

A unified framework for Bregman proximal methods: subgradient, gradient, and accelerated gradient schemes

David H. Gutman* Javier F. Peña†

August 13, 2019

Abstract

We provide a unified framework for analyzing the convergence of Bregman proximal first-order algorithms for convex minimization. Our framework hinges on properties of the convex conjugate and gives novel proofs of the convergence rates of the Bregman proximal subgradient, Bregman proximal gradient, and a new accelerated Bregman proximal gradient algorithm under fairly general and mild assumptions. Our accelerated Bregman proximal gradient algorithm attains the best-known accelerated rate of convergence when suitable relative smoothness and triangle scaling assumptions hold. However, the algorithm requires no prior knowledge of any related smoothness or triangle scaling constants.

1 Introduction

Let \mathbb{E} be a finite dimensional real vector space and $\phi = f + \Psi$ where $f : \mathbb{E} \rightarrow \mathbb{R} \cup \{\infty\}$ and $\Psi : \mathbb{E} \rightarrow \mathbb{R} \cup \{\infty\}$ are closed convex functions. Consider the convex minimization problem

$$\min_{x \in \mathbb{E}} \phi(x). \tag{1}$$

A variety of popular algorithmic approaches for solving (1) are based on the following *proximal map*

$$g \mapsto \arg \min_{y \in \mathbb{E}} \left\{ \langle g, y \rangle + \Psi(y) + \frac{L}{2} \|y - x\|_2^2 \right\}. \tag{2}$$

For example, the proximal gradient method, also known as the forward-backward splitting method [20], generates a sequence $x_k \in \text{dom}(\phi)$, $k = 0, 1, \dots$ via

$$x_{k+1} := \arg \min_{y \in \mathbb{E}} \left\{ \langle \nabla f(x_k), y \rangle + \Psi(y) + \frac{L_k}{2} \|y - x_k\|_2^2 \right\}.$$

*Department of Mathematical Sciences, Carnegie Mellon University, USA, dgutman@andrew.cmu.edu

†Tepper School of Business, Carnegie Mellon University, USA, jfp@andrew.cmu.edu

The focus of this paper is a more general and flexible class of *Bregman proximal first-order methods* based on the *Bregman proximal map*

$$g \mapsto \arg \min_{y \in \mathbb{E}} \{ \langle g, y \rangle + \Psi(y) + LD_h(y, x) \} \quad (3)$$

where $D_h(y, x) := h(y) - h(x) - \langle \nabla h(x), y - x \rangle$ is the Bregman distance [12] generated by some *reference* convex function $h : \mathbb{E} \rightarrow \mathbb{R} \cup \{\infty\}$. The popular class of Euclidean proximal methods based on (2) corresponds to the special case when the reference function is the squared Euclidean norm $h(x) := \frac{1}{2} \|x\|_2^2$. Naturally, Bregman proximal methods rely on the critical assumption that the problem (3) is well-posed and has a computable solution.

The mirror descent method [6, 21, 24] is a well-known instance of a Bregman proximal first-order method when $\Psi = \delta_C$, the indicator function of C , for some closed convex set $C \subseteq \mathbb{E}$. Some more recent instances of Bregman proximal methods include the *NoLips* algorithm introduced by Bauschke, Bolte, and Teboulle [3], which follows a *Bregman proximal gradient* template [2, 8, 32, 33]. This same algorithmic template underlies the *relative gradient scheme* proposed by Lu, Freund, and Nesterov [22]. Both [3] and [22] establish convergence results for the Bregman proximal gradient method by relying on a *Lipschitz-like convexity condition* (LC) as defined in [3] or the equivalent *relative smoothness* condition as defined in [22]. Furthermore, the articles [22] and [32] give stronger linear convergence results for the Bregman proximal gradient method under the additional condition of *relative strong convexity*. Other instances of Bregman proximal methods are the *Bregman proximal subgradient* method [9, 10, 15, 32] and the accelerated Bregman proximal gradient methods for relative smooth functions very recently developed by Hanzely, Richtarik, and Xiao [17]. Bregman proximal methods are a special class of the broader class of *proximal methods* which in turn includes a wide range of algorithms for optimization, see [2, 4, 13, 32, 34] and the many references therein.

1.1 Main contribution: a convex conjugate framework

The central contribution of this paper is a framework to analyze the convergence of Bregman proximal first-order algorithms. Our framework hinges on the *convex conjugate* and can be seen as a natural extension of the approach that we introduced in [16, 29], which was restricted to the Euclidean setting. We rely on standard convex analysis notation and results as presented in [5, 11, 18, 30]. Recall that the convex conjugate of a convex function $F : \mathbb{E} \rightarrow \mathbb{R} \cup \{\infty\}$ is the function $F^* : \mathbb{E}^* \rightarrow \mathbb{R} \cup \{\infty\}$ defined via

$$F^*(u) := \inf_{x \in \mathbb{E}} \{ \langle u, x \rangle - F(x) \}.$$

By construction the convex conjugate function F^* is convex and satisfies the following *Fenchel inequality*: For all $x \in \mathbb{E}$, $u \in \mathbb{E}^*$ we have $F(x) + F^*(u) \geq \langle u, x \rangle$ and $F(x) + F^*(u) = \langle u, x \rangle$ if and only if $u \in \partial F(x)$.

Our convex conjugate framework automatically yields new derivations of convergence rates for the Bregman proximal subgradient method and for the Bregman proximal gradient method (Section 2 and Section 3). In addition, and perhaps most interesting, our convex conjugate framework also applies to a new accelerated Bregman proximal gradient method

(Section 4). The gist of our convex conjugate approach can be summarized as follows. Suppose $y, z \in \text{dom}(\phi)$ and a convex distance function $d : \mathbb{E} \rightarrow \mathbb{R} \cup \{\infty\}$ is differentiable at z and satisfies

$$\phi(y) \leq -\phi^*(-\nabla d(z)) - \langle \nabla d(z), z \rangle + D(z) = -\phi^*(-\nabla d(z)) - d^*(\nabla d(z)). \quad (4)$$

From (4) it immediately follows that $\phi(y) - \phi(x) \leq d(x)$ for all $x \in \text{dom}(\phi)$ since (4) implies

$$\phi(y) \leq \inf_{w \in \mathbb{E}} \{\phi(w) + \langle \nabla D(z), w \rangle\} + \inf_{w \in \mathbb{E}} \{d(w) - \langle \nabla d(z), w \rangle\} \leq \phi(x) + d(x).$$

In the main sections of the paper we show that three classes of Bregman proximal methods (subgradient, gradient, accelerated gradient) generate sequences $x_k, z_k \in \text{dom}(\phi) \cap \text{ri}(\text{dom}(h))$, $k = 0, 1, 2, \dots$ such that (4), or a slight modification of it, holds for $y = x_k$, $z = z_k$, and $d(\cdot) = C_k D_h(\cdot, x_0)$ for some nondecreasing sequence $C_k \in \mathbb{R}_+$, $k = 0, 1, 2, \dots$. More precisely, Theorem 3 shows that (4) holds for the accelerated Bregman proximal gradient method iterates, see (14). Theorem 2 shows that an inequality stronger than (4) holds for the Bregman proximal gradient method iterates, see (11). Theorem 1 shows that a slight variation of (4) holds for the Bregman proximal subgradient method iterates, see (10). In particular, for the Bregman proximal gradient and accelerated Bregman proximal gradient methods Theorem 2 and Theorem 3 yield

$$\phi(x_k) - \phi(x) \leq C_k D_h(x, x_0)$$

for all $x \in \text{dom}(\phi)$. We also get a similar inequality for the Bregman proximal subgradient method. In each case it will be easy to see that the sequence C_k , $k = 0, 1, \dots$ goes to zero under fairly mild and general assumptions. In particular, we show that the sequence C_k is as follows under suitable assumptions on f, ϕ , and h :

- For the Bregman proximal subgradient method $C_k = \mathcal{O}(1/\sqrt{k})$ if the pair (ϕ, h) satisfies the $W[\phi, h]$ boundedness condition as defined in [32]. See Corollary 2.
- For the Bregman proximal gradient method $C_k = \mathcal{O}(1/k)$ if f is L -smooth relative to h for some constant $L > 0$ as defined by [3, 22]. See Corollary 3.
- For the accelerated Bregman proximal gradient method $C_k = \mathcal{O}(1/k^\gamma)$ if f is (L, γ) -smooth relative to h for some constants $L > 0$ and $\gamma > 0$. (See (25).)

The above results yield new derivations of known convergence rates via our convex conjugate approach. However, our main results, namely Theorem 1, Theorem 2, and Theorem 3 hold more broadly. In particular, Theorem 1 only requires the Bregman steps to be *admissible* as defined below. Theorem 2 and Theorem 3 only require the Bregman steps to be admissible and to satisfy a suitable *decrease condition*. None of these three main results requires any further assumptions like Lipschitz continuity or relative smoothness.

1.2 Technical assumptions

We aim to present our developments in as much generality as possible. To that end, throughout the paper we make the blanket Assumption 1 below. We should note that the *admissibility condition* (A.3) is primarily a technicality. This condition is concerned with the choice

of $L > 0$ that guarantees the well-posedness of problem (3). As Example 1(b,c) below illustrates, in many cases problem (3) is readily well-posed and thus the admissibility condition (A.3.i) automatically holds for all $L > 0, g \in \mathbb{E}^*, x \in \text{ri}(\text{dom}(h))$. However, Example 1(a) also illustrates that in some cases the well-posedness of problem (3) may require a more careful choice of $L > 0$.

Assumption 1. The functions $f : \mathbb{E} \rightarrow \mathbb{R} \cup \{\infty\}$, $\Psi : \mathbb{E} \rightarrow \mathbb{R} \cup \{\infty\}$, and $h : \mathbb{E} \rightarrow \mathbb{R} \cup \{\infty\}$ satisfy the following conditions.

