Automatica 162 (2024) 111514

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Distributed empirical risk minimization with differential privacy[☆]

Changxin Liu[a,*], Karl H. Johansson[a], Yang Shi[b]

[a] School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, and Digital Futures, 100 44 Stockholm, Sweden
[b] Department of Mechanical Engineering, University of Victoria, Victoria, B.C. V8W 3P6, Canada

ARTICLE INFO

Article history:
Received 30 November 2022
Received in revised form 30 June 2023
Accepted 7 December 2023
Available online 20 January 2024

Keywords:
Distributed optimization
Empirical risk minimization
Differential privacy
Dual averaging

ABSTRACT

This work studies the distributed empirical risk minimization (ERM) problem under differential privacy (DP) constraint. Standard distributed algorithms achieve DP typically by perturbing all local subgradients with noise, leading to significantly degenerated utility. To tackle this issue, we develop a class of private distributed dual averaging (DDA) algorithms, which activates a fraction of nodes to perform optimization. Such subsampling procedure provably amplifies the DP guarantee, thereby achieving an equivalent level of DP with reduced noise. We prove that the proposed algorithms have utility loss comparable to centralized private algorithms for both general and strongly convex problems. When removing the noise, our algorithm attains the optimal $\mathcal{O}(1/t)$ convergence for non-smooth stochastic optimization. Finally, experimental results on two benchmark datasets are given to verify the effectiveness of the proposed algorithms.

1. Introduction

Consider a group of n nodes, where each node i has a local dataset $D_i = \{\xi_i^{(1)}, \dots, \xi_i^{(q)}\}$ that contains a finite number q of data samples. The nodes are connected via a communication network. They aim to collaboratively solve the empirical risk minimization (ERM) problem, where the machine learning models are trained by minimizing the average of empirical prediction loss over known data samples. Formally, the optimization problem is given by

$$\min_{x\in\mathbb{R}^m} \left\{ F(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + h(x) \right\}, \tag{1}$$

where $f_i(x) = \frac{1}{q}\sum_{j=1}^{q} l_i(x, \xi_i^{(j)})$ represents the empirical risk on node i, $l_i(x, \xi)$ is the loss of the model x over the data instance ξ, and $h(x)$ is the regularization term shared across the nodes. This setup has been commonly considered in machine learning (Lian et al., 2017), where $h(x)$ is used to promote sparsity or model the constraints.

As the loss and its gradient in ERM are characterized by data samples, potential privacy issues arise when the datasets are sensitive (Bassily, Smith, & Thakurta, 2014). In particular, when l_i is the hinge loss, the solution to Problem (1), i.e., support vector machine (SVM), in its dual form typically discloses data points (Bassily et al., 2014). Advanced attacks such as input data reconstruction (Zhu & Han, 2020) and attribute inference (Melis, Song, De Cristofaro, & Shmatikov, 2019) can extract private information from the gradients. To defend privacy attacks, differential privacy (DP) has become prevalent in cryptography and machine learning (Abadi et al., 2016; Dwork, 2006), due to its precise notion and computational simplicity. Informally, DP requires the outcome of an algorithm to remain stable under any possible changes to an individual in the database, and therefore protects individuals from attacks that try to steal the information particular to them. The DP constraint induces a tradeoff between privacy and utility in learning algorithms (Bassily et al., 2014; Chaudhuri, Sarwate, & Sinha, 2012; Kifer, Smith, & Thakurta, 2012; Wang, Ye, & Xu, 2017).

In this work, we are interested in solving Problem (1) while providing rigorous DP guarantee for each data sample in $D := \cup_{i=1}^{n} D_i$.

1.1. Related work

For problems without regularization, i.e., $h \equiv 0$, Huang, Mitra, and Vaidya (2015) developed a differentially private distributed gradient descent (DGD) algorithm by perturbing the local output with Laplace noise. Notably, the learning rate is designed to be linearly decaying such that the sensitivity of the algorithm also

[☆] This work was supported by the Swedish Research Council, Sweden, the Knut and Alice Wallenberg Foundation, Sweden, the Swedish Foundation for Strategic Research, Sweden, and an NSERC Postdoctoral Fellowship. A preliminary version of this work has been reported in the 9th IFAC Workshop on Networked Systems (Liu et al., 2022). This paper was recommended for publication in revised form by Associate Editor André M.H. Teixeira under the direction of Editor Christos G. Cassandras.

[*] Corresponding author.

E-mail addresses: changxin@kth.se (C. Liu), kallej@kth.se (K.H. Johansson), yshi@uvic.ca (Y. Shi).

https://doi.org/10.1016/j.automatica.2024.111514

decreases linearly. Then, one can decompose the prescribed DP parameter ε into a sequence $\{\varepsilon_t\}_{t\geq 1}$, such that $\sum_{\tau=1}^{\infty} \varepsilon_\tau = \varepsilon$ and the operation at each time instant t can be made ε_t-DP. However, such choice of learning rate slows down the convergence dramatically and results in a utility loss in the order of $\mathcal{O}(m/\varepsilon^2)$, where m denotes the dimension of the decision variable. Under the more reasonable learning rate $\Theta(1/\sqrt{t})$, the utility loss can be improved to $\mathcal{O}\left(\sqrt[4]{mn^2/\varepsilon}\right)$ (Han, Topcu, & Pappas, 2016), where n denotes the number of nodes. Along this line of research, Han, Liu, Lin, and Xia (2022), Xiong, Xu, You, Liu, and Wu (2020) and Zhu, Xu, Guan, and Wu (2018) extended the algorithm to time-varying objective functions, and Ding, Zhu, He, Chen, and Guan (2021) advanced the convergence rate to linear based on an additional gradient-tracking scheme. Wang, Zhang, and He (2022) developed a distributed algorithm with DP for stochastic aggregative games. The differentially private distributed optimization problem with coupled equality constraints has been studied in Chen, Chen, Xiang, and Ren (2021). In these works, however, ε-DP is proved only for each iteration, leading to a cumulative privacy loss of $t\varepsilon$ after t iterations. To attenuate the noise effect while ensuring DP, Vlaski and Sayed (2021) constructed topology-aware noise, with which each node perturbs the messages to its neighbors (including itself) with different perturbations whose weighted sum is 0.

For federated learning (FL) with heterogeneous data, Hu, Guo, Li, Pei, and Gong (2020) developed a personalized linear model training algorithm with DP. In Noble, Bellet, and Dieuleveut (2022), general models were considered. In particular, the subsampling of users and local data has been explicitly considered to amplify the DP guarantee and improve the utility.

To tackle regularized learning problems, the alternating direction method of multipliers (ADMM) has been used to design distributed algorithms with DP (Zhang, Ahmad and Wang, 2018; Zhang, Khalili and Liu, 2018; Zhang & Zhu, 2016). However, an explicit tradeoff analysis between privacy and utility was missing. Xiao and Devadas (2021) investigated the privacy guarantee produced not only by random noise injection but also by *mixup* (Zhang, Cisse, Dauphin and Lopez-Paz, 2018), i.e., a random convex combination of inputs. Approximate DP and advanced composition (Kairouz, Oh, & Viswanath, 2015) were used to keep track of the cumulative privacy loss. The privacy–utility tradeoff in linearized ADMM and DGD were captured by the bound $\mathcal{O}\left(m/(\sqrt{n}\varepsilon)\right)$.

To summarize, existing private distributed optimization algorithms applied to Problem (1) typically require each node to make a gradient query to the local dataset at each time instant. Since the sizes of local datasets are considerably smaller than that of the original dataset, local gradient queries have larger sensitivity parameters than that in centralized settings. Therefore, private distributed optimization paradigms in the literature typically employed a larger magnitude of noise to secure the same level of DP, and suffered from relatively low utility. Recently, an asynchronous DGD method with DP was developed in Xu, Zhang, and Wang (2021), which achieved a lower utility loss. The algorithm assumed that each local mini-batch is a subset of data instances uniformly sampled from the overall dataset without replacement, which appears to be restrictive in distributed settings.

1.2. Contribution

We develop a class of differentially private distributed dual averaging (DDA) algorithms for solving Problem (1). At each iteration, a fraction of nodes is activated uniformly at random to perform local stochastic subgradient query and local update with perturbed subgradient. Such subsampling procedure provably amplifies the DP guarantee and therefore helps achieve the same level of DP with weaker noise. To ensure a user-defined level of DP, we provide sufficient conditions on the noise variance in Theorem 1, which admits a smaller bound of variance than existing results.

The properties of the proposed algorithms in terms of convergence and the privacy-utility tradeoff are analyzed. First, a non-asymptotic convergence analysis is conducted for dual averaging with inexact oracles under general choices of hyperparameters, and the results are summarized in Theorem 2. This piece of result illustrates how the lack of global information and the DP noise in private DDA quantitatively affect the convergence, which lays the foundation for subsequent analysis. Then, we investigate the convergence rate of the non-private (noiseless) version of DDA for both strongly convex and general convex objective functions under two sets of hyperparameters in Corollaries 1 and 2, respectively. We remark that Corollary 1 advances the best known convergence rate of DDA for nonsmooth stochastic optimization, i.e., $\mathcal{O}(1/\sqrt{t})$, to $\mathcal{O}(1/t)$. The key to obtaining the improved rate is the use of a new class of parameters.

The privacy–utility tradeoff of the proposed algorithm is examined in Corollaries 3 and 4. In particular, when the objective function is non-smooth and strongly convex, the utility loss is characterized by $\mathcal{O}(m\iota^2/(q^2\varepsilon^2))$, where m, ι, q, ε denote the variable dimension, node sampling ratio, number of samples per node, and DP parameter, respectively. For comparison, we present in Table 1 a comparison of some of the most relevant works.[1]

Finally, we verify the effectiveness of the proposed algorithms via distributed SVM on two open-source datasets. Several comparison results are also presented to support our theoretical findings.

1.3. Outline

The rest of the paper is organized as follows. Section 2 introduces some preliminaries. We present our algorithms and their theoretical properties in Section 3, whose proofs are postponed to Section 4. Some experimental results are given in Section 5. Section 6 concludes the paper.

2. Preliminaries

2.1. Basic setup

We consider the distributed ERM in (1), in which h is a closed convex function with non-empty domain $\text{dom}(h)$. Examples of $h(x)$ include l_1-regularization, i.e., $h(x) = \lambda\|x\|_1$, $\lambda > 0$, and the indicator function of a closed convex set. The regularization term h and the loss functions l_i for all $i = 1, \ldots, n$ satisfy the following assumptions.

Assumption 1. (i) $h(\cdot)$ is a proper closed convex (strongly convex) function with modulus $\mu = 0$ (resp. $\mu > 0$), i.e., for any $x, y \in$ *dom*(h),

$$h(\alpha x + (1-\alpha)y) \leq \alpha h(x) + (1-\alpha)h(y) - \frac{\mu\alpha(1-\alpha)}{2}\|x-y\|^2;$$

(ii) each $l_i(\cdot, \xi_i)$ is convex on *dom*(h).

When $q = 1$, Problem (1) reduces to a deterministic distributed optimization problem. In Problem (1), the information exchange only occurs between connected nodes. Similar to existing research (Duchi, Agarwal, & Wainwright, 2011; Nedic &

[1] Table 1 presents the dependence on variable dimension m, number of nodes n, number of samples q per node, sampling ratio $\iota \propto 1/n$, and ε for utility loss. The work in Xu et al. (2021) considered nonconvex problems, and the results are adapted to convex problems for comparison in Table 1.

Table 1
A comparison of some related works.

