Machine Learning (2020) 109:569–601
https://doi.org/10.1007/s10994-019-05847-6

Communication-efficient distributed multi-task learning with matrix sparsity regularization

Qiang Zhou[1] · Yu Chen[1] · Sinno Jialin Pan[1]

Received: 3 May 2019 / Revised: 20 July 2019 / Accepted: 16 September 2019 / Published online: 7 October 2019

Abstract
This work focuses on distributed optimization for multi-task learning with matrix sparsity regularization. We propose a fast communication-efficient distributed optimization method for solving the problem. With the proposed method, training data of different tasks can be geo-distributed over different local machines, and the tasks can be learned jointly through the matrix sparsity regularization without a need to centralize the data. We theoretically prove that our proposed method enjoys a fast convergence rate for different types of loss functions in the distributed environment. To further reduce the communication cost during the distributed optimization procedure, we propose a data screening approach to safely filter inactive features or variables. Finally, we conduct extensive experiments on both synthetic and real-world datasets to demonstrate the effectiveness of our proposed method.

Keywords Distributed learning · Multi-task learning · Acceleration

1 Introduction

Multi-task learning (MTL) (Caruana 1997) aims to jointly learn multiple machine learning tasks by exploiting their commonality to boost the generalization performance of each task. Similar to many standard machine learning techniques, in MTL, a single machine is assumed to be able to access all training data over different tasks. However, in practice, especially in the context of smart city, training data for different tasks is owned by different organizations and geo-distributed over different local machines, and centralizing the data may result in expensive cost of data transmission and cause privacy and security issues. Take personal-

Editors: Kee-Eung Kim and Jun Zhu.

✉ Sinno Jialin Pan
sinnopan@ntu.edu.sg

Qiang Zhou
zhouqiang@u.nus.edu

Yu Chen
chenyu@ntu.edu.sg

[1] Nanyang Technological University, Singapore, Singapore

ized healthcare as a motivating example. In this context, learning a personalized healthcare prediction model from each user's personal data including his/her profile and various sensor readings from his/her mobile device is considered as a different task. On one hand, the personal data may be too sparse to learn a precise prediction model for each task, and thus MTL is desired. On the other hand, some of the users may not be willing to share their personal data, which results in a failure of applying standard MTL methods. Thus, a distributed MTL algorithm is more preferred. However, if frequent communication is required for the distributed MTL algorithm to obtain an optimal prediction model for each task, users have to pay for expensive cost on data transmission, which is not practical. Therefore, designing a communication-efficient MTL algorithm in the distributed computing environment is crucial to address the aforementioned problem.

Though a number of distributed machine learning frameworks have been proposed, most of them are focused on single task learning problems (Li et al. 2014; Boyd et al. 2011; Jaggi et al. 2014; Ma et al. 2015). In particular, COCOA+ as a general distributed machine learning framework has been proposed for strongly convex learning problems (Smith et al. 2017b; Ma et al. 2015; Jaggi et al. 2014). To handle non-strongly regularizers (e.g., ℓ_1-norm), Smith et al. (2015, 2017b) extended COCOA+ by directly solving the primal problem instead of its dual problem. However, in their proposed method, data needs to be distributed by features rather than instances. In our problem setting, we suppose the training data for different tasks is originally geo-distributed over different machines. In this case, to use the method proposed in Smith et al. (2015, 2017b), one has to first centralize the data of all the tasks and then re-distribute the data w.r.t. different sets of features, which is impractical.

In this paper, different from previous methods, we focus on the MTL formulation with a $\ell_{2,1}$-norm regularization on the weight matrix over all the tasks, and offer a communication-efficient distributed optimization framework to solve it. Specifically, we have two main contributions: (1) We first present an efficient distributed optimization method that enjoys a fast convergence rate for solving the $\ell_{2,1}$-norm regularized MTL problem. To achieve this, we carefully design a subproblem for each local worker by incorporating an extrapolation step on the dual variables. We theoretically prove that with the well-designed local subproblem, our proposed method obtains a faster convergence rate than COCOA+ (Ma et al. 2015; Smith et al. 2017b), especially on ill-conditioned problems. Recently, Ma et al. (2017) also attempted to improve the convergence rate of COCOA+. However, our acceleration scheme is different from theirs. Specifically, with a strongly convex regularizer, the acceleration (Ma et al. 2017) can only be done for Lipschitz continuous losses, while our method is able to improve the convergence rate for both smooth and Lipschitz continuous losses. (2) To further reduce the communication cost at each round when handling extremely high-dimensional data, we propose a dynamic feature screening approach to progressively eliminate the features that are associated with zeros values in the optimal solution. Consequently, the communication cost can be substantially reduced as there are only a few features associated with nonzero values in the solution due to the effect of the sparsity regularization. Note that there exist several data or feature screening approaches for single task learning or MTL problems. We believe that this is the first proposed to reduce communication cost in distributed optimization.

Recently, there have been several attempts at developing distributed optimization frameworks for MTL. Baytas et al. (2016) and Xie et al. (2017) proposed asynchronous proximal gradient based algorithms for distributed MTL. Their proposed methods, however, are communication-heavy as gradients need to be frequently communicated among machines. Wang et al. (2016) proposed a Distributed debiased Sparse Multi-task Lasso (DSML) algorithm. In DSML, there is only one round of communication between the local workers and the master. However, it requires the local workers to perform heavy computation (i.e., estimating a $d \times d$ sparse matrix) to obtain a debiased lasso solution. More importantly, DSML makes a stronger assumption to ensure support recovery. More recently, to provide trade-off between local computation and global communication, COCOA+ has been extended for multi-task relationship learning by Liu et al. (2017). Later, this problem is further studied in Smith et al. (2017a) by considering statistical and systems challenges. Note that our work is different from Liu et al. (2017) and Smith et al. (2017a) in two ways: (1) Our proposed method enjoys a faster convergence rate than that analyzed in Liu et al. (2017) and Smith et al. (2017a) since their rates are the same as COCOA+. (2) We study different MTL models. Specifically, Liu et al. (2017) and Smith et al. (2017a) studied task-relationship based MTL model (Zhang and Yeung 2010) while our problem is feature based MTL. They are different as discussed in Zhang and Yang (2017). Moreover, as our work focuses on feature-based MTL model with sparsity (Obozinski et al. 2010, 2011; Wang and Ye 2015), it enables us to design a tailored feature screening technique to further reduce the communication cost. Unlike our framework, decentralized MTL methods have also been studied in Wang et al. (2018), Bellet et al. (2018), Vanhaesebrouck et al. (2017) and Zhang et al. (2018). However, these approaches may incur heavier communication cost because frequent communications are often required between tasks in MTL.

2 Notation and preliminaries

Throughout this paper, $\mathbf{w} \in \mathbb{R}^{dK}$ and $\mathbf{W} \in \mathbb{R}^{d \times K}$ denote a vector and a matrix, respectively, and $\mathscr{G}$ denotes a set.

- $[m] \overset{\text{def}}{=} \{i \mid 1 \leq i \leq m, i \in \mathbb{N}\}$, $\left\{\mathscr{G}_j\right\}_{j=1}^{d}$: $\mathscr{G}_j \overset{\text{def}}{=} \left\{(k-1)d + j \mid k \in [K]\right\}$, $[x]_+ \overset{\text{def}}{=} \max(x, 0)$.
- w_i and W_{ij}: the ith and (i, j)th entries of $\mathbf{w}$ and $\mathbf{W}$, respectively.
- $\mathbf{W}_{i\cdot}$: the ith row of $\mathbf{W}$, $\mathbf{w}_{\mathscr{G}} \overset{\text{def}}{=} \{w_i \mid i \in \mathscr{G}\}$, $\mathbf{W}_{\mathscr{G}\cdot} \overset{\text{def}}{=} \{\mathbf{W}_{i\cdot} \mid i \in \mathscr{G}\}$.
- $\mathbf{0}$: a vector or matrix with all its entries equal to 0, $\mathbf{I}$: identity matrix.
- $\|\mathbf{w}\| \overset{\text{def}}{=} \sqrt{\langle \mathbf{w}, \mathbf{w} \rangle}$: ℓ_2-norm of $\mathbf{w}$, $\|\mathbf{W}\|_{\mathrm{F}} \overset{\text{def}}{=} \sqrt{\mathrm{tr}[\mathbf{W}^\top \mathbf{W}]}$: Frobenius norm of $\mathbf{W}$.
- $\|\mathbf{w}\|_{2,1} \overset{\text{def}}{=} \sum_{j=1}^{d} \|\mathbf{w}_{\mathscr{G}_j}\|$ and $\|\mathbf{W}\|_{2,1} \overset{\text{def}}{=} \sum_{j=1}^{d} \|\mathbf{W}_{j\cdot}\|$: $\ell_{2,1}$-norm of $\mathbf{w}$ and $\mathbf{W}$, respectively.

Definition 1 A function $f(\cdot)$ is L-Lipschitz continuous with respect to $\|\cdot\|$, if $\forall \mathbf{w}, \widehat{\mathbf{w}} \in \mathbb{R}^d$ it holds that $|f(\widehat{\mathbf{w}}) - f(\mathbf{w})| \leq L\|\widehat{\mathbf{w}} - \mathbf{w}\|$.

Definition 2 A function $f(\cdot)$ is L-smooth with respect to $\|\cdot\|$, if $\forall \mathbf{w}, \widehat{\mathbf{w}} \in \mathbb{R}^d$ it holds that $f(\widehat{\mathbf{w}}) \leq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \widehat{\mathbf{w}} - \mathbf{w} \rangle + L\|\widehat{\mathbf{w}} - \mathbf{w}\|^2/2$.

Definition 3 A function $f(\cdot)$ is γ-strongly convex with respect to $\|\cdot\|$, if $\forall \mathbf{w}, \widehat{\mathbf{w}} \in \mathbb{R}^d$ it holds that $f(\widehat{\mathbf{w}}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \widehat{\mathbf{w}} - \mathbf{w} \rangle + \gamma \|\widehat{\mathbf{w}} - \mathbf{w}\|^2/2$.

Definition 4 For function $f(\cdot)$, its convex conjugate $f^*(\cdot)$ is defined as $f^*(\boldsymbol{\alpha}) \overset{\text{def}}{=} \sup_{\mathbf{w}} \left\{\langle \boldsymbol{\alpha}, \mathbf{w} \rangle - f(\mathbf{w})\right\}$.

Lemma 1 (Hiriart-Urruty and Lemaréchal 1993) *Assume that function f is closed and convex. If f is $(1/\gamma)$-smooth w.r.t. $\|\cdot\|$, then f^* is γ-strongly convex w.r.t. the dual norm $\|\cdot\|_*$.*

3 Problem setup

For simplicity, we consider the setting with K tasks distributed over K workers.[1] For each task k, we have n_k labeled instances $\{\mathbf{x}_i^k, y_i^k\}_{i=1}^{n_k}$ stored locally on worker k, where $\mathbf{x}_i^k \in \mathbb{R}^d$ is the ith input, and y_i^k is the corresponding output. Our goal is to jointly learn different models in terms of $\mathbf{w}^k \in \mathbb{R}^d$, $k \in [K]$ for each task. For ease of presentation, we define

- $n \overset{\text{def}}{=} \sum_{k=1}^K n_k$: the total number of training instances over all the tasks.
- $\mathbf{X}^k \overset{\text{def}}{=} \left[\mathbf{x}_1^k, \ldots, \mathbf{x}_{n_k}^k\right] \in \mathbb{R}^{d \times n_k}$ and $\mathbf{y}^k \overset{\text{def}}{=} \left[y_1^k, \ldots, y_{n_k}^k\right]^\top \in \mathbb{R}^{n_k}$: the input and output for task k.
- $\mathbf{W} \overset{\text{def}}{=} [\mathbf{w}^1, \ldots, \mathbf{w}^K] \in \mathbb{R}^{d \times K}$: the weight matrix over all the tasks.
- $\mathbf{A} \overset{\text{def}}{=} \text{diag}\left(\mathbf{X}^1, \ldots, \mathbf{X}^K\right) \in \mathbb{R}^{dK \times n}$, $\mathbf{w} \overset{\text{def}}{=} [(\mathbf{w}^1)^\top, \ldots, (\mathbf{w}^K)^\top]^\top \in \mathbb{R}^{dK}$.

We focus on the following MTL formulation with sparsity regularization (Obozinski et al. 2010, 2011; Lee et al. 2010; Wang and Ye 2015):

$$\min_{\mathbf{W}} \frac{1}{n} f(\mathbf{w}) + \lambda \left(\rho \|\mathbf{W}\|_{2,1} + \frac{1-\rho}{2} \|\mathbf{W}\|_{\text{F}}^2 \right), \tag{1}$$

where $f(\mathbf{w}) \overset{\text{def}}{=} \sum_{k=1}^K \sum_{i=1}^{n_k} f_{ki}\left(\langle \mathbf{x}_i^k, \mathbf{w}^k \rangle\right)$, $f_{ki}\left(\langle \mathbf{x}_i^k, \mathbf{w}^k \rangle\right)$ is the loss function of the kth task on the ith data point $(\mathbf{x}_i^k, y_i^k)$ and $\rho \in (0, 1)$. The group sparsity regularization $\|\mathbf{W}\|_{2,1}$ aims to improve the generalization performance for each task by selecting important features, whose effect to the overall objective is controlled by the parameter λ. Note that the regularization term $\|\mathbf{W}\|_{\text{F}}^2$ is not only to control the complexity of each linear model but also to facilitate distributed optimization.[2] One can rewrite (1) as the following vectorization form,

$$\min_{\mathbf{w}} \left\{ P(\mathbf{w}) \overset{\text{def}}{=} \frac{1}{n} f(\mathbf{w}) + \lambda g(\mathbf{w}) \right\}, \tag{2}$$

where $g(\mathbf{w}) \overset{\text{def}}{=} \rho \sum_{j=1}^d \|\mathbf{w}_{\mathcal{G}_j}\| + (1-\rho)\|\mathbf{w}\|^2/2$.

3.1 Dual problem

Compared to the primal problem, it is well-known that there is a dual variable associated with each training instance in its dual problem. This property makes the dual problem more tractable for distributed optimization if training instances are stored on different workers. Let $\boldsymbol{\alpha} = [\alpha_1^1, \ldots, \alpha_{n_K}^K]^\top \in \mathbb{R}^n$. As derived in “Appendix A”, the dual problem of (2) is

$$\min_{\boldsymbol{\alpha}} \left\{ D(\boldsymbol{\alpha}) \overset{\text{def}}{=} \frac{1}{n} f^*(-\boldsymbol{\alpha}) + \lambda g^* \left(\frac{\mathbf{A}\boldsymbol{\alpha}}{\lambda n} \right) \right\}, \tag{3}$$

[1] In general, the numbers of tasks and workers can be different.

[2] Note that in this work we assume the regularizer is strongly convex which is the same as in COCOA+. As discussed in Sect. 1, for non-strongly convex regularizer, though an extension of COCOA+ has been proposed in Smith et al. (2015, 2017b), it is not practical for real-world scenarios as data needs to be geo-distributed by features rather than instances over local workers. In fact, our proposed method can also be applied to accelerate the approach proposed in Smith et al. (2015, 2017b). However, how to develop a distributed optimization algorithm when data is geo-distributed by instances and the regularizer of the objective is non-strongly convex is still an open problem. We leave this to our future study.

where $f^*(-\boldsymbol{\alpha}) \overset{\text{def}}{=} \sum_{k=1}^{K} \sum_{i=1}^{n_k} f_{ki}^*(-\alpha_i^k)$, $f_{ki}^*(\cdot)$ is the conjugate function of $f_{ki}(\cdot)$ and

$$g^*\left(\frac{\mathbf{A}\boldsymbol{\alpha}}{\lambda n}\right) \overset{\text{def}}{=} \sum_{j=1}^{d} \left\{ g_j^*\left(\frac{\mathbf{A}_{\mathscr{G}_j \cdot}\boldsymbol{\alpha}}{\lambda n}\right) \overset{\text{def}}{=} \frac{\left[\|\mathbf{A}_{\mathscr{G}_j \cdot}\boldsymbol{\alpha}\| - \rho\lambda n\right]_+^2}{2(1-\rho)\lambda^2 n^2} \right\}.$$

Let $\mathbf{w}_\star$ and $\boldsymbol{\alpha}_\star$ be optimal solutions to (2) and (3), respectively. One can obtain a primal solution $\mathbf{w}(\boldsymbol{\alpha})$ from any dual feasible $\boldsymbol{\alpha}$ via

$$\mathbf{w}(\boldsymbol{\alpha}) \overset{\text{def}}{=} \nabla g^*\big(\mathbf{A}\boldsymbol{\alpha}/(\lambda n)\big). \tag{4}$$

Thus, the duality gap at $\boldsymbol{\alpha}$ is $G(\boldsymbol{\alpha}) \overset{\text{def}}{=} P(\mathbf{w}(\boldsymbol{\alpha})) - (-D(\boldsymbol{\alpha})) = P(\mathbf{w}(\boldsymbol{\alpha})) + D(\boldsymbol{\alpha})$.

4 Efficient distributed optimization

For ease of presentation, we further introduce some additional notations. Let $\{\mathscr{P}_k\}_{k=1}^K$ be a partition of $[n]$ such that $\boldsymbol{\alpha}_{\mathscr{P}_k} \in \mathbb{R}^{n_k}$ are the dual variables associated with the training instances of the kth task. For $k \in [K]$, $\mathbf{A} \in \mathbb{R}^{dK \times n}$ and $\mathbf{z} \in \mathbb{R}^n$, we define

- $\widehat{\mathbf{A}}^k \in \mathbb{R}^{dK \times n}$: $\big(\widehat{\mathbf{A}}^k\big)_{\cdot i} \overset{\text{def}}{=} \mathbf{A}_{\cdot i}$ if $i \in \mathscr{P}_k$, otherwise $\mathbf{0}$.
- $\widehat{\boldsymbol{\alpha}}^k \in \mathbb{R}^n$: $\big(\widehat{\boldsymbol{\alpha}}^k\big)_i \overset{\text{def}}{=} \alpha_i$ if $i \in \mathscr{P}_k$, otherwise 0, $\boldsymbol{\alpha}^k \in \mathbb{R}^{n_k}$: $\boldsymbol{\alpha}^k \overset{\text{def}}{=} \boldsymbol{\alpha}_{\mathscr{P}_k}$, $f_k^*(-\widehat{\boldsymbol{\alpha}}^k) \overset{\text{def}}{=} \sum_{i \in \mathscr{P}_k} f_{ki}^*(-\alpha_i^k)$.

