp(\langle a_i, x\rangle - b_i), \tag{46}$$

for $c, a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ and $\sigma > 0$ is a regularization parameter. The functions $\exp(\langle a_i, x\rangle - b_i)$ can be interpreted as penalty functions for the constraints. Exponential smoothing is dual to entropic regularization which is a common technique for solving large-scale optimal transport problems [15]. Alternatively, the exponential loss function appears in boosting for machine learning [18, 37].

For applying our algorithm we rewrite the cost as $f + g$ for functions $g(x) := \langle c, x\rangle$ and $f(x) := \sum_{i=1}^m \sigma \star \exp(\langle a_i, x\rangle - b_i)$. Assuming that $a_i \in \mathbb{R}^n_+$ with $\sum_{i=1}^m a_i \in \mathbb{R}^n_{++}$, thanks to Example 4.10, f has the anisotropic descent property relative to $\phi = \operatorname{Exp}$ with constant $L = \max_{1\leq i\leq m}\|a_i\|_1/\sigma$. Further assuming that $-c \in \mathbb{R}^n_+ = \operatorname{int}\operatorname{dom}\phi^*$ we also have that

$\emptyset \neq \operatorname{dom} g^* \cap \operatorname{int} \operatorname{dom} \phi_-^*$ and thus our algorithm can be applied. The anisotropic proximal mapping $\operatorname{aprox}_g^{\lambda \star \phi_-}$ of g is separable where each individual proximal mapping amounts to $x_j = \arg\min_{t \in \mathbb{R}} \lambda \star \exp(t - y_j) + c_j t$, which by first-order stationarity yields $x_j = y_j + \lambda \log(c_j)$.

Exponential regularization for linear programs was also considered in [27] using a different reference function. However, this imposes restrictions on the rank of the constraint matrix A and the corresponding feasible set.

Using Exp as the reference function we are able to recover the Sinkhorn algorithm for regularized optimal transport [15] and a classical parallel update algorithm for AdaBoost [13] as special cases of the proposed anisotropic proximal gradient framework. This complements the existing primal interpretation in terms of entropic subspace projections [9, 13]; also see [16].

For marginals $r \in \mathbb{R}^n_{++}$ and $s \in \mathbb{R}^m_{++}$ and transportation cost $C \in \mathbb{R}^{m \times n}$ the regularized dual OT problem takes the form

$$\min_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m} \langle -r, \alpha \rangle + \langle -s, \beta \rangle + \sum_{i=1}^m \sum_{j=1}^n \sigma \star \exp(\langle e_j, \alpha \rangle + \langle e_i, \beta \rangle - C_{ij}). \tag{47}$$

In order to recover the Sinkhorn algorithm we apply our method in a Gauss–Seidel fashion where α^{k+1} and β^{k+1} are obtained by anisotropic proximal gradient steps applied to the objective for fixed β^k respectively α^{k+1}. Then $f(\alpha, \beta^k)$ respectively $f(\alpha^{k+1}, \beta)$ both have the anisotropic descent property relative to $\phi = \operatorname{Exp}$ with constants $\|e_i\|_1/\sigma = \|e_j\|_1/\sigma = 1/\sigma$. In fact, for fixed α or β the objective is separable and anisotropic proximal gradient steps with step-size $1/\sigma$ correspond to exact minimization steps in the same way Euclidean coordinate-wise gradient descent on a quadratic objective corresponds to exact coordinate descent for an appropriately chosen step-size. The exact dual coordinate descent interpretation of Sinkhorn for OT is well known in literature. Alternatively, the anisotropic proximal gradient method can be applied in a joint fashion. Then, the smoothness constant of f becomes $\|e_i\|_1/\sigma + \|e_j\|_1/\sigma = 2/\sigma$.

A drawback of the exponential reference function is the fact that the corresponding pre-conditioner $\nabla \phi^*$ is only defined on the positive orthant. This imposes a "one-sided" structure on the optimization problem in the sense that the coupling vectors $a_i \in \mathbb{R}^n_+$ are constrained to the nonnegative orthant; see Example 4.10. To handle "two-sided" optimization problems, the following lifting can be applied as a workaround: We introduce an additional variable x_- along with the linear sharing constraint $x_- = -x$. We define

$$a_{ij}^+ := \begin{cases} a_{ij} & \text{if } a_{ij} \geq 0 \\ 0 & \text{otherwise,} \end{cases} \qquad a_{ij}^- := \begin{cases} a_{ij} & \text{if } a_{ij} \leq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{48}$$

Analogously, we define c^+, c^-. Then we rewrite the problem equivalently as

$$\min_{x, x_- \in \mathbb{R}^n} \sum_{i=1}^m \sigma \star \exp(\langle a_i^+, x \rangle + \langle a_i^-, x_- \rangle - b_i) + \langle x, c^+ \rangle + \langle x_-, c^- \rangle + \delta_C(x, x_-), \tag{49}$$

where $f(x, x_-) = \sum_{i=1}^m \sigma \star \exp(\langle a_i^+, x \rangle + \langle a_i^-, x_- \rangle - b_i) + \langle x, c^+ \rangle + \langle x_-, c^- \rangle$ and $g(x, x_-) := \delta_C(x, x_-)$ for $C = \{(x, x_-) \in (\mathbb{R}^n)^2 : x_- = -x\}$. We identify the vectors $a_i \in \mathbb{R}^n$ and $a_i^+, a_i^- \in \mathbb{R}^n_+$ as the rows of the matrices A_i, A_i^+, A_i^-. Thanks to Lemma 2.7(i) the x-update can be expressed in terms of the Bregman projection onto the consensus subspace $C^* = \{(x, y) \in$

