An inexact Bregman proximal gradient method and its inertial variant

Lei Yang∗, Kim-Chuan Toh†

May 18, 2022

Abstract

The Bregman proximal gradient (BPG) method, which uses the Bregman distance as a proximity measure in the iterative scheme, has recently been re-developed for minimizing convex composite problems without the global Lipschitz gradient continuity assumption. This makes the BPG appealing for a wide range of applications, and hence it has received growing attention in recent years. However, most existing convergence results are only obtained under the assumption that the involved subproblems are solved exactly, which is not realistic in many applications. For the BPG to be implementable and practical, in this paper, we develop an inexact version of the BPG (denoted by iBPG) by employing a novel two-point-type inexact stopping condition for solving the subproblems. The iteration complexity of \(O(1/k)\) and the convergence of the sequence are also established for our iBPG under some proper conditions. Moreover, we develop an inertial variant of our iBPG (denoted by v-iBPG) and establish the iteration complexity of \(O(1/k^\gamma)\), where \(\gamma \geq 1\) is a restricted relative smoothness exponent. When the smooth part in the objective has a Lipschitz continuous gradient and the kernel function is strongly convex, we have \(\gamma = 2\) and thus the v-iBPG improves the iteration complexity of the iBPG from \(O(1/k)\) to \(O(1/k^2)\), in accordance with the existing results on the accelerated BPG. Finally, some preliminary numerical experiments for solving a relaxation of the quadratic assignment problem are conducted to show the convergence behaviors of our iBPG and v-iBPG under different inexactness settings.

Keywords: Proximal gradient method; Bregman distance; relative smoothness; inexact condition; Nesterov’s acceleration.

1 Introduction

In this paper, we consider the following convex composite problem:

\[
\inf_{x \in \mathbb{Q}} F(x) := P(x) + f(x), \tag{1.1}
\]

where \(\mathbb{Q} \subseteq \mathbb{E}\) is a closed convex set with nonempty interior denoted by \(\text{int} \mathbb{Q}\), and \(\mathbb{E}\) is a real finite dimensional Euclidean space equipped with an inner product \(\langle \cdot, \cdot \rangle\) and its induced norm \(\| \cdot \|\).

∗Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China (yanglei.math@gmail.com).
†Department of Mathematics, and Institute of Operations Research and Analytics, National University of Singapore, Singapore 119076 (mattohkc@nus.edu.sg). The research of this author is supported by the Ministry of Education, Singapore, under Academic Research Fund Tier 1 Grant A00084930000.
The functions $f, P : \mathbb{E} \to (-\infty, \infty]$ are proper closed convex functions and $f$ is differentiable on $\text{int} \ Q$. Problem (1.1) is a generic optimization problem that arises in many application areas such as machine learning, data science, image/signal processing, to mention just a few, and has been extensively studied in the literature.

A classical method for solving problem (1.1) is the proximal gradient (PG) method [22, 30] (also known as the forward-backward splitting method), whose basic iterative step is given by

$$x^{k+1} = \arg\min_{x \in Q} \left\{ P(x) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2} \|x - x^k\|^2 \right\},$$

where $L > 0$ is a constant. This scheme is developed based on the construction of a quadratic upper approximation of the smooth part $f$. Therefore, a central assumption required in the development and analysis of the PG is that $\nabla f$ is Lipschitz continuous on $Q$. Moreover, one can consider the Bregman distance $D_\phi$ associated with a kernel function $\phi$ as a proximity measure (see next section for definition), and naturally generalize the PG to the Bregman proximal gradient (BPG) method (see, for example, [4, 6, 31, 45, 46]), whose basic iterative step reads as follows:

$$x^{k+1} = \arg\min_{x \in Q} \left\{ P(x) + \langle \nabla f(x^k), x - x^k \rangle + LD_\phi(x, x^k) \right\}. \quad (1.2)$$

Comparing to the PG, the BPG with a suitable choice of $D_\phi$ is better able to exploit the underlying geometry/structure of the problem and possibly obtains a more tractable subproblem (1.2) (see, for example, [10, Section 5], [36, Lemma 4]). More importantly, as investigated in recent insightful works [6, 31], the BPG can be developed based on the notion of Lipschitz-like/convexity condition or relative smoothness, which is weaker than the longstanding global Lipschitz gradient continuity assumption. This makes the BPG applicable for a wider range of problems, such as the Poisson inverse problem [6] and the D-optimal design [31], where $f$ involves the logarithm function in the form of relative entropy and log-determinant, respectively. Because of such an appealing potential, the BPG and its variants have gained increasing attention in recent years; see, for example, [5, 6, 12, 21, 23, 31, 44, 47, 50].

Unfortunately, although the underlying geometry/structure can possibly be captured by the kernel function, the subproblem in form of (1.2) still has no closed-form solution in many applications and its computation may be demanding numerically; see examples in [18, 26, 41] and the one in our numerical section. Thus, for the BPG to be implementable and practical, it must allow one to solve the subproblem approximately with progressive accuracy and the corresponding stopping condition should be practically verifiable. However, unlike the inexact PG method which has been widely studied under different inexact criteria (see, for example, [20, 42, 48]), the study on the inexact BPG method is still at an early stage. Kabbadj [28] recently developed an inexact BPG method by allowing an $\varepsilon$-minimizer for each subproblem (1.2). Since the optimal objective value of subproblem (1.2) is generally unknown, such an inexact condition can be difficult, if not impossible, to verify in practice. In [43], Stonyakin et al. proposed a general inexact algorithmic framework for the first-order methods containing the BPG as a special case, where an inexact solution of the subproblem (1.2) is accepted when a certain inexact condition [43, Definition 2.12] holds. In our context, their inexact condition can be covered by our inexact condition (3.2) with $\nu_k \equiv \mu_k \equiv 0$. But we note that their condition could be difficult to verify practically; see an example in Remark 3.1.

Over the last few decades, Nesterov’s series of seminal works [33, 34, 36] (see also [35]) on accelerated gradient methods have inspired numerous accelerated variants of the PG (see, e.g.,
as well as their inexact counterparts (see, e.g., [3, 14, 27, 42, 48]). Naturally, those developments also motivated many researchers to explore whether and how the BPG can be accelerated. A successful attempt was made by Auslender and Teboulle [4, Section 5], who proposed an improved interior gradient algorithm (a Bregman gradient scheme) for solving a special case of problem (1.1) with $P \equiv 0$ and showed that this algorithm can achieve a faster rate of $O(1/k^2)$. Later on, it was extended by Tseng [45, 46] to handle problem (1.1) in a more general composite setting. Other accelerated variants of the BPG can also be found in [29, 45, 46]. Note that all these works [4, 29, 45, 46] just mentioned require the assumptions that $f$ has a globally Lipschitz gradient and that the kernel function of the Bregman distance is strongly convex. Until recently, based on the relative smoothness assumption and a so-called triangle scaling property of the Bregman distance, Hanzely, Richtárik and Xiao [24] developed an accelerated BPG method that attains a rate of $O(1/k^\gamma)$, where $\gamma \in (0, 2]$ is the triangle scaling exponent. Later, a similar result was also derived under a slightly broader smoothness condition by Gutman and Peña [23, Section 3.3]. Very recently, Dragomir et al. [21] further showed that the rate of $O(1/k)$ is indeed optimal for the BPG to solve the class of problems merely satisfying the relative smoothness assumption, and thus the BPG cannot be accelerated if no other regularity condition is imposed. But empirically improving the BPG via an inertial step and a certain adaptive backtracking strategy is still possible, as discussed in [24, 32].

However, among the aforementioned works that have investigated the possibility of improving the BPG, none of them take into account of the possible errors incurred when one solves the involved subproblem inexactly. Indeed, as far as we know, the rigorous study on the inexact inertial BPG method is very limited. Stonyakin et al. [43] have developed an inexact inertial variant of the BPG method in their work. But as we have mentioned in last paragraph, the inexact condition used in [43] can be covered by our inexact condition and may still require improvements in terms of its practical verification. Moreover, we also note that Hien et al. [25] recently proposed a block alternating Bregman majorization-minimization framework with extrapolation, which gives an inertial BPG if only one block of variables is considered, and employed a surrogate function for $P$ to obtain a more tractable auxiliary subproblem which possibly admits a closed-form solution. But such a surrogate function could be nontrivial to find when $P$ is complex, and the auxiliary subproblem should be solved exactly. Thus, the kind of inexactness considered in [25] is very different from what we consider in this paper. Moreover, replacing $P$ by its surrogate could also potentially slow down the BPG method, especially when the surrogate is not a good approximation of $P$.

In this paper, to facilitate the practical implementations of the BPG and its inertial variant, we attempt to develop their inexact counterparts. We provide theoretical insights on how the error incurred in the inexact computation of the subproblem would affect the convergence rate in terms of the objective function value. The inexact condition to be used has a two-point feature and it is motivated by our recent works [19, 49] on a new inexact Bregman proximal point algorithm. The resulting inexact framework is rather broad and can handle different types of errors that may occur when solving the subproblem, and it is also helpful to circumvent the underlying feasibility difficulty in evaluating the Bregman distance when the problem has a complex feasible set; see Remark 3.1 for more details.

The contributions of this paper are summarized as follows.

1. We develop an inexact Bregman proximal gradient (iBPG) method based on a practically verifiable stopping criterion that is distinct from the existing ones in [28, 43]. The iteration
complexity of $O(1/k)$ and the convergence of the sequence are also established for our iBPG under some proper conditions.

2. We develop an inertial variant of our iBPG (denoted by v-iBPG) and establish the iteration complexity of $O(1/k^\gamma)$ ($\gamma \geq 1$) under an additional restricted relative smoothness assumption (see Assumption C). Our result can recover the related results in [4, 24, 45, 46] wherein the subproblem is solved exactly, and as a byproduct, our analysis provides a unified treatment of these existing results which are developed under different conditions. In particular, when the smooth part in the objective has a Lipschitz continuous gradient and the kernel function is strongly convex, we have $\gamma = 2$ and thus the v-iBPG improves the iteration complexity of the iBPG from $O(1/k)$ to $O(1/k^2)$.

3. We conduct some preliminary numerical experiments to evaluate the performances of our iBPG and v-iBPG under different inexactness settings. The computational results empirically verify the necessity of developing inexact versions of those methods and also illustrate that different methods do have different inherent inexactness tolerance requirements, in accordance with our theoretical results.

The rest of this paper is organized as follows. We present notation and preliminaries in Section 2. We then describe our iBPG for solving problem (1.1) and establish the convergence results in Section 3. A possibly faster inertial variant of our iBPG is developed and analyzed in Section 4. Some preliminary numerical results are reported in Section 5, with some concluding remarks given in Section 6.

2 Notation and preliminaries

Assume that $f : E \to (-\infty, \infty]$ is a proper closed convex function. For a given $\nu \geq 0$, the $\nu$-subdifferential of $f$ at $x \in \text{dom} f$ is defined by $\partial_{\nu} f(x) := \{ d \in E : f(y) \geq f(x) + \langle d, y - x \rangle - \nu, \forall y \in E \}$, and when $\nu = 0$, $\partial_{\nu} f$ is simply denoted by $\partial f$, which is referred to as the subdifferential of $f$. The conjugate function of $f$ is the function $f^* : E \to (-\infty, \infty]$ defined by $f^*(y) := \sup \{ \langle y, x \rangle - f(x) : x \in E \}$. A proper closed convex function $f$ is essentially smooth if (i) int dom $f$ is not empty; (ii) $f$ is differentiable on int dom $f$; (iii) $\|\nabla f(x_k)\| \to \infty$ for every sequence $\{x_k\}$ in int dom $f$ converging to a boundary point of int dom $f$; see [40, page 251].

For a vector $x \in \mathbb{R}^n$, $x_i$ denotes its $i$-th entry, $\text{Diag}(x)$ denotes the diagonal matrix whose $i$th diagonal entry is $x_i$, $\|x\|$ denotes its $\ell_2$ (Euclidean) norm. For a matrix $A \in \mathbb{R}^{m \times n}$, $a_{ij}$ denotes its $(i, j)$th entry, $A_{ij}$ denotes its $j$th column, $\|A\|_F$ denotes its Frobenius norm. For a closed convex set $\mathcal{X} \subseteq \mathbb{R}^n$, its indicator function $\delta_{\mathcal{X}}$ is defined by $\delta_{\mathcal{X}}(x) = 0$ if $x \in \mathcal{X}$ and $\delta_{\mathcal{X}}(x) = +\infty$ otherwise.