- (A.1) The functions f and Ψ are closed and convex. Throughout the sequel, we let $\phi := f + \Psi$.
- (A.2) The reference function h is convex and differentiable on $\text{ri}(\text{dom}(h))$ and satisfies $\text{dom}(\Psi) \subseteq \overline{\text{dom}(h)}$ and $\emptyset \neq \text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi) \subseteq \text{ri}(\text{dom}(f))$.
- (A.3) The pair of functions (h, Ψ) satisfies the following *admissibility conditions*:

- (i) For all $g \in \mathbb{E}^*$ and $x \in \text{ri}(\text{dom}(h))$ there exists $L > 0$ such that the Bregman proximal map (3) has a unique solution in $\text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$. When this holds we shall say that L is *admissible* for g at x .
- (ii) There is an oracle that takes as input $g \in \mathbb{E}^*, x \in \text{ri}(\text{dom}(h)), L > 0$ and yields as output either a certificate that L is not admissible for g at x or the unique solution to (3) in $\text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$. Observe that in the latter case the solution to (3) is the unique point $y \in \text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$ that satisfies the optimality conditions

$$g + g^\Psi + L(\nabla h(y) - \nabla h(x)) = 0, \quad g^\Psi \in \partial\Psi(y).$$

Observe that a constraint of the form $x \in C$ for a closed convex set $C \subseteq \mathbb{E}$ can be easily incorporated in the above setting by adding the indicator function δ_C to Ψ . The admissibility condition (A.3.i) can be ensured under suitable assumptions on Ψ and h . In particular, as detailed in [3,32], condition (A.3.i) holds when h is a Legendre function [30] and Ψ is bounded below and satisfies $\text{ri}(\text{dom}(\Psi)) \subseteq \text{ri}(\text{dom}(h))$, see [32, Lemma 2.3]. Furthermore, in concrete applications it is often easy to verify directly the admissibility conditions (A.3.i) and (A.3.ii) as Example 1 shows. For simplicity, Example 1 assumes that $\Psi = 0$. The admissibility properties in Example 1 can be extended to popular choices of regularization functions Ψ such as $\Psi(x) = \lambda\|x\|_2^2/2$ or $\Psi(x) = \lambda\|x\|_1$ for $\lambda > 0$. They can also be extended to popular choices of indicator functions such as $\Psi = \delta_{\Delta_{n-1}}$ for $\Delta_{n-1} := \{x \in \mathbb{R}_+^n : \|x\|_1 = 1\}$.

Example 1. Suppose $\Psi = 0$. The admissibility conditions (A.3.i) and (A.3.ii) hold for the following reference functions $h : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$.

- (a) The Burg entropy function $h(x) := -\sum_{i=1}^n \log(x_i)$. In this case $L > 0$ is admissible for $g \in \mathbb{R}^n$ at $x \in \mathbb{R}_{++}^n = \text{ri}(\text{dom}(h))$ if and only if $-\nabla h(x) + g/L \in \mathbb{R}_{++}^n$ and in this case the solution to (3) is the vector $y \in \mathbb{R}_{++}^n$ defined componentwise as $y_i = 1/(1/x_i + g_i/L)$, $i = 1, \dots, n$.
- (b) The Boltzmann-Shannon entropy function $h(x) := \sum_{i=1}^n x_i \log(x_i)$. In this case any $L > 0$ is admissible for any $g \in \mathbb{R}^n$ at any $x \in \mathbb{R}_{++}^n = \text{ri}(\text{dom}(h))$ and the solution to (3) is the vector $y \in \mathbb{R}_+^n$ defined componentwise as $y_i = e^{\log(x_i) - g_i/L}$, $i = 1, \dots, n$.

- (c) The squared Euclidean function $h(x) := \|x\|_2^2/2$. In this case any $L > 0$ is admissible for any $g \in \mathbb{R}^n$ at any $x \in \mathbb{R}^n = \text{ri}(\text{dom}(h))$ and the solution to (3) is the vector $y = x - g/L$.

To sharpen some of our results, sometimes we will assume that the pair (h, Ψ) satisfies the *sufficient admissibility condition* defined below. Observe that this condition is satisfied by the three reference functions h in Example 1 and the popular choices of Ψ mentioned above. By [32, Lemma 2.3], the sufficient admissibility condition also holds when h is a Legendre function and $\text{ri}(\text{dom}(\Psi)) \subseteq \text{ri}(\text{dom}(h))$.

Definition 1. Let $h : \mathbb{E} \rightarrow \mathbb{R} \cup \{\infty\}$ be a convex function differentiable on $\text{ri}(\text{dom}(h))$ and let $\Psi : \mathbb{E} \rightarrow \mathbb{R} \cup \{\infty\}$ be a closed convex function with $\text{dom}(\Psi) \subseteq \overline{\text{dom}(h)}$ and $\text{dom}(\Psi) \cap \text{ri}(\text{dom}(h)) \neq \emptyset$. The pair (h, Ψ) satisfies the *sufficient admissibility condition* if $L > 0$ is admissible for $g \in \mathbb{E}^*$ at $x \in \text{ri}(\text{dom}(h))$ whenever the function

$$y \mapsto \langle g, y \rangle + \Psi(y) + LD_h(y, x)$$

is bounded below.

We will rely on properties of the convex conjugate [5, 11, 18, 30] and on the following *three-point property* [13, Lemma 3.1] of the Bregman distance induced by h . For all $a \in \text{dom}(h)$ and $b, c \in \text{ri}(\text{dom}(h))$

$$D_h(a, b) + D_h(b, c) = D_h(a, c) - \langle \nabla h(b) - \nabla h(c), a - b \rangle.$$

1.3 Organization of the paper

The main sections of the paper are organized as follows. Section 1.4 presents a key lemma that provides the crux of our approach. Sections 2 through Section 4 detail our convex conjugate approach in the contexts of the Bregman proximal subgradient, Bregman proximal gradient, and accelerated Bregman proximal gradient templates. In the latter case we discuss the connection between our work and the recent work of Hanzely, Richtarik and Xiao [17]. Section 5 shows that a variant of our accelerated Bregman proximal gradient template that includes periodic restart has linear convergence provided that suitable smoothness and functional growth conditions hold. Finally, Section 6 summarizes some numerical experiments on the D-optimal design problem and on the Poisson linear inverse problem. Consistent with the numerical evidence reported in [17], we observe that the accelerated Bregman proximal gradient method converges approximately at a rate $\mathcal{O}(1/k^2)$. Furthermore, our computational experiments provide interesting new numerical evidence that explains this behavior.

1.4 A key lemma

Lemma 1 below provides the crux of our main developments. It concerns a construction that is common to all of the algorithmic schemes that we discuss in the main sections of the paper. Suppose the sequences $y_k, z_k \in \text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$, $g_k \in \partial f(y_k)$, and $t_k \in \mathbb{R}_+$ for $k = 0, 1, 2, \dots$ are such that $1/t_k$ is admissible for g_k at z_k and

$$z_{k+1} = \arg \min \{t_k(\langle g_k, z \rangle + \Psi(z)) + D_h(z, z_k)\}. \quad (5)$$

The optimality conditions enable us to rewrite (5) as

$$t_k(g_k + g_k^\Psi) + \nabla h(z_{k+1}) - \nabla h(z_k) = 0 \quad (6)$$

for some $g_k^\Psi \in \Psi(z_{k+1})$.

Lemma 1. *Suppose $y_k, z_k \in \text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$, $g_k \in \partial f(y_k)$, $g_k^\Psi \in \partial \Psi(z_{k+1})$, and $t_k \in \mathbb{R}_+$ satisfy (6). Then*

$$\begin{aligned} & \frac{\sum_{i=0}^k t_i (f(y_i) + \Psi(z_{i+1}) + \langle g_i, z_{i+1} - y_i \rangle) + D_h(z_{i+1}, z_i)}{\sum_{i=0}^k t_i} \\ &= -\frac{\sum_{i=0}^k t_i (f^*(g_i) + \Psi^*(g_i^\Psi))}{\sum_{i=0}^k t_i} - \left(\frac{1}{\sum_{i=0}^k t_i} D_h(\cdot, z_0) \right)^* (-u_k) \\ &\leq -f^*(v_k) - \Psi^*(w_k) - \left(\frac{1}{\sum_{i=0}^k t_i} D_h(\cdot, z_0) \right)^* (-u_k) \\ &\leq -\phi^*(u_k) - \left(\frac{1}{\sum_{i=0}^k t_i} D_h(\cdot, z_0) \right)^* (-u_k), \end{aligned}$$

where

$$v_k := \frac{\sum_{i=0}^k t_i g_i}{\sum_{i=0}^k t_i}, \quad w_k := \frac{\sum_{i=0}^k t_i g_i^\Psi}{\sum_{i=0}^k t_i}, \quad u_k := v_k + w_k = \frac{1}{\sum_{i=0}^k t_i} (\nabla h(z_0) - \nabla h(z_{k+1})).$$

Proof. The above statements readily follow from identities (7) and (8) below. For $k = 0, 1, \dots$

$$\nabla h(z_0) - \nabla h(z_{k+1}) = \sum_{i=0}^k t_i (g_i + g_i^\Psi) \quad (7)$$

and

$$\begin{aligned} & \sum_{i=0}^k t_i (f(y_i) + \Psi(z_{i+1}) + \langle g_i, z_{i+1} - y_i \rangle) + D_h(z_{i+1}, z_i) \\ &= -\sum_{i=0}^k t_i (f^*(g_i) + \Psi^*(g_i^\Psi)) + \langle \nabla h(z_0) - \nabla h(z_{k+1}), z_{k+1} \rangle + D_h(z_{k+1}, z_0). \end{aligned} \quad (8)$$

Indeed, (7) readily yields

$$u_k = \frac{1}{\sum_{i=0}^k t_i} (\nabla h(z_0) - \nabla h(z_{k+1}))$$

and

$$-\left(\frac{1}{\sum_{i=0}^k t_i} D_h(\cdot, z_0) \right)^* (-u_k) = \frac{\langle \nabla h(z_0) - \nabla h(z_{k+1}), z_{k+1} \rangle + D_h(z_{k+1}, z_0)}{\sum_{i=0}^k t_i}.$$

Therefore (8) implies that

$$\begin{aligned} & \frac{\sum_{i=0}^k t_i (f(y_i) + \Psi(z_{i+1}) + \langle g_i, z_{i+1} - y_i \rangle) + D_h(z_{i+1}, z_i)}{\sum_{i=0}^k t_i} \\ &= -\frac{\sum_{i=0}^k t_i (f^*(g_i) + \Psi^*(g_i^\Psi))}{\sum_{i=0}^k t_i} - \left(\frac{1}{\sum_{i=0}^k t_i} D_h(\cdot, z_0) \right)^* (-u_k). \end{aligned}$$

The subsequent two inequalities follow from the convexity of f^* and Ψ^* and the fact that $\phi = f + \Psi$ respectively.

