

# Bregman Proximal Point Algorithm Revisited: A New Inexact Version and its Inertial Variant

Lei Yang\*      Kim-Chuan Toh†

May 18, 2022

## Abstract

We study a general convex optimization problem, which covers various classic problems in different areas and particularly includes many optimal transport related problems arising in recent years. To solve this problem, we revisit the classic Bregman proximal point algorithm (BPPA) and introduce a new inexact stopping condition for solving the subproblems, which can circumvent the underlying feasibility difficulty often appearing in existing inexact conditions when the problem has a complex feasible set. Our inexact condition also covers several existing inexact conditions as special cases and hence makes our inexact BPPA (iBPPA) more flexible to fit different scenarios in practice. As an application to the standard optimal transport (OT) problem, our iBPPA with the entropic proximal term can bypass some numerical instability issues that usually plague the popular Sinkhorn’s algorithm in the OT community, since our iBPPA does not require the proximal parameter to be very small for obtaining an accurate approximate solution. The iteration complexity of  $O(1/k)$  and the convergence of the sequence are also established for our iBPPA under some mild conditions. Moreover, inspired by Nesterov’s acceleration technique, we develop an inertial variant of our iBPPA, denoted by V-iBPPA, and establish the iteration complexity of  $O(1/k^\lambda)$ , where  $\lambda \geq 1$  is a quadrangle scaling exponent of the kernel function. In particular, when the proximal parameter is a constant and the kernel function is strongly convex with Lipschitz continuous gradient (hence  $\lambda = 2$ ), our V-iBPPA achieves a faster rate of  $O(1/k^2)$  just as existing accelerated inexact proximal point algorithms. Some preliminary numerical experiments for solving the standard OT problem are conducted to show the convergence behaviors of our iBPPA and V-iBPPA under different inexactness settings. The experiments also empirically verify the potential of our V-iBPPA on improving the convergence speed.

**Keywords:** Proximal point algorithm; Bregman distance; inexact condition; Nesterov’s acceleration; optimal transport.

## 1 Introduction

We consider the following convex optimization problem

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \bar{\mathcal{C}}, \quad (1.1)$$

---

\*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China ([yanglei.math@gmail.com](mailto:yanglei.math@gmail.com)).

†Department of Mathematics, and Institute of Operations Research and Analytics, National University of Singapore ([matttohc@nus.edu.sg](mailto:matttohc@nus.edu.sg)). This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 Grant: R-146-000-336-114.

where  $f : \mathbb{E} \rightarrow (-\infty, \infty]$  is a proper closed convex function,  $\mathcal{C} \subseteq \mathbb{E}$  is a nonempty convex open set,  $\bar{\mathcal{C}}$  is the closure of  $\mathcal{C}$  and  $\mathbb{E}$  is a real finite dimensional Euclidean space equipped with an inner product  $\langle \cdot, \cdot \rangle$  and its induced norm  $\| \cdot \|$ . Problem (1.1) can cover a large class of convex optimization problems in various areas. We are particularly interested in optimization problems over the nonnegative octant arising in the area of optimal transport (OT); for example, the standard OT problem (see problem (4.1)), the capacity constrained OT problem and the problem of computing Wasserstein barycenters, just to name a few. All these problems have found many applications and attracted considerable attention in recent years. We refer interested readers to a recent comprehensive survey [42] (mainly from the computational point of view) and references therein for more details on OT and its related problems.

Among different solution methods designed for solving problem (1.1), the proximal point algorithm (PPA) is arguably the most fundamental one that basically generates a sequence  $\{\mathbf{x}^k\}$  via the following scheme

$$\mathbf{x}^{k+1} \approx \arg \min_{\mathbf{x}} \left\{ f(\mathbf{x}) + \gamma_k d(\mathbf{x}, \mathbf{x}^k) : \mathbf{x} \in \bar{\mathcal{C}} \right\}, \quad (1.2)$$

where  $d(\cdot, \cdot)$  is a certain proximity measure,  $\gamma_k > 0$  is a given proximal parameter and “ $\approx$ ” means that  $\mathbf{x}^{k+1}$  is only required to be an approximate solution of (1.2) in some sense (to be specified later) since computing an exact solution of (1.2) is in general too expensive. With the choice of  $d(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|^2$ , the above iterative scheme exactly reduces to the classical (inexact) PPA which enjoys comprehensive convergence results; see, for example, [20, 21, 34, 37, 45, 46]. Besides the half squared Euclidean distance, various researchers have also considered using some other non-Euclidean proximity measures in (1.2); see, for example, [2, 12, 13, 14, 17, 18, 19, 26, 27, 28, 54, 55]. Such an idea stems not only from natural algorithmic generalizations, but also from practical considerations on some specific applications. In particular, we find that, for OT and its related problems, a proper choice of  $d$  (specifically, the entropic proximal term) may capture the geometry/structure of the problem, which allows one to eliminate the constraint  $\mathbf{x} \in \bar{\mathcal{C}}$  during the PPA iterations and leads to a simpler subproblem (1.2). To date, there exist a variety of general proximity measures such as the Bregman distance [8] and the  $\psi$ -divergence [16]; see [3] for a comprehensive study on various proximity measures. In this paper, we will focus on the scheme (1.2) based on the Bregman distance, which has a long history of being incorporated in proximal-type methods and is still very popular nowadays (see, for example, [4, 7, 33, 56]), but some results developed later can also be extended to other proximity measures.

We now consider the scheme (1.2) with the choice of  $d(\mathbf{x}, \mathbf{y}) = \mathcal{D}_\phi(\mathbf{x}, \mathbf{y})$ , namely,

$$\mathbf{x}^{k+1} \approx \arg \min_{\mathbf{x}} \left\{ f(\mathbf{x}) + \gamma_k \mathcal{D}_\phi(\mathbf{x}, \mathbf{x}^k) \right\}, \quad (1.3)$$

where  $\mathcal{D}_\phi(\cdot, \cdot)$  is the Bregman distance associated with the kernel function  $\phi$  (see next section for definition). This scheme is a generic template for an inexact Bregman proximal point algorithm (iBPPA); see, for example, [3, 12, 13, 14, 17, 18, 28, 52]. In particular, we are interested in a class of kernel functions  $\phi$  satisfying certain desirable properties including  $\text{dom } \phi = \bar{\mathcal{C}}$  (see Assumption A(iii)) so that the sequence  $\{\mathbf{x}^k\}$  is forced to stay in  $\mathcal{C}$  and thus the constraint  $\mathbf{x} \in \bar{\mathcal{C}}$  is automatically eliminated in (1.3). But, even without such a constraint, the subproblem (1.3) is still generally nontrivial to solve. Therefore, for the algorithm to be implementable and practical, it must allow one to solve the subproblem approximately and the corresponding stopping condition must be practically verifiable. This consideration then gives rise to different

variants of the iBPPA. In the literature, a commonly used inexact framework is based on the  $\nu$ -subdifferential of  $f$  [9, 28, 55]. Basically, the iterate  $\mathbf{x}^{k+1}$  needs to satisfy

$$0 \in \partial_{\nu_k} f(\mathbf{x}^{k+1}) + \gamma_k (\nabla \phi(\mathbf{x}^{k+1}) - \nabla \phi(\mathbf{x}^k)), \quad (1.4)$$

which allows the approximate computation of the subdifferential of  $f$  at  $\mathbf{x}^{k+1}$ . Another widely used inexact framework was first considered by Rockafellar [46] for the classic PPA and later extended by Eckstein [18] for the BPPA. Specifically, it requires  $\mathbf{x}^{k+1}$  to satisfy

$$\Delta^k \in \partial f(\mathbf{x}^{k+1}) + \gamma_k (\nabla \phi(\mathbf{x}^{k+1}) - \nabla \phi(\mathbf{x}^k)) \quad \text{with} \quad \|\Delta^k\| \leq \eta_k, \quad (1.5)$$

which is typically easier to check than the  $\nu$ -subdifferential-based condition (1.4). However, we should be mindful that both conditions (1.4) and (1.5) implicitly require that the approximate solution  $\mathbf{x}^{k+1}$  must satisfy  $\mathbf{x}^{k+1} \in \text{dom } f$  (for the nonemptiness of  $\partial_{\nu_k} f(\mathbf{x}^{k+1})$  or  $\partial f(\mathbf{x}^{k+1})$ ) and  $\mathbf{x}^{k+1} \in \text{dom } \nabla \phi$  (for the well-definedness of  $\nabla \phi(\mathbf{x}^{k+1})$ ) at the same time. But in practice, such a requirement may be nontrivial to satisfy, especially when  $\text{dom } f$  is not a simple convex set. Thus checking whether condition (1.4) or (1.5) holds could be very expensive, if not impossible. In [52], Solodov and Svaiter proposed another inexact framework for the BPPA for which the stopping condition is more practical and constructive when  $\nabla \phi$  is explicitly invertible. Specifically, this approach requires a triple  $(\mathbf{x}^{k+1}, \mathbf{y}^k, \mathbf{v}^k)$  to satisfy

$$\begin{aligned} \mathbf{x}^{k+1} &= (\nabla \phi)^{-1}(\nabla \phi(\mathbf{x}^k) - \gamma_k^{-1} \mathbf{v}^k), \quad \mathbf{v}^k \in \partial f(\mathbf{y}^k), \\ \mathcal{D}_\phi(\mathbf{y}^k, \mathbf{x}^{k+1}) &\leq \sigma^2 \mathcal{D}_\phi(\mathbf{y}^k, \mathbf{x}^k), \end{aligned} \quad (1.6)$$

where  $\mathbf{y}^k$  is an intermediary point and  $\sigma \in [0, 1)$  is a constant. Note that it needs the exact computation of an element  $\mathbf{v}^k$  in  $\partial f(\mathbf{y}^k)$ , which sometimes could be difficult to satisfy when  $f$  is not a simple function. We should point out that when  $\phi$  is chosen as the classical half squared Euclidean norm, this exactness requirement has been relaxed by allowing an element in  $\partial_\nu f$  for some  $\nu \geq 0$  (see, e.g., [35, 49]), but it is not clear whether such an exactness requirement can be relaxed for a general kernel function.

The aforementioned feasibility difficulty of requiring  $\mathbf{x}^{k+1} \in \text{dom } f \cap \text{dom } \nabla \phi$  in (1.4) or (1.5) and the difficulty of computing an element of  $\partial f(\mathbf{y}^k)$  in (1.6) thus motivate us to propose a new inexact framework (see (3.2)), which relaxes the previous stringent requirements by allowing  $\partial_{\nu_k} f$  and  $\nabla \phi$  to be computed at two slightly different points, respectively. Though the idea is simple, it is surprising that it has not been explored before. Later in Section 4, we show by a concrete application to the standard OT problem that the verification of our inexact condition (3.2) is implementable and more practical. Moreover, our iBPPA with the entropic proximal term can bypass some numerical instability issues that often plague the popular Sinkhorn's algorithm used in the OT community. This is because in contrast to Sinkhorn's algorithm, our iBPPA does not require the proximal parameter to be very small in order to obtain an accurate approximate solution, as evident from our numerical results in Section 6.

Over the last few decades, Nesterov's series of seminal works [38, 39, 41] (see also [40]) on accelerated gradient methods have inspired various extensions and variants, such as the classical accelerated proximal point method of Güler [21] as well as its recent Bregman extension [61], the accelerated interior gradient algorithm of Auslender and Teboulle [3], and the recent inertial variants of the Bregman proximal gradient method in [23, 24]. Motivated by these studies, it is natural for us to explore whether and how our iBPPA can be accelerated. Here, we should point

out that the convergence rate (in terms of the objective function value) of PPA-type methods, including our iBPPA, can usually be improved by simply choosing smaller proximal parameters (see Remark 3.1). However, a smaller proximal parameter often leads to a harder and possibly more ill-conditioned subproblem, which may not be efficiently solvable as in the case of many OT related problems. Therefore, it is important to develop a possibly accelerated variant of our iBPPA without explicitly resorting to using smaller proximal parameters.

The contributions of this paper are summarized as follows.

1. We have proposed a new stopping condition for inexactly solving the subproblems in iBPPA. This condition can circumvent the difficulty of demanding the interior feasibility or requiring the exact computation of  $\partial f$  in existing inexact conditions. Moreover, it is flexible enough to fit different scenarios, and covers conditions (1.4) and (1.5) as special cases. The iteration complexity of  $O(1/k)$  and the convergence of the sequence are also established for our iBPPA under some mild conditions; see Section 3.
2. We have developed an inertial variant of our iBPPA, denoted by V-iBPPA, based on Nesterov's acceleration technique. By making use of the quadrangle scaling property of the Bregman distance (see Definition 5.1), we show that the V-iBPPA possesses an iteration complexity of  $O(1/k^\lambda)$  under a proper inexactness control, where  $\lambda \geq 1$  is a quadrangle scaling exponent; see Theorem 5.2. Moreover, when the proximal parameter is a constant and the kernel function is strongly convex with Lipschitz continuous gradient (hence  $\lambda = 2$ ), our V-iBPPA achieves a faster rate of  $O(1/k^2)$  just like the existing accelerated inexact proximal point algorithms in, for example, [21, 36].
3. We have also conducted numerical experiments to evaluate the performances of our iBPPA and V-iBPPA under different inexactness settings, in comparison to the inexact hybrid proximal extragradient methods of Solodov and Svaiter [49, 52]. The computational results empirically verify the improved performance of our V-iBPPA and demonstrate the promising potential of (V-)iBPPA for solving OT-related problems.

The rest of this paper is organized as follows. We present notation and preliminaries in Section 2. We then describe a new iBPPA for solving (1.1) and establish the convergence results in Section 3. A concrete application of our iBPPA to the standard OT problem is given in Section 4. We next develop an inertial variant of our iBPPA by employing Nesterov's acceleration technique in Section 5. Some preliminary numerical results are reported in Section 6, with some concluding remarks given in Section 7.

## 2 Notation and preliminaries

Assume that  $f : \mathbb{E} \rightarrow (-\infty, \infty]$  is a proper closed convex function. For a given  $\nu \geq 0$ , the  $\nu$ -subdifferential of  $f$  at  $\mathbf{x} \in \text{dom } f := \{\mathbf{x} \in \mathbb{E} : f(\mathbf{x}) < \infty\}$  is defined by  $\partial_\nu f(\mathbf{x}) := \{\mathbf{d} \in \mathbb{E} : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{d}, \mathbf{y} - \mathbf{x} \rangle - \nu, \forall \mathbf{y} \in \mathbb{E}\}$ , and when  $\nu = 0$ ,  $\partial_\nu f$  is simply denoted by  $\partial f$ . The conjugate function of  $f$  is the function  $f^* : \mathbb{E} \rightarrow (-\infty, \infty]$  defined by  $f^*(\mathbf{y}) := \sup \{\langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$ . A proper closed convex function  $f$  is essentially smooth if (i)  $\text{int dom } f$  is not empty; (ii)  $f$  is differentiable on  $\text{int dom } f$ ; (iii)  $\|\nabla f(\mathbf{x}_k)\| \rightarrow \infty$  for every sequence  $\{\mathbf{x}_k\}$  in  $\text{int dom } f$  converging to a boundary point of  $\text{int dom } f$ ; see [44, page 251].

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

Given a proper closed strictly convex function  $\phi : \mathbb{E} \rightarrow (-\infty, \infty]$ , finite at  $\mathbf{x}, \mathbf{y}$  and differentiable at  $\mathbf{y}$ , the Bregman distance [8] between  $\mathbf{x}$  and  $\mathbf{y}$  associated with the kernel function  $\phi$  is defined as

$$\mathcal{D}_\phi(\mathbf{x}, \mathbf{y}) := \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \nabla \phi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

It is easy to see that  $\mathcal{D}_\phi(\mathbf{x}, \mathbf{y}) \geq 0$  and equality holds if and only if  $\mathbf{x} = \mathbf{y}$  due to the strictly convexity of  $\phi$ . When  $\mathbb{E} = \mathbb{R}^n$  and  $\phi(\cdot) = \frac{1}{2}\|\cdot\|^2$ ,  $\mathcal{D}_\phi(\cdot, \cdot)$  recovers the half squared Euclidean distance. Moreover, one can easily verify the following identity.

**Lemma 2.1 (Four points identity).** *Suppose that a proper closed strictly convex function  $\phi : \mathbb{E} \rightarrow (-\infty, \infty]$  is finite at  $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$  and differentiable at  $\mathbf{a}, \mathbf{b}$ . Then,*

$$\langle \nabla \phi(\mathbf{a}) - \nabla \phi(\mathbf{b}), \mathbf{c} - \mathbf{d} \rangle = \mathcal{D}_\phi(\mathbf{c}, \mathbf{b}) + \mathcal{D}_\phi(\mathbf{d}, \mathbf{a}) - \mathcal{D}_\phi(\mathbf{c}, \mathbf{a}) - \mathcal{D}_\phi(\mathbf{d}, \mathbf{b}). \quad (2.1)$$

We next recall the definition of a Bregman function, which plays an important role in the convergence analysis of the Bregman-distance-based method.

**Definition 2.1 (Bregman function [11, Definition 2.1]).** *Let  $\mathcal{S} \subseteq \mathbb{E}$  be a nonempty open convex set with its closure denoted as  $\overline{\mathcal{S}}$ . We say that  $\phi : \overline{\mathcal{S}} \mapsto \mathbb{R}$  is a Bregman function with zone  $\mathcal{S}$  if the following conditions hold.*

- (B1)  $\phi$  is strictly convex and continuous on  $\overline{\mathcal{S}}$ .
- (B2)  $\phi$  is continuously differentiable on  $\mathcal{S}$ .
- (B3) The left partial level set  $\mathcal{L}(\mathbf{y}, \alpha) = \{\mathbf{x} \in \overline{\mathcal{S}} : \mathcal{D}_\phi(\mathbf{x}, \mathbf{y}) \leq \alpha\}$  is bounded for every  $\mathbf{y} \in \mathcal{S}$  and  $\alpha \in \mathbb{R}$ . Moreover, the right partial level set  $\mathcal{R}(\mathbf{x}, \alpha) = \{\mathbf{y} \in \mathcal{S} : \mathcal{D}_\phi(\mathbf{x}, \mathbf{y}) \leq \alpha\}$  is bounded for every  $\mathbf{x} \in \overline{\mathcal{S}}$  and  $\alpha \in \mathbb{R}$ .
- (B4) If  $\{\mathbf{y}^k\} \subseteq \mathcal{S}$  converges to some  $\mathbf{y}^* \in \overline{\mathcal{S}}$ , then  $\mathcal{D}_\phi(\mathbf{y}^*, \mathbf{y}^k) \rightarrow 0$ .
- (B5) (**Convergence consistency**) If  $\{\mathbf{x}^k\} \subseteq \overline{\mathcal{S}}$  and  $\{\mathbf{y}^k\} \subseteq \mathcal{S}$  are two sequences such that  $\{\mathbf{x}^k\}$  is bounded,  $\mathbf{y}^k \rightarrow \mathbf{y}^*$  and  $\mathcal{D}_\phi(\mathbf{x}^k, \mathbf{y}^k) \rightarrow 0$ , then  $\mathbf{x}^k \rightarrow \mathbf{y}^*$ .

Some remarks are in order concerning this definition. The above definition was originally introduced by Censor and Lent [11]. However, it has already been noticed (for example, by Eckstein [18, Section 2]) that the condition on the boundedness of the left partial level set in (B3) is redundant because it follows automatically from the observation that  $\mathcal{L}(\mathbf{y}, 0) = \{\mathbf{y}\}$  for all  $\mathbf{y} \in \mathcal{S}$ , the convexity of  $\mathcal{D}_\phi(\cdot, \mathbf{y})$  and [44, Corollary 8.7.1]. Moreover, Solodov and Svaiter have shown in [52, Theorem 2.4] that the convergence consistency (B5) also holds automatically as a consequence of the other conditions. But for ease of future reference, we still keep the left partial level-boundedness and (B5) in the definition. When  $\mathbb{E} = \mathbb{R}^n$ , two popular Bregman functions are  $\phi(\mathbf{x}) := \frac{1}{2}\|\mathbf{x}\|^2$  with zone  $\mathbb{R}^n$  and  $\phi(\mathbf{x}) := \sum_{i=1}^n x_i(\log x_i - 1)$  with zone  $\mathbb{R}_+^n$ . We refer the reader to [5, 10] and [13, Chapter 2] for more details and examples, as well as a brief historical review on Bregman functions.

Next, we give three supporting lemmas.

**Lemma 2.2** ([43, 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** ([30, Lemma 3.5]). *Suppose that  $\{\lambda_n\}_{n=0}^\infty \subseteq \mathbb{R}_+$  and  $\{\alpha_n\}_{n=0}^\infty \subseteq \mathbb{R}$  are two sequences. Let  $t_n := \sum_{k=0}^n \lambda_k$  and  $\beta_n := t_n^{-1} \sum_{k=0}^n \lambda_k \alpha_k$ . If  $t_n \rightarrow +\infty$ , then*

- (i)  $\liminf_{n \rightarrow +\infty} \alpha_n \leq \liminf_{n \rightarrow +\infty} \beta_n \leq \limsup_{n \rightarrow +\infty} \beta_n \leq \limsup_{n \rightarrow +\infty} \alpha_n$ ;
- (ii) *moreover, if  $\alpha := \lim_{n \rightarrow +\infty} \alpha_n$  exists, then  $\beta_n \rightarrow \alpha$ . (Silverman-Toeplitz theorem).*

**Lemma 2.4.** *Let  $g : \mathbb{E} \rightarrow (-\infty, \infty]$  be a proper closed convex function and  $\phi$  be a convex, essentially smooth function. For any  $\mathbf{y} \in \text{int dom } \phi$  and  $\varepsilon > 0$ , let  $P_{\mathbf{y}}^\varepsilon(\mathbf{x}) := g(\mathbf{x}) + \varepsilon \mathcal{D}_\phi(\mathbf{x}, \mathbf{y})$ . Suppose that an optimal solution (denoted by  $\mathcal{J}_\varepsilon \mathbf{y}$ ) of problem  $\min_{\mathbf{x}} \{P_{\mathbf{y}}^\varepsilon(\mathbf{x})\}$  exists. Then, we have*

$$P_{\mathbf{y}}^\varepsilon(\mathbf{u}) - P_{\mathbf{y}}^\varepsilon(\mathcal{J}_\varepsilon \mathbf{y}) \geq \varepsilon \mathcal{D}_\phi(\mathbf{u}, \mathcal{J}_\varepsilon \mathbf{y}), \quad \forall \mathbf{u} \in \text{dom } g \cap \text{dom } \phi. \quad (2.2)$$

Moreover, if  $g$  is an affine function, then the above inequality holds with equality.