Given a proper closed strictly convex function $\phi : E \to (-\infty, \infty]$, finite at $x, y$ and differentiable at $y$ but not necessarily at $x$, the Bregman distance [15] between $x$ and $y$ associated with the kernel function $\phi$ is defined as

$$D_{\phi}(x, y) := \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle.$$  

It is easy to see that $D_{\phi}(x, y) \geq 0$ and equality holds if and only if $x = y$ due to the strict convexity of $\phi$. When $E := \mathbb{R}^n$ and $\phi(\cdot) := \| \cdot \|^2$, $D_{\phi}(\cdot, \cdot)$ readily recovers the classical squared Euclidean distance. Based on the Bregman distance, we then define the restricted relative smoothness as follows.
**Definition 2.1 (Restricted relative smoothness on \( \mathcal{X} \)).** Let \( f, \phi : \mathbb{E} \to (-\infty, \infty] \) be proper closed convex functions with \( \text{dom } f \supseteq \text{dom } \phi \) and \( f, \phi \) be differentiable on \( \text{int } \text{dom } \phi \). Given a closed convex set \( \mathcal{X} \subseteq \mathbb{E} \) with \( \mathcal{X} \cap \text{int } \text{dom } \phi \neq \emptyset \), we say that \( f \) is \( L \)-smooth relative to \( \phi \) restricted on \( \mathcal{X} \) if there exists \( L \geq 0 \) such that

\[
 f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + LD_{\phi}(y, x), \quad \forall \ x \in \text{int dom } \phi \cap \mathcal{X}, \ y \in \text{dom } \phi \cap \mathcal{X}. \tag{2.1}
\]

The above restricted relative smoothness modifies the original relative smoothness (or the Lipschitz-like/convexity condition) introduced in [6, 31] by imposing a restricted set \( \mathcal{X} \), and it readily reduces to the original notion when \( \mathcal{X} = \mathbb{E} \). Such a restriction would help to extend the notion of the relative smoothness to more choices of \((f, \phi)\) with a proper \( \mathcal{X} \). For example, when \( \nabla f \) is \( L_f \)-Lipschitz continuous on \( \mathbb{R}^n \) and \( \phi \) is \( \mu_{\phi} \)-strongly convex on \( \mathcal{X} \subseteq \mathbb{R}^n \), then it can be verified that \( f \) is \( \frac{L_f}{\mu_{\phi}} \)-smooth relative to \( \phi \) restricted on \( \mathcal{X} \), but \( f \) may not be \( L \)-smooth relative to \( \phi \) according to the original definition in [6, 31] because \( \phi \) may not be strongly convex on its domain. For example, the entropy function \( \phi(x) = \sum_{i=1}^n x_i (\log x_i - 1) \) is \( \frac{1}{\alpha} \)-smooth convex on \([0, \alpha]^n\) with any \( \alpha > 0 \), but it is not strongly convex on \( \text{dom } \phi = \mathbb{R}_+^n \). Therefore, employing the notion of restricted relative smoothness in Definition 2.1 could broaden the possible applications of the Bregman proximal gradient method and its inertial variants. More discussions on the relative smoothness can be found in [6, 31].

In order to establish the rigorous analysis under the relatively smooth setting, we make the following blanket technical assumptions.

**Assumption A.** Problem (1.1) and the kernel function \( \phi \) satisfy the following assumptions.

A1. \( \phi : \mathbb{E} \to (-\infty, \infty] \) is essentially smooth and strictly convex on \( \text{int } \text{dom } \phi \). Moreover, \( \text{dom } \phi = Q \), where \( \text{dom } \phi \) denotes the closure of \( \text{dom } \phi \).

A2. \( P : \mathbb{E} \to (-\infty, \infty] \) is a proper closed convex function with \( \text{dom } P \cap \text{int } Q \neq \emptyset \).

A3. \( f : \mathbb{E} \to (-\infty, \infty] \) is a proper closed convex function with \( \text{dom } f \supseteq \text{dom } \phi \) and \( f \) is differentiable on \( \text{int } \text{dom } \phi \). Moreover, there exists a closed convex set \( \mathcal{X} \supseteq \text{dom } P \cap \text{dom } \phi \) such that \( f \) is \( L \)-smooth relative to \( \phi \) restricted on \( \mathcal{X} \).

A4. \( F^* := \inf \{ F(x) : x \in Q \} > -\infty \), i.e., problem (1.1) is bounded from below.

A5. Each subproblem in the iBPG and its inertial variant is well-defined in the sense that the subproblem has a unique minimizer, which lies in \( \text{int } \text{dom } \phi \).

Note that, in Assumption A1, \( \text{dom } \phi = Q \) implies \( \text{int } \text{dom } \phi = \text{int } Q \) due to the convexity of \( \text{dom } \phi \) and [9, Proposition 3.36(iii)]. Then, one can see from Assumptions A2&3 that

\[
 \text{dom } (P + f) \cap \text{int dom } \phi = \text{dom } P \cap \text{int dom } \phi = \text{dom } P \cap \text{int } Q \neq \emptyset.
\]

This, together with \( \text{dom } \phi = Q \) and [9, Proposition 11.1(iv)] implies that

\[
 F^* = \inf \{ F(x) : x \in Q \} = \inf \{ F(x) : x \in \text{dom } \phi \}. \tag{2.2}
\]

Assumption A5 is standard and is commonly made for ensuring the well-posedness of Bregman-type methods. It can be satisfied when, for example, \( \phi \) is strongly convex. Other sufficient conditions are given in [6, Lemma 2].

Finally, we give four supporting lemmas that will be used in the subsequent analysis. The identity in first lemma is routine to verify and the proofs of last two lemmas are relegated to Appendix A.
Lemma 2.1 (Four points identity). Suppose that a proper closed strictly convex function $\phi : \mathbb{E} \to (-\infty, \infty]$ is finite at $a, b, c, d$ and differentiable at $a, b$. Then,

$$\langle \nabla \phi(a) - \nabla \phi(b), c - d \rangle = D_\phi(c, b) + D_\phi(d, a) - D_\phi(c, a) - D_\phi(d, b). \quad (2.3)$$

Lemma 2.2 ([39, Section 2.2]). Suppose that $\{\alpha_k\}_{k=0}^{\infty} \subseteq \mathbb{R}$ and $\{\beta_k\}_{k=0}^{\infty} \subseteq \mathbb{R}$ are two sequences such that $\{\alpha_k\}$ is bounded from below, $\sum_{k=0}^{\infty} \beta_k < \infty$, and $\alpha_{k+1} \leq \alpha_k + \beta_k$ holds for all $k$. Then, $\{\alpha_k\}$ is convergent.

Lemma 2.3. Let $\{\alpha_k\}_{k=0}^{\infty}$ be a nonnegative sequence. If $\sum_{k=0}^{\infty} \alpha_k < \infty$, then $\frac{1}{k} \sum_{i=0}^{k-1} i\alpha_i \to 0$.

Lemma 2.4. Suppose that Assumption A holds and $\mathcal{X}$ is the closed convex set in Assumption A3. Given any $y \in \text{int dom } \phi \cap \mathcal{X}$, $\lambda \geq 0$, $\eta \geq 0$, $\mu \geq 0$ and $\nu \geq 0$, let $(x^*, \tilde{x}^*)$ be a pair of approximate solutions of

$$\min_{x} P(x) + \langle \nabla f(y), x - y \rangle + \lambda D_\phi(x, y)$$

such that $x^* \in \mathcal{X} \cap \text{int dom } \phi$, $\tilde{x}^* \in \text{dom } P \cap \text{dom } \phi$ and

$$\Delta \in \partial \nu P(\tilde{x}^*) + \nabla f(y) + \lambda (\nabla \phi(x^*) - \nabla \phi(y)) \quad \text{with} \quad \|\Delta\| \leq \eta, \quad D_\phi(\tilde{x}^*, x^*) \leq \mu. \quad (2.4)$$

Then, for any $x \in \text{dom } P \cap \text{dom } \phi$, we have

$$F(\tilde{x}^*) - F(x) \leq \lambda D_\phi(x, y) - \lambda D_\phi(x, x^*) - (\lambda - L)D_\phi(x^*, y) + \eta \|\tilde{x}^* - x\| + \lambda \mu + \nu.$$

3 An inexact Bregman proximal gradient method

In this section, we develop an inexact Bregman proximal gradient (iBPG) method for solving problem (1.1). The complete framework is presented as Algorithm 1.

**Algorithm 1** An inexact Bregman proximal gradient (iBPG) method for solving problem (1.1)

**Input:** Let $\{\eta_k\}_{k=0}^{\infty}$, $\{\mu_k\}_{k=0}^{\infty}$ and $\{\nu_k\}_{k=0}^{\infty}$ be three sequences of nonnegative scalars, and $\mathcal{X} \supseteq \text{dom } P \cap \text{dom } \phi$ be the closed convex set in Assumption A3. Choose $x^0 \in \text{int dom } \phi \cap \mathcal{X}$ and $\tilde{x}^0 \in \text{dom } P \cap \text{dom } \phi$ arbitrarily. Set $k = 0$.

**while** the termination criterion is not met, **do**

**Step 1.** Find a pair $(x^{k+1}, \tilde{x}^{k+1})$ by approximately solving the following problem

$$\min_{x} P(x) + \langle \nabla f(x^k), x - x^k \rangle + L D_\phi(x, x^k), \quad (3.1)$$

such that $x^{k+1} \in \mathcal{X} \cap \text{int dom } \phi$, $\tilde{x}^{k+1} \in \text{dom } P \cap \text{dom } \phi$ and

$$\Delta^k \in \partial \nu_k P(\tilde{x}^{k+1}) + \nabla f(x^k) + L (\nabla \phi(x^{k+1}) - \nabla \phi(x^k)) \quad \text{with} \quad \|\Delta^k\| \leq \eta_k, \quad D_\phi(\tilde{x}^{k+1}, x^{k+1}) \leq \mu_k. \quad (3.2)$$

**Step 2.** Set $k = k + 1$ and go to **Step 1**.

**end while**

**Output:** $(x^k, \tilde{x}^k)$
One can see that, at each iteration, our iBPG in Algorithm 1 allows us to approximately solve the subproblem (3.1) under the inexact condition (3.2). Since the subproblem (3.1) has a unique solution $x^{k\ast} \in \text{dom} P \cap \text{int} \text{dom} \phi$ (by Assumption A5), then condition (3.2) always holds at $x^{k+1} = \tilde{x}^{k+1} = x^{k\ast}$ and hence it is achievable. When setting $\nu_k \equiv \eta_k \equiv \mu_k \equiv 0$, it means that $x^{k+1}$ (equals to $\tilde{x}^{k+1}$) should be an optimal solution of the subproblem, and thus, our iBPG method readily reduces to the exact BPG method studied in [6, 31, 46, 50]. Moreover, when $\nu_k \equiv \mu_k \equiv 0$, condition (3.2) is equivalent to the inexact condition considered in [43, Definition 2.12] for the subproblem (3.1), and hence our iBPG contains the inexact BPG method studied in [43] as a special case.

The inexact condition (3.2) is of two-point type and it is inspired by the condition proposed in the recent works [19, 49] for developing a new inexact Bregman proximal point algorithm. It may look unusual at the first glance, but it actually provides a rather broad inexact framework in which the approximate solutions are handled through $\partial \nu_k P$ (an approximation of $\partial P$), the error term $\Delta^k$ appearing on the left-hand-side of the optimality condition, and the deviation $D_{\phi}(\tilde{x}^{k+1}, x^{k+1})$. Thus, our inexact framework is amenable to different type of errors incurred when one solves the subproblem inexactly. Moreover, the admissible deviation allows $\partial \nu_k P$ and $\nabla \phi$ to be computed at two slightly different points, respectively. Such a simple strategy would help to circumvent the possible feasibility difficulty of requiring $x^{k+1} \in \text{dom} P \cap \text{dom} \nabla \phi$ by the exact BPG method studied in [6, 31, 46, 50] and by the inexact BPG method studied in [43], as exemplified in Remark 3.1. In addition, we have noticed a recent work by Kabbadj [28], who proposed an inexact BPG method by allowing an $\varepsilon$-minimizer for the subproblem (3.1). However, such an inexact condition usually cannot be directly used in practical implementations since the optimal objective value of the subproblem is generally unknown a priori.

