Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 55-13 (2022) 43–48

IFAC Papers Online CONFERENCE PAPER ARCHIVE

# Private Stochastic Dual Averaging for Decentralized Empirical Risk Minimization

**Changxin Liu** * **Karl H. Johansson** * **Yang Shi** **

* *School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, and Digital Futures, 100 44 Stockholm, Sweden (e-mail: {changxin, kallej}@kth.se).*

** *Department of Mechanical Engineering, University of Victoria, Victoria, B.C. V8W 3P6, Canada (e-mail: yshi@uvic.ca)*

**Abstract:**
In this work, we study the decentralized empirical risk minimization problem under the constraint of differential privacy (DP). Based on the algorithmic framework of dual averaging, we develop a novel decentralized stochastic optimization algorithm to solve the problem. The proposed algorithm features the following: *i*) it perturbs the stochastic subgradient evaluated over individual data samples, with which the information about the dataset can be released in a differentially private manner; *ii*) it employs hyperparameters that are more aggressive than conventional decentralized dual averaging algorithms to speed up convergence. The upper bound for the utility loss of the proposed algorithm is proven to be smaller than that of existing methods to achieve the same level of DP. As a by-product, when removing the perturbation, the non-private version of the proposed algorithm attains the optimal $\mathcal{O}(1/t)$ convergence rate for non-smooth stochastic optimization. Finally, experimental results are presented to demonstrate the effectiveness of the algorithm.

*Keywords:* Dual averaging, differential privacy, distributed optimization, convex optimization, large scale optimization problems.

## 1. INTRODUCTION

In decentralized learning, multiple parties aim at training machine learning models collaboratively. Compared to its centralized counterpart, it potentially improves training speed, scalability and robustness. Particularly, efficient optimization algorithms with local computation and peer-to-peer message-passing lie at the core of such decentralized frameworks. Existing approaches consist of decentralized primal-dual algorithms (Jakovetic et al., 2011), consensus-based gradient descent (GD) (Nedic et al., 2010), and consensus-based dual averaging (DA) (Duchi et al., 2011; Colin et al., 2016; Liu et al., 2021). The latter has demonstrated its advantages in promoting sparsity (Xiao, 2009), i.e., explicit feature selection, and handling time-varying networks (Duchi et al., 2011). We focus on DA-based methods in this work.

While the vast amounts of data in the modern society have contributed to the development of high-performance machine learning, they give rise to serious privacy concerns (Fredrikson et al., 2015; Zhu and Han, 2020). For example, it has been verified that the gradients used for training disclose essential properties of the dataset (Bassily et al., 2014), which may result in a reluctance to share useful information. To this end, differential privacy (DP) has been proposed to quantify to what extent the individual privacy in a dataset can be preserved while releasing useful aggregate information about the dataset. Specifically, DP provides rigorous statistical guarantees that the inclusion of an individual in the dataset is almost indistinguishable. Thanks to its powerful features, differentially private data-releasing mechanisms, e.g., noise-adding, have been incorporated in some machine learning algorithms to preserve privacy, e.g., empirical risk minimization (ERM) (Bassily et al., 2014), principal component analysis (Chaudhuri et al., 2012), federated learning (Agarwal et al., 2021), and decentralized learning (Hale and Egerstedt, 2015; Huang et al., 2015).

The DP constraint induces a tradeoff between privacy and utility in learning algorithms. While a number of attempts have been made to establish such a tradeoff in decentralized learning, the obtained upper bounds are arguably not tight enough. For example, Huang et al. (2015) developed a differentially private decentralized GD algorithm by perturbing the local output with Laplace noise. Notably, the learning rate is set exponentially diminishing such that the sensitivity of the algorithm also decreases linearly. By doing so, a summable sequence of privacy budgets can be assigned to individual iterations, making the whole iterative process $\epsilon$-DP. However, such choice of learning rate slows down the convergence dramatically and results in a utility loss in the order of $\mathcal{O}(m/\epsilon^2)$, where $m$ denotes the dimension of the decision variable. Under the more reasonable learning rate $\Theta(1/\sqrt{t})$, the utility loss can be im-

⋆ This work was supported by the Knut and Alice Wallenberg Foundation, the Swedish Foundation for Strategic Research, and the Swedish Research Council.