	Privacy	Noise	Perturbed	Utility upper bound		Non-smooth
			term	Convex	Strongly convex	regularizer
Noble et al. (2022) (FL)	(ε,δ)-DP	Gaussian	Gradient	$\mathcal{O}\left(\frac{\sqrt{m}}{nq\varepsilon}\right)$	$\mathcal{O}\left(\frac{m}{n^2q^2\varepsilon^2}\right)$	No
Huang et al. (2015)	ε-DP	Laplace	Output	–	$\mathcal{O}(\frac{m}{\varepsilon^2})$	No
Xiao and Devadas (2021) (ADMM)	(ε,δ)-DP	Gaussian	Output	$\mathcal{O}\left(\frac{n^{3/2}\varepsilon}{m}\right)$	–	No
Xiao and Devadas (2021) (DGD)	(ε,δ)-DP	Gaussian	Gradient	$\mathcal{O}(\frac{m}{\sqrt{n}\varepsilon})$	–	No
Xu et al. (2021)	(ε,δ)-DP	Gaussian	Gradient	$\mathcal{O}(\frac{\sqrt{m}}{q\varepsilon})$	–	No
This work	(ε,δ)-DP	Gaussian	Gradient	$\mathcal{O}\left(\frac{\sqrt{m\iota}}{q\varepsilon}\right)$	$\mathcal{O}\left(\frac{m\iota^2}{q^2\varepsilon^2}\right)$	Yes

Ozdaglar, 2009), we use a doubly stochastic matrix $W \in [0,1]^{n\times n}$ to encode the network topology and the weights of connected links at time t. In particular, its (i,j)-th entry, w_{ij}, denotes the weight used by i when counting the message from j. When $w_{ij} = 0$, nodes i and j are disconnected.

2.2. Conventional DDA

The DDA algorithm originally proposed by Duchi et al. (2011) can be applied to solve Problem (1). In particular, let $d(\cdot) \geq 0$ be a strongly convex function with modulus 1 on dom(h). Each node, starting with $z_i^{(1)} = 0$, iteratively generates $\{z_i^{(t)}, x_i^{(t)}\}_{t\geq 1}$ according to

$$x_i^{(t)} = \underset{x\in\mathbb{R}^m}{\operatorname{argmin}} \left\{ \langle z_i^{(t)}, x\rangle + t(h(x)) + \gamma_t d(x) \right\} \tag{2}$$

and

$$z_i^{(t+1)} = \sum_{j=1}^{n} w_{ij}\left(z_j^{(t)} + \hat{g}_j^{(t)}\right), \tag{3}$$

where $\{\gamma_t\}_{t\geq 1}$ is a non-decreasing sequence of parameters, w_{ij} is the (i,j)th entry of matrix W, $\hat{g}_j^{(t)} \in \partial l_j(x_j^{(t)}, \xi_j^{(t)})$ denotes the stochastic subgradient of local loss over $x_j^{(t)}$ with $\xi_j^{(t)}$ uniformly sampled from D_j, and $\partial l_j(x_j^{(t)}, \xi_j^{(t)})$ represents the corresponding subdifferential. Throughout the process, each node only passes z_i to its immediate neighbors and updates x_i according to (2). Existing DDA algorithms, when applied to solve Problem (1), converge as $\mathcal{O}(1/\sqrt{t})$ (Colin, Bellet, Salmon, & Clémençon, 2016; Duchi et al., 2011).

2.3. Threat model and DP

In a distributed optimization algorithm, messages bearing information about the local training data are exchanged among the nodes, which leads to privacy risk. In this work, we consider the following two types of attackers.

- *Honest-but-curious nodes* are assumed to follow the algorithm to perform communication and computation. However, they may record the intermediate results to infer the sensitive information about the other nodes.
- *External eavesdroppers* stealthily listen to the private communications between the nodes.

By collecting the confidential messages, the attackers are able to infer private information about the users (Zhu & Han, 2020). To defend them, we employ tools from DP. Indeed, DP has been recognized as the gold standard in quantifying individual privacy preservation for randomized algorithms. It refers to the property of a randomized algorithm that the presence or absence of an individual in a dataset cannot be distinguished based on the output of the algorithm. Formally, we introduce the following definition of DP for distributed optimization algorithms (Zhang, Khalili et al., 2018).

Definition 1. Consider a communication network, in which each node has its own dataset D_i. Let $\{z_i^{(t)} : i = 1,\ldots,n\}$ denote the set of messages exchanged among the nodes at iteration t. A distributed algorithm satisfies (ε,δ)-DP during T iterations, if for every pair of neighboring datasets $D = \cup_{i=1}^{n} D_i$ and $D' = \cup_{i=1}^{n} D_i'$, and for any set of possible outputs $\mathcal{O}$ during T iterations we have

$$Pr[\{z_i^{(t)}, i = 1,\ldots,n\}_{t=1}^{T} \in \mathcal{O}|D]$$
$$\leq e^{\varepsilon} Pr[\{z_i^{(t)}, i = 1,\ldots,n\}_{t=1}^{T} \in \mathcal{O}|D'] + \delta.$$

3. Differentially private DDA algorithm

In this section, we develop the differentially private DDA algorithm, followed by its privacy-preserving and convergence properties.

3.1. Node subsampling in distributed optimization

As explained in Section 1, parallelized local gradient queries in distributed optimization necessitate stronger noise to achieve DP and therefore deteriorate utility. To circumvent this problem, we only activate a random fraction of the nodes at each time instant to perform averaging and local optimization. This allows us to amplify the privacy of the algorithm, and thereby achieving the same level of DP with noise weaker than in existing works.

Definition 2. For every $t \geq 1$, an integer number of $n\iota$ nodes are sampled uniformly at random with some $\iota \in (0,1]$.

The sampling procedure gives rise to a time-varying stochastic communication network. Slightly adjusted to the notation in Section 2.1, we let $W^{(t)} \in [0,1]^{n\times n}$ be a random gossip matrix at time t, where the (i,j)th entry, $w_{ij}^{(t)}$, denotes the weight of the link (i,j) at time t. Denote by $\mathcal{N}^{(t)}$ and $\mathcal{N}_i^{(t)} := \{j | j \neq i, w_{ij}^{(t)} > 0\}$ the set of activated nodes and the set of i's neighbors at time t, respectively. It is worthwhile to point out that $W^{(t)}$ and ι are dependent. That is, we have $w_{ij}^{(t)} > 0$ for $i \in \mathcal{N}^{(t)}$ and $j \in \mathcal{N}_i^{(t)}$, and $w_{ij}^{(t)} = 0$ otherwise.

For the gossip matrix $W^{(t)}$, we assume the following standard condition (Liu, Zhou, Pei, Zhang and Shi, 2022).

Assumption 2. For every $t \geq 1$, (i) $W^{(t)}$ is doubly stochastic[2]; (ii) $W^{(t)}$ is independent of the random events that occur up to time $t-1$; and (iii) there exists a constant $\beta \in (0,1)$ such that

$$\sqrt{\rho\left(\mathbb{E}_W\left[W^{(t)^T}W^{(t)}\right] - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)} \leq \beta, \tag{4}$$

[2] $W^{(t)}\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^T W^{(t)} = \mathbf{1}^T$ where $\mathbf{1}$ denotes the all-one vector of dimensionality n.

Algorithm 1 Differentially Private DDA

```
   Input: $\mu \geq 0$, $a > 0$, a strongly convex function $d$ with modulus 1 on dom($h$), and $T > 0$
   Output: $\tilde{x}_i^{(T)} = A_T^{-1} \sum_{t=1}^{T} a_t x_i^{(t)}$
1: Initialize: set $z_i^{(1)} = 0$ and identify $x_i^{(1)}$ according to (6) for all $i = 1, \ldots, n$
2: for $t = 1, 2, \ldots, T$ do
     For active nodes $i \in \mathcal{N}^{(t)}$:
3:     $\xi_i^{(t)} \sim$ Uniform$\{1, \ldots, q\}$ and $\nu_i(t) \sim \mathcal{N}(0, \sigma^2 I)$
4:     release $\hat{g}_i^{(t)} + \nu_i^{(t)}$
5:     collect $z_j^{(t)} + a_t(\hat{g}_j^{(t)} + \nu_j^{(t)})$ from all nodes $j \in \mathcal{N}_i^{(t)}$
6:     update $z_i^{(t+1)}$ by (5)
7:     compute $x_i^{(t+1)}$ by (6)
     For inactive nodes $i \notin \mathcal{N}^{(t)}$:
8:     set $z_i^{(t+1)} = z_i^{(t)}$ and $x_i^{(t+1)} = x_i^{(t)}$
9: end for
```

where $\rho(\cdot)$ denotes the spectral radius and the expectation $\mathbb{E}[\cdot]$ is taken with respect to the distribution of $W^{(t)}$ at time t.

3.2. Private DDA with stochastic subgradient perturbation

Next, we introduce a differentially private DDA algorithm presented as Algorithm 1.

The update for the local dual variable $z_i^{(t)}$ reads

$$z_i^{(t+1)} = \sum_{j=1}^{n} w_{ij}^{(t)} \left(z_j^{(t)} + a_t \eta_j^{(t)} \zeta_j^{(t)} \right), \tag{5}$$

where $\zeta_j^{(t)} = \hat{g}_j^{(t)} + \nu_j^{(t)}$, $\nu_j^{(t)} \sim \mathcal{N}(0, \sigma^2 I)$, $\hat{g}_j^{(t)} \in \partial l_j(x_j^{(t)}, \xi_j^{(t)})$ with $\xi_j^{(t)}$ uniformly sampled from D_j, $\eta_i^{(t)}$ indicates whether node i performs local update at time t, i.e.,

$$\eta_i^{(t)} = \begin{cases} 1, & \text{if } i \text{ active} \\ 0, & \text{otherwise} \end{cases}$$

and $\{a_t > 0\}_{t \geq 1}$ is a sequence of non-decreasing parameters. The non-decreasing property of $\{a_t\}_{t \geq 1}$ is motivated by that, when the objective exhibits some desirable properties, e.g., strong convexity, assigning heavier weights to fresher subgradients can speed up convergence (Lu, Freund, & Nesterov, 2018; Tao, Li, Pan, & Tao, 2021). In the special case where $a_t = 1$, $\eta_i^{(t)} = 1$ and $\sigma = 0$, (5) reduces to the conventional update in (3).

Equipped with (5), node i, active at time t, can perform a local computation to derive its estimate about the global optimum:

$$x_i^{(t+1)} = \underset{x \in \mathbb{R}^m}{\operatorname{argmin}} \left\{ \langle z_i^{(t+1)}, x \rangle + \iota A_{t+1} h(x) + \gamma_{t+1} d(x) \right\}, \tag{6}$$

where ι is defined in Definition 2, $A_t = \sum_{\tau=1}^{t} a_\tau$ and $\{\gamma_t\}_{t \geq 1}$ is a non-decreasing sequence of positive parameters. By convention, we let $A_0 = a_0 = 0$ and $\gamma_0 = 0$.

For general regularization $h(x)$, the update in (6) requires the knowledge of ι. This requirement is necessary due to technical reasons. More precisely, due to node sampling, the term $\langle z_i^{(t+1)}, x \rangle$ in (6) serves as a linear approximation of $\iota f_i(x)/n$ rather than $f_i(x)/n$ in standard DDA (Duchi et al., 2011). Thus, one scales up $h(x)$ also with ι in (6) in order to solve the original problem in (1). In the special case where $h(x)$ is the indicator function of a convex set, the knowledge of ι is not needed since $\iota h(x) \equiv h(x)$.

The overall procedure is summarized in Algorithm 1. Each node $i = 1, \ldots, n$ initializes $z_i^{(1)} = 0$ in Step 1. At each time instant t, only active nodes $i \in \mathcal{N}^{(t)}$ update $z_i^{(t+1)}$ and $x_i^{(t+1)}$ by following Steps 3–7. In particular, each active node computes and then perturbs the local stochastic subgradient in Step 3 and 4, respectively, followed by the information exchange with neighboring nodes in Step 5. Then, $z_i^{(t+1)}$ and $x_i^{(t+1)}$ are updated in Steps 6 and 7. For inactive nodes at each time instant t, they simply set $z_i^{(t+1)} = z_i^{(t)}$ and $x_i^{(t+1)} = x_i^{(t)}$.

Remark 1. There are two common approaches to achieve DP for optimization methods. The first type disturbs the output of a non-private algorithm (Zhang, Zheng, Mou, & Wang, 2017), and the second type perturbs the subgradient (Bassily et al., 2014; Wang et al., 2017). The former involves recursively estimating the (time-varying) sensitivity of updates (Wang & Nedić, 2023). This makes the propagation of DP noise and its effect on convergence difficult to quantify (Wang & Nedić, 2023). In this work, we adopt the latter approach in Algorithm 1, where we introduce Gaussian noise to perturb the stochastic subgradient $\hat{g}_i$. By leveraging the time-invariant sensitivity of the gradient query, we can effectively conduct both privacy and utility analyses in the presence of non-smooth regularization. It is worth noting that, in this scenario, the step-size scheduling rule allows for control over utility.

