J. Oper. Res. Soc. China (2015) 3:117–138
DOI 10.1007/s40305-015-0073-3

Distributed and Consensus Optimization for Non-smooth Image Reconstruction

Xiao-Jing Ye

Received: 10 October 2014 / Revised: 13 January 2015 / Accepted: 17 January 2015 /
Published online: 18 February 2015

Abstract We present a primal-dual algorithm with consensus constraints that can effectively handle large-scale and complex data in a variety of non-smooth image reconstruction problems. In particular, we focus on the case that the data fidelity term can be decomposed into multiple relatively simple functions and deployed to parallel computing units to cooperatively solve for a consensual solution of the original problem. In this case, the subproblems usually have closed form solution or can be solved efficiently at local computing units, and hence the per-iteration computation complexity is very low. A comprehensive convergence analysis of the algorithm, including convergence rate, is established.

Keywords Distributed optimization · Parallel computing · Consensus · Total variation · Image reconstruction

Mathematics Subject Classification 49N45

1 Introduction

1.1 Problem Formulation

In recent years, there are extensive interests in solving non-smooth image reconstruction problems in the form of

$$\min_{x\in X} \{F(x) + J(Kx)\}, \tag{1.1}$$

X.-J. Ye (✉)
Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, USA
e-mail: xye@gsu.edu

where x is the image to be solved and here represented as a vector in $\mathbb{R}^n$, J is a non-smooth (non-differentiable), proper, convex, and lower semi-continuous (l.s.c.) simple function, K is a linear operator, X is the admissible set of the desired solution that usually describes a box constraint or non-negativity constraint on the solution image. In this paper, we are inspired by the fact that in a variety of real-world applications, the so-called data fidelity term $F(x)$, which models the physical or statistical relations between observed data and the unknown image x, can be decomposed into a number of relatively simple functions F_i for $i = 1, \cdots, m$ as

$$F(x) = \sum_{i=1}^{m} F_i(x). \tag{1.2}$$

Here for simplicity we mean that the proximity operators

$$(I + \alpha\partial J)^{-1}(y) := \underset{y\in\mathrm{dom}(J)}{\arg\min} \left\{ J(z) + \frac{1}{2\alpha}\|z - y\|^2 \right\}, \tag{1.3}$$

$$(I + \alpha\partial F_i)^{-1}(x) := \underset{x\in\mathrm{dom}(F_i)}{\arg\min} \left\{ F_i(z) + \frac{1}{2\alpha}\|z - x\|^2 \right\}, \quad i = 1, \cdots, m \tag{1.4}$$

can be evaluated or solved easily. Here $\|\cdot\|$ is the regular norm induced by inner product $\langle\cdot,\cdot\rangle$ defined on finite dimensional vector spaces (e.g., $\mathbb{R}^n$ and $\mathbb{C}^n$), $\mathrm{dom}(J)$ and $\mathrm{dom}(F_i)$ denote the (convex) domains of functions J and F_i, respectively, and $\alpha > 0$ is a constant. For example, in image reconstruction with the robust total-variation (TV) regularization proposed in [38], K represents gradient operator and $J(\cdot)$ is ℓ_1 (or ℓ_1-like) norm, such that $J(Kx)$ computes TV semi-norm of an image x. In this paper, we always use TV as example for analysis and numerical tests, and the results can be readily extended to other J and K combinations satisfying the conditions specified above. The data fidelity term $F(\cdot)$ can have different formats depending on specific applications and will be described in details later in this section. We also remark here that the proximity operators of J and F_i can be easily evaluated if and only if can be those of their Fenchel dual functions J^* and F_i^*, due to the Moreau's identity [31].

It is worth noting that, despite of the special formulation and simplicity of composition functions J and F_i, optimization (1.1) has a large range of applications in machine learning, statistical analysis, and signal processing besides image reconstruction. For example, in signal processing, K is a linear operator such as wavelet transform [13,14,28,40]. In group lasso [19,24,29], K is an indication matrix corresponding to group labeling, etc. In terms of data fidelity term $F(x)$, we observe that data fidelity term $F(x)$ often yields a decomposition into a sum of relatively simple functions $F_i(x)$ as in (1.2). In particular,

1. Least squares $F(x) = \frac{1}{2}\|Ax - b\|^2$, where $A = (a_1, \cdots, a_m)^{\mathrm{T}}$. This can be decomposed into simple functions $F_i(x) = |a_i^{\mathrm{T}}x - b_i|^2$ for $i = 1, \cdots, m$. Least squares are widely used under implicit assumption that the noise in data follows independent and identically distributed Gaussian.

2. The least ℓ_1 norm $F(x) = \|Ax - b\|_1$. This can be decomposed into simple functions $F_i(x) = |a_i^\mathrm{T} x - b_i|$. The ℓ_1 norm is robust to outliners as it is arising from the modeling of noise by Laplacian distribution (sometimes called double exponential distribution).
3. The Poisson noise where $F(x) = \sum_{i=1}^m \big(a_i^\mathrm{T} x - b_i \log(a_i^\mathrm{T} x + c)\big)$. This is derived from the log-likelihood of independent Poisson noise $b_i \sim \text{Poisson}(a_i^\mathrm{T} x + c)$ for $i = 1, \cdots, m$ where c is the given base intensity.
4. Logistic regression where $F(x) = \sum_{i=1}^m \log\big(1 + \exp(-b_i a_i^\mathrm{T} x)\big)$. For positivity test, a_i is the feature vector of sample i and $b_i \in \{-1, 1\}$ is its class label. The solution x returns a classifier.

The cases above cover a majority of image reconstruction problems in real-world applications. In some other cases, although closed form solutions to the subproblems are not available, one can readily derive a routine (using Newton's method for instance) to calculate the solution easily.

1.2 Our Contribution

The contributions of this paper are in two phases. First, we propose a consensus optimization model to break image reconstruction problem down to multiple subproblems which involve relatively simple functions and can be solved efficiently. In particular, the subproblems usually have closed form solution and can readily handle non-differentiable objectives. Examples include general least squares data fidelity term $\|Ax - b\|^2$, ℓ_1 data term $\|Ax - b\|_1$, data term derived from likelihood function of Poisson distribution and logistic regression for arbitrary matrix A, which cover a large variety of signal/image processing and machine learning applications. Moreover, the computation can be easily carried out in parallel under both centralized and decentralized settings. Second, we integrate the merits of primal-dual formulation of TV to tackle its non-differentiability issue and alternate direction method of multipliers (ADMM) to deal with consensus constraints. Different from classical primal-dual and ADMM methods, the proposed algorithm only approximately solves all the primal variables and dual variables (of both primal-dual formulation of TV term and the Lagrangian of consensus constraints) in each iteration so that the per-iteration complexity remains low. Nevertheless, the iterates generated by the algorithm are still proved to be convergent, and the rate is also given in terms of the (perturbed) gap function value.

1.3 Proposed Algorithm

In this paper, we propose and analyze a fast numerical algorithm for solving the optimization problem (1.1) by introducing consensus constraints. Depending on the restrictions by specific application, there are two types of consensus constraints that can be adopted: centralized version and decentralized version.

Centralized Version If there is a center computing cluster (fusion center) that can efficiently communicate with parallel computing units, or there is a shared memory that can be easily accessed by these units, then one can introduce auxiliary variables w_i as local copy of unit i, and require that all w_i equal to x stored at fusion center [5]. Hence the following consensus minimax problem can be proposed:

$$\min_{x\in X, w_i} \left\{ \sum_{i=1}^{m} F_i(w_i) + \max_{y\in Y} \left[\langle Kx, y\rangle - J^*(y) \right] : w_i = x, i = 1, 2, \cdots, m \right\}, \quad (1.5)$$

where the maximization is due to Fenchel transform $J(Kx) = \max_{y\in Y}[\langle Kx, y\rangle - J^*(y)]$ to overcome the non-differentiability issue of J, which is a commonly used technique for solving TV-based image reconstruction in recent years [8,11,18,23, 44]. More importantly, in this case, the i-th parallel computing unit can store partial data corresponding to F_i, and implement solvers for w_i and Lagrangian multiplier u_i (shown later). In each iteration, the updated w_i and u_i are acquired by the center cluster to compute x and y stored in a shared memory, which are then used by the parallel units to update their w_i and u_i in the next iteration.

Decentralized Version In this case, neither a central computing cluster nor a shared memory is available, and often the communication between parallel units can be restrictively limited. For example, in large-size wireless sensor network $G(V, E)$, the sensor nodes in V (also as computing units) form the network such that each node can only exchange information with a small amount of neighbor nodes due to excessive battery power consumption or data loss in long-distance signal transmission. Namely, nodes k and l can communicate only if $(k, l) \in E$. To properly address this issue, we need to set consensus constraints different from (1.1). In particular, one of the possible choices is to introduce auxiliary variable $x_{ij} (= x_{ji})$ for each edge (i, j) in the undirected network as [39] and impose constraints.

$$w_i = x_{ij}, \quad w_j = x_{ij}, \quad \text{for all } (i, j) \in E. \quad (1.6)$$

Due to this constraint, w_i of all nodes in a connected network should be equal. Hence we can rewrite (1.1) as the following consensus minimax problem:

$$\min_{w_i, x_{ij}} \left\{ \sum_{i=1}^{m} \left[F_i(w_i) + \frac{1}{m} \max_{y_i\in Y} \left[\langle Kw_i, y_i\rangle - J^*(y_i) \right] \right] : w_i = x_{ij}, \quad \forall i \in V, (i, j) \in E \right\}. \quad (1.7)$$

In this case, we can readily show that the proposed algorithm only requires exchanging updates between neighbor nodes. We also point out that each node i will privately compute its own dual variable y_i which may not converge to a consensus. However, this does not affect the consensus of w_i, and it can be readily shown that $y_i \in \partial J(Kw)$ for all i upon convergence, where w is the consensual value of all w_i and x_{ij}. In the end, only w (as w_i retrieved from any node i) is returned by the algorithm as a solution to (1.1).

