

# Local Differential Privacy in Decentralized Optimization

by

Hanshen Xiao

B.S., Mathematics, Tsinghua University (2017)

Submitted to the Department of Electrical Engineering and Computer Science

in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

© Massachusetts Institute of Technology 2019. All rights reserved.

**Signature redacted**

Author .....

Department of Electrical Engineering and Computer Science

August 30, 2019

**Signature redacted**

Certified by .....

Srini Devadas

Professor of Electrical Engineering and Computer Science

Thesis Supervisor

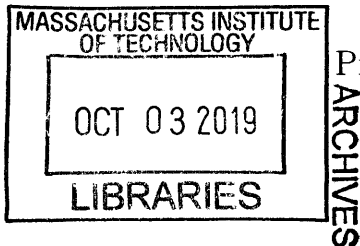
**Signature redacted**

Accepted by .....

Leslie A. Kolodziejski

Professor of Electrical Engineering and Computer Science

Chair, Department Committee on Graduate Students





# Local Differential Privacy in Decentralized Optimization

by

Hanshen Xiao

Submitted to the Department of Electrical Engineering and Computer Science  
on August 30, 2019, in partial fulfillment of the  
requirements for the degree of  
Master of Science in Computer Science and Engineering

## Abstract

Privacy concerns with sensitive data are receiving increasing attention. In this thesis, we study local differential privacy (LDP) in interactive decentralized optimization. Comparing to central differential privacy (DP), where a centralized curator maintains the dataset, LDP is a stronger notion yet with industrial adoption, which allows data of an individual to be privatized before sharing. Consequently, more challenges are encountered to build efficient statistical analyzer in LDP setting.

Towards practical decentralized optimization in LDP, we extend LDP into a more comprehensive notion which provides both worst and average case privacy guarantees. Accordingly, two approaches to sharpen utility-privacy tradeoff are proposed for the worst and the average, respectively: First, cryptographically incorporated with merely linear secret sharing, we show the privacy guarantee can be improved by a factor of  $\sqrt{N'}$  where  $N'$  amongst all  $N$  agents are semi-honest. Second, we take Alternating Direction Method of Multipliers (ADMM), and decentralized (stochastic) gradient descent (D(S)GD) as two concrete examples to propose a framework of first-order based optimization with random local aggregators. We prove such local randomization lead to the same utility guarantee but amplify average LDP by a constant, empirically around 30%. Thorough experiments support our theory.

Thesis Supervisor: Srini Devadas

Title: Professor of Electrical Engineering and Computer Science



## Acknowledgments

First and foremost, I would like to thank my advisor, Srinivas Devadas. During my first two years in MIT, Srinivas always gives me complete freedom and encourages me to study anything I am interested in. Srinivas has been an endless source of creative ideas along with exquisite taste of research.

I am also grateful to Yu Ye. We initiated this project in a brainstorm when both of us did not even know differential privacy at that time. Those discussions at the early stage of the project where we try to formalize the notion of privacy motivates me to think more about the nature of privacy from many interesting perspectives. I would also like to thank Prof. Salil Vadhan, Prof. Adam Smith and Tiancong Chen for many meaningful comments and always being available to answer my questions.

Outside of the thesis, I had countless pleasure to work with Jun Wan, Fengyi Li, Di Wang, Nan Du and Guoqiang Xiao on various topics varying from random graphs, Byzantine protocols and statistical learning. I am deeply grateful to my family for their continuous support.



# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>   | <b>13</b> |
| 1.1      | Problem Statement . . . . .   | 15        |
| 1.2      | Differential Privacy . . . . .  | 16        |
| 1.3      | Motivation . . . . .  | 16        |
| 1.4      | Methodology Overview . . . . .  | 17        |
| 1.5      | Contribution and Organization . . . . .                                 | 18        |
| <b>2</b> | <b>Algorithm Description and Privacy Analysis</b>                       | <b>21</b> |
| 2.1      | Secret-sharing Based Privacy Amplification . . . . .                    | 21        |
| 2.2      | Decentralized Optimization with Randomized Local Aggregations . . . . . | 23        |
| <b>3</b> | <b>Convergence and Utility Tradeoff Analysis</b>                        | <b>29</b> |
| 3.1      | Utility Analysis of Algorithm 1 . . . . .                               | 30        |
| 3.2      | Utility Analysis of Algorithm 2 . . . . .                               | 32        |
| 3.3      | Convergence in a View of Random Stochastic Matrix . . . . .             | 32        |
| <b>4</b> | <b>Conclusion</b>   | <b>37</b> |
| <b>A</b> | <b>Proof of Theorem 2.1.1</b>   | <b>39</b> |
| <b>B</b> | <b>Proof of Lemma 2.2.1</b>   | <b>41</b> |
| <b>C</b> | <b>Proof of Theorem 2.2.1</b>   | <b>43</b> |
| <b>D</b> | <b>Proof of Theorem 2.2.2</b>   | <b>47</b> |

|          |  |           |
|----------|--|-----------|
| <b>E</b> | <b>Simulation Results</b>  | <b>49</b> |
| <b>F</b> | <b>Proof of Theorem 3.1.1</b>  | <b>55</b> |
| <b>G</b> | <b>Proof of Theorem 3.1.2</b>  | <b>59</b> |
| <b>H</b> | <b>Proof of Theorem 3.1.3</b>  | <b>61</b> |
| <b>I</b> | <b>Proof of Theorem 3.2.1</b>  | <b>63</b> |
| <b>J</b> | <b>Utility Analysis of Algorithm 1 under Strongly-convex Assumptions</b> | <b>67</b> |



# List of Figures

|     |  |    |
|-----|--|----|
| 2-1 | Expected Average Privacy Amplification and Performance Comparison<br>with Existing Works . . . . . | 26 |
| 3-1 | Rate of $\prod_{k=1}^K W_k$ towards identical rows (a) . . . . .                                   | 36 |
| 3-2 | Rate of $\prod_{k=1}^K W_k$ towards identical rows (b) . . . . .                                   | 36 |
| E-1 | Figure E-1. (b). Simulation on Graphs $N = 100$ , $ \mathcal{E}  = 200$ . . . . .                  | 51 |
| E-2 | Figure E-2: (c). Simulation on Graphs $N = 10$ , $ \mathcal{E}  = 40$ . . . . .                    | 52 |
| E-3 | Figure E-3: (d). Simulation on Graphs $N = 10$ , $ \mathcal{E}  = 20$ . . . . .                    | 52 |
| E-4 | Figure E-4: (d). Simulation on Graphs $N = 100$ , $ \mathcal{E}  = 200$ . . . . .                  | 52 |



# List of Tables



# Chapter 1

## Introduction

Due to the underlying intensive computation and memory requirement in large-scale machine learning, distributed learning has witnessed tremendous development in recent years. In general, there exist two typical scenarios of distributed optimization. The first one assumes a central server to collect and average out local estimates from each agent to update the global model, for example, federated learning [1],[2],[3]. When a "data fusion" center is costly or infeasible, one recourse is a decentralized approach where each agent broadcasts updates to its neighbors and agents collaboratively approach the global optimum [4],[5],[6],[7].

While there is great interest in advances to accelerate the performance of optimization algorithms, privacy preservation is also equally important to many machine learning tasks, especially in the processing of medical records and financial data. Techniques are required to quantify privacy loss during processing, and differential privacy (DP) is one of the best known rigorous theoretical mechanisms that serves this purpose. There is a large body of DP based Empirical Risk Minimization (ERM) work [8],[9],[10],[11], [12], [13],[14], [15] and DP-based  $k$ -means[16], Bayesian learning [17], identity testing [18] and deep learning [19], [20] works. Given a randomized algorithm, DP offers a provable guarantee against statistical inference that its output is insensitive to a slight change in the input dataset, for example, replacement of a single datapoint. Thus, from outputs observed, it is hard to distinguish the participation of an individual. The notion of DP was initially developed with a centralized view. Under central DP

or distributed learning with trusted central servers, using secure aggregator techniques [21], [22] or subsampling techniques [23] with secrecy of intermediate computation [24], many elegant mechanisms of privacy amplification have been proposed.

However, agents may not trust any other parties to collect their local data. To this end, a stronger notion is local differential privacy (LDP) [25], [26] in the distributed scenario, where each agent can run a randomization procedure locally and the privacy of an individual is still guaranteed even with a malicious collector. LDP has been adopted by Apple, Microsoft and Google as one formal definition of privacy [27], [28], [29]. Nonetheless, in contrast to the central model, LDP is far less studied despite successful deployment in industry [30]. Especially for private optimization, in the central model, the optimization protocol can be viewed as a black box since no information leakage can occur during the execution. Perturbation can be elegantly added only in the objective function, at the beginning, or the output, at the end. Well-known objective/output perturbation methods (e.g. [8]) and follow-up work [31], [32], [33] all capture this idea. For LDP, though one can still apply the above techniques in a non-interactive manner [34],[12],[35], high sample complexity to compensate for accuracy loss may be required.

In an interactive decentralized optimization, additional privacy loss arises from agents' cooperation. To investigate LDP in practice, we need concrete algorithms to support our analysis. In general, there are two types of decentralized optimization. One is (sub)gradient based, such as decentralized (stochastic) gradient descent (GD) methods [7], [6], [36], and EXTRA [37]. The second relies on solving a constrained problem with dual variables to minimize some Lagrangian function, such as Alternative Direction Method of Multipliers (ADMM) [38]. Though both proceed in an iterative manner, the computation of GD in each step can be less expensive compared to ADMM. Nevertheless, for general convex problems, the convergence rate of decentralized GD is  $O(1/\sqrt{K})$  and that of ADMM is  $O(1/K)$  [4], where  $K$  denotes the number of iterations. Under such a framework, agents enrolled in computing only need to share the states of optimization with neighbors. However, privacy loss also arises from such information exchange, since exposed intermediate results can be easily used to learn the sensitive

parameters of the local private functions. Incorporating cryptographic methods, such as (partial) homomorphic encryption [39], [40], [41], can come with high overhead especially in large-scale optimization. Alternatively, under the lens of DP, the most common approach is to apply perturbed local estimates during the update exchange in decentralized algorithms [42], [43], [44], [45], [46], [47], [48]. Although heuristic exploration, such as gradually increasing the step penalty [43], can improve the utility-privacy tradeoff, existing works lack insights into the fundamentals of algorithmic convergence with noise perturbation.

## 1.1 Problem Statement

Consider a decentralized optimization problem across  $N$  agents in a connected network. The network is modeled by an undirected graph  $\mathcal{G}(\mathcal{N}, \mathcal{E})$ . Nodes are indexed as  $\mathcal{N} = \{1, \dots, N\}$  and when two nodes  $i$  and  $j$  are neighbors that can communicate,  $(i, j) \in \mathcal{E}$ . In general, we assume each node holds a function  $f_i(\mathbf{x}_i)$  that we regard as a loss function determined by samples held locally with the parameter  $\mathbf{x}_i$  to be optimized. Throughout the rest of the paper, we always assume that  $f_i(\cdot)$  is a differentiable convex function  $C \rightarrow \mathbb{R}$  and  $\mathbf{x}_i \in C \subset \mathbb{R}^d$ .  $C$  can be viewed as the constraint, assumed to be a closed convex set. In general, we express the objective function to minimize as

$$\min_{\mathbf{x}_{[1:N]}} \sum_{i=1}^N f_i(\mathbf{x}_i), \quad s.t. \quad \sum_{i=1}^N A_i \mathbf{x}_i = \mathbf{c}, \quad (1.1)$$

under a linear constraint. In many learning problems,  $\mathbf{x}_{[1:N]}$  stand for one parameter to be collaboratively optimized, where  $[1 : N]$  is the compact form of  $\{1, 2, \dots, N\}$ . We term the problem as consensus optimization if the constraint requires that all  $\mathbf{x}_i$  be equal, of which the restrain can still be enforced as the linear constraint  $\sum_{i=1}^N A_i \mathbf{x}_i = \mathbf{0}$ , where  $A_i$  includes the information of graph connectivity [49].

## 1.2 Differential Privacy

For a randomized algorithm  $\mathcal{A}$  and a dataset  $\mathcal{D}$  as its input, we call  $\mathcal{D}'$  adjacent to  $\mathcal{D}$  if  $\mathcal{D}$  and  $\mathcal{D}'$  only differ in one data point. In the central model, quantitatively, we say  $\mathcal{A}$  achieves  $\epsilon$ -DP if for any adjacent  $\mathcal{D}$  and  $\mathcal{D}'$ , and any set  $S$  in the domain of  $\mathcal{A}(\cdot)$ ,

$$\Pr[\mathcal{A}(\mathcal{D}) \in S] \leq e^\epsilon \Pr[\mathcal{A}(\mathcal{D}') \in S] + \delta. \quad (1.2)$$

When  $\delta = 0$ , it is termed as pure DP, otherwise it is referred to relaxed DP, where one may apply a stronger composition theorem for accumulated privacy loss [50]. To embed the notion in the decentralized optimization setting, in this paper,  $\mathcal{A}$  corresponds to the optimization algorithm selected, while the functions  $f_{[1:N]}$  behave as the inputs and are the privacy concern. In the local version, each agent does not trust anyone and, in the worst case, all other parties are colluding against some agent  $i$  to learn some sensitive information of  $f_i$ . Following [51], [25], LDP in the context of decentralized optimization can be similarly defined as,

**Definition 1** ( $(\epsilon, \delta)$ -LDP with worst-case guarantee). *A decentralized optimization algorithm  $\mathcal{A}$  is  $(\epsilon, \delta)$ -LDP if for each  $i \in [1 : N]$  and any  $S$  in the output domain of  $\mathcal{A}$ ,*

$$\Pr(\mathcal{A}(\mathcal{D}_i) \in S) \leq e^\epsilon \Pr(\mathcal{A}(\mathcal{D}'_i) \in S) + \delta \quad (1.3)$$

where  $\mathcal{D}_i = (f_1, \dots, \hat{f}_i, \dots, f_N)$  and  $\mathcal{D}'_i = (f_1, \dots, \tilde{f}_i, \dots, f_N)$  only differ in the  $i^{\text{th}}$  entry and  $\hat{f}_i$  and  $\tilde{f}_i$  are arbitrary two objective functions, determined by two possible adjacent datasets of agent  $i$ , respectively.

## 1.3 Motivation

The worst-case nature of  $(\epsilon, \delta)$  (L)DP in formulation (1.2) and (1.4) does not capture the expected privacy guarantee over the distribution of outputs. Accordingly, we develop a output-specific LDP notion, which allows us to more precisely quantify the statistical price of privacy.



**Definition 2** ( $(\epsilon, \delta, \gamma)$ -LDP with worst & average-case guarantee). A decentralized optimization algorithm  $\mathcal{A}$  is  $(\epsilon, \delta, \gamma)$ -LDP if for each  $i \in [1 : N]$  and any output  $\chi$  of  $\mathcal{A}$ ,

$$\Pr(\mathcal{A}(\mathcal{D}_i) = \chi) \leq e^{\epsilon(\chi)} \Pr(\mathcal{A}(\mathcal{D}'_i) = \chi) + \delta(\chi), \quad (1.4)$$

such that  $\sup_{\chi} \epsilon(\chi) \leq \epsilon$  and  $\sup_{\chi} \delta(\chi) \leq \delta$ , while  $\mathbb{E}_{\chi} \epsilon(\chi) \leq \gamma\epsilon$  and  $\mathbb{E}_{\chi} \delta(\chi) \leq \gamma\delta$  for any  $\mathcal{D}_i$  and  $\mathcal{D}'_i$ . The distribution of  $\chi$  depends on the true dataset used in  $\mathcal{A}$ .

The above definition is more comprehensive than (1.4), where the additional parameter  $\gamma$  indeed captures the gap between the worst and average privacy loss. On the other side, to randomize an algorithm, the most common technique in DP is perturbation, such as the well-known Laplace/Gaussian Mechanism, where an independent noise is added. Therefore, it is of sufficient interests to understand the utility-privacy tradeoff and, even more, whether privacy can come for free without utility compromise. Unfortunately, such independent noise will bring utility compromise in general. Lower bounds of utility loss in ERM has been provided in [10], [25], [34], which shows a negative answer that given worst-case (L)DP guarantee, privacy always comes with a non negligible cost. However, it remains open whether there exists certain algorithmically-dependent randomization for privacy amplification in a non-asymptotic view or with respect to (w.r.t.) average loss.

## 1.4 Methodology Overview

We first review the basic updating subroutine in decentralized gradient descent. Let  $\mathbf{x}_i^k$  denote the local estimate of the global optimum for agent  $i$  at round  $k$ . In consensus optimization, a naively noisy form of decentralized GD [7], [6] can be described as,

$$\boxed{\mathbf{x}_i^{k+1} := \sum_{j=1}^N \underbrace{w_{ij}}_A \mathbf{x}_j^k - \underbrace{\eta_{k+1}}_B \nabla f_i(\mathbf{x}_i^k) + \Delta_i^{k+1}} \quad (1.5)$$

where  $w_{ij} \in [0, 1]$  is the weight assigned to  $\mathbf{x}_j^k$  such that  $\sum_{j=1}^N w_{ij} = 1$  and  $\eta_{k+1}$  is the step size at round  $(k + 1)$ .  $\Delta_i^{k+1}$  denotes the noise added, which is assumed to be

Laplace noise. Therefore, Part A in (1.5) is a weighted average of updates collected from the last iteration and Part B corresponds to the gradient step taken. In the typical case (1.5), gradient descent is randomized by resorting to perturbation  $\Delta_i^{k+1}$ . One can apply a stochastic gradient instead, but the subsampling does not amplify privacy [10]. However, we show that carefully designed randomization can reshape the output distribution to produce a sharpened average privacy loss. In Section 2, it is noted that the optimization protocol may take dozens of steps and divergence among  $\mathbf{x}_i^k$ ,  $i \in [1 : N]$ , always exists. Now, for instance, imagine if we use random weights  $\bar{w}_{ij}$  instead in (1.5), which are independently selected across each iteration: Part A becomes a random variable within some neighborhood of  $\mathbf{x}_i^k$ , while Part B remains the same. In the  $(k + 1)$ th iteration, conditional on  $\mathbf{x}_i^k$ ,  $i \in [1 : N]$ , and  $f_i$ ,  $\mathbf{x}_i^{k+1}$  then follows a mixture Laplace distribution with a random mean. Similarly, recalling the updating rule of ADMM with dual method for (1.1), each update relies on solving an optimization problem:

$$\mathbf{x}_i^{k+1} := \arg \min_{\mathbf{x}_i} \mathcal{L}(\mathbf{x}_1^k, \dots, \mathbf{x}_i, \dots, \mathbf{x}_N^k, \boldsymbol{\lambda}^k) + \frac{\rho}{2} \left\| A_i \mathbf{x}_i + \sum_{j \neq i}^N A_j \mathbf{x}_j^k - \mathbf{c} \right\|^2 + \frac{\Gamma}{2} \|\mathbf{x}_i - \mathbf{x}_i^k\|^2 + \Delta_i^{k+1} \quad (1.6)$$

where the Lagrangian function is defined as  $\mathcal{L}(\mathbf{x}_1, \dots, \mathbf{x}_N, \boldsymbol{\lambda}) = \sum_{i=1}^N f_i(\mathbf{x}_i) - \boldsymbol{\lambda}^T (\sum_{i=1}^N A_i \mathbf{x}_i - \mathbf{c})$ . The Lagrangian multiplier  $\boldsymbol{\lambda}^{k+1}$  is updated through  $\boldsymbol{\lambda}^{k+1} := \boldsymbol{\lambda}^k - \zeta (\sum_{i=1}^N A_i \mathbf{x}_i^{k+1} - \mathbf{c})$ . If we allow each agent to independently select random penalties  $\rho$  and  $\Gamma$  across iterations, clearly, a similar mixture Laplace with random mean,  $\mathbf{x}_i^{k+1}$  is produced. While a random mean amplifies the uncertainty, trivially incorporating such an idea in existing algorithms may easily lead to unclear convergence guarantees or cumbersome privacy analysis.

## 1.5 Contribution and Organization

In this paper, we study the LDP of interactive decentralized optimization from both asymptotic and non-asymptotic perspectives.

- (i) We first follow an active line of research on LDP amplification via cryptography. Very recently, via shuffling for anonymity, Erlingsson et al. propose a framework to narrow the gap between central DP and LDP [52], which is further generalized by Cheu et.al [53]. For first-order based decentralized optimizations which satisfy  $(\epsilon, \delta)$ -LDP, in Theorem 2.1.1, we present a framework with merely a linear secret sharing scheme which can satisfy  $(\epsilon/\sqrt{N'}, \delta)$  central DP, where  $N'$  is the number of semi-honest agents amongst all  $N$  nodes in the network.
- (ii) Taking ADMM and D(S)GD as two examples, we propose a framework of decentralized optimization with varying parameters. A unified local privacy analysis is presented in Theorem 2.2.1 and we quantify the privacy amplification constant  $\gamma$  in Theorem 2.2.2. Rigorous analysis on the upper bound of utility loss for proposed algorithms are included in Theorems 3.1.1, 3.1.2, 3.1.3, 3.2.1. Following that, we provide more refined analysis to explain why even with further randomization, the proposed algorithms are with almost the same convergence rate compared to the ones with fixed-parameters in practice. Experiments support the theory. Further analysis on the strongly convex case is shown in Appendix J.



# Chapter 2

## Algorithm Description and Privacy Analysis

### 2.1 Secret-sharing Based Privacy Amplification

Generally speaking, updates from neighbors are aggregated in some linearly weighted form (e.g. Part A in (1.5)) in first-order based decentralized optimization, while the exchange procedure incurs the privacy loss before each aggregation. In the following, we show once the aggregation can be securely computed from the multiple parties, the differential privacy will also be benefited accordingly. Due to the nice structure of the linear form, we consider secret sharing. We first give the details of the algorithm construction, where we take the private DGD as an example. Without loss of generality, we assume the graph is fully connected temporarily to ease the algorithm description in this section.

