Linesearch-free adaptive Bregman proximal gradient for convex minimization without relative smoothness*

Hongjia Ou† Puya Latafat‡ Andreas Themelis†

Abstract

This paper introduces adaptive Bregman proximal gradient algorithms for solving convex composite minimization problems without relying on global relative smoothness or strong convexity assumptions. Building upon recent advances in adaptive stepsize selections, the proposed methods generate stepsizes based on local curvature estimates, entirely eliminating the need for backtracking linesearch. A key innovation is a Bregman generalization of Young's inequality, which allows controlling a critical inner product in terms of the same Bregman distances used in the updates. Our theory applies to problems where the differentiable term is merely *locally* smooth relative to a distance-generating function, without requiring the existence of global moduli or symmetry coefficients. Numerical experiments demonstrate their competitive performance compared to existing approaches across various problem classes.

Keywords. Convex minimization, Bregman proximal gradient method, relative smoothness, adaptive stepsizes, Bregman distance.
AMS subject classifications. 65K05, 90C06, 90C25, 90C30, 49M29.

Contents

*A. Themelis was supported by the JSPS KAKENHI grant number JP24K20737. P. Latafat is a member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA - National Group for Mathematical Analysis, Probability and their Applications) of the Istituto Nazionale di Alta Matematica (INdAM - National Institute of Higher Mathematics).

†Faculty of Information Science and Electrical Engineering (ISEE), Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan. *E-mails:* ou.honjia.069@s.kyushu-u.ac.jp, andreas.themelis@ees.kyushu-u.ac.jp

‡IMT School for Advanced Studies Lucca. *E-mail:* puya.latafat@imtlucca.it

1 Introduction

This work considers structured optimization problems of the form

$$\underset{x\in\overline{C}}{\text{minimize}}\ \varphi(x) \coloneqq f(x) + g(x), \tag{P}$$

where $\overline{C}$ denotes the closure of $C \coloneqq \operatorname{int}\operatorname{dom}\phi$ for a proper 1-coercive function $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ of Legendre type, $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ are proper closed convex, while f is only assumed to be *locally* smooth relative to the *distance-generating function* (dgf). By local smoothness relative to ϕ we refer to the existence, for any compact and convex set $\mathcal{K} \subset \operatorname{int}\operatorname{dom}\phi$, of a constant $L^\phi_{f,\mathcal{K}} \geq 0$ such that $L^\phi_{f,\mathcal{K}}\phi - f$ is convex over $\mathcal{K}$; see Assumption 2.1. When such a constant L^ϕ_f exists independently of $\mathcal{K}$, that is such that $L^\phi_f\phi - f$ is convex on $\operatorname{int}\operatorname{dom}\phi$, then the usual notion of (global) smoothness of f relative to ϕ is recovered.

For convex f, these definitions are respective generalizations of local and global Lipschitz-smoothness of f, that is, local and global Lipschitz continuity of ∇f. Indeed, these are recovered when $\phi = \mathfrak{j}$, where with

$$\mathfrak{j} \coloneqq \tfrac{1}{2}\|\cdot\|^2$$

we denote the squared Euclidean norm.

A standard approach for solving (P) is to use fixed point iterations with the Bregman proximal gradient (BPG) operator

$$x^{k+1} = \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \left\{ f(x^k) + \langle \nabla f(x^k), w - x^k\rangle + g(w) + \tfrac{1}{\gamma_{k+1}} \mathrm{D}_\phi(w, x^k) \right\}, \quad (1.1)$$

where $\gamma_{k+1} > 0$ is the stepsize parameter and

$$\mathrm{D}_\phi(x,y) \coloneqq \begin{cases} \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle & \text{if } (x,y) \in \operatorname{dom}\phi \times \operatorname{int}\operatorname{dom}\phi \\ \infty & \text{otherwise} \end{cases} \quad (1.2)$$

is the Bregman distance associated to ϕ.

Noticing that $\mathrm{D}_{\mathfrak{j}}(x,y) = \frac{1}{2}\|x-y\|^2$ is the squared Euclidean norm, BPG updates (1.1) reduce to standard proximal gradient iterations when $\phi = \mathfrak{j}$. More generally, a well-chosen kernel ϕ can naturally encode the feasible region through its domain, blending the benefits of barrier and operator splitting methods. Beyond this, the Bregman framework addresses important smoothness limitations. In many applications, the differentiable term f lacks Lipschitz-smoothness, ubiquitous requirement for first-order methods, but may exhibit smoothness *relative* to a kernel ϕ other than the squared Euclidean norm [3, 14]. Moreover, even in unconstrained problems and with f enjoying global Lipschitz smoothness, appropriate kernel selections can yield tighter smoothness parameters, enabling larger stepsizes and faster convergence.

1.1 Motivations and related work

Under the assumption that f is globally L_f^ϕ-relatively smooth, the BPG method enjoys a descent property in terms of the Bregman distance, provided the stepsize does not exceed $(1+\alpha)/L_f^\phi$, where $\alpha \in [0,1]$ is the so-called *symmetry coefficient* [3]. However, reliance on global smoothness and symmetry properties typically leads to unnecessarily small stepsizes and slow convergence in practice. This observation is not limited to the Bregman setting, as it also pertains the standard proximal gradient method and first-order algorithms in general. For this reason, even when constant stepsizes based on *global* moduli are employable, time-varying selections reflecting the *local* landscape of the problem can significantly improve algorithmic performance.

Adaptive methods in the Euclidean setting *Backtracking linesearch* is a well-established practice to achieve this; linesearch refers to a trial-and-error procedure that iteratively adjusts the stepsize until a prescribed condition, typically a descent on the cost, is verified. These techniques can significantly accelerate convergence by selecting more effective stepsizes, but they incur higher per-iteration costs due to repeated evaluations, until the needed condition is

met, thereby leading to slower individual iterations. In response to this, [16] introduced an adaptive stepsize selection strategy for the gradient method based on local estimates of Lipschitz moduli that can be derived from available data. Initially limited to smooth minimization, the approach was refined and extended in [12] to accommodate nonsmooth proximable terms, and later further refined in several other flavors [17, 11, 29, 21]. Among these developments, [20] showed that this class of adaptive methods extends to the local Hölderian setting, while [13, 24] proposed accelerated variants à la Nesterov [18].

Advances in the Bregman setting All the above-mentioned works are however limited to the standard proximal gradient setup, dubbed "Euclidean setting" as it is captured by the choice of $\phi = \frac{1}{2}\|\cdot\|^2$ as the square Euclidean norm. Linesearch techniques can be directly extended to Bregman proximal gradient iterations (1.1), where they preserve the same advantages and limitations as in the Euclidean case. In contrast, generalizing other adaptive schemes to the Bregman setting proves significantly more challenging. To the best of our knowledge, the only successful extension in this direction is the BaGRAAL method proposed in [25], which adapts the golden-ratio scheme of [15] to the Bregman context. Remarkably, similarly to its predecessor [25] covers a class of hemivariational inequalities broader than composite minimization. On the other hand, it requires the Bregman kernel ϕ to be strongly convex, with the stepsize parameters explicitly dependent on the corresponding modulus of strong convexity. More importantly, this method derives its stepsizes from *Euclidean* (Lipschitz) estimates, which are not aligned with the underlying Bregman geometry. As we will demonstrate in our simulations, this results in conservative stepsize selections and, consequently, slower convergence in practice.

1.2 Contribution

In this work, we propose two adaptive stepsize selection strategies for Bregman proximal gradient iterations that operate without requiring strong convexity of the Bregman kernel or global Lipschitz smoothness of the differentiable term, thus significantly broadening the scope of applicability of Bregman-based methods. A central technical challenge in the Bregman setting arises from controlling inner product terms, which in the Euclidean case are typically bounded using Young's or Cauchy–Schwarz inequalities. The game changer in our approach is the introduction of a novel Bregman generalization of Young's inequality, which enables a direct and effective handling of inner products in terms of Bregman distances. Extensive numerical experiments confirm that the proposed methodology outperforms existing approaches in terms of convergence speed, robustness across problem classes, and employment of large stepsizes, all under very general working assumptions.

1.3 Preliminaries and notation

The set of natural numbers is $\mathbb{N} := \{0, 1, 2, \ldots\}$, while $\mathbb{R}$, $\mathbb{R}_{++} := (0, \infty)$, and $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ denote the set of real, strictly positive, and extended-real numbers, respectively. For $t \in \mathbb{R}$, we define $[t]_+ := \max\{t, 0\}$. We use $\langle \cdot\,, \cdot \rangle$ to denote the standard inner product on $\mathbb{R}^n$, and for a symmetric and positive definite $\mathbb{R}^{n\times n}$ matrix Q, denoted $Q \in \mathrm{Sym}_{++}(\mathbb{R}^n)$, we let $\|x\|_Q = \sqrt{\langle x, Qx\rangle}$ be

the induced norm. In case Q is the identity matrix, we simply write $\|x\|$. Given a set $\mathcal{D} \subseteq \mathbb{R}^n$, with $\operatorname{int} \mathcal{D}$ we denote its interior.

With $\mathrm{id} : \mathbb{R}^n \to \mathbb{R}^n$ we indicate the identity function, while $\mathfrak{j} : \mathbb{R}^n \to \mathbb{R}$ indicates the square Euclidean norm $\mathfrak{j}(x) = \frac{1}{2}\|x\|^2$. The *domain* and *epigraph* of an extended-real-valued function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ are the sets $\operatorname{dom} h := \{x \in \mathbb{R}^n \mid h(x) < \infty\}$ and $\operatorname{epi} h := \{(x, c) \in \mathbb{R}^n \times \mathbb{R} \mid h(x) \leq c\}$. Function h is said to be: *proper* if $h > -\infty$ and $\operatorname{dom} h \neq \emptyset$; *lower semicontinuous (lsc)* if $\operatorname{epi} h$ is a closed subset of $\mathbb{R}^{n+1}$; *1-coercive* if $\lim_{\|x\|\to\infty} \frac{h(x)}{\|x\|} = \infty$.

The *conjugate* of a proper, lsc, convex function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is the proper, lsc, convex function $h^* : \mathbb{R}^n \to \overline{\mathbb{R}}$ defined by $h^*(\xi) := \sup_{x\in\mathbb{R}^n} \{\langle \xi, x\rangle - h(x)\}$. The *subdifferential* of h at $x \in \operatorname{dom} h$ is the set

$$\partial h(x) := \{u \in \mathbb{R}^n \mid h(x') \geq h(x) + \langle u, x' - x\rangle \ \forall x' \in \mathbb{R}^n\},$$

while $\partial h(x) = \emptyset$ for $x \notin \operatorname{dom} h$. h is differentiable at x iff $\partial h(x)$ is a singleton, and in this case one has that $\partial f(x) = \{\nabla h(x)\}$.

Bregman distance

We next list a few known facts related to Bregman distances as in (1.2).

Fact 1.1 (three-point identity [8, Lem. 3.1])**.** *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper lsc convex function differentiable on $\operatorname{int}\operatorname{dom} h$. For any $x \in \operatorname{dom} h$, and $y, z \in \operatorname{int}\operatorname{dom} h$ the following holds:*

$$\mathrm{D}_\phi(x, z) = \mathrm{D}_\phi(x, y) + \mathrm{D}_\phi(y, z) + \langle x - y, \nabla h(y) - \nabla h(z)\rangle.$$

We say that a proper, lsc, convex function $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *of Legendre type* (or simply *Legendre*) if it is (i) *essentially smooth*, namely differentiable on $\operatorname{int}\operatorname{dom} h \neq \emptyset$ and such that $\|\nabla h(x^k)\| \to \infty$ whenever $\operatorname{int}\operatorname{dom} h \ni x^k \to x \in \operatorname{dom} h \setminus \operatorname{int}\operatorname{dom} h$, and (ii) *essentially strictly convex*, namely strictly convex on every convex subset of $\operatorname{dom} \partial h$.

Fact 1.2. *Let $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be of Legendre type, and let $C := \operatorname{int}\operatorname{dom}\phi$.*

(i) [4, Thm. 3.7(vi)] $\mathrm{D}_\phi(x, \cdot\,)$ *is 1-coercive for any $x \in C$.*

(ii) [4, Thm. 3.8(i)] *If a sequence $\left(x^k\right)_{k\in\mathbb{N}}$ converges to a boundary point of C, then $\mathrm{D}_\phi(x, x^k) \to \infty$ for any $x \in C$.*

(iii) [4, Prop. 2.16], [22, Thm. 26.5] *The conjugate function ϕ^* is continuously differentiable, strictly convex on $\mathbb{R}^n$, and $\nabla\phi^* = \nabla\phi^{-1}$.*

(iv) [4, Thm. 3.7(v)] $\mathrm{D}_\phi(x, y) = \mathrm{D}_{\phi^*}(\nabla\phi(y), \nabla\phi(x))$ *for any $x, y \in C$.*

Finally, we introduce the symbol Δ_ϕ to indicate the *symmetrized Bregman distance*

$$\Delta_\phi(x, y) := \mathrm{D}_\phi(x, y) + \mathrm{D}_\phi(y, x) = \begin{cases} \langle \nabla\phi(x) - \nabla\phi(y), x - y\rangle & \text{if } x, y \in \operatorname{int}\operatorname{dom}\phi \\ \infty & \text{otherwise.} \end{cases} \tag{1.3}$$

2 Problem setting and proposed algorithms

Problem (P) will be studied under the following assumptions.

Assumption 2.1. *The following hold in* (P)*:*

(i) $C = \operatorname{int} \operatorname{dom} \phi$ *for a proper, convex, 1-coercive function* ϕ *of Legendre type, which is twice differentiable with* $\nabla^2 \phi \succ 0$ *on* C*.*

(ii) $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is proper, convex, lsc, and* locally smooth relative to ϕ*: that is,* $\operatorname{dom} \phi \subseteq \operatorname{dom} f$*, and for every convex and compact set* $\mathcal{K} \subset C$ *there exists* $L^{\phi}_{f,\mathcal{K}} > 0$ *such that* $L^{\phi}_{f,\mathcal{K}} \phi - f$ *is convex on* $\mathcal{K}$*.*

(iii) $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is proper, lsc, and convex with* $\operatorname{dom} g \cap C \neq \emptyset$*.*

(iv) *A solution exists:* $\operatorname{arg\,min}_{\overline{C}} \varphi \neq \emptyset$*.*

Beyond convexity, all these basic requirements on f and g are virtually negligible. The local relative smoothness in Assumption 2.1(ii) is tantamount to saying that f is differentiable on C,[1] and that

$$\mathrm{D}_f(x, y) \leq L^{\phi}_{f,\mathcal{K}} \, \mathrm{D}_{\phi}(x, y) \quad \forall x, y \in \mathcal{K}, \tag{2.1}$$

which simplifies to local Lipschitz continuity of ∇f when $\phi = \frac{1}{2}\|x\|^2$. Importantly, note that $\mathcal{K}$ need only be a compact subset of *the interior* of $\operatorname{dom} \phi$, thereby far from boundary points at which ϕ is vertical because of essential smoothness. As such, any convex function f which is, say, twice differentiable on C, enjoys this requirement, even if exhibiting an infinite slope at boundary points of C.

Example 2.2. Let $\phi(x) = x \ln x - x$ be the *Boltzmann-Shannon entropy* on $\mathbb{R}_+$. Any convex function that is twice differentiable on $\mathbb{R}_{++}$, such as $f(x) = \frac{1}{x}$ on $\mathbb{R}_{++}$ and ∞ elsewhere, is locally smooth relative to ϕ as in Assumption 2.1(ii), despite the fact that there may exist no L such that $L\phi - f$ is convex in a neighborhood of 0 (as is the case for the given f). □

As we detail in the following subsection, the basic requirements listed in Assumption 2.1 are enough to guarantee that the proposed adaptive stepsize selection strategies produce iterates x^k such that $\inf_{k \in \mathbb{N}} \varphi(x^k) = \inf_{\overline{C}} \varphi$. Slightly more can be said upon assuming that the Bregman distance generated by ϕ satisfies the following mild additional assumption.

Assumption 2.3 (Bregman with zone C [23, Def. 2.1])**.** *The dgf* ϕ *satisfies the following:*

(i) $\mathrm{D}_{\phi}(x, x^k) \to 0$ *whenever* $C \ni x^k \to x$ *(in particular,* $\operatorname{dom} \phi$ *is closed).*

(ii) $\mathrm{D}_{\phi}(x, \cdot\,)$ *is level bounded for any* $x \in \overline{C} \setminus C$*.*

Assumption 2.3 holds vacuously whenever ϕ has full domain $\mathbb{R}^n$. More generally, it is a standard requirement satisfied by many kernels used in practice; see for instance [3, Rem. 4].