In this paper, we focus on the centralized version (1.5), as the implementation and convergence analysis can be readily modified and applied to the decentralized version (1.7). The proposed algorithm is called COMMON, an abbreviation of consensus minimax optimization, and is summarized in Algorithm 1.

Algorithm 1 Primal-dual algorithm for COnsensus MiniMax OptimizatioN (COMMON) (1.5).

```
Initialize $x^0 = w_i^0 = u_i^0 = 0$ for $i = 1, \cdots, m$, and $y^0 = \bar{y}^0 \in Y$.
for $t = 0, 1, 2, \cdots$ do
    $x^{t+1} = \mathit{\Pi}_X \left( (1 + \alpha\delta m)^{-1} \left( \alpha\delta \sum_{i=1}^{m} (w_i^t + \frac{u_i^t}{\delta}) + x^t - \alpha K^{\mathrm{T}} \bar{y}^t \right) \right),$   (1.8)
    $w_i^{t+1} = \left( I + (\delta + \gamma^{-1})^{-1} \partial F_i \right)^{-1} \left( (\delta + \gamma^{-1})^{-1} (\delta x^{t+1} - u_i^t + \gamma^{-1} w_i^t) \right), \forall i,$   (1.9)
    $u_i^{t+1} = u_i^t - \delta(x^{t+1} - w_i^{t+1}), \forall i,$   (1.10)
    $y^{t+1} = \mathit{\Pi}_Y \left( y^t + \beta K x^{t+1} \right),$   (1.11)
    $\bar{y}^{t+1} = 2y^{t+1} - y^t.$   (1.12)
end for
```

We point out that a major advantage of COMMON is that the subproblems can be easily solved in parallel: $\mathit{\Pi}_X(x)$ corresponds to projection onto X that describes a box constraint $[\xi, \eta]$ or a non-negativity constraint, which can be implemented as component-wise proximity operator $\max(\xi, \min(\eta, x))$ or $\max(0, x)$, respectively; the functions F_i in (1.9) are simple or their proximity operators defined in (1.4) can be easily computed, for which a few examples are shown below; $\mathit{\Pi}_Y$ in (1.11) is again a projection onto Y, which is a simple threshold function for TV regularization, namely, $(\mathit{\Pi}_Y(y))_i = y_i / \max(1, \|y_i\|)$, where $y_i \in \mathbb{R}^2$ is the i-th component of $y \in \mathbb{R}^{2n}$ [44]. Therefore solutions of these subproblems require $O(n)$ computation complexity and hence the algorithm readily scale to problems of larger sizes.

Here we provide some examples that the proximity operator in (1.9) can be evaluated or calculated easily, according to the four cases listed in the previous subsection as follows. Note that simple shifting or change of variables can be combined with the results below to obtain proximity operators $(I + \mu^{-1}\partial F_i)^{-1}$.

1. Least squares. The proximity operator of F_i reduces to solving the following minimization with some given $a \in \mathbb{R}^n$, $b \in \mathbb{R}$, and $\mu > 0$,

$$\min_{w} \left\{ \frac{1}{2} |a^{\mathrm{T}} w - b|^2 + \frac{\mu}{2} \|w\|^2 \right\}, \tag{1.13}$$

   which has a closed form solution $w^* = \frac{ab}{\mu + \|a\|^2}$.

2. The ℓ_1 norm. The proximity operator of F_i reduces to solving the following minimization with some given $a \in \mathbb{R}^n$, $b \in \mathbb{R}$, and $\mu > 0$,

$$\min_w \left\{ |a^{\mathrm{T}} w - b| + \frac{\mu}{2} \|w\|^2 \right\}, \tag{1.14}$$

which has a closed form solution $w^* = -\text{sign}(b) \max(0, |b| - \frac{\|a\|^2}{\mu}) \frac{a}{\|a\|}$ if $a \neq 0$ and 0 otherwise.

3. The Poisson error. The proximity operator of F_i reduces to solving the following minimization with some given $a \in \mathbb{R}^n$, $c \in \mathbb{R}$, and $\mu > 0$,

$$\min_w \left\{ a^{\mathrm{T}} w - \log(a^{\mathrm{T}} w + c) + \frac{\mu}{2} \|w\|^2 \right\}, \tag{1.15}$$

which has a unique closed form solution $w^* = \frac{-b\mu - \|a\|^2 + \sqrt{(\|a\|^2 + b\mu)^2 + 4\mu\|a\|^2}}{2\mu\|a\|^2} a$ if $a \neq 0$ and 0 otherwise.

4. Logistic regression. The proximity operator of F_i reduces to solving the following minimization with some given $a \in \mathbb{R}^n$, $b \in \{-1, 1\}$, and $\mu > 0$,

$$\min_w \left\{ \log\left(1 + \exp(-ba^{\mathrm{T}} w)\right) + \frac{\mu}{2} \|w\|^2 \right\}. \tag{1.16}$$

The problem has the solution of form $w^* = \frac{rba}{\|a\|}$, where the scalar r can be computed from $\min_{r \geqslant 0} \{\log(1 + \mathrm{e}^{-\|a\| r}) + \frac{\mu r^2}{2}\}$ in a few iterations using Newton's method.

As we can see, the updates of w_i and u_i can be carried out in parallel. Since the computation for all variables are either direct or simple, the per-iteration computation complexity of COMMON is very low. In terms of memory cost, COMMON requires auxiliary variables $\{w_i\}_{i=1}^m$ and $\{u_i\}_{i=1}^m$, each of these two has merely the same size of the sensing matrix A and can be distributed to parallel computing units as A, e.g., a_i, w_i, u_i are stored in the i-th unit, for $i = 1, 2, \cdots, m$. In certain applications, the matrix A can be large and sparse, and the variables w_i and u_i can be stored as sparse vectors as well since only the components with the same support of a_i are non-zero or updated. Moreover, the computations involve individual rows a_i of A only, not the columns of A during the operations of A^{T} in traditional optimization methods.

The idea of using consensus constraints can be readily applied to more general cases where variable splitting can help to decouple difficulties in solving the problem as a whole. In those cases, after auxiliary variables are properly introduced, the alternating minimizations of the variables become easy to solve. This technique can efficiently reduce the per-iteration computational cost and potentially improve the performance of an iterative scheme.

1.4 Related Work

In recent years, there has been a large amount of research conducted to deal with the non-smooth optimization problems, i.e., with TV regularization. In particular, many recent advances avoid smoothing the non-differentiable TV norm using ADMM [16,20] and primal-dual methods, mostly under the assumption that $F(x)$ is relatively simple. After introducing auxiliary variable to substitute the Kx in the non-smooth J function in (1.1), ADMM is applied to alternately minimize the objective with respect to the variable x and auxiliary variable, so that the subproblems can be solved easily and overall convergence is fast. See e.g., [6,17,21,22,41,43] and references therein.

More recently, several algorithms that utilize the primal-dual formulation of TV norm have been developed to solve (1.1) where F is a simple function. For example, to solve the image denoising problem modeled by $F(x) = \frac{1}{2}\|x - b\|^2$, where b is an input noisy image, Chambolle [7] provided a semi-implicit gradient descent algorithm using the idea of Lagrange multipliers to solve the dual problem. In [44], an efficient primal-dual hybrid gradient (PDHG) algorithm was proposed to solve (1.1) and applied to image denoising and deblurring where the proximity operator of F can be easily computed using fast Fourier transforms. Different from the method in [7] that only solves the dual problem, PDHG alternately updates the primal and dual variables in each iteration with an adaptive proximal step, which is very efficient for TV-based image reconstruction problems. Moreover, a slightly modified version of the PDHG algorithm is developed [18] and showed equivalent to the split inexact Uzawa method [43]. They also proved a convergence result for PDHG applied to TV denoising with some restrictions on the PDHG step-size parameters. More comprehensive study on the convergence of PDHG is presented in [4,8,23]. In particular, Chambolle et al. [8] established a convergence rate of $\mathcal{O}(L_K/t)$ when the domains of primal and dual variables are both bounded, where t is iteration number and L_K is the Lipschitz constant of the operator K. We remark here that the primal-dual methods are also closely related to the Douglas–Rachford splitting method [15,27] and a pre-conditioned version of ADMM, see e.g., [8,17,23,30] for detailed reviews on the relationship between the primal-dual methods and those algorithms.

