Bregman Douglas-Rachford Splitting Method*

Shiqian Ma[1], Lin Xiao[2], Renbo Zhao[3]

[1]Rice University, [2], [3]University of Iowa

First version: Nov 5, 2021
This version: Sep 4, 2025†

Abstract

In this paper, we propose the Bregman Douglas-Rachford splitting (BDRS) method and its variant Bregman Peaceman-Rachford splitting method for solving maximal monotone inclusion problem. We show that BDRS is equivalent to a Bregman alternating direction method of multipliers (ADMM) when applied to the dual of the problem. A special case of the Bregman ADMM is an alternating direction version of the exponential multiplier method. To the best of our knowledge, algorithms proposed in this paper are new to the literature. We also discuss how to use our algorithms to solve the discrete optimal transport (OT) problem. We prove the convergence of the algorithms under certain assumptions, though we point out that one assumption does not apply to the OT problem.

1 Introduction

This paper studies the celebrated Douglas-Rachford splitting method (DRS) that has a long history dating back to 1956 for solving variational problems arising from numerical PDEs [25]. The DRS later on became a very popular method for finding zeros of the sum of maximal monotone operators:

$$\text{Find } x, \text{ s.t., } 0 \in A(x) + B(x), \tag{1.1}$$

where $A : \mathbb{R}^n \to \mathbb{R}^n$ and $B : \mathbb{R}^n \to \mathbb{R}^n$ are two maximal monotone operators. The problem (1.1) is also known as the monotone inclusion problem. The solution set of (1.1) is denoted as $(A+B)^{-1}(0)$. The DRS has been studied by many researchers under various settings [57, 37, 36, 41, 35, 26, 30, 6, 11, 22, 83, 54, 47, 77, 78, 55, 79, 17, 1]. When A and B are normal cone operators, the problem (1.1) reduces to a feasibility problem which seeks for a point in the intersection of two sets. We refer to the recent survey paper [56] for more details on DRS for feasibility problems. The DRS for

*The three authors started to work on this problem back in 2020, and this draft was ready in Nov 2021. We spent the last four years trying to prove the convergence of the algorithms under the most general setting, but did not succeed. Despite of that, we believe that the current contributions are significant. So we decided to share the results with the community now. Since this draft is from 2021, we apologize that the references are not up to date. We will include more recent works in a later version of the paper.

†The results in Sections 1-5 were written by Shiqian Ma in 2020, and the results in the Appendix were written by Shiqian Ma in Nov 2021. These results formed the first version of the paper which was ready on Nov 5, 2021. Shiqian Ma polished the writing in Sep 2025, which gives the current version.

solving (1.1) can be written as:

$$x^k := J_{\gamma_k B}(z^k) \tag{1.2a}$$
$$y^k := J_{\gamma_k A}(2x^k - z^k) \tag{1.2b}$$
$$z^{k+1} := z^k - x^k + y^k, \tag{1.2c}$$

where $J_T := (I+T)^{-1}$ is called the resolvent of the operator T, and $\gamma_k > 0$ is a parameter.

A well-known application of the DRS is the so-called alternating direction method of multipliers (ADMM), which is usually applied to solving the following convex minimization problem:

$$\min_{u,v} \ f(u) + g(v), \ \text{s.t.}, \ Mu + Nv = b, \ u \in \mathbb{R}^p, v \in \mathbb{R}^q, \tag{1.3}$$

where $b \in \mathbb{R}^m$, $M \in \mathbb{R}^{m\times p}$ and $N \in \mathbb{R}^{m\times q}$, and f and g are both proper, closed and convex functions. The ADMM for solving (1.3) updates the iterates as follows:

$$u^{k+1} := \operatorname*{argmin}_u \ \mathcal{L}_\beta(u, v^k; w^k) \tag{1.4a}$$
$$v^{k+1} := \operatorname*{argmin}_v \ \mathcal{L}_\beta(u^{k+1}, v; w^k) \tag{1.4b}$$
$$w^{k+1} := w^k + \beta(Mu^{k+1} + Nv^{k+1} - b). \tag{1.4c}$$

Here

$$\mathcal{L}_\beta(u, v; w) := f(u) + g(v) + \langle w, Mu + Nv - b\rangle + \frac{\beta}{2}\|Mu + Nv - b\|_2^2, \tag{1.5}$$

is the augmented Lagrangian function for (1.3), where w denotes the Lagrange multiplier, and $\beta > 0$ is a penalty parameter. As proved by Gabay [36], ADMM (1.4) for solving (1.3) is a special case of DRS (1.2) applied to solving the dual problem of (1.3), whose optimality condition is in the form of (1.1). More specifically, the dual problem of (1.3) is given by

$$\min_x \ f^*(-M^\top x) + g^*(-N^\top x) + b^\top x, \tag{1.6}$$

where f^* and g^* are the conjugate functions of f and g, respectively. The optimality condition of (3.6) is:

$$0 \in -M\partial f^*(-M^\top x) - N\partial g^*(-N^\top x) + b, \tag{1.7}$$

which is in the form of (1.1) by defining $A(x) = -M\partial f^*(-M^\top x)$ and $B(x) = -N\partial g^*(-N^\top x) + b$. Appying DRS (1.2) to (1.7) gives the ADMM (1.4).

ADMM has received significant attention due to its applications in signal processing, image processing, semidefinite programming and statistics [22, 43, 90, 91, 59]. It is not possible to exhaust the vast literature on ADMM, and we thus refer the reader to the following survey papers for more details on the theory and applications of ADMM and its variants [14, 31, 40, 60].

The efficiency of the DRS (1.2) relies on the assumption that both $J_{\gamma_k A}(z)$ and $J_{\gamma_k B}(z)$ can be computed easily, and similarly, the efficiency of the ADMM (1.4) replies on the assumption that the two minimization subproblems are easy to solve. As we will discuss later, in certain applications, these computations are not easy (even when $M = N = I$) and more general Bregman distances need to be considered when designing these algorithms. However, to the best of our knowledge, Bregman DRS (BDRS) was not considered in the literature before. There exists one work on Bregman ADMM [89], but this algorithm only applies to some special class of problems and is different from the algorithms that we consider in this paper.

Our contributions. In this paper, we target to design the BDRS algorithm, and our main contributions are as follows.

(i) We design the first BDRS algorithm in the literature, and analyze its connections with several existing methods. We also propose a Bregman Peaceman-Rachford splitting (BPRS) method, which is a close variant of BDRS. We show that BDRS is equivalent to a Bregman ADMM algorithm when applied to its dual.

(ii) We show that if the Bregman distance is generated by the Boltzmann-Shannon entropy, then when applied to the dual problem of linear inequality constrained convex programming problem, our BDRS gives an alternating direction version of the exponential multiplier method. We name this algorithm ADEMM, and this is also a new algorithm, to the best of our knowledge.

(iii) We discuss how to use our BDRS and ADEMM to solve the discrete optimal transport (OT) problem, and discuss how they relate to and why they are better than the Sinkhorn's algorithm.

(iv) We prove the convergence of BDRS under certain assumptions, though we want to point out that one of the assumptions does not apply to the OT problem.

Organization. The rest of this paper is organized as follows. In Section 2 we provide some preliminaries on Bregman distance and algorithms based on it. In Section 3, we propose our BDRS and ADEMM algorithms and discuss their connections with existing algorithms. In Section 4, we propose the BPRS algorithm. In Section 5, we discuss how to apply our proposed algorithms to solving the discrete optimal transport problem. We draw some conclusions in Section 6. In the appendix, we provide the convergence analysis for BPRS and BDRS, both under certain assumptions.

2 Preliminaries and Existing Bregman Algorithms

In this section, we briefly review the basics of Bregman distance and existing algorithms that use Bregman distance. The Bregman distance was first proposed by Bregman [15] in a primal-dual method for solving linearly constrained convex programming problems that involves non-orthogonal projections onto hyperplanes. This method was further studied by Censor and Lent [18], and De Pierro and Iusem [24].

We now introduce the notions of Legendre function and Bregman distance.

Definition 2.1 ([74])**.** *A function h is called a Legendre function, if it is proper, lower semicontinuous, strictly convex and essentially smooth.*

A Legendre function enjoys the following two useful properties:

(i) h is Legendre if and only if its conjugate h^* is Legendre.

(ii) The gradient of a Legendre function h is a bijection from int dom$h$ to int domh^*, and its inverse is the gradient of the conjugate, that is, we have $(\nabla h)^{-1} = \nabla h^*$.

Definition 2.2. *For a Legendre function h, the Bregman distance corresponding to h is defined as*

$$D_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle. \tag{2.1}$$

The following Bregman distances are commonly seen in practice (for more examples, see [52]).

Example 2.1. *(i) Energy: If $h(x) = \frac{1}{2}\|x\|_2$, then $D_h(x, y) = \frac{1}{2}\|x-y\|_2^2$ is the Euclidean distance.*

(ii) Quadratic: If $h(x) = \frac{1}{2}x^\top Lx$ *with matrix* $L \succ 0$*, then* $D_h(x,y) = \frac{1}{2}\|x-y\|_L^2 = \frac{1}{2}(x-y)^\top L(x-y)$.

(iii) Boltzmann-Shannon entropy: If $h(x)$ *is the Boltzmann-Shannon entropy function defined as* $h(x) = \sum_i x_i(\log x_i - 1)$*, then* $D_h(x,y) = \sum_i x_i \log \frac{x_i}{y_i} - x_i + y_i$ *is the Kullback-Leibler (KL) divergence. Note that the domain of* h *is* $\mathrm{dom}\, h = \mathbb{R}^n_{++} = \{x \mid x > 0\}$*. Moreover,* $h^*(x) = \sum_i e^{x_i}$.

(iv) Burg's entropy: If $h(x) = -\sum_i \log x_i$*, then* $D_h(x,y) = -\sum_i \log \frac{x_i}{y_i} + \frac{x_i}{y_i} - 1$*. Note that the domain of* h *is* $\mathrm{dom}\, h = \mathbb{R}^n_{++}$.

We now define a few useful Bregman operators.

Definition 2.3 (Bregman forward operator, Bregman resolvent operator, Bregman reflection operator, Bregman Mann's operator)**.** *Use* h *to denote a Legendre function. The Bregman forward operator for a single-valued operator* T *is defined as*

$$F_T^h := \nabla h^* \circ (\nabla h - T). \tag{2.2}$$

The Bregman resolvent operator of a maximal monotone operator T *is defined as [27]*

$$J_T^h := (\nabla h + T)^{-1} \circ \nabla h. \tag{2.3}$$

The Bregman reflection operator of a maximal monotone operator T *is defined as*

$$R_T^h := \nabla h^* \circ (2\nabla h \circ J_T^h - \nabla h), \tag{2.4}$$

and the Bregman Mann's operator of a maximal monotone operator T *is defined as*

$$M_\alpha^h(T) := \nabla h^* \circ (\alpha \nabla h + (1-\alpha)\nabla h \circ T), \tag{2.5}$$

The notion of Bregman resolvent operator (2.3) was first proposed by Eckstein in [27]. When $h(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, J_T^h reduces to $J_T := (I+T)^{-1}$, because $\nabla h = I$. The definitions of Bregman relection operator (2.4) and Bregman Mann's operator (2.5) are new to the literature to the best of our knowledge.

We now review several classes of Bregman algorithms for both monotone inclusion and convex optimization problems.

2.1 Bregman Gradient Method and Mirror Descent Method

For unconstrained convex minimization problem

$$\min_{x \in \mathcal{X}} \; f(x) \tag{2.6}$$

where f is convex and smooth and $\mathcal{X} \subseteq \mathbb{R}^n$ is a convex set, Nemirovski and Yudin [62] proposed the mirror descent algorithm which iterates as

$$x^{k+1} := \nabla h^*(\nabla h(x^k) - \gamma_k \nabla f(x^k)) \equiv F^h_{\gamma_k \nabla f}(x^k), \tag{2.7}$$

where $\gamma_k > 0$ is the step size, and h is a Legendre function. It was later showed by Beck and Teboulle [7] that the mirror descent algorithm (2.7) can be interpreted as a projected gradient method with a Bregman distance, which can be described as follows:

$$x^{k+1} := \operatorname*{argmin}_{x \in \mathcal{X}} \; \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{\gamma_k} D_h(x, x^k). \tag{2.8}$$

Here we discuss an important setting where S is the probability simplex $S := \{x \in \mathbb{R}^n \mid \sum_i x_i = 1, x \geq 0\}$. In this case, if one applies the projected gradient method using the Euclidean distance to solve (2.6), then each iteration requires a projection onto S. If one uses the Bregman distance generated by the Boltzmann-Shannon entropy, i.e., Example 2.1 (iii), then the solution of (2.8) is given by a simple normalization

$$x^{k+1} := \frac{y^k}{\|y^k\|_1}, \quad \text{with } y^k := x^k \circ e^{-\gamma_k \nabla f(x^k)},$$

which is easier to compute than the projection onto S. Here for vectors a and b, $a \circ b$ is the componentwise multipication, and e^a is the componentwise exponential function.