**Remark 3.1 (Comments on the underlying feasibility difficulty).** To illustrate the underlying feasibility difficulty when solving the subproblem in the BPG method, we give the following example, which is also a test problem in our numerical experiments. Given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, let $P$ be an indicator function $\delta_C$ with $C := \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$, and consider the entropy function $\phi(x) = \sum_{i=1}^n x_i(\log x_i - 1)$ as the kernel function with $\text{dom} \phi = \mathbb{R}_+^n$. Then, the subproblem (3.1) can be expressed as

$$
\min_x \langle \nabla f(x^k), x - x^k \rangle + LD_{\phi}(x, x^k) \quad \text{s.t.} \quad Ax = b, \ x \geq 0.
$$

Thus, when employing the exact BPG method in [6, 31, 46, 50] or the inexact BPG method in [43], one needs to find the exact solution or an inexact solution $\tilde{x}^{k+1}$ that should be strictly contained in $\text{dom} P \cap \text{dom} \nabla \phi = \{x \in \mathbb{R}^n : Ax = b, x > 0\} =: \text{relin} C$ (namely, the relative interior of $C$). However, for most choices of $A$, a subroutine for solving (3.3) may only return a candidate approximate solution $x^{k+1}$ that is located outside $C$. Hence, one has to further perform a proper projection/rounding step at $x^{k+1}$ to compute an approximate solution $\tilde{x}^{k+1}$ in $\text{relin} C$ and then check the inexact condition at $\tilde{x}^{k+1}$. Due to the non-closedness of $\text{relin} C$, this step is indeed difficult to implement especially when some entries of the exact solution are very close to zero. In contrast, our two-point inexact condition requires separately $x^{k+1} \in \text{int} \text{dom} \phi \cap \mathcal{X}$ and $\tilde{x}^{k+1} \in \text{dom} P \cap \text{dom} \phi = C$, where $\mathcal{X} \supseteq \text{dom} P \cap \text{dom} \phi$ can be any closed convex set satisfying Assumption A3. The point $x^{k+1}$ (which is generally located outside $C$) can be easily obtained from the subroutine and the point $\tilde{x}^{k+1}$ can be computed by performing a proper projection/rounding step at $x^{k+1}$ over $C$, which is clearly easier and more practical than computing a point in $\text{relin} C$. In this regard, our iBPG is more favorable than the exact BPG method in [6, 31, 46, 50] and
Theorem 3.1. Suppose that Assumption A holds. Let \( y \) be the inexact BPG in Algorithm 1. Then, the following statements hold.

Proof. The desired result can be easily obtained by applying Lemma 2.4 with \( k = 1 \).

We next establish the convergence of our iBPG. The analysis is inspired by \([6, 31]\), but is more involved due to the two-point inexact condition (3.2). We first give the following sufficient-descent-like property.

Lemma 3.1. Suppose that Assumption A holds. Let \( \{x^k\} \) and \( \{\tilde{x}^k\} \) be the sequences generated by the iBPG in Algorithm 1. Then, for any \( k \geq 0 \) and any \( x \in \text{dom } P \cap \text{dom } \phi \),

\[
F(\tilde{x}^{k+1}) - F(x) \leq LD_\phi(x, x^k) - LD_\phi(x, x^{k-1}) + \eta_k \|\tilde{x}^{k+1} - x\| + L \mu_k + \nu_k.
\]

Proof. The desired result can be easily obtained by applying Lemma 2.4 with \( \tilde{x}^* = \tilde{x}^{k+1} \), \( x^* = x^{k+1}, y = x^k \), \( \lambda = L \), \( \eta = \eta_k \), \( \mu = \mu_k \) and \( \nu = \nu_k \). \( \Box \)

We then have the following results concerning the convergence of the function value.

Theorem 3.1. Suppose that Assumption A holds. Let \( \{x^k\} \) and \( \{\tilde{x}^k\} \) be the sequences generated by the iBPG in Algorithm 1. Then, the following statements hold.

(i) For any \( k \geq 0 \) and any \( x \in \text{dom } P \cap \text{dom } \phi \), we have

\[
F(y) - F(x) \leq \frac{1}{k+1} \left( LD_\phi(x, x^0) + \sum_{i=0}^{k} (\eta_i \|\tilde{x}^{i+1}\| + \eta_i \|x\| + L \mu_i + \nu_i) \right),
\]

and

\[
F(\tilde{x}^{k+1}) - F(x) \leq \frac{1}{k+1} \left( LD_\phi(x, x^0) + \sum_{i=0}^{k} (\eta_i \|\tilde{x}^{i+1}\| + \eta_i \|x\| + L \mu_i + \nu_i + i \xi_i) \right),
\]

where \( \xi_i := \eta_i \|\tilde{x}^{i+1} - \tilde{x}^i\| + L (\mu_i + \mu_{i-1}) + \nu_i \).

(ii) If \( \sum_{i=0}^{k} \eta_i \|\tilde{x}^{i+1}\| + 1 \to 0 \), \( \sum_{i=0}^{k} \mu_i \to 0 \) and \( \sum_{i=0}^{k} \nu_i \to 0 \), then

\[ F(\frac{1}{k} \sum_{i=0}^{k} \tilde{x}^{i+1}) \to F^*. \]

(iii) If \( \sum_{i=0}^{k} i \eta_i \|\tilde{x}^{i+1}\| + \|\tilde{x}^i\| + 1 \to 0 \), \( \sum_{i=0}^{k} i \mu_i \to 0 \) and \( \sum_{i=0}^{k} i \nu_i \to 0 \), then

\[ F(\tilde{x}^k) \to F^*. \]

(iv) (Iteration complexity of iBPG at the averaged iterate) If \( \sum \eta_k (\|\tilde{x}^{k+1}\| + 1) < \infty \), \( \sum \mu_k < \infty \), \( \sum \nu_k < \infty \) and problem (1.1) has an optimal solution \( x^* \in \text{dom } \phi \), then

\[ F\left(\frac{1}{k} \sum_{i=0}^{k} \tilde{x}^{i+1}\right) - F^* \leq O\left(\frac{1}{k}\right). \]

(v) (Iteration complexity of iBPG at the last iterate) If \( \sum \eta_k (\|\tilde{x}^{k+1}\| + \|\tilde{x}^k\| + 1) < \infty \), \( \sum k \mu_k < \infty \), \( \sum k \nu_k < \infty \) and problem (1.1) has an optimal solution \( x^* \in \text{dom } \phi \), then

\[ F(\tilde{x}^k) - F^* \leq O\left(\frac{1}{k}\right). \]
Proof. Statement (i). First, it follows from (3.4) in Lemma 3.1 that, for any $i \geq 0$ and any $x \in \text{dom } P \cap \text{dom } \phi$,

$$F(\tilde{x}^{i+1}) - F(x) \leq LD_{\phi}(x, x^{i}) - LD_{\phi}(x, \tilde{x}^{i+1}) + \eta_{i}(\|\tilde{x}^{i+1}\| + \|x\|) + L\mu_{i} + \nu_{i},$$  \hspace{1cm} (3.7)

Then, for any $k \geq 0$, summing (3.7) from $i = 0$ to $i = k$ yields

$$\sum_{i=0}^{k} F(\tilde{x}^{i+1}) - (k + 1) F(x) \leq LD_{\phi}(x, x^{0}) - LD_{\phi}(x, x^{k+1}) + \sum_{i=0}^{k} (\eta_{i}\|\tilde{x}^{i+1}\| + \eta_{i}\|x\| + L\mu_{i} + \nu_{i}).$$  \hspace{1cm} (3.8)

From this inequality and $F\left(\frac{1}{k+1} \sum_{i=0}^{k} \tilde{x}^{i+1}\right) \leq \frac{1}{k+1} \sum_{i=0}^{k} F(\tilde{x}^{i+1})$ (by the convexity of $F$), we can readily obtain (3.5). Moreover, by setting $x = \tilde{x}^{k}$ in (3.4), we have

$$F(\tilde{x}^{k+1}) - F(\tilde{x}^{k}) \leq LD_{\phi}(\tilde{x}^{k}, x^{k}) - LD_{\phi}(\tilde{x}^{k}, x^{k+1}) + \eta_{k}\|\tilde{x}^{k+1} - \tilde{x}^{k}\| + L\mu_{k} + \nu_{k} \leq \eta_{k}\|\tilde{x}^{k+1} - \tilde{x}^{k}\| + L(\mu_{k} + \mu_{k-1}) + \nu_{k} =: \xi_{k},$$

where the last inequality follows from $D_{\phi}(\tilde{x}^{k}, x^{k}) \leq \mu_{k-1}$ in condition (3.2). Then, for any $i \geq 0$, it holds that

$$iF(\tilde{x}^{i+1}) \leq iF(\tilde{x}^{i}) + i\xi_{i} \iff F(\tilde{x}^{i+1}) \geq (i + 1) F(\tilde{x}^{i+1}) - iF(\tilde{x}^{i}) - i\xi_{i},$$

which further implies that

$$\sum_{i=0}^{k} F(\tilde{x}^{i+1}) \geq (k + 1) F(\tilde{x}^{k+1}) - \sum_{i=0}^{k} i\xi_{i}.$$  

This together with (3.8) implies that

$$(k + 1)(F(\tilde{x}^{k+1}) - F(x)) \leq \sum_{i=0}^{k} F(\tilde{x}^{i+1}) - (k + 1) F(x) + \sum_{i=0}^{k} i\xi_{i} \leq LD_{\phi}(x, x^{0}) - LD_{\phi}(x, x^{k+1}) + \sum_{i=0}^{k} (\eta_{i}\|\tilde{x}^{i+1}\| + \eta_{i}\|x\| + L\mu_{i} + \nu_{i} + i\xi_{i}).$$

Then, dividing the above inequality by $k + 1$, we get (3.6).

Statement (ii). When $\frac{1}{k} \sum_{i=0}^{k-1} \eta_{i} (\|\tilde{x}^{i+1}\| + 1) \rightarrow 0$, $\frac{1}{k} \sum_{i=0}^{k-1} \mu_{i} \rightarrow 0$ and $\frac{1}{k} \sum_{i=0}^{k-1} \nu_{i} \rightarrow 0$, one can see from (3.5) that

$$\limsup_{k \rightarrow \infty} F\left(\frac{1}{k} \sum_{i=0}^{k-1} \tilde{x}^{i+1}\right) \leq F(x), \quad \forall x \in \text{dom } P \cap \text{dom } \phi.$$  

This together with $\frac{1}{k} \sum_{i=0}^{k-1} \tilde{x}^{i+1} \in \text{dom } P \cap \text{dom } \phi$ and (2.2) implies that

$$F^{*} \leq \liminf_{k \rightarrow \infty} F\left(\frac{1}{k} \sum_{i=0}^{k-1} \tilde{x}^{i+1}\right) \leq \limsup_{k \rightarrow \infty} F\left(\frac{1}{k} \sum_{i=0}^{k-1} \tilde{x}^{i+1}\right) \leq F^{*},$$

from which, we can conclude that $F\left(\frac{1}{k} \sum_{i=0}^{k-1} \tilde{x}^{i+1}\right) \rightarrow F^{*}$.

Statement (iii). When $\frac{1}{k} \sum_{i=0}^{k-1} \eta_{i} (\|\tilde{x}^{i+1}\| + \|\tilde{x}^{i}\| + 1) \rightarrow 0$, $\frac{1}{k} \sum_{i=0}^{k-1} i\xi_{i} \rightarrow 0$ and $\frac{1}{k} \sum_{i=0}^{k-1} i\nu_{i} \rightarrow 0$, it is easy to verify that $\frac{1}{k} \sum_{i=0}^{k-1} \eta_{i} \|\tilde{x}^{i+1}\| \rightarrow 0$, $\frac{1}{k} \sum_{i=0}^{k-1} \eta_{i} \|\tilde{x}^{i+1}\| \rightarrow 0$, $\frac{1}{k} \sum_{i=0}^{k-1} \mu_{i} \rightarrow 0$ and $\frac{1}{k} \sum_{i=0}^{k-1} \nu_{i} \rightarrow 0$. Using these facts and (3.6), we have

$$\limsup_{k \rightarrow \infty} F(\tilde{x}^{k}) \leq F(x), \quad \forall x \in \text{dom } P \cap \text{dom } \phi.$$  

9
This then together with (2.2) implies that

\[ F^* \leq \liminf_{k \to \infty} F(\bar{x}^k) \leq \limsup_{k \to \infty} F(\bar{x}^k) \leq F^*, \]

from which we can conclude that \( F(\bar{x}^k) \to F^* \).