However, it is worth noting that all the methods discussed above assume a very special structure of fidelity term F (e.g., simple) to achieve high efficiency. For general convex and continuously differentiable F function, linearization of F can be adopted so that the subproblem is easy to solve. For example, the Bregman operator splitting (BOS) method [43] replaces $F(x)$ by $F(x^t) + \langle \nabla F(x^t), x - x^t \rangle$ and computes approximate solution in each iteration. This method is later shown to be equivalent to inexact Uzawa method proposed in [1]. The BOS algorithm adopts a restrictive step-size policy for convergence, and hence performs less efficiently when compared to variable step sizes based on Barzilai-Borwein method [3] and backtracking as developed in [11,42]. The linearization idea can also be adopted in the primal-dual framework to overcome the non-simplicity of $F(x)$ term (1.1). In particular, Chen et al. [12] developed an accelerated scheme using the Nesterov's idea [34–36] to reach $\mathcal{O}(\frac{L_K}{t} + \frac{L_F}{t^2})$, so that the convergence rate does not suffer too much of the Lipschitz constant L_F due to linearization. It is expected that this rate is optimal based on the observations that the convergence rate of first-order method for smooth objective function $F(x)$ solely is at most $\mathcal{O}(L_F/t^2)$ [37], whereas that of solving $\min_{x\in X}\max_{y\in Y}\langle Kx, y\rangle$ is at most $\mathcal{O}(L_K/t)$ for bounded Y and linear bounded operator K [33]. Nevertheless, these approaches are categorized as first-order methods as they apply the aforementioned approximation of F and require gradient of the F term during the iterations. However, in many applications of signal/image processing and machine learning, the F term is usually structured and can be decomposed to simple functions. The proposed method utilizes this feature to tackle the large F term without approximation of F or computation of its gradient, and the resulting scheme can be well suited for distributed computing.

On the other hand, besides satisfactory performance in solving TV-based image reconstructions, ADMM actually works well empirically for a variety of convex optimization problems that involve equality constraints. Therefore, it is shown to be well suited to distributed and consensus optimization arising in signal and image processing, statistics, and machine learning [5]. In addition, during the past few years there are growing interests in solving consensus problem over network [2,5,9,32]. The goal is to perform consensus averaging or optimization on the network such that all nodes reach the same value upon convergence. The computation is usually required to be decentralized since a fusion center may not be available for the nodes to communicate freely [10,25,26,39]. However, non-smooth image reconstruction problem has not been considered in the literature in this field. As we can see in the previous subsection, in particular the minimizations (1.13)–(1.16) arising in formulation of many real-world applications, the consensus constraints can overcome the issue of dealing with the large-scale complex data fidelity term as a whole, and the original problem can break down to multiple subproblems involving relatively simple minimizations that can be solved easily. More importantly, the resulting algorithm COMMON solves the $F(x)$ term exactly through the consensus approach and avoids linearization of $F(x)$ completely. Therefore, it also enjoys $\mathcal{O}(L_K/t)$ convergence rate. Moreover, COMMON can readily handle non-smooth functions $F(x)$ which appear frequently in ℓ_1 and sparsity-based optimization problems such as (1.14), without approximation of $F(x)$ by smooth functions as in other gradient-based optimization approaches.

2 Convergence Analysis

In this section, we establish the convergence of COMMON and provide an estimate of its convergence rate. To start with, we write the optimization problem (1.5) into the following general form,

$$\min_{x,w} \max_{y,u} \left\{ L(x, w; y, u) + \frac{\delta}{2}\|Bx - w\|^2 \right\}, \tag{2.1}$$

where (x, w) is the pair of primal variables and $x, w_i \in \mathbb{R}^n$ for each $i = 1, \cdots, m$, and (y, u) is the pair of dual variables and $y \in \mathbb{R}^{2n}$ and $u_i \in \mathbb{R}^n$ for each i. The Lagrangian $L(x, w; y, u)$ is defined by

$$L(x, w; y, u) = F(w) + H(x) - J^*(y) + \langle Kx, y\rangle - \langle u, Bx - w\rangle, \tag{2.2}$$

where $H(x)$ is a proper, convex, and l.s.c. function, and B is a bounded linear operator with induced operator norm L_B. Note that if we set $w = (w_1^{\mathrm{T}}, \cdots, w_m^{\mathrm{T}})^{\mathrm{T}}$, $u = (u_1^{\mathrm{T}}, \cdots, u_m^{\mathrm{T}})^{\mathrm{T}}$, $B = (I, \cdots, I)^{\mathrm{T}}$, and $H(x)$ to the indicator function of set X such that $H(x) = 0$ if $x \in X$ and ∞ otherwise, then we return to the consensus minimax problem (1.5). For decentralized version (1.7), one can modify matrix B such that the constraints $w_i = x_{ij}$ for all $i \in V$ and $(i, j) \in E$ can be represented by $Bx = Cw$ where x is composed of all x_{ij} and w of all w_i. In this case, the multiplier u contains all u_{ij} (note here u_{ij} and u_{ji} may be different), each having the same size as x_{ij} and is corresponding to constraint $x_{ij} - w_i = 0$. Then we can apply similar alternating minimization scheme below where every node i will only need to exchange w_i and u_{ij} with its neighbors j and hence the computation is decentralized.

As can be seen, the steps (1.8) to (1.12) in Algorithm 1 implement the following scheme (2.3)–(2.7) to solve the minimax problem (1.5).

$$x^{t+1} = \arg\min_{x \in X} \left\{ H(x) + \langle Kx, \bar{y}^t \rangle + \frac{\delta}{2} \| Bx - w^t - \frac{u^t}{\delta} \|^2 + \frac{1}{2\alpha} \| x - x^t \|^2 \right\}, \tag{2.3}$$

$$w^{t+1} = \arg\min_{w \in W} \left\{ F(w) + \frac{\delta}{2} \| Bx^{t+1} - w - \frac{u^t}{\delta} \|^2 + \frac{1}{2\gamma} \| w - w^t \|^2 \right\}, \tag{2.4}$$

$$u^{t+1} = u^t - \delta (Bx^{t+1} - w^{t+1}), \tag{2.5}$$

$$y^{t+1} = \arg\min_{y \in Y} \left\{ J^*(y) - \langle Kx^{t+1}, y \rangle + \frac{1}{2\beta} \| y - y^t \|^2 \right\}, \tag{2.6}$$

$$\bar{y}^{t+1} = 2y^{t+1} - y^t, \tag{2.7}$$

where X, W, and Y are the domains of variables x, w, and y, respectively. This scheme alternately solves for the primal variables x and w, and dual variables y and u. However, different from traditional primal-dual algorithms which solve the primal variable (x, w) together thoroughly (which usually require an extensive number of inner iterations) before moving onto dual variable (y, u) and vice versa, COMMON updates each variable immediately after a new value of previous variable is computed. This approach avoids inner iterations and ensures that the per-iteration computation complexity remains low. Meanwhile, a convergence analysis needs to be established as those for traditional primal-dual methods do not apply.

To prove convergence and estimate the rate, we first introduce a useful gap function to access solution quality of (2.1). Let $z = (x, w; y, u)$ denote the primal-dual variables, and $Z := X \times W \times Y \times U$ be the domain of $(x, w; y, u)$. Then we define function $Q(\tilde{z}; z)$ for $\tilde{z}, z \in Z$ as follows,

$$Q(\tilde{z}, z) = L(\tilde{x}, \tilde{w}; y, u) - L(x, w; \tilde{y}, \tilde{u}), \tag{2.8}$$

where $L(x, w; y, u)$ is defined in (2.2). Note that $Q(\cdot, z)$ is a convex function for any fixed $z \in Z$. Also $Q(\tilde{z}, z) = -Q(z, \tilde{z})$. Moreover, $Q(\tilde{z}, z) \leqslant 0$ (or $Q(z, \tilde{z}) \geqslant 0$) for all $z \in Z$ if and only if $\tilde{z}$ is a saddle point of $L(x, w; y, u)$. Therefore, it is natural to define the gap function as follows if the feasible set Z is bounded:

$$g(\tilde{z}) = \sup_{z\in Z} Q(\tilde{z}, z). \tag{2.9}$$

In particular, it can be readily shown that $[F(\tilde{x})+J(K\tilde{x})]-[F(x^*)+J(Kx^*)] \leqslant g(\tilde{z})$ for all $\tilde{z} = (\tilde{x}, \tilde{w}; \tilde{y}, \tilde{u}) \in Z$ if x^* is optimal for the primal problem of (2.1). If the feasible set Z is unbounded, the gap function (2.9) is not well defined even if $\tilde{z}$ is close to an optimal solution. In this case, it is shown that there always exists a perturbation vector v such that

$$\tilde{g}(\tilde{z}, v) = \sup_{z\in Z} \{Q(\tilde{z}, z) - \langle v, \tilde{z} - z\rangle\} \tag{2.10}$$

is well defined [30]. As an alternate to (2.9), we will show that the proposed algorithm returns a nearly optimal solution with small gap $\tilde{g}(\tilde{z}, v)$ with small perturbation v in the Z unbounded case.