2.2 Bregman Proximal Gradient Method and Bregman Forward-Backward Splitting

For composite convex minimization problem

$$\min_{x \in S} \; f(x) + g(x) \tag{2.9}$$

where f is convex and smooth and g is convex and possibly nonsmooth, Bregman proximal gradient method was studied in [4, 58, 86], which iterates as:

$$x^{k+1} := \operatorname*{argmin}_x \langle \nabla f(x^k), x - x^k \rangle + g(x) + \frac{1}{\gamma_k} D_h(x, x^k) \tag{2.10a}$$

$$\equiv \operatorname{prox}^h_{\gamma_k g}(\nabla h^*(\nabla h(x^k) - \gamma_k \nabla f(x^k))) \equiv J^h_{\gamma_k \partial g} \circ F^k_{\gamma_k \nabla f}(x^k), \tag{2.10b}$$

where

$$\operatorname{prox}^h_g(z) := \operatorname*{argmin}_x \; g(x) + D_h(x, z) \equiv J^h_{\partial g}(z), \tag{2.11}$$

is called the Bregman proximal map of g with respect to h. An interesting question is how to accelerate the Bregman proximal gradient method using Nesterov's acceleration techniques [63, 64, 65, 8]. When f has an globally Lipschitz gradient, and h is strongly convex, faster algorithms have been given by Auslender and Teboulle [2], and Tseng [86]. However, when these assumptions are weakened, this problem has not been fully addressed in the literature. We refer to [46, 45] for some recent progresses on this topic and the recent paper by Teboulle [85] for more detailed discussions on Bregman proximal gradient method.

When it comes to the monotone inclusion problem (1.1) with B being single-valued, the Bregman forward-backward splitting methods are studied in the literature, which iterates as

$$x^{k+1} := (\nabla h + \gamma_k A)^{-1}(\nabla h(x^k) - \gamma_k B(x^k)) \equiv J^h_{\gamma_k A} \circ F^h_{\gamma_k B}(x^k). \tag{2.12}$$

We refer to the recent paper by Bùi and Combettes [16] and references therein for more discussions on this method.

2.3 Bregman Proximal Point Method and Bregman Augmented Lagrangian method

Another widely used Bregman algorithm is the Bregman proximal point method (PPM). The idea of Bregman PPM can be traced back to [34, 33, 32]. The Bregman PPM in its current form was proposed by Censor and Zenios [19] for convex minimization problem and by Eckstein [27] for

monotone inclusion problem. This method was further studied in [20, 48, 9, 84, 3, 50, 51, 5]. For the maximal monotone inclusion problem $0 \in T(x)$, the Bregman PPM iterates as

$$x^{k+1} := J^h_{\gamma_k T}(x^k) = (\nabla h + \gamma_k T)^{-1} \circ \nabla h(x^k). \tag{2.13}$$

For convex minimization problem (2.6) with f being nonsmooth, the Bregman PPM reduces to

$$x^{k+1} := J^h_{\gamma_k \partial f}(x^k) = \operatorname*{argmin}_x \ f(x) + \frac{1}{\gamma_k} D_h(x, x^k). \tag{2.14}$$

This algorithm generalizes the PPM in Euclidean space [61, 76, 75, 49] to non-quadratic distance. Since it is usually difficult to solve the subproblem in (2.13) and (2.14) exactly, inexact Bregman PPM is studied in the literature [28, 82, 93]. Moreover, the exponential multiplier method proposed by Kort and Bertsekas [53, 87] is known to be a special case of Bregman PPM with a specifically chosen h. We will discuss this method in more details in Section 3.3. Furthermore, the nonlinear rescaling method developed by Polyak [69, 70, 71, 44] is also known to be equivalent to a Bregman PPM with suitably chosen distance and nonlinear penalty functions, as proved by Polyak and Teboulle [72].

Similar to the connections between PPM and augmented Lagnrangain method (ALM) in the Euclidean case, there is a similar connection between Bregman PPM and Bregman ALM, as illustrated by Eckstein [27]. Here we use the following convex optimization problem with linear equality constraint to illustrate the idea:

$$\min_x \ f(x), \ \text{s.t.}, \ Mx = b, x \in \mathcal{X}. \tag{2.15}$$

It can be shown that the Bregman PPM for solving the dual problem of (2.15) is equivalent to the following Bregman ALM for solving (2.15):

$$x^{k+1} \in \operatorname*{argmin}_{x \in \mathcal{X}} \ f(x) + \frac{1}{\gamma_k} h^*(\nabla h(\lambda^k) + \gamma_k(Mx - b)) \tag{2.16a}$$

$$\lambda^{k+1} := \nabla h^*(\nabla h(\lambda^k) + \gamma_k(Mx^{k+1} - b)). \tag{2.16b}$$

One can see that unlike the classical ALM with a quadratic penalty function, the Bregman ALM (2.16) adopts a non-quadratic penalty function h^*. A classical reference on Bregman ALM is due to Bertsekas [10], and a more recent survey is due to Iusem [49]. Some very recent works on this topic include [29, 92].

2.4 Bregman ADMM

A natural extension of the ADMM in Euclidean space is the Bregman ADMM. A Bregman ADMM was studied by Wang and Banerjee [89], in which the authors targeted solving (1.3) using the following Bregman ADMM:

$$u^{k+1} := \operatorname*{argmin}_y \ f(y) + \langle w^k, Mu + Nv^k - b\rangle + \gamma D_h(b - Mu, Nv^k) \tag{2.17a}$$

$$v^{k+1} := \operatorname*{argmin}_z \ g(z) + \langle w^k, Mu^{k+1} + Nv - b\rangle + \gamma D_h(Nv, b - Mu^{k+1}) \tag{2.17b}$$

$$w^{k+1} := w^k + \gamma(Mu^{k+1} + Nv^{k+1} - b). \tag{2.17c}$$

Note that this algorithm requires Nv^k and $b - Mu^{k+1}$ to lie in the domain of h, which is very difficult to guarantee for many useful Bregman distances such as the ones generated by the Boltzmann-Shannon entropy and the Burg's entropy. Thus the applicability of (2.17) is limited.

3 Bregman Douglas-Rachford Splitting Method

In this section, we introduce our BDRS for solving (1.1). A typical iteration of our BDRS for solving (1.1) is as follows:

$$x^k := J^h_{\gamma_k B}(z^k) \tag{3.1a}$$

$$y^k := J^h_{\gamma_k A} \circ \nabla h^*(2\nabla h(x^k) - \nabla h(z^k)) \tag{3.1b}$$

$$z^{k+1} := \nabla h^*(\nabla h(z^k) - \nabla h(x^k) + \nabla h(y^k)), \tag{3.1c}$$

where $\gamma_k > 0$ is a parameter. We now provide some explainations to our BDRS (3.1). Note that the Bregman resolvent operators are always taken to the points in the primal space. For points in the mirror space, we always use ∇h^* to transform them back to the primal space. Note that (3.1) can be written equivalently as

$$\nabla h(z^{k+1}) = \nabla h(z^k) + \nabla h \circ J^h_{\gamma_k A} \circ \nabla h^*(2\nabla h \circ J^h_{\gamma_k B}(z^k) - \nabla h(z^k)) - \nabla h \circ J^h_{\gamma_k B}(z^k). \tag{3.2}$$

Using the Bregman reflection operator (2.4) and the Bregman Mann's operator (2.5), the BDRS (3.2) can be written more compactly as

$$z^{k+1} := M^h_{\frac{1}{2}}(R^h_{\gamma_k A} R^h_{\gamma_k B})(z^k). \tag{3.3}$$

We notice that when $h(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, all the three forms of BDRS, i.e., (3.1), (3.2), and (3.3) reduce exactly to their Euclidean counterparts.

3.1 Application to Convex Minimization: Bregman ADMM

First, we note that if one applies the Bregman ALM to solve the convex optimization problem with linear equality constraints (1.3), then it should iterate as follows:

$$(u^k, v^k) := \operatorname*{argmin}_{u,v} \; f(u) + g(v) + \frac{1}{\gamma_k} h^*(\nabla h(w^k) + \gamma_k(Mu + Nv - b)) \tag{3.4a}$$

$$w^{k+1} := \nabla h^*(\nabla h(w^k) + \gamma_k(Mu^k + Nv^k - b)). \tag{3.4b}$$

As a result, it is easy to see that the Bregman ADMM for solving (1.3) is given by the following updates:

$$u^k := \operatorname*{argmin}_{u} \; f(u) + \frac{1}{\gamma_k} h^*(\nabla h(w^k) + \gamma_k(Mu + Nv^{k-1} - b)) \tag{3.5a}$$

$$v^k := \operatorname*{argmin}_{v} \; g(v) + \frac{1}{\gamma_k} h^*(\nabla h(w^k) + \gamma_k(Mu^k + Nv - b)) \tag{3.5b}$$

$$w^{k+1} := \nabla h^*(\nabla h(w^k) + \gamma_k(Mu^k + Nv^k - b)), \tag{3.5c}$$

where we alternatingly update u^k and v^k in Bregman ALM with the other variable being fixed. To the best of our knowledge, both BDRS (3.1) and Bregman ADMM (3.5) are new to the literature. In the following, we show that the Bregman ADMM (3.5) is actually a direct application of BDRS (3.1) to the dual of (1.3), which is given by

$$\max_w \min_{u,v} f(u) + g(v) + \langle w, Mu + Nv - b\rangle \equiv \max_w -f^*(-M^\top w) - g^*(-N^\top w) - \langle b, w\rangle \tag{3.6a}$$

$$\equiv \min_w f^*(-M^\top w) + g^*(-N^\top w) + \langle b, w\rangle, \tag{3.6b}$$

where w is the dual variable (Lagrange multiplier), and f^* and g^* are the conjugate functions of f and g, respectively. The optimality condition of (3.6) is given by

$$0 \in -M\partial f^*(-M^\top w) - N\partial g^*(-N^\top w) + b, \tag{3.7}$$

which is in the from of (1.1) with $A(w) = -M\partial f^*(-M^\top w)$ and $B(w) = -N\partial g^*(-N^\top w) + b$.

Theorem 3.1. *The BDRS* (3.1) *for solving the dual problem* (3.7) *is equivalent to the Bregman ADMM* (3.5) *for solving the primal problem* (1.3).

Proof. With $A(w) = -M\partial f^*(-M^\top w)$ and $B(w) = -N\partial g^*(-N^\top w) + b$, the BDRS for solving (3.7) can be written as

$$x^k := \operatorname*{argmin}_x \; g^*(-N^\top x) + b^\top x + \frac{1}{\gamma_k} D_h(x, z^k) \tag{3.8a}$$

$$y^k := \operatorname*{argmin}_y \; f^*(-M^\top y) + \frac{1}{\gamma_k} D_h(y, \nabla h^*(2\nabla h(x^k) - \nabla h(z^k))) \tag{3.8b}$$

$$z^{k+1} := \nabla h^*(\nabla h(z^k) - \nabla h(x^k) + \nabla h(y^k)). \tag{3.8c}$$

By using v to denote the dual variable for (3.8a), we obtain that (3.8a) is equivalent to

$$\min_x \max_v \; \langle -N^\top x, v\rangle - g(v) + b^\top x + \frac{1}{\gamma_k} D_h(x, z^k) \tag{3.9a}$$

$$= \max_v \min_x \; \langle -N^\top x, v\rangle - g(v) + b^\top x + \frac{1}{\gamma_k} D_h(x, z^k) \tag{3.9b}$$

$$\equiv \max_v \min_x \; h(x) - \langle \nabla h(z^k) + \gamma_k(Nv - b), x\rangle - \gamma_k g(v) - h(z^k) + \langle \nabla h(z^k), z^k\rangle \tag{3.9c}$$

$$= \max_v -\{\max_x -h(x) + \langle \nabla h(z^k) + \gamma_k(Nv - b), x\rangle + \gamma_k g(v) + h(z^k) - \langle \nabla h(z^k), z^k\rangle\} \tag{3.9d}$$

$$= \max_v -\{h^*(\nabla h(z^k) + \gamma_k(Nv - b)) + \gamma_k g(v) + h(z^k) - \langle \nabla h(z^k), z^k\rangle\} \tag{3.9e}$$

$$\equiv \min_v h^*(\nabla h(z^k) + \gamma_k(Nv - b)) + \gamma_k g(v). \tag{3.9f}$$

Moreover, note that the optimal x in (3.9d) is given by:

$$x := \nabla h^*(\gamma_k(Nv - b) + \nabla h(z^k)). \tag{3.10}$$

Similarly, by using u to denote the dual variable for (3.8b), we obtain that (3.8b) is equivalent to (for ease of presentation, denote $\tilde{y}^k := \nabla h^*(2\nabla h(x^k) - \nabla h(z^k))$):

$$\min_y \max_u \; \langle -M^\top y, u\rangle - f(u) + \frac{1}{\gamma_k} D_h(y, \tilde{y}^k) \tag{3.11a}$$

$$= \max_u \min_y \; \langle -M^\top y, u\rangle - f(u) + \frac{1}{\gamma_k} D_h(y, \tilde{y}^k) \tag{3.11b}$$

$$\equiv \max_u \min_y \; h(y) - \langle \nabla h(\tilde{y}^k) + \gamma_k Mu, y\rangle - \gamma_k f(u) - h(\tilde{y}^k) + \langle \nabla h(\tilde{y}^k), \tilde{y}^k\rangle \tag{3.11c}$$

$$= \max_u -\{\max_y -h(y) + \langle \nabla h(\tilde{y}^k) + \gamma_k Mu, y\rangle + \gamma_k f(u) + h(\tilde{y}^k) - \langle \nabla h(\tilde{y}^k), \tilde{y}^k\rangle\} \tag{3.11d}$$

$$= \max_u -\{h^*(\nabla h(\tilde{y}^k) + \gamma_k Mu) + \gamma_k f(u) + h(\tilde{y}^k) - \langle \nabla h(\tilde{y}^k), \tilde{y}^k\rangle\} \tag{3.11e}$$

$$\equiv \min_u h^*(2\nabla h(x^k) - \nabla h(z^k) + \gamma_k Mu) + \gamma_k f(u). \tag{3.11f}$$