$\mathbb{R}^n \times \mathbb{R}^n \mid x = y\}$ which, thanks to [3, Example 3.16(i)], admits a simple closed form solution $(x, x) = \operatorname{bprox}_{\delta_{C^*}}^{\lambda\phi^*}(y_1, y_2)$ in terms of the geometric mean: $x = \nabla\phi((\nabla\phi^*(y_1) + \nabla\phi^*(y_2))/2)$. Eliminating x_-^k and y^k, y_-^k from the algorithm the x-update can be written compactly as follows:

$$x^{k+1} = x^k - \frac{\lambda}{2}(\nabla\phi^*(A_+^\top \nabla(\sigma \star \operatorname{Exp})(Ax^k - b) + c^+) \\ - \nabla\phi^*(A_-^\top \nabla(\sigma \star \operatorname{Exp})(Ax^k - b) + c^-)). \tag{50}$$

Without loss of generality we may assume that $\sum_{i=1}^m a_i^+ \in \mathbb{R}^n_{++}$ and $\sum_{i=1}^m a_i^- \in \mathbb{R}^n_{++}$ as well as $c^+, c^- \in \mathbb{R}^n_{++}$. Then $\sum_{i=1}^m \sigma \star \exp(\langle a_i^+, x\rangle + \langle a_i^-, x_-\rangle - b_i)$ complies with the requirements in Example 4.10 for $\phi = \operatorname{Exp}$ and $L = \max_{1\leq i\leq m} \|a_i\|_1/\sigma$. Thanks to Proposition 3.7 $\langle x, c^+\rangle + \langle x_-, c^-\rangle$ is anisotropically smooth relative to Exp for any $L > 0$. Invoking Lemma 4.9 we deduce that $f(x, x_-)$ remains anisotropically smooth relative to Exp with constant L. For the special case that $b = c = 0$, $\sigma = 1$ and $A \in [-1, 1]^n$ with $\|a_i\|_1 \leq 1$ the smoothness constant and step-size can be chosen to be $\lambda = L = 1$ and (50) specializes to the parallel update optimization algorithm [13] for AdaBoost [18, 37] with $q_0 = \mathbb{1}$.

8 Conclusion

In this paper we have considered dual space nonlinear preconditioning for forward-backward splitting. The algorithm is an extension of dual space preconditioning for gradient descent [27] to the nonconvex and composite setting. In our case, the method is derived via an anisotropic descent inequality relative to a reference function whose inverse gradient takes the role of a nonlinear preconditioner in the proximal gradient scheme. Building upon recent developments for the Bregman proximal average [43] a calculus for anisotropic smoothness is derived under joint convexity of the dual Bregman distance and practical examples are provided. We prove subsequential convergence of the method using a regularized gap function which vanishes at rate $\mathcal{O}(1/K)$ and we analyze the algorithm's linear convergence under an anisotropic generalization of the proximal PL-inequality [21]. The method generalizes existing classical algorithms for optimal transport and boosting. This provides a dual view onto the framework of entropic subspace projections that is typically used to derive these algorithms. We also discuss examples which go beyond these existing methods.

A Missing proofs

A.1 Proof of Lemma 2.7

The proof is a slight refinement of the proof of [14, Theorem 3.1(ii)] and [43, Proposition 2.16].

Proof. Thanks to [8, Proposition 6.19(vii)] for $L = \operatorname{id} : \mathbb{R}^n \to \mathbb{R}^n$ the constraint qualification implies that

$$0 \in \operatorname{ri}(\operatorname{dom}\phi^* - \operatorname{dom} h^*). \tag{51}$$

"2.7(i)": Let $x \in \operatorname{dom}\phi = \mathbb{R}^n$. Invoking [14, Theorem 3.1(ii)] we obtain

$$x = \operatorname{aprox}_h^{\lambda\star\phi}(x) + \nabla(\lambda \star \phi)^*\big(\operatorname{bprox}_{h^*}^{(\lambda\star\phi)^*}(\nabla(\lambda \star \phi)(x))\big).$$

Note that $(\lambda \star \phi)^* = \lambda\phi^*$, $\nabla(\lambda \star \phi) = \nabla\phi(\lambda^{-1}(\cdot))$ and $\nabla(\lambda \star \phi)^* = \lambda\nabla\phi^*$. Thus we have

$$x = \operatorname{aprox}_h^{\lambda\star\phi}(x) + \lambda\nabla\phi^*\big(\operatorname{bprox}_{h^*}^{\lambda\phi^*}(\nabla\phi(\lambda^{-1}x))\big),$$

as claimed.

"2.7(ii)": By the constraint qualification (51) in view of [8, Proposition 15.7] we have that $h \boxplus \lambda \star \phi \in \Gamma_0(\mathbb{R}^n)$ with $\operatorname{dom}(h \boxplus \lambda \star \phi) = \operatorname{dom} h + \lambda \operatorname{dom} \phi = \operatorname{dom} h + \mathbb{R}^n = \mathbb{R}^n$ and the infimal convolution is exact. Invoking [8, Theorem 15.3] we have that $(h^* + \lambda\phi^*)^* = h \boxplus \lambda \star \phi$. By (i) $\operatorname{aprox}_h^{\lambda\star\phi}$ is single-valued and thanks to [8, Proposition 18.7] we have that $h \boxplus \lambda \star \phi$ is differentiable with $\nabla(h \boxplus \lambda \star \phi)(x) = \nabla(\lambda \star \phi)(x - \operatorname{aprox}_h^{\lambda\star\phi}(x)) = \nabla\phi(\lambda^{-1}(x - \operatorname{aprox}_h^{\lambda\star\phi}(x)))$ for any $x \in \mathbb{R}^n$. By [35, Corollary 25.5.1], since $h \boxplus \lambda \star \phi$ is finite-valued and convex the gradient map $\nabla(h \boxplus \lambda \star \phi) : \mathbb{R}^n \to \mathbb{R}^n$ is actually continuous.