10.1016/j.ifacol.2022.07.233

proved to $\mathcal{O}\left(\sqrt[4]{\frac{mn^2}{\epsilon}}\right)$ (Han et al., 2016), where $n$ denotes the number of nodes. Along this line of research, Zhu et al. (2018); Xiong et al. (2020); Han et al. (2021) extended the algorithm to time-varying objective functions, and Ding et al. (2021) advanced the convergence to linear based on an additional gradient-tracking scheme. In these works, however, $\epsilon$-DP is proven only for each iteration, leading to a cumulative privacy loss of $t\epsilon$ after $t$ rounds of execution. To tackle *regularized* learning problems, the alternating direction method of multipliers (ADMM) has been used to design decentralized algorithms with DP (Zhang and Zhu, 2016; Zhang et al., 2018a). However, an explicit tradeoff analysis between privacy and utility was missing. Recently, Xiao and Devadas (2021) investigated the privacy guarantee produced not only by random noise injection but also by *mixup* (Zhang et al., 2018b), i.e., a random convex combination of inputs. The utility-privacy tradeoff in linearized ADMM and GD-based decentralized algorithms were captured by the bound $\mathcal{O}\left(\frac{m}{\sqrt{n}\epsilon}\right)$. However, there still exists a substantial gap between the available bounds for the decentralized and centralized algorithms (Bassily et al., 2014). For example, a mini-batch version (a fixed number $n$ of training samples) of the results in (Bassily et al., 2014), which corresponds to decentralized learning over complete graphs, gives the much tighter utility bound $\mathcal{O}\left(\frac{m\log^2(q/\delta)\log(1/\delta)}{q^2\epsilon^2}\right)$ to achieve $(\epsilon,\delta)$-DP, where $q \gg 1/\epsilon$ represents the quotient of total number of samples and $n$. This observation naturally motivates an interesting question. *Is a comparable utility bound achievable for the general decentralized setup?* The main theme of the current paper is to answer this question.

In this work, we consider the decentralized regularized ERM problem, where each node has a convex, possibly non-smooth, loss function defined by its own dataset and shares the same strongly convex regularization term. Different from existing approaches, we perturb the stochastic subgradient evaluated over a single data sample with proper noise, based on which a private decentralized DA (DDA) algorithm is developed. To obey $(\epsilon,\delta)$-DP, the proposed algorithm bears utility loss in the order of $\mathcal{O}(\frac{m\log(1/\delta)}{q^2\epsilon^2})$, which is significantly smaller than existing decentralized algorithms with DP. We remark that, when removing subgradient perturbation, the non-private version of our algorithm improves the standard convergence rate $\mathcal{O}(1/\sqrt{t})$ of DDA (Duchi et al., 2011) to the optimal rate $\mathcal{O}(1/t)$ for non-smooth strongly convex problems.

The rest of the paper is organized as follows. Section 2 formulates the problem and introduces some preliminaries. Section 3 presents our algorithm and its theoretical properties. Finally, some experimental results are provided in Section 4.

## 2. PRELIMINARIES

### 2.1 Basic Setup

Consider a decentralized network captured by an undirected graph with weights: $(N, W)$. $N = \{1, \ldots, n\}$ denotes the set of $n$ nodes. $W \in [0,1]^{n\times n}$ is a symmetric doubly stochastic matrix, where the $(i,j)$-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. We denote the set of $i$'s neighbors by $N_i := \{j | j \in N \setminus \{i\}, w_{ij} > 0\}$.

Each node $i$ possesses a local dataset $D_i = \{\xi_i^{(1)}, \ldots, \xi_i^{(q_i)}\}$ that contains a finite number $q_i$ of data samples. The nodes aim to cooperatively solve the following regularized ERM problem

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

where

$$f_i(x) = \frac{1}{q_i}\sum_{j=1}^{q_i} l_i(x, \xi_i^{(j)})$$

with the loss function $l_i(x, \xi_i)$ measuring the accuracy of the learned model (characterized by $x$) over each data sample (denoted by $\xi_i^{(j)}$). $h(x)$ is a regularization term with domain $\mathrm{dom}(h) := \{x \in \mathbb{R}^m | h(x) < +\infty\}$.

Our goal is to solve Problem (1) in a fully decentralized manner, while providing rigorous privacy guarantee for each data sample in $D := \cup_{i\in N} D_i$.

### 2.2 Conventional DDA Algorithm

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

$$x_i(t) = \operatorname*{argmin}_{x\in\mathbb{R}^m} \{\langle z_i(t), x\rangle + t(h(x)) + \gamma(t)d(x)\} \tag{2}$$

and

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

where $\hat{g}_j(t) \in \partial l_j(x_j(t), \xi_j(t))$ with $\xi_j(t) \sim_u D_j$, $\{\gamma(t)\}_{t\geq 1}$ is a non-decreasing sequence of parameters, and $w_{ij}$ is the $(i,j)$-th entry of the mixing matrix $W$. Throughout the process, each node passes $z_i$ to its immediate neighbors and updates $x_i$ according to (2). For non-smooth convex functions, the conventional DDA converges at $\mathcal{O}(1/\sqrt{t})$ (Duchi et al., 2011; Colin et al., 2016).