*Proof.* Since  $\phi$  is essentially smooth, then  $\mathcal{J}_\varepsilon \mathbf{y}$  must lie in  $\text{int dom } \phi$  and satisfy

$$0 \in \partial g(\mathcal{J}_\varepsilon \mathbf{y}) + \varepsilon (\nabla \phi(\mathcal{J}_\varepsilon \mathbf{y}) - \nabla \phi(\mathbf{y})) \iff -\varepsilon (\nabla \phi(\mathcal{J}_\varepsilon \mathbf{y}) - \nabla \phi(\mathbf{y})) \in \partial g(\mathcal{J}_\varepsilon \mathbf{y}).$$

From the convexity of  $g$ , for any  $\mathbf{u} \in \text{dom } g \cap \text{dom } \phi$ ,

$$\begin{aligned} g(\mathbf{u}) &\geq g(\mathcal{J}_\varepsilon \mathbf{y}) - \varepsilon \langle \nabla \phi(\mathcal{J}_\varepsilon \mathbf{y}) - \nabla \phi(\mathbf{y}), \mathbf{u} - \mathcal{J}_\varepsilon \mathbf{y} \rangle \\ &= g(\mathcal{J}_\varepsilon \mathbf{y}) - \varepsilon (\mathcal{D}_\phi(\mathbf{u}, \mathbf{y}) - \mathcal{D}_\phi(\mathbf{u}, \mathcal{J}_\varepsilon \mathbf{y}) - \mathcal{D}_\phi(\mathcal{J}_\varepsilon \mathbf{y}, \mathbf{y})), \end{aligned}$$

where the equality follows from (2.1). Then, rearranging the above inequality results in (2.2). Moreover, when  $g$  is an affine function, it is easy to see that  $g(\mathbf{u}) = g(\mathcal{J}_\varepsilon \mathbf{y}) + \langle \nabla g(\mathcal{J}_\varepsilon \mathbf{y}), \mathbf{u} - \mathcal{J}_\varepsilon \mathbf{y} \rangle$  for any  $\mathbf{u} \in \text{dom } \phi$ . This together with arguments similar to those just presented above implies the equality in (2.2). We completed the proof.  $\square$

Finally, we make some blanket assumptions on our problem (1.1) and the kernel function  $\phi$ , which are essential for guaranteeing the well-definedness of our problem and subproblems as well as the convergence of the presented algorithms.

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

- (i)  $\text{dom } f \cap \mathcal{C}$  is nonempty.
- (ii)  $\rho := \max \{\|\mathbf{x} - \mathbf{y}\| : \mathbf{x}, \mathbf{y} \in \text{dom } f \cap \bar{\mathcal{C}}\} < \infty$ .
- (iii)  $\text{dom } \phi = \bar{\mathcal{C}}$ ,  $\phi$  is a Bregman function with zone  $\mathcal{C}$  and  $\phi$  is essentially smooth.

One can see from Assumption A(i)&(ii) that  $\text{dom}(f + \delta_{\bar{\mathcal{C}}})$  is nonempty and  $f + \delta_{\bar{\mathcal{C}}}$  is level-bounded. Hence, a solution of problem (1.1) exists; see, for example, [47, Theorem 1.9]. Note also that Assumption A(ii) actually requires the feasible set of problem (1.1) to be bounded. This property then ensures the existence of a solution of each subproblem and the boundedness of sequence generated by our algorithm. Some weaker assumptions are possible, but involve a bit more analysis when we deal with the convergence of the iBPPA; see Remark 3.2. Here, we simply impose Assumption A(ii). This assumption can be satisfied by many practical problems, for example, the standard OT problem (4.1) and its various related problems [42].

### 3 A new inexact Bregman proximal point algorithm

In this section, we develop a new inexact Bregman proximal point algorithm (iBPPA) for solving problem (1.1). The complete framework is presented as Algorithm 1.

---

**Algorithm 1** An inexact Bregman proximal point algorithm (iBPPA) for (1.1)

---

**Input:** Let  $\{\gamma_k\}_{k=0}^\infty$ ,  $\{\eta_k\}_{k=0}^\infty$ ,  $\{\mu_k\}_{k=0}^\infty$ ,  $\{\nu_k\}_{k=0}^\infty$  be four sequences of nonnegative scalars. Choose  $\mathbf{x}^0 = \tilde{\mathbf{x}}^0 \in \mathcal{C}$  arbitrarily and a kernel function  $\phi$ . Set  $k = 0$ .

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

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

$$\min_{\mathbf{x}} f(\mathbf{x}) + \gamma_k \mathcal{D}_\phi(\mathbf{x}, \mathbf{x}^k), \quad (3.1)$$

such that  $\mathbf{x}^{k+1} \in \mathcal{C}$ ,  $\tilde{\mathbf{x}}^{k+1} \in \text{dom } f \cap \bar{\mathcal{C}}$  and

$$\begin{aligned} \Delta^k &\in \partial_{\nu_k} f(\tilde{\mathbf{x}}^{k+1}) + \gamma_k (\nabla \phi(\mathbf{x}^{k+1}) - \nabla \phi(\mathbf{x}^k)) \\ &\text{with } \|\Delta^k\| \leq \eta_k, \quad \mathcal{D}_\phi(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^{k+1}) \leq \mu_k. \end{aligned} \quad (3.2)$$

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

**end while**

**Output:**  $(\mathbf{x}^k, \tilde{\mathbf{x}}^k)$

---

In the spirit of the PPA-type method, our iBPPA in Algorithm 1 basically solves the original problem (1.1) via approximately solving a sequence of subproblems (3.1) each involving a Bregman proximal term associated with the kernel function  $\phi$ . Since  $\text{dom } \phi = \bar{\mathcal{C}}$  by Assumption A(iii), the constraint  $\mathbf{x} \in \bar{\mathcal{C}}$  can be removed in (3.1). Moreover, under Assumption A, one can see that, at the  $k$ -th iteration, the solution  $\mathbf{x}^{k,*}$  of subproblem (3.1) exists and lies in  $\mathcal{C}$  ( $= \text{dom } \phi$ ). Indeed, Assumption A(ii) and  $\text{dom } \phi = \bar{\mathcal{C}}$  imply that the objective function in subproblem (3.1) is level-bounded. Thus, a solution exists (by [47, Theorem 1.9]) and must be unique since  $\phi$  is strictly convex (by Assumption A(iii) and condition (B1)). The essential smoothness of  $\phi$  (by Assumption A(iii)) and Assumption A(i) further imply that  $\mathbf{x}^{k,*}$  cannot be at the boundary of  $\mathcal{C}$ . Hence, the subproblem and iterate are well-defined. Our inexact condition (3.2) always holds at  $\mathbf{x}^{k+1} = \tilde{\mathbf{x}}^{k+1} = \mathbf{x}^{k,*}$  and thus it is achievable.

The inexact condition (3.2) is rather broad for covering some existing approximation conditions. When  $\nu_k \equiv \eta_k \equiv \mu_k \equiv 0$ ,  $\mathbf{x}^{k+1}$  ( $= \tilde{\mathbf{x}}^{k+1}$ ) is obviously the exact optimal solution of subproblem (3.1). In this case, our iBPPA reduces to the classical exact BPPA [12, 13, 14, 17]. When  $\eta_k \equiv \mu_k \equiv 0$ , condition (3.2) reduces to condition (1.4) studied in [9, 28, 55]. Moreover, when  $\nu_k \equiv \mu_k \equiv 0$ , condition (3.2) reduces to condition (1.5) studied by Eckstein in [18]. More importantly, the inexact condition (3.2) can bypass the underlying difficulty of demanding interior feasibility, which appears to be often overlooked in the literature.

As we have mentioned in the introduction, to check either condition (1.4) or (1.5), one has to compute an approximate solution  $\mathbf{x}^{k+1}$  that belongs to both  $\text{dom } f$  (for the nonemptiness of  $\partial_{\nu_k} f(\mathbf{x}^{k+1})$  or  $\partial f(\mathbf{x}^{k+1})$ ) and  $\text{dom } \nabla \phi$  (for the well-definedness of  $\nabla \phi(\mathbf{x}^{k+1})$ ). However, in practice, even finding a point in  $\text{dom } f \cap \text{dom } \nabla \phi$  can be nontrivial when  $\text{dom } f$  is not a simple convex set. Thus, in this case, condition (1.4) or (1.5) may no longer be suitable. Our

inexact condition (3.2) allows one to evaluate  $\partial_{\nu_k} f$  and  $\nabla\phi$  at two different points to deal with  $\text{dom } f$  and  $\text{dom } \nabla\phi$  separately. It is also interesting to compare our condition with condition (1.6). Both conditions allow the error tolerance criteria to be checked at two different points. But the mechanisms are different. Our condition (3.2) aims to relax the stringent requirement  $\mathbf{x}^{k+1} \in \text{dom } f \cap \text{dom } \nabla\phi$ , while condition (1.6) inherits the idea of a hybrid approach developed by Solodov and Svaiter [49, 50, 51, 53] (now known as the hybrid proximal extragradient (HPE) method [35, 36]) to use an intermediary point for computing  $\mathbf{x}^{k+1}$ . The latter condition is constructive and does not need the usual summable-error requirement. However, it generally needs the exact computation of an element in  $\partial f$  at an intermediary point. Note that, when  $\phi$  is chosen as the classical squared Euclidean norm, this exactness requirement has been relaxed by allowing an element in  $\partial_{\nu} f$  for some  $\nu \geq 0$  (see, e.g., [35, 49]), but, to our knowledge, it is still not clear whether such a requirement can be relaxed for a general kernel function. This exactness requirement may limit the application of condition (1.6). Moreover, when employing condition (1.6), one has to compute  $\mathbf{x}^{k+1}$  via an extragradient step to guarantee the convergence. In contrast, our condition (3.2) appears to be more straightforward and flexible. Later, we shall illustrate the potential advantages of our condition through a concrete example on the standard OT problem in Section 4.

We next establish the convergence of our iBPPA in Algorithm 1. Our analysis is inspired by several existing works (see, for example, [18, 55]). We start by establishing a sufficient-descent-like property in the following lemma.

**Lemma 3.1 (Sufficient-descent-like property).** *Let  $\{\mathbf{x}^k\}$  and  $\{\tilde{\mathbf{x}}^k\}$  be the sequences generated by the iBPPA in Algorithm 1. Then, for any  $\mathbf{u} \in \text{dom } f \cap \bar{\mathcal{C}}$ ,*

$$\begin{aligned} f(\tilde{\mathbf{x}}^{k+1}) &\leq f(\mathbf{u}) + \gamma_k (\mathcal{D}_\phi(\mathbf{u}, \mathbf{x}^k) - \mathcal{D}_\phi(\mathbf{u}, \mathbf{x}^{k+1}) - \mathcal{D}_\phi(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^k)) \\ &\quad + \langle \Delta^k, \tilde{\mathbf{x}}^{k+1} - \mathbf{u} \rangle + \gamma_k \mu_k + \nu_k. \end{aligned} \quad (3.3)$$

*Proof.* From condition (3.2), there exists a  $\mathbf{d}^{k+1} \in \partial_{\nu_k} f(\tilde{\mathbf{x}}^{k+1})$  such that  $\Delta^k = \mathbf{d}^{k+1} + \gamma_k (\nabla\phi(\mathbf{x}^{k+1}) - \nabla\phi(\mathbf{x}^k))$ . Then, for any  $\mathbf{u} \in \text{dom } f \cap \bar{\mathcal{C}}$ , we see that

$$\begin{aligned} f(\mathbf{u}) &\geq f(\tilde{\mathbf{x}}^{k+1}) + \langle \mathbf{d}^{k+1}, \mathbf{u} - \tilde{\mathbf{x}}^{k+1} \rangle - \nu_k \\ &= f(\tilde{\mathbf{x}}^{k+1}) + \langle \Delta^k - \gamma_k (\nabla\phi(\mathbf{x}^{k+1}) - \nabla\phi(\mathbf{x}^k)), \mathbf{u} - \tilde{\mathbf{x}}^{k+1} \rangle - \nu_k, \end{aligned}$$

which implies that

$$f(\tilde{\mathbf{x}}^{k+1}) \leq f(\mathbf{u}) + \gamma_k \langle \nabla\phi(\mathbf{x}^{k+1}) - \nabla\phi(\mathbf{x}^k), \mathbf{u} - \tilde{\mathbf{x}}^{k+1} \rangle + \langle \Delta^k, \tilde{\mathbf{x}}^{k+1} - \mathbf{u} \rangle + \nu_k.$$

Note from the four points identity (2.1) and  $\mathcal{D}_\phi(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^{k+1}) \leq \mu_k$  in (3.2) that

$$\langle \nabla\phi(\mathbf{x}^{k+1}) - \nabla\phi(\mathbf{x}^k), \mathbf{u} - \tilde{\mathbf{x}}^{k+1} \rangle \leq \mathcal{D}_\phi(\mathbf{u}, \mathbf{x}^k) - \mathcal{D}_\phi(\mathbf{u}, \mathbf{x}^{k+1}) - \mathcal{D}_\phi(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^k) + \mu_k.$$

Combining the above two inequalities, we obtain (3.3).  $\square$

Based on the sufficient-descent-like property, we can estimate the iteration complexity of our iBPPA in terms of the function value as follows.

**Theorem 3.1 (Iteration complexity of the iBPPA).** *Let  $\{\mathbf{x}^k\}$  and  $\{\tilde{\mathbf{x}}^k\}$  be the sequences generated by the iBPPA in Algorithm 1. Then, for any optimal solution  $\mathbf{x}^*$  of problem (1.1), we have*

$$f(\tilde{\mathbf{x}}^N) - f(\mathbf{x}^*) \leq \sigma_{N-1}^{-1} \left( \mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^0) + \sum_{k=0}^{N-1} \mu_k + \sum_{k=0}^{N-1} \gamma_k^{-1} (\rho\eta_k + \nu_k) + \sum_{k=0}^{N-1} \sigma_{k-1} \xi_k \right), \quad (3.4)$$

where  $\sigma_{-1} := 0$ ,  $\sigma_k := \sum_{i=0}^k \gamma_i^{-1}$  and  $\xi_k := f(\tilde{\mathbf{x}}^{k+1}) - f(\tilde{\mathbf{x}}^k) \leq \gamma_k(\mu_{k-1} + \mu_k) + \rho\eta_k + \nu_k$  for every integer  $k \geq 0$ . Moreover, if the summable-error condition that  $\max\{\sum \mu_k, \sum \gamma_k^{-1}\nu_k, \sum \gamma_k^{-1}\eta_k, \sum \sigma_{k-1}\xi_k\} < \infty$  holds, then

$$f(\tilde{\mathbf{x}}^N) - f(\mathbf{x}^*) \leq O\left(\frac{1}{\sum_{k=0}^{N-1} \gamma_k^{-1}}\right).$$

*Proof.* First, we see from (3.3) in Lemma 3.1 with  $\mathbf{u} = \tilde{\mathbf{x}}^k$  that

$$\begin{aligned} \xi_k &:= f(\tilde{\mathbf{x}}^{k+1}) - f(\tilde{\mathbf{x}}^k) \\ &\leq \gamma_k(\mathcal{D}_\phi(\tilde{\mathbf{x}}^k, \mathbf{x}^k) - \mathcal{D}_\phi(\tilde{\mathbf{x}}^k, \mathbf{x}^{k+1}) - \mathcal{D}_\phi(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^k)) + \langle \Delta^k, \tilde{\mathbf{x}}^{k+1} - \tilde{\mathbf{x}}^k \rangle + \gamma_k\mu_k + \nu_k \\ &\leq \gamma_k\mathcal{D}_\phi(\tilde{\mathbf{x}}^k, \mathbf{x}^k) + |\langle \Delta^k, \tilde{\mathbf{x}}^{k+1} - \tilde{\mathbf{x}}^k \rangle| + \gamma_k\mu_k + \nu_k \\ &\leq \gamma_k(\mu_{k-1} + \mu_k) + \rho\eta_k + \nu_k, \end{aligned} \quad (3.5)$$

where the last inequality follows from condition (3.2) and  $\|\tilde{\mathbf{x}}^{k+1} - \tilde{\mathbf{x}}^k\| \leq \rho$  (due to  $\tilde{\mathbf{x}}^{k+1}, \tilde{\mathbf{x}}^k \in \text{dom}f \cap \bar{\mathcal{C}}$  and Assumption A(ii)). Moreover, for any  $k \geq 0$ ,

$$\begin{aligned} f(\tilde{\mathbf{x}}^{k+1}) = f(\tilde{\mathbf{x}}^k) + \xi_k &\implies (\sigma_k - \gamma_k^{-1})f(\tilde{\mathbf{x}}^{k+1}) = \sigma_{k-1}f(\tilde{\mathbf{x}}^k) + \sigma_{k-1}\xi_k \\ &\implies \gamma_k^{-1}f(\tilde{\mathbf{x}}^{k+1}) = \sigma_k f(\tilde{\mathbf{x}}^{k+1}) - \sigma_{k-1}f(\tilde{\mathbf{x}}^k) - \sigma_{k-1}\xi_k. \end{aligned}$$

Summing the above equality from  $k = 0$  to  $k = N - 1$  results in

$$\sum_{k=0}^{N-1} \gamma_k^{-1} f(\tilde{\mathbf{x}}^{k+1}) = \sigma_{N-1} f(\tilde{\mathbf{x}}^N) - \sum_{k=0}^{N-1} \sigma_{k-1} \xi_k. \quad (3.6)$$

Let  $\mathbf{x}^*$  be an arbitrary optimal solution of problem (1.1). Then, using (3.3) with  $\mathbf{u} = \mathbf{x}^*$  again, we see that, for all  $k \geq 0$ ,

$$\begin{aligned} f(\tilde{\mathbf{x}}^{k+1}) - f(\mathbf{x}^*) &\leq \gamma_k(\mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^k) - \mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^{k+1}) - \mathcal{D}_\phi(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^k)) + \langle \Delta^k, \tilde{\mathbf{x}}^{k+1} - \mathbf{x}^* \rangle + \gamma_k\mu_k + \nu_k \\ &\leq \gamma_k(\mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^k) - \mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^{k+1})) + \rho\eta_k + \gamma_k\mu_k + \nu_k, \end{aligned}$$

where the last inequality follows from  $\Delta^k \leq \eta_k$  and  $\|\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^*\| \leq \rho$ . Thus, we get

$$\gamma_k^{-1} f(\tilde{\mathbf{x}}^{k+1}) - \gamma_k^{-1} f(\mathbf{x}^*) \leq \mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^k) - \mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^{k+1}) + \mu_k + \gamma_k^{-1}(\rho\eta_k + \nu_k).$$

Summing the above inequality from  $k = 0$  to  $k = N - 1$ , we obtain that

$$\begin{aligned} &\sum_{k=0}^{N-1} \gamma_k^{-1} f(\tilde{\mathbf{x}}^{k+1}) - \sigma_{N-1} f(\mathbf{x}^*) \\ &\leq \mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^0) - \mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^N) + \sum_{k=0}^{N-1} \mu_k + \sum_{k=0}^{N-1} \gamma_k^{-1}(\rho\eta_k + \nu_k) \\ &\leq \mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^0) + \sum_{k=0}^{N-1} \mu_k + \sum_{k=0}^{N-1} \gamma_k^{-1}(\rho\eta_k + \nu_k). \end{aligned} \quad (3.7)$$

This together with (3.6) implies that

$$\begin{aligned} \sigma_{N-1}(f(\tilde{\mathbf{x}}^N) - f(\mathbf{x}^*)) &= \sum_{k=0}^{N-1} \gamma_k^{-1} f(\tilde{\mathbf{x}}^{k+1}) + \sum_{k=0}^{N-1} \sigma_{k-1} \xi_k - \sigma_{N-1} f(\mathbf{x}^*) \\ &\leq \mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^0) + \sum_{k=0}^{N-1} \mu_k + \sum_{k=0}^{N-1} \gamma_k^{-1}(\rho\eta_k + \nu_k) + \sum_{k=0}^{N-1} \sigma_{k-1} \xi_k. \end{aligned}$$

Dividing the above inequality by  $\sigma_{N-1}$ , we can obtain (3.4). The remaining result readily follows from (3.4) under given conditions. We then complete the proof.  $\square$

**Remark 3.1 (Comments on iteration complexity).** We see from Theorem 3.1 that, under the summable-error condition, the convergence rate of  $\{f(\tilde{\mathbf{x}}^k)\}$  is mainly determined by  $(\sum \gamma_k^{-1})^{-1}$ . Since the choice of  $\{\gamma_k\}$  can be quite flexible, one can obtain different convergence rates of  $\{f(\tilde{\mathbf{x}}^k)\}$ . For example,

- if  $0 < \underline{\gamma} \leq \gamma_k \leq \bar{\gamma} < \infty$ , then  $f(\tilde{\mathbf{x}}^N) - f(\mathbf{x}^*) \leq O(\frac{1}{N})$ ;
- if  $\gamma_k = \frac{1}{1+k}$ , then  $f(\tilde{\mathbf{x}}^N) - f(\mathbf{x}^*) \leq O(\frac{1}{N^2})$ ;
- if  $\gamma_k = \gamma_0 \varrho^k$  with  $0 < \varrho < 1$ , then  $f(\tilde{\mathbf{x}}^N) - f(\mathbf{x}^*) \leq O(\varrho^N)$ .

