Accelerated Bregman proximal gradient methods for relatively smooth convex optimization

Filip Hanzely * Peter Richtárik *†‡ Lin Xiao §

August 8, 2018

Abstract

We consider the problem of minimizing the sum of two convex functions: one is differentiable and relatively smooth with respect to a reference convex function, and the other can be nondifferentiable but simple to optimize. The relatively smooth condition is much weaker than the standard assumption of uniform Lipschitz continuity of the gradients, thus significantly increases the scope of potential applications. We present accelerated Bregman proximal gradient (ABPG) methods that employ the Bregman distance of the reference function as the proximity measure. These methods attain an $O(k^{-\gamma})$ convergence rate in the relatively smooth setting, where $\gamma \in [1, 2]$ is determined by a triangle scaling property of the Bregman distance. We develop adaptive variants of the ABPG method that automatically ensure the best possible rate of convergence and argue that the $O(k^{-2})$ rate is attainable in most cases. We present numerical experiments with three applications: D-optimal experiment design, Poisson linear inverse problem, and relative-entropy nonnegative regression. In all experiments, we obtain numerical certificates showing that these methods do converge with the $O(k^{-2})$ rate.

1 Introduction

Let $\mathbb{R}^n$ be the n-dimensional real Euclidean space endowed with inner product $\langle x, y\rangle = \sum_{i=1}^n x^{(i)} y^{(i)}$ and the Euclidean norm $\|x\| = \sqrt{\langle x, x\rangle}$. We consider optimization problems of the form

$$\underset{x\in C}{\text{minimize}} \ \left\{\phi(x) := f(x) + \Psi(x)\right\}, \tag{1}$$

where C is a closed convex set in $\mathbb{R}^n$, and f and Ψ are proper, closed convex functions. We assume that f is differentiable on an open set that contains the relative interior of C (denoted as $\operatorname{rint} C$). For the development of first-order methods, we also assume that C and Ψ are *simple*, whose precise meaning will be explained in the context of specific algorithms.

First-order methods for solving (1) are often based on the idea of minimizing a simple approximation of the objective ϕ during each iteration. Specifically, in the *proximal gradient method*, we

*Division of Computer, Electrical and Mathematical Sciences and Engineering (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia. Emails: filip.hanzely@kaust.edu.sa, peter.richtarik@kaust.edu.sa

†School of Mathematics, The University of Edinburgh, Edinburgh, United Kingdom.

‡Moscow Institute of Physics and Technology, Dolgoprudny, Russia.

§Microsoft Research, Redmond, Washington, United States. Email: lin.xiao@microsoft.com

start with an initial point $x_0 \in \operatorname{rint} C$ and generate a sequence x_k for $k = 1, 2, \ldots$ with

$$x_{k+1} = \arg\min_{x \in C} \Big\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L_k}{2} \|x - x_k\|^2 + \Psi(x) \Big\}, \tag{2}$$

where $L_k > 0$ for all $k \geq 0$. Here, we use the gradient $\nabla f(x_k)$ to construct a local quadratic approximation of f around x_k while leaving Ψ untouched. Our assumption that C and Ψ are simple means that the minimization problem in (2) can be solved efficiently, especially if it admits a closed-form solution.

Assuming that ϕ is bounded below, convergence of the proximal gradient method can be established if $\phi(x_{k+1}) \leq \phi(x_k)$ for all $k \in \mathbb{N}$. A sufficient condition for this to hold is that the quadratic approximation of f in (2) is an upper approximation (majorization). This is the basic idea behind many general methods for nonlinear optimization. To this end, a common assumption is for the gradient of f to satisfy a uniform Lipschitz condition, i.e., there exists a constant L_f such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\|, \quad \forall\, x, y \in \operatorname{rint} C. \tag{3}$$

This smoothness assumption implies (see, e.g., [19])

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L_f}{2} \|x - y\|^2, \quad \forall\, x \in C,\ y \in \operatorname{rint} C. \tag{4}$$

Therefore, setting $L_k = L_f$ for all $k \in \mathbb{N}$ ensures that the quadratic approximation of f in (2) is always an upper bound of f, which implies $\phi(x_{k+1}) \leq \phi(x_k)$ for all $k \in \mathbb{N}$. Moreover, it can be shown that the proximal gradient method enjoys an $O(k^{-1})$ convergence rate, i.e.,

$$\phi(x_k) - \phi(x) \leq \frac{L_f}{k} \frac{\|x - x_0\|^2}{2}, \quad \forall\, x \in C. \tag{5}$$

See, e.g., [6], [21] and [5, Chapter 10]. Under the same assumption, accelerated proximal gradient methods [19, 2, 6, 26, 21] can achieve a faster $O(k^{-2})$ convergence rate:

$$\phi(x_k) - \phi(x) \leq \frac{4L_f}{(k+2)^2} \frac{\|x - x_0\|^2}{2}, \quad \forall\, x \in C, \tag{6}$$

which is optimal (up to a constant factor) for this class of convex optimization problems [17, 19].

1.1 Relative smoothness

While the uniform smoothness condition (3) is central in the development and analysis of first-order methods, there are many applications where the objective function does not have this property, despite being convex and differentiable. For example, in D-optimal experiment design (e.g., [14, 1]) and Poisson inverse problems (e.g., [11, 7]), the objective functions involve the logarithm in the form of log-determinant or relative entropy, whose gradients may blow up towards the boundary of the feasible region. In order to develop efficient first-order algorithms for solving such problems, the notion of *relative smoothness* was introduced by several recent works [3, 16, 27].

Let h be a strictly convex function that is differentiable on $\operatorname{rint} C$. In fact, we require C is the closure of $\operatorname{dom} h$, i.e., $C = \overline{\operatorname{dom} h}$. The Bregman distance [9] associated with h is defined as

$$D_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle, \quad \forall\, x \in \operatorname{dom} h,\ y \in \operatorname{rint} \operatorname{dom} h.$$

Definition 1. *The function f is called L-*smooth relative *to h on C if there is an $L > 0$ such that*

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + L D_h(x, y), \quad \forall\, x \in C,\ y \in \operatorname{rint} C. \tag{7}$$

As shown in [3] and [16], this notion of relative smoothness is equivalent to the following statements:

- $Lh - f$ is a convex function on $\operatorname{rint} C$.
- If both f and h are twice differentiable, then $\nabla^2 f(x) \preceq L\, \nabla^2 h(x)$ for all $x \in \operatorname{rint} C$.
- $\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L \langle \nabla h(x) - \nabla h(y), x - y \rangle$ for all $x, y \in \operatorname{rint} C$.

The definition of relative smoothness in (7) gives an upper approximation of f that is similar to (4). In fact, (4) is a special case of (7) with $h = (1/2)\|x\|^2$ and $D_h(x, y) = (1/2)\|x - y\|^2$. Therefore it is natural to consider a more general algorithm by replacing the squared Euclidean distance in (2) with a Bregman distance:

$$x_{k+1} = \operatorname*{arg\,min}_{x \in C} \Big\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + L_k D_h(x, x_k) + \Psi(x) \Big\}. \tag{8}$$

Here, our assumption that C and Ψ are simple means that the minimization problem in (8) can be solved efficiently. Similar to the proximal gradient method (2), this algorithm can also be interpreted through operator splitting mechanism: it is the composition of a Bregman proximal step and a Bregman gradient step (see details in [3, Section 3.1]). Therefore, it is called the *Bregman proximal gradient* (BPG) method [25].

Under the relative smoothness condition (7), setting $L_k = L$ ensures that the function being minimized in (8) is a majorization of ϕ, which implies $\phi(x_{k+1}) \leq \phi(x_k)$ for all $k \in \mathbb{N}$. It was first shown in [8] (for the case $\Psi \equiv 0$) that the BGD method has a $O(k^{-1})$ convergence rate:

$$\phi(x_k) - \phi(x) \leq \frac{L}{k} D_h(x, x_0), \quad \forall\, x \in \operatorname{dom} h.$$

This is a generalization of (5). The same convergence rate for the general case (with nontrivial Ψ) is obtained in [3], where the authors also discussed the effect of a symmetry measure for the Bregman distance. Similar results are also obtained in [16] and [27]. In addition, [16] introduced the notion of relative strong convexity and obtained linear convergence of the BPG method when both relative smoothness and relative strong convexity hold. More recently, [12] studied stochastic gradient descent and randomized coordinate descent methods in the relatively smooth setting, and [15] extended this framework to minimize relatively continuous convex functions.

An apparent interesting question is whether we can obtain the accelerated $O(k^{-2})$ convergence rate in the relatively smooth setting [16, 25], which is the focus of our investigation in this paper.

1.2 Contributions and outline

We propose accelerated Bregman proximal gradient (ABPG) methods that attain an $O(k^{-\gamma})$ convergence rate, for some $\gamma \in [1, 2]$, in the relatively smooth setting. More specifically, under the assumption (7), the basic ABPG method produces a sequence $\{x_k\}_{k \in \mathbb{N}}$ satisfying

$$\phi(x_k) - \phi(x) \leq \left(\frac{\gamma}{k + \gamma} \right)^{\gamma} L D_h(x, x_0), \quad \forall\, x \in \operatorname{dom} h. \tag{9}$$

The exact value of γ depends on a *triangle scaling property* of the Bregman distance. For $D_h(x,y) = (1/2)\|x-y\|^2$, we have $\gamma = 2$ and $L = L_f$, hence the result in (9) recovers that in (6).

In Section 2, we define the triangle-scaling property for general Bregman distances, where γ appears as a triangle-scaling exponent (TSE). We estimate the value of γ for some Bregman distances that appear frequently in applications. Moreover, we derive an intrinsic triangle-scaling property that allows us to use $\gamma = 2$ locally for all h that is twice continuously differentiable.

In Section 3, we present the basic ABPG method and prove that it attains the convergence rate in (9). We also give an adaptive variant that can automatically search for the largest possible value of γ. In Section 4, we develop adaptive ABPG methods that automatically adjust an additional gain factor in order to work with the intrinsic TSE $\gamma = 2$ and are capable of obtaining the $O(k^{-2})$ convergence rate. In Section 5, we present an accelerated Bregman dual-averaging algorithm that attains the $O(k^{-\gamma})$ convergence rate.

Finally, in Section 6, we present numerical experiments with three applications: the D-optimal experiment design problem, a Poisson linear inverse problem, and relative-entropy nonnegative regression. In all experiments, the ABPG methods, especially the adaptive variants, demonstrate superior performance compared with the BPG method. Moreover, we obtain numerical certificates that the ABPG methods converge with $O(k^{-2})$ rate in all three applications.

Related work. The relative smoothness condition directly extends the upper approximation property (4) with more general Bregman distances. Nesterov [22] took an alternative approach by extending the Lipschitz condition (3). Specifically, he considered functions with Hölder continuous gradients with a parameter $\nu \in [0,1]$:

$$\|\nabla f(x) - \nabla f(y)\|_* \le L_\nu \|x-y\|^\nu, \quad x, y \in C,$$

and obtained $O(k^{-(1+\nu)/2})$ rate with a universal gradient method and $O(k^{-(1+3\nu)/2})$ rate with accelerated schemes. These methods are called "universal" because they do not assume the knowledge of ν and automatically ensure the best possible rate of convergence. The accelerated $O(k^{-(1+3\nu)/2})$ rate interpolates between $O(k^{-1/2})$ and $O(k^{-2})$ with $\nu \in [0,1]$. There seems to be no simple connection or correspondence between the Hölder smoothness property and the combination of relative smoothness and the triangle scaling property studied in this paper.

Technical assumptions. Development and analysis of optimization methods in the relatively smooth setting require some delicate assumptions in order to cover many interesting applications without loss of rigor. Here we adopt the same assumptions made in [3] regarding problem (1).

Assumption A. *The set $C = \overline{\operatorname{dom} h}$ is convex, and the following statements hold:*

1. *$h : \mathbb{R}^n \to (-\infty, \infty]$ is of Legendre type [24, Section 26]. In other words, it is essentially smooth and strictly convex in $\operatorname{rint} \operatorname{dom} h$. Essential smoothness means that it is differentiable and $\|\nabla h(x_k)\| \to \infty$ for every sequence $\{x_k\}_{k\in\mathbb{N}}$ converging to a boundary point of $\operatorname{dom} h$.*
2. *$f : \mathbb{R}^n \to (-\infty, \infty]$ is a proper and closed convex function, and it is differentiable on $\operatorname{rint} C$.*
3. *$\Psi : \mathbb{R}^n \to (-\infty, \infty]$ is a proper and closed convex function, and $\operatorname{dom} \Psi \cap \operatorname{rint} \operatorname{dom} h \neq \emptyset$.*
4. *$\inf_{x \in C}\{f(x) + \Psi(x)\} > -\infty$, i.e., problem (1) is bounded below.*
5. *The BPG step (8) is well posed, meaning that x_{k+1} is unique and belongs to $\operatorname{rint} \operatorname{dom} h$.*

Sufficient conditions for the well-posedness of (8) are given in [3, Lemma 2]. The same conditions also ensure that our proposed accelerated methods are well-posed.