"2.7(iii)": We have for any $x \in \mathbb{R}^n$ that

$$\begin{aligned}\operatorname{bprox}_{h^*}^{\lambda\phi^*}(\nabla\phi(\lambda^{-1}x)) &= \nabla\phi\big(\lambda^{-1}(x - \operatorname{aprox}_h^{\lambda\star\phi}(x))\big) = \nabla(h \boxplus \lambda \star \phi)(x) \\ &= (\partial(h^* + \lambda\phi^*))^{-1}(x) = (\partial h^* + \lambda\nabla\phi^*)^{-1}(x),\end{aligned}$$

where the first identity follows by 2.7(i), the second by 2.7(ii), the third by $h \boxplus \lambda \star \phi = (h^* + \lambda\phi^*)^*$ and convex conjugacy and the last by the constraint qualification and the sum-rule [36, Corollary 10.9] for the subdifferential: $\partial(h^* + \lambda\phi^*) = \partial h^* + \lambda\nabla\phi^*$.

"2.7(iv)": By Fermat's rule [36, Theorem 10.1] for the convex case $y \in \operatorname{aprox}_h^{\lambda\star\phi}(x) = \arg\min_{y\in\mathbb{R}^n} \lambda \star \phi(x - y) + h(y)$ is equivalent to $\nabla\phi(\lambda^{-1}(x - y)) \in \partial h(y)$. Equivalently this means $y \in (\mathrm{id} + \lambda\nabla\phi^*\partial h)^{-1}(x)$. □

A.2 Proof of Proposition 3.6

The proof is an adaption of [24, Proposition 3.9] for ϕ which are possibly not super-coercive.

Proof. Without loss of generality assume that $L = 1 = \lambda$ by redefining ϕ as $\lambda \star \phi$.

"3.6(i) ⇒ 3.6(ii)": Let $\bar{x} \in \mathbb{R}^n$. Then we have that for any x

$$-h(x) \geq -\phi(x - \bar{x} + \nabla\phi^*(\nabla h(\bar{x}))) + \phi(\nabla\phi^*(\nabla h(\bar{x}))) - h(\bar{x}),$$

which means by definition of the Φ-subdifferential that $\bar{x} - \nabla\phi^*(\nabla h(\bar{x})) \in \partial_\Phi(-h)(\bar{x})$ for $\Phi(x, y) := -\phi(x - y)$. Now take $\bar{y} \in \partial_\Phi(-h)(\bar{x})$. This means in view of Proposition 3.5 that $\bar{x} \in \arg\min_{x\in\mathbb{R}^n} \phi(x - \bar{y}) - h(x)$ and thus by Fermat's rule $0 = \nabla\phi(\bar{x} - \bar{y}) - \nabla h(\bar{x})$. This means that $\bar{y} = \bar{x} - \nabla\phi^*(\nabla h(\bar{x}))$.

"3.6(ii) ⇒ 3.6(i)": Assume that $\partial_\Phi(-h) = \mathrm{id} - \nabla\phi^* \circ \nabla h$. Let $\bar{x} \in \mathbb{R}^n$. Then $\bar{x} - \nabla\phi^*(\nabla h(\bar{x})) \in \partial_\Phi(-h)(\bar{x})$. By definition of the Φ-subdifferential this means the anisotropic descent inequality holds at $\bar{x}$. □

A.3 Proof of Proposition 3.7

The proof combines and adapts the proofs of [24, Proposition 3.10, Corollary 4.6 and Theorem 4.7] for ϕ which are possibly not super-coercive.

Proof. Without loss of generality we assume that $L = 1 = \lambda$ by replacing ϕ with $\lambda \star \phi$.

"3.7(i) ⇒ 3.7(ii)": By definition $h \in \mathcal{C}^1(\mathbb{R}^n)$. Let $\bar{x} \in \mathbb{R}^n$. In view of Proposition 3.6 we have that $\bar{x} - \nabla\phi^*(\nabla h(\bar{x})) \in \partial_\Phi(-h)(\bar{x})$ for $\Phi(x, y) := -\phi(x - y)$. Invoking Proposition 3.5 this

means that $(-h)(\bar{x}) = (-h)^{\Phi\Phi}(\bar{x})$. Since $\bar{x} \in \mathbb{R}^n$ was arbitrary we know by Proposition 3.3 that $-h$ is Φ-convex. By definition of the Φ-biconjugate this means that $h = -(-h)^{\Phi\Phi} = \xi \boxplus \phi$ for $\xi := (-h)^{\Phi}$.

"3.7(ii) $\Rightarrow$ 3.7(i)": Suppose that $h \in \mathcal{C}^1(\mathbb{R}^n)$ and $h = \xi \boxplus \phi$ for some $\xi : \mathbb{R}^n \to \overline{\mathbb{R}}$. Let $\bar{x} \in \mathbb{R}^n$. Since $h \in \mathcal{C}^1(\mathbb{R}^n)$ it is in particular finite-valued and we have $h(\bar{x}) = \inf_{y\in\mathbb{R}^n} \xi(y) + \phi(\bar{x} - y) \in \mathbb{R}$. This means that for any $\varepsilon > 0$ there exists $\bar{y}_\varepsilon$ such that

$$h(\bar{x}) + \varepsilon = \xi(\bar{y}_\varepsilon) + \phi(\bar{x} - \bar{y}_\varepsilon), \tag{52}$$

while for any $x \in \mathbb{R}^n$ we have that $h(x) \leq \xi(\bar{y}_\varepsilon) + \phi(x - \bar{y}_\varepsilon)$. Combining the identities and eliminating $\xi(\bar{y}_\varepsilon)$ we obtain that for any $x \in \mathbb{R}^n$:

$$\phi(x - \bar{y}_\varepsilon) - h(x) + \varepsilon \geq \phi(\bar{x} - \bar{y}_\varepsilon) - h(\bar{x}), \tag{53}$$