Statement (iv). When problem (1.1) has an optimal solution \( x^* \) such that \( x^* \in \text{dom } \phi \), we have that \( F^* = F(x^*) = \min \{ F(x) : x \in Q \} \). Then, applying (3.5) with \( x = x^* \), we get

\[
F \left( \frac{1}{k+1} \sum_{i=0}^{k} \bar{x}^{i+1} \right) - F(x^*) \leq \frac{1}{k+1} \left( LD_\phi(x^*, x^0) + \sum_{i=0}^{k} \left( \eta_i \| \bar{x}^{i+1} \| + \eta_i \| x^* \| + L\mu_i + \nu_i \right) \right).
\]

This, together with \( \sum \eta_i (\| \bar{x}^{k+1} \| + 1) < \infty, \sum \mu_k < \infty \) and \( \sum \nu_k < \infty \), yields \( F \left( \frac{1}{k} \sum_{i=0}^{k-1} \bar{x}^{i+1} \right) - F^* \leq O(k^{-1}) \).

Statement (v). Similar to statement (iv), applying (3.6) with \( x = x^* \), we have

\[
F(\bar{x}^{k+1}) - F(x^*) \leq \frac{1}{k+1} \left( LD_\phi(x^*, x^0) + \sum_{i=0}^{k} \left( \eta_i \| \bar{x}^{i+1} \| + \eta_i \| x^* \| + L\mu_i + \nu_i + \xi_i \right) \right).
\]

Moreover, from \( \sum k\eta_k (\| \bar{x}^{k+1} \| + 1) < \infty, \sum k\mu_k < \infty \) and \( \sum k\nu_k < \infty \), we see that \( \sum \eta_k < \infty, \sum \mu_k < \infty, \sum \nu_k < \infty, \sum \eta_k \| \bar{x}^{k+1} \| < \infty \) and \( \sum k\xi_k < \infty \). This together with the above relation implies that \( F(\bar{x}^k) - F^* \leq O(k^{-1}) \).  

Remark 3.2 (Comments on iteration complexity of iBPG). We see from Theorem 3.1 that, under proper summable-error conditions, the convergence of the function value achieves the rate of \( O(1/k) \) at both the averaged iterate \( \bar{x}^k \) and the last iterate \( x^k \). The latter result requires (unsurprisingly) stronger summability conditions and covers the related complexity results in [6, 31, 50] wherein each subproblem is solved exactly. Next we comment on how the condition \( \sum \eta_k (\| \bar{x}^{k+1} \| + 1) < \infty \) (or \( \sum k\eta_k (\| \bar{x}^{k+1} \| + 1) < \infty \)) can be ensured. If one knows a priori that \( \{ \bar{x}^k \} \) will be bounded (for example, when \( \text{dom } P \cap \text{dom } \phi \) is bounded), then it readily reduces to \( \sum \eta_k < \infty \) (or \( \sum k\eta_k < \infty \)). Moreover, one could also set \( \eta_k := ((\| \bar{x}^{k+1} \| + 1)^{-1}\eta_k \) (or \( \eta_k := ((\| \bar{x}^{k+1} \| + \| \bar{x}^k \| + 1)^{-1}\eta_k \)) for all \( k \geq 0 \), where \( \{ \eta_k \} \) is an arbitrarily given summable nonnegative sequence. This then ensures \( \sum \eta_k (\| \bar{x}^{k+1} \| + 1) < \infty \) (or \( \sum k\eta_k (\| \bar{x}^{k+1} \| + \| \bar{x}^k \| + 1) < \infty \)). In this case, one actually needs to check \( \| \Delta^k \| (\| \bar{x}^{k+1} \| + 1) \leq \eta_k \) (or \( \| \Delta^k \| (\| \bar{x}^{k+1} \| + \| \bar{x}^k \| + 1) \leq \eta_k \), which implies that if \( \{ \bar{x}^k \} \) appears to be unbounded, one may need to drive \( \Delta^k \) to zero more quickly.

Theorem 3.1 gives the convergence rate of our iBPG in terms of the function value. To establish the convergence result for the sequence of iterates, we need to make some additional assumptions on the kernel function \( \phi \), which are often used for studying the convergence of a Bregman-distance-based method (see, for example, [6, 44, 49]) and can be satisfied by, for example, the quadratic kernel function \( \phi(x) := \frac{1}{2} \| x \|^2 \) and the entropy kernel function \( \phi(x) := \sum_{i=1}^{n} x_i (\log x_i - 1) \). More discussions and explanations on functions with these properties can be found in [7] and references therein.

Assumption B. Assume that the kernel function \( \phi \) satisfies the following conditions.

B1. \( \text{dom } \phi = \overline{\text{dom } \phi} = Q, \) i.e., the domain of \( \phi \) is closed.

B2. For any \( x \in \text{dom } \phi \) and \( \alpha \in \mathbb{R} \), the level set \( \{ y \in \text{int } \text{dom } \phi : D_\phi(x, y) \leq \alpha \} \) is bounded.
B3. If \( \{x^k\} \subseteq \text{int dom } \phi \) converges to some \( x^* \in \text{dom } \phi \), then \( \mathcal{D}_\phi(x^*, x^k) \to 0 \).

B4. (Convergence consistency) If \( \{x^k\} \subseteq \text{dom } \phi \) and \( \{y^k\} \subseteq \text{int dom } \phi \) are two sequences such that \( \{x^k\} \) is bounded, \( y^k \to y^* \) and \( \mathcal{D}_\phi(x^k, y^k) \to 0 \), then \( x^k \to y^* \).

With Assumption B, we have the following convergence results.

**Theorem 3.2.** Suppose that Assumption A holds, \( \sum \eta_k < \infty \), \( \sum \mu_k < \infty \) and \( \sum \nu_k < \infty \). Let \( \{x^k\} \) and \( \{\tilde{x}^k\} \) be the sequences generated by the iBPG in Algorithm 1. If \( \{x^k\} \) is bounded and the optimal solution set of problem (1.1) is nonempty, then the following statements hold.

(i) Any cluster point of \( \{\tilde{x}^k\} \) is an optimal solution of problem (1.1).

(ii) If, in addition, Assumption B holds, then the sequences \( \{x^k\} \) and \( \{\tilde{x}^k\} \) converge to the same limit that is an optimal solution of problem (1.1).

**Proof.** Statement (i). Since \( \sum \eta_k < \infty \), \( \sum \mu_k < \infty \) and \( \sum \nu_k < \infty \), it then follows from Lemma 2.3 that \( \frac{1}{k} \sum_{i=0}^{k-1} \eta_i \to 0 \), \( \frac{1}{k} \sum_{i=0}^{k-1} \mu_i \to 0 \) and \( \frac{1}{k} \sum_{i=0}^{k-1} \nu_i \to 0 \). Using these facts and the boundedness of \( \{\tilde{x}^k\} \), we can see from Theorem 3.1(iii) that \( F(\tilde{x}^k) \to F^* \). On the other hand, when the optimal solution set of problem (1.1) is nonempty, we have that \( F^* = \min \{F(x) : x \in Q\} \). Now, suppose that \( \tilde{x}^\infty \) is a cluster point (which must exist since \( \{\tilde{x}^k\} \) is bounded) and \( \{\tilde{x}^{k_i}\} \) is a convergent subsequence such that \( \lim_{i \to \infty} \tilde{x}^{k_i} = \tilde{x}^\infty \). Then,

\[
\min \{F(x) : x \in Q\} = F^* = \lim_{k \to \infty} F(\tilde{x}^k) = \lim_{k_i \to \infty} F(\tilde{x}^{k_i}) \geq F(\tilde{x}^\infty),
\]

where the last inequality follows from the lower semicontinuity of \( F \) (since \( P \) and \( f \) are closed by Assumptions A2&3). This implies that \( F(\tilde{x}^\infty) \) is finite and hence \( \tilde{x}^\infty \in \text{dom } F \). Moreover, since \( \tilde{x}^k \in \text{dom } P \cap \text{dom } \phi \subseteq \text{dom } \phi \), then \( \tilde{x}^\infty \in \text{dom } \phi = Q \). Therefore, \( \tilde{x}^\infty \) is an optimal solution of problem (1.1). This proves statement (i).

**Statement (ii).** Let \( x^\ast \) be an arbitrary optimal solution of problem (1.1). From Assumptions A3 and B1, we see that \( x^\ast \in \text{dom } P \cap \text{dom } f \cap Q = \text{dom } P \cap \text{dom } \phi \). Then, we set \( x = x^\ast \) in (3.4) and obtain after rearranging the resulting inequality that

\[
\mathcal{D}_\phi(x^\ast, x^{k+1}) \\
\leq \mathcal{D}_\phi(x^\ast, x^k) + L^{-1}(F(x^\ast) - F(\tilde{x}^{k+1})) + L^{-1}\eta_k\|x^{k+1} - x^\ast\| + L^{-1}\nu_k + \mu_k \\
\leq \mathcal{D}_\phi(x^\ast, x^{k}) + L^{-1}(\|\tilde{x}^{k+1}\| + \|x^\ast\|)\eta_k + L^{-1}\nu_k + \mu_k, \quad (3.9)
\]

where the last inequality follows from \( F(x^\ast) \leq F(\tilde{x}^{k+1}) \) for all \( k \geq 0 \). Thus, from (3.9), the nonnegativity of \( \mathcal{D}_\phi(x^\ast, x^k) \), the boundedness of \( \{\tilde{x}^k\} \), \( \sum \eta_k < \infty \), \( \sum \mu_k < \infty \), \( \sum \nu_k < \infty \) and Lemma 2.2, we obtain that \( \{\mathcal{D}_\phi(x^\ast, x^k)\} \) is convergent. This then together Assumption B2 implies that the sequence \( \{x^k\} \) is bounded and hence has at least one cluster point. Suppose that \( x^\infty \) is a cluster point and \( \{x^{k_j}\} \) is a convergent subsequence such that \( \lim_{j \to \infty} x^{k_j} = x^\infty \). Then, from \( \mathcal{D}_\phi(\tilde{x}^{k_j}, x^{k_j}) \leq \mu_{k_j} \to 0 \), the boundedness of \( \{\tilde{x}^{k_j}\} \) and Assumption B4, we have that \( \lim_{j \to \infty} \tilde{x}^{k_j} = x^\infty \). This together with statement (i) implies that \( x^\infty \) is an optimal solution of problem (1.1). Moreover, by using (3.9) with \( x^\ast \) replaced by \( x^\infty \), we can conclude that \( \{\mathcal{D}_\phi(x^\infty, x^{k_j})\} \) is convergent. On the other hand, it follows from \( \lim_{j \to \infty} x^{k_j} = x^\infty \) and Assumption B3 that \( \mathcal{D}_\phi(x^\infty, x^{k_j}) \to 0 \). Consequently, we must have that \( \mathcal{D}_\phi(x^\infty, x^{k_j}) \to 0 \). Now, let \( z \) be arbitrary cluster point of \( \{x^k\} \) with a convergent subsequence \( \{x^{k_j}\} \) such that
\( x^k \to z \). Since \( D_\phi(x^\infty, x^k) \to 0 \), then we have \( D_\phi(x^\infty, x^k) \to 0 \). From this and Assumption B4, we see that \( x^\infty = z \). Since \( z \) is arbitrary, we can conclude that \( \lim_{k \to \infty} x^k = x^\infty \). Finally, using this together with the boundedness of \( \{\bar{x}^k\} \), \( D_\phi(\bar{x}^k, x^k) \to 0 \) and Assumption B4, we deduce that \( \{\bar{x}^k\} \) also converges to \( x^\infty \). This completes the proof. \( \square \)

4 An inertial variant of the iBPG method

In this section, we shall develop an inertial variant of our iBPG to obtain a possibly faster convergence speed. Before proceeding, we introduce an additional restricted relative smoothness assumption, which modifies the one proposed by Gutman and Peña in [23, Section 3.3]. This assumption is crucial for developing the complexity of our inertial variant and can subsume the related conditions used in [4, 24, 45, 46] for developing inertial methods.