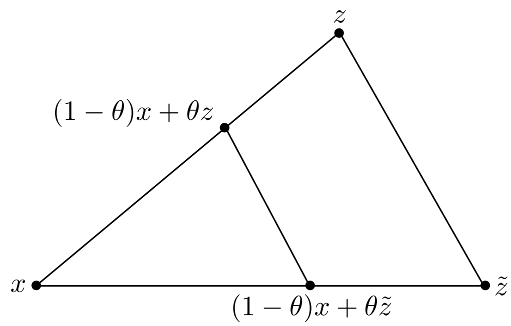


Figure 1: Illustration of different points in the triangle scaling property.

2 Triangle scaling of Bregman distance

In this section, we define the *triangle scaling property* for Bregman distances and discuss two different notions of *triangle scaling exponent* (TSE).

Definition 2. *Let h be a convex function that is differentiable on* $\operatorname{rint}\operatorname{dom} h$*. The Bregman distance D_h has the* triangle scaling property *if there is some $\gamma > 0$ such that for all $x, z, \tilde{z} \in \operatorname{rint}\operatorname{dom} h$,*

$$D_h\big((1-\theta)x + \theta z,\ (1-\theta)x + \theta\tilde{z}\big) \ \le\ \theta^\gamma D_h(z, \tilde{z}), \qquad \forall\, \theta \in [0, 1]. \tag{10}$$

We call γ a (uniform) triangle scaling exponent *(TSE) of D_h.*

Figure 1 gives a geometric illustration of the points involved in the above definition. When $\gamma = 1$, inequality (10) holds if $D_h(x, y)$ is jointly convex in (x, y). This is because

$$D_h\big((1-\theta)x + \theta z,\ (1-\theta)x + \theta\tilde{z}\big) \le (1-\theta)D_h(x, x) + \theta D_h(z, \tilde{z}) = \theta D_h(z, \tilde{z}).$$

Therefore it is useful to study jointly convex Bregman distances. Suppose $h : \mathbb{R} \to (-\infty, \infty]$ is strictly convex and twice continuously differentiable on an open interval in $\mathbb{R}$. Let h'' denotes the second derivative of h. It was shown in [4] that the Bregman distance $D_h(\cdot, \cdot)$ is jointly convex if and only if $1/h''$ is concave. This result applies directly to separable functions which can be written as $h(x) = \sum_{i=1}^n h_i(x^{(i)})$. If $1/h_i''$ is concave for each $i = 1, \ldots, n$, then we conclude that D_h has a uniform TSE of at least 1. Below are some specific examples:

- *The squared Euclidean distance.* Let $h(x) = (1/2)\|x\|_2^2$ and $D_h(x, y) = (1/2)\|x - y\|_2^2$. Obviously, here D_h is jointly convex in its two arguments. But it is also easy to see that

  $$\frac{1}{2}\left\|(1-\theta)x + \theta z - \big((1-\theta)x + \theta\tilde{z}\big)\right\|_2^2 = \frac{1}{2}\|\theta(z - \tilde{z})\|_2^2 = \theta^2 \frac{1}{2}\|z - \tilde{z}\|_2^2.$$

  Therefore the squared Euclidean distance has a uniform TSE $\gamma = 2$, which is much larger than 1 obtained by following the jointly convex argument.

- *The generalized Kullback-Leibler (KL) divergence.* Let h be the negative Boltzmann-Shannon entropy: $h(x) = \sum_{i=1}^n x^{(i)} \log x^{(i)}$ defined on $\mathbb{R}_+^n$. The Bregman distance associated with h is

  $$D_{\mathrm{KL}}(x, y) = \sum_{i=1}^n \left(x^{(i)} \log\left(\frac{x^{(i)}}{y^{(i)}}\right) - x^{(i)} + y^{(i)} \right). \tag{11}$$

Since $1/h_i'' = x^{(i)}$ is linear thus concave for each i, we conclude that $D_{\mathrm{KL}}(x,y)$ is jointly convex in (x,y), which implies that it has a uniform TSE $\gamma = 1$.

- *The Itakura-Saito (IS) distance.* The IS distance is the Bregman distance generated by Burg's entropy $h(x) = \sum_{i=1}^n -\log(x^{(i)})$ with $\operatorname{dom} h = \mathbb{R}^n_{++}$:

$$D_{\mathrm{IS}}(x,y) = \sum_{i=1}^n \left(-\log\left(\frac{x^{(i)}}{y^{(i)}}\right) + \frac{x^{(i)}}{y^{(i)}} - 1 \right). \tag{12}$$

Since $1/h_i'' = (x^{(i)})^2$ is not concave, we conclude that $D_{\mathrm{IS}}(\cdot,\cdot)$ is not jointly convex. Hence if it has a uniform TSE, then it is likely to be less than 1. In fact, it can be easily checked numerically that any $\gamma > 0.5$ is not a uniform TSE for D_{IS}.

We observe that the largest uniform TSEs are quite different for the three popular Bregman distances listed above. An important question is: Are these differences essential such that they lead to different convergence rates if different Bregman distances are used in an accelerated algorithm? It would be ideal to derive an intrinsic characterization that is common for most Bregman distances and essential for convergence analysis of accelerated algorithms.

2.1 The intrinsic triangle-scaling exponent

For any fixed triple $\{x, z, \tilde{z}\} \subset \operatorname{rint}\operatorname{dom} h$, we consider a relaxed version of triangle scaling:

$$D_h\big((1-\theta)x + \theta z,\ (1-\theta)x + \theta\tilde{z}\big) \ \le\ G(x,z,\tilde{z})\,\theta^\gamma D_h(z,\tilde{z}), \qquad \forall\, \theta \in [0,1], \tag{13}$$

where $G(x,z,\tilde{z})$ depends on the triple $\{x,z,\tilde{z}\}$ but does not depend on θ.

Definition 3. *The intrinsic TSE of D_h, denoted γ_{in}, is the largest γ such that* (13) *holds with some finite $G(x,z,\tilde{z})$ for all triples $\{x,z,\tilde{z}\} \subset \operatorname{rint}\operatorname{dom} h$.*

Notice that when θ is bounded away from 0, we can always find sufficiently large $G(x,z,\tilde{z})$ to make the inequality in (13) hold with any value of γ. Therefore, the intrinsic TSE is determined only by the asymptotic behavior of $D_h\big((1-\theta)x+\theta z, (1-\theta)x+\theta\tilde{z}\big)$ when $\theta \to 0$. More precisely, it is the largest γ such that

$$\limsup_{\theta\to 0} \frac{D_h\big((1-\theta)x+\theta z, (1-\theta)x+\theta\tilde{z}\big)}{\theta^\gamma} \ <\ \infty.$$

We show that a broad family of Bregman distances share the same intrinsic TSE $\gamma_{\mathrm{in}} = 2$.

Theorem 1. *If h is convex and twice continuously differentiable on* $\operatorname{rint}\operatorname{dom} h$*, then the intrinsic TSE of the Bregman distance D_h is 2. Specifically, for any $\{x,z,\tilde{z}\} \subset \operatorname{rint}\operatorname{dom} h$, we have*

$$\lim_{\theta\to 0} \frac{D_h\big((1-\theta)x+\theta z,\ (1-\theta)x+\theta\tilde{z}\big)}{\theta^2} = \frac{1}{2}\big\langle \nabla^2 h(x)(z-\tilde{z}), z-\tilde{z}\big\rangle. \tag{14}$$

Proof. Consider the limit in (14), since both the numerator $D_h\big((1-\theta)x+\theta z,\ (1-\theta)x+\theta\tilde{z}\big)$ and the denominator θ^2 converge to zero as $\theta \to 0$, we apply L'Hospital's rule. First, by definition of the Bregman distance, we have

$$\begin{aligned}
& D_h\big((1-\theta)x+\theta z,\ (1-\theta)x+\theta\tilde{z}\big) \\
=\ & D_h\big(x+\theta(z-x),\ x+\theta(\tilde{z}-x)\big) \\
=\ & h\big(x+\theta(z-x)\big) - h\big(x+\theta(\tilde{z}-x)\big) - \big\langle \nabla h\big(x+\theta(\tilde{z}-x)\big), \theta(z-\tilde{z})\big\rangle.
\end{aligned}$$

The derivative of $D_h\big((1-\theta)x+\theta z,\ (1-\theta)x+\theta\tilde{z}\big)$ with respect to θ is

$$\frac{d}{d\theta}D_h\big((1-\theta)x+\theta z,\ (1-\theta)x+\theta\tilde{z}\big) \;=\; A(\theta)-\big\langle\nabla^2 h(x+\theta(\tilde{z}-x))(\tilde{z}-x),\theta(z-\tilde{z})\big\rangle,$$

where

$$A(\theta)=\big\langle\nabla h(x+\theta(z-x)),z-x\big\rangle-\big\langle\nabla h(x+\theta(\tilde{z}-x)),\tilde{z}-x\big\rangle-\big\langle\nabla h(x+\theta(\tilde{z}-x)),z-\tilde{z}\big\rangle.$$

Therefore,

$$\begin{aligned}\lim_{\theta\to 0}\frac{D_h\big((1-\theta)x+\theta z,(1-\theta)x+\theta\tilde{z}\big)}{\theta^2} &= \lim_{\theta\to 0}\frac{A(\theta)-\big\langle\nabla^2 h\big(x+\theta(\tilde{z}-x)\big)(\tilde{z}-x),\theta(z-\tilde{z})\big\rangle}{2\theta}\\ &= \lim_{\theta\to 0}\frac{A(\theta)}{2\theta}-\lim_{\theta\to 0}\frac{\big\langle\nabla^2 h\big(x+\theta(\tilde{z}-x)\big)(\tilde{z}-x),z-\tilde{z}\big\rangle}{2}\\ &= \lim_{\theta\to 0}\frac{A(\theta)}{2\theta}-\frac{1}{2}\big\langle\nabla^2 h(x)(\tilde{z}-x),z-\tilde{z}\big\rangle. \qquad (15)\end{aligned}$$

Notice that

$$\lim_{\theta\to 0}A(\theta)=\big\langle\nabla h(x),z-x\big\rangle-\big\langle\nabla h(x),\tilde{z}-x\big\rangle-\big\langle\nabla h(x),z-\tilde{z}\big\rangle=0,$$

so we apply L'Hospital's rule again:

$$\lim_{\theta\to 0}\frac{A(\theta)}{2\theta}=\frac{\big\langle\nabla^2 h(x)(z-x),z-x\big\rangle-\big\langle\nabla^2 h(x)(\tilde{z}-x),\tilde{z}-x\big\rangle-\big\langle\nabla^2 h(x)(\tilde{z}-x),z-\tilde{z}\big\rangle}{2}.$$

Plugging the last equality into (15) and after some simple algebra, we arrive at (14). □

According to Theorem 1, the three examples we considered earlier, the squared Euclidean distance, the generalized KL-divergence and the IS-distance, share the same intrinsic TSE $\gamma_{\rm in}=2$. Theorem 1 also implies that the largest uniform TSE cannot exceed 2.

2.2 Estimating the triangle-scaling gain

It can be hard to give a general upper bound on the triangle-scaling gain $G(x,z,\tilde{z})$ in (13) that works with the intrinsic TSE. Here we give specific bounds on $G(x,z,\tilde{z})$ for the KL-divergence and the IS-distance.

For the generalized KL-divergence defined in (11), we have

$$\begin{aligned}&D_{\rm KL}\big((1-\theta)x+\theta z,(1-\theta)x+\theta\tilde{z}\big)\\ =\;& \sum_{i=1}^n\bigg(\big((1-\theta)x^{(i)}+\theta z^{(i)}\big)\log\bigg(\frac{(1-\theta)x^{(i)}+\theta z^{(i)}}{(1-\theta)x^{(i)}+\theta\tilde{z}^{(i)}}\bigg)-\theta\big(z^{(i)}-\tilde{z}^{(i)}\big)\bigg)\\ \overset{(*)}{\le}\;& \sum_{i=1}^n\bigg(\big((1-\theta)x^{(i)}+\theta z^{(i)}\big)\frac{\theta(z^{(i)}-\tilde{z}^{(i)})}{(1-\theta)x^{(i)}+\theta\tilde{z}^{(i)}}-\theta\big(z^{(i)}-\tilde{z}^{(i)}\big)\bigg)\\ =\;& \theta\sum_{i=1}^n\bigg(\frac{(1-\theta)x^{(i)}+\theta z^{(i)}}{(1-\theta)x^{(i)}+\theta\tilde{z}^{(i)}}-1\bigg)\big(z^{(i)}-\tilde{z}^{(i)}\big)\\ =\;& \theta^2\sum_{i=1}^n\frac{\big(z^{(i)}-\tilde{z}^{(i)}\big)^2}{(1-\theta)x^{(i)}+\theta\tilde{z}^{(i)}}, \qquad (16)\end{aligned}$$

where the inequality (*) used $\log(\alpha) \le \alpha - 1$ for all $\alpha > 0$. In order to make (13) hold with $\gamma = 2$, we can replace the denominator $(1-\theta)x^{(i)} + \theta\tilde{z}^{(i)}$ in (16) with $\min\{x^{(i)}, \tilde{z}^{(i)}\}$ and set

$$G_{\mathrm{KL}}(x, z, \tilde{z}) = \sum_{i=1}^{n} \frac{\big(z^{(i)} - \tilde{z}^{(i)}\big)^2}{\min\{x^{(i)}, \tilde{z}^{(i)}\}} \Big/ D_{\mathrm{KL}}(z, \tilde{z}). \tag{17}$$

For the IS-distance defined in (12), we can derive a similar bound (using $\log(\alpha) \le \alpha - 1$ again):

$$D_{\mathrm{IS}}\big((1-\theta)x + \theta z, (1-\theta)x + \theta\tilde{z}\big) \le \theta^2 \sum_{i=1}^{n} \frac{\big(z^{(i)} - \tilde{z}^{(i)}\big)^2}{\big((1-\theta)x^{(i)} + \theta z^{(i)}\big)\big((1-\theta)x^{(i)} + \theta\tilde{z}^{(i)}\big)}. \tag{18}$$

To satisfy (13) with $\gamma = 2$, we can set

$$G_{\mathrm{IS}}(x, z, \tilde{z}) = \sum_{i=1}^{n} \frac{\big(z^{(i)} - \tilde{z}^{(i)}\big)^2}{\big(\min\{x^{(i)}, z^{(i)}, \tilde{z}^{(i)}\}\big)^2} \Big/ D_{\mathrm{IS}}(z, \tilde{z}). \tag{19}$$

We note that the two upper bounds in (16) and (18) are asymptotically tight, meaning that they match the limit in (14) as $\theta \to 0$.