### 2.3 Differential Privacy

DP has been recognized as the gold standard in quantifying the individual privacy preservation for randomized algorithms. It refers to the property of a randomized algorithm that the presence or absence of an individual cannot be distinguished based on the output of the algorithm. Formally, DP is defined as follows.

*Definition 1.* (($\varepsilon$, $\delta$)-DP). A randomized algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies ($\varepsilon$, $\delta$)-DP if for every pair of neighboring datasets $D, D' \in \mathcal{D}$, i.e., datasets that exactly differ in one entry, and for any subset $\mathcal{O} \subseteq \mathcal{R}$ we have

$$Pr[\mathcal{A}(D) \in \mathcal{O}] \leq e^{\varepsilon} Pr[\mathcal{A}(D') \in \mathcal{O}] + \delta.$$

If $\delta = 0$, the mechanism is called $\epsilon$-DP.

In a decentralized optimization algorithm such as (2) and (3), new messages bearing information about the local training data are exchanged among the nodes, which gives privacy concerns. Thus, when evaluating the privacy loss in a decentralized and iterative algorithm, the messages broadcast up to time $t$ should be taken as the output of the algorithm. Formally, the DP definition is tailored for decentralized and iterative algorithms as follows.

*Definition 2.* Consider a decentralized network described by $(N, W)$, where each node has its own dataset $D_i$. Let $\{z_i(t), i \in N\}$ denote the set of messages exchanged among the nodes at iteration $t$. A decentralized and iterative algorithm satisfies $(\epsilon, \delta)$-DP during $T$ iterations if for every pair of neighboring datasets $D = \cup_{i\in N} D_i$ and $D' = \cup_{i\in N} D'_i$, and for any set of possible outputs $\mathcal{O}$ during $T$ iterations we have

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

## 3. DIFFERENTIALLY PRIVATE DDA

### *3.1 Private DDA with Subgradient Perturbation*

In the literature, there are two main types of approaches to achieve DP. The first type of approaches disturbs the output of a non-private algorithm (Zhang et al., 2017). However, they cannot be generalized to the setting with non-smooth regularization. The second type perturbs the (sub)gradient used in the optimization algorithm (Bassily et al., 2014). To support non-smooth regularization, we perturb the stochastic subgradient $\hat{g}_i$ with a Gaussian noise vector $\nu_i(t) \sim \mathcal{N}(0, \sigma^2 I)$. The scale of the noise, i.e., $\sigma^2$, shall be calibrated according to the sensitivity of the Gaussian mechanism to fulfill a prescribed privacy budget. Based on this strategy and partially motivated by the conventional DDA method in (3), we develop the following update rule

$$z_i(t+1) = \sum_{j=1}^{n} w_{ij} \left(z_j(t) + a(t)\left(\hat{g}_j(t) + \nu_j(t)\right)\right). \quad (4)$$

where $\hat{g}_j(t) \in \partial l_j(x_j(t), \xi_j(t))$ with $\xi_j(t) \sim_u D_j$, $w_{ij}$ is the $(i,j)$-th element in the mixing matrix $W$, and $\{a(t)\}_{t\geq 1}$ is a sequence of non-decreasing parameters. By setting $a(t) = 1$ and $\nu_i(t) = 0, i = 1, \ldots, n$, (4) reduces to the conventional update in (3). We will show that, when Problem (1) is strongly convex, faster convergence can be attained with proper choices of $\{a(t)\}_{t\geq 1}$. Accordingly, each nodes solves

$$\begin{aligned} &x_i(t+1) \\ &= \operatorname*{argmin}_{x\in\mathbb{R}^m} \left\{\langle z_i(t+1), x\rangle + A(t+1)h(x) + \gamma(t+1)d(x)\right\} \end{aligned} \quad (5)$$

to generate its local estimate about the global optimum, where $A(t) = \sum_{\tau=1}^{t} a(\tau)$ and $\{\gamma(t)\}_{t\geq 1}$ [1] is a non-decreasing sequence of positive parameters. By convention, we let $A(0) = a(0) = 0$ and $\gamma(0) = 0$. The overall procedure is summarized in Algorithm 1.