Moreover, the optimal y in (3.11d) is given by

$$y := \nabla h^*(2\nabla h(x^k) - \nabla h(z^k) + \gamma_k M u). \tag{3.12}$$

By combining (3.9), (3.10), (3.11) and (3.12), we have that the BDRS (3.8) is equivalent to

$$\begin{aligned}
v^k &:= \operatorname*{argmin}_v h^*(\nabla h(z^k) + \gamma_k(Nv - b)) + \gamma_k g(v) && \text{(3.13a)}\\
x^k &:= \nabla h^*(\gamma_k(Nv^k - b) + \nabla h(z^k)) && \text{(3.13b)}\\
u^k &:= \operatorname*{argmin}_u h^*(2\nabla h(x^k) - \nabla h(z^k) + \gamma_k M u) + \gamma_k f(u) && \text{(3.13c)}\\
y^k &:= \nabla h^*(2\nabla h(x^k) - \nabla h(z^k) + \gamma_k M u^k) && \text{(3.13d)}\\
z^{k+1} &:= \nabla h^*(\nabla h(z^k) - \nabla h(x^k) + \nabla h(y^k)). && \text{(3.13e)}
\end{aligned}$$

Reordering the updates in (3.13), we know that (3.13) is equivalent to

$$\begin{aligned}
u^k &:= \operatorname*{argmin}_u h^*(2\nabla h(x^k) - \nabla h(z^k) + \gamma_k M u) + \gamma_k f(u) && \text{(3.14a)}\\
y^k &:= \nabla h^*(2\nabla h(x^k) - \nabla h(z^k) + \gamma_k M u^k) && \text{(3.14b)}\\
z^{k+1} &:= \nabla h^*(\nabla h(z^k) - \nabla h(x^k) + \nabla h(y^k)) && \text{(3.14c)}\\
v^{k+1} &:= \operatorname*{argmin}_v h^*(\nabla h(z^{k+1}) + \gamma_k(Nv - b)) + \gamma_k g(v) && \text{(3.14d)}\\
x^{k+1} &:= \nabla h^*(\gamma_k(Nv^{k+1} - b) + \nabla h(z^{k+1})). && \text{(3.14e)}
\end{aligned}$$

Combining (3.14b) and (3.14c) yields

$$\nabla h(z^{k+1}) = \nabla h(z^k) - \nabla(x^k) + 2\nabla h(x^k) - \nabla h(z^k) + \gamma_k M u^k = \nabla h(x^k) + \gamma_k M u^k. \tag{3.15}$$

From (3.14e) we have

$$\nabla h(z^{k+1}) = \nabla h(x^{k+1}) - \gamma_k(Nv^{k+1} - b), \tag{3.16}$$

which implies

$$2\nabla h(x^{k+1}) - \nabla h(z^{k+1}) = \nabla h(x^{k+1}) + \gamma_k(Nv^{k+1} - b), \tag{3.17}$$

and

$$\nabla h(x^{k+1}) = \nabla h(x^k) + \gamma_k(Mu^k + Nv^{k+1} - b). \tag{3.18}$$

By substituting (3.17) to (3.14a), (3.15) to (3.14d), and combining with (3.18), we know that (3.14) is equivalent to

$$\begin{aligned}
u^k &:= \operatorname*{argmin}_u h^*(\nabla h(x^k) + \gamma_k(Nv^k - b + Mu)) + \gamma_k f(u) && \text{(3.19a)}\\
v^{k+1} &:= \operatorname*{argmin}_v h^*(\nabla h(x^k) + \gamma_k(Mu^k + Nv - b)) + \gamma_k g(v) && \text{(3.19b)}\\
x^{k+1} &:= \nabla h^*(\nabla h(x^k) + \gamma_k(Mu^k + Nv^{k+1} - b)). && \text{(3.19c)}
\end{aligned}$$

This is exactly the same as the Bregman ADMM (3.5). □

3.2 Connection to Variable Metric ADMM

A variable metric ADMM has been studied in [39, 13], which solves the convex minimization problem (1.3) using the following updates:

$$u^k := \operatorname*{argmin}_u \ f(u) + \frac{1}{2\gamma_k}\|\gamma_k(Mu + Nv^{k-1} - b) + Lw^k\|^2_{L^{-1}} \tag{3.20a}$$

$$v^k := \operatorname*{argmin}_v \ g(v) + \frac{1}{2\gamma_k}\|\gamma_k(Mu^k + Nv - b) + Lw^k\|^2_{L^{-1}} \tag{3.20b}$$

$$w^{k+1} := w^k + \gamma_k L^{-1}(Mu^k + Nv^k - b), \tag{3.20c}$$

where $L \succ 0$ is a positive definite matrix, and $\|x\|_L^2 := x^\top Lx$. This algorithm is also known as applying ADMM to (1.3) with preconditioned constraints [38]. We now show that this variable metric ADMM is a special case of our Bregman ADMM (3.5) with a special choice of $h(x) = \frac{1}{2}\|x\|_L^2 := \frac{1}{2}x^\top Lx$, i.e., the quadratic function in Example 2.1 (ii). In this case, we have

$$\nabla h(x) = Lx, \quad h^*(x) = \frac{1}{2}\|x\|^2_{L^{-1}} = \frac{1}{2}x^\top L^{-1}x, \quad \nabla h^*(x) = L^{-1}x. \tag{3.21}$$

Theorem 3.2. *The variable metric ADMM* (3.20) *is a special case of the Bregman ADMM* (3.5) *with h given in Example 2.1 (ii).*

Proof. From (3.21), we know that the Bregman ADMM (3.5) can be written as

$$u^k := \operatorname*{argmin}_u \ f(u) + \frac{1}{2\gamma_k}\|Lw^k + \gamma_k(Mu + Nv^{k-1} - b)\|^2_{L^{-1}} \tag{3.22a}$$

$$v^k := \operatorname*{argmin}_v \ g(v) + \frac{1}{2\gamma_k}\|Lw^k + \gamma_k(Mu^k + Nv - b)\|^2_{L^{-1}} \tag{3.22b}$$

$$w^{k+1} := L^{-1}(Lw^k + \gamma_k(Mu^k + Nv^k - b)), \tag{3.22c}$$

which is the same as the variable metric ADMM (3.20). □

3.3 Alternating Direction Exponential Multiplier Method

In this section, we propose a new algorithm, which is a special case of BDRS, for solving the following linear inequality constrained convex minimization problem:

$$\min_{u,v} \ f(u) + g(v), \ \text{s.t.}, \ Mu + Nv - b \le 0, \ u \in \mathbb{R}^p, v \in \mathbb{R}^q, \tag{3.23}$$

where f and g are both proper, closed and convex functions, and $M \in \mathbb{R}^{m\times p}$, $N \in \mathbb{R}^{m\times q}$, $b \in \mathbb{R}^m$.

One important approach for solving (3.23) is the exponential multiplier method (EMM) that was proposed and studied by Bertsekas, Kort and Tseng [53, 10, 87]. Unlike the usual augmented Lagrangian method that uses a quadratic penalty term, the EMM proposes to use a non-quadratic penalty term. More specifically, the EMM uses the exponential penalty function given by:

$$\psi(t) = e^t - 1. \tag{3.24}$$

By associating a Lagrange multiplier w_j to the j-th constraint in (3.23), the EMM for solving (3.23) iterates as follows [1]:

$$(u^k, v^k) := \operatorname*{argmin}_{u,v} f(u) + g(v) + \frac{1}{\gamma_k} \sum_{j=1}^{m} w_j^k \psi(\gamma_k(M_j^\top u + N_j^\top v - b_j)) \tag{3.25a}$$

$$w_j^{k+1} := w_j^k \nabla\psi(\gamma_k(M_j^\top u^k + N_j^\top v^k - b_j)) = w_j^k e^{\gamma_k(M_j^\top u^k + N_j^\top v^k - b_j)},\ j = 1, \ldots, m, \tag{3.25b}$$

where $\gamma_k > 0$ is a parameter, and $M_j^\top$ denotes the j-th row of M and $N_j^\top$ denotes the j-th row of N. It can be shown that the EMM (3.25) is equivalent to a Bregman proximal point algorithm for solving the dual of (3.23). Note that the dual of (3.23) is given by

$$\max\ d(w),\ \text{s.t.},\ w \geq 0, \tag{3.26}$$

where the dual function $d(w)$ is defined as

$$d(w) := \min_{u,v} \left\{ f(u) + g(v) + \sum_{j=1}^{m} w_j (M_j^\top u + N_j^\top v - b_j) \right\}.$$

It can be shown that the EMM (3.25) for solving the primal problem (3.23) is equivalent to the following algorithm for solving the dual problem (3.26):

$$w^{k+1} := \operatorname*{argmax}_{w \geq 0} \left\{ d(w) - \frac{1}{\gamma_k} \sum_{j=1}^{m} w_j^k \psi^* \left(\frac{w_j}{w_j^k} \right) \right\}, \tag{3.27}$$

where ψ^* is the conjugate function of ψ, which is the entropy function

$$\psi^*(s) = s \log s - s + 1. \tag{3.28}$$

Note that (3.27) is indeed a Bregman proximal point algorithm, because $\sum_{j=1}^{m} w_j^k \psi^* \left(\frac{w_j}{w_j^k} \right) = D_h(w, w^k)$ with h being the Boltzmann-Shannon entropy defined in Example (2.1) (iii).

Note that for some applications, the subproblem (3.25a) in EMM can be difficult to solve. To overcome this difficulty, we propose an alternating direction exponential multiplier method (ADEMM) for solving (3.23). The ADEMM iterates as follows:

$$u^k := \operatorname*{argmin}_{u} f(u) + \frac{1}{\gamma_k} \sum_{j=1}^{m} w_j^k \psi(\gamma_k(M_j^\top u + N_j^\top v^{k-1} - b_j)) \tag{3.29a}$$

$$v^k := \operatorname*{argmin}_{v} g(v) + \frac{1}{\gamma_k} \sum_{j=1}^{m} w_j^k \psi(\gamma_k(M_j^\top u^k + N_j^\top v - b_j)) \tag{3.29b}$$

$$w_j^{k+1} := w_j^k \nabla\psi(\gamma_k(M_j^\top u^k + N_j^\top v^k - b_j)) = w_j^k e^{\gamma_k(M_j^\top u^k + N_j^\top v^k - b_j)},\ j = 1, \ldots, m. \tag{3.29c}$$

[1] Note that the EMM proposed in [53, 10, 87] allows the scalar γ_k to be replaced by a vector whose j-th entry is $[\gamma_k]_j$. In this case, the EMM becomes

$$(u^k, v^k) := \operatorname*{argmin}_{u,v} f(u) + g(v) + \sum_{j=1}^{m} \frac{w_j^k}{\gamma_j^k} \psi([\gamma_k]_j (M_j^\top u + N_j^\top v - b_j))$$

$$w_j^{k+1} := w_j^k \nabla\psi([\gamma_k]_j (M_j^\top u^k + N_j^\top v^k - b_j)) = w_j^k e^{[\gamma_k]_j (M_j^\top u^k + N_j^\top v^k - b_j)},\ j = 1, \ldots, m.$$

Here we use a scalar γ_k for simplicity.

That is, the ADEMM alternatingly minimizes the Lagrangian function with respect to u and v in each iteration. This is exactly in the same spirit of the usual ADMM in Euclidean space with a quadratic penalty. To the best of our knowledge, the ADEMM (3.29) is new to the literature: it is the first alternating direction version of EMM. For some applications, both subproblems (3.29a) and (3.29b) are easier to solve than (3.25a). This is indeed the case as we will see later in the discrete optimal transport problem in Section 5.

Very interestingly, the ADEMM (3.29) for solving the primal problem (3.23) is equivalent to BDRS (3.1) for solving the dual problem (3.26), with h being the Boltzmann-Shannon entropy.

Theorem 3.3. *The ADEMM* (3.29) *for solving the primal problem* (3.23) *is equivalent to BDRS* (3.1) *for solving the dual problem* (3.26)*, with h being the Boltzmann-Shannon entropy.*

Proof. We first note that the dual problem (3.26) is equivalent to

$$\min_{w\geq 0}\ f^*(-M^\top w)+g^*(-N^\top w)+b^\top w, \tag{3.30}$$

whose optimality condition is given by:

$$0\in -M\partial f^*(M^\top w)-N\partial g^*(N^\top w)+b+\partial\mathbb{I}(w\geq 0), \tag{3.31}$$

where $\mathbb{I}(C)$ denotes the indicator function of set C. By letting $A(w)=-M\partial f^*(-M^\top w)+\partial\mathbb{I}(w\geq 0)$ and $B(w)=-N\partial g^*(-N^\top w)+b+\partial\mathbb{I}(w\geq 0)$, we note that (3.31) is in the same form as (1.1) and thus can be solved by BDRS (3.1). The BDRS (3.1) with h being the Boltzmann-Shannon entropy can be written as

$$x^k := \operatorname*{argmin}_{x\geq 0}\ g^*(-N^\top x)+b^\top x+\frac{1}{\gamma_k}D_h(x,z^k) \tag{3.32a}$$

$$y^k := \operatorname*{argmin}_{y\geq 0}\ f^*(-M^\top y)+\frac{1}{\gamma_k}D_h(y,(x^k\circ x^k)/z^k) \tag{3.32b}$$

$$z^{k+1} := \nabla h^*(\nabla h(z^k)-\nabla h(x^k)+\nabla h(y^k)), \tag{3.32c}$$

where $a\circ b$ denotes the elementwise multiplication of vectors a and b, and a/b denotes the elementwise division of vectors a and b.