[1] See [28, Prop. 3.7] or [1, Prop. 2.5].

2.1 Local moduli estimates

Our approach builds upon the Euclidean analyses of [12, 11], and more generally follows the "self-adaptive" rationale of generating stepsizes solely based on past available data, without resorting to inner loops or requiring existence (or knowledge thereof) of any global modulus. However, the involvement of Bregman geometry brings forth several challenges that do not allow for straightforward extensions of these works.

Each iteration revolves around three local estimates: two are Lipschitz-like estimates for ∇f and for the *forward operator*

$$H_k \coloneqq \nabla\phi - \gamma_k \nabla f, \tag{2.2}$$

and one measuring the gap between $\mathrm{D}_\phi(x, y)$ and $\mathrm{D}_\phi(y, x)$ at specific points. This latter measure is superfluous in Euclidean analyses, since the quadratic function $\phi = \mathfrak{j}$ enjoys complete symmetry.

Differentiable function f First, based on (2.1),

$$\ell_k \coloneqq \frac{\Delta_f(x^k, x^{k-1})}{\Delta_\phi(x^k, x^{k-1})} \tag{2.3a}$$

provides an estimate of the relative smoothness modulus on the line segment between the last two consecutive iterates x^{k-1} and x^k. This is the obvious counterpart of (the inverse of) a Barzilai-Borwein stepsize [2], and the local Lipschitz estimate of ∇f used in [12] for the Euclidean case, which indeed matches (2.3a) when $\phi = \mathfrak{j}$.

Forward operator H_k Inferring a Bregman equivalent of a local Lipschitz estimate of the forward operator (2.2) is not as immediate. Indeed, replacing $\|H_k(x^k) - H_k(x^{k-1})\|^2$ with, say, a Bregman term $\mathrm{D}_{\phi^*}(H_k(x^k), H_k(x^{k-1}))$ does not seem to lead to quantities that naturally arise in the analysis. Our solution is more convoluted, and specifically given by

$$\Lambda_{k,\delta} \frac{2\,\mathrm{D}_{\phi^*}\big(\nabla\phi(x^k) + \delta\big[H_k(x^k) - H_k(x^{k-1})\big], \nabla\phi(x^k)\big)}{\delta^2\,\Delta_\phi(x^k, x^{k-1})} \tag{2.3b}$$

depending on some parameter $\delta > 0$ (specified later). Despite its deceptive intricacy, when $\phi = \mathfrak{j}$ is quadratic, and thus so is its conjugate ϕ^*, this estimate recovers the (square) Lipschitz estimate $\frac{\|H_k(x^k) - H_k(x^{k-1})\|^2}{\|x^k - x^{k-1}\|^2}$ of [12, Lem. 2.1(ii)], independently of the parameter δ. However, a judicious choice of δ will be crucial for our convergence analysis in the generality of Assumption 2.1. The expression (2.3b) for more general ϕ owes to the following Bregman extension of the Young inequality

$$\langle x - y, v\rangle \leq \tfrac{1}{\delta}\,\mathrm{D}_\phi(x, y) + \tfrac{1}{\delta}\,\mathrm{D}_{\phi^*}\big(\nabla\phi(y) + \delta v, \nabla\phi(y)\big),$$

see Lemma 3.2 for a precise statement.

Bregman kernel ϕ Lastly, a local *symmetry coefficient*

$$\alpha_k := \frac{\mathrm{D}_\phi(x^k, x^{k-1})}{\mathrm{D}_\phi(x^{k-1}, x^k)} \in (0, \infty) \tag{2.3c}$$

allows us to express

$$\Delta_\phi(x^k, x^{k-1}) = \tfrac{1+\alpha_k}{\alpha_k}\, \mathrm{D}_\phi(x^k, x^{k-1}). \tag{2.4}$$

Note that $\alpha_k > 0$ holds for any k by essential strict convexity of ϕ, regardless of whether or not ϕ has a *global* (strictly positive) *symmetry coefficient*

$$\alpha(\phi) := \inf_{\substack{x,y \in \operatorname{int}\operatorname{dom}\phi \\ x \neq y}} \frac{\mathrm{D}_\phi(x, y)}{\mathrm{D}_\phi(y, x)}. \tag{2.5}$$

Even when it does, a symmetry coefficient based on the global landscape of ϕ may be excessively conservative; instead, the use of local estimates enables the adoption of tighter constants, ultimately leading to larger stepsizes for Bregman proximal gradient iterations (1.1). The interested reader is referred to the recent work [19] for an in-depth analysis of the coefficient $\alpha(\phi)$ for a certain class of Bregman kernels ϕ.

2.2 Proposed algorithms and main results

Based on these three quantities, choose a stepsize γ_{k+1} and proceed with a Bregman proximal gradient update (1.1). We propose the following two options, where we let

$$\rho_{k+1} := \tfrac{\gamma_{k+1}}{\gamma_k}$$

denote the ratio of consecutive stepsizes (so that $\gamma_{k+1} = \rho_{k+1}\gamma_k$):

B-adaPG

set $\hat{\rho}_{k+1} = \sqrt{1+\rho_k}$ and $\delta = 2\hat{\rho}_{k+1}$, and update

$$\rho_{k+1} = \min\left\{\hat{\rho}_{k+1}, \frac{\alpha_k}{1+\alpha_k}\frac{1}{2\hat{\rho}_{k+1}\left[\Lambda_{k,\delta} - (1-\gamma_k\ell_k)\right]_+}\right\} \tag{2.6}$$

and, in case ϕ enjoys a symmetry coefficient $\alpha = \alpha(\phi) > 0$,

B-adaPG$_\alpha$ (if $\alpha(\phi) > 0$)

set $\hat{\rho}_{k+1} = \sqrt{\frac{1+\alpha}{2} + \rho_k}$ and $\delta = \frac{2}{1+\alpha}\hat{\rho}_{k+1}$, and update

$$\rho_{k+1} = \min\left\{\hat{\rho}_{k+1}, \frac{\alpha}{2\hat{\rho}_{k+1}\left[\Lambda_{k,\delta} - (1-\gamma_k\ell_k)\right]_+}\right\}. \tag{2.7}$$

Remark 2.4. We use the convention that $\frac{1}{0} = \infty$, and remind that $[t]_+ = \max\{0, t\}$. In particular, whenever $\Lambda_{k,\delta} \leq 1 - \gamma_k\ell_k$ all updates reduce to $\rho_{k+1} = \hat{\rho}_{k+1}$. It is implied that a starting point $x^0 \in C$ should be provided, as well as two stepsizes $\gamma_0, \gamma_1 > 0$ for the first two iterations. We refer the reader to Section 5.1.1 for a practical initialization strategy, which recasts the one in [12, §2.1.1] in our Bregman setting. □

The following theorem collects the main results for Bregman proximal gradient iterations (1.1) with stepsizes selected according to the above rules.

Theorem 2.5 (summary of main results)**.** *Suppose that Assumption 2.1 holds, and consider the iterates generated by B-adaPG. Then, one always has that*

$$\inf_{k\in\mathbf{N}} \varphi(x^k) = \inf_{\overline{C}} \varphi. \tag{2.8}$$

Moreover,

(i) *If* $C \cap \arg\min_{\overline{C}} \varphi \neq \emptyset$ *(equivalently, if* $C \cap \arg\min \varphi \neq \emptyset$*), then there exists* $x^\star \in C \cap \arg\min \varphi$ *such that* $x^k \to x^\star$.

(ii) *If Assumption 2.3 holds, then* $\left(x^k\right)_{k\in\mathbf{N}}$ *is bounded and admits a unique optimal limit point.*

When ϕ *has symmetry coefficient* $\alpha > 0$*, the same is true for B-adaPG*$_\alpha$.

Any Legendre kernel ϕ with a non-open domain necessarily satisfies $\alpha(\phi) = 0$ [3, Prop. 2]. Consequently, the statement of Theorem 2.5(ii) is only nontrivial for the B-adaPG variant: a positive symmetry coefficient $\alpha(\phi) > 0$, combined with the domain closedness required by Assumption 2.3, forces $\operatorname{dom}\phi = \mathbb{R}^n$. In this case, the stronger conclusion of Theorem 2.5(i) applies instead.

Although the general theoretical results are weaker than in the Euclidean setting, our proposed methods demonstrate significant practical advantages. Extensive numerical simulations in Section 5 confirm that the adaptive choices B-adaPG and B-adaPG$_\alpha$ enable larger stepsizes (even on average) and dramatically speed up convergence, outperforming even aggressive linesearch strategies. Providing a firm theoretical basis for this observed performance, similar to what has been established in the Euclidean case, remains a compelling direction for future research. Nonetheless, we outline below some theoretical refinements that can already be ensured, whose full details are however omitted for the sake of keeping the presentation simple.

Remark 2.6. Some comments are in order for Theorem 2.5.

(i) The setting of Theorem 2.5(i) corresponds to the case in which φ has *unconstrained* minimizers lying in $\operatorname{int}\operatorname{dom}\phi$; that is, the Bregman geometry induced by ϕ does not act as an active *barrier* in the minization problem. In this case, it can actually be shown that the stepsizes γ_k in both algorithmic variants are lower bounded by some $\gamma_{\min} > 0$, a property that can be used to infer a rate

$$\min_{k\leq K} \varphi(x^k) - \inf_{\overline{C}} \varphi \leq O(\tfrac{1}{K+1})$$

for the best-so-far cost; see Lemma 4.1. Such a lower bound $\gamma_{\min}$ can be derived albeit with tedious, very conservative, and not particularly insightful expressions; for this reason, we confine the discussion to this brief comment.

(ii) Concerning Theorem 2.5(ii), it can be shown that the entire sequence converges up to replacing the first element in the minimum defining the update rule of ρ_{k+1} with $(1-\epsilon)\hat{\rho}_{k+1}$ for some $0 < \epsilon \ll 1$ (as opposed to $\epsilon = 0$). Under Assumption 2.3, this slight modification of the stepsize rule guarantees that the sequence $\left(\gamma_k(\varphi(x^k) - \inf_{\overline{C}}\varphi)\right)_{k\in\mathbf{N}}$ converges to zero. This fact

follows from a simple telescoping argument on (3.18), from which sequential convergence can be established arguing as in the proof of Theorem 4.4. We conjecture that the same convergence result holds without this modification of the stepsize update rule, but a formal proof of this fact remains an open problem. □

2.2.1 Comparison with Euclidean methods

The B-adaPG variant can be interpreted as a Bregman analogue of the adaPG update of [12], which reads

$$\rho_{k+1}^{\mathsf{adaPG}} = \min\left\{\hat{\rho}_{k+1},\ \frac{1}{2\sqrt{[\Lambda_k-(1-\gamma_k\ell_k)]_+}}\right\}.$$

Here, ℓ_k and Λ_k are as in (2.3a) and (2.3b) with $\phi = \frac{1}{2}$ (as already mentioned, in this case the latter is independent of the parameter δ and is thus omitted from the subscript). On the other hand, with $\phi = \frac{1}{2}$ (hence $\alpha_k = 1$) (2.6) reads

$$\begin{aligned}\rho_{k+1}^{\mathsf{B\text{-}adaPG}} &= \min\left\{\hat{\rho}_{k+1},\ \frac{1}{4\hat{\rho}_{k+1}[\Lambda_k-(1-\gamma_k\ell_k)]_+}\right\}\\ &\leq \min\left\{\hat{\rho}_{k+1},\ \frac{1}{2\sqrt{[\Lambda_k-(1-\gamma_k\ell_k)]_+}}\right\} = \rho_{k+1}^{\mathsf{adaPG}},\end{aligned}$$

thus introducing a slight conservativsm over the Euclidean predecessor; the inequality owes to the fact that $\rho_{k+1} \leq \hat{\rho}_{k+1}$, hence that $\rho_{k+1}^2 \leq \rho_{k+1}\hat{\rho}_{k+1}$.

Similarly, the B-adaPG$_\alpha$ variant with $\phi = \frac{1}{2}$ (hence $\alpha = 1$) simplifies to a slightly more conservative variant of the update

$$\rho_{k+1}^{\mathsf{adaPG}^{1,\frac{1}{2}}} := \min\left\{\sqrt{1+\rho_k},\ \frac{1}{\sqrt{2[\Lambda_{k,\delta}-(1-\gamma_k\ell_k)]_+}}\right\}$$

of [11, adaPG$^{1,\frac{1}{2}}$], having

$$\rho_{k+1}^{\mathsf{B\text{-}adaPG}_\alpha} = \min\left\{\hat{\rho}_{k+1},\ \frac{1}{2\hat{\rho}_{k+1}[\Lambda_{k,\delta}-(1-\gamma_k\ell_k)]_+}\right\} \leq \rho_{k+1}^{\mathsf{adaPG}^{1,\frac{1}{2}}}$$

(in all occurrences throughout this subsection, $\hat{\rho}_{k+1} = \sqrt{1+\rho_k}$).

Remark 2.7 (quadratic kernels)**.** As explained in Section 2.1, for general ϕ the curvature estimate $\Lambda_{k,\delta}$ as in (2.3b) depends on the Bregman-Young parameter $\delta > 0$. In the analyses of [12, 11], this parameter is optimally chosen as a suitable multiple of the ratio $\rho_{k+1} = \gamma_{k+1}/\gamma_k$, a feasible choice given that the value of $\Lambda_{k,\delta}$ is independent of δ. This is not the case for more general kernels ϕ, whence the above-discussed conservatism originates: the value of γ_{k+1} depends on $\Lambda_{k,\delta}$, and should $\Lambda_{k,\delta}$ in turn depend on γ_{k+1} a circular dependency would arise.

Specializing B-adaPG$_\alpha$ to quadratic $\phi = \frac{1}{2}\|\cdot\|_Q^2$ with $Q \in \mathrm{Sym}_{++}(\mathbb{R}^n)$, this issue does not persist and the tighter analyses of the Euclidean cases are recovered. The corresponding algorithm produces iterates

$$x^{k+1} = \operatorname*{arg\,min}_{x\in\mathbb{R}^n}\left\{\langle\nabla f(x^k), x-x^k\rangle + g(x) + \frac{1}{2\gamma_{k+1}}\|x-x^k\|_Q^2\right\}$$

with stepsizes chosen as

$$\gamma_{k+1} = \gamma_k \min\left\{\sqrt{1+\tfrac{\gamma_k}{\gamma_{k-1}}},\ \frac{1}{\sqrt{2\left[\gamma_k^2 L_k^2 - \gamma_k\ell_k\right]_+}}\right\},$$

where

$$\ell_k = \tfrac{\langle \nabla f(x^k) - \nabla f(x^{k-1}), x^k - x^{k-1}\rangle}{\|x^k - x^{k-1}\|_Q^2} \quad \text{and} \quad L_k = \tfrac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|_{Q^{-1}}}{\|x^k - x^{k-1}\|_Q}.$$

This expression follows from the easily verifiable fact that $\Lambda_{k,\delta} = \gamma_k^2 L_k^2 - 2\gamma_k \ell_k + 1$ (independently of δ) in this case. □

3 Main inequalities

To ease the subsequent discussion, we introduce some convenient notational shorthands and remind some of those already encountered. Relative to the iterates generated by (1.1), for any $k \in \mathbb{N}$ and $x \in \operatorname{dom}\varphi$ we denote

$$P_k(x) \coloneqq \varphi(x^k) - \varphi(x) \tag{3.1}$$

and

$$B_{k+1} \coloneqq \rho_{k+1}\langle x^{k+1} - x^k, H_k(x^k) - H_k(x^{k-1})\rangle, \tag{3.2}$$

where $\rho_{k+1} = \frac{\gamma_{k+1}}{\gamma_k}$ and

$$H_k \coloneqq \nabla\phi - \gamma_k \nabla f \tag{3.3}$$

is the "forward" operator. Due to convexity of g, the BPG updates in (1.1) are characterized by the subgradient inclusion

$$\tilde{\nabla} g(x^{k+1}) \coloneqq \tfrac{\nabla\phi(x^k) - \nabla\phi(x^{k+1})}{\gamma_{k+1}} - \nabla f(x^k) \in \partial g(x^{k+1}). \tag{3.4a}$$

Throughout, we use the notation $\tilde{\nabla} g(x^{k+1})$ to indicate this particular element of the subgradient of g along the iterates. Similarly, we use

$$\begin{aligned}\tilde{\nabla}\varphi(x^{k+1}) &\coloneqq \tilde{\nabla} g(x^{k+1}) + \nabla f(x^{k+1}) \\ &= \tfrac{H_{k+1}(x^k) - H_{k+1}(x^{k+1})}{\gamma_{k+1}} \in \partial\varphi(x^{k+1}).\end{aligned} \tag{3.4b}$$

In light of these, we adopt the notation

$$\tilde{\mathrm{D}}_g(w, x^k) \coloneqq g(w) - g(x^k) - \langle \tilde{\nabla} g(x^k), w - x^k\rangle,$$

and similarly

$$\tilde{\mathrm{D}}_\varphi(w, x^k) \coloneqq \varphi(w) - \varphi(x^k) - \langle \tilde{\nabla}\varphi(x^k), w - x^k\rangle,$$

which are both positive quantities for any $k \in \mathbb{N}$ and $w \in \mathbb{R}^n$.