Recall that we assume $\{\mathbf{X}^k, \mathbf{y}^k\}_{k=1}^K$ to be stored over K local workers. Therefore, it is highly desirable to develop a communication-efficient distributed optimization method to solve (3). Note that one can adopt COCOA+ (Ma et al. 2015; Smith et al. 2017b) to solve the dual problem, which is similar to the idea of adopting COCOA+ for distributed multi-task relationship learning (Liu et al. 2017; Smith et al. 2017a). However, in this way, the convergence rate of such a COCOA+-based approach fails to reach the best one as discussed in Arjevani and Shamir (2015). To address this problem, we present an efficient distributed optimization method to solve (3) with a faster convergence rate compared with the COCOA+-based approach. The high-level idea of the proposed method is summarized in Algorithm 1, and the details are discussed as follows.

Algorithm 1 Efficient Distributed Optimization for (3)

```
Input: $\{\mathbf{x}_i^k, y_i^k\}_{i=1}^{n_k}$, $k \in [K]$ distributed on $K$ workers, strong convexity parameter $\mu$, which will be formally
    defined in Sect. 5.
Initialize: $\boldsymbol{\alpha}_0 \overset{\text{def}}{=} \mathbf{0}$, $\mathbf{u}_1 \overset{\text{def}}{=} \boldsymbol{\alpha}_0$, $\theta_0 \overset{\text{def}}{=} \sqrt{\vartheta\eta}$ if $\mu > 0$ otherwise $\theta_0 \overset{\text{def}}{=} 1$.
for $t = 1$ to $T$ do
  Send $\mathbf{w}(\mathbf{u}_t) = \nabla g^*\big(\mathbf{A}\mathbf{u}_t/(\lambda n)\big)$ to all workers
  for $k \in [K]$ in parallel over workers do
    Update $\boldsymbol{\alpha}_t^k$ via solving (5)
    Send $\mathbf{A}\widehat{\boldsymbol{\alpha}}_t^k$ to the master
  end for
  Set $\theta_t$ via (7)
  Update $\mathbf{A}\mathbf{u}_{t+1}$ via (8)
end for
```

In order to minimize (3) with respect to $\boldsymbol{\alpha}$ in a distributed environment, one needs to design a subproblem for each worker such that the objective value of (3) decreases when each worker minimizes its local subproblem by only accessing its local data. In (3), the term $f^*(\cdot)$ is separable for examples on different workers but $g^*(\cdot)$ is not. Note that $g^*(\cdot)$ is a smooth function. By Definition 2, it has a quadratic upper bound based on a reference point $\mathbf{u}$ that is separable. By making use of this upper bound, one can design a subproblem for each worker such that $D(\boldsymbol{\alpha})$ decreases if each worker minimizes its local subproblem. Let $\eta \overset{\text{def}}{=} (1-\rho)\lambda n^2$. The following subproblem is used for the kth worker at the tth iteration:

$$\widehat{\boldsymbol{\alpha}}_t^k \overset{\text{def}}{=} \underset{\widehat{\boldsymbol{\alpha}}_t^k \in \mathbb{R}^n}{\operatorname{argmin}}\, L_k\big(\widehat{\boldsymbol{\alpha}}_t^k; \widehat{\mathbf{u}}_t^k, \mathbf{w}(\mathbf{u}_t)\big), \tag{5}$$

where $\mathbf{u}_t$ is a reference point at the tth iteration and

$$L_k\big(\widehat{\boldsymbol{\alpha}}_t^k; \widehat{\mathbf{u}}_t^k, \mathbf{w}(\mathbf{u}_t)\big) \overset{\text{def}}{=} \frac{1}{n} f_k^*\big(-\widehat{\boldsymbol{\alpha}}_t^k\big) + \frac{\lambda}{K} g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big) + \frac{1}{n}\Big\langle \mathbf{w}(\mathbf{u}_t), \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{u}}_t^k\big)\Big\rangle + \frac{1}{2\eta}\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{u}}_t^k\big)\big\|^2. \tag{6}$$

It can be proved that $D(\boldsymbol{\alpha}_t) \le \sum_{k=1}^K L_k\big(\widehat{\boldsymbol{\alpha}}_t^k; \widehat{\mathbf{u}}_t^k, \mathbf{w}(\mathbf{u}_t)\big)$ holds for any $\mathbf{u}_t$. Therefore, $D(\boldsymbol{\alpha})$ can be minimized by employing each local worker to solve its own local subproblem 5. With $\mathbf{w}(\mathbf{u}_t)$, each subproblem can be minimized by only accessing the corresponding local data $(\mathbf{X}^k, \mathbf{y}^k)$.

In the literature of distributed optimization, e.g., COCOA+-based approaches (Ma et al. 2015; Smith et al. 2017a, b; Liu et al. 2017), the reference point $\mathbf{u}_t$ is set to be the solution of last iteration $\boldsymbol{\alpha}_{t-1}$. It leads to that the convergence rate of COCOA+-based approachs fails to reach the best one as discussed in Arjevani and Shamir (2015). In contrast, $\mathbf{u}_t$ in our proposed method is set as follows,

$$\mathbf{u}_{t+1} = \boldsymbol{\alpha}_t + \frac{(1-\theta_{t-1})\theta_{t-1}}{\theta_t + \theta_{t-1}^2}\big(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{t-1}\big),$$

where θ_t is the solution of

$$\theta_t^2 = \big(1-\theta_t\big)\theta_{t-1}^2 + \vartheta\eta\theta_t, \tag{7}$$

where $\vartheta \overset{\text{def}}{=} \mu/n$. The definition of $\mathbf{u}_{t+1}$ implies

$$\mathbf{A}\mathbf{u}_{t+1} = \sum_{k=1}^K \bigg\{\mathbf{A}\widehat{\boldsymbol{\alpha}}_t^k + \frac{\theta_{t-1}\big(1-\theta_{t-1}\big)}{\theta_t + \theta_{t-1}^2}\Big(\mathbf{A}\widehat{\boldsymbol{\alpha}}_t^k - \mathbf{A}\widehat{\boldsymbol{\alpha}}_{t-1}^k\Big)\bigg\}. \tag{8}$$

Specifically, $\mathbf{u}_{t+1}$ is obtained based on an extrapolation from $\boldsymbol{\alpha}_t$ and $\boldsymbol{\alpha}_{t-1}$. This is similar to Nesterov's acceleration technique (Nesterov 2013). As we will see, this technique yields a faster convergence rate compared to COCOA+-based approaches (Ma et al. 2015; Smith et al. 2017a, b; Liu et al. 2017). Recently, Zheng et al. (2017) presented an accelerated distributed alternating dual maximization algorithm for single task learning, where an extrapolation is applied on the primal variable for acceleration. For smooth losses, they only proved the accelerated convergence rate in terms of primal suboptimality while we also prove it for duality gap, resulting in a stronger result.

Remark 1 In each iteration of Algorithm 1, $\mathbf{w}(\mathbf{u}_t)$ and $\{\mathbf{A}\widehat{\boldsymbol{\alpha}}_t^k\}_{k=1}^K$ are communicated between master and workers. By the definitions of $\mathbf{A}$ and $\widehat{\boldsymbol{\alpha}}_t^k$, we note that $\big(\mathbf{w}(\mathbf{u}_t)\big)^k \in \mathbb{R}^d$ and $\mathbf{X}^k\boldsymbol{\alpha}_t^k \in \mathbb{R}^d$ are actually communicated between master and the kth worker. Therefore, its communication cost for each iteration is the same as COCOA+ in which $\big(\mathbf{w}(\boldsymbol{\alpha}_t)\big)^k \in \mathbb{R}^d$ and $\mathbf{X}^k\boldsymbol{\alpha}_t^k \in \mathbb{R}^d$ are communicated. Note that $\mathbf{w}(\mathbf{u}_{t+1})$ depends on $\mathbf{A}\boldsymbol{\alpha}_t$ but also $\mathbf{A}\boldsymbol{\alpha}_{t-1}$, therefore we can keep a copy of $\mathbf{A}\boldsymbol{\alpha}_{t-1}$ on the master until iteration t. In this way, no extra communication cost is induced in each iteration by our method for acceleration.

5 Convergence analysis

In this section, we analyze the convergence rate of the proposed method and show that it is faster than COCOA+-based approaches. All the proofs can be found in "Appendix". In our analysis, we assume that all f_{ki}^*, $k \in [K]$, $i \in [n_k]$ are μ-strongly convex ($\mu \geq 0$) with respect to the norm $\|\cdot\|$. According to Lemma 1, it is equivalent to assuming that all f_{ki}, for $k \in [K]$ and $i \in [n_k]$ are $(1/\mu)$-smooth with respect to the norm $\|\cdot\|$. Since μ is allowed to be 0, our analysis also covers the case that all f_{ki}^*, for $k \in [K]$ and $i \in [n_k]$ are only generally convex (i.e., $\mu = 0$), which implies that all f_{ki} for $k \in [K]$ and $i \in [n_k]$ are Lipschitz continuous instead of smooth. To facilitate analysis, we also assume that $L_k\big(\widehat{\boldsymbol{\alpha}}_t^k; \widehat{\mathbf{u}}_t^k, \mathbf{w}(\mathbf{u}_t)\big)$ is exactly solved for any $k \in [K]$ and $t \geq 1$.

By defining $\zeta_t \overset{\text{def}}{=} \theta_t^2/\eta$, (7) becomes

$$\zeta_t = \big(1 - \theta_t\big)\zeta_{t-1} + \vartheta\theta_t. \tag{9}$$

For any $t \geq 1$ and $k \in [K]$, $\widehat{\mathbf{v}}_t^k$ is defined as

$$\widehat{\mathbf{v}}_t^k \overset{\text{def}}{=} \widehat{\boldsymbol{\alpha}}_{t-1}^k + \big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\boldsymbol{\alpha}}_{t-1}^k\big)\big/\theta_t, \quad k \in [K]. \tag{10}$$

In addition, the suboptimality on dual objective function ϵ_D^t is defined as $\epsilon_D^t \overset{\text{def}}{=} D(\boldsymbol{\alpha}_t) - D(\boldsymbol{\alpha}_\star)$, $t \geq 0$. By using the above notations, the following lemma shows that there is an upper bound for the suboptimality ϵ_D^t. As we will see, this is the foundation for analyzing the convergence rate of duality gap.

Lemma 2 *Consider applying Algorithm 1 to solve (3), the following inequality holds for any $t \geq 1$,*

$$\epsilon_D^t + R^t \leq \gamma_t\big(\epsilon_D^0 + R^0\big), \tag{11}$$

where $R^t = \frac{\zeta_t}{2}\sum_{k=1}^K \big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_t^k\big)\big\|^2$, $\gamma_t = \prod_{i=1}^t \big(1 - \theta_i\big)$ for any $t \geq 1$ and $\gamma_0 = 1$.

It can be found that the form of γ_t determines the convergence rate of Algorithm 1. Therefore, next, we study the convergence rate by using the upper bound of γ_t under different settings of the loss function.

5.1 Convergence rate for smooth losses

By applying Lemma 2, the following lemma characterizes the effect of iterations of Algorithm 1 when the loss functions f_{ki}'s are $(1/\mu)$-smooth for any $k \in [K]$ and $i \in [n_k]$.

Lemma 3 *Assume the loss functions f_{ki}'s are $(1/\mu)$-smooth for any $k \in [K]$ and $i \in [n_k]$. If $\theta_0 = \sqrt{\vartheta\eta}$ and $(1 - \rho)\lambda\mu n \leq 1$, then the following inequality holds for any $t \geq 1$*

$$\epsilon_D^t \leq \Big(1 - \sqrt{(1 - \rho)\lambda\mu n}\Big)^t \big(\epsilon_D^0 + R^0\big). \tag{12}$$

Let $\sigma_{\max} \overset{\text{def}}{=} \max_{\boldsymbol{\alpha} \neq 0} \|\mathbf{A}\boldsymbol{\alpha}\|^2/\|\boldsymbol{\alpha}\|^2$. By applying Lemma 3, the next theorem shows the communication complexities for smooth losses in terms of dual objective and duality gap.

Theorem 1 *Assume the loss functions f_{ki}'s are $(1/\mu)$-smooth for any $k \in [K]$ and $i \in [n_k]$. If $\theta_0 = \sqrt{\vartheta\eta}$ and $(1-\rho)\lambda\mu n \leq 1$, then after T iterations in Algorithm 1 with*

$$T \geq \sqrt{\frac{1}{(1-\rho)\lambda\mu n}} \log\left(\left(1+\sigma_{max}\right)\frac{\epsilon_D^0}{\epsilon_D}\right),$$

$D(\boldsymbol{\alpha}_T) - D(\boldsymbol{\alpha}_\star) \leq \epsilon_D$ holds. Furthermore, after T iterations with

$$T \geq \sqrt{\frac{1}{(1-\rho)\lambda\mu n}} \log\left(\left(1+\sigma_{max}\right)\frac{(1-\rho)\lambda\mu n + \sigma_{max}}{(1-\rho)\lambda\mu n}\frac{\epsilon_D^0}{\epsilon_G}\right),$$

it holds that $P(\mathbf{w}(\boldsymbol{\alpha}_T)) - (-D(\boldsymbol{\alpha}_T)) \leq \epsilon_G$.

Following Zhang and Xiao (2017), we define the condition number κ as $\kappa \overset{\text{def}}{=} \max_{k,i} \|\mathbf{x}_i^k\|^2/(\lambda\mu)$. With the above analysis, the communication complexity of our method is linear with respect to $\sqrt{\kappa}$, while it is linear with κ for COCOA+-based approaches (Ma et al. 2015; Smith et al. 2017b). The value of κ is typically the order of n as λ is usually set to the order of $1/n$ (Bousquet and Elisseeff 2002). Therefore, our method is expected to converge faster than COCOA+-based approaches.

5.2 Convergence rate for Lipschitz continuous losses

Next, we present the convergence rate of the Algorithm 1 when the loss function is just general convex and Lipschitz continuous.

Theorem 2 *Assume the loss functions f_{ki}'s are generally convex and L-Lipschitz continuous for any $k \in [K]$, $i \in [n_k]$. If $\theta_0 = 1$, the following inequality holds for any $t \geq 1$*

$$\epsilon_D^t \leq \frac{1}{(t+2)^2}\left(4\epsilon_D^0 + \frac{8L^2\sigma_{max}}{(1-\rho)\lambda n^2}\right). \tag{13}$$

After T iterations in Algorithm 1 with

$$T \geq \sqrt{\frac{8L^2\sigma_{max}}{(1-\rho)\lambda n^2\epsilon_D} + \frac{4\epsilon_D^0}{\epsilon_D}} - 2, \tag{14}$$

it holds that $D(\boldsymbol{\alpha}_T) - D(\boldsymbol{\alpha}_\star) \leq \epsilon_D$.

Remark 2 For generally convex loss function, the dual objective obtained by Algorithm 1 decreses in $O(1/t^2)$ instead of $O(1/t)$ for COCOA+. Therefore, the complexity for obtaining ϵ_D-suboptimal solution is $\sqrt{1/\epsilon_D}$ that is faster than that of COCOA+ (i.e., $1/\epsilon_D$).

6 Further reduce communication cost via dynamic feature screening

In Sect. 4, we present an acceleration method for distributed optimization of (3) that reduces the communication cost in terms of iteration of communications. As discussed in Remark 1, the communication cost of our method in each iteration is linear with the number of features d, that is the same as previous distributed optimization methods for sparsity-regularized problems. It can be expensive for high-dimensional data. To address this issue, we present a method to reduce the communication cost for each iteration by exploiting the sparsity of $\mathbf{w}_\star$ (Bonnefoy et al. 2015; Fercoq et al. 2015; Ndiaye et al. 2017). It is well-known that the $\ell_{2,1}$-norm regularization is able to produce a row sparse pattern on $\mathbf{W}_\star$ (Obozinski et al. 2011, 2010; Yuan et al. 2006; Zou and Hastie 2005). In other words, $(\mathbf{w}_\star)_{\mathscr{G}_j}$ will be $\mathbf{0}$ for most $\mathscr{G}_j$, $j \in [d]$. Thereafter, we refer the jth feature as an *inactive feature* if $(\mathbf{w}_\star)_{\mathscr{G}_j} = \mathbf{0}$, otherwise an *active feature*. The key idea of feature screening is to identify inactive features before sending the updated information to workers (Line 4 in Algorithm 1). In this way, the communication cost can be reduced since it is linear with the number active features.

To identify inactive features, we need to exploit the KKT condition of (2)

$$(\alpha_\star)_i^k \in \partial f_{ki}\big(\langle \mathbf{x}_i^k, \mathbf{w}_\star^k \rangle\big), \forall k \in [K], i \in [n_k], \tag{15}$$

$$\frac{\mathbf{A}_{\mathscr{G}_j\cdot}\boldsymbol{\alpha}_\star}{\lambda n} \in (1-\rho)\big(\mathbf{w}_\star\big)_{\mathscr{G}_j} + \rho\partial\big\|(\mathbf{w}_\star)_{\mathscr{G}_j}\big\|, \forall j \in [d]. \tag{16}$$

By checking the subgradient of $\|\cdot\|$, it implies $(\mathbf{w}_\star)_{\mathscr{G}_j} = \mathbf{0}$ if $\|(\mathbf{w}_\star)_{\mathscr{G}_j}\| < 1$. Combining this fact with (16), we have

$$\big\|\mathbf{A}_{\mathscr{G}_j\cdot}\boldsymbol{\alpha}_\star\big\| < \rho\lambda n \;\Rightarrow\; (\mathbf{w}_\star)_{\mathscr{G}_j} = \mathbf{0}. \tag{17}$$

Algorithm 2 Dynamic Feature Screening for (3)

```
Input: ${\mathbf{A}\widehat{\boldsymbol{\alpha}}_t^k\}_{k=1}^K$
Compute duality gap $G(\boldsymbol{\alpha}_t)$
for all currently active features do
    Identify inactive feature via solving (19)
end for
```

It can be shown that one can obtain the exact optimum even without considering these inactive features during optimization. Therefore, one can reduce the communication cost by discarding these inactive features, thus less information needs to be communicated. To use (17) to identify inactive features, one needs to have $\boldsymbol{\alpha}_\star$ that is unknown before the optimization problem (3) is solved. Next, we show that a feasible set $\mathscr{F}$ can be constructed for $\boldsymbol{\alpha}_\star$ by using the strong convexity of $D(\boldsymbol{\alpha})$.