By introducing v as the dual variable of (3.32a), we obtain that (3.32a) is equivalent to:

$$\begin{aligned}
&\min_x\max_v\langle -N^\top x,v\rangle-g(v)+b^\top x+\frac{1}{\gamma_k}D_h(x,z^k)\\
=&\max_v\min_x\langle -N^\top x,v\rangle-g(v)+b^\top x+\frac{1}{\gamma_k}D_h(x,z^k)\\
\equiv&\max_v -\gamma_k g(v)-h^*(\nabla h(z^k)+\gamma_k(N_j^\top v-b_j))\\
\equiv&\min_v g(v)+\frac{1}{\gamma_k}\sum_j z_j^k e^{\gamma_k(N_j^\top v-b_j)},
\end{aligned}$$

with $\gamma_k(-Nv+b)+\nabla h(x)-\nabla h(z^k)=0$. Similarly, by introducing u as the dual variable of (3.32b),

we obtain that (3.32b) is equivalent to:

$$\begin{aligned}
&\min_y \max_u \langle -M^\top y, u\rangle - f(u) + \frac{1}{\gamma_k} D_h(y, (x^k \circ x^k)/z^k) \\
=&\max_u \min_y \langle -M^\top y, u\rangle - f(u) + \frac{1}{\gamma_k} D_h(y, (x^k \circ x^k)/z^k) \\
=&\max_u -\gamma f(u) - \sum_j (x_j^k \cdot x_j^k / z_j^k) e^{\gamma_k M_j^\top u} \\
\equiv&\min_u f(u) + \frac{1}{\gamma_k} \sum_j (x_j^k \cdot x_j^k / z_j^k) e^{\gamma_k M_j^\top u},
\end{aligned}$$

with $-\gamma_k M u + \nabla h(y) - \nabla h((x^k \circ x^k)./z^k) = 0$. Therefore, (3.32) can be equivalently rewritten as

$$v^k := \operatorname*{argmin}_v g(v) + \frac{1}{\gamma_k} \sum_j z_j^k e^{\gamma_k (N_j^\top v - b_j)} \tag{3.33a}$$

$$x_j^k := z_j^k e^{\gamma_k (N_j^\top v^k - b_j)} \tag{3.33b}$$

$$u^k := \operatorname*{argmin}_u f(u) + \frac{1}{\gamma_k} \sum_j (x_j^k \cdot x_j^k / z_j^k) e^{\gamma_k M_j^\top u} \tag{3.33c}$$

$$y_j^k := (x_j^k \cdot x_j^k / z_j^k) e^{\gamma_k M_j^\top u^k} \tag{3.33d}$$

$$z^{k+1} := \nabla h^*(\nabla h(z^k) - \nabla h(x^k) + \nabla h(y^k)). \tag{3.33e}$$

We have (3.33) can be equivalently rewritten as

$$\begin{aligned}
v^k &:= \operatorname*{argmin}_v g(v) + \frac{1}{\gamma_k} \sum_j z_j^k e^{\gamma_k (N_j^\top v - b_j)} \\
x_j^k &:= z_j^k e^{\gamma_k (N_j^\top v^k - b_j)} \\
u^k &:= \operatorname*{argmin}_u f(u) + \frac{1}{\gamma_k} \sum_j x_j^k e^{\gamma_k (M_j^\top u + N_j^\top v^k - b_j)} \\
z_j^{k+1} &:= \nabla h^*(\nabla h(z_j^k) + \gamma_k (M_j^\top u^k + N_j^\top v^k - b_j)).
\end{aligned}$$

The last equation implies that $\nabla h(z_j^{k+1}) = \nabla h(x_j^k) + \gamma_k M_j^\top u^k$, and therefore, $z_j^{k+1} = x_j^k e^{\gamma_k M_j^\top u^k}$. Thus, the above is equivalent to

$$\begin{aligned}
v^k &:= \operatorname*{argmin}_v g(v) + \frac{1}{\gamma_k} \sum_j x_j^{k-1} e^{\gamma_k (M_j^\top u^{k-1} + N_j^\top v - b_j)} \\
x_j^k &:= x_j^{k-1} e^{\gamma_k (M_j^\top u^{k-1} + N_j^\top v^k - b_j)} \\
u^k &:= \operatorname*{argmin}_u f(u) + \frac{1}{\gamma_k} \sum_j x_j^k e^{\gamma_k (M_j^\top u + N_j^\top v^k - b_j)}.
\end{aligned}$$

This is exactly the ADEMM (3.29). □

We end this section by remarking that the EMM (3.25), being equivalent to a Bregman PPM, is also equivalent to the nonlinear rescaling method developed by Polyak [69, 70, 71, 44] with suitably

chosen nonlinear rescaling function, as proved by Polyak and Teboulle [72]. More specifically, the linearly inequality constrained problem (3.23) is equivalent to:

$$\min_{u,v} \ f(u)+g(v), \ \text{s.t.}, \ \frac{1}{\gamma_k}\psi(\gamma_k(M_j^\top u+N_j^\top v-b_j))\leq 0, j=1,\ldots,m, \tag{3.34}$$

where ψ is defined in (3.24). By associating a Lagrange multiplier w_j to the j-th constraint in (3.34), the Lagrangian function of (3.34) is given by:

$$\mathcal{L}_{NR}(u,v;w):=f(u)+g(v)+\frac{1}{\gamma_k}\sum_j w_j\psi(\gamma_k(M_j^\top u+N_j^\top v-b_j)). \tag{3.35}$$

The nonlinear rescaling method is essentially a Lagrangian multiplier method for solving (3.35) and is given by:

$$(u^k,v^k):=\operatorname*{argmin}_{u,v} \ \mathcal{L}_{NR}(u,v;w^k) \tag{3.36a}$$

$$w_j^{k+1}:=w_j^k\nabla\psi(\gamma_k(M_j^\top u^k+N_j^\top v^k-b_j)), j=1,\ldots,m. \tag{3.36b}$$

Polyak and Griva [71, 44] proposed to use Newton's method to solve a primal-dual system that consists (3.36b) and the KKT system of (3.36a). Note that Newton's method can be employed because function ψ and its derivative ψ' are both twice continuously differentiable. This is one of the main motivations of designing the nonlinear rescaling method. Our ADEMM (3.29) leads to the following alternating direction version of the nonlinear rescaling method:

$$\begin{aligned}
u^k&:=\operatorname*{argmin}_{u,v} \ \mathcal{L}_{NR}(u,v^{k-1};w^k)\\
v^k&:=\operatorname*{argmin}_{u,v} \ \mathcal{L}_{NR}(u^k,v;w^k)\\
w_j^{k+1}&:=w_j^k\nabla\psi(\gamma_k(M_j^\top u^k+N_j^\top v^k-b_j)), j=1,\ldots,m.
\end{aligned}$$

4 Bregman Peaceman-Rachford Splitting and Bregman Double-Backward Method

In this section, we discuss two algorithms that are related to BDRS, namely Bregman Peaceman-Rachford spliting method (BPRS) and Bregman double-backward method (BDBM). The Peaceman-Rachford splitting (PRS) method is another well-studied operator splitting method that was also proposed to solve variational problems arising from numerical PDEs [67]. The PRS can also be applied to solving the monotone inclusion problem (1.1). Using our notions defined in Definition 2.3, the BPRS for solving (1.1) can be written as

$$z^{k+1}:=R^h_{\gamma_k A}R^h_{\gamma_k B}(z^k). \tag{4.1}$$

When $h(x)=\frac{1}{2}\|x\|_2^2$, (4.1) reduces to the original PRS in the Euclidean space. For the convex minimization problem (1.3), it is known that a symmetric ADMM is equivalent to the PRS applied to solving the dual of (1.3) [41]. The symmetric ADMM for solving (1.3) iterates as follows:

$$u^{k+1}:=\operatorname*{argmin}_{u} \ \mathcal{L}_\beta(u,v^k;w^k) \tag{4.2a}$$

$$w^{k+\frac{1}{2}}:=w^k+\beta(Mu^{k+1}+Nv^k-b) \tag{4.2b}$$

$$v^{k+1}:=\operatorname*{argmin}_{v} \ \mathcal{L}_\beta(u^{k+1},v;w^{k+\frac{1}{2}}) \tag{4.2c}$$

$$w^{k+1}:=w^{k+\frac{1}{2}}+\beta(Mu^{k+1}+Nv^{k+1}-b), \tag{4.2d}$$

where the augmented Lagrangian function $\mathcal{L}_\beta$ is defined in (1.5). This symmetric ADMM is equivalent to an alternating linearization method when f and g are both smooth, as studied by Goldfarb, Ma and Scheinberg [42]. For BPRS (4.1), we can prove a similar result.

Theorem 4.1. *For $\gamma_k > 0$, the BPRS (4.1) for solving the dual of (1.3) with $A(x) = -M\partial f^*(-M^\top x)$ and $B(x) = -N\partial g^*(-N^\top x)+b$, is equivalent to the following Bregman symmetric ADMM for solving (1.3):*

$$u^k := \operatorname*{argmin}_u \ f(u) + \frac{1}{\gamma_k} h^*(\nabla h(w^k) + \gamma_k(Mu + Nv^{k-1} - b)) \tag{4.3a}$$

$$w^{k+\frac{1}{2}} := \nabla h^*(\nabla h(w^k) + \gamma_k(Mu^k + Nv^{k-1} - b)) \tag{4.3b}$$

$$v^k := \operatorname*{argmin}_v \ g(v) + \frac{1}{\gamma_k} h^*(\nabla h(w^{k+\frac{1}{2}}) + \gamma_k(Mu^k + Nv - b)) \tag{4.3c}$$

$$w^{k+1} := \nabla h^*(\nabla h(w^{k+\frac{1}{2}}) + \gamma_k(Mu^k + Nv^k - b)). \tag{4.3d}$$

Proof. The proof is similar to the proof of Theorem 3.1. The BPRS (4.1) can be equivalently written as

$$x^k := J^h_{\gamma_k B}(z^k) \tag{4.4a}$$

$$y^k := J^h_{\gamma_k A} \circ \nabla h^*(2\nabla h(x^k) - \nabla h(z^k)) \tag{4.4b}$$

$$z^{k+1} := \nabla h^*(\nabla h(z^k) - 2\nabla h(x^k) + 2\nabla h(y^k)), \tag{4.4c}$$

which is further equivalent to

$$x^k := \operatorname*{argmin}_x \ g^*(-N^\top x) + b^\top x + \frac{1}{\gamma_k} D_h(x, z^k) \tag{4.5a}$$

$$y^k := \operatorname*{argmin}_y \ f^*(-M^\top y) + \frac{1}{\gamma_k} D_h(y, \nabla h^*(2\nabla h(x^k) - \nabla h(z^k))) \tag{4.5b}$$

$$z^{k+1} := \nabla h^*(\nabla h(z^k) - 2\nabla h(x^k) + 2\nabla h(y^k)). \tag{4.5c}$$

By associating v as the dual variable of (4.5a), and u as the dual variable of (4.5b), similar to the proof of Theorem 3.1, it can be shown that (4.5) is equivalent to

$$v^k := \operatorname*{argmin}_v h^*(\nabla h(z^k) + \gamma_k(Nv - b)) + \gamma_k g(v) \tag{4.6a}$$

$$x^k := \nabla h^*(\gamma_k(Nv^k - b) + \nabla h(z^k)) \tag{4.6b}$$

$$u^k := \operatorname*{argmin}_u h^*(2\nabla h(x^k) - \nabla h(z^k) + \gamma_k Mu) + \gamma_k f(u) \tag{4.6c}$$

$$y^k := \nabla h^*(2\nabla h(x^k) - \nabla h(z^k) + \gamma_k Mu^k) \tag{4.6d}$$

$$z^{k+1} := \nabla h^*(\nabla h(z^k) - 2\nabla h(x^k) + 2\nabla h(y^k)). \tag{4.6e}$$

Reordering the updates in (4.6), we know that (4.6) is equivalent to

$$u^k := \operatorname*{argmin}_u h^*(2\nabla h(x^k) - \nabla h(z^k) + \gamma_k Mu) + \gamma_k f(u) \tag{4.7a}$$

$$y^k := \nabla h^*(2\nabla h(x^k) - \nabla h(z^k) + \gamma_k Mu^k) \tag{4.7b}$$

$$z^{k+1} := \nabla h^*(\nabla h(z^k) - 2\nabla h(x^k) + 2\nabla h(y^k)) \tag{4.7c}$$

$$v^{k+1} := \operatorname*{argmin}_v h^*(\nabla h(z^{k+1}) + \gamma_k(Nv - b)) + \gamma_k g(v) \tag{4.7d}$$

$$x^{k+1} := \nabla h^*(\gamma_k(Nv^{k+1} - b) + \nabla h(z^{k+1})). \tag{4.7e}$$

Note that (4.7e) implies

$$\nabla h(z^{k+1}) = \nabla h(x^{k+1}) - \gamma_k(Nv^{k+1} - b), \tag{4.8}$$

which immediately gives

$$\nabla h(z^k) = \nabla h(x^k) - \gamma_k(Nv^k - b). \tag{4.9}$$

Combining (4.7b) and (4.7c) yields

$$\nabla h(z^{k+1}) = \nabla h(y^k) + \gamma_k Mu^k, \tag{4.10}$$

which together with (4.8) gives

$$\nabla h(x^{k+1}) = \nabla h(y^k) + \gamma_k(Mu^k + Nv^{k+1} - b). \tag{4.11}$$

Moreover, (4.7b) implies

$$\nabla h(y^k) = 2\nabla h(x^k) - \nabla h(z^k) + \gamma_k Mu^k = \nabla h(x^k) + \gamma_k(Mu^k + Nv^k - b), \tag{4.12}$$

where the second equality is due to (4.9). Substituting (4.10) into (4.7d), (4.9) into (4.7a), and combining with (4.12) and (4.11), we know that (4.7) is equivalent to

$$\begin{aligned}
u^k &:= \operatorname*{argmin}_u h^*(\nabla h(x^k) + \gamma_k(Mu + Nv^k - b)) + \gamma_k f(u)\\
y^k &:= \nabla h^*(\nabla h(x^k) + \gamma_k(Mu^k + Nv^k - b))\\
v^{k+1} &:= \operatorname*{argmin}_v h^*(\nabla h(y^k) + \gamma_k(Mu^k + Nv - b)) + \gamma_k g(v)\\
x^{k+1} &:= \nabla h^*(\nabla h(y^k) + \gamma_k(Mu^k + Nv^{k+1} - b)),
\end{aligned}$$

which is exactly the same as the Bregman symmetric ADMM (4.3). □

For the linearly inequality constrained problem (3.23), the BPRS with h being the Boltzmann-Shannon entropy leads to a symmetric version of ADEMM.