Assumption C. In addition to Assumption A3 with a closed convex set \( X \supseteq \text{dom } P \cap \text{dom } \phi \), \( f \) also satisfies the following smoothness condition relative to \( \phi \) restricted on \( X \): there exist two constants \( \tau > 0 \) and \( \gamma \geq 1 \) such that, for any \( x, \bar{z} \in \text{dom } P \cap \text{dom } \phi \) and \( z \in \text{int } \text{dom } \phi \cap X \),

\[
D_f((1 - \theta)x + \theta \bar{z}, (1 - \theta)x + \theta z) \leq \tau L \theta^\gamma D_\phi(\bar{z}, z), \quad \forall \theta \in (0, 1]. \tag{4.1}
\]

Here, \( \gamma \) is called the restricted relative smoothness exponent of \( f \) relative to \( \phi \) restricted on \( X \).

We provide below two examples of \((f, \phi)\) satisfying Assumption C.

Example 1. If \( \nabla f \) is \( L_f \)-Lipschitz continuous on \( X \) and \( \phi \) is \( \mu_\phi \)-strongly convex on \( X \), one can verify that

\[
D_f((1 - \theta)x + \theta \bar{z}, (1 - \theta)x + \theta z) \leq \frac{L_f \theta^2}{2} \| \bar{z} - z \|^2 \leq \mu_\phi^{-1} L_f \theta^2 D_\phi(\bar{z}, z).
\]

Thus, (4.1) holds with \( \tau = \mu_\phi^{-1} \), \( L = L_f \) and \( \gamma = 2 \). Here, we would like to point out that, without the restriction on the set \( X \), the above inequality may not hold for the entropy kernel function \( \phi(x) = \sum_i x_i (\log x_i - 1) \) even when \( \nabla f \) is Lipschitz continuous on \( \mathbb{R}^n \), since the entropy kernel is not strongly convex on its whole domain \( \mathbb{R}^n_+ \). In contrast, when taking into consideration of the restriction on the set \( X \) and \( X \) is a bounded and closed subset of \( \mathbb{R}^n_+ \), one can have the above inequality with the exponent \( \gamma = 2 \) for the entropy kernel function. As we shall see later from Theorem 4.1(iii), this would lead to a faster convergence rate of \( O(1/k^2) \). Therefore, introducing the set \( X \) is important to broaden the choices of \((f, \phi)\).

Example 2. If \( f \) is \( L \)-smooth relative to \( \phi \) restricted on \( X \) and the Bregman distance \( D_\phi \) has the so-called triangle scaling property (TSP) [24, Definition 2]) with a triangle scaling exponent \( \gamma \geq 1 \), one can verify that

\[
D_f((1 - \theta)x + \theta \bar{z}, (1 - \theta)x + \theta z) \leq LD_\phi((1 - \theta)x + \theta \bar{z}, (1 - \theta)x + \theta z) \leq L \theta^\gamma D_\phi(\bar{z}, z).
\]

Thus, (4.1) holds with \( \tau = 1 \) and some \( \gamma \geq 1 \) determined by \( D_\phi \). For example, when considering the entropy function \( \phi(x) = \sum_i x_i (\log x_i - 1) \) as a kernel function, \( D_\phi(\cdot, \cdot) \) is jointly convex\(^1\) and hence \( D_\phi \) has the TSP with \( \gamma = 1 \).

\(^1\)More examples on the joint convexity of the Bregman distance can be found in [8].
We also need to specify the conditions on the choice of the parameters sequence \( \{\theta_k\}_{k=-1}^\infty \), which will be used for developing the inertial method in the sequel. Specifically, the sequence of the parameters \( \{\theta_k\}_{k=-1}^\infty \subseteq (0, 1] \) with \( \theta_{-1} = \theta_0 = 1 \) is chosen such that, for all \( k \geq 1 \),

\[
\begin{align*}
\theta_k &\leq \frac{\alpha - 1}{k + \alpha - 1}, \quad \alpha \geq \gamma + 1, \quad (4.2a) \\
\tilde{\theta}_k &:= \frac{1}{\theta_{k-1}} - \frac{1 - \theta_k}{\theta_k} \geq 0, \quad (4.2b)
\end{align*}
\]

where \( \gamma \geq 1 \) is the restricted relative smoothness exponent specified in Assumption C. These conditions are inspired by several works on accelerated methods (see, for example, [17, 24, 45]), and they actually provide a unified and broader framework to choose \( \{\theta_k\} \). Using similar arguments as in [24, Lemma 3], one can show that, for any \( \alpha \geq \gamma + 1 \), \( \theta_k = \frac{\alpha - 1}{k + \alpha - 1} \), \( \forall k \geq 1 \), satisfy conditions (4.2a) and (4.2b). One can also obtain a sequence \( \{\theta_k\} \) by iteratively solving the equality form of (4.2b) via an appropriate root-finding procedure. It should be noted that these conditions allow \( \{\theta_k\} \) to decrease but not too fast. Moreover, we have the following lemma whose proof can be found in Appendix A.

**Lemma 4.1.** For any sequence \( \{\theta_k\}_{k=-1}^\infty \subseteq (0, 1] \) with \( \theta_{-1} = \theta_0 = 1 \) satisfying conditions (4.2a) and (4.2b), we have that \( \theta_k \geq \frac{1}{k+\gamma} \) for all \( k \geq 0 \) and

\[
\sum_{i=0}^k \tilde{\theta}_i \leq 2 + \frac{(k + 1 + \gamma)^\gamma}{\gamma}.
\]

We are now ready to present an inertial variant of our iBPG, denoted by v-iBPG for short. The complete framework is presented as Algorithm 2.

Our v-iBPG in Algorithm 2 is inspired by Hanzely, Richtárik and Xiao’s accelerated Bregman proximal gradient method [24], which extends Auslender and Teboulle’s improved interior gradient algorithm [4, Section 5] and Tseng’s extension [45, 46] to the relatively smooth setting. However, note that none of the methods in [4, 24, 45, 46] allow the subproblem to be solved approximately. Moreover, the analysis in [24] is established based on the relative smoothness condition and a crucial triangle scaling property (TSP) for the Bregman distance. Since TSP may not imply the strong convexity of the kernel function, the convergence results in [24] actually cannot recover the related results in [4, 45, 46] when \( \nabla f \) is \( L \)-Lipschitz continuous. In contrast, our analysis shall use the restricted relative smoothness condition (2.1) together with a more general condition (4.1) in Assumption C that could be implied by the conditions used in [4, 24, 45, 46]. Thus, our subsequent convergence results can readily subsume the related results in [4, 24, 45, 46] when the subproblem is solved exactly.

**Lemma 4.2.** Suppose that Assumptions A and C hold. Let \( \{x^k\}, \{z^k\} \) and \( \tilde{z}^k \) be the sequences generated by the v-iBPG in Algorithm 2. Then, for any \( k \geq 0 \) and any \( x \in \text{dom } P \cap \text{dom } \phi \),

\[
\begin{align*}
&\frac{1 - \theta_{k+1}}{\theta_{k+1}^i} (F(x^{k+1}) - F(x)) + \tau L D_\phi(x, z^{k+1}) \\
&\leq \frac{1 - \theta_k}{\theta_k^i} (F(x^k) - F(x)) + \tau L D_\phi(x, z^k) + e(x)\tilde{\theta}_{k+1} + \theta_k^{1-\gamma}\eta_k\|\tilde{z}^{k+1} - x\| + \tau L \mu_k + \theta_k^{1-\gamma} \nu_k,
\end{align*}
\]

where \( e(x) := F(x) - F^* \) and \( \tilde{\theta}_{k+1} := \theta_k^{1-\gamma} - \theta_{k+1}^{1-\gamma}(1 - \theta_{k+1}) \).
Algorithm 2. An inertial variant of iBPG (v-iBPG) for solving problem (1.1)

Input: Let \( \{\eta_k\}_{k=0}^{\infty}, \{\mu_k\}_{k=0}^{\infty} \) and \( \{\nu_k\}_{k=0}^{\infty} \) be three sequences of nonnegative scalars, and \( \mathcal{X} \supset \text{dom} \ P \cap \text{dom} \ \phi \) be the closed convex set in Assumption A3. Choose \( x^0 \in \text{dom} \ P \cap \text{dom} \ \phi \) and \( z^0 \in \text{int} \text{dom} \ \phi \cap \mathcal{X} \) arbitrarily. Set \( \theta_0 = 1 \) and \( k = 0. \)

while the termination criterion is not met, do

Step 1. Compute \( y^k = (1 - \theta_k)x^k + \theta_k z^k. \)

Step 2. Find a pair \((z^{k+1}, \bar{z}^{k+1})\) by approximately solving the following problem

\[
\begin{align*}
\min_{z} & \ P(z) + \langle \nabla f(y^k), z - y^k \rangle + \tau L \theta_k^{-1} \mathcal{D}_{\phi}(z, z^k), \\
\text{such that} & \ z^{k+1} \in \mathcal{X} \cap \text{int dom} \ \phi, \ \bar{z}^{k+1} \in \text{dom} \ P \cap \text{dom} \ \phi \text{ and} \\
& \Delta^k \in \partial_{\mu_k} P(\bar{z}^{k+1}) + \nabla f(y^k) + \tau L \theta_k^{-1} (\nabla \phi(z^{k+1}) - \nabla \phi(z^k)) \\
& \text{with} \ \|\Delta^k\| \leq \eta_k, \ \mathcal{D}_{\phi}(\bar{z}^{k+1}, z^{k+1}) \leq \mu_k.
\end{align*}
\]

Step 3. Compute \( x^{k+1} = (1 - \theta_k)x^k + \theta_k \bar{z}^{k+1}. \)

Step 4. Choose \( \theta_{k+1} \in (0, 1] \) satisfying conditions (4.2a) and (4.2b).

Step 5. Set \( k = k + 1 \) and go to Step 1.

end while

Output: \( x^k \)

Proof. First, using the similar arguments of deducing (A.3), we have that, for any \( k \geq 0 \) and any \( x \in \text{dom} \ P \cap \text{dom} \ \phi, \)

\[
P(\bar{z}^{k+1}) \leq P(x) - \langle \nabla f(y^k), \bar{z}^{k+1} - x \rangle + \tau L \theta_k^{-1} \mathcal{D}_{\phi}(x, z^k) - \tau L \theta_k^{-1} \mathcal{D}_{\phi}(x, z^{k+1}) \\
- \tau L \theta_k^{-1} \mathcal{D}_{\phi}(\bar{z}^{k+1}, z^k) + \eta_k \|z^{k+1} - x\| + \tau L \theta_k^{-1} \mu_k + \nu_k.
\]

Then, we obtain that

\[
P(\bar{z}^{k+1}) + f(y^k) + \langle \nabla f(y^k), \bar{z}^{k+1} - y^k \rangle \leq P(x) + f(x) + \tau L \theta_k^{-1} \mathcal{D}_{\phi}(x, z^k) - \tau L \theta_k^{-1} \mathcal{D}_{\phi}(x, z^{k+1}) \\
- \tau L \theta_k^{-1} \mathcal{D}_{\phi}(\bar{z}^{k+1}, z^k) + \eta_k \|z^{k+1} - x\| + \tau L \theta_k^{-1} \mu_k + \nu_k.
\]

where the first inequality follows from (4.6) and the second inequality follows from the convexity of \( f. \) Next, we see that

\[
\begin{align*}
F(x^{k+1}) & = P(x^{k+1}) + f(x^{k+1}) = P((1 - \theta_k)x^k + \theta_k \bar{z}^{k+1}) + f(x^{k+1}) \\
& \leq (1 - \theta_k) P(x^k) + \theta_k P(\bar{z}^{k+1}) + f(y^k) + \langle \nabla f(y^k), x^{k+1} - y^k \rangle + D_f(x^{k+1}, y^k) \\
& = (1 - \theta_k) \left[ P(x^k) + f(y^k) + \langle \nabla f(y^k), x^k - y^k \rangle \right] + \theta_k \left[ P(\bar{z}^{k+1}) + f(y^k) + \langle \nabla f(y^k), \bar{z}^{k+1} - y^k \rangle \right] \\
& \quad + D_f((1 - \theta_k)x^k + \theta_k \bar{z}^{k+1}, (1 - \theta_k)x^k + \theta_k z^k) \\
& \leq (1 - \theta_k) F(x^k) + \theta_k F(x) + \tau L \theta_k^{-1} \mathcal{D}_{\phi}(x, z^k) - \tau L \theta_k^{-1} \mathcal{D}_{\phi}(x, z^{k+1}) + \theta_k \eta_k \|\bar{z}^{k+1} - x\| + \tau L \theta_k^{-1} \mu_k + \theta_k \nu_k,
\end{align*}
\]
where the first inequality follows from the convexity of $P$ and the definition of $D_f$, the last inequality follows from (4.1) and (4.7). Now, in the above inequality, subtracting $F(x)$ from both sides, dividing both sides by $\theta_k^\gamma$ and rearranging the resulting relation, we have that for any $k \geq 0$,

$$
\theta_k^{-\gamma}(F(x^{k+1}) - F(x)) + \tau LD_\phi(x, z^{k+1}) \\
\leq \theta_k^{-\gamma}(1 - \theta_k)(F(x^k) - F(x)) + \tau LD_\phi(x, z^k) + \theta_k^{1-\gamma}(\|z^{k+1} - x\|) + \tau \mu_k + \theta_k^{1-\gamma}\nu_k. \quad (4.8)
$$