We next prove (7) and (8). Identity (7) is an immediate consequence of (6). We prove (8) by induction. First, observe that from (6) it follows that

$$\begin{aligned} & t_k (f(y_k) + \Psi(z_{k+1}) + \langle g_k, z_{k+1} - y_k \rangle) + D_h(z_{k+1}, z_k) \\ &= -t_k (f^*(g_k) + \Psi^*(g_k^\Psi)) + \langle t_k (g_k + g_k^\Psi), z_{k+1} \rangle + D_h(z_{k+1}, z_k) \\ &= -t_k (f^*(g_k) + \Psi^*(g_k^\Psi)) + \langle \nabla h(z_k) - \nabla h(z_{k+1}), z_{k+1} \rangle + D_h(z_{k+1}, z_k) \end{aligned} \quad (9)$$

For $k = 0$ identity (8) immediately follows from (9). Suppose (8) holds for k . Then the induction hypothesis, identity (9), and three-point property of D_h imply that

$$\begin{aligned} & \sum_{i=0}^{k+1} t_i (f(y_i) + \Psi(z_{i+1}) + \langle g_i, z_{i+1} - y_i \rangle) + D_h(z_{i+1}, z_i) \\ &= -\sum_{i=0}^k t_i (f^*(g_i) + \Psi^*(g_i^\Psi)) + \langle \nabla h(z_0) - \nabla h(z_{k+1}), z_{k+1} \rangle + D_h(z_{k+1}, z_0) \\ &\quad + t_{k+1} (f(y_{k+1}) + \Psi(z_{k+2}) + \langle g_{k+1}, z_{k+2} - y_{k+1} \rangle) + D_h(z_{k+2}, z_{k+1}) \\ &= -\sum_{i=0}^k t_i (f^*(g_i) + \Psi^*(g_i^\Psi)) + \langle \nabla h(z_0) - \nabla h(z_{k+1}), z_{k+1} \rangle + D_h(z_{k+1}, z_0) \\ &\quad - t_{k+1} (f^*(g_{k+1}) + \Psi^*(g_{k+1}^\Psi)) + \langle \nabla h(z_{k+1}) - \nabla h(z_{k+2}), z_{k+2} \rangle + D_h(z_{k+2}, z_{k+1}) \\ &= -\sum_{i=0}^{k+1} t_i (f^*(g_i) + \Psi^*(g_i^\Psi)) + \langle \nabla h(z_0) - \nabla h(z_{k+2}), z_{k+2} \rangle + D_h(z_{k+2}, z_0). \end{aligned}$$

□

2 Bregman proximal subgradient

We first consider the case when f is convex and we only have a subgradient oracle for f . Algorithm 1 describes a Bregman proximal subgradient template for (1). This algorithmic template has been discussed in [15, 32]. Observe that Step 1 and Step 4 in Algorithm 1 automatically guarantee that $x_k \in \text{ri}(\text{dom}(f))$, $k = 0, 1, \dots$ by conditions (A.2) and (A.3) in Assumption 1.

Algorithm 1 Bregman proximal subgradient template

- 1: **input:** $x_0 \in \text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$
 - 2: **for** $k = 0, 1, 2, \dots$ **do**
 - 3: pick $g_k \in \partial f(x_k)$ and $t_k > 0$ so that $1/t_k$ is admissible for g_k at x_k
 - 4: $x_{k+1} := \arg \min_{x \in \mathbb{E}} \{t_k(\langle g_k, x \rangle + \Psi(x)) + D_h(x, x_k)\}$
 - 5: **end for**
-

Theorem 1. For $k = 0, 1, 2, \dots$ and $u_k := \frac{1}{\sum_{i=0}^k t_i} (\nabla h(x_0) - \nabla h(x_{k+1}))$ the iterates generated by Algorithm 1 satisfy

$$\begin{aligned} & \frac{\sum_{i=0}^k t_i (f(x_i) + \Psi(x_{i+1}) + \langle g_i, x_{i+1} - x_i \rangle) + D_h(x_{i+1}, x_i)}{\sum_{i=0}^k t_i} \\ & \leq -\phi^*(u_k) - \left(\frac{1}{\sum_{i=0}^k t_i} D_h(\cdot, x_0) \right)^* (-u_k). \end{aligned} \quad (10)$$

Proof. This follows by applying Lemma 1 to $z_k = y_k = x_k$ for $k = 0, 1, \dots$ □

Theorem 1 implies the convergence of $\min_{i=0,1,\dots,k} \phi(x_i)$ to $\min_x \phi(x)$ under fairly mild and general conditions as detailed in Corollary 1 and Corollary 2 below. To that end, we will rely on the following type of boundedness condition discussed by Teboulle [32].

Definition 2. The pair (f, h) satisfies the condition $W[f, h]$ on $C \subseteq \text{dom}(f) \cap \text{dom}(h)$ if there exists some $G > 0$ such that for all $x, u \in C, g \in \partial f(x)$, and $t > 0$ the following inequality holds

$$\langle tg, u - x \rangle - D_h(u, x) \leq \frac{G^2 t^2}{2}.$$

As noted by Teboulle [32], the condition $W[f, h]$ holds for $G = L/\sigma$ whenever f is L -Lipschitz and h is σ -strongly convex for some norm on \mathbb{E} . It is also easy to see that the condition $W[f, h]$ holds if f is G -continuous relative to h as defined by Lu [21].

The following result concerns the special case when $\Psi = \delta_C$ for some closed convex set $C \subseteq \text{dom}(f) \cap \text{dom}(h)$. In this case Algorithm 1 is the mirror-descent method for the problem

$$\min_{x \in C} f(x).$$

Corollary 1. Suppose $\Psi = \delta_C$ for some closed convex set $C \subseteq \text{dom}(f) \cap \text{dom}(h)$ and the pair (f, h) satisfies the $W[f, h]$ condition for some $G > 0$ on C . Then the iterates generated by Algorithm 1 satisfy

$$\min_{i=0,\dots,k} (f(x_i) - f(x)) \leq \frac{D_h(x, x_0) + \sum_{i=0}^k t_i^2 G^2 / 2}{\sum_{i=0}^k t_i}.$$

for all $x \in C$.

Proof. In this case $\Psi(x) = 0$ for all $x \in C$ and thus Theorem 1 implies that

$$\begin{aligned} & \frac{\sum_{i=0}^k t_i (f(x_i) + \langle g_i, x_{i+1} - x_i \rangle) + D_h(x_{i+1}, x_i)}{\sum_{i=0}^k t_i} \\ & \leq \min_{x \in C} \{f(x) - \langle u_k, x \rangle\} + \min_x \left\{ \frac{1}{\sum_{i=0}^k t_i} D_h(x, x_0) + \langle u_k, x \rangle \right\}. \end{aligned}$$

Therefore for all $x \in C$

$$\frac{\sum_{i=0}^k t_i (f(x_i) + \langle g_i, x_{i+1} - x_i \rangle) + D_h(x_{i+1}, x_i)}{\sum_{i=0}^k t_i} \leq f(x) + \frac{1}{\sum_{i=0}^k t_i} D_h(x, x_0).$$

Since each $g_i \in \partial f(x_i)$, the convexity of f and $W[f, h]$ condition imply that

$$\begin{aligned} \min_{i=0, \dots, k} (f(x_i) - f(x)) & \leq \frac{D_h(x, x_0) + \sum_{i=0}^k \langle t g_i, x_i - x_{i+1} \rangle - D_h(x_{i+1}, x_i)}{\sum_{i=0}^k t_i} \\ & \leq \frac{D_h(x, x_0) + \sum_{i=0}^k t_i^2 G^2 / 2}{\sum_{i=0}^k t_i}. \end{aligned}$$

□

For general Ψ , we have the following result discussed in [32]. This result is also closely related to some results by Bello-Cruz [9] on the proximal subgradient method.

Corollary 2. *Suppose the pair (ϕ, h) satisfies the $W[\phi, h]$ condition for some $G > 0$ on $\text{dom}(\phi)$. Then the iterates generated by Algorithm 1 satisfy*

$$\min_{i=0, \dots, k} (\phi(x_i) - \phi(x)) \leq \frac{D_h(x, x_0) + \sum_{i=0}^k t_i^2 G^2 / 2}{\sum_{i=0}^k t_i}$$

for all $x \in \text{dom}(\phi)$.