Remark. Suppose h is twice continuously differentiable. If $\|z - \tilde{z}\|$ is small, then by definition of the Bregman distance,

$$D_h(z, \tilde{z}) = \frac{1}{2}\big\langle \nabla^2 h(\tilde{z})(z - \tilde{z}), z - \tilde{z} \big\rangle + o(\|z - \tilde{z}\|^2). \tag{20}$$

Therefore, in the regime of $\theta \to 0$ and $\|z - \tilde{z}\|$ small, a good estimate based on (14) and (20) is

$$G(x, z, \tilde{z}) = O\left(\frac{\big\langle \nabla^2 h(x)(z - \tilde{z}), z - \tilde{z} \big\rangle}{\big\langle \nabla^2 h(\tilde{z})(z - \tilde{z}), z - \tilde{z} \big\rangle} \right).$$

If in addition $\nabla^2 h(x) \preceq \nabla^2 h(\tilde{z})$ or $\|x - \tilde{z}\|$ is small, then it suffices to have $G(x, z, \tilde{z}) = O(1)$.

3 Accelerated Bregman proximal gradient method

In this section, we present the accelerated Bregman proximal gradient (ABPG) method for solving problem (1), and analyze its convergence rate under the uniform triangle-scaling property. Adaptive variants based on the intrinsic TSE are developed in Section 4.

To simplify notation, we define a lower approximation of $\phi(x) = f(x) + \Psi(x)$ by linearizing f at a given point y:

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

If f is L-smooth relative to h (Definition 1), then we have both a lower and an upper approximation:

$$\ell(x|y) \;\le\; \phi(x) \;\le\; \ell(x|y) + L D_h(x, y). \tag{21}$$

Algorithm 1 describes the ABPG method. Its input parameters include a uniform TSE γ of D_h and an initial point $x_0 \in \operatorname{rint} C$. The sequence $\{\theta_k\}_{k \in \mathbb{N}}$ in Algorithm 1 satisfies $0 < \theta_k \le 1$ and

$$\frac{1 - \theta_{k+1}}{\theta_{k+1}^{\gamma}} \le \frac{1}{\theta_k^{\gamma}}, \qquad \forall\, k \ge 0. \tag{22}$$

Algorithm 1: Accelerated Bregman proximal gradient (ABPG) method

input: initial point $x_0 \in \operatorname{rint} C$ and $\gamma \geq 1$.

initialize: $z_0 = x_0$ and $\theta_0 = 1$.

for $k = 0, 1, 2, \ldots$ **do**

1 $y_k = (1 - \theta_k)x_k + \theta_k z_k$

2 $z_{k+1} = \arg\min_{z \in C} \big\{ \ell(z|y_k) + \theta_k^{\gamma-1} L D_h(z, z_k) \big\}$

3 $x_{k+1} = (1 - \theta_k)x_k + \theta_k z_{k+1}$

4 choose $\theta_{k+1} \in (0, 1]$ such that $\frac{1-\theta_{k+1}}{\theta_{k+1}^\gamma} \leq \frac{1}{\theta_k^\gamma}$

end

When $\gamma = 2$ and $\Psi \equiv 0$, Algorithm 1 reduces to the IGA (improved interior gradient algorithm) method in [2], which is an extension of Nesterov's accelerated gradient method in [18] to the Bregman proximal setting. It was shown in [2] that the IGA method attains $O(k^{-2})$ rate of convergence under the uniform Lipschitz condition 3. In this paper, we consider the general case $\gamma \in [1, 2]$ under the much weaker relatively smooth condition.

Using the definition of $\ell(\cdot|\cdot)$, line 2 in Algorithm 1 can be written as

$$z_{k+1} = \arg\min_{x \in C} \Big\{ f(y_k) + \langle \nabla f(y_k), x - y_k \rangle + \theta_k^{\gamma-1} L D_h(x, z_k) + \Psi(x) \Big\}, \tag{23}$$

which is very similar to the BPG step (8). Here the function f is linearized around y_k but the Bregman distance is measured from a different point z_k. Therefore it does not fit into the framework of majorization and the sequence $\phi(x_k)$ may not be monotone decreasing. However, the upper bound in (21) is still crucial to ensure convergence of the algorithm. Under the same assumption that the BPG step is well-posed (Assumption A.5), the ABPG method is also well-posed, meaning that $z_{k+1} \in \operatorname{rint} C$ always and it is unique.

3.1 Convergence analysis of ABPG

We show that the ABPG method converges with a sublinear rate of $O(k^{-\gamma})$. First, we state a basic property of optimization with Bregman distance [10, Lemma 3.2].

Lemma 1. *For any closed convex function $\varphi : \mathbb{R}^n \to (-\infty, \infty]$ and any $z \in \operatorname{rint} \operatorname{dom} h$, if*

$$z_+ = \arg\min_{x \in C} \big\{ \varphi(x) + D_h(x, z) \big\}$$

and h is differentiable at z_+, then

$$\varphi(x) + D_h(x, z) \geq \varphi(z_+) + D_h(z_+, z) + D_h(x, z_+), \quad \forall x \in \operatorname{dom} h.$$

The following lemma establishes a relationship between the two consecutive steps of Algorithm 1. It is an extension of Proposition 1 in [26] , which uses $\gamma = 2$ under the assumption (3).

Lemma 2. *Suppose Assumption A holds, f is L-smooth relative to h on C, and γ is a uniform TSE of D_h. For any $x \in \operatorname{dom} h$, the sequences generated by Algorithm 1 satisfy, for all $k \geq 0$,*

$$\frac{1 - \theta_{k+1}}{\theta_{k+1}^\gamma} \big(\phi(x_{k+1}) - \phi(x) \big) + L D_h(x, z_{k+1}) \;\leq\; \frac{1 - \theta_k}{\theta_k^\gamma} \big(\phi(x_k) - \phi(x) \big) + L D_h(x, z_k). \tag{24}$$

Proof. First, using the upper approximation in (21) and line 1 and line 3 in Algorithm 1, we have

$$\begin{aligned}\phi(x_{k+1}) &\le \ell(x_{k+1}|y_k) + LD_h(x_{k+1}, y_k) \\ &= \ell(x_{k+1}|y_k) + LD_h\big((1-\theta_k)x_k + \theta_k z_{k+1}, (1-\theta_k)x_k + \theta_k z_k\big) \\ &\le \ell(x_{k+1}|y_k) + \theta_k^\gamma LD_h(z_{k+1}, z_k), \end{aligned} \tag{25}$$

where in the last inequality we used the triangle-scaling property (10). Using $x_{k+1} = (1-\theta_k)x_k + \theta_k z_{k+1}$ and convexity of $\ell(\cdot|y_k)$, we have

$$\begin{aligned}\phi(x_{k+1}) &\le (1-\theta_k)\ell(x_k|y_k) + \theta_k \ell(z_{k+1}|y_k) + \theta_k^\gamma LD_h(z_{k+1}, z_k) \\ &= (1-\theta_k)\ell(x_k|y_k) + \theta_k \Big(\ell(z_{k+1}|y_k) + \theta_k^{\gamma-1} LD_h(z_{k+1}, z_k)\Big). \end{aligned} \tag{26}$$

Now applying Lemma 1 with $\varphi(x) = \ell(x|y_k)/(\theta^{\gamma-1}L)$ yields, for any $x \in \operatorname{dom} h$,

$$\ell(z_{k+1}|y_k) + \theta_k^{\gamma-1} LD_h(z_{k+1}, z_k) \;\le\; \ell(x|y_k) + \theta_k^{\gamma-1} LD_h(x, z_k) - \theta_k^{\gamma-1} LD_h(x, z_{k+1}).$$

Hence

$$\begin{aligned}\phi(x_{k+1}) &\le (1-\theta_k)\ell(x_k|y_k) + \theta_k \Big(\ell(x|y_k) + \theta_k^{\gamma-1} LD_h(x, z_k) - \theta_k^{\gamma-1} LD_h(x, z_{k+1})\Big) \\ &= (1-\theta_k)\ell(x_k|y_k) + \theta_k \ell(x|y_k) + \theta_k^\gamma \big(LD_h(x, z_k) - LD_h(x, z_{k+1})\big) \\ &\le (1-\theta_k)\phi(x_k) + \theta_k \phi(x) + \theta_k^\gamma \big(LD_h(x, z_k) - LD_h(x, z_{k+1})\big), \end{aligned}$$

where in the last inequality we used the lower bound in (21). Subtracting $\phi(x)$ from both sides of the inequality above, we obtain

$$\phi(x_{k+1}) - \phi(x) \;\le\; (1-\theta_k)\big(\phi(x_k) - \phi(x)\big) + \theta_k^\gamma \big(LD_h(x, z_k) - LD_h(x, z_{k+1})\big).$$

Dividing both sides by θ_k^γ and rearranging terms yield

$$\frac{1}{\theta_k^\gamma}\big(\phi(x_{k+1}) - \phi(x)\big) + LD_h(x, z_{k+1}) \;\le\; \frac{1-\theta_k}{\theta_k^\gamma}\big(\phi(x_k) - \phi(x)\big) + LD_h(x, z_k). \tag{27}$$

Finally applying the condition (22) gives the desired result. □

Lemma 3. *The sequence $\theta_k = \frac{\gamma}{k+\gamma}$ for $k = 0, 1, 2, \ldots$ satisfies the condition* (22).

Proof. With $\theta_k = \frac{\gamma}{k+\gamma}$, we have

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^\gamma} = \left(1 - \frac{\gamma}{k+1+\gamma}\right)\left(\frac{k+1+\gamma}{\gamma}\right)^\gamma = \frac{(k+1)(k+1+\gamma)^{\gamma-1}}{\gamma^\gamma} \tag{28}$$

and

$$\frac{1}{\theta_k^\gamma} = \left(\frac{k+\gamma}{\gamma}\right)^\gamma = \frac{(k+\gamma)^\gamma}{\gamma^\gamma}. \tag{29}$$

Recall the weighted arithmetic mean and geometric mean inequality (see, e.g., [13, Section 2.5].), i.e., for any positive real numbers a, b, α and β, it holds that

$$a^\alpha b^\beta \le \left(\frac{\alpha a + \beta b}{\alpha + \beta}\right)^{\alpha+\beta}. \tag{30}$$

Setting $a = k+1$, $b = k+1+\gamma$, $\alpha = 1$ and $\beta = \gamma - 1$, we arrive at

$$(k+1)(k+1+\gamma)^{\gamma-1} \leq \left(\frac{k+1+(\gamma-1)(k+1+\gamma)}{1+\gamma-1}\right)^{1+\gamma-1} = (k+\gamma)^{\gamma},$$

which, together with (28) and (29), implies the inequality (22). □

A slightly faster converging sequence θ_k can be obtained by solving the equality in (22). Since there is no closed-form solution in general, we can find θ_{k+1} as the root of

$$\theta^{\gamma} - \theta_k^{\gamma}(1-\theta) = 0 \tag{31}$$

numerically, say, using Newton's method with θ_k as the starting point.

Lemma 4. *Let $\theta_0 = 1$ and θ_{k+1} be the solution to* (31) *for all $k \geq 0$. Then $\theta_k \leq \frac{\gamma}{k+\gamma}$ for all $k \geq 0$.*

Proof. Let $\vartheta_k = \frac{\gamma}{k+\gamma}$ and define another sequence ξ_k such that $\xi_0 = 1$ and

$$\frac{1-\xi_{k+1}}{\xi_{k+1}^{\gamma}} = \frac{1}{\vartheta_k^{\gamma}}, \qquad \forall\, k \geq 0. \tag{32}$$

Notice that the function

$$\omega(\theta) := \frac{1-\theta}{\theta^{\gamma}}$$

is monotone decreasing in θ. Since $\omega(\vartheta_{k+1}) \leq 1/\vartheta_k^{\gamma}$ by Lemma 3 and $\omega(\xi_{k+1}) = 1/\vartheta_k^{\gamma}$ by (32), we have $\xi_{k+1} \leq \vartheta_{k+1}$ for all $k \geq 0$.