Theorem 2.1 *Suppose $(\hat{x}, \hat{w}; \hat{y}, \hat{u})$ is a saddle point of $L(x, w; y, u)$ defined in (2.2), and the parameters satisfy $\alpha, \beta, \gamma, \delta > 0$ and $\alpha\beta L_K^2 < 1$. Then the sequence $\{(x^t, w^t; y^t, u^t)\}_t$ generated by COMMON satisfies the following conditions:*

1. *For any t, the distance from iterate $(x^t, w^t; y^t, u^t)$ to $(\hat{x}, \hat{w}; \hat{y}, \hat{u})$ is bounded:*

$$\frac{1}{2\alpha}\|\hat{x} - x^t\|^2 + \frac{1}{2\gamma}\|\hat{w} - w^t\|^2 + \frac{1}{2\beta}\|\hat{y} - y^t\|^2 + \frac{1}{2\delta}\|\hat{u} - u^t\|^2 \leqslant CD^2(\hat{z}, z^0), \tag{2.11}$$

where the constant $C \leqslant (1 - \sqrt{\alpha\beta}L_K)^{-1}$ and $D^2(\hat{z}, z^0)$ is set to

$$\frac{1}{2}\left(\frac{1}{\alpha} + \delta L_B^2\right)\|\hat{x}-x^0\|^2+\frac{1}{2\gamma}\|\hat{w}-w^0\|^2+\frac{1}{2\beta}\|\hat{y}-y^0\|^2+\frac{1}{2\delta}\|\hat{u}-u^0\|^2. \tag{2.12}$$

2. *There exists a saddle point $(x^*, w^*; y^*, u^*)$ of $L(x, w; y, u)$, such that the entire sequence $(x^t, w^t; y^t, u^t)$ converges to $(x^*, w^*; y^*, u^*)$ as $t \to \infty$.*

Proof We first observe that the optimality conditions of the minimization problems (2.3), (2.4), and (2.6) imply the inequalities for all feasible $x \in X$, $w \in W$, $y \in Y$ as follows, due to the convexity of functions H, F, and J^*:

$$\begin{aligned} H(x^{t+1}) - H(x) \leqslant \Big\langle & K^{\mathrm{T}}\bar{y}^t + \delta B^{\mathrm{T}}((Bx^{t+1} - w^t) - u^t) \\ & + \frac{x^{t+1} - x^t}{\alpha}, x - x^{t+1}\Big\rangle, \end{aligned} \tag{2.13}$$

$$F(w^{t+1}) - F(w) \leqslant \left\langle -\delta(Bx^{t+1} - w^{t+1}) + u^t + \frac{w^{t+1} - w^t}{\gamma}, w - w^{t+1}\right\rangle, \tag{2.14}$$

$$J^*(y^{t+1}) - J^*(y) \leqslant \left\langle -Kx^{t+1} + \frac{y^{t+1} - y^t}{\beta}, y - y^{t+1}\right\rangle. \tag{2.15}$$

Meanwhile, due to the definition of $Q(\tilde{z}, z)$ in (2.8), we have

$$\begin{aligned} Q(z^{t+1}; z) &= L(x^{t+1}, w^{t+1}; y, u) - L(x, w; y^{t+1}, u^{t+1}) \\ &= \Big[F(w^{t+1}) + H(x^{t+1}) - J^*(y) + \langle Kx^{t+1}, y\rangle \\ &\quad - \langle u, Bx^{t+1} - w^{t+1}\rangle\Big] - \Big[\Big(F(w) + H(x) \\ &\quad - J^*(y^{t+1}) + \langle Kx, y^{t+1}\rangle - \langle u^{t+1}, Bx - w\rangle\Big)\Big]. \end{aligned} \tag{2.16}$$

Applying the inequalities (2.13), (2.14), and (2.15) to (2.16), we obtain that

$$\begin{aligned} Q(z^{t+1}; z) \leqslant\ & \langle K^{\mathrm{T}}\bar{y}^t + \delta B^{\mathrm{T}}((Bx^{t+1} - w^t) - u^t), x - x^{t+1}\rangle \\ & - \langle \delta(Bx^{t+1} - w^{t+1}) - u^t, w - w^{t+1}\rangle + \langle -Kx^{t+1}, y - y^{t+1}\rangle \\ & + \langle Kx^{t+1}, y\rangle - \langle u, Bx^{t+1} - w^{t+1}\rangle - \langle Kx, y^{t+1}\rangle + \langle u^{t+1}, Bx - w\rangle \\ & + h_\alpha(x, x^t, x^{t+1}) + h_\gamma(w, w^t, w^{t+1}) + h_\beta(y, y^t, y^{t+1}), \end{aligned} \tag{2.17}$$

where for notation simplicity, we introduced an h function defined by

$$\begin{aligned} h_\alpha(x, x^t, x^{t+1}) &:= \frac{1}{\alpha}\langle x^{t+1} - x^t, x - x^{t+1}\rangle \\ &= \frac{1}{2\alpha}\left(\|x - x^t\|^2 - \|x - x^{t+1}\|^2 - \|x^t - x^{t+1}\|^2\right), \end{aligned} \tag{2.18}$$

and $h_\gamma(w, w^t, w^{t+1})$ and $h_\beta(y, y^t, y^{t+1})$ are defined in a similar manner.

We collect the terms on inner product terms involving y, y^t, $\bar{y}^t$, and y^{t+1} on the right side of (2.17), and be aware of (2.7), to get

$$\begin{aligned} & \langle K^{\mathrm{T}}\bar{y}^t, x - x^{t+1}\rangle - \langle Kx^{t+1}, y - y^{t+1}\rangle + \langle Kx^{t+1}, y\rangle - \langle Kx, y^{t+1}\rangle \\ &= \langle \bar{y}^t, K(x - x^{t+1})\rangle - \langle y^{t+1}, K(x - x^{t+1})\rangle \\ &= \langle y^t - y^{t+1}, K(x - x^{t+1})\rangle + \langle y^t - y^{t-1}, K(x - x^{t+1})\rangle \\ &= -\langle y^{t+1} - y^t, K(x - x^{t+1})\rangle + \langle y^t - y^{t-1}, K(x - x^t)\rangle \\ &\quad -\langle y^t - y^{t-1}, K(x^{t+1} - x^t)\rangle. \end{aligned} \tag{2.19}$$

Then we collect the other inner product terms on the right side of (2.17) and get the following:

$$\begin{aligned} & \langle \delta(Bx^{t+1} - w^t) - u^t, B(x - x^{t+1})\rangle - \langle \delta(Bx^{t+1} - w^{t+1}) - u^t, w - w^{t+1}\rangle \\ & - \langle u, Bx^{t+1} - w^{t+1}\rangle + \langle u^{t+1}, Bx - w\rangle, \end{aligned} \tag{2.20}$$

which can be rewritten as follows due to the relation in (2.5),

$$\begin{aligned}
&\langle -u^{t+1}+\delta(w^{t+1}-w^t), B(x-x^{t+1})\rangle + \langle u^{t+1}, w-w^{t+1}\rangle \\
&-\langle u, Bx^{t+1}-w^{t+1}\rangle + \langle u^{t+1}, Bx-w\rangle \\
=\;&\langle u^{t+1}-u, Bx^{t+1}-w^{t+1}\rangle + \delta\langle w^{t+1}-w^t, B(x-x^{t+1})\rangle \\
=\;&\frac{1}{\delta}\langle u-u^{t+1}, u^{t+1}-u^t\rangle + \delta\langle B(x^{t+1}-x^t), B(x-x^{t+1})\rangle \\
&+\langle (u^{t+1}-u^t)-(u^t-u^{t-1}), B(x-x^{t+1})\rangle \\
=\;&h_\delta(u,u^t,u^{t+1}) + h_{\delta^{-1}}(Bx,Bx^t,Bx^{t+1}) + \langle u^{t+1}-u^t, B(x-x^{t+1})\rangle \\
&-\langle u^t-u^{t-1}, B(x-x^t)\rangle + \langle u^t-u^{t-1}, B(x^{t+1}-x^t)\rangle. \qquad (2.21)
\end{aligned}$$

Now we substitute (2.19) and (2.21) back into the estimate (2.17) and obtain the following inequality:

$$\begin{aligned}
Q(z^{t+1};z) \leqslant\;& h_\alpha(x,x^t,x^{t+1}) + h_\gamma(w,w^t,w^{t+1}) + h_\beta(y,y^t,y^{t+1}) \\
&+ h_\delta(u,u^t,u^{t+1}) + h_{\delta^{-1}}(Bx,Bx^t,Bx^{t+1}) \\
&+ \langle u^{t+1}-u^t, B(x-x^{t+1})\rangle - \langle u^t-u^{t-1}, B(x-x^t)\rangle \\
&- \langle y^{t+1}-y^t, K(x-x^{t+1})\rangle + \langle y^t-y^{t-1}, K(x-x^t)\rangle \\
&- \langle y^t-y^{t-1}, K(x^{t+1}-x^t)\rangle + \langle u^t-u^{t-1}, B(x^{t+1}-x^t)\rangle. \qquad (2.22)
\end{aligned}$$