Proof. The convexity of Ψ and Theorem 1 imply that

$$\begin{aligned} & \frac{\sum_{i=0}^k t_i (\phi(x_i) + \langle g_i + \tilde{g}_i^\Psi, x_{i+1} - x_i \rangle) + D_h(x_{i+1}, x_i)}{\sum_{i=0}^k t_i} \\ & \leq \min_x \{ \phi(x) - \langle u_k, x \rangle \} + \min_x \left\{ \frac{1}{\sum_{i=0}^k t_i} D_h(x, x_0) + \langle u_k, x \rangle \right\} \end{aligned}$$

for any $\tilde{g}_i^\Psi \in \partial \Psi(x_i)$. Hence for all $x \in \text{dom}(\phi)$

$$\frac{\sum_{i=0}^k t_i (\phi(x_i) + \langle g_i + \tilde{g}_i, x_{i+1} - x_i \rangle) + D_h(x_{i+1}, x_i)}{\sum_{i=0}^k t_i} \leq f(x) + \frac{1}{\sum_{i=0}^k t_i} D_h(x, x_0).$$

Since each $g_i + \tilde{g}_i \in \partial\phi(x_i)$, the convexity of ϕ and $W[\phi, h]$ condition imply that

$$\begin{aligned} \min_{i=0,\dots,k} (\phi(x_i) - \phi(x)) &\leq \frac{D_h(x, x_0) + \sum_{i=0}^k \langle t(g_i + \tilde{g}_i), x_i - x_{i+1} \rangle - D_h(x_{i+1}, x_i)}{\sum_{i=0}^k t_i} \\ &\leq \frac{D_h(x, x_0) + \sum_{i=0}^k t_i^2 G^2 / 2}{\sum_{i=0}^k t_i}. \end{aligned}$$

□

In both Corollary 1 and Corollary 2 it is easy to see that if $t_i = 1/\sqrt{k+1}$, $i = 0, 1, \dots, k$ are admissible then for this choice of t_i , $i = 0, 1, \dots, k$ we have

$$\min_{i=0,\dots,k} (\phi(x_i) - \phi(x)) \leq \frac{D_h(x, x_0) + G^2/2}{\sqrt{k+1}}.$$

A closer look at the proof of Corollary 2 also reveals that if $t_i := 1/(i+1)$, $i = 0, 1, \dots$ are admissible then for this choice of t_i , $i = 0, 1, \dots$ we have $\min_{i=0,\dots,k} \phi(x_i) \rightarrow \inf_{x \in \mathbb{E}} \phi(x)$ provided the following weaker version of $W[\phi, h]$ holds: there exist $\gamma > 1$ and $G > 0$ such that for all $x, u \in \text{dom}(\phi) \cap \text{dom}(h)$ and $g \in \partial\phi(x)$

$$\langle tg, u - x \rangle - D_h(u, x) \leq (Gt)^\gamma.$$

Likewise for Corollary 1.

3 Bregman proximal gradient

Next, we consider the case when f is differentiable on $\text{ri}(\text{dom}(f))$ and we have a gradient oracle for f . Algorithm 2 describes a Bregman proximal gradient template for (1). This template has been discussed in [2, 3, 22, 32]. Observe that Step 1 and Step 4 in Algorithm 2 automatically guarantee that $x_k \in \text{ri}(\text{dom}(f))$, $k = 0, 1, \dots$ by conditions (A.2) and (A.3) in Assumption 1.

Algorithm 2 Bregman proximal gradient template

- 1: **input:** $x_0 \in \text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$
 - 2: **for** $k = 0, 1, 2, \dots$ **do**
 - 3: pick $L_k > 0$ admissible for $\nabla f(x_k)$ at x_k
 - 4: $x_{k+1} := \arg \min_{x \in \mathbb{E}} \{ \langle \nabla f(x_k), x \rangle + \Psi(x) + L_k D_h(x, x_k) \}$
 - 5: **end for**
-

The bound (11) in Theorem 2 below is similar to the bound (10) in Theorem 1. The similarity is more salient if we let $t_k := 1/L_k$, $k = 0, 1, \dots$

Theorem 2. For $k = 1, 2, \dots$ and $u_k := \frac{1}{\sum_{i=0}^k 1/L_i} (\nabla h(x_0) - \nabla h(x_{k+1}))$ the iterates generated by Algorithm 2 satisfy

$$\begin{aligned} & \frac{\sum_{i=0}^k (\phi(x_{i+1}) - D_f(x_{i+1}, x_i))/L_i + D_h(x_{i+1}, x_i)}{\sum_{i=0}^k 1/L_i} \\ & \leq -\phi^*(u_k) - \left(\frac{1}{\sum_{i=0}^k 1/L_i} D_h(\cdot, x_0) \right)^* (-u_k). \end{aligned} \quad (11)$$

Proof. Applying Lemma 1 to $z_k = y_k = x_k$, $g_k = \nabla f(x_k)$, and $t_k = 1/L_k$ for $k = 0, 1, 2, \dots$ we get

$$\begin{aligned} & \frac{\sum_{i=0}^k (\phi(x_{i+1}) - D_f(x_{i+1}, x_i))/L_i + D_h(x_{i+1}, x_i)}{\sum_{i=0}^k 1/L_i} \\ & = \frac{\sum_{i=0}^k (f(x_i) + \Psi(x_{i+1}) + \langle g_i, x_{i+1} - x_i \rangle)/L_i + D_h(x_{i+1}, x_i)}{\sum_{i=0}^k 1/L_i} \\ & \leq -\phi^*(u_k) - \left(\frac{1}{\sum_{i=0}^k 1/L_i} D_h(\cdot, x_0) \right)^* (-u_k). \end{aligned}$$

□

Corollary 3. Suppose L_k , $k = 0, 1, \dots$ in Step 3 of Algorithm 2 are chosen so that the following decrease condition holds

$$D_f(x_{k+1}, x_k) \leq L_k D_h(x_{k+1}, x_k). \quad (12)$$

Then

$$\phi(x_k) - \phi(x) \leq \frac{1}{\sum_{i=0}^{k-1} 1/L_i} D_h(x, x_0)$$

for all $x \in \text{dom}(\phi)$.

Proof. Theorem 2 and (12) imply

$$\begin{aligned} \frac{\sum_{i=0}^{k-1} \phi(x_{i+1})/L_i}{\sum_{i=0}^{k-1} 1/L_i} & \leq -\phi^*(u_k) - \left(\frac{1}{\sum_{i=0}^k 1/L_i} D_h(\cdot, x_0) \right)^* (-u_k) \\ & = \inf_{x \in \mathbb{E}} \{ \phi(x) - \langle u_{k-1}, x \rangle \} + \inf_{x \in \mathbb{E}} \left\{ \langle u_{k-1}, x \rangle + \frac{1}{\sum_{i=0}^{k-1} 1/L_i} D_h(x, x_0) \right\} \\ & \leq \phi(x) + \frac{1}{\sum_{i=0}^{k-1} 1/L_i} D_h(x, x_0) \end{aligned}$$

for all $x \in \text{dom}(\phi)$. Furthermore, (12) readily implies that $\phi(x_{i+1}) \leq \phi(x_i)$, $i = 0, 1, 2, \dots$. Therefore

$$\phi(x_k) \leq \frac{\sum_{i=0}^{k-1} \phi(x_{i+1})/L_i}{\sum_{i=0}^{k-1} 1/L_i} \leq \phi(x) - \frac{1}{\sum_{i=0}^{k-1} 1/L_i} D_h(x, x_0)$$

for all $x \in \text{dom}(\phi)$.

□

Consider the case when f is L_f -smooth relative to h on $\text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$ as defined in [3, 22]. This means that $L_f h - f$ is convex on $\text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$ for some constant $L_f > 0$ or equivalently [3, Lemma 1]

$$D_f(y, x) \leq L_f D_h(y, x)$$

for all $x, y \in \text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$. Suppose that in addition $\phi = f + \Psi$ is bounded below and the pair (h, Ψ) satisfies the sufficient admissibility condition (see Definition 1). Thus to ensure (12) we can choose $L_k = L := L_f$ if L_f is known, or more generally $L_k \leq L := \max\{\bar{L}, \alpha L_f\}$ for some $\alpha > 1$ and some initial guess \bar{L} via a standard backtracking procedure. In this case Corollary 3 thus yields the following convergence rate previously established in [3, 22]: for all $x \in \text{dom}(\phi)$

$$\phi(x_k) - \phi(x) \leq \frac{L D_h(x, x_0)}{k}.$$

This convergence rate can be readily sharpened by relying on the recent developments by Lee and Wright [19].

Corollary 4. *Suppose $L_k, k = 0, 1, \dots$ in Step 3 of Algorithm 2 are chosen so that (12) holds. If $L_k \leq L, k = 0, 1, \dots$ for some $L > 0$, $\bar{\phi} := \min_{x \in \mathbb{E}} \phi(x)$ is finite, and $\bar{X} := \arg \min_{x \in \mathbb{E}} \phi(x) \neq \emptyset$ then*

$$\phi(x_k) - \bar{\phi} = o(1/k).$$

Proof. By [19, Lemma 1], it suffices to show that the sequence $\{\phi(x_k) - \bar{\phi}, k = 0, 1, \dots\}$ is non-negative, monotonically decreasing, and summable. Indeed, (12) readily implies that $\{\phi(x_k) - \bar{\phi}, k = 0, 1, \dots\}$ is non-negative and monotonically decreasing. Furthermore, proceeding as in the proof of Corollary 3 it follows that for $k = 1, 2, \dots$

$$\sum_{i=0}^{k-1} (\phi(x_{i+1}) - \bar{\phi}) \leq L \sum_{i=0}^{k-1} (\phi(x_{i+1}) - \bar{\phi}) / L_i \leq L D(\bar{X}, x_0)$$

and so $\{\phi(x_k) - \bar{\phi}, k = 0, 1, \dots\}$ is summable. □

4 Accelerated Bregman proximal gradient

The interesting challenge of devising an accelerated version of Algorithm 2 was posed as an open problem in both [32] and [22]. A solution to this challenge was recently given by Hanzely, Richtarik, and Xiao in [17]. We develop a new accelerated Bregman proximal gradient template as described in Algorithm 3. This algorithmic template shares some similarities with Algorithm ABPG in [17] but there are also some key differences. In particular, Algorithm 3 relies only on the decrease condition (15) at each iteration. The algorithm does not require explicit knowledge of relative smooth or triangle scaling constants. Like Steps 1, 2, and 3 in [17, Algorithm ABPG], the updates of the sequences x_k, y_k, z_k in Steps 6, 8, and 9 of Algorithm 3 follow the same pattern used in the *Improved Interior Gradient Algorithm* (IGA) in [2].