Indeed, it is not hard to see that an arbitrarily fast convergence rate can be achieved with a proper decreasing sequence of  $\{\gamma_k\}$ . However, for a fast decreasing sequence of  $\{\gamma_k\}$ , the corresponding choices of  $\{\mu_k\}$ ,  $\{\nu_k\}$  and  $\{\eta_k\}$  also become more stringent to guarantee the summable-error conditions. Thus, when applying the iBPPA for solving a specific problem, one needs to make a tradeoff between the convergence rate and the tolerable inexactness. In addition, we should mention that condition  $\sum \sigma_{k-1} \xi_k < \infty$  is not as restrictive as it appears. For example, consider the case  $0 < \underline{\gamma} \leq \gamma_k \leq \bar{\gamma} < +\infty$  and for some  $p > 1$ ,  $\mu_k \leq O(k^{-p})$ ,  $\nu_k \leq O(k^{-p})$ ,  $\eta_k \leq O(k^{-p})$ . Then it follows from (3.5) that  $\xi_k \leq \bar{\gamma}(\mu_{k-1} + \mu_k) + \rho\eta_k + \nu_k \leq O(k^{-p})$ . This together with  $\sigma_k := \sum_{i=0}^k \gamma_i^{-1} \leq (k+1)\underline{\gamma}^{-1}$  implies that  $\sum \sigma_{k-1} \xi_k \leq O(\sum k^{1-p})$ . Hence, condition  $\sum \sigma_{k-1} \xi_k < \infty$  holds whenever  $p > 2$ . Moreover, if the function values decrease monotonically along the sequence  $\{\tilde{\mathbf{x}}^k\}$ , as we often observe in our experiments, then  $\xi_k := f(\tilde{\mathbf{x}}^{k+1}) - f(\tilde{\mathbf{x}}^k) \leq 0$  and the condition  $\sum \sigma_{k-1} \xi_k < \infty$  is automatically met.

We next present the main convergence results for our iBPPA.

**Theorem 3.2 (Convergence of the iBPPA).** Suppose that Assumption A holds. Let  $\{\mathbf{x}^k\}$  and  $\{\tilde{\mathbf{x}}^k\}$  be the sequences generated by the iBPPA in Algorithm 1, and  $f^* := \min \{f(\mathbf{x}) : \mathbf{x} \in \bar{\mathcal{C}}\}$ . Then, the following statements hold.

- If  $\sup_k \{\gamma_k\} < \infty$ ,  $\sum \mu_k < \infty$ ,  $\sum \nu_k < \infty$  and  $\sum \eta_k < \infty$ , then  $f(\tilde{\mathbf{x}}^k) \rightarrow f^*$ .
- If  $\sup_k \{\gamma_k\} < \infty$ ,  $\sum \mu_k < \infty$ ,  $\sum \gamma_k^{-1} \nu_k < \infty$  and  $\sum \gamma_k^{-1} \eta_k < \infty$ , then the sequences  $\{\mathbf{x}^k\}$  and  $\{\tilde{\mathbf{x}}^k\}$  converge to the same limit that is an optimal solution of problem (1.1).

*Proof. Statement (i).* Let  $\mathbf{x}^*$  be an arbitrary optimal solution of problem (1.1). Then, from (3.7), we have for any nonnegative integer  $n$ ,

$$\begin{aligned} & \sigma_n^{-1} \sum_{k=0}^n \gamma_k^{-1} f(\tilde{\mathbf{x}}^{k+1}) \\ & \leq f(\mathbf{x}^*) + \sigma_n^{-1} \mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^0) + \sigma_n^{-1} \sum_{k=0}^n \mu_k + \sigma_n^{-1} \sum_{k=0}^n \gamma_k^{-1} (\rho\eta_k + \nu_k). \end{aligned} \quad (3.8)$$

where  $\sigma_n := \sum_{k=0}^n \gamma_k^{-1}$  for  $n = 0, 1, 2, \dots$ . Note that  $\sigma_n \rightarrow +\infty$  since  $\sup_k \{\gamma_k\} < +\infty$ , and  $\rho\eta_k + \nu_k \rightarrow 0$  since  $\sum \nu_k < \infty$  and  $\sum \eta_k < \infty$ . Thus, from Lemma 2.3(ii), we see that  $\sigma_n^{-1} \sum_{k=0}^n \gamma_k^{-1} (\rho\eta_k + \nu_k) \rightarrow 0$ . This together with (3.8),  $\sum \mu_k < \infty$  and Lemma 2.3(i) implies that

$$\liminf_{n \rightarrow \infty} f(\tilde{\mathbf{x}}^{n+1}) \leq \liminf_{n \rightarrow \infty} \sigma_n^{-1} \sum_{k=0}^n \gamma_k^{-1} f(\tilde{\mathbf{x}}^{k+1}) \leq f(\mathbf{x}^*).$$

Note also that  $f(\tilde{\mathbf{x}}^{n+1}) \geq f(\mathbf{x}^*)$  for all  $n$  since  $\tilde{\mathbf{x}}^{n+1} \in \text{dom } f \cap \bar{\mathcal{C}}$ . Then, we have that  $\liminf_{n \rightarrow \infty} f(\tilde{\mathbf{x}}^{n+1}) = f(\mathbf{x}^*)$ . On the other hand,  $\{f(\tilde{\mathbf{x}}^k)\}$  is bounded from below since  $\tilde{\mathbf{x}}^k \in$

$\text{dom } f \cap \bar{\mathcal{C}}$  for all  $k$  and the solution set of problem (1.1) is nonempty (by Assumption A(i)&(ii)). Finally, from (3.5) and Lemma 2.2, together with  $\sup_k \{\gamma_k\} < \infty$  and the summability of  $\{\mu_k\}$ ,  $\{\nu_k\}$ ,  $\{\eta_k\}$ , we see that  $\{f(\tilde{\mathbf{x}}^k)\}$  is convergent and hence  $f(\tilde{\mathbf{x}}^k) \rightarrow f(\mathbf{x}^*)$ . This proves statement (i).

*Statement (ii).* First, since  $\sup_k \{\gamma_k\} < \infty$ , then  $\inf_k \{\gamma_k^{-1}\} > 0$ . This together with  $\sum \gamma_k^{-1} \nu_k < \infty$  and  $\sum \gamma_k^{-1} \eta_k < \infty$  implies that  $\sum \nu_k < \infty$  and  $\sum \eta_k < \infty$ . Thus, statement (i) holds. Since  $\{\tilde{\mathbf{x}}^k\}$  is bounded (due to  $\tilde{\mathbf{x}}^k \in \text{dom } f \cap \bar{\mathcal{C}}$  and Assumption A(ii)), it has at least one cluster point. Suppose that  $\tilde{\mathbf{x}}^\infty$  is a cluster point and  $\{\tilde{\mathbf{x}}^{k_i}\}$  is a convergent subsequence such that  $\lim_{i \rightarrow \infty} \tilde{\mathbf{x}}^{k_i} = \tilde{\mathbf{x}}^\infty$ . Then, from the closedness of  $f$ , we have that  $f(\tilde{\mathbf{x}}^\infty) \leq \liminf_{i \rightarrow \infty} f(\tilde{\mathbf{x}}^{k_i}) = f^*$ . Note that  $\tilde{\mathbf{x}}^\infty \in \text{dom } f \cap \bar{\mathcal{C}}$  since  $\text{dom } f \cap \bar{\mathcal{C}}$  is closed. Hence,  $\tilde{\mathbf{x}}^\infty$  must be an optimal solution of (1.1).

Next, let  $\mathbf{x}^*$  be an arbitrary optimal solution of (1.1). Obviously,  $f(\mathbf{x}^*) \leq f(\tilde{\mathbf{x}}^{k+1})$  for all  $k \geq 0$  since  $\tilde{\mathbf{x}}^{k+1} \in \text{dom } f \cap \bar{\mathcal{C}}$ . By setting  $\mathbf{u} = \mathbf{x}^*$  in (3.3) and recalling  $\|\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^*\| \leq \rho$  (by Assumption A(ii)), we see that

$$\begin{aligned} 0 &\leq \mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^{k+1}) \\ &\leq \mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^k) + \gamma_k^{-1}(f(\mathbf{x}^*) - f(\tilde{\mathbf{x}}^{k+1})) - \mathcal{D}_\phi(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^k) + \mu_k + \gamma_k^{-1}(\rho\eta_k + \nu_k) \\ &\leq \mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^k) + \mu_k + \gamma_k^{-1}(\rho\eta_k + \nu_k). \end{aligned} \quad (3.9)$$

Thus, we can conclude from (3.9),  $\max\{\sum \mu_k, \sum \gamma_k^{-1} \nu_k, \sum \gamma_k^{-1} \eta_k\} < \infty$  and Lemma 2.2 that  $\{\mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^k)\}$  is convergent. From this fact and condition (B3) in Definition 2.1, we further see that  $\{\mathbf{x}^k\}$  is bounded and hence it has at least one cluster point. Suppose that  $\mathbf{x}^\infty$  is a cluster point and  $\{\mathbf{x}^{k_j}\}$  is a convergent subsequence such that  $\lim_{j \rightarrow \infty} \mathbf{x}^{k_j} = \mathbf{x}^\infty$ . Then, from the fact that  $\mathcal{D}_\phi(\tilde{\mathbf{x}}^{k_j}, \mathbf{x}^{k_j}) \leq \mu_{k_j-1} \rightarrow 0$ , the boundedness of  $\{\tilde{\mathbf{x}}^{k_j}\}$  and the convergence consistency of  $\phi$  (see condition (B5) in Definition 2.1), we have that  $\lim_{j \rightarrow \infty} \tilde{\mathbf{x}}^{k_j} = \mathbf{x}^\infty$ . Therefore, from what we have proved in the last paragraph,  $\mathbf{x}^\infty$  is an optimal solution of (1.1). Moreover, by using (3.9) with  $\mathbf{x}^*$  replaced by  $\mathbf{x}^\infty$ , we can conclude that  $\{\mathcal{D}_\phi(\mathbf{x}^\infty, \mathbf{x}^k)\}$  is convergent. On the other hand, it follows from  $\lim_{j \rightarrow \infty} \mathbf{x}^{k_j} = \mathbf{x}^\infty$  and condition (B4) of the Bregman function that  $\mathcal{D}_\phi(\mathbf{x}^\infty, \mathbf{x}^{k_j}) \rightarrow 0$ . Consequently,  $\{\mathcal{D}_\phi(\mathbf{x}^\infty, \mathbf{x}^k)\}$  must converge to zero. Now, let  $\hat{\mathbf{x}}^\infty$  be any cluster point of  $\{\mathbf{x}^k\}$  with a subsequence  $\{\mathbf{x}^{k'_j}\}$  such that  $\mathbf{x}^{k'_j} \rightarrow \hat{\mathbf{x}}^\infty$ . Since  $\mathcal{D}_\phi(\mathbf{x}^\infty, \mathbf{x}^k) \rightarrow 0$ , we have  $\mathcal{D}_\phi(\mathbf{x}^\infty, \mathbf{x}^{k'_j}) \rightarrow 0$ . Using the convergence consistency of  $\phi$  again, we see that  $\mathbf{x}^\infty = \hat{\mathbf{x}}^\infty$ . Since  $\hat{\mathbf{x}}^\infty$  is arbitrary, we can conclude that  $\lim_{k \rightarrow \infty} \mathbf{x}^k = \mathbf{x}^\infty$ . This, together with the boundedness of  $\{\tilde{\mathbf{x}}^k\}$ ,  $\mathcal{D}_\phi(\tilde{\mathbf{x}}^k, \mathbf{x}^k) \rightarrow 0$  and the convergence consistency of  $\phi$ , implies that  $\{\tilde{\mathbf{x}}^k\}$  also converges to  $\mathbf{x}^\infty$ . This completes the proof.  $\square$

**Remark 3.2 (Comments on the boundedness of  $\text{dom } f \cap \bar{\mathcal{C}}$ ).** From the analysis in this section, one can see that the boundedness of  $\text{dom } f \cap \bar{\mathcal{C}}$  in Assumption A(ii) is used to guarantee the existence of solutions of problem (1.1) and the subproblem (3.1), as well as the boundedness of  $\{\tilde{\mathbf{x}}^k\}$ , which is a key fact for developing the convergence of the sequence in Theorem 3.2. Here, we would like to comment on some other (possibly weaker) assumptions in place of the boundedness assumption. Indeed, one could just assume that  $f + \delta_{\bar{\mathcal{C}}}$  is level-bounded and  $\sum |\langle \Delta^k, \tilde{\mathbf{x}}^{k+1} - \tilde{\mathbf{x}}^k \rangle| < \infty$ . The former together with Assumption A(i) will ensure that the original problem and the subproblem have solutions, while the latter, together with  $\sup_k \{\gamma_k\} < \infty$ , the summability of  $\{\mu_k\}$  and  $\{\nu_k\}$ , (3.5) and Lemma 2.2, can ensure that  $\{f(\tilde{\mathbf{x}}^k)\}$  is convergent. Then, the convergence of  $\{f(\tilde{\mathbf{x}}^k)\}$  and the level-boundedness of  $f + \delta_{\bar{\mathcal{C}}}$  further imply that  $\{\tilde{\mathbf{x}}^k\}$  is bounded. With these facts, one can establish the same results as in Theorems 3.1 and 3.2. Note that condition

$\sum |\langle \Delta^k, \tilde{\mathbf{x}}^{k+1} - \tilde{\mathbf{x}}^k \rangle| < \infty$  can often be met without much difficulty. One simple case is when  $\text{dom } f \cap \bar{\mathcal{C}}$  is bounded and  $\sum \eta_k < \infty$ , as considered in this paper. Moreover, when  $\Delta^k \equiv 0$ , as is the case in application to the optimal transport problem (see the next section for more details),  $\sum |\langle \Delta^k, \tilde{\mathbf{x}}^{k+1} - \tilde{\mathbf{x}}^k \rangle| < \infty$  holds trivially. In addition, one could check one more condition  $|\langle \Delta^k, \tilde{\mathbf{x}}^{k+1} - \tilde{\mathbf{x}}^k \rangle| \leq \tilde{\eta}_k$  along with condition (3.2) at each iteration, where  $\{\tilde{\eta}_k\}$  is a given summable nonnegative sequence. This then enforces  $\sum |\langle \Delta^k, \tilde{\mathbf{x}}^{k+1} - \tilde{\mathbf{x}}^k \rangle| < \infty$ .

## 4 Application to the optimal transport problem

In this section, we present a concrete application to the optimal transport (OT) problem to show the potential advantages of our iBPPA in Algorithm 1. The discrete OT problem is a classical optimization problem that has received great attention in recent years. We refer interested readers to a recent comprehensive survey [42] (mainly from the computational point of view) and references therein for more details. Mathematically, the discrete OT problem is given as follows:

$$\min_X \langle C, X \rangle \quad \text{s.t.} \quad X \in \Omega := \{X \in \mathbb{R}^{m \times n} : X \mathbf{e}_n = \mathbf{a}, X^\top \mathbf{e}_m = \mathbf{b}, X \geq 0\}, \quad (4.1)$$

where  $C \in \mathbb{R}_+^{m \times n}$  is a given cost matrix,  $\mathbf{a} := (a_1, \dots, a_m)^\top \in \Sigma_m$  and  $\mathbf{b} := (b_1, \dots, b_n)^\top \in \Sigma_n$  are given probability vectors with  $\Sigma_m$  (resp.  $\Sigma_n$ ) denoting the  $m$  (resp.  $n$ )-dimensional unit simplex, and  $\mathbf{e}_m$  (resp.  $\mathbf{e}_n$ ) denotes the  $m$  (resp.  $n$ )-dimensional vector of all ones. It is obvious that the OT problem (4.1) falls into the form of (1.1) via some simple reformulations and thus our iBPPA in Algorithm 1 is applicable. We will consider the following two cases.

### 4.1 iBPPA with the quadratic proximal term

In this case, we equivalently reformulate (4.1) as

$$\min_X \delta_\Omega(X) + \langle C, X \rangle \quad \text{s.t.} \quad X \in \mathbb{R}^{m \times n}, \quad (4.2)$$

which obviously takes the form of (1.1) with  $f(X) = \delta_\Omega(X) + \langle C, X \rangle$  and  $\mathcal{C} = \mathbb{R}^{m \times n}$ . Then, we can apply our iBPPA with the quadratic kernel function  $\phi(X) = \frac{1}{2} \|X\|_F^2$  to solve (4.2). The subproblem at each iteration takes the following generic form

$$\min_X \delta_\Omega(X) + \langle C, X \rangle + \frac{\gamma}{2} \|X - S\|_F^2 \quad (4.3)$$

for some given  $S \in \mathbb{R}^{m \times n}$  and  $\gamma > 0$ , which is equivalent to

$$\min_X \frac{1}{2} \|X - S + \gamma^{-1} C\|_F^2 \quad \text{s.t.} \quad X \in \Omega. \quad (4.4)$$

Thus, solving the subproblem (4.3) amounts to computing the projection of  $G := S - \gamma^{-1} C$  over  $\Omega$ . To the best of our knowledge, the state-of-the-art method for computing such a projection is the semismooth Newton conjugate gradient (SSNCG) method proposed recently by Li, Sun and Toh [31]. Specifically, they consider the following dual problem of (4.4):

$$\min_{\mathbf{y}} \Psi(\mathbf{y}) := \frac{1}{2} \|\Pi_+(\mathcal{A}^*(\mathbf{y}) + G)\|_F^2 - \langle \mathbf{y}, \mathbf{c} \rangle - \frac{1}{2} \|G\|_F^2 \quad \text{s.t.} \quad \mathbf{y} \in \text{Ran}(\mathcal{A}), \quad (4.5)$$

where  $\mathbf{y} \in \mathbb{R}^{m+n}$  is the dual variable,  $\mathcal{A} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m+n}$  is the linear operator defined by  $\mathcal{A}(X) := [X\mathbf{e}_n; X^\top\mathbf{e}_m]$ ,  $\mathcal{A}^*$  is the adjoint operator of  $\mathcal{A}$ ,  $\text{Ran}(\mathcal{A})$  is the range space of  $\mathcal{A}$ ,  $\Pi_+ : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}_+^{m \times n}$  is the projection operator over  $\mathbb{R}_+^{m \times n}$ , and  $\mathbf{c} := [\mathbf{a}; \mathbf{b}]$ . It is easy to verify that if  $\bar{\mathbf{y}}$  is a solution of the nonsmooth equation

$$\nabla\Psi(\mathbf{y}) = \mathcal{A}\Pi_+(\mathcal{A}^*(\mathbf{y}) + G) - \mathbf{c} = 0, \quad \mathbf{y} \in \text{Ran}(\mathcal{A}),$$

then  $\bar{\mathbf{y}}$  solves (4.5) and  $\bar{X} := \Pi_+(\mathcal{A}^*(\bar{\mathbf{y}}) + G)$  solves (4.4). In view of this, SSNCG is then adapted to solve the above nonsmooth equation. Indeed, started from  $\mathbf{y}^0 \in \text{Ran}(\mathcal{A})$ , SSNCG ensures that the generated sequence  $\{\mathbf{y}^t\}$  always lies in  $\text{Ran}(\mathcal{A})$  and  $\|\nabla\Psi(\mathbf{y}^t)\| \rightarrow 0$  (see [31, Theorem 2]). Thus, in practice, an approximate solution  $X^t := \Pi_+(\mathcal{A}^*(\mathbf{y}^t) + G)$  of (4.4) can be returned when  $\|\nabla\Psi(\mathbf{y}^t)\| \leq \varepsilon$  for a given tolerance  $\varepsilon > 0$ . Extensive numerical results have been reported in [31] to show the high efficiency of SSNCG for computing the projection over  $\Omega$ . Hence, it is natural to use SSNCG as a subroutine for our iBPPA employing the quadratic kernel function.

A possible feasibility issue, however, may occur when one tries to verify the stopping condition for solving the subproblem (4.3) via SSNCG, because an approximate solution  $X^t = \Pi_+(\mathcal{A}^*(\mathbf{y}^t) + G)$  returned by SSNCG may not be exactly feasible (indeed, we only have  $\|\mathcal{A}(X^t) - \mathbf{c}\| \leq \varepsilon$ ). Therefore, an additional projection or rounding procedure may be needed to produce a feasible point in  $\Omega$  when performing a certain inexact rule. But its computation is in general nontrivial especially for a complicated feasible region  $\Omega$ . Fortunately, in our iBPPA, we are able to avoid explicitly computing a feasible point and allow an approximately feasible  $X^t$  to be the next proximal point based on the observations given in the next two paragraphs.