Crucial Value $\lambda_{\max}$: In view of (17) and (15), there exists a crucial value $\lambda_{\max}$ such that $\mathbf{w}_\star = \mathbf{0}$ for any $\lambda \geq \lambda_{\max}$. Let $\mathbf{r} = [f'_{11}(0), \ldots, f'_{Kn_K}(0)] \in \mathbb{R}^n$, (15) implies that $\boldsymbol{\alpha}_\star = \mathbf{r}$ when $\mathbf{w}_\star = \mathbf{0}$. By substituting $\boldsymbol{\alpha}_\star$ into (17), we obtain $\lambda_{\max} = \max_{j\in[d]} \|\mathbf{A}_{\mathscr{G}_j}\mathbf{r}\|/(\rho n)$. It is trivial to obtain a closed form solution $\mathbf{w}_\star = \mathbf{0}$ and $\boldsymbol{\alpha}_\star = \mathbf{r}$ if $\lambda \geq \lambda_{\max}$. Therefore, we only focus on the cases when $\lambda < \lambda_{\max}$.

Feasible Set of $\boldsymbol{\alpha}_\star$: Lemma 1 implies $D(\boldsymbol{\alpha})$ is strongly convex if f_{ki}'s are smooth for all k and i. By using this fact, the dual optimal solution $\boldsymbol{\alpha}_\star$ can be bounded in terms of $\boldsymbol{\alpha}$ and its duality gap $G(\boldsymbol{\alpha})$ as stated in the following lemma.

Lemma 4 *Assume the loss functions f_{ki}'s are $(1/\mu)$-smooth for any $k \in [K]$, $i \in [n_k]$. For any dual feasible solution $\boldsymbol{\alpha}$, it holds that $\boldsymbol{\alpha}_\star \in \mathscr{F} \overset{\text{def}}{=} \big\{\boldsymbol{\theta} \mid \|\boldsymbol{\theta} - \boldsymbol{\alpha}\| \leq \sqrt{2G(\boldsymbol{\alpha})n/\mu}\big\}$.*

By using Lemma 4, (17) can be relaxed as

$$\max_{\boldsymbol{\theta}\in\mathscr{F}} \big\|\mathbf{A}_{\mathscr{G}_j\cdot}\boldsymbol{\theta}\big\| < \rho\lambda n \;\Rightarrow\; (\mathbf{w}_\star)_{\mathscr{G}_j} = \mathbf{0}. \tag{18}$$

In other words, we need to solve the following problem

$$\max_{\boldsymbol{\theta}} \left\|\mathbf{A}_{\mathscr{G}_j}.\boldsymbol{\theta}\right\|, \quad \text{s.t. } \|\boldsymbol{\theta}-\boldsymbol{\alpha}\| \leq \sqrt{2G(\boldsymbol{\alpha})n/\mu}. \tag{19}$$

Although it is non-convex, the global optimum of (19) can be obtained by using the result in Gay (1981). Let us define $\mathbf{H} \in \mathbb{R}^{K\times K}$, $\mathbf{g} \in \mathbb{R}^K$, υ_j, $\mathscr{I}_j$, $\overline{\mathscr{I}_j}$ and $\overline{\mathbf{s}} \in \mathbb{R}^K$ as

- $\mathbf{H} \overset{\text{def}}{=} -\text{diag}\big(2\|\mathbf{X}^1_{j\cdot}\|^2, \ldots, 2\|\mathbf{X}^K_{j\cdot}\|^2\big), \mathbf{g} \overset{\text{def}}{=} -2\big[\|\mathbf{X}^1_{j\cdot}\|\,|\langle\mathbf{X}^1_{j\cdot}, \boldsymbol{\alpha}^1\rangle|, \ldots, \|\mathbf{X}^K_{j\cdot}\|\,|\langle\mathbf{X}^K_{j\cdot}, \boldsymbol{\alpha}^K\rangle|\big]^\top$.
- $\upsilon_j \overset{\text{def}}{=} \max_{k\in[K]} \|\mathbf{X}^k_{j\cdot}\|^2$, $\mathscr{I}_j \overset{\text{def}}{=} \Big\{k \mid \|\mathbf{X}^k_{j\cdot}\|^2 = \upsilon_j, k \in [K]\Big\}$, $\overline{\mathscr{I}_j} \overset{\text{def}}{=} [K] \setminus \mathscr{I}_j$.
- $\overline{s_k} \overset{\text{def}}{=} \frac{\|\mathbf{X}^k_{j\cdot}\|\,|\langle\mathbf{X}^k_{j\cdot},\boldsymbol{\alpha}^k\rangle|}{\upsilon_j - \|\mathbf{X}^k_{j\cdot}\|^2}$ if $k \in \overline{\mathscr{I}_j}$, otherwise $\overline{s_k} \overset{\text{def}}{=} 0$.

By using the above notations, the solution of (19) is given in the following lemma.

Lemma 5 *If $\upsilon_j = 0$, the maximum value of (19) is 0. Otherwise, the upper bound is*

$$\sum_{k=1}^{K} \langle\mathbf{X}^k_{j\cdot}, \boldsymbol{\alpha}^k\rangle^2 + \frac{nG(\boldsymbol{\alpha})}{\mu}\vartheta_\star - \frac{1}{2}\langle\mathbf{g}, \mathbf{s}_\star\rangle\,,$$

where $\vartheta_\star$ and $\mathbf{s}_\star$ are defined as follows: (a) $\vartheta_\star = 2\upsilon_j$ and $\mathbf{s}^\star = \overline{\mathbf{s}} + \widehat{\mathbf{s}}$ if 1) $\exists\,\widehat{\mathbf{s}} \in \mathbb{R}^K$ with $\widehat{\mathbf{s}}_{\mathscr{I}_j} = \mathbf{0}$ and $\|\overline{\mathbf{s}} + \widehat{\mathbf{s}}\| = \sqrt{2G(\boldsymbol{\alpha})n/\mu}$, and 2) $\Big\langle\mathbf{X}^t_{\cdot j}, \boldsymbol{\theta}_t\Big\rangle = 0$, $\forall t \in \mathscr{I}_j$. (b) Otherwise, $\vartheta_\star > 2\upsilon_j$ is solution of $\|\,(\mathbf{H}+\vartheta_\star\mathbf{I})^{-1}\,\mathbf{g}\| = \sqrt{2G(\boldsymbol{\alpha})n/\mu}$, and $\mathbf{s}_\star = -(\mathbf{H}+\vartheta_\star\mathbf{I})^{-1}\mathbf{g}$.

To perform screening every p iterations, one can simply add the following three lines before line 4 in Algorithm 1.

```
– if t%p = 0 then
–     Call Algorithm 2
– end if
```

Costs of Screening: Note that the screening is performed on the master every p iterations.

- By carefully examining the detailed screening rule, the master actually only needs $\mathbf{A}\boldsymbol{\alpha}_t$ when evaluating screening rule. Even without screening, the $\mathbf{A}\boldsymbol{\alpha}_t$ needs to be computed and sent to the master in each iteration as stated in Algorithm 1 and Remark 1. Therefore, the feature screening does not induce extra communication cost.
- Regarding the computational cost, we note that the screening problem is dependent on the number of active features that is at most d (there are less and less feature due to screening). As shown in Lemma 5, the screening problem for each feature is a one dimension variable optimization problem. It either has a closed form solution (Case 1) or can be efficiently solved by using Newton's method (Case 2) that usually takes less than 5 iterations to meet the accuracy 10^{-15}.
- More importantly, by screening out inactive features, it can substantially save optimization problem, especially on local computation. Recall that the local SDCA computation complexity is $O(Hd)$ where H is the local SDCA iteration number and its is usually more than 10^5. Compared to local SDCA computation cost, the cost of screening is negligible.

We note that Ndiaye et al. (2015) also presented a feature screening method for multi-task learning. However, in their work, all tasks are assumed to share the same training data while our method allows each task to has its own training data. Consequently, the feature screening problem (19) becomes non-convex instead of convex, which is different from and more challenging than that studied in Ndiaye et al. (2015). In addition, Wang and Ye (2015) developed a static screening rule that exploits the solution at another regularization parameter and only performs screening before the optimization procedure. By contrast, our screening rule is a dynamic with a weaker assumption to exploit the latest solution to repeatedly perform screening during optimization. Therefore, our screening is more practical and performs better empirically.

Difference between Our Proposed Method and COCOA+ We denote the proposed method by $\mathtt{DMTL}_{\mathcal{S}}$. There are two main differences between $\mathrm{DMTL}_{\mathcal{S}}$ and COCOA+. First, $\mathrm{DMTL}_{\mathcal{S}}$ constructs the subproblem 5 by using the extrapolation of the solutions in last two iterations that is able to achieve accelerated convergence rate. In contrast, COCOA+ only uses the solution of last iteration. Second, $\mathrm{DMTL}_{\mathcal{S}}$ presents a dynamic feature screening method to reduce the communication cost for each iteration by exploiting the sparsity of the model.

7 Experiments

7.1 Experimental setting

In previous sections, we present our method by focusing on distributed MTL. We hereby conduct experiments to show the advantages of the proposed method for MTL. In fact, our approach can also be extended for distributed single task learning (STL) and the details are provided in the "Appendix".

To demonstrate the advantages of $\mathtt{DMTL}_{\mathcal{S}}$, we compare $\mathtt{DMTL}_{\mathcal{S}}$ with a COCOA+-based approach (Ma et al. 2015; Smith et al. 2017b) and its extension MOCHA (Smith et al. 2017a) to solve the dual problem (3). In our experiments, the squared loss is used for regression, and the smoothed hinge loss (Shalev-Shwartz and Zhang 2013) is used for classification with $\mu = 0.5$ for all experiments. It is clear to see that f_{ki} is $(1/\mu)$-smooth. For ease of comparison, the local subproblem is solved by using SDCA (Shalev-Shwartz and Zhang 2013) for all methods. The number of iterations for SDCA is set to $H = 10^4$ for all datasets.

We run all experiments on a local server with 64 worker cores. A distributed environment is simulated on the machine by using distributed platform Petuum (Xing et al. 2015),[3] and workers for each task are assigned to isolated processes that communicate solely through the platform. Regarding the performance, we evaluate the number of communication iterations required by different methods to obtain a solution with prescribed duality gap. Due to the limitation of computational resources, we are not able to perform experiments on a real distributed environment. However, the results (i.e., the number of communication iterations) reported in this paper does not depend on the environment that it runs on. Compared to COCOA+, the additional computation incurred by our method is negligible: the computation complexity of each iteration of COCOA+ is $O(H \times d)$. The additional computations required by our method for acceleration and feature screening is $O(d)$ and $O(d)$, respectively. This cost is negligible compared to that of SDCA because H is usually around 10^5.

We conduct experiments on the following three datasets (Table 1).

Synthetic Data contains $K = 10$ regression tasks and generated by using $y_i^k = \langle \mathbf{x}_i^k, \mathbf{w}^k \rangle + \epsilon$. The number of examples for each task is randomly generated, which ranges from 903 to 1098. $\mathbf{x}_i^k \in \mathbb{R}^{50,000}$ is drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, 0.5\mathbf{I})$. To obtain a $\mathbf{W}$ with row

[3] Note that our method can be implemented in other distributed platforms.

Table 1 Statistics of the datasets for MTL

Dataset	Synthetic	News20	MDS
# Tasks	10	5	22
# Samples	9081	5869	16,967
# Features	50,000	34,967	10,000
Sparsity (%)	100	0.3	0.8

sparsity, we randomly select 400 dimensions from $[d]$ and generate them from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ for all tasks. For each task, extra noise from $\mathcal{N}(\mathbf{0}, 0.5\mathbf{I})$ is added to $\mathbf{W}$.

News20 (Lang 1995) is a collection of around 20,000 documents from 20 different newsgroups. To construct a multi-task learning problem, we create 5 binary classification tasks using data of all the 5 groups from *comp* as positive examples. For the negative examples, we choose data from *misc.forsale, rec.autos, rec.motorcycles, rec.sport.baseball* and *rec.sport.hockey*. The number of training examples for each task ranges from 1163 to 1190, and the number of features is 34,967.

MDS (Blitzer et al. 2007) includes product reviews on 25 domains in Amazon. We use 22 domains each of which has more than 100 examples for multitask binary sentiment classification. To simulate MTL, we randomly select 1000 examples as training data for the domain with more than 1000 examples. Consequently, the number training examples for each domain ranges from 220 to 1000. The number of features of is 10,000.

7.2 Results of faster convergence rate

In order to test the convergence rate of $\texttt{DMTL}_{S}$, we compare it with the COCOA+-based approach to solving (3) under varying values of λ. In view of Sect. 6, we chose $\lambda = 10^{-2}\lambda_{\max}$ and $\lambda = 10^{-3}\lambda_{\max}$ to solve (3). We set $\boldsymbol{\alpha}_0 = \mathbf{0}$ for all methods and $\rho = 0.9$ for all experiments.

Figure 1 shows the comparison results in terms of the numbers of iterations for communication used by $\texttt{DMTL}_{S}$ and COCOA+ to obtain a solution meeting a prescribed duality gap. From the Fig. 1, we can observe that:

- $\texttt{DMTL}_{S}$ is significantly faster than COCOA+ in terms of the number of iterations to meet a prescribed duality gap. Take the synthetic dataset and News20 for example, to obtain a solution atλ=$10^{-3}\lambda_{\max}$ with duality gap 10^{-5}, $\texttt{DMTL}_{S}$ obtains speedups of a factor of 6.64 and 6.94 over COCOA+ on the two datasets, respectively.
- Generally, the speedup obtained by $\texttt{DMTL}_{S}$ is more significant for small values of λ. For example, when $\lambda = 10^{-2}\lambda_{\max}$, $\texttt{DMTL}_{S}$ converges 4.81 and 4.05 times faster than COCOA+ on the synthetic dataset and News20, respectively. In contrast, the speedups is improved up to 7.00 and 5.70 times faster than COCOA+ when $\lambda = 10^{-3}\lambda_{\max}$.
- The improvement is more pronounced when a higher precision is used as the stopping criterion. Take News20 with $\lambda = 10^{-3}\lambda_{\max}$ for example, the speedups of $\texttt{DMTL}_{S}$ over COCOA+ are 4.00, 4.94, 5.70 and 6.94 when the duality gaps are 10^{-2}, 10^{-3}, 10^{-4} and 10^{-5}, respectively.

7.3 Robust to straggler

In Smith et al. (2017a), MOCHA is proposed to improve COCOA+ to handle systems heterogeneity, e.g., straggler. That means some workers are considerably slower than others and

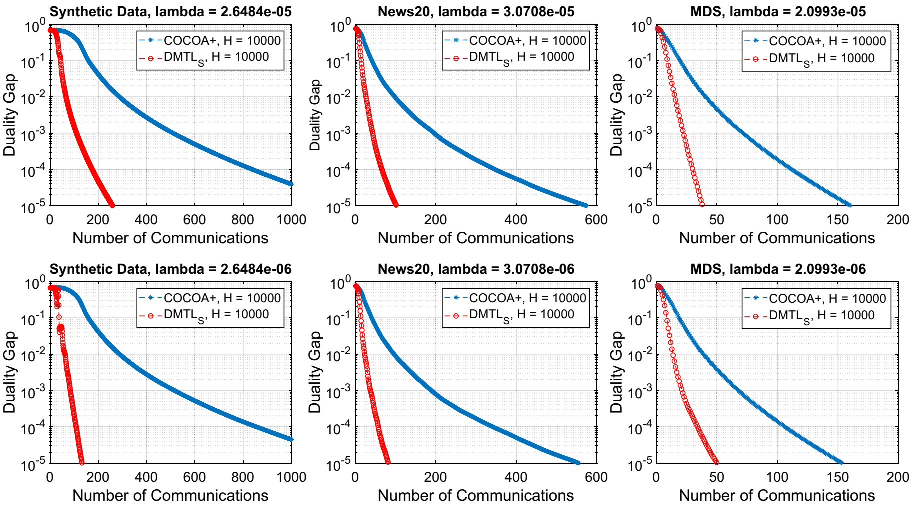

Fig. 1 Duality gap versus communicated iterations on the three datasets for $\lambda = 10^{-2}\lambda_{\max}$ and $\lambda = 10^{-3}\lambda_{\max}$

the stragglers fail to return prescribed accurate solution for some iterations. Here, we compare our method with COCOA+ equipped with handling system heterogeneity as presented in Smith et al. (2017a) on News20 and show that our method converges faster even if there exist stragglers. Specifically, we perform experiments under the setting of Smith et al. (2017a) by using different values of H for different workers to simulate the effect of stragglers. The value of H for each iteration is draw from $[0.9n_{\min}, n_{\min}]$ to simulate *low variability* environment and $[0.5n_{\min}, n_{\min}]$ to simulate *high variability* environment, where $n_{\min} = \min_k n_k$.

As shown in Fig. 2, our method is still able to substantially reduce the number of communication for both low and high variability environments. This shows that empirically DMTL$_S$ is robust to straggler although our analysis assumes that the local subproblem needs to be exactly solved.

7.4 Results of reduced communication cost

To demonstrate the effect of dynamic screening for reducing communication cost, we perform a warm start cross validation experiment on News20 and MDS. Specifically, we solve (3) with 50 various values of λ, $\{\lambda_i\}_{i=1}^{50}$, which are equally distributed on the logarithmic grid from $0.01\lambda_{\max}$ to $0.3\lambda_{\max}$ sequentially (i.e., the solution of λ_i is used as the initialization of λ_{i-1}). To evaluate the total communication cost for the 50 values of λ, we calculate the total number of vectors of dimension d used for communication for each worker. We experiment on the following two settings: 1) DMTL$_S$ *without* dynamic screening (Without DS), and 2) DMTL$_S$ *with* dynamic screening (With DS). Figure 3 presents the total communication cost used by DMTL$_S$ *without* and *with* dynamic screening to solve (3) over $\{\lambda_i\}_{i=1}^{50}$ on News20 and MDS.