Here  $\|\cdot\|_q$  denotes the  $l_q$  norm and  $\|\cdot\|$  denotes the standard  $l_2$  norm for brevity throughout the rest of this thesis. The above scheme relies on the basic fact that  $\sum_{i=1}^N \hat{s}_i^k \equiv \sum_{i=1}^N \sum_{j=1}^N s_{ji}^k \equiv \sum_{i=1}^N \sum_{j=1}^N s_{ij}^k \equiv \sum_{i=1}^N \mathbf{x}_i^k \pmod{p}$ . Besides, in ADMM protocol, it is noted that the aggregation is in the form  $\sum_{i=1}^N A_i \mathbf{x}_i^k$ . One may replace step 1 with ADMM updating procedure (1.6) and implement secret sharing on  $A_i \mathbf{x}_i^k$  instead. A similar private ADMM protocol can be derived as well.

---

Secret-sharing based Private Decentralized Gradient Descent (DGD)

---

**Input:**  $f_{[1:N]}$ ,  $\mathbf{x}_{[1:N]}^0$  and  $p \in \mathbb{Z}$  that  $p > \|\sum_{i=1}^N \mathbf{x}^k\|_\infty$  with overwhelming probability.

**for**  $k = 1, 2, \dots, K$  **do**

**Agents**  $i = 1$  **to**  $N$  **do** in parallel:

    1. Agent  $v_i$  applies an  $(\epsilon, \delta)$ -locally private version of DGD updating rule:  $\mathbf{x}_i^k := \frac{\sum_{j=1}^N \mathbf{x}_j^{k-1}}{N} - \eta_k \nabla f_i(\mathbf{x}_i^{k-1}) + \Delta_i^k$ , where  $\Delta_i^k$  is a Gaussian  $\mathcal{N}(\mathbf{0}, \sigma^2)$ .

    2. Randomly split  $\mathbf{x}_i^k$  into  $N$  shares,  $\mathbf{s}_{[1:N]}^k$ , such that  $\mathbf{x}_i^k = \sum_{j=1}^N \mathbf{s}_{ij}^k \pmod{p}$ .

    3. Agent  $v_i$  sends  $\mathbf{s}_{ij}^k$  to  $v_j$  while keeping  $\mathbf{s}_{ii}^k$  as a secret, left to itself.

    4. Each agent sums up  $\mathbf{s}_{[1:N]}^k$  received as  $\hat{\mathbf{s}}_i^k = \sum_{j=1}^N \mathbf{s}_{ji}^k \pmod{p}$ .

    5.  $v_i$  broadcasts  $\hat{\mathbf{s}}_i^k$

    6. Each agent reconstructs  $\sum_{i=1}^N \mathbf{x}_i^k = \sum_{i=1}^N \hat{\mathbf{s}}_i^k \pmod{p}$

**end for**

---

**Theorem 2.1.1.** *When  $N \geq 3$ , if each node can have secure communication with the rest nodes, the above linear secret sharing based DGD satisfies  $(\epsilon/\sqrt{N'}, \delta)$  central DP where there exist  $N'$  semi-honest nodes across the network.*

**Remark 2.1.1.** *In the federated learning setting, where there may exist several, say  $g$ , curators to collect and aggregate the updates, a similar secret sharing can be implemented as: the update  $\mathbf{x}_i^k$  of each agent is still randomly split into  $g$  shares sent to the  $g$  curators, respectively. When at least one curator is semi-honest, Theorem 2.1.1 holds as well.*

Theorem 2.1.1 implies that through the cooperation, the inputs of semi-honest agents can be securely merged and the privacy is amplified in a central view of the input union. However, it is worth noting that when all other nodes are colluding against some node  $i$ , i.e.,  $N' = 1$ , Theorem 2.1.1 does not provide privacy amplification. In the following, we resort to algorithm-dependent randomization to further improve LDP even under the worst assumptions on nodes across the network.

## 2.2 Decentralized Optimization with Randomized Local Aggregations

We first present the main protocol of modified private ADMM and Decentralized SGD (DSGD) as Algorithm 1 and 2, respectively. Here, we omit the projection step if  $\mathcal{C} \subsetneq \mathbb{R}^d$  since we are interested in arbitrary constraints. Still for simplicity, in this section, we assume  $\mathcal{G}$  is fully connected and only consider the consensus problem temporarily, while we provide a convergence proof for general cases in the next section.

In contrast to previous private ADMM protocols [38], [42], [43], we consider applying a first-order approximation for each  $f_i$ :

$$f_i(\mathbf{x}_i) \approx f_i(\mathbf{x}_i^k) + \nabla f_i(\mathbf{x}_i^k)(\mathbf{x}_i - \mathbf{x}_i^k), \quad (2.1)$$

and we derive a modified private ADMM accordingly as follows. Compared to

---

### Algorithm 1 Modified Private ADMM with First-order Approximation

---

**Input:** Local functions  $f_{[1:N]}$ , step penalty  $\zeta$ .

Initialize  $\mathbf{x}_{[1:N]}^0$  randomly,  $\boldsymbol{\lambda}_{[1:N]}^0 = \mathbf{0}$ . Each agent selects a private constant  $D_i$ .

**for**  $k = 0, 1, 2, \dots, K - 1$  **do**

**Agents**  $i = 1$  **to**  $N$  **do in parallel:**

        Randomly pick two positive diagonal matrices  $\bar{\boldsymbol{\rho}}_i^{k+1}$  and  $\boldsymbol{\Gamma}_i^{k+1}$  such that  $(N - 1)\bar{\boldsymbol{\rho}}_i^{k+1} + \boldsymbol{\Gamma}_i^{k+1} = D_i \cdot \mathbf{I}_p$  and update  $\mathbf{x}_i^{k+1}$  *in parallel:*

$$\mathbf{x}_i^{k+1} := \frac{\boldsymbol{\Gamma}_i^{k+1}}{D_i} \mathbf{x}_i^k + \frac{(N-1)\bar{\boldsymbol{\rho}}_i^{k+1}}{D_i} \frac{\sum_{j \neq i} \mathbf{x}_j^k}{N-1} - \frac{D_i^{-1} \nabla f_i(\mathbf{x}_i^k)}{D_i} + D_i^{-1} \boldsymbol{\lambda}_i^k + \Delta_i^{k+1}. \quad (2.2)$$

        Exchange  $\mathbf{x}_i^{k+1}$  and then update  $\boldsymbol{\lambda}_i^{k+1} := \boldsymbol{\lambda}_i^k - \zeta \sum_{j \neq i} (\mathbf{x}_j^{k+1} - \mathbf{x}_i^{k+1})$ .

**end for**

---

conventional ADMM with updating rule (1.6), Algorithm 1 has reduced complexity, same as that of DGD, which avoids finding the solution of (1.6)<sup>1</sup>. It is clear that Algorithm 1 and 2 share a very similar structure except for the dual variable  $\boldsymbol{\lambda}_i^k$  in ADMM. Intuitively, Term (A) in either (2.2) or (2.3) behaves as a random aggregator

---

<sup>1</sup>Conventional ADMM may encounter considerable computation overhead across each iteration since in general no closed-form optima of (1.6) exists

to merge the updates from the previous iteration and Part **B** corresponds to the effect from the function  $f_i$  on updating  $\mathbf{x}_i^{k+1}$ . Here, we only give two examples to select the random weights where Term (A) is uniformly distributed in some interval. Many variants can be easily derived, and more details can be found in Appendix E.

---

**Algorithm 2** Modified Private Decentralized Stochastic Gradient Descent

---

**Input:** Local functions  $f_i$  and a diminishing sequence  $\{\eta_k\}$

Randomly divide  $N$  agents into  $2K$  groups,  $S_{[1:2K]}$ .

Initialization  $\mathbf{y}_1^0$  and  $\mathbf{y}_2^0$ .

**for**  $k = 0, 1, 2, \dots, K - 1$  **do**

**Agents**  $i$  in  $S_{2k+1}$  **and**  $S_{2k+2}$  **do in parallel :**

        Randomly pick a positive diagonal matrix  $\mathbf{w}_i$  of which the non-zero elements are within  $(0, 1)$ . Then, update  $\mathbf{x}_i$  as:

$$\mathbf{x}_i := \underbrace{\mathbf{w}_i \mathbf{y}_1^k + (\mathbf{I}_d - \mathbf{w}_i) \mathbf{y}_2^k}_A - \underbrace{\eta_{k+1} N \nabla f_i \left( \frac{\mathbf{y}_1^k + \mathbf{y}_2^k}{2} \right)}_B + \Delta_i, \quad (2.3)$$

        Broadcast  $\mathbf{x}_i$  to agents in  $S_{2k+3}$  and  $S_{2k+4}$  where  $\mathbf{y}_1^{k+1} = \frac{1}{|S_{2k+1}|} \sum_{i \in S_{2k+1}} \mathbf{x}_i$  and

$\mathbf{y}_2^{k+1} = \frac{1}{|S_{2k+2}|} \sum_{i \in S_{2k+2}} \mathbf{x}_i$ .

**end for**

---

To quantify the privacy loss, we need to introduce a concept, *sensitivity*, to further specify the noise. In proposed algorithms, the gradients of  $f_i$  are our privacy concern. We say our protocols satisfy  $\mathcal{B}_q$  sensitivity if for any two objective functions  $\hat{f}_i$  and  $\tilde{f}_i$  determined by two neighboring datasets of any agent  $i$ ,  $\sup_{\mathbf{x} \in \mathcal{C}} \left\| \nabla \hat{f}_i(\mathbf{x}) - \nabla \tilde{f}_i(\mathbf{x}) \right\|_q \leq \mathcal{B}_q$ . In the case of ERM,  $\mathcal{B}_\infty = O\left(\frac{1}{\min_i b_i}\right)$  where  $b_i$  denotes the size of dataset held by agent  $i$  and each data point is bounded in  $l_\infty$ . In the following, without loss of generality, we restrict our focus to Algorithm 1 under  $\mathcal{B}_\infty$  sensitivity. One can easily generalize the following analysis to Algorithm 2 and sensitivity  $\mathcal{B}_q$  in other norms. The following lemma, whose proof is in Appendix B, provides a semi closed-form of  $\epsilon(\chi)$ .

**Lemma 2.2.1.** *Given  $\chi = \mathbf{x}_{[1:N]}^{[0:K]}$ , Algorithm 1 satisfies  $\epsilon(\chi)$ -LDP, where*

$$\epsilon(\chi) = \sup_{\hat{f}_i, \tilde{f}_i} \left| \sum_{k=0}^{K-1} \sum_{l=1}^d \log \frac{P(\mathbf{x}_i^{k+1}[l] | \tilde{f}_i, \mathbf{x}_{[1:N]}^{[0:k]})}{P(\mathbf{x}_i^{k+1}[l] | \hat{f}_i, \mathbf{x}_{[1:N]}^{[0:k]})} \right|, \quad (2.4)$$

where  $\mathbf{x}_i^k[l]$  denotes the  $l^{\text{th}}$  coordinate of  $\mathbf{x}_i^k \in \mathbb{R}^d$ .



Following the Laplace mechanism [54],[55], we assume each coordinate of  $\Delta_i^k$  i.i.d. following a Laplace distribution  $\text{Lap}(0, \beta_k)$  with probability density  $P(y) = \frac{\beta_k}{2} e^{-\beta_k |y|}$ . Thus, either in (2.3) or (2.2),  $P(\mathbf{x}_i^{k+1}[l] | f_i, \mathbf{x}_{[1:N]}^{[0:k]})$  follows the same Laplace distribution as that of  $\Delta_i^{k+1}[l]$  except the mean of  $\mathbf{x}_i^{k+1}[l]$  is a random variable uniformly distributed in some interval, denoted by  $\tau_i^{k+1}[l]$  (see Term (A) in (2.2)). To bound the accumulated privacy loss, we consider the composition of the loss on each coordinate  $l$  and iteration  $k$ , denoted by  $\epsilon_i^{k+1}(\mathcal{X}) = \sup_{f_i, \hat{f}_i} \left| \log \left[ P(\mathbf{x}_i^{k+1}[l] | f_i, \mathbf{x}_{[1:N]}^{[0:k]}) / P(\mathbf{x}_i^{k+1}[l] | \hat{f}_i, \mathbf{x}_{[1:N]}^{[0:k]}) \right] \right|$ . The following theorem provides an upper bound of  $\epsilon_i^{k+1}(\mathcal{X})$ .

**Theorem 2.2.1.** *Algorithms 1 achieves  $\epsilon(\mathcal{X}) \leq \sum_{k=0}^{K-1} \sum_{l=1}^d \epsilon_l^{k+1}(\mathcal{X})$ -LDP, <sup>2</sup> where*

$$\epsilon_l^k(\mathcal{X}) \leq \max_{|t| \leq \frac{1}{D_i} \mathcal{B}_\infty} \left| \log \left[ \int_{\tau_i^k[l]} e^{-\beta_k |x_i^k[l]-x|} dx \right] - \log \left[ \int_{\tau_i^k[l]+t} e^{-\beta_k |x_i^k[l]-x|} dx \right] \right|, \quad (2.5)$$

where  $\tau_i^k[l] + t$  implies uniformly moving the interval with  $t$ . Moreover, for arbitrary  $\mathcal{X}$ , the right hand of (2.5) is never bigger than  $\frac{1}{D_i} \beta_k \mathcal{B}_\infty$ . Specifically, when  $\mathbf{x}_i^k[l]$  belongs to  $\tau_i^k[l]$ , it is strictly smaller than  $\frac{1}{D_i} \beta_k \mathcal{B}_\infty$ .

Theorem 2.2.1 indicates that by directly applying Laplace/Gaussian on ADMM or DGD, the worst privacy loss equals to the average one, i.e.,  $\gamma = 1$  in the modified LDP (Definition 2). However, with random aggregation incorporated, there is always a chance that one may achieve a strictly better average privacy loss while the same worst-case guarantee still holds compared with the fixed-parameter case.

We include the proof of the above theorem in Appendix C. As a straightforward corollary of the above theorem, if we fix parameters in Algorithms 1, then the local privacy loss is  $\sum_{k=0}^{K-1} d \frac{1}{D_i} \beta_{k+1} \mathcal{B}_\infty$ , which matches prior results [43], [42], [48]. To conclude, for each  $\epsilon_i^k(\mathcal{X})$ , randomized weights renders a constant privacy reduction expressed as a conditional expectation,  $\gamma_l^k = \mathbb{E}_{\mathbf{x}_i^k} [\epsilon_i^k(\mathcal{X}) / (\frac{1}{D_i} \beta_k \mathcal{B}_\infty) | \mathbf{x}_{[1:N]}^{[0:k-1]}]$ . From (2.5), a longer length of the interval  $\tau_i^k[l]$ , denoted by  $\omega$ , renders a more concentrated mixture distribution compared to pure Laplace. In the following theorem, whose proof

<sup>2</sup>One can obtain a stronger composition form in relaxed  $(\epsilon, \delta)$ -LDP [50] that Algorithms 1 and 2 achieve  $(\sum_{k=0}^K \sum_{l=1}^d \frac{(e^{\epsilon_l^{k+1}(\mathcal{X})} - 1) \epsilon_l^{k+1}(\mathcal{X})}{e^{\epsilon_l^{k+1}(\mathcal{X})} + 1} + \sqrt{-2 \sum_{k=0}^K \sum_{l=1}^d (\epsilon_l^{k+1}(\mathcal{X}))^2 \log(\delta)}) \log(\delta), \delta)$ -LDP, for some  $\delta \in (0, 1)$ .

is in Appendix D, we quantify  $\gamma_l^k$  when  $\omega > \frac{1}{D_i} \mathcal{B}_\infty$ .

**Theorem 2.2.2.** When  $\omega > \frac{1}{D_i} \mathcal{B}_\infty$ ,

$$\gamma_l^k \leq \frac{1}{\frac{1}{D_i} \beta_k \mathcal{B}_\infty} \log \left\{ e^{\frac{1}{D_i} \beta_k \mathcal{B}_\infty} \left[ 1 - 2 \int_0^{\frac{\omega - \frac{1}{D_i} \mathcal{B}_\infty}{2}} \int_0^\omega \Phi(x, y) dy dx \right] + 2 \int_0^{\frac{\omega - \frac{1}{D_i} \mathcal{B}_\infty}{2}} \int_{-\frac{1}{D_i} \mathcal{B}_\infty}^{\omega - \frac{1}{D_i} \mathcal{B}_\infty} \Phi(x, y) dy dx \right\}, \quad (2.6)$$

where  $\Phi(x, y) = \frac{\beta_k}{2\omega} e^{-\beta_k |x-y|}$ .

In Fig. 1 of 2-1, we show the relationship between  $\gamma$ ,  $\omega$  and  $\beta$ , where temporarily the dependence on  $l$  and  $k$  is dropped for brevity and  $\frac{\mathcal{B}_\infty}{D_i}$  is fixed to 0.001. With  $\omega$  ranging from 0.1 to 1 and  $\beta$  from 2 to 10, clearly larger  $\omega$  and  $\beta$ , corresponding to a longer interval length and noise of smaller variance, lead to better privacy amplification.

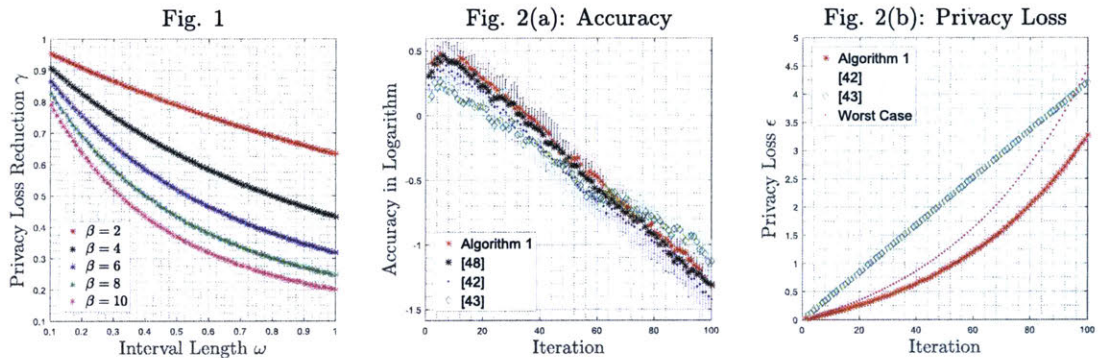


Figure 2-1: Expected Average Privacy Amplification and Performance Comparison with Existing Works

As a comparison, we also test the proposed schemes and state-of-the-art approaches on a regularized logistic regression of the *Adult* dataset, from the UCI machine learning repository [43]. The performance of existing private ADMM [43] (with increasing penalty), [42], [48] and Algorithm 1 is illustrated in Fig. 2 of 2-1, where the communication graph  $\mathcal{G}$  is randomly generated with  $N = 10$  and  $E = |\mathcal{E}| = 20$ . Fig. 2 (a) of 2-1 shows the accuracy in a logarithmic scale, where even with the

first-order approximation, Algorithm 1 has almost the same performance as prior works. Associated privacy loss of Algorithm 1 is presented in Fig. 2 (b) of 2-1. The worst case in Fig. 2 (b) of 2-1 refers to  $\sum_{k=0}^{K-1} d \frac{1}{D_i} \beta_{k+1} \mathcal{B}_\infty$ . On average, we achieve 30% privacy loss reduction. We omit the privacy loss of [48] in Fig. 2 (b) of since it is too loose in this example. We note that the plots of [43], [42] in Fig. 2(b) (and [48]) either require global sensitivity or smoothness of gradients, while we only assume local sensitivity. Full description and results of experiments are included in Appendix E. In Appendix E (Fig. E-3 and E-4), we also show that the parameter randomization in Algorithms 1 and 2 does not bring accuracy compromise, which is guaranteed by the convergence theorems presented in next section.



# Chapter 3

## Convergence and Utility Tradeoff

### Analysis

Throughout this section, we will provide convergence proofs of Algorithm 1 & 2 proposed. *It is worth noting that the following theorems and framework of proofs are invariant to whether a randomized local aggregation is incorporated or not when parameters are in some proper admissible range. The conclusions shown also match the theoretical lower bound of utility loss in an asymptotic view.* With those upper bounds of convergence rate at hand, we will provide stronger evidence to show why the further randomized aggregation does not compromise the performance in practice. To deal with those asymptotic utility bound, we stick to the conventional LDP setting (Definition 1) w.r.t. the worst case.

For simplicity, let  $\mathbf{X}^k = (\mathbf{x}_1^k, \mathbf{x}_2^k, \dots, \mathbf{x}_N^k)$  and  $F(\mathbf{X}^k) = \sum_{i=1}^N f_i(\mathbf{x}_i^k)$ , where accordingly  $\nabla F(\mathbf{X}^k) = (\nabla f_1(\mathbf{x}_1^k), \dots, \nabla f_N(\mathbf{x}_N^k))$ . We first recall some commonly used notions in convex optimization analysis.