We first assume that there is a procedure, denoted by  $\mathcal{G}_\Omega$ , such that for any  $X \geq 0$ , after performing  $\mathcal{G}_\Omega$  on  $X$ , we can obtain that  $\mathcal{G}_\Omega(X) \in \Omega$  and  $\|X - \mathcal{G}_\Omega(X)\|_F \leq c\|\mathcal{A}(X) - \mathbf{c}\|$  for some constant  $c > 0$ . Since  $\Omega$  is a polyhedron, such a procedure is indeed achievable. One natural example is the projection operator denoted by  $\mathcal{P}_\Omega$ . By the Hoffman error bound theorem [25], there must exist a constant  $c > 0$  such that  $\|X - \mathcal{P}_\Omega(X)\|_F \leq c\|\mathcal{A}(X) - \mathbf{c}\|$  for any  $X \geq 0$ . Moreover, one can also consider the rounding procedure in [1, Algorithm 2] as  $\mathcal{G}_\Omega$ , which can be computationally more efficient than the projection.

Next we discuss how the stopping condition (3.2) for the subproblem (4.3) in our iBPPA can be verified. When an approximate solution  $X^t = \Pi_+(\mathcal{A}^*(\mathbf{y}^t) + G) \geq 0$  is returned by SSNCG, with the aid of  $\mathcal{G}_\Omega$ , we have that

$$\|X^t - \mathcal{G}_\Omega(X^t)\|_F \leq c\|\mathcal{A}(X^t) - \mathbf{c}\| = c\|\nabla\Psi(\mathbf{y}^t)\|. \quad (4.6)$$

Thus, for any  $Y \in \Omega$ , we see that

$$\begin{aligned} & \langle -C - \gamma(X^t - S), Y - \mathcal{G}_\Omega(X^t) \rangle \\ &= \gamma \langle G - X^t, Y - \mathcal{G}_\Omega(X^t) \rangle = \gamma \langle \mathcal{A}^*(\mathbf{y}^t) + G - X^t, Y - \mathcal{G}_\Omega(X^t) \rangle \\ &= \gamma \langle \mathcal{A}^*(\mathbf{y}^t) + G - X^t, Y - X^t \rangle + \gamma \langle \mathcal{A}^*(\mathbf{y}^t) + G - X^t, X^t - \mathcal{G}_\Omega(X^t) \rangle \\ &\leq \gamma \langle \mathcal{A}^*(\mathbf{y}^t) + G - X^t, X^t - \mathcal{G}_\Omega(X^t) \rangle \\ &\leq \gamma \|\min\{\mathcal{A}^*(\mathbf{y}^t) + G, 0\}\|_F \|X^t - \mathcal{G}_\Omega(X^t)\|_F \leq \gamma c' c \|\nabla\Psi(\mathbf{y}^t)\|, \end{aligned} \quad (4.7)$$

where the first equality follows from  $G := S - \gamma^{-1}C$ , the second equality follows from  $\langle \mathcal{A}^*(\mathbf{y}^t), Y - \mathcal{G}_\Omega(X^t) \rangle = \langle \mathbf{y}^t, \mathcal{A}(Y) - \mathcal{A}(\mathcal{G}_\Omega(X^t)) \rangle = 0$ , and the first inequality follows from  $X^t = \Pi_+(\mathcal{A}^*(\mathbf{y}^t) + G)$  and  $Y \geq 0$ . The last inequality follows from (4.6) and the fact that  $\{\mathbf{y}^t\}$  is convergent [31,

Theorem 2], and hence  $\|\min\{\mathcal{A}^*(\mathbf{y}^t) + G, 0\}\|_F$  must be bounded from the above by some constant  $c' > 0$ . Then, for any  $\nu \geq 0$  such that  $\gamma c' c \|\nabla\Psi(\mathbf{y}^t)\| \leq \nu$ , we can obtain from (4.7) that

$$0 \in \partial_\nu \delta_\Omega(\mathcal{G}_\Omega(X^t)) + C + \gamma(X^t - S).$$

In view of this relation and (4.6), our inexact condition (3.2) is checkable at the pair of points  $(X^t, \mathcal{G}_\Omega(X^t))$  and it can be satisfied as long as  $\|\nabla\Psi(\mathbf{y}^t)\|$  is sufficiently small. It is worth noting that, though the procedure  $\mathcal{G}_\Omega$  is used in above discussion, it turns out that one does not need to explicitly compute  $\mathcal{G}_\Omega(X^t)$  and a possibly infeasible point  $X^t$  is allowed to be the next proximal point within our framework.

In contrast, the classic inexact conditions  $0 \in \partial_\nu \delta_\Omega(X) + C + \gamma(X - S)$  (condition (1.4)) and  $\text{dist}(0, \partial\delta_\Omega(X) + C + \gamma(X - S)) \leq \eta$  (condition (1.5)) have to be checked at a *single* feasible point. Note that, for any  $Y \in \Omega$ ,

$$\begin{aligned} & \langle -C - \gamma(\mathcal{G}_\Omega(X^t) - S), Y - \mathcal{G}_\Omega(X^t) \rangle \\ &= \langle -C - \gamma(X^t - S), Y - \mathcal{G}_\Omega(X^t) \rangle + \gamma \langle X^t - \mathcal{G}_\Omega(X^t), Y - \mathcal{G}_\Omega(X^t) \rangle \\ &\leq \langle -C - \gamma(X^t - S), Y - \mathcal{G}_\Omega(X^t) \rangle + \gamma c'' \|X^t - \mathcal{G}_\Omega(X^t)\|_F \leq \gamma c(c' + c'') \|\nabla\Psi(\mathbf{y}^t)\|, \end{aligned} \quad (4.8)$$

where the first inequality follows from  $\|Y - \mathcal{G}_\Omega(X^t)\|_F \leq c''$  for some constant  $c'' > 0$  (since  $\Omega$  is bounded) and the last inequality follows from (4.6) and (4.7). Then, for any  $\nu \geq 0$  such that  $\gamma c(c' + c'') \|\nabla\Psi(\mathbf{y}^t)\| \leq \nu$ , the inequality (4.8) implies that

$$0 \in \partial_\nu \delta_\Omega(\mathcal{G}_\Omega(X^t)) + C + \gamma(\mathcal{G}_\Omega(X^t) - S),$$

from which we see that condition (1.4) is verifiable at  $\mathcal{G}_\Omega(X^t)$  and can also be satisfied as long as  $\|\nabla\Psi(\mathbf{y}^t)\|$  is sufficiently small. However, within this framework, one has to compute  $\mathcal{G}_\Omega(X^t)$  explicitly and use it as the next proximal point, which can bring more computational burden.

Next, we consider the hybrid proximal extragradient (HPE) method, which is developed and studied in [35, 36, 49, 50, 51, 53] as a constructive variant of the inexact proximal point algorithm (using the quadratic proximal term). In HPE, a relative error criteria is used for the subproblem involved. In our context, for a given  $\sigma \in [0, 1)$ , one needs to find a triple  $(Y, V, \varepsilon) \in \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n} \times \mathbb{R}_+$  such that

$$V \in \partial_\varepsilon (\delta_\Omega + \langle C, \cdot \rangle)(Y), \quad \|\gamma^{-1}V + Y - S\|_F^2 + 2\gamma^{-1}\varepsilon \leq \sigma^2 \|Y - S\|_F^2.$$

Indeed, recall (4.7), we have that

$$V^t := -\gamma(X^t - S) \in \partial_{\varepsilon_t} \delta_\Omega(\mathcal{G}_\Omega(X^t)) + C = \partial_{\varepsilon_t} (\delta_\Omega + \langle C, \cdot \rangle)(\mathcal{G}_\Omega(X^t))$$

with  $\varepsilon_t := \gamma c' c \|\nabla\Psi(\mathbf{y}^t)\|$ . Thus, the above relative error criterion is verifiable at  $(\mathcal{G}_\Omega(X^t), V^t, \varepsilon_t)$  and can be satisfied whenever

$$\|X^t - \mathcal{G}_\Omega(X^t)\|_F^2 + 2c' c \|\nabla\Psi(\mathbf{y}^t)\| \leq \sigma^2 \|\mathcal{G}_\Omega(X^t) - S\|_F^2.$$

But this criterion may not be easy to check *directly* since the constants  $c$  and  $c'$  are generally unknown. We now recall (4.6) and the fact that  $c^2 \|\nabla\Psi(\mathbf{y}^t)\|^2 \leq \|\nabla\Psi(\mathbf{y}^t)\|$  holds for all sufficiently large  $t$  (since  $\|\nabla\Psi(\mathbf{y}^t)\| \rightarrow 0$  along the sequence generated by SSNCG). Thus, one could check whether  $\|\nabla\Psi(\mathbf{y}^t)\| \leq \tilde{\sigma}^2 \|\mathcal{G}_\Omega(X^t) - S\|_F^2$  for some  $\tilde{\sigma} \in [0, \sigma]$  in order to guarantee the above condition. Note that, in practical implementations, one could simply choose any  $\tilde{\sigma} \in [0, 1)$  since

$\sigma$  can be any number in  $[0, 1)$ . After obtaining such a triple, an extragradient step is performed to compute the new point  $S - \gamma^{-1}V^t$ , which is exactly  $X^t$  in this case (since  $V^t = -\gamma(X^t - S)$ ). Thus, similar to our framework, HPE also allows the possibly infeasible point  $X^t$  to be the next proximal point, but the quantity  $\|\mathcal{G}_\Omega(X^t) - S\|_F^2$  requires the explicit computation of  $\mathcal{G}_\Omega(X^t)$  (if not impossible) for the verification of the relative error criteria, which brings more computational burden. Finally, note that here we only focus on the comparisons with the primal application of the HPE (as presented above), which is more straightforward for problem (4.2) and is closer to our approach.

## 4.2 iBPPA with the entropic proximal term

In this case, we equivalently reformulate (4.1) as

$$\min_X \delta_{\Omega^\circ}(X) + \langle C, X \rangle \quad \text{s.t.} \quad X \geq 0, \quad (4.9)$$

where  $\Omega^\circ := \{X \in \mathbb{R}^{m \times n} : X\mathbf{e}_n = \mathbf{a}, X^\top \mathbf{e}_m = \mathbf{b}\}$  is an affine space. This problem takes the form of (1.1) with  $f(X) = \delta_{\Omega^\circ}(X) + \langle C, X \rangle$  and  $\mathcal{C} = \mathbb{R}_{++}^{m \times n}$ . Then, we apply our iBPPA with the entropy kernel function  $\phi(X) = \sum_{ij} x_{ij}(\log x_{ij} - 1)$  for solving (4.9). The subproblem involved at each iteration takes the following generic form

$$\min_X \delta_{\Omega^\circ}(X) + \langle C, X \rangle + \gamma \mathcal{D}_\phi(X, S)$$

for some given  $S \in \mathbb{R}^{m \times n}$  and  $\gamma > 0$ , which is equivalent to

$$\min_X \langle M, X \rangle + \gamma \sum_{ij} x_{ij}(\log x_{ij} - 1), \quad \text{s.t.} \quad X\mathbf{e}_n = \mathbf{a}, X^\top \mathbf{e}_m = \mathbf{b}, \quad (4.10)$$

where  $M := C - \gamma \log S$ . Note that the constraint  $X \geq 0$  is implicitly imposed by  $\text{dom } \phi = \mathbb{R}_+^{m \times n}$ . Moreover, the subproblem (4.10) has the same form as the entropic regularized OT problem and hence can be readily solved by the popular Sinkhorn's algorithm [42, Section 4.2]. Specifically, let  $K := e^{-M/\gamma}$ . Then, given an arbitrary initial positive vector  $\mathbf{v}^0$ , the iterative scheme is given by

$$\mathbf{u}^t = \mathbf{a} ./ K \mathbf{v}^{t-1}, \quad \mathbf{v}^t = \mathbf{b} ./ K^\top \mathbf{u}^t, \quad (4.11)$$

where  $./$  denotes the entrywise division between two vectors. When a pair  $(\mathbf{u}^t, \mathbf{v}^t)$  is obtained based on a certain stopping criterion, an approximate solution of (4.10) can be recovered by setting  $X^t := \text{Diag}(\mathbf{u}^t) K \text{Diag}(\mathbf{v}^t)$ . Sinkhorn's algorithm in (4.11) only involves matrix-vector multiplications/divisions with  $O(m + n)$  memory complexity and hence can be implemented highly efficiently in practice. However, it should be noted that Sinkhorn's algorithm may suffer from severe numerical instabilities (due to loss of accuracy involving overflow/underflow operations) and very slow convergence speed when the proximal parameter  $\gamma$  takes a small value. The former issue can partially be alleviated by some stabilization techniques (e.g., the log-sum-exp operation [42, Section 4.4]) at the expense of losing some computational efficiency, while the latter is hard to circumvent. Fortunately, in our iBPPA, we have the freedom not to choose a small  $\gamma$  and thus the aforementioned two issues can be avoided. More details on Sinkhorn's algorithm for solving the entropic regularized OT problem can be found in [42, Section 4].

We next discuss how to use Sinkhorn's algorithm as a subroutine in our iBPPA employing the entropic proximal term. Note that an approximate solution  $X^t := \text{Diag}(\mathbf{u}^t) K \text{Diag}(\mathbf{v}^t)$  returned

by Sinkhorn's algorithm is in general not exactly feasible. Thus, some existing inexact conditions such as (1.4) and (1.5) cannot be directly verified at  $X^t$ . Therefore, a certain projection or rounding procedure is needed. Moreover, such a procedure would be more restrictive than that in the case of using the quadratic proximal term because conditions like (1.4) and (1.5) can only be checked at a point in  $\Omega^\circ \cap \mathbb{R}_{++}^{m \times n}$ , that is, the *relative interior* of  $\Omega$ . Therefore, one needs to have a procedure, denoted by  $\mathcal{G}_{\Omega+}$ , such that  $\mathcal{G}_{\Omega+}(X^t) \in \text{relint } \Omega$ , which is in general more difficult to construct than a procedure, denoted by  $\mathcal{G}_\Omega$ , such that  $\mathcal{G}_\Omega(X^t) \in \Omega$ . Fortunately, our iBPPA only requires the latter procedure  $\mathcal{G}_\Omega$ . Recall that  $M = C - \gamma \log S$ ,  $X^t = \text{Diag}(\mathbf{u}^t) K \text{Diag}(\mathbf{v}^t)$  and  $K = e^{-M/\gamma}$ . Then, for any  $Y \in \Omega^\circ$ , we see that

$$\begin{aligned} & \langle -C - \gamma(\log X^t - \log S), Y - \mathcal{G}_\Omega(X^t) \rangle = \langle -M - \gamma \log X^t, Y - \mathcal{G}_\Omega(X^t) \rangle \\ & = \langle -M - \gamma \log(\text{Diag}(\mathbf{u}^t) K \text{Diag}(\mathbf{v}^t)), Y - \mathcal{G}_\Omega(X^t) \rangle \\ & = \langle -\gamma(\log \mathbf{u}^t) \mathbf{e}_n^\top - \gamma \mathbf{e}_m(\log \mathbf{v}^t)^\top, Y - \mathcal{G}_\Omega(X^t) \rangle \\ & = -\gamma \langle \log \mathbf{u}^t, Y \mathbf{e}_n - \mathcal{G}_\Omega(X^t) \mathbf{e}_n \rangle - \gamma \langle \log \mathbf{v}^t, Y^\top \mathbf{e}_m - (\mathcal{G}_\Omega(X^t))^\top \mathbf{e}_m \rangle = 0, \end{aligned}$$

where the last equality follows from  $Y \mathbf{e}_n = \mathbf{a} = \mathcal{G}_\Omega(X^t) \mathbf{e}_n$  and  $Y^\top \mathbf{e}_m = \mathbf{b} = (\mathcal{G}_\Omega(X^t))^\top \mathbf{e}_m$ . This relation implies that

$$0 \in \partial \delta_{\Omega^\circ}(\mathcal{G}_\Omega(X^t)) + C + \gamma(\log X^t - \log S). \quad (4.12)$$

In this case, the quantity  $\Delta^k$  on the left-hand-side of (3.2) is 0. Thus, our inexact condition (3.2) is verifiable at the pair  $(X^t, \mathcal{G}_\Omega(X^t))$  and can be satisfied when  $\mathcal{D}_\phi(\mathcal{G}_\Omega(X^t), X^t)$  is sufficiently small. Moreover, we further have  $\|X^t - \mathcal{G}_\Omega(X^t)\|_F \leq c(\|X^t \mathbf{e}_n - \mathbf{a}\| + \|(X^t)^\top \mathbf{e}_m - \mathbf{b}\|)$  for some  $c > 0$  as in subsection 4.1. Thus, when the feasibility violation  $\|X^t \mathbf{e}_n - \mathbf{a}\| + \|(X^t)^\top \mathbf{e}_m - \mathbf{b}\|$  is small, the quantity  $\mathcal{D}_\phi(\mathcal{G}_\Omega(X^t), X^t)$  is also likely to be small. Indeed, we can observe from Figure 1 that both quantities decrease in tandem. Thus, in practice, one may only check the quantity  $\|X^t \mathbf{e}_n - \mathbf{a}\| + \|(X^t)^\top \mathbf{e}_m - \mathbf{b}\| = \|\mathbf{u}^t \odot K \mathbf{v}^t - \mathbf{a}\|$  without explicitly computing  $\mathcal{G}_\Omega(X^t)$  to save cost.

Note that  $\nabla \phi$  is explicitly invertible in this case and  $V^t := -\gamma(\log X^t - \log S) \in \partial(\delta_{\Omega^\circ} + \langle C, \cdot \rangle)(\mathcal{G}_\Omega(X^t))$  from (4.12). Thus, we see that the relative error condition (1.6) is also checkable, and by some simple manipulations, it can be shown to hold at  $(X^t, \mathcal{G}_\Omega(X^t), V^t)$  when  $\mathcal{D}_\phi(\mathcal{G}_\Omega(X^t), X^t) \leq \sigma^2 \mathcal{D}_\phi(\mathcal{G}_\Omega(X^t), S)$ . Comparing to our framework, the verification of this condition requires one to compute one more quantity  $\mathcal{D}_\phi(\mathcal{G}_\Omega(X^t), S)$  and thus incurs extra cost. Moreover, condition (1.6) generally requires one to compute an element in  $\partial f$  (rather than a larger set  $\partial_\nu f$  for some  $\nu > 0$ ) at an intermediary point and then performs an 'extragradient' step to compute a new proximal point. Such a requirement on an element of  $\partial f$  at some intermediate point may be expensive to satisfy when  $f$  is not simple; see, for example, the class of linear programming problems studied in [15].

Finally, we end this section with a few remarks on some potential numerical issues that may be encountered when employing the inexact condition (1.5). Assume that we have at hand a procedure  $\mathcal{G}_{\Omega+}$  that is able to find a point in the relative interior of  $\Omega$ . Using similar arguments for deducing (4.12), we can get

$$\gamma(\log \mathcal{G}_{\Omega+}(X^t) - \log X^t) \in \partial \delta_{\Omega^\circ}(\mathcal{G}_{\Omega+}(X^t)) + C + \gamma(\log \mathcal{G}_{\Omega+}(X^t) - \log S).$$

Thus, condition (1.5) is verifiable at  $\mathcal{G}_{\Omega+}(X^t)$  and can be satisfied when the error  $\gamma \|\log \mathcal{G}_{\Omega+}(X^t) - \log X^t\|_F = \gamma \|\log(\mathcal{G}_{\Omega+}(X^t)/X^t)\|_F$  is sufficiently small. However, as observed from our experiments, checking the quantity  $\gamma \|\log(\mathcal{G}_{\Omega+}(X^t)/X^t)\|_F$  is numerically less stable than checking the

quantity  $\mathcal{D}_\phi(\mathcal{G}_{\Omega^+}(X^t), X^t)$  in our framework, as one can observe from Figure 1. To better illustrate this issue, we generate some instances of subproblem (4.10) as follows: we set  $m = n = 1000$  and set  $S$  to be a matrix of ones; moreover, we choose  $\gamma \in \{0.1, 0.01, 0.001\}$  and randomly generate  $(\mathbf{a}, \mathbf{b}, C)$  by the same way in subsection 6.1. Then, we apply Sinkhorn’s algorithm and terminate it after some iterations. During the iterations, we record the feasibility accuracy of  $X^t$  as well as the quantities  $\gamma \|\log(\mathcal{G}_{\Omega^+}(X^t)/X^t)\|_F$  and  $\mathcal{D}_\phi(\mathcal{G}_{\Omega^+}(X^t), X^t)$ , where the rounding procedure in [1, Algorithm 2] is chosen as  $\mathcal{G}_{\Omega^+}$ . Moreover, to avoid the possible overflow or underflow in computation, we set  $X^t := \max\{X^t, 10^{-16}\}$  and  $\mathcal{G}_{\Omega^+}(X^t) := \max\{\mathcal{G}_{\Omega^+}(X^t), 10^{-16}\}$  when computing the quantities  $\gamma \|\log(\mathcal{G}_{\Omega^+}(X^t)/X^t)\|_F$  and  $\mathcal{D}_\phi(\mathcal{G}_{\Omega^+}(X^t), X^t)$ . The computational results are presented in Figure 1. One can see that  $\gamma \|\log(\mathcal{G}_{\Omega^+}(X^t)/X^t)\|_F$  always stays at a large value and it hardly decreases as  $X^t$  gets close to the feasible set, especially when  $\gamma$  is small. This is mainly because some entries of  $\mathcal{G}_{\Omega^+}(X^t)/X^t$  could be close to zero and that leads to large negative numbers after performing the log operations. Thus, using  $\gamma \|\log(\mathcal{G}_{\Omega^+}(X^t)/X^t)\|_F \leq \eta$  for some  $\eta \geq 0$  as a stopping criterion (hence condition (1.5)) could be impractical. In contrast, the quantity  $\mathcal{D}_\phi(\mathcal{G}_{\Omega^+}(X^t), X^t)$  decreases much more rapidly to zero as the iteration proceeds. Therefore, it can provide a reliable stopping criterion. This indeed highlights another advantage of our inexact framework with the entropic kernel function.