The main identity in our study is an adaptation to the Bregman setting of the inequality in [12, Lem. 2.2]. The proof closely patterns the one in the reference, and is provided in the appendix for completeness.

Lemma 3.1 (main identity; extension of [12, Lem. 2.2])**.** *Suppose that Assumption 2.1 holds, and starting from $x^0 \in C$ consider Bregman proximal gradient iterations (1.1) with stepsizes $\gamma_k > 0$. Then, for any $x \in \operatorname{dom}\varphi \cap \operatorname{dom}\phi$, $\vartheta_{k+1} \geq 0$, $k \in \mathbb{N}$, it holds that*

$$\begin{aligned}&\mathrm{D}_\phi(x, x^{k+1}) + \gamma_{k+1}(1 + \vartheta_{k+1})P_k(x) + \mathrm{D}_\phi(x^{k+1}, x^k) \\ &= \mathrm{D}_\phi(x, x^k) + \gamma_{k+1}\vartheta_{k+1}P_{k-1}(x) - \rho_{k+1}\vartheta_{k+1}(1 - \gamma_k\ell_k)\,\Delta_\phi(x^k, x^{k-1}) + B_{k+1} \\ &\quad - \gamma_{k+1}\Big\{\mathrm{D}_f(x, x^k) + \tilde{\mathrm{D}}_g(x^{k+1}, x^k) + \tilde{\mathrm{D}}_g(x, x^{k+1}) + \vartheta_{k+1}\,\tilde{\mathrm{D}}_\varphi(x^{k-1}, x^k)\Big\},\end{aligned}$$

where $P_k(x)$, B_{k+1}, and ℓ_k are as in (3.1), (3.2), and (2.3a). In particular, denoting

$$\widehat{\mathcal{U}}_k(x) \coloneqq \mathrm{D}_\phi(x, x^k) + \gamma_k(1+\vartheta_k)P_{k-1}(x) + \mathrm{D}_\phi(x^k, x^{k-1}),$$

one has that

$$\begin{aligned}\widehat{\mathcal{U}}_{k+1}(x) \leq \widehat{\mathcal{U}}_k(x) - \gamma_k(1+\vartheta_k-\rho_{k+1}\vartheta_{k+1})P_{k-1}(x) + B_{k+1} \\ - \left[\tfrac{\alpha_k}{1+\alpha_k} + \rho_{k+1}\vartheta_{k+1}(1-\gamma_k\ell_k)\right]\Delta_\phi(x^k, x^{k-1}).\end{aligned} \tag{3.5}$$

Proof. See Appendix A. □

The following two subsections will be devoted to developing analogues of the Young's inequality allowing us to bound the inner product term B_{k+1} in terms of Bregman distances. These pinpoint the main departure from previous Euclidean analyses, in particular the need to introduce a new parameter $\hat{\rho}_{k+1}$ that generates some unavoidable conservatism; see the discussion after Theorem 2.5.

3.1 Young's inequality in the Bregman sense

Young's inequality is a very simple but powerful tool enabling to bound inner products in terms of sum of squares. Its derivation is elementary, all revolving around the fact that, for any $u, v \in \mathbb{R}^n$ and $\delta > 0$,

$$\langle u, v\rangle = \tfrac{1}{\delta}\langle u, \delta v\rangle = \tfrac{1}{2\delta}\|u\|^2 + \tfrac{\delta}{2}\|v\|^2 - \tfrac{1}{2\delta}\|u-\delta v\|^2.$$

Discarding the negative term leaves us with the familiar bound holding for any $\delta > 0$. The same arguments can be extended beyond quadratic norms to more general Bregman distances by means of the three-point identity of Fact 1.1.

Lemma 3.2 (Young's inequality in the Bregman sense)**.** *Let $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a Legendre and 1-coercive convex function. Then, for any $x \in \operatorname{dom}\phi$, $y \in \operatorname{int}\operatorname{dom}\phi$, $v \in \mathbb{R}^n$, and $\delta > 0$ one has*

$$\langle x-y, v\rangle \leq \tfrac{1}{\delta}\,\mathrm{D}_\phi(x, y) + \tfrac{1}{\delta}\,\mathrm{D}_{\phi^*}\big(\nabla\phi(y) + \delta v, \nabla\phi(y)\big). \tag{3.6}$$

Proof. We have

$$\begin{aligned}\langle x-y, v\rangle &= \tfrac{1}{\delta}\langle x-y, \delta v\rangle \\ &= \tfrac{1}{\delta}\langle x-y, \nabla\phi(\nabla\phi^*\big(\nabla\phi(y)+\delta v\big)) - \nabla\phi(y)\rangle \\ &= \tfrac{1}{\delta}\,\mathrm{D}_\phi(x, y) + \tfrac{1}{\delta}\,\mathrm{D}_\phi\big(y, \nabla\phi^*\big(\nabla\phi(y)+\delta v\big)\big) \\ &\quad - \tfrac{1}{\delta}\,\mathrm{D}_\phi\big(x, \nabla\phi^*\big(\nabla\phi(y)+\delta v\big)\big).\end{aligned}$$

The second equality follows from Legendreness and 1-coercivity of ϕ, ensuring that $(\nabla\phi)^{-1} = \nabla\phi^*$ [22, Thm. 26.5] and that the domain of ϕ^* is $\mathbb{R}^n$ [4, Prop. 2.16]. The third equality is derived from the three point identity of Fact 1.1. Since $\mathrm{D}_\phi \geq 0$, the claimed inequality (3.6) is obtained. □

When $\phi = \mathfrak{j}$, one recovers the usual Young's inequality

$$\langle x-y, v\rangle \leq \tfrac{1}{2\delta}\|x-y\|^2 + \tfrac{\delta}{2}\|v\|^2 =: \psi_{\mathfrak{j}}(\delta).$$

Note that the right-hand side $\psi_j(\delta)$ diverges as $\delta \to 0^+$ and $\delta \to \infty$, and attains a unique minimizer at $\delta = \frac{\|x-y\|}{\|v\|}$. This specific choice of δ leads to the Cauchy-Schwarz inequality

$$\langle x - y, v\rangle \leq \inf_{\delta>0} \psi_j(\delta) = \|x - y\|\|v\|.$$

A similar pattern occurs for the right-hand side in (3.6), although with some complications arising because of the dependency on δ for the second Bregman distance.

Lemma 3.3 (A Cauchy–Schwarz inequality in the Bregman sense)**.** *In the setting of Lemma 3.2, either the upper bound in (3.6) is always decreasing in δ, or it is minimized at a unique $\delta^\star > 0$ which is characterized by the identity*

$$\mathrm{D}_{\phi^*}\big(\nabla\phi(y), \nabla\phi(y) + \delta^\star v\big) = \mathrm{D}_\phi(x, y). \tag{3.7}$$

In this latter case, which is necessarily true when $\operatorname{dom}\phi$ *is open, one has that*

$$\langle x - y, v\rangle \leq \tfrac{1}{\delta^\star}\,\Delta_{\phi^*}\big(\nabla\phi(y) + \delta^\star v, \nabla\phi(y)\big). \tag{3.8}$$

Proof. Let

$$\psi_\phi(\delta) \coloneqq \frac{\mathrm{D}_\phi(x, y) + \mathrm{D}_{\phi^*}\big(\nabla\phi(y) + \delta v, \nabla\phi(y)\big)}{\delta}$$

denote the right-hand side of the Bregman-Young inequality (3.6). A simple computation reveals that its derivative is

$$\begin{aligned}\psi'_\phi(\delta) &= \frac{\langle\nabla\phi^*\big(\nabla\phi(y) + \delta v\big) - y, \delta v\rangle - \mathrm{D}_\phi(x, y) - \mathrm{D}_{\phi^*}\big(\nabla\phi(y) + \delta v, \nabla\phi(y)\big)}{\delta^2}\\ &= \frac{\mathrm{D}_{\phi^*}\big(\nabla\phi(y), \nabla\phi(y) + \delta v\big) - \mathrm{D}_\phi(x, y)}{\delta^2}.\end{aligned}$$

Since ϕ^* is strictly convex, the numerator is strictly increasing. Moreover, as $\mathrm{D}_\phi(x, y) > 0$, it is strictly negative at $\delta = 0$. Therefore, it is either negative for all $\delta > 0$ or it vanishes at a unique $\delta^\star$ as in the statement, which must be the global minimum of ψ_ϕ.

Thus, in this latter case,

$$\langle x - y, v\rangle \leq \inf_{\delta>0} \psi_\phi(\delta) = \frac{\mathrm{D}_\phi(x, y) + \mathrm{D}_{\phi^*}\big(\nabla\phi(y) + \delta^\star v, \nabla\phi(y)\big)}{\delta^\star}, \tag{3.9}$$

which by (3.7) expands to the right-hand side of (3.8).

Finally, if $\operatorname{dom}\phi$ is open, then it follows from [4, Cor. 3.11] that $\mathrm{D}_{\phi^*}\big(\nabla\phi(y), \cdot\big)$ is coercive, and thus the numerator in the expression of $\psi'_\phi(\delta)$ cannot be negative for all δ. □

Note that the right-hand side of (3.8) does depend on x via the parameter $\delta^\star$, as evident by its definition (3.7). This upper bound can be relaxed into a simplified form whenever ϕ has a (strictly positive) *symmetry coefficient* $\alpha = \alpha(\phi) > 0$ as in (2.5). In this case, $\operatorname{dom}\phi$ must be open [3, Prop. 2] and, since $\alpha(\phi) = \alpha(\phi^*)$ [3, Rem. 2(b)], (3.9) can be further expanded into

$$\langle x - y, v\rangle \leq \frac{\mathrm{D}_\phi(x, y) + \frac{1}{\alpha}\,\mathrm{D}_{\phi^*}\big(\nabla\phi(y), \nabla\phi(y) + \delta^\star v\big)}{\delta^\star},$$

which combined with (3.7) results in the following simplified version.

Corollary 3.4. *Consider the setting of Lemma 3.2 with* $\langle x-y, v\rangle > 0$*, and suppose that* ϕ *has symmetry coefficient* $\alpha > 0$*. Then,* $\delta^\star$ *as in (3.7) exists, and one has*

$$\langle x-y, v\rangle \leq \tfrac{1+\alpha}{\alpha}\tfrac{1}{\delta^\star}\,\mathrm{D}_\phi(x,y). \tag{3.10}$$

3.2 Bounding the inner product B_{k+1}

Patterning previous analyses of adaptive stepsizes, our goal is to turn the identity of Lemma 3.1 into a descent-type inequality on some merit function. The bottleneck lies in the inner product term B_{k+1} as in (3.2), which in previous analyses restricted to the Euclidean case was handled via standard Young's or Cauchy-Schwarz bounds.

The Bregman version of Young's inequality given in Lemma 3.2 allows us to replicate these ideas, but with an important caveat. Indeed, the presence of δ within the argument of the Bregman distance D_{ϕ^*} constrains our choice on the parameter, causing a slight departure from the easier Euclidean case in which such complication does not arise. In this subsection, we identify two possible options, each leading to one of the two algorithmic variants B-adaPG and B-adaPG$_\alpha$.

3.2.1 B-adaPG bound

A direct application of the Young's inequality of Lemma 3.2 allows us to bound the inner product B_{k+1} as

$$\begin{aligned} B_{k+1} &\leq \tfrac{\rho_{k+1}}{\delta_{k+1}}\,\mathrm{D}_\phi(x^{k+1}, x^k) \\ &\quad + \tfrac{\rho_{k+1}}{\delta_{k+1}}\,\mathrm{D}_{\phi^*}\big(\nabla\phi(x^k) + \delta_{k+1}\big[H_k(x^k) - H_k(x^{k-1})\big], \nabla\phi(x^k)\big) \end{aligned}$$

for any $\delta_{k+1} > 0$, which in terms of the Lipschitz-like estimate $\Lambda_{k,\delta}$ as in (2.3b) reads

$$\leq \tfrac{\rho_{k+1}}{\delta_{k+1}}\,\mathrm{D}_\phi(x^{k+1}, x^k) + \tfrac{\delta_{k+1}\rho_{k+1}}{2}\Lambda_{k,\delta_{k+1}}\,\Delta_\phi(x^k, x^{k-1}). \tag{3.11}$$

As we will see, the employment of this inequality combined with a specific choice of δ_{k+1} will lead to the stepsize update in the B-adaPG variant.

3.2.2 B-adaPG$_\alpha$ bound

The bound leading to the B-adaPG$_\alpha$ variant follows from the combination of (3.11) and the following lemma, which furnishes a Bregman generalization of the inequality derived in [11, Lem. 2.1].

Lemma 3.5. *Let* $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ *be Legendre, and* $f, g : \mathbb{R}^n \to \overline{\mathbb{R}}$ *be proper, lsc, and convex, with* f *differentiable on* $C \coloneqq \operatorname{int}\operatorname{dom}\phi$ *and* $\operatorname{dom} g \cap C \neq \emptyset$*. Then, denoting* $H_k \coloneqq \nabla\phi - \gamma_k f$*, Bregman proximal gradient iterations (1.1) with stepsizes* $\gamma_k > 0$ *and starting at some* $x^0 \in C$ *satisfy*

$$B_{k+1} \coloneqq \tfrac{\gamma_{k+1}}{\gamma_k}\langle x^{k+1} - x^k, H_k(x^k) - H_k(x^{k-1})\rangle \geq \Delta_\phi(x^{k+1}, x^k) \quad \forall k \geq 1.$$

Proof. Follows by observing that

$$\begin{aligned} &\tfrac{\gamma_{k+1}}{\gamma_k}\big(H_k(x^k) - H_k(x^{k-1})\big) \\ &= \nabla\phi(x^{k+1}) - \nabla\phi(x^k) + \gamma_{k+1}\Big[\overbrace{\tfrac{H_{k+1}(x^k) - \nabla\phi(x^{k+1})}{\gamma_{k+1}}}^{\tilde\nabla g(x^{k+1})} - \overbrace{\tfrac{H_k(x^{k-1}) - \nabla\phi(x^k)}{\gamma_k}}^{\tilde\nabla g(x^k)}\Big], \end{aligned}$$

hence that $B_{k+1} = \Delta_\phi(x^{k+1}, x^k) + \gamma_{k+1}\,\tilde{\mathrm{D}}_g(x^{k+1}, x^k) \geq \Delta_\phi(x^{k+1}, x^k)$. □

We may thus complement (3.11) with a lower bound as

$$\begin{aligned}\Delta_\phi(x^{k+1}, x^k) &\leq B_{k+1} \\ &\leq \tfrac{\rho_{k+1}}{\delta_{k+1}}\,\mathrm{D}_\phi(x^{k+1}, x^k) + \tfrac{\delta_{k+1}\rho_{k+1}}{2}\Lambda_{k,\delta_{k+1}}\,\Delta_\phi(x^k, x^{k-1}).\end{aligned}$$

If ϕ has symmetry coefficient $\alpha > 0$ as in (2.5), then note that

$$\Delta_\phi(x^{k+1}, x^k) \geq (1+\alpha)\,\mathrm{D}_\phi(x^{k+1}, x^k).$$

Combined with the previous inequality, we obtain

$$\Big(1 + \alpha - \tfrac{\rho_{k+1}}{\delta_{k+1}}\Big)\,\mathrm{D}_\phi(x^{k+1}, x^k) \leq \tfrac{\delta_{k+1}\rho_{k+1}}{2}\Lambda_{k,\delta_{k+1}}\,\Delta_\phi(x^k, x^{k-1}),$$

which plugged into (3.11) leads to

$$B_{k+1} \leq \tfrac{(1+\alpha)\delta_{k+1}^2}{(1+\alpha)\delta_{k+1} - \rho_{k+1}}\,\tfrac{\rho_{k+1}}{2}\Lambda_{k,\delta_{k+1}}\,\Delta_\phi(x^k, x^{k-1}), \tag{3.12}$$

holding for any δ_{k+1} such that $(1+\alpha)\delta_{k+1} - \rho_{k+1} > 0$.