- A function  $f(\mathbf{x}) : C \rightarrow \mathbb{R}$  is  $L$ -Lipschitz continuous if for any  $\mathbf{x}, \mathbf{y} \in C$ ,  $|f(\mathbf{x}) - f(\mathbf{y})| \leq L \|\mathbf{x} - \mathbf{y}\|$ .
- A function  $f(\mathbf{x}) : C \rightarrow \mathbb{R}$  is  $M$ -smooth if  $\nabla f(\cdot)$  is  $M$ -Lipschitz continuous: for any  $\mathbf{x}, \mathbf{y} \in C$ ,

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq \sqrt{M} \|\mathbf{x} - \mathbf{y}\|.$$

### 3.1 Utility Analysis of Algorithm 1

We describe the construction of Algorithm 1 in two steps to solve (1.1). First we recall the case of conventional ADMM (1.6) without first-order approximation but incorporating random penalties across iterations. The updating procedure of node  $i$  at the  $(k + 1)$ th iteration becomes

$$\begin{cases} \tilde{\mathbf{x}}_i^{k+1} := \arg \min_{\mathbf{x}_i} f_i(\mathbf{x}_i) - \boldsymbol{\lambda}^{kT} \left( A_i \mathbf{x}_i + \sum_{j \neq i} A_j \mathbf{x}_j^k - \mathbf{c} \right) \\ \quad + \frac{1}{2} \left\| A_i \mathbf{x}_i + \sum_{j \neq i} A_j \mathbf{x}_j^k - \mathbf{c} \right\|_{\boldsymbol{\rho}_i^{k+1}}^2 + \frac{1}{2} \left\| \mathbf{x}_i - \mathbf{x}_i^k \right\|_{\Gamma_i^{k+1}}^2 \\ \mathbf{x}_i^{k+1} = \tilde{\mathbf{x}}_i^{k+1} + \Delta_i^{k+1}. \end{cases} \quad (3.1)$$

and the Lagrangian multiplier is updated accordingly as  $\tilde{\boldsymbol{\lambda}}^{k+1} := \boldsymbol{\lambda}^k - \boldsymbol{\gamma}_i^{k+1} \boldsymbol{\rho}_i^{k+1} (\sum_{i=1}^N A_i \mathbf{x}_i^{k+1} - \mathbf{c})$ ; and  $\boldsymbol{\lambda}^{k+1} := \boldsymbol{\lambda}^k - \boldsymbol{\gamma}_i^{k+1} \boldsymbol{\rho}_i^{k+1} (\sum_{i=1}^N A_i \tilde{\mathbf{x}}_i^{k+1} - \mathbf{c})$ .  $\boldsymbol{\gamma}_i^{k+1} \boldsymbol{\rho}_i^{k+1} = \zeta \cdot \mathbf{I}$  is a global constant set up at the beginning.  $\|z\|_G^2 = z^T G z$ . Let  $\mathbf{u}^{k+1} = [\mathbf{x}_{[1:N]}^{k+1}, \boldsymbol{\lambda}^{k+1}]$  and  $\mathbf{u}^* = [\mathbf{x}_{[1:N]}^*, \boldsymbol{\lambda}]$ , where  $\mathbf{x}_{[1:N]}^*$  stand for the optimum to (1.1) and  $\boldsymbol{\lambda}$  is an arbitrary point in  $\mathbb{R}^d$ .

**Theorem 3.1.1.** *If  $f_{[1:N]}$  are all  $M$ -smooth convex functions, following (J.1), we have*

$$\begin{aligned} & \mathbb{E}[F(\mathbf{X}^{k+1}) - F(\mathbf{X}^*) - \boldsymbol{\lambda}^T A(\mathbf{X}^{k+1} - \mathbf{X}^*)] \\ & \leq \frac{1}{2} \left( \sum_{i=1}^N (\sqrt{M} + D_i + \frac{1}{\zeta}) \mathbb{E}[\|\Delta_i^{k+1}\|^2] + \mathbb{E}[\|\mathbf{u}^k - \mathbf{u}^*\|_G^2 - \|\mathbf{u}^{k+1} - \mathbf{u}^*\|_G^2 - h^{k+1}] \right), \end{aligned} \quad (3.2)$$

where  $G = \text{diag}\{\mathbf{D}_1, \dots, \mathbf{D}_N, \frac{1}{\zeta}\}$ .  $\text{diag}\{\mathbf{D}_1, \dots, \mathbf{D}_N, \frac{1}{\zeta}\}$  denotes a diagonal matrix with elements  $\mathbf{D}_1, \dots, \mathbf{D}_N, \frac{1}{\zeta}$ . Here  $\mathbf{D}_i = D_i \cdot \mathbf{I} = \Gamma_i^{k+1} + A_i^T \boldsymbol{\rho}_i^{k+1} A_i$  is positive definite.  $h^{k+1}$  is some remainder term which is non-negative if for any  $i, j \in [1 : N]$

$$\begin{cases} \frac{D_j}{N^2} (\underline{\rho}_i - \frac{\zeta}{2}) \geq \sigma_{\max, j}^2 \bar{\rho}_i^2 \\ \frac{1}{(N-1)^2} (\underline{\rho}_i - \frac{\zeta}{2})(\underline{\rho}_j - \frac{\zeta}{2}) \geq (\bar{\rho}_i - \underline{\rho}_i)^2 \end{cases} \quad (3.3)$$

where  $\underline{\rho}_i \cdot \mathbf{I} \leq \boldsymbol{\rho}_i^{k+1} \leq \bar{\rho}_i \cdot \mathbf{I}$  for some constants  $0 < \underline{\rho}_i < \bar{\rho}_i$  and  $\sigma_{\max, i}$  is the largest singular value of  $A_i$ .

The proof of Theorem 3.1.1 can be found in Appendix F. With Theorem 3.1.1, we can show the convergence and utility loss of Algorithm 1 w.r.t. the norm  $\|\cdot\|_G$ .

**Theorem 3.1.2.** *Under the same condition in Theorem 3.1.1, let  $\bar{\mathbf{X}}^K = \frac{1}{K} \sum_{k=1}^K \mathbf{X}^k$ ,*

$$\mathbb{E}[F(\bar{\mathbf{X}}^K)] - F(\mathbf{X}^*) \leq \frac{\|\mathbf{X}^0 - \mathbf{X}^*\|_{G_x}^2 + \frac{4}{\zeta} \|\boldsymbol{\lambda}^*\|^2}{2K} + \frac{\sum_{k=1}^K \sum_{i=1}^N (\sqrt{M} + D_i + \frac{1}{\zeta}) \mathbb{E}[\|\boldsymbol{\Delta}_i^k\|^2]}{K}, \quad (3.4)$$

where  $\boldsymbol{\lambda}^*$  is state of  $\boldsymbol{\lambda}^k$  when  $\mathbf{X}^k$  reaches the optimum. Given  $(\epsilon, \delta)$  budget for LDP, the utility loss is  $O(\frac{\sqrt{Nd} \mathcal{B}_\infty \|\mathbf{X}^0 - \mathbf{X}^*\|_{G_x}}{\epsilon})$ . Here  $G_x = \text{diag}\{\mathbf{D}_1, \dots, \mathbf{D}_N\}$ .

The proof of Theorem 3.1.2 can be found in Appendix G. At last, the following results bridge the above utility analysis and the case where we further apply first-order approximation in (J.1) as

$$\mathbf{x}_i^{k+1} := \mathbf{D}_i^{-1} [A_i^T (\boldsymbol{\lambda}^k - \boldsymbol{\rho}_i^{k+1} (\sum_{j \neq i} A_j \mathbf{x}_j^k - \mathbf{c})) + \boldsymbol{\Gamma}_i^{k+1} \mathbf{x}_i^k - \nabla f_i(\mathbf{x}_i^k)] + \boldsymbol{\Delta}_i^{k+1}. \quad (3.5)$$

To quantify the loss from the approximation, we provide the following theorem.

**Theorem 3.1.3.** *Under modified updating procedure (J.17), Theorem 3.1.1 holds with the same setup and  $h^{k+1}$  is non-negative if for any  $i, j \in [1 : N]$ :*

$$\begin{cases} \frac{D_j - M}{N^2} (\underline{\rho}_i - \frac{\zeta}{2}) \geq \sigma_{\max, j}^2 \bar{\rho}_i^2 \\ \frac{1}{(N-1)^2} (\underline{\rho}_i - \frac{\zeta}{2}) (\underline{\rho}_j - \frac{\zeta}{2}) \geq (\bar{\rho}_i - \underline{\rho}_i)^2 \end{cases} \quad (3.6)$$

The proof of Theorem 3.1.3 can be found in Appendix H. In addition, if  $f_{1:N}$  are further assumed to be strongly convex, without perturbation, ADMM can achieve a linear convergence rate. As for private ADMM, indeed, it can be proved that the strongly-convex assumptions will not improve the utility bound obtained in Theorem 3.1.2 asymptotically whereas the number of iterations required is in a logarithmic scale. The details can be found in Appendix J.

## 3.2 Utility Analysis of Algorithm 2

In comparison to ADMM, D(S)GD only captures consensus optimization. When  $\|\nabla f_i\|$  is bounded, the following theorem shows the privacy-utility tradeoff of Algorithm 2. Here we assume  $\mathbf{x}^*$  is the optimum to (1) in the consensus case.

**Theorem 3.2.1.** *Assume that  $f_i(\mathbf{x})$  is convex and  $\|\nabla f_i(\mathbf{x})\|^2$  is bounded by  $G^2$  for each  $i$ . Moreover, let  $\max_{k,j \in \{1,2\}} \mathbb{E}[(\bar{\Delta}_j^k / \eta_k)^2] \leq V^2$ . When we select the step size  $\eta_k = \frac{1}{c\sqrt{k}}$  for some constant  $c$ , then*

$$\begin{aligned} \frac{1}{N} \mathbb{E}[F(\frac{\sum_{k=0}^{K-1} (\mathbf{y}_1^k + \mathbf{y}_2^k)}{2K}) - F(\mathbf{x}^*)] \\ \leq \frac{c\sqrt{K}(\|\mathbf{y}_1^0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_2^0 - \mathbf{x}^*\|^2) + 2c^{-1}(\log K + 2)\sqrt{K+1}(G^2 + V^2)}{4K} \end{aligned} \quad (3.7)$$

In  $\epsilon$ -LDP, Algorithm 2 has utility loss  $\tilde{O}(\frac{Nd^{3/2}\mathcal{B}_\infty}{\sqrt{N\epsilon}})$ . With  $(\epsilon, \delta)$  relaxation, this bound can be sharpened to  $\tilde{O}(\frac{Nd\mathcal{B}_\infty}{\sqrt{N\epsilon}})$ . The  $\tilde{O}$  is the big O notation that ignores logarithmic factors.

The proof of Theorem 3.2.1 can be found in Appendix I. Clearly, both the proposed scheme matches the lower bound of LDP derived in [34].<sup>1</sup>

## 3.3 Convergence in a View of Random Stochastic Matrix

In general, the framework of private decentralized GD proposed can be described as

$$\mathbf{X}^{k+1} = W_{k+1}\mathbf{X}^k - \xi^{k+1}\nabla F(\mathbb{E}[W_{k+1}]\mathbf{X}^k) + \Delta^{k+1}, \quad (3.8)$$

where  $W_{k+1}$  is a random stochastic random determined by the randomized aggregation of each agent. We call  $W$  s stochastic matrix if the sum of entries in each row is 1.  $W$  is further doubly stochastic if  $W^T$  is stochastic as well. When we assume  $\mathbb{E}[W_{k+1}]$  is further

<sup>1</sup>Please note that the setup is a bit different in [34] where their objective function is scaled by  $\frac{1}{N}$ , i.e.,  $F(\mathbf{X}) = \frac{1}{N} \sum_{i=1}^N f_i(\mathbf{x}_i)$ . Thus their utility bound is scaled by  $\frac{1}{N}$  comparing to our case.



doubly stochastic, applying the convexity of  $F(\cdot)$  that  $\langle \nabla F(\mathbb{E}[W_{k+1}]\mathbf{X}^k), \mathbb{E}[W_{k+1}]\mathbf{X}^k - \mathbf{X}^* \rangle \geq F(\mathbb{E}[W_{k+1}]\mathbf{X}^k) - F(\mathbf{X}^*)$ , we have

$$\begin{aligned}
& \mathbb{E}[F(\mathbb{E}[W_{k+1}]\mathbf{X}^k) - F(\mathbf{X}^*)] \\
& \leq \frac{1}{2}(\zeta^{k+1})^{-1}(\mathbb{E}[\|\mathbb{E}[W_{k+1}]\mathbf{X}^k - \mathbf{X}^*\|^2 + \|\xi^{k+1}\nabla F(\mathbb{E}[W_{k+1}]\mathbf{X}^k) - \Delta^{k+1}\|^2 - \|\mathbf{X}^{k+1} - \mathbf{X}^*\|^2]) \\
& \leq \frac{1}{2}(\zeta^{k+1})^{-1}(\mathbb{E}[\|\mathbf{X}^k - \mathbf{X}^*\|^2 - \|\mathbf{X}^{k+1} - \mathbf{X}^*\|^2 + \|\xi^{k+1}\nabla F(\mathbb{E}[W_{k+1}]\mathbf{X}^k)\|^2] + \mathbb{E}[\|\Delta^{k+1}\|^2])
\end{aligned} \tag{3.9}$$

Following the proof of Theorem 3.2.1, we may sum up and average out both sides of (3.9) for  $k = 0, 1, \dots, K-1$ , where the left side  $\frac{1}{K} \sum_{k=0}^{K-1} F(\mathbb{E}[W_{k+1}]\mathbf{X}^k) - F(\mathbf{X}^*) \geq F(\frac{\sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}]\mathbf{X}^k}{K}) - F(\mathbf{X}^*)$ . However, such bound can not be straightforwardly used for measuring utility since  $\mathbf{X}^k$  is not necessarily in consensus, i.e.,  $\mathbf{x}_i^k = \mathbf{x}_j^k$  for any  $i, j \in [1 : N]$ , and thus  $F(\mathbb{E}[W_{k+1}]\mathbf{X}^k) - F(\mathbf{X}^*) \geq 0$  may not hold. However, when we assume  $f_i(\cdot)$  is  $L$ -Lipschitz, such gap can be fixed via

$$\begin{aligned}
& \sum_{i=1}^N f_i\left(\frac{\sum_{k=0}^{K-1} \sum_{i=1}^N \mathbb{E}[W_{k+1}(i, \cdot)^T] \mathbf{x}_i^k}{KN}\right) - F(\mathbf{X}^*) \leq F\left(\frac{\sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}]\mathbf{X}^k}{K}\right) - F(\mathbf{X}^*) \\
& + L \sum_{i=1}^N \left\| \frac{\sum_{k=0}^{K-1} \sum_{i=1}^N \mathbb{E}[W_{k+1}(i, \cdot)^T] \mathbf{x}_i^k}{KN} - \frac{\sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}(i, \cdot)^T] \mathbf{x}_i^k}{K} \right\|,
\end{aligned} \tag{3.10}$$

where  $W_{k+1}(i, \cdot)$  denotes the  $i^{\text{th}}$  row of  $W_{k+1}$ . The above analysis is invariant to the randomness of  $W_k$ . Therefore, with convergence guarantee at hand, it is equivalent to considering the rate of  $\frac{\sum_{k=0}^{K-1} \mathbf{X}^k}{K}$  towards the consensus status. Rewrite  $\frac{\sum_{k=0}^{K-1} \mathbf{X}^k}{K}$  in the following form,

$$\frac{\sum_{k=0}^{K-1} \mathbf{X}^k}{K} = \frac{\sum_{k=0}^{K-1} (\prod_{j=1}^k W_j \mathbf{X}^0 + \sum_{j=0}^k \prod_{l=j+1}^k W_l R^j)}{K} \tag{3.11}$$

where for simplicity we rewrite  $\mathbf{X}^{k+1} = W_{k+1}\mathbf{X}^k + \mathbf{R}^{k+1}$  for some remainder term  $\mathbf{R}^{k+1}$  and  $\prod_{l=j+1}^k W_l = \mathbf{I}$  if  $j+1 > k$ . Let us consider an extreme scenario where we assume the communication graph is fully-connected and  $N$  is even for simplicity: a). We fix  $W_{k+1} = W$ , where  $W[i, j] = \frac{1}{N}$ . Here  $W[i, j]$  denotes the entry in  $W$  which lies in the row number  $i$  and column number  $j$ ; b).  $\mathbb{E}[W_{k+1}] = W$  where  $W[i, j] = r_i^{k+1} \times \frac{2}{N}$

if  $j \leq \frac{N}{2}$ , otherwise  $W[i, j] = (1 - r_i^{k+1}) \times \frac{2}{N}$ . Here  $r_i$  is independently and randomly selected in  $(0, 1)$  by agent  $i$  in iteration  $(k + 1)$ . (b) indeed captures the case where each agent  $i$  independently aggregates  $\mathbf{x}_{[1:N]}^k$  in a form  $r_i^{k+1} \frac{\sum_{i=1}^{N/2} \mathbf{x}_i^k}{N/2} + (1 - r_i^{k+1}) \frac{\sum_{i=N/2+1}^N \mathbf{x}_i^k}{N/2}$ . On the other hand, case (a) is equivalent to the GD in a central model, where  $\mathbf{x}_i^k = \mathbf{x}_j^k$  holds across each  $k$  and  $W$  is a matrix with identical rows, i.e., for any  $i_1, i_2, j \in [1 : N]$ ,  $W[i_1, j] = W[i_2, j]$ . Such  $W$  is with the property that for any vector  $\mathbf{x}$ , entries of  $W\mathbf{x}$  are identical. To analyze the rate of  $\frac{\sum_{k=0}^{K-1} \mathbf{x}^k}{K}$  in (3.11) towards consensus, we introduce the following metric  $\phi(\mathbf{x})$ , which denotes the largest deviation between any two coordinates of  $\mathbf{x}$  in absolute value. Consequently,  $\phi(W\mathbf{x}) = 0$  for arbitrary  $\mathbf{x}$ . To quantify the rate w.r.t. the product of randomized stochastic matrix  $W_k$  towards a matrix with identical rows, we provide the following lemma.

**Lemma 3.3.1.** *For any two independent matrices  $W_1$  and  $W_2$  following the same distribution:  $W_l[i, j] = r_i^l \times \frac{2}{N}$  if  $j \leq \frac{N}{2}$ , otherwise  $W_l[i, j] = (1 - r_i^l) \times \frac{2}{N}$  for some independently random weight  $r_i^l \in (0, 1)$ ,  $l = 1, 2$ . Let  $\tilde{W} = W_1 W_2$ , then for any  $i_1, i_2, j \in [1 : N]$ ,*

$$\mathbb{E}[|\tilde{W}(i_1, j) - \tilde{W}(i_2, j)|] \leq \frac{1}{2} \phi(W_2(:, j)) \quad (3.12)$$

*Proof.* It is noted that  $\tilde{W}(i_1, j) = \sum_{l=1}^N W_1(i_1, l) W_2(l, j)$  and  $\tilde{W}(i_2, j) = \sum_{l=1}^N W_1(i_2, l) W_2(l, j)$ . Without loss of generality, we assume  $W_2(j, 1) = \min_l \{W_2(j, l)\}$ . Thus,

$$\begin{aligned} |\tilde{W}(i_1, j) - \tilde{W}(i_2, j)| &= \left| \sum_{l=1}^N (W_1(i_1, l) - W_1(i_2, l)) W_2(l, j) \right| \\ &= \left| \left(1 - \sum_{l=2}^N W_1(i_1, l) - 1 + \sum_{l=2}^N W_1(i_2, l)\right) W_2(1, j) + \sum_{l=2}^N (W_1(i_1, l) - W_1(i_2, l)) W_2(l, j) \right| \\ &= \left| \sum_{l=2}^N (W_1(i_1, l) - W_1(i_2, l)) (W_2(l, j) - W_2(1, j)) \right| \end{aligned} \quad (3.13)$$

It is noted that  $W_2(l, j) - W_2(1, j) \geq 0$  for  $l \geq 2$  which is no bigger than  $\phi(W_2(:, j))$ . Due to the distribution of  $W_1$ , either  $W_1(i_1, l) - W_1(i_2, l) \geq 0, l \in [1 : \frac{N}{2}]$  and  $W_1(i_1, l) - W_1(i_2, l) \leq 0, l \in [\frac{N}{2} + 1 : N]$ , or  $W_1(i_1, l) - W_1(i_2, l) \leq 0, l \in [1 : \frac{N}{2}]$  and  $W_1(i_1, l) - W_1(i_2, l) \geq 0, l \in [\frac{N}{2} + 1 : N]$ . Therefore, by taking expectation on both sides

of (3.13),

$$\begin{aligned}\mathbb{E}[|\tilde{W}(i_1, j) - \tilde{W}(i_2, j)|] &\leq \frac{N}{2} \mathbb{E}[\max_l W_1(i_1, l) - \min_l W_2(i_2, l)] \mathbb{E}[\phi(W_2(:, j))] \\ &= \left(\frac{3}{4} - \frac{1}{4}\right) \mathbb{E}[\phi(W_2(:, j))] = \frac{1}{2} \mathbb{E}[\phi(W_2(:, j))].\end{aligned}\quad (3.14)$$

□

The above lemma indicates that with randomness in generating  $W_k$ , though we can not guarantee that  $W_k$  is with identical rows, the product of  $W_k$  convergence to a matrix with identical rows in an exponential rate in expectation. On the other hand, it is noted that the expression of  $\sum_{k=0}^{K-1} \mathbf{X}^k / K$  in (3.11) always contains a term  $\frac{R^{k-1}}{K}$ . Therefore, even with fixed  $W$  of identical rows, the  $\phi(\sum_{k=0}^{K-1} \mathbf{X}^k / K)$  is still dominated by  $\phi(\frac{R^{k-1}}{K})$ . Thus, the effect of deviation amongst the rows of  $W_k$  due to randomization on the consensus rate of  $\sum_{k=0}^{K-1} \mathbf{X}^k / K$  can be negligible.

Fig. 3-1 and Fig. 3-2 show the simulation results w.r.t. the average of  $\mathbb{E}[\phi(\left(\prod_{k=1}^K W_k\right)(:, j))]$  across  $j = 1, 2, \dots, N$ . As for parameter selection, we test  $K$  ranging from 1 to 10 over six kinds of randomly generated connected graph where  $N$  denotes the number of nodes and  $E$  denotes the number of edges in the figure legend. We provide two variants of  $W_k$  generation in the two figures, respectively. In Fig. 3-1, we independently select  $W_k(i, i) = r_i^k$  and  $W_k(i, j) = \frac{1-r_i^k}{d_i}$  if node  $i$  and  $j$  are neighbors in the graph; otherwise  $W_k(i, j) = 0$ .  $\{r_i^k\}$  are i.i.d. uniform variables in  $(0, 1)$ . Here  $d_i$  is the degree of node  $i$ . Indeed, Fig. 3-1 captures the parameter selection in Algorithm 1 where  $r_i^k$  corresponds to  $\frac{\mathbf{r}_i^k}{D_i}$ . In Fig. 3-2,  $W_k(i, :)$  are independently generated in a way that we uniformly select two nodes among node  $i$  and its neighbors, of which the indexes are denoted by  $j_1$  and  $j_2$ . Then  $W_k(i, j_1) = \frac{r_i^k}{d_i+1}$  and  $W_k(i, j_2) = \frac{1-r_i^k}{d_i+1}$ , while the rest entries  $W_k(i, j) = \frac{1}{d_i+1}$  if  $i$  and  $j$  are connected, otherwise 0. We also compare the case if we fix  $W_k$  as the expectation of the above two distributions, respectively. In the legend of two figures,  $R$  stands for the random  $W_k$  case while  $F$  represents the fixed case.