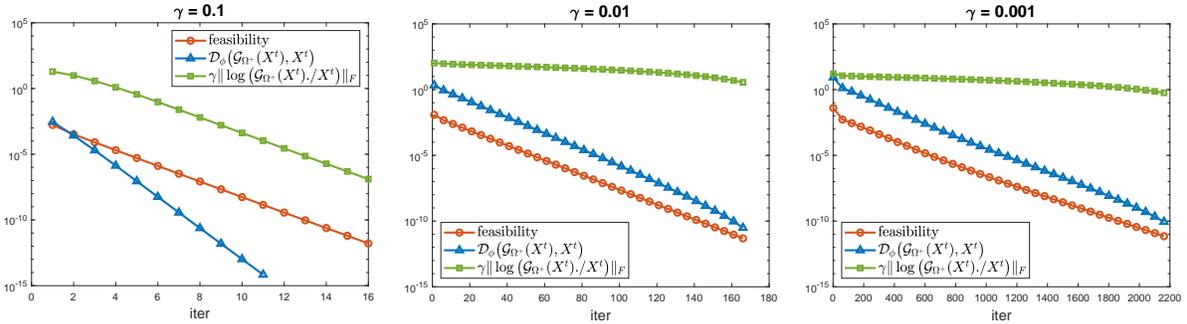


Figure 1: Comparisons between  $\mathcal{D}_\phi(\mathcal{G}_{\Omega^+}(X^t), X^t)$  and  $\gamma \|\log(\mathcal{G}_{\Omega^+}(X^t)/X^t)\|_F$ , where “feasibility” denotes the feasibility accuracy of  $X^t$ , defined as the value of  $\|X^t \mathbf{e}_n - \mathbf{a}\| + \|(X^t)^\top \mathbf{e}_m - \mathbf{b}\|$ .

## 5 An inertial variant of the iBPPA

In this section, we develop an inertial variant of our iBPPA, denoted by V-iBPPA for short. The inspiration comes from Güler’s first classical accelerated proximal point method [21] and its recent Bregman extension [61]. The basic idea used there actually originates from Nesterov’s ingenious technique (called *estimate sequence*) in [39] that has motivated many kinds of accelerated methods (see, for example, [3, 29, 57, 58]). We also adapt such an idea to develop the V-iBPPA to achieve the possible acceleration. Specifically, our estimate sequence of functions  $\{H_k(\mathbf{x})\}_{k=0}^\infty$  are constructed recursively as follows:

$$\begin{aligned}
 H_0(\mathbf{x}) &:= f(\tilde{\mathbf{x}}^0) + \pi \mathcal{D}_\phi(\mathbf{x}, \mathbf{x}^0), \\
 H_{k+1}(\mathbf{x}) &:= (1 - \theta_k) H_k(\mathbf{x}) \\
 &\quad + \theta_k (f(\tilde{\mathbf{x}}^{k+1}) + \gamma_k \langle \nabla \phi(\mathbf{y}^k) - \nabla \phi(\mathbf{x}^{k+1}), \mathbf{x} - \tilde{\mathbf{x}}^{k+1} \rangle - \rho \eta_k - \nu_k),
 \end{aligned} \tag{5.1}$$

where  $\pi$  and  $\gamma_k$  are positive numbers,  $\eta_k$  and  $\nu_k$  are nonnegative numbers,  $\theta_k$  is a number in  $[0, 1)$  (to be specified by (5.9)),  $\mathbf{x}^0 = \tilde{\mathbf{x}}^0 \in \mathcal{C}$  ( $= \text{int dom } \phi$ ) and  $\rho > 0$  is the diameter of the feasible set (by Assumption A(ii)). Resorting to this estimate sequence of functions  $\{H_k(\mathbf{x})\}_{k=0}^\infty$ , we then present the complete framework of the V-iBPPA in Algorithm 2.

---

**Algorithm 2** An inertial variant of iBPPA (V-iBPPA) for (1.1)

---

**Input:** Let  $\{\gamma_k\}_{k=0}^\infty$ ,  $\{\nu_k\}_{k=0}^\infty$ ,  $\{\eta_k\}_{k=0}^\infty$  and  $\{\mu_k\}_{k=0}^\infty$  be four sequences of nonnegative scalars. Choose  $\mathbf{x}^0 = \tilde{\mathbf{x}}^0 = \mathbf{z}^0 \in \mathcal{C}$  arbitrarily and a kernel function  $\phi$ . Set  $k = 0$ .

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

**Step 1.** Choose  $\theta_k \in [0, 1)$  satisfying (5.9) and set  $\mathbf{y}^k = \theta_k \mathbf{z}^k + (1 - \theta_k) \mathbf{x}^k$ .

**Step 2.** Find a pair  $(\mathbf{x}^{k+1}, \tilde{\mathbf{x}}^{k+1})$  by approximately solving the following problem

$$\min_{\mathbf{x}} f(\mathbf{x}) + \gamma_k \mathcal{D}_\phi(\mathbf{x}, \mathbf{y}^k), \quad (5.2)$$

such that  $\mathbf{x}^{k+1} \in \mathcal{C}$ ,  $\tilde{\mathbf{x}}^{k+1} \in \text{dom } f \cap \bar{\mathcal{C}}$  and

$$\begin{aligned} \Delta^k &\in \partial_{\nu_k} f(\tilde{\mathbf{x}}^{k+1}) + \gamma_k (\nabla \phi(\mathbf{x}^{k+1}) - \nabla \phi(\mathbf{y}^k)) \\ \text{with } \|\Delta^k\| &\leq \eta_k, \quad \mathcal{D}_\phi(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^{k+1}) \leq \mu_k. \end{aligned} \quad (5.3)$$

**Step 3.** Set  $H_{k+1}(\mathbf{x})$  by (5.1) and compute  $\mathbf{z}^{k+1} = \arg \min_{\mathbf{x}} \{H_{k+1}(\mathbf{x})\}$ .

**Step 4.** Set  $k = k + 1$  and go to **Step 1**.

**end while**

**Output:**  $(\mathbf{x}^k, \tilde{\mathbf{x}}^k)$

---

Comparing to the iBPPA in Algorithm 1, the V-iBPPA in Algorithm 2 uses an intermediary point  $\mathbf{y}^k$  as the proximal point. When  $\theta_k \equiv 0$ , we have  $\mathbf{y}^k \equiv \mathbf{x}^k$  and the V-iBPPA readily reduces to the iBPPA, while with the special choice of  $\theta_k$  in (5.9), we shall see later that the V-iBPPA enjoys a flexible convergence rate depending on the property of the kernel function and is able to achieve a faster rate in some scenarios. From arguments similar to those following Algorithm 1, the subproblem (5.2) and the inexact condition (5.3) are also well-defined under Assumption A, provided  $\mathbf{y}^k \in \mathcal{C}$  ( $= \text{int dom } \phi$ ). Note from the construction of  $H_k(\mathbf{x})$  in (5.1) that

$$H_k(\mathbf{x}) = L_k(\mathbf{x}) + \pi c_k \mathcal{D}_\phi(\mathbf{x}, \mathbf{x}^0), \quad (5.4)$$

where  $L_k(\cdot)$  is an affine function and  $c_k$  is a positive scalar depending on  $k$ . Since  $\mathcal{D}_\phi(\cdot, \mathbf{x}^0)$  is level-bounded (by condition (B3) in Definition 2.1), then  $H_k(\mathbf{x})$  is level-bounded. Hence, an optimal solution  $\mathbf{z}^k$  of problem  $\min_{\mathbf{x}} \{H_k(\mathbf{x})\}$  exists [47, Theorem 1.9] and must also be unique since  $\phi$  is strictly convex (by condition (B1) in Definition 2.1). The essential smoothness of  $\phi$  (by Assumption A(iii)) further imposes that  $\mathbf{z}^k \in \mathcal{C}$ . This together with  $\mathbf{x}^k \in \mathcal{C}$  ensures that the intermediary point  $\mathbf{y}^k$ , as a convex combination of  $\mathbf{x}^k$  and  $\mathbf{z}^k$ , always lies in  $\mathcal{C}$ . Therefore, Algorithm 2 is well-defined. Here, we would also like to point out that, when using  $\phi(\cdot) = \frac{1}{2} \|\cdot\|^2$  (hence  $\text{dom } \phi = \text{int dom } \phi = \mathbb{E}$ ), one can have more freedom to choose other updating formulas for  $\mathbf{y}^k$ , and they give rise to different variants of the accelerated PPA such as Güler's second accelerated proximal point method [21, Section 6] and the so-called catalyst acceleration method

proposed recently in [32]. Our inexact criterion can also be incorporated into those variants. We will leave this topic for future investigation.

In the following, we shall study the convergence property of our V-iBPPA in Algorithm 2. Since we now use the intermediary point  $\mathbf{y}^k$  as the proximal point in the subproblem (5.2), the analysis for Algorithm 2 turns out to be different from that for Algorithm 1. In particular, all convergence results presented later are in terms of the objective function value, as is the case in most existing works on various accelerated methods. Our analysis is motivated by several existing works (e.g., [3, 21, 24, 61]) that are based on the Nesterov's estimate sequence. Before proceeding, we introduce the following quadrangle scaling property for the Bregman distance.

**Definition 5.1 (Quadrangle scaling property).** *Let  $\phi$  be a proper closed convex function which is differentiable on  $\text{int dom } \phi$ . We say  $\phi$  has the quadrangle scaling property (QSP) if there exist an exponent  $\lambda \geq 1$  and two constants  $\tau_1, \tau_2 > 0$  such that, for any  $\mathbf{a}, \mathbf{c} \in \text{dom } \phi$  and  $\mathbf{b}, \mathbf{d} \in \text{int dom } \phi$ , the following inequality holds for any  $\theta \in [0, 1]$ ,*

$$\mathcal{D}_\phi(\theta\mathbf{a} + (1-\theta)\mathbf{c}, \theta\mathbf{b} + (1-\theta)\mathbf{d}) \leq \tau_1 \theta^\lambda \mathcal{D}_\phi(\mathbf{a}, \mathbf{b}) + \tau_2 (1-\theta)^\lambda \mathcal{D}_\phi(\mathbf{c}, \mathbf{d}). \quad (5.5)$$

Here,  $\lambda$  is called the quadrangle scaling exponent (QSE) of  $\phi$ , and  $\tau_1, \tau_2$  are called the quadrangle scaling constants (QSCs) of  $\phi$ .

Note that when  $\mathbf{c} = \mathbf{d}$ , the QSP reduces to a so-called intrinsic triangle scaling property (TSP) introduced recently in [24, Section 2] for developing accelerated Bregman proximal gradient methods. Thus, our QSP is an extension of the TSP. Two representative examples for the QSP are given as follows.

- If  $\phi$  is  $\mu_\phi$ -strongly convex and  $\nabla\phi$  is  $L_\phi$ -Lipschitz, i.e.,  $\frac{\mu_\phi}{2}\|\mathbf{x}-\mathbf{y}\|^2 \leq \mathcal{D}_\phi(\mathbf{x}, \mathbf{y}) \leq \frac{L_\phi}{2}\|\mathbf{x}-\mathbf{y}\|^2$ , then for any  $\mathbf{a}, \mathbf{c} \in \text{dom } \phi$ ,  $\mathbf{b}, \mathbf{d} \in \text{int dom } \phi$  and  $\theta \in [0, 1]$ ,

$$\begin{aligned} \mathcal{D}_\phi(\theta\mathbf{a} + (1-\theta)\mathbf{c}, \theta\mathbf{b} + (1-\theta)\mathbf{d}) &\leq \frac{L_\phi}{2}\|\theta(\mathbf{a}-\mathbf{b}) + (1-\theta)(\mathbf{c}-\mathbf{d})\|^2 \\ &\leq L_\phi\theta^2\|\mathbf{a}-\mathbf{b}\|^2 + L_\phi(1-\theta)^2\|\mathbf{c}-\mathbf{d}\|^2 = \frac{2L_\phi}{\mu_\phi}\theta^2\frac{\mu_\phi}{2}\|\mathbf{a}-\mathbf{b}\|^2 + \frac{2L_\phi}{\mu_\phi}(1-\theta)^2\frac{\mu_\phi}{2}\|\mathbf{c}-\mathbf{d}\|^2 \\ &\leq \frac{2L_\phi}{\mu_\phi}\theta^2\mathcal{D}_\phi(\mathbf{a}, \mathbf{b}) + \frac{2L_\phi}{\mu_\phi}(1-\theta)^2\mathcal{D}_\phi(\mathbf{c}, \mathbf{d}). \end{aligned}$$

Thus, in this case,  $\phi$  has the QSP with  $\lambda = 2$  and  $\tau_1 = \tau_2 = 2L_\phi/\mu_\phi$ .

- If  $\mathcal{D}_\phi(\cdot, \cdot)$  is jointly convex, which can be satisfied by the entropy kernel function  $\phi(\mathbf{x}) = \sum_i x_i(\log x_i - 1)$  (see [6] for more examples), then for any  $\theta \in [0, 1]$ ,

$$\mathcal{D}_\phi(\theta\mathbf{a} + (1-\theta)\mathbf{c}, \theta\mathbf{b} + (1-\theta)\mathbf{d}) \leq \theta \mathcal{D}_\phi(\mathbf{a}, \mathbf{b}) + (1-\theta) \mathcal{D}_\phi(\mathbf{c}, \mathbf{d})$$

Thus, in this case,  $\phi$  has the QSP with  $\lambda = \tau_1 = \tau_2 = 1$ .

We now start the analysis with a lemma concerning the difference  $H_k(\mathbf{x}) - f(\mathbf{x})$ .

**Lemma 5.1.** *Let the estimate sequence of functions  $\{H_k(\mathbf{x})\}_{k=0}^\infty$  be generated by (5.1). Then, for all  $k \geq 0$ , we have*

$$H_{k+1}(\mathbf{x}) - f(\mathbf{x}) \leq (1-\theta_k)(H_k(\mathbf{x}) - f(\mathbf{x})), \quad \forall \mathbf{x} \in \text{dom } f \cap \bar{\mathcal{C}}.$$

*Proof.* From condition (5.3), there exists a  $\mathbf{d}^{k+1} \in \partial_{\nu_k} f(\tilde{\mathbf{x}}^{k+1})$  such that  $\Delta^k = \mathbf{d}^{k+1} + \gamma_k (\nabla \phi(\mathbf{x}^{k+1}) - \nabla \phi(\mathbf{y}^k))$ . For notational simplicity, let

$$\Xi^k(\mathbf{x}) := \langle \nabla \phi(\mathbf{y}^k) - \nabla \phi(\mathbf{x}^{k+1}), \mathbf{x} - \tilde{\mathbf{x}}^{k+1} \rangle. \quad (5.6)$$

Then, for any  $\mathbf{x} \in \text{dom} f \cap \bar{\mathcal{C}}$ , we see that

$$\begin{aligned} f(\mathbf{x}) &\geq f(\tilde{\mathbf{x}}^{k+1}) + \langle \mathbf{d}^{k+1}, \mathbf{x} - \tilde{\mathbf{x}}^{k+1} \rangle - \nu_k \\ &= f(\tilde{\mathbf{x}}^{k+1}) + \langle \Delta^k - \gamma_k (\nabla \phi(\mathbf{x}^{k+1}) - \nabla \phi(\mathbf{y}^k)), \mathbf{x} - \tilde{\mathbf{x}}^{k+1} \rangle - \nu_k \\ &\geq f(\tilde{\mathbf{x}}^{k+1}) + \gamma_k \Xi^k(\mathbf{x}) + \langle \Delta^k, \mathbf{x} - \tilde{\mathbf{x}}^{k+1} \rangle - \nu_k \\ &\geq f(\tilde{\mathbf{x}}^{k+1}) + \gamma_k \Xi^k(\mathbf{x}) - \rho\eta_k - \nu_k, \end{aligned} \quad (5.7)$$

where the last inequality follows from  $\langle \Delta^k, \mathbf{x} - \tilde{\mathbf{x}}^{k+1} \rangle \geq -\|\mathbf{x} - \tilde{\mathbf{x}}^{k+1}\| \|\Delta^k\| \geq -\rho\eta_k$  due to  $\mathbf{x}, \tilde{\mathbf{x}}^{k+1} \in \text{dom} f \cap \bar{\mathcal{C}}$  and Assumption A. Using (5.7) and the construction of  $H_k(\mathbf{x})$  in (5.1), we see that

$$\begin{aligned} H_{k+1}(\mathbf{x}) - f(\mathbf{x}) &= (1 - \theta_k) H_k(\mathbf{x}) + \theta_k \left( f(\tilde{\mathbf{x}}^{k+1}) + \gamma_k \Xi^k(\mathbf{x}) - \rho\eta_k - \nu_k \right) - f(\mathbf{x}) \\ &= (1 - \theta_k) (H_k(\mathbf{x}) - f(\mathbf{x})) + \theta_k \left( f(\tilde{\mathbf{x}}^{k+1}) + \gamma_k \Xi^k(\mathbf{x}) - \rho\eta_k - \nu_k - f(\mathbf{x}) \right) \\ &\leq (1 - \theta_k) (H_k(\mathbf{x}) - f(\mathbf{x})). \end{aligned}$$

This completes the proof.  $\square$

One can easily see from Lemma 5.1 that, at  $k$ -th iteration, the difference  $H_k(\mathbf{x}) - f(\mathbf{x})$  is reduced by a factor  $1 - \theta_k$ . Then, by induction, we further obtain that

$$H_k(\mathbf{x}) - f(\mathbf{x}) \leq c_k (H_0(\mathbf{x}) - f(\mathbf{x})), \quad \forall \mathbf{x} \in \text{dom} f \cap \bar{\mathcal{C}}, \quad (5.8)$$

where

$$c_0 := 1, \quad c_k := \prod_{i=0}^{k-1} (1 - \theta_i) \quad \text{for } k \geq 1.$$

To further evaluate the reduction in the original objective (that is,  $f(\tilde{\mathbf{x}}^k) - f(\mathbf{x})$ ) based on (5.8), we only need to explore the relation between  $f(\tilde{\mathbf{x}}^k)$  and  $H_k(\mathbf{z}^k)$ , where  $\mathbf{z}^k = \arg \min_{\mathbf{x}} \{H_k(\mathbf{x})\}$  by **Step 3** in Algorithm 2. Indeed, we have the following result.

**Lemma 5.2.** *Let  $\{\mathbf{x}^k\}$  and  $\{\tilde{\mathbf{x}}^k\}$  be the sequences generated by the V-iBPPA in Algorithm 2. Suppose that Assumption A holds,  $\phi$  has the QSP with an exponent  $\lambda \geq 1$  and QSCs  $\tau_1, \tau_2 > 0$ , and  $\theta_k$  is chosen such that*

$$\tau_1 \gamma_k \theta_k^\lambda = \pi c_k (1 - \theta_k). \quad (5.9)$$

*If  $f(\tilde{\mathbf{x}}^k) \leq H_k(\mathbf{z}^k) + \delta_k$  for some  $k \geq 0$  and  $\delta_k \geq 0$ , then*

$$f(\tilde{\mathbf{x}}^{k+1}) \leq H_{k+1}(\mathbf{z}^{k+1}) + (1 - \theta_k) \delta_k + \gamma_k (\mu_k + \tau_2 \mu_{k-1}) + \rho\eta_k + \nu_k.$$

*Proof.* First, from (5.4), Lemma 2.4 and the definition of  $\mathbf{z}^k$  as a minimizer of  $H_k(\cdot)$  (by **Step 3** in Algorithm 2), we see that

$$H_k(\mathbf{z}^{k+1}) = H_k(\mathbf{z}^k) + \pi c_k \mathcal{D}_\phi(\mathbf{z}^{k+1}, \mathbf{z}^k),$$

which, together with the hypothesis of this lemma, implies that

$$H_k(\mathbf{z}^{k+1}) \geq f(\tilde{\mathbf{x}}^k) + \pi c_k \mathcal{D}_\phi(\mathbf{z}^{k+1}, \mathbf{z}^k) - \delta_k. \quad (5.10)$$