As in Algorithm ABPG in [17] and in Algorithm IGA in [2], the gist of achieving acceleration in Algorithm 3 is to generate different sequences for the main iterates, the gradients, and the reference points used in the Bregman proximal gradient steps. (See steps 6, 8, and 9.) This is in sharp contrast to Algorithm 2 that generates a single sequence. The idea of generating different sequences can be traced back to Nesterov's seminal accelerated gradient algorithm [25] and underlies a number of other accelerated first-order algorithms [7, 14, 25–27].

Algorithm 3 Accelerated Bregman proximal gradient template

- 1: **input:** $x_0 \in \text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$; $\theta_0 := 1$ $z_0 := x_0$; $y_0 := x_0$
 - 2: pick $L_0 > 0$ admissible for $\nabla f(x_0)$ at x_0
 - 3: $x_1 := z_1 := \arg \min_{z \in \mathbb{E}} \{\langle \nabla f(x_0), z \rangle + \Psi(z) + L_0 D_h(z, x_0)\}$
 - 4: **for** $k = 1, 2, \dots$ **do**
 - 5: pick $\theta_k \in (0, 1)$ so that L_k is admissible for $\nabla f(y_k)$ at z_k for L_k and y_k as below
 - 6: $y_k := (1 - \theta_k)x_k + \theta_k z_k$
 - 7: $L_k := L_{k-1} \theta_{k-1} (1 - \theta_k) / \theta_k$
 - 8: $z_{k+1} := \arg \min_{z \in \mathbb{E}} \{\langle \nabla f(y_k), z \rangle + \Psi(z) + L_k D_h(z, z_k)\}$
 - 9: $x_{k+1} := (1 - \theta_k)x_k + \theta_k z_{k+1}$
 - 10: **end for**
-

An important feature of Algorithm 3 is the tight connection between the sequences $\{L_k, k = 0, 1, \dots\}$ and $\{\theta_k, k = 0, 1, \dots\}$. Indeed, a straightforward calculation shows that Step 7 in Algorithm 3 can be equivalently stated as follows

$$\theta_k = \frac{1/L_k}{\sum_{i=0}^k 1/L_i}, \quad k = 0, 1, \dots \quad (13)$$

As a consequence, Step 9 can be equivalently stated as

$$x_{k+1} = \frac{\sum_{i=0}^k z_{i+1}/L_i}{\sum_{i=0}^k 1/L_i}, \quad k = 0, 1, \dots$$

Similarly, Step 6 can be equivalently stated as

$$y_{k+1} = \frac{\sum_{i=0}^k z_{i+1}/L_i + z_{k+1}/L_{k+1}}{\sum_{i=0}^{k+1} 1/L_i}, \quad k = 0, 1, \dots$$

The bound (14) in Theorem 3 below has a similar format to the bounds (10) and (11).

Theorem 3. *For $k = 0, 1, 2, \dots$ and $u_k := \theta_k L_k (\nabla h(x_0) - \nabla h(z_{k+1}))$ the iterates generated by Algorithm 3 satisfy*

$$\begin{aligned} & \frac{\sum_{i=0}^k (\phi(z_{i+1}) - D_f(z_{i+1}, y_i))/L_i + D_h(z_{i+1}, z_i)}{\sum_{i=0}^k 1/L_i} \\ & \leq -\phi^*(u_k) - (\theta_k L_k D_h(\cdot, x_0))^*(-u_k). \end{aligned} \quad (14)$$

Proof. From Lemma 1 applied to $g_k = \nabla f(y_k)$, and $t_k = 1/L_k$ for $k = 0, 1, 2, \dots$ and (13) it follows that

$$\begin{aligned}
& \frac{\sum_{i=0}^k (\phi(z_{i+1}) - D_f(z_{i+1}, y_i))/L_i + D_h(z_{i+1}, z_i)}{\sum_{i=0}^k 1/L_i} \\
&= \frac{\sum_{i=0}^k (f(y_i) + \Psi(z_{i+1}) + \langle g_i, z_{i+1} - y_i \rangle)/L_i + D_h(z_{i+1}, z_i)}{\sum_{i=0}^k 1/L_i} \\
&\leq -\phi^*(u_k) - \left(\frac{1}{\sum_{i=0}^k 1/L_i} D_h(\cdot, x_0) \right)^* (-u_k) \\
&= -\phi^*(u_k) - (\theta_k L_k D_h(\cdot, x_0))^* (-u_k).
\end{aligned}$$

□

Corollary 5. *Suppose $L_0 > 0$ and $\theta_k \in (0, 1]$, $k = 0, 1, 2, \dots$ in Algorithm 3 are chosen so that each L_k is admissible for $\nabla f(y_k)$ at z_k and the following decrease condition holds for $k = 0, 1, 2, \dots$*

$$\phi(x_{k+1}) \leq (1 - \theta_k)\phi(x_k) + \theta_k(\phi(z_{k+1}) - D_f(z_{k+1}, y_k) + L_k D_h(z_{k+1}, z_k)). \quad (15)$$

Then for $k = 0, 1, 2, \dots$ the iterates generated by Algorithm 3 satisfy

$$\phi(x_{k+1}) - \phi(x) \leq \theta_k L_k D_h(x, x_0)$$

for all $x \in \text{dom}(\phi)$.

Proof. Theorem 3 together with (13), (15), and a straightforward induction imply that

$$\begin{aligned}
\phi(x_{k+1}) &\leq -\phi^*(u_k) - (\theta_k L_k D_h(\cdot, x_0))^* (-u_k) \\
&= \inf_{x \in \mathbb{E}} \{\phi(x) - \langle u_k, x \rangle\} + \inf_{x \in \mathbb{E}} \{\langle u_k, x \rangle + \theta_k L_k D_h(x, x_0)\} \\
&\leq \phi(x) + \theta_k L_k D_h(x, x_0)
\end{aligned}$$

for all $x \in \text{dom}(\phi)$. □

Our next two results, which are closely related to [17, Theorem 2], gives a rate of convergence for Algorithm 3 that generalizes the iconic $\mathcal{O}(1/k^2)$ convergence rate of Euclidean proximal methods under suitable Lipschitz continuity assumptions [7, 27]. We will rely on the following consequence of the weighted arithmetic geometric mean inequality established in [17, Lemma 3]. If $\gamma > 0$ then for $k = 1, 2, \dots$

$$\prod_{i=1}^k \frac{\gamma}{i + \gamma} \leq \left(\frac{\gamma}{k + \gamma} \right)^\gamma. \quad (16)$$

Proposition 1. *Suppose $L_0 > 0$ and $\theta_k \in (0, 1]$, $k = 0, 1, 2, \dots$ in Algorithm 3 are chosen so that each L_k is admissible for $\nabla f(y_k)$ at z_k and such that the decrease condition (15)*

holds. If $\theta_k \geq \gamma/(k + \gamma)$, $k = 0, 1, \dots$ for some constant $\gamma > 0$ then the iterates generated by Algorithm 3 satisfy

$$\phi(x_{k+1}) - \phi(x) \leq \left(\frac{\gamma}{k + \gamma} \right)^\gamma L_0 D_h(x, x_0) \quad (17)$$

for all $x \in \text{dom}(\phi)$ and $k = 1, 2, \dots$.

Proof. By Corollary 5, it suffices to show that for $k = 1, 2, \dots$

$$\theta_k L_k \leq \left(\frac{\gamma}{k + \gamma} \right)^\gamma L_0. \quad (18)$$

Inequality (18) in turn readily follows from (16), the fact that $\theta_i \geq \gamma/(i + \gamma)$, $i = 0, 1, \dots$, and the observation that for $k = 1, 2, \dots$

$$\theta_k L_k = \prod_{i=1}^k (1 - \theta_i) L_0.$$

□

Proposition 2. *Suppose that there are constants $L > 0$ and $\gamma > 0$ such that for all $x, z, \tilde{z} \in \text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$ and $\theta \in [0, 1]$*

$$D_f((1 - \theta)x + \theta z, (1 - \theta)x + \theta \tilde{z}) - (1 - \theta)D_f(x, (1 - \theta)x + \theta \tilde{z}) \leq \theta^\gamma L D_h(z, \tilde{z}). \quad (19)$$

In addition, suppose ϕ is bounded below and (h, Ψ) satisfy the sufficient admissibility condition. If Algorithm 3 chooses $L_0 := L$ and θ_k via $\theta_0 = 1$ and $\theta_{k+1}^\gamma = (1 - \theta_{k+1})\theta_k^\gamma$, $k = 0, 1, 2, \dots$ then each L_k is admissible for $\nabla f(y_k)$ at z_k and (15) holds for $k = 0, 1, \dots$. Furthermore, the iterates generated by Algorithm 3 satisfy

$$\phi(x_{k+1}) - \phi(x) \leq \left(\frac{\gamma}{k + \gamma} \right)^\gamma L D_h(x, x_0) \quad (20)$$

Proof. To prove that each L_k is admissible for $\nabla f(y_k)$ at z_k and (15) holds for $k = 0, 1, \dots$ it suffices to show that

$$\phi((1 - \theta_k)x_k + \theta_k z) - (1 - \theta_k)\phi(x_k) - \theta_k (\phi(z) - D_f(z, y_k)) \leq \theta_k L_k D_h(z, z_k) \quad (21)$$

for $z \in \text{ri}(\text{dom}(h))$. Indeed, on the one hand (21) can be rewritten as

$$\phi((1 - \theta_k)x_k + \theta_k z) - (1 - \theta_k)\phi(x_k) \leq \theta_k (\langle \nabla f(y_k), z \rangle + L_k D_h(z, z_k) + \Psi(z))$$

and thus L_k is admissible for $\nabla f(y_k)$ at z_k since ϕ is bounded below and (h, Ψ) satisfy the sufficient admissibility condition. On the other hand, (21) applied to $z = z_{k+1}$ readily implies (15).