Clearly, when the graph is sparse, random  $W_k$  and fixed  $W_k$  share almost the same rate towards consensus. As the graph gets denser, the fixed case will have a faster convergence. Especially, the expectation of  $W^k$  is with identical rows when the

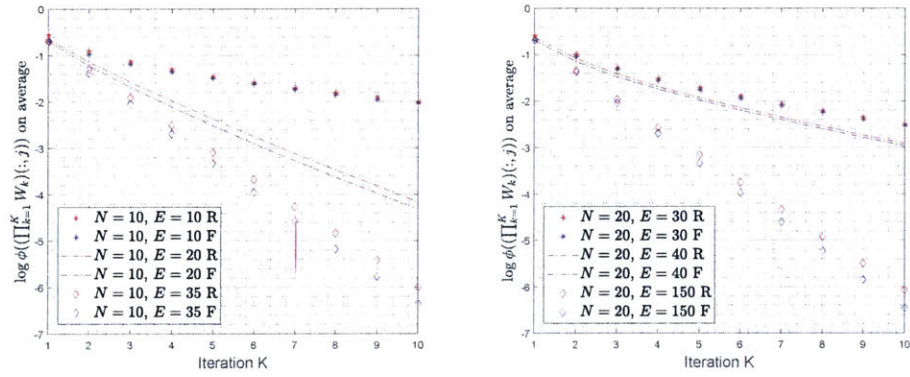


Figure 3-1: Rate of  $\prod_{k=1}^K W_k$  towards identical rows (a)

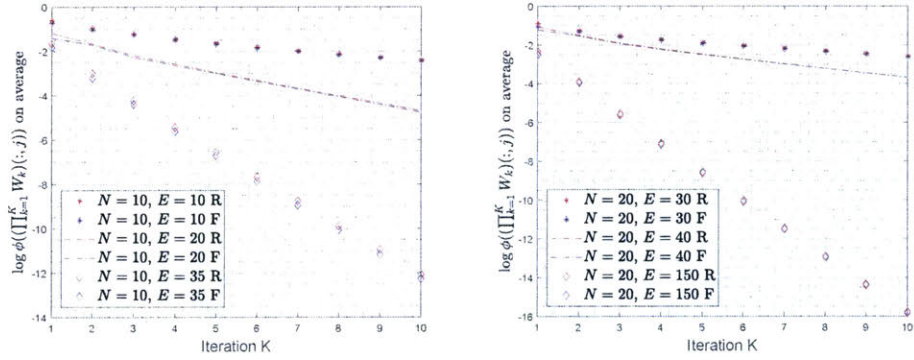


Figure 3-2: Rate of  $\prod_{k=1}^K W_k$  towards identical rows (b)

graph is fully connected. On the other hand, from the simulation, both cases have an exponential decaying rate, which coincides with our theory.

# Chapter 4

## Conclusion

In this thesis, we investigate LDP in interactive decentralized optimization from both asymptotic and non-asymptotic viewpoints. A more comprehensive notion of LDP is developed, which takes the average privacy loss into account. Resort to secret sharing, we present an efficient way to secretly aggregate updates, where the consequent privacy guarantee is improved by a factor of  $\sqrt{N'}$ . Here  $N'$  is the number of semi-honest nodes in the network. Besides, a framework of first-order based decentralized optimization with random aggregators is proposed. We prove such further randomness can incur an amplified average privacy guarantee while directly applying Laplace or Gaussian mechanism on optimization algorithms, as existing works, the worst privacy loss equals the average one. Rigorous utility-privacy analysis has been provided. In addition, via the convergence of random stochastic matrix, we develop more refined evidence to explain why the performance of proposed decentralized optimization algorithm with further random aggregations is almost the same as that without such randomization. Simulation results are provided to validate the theory.



# Appendix A

## Proof of Theorem 2.1.1

Considering  $\mathbf{x}_{i_0}^k$  for arbitrary  $i_0 \in \{1, 2, \dots, N\}$ , based on the definition of secret splitting, one may reconstruct  $\mathbf{x}_{i_0}$  if and only all the  $N$  shares have been collected (and decrypted properly in the encryption case). In the first step of Algorithm 1, where  $(N-1)$  random shares have been distributed to the remaining  $(N-1)$  nodes, if there exists another semi-honest node, denoted by  $v_{i_1}$  with shares  $\mathbf{s}_{i_0 i_1}^k$ , then, in the second step, each node sums up all the shares received as  $\hat{\mathbf{s}}_{[1:N]}^k$  and broadcasts. It is clear that  $\hat{\mathbf{s}}_{i_0}^k = \sum_{j=1}^N \mathbf{s}_{j i_0}^k \bmod p$ , of which the reconstruction requires both  $\mathbf{s}_{i_1 i_0}^k$  and  $\mathbf{s}_{i_0 i_0}^k$ , while  $\mathbf{s}_{i_0 i_0}^k$  is a secret of  $v_{i_0}$  and  $\mathbf{s}_{i_1 i_0}^k$  is a secret between  $v_{i_1}$  and  $v_{i_0}$ . With the assumption that  $v_{i_1}$  is not colluding, for  $v_i$ ,  $i \neq i_1, i_0$ , from  $\hat{\mathbf{s}}_{i_0}^k$ , it is impossible to infer either  $\mathbf{s}_{i_1 i_0}^k$  or  $\mathbf{s}_{i_0 i_0}^k$ . With a similar reasoning, since  $N \geq 3$ , the reconstruction of  $\hat{\mathbf{s}}_{i_0}^k$  is also determined by some  $\mathbf{s}_{i i_0}^k$  for  $i \neq i_0, i_1$ , which is unknown to  $v_{i_1}$ . Thus,  $v_{i_1}$  cannot infer  $\mathbf{s}_{i_0 i_0}^k$  either. In a nutshell, either for  $v_i$ ,  $i \neq i_1, i_0$  or  $v_{i_1}$ , at least one share, i.e.,  $\mathbf{s}_{i_0 i_0}^k$ , cannot be inferred and thus  $\mathbf{x}_{i_0}^k$  is secure to at most  $(N-2)$  colluding nodes. From the above analysis, we can merge all semi-honest nodes, denoted by  $\mathcal{H}$  into one and all other colluding nodes can only infer  $\sum_{i \in \mathcal{H}} \mathbf{x}_i^k$  but nothing else.

On the other hand, the underlying noise  $\sum_{i \in \mathcal{H}} \Delta_i^k$  is in  $\mathcal{N}(\mathbf{0}, N' \sigma^2)$ . Following the Gaussian mechanism, in a central view of  $\sum_{i \in \mathcal{H}} \mathbf{x}_i^k$ , it achieves  $(\frac{\epsilon}{\sqrt{N'}}, \delta)$  DP.





# Appendix B

## Proof of Lemma 2.2.1

It is noted that for an  $\mathbf{X} = \mathbf{x}_{[1:N]}^{[0:K]}$  observed, since there is no prior on the inputs  $\mathcal{D} = \{f_1, \dots, f_i, \dots, f_N\}$  and  $\mathcal{D}' = \{f_1, \dots, \hat{f}_i, \dots, f_N\}$ ,

$$\frac{P(\mathcal{D}|\mathbf{x}_{[1:N]}^{[0:K]})}{P(\mathcal{D}'|\mathbf{x}_{[1:N]}^{[0:K]})} = \frac{P(\mathbf{x}_{[1:N]}^{[0:K]}|\mathcal{D})}{P(\mathbf{x}_{[1:N]}^{[0:K]}|\mathcal{D}')} = \frac{P(\mathbf{x}_{[1:N]}^0|\mathcal{D}) \prod_{k=1}^K P(\mathbf{x}_{[1:N]}^k|\mathcal{D}, \mathbf{x}_{[1:N]}^{[0:k-1]})}{P(\mathbf{x}_{[1:N]}^0|\mathcal{D}') \prod_{k=1}^K P(\mathbf{x}_{[1:N]}^k|\mathcal{D}', \mathbf{x}_{[1:N]}^{[0:k-1]})}, \quad (\text{B.1})$$

It is noted that  $\mathcal{D}$  and  $\mathcal{D}'$  differ in  $f_i$  and  $\hat{f}_i$ , to which the distribution of  $\mathbf{x}_{[1:N]\setminus i}$  is invariant, and  $\mathbf{x}_i^k$  only depends on the private function of agent  $i$  and  $\mathbf{x}_{[1:N]}^{[1:k-1]}$ . On the other hand, the initialization of  $\mathbf{x}_{[1:N]}^0$  is independent of the dataset. Thus, (B.1) can be further simplified as

$$\prod_{k=0}^{K-1} \frac{P(\mathbf{x}_i^{k+1}|f_i, \mathbf{x}_{[1:N]}^{[0:k]})}{P(\mathbf{x}_i^{k+1}|\hat{f}_i, \mathbf{x}_{[1:N]}^{[0:k]})} = \prod_{k=0}^{K-1} \prod_{l=1}^d \frac{P(\mathbf{x}_i^{k+1}[l]|f_i, \mathbf{x}_{[1:N]}^{[0:k]})}{P(\mathbf{x}_i^{k+1}[l]|\hat{f}_i, \mathbf{x}_{[1:N]}^{[0:k]})},$$

since the noise on each dimension is i.i.d. By taking the logarithm of the above equation and recalling the definition of  $\epsilon(\mathbf{X})$  in Definition 1, the lemma follows.



# Appendix C

## Proof of Theorem 2.2.1

When we assume  $\Delta_i^{k+1}$  is a Laplace distribution, the distributions  $P(\mathbf{x}_i^{k+1}[j]|f_i, \mathbf{x}_{[1:N]}^{[0:k]})$  in either Algorithm 1 or 2 share a very similar structure. Both follow a mixture Laplace distribution with a random mean. In Algorithm 1 the mean is randomly distributed in an interval starting from  $\mathbf{x}_i^k - D_i^{-1} \nabla f_i(\mathbf{x}_i^k) + D_i^{-1} \boldsymbol{\lambda}_i^k$  to  $\frac{\sum_{j \neq i} \mathbf{x}_j^k}{N-1} - D_i^{-1} \nabla f_i(\mathbf{x}_i^k) + D_i^{-1} \boldsymbol{\lambda}_i^k$ , while in Algorithm 2 the mean is randomly distributed in an interval starting from  $\frac{2}{N} \mathbf{x}_{i_k}^k + \frac{\sum_{j \neq i_k, i_k} \mathbf{x}_j^k}{N} - \eta_{k+1} \nabla f_i(\frac{\sum_{j=1}^N \mathbf{x}_i^k}{N})$  to  $\frac{2}{N} \mathbf{x}_{i_k}^k + \frac{\sum_{j \neq i_k, i_k} \mathbf{x}_j^k}{N} - \eta_{k+1} \nabla f_i(\frac{\sum_{j=1}^N \mathbf{x}_i^k}{N})$ . Without loss of generality, we focus on Algorithm 1. Since  $\Delta_i^{k+1}$  on each dimension is i.i.d. in  $\text{Lap}(0, \beta^{k+1})$ , recalling Lemma 2.1, at iteration  $k+1$ , the bound on privacy loss in the  $l$ -th dimension can be expressed as

$$\epsilon_i^{k+1}(\mathbf{X}) = \sup_{\hat{f}_i \in \mathcal{F}_i} \left| \log \frac{P(\mathbf{x}_i^{k+1}[l]|f_i, \mathbf{x}_{[1:N]}^{[0:k]})}{P(\mathbf{x}_i^{k+1}[l]|\hat{f}_i, \mathbf{x}_{[1:N]}^{[0:k]})} \right| = \max_{|t| \leq D_i^{-1} \mathcal{B}_\infty} \left| \log \frac{\int_{\tau_i^{k+1}[l]} e^{-\beta_{k+1} |\mathbf{x}_i^{k+1}[l]-Y|} dY}{\int_{\tau_i^{k+1}[l]+t} e^{-\beta_{k+1} |\mathbf{x}_i^{k+1}[l]-Y|} dY} \right|. \quad (\text{C.1})$$

We reformulate this problem as follows. For  $X \in \mathbb{R}$ , we consider

$$\max_{|t| \leq D_i^{-1} \mathcal{B}_\infty} \left| \log \frac{\int_0^\omega \beta_{k+1} e^{-\beta_{k+1} |X-Y|} dY}{\int_t^{t+\omega} \beta_{k+1} e^{-\beta_{k+1} |X-Y|} dY} \right|, \quad (\text{C.2})$$

for some positive numbers  $\omega, \beta_{k+1}$  and  $\mathcal{B}_\infty$ . Here,  $\omega$  corresponds to the length of the interval.

For a fixed  $t$ ,  $|t| \leq \mathcal{B}_\infty$ , if  $X \notin [0, \omega] \cup [t, \omega + t]$ , then

$$\left| \log \frac{\int_0^\omega \beta_{k+1} e^{-\beta_{k+1}|X-Y|} dY}{\int_t^{t+\omega} \beta_{k+1} e^{-\beta_{k+1}|X-Y|} dY} \right| = \left| \log \frac{\int_0^\omega e^{-\beta_{k+1}|X-Y|} dY}{e^{\beta_{k+1}t} \int_0^\omega e^{-\beta_{k+1}|X-Y|} dY} \right| = |\beta_{k+1}t| \leq D_i^{-1} \beta_{k+1} \mathcal{B}_\infty.$$

In the following, without loss of generality, we assume  $X \in [0, \omega]$ , then  $\int_0^\omega \beta_{k+1} e^{-\beta_{k+1}|X-Y|} dY = 2 - e^{-\beta_{k+1}X} - e^{-\beta_{k+1}(\omega-X)}$ . First, supposing that  $X \in [t, \omega+t]$ , then  $\int_t^{\omega+t} \beta_{k+1} e^{-\beta_{k+1}|X-Y|} dY = 2 - e^{-\beta_{k+1}(X-t)} - e^{-\beta_{k+1}(\omega+t-X)}$ . To show

$$e^{-\beta_{k+1}|t|} \leq \frac{2 - e^{-\beta_{k+1}X} - e^{-\beta_{k+1}(\omega-X)}}{2 - e^{-\beta_{k+1}(X-t)} - e^{-\beta_{k+1}(\omega+t-X)}} \leq e^{\beta_{k+1}|t|},$$

it is equivalent to showing

$$\begin{cases} 2e^{\beta_{k+1}|t|} - e^{-\beta_{k+1}X + \beta_{k+1}|t|} - e^{-\beta_{k+1}(\omega-X) + \beta_{k+1}|t|} \geq 2 - e^{-\beta_{k+1}(X-t)} - e^{-\beta_{k+1}(\omega+t-X)}, \\ 2 - e^{-\beta_{k+1}X} - e^{-\beta_{k+1}(\omega-X)} \leq 2e^{\beta_{k+1}|t|} - e^{-\beta_{k+1}(X-t) + \beta_{k+1}|t|} - e^{-\beta_{k+1}(\omega+t-X) + \beta_{k+1}|t|}. \end{cases} \quad (\text{C.3})$$

Due to the symmetry, we merely prove the case that when  $t \geq 0$ , where (C.3) can be rewritten as,

$$\begin{cases} 2e^{\beta_{k+1}t} - e^{-\beta_{k+1}(X-t)} - e^{-\beta_{k+1}(\omega-X-t)} \geq 2 - e^{-\beta_{k+1}(X-t)} - e^{-\beta_{k+1}(\omega+t-X)}, \\ 2 - e^{-\beta_{k+1}X} - e^{-\beta_{k+1}(\omega-X)} \leq 2e^{\beta_{k+1}t} - e^{-\beta_{k+1}(X-2t)} - e^{-\beta_{k+1}(\omega-X)}. \end{cases} \quad (\text{C.4})$$

Clearly, for the first inequality, it suffices to show

$$2(e^{\beta_{k+1}t} - 1) \geq (e^{2\beta_{k+1}t} - 1)e^{-\beta_{k+1}(\omega+t-X)}, \quad (\text{C.5})$$

and it can be further simplified as  $2e^{\beta_{k+1}(\omega+t-X)} \geq e^{\beta_{k+1}t} + 1$ . Such a claim follows clearly as  $\omega - X \geq 0$ . For the second inequality, with similar reasoning, it is equivalent to

$$2e^{\beta_{k+1}X} \geq e^{\beta_{k+1}t} + 1, \quad (\text{C.6})$$

which holds since  $X \geq t$ . At last, we consider  $X \notin [t, t + \omega]$ . Still, due to the symmetry,

we can assume  $t > 0$  and  $X < t$ . Then, it is equivalent to show:

$$\begin{cases} 2e^{\beta_{k+1}t} - e^{-\beta_{k+1}(X-t)} - e^{-\beta_{k+1}(\omega-X)+\beta_{k+1}t} \geq e^{-\beta_{k+1}(t-X)} - e^{-\beta_{k+1}(\omega+t-X)}, \\ 2 - e^{-\beta_{k+1}X} - e^{-\beta_{k+1}(\omega-X)} \leq e^{\beta_{k+1}X} - e^{-\beta_{k+1}(\omega-X)}. \end{cases} \quad (\text{C.7})$$

As for the first inequality, assume that  $g(t) = 2e^{\beta_{k+1}t} - e^{-\beta_{k+1}(X-t)} - e^{-\beta_{k+1}(t-X)} - e^{-\beta_{k+1}(\omega-X)+\beta_{k+1}t} + e^{-\beta_{k+1}(\omega+t-X)}$ . It is noted that when  $t = 0$ ,  $x$  should be also be 0 based on the assumption and  $g(0) = 0$ . On the other hand,

$$\frac{dg}{dt} = \beta_{k+1}(2e^{\beta_{k+1}t} - e^{-\beta_{k+1}(X-t)} + e^{-\beta_{k+1}(t-X)} - e^{-\beta_{k+1}(\omega-X)+\beta_{k+1}t} - e^{-\beta_{k+1}(\omega+t-X)}). \quad (\text{C.8})$$

Since  $X < \omega$ , to show  $g(t)$  is non-decreasing with respect to  $t$ , it suffices to show that,

$$2e^{\beta_{k+1}t} - e^{-\beta_{k+1}(X-t)} + e^{-\beta_{k+1}(t-X)} - e^{-\beta_{k+1}(t-X)+\beta_{k+1}t} - e^{-\beta_{k+1}(t+t-X)} \geq 0.$$

It is clear that  $e^{\beta_{k+1}t} \geq e^{-\beta_{k+1}(X-t)}$  and  $e^{-\beta_{k+1}(t-X)} \geq e^{-\beta_{k+1}(2t-X)}$  as both  $X$  and  $t$  are non-negative. Furthermore,  $e^{\beta_{k+1}t} \geq e^{-\beta_{k+1}(t-X)+\beta_{k+1}t} = e^{\beta_{k+1}X}$  since  $t \geq X$ . Therefore, (C.8) is non-negative. The second inequality of (C.7) is exactly the AM-GM inequality that

$$2 \leq e^{-\beta_{k+1}X} + e^{\beta_{k+1}X}.$$

In a nutshell, we have proven that (C.2) is upper bounded by  $\max_{|t| \leq D_i^{-1} \mathcal{B}_\infty} |t\beta_{k+1}| = \beta_{k+1}D_i^{-1} \mathcal{B}_\infty$ . Moreover, when  $X$  belongs to the intersection of the two intervals,  $(0, \omega)$  and  $(t, \omega + t)$ , the above inequalities are strict, i.e., (C.2) is strictly smaller than  $\beta_{k+1}D_i^{-1} \mathcal{B}_\infty$ , which is the case if we fix all parameters to be constants. Similarly, by replacing  $D_i^{-1}$  with  $\eta_{k+1}$ , we derive the proof for the case of Algorithm 2.



# Appendix D

## Proof of Theorem 2.2.2

We drop all the dependence on  $i$ ,  $k$  and  $l$  for brevity and let  $\alpha = \frac{1}{D_i}$ . Following the normalization in the proof of Theorem 2.1, we still assume  $x$  is a Laplace distribution of which the mean is uniformly distributed in  $[0, \omega]$ , conditional on all prior intermediate outputs. As the corollary of Theorem 2.1,

$$\Theta(x) = \max_{|t| \leq \alpha \mathcal{B}_\infty} \left| \log \frac{\int_0^\omega e^{-\beta|x-y|} dy}{\int_t^{\omega+t} e^{-\beta|x-y|} dy} \right|,$$

where the maximization is achieved when  $t$  either equals to  $\alpha \mathcal{B}_\infty$  or  $-\alpha \mathcal{B}_\infty$ . Here we let  $\alpha = \frac{1}{D_i}$  for brevity. To quantify  $\gamma$ , it suffices to calculate

$$\int_{-\infty}^{\infty} \int_0^\omega \Theta(x) \frac{\beta}{2\omega} e^{-\beta|x-y|} dy dx, \quad (\text{D.1})$$

since the probability density function of  $x$  is  $\int_0^\omega \frac{\beta}{2\omega} e^{-\beta|x-y|} dy$ . With the concavity of  $\log(\cdot)$ , (D.1) is upper bounded by

$$\log \int_{-\infty}^{\infty} \int_0^\omega \max_{t=\pm\alpha \mathcal{B}_\infty} \left\{ \frac{\int_0^\omega e^{-\beta|x-z|} dz}{\int_t^{\omega+t} e^{-\beta|x-z|} dz}, \frac{\int_t^{\omega+t} e^{-\beta|x-z|} dz}{\int_0^\omega e^{-\beta|x-z|} dz} \right\} \frac{\beta}{2\omega} e^{-\beta|x-y|} dx dy. \quad (\text{D.2})$$

Now we take a closer look into  $\max_{t=\pm\alpha \mathcal{B}_\infty} \left\{ \frac{\int_0^\omega e^{-\beta|x-z|} dz}{\int_t^{\omega+t} e^{-\beta|x-z|} dz}, \frac{\int_t^{\omega+t} e^{-\beta|x-z|} dz}{\int_0^\omega e^{-\beta|x-z|} dz} \right\}$ . Still from the corollary of Theorem 2.1, once  $\mathbf{x}^{k+1} \notin [0, \omega]$ ,  $\max_{t=\pm\alpha \mathcal{B}_\infty} \left\{ \frac{\int_0^\omega e^{-\beta|x-z|} dz}{\int_t^{\omega+t} e^{-\beta|x-z|} dz}, \frac{\int_t^{\omega+t} e^{-\beta|x-z|} dz}{\int_0^\omega e^{-\beta|x-z|} dz} \right\} =$

$e^{\alpha\beta\mathcal{B}_\infty}$ .