Theorem 4.2. *For $\gamma_k > 0$, the BPRS* (4.1) *for solving the dual of* (3.23) *with h being the Boltzmann-Shannon entropy and $A(x) = -M\partial f^*(-M^\top x)$ and $B(x) = -N\partial g^*(-N^\top x) + b$, is equivalent to the following Bregman symmetric ADEMM for solving* (3.23)*:*

$$u^k := \operatorname*{argmin}_u f(u) + \frac{1}{\gamma_k}\sum_{j=1}^m w_j^k \psi(\gamma_k(M_j^\top u + N_j^\top v^{k-1} - b_j)) \tag{4.13a}$$

$$w_j^{k+\frac{1}{2}} := w_j^k e^{\gamma_k(M_j^\top u^k + N_j^\top v^k - b_j)},\ j = 1, \ldots, m. \tag{4.13b}$$

$$v^k := \operatorname*{argmin}_v g(v) + \frac{1}{\gamma_j^k}\sum_{j=1}^m w_j^{k+\frac{1}{2}} \psi(\gamma_k(M_j^\top u^k + N_j^\top v - b_j)) \tag{4.13c}$$

$$w_j^{k+1} := w_j^{k+\frac{1}{2}} e^{\gamma_k(M_j^\top u^k + N_j^\top v^k - b_j)},\ j = 1, \ldots, m, \tag{4.13d}$$

where ψ is defined in (3.24)*.*

Proof. The proof is very similar to the proof of Theorem 4.1. We thus omit it for brevity. □

Another closely related splitting method is the double-backward (also called backward-backward) method. The double-backward method was proposed by Passty [66] and later studied by many others including Combettes [21]. The Bregman double-backward method for solving (1.1) is given by

$$z^{k+1} := J^h_{\gamma_k A} J^h_{\gamma_k B}(z^k). \tag{4.14}$$

The theoretical analysis of BDBM has been mainly studied for solving the feasibility problem, i.e., when A and B are both normal cone operators. In this case, the monotone inclusion problem reduces to

$$\text{Find } x, \text{ s.t., } x \in \mathcal{X}_1 \bigcap \mathcal{X}_2, \tag{4.15}$$

where $\mathcal{X}_1$ and $\mathcal{X}_2$ are convex sets, and A and B are normal cone operators of $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. Under the assumption that $A^{-1}(0) \cap B^{-1}(0)$ is not empty, the weak convergence of BDBM (4.14) was proved by Reich [73]. In fact, note that both $J^h_{\gamma_k B}$ and $J^h_{\gamma_k A}$ are quasi-Bregman nonexpansive (QBNE, see [80]), and thus their product is also QBNE [80]. Therefore, we have $\forall u \in A^{-1}(0) \bigcap B^{-1}(0) = \text{Fix}(J^h_{\gamma_k A}) \bigcap \text{Fix}(J^h_{\gamma_k B})$,

$$D_h(u, x_{k+1}) = D_h(u, J^h_{\gamma_k B} J^h_{\gamma_k A}(x_k)) \leq D_h(u, x_k) \leq D_h(u, x_0), \tag{4.16}$$

where $\text{Fix}(A)$ denotes the fixed point set of A. Therefore, (4.16) implies that $D_h(u, x_k)$ is convergent.

For general A and B, however, it may not be reasonable to assume that $A^{-1}(0) \cap B^{-1}(0)$ is not empty, and thus the convergence result in [73] does not apply. This is indeed the case as we will see later for the optimal transport problem in Section 5.

5 Applications to Discrete Optimal Transport

Optimal transport has found many important applications in machine learning and data science recently [88, 68]. In the case of discrete probability measures, one is given two sets of finite number atoms, $\{y_1, y_2, \ldots, y_n\} \subset \mathbb{R}^d$ and $\{z_1, z_2, \ldots, z_n\} \subset \mathbb{R}^d$, and two probability distributions $\mu_n = \sum_{i=1}^n r_i \delta_{y_i}$ and $\nu_n = \sum_{j=1}^n c_j \delta_{z_j}$. Here $r = (r_1, r_2, \ldots, r_n)^\top \in \Delta^n$ and $c = (c_1, c_2, \ldots, c_n)^\top \in \Delta^n$, Δ^n denotes the probability simplex in $\mathbb{R}^n$ and δ_y denotes the Dirac delta function at y. The optimal transport between μ_n and ν_n is obtained by solving the following problem:

$$\min_X \langle C, X \rangle, \text{ s.t., } X\mathbf{1} = r, X^\top \mathbf{1} = c, X \geq 0, \tag{5.1}$$

where $\mathbf{1}$ denotes the n-dimensional all-one vector, $C \in \mathbb{R}^{n \times n}$ is the cost matrix whose (i, j)-th component is $C_{ij} = \|y_i - z_j\|^2$. Note that (5.1) is a linear program (LP) and can be solved by off-the-shelf LP solvers. However, (5.1) appearing in real applications can be very large, and classical LP solvers may suffer scalability issues. In [23], Cuturi suggested to adopt the algorithm proposed by Sinkhorn and Knopp [81] to solve the following approximation of (5.1):

$$\min_X \langle C, X \rangle + \eta h(X), \text{ s.t., } X\mathbf{1} = r, X^\top \mathbf{1} = c, \tag{5.2}$$

where $\eta > 0$ is a penalty parameter, $h(X)$ is the Boltzmann-Shannon entropy, and for matrix X it is defined as: $h(X) = \sum_{ij} X_{ij}(\log X_{ij} - 1)$. That is, an entropy penalty term is added to the objective function with a penalty parameter η. The advantage of the penalized problem is that the nonnegativity constraint $X \geq 0$ is no longer needed because it is implicitly enforced by the entropy function $h(X)$. The algorithm proposed in [81] (from now on, we call it Sinkhorn's algorithm) solves the dual problem (5.2) using a block minimization algorithm. More specifically, using α and

β to denote the Lagrange multipliers associated with the two linear equality constraints of (5.2), the dual problem of (5.2) can be written as:

$$\max_{\alpha,\beta} \min_{X} \ \langle C, X\rangle + \eta h(X) - \langle \alpha, X\mathbf{1} - r\rangle - \langle \beta, X^\top \mathbf{1} - c\rangle \tag{5.3a}$$

$$= \max_{\alpha,\beta} \ \langle \alpha, r\rangle + \langle \beta, c\rangle - \eta \sum_{ij} e^{\frac{1}{\eta}(\alpha_i + \beta_j - C_{ij})}. \tag{5.3b}$$

Note that the X-minimization problem in (5.3a) has a closed-form optimal solution given by

$$X_{ij} = e^{\frac{1}{\eta}(\alpha_i + \beta_j - C_{ij})}, \quad i, j = 1, \ldots, n.$$

By letting $K_{ij} = e^{-C_{ij}/\eta}$, $u_i = e^{\alpha_i/\eta}$, and $v_j = e^{\beta_j/\eta}$, (5.3b) can be equivalently written as:

$$\min_{u,v} \sum_{ij} u_i K_{ij} v_j - \sum_i r_i \log u_i - \sum_j c_j \log v_j. \tag{5.4}$$

The Sinkhorn's algorithm [81, 23] solves (5.4) by alternatingly minimizing u and v with the other variable being fixed, i.e.,

$$u^k := \operatorname*{argmin}_{u} \ \sum_{ij} u_i K_{ij} v_j^{k-1} - \sum_i r_i \log u_i \tag{5.5a}$$

$$v^k := \operatorname*{argmin}_{v} \ \sum_{ij} u_i^k K_{ij} v_j - \sum_j c_j \log v_j. \tag{5.5b}$$

It turns out that both subproblems in (5.5) admit closed-form solutions, and the Sinkhorn's algorithm can be written more compactly as:

$$u^k := r./(Kv^{k-1}) \tag{5.6a}$$

$$v^k := c./(K^\top u^k). \tag{5.6b}$$

The Sinkhorn's algorithm can be implemented very efficiently because the computations in (5.6) are simple. One drawback of the Sinkhorn's algorithm is that it is very difficult to tune the parameter η. Ideally, one wants a small η so that the penalized problem (5.2) is close to the original problem (5.1). However, small η will cause numerical instability of the Sinkhorn's algorithm. It is usually not clear how small η can be without causing numerical issues. Moreover, the Sinkhorn's algorithm only solves the regularized problem (5.2), not the original OT problem (5.1).

Now we discuss how to use our BDRS (or ADEMM) to solve the original OT (5.1). Note that (5.1) can be written in the form of (1.1) by defining

$$A(X) = C + \partial\mathbb{I}(X\mathbf{1} = r) + \partial\mathbb{I}(X \geq 0), \quad B(X) = \partial\mathbb{I}(X^\top \mathbf{1} = c) + \partial\mathbb{I}(X \geq 0). \tag{5.7}$$

Now the BDRS (3.1) can be written as (for the ease of comparison with Sinkhorn's algorithm, we choose γ_k to be a constant $1/\eta$):

$$X^k := \operatorname*{argmin}_{X} \ D_h(X, Z^k), \ \text{s.t.}, \ X^\top \mathbf{1} = c \tag{5.8a}$$

$$Y^k := \operatorname*{argmin}_{Y} \ \langle C, Y\rangle + \eta D_h(Y, X^k \circ X^k./Z^k), \ \text{s.t.}, \ Y\mathbf{1} = r \tag{5.8b}$$

$$Z^{k+1} := Z^k \circ Y^k./X^k. \tag{5.8c}$$

The dual problem of the OT (5.1) is given by:

$$\min_{\alpha,\beta} \ -r^\top \alpha - c^\top \beta, \ \text{s.t.}, \ \alpha_i + \beta_j \le C_{ij}, i, j = 1, \ldots, n. \tag{5.9}$$

The ADEMM (3.29) for solving (5.9) is given by:

$$\alpha^k := \operatorname*{argmin}_{\alpha} \ -r^\top \alpha + \eta \sum_{ij} X_{ij}^k e^{\frac{1}{\eta}(\alpha_i + \beta_j^{k-1} - C_{ij})} \tag{5.10a}$$

$$\beta^k := \operatorname*{argmin}_{\beta} \ -c^\top \beta + \eta \sum_{ij} X_{ij}^k e^{\frac{1}{\eta}(\alpha_i^k + \beta_j - C_{ij})} \tag{5.10b}$$

$$X_{ij}^{k+1} := X_{ij}^k e^{\frac{1}{\eta}(\alpha_i^k + \beta_j^k - C_{ij})}, \ i, j = 1, \ldots, n. \tag{5.10c}$$

By denoting $u = e^{\alpha/\eta}$, $v = e^{\beta/\eta}$, and $K_{ij} = e^{-C_{ij}/\eta}$, (5.10) can be further rewritten as

$$u^k := \operatorname*{argmin}_{u} \ \sum_{ij} u_i X_{ij}^k K_{ij} v_j^{k-1} - \sum_i r_i \log u_i \tag{5.11a}$$

$$v^k := \operatorname*{argmin}_{v} \ \sum_{ij} u_i^k X_{ij}^k K_{ij} v_j - \sum_j c_j \log v_j \tag{5.11b}$$

$$X_{ij}^{k+1} := u_i^k X_{ij}^k K_{ij} v_j^k, \ i, j = 1, \ldots, n. \tag{5.11c}$$

Similar to (5.5), the two subproblems (5.11a) and (5.11b) admit closed-form solutions, and (5.11) can be equivalently written as:

$$u^k := r./((X^k \circ K)v^{k-1}) \tag{5.12a}$$

$$v^k := c./((X^k \circ K)^\top u^k) \tag{5.12b}$$

$$X_{ij}^{k+1} := u_i^k X_{ij}^k K_{ij} v_j^k, \ i, j = 1, \ldots, n. \tag{5.12c}$$

Now comparing ADEMM (5.12) with the Sinkhorn's algorithm (5.6), we see when updating u^k and v^k, (5.12) replaces matrix K in (5.6) by matrix $X^k \circ K$, which is no longer a constant matrix. The matrix X^{k+1} is then updated by (5.12c).

Remark 5.1. *Our BDRS* (5.12) *for solving the OT probelm can be a much better algorithm than the Sinkhorn's algorithm, because we do not require η to be close to 0, and thus resolve the issue of numerical instability of the Sinkhorn's algoirthm. We also see that the computations in* (5.12) *are very simple and can be done in parallel as illustrated in [23] for the Sinkhorn's algorithm. Thus our BDRS can be a much better algorithm than the classical ADMM for solving* (5.1) *which requires projections onto the probability simplex in each iteration.*

We now discuss the BDBM (4.14) for solving the OT problem (5.1). We note that $\operatorname{Fix}(J_{\gamma A}^h J_{\gamma B}^h) \not\subset (A+B)^{-1}(0)$, although $\operatorname{Fix}(R_{\gamma A}^h R_{\gamma B}^h) \subset (A+B)^{-1}(0)$. Therefore, even though we can find a fixed point of $J_{\gamma A}^h J_{\gamma B}^h$ using the BDBM with a constant $\gamma_k = \gamma$, the solution we find may not solve the OT problem (5.1). However, we have the following result regarding the fixed point of $J_{\gamma_k A}^h J_{\gamma_k B}^h$.