showing that $\bar{x} \in \varepsilon\text{-arg}\min\{\phi(\cdot - \bar{y}_\varepsilon) - h\}$. Ekeland's variational principle with $\delta = \sqrt{\varepsilon}$, see [36, Proposition 1.43], yields the existence of a point $\bar{x}_\varepsilon \in \overline{\mathrm{B}}(\bar{x}; \sqrt{\varepsilon})$ with $\phi(\bar{x}_\varepsilon - \bar{y}_\varepsilon) - h(\bar{x}_\varepsilon) \leq \phi(\bar{x} - \bar{y}_\varepsilon) - h(\bar{x})$ and $\bar{x}_\varepsilon \in \arg\min\{\phi(\cdot - \bar{y}_\varepsilon) - h + \sqrt{\varepsilon}\|\cdot - \bar{x}_\varepsilon\|\}$. By Fermat's rule [36, Theorem 10.1] and the fact that $\partial\|\cdot\|(0) = \overline{\mathrm{B}}(0;1)$ we have $0 \in \nabla\phi(\bar{x}_\varepsilon - \bar{y}_\varepsilon) - \nabla h(\bar{x}_\varepsilon) + \overline{\mathrm{B}}(0; \sqrt{\varepsilon})$ and thus

$$\nabla h(\bar{x}_\varepsilon) - \nabla\phi(\bar{x}_\varepsilon - \bar{y}_\varepsilon) =: u_\varepsilon \in \overline{\mathrm{B}}(0; \sqrt{\varepsilon}). \tag{54}$$

Since $\operatorname{rge}\nabla\phi = \operatorname{int}(\operatorname{dom}\phi^*)$ we have that $\nabla h(\bar{x}_\varepsilon) - u_\varepsilon = \nabla\phi(\bar{x}_\varepsilon - \bar{y}_\varepsilon) \in \operatorname{int}(\operatorname{dom}\phi^*)$. Passing to $\varepsilon \to 0$, by continuity, $\operatorname{int}(\operatorname{dom}\phi^*) \ni \nabla h(\bar{x}_\varepsilon) - u_\varepsilon \to \nabla h(\bar{x}) \in \operatorname{int}(\operatorname{dom}\phi^*)$, where the inclusion follows by the constraint qualification $\operatorname{rge}\nabla h \subseteq \operatorname{rge}\nabla\phi = \operatorname{int}(\operatorname{dom}\phi^*)$. Since $\nabla\phi^*$ is continuous relative to $\operatorname{int}(\operatorname{dom}\phi^*)$ we have by rewriting (54)

$$\bar{y}_\varepsilon = \bar{x}_\varepsilon - \nabla\phi^*(\nabla h(\bar{x}_\varepsilon) - u_\varepsilon) \to \bar{x} - \nabla\phi^*(\nabla h(\bar{x})) =: \bar{y},$$

for $\varepsilon \to 0$. Thus, passing to $\varepsilon \to 0$ in (53) we obtain again by continuity that

$$\phi(x - \bar{x} + \nabla\phi^*(\nabla h(\bar{x}))) - h(x) \geq \phi(\nabla\phi^*(\nabla h(\bar{x}))) - h(\bar{x}),$$

which is the anisotropic descent inequality.

Thanks to the previous part of the proof the function ξ in the expression $h = \xi \boxplus \phi$ can be taken as $\xi = (-h)^{\Phi}$ which is proper lsc as it is a pointwise supremum over continuous functions. Recall that $\bar{y}_\varepsilon \to \bar{x} - \nabla\phi^*(\nabla h(\bar{x})) = \bar{y}$ for $\varepsilon \to 0$. Then we have:

$$\inf_{y\in\mathbb{R}^n} \xi(y) + \phi(\bar{x} - y) \leq \xi(\bar{y}) + \phi(\bar{x} - \bar{y}) \leq \lim_{\varepsilon \searrow 0} \xi(\bar{y}_\varepsilon) + \phi(\bar{x} - \bar{y}_\varepsilon) = h(\bar{x}) = \inf_{y\in\mathbb{R}^n} \xi(y) + \phi(\bar{x} - y),$$

where the second inequality holds since ξ is proper lsc and the second last equality holds thanks to (52). This completes the reverse direction of the proof.

Finally, assume that $h \in \Gamma_0(\mathbb{R}^n)$ and $h = \xi \boxplus \phi$ for some $\xi : \mathbb{R}^n \to \overline{\mathbb{R}}$. Then by definition of the convex conjugate we have that $h^* = \xi^* + \phi^*$ and $\xi^* \in \Gamma_0(\mathbb{R}^n)$. By assumption we have $\operatorname{dom}\partial h^* = \operatorname{rge}\partial h \subseteq \operatorname{rge}\nabla\phi = \operatorname{int}\operatorname{dom}\phi^*$ implying that $\emptyset \neq \operatorname{dom} h^* \cap \operatorname{int}\operatorname{dom}\phi^*$. Since $\operatorname{dom} h^* = \operatorname{dom}\xi^* \cap \operatorname{dom}\phi^*$ we have $\emptyset \neq (\operatorname{dom}\xi^* \cap \operatorname{dom}\phi^*) \cap \operatorname{int}\operatorname{dom}\phi^* = \operatorname{dom}\xi^* \cap \operatorname{int}\operatorname{dom}\phi^*$. To show the opposite direction let $h^* = \psi + \phi^*$ for some $\psi \in \Gamma_0(\mathbb{R}^n)$ such that $\operatorname{dom}\psi \cap \operatorname{int}\operatorname{dom}\phi^* \neq \emptyset$. Then we can invoke [8, Proposition 6.19(vii)] and [8, Theorem 15.3] to obtain that $h = h^{**} = (\psi + \phi^*)^* = \psi^* \boxplus \phi \in \Gamma_0(\mathbb{R}^n)$. By Lemma 2.7(ii) $\psi^* \boxplus \phi \in \mathcal{C}^1(\mathbb{R}^n)$ with $\nabla(\psi^* \boxplus \phi) = \nabla\phi \circ (\mathrm{id} - \operatorname{aprox}^{\phi}_{\psi^*})$. This implies that $\operatorname{rge}\partial(\psi^* \boxplus \phi) \subseteq \operatorname{rge}\nabla\phi$. This verifies all claims in 3.7(ii) and concludes the proof. □

A.4 Proof of Example 4.8

Proof. To show the claimed result we first show that $f_i := \ell(\langle a_i, \cdot\rangle)$ is anisotropically smooth relative to ϕ with constant L. The claimed result then follows by verifying joint convexity of D_{ϕ^*} and invoking Lemma 4.5.