3.3. Privacy analysis

To establish the privacy-preserving property of Algorithm 1, we make the following assumption.

Assumption 3. Each $l_i(\cdot, \xi_i)$ is L-Lipschitz, i.e., for any $x, y \in$ *dom*(h),

$$|l_i(x, \xi_i) - l_i(y, \xi_i)| \leq L\|x - y\|.$$

By Assumption 3, we readily have that each $f_i(\cdot)$ is L-Lipschitz. Next, we state the privacy guarantee for Algorithm 1. The proof can be found in Section 4.1.

Theorem 1. *Suppose Assumption 3 is satisfied, and a random fraction of the nodes with ratio $\iota \in (0, 1]$ is active at each time instant. Given parameters q, $\varepsilon \in (0, 1]$, and $\delta_0 \in (0, 1]$, if*

$$\sigma^2 \geq \frac{32\iota^2 L^2 T \log(2/\delta_0)}{q^2 \varepsilon^2}$$

and $T \geq 5q^2\varepsilon^2/(4\iota^2)$, then Algorithm 1 is (ε, δ)-DP for some constant $\delta \in (0, 1]$.

Remark 2. A few remarks on the results in Theorem 1 are in order:

(i) It can be verified from the proof of Theorem 1 that Algorithm 1 is $(\hat{\varepsilon}, \delta)$-DP after $t \leq T$ iterations with

$$\hat{\varepsilon} \leq \sqrt{\frac{3t}{5T}}\varepsilon + \frac{t}{5T}\varepsilon^2. \tag{7}$$

(ii) Theorem 1 emphasizes that, to achieve a prescribed privacy budget during T iterations, the noise variance σ^2 is related to the DP parameters (ε, δ), the Lipschitz constant L of the loss, the number of samples per local dataset, and the iteration number T. Notably, the lower bound for variance is weighted by $\iota^2 \leq 1$, meaning that the same level of DP can be achieved with reduced noise.

3.4. Privacy–utility tradeoff

Next, we perform a non-asymptotic analysis of Algorithm 1, followed by an explicit privacy–utility tradeoff.

Motivated by Duchi et al. (2011), we define an auxiliary sequence of variables:

$$y^{(t+1)} = \operatorname*{argmin}_{x\in\mathbb{R}^m} \left\{ \langle \overline{z}^{(t+1)}, x\rangle + \iota A_{t+1}h(x) + \gamma_{t+1}d(x) \right\}, \tag{8}$$

where $\overline{z}^{(t)} = \frac{1}{n}\sum_{i=1}^{n} z_i^{(t)}$ and $\{z_i^{(t)} : i = 1, \ldots, n\}_{t\geq 1}$ are generated by Algorithm 1. The convergence property of $y^{(t)}$ is summarized in Theorem 2, whose proof is provided in Section 4.2.

Theorem 2. *Suppose Assumptions 1, 2, and 3 are satisfied. For all $t \geq 1$, we have*

$$A_t\mathbb{E}[F(\tilde{y}^{(t)}) - F(x^*)]$$
$$\leq \frac{\gamma_t}{\iota}d(x^*) + \sum_{\tau=1}^{t} \frac{a_\tau^2}{\mu\iota A_\tau + \gamma_\tau}\left(M + \frac{m\iota\sigma^2}{2} + \frac{2\sqrt{m\iota}L\sigma}{1-\beta}\right), \tag{9}$$

where σ is defined in Theorem 1, $\tilde{y}^{(t)} = A_t^{-1}\sum_{\tau=1}^{t} a_\tau y^{(\tau)}$, $M = \iota L^2/2 + 2\sqrt{\iota}L^2/(1-\beta)$, and the expectation is over the randomness of the algorithm.

From the error bound in (9), we observe that the last two terms are contributed by the noise. How the noise affects the error bound is determined in part by the hyperparameters of the algorithm. Next, we first investigate the choices of a_t that lead to optimal convergence rates for Algorithm 1 with $\sigma = 0$; the results for strongly convex and general convex functions are presented in Corollaries 1 and 2, whose proofs are given in Appendices C and D, respectively.

Corollary 1. *Suppose Assumptions 2 and 3 are satisfied. In addition, Assumption 1 holds with $\mu > 0$, i.e., $h(x)$ is μ-strongly convex. If $\sigma = 0$,*

$$a_t = t \quad \text{and} \quad \gamma_t = 0,$$

then for all $t \geq 1$, and $i = 1, \ldots, n$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\tilde{x}_i^{(t)} - x^*\|^2] \leq \frac{16}{t+1}\left(\frac{L^2(\log t + 1)}{\mu^2\iota(1-\beta)^2 t} + \frac{M}{\mu^2\iota}\right), \tag{10}$$

where $\tilde{x}_i^{(t)} = A_t^{-1}\sum_{\tau=1}^{t} a_\tau x_i^{(\tau)}$ and M is a positive constant given in Theorem 2.

Remark 3. Corollary 1 indicates that the non-private version of Algorithm 1, i.e., $\sigma = 0$, attains the optimal convergence rate $\mathcal{O}(1/t)$ when Problem (1) is strongly convex. Compared to the algorithm in Yuan, Hong, Ho, and Jiang (2018), where the authors focused on constrained problems, the proposed algorithm handles general non-smooth regularizers. Furthermore, the results can be extended to the case where each $f_i(x)$ but not $h(x)$ is strongly convex by following a similar idea in Liu, Zhou et al. (2022).

Corollary 2. *Suppose Assumptions 2 and 3 are satisfied. In addition, Assumption 1 holds with $\mu = 0$, i.e., $h(x)$ is general convex. If $\sigma = 0$,*

$$a_t = 1 \quad \text{and} \quad \gamma_t = \gamma\sqrt{t},$$

then for all $t \geq 1$, and $i = 1, \ldots, n$, we have

$$\mathbb{E}\left[F(\tilde{y}^{(t)}) - F(x^*)\right] \leq \frac{d(x^*) + 2\iota M}{\iota\gamma\sqrt{t}}, \tag{11}$$

where $\tilde{y}^{(t)} = t^{-1}\sum_{\tau=1}^{t} y^{(\tau)}$ and M is a positive constant given in Theorem 2. In addition, for all $t \geq 1$, and $i = 1, \ldots, n$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\tilde{x}_i^{(t)} - \tilde{y}^{(t)}\|] \leq \frac{2L\sqrt{\iota}}{\gamma(1-\beta)\sqrt{t}}, \tag{12}$$

where $\tilde{x}_i^{(t)} = t^{-1}\sum_{\tau=1}^{t} x_i^{(\tau)}$.

Under the same hyperparameters, we study the privacy–utility tradeoff of Algorithm 1 with $\sigma \neq 0$ for strongly convex and general convex functions in Corollaries 3 and 4, whose proofs are presented in Appendices E and F, respectively.

Corollary 3. *Suppose Assumptions 2 and 3 are satisfied. In addition, Assumption 1 holds with $\mu > 0$, i.e., $h(x)$ is μ-strongly convex. If*

$$a_t = t, \quad \gamma_t = 0,$$

and $\beta \in (1 - \sqrt{1/e}, 1]$, then for $T = \mathcal{O}\left(\frac{q^2\varepsilon^2}{\iota^3(1-\beta)^2 m\log(1/\delta_0)}\right)$ and $i = 1, \ldots, n$:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\tilde{x}_i^{(T)} - x^*\|^2\right] \leq \mathcal{O}\left(\frac{m\iota^2L^2\log(1/\delta_0)}{\mu^2q^2\varepsilon^2}\right), \tag{13}$$

where $\tilde{x}_i^{(T)} = A_T^{-1}\sum_{t=1}^{T} a_t x_i^{(t)}$, and δ_0 is defined in Theorem 1.

Corollary 4. *Suppose Assumptions 2 and 3 are satisfied. In addition, Assumption 1 holds with $\mu = 0$, i.e., $h(x)$ is general convex. If*

$$a_t = 1, \quad \gamma_t = \gamma\sqrt{t}, \tag{14}$$

and $\iota \leq 1 - \beta$, then the following holds for $T = \mathcal{O}\left(\frac{q^2\varepsilon^2}{\iota^3(1-\beta)^2 m\log(1/\delta_0)}\right)$:

$$\mathbb{E}[F(\tilde{y}^{(T)}) - F(x^*)] \leq \mathcal{O}\left(\frac{\left(L^2 + d(x^*)\right)\sqrt{m\iota\log(1/\delta_0)}}{\gamma q\varepsilon}\right) \tag{15}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\tilde{x}_i^{(T)} - \tilde{y}^{(T)}\|] \leq \frac{L\sqrt{m\iota\log(1/\delta_0)}}{\gamma q\varepsilon}$$

for $i = 1, \ldots, n$, where $\tilde{y}^{(T)} = \frac{1}{T}\sum_{t=1}^{T} y^{(t)}$, $\tilde{x}_i^{(T)} = \frac{1}{T}\sum_{t=1}^{T} x_i^{(t)}$, and δ_0 is defined in Theorem 1.

Corollaries 3 and 4 highlight that the sampling procedure lowers down the utility loss for both strongly convex and general convex problems. In particular, the utility loss in the strongly convex case becomes $\iota^2 \approx 1/n^2$ times smaller than that without sampling. For general convex problems, the utility loss is $\sqrt{\iota}$ times smaller. They also suggest that the number of iterations increases in order to achieve a lower utility loss.

4. Proofs of main results

This section presents the proof of Theorems 1 and 2.

4.1. Proof of Theorem 1

We start by introducing some useful properties of DP (Dwork, 2006; Girgis, Data, Diggavi, Kairouz, & Suresh, 2021; Kairouz et al., 2015).

Lemma 1 (*Gaussian Mechanism*)**.** *Consider the Gaussian mechanism for answering the query $r : \mathcal{D} \to \mathbb{R}^m$:*

$$\mathcal{M} = r(D) + \nu,$$

where $D \in \mathcal{D}$, $\nu \sim \mathcal{N}(0, \sigma^2 I)$. The mechanism $\mathcal{M}$ is $(\sqrt{2\log(2/\delta)}\,\Delta/\sigma, \delta)$-DP where Δ denotes the sensitivity of r, i.e., $\Delta = \sup_{D,D'}\|r(D) - r(D')\|$.

Recall from Algorithm 1 that, at each iteration, $n\iota$ nodes are sampled from n nodes at random, and each activated node randomly selects a data sample from q instances to compute stochastic gradients. Although such a sub-sampling is not uniform, i.e., the subsets of $n\iota$ data samples are not necessarily chosen with equal probability, it still helps amplify the privacy (Girgis et al., 2021, Lemma 10).

Lemma 2 (*Privacy Amplification by Subsampling*)**.** *Suppose $\mathcal{M}$ is an (ε, δ)-DP mechanism. Let $samp_{r_1,r_2} : \mathcal{U}^{r_1} \to \mathcal{U}^{r_2}$ be the subsampling operation that takes a dataset belonging to $\mathcal{U}^{r_1}$ as input and selects uniformly at random a subset of $r_2 \le r_1$ elements from the input dataset. Then, the mechanism*

$$\mathcal{A}(\mathcal{D}) = samp_{n,n\iota}(\mathcal{G}_1, \ldots, \mathcal{G}_n)$$

where $\mathcal{G}_i = samp_{q,1}(\mathcal{M}(\hat{g}_i(\xi_i^{(1)})), \ldots, \mathcal{M}(\hat{g}_i(\xi_i^{(q)})))$ and $\hat{g}_i(\xi)$ is a subgradient of l_i evaluated over data point ξ is $(\ln(1 + \iota(e^{\varepsilon} - 1)/q), \iota\delta/q)$-DP,

Lemma 3 (*Composition of DP*)**.** *Given T randomized algorithms $\mathcal{A}_1, \ldots, \mathcal{A}_\tau, \ldots, \mathcal{A}_T : \mathcal{D} \to \mathcal{R}$, each of which is $(\varepsilon_i, \delta_i)$-DP with $\varepsilon_i \in (0, 0.9]$ and $\delta_i \in (0, 1]$. Then $\mathcal{A} : \mathcal{D} \to \mathcal{R}^t$ with $\mathcal{A}(\mathcal{D}) = (\mathcal{A}_1(\mathcal{D}), \ldots, \mathcal{A}_t(\mathcal{D}))$ is $(\tilde{\varepsilon}, \tilde{\delta})$-DP with*

$$\tilde{\varepsilon} = \sqrt{\sum_{i=1}^{k} 2\varepsilon_i^2 \log(e + \frac{\sqrt{\sum_{i=1}^{k} \varepsilon_i^2}}{\delta'}) + \sum_{i=1}^{k} \varepsilon_i^2}$$

and

$$\tilde{\delta} = 1 - (1 - \delta') \prod_{i=1}^{k} (1 - \delta_i)$$

for any $\delta' \in (0, 1]$.