**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: $\tilde{x}_i(T) = A(T)^{-1} \sum_{\tau=1}^{T} a(\tau) x_i(\tau)$
Initialize: set $z_i(1) = 0$ and identify $x_i(1)$ according to (5) for all $i \in N$
for $t = 1, 2, \ldots, T$ do
    In parallel for agents $i \in N$:
    randomly sample $\xi_i(t) \sim_u D_i$
    generate noise $\nu_i(t) \sim \mathcal{N}(0, \sigma^2 I)$
    collect $z_j(t) + a(t)(\hat{g}_j(t) + \nu_j(t))$ from all agents $j \in N_i$
    update $z_i(t+1)$ by (4)
    compute $x_i(t+1)$ by (5)
end for
```

### *3.2 Privacy Guarantee*

Before proceeding to the privacy guarantee, we make the following assumption.

*Assumption 1.* ($L$-Lipschitz). Each $l_i(\cdot, \xi_i)$ is $L$-Lipschitz, that is, $\forall x, y \in dom(h)$

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

Next, we state the privacy-preserving property of Algorithm 1 in Theorem 1, whose proof can be found in Appendix A.

*Theorem 1.* (Privacy Guarantee). Given $0 < \epsilon \leq 1$ and $0 < \delta \leq 1/3$. If

$$\sigma^2 \geq \frac{12L^2 T \log(1/\delta)}{q^2 \epsilon^2}$$

where $q = \min_{i\in N} q_i$, then Algorithm 1 is $(\epsilon, \delta)$-DP.

Theorem 1 emphasizes that, to achieve a prescribed privacy budget during $T$ iterations, the noise variance $\sigma^2$ depends on the DP parameters $(\epsilon, \delta)$, the Lipschitz constant $L$ of the loss, and the number of samples per local dataset.

### *3.3 Privacy-Utility Tradeoff*

For the convergence of the algorithm, we make the following technical assumptions.

*Assumption 2.* (Spectral Gap). For the symmetric doubly stochastic matrix $W$, we have its second largest singular value, denoted by $\beta = \sigma_2(W)$, smaller than 1.

*Assumption 3.* (Convexity). i) $h(\cdot)$ is a proper closed strongly convex function with modulus $\mu > 0$, 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 $dom(h)$.

Motivated by some existing works (Duchi et al., 2011), we first present the convergence property of an auxiliary sequence $\{y(t)\}_{t\geq 0}$, which then immediately suggests the convergence of the sequence $\{x_i(t) : i \in N\}_{t\geq 1}$ generated by Algorithm 1. In particular, we define

$$y(t) = \operatorname*{argmin}_{x\in\mathbb{R}^m} \left\{\langle \overline{z}(t), x\rangle + A(t)h(x) + \gamma(t)d(x)\right\}, \quad (6)$$

where $\overline{z}(t) = \frac{1}{n}\sum_{i=1}^{n} z_i(t)$ and $\{z_i(t) : i \in N\}_{t\geq 1}$ are generated by Algorithm 1. To streamline the analysis, we

[1] We will show that the parameter $\{\gamma(t)\}_{t\geq 1}$ can be set constant (including 0) for non-smooth strongly convex problems. However, we keep it in (5) to be consistent with the conventional DDA update in (2).

also introduce the following notation: $\zeta_i(t) = \hat{g}_i(t) + \nu_i(t)$, $\overline{\zeta}(t) = n^{-1}\sum_{i=1}^{n}\zeta_i(t)$.

We study the convergence of $\{y(t)\}_{t\geq 0}$ in the following theorem, whose proof is sketched in Appendix B due to limited space.

*Theorem 2.* (Convergence of $\{y(t)\}_{t\geq 0}$). Suppose Assumptions 1, 2, and 3 are satisfied. For all $t \geq 1$, we have

$$\begin{aligned}&\mathbb{E}[F(\tilde{y}(t)) - F(x^*)]\\ &\leq \frac{1}{A(t)}\Big(\gamma(t)d(x^*) + \sum_{\tau=1}^{t}\frac{a(\tau)^2}{\mu A(\tau)+\gamma(\tau)}M \\ &\qquad + \sum_{\tau=1}^{t}\frac{a(\tau)^2}{\mu A(\tau)+\gamma(\tau)}\Big(\frac{m\sigma^2}{2} + \frac{2\sqrt{m}L\sigma}{1-\beta}\Big)\Big)\end{aligned} \tag{7}$$

where $\tilde{y}(t) = A(t)^{-1}\sum_{\tau=1}^{t}a(\tau)y(\tau)$, $M = L^2/2 + 2L^2/(1-\beta)$, and $\sigma$ is defined in Theorem 1.