Note that due to Cauchy-Schwartz inequality and Young's inequality, there is

$$\begin{aligned}
|\langle y^t-y^{t-1}, K(x^{t+1}-x^t)\rangle| &\leqslant L_K\|y^t-y^{t-1}\|\|x^{t+1}-x^t\| \\
&\leqslant \frac{\sqrt{\alpha\beta}L_K}{2\beta}\|y^t-y^{t-1}\|^2 + \frac{\sqrt{\alpha\beta}L_K}{2\alpha}\|x^{t+1}-x^t\|^2, \quad (2.23)
\end{aligned}$$

and similarly that

$$|\langle u^t-u^{t-1}, B(x^{t+1}-x^t)\rangle| \leqslant \frac{1}{2\delta}\|u^t-u^{t-1}\|^2 + \frac{\delta}{2}\|Bx^{t+1}-Bx^t\|^2. \qquad (2.24)$$

Substituting the two estimates above into (2.22) and using the definition of h in (2.18), we obtain

$$\begin{aligned}
Q(z^{t+1};z) \leqslant\;& \frac{1}{2\alpha}(\|x-x^t\|^2 - \|x-x^{t+1}\|^2) - \frac{1-\sqrt{\alpha\beta}L_K}{2\alpha}\|x^t-x^{t+1}\|^2 \\
&+ \frac{1}{2\gamma}(\|w-w^t\|^2 - \|w-w^{t+1}\|^2) - \frac{1}{2\gamma}\|w^t-w^{t+1}\|^2 \\
&+ \frac{1}{2\beta}(\|y-y^t\|^2 - \|y-y^{t+1}\|^2) - \frac{1}{2\beta}(\|y^t-y^{t+1}\|^2 - \sqrt{\alpha\beta}L_K\|y^t-y^{t-1}\|^2) \\
&+ \frac{1}{2\delta}(\|u-u^t\|^2 - \|u-u^{t+1}\|^2) - \frac{1}{2\delta}(\|u^t-u^{t+1}\|^2 - \|u^t-u^{t-1}\|^2) \\
&+ \frac{\delta}{2}\left(\|Bx-Bx^t\|^2 - \|Bx-Bx^{t+1}\|^2\right)
\end{aligned}$$

$$
\begin{aligned}
&+ \langle u^{t+1} - u^t, B(x - x^{t+1})\rangle - \langle u^t - u^{t-1}, B(x - x^t)\rangle \\
&- \langle y^{t+1} - y^t, K(x - x^{t+1})\rangle + \langle y^t - y^{t-1}, K(x - x^t)\rangle. \qquad (2.25)
\end{aligned}
$$

Hence, taking the sum of j from 0 to $t-1$ on both sides, we obtain

$$
\begin{aligned}
\sum_{j=0}^{t-1} Q(z^{j+1}; z) \leqslant{}& \frac{1}{2\alpha}(\|x - x^0\|^2 - \|x - x^t\|^2) - \frac{1 - \sqrt{\alpha\beta}L_K}{2\alpha} \sum_{j=0}^{t-1} \|x^j - x^{j+1}\|^2 \\
&+ \frac{1}{2\gamma}(\|w - w^0\|^2 - \|w - w^t\|^2) - \frac{1}{2\gamma} \sum_{j=0}^{t-1} \|w^j - w^{j+1}\|^2 \\
&+ \frac{1}{2\beta}(\|y - y^0\|^2 - \|y - y^t\|^2) - \frac{1}{2\beta}\|y^{t-1} - y^t\|^2 \\
&- \frac{1 - \sqrt{\alpha\beta}L_K}{2\beta} \sum_{j=0}^{t-1} \|y^j - y^{j-1}\|^2 \\
&+ \frac{1}{2\delta}(\|u - u^0\|^2 - \|u - u^t\|^2) - \frac{1}{2\delta}\|u^t - u^{t-1}\|^2 \\
&+ \frac{\delta}{2}(\|Bx - Bx^0\|^2 - \|Bx - Bx^t\|^2) \\
&+ \langle u^t - u^{t-1}, B(x - x^t)\rangle - \langle y^t - y^{t-1}, K(x - x^t)\rangle, \qquad (2.26)
\end{aligned}
$$

where we use convention that $y^{-1} = y^0$ and $u^{-1} = u^0$. We further note that

$$
|\langle y^t - y^{t-1}, K(x - x^t)\rangle| \leqslant \frac{\sqrt{\alpha\beta}L_K}{2\beta}\|y^t - y^{t-1}\|^2 + \frac{\sqrt{\alpha\beta}L_K}{2\alpha}\|x - x^t\|^2 \qquad (2.27)
$$

and that

$$
|\langle u^t - u^{t-1}, B(x - x^t)\rangle| \leqslant \frac{1}{2\delta}\|u^t - u^{t-1}\|^2 + \frac{\delta}{2}\|Bx - Bx^t\|^2. \qquad (2.28)
$$

Therefore, submitting the above two inequalities into (2.26), we obtain

$$
\begin{aligned}
\sum_{j=0}^{t-1} Q(z^{j+1}; z) \leqslant{}& \frac{1}{2\alpha}\|x - x^0\|^2 - \frac{1 - \sqrt{\alpha\beta}L_K}{2\alpha}\|x - x^t\|^2 \\
&- \sum_{j=0}^{t-1} \frac{1 - \sqrt{\alpha\beta}L_K}{2\alpha}\|x^j - x^{j+1}\|^2 \\
&+ \frac{1}{2\gamma}(\|w - w^0\|^2 - \|w - w^t\|^2) - \frac{1}{2\gamma} \sum_{j=0}^{t-1} \|w^j - w^{j+1}\|^2
\end{aligned}
$$

$$+\frac{1}{2\beta}(\|y-y^0\|^2-\|y-y^t\|^2)-\frac{1-\sqrt{\alpha\beta}L_K}{2\beta}\sum_{j=0}^{t}\|y^j-y^{j+1}\|^2$$
$$+\frac{1}{2\delta}(\|u-u^0\|^2-\|u-u^t\|^2)+\frac{\delta}{2}\|Bx-Bx^0\|^2. \tag{2.29}$$

Substituting a saddle point $\hat{z}=(\tilde{x},\tilde{w};\tilde{y},\tilde{u})$ of Lagrangian L above, which yields $Q(z^{j+1},\hat{z})\geqslant 0, \forall j$, we obtain

$$0\leqslant\frac{1}{2\alpha}\|\hat{x}-x^0\|^2-\frac{1-\sqrt{\alpha\beta}L_K}{2\alpha}\|\hat{x}-x^t\|^2-\sum_{j=0}^{t-1}\frac{1-\sqrt{\alpha\beta}L_K}{2\alpha}\|x^j-x^{j+1}\|^2$$
$$+\frac{1}{2\gamma}(\|\hat{w}-w^0\|^2-\|\hat{w}-w^t\|^2)-\frac{1}{2\gamma}\sum_{j=0}^{t-1}\|w^j-w^{j+1}\|^2$$
$$+\frac{1}{2\beta}(\|\hat{y}-y^0\|^2-\|\hat{y}-y^t\|^2)-\frac{1-\sqrt{\alpha\beta}L_K}{2\beta}\sum_{j=0}^{t}\|y^j-y^{j+1}\|^2$$
$$+\frac{1}{2\delta}(\|\hat{u}-u^0\|^2-\|\hat{u}-u^t\|^2)+\frac{\delta}{2}\|B\hat{x}-Bx^0\|^2, \tag{2.30}$$

since $\alpha\beta L_K^2<1$. Hence the conclusion (2.11) follows.

Now we are left to prove that the entire sequence (x^t,w^t,y^t,u^t) converges to some saddle point. First of all, the boundedness of (x^t,w^t,y^t,u^t) in (2.11) implies the existence of subsequence $(x^{t_k},w^{t_k},y^{t_k},u^{t_k})$ that converges to a limit point (x^*,w^*,y^*,u^*). Furthermore, the estimate in (2.30) implies boundedness of series $\sum_t\|x^{t+1}-x^t\|^2$, $\sum_t\|w^{t+1}-w^t\|^2$, and $\sum_t\|y^{t+1}-y^t\|^2$, from which we conclude that $\lim_t(x^{t+1}-x^t)=0$, $\lim_t(w^{t+1}-w^t)=0$, and $\lim_t(y^{t+1}-y^t)=0$. Hence $\lim_t(Bx^{t+1}-Bx^t)=0$ and $\lim_t(u^{t+1}-u^t)=0$. Substituting (x^t,w^t,y^t,u^t) by $(x^{t_k},w^{t_k},y^{t_k},u^{t_k})$ in (2.3)–(2.6), and taking limit $k\to\infty$, we can see that (x^*,w^*,y^*,u^*) is a saddle point of the minimax problem (2.1).