We next prove (21). The convexity of Ψ and (19) yield

$$\begin{aligned} & \phi((1 - \theta_k)x_k + \theta_k z) - (1 - \theta_k)\phi(x_k) - \theta_k(\phi(z) - D_f(z, y_k)) \\ & \leq f((1 - \theta_k)x_k + \theta_k z) - (1 - \theta_k)f(x_k) + \theta_k D_f(z, (1 - \theta_k)x_k + \theta_k z_k) \\ & = D_f((1 - \theta_k)x_k + \theta_k z, (1 - \theta_k)x_k + \theta_k z_k) - (1 - \theta_k)D_f(x_k, (1 - \theta_k)x_k + \theta_k z_k) \\ & \leq \theta_k^\gamma L D_h(z, z_k). \end{aligned}$$

Next, observe that $\theta_k L_k = \prod_{i=0}^{k-1} (1 - \theta_i) L = \theta_k^\gamma L_0 = \theta_k^\gamma L$ since $\theta_0 = 1$ and $(1 - \theta_i)\theta_i^\gamma = \theta_{i+1}^\gamma$ for $i = 0, 1, \dots, k-1$. Thus (21) follows.

Finally, as shown in [17, Lemma 4], $\theta_0 = 1$ and $\theta_{k+1}^\gamma = (1 - \theta_{k+1})\theta_k^\gamma$ imply that $\theta_k \leq \gamma/(k + \gamma)$. Therefore Corollary 5 and $\theta_k L_k = \theta_k^\gamma L$ imply that the iterates generated by Algorithm 3 satisfy (20). \square

A limitation of Proposition 2 is that in principle it requires knowledge of L and γ to set L_0 and θ_k . Our next result shows that the same rate of convergence can be attained by Algorithm 3 if $L_0 > 0$ is chosen as small as possible and $\theta_k \in (0, 1]$ is chosen as large as possible. As we detail below, the slightly weaker rate (24) holds for more realistic and easily implementable line-search procedures that choose L_0 and θ_k .

Theorem 4. *Suppose (f, h, Ψ) satisfy (19) for all $x, z, \tilde{z} \in \text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$ and $\theta \in [0, 1]$, and in addition ϕ is bounded below and (h, Ψ) satisfy the sufficient admissibility condition. Suppose Algorithm 3 chooses $L_0 > 0$ in Step 2 as small as possible so that (15) holds for $k = 0$ and $\theta_k \in (0, 1)$, $k = 1, 2, \dots$ in Step 5 as large as possible so that L_k is admissible for $\nabla f(y_k)$ at z_k and (15) holds. Then the iterates generated by Algorithm 3 satisfy (20).*

Proof. By Corollary 5 it suffices to show that

$$\theta_k L_k \leq \left(\frac{\gamma}{k + \gamma} \right)^\gamma L. \quad (22)$$

To do so, we rely on the following inequality established in the proof of Proposition 2:

$$\phi((1 - \theta_k)x_k + \theta_k z) - (1 - \theta_k)\phi(x_k) - \theta_k(\phi(z) - D_f(z, y_k)) \leq \theta_k^\gamma L D_h(z, z_k). \quad (23)$$

We next prove (22) by induction. The case $k = 0$, that is, $L_0 \leq L$ readily follows from (23) and the fact that $\theta_0 = 1$. For the main inductive step, suppose (22) holds for $k - 1 \geq 0$ and thus

$$\theta_{k-1} L_{k-1} = \left(\frac{\gamma}{\hat{k} - 1 + \gamma} \right)^\gamma L$$

for some $\hat{k} \geq k$. Inequality (23) implies that L_k is admissible for $\nabla f(y_k)$ at z_k and (15) holds if

$$\theta_k^\gamma L \leq \theta_k L_k = (1 - \theta_k)\theta_{k-1} L_{k-1}.$$

Hence $\theta_k \geq \hat{\theta}$ where $\hat{\theta} \in (0, 1]$ is the root of

$$\frac{1 - \hat{\theta}}{\hat{\theta}^\gamma} = \frac{L}{\theta_{k-1} L_{k-1}} = \left(\frac{\hat{k} - 1 + \gamma}{\gamma} \right)^\gamma.$$

As shown in [17, Lemma 3], the weighted arithmetic mean geometric mean inequality implies that $\hat{\theta} \leq \gamma/(\hat{k} + \gamma)$. Therefore,

$$\theta_k L_k = (1 - \theta_k)\theta_{k-1}L_{k-1} \leq (1 - \hat{\theta})\theta_{k-1}L_{k-1} = \hat{\theta}^\gamma L \leq \left(\frac{\gamma}{\hat{k} + \gamma}\right)^\gamma L \leq \left(\frac{\gamma}{k + \gamma}\right)^\gamma L.$$

Thus (22) holds for k as well. \square

Consider the following more realistic line-search procedures. Suppose we choose L_0 via the following standard binary search procedure: Start with an initial guess $L_0 > 0$ and repeatedly scale L_0 (up or down) by $\alpha > 1$ until (15) just holds for $k = 0$. This kind of procedure will choose $L_0 \leq \alpha L$. Suppose $\theta_k \in (0, 1]$, $k = 1, 2, \dots$ is chosen via the following binary search procedure which is a variant of the approach used in [17, Algorithm ABPG-e]: Set $\theta_k := \gamma_k/(k + \gamma_k)$ for some initial guess $\gamma_k > 0$ and repeatedly increase or decrease γ_k by some sufficiently small $\delta > 0$ until (15) just holds. These two procedures and a straightforward modification of the proof of Theorem 4 imply that for some $\tilde{\gamma} \geq \gamma - \delta$ the iterates generated by Algorithm 3 satisfy

$$\phi(x_k) - \phi(x) \leq \left(\frac{\tilde{\gamma}}{k - 1 + \tilde{\gamma}}\right)^{\tilde{\gamma}} \alpha LD_h(x, x_0) \quad (24)$$

for all $x \in \text{dom}(\phi)$.

Our numerical experiments in Section 6 show that in many cases the above procedure yields $\gamma_k \approx 2$ for k sufficiently large. This behavior of the sequence $\{\gamma_k : k = 1, 2, \dots\}$ implies that the iterate values converge to the optimal value at the iconic $\mathcal{O}(1/k^2)$ convergence rate of accelerated gradient methods.

We conclude this section with a brief discussion of condition (19) that yields the $\mathcal{O}(1/k^\gamma)$ convergence rate of Algorithm 3. Observe that (19) holds if

$$D_f((1 - \theta)x + \theta z, (1 - \theta)x + \theta \tilde{z}) \leq \theta^\gamma LD_h(z, \tilde{z}) \quad (25)$$

for all $x, z, \tilde{z} \in \text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$ and $\theta \in [0, 1]$. Observe that f is L -smooth relative to h on $\text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$ precisely when (25) for $\theta = 1$ and all $x, z, \tilde{z} \in \text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$. (See [3, 22].) Thus when (25) holds for all $x, z, \tilde{z} \in \text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$ and $\theta \in [0, 1]$, we say that f is (L, γ) -smooth relative to h on $\text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$. We note that in the special case when $h(x) = \frac{1}{2}\|x\|_2^2$, if f is L -smooth relative to h then it is automatically $(L, 2)$ -smooth relative to h . In the general case, it is readily apparent that if f is L -smooth relative to h on some $C \subseteq \text{dom}(f) \cap \text{dom}(h)$ and D_h has triangle scaling exponent γ as defined in [17], then f is (L, γ) -smooth relative to h on C .

5 Linear convergence of accelerated Bregman proximal gradient

We next show that some variants of Algorithm 3 that include restart attain an accelerated linear rate of convergence provided that some suitable relative smoothness and functional

growth conditions hold. The algorithmic schemes and proofs in this section follow in a fairly straightforward fashion from the same ideas used in known restart schemes such as those in [23, 27, 28, 31]. We should note that unlike the previous algorithms in the paper, Algorithm 4 and Algorithm 5 below require some additional knowledge about the problem.