Based on the convergence result in Theorem 2, we provide an explicit utility-privacy tradeoff for Algorithm 1 in the following corollary.

*Corollary 1.* (Utility Loss). Suppose the premise of Theorem 2 holds. If

$$a(t) = t \quad \text{and} \quad \gamma(t) = 0, \tag{8}$$

then there exists $T \geq 0$ for all $i \in N$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\tilde{x}_i(T) - x^*\|^2] \leq \mathcal{O}\left(\frac{mL^2\log(1/\delta)}{\mu^2 q^2\epsilon^2}\right) \tag{9}$$

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

**Proof.** We obtain from the update of $A(t)$ in Algorithm 1 that $\sum_{\tau=1}^{t}\frac{a(\tau)^2}{\mu A(\tau)+\gamma(\tau)} = \sum_{\tau=1}^{t}\frac{2\tau^2}{\mu\tau(\tau+1)} \leq \frac{2t}{\mu}$. By (7), we have

$$\begin{aligned}&\mathbb{E}[F(\tilde{y}(t)) - F(x^*)]\\ &\leq \frac{4M}{\mu(t+1)} + \frac{16\sqrt{3m\log(1/\delta)}L^2}{\mu(1-\beta)q\epsilon\sqrt{t+1}} + \frac{24mL^2\log(1/\delta)}{\mu q^2\epsilon^2}\end{aligned} \tag{10}$$

Upon using convexity of $\|\cdot\|^2$ and (B.2), we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\tilde{x}_i(t) - \tilde{y}(t)\|^2] \leq \frac{8(L^2+m\sigma^2)(\log t + 1)}{\mu^2 t(t+1)(1-\beta)^2}. \tag{11}$$

Due to $\mu$-strong convexity of $F$, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|\tilde{x}_i(t) - x^*\|^2 &\leq \frac{2}{n}\sum_{i=1}^{n}\|\tilde{x}_i(t) - \tilde{y}(t)\|^2 + 2\|\tilde{y}(t) - x^*\|^2\\ &\leq \frac{2}{n}\|\tilde{x}_i(t) - \tilde{y}(t)\|^2 + \frac{4}{\mu}\Big(F(\tilde{y}(t)) - F(x^*)\Big),\end{aligned}$$

which together with (10) and (11) gives the desired result.

*Remark 1.* Compared to existing decentralized optimization methods with DP (Huang et al., 2015; Han et al., 2016; Xiao and Devadas, 2021), Algorithm 1 attains a much tighter bound of the utility loss, suggesting a better tradeoff is achieved between privacy and utility.

As an immediate consequence of Corollary 1, we show in Corollary 2 that Algorithm 1 attains $\mathcal{O}(1/t)$ when $\sigma = 0$.

*Corollary 2.* (Rate of Convergence if $\sigma = 0$). Suppose the premise of Theorem 2 holds. If $\sigma = 0$,

$$a(t) = t \quad \text{and} \quad \gamma(t) = 0,$$

then for all $t \geq 1$, and $i \in N$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\tilde{x}_i(t) - x^*\|^2] \leq \frac{16}{t+1}\left(\frac{L^2(\log t + 1)}{\mu^2(1-\beta)^2 t} + \frac{M}{\mu^2}\right), \tag{12}$$

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

**Proof.** The proof is straightforward by adapting the proof of Corollary 1 to the case with $\sigma = 0$.

*Remark 2.* Corollary 2 illustrates that the non-private version of Algorithm 1 attains $\mathcal{O}(1/t)$ convergence when Problem (1) is strongly convex, which is optimal for non-smooth stochastic optimization (Yuan et al., 2018). Compared to the optimal algorithm in (Yuan et al., 2018), Algorithm 1 is robust to non-smooth regularizers, e.g., elastic net.

## 4. EXPERIMENTS

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

### 4.1 Setup

In the experiments, we consider a ring network of $n = 20$ nodes. The corresponding mixing matrix is created with uniform weights. We use the benchmark dataset *epsilon* (Sonnenburg et al., 2006), where the 400,000 data samples are evenly assigned to $n = 20$ working nodes at random. We consider the following regularized SVM problem

$$\min_{x}\left\{F(x) = \frac{1}{n}\sum_{i=1}^{n}f_i(x) + \frac{\mu}{2}\|x\|^2\right\} \tag{13}$$

where $\mu = 0.0005$,

$$f_i(x) = \frac{1}{q}\sum_{j=1}^{q}\max\left\{0, 1 - y_i^{(j)}\left\langle C_i^{(j)}, x\right\rangle\right\}, \tag{14}$$