Now we substitute $z=(x,w;y,u)$ by this saddle point $z^*=(x^*,w^*,y^*,u^*)$ in (2.25), and take the sum of j from t_k to $t-1$ for $t>t_k$ to get

$$0\leqslant\frac{1}{2\alpha}(\|x^*-x^{t_k}\|^2-\|x^*-x^t\|^2)-\frac{1-\sqrt{\alpha\beta}L_K}{2\alpha}\sum_{j=t_k}^{t-1}\|x^j-x^{j+1}\|^2$$
$$+\frac{1}{2\gamma}(\|w^*-w^{t_k}\|^2-\|w^*-w^t\|^2)-\frac{1}{2\gamma}\sum_{j=t_k}^{t-1}\|w^j-w^{j+1}\|^2$$
$$+\frac{1}{2\beta}(\|y^*-y^{t_k}\|^2-\|y^*-y^t\|^2)$$
$$-\frac{1}{2\beta}\sum_{j=t_k}^{t-1}(\|y^j-y^{j+1}\|^2-\sqrt{\alpha\beta}L_K\|y^j-y^{j-1}\|^2)$$

$$
\begin{aligned}
&+\frac{1}{2\delta}(\|u^*-u^{t_k}\|^2-\|u^*-u^t\|^2)-\frac{\delta}{2}(\|Bx^*-w^{t_k}\|^2-\|Bx^*-w^t\|^2)\\
&-\langle y^t-y^{t-1},K(x^*-x^t)\rangle+\langle y^{t_k}-y^{t_k-1},K(x^*-x^{t_k})\rangle.
\end{aligned}
$$

Therefore, letting $k\to\infty$, we know $t_k\to\infty$ and hence terms such as $\|x^*-x^{t_k}\|^2$ and all summations (due to boundedness of infinite sum) above vanish. Hence we conclude that $(x^t,w^t,y^t,u^t)\to(x^*,w^*,y^*,u^*)$ as $t\to\infty$.

In the case that $Z=X\times W\times Y\times U$ is bounded, we can use the gap function (2.9) to access solution quality and derive convergence rate. More precisely, suppose that there are bounds D_X^2, D_W^2, D_Y^2, and D_U^2 that satisfy

$$
\begin{aligned}
&\sup_{x_1,x_2\in X}\|x_1-x_2\|^2\leqslant D_X^2, &&\sup_{w_1,w_2\in W}\|w_1-w_2\|^2\leqslant D_W^2,\\
&\sup_{y_1,y_2\in Y}\|y_1-y_2\|^2\leqslant D_Y^2, &&\sup_{u_1,u_2\in U}\|u_1-u_2\|^2\leqslant D_U^2.
\end{aligned}\tag{2.31}
$$

Then we have the following result which indicates $\mathcal{O}(1/t)$ convergence rate in terms of gap function of averaged iterate $\tilde{z}_1^t$.

Theorem 2.2 *Let $z^t=(x^t,w^t,y^t,u^t)$ be the sequence generated by Algorithm 1, and that $\tilde{z}_1^t=(\sum_{j=1}^t z^j)/t$, then*

$$
g(\tilde{z}_1^t)\leqslant\frac{1}{2t}\left((\alpha^{-1}+\delta L_B^2)D_X^2+\gamma^{-1}D_W^2+\beta^{-1}D_Y^2+\delta^{-1}D_U^2\right).\tag{2.32}
$$

Proof Since $Q(\tilde{z},z)$ is a convex function with respect to $\tilde{z}$ for every fixed z, we know that

$$
tQ(\tilde{z}_1^t,z)\leqslant\sum_{j=0}^{t-1}Q(z^{j+1},z).\tag{2.33}
$$

Then substituting z by z^* in (2.29), we obtain that

$$
\begin{aligned}
tQ(\tilde{z}_1^t,z^*)\leqslant{}&\frac{1}{2\alpha}\|x^*-x^0\|^2+\frac{1}{2\gamma}\|w^*-w^0\|^2+\frac{1}{2\beta}\|y^*-y^0\|^2\\
&+\frac{1}{2\delta}\|u^*-u^0\|^2+\frac{\delta}{2}\|Bx^*-Bx^0\|^2.
\end{aligned}\tag{2.34}
$$

By the definition of domain bounds in (2.31), we conclude with (2.32).

If one of the domains X, W, Y, and U is unbounded, then Z is unbounded. In this case, we use a perturbed gap function (2.10) to estimate the rate of convergence of $\tilde{z}_1^t$ to an optimal solution as $t\to\infty$. First, we derive the following estimate:

Theorem 2.3 *Let $\hat{z}=(\hat{x},\hat{w};\hat{y},\hat{u})$ be a saddle point of* (2.1)*, then*

$$
\tilde{g}(\tilde{z}_1^t,v^t)\leqslant\frac{D^2(\tilde{z}_1^t,z^0)}{t},\tag{2.35}
$$

where the perturbation vector v^t *satisfies*

$$\|v^t\| \leqslant \frac{\sqrt{2}}{t}\left[\sqrt{\alpha^{-1}+\delta L_B^2}+\left(\sqrt{\alpha}(\alpha^{-1}+\delta L_B^2)+2\sqrt{\delta}L_B+2\sqrt{\beta}L_K\right)\sqrt{C}\right.$$
$$\left.+\left(\sqrt{\beta^{-1}}+\sqrt{\gamma^{-1}}+\sqrt{\delta^{-1}}\right)(1+\sqrt{C})\right]D(\hat{z},z^0), \tag{2.36}$$

the distance $D(\cdot,\cdot)$ *is defined as in* (2.12)*, and* $C \leqslant (1-\sqrt{\alpha\beta}L_K)^{-1}$ *is a constant.*

Proof We first note the following identity:

$$\begin{aligned}
&\|x-x^0\|^2-\|x-x^t\|^2\\
&=\|(x-\tilde{x}_1^t)-(x^0-\tilde{x}_1^t)\|^2-\|(x-\tilde{x}_1^t)-(x^t-\tilde{x}_1^t)\|^2\\
&=\|x^0-\tilde{x}_1^t\|^2-\|x^t-\tilde{x}_1^t\|^2+2\langle x^0-x^t,\tilde{x}_1^t-x\rangle,
\end{aligned} \tag{2.37}$$

where $\tilde{x}_1^t=(\sum_{j=1}^t x^j)/t$. Similar definition and identity hold for $\tilde{y}_1^t$, $\tilde{w}_1^t$, and $\tilde{u}_1^t$. Due to convexity of $Q(\cdot,z)$, we deduce from (2.26) that

$$\begin{aligned}
tQ(\tilde{z}_1^t,z) \leqslant\ & \frac{1}{2\alpha}(\|\tilde{x}_1^t-x^0\|^2-\|\tilde{x}_1^t-x^t\|^2)+\frac{1}{\alpha}\langle x^0-x^t,\tilde{x}_1^t-x\rangle\\
&+\frac{1}{2\gamma}(\|\tilde{w}_1^t-w^0\|^2-\|\tilde{w}_1^t-w^t\|^2)+\frac{1}{\gamma}\langle w^0-w^t,\tilde{w}_1^t-w\rangle\\
&+\frac{1}{2\beta}(\|\tilde{y}_1^t-y^0\|^2-\|\tilde{y}_1^t-y^t\|^2)+\frac{1}{\beta}\langle y^0-y^t,\tilde{y}_1^t-y\rangle-\frac{1}{2\beta}\|y^{t-1}-y^t\|^2\\
&+\frac{1}{2\delta}(\|\tilde{u}_1^t-u^0\|^2-\|\tilde{u}_1^t-u^t\|^2)+\frac{1}{\delta}\langle u^0-u^t,\tilde{u}_1^t-u\rangle-\frac{1}{2\delta}\|u^t-u^{t-1}\|^2\\
&+\frac{\delta}{2}(\|B\tilde{x}_1^t-Bx^0\|^2-\|B\tilde{x}_1^t-Bx^t\|^2)+\delta\langle Bx^0-Bx^t,B\tilde{x}_1^t-Bx\rangle\\
&+\langle u^t-u^{t-1},B(x-\tilde{x}_1^t)\rangle+\langle u^t-u^{t-1},B(\tilde{x}_1^t-x^t)\rangle\\
&-\langle y^t-y^{t-1},K(x-\tilde{x}_1^t)\rangle-\langle y^t-y^{t-1},K(\tilde{x}_1^t-x^t)\rangle.
\end{aligned} \tag{2.38}$$

Meanwhile, we have that

$$|\langle u^t-u^{t-1},B(\tilde{x}_1^t-x^t)\rangle| \leqslant \frac{1}{2\delta}\|u^t-u^{t-1}\|^2+\frac{\delta}{2}\|B(\tilde{x}_1^t-x^t)\|^2 \tag{2.39}$$

and that

$$|\langle y^t-y^{t-1},K(\tilde{x}_1^t-x^t)\rangle| \leqslant \frac{\sqrt{\alpha\beta}L_K}{2\beta}\|y^t-y^{t-1}\|^2+\frac{\sqrt{\alpha\beta}L_K}{2\alpha}\|\tilde{x}_1^t-x^t\|^2. \tag{2.40}$$