We have $h^*(t) = (t+1)\log((t+1)/2) + (1-t)\log((1-t)/2)$ with $\operatorname{dom} h^* = [-1,1]$ where $t\log t + (1-t)\log(1-t) = 0$ for $t \in \{0,1\}$. To prove the claimed result we invoke Proposition 3.7 and verify that $(\ell \circ a_i^\top)^* = \psi + \frac{1}{L}\phi^*$ for some $\psi \in \Gamma_0(\mathbb{R}^n)$ and $\operatorname{dom}\psi \cap \operatorname{int}\operatorname{dom}\phi^* \neq \emptyset$:

Owing to [36, Theorem 11.23(b)] the conjugate $(\ell \circ a_i^\top)^*$ amounts to the infimal postcomposition $a_i\ell^*$ of $\ell^*(t) = t\log t + (1-t)\log(1-t)$ by a_i with $\operatorname{dom}\ell^* = [0,1]$ and is given as follows:

$$(a_i\ell^*)(x) = \inf\{\ell^*(t) \mid t \in \mathbb{R} : a_i t = x\} = \begin{cases} h(t) & \text{if } a_i t = x \text{ for some } t \in [0,1], \\ +\infty & \text{otherwise.} \end{cases}$$

Then we have that $\operatorname{dom} a_i\ell^* = \{a_i t \mid t \in [0,1]\} \subseteq \operatorname{dom}\phi^*$ and $\frac{1}{2}a_i \in \operatorname{int}\operatorname{dom}\phi^* \cap \operatorname{ri}(\operatorname{dom} a_i\ell^*)$. Thus we can invoke Lemma 4.7 and verify convexity of $a_i\ell^* \mathbin{\dot{-}} \frac{1}{L}\phi^*$ on the relative interior of $\operatorname{dom} a_i\ell^*$: Equivalently this means that the one-dimensional function $\ell^* \mathbin{\dot{-}} \frac{1}{L}\phi^* \circ a_i$ is convex on $(0,1)$. Since $(h^*)''(t) = 2/(1-t^2)$ and $(\ell^*)''(t) = \frac{1}{t-t^2} > 0$ on $(0,1)$ this is implied by

$$L \geq \frac{a_i^\top \nabla^2\phi^*(a_i t)a_i}{(\ell^*)''(t)} = a_i^\top \operatorname{Diag} 2(t-t^2)/(1-a_i^2t^2)a_i = \sum_{j=1}^m \frac{2(t-t^2)a_{ij}^2}{1-a_{ij}^2t^2},$$

for all $t \in (0,1)$. Clearly, $\frac{2(t-t^2)a_{ij}^2}{1-a_{ij}^2t^2} \leq \frac{2(t-t^2)a_{ij}^2}{1-t^2}$ and $\frac{2(t-t^2)a_{ij}^2}{1-t^2}$ is monotonically increasing on $[0,1)$. For $t \to 1$ the expression is indefinite. Thus we apply L'Hospital's rule and obtain:

$$\lim_{t\to 1}\frac{2(t-t^2)a_{ij}^2}{1-t^2} = \lim_{t\to 1}\frac{2a_{ij}^2(1-2t)}{-2t} = \lim_{t\to 1}\frac{(2t-1)a_{ij}^2}{t} = a_{ij}^2.$$

and thus we have for $t \in (0,1)$

$$\frac{(\phi^* \circ a_i)''(t)}{(h^*)''(t)} = \sum_{j=1}^m \frac{2(t-t^2)a_{ij}^2}{1-a_{ij}^2t^2} \leq \sum_{j=1}^m \frac{2(t-t^2)a_{ij}^2}{1-t^2} \leq \sum_{j=1}^m a_{ij}^2 = \|a_i\|^2 =: L_i.$$

As a consequence f_i is anisotropically smooth relative to $\frac{1}{L_i}\star\phi$. Since $1/(h^*)''(t) = (1-t^2)/2$ is concave by [4, Theorem 3.3(i)] D_{ϕ^*} is jointly convex. Then we can invoke Lemma 4.5 recursively as in Remark 4.6 to show that $f = \frac{1}{m}\sum_{i=1}^m f_i$ is anisotropically smooth relative to $\frac{1}{L}\star\phi$ with $L = \max_{1\leq i\leq m}\|a_i\|^2$. □

A.5 Proof of Example 4.10

Proof. We show that $f_i = \sigma \star \exp(\langle a_i, x\rangle - b_i)$ is anisotropically smooth relative to ϕ with constant L and then invoke Lemma 4.9. To prove the claimed result we invoke Proposition 3.7 and verify that $(\sigma\star\exp(\cdot - b_i)\circ a_i^\top)^* = \psi + \frac{1}{L}\phi^*$ for some $\psi \in \Gamma_0(\mathbb{R}^n)$ and $\emptyset \neq \operatorname{dom}\psi \cap \operatorname{int}\operatorname{dom}\phi^*$:

Owing to [36, Theorem 11.23(b)] the conjugate $(\sigma \star \exp(\cdot - b_i) \circ a_i^\top)^*$ amounts to the infimal postcomposition $a_i\ell$ of $\ell(t) = \sigma h(t) + b_i t$ by a_i with h being the von Neumann entropy and is given as follows:

$$(a_i\ell)(x) = \inf\{\ell(t) \mid t \in \mathbb{R} : a_i t = x\} = \begin{cases} \ell(t) & \text{if } a_i t = x \text{ for some } t \in \mathbb{R}_+, \\ +\infty & \text{otherwise.} \end{cases}$$