Next we prove $\theta_k \leq \vartheta_k$ for all $k \geq 0$ by mathematical induction. This obviously holds for $k = 0$ since $\theta_0 = \vartheta_0 = 1$. Suppose $\theta_k \leq \vartheta_k$ holds for some $k \geq 0$. Then using the facts $\omega(\theta_{k+1}) = 1/\theta_k^{\gamma}$ and $\omega(\xi_{k+1}) = 1/\vartheta_k^{\gamma}$, we obtain $\omega(\theta_{k+1}) \geq \omega(\xi_{k+1})$. Since ω is monotone decreasing, we conclude that $\theta_{k+1} \leq \xi_{k+1}$. Combining with $\xi_{k+1} \leq \vartheta_{k+1}$ obtained above, we have $\theta_{k+1} \leq \vartheta_{k+1}$. This completes the induction. □

Theorem 2. *Suppose Assumption A holds, f is L-smooth relative to h on C, and γ is a uniform TSE of D_h. If $\theta_k \leq \frac{\gamma}{k+\gamma}$ for all $k \geq 0$, then the outputs of Algorithm 1 satisfy, for any $x \in \operatorname{dom} h$,*

$$\phi(x_{k+1}) - \phi(x) \leq \left(\frac{\gamma}{k+\gamma}\right)^{\gamma} L D_h(x, x_0), \qquad \forall\, k \geq 0.$$

Proof. A direct consequence of Lemma 2 is, for any $x \in \operatorname{dom} h$,

$$\frac{1-\theta_k}{\theta_k^{\gamma}}\big(\phi(x_k) - \phi(x)\big) + L D_h(x, z_k) \leq \frac{1-\theta_0}{\theta_0}\big(\phi(x_0) - \phi(x)\big) + L D_h(x, z_0).$$

Combining with (27), we have

$$\frac{1}{\theta_k^{\gamma}}\big(\phi(x_{k+1}) - \phi(x)\big) + L D_h(x, z_{k+1}) \leq \frac{1-\theta_0}{\theta_0}\big(\phi(x_0) - \phi(x)\big) + L D_h(x, z_0).$$

Using $D_h(x, z_{k+1}) \geq 0$ and the initializations $\theta_0 = 1$ and $z_0 = x_0$, we obtain

$$\frac{1}{\theta_k^{\gamma}}\big(\phi(x_{k+1}) - \phi(x)\big) \leq L D_h(x, z_0),$$

which implies

$$\phi(x_{k+1}) - \phi(x) \leq \theta_k^{\gamma} L D_h(x, x_0).$$

It remains to apply the condition $\theta_k \leq \frac{\gamma}{k+\gamma}$. □

Algorithm 2: ABPG method with exponent adaption (ABPG-e)

input: initial point $x_0 \in \operatorname{rint} C$, $\gamma_0 \geq 2$, $\gamma_{\min} > 0$, and $\delta > 0$.

initialize: $z_0 = x_0$, $\gamma_{-1} = \gamma_0$, and $\theta_0 = 1$.

for $k = 0, 1, 2, \ldots$ **do**

  $y_k = (1-\theta_k)x_k + \theta_k z_k$

  repeat for $t = 0, 1, 2, \ldots$

    $\gamma_k = \max\{\gamma_{k-1} - \delta t,\ \gamma_{\min}\}$

    $z_{k+1} = \arg\min_{z \in C}\big\{\ell(z|y_k) + \theta_k^{\gamma_k - 1} L D_h(z, z_k)\big\}$

    $x_{k+1} = (1-\theta_k)x_k + \theta_k z_{k+1}$

  until $f(x_{k+1}) \leq f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k\rangle + \theta_k^{\gamma_k} L D_h(z_{k+1}, z_k)$

  choose $\theta_{k+1} \in (0,1]$ such that $\frac{1-\theta_{k+1}}{\theta_{k+1}^{\gamma_k}} \leq \frac{1}{\theta_k^{\gamma_k}}$

end

3.2 ABPG method with exponent adaption

The best convergence rate of the ABPG method is obtained with the largest uniform TSE for the Bregman distance. Since it is often hard to determine the largest TSE, we present in Algorithm 2 a variant of the ABPG method with automatic exponent adaption, called the ABPG-e method.

This method starts with a large $\gamma_0 \geq 2$. During each iteration k, it reduces γ_k by a small amount $\delta > 0$ until some stopping criterion is satisfied. An obvious choice for the stopping criterion is the local triangle-scaling property

$$D_h(x_{k+1}, y_k) \leq \theta_k^{\gamma_k} D_h(z_{k+1}, z_k), \tag{33}$$

where $x_{k+1} = (1-\theta_k)x_k + \theta_k z_{k+1}$ and $y_k = (1-\theta_k)x_k + \theta_k z_k$. According to the proof of Lemma 2, we can also use the inequality (25) as stopping criterion, which is implied by (33) and the relatively smooth assumption. For convergence, we only need (25) to hold, which can be less conservative than (33). In Algorithm 2, we use the following inequality as the stopping criterion

$$f(x_{k+1}) \leq f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k\rangle + \theta_k^{\gamma_k} L D_h(z_{k+1}, z_k),$$

which is equivalent to (25) (by subtracting $\Psi(x_{k+1})$ from both sides of the inequality). In practice, this condition often leads to much faster convergence than using (33). Computationally, it is slightly more expensive since it needs to evaluate $f(x_{k+1})$ in addition to $\nabla f(y_k)$ during each inner loop, while (33) does not.

The lower bound $\gamma_{\min}$ can be any known uniform TSE, which guarantees that the stopping criterion can always be satisfied. Since $\gamma_{k+1} \leq \gamma_k$ and $\theta_{k+1} \in (0,1)$, we always have $\theta_{k+1}^{\gamma_{k+1}} \geq \theta_{k+1}^{\gamma_k}$. Therefore

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^{\gamma_{k+1}}} \leq \frac{1-\theta_{k+1}}{\theta_{k+1}^{\gamma_k}} \leq \frac{1}{\theta_k^{\gamma_k}}.$$

By replacing inequality (22) with the one above and repeating the analysis in Section 3.1, we obtain the following result.

Algorithm 3: ABPG method with monotone gain adaption

input: initial points $x_0 \in C$, $\gamma > 1$, and $\rho > 1$.

initialize: $z_0 = x_0$, $\theta_0 = 1$, $G_{-1} = 1$.

for $k = 0, 1, 2, \ldots$ **do**

  $y_k = (1-\theta_k)x_k + \theta_k z_k$

  repeat for $t = 0, 1, 2, \ldots$

    $G_k = G_{k-1}\rho^t$

    $z_{k+1} = \arg\min_{z\in C}\{\ell(z|y_k) + G_k\theta_k^{\gamma-1}LD_h(z,z_k)\}$

    $x_{k+1} = (1-\theta_k)x_k + \theta_k z_{k+1}$

  until $f(x_{k+1}) \le f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k\rangle + G_k\theta_k^{\gamma}LD_h(z_{k+1}, z_k)$

  choose $\theta_{k+1} \in (0,1]$ such that $\frac{1-\theta_{k+1}}{\theta_{k+1}^{\gamma}} \le \frac{1}{\theta_k^{\gamma}}$

end

Theorem 3. *Suppose Assumption A holds, f is L-smooth relative to h on C, and $\gamma_{\min}$ is a uniform TSE of D_h. Then the sequences generated by Algorithm 2 satisfy, for any $x \in \operatorname{dom} h$,*

$$\phi(x_{k+1}) - \phi(x) \le \left(\frac{\gamma_k}{k+\gamma_k}\right)^{\gamma_k} LD_h(x, x_0), \qquad \forall\, k\ \ge 0.$$

The convergence rate of ABPG-e is determined by the last value γ_k. Since we only need to satisfy the local triangle-scaling property (33) instead of the uniform condition (10), it is very likely that γ_k is greater than the largest uniform TSE. However, according to Theorem 1, when $k \to \infty$, the limit of γ_k (which always exists) cannot be larger than the intrinsic TSE $\gamma_{\text{in}} = 2$.

4 ABPG methods with gain adaption

In this section, we present and analyze adaptive ABPG methods based on the concept of intrinsic TSE developed in Section 2.1. Instead of searching for the largest uniform TSE as in Algorithm 2, we can replace line 2 in Algorithm 1 by

$$z_{k+1} = \operatorname*{arg\,min}_{z\in C}\left\{\ell(z|y_k) + G_k\theta_k^{\gamma-1}LD_h(z,z_k)\right\}$$

and adjust the gain G_k while keeping $\gamma = \gamma_{\text{in}}$ fixed. Algorithm 3 is such a method with monotone gain adaption, meaning that $G_{k+1} \ge G_k$ for all $k \ge 0$. Let $\rho > 1$ be an adaption parameter. During each iteration k, it finds the smallest integer $t \ge 0$ such that $G_k = G_{k-1}\rho^t$ satisfies the stopping criterion

$$f(x_{k+1}) \le f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k\rangle + G_k\theta_k^{\gamma_k}LD_h(z_{k+1}, z_k), \tag{34}$$

which is implied by the relative smoothness and the local triangle-scaling property

$$D_h(x_{k+1}, y_k) = D_h\big((1-\theta)x_k + \theta z_{k+1}, (1-\theta)x_k + \theta z_k\big) \le G_k\theta^{\gamma}D_h(z_{k+1}, z_k). \tag{35}$$

By definition of the intrinsic TSE, such a G_k always exists for $\gamma = \gamma_{\text{in}}$, i.e., the stopping criterion for gain adaption in Algorithm 3 can always be satisfied.

Theorem 4. *Suppose Assumption A holds, f is L-smooth relative to h on C, and $\gamma = \gamma_{\rm in}$ is the intrinsic TSE of D_h. Then the sequences generated by Algorithm 3 satisfy, for any $x \in \operatorname{dom} h$,*

$$\phi(x_{k+1}) - \phi(x) \le \left(\frac{\gamma}{k+\gamma}\right)^{\gamma} G_k L D_h(x, x_0), \qquad \forall\, k \ge 0. \tag{36}$$

Proof. We follow the same steps as in Section 3.1. In light of (35), the inequality (25) becomes

$$\phi(x_{k+1}) \le \ell(x_{k+1}|y_k) + G_k\theta_k^{\gamma} L D_h(z_{k+1}, z_k), \tag{37}$$

and the inequality (27) becomes

$$\frac{1}{G_k\theta_k^{\gamma}}\big(\phi(x_{k+1}) - \phi(x)\big) + LD_h(x, z_{k+1}) \ \le\ \frac{1-\theta_k}{G_k\theta_k^{\gamma}}\big(\phi(x_k) - \phi(x)\big) + LD_h(x, z_k). \tag{38}$$

Since $\{\theta_k\}_{k\in\mathbb{N}}$ satisfy $\frac{1-\theta_{k+1}}{\theta_{k+1}^{\gamma}} \le \frac{1}{\theta_k^{\gamma}}$ and the algorithm ensures $G_{k+1} \ge G_k$, we have

$$\frac{1-\theta_{k+1}}{G_{k+1}\theta_{k+1}^{\gamma}} \le \frac{1-\theta_{k+1}}{G_k\theta_{k+1}^{\gamma}} \le \frac{1}{G_k\theta_k^{\gamma}}.$$

Thus the following inequality replaces (24) in Lemma 2:

$$\frac{1-\theta_{k+1}}{G_{k+1}\theta_{k+1}^{\gamma}}\big(\phi(x_{k+1}) - \phi(x)\big) + LD_h(x, z_{k+1}) \ \le\ \frac{1-\theta_k}{G_k\theta_k^{\gamma}}\big(\phi(x_k) - \phi(x)\big) + LD_h(x, z_k). \tag{39}$$

The rest of the proof is similar to that for Theorem 2. □

4.1 ABPG method with non-monotone gain adaption

Although the gain adaption loop in Algorithm 3 always exits with a finite G_k that satisfies (34), it could be very large. More importantly, since $\{G_k\}_{k\in\mathbb{N}}$ is monotone non-decreasing, the algorithm may stuck with some large G_k even if much smaller gains would work for later iterations. The convergence rate obtained in Theorem 4 depends on the last G_k, which is also the largest gain up to iteration k.

Algorithm 4 describes another variant of ABPG with an adaptive, non-monotone gain search scheme. At the beginning of each iteration k, a smaller tentative gain, $M_k = \max\{G_{k-1}/\rho,\ G_{\min}\}$ where $\rho > 1$, is first proposed. The gain adaption loop finds the smallest integer $t \ge 0$ such that $G_k = M_k\rho^t$ and the corresponding vectors y_k, z_{k+1}, x_{k+1} satisfy the inequality (34).