Substituting the two inequalities above into (2.38), we obtain

$$\begin{aligned} tQ(\tilde{z}_1^t;\hat{z}) \leqslant{}& D(\tilde{z}_1^t,\hat{z}) - \langle(\alpha^{-1}+\delta B^{\mathrm{T}}B)(x^0-x^t),\tilde{x}_1^t-x\rangle \\ &+\langle B^{\mathrm{T}}(u^t-u^{t-1})-K^{\mathrm{T}}(y^t-y^{t-1}),\tilde{x}_1^t-x\rangle+\langle\gamma^{-1}(w^0-w^t),\tilde{w}_1^t-w\rangle \\ &+\langle\beta^{-1}(y^0-y^t),\tilde{y}_1^t-y\rangle-\langle\delta^{-1}(u^0-u^t),\tilde{u}_1^t-u\rangle, \end{aligned} \tag{2.41}$$

from which the estimate (2.35) follows with v^t defined by

$$v^t = \frac{1}{t}\begin{pmatrix} (\alpha^{-1}+\delta B^{\mathrm{T}}B)(x^0-x^t)+B^{\mathrm{T}}(u^t-u^{t-1})-K^{\mathrm{T}}(y^t-y^{t-1}) \\ \gamma^{-1}(w^0-w^t) \\ \beta^{-1}(y^0-y^t) \\ \delta^{-1}(u^0-u^t) \end{pmatrix}. \tag{2.42}$$

Now we are left to derive the estimate (2.36). By the setting (2.42), we have

$$\begin{aligned} v^t \leqslant{}& \frac{1}{t}\big[(\alpha^{-1}+\delta B^{\mathrm{T}}B)\|x^0-x^t\|+L_B\|u^t-u^{t-1}\|+L_K\|y^t-y^{t-1}\| \\ &+\gamma^{-1}\|w^0-w^t\|^2+\beta^{-1}\|y^0-y^t\|+\delta^{-1}\|u^0-u^0\|)\big]. \end{aligned} \tag{2.43}$$

Note that by the definition of (2.12) and the bound (2.11), there is

$$\begin{aligned} \|x^0-x^t\| &\leqslant \|x^0-\hat{x}\|+\|x^t-\hat{x}\| \\ &\leqslant \sqrt{2(\alpha^{-1}+\delta L_B^2)^{-1}}D(\hat{z},z^0)+\sqrt{2\alpha C}D(\hat{z},z^0) \\ &\leqslant \sqrt{2}\left(\sqrt{(\alpha^{-1}+\delta L_B^2)^{-1}}+\sqrt{\alpha C}\right)D(\hat{z},z^0). \end{aligned} \tag{2.44}$$

Similarly, there are

$$\|w^0-w^t\| \leqslant \sqrt{2\gamma}(1+\sqrt{C})D(\hat{z},z^0), \tag{2.45}$$

$$\|y^0-y^t\| \leqslant \sqrt{2\beta}(1+\sqrt{C})D(\hat{z},z^0), \tag{2.46}$$

$$\|u^0-u^t\| \leqslant \sqrt{2\delta}(1+\sqrt{C})D(\hat{z},z^0). \tag{2.47}$$

Furthermore, it can be readily shown that

$$\|u^t-u^{t-1}\| \leqslant \|u^t-\hat{u}\|+\|u^{t-1}-\hat{u}\| \leqslant 2\sqrt{2\delta C}D(\hat{z},z^0), \tag{2.48}$$

$$\|y^t-y^{t-1}\| \leqslant \|y^t-\hat{y}\|+\|y^{t-1}-\hat{y}\| \leqslant 2\sqrt{2\beta C}D(\hat{z},z^0). \tag{2.49}$$

Substituting estimates (2.44)–(2.49) into (2.43), we conclude with the bound of v^t in (2.36).

3 Numerical Tests

In this section, we present the numerical results of COMMON on the reconstruction of a Shepp-Logan phantom of various sizes n: $2^{12}, 2^{14}, 2^{16}$ (i.e., 64×64, 128×128, and 256×256, respectively). With a pre-computed attenuation matrix A (of two sizes $m = 2^{10}$ and $m = 2^{14}$), and the phantom x, we simulate three types of noises which follow Guassian, Laplacian (double exponential), and Poisson distributions, respectively. For the Gaussian and Laplacian noise, the standard deviation is set to 0.1. That is, noise $b = Ax + n$, where $n \sim N(0, 0.1^2)$ and $n \sim \text{Laplace}(0.1^{-1})$, respectively. For the Poisson case, the noisy data $b_i \sim \text{Poisson}(a_i^{\mathrm{T}} x + 1)$ independently where a_i denotes the i-th row of A. The data fidelity term F in (1.1) is then constructed using the A, b and these three noise models. Now we apply the proposed algorithm to solve (1.1) as $F(x)$ can be decomposed as (1.2) with each F_i defined according to these noise models. For the three noise modes, we choose the weighting parameter of TV term as 10^{-3}, 10, 10^{-1} by empirical experiments, for which the reconstructed image has nearly optimal quality with satisfactory noise-to-ratio level. The distributed computation is simulated in MATLAB R2013b (v8.2) and performed on a desktop computer with Intel Quad-Core 3.7 GHz Processor and 32 GB of memory.

To test the efficiency of dealing with data term with consensus optimization, we compare COMMON with two recent numerical algorithms: Bregman operator splitting (BOS) [43] and BOS with variable step (BOSVS) sizes [11]. BOS and BOSVS are designed to solve non-smooth image reconstruction problem with TV regularization and general data fidelity term. The difference between BOS and classical ADMM (or well known as the split Bregman method in imaging community) is that the data term $F(x)$ is approximated by $F(x^t) + \langle \nabla F(x^t), x - x^t \rangle$ with an additional proximity term $\frac{1}{2\alpha}\|x - x^t\|^2$ in BOS, where $\alpha \leqslant 1/\|\nabla F(x)\|_\infty$ is a fixed step size for guaranteed convergence. The BOSVS algorithm relaxes this restrictive step size bound and uses inexact line search to find optimal step size in each iteration, still with guaranteed convergence. The performance of BOSVS is shown to be much better than BOS in practice. Note that BOS and BOSVS both require that the data fidelity term $F(x)$ is differentiable. Therefore, in the case of Laplacian noise where $F(x) = \|Ax - b\|_1$, we use smooth approximation $F_\epsilon(x) := \sum_{i=1}^m (|a_i^{\mathrm{T}} x - b_i|^2 + \epsilon)^{1/2}$ with $\epsilon = 10^{-6}$ in BOS and BOSVS. For all comparison algorithms, we simply set the termination criterion to $\|x^t - x^{t-1}\|/\|x^t\| < 10^{-5}$ in experiments. We also point out here that the termination criterion of COMMON can be set more sophisticatedly using the perturbed gap function as in Theorem 2.3.

The numerical results of the reconstructions by the three comparison algorithms are given in Tables 1, 2, and 3 below. In each table, the following outputs of the three comparison algorithms are given: the objective value (Obj) of (1.1), total iteration (Itr) number until termination criterion is met, and the relative error (Err) of final output x^t to the original image x, i.e., $\|x^t - x\|/\|x\|$. For each of the two sampling sizes m, we test reconstruction of three different image sizes n. Table 1 shows the performance of BOS, BOSVS, and COMMON on the reconstruction of x given b corrupted by Gaussian noise. As we can see, COMMON consistently returns image of good quality (<3% relative error) with the least number of iterations, while

Table 1 Numerical result on image reconstruction where data is corrupted by Gaussian noise: the iteration number (Itr), the objective function (Obj) values (1.1), and relative errors (Err) to original image of the outputs by comparison algorithms BOS, BOSVS, and COMMON

m	n	BOS			BOSVS			COMMON		
		Obj	Itr	Err	Obj	Itr	Err	Obj	Itr	Err
2^{10}	2^{12}	3.57	324	0.038	3.41	67	0.031	3.37	41	0.029
2^{14}	2^{12}	3.64	167	0.035	3.47	35	0.030	3.46	23	0.027
2^{10}	2^{14}	3.99	647	0.037	3.79	122	0.031	3.88	57	0.023
2^{14}	2^{14}	4.14	402	0.032	3.98	65	0.028	3.94	43	0.022
2^{10}	2^{16}	4.33	805	0.030	4.32	211	0.024	4.32	65	0.023
2^{14}	2^{16}	4.81	648	0.029	4.78	97	0.021	4.79	52	0.020

Table 2 Numerical result on image reconstruction where data is corrupted by Laplacian noise: the iteration number (Itr), the objective function (Obj) values (1.1), and relative errors (Err) to original image of the outputs by comparison algorithms BOS, BOSVS, and COMMON

m	n	BOS			BOSVS			COMMON		
		Obj	Itr	Err	Obj	Itr	Err	Obj	Itr	Err
2^{10}	2^{12}	64.64	1 790	0.033	64.00	274	0.031	64.02	63	0.033
2^{14}	2^{12}	65.58	862	0.031	64.28	145	0.030	64.31	47	0.032
2^{10}	2^{14}	64.70	1 917	0.034	64.26	317	0.031	64.22	51	0.034
2^{14}	2^{14}	66.57	915	0.032	64.77	175	0.027	64.69	41	0.032
2^{10}	2^{16}	67.72	2 134	0.030	64.80	346	0.029	64.81	96	0.029
2^{14}	2^{16}	68.52	1 231	0.029	64.82	185	0.026	64.82	63	0.028

Table 3 Numerical result on image reconstruction where data is corrupted by Poisson noise: the iteration number (Itr), the objective function (Obj) values (1.1) and relative errors (Err) to original image of the outputs by comparison algorithms BOS, BOSVS, and COMMON

m	n	BOS			BOSVS			COMMON		
		Obj	Itr	Err	Obj	Itr	Err	Obj	Itr	Err
2^{10}	2^{12}	−792.4	763	0.052	−794.8	119	0.042	−792.9	64	0.043
2^{14}	2^{12}	−793.9	503	0.050	−796.5	78	0.043	−794.1	41	0.041
2^{10}	2^{14}	−793.7	1 005	0.052	−794.1	187	0.048	−795.6	102	0.043
2^{14}	2^{14}	−793.9	796	0.049	−796.2	74	0.040	−796.2	69	0.042
2^{10}	2^{16}	−795.2	1 257	0.058	−796.4	146	0.042	−796.3	131	0.049
2^{14}	2^{16}	−795.9	730	0.047	−799.1	83	0.039	−798.9	63	0.039

BOSVS takes a bit more iterations to get similar quality, and BOS is much less efficient due to its restrictive step size policy. Similar results appear in the case of Laplace noise and Poisson noise in Tables 2 and 3. In particular, COMMON is much more efficient in comparison to BOS and BOSVS in the Laplace noise case: COMMON tackles the non-differentiability of ℓ_1 norm in $F(x)$ adequately by consensus optimization, such that the ℓ_1 minimization of type (1.14) can be solved exactly, whereas the BOS and BOSVS require smoothing of the singularities of ℓ_1 norm and become less efficient.