Since we assume  $\omega > \alpha\mathcal{B}_\infty$ , it is not hard to observe that

- $x \in [0, \frac{\omega - \alpha\mathcal{B}_\infty}{2}]$ ,  $\max_{t=\pm\alpha\mathcal{B}_\infty} \left\{ \frac{\int_0^\omega e^{-\beta|x-z|} dz}{\int_t^{\omega+t} e^{-\beta|x-z|} dz}, \frac{\int_t^{\omega+t} e^{-\beta|x-z|} dz}{\int_0^\omega e^{-\beta|x-z|} dz} \right\} = \frac{\int_t^{\omega+t} e^{-\beta|x-z|} dz}{\int_0^\omega e^{-\beta|x-z|} dz} \Big|_{t=-\alpha\mathcal{B}_\infty}$  ;
- $x \in [\frac{\omega - \alpha\mathcal{B}_\infty}{2}, \frac{\omega}{2}]$ ,  $\max_{t=\pm\alpha\mathcal{B}_\infty} \left\{ \frac{\int_0^\omega e^{-\beta|x-z|} dz}{\int_t^{\omega+t} e^{-\beta|x-z|} dz}, \frac{\int_t^{\omega+t} e^{-\beta|x-z|} dz}{\int_0^\omega e^{-\beta|x-z|} dz} \right\} = \frac{\int_0^\omega e^{-\beta|x-z|} dz}{\int_t^{\omega+t} e^{-\beta|x-z|} dz} \Big|_{t=-\alpha\mathcal{B}_\infty}$  ;
- $x \in [\frac{\omega}{2}, \frac{\omega + \alpha\mathcal{B}_\infty}{2}]$ ,  $\max_{t=\pm\alpha\mathcal{B}_\infty} \left\{ \frac{\int_0^\omega e^{-\beta|x-z|} dz}{\int_t^{\omega+t} e^{-\beta|x-z|} dz}, \frac{\int_t^{\omega+t} e^{-\beta|x-z|} dz}{\int_0^\omega e^{-\beta|x-z|} dz} \right\} = \frac{\int_0^\omega e^{-\beta|x-z|} dz}{\int_t^{\omega+t} e^{-\beta|x-z|} dz} \Big|_{t=\alpha\mathcal{B}_\infty}$  ;
- $x \in [\frac{\omega + \alpha\mathcal{B}_\infty}{2}, \omega]$ ,  $\max_{t=\pm\alpha\mathcal{B}_\infty} \left\{ \frac{\int_0^\omega e^{-\beta|x-z|} dz}{\int_t^{\omega+t} e^{-\beta|x-z|} dz}, \frac{\int_t^{\omega+t} e^{-\beta|x-z|} dz}{\int_0^\omega e^{-\beta|x-z|} dz} \right\} = \frac{\int_t^{\omega+t} e^{-\beta|x-z|} dz}{\int_0^\omega e^{-\beta|x-z|} dz} \Big|_{t=\alpha\mathcal{B}_\infty}$  .

Thus, fortunately, we can avoid the complicated integral at least in  $x \in [0, \frac{\omega}{2} - \alpha\mathcal{B}_\infty]$ , or  $x \in [\frac{\omega}{2} + \alpha\mathcal{B}_\infty, \omega]$  where it is simplified to  $O(\int_t^{\omega+t} e^{-\beta|x-y|} dy)$ . Now we can split  $\mathbb{R}$  into three parts,  $(-\infty, 0) \cup (\omega, \infty)$ ,  $[0, \frac{\omega - \alpha\mathcal{B}_\infty}{2}] \cup [\frac{\omega + \alpha\mathcal{B}_\infty}{2}, \omega]$  and the rest  $(\frac{\omega - \alpha\mathcal{B}_\infty}{2}, \frac{\omega + \alpha\mathcal{B}_\infty}{2})$ . To avoid the tedious term when  $x \in (\frac{\omega - \alpha\mathcal{B}_\infty}{2}, \frac{\omega + \alpha\mathcal{B}_\infty}{2})$ , we simply use the global upper bound to simplify them to derive a closed-form expression but one may obtain the expression of  $\gamma$  exactly with the same reasoning. Note the symmetry on  $t = \pm\alpha\mathcal{B}_\infty$ , (D.2) is upper bounded by

$$\log \left\{ e^{\omega\alpha\mathcal{B}_\infty} \left[ 1 - 2 \int_0^{(\omega - \alpha\mathcal{B}_\infty)/2} \int_0^\omega \frac{\beta}{2\omega} e^{-\beta|x-y|} dy dx \right] + 2 \int_0^{(\omega - \alpha\mathcal{B}_\infty)/2} \int_{-\alpha\mathcal{B}_\infty}^{\omega - \alpha\mathcal{B}_\infty} \frac{\beta}{2\omega} e^{-\beta|x-y|} dy dx \right\}. \quad (\text{D.3})$$



# Appendix E

## Simulation Results

We test the proposed schemes and state-of-art approaches on regularized empirical risk minimization (ERM) tasks. We use the standard *Adult* dataset from the UCI Machine Learning Repository. For simplicity, we call the task as UCI in the following. In UCI, the dataset consists of demographic records, including age, sex and income etc. in 15 total features. We try to predict whether the annual income of an individual is above  $50k$ . After processing of the data, we remove all individuals with missing values and normalize both columns (features) and rows (individuals) while converting labels  $\{\geq 50k, < 50k\}$  to  $\{0, 1\}$ . The training samples are denoted by  $\{\mathbf{z}_j^i \in \mathbb{R}^{14}, \mathcal{L}_j^i \in \{0, 1\} | i = 1, \dots, N, j = 1, \dots, b_i\}$ . Consistent with [43], [42], we select  $\mathcal{L}(\mathbf{x}) = \log(1 + \exp(-x))$ . Thus,  $N$  agents are collaboratively solving the following logistic regression:

$$\min_{\mathbf{x}} \sum_{i=1}^N f_i(\mathbf{x}) = \sum_{i=1}^N \left( \frac{1}{b_i} \sum_{j=1}^{b_i} \log(1 + \exp(-\mathcal{L}_j^i \mathbf{x}^T \mathbf{z}_j^i)) + \frac{1}{2} \|\mathbf{x}\|^2 \right).$$

UCI is run with different parameter settings. 10 independent runs of each algorithm for comparison are performed and each agent is randomly assigned 100 samples from the dataset. In each run, the communication graph is randomly generated using the given  $N$  and the number of edges  $|\mathcal{E}|$ .

In UCI, four examples (a), (b), (c) and (d) are provided. We uniformly assume that  $D_i = D = 10$  and  $\zeta = 0.5$  in all cases for Algorithm 1. In the case of private

ADMM, previous works all assume fixed parameters in the optimization protocol. In [42], the Lagrangian multiplier at the beginning of each iteration is perturbed, while [48] considers the output perturbation at the end of each iteration. Further, in [43], the authors introduce a sequence of increasing step penalty, which can bring better utility-privacy tradeoff empirically. For [42], [48] with constant fixed penalty, we assume  $\Gamma_i = 0.5D$  and  $\rho_i = \frac{0.5D}{|\mathcal{N}_i|}$ , corresponding to the expectation of the penalty terms in Algorithm 1. Here  $\mathcal{N}_i$  denotes the neighbors of agent  $i$ . As for [43], we follow their setting that  $\Gamma_i^k = 0.5 \times 1.02^k |\mathcal{N}_i|$  and  $\rho_i^k = 0.5 \times 1.02^k$ .<sup>1</sup>

In the privacy part, with the same assumption in [43], we assume  $f_i$  and  $\hat{f}_i$  may only differ in one sample and thus, due the normalization,  $\mathcal{B}_\infty = \frac{1}{b_i} = 0.01$ , and  $\mathbf{J} = \frac{2.8}{DB_i}$ , the Jacobian constant required by [43] in their privacy analysis. It is noted that derivative of  $\mathcal{L}$  is within  $(-1, 0]$ , while the privacy analysis of [48] requires a global sensitivity on that of  $\nabla \mathcal{L}$ . This makes their bound in this example too loose and we omit their privacy loss bound in our simulation. Following the setting of [43], [48], we also use a diminishing noise by selecting  $\beta_k = 1.02^k$ . The results of Example (a) are illustrated in Fig. 2-1, where  $N = 10$ ,  $|\mathcal{E}| = 20$ . The accuracy logarithm defined by  $\log \|\mathbf{x}_i^k - \mathbf{x}_i^*\|/d$ , across 100 iterations averaged across 10 runs. The difference between the best and the worst accuracy over 100 runs is also marked. In example (b), with illustration shown in Fig. E-2, under the same setting, we test algorithms in a large-scale case where  $N = 100$ ,  $|\mathcal{E}| = 200$ .

In Example (c), we present an interesting variant to Algorithm 2. With Theorem 2.3, it is clear that under our framework, a larger interval can produce a better privacy amplification. Instead of the construction in Algorithm 2, one can construct a random aggregator by more aggressively utilizing the divergence among  $\mathbf{x}_{[1:N]}^k$ . For an instance, let  $\widetilde{\mathcal{N}}_i$  denote the neighbors of node  $i$  including  $i$ . To aggregate  $\mathbf{x}_{j \in \mathcal{N}_i}^k$ , for each dimension  $l \in [1 : d]$ , let

$$\alpha_{i,\min}^k = \min_{j \in \widetilde{\mathcal{N}}_i} \mathbf{x}_j^k[l], \quad \text{and} \quad \alpha_{i,\max}^k = \max_{j \in \widetilde{\mathcal{N}}_i} \mathbf{x}_j^k[l].$$

---

<sup>1</sup>We do not optimize the increasing penalty here but we find that in some cases by proper selection, a privacy loss reduction can be achieved empirically at a cost of relatively small utility compromise. Such techniques can also be applied in our algorithms.

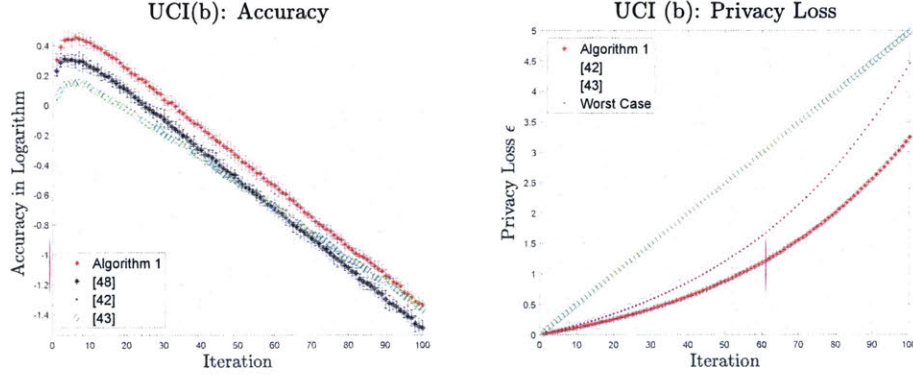


Figure E-1. (b). Simulation on Graphs  $N = 100$ ,  $|\mathcal{E}| = 200$

---

#### Algorithm 2\*

---

**Input:** Local functions  $f_i$  and a diminishing sequence  $\{\eta_k\}$

Initialize  $\mathbf{x}_{[1:N]}^0$ .

**for**  $k = 0, 1, 2, \dots, K - 1$  **do**

**Agents**  $i = 1$  **to**  $N$  **do in parallel :**

**for**  $l = 1, 2, \dots, d$  **do**

            Randomly and independently generating a weight  $w$  within  $(0, 1)$  and then updating  $\mathbf{x}_i^{k+1}[l]$ :

$$\mathbf{x}_i^{k+1}[l] := w \cdot \alpha_{i,\min}^k + (1 - w) \cdot \alpha_{i,\max}^k - \eta_{k+1} \nabla f_i(\mathbf{x}_i^k) + \Delta_i^{k+1}[l]. \quad (\text{E.1})$$

            Exchange  $\mathbf{x}_i^{k+1}$  with Neighbors

**end for**

**end for**

---

We consider the updating subroutine (E.1) in Algorithm 2\*, i.e., each coordinate of  $\mathbf{x}_i^{k+1}[l]$  is uniformly selected between the largest  $\max_{j \in \mathcal{N}_i} \mathbf{x}_j^k[l]$  and the smallest  $\min_{j \in \mathcal{N}_i} \mathbf{x}_j^k[l]$ .

Different from Algorithm 1 and 2 which are controllable, aggregation in Algorithm 2\* will bring some compromise in convergence. Following [45], we also select a diminishing step size, where for [45],  $\eta_k = 0.9^k$  and in Algorithm 2\*,  $\eta_k = 0.93^k$  for balance. Not surprisingly, Algorithm 2\* has a worse convergence at the beginning since information from neighbors is less efficiently merged but finally [45] and Algorithm 2\* achieve almost the same utility loss. However, the privacy loss of Algorithm 2\* is only 30% of [45], as shown in Fig. E-2. Since Algorithm 1 applies a constant step size, it

has better accuracy but worse privacy loss, where all noises are fixed to be  $\beta_k = 1.02^k$ .

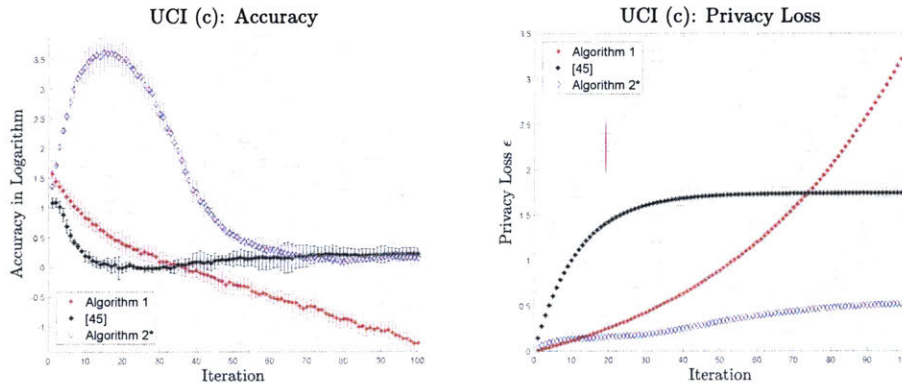


Figure E-2: (c). Simulation on Graphs  $N = 10$ ,  $|\mathcal{E}| = 40$

Finally, we provide the performance of non-private optimization in UCI. With the same parameter setting as before, we test Algorithm 1 without first-order approximation and conventional ADMM with fixed parameters. In addition, we set the step size  $\eta_k = 0.95^k$  and test Algorithm 2 and conventional decentralized GD (DGD) with fixed parameters. The performance is illustrated as follows. In Fig. E-3, the graph is randomly generated with  $N = 10$  and  $|\mathcal{E}| = 20$ , same as Example (a) in Fig. 2-1; while in Fig. E-4,  $N = 100$  and  $|\mathcal{E}| = 200$ , same as Example (b).

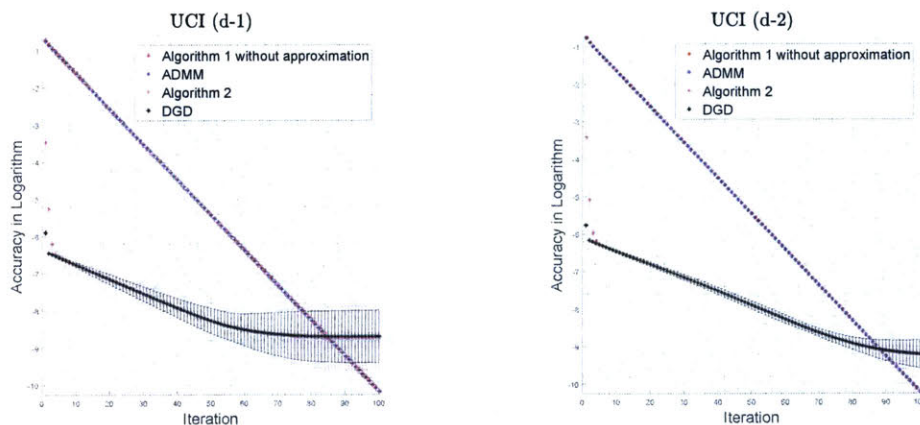


Figure E-3: (d). Simulation on Graphs  $N = 10$ ,  $|\mathcal{E}| = 20$

Figure E-4: (d). Simulation on Graphs  $N = 100$ ,  $|\mathcal{E}| = 200$

From the above, randomization defined in Algorithm 1 and 2 does not incur

accuracy loss, which is consistent with our analysis.



# Appendix F

## Proof of Theorem 3.1.1

Since  $f_i$  is convex,

$$\langle \nabla f_i(\tilde{\mathbf{x}}_i^{k+1}), \tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^* \rangle \geq f_i(\tilde{\mathbf{x}}_i^{k+1}) - f_i(\mathbf{x}_i^*). \quad (\text{F.1})$$

Due to the optimality condition satisfied in (J.1), we have

$$\nabla f_i(\tilde{\mathbf{x}}_i^{k+1}) = A_i^T(\boldsymbol{\lambda}^k - \boldsymbol{\rho}_i^{k+1}(A_i \tilde{\mathbf{x}}_i^{k+1} + \sum_{j \neq i} A_j \mathbf{x}_j^k - \mathbf{c})) + \boldsymbol{\Gamma}_i^{k+1}(\mathbf{x}_i^k - \tilde{\mathbf{x}}_i^{k+1}). \quad (\text{F.2})$$

Also from the KKT condition, for the optimal states  $\mathbf{x}_{[1:N]}^*$ ,  $\sum_{i=1}^N A_i \mathbf{x}_i^* = \mathbf{c}$ . Substitute the above equations into (F.1), we have

$$\begin{aligned} & \langle A_i^T(\boldsymbol{\lambda}^k - \boldsymbol{\rho}_i^{k+1}(A \tilde{\mathbf{X}}^k - \mathbf{c}) + \boldsymbol{\rho}_i^{k+1} A_i(\mathbf{x}_i^k - \tilde{\mathbf{x}}_i^{k+1})), \tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^* \rangle + (\tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^*)^T \boldsymbol{\Gamma}_i^{k+1}(\mathbf{x}_i^k - \tilde{\mathbf{x}}_i^{k+1}) \\ & \geq f_i(\tilde{\mathbf{x}}_i^{k+1}) - f_i(\mathbf{x}_i^*). \end{aligned} \quad (\text{F.3})$$

Here  $A \mathbf{X}^k = \sum_{i=1}^N A_i \mathbf{x}_i^k$ . Let  $\boldsymbol{\lambda}^k = \boldsymbol{\lambda}^k - \boldsymbol{\lambda} + \boldsymbol{\lambda}$ , we can rewrite (F.3) as

$$\begin{aligned} & \langle \boldsymbol{\lambda}^k - \boldsymbol{\lambda}, A_i(\tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^*) \rangle + (\tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^*)^T (\boldsymbol{\Gamma}_i^{k+1} + A_i^T \boldsymbol{\rho}_i^{k+1} A_i)(\mathbf{x}_i^k - \tilde{\mathbf{x}}_i^{k+1}) \\ & - \langle A \tilde{\mathbf{X}}^k - \mathbf{c}, \boldsymbol{\rho}_i^{k+1} A_i(\tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^*) \rangle \geq f_i(\tilde{\mathbf{x}}_i^{k+1}) - f_i(\mathbf{x}_i^*) - \boldsymbol{\lambda}^T A_i(\tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^*). \end{aligned} \quad (\text{F.4})$$

Summing up the above formulas for  $i = 1, 2, \dots, N$ , we have

$$\begin{aligned} & \frac{1}{\zeta} (\langle \lambda^k - \tilde{\lambda}^{k+1}, \lambda^k - \tilde{\lambda}^{k+1} \rangle + \langle \tilde{\lambda}^{k+1} - \lambda, \lambda^k - \tilde{\lambda}^{k+1} \rangle) + \sum_{i=1}^N (\tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^*)^T (\mathbf{\Gamma}_i^{k+1} + A_i^T \boldsymbol{\rho}_i^{k+1} A_i) (\mathbf{x}_i^k - \tilde{\mathbf{x}}_i^{k+1}) \\ & - \langle A \tilde{\mathbf{X}}^k - \mathbf{c}, \sum_{i=1}^N \boldsymbol{\rho}_i^{k+1} A_i (\tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^*) \rangle \geq F(\tilde{\mathbf{X}}^{k+1}) - F(\mathbf{X}_i^*) - \lambda^T A (\tilde{\mathbf{X}}^{k+1} - \mathbf{X}^*). \end{aligned} \quad (\text{F.5})$$

Let the matrix  $\mathbf{G} = \text{blkdiag}\{\mathbf{D}_1, \dots, \mathbf{D}_N, \frac{1}{\zeta}\}$ , where  $\mathbf{D}_i = D_i \cdot \mathbf{I}$ . With the identity:

$\|\mathbf{u}^k - \mathbf{u}^*\|_{\mathbf{G}}^2 - \|\tilde{\mathbf{u}}^{k+1} - \mathbf{u}^*\|_{\mathbf{G}}^2 = 2(\tilde{\mathbf{u}}^{k+1} - \mathbf{u}^*)^T \mathbf{G}(\mathbf{u}^k - \tilde{\mathbf{u}}^{k+1}) + \|\mathbf{u}^k - \tilde{\mathbf{u}}^{k+1}\|_{\mathbf{G}}^2$ , we have

$$\begin{aligned} & \|\mathbf{u}^k - \mathbf{u}^*\|_{\mathbf{G}}^2 - \|\tilde{\mathbf{u}}^{k+1} - \mathbf{u}^*\|_{\mathbf{G}}^2 - \|\mathbf{u}^k - \tilde{\mathbf{u}}^{k+1}\|_{\mathbf{G}}^2 + \frac{2}{\zeta} \|\lambda^k - \tilde{\lambda}^{k+1}\|^2 \\ & - 2\langle A \mathbf{X}^k - \mathbf{c}, \sum_{i=1}^N \boldsymbol{\rho}_i^{k+1} A_i (\tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^*) \rangle \geq 2[F(\tilde{\mathbf{X}}^{k+1}) - F(\mathbf{X}_i^*) - \lambda^T A (\tilde{\mathbf{X}}^{k+1} - \mathbf{X}^*)]. \end{aligned} \quad (\text{F.6})$$

Here  $\mathbf{u}^k = (\mathbf{x}_{[1:N]}^k, \lambda^k)$ ,  $\tilde{\mathbf{u}}^{k+1} = (\tilde{\mathbf{x}}_{[1:N]}^{k+1}, \tilde{\lambda}^{k+1})$  and  $\mathbf{u}^* = (\mathbf{x}_{[1:N]}^*, \lambda)$ . Let  $h^{k+1} = \|\mathbf{u}^k - \tilde{\mathbf{u}}^{k+1}\|_{\mathbf{G}}^2 - \frac{2}{\zeta} \|\lambda^k - \tilde{\lambda}^{k+1}\|^2 + 2\langle A \mathbf{X}^k - \mathbf{c}, \sum_{i=1}^N \boldsymbol{\rho}_i^{k+1} A_i (\tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^*) \rangle$ , which can be further rewritten as

$$\begin{aligned} h^{k+1} &= \sum_{i=1}^N D_i \|\mathbf{x}_i^k - \tilde{\mathbf{x}}_i^{k+1}\|^2 - \frac{1}{\zeta} \|\lambda^k - \tilde{\lambda}^{k+1}\|^2 + 2\langle \sum_{i=1}^N A_i (\mathbf{x}_i^k - \tilde{\mathbf{x}}_i^{k+1} + \tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^*), \sum_{i=1}^N \boldsymbol{\rho}_i^{k+1} A_i (\tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^*) \rangle \\ &= \sum_{i=1}^N D_i \|\mathbf{x}_i^k - \tilde{\mathbf{x}}_i^{k+1}\|^2 - \zeta \|A(\tilde{\mathbf{X}}^{k+1} - \mathbf{X}^*)\|^2 + 2\langle \sum_{i=1}^N A_i (\mathbf{x}_i^k - \tilde{\mathbf{x}}_i^{k+1}), \sum_{i=1}^N \boldsymbol{\rho}_i^{k+1} A_i (\tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^*) \rangle \\ & \quad + 2\rho_0 \|A(\tilde{\mathbf{X}}^{k+1} - \mathbf{X}^*)\|^2 + 2\langle \sum_{i=1}^N A_i (\tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^*), \sum_{i=1}^N (\boldsymbol{\rho}_i^{k+1} - \rho_0 \mathbf{I}) A_i (\tilde{\mathbf{x}}_i^{k+1} - \mathbf{x}_i^*) \rangle. \end{aligned} \quad (\text{F.7})$$

We assume  $\underline{\rho}_i \cdot \mathbf{I} \leq \boldsymbol{\rho}_i^{k+1} \leq \bar{\rho}_i \cdot \mathbf{I}$  and  $\sigma_{\max, i}$  is the largest singular value of  $A_i$ . To guarantee that  $h^{k+1} \geq 0$ , it suffices to let

$$\begin{cases} \frac{D_j}{N^2} (\underline{\rho}_i - \frac{\zeta}{2}) \geq \sigma_{\max, j}^2 \bar{\rho}_i^2, \\ \frac{1}{(N-1)^2} (\underline{\rho}_i - \frac{\zeta}{2}) (\underline{\rho}_j - \frac{\zeta}{2}) \geq (\bar{\rho}_i - \underline{\rho}_i)^2 \end{cases} \quad (\text{F.8})$$



In the following, we generalize the above with respect to  $\mathbf{x}_i^{k+1} = \tilde{\mathbf{x}}_i^{k+1} + \Delta_i^{k+1}$ . It is noted that

$$f_i(\mathbf{x} + \Delta_i^{k+1}) = f_i(\mathbf{x}) + \langle \Delta_i^{k+1}, \nabla f_i(\mathbf{x}) \rangle + \frac{\sqrt{M}}{2} \|\Delta_i^{k+1}\|^2.$$

In addition,  $\mathbb{E}[\|\tilde{\mathbf{u}}^{k+1} - \mathbf{u}^{k+1}\|_G^2] = \|\tilde{\mathbf{u}}^{k+1} - \mathbf{u}^*\|_G^2 + \mathbb{E}[\|\mathbf{u}^{k+1} - \mathbf{u}^*\|_G^2]$  since the mean of noise added is zeros. Putting things together, the claim follows.



# Appendix G

## Proof of Theorem 3.1.2

Summing up (3.2) for  $k = 0, 1, \dots, K - 1$  and applying Jensen inequality, we have

$$\begin{aligned} \mathbb{E}[K(F(\bar{\mathbf{X}}^K) - F(\mathbf{X}^*)) - \lambda^T A \sum_{k=0}^K (\mathbf{X}^{k+1} - \mathbf{X}^*)] &\leq \sum_{k=0}^{K-1} \mathbb{E}[F(\mathbf{X}^{k+1}) - F(\mathbf{X}^*) + \lambda^T A \mathbf{X}^{k+1}] \\ &\leq \frac{1}{2} \sum_{k=0}^{K-1} \sum_{i=1}^N (\sqrt{M} + D_i + \frac{1}{\zeta}) \mathbb{E}[\|\Delta_i^{k+1}\|^2] + \mathbb{E}[\|\mathbf{u}^0 - \mathbf{u}^*\|_G^2] \end{aligned} \quad (\text{G.1})$$

Since  $\lambda^0 = \mathbf{0}$  and  $\lambda$  in  $\mathbf{u}^*$  is an arbitrary point in  $\mathbb{R}^d$ , by letting  $\lambda = \mathbf{0}$ , we have

$$\mathbb{E}[F(\bar{\mathbf{X}}^K) - F(\mathbf{X}^*)] \leq \frac{\|\mathbf{X}^0 - \mathbf{X}^*\|_{G_x}^2}{2K} + \frac{\sum_{k=1}^K \sum_{i=1}^N (\sqrt{M} + D_i + \frac{1}{\zeta}) \mathbb{E}[\|\Delta_i^k\|^2]}{K} \quad (\text{G.2})$$

On the other hand, with KKT condition,  $-\lambda^{*T} A (\bar{\mathbf{X}}^K - \mathbf{X}^*) = -\nabla F(\mathbf{X}^*) (\bar{\mathbf{X}}^K - \mathbf{X}^*)$ .

Applying convexity of  $F(\cdot)$ , we have

$$F(\mathbf{X}^*) \leq F(\bar{\mathbf{X}}^K) - \nabla F(\mathbf{X}^*)^T (\bar{\mathbf{X}}^K - \mathbf{X}^*) \quad (\text{G.3})$$

Thus, by letting  $\lambda = 2\lambda^*$  in (G.1) and adding  $-\lambda^{*T} A (\bar{\mathbf{X}}^K - \mathbf{X}^*)$  on both side of (G.3)

$$\begin{aligned} -\lambda^{*T} A \sum_{k=0}^K (\mathbf{X}^{k+1} - \mathbf{X}^*) &\leq F(\bar{\mathbf{X}}^K) - F(\mathbf{X}^*) - 2\lambda^{*T} A \sum_{k=0}^K (\mathbf{X}^{k+1} - \mathbf{X}^*) \\ &\leq \frac{\|\mathbf{X}^0 - \mathbf{X}^*\|_{G_x}^2 + \frac{4}{\zeta} \|\lambda^*\|^2}{2T} + \frac{\sum_{k=1}^K \sum_{i=1}^N (\sqrt{M} + D_i + \frac{1}{\zeta}) \mathbb{E}[\|\Delta_i^k\|^2]}{K} \end{aligned} \quad (\text{G.4})$$

Putting the upper and lower bounds of  $F(\bar{\mathbf{X}}^K) - F(\mathbf{X}^*)$  together, the claim follows.

# Appendix H

## Proof of Theorem 3.1.3

With the smooth assumptions on  $\nabla f_i$ , i.e., for any  $\mathbf{x}$  and  $\mathbf{y}$ ,

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|^2 \leq M \|\mathbf{x} - \mathbf{y}\|^2,$$

we have the following fact: for any  $\mathbf{z}$

$$f_i(\mathbf{x}) - f_i(\mathbf{y}) \leq \nabla f_i^T(\mathbf{z})(\mathbf{x} - \mathbf{y}) + \frac{M}{2} \|\mathbf{x} - \mathbf{z}\|^2,$$

Therefore, by replacing  $\mathbf{x}_i^{k+1}$  with  $\mathbf{x}_i^k$  in (F.1), all the deductions in Theorem 3.1.1 keep the same except that the term  $\sum_{i=1}^N D_i \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\|^2$  in the expression of  $h^{k+1}$  in (F.7) becomes  $\sum_{i=1}^N (D_i - M) \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\|^2$ . Therefore, by replacing  $D_i$  in (3.3) with  $D_i - M$ , the claim still follows obviously.



# Appendix I

## Proof of Theorem 3.2.1

Without loss of generality, we scale the original objective function by a factor of  $\frac{1}{N}$ : let  $F(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N f_i(\mathbf{x}_i)$  and accordingly the updating rule of Algorithm 2 becomes,

$$\mathbf{x}_i := \mathbf{w}_i \mathbf{y}_1^k + (\mathbf{I}_d - \mathbf{w}_i) \mathbf{y}_2^k - \eta_{k+1} \nabla f_i\left(\frac{\mathbf{y}_1^k + \mathbf{y}_2^k}{2}\right) + \Delta_i. \quad (\text{I.1})$$

For each  $k \in [0 : K - 1]$ ,

$$\begin{aligned} \|\mathbf{y}_1^{k+1} - \mathbf{x}^*\|^2 &= \left\| \frac{1}{|S_{2k+1}|} \sum_{i \in S_{2k+1}} (\mathbf{w}_i \mathbf{y}_1^k + (\mathbf{I}_d - \mathbf{w}_i) \mathbf{y}_2^k - \mathbf{x}^* - \zeta_{k+1} \nabla f_i\left(\frac{\mathbf{y}_1^k + \mathbf{y}_2^k}{2}\right)) \right\|^2 \\ &= \left\| \frac{1}{|S_{2k+1}|} \sum_{i \in S_{2k+1}} (\mathbf{w}_i (\mathbf{y}_1^k - \mathbf{x}^*) + (\mathbf{I}_d - \mathbf{w}_i) (\mathbf{y}_2^k - \mathbf{x}^*)) \right\|^2 \\ &\quad - 2\zeta_{k+1} \left\langle \frac{1}{|S_{2k+1}|} \sum_{i \in S_{2k+1}} (\mathbf{w}_i \mathbf{y}_1^k + (\mathbf{I}_d - \mathbf{w}_i) \mathbf{y}_2^k) - \mathbf{x}^*, \sum_{i \in S_{2k+1}} \frac{\nabla f_i\left(\frac{\mathbf{y}_1^k + \mathbf{y}_2^k}{2}\right) + \zeta_{k+1}^{-1} \Delta_i}{|S_{2k+1}|} \right\rangle \\ &\quad + \left\| \sum_{i \in S_{2k+1}} \frac{1}{|S_{2k+1}|} (\zeta_{k+1} \nabla f_i\left(\frac{\mathbf{y}_1^k + \mathbf{y}_2^k}{2}\right) + \Delta_i) \right\|^2 \end{aligned} \quad (\text{I.2})$$

In the following, we will use the following inequality that, if for  $i \in [1 : N]$ ,  $\omega_i > 0$  and  $\sum_{i=1}^N \omega_i = 1$ , then for arbitrary  $N$  real numbers  $r_{[1:N]}$ , the following holds,

$$\left( \sum_{i=1}^N \omega_i r_i \right)^2 \leq \sum_{i=1}^N \omega_i r_i^2. \quad (\text{I.3})$$

It is noted that in (I.2), the sum of weights of  $(\mathbf{y}_1^k - \mathbf{x}^*)$  and  $(\mathbf{y}_2^k - \mathbf{x}^*)$  is always the identity. With (I.3), by taking expectation on both sides of (I.2), we have

$$\begin{aligned} \mathbb{E}[\|\mathbf{y}_1^{k+1} - \mathbf{x}^*\|^2] \leq & \mathbb{E}\left[\frac{\|\mathbf{y}_1^k - \mathbf{x}^*\|^2 + \|\mathbf{y}_2^k - \mathbf{x}^*\|^2}{2}\right] + (\eta_{k+1}^2 G^2 + \mathbb{E}[\|\bar{\Delta}_1^k\|^2]) \\ & - 2\eta_{k+1} \left\langle \frac{\mathbf{y}_1^k + \mathbf{y}_2^k}{2}, \nabla F\left(\frac{\mathbf{y}_1^k + \mathbf{y}_2^k}{2}\right) \right\rangle, \end{aligned} \quad (\text{I.4})$$

where  $\bar{\Delta}_1^k = \frac{1}{|S_{2k+1}|} \sum_{i \in S_{2k+1}} \Delta_i$ . Here we use the fact that since we randomly divide the agents into  $2K$  subsets and thus for each  $i$ ,

$$\mathbb{E}\left[\frac{1}{|S_i|} \sum_{i \in S_i} \nabla f_i(\mathbf{x})\right] = \nabla F(\mathbf{x})$$

for arbitrary  $\mathbf{x}$ . On the other hand,  $\mathbb{E}\left[\frac{1}{|S_i|} \sum_{i \in S_i} \mathbf{w}_i\right] = \frac{1}{2} \mathbf{I}_d$ , where the selection of  $\mathbf{w}_i$  is independent to the agents grouping scheme.

Similarly, we can derive the similar upper bound of  $\mathbb{E}[\|\mathbf{y}_2^{k+1} - \mathbf{x}^*\|^2]$  that

$$\begin{aligned} \mathbb{E}[\|\mathbf{y}_2^{k+1} - \mathbf{x}^*\|^2] \leq & \mathbb{E}\left[\frac{\|\mathbf{y}_1^k - \mathbf{x}^*\|^2 + \|\mathbf{y}_2^k - \mathbf{x}^*\|^2}{2}\right] + (\eta_{k+1}^2 G^2 + \mathbb{E}[\|\bar{\Delta}_2^k\|^2]) \\ & - 2\eta_{k+1} \left\langle \frac{\mathbf{y}_1^k + \mathbf{y}_2^k}{2}, \nabla F\left(\frac{\mathbf{y}_1^k + \mathbf{y}_2^k}{2}\right) \right\rangle. \end{aligned} \quad (\text{I.5})$$

where  $\bar{\Delta}_2^k = \frac{1}{|S_{2k+2}|} \sum_{i \in S_{2k+2}} \|\Delta\|_i$ . Applying the fact that  $\left\langle \frac{\mathbf{y}_1^k + \mathbf{y}_2^k}{2} - \mathbf{x}^*, \nabla F\left(\frac{\mathbf{y}_1^k + \mathbf{y}_2^k}{2}\right) \right\rangle \geq F\left(\frac{\mathbf{y}_1^k + \mathbf{y}_2^k}{2}\right) - F(\mathbf{x}^*)$  and taking average on both sides of (I.4) and (I.5), we can bound  $F\left(\frac{\mathbf{y}_1^k + \mathbf{y}_2^k}{2}\right) - F(\mathbf{x}^*)$  as,

$$\begin{aligned} F\left(\frac{\mathbf{y}_1^k + \mathbf{y}_2^k}{2}\right) - F(\mathbf{x}^*) \leq & \frac{\eta_{k+1}^{-1}}{4} (\|\mathbf{y}_1^k - \mathbf{x}^*\|^2 + \|\mathbf{y}_2^k - \mathbf{x}^*\|^2 - \|\mathbf{y}_1^{k+1} - \mathbf{x}^*\|^2 - \|\mathbf{y}_2^{k+1} - \mathbf{x}^*\|^2) \\ & + \left( \frac{\eta_{k+1}}{2} G^2 + \frac{\mathbb{E}[\|\bar{\Delta}_1^k\|^2] + \mathbb{E}[\|\bar{\Delta}_2^k\|^2]}{4\eta_{k+1}} \right). \end{aligned} \quad (\text{I.6})$$

Before we can derive a global convergence analysis, we need to give an upper bound on  $\|\mathbf{y}_1^k - \mathbf{x}^*\|$  and  $\|\mathbf{y}_2^k - \mathbf{x}^*\|$  with the initial divergence  $\|\mathbf{y}_1^0 - \mathbf{x}^*\|$ ,  $\|\mathbf{y}_2^0 - \mathbf{x}^*\|$  and the noise  $\mathbb{E}[\|\bar{\Delta}_1^k\|^2]$  and  $\mathbb{E}[\|\bar{\Delta}_2^k\|^2]$ . It is noted that, with rearrangement on (I.6) and the



fact  $F(\mathbf{x}) - F(\mathbf{x}^*) \geq 0$ ,

$$\mathbb{E}[\|\mathbf{y}_1^{k+1} - \mathbf{x}^*\|^2] \leq \mathbb{E}\left[\frac{\|\mathbf{y}_1^k - \mathbf{x}^*\|^2 + \|\mathbf{y}_2^k - \mathbf{x}^*\|^2}{2}\right] + \eta_{k+1}^2(G^2 + \mathbb{E}[\|\eta_{k+1}^{-1}\Delta_1^{k+1}\|^2]). \quad (\text{I.7})$$

When we select  $\eta_k = \frac{1}{c\sqrt{k}}$ ,  $\sum_{k=1}^K \eta_k^2 = \sum_{k=1}^K \frac{1}{c^2 k} \leq \frac{\log K + 1}{c^2}$  since  $k \leq K$ . The above renders an upper bound on  $\mathbb{E}\left[\frac{\|\mathbf{y}_1^{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}_2^{k+1} - \mathbf{x}^*\|^2}{2}\right]$  that

$$\mathbb{E}\left[\frac{\|\mathbf{y}_1^{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}_2^{k+1} - \mathbf{x}^*\|^2}{2}\right] \leq \mathbb{E}\left[\frac{\|\mathbf{y}_1^0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_2^0 - \mathbf{x}^*\|^2}{2}\right] + \frac{\log K + 1}{c^2}(G^2 + V^2),$$

with the assumption  $\max_{k,j \in \{1,2\}} \mathbb{E}[(\bar{\Delta}_j^k / \eta_k)^2] \leq V^2$ . Thus, (I.6) can be further formulated as

$$\begin{aligned} & \sum_{k=0}^{K-1} \frac{\mathbb{E}[F(\frac{\mathbf{y}_1^k + \mathbf{y}_2^k}{2}) - F(\mathbf{x}^*)]}{K} \\ & \leq \frac{\sum_{k=1}^K (\eta_{k+1}^{-1} - \eta_k^{-1})(\|\mathbf{y}_1^k - \mathbf{x}^*\|^2 + \|\mathbf{y}_2^k - \mathbf{x}^*\|^2) + \eta_1^{-1}(\|\mathbf{y}_1^0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_2^0 - \mathbf{x}^*\|^2) + 2 \sum_{k=0}^{K-1} \eta_{k+1}(G^2 + V^2)}{4K} \\ & = O\left(\frac{c\sqrt{K}(\|\mathbf{y}_1^0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_2^0 - \mathbf{x}^*\|^2) + c^{-1}(\log K + 2)\sqrt{K+1}(G^2 + V^2)}{K}\right), \end{aligned} \quad (\text{I.8})$$

and  $\mathbb{E}[\sum_{i=1}^N f_i(\sum_{k=0}^{K-1} \sum_{i=1}^N \mathbf{x}_i^k / NK) - f_i(\mathbf{x}^*)] \leq \sum_{k=0}^{K-1} \sum_{i=1}^N \frac{\mathbb{E}[f_i(\bar{\mathbf{x}}^k)] - f_i(\mathbf{x}^*)}{K}$ . Here we use the trick of SGD proof that selecting such a sequence of decreasing step size. To finally disclose the utility-privacy tradeoff, we specify the parameter of noise. In pure  $\epsilon$ -LDP setting, since the sensitivity is bounded by  $\mathcal{B}_\infty$  in  $l_\infty$ , on each dimension we may add a noise following  $\text{Lap}(0, \frac{\epsilon}{d\eta_k \mathcal{B}_\infty})$  to produce a total  $\epsilon$  loss from  $d$  dimensions. Under the relaxed  $(\epsilon, \delta)$ -DP setting, with the strong composition theorem [50], the variance  $\mathbb{E}[(\Delta_i^k / \eta_k)^2]$  can be reduced to  $\tilde{O}(\frac{K}{N} d (\frac{\sqrt{d} \mathcal{B}_\infty}{\epsilon})^2)$ . Substituting those into (I.8), we complete the proof of the Theorem that

$$\begin{cases} \text{pure } \epsilon - \text{LDP} : \tilde{O}\left(\frac{\sqrt{\|\mathbf{y}_1^0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_2^0 - \mathbf{x}^*\|^2} (G + \frac{\sqrt{K}}{\sqrt{N}} \frac{d^{3/2} \mathcal{B}_\infty}{\epsilon})}{\sqrt{K}}\right) = \tilde{O}\left(\frac{d^{3/2} \mathcal{B}_\infty}{\sqrt{N} \epsilon}\right) \\ \text{relaxed } (\epsilon, \delta) - \text{LDP} : \tilde{O}\left(\frac{d \mathcal{B}_\infty}{\sqrt{N} \epsilon}\right) \end{cases} \quad (\text{I.9})$$



# Appendix J

## Utility Analysis of Algorithm 1 under Strongly-convex Assumptions

For simplicity, we first consider the case via the following updating rule without noise perturbation:

$$\mathbf{x}_i^{k+1} := \arg \min_{\mathbf{x}_i} f_i(\mathbf{x}_i) - \lambda^{kT} \left( A_i \mathbf{x}_i + \sum_{j \neq i} A_j \mathbf{x}_j^k - \mathbf{c} \right) + \frac{1}{2} \left\| A_i \mathbf{x}_i + \sum_{j \neq i} A_j \mathbf{x}_j^k - \mathbf{c} \right\|_{\rho_i^{k+1}}^2 + \frac{1}{2} \left\| \mathbf{x}_i - \mathbf{x}_i^k \right\|_{\Gamma_i^{k+1}}^2, \quad (\text{J.1})$$

and the Lagrangian multiplier is updated accordingly as  $\lambda^{k+1} := \lambda^k - \gamma_i^{k+1} \rho_i^{k+1} (\sum_{i=1}^N A_i \mathbf{x}_i^{k+1} - \mathbf{c})$ . Here we following the notions before and define  $\mathbf{u}^k = (\mathbf{x}_{[1:N]}^k, \lambda^k)$ . In the following, we assume  $f_i$  is  $M_i$ -smooth and  $m_i$ -strongly convex, i.e.,  $m_i \|\mathbf{x} - \mathbf{y}\|^2 \leq (\mathbf{x} - \mathbf{y})^T (\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y}))$ , for any  $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ .

**Lemma J.0.1.** *Following the above updating rule (J.1),  $\mathbf{u}^k$  converges linearly to  $\mathbf{u}^*$  with penalty  $D_i \cdot \mathbf{I} = A_i^T \rho_i^{k+1} A_i + \Gamma_i^{k+1}$ , where  $D_i$  is a constant, if*

$$\alpha < \frac{2m_i}{N \rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2 + \check{\rho}_{i,\max}^{2(k+1)} \sigma_{i,\max}^2}, \quad \rho^0 > \frac{N}{2\alpha}, \quad D_i > \max \left\{ \rho_{i,\max}^{k+1} \sigma_{i,\max}^2, \frac{N \sigma_{i,\max}^2}{\alpha} \right\}, \quad \zeta < 2\rho^0 - \frac{N}{\alpha},$$

for some positive  $\alpha$ ,  $\zeta$  and  $\rho^0$ . Here  $\check{\rho}_{i,\max}^{k+1}$  is the diagonal element of matrix  $\rho_i^{k+1} - \rho^0 \cdot \mathbf{I}$  with the maximal absolute value and  $\sigma_{i,\max}$  is the largest singular value of  $A_i$ . More

specifically,

$$\|\mathbf{u}^k - \mathbf{u}^*\|_G^2 \geq (1 + p) \|\mathbf{u}^{k+1} - \mathbf{u}^*\|_G^2,$$

for some  $p > 0$ , where  $G = \text{diag}(D_1 \cdot \mathbf{I}, \dots, D_N \cdot \mathbf{I}, \zeta \cdot \mathbf{I})$ . The selection of  $p$  is specified in (J.15).

*Proof.* Since the proximal term  $\|\mathbf{x}_i - \mathbf{x}_i^k\|_{\Gamma_i^{k+1}}^2$  is required to be nonnegative, the matrix  $\Gamma_i^{k+1}$  should be positive definite. With  $D_i \cdot \mathbf{I} = A_i^T \rho_i^{k+1} A_i + \Gamma_i^{k+1}$ , we just need to guarantee that the  $D_i$  satisfy  $D_i - \sigma_{\max}(A_i^T \rho_i^{k+1} A_i) > 0$  where  $\sigma_{\max}(Z)$  and  $\sigma_{\min}(Z)$  denote the maximal and the minimal non-zero singular value of  $Z$ , respectively. It leads to  $D_i > \rho_{i,\max}^{k+1} \sigma_{i,\max}^2$  where  $\sigma_{i,\max}$  is the largest singular value of  $A_i$  and  $\rho_{i,\max}^{k+1}$  is the maximum diagonal element of  $\rho_i^{k+1}$ .

To show the linear convergence, it suffices to determine  $p > 0$  such that,

$$\|\mathbf{u}^k - \mathbf{u}^*\|^2 \geq (1 + p) \|\mathbf{u}^{k+1} - \mathbf{u}^*\|^2, \quad (\text{J.2})$$

which can be reformulated as

$$\|\mathbf{u}^k - \mathbf{u}^*\|^2 - \|\mathbf{u}^{k+1} - \mathbf{u}^*\|^2 \geq p \|\mathbf{u}^{k+1} - \mathbf{u}^*\|^2. \quad (\text{J.3})$$

With the strong convexity,

$$\langle \mathbf{x} - \mathbf{y}, \nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y}) \rangle \geq m_i \|\mathbf{x} - \mathbf{y}\|^2. \quad (\text{J.4})$$

And from (J.1), we have

$$\nabla f_i(\mathbf{x}_i^{k+1}) = A_i^T (\lambda^k - \rho_i^{k+1} (A_i \mathbf{x}_i^{k+1} + \sum_{j \neq i} A_j \mathbf{x}_j^k - \mathbf{c})) + \Gamma_i^{k+1} (\mathbf{x}_i^k - \mathbf{x}_i^{k+1}). \quad (\text{J.5})$$

Also from the KKT condition, for the optimal states  $\lambda^*$  and  $\mathbf{x}^* = (\mathbf{x}_1^*, \dots, \mathbf{x}_N^*)$

$$\nabla f_i(\mathbf{x}_i^*) = A_i^T \lambda^*, \quad \sum_{i=1}^N A_i \mathbf{x}_i^* = \mathbf{c}. \quad (\text{J.6})$$

Substituting the above equations into (J.4)

$$(\mathbf{x}_i^{k+1} - \mathbf{x}_i^*)^T (A_i^T (\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*) - A_i^T \boldsymbol{\rho}_i^{k+1} (A_i (\mathbf{x}_i^{k+1} - \mathbf{x}_i^k) + \sum_{j=1}^N A_j (\mathbf{x}_j^k - \mathbf{x}_j^*)) + \boldsymbol{\Gamma}_i^{k+1} (\mathbf{x}_i^k - \mathbf{x}_i^{k+1})) \geq m_i \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2. \quad (\text{J.7})$$

Summing up all for each  $i$ , it is noted that  $\sum_{i=1}^N A_i (\mathbf{x}_i^{k+1} - \mathbf{x}_i^*) = \frac{1}{\zeta} (\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1})$  and

$$\begin{aligned} (\mathbf{u}^{k+1} - \mathbf{u}^*)^T G (\mathbf{u}^k - \mathbf{u}^{k+1}) &= \frac{1}{\zeta} (\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^*)^T (\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1}) + \sum_{i=1}^N (\mathbf{x}_i^{k+1} - \mathbf{x}_i^*)^T (A_i^T \boldsymbol{\rho}_i^{k+1} A_i + \boldsymbol{\Gamma}_i^{k+1}) (\mathbf{x}_i^k - \mathbf{x}_i^{k+1}) \\ &\geq -\frac{1}{\zeta} \|\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1}\|^2 + \sum_{i=1}^N m_i \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 + \left( \sum_{i=1}^N \boldsymbol{\rho}_i^{k+1} A_i (\mathbf{x}_i^{k+1} - \mathbf{x}_i^*) \right)^T \left( \sum_{j=1}^N A_j (\mathbf{x}_j^k - \mathbf{x}_j^*) \right). \end{aligned} \quad (\text{J.8})$$

Here, let the matrix  $G = \text{diag}(\{\mathbf{D}_1, \dots, \mathbf{D}_N, \frac{1}{\zeta}\})$ , where  $\mathbf{D}_i = D_i \cdot \mathbf{I}$ , then it suffices to show  $\|\mathbf{u}^k - \mathbf{u}^*\|_G^2 - \|\mathbf{u}^{k+1} - \mathbf{u}^*\|_G^2 \geq p \|\mathbf{u}^{k+1} - \mathbf{u}^*\|_G^2$ . On the other hand,  $\|\mathbf{u}^k - \mathbf{u}^*\|_G^2 - \|\mathbf{u}^{k+1} - \mathbf{u}^*\|_G^2 = 2(\mathbf{u}^{k+1} - \mathbf{u}^*)^T G (\mathbf{u}^k - \mathbf{u}^{k+1}) + \|\mathbf{u}^k - \mathbf{u}^{k+1}\|_G^2$ . Referring to (J.8), it is equivalent to figure out  $p$  such that,

$$\begin{aligned} 2 \sum_{i=1}^N m_i \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 + 2 \left( \sum_{i=1}^N \boldsymbol{\rho}_i^{k+1} A_i (\mathbf{x}_i^{k+1} - \mathbf{x}_i^*) \right)^T \left( \sum_{i=1}^N A_i (\mathbf{x}_i^k - \mathbf{x}_i^*) \right) + \sum_{i=1}^N D_i \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\|^2 \\ - \frac{1}{\zeta} \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\|^2 \geq p \left( \sum_{i=1}^N D_i \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 + \frac{1}{\zeta} \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^*\|^2 \right). \end{aligned} \quad (\text{J.9})$$

From (J.5) and with the fact that  $\mathbf{x}_i^k - \mathbf{x}_i^* = \mathbf{x}_i^k - \mathbf{x}_i^{k+1} + \mathbf{x}_i^{k+1} - \mathbf{x}_i^*$ , we get

$$\begin{aligned} \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^*\|^2 &\leq \frac{1}{\sigma_{i,\min}^2} \|A_i^T (\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^*)\|^2 \\ &= \frac{1}{\sigma_{i,\min}^2} \left\| \nabla f_i (\mathbf{x}_i^{k+1}) - \nabla f_i (\mathbf{x}_i^*) - A_i^T (\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1}) - D_i (\mathbf{x}_i^k - \mathbf{x}_i^{k+1}) + A_i^T \boldsymbol{\rho}_i^{k+1} \sum_{j=1}^N A_j (\mathbf{x}_j^k - \mathbf{x}_j^*) \right\|^2 \\ &\leq \frac{5}{\sigma_{i,\min}^2} (M_i \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 + \sigma_{i,\max}^2 \|\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1}\|^2 + D_i^2 \|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2 + \\ &\quad \rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2 \left\| \sum_{j=1}^N A_j (\mathbf{x}_j^k - \mathbf{x}_j^{k+1}) \right\|^2 + \rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2 \left\| \sum_{j=1}^N A_j (\mathbf{x}_j^{k+1} - \mathbf{x}_j^*) \right\|^2), \end{aligned} \quad (\text{J.10})$$

where  $\sigma_{i,\min}$  is the smallest nonzero singular value of  $A_i$ . For simplicity,  $\rho_{i,\max}^{2(k+1)} =$

$(\rho_{i,\max}^{k+1})^2$ . Now, we substitute (J.10) to (J.9), and then it can be reformulated as

$$\begin{aligned}
& \sum_i (2m_i - \frac{5M_i p}{\zeta N \sigma_{i,\min}^2} - D_i p) \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 + \sum_i (D_i - \frac{5p D_i^2}{\zeta N \sigma_{i,\min}^2}) \|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2 \\
& - (\frac{1}{\zeta} + \frac{5p}{\zeta N} \sum_{i=1}^N \frac{\sigma_{i,\max}^2}{\sigma_{i,\min}^2}) \|\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1}\|^2 + 2(\sum_{i=1}^N \boldsymbol{\rho}_i^{k+1} A_i(\mathbf{x}_i^{k+1} - \mathbf{x}_i^*))^T (\sum_{j=1}^N A_j(\mathbf{x}_j^k - \mathbf{x}_j^*)) \\
& - \frac{5p}{\zeta N} \sum_{i=1}^N \frac{\rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2}{\sigma_{i,\min}^2} \left( \left\| \sum_{j=1}^N A_j(\mathbf{x}_j^k - \mathbf{x}_j^{k+1}) \right\|^2 + \frac{1}{\zeta^2} \|\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1}\|^2 \right) \geq 0.
\end{aligned} \tag{J.11}$$

Moreover,  $2(\sum_{i=1}^N \boldsymbol{\rho}_i^{k+1} A_i(\mathbf{x}_i^{k+1} - \mathbf{x}_i^*))^T (\sum_{j=1}^N A_j(\mathbf{x}_j^k - \mathbf{x}_j^*))$  can be rewritten as

$$\begin{aligned}
& 2(\sum_{i=1}^N \boldsymbol{\rho}_i^{k+1} A_i(\mathbf{x}_i^{k+1} - \mathbf{x}_i^*))^T (\sum_{j=1}^N A_j(\mathbf{x}_j^k - \mathbf{x}_j^*)) \\
& = 2(\sum_{i=1}^N \boldsymbol{\rho}_i^{k+1} A_i(\mathbf{x}_i^{k+1} - \mathbf{x}_i^*))^T (\sum_{j=1}^N A_j(\mathbf{x}_j^k - \mathbf{x}_j^{k+1})) + \frac{2}{\zeta^2} (\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1})^T \boldsymbol{\rho}^0 (\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1}) + \\
& \frac{2}{\zeta} (\sum_{i=1}^N (\boldsymbol{\rho}_i^{k+1} - \boldsymbol{\rho}^0) A_i(\mathbf{x}_i^{k+1} - \mathbf{x}_i^*))^T (\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1}) \\
& \geq \frac{2\rho^0}{\zeta^2} \|\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1}\|^2 - \sum_{i=1}^N \alpha N \rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2 \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 - \sum_{i=1}^N \frac{N \sigma_{i,\max}^2}{\alpha} \|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2 - \\
& \sum_{i=1}^N \alpha \check{\rho}_{i,\max}^{2(k+1)} \sigma_{i,\max}^2 \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 - \frac{N}{\alpha \zeta^2} \|\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1}\|^2,
\end{aligned} \tag{J.12}$$

where  $\boldsymbol{\rho}^0 = \rho^0 \cdot \mathbf{I}$  and  $\check{\rho}_{i,\max}^{k+1}$  is the maximum diagonal element of matrix  $\boldsymbol{\rho}_i^{k+1} - \boldsymbol{\rho}^0$ .

Further, we have the following AM-GM inequality

$$\left\| \sum_{j=1}^N A_j(\mathbf{x}_j^k - \mathbf{x}_j^{k+1}) \right\|^2 \leq N \sum_{j=1}^N \sigma_{j,\max}^2 \|\mathbf{x}_j^k - \mathbf{x}_j^{k+1}\|^2. \tag{J.13}$$

Combining (29), (30) and (31), we find that it suffices to find out  $p$  such that

$$\begin{aligned}
& \sum_{i=1}^N \left( 2m_i - \frac{5M_i p}{\zeta N \sigma_{i,\min}^2} - D_i p - \alpha N \rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2 - \alpha \check{\rho}_{i,\max}^{2(k+1)} \sigma_{i,\max}^2 \right) \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 + \\
& \sum_{i=1}^N \left( D_i - \frac{N \sigma_{i,\max}^2}{\alpha} - \frac{5p D_i^2}{\zeta N \sigma_{i,\min}^2} - \frac{5p \sigma_{i,\max}^2}{\zeta} \sum_{j=1}^N \frac{\rho_{j,\max}^{2(k+1)} \sigma_{j,\max}^2}{\sigma_{j,\min}^2} \right) \|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2 + \quad (\text{J.14}) \\
& \left( \frac{2\rho^0}{\zeta^2} - \frac{N}{\alpha \zeta^2} - \frac{1}{\zeta} - \frac{5p}{\zeta} \sum_{i=1}^N \left( \frac{\rho_{i,\max}^{2(k+1)}}{\zeta^2} + \frac{1}{N} \right) \frac{\sigma_{i,\max}^2}{\sigma_{i,\min}^2} \right) \|\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1}\|^2 \geq 0.
\end{aligned}$$

Therefore,  $p$  can be selected as

$$\min \left\{ \frac{2m_i - \alpha N \rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2 - \alpha \check{\rho}_{i,\max}^{2(k+1)} \sigma_{i,\max}^2}{\frac{5M_i}{\zeta N \sigma_{i,\min}^2} + D_i}, \frac{D_i - \frac{N \sigma_{i,\max}^2}{\alpha}}{\frac{5D_i^2}{\zeta N \sigma_{i,\min}^2} + \frac{5\sigma_{i,\max}^2}{\zeta} \sum_{j=1}^N \frac{\rho_{j,\max}^{2(k+1)} \sigma_{j,\max}^2}{\sigma_{j,\min}^2}}, \frac{\frac{2\rho^0}{\zeta} - \frac{N}{\alpha \zeta} - 1}{5 \sum_{i=1}^N \left( \frac{\rho_{i,\max}^{2(k+1)}}{\zeta^2} + \frac{1}{N} \right) \frac{\sigma_{i,\max}^2}{\sigma_{i,\min}^2}} \right\}. \quad (\text{J.15})$$

To guarantee that  $p > 0$ , the parameters  $\alpha$ ,  $D_i$ ,  $\rho^0$  and  $\zeta$  should satisfy:

$$\begin{cases} \alpha < \frac{2m_i}{N \rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2 + \check{\rho}_{i,\max}^{2(k+1)} \sigma_{i,\max}^2}, \\ D_i > \max \left\{ \rho_{i,\max}^{k+1} \sigma_{i,\max}^2, \frac{N \sigma_{i,\max}^2}{\alpha} \right\}, \\ \rho^0 > \frac{N}{2\alpha}, \\ \zeta < 2\rho^0 - \frac{N}{\alpha}. \end{cases} \quad (\text{J.16})$$

□

Similar to Theorem 3.1.3, we consider the first-order approximation based updating rule,

$$\mathbf{x}_i^{k+1} := \mathbf{D}_i^{-1} \left[ \mathbf{A}_i^T (\boldsymbol{\lambda}^k - \rho_i^{k+1} \left( \sum_{j \neq i} \mathbf{A}_j \mathbf{x}_j^k - \mathbf{c} \right)) + \boldsymbol{\Gamma}_i^{k+1} \mathbf{x}_i^k - \nabla f_i(\mathbf{x}_i^k) \right]. \quad (\text{J.17})$$

To quantify the loss from the approximation, we provide the following lemma.

**Lemma J.0.2.** *First-order approximation based ADMM, with modified updating procedure (J.17) still enjoys the linear convergence rate with proper penalty selection specified in (J.22)*

*Proof.* Under the strong continuity of both  $f_i$  and its gradient  $\nabla f_i$ , for any  $\mathbf{x}$  and  $\mathbf{y}$ ,

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|^2 \leq M_i \|\mathbf{x} - \mathbf{y}\|^2,$$

and we use the following fact, for any  $\mathbf{z}$

$$f_i(\mathbf{x}) - f_i(\mathbf{y}) \leq \nabla f_i^T(\mathbf{z})(\mathbf{x} - \mathbf{y}) + \frac{M_i}{2} \|\mathbf{x} - \mathbf{z}\|^2,$$

and with strong convexity we have

$$\frac{m_i}{2} \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 + \nabla f_i(\mathbf{x}_i^*)^T(\mathbf{x}_i^{k+1} - \mathbf{x}_i^*) \leq f_i(\mathbf{x}_i^{k+1}) - f_i(\mathbf{x}_i^*) \leq \nabla f_i^T(\mathbf{x}_i^k)(\mathbf{x}_i^{k+1} - \mathbf{x}_i^*) + \frac{M_i}{2} \|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2. \quad (\text{J.18})$$

On the other hand, since  $A_i^T \boldsymbol{\lambda}^* = \nabla f_i(\mathbf{x}_i^*)$ , thus

$$\frac{m_i}{2} \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 \leq (\mathbf{x}_i^{k+1} - \mathbf{x}_i^*)^T (\nabla f(\mathbf{x}_i^k) - A_i^T \boldsymbol{\lambda}^*) + \frac{M_i}{2} \|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2.$$

Recalling (J.17) that  $\nabla f_i(\mathbf{x}_i^k) = A_i^T(\boldsymbol{\lambda}^k - \boldsymbol{\rho}_i^{k+1}(A_i \mathbf{x}_i^{k+1} + \sum_{j \neq i} A_j \mathbf{x}_j^k - \mathbf{c})) + \boldsymbol{\Gamma}_i^{k+1}(\mathbf{x}_i^k - \mathbf{x}_i^{k+1})$ ,

we have the following,

$$\frac{m_i}{2} \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 \leq (\mathbf{x}_i^{k+1} - \mathbf{x}_i^*)^T (A_i^T(\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*) - A_i^T \boldsymbol{\rho}_i^{k+1} \sum_{j=1}^N A_j(\mathbf{x}_j^k - \mathbf{x}_j^*) + \mathbf{D}_i(\mathbf{x}_i^k - \mathbf{x}_i^{k+1})) + \frac{M_i}{2} \|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2. \quad (\text{J.19})$$



Due to the approximation, we have a different bound as

$$\begin{aligned}
\|\lambda^{k+1} - \lambda^*\|^2 &\leq \frac{1}{\sigma_{i,\min}^2} \left\| \nabla f_i(\mathbf{x}_i^k) - \nabla f_i(\mathbf{x}_i^*) - A_i^T(\lambda^k - \lambda^{k+1}) - D_i(\mathbf{x}_i^k - \mathbf{x}_i^{k+1}) + A_i^T \rho_i^{k+1} \sum_{j=1}^N A_j(\mathbf{x}_j^k - \mathbf{x}_j^*) \right\|^2 \\
&\stackrel{(a)}{\leq} \frac{5}{\sigma_{i,\min}^2} (2M_i \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 + \sigma_{i,\max}^2 \|\lambda^k - \lambda^{k+1}\|^2 + (D_i^2 + 2M_i) \|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2 + \\
&\quad \rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2 \left\| \sum_{j=1}^N A_j(\mathbf{x}_j^k - \mathbf{x}_j^{k+1}) \right\|^2 + \rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2 \left\| \sum_{j=1}^N A_j(\mathbf{x}_j^{k+1} - \mathbf{x}_j^*) \right\|^2),
\end{aligned} \tag{J.20}$$

where (a) is from the fact that  $\|\nabla f_i(\mathbf{x}_i^k) - \nabla f_i(\mathbf{x}_i^*)\|^2 \leq M_i \|\mathbf{x}_i^k - \mathbf{x}_i^{k+1} + \mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 \leq 2M_i \|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2 + 2M_i \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2$ . The rest of the proof is similar to that of Lemma J.0.1, and the  $\rho$  can be selected as

$$\min \left\{ \frac{m_i - \alpha N \rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2 - \alpha \rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2}{\frac{10M_i}{\zeta N \sigma_{i,\min}^2} + D_i}, \frac{D_i - \frac{N \sigma_{i,\max}^2}{\alpha} - M_i}{\frac{5(D_i^2 + 2M_i)}{\zeta N \sigma_{i,\min}^2} + \frac{5 \sigma_{i,\max}^2}{\zeta} \sum_{j=1}^N \frac{\rho_{j,\max}^{2(k+1)} \sigma_{j,\max}^2}{\sigma_{j,\min}^2}}, \frac{\frac{2\rho^0}{\zeta} - \frac{N}{\alpha\zeta} - 1}{5 \sum_{i=1}^N \left( \frac{\rho_{i,\max}^{2(k+1)}}{\zeta^2} + \frac{1}{N} \right) \frac{\sigma_{i,\max}^2}{\sigma_{i,\min}^2}} \right\}, \tag{J.21}$$

with parameters:

$$\begin{cases} \alpha < \frac{m_i}{N \rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2 + \rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2}, \\ D_i > \max\{\rho_{i,\max}^{k+1} \sigma_{i,\max}^2, \frac{N \sigma_{i,\max}^2}{\alpha} + M_i\}, \\ \rho^0 > \frac{N}{2\alpha}, \\ \zeta < 2\rho^0 - \frac{N}{\alpha}. \end{cases} \tag{J.22}$$

□

Finally, we consider the perturbation version: an independent noise  $\Delta_i^k$  is added at the end of the updating procedure (J.17) as that of Algorithm 1.