Lemma 4 (*Post-Processing*)**.** *Given a randomized algorithm $\mathcal{A}$ that is (ε, δ)-DP. For arbitrary mapping p from the set of possible outputs of $\mathcal{A}$ to an arbitrary set, $p(\mathcal{A}(\cdot))$ is (ε, δ)-DP.*

We are now in a position to prove Theorem 1.

DP at each time t: We begin by noting that the subgradient perturbation procedure at time t, denoted by $\mathcal{M}_t$, is a Gaussian mechanism whose sensitivity, by Assumption 3, is $\Delta \le 2L$. Based on Lemma 1, $\mathcal{M}_t$ is $(\varepsilon_t, \delta_0)$-DP with

$$\varepsilon_t = \frac{2L\sqrt{2\log(2/\delta_0)}}{\sigma}$$

for any $\delta_0 \in [0, 1]$. Due to the conditions on σ and T, we obtain

$$\varepsilon_t^2 = \frac{8L^2 \log(2/\delta_0)}{\sigma^2} = \frac{q^2\varepsilon^2}{4\iota^2 T} \le 0.2. \tag{16}$$

Denote by $\mathcal{A}_t$ the composition of $\mathcal{M}_t$ and the subsampling procedure. Upon using Lemma 2 and (16), we obtain that $\mathcal{A}_t$ is $(\varepsilon'_t, \iota\delta_0/q)$-DP with

$$\varepsilon'_t = \frac{2\iota\varepsilon_t}{q} \ge \frac{\iota(e^{\varepsilon_t} - 1)}{q} \ge \ln\left(1 + \frac{\iota(e^{\varepsilon_t} - 1)}{q}\right).$$

In addition, because of (16), we get $\varepsilon'_t = 2\iota\varepsilon_t/q \le 2\varepsilon_t \le 0.9$ and

$$\sum_{t=1}^{T} \varepsilon_t'^2 = \frac{4\iota^2}{q^2} \sum_{t=1}^{T} \frac{8L^2 \log(2/\delta_0)}{\sigma^2} = \sum_{t=1}^{T} \frac{\varepsilon^2}{T} \le 1. \tag{17}$$

DP after T **iterations**: Consider the composition of $\mathcal{A}_1, \ldots, \mathcal{A}_\tau, \ldots, \mathcal{A}_T$, denoted by $\mathcal{A}$. Based on the advanced composition rule for DP in Lemma 3, we obtain $\mathcal{A}$ is $(\tilde{\varepsilon}, \tilde{\delta})$-DP with

$$\tilde{\varepsilon} = \sqrt{2\sum_{t=1}^{T} \varepsilon_t'^2 \log(e + \frac{\sqrt{\sum_{t=1}^{T} \varepsilon_t'^2}}{\delta'}) + \sum_{t=1}^{T} \varepsilon_t'^2}$$

and $\tilde{\delta} = 1 - (1 - \delta')(1 - \iota\delta_0)^T$ for any $\delta' \in (0, 1]$. Furthermore, there holds

$$\begin{aligned}
\tilde{\varepsilon} &\le \sqrt{2\sum_{t=1}^{T} \varepsilon_t'^2 \log(e + \frac{\sqrt{\sum_{t=1}^{T} \varepsilon_t'^2}}{\delta'}) + \frac{1}{5}\varepsilon^2} \\
&\le \sqrt{3\sum_{t=1}^{T} \varepsilon_t'^2 + \frac{1}{5}\varepsilon} = \sqrt{\frac{3}{5}\varepsilon^2 + \frac{1}{5}\varepsilon} \\
&\le \varepsilon,
\end{aligned}$$

where we fix $\delta' = \sqrt{\sum_{t=1}^{T} \varepsilon_t'^2} \le 1$ and use (17) to get the second inequality. By setting $\delta = \tilde{\delta}$, we have that $\mathcal{A}$ is (ε, δ)-DP.

DP after postprocessing: The intermediate results $\{\mathbf{z}^{(\tau)}\}_{\tau=1}^{T}$ are computed based on the output of $\mathcal{A}$, i.e., perturbed subgradients. By the post-processing property of DP in Lemma 4, Algorithm 1 also satisfies (ε, δ)-DP specified in Definition 1.

4.2. Proof of Theorem 2

Before proving Theorem 2, we present two useful lemmas whose proofs are given in Appendices A and B.

Lemma 5. *For the sequence $\{x_i^{(t)} : i = 1, \ldots, n\}_{t \ge 1}$ generated by Algorithm 1 and the auxiliary sequence $\{y^{(t)}\}_{t \ge 1}$ defined in (8), one has that for all $t \ge 1$ and $i = 1, \ldots, n$,*

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\|x_i^{(t)} - y^{(t)}\|\right] \le \frac{a_t(L + \sqrt{m}\sigma)\sqrt{\iota}}{(1 - \beta)(\mu\iota A_t + \gamma_t)} \tag{18}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\|x_i^{(t)} - y^{(t)}\|^2\right] \le \frac{a_t^2(L^2 + m\sigma^2)\iota}{(1 - \beta)^2(\mu\iota A_t + \gamma_t)^2}. \tag{19}$$

Lemma 6. *For all $t \ge 1$, we have*

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} \sum_{\tau=1}^{t} a_\tau \left(\frac{1}{\iota}\left\langle \eta_i^{(\tau)}\zeta_i^{(\tau)}, y^{(\tau)} - x^*\right\rangle + h(y^{(\tau)}) - h(x^*)\right) \\
&\le \frac{1}{2} \sum_{\tau=1}^{t} \frac{a_\tau^2}{\iota\,(\mu\iota A_\tau + \gamma_\tau)} \left\|\frac{1}{n} \sum_{i=1}^{n} \eta_i^{(\tau)}\zeta_i^{(\tau)}\right\|^2 + \frac{\gamma_t}{\iota} d(x^*).
\end{aligned} \tag{20}$$

Now we are ready to prove Theorem 2. Upon using $A_t = \sum_{\tau=1}^{t} a_\tau$, convexity of $f := n^{-1} \sum_{i=1}^{n} f_i$, and L-Lipschitz continuity of each f_j, we have

$$\begin{aligned}
&A_t \left(f(\tilde{y}^{(t)}) - f(x^*)\right) \\
&\le \sum_{\tau=1}^{t} a_\tau \left(f(y^{(\tau)}) - f(x^*)\right) \\
&= \frac{1}{n} \sum_{\tau=1}^{t} \sum_{j=1}^{n} a_\tau \left(f_j(y^{(\tau)}) - f_j(x_j^{(\tau)}) + f_j(x_j^{(\tau)}) - f_j(x^*)\right) \\
&\le \frac{1}{n} \sum_{\tau=1}^{t} \sum_{j=1}^{n} a_\tau L\|y^{(\tau)} - x_j^{(\tau)}\| + \frac{1}{n} \sum_{\tau=1}^{t} \phi^{(\tau)},
\end{aligned} \tag{21}$$

where $\phi^{(\tau)} = \sum_{j=1}^n a_\tau \left(f_j(x_j^{(\tau)}) - f_j(x^*)\right)$. Further using convexity of $f_j, j = 1, \ldots, n$, we have

$$
\begin{aligned}
\phi^{(\tau)} &\le \sum_{j=1}^n a_\tau \left\langle g_j^{(\tau)}, x_j^{(\tau)} - x^* \right\rangle \\
&= n a_\tau \left\langle \overline{\zeta}^{(\tau)}, y^{(\tau)} - x^* \right\rangle + \sum_{j=1}^n a_\tau \left(\left\langle g_j^{(\tau)}, x_j^{(\tau)} - y^{(\tau)} \right\rangle \right. \\
&\quad \left. + \left\langle g_j^{(\tau)} - \zeta_j^{(\tau)}, y^{(\tau)} - x^* \right\rangle \right),
\end{aligned} \tag{22}
$$

where $\overline{\zeta}^{(t)} = n^{-1} \sum_{i=1}^n \zeta_i^{(t)}$. Therefore,

$$
\begin{aligned}
& A_t \left(F(\tilde{y}^{(t)}) - F(x^*)\right) \\
&\le A_t \left(f(\tilde{y}^{(t)}) - f(x^*)\right) + \sum_{\tau=1}^t a_\tau \left(h(y^{(\tau)}) - h(x^*)\right) \\
&\le \sum_{\tau=1}^t a_\tau \left(\frac{2}{n} \sum_{j=1}^n L \|x_j^{(\tau)} - y^{(\tau)}\| + h(y^{(\tau)}) - h(x^*) \right) \\
&\quad + \sum_{\tau=1}^t a_\tau \left\langle \overline{\zeta}^{(\tau)}, y^{(\tau)} - x^* \right\rangle \\
&\quad + \frac{1}{n} \sum_{\tau=1}^t \sum_{j=1}^n a_\tau \left\langle g_j^{(\tau)} - \zeta_j^{(\tau)}, y^{(\tau)} - x^* \right\rangle,
\end{aligned} \tag{23}
$$

where we use

$$
\sum_{\tau=1}^t a_\tau \left(h(x^*) - h(y^{(\tau)})\right) \le A_t \left(h(x^*) - h(\tilde{y}^{(t)})\right)
$$

and $F = f + h$ in the first inequality, and use (21), (22) in the last inequality. Due to uniform node sampling with probability ι, we have

$$
\frac{1}{n} \mathbb{E}_\tau \left[\sum_{i \in \mathcal{N}^{(t)}} \zeta_i^{(\tau)} \right] = \frac{\iota}{n} \sum_{i=1}^n \mathbb{E}_\tau \left[\zeta_i^{(\tau)} \right] = \iota \mathbb{E}_\tau \left[\overline{\zeta}^{(\tau)} \right],
$$

where $\mathbb{E}_\tau$ denotes expectation conditioned on $\{x_i^{(\tau)}, i = 1, \ldots, n\}$. Therefore, by putting the conditioned expectation on (23) and using the law of total expectation, we obtain

$$
\begin{aligned}
& A_t \mathbb{E}\left[F(\tilde{y}^{(t)}) - F(x^*)\right] \\
&\le \sum_{\tau=1}^t a_\tau \left(\frac{2}{n} \sum_{j=1}^n L \mathbb{E}[\|x_j^{(\tau)} - y^{(\tau)}\|] + \mathbb{E}[h(y^{(\tau)}) - h(x^*)] \right) \\
&\quad + \frac{1}{\iota} \sum_{\tau=1}^t a_\tau \mathbb{E}\left[\left\langle \frac{1}{n} \sum_{i=1}^n \eta_i^{(\tau)} \zeta_i^{(\tau)}, y^{(\tau)} - x^* \right\rangle \right] \\
&\quad + \frac{1}{n} \sum_{\tau=1}^t \sum_{j=1}^n a_\tau \mathbb{E}\left[\left\langle g_j^{(\tau)} - \zeta_j^{(\tau)}, y^{(\tau)} - x^* \right\rangle \right].
\end{aligned}
$$

Since $\nu_j^{(\tau)}$ and $\hat{g}_j^{(\tau)}$ are independent of $y^{(\tau)}$ and $\mathbb{E}[\hat{g}_j^{(\tau)}] = g_j^{(\tau)}$, we have

$$
\begin{aligned}
& \mathbb{E}\left[\left\langle g_j^{(\tau)} - \zeta_j^{(\tau)}, y^{(\tau)} - x^* \right\rangle \right] \\
&= \mathbb{E}\left[\left\langle g_j^{(\tau)} - \hat{g}_j^{(\tau)}, y^{(\tau)} - x^* \right\rangle \right] + \mathbb{E}\left[\left\langle -\nu_j^{(\tau)}, y^{(\tau)} - x^* \right\rangle \right] \\
&= 0.
\end{aligned} \tag{24}
$$