Throughout this section assume that $\bar{\phi} := \min_x \phi(x) < \infty$ and $\bar{X} := \{x \in \text{dom}(\phi) : \phi(x) = \bar{\phi}\} \neq \emptyset$. Let for $x \in \text{dom}(\phi)$ let $D_h(\bar{X}, x) := \inf_{\bar{x} \in \bar{X}} D_h(\bar{x}, x)$. Suppose f is both L_f -smooth relative to h and μ_f -strongly convex relative to h on $\text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$. That is, both $L_f h - f$ and $f - \mu_f h$ are convex on $\text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$. As discussed in [32] and [22], under these conditions the iterates generated by Algorithm 2 satisfy

$$D_h(\bar{X}, x_k) \leq \left(1 - \frac{\mu_f}{L_f}\right)^k D_h(\bar{X}, x_0)$$

and

$$\phi(x_k) - \bar{\phi} \leq L_f \left(1 - \frac{\mu_f}{L_f}\right)^k D_h(\bar{X}, x_0)$$

provided $L_k = L_f$, $k = 0, 1, \dots$. A straightforward modification of the argument in [32] shows that these inequalities also hold with L_f replaced with $\max\{\bar{L}, \alpha L_f\}$ if L_k is instead chosen via a backtracking procedure that starts with an initial guess \bar{L} for L_f and repeatedly scales it up by $\alpha > 1$ until condition (12) holds.

The above bounds imply that Algorithm 2 yields $x_k \in \text{dom}(\phi)$ with $\phi(x_k) - \bar{\phi} < \epsilon$ in at most

$$k = \mathcal{O}\left(\frac{L_f}{\mu_f} \cdot \log\left(\frac{L_f D_h(\bar{X}, x_0)}{\epsilon}\right)\right)$$

iterations. We next show that under the stronger (L, γ) -relative smoothness assumption and a relative *functional growth* assumption, two variants of Algorithm 3 that include restart achieve a faster linear rate. Note that Algorithm 4 requires knowledge of the optimal value $\bar{\phi}$. On the other hand, Algorithm 5 requires knowledge of a certain *condition number* L/κ_ϕ of ϕ and of a constant $\gamma > 1$ so that f is (L, γ) -smooth relative to h on $\text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$.

Suppose $\bar{\phi} := \min_{x \in \mathbb{E}} \phi(x)$ is finite, $\bar{X} := \arg \min_{x \in \mathbb{E}} \phi(x) \neq \emptyset$, and $\kappa > 0$. Following [23], we shall say that ϕ has κ -*functional growth relative to* h if

$$\phi(x) - \bar{\phi} \geq \kappa D_h(\bar{X}, x)$$

for all $x \in \text{dom}(\phi) \cap \text{ri}(\text{dom}(h))$.

Proposition 3. *Suppose $L, \gamma, \kappa > 0$ are such that f is (L, γ) -smooth relative to h on $\text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$ and ϕ has κ -functional growth relative to h . Then each call to Algorithm 3 in Algorithm 4 halts after at most*

$$k = \gamma \left(\frac{2L}{\kappa}\right)^{1/\gamma} \tag{26}$$

iterations. On the other hand, the sequence of outer iterates $\{w_\ell : \ell = 0, 1, \dots\}$ generated by Algorithm 5 satisfies

$$\phi(w_{\ell+1}) - \bar{\phi} \leq \frac{\phi(w_\ell) - \bar{\phi}}{2}. \tag{27}$$

Algorithm 4 Accelerated Bregman proximal gradient with restart (version 1)

Pick $w_0 \in \text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$

for $\ell = 0, 1, \dots$ **do**

 let $x_0 := w_\ell$ and run Algorithm 3 until

$$\phi(x_k) - \bar{\phi} \leq \frac{\phi(x_0) - \bar{\phi}}{2}$$

 let $w_{\ell+1} := x_k$

end for

Algorithm 5 Accelerated Bregman proximal gradient with restart (version 2)

Pick $w_0 \in \text{ri}(\text{dom}(h)) \cap \text{dom}(\Psi)$

for $\ell = 0, 1, \dots$ **do**

 let $x_0 := w_\ell$ and run Algorithm 3 until

$$k = \gamma \left(\frac{2L}{\kappa} \right)^{1/\gamma}$$

 let $w_{\ell+1} := x_k$

end for

and

$$D_h(\bar{X}, w_{\ell+1}) \leq \frac{D_h(\bar{X}, w_\ell)}{2}.$$

In particular, either Algorithm 4 or Algorithm 5 yields $x_K \in \text{dom}(\phi)$ such that $\phi(x_K) - \bar{\phi} < \epsilon$ after at most

$$K = \mathcal{O} \left(\left(\frac{L}{\kappa} \right)^{1/\gamma} \log \left(\frac{LD_h(\bar{X}, x_0)}{\epsilon} \right) \right)$$

accelerated Bregman proximal gradient iterations.

Proof. Theorem 4 implies that the iterates generated by Algorithm 3 satisfy

$$\phi(x_k) - \bar{\phi} \leq \left(\frac{\gamma}{k-1+\gamma} \right)^\gamma LD_h(\bar{X}, x_0) \leq \frac{L}{\kappa_\phi} \left(\frac{\gamma}{k-1+\gamma} \right)^\gamma (\phi(x_0) - \bar{\phi}).$$

Thus both (26) and (27) follow. In addition, for $\ell = 0, 1, \dots$ the outer iterates generated by Algorithm 5 satisfy

$$\phi(w_{\ell+1}) - \bar{\phi} \leq \left(\frac{\gamma}{k-1+\gamma} \right)^\gamma LD_h(\bar{X}, w_\ell) \leq \frac{\kappa D_h(\bar{X}, w_\ell)}{2}$$

and so

$$D_h(\bar{X}, w_{\ell+1}) \leq \frac{\phi(w_{\ell+1}) - \bar{\phi}}{\kappa} \leq \frac{D_h(\bar{X}, w_\ell)}{2}.$$

□

6 Numerical experiments

We implemented a MATLAB version of Algorithm 2 with line-search to choose L_k . Following the convention in [17], we will refer to this implementation as Algorithm BPG-LS. We also implemented two MATLAB versions of Algorithm 3. The first one sets $L_0 := L_f$ and θ_k via $\theta_0 = 1$ and $\theta_k^\gamma = (1 - \theta_k)\theta_{k-1}^\gamma$, $k = 1, 2, \dots$ assuming that L_f and γ are known. As indicated in Section 4, this version is identical to Algorithm ABPG in [17]. We also implemented a second version of Algorithm 3 with the line-search procedures to choose L_0 and θ_k sketched at the end of Section 4 for $\alpha = 2$ and $\delta = 0.1$. In particular, our implementation sets $\theta_k = \frac{\gamma k}{k + \gamma k}$, $k = 1, 2, \dots$ where $\gamma_k > 0$ is chosen via line-search so that (15) holds. We refer to this version as Algorithm ABPG-LS.

We next report results on some numerical experiments on random instances of two problems that provide interesting tests for Bregman proximal methods. The first one is the D-optimal design problem [1, 22]

$$\min_{x \in \Delta_{n-1}} -\log(\det(HXH^\top))$$

where $X = \text{Diag}(x)$ and $H \in \mathbb{R}^{m \times n}$ with $m < n$ and $\Delta_{n-1} := \{x \in \mathbb{R}_+^n : \|x\|_1 = 1\}$.

The second one is the Poisson linear inverse problem [3]

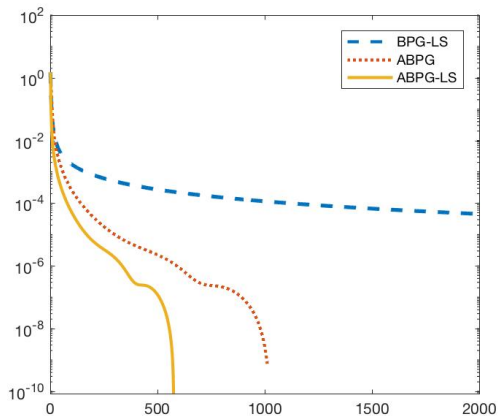
$$\min_{x \in \mathbb{R}_+^n} D_{KL}(b, Ax)$$

where $b \in \mathbb{R}_+^n$ and $A \in \mathbb{R}_+^{m \times n}$ with $m > n$ and $D_{KL}(\cdot, \cdot)$ is the Kullback-Leibler divergence, that is, the Bregman distance associated to the Boltzmann-Shannon entropy function $x \mapsto \sum_{i=1}^n x_i \log(x_i)$.

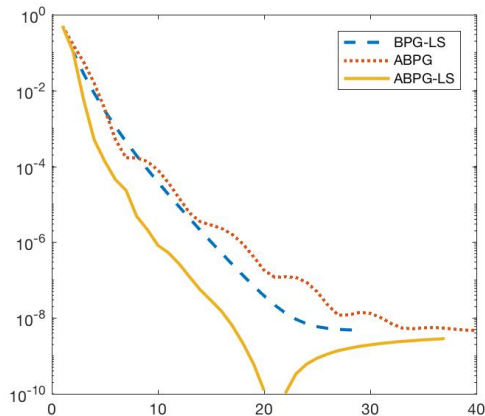
It was shown in [22] that the function $f(x) = -\log(\det(HXH^\top))$ is 1-smooth relative to the Burg entropy $h(x) = -\sum_{i=1}^n \log(x_i)$. On the other hand, it was shown in [3] that the function $x \mapsto D_{KL}(b, Ax)$ is $\|b\|_1$ -smooth relative to $h(x) = -\sum_{i=1}^n \log(x_i)$. Thus we use the Burg entropy $h(x) = -\sum_{i=1}^n \log(x_i)$ as reference function for both problems. The implementation of Algorithm ABPG requires values of L_f and γ as input. We used the values $L_f = 1$ for the D-optimal design problem and $L_f = \|b\|_1$ for the Poisson linear inverse problem which are “safe” as per the above relative smoothness results. For γ , we used the default value $\gamma = 2$. This value is attractive because it yields the accelerated rate $\mathcal{O}(1/k^2)$ but is not safe because as discussed in [17], the Bregman distance for the Burg entropy has a smaller uniform triangle scaling exponent. Nonetheless, as in the experiments reported in [17], the choice of $\gamma = 2$ worked fine in our numerical experiments.