**Theorem J.0.1.** *With the same assumptions as Lemma J.0.2, if the updating procedure further perturbs with an independent noise  $\Delta_i^{k+1}$ , same as Algorithm 1, there*

exists a constant  $a \in (0, 1)$  and residual  $R^k$  such that

$$\|\mathbf{u}^k - \mathbf{u}^*\|_G^2 \leq a^k \|\mathbf{u}^0 - \mathbf{u}^*\|_G^2 + R^k, \quad (\text{J.23})$$

where the expression of  $a$  and  $R^k$  can be found in the following proof.<sup>1</sup> With a total  $\epsilon$  LDP budget,  $\mathbb{E} \|\mathbf{u}^k - \mathbf{u}^*\|^2 = \tilde{O}(\frac{d^3 N \mathcal{B}_\infty^2}{\epsilon^2})$  and under relaxed  $(\epsilon, \delta)$ -LDP, this bound is sharpened to  $\tilde{O}(\frac{d^2 N \mathcal{B}_\infty^2}{\epsilon^2})$ , where we ignore other constants with respect to  $A_i$  and  $f_i$ . Here  $\tilde{O}$  is the big- $O$  that ignores logarithmic factors.

*Proof.* From the updating procedure with noise,

$$\mathbf{x}_i^{k+1} = D_i^{-1} (A_i^T \boldsymbol{\rho}_i^{k+1} (\mathbf{c} - \sum_{j \neq i} A_j \mathbf{x}_j^k) + A_i^T \boldsymbol{\lambda}^k + \boldsymbol{\Gamma}_i^{k+1} \mathbf{x}_i^k - \nabla f_i(\mathbf{x}_i^k)) + \boldsymbol{\Delta}_i^{k+1}. \quad (\text{J.24})$$

We then derive the expression of  $\nabla f(\mathbf{x}_i^k)$  as follows,

$$\nabla f(\mathbf{x}_i^k) = A_i^T \boldsymbol{\lambda}^k - A_i^T \boldsymbol{\rho}_i^{k+1} \sum_{j=1}^N A_j (\mathbf{x}_j^k - \mathbf{x}_j^*) + \mathbf{D}_i (\mathbf{x}_i^k - \mathbf{x}_i^{k+1}) + \mathbf{D}_i \boldsymbol{\Delta}_i^{k+1}. \quad (\text{J.25})$$

It is noted that the only difference, when compared to (J.5), arises from the additional term  $\boldsymbol{\Delta}_i^{k+1}$ . Due to the strong convexity assumed, we conduct a similar reasoning as (J.19) and have the following inequality:

$$\begin{aligned} \frac{m_i}{2} \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 &\leq (\mathbf{x}_i^{k+1} - \mathbf{x}_i^*)^T (A_i^T (\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*) - A_i^T \boldsymbol{\rho}_i^{k+1} \sum_{j=1}^N A_j (\mathbf{x}_j^k - \mathbf{x}_j^*) \\ &\quad + \mathbf{D}_i (\mathbf{x}_i^k - \mathbf{x}_i^{k+1}) + \mathbf{D}_i \boldsymbol{\Delta}_i^{k+1}) + \frac{M_i}{2} \|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2. \end{aligned} \quad (\text{J.26})$$

By summing up over  $i$  from 1 to  $N$  on both sides of (J.26), we have

$$\begin{aligned} \sum_{i=1}^N \frac{m_i}{2} \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 &\leq \sum_{i=1}^N ((\mathbf{x}_i^{k+1} - \mathbf{x}_i^*)^T (A_i^T (\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*) - A_i^T \boldsymbol{\rho}_i^{k+1} \sum_{j=1}^N A_j (\mathbf{x}_j^k - \mathbf{x}_j^*) \\ &\quad + \mathbf{D}_i (\mathbf{x}_i^k - \mathbf{x}_i^{k+1}) + \mathbf{D}_i \boldsymbol{\Delta}_i^{k+1}) + \frac{M_i}{2} \|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2). \end{aligned} \quad (\text{J.27})$$

---

<sup>1</sup>In particular, when  $\lim_{k \rightarrow \infty} \boldsymbol{\Delta}_i^k \rightarrow \mathbf{0}$  for each  $i$ , i.e., a diminishing noise is utilized,  $\lim_{k \rightarrow \infty} R^k \rightarrow 0$ .

By moving the left hand side to the right hand side, and taking the term  $\mathbf{D}_i \Delta_i^{k+1}$  out of the summation, we have

$$\begin{aligned}
& \underbrace{\sum_{i=1}^N ((\mathbf{x}_i^{k+1} - \mathbf{x}_i^*)^T (A_i^T (\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*) - A_i^T \boldsymbol{\rho}_i^{k+1} \sum_{j=1}^N A_j (\mathbf{x}_j^k - \mathbf{x}_j^*) + \mathbf{D}_i (\mathbf{x}_i^k - \mathbf{x}_i^{k+1})))}_{(1)} \\
& \quad \underbrace{\frac{M_i}{2} \|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2 - \frac{m_i}{2} \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2}_{(1)} + \underbrace{\sum_{i=1}^N (\mathbf{x}_i^{k+1} - \mathbf{x}_i^*)^T \mathbf{D}_i \Delta_i^{k+1}}_{(2)} \geq 0.
\end{aligned} \tag{J.28}$$

Therefore, the proof of Lemma J.0.2 is an analysis on term (1). From Lemma J.0.2, there exists  $p > 0$  for parameters within the admissible range defined in (J.16),  $\|\mathbf{u}^k - \mathbf{u}^*\|_G^2 \geq (1+p) \|\mathbf{u}^{k+1} - \mathbf{u}^*\|_G^2$ . Now combining both terms (1) and (2) to show the convergence rate, it still holds with almost the same reasoning except one difference. Due to the noise, the upper bound of  $\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^*\|^2$ , given before as (J.10), becomes

$$\begin{aligned}
& \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^*\|^2 \\
& \leq \frac{1}{\sigma_{i,\min}^2} \|\nabla f_i(\mathbf{x}_i^k) - \nabla f_i(\mathbf{x}_i^*) - A_i^T (\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1}) - \mathbf{D}_i (\mathbf{x}_i^k - \mathbf{x}_i^{k+1}) + A_i^T \boldsymbol{\rho}_i^{k+1} \sum_{j=1}^N A_j (\mathbf{x}_j^k - \mathbf{x}_j^*) + \mathbf{D}_i \Delta_i^{k+1}\|^2 \\
& \leq \frac{6}{\sigma_{i,\min}^2} (2M_i \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 + \|A_i (\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1})\|^2 + (D_i^2 + 2M_i) \|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2 + \\
& \quad \rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2 \|\sum_{j=1}^N A_j (\mathbf{x}_j^k - \mathbf{x}_j^{k+1})\|^2 + \rho_{i,\max}^{2(k+1)} \sigma_{i,\max}^2 \|\sum_{j=1}^N A_j (\mathbf{x}_j^{k+1} - \mathbf{x}_j^*)\|^2 + D_i^2 \|\Delta_i^{k+1}\|^2).
\end{aligned} \tag{J.29}$$

The changes in the constants here slightly change the range of  $p$  selection but do not affect the existence of  $p$  such that

$$\begin{aligned}
\|\mathbf{u}^k - \mathbf{u}^*\|_G^2 & \geq (1+p) \|\mathbf{u}^{k+1} - \mathbf{u}^*\|_G^2 - 2 \sum_{i=1}^N D_i (\mathbf{x}_i^{k+1} - \mathbf{x}_i^*)^T \Delta_i^{k+1} - \frac{6\delta}{\zeta N} \sum_{i=1}^N \frac{D_i^2}{\sigma_{i,\min}^2} \|\Delta_i^{k+1}\|^2 \\
& \geq (1 + (1 - \hat{\epsilon})p) \|\mathbf{u}^{k+1} - \mathbf{u}^*\|_G^2 - \sum_{i=1}^N \left( \frac{6\delta D_i^2}{\zeta N \sigma_{i,\min}^2} + \frac{D_i}{\hat{\epsilon} p} \right) \|\Delta_i^{k+1}\|^2,
\end{aligned} \tag{J.30}$$

where  $\hat{\epsilon} \in (0, 1)$ . Let  $a = \frac{1}{1+(1-\hat{\epsilon})p}$ , then <sup>2</sup>

$$\begin{aligned}
\|\mathbf{u}^{K+1} - \mathbf{u}^*\|_G^2 &\leq a \|\mathbf{u}^K + \mathbf{u}^*\|_G^2 + \sum_{i=1}^N \left( \frac{6pD_i^2}{\zeta N \sigma_{i,\min}^2} + \frac{D_i}{\hat{\epsilon}p} \right) a \|\Delta_i^{K+1}\|^2 \\
&\leq \dots \\
&\leq a^{K+1} \|\mathbf{u}^0 - \mathbf{u}^*\|_G^2 + \sum_{i=1}^N \left( \frac{6pD_i^2}{\zeta N \sigma_{i,\min}^2} + \frac{D_i}{\hat{\epsilon}p} \right) \sum_{k=1}^{K+1} a^k \|\Delta_i^{K+2-k}\|^2 \\
&= a^{K+1} \|\mathbf{u}^0 - \mathbf{u}^*\|_G^2 + R^{K+1}.
\end{aligned} \tag{J.33}$$

At last, we analyze the utility-privacy tradeoff. With respect to the  $\mathcal{B}_\infty$  sensitivity, based on the Laplace mechanism [54], let each coordinate of  $\Delta_i^k$  for any  $k$ , i.i.d. follow  $\text{Lap}(0, \frac{\epsilon D_i}{\mathcal{B}_\infty K d})$  for the composition across  $K$  iterations and  $d$  dimensions in pure  $\epsilon$ -LDP. Substituting the above form into (J.30), we have

$$\|\mathbf{u}^K - \mathbf{u}^*\|_G^2 = O(a^K \|\mathbf{u}^0 - \mathbf{u}^*\|_G^2 + 2Nd \left( \frac{\mathcal{B}_\infty K d}{\epsilon} \right)^2),$$

which is  $\tilde{O}(\frac{N \mathcal{B}_\infty^2 d^3}{\epsilon^2})$  due to the exponential decaying of the first term. Here we omit all other constants to avoid the tedious expression on  $\frac{1}{1-a}$ . Similarly, under  $(\epsilon, \delta)$ -LDP, with the strong composition [50], we only require that each coordinate of  $\Delta_i^k$  for any

---

<sup>2</sup> As a short comment, when  $\lim_{K \rightarrow \infty} \|\Delta_i^K\|^2 \rightarrow 0$ , there exists a constant  $C$  that  $\sum_{i=1}^N \left( \frac{6\delta D_i^2}{\zeta N \sigma_{i,\min}^2} + \frac{D_i}{\hat{\epsilon}p} \right) \|\Delta_i^K\|^2 \leq C \max_i \|\Delta_i^K\|^2$ . Therefore,

$$R^{K+1} \leq C \sum_{k=1}^{K+1} \max_i \|\Delta_i^k\|^2 a^{K+2-k}. \tag{J.31}$$

For any arbitrarily small constant  $z > 0$ , there exists  $k_0$ , such that for any  $K > 2k_0$ ,

$$C \sum_{k=1}^{k_0} \max_i \|\Delta_i^k\|^2 a^{K+1-k} \leq C a^{k_0} \sum_{k=1}^{k_0} \max_i \|\Delta_i^K\|^2 a^{k_0+1-k} < \frac{z}{2}.$$

On the other hand,  $\max_i \|\Delta_i^k\|^2 \leq \frac{z(1-c)}{2Cc}$  for any  $k > k_0$ . Therefore,

$$R^K \leq C \sum_{k=1}^{k_0} \max_i \|\Delta_i^k\|^2 a^{K+1-k} + C \sum_{k=k_0+1}^K \max_i \|\Delta_i^k\|^2 a^{K+1-k} \leq \frac{z}{2} + C \max_i \|\Delta_i^{k_0+1}\|^2 \sum_{k=k_0+1}^K a^{K-k} \leq z. \tag{J.32}$$

$k$ , i.i.d. follow  $\text{Lap}(0, O(\frac{\epsilon D_i}{\mathcal{B}_\infty \sqrt{Kd}}))$  and thus the utility loss is  $\tilde{O}(\frac{N \mathcal{B}_\infty^2 d^2}{\epsilon^2})$ .<sup>3</sup> □

---

<sup>3</sup>Note that in relaxed LDP,  $\delta$  is assumed as a constant and  $\epsilon$  is sufficiently small. Thus we drop the  $\log(\frac{1}{\delta})$  term.



# Bibliography

- [1] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.
- [2] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, pages 4424–4434, 2017.
- [3] Naman Agarwal, Ananda Theertha Suresh, Felix Xinnan X Yu, Sanjiv Kumar, and Brendan McMahan. cpsgd: Communication-efficient and differentially-private distributed sgd. In *Advances in Neural Information Processing Systems*, pages 7575–7586, 2018.
- [4] Wei Shi, Qing Ling, Kun Yuan, Gang Wu, and Wotao Yin. On the linear convergence of the admm in decentralized consensus optimization. *IEEE Trans. Signal Processing*, 62(7):1750–1761, 2014.
- [5] Tsung-Hui Chang, Mingyi Hong, and Xiangfeng Wang. Multi-agent distributed optimization via inexact consensus admm. *IEEE Trans. Signal Processing*, 63(2):482–497, 2015.
- [6] Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.
- [7] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48, 2009.
- [8] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.
- [9] Prateek Jain and Abhradeep Thakurta. Differentially private learning with kernels. *Journal of Machine Learning Research*, 2013.
- [10] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

- [11] Kunal Talwar, Abhradeep Guha Thakurta, and Li Zhang. Nearly optimal private lasso. In *Advances in Neural Information Processing Systems*, pages 3025–3033, 2015.
- [12] Di Wang, Marco Gaboardi, and Jinhui Xu. Empirical risk minimization in non-interactive local differential privacy revisited. In *Advances in Neural Information Processing Systems*, pages 973–982, 2018.
- [13] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, pages 2722–2731, 2017.
- [14] Di Wang and Jinhui Xu. Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. In *AAAI Conference on Artificial Intelligence*, 2019.
- [15] Katrina Ligett, Seth Neel, Aaron Roth, Bo Waggoner, and Steven Z Wu. Accuracy first: Selecting a differential privacy level for accuracy constrained erm. In *Advances in Neural Information Processing Systems*, pages 2566–2576, 2017.
- [16] Dong Su, Jianneng Cao, Ninghui Li, Elisa Bertino, and Hongxia Jin. Differentially private k-means clustering. In *Proceedings of the sixth ACM conference on data and application security and privacy*, pages 26–37. ACM, 2016.
- [17] Mikko Heikkilä, Eemil Lagerspetz, Samuel Kaski, Kana Shimizu, Sasu Tarkoma, and Antti Honkela. Differentially private bayesian learning on distributed data. In *Advances in Neural Information Processing Systems*, pages 3226–3235, 2017.
- [18] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private testing of identity and closeness of discrete distributions. In *Advances in Neural Information Processing Systems*, pages 6878–6891, 2018.
- [19] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321. ACM, 2015.
- [20] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.
- [21] Bargav Jayaraman, Lingxiao Wang, David Evans, and Quanquan Gu. Distributed learning without distress: Privacy-preserving empirical risk minimization. In *Advances in Neural Information Processing Systems*, pages 6346–6357, 2018.
- [22] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Secure multi-party differential privacy. In *Advances in neural information processing systems*, pages 2008–2016, 2015.



- [23] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems*, pages 6277–6287, 2018.
- [24] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 521–532. IEEE, 2018.
- [25] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.
- [26] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. In *Advances in neural information processing systems*, pages 2879–2887, 2014.
- [27] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy loss in apple’s implementation of differential privacy on macos 10.12. 2017.
- [28] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pages 3571–3580, 2017.
- [29] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.
- [30] Matthew Joseph, Aaron Roth, Jonathan Ullman, and Bo Waggoner. Local differential privacy for evolving data. In *Advances in Neural Information Processing Systems*, pages 2375–2384, 2018.
- [31] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1, 2012.
- [32] Prateek Jain and Abhradeep Guha Thakurta. (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, pages 476–484, 2014.
- [33] Shiva Prasad Kasiviswanathan and Hongxia Jin. Efficient private empirical risk minimization for high-dimensional learning. In *International Conference on Machine Learning*, pages 488–497, 2016.
- [34] Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 58–77. IEEE, 2017.

- [35] Di Wang, Adam Smith, and Jinhui Xu. Noninteractive locally private learning of linear models via polynomial approximations. *Proceedings of Machine Learning Research vol*, 98:1–20, 2019.
- [36] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 5330–5340, 2017.
- [37] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.
- [38] Ermin Wei and Asuman Ozdaglar. Distributed alternating direction method of multipliers. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5445–5450. IEEE, 2012.
- [39] Andreea B Alexandru, Konstantinos Gatsis, Yasser Shoukry, Sanjit A Seshia, Paulo Tabuada, and George J Pappas. Cloud-based quadratic optimization with partially homomorphic encryption. *arXiv preprint arXiv:1809.02267*, 2018.
- [40] S. Han, L W. K. Ng, Wan, and V. C. S. Lee. Privacy-preserving gradient-descent methods. *IEEE Transactions on Information Forensics and Security*, 22(6):884–899, 2010.
- [41] Chunlei Zhang, Muaz Ahmad, and Yongqiang Wang. Admm based privacy-preserving decentralized optimization. *IEEE Transactions on Information Forensics and Security*, 14(3):565–580, 2019.
- [42] Tao Zhang and Quanyan Zhu. Dynamic differential privacy for admm-based distributed classification learning. *IEEE Transactions on Information Forensics and Security*, 12(1):172–187, 2017.
- [43] Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. Improving the privacy and accuracy of ADMM-based distributed algorithms. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5796–5805, 10–15 Jul 2018.
- [44] Yuanxiong Guo and Yanmin Gong. Practical collaborative learning for crowdsensing in the internet of things with differential privacy. In *2018 IEEE Conference on Communications and Network Security (CNS)*, pages 1–9. IEEE, 2018.
- [45] Z. Huang, S. Mitra, and N. Vaidya. Differentially private distributed optimization. In *Proceedings of the 2015 International Conference on Distributed Computing and Networking*, volume 4 of *Proceedings of Machine Learning Research*. ACM, 2015.
- [46] S. Han, U. Topcu, and G. J. Pappas. Differentially private distributed constrained optimization. *IEEE Transactions on Automatic Control*, 62(1):50–64, 2017.

- [47] Y. Lou, L. Yu, S. Wang, and P. Yi. Privacy preservation in distributed subgradient optimization algorithms. *IEEE Transactions on Cybernetics*, 48(7):2154–2165, 2018.
- [48] Jiahao Ding, Yanmin Gong, Miao Pan, and Zhu Han. Optimal differentially private admm for distributed machine learning. *arXiv preprint arXiv:1901.02094*, 2019.
- [49] Ali Makhdoumi and Asuman Ozdaglar. Convergence rate of distributed admm over networks. *IEEE Transactions on Automatic Control*, 62(10):5082–5095, 2017.
- [50] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017.
- [51] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)*, 60(2):12, 2013.
- [52] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019.
- [53] Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 375–403. Springer, 2019.
- [54] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, volume 7, pages 94–103, 2007.
- [55] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.