and $\{C_i^{(j)}, y_i^{(j)}\}_{j=1}^{q=20000} := D_i$ are data samples private to node $i$. For the parameters of DP $(\epsilon, \delta)$, we consider $\epsilon \in \{0.2, 0.4, 0.6, 0.8, 1\}$ and $\delta = 0.01$. For Algorithm 1, we set $a(t) = t$ and $\gamma(t) = 20$. For the algorithm in (Duchi et al., 2011), we let $a(t) = 1$ and $\gamma(t) = 20 + \sqrt{\mu t}$.

### 4.2 Results

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). Each experiment is repeated three times; the mean curve of the results is plotted.

We observe from Figure 1 that the private version of Algorithm 1, at the expense of achieving $(1, 0.01)$-DP, presents a slower convergence than its non-private counterpart. As a result, the private algorithm yields a slightly lower testing accuracy. However, they both outperform the algorithm in (Duchi et al., 2011) in terms of convergence speed and testing accuracy. Furthermore, Figure 2 highlights that the utility degenerates when the DP parameter $\epsilon$ decreases. This is because a smaller $\epsilon$ suggests a tighter DP constraint that requires a stronger noise to perturb the subgradient, as revealed in Theorem 1.

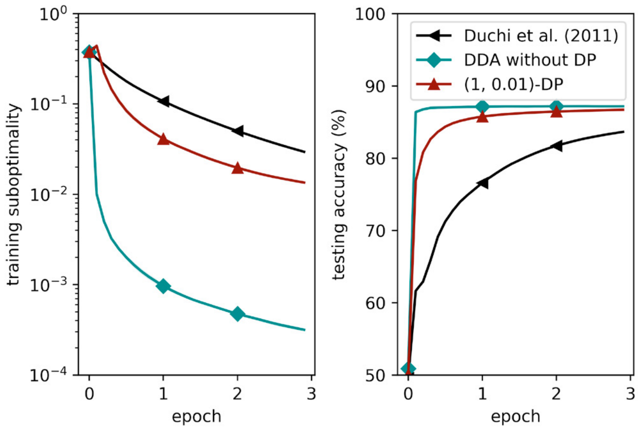

Fig. 1. Performance comparison between Algorithm 1 and (Duchi et al., 2011).

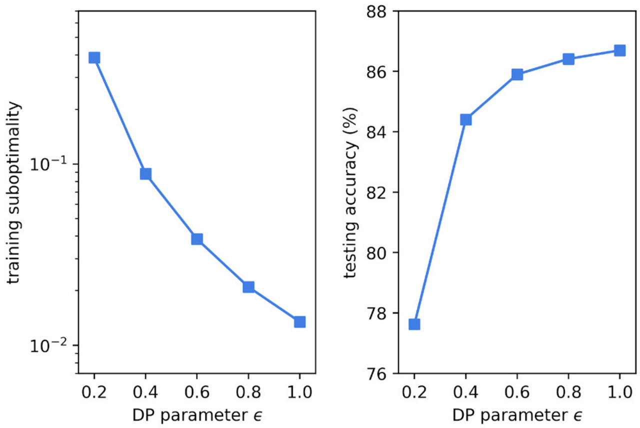

Fig. 2. Privacy–utility tradeoff. The suboptimality and accuracy are evaluated after 3-epoch training.

## REFERENCES

Agarwal, N., Kairouz, P., and Liu, Z. (2021). The skellam mechanism for differentially private federated learning. *Advances in Neural Information Processing Systems*, 34.

Bassily, R., Smith, A., and Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, 464–473. IEEE.

Chaudhuri, K., Sarwate, A., and Sinha, K. (2012). Near-optimal differentially private principal components. *Advances in Neural Information Processing Systems*, 25, 989–997.

Colin, I., Bellet, A., Salmon, J., and Clémençon, S. (2016). Gossip dual averaging for decentralized optimization of pairwise functions. In *International Conference on Machine Learning*, 1388–1396. PMLR.

Ding, T., Zhu, S., He, J., Chen, C., and Guan, X.P. (2021). Differentially private distributed optimization via state and direction perturbation in multi-agent systems. *IEEE Transactions on Automatic Control.*

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

Fredrikson, M., Jha, S., and Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 1322–1333.

Hale, M. and Egerstedt, M. (2015). Differentially private cloud-based multi-agent optimization with constraints. In *2015 American Control Conference (ACC)*, 1235–1240. IEEE.

Han, D., Liu, K., Lin, Y., and Xia, Y. (2021). Differentially private distributed online learning over time-varying digraphs via dual averaging. *International Journal of Robust and Nonlinear Control.*

Han, S., Topcu, U., and Pappas, G.J. (2016). Differentially private distributed constrained optimization. *IEEE Transactions on Automatic Control*, 62(1), 50–64.

Huang, Z., Mitra, S., and Vaidya, N. (2015). Differentially private distributed optimization. In *Proceedings of the 2015 International Conference on Distributed Computing and Networking*, 1–10.

Jakovetic, D., Xavier, J., and Moura, J.M. (2011). Cooperative convex optimization in networked systems: Augmented lagrangian algorithms with directed gossip communication. *IEEE Transactions on Signal Processing*, 59(8), 3889–3902.

Liu, C., Zhou, Z., Pei, J., Zhang, Y., and Shi, Y. (2021). Decentralized composite optimization in stochastic networks: A dual averaging approach with linear convergence. *arXiv preprint arXiv:2106.14075.*