Another major difference between Algorithm 4 and all previous variants is that the sequence $\{\theta_k\}_{k\in\mathbb{N}}$ in Algorithm 4 is generated by solving the equation

$$\frac{1-\theta_{k+1}}{G_{k+1}\theta_{k+1}^{\gamma}} = \frac{1}{G_k\theta_k^{\gamma}}. \tag{40}$$

Since we don't have a priori bounds on the gains G_k, it is hard to characterize how fast θ_k converges to zero. In fact, $\{\theta_k\}_{k\in\mathbb{N}}$ may not be a monotone decreasing sequence. Instead of tracking G_k and θ_k separately, we analyze the convergence of the combined quantity $G_k\theta_k^{\gamma}$. The following simple lemma will be very useful.

Algorithm 4: ABPG method with non-monotone gain adaption (ABPG-g)

input: initial points $x_0 \in C$, $\gamma > 1$, $\rho > 1$ and $G_{\min} > 0$.

initialize: $z_0 = x_0$, $G_{-1} = 1$

for $k = 0, 1, 2, \ldots$ **do**

  $M_k = \max\{G_{k-1}/\rho,\ G_{\min}\}$

  repeat for $t = 0, 1, 2, \ldots$

    $G_k = M_k \rho^t$

    if $k > 0$ **then** compute θ_k by solving $\dfrac{1-\theta_k}{G_k \theta_k^\gamma} = \dfrac{1}{G_{k-1}\theta_{k-1}^\gamma}$

    $y_k = (1-\theta_k)x_k + \theta_k z_k$

    $z_{k+1} = \arg\min_{z \in C} \left\{ \ell(z|y_k) + G_k \theta_k^{\gamma-1} L D_h(z, z_k) \right\}$

    $x_{k+1} = (1-\theta_k)x_k + \theta_k z_{k+1}$

  until $f(x_{k+1}) \le f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k \rangle + G_k \theta_k^\gamma L D_h(z_{k+1}, z_k)$

end

Lemma 5. *For any $\alpha, \beta > 0$ and $\gamma \ge 1$, the following inequality holds:*

$$\alpha^\gamma - \beta^\gamma \ \le\ \gamma(\alpha - \beta)\alpha^{\gamma-1}.$$

Proof. The case of $\gamma = 1$ is obvious. Assume $\gamma > 1$. The desired inequality is equivalent to

$$\alpha^{\gamma-1}\beta \ \le\ \frac{(\gamma-1)\alpha^\gamma + \beta^\gamma}{\gamma} \ =\ \frac{(\gamma-1)\alpha^\gamma + 1 \cdot \beta^\gamma}{(\gamma-1)+1}.$$

Applying the weighted arithmetic and geometric mean inequality (30), we have

$$\frac{(\gamma-1)\alpha^\gamma + 1 \cdot \beta^\gamma}{(\gamma-1)+1} \ \ge\ \Big((\alpha^\gamma)^{\gamma-1} (\beta^\gamma)^1 \Big)^{\frac{1}{\gamma}} \ =\ \alpha^{\gamma-1}\beta,$$

which completes the proof. □

Theorem 5. *Suppose Assumption A holds, f is L-smooth relative to h on C, and $\gamma = \gamma_{\rm in}$ is the intrinsic TSE of D_h. Then the sequences generated by Algorithm 4 satisfy, for any $x \in \operatorname{dom} h$,*

$$\phi(x_{k+1}) - \phi(x) \ \le\ \left(\frac{\gamma}{k+\gamma}\right)^\gamma \overline{G}_k L D_h(x, x_0), \qquad \forall\, k \ge 0, \tag{41}$$

where $\overline{G}_k$ is a weighted geometric mean of the gains at each step:

$$\overline{G}_k \ =\ \left(G_0^\gamma G_1 \cdots G_k\right)^{\frac{1}{k+\gamma}}. \tag{42}$$

Proof. Following the analysis outlined in the proof of Theorem 4, and using the equality (40), we can show that (39) still holds. Then the same arguments in the proof of Theorem 2 lead to

$$\phi(x_{k+1}) - \phi(x) \ \le\ G_k \theta_k^\gamma L D_h(x, x_0). \tag{43}$$

Next we derive an upper bound for $G_k\theta_k^\gamma$. For convenience, let's define for $k = 0, 1, 2, \ldots$,

$$A_k = \frac{1}{G_k\theta_k^\gamma}, \qquad a_{k+1} = \frac{1}{G_{k+1}\theta_{k+1}^{\gamma-1}}.$$

Then (40) implies $a_{k+1} = A_{k+1} - A_k$. Moreover, we have

$$A_{k+1} \;=\; \frac{1}{G_{k+1}\theta_{k+1}^\gamma} \;=\; G_{k+1}^{\frac{1}{\gamma-1}} a_{k+1}^{\frac{\gamma}{\gamma-1}} \;=\; G_{k+1}^{\frac{1}{\gamma-1}}\Big(A_{k+1} - A_k\Big)^{\frac{\gamma}{\gamma-1}}. \tag{44}$$

Applying Lemma 5 with $\alpha = A_{k+1}^{1/\gamma}$ and $\beta = A_k^{1/\gamma}$, we obtain

$$A_{k+1} - A_k \;=\; \Big(A_{k+1}^{\frac{1}{\gamma}}\Big)^\gamma - \Big(A_k^{\frac{1}{\gamma}}\Big)^\gamma \;\le\; \gamma\Big(A_{k+1}^{\frac{1}{\gamma}} - A_k^{\frac{1}{\gamma}}\Big)A_{k+1}^{\frac{\gamma-1}{\gamma}}.$$

Combining with (44) yields

$$A_{k+1} \;=\; G_{k+1}^{\frac{1}{\gamma-1}}\Big(A_{k+1} - A_k\Big)^{\frac{\gamma}{\gamma-1}} \;\le\; G_{k+1}^{\frac{1}{\gamma-1}}\gamma^{\frac{\gamma}{\gamma-1}}\Big(A_{k+1}^{\frac{1}{\gamma}} - A_k^{\frac{1}{\gamma}}\Big)^{\frac{\gamma}{\gamma-1}}A_{k+1}$$

We can eliminate the common factor A_{k+1} on both sides of the above inequality to obtain

$$1 \;\le\; G_{k+1}^{\frac{1}{\gamma-1}}\gamma^{\frac{\gamma}{\gamma-1}}\Big(A_{k+1}^{\frac{1}{\gamma}} - A_k^{\frac{1}{\gamma}}\Big)^{\frac{\gamma}{\gamma-1}},$$

which implies

$$A_{k+1}^{\frac{1}{\gamma}} - A_k^{\frac{1}{\gamma}} \;\ge\; \frac{1}{\gamma\, G_{k+1}^{1/\gamma}}, \qquad k = 0, 1, 2, \ldots.$$

Summing the above inequality from step 0 to $k-1$ and using $A_0 = G_0$, we have

$$A_k^{\frac{1}{\gamma}} \;\ge\; \sum_{t=1}^k \frac{1}{\gamma\, G_t^{1/\gamma}} + A_0^{\frac{1}{\gamma}} \;=\; \sum_{t=1}^k \frac{1}{\gamma\, G_t^{1/\gamma}} + \frac{1}{G_0^{1/\gamma}} \;=\; \frac{1}{\gamma}\left(\sum_{t=1}^k \frac{1}{G_t^{1/\gamma}} + \frac{\gamma}{G_0^{1/\gamma}}\right).$$

Using the weighted arithmetic and geometric mean inequality (e.g., [13, Section 2.5]) gives

$$\sum_{t=1}^k \frac{1}{G_t^{1/\gamma}} + \frac{\gamma}{G_0^{1/\gamma}} \;\ge\; (k+\gamma)\left(\left(\frac{1}{G_0^{1/\gamma}}\right)^\gamma \frac{1}{G_1^{1/\gamma}}\cdots\frac{1}{G_k^{1/\gamma}}\right)^{\frac{1}{k+\gamma}} \;=\; (k+\gamma)\Big(G_0^\gamma G_1\cdots G_k\Big)^{\frac{-1}{\gamma(k+\gamma)}}.$$

Combining the last two inequalities above, we arrive at

$$A_k \;\ge\; \left(\frac{k+\gamma}{\gamma}\right)^\gamma \left(G_0^\gamma G_1\cdots G_k\right)^{\frac{-1}{k+\gamma}}.$$

Therefore,

$$G_k\theta_k^\gamma \;=\; \frac{1}{A_k} \;\le\; \left(\frac{\gamma}{k+\gamma}\right)^\gamma \left(G_0^\gamma G_1\cdots G_k\right)^{\frac{1}{k+\gamma}}.$$

Finally, substituting the inequality above into (43) gives the desired result. □

The geometric mean $\overline{G}_k$ in (42) can be much smaller than $G_{\max} = \max\{G_0, G_1, \ldots, G_k\}$. Thus the convergence rate in (41) can be much faster than the one in (36) where the gains are monotone non-decreasing and thus $G_k = G_{\max}$.

Under the assumption of uniform Lipschitz smoothness (3), Nesterov [21] proposed an accelerated gradient method with non-monotone line search. However, the complexity obtained there still depends on the global Lipschitz constant L_f, more specifically, replacing $\overline{G}_k L$ in (41) with ρL_f when $\gamma = 2$. Our result in (41) can be more tight if the local Lipschitz constants are smaller.

Total number of oracle calls. We follow the approach of [21, Lemma 4]. Notice that each inner loop needs to call a gradient oracle to compute $\nabla f(y_k)$, and also $f(x_{k+1})$ when we use (34) as the stopping criterion for gain adaption. Let $n_i \geq 1$ be the number of calls of the oracle (for $\nabla f(y_k)$) at the ith iteration, for $i = 0, \ldots, k$. Then

$$G_{i+1} = \max\{G_i/\rho, G_{\min}\}\rho^{n_i-1} \geq G_i\rho^{n_i-2}, \qquad i = 0, \ldots, k-1.$$

Thus

$$n_i \leq 2 + \log_\rho \frac{G_{i+1}}{G_i} = 2 + \frac{1}{\ln\rho} \ln \frac{G_{i+1}}{G_i}.$$

Therefore, the total number of oracle calls is

$$N_k = \sum_{i=0}^{k} n_i \leq \sum_{i=0}^{k} \left(2 + \frac{1}{\ln\rho} \ln \frac{G_{i+1}}{G_i}\right) = 2(k+1) + \frac{1}{\ln\rho} \ln \frac{G_k}{G_0}.$$

Roughly speaking, on average each iteration need two oracle calls (unless G_k is very large).

An explicit update rule for θ_k. As an alternative to calculating θ_{k+1} by solving the equation (40), we can also use the following explicit update rule:

$$\frac{1}{\theta_{k+1}} \;=\; \frac{\gamma\alpha_k}{1+\alpha_k(\gamma-1)}\frac{1}{\theta_k} + \frac{1}{1+\alpha_k(\gamma-1)},$$

where $\alpha_k = G_{k+1}/G_k$ for $k = 0, 1, 2, \ldots$. This recursion is obtained by solving a linearized equation of (40). In particular, if $\alpha_k = 1$ for all $k \geq 0$, then this formula produces $\theta_k = \gamma/(k+\gamma)$. The sequence $\{\theta_k\}_{k\in\mathbb{N}}$ generated this way satisfies an inequality obtained by replacing the "=" sign with "$\leq$" in (40). Although Theorem 5 does not apply to this sequence, it often has comparable or even faster performance in practice, especially when the α_k's are close to 1.

4.2 Towards the $O(k^{-2})$ convergence rates

Theorem 1 shows that the intrinsic TSE $\gamma_{\text{in}} = 2$ for all Bregman distances D_h where h is convex and twice continuously differentiable. This covers most Bregman distances of practical interest. For the ABPG-g method (Algorithm 4) to obtain true $O(k^{-2})$ convergence rate, we need to make sure that $\overline{G}_k$ defined in (42) is $O(1)$.

If the sequence $\{z_k\}_{k\in\mathbb{N}}$ converges, according to the remark at the end of Section 2.2, we have

$$G_k \;\leq\; \rho\, G(x_{k+1}, z_{k+1}, z_k) \;=\; O\left(\frac{\langle \nabla^2 h(x_{k+1})(z_{k+1}-z_k), z_{k+1}-z_k\rangle}{\langle \nabla^2 h(z_k)(z_{k+1}-z_k), z_{k+1}-z_k\rangle}\right) \tag{45}$$

when k is large. If in addition $\|x_{k+1} - z_k\|$ is small, certainly if $\{x_k\}$ and $\{z_k\}$ converges to the same point, then $G_k = O(1)$ and so is the geometric mean $\overline{G}_k$ when k is large.