Moreover,

$$
\theta_k^{-\gamma}(F(x^{k+1}) - F(x)) = \theta_{k+1}^{-\gamma}(1 - \theta_{k+1})(F(x^{k+1}) - F(x)) + \theta_{k+1}^{-\gamma}(F(x^{k+1}) - F(x)) \\
\geq \theta_{k+1}^{-\gamma}(1 - \theta_{k+1})(F(x^{k+1}) - F(x)) + \theta_{k+1}^{-\gamma}(F^* - F(x)),
$$

where the last inequality follows from $\theta_{k+1} \geq 0$ (by condition (4.2b)) and $F(x^{k+1}) \geq F^*$ for all $k \geq 0$. Then, combining the above relations, we can obtain (4.5) and complete the proof.

**Theorem 4.1.** Suppose that Assumptions A and C hold. Let $\{x^k\}$, $\{z^k\}$ and $\{\tilde{z}^k\}$ be the sequences generated by the v-iBPG in Algorithm 2. Then, the following statements hold.

(i) For any $k \geq 0$ and any $x \in \text{dom } P \cap \text{dom } \phi$,

$$
F(x^{k+1}) - F(x) \\
\leq \left( \frac{\alpha - 1}{k + \alpha - 1} \right)^\gamma \left( \tau LD_\phi(x, z^0) + e(x)\sum_{i=0}^k \tilde{\theta}_i + \sum_{i=0}^k \theta_i^{1-\gamma}\eta_i(\|z^{i+1}\| + \|x\|) + \tau \mu_i + \theta_i^{1-\gamma}\nu_i \right),
$$

where $e(x) := F(x) - F^*$, $\tilde{\theta}_{k+1} := \theta_k^{-\gamma} - \theta_{k+1}^{-\gamma}(1 - \theta_{k+1})$, and $\alpha$ is as in (4.2a).

(ii) If $k^{-\gamma}\sum_{i=0}^{k-1} \theta_i^{1-\gamma}\eta_i(\|z^{i+1}\| + 1) \to 0$, $k^{-\gamma}\sum_{i=0}^{k-1} \mu_i \to 0$ and $k^{-\gamma}\sum_{i=0}^{k-1} \theta_i^{1-\gamma}\nu_i \to 0$, then

$$
F(x^k) \to F^*.
$$

(iii) (Iteration complexity of v-iBPG) If $\sum \theta_k^{1-\gamma}\eta_k(\|z^{k+1}\| + 1) < \infty$, $\sum \mu_k < \infty$ and $\sum \theta_k^{1-\gamma}\nu_k < \infty$ and problem (1.1) has an optimal solution $x^*$ such that $x^* \in \text{dom } \phi$, then

$$
F(x^k) - F^* \leq O \left( \frac{1}{k^{\gamma}} \right).
$$

**Proof.** Statement (i). First, we see from (4.5) that, for any $i \geq 0$ and any $x \in \text{dom } P \cap \text{dom } \phi$,

$$
\theta_i^{-\gamma}(1 - \theta_{i+1})(F(x^{i+1}) - F(x)) + \tau LD_\phi(x, z^{i+1}) \\
\leq \theta_i^{-\gamma}(1 - \theta_i)(F(x^i) - F(x)) + \tau LD_\phi(x, z^i) + e(x)\tilde{\theta}_{i+1} \\
+ \theta_i^{1-\gamma}\eta_i(\|z^{i+1}\| + \|x\|) + \tau \mu_i + \theta_i^{1-\gamma}\nu_i.
$$

Then, for any $k \geq 1$, summing the above inequality from $i = 0$ to $i = k - 1$ and recalling $\theta_0 = 1$ results in

$$
\theta_k^{-\gamma}(1 - \theta_k)(F(x^k) - F(x)) + \tau LD_\phi(x, z^k) \\
\leq \tau LD_\phi(x, z^0) + e(x)\sum_{i=0}^{k-1} \tilde{\theta}_{i+1} + \sum_{i=0}^{k-1} \left( \theta_i^{1-\gamma}\eta_i(\|z^{i+1}\| + \|x\|) + \tau \mu_i + \theta_i^{1-\gamma}\nu_i \right).
$$

15
Moreover, it is easy to verify that this inequality also holds for $k = 0$ (using (4.8) for $k = 0$). Thus, multiplying the above inequality by $\{\theta_k\}$ and using condition (4.2a), we can obtain the desired result in statement (i).

**Statement (ii).** Note from Lemma 4.1 that $\sum_{i=0}^{k} \hat{\theta}_i \leq 2 + \frac{(k+1+\gamma)^\gamma}{\gamma}$. Thus, one can verify that, for all $k \geq 0$,

$$\left(\frac{\alpha - 1}{k + \alpha - 1}\right)^\gamma \sum_{i=0}^{k} \hat{\theta}_i \leq 2 \left(\frac{\alpha - 1}{k + \alpha - 1}\right)^\gamma + \left(\frac{\alpha - 1}{k + \alpha - 1}\right)^\gamma \frac{(k+1+\gamma)^\gamma}{\gamma} \leq 2 + \frac{(\alpha - 1)^\gamma}{\gamma} \left(\frac{k + 1 + \gamma}{k + \alpha - 1}\right)^\gamma \leq 2 + \frac{\max \{(\alpha - 1)^\gamma, (1 + \gamma)^\gamma\}}{\gamma},$$

where the second inequality follows from $\frac{\alpha - 1}{k + \alpha - 1} \leq 1$ and the last inequality follows from $\frac{k+1+\gamma}{k+\alpha-1} \leq \max \{1, \frac{1+\gamma}{\alpha-1}\}$. Using this fact, $k^{-\gamma} \sum_{i=0}^{k-1} \theta_i^{1-\gamma} \eta_i (\|z_i^{1+1}\| + 1) \to 0$, $k^{-\gamma} \sum_{i=0}^{k-1} \mu_i \to 0$, and $k^{-\gamma} \sum_{i=0}^{k-1} \theta_i^{1-\gamma} \nu_i \to 0$, one can see from statement (i) that

$$\limsup_{k \to \infty} \frac{F(x^k)}{\nu(x)} \leq \frac{F(x)}{\nu(x)} + \left(2 + \frac{\max \{(\alpha - 1)^\gamma, (1 + \gamma)^\gamma\}}{\gamma}\right) e(x), \quad \forall x \in \text{dom} \cap \text{dom} \phi.$$  

This together with (2.2) implies that

$$F^* \leq \liminf_{k \to \infty} F(x^k) \leq \limsup_{k \to \infty} F(x^k) \leq F^*,$$

from which, we can conclude that $F(x^k) \to F^*$ and prove statement (ii).

**Statement (iii).** When problem (1.1) has an optimal solution $x^*$ such that $x^* \in \text{dom} \phi$, we have $F^* = F(x^*) = \min \{F(x) : x \in Q\}$. Then, applying statement (i) with $x = x^*$, we get

$$F(x^{k+1}) - F^* \leq \left(\frac{\alpha - 1}{k + \alpha - 1}\right)^\gamma \left(\tau LD_\phi(x^*, z^0) + \sum_{i=0}^{k} \left(\theta_i^{1-\gamma} \eta_i \|z_i^{1+1}\| + \theta_i^{1-\gamma} \eta_i \|x^*\| + \tau L \mu_i + \theta_i^{1-\gamma} \nu_i\right)\right).$$

This together with $\sum \theta_k^{1-\gamma} \eta_k (\|z_k^{1+1}\| + 1) < \infty$, $\sum \mu_k < \infty$ and $\sum \theta_k^{1-\gamma} \nu_k < \infty$ implies that $F(x^k) - F^* \leq O(k^{-\gamma})$. This completes the proof. \qed

One can see from Theorem 4.1 that our v-iBPG achieves a flexible convergence rate of $O(1/k^\gamma)$ with $\gamma \geq 1$ being a restricted relative smoothness exponent. Thus, when $\gamma > 1$, the v-iBPG indeed improves the $O(1/k)$ convergence rate of the iBPG at both the averaged iterate and the last iterate. Moreover, similar to the discussions in Remark 3.2, the condition that $\sum \theta_k^{1-\gamma} \eta_k (\|z_k^{1+1}\| + 1) < \infty$ simply reduces to $\sum \theta_k^{1-\gamma} \eta_k < \infty$ if, for example, dom $P \cap$ dom $\phi$ is bounded. One could also set $\eta_k := (\|z_k^{1+1}\| + 1)^{-1} \theta_k^{1-\gamma} \eta_k$ for all $k \geq 0$ with an arbitrarily summable nonnegative sequence $\{\eta_k\}$ to ensure this condition.

In particular, when $\gamma = 2$ and dom $P \cap$ dom $\phi$ is bounded, along with the fact $\theta_k = O(1/k)$ (see condition (4.2a) and Lemma 4.1), the summable-error conditions required by
the v-iBPG in Theorem 4.1(iii) (which boils down to \( \max \{ \sum k \eta_k, \sum \mu_k, \sum k \nu_k \} < \infty \)) are indeed weaker than those required by the iBPG in Theorem 3.1(v) (which boils down to \( \max \{ \sum k \eta_k, \sum k \mu_k, \sum k \nu_k \} < \infty \)). Thus, it is interesting to see that the v-iBPG achieves a faster convergence rate at the last iterate, while requiring weaker error controls than the iBPG. But we should be mindful that, since \( \theta_k \) is contained in the subproblem (4.3) of the v-iBPG as the proximal parameter and it goes to zero eventually, we may have a more difficult subproblem to solve when \( \theta_k \) becomes very small.

Before closing this section, we give the subsequential convergence result for the v-iBPG, but the global convergence could be more challenging to establish even when the subproblem is solved exactly and will be left for future research.

**Theorem 4.2** (Subsequential convergence of v-iBPG). Suppose that Assumptions A and C hold. Let \( \{ x^k \} \) and \( \{ \tilde{z}^k \} \) be the sequences generated by the v-iBPG in Algorithm 2. If \( \sum \theta_k^{1-\gamma} \| \tilde{z}^{k+1} \| + 1 < \infty \), \( \sum k \mu_k < \infty \), \( \sum \theta_k^{1-\gamma} k \nu_k < \infty \), the optimal solution set of problem (1.1) is nonempty and \( \{ x^k \} \) is bounded, then any cluster point of \( \{ x^k \} \) is an optimal solution of problem (1.1).