3.3 A merit function for B-adaPG

By bounding the inner product term B_{k+1} with (3.11), the inequality (3.5) in Lemma 3.1 reveals that Bregman proximal gradient iterations (1.1) with arbitrary stepsizes $\gamma_k > 0$ satisfy

$$\begin{aligned}&\widehat{\mathcal{U}}_{k+1}(x) - \tfrac{\rho_{k+1}}{\delta_{k+1}}\,\mathrm{D}_\phi(x^{k+1}, x^k) \\ &\leq \widehat{\mathcal{U}}_k(x) - \tfrac{\rho_k}{\delta_k}\,\mathrm{D}_\phi(x^k, x^{k-1}) - \gamma_k(1 + \vartheta_k - \rho_{k+1}\vartheta_{k+1})P_{k-1}(x) \\ &\quad - \left[\Big(1 - \tfrac{\rho_k}{\delta_k}\Big) - \rho_{k+1}(1+\alpha_k)\tfrac{\frac{\delta_{k+1}}{2}\Lambda_{k,\delta_{k+1}} - \vartheta_{k+1}(1-\gamma_k\ell_k)}{\alpha_k}\right]\mathrm{D}_\phi(x^k, x^{k-1})\end{aligned} \tag{3.13}$$

for any $x \in \operatorname{dom}\varphi \cap \operatorname{dom}\phi$, $\vartheta_k \geq 0$, and $\delta_k > 0$, $k \in \mathbb{N}$. Imposing that the multiplying coefficients of $P_{k-1}(x)$ and $\mathrm{D}_\phi(x^k, x^{k-1})$ in the right-hand side of (3.13) are negative amounts to the following two conditions:

$$\vartheta_{k+1}\rho_{k+1} \leq 1 + \vartheta_k \tag{3.14a}$$

and

$$1 - \tfrac{\rho_k}{\delta_k} \geq \rho_{k+1}\tfrac{1+\alpha_k}{\alpha_k}\big[\tfrac{\delta_{k+1}}{2}\Lambda_{k,\delta_{k+1}} - \vartheta_{k+1}(1-\gamma_k\ell_k)\big]. \tag{3.14b}$$

In line with the analyses of [16, 12], a convenient choice for the parameter δ_{k+1} is $\delta_{k+1} = 2\rho_{k+1}$. However, this choice is not feasible in our more general setting. Specifically, selecting ρ_{k+1} to satisfy (3.14b) requires knowledge of the quantity $\Lambda_{k,\delta_{k+1}}$, which generally depends on δ_{k+1} itself (except in the special case where ϕ is quadratic). As a result, setting $\delta_{k+1} = 2\rho_{k+1}$ would create a circular dependency between the two parameters. To complicate things further, note that the left-hand side of (3.14b) indicates that a constraint $\delta_k > \rho_k$ must be in place in order to ensure the existence of a $\rho_{k+1} > 0$ satisfying the inequality.

In order to resolve this circular dependence, we introduce a parameter $\hat{\rho}_{k+1}$ that shall provide an overestimate

$$\hat{\rho}_{k+1} \geq \rho_{k+1} \tag{3.15}$$

while only being based on information available at iteration k. Its explicit value will be revealed shortly after. Then, we may conveniently select $\delta_k = 2\hat{\rho}_k$ and $\vartheta_k = \hat{\rho}_k$ so that (3.14) simplifies as

$$\rho_{k+1}\hat{\rho}_{k+1} \leq 1 + \hat{\rho}_k \quad \text{and} \quad 1 - \tfrac{\rho_k}{2\hat{\rho}_k} \geq \rho_{k+1}\hat{\rho}_{k+1}\tfrac{1+\alpha_k}{\alpha_k}\big[\Lambda_{k,\delta_{k+1}} - (1-\gamma_k\ell_k)\big],$$

that is,

$$\rho_{k+1} \leq \min\left\{\frac{1+\hat{\rho}_k}{\hat{\rho}_{k+1}}, \frac{\alpha_k}{1+\alpha_k}\frac{1+\frac{\hat{\rho}_k-\rho_k}{\hat{\rho}_k}}{2\hat{\rho}_{k+1}\big[\Lambda_{k,2\hat{\rho}_{k+1}} - (1-\gamma_k\ell_k)\big]_+}\right\}. \tag{3.16}$$

Due to (3.15) the term $\frac{\hat{\rho}_k-\rho_k}{\hat{\rho}_k}$ in the numerator of the second update is always nonnegative. Using this, we show in the next lemma that the update rule of B-adaPG always complies with the above bound. In fact, while it is possible to retain the term $\frac{\hat{\rho}_k-\rho_k}{\hat{\rho}_k}$ in the update of B-adaPG, for the sake of a neater expression we opted to omit it at the cost of introducing slight conservatism.

Lemma 3.6. *Suppose that Assumption 2.1 holds, and consider the iterates generated by Bregman proximal gradient iterations (1.1) with γ_k and $\hat{\rho}_k$ selected according to B-adaPG. Then, both (3.15) and (3.16) are satisfied for any $k \in \mathbb{N}$.*

Moreover, denoting

$$\mathcal{U}_k(x) \coloneqq \mathrm{D}_\phi(x, x^k) + \gamma_k(1+\hat{\rho}_k)P_{k-1}(x) + (1 - \tfrac{\rho_k}{2\hat{\rho}_k})\,\mathrm{D}_\phi(x^k, x^{k-1}), \tag{3.17}$$

one has that

$$\mathcal{U}_{k+1}(x) \leq \mathcal{U}_k(x) - \gamma_k(\overbrace{1+\hat{\rho}_k-\rho_{k+1}\hat{\rho}_{k+1}}^{\geq \hat{\rho}_k - \rho_k \geq 0})P_{k-1}(x) \tag{3.18}$$

holds for any $x \in \operatorname{dom}\varphi \cap \operatorname{dom}\phi$ and $k \in \mathbb{N}$.

Proof. The update (2.6) clearly ensures that $\rho_{k+1} \leq \hat{\rho}_{k+1}$ always holds. Note that the second element in the minimum within (3.16) coincides with that in (2.6); moreover, since $\rho_{k+1} \leq \hat{\rho}_{k+1}$ one has that $\rho_{k+1}\hat{\rho}_{k+1} \leq \hat{\rho}_{k+1}^2 = 1 + \rho_k \leq 1 + \hat{\rho}_k$, altogether confirming that the validity of (3.16).

Observe further that the bound $\hat{\rho}_k \geq \rho_k$ ensures that both elements in the minimum of (2.6) are strictly positive, and thus so are the generated stepsizes γ_k. Finally, (3.18) follows from (3.13) with the specified choices of ϑ_k and δ_k. □

3.4 A merit function for B-adaPG$_\alpha$

In case ϕ has a symmetry coefficient $\alpha > 0$, we may leverage the bound (3.12) for B_{k+1}. By doing so, the inequality (3.5) in Lemma 3.1 becomes

$$\begin{aligned}\widehat{\mathcal{U}}_{k+1}(x) &\leq \widehat{\mathcal{U}}_k(x) - \gamma_k(1+\vartheta_k - \rho_{k+1}\vartheta_{k+1})P_{k-1}(x)\\ &- \left\{\tfrac{\alpha_k}{1+\alpha_k} - \rho_{k+1}\left[\tfrac{(1+\alpha)\delta_{k+1}^2}{(1+\alpha)\delta_{k+1}-\rho_{k+1}}\tfrac{1}{2}\Lambda_{k,\delta_{k+1}} - \vartheta_{k+1}(1-\gamma_k\ell_k)\right]\right\}\Delta_\phi(x^k, x^{k-1}).\end{aligned}$$

Once again, we introduce a parameter $\hat{\rho}_{k+1}$ (to be specified later) that satisfies (3.15) whilst based on information available at iteration k. We can conveniently set $\delta_{k+1} = \vartheta_{k+1} = \frac{2}{1+\alpha}\hat{\rho}_{k+1}$, and use the bound $(1+\alpha)\delta_{k+1} - \rho_{k+1} \geq (1+\alpha)\delta_{k+1} - \hat{\rho}_{k+1}$ to simplify the coefficient of $\Lambda_{k,\delta_{k+1}}$. Combined with the fact that $\frac{\alpha_k}{1+\alpha_k} \geq \frac{\alpha}{1+\alpha}$, the inequality simplifies as

$$\begin{aligned}\mathcal{U}^{\alpha}_{k+1}(x) &\coloneqq \mathrm{D}_{\phi}(x, x^{k+1}) + \mathrm{D}_{\phi}(x^{k+1}, x^{k}) + \gamma_{k+1}\left(1 + \tfrac{2}{1+\alpha}\hat{\rho}_{k+1}\right) P_k(x) \\ &\leq \mathcal{U}^{\alpha}_{k}(x) - \gamma_k(1 + \tfrac{2}{1+\alpha}\hat{\rho}_k - \tfrac{2}{1+\alpha}\hat{\rho}_{k+1}\rho_{k+1})P_{k-1}(x) \\ &\quad - \left\{\tfrac{\alpha}{1+\alpha} - \tfrac{2}{1+\alpha}\hat{\rho}_{k+1}\rho_{k+1}\left[\Lambda_{k,\delta_{k+1}} - (1-\gamma_k\ell_k)\right]\right\}\Delta_{\phi}(x^k, x^{k-1}),\end{aligned}$$

where $\mathcal{U}^{\alpha}_{k}(x)$ corresponds to $\widehat{\mathcal{U}}_k(x)$ as in Lemma 3.1 with $\vartheta_k = \frac{2}{1+\alpha}\hat{\rho}_k$. By imposing negativity of the coefficients of $P_{k-1}(x)$ and $\Delta_{\phi}(x^k, x^{k-1})$ as done in the previous subsection, the following analogue of Lemma 3.6 for B-adaPG$_{\alpha}$ is derived.

Lemma 3.7. *Additionally to Assumption 2.1, suppose that ϕ has symmetry coefficient $\alpha > 0$ and consider the iterates generated by Bregman proximal gradient iterations (1.1) with γ_k and $\hat{\rho}_k$ selected according to B-adaPG$_{\alpha}$. Then, denoting*

$$\mathcal{U}^{\alpha}_{k}(x) \coloneqq \mathrm{D}_{\phi}(x, x^{k}) + \mathrm{D}_{\phi}(x^{k}, x^{k-1}) + \gamma_k\left(1 + \tfrac{2}{1+\alpha}\hat{\rho}_k\right) P_{k-1}(x), \tag{3.19}$$

one has that

$$\mathcal{U}^{\alpha}_{k+1}(x) \leq \mathcal{U}^{\alpha}_{k}(x) - \gamma_k\Big(\overbrace{1 + \tfrac{2}{1+\alpha}\hat{\rho}_k - \tfrac{2}{1+\alpha}\hat{\rho}_{k+1}\rho_{k+1}}^{\geq \frac{2}{1+\alpha}(\hat{\rho}_k - \rho_k) \geq 0}\Big) P_{k-1}(x) \tag{3.20}$$

holds for any $x \in \operatorname{dom}\varphi \cap \operatorname{dom}\phi$ and $k \in \mathbb{N}$.

4 Convergence analysis

This section is devoted to proving Theorem 2.5 in its entirety. We first provide some technical lemmas that will be invoked in the proofs. The first one is a direct consequence of Lemmas 3.6 and 3.7, and its statement closely patterns similar results in the Euclidean setting. The simple proof is given in the appendix.

Lemma 4.1. *Suppose that Assumption 2.1 holds, and consider the iterates generated by Bregman proximal gradient iterations (1.1) with γ_k and $\hat{\rho}_k$ selected according to B-adaPG. Then, with $\mathcal{U}_k$ as in (3.17), the following hold for any $x \in \operatorname{dom}\phi$ satisfying $\varphi(x) \leq \inf_{k\in\mathbb{N}}\varphi(x^k)$:*

(i) *$(\mathcal{U}_k(x))_{k\in\mathbb{N}}$ decreases and converges to a finite value.*

(ii) *$P^{\min}_K(x) \leq \frac{\mathcal{U}_0(x)}{\sum_{k=1}^{K+1}\gamma_k}$ for every $K \geq 1$, where $P^{\min}_K(x) \coloneqq \min_{k\leq K} P_k(x)$.*

When ϕ has symmetry coefficient $\alpha > 0$, all remains true for the updates of B-adaPG$_{\alpha}$, with $\mathcal{U}_k \leftarrow \mathcal{U}^{\alpha}_k$ as in (3.19).

Proof. See Appendix A. □

The existence of $x \in \operatorname{int}\operatorname{dom}\phi$ such that $\varphi(x) \leq \varphi(x^k)$ for all k is fundamental for the validity of Lemma 4.1. When $\operatorname{dom}\phi = \mathbb{R}^n$, thus in the Euclidean

case in particular, this is granted. More generally, the possibility of solutions existing merely on the boundary of $\operatorname{dom}\phi$ renders this statement inapplicable. Nevertheless, this result will prove fundamental in demonstrating Theorem 2.5 in its full generality, and particularly the general claim in (2.8).

The same commentary applies to the following lemma, which assumes that the iterates x^k stay bounded away from the boundary of $\operatorname{dom}\phi$. Roughly speaking, under this assumption it ensures that whenever stepsizes drop below a certain threshold, the updates in both (2.6) and (2.7) reduce to $\gamma_{k+1} = \hat{\rho}_{k+1}\gamma_k$, cf. Remark 2.4, and thus increase. This behavior is at the heart of the careful choice of parameters $\vartheta_k = \hat{\rho}_k$ for B-adaPG and $\vartheta_k = \frac{1+\alpha}{2\alpha}\hat{\rho}_k$ for B-adaPG$_\alpha$.

Lemma 4.2. *Suppose that Assumption 2.1 holds, and consider a sequence $\left(x^k\right)_{k\in\mathbb{N}}$ generated by Bregman proximal gradient iterations (1.1) with stepsizes $\gamma_k > 0$. Suppose that $\left(x^k\right)_{k\in\mathbb{N}}$ is contained in a compact set $\mathcal{K} \subset C$, and consider the ratio*

$$\Lambda_{k,\delta_{k+1}} = \frac{2\,\mathrm{D}_{\phi^*}\left(\nabla\phi(x^k) + \delta_{k+1}\left[H_k(x^k) - H_k(x^{k-1})\right], \nabla\phi(x^k)\right)}{\delta_{k+1}^2\,\Delta_\phi(x^k, x^{k-1})}$$

as in (2.3b), where $(\delta_k)_{k\in\mathbb{N}} \subset \mathbb{R}_{++}$ is a bounded sequence. If $\gamma_k \to 0$, then $\Lambda_{k,\delta_{k+1}} \to 1$.