Theorem 5.2. *For the OT problem* (5.1) *with A and B defined in* (5.7) *and h being the Boltzmann-Shannon entropy, it holds that*

$$\lim_{\gamma_k \to 0} \operatorname{Fix}(J_{\gamma_k A}^h J_{\gamma_k B}^h) \subset (A+B)^{-1}(0).$$

Proof. Assume $x \in \mathrm{Fix}(J^h_{\gamma_k A} J^h_{\gamma_k B})$, that is

$$x = J^h_{\gamma_k A} J^h_{\gamma_k B} x,$$

which can be equivalently written as

$$x = (\nabla h + \gamma_k A)^{-1} \circ \nabla h \circ (\nabla h + \gamma_k B)^{-1} \circ \nabla h(x).$$

This is further equivalent to (note that $\nabla h(x) = \log x$ and $\nabla h^*(x) = e^x$):

$$\begin{aligned}
& (\nabla h + \gamma_k B)\nabla h^* (\nabla h + \gamma_k A)x = \nabla h(x) \\
\iff \quad & (I + \gamma_k B \nabla h^*)(\nabla h(x) + \gamma_k A(x)) = \nabla h(x) \\
\iff \quad & (I + \gamma_k B \nabla h^*) \log(x^k \cdot e^{\gamma_k A(x)}) = \nabla h(x) \\
\iff \quad & \log(x \cdot e^{\gamma_k A(x)}) + \gamma_k B(x \cdot e^{\gamma_k A(x)}) = \nabla h(x) \\
\iff \quad & \log(x \cdot e^{\gamma_k A(x)} \cdot e^{\gamma_k B(x \cdot e^{\gamma_k A(x)})}) = \log(x) \\
\iff \quad & \gamma_k A(x) + \gamma_k B(x \cdot e^{\gamma_k A(x)}) = 0 \\
\iff \quad & A(x) + B(x \cdot e^{\gamma_k A(x)}) = 0.
\end{aligned}$$

By letting $\gamma_k \to 0$, we have $A(x) + B(x) = 0$, i.e., $x \in (A + B)^{-1}(0)$. □

Theorem 5.2 shows that if we let $\gamma_k \to 0$, then the BDBM (4.14) solves the OT problem (5.1).

6 Concluding Remarks

In this paper, we proposed several new algorithms for solving monotone operator inclusion problem and convex minimization problems. The algorithms include BDRS, BPRS, Bregman ADMM, and ADEMM. We discussed their connections with existing algorithms in the literature. We also discussed how to apply our algorithms to solve the discrete optimal transport problem. We proved the convergence of the algorithms under certain assumptions, though we point out that one assumption does not apply to the OT problem. We leave it as a future work to prove the convergence of the proposed algorithms under the most general setting.

References

[1] F. J. Aragón Artacho, Y. Censor, and A. Gibali. The cyclic Douglas-Rachford algorithm with r-sets-Douglas-Rachford operators. *Optimization Methods & Software*, 34(4):875–889, 2019.

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

[3] A. Auslender, M. Teboulle, and S. Ben-Tiba. Interior proximal and multiplier methods based on second order homogeneous kernels. *Mathematics of Operations Research*, 24(3):645–668, 1999.

[4] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42:330–348, 2017.

[5] H. H. Bauschke, J. M. Borwein, and P. L. Combettes. Bregman monotone optimization algorithms. *SIAM J. Control. Optim.*, 42(2):596–636, 2003.

[6] H. H. Bauschke and W. M. Moursi. On the Douglas-Rachford algorithm. *Mathematical Programming Series A*, 164:263–284, 2017.

[7] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methodsfor convex optimization. *Operations Research Letters*, 31:167–175, 2003.

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

[9] A. Ben-Tal and M. Zibulevsky. Penalty/barrier multiplier methods for convex programming problems. *SIAM J. Optimization*, 7(2):347–366, 1997.

[10] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, 1996.

[11] J. M. Borwein and B. Sims. The Douglas-Rachford algorithm in the absence of convexity. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 93–109. Springer, New York, NY, 2011.

[12] R. I. Bot and A. Bohm. An incremental mirror descent subgradient algorithm with random sweeping and proximal step. *Optimization*, 68(1):33–50, 2019.

[13] R. I. Bot, E. R. Csetnek, and D. Meier. Variable metric ADMM for solving variational inequalities with monotone operators over affine sets. In H. H. Bauschke, R. S. Burachik, and D. R. Luke, editors, *Splitting Algorithms, Modern Operator Theory, and Applications*, pages 91–112. Springer, 2019.

[14] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[15] L. Bregman. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[16] M. N. Bùi and P. L. Combettes. Bregman forward-backward operator splitting. *Set-Valued and Variational Analysis*, 2020.

[17] M. N. Bùi and P. L. Combettes. The Douglas-Rachford algorithm converges only weakly. *SIAM Journal on Control and Optimization*, 58(2):1118–1120, 2020.

[18] Y. Censor and A. Lent. An iterative row-action method for interval convex programming. *Journal of Optimization Theory and Applications*, 34:321–353, 1981.

[19] Y. Censor and S. A. Zenios. The proximal minimization algorithm with D-functions. *J. Optim. Theory Appl.*, 73:451–464, 1992.

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

[21] P. L. Combettes. Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, 53:475–504, 2004.

[22] P. L. Combettes and Jean-Christophe Pesquet. A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):564–574, 2007.

[23] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[24] A. R. De Pierro and A. N. Iusem. A relaxed version of Bregman's methodfor convex programming. *J. Optim. Theory Appl.*, 51(3):421–440, 1986.

[25] J. Douglas and H. H. Rachford. On the numerical solution of the heat conduction problem in 2 and 3 space variables. *Transactions of the American Mathematical Society*, 82:421–439, 1956.

[26] J. Eckstein. *Splitting methods for monotone operators with applications to parallel optimization.* PhD thesis, Massachusetts Institute of Technology, 1989.

[27] J. Eckstein. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Mathematics of Operations Research*, 18(1):202–226, 1993.

[28] J. Eckstein. Approximate iterations in Bregman-function-based proximal algorithms. *Mathematical programming*, 83(1):113–123, 1998.

[29] J. Eckstein. A practical general approximation criterion for methods ofmultipliers based on bregman distances. *Mathematical Programming Series A*, 96:61–86, 2003.

[30] J. Eckstein and D. P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992.

[31] J. Eckstein and W. Yao. Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives. *Pacific Journal on Optimization*, 11(4):619–644, 2015.

[32] P. P. B. Eggermont. Multiplicative iterative algorithms for convex programming. *Linear Algebra and Its Applications*, 130:25–42, 1990.

[33] J. Eriksson. An iterative primal-dual algorithm for linear programming. Technical report, Technical Report 85-10, Department of Mathematics, Linkoping University, 1985.

[34] S. Erlander. Entropy in linear programs. *Mathematical Programming*, 21:137–151, 1981.

[35] M. Fortin and R. Glowinski. *Augmented Lagrangian methods: applications to the numerical solution of boundary-value problems.* North-Holland Pub. Co., 1983.

[36] D. Gabay. Applications of the method of multipliers to variational inequalities. In M. Fortin and R. Glowinski, editors, *Augmented Lagrangian Methods: Applications to the Solution of Boundary Value Problems.* North-Hollan, Amsterdam, 1983.

[37] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. *Comp. Math. Appl.*, 2:17–40, 1976.

[38] E. Ghadimi, A. Teixeira, I. Shames, and M. Johansson. Optimal parameterselection for the alternating direction method of multipliers (ADMM): Quadratic problems. *IEEE Trans. Automatic Control*, 60(3):644–658, 2015.

[39] P. Giselsson and S. Boyd. Linear convergence and metric selection for Douglas-Rachford splitting and ADMM. *IEEE Transantions on Automatic Control*, 62(2):532–544, 2017.

[40] R. Glowinski. On alternating direction methods of multipliers: A historical perspective. In W. Fitzgibbon, Y. A. Kuznetsov, P. Neittaanmäki, and O. Pironneau, editors, *Modeling, Simulation and Optimization for Scienceand Technology*, volume 34, pages 59–82. Springer Netherlands, Dordrecht, 2014.

[41] R. Glowinski and P. Le Tallec. *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*. SIAM, Philadelphia, Pennsylvania, 1989.

[42] D. Goldfarb, S. Ma, and K. Scheinberg. Fast alternating linearization methods for minimizing the sum of two convex functions. *Mathematical Programming*, 141(1-2):349–382, 2013.

[43] T. Goldstein and S. Osher. The split Bregman method for L1-regularized problems. *SIAM J. Imaging Sci.*, 2:323–343, 2009.

[44] I. Griva and R. A. Polyak. Primal-dual nonlinear rescaling method with dynamic scaling parameter update. *Math. Program. Ser. A*, 106:237–259, 2006.

[45] D. H. Gutman and J. F. Pena. A unified framework for Bregman proximal methods: subgradient, gradient, and accelerated gradient schemes. *http://www.optimization-online.org/DB_FILE/2018/12/6996.pdf*, 2018.

[46] F. Hanzely, P. Richtárik, and L. Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex minimization. *Computational Optimization and Applications*, 79(2):405–440, 2021.

[47] B. He and X. Yuan. On the convergence rate of Douglas-Rachford operator splitting method. *Mathematical Programming*, 153(2):715–722, 2015.

[48] A. N. Iusem. On some properties of generalized proximal point methods for quadratic and linear programming. *J. Optim. Theory Appl.*, 96:337–362, 1998.

[49] A. N. Iusem. Augmented Lagrangian methods and proximal point methods for convex optimization. *Investigación Operativa*, 8:11–49, 1999.

[50] A. N. Iusem, B. F. Svaiter, and M. Teboulle. Entropy-like proximal methods in convex programming. *Mathematics of Operations Research*, 19(4):790–814, 1994.

[51] A. N. Iusem and M. Teboulle. Convergence rate analysis of nonquadratic proximal methods for convex and linear programming. *Mathematics of Operations Research*, 20(3):657–677, 1995.

[52] K. C. Kiwiel. Proximal minimization methods with generalized Bregman functions. *SIAM Journal on Control and Optimization*, 35(4):1142–1168, 1997.

[53] B. W. Kort and D. P. Bertsekas. A new penalty function method for constrained minimization. In *Proceedings of the IEEE conference on decision and control*, pages 162–166, 1972.

[54] G. Li and T. K. Pong. Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. *Math. Program.*, 159:371–401, 2016.

[55] J. Liang, J. Fadili, and G. Peyré. Local convergence properties of Douglas-Rachford and alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, 172(3):874–913, 2017.

[56] S. B. Lindstrom and B. Sims. Survey: Sixty years of Douglas-Rachford. *Journal of the Australian Mathematical Society*, 110(3):333–370, 2021.

[57] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16:964–979, 1979.

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

[59] S. Ma and N. S. Aybat. Efficient optimization algorithms for robust principal component analysis and its variants. *Proceedings of the IEEE*, 106(8):1411–1426, 2018.

[60] S. Ma and M. Hong. A gentle introduction to ADMM for statistical problems. In *Handbook of Computational Statistics and Data Science*. John Wiley & Sons, 2020.

[61] B. Martinet. Régularisation d'inéquations variationelles par approximations successives. *Rev. Francqise Informat. Recherche Opérationelle*, 4:154–159, 1970.

[62] A. S. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics, John Wiley, 1983.

[63] Y. E. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269:543–547, 1983.

[64] Y. E. Nesterov. *Introductory lectures on convex optimization: A basic course*. Applied Optimization. Kluwer Academic Publishers, Boston, MA, 2004.

[65] Y. E. Nesterov. Smooth minimization for non-smooth functions. *Math. Program. Ser. A*, 103:127–152, 2005.

[66] G. B. Passty. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 72:383–390, 1979.

[67] D. H. Peaceman and H. H. Rachford. The numerical solution of parabolic elliptic differential equations. *SIAM Journal on Applied Mathematics*, 3:28–41, 1955.

[68] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[69] R. A. Polyak. Controlled processes in extremal and equilibrium problems. *VINITI, deposited manuscript, Moscow (in Russian)*, 1986.

[70] R. A. Polyak. The nonlinear rescaling principle in linear programming. Technical Report IBM Research Report RC15030, IBM T. J. Watson Research Center (Yorktown Heights, NY), 1989.

[71] R. A. Polyak and I. Griva. Primal-dual nonlinear rescaling method for convex optimization. *Journal of Optimization Theory and Applications*, 122(1):111–156, 2004.

[72] R. A. Polyak and M. Teboulle. Nonlinear rescaling and proximal-like methods in convex optimization. *Math. Program.*, 76:265–284, 1997.

[73] S. Reich. A weak convergence theorem for the alternating method with Bregman distances. In *Theory and Applications of Nonlinear Operators of Accretive and Monotone Type*, pages 313–318. Marcel Dekker, New York, 1996.

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

[75] R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Math. Oper. Res.*, 1(2):97–116, 1976.

[76] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.*, 14(5):877–898, 1976.

[77] E. K. Ryu. Uniqueness of DRS as the 2 operator resolvent-splitting and impossibility of 3 operator resolvent-splitting. *Mathematical Programming Series A*, 2020.

[78] E. K. Ryu, Y. Liu, and W. Yin. Douglas-Rachford splitting and ADMM for pathological convex optimization. *Computational Optimization and Applications*, 2019.

[79] E. K. Ryu and W. Yin. *Large-Scale Convex Optimization via Monotone Operators*. Cambridge University Press (to be published), 2022.

[80] S. Sabach. Products of finitely many resolvents of maximal monotone mappings in reflexive Banach spaces. *SIAM Journal on Optimization*, 21:1289–1308, 2011.

[81] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21:343–348, 1967.

[82] M. V. Solodov and B. F. Svaiter. An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions. *Mathematics of Operations Research*, 25:214–230, 2000.

[83] B. F. Svaiter. Weak convergence on Douglas-Rachford method. *SIAM Journal on Control and Optimization*, 49(1):280–287, 2011.

[84] M. Teboulle. Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research*, 17(3):670–690, 1992.

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

[86] P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Math. Program. Ser. B*, 125:263–295, 2010.