From Fig. 3, we can observe that:

- The communication cost has been substantially reduced by the proposed dynamic screening because the most inactive features have been progressively identified and discarded during optimization. For example, when the prescribed duality is 10^{-7}, the communi-

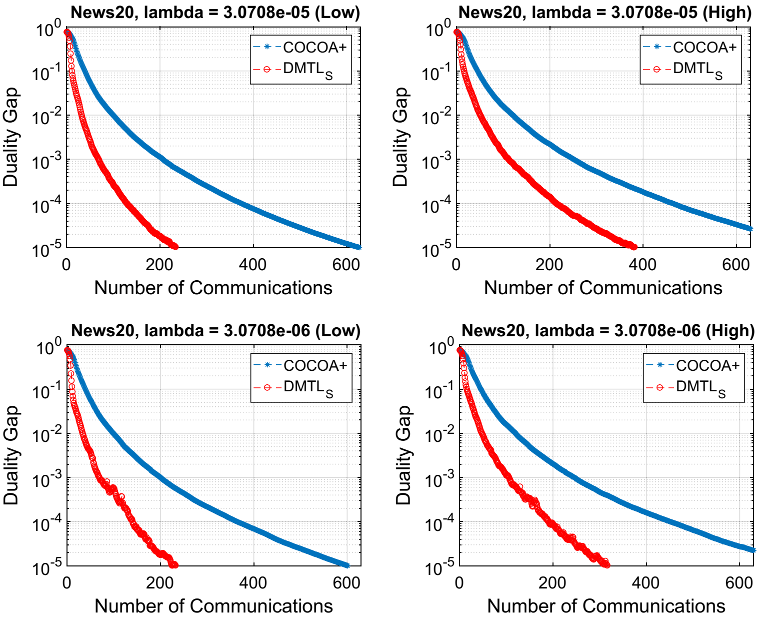

Fig. 2 Duality gap versus communicated iterations on News20 with *systems heterogeneity* for $\lambda = 10^{-2}\lambda_{\max}$ and $\lambda = 10^{-3}\lambda_{\max}$. Here, COCOA+ denotes its original version equipped with handling system heterogeneity as presented in Smith et al. (2017a)

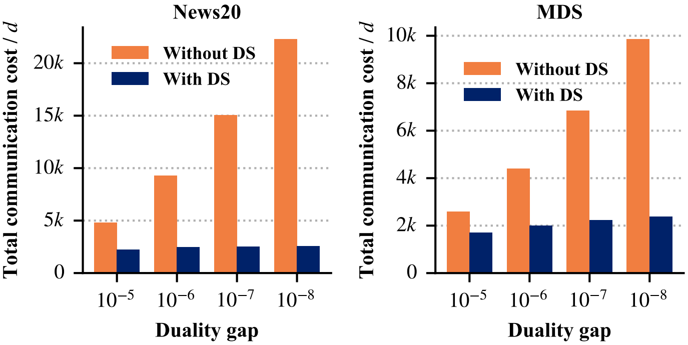

Fig. 3 Effect of dynamic screening for reducing communication cost. Total communication cost (normalized by feature dimension d) used by our method *without* and *with* dynamic screening for solving (3) over $\{\lambda_i\}_{i=1}^{50}$ on News20 and MDS

cation cost reduction by the proposed method is 83.32% and 67.43% on News20 and MDS, respectively.

- This advantage of dynamic screening is more significant when a higer precision is used as the stopping criterion. On News20, the speedup increases from 5.99 to 8.63 when the duality gap changes from 10^{-7} to 10^{-8}. This is because more inactive features can be screened out when a more accurate solution is obtained.
- More importantly, the proposed dynamic screening is more pronounced for the problem with higher dimension. Take the duality gap of 10^{-8} for example, the speedups

obtained by dynamic screening are 8.63 and 4.14 on News20 and MDS, respectively, where News20 is of much higher dimensionality than MDS.

8 Conclusion

In this paper, we present a new distributed optimization method, $\mathtt{DMTL}_S$, for MTL with matrix sparsity regularization. We provide theoretical convergence analysis for $\mathtt{DMTL}_S$. We also propose a data screening method to further reduce the communication cost. We carefully design and conduct extensive experiments on both synthetic and real-world datasets to verify the faster convergence rate and the reduced communication cost of $\mathtt{DMTL}_S$ in comparison with two state-of-the-art baselines, COCOA+ and MOCHA.

Acknowledgements This work is supported by NTU Singapore Nanyang Assistant Professorship (NAP) Grant M4081532.020 and Singapore MOE AcRF Tier-2 Grant MOE2016-T2-2-060.

Appendix A: Dual problem

By introducing z_i^k for each f_{ki}, one can rewrite (2) as

$$\begin{aligned}&\min_{\mathbf{w}} \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} f_{ki}(-z_i^k)\\ &\quad+\lambda\left(\rho\sum_{j=1}^{d}\|\mathbf{w}_{\mathscr{G}_j}\|+\frac{1-\rho}{2}\|\mathbf{w}\|^2\right)\ \text{ s.t. } \langle \mathbf{x}_i^k,\mathbf{w}^k\rangle+z_i^k=0, k\in[K], i\in[n_k].\end{aligned}$$

Let $-\frac{1}{n}\alpha_i^k$ be the Lagrangian multiplier for the (k,i)th constraint. For convenience, let

$$\mathbf{z}=\left[z_1^1,\ldots,z_{n_K}^K\right]^\top\in\mathbb{R}^n \ \text{ and } \ \boldsymbol{\alpha}=\left[\alpha_1^1,\ldots,\alpha_{n_K}^K\right]^\top\in\mathbb{R}^n.$$

Then, the Lagrangian is

$$\begin{aligned}L\big(\mathbf{w},\mathbf{z},\boldsymbol{\alpha}\big)&=\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_K} f_{ki}(-z_i^k)+\lambda\left(\rho\sum_{j=1}^{d}\|\mathbf{w}_{\mathscr{G}_j}\|+\frac{1-\rho}{2}\|\mathbf{w}\|^2\right)\\ &\quad-\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_K}\alpha_i^k\left(\langle\mathbf{x}_i^k,\mathbf{w}^k\rangle+z_i^k\right)\\ &=\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_K}\left(f_{ki}(-z_i^k)-\alpha_i^k z_i^k\right)+\lambda\left(\rho\sum_{j=1}^{d}\|\mathbf{w}_{\mathscr{G}_j}\|+\frac{1-\rho}{2}\|\mathbf{w}\|^2\right)\\ &\quad-\frac{1}{n}\left\langle\mathbf{A}\boldsymbol{\alpha},\mathbf{w}\right\rangle.\end{aligned}$$

The dual problem can be obtained by taking the infimum with respect to both $\mathbf{w}$ and $\mathbf{z}$

$$\inf_{\mathbf{w},\mathbf{z}} L(\mathbf{w},\mathbf{z},\boldsymbol{\alpha}) = \frac{1}{n}\inf_{\mathbf{z}}\sum_{k=1}^{K}\sum_{i=1}^{n_K}\Big(f_{ki}(-z_i^k) - \alpha_i^k z_i^k\Big) \\ + \inf_{\mathbf{w}}\left\{\lambda\bigg(\rho\sum_{j=1}^{d}\|\mathbf{w}_{\mathscr{G}_j}\| + \frac{1-\rho}{2}\|\mathbf{w}\|^2\bigg) - \frac{1}{n}\langle\mathbf{A}\boldsymbol{\alpha},\mathbf{w}\rangle\right\}.$$

where

$$\frac{1}{n}\inf_{\mathbf{z}}\sum_{k=1}^{K}\sum_{i=1}^{n_K}\Big(f_{ki}(-z_i^k) - \alpha_i^k z_i^k\Big) = -\frac{1}{n}\sup_{\mathbf{z}}\sum_{k=1}^{K}\sum_{i=1}^{n_K}\Big(\alpha_i^k z_i^k - f_{ki}(-z_i^k)\Big) \\ = -\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_K} f_{ki}^*(-\alpha_i^k). \tag{20}$$

$$\lambda\inf_{\mathbf{w}}\left\{\rho\sum_{j=1}^{d}\|\mathbf{w}_{\mathscr{G}_j}\| + \frac{1-\rho}{2}\|\mathbf{w}\|^2 - \frac{1}{\lambda n}\langle\mathbf{A}\boldsymbol{\alpha},\mathbf{w}\rangle\right\} \\ = -\lambda\sup_{\mathbf{w}}\left\{\frac{1}{\lambda n}\langle\mathbf{A}\boldsymbol{\alpha},\mathbf{w}\rangle - \bigg(\rho\sum_{j=1}^{d}\|\mathbf{w}_{\mathscr{G}_j}\| + \frac{1-\rho}{2}\|\mathbf{w}\|^2\bigg)\right\} = -\lambda g^*\Big(\frac{\mathbf{A}\boldsymbol{\alpha}}{\lambda n}\Big).$$

Regarding the explicit form of $g^*\big(\frac{\mathbf{A}\boldsymbol{\alpha}}{\lambda n}\big)$, it can be shown that

$$g^*\Big(\frac{\mathbf{A}\boldsymbol{\alpha}}{\lambda n}\Big) = \inf_{\mathbf{w}}\left\{\rho\sum_{j=1}^{d}\|\mathbf{w}_{\mathscr{G}_j}\| + \frac{1-\rho}{2}\|\mathbf{w}\|^2 - \frac{1}{\lambda n}\langle\mathbf{A}\boldsymbol{\alpha},\mathbf{w}\rangle\right\}.$$

The optimality condition of the above problem implies

$$\mathbf{0} \in (1-\rho)\mathbf{w}_{\mathscr{G}_j} + \rho\partial\|\mathbf{w}_{\mathscr{G}_j}\| - \frac{1}{\lambda n}\mathbf{A}_{\mathscr{G}_j\cdot}\boldsymbol{\alpha},\ j\in[d]. \tag{21}$$

The definition of subgradient implies

$$\mathbf{w}_{\mathscr{G}_j} = \mathbf{0}\ \text{ if }\ \frac{1}{\lambda n}\|\mathbf{A}_{\mathscr{G}_j\cdot}\boldsymbol{\alpha}\| < \rho.$$

Otherwise, we have

$$\mathbf{0} = (1-\rho)\mathbf{w}_{\mathscr{G}_j} + \rho\frac{\mathbf{w}_{\mathscr{G}_j}}{\|\mathbf{w}_{\mathscr{G}_j}\|} - \frac{1}{\lambda n}\mathbf{A}_{\mathscr{G}_j\cdot}\boldsymbol{\alpha}.$$

which implies

$$\|\mathbf{w}_{\mathscr{G}_j}\| = \frac{\|\mathbf{A}_{\mathscr{G}_j\cdot}\boldsymbol{\alpha}\| - \rho\lambda n}{(1-\rho)\lambda n}\ \text{ and }\ \mathbf{w}_{\mathscr{G}_j} = \frac{\|\mathbf{A}_{\mathscr{G}_j\cdot}\boldsymbol{\alpha}\| - \rho\lambda n}{(1-\rho)\|\mathbf{A}_{\mathscr{G}_j\cdot}\boldsymbol{\alpha}\|}\frac{1}{\lambda n}\mathbf{A}_{\mathscr{G}_j\cdot}\boldsymbol{\alpha}.$$

Combining these two cases together, we obtain

$$\mathbf{w}_{\mathscr{G}_j} = \frac{\big[\|\mathbf{A}_{\mathscr{G}_j\cdot}\boldsymbol{\alpha}\| - \rho\lambda n\big]_+}{(1-\rho)\|\mathbf{A}_{\mathscr{G}_j\cdot}\boldsymbol{\alpha}\|}\frac{1}{\lambda n}\mathbf{A}_{\mathscr{G}_j\cdot}\boldsymbol{\alpha},\ \forall j\in[d]. \tag{22}$$

Then, the conjugate of $g(\mathbf{w})$ is

$$g^*\Big(\frac{\mathbf{A}\boldsymbol{\alpha}}{\lambda n}\Big) = \sum_{j=1}^{d} \rho\|\mathbf{w}_{\mathcal{G}_j}\| + \frac{1-\rho}{2}\|\mathbf{w}\|^2 - \frac{1}{\lambda n}\langle \boldsymbol{\alpha}, \mathbf{X}\mathbf{w}\rangle = \sum_{j=1}^{d} \frac{\big[\|\mathbf{A}_{\mathcal{G}_j}.\boldsymbol{\alpha}\| - \rho\lambda n\big]_+^2}{2(1-\rho)\lambda^2 n^2}.$$

Therefore, the dual problem of (2) is

$$\max_{\boldsymbol{\alpha}} -\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_K} f_{ki}^*(-\alpha_i^k) - \lambda\sum_{j=1}^{d}\frac{\big[\|\mathbf{A}_{\mathcal{G}_j}.\boldsymbol{\alpha}\| - \rho\lambda n\big]_+^2}{2(1-\rho)\lambda^2 n^2}.$$

Let $\mathbf{w}_\star$ and $\boldsymbol{\alpha}_\star$ denote the primal and dual optimal solutions, respectively. From (20) and (21) the KKT condition of (2) establishes

$$\begin{aligned}&\big(\alpha_\star\big)_i^k \in \partial f_{ki}\big(\langle \mathbf{x}_i^k, \mathbf{w}_\star^k\rangle\big), \forall k \in [K], i \in [n_k] \text{ and}\\ &\frac{1}{\lambda n}\mathbf{A}_{\mathcal{G}_j}.\boldsymbol{\alpha}_\star \in (1-\rho)\mathbf{w}_{\star\mathcal{G}_j} + \rho\partial\|\mathbf{w}_{\star\mathcal{G}_j}\|, j \in [d].\end{aligned}$$

Appendix B: Convergence analysis

To facilitate the proof, we first introduce some useful notations and technical Lemmas. It is easy to verify that $\widehat{\mathbf{u}}_t^k$ can be rewritten as

$$\widehat{\mathbf{u}}_t^k = \widehat{\boldsymbol{\alpha}}_{t-1}^k + \frac{\theta_t\zeta_{t-1}}{\zeta_{t-1}+\vartheta\theta_t}\big(\widehat{\mathbf{v}}_{t-1}^k - \widehat{\boldsymbol{\alpha}}_{t-1}^k\big), \quad k \in [K]. \tag{23}$$

For any $t \geq 0$, we define $\boldsymbol{\beta}_t$ as $\boldsymbol{\beta}_t \overset{\text{def}}{=} \big(\mathbf{u}_t - \boldsymbol{\alpha}_t\big)/\eta \Rightarrow \boldsymbol{\alpha}_t = \mathbf{u}_t - \eta\boldsymbol{\beta}_t \;\; \forall k \in [K]$.

Lemma 6 (Dünner et al. 2016) *Consider the following pair of optimization problems, which are dual to each other:*

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n}\big\{D(\boldsymbol{\alpha}) \overset{\text{def}}{=} f^*(-\boldsymbol{\alpha}) + g^*(\mathbf{A}\boldsymbol{\alpha})\big\} \;\; \textit{and} \;\; \min_{\mathbf{w}\in\mathbb{R}^d}\big\{P(\mathbf{w}) \overset{\text{def}}{=} f(\mathbf{A}\mathbf{w}) + g(\mathbf{w})\big\},$$

where f^* is μ-strongly convex with respect to a norm $\|\cdot\|_{f^*}$ and g^* is $1/\beta$-smooth with respect to a norm $\|\cdot\|_{g^*}$. Let $\sigma_{max} = \max_{\boldsymbol{\alpha}\neq 0}\|\mathbf{A}\boldsymbol{\alpha}\|_{g^*}^2/\|\boldsymbol{\alpha}\|_{f^*}^2$. Suppose an arbitrary optimization algorithm is applied to the first problem and it produces a sequence of (possibly random) iterates $\{\boldsymbol{\alpha}_t\}_{t=0}^{\infty}$ such that there exits $C \in (0, 1]$, $D \geq 0$ such that

$$\mathbb{E}\big[D(\boldsymbol{\alpha}_t) - D(\boldsymbol{\alpha}_\star)\big] \leq (1-C)^t D.$$

Then, for any

$$t \geq \frac{1}{C}\log\frac{D(\sigma_{max}+\mu\beta)}{\mu\beta\epsilon},$$

it holds that $\mathbb{E}\big[P(\mathbf{w}(\boldsymbol{\alpha}_t)) - (-D(\boldsymbol{\alpha}_t))\big] \leq \epsilon$.

Remark 3 This lemma enables us transfer the convergence rate of objective function to the convergence rate of duality gap.