Moreover, recall the definition of  $\Xi^k(\cdot)$  in (5.6), one can verify that

$$\begin{aligned} & (1 - \theta_k) f(\tilde{\mathbf{x}}^k) + \theta_k \left( f(\tilde{\mathbf{x}}^{k+1}) + \gamma_k \Xi^k(\mathbf{z}^{k+1}) - \rho\eta_k - \nu_k \right) \\ & \geq (1 - \theta_k) \left( f(\tilde{\mathbf{x}}^{k+1}) + \gamma_k \Xi^k(\tilde{\mathbf{x}}^k) - \rho\eta_k - \nu_k \right) + \theta_k \left( f(\tilde{\mathbf{x}}^{k+1}) + \gamma_k \Xi^k(\mathbf{z}^{k+1}) - \rho\eta_k - \nu_k \right) \\ & = f(\tilde{\mathbf{x}}^{k+1}) + \gamma_k \Xi^k(\theta_k \mathbf{z}^{k+1} + (1 - \theta_k) \tilde{\mathbf{x}}^k) - \rho\eta_k - \nu_k \\ & = f(\tilde{\mathbf{x}}^{k+1}) + \gamma_k \mathcal{D}_\phi(\theta_k \mathbf{z}^{k+1} + (1 - \theta_k) \tilde{\mathbf{x}}^k, \mathbf{x}^{k+1}) + \gamma_k \mathcal{D}_\phi(\tilde{\mathbf{x}}^{k+1}, \mathbf{y}^k) \\ & \quad - \gamma_k \mathcal{D}_\phi(\theta_k \mathbf{z}^{k+1} + (1 - \theta_k) \tilde{\mathbf{x}}^k, \mathbf{y}^k) - \gamma_k \mathcal{D}_\phi(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^{k+1}) - \rho\eta_k - \nu_k \\ & \geq f(\tilde{\mathbf{x}}^{k+1}) - \gamma_k \mathcal{D}_\phi(\theta_k \mathbf{z}^{k+1} + (1 - \theta_k) \tilde{\mathbf{x}}^k, \mathbf{y}^k) - \gamma_k \mathcal{D}_\phi(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^{k+1}) - \rho\eta_k - \nu_k \\ & \geq f(\tilde{\mathbf{x}}^{k+1}) - \gamma_k \mathcal{D}_\phi(\theta_k \mathbf{z}^{k+1} + (1 - \theta_k) \tilde{\mathbf{x}}^k, \theta_k \mathbf{z}^k + (1 - \theta_k) \mathbf{x}^k) - \gamma_k \mu_k - \rho\eta_k - \nu_k \\ & \geq f(\tilde{\mathbf{x}}^{k+1}) - \gamma_k \tau_1 \theta_k^\lambda \mathcal{D}_\phi(\mathbf{z}^{k+1}, \mathbf{z}^k) - \gamma_k \tau_2 (1 - \theta_k)^\lambda \mathcal{D}_\phi(\tilde{\mathbf{x}}^k, \mathbf{x}^k) - \gamma_k \mu_k - \rho\eta_k - \nu_k \\ & \geq f(\tilde{\mathbf{x}}^{k+1}) - \tau_1 \gamma_k \theta_k^\lambda \mathcal{D}_\phi(\mathbf{z}^{k+1}, \mathbf{z}^k) - \gamma_k (\mu_k + \tau_2 \mu_{k-1}) - \rho\eta_k - \nu_k, \end{aligned} \quad (5.11)$$

where the first inequality follows from (5.7) with  $\mathbf{x} = \tilde{\mathbf{x}}^k$ , the second equality follows from the four points identity (2.1), the third inequality follows from  $\mathbf{y}^k = \theta_k \mathbf{z}^k + (1 - \theta_k) \mathbf{x}^k$  (by **Step 1** in Algorithm 2) and  $\mathcal{D}_\phi(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^{k+1}) \leq \mu_k$  (by condition (5.3)), the second last inequality follows from the QSP of  $\mathcal{D}_\phi$  and the last inequality follows from  $1 - \theta_k \leq 1$  and  $\mathcal{D}_\phi(\tilde{\mathbf{x}}^k, \mathbf{x}^k) \leq \mu_{k-1}$ . Then, we see that

$$\begin{aligned} H_{k+1}(\mathbf{z}^{k+1}) &= (1 - \theta_k) H_k(\mathbf{z}^{k+1}) + \theta_k \left( f(\tilde{\mathbf{x}}^{k+1}) + \gamma_k \Xi^k(\mathbf{z}^{k+1}) - \rho\eta_k - \nu_k \right) \\ &\geq (1 - \theta_k) f(\tilde{\mathbf{x}}^k) + \theta_k \left( f(\tilde{\mathbf{x}}^{k+1}) + \gamma_k \Xi^k(\mathbf{z}^{k+1}) - \rho\eta_k - \nu_k \right) + \pi c_k (1 - \theta_k) \mathcal{D}_\phi(\mathbf{z}^{k+1}, \mathbf{z}^k) - (1 - \theta_k) \delta_k \\ &\geq f(\tilde{\mathbf{x}}^{k+1}) + [\pi c_k (1 - \theta_k) - \tau_1 \gamma_k \theta_k^\lambda] \mathcal{D}_\phi(\mathbf{z}^{k+1}, \mathbf{z}^k) - (1 - \theta_k) \delta_k - \gamma_k (\mu_k + \tau_2 \mu_{k-1}) - \rho\eta_k - \nu_k \\ &\geq f(\tilde{\mathbf{x}}^{k+1}) - (1 - \theta_k) \delta_k - \gamma_k (\mu_k + \tau_2 \mu_{k-1}) - \rho\eta_k - \nu_k, \end{aligned}$$

where the first equality follows from the construction of  $H_k(\mathbf{x})$  in (5.1), the first inequality follows from (5.10), the second inequality follows from (5.11) and the last inequality follows from the choice of  $\theta_k$  in (5.9). This completes the proof.  $\square$

Then, we have the theorem concerning the reduction of the objective value.

**Theorem 5.1.** *Suppose that Assumption A holds,  $\phi$  has the QSP and  $\theta_k$  satisfies (5.9). Let  $\{\mathbf{x}^k\}$  and  $\{\tilde{\mathbf{x}}^k\}$  be the sequences generated by the V-iBPPA in Algorithm 2. Then, for any optimal solution  $\mathbf{x}^*$  of problem (1.1), we have*

$$f(\tilde{\mathbf{x}}^N) - f(\mathbf{x}^*) \leq c_N (f(\tilde{\mathbf{x}}^0) - f(\mathbf{x}^*) + \pi \mathcal{D}_\phi(\mathbf{x}^*, \mathbf{x}^0)) + \delta_N, \quad (5.12)$$

where the error sequence  $\{\delta_k\}_{k=0}^\infty$  satisfies

$$\delta_0 = 0, \quad \delta_{k+1} = (1 - \theta_k) \delta_k + \gamma_k (\mu_k + \tau_2 \mu_{k-1}) + \rho\eta_k + \nu_k, \quad k = 0, 1, \dots \quad (5.13)$$

*Proof.* First, from Lemma 5.2, it is easy to prove by induction that  $f(\tilde{\mathbf{x}}^N) \leq H_N(\mathbf{z}^N) + \delta_N$  for any  $N \geq 0$ . Moreover, note from (5.8) that  $H_N(\mathbf{x}^*) - f(\mathbf{x}^*) \leq c_N (H_0(\mathbf{x}^*) - f(\mathbf{x}^*))$  for

any  $N \geq 0$ . These relations together with the fact that  $\mathbf{z}^N$  is the minimizer of the problem  $\min_{\mathbf{x}}\{H_N(\mathbf{x})\}$  prove the desired result.  $\square$

From the choice of  $\theta_k$  in (5.9), we see that  $0 < \theta_k < 1$  and hence  $c_N \rightarrow 0$ . This together with (5.12) shows that  $f(\tilde{\mathbf{x}}^N)$  converges to  $f^* := \min\{f(\mathbf{x}) : \mathbf{x} \in \bar{\mathcal{C}}\}$  as long as  $\delta_N \rightarrow 0$ . Here,  $c_N$  and  $\delta_N$  naturally determine the convergence rate and thus we must estimate their magnitudes. The following estimate on  $c_N$  extends [21, Lemma 2.2] to a more general setting.

**Lemma 5.3.** *For any  $N \geq 1$ , we have*

$$\left(1 + (\pi/\tau_1)^{\frac{1}{\lambda}} \sum_{k=0}^{N-1} \gamma_k^{-\frac{1}{\lambda}}\right)^{-\lambda} \leq c_N \leq \left(1 + \lambda^{-1}(\pi/\tau_1)^{\frac{1}{\lambda}} \sum_{k=0}^{N-1} \gamma_k^{-\frac{1}{\lambda}}\right)^{-\lambda}. \quad (5.14)$$

Moreover, if  $\sup_k\{\gamma_k\} < \infty$ , then  $c_N = O\left(\left(\sum_{k=0}^{N-1} \gamma_k^{-\frac{1}{\lambda}}\right)^{-\lambda}\right)$ .

*Proof.* First, it is easy to see that  $c_{k+1} = (1 - \theta_k)c_k$  and then  $\theta_k = 1 - c_{k+1}/c_k$  for all  $k \geq 0$ . Substituting this in (5.9) results in

$$\tau_1 \gamma_k (1 - c_{k+1}/c_k)^\lambda = \pi c_{k+1} \iff c_{k+1}^{-1} - c_k^{-1} = (\pi/\tau_1)^{\frac{1}{\lambda}} \gamma_k^{-\frac{1}{\lambda}} c_{k+1}^{\frac{1}{\lambda}-1}. \quad (5.15)$$

Note that  $c_{k+1} \leq c_k$  (since  $\theta_k \in (0, 1)$ ) and  $\lambda \geq 1$  (by definition of QSE). Hence,

$$c_{k+1}^{\frac{1}{\lambda}-1} \left(c_{k+1}^{-\frac{1}{\lambda}} - c_k^{-\frac{1}{\lambda}}\right) = c_{k+1}^{-1} - c_{k+1}^{\frac{1}{\lambda}-1} c_k^{-\frac{1}{\lambda}} \leq c_{k+1}^{-1} - c_k^{-1}.$$

Combing this and (5.15), we see that  $c_{k+1}^{-\frac{1}{\lambda}} - c_k^{-\frac{1}{\lambda}} \leq (\pi/\tau_1)^{\frac{1}{\lambda}} \gamma_k^{-\frac{1}{\lambda}}$ . Summing this inequality from  $k = 0$  to  $k = N - 1$ , we obtain that

$$c_N^{-\frac{1}{\lambda}} \leq 1 + (\pi/\tau_1)^{\frac{1}{\lambda}} \sum_{k=0}^{N-1} \gamma_k^{-\frac{1}{\lambda}},$$

which gives the lower bound on  $c_N$ . On the other hand, it is easy to show by Young's inequality that  $c_{k+1}^{\frac{1}{\lambda}-1} c_k^{-\frac{1}{\lambda}} \leq (1 - \lambda^{-1})c_{k+1}^{-1} + \lambda^{-1}c_k^{-1}$  and thus

$$c_{k+1}^{-1} - c_k^{-1} \leq \lambda c_{k+1}^{\frac{1}{\lambda}-1} \left(c_{k+1}^{-\frac{1}{\lambda}} - c_k^{-\frac{1}{\lambda}}\right).$$

Combing this and (5.15), we see that  $c_{k+1}^{-\frac{1}{\lambda}} - c_k^{-\frac{1}{\lambda}} \geq \lambda^{-1}(\pi/\tau_1)^{\frac{1}{\lambda}} \gamma_k^{-\frac{1}{\lambda}}$ . Summing this inequality from  $k = 0$  to  $k = N - 1$ , we obtain that

$$c_N^{-\frac{1}{\lambda}} \geq 1 + \lambda^{-1}(\pi/\tau_1)^{\frac{1}{\lambda}} \sum_{k=0}^{N-1} \gamma_k^{-\frac{1}{\lambda}},$$

which gives the upper bound on  $c_N$ . The other result follows immediately from (5.14).  $\square$

We immediately have the following proposition.

**Proposition 5.1.** *Suppose that all conditions in Theorem 5.1 and Lemma 5.3 hold. If  $\delta_N \leq O(c_N)$ , then*

$$f(\tilde{\mathbf{x}}^N) - f(\mathbf{x}^*) \leq O\left(\left(\sum_{k=0}^{N-1} \gamma_k^{-\frac{1}{\lambda}}\right)^{-\lambda}\right). \quad (5.16)$$

Notice from Proposition 5.1 that, when the QSE  $\lambda$  is strictly larger than 1, the convergence rate (in terms of the function value) of the V-iBPPA is better than the convergence rate of the iBPPA given in Theorem 3.1 since  $(\sum_{k=0}^{N-1} \gamma_k^{-\frac{1}{\lambda}})^\lambda > \sum_{k=0}^{N-1} \gamma_k^{-1}$  always holds for any  $\lambda > 1$ . When  $\lambda = 2$ , this result recovers the related results in [21, 61] when the subproblem is solved exactly. Moreover, using (5.16) and similar arguments as in Remark 3.1, we see that  $\{f(\tilde{\mathbf{x}}^k)\}$  can also converge to  $f(\mathbf{x}^*)$  arbitrarily fast with a proper decreasing sequence of  $\{\gamma_k\}$ . However, we should be mindful that such a favorable convergence rate comes with the requirement that  $\delta_N \leq O(c_N)$ , which may impose stringent inexact tolerance requirement for each subproblem. An estimate on  $\delta_N$  under certain choices of  $\{\mu_k\}$ ,  $\{\nu_k\}$ ,  $\{\eta_k\}$  is given in the following lemma.

**Lemma 5.4.** *Suppose that  $\{\delta_k\}_{k=0}^\infty$  satisfies (5.13). Then, for all  $N \geq 1$ , we have*

$$\delta_N \leq \frac{1}{(1 + \lambda^{-1}(\pi/\tau_1)^{\frac{1}{\lambda}} \sum_{i=0}^{N-1} \gamma_i^{-\frac{1}{\lambda}})^\lambda} \sum_{k=0}^{N-1} \left(1 + \left(\frac{\pi}{\tau_1}\right)^{\frac{1}{\lambda}} \sum_{i=0}^k \gamma_i^{-\frac{1}{\lambda}}\right)^\lambda \beta_k, \quad (5.17)$$

where  $\beta_k := \gamma_k(\mu_k + \tau_2 \mu_{k-1}) + \rho\eta_k + \nu_k$ . Moreover, suppose that  $\{\gamma_k\}$  is non-increasing and for some  $p > 1$  such that  $p \neq \lambda + 1$ ,

$$\mu_k \leq O\left(\frac{1}{(k+1)^p}\right), \quad \nu_k \leq O\left(\frac{\gamma_k}{(k+1)^p}\right), \quad \eta_k \leq O\left(\frac{\gamma_k}{(k+1)^p}\right), \quad \forall k \geq 0. \quad (5.18)$$

Then, for all  $N \geq 1$ , we have  $\delta_N \leq O\left(\frac{1}{N^{p-1}}\right)$ .

*Proof.* Since  $1 - \theta_k = c_{k+1}/c_k$  for all  $k \geq 0$ , then  $\delta_{k+1}$  can be written as  $\delta_{k+1} = (c_{k+1}/c_k) \delta_k + \beta_k$ . Dividing this equality by  $c_{k+1}$  and rearranging the terms, we have  $\delta_{k+1}/c_{k+1} - \delta_k/c_k = \beta_k/c_{k+1}$ . Thus, summing this equality from  $k = 0$  to  $k = N - 1$  results in  $\delta_N = c_N \sum_{k=0}^{N-1} \beta_k/c_{k+1}$ . Using this together with the lower and upper bounds on  $c_N$  ( $N \geq 1$ ) in (5.14), we obtain (5.17).

Moreover, since  $\{\gamma_k\}$  is non-increasing (hence  $\gamma_k \leq \gamma_0$  for all  $k$ ), we have that  $\sum_{i=0}^{N-1} \gamma_i^{-\frac{1}{\lambda}} \geq \gamma_0^{-\frac{1}{\lambda}} N$  and  $\sum \gamma_i^{-\frac{1}{\lambda}} = \infty$ . The latter further implies that there exists a constant  $a > 0$  such that  $1 + (\pi/\tau_1)^{\frac{1}{\lambda}} \sum_{i=0}^k \gamma_i^{-\frac{1}{\lambda}} \leq a (\pi/\tau_1)^{\frac{1}{\lambda}} \sum_{i=0}^k \gamma_i^{-\frac{1}{\lambda}}$  for any  $k \geq 0$ . On the other hand, one can see from (5.18) that there exist a constant  $a' > 0$  such that  $\beta_k = \gamma_k(\mu_k + \tau_2 \mu_{k-1}) + \rho\eta_k + \nu_k \leq a' \gamma_k / (k+1)^p$  for all  $k \geq 0$ . Thus, substituting these bounds in (5.17) results in

$$\begin{aligned} \delta_N &\leq \frac{\gamma_0 a' a^\lambda \lambda^\lambda}{N^\lambda} \sum_{k=0}^{N-1} \left(\sum_{i=0}^k \gamma_i^{-\frac{1}{\lambda}}\right)^\lambda \frac{\gamma_k}{(k+1)^p} \\ &= \frac{\gamma_0 a' a^\lambda \lambda^\lambda}{N^\lambda} \sum_{k=0}^{N-1} \left(\sum_{i=0}^k \left(\frac{\gamma_k}{\gamma_i}\right)^{\frac{1}{\lambda}}\right)^\lambda \frac{1}{(k+1)^p} \leq \frac{\gamma_0 a' a^\lambda \lambda^\lambda}{N^\lambda} \sum_{k=0}^{N-1} (k+1)^{\lambda-p}. \end{aligned}$$

Note also that there exists a constant  $\tilde{a} > 0$  such that

$$\sum_{k=0}^{N-1} (k+1)^{\lambda-p} = \sum_{k=1}^N k^{\lambda-p} \leq \tilde{a} \int_1^N t^{\lambda-p} dt \leq \tilde{a} (\lambda + 1 - p)^{-1} N^{\lambda+1-p}.$$

Using these relations, we complete the proof.  $\square$

Using the estimates on  $c_N$  and  $\delta_N$ , together with (5.12), we can give the following concrete convergence rate in terms of the function value for our V-iBPPA.

**Theorem 5.2.** *Suppose that all conditions in Theorem 5.1, Lemmas 5.3 and 5.4 hold. Let  $\{\mathbf{x}^k\}$  and  $\{\tilde{\mathbf{x}}^k\}$  be the sequences generated by the V-iBPPA in Algorithm 2. Then, for any optimal solution  $\mathbf{x}^*$  of problem (1.1), we have*

$$f(\tilde{\mathbf{x}}^N) - f(\mathbf{x}^*) \leq O\left(\left(\sum_{k=0}^{N-1} \gamma_k^{-\frac{1}{\lambda}}\right)^{-\lambda}\right) + O\left(\frac{1}{N^{p-1}}\right).$$

*In particular, if  $\gamma_k$  satisfies  $0 < \underline{\gamma} \leq \gamma_k \leq \bar{\gamma} < +\infty$  and  $p > \lambda + 1$ , then we have*

$$f(\tilde{\mathbf{x}}^N) - f(\mathbf{x}^*) \leq O\left(\frac{1}{N^\lambda}\right).$$

Now, we see from Theorem 5.2 that, when  $0 < \underline{\gamma} \leq \gamma_k \leq \bar{\gamma} < +\infty$ , our V-iBPPA enjoys a flexible convergence rate determined by the QSE  $\lambda$  of the kernel function  $\phi$ . Thus, when  $\lambda > 1$ , the V-iBPPA indeed improves the  $O(1/N)$  convergence rate of the iBPPA (see Remark 3.1), and in the particular case  $\lambda = 2$ , the V-iBPPA achieves the rate of  $O(1/N^2)$  common to existing accelerated (inexact) proximal point algorithms; see, for example, [21, 36, 48, 59]. But the choices of  $\{\mu_k\}$ ,  $\{\nu_k\}$ ,  $\{\eta_k\}$  following the way of (5.18) may become more restrictive. For example, for  $\lambda = 2$ , we need  $p > 3$  for the V-iBPPA to achieve the rate of  $O(N^{-2})$ .<sup>1</sup> Before ending this section, some remarks are in order regarding the practical implementations of our V-iBPPA.