Therefore, we obtain from Lemma 6 that

$$
\begin{aligned}
& A_t \mathbb{E}[F(\tilde{y}^{(t)}) - F(x^*)] \\
&\le \frac{1}{2} \sum_{\tau=1}^t \frac{a_\tau^2}{\iota\,(\mu \iota A_\tau + \gamma_\tau)} \mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^n \eta_i^{(\tau)} \zeta_i^{(\tau)} \right\|^2 \right] \\
&\quad + \frac{\gamma_t}{\iota} d(x^*) + \frac{2}{n} \sum_{\tau=1}^t \sum_{j=1}^n L a_\tau \mathbb{E}\left[\|x_j^{(\tau)} - y^{(\tau)}\| \right].
\end{aligned} \tag{25}
$$

Furthermore, we have

$$
\left\| \frac{1}{n} \sum_{i=1}^n \eta_i^{(\tau)} \zeta_i^{(\tau)} \right\|^2 = \frac{1}{n^2} \left\| \sum_{i \in \mathcal{N}^{(t)}} \zeta_i^{(\tau)} \right\|^2 \le \frac{\iota}{n} \sum_{i \in \mathcal{N}^{(t)}} \left\| \zeta_i^{(\tau)} \right\|^2.
$$

Since

$$
\mathbb{E}_\tau \left[\sum_{i \in \mathcal{N}^{(t)}} \left\| \zeta_i^{(\tau)} \right\|^2 \right] = \iota \sum_{i=1}^n \mathbb{E}_\tau \left[\left\| \zeta_i^{(\tau)} \right\|^2 \right],
$$

we remove the conditioning based on the law of total expectation to obtain

$$
\begin{aligned}
& \mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^n \eta_i^{(\tau)} \zeta_i^{(\tau)} \right\|^2 \right] \le \frac{\iota^2}{n} \sum_{i=1}^n \mathbb{E}\left[\left\| \zeta_i^{(\tau)} \right\|^2 \right] \\
&\le \frac{\iota^2}{n} \sum_{i=1}^n \left(\mathbb{E}\left[\left\| \hat{g}_i^{(\tau)} \right\|^2 \right] + \mathbb{E}\left[\left\| \nu_i^{(\tau)} \right\|^2 \right] \right) \\
&\le \iota^2 (L^2 + m\sigma^2),
\end{aligned}
$$

where we use the fact that $\nu_i^{(\tau)}$ is independent of $\hat{g}_i^{(\tau)}$. By plugging the above into (25) and using Lemma 5, we arrive at (9) as desired.

5. Experiments

In this section, we present experimental results of the proposed algorithms.

5.1. Setup

We use the benchmark datasets *epsilon* (Sonnenburg, Rätsch, Schäfer, & Schölkopf, 2006) and *rcv1* (Lewis, Yang, Russell-Rose, & Li, 2004) in the experiments. Some information about the datasets is given in Table 2. We randomly assign the data samples evenly among the $n = 20$ working nodes. The working nodes aim to solve the following regularized SVM problem:

$$
\min_x \left\{ F(x) = \frac{1}{n} \sum_{i=1}^n f_i(x) + h(x) \right\}, \tag{26}
$$

where

$$
f_i(x) = \frac{1}{q} \sum_{j=1}^q \max \left\{ 0, 1 - y_i^{(j)} \left\langle C_i^{(j)}, x \right\rangle \right\}, \tag{27}
$$

$\{C_i^{(j)}, y_i^{(j)}\}_{j=1}^q := D_i$ are data samples private to node i. In the experiment, we consider two choices of the regularizer, i.e., $h(x) = \phi \|x\|_1$ and $h(x) = \mu \|x\|_2^2$ where $\phi > 0$ and $\mu > 0$ will be specified later.

Throughout the experiments, we consider a complete graph with $n = 20$ nodes as the supergraph. Based on it, we consider two edge sampling strategies, that is, 1 or 2 edges are sampled uniformly at random from the set of all edges at each time instant. The corresponding gossip matrices are created with Metropolis weights (Xiao, Boyd, & Kim, 2007).

Table 2
Properties of the datasets.

Datasets	# of samples	# of features
epsilon	400 000	2000
rcv1	677 399	47 236

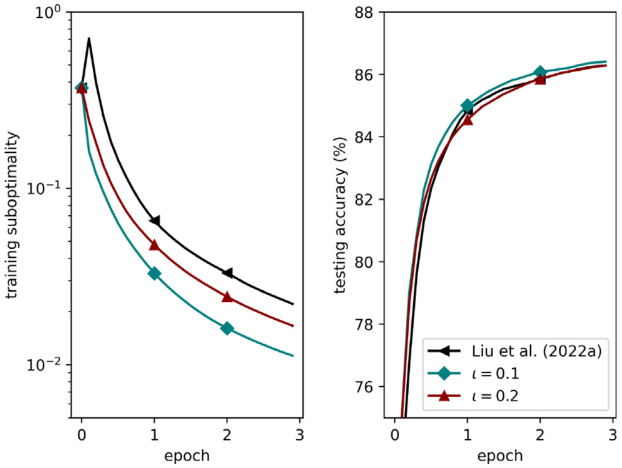


Fig. 1. Performance comparison between Algorithm 1 and Liu, Johansson et al. (2022) for l_2-regularized SVM with $\varepsilon = 0.8$.

Some common parameters used in the two sets of experiments are introduced in the following. For the parameters of DP, we consider $\varepsilon \in \{0.2, 0.4, 0.6, 0.8, 1\}$ and $\delta_0 = 0.01$. The random noises in these two cases are generated accordingly based on Theorem 1. The convergence performance of the algorithm is captured by suboptimality, i.e., $F(n^{-1}\sum_{i=1}^{n}\tilde{x}_i^{(t)}) - F(x^*)$, versus the number of iterations, where the ground truth is obtained by the optimizer SGDClassifier from scikit-learn (Pedregosa et al., 2011).

5.2. Results for l_2-regularized SVM

Set $h(x) = \mu\|x\|^2/2$ with $\mu = 0.0005$. Since the problem is strongly convex, we set $a_t = t$ and $\gamma_t = 20$.

We set $\varepsilon = 0.8$ and compare the convergence performance between (Liu, Johansson and Shi, 2022) and Algorithm 1 under different choices of $\iota \in \{0.1, 0.2\}$. Fig. 1 shows that Algorithm 1 with both choices of ι outperform (Liu, Johansson et al., 2022) in terms of convergence speed and model accuracy. Furthermore, the use of larger ι in Algorithm 1 leads to higher utility loss, which verifies Corollary 3. We observe that selecting a higher number of sampled nodes at each step leads to improved network connectivity as well as increased noise. The findings from Fig. 1 indicate that, in this specific example, the impact of increased noise on convergence performance may outweigh the benefits of enhanced connectivity.

Next, we examine the performance of Liu, Johansson et al. (2022) and Algorithm 1 under a set of DP parameters. The result in Fig. 2 illustrates that increasing the value of ε–indicating a less stringent privacy requirement–results in decreased utility loss across all the methods. This can be attributed to the fact that a smaller value of ε corresponds to a more stringent differential privacy (DP) constraint, necessitating a stronger noise to perturb the subgradient. In addition, the performance gap between Algorithm 1 and Liu, Johansson et al. (2022) is more significant for the case with smaller ε, i.e., a tighter DP requirement.

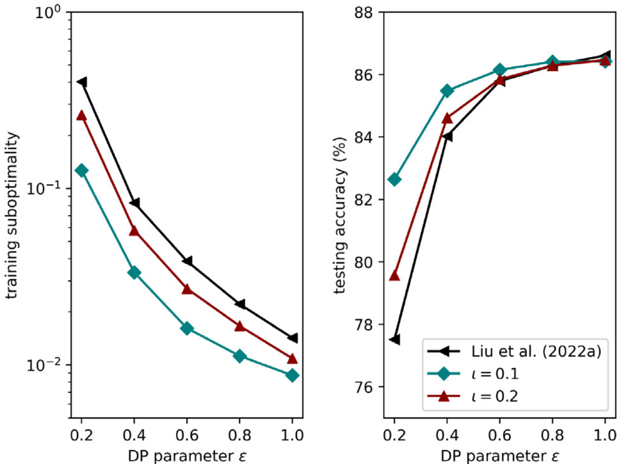


Fig. 2. Privacy–utility tradeoff in l_2-regularized SVM. The suboptimality and accuracy are evaluated after 3-epoch training.

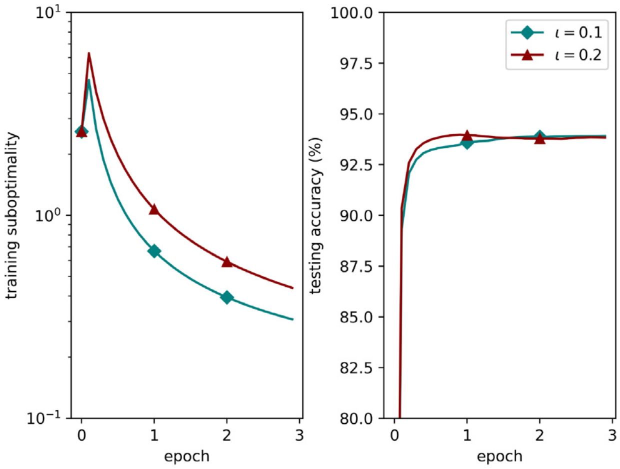


Fig. 3. Performance comparison between Algorithm 1 with different ι for l_1-regularized SVM with $\varepsilon = 0.4$.

5.3. Results for l_1-regularized SVM

Set $h(x) = \phi\|x\|_1$ with $\phi = 0.0005$. In this case, the problem in (26) is convex with a non-smooth regularization term. According to Corollary 2, we set $\gamma_t = 0.01\sqrt{t}$ and $a_t = 1$ in the experiment.

First, we set $\varepsilon = 0.4$ and compare Algorithm 1 under different subsampling ratios. The findings depicted in Fig. 3 illustrate a similar trend: As the subsampling ratio ι decreases, the utility loss diminishes correspondingly. Additionally, we present the results for Algorithm 1 with various DP parameters in Fig. 4. Notably, for both selected subsampling ratios, we observe a degradation in utility as the DP parameter ε decreases.

In summary, the experimental results reveal the effectiveness of the proposed algorithms and validate our theoretical findings.

6. Conclusion

In this work, we presented a class of differentially private DDA algorithms for solving ERM over networks. The proposed algorithms achieve DP by (*i*) randomly activating a fraction of nodes at each time instant and (*ii*) perturbing the stochastic subgradients over individual data samples within activated nodes. We proved

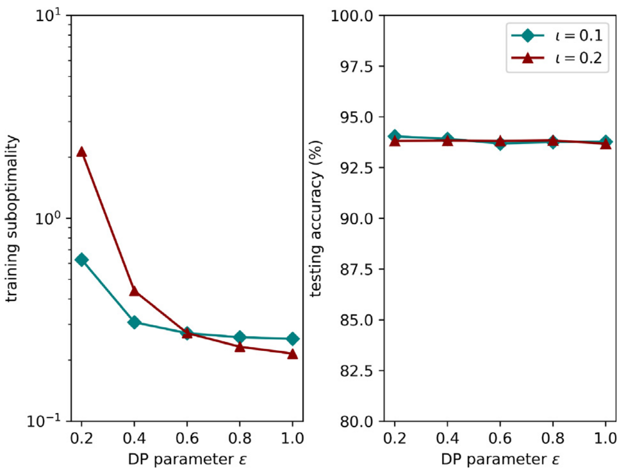


Fig. 4. Privacy–utility tradeoff in l_1-regularized SVM. The suboptimality and accuracy are evaluated after 3-epoch training.

that our algorithms substantially improve over existing ones in terms of utility loss.

There are numerous promising directions for future endeavors. Firstly, an intriguing avenue to explore is the heterogeneous case, where nodes exhibit substantial variations in dataset size and/or Lipschitz constants. Secondly, it is worthwhile to investigate the high probability convergence of the proposed algorithms.