Mironov, I. (2017). Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, 263–275. IEEE.

Nedic, A., Ozdaglar, A., and Parrilo, P.A. (2010). Constrained consensus and optimization in multi-agent networks. *IEEE Transactions on Automatic Control*, 55(4), 922–938.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Sonnenburg, S., Rätsch, G., Schäfer, C., and Schölkopf, B. (2006). Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7, 1531–1565.

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

Xiao, L. (2009). Dual averaging method for regularized stochastic learning and online optimization. *Advances in Neural Information Processing Systems*, 22, 2116–2124.

Xiong, Y., Xu, J., You, K., Liu, J., and Wu, L. (2020). Privacy-preserving distributed online optimization over unbalanced digraphs via subgradient rescaling. *IEEE Transactions on Control of Network Systems*, 7(3), 1366–1378.

Yuan, D., Hong, Y., Ho, D.W., and Jiang, G. (2018). Optimal distributed stochastic mirror descent for strongly convex optimization. *Automatica*, 90, 196–203.

Zhang, C., Ahmad, M., and Wang, Y. (2018a). Admm based privacy-preserving decentralized optimization. *IEEE Transactions on Information Forensics and Security*, 14(3), 565–580.

Zhang, H., Cisse, M., Dauphin, Y.N., and Lopez-Paz, D. (2018b). mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations.*

Zhang, J., Zheng, K., Mou, W., and Wang, L. (2017). Efficient private erm for smooth objectives. *arXiv preprint arXiv:1703.09947.*

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

Zhu, J., Xu, C., Guan, J., and Wu, D.O. (2018). Differentially private distributed online algorithms over time-varying directed networks. *IEEE Transactions on Signal and Information Processing over Networks*, 4(1), 4–17.

Zhu, L. and Han, S. (2020). Deep leakage from gradients. In *Federated learning*, 17–31. Springer.

## Appendix A. PROOF OF THEOREM 1

To track the privacy loss of an iterative algorithm after $T \geq 1$ rounds of iteration, we use Rényi DP (RDP) (Mironov, 2017). To proceed, we present the definition and some main properties of RDP (Mironov, 2017).

*Definition 3.* ($(\alpha, \epsilon)$-RDP). A randomized algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{R}$ is $\rho$-RDP of order $\alpha > 1$, or $(\alpha, \rho)$-RDP for short, if for every pair of neighboring datasets $D, D' \in \mathcal{D}$ we have

$$\mathrm{D}_\alpha(\mathcal{A}(D)||\mathcal{A}(D')) \leq \rho,$$

where $\mathrm{D}_\alpha(\mathcal{A}(D)||\mathcal{A}(D'))$ is the $\alpha$-Rényi divergence between $\mathcal{A}(D)$ and $\mathcal{A}(D')$, i.e.,

$$\begin{aligned}&\mathrm{D}_\alpha(\mathcal{A}(D)||\mathcal{A}(D'))\\ &= \frac{1}{\alpha - 1} \log \int_{\mathcal{R}} Pr[\mathcal{A}(D) = z]^\alpha Pr[\mathcal{A}(D') = z]^{1-\alpha} dz.\end{aligned}$$

*Lemma 1.* (Composition of RDP). Given $T$ randomized algorithms $\mathcal{A}_1, \ldots, \mathcal{A}_\tau, \ldots, \mathcal{A}_T : \mathcal{D} \to \mathcal{R}$, each of which is $(\alpha, \rho(\tau))$-RDP. 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 $(\alpha, \sum_{\tau=1}^{T} \rho(\tau))$-RDP.

*Lemma 2.* (Relation between RDP and DP). If a randomized algorithm is $(\alpha, \rho)$-RDP, then it is $(\rho + \frac{\log(1/\delta)}{\alpha - 1}, \delta)$-DP, $\forall \delta \in (0, 1)$.

*Lemma 3.* (Gaussian Mechanism). Consider the Gaussian mechanism for answering the query $r : \mathcal{D} \to \mathbb{R}^m$:

$$\mathcal{M} = r(D) + \nu, \tag{A.1}$$

where $D \in \mathcal{D}$, $\nu \sim \mathcal{N}(0, \sigma^2 I)$. If $\sigma^2 = \frac{\Delta^2}{2\rho}$ where $\Delta$ denotes the sensitivity of $r$, i.e., $\Delta = \sup_{D,D'} \|r(D) - r(D')\|$, then (A.1) is $(\alpha, \alpha\rho)$-RDP.