**Remark 5.1 (Practical computation on  $\mathbf{z}^{k+1}$ ).** *Note that, at each iteration of our V-iBPPA, one needs to compute  $\mathbf{z}^{k+1}$  as the minimizer of  $H_{k+1}(\mathbf{x})$  in order to form the next intermediary point  $\mathbf{y}^{k+1}$ . Thanks to the favorable construction of  $\{H_k(\mathbf{x})\}_{k=0}^\infty$  in (5.1), we can show that  $\mathbf{z}^{k+1}$  actually admits a closed form expression based on the following observations. Indeed, we see from (5.4), Lemma 2.4 and the definition of  $\mathbf{z}^k$  as a minimizer of  $H_k(\cdot)$  that*

$$H_k(\mathbf{x}) = H_k(\mathbf{z}^k) + \pi c_k \mathcal{D}_\phi(\mathbf{x}, \mathbf{z}^k). \quad (5.19)$$

*Then we can show by using (5.1), (5.9) and (5.19) that*

$$\begin{aligned} \mathbf{z}^{k+1} &= \arg \min_{\mathbf{x}} \{H_{k+1}(\mathbf{x})\} = \arg \min_{\mathbf{x}} \left\{ \tau_1 \theta_k \mathcal{D}_\phi(\mathbf{x}, \mathbf{z}^k) + \langle \nabla \phi(\mathbf{y}^k) - \nabla \phi(\mathbf{x}^{k+1}), \mathbf{x} \rangle \right\} \\ &= \nabla \phi^* \left( \nabla \phi(\mathbf{z}^k) + \tau_1^{-1} \theta_k^{-1} (\nabla \phi(\mathbf{x}^{k+1}) - \nabla \phi(\mathbf{y}^k)) \right), \end{aligned}$$

*where the last equality follows from the optimality condition together with [44, Theorem 26.5] and the fact that  $\phi$  is strictly convex and essentially smooth (by Assumption A(iii)). Therefore, one can compute  $\mathbf{z}^{k+1}$  via the above expression without generating  $H_{k+1}(\mathbf{x})$  explicitly. For example, when  $\phi(\mathbf{x}) = \frac{1}{2} \|\mathbf{x}\|^2$ , we have that  $\phi^*(\mathbf{x}') = \frac{1}{2} \|\mathbf{x}'\|^2$  and  $\mathbf{z}^{k+1} = \mathbf{z}^k + \tau_1^{-1} \theta_k^{-1} (\mathbf{x}^{k+1} - \mathbf{y}^k)$ . Moreover, when  $\phi(\mathbf{x}) = \sum_i x_i (\log x_i - 1)$ , we have that  $\phi^*(\mathbf{x}') = \sum_i e^{x'_i}$  and*

$$\mathbf{z}^{k+1} = \mathbf{z}^k \odot \left( \mathbf{x}^{k+1} ./ \mathbf{y}^k \right)^{\tau_1^{-1} \theta_k^{-1}}. \quad (5.20)$$

**Remark 5.2 (Practical computation on QSE and QSC).** *From the above analysis, one can see that the QSP of a kernel function  $\phi$  is crucial for developing the V-iBPPA, as is the case in [24, 61] using the TSP for deriving their inertial methods. In particular, the choice*

<sup>1</sup>It is worth noting from [21, Section 3] that, when  $\phi(\cdot) = \frac{1}{2} \|\cdot\|^2$  and  $\mu_k \equiv \nu_k \equiv 0$ , a weaker condition  $p > \frac{3}{2}$  is sufficient for guaranteeing the rate of  $O(N^{-2})$ .

of  $\theta_k$  by (5.9) requires the knowledge of the QSE  $\lambda$  as well as the QSC  $\tau_1$ , and  $\lambda$  would also determine the convergence rate (see Theorem 5.2). From the discussions following Definition 5.1, we know that the quadratic kernel function  $\phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$  has  $\lambda = \tau_1 = 2$  which can be readily used in practical computation and grant a rate of  $O(k^{-2})$ , while the entropy kernel function  $\phi(\mathbf{x}) = \sum_i x_i(\log x_i - 1)$  only has  $\lambda = \tau_1 = 1$ , which leads to a rate of  $O(k^{-1})$ . Interestingly, for the entropy kernel function, we observe that, for any  $\theta \in [\epsilon, 1 - \epsilon]$  with a given small  $\epsilon > 0$ ,

$$\begin{aligned} \mathcal{D}_\phi(\theta\mathbf{a} + (1 - \theta)\mathbf{c}, \theta\mathbf{b} + (1 - \theta)\mathbf{d}) &\leq \theta \mathcal{D}_\phi(\mathbf{a}, \mathbf{b}) + (1 - \theta) \mathcal{D}_\phi(\mathbf{c}, \mathbf{d}) \\ &\leq \frac{1}{\theta} \theta^2 \mathcal{D}_\phi(\mathbf{a}, \mathbf{b}) + \frac{1}{1-\theta} (1 - \theta)^2 \mathcal{D}_\phi(\mathbf{c}, \mathbf{d}) \leq \epsilon^{-1} \theta^2 \mathcal{D}_\phi(\mathbf{a}, \mathbf{b}) + \epsilon^{-1} (1 - \theta)^2 \mathcal{D}_\phi(\mathbf{c}, \mathbf{d}), \end{aligned}$$

which implies that the inequality (5.5) holds for any  $\theta \in [\epsilon, 1 - \epsilon]$  with  $\lambda = 2$  and  $\tau_1 = \tau_2 = \epsilon^{-1}$ . This relation is indeed sufficient for studying the convergence behavior of the V-iBPPA within a finite number of iterations (as is the case in practical implementations), because in the analysis (precisely, in (5.11)), we only need the inequality (5.5) to be satisfied at a special  $\theta_k \in [0, 1]$  given by (5.9) and  $\theta_k$  just asymptotically goes to 0. This then motivates us to use  $\lambda = 2$  and  $\tau_1 = \epsilon^{-1}$  for the V-iBPPA with the entropy kernel function to obtain a possibly faster convergence rate when  $\theta_k \geq \epsilon$ , and moreover, we may reset  $\epsilon$  to be a smaller value or simply terminate the algorithm when  $\theta_k < \epsilon$ . But, as observed from our experiments, the choice of  $\tau_1 = \epsilon^{-1}$  seems to be too conservative to achieve a faster speed. Therefore, in our experiments in the next section, we adapt a heuristic strategy to choose  $\tau_1$ . Specifically, we initially set  $\tau_1 = 1$  and then increase it by setting the new  $\tau_1$  to be  $2\tau_1$  if  $\tau_1\theta_k < 0.1$ .

## 6 Numerical experiments

In this section, we conduct some numerical experiments to test our iBPPA and V-iBPPA for solving the standard OT problem (4.1). Our purpose here is to preliminarily show the convergence behaviors of two methods under different inexact settings and evaluate the potential of achieving accelerated performance of the V-iBPPA. More experiments of our iBPPA for solving a class of linear programming problems has been reported in our recent technical report [15]. All experiments in this section are run in MATLAB R2020b on a Windows workstation with Intel Xeon Processor E-2176G@3.70GHz and 64GB of RAM.

### 6.1 Implementation details

One can show that the dual problem of (4.1) is

$$\max_{\mathbf{f}, \mathbf{g}} \langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle \quad \text{s.t.} \quad Z(\mathbf{f}, \mathbf{g}) := C - \mathbf{f}e_n^\top - e_m\mathbf{g}^\top \geq 0, \quad (6.1)$$

and the Karush-Kuhn-Tucker (KKT) system for (4.1) and (6.1) is

$$X\mathbf{e}_n = \mathbf{a}, \quad X^\top\mathbf{e}_m = \mathbf{b}, \quad \langle X, Z(\mathbf{f}, \mathbf{g}) \rangle = 0, \quad X \geq 0, \quad Z(\mathbf{f}, \mathbf{g}) \geq 0, \quad (6.2)$$

where  $\mathbf{f} \in \mathbb{R}^m$  and  $\mathbf{g} \in \mathbb{R}^n$  are the Lagrangian multipliers (or dual variables). Note that the strong duality holds for (4.1) and (6.1), and  $(X, \mathbf{f}, \mathbf{g})$  satisfies the KKT system (6.2) if and only if  $X$  solves (4.1) and  $(\mathbf{f}, \mathbf{g})$  solves (6.1), respectively. Based on (6.2), we define the relative KKT residual for any  $(X, \mathbf{f}, \mathbf{g})$  as follows:

$$\Delta_{\text{kkt}}(X, \mathbf{f}, \mathbf{g}) := \max \{ \Delta_p, \Delta_d, \Delta_c \},$$

where  $\Delta_p := \max \left\{ \frac{\|Xe_n - \mathbf{a}\|}{1 + \|\mathbf{a}\|}, \frac{\|X^\top \mathbf{e}_m - \mathbf{b}\|}{1 + \|\mathbf{b}\|}, \frac{\|\min\{X, 0\}\|_F}{1 + \|X\|_F} \right\}$ ,  $\Delta_d := \frac{\|\min\{Z(\mathbf{f}, \mathbf{g}), 0\}\|_F}{1 + \|C\|_F}$  and  $\Delta_c := \frac{|\langle X, Z(\mathbf{f}, \mathbf{g}) \rangle|}{1 + \|C\|_F}$ . Obviously,  $(X, \mathbf{f}, \mathbf{g})$  is a solution of the KKT system (6.2) if and only if  $\Delta_{\text{kkt}} = 0$ . Thus, it is natural to use  $\Delta_{\text{kkt}}$  to measure the accuracy of an approximate solution returned by a method. We then use  $\Delta_{\text{kkt}}$  to set up the stopping criterion for our iBPPA and V-iBPPA. Specifically, we terminate both methods when

$$\Delta_{\text{kkt}}(X^{k+1}, \mathbf{f}^{k+1}, \mathbf{g}^{k+1}) < \text{Tol}, \quad (6.3)$$

where the value of Tol will be given later, and  $X^{k+1}$  and  $(\mathbf{f}^{k+1}, \mathbf{g}^{k+1})$  are respectively the approximate optimal solutions of the subproblem ((3.1) or (5.2)) and its corresponding dual problem at the  $k$ -th iteration.

For the kernel function  $\phi$ , we adopt two choices:  $\phi(X) = \frac{1}{2}\|X\|_F^2$  (leading to the quadratic proximal term) and  $\phi(X) = \sum_{ij} x_{ij}(\log x_{ij} - 1)$  (leading to the entropic proximal term). For ease of future reference, in the following, we use iPPA/V-iPPA to denote iBPPA/V-iBPPA with the quadratic proximal term and use iEPPA/V-iEPPA to denote iBPPA/V-iBPPA with the entropic proximal term. For V-iPPA and V-iEPPA, the QSE  $\lambda$  and the QSC  $\tau_1$  are chosen based on Remark 5.2. Moreover, from the discussions in Section 4, we have the following facts.

For iPPA/V-iPPA, at the  $k$ -th iteration, the subproblem can be solved by the semismooth Newton conjugate gradient (SSNCG) method and our inexact condition ((3.2) or (5.3)) can be satisfied when  $\|\nabla\Psi_k(\mathbf{y}^{k,t})\|$  is sufficiently small, where  $\nabla\Psi_k$  is the gradient of the dual objective and  $\{\mathbf{y}^{k,t}\}$  is the sequence generated by SSNCG. At the  $k$ -th iteration ( $k \geq 0$ ), we terminate SSNCG when

$$\|\nabla\Psi_k(\mathbf{y}^{k,t})\| \leq \max \{ \Upsilon / (k+1)^p, 10^{-10} \}.$$

For iEPPA/V-iEPPA, at the  $k$ -th iteration, the subproblem can be solved by Sinkhorn's algorithm and our inexact condition ((3.2) or (5.3)) can be satisfied when  $\mathcal{D}_\phi(\mathcal{G}_\Omega(X^{k,t}), X^{k,t})$  is sufficiently small, where  $X^{k,t} := \text{Diag}(\mathbf{u}^{k,t}) K^k \text{Diag}(\mathbf{v}^{k,t})$  with  $\{(\mathbf{u}^{k,t}, \mathbf{v}^{k,t})\}$  generated by (4.11) and  $\mathcal{G}_\Omega$  is a rounding procedure [1, Algorithm 2]. At the  $k$ -th iteration ( $k \geq 0$ ), we terminate Sinkhorn's algorithm when

$$\mathcal{D}_\phi(\mathcal{G}_\Omega(X^{k,t}), X^{k,t}) \leq \max \{ \Upsilon / (k+1)^p, 10^{-10} \}.$$

The above coefficient  $\Upsilon$  controls the initial accuracy for solving the subproblem and, together with  $p$ , would determine the tightness of the tolerance requirement. Generally, for a fixed  $p$ ,  $\Upsilon$  should be neither too small to avoid excessive cost of solving each subproblem, nor too large to avoid unnecessary large number of outer iterations. The optimal choice of  $\Upsilon$  depends on many factors such as the value of  $p$ , the kernel function  $\phi$  and the proximal parameter  $\gamma_k$ . In our experiments, we simply use  $\Upsilon = 1, 10^{-3}$  and  $p = 1.001, 1.01, 1.1, 2.1, 3.1$  without delicate tunings. Moreover, at each iteration, we employ the warm-start strategy to initialize the subroutine (SSNCG or Sinkhorn's algorithm) by the solution obtained at the previous iteration.

For the choice of the proximal parameter  $\gamma_k$ , we simply fix it to be a constant  $\gamma$  throughout the iterations. For iPPA/V-iPPA, we choose  $\gamma \in \{10, 1, 0.1\}$ , and for iEPPA/V-iEPPA, we choose  $\gamma \in \{1, 0.1, 0.01\}$ . It is also possible to adaptively tune  $\gamma_k$ , together with careful tunings of  $\Upsilon$  and  $p$ , to further improve the numerical performance of the whole algorithm, but we will skip such investigations in this paper.

We next discuss how we generate the simulated data. We first generate two discrete probability distributions  $\{(a_i, \mathbf{p}_i) \in \mathbb{R}_+ \times \mathbb{R}^3 : i = 1, \dots, m\}$  and  $\{(b_j, \mathbf{q}_j) \in \mathbb{R}_+ \times \mathbb{R}^3 : j = 1, \dots, n\}$ . Here,  $\mathbf{a} := (a_1, \dots, a_m)^\top$  and  $\mathbf{b} := (b_1, \dots, b_n)^\top$  are probabilities/weights, which are generated

from the uniform distribution on the open interval  $(0, 1)$  and further normalized such that  $\sum_i^m a_i = \sum_j^n b_j = 1$ . Moreover,  $\{\mathbf{p}_i\}$  and  $\{\mathbf{q}_j\}$  are support points whose entries are drawn from a Gaussian mixture distribution via the following MATLAB commands:

```
num = 5; mean = [-20;-10;0;10;20]; sigma(1,1,:) = 5*ones(num,1);
weights = rand(num,1); distrib = gmdistribution(mean,sigma,weights);
```

Then, the cost matrix  $C$  is generated by  $c_{ij} = \|\mathbf{p}_i - \mathbf{q}_j\|^2$  for  $1 \leq i \leq m$  and  $1 \leq j \leq n$  and normalized by dividing (element-wise) by its maximal entry.

As discussed in section 4, the hybrid proximal extragradient (HPE) method and its Bregman generalization using condition (1.6) (denoted by BHPE for short) are applicable for solving the OT problem (4.1) using the same subroutines as our methods. Similarly, an accelerated variant of the HPE (denoted by AHPE for short), developed in [36] based on Nesterov’s acceleration technique, is also applicable. Thus, we include them in our comparisons. The error tolerance constant  $\sigma$  is chosen from  $\{0.999, 0.99, 0.9, 0.5, 0.1\}$ . Moreover, since (4.1) is a linear programming (LP) problem, we can also apply Gurobi 8.0.0 [22] (with default settings) to solve it. It is well known that Gurobi is a powerful commercial package for solving LPs and is able to provide a high quality solution. Therefore, we will use the objective function value obtained by Gurobi as the benchmark in the following figures.

In the following comparisons, we choose  $m = n = 500$  and initialize all methods with  $X^0 := \mathbf{ab}^\top$ . Moreover, we terminate iPPA/V-iPPA/HPE/AHPE when (6.3) holds with  $\text{Tol} < 10^{-7}$  or the number of SSNCG iterations reaches 1000, and terminate iEPPA/V-iEPPA/BHPE when (6.3) holds with  $\text{Tol} < 10^{-5}$  or the number of Sinkhorn iterations reaches 10000.

## 6.2 Comparison results

Figures 2 and 3 show the comparison results of iPPA/V-iPPA/HPE/AHPE and iEPPA/V-iEPPA/BHPE, respectively. In each figure, we plot the “nfval” against the number of SSNCG/Sinkhorn iterations, where “nfval” denotes the normalized function value  $|\langle C, \mathcal{G}_\Omega(X^{k,t}) \rangle - f^*| / |f^*|$ ,  $f^*$  is the highly accurate optimal function value computed by Gurobi and  $X^{k,t}$  is the approximate solution computed by the subroutine at the  $t$ -th inner iteration of the  $k$ -th outer iteration. Moreover, in Tables 1 and 2, we also show the terminating value of  $\Delta_{\text{kkt}}(X^{k+1}, \mathbf{f}^{k+1}, \mathbf{g}^{k+1})$  (denoted by “kkt”), the number of outer iterations (denoted by “out#”), the number of SSNCG/Sinkhorn iterations (denoted by “ssn#”/“sink#”), and the computational time in seconds (denoted by “time”). Note that Sinkhorn’s algorithm itself has been popularly used to approximately solve OT by solving its entropic regularized counterpart (i.e., problem (4.10) with  $C$  in place of  $M$ ). Thus, we also include it in comparison with iEPPA/V-iEPPA/BHPE. From the results, we have several observations as follows.

When  $p = 3.1$  (giving a fast tolerance decay), for (V-)iPPA and (V-)iEPPA, a smaller  $\gamma$  usually leads to a faster convergence speed in terms of the total number of outer iterations incurred. This implies that the choice of  $\gamma$  dominates the convergence rate under a tight tolerance requirement, matching the complexity results in Theorem 3.1 and Proposition 5.1. When  $p$  is smaller, such phenomenon tends to disappear due to the loose accuracy control. But this does not mean worse overall performance. For example, for (V-)iEPPA in Figure 3, the choice of  $p = 1.1$ , along with a relatively large  $\gamma$ , can perform much better. Hence, setting a proper value of  $p$  for faster convergence needs to take into account the choice of  $\gamma$ .

For  $p = 3.1$ , V-iPPA/V-iEPPA always outperforms iPPA/iEPPA, and for  $p = 2.1$ , V-iPPA/V-iEPPA also performs better when  $\gamma$  is large. Indeed, one can see from Tables 1&2, together with Figures 2&3, that, for  $p = 3.1$  (and for  $p = 2.1$  in many cases), V-iPPA/V-iEPPA usually takes less outer iterations to achieve a comparable “kkt”/“nfval” or takes comparable outer iterations to achieve a better “kkt”/“nfval”. This (to some extent) verifies the favorable iteration complexity of V-iPPA/V-iEPPA, as we expect from Remark 5.2. But note that the improvement becomes less significant for a smaller  $\gamma$ , because a small  $\gamma$  would dominate the convergence speed as observed in the last paragraph. For example, when  $\gamma = 0.1$ ,  $\Upsilon = 10^{-3}$  and  $p = 3.1$ , both iPPA and V-iPPA only need 15 outer iterations to obtain a high accuracy solution (“kkt” is about  $10^{-8}$ ) and hence one cannot observe the improvement clearly. On the other hand, when  $p = 1.1$ , the improvement is destroyed by the crude solutions of the subproblems. This matches the results established in Theorem 5.2, which states that improved complexity holds under a sufficiently tight tolerance requirement.

With proper choices of parameters, (V-)iPPA (resp. iEPPA) and (A)HPE (resp. BHPE) can be comparable to each other when measuring “nfval” against the number of SSNCG (resp. Sinkhorn) iterations, as shown in Figures 2 and 3. This is actually reasonable because (V-)iPPA (resp. iEPPA) and (A)HPE (resp. BHPE) essentially use the similar (accelerated) PPA (resp. BPPA) framework but with different stopping criteria for solving the subproblems. Since (A)HPE and BHPE only involve an error tolerance constant  $\sigma \in [0, 1)$ , they are more friendly to parameter tunings, but they may incur non-negligible extra cost on checking the relative error condition.

- As discussed in subsection 4.1, HPE (similarly, AHPE) has to compute a *feasible* intermediary point and thus would need to perform projection/rounding per iteration, while our iPPA can avoid such computations during the iterations. One can also observe from Table 1 that, for each  $\gamma$ , our (V-)iPPA always takes less time than (A)HPE within comparable number of SSNCG iterations. Thus, our (V-)iPPA can be more advantageous for a large-scale problem with a complex polyhedra set.
- As discussed in subsection 4.2, for implementing iEPPA and BHPE, we have to explicitly retrieve an approximate solution  $X^{k,t} := \text{Diag}(\mathbf{u}^{k,t}) K^k \text{Diag}(\mathbf{v}^{k,t})$ , find its projection/rounding  $\mathcal{G}_\Omega(X^{k,t})$  and then compute their Bregman distance  $\mathcal{D}_\phi(\mathcal{G}_\Omega(X^{k,t}), X^{k,t})$ . Moreover, BHPE has to compute one more quantity  $\mathcal{D}_\phi(\mathcal{G}_\Omega(X^{k,t}), X^k)$  and thus incurs extra cost. Since the operation complexity of computing the Bregman distance is roughly  $5mn$ , which is about 2.5 times more than that of Sinkhorn iteration itself (4.11), this extra cost is not negligible. From Table 2, one can also see that, for each  $\gamma$ , our iEPPA usually takes less time than BHPE within a comparable number of Sinkhorn iterations.

Finally, one can see from Figure 3 that Sinkhorn’s algorithm with a relatively large  $\gamma$  is highly efficient for obtaining a rough approximate solution, but when driving  $\gamma$  to a smaller value to obtain a more accurate solution, it rapidly becomes very slow. Moreover, when  $\gamma = 10^{-4}$ , numerical instabilities occur and one needs to carry out the computations of (4.11) via some stabilization techniques (e.g., the *log-sum-exp* technique [42, Section 4.4]) at the expense of losing some computational efficiency. In contrast, under a broad range of tolerance settings, our (V-)iEPPA is able to achieve an approximate solution of reasonable quality even when  $\gamma = 1$ . Thus, we can safely use the efficient iterative scheme (4.11) as a subroutine without worries on possible numerical instabilities. We also notice that the similar framework of iEPPA has

been considered for solving OT in [42, Remark 4.9] and [60]. However, the inexact condition used there is either heuristic (using a fixed number of inner iterations) without the rigorous theoretical guarantee or rather stringent so that it is nontrivial to implement. Thus, our (V-)iEPPA somewhat reduces the gap between the theory and the practical implementation when applying the BPPA-type method for solving OT. We believe that there is still ample room for improving our (V-)iEPPA with a dedicated tolerance adjustment and our (V-)iEPPA has great potential to solve other OT-related problems, which we leave for future research.