It is also worth noting that besides efficiency in convergence speed, COMMON can be readily adopted for distributed computing which can significantly reduced computational time in contrast to traditional methods. Moreover, COMMON can be implemented for decentralized computation when a central cluster core or a shared memory is unavailable in specific applications.

4 Concluding Remarks

We proposed and analyzed an efficient primal-dual algorithm for consensus minimax optimization, called COMMON, to solve a class of non-smooth image reconstruction problems. The algorithm is inspired by the observation that the data fidelity term $F(x)$ can often be expressed as sum of relatively simple functions due to physical modeling of data acquisition in a large number of real-world applications. Therefore, the consensus constraints are introduced such that the computation can be easily deployed for parallel computing and solved efficiently. COMMON iteratively solves the subproblems of primal variables and dual variables, such that the per-iteration complexity is extremely low. Convergence analysis shows that COMMON has guaranteed convergence, and the rate is $\mathcal{O}(1/t)$ in terms of the (perturbed) gap function where t is iteration number.

References

[1] Arrow, K.J., Hurwicz, L., Uzawa, H.: Studies in linear and non-linear programming. In: Chenery, H.B., Johnson, S.M., Karlin, S., Marschak, T., Solow, R.M. (eds.) Stanford Mathematical Studies in the Social Sciences, vol. II. Stanford University Press, Stanford (1958)

[2] Aysal, T.C., Yildiz, M.E., Sarwate, A.D., Scaglione, A.: Broadcast gossip algorithms for consensus. IEEE. Trans. Signal Process. **57**(7), 2748–2761 (2009)

[3] Barzilai, J., Borwein, J.M.: Two-point step size gradient methods. IMA J. Numer. Anal. **8**(1), 141–148 (1988)

[4] Bonettini, S., Ruggiero, V.: On the convergence of primal-dual hybrid gradient algorithms for total variation image restoration. J. Math. Imaging Vision **44**(3), 236–253 (2012)

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

[6] Cai, J.F., Osher, S., Shen, Z.: Split Bregman methods and frame based image restoration. Multiscale Model. Simul. **8**(2), 337–369 (2009/10). doi:10.1137/090753504

[7] Chambolle, A.: An algorithm for total variation minimization and applications. J. Math. Imaging Vision **20**(1–2), 89–97 (2004)

[8] Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vision **40**(1), 120–145 (2011)

[9] Chang, T.H., Hong, M., Wang, X.: Multi-agent distributed optimization via inexact consensus ADMM. arXiv preprint arXiv:1402.6065 (2014)

[10] Chen, I.A., et al.: Fast distributed first-order methods. Ph.D. thesis, Massachusetts Institute of Technology (2012)

[11] Chen, Y., Hager, W.W., Yashtini, M., Ye, X., Zhang, H.: Bregman operator splitting with variable stepsize for total variation image reconstruction. Comput. Optim. Appl. **54**(2), 317–342 (2013). doi:10.1007/s10589-012-9519-2

[12] Chen, Y., Lan, G., Ouyang, Y.: Optimal primal-dual methods for a class of saddle point problems. arXiv preprint arXiv:1309.5548 (2013)

[13] Donoho, D.L., Johnstone, I.M.: Adapting to unknown smoothness via wavelet shrinkage. J. Am. Stat. Assoc. **90**(432), 1200–1224 (1995)

[14] Donoho, D.L., Johnstone, J.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**(3), 425–455 (1994)

[15] Douglas, J., Rachford, H.: On the numerical solution of heat conduction problems in two and three space variables. Trans. Am. Math. Soc. **82**, 421–439 (1956)

[16] Eckstein, J., Bertsekas, D.P.: On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**(1–3), 293–318 (1992)

[17] Esser, E.: Applications of lagrangian-based alternating direction methods and connections to split bregman. CAM Rep. **9**, 31 (2009)

[18] Esser, E., Zhang, X., Chan, T.F.: A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. SIAM J. Imaging Sci. **3**(4), 1015–1046 (2010)

[19] Friedman, J., Hastie, T., Tibshirani, R.: A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:1001.0736 (2010)

[20] Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**(1), 17–40 (1976)

[21] Goldstein, T., Bresson, X., Osher, S.: Geometric applications of the split bregman method: segmentation and surface reconstruction. J. Sci. Comput. **45**(1), 272–293 (2010). doi:10.1007/s10915-009-9331-z

[22] Goldstein, T., Osher, S.: The split Bregman method for l_1-regularized problems. SIAM J. Imaging Sci. **2**(2), 323–343 (2009). doi:10.1137/080725891

[23] He, B., Yuan, X.: Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. SIAM J. Imaging Sci. **5**(1), 119–149 (2012)

[24] Jacob, L., Obozinski, G., Vert, J.P.: Group lasso with overlap and graph lasso. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 433–440. ACM, New York (2009)

[25] Jakovetic, D., Xavier, J., Moura, J.: Fast Distributed Gradient Methods. http://arxiv.org/abs/1112.2972 (2011)

[26] Johansson, B., Speranzon, A., Johansson, M., Johansson, K.H.: On decentralized negotiation of optimal consensus. Automatica **44**(4), 1175–1179 (2008)

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

[28] Mallat, S.: A Wavelet Tour of Signal Processing: The Sparse Way. Academic Press, Burlington (2008)

[29] Meier, L., Van De Geer, S., Bühlmann, P.: The group lasso for logistic regression. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **70**(1), 53–71 (2008)

[30] Monteiro, R.D., Svaiter, B.F.: Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. Optimization **2713**, 1 (2010). (Online preprint)

[31] Moreau, J.J.: Fonctions convexes duales et points proximaux dans un espace hilbertien. C. R. Acad. Sci. Paris **255**, 2897–2899 (1962)

[32] Nedic, A., Ozdaglar, A., Parrilo, P.A.: Constrained consensus and optimization in multi-agent networks. IEEE Trans. Autom. Control **55**(4), 922–938 (2010)

[33] Nemirovsky, A.: Information-based complexity of linear operator equations. J. Complex. **8**(2), 153–175 (1992)

[34] Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence o (1/k2). Doklady AN SSSR **269**, 543–547 (1983)

[35] Nesterov, Y.: Nonsmooth convex optimization. In: Introductory Lectures on Convex Optimization, pp. 111–170. Springer, New York (2004)

[36] Nesterov, Y.: Primal-dual subgradient methods for convex problems. Math. Program. **120**(1, Ser. B), 221–259 (2009). doi:10.1007/s10107-007-0149-x

[37] Nesterov, Y., Nesterov, I.U.E.: Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Springer, New York (2004)

[38] Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**(1), 259–268 (1992)

[39] Shi, W., Ling, Q., Yuan, K., Wu, G., Yin, W.: On the linear convergence of the ADMM in decentralized consensus optimization. IEEE Trans. Signal Process. **62**, 4613–4617 (2013)

[40] Starck, J.L., Murtagh, F., Fadili, J.M.: Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity. Cambridge University Press, Cambridge (2010)

[41] Wu, C., Tai, X.C.: Augmented Lagrangian method, dual methods, and split Bregman iteration for ROF, vectorial TV, and high order models. SIAM J. Imaging Sci. **3**(3), 300–339 (2010). doi:10.1137/090767558

[42] Ye, X., Chen, Y., Huang, F.: Computational acceleration for mr image reconstruction in partially parallel imaging. IEEE Trans. Med. Imaging **30**(5), 1055–1063 (2011). doi:10.1109/TMI.2010.2073717

[43] Zhang, X., Burger, M., Osher, S.: A unified primal-dual algorithm framework based on Bregman iteration. J. Sci. Comput. **46**(1), 20–46 (2011). doi:10.1007/s10915-010-9408-8

[44] Zhu, M., Chan, T.: An efficient primal-dual hybrid gradient algorithm for total variation image restoration. UCLA CAM Report pp. 08–34 (2008)