Lemma 7 *For any $t \geq 1$, the following identities hold*

$$\frac{\theta_t \zeta_{t-1}}{\zeta_t}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_{t-1}^k\big) = \big(\widehat{\boldsymbol{\alpha}}_{t-1}^k - \widehat{\mathbf{u}}_t^k\big) \tag{24}$$

$$\zeta_t \widehat{\mathbf{v}}_t^k = (1-\theta_t)\zeta_{t-1}\widehat{\mathbf{v}}_{t-1}^k + \vartheta\theta_t \widehat{\mathbf{u}}_t^k - \theta_t \widehat{\boldsymbol{\beta}}_t^k \tag{25}$$

$$\begin{aligned}&\frac{\zeta_t}{2}\Big(\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_t^k\big)\big\|^2 - \big\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_t^k\big)\big\|^2\Big)\\ &\quad - \frac{(1-\theta_t)\zeta_{t-1}}{2}\Big(\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2 - \big\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2\Big)\\ &= \frac{\vartheta\theta_t}{2}\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big)\big\|^2 + \theta_t\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big), \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\rangle \end{aligned}\tag{26}$$

$$\begin{aligned}&\frac{\zeta_t}{2}\big\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_t^k\big)\big\|^2 - (1-\theta_t)\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_{t-1}^k - \widehat{\mathbf{u}}_t^k\big), \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\rangle - \frac{\eta}{2}\big\|\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\|^2\\ &= \frac{(1-\theta_t)\zeta_{t-1}}{2}\Big(1 - \frac{\vartheta\theta_t}{\zeta_t}\Big)\big\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2. \end{aligned}\tag{27}$$

Proof First, we show that (24) can be proved by using the definition of $\widehat{\mathbf{u}}_t^k$ and ζ_t.

$$\begin{aligned}&(\zeta_{t-1} + \vartheta\theta_t)\widehat{\mathbf{u}}_t^k = (\zeta_{t-1} + \vartheta\theta_t)\widehat{\boldsymbol{\alpha}}_{t-1}^k + \theta_t\zeta_{t-1}\big(\widehat{\mathbf{v}}_{t-1}^k - \widehat{\boldsymbol{\alpha}}_{t-1}^k\big)\\ &\quad\Rightarrow\ \big((1-\theta_t)\zeta_{t-1} + \vartheta\theta_t\big)\widehat{\mathbf{u}}_t^k + \theta_t\zeta_{t-1}\widehat{\mathbf{u}}_t^k = \big((1-\theta_t)\zeta_{t-1} + \vartheta\theta_t\big)\widehat{\boldsymbol{\alpha}}_{t-1}^k + \theta_t\zeta_{t-1}\widehat{\mathbf{v}}_{t-1}^k\\ &\quad\Rightarrow\ \theta_t\zeta_{t-1}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_{t-1}^k\big) = \zeta_t\big(\widehat{\boldsymbol{\alpha}}_{t-1}^k - \widehat{\mathbf{u}}_t^k\big),\end{aligned}$$

which implies $\theta_t\zeta_{t-1}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_{t-1}^k\big)/\zeta_t = \big(\widehat{\boldsymbol{\alpha}}_{t-1}^k - \widehat{\mathbf{u}}_t^k\big)$. Next, (25) can be shown by using $\widehat{\boldsymbol{\beta}}_t^k$ and $\zeta_t = \theta_t^2/\eta$. Following from the definition of $\widehat{\mathbf{v}}_t^k$, we have

$$\widehat{\mathbf{v}}_t^k = \widehat{\boldsymbol{\alpha}}_{t-1}^k + \frac{1}{\theta_t}\big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\boldsymbol{\alpha}}_{t-1}^k\big) = \widehat{\boldsymbol{\alpha}}_{t-1}^k + \frac{1}{\theta_t}\big(\widehat{\mathbf{u}}_t^k - \eta\widehat{\boldsymbol{\beta}}_t^k - \widehat{\boldsymbol{\alpha}}_{t-1}^k\big) = \frac{1}{\theta_t}\big(\widehat{\mathbf{u}}_t^k - (1-\theta_t)\widehat{\boldsymbol{\alpha}}_{t-1}^k\big) - \frac{\theta_t}{\zeta_t}\widehat{\boldsymbol{\beta}}_t^k,$$

which implies

$$\begin{aligned}\zeta_t\widehat{\mathbf{v}}_t^k &= \frac{1}{\theta_t}\big(\zeta_t\widehat{\mathbf{u}}_t^k - \zeta_t(1-\theta_t)\widehat{\boldsymbol{\alpha}}_{t-1}^k\big) - \theta_t\widehat{\boldsymbol{\beta}}_t^k\\ &= \frac{1-\theta_t}{\theta_t}\Big(\frac{\zeta_t}{1-\theta_t}\widehat{\mathbf{u}}_t^k - \zeta_t\widehat{\boldsymbol{\alpha}}_{t-1}^k\Big) - \theta_t\widehat{\boldsymbol{\beta}}_t^k\\ &= \frac{1-\theta_t}{\theta_t}\Big(\frac{(1-\theta_t)\zeta_{t-1} + \vartheta\theta_t}{1-\theta_t}\widehat{\mathbf{u}}_t^k - \zeta_t\widehat{\boldsymbol{\alpha}}_{t-1}^k\Big) - \theta_t\widehat{\boldsymbol{\beta}}_t^k\\ &= \frac{1-\theta_t}{\theta_t}\Big((\zeta_{t-1} + \vartheta\theta_t)\widehat{\mathbf{u}}_t^k - \zeta_t\widehat{\boldsymbol{\alpha}}_{t-1}^k\Big) - \frac{1-\theta_t}{\theta_t}\Big(\vartheta\theta_t - \frac{\vartheta\theta_t}{1-\theta_t}\Big)\widehat{\mathbf{u}}_t^k - \theta_t\widehat{\boldsymbol{\beta}}_t^k\\ &= (1-\theta_t)\zeta_{t-1}\widehat{\mathbf{v}}_{t-1}^k + \vartheta\theta_t\widehat{\mathbf{u}}_t^k - \theta_t\widehat{\boldsymbol{\beta}}_t^k.\end{aligned}$$

To proved (26), we need to use (24) and (25). By using the definition of ζ_t and (25), one can show that

$$\begin{aligned}&\frac{\zeta_t}{2}\Big(\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_t^k\big)\big\|^2 - \big\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_t^k\big)\big\|^2\Big)\\ &\quad = \frac{1}{2\zeta_t}\Big(\big\|\mathbf{A}\big(\zeta_t\widehat{\boldsymbol{\alpha}}_\star^k - \zeta_t\widehat{\mathbf{v}}_t^k\big)\big\|^2 - \big\|\mathbf{A}\big(\zeta_t\widehat{\mathbf{u}}_t^k - \zeta_t\widehat{\mathbf{v}}_t^k\big)\big\|^2\Big)\\ &\quad = \frac{1}{2\zeta_t}\Big(\Big\|\mathbf{A}\Big(\big((1-\theta_t)\zeta_{t-1} + \vartheta\theta_t\big)\widehat{\boldsymbol{\alpha}}_\star^k - \big((1-\theta_t)\zeta_{t-1}\widehat{\mathbf{v}}_{t-1}^k + \vartheta\theta_t\widehat{\mathbf{u}}_t^k - \theta_t\widehat{\boldsymbol{\beta}}_t^k\big)\Big)\Big\|^2\end{aligned}$$

$$
\begin{aligned}
&- \Big\| \mathbf{A}\Big(\big((1-\theta_t)\zeta_{t-1} + \vartheta\theta_t\big)\widehat{\mathbf{u}}_t^k - \big((1-\theta_t)\zeta_{t-1}\widehat{\mathbf{v}}_{t-1}^k + \vartheta\theta_t\widehat{\mathbf{u}}_t^k - \theta_t\widehat{\boldsymbol{\beta}}_t^k\big)\Big)\Big\|^2\Big) \\
&= \frac{1}{2\zeta_t}\Big(\Big\| (1-\theta_t)\zeta_{t-1}\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_{t-1}^k\big) + \vartheta\theta_t\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big) + \theta_t\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k \Big\|^2 \\
&- \Big\| (1-\theta_t)\zeta_{t-1}\mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_{t-1}^k\big) + \theta_t\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k \Big\|^2\Big) \\
&= \frac{1}{2\zeta_t}\Big((1-\theta_t)^2\zeta_{t-1}^2 \big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2 + \vartheta^2\theta_t^2 \big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big)\big\|^2 \\
&- (1-\theta_t)^2\zeta_{t-1}^2 \big\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2 \\
&+ 2\vartheta\theta_t(1-\theta_t)\zeta_{t-1}\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_{t-1}^k\big), \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big)\big\rangle \\
&+ 2\theta_t(1-\theta_t)\zeta_{t-1}\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_{t-1}^k\big), \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\rangle \\
&+ 2\vartheta\theta_t^2\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big), \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\rangle - 2\theta_t(1-\theta_t)\zeta_{t-1}\big\langle \mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_{t-1}^k\big), \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\rangle\Big) \\
&= \frac{1}{2\zeta_t}\big(1-\theta_t\big)\zeta_{t-1}\big(\zeta_t - \vartheta\theta_t\big)\Big(\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2 - \big\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2\Big) \\
&+ \frac{1}{2\zeta_t}\vartheta\theta_t\big(\zeta_t - (1-\theta_t)\zeta_{t-1}\big)\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big)\big\|^2 \\
&+ \frac{1}{\zeta_t}\vartheta\theta_t(1-\theta_t)\zeta_{t-1}\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_{t-1}^k\big), \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big)\big\rangle \\
&+ \frac{1}{\zeta_t}\theta_t(1-\theta_t)\zeta_{t-1}\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big), \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\rangle + \frac{1}{\zeta_t}\vartheta\theta_t^2\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big), \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\rangle \\
&= \frac{(1-\theta_t)\zeta_{t-1}}{2}\Big(\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2 - \big\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2\Big) + \frac{\vartheta\theta_t}{2}\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big)\big\|^2 \\
&- \frac{\vartheta\theta_t(1-\theta_t)\zeta_{t-1}}{2\zeta_t}\Big(\Big\langle \mathbf{A}\Big(\big(\widehat{\boldsymbol{\alpha}}_\star^k + \widehat{\mathbf{u}}_t^k - 2\widehat{\mathbf{v}}_{t-1}^k\big) + \big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big) \\
&- 2\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_{t-1}^k\big)\Big), \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big)\Big\rangle\Big) \\
&+ \frac{\theta_t}{\zeta_t}\big((1-\theta_t)\zeta_{t-1} + \vartheta\theta_t\big)\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big), \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\rangle \\
&= \frac{(1-\theta_t)\zeta_{t-1}}{2}\Big(\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2 - \big\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2\Big) \\
&+ \frac{\vartheta\theta_t}{2}\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big)\big\|^2 + \theta_t\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big), \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\rangle,
\end{aligned}
$$

which can be rewritten as

$$
\begin{aligned}
&\frac{\zeta_t}{2}\Big(\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_t^k\big)\big\|^2 - \big\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_t^k\big)\big\|^2\Big) \\
&\quad - \frac{(1-\theta_t)\zeta_{t-1}}{2}\Big(\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2 - \big\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2\Big) \\
&\quad = \frac{\vartheta\theta_t}{2}\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big)\big\|^2 + \theta_t\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big), \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\rangle.
\end{aligned}
$$

Finally, we prove (27) by using (24) and (25).

$$\begin{aligned}
\frac{\zeta_t}{2}\left\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k-\widehat{\mathbf{v}}_t^k\big)\right\|^2 &= \frac{\zeta_t}{2}\left\|\mathbf{A}\big(\zeta_t\widehat{\mathbf{u}}_t^k-\zeta_t\widehat{\mathbf{v}}_t^k\big)\right\|^2\\
&= \frac{1}{2\zeta_t}\Big\|\mathbf{A}\Big(\big((1-\theta_t)\zeta_{t-1}+\vartheta\theta_t\big)\widehat{\mathbf{u}}_t^k\\
&\quad -(1-\theta_t)\zeta_{t-1}\widehat{\mathbf{v}}_{t-1}^k-\vartheta\theta_t\widehat{\mathbf{u}}_t^k+\theta_t\widehat{\boldsymbol{\beta}}_t^k\Big)\Big\|^2\\
&= \frac{1}{2\zeta_t}\left\|\mathbf{A}\big((1-\theta_t)\zeta_{t-1}\big(\widehat{\mathbf{u}}_t^k-\widehat{\mathbf{v}}_{t-1}^k\big)+\theta_t\widehat{\boldsymbol{\beta}}_t^k\big)\right\|^2\\
&= \frac{(1-\theta_t)^2\zeta_{t-1}^2}{2\zeta_t}\left\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k-\widehat{\mathbf{v}}_{t-1}^k\big)\right\|^2\\
&\quad +\frac{\theta_t(1-\theta_t)\zeta_{t-1}}{\zeta_t}\big\langle\mathbf{A}\big(\widehat{\mathbf{u}}_t^k-\widehat{\mathbf{v}}_{t-1}^k\big),\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\rangle+\frac{\theta_t^2}{2\zeta_t}\left\|\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\right\|^2.
\end{aligned}$$

By using (24), we obtain

$$\begin{aligned}
&\frac{\zeta_t}{2}\left\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k-\widehat{\mathbf{v}}_t^k\big)\right\|\\
&= (1-\theta_t)\frac{\zeta_{t-1}(\zeta_t-\vartheta\theta_t)}{2\zeta_t}\left\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k-\widehat{\mathbf{v}}_{t-1}^k\big)\right\|^2+(1-\theta_t)\big\langle\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_{t-1}^k-\widehat{\mathbf{u}}_t^k\big),\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\rangle\\
&\quad+\frac{\eta}{2}\left\|\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\right\|^2\\
&= (1-\theta_t)\frac{\zeta_{t-1}}{2}\Big(1-\frac{\vartheta\theta_t}{\zeta_t}\Big)\left\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k-\widehat{\mathbf{v}}_{t-1}^k\big)\right\|^2\\
&\quad+(1-\theta_t)\big\langle\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_{t-1}^k-\widehat{\mathbf{u}}_t^k\big),\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\rangle+\frac{\eta}{2}\left\|\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\right\|^2.
\end{aligned}$$

This completes the proof. □

B.1 Proof of Lemma 2

Lemma 2 *Consider applying Algorithm 1 to solve (3), the following inequality holds for any $t \geq 1$,*

$$\epsilon_D^t + R^t \leq \gamma_t\big(\epsilon_D^0 + R^0\big), \tag{11}$$

where $R^t = \frac{\zeta_t}{2}\sum_{k=1}^K \left\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_t^k\big)\right\|^2$, $\gamma_t = \prod_{i=1}^t \big(1-\theta_i\big)$ *for any* $t \geq 1$ *and* $\gamma_0 = 1$.

Proof Following from the optimality condition of $\widehat{\boldsymbol{\alpha}}_t^k$, the following holds for any $k \in [K]$

$$\begin{aligned}
&\mathbf{0} \in L_k\big(\widehat{\boldsymbol{\alpha}}_t^k;\widehat{\mathbf{u}}_t^k,\mathbf{w}(\mathbf{u}_t)\big)\\
&\Rightarrow\ \mathbf{0} \in -\frac{1}{n}\partial f_k^*\big(-\widehat{\boldsymbol{\alpha}}_t^k\big)+\frac{1}{n}\big(\widehat{\mathbf{A}}^k\big)^\top\nabla g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big)+\frac{1}{\eta}\big(\widehat{\mathbf{A}}^k\big)^\top\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_t^k-\widehat{\mathbf{u}}_t^k\big)\\
&\Rightarrow -\frac{1}{n}\big(\widehat{\mathbf{A}}^k\big)^\top\nabla g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big)-\frac{1}{\eta}\big(\widehat{\mathbf{A}}^k\big)^\top\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_t^k-\widehat{\mathbf{u}}_t^k\big)\in-\frac{1}{n}\partial f_k^*\big(-\widehat{\boldsymbol{\alpha}}_t^k\big).
\end{aligned}\tag{28}$$

By using the fact f^* is μ-strongly convex, the following inequality holds for any $\mathbf{z} \in \mathbb{R}^n$

$$\begin{aligned}\frac{1}{n}f^*(-\boldsymbol{\alpha}_t) &\le \frac{1}{n}f^*(-\mathbf{z}) - \frac{1}{n}\langle \partial f^*(-\boldsymbol{\alpha}_t), \boldsymbol{\alpha}_t - \mathbf{z}\rangle - \frac{\mu}{2n}\|\mathbf{z}-\boldsymbol{\alpha}_t\|^2 \\ &= \frac{1}{n}f^*(-\mathbf{z}) - \frac{1}{n}\sum_{k=1}^{K}\big\langle \partial f_k^*\big(-\widehat{\boldsymbol{\alpha}}_t^k\big), \widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{z}}^k\big\rangle - \frac{\vartheta}{2}\|\mathbf{z}-\boldsymbol{\alpha}_t\|^2.\end{aligned}$$

Substituting (28) into the above inequality, we obtain

$$\begin{aligned}\frac{1}{n}f^*(-\boldsymbol{\alpha}_t) \le\ & \frac{1}{n}f^*(-\mathbf{z}) - \frac{1}{n}\sum_{k=1}^{K}\Big\langle \big(\widehat{\mathbf{A}}^k\big)^\top \nabla g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big), \big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{z}}^k\big)\Big\rangle \\ & - \frac{1}{\eta}\sum_{k=1}^{K}\big\langle \big(\widehat{\mathbf{A}}^k\big)^\top \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{u}}_t^k\big), \big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{z}}^k\big)\big\rangle - \frac{\vartheta}{2}\|\mathbf{z}-\boldsymbol{\alpha}_t\|^2 \\ & = \frac{1}{n}f^*(-\mathbf{z}) - \frac{1}{n}\sum_{k=1}^{K}\Big\langle \nabla g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big), \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{z}}^k\big)\Big\rangle - \frac{1}{\eta}\sum_{k=1}^{K}\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{u}}_t^k\big), \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{z}}^k\big)\big\rangle \\ & - \frac{\vartheta}{2}\|\mathbf{z}-\boldsymbol{\alpha}_t\|^2 \\ & = \frac{1}{n}f^*(-\mathbf{z}) - \frac{1}{n}\Big\langle \nabla g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big), \mathbf{A}\big(\boldsymbol{\alpha}_t - \mathbf{z}\big)\Big\rangle - \frac{1}{\eta}\sum_{k=1}^{K}\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{u}}_t^k\big), \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{z}}^k\big)\big\rangle \\ & - \frac{\vartheta}{2}\|\mathbf{z}-\boldsymbol{\alpha}_t\|^2.\end{aligned}$$