We are now in a position to prove Theorem 1.

Consider the Gaussian mechanism

$$\mathcal{M}_t = \hat{\mathbf{g}}(t) + \boldsymbol{\nu}(t) \tag{A.2}$$

where

$$\hat{\mathbf{g}}(t) = \begin{bmatrix} \hat{g}_1(t) \\ \vdots \\ \hat{g}_n(t) \end{bmatrix}, \quad \boldsymbol{\nu}(t) = \begin{bmatrix} \nu_1(t) \\ \vdots \\ \nu_n(t) \end{bmatrix}.$$

Its sensitivity can be derived as

$$\Delta(t) = \frac{1}{q} \sup_{D,D'} \|\hat{\mathbf{g}}_D(t) - \hat{\mathbf{g}}_{D'}(t)\| \leq \frac{2L}{q}$$

where $\hat{\mathbf{g}}_D(t), \hat{\mathbf{g}}_{D'}(t)$ denote the subgradient $\hat{\mathbf{g}}(t)$ evaluated over the two neighboring datasets $D, D'$, respectively. From Lemma 3, the Gaussian mechanism in (A.2) at every iteration $t$ is $(\alpha, \alpha\rho(t))$-RDP with

$$\rho(t) = \frac{\epsilon^2}{6 \log(1/\delta) T}. \tag{A.3}$$

By the post-processing theorem, Algorithm 1 also satisfies $(\alpha, \alpha\rho(t))$-RDP at every iteration $t$. Upon using Lemma 1, we further obtain, after $T$ iterations, Algorithm 1 is $(\alpha, \alpha\rho)$-RDP with $\rho = \epsilon^2/(6\log(1/\delta))$. Therefore, based on Lemma 2, Algorithm 1 is $(\epsilon', \delta)$-DP with

$$\epsilon' = \frac{\alpha\epsilon^2}{6\log(1/\delta)} + \frac{\log(1/\delta)}{\alpha - 1} \leq \epsilon \left( \frac{1}{6} + \frac{1}{2} + \frac{1}{3} \right) = \epsilon,$$

where we use $\alpha = 1 + 3\log(1/\delta)/\epsilon$, $\epsilon^2 \leq \epsilon \leq 1$ and $0 < \delta \leq 1/3$ to obtain the inequality.

## Appendix B. PROOF SKETCH OF THEOREM 2

Before proving Theorem 2, we present two useful lemmas whose proofs are omitted for brevity.

*Lemma 4.* For the sequence $\{x_i(t) : i \in N\}_{t \geq 1}$ generated by Algorithm 1 and the auxiliary sequence $\{y(t)\}_{t \geq 1}$ defined in (6), one has that for all $t \geq 1$ and $i \in N$,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\|x_i(t) - y(t)\|\right] \leq \frac{a(t)(L + \sqrt{m}\sigma)}{(1 - \beta)(\mu A(t) + \gamma(t))} \tag{B.1}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\|x_i(t) - y(t)\|^2\right] \leq \frac{a(t)^2(L^2 + m\sigma^2)}{(1 - \beta)^2(\mu A(t) + \gamma(t))^2}. \tag{B.2}$$

*Lemma 5.* For all $t \geq 1$, we have

$$\begin{aligned}&\sum_{\tau=1}^{t} a(\tau) \left( \langle \bar{\zeta}(\tau), y(\tau) - x^* \rangle + h(y(\tau)) - h(x^*) \right)\\ &\leq \frac{1}{2} \sum_{\tau=1}^{t} \frac{a(\tau)^2}{\mu A(\tau) + \gamma(\tau)} \|\bar{\zeta}(\tau)\|^2 + \gamma(t) d(x^*).\end{aligned} \tag{B.3}$$

We are ready to prove Theorem 2.

Following the procedure in (Duchi et al., 2011, Theorem 1), we can use the convexity of $f = \frac{1}{n}\sum_{i=1}^{n} f_i$, and the Lipschitz continuity of $f_j, j \in N$, and Lemma 5 (an improved result over (Duchi et al., 2011, Lemma 2)) to obtain

$$\begin{aligned}&A(t)\left[F(\tilde{y}(t)) - F(x^*)\right]\\ &\leq A(t)\left[f(\tilde{y}(t)) - f(x^*)\right] + \sum_{\tau=1}^{t} a(\tau)\left(h(y(\tau)) - h(x^*)\right)\\ &\leq \frac{1}{n} \sum_{j=1}^{n} \sum_{\tau=1}^{t} L a(\tau) \|x_j(\tau) - y(\tau)\|\\ &\quad + \frac{1}{2} \sum_{\tau=1}^{t} \frac{a(\tau)^2}{\mu A(\tau) + \gamma(\tau)} \|\bar{\zeta}(\tau)\|^2 + \gamma(t) d(x^*)\\ &\quad + \frac{1}{n} \sum_{\tau=1}^{t} \sum_{j=1}^{n} a(\tau) \Big( \langle \zeta_j(\tau), x_j(\tau) - y(t) \rangle\\ &\qquad + \langle g_j(\tau) - \zeta_j(\tau), x_j(\tau) - x^* \rangle \Big).\end{aligned}$$

Next, we exploit the statistical independence and (B.1) to bound the right-hand side of the above equation in expectation, arriving at (7) as desired.