For concrete discussion, we consider relatively smooth optimization with the generalized KL-divergence and the IS-distance. If all coordinates of x_k and z_k are bounded away from zero, then we can easily bound G_k using (17) or (19), thus obtain the $O(k^{-2})$ rate. A particularly interesting case is when some of the coordinates $x_k^{(i)} \to 0$. Since $x_{k+1} = (1-\theta_k)x_k + \theta_k z_{k+1}$, we know that the sequence $\{x_k\}$ is obtained by taking convex combinations of the sequence $\{z_k\}$. If both sequences

Algorithm 5: Accelerated Bregman dual averaging (ABDA) method

input: initial point $z_0 \in \operatorname{rint} C$ and $\gamma > 1$.

initialize: $x_0 = z_0$, $\psi_0(x) \equiv 0$, and $\theta_0 = 1$.

for $k = 0, 1, 2, \ldots$ **do**

1 $y_k := (1-\theta_k)x_k + \theta_k z_k$

2 $\psi_{k+1}(x) := \psi_k(x) + \theta_k^{1-\gamma}\ell(x|y_k)$

3 $z_{k+1} := \arg\min_{z\in C}\bigl\{\psi_{k+1}(z) + Lh(z)\bigr\}$

4 $x_{k+1} := (1-\theta_k)x_k + \theta_k z_{k+1}$

5 find $\theta_{k+1} \in (0,1]$ such that $\frac{1-\theta_{k+1}}{\theta_{k+1}^{\gamma}} = \frac{1}{\theta_k^{\gamma}}$

end

converge and some coordinates $x_k^{(i)} \to 0$, we must also have the corresponding $z_k^{(i)} \to 0$, indeed at the same or a faster rate because $\theta_k \to 0$. Therefore, even though the diagonal entries of the Hessian $\nabla_{ii}^2 h(x_k) \to \infty$ (as $1/x_k^{(i)}$ for KL-divergence or $1/(x_k^{(i)})^2$ for IS-distance), the corresponding entries $\nabla_{ii}^2 h(z_k) \to \infty$ at the same or a faster rate. Hence, according to (45), we still have $G_k = O(1)$.

While a formal proof of the above arguments may require additional technical assumptions and more careful analysis, we note that the sequence $\{G_k\}$ is readily available as part of the computation and we can easily check the magnitude of $\overline{G}_k$. When it is small, we obtain a numerical certificate that the algorithm did converge with the $O(k^{-2})$ rate. This is exactly what we observe in the numerical experiments in Section 6.

5 Accelerated Bregman dual averaging method

In this section, we present an accelerated Bregman dual averaging (ABDA) method under the relative smoothness assumption. This method extends Nesterov's accelerated dual averaging method ([20] and [26, Algorithm 3]) to the relatively smooth setting. Here we focus on a simple variant in Algorithm 5 based on the uniform triangle-scaling property, although it is also possible to develop more sophisticated variants with automatic exponent or gain adaption.

Line 2 in Algorithm 5 defines a sequence of functions $\{\psi_k\}_{k\in\mathbb{N}}$ starting with $\psi_0 \equiv 0$:

$$\psi_{k+1}(x) := \psi_k(x) + \theta_k^{1-\gamma}\ell(x|y_k). \tag{46}$$

In other words, ψ_{k+1} is a weighted sum of the lower approximations in (21) constructed at $y_0, \ldots, y_k$:

$$\psi_{k+1}(x) = \sum_{t=0}^{k} \theta_t^{1-\gamma}\ell(x|y_t). \tag{47}$$

Line 3 in Algorithm 5 can be written as

$$z_{k+1} = \operatorname*{arg\,min}_{z\in C}\bigl\{\langle g_k, z\rangle + \vartheta_k\Psi(z) + Lh(z)\bigr\} \tag{48}$$

where

$$g_k = \sum_{t=1}^{k}\theta_t^{1-\gamma}\nabla f(y_t), \qquad \vartheta_k = \sum_{t=1}^{k}\theta_t^{1-\gamma}.$$

When implementing Algorithm 5, we only need to keep track of g_k and ϑ_k, and there is no need to maintain the abstract form of $\psi_k(x)$. Here our assumption of C and Ψ being simple means that the minimization problem in (48) can be solved efficiently. This requirement is equivalent to that for the BPG method (8) and all variants of the ABPG methods in this paper.

Algorithm 5 (line 5) requires the sequence $\{\theta_k\}_{k\in\mathbb{N}}$ satisfy

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^{\gamma}} = \frac{1}{\theta_k^{\gamma}}, \qquad \forall\, k \geq 0. \tag{49}$$

Under this condition, we can show

$$\vartheta_k = \sum_{i=0}^{k} \theta_i^{1-\gamma} = \frac{1}{\theta_k^{\gamma}}. \tag{50}$$

To see this, we use induction. Clearly it holds for $k=0$ if we choose $\theta_0 = 1$. Suppose it holds for some $k \geq 0$, then in light of (50) and (49),

$$\vartheta_{k+1} = \sum_{i=0}^{k+1} \frac{1}{\theta_i^{\gamma-1}} = \frac{1}{\theta_k^{\gamma}} + \frac{1}{\theta_{k+1}^{\gamma-1}} = \frac{1-\theta_{k+1}}{\theta_{k+1}^{\gamma}} + \frac{1}{\theta_{k+1}^{\gamma-1}} = \frac{1-\theta_{k+1}+\theta_{k+1}}{\theta_{k+1}^{\gamma}} = \frac{1}{\theta_{k+1}^{\gamma}}.$$

Therefore the inequality (50) holds for all $k \geq 0$.

To analyze the convergence of Algorithm 5, we need the following simple variant of Lemma 1.

Lemma 6. *Suppose h is convex and differentiable on* $\operatorname{rint} C$*. For any closed convex function φ, if*

$$z = \operatorname*{arg\,min}_{x\in C} \bigl\{\varphi(x) + h(x)\bigr\}$$

and h is differentiable at z, then

$$\varphi(x) + h(x) \geq \varphi(z) + h(z) + D_h(x,z), \quad \forall\, x \in \operatorname{dom} h.$$

Lemma 7. *Suppose Assumption A holds, f is L-smooth relative to h on C, and γ is a uniform TSE of D_h. Then the sequences generated by Algorithm 5 satisfy, for all $x \in \operatorname{dom} h$ and all $k \geq 1$,*

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^{\gamma}}\phi(x_{k+1}) - \psi_{k+1}(z_{k+1}) - Lh(z_{k+1}) \;\leq\; \frac{1-\theta_k}{\theta_k^{\gamma}}\phi(x_k) - \psi_k(z_k) - Lh(z_k). \tag{51}$$

Proof. We can start with the inequality (26):

$$\begin{aligned}
\phi(x_{k+1}) \;&\leq\; (1-\theta_k)\ell(x_k|y_k) + \theta_k \ell(z_{k+1}|y_k) + \theta_k^{\gamma} L D_h(z_{k+1}, z_k) \\
&=\; (1-\theta_k)\ell(x_k|y_k) + \theta_k^{\gamma}\left(\theta_k^{1-\gamma}\ell(z_{k+1}|y_k) + LD_h(z_{k+1}, z_k)\right) \\
&\leq\; (1-\theta_k)\phi(x_k) + \theta_k^{\gamma}\left(\theta_k^{1-\gamma}\ell(z_{k+1}|y_k) + LD_h(z_{k+1}, z_k)\right).
\end{aligned} \tag{52}$$

Notice that for $k \geq 1$, z_k is the minimizer of $\psi_k(z) + Lh(z)$ over C. We use Lemma 6 to obtain

$$\psi_k(z_k) + Lh(z_k) + LD_h(z_{k+1}, z_k) \leq \psi_k(z_{k+1}) + Lh(z_{k+1}),$$

which gives

$$LD_h(z_{k+1}, z_k) \le \psi_k(z_{k+1}) + Lh(z_{k+1}) - \psi_k(z_k) - Lh(z_k). \tag{53}$$

Combining the inequalities (52) and (53), we obtain

$$\begin{aligned} \phi(x_{k+1}) &\le (1-\theta_k)\phi(x_k) + \theta_k^\gamma \Big(\theta_k^{1-\gamma}\ell(z_{k+1}|y_k) + \psi_k(z_{k+1}) + Lh(z_{k+1}) - \psi_k(z_k) - Lh(z_k)\Big) \\ &= (1-\theta_k)\phi(x_k) + \theta_k^\gamma\big(\psi_{k+1}(z_{k+1}) + Lh(z_{k+1}) - \psi_k(z_k) - Lh(z_k)\big), \end{aligned}$$

where in the last equality we used recursive definition of ψ_{k+1} in (46). Dividing both sides of the above inequality by θ_k^γ, we have

$$\frac{1}{\theta_k^\gamma}\phi(x_{k+1}) \le \frac{1-\theta_k}{\theta_k^\gamma}\phi(x_k) + \psi_{k+1}(z_{k+1}) + Lh(z_{k+1}) - \psi_k(z_k) - Lh(z_k).$$

Using (49) and rearranging terms gives the desired result (51), which holds for $k \ge 1$. □

Theorem 6. *Suppose Assumption A holds, f is L-smooth relative to h on C, and γ is a uniform TSE of D_h. The sequences generated by Algorithm 5 satisfy:*

(a) if $z_0 = \arg\min_{z\in C} h(z)$, then for any $x \in \operatorname{dom} h$,

$$\phi(x_{k+1}) - \phi(x) \le \left(\frac{\gamma}{k+\gamma}\right)^\gamma L\big(h(x) - h(z_0)\big), \qquad \forall\, k \ge 0; \tag{54}$$

(b) otherwise, for any $x \in \operatorname{dom} h$,

$$\phi(x_{k+1}) - \phi(x) \le \left(\frac{\gamma}{k+\gamma}\right)^\gamma L\big(h(x) - h(z_1) + D_h(z_1, z_0)\big), \qquad \forall\, k \ge 0. \tag{55}$$

Proof. If $z_0 = \arg\min_{z\in C} h(z)$, we use the definition $\psi_0 \equiv 0$ to conclude that

$$z_0 = \underset{z\in C}{\arg\min}\big\{\psi_0(z) + Lh(z)\big\}.$$

In this case, we can extend the result of Lemma 7 to hold for all $k \ge 0$. Applying the inequality (51) for iterations $0, 1, \ldots, k$, we obtain

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^\gamma}\phi(x_{k+1}) - \psi_{k+1}(z_{k+1}) - Lh(z_{k+1}) \le \frac{1-\theta_0}{\theta_0^\gamma}\phi(x_0) - \psi_0(z_0) - Lh(z_0) = -Lh(z_0),$$

where we used $\theta_0 = 1$ and $\psi_0 \equiv 0$. Next using (49) and rearranging terms, we have

$$\begin{aligned} \frac{1}{\theta_k^\gamma}\phi(x_{k+1}) &\le \psi_{k+1}(z_{k+1}) + Lh(z_{k+1}) - Lh(z_0) \\ &\le \psi_{k+1}(x) + Lh(x) - Lh(x_0) \qquad (56) \\ &= \sum_{t=0}^{k}\theta_t^{1-\gamma}\ell(x|y_t) + L\big(h(x) - h(z_0)\big) \\ &\le \sum_{t=0}^{k}\theta_t^{1-\gamma}\phi(x) + L\big(h(x) - h(z_0)\big) \\ &= \frac{1}{\theta_k^\gamma}\phi(x) + L\big(h(x) - h(z_0)\big), \qquad (57) \end{aligned}$$

where the second inequality used the fact that z_{k+1} is the minimizer of $\psi_{k+1}(z) + Lh(z)$, the third inequality used $\ell(x|y_t) \le \phi(x)$, and the last equality used (50). Rearranging terms of (57) yields

$$\phi(x_{k+1}) - \phi(x) \le \theta_k^\gamma L\big(h(x) - h(z_0)\big).$$

According to Lemma 4, we have $\theta_k \le \frac{\gamma}{k+\gamma}$ if (49) holds, which gives (54).