By using the fact that g^* is $1/(1-\rho)$-smooth and convex, the following inequality holds for any $\mathbf{z} \in \mathbb{R}^n$

$$\begin{aligned}&\lambda g^*\Big(\frac{\mathbf{A}\boldsymbol{\alpha}_t}{\lambda n}\Big) \\ &\le \lambda g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big) + \lambda\Big\langle \nabla g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big), \frac{\mathbf{A}(\boldsymbol{\alpha}_t - \mathbf{u}_t)}{\lambda n}\Big\rangle + \frac{\lambda}{2(1-\rho)}\Big\|\frac{\mathbf{A}(\boldsymbol{\alpha}_t - \mathbf{u}_t)}{\lambda n}\Big\|^2 \\ &\le \lambda g^*\Big(\frac{\mathbf{A}\mathbf{z}}{\lambda n}\Big) - \lambda\Big\langle \nabla g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big), \frac{\mathbf{A}(\mathbf{z} - \mathbf{u}_t)}{\lambda n}\Big\rangle \\ &\quad + \lambda\Big\langle \nabla g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big), \frac{\mathbf{A}(\boldsymbol{\alpha}_t - \mathbf{u}_t)}{\lambda n}\Big\rangle + \frac{\lambda}{2(1-\rho)}\Big\|\frac{\mathbf{A}(\boldsymbol{\alpha}_t - \mathbf{u}_t)}{\lambda n}\Big\|^2 \\ &= \lambda g^*\Big(\frac{\mathbf{A}\mathbf{z}}{\lambda n}\Big) - \frac{1}{n}\Big\langle \nabla g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big), \mathbf{A}(\mathbf{z} - \boldsymbol{\alpha}_t)\Big\rangle + \frac{\lambda}{2(1-\rho)}\Big\|\frac{\mathbf{A}(\boldsymbol{\alpha}_t - \mathbf{u}_t)}{\lambda n}\Big\|^2 \\ &\le \lambda g^*\Big(\frac{\mathbf{A}\mathbf{z}}{\lambda n}\Big) - \frac{1}{n}\Big\langle \nabla g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big), \mathbf{A}(\mathbf{z} - \boldsymbol{\alpha}_t)\Big\rangle + \frac{\lambda}{2(1-\rho)\lambda^2 n^2}\sum_{k=1}^{K}\big\|\mathbf{A}(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{u}}_t^k)\big\|^2 \\ &= \lambda g^*\Big(\frac{\mathbf{A}\mathbf{z}}{\lambda n}\Big) - \frac{1}{n}\Big\langle \nabla g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big), \mathbf{A}(\mathbf{z} - \boldsymbol{\alpha}_t)\Big\rangle + \frac{1}{2\eta}\sum_{k=1}^{K}\big\|\mathbf{A}(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{u}}_t^k)\big\|^2,\end{aligned}$$

where the last inequality is obtained by using the fact that $\mathbf{A}$ is a block diagonal matrix. Thus,

$$
\begin{aligned}
D(\boldsymbol{\alpha}_t) &= \frac{1}{n} f^*(-\boldsymbol{\alpha}_t) + \lambda g^*\Big(\frac{\mathbf{A}\boldsymbol{\alpha}_t}{\lambda n}\Big) \\
&\le \frac{1}{n} f^*(-\mathbf{z}) - \frac{1}{n}\Big\langle \nabla g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big), \mathbf{A}\big(\boldsymbol{\alpha}_t - \mathbf{z}\big)\Big\rangle - \frac{1}{\eta}\sum_{k=1}^{K}\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{u}}_t^k\big), \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{z}}^k\big)\big\rangle \\
&\quad - \frac{\vartheta}{2}\|\mathbf{z} - \boldsymbol{\alpha}_t\|^2 + \lambda g^*\Big(\frac{\mathbf{A}\mathbf{z}}{\lambda n}\Big) - \frac{1}{n}\Big\langle \nabla g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big), \mathbf{A}(\mathbf{z} - \boldsymbol{\alpha}_t)\Big\rangle \\
&\quad + \frac{1}{2\eta}\sum_{k=1}^{K}\big\|\mathbf{A}(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{u}}_t^k)\big\|^2 \\
&= D(\mathbf{z}) - \frac{1}{\eta}\sum_{k=1}^{K}\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{u}}_t^k\big), \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{u}}_t^k + \widehat{\mathbf{u}}_t^k - \widehat{\mathbf{z}}^k\big)\big\rangle \\
&\quad + \frac{1}{2\eta}\sum_{k=1}^{K}\big\|\mathbf{A}(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{u}}_t^k)\big\|^2 - \frac{\vartheta}{2}\|\mathbf{z} - \boldsymbol{\alpha}_t\|^2 \\
&= D(\mathbf{z}) - \frac{1}{\eta}\sum_{k=1}^{K}\big\langle \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{u}}_t^k\big), \mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{z}}^k\big)\big\rangle - \frac{1}{2\eta}\sum_{k=1}^{K}\big\|\mathbf{A}(\widehat{\boldsymbol{\alpha}}_t^k - \widehat{\mathbf{u}}_t^k)\big\|^2 \\
&\quad - \frac{\vartheta}{2}\|\mathbf{z} - \boldsymbol{\alpha}_t\|^2 \\
&= D(\mathbf{z}) - \sum_{k=1}^{K}\big\langle \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k, \mathbf{A}\big(\widehat{\mathbf{z}}^k - \widehat{\mathbf{u}}_t^k\big)\big\rangle - \frac{\eta}{2}\sum_{k=1}^{K}\big\|\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\|^2 - \frac{\vartheta}{2}\|\mathbf{z} - \boldsymbol{\alpha}_t\|^2,
\end{aligned}
$$

which implies

$$
D(\mathbf{z}) \ge D(\boldsymbol{\alpha}_t) + \sum_{k=1}^{K}\big\langle \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k, \mathbf{A}\big(\widehat{\mathbf{z}}^k - \widehat{\mathbf{u}}_t^k\big)\big\rangle + \frac{\eta}{2}\sum_{k=1}^{K}\big\|\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\|^2 + \frac{\vartheta}{2}\|\mathbf{z} - \boldsymbol{\alpha}_t\|^2. \tag{29}
$$

Substituting $\mathbf{z} = \boldsymbol{\alpha}_\star$ and $\mathbf{z} = \boldsymbol{\alpha}_{t-1}$ into (29), we obtain

$$
\begin{aligned}
D(\boldsymbol{\alpha}_\star) &\ge D(\boldsymbol{\alpha}_t) + \sum_{k=1}^{K}\big\langle \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k, \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big)\big\rangle + \frac{\eta}{2}\sum_{k=1}^{K}\big\|\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\|^2 + \frac{\vartheta}{2}\|\boldsymbol{\alpha}_\star - \boldsymbol{\alpha}_t\|^2 \\
D(\boldsymbol{\alpha}_{t-1}) &\ge D(\boldsymbol{\alpha}_t) + \sum_{k=1}^{K}\big\langle \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k, \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_{t-1}^k - \widehat{\mathbf{u}}_t^k\big)\big\rangle + \frac{\eta}{2}\sum_{k=1}^{K}\big\|\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\|^2 + \frac{\vartheta}{2}\|\boldsymbol{\alpha}_{t-1} - \boldsymbol{\alpha}_t\|^2.
\end{aligned}
$$

Combining these two inequalities together with coefficients θ_t and $(1-\theta_t)$, respectively, we obtain

$$
\begin{aligned}
&\theta_t D(\boldsymbol{\alpha}_\star) + (1-\theta_t) D(\boldsymbol{\alpha}_{t-1}) \\
&\quad \ge D(\boldsymbol{\alpha}_t) + \sum_{k=1}^{K}\big\langle \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k, \mathbf{A}\big(\theta_t\widehat{\boldsymbol{\alpha}}_\star^k + (1-\theta_t)\widehat{\boldsymbol{\alpha}}_{t-1}^k - \widehat{\mathbf{u}}_t^k\big)\big\rangle + \frac{\eta}{2}\sum_{k=1}^{K}\big\|\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\big\|^2 \\
&\qquad + \frac{\vartheta\theta_t}{2}\|\boldsymbol{\alpha}_\star - \boldsymbol{\alpha}_t\|^2 + \frac{\vartheta(1-\theta_t)}{2}\|\boldsymbol{\alpha}_{t-1} - \boldsymbol{\alpha}_t\|^2,
\end{aligned}
$$

which is equivalent to

$$\begin{aligned}
D(\boldsymbol{\alpha}_t) - D(\boldsymbol{\alpha}_\star) \leq & (1-\theta_t)\big(D(\boldsymbol{\alpha}_{t-1}) - D(\boldsymbol{\alpha}_\star)\big) - \sum_{k=1}^{K} \big\langle \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k, \mathbf{A}\big(\theta_t \widehat{\boldsymbol{\alpha}}_\star^k + (1-\theta_t)\widehat{\boldsymbol{\alpha}}_{t-1}^k - \widehat{\mathbf{u}}_t^k\big)\big\rangle \\
& - \frac{\eta}{2}\sum_{k=1}^{K} \|\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\|^2 - \frac{\vartheta\theta_t}{2}\|\boldsymbol{\alpha}_\star - \boldsymbol{\alpha}_t\|^2 - \frac{\vartheta(1-\theta_t)}{2}\|\boldsymbol{\alpha}_{t-1} - \boldsymbol{\alpha}_t\|^2 \\
& = (1-\theta_t)\big(D(\boldsymbol{\alpha}_{t-1}) - D(\boldsymbol{\alpha}_\star)\big) - (1-\theta_t)\sum_{k=1}^{K} \big\langle \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k, \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_{t-1}^k - \widehat{\mathbf{u}}_t^k\big)\big\rangle \\
& - \theta_t \sum_{k=1}^{K} \big\langle \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k, \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big)\big\rangle \\
& - \frac{\eta}{2}\sum_{k=1}^{K} \|\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\|^2 - \frac{\vartheta\theta_t}{2}\|\boldsymbol{\alpha}_\star - \boldsymbol{\alpha}_t\|^2 - \frac{\vartheta(1-\theta_t)}{2}\|\boldsymbol{\alpha}_{t-1} - \boldsymbol{\alpha}_t\|^2.
\end{aligned}$$

Substituting (26) into the above inequality, we obtain

$$\begin{aligned}
& D(\boldsymbol{\alpha}_t) - D(\boldsymbol{\alpha}_\star) \\
& \quad \leq (1-\theta_t)\big(D(\boldsymbol{\alpha}_{t-1}) - D(\boldsymbol{\alpha}_\star)\big) - (1-\theta_t)\sum_{k=1}^{K} \big\langle \mathbf{A}\widehat{\boldsymbol{\beta}}_t^k, \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_{t-1}^k - \widehat{\mathbf{u}}_t^k\big)\big\rangle \\
& \quad - \frac{\vartheta\theta_t}{2}\sum_{k=1}^{K} \big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{u}}_t^k\big)\big\|^2 \\
& \quad - \frac{\zeta_t}{2}\sum_{k=1}^{K}\Big(\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_t^k\big)\big\|^2 - \big\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_t^k\big)\big\|^2\Big) - \frac{\eta}{2}\sum_{k=1}^{K}\|\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\|^2 \\
& \quad - \frac{\vartheta(1-\theta_t)}{2}\|\boldsymbol{\alpha}_{t-1} - \boldsymbol{\alpha}_t\|^2 \\
& \quad + \frac{(1-\theta_t)\zeta_{t-1}}{2}\sum_{k=1}^{K}\Big(\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2 - \big\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2\Big) - \frac{\vartheta\theta_t}{2}\|\boldsymbol{\alpha}_\star - \boldsymbol{\alpha}_t\|^2,
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
& \big(D(\boldsymbol{\alpha}_t) - D(\boldsymbol{\alpha}_\star)\big) + \frac{\zeta_t}{2}\sum_{k=1}^{K}\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_t^k\big)\big\|^2 \\
& \quad \leq (1-\theta_t)\Bigg(D(\boldsymbol{\alpha}_{t-1}) - D(\boldsymbol{\alpha}_\star) + \frac{\zeta_{t-1}}{2}\sum_{k=1}^{K}\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2 \\
& \quad -\frac{\zeta_{t-1}}{2}\sum_{k=1}^{K}\big\|\mathbf{A}\big(\widehat{\mathbf{u}}_t^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2\Bigg)
\end{aligned}$$

$$+\frac{\zeta_t}{2}\sum_{k=1}^{K}\left\|\mathbf{A}\left(\widehat{\mathbf{u}}_t^k-\widehat{\mathbf{v}}_t^k\right)\right\|^2-(1-\theta_t)\sum_{k=1}^{K}\left\langle\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k,\mathbf{A}\left(\widehat{\boldsymbol{\alpha}}_{t-1}^k-\widehat{\mathbf{u}}_t^k\right)\right\rangle-\frac{\eta}{2}\sum_{k=1}^{K}\left\|\mathbf{A}\widehat{\boldsymbol{\beta}}_t^k\right\|^2$$
$$-\frac{\vartheta\theta_t}{2}\sum_{k=1}^{K}\left\|\mathbf{A}\left(\widehat{\boldsymbol{\alpha}}_\star^k-\widehat{\mathbf{u}}_t^k\right)\right\|^2-\frac{\vartheta\theta_t}{2}\|\boldsymbol{\alpha}_\star-\boldsymbol{\alpha}_t\|^2-\frac{\vartheta(1-\theta_t)}{2}\|\boldsymbol{\alpha}_{t-1}-\boldsymbol{\alpha}_t\|^2.$$

Substituting (27) into the above inequality, we obtain

$$\begin{aligned}
&\left(D(\boldsymbol{\alpha}_t)-D(\boldsymbol{\alpha}_\star)\right)+\frac{\zeta_t}{2}\sum_{k=1}^{K}\left\|\mathbf{A}\left(\widehat{\boldsymbol{\alpha}}_\star^k-\widehat{\mathbf{v}}_t^k\right)\right\|^2\\
&\quad\le(1-\theta_t)\Big(D(\boldsymbol{\alpha}_{t-1})-D(\boldsymbol{\alpha}_\star)+\frac{\zeta_{t-1}}{2}\sum_{k=1}^{K}\left\|\mathbf{A}\left(\widehat{\boldsymbol{\alpha}}_\star^k-\widehat{\mathbf{v}}_{t-1}^k\right)\right\|^2\\
&\qquad-\frac{\zeta_{t-1}}{2}\sum_{k=1}^{K}\left\|\mathbf{A}\left(\widehat{\mathbf{u}}_t^k-\widehat{\mathbf{v}}_{t-1}^k\right)\right\|^2\Big)\\
&\qquad+\frac{(1-\theta_t)\zeta_{t-1}}{2}\Big(1-\frac{\vartheta\theta_t}{\zeta_t}\Big)\left\|\mathbf{A}\left(\widehat{\mathbf{u}}_t^k-\widehat{\mathbf{v}}_{t-1}^k\right)\right\|^2\\
&\qquad-\frac{\vartheta\theta_t}{2}\sum_{k=1}^{K}\left\|\mathbf{A}\left(\widehat{\boldsymbol{\alpha}}_\star^k-\widehat{\mathbf{u}}_t^k\right)\right\|^2-\frac{\vartheta\theta_t}{2}\|\boldsymbol{\alpha}_\star-\boldsymbol{\alpha}_t\|^2\\
&\qquad-\frac{\vartheta(1-\theta_t)}{2}\|\boldsymbol{\alpha}_{t-1}-\boldsymbol{\alpha}_t\|^2\\
&\quad=(1-\theta_t)\Big(D(\boldsymbol{\alpha}_{t-1})-D(\boldsymbol{\alpha}_\star)+\frac{\zeta_{t-1}}{2}\sum_{k=1}^{K}\left\|\mathbf{A}\left(\widehat{\boldsymbol{\alpha}}_\star^k-\widehat{\mathbf{v}}_{t-1}^k\right)\right\|^2\Big)\\
&\qquad-\frac{(1-\theta_t)\zeta_{t-1}}{2}\frac{\vartheta\theta_t}{\zeta_t}\left\|\mathbf{A}\left(\widehat{\mathbf{u}}_t^k-\widehat{\mathbf{v}}_{t-1}^k\right)\right\|^2\\
&\qquad-\frac{\vartheta\theta_t}{2}\sum_{k=1}^{K}\left\|\mathbf{A}\left(\widehat{\boldsymbol{\alpha}}_\star^k-\widehat{\mathbf{u}}_t^k\right)\right\|^2-\frac{\vartheta\theta_t}{2}\|\boldsymbol{\alpha}_\star-\boldsymbol{\alpha}_t\|^2-\frac{\vartheta(1-\theta_t)}{2}\|\boldsymbol{\alpha}_{t-1}-\boldsymbol{\alpha}_t\|^2,
\end{aligned}$$

which can be rewritten as

$$\begin{aligned}
&\left(D(\boldsymbol{\alpha}_t)-D(\boldsymbol{\alpha}_\star)\right)+\frac{\zeta_t}{2}\sum_{k=1}^{K}\left\|\mathbf{A}\left(\widehat{\boldsymbol{\alpha}}_\star^k-\widehat{\mathbf{v}}_t^k\right)\right\|^2\\
&\quad\le(1-\theta_t)\Big(D(\boldsymbol{\alpha}_{t-1})-D(\boldsymbol{\alpha}_\star)+\frac{\zeta_{t-1}}{2}\sum_{k=1}^{K}\left\|\mathbf{A}\left(\widehat{\boldsymbol{\alpha}}_\star^k-\widehat{\mathbf{v}}_{t-1}^k\right)\right\|^2\Big)\\
&\qquad-\frac{(1-\theta_t)\zeta_{t-1}}{2}\frac{\vartheta\theta_t}{\zeta_t}\left\|\mathbf{A}\left(\widehat{\mathbf{u}}_t^k-\widehat{\mathbf{v}}_{t-1}^k\right)\right\|^2\\
&\qquad-\frac{\vartheta\theta_t}{2}\sum_{k=1}^{K}\left\|\mathbf{A}\left(\widehat{\boldsymbol{\alpha}}_\star^k-\widehat{\mathbf{u}}_t^k\right)\right\|^2-\frac{\vartheta\theta_t}{2}\|\boldsymbol{\alpha}_\star-\boldsymbol{\alpha}_t\|^2-\frac{\vartheta(1-\theta_t)}{2}\|\boldsymbol{\alpha}_{t-1}-\boldsymbol{\alpha}_t\|^2.
\end{aligned}$$

This implies

$$\big(D(\boldsymbol{\alpha}_t) - D(\boldsymbol{\alpha}_\star)\big) + \sum_{k=1}^{K} \frac{\zeta_t}{2} \big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_t^k\big)\big\|^2$$

$$\leq (1-\theta_t)\Big(D(\boldsymbol{\alpha}_{t-1}) - D(\boldsymbol{\alpha}_\star) + \frac{\zeta_{t-1}}{2} \sum_{k=1}^{K} \big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_{t-1}^k\big)\big\|^2\Big).$$

Applying the above inequality for $i = 1$ to t, we obtain

$$\big(D(\boldsymbol{\alpha}_t) - D(\boldsymbol{\alpha}_\star)\big) + \frac{\zeta_t}{2} \sum_{k=1}^{K} \big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_t^k\big)\big\|^2$$

$$\leq \prod_{i=1}^{t}(1-\theta_i)\Big(D(\boldsymbol{\alpha}_0) - D(\boldsymbol{\alpha}_\star) + \frac{\zeta_0}{2} \big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_0^k\big)\big\|^2\Big),$$

By using the definition of γ_t, the above inequality can be rewritten as

$$\big(D(\boldsymbol{\alpha}_t) - D(\boldsymbol{\alpha}_\star)\big) + \frac{\zeta_t}{2} \sum_{k=1}^{K} \big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_t^k\big)\big\|^2 \leq \gamma_t \Big(D(\boldsymbol{\alpha}_0) - D(\boldsymbol{\alpha}_\star) + \frac{\zeta_0}{2} \big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_0^k\big)\big\|^2\Big).$$

This completes the proof. □

B.2 Convergence analysis for smooth losses

B.2.1 Proof of Lemma 3

Lemma 3 *Assume the loss functions f_{ki}'s are $(1/\mu)$-smooth for any $k \in [K]$ and $i \in [n_k]$. If $\theta_0 = \sqrt{\vartheta\eta}$ and $(1-\rho)\lambda\mu n \leq 1$, then the following inequality holds for any $t \geq 1$*

$$\epsilon_D^t \leq \Big(1 - \sqrt{(1-\rho)\lambda\mu n}\Big)^t \big(\epsilon_D^0 + R^0\big). \tag{12}$$

Proof It can be proved by using Lemma 2. From Lemma 1, we know that f_{ki} are μ-strongly convex for any $k \in [K], i \in [n_k]$ since f_{ki} is $(1/\mu)$-smooth. If $\zeta_{t-1} \geq \vartheta$, then $\zeta_t = (1-\theta_t)\zeta_{t-1} + \vartheta\theta_t \geq (1-\theta_t)\vartheta + \theta_t\vartheta = \vartheta$. There we have $\zeta_t \geq \vartheta$ holds for any $t \geq 1$ since $\zeta_0 \geq \vartheta$. Hence,

$$\theta_t^2/\eta = \zeta_t \Rightarrow \theta_t \geq \sqrt{\eta\zeta_t} \Rightarrow \theta_t \geq \sqrt{\eta\vartheta} = \sqrt{(1-\rho)\lambda\mu n}.$$

Then, γ_t can be bounded

$$\gamma_t = \prod_{i=1}^{t}(1-\theta_i) \leq \big(1 - \sqrt{(1-\rho)\lambda\mu n}\big)^t.$$

Substituting this result and $\mathbf{v}_0 = \boldsymbol{\alpha}_0$ into (11), we obtain

$$D(\boldsymbol{\alpha}_t) - D(\boldsymbol{\alpha}_\star) \leq \big(1 - \sqrt{(1-\rho)\lambda\mu n}\big)^t \Big(D(\boldsymbol{\alpha}_0) - D(\boldsymbol{\alpha}_\star) + \frac{\zeta_0}{2} \big\|\mathbf{A}(\boldsymbol{\alpha}_\star - \boldsymbol{\alpha}_0)\big\|^2\Big).$$

This completes the proof. □

B.2.2 Proof of Theorem 1

Theorem 1 *Assume the loss functions f_{ki}'s are $(1/\mu)$-smooth for any $k \in [K]$ and $i \in [n_k]$. If $\theta_0 = \sqrt{\vartheta\eta}$ and $(1-\rho)\lambda\mu n \leq 1$, then after T iterations in Algorithm 1 with*

$$T \geq \sqrt{\frac{1}{(1-\rho)\lambda\mu n}} \log\left(\left(1+\sigma_{max}\right)\frac{\epsilon_D^0}{\epsilon_D}\right),$$

$D\big(\boldsymbol{\alpha}_T\big) - D\big(\boldsymbol{\alpha}_\star\big) \leq \epsilon_D$ holds. Furthermore, after T iterations with

$$T \geq \sqrt{\frac{1}{(1-\rho)\lambda\mu n}} \log\left(\left(1+\sigma_{max}\right)\frac{(1-\rho)\lambda\mu n + \sigma_{max}}{(1-\rho)\lambda\mu n}\frac{\epsilon_D^0}{\epsilon_G}\right),$$

it holds that $P\big(\mathbf{w}(\boldsymbol{\alpha}_T)) - (-D(\boldsymbol{\alpha}_T)) \leq \epsilon_G$.