Table 1: Comparisons among iPPA, V-iPPA, HPE and AHPE. In the table, “out#” denotes the number of outer iterations, “ssn#” denotes the the number of SSNCG iterations, and “-” means that the number of SSNCG iterations reaches 1000.

method	$\gamma = 10$				$\gamma = 1$				$\gamma = 0.1$			
	kkt	out#	ssn#	time	kkt	out#	ssn#	time	kkt	out#	ssn#	time
iPPA ( $\Upsilon = 1$ )												
$p = 1.001$	2.35e-3	363	-	7.9	3.55e-4	323	-	7.5	2.74e-3	329	-	7.9
$p = 1.01$	1.27e-3	344	-	7.6	2.79e-3	335	-	7.5	8.14e-4	340	-	7.9
$p = 1.1$	4.44e-4	385	-	7.3	1.16e-4	372	-	7.4	9.96e-4	365	-	7.6
$p = 2.1$	4.67e-7	558	-	7.5	9.60e-8	475	924	6.9	3.26e-7	526	-	7.6
$p = 3.1$	3.20e-7	533	-	7.8	9.26e-8	126	284	2.2	8.06e-8	75	241	2.1
V-iPPA ( $\Upsilon = 1$ )												
$p = 1.001$	2.71e-3	228	-	7.3	1.43e-3	216	-	7.6	1.87e-3	232	-	8.1
$p = 1.01$	1.34e-3	233	-	7.2	3.52e-3	218	-	7.5	8.91e-4	223	-	7.9
$p = 1.1$	6.06e-4	244	-	7.3	7.02e-4	238	-	7.5	1.61e-3	233	-	8.0
$p = 2.1$	5.70e-7	375	-	7.1	3.19e-6	366	-	7.5	4.59e-6	322	-	8.3
$p = 3.1$	9.27e-8	143	475	3.5	6.56e-8	81	270	2.1	3.66e-8	80	303	2.6
iPPA ( $\Upsilon = 10^{-3}$ )												
$p = 1.001$	7.81e-7	549	-	7.3	8.26e-8	330	643	4.8	7.65e-8	466	871	7.0
$p = 1.01$	3.47e-7	552	-	7.3	9.78e-8	304	582	4.3	9.99e-8	299	596	4.6
$p = 1.1$	4.53e-7	541	-	7.4	6.52e-8	227	455	3.3	7.67e-8	219	452	3.6
$p = 2.1$	4.53e-7	444	-	7.6	9.16e-8	125	327	2.5	3.87e-8	24	154	1.3
$p = 3.1$	8.45e-7	291	-	7.7	9.16e-8	125	398	3.2	8.79e-8	15	188	1.5
V-iPPA ( $\Upsilon = 10^{-3}$ )												
$p = 1.001$	1.76e-6	368	-	7.3	1.69e-7	377	-	7.8	9.15e-8	301	874	7.4
$p = 1.01$	4.93e-7	380	-	7.2	8.88e-8	338	880	6.8	3.11e-7	361	-	8.3
$p = 1.1$	3.52e-7	371	-	7.2	6.85e-8	360	925	7.3	9.85e-8	163	513	4.3
$p = 2.1$	9.28e-8	140	583	4.1	9.63e-8	55	236	1.8	6.11e-8	28	176	1.5
$p = 3.1$	9.58e-8	139	775	5.9	8.69e-8	45	259	2.0	5.78e-8	15	154	1.3
HPE												
$\sigma = 0.999$	7.58e-7	328	-	10.6	9.16e-8	125	412	4.4	9.40e-8	13	162	1.8
$\sigma = 0.99$	7.58e-7	326	-	10.5	9.16e-8	125	416	4.4	9.40e-8	13	162	1.8
$\sigma = 0.9$	7.58e-7	319	-	10.4	9.16e-8	125	423	4.5	9.40e-8	13	180	1.9
$\sigma = 0.5$	8.29e-7	292	-	10.5	9.16e-8	125	450	4.8	9.17e-8	13	201	2.1
$\sigma = 0.1$	1.44e-6	217	-	10.6	9.16e-8	125	514	5.6	9.31e-8	13	277	2.9
AHPE												
$\sigma = 0.999$	9.78e-8	105	697	7.2	7.78e-8	33	251	2.7	8.40e-8	11	187	2.0
$\sigma = 0.99$	9.78e-8	105	697	7.1	7.78e-8	33	251	2.6	8.40e-8	11	187	2.0
$\sigma = 0.9$	9.56e-8	105	696	7.1	7.49e-8	33	259	2.7	8.89e-8	11	170	1.9
$\sigma = 0.5$	9.01e-8	105	751	7.7	7.76e-8	33	266	2.8	9.89e-8	10	225	2.3
$\sigma = 0.1$	8.99e-8	105	874	9.2	7.71e-8	33	309	3.3	9.00e-8	10	300	3.1

## 7 Concluding remarks

In this paper, we propose a new inexact Bregman proximal point algorithm (iBPPA) for solving a general class of convex problems. Compared to existing iBPPAs, we introduce a more flexible stopping condition for solving the subproblems to circumvent the underlying feasibility issue that often appears, but overlooked, in existing inexact conditions when the problem has a complicated feasible set. Our inexact condition also covers some existing inexact conditions as special cases. The iteration complexity of  $O(1/k)$  and the convergence of the sequence are established for our iBPPA under some mild conditions. In addition, we successfully develop an

Table 2: Comparisons among iEPPA, V-iEPPA and BHPE. In the table, “out#” denotes the number of outer iterations, “sink#” denotes the the number of Sinkhorn iterations, and “-” means that the number of Sinkhorn iterations reaches 10000.

method	$\gamma = 1$				$\gamma = 0.1$				$\gamma = 0.01$			
	kkt	out#	sink#	time	kkt	out#	sink#	time	kkt	out#	sink#	time
iEPPA ( $\Upsilon = 1$ )												
$p = 1.001$	1.00e-5	5800	5800	46.1	1.00e-5	581	581	4.5	1.00e-5	747	2383	13.7
$p = 1.01$	1.00e-5	5800	5800	45.3	1.00e-5	581	581	4.5	9.94e-6	709	2311	13.2
$p = 1.1$	1.00e-5	5800	5800	45.2	9.99e-6	584	586	4.5	9.96e-6	428	1511	8.5
$p = 2.1$	5.51e-5	1860	-	53.1	9.98e-6	581	5808	29.2	9.96e-6	59	1149	5.7
$p = 3.1$	8.53e-4	308	-	48.5	1.16e-4	116	-	48.5	1.01e-5	58	-	49.0
V-iEPPA ( $\Upsilon = 1$ )												
$p = 1.001$	9.95e-6	1007	1007	15.3	3.79e-4	963	-	56.1	4.71e-5	432	-	51.4
$p = 1.01$	9.95e-6	1007	1007	15.1	1.64e-4	912	-	55.8	1.53e-4	376	-	50.2
$p = 1.1$	9.95e-6	1007	1007	15.2	8.16e-5	994	-	56.2	7.33e-5	276	-	49.0
$p = 2.1$	9.99e-6	1029	3192	25.7	9.99e-6	106	1920	10.3	9.37e-6	34	3345	16.0
$p = 3.1$	5.29e-5	246	-	50.3	9.78e-6	106	8710	42.7	9.63e-6	19	2024	9.8
iEPPA ( $\Upsilon = 10^{-3}$ )												
$p = 1.001$	1.81e-5	4195	-	59.9	1.42e-5	485	-	49.4	9.86e-6	59	8559	40.5
$p = 1.01$	2.02e-5	3838	-	58.4	1.56e-5	462	-	48.8	9.86e-6	59	8720	41.1
$p = 1.1$	6.09e-5	1747	-	52.4	2.72e-5	293	-	48.6	1.05e-5	57	-	47.0
$p = 2.1$	1.13e-3	255	-	44.0	2.00e-4	81	-	45.6	5.11e-5	20	-	48.7
$p = 3.1$	1.85e-3	184	-	18.3	3.52e-4	56	-	28.8	7.35e-5	16	-	39.0
V-iEPPA ( $\Upsilon = 10^{-3}$ )												
$p = 1.001$	9.99e-6	1031	3281	26.2	1.00e-5	104	3240	16.7	8.87e-6	20	4206	20.0
$p = 1.01$	1.00e-5	1028	3422	26.7	1.00e-5	104	3305	17.0	8.86e-6	20	4247	20.2
$p = 1.1$	1.00e-5	1029	5010	34.4	9.94e-6	105	4154	21.0	8.80e-6	20	4664	22.1
$p = 2.1$	7.85e-5	181	-	48.6	2.65e-5	47	-	48.0	1.14e-5	18	-	48.7
$p = 3.1$	2.05e-4	110	-	25.7	4.38e-5	35	-	34.2	1.80e-5	13	-	44.0
BHPE												
$\sigma = 0.999$	1.00e-5	5800	5800	55.4	9.99e-6	603	657	6.0	9.23e-6	187	3932	25.1
$\sigma = 0.99$	1.00e-5	5800	6259	58.2	9.98e-6	580	1150	9.1	9.94e-6	59	1280	8.2
$\sigma = 0.9$	2.46e-5	3163	-	72.8	9.98e-6	581	5206	34.9	9.80e-6	59	1894	12.2
$\sigma = 0.5$	1.34e-4	1051	-	68.4	2.65e-5	299	-	65.0	9.85e-6	59	5910	37.6
$\sigma = 0.1$	7.56e-4	334	-	67.9	1.41e-4	102	-	65.7	3.33e-5	26	-	63.8

inertial variant of our iBPPA (denoted by V-iBPPA) based on Nesterov’s acceleration technique. Specifically, when the proximal parameter  $\gamma_k$  satisfies that  $0 < \underline{\gamma} \leq \gamma_k \leq \bar{\gamma} < \infty$ , the V-iBPPA enjoys an iteration complexity of  $O(1/k^\lambda)$ , where  $\lambda \geq 1$  is a quadrangle scaling exponent of the kernel function. Thus, if  $\lambda$  is strictly larger than 1, the V-iBPPA achieves acceleration. Some preliminary experiments for solving the standard OT problem are conducted to illustrate the influence of the inexact settings on the convergence behaviors of our iBPPA and V-iBPPA. The experiments also empirically verify the potential of the V-iBPPA on improving the convergence speed.

## Acknowledgments

We thank the editor and referees for their valuable suggestions and comments, which have helped to improve the quality of this paper. We also thank Professor Yair Censor for bringing to our attention the references [10, 13] on Bregman functions.

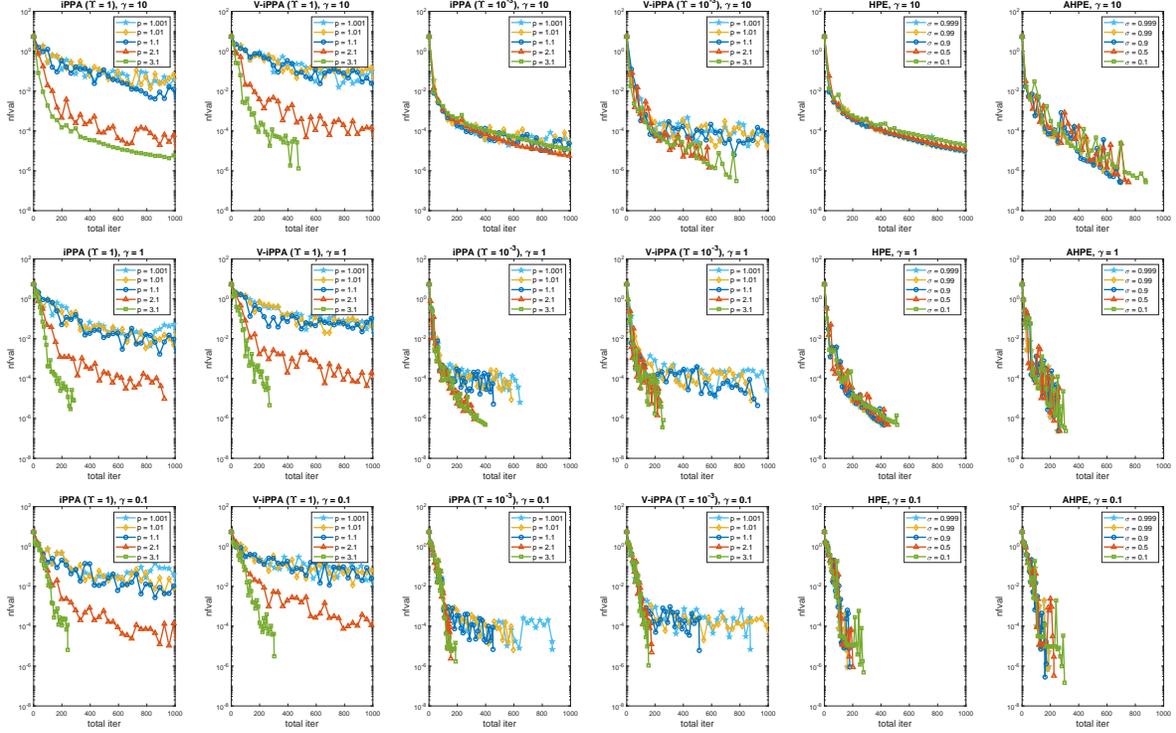


Figure 2: Comparisons among iPPA, V-iPPA, HPE and AHPE.

## 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 30*, pages 1964–1974, 2017.
- [2] A. Auslender and M. Haddou. An interior-proximal method for convex linearly constrained problems and its extension to variational inequalities. *Math. Program.*, 71(1):77–100, 1995.
- [3] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim.*, 16(3):697–725, 2006.
- [4] 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.
- [5] H.H. Bauschke and J.M. Borwein. Legendre functions and the method of random Bregman projections. *J. Convex Anal.*, 4(1):27–67, 1997.
- [6] 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.
- [7] 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.

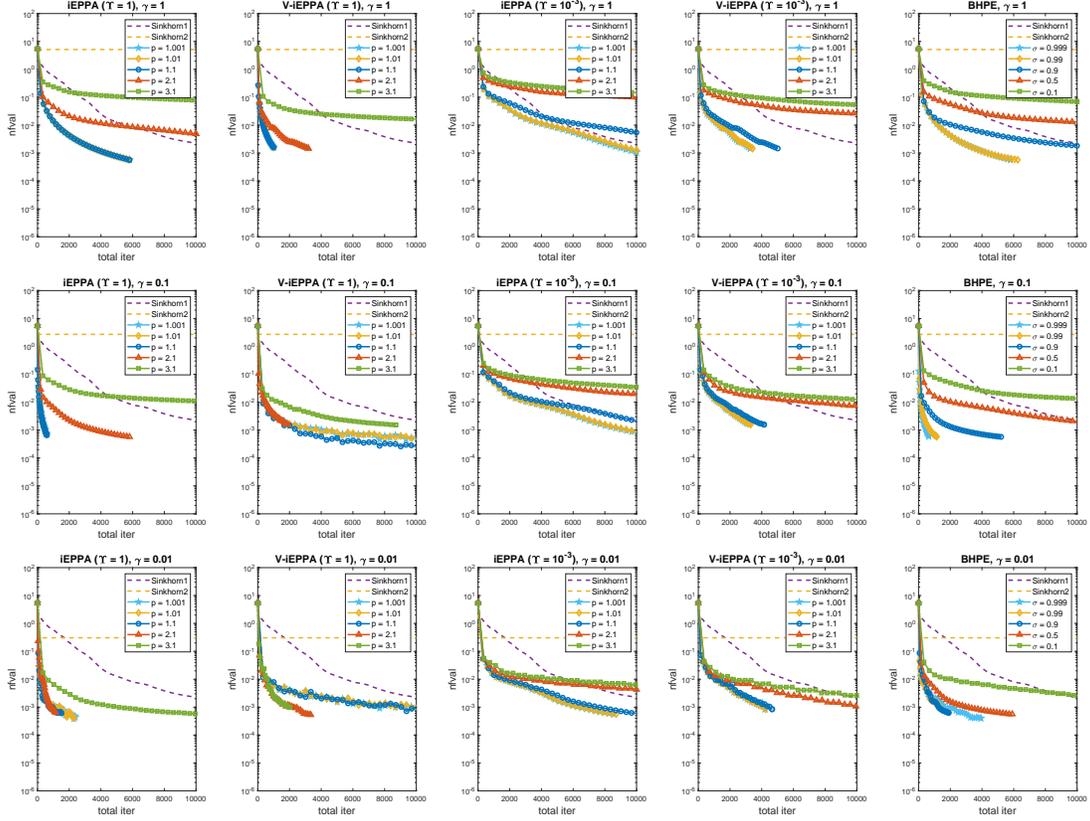


Figure 3: Comparisons among iEPPA, V-iEPPA and BHPE. As benchmarks, “Sinkhorn1” and “Sinkhorn2” denotes Sinkhorn’s algorithm with  $\gamma = 10^{-4}$  and  $\gamma$  given in each title, respectively.

[8] 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.

[9] R.S. Burachik, A.N. Iusem, and B.F. Svaiter. Enlargement of monotone operators with applications to variational inequalities. *Set-Valued Analysis*, 5(2):159–180, 1997.

[10] D. Butnariu, C. Byrne, and Y. Censor. Redundant axioms in the definition of Bregman functions. *J. Convex Anal.*, 10(1):245–254, 2003.

[11] Y. Censor and A. Lent. An iterative row-action method for interval convex programming. *J. Optim. Theory Appl.*, 34(3):321–353, 1981.

[12] Y. Censor and S.A. Zenios. Proximal minimization algorithm with  $D$ -functions. *J. Optim. Theory Appl.*, 73(3):451–464, 1992.

[13] Y. Censor and S.A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, New York, 1997.

[14] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM J. Optim.*, 3(3):538–543, 1993.

- [15] 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.
- [16] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Stud. Sci. Math. Hung.*, 2:229–318, 1967.
- [17] J. Eckstein. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Math. Oper. Res.*, 18(1):202–226, 1993.
- [18] J. Eckstein. Approximate iterations in Bregman-function-based proximal algorithms. *Math. Program.*, 83(1-3):113–123, 1998.
- [19] P.P.B Eggermont. Multiplicative iterative algorithms for convex programming. *Linear Algebra Appl.*, 130:25–42, 1990.
- [20] O. Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM J. Control Optim.*, 29(2):403–419, 1991.
- [21] O. Güler. New proximal point algorithms for convex minimization. *SIAM J. Optim.*, 2(4):649–664, 1992.
- [22] Inc. Gurobi Optimization. Gurobi Optimizer Reference Manual, 2018.
- [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] A.J. Hoffman. On approximate solutions of systems of linear inequalities. *J. Res. Natl. Bur. Stand.*, 49(4):263–265, 1952.
- [26] A.N. Iusem, B.F. Svaiter, and M. Teboulle. Entropy-like proximal methods in convex programming. *Math. Oper. Res.*, 19(4):790–814, 1994.
- [27] A.N. Iusem and M. Teboulle. Convergence rate analysis of nonquadratic proximal methods for convex and linear programming. *Math. Oper. Res.*, 20(3):657–677, 1995.
- [28] K.C. Kiwiel. Proximal minimization methods with generalized Bregman functions. *SIAM J. Control Optim.*, 35(4):1142–1168, 1997.
- [29] G. Lan, Z. Lu, and R.D.C. Monteiro. Primal-dual first-order methods with  $\mathcal{O}(1/\epsilon)$  iteration-complexity for cone programming. *Math. Program.*, 126(1):1–29, 2011.
- [30] B Lemaire. On the convergence of some iterative methods for convex minimization. In *Recent Developments in Optimization*, pages 252–268. Springer, 1995.
- [31] X. Li, D.F. Sun, and K.-C. Toh. On the efficient computation of a generalized Jacobian of the projector over the Birkhoff polytope. *Math. Program.*, 179(1-2):419–446, 2020.
- [32] H. Lin, J. Mairal, and Z. Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *J. Mach. Learn. Res.*, 18(1):7854–7907, 2017.

- [33] H. Lu, R.M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim.*, 28(1):333–354, 2018.
- [34] B. Martinet. Régularisation d’inéquations variationnelles par approximations successives. *Rev. Française Inf. Rech. Oper.*, 4:154–159, 1970.
- [35] R.D.C. Monteiro and B.F. Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM J. Optim.*, 20(6):2755–2787, 2010.
- [36] R.D.C. Monteiro and B.F. Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM J. Optim.*, 23(2):1092–1125, 2013.
- [37] J.-J. Moreau. Proximité et dualité dans un espace Hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.
- [38] Y. Nesterov. A method of solving a convex programming problem with convergence rate  $O(1/k^2)$ . *Sov. Math. Dokl.*, 27(2):372–376, 1983.
- [39] 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.
- [40] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2003.
- [41] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.
- [42] G. Peyré and M. Cuturi. Computational optimal transport. *Found. Trends Mach. Learn.*, 11(5-6):355–607, 2019.
- [43] B.T. Polyak. *Introduction to optimization*. Optimization Software Inc., New York, 1987.
- [44] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.
- [45] R.T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Math. Oper. Res.*, 1(2):97–116, 1976.
- [46] R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.*, 14(5):877–898, 1976.
- [47] R.T. Rockafellar and R.J-B. Wets. *Variational Analysis*. Springer, 1998.
- [48] 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.
- [49] M.V. Solodov and B.F. Svaiter. A hybrid approximate extragradient – proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, 7(4):323–345, 1999.
- [50] M.V. Solodov and B.F. Svaiter. A hybrid projection-proximal point algorithm. *J. Convex Anal.*, 6(1):59–70, 1999.

- [51] M.V. Solodov and B.F. Svaiter. Error bounds for proximal point subproblems and associated inexact proximal point algorithms. *Math. Program.*, 88(2):371–389, 2000.
- [52] M.V. Solodov and B.F. Svaiter. An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions. *Math. Oper. Res.*, 25(2):214–230, 2000.
- [53] M.V. Solodov and B.F. Svaiter. A unified framework for some inexact proximal point algorithms. *Numer. Funct. Anal. Optim.*, 22(7-8):1013–1035, 2001.
- [54] M. Teboulle. Entropic proximal mappings with applications to nonlinear programming. *Math. Oper. Res.*, 17(3):670–690, 1992.
- [55] M. Teboulle. Convergence of proximal-like algorithms. *SIAM J. Optim.*, 7(4):1069–1083, 1997.
- [56] M. Teboulle. A simplified view of first order methods for optimization. *Math. Program.*, 170(1):67–96, 2018.
- [57] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *Technical report*, 2008.
- [58] P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Math. Program.*, 125(2):263–295, 2010.
- [59] S. Villa, S. Salzo, L. Baldassarre, and A. Verri. Accelerated and inexact forward-backward algorithms. *SIAM J. Optim.*, 23(3):1607–1633, 2013.
- [60] Y. Xie, X. Wang, R. Wang, and H. Zha. A fast proximal point method for computing exact Wasserstein distance. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, pages 433–453, 2020.
- [61] S. Yan and N. He. Bregman augmented Lagrangian and its acceleration. *arXiv:2002.06315*, 2020.