Proof. For notational conciseness, let us denote $u^k := \nabla\phi(x^k) - \nabla\phi(x^{k-1})$ and $v^k := \nabla f(x^k) - \nabla f(x^{k-1})$, so that $H_k(x^k) - H_k(x^{k-1}) = u^k - \gamma_k v^k$. For any $k \in \mathbb{N}$ there exists ξ^k on the line segment between $\nabla\phi(x^k)$ and $\nabla\phi(x^k) + \delta_{k+1}(H_k(x^k) - H_k(x^{k-1}))$ and η^k on the line segment between $\nabla\phi(x^k)$ and $\nabla\phi(x^{k-1})$ such that

$$\begin{aligned}
\Lambda_{k,\delta_{k+1}} &= \frac{2\,\mathrm{D}_{\phi^*}\left(\nabla\phi(x^k) + \delta_{k+1}\left[H_k(x^k) - H_k(x^{k-1})\right], \nabla\phi(x^k)\right)}{\delta_{k+1}^2\,\Delta_\phi(x^k, x^{k-1})} \\
&= \frac{\langle \nabla^2\phi^*(\xi_k)\left[H_k(x^k) - H_k(x^{k-1})\right], H_k(x^k) - H_k(x^{k-1})\rangle}{\langle \nabla^2\phi^*(\eta_k)u^k, u^k\rangle} \\
&= \frac{\langle \nabla^2\phi^*(\xi_k)u^k, u^k\rangle}{\langle \nabla^2\phi^*(\eta_k)u^k, u^k\rangle} - 2\gamma_k\frac{\langle \nabla^2\phi^*(\xi_k)u^k, v^k\rangle}{\langle \nabla^2\phi^*(\eta_k)u^k, u^k\rangle} + \gamma_k^2\frac{\langle \nabla^2\phi^*(\xi_k)v^k, v^k\rangle}{\langle \nabla^2\phi^*(\eta_k)u^k, u^k\rangle}.
\end{aligned} \tag{4.1}$$

Since ϕ is twice continuously differentiable with $\nabla^2\phi \succ 0$ on C, and $\mathcal{K} \subset C$ is compact and convex, one has that

$$L_{\phi,\mathcal{K}} := \sup_{\mathcal{K}}\|\nabla^2\phi\| = \left(\inf_{\nabla\phi(\mathcal{K})}\lambda_{\min}(\nabla^2\phi^*)\right)^{-1}$$

is finite ($L_{\phi,\mathcal{K}}$ being the Lipschitz modulus of $\nabla\phi$ on $\mathcal{K}$). As such, one has that $\langle \nabla^2\phi^*(\eta_k)u^k, u^k\rangle \geq L_{\phi,\mathcal{K}}^{-1}\|u^k\|^2$ for all k. Moreover, letting $L_{f,\mathcal{K}}^{\phi}$ denote a smoothness modulus of f relative to ϕ on $\mathcal{K}$, ensured to exist by Assumption 2.1(ii), we infer from [1, Prop. 2.5(ii)] that $\|v^k\| \leq L_{f,\mathcal{K}}^{\phi}L_{\phi,\mathcal{K}}\|u^k\|$ holds for all k. Therefore,

$$\begin{aligned}
|\Lambda_{k,\delta_{k+1}} - 1| \leq{}& \|\nabla^2\phi^*(\xi_k) - \nabla^2\phi^*(\eta_k)\|L_{\phi,\mathcal{K}} + 2\gamma_k L_{f,\mathcal{K}}^{\phi}L_{\phi,\mathcal{K}}^2\|\nabla^2\phi^*(\xi_k)\| \\
&+ \gamma_k^2 L_{\phi,\mathcal{K}}^3(L_{f,\mathcal{K}}^{\phi})^2\|\nabla^2\phi^*(\xi_k)\|.
\end{aligned}$$

If $\gamma_k \to 0$, then eventually $\gamma_k < \frac{1}{L_{f,\mathcal{K}}^{\phi}}$ and standard results ensure that $\left(x^k\right)_{k\in\mathbb{N}}$ converges to some point $x \in \mathcal{K}$. In this case, $\left(\xi^k\right)_{k\in\mathbb{N}}$ and $\left(\eta^k\right)_{k\in\mathbb{N}}$ converge to $\nabla\phi(x)$ (the former sequence because $(\delta_k)_{k\in\mathbb{N}}$ is bounded), and by continuity of $\nabla^2\phi^*$ on $\mathcal{K}$ the right-hand side in the above inequality vanishes. □

The following is another auxiliary result that considers iterates (1.1) that remain bounded away from the boundary of $\operatorname{dom}\phi$. It essentially states that whenever a subsequence $\left(x^k\right)_{k\in K}$ converges to a solution, then also the shifted subsequence $\left(x^{k+1}\right)_{k\in K}$ does provided that the corresponding stepsizes do not diverge.

Lemma 4.3. *Additionally to Assumption 2.1 suppose that* $\arg\min_C \varphi \neq \emptyset$. *Let a sequence* $\left(y^k\right)_{k\in\mathbb{N}}$ *contained in* C *and converging to a point* $y^\star \in \arg\min_C \varphi$ *be fixed, and for every* $k \in \mathbb{N}$ *let*

$$\bar{y}^k := \arg\min_w \left\{ g(w) + \langle \nabla f(y^k), w - y^k\rangle + \tfrac{1}{\gamma_{k+1}} \mathrm{D}_\phi(w, y^k) \right\} \tag{4.2}$$

where $(\gamma_k)_{k\in\mathbb{N}}$ *is a bounded sequence of strictly positive stepsizes. Then,* $\bar{y}^k \to y^\star$ *and* $\gamma_{k+1}(\varphi(\bar{y}^k) - \min\varphi) \to 0$.

Proof. Consider the Bregman proximal operator $\overleftarrow{\operatorname{prox}}^\phi_g : \operatorname{int}\operatorname{dom}\phi \rightrightarrows \operatorname{dom}\phi$ defined as

$$\overleftarrow{\operatorname{prox}}^\phi_g(y) := \arg\min_{w\in\mathbb{R}^n} \{g(w) + \mathrm{D}_\phi(w, y)\}$$

(note that $\overleftarrow{\operatorname{prox}}^\phi_g(y) \in \operatorname{int}\operatorname{dom}\phi \subseteq \operatorname{dom}\phi$ for any $y \in \operatorname{int}\operatorname{dom}\phi$, owing to Assumption 2.1). The mapping $\overleftarrow{\operatorname{prox}}^\phi_{\gamma_{k+1}g} \circ \nabla\phi^*$ is $\nabla\phi$-firmly-nonexpansive, in the sense that

$$\Delta_\phi(\bar{y}_1, \bar{y}_2) \leq \langle \eta_1 - \eta_2, \bar{y}_1 - \bar{y}_2\rangle \tag{4.3}$$

holds for any $\eta_i \in \mathbb{R}^2$ and $\bar{y}_i = \overleftarrow{\operatorname{prox}}^\phi_{\gamma_{k+1}g} \circ \nabla\phi^*(\eta_i)$, $i = 1, 2$; a proof of this fact can be found in [26, Lem. 4.2] (see also [27, Thm. 4.9] for its *equivalence* to convexity of g). Note that (4.2) can equivalently be written as

$$\bar{y}^k = \overleftarrow{\operatorname{prox}}^\phi_{\gamma_{k+1}g} \circ \nabla\phi^*(H_{k+1}(y^k)),$$

where we remind that $H_{k+1} = \nabla\phi - \gamma_{k+1}\nabla f$. Using $\nabla\phi$-firm nonexpansiveness and recalling that $y^\star = \overleftarrow{\operatorname{prox}}^\phi_{\gamma_{k+1}g} \circ \nabla\phi^*(H_{k+1}(y^\star))$ we have

$$\begin{aligned}
\Delta_\phi(\bar{y}^k, y^\star) &\overset{(4.3)}{\leq} \langle H_{k+1}(y^k) - H_{k+1}(y^\star), \bar{y}^k - y^\star\rangle \\
&\leq \Big(\|\nabla\phi(y^k) - \nabla\phi(y^\star)\| + \gamma_{k+1}\|\nabla f(y^k) - \nabla f(y^\star)\| \Big) \|\bar{y}^k - y^\star\|,
\end{aligned}$$

which vanishes as $k \to \infty$. Then, by the essential smoothness of h, it follows that $\bar{y}^k \to y^\star$. Moreover, by subgradient inequality, for every $k \in \mathbb{N}$ it holds that

$$\begin{aligned}
0 \leq \gamma_{k+1}(\varphi(\bar{y}^k) - \min\varphi) &= \gamma_{k+1}(f(\bar{y}^k) + g(\bar{y}^k) - \min\varphi) \\
&\leq \gamma_{k+1}(f(\bar{y}^k) - f(y^\star)) \\
&\quad - \langle \nabla\phi(y^k) - \gamma_{k+1}\nabla f(y^k) - \nabla\phi(\bar{y}^k), y^\star - \bar{y}^k\rangle.
\end{aligned}$$

The proof then follows from continuity of f and the fact that the inner product vanishes, since both y^k and $\bar{y}^k$ converge to $y^\star$. □

In the remainder of this section, we delve into the proof of Theorem 2.5. We begin by establishing Theorem 2.5(i), which addresses the unconstrained-like setting where all three lemmas introduced above are directly applicable. Despite the simplifying assumptions, this part of the proof is the most technically involved. Once established, the general result in (2.8), and ultimately Theorem 2.5(ii), will follow as comparatively simpler corollaries.

4.1 Proof of Theorem 2.5(i)

This entire subsection is dedicated to proving the following result, which covers Theorem 2.5(i) under more general conditions; this higher degree of generality will serve as a fundamental intermediate step in the treatment of the more general setting.

Theorem 4.4. *Additionally to Assumption 2.1, suppose that*

$$\text{there exists } x \in C \text{ such that } \varphi(x) \leq \inf_{k\in\mathbb{N}} \varphi(x^k) \tag{4.4}$$

holds for the iterates x^k generated by B-adaPG.. Then, the sequence $\left(x^k\right)_{k\in\mathbb{N}}$ converges to a solution $x^\star \in \arg\min_C \varphi$ and $\varphi(x^k) \to \inf_{\overline{C}} \varphi$.

When ϕ has symmetry coefficient $\alpha > 0$, the same is true for B-adaPG$_\alpha$.

The fact that this result subsumes Theorem 2.5(i) is obvious by observing that the validity of Theorem 4.4 implies that any x complying with (4.4) must necessarily belong to $\arg\min_{\overline{C}} \varphi \cap C$. The proof of Theorem 4.4 will be carried out via intermediate claims. In what follows, we consider $\mathcal{U}_k$ as in (3.17) in case of B-adaPG; under the needed symmetry assumption, the same proof applies to B-adaPG$_\alpha$ as well by simply replacing $\mathcal{U}_k \leftarrow \mathcal{U}_k^\alpha$.

Claim 1. *There exists a compact set $\mathcal{K} \subset C$ containing all the iterates x^k.*

Proof. In this case, Lemma 4.1(i) implies that

$$\mathrm{D}_\phi(x, x^k) \leq \mathcal{U}_k(x) \leq \mathcal{U}_0(x)$$

holds for any $k \in \mathbb{N}$. The assertion then follows from Fact 1.2(i). □

Claim 2. $\gamma_k \not\to 0$.

Proof. To arrive to a contradiction, suppose that $\gamma_k \to 0$. Then, Lemma 4.2 implies that eventually the quantity $[\Lambda_{k,\delta} - (1 - \gamma_k \ell_k)]_+$ appearing in both (2.6) and (2.7) vanishes. Since $\left(x^k\right)_{k\in\mathbb{N}}$ is bounded and bounded away from the boundary of C, regardless of whether or not ϕ has a symmetry coefficient $\alpha > 0$, it holds that $\inf_{k\in\mathbb{N}} \alpha_k > 0$; furthermore, for both B-adaPG and B-adaPG$_\alpha$ note that $\hat{\rho}_{k+1} \leq \sqrt{1+\rho_k} \leq \sqrt{1+\hat{\rho}_k}$, implying that $\hat{\rho}_k \leq \frac{1+\sqrt{5}}{2}$ holds for any k. As such, in both (2.6) and (2.7) eventually the second element in the minimum is infinite, implying that the update reduces to $\gamma_{k+1} = \hat{\rho}_{k+1}\gamma_k$ is divergent, a contradiction. □

Claim 3. $\inf_{k\in\mathbb{N}} \varphi(x^k) = \inf_{\overline{C}} \varphi$ *(in particular, necessarily $x \in \arg\min_C \varphi$).*

Proof. Having shown that $\gamma_k \not\to 0$, it follows that $\sum_{k\in\mathbb{N}} \gamma_k = \infty$. Lemma 4.1(ii) then implies that $\inf_{k\in\mathbb{N}} \varphi(x^k) = \varphi(x)$. From the arbitrariness of the point x as in (4.4) we conclude that necessarily $x \in \arg\min_C \varphi \subseteq \arg\min_{\overline{C}} \varphi$, with inclusion holding by virtue of [5, Prop. 11.1(iv)]. Hence $\inf_{k\in\mathbb{N}} \varphi(x^k) = \inf_{\overline{C}} \varphi$. □

Claim 4. *There exists exactly one optimal limit point $x^\star \in \arg\min_C \varphi$.*

Proof. The existence of an optimal limit point is guaranteed by the previous claim, since $\left(x^k\right)_{k\in\mathbb{N}}$ is bounded (and bounded away from the boundary of C) and φ is lsc. Consider two optimal limit points $\bar{x}_1$ and $\bar{x}_2$, and two corresponding subsequences $\left(x^k\right)_{k\in K_1} \to \bar{x}_1$ and $\left(x^k\right)_{k\in K_2} \to \bar{x}_2$. Notice that

$$\mathcal{U}_k(\bar{x}_1) - \mathcal{U}_k(\bar{x}_2) = \mathrm{D}_\phi(\bar{x}_1, x^k) - \mathrm{D}_\phi(\bar{x}_2, x^k)$$

converges to some finite value U, because both $\mathcal{U}_k(\bar{x}_1)$ and $\mathcal{U}_k(\bar{x}_2)$ are convergent. Considering the limits along $k \in K_1$ and $k \in K_2$ yields that

$$U = \mathrm{D}_\phi(\bar{x}_1, \bar{x}_1) - \mathrm{D}_\phi(\bar{x}_2, \bar{x}_1) = \mathrm{D}_\phi(\bar{x}_1, \bar{x}_2) - \mathrm{D}_\phi(\bar{x}_2, \bar{x}_2),$$

hence that $-\mathrm{D}_\phi(\bar{x}_2, \bar{x}_1) = \mathrm{D}_\phi(\bar{x}_1, \bar{x}_2)$. Thus $\Delta_\phi(\bar{x}_1, \bar{x}_2) = 0$, implying that $\bar{x}_1 = \bar{x}_2$. □

In light of the previous claim, the proof of Theorem 4.4 is completed once we show that $x^k \to x^\star$. To this end, owing to the fact that $\mathrm{D}_\phi(x^\star, x^k) \leq \mathcal{U}_k(x^\star)$ it will suffice that $U \coloneqq \lim_{k\to\infty} \mathcal{U}_k(x^\star)$ is zero, (existence and finiteness of U is ensured by Lemma 4.1(i)).

- Let us consider a subsequence $\left(x^k\right)_{k\in K}$ converging to $x^\star$. If the corresponding subsequence $(\gamma_{k+1})_{k\in K}$ is bounded, then it follows from Lemma 4.3 that $x^{k+1} \to x^\star$ and $\gamma_{k+1}P_k(x^\star) \to 0$ as well, and from the expression of $\mathcal{U}_k$ together with the fact that $\hat{\rho}_{k+1}$ is bounded and $\inf_k \alpha_k > 0$, it is clear that $\mathcal{U}_{k+1}(x^\star) \to 0$ as $K \ni k \to \infty$.
- In what follows, let us instead consider the complementary case in which for any subsequence x^k converging to $x^\star$ the correponding subsequence of stepsizes $(\gamma_{k+1})_{k\in K}$ is divergent. In this case, note that since $\gamma_{k+1}P_k(x^\star) \leq \mathcal{U}_{k+1}(x^\star) \leq \mathcal{U}_0(x^\star)$, we have

$$\begin{aligned}\left(x^k\right)_{k\in K} \to x^\star &\Leftrightarrow (\gamma_{k+1})_{k\in K} \to \infty \Leftrightarrow (P_k(x^\star))_{k\in K} \to 0\\ &\Rightarrow (\gamma_k)_{k\in K} \to \infty\\ &\Leftrightarrow \left(x^{k-1}\right)_{k\in K} \to x^\star \Leftrightarrow (P_{k-1}(x^\star))_{k\in K} \to 0,\end{aligned} \tag{4.5}$$

  where the right implication follows from the fact that $\gamma_{k+1} \leq \hat{\rho}_{k+1}\gamma_k \leq \rho_{\max}\gamma_k$, for some $\rho_{\max} \leq \frac{1+\sqrt{5}}{2}$. Since an optimal limit exists, it follows that $\sup_{k\in\mathbb{N}} \gamma_k = \infty$.

  In what follows, we construct a specific subsequence $K \coloneqq \{k_0, k_1, \ldots\}$ along which $(\gamma_k)_{k\in K}$ diverges. In doing so, we expand upon the arguments in the proof of [12, Thm. 2.3] to account for the complications of the non-Euclidean setting investigated here. For $i \geq 0$ let

$$k_{i+1} = \min\left\{k \geq k_i \mid \gamma_k \geq \rho_{\max}\gamma_{k_i}\right\}. \tag{4.6}$$

  Then, $(\gamma_{k_i})_{i\in\mathbb{N}} \to \infty$, which by (4.5) implies that

$$\lim_{i\to\infty} x^{k_i - s} = x^\star \quad \forall s = 1, 2, 3 \tag{4.7}$$

  (in fact, for any $s \in \mathbb{N}$). In light of Lemma 4.1(i) we have that

$$U = \lim_{k\to\infty} \mathcal{U}_k = \lim_{i\to\infty} \mathcal{U}_{k_i-1}(x^\star) = \lim_{i\to\infty} \gamma_{k_i-1}(1 + \vartheta_{k_i-1})P_{k_i-2}(x^\star). \tag{4.8}$$

We proceed to show that $\gamma_{k-1}P_{k-2}(x^\star)$ converges to zero along the same subsequence. For every $i \in \mathbb{N}$, note that

$$\rho_{k_i} > 1 \quad \text{and} \quad k_{i-1} \leq k_i - 1 \tag{4.9a}$$

by minimality of k_i and the fact that $\rho_{\max} > 1$, hence that

$$\text{either} \quad \rho_{k_i-1} > 1 \quad \text{or} \quad \big(k_i - 1 \notin K \text{ and thus } k_{i-1} \leq k_i - 2\big). \tag{4.9b}$$

These combined imply also that

$$\rho_{k_i-1} \geq \rho_{\max}^{-1}. \tag{4.9c}$$

Indeed, if not, then $\rho_{k_i-1} < \rho_{\max}^{-1} < 1$, implying by (4.9b) that $k_{i-1} \leq k_i - 2$; this would lead to the contradiction

$$\rho_{\max}\gamma_{k_i-1} \overset{(4.6)}{\leq} \gamma_{k_i} \leq \rho_{\max}\gamma_{k_i-1} = \rho_{\max}\rho_{k_i-1}\gamma_{k_i-2} \overset{\perp}{<} \gamma_{k_i-2} \leq \rho_{\max}\gamma_{k_i-1},$$

where "$\perp$" marks where the contradictory inequality is used, and the last inequality follows both in case $k_i - 2 = k_{i-1}$ (since $\rho_{\max} > 1$) or when $k_{i-1} < k_i - 2 < k_i$ (from minimality in the definition of k_i). Therefore, (4.9c) holds true, implying in particular that the sequence

$$\Big(\Lambda_{k_i-2,2\hat{\rho}_{k_i-1}}\Big)_{i\in\mathbb{N}} \quad \text{is bounded,} \tag{4.10}$$

for otherwise the second term in either (2.6) or (2.7) would vanish, owing to the fact that $\hat{\rho}_k$ is lower bounded by $\sqrt{1/2}$ and similarly α_k is bounded away from zero.