We have $\operatorname{dom} a_i\ell = \{a_i t \mid t \in \mathbb{R}_+\} \subseteq \operatorname{dom} H$ and since $a_i \in \mathbb{R}^n_{++}$ we have $\operatorname{ri}\operatorname{dom}(a_i\ell) \cap \operatorname{int}\operatorname{dom}\phi^* \neq \emptyset$. Therefore we can invoke Lemma 4.7 to show convexity of $a_i\ell \mathbin{\dot{-}} \frac{1}{L}H$ on $\operatorname{ri}\operatorname{dom}(a_i\ell)$. Equivalently this means that the one-dimensional function $\ell \mathbin{\dot{-}} \frac{1}{L}H \circ a_i$ is convex on $\mathbb{R}_{++}$. This is implied by

$$\ell''(t) - \frac{1}{L}a_i^\top \nabla^2 H(a_i t)a_i = \sigma h''(t) - \frac{1}{L}a_i^\top \nabla^2 H(a_i t)a_i = \frac{\sigma}{t} - \frac{1}{L}\|a_i\|_1 \frac{1}{t} = \frac{1}{t}(\sigma - \|a_i\|_1/L) \geq 0,$$

for all $t > 0$. This is guaranteed by the choice of $L = \max_{1\leq i\leq m} \|a_i\|_1/\sigma$. Applying Lemma 4.9 recursively we obtain the claimed result. □

A.6 Proof of Proposition 5.11

The proof is an adaption of the proof of [24, Theorem 4.3] for ϕ which are possibly not super-coercive.

Proof. Without loss of generality assume that $\mu = 1$ by replacing ϕ with $\mu^{-1} \star \phi$. Choose $\Phi(x, y) = +\mu \star \phi(x - y)$.

"5.11(i) ⇒ 5.11(ii)": Let $\bar{v} \in \operatorname{int}\operatorname{dom}\phi^* = \operatorname{rge}\nabla\phi$. Thanks to the constraint qualification we have $\bar{v} \in \operatorname{rge}\partial h$. Thus there exists $\bar{x} \in \operatorname{dom}\partial h$ such that $\bar{v} \in \partial h(\bar{x})$ and the subgradient inequality (36) holds true for $(\bar{x}, \bar{v})$. Then the anisotropic subgradient inequality means by definition that $\bar{y} := \bar{x} - \nabla\phi^*(\bar{v}) \in \partial_\Phi h(\bar{x})$. Invoking Proposition 3.5 we have $h(\bar{x}) + h^\Phi(\bar{y}) = \phi(\bar{x} - \bar{y})$ and $\bar{x} \in \partial_\Phi h^\Phi(\bar{y})$ where the latter means by definition that $\bar{y} \in \arg\max \phi(\bar{x} - \cdot) - h^\Phi$. Combined these yield

$$\sup_{y\in\mathbb{R}^n} \phi(\bar{x} - y) - h^\Phi(y) = \phi(\bar{x} - \bar{y}) - h^\Phi(\bar{y}) = h(\bar{x}). \tag{55}$$

Fenchel duality yields $\phi(\bar{x} - y) = \phi^{**}(\bar{x} - y) = \sup_{v\in\mathbb{R}^n}\langle \bar{x} - y, v\rangle - \phi^*(v)$. Define $q(y) := \sup_{v\in\mathbb{R}^n} \xi(y, v)$ for $\xi(y, v) := \langle \bar{x} - y, v\rangle - \phi^*(v) - h^\Phi(y)$. Then we can rewrite the supremum in (55) in terms of the joint supremum $h(\bar{x}) = \sup_{y\in\mathbb{R}^n} q(y) = q(\bar{y})$ and in particular $\bar{y} \in \arg\max q$. We have $\bar{y} = \bar{x} - \nabla\phi^*(\bar{v})$ or equivalently $\bar{v} = \nabla\phi(\bar{x} - \bar{y})$ which via convex conjugacy implies that $\bar{v} \in \arg\max\langle \bar{x} - \bar{y}, \cdot\rangle - \phi^*$. This shows that $\bar{v} \in \arg\max \xi(\bar{y}, \cdot)$. Define $p(v) = \sup_{y\in\mathbb{R}^n} \xi(y, v)$. Then [36, Proposition 1.35] yields that $(\bar{y}, \bar{v}) \in \arg\max \xi$ as well as $\bar{v} \in \arg\max p$. Overall this yields

$$\begin{aligned} h(\bar{x}) &= \sup_{y\in\mathbb{R}^n} q(y) = \sup_{v\in\mathbb{R}^n} p(v) = p(\bar{v}) \\ &= \langle \bar{x}, \bar{v}\rangle - \phi^*(\bar{v}) + \sup_{y\in\mathbb{R}^n} \langle y, -\bar{v}\rangle - h^\Phi(y) = \langle \bar{x}, \bar{v}\rangle - \phi^*(\bar{v}) + (h^\Phi)^*(-\bar{v}), \end{aligned}$$

which combined with the fact that $\bar{v} \in \partial h(\bar{x})$ results in

$$h^*(\bar{v}) = \langle \bar{x}, \bar{v}\rangle - h(\bar{x}) = \phi^*(\bar{v}) - (h^\Phi)^*_-(\bar{v}) \in \mathbb{R}.$$

Since $\bar{v} \in \operatorname{int}\operatorname{dom}\phi^*$ was arbitrary we have $\phi^* \mathbin{\dot{-}} h^* \equiv (h^\Phi)^*_-$ on $\operatorname{int}\operatorname{dom}\phi^*$ and thus $\phi^* \mathbin{\dot{-}} h^*$ is convex on $\operatorname{int}\operatorname{dom}\phi^*$.