Appendix A. Proof of Lemma 5

Notation: To facilitate the presentation, we introduce the following notation. Define $\boldsymbol{W}^{(t)} = W^{(t)} \otimes I$. Given a real-valued random vector x, we let

$$\|x\|_{\mathbb{E}} = \sqrt{\mathbb{E}[\|x\|^2]}. \tag{A.1}$$

Accordingly, for a square random matrix W, we denote $\|W\|_{\mathbb{E}} = \sup_{\|x\|_{\mathbb{E}}=1} \|Wx\|_{\mathbb{E}}$. Denote $\mathbf{y}^{(t)} = \mathbf{1} \otimes y^{(t)}$ and

$$\mathbf{z}^{(t)} = \begin{bmatrix} z_1^{(t)} \\ \vdots \\ z_n^{(t)} \end{bmatrix}, \mathbf{x}^{(t)} = \begin{bmatrix} x_1^{(t)} \\ \vdots \\ x_n^{(t)} \end{bmatrix}, \boldsymbol{\zeta}^{(t)} = \begin{bmatrix} \zeta_1^{(t)} \\ \vdots \\ \zeta_n^{(t)} \end{bmatrix}, \boldsymbol{\eta}^{(t)} = \begin{bmatrix} \eta_1^{(t)} \\ \vdots \\ \eta_n^{(t)} \end{bmatrix}. \tag{A.2}$$

Proof. This proof consists of three parts. First, we prove

$$\|\mathbf{x}^{(t)} - \mathbf{y}^{(t)}\| \le \frac{1}{\gamma_t + \mu\iota A_t} \left\|\mathbf{z}^{(t)} - \mathbf{1} \otimes \bar{z}^{(t)}\right\|, \tag{A.3}$$

where $\bar{z}(t) = n^{-1} \sum_{i=1}^n z_i(t)$. Second, we prove

$$\left\| \frac{\mathbf{z}^{(t)}}{a_t} - \frac{\mathbf{1} \otimes \bar{z}^{(t)}}{a_t} \right\|_{\mathbb{E}}^2 \le \frac{n\iota(L^2 + m\sigma^2)}{(1-\beta)^2}. \tag{A.4}$$

Finally, we conclude the proof using these two inequalities.

Part (i) Following Liu, Zhou et al. (2022, Lemma 5), we have $\iota A_\tau h(x) + \gamma_\tau d(x)$ is $(\mu\iota A_\tau + \gamma_\tau)$-strongly convex, and Ψ_τ^*, defined by

$$\Psi_\tau^*(w) = \sup_x \left\{ \langle w, x \rangle - \iota A_\tau h(x) - \gamma_\tau d(x) \right\}, \tag{A.5}$$

has $(\mu\iota A_\tau + \gamma_\tau)^{-1}$-Lipschitz continuous gradients. Let $\theta^{(t)} = n^{-1} \sum_{i=1}^n \eta_i^{(t)} \zeta_i^{(t)}$ and $\circ$ be the Hadamard product. Due to

$$\begin{aligned}
\bar{z}^{(t+1)} &= \frac{1}{n}(\mathbf{1}^T \otimes I)\mathbf{z}^{(t+1)} \\
&= \frac{\mathbf{1}^T \otimes I}{n} \boldsymbol{W}^{(t)} \left(\mathbf{z}^{(t)} + a_t \boldsymbol{\eta}^{(t)} \circ \boldsymbol{\zeta}^{(t)}\right) \\
&= \frac{(\mathbf{1}^T \otimes I)}{n} \left(\mathbf{z}^{(t)} + a_t \boldsymbol{\eta}^{(t)} \circ \boldsymbol{\zeta}^{(t)}\right) \\
&= \bar{z}^{(t)} + a_t \theta^{(t)},
\end{aligned}$$

we have $y^{(\tau)} = \nabla \Psi_\tau^* \left(-\sum_{k=1}^{\tau-1} a_k \theta^{(k)}\right) = \nabla \Psi_\tau^* \left(-z^{(t)}\right)$. In addition, $x_i^{(\tau)} = \nabla \Psi_\tau^*(-z_i^{(\tau)})$, $\forall i = 1, \ldots, n$, which gives us (A.3).

Part (ii) When $t = 1$, since $z_i^{(1)} = 0$ for all i, we have $\bar{z}^{(1)} = 0$ and therefore (A.4) satisfied. Next, we consider the case with $t \ge 1$.

Let $\tilde{\mathbf{z}}^{(t+1)} = \mathbf{z}^{(t+1)} - \mathbf{1} \otimes \bar{z}^{(t+1)}$. Then,

$$\tilde{\mathbf{z}}^{(\tau+1)} = \tilde{\boldsymbol{W}}^{(\tau)} \left(\tilde{\mathbf{z}}^{(\tau)} + a_\tau \left(\boldsymbol{\eta}^{(\tau)} \circ \boldsymbol{\zeta}^{(\tau)} - \mathbf{1} \otimes \theta^{(\tau)}\right)\right), \tag{A.6}$$

where $\tilde{\boldsymbol{W}}^{(\tau-1)} = \tilde{W}^{(\tau-1)} \otimes I$ and $\tilde{W}^{(\tau-1)} = W^{(\tau-1)} - \frac{\mathbf{1}\mathbf{1}^T}{n}$. By iterating (A.6), we obtain

$$\begin{aligned}
\tilde{\mathbf{z}}^{(t)} &= \sum_{\tau=1}^{t-1} \tilde{\boldsymbol{W}}^{(\tau,t-1)} a_\tau \left(\boldsymbol{\eta}^{(\tau)} \circ \boldsymbol{\zeta}^{(\tau)} - \mathbf{1} \otimes \theta^{(\tau)}\right) \\
&= \sum_{\tau=1}^{t-1} \tilde{\boldsymbol{W}}^{(\tau,t-1)} a_\tau \left(\boldsymbol{\eta}^{(\tau)} \circ \boldsymbol{\zeta}^{(\tau)}\right),
\end{aligned} \tag{A.7}$$

where $\tilde{\boldsymbol{W}}^{(\tau,t-1)} = \tilde{\boldsymbol{W}}^{(t-1)} \ldots \tilde{\boldsymbol{W}}^{(\tau)}$. Since $\mathbf{1} \otimes \theta^{(\tau)} = (\mathbf{1} \otimes I)\theta^{(\tau)}$ and $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$, we have $\left(\tilde{W}^{(\tau,t-1)} \otimes I\right)(\mathbf{1} \otimes \theta^{(\tau)}) = \left(\left(\tilde{W}^{(\tau,t-1)}\mathbf{1}\right) \otimes I\right) \theta^{(\tau)} = 0$. Therefore, (A.7) can be rewritten as, $\forall t \ge 1$,

$$\tilde{\mathbf{z}}^{(t)} = \sum_{\tau=1}^{t-1} \tilde{\boldsymbol{W}}^{(\tau,t-1)} a_\tau \left(\boldsymbol{\eta}^{(\tau)} \circ \boldsymbol{\zeta}^{(\tau)}\right).$$

Upon taking the norm defined in (A.1) on both sides and using the Minkowski inequality (Gut, 2013), we obtain

$$\|\tilde{\mathbf{z}}^{(t)}\|_{\mathbb{E}} \le \sum_{\tau=1}^{t-1} a_\tau \left\| \tilde{\boldsymbol{W}}^{(\tau,t-1)} \left(\boldsymbol{\eta}^{(\tau)} \circ \boldsymbol{\zeta}^{(\tau)}\right) \right\|_{\mathbb{E}}. \tag{A.8}$$

Consider

$$\begin{aligned}
&\left\| \tilde{\boldsymbol{W}}^{(\tau,t-1)} \left(\boldsymbol{\eta}^{(\tau)} \circ \boldsymbol{\zeta}^{(\tau)}\right) \right\|_{\mathbb{E}}^2 \\
&\overset{(i)}{=} \mathbb{E}\Big[\left(\boldsymbol{\eta}^{(\tau)} \circ \boldsymbol{\zeta}^{(\tau)}\right)^T (\tilde{\boldsymbol{W}}^{(\tau,t-2)})^T \left((\tilde{\boldsymbol{W}}^{(t-1)})^T \tilde{\boldsymbol{W}}^{(t-1)}\right) \\
&\qquad \times \tilde{\boldsymbol{W}}^{(\tau,t-2)} \left(\boldsymbol{\eta}^{(\tau)} \circ \boldsymbol{\zeta}^{(\tau)}\right) \Big] \\
&\overset{(ii)}{\le} \left\| \tilde{\boldsymbol{W}}^{(\tau,t-2)} \left(\boldsymbol{\eta}^{(\tau)} \circ \boldsymbol{\zeta}^{(\tau)}\right) \right\|_{\mathbb{E}}^2 \rho \left(\mathbb{E}_W \left[(\tilde{W}^{(t-1)})^T \tilde{W}^{(t-1)}\right]\right) \\
&\overset{(iii)}{\le} \beta^2 \left\| \tilde{\boldsymbol{W}}^{(\tau,t-2)} \left(\boldsymbol{\eta}^{(\tau)} \circ \boldsymbol{\zeta}^{(\tau)}\right) \right\|_{\mathbb{E}}^2 \\
&\overset{(iv)}{\le} \beta^{2(t-\tau-1)} \left\| \tilde{\boldsymbol{W}}^{(\tau)} \left(\boldsymbol{\eta}^{(\tau)} \circ \boldsymbol{\zeta}^{(\tau)}\right) \right\|_{\mathbb{E}}^2,
\end{aligned}$$

where we use $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ and that $\tilde{W}^{(t-1)}$ is independent of the random events that occur up to time $t-2$ to obtain (i) and (ii), respectively, (iii) is due to $\mathbb{E}[(\tilde{W}^{(t-1)})^T \tilde{W}^{(t-1)}] = \mathbb{E}[(W^{(t-1)})^T W^{(t-1)}] - \frac{\mathbf{1}\mathbf{1}^T}{n}$ and (iv) is by iteration. Therefore, we get from (A.8) that

$$\begin{aligned}\|\tilde{\mathbf{z}}^{(t)}\|_{\mathbb{E}} &\le \sum_{\tau=1}^{t-1}\beta^{t-\tau-1}a_\tau\left\|\tilde{\boldsymbol{W}}^{(\tau)}\left(\boldsymbol{\eta}^{(\tau)}\circ\boldsymbol{\zeta}^{(\tau)}\right)\right\|_{\mathbb{E}}\\ &\le \sum_{\tau=1}^{t-1}\beta^{t-\tau-1}a_\tau\left\|\tilde{\boldsymbol{W}}^{(\tau)}\right\|_{\mathbb{E}}\left\|\boldsymbol{\eta}^{(\tau)}\circ\boldsymbol{\zeta}^{(\tau)}\right\|_{\mathbb{E}}\\ &\le \sum_{\tau=1}^{t-1}\beta^{t-\tau-1}a_\tau\left\|\boldsymbol{\eta}^{(\tau)}\circ\boldsymbol{\zeta}^{(\tau)}\right\|_{\mathbb{E}},\end{aligned}$$

where the last inequality is due to $\left\|\tilde{\boldsymbol{W}}^{(\tau)}\right\|_{\mathbb{E}} \le 1$, leading to

$$\begin{aligned}\|\tilde{\mathbf{z}}^{(t)}\|_{\mathbb{E}}^2 &\le \left(\sum_{\tau=1}^{t-1}a_\tau\beta^{t-\tau-1}\left\|\boldsymbol{\eta}^{(\tau)}\circ\boldsymbol{\zeta}^{(\tau)}\right\|_{\mathbb{E}}\right)^2\\ &\le \left(\sum_{\tau=1}^{t-1}\left(\beta^{\frac{t-\tau-1}{2}}\right)^2\right)\left(\sum_{\tau=1}^{t-1}\left(a_\tau\beta^{\frac{t-\tau}{2}}\left\|\boldsymbol{\eta}^{(\tau)}\circ\boldsymbol{\zeta}^{(\tau)}\right\|_{\mathbb{E}}\right)^2\right)\\ &\le \frac{1}{1-\beta}\sum_{\tau=1}^{t}a_\tau^2\beta^{t-\tau-1}\left\|\boldsymbol{\eta}^{(\tau)}\circ\boldsymbol{\zeta}^{(\tau)}\right\|_{\mathbb{E}}^2,\end{aligned}$$

where the second inequality is due to $(w+v)^2 \le 2w^2+2v^2$. Upon dividing both sides by a_t^2 and using $0 < a_\tau \le a_t, \forall\tau \le t$, we have

$$\left\|\frac{\tilde{\mathbf{z}}^{(t)}}{a_t}\right\|_{\mathbb{E}}^2 \le \frac{1}{1-\beta}\sum_{\tau=1}^{t}\frac{a_\tau^2}{a_t^2}\beta^{t-\tau-1}\left\|\boldsymbol{\eta}^{(\tau)}\circ\boldsymbol{\zeta}^{(\tau)}\right\|_{\mathbb{E}}^2 \le \frac{n\iota(L^2+m\sigma^2)}{(1-\beta)^2}.$$