Figure 1 depicts the convergence of Algorithms BPG-LS, ABPG, and ABPG-LS on two typical random instances $H \in \mathbb{R}^{100 \times 250}$ and $H \in \mathbb{R}^{200 \times 300}$ for the D-optimal design problem. The suboptimality gap is measured relative to the smallest objective value attained by the three algorithms, which was ABPG-LS in all cases. The entries of the instances H for this problem are independent draws from the standard normal distribution.

Figure 2 depicts similar convergence results on typical random instances $A \in \mathbb{R}^{250 \times 100}$, $b \in \mathbb{R}^{250}$ and $A \in \mathbb{R}^{300 \times 200}$, $b \in \mathbb{R}^{300}$ for the Poisson linear inverse problem. In this case the entries of A and of b are independent draws from the uniform distribution on $[0, 1]$.

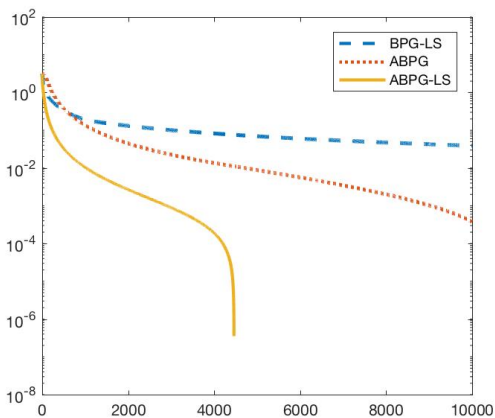


(a) $m = 100, n = 250$

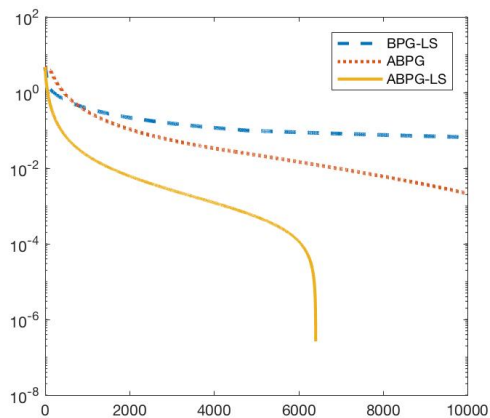


(b) $m = 200, n = 300$

Figure 1: Suboptimality gap on typical instances of the D-optimal design problem.



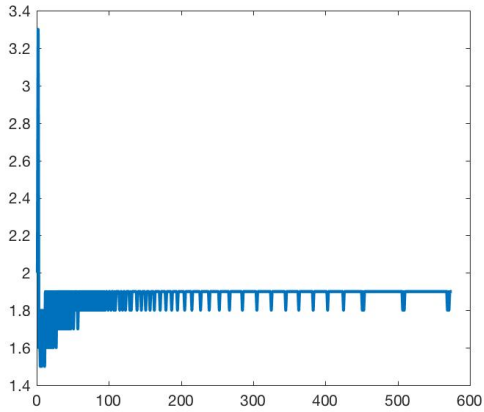
(a) $m = 250, n = 100$



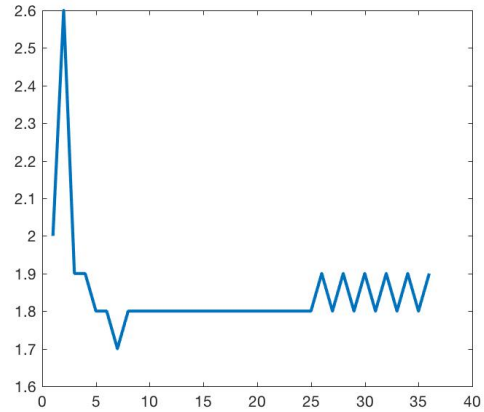
(b) $m = 300, n = 200$

Figure 2: Suboptimality gap on typical instances of the Poisson linear inverse problem.

The numerical experiments demonstrate that the convergence rates of the algorithms BPG-LS, ABPG, and ABPG-LS usually follow the pattern one would expect: In most cases Algorithm BPG-LS is the slowest while ABPG-LS is the fastest. An exception occurs in the easier 200×300 D-optimal design instances where BPG-LS performs as well as ABPG or better. As noted in [17] this can be attributed to the better conditioning of these instances and the linear convergence property of Algorithm 2. Figure 3 and Figure 4 depict an interesting phenomenon that we observed in our experiments. These figures display plots of the values of γ_k throughout the execution of Algorithm ABPG-LS in the four instances discussed above. In all of these cases it is evident that γ_k hovers near 2. Since the algorithm sets $\theta_k = \gamma_k / (k + \gamma_k)$, these values of γ_k imply that Algorithm ABPG-LS approximately attains the iconic $\mathcal{O}(1/k^2)$ convergence rate of accelerated gradient methods. This numerical evidence is striking and consistent with the results reported in [17].

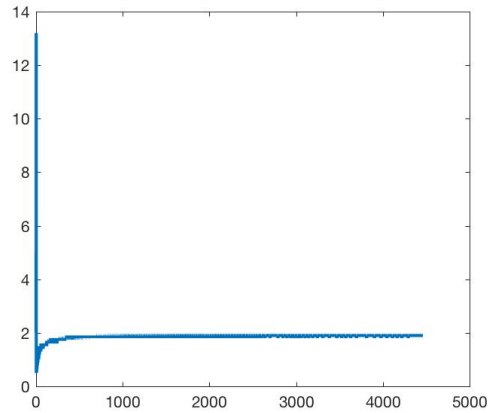


(a) $m = 400, n = 250$

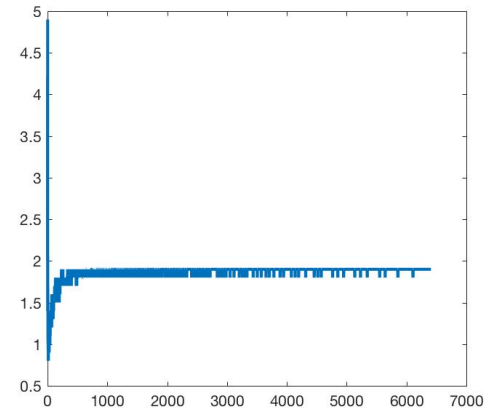


(b) $m = 200, n = 100$

Figure 3: Sequence $\{\gamma_k : k = 1, 2, \dots\}$ in ABPG-LS for typical instances of D-design optimal problem.



(a) $m = 400, n = 250$



(b) $m = 200, n = 100$

Figure 4: Sequence $\{\gamma_k : k = 1, 2, \dots\}$ in ABPG-LS for typical instances of Poisson linear inverse problem.

References

- [1] C. Atwood. Optimal and efficient designs of experiments. *The Annals of Mathematical Statistics*, pages 1570–1602, 1969.
- [2] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16(3):697–725, 2006.
- [3] H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2016.

- [4] H. Bauschke and J. Borwein. Legendre functions and the method of random Bregman projections. *Journal of Convex Analysis*, 4(1):27–67, 1997.
- [5] H. Bauschke and P. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.
- [6] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.
- [7] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- [8] S. Becker, E. Candès, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, 2011.
- [9] J. Bello-Cruz. On proximal subgradient splitting method for minimizing the sum of two nonsmooth convex functions. *Set-Valued and Variational Analysis*, 25(2):245–263, 2017.
- [10] J. Bello-Cruz and T. Nghia. On the convergence of the forward–backward splitting method with line-searches. *Optimization Methods and Software*, 31(6):1209–1238, 2016.
- [11] J. Borwein and A. Lewis. *Convex analysis and nonlinear optimization*. Springer, New York, 2000.
- [12] L. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.
- [13] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.
- [14] D. Drusvyatskiy, M. Fazel, and S. Roy. An optimal first order method based on optimal quadratic averaging. *SIAM Journal on Optimization*, 28(1):251–271, 2018.
- [15] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT*, pages 14–26, 2010.
- [16] D. Gutman and J. Peña. Convergence rates of proximal gradient methods via the convex conjugate. *SIAM Journal on Optimization*, *To Appear*, 2018.
- [17] F. Hanzely, P. Richtarik, and L. Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *arXiv preprint arXiv:1808.03045*, 2018.
- [18] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer–Verlag, Berlin, 1993.
- [19] C. Lee and S. Wright. First-order algorithms converge faster than $\mathcal{O}(1/k)$ on convex problems. *arXiv preprint arXiv:1812.08485*, 2018.

- [20] P. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.
- [21] H. Lu. “Relative-continuity” for non-Lipschitz non-smooth convex optimization using stochastic (or deterministic) mirror descent. *arXiv preprint arXiv:1710.04718*, 2017.
- [22] H. Lu, R. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.
- [23] I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, pages 1–39, 2018.
- [24] A. Nemirovsky and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.
- [25] Y. Nesterov. A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. Doklady AN SSSR (in Russian). (*English translation. Soviet Math. Dokl.*), 269:543–547, 1983.
- [26] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Kluwer Academic Publishers, 2004.
- [27] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.
- [28] B. O’Donoghue and E. J. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15:715–732, 2015.
- [29] J. Peña. Convergence of first-order methods via the convex conjugate. *Operations Research Letters*, 45:561–564, 2017.
- [30] T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.
- [31] V. Roulet and A. d’Aspremont. Sharpness, restart and acceleration. In *Advances in Neural Information Processing Systems*, pages 1119–1129, 2017.
- [32] M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, pages 1–30, 2018.
- [33] P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125(2):263–295, 2010.
- [34] Y. Zhou, Y. Liang, and L. Shen. A unified approach to proximal algorithms using Bregman distance. Technical report, Tech. report, 2016.