"5.11(ii) ⇒ 5.11(i)": Thanks to the constraint qualification we have $\operatorname{dom} h^* \supseteq \operatorname{dom}\partial h^* = \operatorname{rge}\partial h \supseteq \operatorname{int}\operatorname{dom}\phi^*$ and thus $\phi^* + (-h^*)$ is finite-valued and convex on $\operatorname{int}\operatorname{dom}\phi^*$. This implies that $-h^*$ is finite-valued on $\operatorname{int}\operatorname{dom}\phi^*$. Since ϕ^* is smooth on $\operatorname{int}\operatorname{dom}\phi^*$ we can invoke [36, Exercise 8.20(b)] to show that $-h^*$ is regular on $\operatorname{int}\operatorname{dom}\phi^*$. Invoking [36, Exercise 8.8(c)]

we obtain that $\partial(\phi^* + (-h^*)) = \widehat{\partial}(\phi^* + (-h^*)) = \nabla\phi^* + \widehat{\partial}(-h^*)$ on $\operatorname{int}\operatorname{dom}\phi$. Since, in addition, $h^* \in \Gamma_0(\mathbb{R}^n)$, both, $\pm h^*$ are regular with $\widehat{\partial}(\pm h^*)$ nonempty on $\operatorname{int}\operatorname{dom}\phi^*$. Invoking [36, Theorem 9.18] this means that h^* is smooth on $\operatorname{int}\operatorname{dom}\phi^*$ and as such so is $g := \phi^* \dot{-} h^*$. Fix $(\bar{x}, \bar{v}) \in \operatorname{gph}\partial h \cap (\mathbb{R}^n \times \operatorname{int}\operatorname{dom}\phi^*)$ and let $\bar{y} := \bar{x} - \nabla\phi^*(\bar{v})$. Then $\nabla\phi(\bar{x} - \bar{y}) = \bar{v} \in \partial h(\bar{x})$ and since $h \in \Gamma_0(\mathbb{R}^n)$ using convex conjugacy we have $\bar{x} = \nabla h^*(\nabla\phi(\bar{x} - \bar{y}))$. Since $h^* \equiv \phi^* \dot{-} g$ on $\operatorname{dom}\phi^*$ and $\operatorname{dom}\phi^* = \operatorname{dom}(\phi^* \dot{-} g)$ we have

$$
\begin{aligned}
h(x) = h^{**}(x) &= \sup_{v\in\mathbb{R}^n} \langle v, x\rangle - h^*(v) \\
&\geq \sup_{v\in\operatorname{dom}\phi^*} \langle v, x\rangle - h^*(v) = \sup_{v\in\mathbb{R}^n} \langle v, x\rangle - (\phi^* \dot{-} g)(v) = (\phi^* \dot{-} g)^*(x) && (56)\\
&= \sup_{y\in\mathbb{R}^n} \phi(x - y) \dot{-} g_-^*(y) && (57)\\
&\geq \phi(x - \bar{y}) \dot{-} g_-^*(\bar{y}), && (58)
\end{aligned}
$$

where (57) follows by Hiriart-Urruty [19, Theorem 2.2]. Smoothness of h^* and g on $\operatorname{int}\operatorname{dom}\phi^*$ yields:

$$
\bar{x} = \nabla h^*(\nabla\phi(\bar{x} - \bar{y})) = \nabla(\phi^* - g)(\nabla\phi(\bar{x} - \bar{y})) = \bar{x} - \bar{y} - \nabla g(\nabla\phi(\bar{x} - \bar{y})),
$$

and therefore $\nabla g(\nabla\phi(\bar{x} - \bar{y})) = -\bar{y}$. Recall that $\nabla\phi(\bar{x} - \bar{y}) = \bar{v} \in \partial h(\bar{x})$. Then we have thanks to convex conjugacy

$$
\begin{aligned}
h(\bar{x}) &= \langle \nabla\phi(\bar{x} - \bar{y}), \bar{x}\rangle - h^*(\nabla\phi(\bar{x} - \bar{y})) = \langle \nabla\phi(\bar{x} - \bar{y}), \bar{x}\rangle - \phi^*(\nabla\phi(\bar{x} - \bar{y})) + g(\nabla\phi(\bar{x} - \bar{y}))\\
&= \langle \nabla\phi(\bar{x} - \bar{y}), \bar{x} - \bar{y}\rangle - \phi^*(\nabla\phi(\bar{x} - \bar{y})) - \big(\langle \nabla\phi(\bar{x} - \bar{y}), -\bar{y}\rangle - g(\nabla\phi(\bar{x} - \bar{y}))\big)\\
&= \phi(\bar{x} - \bar{y}) - g_-^*(\bar{y}).
\end{aligned}
$$

By combining this result with (58) we obtain

$$
h(x) \geq h(\bar{x}) + \phi(x - \bar{y}) - \phi(\bar{x} - \bar{y}) = h(\bar{x}) + \phi(x - \bar{x} + \nabla\phi^*(\bar{v})) - \phi(\nabla\phi^*(\bar{v})). \qquad \square
$$

References

[1] E. J. Balder. "An extension of duality-stability relations to nonconvex optimization problems". In: *SIAM Journal on Control and Optimization* 15.2 (1977), pp. 329–343.

[2] H. Bauermeister, E. Laude, T. Möllenhoff, M. Moeller, and D. Cremers. "Lifting the convex conjugate in Lagrangian relaxations: a tractable approach for continuous Markov random fields". In: *SIAM Journal on Imaging Sciences* 15.3 (2022), pp. 1253–1281.

[3] H. H. Bauschke and P. L. Combettes. "Iterating Bregman Retractions". In: *SIAM Journal on Optimization* 13.4 (2003), pp. 1159–1173. ISSN: 1052-6234.

[4] H. H. Bauschke and J. M. Borwein. "Joint and separate convexity of the Bregman distance". In: *Studies in Computational Mathematics*. Vol. 8. Elsevier, 2001, pp. 23–36.

[5] H. H. Bauschke, J. Bolte, J. Chen, M. Teboulle, and X. Wang. "On Linear Convergence of Non-Euclidean Gradient Methods without Strong Convexity and Lipschitz Gradient Continuity". In: *Journal of Optimization Theory and Applications* 182.3 (2019), pp. 1068–1087.

[6] H. H. Bauschke, J. Bolte, and M. Teboulle. "A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications". In: *Mathematics of Operations Research* 42.2 (2017), pp. 330–348.

[7] H. H. Bauschke and J. M. Borwein. "Legendre functions and the method of random Bregman projections". In: *Journal of Convex Analysis* 4.1 (1997), pp. 27–67.

[8] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Science & Business Media, 2011.

[9] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. "Iterative Bregman projections for regularized transportation problems". In: *SIAM Journal on Scientific Computing* 37.2 (2015), A1111–A1138.