If $z_0 \ne \arg\min_{z\in C} h(z)$, then we can only apply (51) for $k \ge 1$ to obtain

$$\begin{aligned}
\frac{1}{\theta_k^\gamma}\phi(x_{k+1}) - \psi_{k+1}(z_{k+1}) - Lh(z_{k+1}) &\le \frac{1-\theta_1}{\theta_1^\gamma}\phi(x_1) - \psi_1(z_1) - Lh(z_1) \\
&= \frac{1}{\theta_0^\gamma}\phi(x_1) - \theta_0^{1-\gamma}\ell(z_1|y_0) - Lh(z_1) \\
&= \phi(z_1) - \ell(z_1|z_0) - Lh(z_1) \\
&\le LD_h(z_1, z_0) - Lh(z_1),
\end{aligned}$$

where the first equality used (49), the second equality used $\theta_0 = 1$, $y_0 = z_0$ and $x_1 = z_1$, and the last inequality is due to relative smoothness: $\phi(z_1) \le \ell(z_1|z_0) + LD_h(z_1, z_0)$. Therefore,

$$\begin{aligned}
\frac{1}{\theta_k^\gamma}\phi(x_{k+1}) &\le \psi_{k+1}(z_{k+1}) + Lh(z_{k+1}) + LD_h(z_1, z_0) - Lh(z_1) \\
&\le \frac{1}{\theta_k^\gamma}\phi(x) + L\big(h(x) - h(z_1) + D_h(z_1, z_0)\big),
\end{aligned}$$

where the last inequality repeats the arguments from (56) to (57). Rearranging terms leads to

$$\phi(x_{k+1}) - \phi(x) \ \le\ \theta_k^\gamma L\big(h(x) - h(z_1) + D_h(z_1, z_0)\big),$$

and further applying Lemma 4 gives the desired result (55). □

As a sanity check, we show that the right-hand-side of (55) is strictly positive for any $x \in \operatorname{dom} h$ such that $\phi(x) < \phi(z_1) + LD_h(x, z_1)$. We exploit the fact that $z_1 = \arg\min_{z\in C}\{\ell(z|z_0) + Lh(z)\}$. Using Lemma 6, we have

$$\ell(z_1|z_0) + Lh(z_1) \le \ell(x|z_0) + Lh(x) - LD_h(x, z_1),$$

which implies

$$L\big(h(x) - h(z_1)\big) \ge LD_h(x, z_1) + \ell(z_1|z_0) - \ell(x|z_0).$$

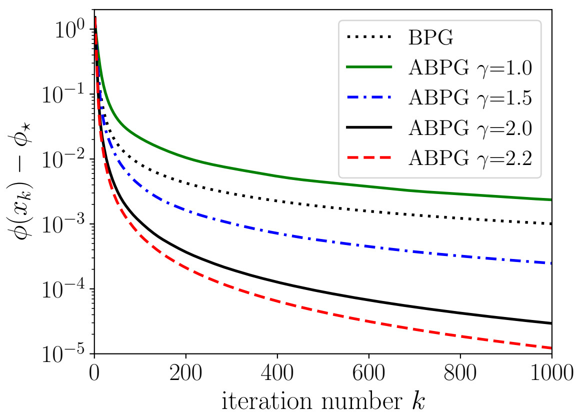
Then we have

$$\begin{aligned}
L\big(h(x) - h(z_1) + D_h(z_1, z_0)\big) &\ge LD_h(x, z_1) + \ell(z_1|z_0) - \ell(x|z_0) + LD_h(z_1, z_0) \\
&= LD_h(x, z_1) + \big(\ell(z_1|z_0) + LD_h(z_1, z_0)\big) - \ell(x|z_0) \\
&\ge LD_h(x, z_1) + \phi(z_1) - \ell(x|z_0) \\
&\ge LD_h(x, z_1) + \phi(z_1) - \phi(x),
\end{aligned}$$

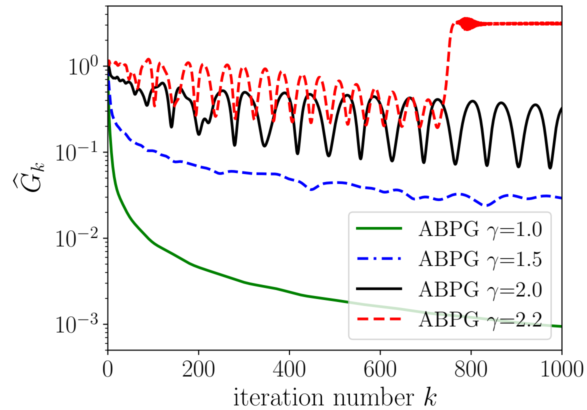
where the second inequality used the upper bound in (21), and the last inequality used the lower bound in (21). Therefore, for any x such that $\phi(x) < \phi(z_1) + LD_h(x, z_1)$, we have

$$L\big(h(x) - h(z_1) + D_h(z_1, z_0)\big) \ > \ LD_h(x, z_1) \ \ge\ 0.$$

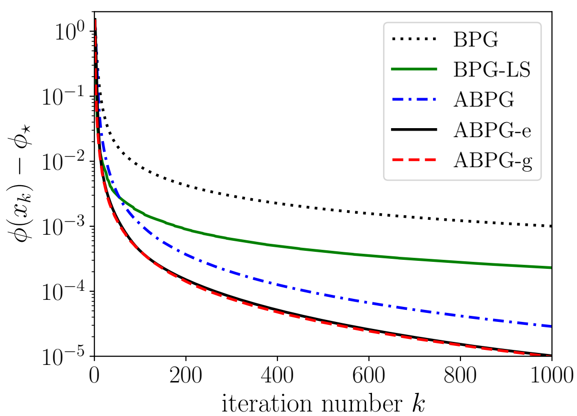
This completes the proof.



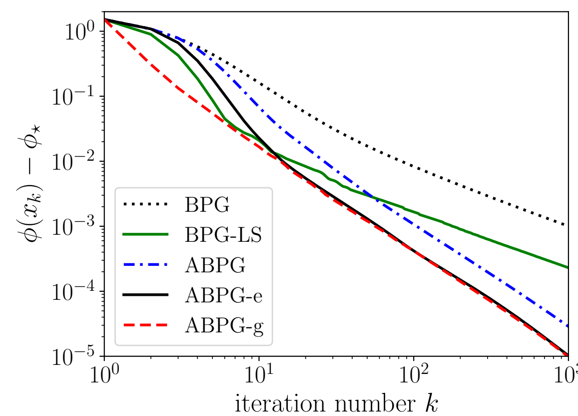
(a) ABPG method with different TSE γ.



(b) Local triangle scaling gain $\widehat{G}_k$.



(c) Adaptive ABPG with $\gamma = 2$ (log-linear plot).



(d) Adaptive ABPG with $\gamma = 2$ (log-log plot).

Figure 2: D-optimal design: random problem instance with $m = 80$ and $n = 200$.

6 Numerical experiments

We consider three applications of relatively smooth convex optimization: D-optimal experiment design, Poisson linear inverse problem, and relative-entropy nonnegative regression. For each application, we compare the algorithms developed in this paper with the BPG method (8) and demonstrate significant performance improvement. Our implementations and experiments are shared through an open-source repository at `https://github.com/Microsoft/accbpg`.

6.1 D-optimal experiment design

Given n vectors $v_1, \ldots, v_n \in \mathbb{R}^m$ where $n \geq m + 1$, the D-optimal design problem is

$$\begin{array}{ll} \text{minimize} & f(x) := -\log\det\left(\sum_{i=1}^n x^{(i)} v_i v_i^T\right) \\ \text{subject to} & \sum_{i=1}^n x^{(i)} = 1 \\ & x^{(i)} \geq 0, \quad i = 1, \ldots, n. \end{array} \tag{58}$$

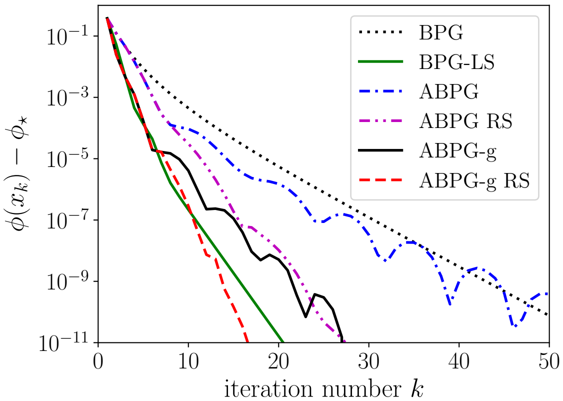


Figure 3: D-optimal design: random problem instance with $m = 80$ and $n = 120$.

In the form of problem (1), we have $\Psi \equiv 0$, $\phi(x) \equiv f(x)$, and C is the standard simplex in $\mathbb{R}^n$. In statistics, this problem corresponds to maximizing the determinant of the Fisher information matrix (e.g., [14, 1]). It is shown in [16] that f defined in (58) is 1-smooth relative to Burg's entropy $h(x) = -\sum_{i=1}^n \log(x^{(i)})$ on $\mathbb{R}^n_+$. In this case, D_h is the IS-distance defined in (12).

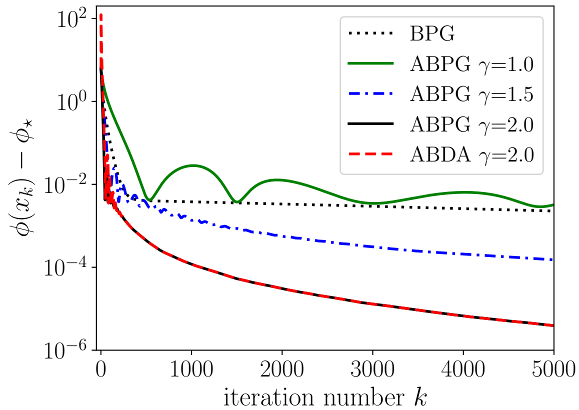
In our first experiment, we set $m = 80$ and $n = 200$ and generated n random vectors in $\mathbb{R}^m$, where the entries of the vectors were generated following independent Gaussian distributions with zero mean and unit variance. The results are shown in Figure 2.

Figure 2(a) shows the reduction of optimality gap by the BPG method (8) and the ABPG method (Algorithm 1) with four different values of γ. For $\gamma = 1$, the ABPG method converges with $O(k^{-1})$ rate, but is slower than the BPG method. When we increase γ to 1.5 and then 2, the ABPG method is significantly faster than BPG. Interestingly, ABPG still converges with $\gamma = 2.2$ (which is larger than the intrinsic TSE $\gamma_{\text{in}} = 2$) and is even faster than with $\gamma = 2$. To better understand this phenomenon, we plot the local triangle-scaling gain

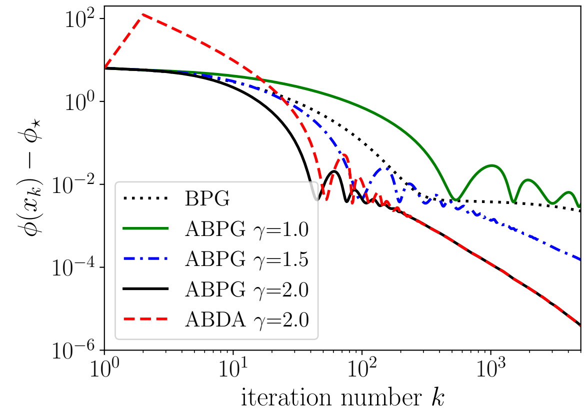
$$\widehat{G}_k = \frac{D_h(x_{k+1}, y_k)}{\theta^\gamma D_h(z_{k+1}, z_k)} = \frac{D_h((1-\theta)x_k + \theta z_{k+1}, (1-\theta)x_k + \theta z_k)}{\theta^\gamma D_h(z_{k+1}, z_k)}. \tag{59}$$

Figure 2(b) shows that for $\gamma = 1.0$ and 1.5, $\widehat{G}_k$ is mostly much smaller than 1. For $\gamma = 2$, $\widehat{G}_k$ is much closer to 1 but always less than 1. This gives a numerical certificate that the ABPG method converged with $O(k^{-2})$ rate. For $\gamma = 2.2$, $\widehat{G}_k$ stayed close to 1 for the first 700 iterations and then jumped to 3 and stayed around. The method diverges with larger value of γ. We didn't plot the ABDA method (Algorithm 5) because it overlaps with ABPG for the same value of γ when the initial point is taken as the center of the simplex, see part (a) of Theorem 6.

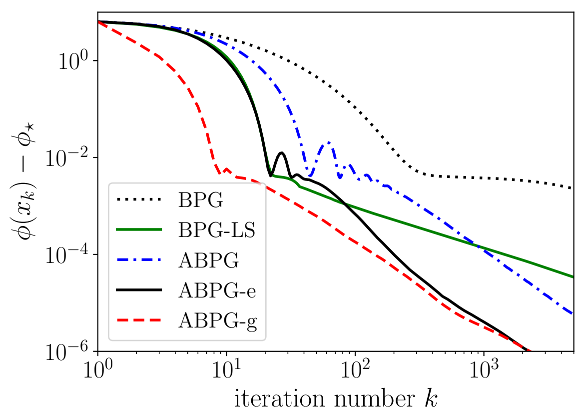
Figure 2(c) compares the basic BPG and ABPG methods with their adaptive variants. The BPG-LS method is a variant of BPG equipped with the same adaptive line-search scheme in Algorithm 4 (see also [21, Method 3.3]). For all variants of ABPG, we set $\gamma = \gamma_{\text{in}} = 2$. For BPG-LS and ABPG-g, we set $\rho = 1.5$ for adjusting the gain G_k. The adaptive variants converged faster than their respective basic versions. Figure 2(d) shows the same results in log-log scale. We can clearly see the different slopes of the BPG variants and ABPG variants, demonstrating their $O(k^{-1})$ and $O(k^{-2})$ convergence rates respectively. For ABPG-e, we started with $\gamma_0 = 3$ and it eventually settled down to $\gamma = 2$, which is reflected in its gradual change of slope in Figure 2(d).



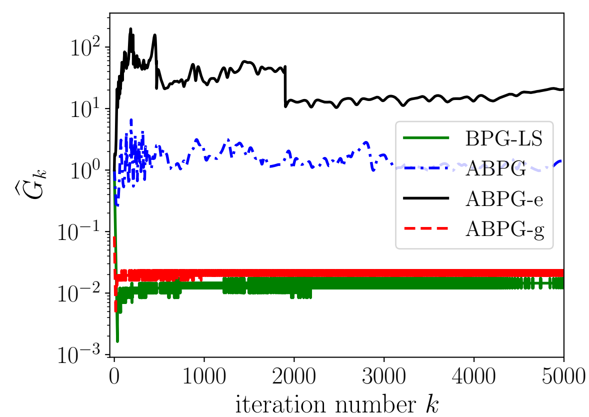
(a) ABPG (varying γ) and ABDA ($\gamma = 2$) methods.



(b) Same results in (a) in log-log plot.



(c) Adaptive ABPG methods $\gamma = 2$ (log-log plot).



(d) Local triangle-scaling gain $\widehat{G}_k$.