[87] P. Tseng and D. P. Bertsekas. One the convergence of the exponential multiplier method for convex programming. *Mathematical Programming*, 60:1–19, 1993.

[88] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[89] H. Wang and A. Banerjee. Bregman alternating direction method of multipliers. In *NIPS*, 2014.

[90] Y. Wang, J. Yang, W. Yin, and Y. Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Sciences*, 1(3):248–272, 2008.

[91] Z. Wen, D. Goldfarb, and W. Yin. Alternating direction augmented Lagrangian methods for semidefinite programming. *Mathematical Programming Computation*, 2:203–230, 2010.

[92] S. Yan and N. He. Bregman augmented Lagrangian and its acceleration. *https://arxiv.org/abs/2002.06315*, 2020.

[93] L. Yang and K.-C. Toh. Bregman proximal point algorithm revisited: A new inexact version and its variant. *https://arxiv.org/abs/2105.10370*, 2021.

A The Convergence Analysis of BPRS

Note that an interesting observation to the operators defined in Definition 2.3 is that

$$R_T^h = F_T^h \circ J_T^h.$$

Therefore the BPRS (4.1) is equivalent to

$$z^{k+1} = F_{\gamma_k A}^h \circ J_{\gamma_k A}^h \circ F_{\gamma_k B}^h \circ J_{\gamma_k B}^h(z^k).$$

Let $x^k = J_{\gamma_k A}^h \circ F_{\gamma_k B}^h \circ J_{\gamma_k B}^h(z^k)$, then the above equation becomes

$$x^{k+1} = J_{\gamma_k A}^h \circ F_{\gamma_k B}^h \circ J_{\gamma_k B}^h \circ F_{\gamma_k A}^h(x^k). \tag{A.1}$$

Therefore, the BPRS is the composition of two Bregman forward-backward operators. Moreover it can be shown that

$$\mathrm{Fix}(J_{\gamma_k A}^h \circ F_{\gamma_k B}^h \circ J_{\gamma_k B}^h \circ F_{\gamma_k A}^h) = (A+B)^{-1}(0). \tag{A.2}$$

A.1 The convergence of BPRS when both f and g are relatively smooth.

In this section, we provide a convergence analysis of BPRS when both f and g are relative smooth functions with respect to D_h, i.e.,

$$f(x) \le f(y) + \langle \nabla f(y), x - y\rangle + L D_h(x,y)$$
$$g(x) \le g(y) + \langle \nabla g(y), x - y\rangle + L D_h(x,y),$$

where $L > 0$ is the relative smoothness parameter. We consider the smooth problem

$$\min_x \; F(x) = f(x) + g(x) \Longleftrightarrow 0 \in A(x) + B(x), \tag{A.3}$$

where $A = \nabla f$ and $B = \nabla g$. Our convergence result is summarized in Theorem A.1.

Theorem A.1. *Assume f and g are relative smooth functions with respect to D_h with parameter L. BPRS with $\gamma_k = \gamma \le 1/L$ for solving* (A.3) *globally converges to $(A+B)^{-1}(0)$.*

Proof. BPRS with $\gamma_k = \gamma \le 1/L$ for solving (A.3) can be rewritten as

$$y^k := \operatorname*{argmin}_y \; f(x^k) + \langle \nabla f(x^k), y - x^k\rangle + g(y) + \frac{1}{\gamma} D_h(y, x^k) \tag{A.4a}$$

$$x^{k+1} := \operatorname*{argmin}_x \; g(y^k) + \langle \nabla g(y^k), x - y^k\rangle + f(x) + \frac{1}{\gamma} D_h(x, y^k). \tag{A.4b}$$

It is easy to obtain that:

$$\begin{aligned} F(y^k) \leq & F(x^k) - \frac{1}{\gamma} D_h(x^k, y^k) \\ F(x^{k+1}) \leq & F(y^k) - \frac{1}{\gamma} D_h(y^k, x^{k+1}), \end{aligned}$$

which further yields

$$F(x^{k+1}) \leq F(x^k) - \frac{1}{\gamma} D_h(x^k, y^k) - \frac{1}{\gamma} D_h(y^k, x^{k+1}).$$

This inequality shows that $F(x^k)$ monotonically decreases. Moreover, by telescoping sum, we obtain:

$$\frac{1}{\gamma} \sum_{k=0}^{N} \Big(D_h(x^k, y^k) + D_h(y^k, x^{k+1}) \Big) \leq F(x^0) - F(x^*).$$

This indicates

$$\lim_{k\to\infty} D_h(x^k, y^k) = 0, \text{ and } \lim_{k\to\infty} D_h(y^k, x^{k+1}) = 0.$$

This further implies

$$\lim_{k\to\infty} \|x^k - y^k\| = 0, \quad \lim_{k\to\infty} \|x^{k+1} - y^k\| = 0, \quad \lim_{k\to\infty} \|x^k - x^{k+1}\| = 0.$$

This shows that $\{x^k\}$ is a Cauchy sequence. Therefore, $\{x^k\}$ is bounded and convergent. From (A.1) and (A.2) we know that $\{x^k\}$ converges to $(A+B)^{-1}(0)$. □

A.2 The convergence of BPRS when both f and g are non-smooth functions.

Note that even when both f and g are nonsmooth, we can still apply (A.1) with $A = \partial f$ and $B = \partial g$. In this case, (A.1) reduces to (with $x^0 := J^h_{\gamma_k A} \circ F^h_{\gamma_k B} \circ J^h_{\gamma_k B}(z^0)$)

$$x^{k+1} = (\nabla h + \gamma_k \partial f)^{-1}(\nabla h - \gamma_k \partial g)(\nabla h + \gamma_k \partial g)^{-1}(\nabla h - \gamma_k \partial f)(x^k). \tag{A.5}$$

In this section we prove the convergence of the BPRS for solving

$$\min_x \ f(x) + g(x) \tag{A.6}$$

when both f and g are nonsmooth functions, under the assumption that $\mathrm{im}(\nabla h^*) \subset \mathrm{dom}\ f \cap \mathrm{dom}\ g$. This assumption was used in [12]. Note that according to (A.5), the BPRS for solving (A.6) can be written as

$$\bar{x}^k := \nabla h^*(\nabla h(w^k) - \gamma_k \partial f(w^k)) \tag{A.7a}$$

$$x^{k+1} := \operatorname*{argmin}_x \ g(x) + \frac{1}{\gamma_k} D_h(x, \bar{x}^k) \tag{A.7b}$$

$$\bar{w}^{k+1} := \nabla h^*(\nabla h(x^{k+1}) - \gamma_k \partial g(x^{k+1})) \tag{A.7c}$$

$$w^{k+1} := \operatorname*{argmin}_w \ f(w) + \frac{1}{\gamma_k} D_h(w, \bar{w}^{k+1}). \tag{A.7d}$$

Note that this is an alternating Bregman proximal subgradient method. Bot and Bohm [12] analyzed the convergence of Bregman proximal subgradient method, which consists only one step of (A.7).

We will use the *Three point identity:*

$$D_h(x,y) + D_h(y,z) = D_h(x,z) - \langle \nabla h(y) - \nabla h(z), x - y\rangle. \tag{A.8}$$

Our main result of the convergence of BPRS is summarized in Theorem A.2.

Theorem A.2. *Assume* $im(\nabla h^*) \subset dom\ f \cap dom\ g$, $\|\partial f\|_\infty \le G$, $\|\partial g\|_\infty \le G$, h *is* σ*-strongly convex over the simplex, and* ∇h^* *is Lipschitz continuous. Then BPRS* (A.7) *with* $\gamma_k = 1/\sqrt{k}$ *for solving* (A.6) *converges sublinearly.*

Proof. First, we have

$$\begin{aligned}
D_h(y,\bar{x}^k) &\overset{(A.8)}{=\!=} D_h(y,w^k) - D_h(\bar{x}^k,w^k) - \langle \nabla h(\bar{x}^k) - \nabla h(w^k), y - \bar{x}^k\rangle \\
&= D_h(y,w^k) - D_h(\bar{x}^k,w^k) + \gamma_k\langle \partial f(w^k), y - \bar{x}^k\rangle \\
&= D_h(y,w^k) - D_h(\bar{x}^k,w^k) + \gamma_k\langle \partial f(w^k), y - w^k\rangle - \gamma_k\langle \partial f(w^k), \bar{x}^k - w^k\rangle \\
&\le D_h(y,w^k) - D_h(\bar{x}^k,w^k) + \gamma_k(f(y) - f(w^k)) + \gamma_k\|\partial f(w^k)\|_\infty\|\bar{x}^k - w^k\|_1 \\
&\le D_h(y,w^k) - D_h(\bar{x}^k,w^k) + \gamma_k(f(y) - f(w^k)) + \frac{\gamma_k^2}{\sigma}\|\partial f(w^k)\|_\infty^2 + \frac{\sigma}{4}\|\bar{x}^k - w^k\|_1^2 \\
&\overset{(*)}{\le} D_h(y,w^k) - D_h(\bar{x}^k,w^k) + \gamma_k(f(y) - f(w^k)) + \frac{\gamma_k^2}{\sigma}\|\partial f(w^k)\|_\infty^2 + \frac{1}{2}D_h(\bar{x}^k,w^k) \\
&\overset{(**)}{\le} D_h(y,w^k) - \frac{1}{2}D_h(\bar{x}^k,w^k) + \gamma_k(f(y) - f(w^k)) + \frac{\gamma_k^2}{\sigma}G^2,
\end{aligned} \tag{A.9}$$

where (*) is due to the σ-strong convexity of h over the simplex (also note that both $\bar{x}^k$ and w^k are on simplex), in (**) we used $\|\partial f\|_\infty \le G$.

Now, the optimality condition of (A.7b) is:

$$0 \in \gamma_k \partial g(x^{k+1}) + \nabla h(x^{k+1}) - \nabla h(\bar{x}^k),$$

which implies

$$\gamma_k(g(y) - g(x^{k+1})) \ge \langle \nabla h(\bar{x}^k) - \nabla h(x^{k+1}), y - x^{k+1}\rangle.$$

Using the three-point identity we have

$$\gamma_k(g(y) - g(x^{k+1})) \ge D_h(y,x^{k+1}) + D_h(x^{k+1},\bar{x}^k) - D_h(y,\bar{x}^k),$$

or, equivalently,

$$\gamma_k(g(x^{k+1}) - g(y)) + D_h(y,x^{k+1}) \le D_h(y,\bar{x}^k) - D_h(x^{k+1},\bar{x}^k). \tag{A.10}$$

Combining (A.9) and (A.10), we have

$$\gamma_k(g(x^{k+1}) - g(y)) + \gamma_k(f(w^k) - f(y)) + D_h(y,x^{k+1}) \le D_h(y,w^k) + \frac{\gamma_k^2}{\sigma}G^2 - D_h(x^{k+1},\bar{x}^k) - \frac{1}{2}D_h(\bar{x}^k,w^k).$$

By adding and subtracting $f(x^{k+1})$, we obtain

$$\begin{aligned}
&\gamma_k(f(x^{k+1}) + g(x^{k+1}) - f(y) - g(y)) + \gamma_k(f(w^k) - f(x^{k+1})) + D_h(y,x^{k+1}) \\
\le& D_h(y,w^k) + \frac{\gamma_k^2}{\sigma}G^2 - D_h(x^{k+1},\bar{x}^k) - \frac{1}{2}D_h(\bar{x}^k,w^k),
\end{aligned}$$

which immediately implies

$$
\begin{aligned}
&\gamma_k(f(x^{k+1})+g(x^{k+1})-f(y)-g(y))+D_h(y,x^{k+1})\\
\le& D_h(y,w^k)+\frac{\gamma_k^2}{\sigma}G^2-D_h(x^{k+1},\bar{x}^k)-\frac{1}{2}D_h(\bar{x}^k,w^k)+\gamma_k G\|w^k-x^{k+1}\|_1\\
\overset{(*)}{\le}& D_h(y,w^k)+\frac{\gamma_k^2}{\sigma}G^2-D_h(x^{k+1},\bar{x}^k)-\frac{1}{2}D_h(\bar{x}^k,w^k)+\gamma_k G\|w^k-\bar{x}^k\|_1+\gamma_k G\|\bar{x}^k-x^{k+1}\|_1\\
\overset{(**)}{\le}& D_h(y,w^k)+\frac{\gamma_k^2}{\sigma}G^2-D_h(x^{k+1},\bar{x}^k)-\frac{1}{2}D_h(\bar{x}^k,w^k)+\gamma_k G\|w^k-\bar{x}^k\|_1+\frac{\gamma_k^2G^2}{2\sigma}+D_h(x^{k+1},\bar{x}^k)\\
=& D_h(y,w^k)+\frac{\gamma_k^2}{\sigma}G^2-\frac{1}{2}D_h(\bar{x}^k,w^k)+\gamma_k G\|w^k-\bar{x}^k\|_1+\frac{\gamma_k^2G^2}{2\sigma}\\
\overset{(***)}{\le}& D_h(y,w^k)+\frac{\gamma_k^2}{\sigma}G^2-\frac{1}{2}D_h(\bar{x}^k,w^k)+\frac{\gamma_k^2G^2}{\sigma}+\frac{\sigma}{4}\|w^k-\bar{x}^k\|_1^2+\frac{\gamma_k^2G^2}{2\sigma}\\
\overset{(****)}{\le}& D_h(y,w^k)+\frac{\gamma_k^2}{\sigma}G^2+\frac{\gamma_k^2G^2}{\sigma}+\frac{\gamma_k^2G^2}{2\sigma}\\
=& D_h(y,w^k)+\frac{5\gamma_k^2}{2\sigma}G^2,
\end{aligned}
$$

where (*) is the triangle inequality, (**) is due to Young's inequality and the σ-strong convexity of h, (***) is due to Young's inequality, and (****) is due to the σ-strong convexity of h. Note that so far we only dealt with (A.7a)-(A.7b), and we obtained

$$\gamma_k(f(x^{k+1})+g(x^{k+1})-f(y)-g(y))+D_h(y,x^{k+1})\le D_h(y,w^k)+\frac{5\gamma_k^2}{2\sigma}G^2. \tag{A.11}$$