Proof It is easy to see that $D(\boldsymbol{\alpha})$ is ϑ-strongly convex since f_{ki} is $(1/\mu)$-smooth for any $k \in [K]$, $i \in [n_k]$. It implies

$$D(\boldsymbol{\alpha}_0) \geq D(\boldsymbol{\alpha}_\star) + \frac{\vartheta}{2}\|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}_\star\|^2 \Rightarrow \|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}_\star\|^2 \leq \frac{2}{\vartheta}\big(D(\boldsymbol{\alpha}_0) - D(\boldsymbol{\alpha}_\star)\big) = \frac{2}{\vartheta}\epsilon_D^0.$$

By using this result, (12) can be rewrite as

$$\begin{aligned}
\epsilon_D^t = D(\boldsymbol{\alpha}_t) - D(\boldsymbol{\alpha}_\star) &\leq \big(1 - \sqrt{(1-\rho)\lambda\mu n}\big)^t\Big(D(\boldsymbol{\alpha}_0) - D(\boldsymbol{\alpha}_\star) + \frac{\zeta_0}{2}\|\mathbf{A}(\boldsymbol{\alpha}_\star - \boldsymbol{\alpha}_0)\|^2\Big)\\
&\leq \big(1 - \sqrt{(1-\rho)\lambda\mu n}\big)^t\Big(\epsilon_D^0 + \frac{\zeta_0}{2}\sigma_{\max}\|\boldsymbol{\alpha}_\star - \boldsymbol{\alpha}_0\|^2\Big)\\
&\leq \big(1 - \sqrt{(1-\rho)\lambda\mu n}\big)^t\Big(\epsilon_D^0 + \frac{\zeta_0}{2}\sigma_{\max}\frac{2}{\vartheta}\epsilon_D^0\Big)\\
&= \big(1 - \sqrt{(1-\rho)\lambda\mu n}\big)^t(1+\sigma_{\max})\epsilon_D^0\\
&\leq \exp(-t\sqrt{(1-\rho)\lambda\mu n})(1+\sigma_{\max})\epsilon_D^0,
\end{aligned}$$

where the last upper bound will be smaller than ϵ_D if

$$t \geq \sqrt{\frac{1}{(1-\rho)\lambda\mu n}} \log\left((1+\sigma_{\max})\frac{\epsilon_D^0}{\epsilon_D}\right).$$

By applying Lemma 6, we know that for any

$$\begin{aligned}
T &\geq \sqrt{\frac{1}{(1-\rho)\lambda\mu n}} \log\frac{(1+\sigma_{\max})\epsilon_D^0\left(\frac{\sigma_{\max}}{(1-\rho)\lambda n^2} + \vartheta\right)}{\vartheta\epsilon_G}\\
\Rightarrow\ T &\geq \sqrt{\frac{1}{(1-\rho)\lambda\mu n}} \log\left((1+\sigma_{\max})\frac{(1-\rho)\lambda\mu n + \sigma_{\max}}{(1-\rho)\lambda\mu n}\frac{\epsilon_D^0}{\epsilon_G}\right),
\end{aligned}$$

it holds that $D(\boldsymbol{\alpha}_T) - P\big(\mathbf{w}(\boldsymbol{\alpha}_T)) \leq \epsilon_G$. □

B.3 Convergence analysis for Lipschitz continuous losses: Proof of Theorem 2

Theorem 2 *Assume the loss functions f_{ki}'s are generally convex and L-Lipschitz continuous for any $k \in [K]$, $i \in [n_k]$. If $\theta_0 = 1$, the following inequality holds for any $t \geq 1$*

$$\epsilon_D^t \leq \frac{1}{(t+2)^2}\Big(4\epsilon_D^0 + \frac{8L^2\sigma_{max}}{(1-\rho)\lambda n^2}\Big). \tag{13}$$

After T iterations in Algorithm 1 with

$$T \geq \sqrt{\frac{8L^2\sigma_{max}}{(1-\rho)\lambda n^2\epsilon_D} + \frac{4\epsilon_D^0}{\epsilon_D}} - 2, \tag{14}$$

it holds that $D\big(\boldsymbol{\alpha}_T\big) - D\big(\boldsymbol{\alpha}_\star\big) \leq \epsilon_D$.

Proof It can be proved by using Lemma 2. It is easy to see that f_{ki}^* are general convex (i.e., $\mu = 0$) since f_{ki} are L-Lipschitz continuous for any $k \in [K], i \in [n_k]$. By using the definition of ζ_t and the fact that $\mu = 0$, we obtain $\gamma_t = (1-\theta_t)\gamma_{t-1} = \zeta_t/\zeta_{t-1}\gamma_{t-1}$. Applying the above identity from $i = 1$ to t, we obtain $\gamma_t = \lambda_0\zeta_t/\zeta_0 = \zeta_t/\zeta_0$. In addition, we can obtain $\theta_t = \big(\gamma_{t-1} - \gamma_t\big)/\gamma_{t-1}$ from $\gamma_t = (1-\theta_t)\gamma_{t-1}$. Therefore,

$$\begin{aligned}\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}} = \frac{2\gamma_{t-1} - 2\sqrt{\gamma_t\gamma_{t-1}}}{2\gamma_{t-1}\sqrt{\gamma_t}} &\geq \frac{2\gamma_{t-1} - (\gamma_{t-1} + \gamma_t)}{2\gamma_{t-1}\sqrt{\gamma_t}} \\ &= \frac{\theta_t}{2\sqrt{\gamma_t}} = \frac{\theta_t}{2\sqrt{\zeta_t/\zeta_0}}.\end{aligned}$$

By using $\theta_t^2/\eta = \zeta_t$, we obtain $1/\gamma_t - 1/\gamma_{t-1} \geq 0.5\sqrt{\eta\zeta_0} = 0.5\sqrt{\zeta_0(1-\rho)\lambda n^2}$. Combing the above inequality from $i = 1$ to $i = t$, we obtain

$$\frac{1}{\sqrt{\gamma_t}} - \frac{1}{\sqrt{\gamma_0}} \geq \frac{t}{2}\sqrt{\eta\zeta_0} \Rightarrow \gamma_t \leq \frac{4}{\big(t\sqrt{\eta\zeta_0} + 2\big)^2} = \frac{4}{\big(t\sqrt{\zeta_0(1-\rho)\lambda n^2} + 2\big)^2}.$$

Substituting this results into (11), we obtain

$$D(\boldsymbol{\alpha}_t) - D(\boldsymbol{\alpha}_\star) \leq \frac{4}{\big(t\sqrt{\zeta_0(1-\rho)\lambda n^2} + 2\big)^2}\Big(D(\boldsymbol{\alpha}_0) - D(\boldsymbol{\alpha}_\star) + \frac{\zeta_0}{2}\sum_{k=1}^{K}\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\mathbf{v}}_0^k\big)\big\|^2\Big)$$

Since $\theta_0 = 1$, we have $\zeta_0 = \theta_0^2/\eta = 1/((1-\rho)\lambda n^2)$. Substituting the value of ζ_0 into (13), we obtain

$$\begin{aligned}D(\boldsymbol{\alpha}_t) - D(\boldsymbol{\alpha}_\star) &\leq \frac{4}{\big(t\sqrt{\zeta_0(1-\rho)\lambda n^2} + 2\big)^2}\left(D(\boldsymbol{\alpha}_0) - D(\boldsymbol{\alpha}_\star) + \frac{\zeta_0}{2}\sum_{k=1}^{K}\big\|\mathbf{A}\big(\widehat{\boldsymbol{\alpha}}_\star^k - \widehat{\boldsymbol{\alpha}}_0^k\big)\big\|^2\right) \\ &= \frac{4}{(t+2)^2}\Big(\epsilon_D^0 + \frac{1}{2(1-\rho)\lambda n^2}\|\mathbf{A}(\boldsymbol{\alpha}_\star - \boldsymbol{\alpha}_0)\|^2\Big) \\ &\leq \frac{4}{(t+2)^2}\Big(\epsilon_D^0 + \frac{1}{2(1-\rho)\lambda n^2}\sigma_{\max}\|\boldsymbol{\alpha}_\star - \boldsymbol{\alpha}_0\|^2\Big) \\ &\leq \frac{1}{(t+2)^2}\Big(4\epsilon_D^0 + \frac{8L^2\sigma_{\max}}{(1-\rho)\lambda n^2}\Big),\end{aligned}$$

where the last upper bound will be smaller than ϵ_D if

$$T \geq \sqrt{\frac{8L^2\sigma_{\max}}{(1-\rho)\lambda n^2\epsilon_D} + \frac{4\epsilon_D^0}{\epsilon_D}} - 2.$$

This completes the proof. □

Appendix C: More details of dynamic feature screening

C.1 Proof of Lemma 4

Lemma 4 *Assume the loss functions f_{ki}'s are $(1/\mu)$-smooth for any $k \in [K]$, $i \in [n_k]$. For any dual feasible solution $\boldsymbol{\alpha}$, it holds that $\boldsymbol{\alpha}_\star \in \mathscr{F} \overset{\text{def}}{=} \left\{\boldsymbol{\theta} \mid \|\boldsymbol{\theta} - \boldsymbol{\alpha}\| \leq \sqrt{2G(\boldsymbol{\alpha})n/\mu}\right\}$.*

Proof Since f_{ki} are $(1/\mu)$-smooth for any $k \in [K]$, $i \in [n_k]$, it implies that $D(\boldsymbol{\alpha})$ is (μ/n)-strongly convex.

$$\begin{aligned}
D(\boldsymbol{\alpha}) &\overset{a}{\geq} D\left(\boldsymbol{\alpha}_\star\right) + \left\langle \partial D\left(\boldsymbol{\alpha}_\star\right), \boldsymbol{\alpha} - \boldsymbol{\alpha}_\star \right\rangle + \frac{\mu}{2n}\left\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\star\right\|^2 \\
\Rightarrow -D(\boldsymbol{\alpha}) &\leq -D\left(\boldsymbol{\alpha}_\star\right) - \left\langle \partial D\left(\boldsymbol{\alpha}_\star\right), \boldsymbol{\alpha} - \boldsymbol{\alpha}_\star \right\rangle - \frac{\mu}{2n}\left\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\star\right\|^2 \\
\Rightarrow -D(\boldsymbol{\alpha}) &\overset{b}{\leq} P(\mathbf{w}(\boldsymbol{\alpha})) - \left\langle \partial D\left(\boldsymbol{\alpha}_\star\right), \boldsymbol{\alpha} - \boldsymbol{\alpha}_\star \right\rangle - \frac{\mu}{2n}\left\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\star\right\|^2 \\
\Rightarrow -D(\boldsymbol{\alpha}) &\overset{c}{\leq} P(\mathbf{w}(\boldsymbol{\alpha})) - \frac{\mu}{2n}\left\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\star\right\|^2,
\end{aligned}$$

where (a) follows from $D(\boldsymbol{\alpha})$ is (μ/n)-strongly convex , (b) is obtained by applying the weakly duality theorem, and (c) follows from the optimality of $\boldsymbol{\alpha}_\star$. Therefore, we obtain

$$\left\|\boldsymbol{\alpha}_\star - \boldsymbol{\alpha}\right\| \leq \sqrt{2nG(\boldsymbol{\alpha})/\mu} \Rightarrow \boldsymbol{\alpha}_\star \in \mathscr{B}(\boldsymbol{\alpha}, \sqrt{2nG(\boldsymbol{\alpha})/\mu}).$$

This completes the proof. □

Before proving Lemma 5, we first introduce the following lemma.

Lemma 8 (Gay 1981) *Let us consider the following minimization problem*

$$\min_{\mathbf{s}\in\mathbb{R}^n} \left\{\psi(\mathbf{s}) \overset{\text{def}}{=} \frac{1}{2}\langle \mathbf{s}, \mathbf{H}\mathbf{s}\rangle + \langle \mathbf{g}, \mathbf{s}\rangle\right\} \text{ s.t. } \|\mathbf{D}\mathbf{s}\| \leq \delta, \tag{30}$$

where $\mathbf{H} \in \mathbb{R}^{n\times n}$ be a symmetric matrix, $\mathbf{D} \in \mathbb{R}^{n\times n}$ is an nonsingular matrix, and $\delta > 0$. Then, $\mathbf{s}_\star$ minimizes $\psi(\mathbf{s})$ over the constraint set if and only if there exists a $\vartheta_\star \geq 0$ such that

$$\mathbf{H} + \vartheta_\star \mathbf{D}^\top \mathbf{D} \succeq 0 \tag{31}$$

$$\left(\mathbf{H} + \vartheta_\star \mathbf{D}^\top \mathbf{D}\right)\mathbf{s}_\star = -\mathbf{g} \tag{32}$$

$$\|\mathbf{D}\mathbf{s}_\star\| = \delta \;\; \textit{if}\; \vartheta_\star > 0. \tag{33}$$

This $\vartheta_\star$ is unique.

Next, we prove Lemma 5 by using Lemma 8.

Lemma 5 *If* $\upsilon_j = 0$, *the maximum value of* (19) *is* 0. *Otherwise, the upper bound is*

$$\sum_{k=1}^{K} \left\langle \mathbf{X}_{j\cdot}^{k}, \boldsymbol{\alpha}^{k} \right\rangle^{2} + \frac{nG(\boldsymbol{\alpha})}{\mu} \vartheta_{\star} - \frac{1}{2} \left\langle \mathbf{g}, \mathbf{s}_{\star} \right\rangle,$$

where $\vartheta_\star$ *and* $\mathbf{s}_\star$ *are defined as follows: (a)* $\vartheta_\star = 2\upsilon_j$ *and* $\mathbf{s}^\star = \overline{\mathbf{s}} + \widehat{\mathbf{s}}$ *if 1)* $\exists\, \widehat{\mathbf{s}} \in \mathbb{R}^K$ *with* $\widehat{\mathbf{s}}_{\mathscr{I}_j} = \mathbf{0}$ *and* $\|\overline{\mathbf{s}} + \widehat{\mathbf{s}}\| = \sqrt{2G(\boldsymbol{\alpha})n/\mu}$, *and 2)* $\left\langle \mathbf{X}_{\cdot j}^{t}, \boldsymbol{\theta}_{t} \right\rangle = 0, \forall t \in \mathscr{I}_j$. *(b) Otherwise,* $\vartheta_\star > 2\upsilon_j$ *is solution of* $\|\, (\mathbf{H} + \vartheta_\star \mathbf{I})^{-1}\, \mathbf{g}\| = \sqrt{2G(\boldsymbol{\alpha})n/\mu}$, *and* $\mathbf{s}_\star = -(\mathbf{H} + \vartheta_\star \mathbf{I})^{-1}\mathbf{g}$.