Figure 4: Poisson linear inverse problem: random instance with $m = 200$ and $n = 100$.

Figure 3 shows the comparison of different methods on another random problem instance with $m = 80$ and $n = 120$. All methods converge much faster and reach very high precision. In particular, BPG and BPG-LS look to have linear convergence. This indicates that this problem instance is much better conditioned and the objective function may be strongly convex relative to Burg's entropy. In this case, it is shown in [16] that the BPG method attains linear convergence. The ABPG and ABPG-g methods demonstrate periodic non-monotone behavior. A well-known technique to avoid such oscillations and attain fast linear convergence is to restart the algorithm whenever the function value starts to increase [23]. We applied restart (RS) to both ABPG and ABPG-g, which resulted in a much faster convergence as shown in Figure 3.

6.2 Poisson linear inverse problem

In Poisson inverse problems (e.g., [11, 7]), we are given a nonnegative observation matrix $A \in \mathbb{R}_+^{m \times n}$ and a noisy measurement vector $b \in \mathbb{R}_{++}^m$, and the goal is to reconstruct the signal $x \in \mathbb{R}_+^n$ such that $Ax \approx b$. A natural measure of closeness of two nonnegative vectors is the KL-divergence defined

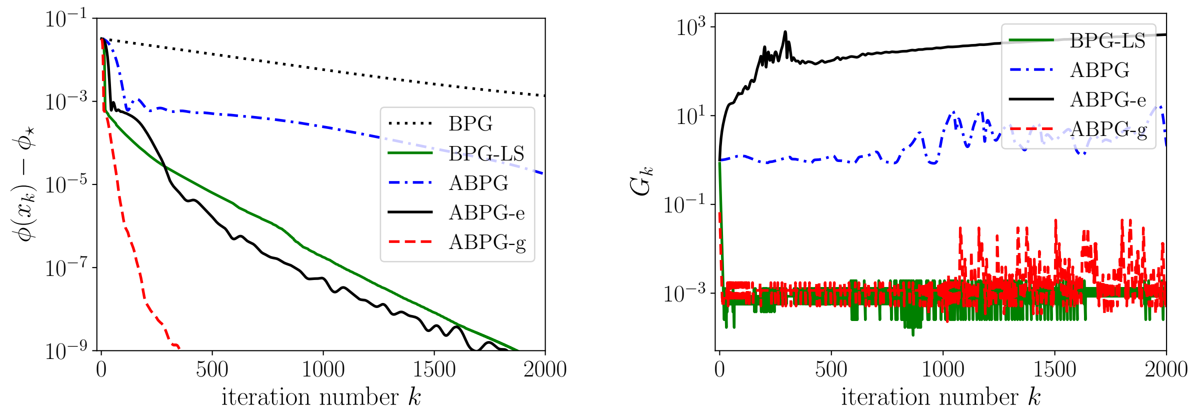


Figure 5: Poisson linear inverse problem: random instance with $m = 100$ and $n = 1000$.

in (11). In particular, minimizing $D_{\mathrm{KL}}(b, Ax)$ corresponds to maximizing the Poisson log-likelihood function. We consider problems of the form

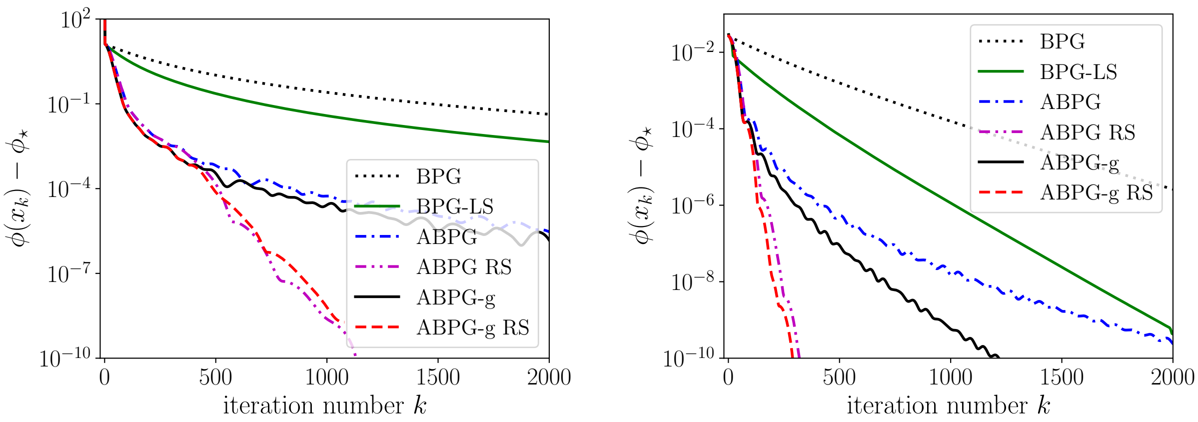
$$\underset{x\in\mathbb{R}^n_+}{\text{minimize}} \;\; \phi(x) := D_{\mathrm{KL}}(b, Ax) + \Psi(x),$$

where $\Psi(x)$ is a simple regularization function. It is shown in [3] that $f(x) = D_{\mathrm{KL}}(b, Ax)$ is L-smooth relative to $h(x) = -\sum_{i=1}^n \log(x^{(i)})$ on $\mathbb{R}^n_+$ for any $L \geq \|b\|_1 = \sum_{i=1}^m b^{(i)}$. Therefore, in the BPG and ABPG methods, we use again the IS-distance D_{IS} in (12) as the proximity measure.

Figure 4 shows our computational results for a randomly generated instance with $m = 200$ and $n = 100$ and $\Psi \equiv 0$ (no regularization). The entries of A and b are generated following independent uniform distribution over the interval $[0, 1]$.

Figure 4(a) shows the reduction of objective gap by BPG and ABPG with $\gamma = 1.0$, 1.5 and 2.0, as well as the ABDA method (Algorithm 5). ABPG and ABDA with $\gamma = 2$ mostly overlap each other in this figure. Figure 4(b) plots the same results in log-log scale, which reveals that ABPG and ABDA (both with $\gamma = 2$) behave quite differently in the beginning. The ABDA method has a jump of objective value at $k = 1$ because $z_0 \neq \arg\min_{z\in C} h(z)$, and its convergence rate is governed by part (b) of Theorem 6. In fact, for $C = \mathbb{R}^n_+$, Burg's entropy $h(x) = -\sum_{i=1}^n \log(x^{(i)})$ is unbounded below as $\|x\| \to \infty$. In contrast, for the D-optimal design problem in Section 6.1, C is the standard simplex, and if we choose $z_0 = x_0 = (1/n, \ldots, 1/n)$ then $z_0 = \arg\min_{z\in C} h(z)$. In that case, we can show that ABPG and ABDA are equivalent when $\Psi \equiv 0$.

Figure 4(c) compares the basic and adaptive variants of BPG and ABPG. For the ABPG and ABPG-g methods, we set $\gamma = \gamma_{\mathrm{in}} = 2$. For ABPG-e, we start with $\gamma_0 = 3$, and the final $\gamma_k = 2.8$ after $k = 5000$ iterations ($\delta = 0.2$ in Algorithm 2). Although ABPG-e uses a much larger γ most of the time, we see ABPG-g converges faster than ABPG-e in the beginning and they eventually become similar. This can be explained through the effective triangle-scaling gains plotted in Figure 4(d). For ABPG and ABPG-e, the effective gains plotted are $\widehat{G}_k$ defined in (59). For BPG-LS and ABPG-g, we plot the G_k's which are adjusted directly in the algorithms. For ABPG-g, $G_k \approx 0.025$ most of the time. The effective $\widehat{G}_k$ for ABPG-e is almost 1000 times larger, which counters the large value of γ used. The sudden reduction of $\widehat{G}_k$ around $k = 2000$ is when γ is reduced from 3 to 2.8. We expect $\gamma_k \to 2$ as k continues to increase.



(a) $m = 1000$ and $n = 100$.

(b) $m = 100$ and $n = 1000$.

Figure 6: Two random instances of relative entropy nonnegative regression ($\gamma = 2$ for ABPG).

Figure 5 shows the results for a randomly generated instance with $m = 100$ and $n = 1000$. In this case, since $m < n$, we added a regularization $\Psi(x) = (\lambda/2)\|x\|^2$ with $\lambda = 0.001$. ABPG-g has the best performance. Again we observe that $G_k \ll 1$ most of the time, which gives a numerical certificate that the ABPG methods do converge with $O(k^{-2})$ rate.

6.3 Relative-entropy nonnegative regression

An alternative approach for solving the nonnegative linear inverse problem described in Section 6.2 is to minimize $D_{\mathrm{KL}}(Ax, b)$, i.e.,

$$\underset{x \in \mathbb{R}^n_+}{\text{minimize}} \;\; \phi(x) := D_{\mathrm{KL}}(Ax, b) + \Psi(x).$$

In this case, it is shown in [3] that $f(x) = D_{\mathrm{KL}}(Ax, b)$ is L-smooth relative to the Boltzmann-Shannon entropy $h(x) = \sum_{i=1}^n x^{(i)} \log(x^{(i)})$ on $\mathbb{R}^n_+$ for any L such that

$$L \;\geq\; \max_{1 \leq j \leq n} \sum_{i=1}^m A_{ij} \;=\; \max_{1 \leq j \leq n} \|A_{:j}\|_1$$

where $A_{:j}$ denotes the jth column of A. Therefore, in the BPG and ABPG methods, we use the KL-divergence D_{KL} defined in (11) as the proximity measure. In our experiment, we apply ℓ_1-regularization $\Psi(x) = \lambda\|x\|_1$ with $\lambda = 0.001$.

Figure 6(a) shows the results for a randomly generated instance with $m = 1000$ and $n = 100$. For all variants of the ABPG method, we set $\gamma = \gamma_{\mathrm{in}} = 2$. Since the accelerated methods demonstrate oscillations in objective value, we tried the restart (RS) trick [23] and obtained faster convergence with apparent linear rate. Figure 6(b) shows the results for a random instance with $m = 100$ and $n = 1000$. In this case, we clearly see linear convergence of the BPG and BPG-LS methods. Again, ABPG methods with restart achieve the fastest convergence. For the ABPG-g method, we always obtain small gains G_k at each step. Therefore their geometric mean $\overline{G}_k$ is also small, which serves as a certificate of the $O(k^{-2})$ convergence rate for this problem instance.

Acknowledgments

We thank Haihao Lu, Robert Freund and Yurii Nesterov for helpful conversations.

References

[1] C. L. Atwood. Optimal and efficient designs of experiments. *The Annals of Mathematical Statistics*, 40(5):1570–1602, 1969.

[2] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16(3):697–725, 2006.

[3] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent Lemma beyond Lipschitz gradient continuity: first-order method revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[4] H. H. Bauschke and J. M. Borwein. Joint and separate convexity of the Bregman distance. In D. Butnariu, Y. Censor, and S. Reich, editors, *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications (Haifa 2000)*, pages 23–26. Elsevier, 2001.

[5] A. Beck. *First-Order Methods in Optimization.* MOS-SIAM Series on Optimization. SIAM, 2017.

[6] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[7] M. Bertero, P. Boccacci, G. Desiderà, and G. Vicidomini. Image deblurring with Poisson data: from cells to galaxies. *Inverse Problems*, 25(12), 2009.

[8] B. Birnbaum, N. R. Devanur, and L. Xiao. Distributed algorithms via gradient descent for Fisher markets. In *Proceedings of the 12th ACM conference on Electronic Commerce*, pages 127–136, San Jose, California, USA, June 2011.

[9] L. M. Bregman. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Comput. Math. and Math. Phys.*, 7:200–217, 1967.

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

[11] I. Csiszár. Why least squares and maximum entropy? an axiomatic approach to inference for linear iverse problems. *The Annals of Statistics*, 19(4):2032–2066, 1991.

[12] F. Hanzely and P. Richtárik. Fastest rates for stochastic mirror descent methods. arXiv preprint arXiv:1803.07374, 2018.

[13] G. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities.* Cambridge University Press, 2nd edition, 1952.

[14] J. Kiefer and J. Wolfowitz. Optimal design in regression problems. *The Annals of Mathematical Statistics*, 30(2):271–294, 1959.

[15] H. Lu. "Relative-continuity" for non-Lipschitz non-smooth convex optimization using stochastic (or deterministic) mirror descent. arXiv preprint arXiv:1710.04718, 2017.

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

[17] A. Nemirovsky and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. J. Wiley & Sons, New York, 1983.

[18] Y. Nesterov. On an approach to the construction of optimal methods of minimization of smooth convex functions. *Èkonom. i. Mat. Metody*, 24:509–517, 1988.

[19] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, Boston, 2004.

[20] Y. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.

[21] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming, Ser. B*, 140:125–161, 2013.

[22] Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming, Ser. A*, 152:381–404, 2015.

[23] B. O'Donoghue and E. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015.

[24] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[25] M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming, Ser. B*, 170:67–96, 2018.

[26] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Unpublished manuscript, 2008.

[27] Y. Zhou, Y. Liang, and L. Shen. A unified approach to proximal algorithms using Bregman distance. Technical Report, 2016.