Part (iii) Based on (A.3) and (A.4), we have

$$\frac{n\iota(L^2+m\sigma^2)a_t^2}{(1-\beta)^2(\mu\iota A_t+\gamma_t)^2} \ge \|\mathbf{x}^{(t)}-\mathbf{y}^{(t)}\|_{\mathbb{E}}^2 \ge \sum_{i=1}^{n}\|x_i^{(t)}-y^{(t)}\|_{\mathbb{E}}^2.$$

By the Jensen's inequality, we have

$$\mathbb{E}\left[\|\mathbf{x}^{(t)}-\mathbf{y}^{(t)}\|\right] \le \sqrt{\|\mathbf{x}^{(t)}-\mathbf{y}^{(t)}\|_{\mathbb{E}}^2} \le \frac{\sqrt{n\iota}(L+\sqrt{m}\sigma)a_t}{(\gamma_t+\mu\iota A_t)(1-\beta)}.$$

This together with the bound between l_1 and l_2-norms, i.e., $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\|x_i^{(t)}-y^{(t)}\| \le \|\mathbf{x}^{(t)}-\mathbf{y}^{(t)}\|$ yields

$$\sum_{i=1}^{n}\mathbb{E}[\|x_i^{(t)}-y^{(t)}\|] \le \frac{n\iota(L+\sqrt{m}\sigma)a_t}{(\gamma_t+\mu\iota A_t)(1-\beta)}. \tag{A.9}$$

Appendix B. Proof of Lemma 6

Proof. Recall (A.5)

$$\Psi_\tau^*(w) = \sup_x\left\{\langle w, x\rangle - \iota A_\tau h(x) - \gamma_\tau d(x) := R_\tau(x, w)\right\}.$$

Since $-R_\tau(\cdot, w)$ is $(\mu\iota A_\tau+\gamma_\tau)$-strongly convex, we have

$$\nabla\Psi_\tau^*\left(-\sum_{k=1}^{\tau-1}a_k\theta^{(k)}\right) = y^{(\tau)},$$

where $\theta^{(t)} = n^{-1}\sum_{i=1}^{n}\eta_i^{(t)}\zeta_i^{(t)}$, and Ψ_τ^* has $(\mu\iota A_\tau+\gamma_\tau)^{-1}$-Lipschitz continuous gradients, see, e.g., Liu, Zhou et al. (2022, Lemma 5), implying

$$\begin{aligned}&\Psi_\tau^*\left(-\sum_{k=1}^{\tau}a_k\theta^{(k)}\right)\\ &\le \Psi_\tau^*\left(-\sum_{k=1}^{\tau-1}a_k\theta^{(k)}\right) - a_\tau\langle y^{(\tau)}, \theta^{(\tau)}\rangle + \frac{a_\tau^2\left\|\theta^{(\tau)}\right\|^2}{2(\mu\iota A_\tau+\gamma_\tau)}.\end{aligned} \tag{B.1}$$

Upon using γ_τ is non-decreasing and $d(x) \ge 0$, we have

$$\begin{aligned}\Psi_\tau^*\left(-\sum_{k=1}^{\tau-1}a_k\theta^{(k)}\right) &= R_\tau\left(y^{(\tau)}, -\sum_{k=1}^{\tau-1}a_k\theta^{(k)}\right)\\ &= R_{\tau-1}\left(y^{(\tau)}, -\sum_{k=1}^{\tau-1}a_k\theta^{(k)}\right) - \iota a_\tau h(y^{(\tau)})\\ &\quad + (\gamma_{\tau-1}-\gamma_\tau)d(y^{(\tau)})\\ &\le R_{\tau-1}\left(y^{(\tau)}, -\sum_{k=1}^{\tau-1}a_k\theta^{(k)}\right) - \iota a_\tau h(y^{(\tau)})\\ &\le \Psi_{\tau-1}^*\left(-\iota\sum_{k=1}^{\tau-1}a_k\theta^{(k)}\right) - \iota a_\tau h(y^{(\tau)}).\end{aligned} \tag{B.2}$$

Upon plugging (B.2) into (B.1) and summing up the resultant inequality from $\tau = 1$ to $\tau = t$, we have

$$\begin{aligned}&\sum_{\tau=1}^{t}a_\tau\left(\langle\theta^{(\tau)}, y^{(\tau)}\rangle + \iota h(y^{(\tau)})\right)\\ &\le \Psi_0^*(0) - \Psi_\tau^*\left(-\sum_{k=1}^{\tau}a_k\theta^{(k)}\right) + \sum_{\tau=1}^{t}\frac{a_\tau^2\|\theta^{(\tau)}\|^2}{2(\mu\iota A_\tau+\gamma_\tau)}.\end{aligned}$$

Note that $y^{(1)} = \nabla\Psi_1^*(0)$, $A_0 = 0$ and $\gamma_0 = 0$ by definition, implying that $\Psi_0^*(0) = 0$. Further considering

$$\begin{aligned}&\sum_{\tau=1}^{t}a_\tau\langle\theta^{(\tau)}, -x^*\rangle \le \iota A_t h(x^*) + \gamma_t d(x^*)\\ &\quad + \sup_x\left\{-\sum_{k=1}^{\tau}a_k\langle\theta^{(k)}, x\rangle - \iota A_t h(x) - \gamma_t d(x)\right\}\\ &\le \Psi_t^*\left(-\sum_{k=1}^{\tau}a_k\theta^{(k)}\right) + \iota A_t h(x^*) + \gamma_t d(x^*),\end{aligned}$$

we obtain

$$\begin{aligned}&\sum_{\tau=1}^{t}a_\tau\left(\langle\theta^{(\tau)}, y^{(\tau)}-x^*\rangle + \iota\left(h(y^{(\tau)})-h(x^*)\right)\right)\\ &\le \sum_{\tau=0}^{t}\frac{a_\tau^2\|\theta^{(\tau)}\|^2}{2(\mu\iota A_\tau+\gamma_\tau)} + \gamma_t d(x^*).\end{aligned} \tag{B.3}$$

Dividing both sides by $\iota > 0$ leads to the desired inequality.

Appendix C. Proof of Corollary 1

Proof. We obtain from the update of A_t in Algorithm 1 that

$$\sum_{\tau=1}^{t}\frac{a_\tau^2}{\mu\iota A_\tau+\gamma_\tau} = \sum_{\tau=1}^{t}\frac{2\tau^2}{\mu\iota\tau(\tau+1)} \le \frac{2t}{\mu\iota}. \tag{C.1}$$

Due to μ-strong convexity of F, we have

$$\begin{aligned}\sum_{i=1}^{n}\|\tilde{x}_i^{(t)}-x^*\|^2 &\le 2\sum_{i=1}^{n}\|\tilde{x}_i^{(t)}-\tilde{y}^{(t)}\|^2 + 2n\|\tilde{y}^{(t)}-x^*\|^2\\ &\le 2\|\tilde{x}_i^{(t)}-\tilde{y}^{(t)}\|^2 + \frac{4n}{\mu}\left(F(\tilde{y}^{(t)})-F(x^*)\right).\end{aligned} \tag{C.2}$$

Upon using convexity of the l_2-norm and (19), we obtain

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}\left[\|\tilde{x}_i^{(t)}-\tilde{y}^{(t)}\|^2\right] &\le \mathbb{E}\left[\frac{1}{A_t}\sum_{\tau=0}^{t}\sum_{i=1}^{n}a_\tau\|x_i^{(\tau)}-y^{(\tau)}\|^2\right]\\ &\le \frac{1}{A_t}\sum_{\tau=1}^{t}\frac{a_\tau^3}{(\mu\iota A_\tau+\gamma_\tau)^2}\left(\frac{\iota n L^2}{(1-\beta)^2}\right) \qquad \text{(C.3)}\\ &\le \frac{4nL^2}{\mu^2\iota A_t(1-\beta)^2}\sum_{\tau=1}^{t}\frac{1}{a_\tau}\le\frac{4nL^2(\log t+1)}{\mu^2\iota A_t(1-\beta)^2}.\end{aligned}$$

By putting expectations on (C.2) and using (C.3) and (9), we arrive at (10) as desired.

Appendix D. Proof of Corollary 2

Proof. Under the choices of $a_t = 1$ and $\gamma_t = \gamma\sqrt{t}$, we have

$$\sum_{\tau=1}^{t}\frac{a_\tau^2}{\mu\iota A_\tau+\gamma_t}=\frac{1}{\gamma}\sum_{\tau=1}^{t}\frac{1}{\sqrt{\tau}}\le\frac{2}{\gamma}\sqrt{t}.$$

Therefore, by using (9) and (18), we obtain for all $t \ge 1$,

$$\mathbb{E}\left[F(\tilde{y}^{(t)})-F(x^*)\right]\le\frac{d(x^*)+2\iota M}{\iota\gamma\sqrt{t}}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|x_i^{(t)}-y^{(t)}\|\right]\le\frac{L\sqrt{\iota}}{\gamma(1-\beta)\sqrt{t}},\quad i=1,\ldots,n$$

respectively. Since l_2-norm is convex, we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\tilde{x}_i^{(t)}-\tilde{y}^{(t)}\|]\le\frac{1}{t}\sum_{\tau=1}^{t}\frac{L\sqrt{\iota}}{\gamma(1-\beta)\sqrt{t}}\le\frac{2L\sqrt{\iota}}{\gamma(1-\beta)\sqrt{t}}. \qquad \text{(D.1)}$$

Appendix E. Proof of Corollary 3

Proof. By (C.1), we have

$$\begin{aligned}\mathbb{E}[F(\tilde{y}^{(T)})-F(x^*)]\le{}&\frac{4M}{\mu\iota(T+1)}+\frac{80m\log(2/\delta_0)\iota^2L^2}{\mu q^2\varepsilon^2}\\ &+\frac{16\sqrt{10m\log(2/\delta_0)}\iota L^2}{\mu q\varepsilon(1-\beta)\sqrt{T+1}},\end{aligned}$$

where M is a positive constant defined in Theorem 2. Similar to (C.3), we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\tilde{x}_i^{(T)}-\tilde{y}^{(T)}\|^2\right]\le\frac{8(L^2+m\sigma^2)(\log T+1)}{\mu^2\iota T(T+1)(1-\beta)^2}.$$

Following (C.2), we set $T=\mathcal{O}\left(\frac{q^2\varepsilon^2}{\iota^3(1-\beta)^2 m\log(1/\delta_0)}\right)$ to obtain (13).

Appendix F. Proof of Corollary 4

Proof. By the specific choices of parameters in (14), we obtain from (9) that

$$\begin{aligned}&\mathbb{E}[F(\tilde{y}^{(T)})-F(x^*)]\\ &\le\frac{d(x^*)+2\iota M}{\gamma\iota\sqrt{T}}+\frac{40m\iota^3L^2\log(2/\delta_0)}{\gamma q^2\varepsilon^2}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\\ &\quad+\frac{4\sqrt{10m\iota\log(2/\delta_0)}\iota L^2}{\gamma q\varepsilon(1-\beta)\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\\ &\le\frac{d(x^*)+2\iota M}{\gamma\iota\sqrt{T}}+\frac{8\sqrt{10m\iota\log(2/\delta_0)}\iota L^2}{\gamma q\varepsilon(1-\beta)}\\ &\quad+\frac{80m\iota^3L^2\log(2/\delta_0)\sqrt{T}}{\gamma q^2\varepsilon^2},\end{aligned}$$

where the last inequality is due to $\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\le 1+\int_1^T\frac{1}{\sqrt{t}}dt\le 2\sqrt{t}$. Upon using $\iota\le 1-\beta$ and $T=\mathcal{O}\left(\frac{q^2\varepsilon^2}{\iota^3(1-\beta)^2 m\log(1/\delta_0)}\right)$, we arrive at (15). In addition, similar to (D.1), we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\tilde{x}_i^{(T)}-\tilde{y}^{(T)}\|]\\ &\le\frac{2(L+\sqrt{m}\sigma)\sqrt{\iota}}{\gamma(1-\beta)\sqrt{T}}\le\mathcal{O}\left(\frac{L\sqrt{m\iota\log(1/\delta_0)}}{\gamma q\varepsilon}\right).\end{aligned}$$

References

Abadi, Martin, Chu, Andy, Goodfellow, Ian, McMahan, H Brendan, Mironov, Ilya, Talwar, Kunal, et al. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security* (pp. 308–318).