Proof Let $\mathbf{z} = \boldsymbol{\theta} - \boldsymbol{\alpha}$, then (19) is equivalent to

$$\max_{\mathbf{z}} \left\| \mathbf{A}_{\mathscr{G}_j \cdot} (\mathbf{z} + \boldsymbol{\alpha}) \right\|^2 \ \text{ s.t. } \|\mathbf{z}\| \le \sqrt{2G(\boldsymbol{\alpha})n/\mu},$$

The objective can be relaxed as following

$$\begin{aligned}
\left\| \mathbf{A}_{\mathscr{G}_j \cdot} (\mathbf{z} + \boldsymbol{\alpha}) \right\|^2 &= \sum_{k=1}^{K} \left\langle \mathbf{X}_{j\cdot}^{k}, (\mathbf{z}^{k} + \boldsymbol{\alpha}^{k}) \right\rangle^{2}, \\
&= \sum_{k=1}^{K} \left(\left\langle \mathbf{X}_{j\cdot}^{k}, \mathbf{z}^{k} \right\rangle^{2} + 2\left\langle \mathbf{X}_{j\cdot}^{k}, \mathbf{z}^{k} \right\rangle \left\langle \mathbf{X}_{j\cdot}^{k}, \boldsymbol{\alpha}^{k} \right\rangle + \left\langle \mathbf{X}_{j\cdot}^{k}, \boldsymbol{\alpha}^{k} \right\rangle^{2} \right), \\
&\le \sum_{k=1}^{K} \left(\left\| \mathbf{X}_{j\cdot}^{k} \right\|^{2} \|\mathbf{z}^{k}\|^{2} + 2\left| \left\langle \mathbf{X}_{j\cdot}^{k}, \boldsymbol{\alpha}^{k} \right\rangle \right| \left\| \mathbf{X}_{j\cdot}^{k} \right\| \|\mathbf{z}^{k}\| \right) + \sum_{k=1}^{K} \left\langle \mathbf{X}_{j\cdot}^{k}, \boldsymbol{\alpha}^{k} \right\rangle^{2}.
\end{aligned}$$

Let $\mathbf{s} \in \mathbb{R}^K$ with $s_t = \|\mathbf{z}^k\|$, we then define $\psi(\mathbf{s})$ as $\psi(\mathbf{s}) = \frac{1}{2}\langle \mathbf{s}, \mathbf{H}\mathbf{s} \rangle + \langle \mathbf{g}, \mathbf{s} \rangle$. By using the relaxed objective function, (19) becomes

$$\max_{\|\mathbf{s}\| \le \sqrt{2G(\boldsymbol{\alpha})n/\mu}} -\psi(\mathbf{s}) + \sum_{k=1}^{K} \left\langle \mathbf{X}_{j\cdot}^{k}, \boldsymbol{\alpha}^{k} \right\rangle^{2} = - \min_{\|\mathbf{s}\| \le \sqrt{2G(\boldsymbol{\alpha})n/\mu}} \psi(\mathbf{s}) + \sum_{k=1}^{K} \left\langle \mathbf{X}_{j\cdot}^{k}, \boldsymbol{\alpha}^{k} \right\rangle^{2},$$

where $\min_{\|\mathbf{s}\| \le \sqrt{2G(\boldsymbol{\alpha})n/\mu}} \psi(\mathbf{s})$ can be rewritten in the form of (30) by defining $\mathbf{D} = \mathbf{I}$ and $\delta = \sqrt{2G(\boldsymbol{\alpha})n/\mu}$. Then, Lemma 8 implies there exists a unique $\vartheta_\star$ such that

$$\mathbf{H} + \vartheta_\star \mathbf{I} \succeq \mathbf{0} \Rightarrow \vartheta_\star \ge \max_{k \in [K]} 2\left\| \mathbf{X}_{j\cdot}^{k} \right\|^{2} \Rightarrow \vartheta_\star \ge 2\upsilon_j,$$

which implies $\vartheta_\star \ge> 0$ since $\upsilon_j > 0$. Then, the problem can be considered as two cases $\vartheta_\star = 2\upsilon_j$ and $\vartheta_\star \ge 2\upsilon_j$. Given $\vartheta_\star$ and $\mathbf{s}_\star$, $\psi(\mathbf{s}_\star)$ can be formulated by using (32) and (33)

$$\begin{aligned}
\psi\,(\mathbf{s}_\star) = \frac{1}{2}\left\langle \mathbf{s}_\star, \mathbf{H}\mathbf{s}_\star \right\rangle + \left\langle \mathbf{g}, \mathbf{s}_\star \right\rangle &= \frac{1}{2}\left\langle \mathbf{s}_\star, (\mathbf{H} + \vartheta_\star \mathbf{I})\, \mathbf{s}_\star \right\rangle + \left\langle \mathbf{g}, \mathbf{s}_\star \right\rangle - \frac{\vartheta_\star}{2}\|\mathbf{s}_\star\|^2 \\
&= -\frac{1}{2}\left\langle \mathbf{s}_\star, \mathbf{g} \right\rangle + \left\langle \mathbf{g}, \mathbf{s}_\star \right\rangle - \frac{\vartheta_\star}{2}\delta^2 = \frac{1}{2}\left\langle \mathbf{g}, \mathbf{s}_\star \right\rangle - \frac{nG(\boldsymbol{\alpha})}{\mu}\vartheta_\star,
\end{aligned}$$

which implies the upper bound of (19) is

$$\sum_{k=1}^{K} \left\langle \mathbf{X}_{j\cdot}^{k}, \boldsymbol{\alpha}^{k} \right\rangle^{2} - \psi\,(\mathbf{s}_\star) = \sum_{k=1}^{K} \left\langle \mathbf{X}_{j\cdot}^{k}, \boldsymbol{\alpha}^{k} \right\rangle^{2} + \frac{nG(\boldsymbol{\alpha})}{\mu}\vartheta_\star - \frac{1}{2}\left\langle \mathbf{g}, \mathbf{s}_\star \right\rangle.$$

Next, we show the values of $\mathbf{s}_\star$ when $\vartheta_\star = 2\upsilon_j$ and $\vartheta_\star \ge 2\upsilon_j$, respectively.

Table 2 Statistics of the datasets for STL

Dataset	# Samples	# Features	Sparsity (%)
RCV1	677,399	47,236	1.5e−3
URL	2,396,130	3,231,961	3.5e−5

Case 1: $\vartheta_\star = 2\upsilon_j$. In this case, (32) and (33) imply $(\mathbf{H} + 2\upsilon_j\mathbf{I})\mathbf{s}_\star = -\mathbf{g}$ and $\|\mathbf{s}_\star\| = \delta$ that is equivalent to Therefore, if all above conditions hold then $\vartheta_\star = 2\upsilon_j$, otherwise we have $\vartheta_\star$ that is discussed in the following.
Case 2: $\vartheta_\star > 2\upsilon_j$. In this case, $\mathbf{H} + \vartheta_\star\mathbf{I}$ is an invertible matrix. From (32) and (33), we obtain

$$(\mathbf{H} + \vartheta_\star)\,\mathbf{s}_\star = -\mathbf{g} \quad \text{and} \quad \|\mathbf{s}_\star\| = \delta,$$

which implies $\mathbf{s}_\star = -\left(\mathbf{H} + \vartheta_\star\mathbf{I}\right)^{-1}\mathbf{g}$ and $\left\|\left(\mathbf{H} + \vartheta_\star\mathbf{I}\right)^{-1}\mathbf{g}\right\| = \sqrt{2G(\boldsymbol{\alpha})n/\mu}$. This completes the proof. □

Lemma 5 shows that there exists a global optimum $\vartheta_\star$, however, we need some algorithm to obtain the value for the case of $\vartheta_\star > 2\upsilon_j$. Note that $\vartheta_\star \in (2\upsilon_j, \infty)$ is the unique solution of

$$\varphi(\vartheta) = \frac{1}{\|\left(\mathbf{H} + \vartheta\mathbf{I}\right)^{-1}\mathbf{g}\|} - \sqrt{\frac{\mu}{2G(\boldsymbol{\alpha})n}} = 0.$$

The above equation can be efficiently solved by using Newton's method. Besides Newton's method, $\vartheta_\star$ can also be efficiently solved by using bisection method.

Appendix D: More details on single task learning

In this section, we provide more details on the extension of our method to single task learning. Specifically, we consider the following ℓ_1-norm regularized learning problem [i.e., elastic net (Zou and Hastie 2005)]

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} f_{ki}\left(\langle \mathbf{x}_i^k, \mathbf{w}\rangle\right) + \lambda\Big(\rho\|\mathbf{w}\|_1 + \frac{1-\rho}{2}\|\mathbf{w}\|^2\Big). \tag{34}$$

Then, the local subproblem for each worker is

$$\begin{aligned} L_k\big(\widehat{\boldsymbol{\alpha}}^k; \mathbf{u}_t\big) \overset{\text{def}}{=} & \frac{1}{n} f_k^*\big(-\widehat{\boldsymbol{\alpha}}^k\big) + \frac{1}{n}\Big\langle \nabla g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big), \mathbf{A}\big(\widehat{\boldsymbol{\alpha}}^k - \widehat{\mathbf{u}}_t^k\big)\Big\rangle \\ & + \frac{\sigma'}{2\eta}\big\|\mathbf{A}(\widehat{\boldsymbol{\alpha}}^k - \widehat{\mathbf{u}}_t^k)\big\|^2 + \frac{\lambda}{K} g^*\Big(\frac{\mathbf{A}\mathbf{u}_t}{\lambda n}\Big), \end{aligned}$$

where $\eta \overset{\text{def}}{=} (1-\rho)\lambda n^2$ and a safe value for σ' is $\sigma' = K$ (Ma et al. 2015). We compare the performance of our method with COCOA+ on two datasets (Table 2) with smoothed hinge loss (Shalev-Shwartz and Zhang 2013)

$$f_{ki}(z_i^k) = \begin{cases} 0 & \text{if } y_i^k z_i^k \geq 1 \\ 1 - y_i^k z_i^k - \frac{\mu}{2} & \text{if } y_i^k z_i^k \leq 1 - \mu \\ \frac{1}{2\mu}\big(1 - y_i^k z_i^k\big)^2 & \text{otherwise} \end{cases}$$

where μ is set to $\mu = 0.5$.

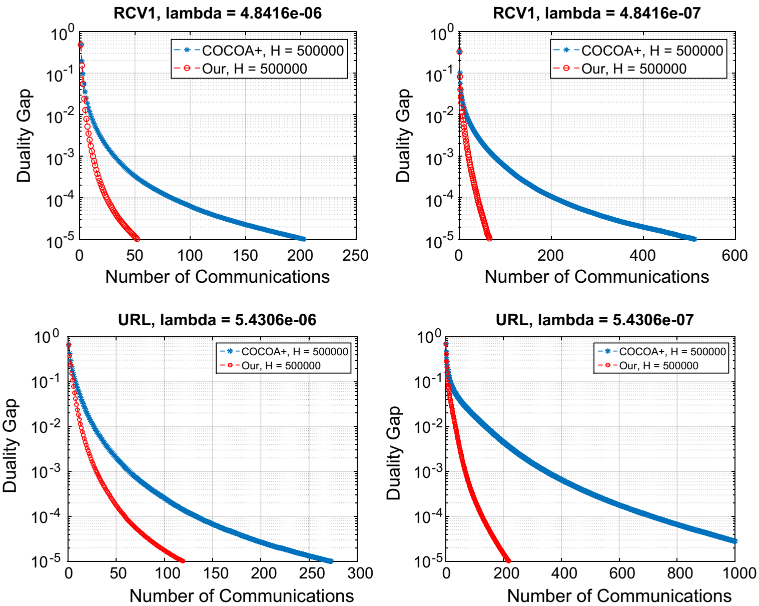

Fig. 4 Duality gap versus communicated iterations for $\lambda = 10^{-2}\lambda_{\max}$ and $\lambda = 10^{-3}\lambda_{\max}$

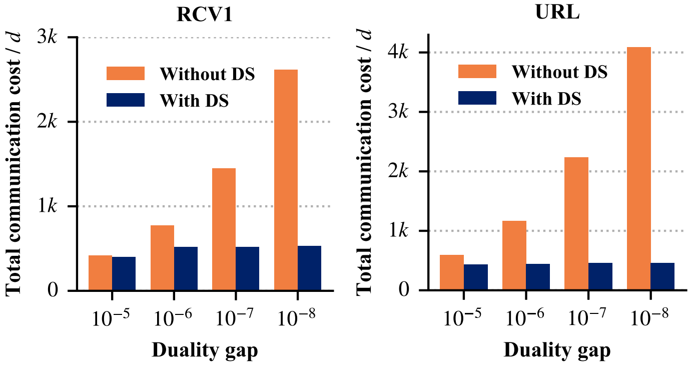

Fig. 5 Effects of dynamic screening for reducing communication costs. Total communication costs (normalized by feature dimension d) used by the proposed method *without* and *with* dynamic screening for solving (3) over $\{\lambda_i\}_{i=1}^{50}$ on RCV1 and URL

We compare the performance of our method with COCOA+ on two datasets in Table 2 with smoothed hinge loss (Shalev-Shwartz and Zhang 2013). In our experiments, 8 workers are used (i.e., $K = 8$) and $\rho = 0.9$ for both datasets. The SDCA (Shalev-Shwartz and Zhang 2013) is used as local solver for both methods and H is set to $H = 5 \times 10^5$. We evaluate two methods for $\lambda = 10^{-2}\lambda_{\max}$ and $\lambda = 10^{-3}\lambda_{\max}$. Figure 4 shows the comparison in terms of the number iterations for communication used by our method and COCOA+ to obtain a solution meeting a prescribed duality gap. In addition, we also evaluate the effect of dynamic screening for further reduced communication cost. The setting is the same as that presented in Sect. 7.4. Figure 5 presents the total communication cost used by our method *without* and *with* dynamic screening to solve (34) on RCV1 and URL. As observed, the proposed method performs as well as it works for MTL.

References

Arjevani, Y., & Shamir, O. (2015). Communication complexity of distributed convex learning and optimization. In *Proceedings of NIPS*.

Baytas, I. M., Yan, M., Jain, A. K., & Zhou, J. (2016). Asynchronous multi-task learning. In *Proceedings of ICDM*.

Bellet, A., Guerraoui, R., Taziki, M., & Tommasi, M. (2018). Personalized and private peer-to-peer machine learning. In: *Proceedings of AISTATS*.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Wortman, J. (2007). Learning bounds for domain adaptation. In *Proceedings of NIPS*.

Bonnefoy, A., Emiya, V., Ralaivola, L., & Gribonval, R. (2015). Dynamic screening: Accelerating first-order algorithms for the lasso and group-lasso. *IEEE Transactions on Signal Processing*, *63*(19), 5121–5132.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, *2*, 499–526.

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

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*(1), 41–75.

Dünner, C., Forte, S., Takác, M., & Jaggi, M. (2016). Primal–dual rates and certificates. In *Proceedings of ICML* (pp 783–792).

Fercoq, O., Gramfort, A., & Salmon, J. (2015). Mind the duality gap: Safer rules for the lasso. In *Proceedings of ICML*.

Gay, D. (1981). Computing optimal locally constrained steps. *SIAM Journal on Scientific and Statistical Computing*, *2*(2), 186–197.

Hiriart-Urruty, J. B., & Lemaréchal, C. (1993). *Convex analysis and minimization algorithms II: Advanced theory and bundle methods*. Berlin: Springer.

Jaggi, M., Smith, V., Takác, M., Terhorst, J., Krishnan, S., Hofmann, T., et al. (2014). Communication-efficient distributed dual coordinate ascent. In *Proceedings of NIPS*.

Lang, K. (1995). Newsweeder: Learning to filter netnews. In *Proceedings of ICML*.

Lee, S., Zhu, J., & Xing, E. P. (2010). Adaptive multi-task lasso: With application to eQTL detection. In *Proceedings of NIPS*.

Li, M., Andersen, D. G., Smola, A. J., & Yu, K. (2014). Communication efficient distributed machine learning with the parameter server. In *Proceedings of NIPS*.

Liu, S., Pan, S. J., & Ho, Q. (2017). Distributed multi-task relationship learning. In *Proceedings of SIGKDD*

Ma, C., Jaggi, M., Curtis, F. E., Srebro, N., & Takáč, M. (2017). An accelerated communication-efficient primal–dual optimization framework for structured machine learning. arXiv preprint arXiv:1711.05305.

Ma, C., Smith, V., Jaggi, M., Jordan, M. I., Richtárik, P., & Takác, M. (2015). Adding vs. averaging in distributed primal–dual optimization. In *Proceedings of ICML*.

Ndiaye, E., Fercoq, O., Gramfort, A., & Salmon, J. (2015). Gap safe screening rules for sparse multi-task and multi-class models. In *Proceedings of NIPS*.

Ndiaye, E., Fercoq, O., Gramfort, A., & Salmon, J. (2017). Gap safe screening rules for sparsity enforcing penalties. *Journal of Machine Learning Research*, *18*, 128:1–128:33.

Nesterov, Y. (2013). *Introductory lectures on convex optimization: A basic course*. Berlin: Springer.

Obozinski, G., Taskar, B., & Jordan, M. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, *20*(2), 231–252.

Obozinski, G., Wainwright, M. J., & Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, *39*(1), 1–47.

Shalev-Shwartz, S., & Zhang, T. (2013). Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, *14*(1), 567–599.

Smith, V., Chiang, C., Sanjabi, M., & Talwalkar, A. S. (2017a). Federated multi-task learning. In *Proceedings of NIPS*.

Smith, V., Forte, S., Jordan, M. I., & Jaggi, M. (2015). L1-regularized distributed optimization: A communication-efficient primal–dual framework. CoRR arXiv:1512.04011.

Smith, V., Forte, S., Ma, C., Takáč, M., Jordan, M. I., & Jaggi, M. (2017b). Cocoa: A general framework for communication-efficient distributed optimization. *Journal of Machine Learning Research*, *18*, 230:1–230:49.

Vanhaesebrouck, P., Bellet, A., & Tommasi, M. (2017). Decentralized collaborative learning of personalized models over networks. In *Proceedings of AISTATS*.

Wang, J., Kolar, M., & Srebro, N. (2016). Distributed multi-task learning. In *Proceedings of AISTATS*.

Wang, J., & Ye, J. (2015). Safe screening for multi-task feature learning with multiple data matrices. In *Proceedings of ICML*.

Wang, W., Wang, J., Kolar, M., & Srebro, N. (2018). Distributed stochastic multi-task learning with graph regularization. arXiv preprint arXiv:1802.03830.

Xie, L., Baytas, I. M., Lin, K., & Zhou, J. (2017). Privacy-preserving distributed multi-task learning with asynchronous updates. In *Proceedings of SIGKDD*.

Xing, E. P., Ho, Q., Dai, W., Kim, J. K., Wei, J., Lee, S., et al. (2015). Petuum: A new platform for distributed machine learning on big data. *IEEE Transactions on Big Data*, *1*(2), 49–67.

Yuan, M., Ekici, A., Lu, Z., & Monteiro, R. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, *68*(1), 49–67.

Zhang, C., Zhao, P., Hao, S., Soh, Y. C., Lee, B., Miao, C., et al. (2018). Distributed multi-task classification: A decentralized online learning approach. *Machine Learning*, *107*(4), 727–747.

Zhang, Y., & Xiao, L. (2017). Stochastic primal–dual coordinate method for regularized empirical risk minimization. *Journal of Machine Learning Research*, *18*, 18:1–18:42.

Zhang, Y., & Yang, Q. (2017). A survey on multi-task learning. CoRR arXiv:1707.08114.

Zhang, Y., & Yeung, D. Y. (2010). A convex formulation for learning task relationships in multi-task learning. In *Proceedings of UAI*.

Zheng, S., Wang, J., Xia, F., Xu, W., & Zhang, T. (2017). A general distributed dual coordinate optimization framework for regularized loss minimization. *Journal of Machine Learning Research*, *18*, 115:1–115:52.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, *67*(2), 301–320.