Let $v^k \coloneqq H_{k-1}(x^{k-1}) - H_{k-1}(x^k) \in \partial\varphi(x^k)$, and observe that

$$\begin{aligned}
\gamma_k P_{k-1}(x^\star) &\leq \rho_k \langle x^\star - x^{k-1}, -v^{k-1}\rangle \\
&\leq \tfrac{\rho_k}{2\hat{\rho}_k} \mathrm{D}_\phi(x^\star, x^{k-1}) + \tfrac{\rho_k}{2\hat{\rho}_k} \mathrm{D}_{\phi^*}\big(\nabla\phi(x^{k-1}) - 2\hat{\rho}_k v^{k-1}\big), \nabla\phi(x^{k-1})\big) \\
&= \tfrac{\rho_k}{2\hat{\rho}_k} \mathrm{D}_\phi(x^\star, x^{k-1}) + 2\rho_k\hat{\rho}_k^2\Lambda_{k-1,2\hat{\rho}_k}\, \Delta_\phi(x^{k-1}, x^{k-2}) \\
&\leq \tfrac{1}{2} \mathrm{D}_\phi(x^\star, x^{k-1}) + 2\rho_{\max}^3\Lambda_{k-1,2\hat{\rho}_k}\, \Delta_\phi(x^{k-1}, x^{k-2})
\end{aligned}$$

holds for any k, where we used the Bregman-Young inequality (3.6) with parameter $\delta = \hat{\rho}_k$ in the second inequality, and $\rho_k \leq \hat{\rho}_k \leq \rho_{\max}$ in the last one. Recall that $\big(x^{k_i-3}\big)_{i\in\mathbb{N}} \to x^\star$ by (4.7). Thus,

$$\gamma_{k_i-1}P_{k_i-2}(x^\star) \leq \underbrace{\tfrac{1}{2}\mathrm{D}_\phi(x^\star, x^{k_i-2})}_{\to 0 \text{ by } (4.7)} + 2\rho_{\max}^3 \underbrace{\Lambda_{k_i-2,2\rho_{k_i-1}}}_{\text{bounded}} \underbrace{\Delta_\phi(x^{k_i-2}, x^{k_i-3})}_{\to 0 \text{ by } (4.7)}.$$

Using this in (4.8) and noting that θ_{k_i-1} being a multiple of $\hat{\rho}_{k_i-1}$ in both B-adaPG and B-adaPG$_\alpha$ is bounded, completes the proof.

4.2 Proof of Eq. (2.8)

In this subsection we prove that the iterates generated by either B-adaPG or B-adaPG$_\alpha$ in the generality of Assumption 2.1 (in addition to ϕ having a symmetry coefficient $\alpha > 0$ in the latter case) are such that $\inf_{k\in\mathbb{N}} \varphi(x^k) = \inf_{\overline{C}} \varphi$.

To see this, contrary to the claim suppose that $\inf_{k\in\mathbb{N}} \varphi(x^k) > \inf_{\overline{C}} \varphi$. Then, since $\inf_{\overline{C}} \varphi = \inf_C \varphi$ by [5, Prop. 11.1(iv)], there exists $x \in C$ such that $\varphi(x^k) \geq \varphi(x)$ holds for all $k \in \mathbb{N}$. Invoking Theorem 4.4 yields a contradiction.

4.3 Proof of Theorem 2.5(ii)

In this subsection we prove that whenever Assumptions 2.1 and 2.3 are satisfied, the iterates generated by either B-adaPG are bounded and admit exactly one optimal limit point.[2] We can without loss of generality assume that $\arg\min_C \varphi = \emptyset$, for otherwise a stronger result is already covered by Theorem 2.5(i).

Under Assumption 2.3, $\operatorname{dom}\phi = \overline{C}$, and therefore in Lemma 4.1(i) one can take any $x \in \arg\min_{\overline{C}} \varphi$. In particular, $\mathrm{D}_\phi(x, x^k) \leq \mathcal{U}_k(x) \leq \mathcal{U}_0(x)$ holds for any k. Since $\mathrm{D}_\phi(x, \cdot)$ is level bounded by Assumption 2.3(ii), boundedness of $\left(x^k\right)_{k\in\mathbb{N}}$ follows. Moreover, we know from (2.8) that $\varphi(x^k) \to \inf_{\overline{C}} \varphi$, and therefore an optimal limit point $x^\star$ exists. To assess its uniqueness, we can argue similarly to the proof of Claim 4, with the minor catch that now such limit points are on the boundary of C. Considering two optimal limit points $\bar{x}_1$ and $\bar{x}_2$, and two corresponding subsequences $\left(x^k\right)_{k\in K_1} \to \bar{x}_1$ and $\left(x^k\right)_{k\in K_2} \to \bar{x}_2$, we still have that

$$\mathcal{U}_k(\bar{x}_1) - \mathcal{U}_k(\bar{x}_2) = \mathrm{D}_\phi(\bar{x}_1, x^k) - \mathrm{D}_\phi(\bar{x}_2, x^k)$$

converges to some finite value U. Again by considering the limit along the two subsequences $k \in K_1$ and $k \in K_2$, Assumption 2.3(i) yields that

$$\lim_{k\in K_2} \mathrm{D}_\phi(\bar{x}_1, x^k) = 0 = \lim_{k\in K_1} \mathrm{D}_\phi(\bar{x}_2, x^k),$$

which by [23, Thm. 2.4] implies that $\bar{x}_1 = \bar{x}_2$.

5 Numerical experiments

In this section, we evaluate the performance of the proposed algorithms on a series of standard simulation problems. Except for the Euclidean simulations of Section 5.5 that exploit available Julia code,[3] all experiments were conducted using MATLAB R2022b.

In each test problem, we compare only those algorithms that are compatible with the problem's structure and domain (see Table 1). In the convergence plots we report the cost against the number of calls to the (Bregman) proximal gradient oracle; except for ABPG-g, in all compared algorithms this coincides with the iteration count. For better visualization and comparison across different methods, the cost profiles are normalized as $\frac{\varphi(x^k)-\min_{\overline{C}}\varphi}{\varphi(x^0)-\min_{\overline{C}}\varphi}$. The value of $\min_{\overline{C}}\varphi$ is retrieved numerically by running Bregman proximal gradient with linesearch B-PG-ls starting at the best iterate attained by all the tested algorithms.[4]

We also plot the stepsizes in a window of consecutive iterations for all adaptive Bregman methods. In test problems where a global relative smoothness constant L_f^ϕ is available, the stepsize plots are normalized by $1/L_f^\phi$.

[2] As commented after Theorem 2.5, there is no loss of generality in considering only B-adaPG and disregard the claim for B-adaPG$_\alpha$.

[3] https://github.com/pylat/adaptive-proximal-algorithms [12, §4]

[4] B-PG-ls was selected because it is the only Bregman method among those considered that guarantees a decrease in the cost function at every iteration.

		L_f^ϕ-smad	ϕ str cvx	$\alpha(\phi) > 0$	$C = \mathbb{R}^n$
B-adaPG					
B-adaPG$_\alpha$				✗	
B-PG-ls					
BaGRAAL	[25]		✗		
ABPG-g	[10]	✗			
PG-ls					✗
adaPG	[12]				✗
adaPG$^{1,\frac{1}{2}}$	[11]				✗

Table 1: List of algorithms used in the numerical experiments of this section and their standing requirements. The term L_f^ϕ-smad, short for L_f^ϕ-*smooth adaptable*, is borrowed from [6] to denote global smoothness of f relative to ϕ with (known) constant L_f^ϕ.

5.1 Compared algorithms

Our stepsize selection B-adaPG and is compared against other Bregman methods and, when applicable, the variant B-adaPG$_\alpha$ and Euclidean strategies. A list of all the algorithms is synopsized in Table 1, together with a schematic summary of the requirements for each. More detailed descriptions are provided in the following subsections.

5.1.1 Proposed adaptive methods (B-adaPG and B-adaPG$_\alpha$)

The adaptive stepsize selection B-adaPG is tested on all problems, as the generality of Assumption 2.1 suffices for its applicability. The variant B-adaPG$_\alpha$ is only tested on those instances in which ϕ has a symmetry coefficient $\alpha(\phi) > 0$.

The initial stepsizes are chosen following the strategy proposed in [12, §2.1.1] for the Euclidean setting, noting that the same advantages extend naturally to the more general Bregman framework considered here. We first generate a trial point $\tilde{x}$ by performing a single Bregman proximal gradient step from the initial point x^0, using a stepsize γ_{init} (chosen as $\gamma_{\text{init}} = 1/L_f^\phi$ whenever a global smoothness modulus L_f^ϕ exists). Then, using x^0 and $\tilde{x}$, we compute a local relative smoothness constant ℓ_0 via (2.3a), and use its reciprocal $\gamma_0 = \frac{1}{\ell_0}$ as a refined initial stepsize. If γ_0 is significantly smaller than γ_{init} (say, $\gamma_0 < 0.1\gamma_{\text{init}}$), we reset γ_0 to γ_{init} and repeat the initialization procedure until a reasonable stepsize γ_0 is obtained. We then proceed to select γ_{-1} small enough such that $\gamma_0\hat{\rho}_0 \geq 1/2\ell_0$ (see [12, §2.1.1] for details).

The overhead of gradient evaluations caused by this selection is fairly accounted for in the plots. The same initialization is also chosen for the linesearch methods described next.

5.1.2 Linesearch methods (B-PG-ls and PG-ls)

B-PG-ls is a standard Bregman proximal gradient method equipped with a linesearch procedure. It is applicable to any problem satisfying Assumption 2.1 (in fact, even when ϕ is not twice differentiable). PG-ls denotes its Euclidean counterpart, which applies in the unconstrained setting $C = \mathbb{R}^n$.

At each iteration, both methods perform a tentative update and then evaluate whether the objective function has decreased sufficiently: if the condition is met, the update is accepted and the next iteration proceeds; otherwise, the stepsize is reduced and a new trial is initiated. To reduce the number of failed attempts, the trial stepsize is initialized close to the last accepted one. In our experiments, we warm-start the stepsize as 1.2 times the previously accepted value. This modest increase helps avoid overly conservative behavior and significantly improves performance by enabling the stepsize to recover from previously small values.

Note that each iteration of linesearch-based methods involves additional function evaluations to determine an acceptable stepsize. These overheads are *not* reflected in our metrics which only count the number of gradient evaluations; remarkably, even without factoring in the additional function evaluations incurred by linesearch, our method consistently achieves superior performance.

5.1.3 Bregman adaptive Golden ratio algorithm (BaGRAAL)

The adaptive method BaGRAAL proposed in [25, Alg. 3] is the Bregman extension of aGraal [15]. It applies to the more general setting of variational inequalities. BaGRAAL is applicable to problems where the Bregman kernel is strongly convex. At each iteration, the algorithm computes an adaptive stepsize based on a local Lipschitz estimate (in our notation):

$$\gamma_{k+1} = \min\left\{\rho\gamma_k,\ \frac{\sigma_\phi \nu \rho_k}{4\gamma_k}\cdot\frac{\|x^k - x^{k-1}\|^2}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|^2},\ \gamma_{\max}\right\},$$

where $\rho_{k+1} = \frac{\gamma_{k+1}\nu}{\gamma_k}$, $\rho \in [1, \frac{1}{\nu} + \frac{1}{\nu^2}]$ and σ_ϕ is the strong convexity parameter of the Bregman kernel ϕ. We refer to [25, Alg. 3] for details of the iterates. Following the choices made in [25], we used $\nu = 1.5$, $\rho = \frac{1}{\nu} + \frac{1}{\nu^2}$, and maximum stepsize $\gamma_{\max} = 10^6$. As suggested in [25], the initial stepsize γ_0 is determined by introducing a small random perturbation to the starting point x^0 to obtain a nearby point $\bar{x}^0$, and then computing the local Lipschitz estimate $L_0 = \frac{\|\nabla f(x^0) - \nabla f(\bar{x}^0)\|_2}{\|x^0 - \bar{x}^0\|_2}$; the initial stepsize is then set as $\gamma_0 = \frac{1}{L_0}$.

In all simulations, this method performs consistently worse compared to the other adaptive Bregman algorithms. On the one hand, this can be attributed to its broader applicability beyond minimization problems; on the other, we believe the culprit lies in its reliance on a *Lipschitz*-based stepsize update, computed as a ratio of *Euclidean norms*, while the actual updates are carried out in the Bregman geometry. In contrast, the proposed B-adaPG and B-adaPG$_\alpha$ schemes leverage *purely Bregman-based estimates*, providing a more faithful description of the problem landscape and iterations updates.

5.1.4 Accelerated BPG with gain adaptation (ABPG-g)

The ABPG-g algorithm is an adaptive variant of the ABPG method, both proposed in [10]. These methods exploit the so-called *triangle scaling property* under Bregman geometry to achieve a convergence rate of $\mathcal{O}(k^{-\gamma})$ via an extrapolation step, where $\gamma \in (0, 2]$ is known as the *triangle scaling exponent* (TSE). This class of methods is applicable when the global relative smoothness constant exists and is known.

Compared to the ABPG method, ABPG-g achieves faster convergence by enforcing an optimal exponent $\gamma_{\text{in}} = 2$, enabled by dynamically adjusting a certain "gain" coefficient. For this reason, we only compare against the latter. The dynamic adjustment is validated via a linesearch process. Differently from B-PG-ls and PG-ls which only involve additional cost evaluations, every failed attempt of the backtracking in ABPG-g incurs an extra full Bregman proximal gradient computation, which is accounted for in the cost plots. We refer to [10, Alg. 3] for further details of the iteration process, where the parameters are here set as $\gamma = 2$, $\rho = 1.1$, and $G_{\min} = 10^{-3}$.

5.2 Unconstrained minimization with Hessian norm growing as a polynomial

As a benchmark to test all algorithms in Table 1, we consider the problem proposed in [14, §2] of minimizing a smooth function whose Hessian grows polynomially in norm. The problem is formulated as

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \tfrac{1}{4}\|Ax-b\|_4^4 + \tfrac{1}{2}\|Cx-d\|_2^2, \tag{5.1}$$

where $A, C \in \mathbb{R}^{m\times n}$ are nonzero matrices, and $b, d \in \mathbb{R}^m$. The cost function is not smooth relative to the Euclidean kernel $\mathfrak{j}$ (i.e., its gradient is not globally Lipschitz differentiable); instead, it is smooth relative to

$$\phi(x) = \tfrac{1}{4}\|x\|_2^4 + \tfrac{1}{2}\|x\|_2^2, \tag{5.2}$$

with modulus

$$L_f^{\phi} = 3\|A\|^4 + 6\|A\|^3\|b\|_2 + 3\|A\|^2\|b\|_2^2 + \|C\|^2,$$

see [14, p. 339]. This kernel ϕ has a symmetry coefficient $\alpha(\phi) = 2 - \sqrt{3}$ [19, Tab. 1 and Thm. 5.2]. On the one hand, since the minimization is carried over the whole space $\mathbb{R}^n$ (as opposed to a proper convex subset $\overline{C}$), the problem can be addressed with standard (proximal) gradient iterations with suitably chosen stepsizes. On the other hand, the smoothness relative to ϕ and the absence thereof relative to $\mathfrak{j}$ indicate that employing Bregman algorithms exploiting this tailored kernel should prove beneficial. Our simulations confirm this intuition, demonstrating the utility of Bregman algorithms even in the unconstrained setting.