Similarly, for (A.7c)-(A.7d), we have

$$\gamma_k(f(w^{k+1})+g(w^{k+1})-f(y)-g(y))+D_h(y,w^{k+1})\le D_h(y,x^{k+1})+\frac{5\gamma_k^2}{2\sigma}G^2. \tag{A.12}$$

Now summing up (A.11) and (A.12) we have (denote $L=\frac{5G^2}{2\sigma}$)

$$\gamma_k(f(x^{k+1})+g(x^{k+1})+f(w^{k+1})+g(w^{k+1})-2(f(y)+g(y)))+D_h(y,w^{k+1})\le D_h(y,w^k)+2L\gamma_k^2. \tag{A.13}$$

Define $z^{k+1}:=(x^{k+1}+w^{k+1})/2$. From the convexity of f and g, we have

$$\gamma_k(f(z^{k+1})+g(z^{k+1})-(f(y)+g(y)))+\frac{1}{2}D_h(y,w^{k+1})\le\frac{1}{2}D_h(y,w^k)+L\gamma_k^2. \tag{A.14}$$

Now summing up (A.14) for $k=0,1,\dots,N-1$, we have

$$\sum_{k=0}^{N-1}\gamma_k(f(z^{k+1})+g(z^{k+1})-(f(y)+g(y)))\le\frac{1}{2}D_h(y,w^0)+L\sum_{k=0}^{N-1}\gamma_k^2. \tag{A.15}$$

Therefore, we have

$$\min_{0\le k\le N-1}\Big(f(z^{k+1})+g(z^{k+1})\Big)-(f(y)+g(y))\le\frac{D_h(y,w^0)}{2\sum_{k=0}^{N-1}\gamma_k}+\frac{L\sum_{k=0}^{N-1}\gamma_k^2}{\sum_{k=0}^{N-1}\gamma_k}. \tag{A.16}$$

By choosing $y=x^*$ and using $\gamma_k=\frac{1}{\sqrt{k}}$, we obtain

$$\lim_{N\to+\infty}\min_{0\le k\le N-1}\Big(f(z^{k+1})+g(z^{k+1})\Big)=f(x^*)+g(x^*).$$

Moreover, the convergence rate is sublinear. □

Remark A.3. *Note that we required that ∇h^* is Lipschitz continuous. This is satisfied if h is the entropy function: $h(x) = \sum_i (x_i \log x_i - x_i)$. In this case, $\nabla h(x) = \log x$, and the conjugate function on probability simplex is given by:*

$$h^*(x) = \sup_{e^\top y = 1} x^\top y - h(y).$$

It is easy to verify that the optimal y is given by $y_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$, and

$$h^*(x) = \log \sum_i e^{x_i} + 1. \tag{A.17}$$

Note that this is $\log \sum \exp$ function, and it is known to have Lipschitz continuous gradient. And, we have $[\nabla h^*(x)]_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$.

B Convergence of BDRS

In this section we consider the same problem as in Section A.2 under the same assumptions in Theorem A.2. Note that BDRS can be equivalently written as

$$w^k := \operatorname*{argmin}_w \; f(w) + \frac{1}{\gamma_k} D_h(w, x^k) \tag{B.1a}$$
$$\bar{x}^k := \nabla h^*(2\nabla h(w^k) - \nabla h(x^k)) \tag{B.1b}$$
$$z^k := \operatorname*{argmin}_z \; g(z) + \frac{1}{\gamma_k} D_h(z, \bar{x}^k) \tag{B.1c}$$
$$y^k := \nabla h^*(2\nabla h(z^k) - \nabla h(\bar{x}^k)) \tag{B.1d}$$
$$x^{k+1} := \nabla h^* \left(\frac{1}{2}\nabla h(x^k) + \frac{1}{2}\nabla h(y^k) \right). \tag{B.1e}$$

This is equivalent to the following when considering problem (A.6):

$$w^k := \operatorname*{argmin}_w \; f(w) + \frac{1}{\gamma_k} D_h(w, x^k) \tag{B.2a}$$
$$\bar{x}^k := \nabla h^*(\nabla h(w^k) - \gamma_k \partial f(w^k)) \tag{B.2b}$$
$$z^k := \operatorname*{argmin}_z \; g(z) + \frac{1}{\gamma_k} D_h(z, \bar{x}^k) \tag{B.2c}$$
$$y^k := \nabla h^*(\nabla h(z^k) - \gamma_k \partial g(z^k)) \tag{B.2d}$$
$$x^{k+1} := \nabla h^* \left(\frac{1}{2}\nabla h(x^k) + \frac{1}{2}\nabla h(y^k) \right). \tag{B.2e}$$

Theorem B.1. *Assume $\mathrm{im}(\nabla h^*) \subset \mathrm{dom}\, f \cap \mathrm{dom}\, g$, $\|\partial f\|_\infty \le G$, $\|\partial g\|_\infty \le G$, h is σ-strongly convex over the simplex, and ∇h^* is Lipschitz continuous. Then BDRS* (B.2) *with $\gamma_k = 1/\sqrt{k}$ for solving* (A.6) *converges sublinearly.*

Note that (B.2a) is one step of Bregman PPA, which yields

$$\gamma_k (f(y) - f(x^{k+1})) \ge D_h(y, x^{k+1}) - D_h(y, x^k) + D_h(x^{k+1}, x^k), \quad \forall y. \tag{B.3}$$

(B.2b) is one step of the mirror subgradient method, which yields

$$\gamma_k(f(x^k) - f(y)) \le D_h(y, x^k) - D_h(y, x^{k+1}) - \frac{1}{2}D_h(x^{k+1}, x^k) + \frac{\gamma_k^2 \|f'(x^k)\|_*^2}{2\sigma}, \quad \forall y. \tag{B.4}$$

Apply (B.4) to (B.2b) and (B.2d), we have (we ignored the constant coefficient of γ_k^2)

$$\gamma_k(f(w^k) - f(y)) \le D_h(y, w^k) - D_h(y, \bar{x}^k) - \frac{1}{2}D_h(\bar{x}^k, w^k) + \gamma_k^2$$
$$\gamma_k(g(z^k) - g(y)) \le D_h(y, z^k) - D_h(y, y^k) - \frac{1}{2}D_h(y^k, z^k) + \gamma_k^2.$$

Apply (B.3) to (B.2a) and (B.2c), we have

$$\gamma_k(f(w^k) - f(y)) \le D_h(y, x^k) - D_h(y, w^k) - D_h(w^k, x^k)$$
$$\gamma_k(g(z^k) - g(y)) \le D_h(y, \bar{x}) - D_h(y, z^k) - D_h(z^k, \bar{x}^k).$$

Combining these four inequalities, we have

$$\begin{aligned} &2\gamma_k(f(w^k) + g(z^k) - f(y) - g(y)) \\ \le &D_h(y, x^k) - D_h(y, y^k) - \frac{1}{2}D_h(y^k, z^k) - \frac{1}{2}D_h(\bar{x}^k, w^k) - D_h(w^k, x^k) - D_h(z^k, \bar{x}^k) + \gamma_k^2. \end{aligned} \tag{B.5}$$

Note that (B.2e) is equivalent to

$$\nabla h(x^{k+1}) - \nabla h(y^k) = \nabla h(x^k) - \nabla h(x^{k+1}). \tag{B.6}$$

Using the 3-point identity (A.8) twice, we have

$$D_h(y, x^{k+1}) = D_h(y, y^k) - D_h(x^{k+1}, y^k) - \langle \nabla h(x^{k+1}) - \nabla h(y^k), y - x^{k+1} \rangle$$
$$\langle \nabla h(x^k) - \nabla h(x^{k+1}), y - x^{k+1} \rangle = D_h(y, x^{k+1}) + D_h(x^{k+1}, x^k) - D_h(y, x^k).$$

Combining these two identities with (B.6), we obtain

$$D_h(y, x^{k+1}) = D_h(y, y^k) - D_h(x^{k+1}, y^k) - D_h(y, x^{k+1}) - D_h(x^{k+1}, x^k) + D_h(y, x^k),$$

which is equivalent to

$$2D_h(y, x^{k+1}) = D_h(y, x^k) + D_h(y, y^k) - D_h(x^{k+1}, y^k) - D_h(x^{k+1}, x^k). \tag{B.7}$$

Combining (B.5) and (B.7), we have

$$\begin{aligned} &2\gamma_k(f(w^k) + g(z^k) - f(y) - g(y)) + 2D_h(y, x^{k+1}) \\ \le &2D_h(y, x^k) - \frac{1}{2}D_h(y^k, z^k) - \frac{1}{2}D_h(\bar{x}^k, w^k) - D_h(w^k, x^k) - D_h(z^k, \bar{x}^k) - D_h(x^{k+1}, y^k) - D_h(x^{k+1}, x^k) + \gamma_k^2. \end{aligned} \tag{B.8}$$

Under the assumption that $\mathrm{im}(\nabla h^*) \subset \mathrm{dom}\, f \cap \mathrm{dom}\, g$, we know $f(x^{k+1})$ and $g(x^{k+1})$ are finite values. By adding and subtracting $f(x^{k+1}) + g(x^{k+1})$ to (B.8), we get

$$\begin{aligned} &2\gamma_k(f(x^{k+1}) + g(x^{k+1}) - f(y) - g(y)) + 2D_h(y, x^{k+1}) \\ \le &2D_h(y, x^k) - \frac{1}{2}D_h(y^k, z^k) - \frac{1}{2}D_h(\bar{x}^k, w^k) - D_h(w^k, x^k) - D_h(z^k, \bar{x}^k) - D_h(x^{k+1}, y^k) - D_h(x^{k+1}, x^k) \\ &+ 2\gamma_k(f(x^{k+1}) - f(w^k) + g(x^{k+1}) - g(z^k)) + \gamma_k^2. \end{aligned} \tag{B.9}$$

Now we only need to show that $2\gamma_k(f(x^{k+1}) - f(w^k) + g(x^{k+1}) - g(z^k))$ can be bounded by those negative terms on the right-hand-side. We have

$$\begin{aligned}
&2\gamma_k(f(x^{k+1}) - f(w^k) + g(x^{k+1}) - g(z^k)) \\
\leq &2\gamma_k G(\|x^{k+1} - w^k\|_1 + \|x^{k+1} - z^k\|_1) \\
\leq &2\gamma_k G(\|x^{k+1} - w^k\|_1 + \|z^k - w^k\|_1 + \|x^{k+1} - w^k\|_1) \\
= &2\gamma_k G(2\|x^{k+1} - w^k\|_1 + \|z^k - w^k\|_1) \\
\leq &2\gamma_k G(2\|x^k - x^{k+1}\|_1 + 2\|w^k - x^k\|_1 + \|z^k - w^k\|_1) \\
\leq &2\gamma_k G(2\|x^k - x^{k+1}\|_1 + 2\|w^k - x^k\|_1 + \|z^k - \bar{x}^k\|_1 + \|\bar{x}^k - w^k\|_1)
\end{aligned}$$

Now apply Young's inequality to these four terms above, we obtain

$$\begin{aligned}
&2\gamma_k(f(x^{k+1}) - f(w^k) + g(x^{k+1}) - g(z^k)) \\
\leq &2\gamma_k G(2\|x^k - x^{k+1}\|_1 + 2\|w^k - x^k\|_1 + \|z^k - \bar{x}^k\|_1 + \|\bar{x}^k - w^k\|_1) \\
\leq &\left(\frac{8\gamma_k^2 G^2}{\sigma} + \frac{\sigma}{2}\|x^{k+1} - x^k\|_1^2\right) + \left(\frac{8\gamma_k^2 G^2}{\sigma} + \frac{\sigma}{2}\|w^k - x^k\|_1^2\right) \\
&\left(\frac{2\gamma_k^2 G^2}{\sigma} + \frac{\sigma}{2}\|z^k - \bar{x}^k\|_1^2\right) + \left(\frac{4\gamma_k^2 G^2}{\sigma} + \frac{\sigma}{4}\|\bar{x}^k - w^k\|_1^2\right) \\
\leq &\gamma_k^2 + D_h(x^{k+1}, x^k) + D_h(w^k, x^k) + D_h(z^k, \bar{x}^k) + \frac{1}{2}D_h(\bar{x}^k, w^k),
\end{aligned}$$

where in the last inequality we omitted the coefficient of γ_k^2, and we used the σ-strong convexity of h. Combining this inequality with (B.9), we get

$$2\gamma_k(f(x^{k+1}) + g(x^{k+1}) - f(y) - g(y)) + 2D_h(y, x^{k+1}) \leq 2D_h(y, x^k) + \gamma_k^2.$$

Now summing this over $k = 0, 1, \ldots, N-1$, we get

$$\sum_{k=0}^{N-1} \gamma_k(f(x^{k+1}) + g(x^{k+1}) - f(y) - g(y)) \leq D_h(y, x^0) + \sum_{k=0}^{N-1} \gamma_k^2.$$

That is,

$$\min_{0\leq k\leq N-1} \left(f(x^{k+1}) + g(x^{k+1})\right) - (f(y) + g(y)) \leq \frac{D_h(y, x^0)}{\sum_{k=0}^{N-1} \gamma_k} + \frac{\sum_{k=0}^{N-1} \gamma_k^2}{\sum_{k=0}^{N-1} \gamma_k}.$$

By choosing $y = x^*$ and using $\gamma_k = \frac{1}{\sqrt{k}}$, we obtain the convergence and the sublinear rate.