[10] B. Birnbaum, N. R. Devanur, and L. Xiao. "Distributed algorithms via gradient descent for Fisher markets". In: *Proceedings of the 12th ACM conference on Electronic commerce*. 2011, pp. 127–136.

[11] L. M. Briceño-Arias, G. Chierchia, E. Chouzenoux, and J.-C. Pesquet. "A random block-coordinate Douglas–Rachford splitting method with low computational complexity for binary logistic regression". In: *Computational Optimization and Applications* 72.3 (2019), pp. 707–726.

[12] J. V. Burke and T. Hoheisel. "Epi-convergent smoothing with applications to convex composite functions". In: *SIAM Journal on Optimization* 23.3 (2013), pp. 1457–1479.

[13] M. Collins, R. E. Schapire, and Y. Singer. "Logistic regression, AdaBoost and Bregman distances". In: *Machine Learning* 48.1 (2002), pp. 253–285.

[14] P. L. Combettes and N. N. Reyes. "Moreau's decomposition in Banach spaces". In: *Mathematical Programming* 139.1-2 (2013), pp. 103–114.

[15] M. Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in neural information processing systems* 26 (2013).

[16] S. Della Pietra, V. Della Pietra, and J. Lafferty. *Duality and auxiliary functions for Bregman distances*. Tech. rep. CMU-CS-01-109. School of Computer Science, Carnegie Mellon University, 2001.

[17] S. Dolecki and S. Kurcyusz. "On Φ-Convexity in Extremal Problems". In: *SIAM Journal on Control and Optimization* 16.2 (1978), pp. 277–300.

[18] Y. Freund and R. E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139.

[19] J.-B. Hiriart-Urruty. "A general formula on the conjugate of the difference of functions". In: *Canadian Mathematical Bulletin* 29.4 (1986), pp. 482–485.

[20] J.-B. Hiriart-Urruty and M.-L. Mazure. "Formulations variationnelles de l'addition parallele et de la soustraction parallelle d'opérateurs semi-définis positifs". In: *CR Acad. Sci. Paris Sér. I Math* 302 (1986), pp. 527–530.

[21] H. Karimi, J. Nutini, and M. Schmidt. "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition". In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer. 2016, pp. 795–811.

[22] T. Larsson and M. Patriksson. “A class of gap functions for variational inequalities”. In: *Mathematical Programming* 64.1 (1994), pp. 53–79.

[23] E. Laude. “Lower envelopes and lifting for structured nonconvex optimization”. PhD thesis. Technical University of Munich, 2021.

[24] E. Laude, A. Themelis, and P. Patrinos. “Conjugate dualities for relative smoothness and strong convexity under the light of generalized convexity”. In: *arXiv preprint arXiv:2112.08886* (2021).

[25] E. Laude, T. Wu, and D. Cremers. “Optimization of Inf-Convolution Regularized Nonconvex Composite Problems”. In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 547–556.

[26] H. Lu, R. M. Freund, and Y. Nesterov. “Relatively smooth convex optimization by first-order methods, and applications”. In: *SIAM Journal on Optimization* 28.1 (2018), pp. 333–354.

[27] C. J. Maddison, D. Paulin, Y. W. Teh, and A. Doucet. “Dual space preconditioning for gradient descent”. In: *SIAM Journal on Optimization* 31.1 (2021), pp. 991–1016.

[28] A. Maignan and T. C. Scott. “Fleshing out the generalized Lambert W function”. In: *ACM Communications in Computer Algebra* 50.2 (2016), pp. 45–60.

[29] I. Mező and Á. Baricz. “On the generalization of the Lambert W function”. In: *Transactions of the American Mathematical Society* 369.11 (2017), pp. 7917–7934.

[30] J.-J. Moreau. “Fonctionnelles convexes”. In: *Séminaire Jean Leray* (1966), pp. 1–108.

[31] J.-J. Moreau. “Inf-convolution, sous-additivité, convexité des fonctions numériques”. In: *Journal de Mathématiques Pures et Appliquées* (1970), pp. 33–41.

[32] P. Patrinos and A. Bemporad. “Proximal Newton methods for convex composite optimization”. In: *52nd IEEE Conference on Decision and Control*. IEEE. 2013, pp. 2358–2363.

[33] J.-P. Penot and M. Volle. “On strongly convex and paraconvex dualities”. In: *Generalized convexity and fractional programming with economic applications*. Springer, 1990, pp. 198–218.

[34] R. T. Rockafellar. “Augmented Lagrange multiplier functions and duality in nonconvex programming”. In: *SIAM Journal on Optimization* 12.2 (1974), pp. 268–285.

[35] R. T. Rockafellar. *Convex Analysis*. New Jersey: Princeton University Press, 1970.

[36] R. T. Rockafellar and R. J. Wets. *Variational Analysis*. New York: Springer, 1998.

[37] R. E. Schapire and Y. Singer. “Improved boosting algorithms using confidence-rated predictions”. In: *Proceedings of the eleventh annual conference on Computational learning theory*. 1998, pp. 80–91.

[38] L. Stella, A. Themelis, and P. Patrinos. “Forward–backward quasi-Newton methods for nonsmooth optimization problems”. In: *Computational Optimization and Applications* 67.3 (2017), pp. 443–487.

[39] P. D. Tao and L. H. An. “Convex analysis approach to DC programming: theory, algorithms and applications”. In: *Acta mathematica vietnamica* 22.1 (1997), pp. 289–355.

[40] M. Teboulle. "Entropic Proximal Mappings with Applications to Nonlinear Programming". In: *Mathematics of Operations Research* 17.3 (1992), pp. 670–690.

[41] A. Themelis, B. Hermans, and P. Patrinos. "A new envelope function for nonsmooth DC optimization". In: *2020 59th IEEE Conference on Decision and Control (CDC)*. 2020, pp. 4697–4702.

[42] C. Villani. *Optimal Transport: Old and New*. Springer, 2008.

[43] X. Wang and H. H. Bauschke. "The Bregman proximal average". In: *SIAM Journal on Optimization* 32.2 (2022), pp. 1379–1401.