The matrices A, C are generated with independent identically distributed entries drawn from the uniform distribution on $[0, 1]$, and the corresponding response vectors b, d are constructed by adding scaled uniform noise to the exact linear outputs. Comparisons were conducted across problems of varying sizes using synthetic data.

As evident from Fig. 5.2.1 (top row), B-adaPG and B-adaPG$_\alpha$ emerge as clear winners in being able to adjust the stepsizes more effectively than with the trial-and-error process of the linesearch. The slow convergence of the accelerated algorithm ABPG-g owes to the high inner iteration cost for adjusting the parameters, which involves calls to the Bregman proximal gradient oracle (the method is the fastest when measured purely in terms of iteration count).

The bottom row in Fig. 5.2.1 illustrates the stepsize behavior of the adaptive Bregman methods. The stepsizes produced by B-adaPG, B-adaPG$_\alpha$, and B-PG-ls

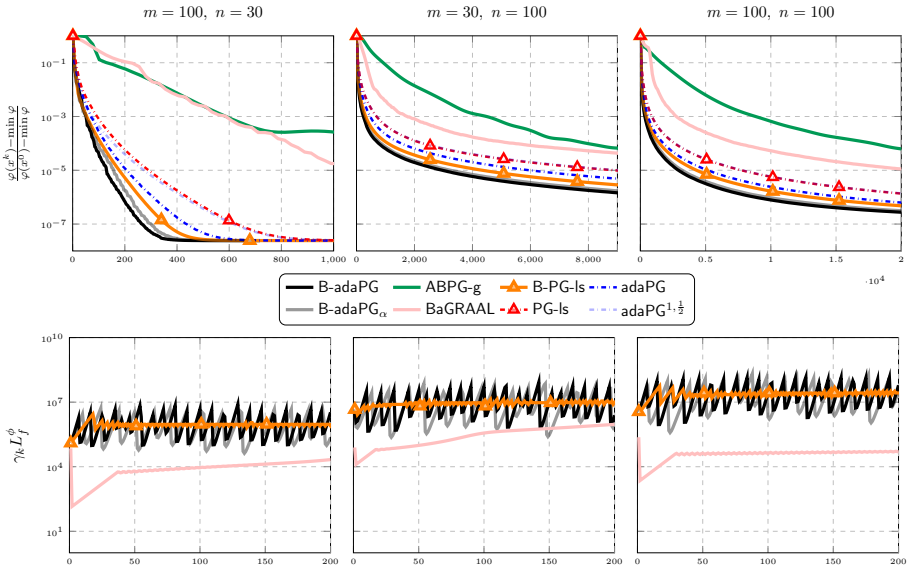

Figure 5.2.1: Hessian growing as a polynomial in ℓ_2 norm. Top: performance comparisons among all algorithms listed in Table 1 in terms of cost. Bottom: stepsize variation (normalized by L_f^ϕ) for Bregman methods with adaptive stepsizes in a window of the first 200 iterations.

oscillate around comparable averages, namely 6 to 7 orders of magnitude larger than the conservative baseline $1/L_f^\phi$. Among these, B-adaPG and B-adaPG$_\alpha$ exhibit notably higher variability, echoing similar empirical observations for their Euclidean counterparts [12, 11, 20]. Such oscillatory stepsize patterns, once regarded as a side effect, have recently received theoretical justification for their efficiency in the unconstrained Euclidean case [9].

5.3 Relative-entropy nonnegative regression

We next test the efficacy of the algorithms when a kernel function ϕ without full domain is employed. The corresponding problems will thus be constrained on (the closure of) $\operatorname{dom}\phi$. This simulation is adapted from [3, §5.3] and concerns a nonegative Poisson linear inverse problem. The problem is formulated as:

$$\underset{x\in\mathbb{R}^n_+}{\text{minimize}}\ \mathrm{KL}(Ax \mid b) + \lambda\|x\|_1, \tag{5.3}$$

where $A \in \mathbb{R}^{m\times n}_+$ and $b \in \mathbb{R}^m_{++}$. The KL-divergence is defined as

$$\mathrm{KL}(x \mid y) := \sum_{i=1}^n \left(x^i \ln \tfrac{x^i}{y^i} - x^i + y^i \right),$$

and is precisely the Bregman distance $\mathrm{D}_\phi(x, y)$ with ϕ being the *Boltzmann-Shannon entropy*

$$\phi(x) = \sum_{i=1}^n x^i \ln x^i. \tag{5.4}$$

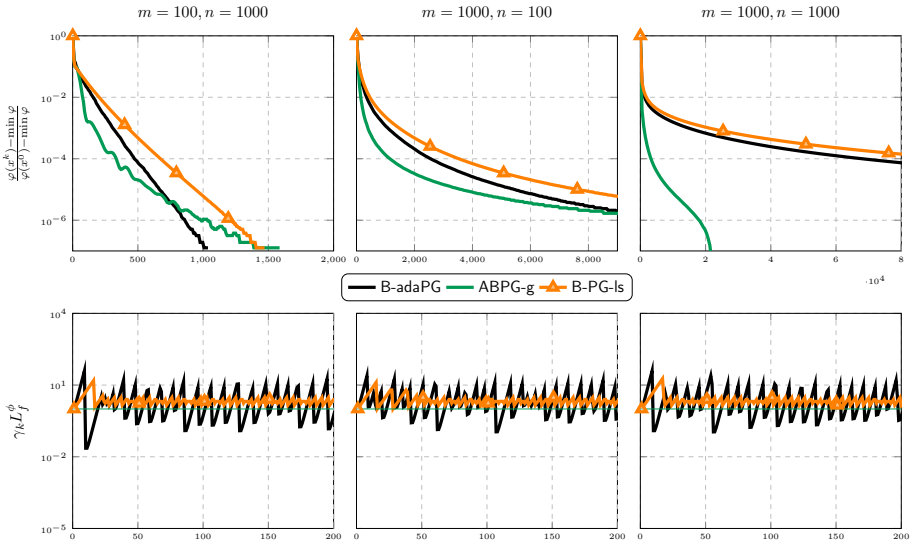

Figure 5.3.1: KL-divergence nonnegative regression. Top: convergence in terms of cost. Bottom: stepsize magnitudes in a window of the first 200 iterations (normalized by $\frac{1}{L_f^\phi}$).

As such, $f(x) \coloneqq \mathrm{KL}(Ax \mid b) = \mathrm{D}_\phi(Ax, b)$ is L_f^ϕ-smooth relative to ϕ with

$$L_f^\phi = \max_{1 \le j \le n} \|A_{:j}\|_1,$$

where $A_{:j}$ denotes the j-th column of A. Note that the Boltzmann-Shannon entropy complies with Assumption 2.3 [3, Rem. 4], and admits a uniform TSE of 1 [10, §2].

For each experiment, the matrix A is sampled with i.i.d. entries from $[0, 1]$ and normalized to sum to one. The response vector b is formed by perturbing the exact output with scaled uniform noise and is strictly positive. We set $\lambda = 0.001$ and generate synthetic data with varying dimensions to compare the performance of different algorithms. As evident from Fig. 5.3.1 (top row), the accelerated algorithms ABPG and ABPG-g perform remarkably well in this problem setup, while our adaptive stepsize selection strategies perform slightly better than the linesearch method. The bottom row offers a snapshot of the stepsize magnitude of adaptive methods on the first 200 iterations, showcasing a fluctuating trend consistent with the observation in Fig. 5.2.1. This time, the stepsizes oscillate around a value slightly higher than $\frac{1}{L_f^\phi}$.

5.4 Relative-entropy barrier minimization on the simplex

This problem involves the minimization of a generalized volumetric barrier function over the probability simplex:

$$\operatorname*{minimize}_{x \in \Delta_n} f(x) \coloneqq \ln \det\left(HX^{-1}H^T\right),$$

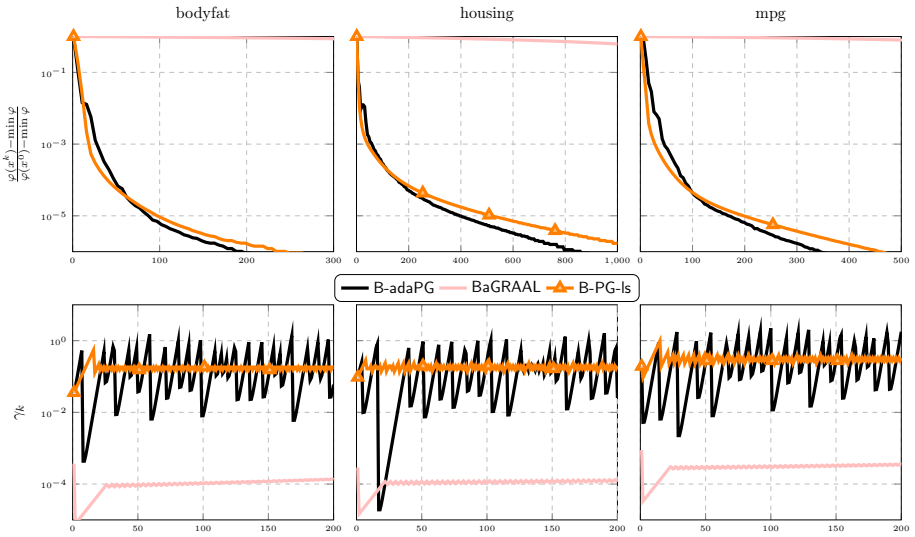

Figure 5.4.1: Relative entropy barrier minimization on the simplex using LIBSVM datasets. Top: convergence in terms of cost. Bottom: stepsize magnitudes in a window of the first 200 iterations.

where $H \in \mathbb{R}^{m\times n}$ with $n \geq m+1$, $X = \operatorname{diag}(x)$, and Δ_n is the probability simplex

$$\Delta_n := \left\{x \in \mathbb{R}^n_+ \mid \textstyle\sum_{i=1}^n x_i = 1\right\}.$$

As observed in [14], f is smooth relative to the *Burg entropy* $x \mapsto -\sum_{i=1}^n \ln x_i$. However, the corresponding Bregman projection onto Δ_n is not available in closed form. For this reason, we instead adopt the Boltzmann-Shannon entropy (5.4) as in the previous experiments, whose associated Bregman projection onto Δ_n does admit a simple closed-form expression: for any $v \in \mathbb{R}^n$ and $y \in \mathbb{R}^n_{++}$, the minimizer of

$$\operatorname*{arg\,min}_{x\in\Delta_n} \{\langle v, x\rangle + \mathrm{D}_\phi(x, y)\}$$

is given by

$$x_i = \frac{y_i e^{-v_i}}{\sum_j y_j e^{-v_j}}, \quad i = 1, \ldots, n. \tag{5.5}$$

Formally, the problem is cast in form (P) as

$$\operatorname*{minimize}_{x\in\mathbb{R}^+_n} \ln\det\left(HX^{-1}H^T\right) + \delta_{\Delta_n}(x),$$

where $\delta_{\Delta_n} : \mathbb{R}^n \to \overline{\mathbb{R}}$ is the *indicator function* of the set Δ_n, namely such that $\delta_{\Delta_n}(x) = 0$ for $x \in \Delta_n$ and $\delta_{\Delta_n}(x) = \infty$ otherwise.

Only algorithms able to cope with lack of relative smoothness and full domain from Table 1 are employed. We conducted experiments on regression datasets from the LIBSVM repository [7], aiming to evaluate its ability to identify the most relevant data points for predicting associated labels; specifically, the bodyfat dataset ($n = 252$, $m = 14$), the mpg dataset ($n = 392$, $m = 7$), and the housing dataset ($n = 506$, $m = 13$). Also in this case, the proposed adaptive method B-adaPG exhibits superior performance over the linearch variant.

We also remark that all solutions in the considered problems exhibit high sparsity (bodyfat: 83.73%, mpg: 96.43%, housing: 92.29%), and thus lie on the boundary of $\operatorname{dom}\phi$. The bottom plots in Fig. 5.4.1 nonetheless demonstrate that the proposed B-adaPG generates stepsizes that stay bounded away from zero, although not theoretically guaranteed by Theorem 2.5(ii). Whether a rigorous confirmation of this trend can be established is currently under investigation.

5.5 Euclidean problems

We finally assess the proposed adaptive selection strategies in purely Euclidean settings, namely, in which the smooth function has a globally Lipschitz continuous gradient. To this end, we consider a standard lasso problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \tfrac{1}{2}\|Ax-b\|^2+\lambda\|x\|_1,$$

where $A\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^m$, and $\lambda=0.01$ is the regularization parameter promoting sparsity. Clearly, $f(x)=\frac{1}{2}\|Ax-b\|^2$ has L_f-Lipschitz gradient with $L_f=\|A\|^2$. We study two setups, each developed in a dedicated subsection.

5.5.1 Bregman vs Euclidean updates

First, we consider the "aggressive" kernel ϕ (5.2), namely the Euclidean $\mathfrak{j}$ augmented by a quadratic function as in Section 5.2:

$$\phi(x)=\tfrac{1}{4}\|x\|_2^4+\tfrac{1}{2}\|x\|_2^2.$$

By doing so, f is smooth relative to ϕ as well, and with same constant $L_f^\phi=L_f$; however, the difference $L_f\phi-f$ is strictly convex, whereas $L_f\mathfrak{j}-f$ is not. This test is meant to compare the performance of Bregman vs Euclidean updates.

As shown in the results of Fig. 5.5.1, Bregman updates seem to outperform purely Euclidean proximal gradient steps. While surprising, this phenomenon can be attributed to the "higher curvature" of $L_f\phi-f$ with respect to that of $L_f\mathfrak{j}-f$, a behavior that we find worthy of future investigations.

5.5.2 Conservatism when $\phi=\mathfrak{j}$

As discussed in Section 2.2, the Bregman analysis investigated here introduces some conservatism; that is, when specialized to the Euclidean kernel $\phi=\mathfrak{j}$, B-adaPG and B-adaPG$_\alpha$ reduce to *dampened* variants of [12, adaPG] and [11, adaPG$^{1,\frac{1}{2}}$]. This second test investigates how the choice $\phi=\mathfrak{j}$ penalizes the performance with respect to adaPG and adaPG$^{1,\frac{1}{2}}$.

Our experiments are based on the Julia code provided in [12], and the test problems are sourced from the LIBSVM dataset [7]. We added our methods into the original test framework and conducted numerical experiments on the same problems. As Fig. 5.5.2 demonstrates, although the discussed differences do have some impact on the algorithms' performance, the outcomes remain acceptable. In fact, to some extent there doesn't appear to be a clear winner. This counterintuitive observation can be attributed to the fact that a small stepsize at an iteration k may trigger a larger stepsize at the next one. We believe that also this aspect is an interesting direction for future research.

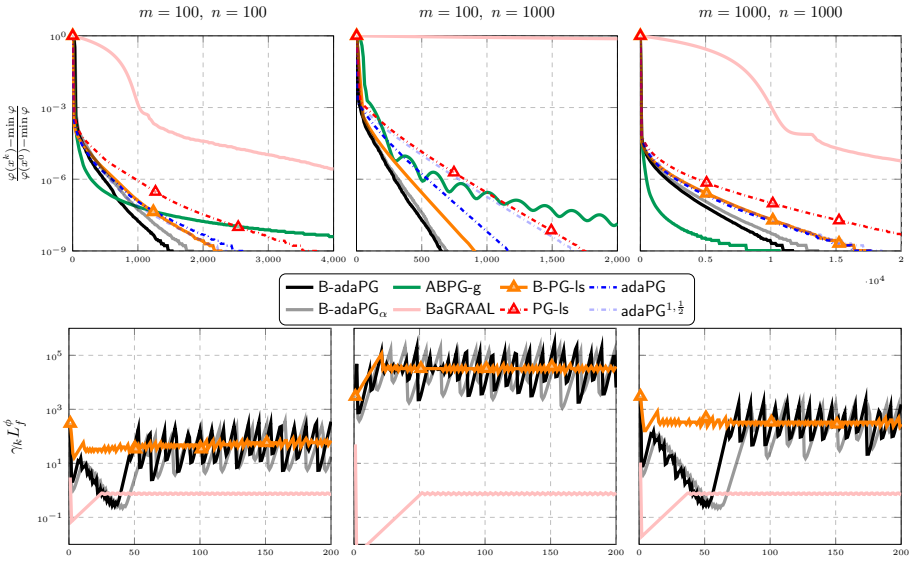

Figure 5.5.1: Random lasso problem with quartic kernel ϕ (5.2). Top: convergence in terms of cost. Bottom: stepsizes normalized by L_f^ϕ in a window of the first 200 iterations.