**Proof.** If the optimal solution set of problem (1.1) is nonempty and \( \{ x^k \} \) is bounded, we have that \( F^* = \min \{ F(x) : x \in Q \} \) and \( \{ x^k \} \) has at least one cluster point. Suppose that \( x^\infty \) is a cluster point and \( \{ x^{k_i} \} \) is a convergent subsequence such that \( \lim_{i \to \infty} x^{k_i} = x^\infty \). Then, from Theorem 4.1(iii) and the lower semicontinuity of \( F \) (since \( P \) and \( f \) are closed by Assumptions A2&3), we see that

\[
\min \{ F(x) : x \in Q \} = F^* = \lim_{k \to \infty} F(x^k) \geq F(x^\infty).
\]

This implies that \( F(x^\infty) \) is finite and hence \( x^\infty \in \text{dom} F \). Moreover, since \( x^k \in \text{dom} P \cap \text{dom} \phi \subseteq \text{dom} \phi \), then \( x^\infty \in \text{dom} \phi = Q \). Therefore, \( x^\infty \) is an optimal solution of problem (1.1). This completes the proof. \( \square \)

5 Numerical experiments

In this section, we conduct some numerical experiments to test our iBPG and v-iBPG for solving a relaxation of the quadratic assignment problem (QAP). Our purpose here is to preliminarily show the influence of the inexactness errors on the convergence behaviors of different methods. All experiments in this section are run in MATLAB R2016b on a Windows workstation with Intel Xeon Processor E-2176G@3.70GHz and 64GB of RAM.

The QAP is a classical discrete optimization problem, which is given by

\[
\min_X \langle X, AXB \rangle \quad \text{s.t.} \quad Xe = e, \ X^T e = e, \ X \in \{0,1\}^{n \times n},
\]

where \( A, B \in \mathbb{R}^{n \times n} \) are given symmetric matrices, \( e \in \mathbb{R}^n \) is the vector of all ones and \( \{0,1\}^{n \times n} \) denotes the set of matrices with only 0 or 1 entries. It is well-known that the above QAP is NP-hard and various relaxations have been proposed to obtain its lower bounds. Among them, Anstreicher and Brixius [2] have shown that a reasonably good lower bound for the above QAP can be obtained by solving the following convex quadratic programming problem:

\[
\min_X \langle X, \mathcal{H}(X) \rangle \quad \text{s.t.} \quad X \in \Omega := \{ X \in \mathbb{R}^{n \times n} : Xe = e, \ X^T e = e, \ X \geq 0 \},
\] (5.1)
where $\mathcal{H} : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$ is a self-adjoint positive semidefinite linear operator defined by $\mathcal{H}(X) := AXB - SX - XT$, and $S, T \in \mathbb{R}^{n \times n}$ are symmetric matrices that are derived from the spectral decompositions of $A$ and $B$, respectively; the details can be found in [2, Section 4]. Note that problem (5.1) falls into the form of (1.1) via some simple reformulations, and thus our iBPG and v-iBPG are applicable.

### 5.1 Implementations of iBPG and v-iBPG

To apply the iBPG and v-iBPG, we equivalently reformulate (5.1) as

$$
\min_X \delta_{\Omega^e}(X) + \langle X, \mathcal{H}(X) \rangle \quad \text{s.t.} \quad X \geq 0, \tag{5.2}
$$

where $\Omega^e := \{ X \in \mathbb{R}^{n \times n} : Xe = e, X^\top e = e \}$ is an affine space. Clearly, this problem takes the form of (1.1) with $P(X) := \delta_{\Omega^e}(X)$, $f(X) := \langle X, \mathcal{H}(X) \rangle$ and $Q := \mathbb{R}^{n \times n}_+$. Then, we consider the entropy kernel function $\phi(X) = \sum_{ij} x_{ij}(\log x_{ij} - 1)$ whose domain is $\mathbb{R}^{n \times n}_+$ and choose $X = [0,1]^{n \times n}$ which contains $\text{dom} \ P \cap \text{dom} \phi$. One can easily verify that $\phi$ is 1-strongly convex on $X$. This, together with the Lipschitz continuity of $\nabla f$, implies that $f$ is $L$-smooth relative to $\phi$ restricted on $X$ with $L = 2\|\mathcal{H}\|$. Using these facts, one can further see from Example 1 in Section 4 that $(f, \phi)$ satisfies Assumption C with $\tau = 1$ and $\gamma = 2$. Thus, we can apply our iBPG and v-iBPG with the entropy kernel function, denoted in short by iBPG-ent and v-iBPG-ent respectively, to solve problem (5.2) (hence problem (5.1)). Moreover, by Theorem 4.1(iii), our v-iBPG-ent can achieve a faster rate of $O(1/k^2)$ complexity. The acceleration of v-iBPG-ent over iBPG-ent is also empirically verified from the subsequent numerical results presented in Figures 1 and 2.

The subproblem at the $k$-th iteration ($k \geq 0$) takes the following form

$$
\min_X \delta_{\Omega^e}(X) + \langle C^k, X \rangle + \varepsilon_k D_\phi(X, S^k),
$$

where $C^k := 2\mathcal{H}(S^k)$, $S^k \in \mathbb{R}^{n \times n}$ and $\varepsilon_k > 0$ are specified in (3.1) or (4.3). This problem is further equivalent to

$$
\min_X \langle M^k, X \rangle + \varepsilon_k \sum_{ij} x_{ij}(\log x_{ij} - 1), \quad \text{s.t.} \quad Xe = e, \quad X^\top e = e, \tag{5.3}
$$

where $M^k := C^k - \varepsilon_k \log S^k$. Problem (5.3) has the same form as the entropic regularized optimal transport problem and hence can be readily solved by the popular Sinkhorn’s algorithm; see [38, Section 4.2] for more details. Specifically, let $\Xi^k := e^{-M^k/\varepsilon_k}$. Then, given an arbitrary initial positive vector $u^{k,0}$, the iterative scheme is given by

$$
u^{k,t} = e./((\Xi^k)^\top u^{k,t}), \quad v^{k,t} = e./((\Xi^k)^\top u^{k,t}),
$$

where ‘./’ denotes the entrywise division between two vectors. When a pair $(u^{k,t}, v^{k,t})$ is obtained based on a certain stopping criterion, an approximate solution of (5.3) can be recovered by setting $X^{k,t} := \text{Diag}(u^{k,t}) \Xi^k \text{Diag}(v^{k,t})$. Note that $X^{k,t}$ is in general not exactly feasible. Thus, a proper projection or rounding procedure is needed for the verification of the inexact condition. Let $\mathcal{G}_\Omega$ be a rounding procedure given in [1, Algorithm 2] so that $X^{k,t} := \mathcal{G}_\Omega(X^{k,t}) \in \Omega$. Then, using similar arguments as in [49, Section 4.2], we have

$$
0 \in \partial \delta_{\Omega^e}(\widehat{X}^{k,t}) + C^k + \varepsilon_k \left( \log X^{k,t} - \log S^k \right).
$$
From this relation, we see that our inexact condition (3.2) or (4.4) is verifiable at the pair \((X^{k,t}, \tilde{X}^{k,t})\), and can be satisfied when the Bregman distance \(D_\phi(\tilde{X}^{k,t}, X^{k,t})\) is smaller than a specified tolerance parameter \(\mu_k\). Here, we would like to remind the reader that, in this case, no error occurs on the left-hand-side of the optimality condition (i.e., \(\Delta^k = 0\)) and the computation of \(\partial \delta_\Omega\) (i.e., \(\nu^k = 0\)), but this is really depends on the problem and the subroutine used to solve the subproblem. Indeed, different subroutines may incur errors at different parts of the optimality condition. We refer the reader to an example in [19, Section 3.2] where the \(\nu\)-subdifferential is incurred when employing a block coordinate descent method to solve the subproblem in the entropic proximal point algorithm. Therefore, a flexible inexact condition in form of (3.2) or (4.4) that allows different type of errors would be useful and necessary to fit different situations in practice.

In view of the above, we employ Sinkhorn’s algorithm as a subroutine in our iBPG-ent and v-iBPG-ent, and as guided by our inexact condition, at the \(k\)-th iteration \((k \geq 0)\), we will terminate Sinkhorn’s algorithm when

\[
D_\phi(\tilde{X}^{k,t}, X^{k,t}) \leq \max \left\{ \frac{1}{(k+1)^p}, 10^{-10} \right\}.
\]

where \(p > 0\) controls the tightness of the inexact tolerance requirement and will be specified later.

Moreover, in v-iBPG-ent, we choose \(\theta_k = \frac{\alpha - 1}{k+\alpha - 1}\) with \(\alpha = 5\) for all \(k \geq 1\). The test instances \((A, B)\) are obtained from the QAP Library [16]. Here, we choose lipa90b \((n = 90)\) and wil100 \((n = 100)\). We initialize all methods with \(X^0 := ee^T\). To maximize the effect of inexactness, we simply employ the cold-start strategy to initialize Sinkhorn’s algorithm. Specifically, at each iteration, we always initialize Sinkhorn’s algorithm with \(v^{k,0} = e\). Moreover, we terminate the iBPG-ent and v-iBPG-ent when the number of Sinkhorn iterations reaches \(5 \times 10^5\).

Since problem (5.1) is a quadratic programming (QP) problem, we also apply Gurobi (with default settings) to solve it. It is well known that Gurobi is a powerful commercial solver for QPs and is able to obtain high quality solutions. Therefore, we will use the objective function value obtained by Gurobi as the benchmark to evaluate the qualities of the solutions obtained by our methods.

### 5.2 Comparison results

Figures 1 and 2 show the comparison results between iBPG-ent and v-iBPG-ent on lipa90b and wil100, respectively. In each figure, we plot the normalized function value (denoted by “nfval”) against the number of outer iterations (denoted by “out#”) and inner iterations (namely, Sinkhorn iterations, denoted by “sink#”), respectively. Here, “nfval” is computed by \(|\langle \mathcal{G}_\Omega(X^k), \mathcal{H}(\mathcal{G}_\Omega(X^k)) \rangle - f^* \| f^* |\) and \(|\langle \mathcal{G}_\Omega(X^{k,t}), \mathcal{H}(\mathcal{G}_\Omega(X^{k,t})) \rangle - f^* \| f^* |\), respectively, where \(f^*\) is the optimal function value computed by Gurobi, \(X^k\) is the approximate solution obtained at the \(k\)-th outer iteration and \(X^{k,t}\) is the approximate solution computed by Sinkhorn’s algorithm at the \(t\)-th inner iteration of the \(k\)-th outer iteration. Note that the computation mainly lies in the Sinkhorn iteration and hence the total computational cost is directly proportional to the number of Sinkhorn iterations. From the results, we have several observations as follows.

The overall performance (namely, “nfval” vs “sink#”) of either iBPG-ent or v-iBPG-ent is unsurprisingly affected by the value of \(p\) which determines the tightness of the inexact tolerance requirement. Given a fixed number of inner iterations, a smaller \(p\) will usually lead to a larger number of outer iterations, as shown in the figures, since the stopping criterion (5.4) is easier to
Figure 1: Comparisons between iBPG-ent and v-iBPG-ent on lipa90b. For the better visualization, we only show plots of “nfval” vs “out#” within $2 \times 10^4$ outer iterations.

satisfy for a smaller $p$. However, when $p$ is too small, which means that the errors in solving the subproblems may be too loose, the method may not converge to a high precision solution (see Figures 1 and 2 for $p = 0.1$, where “nfval” stagnates before reaching the level of $10^{-6}$). This observation matches our theoretical results, where the convergence (rate) is only guaranteed under certain requirements on the inexact errors. On the other hand, the choice of $p = 3.1$ yields the fastest tolerance decay, but does not give the best overall computational efficiency for any method, mainly due to the excessive cost of solving each subproblem. This shows that a tighter tolerance requirement does not necessarily lead to a better overall efficiency and there is a trade-off between the number of outer and inner iterations. In both Figures 1 and 2, choosing $p = 1.1$ leads to the best overall efficiency. Moreover, one can also see that the v-iBPG-ent always outperforms the iBPG-ent, even for $p = 0.1$. This empirically verifies the improved iteration complexity of the v-iBPG-ent under similar tolerance requirements for the iBPG-ent, as we can expect from Theorem 4.1(iii) with $\gamma = 2$ in this case.

In summary, allowing the subproblem to be solved approximately is necessary and important for the BPG and its inertial variant to be implementable and practical. Different methods also have different inherent inexactness tolerance requirements. The study of such phenomena can deepen our understanding of these methods and guide us to choose proper parameters in practical implementations.
6 Concluding remarks

In this paper, we develop an inexact Bregman proximal gradient (iBPG) method based on a novel two-point inexact stopping condition, and establish the iteration complexity of $O(1/k)$ as well as the convergence of the sequence under some proper conditions. To improve the convergence speed, we further develop an inertial variant of our iBPG (denoted by v-iBPG) and show that it has the iteration complexity of $O(1/k^\gamma)$, where $\gamma \geq 1$ is a restricted relative smoothness exponent. Thus, when $\gamma > 1$, the v-iBPG can improve the $O(1/k)$ convergence rate of the iBPG. We also conduct some preliminary numerical experiments for solving a relaxation of the quadratic assignment problem to show the convergence behaviors of the iBPG and v-iBPG under different inexactness settings.