Bassily, Raef, Smith, Adam, & Thakurta, Abhradeep (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science* (pp. 464–473). IEEE.

Chaudhuri, Kamalika, Sarwate, Anand, & Sinha, Kaushik (2012). Near-optimal differentially private principal components. *Advances in Neural Information Processing Systems*, *25*, 989–997.

Chen, Fei, Chen, Xiaozheng, Xiang, Linying, & Ren, Wei (2021). Distributed economic dispatch via a predictive scheme: heterogeneous delays and privacy preservation. *Automatica*, *123*, Article 109356.

Colin, Igor, Bellet, Aurélien, Salmon, Joseph, & Clémençon, Stéphan (2016). Gossip dual averaging for decentralized optimization of pairwise functions. In *International conference on machine learning* (pp. 1388–1396). PMLR.

Ding, Tie, Zhu, Shanying, He, Jianping, Chen, Cailian, & Guan, Xinping (2021). Differentially private distributed optimization via state and direction perturbation in multiagent systems. *IEEE Transactions on Automatic Control*, *67*(2), 722–737.

Duchi, John C., Agarwal, Alekh, & Wainwright, Martin J. (2011). Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, *57*(3), 592–606.

Dwork, Cynthia (2006). Differential privacy. In *International colloquium on automata, languages, and programming* (pp. 1–12). Springer.

Girgis, Antonious M, Data, Deepesh, Diggavi, Suhas, Kairouz, Peter, & Suresh, Ananda Theertha (2021). Shuffled model of federated learning: Privacy, accuracy and communication trade-offs. *IEEE Journal on Selected Areas in Information Theory*, *2*(1), 464–478.

Gut, Allan (2013). *Probability: a graduate course, Vol. 75*. Springer Science & Business Media.

Han, Dongyu, Liu, Kun, Lin, Yeming, & Xia, Yuanqing (2022). Differentially private distributed online learning over time-varying digraphs via dual averaging. *International Journal of Robust and Nonlinear Control*, *32*(5), 2485–2499.

Han, Shuo, Topcu, Ufuk, & Pappas, George J. (2016). Differentially private distributed constrained optimization. *IEEE Transactions on Automatic Control*, *62*(1), 50–64.

Hu, Rui, Guo, Yuanxiong, Li, Hongning, Pei, Qingqi, & Gong, Yanmin (2020). Personalized federated learning with differential privacy. *IEEE Internet of Things Journal*, *7*(10), 9530–9539.

Huang, Zhenqi, Mitra, Sayan, & Vaidya, Nitin (2015). Differentially private distributed optimization. In *Proceedings of the 2015 international conference on distributed computing and networking* (pp. 1–10).

Kairouz, Peter, Oh, Sewoong, & Viswanath, Pramod (2015). The composition theorem for differential privacy. In *International conference on machine learning* (pp. 1376–1385). PMLR.

Kifer, Daniel, Smith, Adam, & Thakurta, Abhradeep (2012). Private convex empirical risk minimization and high-dimensional regression. In *Proceedings of the 25th annual conference on learning theory, Vol. 23* (pp. 25.1–25.40). PMLR.

Lewis, David D., Yang, Yiming, Russell-Rose, Tony, & Li, Fan (2004). RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, *5*, 361–397.

Lian, Xiangru, Zhang, Ce, Zhang, Huan, Hsieh, Cho-Jui, Zhang, Wei, & Liu, Ji (2017). Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in neural information processing systems* (pp. 5330–5340).

Liu, Changxin, Johansson, Karl H., & Shi, Yang (2022). Private stochastic dual averaging for decentralized empirical risk minimization. *IFAC-PapersOnLine*, *55*(13), 43–48.

Liu, Changxin, Zhou, Zirui, Pei, Jian, Zhang, Yong, & Shi, Yang (2022). Decentralized composite optimization in stochastic networks: A dual averaging approach with linear convergence. *IEEE Transactions on Automatic Control*, *68*(8), 4650–4665.

Lu, Haihao, Freund, Robert M., & Nesterov, Yurii (2018). Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, *28*(1), 333–354.

Melis, Luca, Song, Congzheng, De Cristofaro, Emiliano, & Shmatikov, Vitaly (2019). Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)* (pp. 691–706). IEEE.

Nedic, Angelia, & Ozdaglar, Asuman (2009). Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, *54*(1), 48–61.

Noble, Maxence, Bellet, Aurélien, & Dieuleveut, Aymeric (2022). Differentially private federated learning on heterogeneous data. In *International conference on artificial intelligence and statistics* (pp. 10110–10145). PMLR.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Sonnenburg, Sören, Rätsch, Gunnar, Schäfer, Christin, & Schölkopf, Bernhard (2006). Large scale multiple kernel learning. *Journal of Machine Learning Research*, *7*, 1531–1565.

Tao, Wei, Li, Wei, Pan, Zhisong, & Tao, Qing (2021). Gradient descent averaging and primal-dual averaging for strongly convex optimization. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 35* (pp. 9843–9850).

Vlaski, Stefan, & Sayed, Ali H. (2021). Graph-homomorphic perturbations for private decentralized learning. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5240–5244). IEEE.

Wang, Yongqiang, & Nedić, Angelia (2023). Tailoring gradient methods for differentially-private distributed optimization. *IEEE Transactions on Automatic Control*.

Wang, Di, Ye, Minwei, & Xu, Jinhui (2017). Differentially private empirical risk minimization revisited: faster and more general. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 2719–2728).

Wang, Jimin, Zhang, Ji-Feng, & He, Xingkang (2022). Differentially private distributed algorithms for stochastic aggregative games. *Automatica*, *142*, Article 110440.

Xiao, Lin, Boyd, Stephen, & Kim, Seung-Jean (2007). Distributed average consensus with least-mean-square deviation. *Journal of Parallel and Distributed Computing*, *67*(1), 33–46.

Xiao, Hanshen, & Devadas, Srinivas (2021). Towards understanding practical randomness beyond noise: Differential privacy and mixup. *Cryptology ePrint Archive*.

Xiong, Yongyang, Xu, Jinming, You, Keyou, Liu, Jianxing, & Wu, Ligang (2020). Privacy-preserving distributed online optimization over unbalanced digraphs via subgradient rescaling. *IEEE Transactions on Control of Network Systems*, *7*(3), 1366–1378.

Xu, Jie, Zhang, Wei, & Wang, Fei (2021). A(DP)^{2}SGD: Asynchronous decentralized parallel stochastic gradient descent with differential privacy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yuan, Deming, Hong, Yiguang, Ho, Daniel W. C., & Jiang, Guoping (2018). Optimal distributed stochastic mirror descent for strongly convex optimization. *Automatica*, *90*, 196–203.

Zhang, Chunlei, Ahmad, Muaz, & Wang, Yongqiang (2018). ADMM based privacy-preserving decentralized optimization. *IEEE Transactions on Information Forensics and Security*, *14*(3), 565–580.

Zhang, Hongyi, Cisse, Moustapha, Dauphin, Yann N, & Lopez-Paz, David (2018). Mixup: Beyond empirical risk minimization. In *International conference on learning representations*.

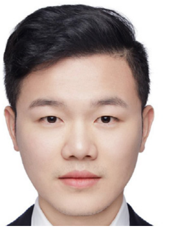
Zhang, Xueru, Khalili, Mohammad Mahdi, & Liu, Mingyan (2018). Improving the privacy and accuracy of ADMM-based distributed algorithms. In *International conference on machine learning* (pp. 5796–5805). PMLR.

Zhang, Jiaqi, Zheng, Kai, Mou, Wenlong, & Wang, Liwei (2017). Efficient private ERM for smooth objectives. In *Proceedings of the 26th international joint conference on artificial intelligence* (pp. 3922–3928).

Zhang, Tao, & Zhu, Quanyan (2016). Dynamic differential privacy for ADMM-based distributed classification learning. *IEEE Transactions on Information Forensics and Security*, *12*(1), 172–187.

Zhu, Ligeng, & Han, Song (2020). Deep leakage from gradients. In *Federated learning* (pp. 17–31). Springer.

Zhu, Junlong, Xu, Changqiao, Guan, Jianfeng, & Wu, Dapeng Oliver (2018). Differentially private distributed online algorithms over time-varying directed networks. *IEEE Transactions on Signal and Information Processing over Networks*, *4*(1), 4–17.

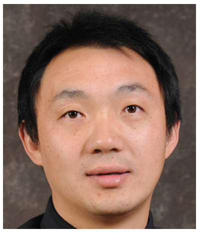


Changxin Liu received the Ph.D. degree in mechanical engineering from the University of Victoria, Victoria, BC, Canada, in 2021. He is currently a Postdoctoral Researcher with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. His current research interests include control, optimization, and learning of networked systems, and applications in real–world cyber–physical systems.

Dr. Liu was a recipient of the NSERC Postdoctoral Fellowship in 2023. He currently serves as Associate Editor for Circuits, Systems, and Signal Processing.

Karl H. Johansson received the M.Sc. and Ph.D. degrees in electrical engineering from Lund University, Lund, Sweden, in 1992 and 1997, respectively. He is a Professor with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. He has held visiting positions at UC Berkeley, Caltech, NTU, HKUST Institute of Advanced Studies, and NTNU. His research interests include networked control systems, cyber–physical systems, and applications in transportation, energy, and automation. He has served on the IEEE Control Systems Society Board of Governors, the IFAC Executive Board, and the European Control Association Council.

Dr. Johansson was a recipient of several best paper awards and other distinctions from IEEE and ACM. He has been awarded Distinguished Professor with the Swedish Research Council and Wallenberg Scholar with the Knut and Alice Wallenberg Foundation. He was also a recipient of the Future Research Leader Award from the Swedish Foundation for Strategic Research and the Triennial Young Author Prize from IFAC. He is Fellow of the Royal Swedish Academy of Engineering Sciences, and he is IEEE Control Systems Society Distinguished Lecturer.



Yang Shi received his B.Sc. and Ph.D. degrees in mechanical engineering and automatic control from Northwestern Polytechnical University, Xi'an, China, in 1994 and 1998, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Alberta, Edmonton, AB, Canada, in 2005. He was a Research Associate in the Department of Automation at Tsinghua University during 1998–2000. From 2005 to 2009, he was an Assistant Professor and Associate Professor in the Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SK, Canada. In 2009, he joined the University of Victoria, and now he is a Professor in the Department of Mechanical Engineering, University of Victoria (UVic), Victoria, BC, Canada. His current research interests include networked and distributed systems, model predictive control (MPC), cyber–physical systems (CPS), robotics and mechatronics, navigation and control of autonomous systems (AUV and UAV), and energy system applications.

Dr. Shi received the University of Saskatchewan Student Union Teaching Excellence Award in 2007, and the Faculty of Engineering Teaching Excellence Award in 2012 at UVic. He is the recipient of the JSPS Invitation Fellowship (short-term) in 2013, the UVic Craigdarroch Silver Medal for Excellence in Research in 2015, the 2017 IEEE Transactions on Fuzzy Systems Outstanding Paper Award, the Humboldt Research Fellowship for Experienced Researchers in 2018; CSME Mechatronics Medal (2023); IEEE Dr.-Ing. Eugene Mittelmann Achievement Award (2023). He is IFAC Council Member; VP on Conference Activities of IEEE IES and the Chair of IEEE IES Technical Committee on Industrial Cyber-Physical Systems. Currently, he is Co-Editor-in-Chief of IEEE Transactions on Industrial Electronics, and Editor-in-Chief of IEEE Canadian Journal of Electrical and Computer Engineering; he also serves as Associate

Editor for Automatica, IEEE Transactions on Automatic Control, Annual Review in Controls, etc. He is a Distinguished Lecturer of IES. He is a Fellow of IEEE, ASME, CSME, Engineering Institute of Canada (EIC), Canadian Academy of Engineering (CAE), and a registered Professional Engineer in British Columbia, Canada.