6 Conclusions

This paper introduced new adaptive stepsize strategies for Bregman proximal gradient algorithms that eliminate the need for traditional backtracking procedures. The proposed methods determine stepsizes dynamically based solely on certain local curvature estimates derived from gradients at the current and previous iterations. This approach enables large stepsizes, often several orders of magnitude larger than their constant stepsize counterparts, leading to fast convergence while maintaining theoretical convergence guarantees. Notably, the theoretical analysis operates under minimal assumptions, requiring only local relative smoothness for the differentiable term, and local strong convexity for the Bregman kernel, rather than their global counterparts, and is thus also agnostic to any such global moduli. A key technical step in our analysis is the development of a Bregman generalization of Young's inequality, which, despite its simplicity, proves essential to the analysis, and is interesting in its own right.

When specialized to $\phi = \jmath$, the proposed algorithms recover the Euclidean counterparts in [12, 11] up to some slight conservatism. Regardless, as shown in the simulations, the flexibility to accommodate arbitrary 1-coercive and Legendre kernels ϕ has remarkable practical advantages even when $\operatorname{dom}\phi = \mathbb{R}^n$.

Some important theoretical questions remain open. Extensive numerical evidence highlights that the proposed adaptive methods generate stepsizes that stay bounded away from zero, even when approaching boundary points of the domain of the kernel function, and fluctuate around values attained by aggressive linesearch routines. This trend is consistent with observations documented in previous studies in the Euclidean setting, but theoretical confirmations in the more general Bregman setup currently do not have a definite answer.

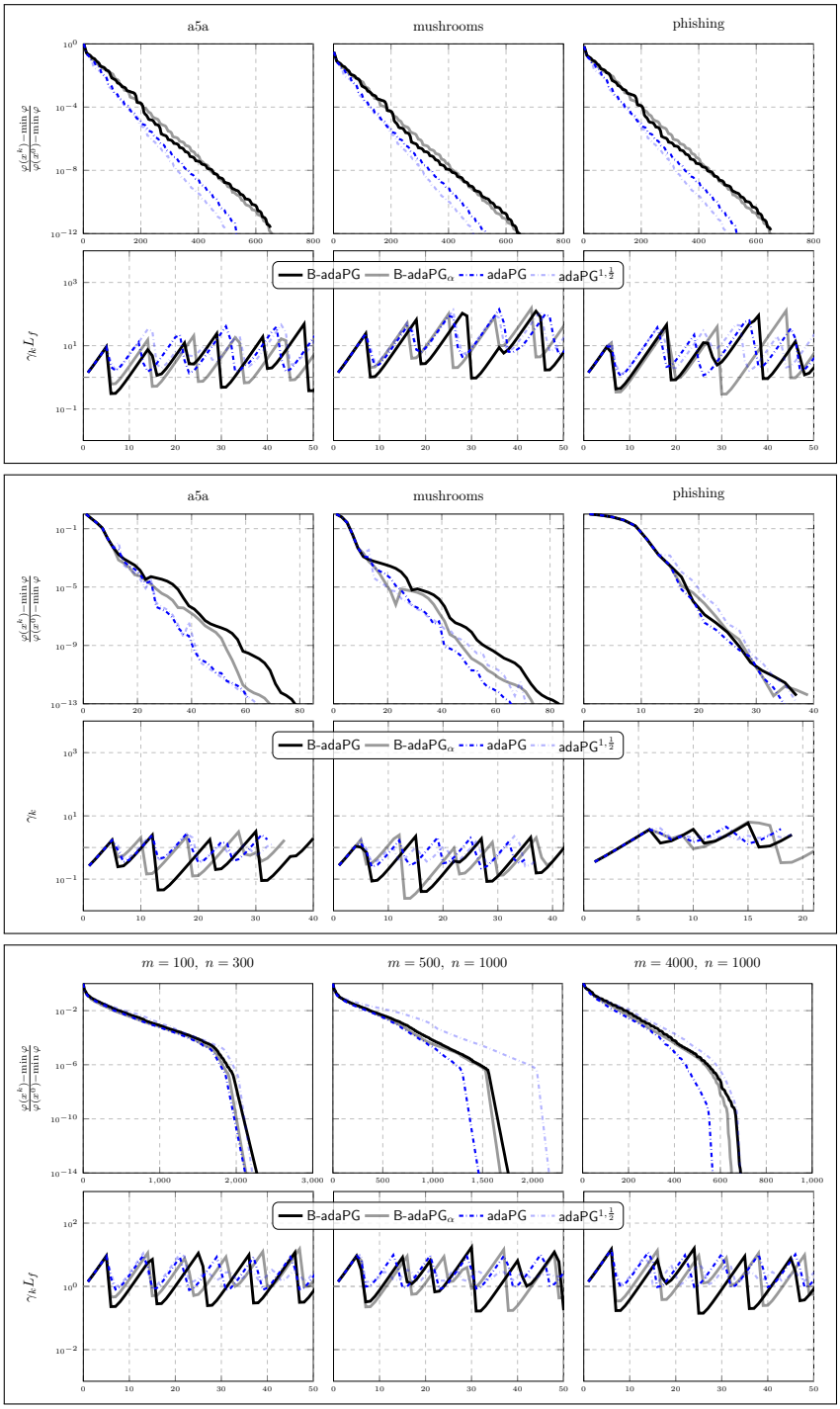

Figure 5.5.2: Comparisons with Euclidean adaptive methods when using kernel $\phi = j$ on sparse logistic regression (top rows), cubic regularization (middle rows), and lasso problems (bottom rows).

A Omitted proofs

Proof of Lemma 3.1. Based on the subgradient characterization (3.4b) and the definition of ℓ_k in (2.3a), we have

$$\begin{aligned} 0 &= \varphi(x^{k-1}) - \varphi(x^k) - \tfrac{1}{\gamma_k}\langle H_k(x^{k-1}) - H_k(x^k), x^{k-1} - x^k\rangle - \tilde{\mathrm{D}}_\varphi(x^{k-1}, x^k) \\ &= P_{k-1}(x) - P_k(x) - \tfrac{1-\gamma_k\ell_k}{\gamma_k}\,\Delta_\phi(x^k, x^{k-1}) - \tilde{\mathrm{D}}_\varphi(x^{k-1}, x^k). \end{aligned} \tag{A.1}$$

Furthermore, by leveraging the subgradient characterization (3.4a) and applying the three point identity of Fact 1.1,

$$\begin{aligned} 0 &= g(x) - g(x^{k+1}) - \langle\tilde{\nabla} g(x^{k+1}), x - x^{k+1}\rangle - \tilde{\mathrm{D}}_g(x, x^{k+1}) \\ &= g(x) - g(x^{k+1}) + \langle\nabla f(x^k), x - x^{k+1}\rangle \\ &\quad - \tfrac{1}{\gamma_{k+1}}\langle\nabla\phi(x^k) - \nabla\phi(x^{k+1}), x - x^{k+1}\rangle - \tilde{\mathrm{D}}_g(x, x^{k+1}) \\ &= g(x) - g(x^{k+1}) + \underbrace{\langle\nabla f(x^k), x - x^{k+1}\rangle}_{\text{(A)}} - \tilde{\mathrm{D}}_g(x, x^{k+1}) \\ &\quad + \tfrac{1}{\gamma_{k+1}}\,\mathrm{D}_\phi(x, x^k) - \tfrac{1}{\gamma_{k+1}}\,\mathrm{D}_\phi(x, x^{k+1}) - \tfrac{1}{\gamma_{k+1}}\,\mathrm{D}_\phi(x^{k+1}, x^k). \end{aligned} \tag{A.2}$$

We next proceed to expand the term (A) as

$$\begin{aligned} \text{(A)} &= \langle\nabla f(x^k), x - x^k\rangle - \langle\nabla f(x^k), x^{k+1} - x^k\rangle \\ &= \langle\nabla f(x^k), x - x^k\rangle + \tfrac{1}{\gamma_k}\langle H_k(x^{k-1}) - \nabla\phi(x^k), x^{k+1} - x^k\rangle \\ &\quad + \tfrac{1}{\gamma_k}\langle H_k(x^{k-1}) - H_k(x^k), x^k - x^{k+1}\rangle \\ &= \Big[f(x) - f(x^k) - \mathrm{D}_f(x, x^k)\Big] + \Big[g(x^{k+1}) - g(x^k) - \tilde{\mathrm{D}}_g(x^{k+1}, x^k)\Big] \\ &\quad + \tfrac{1}{\gamma_{k+1}}B_{k+1}, \end{aligned}$$

which combined with (A.2) gives

$$\begin{aligned} 0 = \; &- P_k(x) + \tfrac{1}{\gamma_{k+1}}\,\mathrm{D}_\phi(x, x^k) - \tfrac{1}{\gamma_{k+1}}\,\mathrm{D}_\phi(x, x^{k+1}) - \tfrac{1}{\gamma_{k+1}}\,\mathrm{D}_\phi(x^{k+1}, x^k) \\ &- \mathrm{D}_f(x, x^k) - \tilde{\mathrm{D}}_g(x^{k+1}, x^k) + \tfrac{1}{\gamma_{k+1}}B_{k+1} - \tilde{\mathrm{D}}_g(x, x^{k+1}). \end{aligned}$$

As done in [12], we may now add (A.2) to (A.1) scaled by ϑ_{k+1} and multiply everything by γ_{k+1} to obtain

$$\begin{aligned} 0 = \; &- \gamma_{k+1}P_k(x) + \mathrm{D}_\phi(x, x^k) - \mathrm{D}_\phi(x, x^{k+1}) - \mathrm{D}_\phi(x^{k+1}, x^k) \\ &- \gamma_{k+1}\,\mathrm{D}_f(x, x^k) - \gamma_{k+1}\,\tilde{\mathrm{D}}_g(x^{k+1}, x^k) + B_{k+1} - \gamma_{k+1}\,\tilde{\mathrm{D}}_g(x, x^{k+1}) \\ &+ \gamma_{k+1}\vartheta_{k+1}\left(P_{k-1}(x) - P_k(x) - \tfrac{1-\gamma_k\ell_k}{\gamma_k}\,\Delta_\phi(x^k, x^{k-1}) - \tilde{\mathrm{D}}_\varphi(x^{k-1}, x^k)\right). \end{aligned}$$

After suitably rearranging, the claimed identity is obtained.

The inequality follows by neglecting the terms between curly brackets, and further using the identity $\mathrm{D}_\phi(x^k, x^{k-1}) = \frac{\alpha_k}{1+\alpha_k}\,\Delta_\phi(x^k, x^{k-1})$. □

Proof of Lemma 4.1. It suffices to prove the claim for B-adaPG; the case of B-adaPG$_\alpha$ under the needed assumptions follows by simply replacing $\mathcal{U}_k \leftarrow \mathcal{U}_k^\alpha$. The assumption on x ensures that $P_k(x) \geq 0$ holds for all k, implying both that

$\mathcal{U}_k(x) \geq 0$ and the claimed monotonic decrease of $(\mathcal{U}_k(x))_{k\in\mathbb{N}}$ with finite limit by virtue of Lemmas 3.6 and 3.7. More precisely, one has that

$$0 \leq \mathcal{U}_{k+1}(x) \leq \mathcal{U}_k(x) - \gamma_k(1 + q\hat{\rho}_k - q\hat{\rho}_{k+1}\rho_{k+1})P_{k-1}(x),$$

where $q = 1$ for B-adaPG and $q = \frac{1+\alpha}{2\alpha}$ for B-adaPG$_\alpha$. A telescoping argument yields that

$$\begin{aligned} P_K^{\min}(x)\sum_{k=1}^{K}\gamma_k(1 + q\hat{\rho}_k - q\hat{\rho}_{k+1}\rho_{k+1}) &\leq \sum_{k=1}^{K}\gamma_k(1 + q\hat{\rho}_k - q\hat{\rho}_{k+1}\rho_{k+1})P_{k-1}(x) \\ &\leq \mathcal{U}_1(x) - \mathcal{U}_{K+1}(x) \\ &\leq \mathcal{U}_1(x) - \gamma_{K+1}(1 + q\hat{\rho}_{K+1})P_K^{\min}(x), \end{aligned}$$

hence that

$$\begin{aligned} P_K^{\min}(x) &\leq \frac{\mathcal{U}_1(x)}{\sum_{k=1}^{K}(\gamma_k + q\gamma_k\hat{\rho}_k - q\gamma_{k+1}\hat{\rho}_{k+1}) + \gamma_{K+1} + q\gamma_{K+1}\hat{\rho}_{K+1}} \\ &\leq \frac{\mathcal{U}_1(x)}{\gamma_1\hat{\rho}_1 + \sum_{k=1}^{K+1}\gamma_k}. \end{aligned}$$

Since $\mathcal{U}_1(x) \leq \mathcal{U}_0(x)$, the claimed inequality follows. □

References

[1] Masoud Ahookhosh, Andreas Themelis, and Panagiotis Patrinos. A Bregman forward-backward linesearch algorithm for nonconvex composite optimization: Superlinear convergence to nonisolated local minima. *SIAM Journal on Optimization*, 31(1):653–685, 2021.

[2] Jonathan Barzilai and Jonathan M. Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, jan 1988.

[3] Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[4] Heinz H. Bauschke and Jonathan M. Borwein. Legendre functions and the method of random Bregman projections. *Journal of Convex Analysis*, 4(1):27–67, 1997.

[5] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, 2017.

[6] Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

[7] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[8] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[9] Benjamin Grimmer, Kevin Shu, and Alex L. Wang. Composing optimized stepsize schedules for gradient descent. *arXiv:2410.16249*, 2024.

[10] Filip Hanzely, Peter Richtárik, and Lin Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *Computational Optimization and Applications*, 79(2):405–440, 2021.

[11] Puya Latafat, Andreas Themelis, and Panagiotis Patrinos. On the convergence of adaptive first order methods: Proximal gradient and alternating minimization algorithms. In *Proceedings of the 6th Annual Learning for Dynamics & Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pages 197–208. PMLR, 2024.

[12] Puya Latafat, Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Adaptive proximal algorithms for convex optimization under local Lipschitz continuity of the gradient. *Mathematical Programming*, 2024.

[13] Tianjiao Li and Guanghui Lan. A simple uniformly optimal method without line search for convex optimization. *Mathematical Programming*, pages 1–38, 2025.

[14] Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[15] Yura Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184(1):383–410, 2020.

[16] Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6702–6712, 13- 2020.

[17] Yura Malitsky and Konstantin Mishchenko. Adaptive proximal gradient method for convex optimization. In *Advances in NeurIPS*, volume 37, pages 100670–100697, 2024.

[18] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27, 1983.

[19] Max Nilsson and Pontus Giselsson. The symmetry coefficient of positively homogeneous functions. 2025.

[20] Konstantinos Oikonomidis, Emanuel Laude, Puya Latafat, Andreas Themelis, and Panagiotis Patrinos. Adaptive proximal gradient methods are universal without approximation. In *Proceedings of the 41st ICML*, volume 235, pages 38663–38682. PMLR, 2024.

[21] Hongjia Ou and Andreas Themelis. Safeguarding adaptive methods: Global convergence of Barzilai-Borwein and other stepsize choices. In *10th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 2802–2807, jul 2024.

[22] Ralph T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[23] Mikhail V. Solodov and Benar F. Svaiter. An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions. *Mathematics of Operations Research*, 25(2):214–230, 2000.

[24] Jaewook J. Suh and Shiqian Ma. An adaptive and parameter-free Nesterov's accelerated gradient method for convex optimization. *arXiv: 2505.11670*, 2025.

[25] Matthew K. Tam and Daniel J. Uteda. Bregman-golden ratio algorithms for variational inequalities. *Journal of Optimization Theory and Applications*, 199(3):993–1021, 2023.

[26] Xianfu Wang and Heinz H. Bauschke. The Bregman proximal average. *SIAM Journal on Optimization*, 32(2):1379–1401, 2022.

[27] Ziyuan Wang and Andreas Themelis. Bregman level proximal subdifferentials and new characterizations of Bregman proximal operators. *arXiv: 2506.07333*, 2025.

[28] Ziyuan Wang, Andreas Themelis, Hongjia Ou, and Xianfu Wang. A mirror inertial forward–reflected–backward splitting: Convergence analysis beyond convexity and Lipschitz smoothness. *Journal of Optimization Theory and Applications*, 203(2):1127–1159, nov 2024.

[29] Danqing Zhou, Shiqian Ma, and Junfeng Yang. Adabb: Adaptive Barzilai-Borwein method for convex optimization. *Mathematics of Operations Research*, 2025.