Appendix A Proofs of supporting lemmas

A.1 Proof of Lemma 2.3

For any $\varepsilon > 0$, we have from $\sum_{k=0}^{\infty} \alpha_k < \infty$ that there exists $\tilde{N} \geq 0$ such that $\sum_{k=\tilde{N}}^{\infty} \alpha_k \leq \varepsilon/2$.

Now, let $N_\varepsilon := \max \left\{ \tilde{N} + 1, 2\varepsilon^{-1} \sum_{i=0}^{\tilde{N}-1} i\alpha_i \right\}$. Then, for any $k \geq N_\varepsilon$, we see that

$$\frac{1}{k} \sum_{i=0}^{k-1} i\alpha_i = \frac{1}{k} \sum_{i=0}^{\tilde{N}-1} i\alpha_i + \frac{1}{k} \sum_{i=\tilde{N}}^{k-1} i\alpha_i \leq \frac{1}{N_\varepsilon} \sum_{i=0}^{\tilde{N}-1} i\alpha_i + \sum_{i=\tilde{N}}^{k-1} \alpha_i \leq \varepsilon.$$ 

Since $\varepsilon$ is arbitrary, we can conclude that $\frac{1}{k} \sum_{i=0}^{k-1} i\alpha_i \to 0$. 

Figure 2: Comparisons between iBPG-ent and v-iBPG-ent on wil100. For the better visualization, we only show plots of “nfval” vs “out#” within $2 \times 10^4$ outer iterations.
A.2 Proof of Lemma 2.4

First, from condition (2.4), there exists \( d \in \partial_\nu P(\bar{x}^*) \) such that \( \Delta = d + \nabla f(y) + \lambda (\nabla \phi(x^*) - \nabla \phi(y)) \). Then, for any \( x \in \text{dom} P \cap \text{dom} \phi \), we see that

\[
P(x) \geq P(\bar{x}^*) + \langle d, x - \bar{x}^* \rangle - \nu = P(\bar{x}^*) + \langle \Delta - \nabla f(y) - \lambda (\nabla \phi(x^*) - \nabla \phi(y)), x - \bar{x}^* \rangle - \nu,
\]

which implies that

\[
P(\bar{x}^*) \leq P(x) - \langle \nabla f(y), \bar{x}^* - x \rangle + \lambda (\nabla \phi(x^*) - \nabla \phi(y)), x - \bar{x}^* \rangle + \eta \| \bar{x}^* - x \| + \nu. \tag{A.1}
\]

Note from the four points identity (2.3) that

\[
\langle \nabla \phi(x^*) - \nabla \phi(y), x - \bar{x}^* \rangle = D_\phi(x, y) - D_\phi(x, x^*) - D_\phi(\bar{x}^*, y) + D_\phi(\bar{x}^*, x^*) \leq D_\phi(x, y) - D_\phi(x, x^*) - D_\phi(\bar{x}^*, y) + \mu. \tag{A.2}
\]

Thus, combining (A.1) and (A.2), we obtain that

\[
P(\bar{x}^*) \leq P(x) - \langle \nabla f(y), \bar{x}^* - x \rangle + \lambda D_\phi(x, y) - \lambda D_\phi(x, x^*) - \lambda D_\phi(\bar{x}^*, y) + \eta \| \bar{x}^* - x \| + \lambda \mu + \nu. \tag{A.3}
\]

On the other hand, since \( f \) is \( L \)-smooth relative to \( \phi \) restricted on \( \mathcal{X} \) (by Assumption A3) and \( f \) is convex with \( \text{dom} f \supseteq \text{dom} \phi \), then

\[
f(\bar{x}^*) \leq f(y) + \langle \nabla f(y), \bar{x}^* - y \rangle + LD_\phi(\bar{x}^*, y),
\]

\[
f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle.
\]

Summing the above two inequalities, we obtain that

\[
f(\bar{x}^*) \leq f(x) + \langle \nabla f(y), \bar{x}^* - x \rangle + LD_\phi(\bar{x}^*, y). \tag{A.4}
\]

Thus, summing (A.3) and (A.4), one can obtain the desired result.

A.3 Proof of Lemma 4.1

We first show by induction that \( \theta_k \geq \frac{1}{k+\gamma} \) for all \( k \geq 0 \). This obviously holds for \( k = 0 \) since \( \theta_0 = 1 \). Suppose that \( \theta_k \geq \frac{1}{k+\gamma} \) holds for some \( k \geq 0 \). Since \( 0 < \theta_k \leq 1 \) and \( \gamma \geq 1 \), then \( (1 - \theta_k)^\gamma \leq 1 - \theta_k \). Using this and (4.2b), we see that

\[
\frac{(1 - \theta_{k+1})^\gamma}{\theta_{k+1}^\gamma} \leq \frac{1}{\theta_k^\gamma} \iff \frac{1 - \theta_{k+1}}{\theta_{k+1}} \leq \frac{1}{\theta_k},
\]

from which, we have that \( \theta_{k+1} \geq \frac{\theta_k}{\theta_{k+1}} = \frac{1}{k+1+\gamma} \). This completes the induction. Using this fact, we further get

\[
\sum_{i=0}^{k} \frac{1}{\theta_{i+1}^\gamma} - \frac{1 - \theta_i}{\theta_i^\gamma} = \frac{1}{\theta_k^\gamma} - \sum_{i=0}^{k} \frac{1}{\theta_i^\gamma-1} \leq 2 + \sum_{i=1}^{k} \frac{1}{\theta_i^{\gamma-1}} \leq 2 + \sum_{i=1}^{k} (i + \gamma)^{\gamma-1} \leq 2 + \int_0^{k+1} (t + \gamma)^{\gamma-1} dt \leq 2 + \frac{(k+1+\gamma)^\gamma}{\gamma}.
\]

This completes the proof.
References

[1] J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[2] K.M. Anstreicher and N.W. Brixius. A new bound for the quadratic assignment problem based on convex quadratic programming. *Math. Program.*, 89(3):341–357, 2001.

[3] J-F Aujol and Ch Dossal. Stability of over-relaxations for the forward-backward algorithm, application to FISTA. *SIAM J. Optim.*, 25(4):2408–2433, 2015.

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

[5] H.H. Bauschke, J. Bolte, J. Chen, M. Teboulle, and X. Wang. On linear convergence of non-Euclidean gradient methods without strong convexity and Lipschitz gradient continuity. *J. Optim. Theory Appl.*, 182(3):1068–1087, 2019.

[6] H.H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Math. Oper. Res.*, 42(2):330–348, 2017.

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

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

[9] H.H. Bauschke and P.L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, volume 408. Springer, 2011.

[10] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, 2003.

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

[12] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM J. Optim.*, 28(3):2131–2151, 2018.

[13] S. Bonettini, F. Porta, and V. Ruggiero. A variable metric forward-backward method with extrapolation. *SIAM J. Sci. Comput.*, 38(4):A2558–A2584, 2016.

[14] S. Bonettini, S. Rebegoldi, and V. Ruggiero. Inertial variable metric techniques for the inexact forward-backward algorithm. *SIAM J. Sci. Comput.*, 40(5):A3180–A3210, 2018.

[15] L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.*, 7(3):200–217, 1967.
[16] R.E. Burkard, S.E. Karisch, and F. Rendl. QAPLIB – a quadratic assignment problem library. *J. Glob. Optim.*, 10(4):391–403, 1997.

[17] A. Chambolle and Ch Dossal. On the convergence of the iterates of the “fast iterative shrinkage/thresholding algorithm”. *J. Optim. Theory Appl.*, 166(3):968–982, 2015.

[18] H.-H. Chao and L. Vandenberghe. Entropic proximal operators for nonnegative trigonometric polynomials. *IEEE Trans. Signal Process.*, 66(18):4826–4838, 2018.

[19] H. Chu, L. Liang, K.-C. Toh, and L. Yang. An efficient implementable inexact entropic proximal point algorithm for a class of linear programming problems. *arXiv:2011.14312*, 2020.

[20] P.L. Combettes and V.R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(4):1168–1200, 2005.

[21] R.-A. Dragomir, A. Taylor, A. d’Aspremont, and J. Bolte. Optimal complexity and certification of Bregman first-order methods. *To appear in Math. Program.*, 2021.

[22] M. Fukushima and H. Mine. A generalized proximal point algorithm for certain non-convex minimization problems. *Int. J. Syst. Sci.*, 12(8):989–1000, 1981.

[23] D.H. Gutman and J.F. Peña. Perturbed Fenchel duality and first-order methods. *To appear in Math. Program.*, 2022.

[24] F. Hanzely, P. Richtárik, and L. Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *Comput. Optim. Appl.*, 79(2):405–440, 2021.

[25] L.T.K. Hien, D.N. Phan, N. Gillis, M. Ahookhosh, and P. Patrinos. Block Bregman majorization minimization with extrapolation. *SIAM Journal on Mathematics of Data Science*, 4(1):1–25, 2022.

[26] K. Jiang, W. Si, C. Chen, and C. Bao. Efficient numerical methods for computing the stationary states of phase field crystal models. *SIAM J. Sci. Comput.*, 42(6):B1350–B1377, 2020.

[27] K. Jiang, D.F. Sun, and K.-C. Toh. An inexact accelerated proximal gradient method for large scale linearly constrained convex SDP. *SIAM J. Optim.*, 22(3):1042–1064, 2012.

[28] S. Kabbadj. Inexact version of Bregman proximal gradient algorithm. In *Abstract and Applied Analysis*, volume 2020, 2020.

[29] G. Lan, Z. Lu, and R.D.C. Monteiro. Primal-dual first-order methods with $O(1/\epsilon)$ iteration-complexity for cone programming. *Math. Program.*, 126(1):1–29, 2011.

[30] P.L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16(6):964–979, 1979.

[31] H. Lu, M.R. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim.*, 28(1):333–354, 2018.
[32] M.C. Mukkamala, P. Ochs, T. Pock, and S. Sabach. Convex-concave backtracking for inertial Bregman proximal gradient algorithms in nonconvex optimization. *SIAM J. Math. Data Sci.*, 2(3):658–682, 2020.

[33] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Math. Dokl.*, 27(2):372–376, 1983.

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

[35] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2003.

[36] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.

[37] Y. Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140(1):125–161, 2013.

[38] G. Peyré and M. Cuturi. Computational optimal transport. *Found. Trends Mach. Learn.*, 11(5-6):355–607, 2019.

[39] B.T. Polyak. *Introduction to Optimization*. Optimization Software Inc., New York, 1987.

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

[41] M. Romain and A. d’Aspremont. A Bregman method for structure learning on sparse directed acyclic graphs. *arXiv preprint arXiv:2011.02764*, 2020.

[42] M. Schmidt, N. Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

[43] F. Stonyakin, A. Tyurin, A. Gasnikov, P. Dvurechensky, A. Agafonov, D. Dvinskikh, M. Alkousa, D. Pasechnyuk, S. Artamonov, and V. Piskunova. Inexact model: A framework for optimization and variational inequalities. *To appear in Optimization Methods and Software*, 2021.

[44] M. Teboulle. A simplified view of first order methods for optimization. *Math. Program.*, 170(1):67–96, 2018.

[45] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *Technical report*, 2008.

[46] P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Math. Program.*, 125(2):263–295, 2010.

[47] Q. Van Nguyen. Forward-backward splitting with Bregman distances. *Vietnam J. Math.*, 45(3):519–539, 2017.

[48] S. Villa, S. Salzo, L. Baldassarre, and A. Verri. Accelerated and inexact forward-backward algorithms. *SIAM J. Optim.*, 23(3):1607–1633, 2013.
[49] L. Yang and K.-C. Toh. Bregman proximal point algorithm revisited: A new inexact version and its variant. arXiv preprint arXiv:2105.10370, 2021.

[50] Y. Zhou, Y. Liang, and L. Shen. A simple convergence analysis of Bregman proximal gradient algorithm. Comput. Optim. Appl., 73(3):903–912, 2019.